ETH Lecture 401-2673-00L Numerical Methods for CSE

Numerical Methods for Computational Science and Engineering

Prof. R. Hiptmair, SAM, ETH Zurich

(with contributions from Prof. P. Arbenz and Dr. V. Gradinaru)

Autumn Term 2025, Version of November 23, 2025
(C) Seminar für Angewandte Mathematik, ETH Zürich

Link to the current version of this lecture document

Always under construction!

The online version will always be work in progress and subject to change.

(Nevertheless, structure and main contents can be expected to be stable)

Do not print before the end of term!

Contents

0 Introduction **8**
- 0.1 Course Fundamentals . . . 8
  - 0.1.1 Focus of this Course . . . 8
  - 0.1.2 Goals . . . 11
  - 0.1.3 Literature . . . 11
- 0.2 Teaching Style and Model . . . 12
  - 0.2.1 Flipped Classroom . . . 12
    - 0.2.1.1 Course Videos . . . 12
    - 0.2.1.2 Following the Course . . . 23
  - 0.2.2 Clarifications and Frank Words . . . 24
  - 0.2.3 Requests . . . 25
  - 0.2.4 Assignments . . . 27
  - 0.2.5 Information on Examinations . . . 28
    - 0.2.5.1 For the Course 401-2673-00L Numerical Methods for CSE (BSc CSE) . . . 28
    - 0.2.5.2 For the Course 401-0663-00L Numerical Methods for CS (BSc Informatik) 29
- 0.3 Programming in C++ . . . 30
  - 0.3.1 Function Arguments and Overloading . . . 30
  - 0.3.2 Templates . . . 32
  - 0.3.3 Function Objects and Lambda Functions . . . 33
  - 0.3.4 Multiple Return Values . . . 36
  - 0.3.5 A Vector Class . . . 37
  - 0.3.6 Complex numbers in C++ . . . 48
- 0.4 Prerequisite Mathematical Knowledge . . . 50
  - 0.4.1 Basics . . . 50
  - 0.4.2 Complex Numbers . . . 51
  - 0.4.3 Trigonometric Functions . . . 51
  - 0.4.4 Linear Algebra and Analysis . . . 51

1 Computing with Matrices and Vectors **53**
- 1.1 Fundamentals . . . 54
  - 1.1.1 Notations . . . 54
  - 1.1.2 Classes of Matrices . . . 56
- 1.2 Software and Libraries . . . 58
  - 1.2.1 EIGEN . . . 58
  - 1.2.2 PYTHON . . . 64
  - 1.2.3 (Dense) Matrix Storage Formats . . . 65
- 1.3 Basic Linear Algebra Operations . . . 70
  - 1.3.1 Elementary Matrix-Vector Calculus . . . 70
  - 1.3.2 BLAS – Basic Linear Algebra Subprograms . . . 76
- 1.4 Computational Effort . . . 82

1.4.1 (Asymptotic) Computational Complexity . . . 83
1.4.2 Cost of Basic Linear-Algebra Operations . . . 84
1.4.3 Improving Complexity in Numerical Linear Algebra: Some Tricks . . . 86
1.5 Machine Arithmetic and Consequences . . . 91
1.5.1 Experiment: Loss of Orthogonality . . . 91
1.5.2 Machine Numbers . . . 94
1.5.3 Roundoff Errors . . . 97
1.5.4 Cancellation . . . 102
1.5.5 Numerical Stability . . . 118

2 Direct Methods for (Square) Linear Systems of Equations 126
2.1 Introduction: Linear Systems of Equations (LSE) . . . 127
2.2 Theory: Linear Systems of Equations (LSE) . . . 130
2.2.1 LSE: Existence and Uniqueness of Solutions . . . 130
2.2.2 Sensitivity/Conditioning of Linear Systems . . . 131
2.3 Gaussian Elimination (GE) . . . 135
2.3.1 Basic Algorithm . . . 135
2.3.2 LU-Decomposition . . . 143
2.3.3 Pivoting . . . 151
2.4 Stability of Gaussian Elimination . . . 158
2.5 Survey: Elimination Solvers for Linear Systems of Equations . . . 165
2.6 Exploiting Structure when Solving Linear Systems . . . 170
2.7 Sparse Linear Systems . . . 179
2.7.1 Sparse Matrix Storage Formats . . . 180
2.7.2 Sparse Matrices in EIGEN . . . 183
2.7.3 Direct Solution of Sparse Linear Systems of Equations . . . 191
2.7.4 LU-Factorization of Sparse Matrices . . . 194
2.7.5 Banded Matrices [DR08, Sect. 3.7] . . . 199
2.8 Stable Gaussian Elimination Without Pivoting . . . 206

3 Direct Methods for Linear Least Squares Problems 214
3.0.1 Overdetermined Linear Systems of Equations: Examples . . . 215
3.1 Least Squares Solution Concepts . . . 218
3.1.1 Least Squares Solutions: Definition . . . 219
3.1.2 Normal Equations . . . 222
3.1.3 Moore-Penrose Pseudoinverse . . . 228
3.1.4 Sensitivity of Least Squares Problems . . . 230
3.2 Normal Equation Methods [DR08, Sect. 4.2], [Han02, Ch. 11] . . . 232
3.3 Orthogonal Transformation Methods [DR08, Sect. 4.4.2] . . . 236
3.3.1 Transformation Idea . . . 236
3.3.2 Orthogonal/Unitary Matrices . . . 238
3.3.3 QR-Decomposition [Han02, Sect. 13], [Gut09, Sect. 7.3] . . . 238
3.3.3.1 QR-Decomposition: Theory . . . 239
3.3.3.2 Computation of QR-Decomposition . . . 242
3.3.3.3 QR-Decomposition: Stability . . . 251
3.3.3.4 QR-Decomposition in EIGEN . . . 252
3.3.4 QR-Based Solver for Linear Least Squares Problems . . . 254
3.3.5 Modification Techniques for QR-Decomposition . . . 259
3.3.5.1 Rank-1 Modifications . . . 259
3.3.5.2 Adding a Column . . . 261
3.3.5.3 Adding a Row . . . 264
3.4 Singular Value Decomposition (SVD) . . . 266

3.4.1 SVD: Definition and Theory . . . 266
3.4.2 SVD in EIGEN . . . 271
3.4.3 Solving General Least-Squares Problems by SVD . . . 274
3.4.4 SVD-Based Optimization and Approximation . . . 277
3.4.4.1 Norm-Constrained Extrema of Quadratic Forms . . . 277
3.4.4.2 Best Low-Rank Approximation . . . 280
3.4.4.3 Principal Component Data Analysis (PCA) . . . 286
3.5 Total Least Squares . . . 298
3.6 Constrained Least Squares . . . 299
3.6.1 Solution via Lagrangian Multipliers . . . 300
3.6.2 Solution via SVD . . . 302

4 Filtering Algorithms 305
4.1 Filters and Convolutions . . . 306
4.1.1 Discrete Finite Linear Time-Invariant Causal Channels/Filters . . . 306
4.1.2 LT-FIR Linear Mappings . . . 309
4.1.3 Discrete Convolutions . . . 312
4.1.4 Periodic Convolutions . . . 316
4.2 Discrete Fourier Transform (DFT) . . . 321
4.2.1 Diagonalizing Circulant Matrices . . . 321
4.2.2 Discrete Convolution via Discrete Fourier Transform . . . 329
4.2.3 Frequency filtering via DFT . . . 331
4.2.4 Real DFT . . . 338
4.2.5 Two-dimensional DFT . . . 340
4.2.6 Semi-discrete Fourier Transform [QSS00, Sect. 10.11] . . . 349
4.3 Fast Fourier Transform (FFT) . . . 359
4.4 Trigonometric Transformations . . . 368
4.4.1 Sine transform . . . 368
4.4.2 Cosine transform . . . 374
4.5 Toeplitz Matrix Techniques . . . 376
4.5.1 Matrices with Constant Diagonals . . . 376
4.5.2 Toeplitz Matrix Arithmetic . . . 379
4.5.3 The Levinson Algorithm . . . 380

5 Data Interpolation and Data Fitting in 1D 384
5.1 Abstract Interpolation (AI) . . . 384
5.2 Global Polynomial Interpolation . . . 392
5.2.1 Uni-Variate Polynomials . . . 392
5.2.2 Polynomial Interpolation: Theory . . . 394
5.2.3 Polynomial Interpolation: Algorithms . . . 398
5.2.3.1 Multiple evaluations . . . 398
5.2.3.2 Single evaluation . . . 401
5.2.3.3 Extrapolation to Zero . . . 404
5.2.3.4 Newton Basis and Divided Differences . . . 408
5.2.4 Polynomial Interpolation: Sensitivity . . . 414
5.3 Shape-Preserving Interpolation . . . 418
5.3.1 Shape Properties of Functions and Data . . . 419
5.3.2 Piecewise Linear Interpolation . . . 421
5.3.3 Cubic Hermite Interpolation . . . 423
5.3.3.1 Definition and Algorithms . . . 423
5.3.3.2 Local Monotonicity-Preserving Hermite Interpolation . . . 426
5.4 Splines . . . 430

5.4.1 Spline Function Spaces . . . 430
5.4.2 Cubic-Spline Interpolation . . . 431
5.4.3 Structural Properties of Cubic Spline Interpolants . . . 436
5.4.4 Shape Preserving Spline Interpolation . . . 440
5.5 Algorithms for Curve Design . . . 445
5.5.1 CAD Task: Curves from Control Points . . . 445
5.5.2 Bezier Curves . . . 448
5.5.3 Spline Curves . . . 452
5.6 Trigonometric Interpolation . . . 456
5.6.1 Trigonometric Polynomials . . . 456
5.6.2 Reduction to Lagrange Interpolation . . . 457
5.6.3 Equidistant Trigonometric Interpolation . . . 459
5.7 Least Squares Data Fitting . . . 464

6 Approximation of Functions in 1D 473
6.1 Introduction . . . 473
6.2 Approximation by Global Polynomials . . . 476
6.2.1 Polynomial Approximation: Theory . . . 477
6.2.2 Error Estimates for Polynomial Interpolation . . . 483
6.2.2.1 Convergence of Interpolation Errors . . . 483
6.2.2.2 Interpolands of Finite Smoothness . . . 487
6.2.2.3 Analytic Interpolands . . . 492
6.2.3 Chebychev Interpolation . . . 499
6.2.3.1 Motivation and Definition . . . 500
6.2.3.2 Chebychev Interpolation Error Estimates . . . 503
6.2.3.3 Chebychev Interpolation: Computational Aspects . . . 510
6.3 Mean Square Best Approximation . . . 515
6.3.1 Abstract Theory . . . 515
6.3.1.1 Mean Square Norms . . . 515
6.3.1.2 Normal Equations . . . 516
6.3.1.3 Orthonormal Bases . . . 518
6.3.2 Polynomial Mean Square Best Approximation . . . 520
6.4 Uniform Best Approximation . . . 526
6.5 Approximation by Trigonometric Polynomials . . . 530
6.5.1 Approximation by Trigonometric Interpolation . . . 531
6.5.2 Trigonometric Interpolation Error Estimates . . . 532
6.5.3 Trigonometric Interpolation of Analytic Periodic Functions . . . 539
6.6 Approximation by Piecewise Polynomials . . . 545
6.6.1 Piecewise Polynomial Lagrange Interpolation . . . 546
6.6.2 Cubic Hermite Interpolation: Error Estimates . . . 549
6.6.3 Cubic Spline Interpolation: Error Estimates [Han02, Ch. 47] . . . 551

7 Numerical Quadrature 555
7.1 Introduction . . . 555
7.2 Quadrature Formulas – Quadrature Rules . . . 557
7.3 Polynomial Quadrature Formulas . . . 561
7.4 Gauss Quadrature . . . 564
7.4.1 Order of a Quadrature Rule . . . 564
7.4.2 Maximal-Order Quadrature Rules . . . 567
7.4.3 Quadrature Error Estimates . . . 575
7.5 Composite Quadrature . . . 580
7.6 Adaptive Quadrature . . . 588

8 Iterative Methods for Non-Linear Systems of Equations 598
8.1 Introduction . . . 598
8.2 Iterative Methods . . . 601
8.2.1 Fundamental Concepts . . . 601
8.2.2 Speed of Convergence . . . 605
8.2.3 Termination Criteria/Stopping Rules . . . 610
8.3 Fixed-Point Iterations . . . 614
8.3.1 Consistent Fixed-Point Iterations . . . 615
8.3.2 Convergence of Fixed-Point Iterations . . . 616
8.4 Finding Zeros of Scalar Functions . . . 623
8.4.1 Bisection . . . 623
8.4.2 Model Function Methods . . . 625
8.4.2.1 Newton Method in the Scalar Case . . . 625
8.4.2.2 Special One-Point Methods . . . 629
8.4.2.3 Multi-Point Methods . . . 633
8.4.3 Asymptotic Efficiency of Iterative Methods for Zero Finding . . . 638
8.5 Newton's Method in $\mathbb{R}^n$ . . . 642
8.5.1 The Newton Iteration . . . 642
8.5.2 Convergence of Newton's Method . . . 654
8.5.3 Termination of Newton Iteration . . . 657
8.5.4 Damped Newton Method . . . 659
8.6 Quasi-Newton Method . . . 663
8.7 Non-linear Least Squares [DR08, Ch. 6] . . . 670
8.7.1 (Damped) Newton Method . . . 673
8.7.2 Gauss-Newton Method . . . 675
8.7.3 Trust Region Method (Levenberg-Marquardt Method) . . . 679

9 Computation of Eigenvalues and Eigenvectors 682
9.1 Theory of eigenvalue problems . . . 685
9.2 "Direct" Eigensolvers . . . 687
9.3 Power Methods . . . 690
9.3.1 Direct power method . . . 690
9.3.2 Inverse Iteration [DR08, Sect. 7.6], [QSS00, Sect. 5.3.2] . . . 697
9.3.3 Preconditioned inverse iteration (PINVIT) . . . 707
9.3.4 Subspace iterations . . . 710
9.3.4.1 Orthogonalization . . . 714
9.3.4.2 Ritz projection . . . 717
9.4 Krylov Subspace Methods . . . 721

10 Krylov Methods for Linear Systems of Equations 733
10.1 Descent Methods [QSS00, Sect. 4.3.3] . . . 734
10.1.1 Quadratic minimization context . . . 734
10.1.2 Abstract steepest descent . . . 735
10.1.3 Gradient method for s.p.d. linear system of equations . . . 736
10.1.4 Convergence of the gradient method . . . 737
10.2 Conjugate gradient method (CG) [Han02, Ch. 9], [DR08, Sect. 13.4], [QSS00, Sect. 4.3.4] . . . 741
10.2.1 Krylov spaces . . . 742
10.2.2 Implementation of CG . . . 743
10.2.3 Convergence of CG . . . 746
10.3 Preconditioning [DR08, Sect. 13.5], [Han02, Ch. 10], [QSS00, Sect. 4.3.5] . . . 750
10.4 Survey of Krylov Subspace Methods . . . 756
10.4.1 Minimal residual methods . . . 756

10.4.2 Iterations with short recursions [QSS00, Sect. 4.5] . . . 757

11 Numerical Integration – Single Step Methods 761
11.1 Initial-Value Problems (IVPs) for Ordinary Differential Equations (ODEs) . . . 762
11.1.1 Ordinary Differential Equations (ODEs) . . . 762
11.1.2 Mathematical Modeling with Ordinary Differential Equations: Examples . . . 764
11.1.3 Theory of Initial-Value-Problems (IVPs) . . . 769
11.1.4 Evolution Operators . . . 773
11.2 Introduction: Polygonal Approximation Methods . . . 776
11.2.1 Explicit Euler method . . . 777
11.2.2 Implicit Euler method . . . 779
11.2.3 Implicit midpoint method . . . 781
11.3 General Single-Step Methods . . . 783
11.3.1 Definition . . . 783
11.3.2 (Asymptotic) Convergence of Single-Step Methods . . . 787
11.4 Explicit Runge-Kutta Single-Step Methods (RKSSMs) . . . 796
11.5 Adaptive Stepsize Control . . . 803
11.5.1 The Need for Timestep Adaptation . . . 803
11.5.2 Local-in-Time Stepsize Control . . . 805
11.5.3 Embedded Runge-Kutta Methods . . . 812

12 Single-Step Methods for Stiff Initial-Value Problems 819
12.1 Model Problem Analysis . . . 820
12.2 Stiff Initial-Value Problems . . . 834
12.3 Implicit Runge-Kutta Single-Step Methods . . . 840
12.3.1 The Implicit Euler Method for Stiff IVPs . . . 840
12.3.2 Collocation Single-Step Methods . . . 841
12.3.3 General Implicit Runge-Kutta Single-Step Methods (RK-SSMs) . . . 845
12.3.4 Model Problem Analysis for Implicit Runge-Kutta Single-Step Methods (IRK-SSMs) 847
12.4 Semi-Implicit Runge-Kutta Methods . . . 855
12.5 Splitting Methods . . . 858

Index 865
Symbols . . . 878
Examples . . . 880
Glossary . . . 888

Chapter 0

Introduction

0.1 Course Fundamentals

0.1.1 Focus of this Course

Emphasis is put

- ▷ on **algorithms** (principles, computational cost, scope, and limitations),
- ▷ on (efficient and stable) **implementation** in **C++** based on the numerical linear algebra template library EIGEN, a **D**omain **S**pecific **L**anguage (DSL) embedded into C++.
- ▷ on **numerical experiments** (design and interpretation).

§0.1.1.1 (Aspects outside the scope of this course) No emphasis will be put on

- theory and proofs (unless essential for derivation and understanding of algorithms).

  ☞ 401-3651-00L Numerical Methods for Elliptic and Parabolic Partial Differential Equations
  401-3652-00L Numerical Methods for Hyperbolic Partial Differential Equations
  (both courses offered in BSc Mathematics)

- hardware aware implementation (cache hierarchies, CPU pipelining, vectorization, etc.)

  ☞ 263-0007-00L Advanced System Lab (How To Write Fast Numerical Code, Prof. M. Püschel, D-INFK)

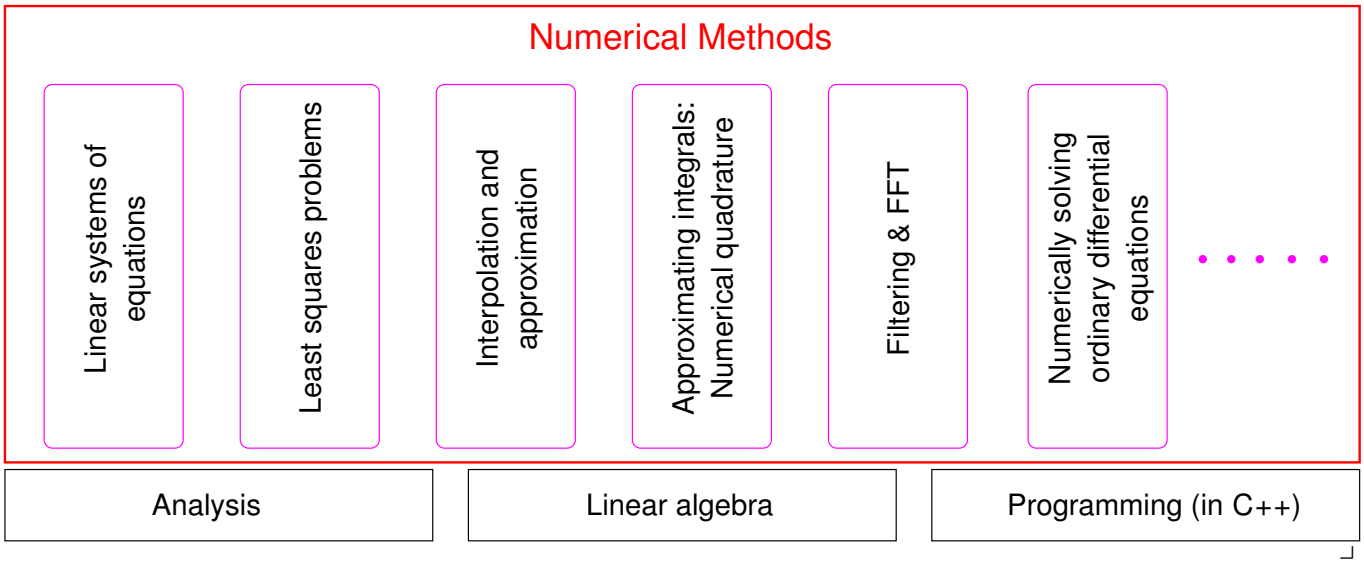
- issues of high-performance computing (HPC, shard and distributed memory parallelisation, vectorization)

  ☞ 151-0107-20L High Performance Computing for Science and Engineering (HPCSE, Prof. P. Koumoutsakos, D-MAVT)
  263-2800-00L Design of Parallel and High-Performance Computing (Prof. T. Höfler, D-INFK)

However, note that these other courses partly rely on knowledge of elementary numerical methods, which is covered in this course. ⌟

Contents

§0.1.1.2 (Prequisites) This course will take for granted basic knowledge of linear algebra, calculus, and programming, that you should have acquired during your first year at ETH.



§0.1.1.3 (Numerical methods: A motley toolbox)

This course discusses **elementary numerical methods and techniques**

They are vastly different in terms of ideas, design, analysis, and scope of application. They are the items in a **toolbox**, some only loosely related by the common purpose of being building blocks for codes for numerical simulation.

Do not expect much coherence between the chapters of this course!

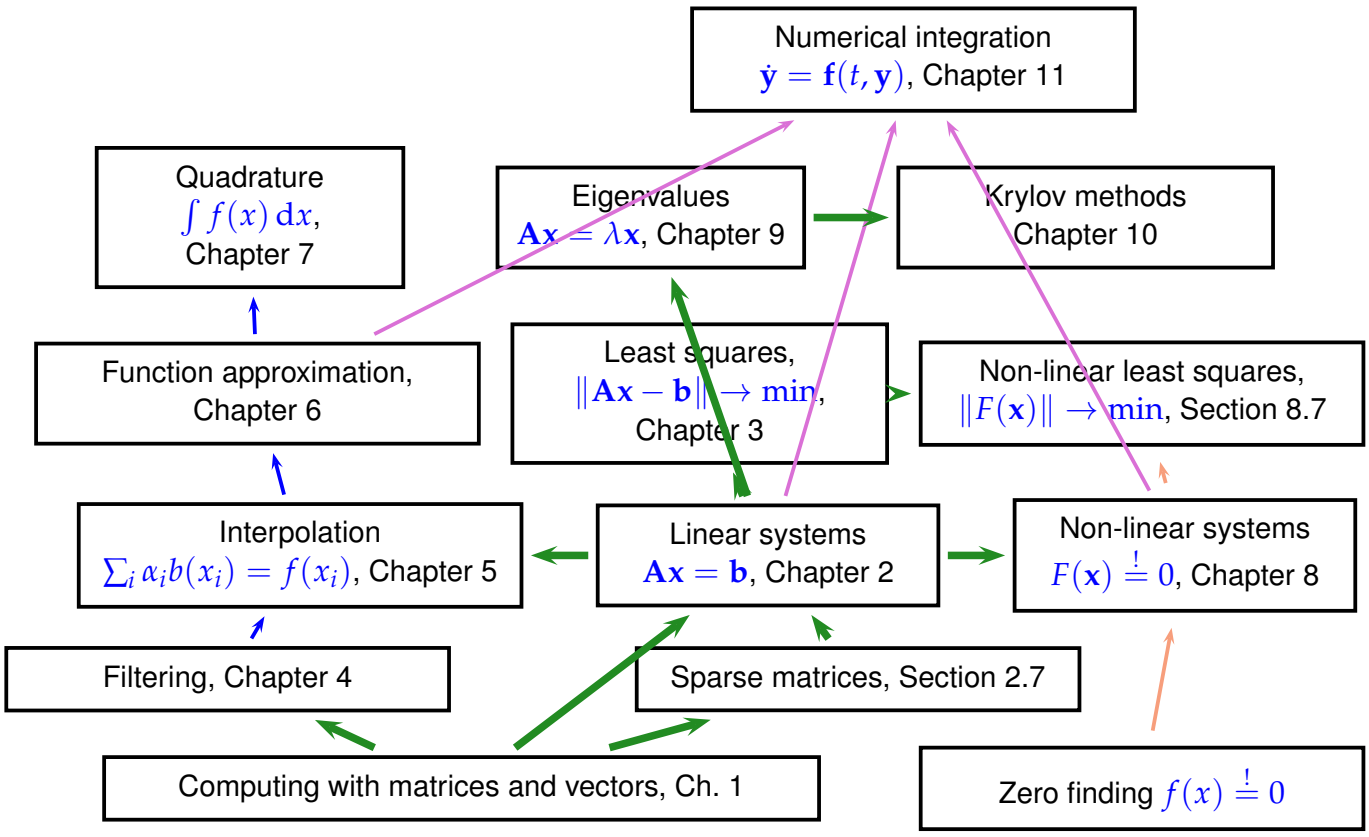
Fig. 1

A purpose-oriented notion of "Numerical methods for CSE":

A: "Stop putting a hammer, a level, and duct tape in one box! They have nothing to do with each other!"

B: "I might need any of these tools when fixing something about the house"

§0.1.1.4 (Dependencies of topics) Despite the diverse nature of the individual topics covered in this course, some depend on others for providing essential building blocks. The following directed graph tries to capture these relationships. The arrows have to be read as "uses results or algorithms of".



Any one-semester course "Numerical methods for CSE" will cover only selected chapters and sections of this document. Only topics addressed in **class** or in **homework problems** will be relevant for exams!

§0.1.1.5 (Relevance of this course) I am a student of computer science. After the exam, may I safely forget everything I have learned in this mandatory "numerical methods" course? **No**, because it is highly likely that other courses or projects will rely on the contents of this course:

singular value decomposition least squares	⟹ Computational statistics, machine learning
function approximation numerical quadrature numerical integration	⟹ machine learning, Numerical methods for PDEs
interpolation least squares	⟹ Computer graphics
eigensolvers sparse linear systems	⟹ Graph theoretic algorithms
numerical integration	⟹ Computer animation, robotics

and many more applications of fundamental numerical methods . . ..

Hardly anyone will need everything covered in this course, but *most of you will need something*.

⌟

0.1.2 Goals

This course is meant to impart

- knowledge of some fundamental algorithms forming the basis of numerical simulations,
- familiarity with essential terms in numerical mathematics and the techniques used for the analysis of numerical algorithms
- the skill to choose the appropriate numerical methods for concrete problems,
- the ability to interpret numerical results,
- proficiency in implementing numerical algorithms efficiently in C++, using numerical libraries.

Indispensable:	Learning by doing (➔ exercises)

0.1.3 Literature

Parts of the following textbooks may be used as supplementary reading for this course. References to relevant sections will be provided in the course material.

Studying extra literature is not important for following this course!

- [AG11] U. ASCHER AND C. GREIF, *A First Course in Numerical Methods*, SIAM, Philadelphia, 2011.

  Comprehensive introduction to numerical methods with an algorithmic focus based on MATLAB. (Target audience: students of engineering subjects)

- [DR08] W. DAHMEN AND A. REUSKEN, *Numerik für Ingenieure und Naturwissenschaftler*, Springer, Heidelberg, 2006.

  Good reference for large parts of this course; provides a lot of simple examples and lucid explanations, but also rigorous mathematical treatment.
  (Target audience: undergraduate students in science and engineering)
  Available for download as PDF

- [Han02] M. HANKE-BOURGEOIS, *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens*, Mathematische Leitfäden, B.G. Teubner, Stuttgart, 2002.

  Gives detailed description and mathematical analysis of algorithms and relies on MATLAB. Profound treatment of theory way beyond the scope of this course. (Target audience: undergraduates in mathematics)

- [QSS00] A. QUARTERONI, R. SACCO, AND F. SALERI, *Numerical mathematics*, vol. 37 of Texts in Applied Mathematics, Springer, New York, 2000.

Classical introductory numerical analysis text with many examples and detailed discussion of algorithms. (Target audience: undergraduates in mathematics and engineering)
Can be obtained from website.

- [DH03] P. Deuflhard and A. Hohmann, *Numerische Mathematik. Eine algorithmisch orientierte Einführung*, DeGruyter, Berlin, 1 ed., 1991.

  Modern discussion of numerical methods with profound treatment of theoretical aspects (Target audience: undergraduate students in mathematics).

- [GGK14]: W.. Gander, M.J. Gander, and F. Kwok, *Scientific Computing*, Text in Computational Science and Engineering, springer, 2014.

  Comprehensive treatment of elementary numerical methods with an algorithmic focus.

D-INFK maintains a webpage with links to some of these books.

Essential prerequisite for this course is a solid knowledge in linear algebra and calculus. Familiarity with the topics covered in the first semester courses is taken for granted, see

- [NS02] K. Nipp and D. Stoffer, *Lineare Algebra*, vdf Hochschulverlag, Zürich, 5 ed., 2002.
- [Gut09] M. Gutknecht, *Lineare algebra*, lecture notes, SAM, ETH Zürich, 2009, available online.
- [Str09] M. Struwe, *Analysis für Informatiker*. Lecture notes, ETH Zürich, 2009, available online.

0.2 Teaching Style and Model

0.2.1 Flipped Classroom

This course will depart from the usual academic teaching arrangement centering around classes taught by a lecturer addressing an audience in a lecture hall.

A flipped-classroom course

This course will follow the **flipped-classroom** paradigm:

Learning by **self-study** guided by

instruction videos tablet notes	lecture notes	interactive Q&A sessions	homeworks tutorial classes

All the course material will be published online through the course Moodle Page. All notes jotted down by the lecturer during the creation of videos or during the Q&A sessions will be made available as PDF.

0.2.1.1 Course Videos

In the flipped-classroom teaching model regular lectures will be replaced with pre-recorded videos. These videos are not commercial-grade clips, but resemble video recordings from a standard classroom setting; they convey the development of the material on a tablet accompanied by the lecturer's voice.

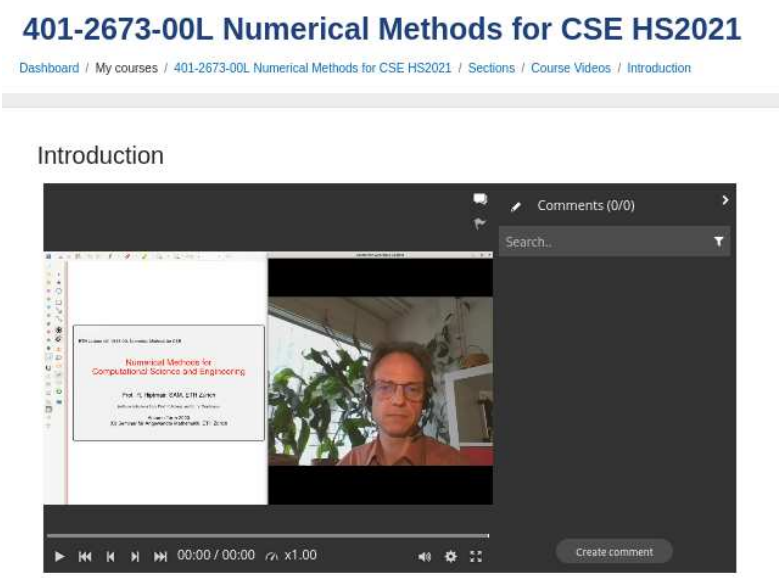


Fig. 2

The Videos will be published through

1. the course Moodle Page (see beside),
2. and as `.mp4`-files on PolyBox (password required).

Every video comes with a PDF containing the tablet notes taken during the creation of the video. However, the PDF may have been *corrected, updated, or supplemented* later.

§0.2.1.2 ("Pause" and "fast forward") Videos have *two big advantages*:

Fig. 3

You can stop a video at any time, whenever

- you need more time to think,
- you want to look up related information,
- you want to work for yourself.

Make use of this possibility!

The video portal also allows you to play the videos at *1.5× speed*. This can be useful, if the current topic is very clear to you. You can also *skip* entire *parts* using the scroll bar. The same functionality (fast playing and skipping) is offered by most video players, for instance the VLC media player. ⌟

§0.2.1.3 (Review questions) Most lecture units (corresponding to a video) are accompanied with a list of review questions. You should try to answer them off the top of your head *without consulting any written material* shortly after you have finished studying the unit .

In case you are utterly clueless about how to approach a review question, you probably need to refresh some of the unit's topics.

⌟

§0.2.1.4 (List of available tutorial videos) This is the list of available video tutorials as of November 23, 2025:

1. Video tutorial for Chapter 0 "Introduction": (16 minutes) Download link, tablet notes
2. Video tutorial for Section 1.1.1 "Notations and Classes of Matrices": (7 minutes) Download link, tablet notes

   → review questions 1.1.2.9
3. Video tutorial for Section 1.2.1 "EIGEN ": (11 minutes) Download link, tablet notes

→ review questions 1.2.1.14

4. Video tutorial for Section 1.2.3 "(Dense) Matrix Storage Formats": (10 minutes) Download link, tablet notes

→ review questions 1.2.3.11

5. Video tutorial for Section 1.4 "Computational Effort": (29 minutes) Download link, tablet notes

→ review questions 1.4.3.11

6. Video tutorial for Section 1.5 "Machine Arithmetic and Consequences": (16 minutes) Download link, tablet notes

→ review questions 1.5.3.18

7. Video tutorial for Section 1.5.4 "Cancellation": (22 minutes) Download link, tablet notes

→ review questions 1.5.4.33

8. Video tutorial for Section 1.5.5 "Numerical Stability": (17 minutes) Download link, tablet notes

→ review questions 1.5.5.23

9. Video tutorial for Section 2.1 & Section 2.2.1 "Introduction and Theory: Linear Systems of Equations (LSEs)": (6 minutes) Download link, tablet notes

→ review questions 2.2.1.7

10. Video tutorial for Ex. 2.1.0.3 "Nodal Analysis of Linear Electric Circuits": (8 minutes) Download link, tablet notes

→ review questions 2.1.0.8

11. Video tutorial for Section 2.2.2 "Sensitivity of Linear Systems": (15 minutes) Download link, tablet notes

→ review questions 2.2.2.12

12. Video tutorial for Section 2.3 & Section 2.5 "Gaussian Elimination": (17 minutes) Download link, tablet notes

→ review questions 2.3.2.21

13. Video tutorial for Section 2.6 "Exploiting Structure when Solving Linear Systems": (17 minutes) Download link, tablet notes

→ review questions 2.6.0.25

14. Video tutorial for Section 2.7.1 "Sparse Matrix Storage Formats": (10 minutes) Download link, tablet notes

→ review questions 2.7.1.5

15. Video tutorial for Section 2.7.2 "Sparse Matrices in EIGEN ": (6 minutes) Download link, tablet notes

→ review questions 2.7.2.17

16. Video tutorial for Section 2.7.3 "Direct Solution of Sparse Linear Systems of Equations": (10 minutes) Download link, tablet notes

→ review questions>2.7.3.7

17. Video tutorial for Section 3.0.1 "Overdetermined Linear Systems of Equations: Examples": (12 minutes) Download link, tablet notes

→ review questions 3.0.1.11

18. Video tutorial for Section 3.1.1 "Least Squares Solutions": (9 minutes) Download link, tablet notes

→ review questions 3.1.1.14

19. Video tutorial for Section 3.1.2 "Normal Equations": (16 minutes) Download link, tablet notes

→ review questions 3.1.2.23

20. Video tutorial for Section 3.1.3 "Moore-Penrose Pseudoinverse": (8 minutes) Download link, tablet notes

→ review questions 3.1.3.8

21. Video tutorial for Section 3.2 "Normal Equation Methods": (12 minutes) Download link, tablet notes

→ review questions 3.2.0.11

22. Video tutorial for Section 3.3 "Orthogonal Transformation Methods": (10 minutes) Download link, tablet notes

→ review questions 3.3.2.3

23. Video tutorial for Section 3.3.3.1 "QR-Decomposition: Theory": (11 minutes) Download link, tablet notes

→ review questions 3.3.3.8

24. Video tutorial for Section 3.3.3.2 & Section 3.3.3.4 "Computation of QR-Decomposition, QR-Decomposition in EIGEN ": (32 minutes) Download link, tablet notes

→ review questions 3.3.3.29

25. Video tutorial for Section 3.3.4 "QR-Based Solver for Linear Least Squares Problems": (9 minutes) Download link, tablet notes

→ review questions 3.3.4.8

26. Video tutorial for Section 3.3.5 "Modification Techniques for QR-Decomposition": (25 minutes) Download link, tablet notes

→ review questions 3.3.5.7

27. Video tutorial for Section 3.4.1 "Singular Value Decomposition: Definition and Theory": (13 minutes) Download link, tablet notes

→ review questions 3.4.1.15

28. Video tutorial for Section 3.4.2 "SVD in EIGEN ": (9 minutes) Download link, tablet notes

→ review questions 3.4.2.10

29. Video tutorial for Section 3.4.3 "Solving General Least-Squares Problems by SVD": (14 minutes) Download link, tablet notes

→ review questions 3.4.3.17

30. Video tutorial for Section 3.4.4.1 "Norm-Constrained Extrema of Quadratic Forms": (11 minutes) Download link, tablet notes

→ review questions 3.4.4.13

31. Video tutorial for Section 3.4.4.2 "Best Low-Rank Approximation": (13 minutes) Download link, tablet notes

→ review questions 3.4.4.25

32. Video tutorial for Section 3.4.4.3 "Principal Component Data Analysis (PCA)": (28 minutes) Download link, tablet notes

→ review questions 3.4.4.51

33. Video tutorial for Section 3.6 "Constrained Least Squares": (23 minutes) Download link, tablet notes

→ review questions 3.6.2.1

34. Video tutorial for Section 4.1.1 "Discrete Finite Linear Time-Invariant Causal Channels/Filters": (11 minutes) Download link, tablet notes

→ review questions 4.1.1.13

35. Video tutorial for Section 4.1.2 "LT-FIR Linear Mappings": (12 minutes) Download link, tablet notes

→ review questions 4.1.2.10

36. Video tutorial for Section 4.1.3 "Discrete Convolutions": (9 minutes) Download link, tablet notes

→ review questions 4.1.3.11

37. Video tutorial for Section 4.1.4 "Periodic Convolutions": (12 minutes) Download link, tablet notes

→ review questions 4.1.4.19

38. Video tutorial for Section 4.2.1 "Diagonalizing Circulant Matrices": (17 minutes) Download link, tablet notes

→ review questions 4.2.1.23

39. Video tutorial for Section 4.2.2 "Discrete Convolution via DFT": (7 minutes) Download link, tablet notes

→ review questions 4.2.2.6

40. Video tutorial for Section 4.2.3 "Frequency filtering via DFT": (20 minutes) Download link, tablet notes

→ review questions 4.2.3.11

41. Video tutorial for Section 4.2.5 "Two-Dimensional DFT": (20 minutes) Download link, tablet notes

→ review questions 4.2.5.24

42. Video tutorial for Section 4.3 "Fast Fourier Transform (FFT)": (16 minutes) Download link, tablet notes

→ review questions 4.3.0.14

43. Video tutorial for Section 4.5 "Toeplitz Matrix Techniques": (20 minutes) Download link, tablet notes

→ review questions 4.5.3.9

44. Video tutorial for Section 5.1 "Abstract Interpolation": (16 minutes) Download link, tablet notes

→ review questions 5.1.0.27

45. Video tutorial for Section 5.2.1 "Uni-Variate Polynomials": (7 minutes) Download link, tablet notes

→ review questions 5.2.1.8

46. Video tutorial for Section 5.2.2 "Polynomial Interpolation: Theory": (6 minutes) Download link, tablet notes

→ review questions 5.2.2.19

47. Video tutorial for Section 5.2.3 "Polynomial Interpolation: Algorithms": (18 minutes) Download link, tablet notes

→ review questions 5.2.3.14

48. Video tutorial for Section 5.2.3.3 "Extrapolation to Zero": (12 minutes) Download link, tablet notes

→ review questions 5.2.3.20

49. Video tutorial for Section 5.2.3.4 "Newton Basis and Divided Differences": (17 minutes) Download link, tablet notes

→ review questions 5.2.3.39

50. Video tutorial for Section 5.2.4 "Polynomial Interpolation: Sensitivity": (13 minutes) Download link, tablet notes

→ review questions 5.2.4.16

51. Video tutorial for Section 5.3 "Shape-Preserving Interpolation": (23 minutes) Download link, tablet notes

→ review questions 5.3.3.19

52. Video tutorial for Section 5.4.1 "Spline Function Spaces": (9 minutes) Download link, tablet notes

→ review questions 5.4.1.5

53. Video tutorial for Section 5.4.2 "Cubic Spline Interpolation": (14 minutes) Download link, tablet notes

→ review questions 5.4.2.16

54. Video tutorial for Section 5.4.3 "Structural Properties of Cubic Spline Interpolants": (12 minutes) Download link, tablet notes

→ review questions 5.4.3.11

55. Video tutorial for Section 5.6 "Trigonometric Interpolation": (14 minutes) Download link, tablet notes

→ review questions 5.6.3.10

56. Video tutorial for Section 5.7 "Least Squares Data Fitting": (13 minutes) Download link, tablet notes

→ review questions 5.7.0.31

57. Video tutorial for Section 6.1 "Approximation of Functions in 1D: Introduction": (7 minutes) Download link, tablet notes

→ review questions 6.1.0.9

58. Video tutorial for Section 6.2 "Polynomial Approximation: Theory": (13 minutes) Download link, tablet notes

→ review questions 6.2.1.28

59. Video tutorial for Section 6.2.2 "Error Estimates for Polynomial Interpolation": (12 minutes) Download link, tablet notes

→ review questions 6.2.2.14

60. Video tutorial for Section 6.2.2.2 "Error Estimates for Polynomial Interpolation: Interpolands of Finite Smoothness": (17 minutes) Download link, tablet notes

→ review questions 6.2.2.34

61. Video tutorial for Section 6.2.2.3 "Error Estimates for Polynomial Interpolation: Analytic Interpolands": (27 minutes) Download link, tablet notes

→ review questions 6.2.2.69

62. Video tutorial for Section 6.2.3.1 "Chebychev Interpolation: Motivation and Definition": (21 minutes) Download link, tablet notes

→ review questions 6.2.3.13

63. Video tutorial for Section 6.2.3.2 "Chebychev Interpolation Error Estimates": (14 minutes) Download link, tablet notes

→ review questions 6.2.3.30

64. Video tutorial for Section 6.2.3.3 "Chebychev Interpolation: Computational Aspects": (11 minutes) Download link, tablet notes

→ review questions 6.2.3.44

65. Video tutorial for Section 6.5.1 "Approximation by Trigonometric Interpolation": (5 minutes) Download link, tablet notes

→ review questions 6.5.1.6

66. Video tutorial for Section 6.5.2 "Trigonometric Interpolation Error Estimates": (14 minutes) Download link, tablet notes

→ review questions 6.5.2.26

67. Video tutorial for Section 6.5.3 "Trigonometric Interpolation of Analytic Periodic Functions": (16 minutes) Download link, tablet notes

→ review questions 6.5.3.18

68. Video tutorial for Section 6.6.1 "Piecewise Polynomial Lagrange Interpolation": (17 minutes) Download link, tablet notes

69. Video tutorial for Section 6.6.2 "Cubic Hermite and Spline Interpolation: Error Estimates": (10 minutes) Download link, tablet notes

70. Video tutorial for Section 7.1 "Numerical Quadrature: Introduction": (4 minutes) Download link, tablet notes

→ review questions 7.1.0.5

71. Video tutorial for Section 7.2 "Quadrature Formulas/Rules": (13 minutes) Download link, tablet notes

→ review questions 7.2.0.15

72. Video tutorial for Section 7.3 "Polynomial Quadrature Formulas": (9 minutes) Download link, tablet notes

→ review questions 7.3.0.12

73. Video tutorial for Section 7.4.1 "Order of a Quadrature Rule": (9 minutes) Download link, tablet notes

→ review questions 7.4.1.12

74. Video tutorial for Section 7.4.2 "Maximal-Order Quadrature Rules": (16 minutes) Download link, tablet notes

→ review questions 7.4.2.27

75. Video tutorial for Section 7.4.3 "(Gauss-Legendre) Quadrature Error Estimates": (18 minutes) Download link, tablet notes

→ review questions 7.4.3.16

76. Video tutorial for Section 7.5 "Composite Quadrature": (18 minutes) Download link, tablet notes

→ review questions 7.5.0.26

77. Video tutorial for Section 7.6 "Adaptive Quadrature": (13 minutes) Download link, tablet notes

→ review questions 7.6.0.20

78. Video tutorial for Section 8.1 "Iterative Methods for Non-Linear Systems of Equations: Introduction": (6 minutes) Download link, tablet notes

→ review questions 8.1.0.6

79. Video tutorial for Section 8.2.1 "Iterative Methods: Fundamental Concepts": (6 minutes) Download link, tablet notes

→ review questions 8.2.1.11

80. Video tutorial for Section 8.2.2 "Iterative Methods: Speed of Convergence": (15 minutes) Download link, tablet notes

→ review questions 8.2.2.16

81. Video tutorial for Section 8.2.3 "Iterative Methods: Termination Criteria/Stopping Rules": (14 minutes) Download link, tablet notes

→ review questions 8.2.3.10

82. Video tutorial for Section 8.3 "Fixed-Point Iterations": (12 minutes) Download link, tablet notes

→ review questions 8.3.2.21

83. Video tutorial for Section 8.4.1 "Finding Zeros of Scalar Functions: Bisection": (7 minutes) Download link, tablet notes

→ review questions 8.4.1.4

84. Video tutorial for Section 8.4.2.1 "Newton Method in the Scalar Case": (20 minutes) Download link, tablet notes

→ review questions 8.4.2.16

85. Video tutorial for Section 8.4.2.3 "Multi-Point Methods": (12 minutes) Download link, tablet notes

→ review questions 8.4.2.41

86. Video tutorial for Section 8.4.3 "Asymptotic Efficiency of Iterative Methods for Zero Finding": (10 minutes) Download link, tablet notes

→ review questions 8.4.3.15

87. Video tutorial for Section 8.5.1 "The Newton Iteration in $\mathbb{R}^n$ (I)": (10 minutes) Download link, tablet notes

→ review questions 8.5.1.46

88. Video tutorial for § 8.5.1.15 "Multi-dimensional Differentiation": (20 minutes) Download link, tablet notes

89. Video tutorial for Section 8.5.1 "The Newton Iteration in $\mathbb{R}^n$ (II)": (15 minutes) Download link, tablet notes

90. Video tutorial for Section 8.5.2 "Convergence of Newton's Method": (9 minutes) Download link, tablet notes

→ review questions 8.5.2.8

91. Video tutorial for Section 8.5.3 "Termination of Newton Iteration": (7 minutes) Download link, tablet notes

→ review questions 8.5.3.9

92. Video tutorial for Section 8.5.4 "Damped Newton Method": (11 minutes) Download link, tablet notes

→ review questions 8.5.4.8

93. Video tutorial for Section 8.6 "Quasi-Newton Method": (15 minutes) Download link, tablet notes

→ review questions 8.6.0.22

94. Video tutorial for Section 8.7 "Non-linear Least Squares": (7 minutes) Download link, tablet notes

→ review questions 8.7.0.10

95. Video tutorial for Section 8.7.1 "Non-linear Least Squares: (Damped) Newton Method": (13 minutes) Download link, tablet notes

→ review questions 8.7.1.9

96. Video tutorial for Section 8.7.2 "(Trust-region) Gauss-Newton Method": (13 minutes) Download link, tablet notes

→ review questions 8.7.3.3

97. Video tutorial for Section 11.1: Initial-Value Problems (IVPs) for Ordinary Differential Equations (ODEs): (35 minutes) Download link, tablet notes

→ review questions 11.1.4.8

98. Video tutorial for Section 11.2: Introduction: Polygonal Approximation Methods: (17 minutes) Download link, tablet notes

→ review questions 11.2.3.4

99. Video tutorial for Section 11.3: General Single-Step Methods: (14 minutes) Download link, tablet notes

→ review questions 11.3.1.17

100. Video tutorial for Section 11.3.2:(Asymptotic) Convergence of Single-Step Methods: (20 minutes) Download link, tablet notes

→ review questions 11.3.2.34

101. Video tutorial for Section 11.4: Explicit Runge-Kutta Single-Step Methods (RKSSMs): (27 minutes) Download link, tablet notes

→ review questions 11.4.0.20

102. Video tutorial for Section 11.5: Adaptive Stepsize Control: (32 minutes) Download link, tablet notes

→ review questions 11.5.3.10

103. Video tutorial for Section 12.1:Model Problem Analysis: (40 minutes) Download link, tablet notes

→ review questions 12.1.0.54

104. Video tutorial for Section 12.2: Stiff Initial-Value Problems: (24 minutes) Download link, tablet notes

→ review questions 12.2.0.17

105. Video tutorial for Section 12.3: Implicit Runge-Kutta Single-Step Methods: (50 minutes) Download link, tablet notes

→ review questions 12.3.4.23

106. Video tutorial for Section 12.4: Semi-Implicit Runge-Kutta Methods: (13 minutes) Download link, tablet notes

→ review questions 12.4.0.10

107. Video tutorial for Section 12.5: Splitting Methods: (21 minutes) Download link, tablet notes

→ review questions 12.5.0.14

⌟

Necessary corrections and updates of the lecture document will sometimes lead to changes in the numbering of paragraphs and formulas, which, of course, cannot be applied to the recorded videos.
However, these changes will be taken into account into the tablet notes supplied for every video.

0.2.1.2 Following the Course

Weekly study assignments

- For every week there is a list of course units and associated videos published on the course Moodle Page.
- The corresponding contents **must** be studied in that same week.

§0.2.1.6 (How to organize your learning)

☛ **Develop a routine**: Plan *fixed slots*, with a total duration of *four hours*, for studying for the course material in your weekly calendar. This does not include homework.

☛ **Choose a stable setting**, in which you can really concentrate (quiet area, headphones, coffee, etc.)

☛ **Take breaks**, when concentration is declining, usually after 20 to 45 minutes, but *avoid online distractions* during breaks.

You must not procrastinate!

Do not put off studying for this course. Dependencies between the topics will make it very hard to catch up.

§0.2.1.7 ("Personalized learning") The flipped classroom model allows students to pursue their preferred ways of studying. The following approaches can be tried.

- **Traditional**: You watch the assigned videos similar to attending a conventional classroom lecture. Afterwards digest the material based on the tablet notes and/or the lecture document. Finally, answer the review questions and look up more information in the lecture document.
- **Reading-centered**: You work through the unit reading the tablet notes, and, sometimes, related sections of the lecture document. You occasionally watch parts of the videos, in case some considerations and arguments have not become clear to you already.

Fig. 4

Collaborative studying is encouraged:

- You may watch course videos together with classmates.
- You may meet to discuss course units.
- You may solve homework problems in a group assigning different parts to different members.

☛ Explaining to others is a great way to deepen understanding.

☛ It is easy to sustain motivation and avoid distraction in a peer study group.

§0.2.1.8 (Question and Answer (Q&A) sessions) The lecturer will offer a two-hour so-called **Q&A session** almost every week during the teaching period, but not in the weeks in which term exams will be held. These Q&A sessions will be devoted to

- discussing and answering questions asked by the participants of the course,
- presenting solutions of review questions, and
- offering additional explanations for some parts of the course.

Questions can be asked right during the Q&A session, but participants of the course are encouraged to submit general or specific questions of comments beforehand.

Questions/comments can be posted in dedicated DISCUNA chat channels (folder "Q&A Channels", community "NumCSE <YEAR>"), which will be set up for each week in which a regular Q&A session will take place.
(Consult the DISCUNA "Help Center" accessible through the sidebar to learn how to post to DISCUNA chats.)

It is highly desirable that questions are *submitted* at least a few hours *before* the start of the Q&A session so that the lecturer has the opportunity to structure his or her answer.

Tablet notes of the Q&A sessions will be made available for download. ⌟

0.2.2 Clarifications and Frank Words

§0.2.2.1 ("Lecture notes")

The PDF you are reading is referred to as lecture document and is an important source of information, but

> this course document is neither a textbook nor comprehensive lecture notes.
> They are meant to supplement and be supplemented by explanations given in the videos.

Some pieces of advice:

- The lecture document is only partly designed to be self-contained and can/should be studied *in parts* in addition to attending to watching the course videos and/or reading the tablet notes.
- This text is not meant for mere reading, but for working with,
- Turn pages all the time and follow the numerous cross-references,
- study the relevant section of the course material when doing homework problems,
- You may study referenced literature to refresh prerequisite knowledge and for alternative presentation of the material (from a different angle, maybe), but be careful about not getting confused or distracted by information overload.

⌟

§0.2.2.2 (Comprehension is a process . . .)

- This course will require

  hard work – perseverance – patience

- Do **not** expect to understand everything at once. Most students will

- understand about one third of the material when watching videos and studying the course material
- understand another third when making a *serious effort* to solve the homework problems,
- hopefully understand the remaining third when studying for the main examination after the end of the course.

Perseverance will be rewarded!

⌟

§0.2.2.3 (Expected workload)

(I) You are a student in the BSc/MSc programme of Computational Science and Engineering (CSE) or others. Then you are taking the *full version* of the course (401-2673-*), which is endowed with **9 ECTS credits**, which roughly corresponds to a total workload of 270 hours:

$$270 \text{ hours} = \underbrace{180 \text{ hours}}_{\text{during term}} + \underbrace{90 \text{ hours}}_{\text{exam preparation}} ,$$

which amounts to a massive

$$\text{average workload} \approx 10 - 14 \text{ hours per week.}$$

(II) If you are a student in the BSc Computer Science you are offered a *trimmed version* of the course, which is worth **7 ECTS credits**. Though a very loose relationship, this roughly indicates a total workload of 180 hours:

$$180 \text{ hours} = \underbrace{110 \text{ hours}}_{\text{during term}} + \underbrace{70 \text{ hours}}_{\text{exam preparation}} .$$

This indicates that you should brace for an

$$\text{average workload} \approx 7 - 9 \text{ hours per week.}$$

For both versions of the course your efforts have to be split between

- watching videos and/or studying the course material,
- and solving homework problems,
- attending Q&A sessions and tutorials.

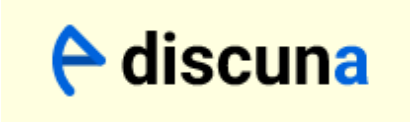
where homework may keep you busy for $5 - 7$ hours every week for the full version.

Of course, all these are averages and the workload may vary between different weeks. ⌟

0.2.3 Requests

The lecturers very much welcome and, putting it even more strongly, rather depend on feedback and suggestions of the students taking the course for continuous improvement of the course contents and presentation. Therefore all participants are strongly encouraged to get involved actively and contribute in the following ways:

§0.2.3.1 (Reporting errors) As the documents for this course will always be in a state of flux, they will inevitably and invariably teem with small errors, mainly typos and omissions.



For error reporting we use the DISCUNA online collaboration platform that runs in the browser.

DISCUNA allows to attach various types of annotations to shared PDF documents, see the DISCUNA "Help Center" or this instruction video.

> Please report errors in the lecture material through the DISCUNA NumCSE Community to which various course-related documents have already been uploaded.

In the beginning of the teaching period you receive a join link of the form `https://app.discuna.com/invite/<JOIN CODE>`. Open the link in a web browser and it will take you to the DISCUNA community page.

To report an error,

1. select the corresponding PDF document (chapter of the lecture document of homework problem) in the left sidebar,
2. press the prominent white-on-blue +-button in the right sidebar,
3. click on the displayed PDF where the error is located,
4. then in the pop-up window choose the "Error" category,
5. and add a title and,
6. if the title does not tell everything, a short description.

In case you cannot or do not want to link an error to a particular point in the PDF, you may just click on the title page of the respective chapter. Then, please precisely specify the concerned section and the number of the paragraph, remark, equation etc. *Do not give page numbers* as they may change with updates to the documents.

Note that chapter PDFs and homework problem files will gradually be added to the DISCUNA NumCSE community. Hence, the final chapters will not be accessible in the beginning of the course. ⌟

§0.2.3.2 (Pointing out technical problems) The DISCUNA NumCSE Community is equipped with a **chat channel "Technical Problems"**. In case you encounter a problem affecting the videos, the course web pages, or the PDF documents supplied online, say, severely distorted or missing audio tracks or a faulty link, instantly post a comment to this channel with a short description of the problem. You can do this after clicking on the channel name in the left sidebar in the community ⌟

§0.2.3.3 (Providing comments and suggestions) The **chat channel "General Comments"** of the DISCUNA NumCSE Community is meant for letting the lecturer know about weaknesses of the contents, structure, and presentation of the course and how they can be remedied. Your statements should be constructive and address specific parts or aspects of the course.

Regularly, students attending the course remark that they have found online resources like instruction videos that they think present some of the course material in a much clearer and better structured way. It is important that you tell the lecturer about those online resources so that he can include pointers to them and get inspiration. Use the "General Comments" channel also for this purpose. State clearly, which part of the course you are referring to, and briefly explain why the online resource is superior or a valuable supplement. ⌟

§0.2.3.4 (Asking/posting questions) Whenever a question comes up while you are studying for the course or trying to solve homework problems and that question lingers, it is probably connected to an issue that also bothers other students. What to do, in case you are not able to attend the Q&A session?

Please post arising question to the DISCUNA Q&A channels even if you do not attend the Q&A session! See also § 0.2.1.8

This has the benefit of

- initiating a discussion of the question that may also be relevant for other students, and
- will make it possible for you to find an answer in the Q&A tablet notes.

Tongue in cheek:

There is no question too stupid to be worth asking!

The most stupid practice is to hesitate to ask questions!

0.2.4 Assignments

A steady and persistent effort spent on homework problems is essential for success in this course.

You should expect to spend 3-5 hours per week on trying to solve the homework problems. Since many involve small coding projects, the time it will take an individual student to arrive at a solution is hard to predict.

For the sake of efficiency:
Avoid coding errors (bugs) in your homework coding projects!

The problems are published online together with plenty of hints. A master solution will also be made available, but it is foolish to read the master solution parallel to working on a problem sheet, because *trying to find the solution on one's own is essential for developing problem solving skills*, though it may occasionally be frustrating.

§0.2.4.1 (Homeworks and tutors' corrections)

- The weekly assignments will be a few problems from the **NCSE Problem Collection** available online as PDF, see course Moodle page for the link. The particular problems to be solved will be communicated through that Moodle page every week.

  Please note that this problem collection is being extended throughout the semester. Thus, make sure that you obtain the most current version every week. A polybox link will also be distributed; if you install the Polybox Client the most current version of all course documents will always be uploaded to your machine.
- Some or all of the problems of an assignment sheet will be discussed in the tutorial classes at least one week after the problems have been assigned.
- Your tutors are happy to examine your solutions and give you feedback : You may either hand them your solution papers during the tutorial session (put your name on every sheet and clearly mark the problems you want to be inspected) or upload a scan/photo through the CODEEXPERT upload interface, see § 0.2.4.2 below. You are encouraged to hand in incomplete and wrong solutions, so that you can receive valuable feedback even on incomplete or failed attempts.

✦ Your tutors will automatically have access to all your homework codes, see § 0.2.4.2 below.

⌟

§0.2.4.2 (CODEEXPERT C++ online IDE and testing evironment)

[code]expert

CODEEXPERT has been developed at ETH as on online IDE for small programming assignment and coding homeworks. It will be used in this course for all C++ homework problems.

Please study the documentation!

CODEEXPERT also offers the possibility of uploading any files to a private area (connected with a homework problem) that, beside you, only the tutor in charge of your exercise group can access. This is the preferred option for sharing (scans/photos of) your solutions of homework problem with your tutor.

Note that CODEEXPERT will also be using for the coding problems of the main examination.

⌟

§0.2.4.4 (CODEEXPERT synchronization with local folder) If you prefer to use your own editor locally on your computer, synchronization between the online CODEEXPERT repository and your local folder is available via Code Expert Sync tool. Follow the instructions here.

The working pipeline is:

Sync from CODEEXPERT platform ⟶ Edit locally ⟶ Sync with CODEEXPERT and run/test ⟶ continue editting locally ...

⌟

0.2.5 Information on Examinations

0.2.5.1 For the Course 401-2673-00L Numerical Methods for CSE (BSc CSE)

§0.2.5.1 (Examinations during the teaching period) From the ETH course directory:

> An optional 30-minutes **mid-term exam** and an optional 30-minutes **end-term exam** will be held during the teaching period. The grades of these interim examinations will be taken into account through a *BONUS* of up to 30% for the final grade.

The term exams will be conducted as *closed book* examinations *on paper*. The dates of the exams will be communicated in the beginning of the term and published on the course webpage. The term exams can neither be repeated nor be taken remotely.

The final grade is computed according to the formula

$$G := 0.25 \cdot \lceil 4 \cdot \max\{G_s, 0.85G_s + 0.15g_m, 0.85G_s + 0.15g_e, 0.7G_s + 0.15g_m + 0.15g_e\} \rceil \,, \quad (0.2.5.2)$$

$G_s \hat{=}$ grade in main exam, $g_m \hat{=}$ mid-term grade, $g_e \hat{=}$ end-term grade,

where $\lceil x \rceil$ designates the smallest integer $\geq x$. ⌟

§0.2.5.3 (Main examination during exam session)

- Three-hour written examination involving coding problems to be done at the computer. The date of the exam will be set and communicated by the ETH exam office, and will also be published on the course webpage.
- The coding part of the exam has to be done using CODEEXPERT.
- Subjects of examination:
  - All topics that have been addressed in a video listed on the course Moodle page or in any assigned homework problem

  The lecture document contains much more material than covered in class. All these extra topics are not relevant for the exam.
- Lecture document (as PDF), the EIGEN documentation, and the online C++ REFERENCE PAGES will be available PDF during the examination. The corresponding final version of the lecture document will be made available at least two weeks before the exam.
- No other materials may be used during the exam.
- The homework problem collection cannot be accessed during the exam.
- The exam questions will be asked in English.
- In case you come to the conclusion that you have too little time to prepare for the main exam a few weeks before the exam, contemplate withdrawing in order not to squander an attempt.

⌟

§0.2.5.4 (Repeating the main exam)

- Bonus points earned in term exams in last year's course can be taken into account for this course's main exam.
- If you want to take this option, please declare this intention by email to the course organizers before the mid-term exam. Otherwise, your bonus will be based on the results of this year's term exams.

⌟

0.2.5.2 For the Course 401-0663-00L Numerical Methods for CS (BSc Informatik)

§0.2.5.5 (Homework bonus) During the teaching period every week quizzes and exercices similar to those that will appear in the final exam are published on the moodle page of the lecture. These are open for answers for about one week and students are expected to answer them within this time. Answering them later is not possible. Correct answers are awarded "**semester points**" that are defined for each questions. Hence, each student has the possibility to accumulate such points during the semester.

Grade bonus
The grade achieved in the final exam will be raised by 0.25 for all students who have earned at least 75% of the "semester points".

⌟

§0.2.5.7 (Main (session) examination)

- Most of the exam questions are quizzes and exercices simmilar to those assigned as homework.
- The exam will mainly comprise either multiple choice tasks or tasks where you have to type the answer in an answer box.

- The exam may contain questions addressing coding. You may be asked to find and correct errors in code snippets or supplement missing parts of a code.
- Some tasks may require programming in order to be answered, though these codes will NOT be checked, but only the correctness of the final result will matter.
- Visual code studio can be used as an editor during the exam, but only the codes submitted through CODEEXPERT will be saved during the exam and taken into account for grading.

⌟

0.3 Programming in C++

C++20 is the *current* ANSI/ISO standard for the programming language C++. On the one hand, it offers a wealth of features and possibilities. On the other hand, this can be confusing and even be prone to inconsistencies. A major cause of inconsistent design is the requirement with backward compatibility with the C programming language and the earlier standard C++ 98.

However, C++ has become the main language in computational science and engineering and high performance computing. Therefore this course relies on C++ to discuss the implementation of numerical methods.

In fact C++ is a blend of different programming paradigms:

- an object oriented core providing classes, inheritance, and runtime polymorphism,
- a powerful template mechanism for parametric types and partial specialization, enabling *template meta-programming* and compile-time polymorphism,
- a collection of abstract data containers and basic algorithms provided by the Standard Template Libary (STL).

Supplementary literature. A popular book for learning C++ that has been upgraded to include the C++11 standard is [LLM12].

The book [Jos12] gives a comprehensive presentation of the new features of C++11 compared to earlier versions of C++.

There are plenty of online reference pages for C++, for instance `http://en.cppreference.com` and `http://www.cplusplus.com/`.

The following sections highlight a few particular aspects of C++ that may be important for code development in this course.

The version of the course for BSc students of Computer Science includes a two-week introduction to C++ in the beginning of the course.

0.3.1 Function Arguments and Overloading

§0.3.1.1 (Function overloading, [LLM12, Sect. 6.4]) Argument types are an integral part of a function declaration in C++. Hence the following functions are different

```
int* f(int);      // use this in the case of a single numeric argument
double f(int *); // use only, if pointer to a integer is given
void f(const MyClass &);  // use when called for a MyClass object
```

and the compiler selects the function to be used depending on the type of the arguments following rather sophisticated rules, refer to overload resolution rules. Complications arise, because implicit type conversions have to be taken into account. In case of ambiguity a compile-time error will be triggered. Functions cannot be distinguished by return type!

For member functions (methods) of classes an additional distinction can be introduced by the **const** specifier:

```
struct MyClass {
  double f(double);       // use for a mutable object of type MyClass
  double f(double) const; // use this version for a constant object
  ...
};
```

The second version of the method `f` is invoked for *constant objects* of type **MyClass**. ⌟

§0.3.1.2 (Operator overloading [LLM12, Chapter 14]) In C++ unary and binary operators like `=`, `==`, `+`, `-`, `*`, `/`, `+=`, `-=`, `*=`, `/=`, `%`, `&&`, `||`, «, », etc. are regarded as functions with a fixed number of arguments (one or two). For built-in numeric and logic types they are defined already. They can be extended to any other type, for instance

```
MyClass operator +(const MyClass &,const MyClass &);
MyClass operator +(const MyClass &,double);
MyClass operator +(const MyClass &);    // unary + !
```

The same selection rules as for function overloading apply. Of course, operators can also be introduced as class member functions.

C++ gives complete freedom to overload operators. However, the semantics of the new operators should be close to the customary use of the operator. ⌟

§0.3.1.3 (Passing arguments by value and by reference [LLM12, Sect. 6.2]) Consider a generic function declared as follows:

```
void f(MyClass x); // Argument x passed by value.
```

When `f` is invoked, a *temporary copy* of the argument is created through the copy constructor or the move constructor of **MyClass**. The new temporary object is a *local variable* inside the function body.

When a function is declared as follows

```
void f(MyClass &x); // Argument x passed by reference.
```

then the argument is passed to the scope of the function and can be changed inside the function. *No copies* are created. If one wants to avoid the creation of temporary objects, which may be costly, but also wants to indicate that the argument will not be modified inside `f`, then the declaration should read

```
void f(const MyClass &x); // Argument x passed by constant referene.
```

New in C++11 is move semantics, enabled in the following definition

```
void f(const MyClass &&x); // Optional shallow copy
```

In this case, if the scope of the object passed as the argument is merely the function or **std::move()** tags it as disposable, the move constructor of **MyClass** is invoked, which will usually do a *shallow copy* only. Refer to Code 0.3.5.10 for an example. ⌟

0.3.2 Templates

§0.3.2.1 (Function templates) The template mechanism supports parameterization of definitions of classes and functions by type. An example of a function templates is

```
template <typename ScalarType,typename VectorType>
VectorType saxpy(ScalarType alpha,const VectorType &x,const
   VectorType &y)
  { return(alpha*x+y); }
```

Depending on the concrete type of the arguments the compiler will instantiate particular versions of this function, for instance `saxpy<float,double>`, when `alpha` is of type `float` and both `x` and `y` are of type `double`. In this case the return type will be **double**.

For the above example the compiler will be able to deduce the types `ScalarType` and `VectorType` from the arguments. The programmer can also specify the types directly through the `< >`-syntax as in

```
saxpy<double,double>(a,x,y);
```

If an instantiation for all arguments of type `double` is desired. In case, the arguments do not supply enough information about the type parameters, specifying (some of) them through `< >` is mandatory. ⌟

§0.3.2.2 (Class templates) A class template defines a class depending on one or more type parameters, for instance

```
template <typename T>
class MyClsTempl {
public:
  using parm_t = T;          // T-dependent type
  MyClsTempl(void);          // Default constructor
  MyClsTempl(const T&);      // Constructor with an argument
  template <typename U>
  T memfn(const T&,const U&) const; // Templated member function
private:
  T *ptr;    // Data member, T-pointer
};
```

Types **MyClsTempl<T>** for a concrete choice of **T** are instantiated when a corresponding object is declared, for instance via

```
double x = 3.14;
MyClass myobj; // Default construction of an object
MyClsTempl<double> tinstd; // Instantiation for T = double
MyClsTempl<MyClass> mytinst(myobj); // Instantiation for T = MyClass
MyClass ret = mytinst.memfn(myobj,x); // Instantiation of member
   function for U = double, automatic type deduction
```

The types spawned by a template for different parameter types have nothing to do with each other. ⌟

Requirements on parameter types

The parameter types for a template have to provide all type definitions, member functions, operators, and data to make possible the instantiation ("compilation") of the class of function template.

0.3.3 Function Objects and Lambda Functions

A function object is an object of a type that provides an overloaded "function call" **operator** `()`. Function objects can be implemented in two different ways:

(I) through special classes like the following that realizes a function $\mathbb{R} \mapsto \mathbb{R}$

```
class MyFun {
public:
  ...
  double operator (double x) const; // Evaluation operator
  ...
};
```

The evaluation operator can take more than one argument and need not be declared `const`.

(II) through lambda functions, an "anonymous function" defined as

```
[<capture list>] (<arguments>) -> <return type> { body; }
```

where <capture list> is a list of variables from the local scope to be passed to the lambda function; an `&` indicates passing by reference,
<arguments> is a comma separated list of function arguments complete with types,
<return type> is an *optional* return type; often the compiler will be able to deduce the return type from the definition of the function.

Function classes should be used, when the function is needed in different places, whereas lambda functions for short functions intended for single use.

C++ code 0.3.3.1: Demonstration of use of lambda function ↪ GITLAB

```
1  int main() {
2    // initialize a vector from an initializer list
3    std::vector<double> v({1.2,2.3,3.4,4.5,5.6,6.7,7.8});
4    // A vector of the same length
5    std::vector<double> w(v.size());
6    // Do cumulative summation of v and store result in w
7    double sum = 0;
8    std::transform(v.begin(),v.end(),w.begin(),
9              [&sum] (double x) { sum += x; return sum;});
10   cout << "sum = " << sum << ", w = [ ";
11   for(auto x: w) cout << x << ' '; cout << ']' << endl;
12   return(0);
13 };
```

In this code the lambda function captures the local variable `sum` by reference, which enables the lambda function to change its value in the surrounding scope.

§0.3.3.2 (Function type wrappers) The special class **std::**`function` provides types for general polymorphic function wrappers.

```
std::function<return type(arg types)>
```

C++ code 0.3.3.3: Use of `std::function` ⇝ GITLAB

```
double binop(double arg1,double arg2) { return (arg1/arg2); }

void stdfunctiontest(void) {
  // Vector of objects of a particular signature
  std::vector<std::function<double(double,double)>> fnvec;
  // Store reference to a regular function
  fnvec.push_back(binop);
  // Store a lambda function
  fnvec.push_back([] (double x,double y)  -> double { return y/x; });
  for (auto fn : fnvec) { std::cout << fn(3,2) << std::endl; }
}
```

In this example an object of type `std::function<double(double,double)>` can hold a regular function taking two **double** arguments and returning another **double** or a lambda function with the same signature. Guess the output of `stdfunctiontest`! ⌟

§0.3.3.4 (Recorder objects) In the case of routines that perform some numerical computations we are often interested in the final result only. Occasionally we may also want to screen intermediate results. The following example demonstrates the use of an optional object for collecting information while the function is being executed. If no such object is supplied, an idle lambda function is passed, which incurs absolutely no runtime overhead.

C++ code 0.3.3.5: An example of a function taking a recorder object. ⇝ GITLAB

```
template <typename RECORDER = std::function<void(int, int)>>
unsigned int myloopfunction(
    unsigned int n, unsigned int val = 1,
    RECORDER &&rec = [](int, int) -> void {}) {
  for (unsigned int i = 0; i < n; ++i) {
    rec(i, val); // Removed by the compiler for the default argument
    if (val % 2 == 0) {
      val /= 2;
    } else {
      val *= 3;
      val++;
    }
  }
  rec(n, val);
  return val;
}
```

C++ code 0.3.3.6: Calling `myloopfunction()` ⇝ GITLAB

```
  std::cout << "myloopfunction(10, 1) = " << myloopfunction(10, 1) << std::endl;
  // Run with recorder
  std::vector<std::pair<int, int>> store{};
  std::cout << "myloopfunction(10, 1) = "
            << myloopfunction(10, 1,
                              [&store](int n, int val) -> void {
                                store.emplace_back(n, val);
                              })
            << std::endl;
  std::cout << "History:" << std::endl;
  for (const auto& i : store) {
    std::cout << i.first << " -> " << i.second << std::endl;
```

```
14    }
```

⌟

§0.3.3.7 (Captures of lambda functions)

- Putting the name of a variable in the current scope in a lambda's capture list, makes that variable accessible inside the lambda's body as a *const reference*. The variables are *immutable* inside the lambda's body.
- Capturing a local variable as *non-const reference* prepend the variable name with `&`. That variable's value can be changed by the lambda.
- The capture list `[=]` captures all local variables as const references.
- Conversely, the capture list `[&]` means that all variables in the local scope are captured by non-const reference and can be changed by the lambda function.

⌟

§0.3.3.8 (Lambda functions inside member functions) To access class methods or class variables in a lambda function inside a member function of a class you have to capture the current object as *const reference* by putting **this** or ⋆**this** in the capture list.

C++ code 0.3.3.9: A lambda function inside a member function. ↬ GITLAB

```
2  struct X {
3    explicit X(int N) : N_(N) {}
4    [[nodiscard]] unsigned int mod(unsigned int n) const { return N_ % n; }
5    bool modmatch(const std::vector<int> &nums, unsigned int n);
6    private:
7      int N_;
8  };
9
10 bool X::modmatch(const std::vector<int> &nums, unsigned int n) {
11   auto it = std::find_if(nums.begin(), nums.end(), [this, n](int k) -> bool {
12     N_++; return (N_ != 0) and ((k % n) == mod(n));
13   });
14   return (it != nums.end());
15 }
```

⌟

§0.3.3.10 (Recursions based on lambda functions) Lambda functions offer an elegant way to implment recursive algorithms locally inside a function. Note that

- you have to capture the lambda function itself by reference,
- and that you cannot use **auto** for automatic compile-time type deduction of that lambda function!

C++ code 0.3.3.11: Recursively calling a lambda function ↬ GITLAB

```
2  int main() {
3    int n_calls = 0;
4    std::function<int(int)> factorial  = [&factorial, &n_calls](int n) -> int {
5      n_calls++;
6      if (n == 0) {
```

```
7       return 1;
8     }
9     return n * factorial(n - 1);
10 };
11   std::cout << "10! = " << factorial(10) << std::endl;
12   return 0;
13 }
```

0.3.4 Multiple Return Values

In PYTHON it is customary to return several variables from a function call, which, in fact, amounts to returning a tuple of mixed-type objects:

```
1 def f(a, b):
2     return min(a, b), max(a, b), (a+b)/2
3 x, y, z = f(1, 2)
```

In C++ this is also possible by using the tuple utility. For instance, the following function computes the mimimal and maximal element of a vector and also returns its cumulative sum. It returns all these values.

C++ code 0.3.4.1: Function with multiple return values ↪ GITLAB

```
1  template<typename T>
2  std::tuple<T,T,std::vector<T>> extcumsum(const std::vector<T> &v) {
3    // Local summation variable captured by reference by lambda function
4    T sum{};
5    // temporary vector for returning cumulative sum
6    std::vector<T> w{};
7    // cumulative summation
8    std::transform(v.cbegin(),v.cend(),back_inserter(w),
9              [&sum] (T x) { sum += x; return(sum); });
10   return (std::make_tuple(*std::min_element(v.cbegin(), v.cend()),
11                           *std::max_element(v.cbegin(), v.cend()),
12                           std::move(w)));
13 }
```

This code snippet shows how to extract the individual components of the tuple returned by the previous function.

C++ code 0.3.4.2: Calling a function with multiple return values ↪ GITLAB

```
1  int main () {
2    // initialize a vector from an initializer list
3    std::vector<double> v({1.2,2.3,3.4,4.5,5.6,6.7,7.8});
4    // Variables for return values
5    double minv,maxv;  // Extremal elements
6    std::vector<double> cs; // Cumulative sums
7    std::tie(minv,maxv,cs) = extcumsum(v);
8    cout << "min = " << minv << ", max = " << maxv << endl;
9    cout << "cs = [ "; for(double x: cs) cout << x << ' '; cout << "]" << endl;
10   return(0);
11 }
```

Be careful: many temporary objects might be created! A demonstration of this hidden cost is given in Exp. 0.3.5.27. From C++17 a more compact syntax is available:

C++ code 0.3.4.3: Calling a function with multiple return values ↪ GITLAB

```
int main () {
  // initialize a vector from an initializer list
  std::vector<double> v({1.2,2.3,3.4,4.5,5.6,6.7,7.8});
  // Definition of variables and assignment of return values all at once
  auto [minv, maxv, cs] = extcumsum(v);
  cout << "min = " << minv << ", max = " << maxv << endl;
  cout << "cs = [ "; for(double x: cs) cout << x << ' '; cout << "]" << endl;
  return(0);
}
```

Remark 0.3.4.4 ("auto" considered harmful) C++ is a strongly typed programming language and every variable must have a precise type. However, the developer of templated classes and functions may not know the type of some variables in advance, because it can be deduced only after instantiation through the compiler. The **auto** keyword has been introduced to handle this situation.

There is a temptation to use **auto** profligately, because it is convenient, in particular when *using* templated data types. However, this denies a major benefit of types, consistency checking at compile time and, as a developer, one may eventually lose track of the types completely, which can lead to errors that are hard to detect.

Thus, the use of **auto** should be avoided, unless in the following situations:

- for variables inside templated functions or classes, whose precise type will only become clear during instantiation,
- for lambda functions, see Section 0.3.3,
- for return values of templated library (member) functions, whose type is "impossible to deduce" by the user. An example is expression templates in EIGEN, refer to Rem. 1.2.1.11 below.

⌟

0.3.5 A Vector Class

Since C++ is an object oriented programming language, datatypes defined by classes play a pivotal role in every C++ program. Here, we demonstrate the main ingredients of a class definition and other important facilities of C++ for the class **MyVector** meant for objects representing vectors from $\mathbb{R}^n$. The codes can be found in ↪ GITLAB. A similar vector class is presented in [Fri19, Ch. 6].

C++ 11 class 0.3.5.1: Definition of a simple vector class MyVector ↪ GITLAB

```
namespace myvec {
class MyVector {
public:
  using value_t = double;
  // Constructor creating constant vector, also default constructor
  explicit MyVector(std::size_t n = 0,double val = 0.0);
  // Constructor:  initialization from an STL container
  template <typename Container>  MyVector(const Container &v);
  // Constructor:  initialization from an STL iterator range
  template <typename Iterator> MyVector(Iterator first,Iterator last);
  // Copy constructor, computational cost O(n)
```

```
12    MyVector(const MyVector &mv);
13    // Move constructor, computational cost O(1)
14    MyVector(MyVector &&mv);
15    // Copy assignment operator, computational cost O(n)
16    MyVector &operator = (const MyVector &mv);
17    // Move assignment operator, computational cost O(1)
18    MyVector &operator = (MyVector &&mv);
19    // Destructor
20    virtual ~MyVector(void);
21    // Type conversion to STL vector
22    operator std::vector<double> () const;
23
24    // Returns length of vector
25    std::size_t size(void) const { return n; }
26    // Access operators:  rvalue & lvalue, with range check
27    double operator [] (std::size_t i) const;
28    double &operator [] (std::size_t i);
29    // Comparison operators
30    bool operator == (const MyVector &mv) const;
31    bool operator != (const MyVector &mv) const;
32    // Transformation of a vector by a function ℝ → ℝ
33    template <typename Functor>
34    MyVector &transform(Functor &&f);
35
36    // Overloaded arithmetic operators
37    // In place vector addition:  x += y;
38    MyVector &operator +=(const MyVector &mv);
39    // In place vector subtraction:  x-= y;
40    MyVector &operator -=(const MyVector &mv);
41    // In place scalar multiplication:  x *= a;
42    MyVector &operator *=(double alpha);
43    // In place scalar division:  x /= a;
44    MyVector &operator /=(double alpha);
45    // Vector addition
46    MyVector operator + (MyVector mv) const;
47    // Vector subtraction
48    MyVector operator - (const MyVector &mv) const;
49    // Scalar multiplication from right and left:  x = a*y; x = y*a
50    MyVector operator * (double alpha) const;
51    friend MyVector operator * (double alpha,const MyVector &);
52    // Scalar divsion:  x = y/a;
53    MyVector operator / (double alpha) const;
54
55    // Euclidean norm
56    [[nodiscard]] double norm(void) const;
57    // Euclidean inner product
58    double operator *(const MyVector &) const;
59    // Output operator
60    friend std::ostream &
61    operator << (std::ostream &,const MyVector &mv);
62
63    static bool dbg; // Flag for verbose output
64    // Non-const static class variables deprecated by C++ core guidelines!
65  private:
66    std::size_t n {0}; // Length of vector
67    double *data {nullptr};  // data array (standard C array)
68  };
69  } // namespace myvec
```

Note the use of a public static data member `dbg` in Line 63 that can be used to control debugging output

by setting `MyVector::dbg = true` or `MyVector::dbg = false`.

Remark 0.3.5.2 (Contiguous arrays in C++) The class **MyVector** uses a C-style array and dynamic memory management with **new** and **delete** to store the vector components. This is for demonstration purposes only and not recommended.

Arrays in C++

In C++ use the STL container **std::vector<T>** for storing data in contiguous memory locations. Exception: use **std::array<T>**, if the number of elements is known at compile time.

⌟

§0.3.5.4 (Member and friend functions of MyVector ⇝ GITLAB)

C++ code 0.3.5.5: Constructor for constant vector, also default constructor, see Line 6 in Code 0.3.5.1 ⇝ GITLAB

```
1 MyVector::MyVector(std::size_t _n,double _a):n(_n),data(nullptr) {
2   if (dbg) cout << "{Constructor MyVector(" << _n
3                 << ") called" << '}' << endl;
4   if (n > 0) data = new double [_n];
5   for (std::size_t l=0;l<n;++l) data[l] = _a;
6 }
```

This constructor can also serve as default constructor (a constructor that can be invoked without any argument), because defaults are supplied for all its arguments.

The following two constructors initialize a vector from sequential containers according to the conventions of the STL.

C++ code 0.3.5.6: Templated constructors copying vector entries from an STL container ⇝ GITLAB

```
1 template <typename Container>
2 MyVector::MyVector(const Container &v):n(v.size()),data(nullptr) {
3   if (dbg) cout << "{MyVector(length " << n
4                 << ") constructed from container" << '}' << endl;
5   if (n > 0) {
6     double *tmp = (data = new double [n]);
7     for(auto i: v) *tmp++ = i;  // foreach loop
8   }
9 }
```

Note the use of the new C++ 11 facility of a "foreach loop" iterating through a container in Line 7.

C++ code 0.3.5.7: Constructor initializing vector from STL iterator range ⇝ GITLAB

```
1 template <typename Iterator>
2 MyVector::MyVector(Iterator first,Iterator last):n(0),data(nullptr) {
3   n = std::distance(first,last);
4   if (dbg) cout << "{MyVector(length " << n
5                 << ") constructed from range" << '}' << endl;
6   if (n > 0) {
7     data = new double [n];
```

```
8       std::copy(first,last,data);
9     }
10  }
```

The use of these constructors is demonstrated in the following code

C++ code 0.3.5.8: Initialization of a MyVector object from an STL vector ↪ GITLAB

```
1  int main() {
2    myvec::MyVector::dbg = true;
3    std::vector<int> ivec = { 1,2,3,5,7,11,13 }; // initializer list
4    myvec::MyVector v1(ivec.cbegin(),ivec.cend());
5    myvec::MyVector v2(ivec);
6    myvec::MyVector vr(ivec.crbegin(),ivec.crend());
7    cout << "v1 = " << v1 << endl;
8    cout << "v2 = " << v2 << endl;
9    cout << "vr = " << vr << endl;
10   return(0);
11 }
```

The following output is produced:

```
{MyVector(length 7) constructed from range}
{MyVector(length 7) constructed from container}
{MyVector(length 7) constructed from range}
v1 = [ 1,2,3,5,7,11,13 ]
v2 = [ 1,2,3,5,7,11,13 ]
vr = [ 13,11,7,5,3,2,1 ]
{Destructor for MyVector(length = 7)}
{Destructor for MyVector(length = 7)}
{Destructor for MyVector(length = 7)}
```

The copy constructor listed next relies on the *STL algorithm* **std::**`copy` to copy the elements of an existing object into a newly created object. This takes n operations.

C++ code 0.3.5.9: Copy constructor ↪ GITLAB

```
1  MyVector::MyVector(const MyVector &mv):n(mv.n),data(nullptr) {
2     if (dbg) cout << "{Copy construction of MyVector(length "
3                   << n << ")" << '}' << endl;
4     if (n > 0) {
5       data = new double [n];
6       std::copy_n(mv.data,n,data);
7     }
8  }
```

An important new feature of C++11 is move semantics which helps avoid expensive copy operations. The following implementation just performs a shallow copy of pointers and, thus, for large n is much cheaper than a call to the copy constructor from Code 0.3.5.9. The source vector is left in an empty vector state.

C++ code 0.3.5.10: Move constructor ↪ GITLAB

```
1  MyVector::MyVector(MyVector &&mv):n(mv.n),data(mv.data) {
2     if (dbg) cout << "{Move construction of MyVector(length "
3                   << n << ")" << '}' << endl;
```

```
4     mv.data = nullptr; mv.n = 0; // Reset victim of data theft
5   }
```

The following code demonstrates the use of **std::move()** to mark a vector object as disposable and allow the compiler the use of the move constructor. The code also uses left multiplication with a scalar, see Code 0.3.5.23.

C++ code 0.3.5.11: Invocation of copy and move constructors ⇝ GITLAB

```
1  int main() {
2    myvec::MyVector::dbg = true;
3    myvec::MyVector v1(std::vector<double>(
4      {1.2,2.3,3.4,4.5,5.6,6.7,7.8,8.9}));
5    myvec::MyVector v2(2.0*v1); // Scalar multiplication
6    myvec::MyVector v3(std::move(v1));
7    cout << "v1 = " << v1 << endl;
8    cout << "v2 = " << v2 << endl;
9    cout << "v3 = " << v3 << endl;
10   return (0);
11 }
```

This code produces the following output. We observe that `v1` is empty after its data have been "stolen" by `v2`.

```
{MyVector(length 8) constructed from container}
{operator a*, MyVector of length 8}
{Copy construction of MyVector(length 8)}
{operator *=, MyVector of length 8}
{Move construction of MyVector(length 8)}
{Destructor for MyVector(length = 0)}
{Move construction of MyVector(length 8)}
v1 = [ ]
v2 = [ 2.4,4.6,6.8,9,11.2,13.4,15.6,17.8 ]
v3 = [ 1.2,2.3,3.4,4.5,5.6,6.7,7.8,8.9 ]
{Destructor for MyVector(length = 8)}
{Destructor for MyVector(length = 8)}
{Destructor for MyVector(length = 0)}
```

We observe that the object `v1` is reset after having been moved to `v3`.

Use `std::move` only for special purposes like above and only if an object has a move constructor. Otherwise a 'move' will trigger a plain copy operation. In particular, do not use `std::move` on objects at the end of their scope, e.g., within `return` statements.

The next operator effects copy assignment of an rvalue **MyVector** object to an lvalue **MyVector**. This involves $O(n)$ operations.

C++ code 0.3.5.12: Copy assignment operator ⇝ GITLAB

```
1  MyVector &MyVector::operator = (const MyVector &mv) {
2    if (dbg) cout << "{Copy assignment of MyVector(length "
3                  << n << "<-" << mv.n << ")"<< '}' << endl;
4    if (this == &mv) return(*this);
5    if (n != mv.n) {
6      n = mv.n;
```

```
    if (data != nullptr) delete [] data;
    if (n > 0) data = new double [n]; else data = nullptr;
  }
  if (n > 0) std::copy_n(mv.data,n,data);
  return(*this);
}
```

The move semantics is realized by an assignment operator relying on shallow copying.

C++ code 0.3.5.13: Move assignment operator ⇝ GITLAB

```
MyVector &MyVector::operator = (MyVector &&mv) {
  if (dbg) cout << "{Move assignment of MyVector(length "
                << n << "<-" << mv.n << ")"<< '}' << endl;
   if (data != nullptr) delete [] data;
   n = mv.n; data = mv.data;
   mv.n = 0; mv.data = nullptr;
   return(*this);
}
```

The destructor releases memory allocated by **new** during construction or assignment.

C++ code 0.3.5.14: Destructor: releases allocated memory ⇝ GITLAB

```
MyVector::~MyVector(void) {
  if (dbg) cout << "{Destructor for MyVector(length = "
                << n << ")" << '}' << endl;
  if (data != nullptr) delete [] data;
}
```

The **operator** keyword is also use to define implicit type conversions.

C++ code 0.3.5.15: Type conversion operator: copies contents of vector into STL vector ⇝ GITLAB

```
MyVector::operator std::vector<double> () const {
  if (dbg) cout << "{Conversion to std::vector, length = " << n << '}' << endl;
  return std::vector<double>(data,data+n);
}
```

The bracket operator `[]` can be used to fetch and set vector components. Note that index range checking is performed; an *exception* is thrown for invalid indices. The following code also gives an example of operator overloading as discussed in § 0.3.1.2.

C++ code 0.3.5.16: rvalue and lvalue access operators ⇝ GITLAB

```
double MyVector::operator [] (std::size_t i) const {
  if (i >= n) throw(std::logic_error("[] out of range"));
  return data[i];
}

double &MyVector::operator [] (std::size_t i) {
  if (i >= n) throw(std::logic_error("[] out of range"));
  return data[i];
}
```

Componentwise direct comparison of vectors. Can be dangerous in numerical codes,*cf.* Rem. 1.5.3.15.

C++ code 0.3.5.17: Comparison operators ⇝ GITLAB

```
bool MyVector::operator == (const MyVector &mv) const
{
  if (dbg) cout << "{Comparison ==: " << n << " <-> " << mv.n << '}' << endl;
  if (n != mv.n) return(false);
  else {
    for(std::size_t l=0;l<n;++l)
      if (data[l] != mv.data[l]) return(false);
  }
  return(true);
}

bool MyVector::operator != (const MyVector &mv) const {
  return !(*this == mv);
}
```

The `transform` method applies a function to every vector component and overwrites it with the value returned by the function. The function is passed as an object of a type providing a `()`-operator that accepts a single argument convertible to `double` and returns a value convertible to `double`.

C++ code 0.3.5.18: Transformation of a vector through a functor `double` → `double` ⇝ GITLAB

```
template <typename Functor>
MyVector &MyVector::transform(Functor &&f) {
  for(std::size_t l=0;l<n;++l) data[l] = f(data[l]);
  return(*this);
}
```

The following code demonstrates the use of the `transform` method in combination with

1. a function object of the following type

C++ code 0.3.5.19: A functor type

```
struct SimpleFunction {
  SimpleFunction(double _a = 1.0):cnt(0),a(_a) {}
  double operator () (double x) { cnt++; return(x+a); }
  int cnt;          // internal counter
  const double a;  // increment value
};
```

2. a lambda function defined directly inside the call to `transform`.

C++ code 0.3.5.20: transformation of a vector via a functor object

```
int main() {
  myvec::MyVector::dbg = false;
  double a = 2.0; // increment
  int cnt = 0;    // external counter used by lambda function
  myvec::MyVector mv(std::vector<double>(
    {1.2,2.3,3.4,4.5,5.6,6.7,7.8,8.9}));
  mv.transform([a,&cnt] (double x) { cnt++; return(x+a); });
  cout << cnt << " operations, mv transformed = " << mv << endl;
```

```
  SimpleFunction trf(a); mv.transform(trf);
  cout << trf.cnt << " operations, mv transformed = " << mv << endl;
  mv.transform(SimpleFunction(-4.0));
  cout << "Final vector = " << mv << endl;
  return(0);
}
```

The output is

```
8 operations, mv transformed = [ 3.2,4.3,5.4,6.5,7.6,8.7,9.8,10.9 ]
8 operations, mv transformed = [ 5.2,6.3,7.4,8.5,9.6,10.7,11.8,12.9 ]
Final vector = [ 1.2,2.3,3.4,4.5,5.6,6.7,7.8,8.9 ]
```

Operator overloading provides the "natural" vector operations in $\mathbb{R}^n$ both in place and with a new vector created for the result.

C++ code 0.3.5.21: In place arithmetic operations (one argumnt) ↪ GITLAB

```
MyVector &MyVector::operator +=(const MyVector &mv) {
  if (dbg) cout << "{operator +=, MyVector of length "
               << n << '}' << endl;
  if (n != mv.n) throw(std::logic_error("+=: vector size mismatch"));
  for(std::size_t l=0;l<n;++l) data[l] += mv.data[l];
  return(*this);
}

MyVector &MyVector::operator -=(const MyVector &mv) {
  if (dbg) cout << "{operator -=, MyVector of length "
               << n << '}' << endl;
  if (n != mv.n) throw(std::logic_error("-=: vector size mismatch"));
  for(std::size_t l=0;l<n;++l) data[l] -= mv.data[l];
  return(*this);
}

MyVector &MyVector::operator *=(double alpha) {
  if (dbg) cout << "{operator *=, MyVector of length "
               << n << '}' << endl;
  for(std::size_t l=0;l<n;++l) data[l] *= alpha;
  return(*this);
}

MyVector &MyVector::operator /=(double alpha) {
  if (dbg) cout << "{operator /=, MyVector of length "
               << n << '}' << endl;
  for(std::size_t l=0;l<n;++l) data[l] /= alpha;
  return(*this);
}
```

C++ code 0.3.5.22: Binary arithmetic operators (two arguments) ↪ GITLAB

```
MyVector MyVector::operator + (MyVector mv) const {
  if (dbg) cout << "{operator +, MyVector of length "
               << n << '}' << endl;
  if (n != mv.n) throw(std::logic_error("+: vector size mismatch"));
  mv += *this;
  return(mv);
}

```

```
9  MyVector MyVector::operator - (const MyVector &mv) const {
10   if (dbg) cout << "{operator -, MyVector of length "
11                 << n << '}' << endl;
12   if (n != mv.n) throw(std::logic_error("+: vector size mismatch"));
13   MyVector tmp(*this); tmp -= mv;
14   return(tmp);
15 }
16
17 MyVector MyVector::operator * (double alpha) const {
18   if (dbg) cout << "{operator *a, MyVector of length "
19                 << n << '}' << endl;
20   MyVector tmp(*this); tmp *= alpha;
21   return(tmp);
22 }
23
24 MyVector MyVector::operator / (double alpha) const {
25   if (dbg) cout << "{operator /, MyVector of length " << n << '}' << endl;
26   MyVector tmp(*this); tmp /= alpha;
27   return(tmp);
28 }
```

C++ code 0.3.5.23: Non-member function for left multiplication with a scalar ↪ GITLAB

```
1 MyVector operator * (double alpha,const MyVector &mv) {
2   if (MyVector::dbg) cout << "{operator a*, MyVector of length "
3                           << mv.n << '}' << endl;
4   MyVector tmp(mv); tmp *= alpha;
5   return(tmp);
6 }
```

C++ code 0.3.5.24: Euclidean norm ↪ GITLAB

```
1 double MyVector::norm(void) const {
2   if (dbg) cout << "{norm: MyVector of length " << n << '}' << endl;
3   double s = 0;
4   for(std::size_t l=0;l<n;++l) s += (data[l]*data[l]);
5   return(std::sqrt(s));
6 }
```

Adopting the notation in some linear algebra texts, the operator $*$ has been chosen to designate the Euclidean inner product:

C++ code 0.3.5.25: Euclidean inner product ↪ GITLAB

```
1 double MyVector::operator *(const MyVector &mv) const {
2   if (dbg) cout << "{dot *, MyVector of length " << n << '}' << endl;
3   if (n != mv.n) throw(std::logic_error("dot: vector size mismatch"));
4   double s = 0;
5   for(std::size_t l=0;l<n;++l) s += (data[l]*mv.data[l]);
6   return(s);
7 }
```

At least for debugging purposes every reasonably complex class should be equipped with output functionality.

C++ code 0.3.5.26: Non-member function output operator ⇝ GITLAB

```
std::ostream &operator << (std::ostream &o,const MyVector &mv) {
  o << "[ ";
  for(std::size_t l=0;l<mv.n;++l)
    o << mv.data[l] << (l==mv.n-1?' ':',');
  return(o << "]");
}
```

⌟

EXPERIMENT 0.3.5.27 ("Behind the scenes" of MyVector arithmetic) The following code highlights the use of operator overloading to obtain readable and compact expressions for vector arithmetic.

C++ code 0.3.5.28:

```
int main() {
  myvec::MyVector::dbg = true;
  myvec::MyVector x(std::vector<double>({1.2,2.3,3.4,4.5,5.6,6.7,7.8,8.9}));
  myvec::MyVector y(std::vector<double>({2.1,3.2,4.3,5.4,6.5,7.6,8.7,9.8}));
  auto z = x+(x*y)*x+2.0*y/(x-y).norm();
}
```

We run the code and trace calls. This is printed to the console:

```
{MyVector(length 8) constructed from container}
{MyVector(length 8) constructed from container}
{dot *, MyVector of length 8}
{operator a*, MyVector of length 8}
{Copy construction of MyVector(length 8)}
{operator *=, MyVector of length 8}
{operator +, MyVector of length 8}
{operator +=, MyVector of length 8}
{Move construction of MyVector(length 8)}
{operator a*, MyVector of length 8}
{Copy construction of MyVector(length 8)}
{operator *=, MyVector of length 8}
{operator -, MyVector of length 8}
{Copy construction of MyVector(length 8)}
{operator -=, MyVector of length 8}
{norm: MyVector of length 8}
{operator /, MyVector of length 8}
{Copy construction of MyVector(length 8)}
{operator /=, MyVector of length 8}
{operator +, MyVector of length 8}
{operator +=, MyVector of length 8}
{Move construction of MyVector(length 8)}
{Destructor for MyVector(length = 0)}
{Destructor for MyVector(length = 8)}
{Destructor for MyVector(length = 8)}
{Destructor for MyVector(length = 8)}
{Destructor for MyVector(length = 0)}
{Destructor for MyVector(length = 8)}
{Destructor for MyVector(length = 8)}
```

```
{Destructor for MyVector(length = 8)}
```

Several temporary objects are created and destroyed and quite a few *copy operations* take place. The situation would be worse unless move semantics was available; if we had not supplied a move constructor, a few more copy operations would have been triggered. Even worse, the frequent copying of data runs a high risk of cache misses. This is certainly not an efficient way to do elementary vector operations though it looks elegant at first glance. ⌟

EXAMPLE 0.3.5.29 (Gram-Schmidt orthonormalization based on MyVector implementation) Gram-Schmidt orthonormalization has been taught in linear algebra and its theory will be revisited in § 1.5.1.1. Here we use this simple algorithm from linear algebra to demonstrate the use of the vector class **MyVector** defined in Code 0.3.5.1.

The templated function `gramschmidt` takes a sequence of vectors stored in a **std::vector** object. The actual vector type is passed as a template parameter. It has to supply `size()` and `norm()` member functions as well as in place arithmetic operations `-=`, `/` and `=`. Note the use of the highlighted methods of the **std::vector** class.

C++ code 0.3.5.30: templated function for Gram-Schmidt orthonormalization ↪ GITLAB

```
1  template <typename Vec>
2  std::vector<Vec> gramschmidt(const std::vector<Vec> &A,double eps=1E-14) {
3    const int k = A.size();   // no.  of vectors to orthogonalize
4    const int n = A[0].size(); // length of vectors
5    cout << "gramschmidt orthogonalization for " << k << ' ' << n << "-vectors" <<
         endl;
6    std::vector<Vec> Q({A[0]/A[0].norm()}); // output vectors
7    for(int j=1;(j<k) && (j<n);++j) {
8      Q.push_back(A[j]);
9      for(int l=0;l<j;++l) Q.back() -= (A[j]*Q[l])*Q[l];
10     if (Q.back().norm() < eps*A[j].norm()) { // premature termination ?
11       Q.pop_back(); break;
12     }
13     Q.back() /= Q.back().norm(); // normalization
14   }
15   return (Q); // return at end of local scope
16 }
```

This driver program calls a function that initializes a sequence of vectors and then orthonormalizes them by means of the Gram-Schmidt algorithm. Eventually orthonormality of the computed vectors is tested. Please pay attention to

- the use of auto to avoid cumbersome type declarations,
- the for loops following the "foreach" syntax.
- automatic indirect template type deduction for the templated function `gramschmidt` from its argument. In Line 6 the function gramschmidt<MyVector> is instantiated.

C++ code 0.3.5.31: Driver code for Gram-Schmidt orthonormalization

```
1  int main() {
2    myvec::MyVector::dbg = false;
3    const int n = 7; const int k = 7;
4    auto A(initvectors(n,k,[] (int i,int j)
5           { return std::min(i+1,j+1); }));
6    auto Q(gramschmidt(A)); // instantiate template for MyVector
```

```
  cout << "Set of vectors to be orthonormalized:" << endl;
  for (const auto &a: A) { cout << a << endl; }
  cout << "Output of Gram-Schmidt orthonormalization: " << endl;
  for (const auto &q: Q) { cout << q << endl; }
  cout << "Testing orthogonality:" << endl;
  for (const auto &qi: Q) {
    for (const auto &qj: Q)
      cout << std::setprecision(3) << std::setw(9) << qi*qj << ' ';
    cout << endl; }
  return(0);
}
```

This initialization function takes a functor argument as discussed in Section 0.3.3.

C++ code 0.3.5.32: Initialization of a set of vectors through a functor with two arguments

```
template<typename Functor>
  std::vector<myvec::MyVector>
    initvectors(std::size_t n,std::size_t k,Functor &&f) {
  std::vector<MyVector> A{};
  for(int j=0;j<k;++j) {
    A.push_back(MyVector(n));
    for(int i=0;i<n;++i)
      (A.back())[i] = f(i,j);
    }
  return(A);
}
```

0.3.6 Complex numbers in C++

§0.3.6.1 (Data types for complex numbers) The fundamental data type for complex numbers is

```
using complex = std::complex<T>;
```

where the template argument `T` must be a floating point type like **double** or **float**. Then he type **complex**

- t supports all basic arithmetic operations $+, -, *, /$,
- provides the member functions `real()` and `imag()` for extracting real and imaginary parts,
- and can be passed to **std::abs()** and **std::arg()** to get the modulus $|z|$ and the argument $\varphi \in [-\pi, \pi]$ of the complex number $z = |z| \exp(i\varphi)$.

Complex conjugation can be done by calling **std::conj()** for a complex number.

§0.3.6.2 (Initialization of complex numbers) The value of a variable of type **std::complex<double>** can be initialized

- by calling the standard constructor and supplying real and imaginary part: `x = std::complex<double>(x,y)`, where `x,y` are of a numeric type that can be converted to **double**. If the second argument is omitted, the imaginary part is set to zero.
- by providing a **complex literal**, `x = 1.0+1.0i`. This entails the directive **using namespace std::complex_literals**.

- by specifying the modulus $r \geq 0$ and argument $\varphi \in \mathbb{R}$ and calling `std::polar()`: `x = std::polar(r,phi)`. Arguments are always given in radians.

⌟

§0.3.6.3 (Functions with complex arguments) All standard mathematical functions like exp, sin, cos, sinh, and cosh can be supplied with complex arguments.

Note that the definition of log and of square roots for complex argument entails specifying a **branch cut**. The default choice for the built-in functions is the negative real line. For instance this means that `std::sqrt(z)` for a complex number z will always have non-negative real part. ⌟

The following code demonstrates the handling of complex numbers.

C++ code 0.3.6.4: Data types and operations for complex numbers ↪ GITLAB

```
#include <complex>
#include <iostream>
#include <numbers>
using complex = std::complex<double>;
using namespace std::complex_literals;
int main(int /*argc*/, char** /*argv*/) {
  std::cout << "Demo: Complex numbers in C++" << std::endl;
  // This initialization requires std::complex_literals
  complex z = 0.5; // Std constructor, real part only
  z += 0.5 + 1.0i;
  // Various elementary operations, see
  // https://en.cppreference.com/w/cpp/numeric/complex
  std::cout << "z = " << z << ", Re(z) = " << z.real()
            << ", Im(z) = " << z.imag() << "| z | = " << std::abs(z)
            << ", arg(z) = " << std::arg(z) << ", conj(z) = " << std::conj(z)
            << std::endl;
  complex w = std::polar(1.0, std::numbers::pi / 4.0);
  std::cout << "w = " << w << std::endl;
  std::cout << "exp(z) = " << std::exp(z)
            << ", abs(exp(z)) = " << std::abs(std::exp(z)) << " = "
            << std::exp(z.real()) << std::endl;
  std::cout << "sqrt(z) = " << std::sqrt(z)
            << ", arg(sqrt(z)) = " << std::arg(std::sqrt(z)) << std::endl;

  return 0;
}
```

Terminal output:

```
Demo: Complex numbers in C++
z = (1,1), Re(z) = 1, Im(z) = 1| z | = 1.41421, arg(z) = 0.785398, conj(z)
    = (1,-1)
w = (0.707107,0.707107)
exp(z) = (1.46869,2.28736), abs(exp(z)) = 2.71828 = 2.71828
sqrt(z) = (1.09868,0.45509), arg(sqrt(z)) = 0.392699
```

0.4 Prerequisite Mathematical Knowledge

0.4.1 Basics

In school you should have learned basic facts of real analysis of one variable, in particular, about differentiation, integration, and fundamental special functions.

§0.4.1.1 (Power functions and roots)

$$x^0 := 1\,, \quad x^{m+n} = x^m x^n\,, \quad x^{-n} = \frac{1}{x^n} \quad \forall x \in \mathbb{R}\setminus\{0\}\,, \quad m,n \in \mathbb{Z}\,, \tag{0.4.1.2}$$

$$x^{a+b} = x^a x^b\,, \quad x^{ab} = (x^a)^b \quad \forall x > 0\,, \quad a,b \in \mathbb{R}\,, \tag{0.4.1.3}$$

$$\frac{d}{dx}\{x \mapsto x^n\} = nx^{n-1}\,, \quad x \neq 0,\ n \in \mathbb{N} \quad , \quad \frac{d}{dx}\{x \mapsto x^a\} = ax^{a-1}\,, \quad x > 0,\ a \in \mathbb{R}\,, \tag{0.4.1.4}$$

$$\int x^a \,\mathrm{d}x = \frac{x^{a+1}}{a+1} + C\,, \quad a \in \mathbb{R}\setminus\{-1\}\,, \tag{0.4.1.5}$$

where the last integral can only cover subsets of $\mathbb{R}^+$ unless $a \in \mathbb{N}$. The notation in (0.4.1.5) expresses that $x \mapsto \frac{x^{a+1}}{a+1}$ in the principal (ger.: Stammfunktion) of $x \mapsto x^a$. ⌟

§0.4.1.6 (Exponential functions and logarithms) In this course $\log$ always stands for the logarithm with respect to basis $e = 2.71828\ldots$.

$$\exp(x) = e^x\,, \quad x \in \mathbb{R} \quad , \quad \log(\exp(x)) = x \quad \forall x \in \mathbb{R} \quad , \quad a^x := \exp(x\log(a))\,, \quad x \in \mathbb{R},\ a > 0\,. \tag{0.4.1.7}$$

Calculus of exponential functions and logarithms:

$$\exp(x+y) = \exp(x)\exp(y)\,, \quad \exp(-x) = \frac{1}{\exp(x)} \quad \forall x,y \in \mathbb{R}\,, \tag{0.4.1.8}$$

$$\log(xy) = \log(x) + \log(y)\,, \quad \log(x/y)\log(x) - \log(y) \quad \forall x,y > 0\,,$$

$$\exp(nx) = \exp(x)^n\,, \quad \exp(ax) = \exp(x)^a \quad \forall x \in \mathbb{R},\ n \in \mathbb{Z},\ a > 0\,. \tag{0.4.1.9}$$

Differentiation and integration:

$$\frac{d}{dx}\{x \mapsto \exp(x)\} = \exp(x)\,, \quad x \in \mathbb{R} \quad , \quad \frac{d}{dx}\{x \mapsto \log(x)\} = \frac{1}{x}\,, \quad x > 0\,, \tag{0.4.1.10}$$

$$\int \exp(x)\,\mathrm{d}x = \exp(x) + C \quad , \quad \int \log(x)\,\mathrm{d}x = x\log(x) - x + C\,, \tag{0.4.1.11}$$

where, of course, the logarithm can only be integrated over subsets of $\mathbb{R}^+$. ⌟

§0.4.1.12 (Rules for differentiation and integration) Assuming sufficient smoothness of the involved functions $f : I \subset \mathbb{R} \to \mathbb{R}$ and $g : D \subset \mathbb{R} \to \mathbb{R}$ and that products and compositions are well-defined, we have, writing $f'(x) := \frac{df}{dx}(x)$, $g'(x) := \frac{dg}{dx}(x)$,

1D product rule: $$\frac{d}{dx}\{x \mapsto f(x)g(x)\} = f'(x)g(x) + f(x)g'(x)\,, \tag{0.4.1.13}$$

1D chain rule: $$\frac{d}{dx}\{x \mapsto f(g(x))\} = f'(g(x))g(x)\,. \tag{0.4.1.14}$$

This implies the following standard integration techniques:

Integration by substituion: $$\int_a^b f(g(x))g'(x)\,\mathrm{d}x = \int_{g(a)}^{g(b)} f(y)\mathrm{d}y\,, \tag{0.4.1.15}$$

integration by parts: $$\int f(x)g'(x)\,\mathrm{d}x = -\int f'(x)g(x)\,\mathrm{d}x + f(x)g(x)\,. \qquad (0.4.1.16)$$

Taylor expansion formula in one dimension for a function that is $m+1$ times continuously differentiable in a neighborhood of x_0

$$f(x_0+h) = \sum_{k=0}^{m} \frac{1}{k!} f^{(k)}(x_0)h^k + R_m(x_0,h)\,, \quad R_m(x_0,h) = \frac{1}{(m+1)!} f^{(m+1)}(\xi)h^{m+1}\,, \qquad (1.5.4.28)$$

for some $\xi \in [\min\{x_0, x_0+h\}, \max\{x_0, x_0+h\}]$, and for all sufficiently small $|h|$. ⌟

0.4.2 Complex Numbers

We write $\imath$ for the imaginary unit: $\operatorname{Re}\imath = 0$, $\operatorname{Im}\imath = 1$, $\imath^2 = -1$. Then every complex number $z \in \mathbb{C}$ can be identified with a pair $(x,y) \in \mathbb{R}^2$ of real numbers via $z = x + \imath y$.

Multiplication: $(x+\imath y)(u+\imath v) = (xu - yv) + \imath(xv + yu) \quad \forall x,y,u,v \in \mathbb{R}\,,$

Complex conjugation: $\overline{x+\imath y} = x - \imath y \quad \forall x,y \in \mathbb{R}\,,$

Modulus: $|z|^2 := \operatorname{Re}\{z\bar{z}\}\,, \quad |Z| = |z||w| \quad \forall z,w \in \mathbb{C}\,,$

Division: $\dfrac{w}{z} = \dfrac{w\bar{z}}{|z|^2} \quad \forall w,z \in \mathbb{C},\ z \neq 0\,.$

Many mathematical functions can be extended to complex arguments and many calculus rules will remain valid for them, in particular the formulas involving the functions exp, sin, cos.

Euler's formula: $$\exp(\imath t) = \cos(t) + \imath\sin(t) \quad \forall t \in \mathbb{R}\,. \qquad (0.4.2.1)$$

Some parts of this course will rely on sophisticated results from complex analysis, that is, the field of mathematics studying functions $\mathbb{C} \to \mathbb{C}$. This results will be recalled when needed.

0.4.3 Trigonometric Functions

Trigonometric functions can be defined via the complex exponential function:

$$\cos(z) = \tfrac{1}{2}(\exp(\imath z) + \exp(-\imath z)) \quad , \quad \sin(z) = \tfrac{1}{2\imath}(\exp(\imath z) - \exp(-\imath z))\,, \quad z \in \mathbb{C}\,. \qquad (0.4.3.1)$$

This implies

$$\sin^2 z + \cos^2 z = 1 \quad \forall z \in \mathbb{C} \quad , \quad 1 + \tan^2 z = \frac{1}{\sin^2 z} \quad \forall z \in \mathbb{C} \setminus \pi\mathbb{Z}\,.$$

Addition formulas:

$$\sin(z \pm w) = \sin z \cos w \pm \cos z \sin w \quad , \quad \cos(z \pm w) = \cos z \cos w \mp \sin z \sin w \quad \forall z,w \in \mathbb{C}\,,$$

and, whenever defined

$$\tan(z \pm w) = \frac{\tan z \pm \tan w}{1 \mp \tan z \tan w} \quad , \quad \arctan z \pm \arctan w = \arctan\left(\frac{z \pm w}{1 \mp zw}\right)\,.$$

0.4.4 Linear Algebra and Analysis

This course takes for granted that participants have been educated in the foundations of linear algebra by attending corresponding first-year introductory courses, for instance

- 401-0151-00L Lineare Algebra
- 401-0231-10L Analysis 1
- 401-0232-10L Analysis 2

Quite a few concepts and techniques introduced in those courses will be needed for and *will be taken for granted* in the current course.

Bibliography

[AG11] Uri M. Ascher and Chen Greif. *A first course in numerical methods*. Vol. 7. Computational Science & Engineering. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2011, pp. xxii+552. DOI: 10.1137/1.9780898719987 (cit. on p. 11).

[DR08] W. Dahmen and A. Reusken. *Numerik für Ingenieure und Naturwissenschaftler*. Heidelberg: Springer, 2008 (cit. on p. 11).

[DH03] P. Deuflhard and A. Hohmann. *Numerical Analysis in Modern Scientific Computing*. Vol. 43. Texts in Applied Mathematics. Springer, 2003 (cit. on p. 12).

[Fri19] F. Friedrich. *Datenstrukturen und Algorithmen*. Lecture slides. 2019 (cit. on p. 37).

[GGK14] W. Gander, M.J. Gander, and F. Kwok. *Scientific Computing*. Vol. 11. Texts in Computational Science and Engineering. Heidelberg: Springer, 2014 (cit. on p. 12).

[Gut09] M.H. Gutknecht. *Lineare Algebra*. Lecture Notes. SAM, ETH Zürich, 2009 (cit. on p. 12).

[Han02] M. Hanke-Bourgeois. *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens*. Mathematische Leitfäden. Stuttgart: B.G. Teubner, 2002 (cit. on p. 11).

[Jos12] N.M. Josuttis. *The C++ Standard Library*. Boston, MA: Addison-Wesley, 2012 (cit. on p. 30).

[LLM12] S. Lippman, J. Lajoie, and B. Moo. *C++ Primer*. 5th. Boston: Addison-Wesley, 2012 (cit. on pp. 30, 31).

[NS02] K. Nipp and D. Stoffer. *Lineare Algebra*. 5th ed. Zürich: vdf Hochschulverlag, 2002 (cit. on p. 12).

[QSS00] A. Quarteroni, R. Sacco, and F. Saleri. *Numerical mathematics*. Vol. 37. Texts in Applied Mathematics. New York: Springer, 2000 (cit. on p. 11).

[Str09] M. Struwe. *Analysis für Informatiker*. Lecture notes, ETH Zürich. 2009 (cit. on p. 12).

Chapter 1

Computing with Matrices and Vectors

§1.0.0.1 (Prerequisite knowledge for Chapter 1) The reader must master the basics of linear vector and matrix calculus as covered in every introductory course on linear algebra [NS02, Ch. 2].

On a few occasions we will also need results of 1D real calculus like Taylor's formula [Str09, Sect. 5.5]. ⌟

§1.0.0.2 (Levels of operations in simulation codes) The lowest level of real arithmetic available on computers are the elementary operations "+", "−", "∗", "\", "^", usually implemented in hardware. The next level comprises computations on finite arrays of real numbers, the elementary linear algebra operations (BLAS). On top of them we build complex algorithms involving iterations and approximations.

Complex iterative/recursive/approximative algorithms
Linear algebra operations on arrays (BLAS)
Elementary operations in $\mathbb{R}$

Hardly ever anyone will contemplate implementing elementary operations on binary data formats; similarly, well tested and optimised code libraries should be used for all elementary linear algebra operations in simulation codes. This chapter will introduce you to such libraries and how to use them smartly. ⌟

Contents

1.1 Fundamentals . . . 54
1.1.1 Notations . . . 54
1.1.2 Classes of Matrices . . . 56
1.2 Software and Libraries . . . 58
1.2.1 EIGEN . . . 58
1.2.2 PYTHON . . . 64
1.2.3 (Dense) Matrix Storage Formats . . . 65
1.3 Basic Linear Algebra Operations . . . 70
1.3.1 Elementary Matrix-Vector Calculus . . . 70
1.3.2 BLAS – Basic Linear Algebra Subprograms . . . 76
1.4 Computational Effort . . . 82
1.4.1 (Asymptotic) Computational Complexity . . . 83
1.4.2 Cost of Basic Linear-Algebra Operations . . . 84
1.4.3 Improving Complexity in Numerical Linear Algebra: Some Tricks . . . 86
1.5 Machine Arithmetic and Consequences . . . 91
1.5.1 Experiment: Loss of Orthogonality . . . 91
1.5.2 Machine Numbers . . . 94
1.5.3 Roundoff Errors . . . 97
1.5.4 Cancellation . . . 102

1.1 Fundamentals

1.1.1 Notations

Video tutorial for Section 1.1.1 "Notations and Classes of Matrices": (7 minutes)
Download link, tablet notes

→ review questions 1.1.2.9

The notations in this course try to adhere to established conventions. Since these may not be universal, idiosyncrasies cannot be avoided completely. Notations in textbooks may be different, beware!

Many considerations apply to real (field $\mathbb{R}$) and complex (field $\mathbb{C}$) numbers alike. Therefore we adopt the notatin $\mathbb{K}$ for a generic field of numbers. Thus, in this course, $\mathbb{K}$ will designate either $\mathbb{R}$ (real numbers) or $\mathbb{C}$ (complex numbers); complex arithmetic [Str09, Sect. 2.5] plays a crucial role in many applications, for instance in signal processing.

§1.1.1.1 (Notations for vectors)

- Vectors = are n-tuples ($n \in \mathbb{N}$) with components $\in \mathbb{K}$.

  vector = one-dimensional array (of real/complex numbers)

- Default in this lecture: vectors = column vectors

$$\underbrace{\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{K}^n}_{\text{column vector}} \quad \Bigg| \quad \underbrace{\begin{bmatrix} x_1 \cdots x_n \end{bmatrix} \in \mathbb{K}^{1,n}}_{\text{row vector}}$$

$\mathbb{K}^n \triangleq$ vector space of *column vectors* with n components in $\mathbb{K}$.

> "Linear algebra convention": Unless stated otherwise, in mathematical formulas vector components are indexed from 1!

✎ notation for column vectors: **bold** small roman letters, e.g. $\mathbf{x}, \mathbf{y}, \mathbf{z}$

- Transposing: $\begin{cases} \text{column vector} & \mapsto & \text{row vector} \\ \text{row vector} & \mapsto & \text{column vector} \end{cases}$.

$$\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}^\top = \begin{bmatrix} x_1 \cdots x_n \end{bmatrix} \quad , \quad \begin{bmatrix} x_1 \cdots x_n \end{bmatrix}^\top = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

✎ Notation for row vectors: $\mathbf{x}^\top, \mathbf{y}^\top, \mathbf{z}^\top$

- Addressing vector components:

✎ two notations:

$$\begin{array}{rcl} \mathbf{x} = [x_1, \ldots, x_n]^\top & \rightarrow & x_i, \quad i = 1, \ldots, n \\ \mathbf{x} \in \mathbb{K}^n & \rightarrow & (\mathbf{x})_i, \quad i = 1, \ldots, n \end{array}$$

- Selecting sub-vectors:

  ✎ notation: $\mathbf{x} = [x_1 \ldots x_n]^\top \succ (\mathbf{x})_{k:l} = [x_k, \ldots, x_l]^\top, 1 \le k \le l \le n$

  ✎ notations like $1 \le k, \ell \le n$ where it is clear from the context that k and ℓ designate integer indices mean "for all $k, \ell \in \{1, \ldots, n\}$".

- j-th unit vector: $\mathbf{e}_j = [0, \ldots, 1, \ldots, 0]^\top$, $(\mathbf{e}_j)_i = \delta_{ij}$, $i, j = 1, \ldots, n$.

  ✎ notation: Kronecker symbol, also called "Kronecker delta", and defined as $\delta_{ij} := 1$, if $i = j$, $\delta_{ij} := 0$, if $i \neq j$.

⌟

§1.1.1.2 (Notations and notions for matrices)

- Matrices = two-dimensional arrays of real/complex numbers

$$\mathbf{A} := \begin{bmatrix} a_{11} & \ldots & a_{1m} \\ \vdots & & \vdots \\ a_{n1} & \ldots & a_{nm} \end{bmatrix} \in \mathbb{K}^{n,m}, \quad n, m \in \mathbb{N}.$$

vector space of $n \times m$-matrices: ($n \hat{=}$ number of rows, $m \hat{=}$ number of columns)

  ✎ notation for matrices: **bold** CAPITAL roman letters, e.g., $\mathbf{A}, \mathbf{S}, \mathbf{Y}$
  Special cases: $\mathbb{K}^{n,1} \leftrightarrow$ column vectors, $\mathbb{K}^{1,n} \leftrightarrow$ row vectors

- Writing a matrix as a tuple of its columns or rows

$$\mathbf{c}_i \in \mathbb{K}^n, \ i = 1, \ldots, m \ \triangleright \ \mathbf{A} := [\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_m] \in \mathbb{K}^{n,m},$$

$$\mathbf{r}_i \in \mathbb{K}^m, \ i = 1, \ldots, n \ \triangleright \ \mathbf{A} := \begin{bmatrix} \mathbf{r}_1^\top \\ \vdots \\ \mathbf{r}_n^\top \end{bmatrix} \in \mathbb{K}^{n,m}.$$

- Addressing matrix entries & sub-matrices (✎ notations):

$$\mathbf{A} := \begin{bmatrix} a_{11} & \ldots & a_{1m} \\ \vdots & & \vdots \\ a_{n1} & \ldots & a_{nm} \end{bmatrix}$$

→ matrix entry/matrix element $(\mathbf{A})_{i,j} := a_{ij}$, $1 \le i \le n, 1 \le j \le m$,
→ i-th row, $1 \le i \le n$: $(\mathbf{A})_{i,:} := [a_{i,1}, \ldots, a_{i,m}]$,
→ j-th column, $1 \le j \le m$: $(\mathbf{A})_{:,j} := [a_{1,j}, \ldots, a_{n,j}]^\top$,
→ matrix block (sub-matrix) $(\mathbf{A})_{k:\ell,r:s} := \left[a_{ij}\right]_{\substack{i=k,\ldots,\ell \\ j=r,\ldots,s}}$, $1 \le k \le \ell \le n$, $1 \le r \le s \le m$.

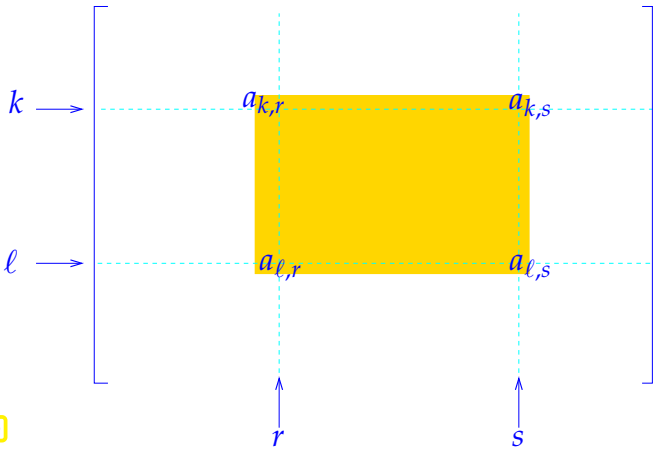


Fig. 5

The colon (:) range notation is inspired by MATLAB's/PYTHON's matrix addressing conventions. $(\mathbf{A})_{k:l,r:s}$ is a matrix of size $(l-k+1) \times (s-r+1)$.

Note that in PYTHON the : notation describes slightly different ranges: the end value is excluded.

✦ Transposed matrix:

$$\mathbf{A}^\top = \begin{bmatrix} a_{11} & \dots & a_{1m} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{nm} \end{bmatrix}^\top := \begin{bmatrix} a_{11} & \dots & a_{n1} \\ \vdots & & \vdots \\ a_{1m} & \dots & a_{mn} \end{bmatrix} \in \mathbb{K}^{m,n} .$$

✦ Adjoint matrix (Hermitian transposed):

$$\mathbf{A}^{\mathrm{H}} := \begin{bmatrix} a_{11} & \dots & a_{1m} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{nm} \end{bmatrix}^{\mathrm{H}} := \begin{bmatrix} \bar{a}_{11} & \dots & \bar{a}_{n1} \\ \vdots & & \vdots \\ \bar{a}_{1m} & \dots & \bar{a}_{mn} \end{bmatrix} \in \mathbb{K}^{m,n} .$$

✎ notation: $\bar{a}_{ij} = \mathfrak{Re}(a_{ij}) - i\mathfrak{Im}(a_{ij})$ complex conjugate of a_{ij}. Of course, for $\mathbf{A} \in \mathbb{R}^{n,m}$ we have $\mathbf{A}^{\mathrm{H}} = \mathbf{A}^\top$.

⌟

1.1.2 Classes of Matrices

Most matrices occurring in mathematical modelling have a special structure. This section presents a few of these. More will come up throughout the remainder of this chapter; see also [AG11, Sect. 4.3].

§1.1.2.1 (Special matrices) Terminology and notations for a few very special matrices:

Identity matrix: $$\mathbf{I} := \mathbf{I}_n := \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix} \in \mathbb{K}^{n,n} ,$$

Zero matrix: $$\mathbf{O} := \mathbf{O}_{n,m} := \begin{bmatrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{bmatrix} \in \mathbb{K}^{n,m} ,$$

Diagonal matrix: $$\operatorname{diag}(d_1, \dots, d_n) := \begin{bmatrix} d_1 & & 0 \\ & \ddots & \\ 0 & & d_n \end{bmatrix} \in \mathbb{K}^{n,n} , \quad d_j \in \mathbb{K} , \quad j = 1, \dots, n .$$

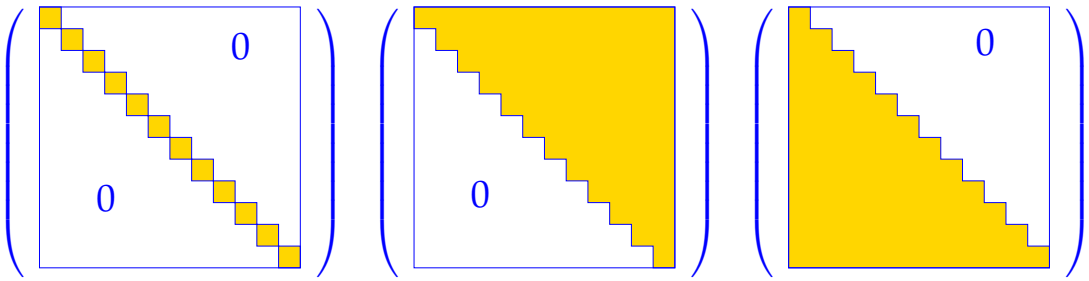
The creation of special matrices can usually be done by special commands or functions in the various languages or libraries dedicated to numerical linear algebra, see § 1.2.1.3. ⌟

§1.1.2.2 (Diagonal and triangular matrices) A little terminology to quickly refer to matrices whose non-zero entries occupy special locations:

Definition 1.1.2.3. Types of matrices

A matrix $\mathbf{A} = [a_{ij}] \in \mathbb{K}^{m,n}$ is a
- diagonal matrix, if $a_{ij} = 0$ for $i \neq j$,
- upper triangular matrix, if $a_{ij} = 0$ for $i > j$,
- lower triangular matrix, if $a_{ij} = 0$ for $i < j$.

A triangular matrix is normalized, if $a_{ii} = 1$, $i = 1, \dots, \min\{m, n\}$.



diagonal matrix upper triangular lower triangular

⌟

§1.1.2.4 (Symmetric matrices)

Definition 1.1.2.5. Hermitian/symmetric matrices

A matrix $\mathbf{M} \in \mathbb{K}^{n,n}$, $n \in \mathbb{N}$, is Hermitian, if $\mathbf{M}^{\mathrm{H}} = \mathbf{M}$. If $\mathbb{K} = \mathbb{R}$, the matrix is called symmetric.

Definition 1.1.2.6. Symmetric positive definite (s.p.d.) matrices → [DR08, Def. 3.31], [QSS00, Def. 1.22]

$\mathbf{M} \in \mathbb{K}^{n,n}$, $n \in \mathbb{N}$, is **symmetric (Hermitian) positive definite** (s.p.d.), if

$$\mathbf{M} = \mathbf{M}^{\mathrm{H}} \quad \text{and} \quad \forall \mathbf{x} \in \mathbb{K}^n: \ \mathbf{x}^{\mathrm{H}}\mathbf{M}\mathbf{x} > 0 \ \Leftrightarrow \ \mathbf{x} \neq 0 \, .$$

If $\mathbf{x}^{\mathrm{H}}\mathbf{M}\mathbf{x} \geq 0$ for all $\mathbf{x} \in \mathbb{K}^n$ ▷ $\mathbf{M}$ **positive semi-definite**.

Lemma 1.1.2.7. Necessary conditions for s.p.d. → [DR08, Satz 3.33], [QSS00, Prop. 1.18]

For a symmetric/Hermitian positive definite matrix $\mathbf{M} = \mathbf{M}^{\mathrm{H}} \in \mathbb{K}^{n,n}$ holds true:

1. $m_{ii} > 0, i = 1, \ldots, n$,
2. $m_{ii}m_{jj} - |m_{ij}|^2 > 0 \quad \forall 1 \leq i < j \leq n$,
3. *all eigenvalues of $\mathbf{M}$ are positive.* *(← also sufficient for symmetric/Hermitian $\mathbf{M}$)*

⌟

Remark 1.1.2.8 (S.p.d. Hessians) Recall from analysis: in an isolated local minimum x^* of a C^2-function $f : \mathbb{R}^n \mapsto \mathbb{R}$ ➤ Hessian $\mathrm{D}^2 f(x^*)$ s.p.d. (see Def. 8.5.1.18 for the definition of the Hessian)

To compute the minimum of a C^2-function iteratively by means of Newton's method (→ Sect. 8.5) a linear system of equations with the s.p.d. Hessian as system matrix has to be solved in each step.

The solutions of many equations in science and engineering boils down to finding the minimum of some (energy, entropy, etc.) function, which accounts for the prominent role of s.p.d. linear systems in applications. ⌟

Review question(s) 1.1.2.9 (Notations, matrix-vector calculus, and special matrices)

(Q1.1.2.9.A) Give a compact notation for the row vector containing the diagonal entries of a square matrix $\mathbf{S} \in \mathbb{R}^{n,n}$, $n \in \mathbb{N}$.

(Q1.1.2.9.B) How can you write down the $s \times s$-submatrix, $s \in \mathbb{N}$, in the upper right corner of $\mathbf{C} \in \mathbb{R}^{n,m}$, $n, m \geq s$.

(Q1.1.2.9.C) We consider two matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n,m}$, both with at most $N \in \mathbb{N}$ non-zero entries. What is the maximal number of non-zero entries of $\mathbf{A} + \mathbf{B}$?

(Q1.1.2.9.D) A matrix $\mathbf{A} \in \mathbb{R}^{n,m}$ enjoys the following property (banded matrix):

$$i \in \{1,\ldots,n\},\ j \in \{1,\ldots,m\},\ i-j \notin \{-B_-,\ldots,B_+\} \quad \Rightarrow \quad (\mathbf{A})_{ij} = 0\,,$$

for given $B_-, B_+ \in \mathbb{N}_0$. What is the maximal number of non-zero entries of $\mathbf{A}$.

(Q1.1.2.9.E) A matrix $\mathbf{A}$ with real entries is known to be **skew-symmetric**: $\mathbf{A}^\top = -\mathbf{A}$ What does this tell us about $\mathbf{A}$ and its entries?

(Q1.1.2.9.F) What is the dimension of the vector space of

- symmetric matrices $\in \mathbb{R}^{n,n}$?
- skew-symmetric matrices $\in \mathbb{R}^{n,n}$, see Question (**Q**1.1.2.9.E)?
- lower triangular matrices $\in \mathbb{R}^{n,n}$?
- diagonal matrices $\in \mathbb{R}^{n,n}$?

△

1.2 Software and Libraries

Whenever algorithms involve matrices and vectors (in the sense of linear algebra) it is advisable to rely on suitable code libraries or numerical programming environments.

1.2.1 EIGEN

Video tutorial for Section 1.2.1 "EIGEN ": (11 minutes) Download link, tablet notes

→ review questions 1.2.1.14

Currently, the most widely used programming language for the development of new simulation software in scientific and industrial high-performance computing is C++. In this course we are going to use and discuss EIGEN as an example for a C++ library for numerical linear algebra ("embedded" domain specific language: DSL).

EIGEN is a header-only C++ template library designed to enable easy, natural and efficient numerical linear algebra: it provides data structures and a wide range of operations for matrices and vectors, see below. EIGEN also implements many more fundamental algorithms documentation page or the discussion below).

EIGEN relies on expression templates to allow the efficient evaluation of complex expressions involving matrices and vectors. Refer to the example given in the EIGEN documentation for details.

➥ Link to an "EIGEN Cheat Sheet" (quick reference relating to MATLAB commands)

§1.2.1.1 (Matrix and vector data types in EIGEN) A generic matrix data type is given by the templated class

```
Eigen::Matrix<typename Scalar,
              int RowsAtCompileTime, int ColsAtCompileTime>
```

Here **Scalar** is the underlying scalar type of the matrix entries, which must support the usual operations '+','-','*','/', and '+=', '*=', '', etc. Usually the scalar type will be either double, float, or complex<>. The cardinal template arguments `RowsAtCompileTime` and `ColsAtCompileTime` can pass a fixed size of the matrix, if it is known at compile time. There is a specialization selected by the template argument `Eigen::Dynamic` supporting variable size "dynamic" matrices.

C++ code 1.2.1.2: Vector type and their use in EIGEN ⇝ GITLAB

```
1  #include <Eigen/Dense>
2
3  template<typename Scalar>
4  void eigenTypeDemo(unsigned int dim)
5  {
6    // General dynamic (variable size) matrices
7    using dynMat_t = Eigen::Matrix<Scalar,Eigen::Dynamic,Eigen::Dynamic>;
8    // Dynamic (variable size) column vectors
9    using dynColVec_t = Eigen::Matrix<Scalar,Eigen::Dynamic,1>;
10   // Dynamic (variable size) row vectors
11   using dynRowVec_t = Eigen::Matrix<Scalar,1,Eigen::Dynamic>;
12   using index_t = typename dynMat_t::Index;
13   using entry_t = typename dynMat_t::Scalar;
14
15   // Declare vectors of size 'dim'; not yet initialized
16   dynColVec_t colvec(dim);
17   dynRowVec_t rowvec(dim);
18   // Initialisation through component access
19   for(index_t i=0; i< colvec.size(); ++i) colvec[i] = static_cast<Scalar>(i);
20   for(index_t i=0; i< rowvec.size(); ++i) rowvec[i] = static_cast<Scalar>(1)/(i+1);
21   colvec[0] = static_cast<Scalar>(3.14);
22   rowvec[dim-1] = static_cast<Scalar>(2.718);
23   // Form tensor product, a matrix, see Section 1.3.1
24   dynMat_t vecprod = colvec*rowvec;
25   const int nrows = vecprod.rows();
26   const int ncols = vecprod.cols();
27 }
```

Note that in Line 24 we could have relied on automatic type deduction via **auto** `vectprod = ...`. However, as argued in Rem. 0.3.4.4 often it is safer to forgo this option and specify the type directly

The following convenience data types are provided by EIGEN, see EIGEN documentation:

- **MatrixXd** ≙ generic variable size matrix with **d**ouble precision entries
- **VectorXd**, **RowVectorXd** ≙ dynamic column and row vectors
  (**=** dynamic matrices with one dimension equal to 1)
- **MatrixNd** with $N = 2,3,4$ for small fixed size square $N \times N$-matrices (type `double`)
- **VectorNd** with $N = 2,3,4$ for small column vectors with fixed length N.

The `d` in the type name may be replaced with `i` (for `int`), `f` (for `float`), and `cd` (for `complex<double>`) to select another basic scalar type.

All matrix type feature the methods `cols()`, `rows()`, and `size()` telling the number of columns, rows, and total number of entries.

Access to individual matrix entries and vector components, both as Rvalue and Lvalue, is possible through the `()`-operator taking two arguments of type `index_t`. If only one argument is supplied, the matrix is accessed as a linear array according to its memory layout. For vectors, that is, matrices where one

dimension is fixed to 1, the `[]`-operator can replace `()` with one argument, see Line 22 of Code 1.2.1.2. ⌟

§1.2.1.3 (Initialization of *dense* matrices in EIGEN, EIGEN documentation) The entry access operator `(int i,int j)` allows the most direct setting of matrix entries; there is hardly any runtime penalty.

Of course, in EIGEN dedicated functions take care of the initialization of the special matrices introduced in § 1.1.2.1:

```
Eigen::MatrixXd I = Eigen::MatrixXd::Identity(n,n);
Eigen::MatrixXd O = Eigen::MatrixXd::Zero(n,m);
Eigen::MatrixXd D = d_vector.asDiagonal();
```

C++ code 1.2.1.4: Initializing special matrices in EIGEN, ⇝ GITLAB

```
1  #include <Eigen/Dense >
2  // Just allocate space for matrix, no initialisation
3  Eigen::MatrixXd A(rows,cols);
4  // Zero matrix.  Similar to matlab command zeros(rows, cols);
5  Eigen::MatrixXd B = MatrixXd::Zero(rows, cols);
6  // Ones matrix.  Similar to matlab command ones(rows, cols);
7  Eigen::MatrixXd C = MatrixXd::Ones(rows, cols);
8  // Matrix with all entries same as value.
9  Eigen::MatrixXd D = MatrixXd::Constant(rows, cols, value);
10 // Random matrix, entries uniformly distributed in [0,1]
11 Eigen::MatrixXd E = MatrixXd::Random(rows, cols);
12 // (Generalized) identity matrix, 1 on main diagonal
13 Eigen::MatrixXd I = MatrixXd::Identity(rows,cols);
14 std::cout << "size of A = (" << A.rows() << ',' << A.cols() << ')' << std::endl;
```

A versatile way to initialize a matrix relies on a combination of the operators « and `,`, which allows the construction of a matrix from blocks, see ⇝ GITLAB, function `blockinit()`.

```
MatrixXd mat3(6,6);
mat3 <<
    MatrixXd::Constant(4,2,1.5), // top row, first block
    MatrixXd::Constant(4,3,3.5), // top row, second block
    MatrixXd::Constant(4,1,7.5), // top row, third block
    MatrixXd::Constant(2,4,2.5), // bottom row, left block
    MatrixXd::Constant(2,2,4.5); // bottom row, right block
```

The matrix is filled top to bottom left to right, block dimensions have to match (like in MATLAB). ⌟

§1.2.1.5 (Access to submatrices in EIGEN, EIGEN documentation) The method `block(int i,int j,int p,int q)` returns a reference to the submatrix with upper left corner at position (i,j) and size $p \times q$.

The methods `row(int i)` and `col(int j)` provide a reference to the corresponding row and column of the matrix. Even more specialised access methods are

`topLeftCorner(p,q)`, `bottomLeftCorner(p,q)`,
`topRightCorner(p,q)`, `bottomRightCorner(p,q)`,
`topRows(q)`, `bottomRows(q)`,
`leftCols(p)`, and `rightCols(q)`,

with obvious purposes.

C++ code 1.2.1.6: Demonstration code for access to matrix blocks in EIGEN ⇝ GITLAB

```
2  template<typename MatType>
3  void blockAccess(Eigen::MatrixBase<MatType> &M)
4  {
5    using index_t = typename Eigen::MatrixBase<MatType>::Index;
6    const index_t nrows(M.rows()); // No.  of rows
7    const index_t ncols(M.cols()); // No.  of columns
8
9    cout << "Matrix M = " << endl << M << endl; // Print matrix
10   // Block size half the size of the matrix
11   const index_t p = nrows/2;
12   const index_t q = ncols/2;
13   // Output submatrix with left upper entry at position (i,i)
14   for(index_t i=0; i < std::min(p,q); i++) {
15     cout << "Block (" << i << ',' << i << ',' << p << ',' << q
16          << ") = " << M.block(i,i,p,q) <<  endl;
17   }
18   // l-value access:  modify sub-matrix by adding a constant
19   M.block(1,1,p,q) += Eigen::MatrixBase<MatType>::Constant(p,q,1.0);
20   cout << "M = " << endl << M << endl;
21   // r-value access:  extract sub-matrix
22   const MatrixXd B = M.block(1,1,p,q);
23   cout << "Isolated modified block = " << endl << B << endl;
24   // Special sub-matrices
25   cout << p << " top rows of m = " << M.topRows(p) << endl;
26   cout << p << " bottom rows of m = " << M.bottomRows(p) << endl;
27   cout << q << " left cols of m = " << M.leftCols(q) << endl;
28   cout << q << " right cols of m = " << M.rightCols(p) << endl;
29   // r-value access to upper triangular part
30   const MatrixXd T = M.template triangularView<Upper>(); //
31   cout << "Upper triangular part = " << endl << T << endl;
32   // l-value access to upper triangular part
33   M.template triangularView<Lower>() *= -1.5; //
34   cout << "Matrix M = " << endl << M << endl;
35 }
```

- Note that the function `blockAccess()` is templated and that the matrix argument passed through `M` has a type derived from **Eigen::MatrixBase**. The deeper reason for this alien looking signature of `blockAccess()` is explained in EIGEN documentation.
- EIGEN offers views for access to triangular parts of a matrix, see Line 30 and Line 33, according to

  `M.triangularView<XX>()`

  where `XX` can stand for one of the following: Upper, Lower, StrictlyUpper, StrictlyLower, UnitUpper, UnitLower, see EIGEN documentation.
- For column and row vectors references to sub-vectors can be obtained by the methods `head(int length)`, `tail(int length)`, and `segment(int pos, int length)`.

Note: Unless the preprocessor switch `NDEBUG` is set, EIGEN performs range checks on all indices. ⌟

§1.2.1.7 (Componentwise operations in EIGEN) Running out of overloadable operators, EIGEN uses the Array concept to furnish entry-wise operations on matrices. An EIGEN-Array contains the same data as a matrix, supports the same methods for initialisation and access, but replaces the operators of matrix arithmetic with entry-wise actions. Matrices and arrays can be converted into each other by the `array()` and `matrix()` methods, see EIGEN documentation for details. Information about functions that enable entry-wise operation is available in the EIGEN documentation.

C++ code 1.2.1.8: Using Array in EIGEN ⇝ GITLAB

```
2  void inline matArray(int nrows,int ncols){
3    Eigen::MatrixXd m1(nrows,ncols);
4    Eigen::MatrixXd m2(nrows,ncols);
5    for(int i = 0; i < m1.rows(); i++) {
6      for(int j = 0; j < m1.cols(); j++) {
7        m1(i,j) = static_cast<double>((i+1))/(j+1);
8        m2(i,j) = static_cast<double>((j+1))/(i+1);
9      }
10   }
11   // Entry-wise product, not a matrix product
12   const Eigen::MatrixXd m3 = (m1.array() * m2.array()).matrix();
13   // Explicit entry-wise operations on matrices are possible
14   const Eigen::MatrixXd m4(m1.cwiseProduct(m2));
15   // Entry-wise logarithm
16   cout << "Log(m1) = " << endl << log(m1.array()) << endl;
17   // Entry-wise boolean expression, true cases counted
18   cout << (m1.array() > 3).count() << " entries of m1 > 3" << endl;
19 }
```

The application of a functor (→ Section 0.3.3) to all entries of a matrix can also be done via the `unaryExpr()` method of a matrix:

```
// Apply a lambda function to all entries of a matrix
auto fnct = [](double x) { return (x+1.0/x); };
cout << "f(m1) = " << endl << m1.unaryExpr(fnct) << endl;
```

⌟

§1.2.1.9 (Reduction operations in EIGEN) According to EIGEN's terminology, reductions are operations that access all entries of a matrix and accumulate some information in the process EIGEN documentation. A typical example is the summation of the entries.

C++ code 1.2.1.10: Summation reduction in EIGEN ⇝ GITLAB

```
2  template <class Matrix>
3  void sumEntries(Eigen::MatrixBase<Matrix> &M) {
4    using Scalar = typename Eigen::MatrixBase<Matrix>::Scalar;
5    // Compute sum of all entries
6    const Scalar s = M.sum();
7    // Row-wise and column-wise sum of entries:  results are vectors
8    const Eigen::Matrix<Scalar, 1, Eigen::Dynamic> colsums{M.rowwise().sum()};
9    const Eigen::Matrix<Scalar, Eigen::Dynamic, 1> rowsums{M.colwise().sum()};
10   std::cout << M.rows() << 'x' << M.cols() << "-matrix: " << colsums.sum()
11             << " = " << rowsums.sum() << " = " << s << std::endl;
12 }
```

⌟

Remark 1.2.1.11 ('auto' in EIGEN codes) The expression template programming model (→ explanations from EIGEN documentation) relies on complex intermediary data types hidden from the user . They support the efficient evaluation of complex expressions EIGEN documentation. Let us look at the following two code snippets that assume that both `M` and `R` are of type **Eigen::MatrixXd**.

Code I:

```
auto D = M.diagonal().asDiagonal(); R = D.inverse();
```

Code II:

```
Eigen::MatrixXd D = M.diagonal().asDiagonal(); R = D.inverse();
```

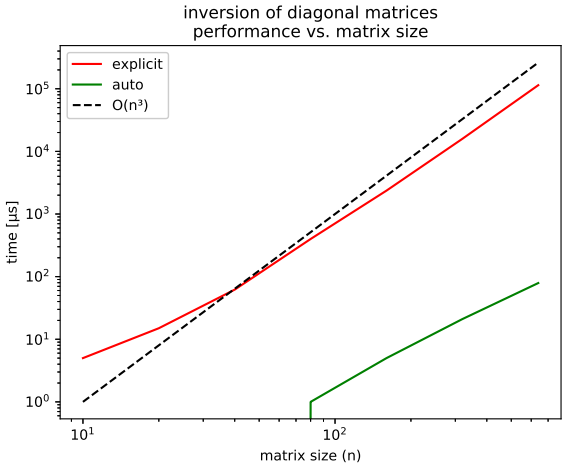


Fig. 6

(Quad-Core Intel Core i7 @ 3.1 GHz L2 256 KB, L3 8MB, Mem 16 GB, macOS 10.15.5, clang version 11.0.3 -02, NDEBUG)

We observe that for large matrices Code I ("auto", —, values $\leq 0.5\mu s$ suppressed) runs much faster than Code II ("explicit", —) though they are "algebraically equivalent".

The reason is that in Code I `D` is of a complex type that preserves the information that the matrix is diagonal. Of course, inverting a diagonal matrix is cheap. Conversely forcing `D` to be of type **Eigen::MatrixXd** loses this information and the expensive `invert()` method for a generic densely populated matrix is invoked.

This is one of the exceptions to Rem. 0.3.4.4: for variables holding the result of EIGEN expressions **auto** is recommended. ⌟

Remark 1.2.1.12 (EIGEN-based code: debug mode and release mode) If you want a C++ code built using the EIGEN library run fast, for instance, for large computations or runtime measurements, you should compile in release mode, that is, with the compiler switches `-O2 -DNDEBUG` (for gcc or clang). In a **cmake**-based build system you can achieve this by setting the flag `CMAKE_BUILD_TYPE` to "Release".

The default setting for EIGEN is debug mode, which makes EIGEN do a lot of consistency checking and considerably slows down execution of a code.

For "production runs" EIGEN-based codes must be compiled in release mode! ⌟

Remark 1.2.1.13 (EIGEN in use)

- ☞ EIGEN is used as one of the base libraries for the Robot Operating System (ROS), an open source project with strong ETH participation.
- ☞ The geometry processing library libigl uses EIGEN as its basic linear algebra engine. It is being used and developed at ETH Zurich, at the Interactive Geometry Lab and Advanced Technologies Lab.

⌟

Review question(s) 1.2.1.14 (EIGEN)

For the following questions you may consult the EIGEN documentation.

(Q1.2.1.14.A) Outline a C++ function

```
template <typename Matrix>
  void replaceWithId(Eigen::DenseBase<Matrix> &M);
```

that checks whether the matrix is an $n \times n$-matrix with even $n \in \mathbb{N}$ and then replaces its upper right $n/2 \times n/2$-block with an identity matrix. Do not use any C++ loops.

(Q1.2.1.14.B) Given an **Eigen::VectorXd** object `v` ($\leftrightarrow \mathbf{v} \in \mathbb{R}^n$), sketch a C++ code snippet that replaces it with a vector $\widetilde{\mathbf{v}}$ defined by

$$(\widetilde{\mathbf{v}})_i := \begin{cases} (\mathbf{v})_n & \text{for } i = 1\,, \\ (\mathbf{v})_{i-1} & \text{for } i = 2,\ldots,n\,. \end{cases}$$

Do not use C++ loops. Can you see a problem?

(Q1.2.1.14.C) Given a matrix $\mathbf{M} \in \mathbb{R}^{m,n}$ stored in an **Eigen::MatrixXd** object `M` write down a C++ code snipper that initializes another variable `Mext` of type **Eigen::MatrixXd** corresponding to

$$\widetilde{\mathbf{M}} := \begin{bmatrix} \mathbf{M} & \mathbf{0} \\ \mathbf{0}^\top & 1 \end{bmatrix} \in \mathbb{R}^{m+1,n+1}$$

using EIGEN's << matrix construction operator EIGEN documentation.

(Q1.2.1.14.D) Learn about the methods `head()` and `tail()` from the EIGEN documentation and express them by means of the `block()` method applied to the same variable.

△

1.2.2 PYTHON

PYTHON is a widely used general-purpose and open source programming language. Together with the packages like NUMPY and MATPLOTLIB it delivers similar functionality like MATLAB for free. For interactive computing IPYTHON can be used. All those packages belong to the SCIPY ecosystem.

PYTHON features a good documentation and several scientific distributions are available (e.g. Anaconda, Enthought) which contain the most important packages. On most Linux-distributions the SCIPY ecosystem is also available in the software repository, as well as many other packages including for example the Spyder IDE delivered with Anaconda.

A good introduction tutorial to numerical PYTHON are the SCIPY-lectures. The full documentation of NUMPY and SCIPY can be found here. For former MATLAB-users there's also a guide. The scripts in this lecture notes follow the official PYTHON style guide.

Note that in PYTHON we have to import the numerical packages explicitly before use. This is normally done at the beginning of the file with lines like **import** `numpy` **as** `np` and **from** `matplotlib` **import** `pyplot` **as** `plt`. Those import statements are often skipped in this lecture notes to focus on the actual computations. But you can always assume the import statements as given here, e.g. `np.ravel(A)` is a call to a NUMPY function and `plt.loglog(x, y)` is a call to a MATPLOTLIB pyplot function.

PYTHON is not used in the current version of the lecture. Nevertheless a few PYTHON codes are supplied in order to convey similarities and differences to implementations in MATLAB and C++.

§1.2.2.1 (Matrices and Vectors in PYTHON) The basic numeric data type in PYTHON are NUMPY's n-dimensional arrays. Vectors are normally implemented as 1D arrays and no distinction is made between row and column vectors. Matrices are represented as 2D arrays.

- ☞ `v = np.array([1, 2, 3])` creates a 1D array with the three elements 1, 2 and 3.
- ☞ `A = np.array([[1, 2], [3, 4]]` creates a 2D array.
- ☞ `A.shape` gives the n-dimensional size of an array.
- ☞ `A.size` gives the total number of entries in an array.

Note: There's also a matrix class in NUMPY with different semantics but its use is officially discouraged and it might even be removed in future release. ⌟

§1.2.2.2 (Manipulating arrays in PYTHON) There are many possibilities listed in the documentation how to create, index and manipulate arrays.

An important difference to MATLAB is, that all arithmetic operations are normally performed element-wise, e.g. `A * B` is not the matrix-matrix product but element-wise multiplication (in MATLAB: `A.*A`). Also `A * v` does a broadcasted element-wise product. For the matrix product one has to use `np.dot(A, B)` or `A.dot(B)` explicitly. ⌟

1.2.3 (Dense) Matrix Storage Formats

Video tutorial for Section 1.2.3 "(Dense) Matrix Storage Formats": (10 minutes) Download link, tablet notes

→ review questions 1.2.3.11

All numerical libraries store the entries of a (generic = *dense*) matrix $\mathbf{A} \in \mathbb{K}^{m,n}$ in a linear array of length mn (or longer). Accessing entries entails suitable index computations.

Two natural options for "vectorisation" of a matrix: *row major, column major*

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

Row major (C-arrays, bitmaps, Python):

A_arr	1	2	3	4	5	6	7	8	9

Column major (Fortran, MATLAB, EIGEN):

A_arr	1	4	7	2	5	8	3	6	9

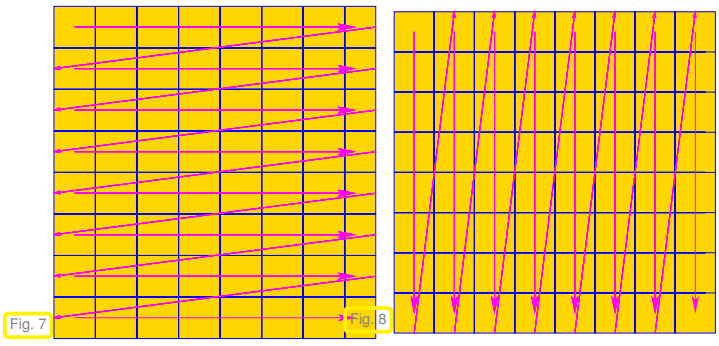
Access to entry $(\mathbf{A})_{ij}$ of $\mathbf{A} \in \mathbb{K}^{n,m}$, $i = 1,\ldots,n$, $j = 1,\ldots,m$:

row major:

$$(\mathbf{A})_{ij} \leftrightarrow \texttt{A_arr(m*(i-1)+(j-1))}$$

column major:

$$(\mathbf{A})_{ij} \leftrightarrow \texttt{A_arr(n*(j-1)+(i-1))}$$



row major column major

EXAMPLE 1.2.3.1 (Accessing matrix data as a vector) In EIGEN the single index access operator relies on the linear data layout:

In EIGEN the data layout can be controlled by a template argument; default is column major.

C++ code 1.2.3.2: Single index access of matrix entries in EIGEN ↪ GITLAB

```
2 void storageOrder(int nrows=6,int ncols=7)
3 {
4   cout << "Different matrix storage layouts in Eigen" << endl;
5   // Template parameter ColMajor selects column major data layout
6   Matrix<double,Dynamic,Dynamic,ColMajor> mcm(nrows,ncols);
7   // Template parameter RowMajor selects row major data layout
8   Matrix<double,Dynamic,Dynamic,RowMajor> mrm(nrows,ncols);
```

```
  // Direct initialization; lazy option:  use int as index type
  for (int l=1,i= 0; i< nrows; i++) {
    for (int j= 0; j< ncols; j++,l++) {
      mcm(i,j) = mrm(i,j) = l;
    }
  }

  cout << "Matrix mrm = " << endl << mrm << endl;
  cout << "mcm linear = ";
  for (int l=0;l < mcm.size(); l++) {
    cout << mcm(l) << ',';
  }
  cout << endl;

  cout << "mrm linear = ";
  for (int l=0;l < mrm.size(); l++) {
    cout << mrm(l) << ',';
  }
  cout << endl;
}
```

The function call `storageOrder(3,3)`, *cf.* Code 1.2.3.2 yields the output

```
Different matrix storage layouts in Eigen
Matrix mrm =
1 2 3
4 5 6
7 8 9
mcm linear = 1,4,7,2,5,8,3,6,9,
mrm linear = 1,2,3,4,5,6,7,8,9,
```

In PYTHON the default data layout is row major, but it can be explicitly set. Further, array transposition does not change any data, but only the memory order and array shape.

PYTHON-code 1.2.3.3: Storage order in PYTHON

```
# array creation
A = np.array([[1, 2], [3, 4]])  # default (row major) storage
B = np.array([[1, 2], [3, 4]], order='F')  # column major storage

# show internal storage
np.ravel(A, 'K')  # array elements as stored in memory:  [1, 2, 3, 4]
np.ravel(B, 'K')  # array elements as stored in memory:  [1, 3, 2, 4]

# nothing happens to the data on transpose, just the storage order
    changes
np.ravel(A.T, 'K')  # array elements as stored in memory:  [1, 2, 3, 4]
np.ravel(B.T, 'K')  # array elements as stored in memory:  [1, 3, 2, 4]

# storage order can be accessed by checking the array's flags
A.flags['C_CONTIGUOUS']  # True
B.flags['F_CONTIGUOUS']  # True
A.T.flags['F_CONTIGUOUS']  # True
B.T.flags['C_CONTIGUOUS']  # True
```

Remark 1.2.3.4 (Vectorisation of a matrix) Mapping a column-major matrix to a column vector with the same number of entries is called vectorization or linearization in numerical linear algebra, in symbols

$$\mathrm{vec} : \mathbb{K}^{n,m} \to \mathbb{K}^{n\cdot m} \quad , \quad \mathrm{vec}(\mathbf{A}) := \begin{bmatrix} (\mathbf{A})_{:,1} \\ (\mathbf{A})_{:,2} \\ \vdots \\ (\mathbf{A})_{:,m} \end{bmatrix} \in \mathbb{R}^{n\cdot m} \,. \tag{1.2.3.5}$$

⌟

Remark 1.2.3.6 (Reshaping matrices in EIGEN) If you need a reshaped view of a matrix' data in EIGEN you can obtain it via the raw data vector belonging to the matrix. Then use this information to create a matrix view by means of `Map` → documentation.

C++ code 1.2.3.7: Demonstration on how reshape a matrix in EIGEN ↬ GITLAB

```
2  template<typename MatType>
3  void reshapetest(MatType &M)
4  {
5    using index_t = typename MatType::Index;
6    using entry_t = typename MatType::Scalar;
7    const index_t nsize(M.size());
8
9   // reshaping possible only for matrices with non-prime dimensions
10   if ((nsize %2) == 0) {
11     entry_t *Mdat = M.data(); // raw data array for M
12     // Reinterpretation of data of M
13     Map<Eigen::Matrix<entry_t,Dynamic,Dynamic>> R(Mdat,2,nsize/2);
14     // (Deep) copy data of M into matrix of different size
15     const Eigen::Matrix<entry_t,Dynamic,Dynamic> S =
16       Map<Eigen::Matrix<entry_t,Dynamic,Dynamic>>(Mdat,2,nsize/2);
17
18     cout << "Matrix M = " << endl << M << endl;
19     cout << "reshaped to " << R.rows() << 'x' << R.cols()
20       << " = " << endl << R << endl;
21     // Modifying R affects M, because they share the data space !
22     R *= -1.5;
23     cout << "Scaled (!) matrix M = " << endl << M << endl;
24     // Matrix S is not affected, because of deep copy
25     cout << "Matrix S = " << endl << S << endl;
26    }
27 }
```

This function has to be called with a mutable (l-value) matrix type object. A sample output is printed next:

```
1  Matrix M =
2    0 -1 -2 -3 -4 -5 -6
3    1  0 -1 -2 -3 -4 -5
4    2  1  0 -1 -2 -3 -4
5    3  2  1  0 -1 -2 -3
6    4  3  2  1  0 -1 -2
7    5  4  3  2  1  0 -1
8  reshaped to 2x21 =
9    0  2  4 -1  1  3 -2  0  2 -3 -1  1 -4 -2  0 -5 -3 -1 -6 -4 -2
10   1  3  5  0  2  4 -1  1  3 -2  0  2 -3 -1  1 -4 -2  0 -5 -3 -1
11 Scaled (!) matrix M =
12   -0  1.5    3  4.5    6  7.5    9
```

```
 −1.5   −0  1.5   3  4.5   6  7.5
   −3 −1.5   −0 1.5   3 4.5   6
 −4.5   −3 −1.5  −0 1.5   3 4.5
   −6 −4.5   −3 −1.5  −0 1.5   3
 −7.5   −6 −4.5  −3 −1.5  −0 1.5
Matrix S =
 0  2  4 −1  1  3 −2  0  2 −3 −1  1 −4 −2  0 −5 −3 −1 −6 −4 −2
 1  3  5  0  2  4 −1  1  3 −2  0  2 −3 −1  1 −4 −2  0 −5 −3 −1
```

Remark 1.2.3.8 (NUMPY function `reshape`) NUMPY offers the function `np.reshape` for changing the dimensions of a matrix $\mathbf{A} \in \mathbb{K}^{m,n}$:

```
# read elements of A in row major order (default)
B = np.reshape(A, (k, l))   # error, in case kl ≠ mn
B = np.reshape(A, (k, l), order='C')   # same as above
# read elements of A in column major order
B = np.reshape(A, (k, l), order='F')
# read elements of A as stored in memory
B = np.reshape(A, (k, l), order='A')
```

This command will create an $k \times l$-array by reinterpreting the array of entries of `A` as data for an array with `k` rows and `l` columns. The order in which the elements of `A` are be read can be set by the `order` argument to row major (default, `'C'`), column major (`'F'`) or `A`'s internal storage order, i.e. row major if `A` is row major or column major if `A` is column major (`'A'`).

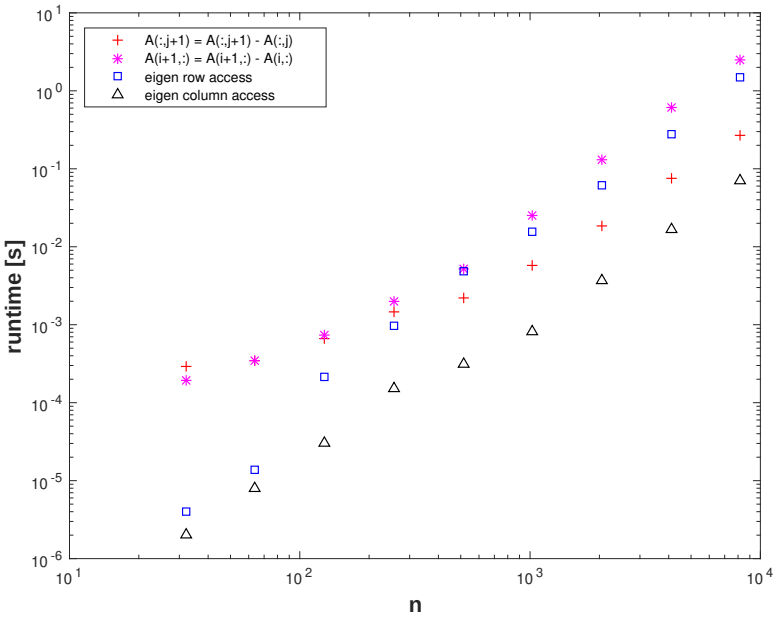
EXPERIMENT 1.2.3.9 (Impact of matrix data access patterns on runtime) Modern CPU feature several levels of memories (registers, L1 cache, L2 cache, ..., main memory) of different latency, bandwidth, and size. Frequently accessing memory locations with widely different addresses results in many cache misses and will considerably slow down the CPU.

The following C++ code sequentially runs through the entries of a *column major* matrix (EIGEN's default) in two ways and measures the (average) time required for the loops to complete. It relies on the **std::chrono** library C++ reference.

C++ code 1.2.3.10: Timing for row and column oriented matrix access for EIGEN ⇝ GITLAB

```
void rowcolaccesstiming()
{
  constexpr size_t K = 3; // Number of repetitions
  constexpr index_t N_min = 5; // Smallest matrix size 32
  constexpr index_t N_max = 13; // Scan until matrix size of 8192
  index_t n = (1UL << static_cast<size_t>(N_min));
  Eigen::MatrixXd times(N_max-N_min+1,3);

  for(index_t l=N_min; l<= N_max; l++, n*=2) {
    Eigen::MatrixXd A = Eigen::MatrixXd::Random(n,n);
    value_t t1 = 1000.0;
    for(size_t k=0;k<K;k++) {
      auto tic =  high_resolution_clock::now();
      for(index_t j=0; j < n-1; j++) {
        A.row(j+1) -= A.row(j); // row access
      }
      auto toc = high_resolution_clock::now();
      const value_t t =
```

```
            static_cast<value_t>(duration_cast<microseconds>(toc-tic).count())/1E6;
      t1 = std::min(t1,t);
    }
    value_t t2 = 1000.0;
    for(size_t k=0;k<K;k++) {
      auto tic = high_resolution_clock::now();
      for(index_t j=0; j < n-1; j++) {
        A.col(j+1) -= A.col(j); //column access
      }
      auto toc = high_resolution_clock::now();
      const value_t t =
          static_cast<value_t>(duration_cast<microseconds>(toc-tic).count())/1E6;
      t2 = std::min(t2,t);
    }
    times(l-N_min,0) = static_cast<value_t>(n);
    times(l-N_min,1) = t1;
    times(l-N_min,2) = t2;
  }
  std::cout << times << std::endl;
}
```



◁ Plot of average runtimes as measured with code Code 1.2.3.10.

Platform:

- ubuntu 14.04 LTS
- i7-3517U CPU @ 1.90GHz
- L1 32 KB, L2 256 KB, L3 4096 KB, Mem 8 GB
- gcc 4.8.4, -O3, -DNDEBUG

The compiler flags `-O3` and `-DNDEBUG` are essential. The C++ code would be significantly slower if the default compiler options were used!

We observe a blatant discrepancy of CPU time required for accessing entries of a matrix in rowwise or columnwise fashion. This reflects the impact of features of the unterlying hardware architecture, like cache size and memory bandwidth:

Interpretation of timings: Since standard matrices in EIGEN are stored column major all the matrix elements in a column occupy contiguous memory locations, which will all reside in the cache together. Hence, column oriented access will mainly operate on data in the cache even for large matrices. Conversely, row oriented access addresses matrix entries that are stored in distant memory locations, which incurs frequent cash misses (cache thrashing).

The impact of hardware architecture on the performance of algorithms will **not** be taken into account in this course, because hardware features tend to be both intricate and ephemeral. However, for modern high performance computing it is essential to adapt implementations to the hardware on which the code is supposed to run. ⌟

Review question(s) 1.2.3.11 (Dense matrix storage formats)

(Q1.2.3.11.A) Write *efficient* elementary C++ loops that realize the matrix×vector product $\mathbf{Mx}$, $\mathbf{m} \in \mathbb{R}^{m,n}$, $\mathbf{x} \in \mathbb{R}^n$, where $\mathbf{M}$ is stored in an **Eigen::MatrixXd** object `M` and $\mathbf{x}$ given as a **Eigen::VectorXd** object `x`. Assume the default (Column major) memory layout for `M`. Discuss the memory access pattern.

(Q1.2.3.11.B) A black-box function has the following signature:

```
template <typename Vector>
double processVector(const Eigen::DenseBase<Vector> &v);
```

It is known that it accesses each vector entry only once.

Recall that the vectorisation of a matrix $\mathbf{A} \in \mathbb{K}^{n,m}$is defined as

$$\mathrm{vec} : \mathbb{K}^{n,m} \to \mathbb{K}^{n\cdot m} \quad , \quad \mathrm{vec}(\mathbf{A}) := \begin{bmatrix} (\mathbf{A})_{:,1} \\ (\mathbf{A})_{:,2} \\ \vdots \\ (\mathbf{A})_{:,m} \end{bmatrix} \in \mathbb{R}^{n\cdot m} \,. \tag{1.2.3.5}$$

Given a matrix $\mathbf{A} \in \mathbb{R}^{n,m}$ stored in an **Eigen::MatrixXd** object `A` (*column major* memory layout), how can you efficiently realize the following function calls in C++:

- `processVector`$(\mathrm{vec}(\mathbf{A}))$,
- `processVector`$(\mathrm{vec}(\mathbf{A}^\top))$?

△

1.3 Basic Linear Algebra Operations

First we refresh the basic rules of vector and matrix calculus. Then we will learn about a very old programming interface for simple dense linear algebra operations.

1.3.1 Elementary Matrix-Vector Calculus

What you should know from linear algebra [NS02, Sect. 2.2]:

- ✦ vector space operations in matrix space $\mathbb{K}^{m,n}$ (addition, multiplication with scalars)
- ✦ dot product: $\mathbf{x},\mathbf{y} \in \mathbb{K}^n$, $n \in \mathbb{N}$: $\mathbf{x}\cdot\mathbf{y} := \mathbf{x}^{\mathrm{H}}\mathbf{y} = \sum_{i=1}^{n} \bar{x}_i y_i \in \mathbb{K}$
  (in EIGEN: `x.dot(y)` or `x.adjoint()*y`, `x`,`y` ≙ column vectors)
- ✦ tensor product: $\mathbf{x} \in \mathbb{K}^m, \mathbf{y} \in \mathbb{K}^n$, $n \in \mathbb{N}$: $\mathbf{x}\mathbf{y}^{\mathrm{H}} = \left(x_i \bar{y}_j\right)_{\substack{i=1,\ldots,m \\ j=1,\ldots,n}} \in \mathbb{K}^{m,n}$
  (in EIGEN: `x*y.adjoint()`, `x`,`y` ≙ column vectors)
- ✦ All are special cases of the matrix product:

$$\mathbf{A} \in \mathbb{K}^{m,n} \,, \quad \mathbf{B} \in \mathbb{K}^{n,k} : \qquad \mathbf{AB} = \left[\sum_{j=1}^{n} a_{ij} b_{jl}\right]_{\substack{i=1,\ldots,m \\ l=1,\ldots,k}} \in \mathbb{K}^{m,k} \,. \tag{1.3.1.1}$$

Recall from linear algebra basic properties of the matrix product: for all $\mathbb{K}$-matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}$ (of suitable sizes), $\alpha, \beta \in \mathbb{K}$

$$\begin{aligned} \text{associative:} &\quad (\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC}) \,, \\ \text{bi-linear:} &\quad (\alpha\mathbf{A} + \beta\mathbf{B})\mathbf{C} = \alpha(\mathbf{AC}) + \beta(\mathbf{BC}) \,, \quad \mathbf{C}(\alpha\mathbf{A} + \beta\mathbf{B}) = \alpha(\mathbf{CA}) + \beta(\mathbf{CB}) \,, \\ \text{non-commutative:} &\quad \mathbf{AB} \neq \mathbf{BA} \quad \text{in general} \,. \end{aligned}$$

§1.3.1.2 (Visualisation of (special) matrix products) Dependency of an entry of a product matrix:

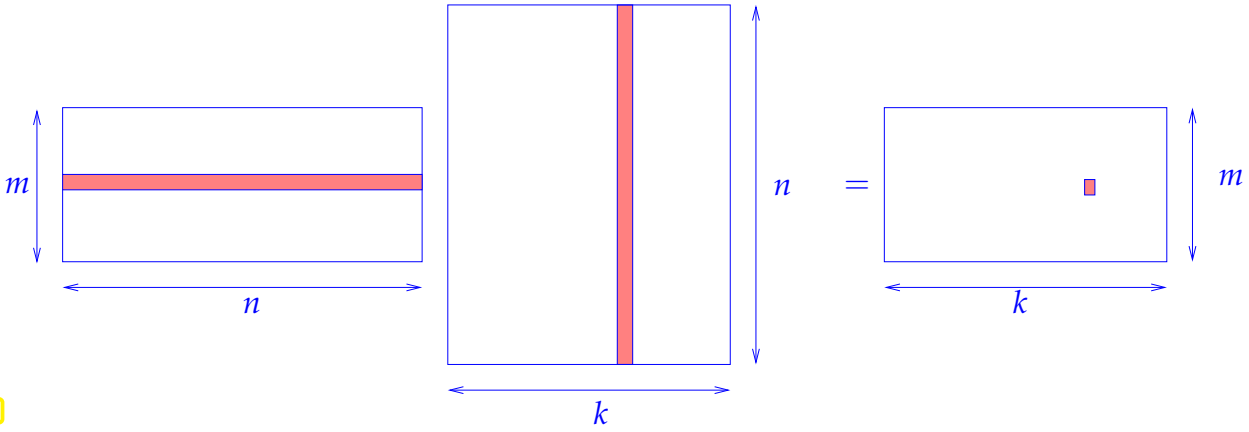
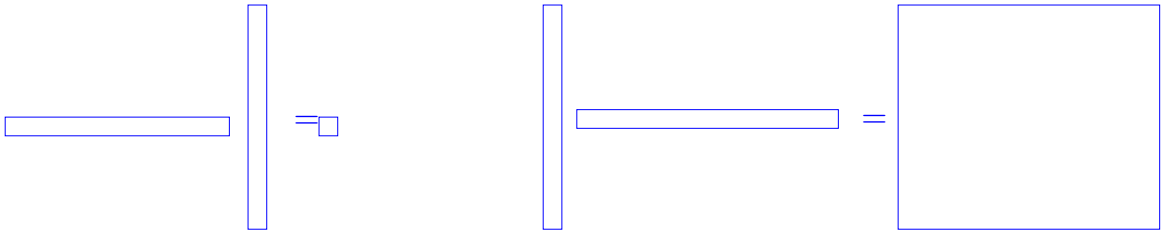


Fig. 10



dot product tensor product

⌟

Remark 1.3.1.3 (Row-wise & column-wise view of matrix product) To understand what is going on when forming a matrix product, it is often useful to decompose it into matrix×vector operations in one of the following two ways:

$\mathbf{A} \in \mathbb{K}^{m,n}$, $\mathbf{B} \in \mathbb{K}^{n,k}$:

$$\mathbf{AB} = \begin{bmatrix} \mathbf{A}(\mathbf{B})_{:,1} & \dots & \mathbf{A}(\mathbf{B})_{:,k} \end{bmatrix} \quad , \quad \mathbf{AB} = \begin{bmatrix} (\mathbf{A})_{1,:}\mathbf{B} \\ \vdots \\ (\mathbf{A})_{m,:}\mathbf{B} \end{bmatrix} . \tag{1.3.1.4}$$

matrix assembled from columns matrix assembled from rows

For notations refer to Sect. 1.1.1. ⌟

Remark 1.3.1.5 (Understanding the structure of product matrices) A "mental image" of matrix multiplication is useful for telling special properties of product matrices.

For instance, zero blocks of the product matrix can be predicted easily in the following situations using the idea explained in Rem. 1.3.1.3 (try to understand how):

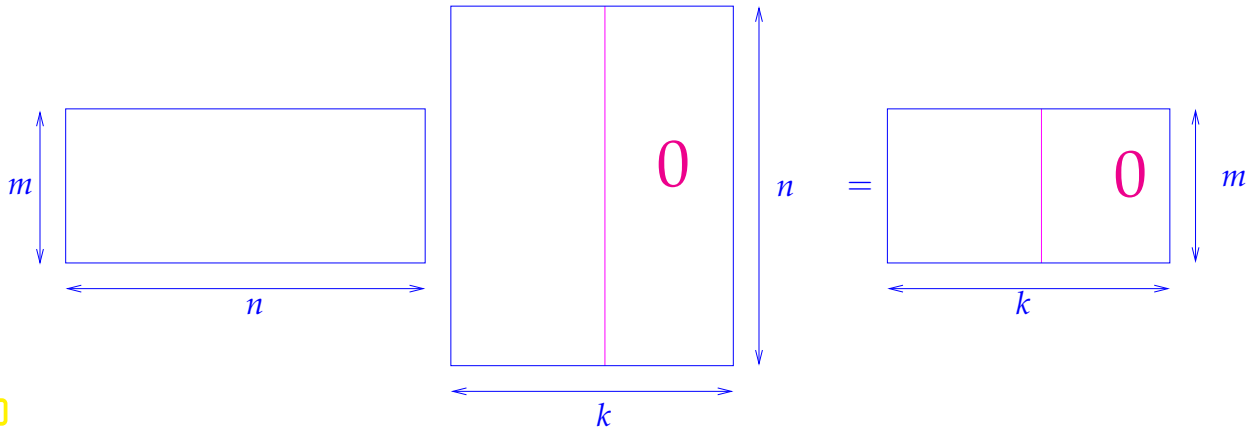


Fig. 11

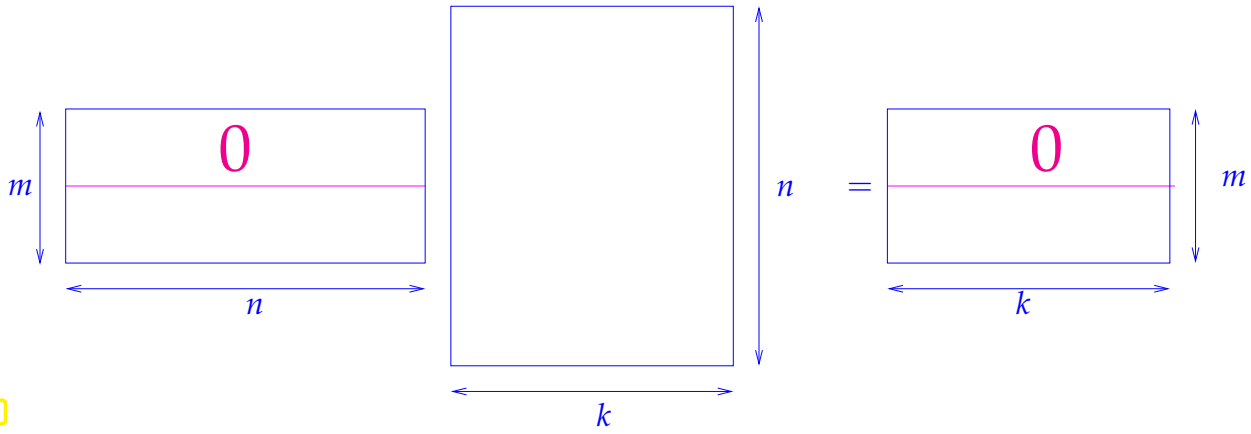


Fig. 12

A clear understanding of matrix multiplication enables you to "see", which parts of a matrix factor matter in a product:

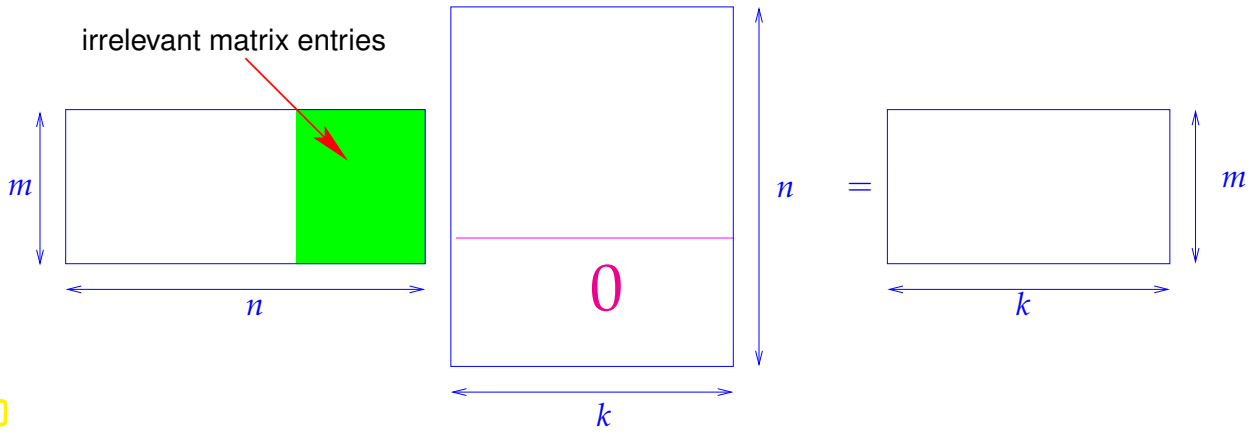
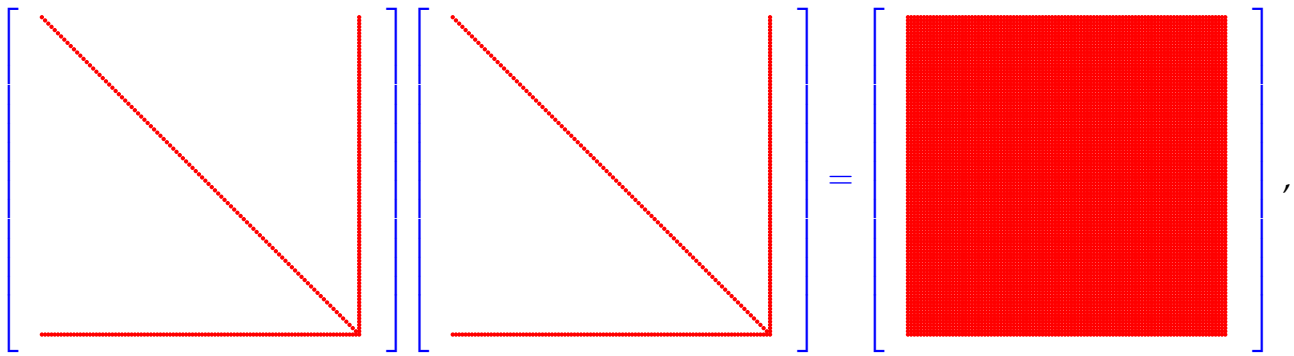
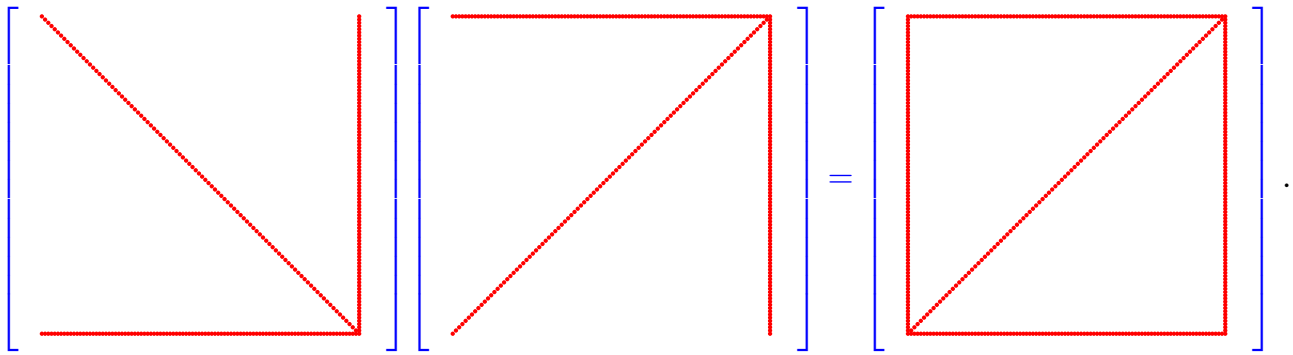


Fig. 13

"Seeing" the structure/pattern of a matrix product:



,



These nice renderings of the so-called patterns of matrices, that is, the distribution of their non-zero entries have been created by a special plotting command `spy()` of matplotlibcpp.

C++ code 1.3.1.6: Visualizing the structure of matrices in EIGEN ⇝ GITLAB

```
#include "matplotlibcpp.h" // Tools for plotting, see
    https://github.com/lava/matplotlib-cpp
#include <Eigen/Dense>
#include <string>
namespace plt = matplotlibcpp;
using namespace Eigen;

// Produce spy-plot of a dense Eigen matrix.
void spy(const Eigen::MatrixXd &M, const std::string &fname) {
  plt::figure();
  plt::spy(M, {{"marker", "o"}, {"markersize", "2"}, {"color", "b"}});
  plt::title("nnz = " + std::to_string(M.nonZeros()));
  plt::savefig(fname);
}

int main() {
  int n = 100;
  MatrixXd A(n, n), B(n, n);
  A.setZero();
  B.setZero();
  // Initialize matrices, see Fig. 13
  A.diagonal() = VectorXd::LinSpaced(n, 1, n);
  A.col(n - 1) = VectorXd::LinSpaced(n, 1, n);
  A.row(n - 1) = RowVectorXd::LinSpaced(n, 1, n);
  B = A.colwise().reverse();
  // Matrix products
  MatrixXd C = A * A, D = A * B;
  spy(A, "Aspy_cpp.eps"); // Sparse arrow matrix
  spy(B, "Bspy_cpp.eps"); // Sparse arrow matrix
  spy(C, "Cspy_cpp.eps"); // Fully populated matrix
  spy(D, "Dspy_cpp.eps"); // Sparse "framed" matrix
  return 0;
}
```

This code also demonstrates the use of `diagonal()`, `col()`, `row()` for L-value access to parts of a matrix.

PYTHON/MATPLOTLIB-command for visualizing the structure of a matrix: `plt.spy(M)`

PYTHON-code 1.3.1.7: Visualizing the structure of matrices in PYTHON

```
n = 100
```

```
A = np.diag(np.mgrid[:n])
A[:, -1] = A[-1, :] = np.mgrid[:n]
plt.spy(A)
plt.spy(A[::-1, :])
plt.spy(np.dot(A, A))
plt.spy(np.dot(A, B))
```

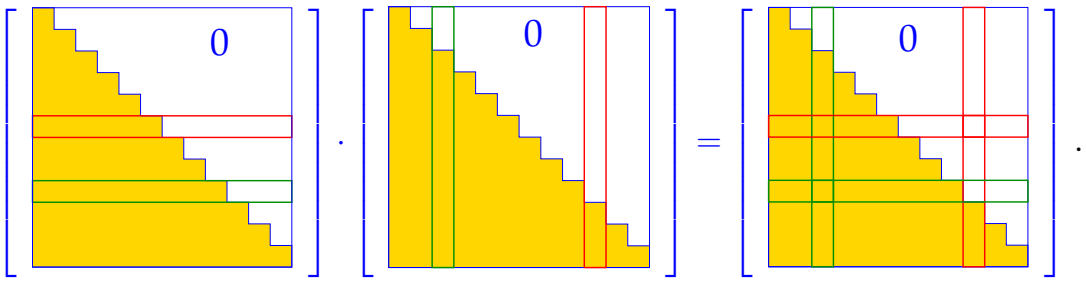
Remark 1.3.1.8 (Multiplying triangular matrices) The following result is useful when dealing with matrix decompositions that often involve triangular matrices.

Lemma 1.3.1.9. Group of regular diagonal/triangular matrices

$$\mathbf{A},\mathbf{B} \begin{Bmatrix} \text{diagonal} \\ \text{upper triangular} \\ \text{lower triangular} \end{Bmatrix} \Rightarrow \mathbf{AB} \text{ and } \mathbf{A}^{-1} \begin{Bmatrix} \text{diagonal} \\ \text{upper triangular} \\ \text{lower triangular} \end{Bmatrix}.$$

(assumes that $\mathbf{A}$ is regular)

"Proof by visualization" → Rem. 1.3.1.5



EXPERIMENT 1.3.1.10 (Scaling a matrix) Scaling = multiplication with diagonal matrices (with non-zero diagonal entries):

It is important to know the different effect of multiplying with a diagonal matrix from left or right:

- multiplication with diagonal matrix *from left* ➤ row scaling

$$\begin{bmatrix} d_1 & 0 & & 0 \\ 0 & d_2 & & 0 \\ & & \ddots & \\ 0 & 0 & & d_n \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & & a_{2m} \\ \vdots & & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nm} \end{bmatrix} = \begin{bmatrix} d_1a_{11} & d_1a_{12} & \dots & d_1a_{1m} \\ d_2a_{21} & d_2a_{22} & \dots & d_2a_{2m} \\ \vdots & & & \vdots \\ d_na_{n1} & d_na_{n2} & \dots & d_na_{nm} \end{bmatrix} = \begin{bmatrix} d_1(\mathbf{A})_{1,:} \\ \vdots \\ d_n(\mathbf{A})_{n,:} \end{bmatrix}.$$

- multiplication with diagonal matrix *from right* ➤ column scaling

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & & a_{2m} \\ \vdots & & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nm} \end{bmatrix} \begin{bmatrix} d_1 & 0 & & 0 \\ 0 & d_2 & & 0 \\ & & \ddots & \\ 0 & 0 & & d_m \end{bmatrix} = \begin{bmatrix} d_1a_{11} & d_2a_{12} & \dots & d_ma_{1m} \\ d_1a_{21} & d_2a_{22} & \dots & d_ma_{2m} \\ \vdots & & & \vdots \\ d_1a_{n1} & d_2a_{n2} & \dots & d_ma_{nm} \end{bmatrix}$$

$$= \begin{bmatrix} d_1(\mathbf{A})_{:,1} & \dots & d_m(\mathbf{A})_{:,m} \end{bmatrix}.$$

Multiplication with a scaling matrix $\mathbf{D} = \operatorname{diag}(d_1,\ldots,d_n) \in \mathbb{R}^{n,n}$ in EIGEN can be realised in three ways, see Code 1.3.1.11, Line 9-Line 11, Line 13, and Line 15.

Measured runtimes (Ubuntu Linux 14.04 LTS, Intel Core(TM) i7-3517U CPU @ 1.90GHz × 4, 64-bit, gcc 4.8.4, -O3 -DNDEBUG) ▷

The code will be slowed down massively in case a temporary dense matrix is created inadvertently. Notice that EIGEN's expression templates avoid this pointless effort, see Line 15.

Timings for different ways to do scaling

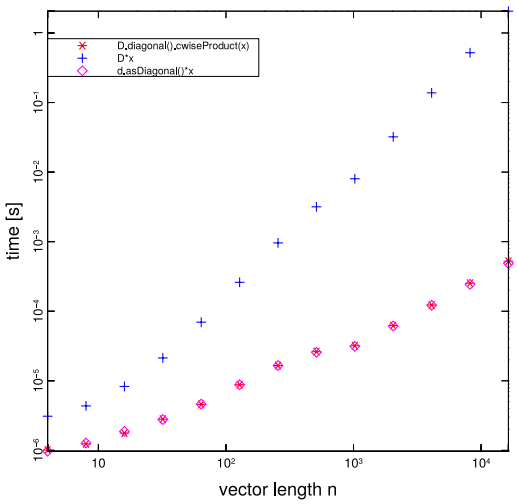


Fig. 14

C++ code 1.3.1.11: Timing multiplication with scaling matrix in EIGEN ↪ GITLAB

```
2  int nruns = 3, minExp = 2, maxExp = 14;
3  MatrixXd tms(maxExp-minExp+1,4);
4  for(int i = 0; i <= maxExp-minExp; ++i){
5    Timer tbad, tgood, topt;  // timer class
6    int n = std::pow(2, minExp + i);
7    VectorXd d = VectorXd::Random(n,1), x = VectorXd::Random(n,1), y(n);
8    for(int j = 0; j < nruns; ++j) {
9      MatrixXd D = d.asDiagonal(); //
10     // matrix vector multiplication
11     tbad.start();  y = D*x; tbad.stop(); //
12     // componentwise multiplication
13     tgood.start(); y= d.cwiseProduct(x); tgood.stop(); //
14     // matrix multiplication optimized by Eigen
15     topt.start();  y = d.asDiagonal()*x; topt.stop(); //
16   }
17   tms(i,0)=n;
18   tms(i,1)=tgood.min(); tms(i,2)=tbad.min(); tms(i,3)= topt.min();
19 }
```

Hardly surprising, the component-wise multiplication of the two vectors is way faster than the intermittent initialisation of a diagonal matrix (main populated by zeros) and the computation of a matrix×vector product. Nevertheless, such blunders keep on haunting numerical codes. Do not rely solely on EIGEN optimizations! ⌟

Remark 1.3.1.12 (Row and column transformations) Simple operations on rows/columns of matrices, *cf.* what was done in Exp. 1.2.3.9, can often be expressed as multiplication with special matrices: For instance, given $\mathbf{A} \in \mathbb{K}^{n,m}$ we obtain $\mathbf{B}$ by adding row $(\mathbf{A})_{j,:}$ to row $(\mathbf{A})_{j+1,:}$, $1 \le j < n$.

Realisation through matrix product ◣

$$\mathbf{B} = \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & 1 & 1 & \\ & & & & \ddots \\ & & & & & 1 \end{bmatrix} \mathbf{A} .$$

The matrix multiplying $\mathbf{A}$ from the left is a specimen of a transformation matrix, a matrix that coincides with the identity matrix $\mathbf{I}$ except for a single off-diagonal entry.

left-multiplication right-multiplication	with transformation matrices	➙ row transformations column transformations

Row/column transformations will play a central role in Section 2.3. ⌟

§1.3.1.13 (Block matrix product) Given matrix dimensions $M, N, K \in \mathbb{N}$ block sizes $1 \leq n < N$ ($n' := N - n$), $1 \leq m < M$ ($m' := M - m$), $1 \leq k < K$ ($k' := K - k$) we start from the following matrices:

$$\begin{matrix} \mathbf{A}_{11} \in \mathbb{K}^{m,n} & \mathbf{A}_{12} \in \mathbb{K}^{m,n'} \\ \mathbf{A}_{21} \in \mathbb{K}^{m',n} & \mathbf{A}_{22} \in \mathbb{K}^{m',n'} \end{matrix} \quad , \quad \begin{matrix} \mathbf{B}_{11} \in \mathbb{K}^{n,k} & \mathbf{B}_{12} \in \mathbb{K}^{n,k'} \\ \mathbf{B}_{21} \in \mathbb{K}^{n',k} & \mathbf{B}_{22} \in \mathbb{K}^{n',k'} \end{matrix} .$$

This matrices serve as sub-matrices or matrix blocks and are assembled into larger matrices

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \in \mathbb{K}^{M,N} \quad , \quad \mathbf{B} = \begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{bmatrix} \in \mathbb{K}^{N,K} .$$

It turns out that the matrix product $\mathbf{AB}$ can be computed by the same formula as the product of simple 2×2-matrices:

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{11}\mathbf{B}_{11} + \mathbf{A}_{12}\mathbf{B}_{21} & \mathbf{A}_{11}\mathbf{B}_{12} + \mathbf{A}_{12}\mathbf{B}_{22} \\ \mathbf{A}_{21}\mathbf{B}_{11} + \mathbf{A}_{22}\mathbf{B}_{21} & \mathbf{A}_{21}\mathbf{B}_{12} + \mathbf{A}_{22}\mathbf{B}_{22} \end{bmatrix} . \tag{1.3.1.14}$$

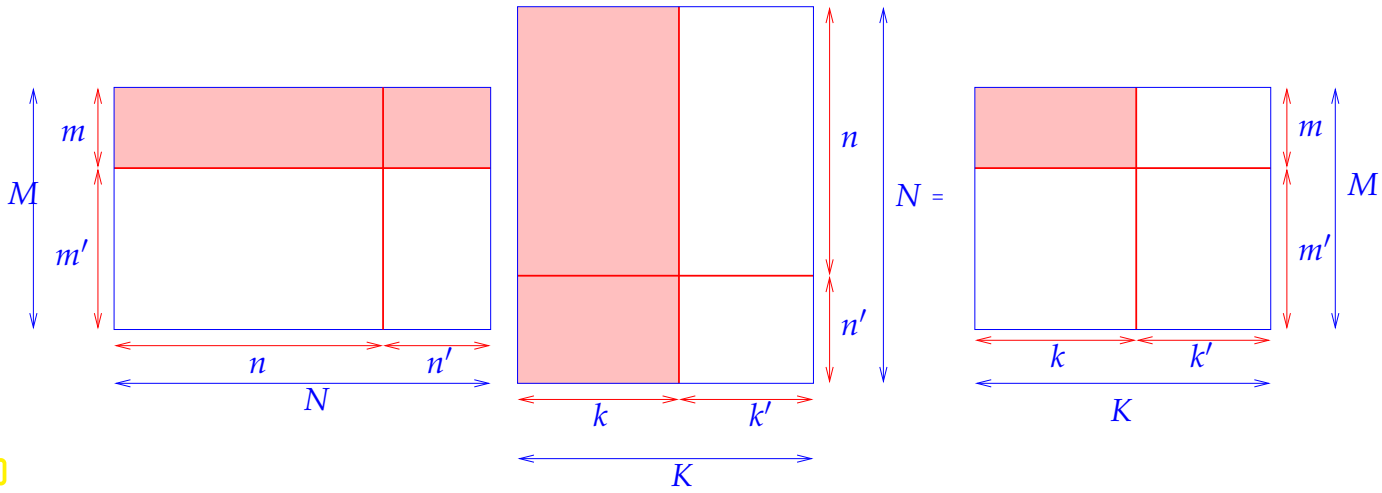


Fig. 15

Bottom line: one can compute with block-structured matrices in *almost* (∗) the same ways as with matrices with real/complex entries, see [QSS00, Sect. 1.3.3].

(∗): you must not use the commutativity of multiplication (because matrix multiplication is not commutative).

⌟

1.3.2 BLAS – Basic Linear Algebra Subprograms

BLAS (Basic Linear Algebra Subprograms) is a specification (API) that prescribes a set of low-level routines for performing common linear algebra operations such as vector addition, scalar multiplication, dot products, linear combinations, and matrix multiplication. They are the de facto low-level routines for linear algebra libraries (Wikipedia).

The BLAS API is standardised by the BLAS technical forum and, due to its history dating back to the 70s, follows conventions of FORTRAN 77, see the Quick Reference Guide for examples. However, wrappers for

other programming languages are available. CPU manufacturers and/or developers of operating systems usually supply highly optimised implementations:

- OpenBLAS: open source implementation with some general optimisations, available under BSD license.
- ATLAS (Automatically Tuned Linear Algebra Software): open source BLAS implementation with auto-tuning capabilities. Comes with C and FORTRAN interfaces and is included in Linux distributions.
- Intel MKL (Math Kernel Library): commercial highly optimised BLAS implemetation available for all Intel CPUs. Used by most proprietory simulation software and also MATLAB.

EXPERIMENT 1.3.2.1 (Multiplying matrices in EIGEN)

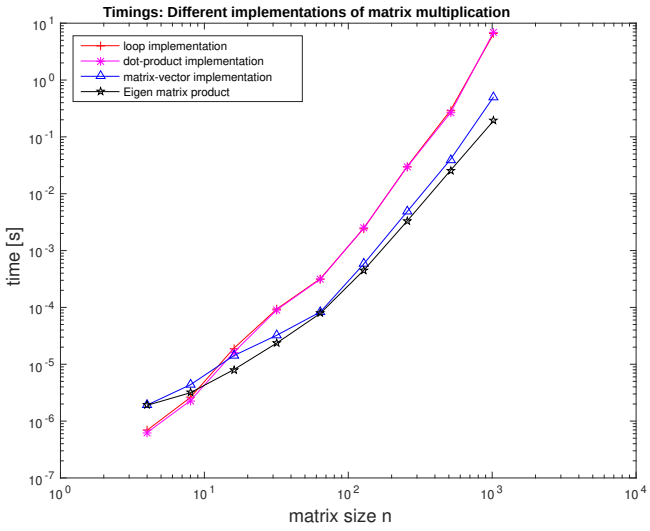
The following EIGEN-based C++ code performs a multiplication of densely populated matrices in three different ways:

1. Direct implementation of three nested loops
2. Realization by matrix×vector products
3. Use of buit-in matrix multiplication of EIGEN

C++ code 1.3.2.2: Timing different implementations of matrix multiplication in EIGEN
↪ GITLAB

```
2  void mmtiming() {
3    int nruns = 3, minExp = 2, maxExp = 10;
4    MatrixXd timings(maxExp - minExp + 1, 5);
5    for (int p = 0; p <= maxExp - minExp; ++p) {
6      Timer t1, t2, t3, t4; // timer class
7      int n = std::pow(2, minExp + p);
8      MatrixXd A = MatrixXd::Random(n, n);
9      MatrixXd B = MatrixXd::Random(n, n);
10     MatrixXd C = MatrixXd::Zero(n, n);
11     for (int q = 0; q < nruns; ++q) {
12       // Loop based implementation no template magic
13       t1.start();
14       for (int i = 0; i < n; ++i)
15         for (int j = 0; j < n; ++j)
16           for (int k = 0; k < n; ++k)
17             C(i, j) += A(i, k) * B(k, j);
18       t1.stop();
19       // dot product based implementation little template magic
20       t2.start();
21       for (int i = 0; i < n; ++i)
22         for (int j = 0; j < n; ++j)
23           C(i, j) = A.row(i).dot(B.col(j));
24       t2.stop();
25       // matrix-vector based implementation middle template magic
26       t3.start();
27       for (int j = 0; j < n; ++j)
28         C.col(j) = A * B.col(j);
29       t3.stop();
30       // Eigen matrix multiplication template magic optimized
31       t4.start();
32       C = A * B;
33       t4.stop();
34     }
35     timings(p, 0) = n;
```

```
36      timings(p, 1) = t1.min();
37      timings(p, 2) = t2.min();
38      timings(p, 3) = t3.min();
39      timings(p, 4) = t4.min();
40    }
41    std::cout << std::scientific << std::setprecision(3) << timings << std::endl;
```



Platform:

- ubuntu 14.04 LTS
- i7-3517U CPU @ 1.90GHz
- L1 32 KB, L2 256 KB, L3 4096 KB, Mem 8 GB
- gcc 4.8.4, -O3

In EIGEN we can achieve some gain in execution speed by relying on compact matrix/vector operations that invoke efficient EIGEN built-in functions. However, compiler optimizations make plain and simple loops almost competitive.

BLAS routines are grouped into "levels" according to the amount of data and computation involved (asymptotic complexity, see Section 1.4.1 and [GV89, Sect. 1.1.12]):

- **Level 1**: vector operations such as scalar products and vector norms.
  asymptotic complexity $O(n)$, (with $n \hat{=}$ vector length),
  e.g.: dot product: $\rho = \mathbf{x}^\top \mathbf{y}$
- **Level 2**: vector-matrix operations such as matrix-vector multiplications.
  asymptotic complexity $O(mn)$,(with $(m,n) \hat{=}$ matrix size),
  e.g.: matrix×vector multiplication: $\mathbf{y} = \alpha \mathbf{A}\mathbf{x} + \beta \mathbf{y}$
- **Level 3**: matrix-matrix operations such as matrix additions or multiplications.
  asymptotic complexity often $O(nmk)$,(with $(n,m,k) \hat{=}$ matrix sizes),
  e.g.: matrix product: $\mathbf{C} = \mathbf{A}\mathbf{B}$

Syntax of BLAS calls:
The functions have been implemented for different types, and are distinguished by the first letter of the function name. E.g. *sdot* is the dot product implementation for single precision and *ddot* for double precision.

- **BLAS LEVEL 1**: vector operations, asymptotic complexity $O(n)$, $n \hat{=}$ vector length
  - dot product $\rho = \mathbf{x}^\top \mathbf{y}$

$$\texttt{xDOT(N,X,INCX,Y,INCY)}$$

    - x $\in \{\mathrm{S}, \mathrm{D}\}$, scalar type: S $\hat{=}$ type `float`, D $\hat{=}$ type `double`
    - $N \hat{=}$ length of vector (modulo stride INCX)
    - $X \hat{=}$ vector $\mathbf{x}$: array of type x
    - `INCX` $\hat{=}$ stride for traversing vector X

- $Y \triangleq$ vector $\mathbf{y}$: array of type x
- INCY $\triangleq$ stride for traversing vector Y

- vector operations $\mathbf{y} = \alpha\mathbf{x} + \mathbf{y}$

```
xAXPY(N,ALPHA,X,INCX,Y,INCY)
```

- x $\in \{S, D, C, Z\}$, S $\triangleq$ type `float`, D $\triangleq$ type `double`, C $\triangleq$ type `complex`
- $N \triangleq$ length of vector (modulo stride INCX)
- ALPHA $\triangleq$ scalar α
- $X \triangleq$ vector $\mathbf{x}$: array of type x
- INCX $\triangleq$ stride for traversing vector X
- $Y \triangleq$ vector $\mathbf{y}$: array of type x
- INCY $\triangleq$ stride for traversing vector Y

✦ **BLAS LEVEL 2**: matrix-vector operations, asymptotic complexity $O(mn)$, $(m,n) \triangleq$ matrix size

- matrix×vector multiplication $\mathbf{y} = \alpha\mathbf{A}\mathbf{x} + \beta\mathbf{y}$

```
xGEMV(TRANS,M,N,ALPHA,A,LDA,X,
      INCX,BETA,Y,INCY)
```

- x $\in \{S, D, C, Z\}$, scalar type: S $\triangleq$ type `float`, D $\triangleq$ type `double`, C $\triangleq$ type `complex`
- $M, N \triangleq$ size of matrix $\mathbf{A}$
- ALPHA $\triangleq$ scalar parameter α
- A $\triangleq$ matrix $\mathbf{A}$ stored in *linear array* of length $M \cdot N$ (column major arrangement)

$$(\mathbf{A})_{i,j} = \mathrm{A}[N * (j-1) + i] \,.$$

- LDA $\triangleq$ "leading dimension" of $\mathbf{A} \in \mathbb{K}^{n,m}$, that is, the number n of rows.
- $X \triangleq$ vector $\mathbf{x}$: array of type x
- INCX $\triangleq$ stride for traversing vector X
- BETA $\triangleq$ scalar paramter β
- $Y \triangleq$ vector $\mathbf{y}$: array of type x
- INCY $\triangleq$ stride for traversing vector Y

- **BLAS LEVEL 3**: matrix-matrix operations, asymptotic complexity $O(mnk)$, $(m,n,k) \triangleq$ matrix sizes

  - matrix×matrix multiplication $\mathbf{C} = \alpha\mathbf{A}\mathbf{B} + \beta\mathbf{C}$

```
xGEMM(TRANSA,TRANSB,M,N,K,
```

```
ALPHA,A,LDA,X,B,LDB,
BETA,C,LDC)
```

(☞ meaning of arguments as above)

Remark 1.3.2.3 (BLAS calling conventions) The BLAS calling syntax seems queer in light of modern object oriented programming paradigms, but it is a legacy of FORTRAN77, which was (and partly still is) the programming language, in which the BLAS routines were coded.

It is a very common situation in scientific computing that one has to rely on old codes and libraries implemented in an old-fashioned style. ⌟

EXAMPLE 1.3.2.4 (Calling BLAS routines from C/C++) When calling BLAS library functions from C, all arguments have to be passed by reference (as pointers), in order to comply with the argument passing mechanism of FORTRAN77, which is the model followed by BLAS.

C++-code 1.3.2.5: BLAS-based SAXPY operation in C++

```
#define daxpy_ cblas_daxpy
#include <iostream>
#include <vector>

// Definition of the required BLAS function.  This is usually done
// in a header file like blas.h that is included in the EIGEN3
// distribution
extern "C" {
int daxpy_(const int* n, const double* da, const double* dx, const int* incx,
           double* dy, const int* incy);
}

using std::cout;
using std::endl;

int main() {
  cout << "Demo code for NumCSE course: call basic BLAS routines from C++"
       << endl;
  const int n = 5;     // length of vector
  const int incx = 1;  // stride
  const int incy = 1;  // stride
  const double alpha = 2.5;  // scaling factor

  // Allocated raw arrays of doubles
  std::vector<double> x(n);
  std::vector<double> y(n);

  for (size_t i = 0; i < n; i++) {
    x[i] = 3.1415 * static_cast<double>(i);
    y[i] = 1.0 / static_cast<double>(i + 1);
  }

  cout << "x=[";
  for (size_t i = 0; i < n; i++) {
    cout << x[i] << ' ';
  }
  cout << "]" << endl;
  cout << "y=[";
  for (size_t i = 0; i < n; i++) {
    cout << y[i] << ' ';
```

```
41    }
42    cout << "]" << endl;
43
44    // Call the BLAS library function passing pointers to all arguments
45    // (Necessary when calling FORTRAN routines from C
46    daxpy_(&n, &alpha, x.data(), &incx, y.data(), &incy);
47
48    cout << "y = " << alpha << " * x + y = [";
49    for (int i = 0; i < n; i++) {
50      cout << y[i] << ' ';
51    }
52    cout << "]" << endl;
53    return (0);
54  }
```

When using EIGEN in a mode that includes an external BLAS library, all this calls are wrapped into EIGEN methods. ⌟

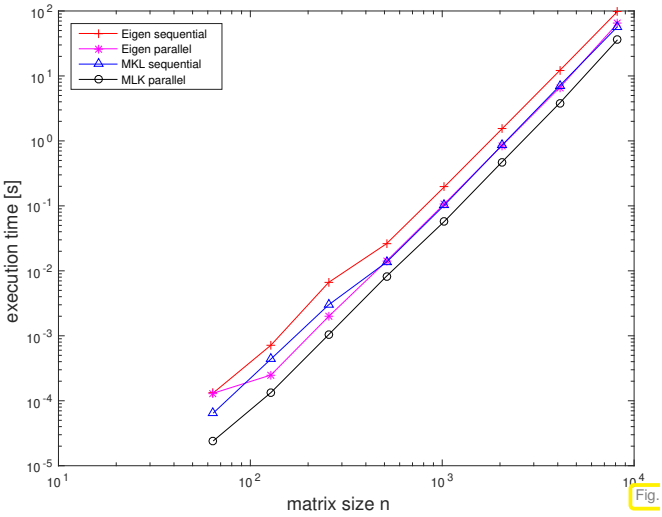
EXAMPLE 1.3.2.6 (Using Intel Math Kernel Library (Intel MKL) from EIGEN) The Intel Math Kernel Library is a highly optimized math library for Intel processors and can be called directly from EIGEN, see EIGEN documentation on "Using Intel® Math Kernel Library from Eigen".

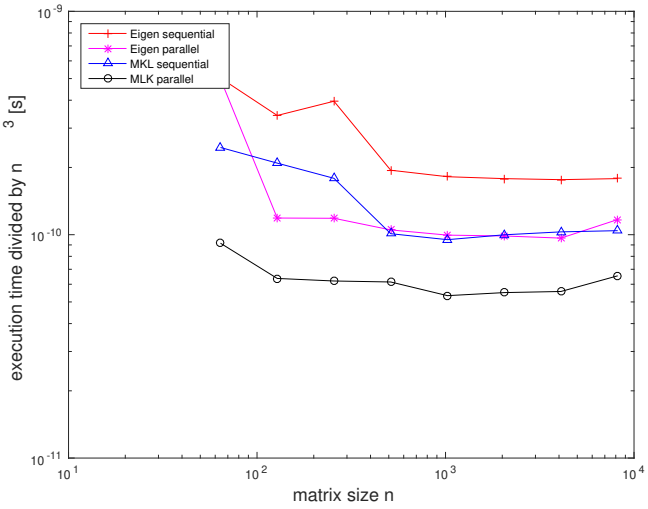
C++-code 1.3.2.7: Timing of matrix multiplication in EIGEN for MKL comparison ↪ GITLAB

```
2  //!  script for timing different implementations of matrix
       multiplications
3  void mmeigenmkl(){
4    int nruns = 3, minExp = 6, maxExp = 13;
5    MatrixXd timings(maxExp-minExp+1,2);
6    for(int p = 0; p <= maxExp-minExp; ++p){
7       Timer t1; // timer class
8       int n = std::pow(2, minExp + p);
9      MatrixXd A = MatrixXd::Random(n,n);
10     MatrixXd B = MatrixXd::Random(n,n);
11     MatrixXd C = MatrixXd::Zero(n,n);
12     for(int q = 0; q < nruns; ++q){
13           t1.start();
14           C = A * B;
15           t1.stop();
16      }
17      timings(p,0)=n; timings(p,1)=t1.min();
18   }
19   std::cout << std::scientific << std::setprecision(3) << timings << std::endl;
20 }
```

Timing results:

n	EIGEN sequential [s]	EIGEN parallel [s]	MKL sequential [s]	MKL parallel [s]
64	1.318e-04	1.304e-04	6.442e-05	2.401e-05
128	7.168e-04	2.490e-04	4.386e-04	1.336e-04
256	6.641e-03	1.987e-03	3.000e-03	1.041e-03
512	2.609e-02	1.410e-02	1.356e-02	8.243e-03
1024	1.952e-01	1.069e-01	1.020e-01	5.728e-02
2048	1.531e+00	8.477e-01	8.581e-01	4.729e-01
4096	1.212e+01	6.635e+00	7.075e+00	3.827e+00
8192	9.801e+01	6.426e+01	5.731e+01	3.598e+01





Timing environment:

- ubuntu 14.04 LTS
- i7-3517U CPU @ 1.90GHz
- L1 32 KB, L2 256 KB, L3 4096 KB, Mem 8 GB
- gcc 4.8.4, -O3

1.4 Computational Effort

Video tutorial for Section 1.4 "Computational Effort": (29 minutes) Download link, tablet notes

→ review questions 1.4.3.11

Large scale numerical computations require immense resources and execution time of numerical codes often becomes a central concern. Therefore, much emphasis has to be put on

1. designing algorithms that produce a desired result with (nearly) minimal computational effort (defined precisely below),
2. exploit possibilities for parallel and vectorised execution,
3. organising algorithms in order to make them fit memory hierarchies ("cache-aware"),
4. implementing codes that make optimal use of hardware resources and capabilities,

While Item 2–Item 4 are out of the scope of this course and will be treated in more advanced lectures, Item 1 will be a recurring theme.

The following definition encapsulates what is regarded as a measure for the "cost" of an algorithm in computational mathematics.

Definition 1.4.0.1. Computational effort

The **computational effort**/**computational cost** required/incurred by a numerical code amounts to the number of elementary operations (additions,subtractions,multiplications,divisions,square roots) executed in a run.

§1.4.0.2 (What computational effort does not tell us) Fifty years ago counting elementary operations provided good predictions of runtimes, but nowadays this is no longer true.

"Computational effort $\not\sim$ runtime"

The computational effort involved in a run of a numerical code is only loosely related to overall execution time on modern computers.

This is conspicuous in Exp. 1.2.3.9, where algorithms incurring exactly the same computational effort took different times to execute.

The reason is that on today's computers a key bottleneck for fast execution is latency and bandwidth of memory, *cf.* the discussion at the end of Exp. 1.2.3.9 and [KW03]. Thus, concepts like I/O-complexity [AV88; GJ10] might be more appropriate for gauging the efficiency of a code, because they take into account the pattern of memory access. ⌟

1.4.1 (Asymptotic) Computational Complexity

The concept of computational effort from Def. 1.4.0.1 is still useful in a particular context:

Definition 1.4.1.1. (Asymptotic) complexity

The **asymptotic (computational) complexity** of an algorithm characterises the worst-case dependence of its computational effort ($\rightarrow$ Def. 1.4.0.1) on one or more problem size parameter(s) when these tend to ∞.

- *Problem size parameters* in numerical linear algebra usually are the lengths and dimensions of the vectors and matrices that an algorithm takes as inputs.
- *Worst case* indicates that the maximum effort over a set of admissible data is taken into account.

When dealing with asymptotic complexities a mathematical formalism comes handy:

Definition 1.4.1.2. Landau symbol [AG11, p. 7]

We write $F(n) = O(G(n))$ for two functions $F, G : \mathbb{N} \to \mathbb{R}$, if there exists a constant $C > 0$ and $n_* \in \mathbb{N}$ such that

$$F(n) \leq C\, G(n) \quad \forall n \geq n_* \,.$$

More generally, $F(n_1, \ldots, n_k) = O(G(n_1, \ldots, n_k))$ for two functions $F, G : \mathbb{N}^k \to \mathbb{R}$ implies the existence of a constant $C > 0$ and a threshold value $n_* \in \mathbb{N}$ such that

$$F(n_1, \ldots, n_k) \leq C G(n_1, \ldots, n_k) \quad \forall n_1, \ldots, n_k \in \mathbb{N} \,, \quad n_\ell \geq n_*, \; \ell = 1, \ldots, k \,.$$

Remark 1.4.1.3 (Meaningful "O-bounds" for complexity) Of course, the definition of the Landau symbol leaves ample freedom for stating meaningless bounds; an algorithm that runs with linear complexity $O(n)$ can be correctly labelled as possessing $O(\exp(n))$ complexity.

Yet, whenever the Landau notation is used to describe asymptotic complexities, the bounds have to be sharp in the sense that no function with slower asymptotic growth will be possible inside the O. To make this precise we stipulate the following.

Sharpness of a complexity bound

Whenever the asymptotic complexity of an algorithm is stated as $O(n^\alpha \log^\beta n \exp(\gamma n^\delta))$ with non-negative parameters $\alpha, \beta, \gamma, \delta \geq 0$ in terms of the problem size parameter n, we take for granted that choosing a smaller value for any of the parameters will no longer yield a valid (or provable) asymptotic bound.

In particular

- complexity $O(n)$ means that the complexity is not $O(n^\alpha)$ for any $\alpha < 1$,
- complexity $O(\exp(n))$ excludes asymptotic complexity $O(n^p)$ for any $p \in \mathbb{R}$.

Terminology: If the asymptotic complexity of an algorithm is $O(n^p)$ with $p = 1, 2, 3$ we say that it is of "linear", "quadratic", and "cubic" complexity, respectively.

⌟

Remark 1.4.1.5 (Relevance of asymptotic complexity) § 8.4.3.14 warned us that computational effort and, thus, asymptotic complexity, of an algorithm for a concrete problem on a particular platform may not have much to do with the actual runtime (the blame goes to memory hierarchies, internal pipelining, vectorisation, etc.).

Then, why do we pay so much attention to asymptotic complexity in this course?

To a certain extent, the asymptotic complexity allows to predict the *dependence of the runtime* of a particular implementation of an algorithm *on the problem size* (for large problems).

For instance, an algorithm with asymptotic complexity $O(n^2)$ is likely to take $4\times$ as much time when the problem size is doubled. ⌟

§1.4.1.6 (Concluding polynomial complexity from runtime measurements)

Available: "Measured runtimes" $t_i = t_i(n_i)$ for different values $n_1, n_2, \ldots, n_N$, $n_i \in \mathbb{N}$, of the problem size parameter

Conjectured: power law dependence $t_i \approx C n_i^\alpha$ (also "algebraic dependence"), $\alpha \in \mathbb{R}$

How can we glean evidence that supports or refutes our conjecture from the data? Look at the data in doubly logarithmic scale!

$$t_i = C n_i^\alpha \quad \Rightarrow \quad \log(t_i) \approx \log C + \alpha \log(n_i) \,, \quad i = 1, \ldots, N \,.$$

If the conjecture holds true, then the points (n_i, t_i) will approximately lie on a *straight line* with slope α in a doubly logarithmic plot (which can be created in PYTHON by the `lmatplotlib.pyplot.loglog` plotting command.

➢ Offers a quick "visual test" of conjectured asymptotic complexity

More rigorous: Perform linear regression on $(\log n_i, \log t_i)$, $i = 1, \ldots, N$ (→ Chapter 3)

⌟

1.4.2 Cost of Basic Linear-Algebra Operations

Performing elementary BLAS-type operations through simple (nested) loops, we arrive at the following obvious complexity bounds:

operation	description	#mul/div	#add/sub	asymp. complexity
dot product	$(\mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^n) \mapsto \mathbf{x}^H\mathbf{y}$	n	$n-1$	$O(n)$
tensor product	$(\mathbf{x} \in \mathbb{R}^m, \mathbf{y} \in \mathbb{R}^n) \mapsto \mathbf{x}\mathbf{y}^H$	nm	0	$O(mn)$
Matrix×vector	$(\mathbf{x} \in \mathbb{R}^n, \mathbf{A} \in \mathbb{R}^{m,n}) \mapsto \mathbf{A}\mathbf{x}$	mn	$(n-1)m$	$O(mn)$
matrix product$^{(*)}$	$(\mathbf{A} \in \mathbb{R}^{m,n}, \mathbf{B} \in \mathbb{R}^{n,k}) \mapsto \mathbf{A}\mathbf{B}$	mnk	$mk(n-1)$	$O(mnk)$

EXPERIMENT 1.4.2.1 (Runtimes of elementary linear algebra operations in EIGEN)

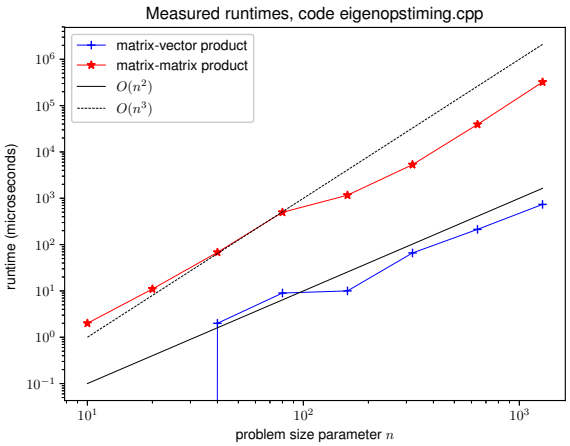


Fig. 19

- Timing code ↪ GITLAB, $m = n$
- Linux kernel 5.5.9-100.fc30.x86_64, Fedora 30
- gcc 9.3.1 -O2, Release mode
- Intel(R) Core(TM) i7-7600U CPU @ 2.80GHz

Runtime data points approximately on lines in *doubly logarithmic plot*, see § 1.4.1.6: Asymptotic behavior of measured runtimes match predictions from above table for large n.

Remark 1.4.2.2 ("Fast" matrix multiplication)

$(*)$: The $O(mnk)$ complexity bound applies to "straightforward" matrix multiplication according to (1.3.1.1).

For $m = n = k$ there are (sophisticated) variants with better asymptotic complexity, e.g., the divide-and-conquer Strassen algorithm [Str69] with asymptotic complexity $O(n^{\log_2 7})$:

Start from $\mathbf{A}, \mathbf{B} \in \mathbb{K}^{n,n}$ with $n = 2\ell$, $\ell \in \mathbb{N}$. The idea relies on the block matrix product (1.3.1.14) with $\mathbf{A}_{ij}, \mathbf{B}_{ij} \in \mathbb{K}^{\ell,\ell}$, $i,j \in \{1,2\}$. Let $\mathbf{C} := \mathbf{A}\mathbf{B}$ be partitioned accordingly: $\mathbf{C} = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix}$. Then tedious elementary computations reveal

$$\begin{aligned} \mathbf{C}_{11} &= \mathbf{Q}_0 + \mathbf{Q}_3 - \mathbf{Q}_4 + \mathbf{Q}_6 \,, \\ \mathbf{C}_{21} &= \mathbf{Q}_1 + \mathbf{Q}_3 \,, \\ \mathbf{C}_{12} &= \mathbf{Q}_2 + \mathbf{Q}_4 \,, \\ \mathbf{C}_{22} &= \mathbf{Q}_0 + \mathbf{Q}_2 - \mathbf{Q}_1 + \mathbf{Q}_5 \,, \end{aligned}$$

where the $\mathbf{Q}_k \in \mathbb{K}^{\ell,\ell}$, $k = 0,\ldots,6$ are obtained from

$$\begin{aligned} \mathbf{Q}_0 &= (\mathbf{A}_{11} + \mathbf{A}_{22}) * (\mathbf{B}_{11} + \mathbf{B}_{22}) \,, \\ \mathbf{Q}_1 &= (\mathbf{A}_{21} + \mathbf{A}_{22}) * \mathbf{B}_{11} \,, \\ \mathbf{Q}_2 &= \mathbf{A}_{11} * (\mathbf{B}_{12} - \mathbf{B}_{22}) \,, \\ \mathbf{Q}_3 &= \mathbf{A}_{22} * (-\mathbf{B}_{11} + \mathbf{B}_{21}) \,, \\ \mathbf{Q}_4 &= (\mathbf{A}_{11} + \mathbf{A}_{12}) * \mathbf{B}_{22} \,, \\ \mathbf{Q}_5 &= (-\mathbf{A}_{11} + \mathbf{A}_{21}) * (\mathbf{B}_{11} + \mathbf{B}_{12}) \,, \\ \mathbf{Q}_6 &= (\mathbf{A}_{12} - \mathbf{A}_{22}) * (\mathbf{B}_{21} + \mathbf{B}_{22}) \,. \end{aligned}$$

Beside a considerable number of matrix additions (computational effort $O(n^2)$) it takes only **7** multiplications of matrices of size $n/2$ to compute $\mathbf{C}$! Strassen's algorithm boils down to the *recursive application* of these formulas for $n = 2^k$, $k \in \mathbb{N}$. The asymptotic complexity of $O(n^{\log_2 7})$ for $n \to \infty$ the follows from the master theorem on divide-and-conquer algorithms.

A refined algorithm of this type can achieve complexity $O(n^{2.36})$, see [CW90]. ⌟

1.4.3 Improving Complexity in Numerical Linear Algebra: Some Tricks

In computations involving matrices and vectors complexity of algoritms can often be reduced by performing the operations in a particular order:

EXAMPLE 1.4.3.1 (Efficient associative matrix multiplication) We consider the multiplication with a rank-1-matrix. Matrices with rank 1 can always be obtained as the tensor product of two vectors, that is, the matrix product of a column vector and a row vector. Given $\mathbf{a} \in \mathbb{K}^m$, $\mathbf{b} \in \mathbb{K}^n$, $\mathbf{x} \in \mathbb{K}^n$ we may compute the vector $\mathbf{y} = \mathbf{a}\mathbf{b}^\top\mathbf{x}$ in two ways:

$$\mathbf{y} = \left(\mathbf{a}\mathbf{b}^\top\right)\mathbf{x}\,. \qquad (1.4.3.2) \qquad\qquad \mathbf{y} = \mathbf{a}\left(\mathbf{b}^\top\mathbf{x}\right)\,. \qquad (1.4.3.3)$$

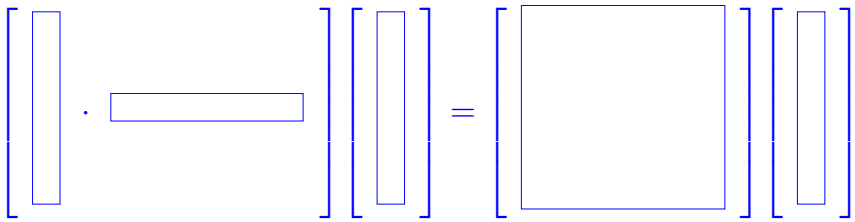
```
T = (a*b.transpose())*x;
```

```
t = a*b.dot(x);
```

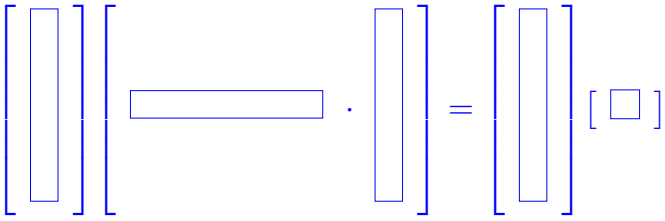
➤ complexity $O(mn)$

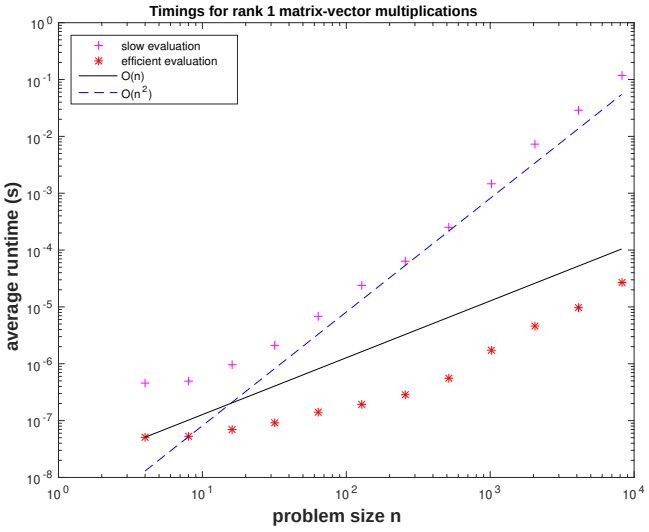
➤ complexity $O(n+m)$ ("linear complexity")

Visualization of evaluation according to (1.4.3.2):



Visualization of evaluation according to (1.4.3.3):





◁ average runtimes for efficient/inefficient matrix×vector multiplication with rank-1 matrices , see § 1.4.1.6 for the rationale behind choosing a *doubly logarithmic plot*.

Platform:

- ubuntu 14.04 LTS
- i7-3517U CPU @ 1.90GHz
- L1 32 KB, L2 256 KB, L3 4096 KB,
- 8 GB main memory
- gcc 4.8.4, -O3

C++ code 1.4.3.4: EIGEN code for Ex. 1.4.3.1 ↪ GITLAB

```
//!  This function compares the runtimes for the multiplication
//!  of a vector with a rank-1 matrix ab^T, a,b ∈ R^n
//!  using different associative evaluations.
//!  Runtime measurements consider minimal time for
//!  several (nruns) runs
MatrixXd dottenstiming() {
  const int nruns = 3, minExp = 2, maxExp = 13;
  // Matrix for storing recorded runtimes
  MatrixXd timings(maxExp-minExp+1,3);
  for(int i = 0; i <= maxExp-minExp; ++i){
    Timer tfool, tsmart; // Timer objects
    const int n = std::pow(2, minExp + i);
    VectorXd a = VectorXd::LinSpaced(n,1,n);
    VectorXd b = VectorXd::LinSpaced(n,1,n).reverse();
    VectorXd x = VectorXd::Random(n,1), y(n);
    for(int j = 0; j < nruns; ++j){
      // Grossly wasteful evaluation
      tfool.start(); y = (a*b.transpose())*x;      tfool.stop();
      // Efficient implementation
      tsmart.start(); y = a * b.dot(x); tsmart.stop();
    }
    timings(i,0)=n;
    timings(i,1)=tsmart.min(); timings(i,2)=tfool.min();
  }
  return timings;
}
```

Complexity can sometimes be reduced by reusing intermediate results.

EXAMPLE 1.4.3.5 (Hidden summation) The asymptotic complexity of the EIGEN code

```
Eigen::MatrixXd AB = A*B.transpose();
y = AB.triangularView<Eigen::Upper>()*x;
```

when supplied with two low-rank matrices $\mathbf{A},\mathbf{B} \in \mathbb{K}^{n,p}$, $p \ll n$, in terms of $n \to \infty$ obviously is $O(n^2)$, because an intermediate $n \times n$-matrix $\mathbf{A}\mathbf{B}^T$ is built.

First, consider the case of a tensor product (= rank-1) matrix, that is, $p = 1$, $\mathbf{A} \leftrightarrow \mathbf{a} = [a_1,\ldots,a_n]^\top \in \mathbb{K}^n$, $\mathbf{B} \leftrightarrow \mathbf{b} = [b_1,\ldots,b_n] \in \mathbb{K}^n$. Then

$$\mathbf{y} = \operatorname{triu}(\mathbf{a}\mathbf{b}^T)\mathbf{x} = \begin{bmatrix} a_1b_1 & a_1b_2 & \ldots & & \ldots & a_1b_n \\ 0 & a_2b_2 & a_2b_3 & \ldots & \ldots & a_2b_n \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & \vdots \\ 0 & \ldots & & \ldots & 0 & a_nb_n \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ \\ \\ \vdots \\ x_n \end{bmatrix}$$

$$= \begin{bmatrix} a_1 & & \\ & \ddots & \\ & & a_n \end{bmatrix} \left(\underbrace{\begin{bmatrix} 1 & 1 & \dots & & \dots & 1 \\ 0 & 1 & 1 & \dots & \dots & 1 \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & \vdots \\ 0 & \dots & & \dots & 0 & 1 \end{bmatrix}}_{\mathbf{T}} \left(\begin{bmatrix} b_1 & & \\ & \ddots & \\ & & b_n \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ \\ \\ \vdots \\ x_n \end{bmatrix} \right) \right) .$$

The brackets indicate the order of the matrix×vector multiplications. Thus, the core problem is the fast multiplication of a vector with an upper triangular matrix $\mathbf{T}$ described in EIGEN syntax by `Eigen::MatrixXd::Ones(n,n).triangularView<Eigen::Upper>()`. Note that multiplication of a vector $\mathbf{x}$ with $\mathbf{T}$ yields a vector of partial sums of components of $\mathbf{x}$ starting from last component:

$$\begin{bmatrix} 1 & 1 & \dots & & \dots & 1 \\ 0 & 1 & 1 & \dots & \dots & 1 \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & \vdots \\ 0 & \dots & & \dots & 0 & 1 \end{bmatrix} \begin{bmatrix} v_1 \\ \vdots \\ \\ \\ \vdots \\ v_n \end{bmatrix} = \begin{bmatrix} s_n \\ s_{n-1} \\ \vdots \\ \\ \vdots \\ s_1 \end{bmatrix}, \quad s_j := \sum_{k=n-j+1}^{n} v_k .$$

This can be achieved by invoking the special C++ command `std::partial_sum` from the C++ standard library (documentation). We also observe that

$$\mathbf{AB}^T = \sum_{\ell=1}^{p} (\mathbf{A})_{:,\ell} ((\mathbf{B})_{:,\ell})^\top ,$$

so that the computations for the special case $p = 1$ discussed above can simply be reused p times!

C++ code 1.4.3.6: Efficient multiplication with the upper diagonal part of a rank-p-matrix in EIGEN ↪ GITLAB

```
2  //!  Computation of y = triu(AB^T)x
3  //!  Efficient implementation with backward cumulative sum
4  //!  (partial_sum)
5  template<class Vec, class Mat>
6  void lrtrimulteff(const Mat& A, const Mat& B, const Vec& x, Vec& y){
7    const int n = A.rows();
8    const int p = A.cols();
9    assert( n == B.rows() && p == B.cols()); // size mismatch
10   for(int l = 0; l < p; ++l){
11     Vec tmp = (B.col(l).array() * x.array()).matrix().reverse();
12     std::partial_sum(tmp.begin(), tmp.end(), tmp.begin());
13     y += (A.col(l).array() * tmp.reverse().array()).matrix();
14   }
15 }
```

This code enjoys the obvious complexity of $O(pn)$ for $p, n \to \infty$, $p < n$. The code offers an example of a function templated with its argument types, see § 0.3.2.1. The types **Vec** and **Mat** must fit the concept of EIGEN vectors/matrices. ⌟

The next concept from linear algebra is important in the context of computing with multi-dimensional arrays.

Definition 1.4.3.7. Kronecker product

The **Kronecker product** $\mathbf{A} \otimes \mathbf{B}$ of two matrices $\mathbf{A} \in \mathbb{K}^{m,n}$ and $\mathbf{B} \in \mathbb{K}^{l,k}$, $m,n,l,k \in \mathbb{N}$, is the $(ml) \times (nk)$-matrix

$$\mathbf{A} \otimes \mathbf{B} := \begin{bmatrix} (\mathbf{A})_{1,1}\mathbf{B} & (\mathbf{A})_{1,2}\mathbf{B} & \dots & \dots & (\mathbf{A})_{1,n}\mathbf{B} \\ (\mathbf{A})_{2,1}\mathbf{B} & (\mathbf{A})_{2,2}\mathbf{B} & & & \vdots \\ \vdots & \vdots & & & \vdots \\ \vdots & \vdots & & & \vdots \\ (\mathbf{A})_{m,1}\mathbf{B} & (\mathbf{A})_{m,2}\mathbf{B} & \dots & \dots & (\mathbf{A})_{m,n}\mathbf{B} \end{bmatrix} \in \mathbb{K}^{ml,nk} .$$

EXAMPLE 1.4.3.8 (Multiplication of Kronecker product with vector) The function $(\mathbf{A} \otimes \mathbf{B})\mathbf{x}$ when invoked with two matrices $\mathbf{A} \in \mathbb{K}^{m,n}$ and $\mathbf{B} \in \mathbb{K}^{l,k}$ and a vector $\mathbf{x} \in \mathbb{K}^{nk}$, will suffer an asymptotic complexity of $O(m \cdot n \cdot l \cdot k)$, determined by the size of the intermediate dense matrix $\mathbf{A} \otimes \mathbf{B} \in \mathbb{K}^{ml,nk}$.

Using the partitioning of the vector $\mathbf{x}$ into n equally long sub-vectors

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}^1 \\ \mathbf{x}^2 \\ \vdots \\ \mathbf{x}^n \end{bmatrix} \quad , \quad \mathbf{x}^j \in \mathbb{K}^k ,$$

we find the representation

$$(\mathbf{A} \otimes \mathbf{B})\mathbf{x} = \begin{bmatrix} (\mathbf{A})_{1,1}\mathbf{B}\mathbf{x}^1 + (\mathbf{A})_{1,2}\mathbf{B}\mathbf{x}^2 + \cdots + (\mathbf{A})_{1,n}\mathbf{B}\mathbf{x}^n \\ (\mathbf{A})_{2,1}\mathbf{B}\mathbf{x}^1 + (\mathbf{A})_{2,2}\mathbf{B}\mathbf{x}^2 + \cdots + (\mathbf{A})_{2,n}\mathbf{B}\mathbf{x}^n \\ \vdots \\ \vdots \\ (\mathbf{A})_{m,1}\mathbf{B}\mathbf{x}^1 + (\mathbf{A})_{m,2}\mathbf{B}\mathbf{x}^2 + \cdots + (\mathbf{A})_{m,n}\mathbf{B}\mathbf{x}^n \end{bmatrix} .$$

The idea is to form the products $\mathbf{B}\mathbf{x}^j$, $j = 1, \dots, n$, once, and then combine them linearly with coefficients given by the entries in the rows of $\mathbf{A}$:

C++ code 1.4.3.9: Efficient multiplication of Kronecker product with vector in EIGEN
➥ GITLAB

```
2  template <class Matrix, class Vector>
3  Vector kronmultv(const Matrix &A, const Matrix &B, const Vector &x){
4      const size_t m = A.rows();
5      const size_t n = A.cols();
6      const size_t l = B.rows();
7      const size_t k = B.cols();
8      // 1st matrix mult.  computes the products Bx^j
9      // 2nd matrix mult.  combines them linearly with the coefficients of
           A
10     Matrix t = B * Matrix::Map(x.data(),k,n) * A.transpose(); //
11     return Matrix::Map(t.data(), m*l, 1);
12 }
```

Recall the reshaping of a matrix in EIGEN in order to understand this code: Rem. 1.2.3.6. Note a new twist: Here the `Map()` member function of an EIGEN data type is used, where `X::Map(<args>)` is roughly equivalent to `Eigen::Map<X>(<args>)`.

The asymptotic complexity of this code is determined by the two matrix multiplications in Line 10. This yields the asymptotic complexity $O(lkn + mnl)$ for $l,k,m,n \to \infty$.

PYTHON-code 1.4.3.10: Efficient multiplication of Kronecker product with vector in PYTHON

```
1 def kronmultv(A, B, x):
2     n, k = A.shape[1], B.shape[1]
3     assert x.size == n * k, 'size mismatch'
4     xx = np.reshape(x, (n, k))
5     Z = np.dot(xx, B.T)
6     yy = np.dot(A, Z)
7     return np.ravel(yy)
```

Note that different reshaping is used in the PYTHON code due to the default row major storage order. ⌟

Review question(s) 1.4.3.11 (Computational effort)

(Q1.4.3.11.A) Explain why the classical concept of "computational effort" (= computational cost) is only loosely related to the runtime of a concrete implementation of an algorithm.

(Q1.4.3.11.B) We are given two dense matrices $\mathbf{A},\mathbf{B} \in \mathbb{R}^{n,p}$, $n,p \in \mathbb{N}$, $p < n$ fixed. What is the asmptotic complexity of each of the following two lines of code in terms of $n,p \to \infty$?

```
Eigen::MatrixXd AB = A*B.transpose();
y = AB.triangularView<Eigen::Upper>()*x;
```

(Q1.4.3.11.C) Given a vector $\mathbf{u} \in \mathbb{R}^n$, $n \in \mathbb{N}$, we consider the matrix

$$\mathbf{A} \in \mathbb{R}^{n,n}: \quad (\mathbf{A})_{i,j} = u_i + u_j + u_i u_j\,, \quad i,j \in \{1,\ldots,n\}\,.$$

Outline an efficient algorithm for computing the matrix-vector product $\mathbf{Ax}$, $\mathbf{x} \in \mathbb{R}^n$. What is its asymptotic complexity for $n \to \infty$?

(Q1.4.3.11.D) [Matrix×vector multiplication involving Kronecker product, *cf.* Ex. 1.4.3.8] How do you have to modify the implementation of the function `kronmultv()` from Ex. 1.4.3.8 to ensure that it is still efficient even when called with long column or row vectors as arguments `A` and `B`, in particular in the cases $\mathbf{A} \in \mathbb{R}^{n,1}$, $\mathbf{B} \in \mathbb{R}^{k,1}$ or $\mathbf{A} \in \mathbb{R}^{1,n}$, $\mathbf{B} \in \mathbb{R}^{1,k}$, $n,k \gg 1$?

C++ code 1.4.3.9: (Potentially not so) efficient multiplication of Kronecker product with vector in EIGEN ↬ GITLAB

```
2  template <class Matrix, class Vector>
3  Vector kronmultv(const Matrix &A, const Matrix &B, const Vector &x){
4      const size_t m = A.rows();
5      const size_t n = A.cols();
6      const size_t l = B.rows();
7      const size_t k = B.cols();
8      // 1st matrix mult.  computes the products Bx^j
9      // 2nd matrix mult.  combines them linearly with the coefficients
           of A
10     Matrix t = B * Matrix::Map(x.data(),k,n) * A.transpose(); //
11     return Matrix::Map(t.data(), m*l, 1);
12 }
```

(Q1.4.3.11.E) [Multiplication with a "kernel matrix"] Ex. 1.4.3.5 gave us efficient functions

```
Eigen::VectorXd mvtriutp(const Eigen::Vector &a,
  const Eigen::Vector &b, const Eigen::Vector &x);
```

```
Eigen::VectorXd mvtriltp(const Eigen::Vector &a,
  const Eigen::Vector &b, const Eigen::Vector &x);
```

for evaluating

$$\mathbf{y} = \mathrm{triu}(\mathbf{a}\mathbf{b}^\top)\mathbf{x} \quad \text{and} \quad \mathbf{y} = \mathrm{tril}(\mathbf{a}\mathbf{b}^\top)\mathbf{x}\,, \quad \mathbf{a},\mathbf{b},\mathbf{x} \in \mathbb{R}^n\,.$$

for large $n \in \mathbb{N}$. Here, triu and tril extract the upper and lower triangular part of a matrix, respectively.

Explain how this function can be used for the efficient computation of the matrix×vector product

$$\left[e^{\alpha|k-j|}\right]_{k,j=1}^{n}\mathbf{x}\,, \quad \mathbf{x} \in \mathbb{R}^n\,, \quad \alpha \in \mathbb{R}\,.$$

(Q1.4.3.11.F) [Initialization of a "kernel matrix"] A call to the black-box function **double** `f(`**unsigned int** `i)` might be expensive. Outline an efficient C++ code for the initialization of the matrix

$$[\mathtt{f}(|k-j|)]_{k.j=1}^{n} \in \mathbb{R}^{n,n}\,.$$

△

1.5 Machine Arithmetic and Consequences

Video tutorial for Section 1.5 "Machine Arithmetic and Consequences": (16 minutes) Download link, tablet notes

→ review questions 1.5.3.18

1.5.1 Experiment: Loss of Orthogonality

§1.5.1.1 (Gram-Schmidt orthogonalisation) From linear algebra [NS02, Sect. 4.4] or Ex. 0.3.5.29 we recall the fundamental algorithm of Gram-Schmidt orthogonalisation of an ordered finite set $\{\mathbf{a}^1,\ldots,\mathbf{a}^k\}$, $k \in \mathbb{N}$, of vectors $\mathbf{a}^\ell \in \mathbb{K}^n$:

Input: $\{\mathbf{a}^1,\ldots,\mathbf{a}^k\} \subset \mathbb{K}^n$

1: $\mathbf{q}^1 := \frac{\mathbf{a}^1}{\|\mathbf{a}^1\|_2}$ *% 1st output vector*
2: **for** $j = 2,\ldots,k$ **do**
{ *% Orthogonal projection*
3: $\mathbf{q}^j := \mathbf{a}^j$
4: **for** $\ell = 1,2,\ldots,j-1$ **do** (GS)
5: { $\mathbf{q}^j \leftarrow \mathbf{q}^j - \mathbf{a}^j \cdot \mathbf{q}^\ell \mathbf{q}^\ell$ }
6: **if** ($\mathbf{q}^j = \mathbf{0}$) **then** STOP
7: **else** { $\mathbf{q}^j \leftarrow \frac{\mathbf{q}^j}{\|\mathbf{q}^j\|_2}$ }
8: }

Output: $\{\mathbf{q}^1,\ldots,\mathbf{q}^j\}$

In linear algebra we have learnt that, if it does not STOP prematurely, this algorithm will compute orthonormal vectors $\mathbf{q}^1,\ldots,\mathbf{q}^k$ satisfying

$$\mathrm{Span}\{\mathbf{q}^1,\ldots,\mathbf{q}^\ell\} = \mathrm{Span}\{\mathbf{a}^1,\ldots,\mathbf{a}^\ell\}\,, \tag{1.5.1.2}$$

for all $\ell \in \{1,\ldots,k\}$.

More precisely, if $\mathbf{a}^1,\ldots,\mathbf{a}^\ell$, $\ell \le k$, are linearly independent, then the Gram-Schmidt algorithm will not terminate before the $\ell+1$-th step.

✎ Notation: $\|\cdot\|_2 \triangleq$ Euclidean norm of a vector $\in \mathbb{K}^n$

The following code implements the Gram-Schmidt orthonormalization of a set of vectors passed as the columns of a matrix $\mathbf{A} \in \mathbb{R}^{n,k}$. The template paramter **Matrix** should match the concept of a matrix type in EIGEN like **Eigen::MatrixXd** or

C++ code 1.5.1.3: Gram-Schmidt orthogonalisation in EIGEN ⇝ GITLAB

```
2  template <class Matrix> Matrix gramschmidt(const Matrix &A) {
3    Matrix Q = A;
4    // First vector just gets normalized, Line 1 of (GS)
5    Q.col(0).normalize();
6    for (unsigned int j = 1; j < A.cols(); ++j) {
7      // Replace inner loop over each previous vector in Q with fast
8      // matrix-vector multiplication (Lines 4, 5 of (GS))
9      Q.col(j) -= Q.leftCols(j) *
10                 (Q.leftCols(j).adjoint() * A.col(j)); //
11     // Normalize vector, if possible.
12     // Otherwise colums of A must have been linearly dependent
13     if (Q.col(j).norm() > 10e-9 * A.col(j).norm()) { //
14       Q.col(j).normalize(); // Line 7 of (GS)
15     } else {
16       std::cerr << "Gram-Schmidt failed: A has lin. dep columns." << std::endl;
17       break;
18     }
19   }
20   return Q;
21 }
```

We will soon learn the rationale behind the odd test in Line 13.

In PYTHON the same algorithm can be implemented as follows:

PYTHON-code 1.5.1.4: Gram-Schmidt orthogonalisation in PYTHON

```
1  def gramschmidt(A):
2      _, k = A.shape
3      Q = A[:, [0]] / np.linalg.norm(A[:, 0])
4      for j in range(1, k):
5          q = A[:, j] - np.dot(Q, np.dot(Q.T, A[:, j]))
6          nq = np.linalg.norm(q)
7          if nq < 1e-9 * np.linalg.norm(A[:, j]):
8              break
9          Q = np.column_stack([Q, q / nq])
10     return Q
```

Note the different loop range due to the zero-based indexing in PYTHON.

⌟

EXPERIMENT 1.5.1.5 (Unstable Gram-Schmidt orthonormalization) If $\{\mathbf{a}^1,\ldots,\mathbf{a}^k\}$ are linearly independent we expect the output vectors $\mathbf{q}^1,\ldots,\mathbf{q}^k$ to be orthonormal:

$$(\mathbf{q}^\ell)^\top \mathbf{q}^m = \delta_{\ell,m}\,, \quad \ell,m \in \{1,\ldots,k\}\,. \tag{1.5.1.6}$$

This property can be easily tested numerically, for instance by computing $\mathbf{Q}^\top\mathbf{Q}$ for a matrix $\mathbf{Q} = [\mathbf{q}^1,\ldots,\mathbf{q}^k] \in \mathbb{R}^{n,k}$.

C++ code 1.5.1.7: Wrong result from Gram-Schmidt orthogonalisation EIGEN ⇝ GITLAB

```
2  void gsroundoff(MatrixXd& A){
3    // Gram-Schmidt orthogonalization of columns of A, see Code 1.5.1.3
4    MatrixXd Q = gramschmidt(A);
5    // Test orthonormality of columns of Q, which should be an
6    // orthogonal matrix according to theory
```

```
7   cout << setprecision(4) << fixed << "I = "
8        << endl << Q.transpose()*Q << endl;
9   // EIGEN's stable internal Gram-Schmidt orthogonalization by
10  // QR-decomposition, see Rem. 1.5.1.9 below
11  HouseholderQR<MatrixXd> qr(A.rows(),A.cols()); //
12  qr.compute(A); MatrixXd Q1 = qr.householderQ(); //
13  // Test orthonormality
14  cout << "I1 = " << endl << Q1.transpose()*Q1 << endl;
15  // Check orthonormality and span property (1.5.1.2)
16  const MatrixXd R1 = qr.matrixQR().triangularView<Upper>();
17  cout << scientific << "A-Q1*R1 = " << endl << A-Q1*R1 << endl;
18 }
```

We test the orthonormality of the output vectors of Gram-Schmidt orthogonalization for a special matrix $\mathbf{A} \in \mathbb{R}^{10,10}$, a so-called Hilbert matrix, defined by $(\mathbf{A})_{i,j} = (i+j-1)^{-1}$. Then Code 1.5.1.7 produces the follwing output:

```
I =
   1.0000    0.0000   -0.0000    0.0000   -0.0000    0.0000   -0.0000   -0.0000   -0.0000   -0.0000
   0.0000    1.0000   -0.0000    0.0000   -0.0000    0.0000   -0.0000   -0.0000   -0.0000   -0.0000
  -0.0000   -0.0000    1.0000    0.0000   -0.0000    0.0000   -0.0000   -0.0000   -0.0000   -0.0000
   0.0000    0.0000    0.0000    1.0000   -0.0000    0.0000   -0.0000   -0.0000   -0.0000   -0.0000
  -0.0000   -0.0000   -0.0000   -0.0000    1.0000    0.0000   -0.0008   -0.0007   -0.0007   -0.0006
   0.0000    0.0000    0.0000    0.0000    0.0000    1.0000   -0.0540   -0.0430   -0.0360   -0.0289
  -0.0000   -0.0000   -0.0000   -0.0000   -0.0008   -0.0540    1.0000    0.9999    0.9998    0.9996
  -0.0000   -0.0000   -0.0000   -0.0000   -0.0007   -0.0430    0.9999    1.0000    1.0000    0.9999
  -0.0000   -0.0000   -0.0000   -0.0000   -0.0007   -0.0360    0.9998    1.0000    1.0000    1.0000
  -0.0000   -0.0000   -0.0000   -0.0000   -0.0006   -0.0289    0.9996    0.9999    1.0000    1.0000
```

Obviously, the vectors produced by the function `gramschmidt` fail to be orthonormal, contrary to the predictions of rigorous results from linear algebra!

However, Line 11, Line 12 of Code 1.5.1.7 demonstrate another way to orthonormalize the columns of a matrix using EIGEN's built-in class template **HouseholderQR** (more details in Section 3.3.3).

```
I1 =
 1.0000 -0.0000  0.0000 -0.0000 -0.0000 -0.0000 -0.0000 -0.0000 -0.0000  0.0000
-0.0000  1.0000 -0.0000  0.0000  0.0000 -0.0000 -0.0000 -0.0000  0.0000 -0.0000
 0.0000 -0.0000  1.0000 -0.0000 -0.0000  0.0000  0.0000  0.0000  0.0000  0.0000
-0.0000  0.0000 -0.0000  1.0000  0.0000 -0.0000 -0.0000  0.0000  0.0000  0.0000
-0.0000  0.0000 -0.0000  0.0000  1.0000 -0.0000  0.0000 -0.0000 -0.0000  0.0000
-0.0000 -0.0000  0.0000 -0.0000 -0.0000  1.0000 -0.0000 -0.0000  0.0000 -0.0000
-0.0000 -0.0000  0.0000 -0.0000  0.0000 -0.0000  1.0000  0.0000  0.0000 -0.0000
-0.0000 -0.0000  0.0000  0.0000 -0.0000 -0.0000  0.0000  1.0000 -0.0000  0.0000
-0.0000  0.0000  0.0000  0.0000 -0.0000  0.0000  0.0000 -0.0000  1.0000 -0.0000
 0.0000 -0.0000  0.0000  0.0000  0.0000 -0.0000 -0.0000  0.0000 -0.0000  1.0000
```

Now we observe apparently perfect orthogonality (1.5.1.6) of the columns of the matrix `Q1` in Code 1.5.1.7. Obviously, there is another algorithm that reliably yields the theoretical output of Gram-Schmidt orthogonalization. There is no denying that it is possible to compute Gram-Schmidt orthonormalization in a "clean" way. ⌟

"Computers cannot compute"

Computers cannot compute "properly" in $\mathbb{R}$: numerical computations may not respect the laws of analysis and linear algebra!

This introduces an important new aspect in the study of numerical algorithms.

Remark 1.5.1.9 (Stable orthonormalization by QR-decomposition) In Code 1.5.1.7 we saw the use of the EIGEN class `HousholderQR<MatrixType>` for the purpose of Gram-Schmidt orthogonalisation.

The underlying theory and algorithms will be explained later in Section 3.3.3. There we will have the following insight:

➢ Up to signs the columns of the matrix $\mathbf{Q}$ available from the QR-decomposition of $\mathbf{A}$ are the same vectors as produced by the Gram-Schmidt orthogonalisation of the columns of $\mathbf{A}$.

! Code 1.5.1.7 demonstrates a case where a desired result can be obtained by two *algebraically equivalent* computations, that is, they yield the same result in a mathematical sense. Yet, when implemented on a computer, the results can be *vastly different*. One algorithm may produce junk ("unstable algorithm"), whereas the other lives up to the expectations ("stable algorithm")

Supplement to Exp. 1.5.1.5: despite its ability to produce orthonormal vectors, we get as output for `D=A-Q1*R1` in Code 1.5.1.7:

```
D =
 2.2204e-16  3.3307e-16  3.3307e-16  1.9429e-16  1.9429e-16  5.5511e-17  1.3878e-16  6.9389e-17  8.3267e-17  9.7145e-17
 0.0000e+00  1.1102e-16  8.3267e-17  5.5511e-17  0.0000e+00  5.5511e-17 -2.7756e-17  0.0000e+00  0.0000e+00  4.1633e-17
-5.5511e-17  5.5511e-17  2.7756e-17  5.5511e-17  0.0000e+00  0.0000e+00  0.0000e+00 -1.3878e-17  1.3878e-17  1.3878e-17
 0.0000e+00  5.5511e-17  2.7756e-17  2.7756e-17  0.0000e+00  1.3878e-17 -1.3878e-17  0.0000e+00  1.3878e-17  2.7756e-17
 0.0000e+00  2.7756e-17  0.0000e+00  1.3878e-17  1.3878e-17  1.3878e-17  0.0000e+00  1.3878e-17  1.3878e-17  4.1633e-17
-2.7756e-17  2.7756e-17  1.3878e-17  4.1633e-17  2.7756e-17  1.3878e-17  0.0000e+00 -1.3878e-17  2.7756e-17  2.7756e-17
 0.0000e+00  2.7756e-17  0.0000e+00  2.7756e-17  2.7756e-17  1.3878e-17  0.0000e+00  1.3878e-17  2.7756e-17  2.0817e-17
 0.0000e+00  2.7756e-17  2.7756e-17  1.3878e-17  1.3878e-17  1.3878e-17  0.0000e+00  1.3878e-17  2.0817e-17  2.7756e-17
 1.3878e-17  1.3878e-17  1.3878e-17  2.7756e-17  1.3878e-17  0.0000e+00 -1.3878e-17  6.9389e-18 -6.9389e-18  1.3878e-17
 0.0000e+00  2.7756e-17  1.3878e-17  1.3878e-17  1.3878e-17  0.0000e+00  0.0000e+00  0.0000e+00  1.3878e-17  1.3878e-17
```

➥ The computed QR-decomposition apparently fails to meet the exact algebraic requirements stipulated by Thm. 3.3.3.4. However, note the tiny size of the "defect". ⌟

1.5.2 Machine Numbers

§1.5.2.1 (The finite and discrete set of machine numbers) The reason, why computers must fail to execute exact computations with real numbers is clear:

Computer = finite automaton ➢ can handle only *finitely many* numbers, not $\mathbb{R}$

machine numbers, set $\mathbb{M}$

Essential property: $\mathbb{M}$ is a finite, discrete subset of $\mathbb{R}$ (its numbers separated by gaps)

The set of machine numbers $\mathbb{M}$ cannot be closed under elementary arithmetic operations $+, -, \cdot, /$, that is, when adding, multiplying, etc., two machine numbers the result may not belong to $\mathbb{M}$.

The results of elementary operations with operands in $\mathbb{M}$ have to be mapped back to $\mathbb{M}$, an operation called rounding.

roundoff errors (*ger.:* Rundungsfehler) are inevitable

⌟

The impact of roundoff means that mathematical identities may not carry over to the computational realm. As we have seen above in Exp. 1.5.1.5

Computers cannot compute "properly" !

numerical computations $\neq$ **analysis linear algebra**

This introduces a *new* and *important* aspect in the study of numerical algorithms!

§1.5.2.2 (Internal representation of machine numbers) Now we give a brief sketch of the internal structure of machine numbers $\in \mathbb{M}$. The main insight will be that

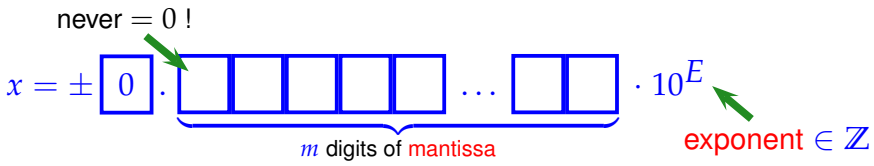
"Computers use floating point numbers (scientific notation)"

EXAMPLE 1.5.2.3 (Decimal floating point numbers) Some 3-digit normalized decimal floating point numbers:

valid: $0.723 \cdot 10^2$, $0.100 \cdot 10^{-20}$, $-0.801 \cdot 10^5$

invalid: $0.033 \cdot 10^2$, $1.333 \cdot 10^{-4}$, $-0.002 \cdot 10^3$

General form of an m-digit normalized decimal floating point number:



Of course, computers are restricted to a *finite range* of exponents:

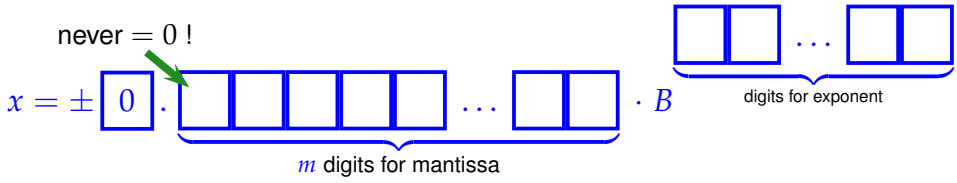
Definition 1.5.2.4. Machine numbers/floating point numbers → [AG11, Sect. 2.2]

Given
- ☞ basis $B \in \mathbb{N} \setminus \{1\}$,
- ☞ exponent range $\{e_{\min}, \ldots, e_{\max}\}$, $e_{\min}, e_{\max} \in \mathbb{Z}$, $e_{\min} < e_{\max}$,
- ☞ number $m \in \mathbb{N}$ of digits (for mantissa),

the corresponding set of machine numbers is

$$\mathbb{M} := \{d \cdot B^E : d = i \cdot B^{-m}, i = B^{m-1}, \ldots, B^m - 1, E \in \{e_{\min}, \ldots, e_{\max}\}\}$$

machine number $\in \mathbb{M}$:



Remark 1.5.2.5 (Extremal numbers in $\mathbb{M}$) Clearly, there is a largest element of $\mathbb{M}$ and two that are closest to zero. These are mainly determined by the range for the exponent E, *cf.* Def. 1.5.2.4.

Largest machine number (in modulus) : $x_{\max} = \max|\mathbb{M}| = (1 - B^{-m}) \cdot B^{e_{\max}}$

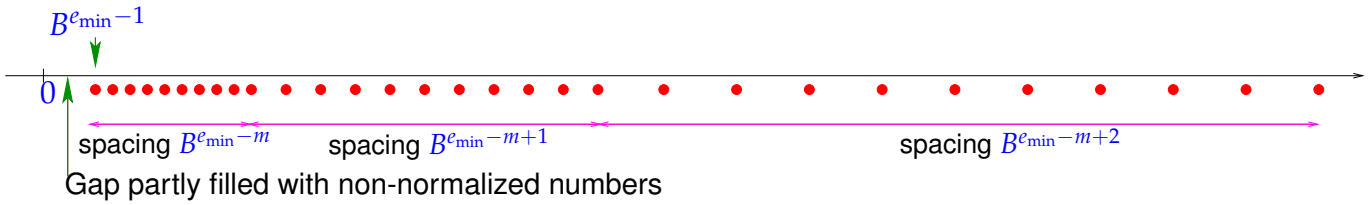
Smallest machine number (in modulus) : $x_{\min} = \min|\mathbb{M}| = B^{-1} \cdot B^{e_{\min}}$

In C++ these extremal machine numbers are accessible through the

`std::numeric_limits<double>::max()`
and `std::numeric_limits<double>::min()`

functions. Other properties of arithmetic types can be queried accordingly from the numeric_limits header. ⌟

Remark 1.5.2.6 (Distribution of machine numbers) From Def. 1.5.2.4 it is clear that there are equi-spaced sections of $\mathbb{M}$ and that the gaps between machine numbers are bigger for larger numbers, see also [AG11, Fig. 2.3].



Non-normalized numbers violate the lower bound for the mantissa i in Def. 1.5.2.4. ⌟

§1.5.2.7 (IEEE standard 754 for machine numbers → [Ove01], [AG11, Sect. 2.4], → link) No surprise: for modern computers $B = 2$ (binary system), the other parameters of the universally implemented machine number system are

single precision : $m = 24^*, E \in \{-125, \ldots, 128\}$ ➢ 4 bytes

double precision : $m = 53^*, E \in \{-1021, \ldots, 1024\}$ ➢ 8 bytes

$*$: including bit indicating sign

The standardisation of machine numbers is important, because it ensures that the same numerical algorithm, executed on different computers will nevertheless produce the same result. ⌟

Remark 1.5.2.8 (Special cases in IEEE standard)

The IEEE standard makes provisions for exceptions triggered by overflow or invalid floating point operations. The following C++ code snippet shows cases when these "flags" are raised.

```
const double x = exp(1000);
const double y = 3/x;
const double z = x*sin(M_PI);
const double w = x*log(1);
cout << x << endl << y << endl << z << endl << w << endl;
```

Output:

```
1 inf
2 0
3 inf
4 -nan
```

$E = e_{\max}$, $M \neq 0$ ≙ NaN = *Not a number* → exception
$E = e_{\max}$, $M = 0$ ≙ Inf = *Infinity* → overflow
$E = 0$ ≙ Non-normalized numbers → underflow
$E = 0$, $M = 0$ ≙ number 0

In C++ these flags can be tested with the functions **std::isnan()** C++ reference and **std::isinf()** C++ reference. ⌟

§1.5.2.9 (Characteristic parameters of IEEE floating point numbers (double precision))

☞ C++ does not always fulfill the requirements of the IEEE 754 standard and it needs to be checked with `std::numeric_limits<T>::is_iec559`.

C++ code 1.5.2.10: Querying characteristics of `double` numbers ↪ GITLAB

```
2
3  #include <iomanip>
4  #include <iostream>
5  #include <limits>
6
7  using std::cout;
8  using std::endl;
9  using std::numeric_limits;
10
11 int main(){
12      cout << numeric_limits<double>::is_iec559 << endl
13      << std::defaultfloat << numeric_limits<double>::min() << endl
14      << std::hexfloat << numeric_limits<double>::min() << endl
15      << std::defaultfloat << numeric_limits<double>::max() << endl
16      << std::hexfloat << numeric_limits<double>::max() << endl;
17 }
```

Output:

```
1 true
2 2.22507e-308
3 0010000000000000
4 1.79769e+308
5 7fefffffffffffff
```

⌟

1.5.3 Roundoff Errors

EXPERIMENT 1.5.3.1 (Input errors and roundoff errors) The following computations would always result in 0, if done in exact arithmetic.

C++ code 1.5.3.2: Demonstration of roundoff errors ↪ GITLAB

```
2  #include <iostream>
3  int main(){
4      std::cout.precision(15);
5      double a = 4.0/3.0;
6      double b = a-1;
7      double c = 3*b;
8      double e = 1-c;
9      std::cout << e << std::endl;
10     a = 1012.0/113.0; b = a-9; c = 113*b; e = 5+c;
11     std::cout << e << std::endl;
12     a = 83810206.0/6789.0; b = a-12345; c = 6789*b; e = c-1;
13     std::cout << e << std::endl;
14 }
```

Output:

```
1 2.22044604925031e-16
2 6.75015598972095e-14
3 -1.60798663273454e-09
```

Can you devise a similar calculation, whose result is even farther off zero? Apparently the rounding that inevitably accompanies arithmetic operations in $\mathbb{M}$ can lead to results that are far away from the true result. ⌟

For the discussion of errors introduced by rounding we need important notions.

Definition 1.5.3.3. Absolute and relative error → [AG11, Sect. 1.2]

Let $\widetilde{x} \in \mathbb{K}$ be an approximation of $x \in \mathbb{K}$. Then its absolute error is given by

$$\epsilon_{\text{abs}} := |x - \widetilde{x}| ,$$

and its relative error is defined as

$$\epsilon_{\text{rel}} := \frac{|x - \widetilde{x}|}{|x|} .$$

Remark 1.5.3.4 (Relative error and number of correct digits) The number of correct (significant, valid) digits of an approximation $\widetilde{x}$ of $x \in \mathbb{K}$ is defined through the relative error:

$$\text{If } \epsilon_{\text{rel}} := \frac{|x-\widetilde{x}|}{|x|} \leq 10^{-\ell}, \text{ then } \widetilde{x} \text{ has } \ell \text{ correct digits, } \ell \in \mathbb{N}_0$$

To see this write write as base-10 floating point numbers

$$\begin{aligned} x &= d_1.d_2d_3\ldots d_m d_{m+1}\ldots d_n \cdot 10^E , \\ \widetilde{x} &= d_1.d_2d_3\ldots d_m \widetilde{d}_{m+1}\ldots \widetilde{d}_n \cdot 10^E , \end{aligned} \qquad d_i, \widetilde{d}_i \in \{0,\ldots,10\} , \quad d_1 \neq 0 , d_{m+1} \neq \widetilde{d}_{m+1} .$$

This means that $\widetilde{x}$ has m correct digits. We compute the relative error

$$\epsilon = \frac{|x - \widetilde{x}|}{|x|} = \frac{0.\overbrace{0\ldots0}^{m}\delta_{m+1}\delta_{m+1}\ldots\delta_n}{d_1.d_2d_3\ldots d_m\widetilde{d}_{m+1}\ldots\widetilde{d}_n} , \quad \delta_{m+1} \neq 0$$

Obviously we have $\epsilon \approx 10^{-m}$. ⌟

§1.5.3.5 (Floating point operations) We may think of the elementary binary operations $+, -, *, /$ in $\mathbb{M}$ comprising two steps:

❶ Compute the exact result of the operation.

❷ Perform rounding of the result of ❶ to map it back to $\mathbb{M}$.

Definition 1.5.3.6. Correct rounding

Correct rounding ("rounding up") is given by the function

$$\text{rd} : \left\{ \begin{array}{rcl} \mathbb{R} & \to & \mathbb{M} \\ x & \mapsto & \max \operatorname{argmin}_{\widetilde{x}\in\mathbb{M}} |x - \widetilde{x}| . \end{array} \right.$$

(Recall that $\operatorname{argmin}_x F(x)$ is the set of arguments of a real valued function F that makes it attain its (global) minimum.)

Of course, ❶ above is not possible in a strict sense, but the effect of both steps can be realised and yields

a floating point realization of $\star \in \{+,-,\cdot,/\}$.

✎ Notation: write $\widetilde{\star}$ for the floating point realization of $\star \in \{+,-,\cdot,/\}$:

Then ❶ and ❷ may be summed up into

$$\text{For} \quad \star \in \{+,-,\cdot,/\}: \qquad x \,\widetilde{\star}\, y := \mathrm{rd}(x \star y) \,.$$

⌟

Remark 1.5.3.7 (Breakdown of associativity) As a consequence of rounding addition $\widetilde{+}$ and multiplication $\widetilde{*}$ as implemented on computers fail to be associative. They will usually be commutative, though this is not guaranteed. ⌟

§1.5.3.8 (Estimating roundoff errors → [AG11, p. 23]) Let us denote by EPS the largest relative error (→ Def. 1.5.3.3) incurred through rounding:

$$\mathrm{EPS} := \max_{x \in I} \frac{|\mathrm{rd}(x) - x|}{|x|} \,, \tag{1.5.3.9}$$

where $I = [\min|\mathbb{M}|, \max|\mathbb{M}|] \cap \mathbb{M}$ is the range of positive machine numbers.

For machine numbers according to Def. 1.5.2.4 EPS can be computed from the defining parameters B (base) and m (length of mantissa) [AG11, p. 24]:

$$\mathrm{EPS} = \tfrac{1}{2} B^{1-m} \,. \tag{1.5.3.10}$$

⌟

However, when studying roundoff errors, we do not want to delve into the intricacies of the internal representation of machine numbers. This can be avoided by just using a single bound for the relative error due to rounding, and, thus, also for the relative error potentially suffered in each elementary operation.

Assumption 1.5.3.11. "Axiom" of roundoff analysis

There is a small positive number EPS, the machine precision, such that for the elementary arithmetic operations $\star \in \{+,-,\cdot,/\}$ and "hard-wired" functions* $f \in \{\exp, \sin, \cos, \log, \ldots\}$ holds

$$x \,\widetilde{\star}\, y = (x \star y)(1+\delta) \quad , \quad \widetilde{f}(x) = f(x)(1+\delta) \quad \forall x, y \in \mathbb{M} \,,$$

with $|\delta| < \mathrm{EPS}$.

*: this is an ideal, which may not be accomplished even by modern CPUs.

relative roundoff errors of elementary steps in a program bounded by machine precision **!**

EXAMPLE 1.5.3.12 (Machine precision for IEEE standard) C++ tells the machine precision as following:

C++ code 1.5.3.13: Finding out EPS in C++ ⇝ GITLAB

```
2 #include <iostream>
3 #include <limits> // get various properties of arithmetic types
```

```
4 int main() {
5   std::cout.precision(15);
6   std::cout << std::numeric_limits<double>::epsilon() << std::endl;
7 }
```

Output:

```
1 2.22044604925031e-16
```

Knowing the machine precision can be important for checking the validity of computations or coding termination conditions for iterative approximations. ⌟

EXPERIMENT 1.5.3.14 (Adding EPS to 1)

```
cout.precision(25);
const double eps =
   std::numeric_limits<double>::epsilon();
cout << std::fixed << 1.0 + 0.5*eps << endl
      << 1.0 - 0.5*eps << endl
      << (1.0 + 2/eps) - 2/eps << endl;
```

Output:

```
1 1.0000000000000000000000000
2 0.9999999999999998889776975
3 0.0000000000000000000000000
```

In fact, the following "definition" of EPS is sometimes used:

EPS is the smallest positive number $\in \mathbb{M}$ for which $1 \tilde{+} \mathrm{EPS} \neq 1$ (in $\mathbb{M}$):

We find that $1 \tilde{+} \mathrm{EPS} = 1$ actually complies with the "axiom" of roundoff error analysis, Ass. 1.5.3.11:

$$1 = (1+\mathrm{EPS})(1+\delta) \quad \Rightarrow \quad |\delta| = \left|\frac{\mathrm{EPS}}{1+\mathrm{EPS}}\right| < \mathrm{EPS}\,,$$

$$\frac{2}{\mathrm{EPS}} = (1+\frac{2}{\mathrm{EPS}})(1+\delta) \quad \Rightarrow \quad |\delta| = \left|\frac{\mathrm{EPS}}{2+\mathrm{EPS}}\right| < \mathrm{EPS}\,.$$

⌟

Do we have to worry about these tiny roundoff errors **?**

YES (→ Exp. 1.5.1.5):

- accumulation of roundoff errors
- amplification of roundoff errors

Remark 1.5.3.15 (Testing equality with zero)

Since results of numerical computations are almost always polluted by roundoff errors: Tests like `if (x == 0)` are pointless and even dangerous, if x contains the result of a numerical computation.

Remedy: Test `if (abs(x) < tol*s) ...`,
$s \mathrel{\hat{=}}$ positive number, compared to which $|x|$ should be small.
`tol` $\mathrel{\hat{=}}$ "suitable" tolerance, often `tol` $\approx$ EPS

We saw a first example of this practise in Code 1.5.1.3, Line 13.

To motivate this rule, think of a code where the number stored in `x` is a length. One may have chosen μm as reference unit and then `x=0.1`. Another user prefers km as length unit, which means that `x=1.0E-10`

in this case. Just comparing `x` to a fixed threshold will lead to a different behavior of the code depending on the choice of physical units, which is certainly not desirable. In general numerical codes should largely be insensitive to the choice of physical units, a property called **scaling invariance**. From these considerations we also conclude that a guideline for choosing the comparison variable `s` is that it should represent a quantity with the same physical units. ⌟

Remark 1.5.3.16 (Overflow and underflow) Since the set of machine numbers $\mathbb{M}$ is a finite set, the result of an arithmetic operation can lie outside the range covered by it. In this case we have to deal with

overflow $\hat{=}$ $|\text{result of an elementary operation}| > \max\{\mathbb{M}\}$
$\hat{=}$ IEEE standard $\Rightarrow$ `Inf`
underflow $\hat{=}$ $0 < |\text{result of an elementary operation}| < \min\{|\mathbb{M} \setminus \{0\}|\}$
$\hat{=}$ IEEE standard $\Rightarrow$ use non-normalized numbers (!)

The Axiom of roundoff analysis Ass. 1.5.3.11 does *not hold* once non-normalized numbers are encountered:

C++ code 1.5.3.17: Demonstration of over-/underflow ↪ GITLAB

```
2  #include <cmath> //define _USE_MATH_DEFINES to access M_PI
3  #include <iostream>
4  #include <limits>
5
6  using std::cout;
7  using std::numeric_limits;
8  using std::endl;
9
10 int main(){
11     cout.precision(15);
12     const double min = numeric_limits<double>::min();
13     const double res1 = M_PI*min/123456789101112;
14     const double res2 = res1*123456789101112/min;
15     cout << res1 << endl << res2 << endl;
16 }
```

Output:

```
1 5.68175492717434e-322
2 3.15248510554597
```

Try to avoid underflow and overflow

A simple example teaching how to avoid overflow during the computation of the norm of a 2D vector [AG11, Ex. 2.9]:

$$r = \sqrt{x^2 + y^2}$$

straightforward evaluation: overflow, when $|x| > \sqrt{\max|\mathbb{M}|}$ or $|y| > \sqrt{\max|\mathbb{M}|}$.

$$r = \begin{cases} |x|\sqrt{1 + (y/x)^2} & \text{, if } |x| \geq |y| \text{ ,} \\ |y|\sqrt{1 + (x/y)^2} & \text{, if } |y| > |x| \text{ .} \end{cases}$$

➢ no overflow! ⌟

Review question(s) 1.5.3.18 (Machine arithmetic)

(Q1.5.3.18.A) What is the order of magnitude of the machine precision `EPS` for the **double** floating point type?

(Q1.5.3.18.B) What is the "Axiom of roundoff analysis"?

(Q1.5.3.18.C) The set of two-digit decimal numbers is

$$\mathbb{D}_2 := \{\pm 0.xy \cdot 10^e : x,y \in \{0,\ldots,9\},\ x \neq 0,\ e \in \mathbb{Z}\} \cup \{0\} .$$

Give a sharp bound for the relative error of (correct) rounding in $\mathbb{D}_2$.

(Q1.5.3.18.D) Given two variables of type **Eigen::VectorXd**, how can you safely check in a C++ code that the vectors they describe are

- linearly dependent,
- orthogonal?

(Q1.5.3.18.E) [An obfuscated C++ function] What is the purpose of the following function? What is the idea behind Line 19

C++ code 1.5.3.19: ↪ GITLAB

```
2  double fun(double x) {
3    bool neg = false;
4    if (x < 0) {
5      neg = true;
6      x = -x;
7    }
8    unsigned int f = 0;
9    while (x > 1) {
10     f++;
11     x /= 2.0;
12   }
13   double v = 1.0 + x;
14   double num = x;
15   double den = 1.0;
16   for (int i = 2; true; ++i) {
17     const double s = (num *= x) / (den *= i);
18     if (s == 0.0) {
19       break; //
20     }
21     v += s;
22   }
23   while (f-- > 0) {
24     v = v * v;
25   }
26   if (neg) {
27     v = 1.0 / v;
28   }
29   return v;
30 }
```

△

1.5.4 Cancellation

Video tutorial for Section 1.5.4 "Cancellation": (22 minutes) Download link, tablet notes

→ review questions 1.5.4.33

In general, predicting the impact of roundoff errors on the result of a multi-stage computation is very difficult, if possible at all. However, there is a constellation that is particularly prone to dangerous *amplification of roundoff errors* and still can be detected easily.

EXAMPLE 1.5.4.1 (Computing the zeros of a quadratic polynomial) The following simple EIGEN code computes the real roots of a quadratic polynomial $p(\xi) = \xi^2 + \alpha\xi + \beta$ by the discriminant formula

$$p(\xi_1) = p(\xi_2) = 0\,, \quad \xi_{1/2} = \frac{1}{2}\Big(-\alpha \pm \sqrt{D}\Big)\,, \text{ if } \quad D := \alpha^2 - 4\beta \geq 0\,. \tag{1.5.4.2}$$

C++ code 1.5.4.3: Discriminant formula for the real roots of $p(\xi) = \xi^2 + \alpha\xi + \beta$ ➙ GITLAB

```
//!  C++ function computing the zeros of a quadratic polynomial
//!  ξ → ξ^2 + αξ + β by means of the familiar discriminant
//!  formula ξ_{1,2} = ½(−α ± √(α^2 − 4β)).  However
//!  this implementation is vulnerable to round-off!  The zeros are
//!  returned in a column vector
inline Vector2d zerosquadpol(double alpha, double beta) {
  Vector2d z;
  const double D = std::pow(alpha, 2) - 4 * beta; // discriminant
  if (D >= 0) {
    // The famous discriminant formula
    const double wD = std::sqrt(D);
    z << (-alpha - wD) / 2, (-alpha + wD) / 2; //
  }
  else {
    throw std::runtime_error( "no real zeros" );
  }
  return z;
}
```

This formula is applied to the quadratic polynomial $p(\xi) = (\xi - \gamma)(\xi - \frac{1}{\gamma})$ after its coefficients α, β have been computed from γ, which will have introduced *small* relative roundoff errors (of size EPS).

C++ code 1.5.4.4: Testing the accuracy of computed roots of a quadratic polynomial ➙ GITLAB

```
//!  Eigen Function for testing the computation of the zeros of a
    parabola
void compzeros() {
  int n = 100;
  MatrixXd res(n, 4);
  VectorXd gamma = VectorXd::LinSpaced(n, 2, 992);
  for (int i = 0; i < n; ++i) {
    double alpha = -(gamma(i) + 1. / gamma(i));
    double beta = 1.;
    Vector2d z1 = zerosquadpol(alpha, beta);
    Vector2d z2 = zerosquadpolstab(alpha, beta);
    double ztrue = 1. / gamma(i), z2true = gamma(i);
    res(i, 0) = gamma(i);
    res(i, 1) = std::abs((z1(0) - ztrue) / ztrue);
    res(i, 2) = std::abs((z2(0) - ztrue) / ztrue);
    res(i, 3) = std::abs((z1(1) - z2true) / z2true);
  }
```

Plot of relative errors Def. 1.5.3.3 ▷

We observe that roundoff incurred during the computation of α and β leads to "wrong" roots.
For large γ the computed small root may be fairly inaccurate as regards its *relative error*, which can be several orders of magnitude larger than machine precision EPS.

The large root always enjoys a small relative error about the size of EPS.

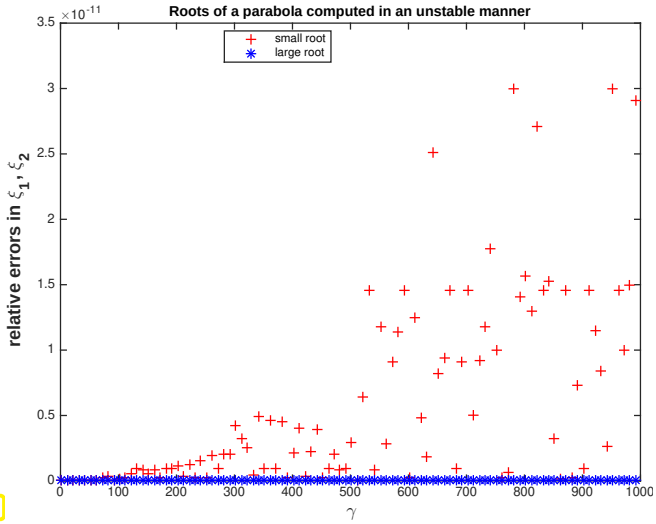


Fig. 21

In order to understand why the small root is much more severely affected by roundoff, note that its computation involves the subtraction of two large numbers, if γ is large. This is the typical situation, in which cancellation occurs. ⌟

§1.5.4.5 (Visualisation of cancellation effect) We look at the *exact* subtraction of two almost equal positive numbers both of which have small relative errors (red boxes) with respect to some desired exact value (indicated by blue boxes). The result of the subtraction will be small, but the errors may add up during the subtraction, ultimately constituting a large fraction of the result.

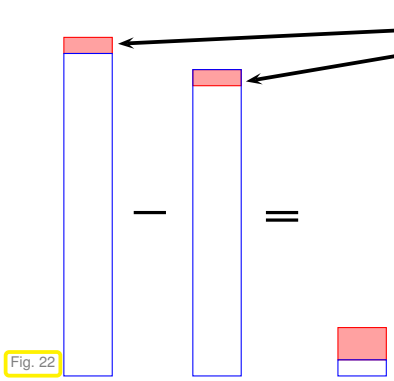


Fig. 22

Cancellation

≙ Subtraction of almost equal numbers

(➤ extreme *amplification* of *relative* errors)

(◁ Roundoff error introduced by subtraction itself is negligible.)

⌟

EXAMPLE 1.5.4.6 (Cancellation in decimal system) We consider two positive numbers x, y of about the same size afflicted with relative errors $\approx 10^{-7}$. This means that their seventh decimal digits are perturbed, here indicated by $*$. When we subtract the two numbers the perturbed digits are shifted to the left, resulting in a possible relative error of $\approx 10^{-3}$:

$$\begin{array}{rcll} x & = & 0.123467* & \leftarrow \text{7th digit perturbed} \\ y & = & 0.123456* & \leftarrow \text{7th digit perturbed} \\ \hline x-y & = & 0.000011* = 0.11*000 \cdot 10^{-4} & \leftarrow \text{3rd digit perturbed} \end{array}$$

padded zeroes

Again, this example demonstrates that cancellation wreaks havoc through error amplification, not through the roundoff error due to the subtraction. ⌟

EXAMPLE 1.5.4.7 (Cancellation when evaluating difference quotients → [DR08, Sect. 8.2.6], [AG11, Ex. 1.3]) From analysis we know that the derivative of a differentiable function $f : I \subset \mathbb{R} \to \mathbb{R}$ at a point

$x \in I$ is the limit of a difference quotient

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} .$$

This suggests the following approximation of the derivative by a difference quotient with small but finite $h > 0$

$$f'(x) \approx \frac{f(x+h) - f(x)}{h} \quad \text{for} \quad |h| \ll 1 .$$

Results from analysis tell us that the approximation error should tend to zero for $h \to 0$. More precise quantitative information is provided by the Taylor formula for a twice continuously differentiable function [AG11, p. 5]

$$f(x+h) = f(x) + f'(x)h + \tfrac{1}{2} f''(\xi) h^2 \quad \text{for some} \quad \xi = \xi(x,h) \in [\min\{x, x+h\}, \max\{x, x+h\}] , \tag{1.5.4.8}$$

from which we infer

$$\frac{f(x+h) - f(x)}{h} - f'(x) = \tfrac{1}{2} h f''(\xi) \quad \text{for some} \quad \xi = \xi(x,h) \in [\min\{x, x+h\}, \max\{x, x+h\}] . \tag{1.5.4.9}$$

We investigate the approximation of the derivative by difference quotients for $f = \exp$, $x = 0$, and different values of $h > 0$:

C++ code 1.5.4.10: Difference quotient approximation of the derivative of exp ➙ GITLAB

```
// Difference quotient approximation
// of the derivative of exp
void diffq() {
  double h = 0.1;
  const double x = 0.0;
  for (int i = 1; i <= 16; ++i) {
    const double df = (exp(x + h) - exp(x)) / h;
    cout << setprecision(14) << fixed;
    cout << setw(5) << -i << setw(20) << abs(df
        - 1) << endl;
    h /= 10;
  }
}
```

Measured relative errors ▷

$\log_{10}(h)$	relative error
-1	0.05170918075648
-2	0.00501670841679
-3	0.00050016670838
-4	0.00005000166714
-5	0.00000500000696
-6	0.00000049996218
-7	0.00000004943368
-8	0.00000000607747
-9	0.00000008274037
-10	0.00000008274037
-11	0.00000008274037
-12	0.00008890058234
-13	0.00079927783736
-14	0.00079927783736
-15	0.11022302462516
-16	1.00000000000000

We observe an initial decrease of the relative approximation error followed by a steep increase when h drops below 10^{-8}.

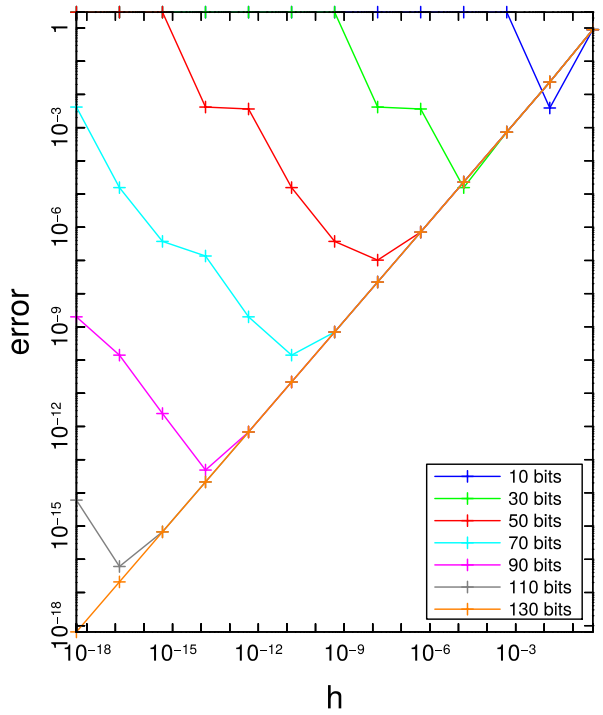


Fig. 23

That the observed errors are really due to round-off errors is confirmed by the numerical results reported besides, using a variable precision floating point module of EIGEN, the MPFRC++ Support module, which is no longer available now.

The C++ used to generate these results can be found in ➙ GITLAB.

Obvious culprit for what we see in Fig. 23: cancellation when computing the numerator of the difference quotient for small $|h|$ leads to a strong amplification of inevitable errors introduced by the evaluation of the transcendent exponential function.

We witness the competition of two opposite effects: Smaller h results in a better approximation of the derivative by the difference quotient, but the impact of cancellation is the stronger the smaller $|h|$.

$$\left.\begin{array}{l}\text{Approximation error } f'(x) - \dfrac{f(x+h)-f(x)}{h} \to 0 \\ \text{Impact of roundoff} \to \infty\end{array}\right\} \text{ as } h \to 0\,.$$

In order to provide a rigorous underpinning for our conjecture, in this example we embark on our first roundoff error analysis merely based on the "Axiom of roundoff analysis" Ass. 1.5.3.11: As in the computational example above we study the approximation of $f'(x) = e^x$ for $f = \exp$, $x \in \mathbb{R}$.

$$\begin{aligned}\texttt{df} &= \frac{e^{x+h}(1+\delta_1) - e^x(1+\delta_2)}{h} \\ &= e^x\left(\frac{e^h-1}{h} + \frac{\delta_1 e^h - \delta_2}{h}\right) \\ \Rightarrow\ |\texttt{df}| &\le e^x\left(\frac{e^h-1}{h} + \texttt{eps}\,\frac{1+e^h}{h}\right)\end{aligned}$$

correction factors take into account roundoff:
(→ "'axiom of roundoff analysis", Ass. 1.5.3.11)

$$|\delta_1|, |\delta_2| \le \texttt{eps}\,.$$

$1 + O(h)$ $\quad O(h^{-1})$ for $h \to 0$

(Note that the estimate for the term $(e^h - 1)/h$ is a particular case of (1.5.4.9).)

relative error: $\left|\dfrac{e^x - \texttt{df}}{e^x}\right| \approx h + \dfrac{2\texttt{eps}}{h} \to \min$ for $h = \sqrt{2\,\texttt{eps}}$.

In double precision: $\sqrt{2\mathrm{eps}} = 2.107342425544702 \cdot 10^{-8}$ ⌟

Remark 1.5.4.11 (Cancellation during the computation of relative errors) In the numerical experiment of Ex. 1.5.4.7 we computed the relative error of the result by subtraction, see Code 1.5.4.10. Of course, massive cancellation will occur! Do we have to worry?

In this case cancellation can be tolerated, because we are *interested only* in the magnitude of the relative error. Even if it was affected itself by a large relative error, this information is still not compromised.

For example, if the relative error has the exact value 10^{-8}, but can be computed only with a huge relative error of 10%, then the perturbed value would still be in the range $[0.9 \cdot 10^{-8}, 1.1 \cdot 10^{-8}]$. Therefore it will still have the correct magnitude and still permit us to conclude the number of valid digits correctly. ⌟

Remark 1.5.4.12 (Cancellation in Gram-Schmidt orthogonalisation of Exp. 1.5.1.5) The Hilbert matrix $\mathbf{A} \in \mathbb{R}^{10,10}$, $(\mathbf{A})_{i,j} = (i+j-1)^{-1}$, considered in Exp. 1.5.1.5 has columns that are almost linearly dependent.

Cancellation when computing orthogonal projection of vector $\mathbf{a}$ onto space spanned by vector $\mathbf{b}$ ▷

$$\mathbf{p} = \mathbf{a} - \frac{\mathbf{a} \cdot \mathbf{b}}{\mathbf{b} \cdot \mathbf{b}} \mathbf{b} \, .$$

If $\mathbf{a}$, $\mathbf{b}$ point in almost the same direction, $\|\mathbf{p}\| \ll \|\mathbf{a}\|, \|\mathbf{b}\|$, so that a "tiny" vector $\mathbf{p}$ is obtained by subtracting two "long" vectors, which implies cancellation.

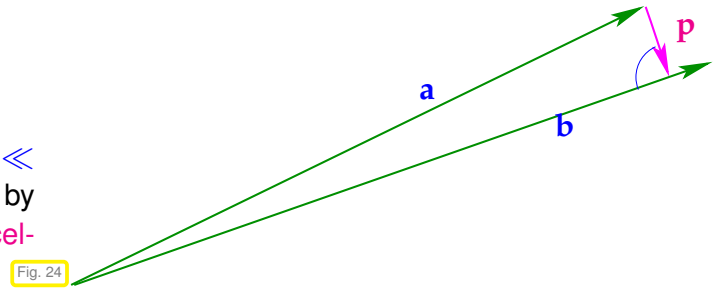


Fig. 24

This can happen in Line 10 of Code 1.5.1.3.
⌟

EXAMPLE 1.5.4.13 (Cancellation: roundoff error analysis) We consider a simple arithmetic expression written in two ways:

$$a^2 - b^2 = (a+b)(a-b) \, , \quad a, b \in \mathbb{R}; .$$

We evaluate this term by means of two algebraically equivalent algorithms for the input data $a = 1.3$, $b = 1.2$ in 2-digit decimal arithmetic with standard rounding. ("Algebraically equivalent" means that two algorithms will produce the same results in the absence of roundoff errors.

Algorithm A	Algorithm B
$x := a \tilde{\cdot} a = 1.7$ (rounded)	$x := a \tilde{+} b = 2.5$ (exact)
$y := b \tilde{\cdot} b = 1.4$ (rounded)	$y := a \tilde{-} b = 0.1$ (exact)
$x \tilde{-} y = 0.30$ (exact)	$x * y = 0.25$ (exact)

Algorithm B produces the exact result, whereas Algorithm A fails to do so. Is this pure coincidence or an indication of the superiority of algorithm B? This question can be answered by roundoff error analysis. We demonstrate the approach for the two algorithms A & B and general input $a, b \in \mathbb{R}$.

Roundoff error analysis heavily relies on Ass. 1.5.3.11 and dropping terms of "higher order" in the machine precision, that is terms that behave like $O(\mathrm{EPS}^q)$, $q > 1$. It involves introducing the relative roundoff error for every elementary operation through a factor $(1+\delta)$, $|\delta| \leq \mathrm{EPS}$.

Algorithm A:

$$x = a^2(1+\delta_1) \, , y = b^2(1+\delta_2)$$

$$\widetilde{f} = (a^2(1+\delta_1) - b^2(1+\delta_2))(1+\delta_3) = f + a^2\delta_1 - b^2\delta_2 + (a^2-b^2)\delta_3 + O(\mathrm{EPS}^2)$$

$$\frac{|\widetilde{f}-f|}{|f|} \leq \mathrm{EPS}\frac{a^2+b^2+|a^2-b^2|}{|a^2-b^2|} + O(\mathrm{EPS}^2) = \mathrm{EPS}\left(1+\frac{|a^2+b^2|}{|a^2-b^2|}\right) + O(\mathrm{EPS}^2)\,. \tag{1.5.4.14}$$

will be neglected

For $a \approx b$ the relative error of the result of Algorithm A will be much larger than the machine precision EPS. This reflects cancellation in the last subtraction step.

Algorithm B:

$$x = (a+b)(1+\delta_1)\,,\ y = (a-b)(1+\delta_2)$$
$$\widetilde{f} = (a+b)(a-b)(1+\delta_1)(1+\delta_2)(1+\delta_3) = f + (a^2-b^2)(\delta_1+\delta_2+\delta_3) + O(\mathrm{EPS}^2)$$

$$\frac{|\widetilde{f}-f|}{|f|} \leq |\delta_1+\delta_2+\delta_3| + O(\mathrm{EPS}^2) \leq 3\mathrm{EPS} + O(\mathrm{EPS}^2)\,. \tag{1.5.4.15}$$

Relative error of the result of Algorithm B is always $\approx$ EPS **!**

In this example we see a general guideline at work:

> If inevitable, subtractions prone to cancellation should be done as early as possible.

The reason is that input data and and initial intermediate results are usually not as much tainted by roundoff errors as numbers computed after many steps. ⌟

§1.5.4.16 (Avoiding disastrous cancellation) The following examples demonstrate a few fundamental techniques for steering clear of cancellation by using alternative formulas that yield the same value (in exact arithmetic), but do not entail subtracting two numbers of almost equal size.

EXAMPLE 1.5.4.17 (Stable discriminant formula → Ex. 1.5.4.1, [AG11, Ex. 2.10]) If ξ_1 and ξ_2 are the two roots of the quadratic polynomial $p(\xi) = \xi^2 + \alpha\xi + \beta$, then $\xi_1 \cdot \xi_2 = \beta$ (Vieta's formula). Thus once we have computed a root, we can obtain the other by simple division.

Idea:

1. Depending on the sign of α compute "stable root" without cancellation.
2. Compute other root from Vieta's formula (avoiding subtraction)

C++ code 1.5.4.18: Stable computation of real roots of a quadratic polynomial ↪ GITLAB

```
2 //!  C++ function computing the zeros of a quadratic polynomial
3 //!  ξ → ξ^2 + αξ + β by means of the familiar discriminant
4 //!  formula ξ_{1,2} = ½(−α ± √(α^2 − 4β)).
5 //!  This is a stable implementation based on Vieta's theorem.
6 //!  The zeros are returned in a column vector
7 Eigen::VectorXd zerosquadpolstab(double alpha, double beta) {
8   Eigen::Vector2d z(2);
9   const double D = std::pow(alpha, 2) - 4 * beta; // discriminant
```

```
10   if (D >= 0) {
11     const double wD = std::sqrt(D);
12     // Use discriminant formula only for zero far away from 0
13     // in order to avoid cancellation.  For the other zero
14     // use Vieta's formula.
15     if (alpha >= 0) {
16       const double t = 0.5 * (-alpha - wD); //
17       z << t, beta / t;
18     } else {
19       const double t = 0.5 * (-alpha + wD); //
20       z << beta / t, t;
21     }
22   }
23   else {
24     throw std::runtime_error( "no real zeros" );
25   }
26   return z;
27 }
```

➥ Invariably, we add numbers with the same sign in Line 16 and Line 19.

Numerical experiment based on the driver code Code 1.5.4.4.

Observation:

The new code can also compute the small root of the polynomial $p(\xi) = (\xi - \gamma)(\xi - \frac{1}{\gamma})$ (expanded in monomials) with a relative error $\approx$ EPS.

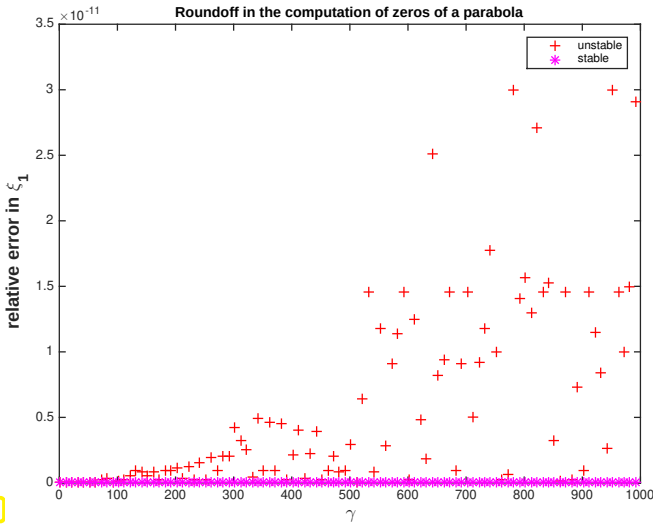


Fig. 25

⌟

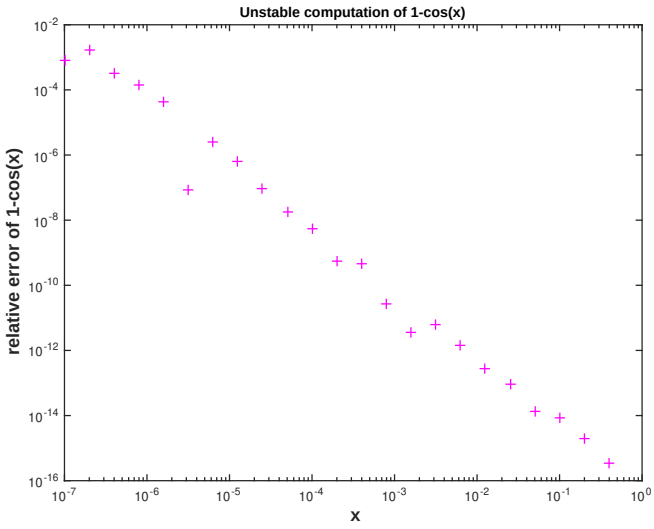
EXAMPLE 1.5.4.19 (Exploiting trigonometric identities to avoid cancellation) The task is to evaluate the integral

$$\int_0^x \sin t \, \mathrm{d}t = \underbrace{1 - \cos x}_{I} = \underbrace{2\sin^2(x/2)}_{II} \quad \text{for} \quad 0 < x \ll 1 \,, \tag{1.5.4.20}$$

and this can be done by the two different formulas I and II.

Relative error of expression I (`1-cos(x)`) with respect to equivalent expression II (`2*sin(x/2)^{2}`) ▷

Expression I is affected by cancellation for $|x| \ll 1$, since then $\cos x \approx 1$, whereas expression II can be evaluated with a relative error $\approx$ EPS for all x.



Fi

⌟

Analytic manipulations offer ample opportunity to rewrite expressions in equivalent form immune to cancellation.

EXAMPLE 1.5.4.21 (Switching to equivalent formulas to avoid cancellation)

Now we see an example of a computation allegedly dating back to Archimedes, who tried to approximate the area of a circle by the areas of inscribed regular polygons.

Approximation of a circle by a regular n-gon, $n \in \mathbb{N}$ ▷

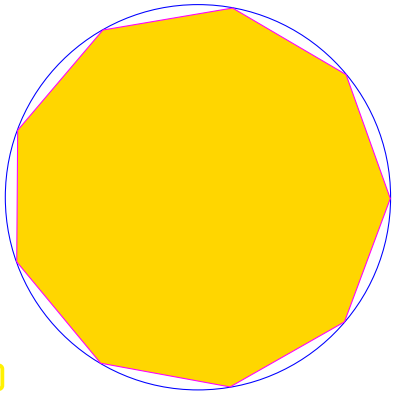


Fig. 27

We focus on the unit circle. The area of the inscribed n-gon is

$$A_n = n \cos\frac{\alpha_n}{2} \sin\frac{\alpha_n}{2} = \frac{n}{2}\sin\alpha_n = \frac{n}{2}\sin\left(\frac{2\pi}{n}\right).$$

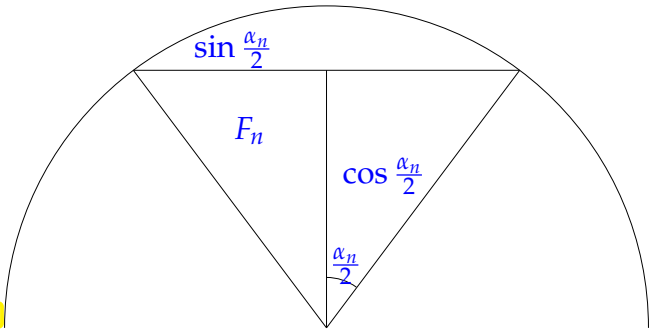


Fig. 28

Recursion formula for A_n derived from

$$\sin\frac{\alpha_n}{2} = \sqrt{\frac{1-\cos\alpha_n}{2}} = \sqrt{\frac{1-\sqrt{1-\sin^2\alpha_n}}{2}},$$

Initial approximation: $A_6 = \frac{3}{2}\sqrt{3}$.

C++ code 1.5.4.22: Tentative computation of circumference of regular polygon ➙ GITLAB

```
2 //!  Approximation of Pi by approximating the circumference of a
3 //!  regular polygon
4 MatrixXd ApproxPlinstable(double tol = 1e-8, unsigned int maxIt = 50){
```

```
5   double s=sqrt(3)/2.;
6   double An=3.*s;// initialization (hexagon case)
7   unsigned int n = 6;
8   unsigned int it = 0;
9   MatrixXd res(maxIt,4); // matrix for storing results
10  res(it,0) = n; res(it,1) = An;
11  res(it,2) = An - M_PI; res(it,3)=s;
12  while( it < maxIt && s > tol ){// terminate when s is 'small enough'
13     s = sqrt((1.- sqrt(1.-s*s))/2.);// recursion for area
14     n *= 2; An = n/2.*s;     // new estimate for circumference
15     ++it;
16     res(it,0) =n; res(it,1) =An;// store results and (absolute) error
17     res(it,2) = An - M_PI; res(it,3)=s;
18  }
19  return res.topRows(it);
20 }
```

The approximation deteriorates after applying the recursion formula many times:

n	A_n	$A_n - \pi$	$\sin \alpha_n$
6	2.598076211353316	-0.543516442236477	0.866025403784439
12	3.000000000000000	-0.141592653589794	0.500000000000000
24	3.105828541230250	-0.035764112359543	0.258819045102521
48	3.132628613281237	-0.008964040308556	0.130526192220052
96	3.139350203046872	-0.002242450542921	0.065403129230143
192	3.141031950890530	-0.000560702699263	0.032719082821776
384	3.141452472285344	-0.000140181304449	0.016361731626486
768	3.141557607911622	-0.000035045678171	0.008181139603937
1536	3.141583892148936	-0.00000876144085 7	0.004090604026236
3072	3.141590463236762	-0.00000219035303 1	0.002045306291170
6144	3.141592106043048	-0.00000054754674 5	0.001022653680353
12288	3.141592516588155	-0.00000013700163 8	0.000511326906997
24576	3.141592618640789	-0.00000003494900 4	0.000255663461803
49152	3.141592645321216	-0.00000000826857 7	0.000127831731987
98304	3.141592645321216	-0.00000000826857 7	0.000063915865994
196608	3.141592645321216	-0.00000000826857 7	0.000031957932997
393216	3.141592645321216	-0.00000000826857 7	0.000015978966498
786432	3.141593669849427	0.000001016259634	0.000007989485855
1572864	3.141592303811738	-0.000000349778055	0.000003994741190
3145728	3.141608696224804	0.000016042635011	0.000001997381017
6291456	3.141586839655041	-0.000005813934752	0.000000998683561
12582912	3.141674265021758	0.000081611431964	0.000000499355676
25165824	3.141674265021758	0.000081611431964	0.000000249677838
50331648	3.143072740170040	0.001480086580246	0.000000124894489
100663296	3.159806164941135	0.018213511351342	0.000000062779708
201326592	3.181980515339464	0.040387861749671	0.000000031610136
402653184	3.354101966249685	0.212509312659892	0.000000016660005
805306368	4.242640687119286	1.101048033529493	0.000000010536712
1610612736	6.000000000000000	2.858407346410207	0.000000007450581

Where does cancellation occur in Line 13 of Code 1.5.4.22? Since $s \ll 1$, computing $1 - s$ will not trigger cancellation. However, the subtraction $1 - \sqrt{1 - s^2}$ will, because $\sqrt{1 - s^2} \approx 1$ for $s \ll 1$:

$$\sin\frac{\alpha_n}{2} = \sqrt{\frac{1-\cos\alpha_n}{2}} = \sqrt{\frac{1-\sqrt{1-\sin^2\alpha_n}}{2}}$$

For $\alpha_n \ll 1$: $\sqrt{1-\sin^2\alpha_n} \approx 1$

Cancellation here!

We arrive at an *equivalent* formula not vulnerable to cancellation essentially using the identity $(a+b)(a-b) = a^2 - b^2$ in order to eliminate the difference of square roots in the numerator.

$$\begin{aligned}\sin\frac{\alpha_n}{2} &= \sqrt{\frac{1-\sqrt{1-\sin^2\alpha_n}}{2}} = \sqrt{\frac{1-\sqrt{1-\sin^2\alpha_n}}{2}\cdot\frac{1+\sqrt{1-\sin^2\alpha_n}}{1+\sqrt{1-\sin^2\alpha_n}}} \\ &= \sqrt{\frac{1-(1-\sin^2\alpha_n)}{2(1+\sqrt{1-\sin^2\alpha_n})}} = \frac{\sin\alpha_n}{\sqrt{2\left(1+\sqrt{1-\sin^2\alpha_n}\right)}}\,.\end{aligned}$$

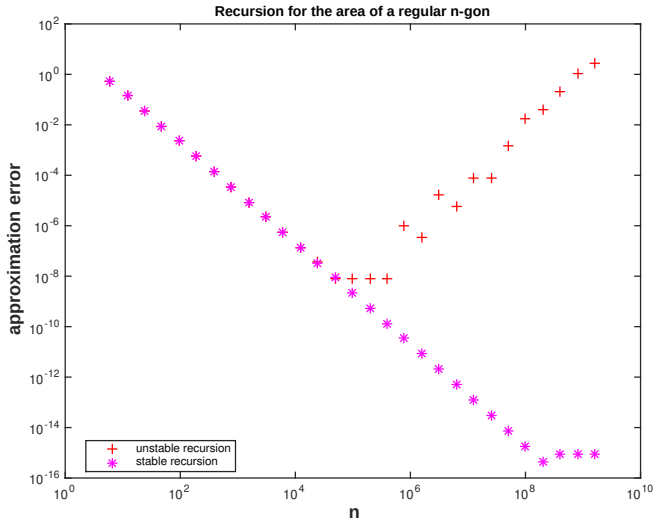
C++ code 1.5.4.23: Stable recursion for area of regular n-gon ↪ GITLAB

```
//!  Approximation of Pi by approximating the circumference of a
//!  regular polygon
MatrixXd apprpistable(double tol = 1e-8, unsigned int maxIt = 50){
  double s=sqrt(3)/2.; double An=3.*s;// initialization (hexagon case)
  unsigned int n = 6;
  unsigned int it = 0;
  MatrixXd res(maxIt,4); // matrix for storing results
  res(it,0) = n; res(it,1) = An;
  res(it,2) = An - M_PI; res(it,3)=s;
  while( it < maxIt && s > tol ){// terminate when s is 'small enough'
    s = s/sqrt(2*(1+sqrt((1+s)*(1-s))));// Stable recursion without
       cancellation
    n *= 2; An = n/2.*s;     // new estimate for circumference
    ++it;
    res(it,0) =n; res(it,1) =An;// store results and (absolute) error
    res(it,2) = An - M_PI; res(it,3)=s;
  }
  return res.topRows(it);
}
```

Using the stable recursion, we observe better approximation for polygons with more corners:

n	A_n	$A_n - \pi$	$\sin \alpha_n$
6	2.598076211353316	-0.543516442236477	0.866025403784439
12	3.000000000000000	-0.141592653589793	0.500000000000000
24	3.105828541230249	-0.035764112359544	0.258819045102521
48	3.132628613281238	-0.008964040308555	0.130526192220052
96	3.139350203046867	-0.002242450542926	0.065403129230143
192	3.141031950890509	-0.000560702699284	0.032719082821776
384	3.141452472285462	-0.000140181304332	0.016361731626487
768	3.141557607911857	-0.000035045677936	0.008181139603937
1536	3.141583892148318	-0.000008761441475	0.004090604026235
3072	3.141590463228050	-0.000002190361744	0.002045306291164
6144	3.141592105999271	-0.000000547590522	0.001022653680338
12288	3.141592516692156	-0.000000136897637	0.000511326907014
24576	3.141592619365383	-0.000000034224410	0.000255663461862
49152	3.141592645033690	-0.000000008556103	0.000127831731976
98304	3.141592651450766	-0.000000002139027	0.000063915866118
196608	3.141592653055036	-0.000000000534757	0.000031957933076
393216	3.141592653456104	-0.000000000133690	0.000015978966540
786432	3.141592653556371	-0.000000000033422	0.000007989483270
1572864	3.141592653581438	-0.000000000008355	0.000003994741635
3145728	3.141592653587705	-0.000000000002089	0.000001997370818
6291456	3.141592653589271	-0.000000000000522	0.000000998685409
12582912	3.141592653589663	-0.000000000000130	0.000000499342704
25165824	3.141592653589761	-0.000000000000032	0.000000249671352
50331648	3.141592653589786	-0.000000000000008	0.000000124835676
100663296	3.141592653589791	-0.000000000000002	0.000000062417838
201326592	3.141592653589794	0.000000000000000	0.000000031208919
402653184	3.141592653589794	0.000000000000001	0.000000015604460
805306368	3.141592653589794	0.000000000000001	0.000000007802230
1610612736	3.141592653589794	0.000000000000001	0.000000003901115

Plot of errors for approximations of π as computed by the two algebraically equivalent recursion formulas▷



Observation, *cf.* Ex. 1.5.4.7

Amplified roundoff errors due to cancellation supersedes approximation error for $n \geq 10^5$.

Roundoff errors merely of magnitude EPS in the case of stable recursion

⌟

EXAMPLE 1.5.4.24 (Summation of exponential series)

In principle, the function value $\exp(x)$ can be approximated up to any accuracy by summing sufficiently many terms of the globally convergent exponential series.

$$\exp(x) = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$
$$= 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \dots .$$

C++ code 1.5.4.25: Summation of exponential series ↪ GITLAB

```
2  double expeval(double x,
3                 double tol=1e-8){
4    // Initialization
5    double y = 1.0;
6    double term = 1.0;
7    uint64_t k = 1;
8    // Termination criterion
9    while(abs(term) > tol*y) {
10     term *= x/static_cast<double>(k);   // next summand
11     y += term; // Summation
12     ++k;
13   }
14   return y;
15 }
```

Results for $\texttt{tol} = 10^{-8}$, $\widetilde{\exp}$ designates the approximate value for $\exp(x)$ returned by the function from Code 1.5.4.25. Rightmost column lists relative errors, which tells us the number of valid digits in the approximate result.

x	Approximation $\widetilde{\exp}(x)$	$\exp(x)$	$\frac{\lvert\exp(x)-\widetilde{\exp}(x)\rvert}{\exp(x)}$
-20	6.1475618242e-09	2.0611536224e-09	1.982583033727893
-18	1.5983720359e-08	1.5229979745e-08	0.049490585500089
-16	1.1247503300e-07	1.1253517472e-07	0.000534425951530
-14	8.3154417874e-07	8.3152871910e-07	0.000018591829627
-12	6.1442105142e-06	6.1442123533e-06	0.000000299321453
-10	4.5399929604e-05	4.5399929762e-05	0.000000003501044
-8	3.3546262812e-04	3.3546262790e-04	0.000000000662004
-6	2.4787521758e-03	2.4787521767e-03	0.000000000332519
-4	1.8315638879e-02	1.8315638889e-02	0.000000000530724
-2	1.3533528320e-01	1.3533528324e-01	0.000000000273603
0	1.0000000000e+00	1.0000000000e+00	0.000000000000000
2	7.3890560954e+00	7.3890560989e+00	0.000000000479969
4	5.4598149928e+01	5.4598150033e+01	0.000000001923058
6	4.0342879295e+02	4.0342879349e+02	0.000000001344248
8	2.9809579808e+03	2.9809579870e+03	0.000000002102584
10	2.2026465748e+04	2.2026465795e+04	0.000000002143799
12	1.6275479114e+05	1.6275479142e+05	0.000000001723845
14	1.2026042798e+06	1.2026042842e+06	0.000000003634135
16	8.8861105010e+06	8.8861105205e+06	0.000000002197990
18	6.5659968911e+07	6.5659969137e+07	0.000000003450972
20	4.8516519307e+08	4.8516519541e+08	0.000000004828737

Observation:

Large relative approximation errors for $x \ll 0$.

For $x \ll 0$ we have $\exp(x)| \ll 1$, but this value is computed by summing large numbers of opposite sign.

Terms summed up for $x = -20$ ▷

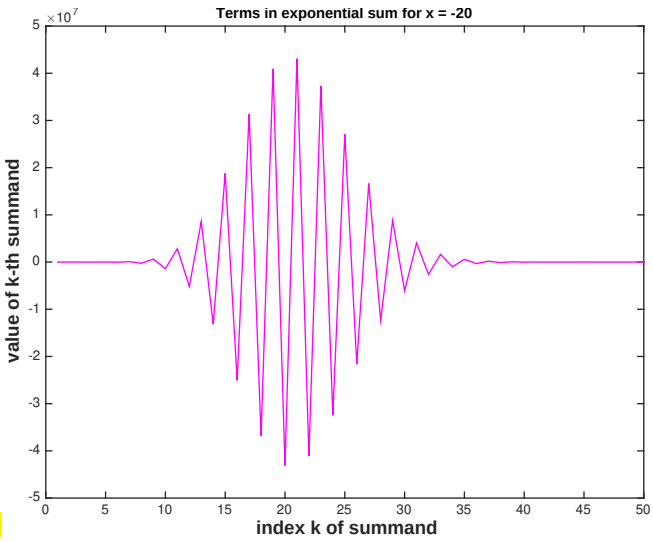


Fig. 30

Remedy: Cancellation can be avoided by using identity

$$\exp(x) = \frac{1}{\exp(-x)} \quad \text{, if} \quad x < 0 \,.$$

EXAMPLE 1.5.4.26 (Trade cancellation for approximation) In a computer code we have to provide a routine for the evaluation of the "hidden difference quotient"

$$I(a) := \int_0^1 e^{at}\,\mathrm{d}t = \frac{\exp(a)-1}{a} \quad \text{for any} \quad a > 0\,, \tag{1.5.4.27}$$

cf. the discussion of cancellation in the context of numerical differentiation in Ex. 1.5.4.7. There we observed massive cancellation.

Trick. Recall the **Taylor expansion** formula in one dimension for a function that is $m+1$ times continuously differentiable in a neighborhood of x_0 [Str09, Satz 5.5.1]

$$f(x_0+h) = \sum_{k=0}^{m} \frac{1}{k!} f^{(k)}(x_0)h^k + R_m(x_0,h)\,, \quad R_m(x_0,h) = \frac{1}{(m+1)!} f^{(m+1)}(\xi)h^{m+1}\,, \tag{1.5.4.28}$$

for some $\xi \in [\min\{x_0, x_0+h\}, \max\{x_0, x_0+h\}]$, and for all sufficiently small $|h|$. Here $R(x_0,h)$ is called the remainder term and $f^{(k)}$ denotes the k-th derivative of f.

Cancellation in (1.5.4.27) can be avoided by replacing $\exp(a)$, $a > 0$, with a suitable Taylor expansion of $a \mapsto e^a$ around $a = 0$ and then dividing by a:

$$\frac{\exp(a)-1}{a} = \sum_{k=0}^{m} \frac{1}{(k+1)!} a^k + R_m(a)\,, \quad R_m(a) = \frac{1}{(m+1)!}\exp(\xi)a^m \text{ for some } 0 \le \xi \le a\,.$$

Then use as an approximation the point value of the Taylor polynomial

$$I(a) \approx \widetilde{I}_m(a) := \sum_{k=0}^{m} \frac{1}{(k+1)!} a^k\,.$$

For a similar discussion see [AG11, Ex. 2.12].

Issue: A finite Taylor sum usually offers only an approximation and we incur an approximation error. This begs the question how to choose the number m of terms to be retained in the Taylor expansion. We have to pick m large enough such that the relative approximation error remains below a prescribed threshold

`tol`. To *estimate* the relative approximation error, we use the expression for the remainder together with the simple estimate $(\exp(a)-1)/a > 1$ for all $a > 0$:

$$\text{rel. err.} = \frac{\left|I(a) - \widetilde{I}_m(a)\right|}{|I(a)|} = \frac{(e^a - 1)/a - \sum_{k=0}^{m} \frac{1}{(k+1)!} a^k}{(e^a-1)/a}$$
$$\leq \frac{1}{(m+1)!}\exp(\xi)a^m \leq \frac{1}{(m+1)!}\exp(a)a^m \,.$$

For $a = 10^{-3}$ we get

m	1	2	3	4	5
	1.0010e-03	5.0050e-07	1.6683e-10	4.1708e-14	8.3417e-18

Hence, keeping $m = 3$ terms is enough for achieving about 10 valid digits.

Relative error of unstable formula `(exp(a)-1.0)/a` and relative error, when using a Taylor expansion approximation for small a ▷

```
if (abs(a) < 1E-3)
  v = 1.0 + (1.0/2 + 1.0/6*a)*a;
else
  v = (exp(a)-1.0)/a;
end
```

Error computed by comparison with the PYTHON library function `numpy.expm1()` that provides a stable implementation of $\exp(x) - 1$.

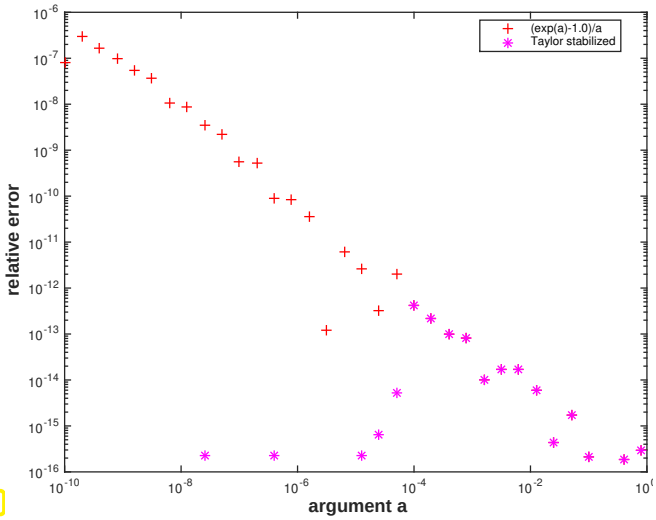


Fig. 31

⌟

EXAMPLE 1.5.4.29 (Complex step differentiation [LM67]) This is a technique from complex analysis that can be applied to real-valued analytic functions. Let $f : I \to \mathbb{R}$, $I \subset \mathbb{R}$ an interval, be analytic in a neighborhood of $x_0 \in I$, which means that it can be written as a convergent power series there, see also Def. 6.2.2.48:

$$f(x) = \sum_{j=0}^{\infty} a_j (x - x_0)^j \quad \forall x : |x - x_0| < \rho \quad \text{and some} \quad \rho > 0,\ a_j \in \mathbb{R} \,. \tag{1.5.4.30}$$

Note that f is infinitely many times differentiable in a neighborhood of x_0 and that its derivatives satisfy $f^{(n)}(x_0) = n!a_n \in \mathbb{R}$, $n \in \mathbb{N}_0$.

A power series like in (6.2.2.36) also makes sense for $x \in \mathbb{C}$, $|x - x_0| < \rho$! Thus, for $0 < h < \rho$ we can approximate f in a neighborhood of x_0 by means of a *complex* Taylor polynomial

$$f(x_0 + \imath h) = f(x_0) + f'(x_0)\imath h - f''(x_0)h^2 + O(h^3) \quad \text{for} \quad h \in \mathbb{R} \to 0 \,. \tag{1.5.4.31}$$

Trick. Take the imaginary part on both sides of (1.5.4.31) using that all derivatives are real:

$$\text{Im}\, f(x_0 + \imath h) = h f'(x_0) + O(h^3) \quad \text{for} \quad h \in \mathbb{R} \to 0 \,.$$

As a consequence we obtain the approximation

$$f'(x_0) = \frac{\text{Im}\, f(x_0 + \imath h)}{h} + O(h^2) \quad \text{for} \quad h \in \mathbb{R} \to 0 \,,$$

which suggests that we may rely on the *cancellation-free* expression

$$f'(x_0) \approx \frac{\operatorname{Im} f(x_0 + \imath h)}{h}$$

for $h \approx \sqrt{\mathrm{EPS}}$ to compute the derivative of f in x_0. ⌟

Remark 1.5.4.32 (A broader view of cancellation) Cancellation can be viewed as a particular case of a situation, in which severe amplification of relative errors is possible. Consider a function

$$F : \mathbb{R}^n \to \mathbb{R} \quad \text{of class } C^2\text{, twice continuously differentiable,}$$

and an argument vector $\mathbf{x} := [x_1, \ldots, x_n]^\top \neq \mathbf{0}$, for which

$$F(x_1, \ldots, x_n) \neq 0 \quad , \quad F(x_1, \ldots, x_n) \approx 0 \quad , \quad \mathbf{grad}\, F(x_1, \ldots, x_n) \neq \mathbf{0} \, .$$

We supply arguments $\widetilde{x}_i \in \mathbb{R}$ with small relative errors ϵ_i, $\widetilde{x}_i = x_i(1+\epsilon_i)$, $i = 1, \ldots, n$, and study the resulting relative error δ of the result

$$\delta := \frac{|F(\widetilde{x}_1, \ldots, \widetilde{x}_n) - F(x_1, \ldots, x_n)|}{|F(x_1, \ldots, x_n)|} \, .$$

Thanks to the smoothness of F, we can employ multi-dimensional Taylor approximation

$$F(x_1(1+\epsilon_1), \ldots, x_n(1+\epsilon_n)) = F(x_1, \ldots, x_n) + \mathbf{grad}\, F(x_1, \ldots, x_n)^\top \begin{bmatrix} \epsilon_1 x_1 \\ \vdots \\ \epsilon_n x_n \end{bmatrix} + R(\mathbf{x}, \boldsymbol{\epsilon}) \, ,$$

$$\text{with remainder} \qquad R(\mathbf{x}, \boldsymbol{\epsilon}) = O(\epsilon_1^2 + \cdots + \epsilon_n^2) \quad \text{for} \quad \epsilon_i \to 0 \, .$$

This yields

$$\delta = \frac{\left| \mathbf{grad}\, F(x_1, \ldots, x_n)^\top \begin{bmatrix} \epsilon_1 x_1 \\ \vdots \\ \epsilon_n x_n \end{bmatrix} + R(\mathbf{x}, \boldsymbol{\epsilon}) \right|}{|F(x_1, \ldots, x_n)|} \, .$$

If $|\epsilon_i| \ll 1$, we can neglect the remainder term and obtain the first-order approximation (indicated by $\doteq$)

$$\delta \doteq \frac{1}{|F(x_1, \ldots, x_n)|} \left| \mathbf{grad}\, F(x_1, \ldots, x_n)^\top \begin{bmatrix} \epsilon_1 x_1 \\ \vdots \\ \epsilon_n x_n \end{bmatrix} \right| \, .$$

In case $|(\mathbf{grad}\, F(x_1, \ldots, x_n))_i| \gg |F(\mathbf{x})|$, $|x_i| \gg 0$ for some $i \in \{1, \ldots, n\}$, we can thus encounter $\delta \gg \max_j |\epsilon_j|$, which indicates a potentially massive amplification of relative errors.

- "Classical cancellation" as discussed in § 1.5.4.5 fits this setting and corresponds to the special choice $F : \mathbb{R}^2 \to \mathbb{R}$, $F(x_1, x_2) := x_1 - x_2$.
- The effect found above can be observed for the simple trigonometric functions sin and cos!

⌟

Review question(s) 1.5.4.33.

(Q1.5.4.33.A) Give an expression for

$$I(a,b) := \int_a^b 1/x \, \mathrm{d}x$$

that allows cancellation-free evaluation for integration bounds $1 \ll a \approx b$.

(Q1.5.4.33.B) For integration bounds $a, b > 1$, $a \approx b$, propose a numerically sound way of computing

$$I(a,b) := \int_a^b \frac{1}{1+x^2} \, \mathrm{d}x \, .$$

Hints.

$$\frac{d}{dx}\{x \mapsto \arctan(x)\} = \frac{1}{1+x^2} \, ,$$
$$\tan(\alpha - \beta) = \frac{\tan(\alpha) - \tan\beta}{1 + \tan(\alpha)\tan(\beta)} \, .$$

(Q1.5.4.33.C) What is the problem with the C++ expression

```
y = std::log(std::cosh(x));
```

where `x` is of type **double**? Rewrite this line of code into an algebraically equivalent one so that problem does no longer occur.

(Q1.5.4.33.D) [Harmless cancellation] Discuss the impact of round-off errors and cancellation for the C++ expression

```
f = x + std::sqrt(1-x*x);
```

where `x` is of type **double** and in the interval $[-1,1]$.

(Q1.5.4.33.E) [Error amplification without subtraction] For what values of $0 < x < 2$ may the following expression yield a result with a huge relative error, even if we assume that the variable `x` contains an "exact value"?

```
f = std::cos(M_PI/2.0*x)/std::log(x);
```

Suggest a way to rewrite the expression that avoids this instability.

(Q1.5.4.33.F) [Error amplification by multiplication] We multiply two quantities $\widetilde{x} \in \mathbb{R}$ and $\widetilde{y} \in \mathbb{R}$, which carry relative errors bounded by machine precision, that is

$$\widetilde{x} = x(1+\delta_x) \quad , \quad \widetilde{y} = y(1+\delta_y) \, , \quad |\delta_x|, |\delta_y| \leq \mathrm{EPS} \, ,$$

where $x, y \in \mathbb{R}$ designate the "exact values". Give a bound for the relative error of the product $\widetilde{x} \cdot \widetilde{y}$ and discuss whether it can be much larger than EPS for particular values of x and y.

△

1.5.5 Numerical Stability

Video tutorial for Section 1.5.5 "Numerical Stability": (17 minutes) Download link, tablet notes

→ review questions 1.5.5.23

We have seen that a particular "problem" can be tackled by different "algorithms", which produce different results due to roundoff errors. This section will clarify what distinguishes a "good" algorithm from a rather abstract point of view.

§1.5.5.1 (The "problem")

A mathematical notion of "problem":

- data space X, usually $X \subset \mathbb{R}^n$
- result space Y, usually $Y \subset \mathbb{R}^m$
- mapping (problem function) $F : X \mapsto Y$

A problem is a well defined *function* that assigns to each datum a result.

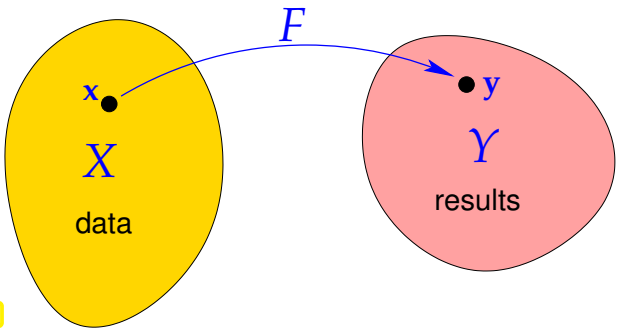


Fig. 32

Note: In this course, both the data space X and the result space Y will always be subsets of finite dimensional vector spaces.

EXAMPLE 1.5.5.2 (The "matrix×vector-multiplication problem") We consider the "problem" of computing the product $\mathbf{Ax}$ for a given matrix $\mathbf{A} \in \mathbb{K}^{m,n}$ and a given vector $\mathbf{x} \in \mathbb{K}^n$.

➢
- Data space $X = \mathbb{K}^{m,n} \times \mathbb{K}^n$ (input is a matrix and a vector)
- Result space $Y = \mathbb{R}^m$ (space of column vectors)
- Problem function $F : X \to Y$, $F(\mathbf{a}, \mathbf{x}) := \mathbf{Ax}$

⌟

§1.5.5.3 (Norms on spaces of vectors and matrices) Norms provide tools for measuring errors. Recall from linear algebra and calculus [NS02, Sect. 4.3], [Gut09, Sect. 6.1]:

Definition 1.5.5.4. Norm

X = vector space over field $\mathbb{K}$, $\mathbb{K} = \mathbb{C}, \mathbb{R}$. A map $\|\cdot\| : X \mapsto \mathbb{R}_0^+$ is a norm on X, if it satisfies

(i) $\forall \mathbf{x} \in X$: $\mathbf{x} \neq 0 \Leftrightarrow \|\mathbf{x}\| > 0$ (definite),
(ii) $\|\lambda \mathbf{x}\| = |\lambda| \|\mathbf{x}\| \quad \forall \mathbf{x} \in X, \lambda \in \mathbb{K}$ (homogeneous),
(iii) $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\| \quad \forall \mathbf{x}, \mathbf{y} \in X$ (triangle inequality).

Examples: (for vector space $\mathbb{K}^n$, vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T \in \mathbb{K}^n$)

name	:	definition	EIGEN function
Euclidean norm	:	$\|\mathbf{x}\|_2 := \sqrt{\lvert x_1\rvert^2 + \cdots + \lvert x_n\rvert^2}$	`x.norm()`
1-norm	:	$\|\mathbf{x}\|_1 := \lvert x_1\rvert + \cdots + \lvert x_n\rvert$	`x.lpNorm<1>()`
∞-norm, max norm	:	$\|\mathbf{x}\|_\infty := \max\{\lvert x_1\rvert, \ldots, \lvert x_n\rvert\}$	`x.lpNorm<Eigen::Infinity>()`

Remark 1.5.5.5 (Inequalities between vector norms) All norms on the vector space $\mathbb{K}^n$, $n \in \mathbb{N}$, are equivalent in the sense that for arbitrary two norms $\|\cdot\|_1$ and $\|\cdot\|_2$ we can always find a constant $C > 0$ such that

$$\|\mathbf{v}\|_1 \leq C \|\mathbf{v}\|_2 \quad \forall \mathbf{v} \in \mathbb{K}^n \,. \tag{1.5.5.6}$$

Of course, the constant C will usually depend on n and the norms under consideration.

For the vector norms introduced above, explicit expressions for the constants "C" are available: for all $\mathbf{x} \in \mathbb{K}^n$

$$\|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1 \leq \sqrt{n}\|\mathbf{x}\|_2 , \tag{1.5.5.7}$$
$$\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_2 \leq \sqrt{n}\|\mathbf{x}\|_\infty , \tag{1.5.5.8}$$
$$\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_1 \leq n\|\mathbf{x}\|_\infty . \tag{1.5.5.9}$$

⌟

The matrix space $\mathbb{K}^{m,n}$ is a vector space, of course, and can also be equipped with various norms. Of particular importance are norms *induced by vector norms* on $\mathbb{K}^n$ and $\mathbb{K}^m$.

Definition 1.5.5.10. Matrix norm

Given vector norms $\|\cdot\|_x$ and $\|\cdot\|_y$ on $\mathbb{K}^n$ and $\mathbb{K}^m$, respectively, the associated matrix norm is defined by

$$\mathbf{M} \in \mathbb{R}^{m,n}: \quad \|\mathbf{M}\| := \sup_{\mathbf{x}\in\mathbb{R}^n\setminus\{0\}} \frac{\|\mathbf{M}\mathbf{x}\|_y}{\|\mathbf{x}\|_x} .$$

By virtue of definition the matrix norms enjoy an important property, they are sub-multiplicative:

$$\forall \mathbf{A} \in \mathbb{K}^{n,m}, \mathbf{B} \in \mathbb{K}^{m,k}: \quad \|\mathbf{A}\mathbf{B}\| \leq \|\mathbf{A}\|\|\mathbf{B}\| . \tag{1.5.5.11}$$

✎ notations for matrix norms for *quadratic matrices* associated with standard vector norms:

$$\|\mathbf{x}\|_2 \to \|\mathbf{M}\|_2 , \quad \|\mathbf{x}\|_1 \to \|\mathbf{M}\|_1 , \quad \|\mathbf{x}\|_\infty \to \|\mathbf{M}\|_\infty$$

EXAMPLE 1.5.5.12 (Matrix norm associated with ∞-norm and 1-norm) Rather simple formulas are available for the matrix norms induced by the vector norms $\|\cdot\|_\infty$ and $\|\cdot\|_1$

e.g. for $\mathbf{M} = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix} \in \mathbb{K}^{2,2}$:

$$\begin{aligned} \|\mathbf{M}\mathbf{x}\|_\infty &= \max\{|m_{11}x_1 + m_{12}x_2|, |m_{21}x_1 + m_{22}x_2|\} \\ &\leq \max\{|m_{11}| + |m_{12}|, |m_{21}| + |m_{22}|\}\, \|x\|_\infty , \\ \|\mathbf{M}\mathbf{x}\|_1 &= |m_{11}x_1 + m_{12}x_2| + |m_{21}x_1 + m_{22}x_2| \\ &\leq \max\{|m_{11}| + |m_{21}|, |m_{12}| + |m_{22}|\}(|x_1| + |x_2|) . \end{aligned}$$

For general $\mathbf{M} = \left[m_{ij}\right] \in \mathbb{K}^{m,n}$

➢ matrix norm $\leftrightarrow \|\cdot\|_\infty$ = row sum norm $\|\mathbf{M}\|_\infty := \max_{i=1,\ldots,m} \sum_{j=1}^{n} |m_{ij}|$, (1.5.5.13)

➢ matrix norm $\leftrightarrow \|\cdot\|_1$ = column sum norm $\|\mathbf{M}\|_1 := \max_{j=1,\ldots,n} \sum_{i=1}^{m} |m_{ij}|$. (1.5.5.14)

⌟

Sometimes special formulas for the Euclidean matrix norm come handy [GV89, Sect. 2.3.3]:

Lemma 1.5.5.15. Formula for Euclidean norm of a Hermitian matrix

$$\mathbf{A} \in \mathbb{K}^{n,n}, \mathbf{A} = \mathbf{A}^H \quad \Rightarrow \quad \|\mathbf{A}\|_2 = \max_{\mathbf{x}\neq 0} \frac{|\mathbf{x}^H \mathbf{A}\mathbf{x}|}{\|\mathbf{x}\|_2^2} .$$

Proof. Recall from linear algebra: Hermitian matrices (a special class of normal matrices) enjoy unitary similarity to diagonal matrices:

$$\exists \mathbf{U} \in \mathbb{K}^{n,n}, \text{ diagonal } \mathbf{D} \in \mathbb{R}^{n,n}: \quad \mathbf{U}^{-1} = \mathbf{U}^H \quad \text{and} \quad \mathbf{A} = \mathbf{U}^H \mathbf{D} \mathbf{U} .$$

Since multiplication with an unitary matrix preserves the 2-norm of a vector, we conclude

$$\|\mathbf{A}\|_2 = \left\|\mathbf{U}^H \mathbf{D} \mathbf{U}\right\|_2 = \|\mathbf{D}\|_2 = \max_{i=1,\ldots,i} |d_i| , \quad \mathbf{D} = \operatorname{diag}(d_1, \ldots, d_n) .$$

On the other hand, for the same reason:

$$\max_{\|\mathbf{x}\|_2=1} \mathbf{x}^H \mathbf{A}\mathbf{x} = \max_{\|\mathbf{x}\|_2=1} (\mathbf{U}\mathbf{x})^H \mathbf{D} (\mathbf{U}\mathbf{x}) = \max_{\|\mathbf{y}\|_2=1} \mathbf{y}^H \mathbf{D}\mathbf{y} = \max_{i=1,\ldots,i} |d_i| .$$

Hence, both expressions in the statement of the lemma agree with the largest modulus of eigenvalues of $\mathbf{A}$.

□

Corollary 1.5.5.16. Euclidean matrix norm and eigenvalues

For $\mathbf{A} \in \mathbb{K}^{m,n}$ the Euclidean matrix norm $\|\mathbf{A}\|_2$ is the square root of the largest (in modulus) eigenvalue of $\mathbf{A}^H\mathbf{A}$.

For a normal matrix $\mathbf{A} \in \mathbb{K}^{n,n}$ (that is, $\mathbf{A}$ satisfies $\mathbf{A}^H\mathbf{A} = \mathbf{A}\mathbf{A}^H$) the Euclidean matrix norm agrees with the modulus of the largest eigenvalue.

⌟

§1.5.5.17 ((Numerical) algorithm) When we talk about an "algorithm" we have in mind a concrete code function in MATLAB or C++; the only way to describe an algorithm is through a piece of code. We assume that this function defines another mapping $\widetilde{F} : X \to Y$ on the data space of the problem. Of course, we can only feed data to the MATLAB/C++-function, if they can be represented in the set $\mathbb{M}$ of machine numbers. Hence, implicit in the definition of $\widetilde{F}$ is the assumption that input data are subject to *rounding* before passing them to the code function proper.

Problem	Algorithm
$F : X \subset \mathbb{R}^n \to Y \subset \mathbb{R}^m$	$\widetilde{F} : X \to \widetilde{Y} \subset \mathbb{M}$

⌟

§1.5.5.18 (Stable algorithm → [AG11, Sect. 1.3]) [Stable algorithm]

- We study a problem (→ § 1.5.5.1) $F : X \to Y$ on data space X into result space Y.
- We assume that both X and Y are equipped with norms $\|\cdot\|_X$ and $\|\cdot\|_Y$, respectively (→ Def. 1.5.5.4).
- We consider a concrete algorithm $\widetilde{F} : X \to Y$ according to § 1.5.5.17.

We write $w(\mathbf{x})$, $\mathbf{x} \in X$, for the computational effort ($\rightarrow$ Def. 1.4.0.1, "number of elementary operations") required by the algorithm for input $\mathbf{x}$.

Definition 1.5.5.19. Stable algorithm

An algorithm $\widetilde{F}$ for solving a problem $F : X \mapsto Y$ is numerically stable if for all $\mathbf{x} \in X$ its result $\widetilde{F}(\mathbf{x})$ (possibly affected by roundoff) is the exact result for "slightly perturbed" data:

$$\exists C \approx 1: \quad \forall \mathbf{x} \in X: \quad \exists \widetilde{\mathbf{x}} \in X: \quad \|\mathbf{x} - \widetilde{\mathbf{x}}\|_X \leq C w(\mathbf{x}) \,\mathrm{EPS}\, \|\mathbf{x}\|_X \quad \wedge \quad \widetilde{F}(\mathbf{x}) = F(\widetilde{\mathbf{x}}) \,.$$

Here EPS should be read as machine precision according to the "Axiom" of roundoff analysis Ass. 1.5.3.11.

Illustration of Def. 1.5.5.19 ▷
($\mathbf{y} \mathrel{\widehat{=}}$ exact result for exact data $\mathbf{x}$)

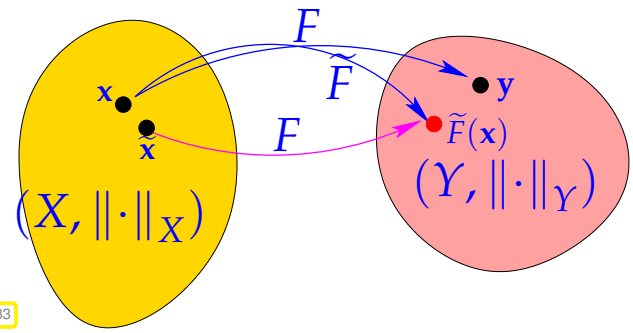


Fig. 33

Terminology:
Def. 1.5.5.19 introduces stability in the sense of backward error analysis

Sloppily speaking, the impact of roundoff $(*)$ on a *stable algorithm* is of the same order of magnitude as the effect of the inevitable perturbations due to rounding of the input data.

➢ For stable algorithms roundoff errors are "harmless".

$(*)$ In some cases the definition of $\widetilde{F}$ will also involve some approximations as in Ex. 1.5.4.26. Then the above statement also includes approximation errors. ⌟

EXAMPLE 1.5.5.20 (Testing stability of matrix×vector multiplication) Assume you are given a black box implementation of a function

```
VectorXd mvmult(const MatrixX &A,const VectorXd &x)
```

that purports to provide a stable implementation of $\mathbf{Ax}$ for $\mathbf{A} \in \mathbb{K}^{m,n}$, $\mathbf{x} \in \mathbb{K}^n$, *cf.* Ex. 1.5.5.2. How can we verify this claim for particular data. Both, $\mathbb{K}^{m,n}$ and $\mathbb{K}^n$ are equipped with the Euclidean norm.

The task is, given $\mathbf{y} \in \mathbb{K}^n$ as returned by the function, to find conditions on $\mathbf{y}$ that ensure the existence of a $\widetilde{\mathbf{A}} \in \mathbb{K}^{m,n}$ such that

$$\widetilde{\mathbf{A}}\mathbf{x} = \mathbf{y} \quad \text{and} \quad \left\|\widetilde{\mathbf{A}} - \mathbf{A}\right\|_2 \leq Cmn \,\mathrm{EPS}\, \|\mathbf{A}\|_2 \,, \tag{1.5.5.21}$$

for a small constant ≈ 1.

In fact we can choose (easy computation)

$$\widetilde{\mathbf{A}} = \mathbf{A} + \mathbf{z}\mathbf{x}^T \quad , \quad \mathbf{z} := \frac{\mathbf{y} - \mathbf{Ax}}{\|\mathbf{x}\|_2^2} \in \mathbb{K}^m \,,$$

and we find

$$\left\|\widetilde{\mathbf{A}}-\mathbf{A}\right\|_2 = \left\|\mathbf{z}\mathbf{x}^T\right\|_2 = \sup_{\mathbf{w}\in\mathbb{K}^n\setminus\{\mathbf{0}\}} \frac{\mathbf{x}\cdot\mathbf{w}\|\mathbf{z}\|_2}{\|\mathbf{w}\|_2} \le \|\mathbf{x}\|_2\|\mathbf{z}\|_2 = \frac{\|\mathbf{y}-\mathbf{A}\mathbf{x}\|_2}{\|\mathbf{x}\|_2}\,.$$

Hence, in principle stability of an algorithm for computing $\mathbf{A}\mathbf{x}$ is confirmed, if for every $\mathbf{x}\in\mathbb{R}^n$ the computed result $\widetilde{\mathbf{y}} = \texttt{mvmult}(\mathbf{A},\mathbf{x})$ satisfies

$$\|\widetilde{\mathbf{y}}-\mathbf{A}\mathbf{x}\|_2 \le C\,mn\,\mathrm{EPS}\,\|\mathbf{x}\|_2\|\mathbf{A}\|_2\,,$$

with a small constant $C>0$ independent of data and problem size.

⌟

Remark 1.5.5.22 (Numerical stability and sensitive dependence on data)

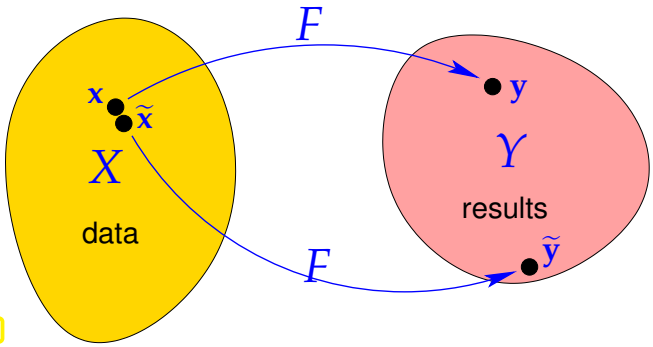


Fig. 34

A problem shows sensitive dependence on the data, if small perturbations of input data lead to large perturbations of the output. Such problems are also called **ill-conditioned**. For such problems stability of an algorithm is easily accomplished.

◁ "Mental image": ill-conditioned problem: slightly different data (w.r.t. $\|\cdot\|_X$) yield vastly different results ($\|\mathbf{y}-\widetilde{\mathbf{y}}\|_Y$ large).

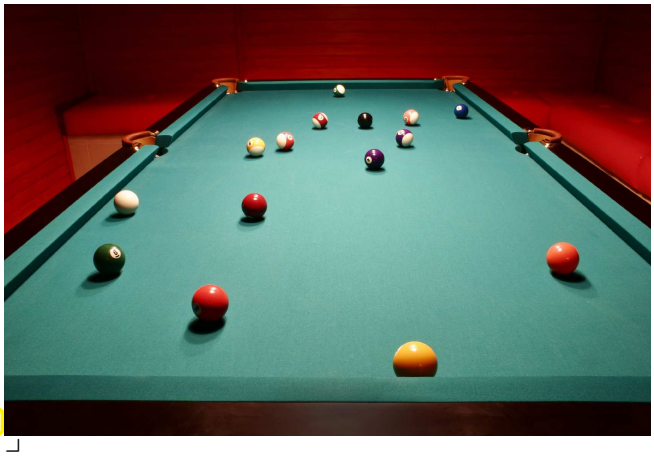


Fig. 35

Example: The problem is the prediction of the position of the billard ball after ten bounces given the initial position, velocity, and spin.

It is well known, that tiny changes of the initial conditions can shift the final location of the ball to virtually any point on the table: the billard problem is chaotic.

Hence, a stable algorithm for its solution may just output a *fixed* or *random* position without even using the initial conditions!

⌟

Review question(s) 1.5.5.23 (Numerical stability)

(Q1.5.5.23.A) We consider the problem of multiplying the Kronecker product of two real $n\times n$ matrices with a vector. Give the formula for the problem mapping and characterize the (largest possible))data space and result space.

(Q1.5.5.23.B) Fill in the blanks in the following definition of a stable algorithm:

Definition 1.5.5.19. Stable algorithm

An algorithm $\widetilde{F}$ for solving a problem $F : X \mapsto Y$ is **numerically stable** if for all $\mathbf{x} \in X$ its result $\widetilde{F}(\mathbf{x})$ (possibly affected by roundoff) is the exact result for "slightly perturbed" data:

$$\exists C \approx 1\colon \quad \forall \square \colon \quad \exists \square \colon \quad \left\| \square \right\|_X \leq C w(\mathbf{x})\, \mathrm{EPS} \|\mathbf{x}\|_X \quad \wedge \quad \widetilde{F}\left(\square \right) = F\left(\square \right).$$

(Q1.5.5.23.C) Suppose you have to examine a black-box function

```
double add(double x, double y);
```

that just adds the two numbers given as arguments. Derive conditions on the returned result that, when satisfied, imply the stability of the implementation of `add()`. Of course, the norm on $\mathbb{R}$ is just $|\cdot|$.

(Q1.5.5.23.D) [A stable implementation?] Show that

```
auto f = [](double x) -> double
   { return std::sin(std::numbers::pi * std::sqrt(x));};
```

provides a stable implementation of $x \mapsto \sin(\pi\sqrt{x})$, $x \in \mathbb{R}$.

△

Learning Outcomes

Principal take-home knowledge and skills from this chapter:

- Learning by doing: Knowledge about the syntax of fundamental operations on matrices and vectors in EIGEN.
- Understanding of the concepts of computational effort/cost and asymptotic complexity in numerics.
- Awareness of the asymptotic complexity of basic linear algebra operations
- Ability to determine the (asymptotic) computational effort for a concrete (numerical linear algebra) algorithm.
- Ability to manipulate simple expressions involving matrices and vectors in order to reduce the computational cost for their evaluation.
- Knowledge about round-off and machine precision.
- Familiarity with the phenomenon of "cancellation": cause, effect, remedies, and tricks

Bibliography

[AV88] A. Aggarwal and J.S. Vitter. “The input/output complexity of sorting and related problems”. In: *Communications of the ACM* 31.9 (1988), pp. 1116–1127 (cit. on p. 83).

[AG11] Uri M. Ascher and Chen Greif. *A first course in numerical methods*. Vol. 7. Computational Science & Engineering. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2011, pp. xxii+552. DOI: 10.1137/1.9780898719987 (cit. on pp. 56, 83, 95, 96, 98, 99, 101, 104, 105, 108, 115, 121).

[CW90] D. Coppersmith and S. Winograd. “Matrix multiplication via arithmetic progression”. In: *J. Symbgolic Computing* 9.3 (1990), pp. 251–280 (cit. on p. 86).

[DR08] W. Dahmen and A. Reusken. *Numerik für Ingenieure und Naturwissenschaftler*. Heidelberg: Springer, 2008 (cit. on pp. 57, 104).

[GV89] G.H. Golub and C.F. Van Loan. *Matrix computations*. 2nd. Baltimore, London: John Hopkins University Press, 1989 (cit. on pp. 78, 120).

[GJ10] Gero Greiner and Riko Jacob. “The I/O Complexity of Sparse Matrix Dense Matrix Multiplication”. In: *LATIN 2010: THEORETICAL INFORMATICS*. Ed. by LopezOrtiz, A. Vol. 6034. Lecture Notes in Computer Science. Microsoft Res; Yahoo Res; Univ Waterloo. 2010, 143–156. DOI: {10.1007/978-3-642-12200-2_14} (cit. on p. 83).

[Gut09] M.H. Gutknecht. *Lineare Algebra*. Lecture Notes. SAM, ETH Zürich, 2009 (cit. on p. 119).

[KW03] M. Kowarschik and C. Weiss. “An Overview of Cache Optimization Techniques and Cache-Aware Numerical Algorithms”. In: *Algorithms for Memory Hierarchies*. Vol. 2625. Lecture Notes in Computer Science. Heidelberg: Springer, 2003, pp. 213–232 (cit. on p. 83).

[LM67] J. N. Lyness and C. B. Moler. “Numerical differentiation of analytic functions”. In: *SIAM J. Numer. Anal.* 4 (1967), pp. 202–210 (cit. on p. 116).

[NS02] K. Nipp and D. Stoffer. *Lineare Algebra*. 5th ed. Zürich: vdf Hochschulverlag, 2002 (cit. on pp. 53, 70, 91, 119).

[Ove01] M.L. Overton. *Numerical Computing with IEEE Floating Point Arithmetic*. Philadelphia, PA: SIAM, 2001 (cit. on p. 96).

[QSS00] A. Quarteroni, R. Sacco, and F. Saleri. *Numerical mathematics*. Vol. 37. Texts in Applied Mathematics. New York: Springer, 2000 (cit. on pp. 57, 76).

[Str69] V. Strassen. “Gaussian elimination is not optimal”. In: *Numer. Math.* 13 (1969), pp. 354–356 (cit. on p. 85).

[Str09] M. Struwe. *Analysis für Informatiker*. Lecture notes, ETH Zürich. 2009 (cit. on pp. 53, 54, 115).

[Van00] Charles F. Van Loan. “The ubiquitous Kronecker product”. In: *J. Comput. Appl. Math.* 123.1-2 (2000), pp. 85–100. DOI: 10.1016/S0377-0427(00)00393-9.

Chapter 2

Direct Methods for (Square) Linear Systems of Equations

§2.0.0.1 (Required prior knowledge for Chapter 2) Also this chapter heavily relies on concepts and techniques from linear algebra as taught in the 1st semester introductory course. Knowledge of the following topics from linear algebra will be taken for granted and they should be refreshed in case of gaps:

- Operations involving matrices and vectors [NS02, Ch. 2], already covered in Chapter 1
- Computations with block-structured matrices, *cf.* § 1.3.1.13
- Linear systems of equations: existence and uniqueness of solutions [NS02, Sects. 1.2, 3.3]
- Gaussian elimination [NS02, Ch. 2]
- LU-decomposition and its connection with Gaussian elimination [NS02, Sect. 2.4]

Contents

2.1 Introduction: Linear Systems of Equations (LSE) . . . 127
2.2 Theory: Linear Systems of Equations (LSE) . . . 130
2.2.1 LSE: Existence and Uniqueness of Solutions . . . 130
2.2.2 Sensitivity/Conditioning of Linear Systems . . . 131
2.3 Gaussian Elimination (GE) . . . 135
2.3.1 Basic Algorithm . . . 135
2.3.2 LU-Decomposition . . . 143
2.3.3 Pivoting . . . 151
2.4 Stability of Gaussian Elimination . . . 158
2.5 Survey: Elimination Solvers for Linear Systems of Equations . . . 165
2.6 Exploiting Structure when Solving Linear Systems . . . 170
2.7 Sparse Linear Systems . . . 179
2.7.1 Sparse Matrix Storage Formats . . . 180
2.7.2 Sparse Matrices in EIGEN . . . 183
2.7.3 Direct Solution of Sparse Linear Systems of Equations . . . 191
2.7.4 LU-Factorization of Sparse Matrices . . . 194
2.7.5 Banded Matrices [DR08, Sect. 3.7] . . . 199
2.8 Stable Gaussian Elimination Without Pivoting . . . 206

2.1 Introduction: Linear Systems of Equations (LSE)

§2.1.0.1 (The problem: solving a linear system) What is "the problem" considered in this chapter, when we apply the notion of "Problem" introduced in § 1.5.5.1, that is, which functions do "Direct Methods for Linear Systems of Equations (LSE)" attempt to evaluate and what suitable data spaces X and result spaces Y?

Input/data : square matrix $\mathbf{A} \in \mathbb{K}^{n,n}$, vector $\mathbf{b} \in \mathbb{K}^n$, $n \in \mathbb{N}$ ➢ data space $X = \mathbb{K}^{n,n} \times \mathbb{K}^n$

Output/result : solution vector $\mathbf{x} \in \mathbb{K}^n$: $\boxed{\mathbf{A}\mathbf{x} = \mathbf{b}}$ ← (square) **linear system of equations** (LSE)

➢ result space $Y = \mathbb{K}^n$

(Terminology: $\mathbf{A} \hat{=}$ system matrix/coefficient matrix, $\mathbf{b} \hat{=}$ right hand side vector)

Linear systems with rectangular system matrices $\mathbf{A} \in \mathbb{K}^{m,n}$, called "overdetermined" for $m > n$, and "underdetermined" for $m < n$ will be treated in Chapter 3. ⌟

Remark 2.1.0.2 (LSE: key components of mathematical models in many fields) Linear systems of equations are ubiquitous in computational science: they are encountered

- with discrete linear models in network theory (see Ex. 2.1.0.3), control, statistics;
- in the case of *discretized* boundary value problems for ordinary and partial differential equations (→ course "Numerical methods for partial differential equations", 4th semester);
- as a result of linearization (e.g, "Newton's method" → Section 8.5).

⌟

EXAMPLE 2.1.0.3 (Nodal analysis of (linear) electric circuit [QSS00, Sect. 4.7.1])

Now we study a very important application of numerical simulation, where (large, sparse) linear systems of equations play a central role: Numerical circuit analysis. We begin with *linear circuits* in the *frequency domain*, which are directly modelled by complex linear systems of equations. In later chapters we will tackle circuits with non-linear elements, see Ex. 8.1.0.1, and, finally, will learn about numerical methods for computing the transient (time-dependent) behavior of circuits, see Ex. 11.1.2.11.

Modeling of simple linear circuits takes only elementary physical laws as covered in any introductory course of physics (or even in secondary school physics). There is no sophisticated physics or mathematics involved. Circuits are composed of so-called circuit elements connected by (ideal) wires.

A circuit diagram ▷

●: **Nodes** that is, junctions of wires

We *number* the nodes $1, \ldots, n$ and write I_{kj} (physical units, $[I_{kj}] = 1\mathrm{A}$) for the electric current flowing from node k to node j. Currents have a sign: $I_{kj} = -I_{jk}$

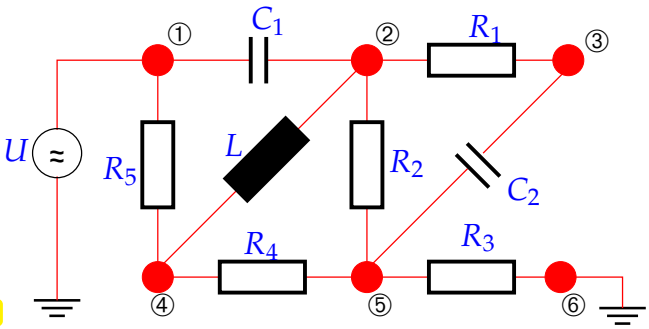


Fig. 36

The most fundamental relationship is the **Kirchhoff current law** (KCL) that demands that the sum of node currents vanishes:

$$\forall k \in \{1,\ldots,n\}: \quad \sum_{j=1}^{n} I_{kj} = 0 \ . \tag{2.1.0.4}$$

The unknowns of the model are the nodal potentials U_k, $k = 1,\ldots,n$. (Some of them may be known, for instance those for grounded nodes: ⑥ in Fig. 36, and nodes connected to voltage sources: ① in Fig. 36.) The difference of the nodal potentials of two connected nodes is called the branch voltage.

The circuit elements are characterized by current-voltage relationships, so-called constitutive relations, here given in frequency domain for angular frequency $\omega > 0$ (physical units $[\omega] = 1\mathrm{s}^{-1}$). We consider only the following simple circuit elements:

- Ohmic resistor: $I = \dfrac{U}{R}$, $[R] = 1\mathrm{VA}^{-1}$
- capacitor: $I = \imath\omega C U$, capacitance $[C] = 1\mathrm{AsV}^{-1}$
- coil/inductor : $I = \dfrac{U}{\imath\omega L}$, inductance $[L] = 1\mathrm{VsA}^{-1}$

$\Rightarrow \quad I_{kj} = \begin{cases} R^{-1}(U_k - U_j) \ , \\ \imath\omega C(U_k - U_j) \ , \\ -\imath\omega^{-1}L^{-1}(U_k - U_j) \ . \end{cases}$

✎ notation: $\imath \triangleq$ imaginary unit "$\imath := \sqrt{-1}$", $\imath = \exp(\imath\pi/2)$, $\imath^2 = -1$

Here we face the special case of a linear circuit: all relationships between branch currents and voltages are of the form

$$I_{kj} = \alpha_{kj}(U_k - U_j) \quad \text{with} \quad \alpha_{kj} \in \mathbb{C} \ . \tag{2.1.0.5}$$

The concrete value of α_{kj} is determined by the circuit element connecting node k and node j.

These constitutive relations are derived by assuming a harmonic time-dependence of all quantities, which is termed circuit analysis in the frequency domain (AC-mode).

$$\text{voltage:} \quad u(t) = \mathrm{Re}\{U\exp(\imath\omega t)\} \quad , \quad \text{current:} \quad i(t) = \mathrm{Re}\{I\exp(\imath\omega t)\} \ . \tag{2.1.0.6}$$

Here $U, I \in \mathbb{C}$ are called complex amplitudes. This implies for temporal derivatives (denoted by a dot):

$$\frac{du}{dt}(t) = \mathrm{Re}\{\imath\omega U\exp(\imath\omega t)\} \quad , \quad \frac{di}{dt}(t) = \mathrm{Re}\{\imath\omega I\exp(\imath\omega t)\} \ . \tag{2.1.0.7}$$

For a capacitor the total charge is proportional to the applied voltage:

$$q(t) = Cu(t) \quad \overset{i(t) = \frac{dq}{dt}(t)}{\Rightarrow} \quad i(t) = C\frac{du}{dt}(t) \ .$$

For a coil the voltage is proportional to the rate of change of current: $u(t) = L\frac{di}{dt}(t)$. Combined with (2.1.0.6) and (2.1.0.7) this leads to the above constitutive relations.

Now we combine the constitutive relations with the Kirchhoff current law (2.1.0.4). We end up with a linear system of equations!

$$\begin{aligned}
&② : \quad \imath\omega C_1(U_2 - U_1) + R_1^{-1}(U_2 - U_3) - \imath\omega^{-1}L^{-1}(U_2 - U_4) + R_2^{-1}(U_2 - U_5) &&= 0 \ , \\
&③ : \quad R_1^{-1}(U_3 - U_2) + \imath\omega C_2(U_3 - U_5) &&= 0 \ , \\
&④ : \quad R_5^{-1}(U_4 - U_1) - \imath\omega^{-1}L^{-1}(U_4 - U_2) + R_4^{-1}(U_4 - U_5) &&= 0 \ , \\
&⑤ : \quad R_2^{-1}(U_5 - U_2) + \imath\omega C_2(U_5 - U_3) + R_4^{-1}(U_5 - U_4) + R_3^{-1}(U_5 - U_6) &&= 0 \ ,
\end{aligned}$$

$$U_1 = U \quad , \quad U_6 = 0 \, .$$

We do not get equations for the nodes ① and ⑥, because these nodes are connected to the "outside world" so that the Kirchhoff current law (2.1.0.4) does not hold (from a local perspective). This is fitting, because the voltages in these nodes are known anyway.

$$\begin{bmatrix} \imath\omega C_1 + \frac{1}{R_1} - \frac{i}{\omega L} + \frac{1}{R_2} & -\frac{1}{R_1} & \frac{i}{\omega L} & -\frac{1}{R_2} \\ -\frac{1}{R_1} & \frac{1}{R_1} + \imath\omega C_2 & 0 & -\imath\omega C_2 \\ \frac{i}{\omega L} & 0 & \frac{1}{R_5} - \frac{i}{\omega L} + \frac{1}{R_4} & -\frac{1}{R_4} \\ -\frac{1}{R_2} & -\imath\omega C_2 & -\frac{1}{R_4} & \frac{1}{R_2} + \imath\omega C_2 + \frac{1}{R_4} + R_3^{-1} \end{bmatrix} \begin{bmatrix} U_2 \\ U_3 \\ U_4 \\ U_5 \end{bmatrix} = \begin{bmatrix} \imath\omega C_1 U \\ 0 \\ \frac{1}{R_5} U \\ 0 \end{bmatrix}$$

This is a linear system of equations with *complex* coefficients: $\mathbf{A} \in \mathbb{C}^{4,4}$, $\mathbf{b} \in \mathbb{C}^4$. For the algorithms to be discussed below this does not matter, because they work alike for real and complex numbers. ⌟

Review question(s) 2.1.0.8 (Nodal analysis of linear electric circuits)

(Q2.1.0.8.A) [A simple resistive circuit]

In the electric circuit drawn beside all resistors have the same resistance $R > 0$.
Which nodal potentials are already known?
Derive the linear system of equations for the remaining unkown nodal potentials.

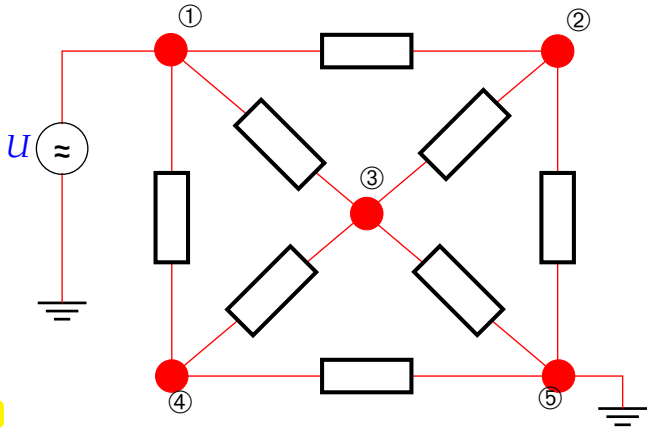


Fig. 37

(Q2.1.0.8.B) [Current source] The voltage source with strength U in Fig. 37 is replaced with a current source, which drives a known current I through the circuit branch it is attached to.

Which linear system of equations has to be solved in order to determine the unknown nodal potentials U_1, U_2, U_3, U_4?

(Q2.1.0.8.C) A linear mapping $\mathsf{L} : \mathbb{R}^n \to \mathbb{R}^n$ is represented by the matrix $\mathbf{A} \in \mathbb{R}^{n,n}$ with respect to the standard basis of $\mathbb{R}^n$ comprising Cartesian coordinate vectors $\mathbf{e}_\ell$, $\ell = 1, \ldots, n$.

Explain, how one can compute the matrix representation of L with respect to the basis

$$\left\{ \begin{bmatrix} 2 \\ 1 \\ 0 \\ \vdots \\ \vdots \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 2 \\ 1 \\ 0 \\ \vdots \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 2 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \cdots \begin{bmatrix} 0 \\ \vdots \\ \vdots \\ 0 \\ 1 \\ 2 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ \vdots \\ \vdots \\ \vdots \\ 0 \\ 1 \\ 2 \end{bmatrix} \right\}$$

by merely solving n linear systems of equations and forming matrix products.

△

2.2 Theory: Linear Systems of Equations (LSE)

2.2.1 LSE: Existence and Uniqueness of Solutions

The following concepts and results are known from linear algebra [NS02, Sect. 1.2], [Gut09, Sect. 1.3]:

Definition 2.2.1.1. Invertible matrix → [NS02, Sect. 2.3]

$$\mathbf{A} \in \mathbb{K}^{n,n} \quad \text{invertible/regular} \quad :\Leftrightarrow \quad \exists_1 \mathbf{B} \in \mathbb{K}^{n,n}: \quad \mathbf{AB} = \mathbf{BA} = \mathbf{I} \,.$$

$\mathbf{B}$ is called the **inverse** of $\mathbf{A}$, (✎ notation $\mathbf{B} = \mathbf{A}^{-1}$)

Now, recall a few notions from linear algebra needed to state criteria for the invertibility of a matrix.

Definition 2.2.1.2. Image space and kernel of a matrix

Given $\mathbf{A} \in \mathbb{K}^{m,n}$, the range/image (space) of $\mathbf{A}$ is the subspace of $\mathbb{K}^m$ spanned by the columns of $\mathbf{A}$

$$\mathcal{R}(\mathbf{A}) := \{\mathbf{Ax},\ \mathbf{x} \in \mathbb{K}^n\} \subset \mathbb{K}^m \,.$$

The kernel/nullspace of $\mathbf{A}$ is

$$\mathcal{N}(\mathbf{A}) := \{\mathbf{z} \in \mathbb{K}^n :\ \mathbf{Az} = \mathbf{0}\} \,.$$

Definition 2.2.1.3. Rank of a matrix → [NS02, Sect. 2.4], [QSS00, Sect. 1.5]

The rank of a matrix $\mathbf{M} \in \mathbb{K}^{m,n}$, denoted by $\operatorname{rank}(\mathbf{M})$, is the maximal number of linearly independent rows/columns of $\mathbf{M}$. Equivalently, $\operatorname{rank}(\mathbf{A}) = \dim \mathcal{R}(\mathbf{A})$.

Theorem 2.2.1.4. Criteria for invertibility of matrix → [NS02, Sect. 2.3 & Cor. 3.8]

A square matrix $\mathbf{A} \in \mathbb{K}^{n,n}$ is invertible/regular if one of the following equivalent conditions is satisfied:

1. $\exists \mathbf{B} \in \mathbb{K}^{n,n}: \quad \mathbf{BA} = \mathbf{AB} = \mathbf{I}$,
2. $\mathbf{x} \mapsto \mathbf{Ax}$ *defines an endomorphism of* $\mathbb{K}^n$,
3. *the columns of* $\mathbf{A}$ *are linearly independent (full column rank),*
4. *the rows of* $\mathbf{A}$ *are linearly independent (full row rank),*
5. $\det \mathbf{A} \neq 0$ *(non-vanishing determinant),*
6. $\operatorname{rank}(\mathbf{A}) = n$ *(full rank).*

§2.2.1.5 (Solution of a LSE as a "problem", recall § 2.1.0.1) Linear algebra give us a *formal* way to denote solution of LSE:

$$\mathbf{A} \in \mathbb{K}^{n,n} \text{ regular} \quad \& \quad \mathbf{Ax} = \mathbf{b} \quad \Rightarrow \quad \mathbf{x} = \mathbf{A}^{-1}\mathbf{b} \,.$$

($\mathbf{A}^{-1}$: inverse matrix)

Now recall our notion of "problem" from § 1.5.5.1 as a function F mapping data in a data space X to a result in a result space Y. Concretely, for $n \times n$ linear systems of equations:

$$F: \begin{cases} X := \mathbb{K}^{n,n}_* \times \mathbb{K}^n & \to \quad Y := \mathbb{K}^n \\ (\mathbf{A}, \mathbf{b}) & \mapsto \quad \mathbf{A}^{-1}\mathbf{b} \end{cases}$$

✎ notation: (open) set of regular matrices $\subset \mathbb{K}^{n,n}$:

$$\mathbb{K}^{n,n}_* := \{\mathbf{A} \in \mathbb{K}^{n,n} : \mathbf{A} \text{ regular/invertible} \to \text{Def. 2.2.1.1}\} .$$

⌟

Remark 2.2.1.6 (The inverse matrix and solution of a LSE) In principle, in EIGEN the inverse of a matrix $\mathbf{A}$ is available through the member function `inverse()` of matrix type, see EIGEN documentation. However, there are only a few case that always involve fixed-size small matrices, where the actual computation of the inverse of a matrix is warranted. The general advice is the following:

Avoid computing the inverse of a matrix (which can almost always be avoided)!
In particular, never ever even contemplate using `x = A.inverse()*b` to solve the linear system of equations $\mathbf{A}\mathbf{x} = \mathbf{b}$, *cf.* Exp. 2.4.0.14. The next sections present a sound way to do this.

Another reason for this advice is given in Exp. 2.4.0.14. ⌟

Review question(s) 2.2.1.7 (LSE: Existence and Uniqueness of Solutions)

(Q2.2.1.7.A) [Diagonal linear systems of equations] How can you tell that a square linear system of equations with a *diagonal* system matrix has a unique solution?

(Q2.2.1.7.B) Outline a practical algorithm that checks whether an upper triangular matrix $\mathbf{A} \in \mathbb{R}^{n,n}$, $n \in \mathbb{N}$, is invertible.

(Q2.2.1.7.C) A square matrix $\mathbf{A} \in \mathbb{R}^{n,n}$ has the following entries

$$(\mathbf{A})_{i,j} = \begin{cases} 1 & \text{, if } j \geq i \text{ ,} \\ \alpha & \text{, if } i = n, j = 1 \text{ ,} \\ 0 & \text{elsewhere,} \end{cases} \qquad i,j \in \{1,\ldots,n\} \text{ ,} \quad \alpha \in \mathbb{R} .$$

For what values of α is $\mathbf{A}$ regular?

△

2.2.2 Sensitivity/Conditioning of Linear Systems

The **sensitivity**/samterm*conditioning of a problem (for given data) gauges the impact of small perturbations of the data on the result.

Before we examine sensitivity for linear systems of equations, we look at the simpler problem of matrix×vector multiplication.

EXAMPLE 2.2.2.1 (Sensitivity of linear mappings) For a *fixed* given regular $\mathbf{A} \in \mathbb{K}^{n,n}$ we study the problem map

$$F : \mathbb{K}^n \to \mathbb{K}^n \quad , \quad \mathbf{x} \mapsto \mathbf{A}\mathbf{x} ,$$

that is, now we consider only the vector $\mathbf{x}$ as data.

Goal: Estimate relative perturbations in $F(\mathbf{x})$ due to relative perturbations in $\mathbf{x}$.

We assume that $\mathbb{K}^n$ is equipped with *some* vector norm (→ Def. 1.5.5.4) and we use the induced matrix norm (→ Def. 1.5.5.10) on $\mathbb{K}^{n,n}$. Using linearity and the elementary estimate $\|\mathbf{M}\mathbf{x}\| \leq \|\mathbf{M}\|\|\mathbf{x}\|$, which is a direct consequence of the definition of an induced matrix norm, we obtain

$$\begin{aligned} \mathbf{A}\mathbf{x} = \mathbf{y} \quad &\Rightarrow \quad \|\mathbf{x}\| \leq \left\|\mathbf{A}^{-1}\right\|\|\mathbf{y}\| \\ \mathbf{A}(\mathbf{x}+\Delta\mathbf{x}) = \mathbf{y}+\Delta\mathbf{y} \quad &\Rightarrow \quad \mathbf{A}\Delta\mathbf{x} = \Delta\mathbf{y} \quad \Rightarrow \quad \|\Delta\mathbf{y}\| \leq \|\mathbf{A}\|\|\Delta\mathbf{x}\| \end{aligned}$$

$$\Rightarrow \quad \frac{\|\Delta \mathbf{y}\|}{\|\mathbf{y}\|} \leq \frac{\|\mathbf{A}\|\|\Delta \mathbf{x}\|}{\|\mathbf{A}^{-1}\|^{-1}\|\mathbf{x}\|} = \|\mathbf{A}\|\left\|\mathbf{A}^{-1}\right\| \frac{\|\Delta \mathbf{x}\|}{\|\mathbf{x}\|} \,. \tag{2.2.2.2}$$

relative perturbation in result — relative perturbation in data

We have found that the quantity $\|\mathbf{A}\|\left\|\mathbf{A}^{-1}\right\|$ bounds amplification of relative errors in the argument vector in a matrix×vector-multiplication with the matrix $\mathbf{A}$. ⌟

Now we study the sensitivity of the problem of finding the solution of a linear system of equations $\mathbf{A}\mathbf{x} = \mathbf{b}$, $\mathbf{A} \in \mathbb{R}^{n,n}$ regular, $\mathbf{b} \in \mathbb{R}^n$, see § 2.1.0.1. We write $\widetilde{\mathbf{x}}$ for the solution of the perturbed linear system.

Question: To what extent do perturbations in the data $\mathbf{A}$, $\mathbf{b}$ cause a

$$\text{(normwise) relative error:} \quad \epsilon_r := \frac{\|\mathbf{x} - \widetilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \quad ?$$

($\|\cdot\| \mathrel{\hat{=}}$ suitable vector norm, e.g., maximum norm $\|\cdot\|_\infty$)

Perturbed linear system:

$$\mathbf{A}\mathbf{x} = \mathbf{b} \;\leftrightarrow\; (\mathbf{A} + \Delta\mathbf{A})\widetilde{\mathbf{x}} = \mathbf{b} + \Delta\mathbf{b} \;\blacktriangleright\; (\mathbf{A} + \Delta\mathbf{A})(\widetilde{\mathbf{x}} - \mathbf{x}) = \Delta\mathbf{b} - \Delta\mathbf{A}\mathbf{x} \,. \tag{2.2.2.3}$$

Theorem 2.2.2.4. Conditioning of LSEs → **[QSS00, Thm. 3.1], [GGK14, Thm 3.5]**

If $\mathbf{A}$ *regular,* $\|\Delta\mathbf{A}\| < \left\|\mathbf{A}^{-1}\right\|^{-1}$ *and* (2.2.2.3)*, then*

(i) $\mathbf{A} + \Delta\mathbf{A}$ *is regular/invertible,*

(ii) If $\mathbf{A}\mathbf{x} = \mathbf{b}$, $(\mathbf{A} + \Delta\mathbf{A})\widetilde{\mathbf{x}} = \mathbf{b} + \Delta\mathbf{b}$, *then*

$$\frac{\|\mathbf{x} - \widetilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq \frac{\left\|\mathbf{A}^{-1}\right\|\|\mathbf{A}\|}{1 - \left\|\mathbf{A}^{-1}\right\|\|\mathbf{A}\|\|\Delta\mathbf{A}\|/\|\mathbf{A}\|} \left(\frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|} + \frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|} \right) .$$

relative error of data — *relative perturbations*

The proof is based on the following fundamental result:

Lemma 2.2.2.5. Perturbation lemma → **[QSS00, Thm. 1.5]**

$$\mathbf{B} \in \mathbb{R}^{n,n},\ \|\mathbf{B}\| < 1 \quad \Rightarrow \quad \mathbf{I} + \mathbf{B} \text{ regular} \quad \wedge \quad \left\|(\mathbf{I} + \mathbf{B})^{-1}\right\| \leq \frac{1}{1 - \|\mathbf{B}\|} .$$

Proof. We start with the △-inequality

$$\|(\mathbf{I} + \mathbf{B})\mathbf{x}\| \geq (\|\mathbf{x}\| - \|\mathbf{B}\mathbf{x}\|) \geq \underbrace{(1 - \|\mathbf{B}\|)}_{\neq 0}\|\mathbf{x}\| \quad \forall \mathbf{x} \in \mathbb{R}^n \,. \tag{2.2.2.6}$$

We conclude that $\mathbf{I} + \mathbf{B}$ must have trivial kernel $\mathcal{N}(\mathbf{I} + \mathbf{B}) = \{\mathbf{0}\}$, which implies that the square matrix $\mathbf{I} + \mathbf{B}$ is regular. We continue using this fact, the definition of the matrix norm, and (2.2.2.6):

$$\left\|(\mathbf{I} + \mathbf{B})^{-1}\right\| = \sup_{\mathbf{x} \in \mathbb{R}^n \setminus \{0\}} \frac{\left\|(\mathbf{I} + \mathbf{B})^{-1}\mathbf{x}\right\|}{\|\mathbf{x}\|} = \sup_{\mathbf{y} \in \mathbb{R}^n \setminus \{0\}} \frac{\|\mathbf{y}\|}{\|(\mathbf{I} + \mathbf{B})\mathbf{y}\|} \leq \frac{1}{1 - \|\mathbf{B}\|} \,.$$

□

Proof. (of Thm. 2.2.2.4) We use a slightly generalized version of Lemma 2.2.2.5, which gives us

$$\left\|(\mathbf{A}+\Delta\mathbf{A})^{-1}\right\| \leq \frac{\left\|\mathbf{A}^{-1}\right\|}{1-\left\|\mathbf{A}^{-1}\Delta\mathbf{A}\right\|} .$$

We combine this estimate with (2.2.2.3):

$$\|\Delta\mathbf{x}\| \leq \frac{\left\|\mathbf{A}^{-1}\right\|}{1-\left\|\mathbf{A}^{-1}\Delta\mathbf{A}\right\|}(\|\Delta\mathbf{b}\|+\|\Delta\mathbf{A}\mathbf{x}\|) \leq \frac{\left\|\mathbf{A}^{-1}\right\|\|\mathbf{A}\|}{1-\left\|\mathbf{A}^{-1}\right\|\|\Delta\mathbf{A}\|}\left(\frac{\|\Delta\mathbf{b}\|}{\|\mathbf{A}\|\|\mathbf{x}\|}+\frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|}\right)\|\mathbf{x}\| .$$

□

Note that the term $\|\mathbf{A}\|\left\|\mathbf{A}^{-1}\right\|$ occurs frequently. Therefore it has been given a special name:

Definition 2.2.2.7. Condition (number) of a matrix

Condition (number) of a matrix $\mathbf{A} \in \mathbb{R}^{n,n}$: $\quad \operatorname{cond}(\mathbf{A}) := \left\|\mathbf{A}^{-1}\right\|\|\mathbf{A}\|$

Note: $\operatorname{cond}(\mathbf{A})$ depends on the matrix norm $\|\cdot\|$!

Rewriting estimate of Thm. 2.2.2.4 with $\Delta\mathbf{b} = 0$,

$$\epsilon_r := \frac{\|\mathbf{x}-\widetilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq \frac{\operatorname{cond}(\mathbf{A})\delta_A}{1-\operatorname{cond}(\mathbf{A})\delta_A} , \quad \delta_A := \frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|} . \tag{2.2.2.8}$$

From (2.2.2.8) we conclude important messsages of $\operatorname{cond}(\mathbf{A})$:

- If $\operatorname{cond}(\mathbf{A}) \gg 1$, *small perturbations* in $\mathbf{A}$ can lead to *large relative errors* in the solution of the LSE.
- If $\operatorname{cond}(\mathbf{A}) \gg 1$, a stable algorithm ($\rightarrow$ Def. 1.5.5.19) can produce solutions with large relative error !

Recall Thm. 2.2.2.4: for regular $\mathbf{A} \in \mathbb{K}^{n,n}$, small $\Delta\mathbf{A}$, generic vector/matrix norm $\|\cdot\|$

$$\begin{array}{c}\mathbf{A}\mathbf{x}=\mathbf{b}\\ (\mathbf{A}+\Delta\mathbf{A})\widetilde{\mathbf{x}}=\mathbf{b}+\Delta\mathbf{b}\end{array} \Rightarrow \frac{\|\mathbf{x}-\widetilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq \frac{\operatorname{cond}(\mathbf{A})}{1-\operatorname{cond}(\mathbf{A})\|\Delta\mathbf{A}\|/\|\mathbf{A}\|}\left(\frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}+\frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|}\right) . \tag{2.2.2.9}$$

$\operatorname{cond}(\mathbf{A}) \gg 1$ ➢ small relative changes of data $\mathbf{A}, \mathbf{b}$ *may* effect huge relative changes in solution.

$\operatorname{cond}(\mathbf{A})$ indicates sensitivity of "LSE problem" $(\mathbf{A}, \mathbf{b}) \mapsto \mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$
(as "amplification factor" of (worst-case) relative perturbations in the data $\mathbf{A}, \mathbf{b}$).

Terminology:

Small changes of data $\Rightarrow$ small perturbations of result : well-conditioned problem
Small changes of data $\Rightarrow$ large perturbations of result : ill-conditioned problem

Note: sensitivity gauge depends on the chosen norm **!**

EXAMPLE 2.2.2.10 (Intersection of lines in 2D) Solving a 2×2 linear system of equations amounts to finding the intersection of two lines in the coordinate plane: This relationship allows a geometric view of "sensitivity of a linear system", when using the distance metric (Euclidean vector norm).

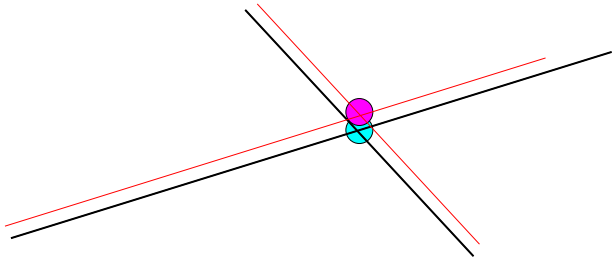
Remember the Hessian normal form of a straight line in the plane. We are given the Hessian normal forms of two lines L_1 and L_2 and want to compute the coordinate vector $\mathbf{x} \in \mathbb{R}^2$ of the point in which they intersect:

$$L_i = \{\mathbf{x} \in \mathbb{R}^2 : \mathbf{x}^T \mathbf{n}_i = d_i\}\,, \quad \mathbf{n}_i \in \mathbb{R}^2,\, d_i \in \mathbb{R}\,, \quad i = 1,2\,.$$

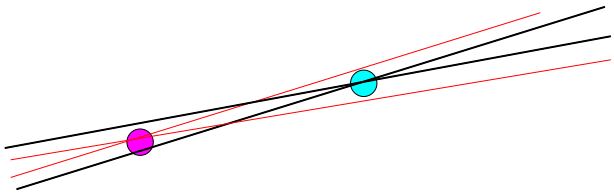
◀ LSE for finding intersection: $\underbrace{\begin{bmatrix} \mathbf{n}_1^T \\ \mathbf{n}_2^T \end{bmatrix}}_{=:\mathbf{A}} \mathbf{x} = \underbrace{\begin{bmatrix} d_1 \\ d_2 \end{bmatrix}}_{=:\mathbf{b}}$,

where the $\mathbf{n}_i$ are (unit) direction vectors, and the $d_i \in \mathbb{R}$ give the (signed) distance to the origin.

Now we perturb the right-hand side vector $\mathbf{b}$ and wonder how this will impact the intersection points. The situation is illustrated by the following two pictures, in which the original and perturbed lines are drawn in black and red, respectively.



nearly orthogonal intersection: well-conditioned



glancing intersection: ill-conditioned

Obviously, if the lines are almost parallel, a small shift in their position will lead to a big shift of the intersection point.

The following EIGEN-based C++ code investigates condition numbers for the matrix $\begin{bmatrix} 1 & \cos\varphi \\ 0 & \sin\varphi \end{bmatrix}$ that can arise when computing the intersection of two lines enclosing the angle φ. As usual the directive **using namespace** `Eigen;` was given in the beginning of the file.

C++-code 2.2.2.11: condition numbers of 2×2 matrices ↪ GITLAB

```
2   VectorXd phi = VectorXd::LinSpaced(50, M_PI / 200, M_PI / 2);
3   MatrixXd res(phi.size(), 3);
4   Matrix2d A;
5   A(0, 0) = 1;
6   A(1, 0) = 0;
7   for (int i = 0; i < phi.size(); ++i) {
8     A(0, 1) = std::cos(phi(i));
9     A(1, 1) = std::sin(phi(i));
10    // L2 condition number is the quotient of the maximal
11    // and minimal singular value of A
12    JacobiSVD<MatrixXd> svd(A);
13    double C2 = svd.singularValues()(0) / //
14                svd.singularValues()(svd.singularValues().size() - 1);
15    // L-infinity condition number
16    double Cinf =
```

```
17          A.inverse().cwiseAbs().rowwise().sum().maxCoeff() *
18          A.cwiseAbs().rowwise().sum().maxCoeff(); //
19      res(i, 0) = phi(i);
20      res(i, 1) = C2;
21      res(i, 2) = Cinf;
22    }
```

In Line 13 we compute the condition number of $\mathbf{A}$ with respect to the Euclidean vector norm using special EIGEN built-in functions.

Line 18 evaluated the condition number of a matrix for the maximum norm, recall Ex. 1.5.5.12.

We clearly observe a blow-up of $\mathrm{cond}(\mathbf{A})$ (with respect to the Euclidean vector norms) as the angle enclosed by the two lines shrinks.

This corresponds to a large sensitivity of the location of the intersection point in the case of glancing incidence.

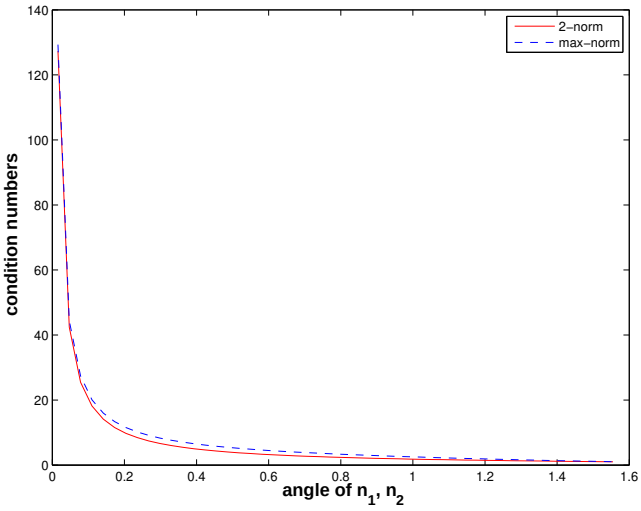


Fig. 38

Heuristics for predicting large $\mathrm{cond}(\mathbf{A})$

$$\mathrm{cond}(\mathbf{A}) \gg 1 \quad \leftrightarrow \quad \text{columns/rows of } \mathbf{A} \text{ "almost linearly dependent"}$$

Review question(s) 2.2.2.12 (Sensitivity of linear systems)

(Q2.2.2.12.A) Analyze the sensitivity of a linear system of equations with *diagonal* system matrix relying on the Euclidean vector and matrix norms. Consider perturbations of the right-hand side vector and of the diagonal elements and investigate the amplification of relative errors.

△

2.3 Gaussian Elimination (GE)

2.3.1 Basic Algorithm

The problem of solving a linear system of equations is rather special compared to many other numerical tasks:

! An exceptional feature of linear systems of equations (LSE) is that its "exact" solution computable with finitely many elementary operations.

The algorithm is **Gaussian elimination (GE)** ($\rightarrow$ secondary school, linear algebra,)

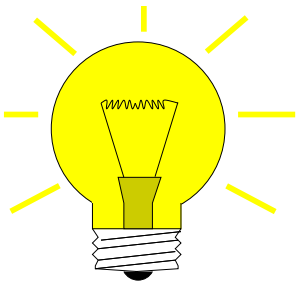
Familiarity with the algorithm of Gaussian elimination for a square linear system of equations will be taken for granted.

Supplementary literature. In case you cannot remember the main facts about Gaussian elimination, very detailed accounts and examples can be found in

- M. Gutknecht's lecture notes [Gut09, Ch. 1],
- the textbook by Nipp & Stoffer [NS02, Ch. 1],
- the numerical analysis text by Quarteroni et al. [QSS00, Sects. 3.2 & 3.3],
- the textbook by Ascher & Greif [AG11, Sect. 5.1],

and, to some extent, below, see Ex. 2.3.1.1.

Wikipedia: Although the method is named after mathematician **Carl Friedrich Gauss**, the earliest presentation of it can be found in the important Chinese mathematical text *Jiuzhang suanshu* or The Nine Chapters on the Mathematical Art, dated approximately 150 B.C., and commented on by **Liu Hui** in the 3rd century.



The idea of Gaussian elimination is the transformation of a linear system of equations into a "simpler", but equivalent LSE by means of successive (invertible) *row transformations*.

Rem. 1.3.1.12: row transformations $\leftrightarrow$ left-multiplication with transformation matrix

Obviously, left multiplication with a regular matrix does not affect the solution of an LSE: for any *regular* $\mathbf{T} \in \mathbb{K}^{n,n}$

$$\mathbf{Ax} = \mathbf{b} \quad \Rightarrow \quad \mathbf{A}'\mathbf{x} = \mathbf{b}' \quad \text{, if} \quad \mathbf{A}' = \mathbf{TA},\ \mathbf{b}' = \mathbf{Tb}\ .$$

So we may try to convert the linear system of equations to a form that can be solved more easily by multiplying with regular matrices from left, which boils down to applying row transformations. A suitable target format is a diagonal linear system of equations, for which all equations are completely decoupled. This is the gist of Gaussian elimination.

EXAMPLE 2.3.1.1 (Gaussian elimination)

Stage ① (Forward) elimination:

$$\begin{bmatrix} 1 & 1 & 0 \\ 2 & 1 & -1 \\ 3 & -1 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 4 \\ 1 \\ -3 \end{bmatrix} \quad \longleftrightarrow \quad \begin{array}{rcrcrcr} x_1 & + & x_2 & & & = & 4 \\ 2x_1 & + & x_2 & - & x_3 & = & 1 \\ 3x_1 & - & x_2 & - & x_3 & = & -3 \end{array} .$$

$$\begin{bmatrix} 1 & 1 & 0 \\ 2 & 1 & -1 \\ 3 & -1 & -1 \end{bmatrix} \begin{bmatrix} 4 \\ 1 \\ -3 \end{bmatrix} \rightarrowtail \begin{bmatrix} \mathbf{1} & \mathbf{1} & 0 \\ 0 & -1 & -1 \\ 3 & -1 & -1 \end{bmatrix} \begin{bmatrix} 4 \\ -7 \\ -3 \end{bmatrix}$$

$$\rightarrowtail \begin{bmatrix} \mathbf{1} & \mathbf{1} & 0 \\ 0 & -1 & -1 \\ 0 & -4 & -1 \end{bmatrix} \begin{bmatrix} 4 \\ -7 \\ -15 \end{bmatrix} \rightarrowtail \underbrace{\begin{bmatrix} 1 & 1 & 0 \\ 0 & \mathbf{-1} & -1 \\ 0 & 0 & 3 \end{bmatrix}}_{=\mathbf{U}} \begin{bmatrix} 4 \\ -7 \\ 13 \end{bmatrix}$$

■ = pivot row, pivot element **bold**.

▶ We have transformed the LSE to **upper triangular form**

Stage ② Solve by **back substitution**:

$$\begin{array}{rcrcrcl} x_1 & + & x_2 & & & = & 4 \\ & - & x_2 & - & x_3 & = & -7 \\ & & & & 3x_3 & = & 13 \end{array} \quad \Rightarrow \quad \begin{array}{rcl} x_3 & = & \frac{13}{3} \\ x_2 & = & 7 - \frac{13}{3} = \frac{8}{3} \\ x_1 & = & 4 - \frac{8}{3} = \frac{4}{3} \, . \end{array}$$

More detailed examples are given in [Gut09, Sect. 1.1], [NS02, Sect. 1.1]. ⌟

More general:

$$\begin{array}{ccccccccc} a_{11}\,x_1 & + & a_{12}\,x_2 & + & \cdots & + & a_{1n}\,x_n & = & b_1 \\ a_{21}\,x_1 & + & a_{22}\,x_2 & + & \cdots & + & a_{2n}\,x_n & = & b_2 \\ \vdots & & \vdots & & \vdots & & \vdots & \vdots & \vdots \\ \vdots & & \vdots & & \vdots & & \vdots & \vdots & \vdots \\ a_{n1}\,x_1 & + & a_{n2}\,x_2 & + & \cdots & + & a_{nn}\,x_n & = & b_n \end{array}$$

- i-th row - $l_{i1}\cdot$ 1st row (pivot row), $l_{i1} := {a_{i1}}/{a_{11}}$, $i = 2, \ldots, n$

$$\begin{array}{ccccccccc} a_{11}\,x_1 & + & a_{12}\,x_2 & + & \cdots & + & a_{1n}\,x_n & = & b_1 \\ & & a_{22}^{(1)}\,x_2 & + & \cdots & + & a_{2n}^{(1)}\,x_n & = & b_2^{(1)} \\ & & \vdots & & \vdots & & \vdots & \vdots & \vdots \\ & & \vdots & & \vdots & & \vdots & \vdots & \vdots \\ & & a_{n2}^{(1)}\,x_2 & + & \cdots & + & a_{nn}^{(1)}\,x_n & = & b_n^{(1)} \end{array}$$

with

$$\begin{aligned} a_{ij}^{(1)} &= a_{ij} - a_{1j}\,l_{i1}, \qquad i,j = 2,\ldots,n \, , \\ b_i^{(1)} &= b_i - b_1\,l_{i1}, \qquad i = 2,\ldots,n \, . \end{aligned}$$

- i-th row - $l_{i1}\cdot$ 2nd row (pivot row), $l_{i2} := {a_{i2}^{(1)}}/{a_{22}^{(1)}}$, $i = 3, \ldots, n$.

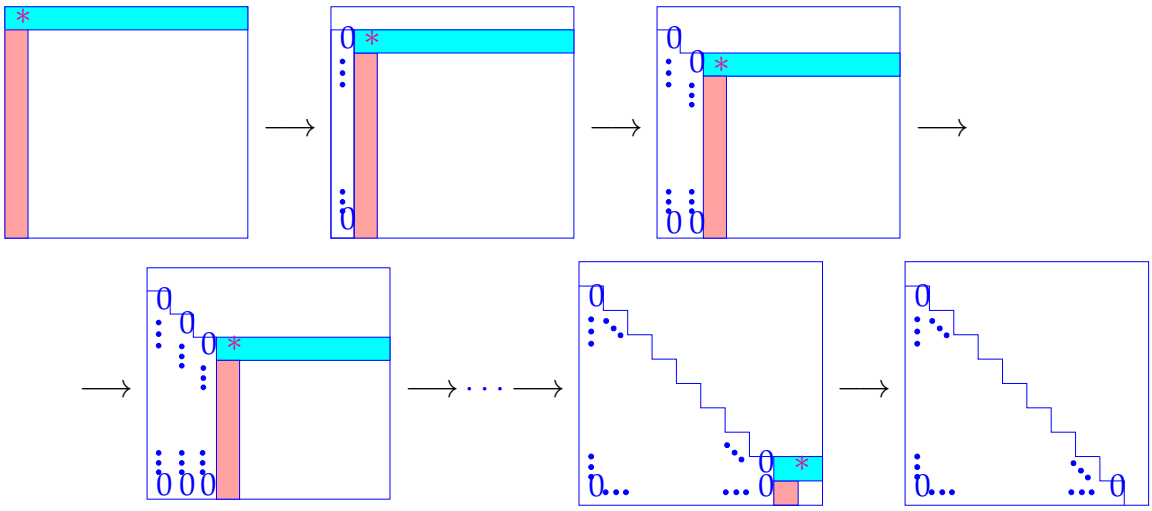
$$\begin{array}{ccccccccccc} a_{11}\,x_1 & + & a_{12}\,x_2 & + & a_{13}\,x_3 & + & \cdots & + & a_{1n}\,x_n & = & b_1 \\ & & a_{22}^{(1)}\,x_2 & + & a_{23}^{(1)}\,x_3 & + & \cdots & + & a_{2n}^{(1)}\,x_n & = & b_2^{(1)} \\ & & & & a_{33}^{(2)}\,x_3 & + & \cdots & + & a_{3n}^{(2)}\,x_n & = & b_3^{(2)} \\ & & & & \vdots & & \vdots & \vdots & \vdots & \vdots & \vdots \\ & & & & a_{n3}^{(2)}\,x_3 & + & \cdots & + & a_{nn}^{(2)}\,x_n & = & b_n^{(2)} \end{array}$$

After $n-1$ steps: linear systems of equations in upper triangular form

$$\begin{array}{ccccccccccc} a_{11}\,x_1 & + & a_{12}\,x_2 & + & a_{13}\,x_3 & + & \cdots & + & a_{1n}\,x_n & = & b_1 \\ & & a_{22}^{(1)}\,x_2 & + & a_{23}^{(1)}\,x_3 & + & \cdots & + & a_{2n}^{(1)}\,x_n & = & b_2^{(1)} \\ & & & & a_{33}^{(2)}\,x_3 & + & \cdots & + & a_{3n}^{(2)}\,x_n & = & b_3^{(2)} \\ & & & & & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots \\ & & & & & & \ddots & \ddots & \vdots & \vdots & \vdots \\ & & & & & & & & a_{nn}^{(n-1)}\,x_n & = & b_n^{(n-1)} \end{array}$$

Terminology: $a_{11}, a_{22}^{(1)}, a_{33}^{(2)}, \ldots, a_{n-1,n-1}^{(n-2)}$ = **pivots/pivot elements**

Graphical depiction:



$*$ $\hat{=}$ the pivot entry (necessarily $\neq 0$, which we assume here), = pivot row

In k-th step (starting from $\mathbf{A} \in \mathbb{K}^{n,n}$, $1 \leq k < n$, pivot row $\mathbf{a}_{k.}^{\top}$):

transformation: $\mathbf{Ax} = \mathbf{b}$ ➤ $\mathbf{A}'\mathbf{x} = \mathbf{b}'$.

with

$$a'_{ij} := \begin{cases} a_{ij} - \frac{a_{ik}}{a_{kk}} a_{kj} & \text{for } k < i, j \leq n \text{ ,} \\ 0 & \text{for } k < i \leq n, j = k \text{ ,} \\ a_{ij} & \text{else,} \end{cases} \quad b'_i := \begin{cases} b_i - \frac{a_{ik}}{a_{kk}} b_k & \text{for } k < i \leq n \text{ ,} \\ b_i & \text{else.} \end{cases} \tag{2.3.1.2}$$

multipliers l_{ik}

§2.3.1.3 (Gaussian elimination: algorithm) Here we give a direct EIGEN implementation of Gaussian elimination for LSE $\mathbf{Ax} = \mathbf{b}$ (grossly inefficient!).

C++ code 2.3.1.4: Solving LSE $\mathbf{Ax} = \mathbf{b}$ with Gaussian elimination ↪ GITLAB

```
2  //!  Gauss elimination without pivoting, x = A\b
3  //!  A must be an n×n-matrix, b an n-vector
4  //!  The result is returned in x
5  void gausselimsolve(const MatrixXd &A, const VectorXd& b,
6                      VectorXd& x) {
7    const Index n = A.rows();
8    MatrixXd Ab(n,n+1); // Augmented matrix [A,b]
9    Ab << A, b; //
10   // Forward elimination (cf.  step ① in Ex. 2.3.1.1)
11   for(Index i = 0; i < n-1; ++i) {
12     const double pivot = Ab(i,i);
13     for(Index k = i+1; k < n; ++k) {
14       const double fac = Ab(k, i) / pivot; // the multiplier
15       Ab.block(k,i+1,1,n-i)-= fac * Ab.block(i,i+1,1,n-i); //
16     }
17   }
18   // Back substitution (cf.  step ② in Ex. 2.3.1.1)
19   Ab(n-1,n) = Ab(n-1,n) / Ab(n-1,n-1);
20   for(Index i = n-2; i >= 0; --i) {
21     for(Index l = i+1; l < n; ++l) {
22       Ab(i,n) -= Ab(l,n)*Ab(i,l);
23     }
```

```
24      Ab(i,n) /= Ab(i,i);
25    }
26    x = Ab.rightCols(1); // Solution in rightmost column!
27  }
```

- In Line 9 the right hand side vector set as last column of matrix, which facilitates simultaneous row transformations of matrix and r.h.s.
- In Line 14 the variable `fac` is the multiplier from (2.3.1.2).
- In Line 26 we extract solution from last column of the transformed matrix.

⌟

§2.3.1.5 (Computational effort of Gaussian elimination) We examine Code 2.3.1.4.

- Forward elimination involves three nested loops (note that the compact vector operation in Line 15 involves another loop from $i+1$ to m).
- Back substitution can be done with two nested loops.

computational cost ($\leftrightarrow$ number of elementary operations) of Gaussian elimination [NS02, Sect. 1.3]:

$$\begin{aligned} &\text{forward elimination :} && \sum_{i=1}^{n-1}(n-i)(2(n-i)+3) = n(n-1)(\tfrac{2}{3}n+\tfrac{7}{6}) \text{ Ops. ,} \\ &\text{back substitution :} && \sum_{i=1}^{n} 2(n-i)+1 = n^2 \text{ Ops. .} \end{aligned} \tag{2.3.1.6}$$

asymptotic complexity ($\rightarrow$ Section 1.4) of Gaussian elimination (without pivoting) for generic LSE $\mathbf{Ax}=\mathbf{b}$, $\mathbf{A}\in\mathbb{R}^{n,n}$ $= \frac{2}{3}n^3 + O(n^2) = O(n^3)$

⌟

EXPERIMENT 2.3.1.7 (Runtime of Gaussian elimination) In this experiment we compare the efficiency of our hand-coded Gaussian elimination with that of library functions.

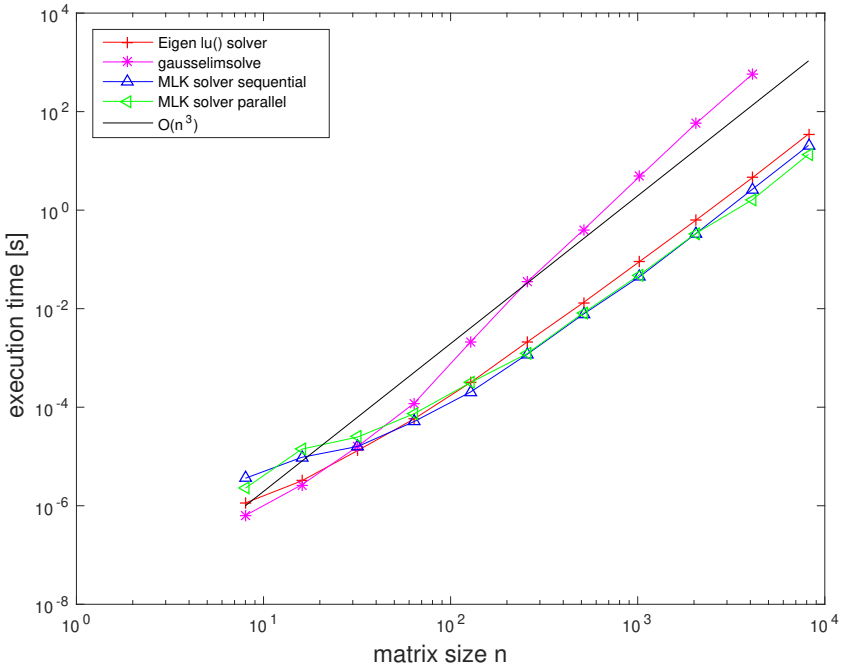
C++ code 2.3.1.8: Measuring runtimes of Code 2.3.1.4 vs. EIGEN lu()-operator vs. MKL
➽ GITLAB

```
2  //!  Eigen code for timing numerical solution of linear systems
3  MatrixXd gausstiming(){
4    std::vector<int> n = {8,16,32,64,128,256,512,1024,2048,4096,8192};
5    int nruns = 3;
6    MatrixXd times(n.size(),3);
7    for(int i = 0; i < n.size(); ++i){
8      Timer t1, t2;  // timer class
9      MatrixXd A = MatrixXd::Random(n[i],n[i]) + n[i]*MatrixXd::Identity(n[i],n[i]);
10     VectorXd b = VectorXd::Random(n[i]);
11     VectorXd x(n[i]);
12     for(int j = 0; j < nruns; ++j){
13       t1.start();  x = A.lu().solve(b);  t1.stop();// Eigen implementation
14       #ifndef EIGEN_USE_MKL_ALL   // only test own algorithm without MKL
15       if(n[i] <= 4096)       // Prevent long runs
16         t2.start();    gausselimsolve(A,b,x);   t2.stop();     // own gauss
             elimination
17       #endif
18     }
19     times(i,0) = n[i]; times(i,1) = t1.min(); times(i,2) = t2.min();
20   }
21   return times;
22 }
```

Platform:

- ubuntu 14.04 LTS
- i7-3517U CPU @ 1.90GHz Ã— 4
- L1 32 KB, L2 256 KB, L3 4096 KB, Mem 8 GB
- gcc 4.8.4, -O3

EIGEN is about two orders of magnitude faster than a direct implementation, MKL is even faster.



n	Code 2.3.1.4 [s]	EIGEN lu() [s]	MKL sequential [s]	MKL parallel [s]
8	6.340e-07	1.140e-06	3.615e-06	2.273e-06
16	2.662e-06	3.203e-06	9.603e-06	1.408e-05
32	1.617e-05	1.331e-05	1.603e-05	2.495e-05
64	1.214e-04	5.836e-05	5.142e-05	7.416e-05
128	2.126e-03	3.180e-04	2.041e-04	3.176e-04
256	3.464e-02	2.093e-03	1.178e-03	1.221e-03
512	3.954e-01	1.326e-02	7.724e-03	8.175e-03
1024	4.822e+00	9.073e-02	4.457e-02	4.864e-02
2048	5.741e+01	6.260e-01	3.347e-01	3.378e-01
4096	5.727e+02	4.531e+00	2.644e+00	1.619e+00
8192	-	3.510e+01	2.064e+01	1.360e+01

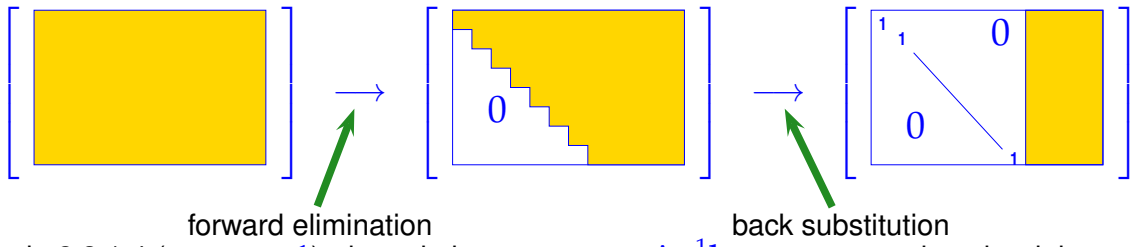
Never implement Gaussian elimination yourself **!**

use numerical libraries (LAPACK/MKL) or EIGEN **!**

A concise list of libraries for numerical linear algebra and related problems can be found here.

Remark 2.3.1.9 (Gaussian elimination for non-square matrices) In Code 2.3.1.4: the right hand side vector $\mathbf{b}$ was first appended to matrix $\mathbf{A}$ as rightmost column, and then forward elimination and back substitution were carried out on the resulting matrix. This can be generalized to a Gaussian elimination for rectangular matrices $\mathbf{A} \in \mathbb{K}^{n,n+1}$!

Consider a "fat matrix" $\mathbf{A} \in \mathbb{K}^{n,m}$, $m > n$:



Recall Code 2.3.1.4 ($m = n+1$): the solution vector $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$ was recovered as the rightmost column of the augmented matrix $(\mathbf{A},\mathbf{b})$ after forward elimination and back substitution. In the above cartoon it would be contained in the yellow part of the matrix on the right.

With this technique we have an efficient way for *simultaneously* solving of LSEs with multiple right hand sides . These multiple right-hand side can be passed as the column of a matrix $\mathbf{B}$ and the problem of solving an LSE for several right-hand-side vectors can be stated as follows:

Given regular $\mathbf{A} \in \mathbb{K}^{n,n}$, $\mathbf{B} \in \mathbb{K}^{n,k}$, seek $\mathbf{X} \in \mathbb{K}^{n,k}$ such that

$$\mathbf{AX} = \mathbf{B} \quad \Leftrightarrow \quad \mathbf{X} = \mathbf{A}^{-1}\mathbf{B}$$

Usually library functions meant to solve LSEs also accept a matrix instead of a right-hand-side vector and then return a matrix of solution vectors. For instance, in EIGEN the following function call accomplishes this:

```
Eigen::MatrixXd X = A.lu().solve(B);
```

Its asymptotic complexityis $O(n^2(n+k))$ for $n,k \to \infty$.

C++ code 2.3.1.10: Gaussian elimination with multiple r.h.s. → Code 2.3.1.4 ↬ GITLAB

```
//!  Gauss elimination without pivoting, X = A^{-1}B
//!  A must be an n×n-matrix, B an n×m-matrix
//!  Result is returned in matrix X
void gausselimsolvemult(const MatrixXd &A, const MatrixXd& B,
                MatrixXd& X) {
  const Eigen::Index n = A.rows();
  const Eigen::Index m = B.cols();
  MatrixXd AB(n, n+m); // Augmented matrix [A,B]
  AB << A, B;
  // Forward elimination, do not forget the B part of the Matrix
  for(Eigen::Index i = 0; i < n-1; ++i){
    const double pivot = AB(i,i);
    for(Eigen::Index k = i+1; k < n; ++k){
      const double fac = AB(k,i)/pivot;
      AB.block(k,i+1,1,m+n-i-1)-= fac * AB.block(i,i+1,1,m+n-i-1);
    }
  }
  // Back substitution
  AB.block(n-1, n,1, m) /= AB(n-1,n-1);
  for(Eigen::Index i = n-2; i >= 0; --i) {
    for(Eigen::Index l = i+1; l < n; ++l) {
      AB.block(i,n,1,m) -= AB.block(l,n,1,m)*AB(i,l);
    }
    AB.block(i,n,1,m) /= AB(i,i);
  }
  X = AB.rightCols(m);
}
```

⌟

Concerning the next two remarks: For understanding or analyzing special variants of Gaussian elimination, it is useful to be aware of

- the effects of elimination steps on the level of *matrix blocks*, *cf.* § 1.3.1.13,
- and of the *recursive nature* of Gaussian elimination.

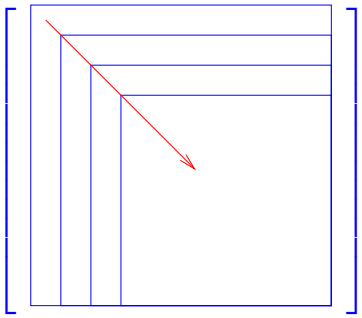
Remark 2.3.1.11 (Gaussian elimination via rank-1 modifications) We can view Gaussian elimination from the perspective of matrix block operations: Then the first step of Gaussian elimination with pivot $\alpha \neq 0$), *cf.* (2.3.1.2), can be expressed as

$$\mathbf{A} := \left[\begin{array}{c|c} \alpha & \mathbf{c}^\top \\ \hline \mathbf{d} & \mathbf{C} \end{array}\right] \rightarrow \mathbf{A}' := \left[\begin{array}{c|c} \alpha & \mathbf{c}^\top \\ \hline \mathbf{0} & \mathbf{C}' := \mathbf{C} - \frac{\mathbf{d}\mathbf{c}^\top}{\alpha} \end{array}\right] . \qquad (2.3.1.12)$$

rank-1 modification of $\mathbf{C}$

Terminology: Adding a tensor product of two vectors to a matrix is called a rank-1 modification of that matrix, see also § 2.6.0.12 below.

Notice that the transformation (2.3.1.12) is applied to the resulting lower-right block $\mathbf{C}'$ in the next elimination step. Thus Gaussian elimination can be realized by successive rank-1 modifications applied to smaller and smaller lower-right blocks of the matrix. An implementation in this spirit is given in Code 2.3.1.13.



r.h.s. $\mathbf{b} \sim$ `A(:,end)` ➤

C++ code 2.3.1.13: GE by rank-1 modification ➙ GITLAB

```
2  //!  in-situ Gaussian elimination, no pivoting
3  //!  right hand side in rightmost column of A
4  //!  back substitution is not done in this code!
5  void blockgs(Eigen::MatrixXd &A){
6    const Eigen::Index n = A.rows();
7    for(Eigen::Index i = 1; i < n; ++i){
8      // rank-1 modification of C
9      A.bottomRightCorner(n-i,n-i+1) -=
           A.col(i-1).tail(n-i) * A.row(i-1).tail(n-i+1) /
           A(i-1,i-1);
10     A.col(i - 1).tail(n - i).setZero(); // set d = 0
11   }
12 }
```

In this code the Gaussian elimination is carried out in situ: the matrix `A` is replaced with the transformed matrices during elimination. If the matrix is not needed later this offers maximum efficiency. An in-situ LU-decomposition as described in Rem. 2.3.2.11 could also be performed by Code 2.3.1.13 after a modification of Line 10. ⌟

Remark 2.3.1.14 (Block Gaussian elimination) Recall the "principle" from § 1.3.1.13: deal with block matrices ("matrices of matrices") like regular matrices (except for commutativity of multiplication!). This suggests a block view of Gaussian elimination:

Given: regular matrix $\mathbf{A} \in \mathbb{K}^{n,n}$ with sub-matrices/blocks

$$\mathbf{A}_{11} := (\mathbf{A})_{1:k,1:k} \quad , \quad \mathbf{A}_{12} = (\mathbf{A})_{1:k,k+1,n}, \qquad k < n,$$
$$\mathbf{A}_{21} := (\mathbf{A})_{k+1:n,1:k} \quad , \quad \mathbf{A}_{22} = (\mathbf{A})_{k+1:n,k+1,n},$$

and a right-hand-side vector $\mathbf{b} \in \mathbb{K}^n$ split into $\mathbf{b}_1 = (\mathbf{b})_{1:k}$, $\mathbf{b}_2 = (\mathbf{b})_{k+1:n}$

We apply the usual row transformations from (2.3.1.2) on the level of matrix block to this block-partitioned linear system using $\mathbf{A}_{11}$ as pivot block. Of course, we have to assume that $\mathbf{A}_{11}$ is invertible, generalizing the assumption that eligible pivot elements must not be zero. Again, the manipulations can be broken down into an elimination step ❶ and a backsubstitution step ❷.

$$\left[\begin{array}{cc|c} \mathbf{A}_{11} & \mathbf{A}_{12} & \mathbf{b}_1 \\ \mathbf{A}_{21} & \mathbf{A}_{22} & \mathbf{b}_2 \end{array}\right] \xrightarrow{❶} \left[\begin{array}{cc|c} \mathbf{A}_{11} & \mathbf{A}_{12} & \mathbf{b}_1 \\ 0 & \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12} & \mathbf{b}_2 - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{b}_1 \end{array}\right]$$
$$\xrightarrow{❷} \left[\begin{array}{cc|c} \mathbf{I} & 0 & \mathbf{A}_{11}^{-1}(\mathbf{b}_1 - \mathbf{A}_{12}\mathbf{S}^{-1}\mathbf{b}_S) \\ 0 & \mathbf{I} & \mathbf{S}^{-1}\mathbf{b}_S \end{array}\right];,$$

where we abbreviated $\mathbf{S} := \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}$, a matrix known as Schur complement, see Rem. 2.3.2.19, and $\mathbf{b}_S := \mathbf{b}_2 - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{b}_1$.

We can read off the solution of the block-partitioned linear system from the above Gaussian elimination:

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix} \quad \Rightarrow \quad \begin{aligned} \mathbf{x}_2 &= \mathbf{S}^{-1}\mathbf{b}_S\,, \\ \mathbf{x}_1 &= \mathbf{A}_{11}^{-1}(\mathbf{b}_1 - \mathbf{A}_{12}\mathbf{S}^{-1}\mathbf{b}_S)\,. \end{aligned} \tag{2.3.1.15}$$

⌟

2.3.2 LU-Decomposition

A matrix factorization (*ger.* Matrixzerlegung) expresses a general matrix $\mathbf{A}$ as product of two *special* (factor) matrices. Requirements for these special matrices define the matrix factorization. Matrix factorizations come with the mathematical issue of existence & uniqueness, and pose the numerical challenge of finding algorithms for computing the factor matrices (efficiently and stably).

Matrix factorizations

- ☞ often capture the essence of algorithms in compact form (here: Gaussian elimination),
- ☞ are important building blocks for complex algorithms,
- ☞ are key theoretical tools for algorithm analysis.

In this section the forward elimination step of Gaussian elimination will be related to a special matrix factorization, the so-called LU-decomposition or LU-factorization.

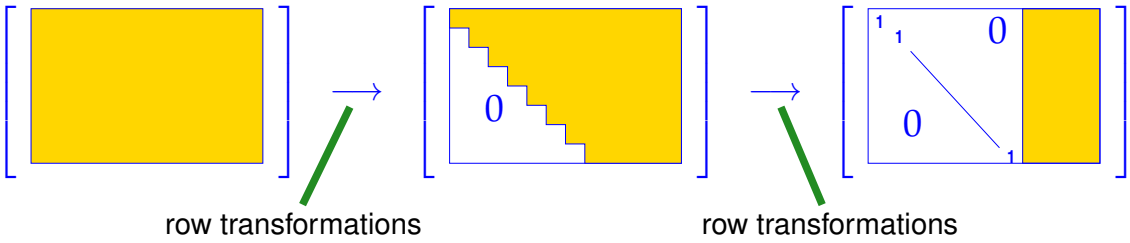
Supplementary literature. The LU-factorization should be well known from the introductory linear algebra course. In case you need to refresh your knowledge, please consult one of the following:

- textbook by Nipp & Stoffer [NS02, Sect. 2.4],
- book by M. Hanke-Bourgeois [Han02, p. II.4],
- linear algebra lecture notes by M. Gutknecht [Gut09, Sect. 3.1],
- textbook by Quarteroni et al. [QSS00, Sect.3.3.1],
- Sect. 3.5 of the book by Dahmen & Reusken,

- Sect. 5.1 of the textbook by Ascher & Greif [AG11].

See also (2.3.2.1) below.

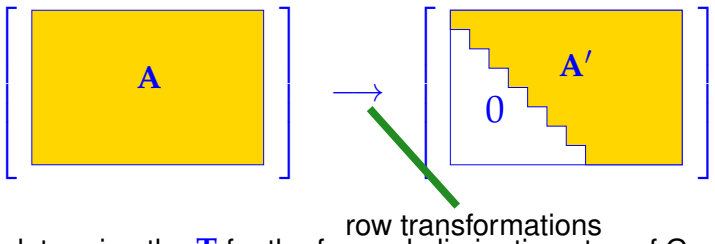
Recall the gist of Gaussian elimination split into the two steps of forward elimination and backsubstitution with (multiple) right-hand-side vectors appended to the coefficient matrix as rightmost columns:



Here: row transformation = adding a multiple of a matrix row to another row, or multiplying a row with a non-zero scalar (number) swapping two rows (more special row transformations are discussed in Rem. 1.3.1.12)

Note: Row transformations preserve regularity of a matrix and, thus, are suitable for transforming linear systems of equations: they will not affect the solution when applied to both the coefficient matrix and right-hand-side vector.

Rem. 1.3.1.12: row transformations can be realized by multiplication *from left* with suitable transformation matrices. When multiplying these transformation matrices we can emulate the effect to successive row transformations through left multiplication with a matrix $\mathbf{T}$:



$$\Leftrightarrow \quad \mathbf{TA} = \mathbf{A}' \,.$$

Now we want to determine the $\mathbf{T}$ for the forward elimination step of Gaussian elimination.

EXAMPLE 2.3.2.1 (Gaussian elimination and LU-factorization → [NS02, Sect. 2.4], [Han02, p. II.4], [Gut09, Sect. 3.1]) We revisit the LSE from Ex. 2.3.1.1 and carry out (forward) Gaussian elimination:

$$\begin{bmatrix} 1 & 1 & 0 \\ 2 & 1 & -1 \\ 3 & -1 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 4 \\ 1 \\ -3 \end{bmatrix} \quad \longleftrightarrow \quad \begin{array}{rcrcrcr} x_1 & + & x_2 & & & = & 4 \\ 2x_1 & + & x_2 & - & x_3 & = & 1 \\ 3x_1 & - & x_2 & - & x_3 & = & -3 \end{array} .$$

$$\begin{bmatrix} 1 & & \\ & 1 & \\ & & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 \\ 2 & 1 & -1 \\ 3 & -1 & -1 \end{bmatrix} \begin{bmatrix} 4 \\ 1 \\ -3 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & & \\ 2 & 1 & \\ & 0 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{1} & 1 & 0 \\ 0 & -1 & -1 \\ 3 & -1 & -1 \end{bmatrix} \begin{bmatrix} 4 \\ -7 \\ -3 \end{bmatrix} \rightarrow$$

$$\begin{bmatrix} 1 & & \\ 2 & 1 & \\ 3 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{1} & 1 & 0 \\ 0 & -1 & -1 \\ 0 & -4 & -1 \end{bmatrix} \begin{bmatrix} 4 \\ -7 \\ -15 \end{bmatrix} \rightarrow \underbrace{\begin{bmatrix} 1 & & \\ 2 & 1 & \\ 3 & 4 & 1 \end{bmatrix}}_{=:\mathbf{L}} \underbrace{\begin{bmatrix} 1 & 1 & 0 \\ 0 & \mathbf{-1} & -1 \\ 0 & 0 & 3 \end{bmatrix}}_{=:\mathbf{U}} \begin{bmatrix} 4 \\ -7 \\ 13 \end{bmatrix}$$

As before we highlight the pivot rows with ▭ and write the pivot element in **bold**. In addition, we let the *negative* multipliers take the places of matrix entries made to vanish; we color these entries red.

After this replacement we make the "surprising" observation that $\mathbf{A} = \mathbf{L}\,\mathbf{U}$! ⌟

The link between Gaussian elimination and matrix factorization, an explanations for the observation made in Ex. 2.3.2.1, becomes clear by recalling that row transformations result from multiplications with elimination matrices:

$$a_1 \neq 0 \quad \blacktriangleright \quad \begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \\ -\frac{a_2}{a_1} & 1 & & & 0 \\ -\frac{a_3}{a_1} & & \ddots & & \\ \vdots & & & \ddots & \\ -\frac{a_n}{a_1} & 0 & & & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} a_1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} . \tag{2.3.2.2}$$

$\blacktriangleright$ $n-1$ steps of Gaussian forward elimination immediately give rise to a matrix factorization (non-zero pivot elements assumed)

$$\mathbf{A} = \mathbf{L}_1 \cdot \cdots \cdot \mathbf{L}_{n-1} \mathbf{U} \ , \quad \text{with} \quad \begin{array}{l} \text{elimination matrices } \mathbf{L}_i,\ i = 1, \ldots, n-1 \ , \\ \text{upper triangular matrix } \mathbf{U} \in \mathbb{R}^{n,n} \ . \end{array}$$

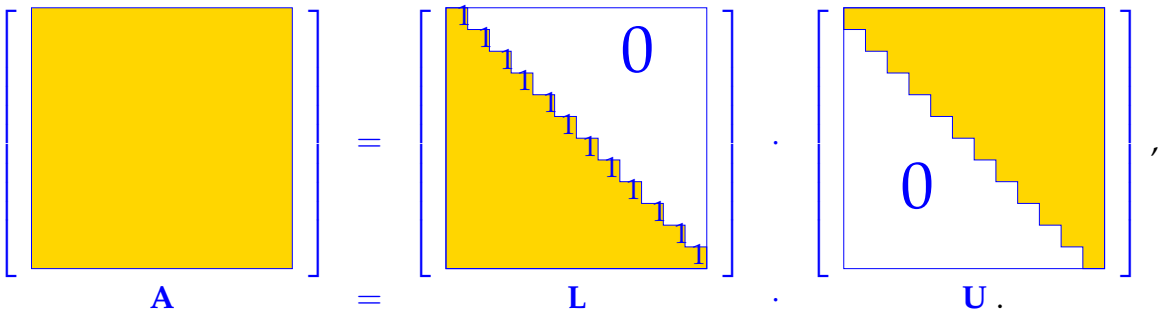
$$\begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \\ l_2 & 1 & & & 0 \\ l_3 & & \ddots & & \\ \vdots & & & \ddots & \\ l_n & 0 & & & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & & & 0 \\ 0 & h_3 & 1 & & \\ \vdots & \vdots & & \ddots & \\ 0 & h_n & 0 & & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \\ l_2 & 1 & & & 0 \\ l_3 & h_3 & 1 & & \\ \vdots & \vdots & & \ddots & \\ l_n & h_n & 0 & & 1 \end{bmatrix}$$

$\blacktriangleright$ The matrix products $\mathbf{L}_1 \cdot \cdots \cdot \mathbf{L}_{n-1}$ yield normalized lower triangular matrices, whose entries are the multipliers $-\frac{a_{ik}}{a_{kk}}$ from (2.3.1.2) $\rightarrow$ Ex. 2.3.1.1.

The matrix factorization that "automatically" emerges during Gaussian forward elimination has a special name:

Definition 2.3.2.3. LU-decomposition/LU-factorization

Given a square matrix $\mathbf{A} \in \mathbb{K}^{n,n}$, an *upper triangular matrix* $\mathbf{U} \in \mathbb{K}^{n,n}$ and a *normalized lower triangular matrix* ($\rightarrow$ Def. 1.1.2.3) form an LU-decomposition/LU-factorization of $\mathbf{A}$, if $\mathbf{A} = \mathbf{LU}$.



Using this notion we can summarize what we have learned from studying elimination matrices:

The (forward) Gaussian elimination (without pivoting), for $\mathbf{Ax} = \mathbf{b}$, $\mathbf{A} \in \mathbb{R}^{n,n}$, if possible, is *algebraically equivalent* to an LU-factorization/LU-decomposition $\mathbf{A} = \mathbf{LU}$ of $\mathbf{A}$ into a normalized lower triangular matrix $\mathbf{L}$ and an upper triangular matrix $\mathbf{U}$, [DR08, Thm. 3.2.1], [NS02, Thm. 2.10], [Gut09, Sect. 3.1].

Algebraically equivalent $\hat{=}$ when carrying out the forward elimination in situ as in Code 2.3.1.4 and storing the multipliers in a lower triangular matrix as in Ex. 2.3.2.1, then the latter will contain the $\mathbf{L}$-factor and the original matrix will be replaced with the $\mathbf{U}$-factor.

Lemma 2.3.2.4. Existence of LU-decomposition

The LU-decomposition of $\mathbf{A} \in \mathbb{K}^{n,n}$ exists, if all submatrices $(\mathbf{A})_{1:k,1:k}$, $1 \le k \le n$, are regular.

Proof. We adopt a block matrix perspective ($\rightarrow$ § 1.3.1.13) and employ induction w.r.t. n:

$n = 1$: assertion trivial

$n-1 \rightarrow n$: Induction hypothesis ensures existence of normalized lower triangular matrix $\widetilde{\mathbf{L}}$ and regular upper triangular matrix $\widetilde{\mathbf{U}}$ such that $\widetilde{\mathbf{A}} = \widetilde{\mathbf{L}}\widetilde{\mathbf{U}}$, where $\widetilde{\mathbf{A}}$ is the upper left $(n-1) \times (n-1)$ block of $\mathbf{A}$:

$$\left[\begin{array}{c|c} \widetilde{\mathbf{A}} & \mathbf{b} \\ \hline \mathbf{a}^\top & \alpha \end{array}\right] = \left[\begin{array}{c|c} \widetilde{\mathbf{L}} & 0 \\ \hline \mathbf{x}^\top & 1 \end{array}\right] \left[\begin{array}{c|c} \widetilde{\mathbf{U}} & \mathbf{y} \\ \hline 0 & \xi \end{array}\right] =: \mathbf{LU} \,.$$

Then solve

❶ $\widetilde{\mathbf{L}}\mathbf{y} = \mathbf{b}$ $\rightarrow$ provides $\mathbf{y} \in \mathbb{K}^n$,

❷ $\mathbf{x}^\top \widetilde{\mathbf{U}} = \mathbf{a}^\top$ $\rightarrow$ provides $\mathbf{x} \in \mathbb{K}^n$,

❸ $\mathbf{x}^\top \mathbf{y} + \xi = \alpha$ $\rightarrow$ provides $\xi \in \mathbb{K}$.

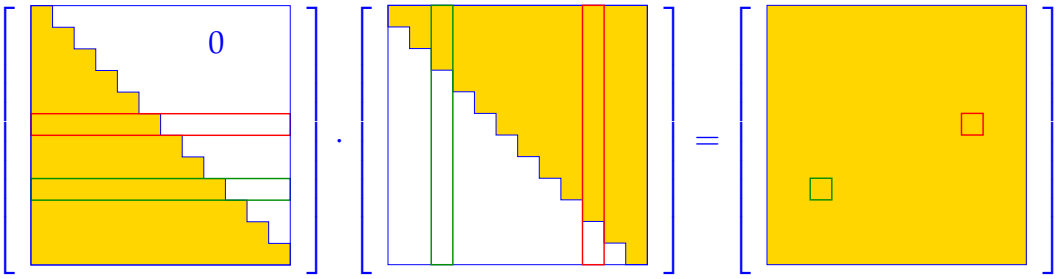
Regularity of $\mathbf{A}$ involves $\xi \neq 0$ (why?) so that $\mathbf{U}$ will be regular, too. □

§2.3.2.5 (Uniqueness of LU-decomposition) Regular upper triangular matrices and normalized lower triangular matrices form *matrix groups* ($\rightarrow$ Lemma 1.3.1.9). Their only common element is the identity matrix.

$$\mathbf{L}_1\mathbf{U}_1 = \mathbf{L}_2\mathbf{U}_2 \quad \Rightarrow \quad \mathbf{L}_2^{-1}\mathbf{L}_1 = \mathbf{U}_2\mathbf{U}_1^{-1} = \mathbf{I} \,.$$

Since inverses of matrices are unique, so are the LU-factors: $\mathbf{U}_1 = \mathbf{U}_2$ and $\mathbf{L}_1 = \mathbf{L}_2$. ⌟

§2.3.2.6 (Basic algorithm for computing LU-decomposition) There are direct ways to determine the factor matrices of the LU-decomposition [Gut09, Sect. 3.1], [QSS00, Sect. 3.3.3] and, of course, they are closely related to forward Gaussian elimination. To derive the algorithm we study the entries of the product of a normalized lower triangular and an upper triangular matrix, see Def. 1.1.2.3:



Taking into account the zero entries known a priori, we arrive at

$$\mathbf{LU} = \mathbf{A} \quad \Rightarrow \quad a_{ik} = \sum_{j=1}^{\min\{i,k\}} l_{ij}u_{jk} = \begin{cases} \sum_{j=1}^{i-1} l_{ij}u_{jk} + 1 \cdot u_{ik} & \text{, if } i \le k \,, \\ \sum_{j=1}^{k-1} l_{ij}u_{jk} + l_{ik}u_{kk} & \text{, if } i > k \,. \end{cases} \tag{2.3.2.7}$$

This reveals how to compute the entries of $\mathbf{L}$ and $\mathbf{U}$ sequentially. We start with the top row of $\mathbf{U}$, which agrees with that of $\mathbf{A}$, and then work our way towards to bottom right corner:

➤
- row by row computation of $\mathbf{U}$
- column by column computation of $\mathbf{L}$

Entries of $\mathbf{A}$ can be replaced with those of $\mathbf{L}$, $\mathbf{U}$!
(so-called in situ/in place computation)

(Crout's algorithm, [Gut09, Alg. 3.1])

≙ rows of $\mathbf{U}$

≙ columns of $\mathbf{L}$

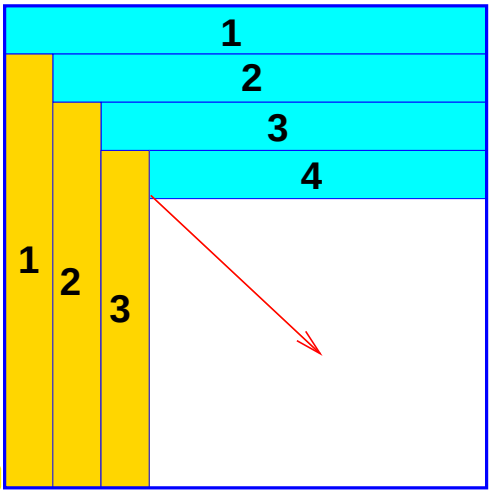


Fig. 40

The following code follows this sequential computation scheme:

C++ code 2.3.2.8: LU-factorization ↪ GITLAB

```
//!  Algorithm of Crout:  LU-factorization of A ∈ K^{n,n}
std::pair<MatrixXd, MatrixXd> lufak(const MatrixXd &A) {
  const Index n = A.rows();
  assert(n == A.cols()); // Ensure matrix is square
  MatrixXd L{MatrixXd::Identity(n, n)};
  MatrixXd U{MatrixXd::Zero(n, n)};
  for (Index k = 0; k < n; ++k) {
    // Compute row of U
    for (Index j = k; j < n; ++j) {
      U(k, j) = A(k, j) - (L.block(k, 0, 1, k) * U.block(0, j, k, 1))(0, 0);
    }
    // Compute column of L
    for (Index i = k + 1; i < n; ++i) {
      L(i, k) = (A(i, k) - (L.block(i, 0, 1, k) * U.block(0, k, k, 1))(0, 0)) /
                U(k, k);
    }
  }
  return { L, U };
}
```

It is instructive to compare this code with a simple implementation of the matrix product of a normalized lower triangular and an upper triangular matrix. From this perspective the LU-factorization looks like the "inversion" of matrix multiplication:

C++ code 2.3.2.9: matrix multiplication $\mathbf{L} \cdot \mathbf{U}$ ↪ GITLAB

```
//!  Multiplication of normalized lower/upper triangular matrices
MatrixXd lumult(const MatrixXd &L, const MatrixXd &U) {
  const Eigen::Index n = L.rows();
  assert(n == L.cols() && n == U.cols() && n == U.rows());
  MatrixXd A{MatrixXd::Zero(n, n)};
  for (Eigen::Index k = 0; k < n; ++k) {
    for (Eigen::Index j = k; j < n; ++j) {
      A(k, j) = U(k, j) + (L.block(k, 0, 1, k) * U.block(0, j, k, 1))(0, 0);
    }
    for (Eigen::Index i = k + 1; i < n; ++i) {
      A(i, k) =
          (L.block(i, 0, 1, k) * U.block(0, k, k, 1))(0, 0) + L(i, k) * U(k, k);
    }
  }
```

```
16   return A;
17 }
```

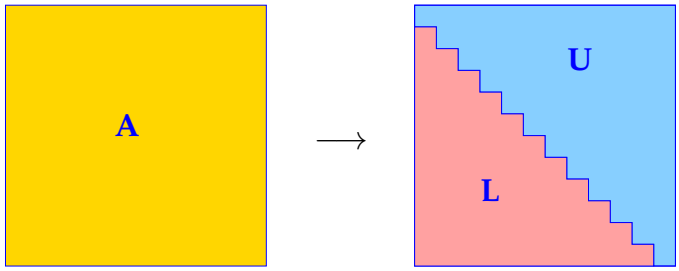
Observe: Solving for entries `L(i,k)` of $\mathbf{L}$ and `U(k,j)` of $\mathbf{U}$ in the multiplication of an upper triangular and normalized lower triangular matrix (→ Code 2.3.2.9) yields the algorithm for LU-factorization (→ Code 2.3.2.8). ⌟

The computational cost of LU-factorization is immediate from Code 2.3.2.8 and the same as for Gaussian elimination, *cf.* § 2.3.1.5:

Asymptotic complexity of LU-factorization of $\mathbf{A} \in \mathbb{R}^{n,n}$

$$= \frac{2}{3}n^3 + O(n^2) = O(n^3) \text{ for } n \to \infty \tag{2.3.2.10}$$

Remark 2.3.2.11 (In-situ LU-decomposition) "In situ" is Latin and means "in place". Many library routines provide routines that *overwrite* the matrix $\mathbf{A}$ with its LU-factors in order to save memory when the original matrix is no longer needed. This is possible because the number of unknown entries of the LU-factors combined exactly agrees with the number of entries of $\mathbf{A}$. The convention is to replace the strict lower-triangular part of $\mathbf{A}$ with $\mathbf{L}$, and the upper triangular part with $\mathbf{U}$:



⌟

Remark 2.3.2.12 (Recursive LU-factorization) Recall Rem. 2.3.1.11 and the recursive view of Gaussian elimination it suggests, because in (2.3.1.12) an analoguous row transformation can be applied to the remaining right-lower block $\mathbf{C}'$.

In light of the close relationship between Gaussian elimination and LU-factorization there will also be a recursive version of LU-factorization.

The following code implements the recursive in situ (in place) LU-decomposition of $\mathbf{A} \in \mathbb{R}^{n,n}$ (without pivoting). It is closely related to Code 2.3.1.13, but now both $\mathbf{L}$ and $\mathbf{U}$ are stored in place of $\mathbf{A}$:

C++ code 2.3.2.13: Recursive LU-factorization ↬ GITLAB

```
2  //!  in situ recursive LU-factorization
3  MatrixXd lurec(const MatrixXd &A){
4       const Eigen::Index n = A.rows();
5       MatrixXd result(n,n);
6       if(n > 1){
7            const VectorXd fac = A.col(0).tail(n-1) / A(0,0);//
8            result.bottomRightCorner(n-1,n-1) = lurec( A.bottomRightCorner(n-1,n-1)
                 - fac * A.row(0).tail(n-1) );//
9            result.row(0) = A.row(0); result.col(0).tail(n-1) = fac;
10           return result;
11      }
12      return A;
13 }
```

Refer to (2.3.1.12) to understand `lurec`: the rank-1 modification of the lower $(n-1)\times(n-1)$-block of the matrix is done in Line 7-Line 8 of the code.

C++ code 2.3.2.14: Driver for recursive LU-factorization of Code 2.3.2.13 ↪ GITLAB

```
2  //!  post-processing:  extract L and U
3  void lurecdriver(const MatrixXd &A, MatrixXd &L, MatrixXd &U){
4       const MatrixXd A_dec = lurec(A);
5       // post-processing:
6       //extract L and U
7       U = A_dec.triangularView<Upper>();
8       L.setIdentity();
9       L += A_dec.triangularView<StrictlyLower>();
10 }
```

§2.3.2.15 (Using LU-factorization to solve a linear system of equations) An intermediate LU-factorization paves the way for a *three-stage* procedure for solving an $n\times n$ linear system of equations.

$\mathbf{Ax}=\mathbf{b}$:
① LU-decomposition $\mathbf{A}=\mathbf{LU}$, #elementary operations $=\frac{1}{3}n(n-1)(n+1)$
② forward substitution, solve $\mathbf{Lz}=\mathbf{b}$, #elementary operations $=\frac{1}{2}n(n-1)$
③ backward substitution, solve $\mathbf{Ux}=\mathbf{z}$, #elementary operations $=\frac{1}{2}n(n+1)$

➢ The asymptotic complexity of the complete three-stage algorithm is (in leading order) the same as for Gaussian elimination (The bulk of computational cost is incurred in the factorization step ①).

However, the perspective of LU-factorization reveals that the solution of linear systems of equations can be split into two separate phases with different asymptotic complexity in terms of the number n of unknowns:

setup phase (factorization) Cost: $O(n^3)$	+	**elimination phase** (forward/backward substition) Cost: $O(n^2)$

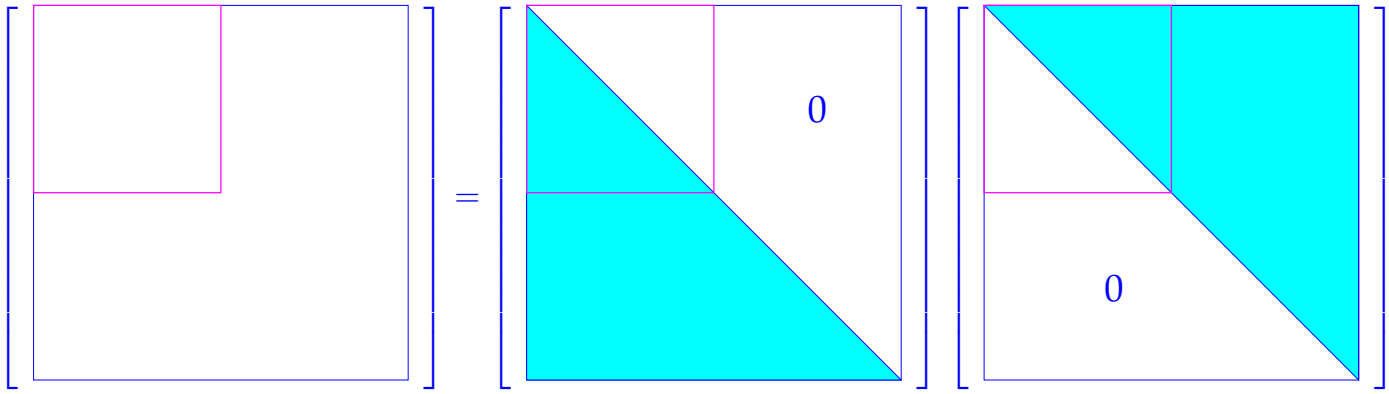
Remark 2.3.2.16 (Rationale for using LU-decomposition in algorithms) Gauss elimination and LU-factorization for the solution of a linear system of equations (→ § 2.3.2.15) are equivalent and only differ in the ordering of the steps.

Then, why is it important to know about LU-factorization?

Because in the case of LU-factorization the expensive forward elimination and the less expensive (forward/backward) substitutions are separated, which sometimes can be exploited to reduce computational cost, as highlighted in Rem. 2.5.0.10 below.

Remark 2.3.2.17 ("'Partial LU-decompositions" of principal minors) The algorithm from § 2.3.2.6 reveals that the the computation of the LU-decomposition of a matrix proceeds from top-left to bottom-right. This implies the locality property discussed in this remark. To understands its heading we remind that a principal minor refers to the left upper block of a matrix

The following "visual rule" help identify the structure of the LU-factors of a matrix.



(2.3.2.18)

The left-upper blocks of both $\mathbf{L}$ and $\mathbf{U}$ in the LU-factorization of $\mathbf{A}$ depend only on the corresponding left-upper block of $\mathbf{A}$! ⌟

Remark 2.3.2.19 (Block LU-factorization) In the spirit of § 1.3.1.13 we can also adopt a matrix-block perspective of LU-factorization. This is a natural idea in light of the close connection between matrix multiplication and matrix factorization, *cf.* the relationship between matrix factorization and matrix multiplication found in § 2.3.2.6:

Block matrix multiplication (1.3.1.14) $\cong$ block LU-decomposition:

We consider a block-partitioned matrix

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \quad , \quad \begin{matrix} \mathbf{A}_{11} \in \mathbb{K}^{n,n} \text{ regular} , & \mathbf{A}_{12} \in \mathbb{K}^{n,m} \\ \mathbf{A}_{21} \in \mathbb{K}^{m,n} , & \mathbf{A}_{22} \in \mathbb{K}^{m,m} . \end{matrix}$$

The block LU-decomposition arises from the block Gaussian forward elimination of Rem. 2.3.1.14 in the same way as the standard LU-decomposition is spawned by the entry-wise Gaussian elimination:

With

$$\underbrace{\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & 0 \\ \mathbf{A}_{21}\mathbf{A}_{11}^{-1} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ 0 & \mathbf{S} \end{bmatrix}}_{\text{block LU-factorization}} \quad , \quad \text{with Schur complement} \quad \boxed{\mathbf{S} := \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}} \, . \tag{2.3.2.20}$$

Under the assumption that $\mathbf{A}_{11}$ is invertible, the Schur complement matrix $\mathbf{S}$ is invertible, if and only if this holds for $\mathbf{A}$. ⌟

Review question(s) 2.3.2.21 (Gaussian elimination and LU-decomposition)

(Q2.3.2.21.A) Performing Gaussian elimination by hand compute the solution of the following 4×4 linear system of equations

$$\begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} .$$

(Q2.3.2.21.B) Give an example of a 2×2-matrix, for which there does **not** exist an LU-decomposition.

(Q2.3.2.21.C) Assume that one of the LU-factors of a square matrix $\mathbf{A} \in \mathbb{R}^{n,n}$ is diagonal. What properties of $\mathbf{A}$ can you infer?

(Q2.3.2.21.D) Suppose the LU-factors $\mathbf{L}, \mathbf{U} \in \mathbb{R}^{n,n}$ of a square matrix $\mathbf{A} \in \mathbb{R}^{n,n}$ exist and have been computed already. Sketch an algorithm for computing the determinant $\det \mathbf{A}$.

From linear algebra remember that the determinant of the product of two square matrices is the product of the determinants of its factors.

(Q2.3.2.21.E) Compute the block LU-decomposition of the partitioned matrix

$$\mathbf{A} = \begin{bmatrix} \mathbf{I}_k & \mathbf{B}^\top \\ \mathbf{B} & \mathbf{O} \end{bmatrix} \in \mathbb{R}^{n,n}, \quad \mathbf{B} \in \mathbb{R}^{n-k,k}, \quad k \in \{1, \ldots, n-1\}.$$

When is this matrix regular?

(Q2.3.2.21.F) Predict the asymptotic computational effort of an efficient algorithm for computing the block LU-decomposition of

$$\mathbf{A} = \begin{bmatrix} \mathbf{I}_k & \mathbf{B}^\top \\ \mathbf{B} & \mathbf{O} \end{bmatrix} \in \mathbb{R}^{n,n}, \quad \mathbf{B} \in \mathbb{R}^{n-k,k}, \quad k \in \{1, \ldots, n-1\},$$

in terms of $n, k \to \infty$.

(Q2.3.2.21.G) What is the inverse of the block matrix

$$\begin{bmatrix} \mathbf{O} & \mathbf{A} \\ \mathbf{A}^\top & \mathbf{O} \end{bmatrix} \in \mathbb{R}^{2n,2n}, \quad \mathbf{A} \in \mathbb{R}^{n,n} \text{ } \textit{regular} \text{ ?}$$

Use block Gaussian elimination to find it and express it in terms of $\mathbf{A}^{-1}$.

(Q2.3.2.21.H) [Schur complement] Consider the following block partitioning of a matrix $\mathbf{A} \in \mathbb{R}^{n+m,n+m}$

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}, \quad \begin{array}{ll} \mathbf{A}_{11} \in \mathbb{R}^{n,n}, & \mathbf{A}_{12} \in \mathbb{R}^{n,m}, \\ \mathbf{A}_{21} \in \mathbb{R}^{m,n}, & \mathbf{A}_{22} \in \mathbb{R}^{m,m}. \end{array}$$

We assume that $\mathbf{A}_{11}$ is *regular*, which renders the **Schur complement**

$$\mathbf{S} := \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12} \in \mathbb{R}^{m,m}$$

well-defined.

Show that $\mathbf{A}$ is singular, if and only if $\mathbf{S}$ is singular :

$$\mathcal{N}(\mathbf{A}) \neq \{\mathbf{0}\} \iff \mathcal{N}(\mathbf{S}) \neq \{\mathbf{0}\}.$$

Hint.

- First show that if $\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \in \mathcal{N}(\mathbf{A})$, then $\mathbf{S}\mathbf{y} = \mathbf{0}$.
- For $\mathbf{y} \in \mathbb{R}^m$ such that $\mathbf{S}\mathbf{y} = \mathbf{0}$ consider $\begin{bmatrix} -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{y} \\ \mathbf{y} \end{bmatrix}$.

△

2.3.3 Pivoting

We know from linear algebra [NS02, Sect. 1.1] that sometimes we have to swap rows of a linear system of equations in order to carry out Gaussian elimination without encountering a division by zero. Here is a

2×2 example:

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \qquad \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_2 \\ b_1 \end{bmatrix}$$

breakdown of Gaussian elimination pivot element $= 0$ — Gaussian elimination feasible

Remedy (in linear algebra): **Pivoting** $\hat{=}$ avoid zero pivot elements by swapping rows.

EXAMPLE 2.3.3.1 (Pivoting and numerical stability → [DR08, Example 3.2.3]) Gaussian elimination for the 2×2 linear system of equations studied in this example will never lead to a division by zero. Nevertheless, Gaussian elimination runs into problems.

```
2  MatrixXd A(2, 2);
3  A << 5.0e-17, 1.0, 1.0, 1.0;
4  VectorXd b(2);
5  VectorXd x2(2);
6  b << 1.0, 2.0;
7  const VectorXd x1 = A.fullPivLu().solve(b);
8  gausselimsolve::gausselimsolve(A, b, x2); // see
       Code 2.3.1.10
9  const auto [L,U] = lufak::lufak(A);    // see
       Code 2.3.2.8
10 const VectorXd z = L.lu().solve(b);
11 const VectorXd x3 = U.lu().solve(z);
12 std::cout << "x1 = \n"
13      << x1 << "\nx2 = \n"
14      << x2 << "\nx3 = \n"
15      << x3 << std::endl;
```

Output:

```
1 x1 =
2 1
3 1
4 x2 =
5 0
6 1
7 x3 =
8 0
9 1
```

We get different results from EIGEN built-in linear solver and out hand-crafted Gaussian elimination! Let's see what we should expect as an "exact solution":

$$\mathbf{A} = \begin{bmatrix} \epsilon & 1 \\ 1 & 1 \end{bmatrix} \quad , \quad \mathbf{b} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \quad \Rightarrow \quad \mathbf{x} = \begin{bmatrix} \dfrac{1}{1-\epsilon} \\ \dfrac{1-2\epsilon}{1-\epsilon} \end{bmatrix} \approx \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \text{for } |\epsilon| \ll 1 \,.$$

What is wrong with EIGEN? To make sense of our observations we have to rely on our insights into roundoff errors gained in Section 1.5.3. Armed with knowledge about the behavior of machine numbers and roundoff errors we can understand what is going on in this example:

❶ We "simulate" floating point arithmetic for straightforward LU-factorization: if $\epsilon \leq \frac{1}{2}\mathrm{EPS}$, $\mathrm{EPS} \hat{=}$ machine precision,

$$\blacktriangleright \quad \mathbf{L} = \begin{bmatrix} 1 & 0 \\ \epsilon^{-1} & 1 \end{bmatrix} \quad , \quad \mathbf{U} = \begin{bmatrix} \epsilon & 1 \\ 0 & 1-\epsilon^{-1} \end{bmatrix} \overset{(*)}{=} \widetilde{\mathbf{U}} := \begin{bmatrix} \epsilon & 1 \\ 0 & -\epsilon^{-1} \end{bmatrix} \quad \text{in } \mathbb{M} \,! \qquad (2.3.3.2)$$

$(*)$: because $1 \widetilde{+} 2/\mathrm{EPS} = 2/\mathrm{EPS}$, see Exp. 1.5.3.14.

$\blacktriangleright$ The solution of $\mathbf{L}\widetilde{\mathbf{U}}\mathbf{x} = \mathbf{b}$ is $\mathbf{x} = \begin{bmatrix} 2\epsilon \\ 1-2\epsilon \end{bmatrix} \approx \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, which is a meaningless result!

❷ Let's conduct an LU-factorization in $\mathbb{M}$ after *swapping rows*:

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ \epsilon & 1 \end{bmatrix} \quad \Rightarrow \quad \mathbf{L} = \begin{bmatrix} 1 & 0 \\ \epsilon & 1 \end{bmatrix} , \quad \mathbf{U} = \begin{bmatrix} 1 & 1 \\ 0 & 1-\epsilon \end{bmatrix} = \widetilde{\mathbf{U}} := \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \quad \text{in } \mathbb{M} \,. \qquad (2.3.3.3)$$

The solution of $\mathbf{L}\widetilde{\mathbf{U}}\mathbf{x} = \mathbf{b}$ is $\mathbf{x} = \begin{bmatrix} 1+2\epsilon \\ 1-2\epsilon \end{bmatrix}$, which is a sufficiently accurate result!

From Section 1.5.5, Def. 1.5.5.19 remember the concept of numerical stability, see also [DR08, Sect. 2.3]. An LU-decomposition computed in $\mathbb{M}$ is stable, if it is the exact LU-decomposition of a slightly perturbed matrix. Is this satisfied for the LU-decompositions obtained in ❶ and ❷?

❶, no row swapping, → (2.3.3.2): $\mathbf{L}\widetilde{\mathbf{U}} = \mathbf{A} + \mathbf{E}$ with $\mathbf{E} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$ unstable !

❷, after row swapping, → (2.3.3.3): $\mathbf{L}\widetilde{\mathbf{U}} = \widetilde{\mathbf{A}} + \mathbf{E}$ with $\mathbf{E} = \begin{bmatrix} 0 & 0 \\ 0 & \epsilon \end{bmatrix}$ stable !

Clearly, swapping rows is necessary for being able to stably compute the LU-decomposition in floating point arithmetic. ⌟

Suitable pivoting is essential for controlling the impact of roundoff errors on Gaussian elimination (→ Section 1.5.5, [NS02, Sect. 2.5])

The main rationale behind pivoting in numerical linear algebra is not to steer clear of division by zero, but to ensure numerical stability of Gaussian elimination.

§2.3.3.4 (Partial pivoting) In linear algebra is was easy to decide when pivoting should be done. We just had to check whether a potential pivot element was equal to zero. The situation is murky in numerical linear algebra because

(i) a test `== 0.0` is meaningless in floating point computations Rem. 1.5.3.15,

(ii) the goal of numerical stability is hard to quantify.

Nevertheless there is a very successful strategy and it is known as **partial pivoting**: Writing $a_{i,j}$ $i,j \in \{k,\ldots,n\}$ for the elements of the intermediate matrix obtained after $k < n$ steps of Gaussian elimination applied to an $n \times n$ LSE, we choose the index j of the next pivot row as follows:

$$j \in \{k,\ldots,n\} \quad \text{such that} \quad \frac{|a_{j,i}|}{\max\{|a_{j,l}|,\ l = k,\ldots,n\}} \to \max \tag{2.3.3.5}$$

In a sense, we choose the *relatively largest* pivot element compared to the other entries in the same row [NS02, Sect. 2.5]. ⌟

EXAMPLE 2.3.3.6 (Gaussian elimination with pivoting for 3×3-matrix) The following sequence of matrices is produced by Gaussian elimination with partial pivoting:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 2 \\ 2 & -3 & 2 \\ 1 & 24 & 0 \end{bmatrix} \overset{❶}{\to} \begin{bmatrix} 2 & -3 & 2 \\ 1 & 2 & 2 \\ 1 & 24 & 0 \end{bmatrix} \overset{❷}{\to} \begin{bmatrix} 2 & -3 & 2 \\ 0 & 3.5 & 1 \\ 0 & 25.5 & -1 \end{bmatrix} \overset{❸}{\to} \begin{bmatrix} 2 & -7 & 2 \\ 0 & 25.5 & -1 \\ 0 & 3.5 & 1 \end{bmatrix} \overset{❹}{\to} \begin{bmatrix} 2 & -7 & 2 \\ 0 & 25.5 & -1 \\ 0 & 0 & 1.373 \end{bmatrix}$$

❶: swap rows 1 & 2.
❷: elimination with top row as pivot row
❸: swap rows 2 & 3
❹: elimination with 2nd row as pivot row ⌟

§2.3.3.7 (Algorithm: Gaussian elimination with partial pivoting)

C++ code 2.3.3.8: Gaussian elimination with pivoting: extension of Code 2.3.1.4 ↪ GITLAB

```
2  //!  Solving an LSE Ax = b by Gaussian elimination with partial pivoting
3  //!  A must be an n×n-matrix, b an n-vector
4  void gepiv(const MatrixXd &A, const VectorXd& b, VectorXd& x){
5      const Eigen::Index n = A.rows();
6      MatrixXd Ab(n,n+1);
7      Ab << A, b;    //
8      // Forward elimination by rank-1 modification, see Rem. 2.3.1.11
9      for(Eigen::Index k = 0; k < n-1; ++k){
10         Eigen::Index j = -1; // j = pivot row index
11         // p = relatively largest pivot
12         const double p = ( Ab.col(k).tail(n-k).cwiseAbs().cwiseQuotient(
               Ab.block(k,k,n-k,n-k).cwiseAbs().rowwise().maxCoeff() )
               ).maxCoeff(&j);//
13         if( p <  std::numeric_limits<double>::epsilon() *
               Ab.block(k,k,n-k,n-k).norm() ) {
14           throw std::logic_error("nearly singular");//
15         }
16         Ab.row(k).tail(n-k+1).swap(Ab.row(k+j).tail(n-k+1));//
17         Ab.bottomRightCorner(n-k-1,n-k) -= Ab.col(k).tail(n-k-1) *
               Ab.row(k).tail(n-k) / Ab(k,k);//
18     }
19     // Back substitution (same as in Code 2.3.1.4)
20     Ab(n-1,n) = Ab(n-1,n) / Ab(n-1,n-1);
21     for(Eigen::Index i = n-2; i >= 0; --i){
22         for(Eigen::Index l = i+1; l < n; ++l){
23             Ab(i,n) -= Ab(l,n)*Ab(i,l);
24         }
25         Ab(i,n) /= Ab(i,i);
26     }
27     x = Ab.rightCols(1); //
28 }
```

(The pivot row index j is chosen in Line 12 of the code.)

Explanations to Code 2.3.3.8:

Line 7: Augment matrix $\mathbf{A}$ by right hand side vector $\mathbf{b}$, see comments on Code 2.3.1.4 for explanations.

Line 12: Select index j for pivot row according to the recipe of partial pivoting, see (2.3.3.5).

Note: Inefficient implementation above (too many comparisons)! Try to do better!

Line 14: If the pivot element is still very small relative to the norm of the matrix, then we have encountered an entire column that is close to zero. Gaussian elimination may not be possible in a stable fashion for this matrix; warn user and terminate.

Line 16: A way to swap rows of a matrix in EIGEN.

Line 17: Forward elimination by means of rank-1-update, see (2.3.1.12).

Line 27: As in Code 2.3.1.4: after back substitution last column of augmented matrix supplies solution $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$.

⌟

§2.3.3.9 (Algorithm: LU-factorization with pivoting) Recall the close relationship between Gaussian elimination and LU-factorization

➢ LU-factorization with pivoting? Of course, just by rearranging the operations of Gaussian forward elimination with pivoting.

EIGEN-based code for in place LU-factorization of $\mathbf{A} \in \mathbb{R}^{n,n}$ with partial pivoting:

C++ code 2.3.3.10: LU-factorization with partial pivoting ↪ GITLAB

```
2  void lupiv(MatrixXd &A){//insitu
3      const Eigen::Index n = A.rows();
4      for(int k = 0; k < n-1; ++k){
5          int j = -1; //j = pivot row index
6          // p = relativly largest pivot
7          const double p = ( A.col(k).tail(n-k).cwiseAbs().cwiseQuotient(
               A.block(k,k,n-k,n-k).cwiseAbs().rowwise().maxCoeff() ) ).maxCoeff(&j);
               //
8          if( p <  std::numeric_limits<double>::epsilon() *
               A.block(k,k,n-k,n-k).norm() ) { //
9            throw std::logic_error("nearly singular");
10         }
11         A.row(k).tail(n-k-1).swap(A.row(k+j).tail(n-k-1));//
12         const  VectorXd fac = A.col(k).tail(n-k-1) / A(k,k);//
13         A.bottomRightCorner(n-k-1,n-k-1) -= fac * A.row(k).tail(n-k-1);//
14         A.col(k).tail(n-k-1) = fac;//
15     }
16 }
```

Notice that the recursive call is omitted as in Rem. 2.3.1.11.

Explanations to Code 2.3.3.10:

Line 7: Find the *relatively largest* pivot element p and the index j of the corresponding row of the matrix, see (2.3.3.5)

Line 8: If the pivot element is still very small relative to the norm of the matrix, then we have encountered an entire column that is close to zero. The matrix is (close to) singular and LU-factorization does not exist.

Line 11: Swap the first and the j-th row of the matrix.

Line 12: Initialize the vector of multiplier.

Line 13: Call the routine for the lower right $(n-1) \times (n-1)$-block of the matrix after subtracting suitable multiples of the first row from the other rows, *cf.* Rem. 2.3.1.11 and Rem. 2.3.2.12.

Line 14: Reassemble the parts of the LU-factors. The vector of multipliers yields a column of $\mathbf{L}$, see Ex. 2.3.2.1. ⌟

Remark 2.3.3.11 (Rationale for partial pivoting policy (2.3.3.5) → **[NS02, Page 47])** Why do we select the *relatively* largest pivot element in (2.3.3.5)? Because we aim for an algorithm for Gaussian elimination/LU-decomposition that possesses the highly desirable scale-invariance property. Loosely speaking, the algorithm should not make different decisions on pivoting when we multiply the LSE with a regular diagonal matrix from the left. Let us take a closer look at a 2×2 example:

Scale linear system of equations from Ex. 2.3.3.1:

$$\begin{bmatrix} 2/\epsilon & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \epsilon & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 & 2/\epsilon \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2/\epsilon \\ 2 \end{bmatrix} := \widetilde{\mathbf{b}}$$

No row swapping would be triggered, if absolutely largest pivot element was used to select the pivot row:

$$\begin{bmatrix} 2 & 2/\epsilon \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 2 & 2/\epsilon \\ 0 & 1-2/\epsilon \end{bmatrix} \doteq \underbrace{\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}}_{\widetilde{\mathbf{L}}} \underbrace{\begin{bmatrix} 2 & 2/\epsilon \\ 0 & -2/\epsilon \end{bmatrix}}_{\widetilde{\mathbf{U}}} \quad \text{in } \mathbb{M} \,.$$

Using the rules of arithmetic in $\mathbb{M}$ ($\rightarrow$ Exp. 1.5.3.14), we find

$$\widetilde{\mathbf{U}}^{-1}\widetilde{\mathbf{L}}^{-1}\widetilde{\mathbf{b}} \doteq -\frac{1}{2}\begin{bmatrix} -1 & -1 \\ 0 & \epsilon \end{bmatrix} \frac{2}{\epsilon} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \doteq \begin{bmatrix} 0 \\ 1 \end{bmatrix} ,$$

which is not an acceptable result.

⌟

§2.3.3.12 (Theory of pivoting) We view pivoting from the perspective of matrix operations and start with a matrix view of row swapping.

Definition 2.3.3.13. Permutation matrix

An n-permutation, $n \in \mathbb{N}$, is a bijective mapping $\pi : \{1,\ldots,n\} \mapsto \{1,\ldots,n\}$. The corresponding permutation matrix $\mathbf{P}_\pi \in \mathbb{K}^{n,n}$ is defined by

$$(\mathbf{P}_\pi)_{ij} = \begin{cases} 1 & \text{, if } j = \pi(i) \,, \\ 0 & \text{else.} \end{cases}$$

Example: permutation $(1,2,3,4) \mapsto (1,3,2,4) \;\widehat{=}\; \mathbf{P} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$.

Note:
- $\mathbf{P}^\top = \mathbf{P}^{-1}$ for any permutation matrix $\mathbf{P}$ ($\rightarrow$ permutation matrices orthogonal/unitary)
- left-multiplication $\mathbf{P}_\pi\mathbf{A}$ effects π-permutation *of rows* of $\mathbf{A} \in \mathbb{K}^{n,m}$
- right-multiplication $\mathbf{A}\mathbf{P}_\pi$ effects π-permutation *of columns* of $\mathbf{A} \in \mathbb{K}^{m,n}$

Lemma 2.3.3.14. Existence of LU-factorization with pivoting $\rightarrow$ [DR08, Thm. 3.25], [Han02, Thm. 4.4]

For any regular $\mathbf{A} \in \mathbb{K}^{n,n}$ there is a permutation matrix ($\rightarrow$ Def. 2.3.3.13) $\mathbf{P} \in \mathbb{K}^{n,n}$, a normalized lower triangular matrix $\mathbf{L} \in \mathbb{K}^{n,n}$, and a regular upper triangular matrix $\mathbf{U} \in \mathbb{K}^{n,n}$ ($\rightarrow$ Def. 1.1.2.3), such that $\mathbf{PA} = \mathbf{LU}$.

Proof. (by induction)

Every regular matrix $\mathbf{A} \in \mathbb{K}^{n,n}$ admits a row permutation encoded by the permutation matrix $\mathbf{P} \in \mathbb{K}^{n,n}$, such that $\mathbf{A}' := (\mathbf{A})_{1:n-1,1:n-1}$ is regular (why ?).

By induction assumption there is a permutation matrix $\mathbf{P}' \in \mathbb{K}^{n-1,n-1}$ such that $\mathbf{P}'\mathbf{A}'$ possesses a LU-factorization $\mathbf{A}' = \mathbf{L}'\mathbf{U}'$. There are $\mathbf{x},\mathbf{y} \in \mathbb{K}^{n-1}$, $\gamma \in \mathbb{K}$ such that

$$\begin{bmatrix} \mathbf{P}' & 0 \\ 0 & 1 \end{bmatrix} \mathbf{PA} = \begin{bmatrix} \mathbf{P}' & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{A}' & \mathbf{x} \\ \mathbf{y}^\top & \gamma \end{bmatrix} = \begin{bmatrix} \mathbf{L}'\mathbf{U}' & \mathbf{x} \\ \mathbf{y}^\top & \gamma \end{bmatrix} = \begin{bmatrix} \mathbf{L}' & 0 \\ \mathbf{c}^\top & 1 \end{bmatrix} \begin{bmatrix} \mathbf{U} & \mathbf{d} \\ 0 & \alpha \end{bmatrix} ,$$

if we choose

$$\mathbf{d} = (\mathbf{L}')^{-1}\mathbf{x} \quad , \quad \mathbf{c} = (\mathbf{u}')^{-T}\mathbf{y} \quad , \quad \alpha = \gamma - \mathbf{c}^{\top}\mathbf{d} \, ,$$

which is always possible. □ ⌟

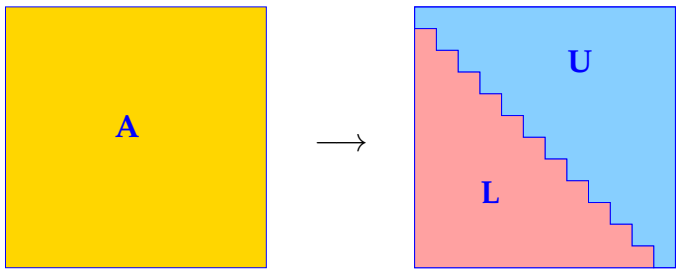
EXAMPLE 2.3.3.15 (Ex. 2.3.3.6 cnt'd) Let us illustrate the assertion of Lemma 2.3.3.14 for the small 3×3 LSE from Ex. 2.3.3.6:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 2 \\ 2 & -3 & 2 \\ 1 & 24 & 0 \end{bmatrix} \overset{❶}{\to} \begin{bmatrix} 2 & -3 & 2 \\ 1 & 2 & 2 \\ 1 & 24 & 0 \end{bmatrix} \overset{❷}{\to} \begin{bmatrix} 2 & -3 & 2 \\ 0 & 3.5 & 1 \\ 0 & 25.5 & -1 \end{bmatrix} \overset{❸}{\to} \begin{bmatrix} 2 & -7 & 2 \\ 0 & 25.5 & -1 \\ 0 & 3.5 & 1 \end{bmatrix} \overset{❹}{\to} \begin{bmatrix} 2 & -7 & 2 \\ 0 & 25.5 & -1 \\ 0 & 0 & 1.373 \end{bmatrix}$$

$$\mathbf{U} = \begin{bmatrix} 2 & -3 & 2 \\ 0 & 25.5 & -1 \\ 0 & 0 & 1.1373 \end{bmatrix}, \quad \mathbf{L} = \begin{bmatrix} 1 & 0 & 0 \\ 0.5 & 1 & 0 \\ 0.5 & 0.1373 & 1 \end{bmatrix}, \quad \mathbf{P} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}.$$

Two permutations: in step ❶ swap rows #1 and #2, in step ❸ swap rows #2 and #3. Apply these swaps to the identity matrix and you will recover $\mathbf{P}$. See also [DR08, Ex. 3.30]. ⌟

§2.3.3.16 (LU-decomposition in EIGEN) EIGEN provides various functions for computing the LU-decomposition of a given matrix. They all perform the factorization in-situ → Rem. 2.3.2.11:



The resulting matrix can be retrieved and used to recover the LU-factors, as demonstrated in the next code snippet. Note that the method `matrixLU` returns just a single matrix, from which the LU-factors have to be extracted using special view methods.

C++ code 2.3.3.17: Performing explicit LU-factorization in EIGEN ↬ GITLAB

```
2  const Eigen::MatrixXd::Index n = A.cols();
3  assert(n == A.rows()); // ensure square matrix
4  const Eigen::PartialPivLU<MatrixXd> lu(A);
5  // Normalized lower-triangule factor
6  MatrixXd L = MatrixXd::Identity(n, n);
7  L.triangularView<StrictlyLower>() += lu.matrixLU();
8  // Upper triangular factor
9  const MatrixXd U = lu.matrixLU().triangularView<Upper>();
10 // Permutation matrix, see Def. 2.3.3.13
11 const MatrixXd P = lu.permutationP();
```

Note that for solving a linear system of equations by means of LU-decomposition (the standard algorithm) we never have to extract the LU-factors. ⌟

Remark 2.3.3.18 (Row swapping commutes with forward elimination) Any kind of pivoting only involves comparisons and row/column permutations, but no arithmetic operations on the matrix entries. This makes the following observation plausible:

The LU-factorization of $\mathbf{A} \in \mathbb{K}^{n,n}$ with partial pivoting by § 2.3.3.9 is *numerically equivalent* to the LU-factorization of $\mathbf{PA}$ without pivoting ($\rightarrow$ Code in § 2.3.2.6), when $\mathbf{P}$ is a permutation matrix gathering the row swaps entailed by partial pivoting.

numerically equivalent $\hat{=}$ same result when executed with the same machine arithmetic

The above statement means that whenever we study the impact of roundoff errors on LU-factorization it is safe to consider only the basic version without pivoting, because we can always assume that row swaps have been conducted beforehand. ⌟

2.4 Stability of Gaussian Elimination

It will turn out that when investigating the stability of algorithms meant to solve linear systems of equations, a key quantity is the residual.

Definition 2.4.0.1. Residual

Given an approximate solution $\widetilde{\mathbf{x}} \in \mathbb{K}^n$ of the LSE $\mathbf{Ax} = \mathbf{b}$ ($\mathbf{A} \in \mathbb{K}^{n,n}$, $\mathbf{b} \in \mathbb{K}^n$), its residual is the vector

$$\mathbf{r} = \mathbf{b} - \mathbf{A}\widetilde{\mathbf{x}} .$$

§2.4.0.2 (Probing stability of a direct solver for LSE) Assume that you have downloaded a direct solver for a general (dense) linear system of equations $\mathbf{Ax} = \mathbf{b}$, $\mathbf{A} \in \mathbb{K}^{n,n}$ regular, $\mathbf{b} \in \mathbb{K}^n$. When given the data $\mathbf{A}$ and $\mathbf{b}$ it returns the perturbed solution $\widetilde{\mathbf{x}}$. How can we tell that $\widetilde{\mathbf{x}}$ is the exact solution of a linear system with slightly perturbed data (in the sense of a tiny relative error of size $\approx$ EPS, EPS the machine precision, see § 1.5.3.8). That is, how can we tell that $\widetilde{\mathbf{x}}$ is an acceptable solution in the sense of backward error analysis, *cf.* Def. 1.5.5.19:

Definition 1.5.5.19. Stable algorithm

An algorithm $\widetilde{F}$ for solving a problem $F : X \mapsto Y$ is numerically stable if for all $\mathbf{x} \in X$ its result $\widetilde{F}(\mathbf{x})$ (possibly affected by roundoff) is the exact result for "slightly perturbed" data:

$$\exists C \approx 1: \quad \forall \mathbf{x} \in X: \quad \exists \widetilde{\mathbf{x}} \in X: \quad \|\mathbf{x} - \widetilde{\mathbf{x}}\|_X \leq C w(\mathbf{x}) \,\mathrm{EPS} \|\mathbf{x}\|_X \quad \wedge \quad \widetilde{F}(\mathbf{x}) = F(\widetilde{\mathbf{x}}) .$$

A question similar to the one we ask now for Gaussian elimination was answered in Ex. 1.5.5.20 for the operation of matrix×vector multiplication.

We can alter either side of the linear system of equations in order to restore $\widetilde{\mathbf{x}}$ as solution:

❶ $\mathbf{x} - \widetilde{\mathbf{x}}$ accounted for by perturbation of right hand side:

$$\begin{aligned} \mathbf{Ax} &= \mathbf{b} \\ \mathbf{A}\widetilde{\mathbf{x}} &= \mathbf{b} + \Delta\mathbf{b} \end{aligned} \quad \Rightarrow \quad \Delta\mathbf{b} = \mathbf{A}\widetilde{\mathbf{x}} - \mathbf{b} =: -\mathbf{r} \quad \text{(residual, Def. 2.4.0.1)} .$$

Hence, $\widetilde{\mathbf{x}}$ can be accepted as a solution, if $\dfrac{\|\mathbf{r}\|}{\|\mathbf{b}\|} \leq Cn^3 \cdot \mathrm{EPS}$, for some small constant $C \approx 1$, see Def. 1.5.5.19. Here, $\|\cdot\|$ can be *any* vector norm on $\mathbb{K}^n$.

❷ $\mathbf{x} - \widetilde{\mathbf{x}}$ accounted for by perturbation of system matrix:

$$\mathbf{A}\mathbf{x} = \mathbf{b} \quad , \quad (\mathbf{A} + \Delta\mathbf{A})\widetilde{\mathbf{x}} = \mathbf{b}$$

[try perturbation $\Delta\mathbf{A} = \mathbf{u}\widetilde{\mathbf{x}}^{\mathrm{H}}$, $\mathbf{u} \in \mathbb{K}^n$]

$$\blacktriangleright \quad \mathbf{u} = \frac{\mathbf{r}}{\|\mathbf{x}\|_2^2} \quad \Rightarrow \quad \Delta\mathbf{A} = \frac{\mathbf{r}\widetilde{\mathbf{x}}^{\mathrm{H}}}{\|\mathbf{x}\|_2^2} \, .$$

As in Ex. 1.5.5.20 we find

$$\frac{\|\Delta\mathbf{A}\|_2}{\|\mathbf{A}\|_2} = \frac{\|\mathbf{r}\|}{\|\mathbf{A}\|_2\|\widetilde{\mathbf{x}}\|_2} \leq \frac{\|\mathbf{r}\|_2}{\|\mathbf{A}\widetilde{\mathbf{x}}\|_2} \, . \tag{2.4.0.3}$$

Thus, $\widetilde{\mathbf{x}}$ is ok in the sense of backward error analysis, if $\dfrac{\|\mathbf{r}\|}{\|\mathbf{A}\widetilde{\mathbf{x}}\|} \leq Cn^3 \cdot \mathrm{EPS}$.

A stable algorithm for solving an LSE yields a residual $\mathbf{r} := \mathbf{b} - \mathbf{A}\widetilde{\mathbf{x}}$ small (in norm) relative to $\mathbf{b}$.

⌟

Now that we know when to accept a vector as solution of a linear system of equations, we can explore whether an implementation of Gaussian elimination (with some pivoting strategy) in floating point arithmetic actually delivers acceptable solutions. Given the several levels of nested loops occurring in algorithms for Gaussian elimination, it is not surprising that the roundoff error analysis of Gaussian elimination based on Ass. 1.5.3.11 is rather involved. Here we merely summarise the results:

The analysis can be simplified by using the fact that equivalence of Gaussian elimination and LU-factorization extends to machine arithmetic, *cf.* Section 2.3.2

Lemma 2.4.0.4. Equivalence of Gaussian elimination and LU-factorization

The following algorithms for solving the LSE $\mathbf{A}\mathbf{x} = \mathbf{b}$ ($\mathbf{A} \in \mathbb{K}^{n,n}$, $\mathbf{b} \in \mathbb{K}^n$) are numerically equivalent:

1. *Gaussian elimination (forward elimination and back substitution) without pivoting, see Algorithm 2.3.1.3.*
2. *LU-factorization of $\mathbf{A}$ ($\rightarrow$ Code in § 2.3.2.6) followed by forward and backward substitution, see Algorithm 2.3.2.15.*

In Rem. 2.3.3.18 we learned that pivoting can be taken into account by a prior permutation of the rows of the linear system of equations. Since permutations do not introduce any roundoff errors, it is thus sufficient to consider LU-factorization without pivoting.

A profound roundoff analysis of Gaussian elimination/LU-factorization can be found in [GV89, Sect. 3.3 & 3.5] and [Hig02, Sect. 9.3]. A less rigorous, but more lucid discussion is given in [TB97, Lecture 22]. Here we only quote a result due to Wilkinson, [Hig02, Thm. 9.5]:

Theorem 2.4.0.5. Stability of Gaussian elimination with partial pivoting

Let $\mathbf{A} \in \mathbb{R}^{n,n}$ be regular and $\mathbf{A}^{(k)} \in \mathbb{R}^{n,n}$, $k = 1, \ldots, n-1$, denote the intermediate matrix arising in the k-th step of § 2.3.3.9 (Gaussian elimination with partial pivoting) when carried out with exact arithmetic.
For the approximate solution $\widetilde{\mathbf{x}} \in \mathbb{R}^n$ of the LSE $\mathbf{Ax} = \mathbf{b}$, $\mathbf{b} \in \mathbb{R}^n$, computed as in § 2.3.3.9 (based on machine arithmetic with machine precision EPS, → Ass. 1.5.3.11) there is $\Delta\mathbf{A} \in \mathbb{R}^{n,n}$ with

$$\|\Delta\mathbf{A}\|_\infty \leq n^3 \frac{3\mathrm{EPS}}{1-3n\mathrm{EPS}}\, \rho\, \|\mathbf{A}\|_\infty\,, \quad \rho := \frac{\max\limits_{i,j,k} |(\mathbf{A}^{(k)})_{ij}|}{\max\limits_{i,j} |(\mathbf{A})_{ij}|}\,,$$

such that $(\mathbf{A} + \Delta\mathbf{A})\widetilde{\mathbf{x}} = \mathbf{b}$.

ρ "small" ➥ Gaussian elimination with partial pivoting is stable (→ Def. 1.5.5.19)

If ρ is "small", the computed solution of a LSE can be regarded as the exact solution of a LSE with "slightly perturbed" system matrix (perturbations of size $O(n^3\mathrm{EPS})$).

Bad news: exponential growth $\rho \sim 2^n$ is possible !

EXAMPLE 2.4.0.6 (Wilkinson's counterexample) We confirm the bad news by means of a famous example, known as the so-called Wilkinson matrix.

$$a_{ij} = \begin{cases} 1 & \text{, if } i = j \vee j = n\,, \\ -1 & \text{, if } i > j\,, \\ 0 & \text{else.} \end{cases}\,, \qquad \text{n=10:} \quad \mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ -1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ -1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ -1 & -1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 & 1 \\ -1 & -1 & -1 & -1 & -1 & 1 & 0 & 0 & 0 & 1 \\ -1 & -1 & -1 & -1 & -1 & -1 & 1 & 0 & 0 & 1 \\ -1 & -1 & -1 & -1 & -1 & -1 & -1 & 1 & 0 & 1 \\ -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & 1 \end{bmatrix}$$

Partial pivoting does not trigger row permutations !

$$\blacktriangleright \quad \mathbf{A} = \mathbf{LU}\,, \quad l_{ij} = \begin{cases} 1 & \text{, if } i = j\,, \\ -1 & \text{, if } i > j\,, \\ 0 & \text{else} \end{cases} \quad u_{ij} = \begin{cases} 1 & \text{, if } i = j\,, \\ 2^{i-1} & \text{, if } j = n\,, \\ 0 & \text{else.} \end{cases}$$

$\blacktriangleright$ We find an exponential blow-up of entries of $\mathbf{U}$!

The following code relies on EIGEN's Gaussian elimination solver and applies it to the Wilkinson matrix:

C++ code 2.4.0.7: Gaussian elimination for "Wilkinson system" in EIGEN ⇝ GITLAB

```
2 MatrixXd res(100,2);
```

```
for(int n = 10; n <= 100*10; n += 10){
      MatrixXd A(n,n); A.setIdentity();
      A.triangularView<StrictlyLower>().setConstant(-1);
      A.rightCols<1>().setOnes();
      VectorXd x = VectorXd::Constant(n,-1).binaryExpr( VectorXd::LinSpaced(n,1,n),
          [](double x, double y){return pow(x,y);});
      double relerr = (A.lu().solve(A*x)-x).norm()/x.norm();
      res(n/10-1,0) = n; res(n/10-1,1) = relerr;
}
// ... different solver(e.g. colPivHouseholderQr()), plotting
```

The measured relative errors are displayed in the following plots alongside the Euclidean condition numbers of the Wilkinson matrices.

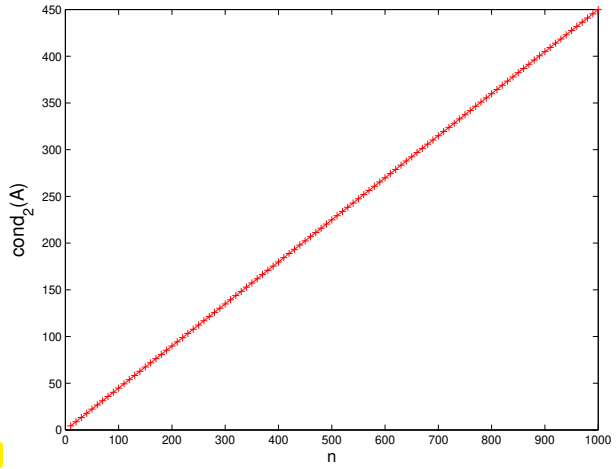


Fig. 41

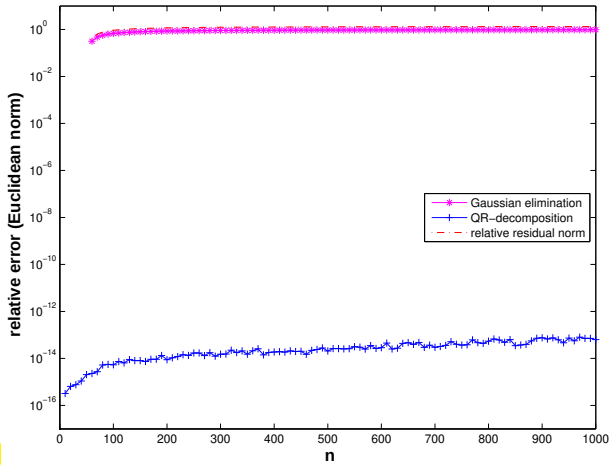


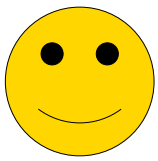
Fig. 42

Large relative errors in solution,
$\updownarrow (*)$
though $\mathrm{cond}_2(\mathbf{A})$ is small!

Evidence of Instability of Gaussian elimination!

$(*)$ If $\mathrm{cond}_2(\mathbf{A})$ was huge, then big errors in the solution of a linear system can be caused by small perturbations of either the system matrix or the right hand side vector, see (2.4.0.13) and the message of Thm. 2.2.2.4, (2.2.2.8). In this case, a stable algorithm can obviously produce a grossly "wrong" solution, as was already explained after (2.2.2.8).
Hence, lack of stability of Gaussian elimination will only become apparent for linear systems with well-conditioned system matrices.

The observations made in this example match Thm. 2.4.0.5, because due to the exponential blow-up of the entries of $\mathbf{U}$ with increasing matrix size we encounter an exponential growth of $\rho = \rho(n)$, see Ex. 2.4.0.6. ⌟



Observation: In practice ρ (almost) always grows only mildly (like $O(\sqrt{n})$) with n

Discussion in [TB97, Lecture 22]: growth factors larger than the order$O(\sqrt{n})$ are exponentially rare in certain relevant classes of *random matrices*.

EXAMPLE 2.4.0.8 (Stability by small random perturbations) Spielman and Teng [ST96] discovered that a tiny relative random perturbation of the Wilkinson matrix on the scale of the machine precision EPS (→ § 1.5.3.8) already remedies the instability of Gaussian elimination.

C++ code 2.4.0.9: Stabilization of Gaussian elimination with partial pivoting by small random perturbations ↬ GITLAB

```
2  //!  Curing Wilkinson's counterexample by random perturbation
3  MatrixXd res(20, 3);
4  mt19937 gen(42); // seed
5  // normal distribution, mean = 0.0, stddev = 1.0
6  std::normal_distribution<> bellcurve;
7  for (int n = 10; n <= 10 * 20; n += 10) {
8    // Build Wilkinson matrix
9    MatrixXd A(n, n); A.setIdentity();
10   A.triangularView<StrictlyLower>().setConstant(-1);
11   A.rightCols<1>().setOnes();
12   // imposed solution
13   VectorXd x = VectorXd::Constant(n, -1).binaryExpr(
14       VectorXd::LinSpaced(n, 1, n),
15       [](double x, double y) { return pow(x, y); });
16   double relerr = (A.lu().solve(A * x) - x).norm() / x.norm();
17   // Randomly perturbed Wilkinson matrix by matrix with iid
18   // N(0,eps) distributed entries
19   MatrixXd Ap = A.unaryExpr([&](double x) {
20     return x + numeric_limits<double>::epsilon() * bellcurve(gen);
21   });
22   double relerrp = (Ap.lu().solve(Ap * x) - x).norm() / x.norm();
23   res(n / 10 - 1, 0) = n;
24   res(n / 10 - 1, 1) = relerr;
25   res(n / 10 - 1, 2) = relerrp;
26 }
```

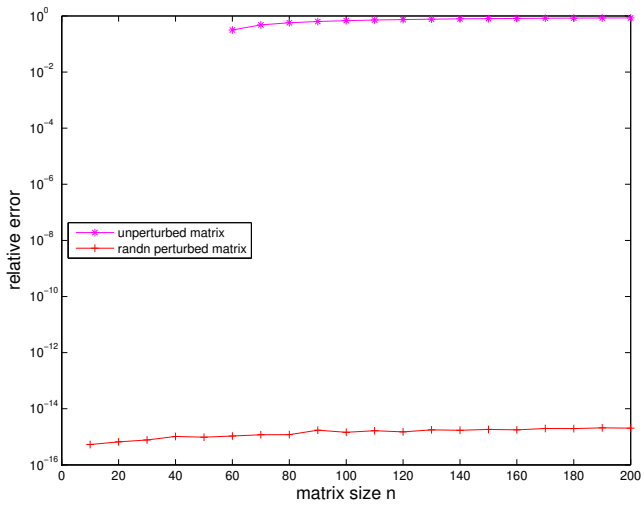


Fig. 43

Recall the statement made above about "improbability" of matrices for which Gaussian elimination with partial pivoting is unstable. This is now matched by the observation that a *tiny random* perturbation of the matrix (almost certainly) cures the problem. This is investigated by the brand-new field of smoothed analysis of numerical algorithms, see [SST06].

Gaussian elimination/LU-factorization with partial pivoting is stable $(*)$
(for all practical purposes) !

$(*)$: stability refers to maximum norm $\|\cdot\|_\infty$.

EXPERIMENT 2.4.0.10 (Conditioning and relative error → Exp. 2.4.0.11 cnt'd)

In the discussion of numerical stability (→ Def. 1.5.5.19, Rem. 1.5.5.22) we have seen that a stable algorithm may produce results with large errors for ill-conditioned problems. The conditioning of the problem of solving a linear system of equations is determined by the condition number (→ Def. 2.2.2.7) of the system

matrix, see Thm. 2.2.2.4.

Hence, for an ill-conditioned linear system, whose system matrix has a huge condition number, (stable) Gaussian elimination may return "solutions" with large errors. This will be demonstrated in this experiment.

Numerical experiment with *nearly singular matrix* from Exp. 2.4.0.11

$$\mathbf{A} = \mathbf{u}\mathbf{v}^T + \epsilon\mathbf{I}\,,$$
$$\mathbf{u} = \tfrac{1}{3}(1,2,3,\ldots,10)^T\,,$$
$$\mathbf{v} = (-1,\tfrac{1}{2},-\tfrac{1}{3},\tfrac{1}{4},\ldots,\tfrac{1}{10})^T$$

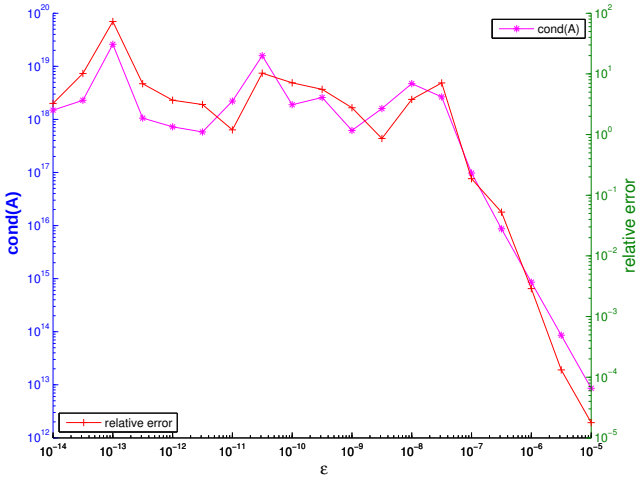


Fig. 44

The practical stability of Gaussian elimination for $\mathbf{A}\mathbf{x} = \mathbf{b}$ is reflected by the size of a particular vector, the residual $\mathbf{r} := \mathbf{b} - \mathbf{A}\widetilde{\mathbf{x}}$, $\widetilde{\mathbf{x}}$ the computed solution, that can easily be computed after the elimination solver has finished:

> In practice *Gaussian elimination/LU-factorization with partial pivoting* produces "relatively small residual vectors"

This is confirmed by a simple consideration analogous to § 2.4.0.2:

$$(\mathbf{A}+\Delta\mathbf{A})\widetilde{\mathbf{x}} = \mathbf{b} \quad\Rightarrow\quad \mathbf{r} = \mathbf{b} - \mathbf{A}\widetilde{\mathbf{x}} = \Delta\mathbf{A}\widetilde{\mathbf{x}} \quad\Rightarrow\quad \|\mathbf{r}\| \le \|\Delta\mathbf{A}\|\|\widetilde{\mathbf{x}}\|\,,$$

for any vector norm $\|\cdot\|$. This means that, if a direct solver for an LSE is stable in the sense of backward error analysis, that is, the perturbed solution could be obtained as the exact solution for a slightly relatively perturbed system matrix, then the *residual will be (relatively) small*.

EXPERIMENT 2.4.0.11 (Small residuals by Gaussian elimination) Gaussian elimination works miracles in terms of delivering small residuals! To demonstrate this we study a numerical experiment with *nearly singular matrix*.

$$\mathbf{A} = \mathbf{u}\mathbf{v}^T + \epsilon\mathbf{I}\,,$$
singular rank-1 matrix

with
$$\mathbf{u} = \tfrac{1}{3}(1,2,3,\ldots,10)^T\,,$$
$$\mathbf{v} = (-1,\tfrac{1}{2},-\tfrac{1}{3},\tfrac{1}{4},\ldots,\tfrac{1}{10})^T$$

C++ code 2.4.0.12: Small residuals for Gauss elimination ↪ GITLAB

```
2  int n = 10;
3  // Initialize vectors u and v
4  VectorXd u = VectorXd::LinSpaced(n, 1, n) / 3.0;
5  VectorXd v = u.cwiseInverse().array() *
6               VectorXd::LinSpaced(n, 1, n)
7                   .unaryExpr([](double x) { return pow(-1, x); })
8                   .array();
9  VectorXd x = VectorXd::Ones(n);
10 double nx = x.lpNorm<Infinity>();
```

```
11  VectorXd expo = VectorXd::LinSpaced(19, -5, -14);
12  Eigen::MatrixXd res(expo.size(), 4);
13  for (int i = 0; i <= expo.size(); ++i) {
14    // Build coefficient matrix A
15    double epsilon = std::pow(10, expo(i));
16    MatrixXd A = u * v.transpose() + epsilon * MatrixXd::Identity(n, n);
17    VectorXd b = A * x;                 // right-hand-side vector
18    double nb = b.lpNorm<Infinity>(); // maximum norm
19    VectorXd xt = A.lu().solve(b);    // Gaussian elimination
20    VectorXd r = b - A * xt;          // residual vector
21    res(i, 0) = epsilon;
22    res(i, 1) = (x - xt).lpNorm<Infinity>() / nx;
23    res(i, 2) = r.lpNorm<Infinity>() / nb;
24    // L-infinity condition number
25    res(i, 3) = A.inverse().cwiseAbs().rowwise().sum().maxCoeff() *
26                A.cwiseAbs().rowwise().sum().maxCoeff();
27  }
```

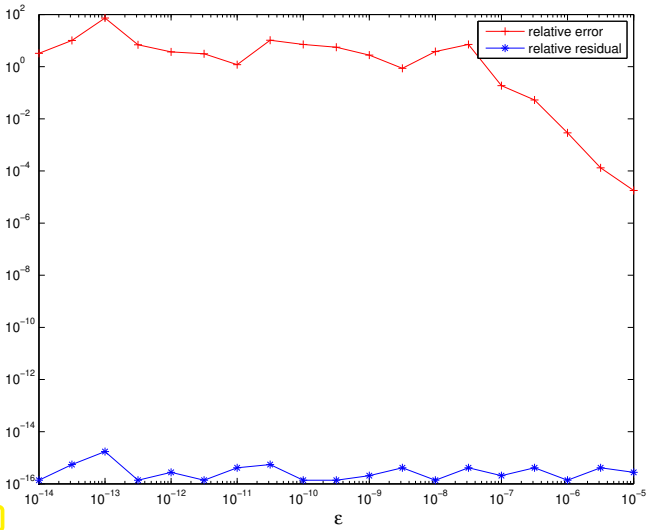


Fig. 45

Observations (w.r.t $\|\cdot\|_\infty$-norm)

- for $\epsilon \ll 1$ large relative error in computed solution $\widetilde{\mathbf{x}}$
- small residuals for any ϵ

How can a *large* relative error be reconciled with a *small* relative residual**?** We continue a discussion that we already started in Ex. 2.4.0.6:

$$\mathbf{A}\mathbf{x} = \mathbf{b} \quad \leftrightarrow \quad \mathbf{A}\widetilde{\mathbf{x}} \approx \mathbf{b}$$

$$\left\{ \begin{array}{lll} \mathbf{A}(\mathbf{x}-\widetilde{\mathbf{x}}) = \mathbf{r} & \Rightarrow & \|\mathbf{x}-\widetilde{\mathbf{x}}\| \le \left\|\mathbf{A}^{-1}\right\|\|\mathbf{r}\| \\ \mathbf{A}\mathbf{x} = \mathbf{b} & \Rightarrow & \|\mathbf{b}\| \le \|\mathbf{A}\|\|\mathbf{x}\| \end{array} \right. \quad \Rightarrow \quad \frac{\|\mathbf{x}-\widetilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \le \|\mathbf{A}\|\left\|\mathbf{A}^{-1}\right\|\frac{\|\mathbf{r}\|}{\|\mathbf{b}\|}\,. \tag{2.4.0.13}$$

➢ If $\mathrm{cond}(\mathbf{A}) := \|\mathbf{A}\|\left\|\mathbf{A}^{-1}\right\| \gg 1$, then a small relative residual may not imply a small relative error. Also recall the discussion in Exp. 2.4.0.10. ⌟

EXPERIMENT 2.4.0.14 (Instability of multiplication with inverse) An important justification for Rem. 2.2.1.6 that advised us not to compute the inverse of a matrix in order to solve a linear system of equations is conveyed by this experiment. We again consider the nearly singular matrix from Exp. 2.4.0.11.

$$\mathbf{A} = \mathbf{u}\mathbf{v}^T + \epsilon\mathbf{I}\,,$$

singular rank-1 matrix

with $\mathbf{u} = \frac{1}{3}(1,2,3,\ldots,10)^T$, $\mathbf{v} = (-1,\frac{1}{2},-\frac{1}{3},\frac{1}{4},\ldots,\frac{1}{10})^T$

C++ code 2.4.0.15: Instability of multiplication with inverse ➙ GITLAB

```
2  int n = 10;
```

```
VectorXd u = VectorXd::LinSpaced(n,1,n) / 3.0;
VectorXd v = u.cwiseInverse().array() * VectorXd::LinSpaced(n,1,n).unaryExpr(
    [](double x){return pow(-1,x);} ).array();
VectorXd x = VectorXd::Ones(n);
VectorXd expo = VectorXd::LinSpaced(19,-5,-14);
MatrixXd res(expo.size(),4);
for(int i = 0; i < expo.size(); ++i){
      double epsilon = std::pow(10, expo(i));
      MatrixXd A = u*v.transpose() + epsilon * (MatrixXd::Random(n,n) +
          MatrixXd::Ones(n,n))/2;
      VectorXd b = A * x;
      double nb = b.lpNorm<Infinity>();
      VectorXd xt = A.lu().solve(b);      // stable solving
      VectorXd r = b - A*xt;   // residual
      MatrixXd B = A.inverse();
      VectorXd xi = B*b;  // solving via inverse
      VectorXd ri = b - A*xi; // residual
      MatrixXd R = MatrixXd::Identity(n,n) - A*B; // residual
      res(i,0) = epsilon; res(i,1) = (r).lpNorm<Infinity>()/nb;
      res(i,2) = ri.lpNorm<Infinity>()/nb;
      // L-infinity condition number
      res(i,3) = R.lpNorm<Infinity>() / B.lpNorm<Infinity>();
}
```

The computation of the inverse is affected by roundoff errors, but does not benefit from the same favorable cancellation of roundoff errors as Gaussian elimination.

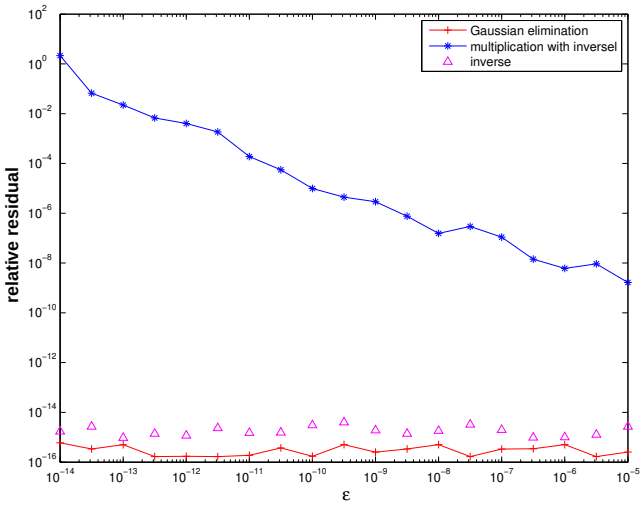


Fig. 46

2.5 Survey: Elimination Solvers for Linear Systems of Equations

All direct (∗) solver algorithms for square linear systems of equations $\mathbf{Ax} = \mathbf{b}$ with given matrix $\mathbf{A} \in \mathbb{K}^{n,n}$, right hand side vector $\mathbf{b} \in \mathbb{K}^n$ and unknown $\mathbf{x} \in \mathbb{K}^n$ rely on variants of Gaussian elimination with pivoting, see Section 2.3.3. Sophisticated, optimised and verified implementations are available in numerical libraries like LAPACK/MKL.

(∗): a direct solver terminates after a predictable finite number of elementary operations for every admissible input.

Never contemplate implementing a general solver for linear systems of equations!

If possible, use algorithms from numerical libraries! (→ Exp. 2.3.1.7)

Therefore, familiarity with details of Gaussian elimination is not required, but one must know when and how to use the library functions and one must be able to assess the computational effort they involve.

§2.5.0.1 (Computational effort for direct elimination) We repeat the reasoning of § 2.3.1.5: Gaussian elimination for a general (dense) matrix invariably involves *three nested loops* of length n, see Code 2.3.1.4, Code 2.3.3.8.

Theorem 2.5.0.2. Cost of Gaussian elimination → § 2.3.1.5

Given a linear system of equations $\mathbf{Ax} = \mathbf{b}$, $\mathbf{A} \in \mathbb{K}^{n,n}$ regular, $\mathbf{b} \in \mathbb{K}^n$, $n \in \mathbb{N}$, the asymptotic computational effort (→ Def. 1.4.0.1) for its direct solution by means of Gaussian elimination in terms of the problem size parameter n is $O(n^3)$ for $n \to \infty$.

The constant hidden in the Landau symbol can be expected to be rather small (≈ 1) as is clear from (2.3.1.6).

The cost for solving are substantially lower, if certain properties of the matrix $\mathbf{A}$ are known. This is clear, if $\mathbf{A}$ is diagonal or orthogonal/unitary. It is also true for triangular matrices (→ Def. 1.1.2.3), because they can be solved by simple back substitution or forward elimination. We recall the observation made in see § 2.3.2.15.

Theorem 2.5.0.3. Cost for solving triangular systems → § 2.3.1.5

In the setting of Thm. 2.5.0.2, the asymptotic computational effort for solving a triangular linear system of equations is $O(n^2)$ for $n \to \infty$.

⌟

§2.5.0.4 (Direct solution of linear systems of equations in EIGEN) EIGEN supplies a rich suite of functions for matrix decompositions and solving LSEs, see EIGEN documentation. The default solver is Gaussian elimination with partial pivoting, accessible through the methods `lu()` and `solve()` of dense matrix types:

Given: system/coefficient matrix $\mathbf{A} \in \mathbb{K}^{n,n}$ regular ↔ A ($n \times n$ EIGEN matrix)
right hand side vectors $\mathbf{B} \in \mathbb{K}^{n,\ell}$ ↔ B ($n \times \ell$ EIGEN matrix)
(corresponds to multiple right hand sides, *cf.* Code 2.3.1.10)

linear algebra	EIGEN
$\mathbf{X} = \mathbf{A}^{-1}\mathbf{B} = \left[\mathbf{A}^{-1}(\mathbf{B})_{:,1}, \ldots, \mathbf{A}^{-1}(\mathbf{B})_{:,\ell}\right]$	`X = A.lu().solve(B);`

The following variant overwrites the right-hand side matrix `B` (in-situ elimination): `A.lu().solveInPlace(B);`

Concerning computational effort, the detailed information given in § 2.3.2.15 can be summarized as

$$\text{cost(X = A.lu().solve(B))} = = O(n^3 + ln^2) \quad \text{for } n, l \to \infty$$

⌟

Remark 2.5.0.5 (Communicating special properties of system matrices in EIGEN) Sometimes, the coefficient matrix of a linear system of equations is known to have certain analytic properties that a direct solver can exploit to perform elimination more efficiently. These properties may even be impossible to

detect by an algorithm, because matrix entries that should vanish exactly might have been perturbed due to roundoff.

Thus one needs to pass EIGEN these informations as follows:

```
// A is lower triangular
x = A.triangularView<Eigen::Lower>().solve(b);
// A is upper triangular
x = A.triangularView<Eigen::Upper>().solve(b);
// A is hermitian / self adjoint and positive definite
x = A.selfadjointView<Eigen::Upper>().llt().solve(b);
// A is hermiatin / self adjoint (poitive or negative semidefinite)
x = A.selfadjointView<Eigen::Upper>().ldlt().solve(b);
```

The methods `llt()` and `ldlt()` rely on special factorizations for *symmetric matrices*, see § 2.8.0.14 below. ⌟

EXPERIMENT 2.5.0.6 (Standard EIGEN lu() operator versus triangularView()) In this numerical experiment we study the gain in efficiency achievable by make the direct solver aware of important matrix properties.

C++ code 2.5.0.7: Direct solver applied to a upper triangular matrix ↪ GITLAB

```
//!  Eigen code:  assessing the gain from using special properties
//!  of system matrices in Eigen
MatrixXd timing(){
    std::vector<int> n = {16,32,64,128,256,512,1024,2048,4096,8192};
    const int nruns = 3;
    MatrixXd times(n.size(),3);
    for(unsigned int i = 0; i < n.size(); ++i){
        Timer t1;
        Timer t2; // timer class
        MatrixXd A = VectorXd::LinSpaced(n[i],1,n[i]).asDiagonal() ;
        A += MatrixXd::Ones(n[i],n[i]).triangularView<Upper>();
        const VectorXd b = VectorXd::Random(n[i]);
        VectorXd x1(n[i]);
        VectorXd x2(n[i]);
        for(int j = 0; j < nruns; ++j){
            t1.start();    x1 = A.lu().solve(b);    t1.stop();
            t2.start();    x2 = A.triangularView<Upper>().solve(b); t2.stop();
        }
        times(i,0) = n[i]; times(i,1) = t1.min(); times(i,2) = t2.min();
    }
    return times;
}
```

Observation: ▷

Being told that only the upper triangular part of the matrix needs to be taken into account, Gaussian elimination reduces to cheap backward elimination, which is much faster than full elimination, *cf.* Thm. 2.5.0.2 vs. Thm. 2.5.0.3.

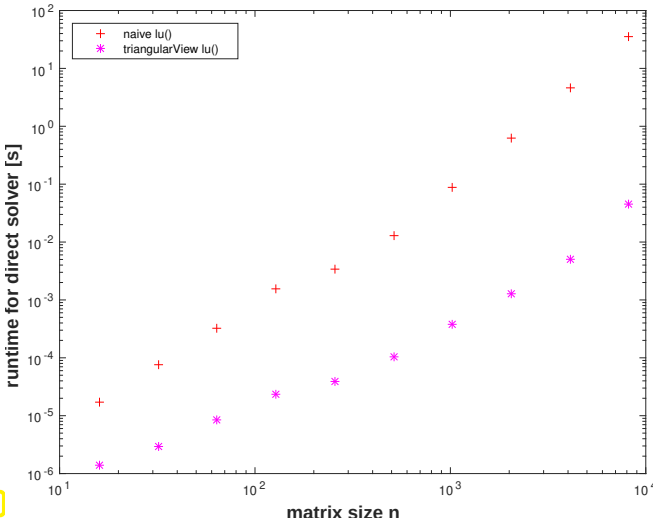


Fig. 47

⌟

§2.5.0.8 (Direct solvers for LSE in EIGEN) Invocation of direct solvers in EIGEN is a two stage process:

❶ Request a decomposition (LU,QR,LDLT) of the matrix and store it in a temporary "decomposition object".

❷ Perform backward & forward substitutions by calling the **`solve()`** method of the decomposition object.

The general format for invoking linear solvers in EIGEN is as follows:

```
Eigen::SolverType<Eigen::MatrixXd> solver(A);
Eigen::VectorXd x = solver.solve(b);
```

This can be reduced to one line, as the solvers can also be used as methods acting on matrices:

```
Eigen::VectorXd x = A.solverType().solve(b);
```

A full list of solvers can be found in the EIGEN documentation. The next code demonstrates a few of the available decompositions that can serve as the basis for a linear solver:

C++-code 2.5.0.9: EIGEN based function solving a LSE ↬ GITLAB

```
// Gaussian elimination with partial pivoting, Code 2.3.3.8
inline void lu_solve(const MatrixXd &A, const VectorXd &b, VectorXd &x) {
  x = A.lu().solve(b); // 'lu()' is short for 'partialPivLu()'
}

// Gaussian elimination with total pivoting
inline void fullpivlu_solve(const MatrixXd &A, const VectorXd &b, VectorXd &x) {
  x = A.fullPivLu().solve(b); // total pivoting
}

// An elimination solver based on Householder transformations
inline void qr_solve(const MatrixXd &A, const VectorXd &b, VectorXd &x) {
  const Eigen::HouseholderQR<MatrixXd> solver(A); // see Section 3.3.3
  x = solver.solve(b);
}

// Use singular value decomposition (SVD)
inline void svd_solve(const MatrixXd &A, const VectorXd &b, VectorXd &x) {
  // SVD based solvers, see Section 3.4
  x = A.jacobiSvd(Eigen::ComputeThinU | Eigen::ComputeThinV).solve(b);
}
```

The different decompositions trade speed for stability and accuracy: fully pivoted and QR-based decompositions also work for nearly singular matrices, for which the standard LU-factorization may non longer be reliable. ⌟

Remark 2.5.0.10 (Many sequential solutions of LSE) As we have seen in Code 2.5.0.9, EIGEN provides functions that return decompositions of matrices. For instance, we can get an object "containing" the LU-decomposition (→ Section 2.3.2) of a matrix by the following commands:

```
Eigen::MatrixXd A(n,n); // A dense square matrix object
......
auto ludec = A.lu(); // Perform LU-decomposition and store the
   factors.
```

Based on the *precomputed decompositions*, a linear system of equations with coefficient matrix $\mathbf{A} \in \mathbb{K}^{n,n}$ can be solved with asymptotic computational effort $O(n^2)$, *cf.* § 2.3.2.15.

The following example illustrates a special situation, in which matrix decompositions can curb computational cost:

C++ code 2.5.0.11: Wasteful approach! ⇝ GITLAB

```
// Setting:  N ≫ 1,
// large matrix A ∈ K^{n,n}
for(int j = 0; j < N; ++j){
     x = A.lu().solve(b);
     b = some_function(x);
}
```

computational effort $O(Nn^3)$

C++ code 2.5.0.12: Smart approach! ⇝ GITLAB

```
// Setting:  N ≫ 1,
// large matrix A ∈ K^{n,n}
auto A_lu_dec = A.lu();
for(int j = 0; j < N; ++j){
     x = A_lu_dec.solve(b);
     b = some_function(x);
}
```

computational effort $O(n^3 + Nn^2)$ ⌟

EXAMPLE 2.5.0.13 (Reuse of LU-decomposition in inverse power iteration) A concrete example is the so-called inverse power iteration, see Chapter 9, for which a skeleton code is given next. It computes the iterates

$$\mathbf{x}^* := \mathbf{A}^{-1}\mathbf{x}^{(k)}\,, \quad \mathbf{x}^{(k+1)} := \frac{\mathbf{x}^*}{\|\mathbf{x}^*\|_2}\,, \quad k = 0, 1, 2, \ldots\,, \tag{2.5.0.14}$$

C++-code 2.5.0.15: Efficient implementation of inverse power method in EIGEN ⇝ GITLAB

```
template<class VecType, class MatType>
VecType invpowit(const Eigen::MatrixBase<MatType> &A,double tol)
{
  using index_t = typename MatType::Index;
  // Make sure that the function is called with a square matrix
  const index_t n = A.cols();
  const index_t m = A.rows();
  eigen_assert(n == m);
  // Request LU-decomposition
  auto A_lu_dec = A.lu();
  // Initial guess for inverse power iteration
  VecType xo = VecType::Zero(n);
  VecType xn = VecType::Random(n);
  // Normalize vector
  xn /= xn.norm();
```

```
17   // Terminate if relative (normwise) change below threshold
18   while ((xo-xn).norm() > xn.norm()*tol) {
19     xo = xn;
20     xn = A_lu_dec.solve(xo);
21     xn /= xn.norm();
22   }
23   return(xn);
24 }
```

The use of `Eigen::MatrixBase<MatType>` makes it possible to call `invpowit` with an expression argument:

```
2 const MatrixXd A = MatrixXd::Random(n,n);
3 const MatrixXd B = MatrixXd::Random(n,n);
4 const auto ev = invpowit<VectorXd>(A+B, tol);
```

This is necessary, because `A+B` will spawn an auxiliary object of a "strange" type determined by the expression template mechanism. ⌟

Remark 2.5.0.16 (Access to LU-factors in EIGEN) LU-decomposition objects available in EIGEN provide access to the computed LU-factors $\mathbf{L}$ and $\boldsymbol{U}$ through a member function `matrixLU()`. This returns a matrix object with $\mathbf{L}$ stored in its strictly lower triangular part, and $\mathbf{U}$ in its upper triangular part.

However note that EIGEN's algorithms for LU-factorization invariably employ (partial) pivoting for the sake of numerical stability, see Section 2.3.3 for a discussion. This has the effect that the LU-factors of a matrix $\mathbf{A} \in \mathbb{R}^{n,n}$ are actually those for a matrix $\mathbf{PA}$, where $\mathbf{P}$ is a permutation matrix as stated in Lemma 2.3.3.14. Thus `matrixLU()` provides the LU-factorization of $\mathbf{A}$ after some row permutation.

C++ code 2.5.0.17: Retrieving the LU-factors from an EIGEN lu object ↪ GITLAB

```
2  std::pair<Eigen::MatrixXd, Eigen::MatrixXd>
3  lufak_eigen(const Eigen::MatrixXd &A) {
4    // Compute LU decomposition
5    auto ludec = A.lu();
6    // The LU-factors are computed by in-situ LU-decomposition,
7    // see Rem. 2.3.2.11, and are stored in a dense matrix of
8    // the same size as A
9    Eigen::MatrixXd L { ludec.matrixLU().triangularView<Eigen::UnitLower>() };
10   const Eigen::MatrixXd U { ludec.matrixLU().triangularView<Eigen::Upper>() };
11   // EIGEN employs partial pivoting, see § 2.3.3.7, which can be viewed
12   // as a prior permutation of the rows of A. We apply the inverse of
       this
13   // permutation to the L-factor in order to achieve A = LU.
14   L.applyOnTheLeft(ludec.permutationP().inverse());
15   // Return LU-factors as members of a 2-tuple.
16   return { L , U };
17 }
```

⌟

2.6 Exploiting Structure when Solving Linear Systems

By "structure" of a linear system we mean prior knowledge that

- ✦ either certain entries of the system matrix vanish,

✦ or the system matrix is generated by a particular formula.

§2.6.0.1 (Triangular linear systems) Triangular linear systems are linear systems of equations whose system matrix is a *triangular matrix* (→ Def. 1.1.2.3).

Thm. 2.5.0.3 tells us that (dense) triangular linear systems can be solved by backward/forward elimination with $O(n^2)$ asymptotic computational effort ($n \hat{=}$ number of unknowns) compared to an asymptotic complexity of $O(n^3)$ for solving a generic (dense) linear system of equations (→ Thm. 2.5.0.2, Exp. 2.5.0.6).

This is the simplest case where exploiting special structure of the system matrix leads to faster algorithms for the solution of a special class of linear systems. ⌟

§2.6.0.2 (Block elimination) Remember that thanks to the possibility to compute the matrix product in a block-wise fashion (→ § 1.3.1.13), Gaussian elimination can be conducted on the level of matrix blocks. We recall Rem. 2.3.1.14 and Rem. 2.3.2.19.

For $k, \ell \in \mathbb{N}$ consider the block partitioned square $n \times n$, $n := k + \ell$, linear system

$$\begin{bmatrix} \mathbf{A}_{11}\mathbf{A}_{12} \\ \mathbf{A}_{21}\mathbf{A}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix}, \quad \begin{array}{l} \mathbf{A}_{11} \in \mathbb{K}^{k,k}, \mathbf{A}_{12} \in \mathbb{K}^{k,\ell}, \mathbf{A}_{21} \in \mathbb{K}^{\ell,k}, \mathbf{A}_{22} \in \mathbb{K}^{\ell,\ell}, \\ \mathbf{x}_1 \in \mathbb{K}^k, \mathbf{x}_2 \in \mathbb{K}^\ell, \mathbf{b}_1 \in \mathbb{K}^k, \mathbf{b}_2 \in \mathbb{K}^\ell. \end{array} \tag{2.6.0.3}$$

Using block matrix multiplication (applied to the matrix×vector product in (2.6.0.3)) we find an equivalent way to write the block partitioned linear system of equations:

$$\begin{array}{rcl} \mathbf{A}_{11}\mathbf{x}_1 + \mathbf{A}_{12}\mathbf{x}_2 &=& \mathbf{b}_1, \\ \mathbf{A}_{21}\mathbf{x}_1 + \mathbf{A}_{22}\mathbf{x}_2 &=& \mathbf{b}_2. \end{array} \tag{2.6.0.4}$$

We assume that $\mathbf{A}_{11}$ is *regular* (invertible) so that we can solve for $\mathbf{x}_1$ from the first equation.

➢ By elementary algebraic manipulations ("block Gaussian elimination") we find

$$\begin{aligned} \mathbf{x}_1 &= \mathbf{A}_{11}^{-1}(\mathbf{b}_1 - \mathbf{A}_{12}\mathbf{x}_2), \\ \underbrace{(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})}_{\text{Schur complement}}\mathbf{x}_2 &= \mathbf{b}_2 - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{b}_1, \end{aligned}$$

The resulting $\ell \times \ell$ linear system of equations for the unknown vector $\mathbf{x}_2$ is called the **Schur complement system** for (2.6.0.3).

Unless $\mathbf{A}$ has a special structure that allows the efficient solution of linear systems with system matrix $\mathbf{A}_{11}$, the Schur complement system is mainly of theoretical interest. ⌟

EXAMPLE 2.6.0.5 (Linear systems with arrow matrices) From $n \in \mathbb{N}$, a diagonal matrix $\mathbf{D} \in \mathbb{K}^{n,n}$, $\mathbf{c} \in \mathbb{K}^n$, $\mathbf{b} \in \mathbb{K}^n$, and $\alpha \in \mathbb{K}$, we can build an $(n+1) \times (n+1)$ arrow matrix.

$$\mathbf{A} = \left[\begin{array}{c|c} \mathbf{D} & \mathbf{c} \\ \hline \mathbf{b}^\top & \alpha \end{array}\right] \tag{2.6.0.6}$$

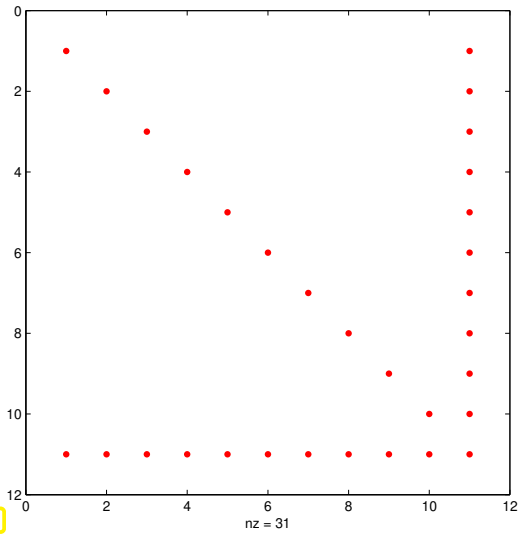


Fig. 48

We can apply the block partitioning (2.6.0.3) with $k = n$ and $\ell = 1$ to a linear system $\mathbf{Ax} = \mathbf{y}$ with system matrix $\mathbf{A}$ and obtain $\mathbf{A}_{11} = \mathbf{D}$, which can be inverted easily, provided that all diagonal entries of $\mathbf{D}$ are different from zero. In this case

$$\mathbf{Ax} = \begin{bmatrix} \mathbf{D} & \mathbf{c} \\ \mathbf{b}^\top & \alpha \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \xi \end{bmatrix} = \mathbf{y} := \begin{bmatrix} \mathbf{y}_1 \\ \eta \end{bmatrix}, \tag{2.6.0.7}$$

$$\begin{aligned} \xi &= \frac{\eta - \mathbf{b}^T \mathbf{D}^{-1} \mathbf{y}_1}{\alpha - \mathbf{b}^\top \mathbf{D}^{-1} \mathbf{c}}, \\ \mathbf{x}_1 &= \mathbf{D}^{-1}(\mathbf{y}_1 - \xi \mathbf{c}) . \end{aligned} \tag{2.6.0.8}$$

These formulas make sense, if $\mathbf{D}$ is regular *and* $\alpha - \mathbf{b}^\top \mathbf{D}^{-1} \mathbf{c} \neq 0$, which is another condition for the invertibility of $\mathbf{A}$.

Using the formula (2.6.0.8) we can solve the linear system (2.6.0.7) with an asymptotic complexity $O(n)$! This superior speed compared to Gaussian elimination applied to the (dense) linear system is evident in runtime measurements.

C++ code 2.6.0.9: Dense Gaussian elimination applied to arrow system ↪ GITLAB

```
VectorXd arrowsys_slow(const VectorXd &d,
                       const VectorXd &c,
                       const VectorXd &b, double alpha,
                       const VectorXd &y) {
  const Eigen::Index n = d.size();
  MatrixXd A(n + 1, n + 1); // Empty dense matrix
  A.setZero();              // Initialize with all zeros.
  A.diagonal().head(n) = d; // Initialize matrix diagonal from a vector.
  A.col(n).head(n) = c;     // Set rightmost column c.
  A.row(n).head(n) = b;     // Set bottom row b^T.
  A(n, n) = alpha;          // Set bottom-right entry α.
  return A.lu().solve(y);   // Gaussian elimination
}
```

Asymptotic complexity $O(n^3)$!
(Due to the serious blunder of accidentally creating a matrix full of zeros, *cf.* Exp. 1.3.1.10.)

C++ code 2.6.0.10: Solving an arrow system according to (2.6.0.8) ➙ **GITLAB**

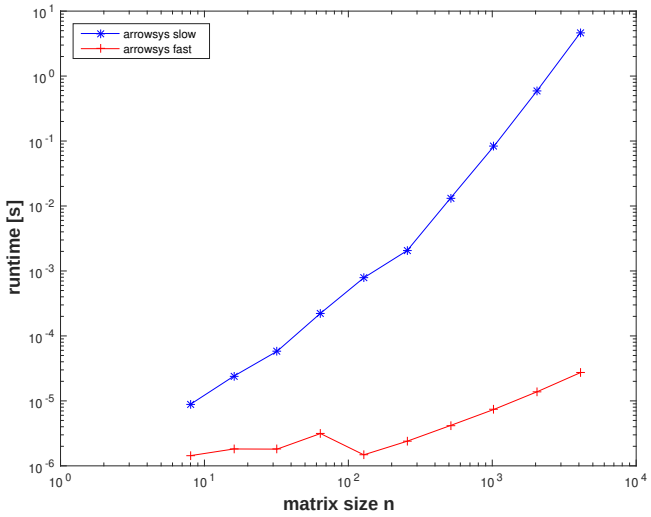
```
VectorXd arrowsys_fast(const VectorXd &d,
                       const VectorXd &c,
                       const VectorXd &b, double alpha,
                       const VectorXd &y) {
  const Eigen::Index n = d.size();
  const VectorXd z = c.array() / d.array();  // z = D^{-1}c
  const VectorXd w =
      y.head(n).array() / d.array();     // w = D^{-1}y_1
  const double den = alpha - b.dot(z);  // denominator in (2.6.0.8)
  // Check for (relatively!)  small denominator
  if (std::abs(den) < std::numeric_limits<double>::epsilon() *
                      (std::abs(b.dot(z)) + std::abs(alpha))) {
    throw std::runtime_error("Nearly singular system");
  }
  const double xi = (y(n) - b.dot(w)) / den;
  return (Eigen::VectorXd(n + 1) << w - xi * z, xi).finished();
}
```

Asymptotic complexity $O(n)$ for $n \to \infty$!

Code for Runtime measurements can be obtained from ➙ GITLAB.

(Intel i7-3517U CPU @ 1.90GHz, 64-bit, Ubuntu Linux 14.04 LTS, gcc 4.8.4, -O3)

No comment! ▷



Remark 2.6.0.11 (Sacrificing numerical stability for efficiency) The vector based implementation of the solver of Code 2.6.0.10 can be vulnerable to roundoff errors, because, upon closer inspection, the algorithm turns out to be equivalent to Gaussian elimination without pivoting, *cf.* Section 2.3.3, Ex. 2.3.3.1.

Caution: stability at risk in Code 2.6.0.10

Yet, there are classes of matrices for which Gaussian elimination without pivoting is guaranteed to be stable. For such matrices algorithms like that of Code 2.6.0.10 are safe. ⌟

§2.6.0.12 (Solving LSE subject to low-rank modification of system matrix) Given a regular matrix $\mathbf{A} \in \mathbb{K}^{n,n}$, let us assume that at some point in a code we are in a position to solve any linear system $\mathbf{Ax} = \mathbf{b}$ "fast", because

- either $\mathbf{A}$ has a favorable structure, eg. triangular, see § 2.6.0.1,
- or an LU-decomposition of $\mathbf{A}$ is already available, see § 2.3.2.15.

Now, a $\widetilde{\mathbf{A}}$ is obtained by changing a *single entry* of $\mathbf{A}$:

$$\mathbf{A}, \widetilde{\mathbf{A}} \in \mathbb{K}^{n,n}: \quad \widetilde{a}_{ij} = \begin{cases} a_{ij} & \text{, if } (i,j) \neq (i^*,j^*) \text{ ,} \\ z + a_{ij} & \text{, if } (i,j) = (i^*,j^*) \text{ ,} \end{cases} \quad , \quad i^*, j^* \in \{1,\ldots,n\} \text{ .} \tag{2.6.0.13}$$

$$\blacktriangleright \quad \boxed{\widetilde{\mathbf{A}} = \mathbf{A} + z \cdot \mathbf{e}_{i^*} \mathbf{e}_{j^*}^T} \text{ .} \tag{2.6.0.14}$$

(Recall: $\mathbf{e}_i \triangleq i$-th unit vector.) The question is whether we can reuse some of the computations spent on solving $\mathbf{A}\mathbf{x} = \mathbf{b}$ in order to solve $\widetilde{\mathbf{A}}\widetilde{\mathbf{x}} = \mathbf{b}$ with less effort than entailed by a direct Gaussian elimination from scratch.

We may also consider a matrix modification affecting a single row: Changing a single row: given $\mathbf{z} \in \mathbb{K}^n$

$$\mathbf{A}, \widetilde{\mathbf{A}} \in \mathbb{K}^{n,n}: \quad \widetilde{a}_{ij} = \begin{cases} a_{ij} & \text{, if } i \neq i^* \text{ ,} \\ (\mathbf{z})_j + a_{ij} & \text{, if } i = i^* \text{ ,} \end{cases} \quad , \quad i^*, j^* \in \{1,\ldots,n\} \text{ .}$$

$$\blacktriangleright \quad \boxed{\widetilde{\mathbf{A}} = \mathbf{A} + \mathbf{e}_{i^*} \mathbf{z}^T} \text{ .} \tag{2.6.0.15}$$

Both matrix modifications (2.6.0.13) and (2.6.0.15) represent rank-1-modifications of $\mathbf{A}$. A generic rank-1-modification reads

$$\mathbf{A} \in \mathbb{K}^{n,n} \quad \mapsto \quad \widetilde{\mathbf{A}} := \mathbf{A} + \boxed{\mathbf{u}\mathbf{v}^H} \text{ , } \mathbf{u}, \mathbf{v} \in \mathbb{K}^n \text{ .} \tag{2.6.0.16}$$

general rank-1-matrix

Trick: Block elimination for an extended linear system, see § 2.6.0.2

As in Ex. 2.3.1.1 we carry out Gaussian elimination for the first column of the bock-partitioned linear system

$$\begin{bmatrix} -1 & \mathbf{v}^{\mathrm{H}} \\ \mathbf{u} & \mathbf{A} \end{bmatrix} \begin{bmatrix} \xi \\ \widetilde{\mathbf{x}} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{b} \end{bmatrix} \tag{2.6.0.17}$$

[G.e. on first column] ➤ $$\begin{bmatrix} -1 & \mathbf{v}^{\mathrm{H}} \\ \mathbf{0} & \mathbf{A} + \mathbf{u}\mathbf{v}^{\mathrm{H}} \end{bmatrix} \begin{bmatrix} \xi \\ \widetilde{\mathbf{x}} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{b} \end{bmatrix} \quad \blacktriangleright \quad \widetilde{\mathbf{A}}\widetilde{\mathbf{x}} = \mathbf{b} \text{ !} \tag{2.6.0.18}$$

Hence, we have solved the modified LSE, once we have found the component $\widetilde{\mathbf{x}}$ of the solution of the linear system (2.6.0.17).

Now we swap (block) rows and columns and consider the block-partitioned linear system

$$\begin{bmatrix} \mathbf{A} & \mathbf{u} \\ \mathbf{v}^{\mathrm{H}} & -1 \end{bmatrix} \begin{bmatrix} \widetilde{\mathbf{x}} \\ \xi \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ 0 \end{bmatrix} \text{ .}$$

We do (block) Gaussian elimination on the first (block))column again, which yields the Schur complement system

$$(1 + \mathbf{v}^{\mathrm{H}}\mathbf{A}^{-1}\mathbf{u})\xi = \mathbf{v}^{\mathrm{H}}\mathbf{A}^{-1}\mathbf{b} \text{ .} \tag{2.6.0.19}$$

$$\blacktriangleright \quad \mathbf{A}\widetilde{\mathbf{x}} = \mathbf{b} - \frac{\mathbf{u}\mathbf{v}^{\mathrm{H}}\mathbf{A}^{-1}}{1 + \mathbf{v}^{\mathrm{H}}\mathbf{A}^{-1}\mathbf{u}}\mathbf{b} \text{ .} \tag{2.6.0.20}$$

The generalization of this formula to rank-k-perturbations if given in the following lemma:

Lemma 2.6.0.21. Sherman-Morrison-Woodbury formula

For regular $\mathbf{A} \in \mathbb{K}^{n,n}$, and $\mathbf{U},\mathbf{V} \in \mathbb{K}^{n,k}$, $n,k \in \mathbb{N}$, $k \leq n$, holds

$$(\mathbf{A}+\mathbf{U}\mathbf{V}^H)^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}(\mathbf{I}+\mathbf{V}^H\mathbf{A}^{-1}\mathbf{U})^{-1}\mathbf{V}^H\mathbf{A}^{-1} ,$$

if $\mathbf{I}+\mathbf{V}^H\mathbf{A}^{-1}\mathbf{U}$ *is regular.*

Proof. Straightforward algbra:

$$\left(\mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}(\mathbf{I}+\mathbf{V}^H\mathbf{A}^{-1}\mathbf{U})^{-1}\mathbf{V}^H\mathbf{A}^{-1}\right)(\mathbf{A}+\mathbf{U}\mathbf{V}^H) =$$
$$\mathbf{I} - \mathbf{A}^{-1}\mathbf{U}\underbrace{(\mathbf{I}+\mathbf{V}^H\mathbf{A}^{-1}\mathbf{U})^{-1}(\mathbf{I}+\mathbf{V}^H\mathbf{A}^{-1}\mathbf{U})}_{=\mathbf{I}}\mathbf{V}^H + \mathbf{A}^{-1}\mathbf{U}\mathbf{V}^H = \mathbf{I} .$$

Uniqueness of the inverse settles the case. □

We use this result to solve $\widetilde{\mathbf{A}}\widetilde{\mathbf{x}} = \mathbf{b}$ with $\widetilde{\mathbf{A}}$ from (2.6.0.16) more efficiently than straightforward elimination could deliver, provided that the LU-factorisation $\mathbf{a} = \mathbf{L}\mathbf{U}$ is already known. We apply Lemma 2.6.0.21 for $k = 1$ and get

$$\widetilde{\mathbf{x}} = \mathbf{A}^{-1}\mathbf{b} - \frac{\mathbf{A}^{-1}\mathbf{u}(\mathbf{v}^H(\mathbf{A}^{-1}\mathbf{b}))}{1+\mathbf{v}^H(\mathbf{A}^{-1}\mathbf{u})} . \tag{2.6.0.22}$$

We have to solve two linear systems of equations with system matrix $\mathbf{A}$, which is "cheap" provided that the LU-decomposition of $\mathbf{A}$ is available. This is another case, where precomputing the LU-decomposition pays off.

Assuming that `lu` passes an object that contains an LU-decomposition of $\mathbf{A} \in \mathbb{R}^{n,n}$, the following code demonstrates an efficient implementation with asymptotic complexity $O(n^2)$ for $n \to \infty$ due to the backward/forward substitutions in Lines 7-8.

C++ code 2.6.0.23: Solving a rank-1 modified LSE ↪ GITLAB

```
2  // Solving rank-1 updated LSE based on (2.6.0.22)
3  template <class LUDec>
4  Eigen::VectorXd smw( const Eigen::VectorXd &u, const Eigen::VectorXd &v,
5      const LUDec &lu, const Eigen::VectorXd &b) {
6    const double singfac = 1.0E6; // Do not lose more than 10 digits
7    const Eigen::VectorXd z = lu.solve(b); // z = A^-1 b
8    const Eigen::VectorXd w = lu.solve(u); // w = A^-1 u
9    const double alpha = 1.0 + v.dot(w); // Compute denominator of (2.6.0.22)
10   const double beta = v.dot(z);        // Factor for numerator of (2.6.0.22)
11   if (std::abs(alpha) <= singfac * std::numeric_limits<double>::epsilon()) {
12     throw std::runtime_error("A nearly singular");
13   }
14   return (z - w * beta / alpha); // see (2.6.0.22)
15 }
```

In Rem. 1.5.3.15 we were told that the test whether a numerical result was zero should be done by comparing with another quantity. Then why is this advice not heeded in the above code. The reason is the line `alpha = 1.0 + v.dot(w)`. Imagine this results in a value $\alpha = 10 \cdot \mathrm{EPS}$. In this case

the computation of α would involve massive cancellation, see Section 1.5.4, and the result would probably have a huge relative error ≈ 0.1. This would completely destroy the accuracy of the final result, regardless of the size of any other quantity computed in the code. Therefore, it is advisable to check the absolute size of α. ⌟

EXAMPLE 2.6.0.24 (Resistance to currents map) Many lineare systems with system matrices that differ in a single entry only have to be solved when we want to determine the dependence of the total impedance of a (linear) circuit from the parameters of a single component.

Large (linear) electric circuit (modelling → Ex. 2.1.0.3) ▷

Sought:

Dependence of (certain) branch currents on "continuously varying" resistance R_x

(➢ currents for many different values of R_x)

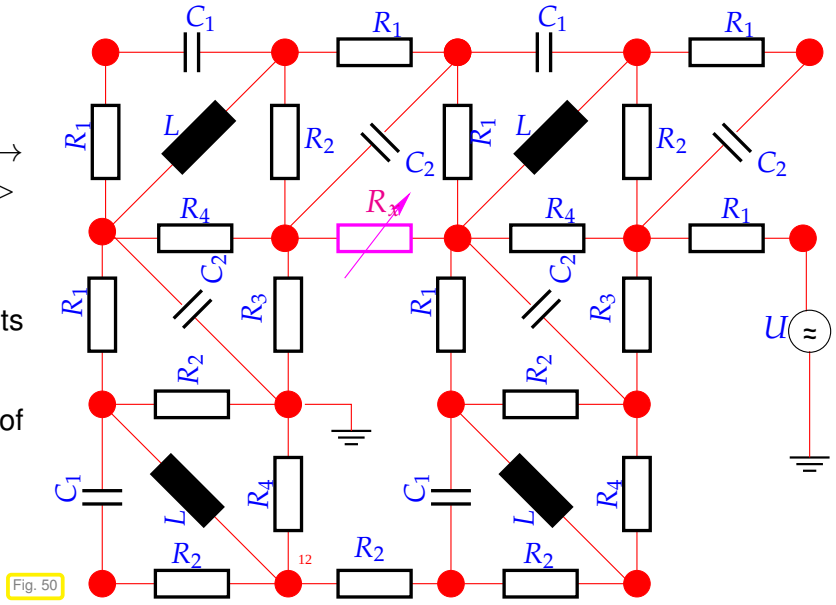


Fig. 50

Only *a few* entries of the nodal analysis matrix $\mathbf{A}$ (→ Ex. 2.1.0.3) are affected by variation of R_x! (If R_x connects nodes i & j ⇒ only entries $a_{ii}, a_{jj}, a_{ij}, a_{ji}$ of $\mathbf{A}$ depend on R_x) ⌟

Review question(s) 2.6.0.25 (Exploiting structure when solving linear systems of equations)

(Q2.6.0.25.A) Compute the block LU-decomposition for the arrow matrix

$$\mathbf{A} = \left[\begin{array}{c|c} \mathbf{D} & \mathbf{c} \\ \hline \mathbf{b}^\top & \alpha \end{array}\right], \quad \begin{array}{l} \mathbf{D} \in \mathbb{R}^{n,n} \text{ regular, } \textit{diagonal}\,, \\ \mathbf{c}, \mathbf{b} \in \mathbb{R}^n\,, \\ \alpha \in \mathbb{R}\,, \end{array}$$

according to the indicated (and natural) partitioning of the matrix.

(Q2.6.0.25.B) Sketch an efficient algorithm for solving the LSE

$$\left[\begin{array}{c|c} \mathbf{I}_{n-1} & \mathbf{c} \\ \hline \mathbf{0}^\top & \alpha \end{array}\right] \mathbf{x} = \mathbf{b}\,, \quad \mathbf{b} \in \mathbb{R}^n\,, \quad \mathbf{c} \in \mathbb{R}^{n-1}\,, \quad \alpha > 0\,.$$

(Q2.6.0.25.C) Given a matrix $\mathbf{A} \in \mathbb{R}^{n,n}$ find rank-1 modifications that replace its i-th row or column with a given vector $\mathbf{w} \in \mathbb{R}^n$.

(Q2.6.0.25.D) Given a regular matrix $\mathbf{A} \in \mathbb{R}^{n,n}$ and $\boldsymbol{b} \in \mathbb{R}^n$, we want to solve many linear systems of the form $\widetilde{\mathbf{A}}(\xi)\mathbf{x} = \mathbf{b}$, where $\widetilde{\mathbf{A}}(\xi)$ is obtained by adding $\xi \in \mathbb{R}$ to every entry of $\mathbf{A}$.

Propose an efficient implementation for a C++ class

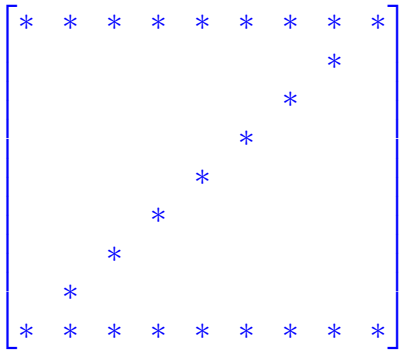
```
class ConstModMatSolveLSE {
  public:
  ConstModMatSolveLSE(const Eigen::MatrixXd &A,
  const Eigen::VectorXd &b);
  Eigen::VectorXd solvemod(double xi) const;
  private:
  ...
};
```

that serves this purpose. Do not forget to test for near-singularity of the matrix $\widetilde{\mathbf{A}}(\xi)$!

(Q2.6.0.25.E) [“Z-shaped” matrix] Let $\mathbf{A} \in \mathbb{R}^{n,n}$ be “Z-shaped”

$$(\mathbf{A})_{i,j} = 0 \text{ ,if } \quad i \in \{2,\ldots,n-1\}\,, i+j \neq n+1 \quad ,$$

e.g. $\mathbf{A} =$



.

1. Outline an efficient algorithm for solving a linear system of equations $\mathbf{A}\mathbf{x} = \mathbf{b}$, $\mathbf{b} \in \mathbb{R}^{n,n}$.
2. Give a sufficient and necessary condition for $\mathbf{A}$ being regular/invertible.

(Q2.6.0.25.F) [Loss of stability] By direct block-wise Gaussian elimination we found the following solution formulas for a block-partitioned linear system of equations with $\mathbf{D} \in \mathbb{R}^{n,n}$, $\mathbf{c}, \mathbf{b} \in \mathbb{R}^n$, $\alpha \in \mathbb{R}$, $\mathbf{y} \in \mathbb{R}^{n+1}$:

$$\mathbf{A}\mathbf{x} = \begin{bmatrix} \mathbf{D} & \mathbf{c} \\ \mathbf{b}^\top & \alpha \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \xi \end{bmatrix} = \mathbf{y} := \begin{bmatrix} \mathbf{y}_1 \\ \eta \end{bmatrix}\,, \tag{2.6.0.7}$$

$$\xi = \frac{\eta - \mathbf{b}^T\mathbf{D}^{-1}\mathbf{y}_1}{\alpha - \mathbf{b}^\top\mathbf{D}^{-1}\mathbf{c}}\,, \qquad \mathbf{x}_1 = \mathbf{D}^{-1}(\mathbf{y}_1 - \xi\mathbf{c})\,. \tag{2.6.0.8}$$

Use these formulas to compute the solution of the 2×2 linear system of equations

$$\begin{bmatrix} \delta & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ \xi \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}\,,$$

assuming $|\delta| < \frac{1}{2}\mathrm{EPS}$ and using floating point arithmetic.

Hint. Remember that, if $|\delta| < \frac{1}{2}\mathrm{EPS}$, in floating point arithmetic

$$1\tilde{+}\delta = 1 \quad \text{and} \quad 2\tilde{+}\delta^{-1} = \delta^{-1}\,.$$

This is compatible with the "Axiom" of roundoff analysis Ass. 1.5.3.11

(Q2.6.0.25.G) [A banded linear system] Sketch an efficient algorithm for the solution of the $n \times n$ linear system of equations

$$\begin{bmatrix} 1 & 0 & \dots & \dots & 0 & 1 \\ 1 & 1 & 0 & \dots & & 0 \\ 0 & \ddots & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 1 & 1 & 0 \\ 0 & \dots & \dots & 0 & 1 & 1 \end{bmatrix} \mathbf{x} = \mathbf{b} \in \mathbb{R}^n\,.$$

When will a solution exists for every right-hand side vector $\mathbf{b}$?

Hint. You may first perform Gaussian elimination for $n = 7$ and $n = 8$ or use the LU-decomposition

$$\begin{bmatrix} 1 & 0 & \dots & \dots & 0 & 1 \\ 1 & 1 & 0 & \dots & & 0 \\ 0 & \ddots & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 1 & 1 & 0 \\ 0 & \dots & \dots & 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & \dots & \dots & 0 & 0 \\ 1 & 1 & 0 & \dots & & 0 \\ 0 & \ddots & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 1 & 1 & 0 \\ 0 & \dots & \dots & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & \dots & & \dots & 0 & 1 \\ 0 & 1 & \ddots & & & \vdots & -1 \\ \vdots & \ddots & \ddots & & & & 1 \\ & & & \ddots & \ddots & \vdots & -1 \\ & & & \ddots & \ddots & 0 & \vdots \\ \vdots & & & & \ddots & \ddots & (-1)^{n-2} \\ 0 & \dots & & & \dots & 0 & 1+(-1)^{n-1} \end{bmatrix}.$$

△

2.7 Sparse Linear Systems

We start with a (rather fuzzy) classification of matrices according to their numbers of zeros:

Dense(ly populated) matrices) ◆ sparse(ly populated) matrices

Notion 2.7.0.1. Sparse matrix

$\mathbf{A} \in \mathbb{K}^{m,n}$, $m,n \in \mathbb{N}$, is sparse, if

$$\mathrm{nnz}(\mathbf{A}) := \#\{(i,j) \in \{1,\ldots,m\} \times \{1,\ldots,n\} : a_{ij} \neq 0\} \ll mn \,.$$

Sloppy parlance: matrix sparse :⇔ "almost all" entries $= 0$ /"only a few percent of" entries $\neq 0$

J.H. Wilkinson's informal working definition for a developer of simulation codes:

Notion 2.7.0.2. Sparse matrix

A matrix with enough zeros that it pays to take advantage of them should be *treated* as sparse.

A more rigorous "mathematical" definition:

Definition 2.7.0.3. Families of sparse matrices

Given a strictly increasing sequences $m : \mathbb{N} \mapsto \mathbb{N}$, $n : \mathbb{N} \mapsto \mathbb{N}$, a family $(\mathbf{A}^{(l)})_{l\in\mathbb{N}}$ of matrices with $\mathbf{A}^{(l)} \in \mathbb{K}^{m_l,n_l}$ is sparse (opposite: dense), if

$$\lim_{l\to\infty} \frac{\mathrm{nnz}(\mathbf{A}^{(l)})}{n_l m_l} = 0 \,.$$

Simple example: families of diagonal matrices (→ Def. 1.1.2.3) of increasing size.

EXAMPLE 2.7.0.4 (Sparse LSE in circuit modelling) See Ex. 2.1.0.3 for the description of a linear electric circuit by means of a linear system of equations for nodal voltages. For large circuits the system matrices will invariably be huge and sparse.

Modern electric circuits (VLSI chips):
$10^5 - 10^7$ circuit elements

- Each element is connected to only *a few* nodes
- Each node is connected to only *a few* elements

[In the case of a linear circuit]

nodal analysis ➤ sparse circuit matrix
(Impossible to even store as dense matrices)

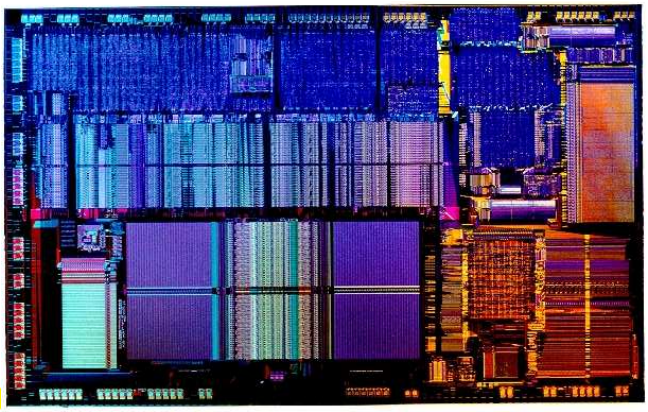


Fig. 51

Remark 2.7.0.5 (Sparse matrices from the discretization of linear partial differential equations) Another important context in which sparse matrices usually arise:

☛ spatial discretization of linear boundary value problems for partial differential equations by means of finite element (FE), finite volume (FV), or finite difference (FD) methods ($\rightarrow$ 4th semester course "Numerical methods for PDEs").

⌟

2.7.1 Sparse Matrix Storage Formats

Sparse matrix storage formats for storing a "sparse matrix" $\mathbf{A} \in \mathbb{K}^{m,n}$ are designed to achieve two objectives:

❶ Amount of memory required is only slightly more than $\mathrm{nnz}(\mathbf{A})$ scalars.

❷ Computational effort for matrix$\times$vector multiplication is proportional to $\mathrm{nnz}(\mathbf{A})$.

In this section we see a few schemes used by numerical libraries.

§2.7.1.1 (Triplet/coordinate list (COO) format) In the case of a sparse matrix $\mathbf{A} \in \mathbb{K}^{m,n}$, this format stores triplets $(i,j,\alpha_{i,j})$, $1 \le i \le m$, $1 \le j \le n$:

```
struct Triplet {
  size_t i; // row index
  size_t j; // column index
  scalar_t a; // additive contribution to matrix entry
};
using TripletMatrix = std::vector<Triplet>;
```

Here **scalar_t** is the underlying scalar type, either **float**, **double**, or **std::complex<double>**.

The vector of triplets in a **TripletMatrix** has size $\geq \mathrm{nnz}(\mathbf{a})$. We write "$\geq$", because *repetitions* of index pairs (i,j) are *allowed*. The matrix entry $(\mathbf{A}))i,j$ is defined to be the *sum* of all values $\alpha_{i,j}$ associated with the index pair (i,j). The next code clearly demonstrates this summation.

C++-code 2.7.1.2: Matrix$\times$vector product $\mathbf{y} = \mathbf{A}\mathbf{x}+\mathbf{y}$ in triplet format

```
1 void multTriplMatvec(const TripletMatrix &A,
2 const vector<scalar_t> &x,
3 vector<scalar_t> &y)
4 for (size_t k=0; k<A.size(); k++) {
5   y[A[k].i] += A[k].a*x[A[k].j];
6 }
```

Note that this code assumes that the result vector y has the appropriate length; no index checks are performed.

Code 2.7.1.2: computational effort is proportional to the number of triplets. (This might be much larger than $\mathrm{nnz}(\mathbf{A})$ in case of many repetitions of triplets.) ⌟

Remark 2.7.1.3 (The zoo of sparse matrix formats) Special sparse matrix storage formats store *only* non-zero entries:

- Compressed Row Storage (CRS)
- Compressed Column Storage (CCS) $\rightarrow$ used by MATLAB
- Block Compressed Row Storage (BCRS)
- Compressed Diagonal Storage (CDS)

- Jagged Diagonal Storage (JDS)
- Skyline Storage (SKS)

All of these formats achieve the two objectives stated above. Some have been designed for sparse matrices with additional structure or for seamless cooperation with direct elimination algorithms (JDS,SKS). ⌟

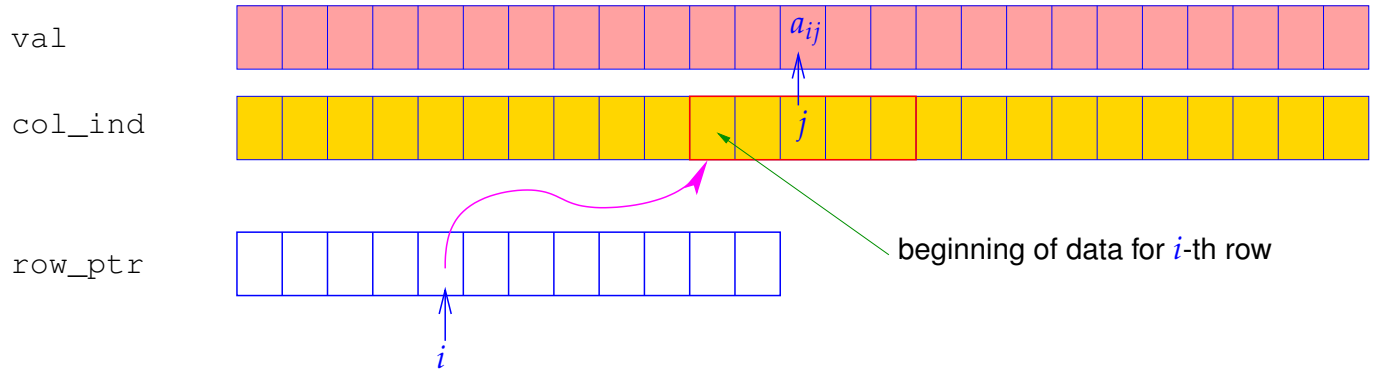
§2.7.1.4 (Compressed row-storage (CRS) format) The CRS format for a sparse matrix $\mathbf{A} = [a_{ij}] \in \mathbb{K}^{n,n}$ keeps the data in three *contiguous* arrays:

`std::vector<scalar_t>`	`val`	size $\geq \mathrm{nnz}(\mathbf{A}) := \#\{(i,j) \in \{1,\ldots,n\}^2, a_{ij} \neq 0\}$
`std::vector<size_t>`	`col_ind`	size = `val.size()`
`std::vector<size_t>`	`row_ptr`	size $n+1$ & `row_ptr`$[n+1]$ = `val.size()` (sentinel value)

As above we write $\mathrm{nnz}(\mathbf{A}) \triangleq$ (**n**umber of **n**on**z**eros) of $\mathbf{A}$

Access to matrix entry $a_{ij} \neq 0$, $1 \leq i,j \leq n$ ("mathematical indexing")

$$\texttt{val}[k] = a_{ij} \quad \Leftrightarrow \quad \begin{cases} \texttt{col_ind}[k] = j\,, \\ \texttt{row_ptr}[i] \leq k < \texttt{row_ptr}[i+1]\,, \end{cases} \qquad 1 \leq k \leq \mathrm{nnz}(\mathbf{A})\,.$$



$$\mathbf{A} = \begin{bmatrix} 10 & 0 & 0 & 0 & -2 & 0 \\ 3 & 9 & 0 & 0 & 0 & 3 \\ 0 & 7 & 8 & 7 & 0 & 0 \\ 3 & 0 & 8 & 7 & 5 & 0 \\ 0 & 8 & 0 & 9 & 9 & 13 \\ 0 & 4 & 0 & 0 & 2 & -1 \end{bmatrix}$$

`val`-vector:

10	-2	3	9	3	7	8	7	3 ... 9	13	4	2	-1

`col_ind`-array:

1	5	1	2	6	2	3	4	1 ... 5	6	2	5	6

`row_ptr`-array:

1	3	6	9	13	17	20

Variant: diagonal CRS format (matrix diagonal stored in separate array)
The CCS format is equivalent to CRS format for the transposed matrix. ⌟

Review question(s) 2.7.1.5 (Sparse Matrix Storage Formats)

(Q2.7.1.5.A) Explain why access to the source code of a function that computes the matrix×vector product for a particular sparse-matrix storage format (encapsulated in a C++ class) already gives you full information about that format.

(Q2.7.1.5.B) Let a matrix $\mathbf{A} \in \mathbb{R}^{n,n}$ be given in COO/triplet format and by an **TripletMatrix** object `A`:

```
struct Triplet {
  size_t i; // row index
  size_t j; // column index
```

```
  scalar_t a; // additive contribution to matrix entry
};
using TripletMatrix = std::vector<Triplet>;
```

Outline the implementation of a function

```
Eigen::VectorXd mvTridiPart(const TripletMatrix &A,
const Eigen::VectorXd &x);
```

that computes $\boldsymbol{y} := \widetilde{\mathbf{A}}\mathbf{x}$, where $\widetilde{\mathbf{A}} \in \mathbb{R}^{n,n}$ is defined as

$$\left(\widetilde{\mathbf{A}}\right)_{i,j} = \begin{cases} (\mathbf{A})_{i,j} & \text{, if } |i-j| \leq 1\,, \\ 0 & \text{else,} \end{cases} \qquad i,j \in \{1,\ldots,n\}\,.$$

(Q2.7.1.5.C) Let a matrix $\mathbf{A} \in \mathbb{R}^{n,n}$ be given in COO/triplet format and by an **TripletMatrix** object `A`:

```
struct Triplet {
  size_t i; // row index
  size_t j; // column index
  scalar_t a; // additive contribution to matrix entry
};
using TripletMatrix = std::vector<Triplet>;
```

Sketch a code that builds a **TripletMatrix** object corresponding to the matrix

$$\mathbf{B} := \begin{bmatrix} \mathbf{I}_n & \mathbf{A} \\ \mathbf{A}^\top & \mathbf{O} \end{bmatrix} \in \mathbb{R}^{2n,2n}\,.$$

(Q2.7.1.5.D) Assume that a sparse matrix in CRS format is represented by an object of the type

```
struct CRSMatrix {
  std::vector<double> val;
  std::vector<std::size_t> col_ind;
  std::vector<std::size_t> row_ptr;
};
```

Describe the implementation of a C++ function

```
CRSMatrix makeSecondDiffMat(unsigned int n);
```

that creates the **CRSMatrix** object for $\mathbf{A} \in \mathbb{R}^{n,n}$ defined as

$$(\mathbf{A})_{i,j} = \begin{cases} 2 & \text{, if } i = j\,, \\ -1 & \text{, if } |i-j| = 1\,, \\ 0 & \text{else,} \end{cases} \qquad i,j \in \{1,\ldots,n\}\,.$$

(Q2.7.1.5.E) For a given matrix $\mathbf{A} \in \mathbb{R}^{m,n}$, $m,n \in \mathbb{N}$, we define the square matrix

$$\mathbf{W}_\mathbf{A} := \begin{bmatrix} \mathbf{O}_{m,m} & \mathbf{A} \\ \mathbf{A}^\top & \mathbf{O}_{n,n} \end{bmatrix} \in \mathbb{R}^{m+n,m+n}\,.$$

Outline the implementation of an efficient C++ function

```
void crsAtoW(std::vector<double> &val,
std::vector<unsigned int> &col_ind,
std::vector<unsigned int> &row_ptr);
```

whose arguments supply the three vectors defining the matrix $\mathbf{A}$ in CRS format and which overwrites them with the corresponding vectors of the CRS-format description of $\mathbf{W}_{\mathbf{A}}$.

Remember that the CRS format of a matrix $\mathbf{A} \in \mathbb{R}^{m,n}$ is defined by

$$\texttt{val}[k] = (\mathbf{A})_{i,j} \quad \Leftrightarrow \quad \begin{cases} \texttt{col_ind}[k] = j\,, \\ \texttt{row_ptr}[i] \leq \texttt{k} < \texttt{row_ptr}[i+1]\,, \end{cases} \quad 1 \leq k \leq \mathrm{nnz}(\mathbf{A})\,.$$

It may be convenient to use **std::vector::**`resize(n)` that resizes a vector so that it contains `n` elements. If `n` is smaller than the current container size, the content is reduced to its first `n` elements, removing those beyond (and destroying them). If `n` is greater than the current container size, the content is expanded by inserting at the end as many elements as needed to reach a size of `n` using their default value.

△

2.7.2 Sparse Matrices in EIGEN

Eigen can handle sparse matrices in the standard Compressed Row Storage (CRS) and Compressed Column Storage (CCS) format, see § 2.7.1.4 and the EIGEN documentation:

```
#include <Eigen/Sparse>
Eigen::SparseMatrix<int, Eigen::ColMajor> Asp(rows,cols); // CCS
   format
Eigen::SparseMatrix<double, Eigen::RowMajor> Bsp(rows,cols); // CRS
   format
```

Usually sparse matrices in CRS/CCS format must not be filled by setting entries through index-pair access, because this would entail frequently moving big chunks of memory. The matrix should first be assembled in triplet format (→ EIGEN documentation), from which a sparse matrix is built. EIGEN offers special data types and facilities for handling triplets.

```
std::vector <Eigen::Triplet <double >   > triplets;
// ..  fill the std::vector triplets ..
Eigen::SparseMatrix<double, Eigen::RowMajor> spMat(rows, cols);
spMat.setFromTriplets(triplets.begin(), triplets.end());
```

A triplet object can be initialized as demonstrated in the following example:

```
unsigned int row_idx = 2;
unsigned int col_idx = 4;
double value = 2.5;
Eigen::Triplet<double> triplet(row_idx,col_idx,value);
std::cout << '(' << triplet.row() << ',' << triplet.col()
<< ',' << triplet.value() << ')' << std::endl;
```

As shown, a **Triplet** object offers the access member functions `row()`, `col()`, and `value()` to fetch the row index, column index, and scalar value stored in a `Triplet`.

The statement that entry-wise initialization of sparse matrices is not efficient has to be qualified in Eigen. Entries can be set, provided that *enough space* for each row (in `RowMajor` format) is reserved in advance. This done by the `reserve()` method that takes an integer vector of maximal expected numbers of non-zero entries per row:

C++-code 2.7.2.1: Accessing entries of a sparse matrix: potentially inefficient! ➙ GITLAB

```
unsigned int rows,cols,max_no_nnz_per_row;
.....
Eigen::SparseMatrix<double, Eigen::RowMajor> mat(rows, cols);
mat.reserve(Eigen::RowVectorXi::Constant(rows, max_no_nnz_per_row));
// do many (incremental) initializations
for (int i = 0; i < rows; ++i) {
  mat.insert(i, i) = -1.0; // only for matrix entries not yet set!
  mat.insert(i, i + 1) = 1.0;
  mat.coeffRef(i, 2 * i) -= 1.0; // access entry possibly not set yet
}
mat.makeCompressed(); // squeeze out zeros
```

`insert(i,j)` sets an entry of the sparse matrix, which is rather efficient, provided that enough space has be `reserve`d. `coeffRef(i,j)` gives l-value and r-value access to any matrix entry, creating a non-zero entry, if needed: costly!

The usual matrix operations are supported for sparse matrices; addition and subtraction may involve only sparse matrices stored in the *same format*. These operations may incur large hidden costs and have to be used with care!

EXPERIMENT 2.7.2.2 (Initialization of sparse matrices in Eigen) We study the runtime behavior of the initialization of a sparse matrix in Eigen. We use the methods described above. The code is available from ➙ GITLAB.

Runtimes (in ms) for the initialization of a banded matrix (with 5 non-zero diagonals, that is, a maximum of 5 non-zero entries per row) using different techniques in Eigen.

Green line: timing for entry-wise initialization with only 4 non-zero entries per row reserved in advance.

(OS: Ubuntu Linux 14.04, CPU: Intel i5@1.80 Ghz, Compiler: g++-4.8.2, -O2)

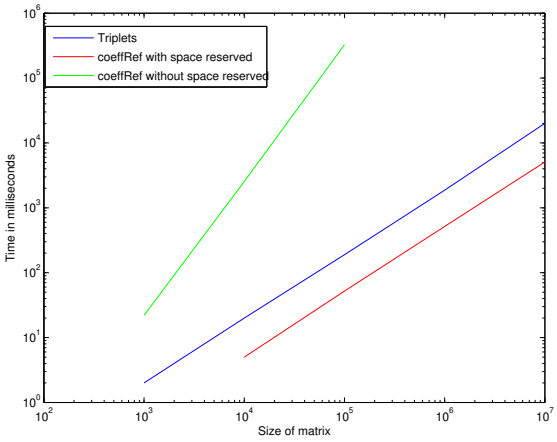


Fig. 52

Observation: insufficient advance allocation of memory massively slows down the set-up of a sparse matrix in the case of direct entry-wise initialization.

Reason: Massive internal copying of data is required to created space for "unexpected" entries. ⌟

Remark 2.7.2.3 (Extracting triplets from Eigen::SparseMatrix) Given an **Eigen::SparseMatrix** object `A` describing a generic sparse matrix $\mathbf{A} \in \mathbb{R}^{n,n}$ we have to create another **Eigen::SparseMatrix** object `B` for a matrix $\mathbf{B} \in \mathbb{R}^{n,n}$, which agrees with $\mathbf{A}$ except that the entries in its first sub-diagonal are equal to the entries of a vector $\mathbf{v} \in \mathbb{R}^{n-1}$:

$$(\mathbf{B})_{i,j} = \begin{cases} (\mathbf{v})_j & \text{, if } i = j+1 \text{ ,} \\ (\mathbf{A})_{i,j} & \text{else,} \end{cases} \qquad i,j \in \{1,\ldots,n\} \text{ .}$$

How can this be done efficiently? By temporarily creating a triplet representation of $\mathbf{A}$, manipulating a few triplets, and then using `makeCompressed()` to obtain `B`.

Thus we need a way to extract a triplet vector from an **Eigen::SparseMatrix** object. This is done by the following fairly complicated code (→ EIGEN documentation):

C++ code 2.7.2.4: Extracting triplets from a Eigen::SparseMatrix ↪ GITLAB

```
2  template <typename Scalar>
3  std::vector<Eigen::Triplet<Scalar>>
4  convertToTriplets(Eigen::SparseMatrix<Scalar> &A) {
5    // Empty vector of triplets to be grown in the following loop
6    std::vector<Eigen::Triplet<Scalar>> triplets{};
7    // Loop over row/columns (depending on column/row major format
8    for (int k = 0; k < A.outerSize(); ++k) {
9      // Loop over inner dimension and obtain triplets corresponding
10     // to non-zero entries.
11     for (typename Eigen::SparseMatrix<Scalar>::InnerIterator it(A, k); it;
12          ++it) {
13       // Retrieve triplet data from iterator
14       triplets.emplace_back(it.row(), it.col(), it.value());
15     }
16   }
17   return triplets;
18 }
```

⌟

EXAMPLE 2.7.2.5 (Smoothing of a triangulation) This example demonstrates that sparse linear systems of equations naturally arise in the handling of triangulations.

Definition 2.7.2.6. Planar triangulation

A planar triangulation (mesh) $\mathcal{M}$ consists of a set $\mathcal{N}$ of $N \in \mathbb{N}$ distinct points $\in \mathbb{R}^2$ and a set $\mathcal{T}$ of triangles with vertices in $\mathcal{N}$, such that the following two conditions are satisfied:

1. the interiors of the triangles are mutually disjoint ("no overlap"),
2. for every two *closed* distinct triangles $\in \mathcal{T}$ their intersection satisfies exactly one of the following conditions:
   (a) it is empty
   (b) it is exactly one vertex from $\mathcal{N}$,
   (c) it is a common edge of both triangles

The points in $\mathcal{N}$ are also called the nodes of the mesh, the triangles the cells, and all line segments connecting two nodes and occurring as a side of a triangle form the set of edges. We always assume a

consecutive numbering of the nodes and cells of the triangulation (starting from 1, MATLAB's convention).

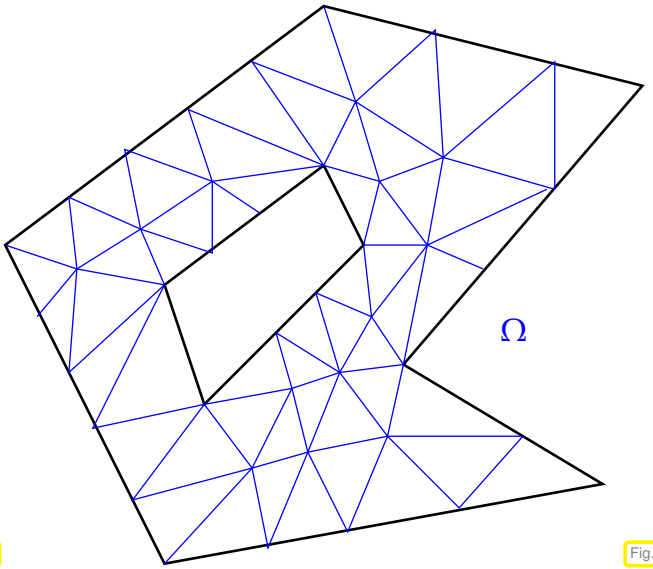


Fig. 53

Valid planar triangulation

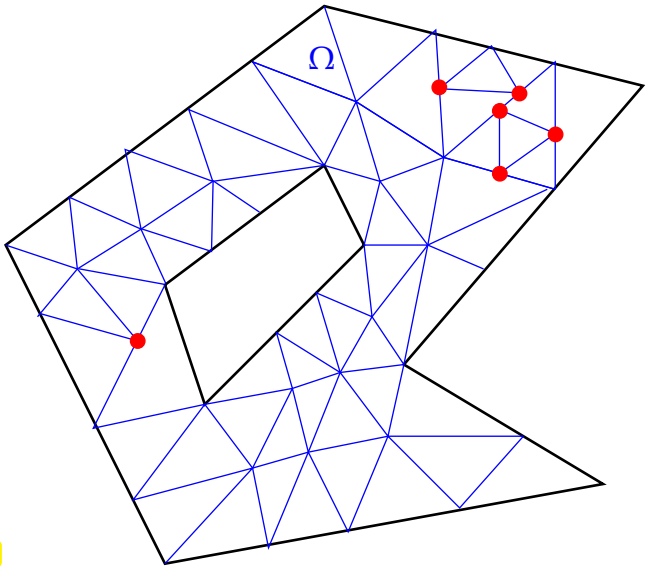
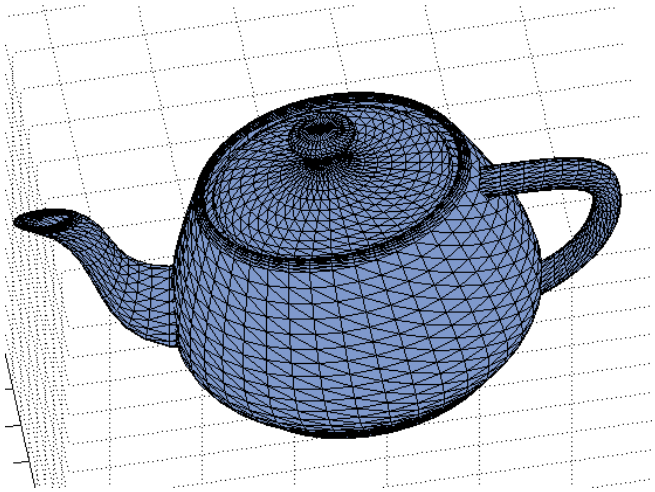
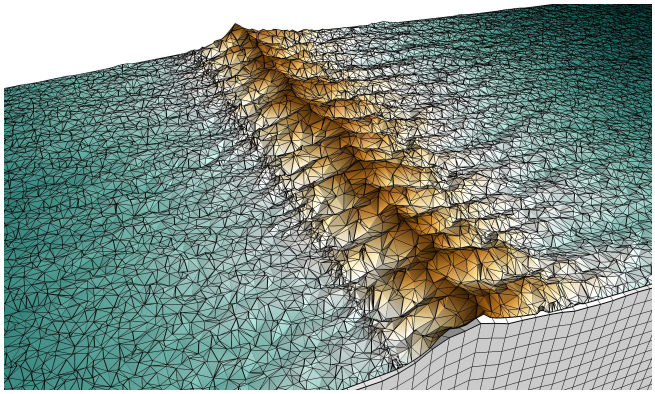


Fig. 54

Mesh with "illegal" hanging nodes

Triangulations are of fundamental importance for computer graphics, landscape models, geodesy, and numerical methods. They need not be planar, but the algorithmic issues remain the same.





Common data structure for describing a triangulation with N nodes and M cells:

- column vector $\mathbf{x} \in \mathbb{R}^N$: x-coordinates of nodes
- column vector $\mathbf{y} \in \mathbb{R}^N$: y-coordinates of nodes
- $M \times 3$-matrix $\mathbf{T}$ whose *rows* contain the index numbers of the vertices of the cells.
  (This matrix is a so-called triangle-node incidence matrix.)

PYTHON's visualization add-on provides the function `matplotlib.pyplot.triplot`, which can be used to draw planar triangulations.

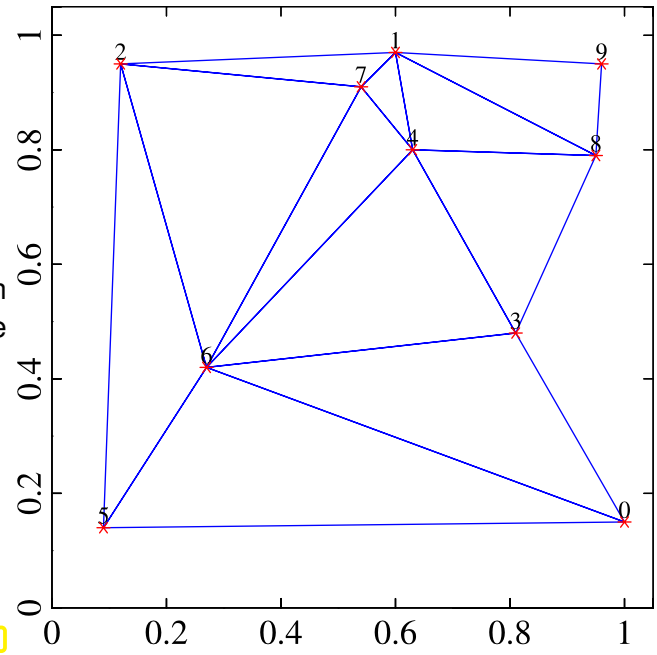


Fig. 55

The cells of a mesh may be rather distorted triangles (with very large and/or small angles), which is usually not desirable. We study an algorithm for smoothing a mesh without changing the planar domain covered by it.

Definition 2.7.2.7. Boundary edge

Every edge that is adjacent to only one cell is a boundary edge of the triangulation. Nodes that are endpoints of boundary edges are boundary nodes.

✎ Notation: $\Gamma \subset \{1,\ldots,N\} \hat{=}$ set of indices of boundary nodes.

✎ Notation: $\mathbf{p}^i = [p_1^i, p_2^i]^\top \in \mathbb{R}^2 \hat{=}$ coordinate vector of node $\sharp i$, $i = 1,\ldots,N$

We define

$$S(i) := \{j \in \{1,\ldots,N\} : \text{ nodes } i \text{ and } j \text{ are connected by an edge}\}\,, \tag{2.7.2.8}$$

as the set of node indices of the "neighbours" of the node with index number i.

Definition 2.7.2.9. Smoothed triangulation

A triangulation is called smoothed, if

$$\mathbf{p}^i = \frac{1}{\sharp S(i)} \sum_{j \in S(i)} \mathbf{p}^j \tag{2.7.2.10}$$

$$\Updownarrow$$

$$\sharp S(i) \left(\mathbf{p}^i\right)_d = \sum_{j \in S(i)} \left(\mathbf{p}^j\right)_d\,, \quad d = 1,2\,, \quad \text{for all} \quad i \in \{1,\ldots,N\} \setminus \Gamma\,,$$

that is, every interior node is located in the center of gravity of its neighbours.

The relations (2.7.2.10) correspond to the lines of a sparse linear system of equations! In order to state it,

we insert the coordinates of all nodes into a column vector $\mathbf{z} \in \mathbb{K}^{2N}$, according to

$$z_i = \begin{cases} p_1^i & \text{, if } 1 \le i \le N \text{ ,} \\ p_2^{i-N} & \text{, if } N+1 \le i \le 2N \text{ .} \end{cases} \tag{2.7.2.11}$$

For the sake of ease of presentation, in the sequel we *assume* (which is not the case in usual triangulation data) that interior nodes have index numbers smaller than that of boundary nodes.

From (2.7.2.8) we infer that the system matrix $\mathbf{C} \in \mathbb{R}^{2n,2N}$, $n := N - \sharp\Gamma$, of that linear system has the following structure:

$$\mathbf{C} = \begin{bmatrix} \mathbf{A} & \mathbf{O} \\ \mathbf{O} & \mathbf{A} \end{bmatrix} \text{ , } \quad (\mathbf{A})_{i,j} = \begin{cases} \sharp S(i) & \text{, if } i = j \text{ ,} \\ -1 & \text{, if } j \in S(i) \text{ ,} \\ 0 & \text{else,} \end{cases} \quad \begin{matrix} i \in \{1,\ldots,n\} \text{ ,} \\ j \in \{1,\ldots,N\} \text{ .} \end{matrix} \tag{2.7.2.12}$$

$$\blacktriangleright \quad (2.7.2.10) \quad \Leftrightarrow \quad \mathbf{Cz} = \mathbf{0} \text{ .} \tag{2.7.2.13}$$

➢ $\mathrm{nnz}(\mathbf{A}) \le$ number of edges of $\mathcal{M}$ + number of interior nodes of $\mathcal{M}$.
➢ The matrix $\mathbf{C}$ associated with $\mathcal{M}$ according to (2.7.2.12) is clearly sparse.
➢ The sum of the entries in every row of $\mathbf{C}$ vanishes.

We partition the vector $\mathbf{z}$ into coordinates of nodes in the interior and of nodes on the boundary

$$\mathbf{z}^T = \begin{bmatrix} \mathbf{z}_1^{\mathrm{int}} \\ \mathbf{z}_1^{\mathrm{bd}} \\ \mathbf{z}_2^{\mathrm{int}} \\ \mathbf{z}_2^{\mathrm{bd}} \end{bmatrix} := \left[z_1,\ldots,z_n,z_{n+1},\ldots,z_N,z_{N+1},\ldots,z_{N+n},z_{N+n+1},\ldots,z_{2N}\right]^\top \text{ .}$$

This induces the following block partitioning of the linear system (2.7.2.13):

$$\begin{bmatrix} \mathbf{A}_{\mathrm{int}} & \mathbf{A}_{\mathrm{bd}} & \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} & \mathbf{A}_{\mathrm{int}} & \mathbf{A}_{\mathrm{bd}} \end{bmatrix} \begin{bmatrix} \mathbf{z}_1^{\mathrm{int}} \\ \mathbf{z}_1^{\mathrm{bd}} \\ \mathbf{z}_2^{\mathrm{int}} \\ \mathbf{z}_2^{\mathrm{bd}} \end{bmatrix} = \mathbf{0} \text{ , } \quad \begin{matrix} \mathbf{A}_{\mathrm{int}} \in \mathbb{R}^{n,n} \text{ ,} \\ \mathbf{A}_{\mathrm{bd}} \in \mathbb{R}^{n,N-n} \text{ .} \end{matrix}$$

$$\Updownarrow$$

$$\left[\begin{array}{cccc} \mathbf{A}_{\mathrm{int}} & \mathbf{A}_{\mathrm{bd}} & & \\ & & \mathbf{A}_{\mathrm{int}} & \mathbf{A}_{\mathrm{bd}} \end{array}\right] \left[\begin{array}{c} \\ \\ \\ \\ \end{array}\right] = \mathbf{0} \text{ .} \tag{2.7.2.14}$$

The linear system (2.7.2.14) holds the key to the algorithmic realization of mesh smoothing; when smoothing the mesh

(i) the node coordinates belonging to interior nodes have to be adjusted to satisfy the equilibrium condition (2.7.2.10), they are unknowns,

(ii) the coordinates of nodes located on the boundary are fixed, that is, their values are known.

unknown $\mathbf{z}_1^{\mathrm{int}}, \mathbf{z}_2^{\mathrm{int}}$, (yellow in (2.7.2.14)) known $\mathbf{z}_1^{\mathrm{bd}}, \mathbf{z}_2^{\mathrm{bd}}$ (pink in (2.7.2.14))

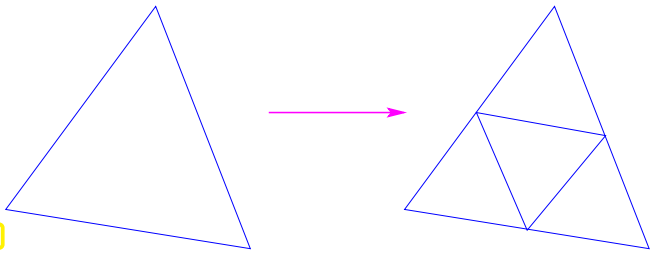
$$\text{(2.7.2.13)/(2.7.2.14)} \quad \Leftrightarrow \quad \mathbf{A}_{\mathrm{int}}\left[\mathbf{z}_1^{\mathrm{int}}\ \mathbf{z}_2^{\mathrm{int}}\right] = -\left[\mathbf{A}_{\mathrm{bd}}\mathbf{z}_1^{\mathrm{bd}}\ \mathbf{A}_{\mathrm{bd}}\mathbf{z}_2^{\mathrm{bd}}\right]. \tag{2.7.2.15}$$

This is a square linear system with an $n \times n$ system matrix, to be solved for two different right hand side vectors. The matrix $\mathbf{A}_{\mathrm{int}}$ is also known as the matrix of the combinatorial graph Laplacian.

We examine the sparsity pattern of the system matrices $\mathbf{A}_{\mathrm{int}}$ for a sequence of triangulations created by regular refinement.

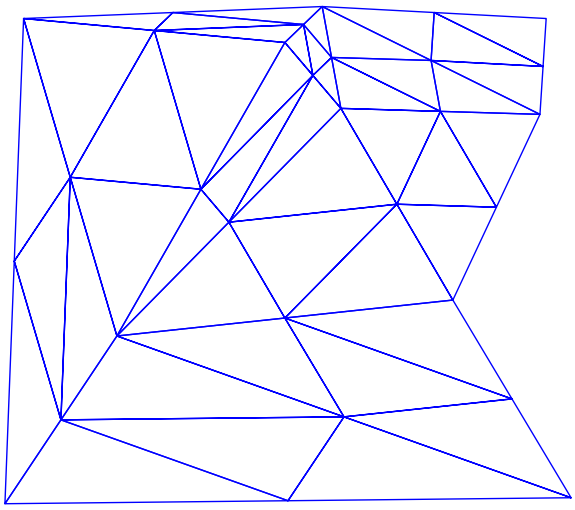
Definition 2.7.2.16. Regular refinemnent of a planar triangulation

The planar triangulation with cells obtained by splitting all cells of a planar triangulation $\mathcal{M}$ into four congruent triangles is called the regular refinement of $\mathcal{M}$.

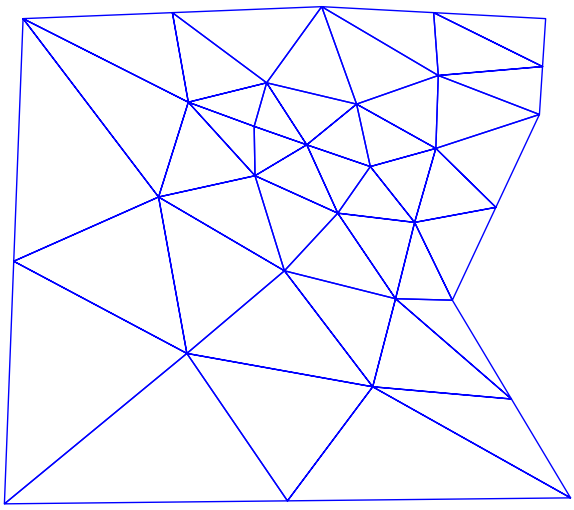


We start from the triangulation of Fig. 55 and in turns perform regular refinement and smoothing (left $\leftrightarrow$ after refinement, right $\leftrightarrow$ after smoothing)

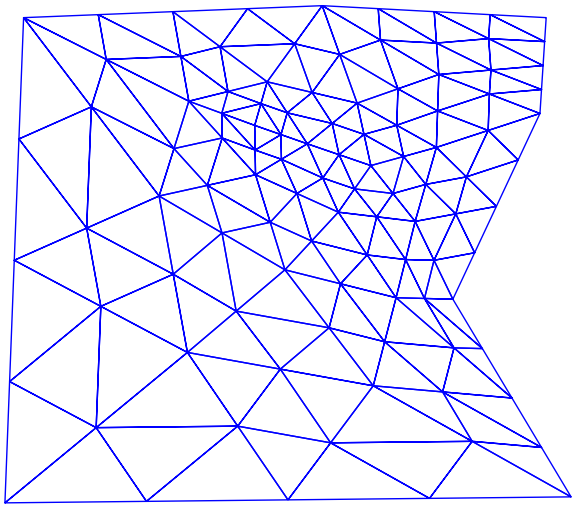
Refined mesh level 1



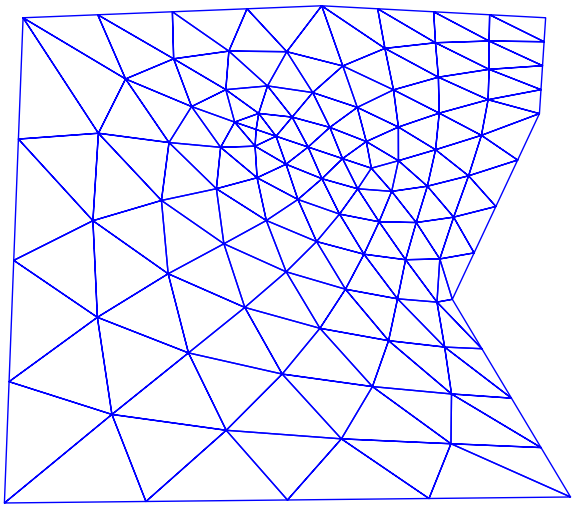
Smoothed mesh level 1



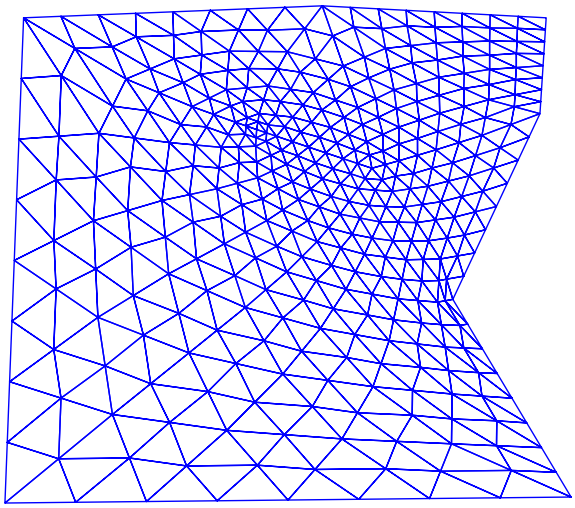
Refined mesh level 2



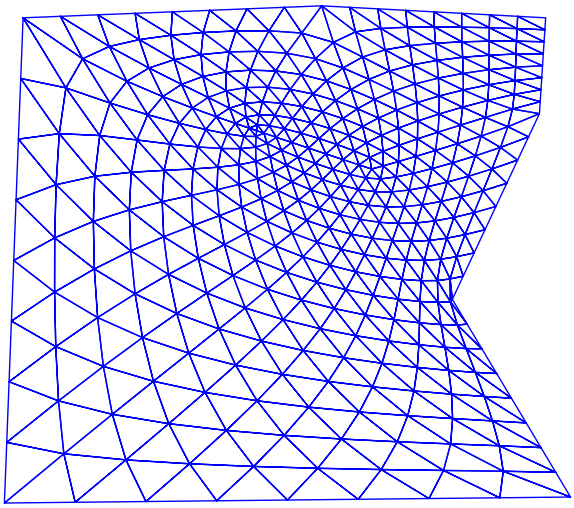
Smoothed mesh level 2



Refined mesh level 3



Smoothed mesh level 3



Below we give spy plots of the system matrices $\mathbf{A}_{\text{int}}$ for the first three triangulations of the sequence:

level 1: 30 points, 73 edges, 44 cells

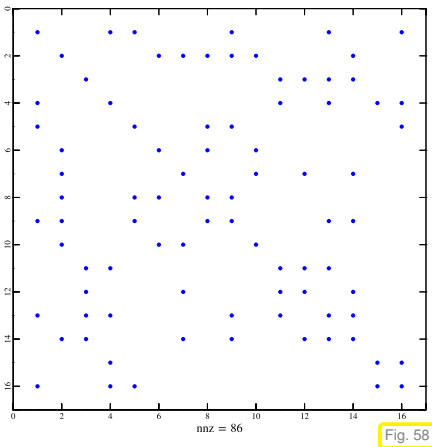


Fig. 57

level 2: 103 points, 278 edges, 176 cells

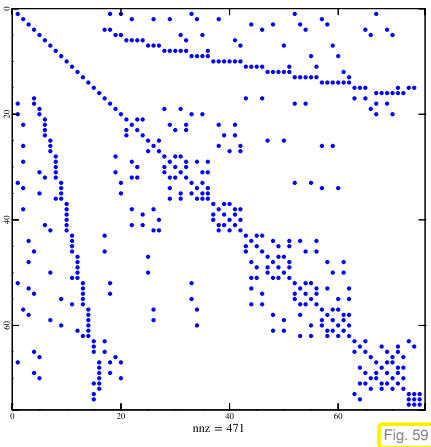


Fig. 58

level 3: 381 points, 1084 edges, 704 cells

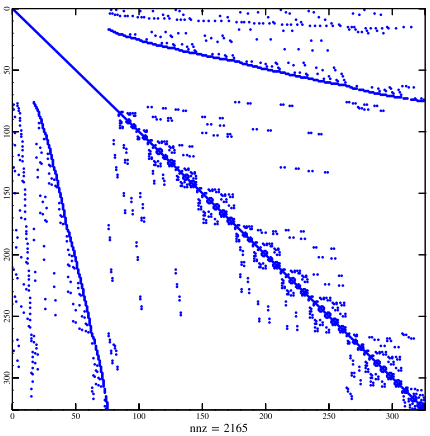


Fig. 59

Review question(s) 2.7.2.17 (Sparse matrices in EIGEN)

(Q2.7.2.17.A) How would you implement the method `setFromTriplets()` of **Eigen::SparseMatrix<double>** in order to achieve an asymptotic complexity $O(\sharp\text{triplets})$?

△

2.7.3 Direct Solution of Sparse Linear Systems of Equations

Efficient Gaussian elimination for sparse matrices requires sophisticated algorithms that are encapsulated in special types of solvers in EIGEN. Their calling syntax remains unchanged, however:

```
Eigen::SolverType<Eigen::SparseMatrix<double>> solver(A);
Eigen::VectorXd x = solver.solve(b);
```

The standard sparse solver is `SparseLU`.

C++ code 2.7.3.1: Function for solving a sparse LSE with EIGEN ⇝ GITLAB

```
using SparseMatrix = Eigen::SparseMatrix<double>;
// Perform sparse elimination
inline void sparse_solve(const SparseMatrix &A, const VectorXd &b, VectorXd &x) {
  const Eigen::SparseLU<SparseMatrix> solver(A);
  if (solver.info() != Eigen::Success) {
    throw std::runtime_error("Matrix factorization failed");
  }
  x = solver.solve(b);
}
```

The constructor of the `solver` object builds the actual sparse LU-decomposition. The `solve` method then does forward and backward elimination, *cf.* § 2.3.2.15. It can be called multiple times, see Rem. 2.5.0.10. For more sample codes see ⇝ GITLAB.

EXPERIMENT 2.7.3.2 (Sparse elimination for arrow matrix) In Ex. 2.6.0.5 we saw that applying the standard lu() solver to a sparse arrow matrix results in an extreme waste of computational resources.

Yet, EIGEN can do much better**!** The main mistake was the creation of a dense matrix instead of storing the arrow matrix in sparse format. There are EIGEN solvers which rely on particular sparse elimination techniques. They still rely of Gaussian elimination with (partial) pivoting (→ Code 2.3.3.8), but take pains to operate on non-zero entries only. This can greatly boost the speed of the elimination.

C++ code 2.7.3.3: Invoking sparse elimination solver for arrow matrix ⇝ GITLAB

```
template <class solver_t>
VectorXd arrowsys_sparse(const VectorXd &d,
                         const VectorXd &c,
                         const VectorXd &b, double alpha,
                         const VectorXd &y) {
    const Eigen::Index n = d.size();
    SparseMatrix<double> A(n+1, n+1); // default:  column-major
    VectorXi reserveVec = VectorXi::Constant(n+1, 2); // nnz per col
    reserveVec(n) = static_cast<int>(n+1); // last full col
    A.reserve(reserveVec);
    for(int j = 0; j < n; ++j){   // initalize along cols for efficiency
        A.insert(j,j) = d(j);     // diagonal entries
        A.insert(n,j) = b(j);     // bottom row entries
    }
    for(int i = 0; i < n; ++i){
        A.insert(i,n) = c(i);     // last col
    }
    A.insert(n,n) = alpha;        // bottomRight entry
    A.makeCompressed();
```

```
21        return solver_t(A).solve(y);
22    }
```

Observation:

The sparse elimination solver is several orders of magnitude faster than `lu()` operating on a dense matrix.

The sparse solver is still slower than Code 2.6.0.10. The reason is that it is a general algorithm that has to keep track of non-zero entries and has to be prepared to do pivoting.

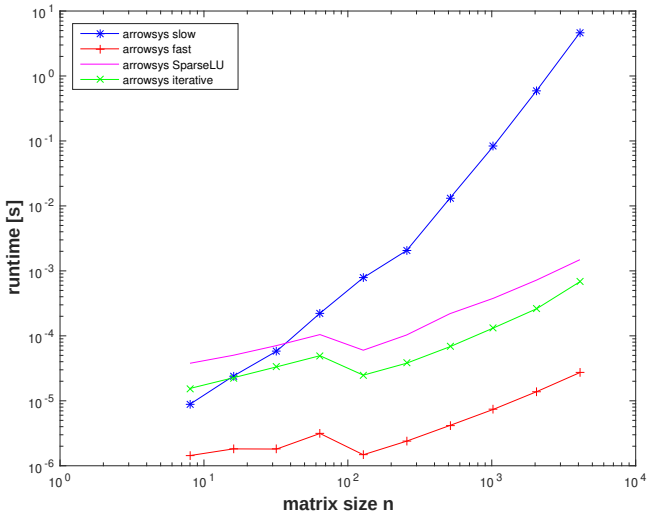


Fig.

EXPERIMENT 2.7.3.4 (Timing sparse elimination for the combinatorial graph Laplacian) We consider a sequence of planar triangulations created by successive regular refinement (→ Def. 2.7.2.16) of the planar triangulation of Fig. 55, see Ex. 2.7.2.5. We use different EIGEN and MKL sparse solver for the linear system of equations (2.7.2.15) associated with each mesh.

Timing results ▷

Platform:

- ubuntu 14.04 LTS
- i7-3517U CPU @ 1.90GHz
- L1 32 KB, L2 256 KB, L3 4096 KB, Mem 8 GB
- gcc 4.8.4, -O3

We observe an *empirical* asymptotic complexity (→ Def. 1.4.1.1) of $O(n^{1.5})$, way better than the asymptotic complexity of $O(n^3)$ expected for Gaussian elimination in the case of dense matrices.

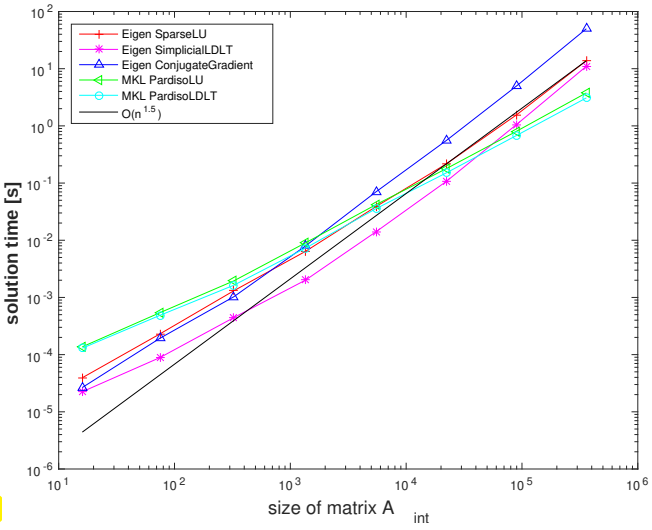


Fig. 61

When solving linear systems of equations directly **dedicated sparse elimination solvers** from *numerical libraries* have to be used!

System matrices are passed to these algorithms in sparse storage formats (→ Section 2.7.1) to convey information about zero entries.

STOP Never ever even think about implementing a general sparse elimination solver by yourself!

For an survey of sparse solvers available in EIGEN see EIGEN documentation.

§2.7.3.5 (Implementations of sparse solvers) Widely used implementations of sparse solvers are:

→ SuperLU (http://www.cs.berkeley.edu/~demmel/SuperLU.html),
→ UMFPACK (https://en.wikipedia.org/wiki/UMFPACK), used by MATLAB's \,
→ PARDISO [SG04] (http://www.pardiso-project.org/), incorporated into MKL

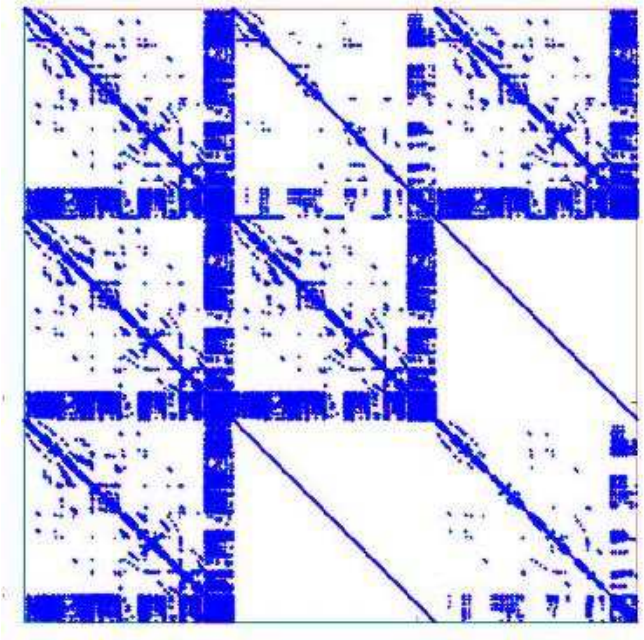


Fig. 62

◁ fill-in (→ Def. 2.7.4.3) during sparse elimination with PARDISO

PARDISO has been developed by Prof. O. Schenk and his group (formerly University of Basel, now USI Lugano).

C++-code 2.7.3.6: Example code demonstrating the use of PARDISO with EIGEN ⇢ GITLAB

```
2  void solveSparsePardiso(size_t n) {
3    using SpMat = Eigen::SparseMatrix<double>;
4    // Initialize a sparse matrix
5    const SpMat M = initSparseMatrix<SpMat>(n);
6    const Eigen::VectorXd b = Eigen::VectorXd::Random(n);
7    Eigen::VectorXd x(n);
8    // Initalization of the sparse direct solver based on the Pardiso
         library with
9    // directly passing the matrix M to the solver Pardiso is part of the
         Intel
10   // MKL library, see also Ex. 1.3.2.6
11   Eigen::PardisoLU<SpMat> solver(M);
12   // The checks of Code 2.7.3.1 are omitted
13   // solve the LSE
14   x = solver.solve(b);
15 }
```

Required is **#include** <Eigen/PardisoSupport>, the compilation flag -DEIGEN_USE_MKL_ALL, and the inclusion of MKL libraries during the linking phase.

```
1 COMPILER = clang++
2 FLAGS = -std=c++11 -m64 -I/usr/include/eigen3 -I${MKLROOT}/include -O3 -DNDEBUG
3
4 # Intel(R) MKL 11.3.2, Linux, None, GNU C/C++, Intel(R) 64, Static, LP64,
    Sequential)
5 FLAGS_LINK = -Wl,--start-group ${MKLROOT}/lib/intel64/libmkl_intel_lp64.a \
6 ${MKLROOT}/lib/intel64/libmkl_core.a ${MKLROOT}/lib/intel64/libmkl_sequential.a \
7 -Wl,--end-group -lpthread -lm -ldl
8
```

```
all: main.cpp
$(COMPILER) $(FLAGS) -DEIGEN_USE_MKL_ALL $< -o main $(FLAGS_LINK)
```

⌟

Review question(s) 2.7.3.7 (Direct solution of sparse linear systems of equations)

(Q2.7.3.7.A) In Code 2.7.3.1 we checked `(solver.info()!= Eigen::Success)`, where `solver` was of type `Eigen::SparseLU<Eigen::SparseMatrix>`. Can you explain, why it is a good idea to include this test.

(Q2.7.3.7.B) What are the benefits of storing a matrix in CRS or CCS format?

△

2.7.4 LU-Factorization of Sparse Matrices

In Section 2.7.1 we have seen, how sparse matrices can be stored requiring $O(\mathrm{nnz}(\mathbf{A}))$ memory.

However, simple examples show that the product of sparse matrices need not be sparse, which means that the multiplication of large sparse matrices will usually require an effort way bigger than the sum of the numbers of their non-zero entries.

What is the situation concerning the solution of square linear systems of equations with sparse system matrices**?** Generically, we have to brace for a computational effort $O(n^3)$ for matrix size $n \to \infty$. Yet Section 2.7.3 sends the message that a better asymptotic complexity can often be achieved, if the sparse matrix has a particular structure and sophisticated library routines are used. In this section, we examine some aspects of Gaussian elimination $\leftrightarrow$ LU-factorisation when applied in a sparse-matrix context.

EXAMPLE 2.7.4.1 (LU-factorization of sparse matrices) We examine the following "sparse" matrix with a typical structure and inspect the pattern of the LU-factors returned by EIGEN, see Code 2.7.4.2.

$$\mathbf{A} = \left[\begin{array}{cccc|cccc} 3 & -1 & & & -1 & & & \\ -1 & \ddots & \ddots & & & \ddots & & \\ & \ddots & \ddots & -1 & & & \ddots & \\ & & -1 & 3 & & & & -1 \\ \hline -1 & & & & 3 & -1 & & \\ & \ddots & & & -1 & \ddots & \ddots & \\ & & \ddots & & & \ddots & \ddots & -1 \\ & & & -1 & & & -1 & 3 \end{array}\right] \in \mathbb{R}^{n,n}, n \in \mathbb{N}$$

C++ code 2.7.4.2: Visualizing LU-factors of a sparse matrix ↪ GITLAB

```
// Build matrix
const Eigen::Index n = 100;
RowVectorXd diag_el(5);
diag_el << -1, -1, 3, -1, -1;
VectorXi diag_no(5);
diag_no << -n, -1, 0, 1, n;
MatrixXd B = diag_el.replicate(2 * n, 1);
B(n - 1, 1) = 0;
B(n, 3) = 0; // delete elements
// A custom function from the Utils folder
const SparseMatrix<double> A = spdiags(B, diag_no, 2 * n, 2 * n);
// It is not possible to access the LU-factors in the case of
// EIGEN's LU-decomposition for sparse matrices.
// Therefore we have to resort to the dense version.
auto solver = MatrixXd(A).lu();
MatrixXd L = MatrixXd::Identity(2 * n, 2 * n);
```

```
18  L += solver.matrixLU().triangularView<StrictlyLower>();
19  const MatrixXd U = solver.matrixLU().triangularView<Upper>();
20  // Plotting
21  spy(A, "Sparse matrix", "sparseA_cpp.eps");
22  spy(L, "Sparse matrix: L factor", "sparseL_cpp.eps");
23  spy(U, "Sparse matrix: U factor", "sparseU_cpp.eps");
```

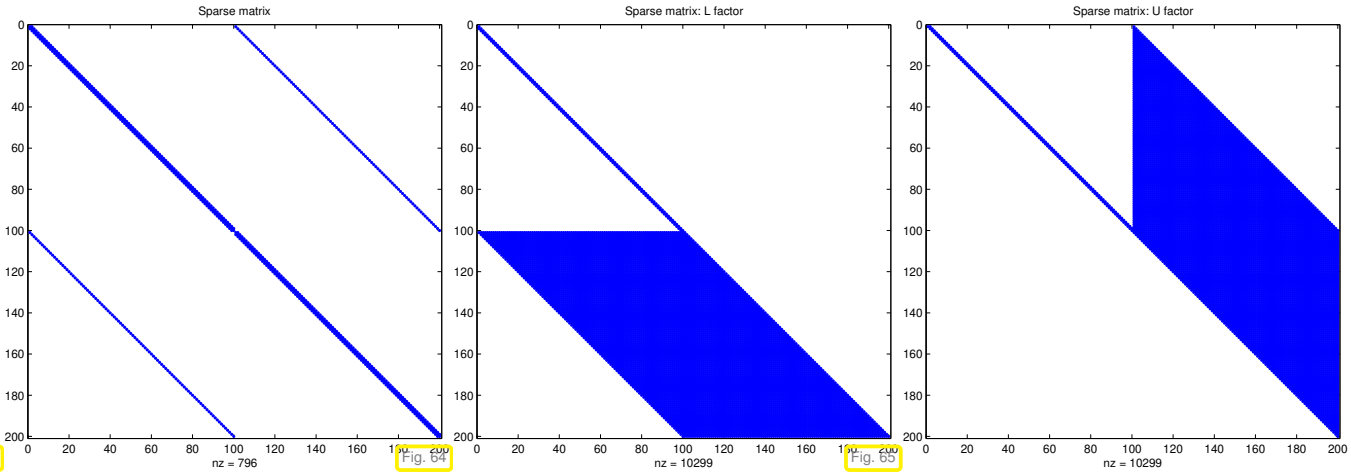


Fig. 63 Fig. 64 Fig. 65

Observation:	$\mathbf{A}$ sparse $\not\Rightarrow$ LU-factors sparse

Of course, in case the LU-factors of a sparse matrix possess many more non-zero entries than the matrix itself, the effort for solving a linear system with direct elimination will increase significantly. This can be quantified by means of the following concept:

Definition 2.7.4.3. Fill-in

Let $\mathbf{A} = \mathbf{LU}$ be an LU-factorization (→ Section 2.3.2) of $\mathbf{A} \in \mathbb{K}^{n,n}$. If $l_{ij} \neq 0$ or $u_{ij} \neq 0$ though $a_{ij} = 0$, then we encounter fill-in at position (i,j).

EXAMPLE 2.7.4.4 (Sparse LU-factors) Ex. 2.7.4.1 ➢ massive fill-in can occur for sparse matrices

This example demonstrates that fill-in can largely be avoided, if the matrix has favorable structure. In this case a LSE with this particular system matrix $\mathbf{A}$ can be solved efficiently, that is, with a computational effort $O(\mathrm{nnz}(\mathbf{A}))$ by Gaussian elimination.

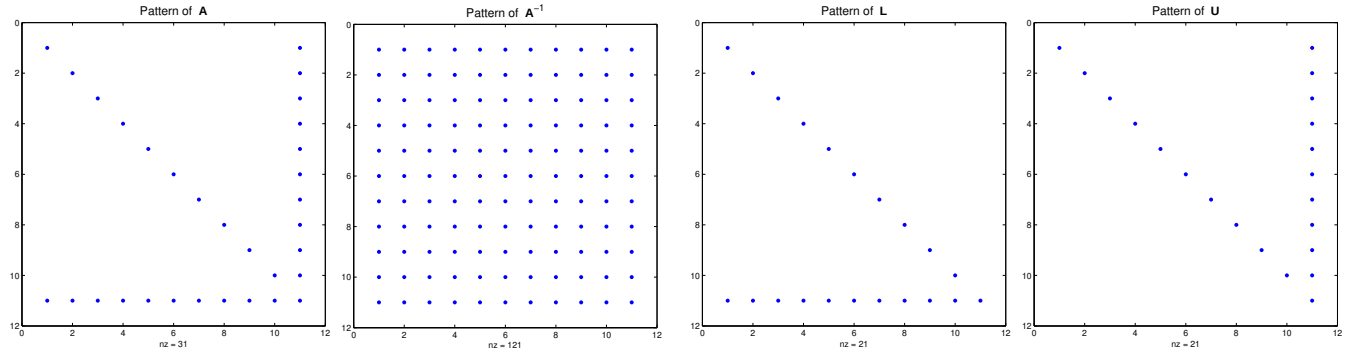
C++ code 2.7.4.5: LU-factorization of sparse matrix ⇝ GITLAB

```
2   // Build matrix
3   MatrixXd A(11, 11);
4   A.setIdentity();
5   A.col(10).setOnes();
6   A.row(10).setOnes();
7   // A.reverseInPlace(); // used inEx. 2.7.4.6
8   auto solver = A.lu();
9   MatrixXd L = MatrixXd::Identity(11, 11);
10  L += solver.matrixLU().triangularView<StrictlyLower>();
11  const MatrixXd U = solver.matrixLU().triangularView<Upper>();
12  const MatrixXd Ainv = A.inverse();
13  // Plotting
```

```
14  spy(A, "Pattern of A", "Apat_cpp.eps");
15  spy(L, "Pattern of L", "Lpat_cpp.eps");
16  spy(U, "Pattern of U", "Upat_cpp.eps");
17  spy(Ainv, "Pattern of A^{-1}", "Ainvpat_cpp.eps");
```

$\mathbf{A}$ is called an "arrow matrix", see the pattern of non-zero entries below and Ex. 2.6.0.5.

Recalling Rem. 2.3.2.17 it is easy to see that the LU-factors of $\mathbf{A}$ will be sparse and that their sparsity patterns will be as depicted below. Observe that despite sparse LU-factors, $\mathbf{A}^{-1}$ will be densely populated.



$\mathbf{L}$, $\mathbf{U}$ sparse $\not\Longrightarrow$ $\mathbf{A}^{-1}$ sparse **!**

Besides stability and efficiency issues, see Exp. 2.4.0.11, this is another reason why using `x = A.inverse()*y` instead of `y = A.lu().solve(b)` is usually a major blunder. ⌟

EXAMPLE 2.7.4.6 (LU-decomposition of flipped "arrow matrix") Recall the discussion in Ex. 2.6.0.5. Here we look at an arrow matrix in a slightly different form:

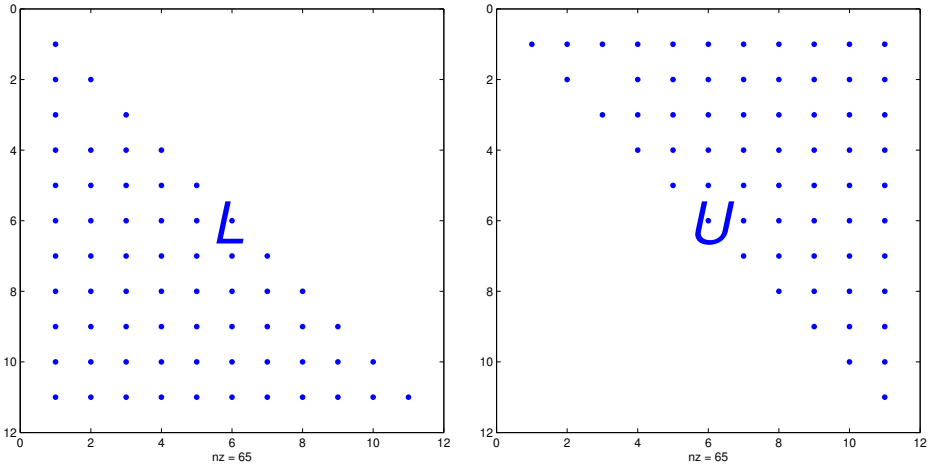
$$\mathbf{M} = \left[\begin{array}{c|c} \alpha & \mathbf{b}^\top \\ \hline \mathbf{c} & \mathbf{D} \end{array}\right] , \quad \begin{array}{l} \alpha \in \mathbb{R} , \\ \mathbf{b}, \mathbf{c} \in \mathbb{R}^{n-1} , \\ \mathbf{D} \in \mathbb{R}^{n-1,n-1} \text{ regular diagonal matrix, } \rightarrow \text{ Def. 1.1.2.3} \end{array} \tag{2.7.4.7}$$

Run the algorithm from § 2.3.2.6 (LU decompisition without pivoting):

- LU-decomposition dense factor matrices with $O(n^2)$ non-zero entries.
- asymptotic computational cost: $O(n^3)$

Output of modified Code 2.7.4.5:

Obvious fill-in (→ Def. 2.7.4.3)



Now it comes as a surprise that the arrow matrix $\mathbf{A}$ from Ex. 2.6.0.5, (2.6.0.6) has sparse LU-factors!

Arrow matrix (2.6.0.6) ▷

$$\mathbf{A} = \begin{bmatrix} \mathbf{D} & \mathbf{c} \\ \mathbf{b}^\top & \alpha \end{bmatrix}$$

$$\mathbf{A} = \underbrace{\begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{b}^\top\mathbf{D}^{-1} & 1 \end{bmatrix}}_{=:\mathbf{L}} \cdot \underbrace{\begin{bmatrix} \mathbf{D} & \mathbf{c} \\ 0 & \sigma \end{bmatrix}}_{=:\mathbf{U}}, \quad \sigma := \alpha - \mathbf{b}^\top\mathbf{D}^{-1}\mathbf{c}\,.$$

➢ In this case LU-factorisation is possible without fill-in, cost merely $O(n)$!

Idea: Transform $\mathbf{A}$ into $\mathbf{A}$ by row and column permutations before performing LU-decomposition.

Details: Apply a cyclic permutation of rows/columns:

- 1st row/column → n-th row/column

- i-th row/column $\to i-1$-th row/column, $i = 2,\ldots,n$

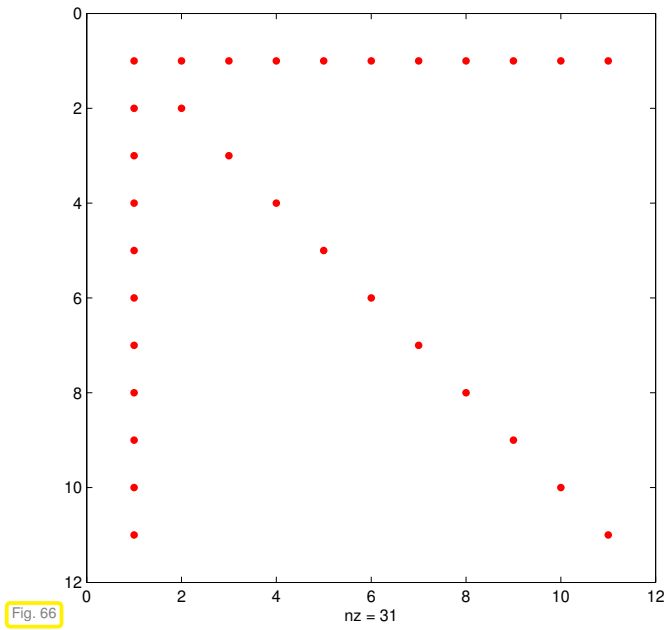


Fig. 66

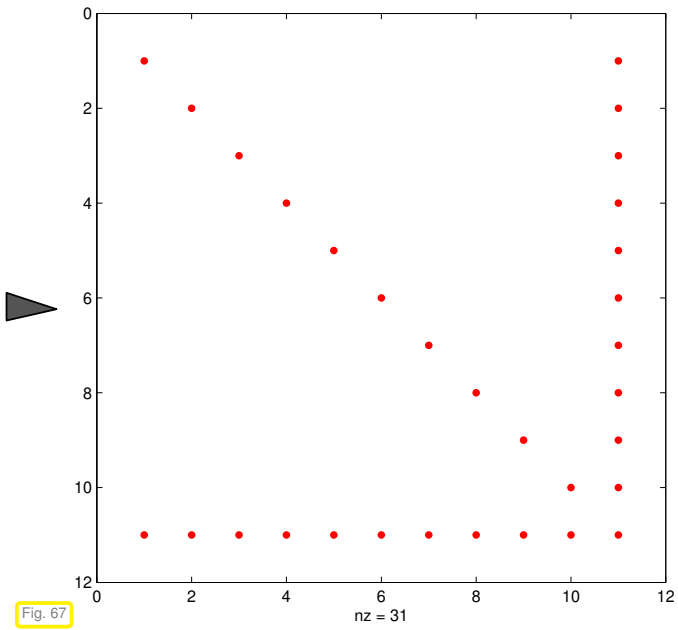


Fig. 67

➢ Then LU-factorization (*without pivoting*) of the resulting matrix requires $O(n)$ operations.

C++ code 2.7.4.8: Permuting arrow matrix, see Fig. 66, Fig. 67 ⇢ GITLAB

```
2  MatrixXd A(11, 11);
3  A.setIdentity();
4  A.col(0).setOnes();
5  A.row(0) = RowVectorXd::LinSpaced(11, 11, 1);
6  // Permutation matrix (→ Def. 2.3.3.13) encoding cyclic
7  // permutation
8  MatrixXd P(11, 11);
9  P.setZero();
10 P.topRightCorner(10, 10).setIdentity();
11 P(10, 0) = 1;
12 spy(A, "A", "InvArrowSpy_cpp.eps");
13 spy(P * A * P.transpose(), "permuted A", "ArrowSpy_cpp.eps");
```

EXAMPLE 2.7.4.9 (Pivoting destroys sparsity) In Ex. 2.7.4.6 we found that permuting a matrix can make it amenable to Gaussian elimination/LU-decomposition with much less fill-in ($\to$ Def. 2.7.4.3). However, recall from Section 2.3.3 that pivoting, which may be essential for achieving numerical stability, amounts to permuting the rows (or even columns) of the matrix. Thus, we may face the awkward situation that pivoting tries to reverse the very permutation we applied to minimize fill-in! The next example shows that this can happen for an arrow matrix.

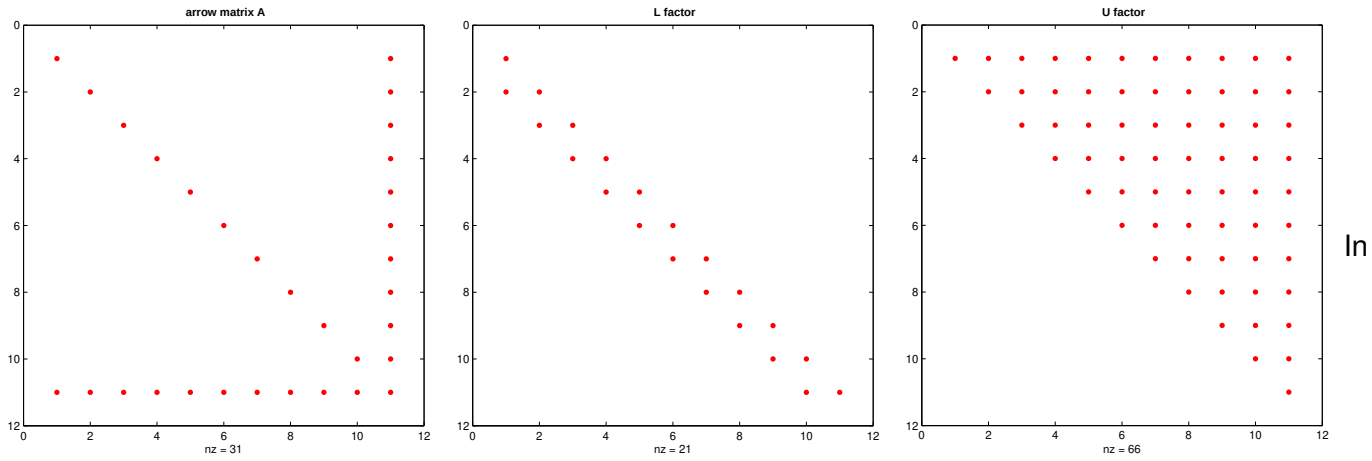
C++ code 2.7.4.10: fill-in due to pivoting ⇢ GITLAB

```
2  // Study of fill-in with LU-factorization due to pivoting
3  Eigen::MatrixXd A(11, 11);
4  A.setZero();
5  A.diagonal() = Eigen::VectorXd::LinSpaced(11, 1, 11).cwiseInverse();
6  A.col(10).setConstant(2);
7  A.row(10).setConstant(2);
```

```
8   auto solver = A.lu();
9   Eigen::MatrixXd L = Eigen::MatrixXd::Identity(11, 11);
10  L += solver.matrixLU().triangularView<Eigen::StrictlyLower>();
11  const Eigen::MatrixXd U = solver.matrixLU().triangularView<Eigen::Upper>();
12  // Plotting
13  spy(A, "Arrow matrix A", "fillinpivotA.eps");
14  spy(L, "L factor", "fillinpivotL.eps");
15  spy(U, "U factor", "fillinpivotU.eps");
16  std::cout << A << std::endl;
```

$$\mathbf{A} = \begin{bmatrix} 1 & & & 2 \\ & \frac{1}{2} & & 2 \\ & & \ddots & \vdots \\ & & \frac{1}{10} & 2 \\ 2 & \cdots & & 2 \end{bmatrix} \quad \rightarrow \quad \text{arrow matrix, Ex. 2.7.4.4}$$

The distributions of non-zero entries of the computed LU-factors ("spy-plots") are as follows:



$\mathbf{A}$ $\mathbf{L}$ $\mathbf{U}$

In this case the solution of a LSE with system matrix $\mathbf{A} \in \mathbb{R}^{n,n}$ of the above type by means of Gaussian elimination with partial pivoting would incur costs of $O(n^3)$. ⌟

2.7.5 Banded Matrices [DR08, Sect. 3.7]

Banded matrices are a special class of sparse matrices (→ Notion 2.7.0.1 with extra structure:

Definition 2.7.5.1. Bandwidth

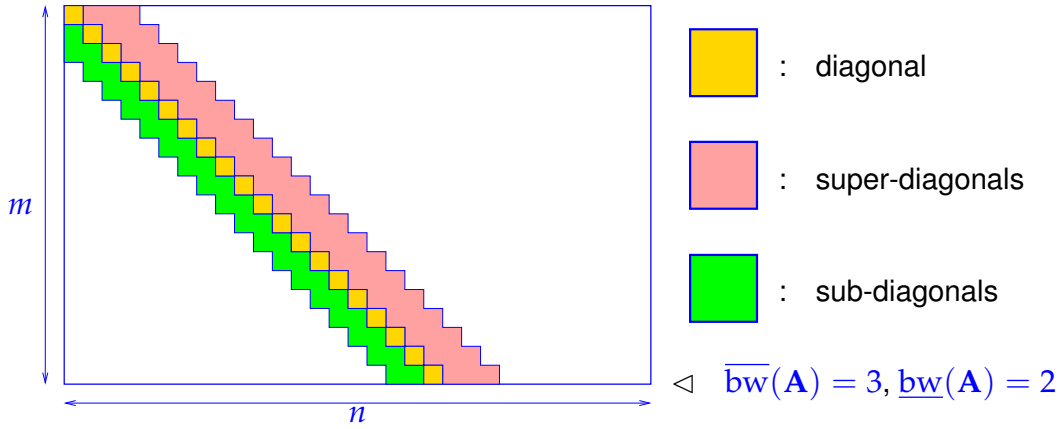
For $\mathbf{A} = (a_{ij})_{i,j} \in \mathbb{K}^{m,n}$ we call

$$\overline{\mathrm{bw}}(\mathbf{A}) := \min\{k \in \mathbb{N}: j - i > k \Rightarrow a_{ij} = 0\} \quad \text{upper bandwidth ,}$$
$$\underline{\mathrm{bw}}(\mathbf{A}) := \min\{k \in \mathbb{N}: i - j > k \Rightarrow a_{ij} = 0\} \quad \text{lower bandwidth .}$$

$\mathrm{bw}(\mathbf{A}) := \overline{\mathrm{bw}}(\mathbf{A}) + \underline{\mathrm{bw}}(\mathbf{A}) + 1$ = **bandwidth** of $\mathbf{A}$.

- $\mathrm{bw}(\mathbf{A}) = 1 \quad \triangleright \quad \mathbf{A}$ diagonal matrix, → Def. 1.1.2.3
- $\overline{\mathrm{bw}}(\mathbf{A}) = \underline{\mathrm{bw}}(\mathbf{A}) = 1 \quad \triangleright \quad \mathbf{A}$ tridiagonal matrix

- More general: $\mathbf{A} \in \mathbb{R}^{n,n}$ with $\mathrm{bw}(\mathbf{A}) \ll n \triangleq$ banded matrix



◁ $\overline{\mathrm{bw}}(\mathbf{A}) = 3, \underline{\mathrm{bw}}(\mathbf{A}) = 2$

for banded matrix $\mathbf{A} \in \mathbb{K}^{m,n}$: $\mathrm{nnz}(\mathbf{A}) \leq \min\{m,n\}\,\mathrm{bw}(\mathbf{A})$

We now examine a generalization of the concept of a banded matrix that is particularly useful in the context of Gaussian elimination:

Definition 2.7.5.2. Matrix envelope

For $\mathbf{A} \in \mathbb{K}^{n,n}$ define

row bandwidth $\mathrm{bw}_i^R(\mathbf{A}) := \max\{0, i-j : a_{ij} \neq 0, 1 \leq j \leq n\}, i \in \{1,\ldots,n\}$

column bandwidth $\mathrm{bw}_j^C(\mathbf{A}) := \max\{0, j-i : a_{ij} \neq 0, 1 \leq i \leq n\}, j \in \{1,\ldots,n\}$

envelope $\mathrm{env}(\mathbf{A}) := \left\{ (i,j) \in \{1,\ldots,n\}^2 : \begin{array}{l} i - \mathrm{bw}_i^R(\mathbf{A}) \leq j \leq i\,, \\ j - \mathrm{bw}_j^C(\mathbf{A}) \leq i \leq j \end{array} \right\}$

EXAMPLE 2.7.5.3 (Envelope of a matrix) We give an example illustrating Def. 2.7.5.2.

$$\mathbf{A} = \begin{bmatrix} * & 0 & * & 0 & 0 & 0 & 0 \\ 0 & * & 0 & 0 & * & 0 & 0 \\ * & 0 & * & 0 & 0 & 0 & * \\ 0 & 0 & 0 & * & * & 0 & * \\ 0 & * & 0 & * & * & * & 0 \\ 0 & 0 & 0 & 0 & * & * & 0 \\ 0 & 0 & * & * & 0 & 0 & * \end{bmatrix} \quad \begin{array}{l} \mathrm{bw}_1^R(A) = 0 \\ \mathrm{bw}_2^R(A) = 0 \\ \mathrm{bw}_3^R(A) = 2 \\ \mathrm{bw}_4^R(A) = 0 \\ \mathrm{bw}_5^R(A) = 3 \\ \mathrm{bw}_6^R(A) = 1 \\ \mathrm{bw}_7^R(A) = 4 \end{array}$$

$\mathrm{env}(A)$ = red entries

$*$ ≙ non-zero matrix entry $a_{ij} \neq 0$

Starting from a "spy-plot", it is easy to find the evelope:

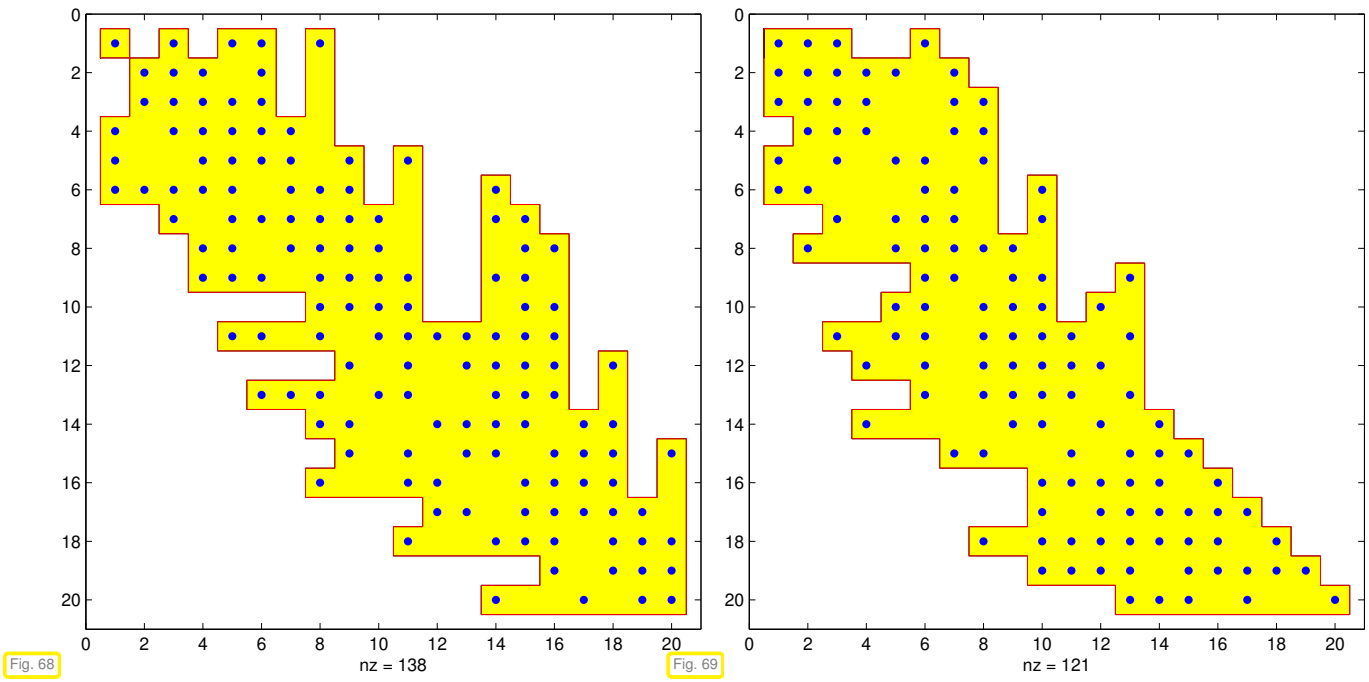


Fig. 68 Fig. 69

Note: the envelope of the arrow matrix from Ex. 2.7.4.4 is just the set of index pairs of its non-zero entries. Hence, the following theorem provides another reason for the sparsity of the LU-factors in that example. ⌟

Theorem 2.7.5.4. Envelope and fill-in → **[QSS00, Sect. 3.9]**

If $\mathbf{A} \in \mathbb{K}^{n,n}$ *is regular with LU-factorization* $\mathbf{A} = \mathbf{LU}$*, then fill-in (→ Def. 2.7.4.3) is confined to* $\mathrm{env}(\mathbf{A})$.

Gaussian elimination *without pivoting*

Proof. (by induction, version I) Examine first step of Gaussian elimination without pivoting, $a_{11} \neq 0$

$$\mathbf{A} = \begin{bmatrix} a_{11} & \mathbf{b}^\top \\ \mathbf{c} & \tilde{\mathbf{A}} \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 0 \\ -\frac{\mathbf{c}}{a_{11}} & \mathbf{I} \end{bmatrix}}_{\mathbf{L}^{(1)}} \underbrace{\begin{bmatrix} a_{11} & \mathbf{b}^\top \\ 0 & \tilde{\mathbf{A}} - \frac{\mathbf{c}\mathbf{b}^\top}{a_{11}} \end{bmatrix}}_{\mathbf{U}^{(1)}}$$

$$\text{If } (i,j) \notin \mathrm{env}(\mathbf{A}) \quad \Rightarrow \quad \begin{cases} c_{i-1} = 0 & \text{, if } i > j\,, \\ b_{j-1} = 0 & \text{, if } i < j\,. \end{cases}$$

$$\Rightarrow \quad \mathrm{env}(\mathbf{L}^{(1)}) \subset \mathrm{env}(\mathbf{A}), \quad \mathrm{env}(\mathbf{U}^{(1)}) \subset \mathrm{env}(\mathbf{A})\,.$$

Moreover, $\mathrm{env}(\tilde{\mathbf{A}} - \frac{\mathbf{c}\mathbf{b}^\top}{a_{11}}) = \mathrm{env}((\mathbf{A})_{2:n,2:n})$ □

Proof. (by induction, version II) Use block-LU-factorization, *cf.* Rem. 2.3.2.19 and proof of Lemma 2.3.2.4:

$$\left[\begin{array}{c|c} \widetilde{\mathbf{A}} & \mathbf{b} \\ \hline \mathbf{c}^\top & \alpha \end{array}\right] = \left[\begin{array}{c|c} \widetilde{\mathbf{L}} & 0 \\ \hline \mathbf{l}^\top & 1 \end{array}\right] \left[\begin{array}{c|c} \widetilde{\mathbf{U}} & \mathbf{u} \\ \hline 0 & \xi \end{array}\right] \quad \Rightarrow \quad \begin{array}{l} \widetilde{\mathbf{U}}^\top \mathbf{l} = \mathbf{c}\,, \\ \widetilde{\mathbf{L}}\mathbf{u} = \mathbf{b}\,. \end{array} \tag{2.7.5.5}$$

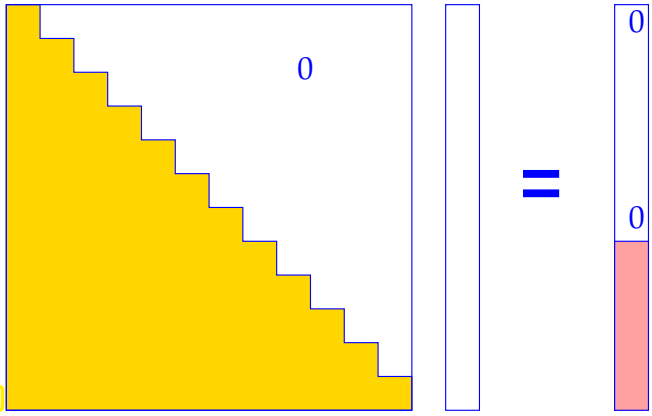


Fig. 70

From Def. 2.7.5.2:

If $m_n^R(\mathbf{A}) = m$, then $c_1, \ldots, c_{n-m} = 0$ (entries of $\mathbf{c}$ from (2.7.5.5))

If $m_n^C(\mathbf{A}) = m$, then $b_1, \ldots, b_{n-m} = 0$ (entries of $\mathbf{b}$ from (2.7.5.5))

$\lhd$ for lower triagular LSE:

If $c_1, \ldots, c_k = 0$ then $l_1, \ldots, l_k = 0$
If $b_1, \ldots, b_k = 0$, then $u_1, \ldots, u_k = 0$

$\Downarrow$

assertion of the theorem $\square$

Thm. 2.7.5.4 immediately suggests a policy for saving cmputational effort when solving linear system whose system matrix $\mathbf{A} \in \mathbb{K}^{n,n}$ is sparse due to *small envelope*:

$$\sharp \mathrm{env}(\mathbf{A}) \ll n^2 :$$

Policy: Confine elimination to envelope!

Details will be given now:

Envelope-aware LU-factorization:

C++ code 2.7.5.6: Computing row bandwidths, → Def. 2.7.5.2 ➽ GITLAB

```
//!  computes rowbandwidth numbers m_i^R(A) of A (sparse
//!  matrix) according to Def. 2.7.5.2
template <class numeric_t>
VectorXi rowbandwidth(const SparseMatrix<numeric_t> &A) {
  VectorXi m = VectorXi::Zero(A.rows());
  for (int k = 0; k < A.outerSize(); ++k) {
    for (typename SparseMatrix<numeric_t>::InnerIterator it(A, k); it; ++it) {
      m(it.row()) =
          std::max<VectorXi::Scalar>(m(it.row()), it.row() - it.col());
    }
  }
  return m;
}
//!  computes row bandwidth numbers m_i^R(A) of A (dense
//!  matrix) according to Def. 2.7.5.2
template <class Derived>
VectorXi rowbandwidth(const MatrixBase<Derived> &A) {
  VectorXi m = VectorXi::Zero(A.rows());
  for (int i = 1; i < A.rows(); ++i) {
    for (int j = 0; j < i; ++j) {
      if (A(i, j) != 0) {
        m(i) = i - j;
        break;
      }
    }
  }
  return m;
}
```

C++ code 2.7.5.7: Envelope aware forward substitution ↪ GITLAB

```
//!  evelope aware forward substitution for Lx = y
//!  (L = lower triangular matrix)
//!  argument mr:  row bandwidth vector
VectorXd substenv(const MatrixXd &L, const VectorXd &y, const VectorXi &mr) {
  const Eigen::Index n = L.cols();
  VectorXd x(n);
  x(0) = y(0) / L(0, 0);
  for (Eigen::Index i = 1; i < n; ++i) {
    if (mr(i) > 0) {
      const double zeta =
          L.row(i).segment(i - mr(i), mr(i)) * x.segment(i - mr(i), mr(i));
      x(i) = (y(i) - zeta) / L(i, i);
    } else {
      x(i) = y(i) / L(i, i);
    }
  }
  return x;
}
```

Asymptotic complexity of envelope aware forward substitution, *cf.* § 2.3.2.15, for $\mathbf{Lx} = \mathbf{y}$, $\mathbf{L} \in \mathbb{K}^{n,n}$ regular lower triangular matrix is

$$O(\#\mathrm{env}(\mathbf{L}))\ !$$

By block LU-factorization ($\rightarrow$ Rem. 2.3.2.19) we find

$$\left[\begin{array}{c|c} (\mathbf{A})_{1:n-1,1:n-1} & (\mathbf{A})_{1:n-1,n} \\ \hline (\mathbf{A})_{n,1:n-1} & (\mathbf{A})_{n,n} \end{array}\right] = \left[\begin{array}{c|c} \mathbf{L}_1 & 0 \\ \hline \mathbf{l}^\top & 1 \end{array}\right] \left[\begin{array}{c|c} \mathbf{U}_1 & \mathbf{u} \\ \hline 0 & \gamma \end{array}\right], \tag{2.7.5.8}$$

$$\Rightarrow \quad (\mathbf{A})_{1:n-1,1:n-1} = \mathbf{L}_1\mathbf{U}_1\,,\quad \mathbf{L}_1\mathbf{u} = (\mathbf{A})_{1:n-1,n}\,,\quad \mathbf{U}_1^\top\mathbf{l} = (\mathbf{A})_{n,1:n-1}^\top\,,\quad \mathbf{l}^\top\mathbf{u} + \gamma = (\mathbf{A})_{n,n}\,. \tag{2.7.5.9}$$

C++ code 2.7.5.10: Envelope aware recursive LU-factorization ↪ GITLAB

```
//!  envelope aware recursive LU-factorization
//!  of structurally symmetric matrix
void luenv(const MatrixXd &A, MatrixXd &L, MatrixXd &U) {
  const Eigen::Index n = A.cols();
  assert(n == A.rows() && "A must be square");
  if (n == 1) {
    L.setIdentity();
    U = A;
  } else {
    VectorXi mr = rowbandwidth(A);  // = colbandwidth thanks to symmetry
    MatrixXd L1(n - 1, n - 1);
    MatrixXd U1(n - 1, n - 1);
    luenv(A.topLeftCorner(n - 1, n - 1), L1, U1);
    VectorXd u = substenv(L1, A.col(n - 1).head(n - 1), mr);
    VectorXd l =
        substenv(U1.transpose(), A.row(n - 1).head(n - 1).transpose(), mr);
    const double gamma = mr(n - 1) > 0 ?
      A(n - 1, n - 1) - l.tail(mr(n - 1)).dot(u.tail(mr(n - 1))) :
      A(n - 1, n - 1);
    L.topLeftCorner(n - 1, n - 1) = L1;
    L.col(n - 1).setZero();
```

```
23      L.row(n - 1).head(n - 1) = l.transpose();
24      L(n - 1, n - 1) = 1;
25      U.topLeftCorner(n - 1, n - 1) = U1;
26      U.col(n - 1).head(n - 1) = u;
27      U.row(n - 1).setZero();
28      U(n - 1, n - 1) = gamma;
29    }
30  }
```

Implementation of envelope aware recursive LU-factorization (no pivoting !)

Assumption: $\mathbf{A} \in \mathbb{K}^{n,n}$ is structurally symmetric

➤ Asymptotic complexity ($\mathbf{A} \in \mathbb{K}^{n,n}$) $O(n \cdot \#\mathrm{env}(\mathbf{A}))$ for $n \to \infty$.

Definition 2.7.5.11. Structurally symmetric matrix

$\mathbf{A} \in \mathbb{K}^{n,n}$ is structurally symmetric, if

$$(\mathbf{A})_{i,j} \neq 0 \quad \Leftrightarrow \quad (\mathbf{A})_{j,i} \neq 0 \quad \forall i,j \in \{1,\ldots,n\} .$$

Since by Thm. 2.7.5.4 fill-in is confined to the envelope, we need store only the matrix entries a_{ij}, $(i,j) \in \mathrm{env}(\mathbf{A})$ when computing (in situ) LU-factorization of *structurally symmetric* $\mathbf{A} \in \mathbb{K}^{n,n}$

➤ Storage required: $n + 2\sum_{i=1}^{n} m_i(\mathbf{A})$ floating point numbers
➤ terminology: envelope oriented matrix storage

EXAMPLE 2.7.5.12 (Envelope oriented matrix storage) Linear envelope oriented matrix storage of *symmetric* $\mathbf{A} = \mathbf{A}^\top \in \mathbb{R}^{n,n}$:

Two arrays:
`scalar_t * val` size P,
`size_t * dptr` size n

$$P := n + \sum_{i=1}^{n} m_i(A) . \qquad (2.7.5.13)$$

$$\mathbf{A} = \begin{bmatrix} * & 0 & * & 0 & 0 & 0 & 0 \\ 0 & * & 0 & 0 & * & 0 & 0 \\ * & 0 & * & 0 & 0 & 0 & * \\ 0 & 0 & 0 & * & * & 0 & * \\ 0 & * & 0 & * & * & * & 0 \\ 0 & 0 & 0 & 0 & * & * & 0 \\ 0 & 0 & * & * & 0 & 0 & * \end{bmatrix}$$

Indexing rule:

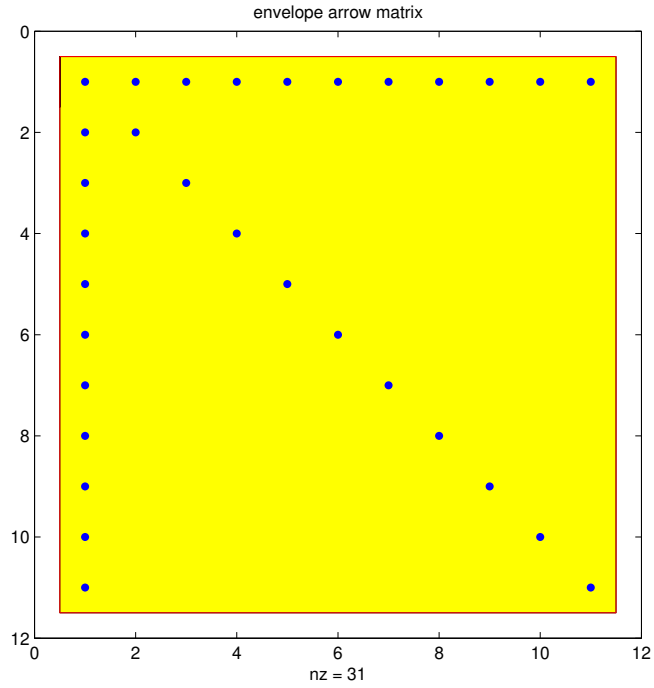
$$\texttt{dptr}[j] = k \quad \Updownarrow \quad \texttt{val}[k] = a_{jj}$$

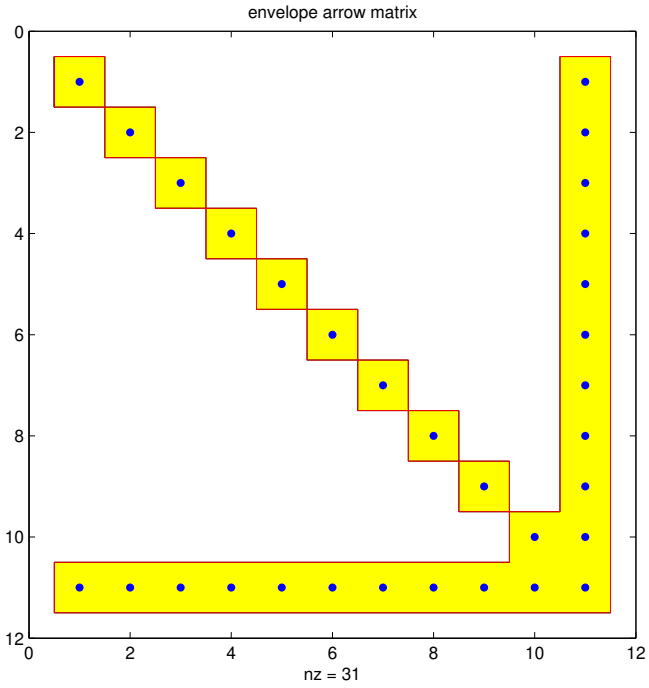
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
val		a_{11}	a_{22}	a_{31}	a_{32}	a_{33}	a_{44}	a_{52}	a_{53}	a_{54}	a_{55}	a_{65}	a_{66}	a_{73}	a_{74}	a_{75}	a_{76}	a_{77}
dptr	0	1	2	5	6	10	12	17										

Minimizing bandwidth/envelope:

Goal: Minimize $m_i(\mathbf{A})$,$\mathbf{A} = (a_{ij}) \in \mathbb{R}^{N,N}$, by permuting rows/columns of $\mathbf{A}$

EXAMPLE 2.7.5.14 (Reducing bandwidth by row/column permutations) Recall: cyclic permutation of rows/columns of arrow matrix applied in Ex. 2.7.4.6. This can be viewed as a drastic shrinking of the envelope:





Another example:Reflection at cross diagonal ➤ reduction of $\#\mathrm{env}(\mathbf{A})$

$$\begin{bmatrix} * & 0 & 0 & * & * & * \\ 0 & * & 0 & 0 & 0 & 0 \\ 0 & 0 & * & 0 & 0 & 0 \\ * & 0 & 0 & * & * & * \\ * & 0 & 0 & * & * & * \\ * & 0 & 0 & * & * & * \end{bmatrix} \longrightarrow \begin{bmatrix} * & * & * & 0 & 0 & * \\ * & * & * & 0 & 0 & * \\ * & * & * & 0 & 0 & * \\ 0 & 0 & 0 & * & 0 & 0 \\ 0 & 0 & 0 & 0 & * & 0 \\ * & * & * & 0 & 0 & * \end{bmatrix}$$

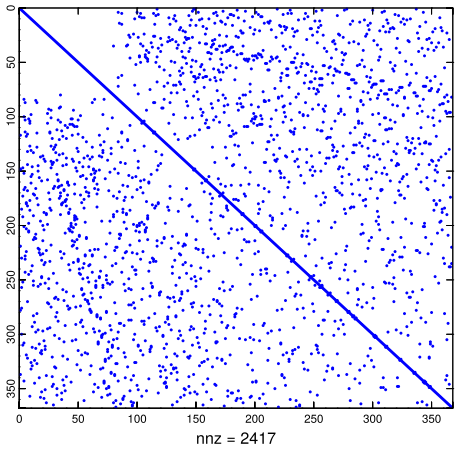
$$i \leftarrow N+1-i$$

$$\#\mathrm{env}(\mathbf{A}) = 30 \#\mathrm{env}(\mathbf{A}) = 22$$

EXAMPLE 2.7.5.15 (Reducing fill-in by reordering)

Envelope reducing permutations are at the heart of all modern sparse solvers (→ § 2.7.3.5). They employ elaborate algorithms for the analysis of matrix graphs, that is, the connections between components of the vector of unknowns defined by nonzero entries of the matrix. For further discussion see [AG11, Sect. 5.7].

EIGEN supplies a few ordering methods for sparse matrices. These methods use permutations to aim for minimal bandwidth/envelop of a given sparse matrix. We study an example with a 347×347 matrix $\mathbf{M}$ originating in the numerical solution of partial differential equations, *cf.* Rem. 2.7.0.5.



Pattern of $\mathbf{M}$ ➥

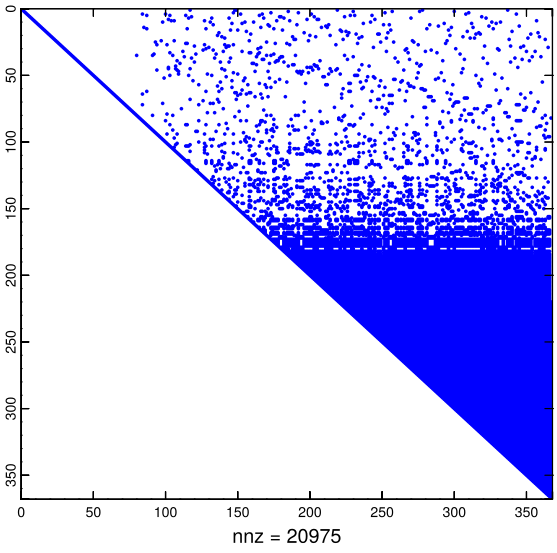
(Here: no row swaps from pivoting !)

C++ code 2.7.5.16: preordering in EIGEN ⇝ GITLAB

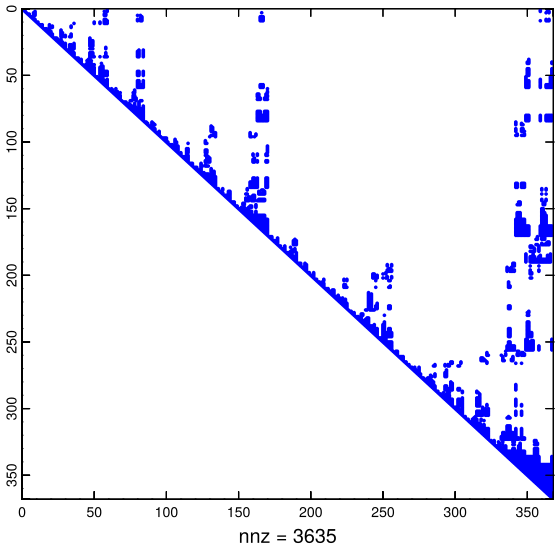
```
2  // L and U cannot be extracted from SparseLU -> LDLT
3  const SimplicialLDLT<SpMat_t, Lower, AMDOrdering<int> > solver1(M);
4  const SimplicialLDLT<SpMat_t, Lower, NaturalOrdering<int> > solver2(M);
5  const MatrixXd U1 =
6      solver1.matrixU() *
```

```
7       MatrixXd::Identity(
8           M.rows(), M.cols()); // explicit conversion fixes occasional
                seqfault
9   const MatrixXd U2 = MatrixXd(solver2.matrixU());
10  // Plotting
11  spy(M, "Sparse matrix M", "MSpy.eps");
12  spy(U1, "U factor (approximate minimum degree)", "AMDUSpy.eps");
13  spy(U2, "U factor (no reordering)", "NaturalUSpy.eps");
```

Examine patterns of LU-factors (→ Section 2.3.2) after reordering:



no reordering



approximate minimum degree

2.8 Stable Gaussian Elimination Without Pivoting

Recall some insights gained, and examples and experimnents seen so far in this chapter:

- Thm. 2.7.5.4 ➢ special structure of the matrix helps avoid fill-in in Gaussian elimination/LU-factorization *without* pivoting.
- Ex. 2.7.4.9 ➢ pivoting can trigger huge fill-in that would not occur without it.
- Ex. 2.7.5.15 ➢ fill-in reducing effect of reordering can be thwarted by later row swapping in the course of pivoting.
- **BUT** pivoting is essential for stability of Gaussian elimination/LU-factorization → Ex. 2.3.3.1.

It would be very desirable to have *a priori criteria*, when Gaussian elimination/LU-factorization remains stable even without pivoting. This can help avoid the extra work for partial pivoting and makes it possible to exploit structure without worrying about stability.

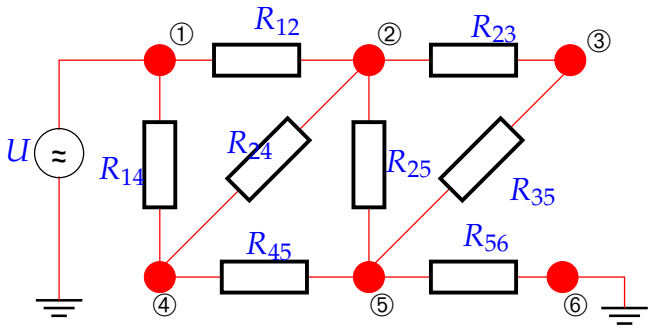
This section will introduce classes of matrices that allow Gaussian elimination without pivoting. Fortunately, linear systems of equations featuring system matrices from these classes are very common in applications.

EXAMPLE 2.8.0.1 (Diagonally dominant matrices from nodal analysis → Ex. 2.1.0.3)

Consider:

electrical circuit entirely composed of Ohmic resistors.

Circuit equations from nodal analysis, see Ex. 2.1.0.3:



$$
\begin{aligned}
② &: \quad R_{12}^{-1}(U_2-U_1)+R_{23}^{-1}(U_2-U_3)+R_{24}^{-1}(U_2-U_4)+R_{25}^{-1}(U_2-U_5) = 0\,,\\
③ &: \quad R_{23}^{-1}(U_3-U_2)+R_{35}^{-1}(U_3-U_5) = 0\,,\\
④ &: \quad R_{14}^{-1}(U_4-U_1)+R_{24}^{-1}(U_4-U_2)+R_{45}^{-1}(U_4-U_5) = 0\,,\\
⑤ &: \quad R_{25}^{-1}(U_5-U_2)+R_{35}^{-1}(U_5-U_3)+R_{45}^{-1}(U_5-U_4)+R_{56}(U_5-U_6) = 0\,,\\
&\qquad U_1 = U \quad , \quad U_6 = 0\,.
\end{aligned}
$$

$$
\begin{bmatrix}
\frac{1}{R_{12}}+\frac{1}{R_{23}}+\frac{1}{R_{24}}+\frac{1}{R_{25}} & -\frac{1}{R_{23}} & -\frac{1}{R_{24}} & -\frac{1}{R_{25}} \\
-\frac{1}{R_{23}} & \frac{1}{R_{23}}+\frac{1}{R_{35}} & 0 & -\frac{1}{R_{35}} \\
-\frac{1}{R_{24}} & 0 & \frac{1}{R_{24}}+\frac{1}{R_{45}} & -\frac{1}{R_{45}} \\
-\frac{1}{R_{25}} & -\frac{1}{R_{35}} & -\frac{1}{R_{45}} & \frac{1}{R_{22}}+\frac{1}{R_{35}}+\frac{1}{R_{45}}+\frac{1}{R_{56}}
\end{bmatrix}
\begin{bmatrix} U_2 \\ U_3 \\ U_4 \\ U_5 \end{bmatrix}
=
\begin{bmatrix} \frac{1}{R_{12}} \\ 0 \\ \frac{1}{R_{14}} \\ 0 \end{bmatrix} U
$$

➢ The matrix $\mathbf{A} \in \mathbb{R}^{n,n}$ arising from nodal analysis satisfies

- $\mathbf{A} = \mathbf{A}^\top \quad , \quad a_{kk} > 0 \quad , \quad a_{kj} \le 0$ for $k \neq j$, (2.8.0.2)
- $\sum_{j=1}^{n} a_{kj} \ge 0\,, \quad k = 1,\ldots,n\,,$ (2.8.0.3)
- $\mathbf{A}$ is regular. (2.8.0.4)

All these properties are obvious except for the fact that $\mathbf{A}$ is regular.

Proof of (2.8.0.4): By Thm. 2.2.1.4 it suffices to show that the nullspace of $\mathbf{A}$ is trivial: $\mathbf{A}\mathbf{x} = 0 \;\Rightarrow\; \mathbf{x} = 0$. So we pick $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{A}\mathbf{x} = 0$, and denote by $i \in \{1,\ldots,n\}$ the index such that

$$|x_i| = \max\{|x_j|,\ j = 1,\ldots,n\}\,.$$

Intermediate goal: show that all entries of $\mathbf{x}$ are the same:

$$\mathbf{A}\mathbf{x} = 0 \;\Rightarrow\; x_i = \sum_{j\neq i} \frac{a_{ij}}{a_{ii}} x_j \;\Rightarrow\; |x_i| \le \sum_{j\neq i} \frac{|a_{ij}|}{|a_{ii}|} |x_j|\,. \tag{2.8.0.5}$$

By (2.8.0.3) and the sign condition from (2.8.0.2) we conclude

$$\sum_{j\neq i} \frac{|a_{ij}|}{|a_{ii}|} \le 1\,. \tag{2.8.0.6}$$

Hence, (2.8.0.6) combined with the above estimate (2.8.0.5) that tells us that the maximum is smaller equal than a mean implies $|x_j| = |x_i|$ for all $j = 1,\ldots,n$. Finally, the sign condition $a_{kj} \le 0$ for $k \neq j$ enforces the same sign of all x_i. Thus, we conclude, w.l.o.g., $x_1 = x_2 = \cdots = x_n$. As

$$\exists i \in \{1,\ldots,n\}\colon \quad \sum_{j=1}^{n} a_{ij} > 0 \quad \text{(strict inequality)}\,,$$

$\mathbf{A}\mathbf{x} = 0$ is only possible for $\mathbf{x} = 0$. ⌟

§2.8.0.7 (Diagonally dominant matrices)

Definition 2.8.0.8. Diagonally dominant matrix → **[QSS00, Def. 1.24]**

$\mathbf{A} \in \mathbb{K}^{n,n}$ is diagonally dominant, if

$$\forall k \in \{1,\ldots,n\}: \quad \sum_{j\neq k} |a_{kj}| \leq |a_{kk}| \ .$$

The matrix $\mathbf{A}$ is called strictly diagonally dominant, if

$$\forall k \in \{1,\ldots,n\}: \quad \sum_{j\neq k} |a_{kj}| < |a_{kk}| \ .$$

Lemma 2.8.0.9. LU-factorization of diagonally dominant matrices

$$\mathbf{A} \text{ } \textit{regular, diagonally dominant with positive diagonal} \quad \Rightarrow \quad \left\{ \begin{array}{c} \mathbf{A} \text{ } \textit{has LU-factorization} \\ \Updownarrow \\ \textit{Gaussian elimination feasible without pivoting}^{(*)} \end{array} \right.$$

$(*)$: In fact, when we apply partial pivoting to a diagonally dominant matrix it will trigger not a single row permutation, because (2.3.3.5) will always be satisfied for $j = k$!

➢ We can dispense with pivoting without compromising stability.

Proof.(of Lemma 2.8.0.9). Appealing to (2.3.1.12) we rely on *induction* w.r.t. n:

It is clear that partial pivoting in the first step selects a_{11} as pivot element, *cf.* (2.3.3.5). Thus after the 1st step of elimination we obtain the modified entries

$$a_{ij}^{(1)} = a_{ij} - \frac{a_{i1}}{a_{11}} a_{1j} \ , \quad i,j = 2,\ldots,n \quad \Rightarrow \quad a_{ii}^{(1)} > 0 \ ,$$

which we conclude from diagonal dominance. That also permits us to infer

$$\begin{aligned} |a_{ii}^{(1)}| - \sum_{\substack{j=2 \\ j\neq i}}^{n} |a_{ij}^{(1)}| &= \left| a_{ii} - \frac{a_{i1}}{a_{11}} a_{1i} \right| - \sum_{\substack{j=2 \\ j\neq i}}^{n} \left| a_{ij} - \frac{a_{i1}}{a_{11}} a_{1j} \right| \\ &\geq a_{ii} - \frac{|a_{i1}||a_{1i}|}{a_{11}} - \sum_{\substack{j=2 \\ j\neq i}}^{n} |a_{ij}| - \frac{|a_{i1}|}{a_{11}} \sum_{\substack{j=2 \\ j\neq i}}^{n} |a_{1j}| \\ &\geq a_{ii} - \frac{|a_{i1}||a_{1i}|}{a_{11}} - \sum_{\substack{j=2 \\ j\neq i}}^{n} |a_{ij}| - |a_{i1}| \frac{a_{11} - |a_{1i}|}{a_{11}} \geq a_{ii} - \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \geq 0 \ . \end{aligned}$$

$\mathbf{A}$ regular, diagonally dominant ⇒ partial pivoting according to (2.3.3.5) selects i-th row in i-th step. ⌟

§2.8.0.10 (Gaussian elimination for symmetric positive definite (s.p.d.) matrices) The class of symmetric positive definite (s.p.d.) matrices has been defined in Def. 1.1.2.6. They permit stable Gaussian elimintation without pivoting:

Theorem 2.8.0.11. Gaussian elimination for s.p.d. matrices

Every symmetric/Hermitian positive definite matrix (s.p.d. → Def. 1.1.2.6) possesses an LU-decomposition (→ Section 2.3.2).

Equivalent to the assertion of the theorem is the assertion that for s.p.d. matrices Gaussian elimination is feasible *without pivoting*.

In fact, this theorem is a corollary of Lemma 2.3.2.4, because all principal minors of an s.p.d. matrix are s.p.d. themselves. However, we outline an alternative self-contained proof:

Proof. (of Thm. 2.8.0.11) we pursue a proof by *induction* with respect to the matrix size n. The assertion in the case $n = 1$ is obviously true.

For the induction argument $n-1 \Rightarrow n$ consider the first step of the elimination algorithm

$$\mathbf{A} = \begin{bmatrix} a_{11} & \mathbf{b}^\top \\ \mathbf{b} & \widetilde{\mathbf{A}} \end{bmatrix} \xrightarrow[\text{Gaussian elimination}]{\text{1. step}} \begin{bmatrix} a_{11} & \mathbf{b}^\top \\ 0 & \widetilde{\mathbf{A}} - \frac{\mathbf{b}\mathbf{b}^\top}{a_{11}} \end{bmatrix} .$$

This step has not problem, because all diagonal entries of an s.p.d. matrix are strictly positive.

The induction requires us to show that the right-lower block $\widetilde{\mathbf{A}} - \frac{\mathbf{b}\mathbf{b}^\top}{a_{11}} \in \mathbb{R}^{n-1,n-1}$ is also symmetric and positive definite. Its symmetry is evident, but the demonstration of the s.p.d. property relies on a trick: As $\mathbf{A}$ ist s.p.d. (→ Def. 1.1.2.6), for every $\boldsymbol{y} \in \mathbb{R}^{n-1} \setminus \{0\}$

$$0 < \begin{bmatrix} -\frac{\mathbf{b}^\top \mathbf{y}}{a_{11}} \\ \mathbf{y} \end{bmatrix}^\top \begin{bmatrix} a_{11} & \mathbf{b}^\top \\ \mathbf{b} & \widetilde{\mathbf{A}} \end{bmatrix} \begin{bmatrix} -\frac{\mathbf{b}^\top \mathbf{y}}{a_{11}} \\ \mathbf{y} \end{bmatrix} = \boldsymbol{y}^\top (\widetilde{\mathbf{A}} - \frac{\mathbf{b}\mathbf{b}^\top}{a_{11}}) \boldsymbol{y} . \tag{2.8.0.12}$$

We conclude that $\widetilde{\mathbf{A}} - \frac{\mathbf{b}\mathbf{b}^\top}{a_{11}}$ positive definite. Thus, according to the induction hypothesis, Gaussian elimination without pivoting can now be applied to that right-lower block.

□

The proof can also be based on the identities

$$\left[\begin{array}{c|c} (\mathbf{A})_{1:n-1,1:n-1} & (\mathbf{A})_{1:n-1,n} \\ \hline (\mathbf{A})_{n,1:n-1} & (\mathbf{A})_{n,n} \end{array}\right] = \left[\begin{array}{c|c} \mathbf{L}_1 & 0 \\ \hline \mathbf{l}^\top & 1 \end{array}\right] \left[\begin{array}{c|c} \mathbf{U}_1 & \mathbf{u} \\ \hline 0 & \gamma \end{array}\right] , \tag{2.7.5.8}$$

$$\Rightarrow \quad (\mathbf{A})_{1:n-1,1:n-1} = \mathbf{L}_1\mathbf{U}_1 , \quad \mathbf{L}_1\mathbf{u} = (\mathbf{A})_{1:n-1,n} , \quad \mathbf{U}_1^\top \mathbf{l} = (\mathbf{A})_{n,1:n-1}^\top , \quad \mathbf{l}^\top \mathbf{u} + \gamma = (\mathbf{A})_{n,n} ,$$

noticing that the principal minor $(\mathbf{A})_{1:n-1,1:n-1}$ is also s.p.d. This allows a simple induction argument.

Note: no pivoting required (→ Section 2.3.3)
(partial pivoting always picks current pivot row)

The next result gives a useful criterion for telling whether a given symmetric/Hermitian matrix is s.p.d.:

Lemma 2.8.0.13. Diagonal dominance and definiteness

A diagonally dominant Hermitian/symmetric matrix with non-negative diagonal entries is positive semi-*definite.*
A strictly *diagonally dominant Hermitian/symmetric matrix with positive diagonal entries is positive definite.*

Proof. For $\mathbf{A} = \mathbf{A}^{\mathrm{H}}$ diagonally dominant, use inequality between arithmetic and geometric mean (AGM) $ab \le \frac{1}{2}(a^2 + b^2)$:

$$\begin{aligned}
\mathbf{x}^{\mathrm{H}}\mathbf{A}\mathbf{x} &= \sum_{i=1}^{n}\Big(a_{ii}|x_i|^2 + \sum_{i\neq j} a_{ij}\bar{x}_i x_j\Big) \ge \sum_{i=1}^{n}\Big(a_{ii}|x_i|^2 - \sum_{i\neq j}|a_{ij}||x_i||x_j|\Big)\\
&\overset{\text{AGM}}{\ge} \sum_{i=1}^{n} a_{ii}|x_i|^2 - \tfrac{1}{2}\sum_{i\neq j}|a_{ij}|(|x_i|^2 + |x_j|^2)\\
&\ge \tfrac{1}{2}\Big(\sum_{i=1}^{n}\{a_{ii}|x_i|^2 - \sum_{j\neq i}|a_{ij}||x_i|^2\}\Big) + \tfrac{1}{2}\Big(\sum_{j=1}^{n}\{a_{ii}|x_j|^2 - \sum_{i\neq j}|a_{ij}||x_j|^2\}\Big)\\
&\ge \sum_{i=1}^{n}|x_i|^2\Big(a_{ii} - \sum_{j\neq i}|a_{ij}|\Big) \ge 0\,.
\end{aligned}$$

⌟

§2.8.0.14 (Cholesky decomposition)

Lemma 2.8.0.15. Cholesky decomposition for s.p.d. matrices → [Gut09, Sect. 3.4], [Han02, Sect. II.5], [QSS00, Thm. 3.6]

For any s.p.d. $\mathbf{A} \in \mathbb{K}^{n,n}$, $n \in \mathbb{N}$, there is a unique upper triangular matrix $\mathbf{R} \in \mathbb{K}^{n,n}$ with $r_{ii} > 0$, $i = 1,\ldots,n$, such that $\mathbf{A} = \mathbf{R}^{\mathrm{H}}\mathbf{R}$ (Cholesky decomposition).

Proof. Thm. 2.8.0.11, Lemma 2.3.2.4 ensure the existence of a *unique LU-decomposition* of $\mathbf{A}$: $\mathbf{A} = \mathbf{L}\mathbf{U}$, which we can rewrite as follows:

$$\mathbf{A} = \mathbf{L}\mathbf{D}\widetilde{\mathbf{U}}\quad,\qquad \begin{array}{l}\mathbf{D} \mathrel{\widehat{=}} \text{diagonal of } \mathbf{U}\,,\\ \widetilde{\mathbf{U}} \mathrel{\widehat{=}} \textit{normalized} \text{ upper triangular matrix} \to \text{Def. 1.1.2.3}\end{array}$$

Due to the uniqueness of the LU-decomposition we infer

$$\mathbf{A} = \mathbf{A}^{\top} \;\Rightarrow\; \mathbf{U} = \mathbf{D}\mathbf{L}^{\top} \;\Rightarrow\; \boxed{\mathbf{A} = \mathbf{L}\mathbf{D}\mathbf{L}^{\top}}\,,$$

with unique $\mathbf{L}$, $\mathbf{D}$ (diagonal matrix)

$$\mathbf{x}^{\top}\mathbf{A}\mathbf{x} > 0 \quad \forall \mathbf{x} \neq 0 \;\Rightarrow\; \mathbf{y}^{\top}\mathbf{D}\mathbf{y} > 0 \quad \forall \mathbf{y} \neq 0\,.$$

➤ The diagonal matrix D has a positive diagonal and, hence, we can take its "square root" and choose $\mathbf{R} := \sqrt{\mathbf{D}}\mathbf{L}^{\top}$.

□

We find formulas analogous to (2.3.2.7)

$$\mathbf{R}^{\mathrm{H}}\mathbf{R} = \mathbf{A} \;\Rightarrow\; a_{ik} = \sum_{j=1}^{\min\{i,k\}} \bar{r}_{ji}r_{jk} = \begin{cases} \sum_{j=1}^{i-1}\bar{r}_{ji}r_{jk} + \bar{r}_{ii}r_{ik} & \text{, if } i < k\,,\\ \sum_{j=1}^{i-1}|r_{ji}|^2 + r_{ii}^2 & \text{, if } i = k\,.\end{cases} \tag{2.8.0.16}$$

C++ code 2.8.0.17: Simple Cholesky factorization ⇝ GITLAB

```
2 //!  simple Cholesky factorization
```

```
3  void cholfac(const MatrixXd &A, MatrixXd &R){
4      const Eigen::Index n = A.rows();
5      R = A;
6      for(Eigen::Index k = 0; k < n; ++k){
7          for(Eigen::Index j = k+1; j < n; ++j) {
8              R.row(j).tail(n-j) -= R.row(k).tail(n-j)*R(k,j)/R(k,k);
9          }
10         R.row(k).tail(n-k) /= std::sqrt(R(k,k));
11     }
12     R.triangularView<StrictlyLower>().setZero();
13 }
```

Cost of Cholesky decomposition

The asymptotic computational cost (# elementary arithmetic operations) of computing the Cholesky decomposition of an $n \times n$ s.p.d. matrix is $\frac{1}{6}n^3 + O(n^2)$ for matrix size $n \to \infty$.

This is "half the costs" of computing a general LU-factorization, *cf.* Code in § 2.3.2.6, but this does not mean "twice as fast" in a concrete implementation, because memory access patterns will have a crucial impact, see Rem. 1.4.1.5.

Gains of efficiency hardly justify the use of Cholesky decomposition in modern numerical algorithms. Savings in memory compared to standard LU-factorization (only one factor $\mathbf{R}$ has to be stored) offer a stronger reason to prefer the Cholesky decomposition. ⌟

§2.8.0.19 (Cholesky-type decompositions in EIGEN) Hardly surprising, EIGEN provides library routines for the computation of the (generalized) Cholesky decomposition of an symmetric (positive definite) matrix. For dense or sparse matrices these are the methods (→ EIGEN documentation)

- `LLT()` for computing a genuine Cholesky decomposition,
- `LDLT()` for computing a factorization $\mathbf{A} = \mathbf{L}\mathbf{D}\mathbf{L}^\top$ with a normalized lower-triangular matrix $\mathbf{L}$ and a diagonal matrix $\mathbf{D}$.

These methods are invoked like all other matrix decomposition methods, refer to § 2.5.0.8, where `solverType` is to be replaced with either `LLT` or `LDLT`. Rem. 2.5.0.10 also applies. The LDLT-decomposition can be attempted for any symmetric matrix, but need not exist. ⌟

§2.8.0.20 (Numerical stability of Cholesky decomposition) The computation of Cholesky-factorization by means of the algorithm of Code 2.8.0.17 is numerically stable (→ Def. 1.5.5.19)!

To understand this recall Thm. 2.4.0.5: Numerical instability of Gaussian elimination (with any kind of pivoting) manifests itself in massive growth of the entries of the intermediate matrices $\mathbf{A}^{(k)}$ arising during elimination. Then use the relationship between LU-factorization and Cholesky decomposition, which tells us that we only have to monitor the growth of entries of intermediate upper triangular "Cholesky factorization matrices" $\mathbf{A} = (\mathbf{R}^{(k)})^{\mathrm{H}}\mathbf{R}^{(k)}$. We consider the Euclidean vector norm/matrix norm (→ Def. 1.5.5.10) $\|\cdot\|_2$

$$\mathbf{A} = \mathbf{R}^{\mathrm{H}}\mathbf{R} \quad \Rightarrow \quad \|\mathbf{A}\|_2 = \sup_{\|\mathbf{x}\|_2=1} \mathbf{x}^{\mathrm{H}}\mathbf{R}^{\mathrm{H}}\mathbf{R}\mathbf{x} = \sup_{\|\mathbf{x}\|_2=1} (\mathbf{R}\mathbf{x})^{\mathrm{H}}(\mathbf{R}\mathbf{x}) = \|\mathbf{R}\|_2^2 .$$

➤ For all intermediate Cholesky factorization matrices holds: $\left\|(\mathbf{R}^{(k)})^{\mathrm{H}}\right\|_2 = \left\|\mathbf{R}^{(k)}\right\|_2 = \|\mathbf{A}\|_2^{1/2}$! Of course, this rules out a blowup of entries of the $\mathbf{R}^{(k)}$.

Computation of the Cholesky decomposition largely agrees with the computation of LU-factorization (without pivoting). Using the latter together with forward and backward substitution (→ Section 2.3.2) to solve a linear system of equations is algebraically and numerically equivalent to using Gaussian elimination without pivoting. From these equivalences we conclude:

Solving LSE with s.p.d. system matrix via
Cholesky decomposition **+** forward & backward substitution
is numerically stable (→ Def. 1.5.5.19)

⇕

Gaussian elimination for s.p.d. matrices

Gaussian elimination *without pivoting* is a *numerically stable* way to solve LSEs with s.p.d. system matrix.

⌟

Learning Outcomes

Principal take-home knowledge and skills from this chapter:

- A clear understanding of the algorithm of Gaussian elimination with and without pivoting (prerequisite knowledge from linear algebra)
- Insight into the relationship between Gaussian elimination and LU-decomposition and the algorithmic relevance of LU-decomposition
- Awareness of the asymptotic complexity of dense Gaussian elimination, LU-decomposition, and elimination for special matrices
- Familiarity with "sparse matrices": notion, data structures, initialization, benefits
- Insight into the reduced computational complexity of the direct solution of sparse linear systems of equations with special structural properties.

Bibliography

[AG11] Uri M. Ascher and Chen Greif. *A first course in numerical methods*. Vol. 7. Computational Science & Engineering. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2011, pp. xxii+552. DOI: 10.1137/1.9780898719987 (cit. on pp. 136, 144, 205).

[DR08] W. Dahmen and A. Reusken. *Numerik für Ingenieure und Naturwissenschaftler*. Heidelberg: Springer, 2008 (cit. on pp. 145, 152, 153, 156, 157, 199).

[GGK14] W. Gander, M.J. Gander, and F. Kwok. *Scientific Computing*. Vol. 11. Texts in Computational Science and Engineering. Heidelberg: Springer, 2014 (cit. on p. 132).

[GV89] G.H. Golub and C.F. Van Loan. *Matrix computations*. 2nd. Baltimore, London: John Hopkins University Press, 1989 (cit. on p. 159).

[Gut09] M.H. Gutknecht. *Lineare Algebra*. Lecture Notes. SAM, ETH Zürich, 2009 (cit. on pp. 130, 136, 137, 143–147, 210).

[Han02] M. Hanke-Bourgeois. *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens*. Mathematische Leitfäden. Stuttgart: B.G. Teubner, 2002 (cit. on pp. 143, 144, 156, 210).

[Hig02] N.J. Higham. *Accuracy and Stability of Numerical Algorithms*. 2nd ed. Philadelphia, PA: SIAM, 2002 (cit. on p. 159).

[NS02] K. Nipp and D. Stoffer. *Lineare Algebra*. 5th ed. Zürich: vdf Hochschulverlag, 2002 (cit. on pp. 126, 130, 136, 137, 139, 143–145, 151, 153, 155).

[QSS00] A. Quarteroni, R. Sacco, and F. Saleri. *Numerical mathematics*. Vol. 37. Texts in Applied Mathematics. New York: Springer, 2000 (cit. on pp. 127, 130, 132, 136, 143, 146, 201, 208, 210).

[SST06] A. Sankar, D.A. Spielman, and S.-H. Teng. "Smoothed analysis of the condition numbers and growth factors of matrices". In: *SIAM J. Matrix Anal. Appl.* 28.2 (2006), pp. 446–476 (cit. on p. 162).

[SG04] O. Schenk and K. Gärtner. "Solving Unsymmetric Sparse Systems of Linear Equations with PARDISO". In: *J. Future Generation Computer Systems* 20.3 (2004), pp. 475–487 (cit. on p. 193).

[ST96] D.A. Spielman and Shang-Hua Teng. "Spectral partitioning works: planar graphs and finite element meshes". In: *Foundations of Computer Science, 1996. Proceedings., 37th Annual Symposium on*. Oct. 1996, pp. 96–105. DOI: 10.1109/SFCS.1996.548468 (cit. on p. 161).

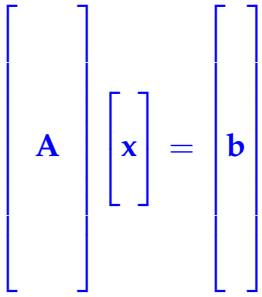
[TB97] L.N. Trefethen and D. Bau. *Numerical Linear Algebra*. Philadelphia, PA: SIAM, 1997 (cit. on pp. 159, 161).

Chapter 3

Direct Methods for Linear Least Squares Problems

In this chapter we study numerical methods for **overdetermined** (OD) linear systems of equations, that is, linear systems with a "tall" rectangular system matrix

$$\mathbf{x} \in \mathbb{R}^n\text{: } \text{“}\mathbf{A}\mathbf{x} = \mathbf{b}\text{”}\,, \qquad (3.0.0.1)$$
$$\mathbf{b} \in \mathbb{R}^m\,,\quad \mathbf{A} \in \mathbb{R}^{m,n}\,,\quad m \geq n\,.$$



We point out that, in contrast to Chapter 1, Chapter 2, we will restrict ourselves to **real** linear systems in this chapter.

Note that the quotation marks in (3.0.0.1) indicate that this is not a well-defined problem in the sense of § 1.5.5.1; $\mathbf{A}\mathbf{x} = \mathbf{b}$ does no define a mapping $(\mathbf{A}, \mathbf{b}) \mapsto \mathbf{x}$, because

- such a vector $\mathbf{x} \in \mathbb{R}^n$ may not exist,
- and, even if it exists, it may not be unique.

Therefore, first we have to establish a crisp concept of that we mean by a "solution" of (3.0.0.1).

Contents

3.0.1 Overdetermined Linear Systems of Equations: Examples . . . 215
3.1 Least Squares Solution Concepts . . . 218
3.1.1 Least Squares Solutions: Definition . . . 219
3.1.2 Normal Equations . . . 222
3.1.3 Moore-Penrose Pseudoinverse . . . 228
3.1.4 Sensitivity of Least Squares Problems . . . 230
3.2 Normal Equation Methods [DR08, Sect. 4.2], [Han02, Ch. 11] . . . 232
3.3 Orthogonal Transformation Methods [DR08, Sect. 4.4.2] . . . 236
3.3.1 Transformation Idea . . . 236
3.3.2 Orthogonal/Unitary Matrices . . . 238
3.3.3 QR-Decomposition [Han02, Sect. 13], [Gut09, Sect. 7.3] . . . 238
3.3.4 QR-Based Solver for Linear Least Squares Problems . . . 254
3.3.5 Modification Techniques for QR-Decomposition . . . 259

3.4 Singular Value Decomposition (SVD) 266
3.4.1 SVD: Definition and Theory 266
3.4.2 SVD in EIGEN 271
3.4.3 Solving General Least-Squares Problems by SVD 274
3.4.4 SVD-Based Optimization and Approximation 277
3.5 Total Least Squares 298
3.6 Constrained Least Squares 299
3.6.1 Solution via Lagrangian Multipliers 300
3.6.2 Solution via SVD 302

3.0.1 Overdetermined Linear Systems of Equations: Examples

Video tutorial for Section 3.0.1 "Overdetermined Linear Systems of Equations: Examples": (12 minutes) Download link, tablet notes

→ review questions 3.0.1.11

You may think that overdetermined linear systems of equations are exotic, but this is not true. Rather they are very common in mathematical models.

EXAMPLE 3.0.1.1 (Linear parameter estimation in 1D) From first principles it is *known* that two physical quantities $x \in \mathbb{R}$ and $y \in \mathbb{R}$ (e.g., pressure and density of an ideal gas) are related by a linear relationship

$$y = \alpha x + \beta \quad \text{for some unknown coefficients/parameters} \quad \alpha, \beta \in \mathbb{R} \,. \tag{3.0.1.2}$$

We carry out $m \in \mathbb{N}$ measurements that yield pairs $(x_i, y_i) \in \mathbb{R}^2$, $i = 1, \ldots, m$, $m \geq 2$. If the measurements were perfect, we could expect that there exist $\alpha, \beta \in \mathbb{R}$ such that $y_i = \alpha x_i + \beta$ for all $i = 1, \ldots, m$. This is an overdetermined linear system of equations of the form (3.0.0.1):

$$\begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ \vdots & \vdots \\ x_m & 1 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_m \end{bmatrix} \quad \leftrightarrow \quad \mathbf{A}\mathbf{x} = \mathbf{b} \,, \quad \mathbf{A} \in \mathbb{R}^{m,2} \,, \mathbf{b} \in \mathbb{R}^m \,, \mathbf{x} \in \mathbb{R}^2 \,. \tag{3.0.1.3}$$

In practice inevitable ("random") *measurement errors* will affect the y_is, push the vector $\mathbf{b}$ out of the range/image $\mathcal{R}(\mathbf{A})$ of $\mathbf{A}$ (→ Def. 2.2.1.2), and thwart the solvability of (3.0.1.3). Assuming independent and randomly distributed measurement errors in the y_i, for $m > 2$ the probability that a solution $\begin{bmatrix} \alpha \\ \beta \end{bmatrix}$ exists is actually zero, see Rem. 3.1.0.2. ⌟

EXAMPLE 3.0.1.4 (Linear regression: Parameter estimation for a linear model) Ex. 3.0.1.1 can be generalized to higher dimensions:

Given: *measured* data points $(\mathbf{x}_i, y_i)$, $\mathbf{x}_i \in \mathbb{R}^n$, $y_i \in \mathbb{R}$, $i = 1, \ldots, m$, $m \geq n+1$ (y_i, $\mathbf{x}_i$ affected by measurement errors).

Known: without measurement errors data would satisfy an affine linear relationship $y = \mathbf{a}^\top \mathbf{x} + \beta$, for some $\mathbf{a} \in \mathbb{R}^n$, $c \in \mathbb{R}$.

Plugging in the measured quantities gives $y_i = \mathbf{a}^\top \mathbf{x}_i + \beta$, $i = 1, \ldots, m$, a linear system of equations of

the form

$$\begin{bmatrix} \mathbf{x}_1^\top & 1 \\ \vdots & \vdots \\ \mathbf{x}_m^\top & 1 \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ \beta \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} \quad \leftrightarrow \quad \mathbf{A}\mathbf{x} = \mathbf{b}\,, \quad \mathbf{A} \in \mathbb{R}^{m,n+1}\,, \mathbf{b} \in \mathbb{R}^m\,, \mathbf{x} \in \mathbb{R}^{n+1}\,, \tag{3.0.1.5}$$

which is an overdetermined LSE, in case $m > n+1$. ⌟

EXAMPLE 3.0.1.6 (Measuring the angles of a triangle [NS02, Sect. 5.1]) We measure the angles of a planar triangle and obtain $\widetilde{\alpha}, \widetilde{\beta}, \widetilde{\gamma}$ (in radians). In the case of perfect measurements the true angles α, β, γ would satisfy

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix} = \begin{bmatrix} \widetilde{\alpha} \\ \widetilde{\beta} \\ \widetilde{\gamma} \\ \pi \end{bmatrix}\,. \tag{3.0.1.7}$$

Measurement errors will inevitably make the measured angles fail to add up to π so that (3.0.1.7) will not have a solution $[\alpha, \beta, \gamma]^\top$.

Then, why should we add this last equation? This is suggested by a tenet of data science that reads "You cannot afford not to use any piece of information available". It turns out that solving (3.0.1.7) "in a suitable way" as discussed below in Section 3.1.1 enhances cancellation of measurement errors and gives better estimates for the angles. We will not discuss this here and refer to statistics for an explanation.

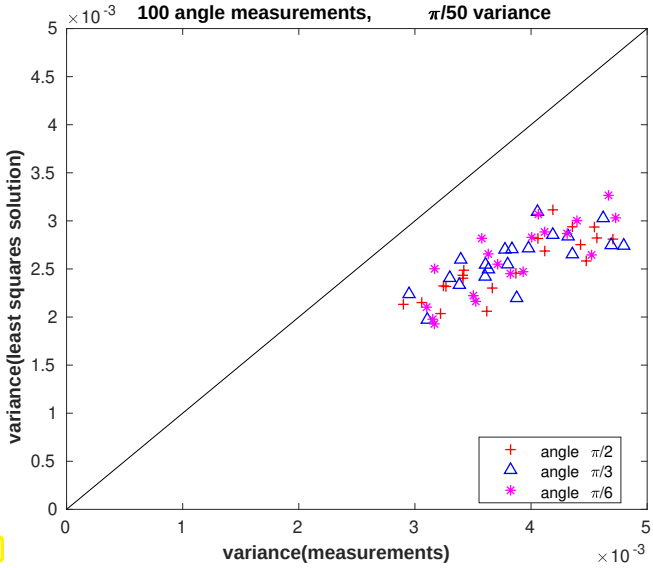


Fig. 71

Here we just report the results of a numerical experiment:
We consider the triangle with angles $\pi/2$, $\pi/3$, and $\pi/6$. Synthetic "measurement errors" are introduced by adding a normally distributed random perturbation with mean 0 and standard deviation $\pi/50$ to the exact values of the angles, yielding $\widetilde{\alpha}$, $\widetilde{\beta}$, and $\widetilde{\gamma}$.
For 100 "measurements" we compute the variance of the raw angles and that of the estimates obtained by solving (3.0.1.7) in least squares sense ($\rightarrow$ Section 3.1.1). These variances are plotted for many different "runs".

We observe that in most runs the variance of the estimates from (3.0.1.7) are smaller than those of the raw data. ⌟

EXAMPLE 3.0.1.8 (Angles in a triangulation)

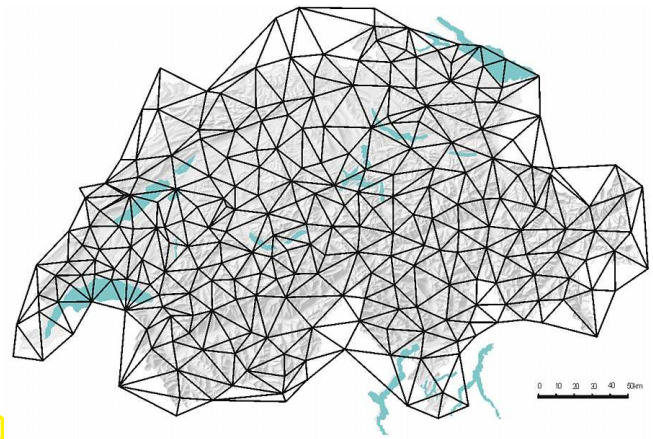


Fig. 72

In Ex. 2.7.2.5 we learned about the concept and data structures for planar triangulations $\rightarrow$ Def. 2.7.2.6. Such triangulations have been and continue to be of fundament importance for geodesy. In particular before distances could be measured accurately by means of lasers, triangulations were indispensable, because angles could already be determined with high precision. *C.F. Gauss* pioneered both the use of triangulations in geodesy and the use of the least squares method to deal with measurement errors $\rightarrow$ Wikipedia.

Die Grundlagen seines Verfahrens hatte Gauss schon 1795 im Alter von 18 Jahren entwickelt. Basis war eine Idee von Pierre-Simon Laplace, die Beträge von Fehlern aufzusummieren, so dass sich die Fehler zu Null addieren. Gauss nahm stattdessen die Fehlerquadrate und konnte die künstliche Zusatzanforderung an die Fehler weglassen.

Gauss benutzte dann das Verfahren intensiv bei seiner Vermessung des Königreichs Hannover durch Triangulation. 1821 und 1823 erschien die zweiteilige Arbeit sowie 1826 eine Ergänzung zur *Theoria combinationis observationum erroribus minimis obnoxiae* (Theorie der den kleinsten Fehlern unterworfenen Kombination der Beobachtungen), in denen Gauss eine Begründung liefern konnte, weshalb sein Verfahren im Vergleich zu den anderen so erfolgreich war: Die Methode der kleinsten Quadrate ist in einer breiten Hinsicht optimal, also besser als andere Methoden.

We now extend Ex. 3.0.1.6 to planar triangulations, for which *measured* values for all internal angles are available. We obtain an overdetermined system of equations by combining the following linear relations:

1. each angle is supposed to be equal to its measured value,
2. the sum of interior angles is π for every triangle,
3. the sum of the angles at an interior node is 2π.

If the planar triangulation has N_0 interior vertices and M cells, then we end up with $4M + N_0$ equations for the $3M$ unknown angles. ⌟

EXAMPLE 3.0.1.9 ((Relative) point locations from distances [GGK14, Sect. 6.1]) Consider n points located on the real axis at unknown locations $x_i \in \mathbb{R}$, $i = 1,\dots,n$. At least we know that $x_i < x_{i+1}$, $i = 1,\dots,n-1$.

We *measure* the $m := \binom{n}{2} = \frac{1}{2}n(n-1)$ pairwise distances $d_{ij} := |x_i - x_j|$, $i,j \in \{1,\dots,n\}$, $i \neq j$. They are connected to the point positions by the overdetermined linear system of equations

$$\begin{array}{c} x_i - x_j = d_{ij}\,, \\ 1 \le j < i \le n\,. \\ \updownarrow \\ \mathbf{A}\mathbf{x} = \mathbf{b} \end{array} \quad \leftrightarrow \quad \begin{bmatrix} -1 & 1 & 0 & \dots & & \dots & 0 \\ -1 & 0 & 1 & 0 & & & \\ \vdots & & \ddots & \ddots & & & \\ \vdots & & & \ddots & \ddots & & \\ -1 & \dots & & & & 0 & 1 \\ 0 & -1 & 1 & & & & 0 \\ \vdots & & & & & & \vdots \\ \vdots & & & & & & \vdots \\ 0 & \dots & & & & -1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} d_{12} \\ d_{13} \\ \vdots \\ \vdots \\ d_{1n} \\ d_{23} \\ \vdots \\ \vdots \\ d_{n-1,n} \end{bmatrix} \tag{3.0.1.10}$$

Note that we can never expect a unique solution for $\mathbf{x} \in \mathbb{R}^n$, because adding a multiple of $[1,1,\dots,1]^\top$ to any solution will again yield a solution, because $\mathbf{A}$ has a non-trivial kernel: $\mathcal{N}(\mathbf{A}) = [1,1,\dots,1]^\top$. Non-uniqueness can be cured by setting $x_1 := 0$, thus removing one component of $\mathbf{x}$.

If the measurements were perfect, we could then find $x_2,\dots,x_n$ from $d_{i-1,i}$, $i = 2,\dots,n$ by solving a standard (square) linear system of equations. However, as in Ex. 3.0.1.6, using much more information through the overdetermined system (3.0.1.10) helps curb measurement errors. ⌟

Review question(s) 3.0.1.11 (Overdetermined Linear Systems of Equations: Examples)

(Q3.0.1.11.A) The mass of three different items is measures in all possible combinations. Find the overdetermined linear system of equations (LSE) that has to be "solved" to obtain estimates for the three masses. What would be the size of the corresponding overdetermined LSE for $m \in \mathbb{N}$ masses?

(Q3.0.1.11.B) A time-harmonic current with frequency $f > 0$ can be written as

$$I(t) = a\cos(2\pi ft) + b\sin(2\pi ft)\,, \quad t \in \mathbb{R}\,, \quad a,b \in \mathbb{R}\,.$$

It is measured at times $t_k = \frac{k}{mf}$, $k = 0,\ldots,m-1$, $m > 2$, and we obtain values $I_k \approx I(t_k)$. Which overdetermined linear system of equations can be used to estimate the coefficients a, b?

(Q3.0.1.11.C) [A $\int_0^1 e^x \,\mathrm{d}e$-type problem] We know the solution $\mathbf{x} \in \mathbb{R}^n$ and the right-hand-side vector $\mathbf{b} \in \mathbb{R}^n$ of the $n \times n$ (Toeplitz) tridiagonal linear system of equations

$$\begin{bmatrix}
\alpha & \beta & 0 & \ldots & & & & & \ldots & 0 \\
\beta & \alpha & \ddots & & & & & & & \vdots \\
0 & \beta & \ddots & \ddots & & & & & & \\
\vdots & & \ddots & \ddots & \ddots & & & & & \\
 & & & \ddots & \ddots & \ddots & & & & \\
 & & & & \ddots & \ddots & \ddots & & & \\
 & & & & & \ddots & \ddots & \ddots & \ddots & \\
 & & & & & & \ddots & \ddots & \ddots & \\
 & & & & & & & \ddots & & \vdots \\
 & & & & & & & \ddots & \ddots & 0 \\
\vdots & & & & & & & & \beta & \alpha & \beta \\
0 & \ldots & & & & & & \ldots & 0 & \beta & \alpha
\end{bmatrix} \mathbf{x} = \mathbf{b}\,.$$

Which overdetermined linear system of equations of maximal size has the vector $[\alpha,\beta]^\top \in \mathbb{R}^2$ as its solution?

(Q3.0.1.11.D) [Visualization of 2D linear regression] We know that solving a 2×2 linear system of equations can be visualized as finding the intersection of two lines in the plane, lines given as solution sets of the rows of the linear system, recall Ex. 2.2.2.10.

Develop a similar visualization for the task of finding the solution of an overdetermined linear system of equations $\mathbf{Ax} = \mathbf{b}$, $\mathbf{A} \in \mathbb{R}^{m,2}$ ($m > 2$), $\mathbf{b} \in \mathbb{R}^m$.

△

3.1 Least Squares Solution Concepts

Video tutorial for Section 3.1.1 "Least Squares Solutions": (9 minutes) Download link, tablet notes

→ review questions 3.1.1.14

Throughout we consider the (possibly overdetermined) linear system of equations

$$\mathbf{x} \in \mathbb{R}^n\!: \quad \text{“}\mathbf{Ax} = \mathbf{b}\text{”}\,, \quad \mathbf{b} \in \mathbb{R}^m\,, \quad \mathbf{A} \in \mathbb{R}^{m,n}\,, \quad m \geq n\,. \tag{3.0.0.1}$$

Recall from linear algebra that $\mathbf{Ax} = \mathbf{b}$ has a solution, if and only if the right hand side vector $\mathbf{b}$ lies in the image (range space, → Def. 2.2.1.2) of the matrix $\mathbf{A}$:

$$\exists \mathbf{x} \in \mathbb{R}^n\!: \quad \mathbf{Ax} = \mathbf{b} \quad \Leftrightarrow \quad \mathbf{b} \in \mathcal{R}(\mathbf{A})\,. \tag{3.1.0.1}$$

✎ Notation for important subspaces associated with a matrix $\mathbf{A} \in \mathbb{K}^{m,n}$ (→ Def. 2.2.1.2)

$$\text{image/range:} \quad \mathcal{R}(\mathbf{A}) := \{\mathbf{A}\mathbf{x},\ \mathbf{x} \in \mathbb{K}^n\} \subset \mathbb{K}^m\ ,$$
$$\text{kernel/nullspace:} \quad \mathcal{N}(\mathbf{A}) := \{\mathbf{x} \in \mathbb{K}^n\colon \mathbf{A}\mathbf{x} = \mathbf{0}\}\ .$$

Remark 3.1.0.2 (Consistent right hand side vectors are highly improbable) If $\mathcal{R}(\mathbf{A}) \neq \mathbb{R}^m$, then "almost all" perturbations of $\mathbf{b}$ (e.g., due to measurement errors) will destroy $\mathbf{b} \in \mathcal{R}(\mathbf{A})$, because $\mathcal{R}(\mathbf{A})$ is a "set of measure zero" in $\mathbb{R}^m$. ⌟

3.1.1 Least Squares Solutions: Definition

Definition 3.1.1.1. Least squares solution

For given $\mathbf{A} \in \mathbb{R}^{m,n}$, $\mathbf{b} \in \mathbb{R}^m$ the vector $\mathbf{x} \in \mathbb{R}^n$ is a least squares solution of the linear system of equations $\mathbf{A}\mathbf{x} = \mathbf{b}$, if

$$\mathbf{x} \in \operatorname*{argmin}_{\mathbf{y}\in\mathbb{R}^n} \|\mathbf{A}\mathbf{y} - \mathbf{b}\|_2^2\ ,$$
$$\Updownarrow$$
$$\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 = \min_{\mathbf{y}\in\mathbb{R}^n} \|\mathbf{A}\mathbf{y} - \mathbf{b}\|_2^2 = \min_{y_1,\ldots,y_n\in\mathbb{R}} \sum_{i=1}^m \left(\sum_{j=1}^n (\mathbf{A})_{i,j} y_j - (\mathbf{b})_i\right)^2\ .$$

➡ A least squares solution is any vector $\mathbf{x}$ that minimizes the Euclidean norm of the residual $\mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x}$, see Def. 2.4.0.1.

We write $\mathrm{lsq}(\mathbf{A},\mathbf{b})$ for the set of least squares solutions of the linear system of equations $\mathbf{A}\mathbf{x} = \mathbf{b}$, $\mathbf{A} \in \mathbb{R}^{m,n}$, $\mathbf{b} \in \mathbb{R}^m$:

$$\mathrm{lsq}(\mathbf{A},\mathbf{b}) := \{\mathbf{x} \in \mathbb{R}^n\colon \mathbf{x} \text{ is a least squares solution of } \mathbf{A}\mathbf{x} = \mathbf{b}\} \subset \mathbb{R}^n\ . \tag{3.1.1.2}$$

§3.1.1.3 (Least squares solutions and " true" solutions of LSE) The concept of least squares solutions is a genuine *generalization* of what is regarded as a solution of linear system of equations in linear algebra: Clearly, for a square linear system of equations with regular system matrix *the* least squares solution agrees with the "true" solution:

$$\mathbf{A} \in \mathbb{R}^{n,n} \text{ regular} \quad \Rightarrow \quad \mathrm{lsq}(\mathbf{A},\mathbf{b}) = \{\mathbf{A}^{-1}\mathbf{b}\} \quad \forall \mathbf{b} \in \mathbb{R}^n\ . \tag{3.1.1.4}$$

Also for $\mathbf{A} \in \mathbb{R}^{m,n}$, $m > n$, the set $\mathrm{lsq}(\mathbf{A},\mathbf{b})$ contains only "true" solutions of $\mathbf{A}\mathbf{x} = \mathbf{b}$, if $\mathbf{b} \in \mathcal{R}(\mathbf{A})$, because in this case the smallest residual is $\mathbf{0}$. ⌟

Next, we examine least squares solutions from a geometric perspective.

EXAMPLE 3.1.1.5 (linear regression → [DR08, Ex. 4.1]) We consider the problem of parameter estimation for a linear model from Ex. 3.0.1.4:

Given: *measured* data points $(\mathbf{x}_i, y_i)$, $\mathbf{x}_i \in \mathbb{R}^n$, $y_i \in \mathbb{R}$, $i = 1,\ldots,m$, $m \geq n+1$
(y_i, $\mathbf{x}_i$ affected by measurement errors).

Known: without measurement errors data would satisfy affine linear relationship

$$y = \mathbf{a}^\top \mathbf{x} + \beta \,, \tag{3.1.1.6}$$

from some parameters $\mathbf{a} \in \mathbb{R}^n$, $\beta \in \mathbb{R}$.

Solving the overdetermined linear system of equations in least squares sense we obtain a *least squares estimate* for the parameters $\mathbf{a}$ and β:

$$(\mathbf{a}, \beta) = \operatorname*{argmin}_{\mathbf{a}\in\mathbb{R}^n, \beta\in\mathbb{R}} \sum_{i=1}^{m} |y_i - \mathbf{a}^\top \mathbf{x}_i - \beta|^2 \tag{3.1.1.7}$$

In statistics, solving (3.1.1.7) is known as linear regression

Linear regression for $n = 1, m = 8$: "fitting" a regression line to data points ▷

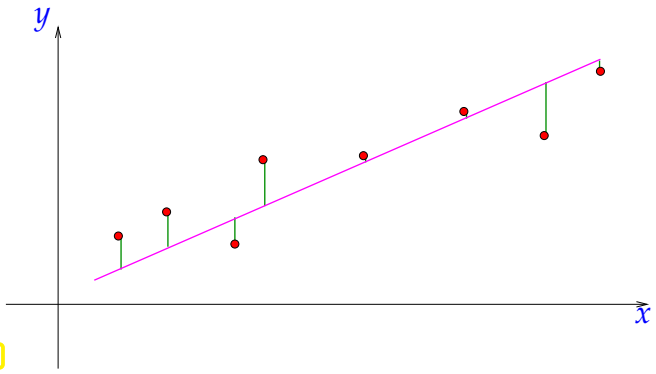


Fig. 73

⌟

§3.1.1.8 (The geometry of least squares problems) A geometric "proof" for the existence of least squares solutions ($\mathbb{R} = \mathbb{R}$)

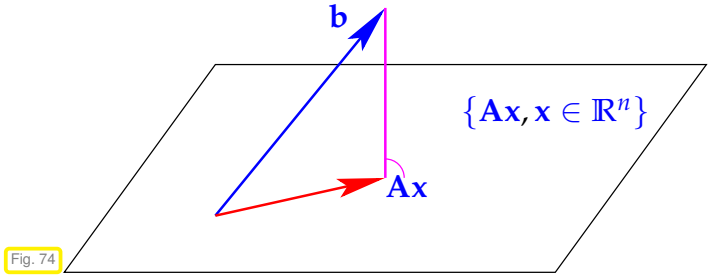


Fig. 74

◁ For a least squares solution $\mathbf{x} \in \mathbb{R}^n$ the vector $\mathbf{Ax} \in \mathbb{R}^m$ is the unique orthogonal projection of $\mathbf{b}$ onto

$$\mathcal{R}(\mathbf{A}) = \operatorname{Span}\{(\mathbf{A})_{:,1}, \ldots, (\mathbf{A})_{:,n}\} \,,$$

because the orthogonal projection provides the nearest (w.r.t. the Euclidean distance) point to $\mathbf{b}$ in the subspace (hyperplane) $\mathcal{R}(\mathbf{A})$.

From this geometric consideration we conclude that $\operatorname{lsq}(\mathbf{A}, \mathbf{b})$ is the space of solutions of $\mathbf{Ax} = \mathbf{b}^*$, where $\mathbf{b}^*$ is the orthogonal projection of $\mathbf{b}$ onto $\mathcal{R}(\mathbf{A})$. Since the set of solutions of a linear system of equations invariably is an affine space, this argument teaches that $\operatorname{lsq}(\mathbf{A}, \mathbf{b})$ is an affine subspace of $\mathbb{R}^n$!

⌟

Geometric intuition yields the following insight:

Theorem 3.1.1.9. Existence of least squares solutions

For any $\mathbf{A} \in \mathbb{R}^{m,n}$, $\mathbf{b} \in \mathbb{R}^m$ a least squares solution of $\mathbf{Ax} = \mathbf{b}$ ($\to$ Def. 3.1.1.1) exists.

Proof. The function $F : \mathbb{R}^n \to \mathbb{R}$, $F(\mathbf{x}) := \|\mathbf{b} - \mathbf{Ax}\|_2^2$ is continuous, bounded from below by 0 and $F(\mathbf{x}) \to \infty$ for $\|\mathbf{x}\| \to \infty$. Hence, there must be an $\mathbf{x}* \in \mathbb{R}^n$ for which it attains its minimum. □

§3.1.1.10 (Least squares solution as maximum-likelihood estimator → [DR08, Sect. 4.5]) Extending the considerations of Ex. 3.0.1.4, a generic *linear parameter estimation* problem seeks to determine the unknown parameter vector $\mathbf{x} \in \mathbb{R}^n$ from the linear relationship $\mathbf{Ax} = \mathbf{y}$, where $\mathbf{A} \in \mathbb{R}^m$ is known and $\mathbf{y} \in \mathbb{R}^m$ is accessible through measurements.

Unfortunately, $\mathbf{y}$ is affected by measurement errors, Thus we model it as a random vector $\mathbf{y} = \mathbf{y}(\omega)$, $\omega \in \Omega$, Ω the set of outcomes from a probability space.

The measurement errors in different components of $\mathbf{y}$ are supposed be *unbiased* (expectation $= 0$), *independent*, and *identically normally distributed* with variance σ^2, $\sigma > 0$, which means

$$\mathbf{y}(\omega) = \mathbf{A}\mathbf{x} + \mathbf{e}(\omega) , \quad \omega \in \Omega , \tag{3.1.1.11}$$

where the probability distribution of $\mathbf{e}$ satisfies

$$\mathbb{P}((\mathbf{e})_\ell \in I_\ell,\ \ell = 1,\ldots,m) = \prod_{\ell=1}^{m} \frac{1}{\sigma\sqrt{2\pi}} \int_{I_\ell} \exp\left(-\tfrac{1}{2}\frac{z^2}{\sigma^2}\right) \mathrm{d}z , \quad I_\ell \subset \mathbb{R} . \tag{3.1.1.12}$$

From (3.1.1.11) we infer that the probability density of $\mathbf{y}$ is

$$L(\mathbf{x};\mathbf{y}) = \prod_{\ell=1}^{m} \exp\left(-\tfrac{1}{2}\left(\frac{y_\ell - (\mathbf{A}\mathbf{x})_\ell}{\sigma}\right)^2\right) = \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2\right) . \tag{3.1.1.13}$$

The last identity follows from $\exp(x)\exp(y) = \exp(x+y)$. This probability density function $\mathbf{y} \mapsto L(\mathbf{x};\mathbf{y})$ is called the **likelihood** of $\mathbf{y}$, and the notation emphasizes its dependence on the parameters $\mathbf{x}$.

Assume that we are given a measurement (realization/sample) $\mathbf{b}$ of $\mathbf{y}$. The maximum likelihood principle then suggests that we choose the parameters so that the probability density of $\mathbf{y}$ becomes maximal at $\mathbf{b}$:

$$\mathbf{x}^* \in \mathbb{R}^n \quad \text{such that} \quad \mathbf{x}^* = \operatorname*{argmax}_{\mathbf{x}\in\mathbb{R}^n} L(\mathbf{x};\mathbf{b}) = \operatorname*{argmax}_{\mathbf{x}\in\mathbb{R}^n} \exp(-\frac{1}{2\sigma^2}\|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2) .$$

Obviously, due to the monotonicity of $\xi \mapsto \exp(\xi)$, $\mathbf{x}^*$ is a least squares solution of $\mathbf{A}\mathbf{x} = \mathbf{b}$ according to Def. 3.1.1.1:

$$\mathbf{x}^* \in \mathbb{R}^n \quad \text{such that} \quad \mathbf{x}^* = \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 .$$

⌟

Review question(s) 3.1.1.14 (Least squares solution: Definition)

(Q3.1.1.14.A) Describe $\mathbf{A} \in \mathbb{R}^{2,2}$ and $\mathbf{b} \in \mathbb{R}^2$ so that $\mathrm{lsq}(\mathbf{A},\mathbf{b})$ contains more than a single vector.

(Q3.1.1.14.B) What is $\mathrm{lsq}(\mathbf{A},\mathbf{0})$ for $\mathbf{A} \in \mathbb{R}^{m,n}$?

(Q3.1.1.14.C) Given a matrix $\mathbf{B} \in \mathbb{R}^{m,n}$, a vector $\mathbf{c} \in \mathbb{R}^m$, and $\lambda > 0$, define

$$\{\mathbf{x}^*\} := \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{B}\mathbf{x} - \mathbf{c}\|_2^2 + \lambda\|\mathbf{x}\|_2^2 \subset \mathbb{R}^n .$$

State an overdetermined linear system of equations $\mathbf{A}\mathbf{x} = \mathbf{b}$, of which $\mathbf{x}^*$ is a least-squares solution.

(Q3.1.1.14.D) [The geometry of least-squares problems]

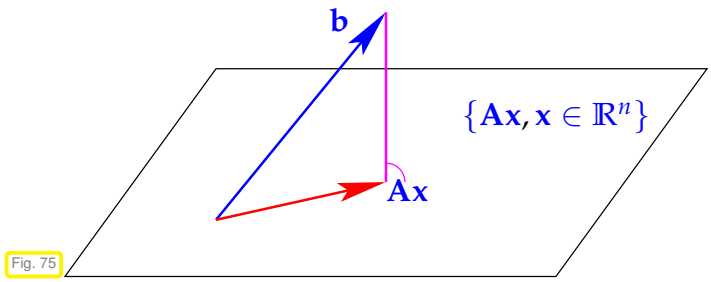


Fig. 75

Explain why the figure beside illustrates the concept of a least-squares solution of an overdetermined linear system of equations $\mathbf{A}\mathbf{x} = \mathbf{b}$, $\mathbf{A} \in \mathbb{R}^{m,n}$, $m > n$.

△

3.1.2 Normal Equations

Video tutorial for Section 3.1.2 "Normal Equations": (16 minutes) Download link, tablet notes

→ review questions 3.1.2.23

Appealing to the geometric intuition gleaned from Fig. 74 we infer the orthogonality of $\mathbf{b} - A\mathbf{x}$, $\mathbf{x}$ a least squares solution of the overdetermined linear systems of equations $\mathbf{A}\mathbf{x} = \mathbf{b}$, to all columns of $\mathbf{A}$:

$$\mathbf{b} - \mathbf{A}\mathbf{x} \perp \mathcal{R}(\mathbf{A}) \quad \Leftrightarrow \quad \mathbf{b} - \mathbf{A}\mathbf{x} \perp (\mathbf{A})_{:,j}\,, \quad j = 1,\ldots,n \quad \Leftrightarrow \quad \mathbf{A}^\top(\mathbf{b} - \mathbf{A}\mathbf{x}) = \mathbf{0}\,.$$

Surprisingly, we have found a square linear system of equations satisfied by the least squares solution. The next theorem gives the formal statement is this discovery. It also completely characterizes $\mathrm{lsq}(\mathbf{A},\mathbf{b})$ and reveals a way to compute this set.

Theorem 3.1.2.1. Obtaining least squares solutions by solving normal equations

The vector $\mathbf{x} \in \mathbb{R}^n$ is a least squares solution (→ Def. 3.1.1.1) of the linear system of equations $\mathbf{A}\mathbf{x} = \mathbf{b}$, $\mathbf{A} \in \mathbb{R}^{m,n}$, $\mathbf{b} \in \mathbb{R}^m$, if and only if *it solves the normal equations (NEQ)*

$$\boxed{\mathbf{A}^\top\mathbf{A}\mathbf{x} = \mathbf{A}^\top\mathbf{b}}\,. \tag{3.1.2.2}$$

Note that the normal equations (3.1.2.2) are an $n \times n$ square linear system of equations with a *symmetric positive semi-definite* coefficient matrix:

$$\begin{bmatrix} \mathbf{A}^\top \end{bmatrix}\begin{bmatrix} \mathbf{A} \end{bmatrix}\begin{bmatrix} \mathbf{x} \end{bmatrix} = \begin{bmatrix} \mathbf{A}^\top \end{bmatrix}\begin{bmatrix} \mathbf{b} \end{bmatrix},$$

$$\Leftrightarrow \quad \begin{bmatrix} \mathbf{A}^\top\mathbf{A} \end{bmatrix}\begin{bmatrix} \mathbf{x} \end{bmatrix} = \begin{bmatrix} \mathbf{A}^\top \end{bmatrix}\begin{bmatrix} \mathbf{b} \end{bmatrix}.$$

Proof. (of Thm. 3.1.2.1)

❶: We first show that a least squares solution satisfies the normal equations. Let $\mathbf{x} \in \mathbb{R}^n$ be a least squares solutions according to Def. 3.1.1.1. Pick an arbitrary $\mathbf{d} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ and define the function

$$\varphi_{\mathbf{d}} : \mathbb{R} \to \mathbb{R} \quad , \quad \varphi_{\mathbf{d}}(\tau) := \|\mathbf{A}(\mathbf{x} + \tau\mathbf{d}) - \mathbf{b}\|_2^2\,. \tag{3.1.2.3}$$

We find the equivalent expression

$$\varphi_{\mathbf{d}}(\tau) = \tau^2\mathbf{d}^\top\mathbf{A}^\top\mathbf{A}\mathbf{d} + 2\tau\mathbf{d}^\top\mathbf{A}^\top(\mathbf{A}\mathbf{x} - \mathbf{b}) + \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2\,,$$

which shows that $\tau \mapsto \varphi_{\mathbf{d}}(\tau)$ is a smooth (C^∞) function.

Moreover, since every $\mathbf{x} \in \mathrm{lsq}(\mathbf{A},\mathbf{b})$ is a minimizer of $\mathbf{y} \mapsto \|\mathbf{A}\mathbf{y}-\mathbf{b}\|_2^2$, we conclude that $\tau \mapsto \varphi_{\mathbf{d}}(\tau)$ has a global minimum in $\tau = 0$. Necessarily,

$$\frac{d\varphi_{\mathbf{d}}}{d\tau}_{|\tau=0} = 2\mathbf{d}^\top\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{b}) = 0\,.$$

Since this holds for *any* vector $\mathbf{d} \neq 0$, we conclude (set $\mathbf{d}$ equal to all the Euclidean unit vectors in $\mathbb{R}^n$)

$$\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{b}) = \mathbf{0}\,,$$

which is equivalent to the normal equations (3.1.2.2).

❷: Let $\mathbf{x}$ be a solution of the normal equations. Then we find by tedious but straightforward computations

$$\begin{aligned}
&\|\mathbf{A}\mathbf{y}-\mathbf{b}\|_2^2 - \|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2\\
&\quad = \mathbf{y}^\top\mathbf{A}^\top\mathbf{A}\mathbf{y} - 2\mathbf{y}^\top\mathbf{A}^\top\mathbf{b} + \mathbf{b}^\top\mathbf{b} - \mathbf{x}^\top\mathbf{A}^\top\mathbf{A}\mathbf{x} + 2\mathbf{x}^\top\mathbf{A}^\top\mathbf{b} - \mathbf{b}^\top\mathbf{b}\\
&\overset{(3.1.2.2)}{=} \mathbf{y}^\top\mathbf{A}^\top\mathbf{A}\mathbf{y} - 2\mathbf{y}^\top\mathbf{A}^\top\mathbf{A}\mathbf{x} + \mathbf{x}^\top\mathbf{A}^\top\mathbf{A}\mathbf{x}\\
&\quad = (\mathbf{y}-\mathbf{x})^\top\mathbf{A}^\top\mathbf{A}(\mathbf{y}-\mathbf{x}) = \|\mathbf{A}(\mathbf{x}-\mathbf{y})\|_2^2 \geq 0\,.\\
&\qquad\qquad \Longrightarrow \quad \|\mathbf{A}\mathbf{y}-\mathbf{b}\| \geq \|\mathbf{A}\mathbf{x}-\mathbf{b}\|\,.
\end{aligned}$$

Since this holds for *any* $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{x}$ must be a global minimizer of $\mathbf{y} \mapsto \|\mathbf{A}\mathbf{y}-\mathbf{b}\|$! □

EXAMPLE 3.1.2.4 (Normal equations for some examples from Section 3.0.1) Given $\mathbf{A}$ and $\mathbf{b}$ it takes only elementary linear algebra operations to form the normal equations

$$\mathbf{A}^\top\mathbf{A}\mathbf{x} = \mathbf{A}^\top\mathbf{b}\,. \tag{3.1.2.2}$$

- For Ex. 3.0.1.1 (1D linear regression), $\mathbf{A} \in \mathbb{R}^{m,2}$ given in

$$\begin{bmatrix} x_1 & 1\\ x_2 & 1\\ \vdots & \vdots\\ & \\ \vdots & \vdots\\ x_m & 1 \end{bmatrix}\begin{bmatrix}\alpha\\ \beta\end{bmatrix} = \begin{bmatrix} y_1\\ y_2\\ \vdots\\ \\ \vdots\\ y_m\end{bmatrix} \quad \leftrightarrow \quad \mathbf{A}\mathbf{x} = \mathbf{b}\,,\quad \mathbf{A}\in\mathbb{R}^{m,2}\,,\mathbf{b}\in\mathbb{R}^m\,,\mathbf{x}\in\mathbb{R}^2\,, \tag{3.0.1.3}$$

  we obtain the normal equations linear system

$$\begin{bmatrix} x_1 & x_2 & \dots & \dots & x_m\\ 1 & 1 & \dots & \dots & 1\end{bmatrix}\begin{bmatrix} x_1 & 1\\ x_2 & 1\\ \vdots & \vdots\\ & \\ \vdots & \vdots\\ x_m & 1\end{bmatrix}\begin{bmatrix}\alpha\\ \beta\end{bmatrix} = \begin{bmatrix}\|\mathbf{x}\|_2^2 & \mathbf{1}^\top\mathbf{x}\\ \mathbf{1}^\top\mathbf{x} & m\end{bmatrix}\begin{bmatrix}\alpha\\ \beta\end{bmatrix} = \begin{bmatrix}\mathbf{x}^\top\mathbf{y}\\ \mathbf{1}^\top\mathbf{y}\end{bmatrix}\,,$$

  with $\mathbf{1} = [1,\dots,1]^\top$.

- In the case of Ex. 3.0.1.4 (multi-dimensional linear regression) and the overdetermined $m \times (n+1)$ linear system

$$\begin{bmatrix} \mathbf{x}_1^\top & 1 \\ \vdots & \vdots \\ \mathbf{x}_m^\top & 1 \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ \beta \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} \quad \leftrightarrow \quad \mathbf{A}\mathbf{x} = \mathbf{b}\,, \quad \mathbf{A} \in \mathbb{R}^{m,n+1}\,, \mathbf{b} \in \mathbb{R}^m\,, \mathbf{x} \in \mathbb{R}^{n+1}\,, \tag{3.0.1.5}$$

the normal equations read

$$\begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \dots & \dots & \mathbf{x}_m \\ 1 & 1 & \dots & \dots & 1 \end{bmatrix} \begin{bmatrix} \mathbf{x}_1^\top & 1 \\ \vdots & \vdots \\ \mathbf{x}_m^\top & 1 \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ \beta \end{bmatrix} = \begin{bmatrix} \mathbf{X}\mathbf{X}^\top & \mathbf{X}\mathbf{1} \\ \mathbf{1}^\top\mathbf{X}^\top & m \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ \beta \end{bmatrix} = \begin{bmatrix} \mathbf{X}\mathbf{y} \\ \mathbf{1}^\top\mathbf{y} \end{bmatrix}\,,$$

where $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_m] \in \mathbb{R}^{n,m}$.

⌟

Remark 3.1.2.5 (Normal equations from gradient) We consider the function

$$J : \mathbb{R}^n \to \mathbb{R} \quad , \quad J(\mathbf{y}) := \|\mathbf{A}\mathbf{y} - \mathbf{b}\|_2^2\,. \tag{3.1.2.6}$$

As above, using elementary identities for the Euclidean inner product on $\mathbb{R}^m$, J can be recast as

$$\begin{aligned} J(\mathbf{y}) &= \mathbf{y}^\top\mathbf{A}^\top\mathbf{A}\mathbf{y} - 2\mathbf{b}^\top\mathbf{A}\mathbf{y} + \mathbf{b}^\top\mathbf{b} \\ &= \sum_{i=1}^n \sum_{j=1}^n (\mathbf{A}^\top\mathbf{A})_{ij} y_i y_j - 2\sum_{i=1}^m \sum_{j=1}^n b_i (\mathbf{A})_{ij} y_j + \sum_{i=1}^m b_i^2\,, \qquad \mathbf{y} = [y_1, \dots, y_n]^\top \in \mathbb{R}^n\,. \end{aligned}$$

Obviously, J is a multivariate polynomial in $y_1, \dots, y_n$. As such, J is an infinitely differentiable function, $J \in C^\infty(\mathbb{R}^n, \mathbb{R})$, see [Str09, Bsp. 7.1.5]. The gradient of J vanishes where J attains extremal values [Str09, Satz 7.5.3]. Thus, $\mathbf{x} \in \mathbb{R}^n$ is a minimizer of J only if

$$\mathbf{grad}\, J(\mathbf{x}) = 2\mathbf{A}^\top\mathbf{A}\mathbf{x} - 2\mathbf{A}^\top\mathbf{b} = \mathbf{0}\,. \tag{3.1.2.7}$$

This formula for the gradient of J can easily be confirmed by computing the partial derivatives $\frac{\partial J}{\partial y_i}$ from the above explicit formula. Observe that (3.1.2.7) is equivalent to the normal equations (3.1.2.2). ⌟

§3.1.2.8 (The linear least squares problem (→ § 1.5.5.1)) Thm. 3.1.2.1 together with Thm. 3.1.1.9 already confirms that the normal equations will always have a solution and that $\mathrm{lsq}(\mathbf{A}, \mathbf{b})$ is a subspace of $\mathbb{R}^n$ parallel to $\mathcal{N}(\mathbf{A}^\top\mathbf{A})$. The next theorem gives even more detailed information.

Theorem 3.1.2.9. Kernel and range of $\mathbf{A}^\top\mathbf{A}$

For $\mathbf{A} \in \mathbb{R}^{m,n}$, $m \geq n$, holds

$$\mathcal{N}(\mathbf{A}^\top\mathbf{A}) = \mathcal{N}(\mathbf{A})\,, \tag{3.1.2.10}$$
$$\mathcal{R}(\mathbf{A}^\top\mathbf{A}) = \mathcal{R}(\mathbf{A}^\top)\,. \tag{3.1.2.11}$$

For the proof we need an basic result from linear algebra:

Lemma 3.1.2.12. Kernel and range of (Hermitian) transposed matrices

For any matrix $\mathbf{A} \in \mathbb{K}^{m,n}$ holds

$$\mathcal{N}(\mathbf{A}) = \mathcal{R}(\mathbf{A}^{\mathrm{H}})^\perp \quad , \quad \mathcal{N}(\mathbf{A})^\perp = \mathcal{R}(\mathbf{A}^{\mathrm{H}})\,.$$

✎ Notation: Orthogonal complement of a subspace $V \subset \mathbb{K}^k$:

$$V^\perp := \{\mathbf{x} \in \mathbb{K}^k : \mathbf{x}^H \mathbf{y} = 0 \ \forall \mathbf{y} \in V\} .$$

Proof. (of Thm. 3.1.2.9)
❶: We first show (3.1.2.10)

$$\mathbf{z} \in \mathcal{N}(\mathbf{A}^\top \mathbf{A}) \Leftrightarrow \mathbf{A}^\top \mathbf{A}\mathbf{z} = \mathbf{0} \Rightarrow \mathbf{z}^\top \mathbf{A}^\top \mathbf{A}\mathbf{z} = \|\mathbf{A}\mathbf{z}\|_2^2 = 0 \Leftrightarrow \mathbf{A}\mathbf{z} = \mathbf{0} ,$$
$$\mathbf{A}\mathbf{z} = \mathbf{0} \Rightarrow \mathbf{A}^\top \mathbf{A}\mathbf{z} = \mathbf{0} \Leftrightarrow \mathbf{z} \in \mathcal{N}(\mathbf{A}^\top \mathbf{A}) .$$

❷: The relationship (3.1.2.11) follows from (3.1.2.10) and Lemma 3.1.2.12:

$$\mathcal{R}(\mathbf{A}^\top) \overset{\text{Lemma 3.1.2.12}}{=} \mathcal{N}(\mathbf{A})^\perp \overset{(3.1.2.10)}{=} \mathcal{N}(\mathbf{A}^\top \mathbf{A})^\perp \overset{\text{Lemma 3.1.2.12}}{=} \mathcal{R}(\mathbf{A}^\top \mathbf{A}) .$$

□

Corollary 3.1.2.13. Uniqueness of least squares solutions

If $m \geq n$ and $\mathcal{N}(\mathbf{A}) = \{\mathbf{0}\}$, then the linear system of equations $\mathbf{A}\mathbf{x} = \mathbf{b}$, $\mathbf{A} \in \mathbb{R}^{m,n}$, $\mathbf{b} \in \mathbb{R}^m$, has a unique least squares solution (→ 3.1.1.1)

$$\mathbf{x} = (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{b} , \tag{3.1.2.14}$$

that can be obtained by solving the normal equations (3.1.2.2).

Note that $\mathbf{A}^\top \mathbf{A}$ is symmetric positive definite (→ Def. 1.1.2.6), if $\mathcal{N}(\mathbf{A}) = \{\mathbf{0}\}$.

Remark 3.1.2.15 (Full-rank condition (→ Def. 2.2.1.3)) For a matrix $\mathbf{A} \in \mathbb{R}^{m,n}$ with $m \geq n$ is equivalent

$$\mathcal{N}(\mathbf{A}) = \{\mathbf{0}\} \iff \operatorname{rank}(\mathbf{A}) = n . \tag{3.1.2.16}$$

Hence the assumption $\mathcal{N}(\mathbf{A}) = \{\mathbf{0}\}$ of Cor. 3.1.2.13 is also called a full-rank condition (FRC), because the rank of $\mathbf{A}$ is maximal. ⌟

EXAMPLE 3.1.2.17 (Meaning of full-rank condition for linear models) We revisit the parameter estimation problem for a linear model.

- For Ex. 3.0.1.1, $\mathbf{A} \in \mathbb{R}^{m,2}$ given in (3.0.1.3) it is easy to see

$$\operatorname{rank}\left(\begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ & \\ \vdots & \vdots \\ x_m & 1 \end{bmatrix}\right) = 2 \quad \Leftrightarrow \quad \exists i, j \in \{1, \ldots, m\} : x_i \neq x_j ,$$

that is, the manifest condition, that the all points (x_i, y_i) have the same x-coordinate.

1D linear regression fails, in case all data points lie on a vertical line in the $x-y$-plane. ▷

It goes without saying that no meaningful regression line can be found in this case.

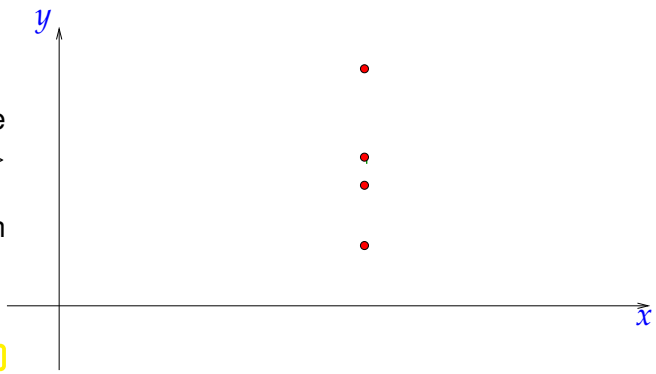


Fig. 76

- In the case of Ex. 3.0.1.4 and the overdetermined $m \times (n+1)$ linear system (3.0.1.5), we find

$$\operatorname{rank}\left(\begin{bmatrix} \mathbf{x}_1^\top & 1 \\ \vdots & \vdots \\ \mathbf{x}_m^\top & 1 \end{bmatrix}\right) = n+1 \quad \Leftrightarrow \quad \begin{array}{l}\text{There is a subset of } n+1 \text{ points} \\ \mathbf{x}_{i_1}, \dots \mathbf{x}_{i_{n+1}} \text{ such that } \{\mathbf{x}_{i_1}, \dots \mathbf{x}_{i_{n+1}}\} \\ \text{spans a non-degenerate } n\text{-simplex.}\end{array}$$

⌟

Remark 3.1.2.18 (Rank defect in linear least squares problems) In case the system matrix $\mathbf{A} \in \mathbb{R}^{m,n}$, $m \geq n$, of an overdetermined linear system arising from a mathematical model fails to have full rank, it hints at inadequate modelling:

In this case parameters are redundant, because different sets of parameters yield the same output quantities: the parameters are not "observable"/"identifiable". ⌟

Remark 3.1.2.19 (Hesse matrix of least squares functional) For the least squares functional

$$J : \mathbb{R}^n \to \mathbb{R} \quad , \quad J(\boldsymbol{y}) := \|\mathbf{A}\mathbf{y} - \mathbf{b}\|_2^2 \, . \tag{3.1.2.6}$$

and its explicit form as polynomial in the vector components y_j we find the Hessian ($\to$ Def. 8.5.1.18, [Str09, Satz 7.5.3]) of J:

$$\mathrm{H}J(\mathbf{y}) = \left[\frac{\partial^2 J}{\partial y_i \partial y_j}(\mathbf{y})\right]_{i,k=1}^{n} = 2\mathbf{A}^\top\mathbf{A} \, . \tag{3.1.2.20}$$

Thm. 3.1.2.9 implies that $\mathbf{A}^\top\mathbf{A}$ is positive definite ($\to$ Def. 1.1.2.6) if and only if $\mathcal{N}(\mathbf{A}) = \{\mathbf{0}\}$.

Therefore, by [Str09, Satz 7.5.3], under the full-rank condition J has a positive definite Hessian everywhere, and a minimum at every stationary point of its gradient, that is, at every solution of the normal equations. ⌟

Remark 3.1.2.21 (Convex least squares functional) Another result from analysis tells us that real-valued C^1-functions on $\mathbb{R}^n$ whose Hessian has positive eigenvalues uniformly bounded away from zero are strictly convex. Hence, if $\mathbf{A}$ has full rank, the least squares functional J from (3.1.2.6) is a strictly convex function.

Visualization of a least squares functional $J : \mathbb{R}^2 \to \mathbb{R}$ for $n = 2$ ▷

Under the full-rank condition the graph of J is a paraboloid with $J(\mathbf{y}) \to \infty$ for $\|\mathbf{y}\| \to \infty$

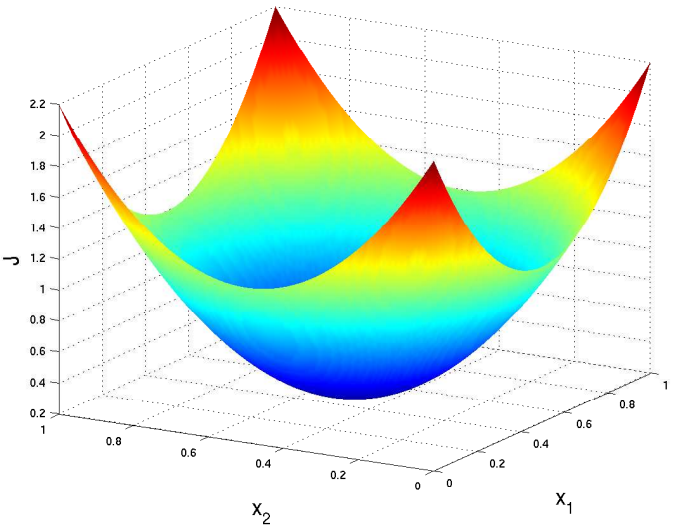


Fig. 77

⌟

Now we are in a position to state precisely what we mean by solving an overdetermined ($m \geq n$!) linear system of equations $\mathbf{Ax} = \mathbf{b}$, $\mathbf{A} \in \mathbb{R}^{m,n}$, $\mathbf{b} \in \mathbb{R}^m$, provided that $\mathbf{A}$ has full (maximal) rank, *cf.* (3.1.2.16).

(Full rank linear) least squares problem: [DR08, Sect. 4.2]

given: $\mathbf{A} \in \mathbb{R}^{m,n}$, $m, n \in \mathbb{N}$, $m \geq n$, $\operatorname{rank}(\mathbf{A}) = n$, $\mathbf{b} \in \mathbb{R}^m$,

find: *unique* $\mathbf{x} \in \mathbb{R}^n$ such that

$$\|\mathbf{Ax} - \mathbf{b}\|_2 = \min\{\|\mathbf{Ay} - \mathbf{b}\|_2 : \mathbf{y} \in \mathbb{R}^n\} \quad (3.1.2.22)$$
$$\Updownarrow$$
$$\mathbf{x} = \operatorname*{argmin}_{\mathbf{y} \in \mathbb{R}^n} \|\mathbf{Ay} - \mathbf{b}\|_2$$

✎ A sloppy notation for the minimization problem (3.1.2.22) is $\|\mathbf{Ax} - \mathbf{b}\|_2 \to \min$ ⌟

Review question(s) 3.1.2.23 (Normal equations)

(Q3.1.2.23.A) Compute the system matrix and the right-hand side vector for the normal equations for 1D linear regression, which led to the overdetermined linear system of equations

$$\begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ & \\ \vdots & \vdots \\ x_m & 1 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \\ \vdots \\ y_m \end{bmatrix},$$

(Q3.1.2.23.B) Let $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\} \subset \mathbb{R}^n$, $k < n$, be a basis of a subspace $V \subset \mathbb{R}^n$. Give a formula for the point $\mathbf{x} \in V$ with smallest Euclidean distance from a given point $\mathbf{p} \in \mathbb{R}^n$. Why is the basis property of $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\} \subset \mathbb{R}^n$ important?

(Q3.1.2.23.C) Characterize the set of least squares solutions $\operatorname{lsq}(\mathbf{A}, \mathbf{b})$, if $\mathbf{A} \in \mathbb{R}^{m,n}$, $m \geq n$, has *orthonormal columns* and $\mathbf{b} \in \mathbb{R}^m$ is an arbitrary vector.

(Q3.1.2.23.D) Let $\mathbf{A} \in \mathbb{R}^{m,n}$, $m \geq n$, have full rank: $\operatorname{rank}(\mathbf{A}) = n$. Show that the mapping

$$\mathsf{P} : \mathbb{R}^m \to \mathbb{R}^m \quad , \quad \mathsf{P}(\mathbf{y}) := \mathbf{A}(\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{y} \, , \quad \mathbf{y} \in \mathbb{R}^m \, ,$$

is an orthogonal projection onto $\mathcal{R}(\mathbf{A})$. This entails proving two properties

(I) $\mathsf{P} \circ \mathsf{P} = \mathsf{P}$ (projection property),

(II) $\mathsf{P}(\mathbf{y}) - \mathbf{y} \perp \mathcal{R}(\mathbf{A})$ for all $\mathbf{y} \in \mathbb{R}^m$.

(Q3.1.2.23.E) [Ridge regression] For a given matrix $\mathbf{A} \in \mathbb{R}^{m,n}$, $m > n$, and vector $\mathbf{b} \in \mathbb{R}^m$, derive a linear system of equations satisfied by any

$$\mathbf{x}^* \in \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \lambda\|\mathbf{x}\|_2^2 ,$$

where $\lambda > 0$ is a fixed constant.

Show that the linear system you have obtained will always possess a unique solution for *any* $\mathbf{A} \in \mathbb{R}^{m,n}$.

(Q3.1.2.23.F) How do the normal equations belonging to an overdetermined linear system of equations $\mathbf{A}\mathbf{x} = \mathbf{b}$, $\mathbf{A} \in \mathbb{R}^{m,n}$, $m > n$, change, when we permute the rows?

Hint. Permuting the rows of an LSE amounts to left-multiplication of the system matrix and the right-hand-side vector with a permutation matrix.

△

3.1.3 Moore-Penrose Pseudoinverse

Video tutorial for Section 3.1.3 "Moore-Penrose Pseudoinverse": (8 minutes) Download link, tablet notes

→ review questions 3.1.3.8

As we have seen in Ex. 3.0.1.9, there can be many least squares solutions of $\mathbf{A}\mathbf{x} = \mathbf{b}$, in case $\mathcal{N}(\mathbf{A}) \neq \{\mathbf{0}\}$. We can impose another condition to single out a unique element of $\mathrm{lsq}(\mathbf{A}, \mathbf{b})$:

Definition 3.1.3.1. Generalized solution of a linear system of equations

The generalized solution $\mathbf{x}^\dagger \in \mathbb{R}^n$ of a linear system of equations $\mathbf{A}\mathbf{x} = \mathbf{b}$, $\mathbf{A} \in \mathbb{R}^{m,n}$, $\mathbf{b} \in \mathbb{R}^m$, is defined as

$$\mathbf{x}^\dagger := \operatorname{argmin}\{\|\mathbf{x}\|_2 : \mathbf{x} \in \mathrm{lsq}(\mathbf{A}, \mathbf{b})\} . \tag{3.1.3.2}$$

➡ The generalized solution is the least squares solution with minimal norm.

§3.1.3.3 (Reduced normal equations) Elementary geometry teaches that the minimal norm element of an affine subspace L (a plane) in Euclidean space is the orthogonal projection of $\mathbf{0}$ onto L.

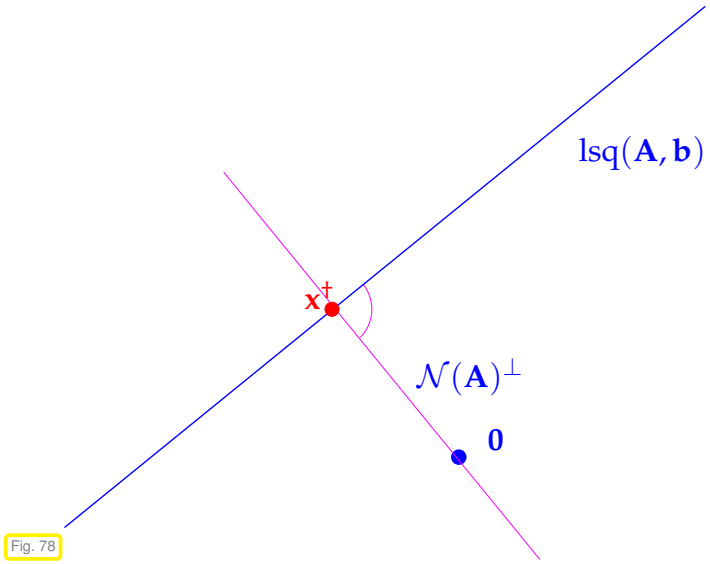


Fig. 78

◁ visualization:

The minimal norm element $\mathbf{x}^\dagger$ of the affine space $\mathrm{lsq}(\mathbf{A},\mathbf{b}) \subset \mathbb{R}^n$ belongs to the subspace of $\mathbb{R}^n$ that is *orthogonal to* $\mathrm{lsq}(\mathbf{A},\mathbf{b})$.

Since the space of least squares solutions of $\mathbf{Ax} = \mathbf{b}$ is an affine subspace parallel to $\mathcal{N}(\mathbf{A})$,

$$\mathrm{lsq}(\mathbf{A},\mathbf{b}) = \mathbf{x}^0 + \mathcal{N}(\mathbf{A}) \;, \quad \mathbf{x}^0 \text{ solves normal equations,} \tag{3.1.3.4}$$

the generalized solution $\mathbf{x}^\dagger$ of $\mathbf{Ax} = \mathbf{b}$ according to Def. 3.1.3.1 is contained in $\mathcal{N}(\mathbf{A})^\perp$. Therefore, given a basis $\{\mathbf{v}_1,\ldots,\mathbf{v}_k\} \subset \mathbb{R}^n$ of $\mathcal{N}(\mathbf{A})^\perp$, $k := \dim \mathcal{N}(\mathbf{A})^\perp = n - \dim \mathcal{N}(\mathbf{A})$, we can find $\mathbf{y} \in \mathbb{R}^k$ such that

$$\mathbf{x}^\dagger = \mathbf{V}\mathbf{y} \quad \text{with} \quad \mathbf{V} := [\mathbf{v}_1,\ldots,\mathbf{v}_k] \in \mathbb{R}^{n,k} \;.$$

Plugging this representation into the normal equations and multiplying with $\mathbf{V}^\top$ yields the reduced normal equations

$$\mathbf{V}^\top\mathbf{A}^\top\mathbf{A}\mathbf{V}\,\mathbf{y} = \mathbf{V}^\top\mathbf{A}^\top\mathbf{b} \tag{3.1.3.5}$$

$$\Updownarrow$$

$$\begin{bmatrix}\mathbf{V}^\top\end{bmatrix}\begin{bmatrix}\mathbf{A}^\top\end{bmatrix}\begin{bmatrix}\mathbf{A}\end{bmatrix}\begin{bmatrix}\mathbf{V}\end{bmatrix}\begin{bmatrix}\mathbf{y}\end{bmatrix} = \begin{bmatrix}\mathbf{V}^\top\end{bmatrix}\begin{bmatrix}\mathbf{A}^\top\end{bmatrix}\begin{bmatrix}\mathbf{b}\end{bmatrix} \;.$$

The very construction of $\mathbf{V}$ ensures $\mathcal{N}(\mathbf{AV}) = \{\mathbf{0}\}$ so that, by Thm. 3.1.2.9 the $k \times k$ linear system of equations (3.1.3.5) has a unique solution. The next theorem summarizes our insights:

Theorem 3.1.3.6. Formula for generalized least-squares solution

Given $\mathbf{A} \in \mathbb{R}^{m,n}$, $\mathbf{b} \in \mathbb{R}^m$, the generalized solution $\mathbf{x}^\dagger$ of the linear system of equations $\mathbf{Ax} = \mathbf{b}$ is given by

$$\mathbf{x}^\dagger = \mathbf{V}(\mathbf{V}^\top\mathbf{A}^\top\mathbf{A}\mathbf{V})^{-1}(\mathbf{V}^\top\mathbf{A}^\top\mathbf{b}) \;,$$

where $\mathbf{V}$ is any *matrix whose columns form a basis of $\mathcal{N}(\mathbf{A})^\perp$.*

Terminology: The matrix

$$\mathbf{A}^{\dagger} := \mathbf{V}(\mathbf{V}^{\top}\mathbf{A}^{\top}\mathbf{A}\mathbf{V})^{-1}\mathbf{V}^{\top}\mathbf{A}^{\top} \in \mathbb{R}^{n,m}$$

is called the Moore-Penrose pseudoinverse of $\mathbf{A}$. If $\mathcal{N}\mathbf{A} = \{\mathbf{0}\}$, then the formula simplifies to $\mathbf{A}^{\dagger} = (\mathbf{A}^{\top}\mathbf{A})^{-1}\mathbf{A}^{\top}$.

✎ notation: $\mathbf{A}^{\dagger} \in \mathbb{R}^{n,m} \triangleq$ pseudoinverse of $\mathbf{A} \in \mathbb{R}^{m,n}$

Note that the Moore-Penrose pseudoinverse does not depend on the choice of $\mathbf{V}$. ⌟

Armed with the concept of generalized solution and the knowledge about its existence and uniqueness we can state the most general linear least squares problem:

(General linear) least squares problem:

given: $\mathbf{A} \in \mathbb{R}^{m,n}$, $m, n \in \mathbb{N}$, $\mathbf{b} \in \mathbb{R}^{m}$,
find: $\mathbf{x}^{\dagger} \in \mathbb{R}^{n}$ such that

$$\text{(i)} \quad \left\|\mathbf{A}\mathbf{x}^{\dagger} - \mathbf{b}\right\|_2 = \min\{\|\mathbf{A}\mathbf{y} - \mathbf{b}\|_2 \colon \mathbf{y} \in \mathbb{R}^{n}\}, \tag{3.1.3.7}$$

$$\text{(ii)} \quad \left\|\mathbf{x}^{\dagger}\right\|_2 \text{ is minimal under the condition (i).}$$

Review question(s) 3.1.3.8 (Moore-Penrose pseudoinverse)

(Q3.1.3.8.A) Let $\mathbf{A} \in \mathbb{R}^{m,n}$ with non-trivial nullspace $\mathcal{N}(\mathbf{A}) \neq \{\mathbf{0}\}$, $k := \dim \mathcal{N}(\mathbf{A})$. The columns of the matrix $\mathbf{V} \in \mathbb{R}^{n,n-k}$ provide a *basis* for the orthogonal complement $\mathcal{N}(\mathbf{A})^{\perp}$. Show that the matrix $\mathbf{V}^{\top}\mathbf{A}^{\top}\mathbf{A}\mathbf{V} \in \mathbb{R}^{n-k,n-k}$ is regular.

(Q3.1.3.8.B) Given $\mathbf{A} \in \mathbb{R}^{m,n}$ and a basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$, $k \leq n$, of the orthogonal complement $\mathcal{N}(\mathbf{A})^{\perp}$, show that the Moore-Penrose pseudoinverse

$$\mathbf{A}^{\dagger} := \mathbf{V}(\mathbf{V}^{\top}\mathbf{A}^{\top}\mathbf{A}\mathbf{V})^{-1}\mathbf{V}^{\top}\mathbf{A}^{\top}, \quad \mathbf{V} := [\mathbf{v}_1, \ldots, \mathbf{v}_k] \in \mathbb{R}^{n,k}$$

does **not** depend on the choice of the basis vectors $\mathbf{v}_{\ell}$.

(Q3.1.3.8.C) [Pseudoinverse of a vector] What is the pseudoinverse of a non-zero column vector $\mathbf{a} \in \mathbb{R}^{m}$, when it is regarded as a $m \times 1$-matrix, $m \in \mathbb{N}$?

What is the pseudoinverse of a non-zero row vector $\mathbf{a}$, $\mathbf{a}^{\top} \in \mathbb{R}^{n}$, when it is regarded as a $1 \times n$-matrix, $n \in \mathbb{N}$?

△

3.1.4 Sensitivity of Least Squares Problems

Consider the full-rank linear least squares problem introduced in (3.1.2.22):

$$\succ \quad \text{data} \quad (\mathbf{A}, \mathbf{b}) \in \mathbb{R}^{m,n} \times \mathbb{R}^{m} \quad , \quad \text{result } \mathbf{x} = \operatorname{argmin}_{\mathbf{y} \in \mathbb{R}^{n}} \|\mathbf{A}\mathbf{y} - \mathbf{b}\|_2 \in \mathbb{R}^{n}$$

On data space and result space we use the Euclidean norm (2-norm $\|\cdot\|_2$) and the associated matrix norm, see § 1.5.5.3.

Recall Section 2.2.2, where we discussed the sensitivity of solutions of square linear systems, that is, the impact of perturbations in the problem data on the result. Now we study how (small) changes in $\mathbf{A}$

and $\mathbf{b}$ affect the unique ($\rightarrow$ Cor. 3.1.2.13) least solution $\mathbf{x}$ of $\mathbf{A}\mathbf{x} = \mathbf{b}$ in the case of $\mathbf{A}$ with full rank ($\Leftrightarrow$ $\mathcal{N}(\mathbf{A}) = \{\mathbf{0}\}$)

Note: If the matrix $\mathbf{A} \in \mathbb{R}^{m,n}$, $m \geq n$, has full rank, then there is a $c > 0$ such that $\mathbf{A} + \Delta\mathbf{A}$ still has full rank for all $\Delta\mathbf{A} \in \mathbb{R}^{m,n}$ with $\|\Delta\mathbf{A}\|_2 < c$. Hence, "sufficiently small" perturbations will not destroy the full-rank property of $\mathbf{A}$. This is a generalization of the Perturbation Lemma 2.2.2.5.

For square linear systems the condition number of the system matrix ($\rightarrow$ Def. 2.2.2.7) provided the key gauge of sensitivity. To express the sensitivity of linear least squares problems we also generalize this concept:

Definition 3.1.4.1. Generalized condition number of a matrix

Given $\mathbf{A} \in \mathbb{K}^{m,n}$, $m \geq n$, $\operatorname{rank}(\mathbf{A}) = n$, we define its generalized (Euclidean) condition number as

$$\operatorname{cond}_2(\mathbf{A}) := \sqrt{\frac{\lambda_{\max}(\mathbf{A}^{\mathrm{H}}\mathbf{A})}{\lambda_{\min}(\mathbf{A}^{\mathrm{H}}\mathbf{A})}}\,.$$

✎ notation: $\lambda_{\min}(\mathbf{A}) \mathrel{\hat{=}}$ smallest (in modulus) eigenvalue of matrix $\mathbf{A}$
$\lambda_{\max}(\mathbf{A}) \mathrel{\hat{=}}$ largest (in modulus) eigenvalue of matrix $\mathbf{A}$

For a square regular matrix this agrees with its condition number according to Def. 2.2.2.7, which follows from Cor. 1.5.5.16.

Theorem 3.1.4.2. Sensitivity of full-rank linear least squares problem

For $m \geq n$, $\mathbf{A} \in \mathbb{R}^{m,n}$, $\operatorname{rank}(\mathbf{A}) = n$, let $\mathbf{x} \in \mathbb{R}^n$ be the solution of the least squares problem $\|\mathbf{A}\mathbf{x} - \mathbf{b}\| \to \min$ and $\widehat{\mathbf{x}}$ the solution of the perturbed least squares problem $\|(\mathbf{A} + \Delta\mathbf{A})\widehat{\mathbf{x}} - \mathbf{b}\| \to \min$. Then

$$\frac{\|\mathbf{x} - \widehat{\mathbf{x}}\|_2}{\|\mathbf{x}\|_2} \lesssim \left(2\operatorname{cond}_2(\mathbf{A}) + \operatorname{cond}_2^2(\mathbf{A})\frac{\|\mathbf{r}\|_2}{\|\mathbf{A}\|_2\|\mathbf{x}\|_2}\right)\frac{\|\Delta\mathbf{A}\|_2}{\|\mathbf{A}\|_2}$$

holds, where $\mathbf{r} = \mathbf{A}\mathbf{x} - \mathbf{b}$ is the residual.

This means: if $\|\mathbf{r}\|_2 \ll 1$ ➤ condition of the least squares problem $\approx \operatorname{cond}_2(\mathbf{A})$
if $\|\mathbf{r}\|_2$ "large" ➤ condition of the least squares problem $\approx \operatorname{cond}_2^2(\mathbf{A})$

For instance, in a linear parameter estimation problem ($\rightarrow$ Ex. 3.0.1.4) a small residual will be the consequence of small measurement errors.

Review question(s) 3.1.4.3 (Sensitivity of least-squares problems)

(Q3.1.4.3.A) Interpret the assertion of the theorem

Theorem 3.1.4.2. Sensitivity of full-rank linear least squares problem

For $m \geq n$, $\mathbf{A} \in \mathbb{R}^{m,n}$, $\operatorname{rank}(\mathbf{A}) = n$, let $\mathbf{x} \in \mathbb{R}^n$ be the solution of the least squares problem $\|\mathbf{A}\mathbf{x} - \mathbf{b}\| \to \min$ and $\widehat{\mathbf{x}}$ the solution of the perturbed least squares problem $\|(\mathbf{A} + \Delta\mathbf{A})\widehat{\mathbf{x}} - \mathbf{b}\| \to \min$. Then

$$\frac{\|\mathbf{x} - \widehat{\mathbf{x}}\|_2}{\|\mathbf{x}\|_2} \lesssim \left(2\operatorname{cond}_2(\mathbf{A}) + \operatorname{cond}_2^2(\mathbf{A})\frac{\|\mathbf{r}\|_2}{\|\mathbf{A}\|_2\|\mathbf{x}\|_2}\right)\frac{\|\Delta\mathbf{A}\|_2}{\|\mathbf{A}\|_2}$$

holds, where $\mathbf{r} = \mathbf{A}\mathbf{x} - \mathbf{b}$ is the residual.

for the case $n = 1$.

Hint. Recall the definition

Definition 3.1.4.1. Generalized condition number of a matrix

Given $\mathbf{A} \in \mathbb{K}^{m,n}$, $m \geq n$, $\operatorname{rank}(\mathbf{A}) = n$, we define its generalized (Euclidean) condition number as

$$\operatorname{cond}_2(\mathbf{A}) := \sqrt{\frac{\lambda_{\max}(\mathbf{A}^H\mathbf{A})}{\lambda_{\min}(\mathbf{A}^H\mathbf{A})}}\,.$$

△

3.2 Normal Equation Methods [DR08, Sect. 4.2], [Han02, Ch. 11]

Video tutorial for Section 3.2 "Normal Equation Methods": (12 minutes) Download link, tablet notes

→ review questions 3.2.0.11

Given $\mathbf{A} \in \mathbb{R}^{m,n}$, $m \geq n$, $\operatorname{rank}(\mathbf{A}) = n$, $\mathbf{b} \in \mathbb{R}^m$, we introduce a first practical numerical method to determine the unique least squares solution (→ Def. 3.1.1.1) of the overdetermined linear system of equations $\mathbf{A}\mathbf{x} = \mathbf{b}$.

In fact, Cor. 3.1.2.13 suggests a simple algorithm for solving linear least squares problems of the form (3.1.2.22) satisfying the full (maximal) rank condition $\operatorname{rank}(\mathbf{A}) = n$: it boils down to solving the normal equations (3.1.2.2):

Algorithm: Normal equation method to solve full-rank least squares problem $\mathbf{Ax} = \mathbf{b}$

❶ Compute regular matrix $\mathbf{C} := \mathbf{A}^\top \mathbf{A} \in \mathbb{R}^{n,n}$.
❷ Compute right hand side vector $\mathbf{c} := \mathbf{A}^\top \mathbf{b}$.
❸ Solve s.p.d. (→ Def. 1.1.2.6) linear system of equations: $\mathbf{Cx} = \mathbf{c}$ → § 2.8.0.14

Definition 1.1.2.6. Symmetric positive definite (s.p.d.) matrices → [DR08, Def. 3.31], [QSS00, Def. 1.22]

$\mathbf{M} \in \mathbb{K}^{n,n}$, $n \in \mathbb{N}$, is symmetric (Hermitian) positive definite (s.p.d.), if

$$\mathbf{M} = \mathbf{M}^{\mathrm{H}} \quad \text{and} \quad \forall \mathbf{x} \in \mathbb{K}^n: \ \mathbf{x}^{\mathrm{H}}\mathbf{M}\mathbf{x} > 0 \ \Leftrightarrow \ \mathbf{x} \neq 0 \, .$$

If $\mathbf{x}^{\mathrm{H}}\mathbf{M}\mathbf{x} \geq 0$ for all $\mathbf{x} \in \mathbb{K}^n$ ▷ $\mathbf{M}$ positive semi-definite.

The s.p.d. property of $\mathbf{C}$ is an immediate consequence of the equivalence $\operatorname{rank}(\mathbf{A}) = n \Leftrightarrow \mathcal{N}(\mathbf{A}) = \{0\}$:

$$\mathbf{x}^\top \mathbf{C}\mathbf{x} = \mathbf{x}^\top \mathbf{A}^\top \mathbf{A}\mathbf{x} = (\mathbf{A}\mathbf{x})^\top (\mathbf{A}\mathbf{x}) = \|\mathbf{A}\mathbf{x}\|_2^2 > 0 \ \Leftrightarrow \ \mathbf{x} \neq \mathbf{0} \, .$$

The above algorithm can be realized in EIGEN in a single line of code:

C++ code 3.2.0.1: Solving a linear least squares problem via normal equations ➙ GITLAB

```
2  //!  Solving the overdetermined linear system of equations
3  //!  Ax = b by solving normal equations (3.1.2.2)
4  //!  The least squares solution is returned by value
5  VectorXd normeqsolve(const MatrixXd &A,const VectorXd &b) {
6    if (b.size() != A.rows()) {
7      throw runtime_error("Dimension mismatch");
8    }
9    // Use Cholesky factorization for s.p.d. system matrix, § 2.8.0.14
10   VectorXd x = (A.transpose()*A).llt().solve(A.transpose()*b);
11   return x;
12 }
```

By Thm. 2.8.0.11, for the s.p.d. matrix $\mathbf{A}^\top \mathbf{A}$ Gaussian elimination remains stable even without pivoting. This is taken into account by requesting the Cholesky decomposition of $\mathbf{A}^\top \mathbf{A}$ by calling the method `llt()`.

§3.2.0.2 (Asymptotic complexity of normal equation method) The problem size parameters for the linear least squares problem (3.1.2.22) are the matrix dimensions $m, n \in \mathbb{N}$, where n small & fixed, $n \ll m$, is common.

In Section 1.4.2 and Thm. 2.5.0.2 we discussed the asymptotic complexity of the operations involved in step ❶-❸ of the normal equation method:

$$\left.\begin{array}{ll} \text{step ❶:} & \text{cost } O(mn^2) \\ \text{step ❷:} & \text{cost } O(nm) \\ \text{step ❸:} & \text{cost } O(n^3) \end{array}\right\} \ \blacktriangleright \ \text{cost } O(n^2 m + n^3) \quad \text{for } m, n \to \infty \, .$$

Note that for small fixed n, $n \ll m$, $m \to \infty$ the computational effort scales linearly with m. ⌟

Remark 3.2.0.3 (Conditioning of normal equations [DR08, pp. 128])

The solution of least squares problems via the normal equation method is *vulnerable to instability*; immediate from Def. 3.1.4.1:

$$\operatorname{cond}_2(\mathbf{A}^{\mathrm{H}}\mathbf{A}) = \operatorname{cond}_2(\mathbf{A})^2 .$$

Recall from Thm. 2.2.2.4: $\operatorname{cond}_2(\mathbf{A}^{\mathrm{H}}\mathbf{A})$ governs amplification of (roundoff) errors in $\mathbf{A}^{\top}\mathbf{A}$ and $\mathbf{A}^{\top}\mathbf{b}$ when solving normal equations (3.1.2.2).

➢ For fairly ill-conditioned $\mathbf{A}$ using the normal equations (3.1.2.2) to solve the linear least squares problem from Def. 3.1.1.1 numerically may run the risk of huge amplification of roundoff errors incurred during the computation of the right hand side $\mathbf{A}^{\mathrm{H}}\mathbf{b}$: potential instability (→ Def. 1.5.5.19) of normal equation approach.

⌟

EXAMPLE 3.2.0.4 (Roundoff effects in normal equations → [DR08, Ex. 4.12]) In this example we witness loss of information in the computation of $\mathbf{A}^{\mathrm{H}}\mathbf{A}$.

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ \delta & 0 \\ 0 & \delta \end{bmatrix} \quad \Rightarrow \quad \mathbf{A}^{\top}\mathbf{A} = \begin{bmatrix} 1+\delta^2 & 1 \\ 1 & 1+\delta^2 \end{bmatrix}$$

Exp. 1.5.3.14: If $\delta \approx \sqrt{\mathrm{EPS}}$, then $1+\delta^2 = 1$ in $\mathbb{M}$ (set of machine numbers, see Def. 1.5.2.4). Hence the computed $\mathbf{A}^{\top}\mathbf{A}$ will fail to be regular, though $\operatorname{rank}(\mathbf{A}) = 2$, $\operatorname{cond}_2(\mathbf{A}) \approx \sqrt{\mathrm{EPS}}$.

C++-code 3.2.0.5: Computation of numerical rank of a matrix ➽ GITLAB

```
2  int main() {
3    MatrixXd A(3,2);
4    // Inquire about machine precision → Ex. 1.5.3.12
5    const double eps = std::numeric_limits<double>::epsilon();
6    // « initialization of matrix → § 1.2.1.3
7    A << 1, 1, sqrt(eps), 0, 0, sqrt(eps);
8    // Output rank of A^TA
9    std::cout << "Rank of A: " << A.fullPivLu().rank() << std::endl
10             << "Rank of A^TA: "
11             << (A.transpose() * A).fullPivLu().rank() << std::endl;
12   return 0;
13 }
```

Output:

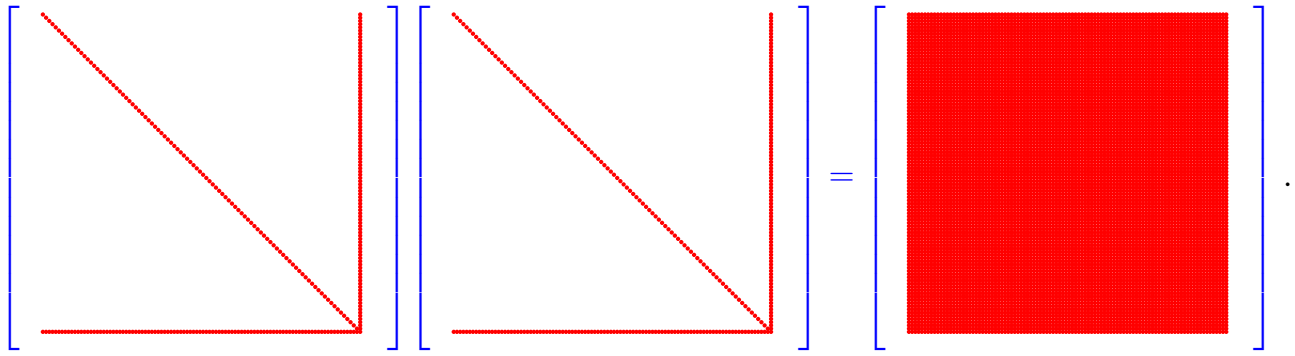
```
1 Rank of A: 2
2 Rank of A^T*A: 1
```

⌟

Remark 3.2.0.6 (Loss of sparsity when forming normal equations) Another *reason not to compute* $\mathbf{A}^{\mathrm{H}}\mathbf{A}$, when both m, n large:

$$\mathbf{A} \text{ sparse} \quad \not\Rightarrow \quad \mathbf{A}^{\top}\mathbf{A} \text{ sparse}$$

Example from Rem. 1.3.1.5: "Arrow matrices"



Consequences for normal equation method, if both m, n large:

- Potential memory overflow, when computing $\mathbf{A}^\top\mathbf{A}$
- Squanders possibility to use efficient sparse direct elimination techniques, see Section 2.7.4

This situation is faced in Ex. 3.0.1.9, Ex. 3.0.1.8. ⌟

§3.2.0.7 (Extended normal equations)

There is a way to avoid the computation of the system matrix $\mathbf{A}^\top\mathbf{A} \in \mathbb{R}^{n,n}$ of the normal equation for the overdetermined linear system of equations $\mathbf{Ax} = \mathbf{b}$, $\mathbf{A} \in \mathbb{R}^{m,n}$, $m \geq n$. The trick is to extend the normal equations,

$$\mathbf{A}^\top\mathbf{Ax} = \mathbf{A}^\top\mathbf{b}\,, \tag{3.1.2.2}$$

by introducing the introduce residual $\mathbf{r} := \mathbf{Ax} - \mathbf{b} \in \mathbb{R}^m$ as new (auxiliary) unknown:

$$\mathbf{A}^\top\mathbf{Ax} = \mathbf{A}^\top\mathbf{b} \quad \Leftrightarrow \quad \mathbf{B}\begin{bmatrix}\mathbf{r}\\ \mathbf{x}\end{bmatrix} := \begin{bmatrix}-\mathbf{I}_m & \mathbf{A}\\ \mathbf{A}^\top & \mathbf{O}\end{bmatrix}\begin{bmatrix}\mathbf{r}\\ \mathbf{x}\end{bmatrix} = \begin{bmatrix}\mathbf{b}\\ \mathbf{0}\end{bmatrix}. \tag{3.2.0.8}$$

The benefit of using (3.2.0.8) instead of the standard normal equations (3.1.2.2) is that *sparsity is preserved*. However, the conditioning of the system matrix in (3.2.0.8) is not better than that of $\mathbf{A}^\top\mathbf{A}$.

A more general substitution $\mathbf{q} := \alpha^{-1}(\mathbf{Ax} - \mathbf{b})$ with $\alpha > 0$ may even improve the conditioning for suitably chosen parameter $\alpha > 0$:

$$\mathbf{A}^\top\mathbf{Ax} = \mathbf{A}^\top\mathbf{b} \quad \Leftrightarrow \quad \mathbf{B}_\alpha\begin{bmatrix}\mathbf{q}\\ \mathbf{x}\end{bmatrix} := \begin{bmatrix}-\alpha\mathbf{I} & \mathbf{A}\\ \mathbf{A}^\top & 0\end{bmatrix}\begin{bmatrix}\mathbf{q}\\ \mathbf{x}\end{bmatrix} = \begin{bmatrix}\mathbf{b}\\ 0\end{bmatrix}. \tag{3.2.0.9}$$

For $m, n \gg 1$, $\mathbf{A}$ sparse, both (3.2.0.8) and (3.2.0.9) lead to large sparse linear systems of equations, amenable to sparse direct elimination techniques, see Section 2.7.4. ⌟

EXAMPLE 3.2.0.10 (Conditioning of the extended normal equations)

In this example we explore empirically how the Euclidean condition number of the extended normal equations (3.2.0.9) is influenced by the coice of α
Consider (3.2.0.8), (3.2.0.9) for

$$\mathbf{A} = \begin{bmatrix} 1+\epsilon & 1 \\ 1-\epsilon & 1 \\ \epsilon & \epsilon \end{bmatrix} .$$

Plot of different condition numbers
in dependence on ϵ
(Here $\alpha = \epsilon\|\mathbf{A}\|_2/\sqrt{2}$)

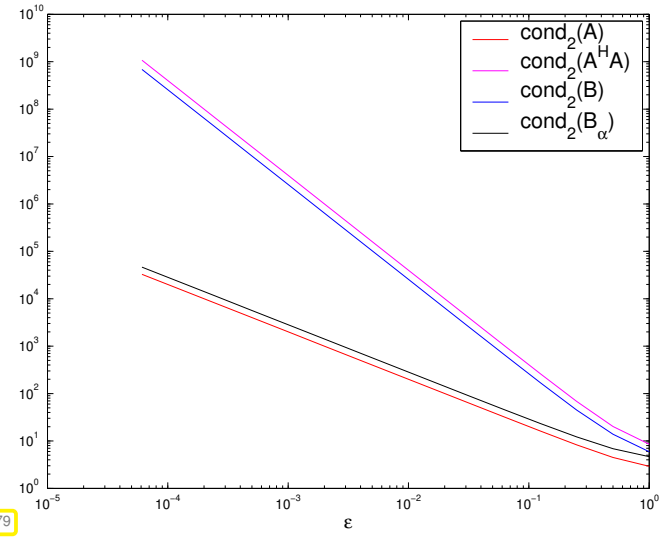


Fig. 79

Review question(s) 3.2.0.11 (Normal equation methods)

(Q3.2.0.11.A) We consider the overdetermined linear system of equations

$$\mathbf{A}\mathbf{x} = \mathbf{b} , \quad \mathbf{A} \in \mathbb{R}^{m,n}, \; m \geq n, \; \mathbf{b} \in \mathbb{R}^m . \tag{3.2.0.12}$$

We augment it by another equation and get another overdetermined linear system of equations

$$\begin{bmatrix} \mathbf{A} \\ \mathbf{v}^\top \end{bmatrix} \widetilde{\mathbf{x}} = \begin{bmatrix} \mathbf{b} \\ \beta \end{bmatrix} , \quad \mathbf{v} \in \mathbb{R}^n , \quad \beta \in \mathbb{R} . \tag{3.2.0.13}$$

How are the normal equations of (3.2.0.12) and (3.2.0.13) related?

(Q3.2.0.11.B) Discuss how the coefficient matrices and right-hand side vectors of the normal equations belonging to the two overdetermined linear system of equations

$$\mathbf{A}\mathbf{x} = \mathbf{b} , \quad \mathbf{A} \in \mathbb{R}^{m,n}, \; m \geq n, \; \mathbf{b} \in \mathbb{R}^m , \tag{3.2.0.14}$$

$$(\mathbf{A} + \mathbf{u}\mathbf{v}^\top)\widetilde{\mathbf{x}} = \mathbf{b} , \quad \mathbf{u} \in \mathbb{R}^m , \quad \mathbf{v} \in \mathbb{R}^n , \tag{3.2.0.15}$$

are related.

△

3.3 Orthogonal Transformation Methods [DR08, Sect. 4.4.2]

Video tutorial for Section 3.3 "Orthogonal Transformation Methods": (10 minutes)
Download link, tablet notes

→ review questions 3.3.2.3

3.3.1 Transformation Idea

We consider the full-rank linear least squares problem (3.1.2.22)

$$\text{given} \quad \mathbf{A} \in \mathbb{R}^{m,n} , \mathbf{b} \in \mathbb{R}^m \quad \text{find} \quad \mathbf{x} = \operatorname*{argmin}_{\mathbf{y}\in\mathbb{R}^n} \|\mathbf{A}\mathbf{y} - \mathbf{b}\|_2 . \tag{3.1.2.22}$$

Setting: $m \geq n$ and $\mathbf{A}$ has full (maximum) rank: $\operatorname{rank}(\mathbf{A}) = n$.

§3.3.1.1 (Generalizing the policy underlying Gaussian elimination) Recall the rationale behind Gaussian elimination (→ Section 2.3, Ex. 2.3.1.1)

➥ By row transformations convert LSE $\mathbf{Ax} = \mathbf{b}$ to *equivalent* (in terms of set of solutions) LSE $\mathbf{Ux} = \widetilde{\mathbf{b}}$, which is easier to solve because it has triangular form.

How to adapt this policy to linear least squares problem (3.1.2.22) **?**

Two questions: ❶ What linear least squares problems are "easy to solve" ?

❷ How can we arrive at them by *equivalent transformations* of (3.1.2.22) ?

Here we call two overdetermined linear systems $\mathbf{Ax} = \mathbf{b}$ and $\widetilde{\mathbf{A}}\mathbf{x} = \widetilde{\mathbf{b}}$ equivalent in the sense of (3.1.2.22), if both have the same set of least squares solutions: $\mathrm{lsq}(\mathbf{A},\mathbf{b}) = \mathrm{lsq}(\widetilde{\mathbf{A}},\widetilde{\mathbf{b}})$, see (3.1.1.2). ⌟

§3.3.1.2 (Triangular linear least squares problems)

The answer to question ❶ is the same as for LSE/Gaussian elimination:

Linear least squares problems (3.1.2.22) with upper *triangular* $\mathbf{A}$ are easy to solve!

$$\left\| \begin{bmatrix} & \\ & \mathbf{A} \\ & \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} - \begin{bmatrix} b_1 \\ \vdots \\ \\ \\ \vdots \\ b_m \end{bmatrix} \right\|_2 \to \min \quad \overset{(*)}{\Longrightarrow} \quad \mathbf{x} = \begin{bmatrix} & \\ & \mathbf{R} \\ & \end{bmatrix}^{-1} \begin{bmatrix} b_1 \\ \vdots \\ \vdots \\ \vdots \\ b_n \end{bmatrix}$$

$\mathbf{x} \hat{=}$ least squares solution

How can we draw the conclusion $(*)$? Obviously, the components $n+1,\ldots,m$ of the vector inside the norm are fixed and do not depend on $\mathbf{x}$. All we can do is to make the first components $1,\ldots,n$ vanish, by choosing a suitable $\mathbf{x}$, see [DR08, Thm. 4.13]. Obviously, $\mathbf{x} = \mathbf{R}^{-1}(\mathbf{b})_{1:n}$ accomplishes this.

Note: since $\mathbf{A}$ has full rank n, the upper triangular part $\mathbf{R} \in \mathbb{R}^{n,n}$ of $\mathbf{A}$ is regular! ⌟

Answer to question ❷:

Idea: If we have a (transformation) matrix $\mathbf{T} \in \mathbb{R}^{m,m}$ satisfying

$$\|\mathbf{Ty}\|_2 = \|\mathbf{y}\|_2 \quad \forall \mathbf{y} \in \mathbb{R}^m\,, \tag{3.3.1.3}$$

then $\operatorname{argmin}_{\mathbf{y}\in\mathbb{R}^n} \|\mathbf{Ay}-\mathbf{b}\|_2 = \operatorname{argmin}_{\mathbf{y}\in\mathbb{R}^n} \left\|\widetilde{\mathbf{A}}\mathbf{y}-\widetilde{\mathbf{b}}\right\|_2$,

where $\widetilde{\mathbf{A}} = \mathbf{TA}$ and $\widetilde{\mathbf{b}} = \mathbf{Tb}$.

The next section will characterize the class of eligible transformation matrices $\mathbf{T}$.

3.3.2 Orthogonal/Unitary Matrices

Definition 3.3.2.1. Unitary and orthogonal matrices → [Gut09, Sect. 2.8]

- $\mathbf{Q} \in \mathbb{K}^{n,n}$, $n \in \mathbb{N}$, is unitary, if $\mathbf{Q}^{-1} = \mathbf{Q}^H$.
- $\mathbf{Q} \in \mathbb{R}^{n,n}$, $n \in \mathbb{N}$, is orthogonal, if $\mathbf{Q}^{-1} = \mathbf{Q}^T$.

Theorem 3.3.2.2. Preservation of Euclidean norm

A matrix is unitary/orthogonal, if and only if the associated linear mapping preserves the 2-norm:

$$\mathbf{Q} \in \mathbb{K}^{n,n} \quad \textit{unitary} \quad \Leftrightarrow \quad \|\mathbf{Q}\mathbf{x}\|_2 = \|\mathbf{x}\|_2 \quad \forall \mathbf{x} \in \mathbb{K}^n \,.$$

From Thm. 3.3.2.2 we immediately conclude that, if a matrix $\mathbf{Q} \in \mathbb{K}^{n,n}$ is unitary/orthogonal, then

- all rows/columns (regarded as vectors $\in \mathbb{K}^n$) have Euclidean norm $= 1$,
- all rows/columns are pairwise orthogonal (w.r.t. Euclidean inner product),
- $|\det \mathbf{Q}| = 1$, $\|\mathbf{Q}\|_2 = 1$, and all eigenvalues $\in \{z \in \mathbb{C}: |z| = 1\}$.
- $\|\mathbf{Q}\mathbf{A}\|_2 = \|\mathbf{A}\|_2$ for any matrix $\mathbf{A} \in \mathbb{K}^{n,m}$

Review question(s) 3.3.2.3 (Orthogonal transformations)

(Q3.3.2.3.A) [Orthogonal matrices in $\mathbb{R}^2$] Give a full characterization of all orthogonal matrices $\mathbf{Q} \in \mathbb{R}^{2,2}$.

Hint. The subspace of $\mathbb{R}^2$ spanned by all vectors orthogonal (w.r.t the Euclidean inner product) to a given vector $\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$ is spanned by $\begin{bmatrix} -u_2 \\ u_1 \end{bmatrix}$.

(Q3.3.2.3.B) [Polarization identity] Prove the following variant of a polarization identity:

$$\mathbf{x}^\top \mathbf{y} = \frac{1}{4}\left(\|\mathbf{x}+\mathbf{y}\|_2^2 - \|\mathbf{x}-\mathbf{y}\|_2^2\right) \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^2 \,.$$

(Q3.3.2.3.C) Based on the result of Question (Q3.3.2.3.A) find an orthogonal matrix $\mathbf{Q} \in \mathbb{R}^{2,2}$, such that

$$\mathbf{Q}\begin{bmatrix} 0 & a_{12} \\ 1 & a_{22} \end{bmatrix} = \begin{bmatrix} * & * \\ 0 & * \end{bmatrix}, \quad a_{12}, a_{22} \in \mathbb{R} \,.$$

Here $*$ stands for an arbitrary matrix entry.

△

3.3.3 QR-Decomposition [Han02, Sect. 13], [Gut09, Sect. 7.3]

This section will answer the question whether and how it is possible to find orthogonal transformations that convert any given matrix $\mathbf{A} \in \mathbb{R}^{m,n}$, $m \geq n$, $\operatorname{rank}(\mathbf{A}) = n$, to upper triangular form, as required for the application of the "equivalence transformation idea" to full-rank linear least squares problems.

3.3.3.1 QR-Decomposition: Theory

Video tutorial for Section 3.3.3.1 "QR-Decomposition: Theory": (11 minutes) Download link, tablet notes

→ review questions 3.3.3.8

§3.3.3.1 (Gram-Schmidt orthogonalisation recalled → § 1.5.1.1)

Input: $\{\mathbf{a}^1,\dots,\mathbf{a}^n\} \subset \mathbb{K}^m$
Output: $\{\mathbf{q}^1,\dots,\mathbf{q}^n\}$ (assuming no premature termination!)

```
1: q^1 := a^1/‖a^1‖_2;   % 1st output vector
2: for j = 2,...,n do
     { % Orthogonal projection
3:     q^j := a^j;
4:     for ℓ = 1,2,...,j−1 do                    (GS)
5:       { q^j ← q^j − ⟨a^j, q^ℓ⟩ q^ℓ; }
6:     if ( q^j = 0 ) then STOP
7:     else { q^j ← q^j/‖q^j‖_2; }
8:   }
```

Theorem 3.3.3.2. Span property of G.S. vectors

If $\{\mathbf{a}^1,\dots,\mathbf{a}^n\} \subset \mathbb{R}^m$ is linearly independent, then Algorithm (GS) *computes orthonormal vectors $\mathbf{q}^1,\dots,\mathbf{q}^n \in \mathbb{R}^m$ satisfying*

$$\operatorname{Span}\{\mathbf{q}^1,\dots,\mathbf{q}^\ell\} = \operatorname{Span}\{\mathbf{a}^1,\dots,\mathbf{a}^\ell\}\,, \tag{1.5.1.2}$$

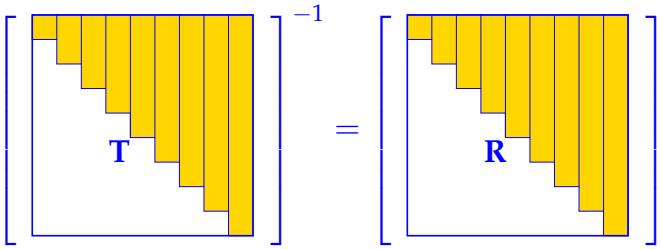
for all $\ell \in \{1,\dots,n\}$.

The span property (1.5.1.2) can be made more explicit in terms of the existence of linear combinations

$$\begin{aligned}
\mathbf{q}^1 &= t_{11}\mathbf{a}^1\\
\mathbf{q}^2 &= t_{12}\mathbf{a}^1 + t_{22}\mathbf{a}^2\\
\mathbf{q}^3 &= t_{13}\mathbf{a}^1 + t_{23}\mathbf{a}^2 + t_{33}\mathbf{a}^3\\
&\vdots\\
\mathbf{q}^n &= t_{1n}\mathbf{a}^1 + t_{2n}\mathbf{a}^2 + \cdots + t_{nn}\mathbf{a}^n\,.
\end{aligned}
\quad\Longleftrightarrow\quad \exists \mathbf{T} \in \mathbb{R}^{n,n} \text{ upper triangular: } \mathbf{Q} = \mathbf{A}\mathbf{T}\,, \tag{3.3.3.3}$$

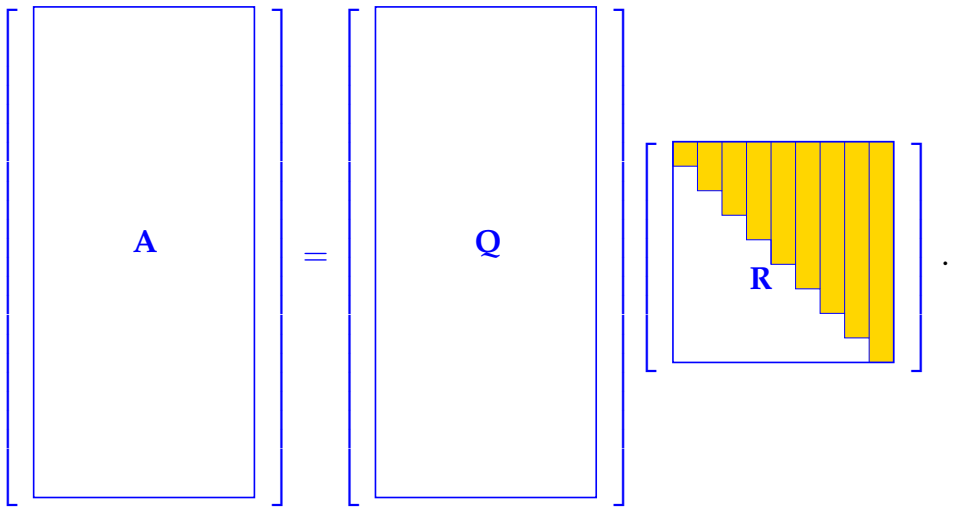
where $\mathbf{Q} = [\mathbf{q}_1,\dots,\mathbf{q}_n] \in \mathbb{R}^{m,n}$ (with orthonormal columns), $\mathbf{A} = [\mathbf{a}_1,\dots,\mathbf{a}_n] \in \mathbb{R}^{m,n}$. Note that thanks to the linear independence of $\{\mathbf{a}^1,\dots,\mathbf{a}^k\}$ and $\{\mathbf{q}^1,\dots,\mathbf{q}^k\}$, the matrix $\mathbf{T} = (t_{ij})_{i,j=1}^k \in \mathbb{R}^{k,k}$is regular ("non-existent" t_{ij} are set to zero, of course).

Recall from Lemma 1.3.1.9 that inverses of regular upper triangular matrices are upper triangular again.



Thus, by (3.3.3.3), we have found an *upper triangular* $\mathbf{R} := \mathbf{T}^{-1} \in \mathbb{R}^{n,n}$ such that

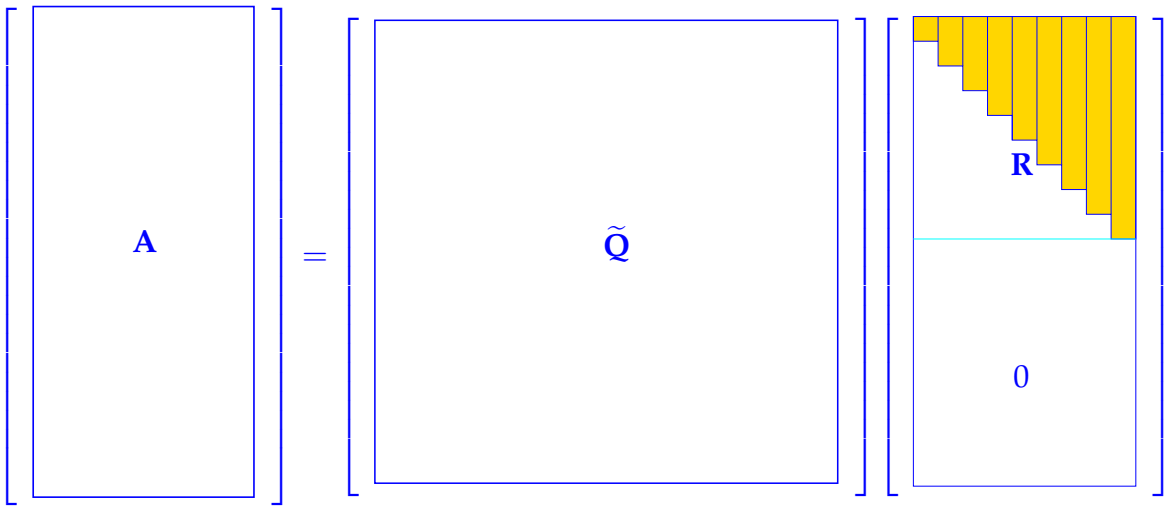
$$\mathbf{A} = \mathbf{QR} \quad \leftrightarrow$$



Next "augmentation by zero": add $m - n$ zero rows at the bottom of $\mathbf{R}$ and complement columns of $\mathbf{Q}$ to an orthonormal basis of $\mathbb{R}^m$, which yields an orthogonal matrix $\widetilde{\mathbf{Q}} \in \mathbb{R}^{m,m}$:

$$\blacktriangleright \quad \mathbf{A} = \widetilde{\mathbf{Q}} \begin{bmatrix} \mathbf{R} \\ 0 \end{bmatrix}$$

$$\updownarrow$$



$$\Leftrightarrow \quad \widetilde{\mathbf{Q}}^\top \mathbf{A} = \begin{bmatrix} \mathbf{R} \\ 0 \end{bmatrix} .$$

⌟

Thus the algorithm of Gram-Schmidt orthonormalization "proves" the following theorem.

Theorem 3.3.3.4. QR-decomposition → [NS02, Satz 5.2], [Gut07, Sect. 7.3]

For any matrix $\mathbf{A} \in \mathbb{K}^{n,k}$ *with* $\operatorname{rank}(\mathbf{A}) = k$ *there exists*

(i) a unique *Matrix* $\mathbf{Q}_0 \in \mathbb{R}^{n,k}$ *that satisfies* $\mathbf{Q}_0^{\mathrm{H}}\mathbf{Q}_0 = \mathbf{I}_k$, *and a* unique *upper triangular Matrix* $\mathbf{R}_0 \in \mathbb{K}^{k,k}$ *with* $(\mathbf{R})_{i,i} > 0$, $i \in \{1,\ldots,k\}$, *such that*

$$\mathbf{A} = \mathbf{Q}_0 \cdot \mathbf{R}_0 \qquad (\textbf{\textit{"economical" QR-decomposition}}) ,$$

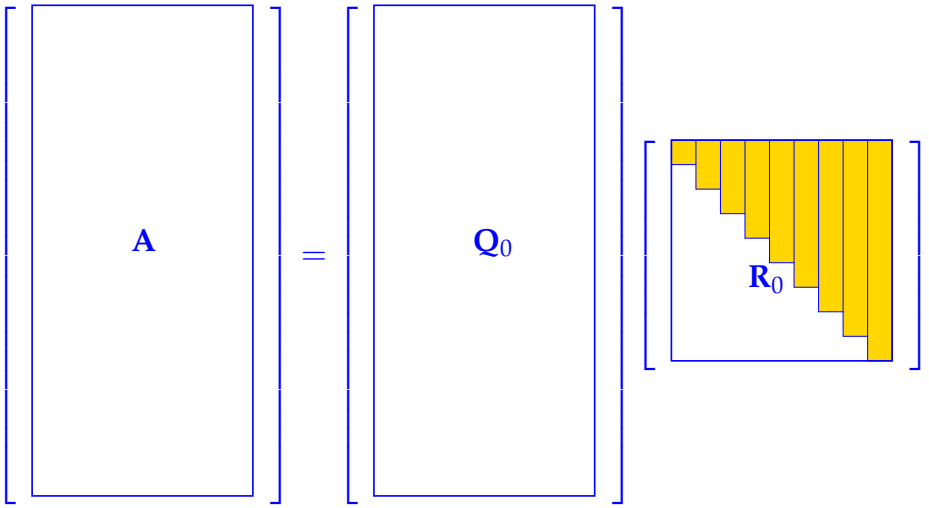
(ii) a unitary Matrix $\mathbf{Q} \in \mathbb{K}^{n,n}$ *and a* unique *upper triangular* $\mathbf{R} \in \mathbb{K}^{n,k}$ *with* $(\mathbf{R})_{i,i} > 0$, $i \in \{1,\ldots,n\}$, *such that*

$$\mathbf{A} = \mathbf{Q} \cdot \mathbf{R} \qquad (\textbf{\textit{full QR-decomposition}}) .$$

If $\mathbb{K} = \mathbb{R}$ *all matrices will be real and* $\mathbf{Q}$ *is then orthogonal.*

Visualisation: "economical" QR-decomposition: $\mathbf{Q}_0^{\mathrm{H}}\mathbf{Q}_0 = \mathbf{I}_k$ (orthonormal columns),

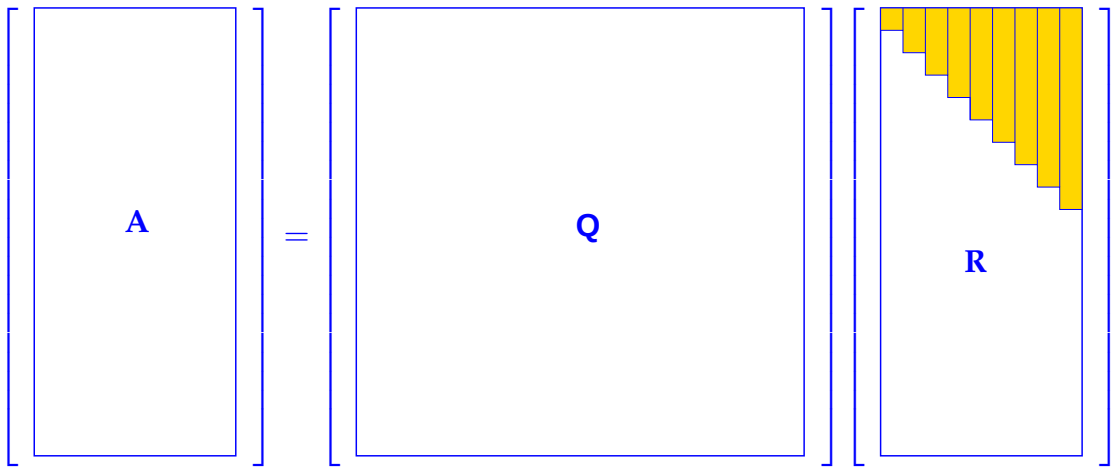
$$\mathbf{A} = \mathbf{Q}_0\mathbf{R}_0 , \quad \mathbf{Q}_0 \in \mathbb{K}^{n,k} , \quad \mathbf{R}_0 \in \mathbb{K}^{k,k} \text{ upper triangular ,}$$



. (3.3.3.5)

Visualisation: full QR-decompisiton: $\mathbf{Q}^{\mathrm{H}}\mathbf{Q} = \mathbf{Q}\mathbf{Q}^{\mathrm{H}} = \mathbf{I}_n$ (orthogonal matrix),

$$\mathbf{A} = \mathbf{Q}\mathbf{R} , \quad \mathbf{Q} \in \mathbb{K}^{n,n} , \quad \mathbf{R} \in \mathbb{K}^{n,k} ,$$



. (3.3.3.6)

For square $\mathbf{A}$, that is, $n = k$, both QR-decompositions coincide.

Corollary 3.3.3.7. Uniqueness of QR-factorization

The "economical" QR-factorization (3.3.3.1) *of* $\mathbf{A} \in \mathbb{K}^{m,n}$, $m \geq n$, *with* $\operatorname{rank}(\mathbf{A}) = n$ *is unique, if we demand* $(\mathbf{R}_0)_{ii} > 0$, $i = 1, \ldots, n$.

Proof. We observe that $\mathbf{R}$ is regular, if $\mathbf{A}$ has full rank n. Since the regular upper triangular matrices form a group under multiplication:

$$\mathbf{Q}_1\mathbf{R}_1 = \mathbf{Q}_2\mathbf{R}_2 \quad \Rightarrow \quad \mathbf{Q}_1 = \mathbf{Q}_2\mathbf{R} \quad \text{with upper triangular } \mathbf{R} := \mathbf{R}_2\mathbf{R}_1^{-1} .$$

$$\blacktriangleright \quad \mathbf{I} = \mathbf{Q}_1^{\mathrm{H}}\mathbf{Q}_1 = \mathbf{R}^{\mathrm{H}}\underbrace{\mathbf{Q}_2^{\mathrm{H}}\mathbf{Q}_2}_{=\mathbf{I}}\mathbf{R} = \mathbf{R}^{\mathrm{H}}\mathbf{R} .$$

The assertion follows by uniqueness of Cholesky decomposition, Lemma 2.8.0.15. □

Review question(s) 3.3.3.8 (QR-Decomposition: Theory)

(Q3.3.3.8.A) Assume that $\mathbf{A} \in \mathbb{R}^{m,n}$, $m \geq n$, satisfies that $\mathbf{A}^\top\mathbf{A}$ is a diagonal matrix with positive diagonal entries. Describe the economical QR-decomposition of $\mathbf{A}$.

(Q3.3.3.8.B) What is the QR-decomposition of a "right-lower triangular" square matrix $\mathbf{A} \in \mathbb{R}^{n,n}$ ($(\mathbf{A})_{i,j} = 0$ for $i + j \leq n$)?

(Q3.3.3.8.C) What is the $\mathbf{R}$ in the full QR-decomposition $\mathbf{A} = \mathbf{QR}$ of a tensor product matrix $\mathbf{A} = \mathbf{u}\mathbf{v}^\top$, $\mathbf{u} \in \mathbb{R}^m$, $\mathbf{v} \in \mathbb{R}^n$, $m, n \in \mathbb{N}$, $m \geq n$.

Hint. $\operatorname{rank}(\mathbf{A}) = \operatorname{rank}(\mathbf{R})$.

(Q3.3.3.8.D) Explain why the full QR-decomposition/QR-factorization of $\mathbf{A} \in \mathbb{R}^{m,n}$, $m > n$, cannot be unique, even if we demand $(\mathbf{R})_{ii} > 0$, $i = 1, \ldots, n$.

△

3.3.3.2 Computation of QR-Decomposition

Video tutorial for Section 3.3.3.2 & Section 3.3.3.4 "Computation of QR-Decomposition, QR-Decomposition in EIGEN": (32 minutes) Download link, tablet notes

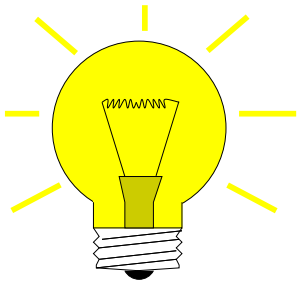
→ review questions 3.3.3.29

In theory, Gram-Schmidt orthogonalization (GS) can be used to compute the QR-factorization of a matrix $\mathbf{A} \in \mathbb{R}^{m,n}$, $m \geq n$, $\operatorname{rank}(\mathbf{A}) = n$. However, as we saw in Exp. 1.5.1.5, Gram-Schmidt orthogonalization in the form of Code 1.5.1.3 is *not* a *stable* algorithm.

There is a stable way to compute QR-decompositions, based on the accumulation of orthogonal transformations.

Corollary 3.3.3.9. Composition of orthogonal transformations

The product of two orthogonal/unitary matrices of the same size is again orthogonal/unitary.



Idea: find simple orthogonal (row) transformations rendering certain matrix elements zero:

$$\mathbf{Q}\begin{pmatrix} \blacksquare \end{pmatrix} = \begin{bmatrix} 0 \\ \blacksquare \end{bmatrix} \quad \text{with} \quad \mathbf{Q}^\top = \mathbf{Q}^{-1}\,.$$

Recall that this "annihilation of column entries" is the key operation in Gaussian forward elimination, where it is achieved by means of non-unitary row transformations, see Sect. 2.3.2. Now we want to find a counterpart of Gaussian elimination based on unitary row transformations on behalf of numerical stability.

EXAMPLE 3.3.3.10 ("Annihilating" orthogonal transformations in 2D) In 2D there are two possible orthogonal transformations make 2nd component of $\mathbf{a} \in \mathbb{R}^2$ vanish, which, in geometric terms, amounts to mapping the vector onto the x_1-axis.

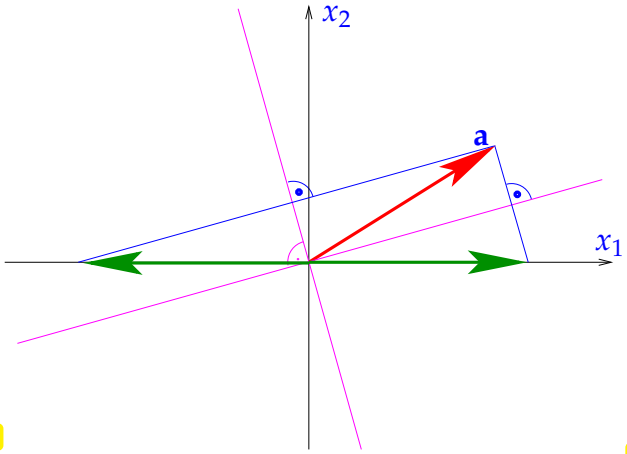


Fig. 80

reflections at angle bisector,

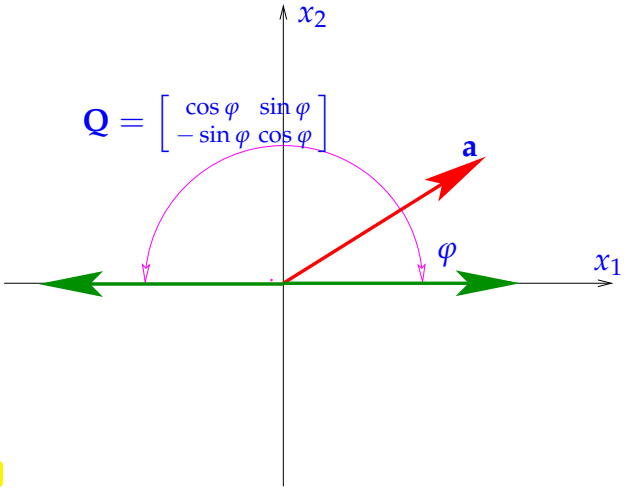


Fig. 81

rotations turning $\mathbf{a}$ onto x_1-axis.

Note that in each case we have *two different* length-preserving lineare mappings at our disposal. This flexibility will be important for curbing the impact of roundoff. ⌟

Both reflections and rotations are actually used in library routines and both are discussed in the sequel:

§3.3.3.11 (Householder reflections → [GV13, Sect. 5.1.2]) The following so-called Householder matrices (HHM) effect the reflection of a vector into a multiple of the first unit vector with the same length:

$$\mathbf{Q} = \mathbf{H}(\mathbf{v}) := \mathbf{I} - 2\frac{\mathbf{v}\mathbf{v}^\top}{\mathbf{v}^\top\mathbf{v}} \quad \text{with} \quad \mathbf{v} = \mathbf{a} \pm \|\mathbf{a}\|_2\mathbf{e}_1\,, \tag{3.3.3.12}$$

where $\mathbf{e}_1$ is the first Cartesian basis vector. Orthogonality of these matrices can be established by direct computation.

Fig. 82 depicts a "geometric derivation" of Householder reflections mapping $\mathbf{a} \to \mathbf{b}$, assuming $\|\mathbf{a}\|_2 = \|\mathbf{b}\|_2$. We accomplish this by a reflection at the hyperplane with normal vector $\mathbf{b} - \mathbf{a}$.

Given $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ with $\|\mathbf{a}\|_2 = \|\mathbf{b}\|_2$, the difference vector $\mathbf{v} := \mathbf{a} - \mathbf{b}$ is orthogonal to the bisector.

$$\begin{aligned}\mathbf{b} = \mathbf{a} - (\mathbf{a} - \mathbf{b}) &= \mathbf{a} - \mathbf{v}\frac{\mathbf{v}^\top \mathbf{v}}{\mathbf{v}^\top \mathbf{v}} \\ &= \mathbf{a} - 2\mathbf{v}\frac{\mathbf{v}^\top \mathbf{a}}{\mathbf{v}^\top \mathbf{v}} = \mathbf{a} - 2\frac{\mathbf{v}\mathbf{v}^\top}{\mathbf{v}^\top \mathbf{v}}\mathbf{a} = \mathbf{H}(\mathbf{v})\mathbf{a}\,,\end{aligned}$$

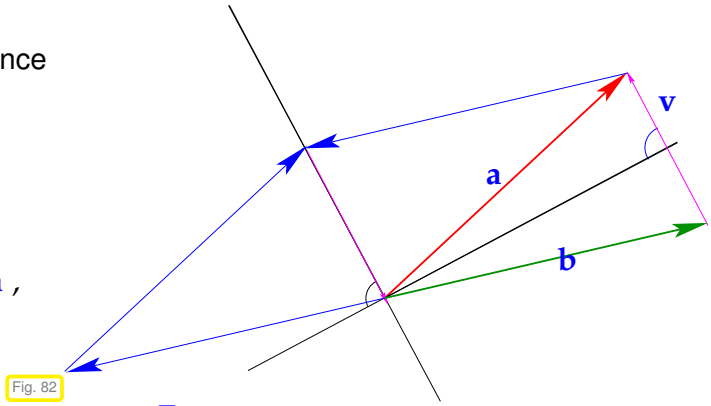


Fig. 82

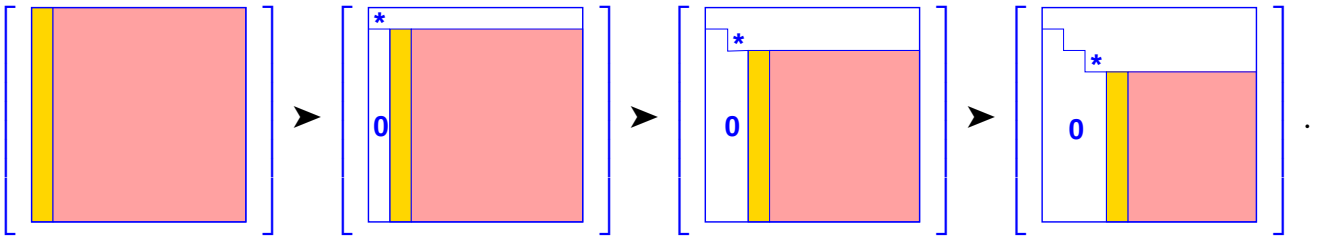
because, due to orthogonality $(\mathbf{a} - \mathbf{b}) \perp (\mathbf{a} + \mathbf{b})$ ($\Leftrightarrow (\mathbf{a} - \mathbf{b})^\top (\mathbf{a} + \mathbf{b}) = 0$),

$$\mathbf{v}^\top \mathbf{v} = (\mathbf{a} - \mathbf{b})^\top (\mathbf{a} - \mathbf{b}) = (\mathbf{a} - \mathbf{b})^\top (\mathbf{a} - \mathbf{b} + \mathbf{a} + \mathbf{b}) = 2(\mathbf{a} - \mathbf{b})^\top \mathbf{a} = 2\mathbf{v}^\top \mathbf{a}\,.$$

As a consequence, we have for the Householder matrix $\mathbf{H}(\mathbf{v})$ from (3.3.3.12)

$$\mathbf{H}(\mathbf{v})\mathbf{a} = \pm\|\mathbf{a}\|_2 \mathbf{e}_1\text{:} \qquad \mathbf{H}(\mathbf{v}) \begin{bmatrix} * \\ * \\ \vdots \\ * \end{bmatrix} = \begin{bmatrix} \pm\|\mathbf{a}\|_2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Hence, suitable successive Householder transformations determined by the leftmost column ("target column") of shrinking bottom right matrix blocks can be used to achieve upper triangular form $\mathbf{R}$. The following series of figures visualizes the gradual annihilation of the lower triangular matrix part for a square matrix:



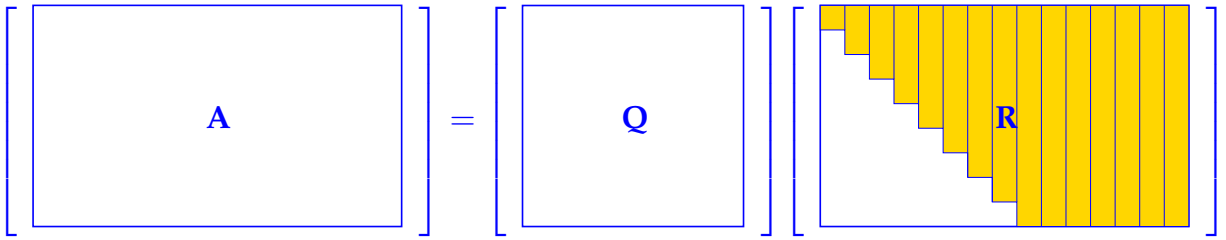
= "target column $\mathbf{a}$" (determines unitary transformation),
= modified in course of transformations.

Writing $\mathbf{Q}_\ell$ for the Householder matrix used in the ℓ-th factorization

$$\mathbf{Q}_{n-1}\mathbf{Q}_{n-2} \cdot \cdots \cdot \mathbf{Q}_1 \mathbf{A} = \mathbf{R}\,,$$

QR-decomposition (QR-factorization) of $\mathbf{A} \in \mathbb{C}^{n,n}$: $\mathbf{A} = \mathbf{Q}\mathbf{R}$, $\mathbf{Q} := \mathbf{Q}_1^\top \cdot \cdots \cdot \mathbf{Q}_{n-1}^\top$ orthogonal matrix, $\mathbf{R}$ upper triangular matrix.

Remark 3.3.3.13 (QR-decomposition of "fat" matrices) We can also apply successive Householder transformation as outlined in § 3.3.3.11 to a matrix $\mathbf{A} \in \mathbb{R}^{m,n}$ with $m < n$. If the first m columns of $\mathbf{A}$ are linearly independent, we obtain another variant of the QR-decomposition:



$$\mathbf{A} = \mathbf{QR} \quad , \quad \mathbf{Q} \in \mathbb{R}^{m,m} \, , \quad \mathbf{R} \in \mathbb{R}^{m,n} \, ,$$

where $\mathbf{Q}$ is orthogonal, $\mathbf{R}$ upper triangular, that is, $(\mathbf{R})_{i,j} = 0$ for $i > j$. ⌟

Remark 3.3.3.14 (Stable implementation of Householder reflections) In (3.3.3.12) the computation of the vector $\mathbf{v}$ can be prone to cancellation (→ Section 1.5.4), if the vector $\mathbf{a}$ encloses a very small angle with the first unit vector, because in this case $\mathbf{v}$ can be very small and beset with a huge relative error. For instance, this occurs for

$$\mathbf{a} = \begin{bmatrix} 1 \\ \delta \\ \vdots \\ \delta \end{bmatrix} \, , \quad \delta \approx 0 \quad \Rightarrow \quad \mathbf{v} = \mathbf{a} - \|\mathbf{a}\| \mathbf{e}_1 = \begin{bmatrix} \xi \\ \delta \\ \vdots \\ \delta \end{bmatrix} \quad \text{with} \quad \xi \approx 0 \, .$$

This is a concern, because in the formula for the Householder matrix,

$$\mathbf{H}(\mathbf{v}) := \mathbf{I} - 2 \frac{\mathbf{v}\mathbf{v}^\top}{\mathbf{v}^\top \mathbf{v}} \, ,$$

$\mathbf{v}$ is normalized to unit length (division by $\|\mathbf{v}\|_2^2$), and then a large absolute error might result.

Fortunately, two choices for $\mathbf{v}$ are possible in (3.3.3.12) and at most one can be affected by cancellation. The right choice is

$$\mathbf{v} = \begin{cases} \mathbf{a} + \|\mathbf{a}\|_2 \mathbf{e}_1 & \text{, if } a_1 > 0 \, , \\ \mathbf{a} - \|\mathbf{a}\|_2 \mathbf{e}_1) & \text{, if } a_1 \leq 0 \, . \end{cases}$$

See [Hig02, Sect. 19.1] and [GV13, Sect. 5.1.3] for a discussion. ⌟

§3.3.3.15 (Givens rotations → [Han02, Sect. 14], [GV13, Sect. 5.1.8]) The 2D rotation displayed in Fig. 81 can be embedded in an identity matrix. Thus, the following orthogonal transformation, a Givens rotation, annihilates the k-th component of a vector $\mathbf{a} = [a_1, \ldots, a_n]^\top \in \mathbb{R}^n$. Here γ stands for $\cos(\varphi)$ and σ for $\sin(\varphi)$, φ the angle of rotation, see Fig. 81.

$$\mathbf{G}_{1k}(a_1, a_k)\mathbf{a} := \begin{bmatrix} \gamma & \cdots & \sigma & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots \\ -\sigma & \cdots & \gamma & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_k \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} a_1^{(1)} \\ \vdots \\ 0 \\ \vdots \\ a_n \end{bmatrix} \quad \text{for} \quad \begin{aligned} \gamma &= \frac{a_1}{\sqrt{|a_1|^2 + |a_k|^2}} \, , \\ \sigma &= \frac{a_k}{\sqrt{|a_1|^2 + |a_k|^2}} \, . \end{aligned} \tag{3.3.3.16}$$

Orthogonality (→ Def. 6.3.1.2) of $\mathbf{G}_{1k}(a_1, a_k)$ is verified immediately. Again, we have two options for an annihilating rotation, see Ex. 3.3.3.10. It will always be possible to choose one that avoids overflow [GV13, Sect. 5.1.8], see Code 3.3.3.17 for details.

C++ code 3.3.3.17: Stable Givens rotation of a 2D vector, [GV13, Alg. 5.1.3] ⇝ GITLAB

```
// plane (2D) Givens rotation avoiding cancellation
// Computes orthogonal G ∈ R^{2,2} with G^T a = [r 0] =: x, r = ±‖a‖_2
void planerot(const Eigen::Vector2d& a, Eigen::Matrix2d& G,
              Eigen::Vector2d& x) {
  int sign{1};
  const double anorm = a.norm();
  if (anorm != 0.0) { //
    double s; // s ↔ σ
```

```
    double c; // c ↔ γ
    if (std::abs(a[1]) > std::abs(a[0])) {  // Avoid overflow
      const double t = -a[0] / a[1];
      s = 1.0 / std::sqrt(1.0 + t * t);
      c = s * t;
      sign = -1;
    } else {
      const double t = -a[1] / a[0];
      c = 1.0 / std::sqrt(1.0 + t * t);
      s = c * t;
    }
    G << c, s, -s, c;  // Form 2×2 Givens rotation matrix
  } else {
    G.setIdentity();
  }
  x << (sign * anorm), 0.0;
}
```

We validate the implementation by straightforward computations using the variable names of Code 3.3.3.17 and $\mathbf{a} = \begin{bmatrix} a_0 \\ a_1 \end{bmatrix}$:

- Case $|a_1| \geq |a_0|$: $t = -\frac{a_0}{a_1}$, $s = \frac{1}{\sqrt{1+t^2}}$, $c = st$

$$\begin{bmatrix} c & s \\ -s & c \end{bmatrix}^\top \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} ca_0 - sa_1 \\ sa_0 + ca_1 \end{bmatrix} = \begin{bmatrix} sta_0 - sa_1 \\ sa_0 + sta_1 \end{bmatrix}$$
$$= \frac{1}{\sqrt{1+t^2}} \begin{bmatrix} ta_0 - a_1 \\ a_0 + ta_1 \end{bmatrix} = \frac{|a_1|}{\|\mathbf{a}\|_2} \begin{bmatrix} -\frac{a_0^2}{a_1} - a_1 \\ a_0 - a_0 \end{bmatrix} = \begin{bmatrix} \operatorname{sgn}(a_1)\|\mathbf{a}\|_2 \\ 0 \end{bmatrix}.$$

- Case $|a_0| > |a_1|$: $t = -\frac{a_1}{a_0}$, $c = \frac{1}{\sqrt{1+t^2}}$, $s = ct$

$$\begin{bmatrix} c & s \\ -s & c \end{bmatrix}^\top \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} ca_0 - cta_1 \\ cta_0 + ca_1 \end{bmatrix} = \frac{1}{\sqrt{1+t^2}} \begin{bmatrix} a_0 - ta_1 \\ ta_0 + a_1 \end{bmatrix} = \frac{|a_0|}{\|\mathbf{a}\|_2} \begin{bmatrix} a_0 + \frac{a_1^2}{a_0} \\ -a_1 + a_1 \end{bmatrix} = \begin{bmatrix} \operatorname{sgn}(a_0)\|\mathbf{a}\|_2 \\ 0 \end{bmatrix}.$$

So far, we know how to annihilate a single component of a vector by means of a Givens rotation that targets that component and some other (the first in (3.3.3.16)). However, for the sake of QR-decomposition we aim to map *all* components to zero except for the first.

☞ This can be achieved by $n-1$ *successive* Givens rotations, see also Code 3.3.3.19

$$\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ \vdots \\ a_n \end{bmatrix} \xrightarrow{\mathbf{G}_{12}(a_1,a_2)} \begin{bmatrix} a_1^{(1)} \\ 0 \\ a_3 \\ \vdots \\ a_n \end{bmatrix} \xrightarrow{\mathbf{G}_{13}(a_1^{(1)},a_3)} \begin{bmatrix} a_1^{(2)} \\ 0 \\ 0 \\ a_4 \\ \vdots \\ a_n \end{bmatrix} \xrightarrow{\mathbf{G}_{14}(a_1^{(2)},a_4)} \cdots \xrightarrow{\mathbf{G}_{1n}(a_1^{(n-2)},a_n)} \begin{bmatrix} a_1^{(n-1)} \\ 0 \\ \vdots \\ \\ \vdots \\ 0 \end{bmatrix} \tag{3.3.3.18}$$

✎ Notation: $\mathbf{G}_{ij}(a_1, a_2) \hat{=}$ Givens rotation (3.3.3.16) modifying rows i and j of the matrix.

C++11 code 3.3.3.19: Roating a vector onto the x_1-axis by successive Givens transformation
➽ **GITLAB**

```
// Orthogonal transformation of a (column) vector into a multiple of
// the first unit vector by successive Givens transformations
// Note that the output vector could be computed much more efficiently!
void givenscoltrf(const VectorXd &aIn, MatrixXd &Q, VectorXd &aOut) {
  const Eigen::Index n = aIn.size();
  // Assemble rotations in a dense matrix Q
  // For (more efficient) alternatives see Rem. 3.3.3.22
  Q.setIdentity(); // Start from Q = I
  Matrix2d G;
  aOut = aIn;
  for (Eigen::Index j = 1; j < n; ++j) {
    const double a0 = aOut[0];
    const double a1 = aOut[j];
    // Determine entries of 2D rotation matrix, see Code 3.3.3.17
    double s = NAN; // s ↔ σ
    double c = NAN; // c ↔ γ
    if (a1 != 0.0) {
      if (std::abs(a1) > std::abs(a0)) { // Avoid overflow
        const double t = -a0 / a1;
        s = 1.0 / std::sqrt(1.0 + t * t);
        c = s * t;
      } else {
        const double t = -a1 / a0;
        c = 1.0 / std::sqrt(1.0 + t * t);
        s = c * t;
      }
      G << c, s, -s, c; // Form 2 × 2 Givens rotation matrix
    } else {            // No rotation required
      G.setIdentity();
    }
    // select 1st and jth element of aOut and use the Map function
    // to prevent copying; equivalent to aOut([1,j]) in MATLAB
    Map<VectorXd, 0, InnerStride<>> aOutMap(aOut.data(), 2, InnerStride<>(j));
    aOutMap = G.transpose() * aOutMap;
    // select 1st and jth column of Q (Q(:,[1,j]) in MATLAB)
    Map<MatrixXd, 0, OuterStride<>> QMap(Q.data(), n, 2, OuterStride<>(j * n));
    // Accumulate orthogonal transformations in a dense matrix; just
       done for
    // demonstration purposes!  See Rem. 3.3.3.22
    QMap = QMap * G;
  }
}
```

Armed with these compound Givens rotations we can proceed as in the case of Householder reflections to accomplish the orthogonal transformation of a full-rank matrix to upper triangular form, see

C++11 code 3.3.3.20: QR-decomposition by successive Givens rotations ⇝ GITLAB

```
//!  QR decomposition of square matrix A by successive Givens
//!  transformations
void qrgivens(const MatrixXd &A, MatrixXd &Q, MatrixXd &R) {
  const Eigen::Index n = A.rows();
  // Assemble rotations in a dense matrix.
  // For (more efficient) alternatives see Rem. Rem. 3.3.3.22
  Q.setIdentity();
  Matrix2d G;
  Vector2d tmp;
  Vector2d xDummy;
  R = A; // In situ transformation
```

```
13    for (Eigen::Index i = 0; i < n - 1; ++i) {
14      for (Eigen::Index j = n - 1; j > i; --j) {
15        tmp(0) = R(j - 1, i);
16        tmp(1) = R(j, i);
17        planerot(tmp, G, xDummy); // see Code 3.3.3.17
18        R.block(j - 1, 0, 2, n) = G.transpose() * R.block(j - 1, 0, 2, n);
19        Q.block(0, j - 1, n, 2) = Q.block(0, j - 1, n, 2) * G;
20      }
21    }
22  }
```

⌟

Remark 3.3.3.21 (Testing `!= 0.0` in Code 3.3.3.17) In light of the guideline "do not test floating point number for exact equality" from Rem. 1.5.3.15 the test **if** `(anorm != 0.0)` Line 8 looks inappropriate. However, its sole purpose is to avoid division by zero and the code will work well even if `anorm`≈ 0. ⌟

Remark 3.3.3.22 (Storing orthogonal transformations) When doing successive orthogonal transformations as in the case of QR-decomposition by means of Householder reflections ($\rightarrow$ § 3.3.3.11) or Givens rotations ($\rightarrow$ § 3.3.3.15) it would be prohibitively expensive to assemble and even multiply the transformation matrices!

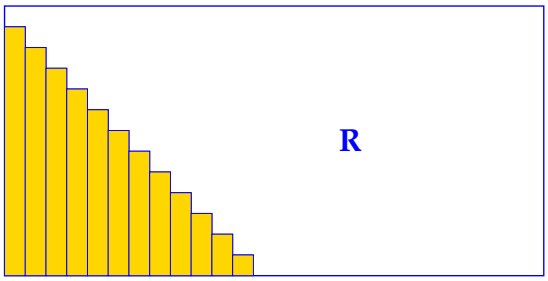
> The matrices for the orthogonal transformation are never built in codes!
> The transformations are stored in a compressed format.

Therefore, we stress that Code 3.3.3.20 is meant for demonstration purposes only, because the construction of the Q-factor matrix would never be done in this way in a well-designed numerical code.

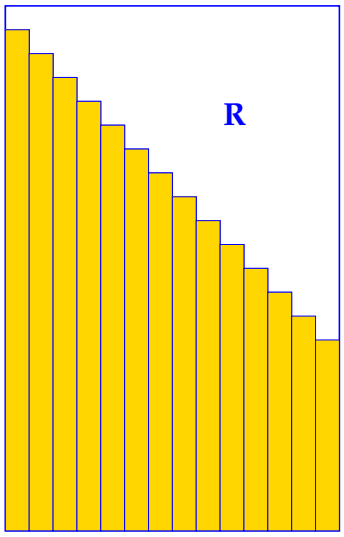
❶ In the case of Householder reflections $H(\mathbf{v}) \in \mathbb{R}^{m,m}$ (3.3.3.12), see [**GOV13**],

➤ store only the last $n-1$ components of the normalized vector $\mathbf{v} \in \mathbb{R}^m$

For QR-decomposition of a matrix $\mathbf{A} \in \mathbb{R}^{m,n}$, by means of successive Householder reflections $\mathbf{H}(\mathbf{v}_1) \cdot \cdots \cdot \mathbf{H}(\mathbf{v}_k)$, $k := \min\{m,n\}$, we store the bottom parts of the vectors $\mathbf{v}_j \in \mathbb{R}^{m-j+1}$, $j = 1,\ldots,k$, whose lengths decrease, in place of the "annihilated" lower triangular part of $\mathbf{A}$, which yields an in-situ QR-factorization.



↑ Case $m < n$



Case $m > n \rightarrow$

$\leftrightarrow$ space for Householder vectors

❷ In the case of Givens rotations, for a single rotation $G_{i,j}(a_1,a_2)$ we need store only the row indices (i,j) and rotation angle [Ste76], [GV13, Sect. 5.1.11]. The latter is subject to a particular encoding scheme:

$$\text{for} \quad \mathbf{G} = \begin{bmatrix} \gamma & \sigma \\ -\sigma & \gamma \end{bmatrix} \quad \Rightarrow \quad \text{store} \quad \rho := \begin{cases} 1_{\mathbb{M}} & \text{, if } \gamma = 0 \text{ ,} \\ \frac{1}{2}\operatorname{sign}(\gamma)\sigma & \text{, if } |\sigma| < |\gamma| \text{ ,} \\ 2\operatorname{sign}(\sigma)/\gamma & \text{, if } |\sigma| \geq |\gamma| \text{ ,} \end{cases} \tag{3.3.3.23}$$

$$\text{which means} \quad \begin{cases} \rho = 1_{\mathbb{M}} & \Rightarrow \quad \gamma = 0 \text{ ,} \quad \sigma = 1 \text{ ,} \\ |\rho| < 1 & \Rightarrow \quad \sigma = 2\rho \text{ ,} \quad \gamma = \sqrt{1-\sigma^2} \text{ ,} \\ |\rho| > 1 & \Rightarrow \quad \gamma = 2/\rho \text{ ,} \quad \sigma = \sqrt{1-\gamma^2} \text{ .} \end{cases} \tag{3.3.3.24}$$

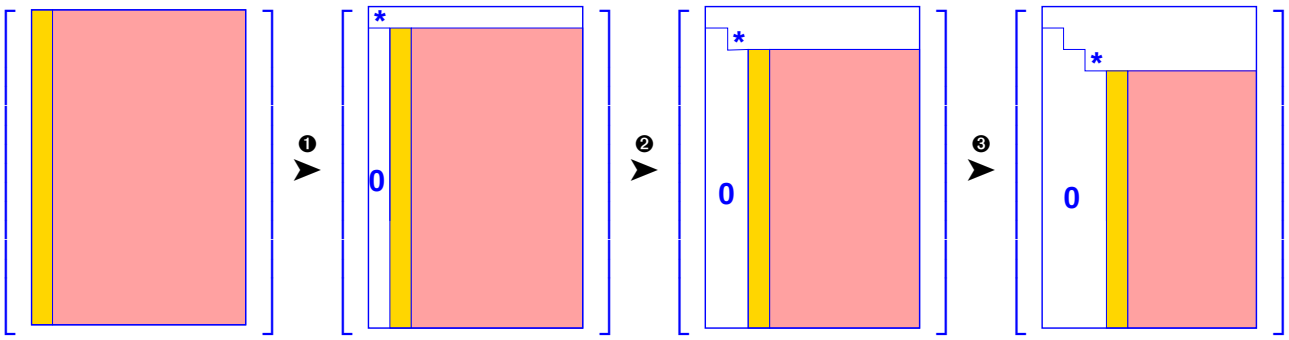
Here $1_{\mathbb{M}<}$ alludes to the fact that the number 1.0 can be represented exactly in machine number systems.

Then store $\mathbf{G}_{ij}(a,b)$ as triple (i,j,ρ). The parameter ρ forgets the sign of the matrix $\mathbf{G}_{ij}$, so the signs of the corresponding rows in the transformed matrix $\mathbf{R}$ have to be changed accordingly. The rationale behind the above convention is to curb the impact of roundoff errors, because when we recover γ,σ by taking the square root of a difference we never subtract two numbers of equal size; *cancellation is avoided*.

Summing up, when asking to "compute the economical/full QR-decomposition" $\mathbf{A} = \mathbf{QR}$ of a matrix $\mathbf{A}$, we request the upper triangular matrix $\mathbf{R}$ plus $\mathbf{Q}$ in a format that permits us to evaluate $\mathbf{Q}\times$vector efficiently.
⌟

§3.3.3.25 (Computational cost of computing QR-decompositions) How many elementary operations are asymptotically involved in the computation of the R-factor of the QR-decomposition of a "tall" matrix $\mathbf{A} \in \mathbb{R}^{m,n}$, $m \geq n$ based on Householder reflections as explained in § 3.3.3.11?

Obviously, the creation of zeros in the lower triangular part of $\mathbf{A}$ can be accomplished by n Householder transformation steps, as is illustrated by the following figure.



Note that the multiplication of a vector $\mathbf{w} \in \mathbb{R}^m$ with a Householder matrix $\mathbf{H}(\mathbf{v}) := \mathbf{I} - 2\mathbf{v}\mathbf{v}^\top$, $\mathbf{v} \in \mathbb{R}^m$, $\|\mathbf{v}\|_2 = 1$, takes only $2m$ operations, *cf.* Ex. 1.4.3.1.

Next, we examine the elementary matrix vector operations involved in orthogonal transformation of $\mathbf{A}$ into an upper triangular matrix $\mathbf{R} \in \mathbb{R}^{m,n}$.

Step ❶:	Householder matrix $\times$ $n-1$ remaining matrix cols. of size m,	cost = $2m(n-1)$
Step ❷:	Householder matrix $\times$ $n-2$ remaining matrix cols. of size $m-1$,	cost = $2(m-1)(n-2)$
Step ❸:	Householder matrix $\times$ $n-3$ remaining matrix cols. of size $m-2$,	cost = $2(m-2)(n-3)$
	$\vdots$	

We see that the combined number of entries of the -colored matrix blocks in the above figures is proportional to the total work.

$$\begin{aligned} \operatorname{cost}(\text{R-factor of } \mathbf{A} \text{ by Householder trf.}) &= \sum_{k=1}^{n-1} 2(m-k+1)(n-k) \\ &= O(mn^2) \quad \text{for} \quad m,n \to \infty \text{ .} \end{aligned} \tag{3.3.3.26}$$

⌟

Remark 3.3.3.27 (QR-decomposition of banded matrices) The advantage of Givens rotations is its selectivity, which can be exploited for banded matrices, see Section 2.7.5.

Definition 2.7.5.1. Bandwidth

For $\mathbf{A} = (a_{ij})_{i,j} \in \mathbb{K}^{m,n}$ we call

$$\overline{\mathrm{bw}}(\mathbf{A}) := \min\{k \in \mathbb{N}: j - i > k \Rightarrow a_{ij} = 0\} \text{ upper bandwidth ,}$$
$$\underline{\mathrm{bw}}(\mathbf{A}) := \min\{k \in \mathbb{N}: i - j > k \Rightarrow a_{ij} = 0\} \text{ lower bandwidth .}$$

$\mathrm{bw}(\mathbf{A}) := \overline{\mathrm{bw}}(\mathbf{A}) + \underline{\mathrm{bw}}(\mathbf{A}) + 1$ = **bandwidth** of $\mathbf{A}$.

Specific case: Orthogonal transformation of an $n \times n$ *tridiagonal* matrix to upper triangular form, that is, the annihilation of the sub-diagonal, by means of successive Givens rotations:

$$\begin{bmatrix} * & * & 0 & 0 & 0 & 0 & 0 & 0 \\ * & * & * & 0 & 0 & 0 & 0 & 0 \\ 0 & * & * & * & 0 & 0 & 0 & 0 \\ 0 & 0 & * & * & * & 0 & 0 & 0 \\ 0 & 0 & 0 & * & * & * & 0 & 0 \\ 0 & 0 & 0 & 0 & * & * & * & 0 \\ 0 & 0 & 0 & 0 & 0 & * & * & * \\ 0 & 0 & 0 & 0 & 0 & 0 & * & * \end{bmatrix} \xrightarrow{\mathbf{G}_{12}} \begin{bmatrix} * & * & * & 0 & 0 & 0 & 0 & 0 \\ 0 & * & * & 0 & 0 & 0 & 0 & 0 \\ 0 & * & * & * & 0 & 0 & 0 & 0 \\ 0 & 0 & * & * & * & 0 & 0 & 0 \\ 0 & 0 & 0 & * & * & * & 0 & 0 \\ 0 & 0 & 0 & 0 & * & * & * & 0 \\ 0 & 0 & 0 & 0 & 0 & * & * & * \\ 0 & 0 & 0 & 0 & 0 & 0 & * & * \end{bmatrix} \xrightarrow{\mathbf{G}_{23}\cdots\mathbf{G}_{n-1,n}} \begin{bmatrix} * & * & * & 0 & 0 & 0 & 0 & 0 \\ 0 & * & * & * & 0 & 0 & 0 & 0 \\ 0 & 0 & * & * & * & 0 & 0 & 0 \\ 0 & 0 & 0 & * & * & * & 0 & 0 \\ 0 & 0 & 0 & 0 & * & * & * & 0 \\ 0 & 0 & 0 & 0 & 0 & * & * & * \\ 0 & 0 & 0 & 0 & 0 & 0 & * & * \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & * \end{bmatrix}$$

$*$ $\hat{=}$ entry set to zero by Givens rotation, $*$ $\hat{=}$ new non-zero entry ("fill-in" → Def. 2.7.4.3).

This is a manifestation of a more general result, see Def. 2.7.5.1 for notations:

Theorem 3.3.3.28. QR-decomposition "preserves bandwidth"

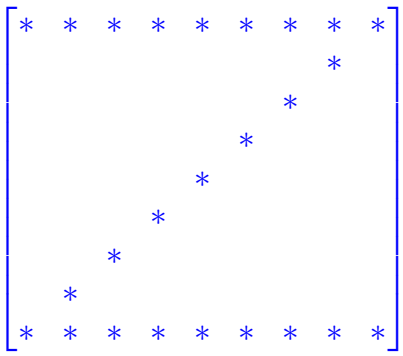
If $\mathbf{A} = \mathbf{QR}$ is the QR-decomposition of a regular matrix $\mathbf{A} \in \mathbb{R}^{n,n}$, then $\mathrm{bw}(\mathbf{R}) \leq \mathrm{bw}(\mathbf{A})$.

Studying the algorithms sketched above for tridiagonal matrices, we find that a total of at most $n \cdot \mathrm{bw}(\mathbf{A})$ Givens rotations is required or computing the QR-decomposition. Each of them acts on $O(\mathrm{bw}(\mathbf{A}))$ non-zero entries of the matrix, which leads to an asymptotic total computational effort of $O(n \cdot \mathrm{bw}(\mathbf{A})^2)$ for $n \to \infty$. ⌟

Review question(s) 3.3.3.29 (Computation of QR-decompositions)

(Q3.3.3.29.A) Let $\mathbf{A} \in \mathbb{R}^{n,n}$ be "Z-shaped"

$$(\mathbf{A})_{i,j} = 0 \text{ ,if } \quad i \in \{2, \ldots, n-1\}\,, i + j \neq n + 1 \quad ,$$

e.g. $\mathbf{A} =$  .

1. Give a sequence of Givens rotations that convert $\mathbf{A}$ into upper triangular form.
2. Think about an efficient way to deploy orthogonal transformation techniques for the efficient solution of a linear system of equations $\mathbf{Ax} = \mathbf{b}$, $\mathbf{b} \in \mathbb{R}^n$.

(Q3.3.3.29.B) The matrix $\mathbf{A} \in \mathbb{R}^{n,n}$, $n \in \mathbb{N}$, is upper triangular except for a single non-zero entry in position $(n,1)$:

$$(\mathbf{A})_{i,j} = 0 \quad \text{, if} \quad i > j \quad \text{and} \quad (i,j) \neq (n,1) \,.$$

Which sequence of Givens rotations (of minimal length) can be used to compute the QR-decomposition of $\mathbf{A}$?

(Q3.3.3.29.C) [Householder matrices] What is a Householder matrix and what are its properties (regularity, orthogonality, symmetry, rank, kernel, range)?

△

3.3.3.3 QR-Decomposition: Stability

In numerical linear algebra orthogonal transformation methods usually give rise to reliable algorithms, thanks to the norm-preserving property of orthogonal transformations.

§3.3.3.30 (Stability of unitary/orthogonal transformations) We consider the mapping (the "transformation" induced by $\mathbf{Q}$)

$$F : \mathbb{K}^n \to \mathbb{K}^n \,, \quad F(\mathbf{x}) := \mathbf{Qx} \,, \quad \mathbf{Q} \in \mathbb{K}^{n,n} \text{ unitary/orthogonal} \,.$$

We are interested in the sensitivity of F, that is, the impact of relative errors in the data vector $\mathbf{x}$ on the output vector $\mathbf{y} := F(\mathbf{x})$.

We study the output for a perturbed input vector:

$$\left.\begin{array}{rcl} \mathbf{Qx} = \mathbf{y} & \Rightarrow & \|\mathbf{x}\|_2 = \|\mathbf{y}\|_2 \\ \mathbf{Q}(\mathbf{x}+\Delta\mathbf{x}) = \mathbf{y}+\Delta\mathbf{y} \quad \Rightarrow \quad \mathbf{Q}\Delta\mathbf{x} = \Delta\mathbf{y} & \Rightarrow & \|\Delta\mathbf{y}\|_2 = \|\Delta\mathbf{x}\|_2 \end{array}\right\} \quad \blacktriangleright \quad \frac{\|\Delta\mathbf{y}\|_2}{\|\mathbf{y}\|_2} = \frac{\|\Delta\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \,.$$

We conclude, that unitary/orthogonal transformations do *not* cause any *amplification of relative errors* in the data vectors.

Of course, this also applies to the "solution" of square linear systems with orthogonal coefficient matrix $\mathbf{Q} \in \mathbb{R}^{n,n}$, which, by Def. 6.3.1.2, boils down to multiplication of the right hand side vector with $\mathbf{Q}^{\mathrm{H}}$. ⌟

Remark 3.3.3.31 (Conditioning of conventional row transformations) Gaussian elimination as presented in § 2.3.1.3 converts a matrix to upper triangular form by elementary row transformations. Those add a scalar multiple of a row of the matrix to another matrix and amount to left-multiplication with matrices

$$\mathbf{T} := \mathbf{I}_n + \mu \mathbf{e}_j \mathbf{e}_i^\top \,, \quad \mu \in \mathbb{K} \,, \quad i,j \in \{1,\ldots,n\},\ i \neq j \,. \tag{3.3.3.32}$$

However, these transformations can lead to a massive amplification of relative errors, which, by virtue of Ex. 2.2.2.1 can be linked to large condition numbers of $\mathbf{T}$.

This accounts for fact that the computation of LU-decompositions by means of Gaussian elimination might not be stable, see Ex. 2.4.0.6. ⌟

EXPERIMENT 3.3.3.33 (Conditioning of conventional row transformations, Rem. 3.3.3.31 cnt'd)

Study in 2D:

2×2 row transformation matrix, (*cf.* elimination matrices of Gaussian elimination.

$$\mathbf{T}(\mu) = \begin{pmatrix} 1 & 0 \\ \mu & 1 \end{pmatrix}$$

Euclidean condition numbers of $\mathbf{T}(\mu)$ ▷

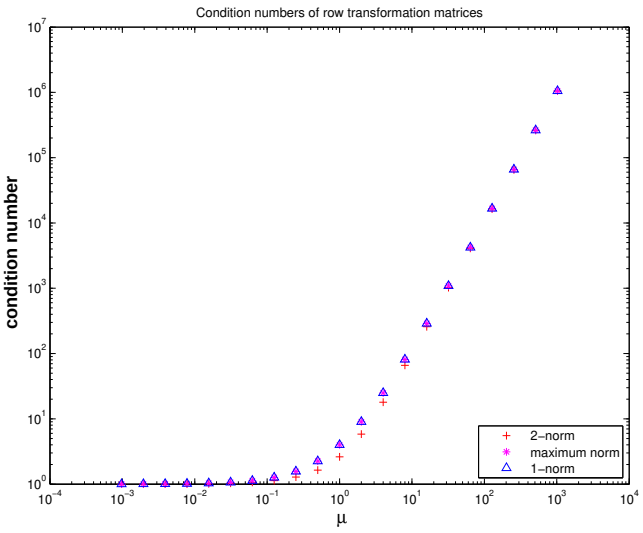


Fig. 83

⌟

The perfect conditioning of orthogonal transformation prevents the destructive build-up of roundoff errors.

Theorem 3.3.3.34. Stability of Householder QR [Hig02, Thm. 19.4]

Let $\widetilde{\mathbf{R}} \in \mathbb{R}^{m,n}$ be the R-factor of the QR-decomposition of $\mathbf{A} \in \mathbb{R}^{m,n}$ computed by means of successive Householder reflections (→ § 3.3.3.11). Then there exists an orthogonal $\mathbf{Q} \in \mathbb{R}^{m,m}$ such that

$$\mathbf{A} + \Delta\mathbf{A} = \mathbf{Q}\widetilde{\mathbf{R}} \quad \text{with} \quad \|\Delta\mathbf{A}\|_2 \leq \frac{cmn\,\mathrm{EPS}}{1 - cmn\,\mathrm{EPS}}\|\mathbf{A}\|_2 \,, \tag{3.3.3.35}$$

where EPS is the machine precision and $c > 0$ a small constant independent of $\mathbf{A}$.

3.3.3.4 QR-Decomposition in EIGEN

EIGEN offers several classes dedicated to computing QR-type decompositions of matrices, for instance **HouseholderQR**. Internally the QR-decomposition is stored in compressed format as explained in Rem. 3.3.3.22. Its computation is triggered by the constructor.

C++-code 3.3.3.36: QR-decompositions in EIGEN ⇝ GITLAB

```
2  # include <Eigen/QR>
3
4  // Computation of full QR-decomposition (3.3.3.1),
5  // dense matrices built for both QR-factors (expensive!)
6  inline std::pair<MatrixXd,MatrixXd> qr_decomp_full(const MatrixXd& A) {
7    const Eigen::HouseholderQR<MatrixXd> qr(A);
8    const MatrixXd Q = qr.householderQ(); //
9    const MatrixXd R = qr.matrixQR().template triangularView<Eigen::Upper>();
10   return {Q,R};
11 }
12
13 // Computation of economical QR-decomposition (3.3.3.1),
14 // dense matrix built for Q-factor (possibly expensive!)
15 inline std::pair<MatrixXd,MatrixXd> qr_decomp_eco(const MatrixXd& A) {
```

```
16   using index_t = MatrixXd::Index;
17   const index_t m = A.rows();
18   const index_t n = A.cols();
19   const Eigen::HouseholderQR<MatrixXd> qr(A);
20   const MatrixXd Q = (qr.householderQ()*MatrixXd::Identity(m,n)); //
21   const MatrixXd R = qr.matrixQR().block(0,0,n,n).template
         triangularView<Eigen::Upper>(); //
22   return {Q,R};
23 }
```

Note that the method `householderQ` returns the Q-factor in compressed format, refer to Rem. 3.3.3.22. Assignment to a matrix will convert it into a (dense) matrix format, see Line 8; only then the actual computation of the matrix entries is performed. It can also be multiplied with another matrix of suitable size, which is used in Line 20 to extract the Q-factor $\mathbf{Q}_0 \in \mathbb{R}^{m,n}$ of the economical QR-decomposition (3.3.3.1).

The matrix returned by the method `matrixQR()` gives access to a matrix storing the QR-factors in compressed form. Its upper triangular part provides $\mathbf{R}$, see Line 21.

§3.3.3.37 (Economical versus full QR-decomposition) The distinction of Thm. 3.3.3.4 between economical and full QR-decompositions of a "tall" matrix $\mathbf{A} \in \mathbb{R}^{m,n}$, $m > n$, becomes blurred on the algorithmic level. If all we want is a representation of the Q-factor as a product of orthogonal transformations as discussed in Rem. 3.3.3.22, exactly the same computations give us both types of QR-decompositions, because, of course, the bottom zero block of $\mathbf{R}$ need not be stored.

> The same computations yield both full and economical QR-decompositions with Q-factors in product form.

This is clearly reflected in Code 3.4.2.1. Thus, in the derivation of algorithms we choose either type of QR-decomposition, whichever is easier to understand. ⌟

§3.3.3.38 (Cost of QR-decomposition in EIGEN) A close inspection of the algorithm for the computation of QR-decompositions of $\mathbf{A} \in \mathbb{R}^{m,n}$ by successive Householder reflections ($\rightarrow$ § 3.3.3.11) reveals, that n transformations costing $\sim mn$ operations each are required.

Asymptotic complexity of Householder QR-decomposition

The computational effort for **HouseholderQR()** of $\mathbf{A} \in \mathbb{R}^{m,n}$, $m > n$, is $O(mn^2)$ for $m, n \to \infty$.

⌟

EXPERIMENT 3.3.3.40 (Asymptotic complexity of Householder QR-factorization) We empirically investigate the (asymptotic) complexity of QR-factorization algorithms in EIGEN through runtime measurements.

C++-code 3.3.3.41: timing QR-factorizations in EIGEN ↬ GITLAB

```
2   int nruns = 3, minExp = 2, maxExp = 6;
3   MatrixXd tms(maxExp-minExp+1,4);
4   for(int i = 0; i <= maxExp-minExp; ++i){
5     Timer t1, t2, t3;    // timer class
6     int n = std::pow(2, minExp + i); int m = n*n;
7     // Initialization of matrix A
8     MatrixXd A(m,n); A.setZero();
9     A.setIdentity(); A.block(n,0,m-n,n).setOnes();
10    A += VectorXd::LinSpaced(m,1,m) * RowVectorXd::LinSpaced(n,1,n);
```

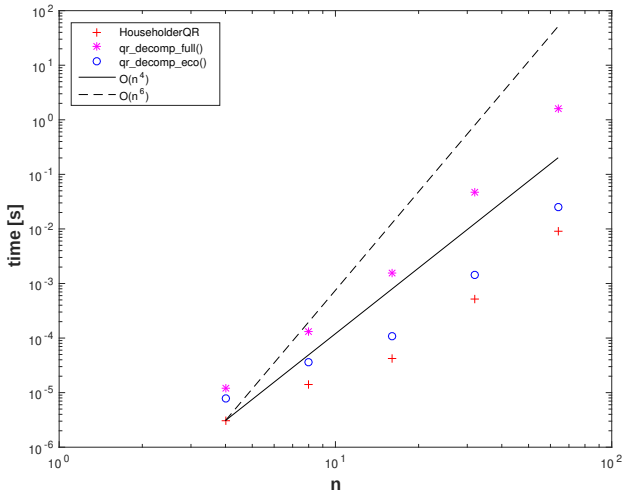
```
11      for(int j = 0; j < nruns; ++j) {
12        // plain QR-factorization in the constructor
13        t1.start(); HouseholderQR<MatrixXd> qr(A); t1.stop();
14        // full decomposition
15        t2.start(); std::pair<MatrixXd,MatrixXd> QR2 = qr_decomp_full(A); t2.stop();
16        // economic decomposition
17        t3.start(); std::pair<MatrixXd,MatrixXd> QR3 = qr_decomp_eco(A); t3.stop();
18      }
19      tms(i,0)=n;
20      tms(i,1)=t1.min(); tms(i,2)=t2.min(); tms(i,3)= t3.min();
21    }
```

Timings for

- plain QR-factorization in the constructor of **HouseholderQR**,
- invocation of function `qr_decomp_full()`, see Code 3.4.2.1,
- call to `qr_decomp_eco()` from Code 3.4.2.1.

Platform:

- ubuntu 14.04 LTS
- CPU i7-3517U, 1.90GHZ, 4 cores
- L1 32 KB, L2 256 KB, L3 4096 KB, Mem 8 GB
- gcc 4.8.4, -O3



The runtimes for the QR-factorization of $\mathbf{A} \in \mathbb{R}^{n^2,n}$ behave like $O(n^2 \cdot n)$ for large n.

3.3.4 QR-Based Solver for Linear Least Squares Problems

Video tutorial for Section 3.3.4 "QR-Based Solver for Linear Least Squares Problems": (9 minutes) Download link, tablet notes

→ review questions 3.3.4.8

The QR-decomposition introduced in Section 3.3.3, Thm. 3.3.3.4, paves the way for the practical algorithmic realization of the "equivalent orthonormal transformation to upper triangular form"-idea from Section 3.3.1.

We consider the full-rank linear least squares problem (3.1.2.22): Given $\mathbf{A} \in \mathbb{R}^{m,n}$, $m \geq n$, $\mathrm{rank}(\mathbf{A}) = n$,

$$\text{seek} \quad \mathbf{x} \in \mathbb{R}^n \quad \text{such that} \quad \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2 \to \min \ .$$

We assume that we are given a
QR-decomposition: $\mathbf{A} = \mathbf{Q}\mathbf{R}$, $\mathbf{Q} \in \mathbb{R}^{m,m}$ orthogonal, $\mathbf{R} \in \mathbb{R}^{m,n}$ (regular) upper triangular matrix.

We apply the orthogonal 2-norm preserving (→ Thm. 3.3.2.2) transformation encoded in $\mathbf{Q}$ to $\mathbf{A}\mathbf{x} - \mathbf{b}$, the vector inside the 2-norm to be minimized:

$$\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2 = \left\|\mathbf{Q}(\mathbf{R}\mathbf{x} - \mathbf{Q}^\top \mathbf{b})\right\|_2 = \left\|\mathbf{R}\mathbf{x} - \widetilde{\mathbf{b}}\right\|_2 \quad , \quad \widetilde{\mathbf{b}} := \mathbf{Q}^\top \mathbf{b} \ .$$

Thus, we have obtained an *equivalent* triangular linear least squares problem:

$$\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2 \to \min \quad \Leftrightarrow \quad \left\| \begin{bmatrix} \mathbf{R}_0 \\ \mathbf{0} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} - \begin{bmatrix} \widetilde{b}_1 \\ \vdots \\ \\ \\ \vdots \\ \widetilde{b}_m \end{bmatrix} \right\|_2 \to \min \ .$$

$$\mathbf{x} = \begin{bmatrix} \mathbf{R}_0 \end{bmatrix}^{-1} \begin{bmatrix} \widetilde{b}_1 \\ \vdots \\ \widetilde{b}_n \end{bmatrix} \ , \quad \text{with residual} \quad \mathbf{r} = \mathbf{Q} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \widetilde{b}_{n+1} \\ \vdots \\ \widetilde{b}_m \end{bmatrix} .$$

Note: by Thm. 3.3.2.2 the norm of the residual is readily available: $\|\mathbf{r}\|_2 = \sqrt{\widetilde{b}_{n+1}^2 + \cdots + \widetilde{b}_m^2}$.

C++-code 3.3.4.1: QR-based solver for full rank linear least squares problem (3.1.2.22)
➙ GITLAB

```
// Solution of linear least squares problem (3.1.2.22) by means of
    QR-decomposition
// Note:  A ∈ ℝ^{m,n} with m > n, rank(A) = n is assumed
// Least squares solution returned in x, residual norm as return value
double qrlsqsolve(const MatrixXd& A, const VectorXd& b,
                  VectorXd& x) {
  const unsigned m = A.rows();
  const unsigned n = A.cols();

  MatrixXd Ab(m, n + 1); Ab << A, b; // Form extended matrix [A,b]

  // QR-decomposition of extended matrix automatically transforms b
  MatrixXd R = Ab.householderQr().matrixQR().template
    triangularView<Eigen::Upper>(); //

  MatrixXd R_nn = R.block(0, 0, n, n); // R-factor R_0
  // Compute least squares solution x = (R)^{-1}_{1:n,1:n}(Q^T b)_{1:n}
  x = R_nn.template triangularView<Eigen::Upper>().solve(R.block(0, n, n, 1));
  return R(n, n); // residual norm = ||Ax̂ - b||_2 (why ?)
}
```

Discussion of (some) details of implementation in Code 3.3.4.1:

- The QR-decomposition is computed in a numerically stable way by means of Householder reflections (→ § 3.3.3.11) by EIGEN's built-in function **householderQR** available for matrix type. The

computational cost of this function when called for an $m \times n$ matrix is, asymptotically for $m, n \to \infty$, $O(n^2 m)$.

- Line 10: We perform the QR-decomposition of the extended matrix $[\mathbf{A}, \mathbf{b}]$ with $\mathbf{b}$ as rightmost column. Thus, the orthogonal transformations are automatically applied to $\mathbf{b}$; the augmented matrix is converted into $[\mathbf{R}, \mathbf{Q}^\top \mathbf{b}]$, the data of the equivalent upper triangular linear least squares problem. Thus, actually, no information about $\mathbf{Q}$ needs to be stored, if one is interested in the least squares solution $\mathbf{x}$ only.

  The idea is borrowed from Gaussian elimination, see Code 2.3.1.4, Line 9.

- Line 14: `MatrixQR()` returns the compressed QR-factorization as a matrix, where the R-factor $\mathbf{R} \in \mathbb{R}^{m,n}$ is contained in the upper triangular part, whose top n rows give $\mathbf{R}_0$ from see (3.3.3.1).

- Line 19: the components $(\mathbf{b})_{n+2:m}$ of the vector $\mathbf{b}$ (treated as rightmost column of the augmented matrix) are annihilated when computing the QR-decomposition (by final Householder reflection): $\left(\mathbf{Q}^\top [\mathbf{A}, \mathbf{b}]\right)_{n+2:m,n} = 0$. Hence, $\left(\mathbf{Q}^\top [\mathbf{A}, \mathbf{b}]\right)_{n+1,n+1} = \left\| (\widetilde{\mathbf{b}})_{n+1:m} \right\|_2$, which gives the norm of the residual.

➤ A QR-based algorithm is implemented in the `solve()` method available for EIGEN's QR-decomposition, see Code 3.3.4.2.

C++ code 3.3.4.2: EIGEN's built-in QR-based linear least squares solver ↬ GITLAB

```
2 // Solving a full-rank least squares problem ||Ax-b||_2 -> min in EIGEN
3 double lsqsolve_eigen(const MatrixXd& A, const VectorXd& b,
4                        VectorXd& x) {
5   x = A.householderQr().solve(b);
6   return ((A*x-b).norm());
7 }
```

Remark 3.3.4.3 (QR-based solution of linear systems of equations) Applying the QR-based algorithm for full-rank linear least squares problems in the case $m = n$, that is, to a square linear system of equations $\mathbf{Ax} = \mathbf{b}$ with a regular coefficient matrix , will compute the solution $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$. In a sense, the QR-decomposition offers an alternative to Gaussian elimination/LU-decomposition discussed in § 2.3.2.15.

The steps for solving a linear system of equations $\mathbf{Ax} = \mathbf{b}$ by means of QR-decomposition are as follows:

$\mathbf{Ax} = \mathbf{b}$:
① QR-decomposition $\mathbf{A} = \mathbf{QR}$, computational costs $\frac{2}{3}n^3 + O(n^2)$
(about twice as expensive as LU-decomposition without pivoting)
② orthogonal transformation $\mathbf{z} = \mathbf{Q}^\top \mathbf{b}$, computational costs $4n^2 + O(n)$
(in the case of *compact storage* of reflections/rotations)
③ Backward substitution, solve $\mathbf{Rx} = \mathbf{z}$, computational costs $\frac{1}{2}n(n+1)$

Benefit: we can utterly dispense with any kind of pivoting:

✌ Computing the generalized QR-decomposition $\mathbf{A} = \mathbf{QR}$ by means of Householder reflections or Givens rotations is (numerically stable) for any $\mathbf{A} \in \mathbb{C}^{m,n}$.

✌ For *any* regular system matrix an LSE can be solved by means of
QR-decomposition + orthogonal transformation + backward substitution
in a stable manner.

Drawback: QR-decomposition can hardly ever avoid massive fill-in (→ Def. 2.7.4.3) also in situations, where LU-factorization greatly benefits from Thm. 2.7.5.4. ⌟

Remark 3.3.4.4 (QR-based solution of banded LSE) From Rem. 3.3.3.27, Thm. 3.3.3.28, we know that that particular situtation, in which QR-decomposition can avoid fill-in (→ Def. 2.7.4.3) is the case of banded matrices, see Def. 2.7.5.1. For a banded $n \times n$ linear systems of equations with small fixed bandwidth $\mathrm{bw}(\mathbf{A}) \leq O(1)$ we incur an

➢ asymptotic computational effort: $O(n)$ for $n \to \infty$

The following code uses QR-decomposition computed by means of selective Givens rotations (→ § 3.3.3.15) to solve a tridiagonal linear system of equations $\mathbf{Ax} = \mathbf{b}$

$$\mathbf{A} = \begin{bmatrix} d_1 & c_1 & 0 & & \dots & & 0 \\ e_1 & d_2 & c_2 & & & & \vdots \\ 0 & e_2 & d_3 & c_3 & & & \\ \vdots & & \ddots & \ddots & \ddots & & c_{n-1} \\ 0 & & \dots & & 0 & e_{n-1} & d_n \end{bmatrix}$$

The matrix is passed in the form of three vectors $\mathbf{e}, \mathbf{c}, \mathbf{d}$ giving the entries in the non-zero bands.

C++ code 3.3.4.5: Solving a tridiagonal system by means of QR-decomposition ↬ GITLAB

```
//!  @brief Solves the tridiagonal system Ax = b with QR-decomposition
//!  @param[in] d Vector of dim n; the diagonal elements
//!  @param[in] c Vector of dim n-1; the lower diagonal elements
//!  @param[in] e Vector of dim n-1; the upper diagonal elements
//!  @param[in] b Vector of dim n; the rhs.
//!  @param[out] x Vector of dim n
VectorXd tridiagqr(VectorXd c, VectorXd d, VectorXd e, VectorXd& b){
  const Eigen::Index n = d.size();
  // resize the vectors c and d to correct length if needed
  c.conservativeResize(n); e.conservativeResize(n);
  const double t = d.norm() + e.norm() + c.norm();
  Matrix2d R;
  Vector2d z;
  Vector2d tmp;
  for(Eigen::Index k = 0; k < n-1; ++k){
    tmp(0) = d(k); tmp(1) = e(k);
    // Use givensrotation to set the entries below the diagonal
    // to zero
    planerot(tmp, R, z); // see Code 3.3.3.17
    if( std::abs(z(0))/t < std::numeric_limits<double>::epsilon() ) {
      throw std::runtime_error("A nearly singular");
    }
    // Update all other entries of the matrix and rhs.  which
    // were affected by the givensrotation
    d(k) = z(0);
    b.segment(k,2).applyOnTheLeft(R);   // rhs.
    // Block of the matrix affected by the givensrotation
    Matrix2d Z;
    Z << c(k),      0,  d(k+1), c(k+1);
    Z.applyOnTheLeft(R);
    // Write the transformed block back to the corresponding places
    c.segment(k,2) = Z.diagonal(); d(k+1) = Z(1,0); e(k) = Z(0,1);
  }
  // Note that the e is now above d and c
  // Backsubstitution acting on upper triangular matrix
  // with upper bandwidth 2 (stored in vectors).
  VectorXd x(n);
  // last row
  x(n-1) = b(n-1)/d(n-1);
  if(n >= 2) {
    // 2nd last row
```

```
43      x(n-2) = (b(n-2)-c(n-2)*x(n-1))/d(n-2);
44      // remaining rows
45      for(Eigen::Index i = n-3; i >= 0; --i) {
46        x(i) = ( b(i) - c(i) * x(i+1) - e(i)*x(i+2) ) / d(i);
47      }
48    }
49    return x;
50  }
```

⌟

EXAMPLE 3.3.4.6 (Stable solution of LSE by means of QR-decomposition) Aiming to confirm the claim of superior stability of QR-based approaches (→ Rem. 3.3.4.3, § 3.3.3.30) we revisit Wilkinson's counterexample from Ex. 2.4.0.6 for which Gaussian elimination with partial pivoting does not yield an acceptable solution.

Wilkinson matrix $\mathbf{A} \in \mathbb{R}^{n,n}$

$$(\mathbf{A})_{i,j} := \begin{cases} -1 & \text{for } i > j,\ j < n\ , \\ 1 & \text{for } i = j\ , \\ 0 & \text{for } i < j,\ j < n\ , \\ 1 & \text{for } j = n\ . \end{cases}$$

QR-decomposition produces perfect solution ▷

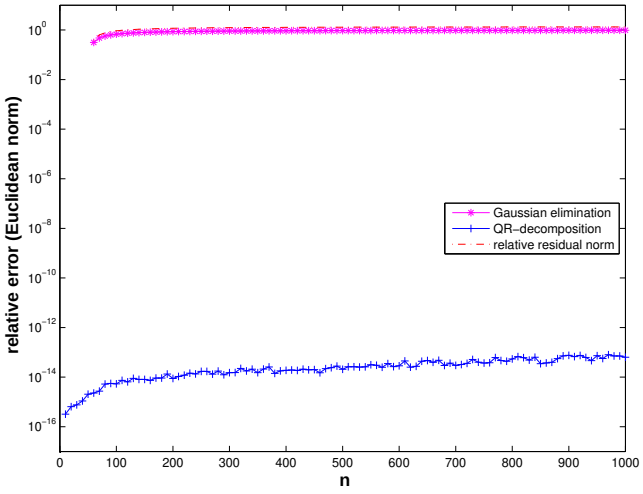


Fig. 85

⌟

Let us summarize the pros and cons of orthogonal transformation techniques for linear least squares problems:

Normal equations vs. orthogonal transformations method

Superior numerical stability (→ Def. 1.5.5.19) of orthogonal transformations methods:

- Use orthogonal transformations methods for least squares problems (3.1.3.7), whenever $\mathbf{A} \in \mathbb{R}^{m,n}$ *dense* and n small.

SVD/QR-factorization cannot exploit sparsity:

- Use normal equations in the expanded form (3.2.0.8)/(3.2.0.9), when $\mathbf{A} \in \mathbb{R}^{m,n}$ sparse (→ Notion 2.7.0.1) and m, n big.

Review question(s) 3.3.4.8 (QR-Based Solver for Linear Least-Squares Problems)

(Q3.3.4.8.A) Given $\mathbf{A} \in \mathbb{R}^{m,n}$, $m > n$, $\operatorname{rank}(\mathbf{A}) = n$, $\mathbf{b} \in \mathbb{R}^m$, let the full QR-decomposition

$$[\mathbf{A}\ \mathbf{b}] = \widetilde{\mathbf{Q}}\widetilde{\mathbf{R}}\ , \quad \begin{array}{l} \widetilde{\mathbf{Q}} \in \mathbb{R}^{m,m} \text{ orthogonal}\ , \\ \widetilde{\mathbf{R}} \in \mathbb{R}^{m,n+1} \text{ upper triangular}\ , \end{array}$$

of the augmented matrix $[\mathbf{A}\ \mathbf{b}] \in \mathbb{R}^{m,n+1}$ be given.

- How can you compute the unique least-squares solution $\mathbf{x}^* \in \mathbb{R}^n$ of $\mathbf{Ax} = \mathbf{b}$ using $\widetilde{\mathbf{Q}}$ and $\widetilde{\mathbf{R}}$?
- Explain why $\|\mathbf{Ax}^* - \mathbf{b}\|_2 = (\mathbf{R})_{n+1,n+1}$.

(Q3.3.4.8.B) Describe how the QR-decomposition of a *tridiagonal* matrix $\mathbf{A} \in \mathbb{R}^{n,n}$ can be computed with an asymptotic effort of $O(n)$ for $n \to \infty$.

△

3.3.5 Modification Techniques for QR-Decomposition

Video tutorial for Section 3.3.5 "Modification Techniques for QR-Decomposition": (25 minutes)
Download link, tablet notes

→ review questions 3.3.5.7

In § 2.6.0.12 we faced the task of solving a square linear system of equations $\widetilde{\mathbf{A}}\mathbf{x} = \mathbf{b}$ efficiently, whose coefficient matrix $\widetilde{\mathbf{A}}$ was a (rank-1) perturbation of $\mathbf{A}$, for which an LU-decomposition was available. Lemma 2.6.0.21 showed a way to reuse the information contained in the LU-decomposition.

A similar task can be posed for the QR-decomposition: Assume that a QR-decomposition (→ Thm. 3.3.3.4) of a matrix $\mathbf{A} \in \mathbb{R}^{m,n}$, $m \geq n$, has already been computed. However, now we have to solve a full-rank linear least squares problem $\left\|\widetilde{\mathbf{A}}\mathbf{x} - \mathbf{b}\right\|_2 \to \min$ with $\widetilde{\mathbf{A}} \in \mathbb{R}^{m,n}$, which is a "slight" perturbation of $\mathbf{A}$. If we aim to use orthogonalization techniques it would be desirable to compute a QR-decomposition of $\widetilde{\mathbf{A}}$ with recourse to the QR-decomposition of $\mathbf{A}$.

Remark 3.3.5.1 (Economical vs. full QR-decomposition) We remind of § 3.3.3.37: The precise type of QR-decomposition, whether full or economical, does not matter, since all algorithms will store the Q-factors as products of orthogonal transformations.

Thus, below we will select that type of QR-decomposition, which allows an easier derivation of an algorithm, which will be the full QR-decomposition. ⌟

3.3.5.1 Rank-1 Modifications

For $\mathbf{A} \in \mathbb{R}^{m,n}$, $m \geq n$, $\operatorname{rank}(\mathbf{A}) = n$, we consider the rank-1 modification, *cf.* (2.6.0.16),

$$\mathbf{A} \quad \longrightarrow \quad \widetilde{\mathbf{A}} := \mathbf{A} + \mathbf{u}\mathbf{v}^\top , \quad \mathbf{u} \in \mathbb{R}^m , \quad \mathbf{v} \in \mathbb{R}^n . \tag{3.3.5.2}$$

Remember from § 2.6.0.12, (2.6.0.13), (2.6.0.15) that changing a single entry, row, or column of $\mathbf{A}$ can be achieved through special rank-1 perturbations.

Given a full QR-decomposition according to Thm. 3.3.3.4, $\mathbf{A} = \mathbf{QR} = \mathbf{Q}\begin{bmatrix}\mathbf{R}_0\\ \mathbf{O}\end{bmatrix}$, $\mathbf{Q} \in \mathbb{R}^{m,m}$ orthogonal (stored in some implicit format as *product of orthogonal transformations*, see Rem. 3.3.3.22), $\mathbf{R} \in \mathbb{R}^{m,n}$ and $\mathbf{R}_0 \in \mathbb{R}^{n,n}$ upper triangular, the goal is to find an efficient algorithm that yields a QR-decomposition of $\widetilde{\mathbf{A}}$: $\widetilde{\mathbf{A}} = \widetilde{\mathbf{Q}}\widetilde{\mathbf{R}}$, $\widetilde{\mathbf{Q}} \in \mathbb{R}^{m,m}$ a product of orthogonal transformations, $\widetilde{\mathbf{R}} \in \mathbb{R}^{m,n}$ upper triangular.

Step ❶: compute $\mathbf{w} = \mathbf{Q}^\top\mathbf{u} \in \mathbb{R}^m$.

Observe that $\mathbf{A} + \mathbf{u}\mathbf{v}^\top = \mathbf{Q}(\mathbf{R} + \mathbf{w}\mathbf{v}^\top)$, because $\mathbf{Q}^\top\mathbf{Q} = \mathbf{I}_m$.

➢ Computational effort = $O(mn)$, if $\mathbf{Q}$ stored in suitable (compressed) format, *cf.* Rem. 3.3.3.22.

Step ❷: Orthogonally transform $\mathbf{w} \to \|\mathbf{w}\|\mathbf{e}_1$, $\mathbf{e}_1 \in \mathbb{R}^m \hat{=}$ 1st coordinate vector.

This can be done by applying $m-1$ Givens rotations to be employed in the following order:

$$\mathbf{w} = \begin{bmatrix} * \\ * \\ \vdots \\ * \\ * \\ * \\ * \end{bmatrix} \xrightarrow{\mathbf{G}_{m-1,m}} \begin{bmatrix} * \\ * \\ \vdots \\ * \\ * \\ * \\ 0 \end{bmatrix} \xrightarrow{\mathbf{G}_{m-2,m-1}} \begin{bmatrix} * \\ * \\ \vdots \\ * \\ * \\ 0 \\ 0 \end{bmatrix} \xrightarrow{\mathbf{G}_{m-3,m-2}} \cdots \xrightarrow{\mathbf{G}_{12}} \begin{bmatrix} * \\ 0 \\ \vdots \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Of course, these transformations also have to act on $\mathbf{R} \in \mathbb{R}^{m,n}$ and they will affect $\mathbf{R}$ by creating a single non-zero subdiagonal by linearly combining pairs of adjacent rows from bottom to top:

$$\mathbf{R} = \begin{bmatrix} * & * & \cdots & * & * & * & * \\ 0 & * & \cdots & * & * & * & * \\ \vdots & & \ddots & & & & \vdots \\ 0 & \cdots & 0 & * & * & * & * \\ 0 & \cdots & 0 & 0 & * & * & * \\ 0 & \cdots & 0 & 0 & 0 & * & * \\ 0 & \cdots & 0 & 0 & 0 & 0 & * \\ 0 & \cdots & 0 & 0 & 0 & 0 & 0 \\ 0 & \cdots & 0 & 0 & 0 & 0 & 0 \\ \vdots & & & & & & \vdots \\ 0 & \cdots & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \xrightarrow{\mathbf{G}_{n,n+1}} \begin{bmatrix} * & * & \cdots & * & * & * & * \\ 0 & * & \cdots & * & * & * & * \\ \vdots & \ddots & & & & \vdots & \\ 0 & \cdots & 0 & * & * & * & * \\ 0 & \cdots & 0 & 0 & * & * & * \\ 0 & \cdots & 0 & 0 & 0 & * & * \\ 0 & \cdots & 0 & 0 & 0 & 0 & * \\ 0 & \cdots & 0 & 0 & 0 & 0 & * \\ 0 & \cdots & 0 & 0 & 0 & 0 & 0 \\ \vdots & & & & & & \vdots \\ 0 & \cdots & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \xrightarrow{\mathbf{G}_{n-1,n}}$$

$$\longrightarrow \begin{bmatrix} * & * & \cdots & * & * & * & * \\ 0 & * & \cdots & * & * & * & * \\ \vdots & & \ddots & & & & \vdots \\ 0 & \cdots & 0 & * & * & * & * \\ 0 & \cdots & 0 & 0 & * & * & * \\ 0 & \cdots & 0 & 0 & 0 & * & * \\ 0 & \cdots & 0 & 0 & 0 & * & * \\ 0 & \cdots & 0 & 0 & 0 & 0 & * \\ 0 & \cdots & 0 & 0 & 0 & 0 & 0 \\ \vdots & & & & & & \vdots \\ 0 & \cdots & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \xrightarrow{\mathbf{G}_{n-2,n-1}} \cdots \xrightarrow{\mathbf{G}_{1,2}} \begin{bmatrix} * & * & \cdots & * & * & * & * \\ * & * & \cdots & * & * & * & * \\ & & \ddots & & & & \vdots \\ 0 & \cdots & * & * & * & * & * \\ 0 & \cdots & 0 & * & * & * & * \\ 0 & \cdots & 0 & 0 & * & * & * \\ 0 & \cdots & 0 & 0 & 0 & * & * \\ 0 & \cdots & 0 & 0 & 0 & 0 & * \\ 0 & \cdots & 0 & 0 & 0 & 0 & 0 \\ \vdots & & & & & & \vdots \\ 0 & \cdots & 0 & 0 & 0 & 0 & 0 \end{bmatrix} =: \mathbf{R}_1 .$$

We see $(\mathbf{R}_1)_{i,j} = 0$, if $i > j+1$. It is a so-called upper Hessenberg matrix. This is also true of $\mathbf{R}_1 + \|\mathbf{w}\|_2 \mathbf{e}_1 \mathbf{v}^\top \in \mathbb{R}^{n,n}$, because only the top row of the matrix $\mathbf{e}_1 \mathbf{v}^\top$ is non-zero. Therefore, if $\mathbf{Q}_1 \in \mathbb{R}^{n,n}$ collects all $m-1$ orthogonal transformations used in Step ❷, then

$$\mathbf{A} + \mathbf{u}\mathbf{v}^\top = \mathbf{Q}\mathbf{Q}_1^\top \big(\underbrace{\mathbf{R}_1 + \|\mathbf{w}\|_2 \mathbf{e}_1 \mathbf{v}^\top}_{\text{upper Hessenberg matrix}} \big) \quad \text{with orthogonal } \mathbf{Q}_1 := \mathbf{G}_{12} \cdot \cdots \cdot \mathbf{G}_{n-1,n} .$$

➢ Computational effort = $O(n + n^2) = O(n^2)$ for $n \to \infty$

Step ❸: Convert $\mathbf{R}_1 + \|\mathbf{w}\|_2 \mathbf{e}_1 \mathbf{v}^\top \in \mathbb{R}^{n,n}$ into upper triangular form by $n-1$ successive Givens rotations

applied to rows of this matrix.

$$\mathbf{R}_1 + \|\mathbf{w}\|_2 \mathbf{e}_1 \mathbf{v}^\top = \begin{bmatrix} * & * & \cdots & * & * & * & * \\ * & * & \cdots & * & * & * & * \\ & & \ddots & & & & \\ 0 & \cdots & * & * & * & * & * \\ 0 & \cdots & 0 & * & * & * & * \\ 0 & \cdots & 0 & 0 & * & * & * \\ 0 & \cdots & 0 & 0 & 0 & * & * \\ 0 & \cdots & 0 & 0 & 0 & 0 & * \\ 0 & \cdots & 0 & 0 & 0 & 0 & 0 \\ \vdots & & & & & & \vdots \\ 0 & \cdots & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \xrightarrow{\mathbf{G}_{12}} \begin{bmatrix} * & * & \cdots & * & * & * & * \\ 0 & * & \cdots & * & * & * & * \\ & & \ddots & & & & \\ 0 & \cdots & * & * & * & * & * \\ 0 & \cdots & 0 & * & * & * & * \\ 0 & \cdots & 0 & 0 & * & * & * \\ 0 & \cdots & 0 & 0 & 0 & * & * \\ 0 & \cdots & 0 & 0 & 0 & 0 & * \\ 0 & \cdots & 0 & 0 & 0 & 0 & 0 \\ \vdots & & & & & & \vdots \\ 0 & \cdots & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \xrightarrow{\mathbf{G}_{23}} \cdots$$

$$\xrightarrow{\mathbf{G}_{n-1,n}} \begin{bmatrix} * & * & \cdots & * & * & * & * \\ 0 & * & \cdots & * & * & * & * \\ & & \ddots & & & & \\ 0 & \cdots & 0 & * & * & * & * \\ 0 & \cdots & 0 & 0 & * & * & * \\ 0 & \cdots & 0 & 0 & 0 & * & * \\ 0 & \cdots & 0 & 0 & 0 & 0 & * \\ 0 & \cdots & 0 & 0 & 0 & 0 & * \\ 0 & \cdots & 0 & 0 & 0 & 0 & 0 \\ \vdots & & & & & & \vdots \\ 0 & \cdots & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \xrightarrow{\mathbf{G}_{n,n+1}} \begin{bmatrix} * & * & \cdots & * & * & * & * \\ 0 & * & \cdots & * & * & * & * \\ & & \ddots & & & & \\ 0 & \cdots & 0 & * & * & * & * \\ 0 & \cdots & 0 & 0 & * & * & * \\ 0 & \cdots & 0 & 0 & 0 & * & * \\ 0 & \cdots & 0 & 0 & 0 & 0 & * \\ 0 & \cdots & 0 & 0 & 0 & 0 & 0 \\ 0 & \cdots & 0 & 0 & 0 & 0 & 0 \\ \vdots & & & & & & \vdots \\ 0 & \cdots & 0 & 0 & 0 & 0 & 0 \end{bmatrix} =: \widetilde{\mathbf{R}} \,.$$

$$(\mathbf{G}_{n,n+1}\mathbf{G}_{n-2,n-1} \cdots \cdots \cdots \mathbf{G}_{23}\mathbf{G}_{12})(\mathbf{R}_1 + \|\mathbf{w}\|_2 \mathbf{e}_1 \mathbf{v}^\top) = \widetilde{\mathbf{R}} \quad \text{(upper triangular!)} \,. \tag{3.3.5.3}$$

Since we need $n-1$ Givens rotations acting on matrix rows of length n:

➢ Computational effort = $O(n^2)$ for $n \to \infty$

► $\widetilde{\mathbf{A}} = \mathbf{A} + \mathbf{u}\mathbf{v}^\top = \widetilde{\mathbf{Q}}\widetilde{\mathbf{R}}$ with $\widetilde{\mathbf{Q}} = \mathbf{Q}\mathbf{Q}_1^\top \mathbf{G}_{12}^\top \mathbf{G}_{23}^\top \cdot \cdots \cdot \mathbf{G}_{n-1,n-2}^\top \mathbf{G}_{n,n-1}^\top$.

➢ Total asymptotic total computational effort = $O(mn + n^2)$ for $m, n \to \infty$

For large n this is much cheaper than the cost $O(n^2 m)$ for computing the QR-decomposition of $\widetilde{\mathbf{A}}$ from scratch. Moreover, we avoid forming and storing the matrix $\widetilde{\mathbf{A}}$.

3.3.5.2 Adding a Column

We obtain an augmented matrix $\widetilde{\mathbf{A}} \in \mathbb{R}^{m,n+1}$ by inserting a column $\mathbf{v}$ into $\mathbf{A} \in \mathbb{R}^{m,n}$ at an arbitrary location.

$$\mathbf{A} \in \mathbb{R}^{m,n} \quad \longrightarrow \quad \widetilde{\mathbf{A}} = [\mathbf{a}_1, \ldots, \mathbf{a}_{k-1}, \mathbf{v}, \mathbf{a}_k, \ldots, \mathbf{a}_n] \in \mathbb{R}^{m,n+1} \,, \quad \mathbf{v} \in \mathbb{R}^m, \; \mathbf{a}_j := (\mathbf{A})_{:,j} \,. \tag{3.3.5.4}$$

On the level of matrix-vector arithmetic the following explanations are easier for the full QR-decompositions, *cf.* Rem. 3.3.5.1.

Given: full QR-decomposition of $\mathbf{A}$: $\mathbf{A} = \mathbf{Q}\mathbf{R}$, $\mathbf{Q} \in \mathbb{R}^{m,m}$ orthogonal (stored as product of $O(n)$ orthogonal transformations), $\mathbf{R} \in \mathbb{R}^{m,n}$ upper triangular.

Sought: full QR-decomposition of $\widetilde{\mathbf{A}}$ from (3.3.5.4): $\widetilde{\mathbf{A}} = \widetilde{\mathbf{Q}}\widetilde{\mathbf{R}}$, $\widetilde{\mathbf{Q}} \in \mathbb{R}^{m,m}$ with orthonormal columns, stored as product of $O(n)$ orthogonal transformations, $\widetilde{\mathbf{R}} \in \mathbb{R}^{m,n+1}$ upper triangular, computed efficiently.

As preparation we point out that left-multiplication of a matrix with another matrix can be understood as forming multiple matrix×vector products:

$$\mathbf{A} = \mathbf{Q}\mathbf{R} \quad \Leftrightarrow \quad \mathbf{Q}^\top \mathbf{A} = \left[\mathbf{Q}^\top \mathbf{a}_1, \ldots, \mathbf{Q}^\top \mathbf{a}_n\right] = \mathbf{R} =: \left[\begin{array}{c} \mathbf{R}_0 \\ \hline \mathbf{O} \end{array}\right], \quad \mathbf{R}_0 \in \mathbb{R}^{n,n},$$

that is, $\left(\mathbf{Q}^\top \mathbf{a}_j\right)_\ell = 0$ for $\ell > j$. We immediately infer

$$\mathbf{Q}^\top \widetilde{\mathbf{A}} = \left[\mathbf{Q}^\top \mathbf{a}_1, \ldots, \mathbf{Q}^\top \mathbf{a}_{k-1}, \mathbf{Q}^\top \mathbf{v}, \mathbf{Q}^\top \mathbf{a}_k, \ldots, \mathbf{Q}^\top \mathbf{a}_n\right] = \left[\begin{array}{ccccc} \ast & \cdots & \ast & \ast & \ast \\ & \ddots & \vdots & \vdots & \vdots \\ & & \ast & \ast & \ast \\ \hline & & & \ast & \ast \\ & & & \ast & \ast \end{array}\right] =: \mathbf{W} \in \mathbb{R}^{m,n+1},$$

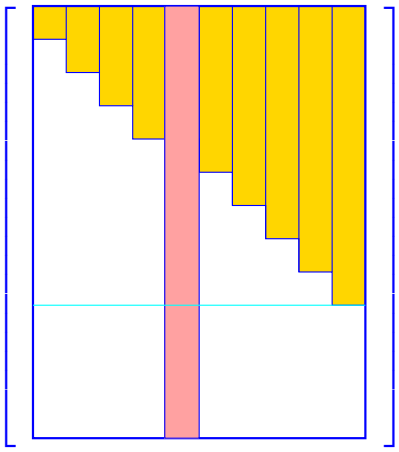
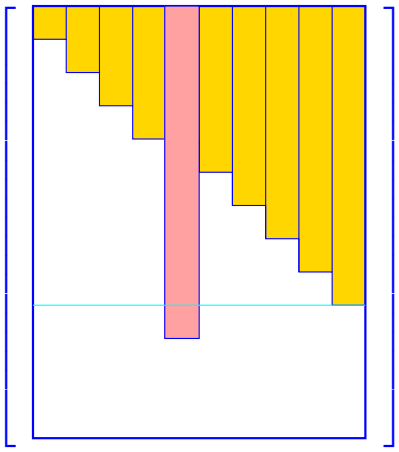
which suggests the following three-step algorithm.

Step ❶: compute $\mathbf{w} = \mathbf{Q}^\top \mathbf{v} \in \mathbb{R}^m$.

➢ Computational effort = $O(mn)$ for $m, n \to \infty$, if $\mathbf{Q}$ stored in suitable (compressed) format.

Step ❷: Annihilate bottom $m-n-1$ components of $\mathbf{w}$ $\leftrightarrow$ rows of $\mathbf{W}$ (, if $m > n+1$).

This can be done by $m-n-1$ Givens rotations targeting adjacent rows of $\mathbf{W}$ bottom $\to$ top:

$\mathbf{W} =$  $\xrightarrow{\mathbf{G}_{m-1,m}}$ $\cdots$ $\xrightarrow{\mathbf{G}_{n+1,n+2}}$  $=: \mathbf{T}$.

➢ Computational effort = $O(m-n)$ for $m, n \to \infty$.

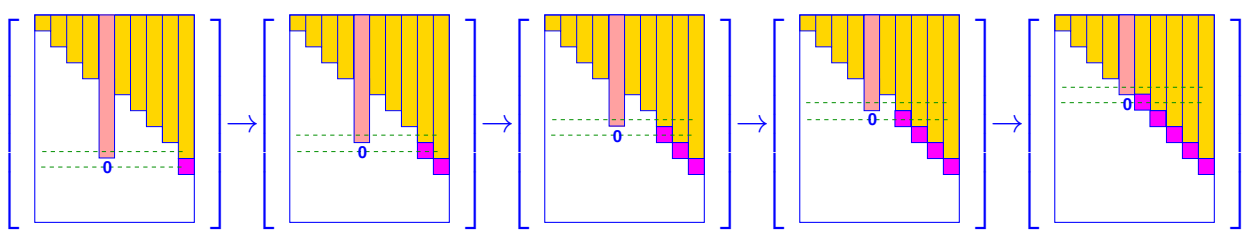
Writing $\mathbf{Q}_2^\top := \mathbf{G}_{n+1,n+2} \cdot \cdots \cdot \mathbf{G}_{m-1,m} \in \mathbb{R}^{m,m}$ for the orthogonal matrix representing the product of Givens rotations, we find

$$\mathbf{Q}_2^\top \mathbf{Q}^\top \widetilde{\mathbf{A}} = \mathbf{T} .$$

Step ❸: Transform $\mathbf{T}$ to upper triangular form.

We accomplish this by applying $n+1-k$ successive Givens rotations from bottom to top in the following fashion.

$\mathbf{T} =$ $\begin{bmatrix} * & \cdots & * & \cdots & * \\ 0 & * & * & & \vdots \\ \vdots & \ddots & \vdots & & \\ & & * & \ddots & \vdots \\ & & * & * & * \\ & & * & & * \\ \vdots & & * & & 0 \\ 0 & \cdots & 0 & \cdots & \vdots \\ \vdots & & \vdots & & \vdots \\ 0 & \cdots & 0 & \cdots & 0 \end{bmatrix}$ $\xrightarrow{\mathbf{G}_{n,n+1}}$ $\cdots$ $\xrightarrow{\mathbf{G}_{k,k+1}}$ $\begin{bmatrix} * & \cdots & * & \cdots & * \\ 0 & * & * & & \vdots \\ \vdots & \ddots & \vdots & & \\ & & * & & \\ & & 0 & \ddots & \\ & & 0 & * & \vdots \\ \vdots & & 0 & & * \\ 0 & \cdots & 0 & \cdots & 0 \\ \vdots & & \vdots & & \\ 0 & \cdots & 0 & \cdots & 0 \end{bmatrix}$



▬ ▬ ▬ ▪ $\hat{=}$ target rows of Givens rotations, ■ $\hat{=}$ new entries $\neq 0$

➢ Computational effort for this step = $O\big((n-k)^2\big)$ for $n \to \infty$

➢ Total asymptotic computational cost = $O(n^2+m)$ for $m,n \to \infty$

Again, for large m,n this is significantly cheaper that forming the matrix $\widetilde{\mathbf{A}}$ and then computing its (economical) QR-decomposition.

3.3.5.3 Adding a Row

Again, the perspective of the full QR-decomposition is preferred for didactic reasons, *cf.* Rem. 3.3.5.1.

We are given a matrix $\mathbf{A} \in \mathbb{R}^{m,n}$ of which a full QR-decomposition ($\to$ Thm. 3.3.3.4) $\mathbf{A} = \mathbf{QR}$, $\mathbf{Q} \in \mathbb{R}^{m,m}$ orthogonal, $\mathbf{R} \in \mathbb{R}^{m,n}$ upper triangular, is already available, maybe only in encoded form ($\to$ Rem. 3.3.3.22).

We add another row to the matrix $\mathbf{A}$ in arbitrary position $k \in \{1,\ldots,m\}$ and obtain

$$\mathbf{A} \in \mathbb{R}^{m,n} \quad \mapsto \quad \widetilde{\mathbf{A}} = \begin{bmatrix} (\mathbf{A})_{1,:} \\ \vdots \\ (\mathbf{A})_{k-1,:} \\ \mathbf{v}^T \\ (\mathbf{A})_{k,:} \\ \vdots \\ (\mathbf{A})_{m,:} \end{bmatrix}, \quad \text{with given} \quad \mathbf{v} \in \mathbb{R}^n \,. \tag{3.3.5.5}$$

Task: Find an algorithm for the *efficient* computation of the QR-decomposition $\widetilde{\mathbf{A}} = \widetilde{\mathbf{Q}}\widetilde{\mathbf{R}}$ of $\widetilde{\mathbf{A}}$ from (3.3.5.5) , $\widetilde{\mathbf{Q}} \in \mathbb{R}^{m+1,m+1}$ orthogonal (as a product of orthogonal transformations), $\widetilde{\mathbf{R}} \in \mathbb{K}^{m+1,n+1}$ upper triangular.

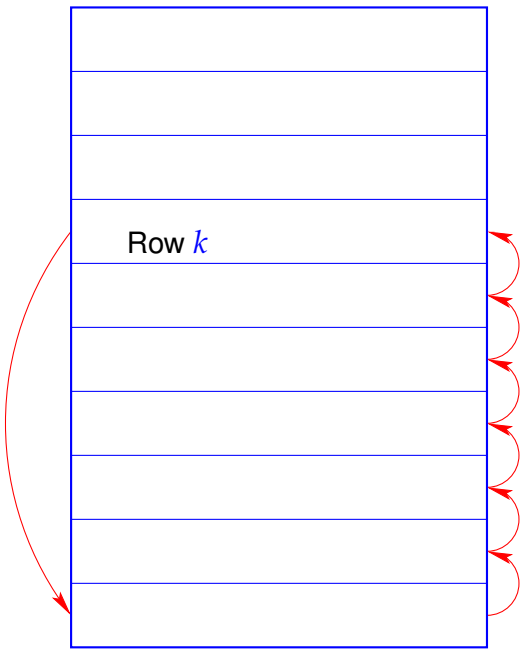
Step ①: Move new row to the bottom.

Employ partial **cyclic permutation** of rows of $\widetilde{\mathbf{A}}$:

row $m+1 \leftarrow$ row k , row $i \leftarrow$ row $i+1$, $i = k,\ldots,m$.

which can be achieved by multiplying from the right with the following (orthogonal!) **permutation matrix** $\mathbf{P} \in \mathbb{R}^{m+1,m+1}$

$$\mathbf{P}^\top := \begin{bmatrix} 1 & 0 & \cdots & & & \cdots & 0 \\ 0 & \ddots & & & & & \vdots \\ & & 1 & 0 & & & 0 \\ \vdots & & & 0 & 0 & & 1 \\ \vdots & & & 1 & 0 & \ddots & 0 \\ & & & & \ddots & \ddots & \vdots \\ 0 & \cdots & & & & 1 & 0 \end{bmatrix} \in \mathbb{R}^{m+1,m+1} \,.$$



$$\mathbf{P}\widetilde{\mathbf{A}} = \begin{bmatrix}\mathbf{A}\\ \mathbf{v}^\top\end{bmatrix} \quad \blacktriangleright \quad \begin{bmatrix}\mathbf{Q}^\top & 0\\ 0 & 1\end{bmatrix}\mathbf{P}\widetilde{\mathbf{A}} = \begin{bmatrix}\mathbf{R}\\ \mathbf{v}^\top\end{bmatrix} = \begin{bmatrix} \mathbf{R} \\ \\ \\ \\ \mathbf{v}^\top \end{bmatrix} =: \mathbf{T} \in \mathbb{R}^{m+1,n} .$$

This step is a mere bookkeeping operation and does not involve any computations.

Step ②: Restore upper triangular form through Givens rotations (→ § 3.3.3.15)

Successively target bottom row and rows from the top to turn leftmost entries of bottom row into zeros. Here demonstrated for the case $m = n$

$$\mathbf{T} = \begin{bmatrix} * & \cdots & & & \cdots & * \\ 0 & * & & & & \vdots \\ \vdots & 0 & \ddots & & & \\ \vdots & \vdots & & * & & \vdots \\ 0 & 0 & & 0 & * & * \\ 0 & 0 & & 0 & 0 & * \\ * & \cdots & \cdots & * & * & * \end{bmatrix} \xrightarrow{\mathbf{G}_{1,m+1}} \begin{bmatrix} * & \cdots & & & \cdots & * \\ 0 & * & & & & \vdots \\ \vdots & 0 & \ddots & & & \\ \vdots & \vdots & & * & & \vdots \\ 0 & 0 & & 0 & * & * \\ 0 & 0 & & 0 & 0 & * \\ 0 & * & \cdots & * & * & * \end{bmatrix} \xrightarrow{\mathbf{G}_{2,m+1}} \cdots$$

$$\cdots \xrightarrow{\mathbf{G}_{m-1,m+1}} \begin{bmatrix} * & \cdots & & & \cdots & * \\ 0 & * & & & & \vdots \\ \vdots & 0 & \ddots & & & \\ \vdots & \vdots & & * & & \vdots \\ 0 & 0 & & 0 & * & * \\ 0 & 0 & & 0 & 0 & * \\ 0 & \cdots & & 0 & 0 & * \end{bmatrix} \xrightarrow{\mathbf{G}_{m,m+1}} \begin{bmatrix} * & \cdots & & & \cdots & * \\ 0 & * & & & & \vdots \\ \vdots & 0 & \ddots & & & \\ \vdots & \vdots & & * & & \vdots \\ 0 & 0 & & 0 & * & * \\ 0 & 0 & & 0 & 0 & * \\ 0 & \cdots & & & 0 & 0 \end{bmatrix} := \widetilde{\mathbf{R}} \quad (3.3.5.6)$$

➢ Computational effort for this step = $O(n^2)$ for $n \to \infty$

Finally, setting $\mathbf{Q}_1 = \mathbf{G}_{m,m+1} \cdot \cdots \cdot \mathbf{G}_{1,m+1}$ the final QR-decomposition reads

$$\widetilde{\mathbf{A}} = \mathbf{P}^T \begin{pmatrix}\mathbf{Q} & 0\\ 0 & 1\end{pmatrix} \mathbf{Q}_1^\top \widetilde{\mathbf{R}} = \widetilde{\mathbf{Q}}\widetilde{\mathbf{R}} \quad \text{with orthogonal } \widetilde{\mathbf{Q}} \in \mathbb{K}^{m+1,m+1} ,$$

because the product of orthogonal matrices is again orthogonal. Of course, we know that $\widetilde{Q}$ is never formed in an algorithm but kept as a sequence of orthogonal transformations.

Review question(s) 3.3.5.7 (Modification teachniques for QR-decompositions)

(Q3.3.5.7.A) Explain why, as far as the use of the QR-decomposition in numerical methods is concerned, the distinction between full and economical versions does not matter.

(Q3.3.5.7.B) [QR-update after dropping a column] Assume that $\widetilde{\mathbf{A}}$ arises from $\mathbf{A} \in \mathbb{R}^{m,n}$, $m \geq n$, by dropping the k-th column, $k \in \{1,\ldots,n\}$. How can a QR-decomposition of $\widetilde{\mathbf{A}}$ be computed based on a QR-decomposition $\mathbf{A} = \mathbf{QR}$ of $\mathbf{A}$?

Hint. Examine the structure of $\mathbf{Q}\widetilde{\mathbf{A}}$.

(Q3.3.5.7.C) [Update of QR-decomposition when modifying a single entry] Assume that the (full) QR-decomposition $\mathbf{A} = \mathbf{QR}$ of $\mathbf{A} \in \mathbb{R}^{m,n}$, $m \geq n$, is available. Describe the algorithm for computing the QR-decomposition of $\widetilde{\mathbf{A}} \in \mathbb{R}^{m,n}$ that arises from setting a single entry of $\mathbf{A}$ at position (ℓ,k), $\ell \in \{1,\ldots,m\}$, $k \in \{1,\ldots,n\}$ to zero.

Exception. You may look at the lecture notes to answer this question.

△

3.4 Singular Value Decomposition (SVD)

Beside the QR-decomposition of a matrix $\mathbf{A} \in \mathbb{R}^{m,n}$ there are other factorizations based on orthogonal transformations. The most important among them is the singular value decomposition (SVD), which can be used to tackle linear least squares problems and many other optimization problems beyond, see [Kal96].

3.4.1 SVD: Definition and Theory

Video tutorial for Section 3.4.1 "Singular Value Decomposition: Definition and Theory": (13 minutes) Download link, tablet notes

→ review questions 3.4.1.15

Theorem 3.4.1.1. Singular value decomposition → **[NS02, Thm. 9.6], [Gut09, Thm. 11.1]**

For any $\mathbf{A} \in \mathbb{K}^{m,n}$ there are unitary/orthogonal *matrices $\mathbf{U} \in \mathbb{K}^{m,m}$, $\mathbf{V} \in \mathbb{K}^{n,n}$ and a (generalized) diagonal $^{(*)}$ matrix $\mathbf{\Sigma} = \operatorname{diag}(\sigma_1,\ldots,\sigma_p) \in \mathbb{R}^{m,n}$, $p := \min\{m,n\}$, $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$ such that*

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{H}} .$$

Terminology $(*)$: A matrix $\mathbf{\Sigma}$ is called a generalized diagonal matrix, if $(\mathbf{\Sigma})_{i,j} = 0$, if $i \neq j$, $1 \leq i \leq m$, $1 \leq j \leq n$. We still use the diag operator to create it from a vector.

Proof. (of Thm. 3.4.1.1, by induction)

To start the induction note that the assertion of the theorem is immediate for $n = 1$ or $m = 1$.

For the induction step $(n-1,m-1) \Rightarrow (m,n)$ first remember from analysis [Str09, Thm. 4.2.3]: Continuous real-valued functions attain extremal values on compact sets (here the unit ball $\{\mathbf{x} \in \mathbb{K}^n : \|\mathbf{x}\|_2 \leq 1\}$). In particular, consider the function $\mathbf{v} \in \mathbb{K}^n \mapsto \|\mathbf{A}\mathbf{v}\| \in \mathbb{R}$. This function will attain its maximal value $\|\mathbf{A}\|_2$ on $\{\mathbf{v} \in \mathbb{K}^2 : \|\mathbf{v}\|_2 \leq 1\}$ for at least one vector $\mathbf{x}$, $\|\mathbf{x}\| = 1$:

$$\blacktriangleright \quad \exists \mathbf{x} \in \mathbb{K}^n,\ \mathbf{y} \in \mathbb{K}^m ,\quad \|\mathbf{x}\| = \|\mathbf{y}\|_2 = 1 : \quad \mathbf{A}\mathbf{x} = \sigma\mathbf{y} ,\quad \sigma = \|\mathbf{A}\|_2 ,$$

where we used the definition of the matrix 2-norm, see Def. 1.5.5.10. By Gram-Schmidt orthogonalization or a similar procedure we can extend the single unit vectors $\mathbf{x}$ and $\mathbf{y}$ to orthonormal bases of $\mathbb{K}^n$ and $\mathbb{K}^m$, respectivelt: $\exists \widetilde{\mathbf{V}} \in \mathbb{K}^{n,n-1}$, $\widetilde{\mathbf{U}} \in \mathbb{K}^{m,m-1}$ such that

$$\mathbf{V} = [\mathbf{x}\ \widetilde{\mathbf{V}}] \in \mathbb{K}^{n,n} \quad , \quad \mathbf{U} = [\mathbf{y}\ \widetilde{\mathbf{U}}] \in \mathbb{K}^{m,m} \quad \text{are orthogonal.}$$

$$\blacktriangleright \quad \mathbf{U}^{\mathrm{H}}\mathbf{A}\mathbf{V} = \begin{bmatrix} \mathbf{y} & \widetilde{\mathbf{U}} \end{bmatrix}^{\mathrm{H}} \mathbf{A} \begin{bmatrix} \mathbf{x} & \widetilde{\mathbf{V}} \end{bmatrix} = \left[\begin{array}{c|c} \mathbf{y}^{\mathrm{H}}\mathbf{A}\mathbf{x} & \mathbf{y}^{\mathrm{H}}\mathbf{A}\widetilde{\mathbf{V}} \\ \hline \widetilde{\mathbf{U}}^{\mathrm{H}}\mathbf{A}\mathbf{x} & \widetilde{\mathbf{U}}^{\mathrm{H}}\mathbf{A}\widetilde{\mathbf{V}} \end{array}\right] = \left[\begin{array}{c|c} \sigma & \mathbf{w}^{\mathrm{H}} \\ \hline \mathbf{0} & \mathbf{B} \end{array}\right] =: \mathbf{A}_1 \,.$$

For the induction argument we have to show that $\mathbf{w} = \mathbf{0}$. Since

$$\left\| \mathbf{A}_1 \begin{bmatrix} \sigma \\ \mathbf{w} \end{bmatrix} \right\|_2^2 = \left\| \begin{bmatrix} \sigma^2 + \mathbf{w}^{\mathrm{H}}\mathbf{w} \\ \mathbf{B}\mathbf{w} \end{bmatrix} \right\|_2^2 = (\sigma^2 + \mathbf{w}^{\mathrm{H}}\mathbf{w})^2 + \|\mathbf{B}\mathbf{w}\|_2^2 \geq (\sigma^2 + \mathbf{w}^{\mathrm{H}}\mathbf{w})^2 \,,$$

we conclude

$$\|\mathbf{A}_1\|_2^2 = \sup_{0 \neq \mathbf{x} \in \mathbb{K}^n} \frac{\|\mathbf{A}_1\mathbf{x}\|_2^2}{\|\mathbf{x}\|_2^2} \geq \frac{\left\|\mathbf{A}_1 \binom{\sigma}{\mathbf{w}}\right\|_2^2}{\left\|\binom{\sigma}{\mathbf{w}}\right\|_2^2} \geq \frac{(\sigma^2 + \mathbf{w}^{\mathrm{H}}\mathbf{w})^2}{\sigma^2 + \mathbf{w}^{\mathrm{H}}\mathbf{w}} = \sigma^2 + \mathbf{w}^{\mathrm{H}}\mathbf{w} \,. \tag{3.4.1.2}$$

We exploit that multiplication with orthogonal matrices either from right or left does not affect the Euclidean matrix norm:

$$\sigma^2 = \|\mathbf{A}\|_2^2 = \left\|\mathbf{U}^{\mathrm{H}}\mathbf{A}\mathbf{V}\right\|_2^2 = \|\mathbf{A}_1\|_2^2 \overset{(3.4.1.2)}{\Longrightarrow} \|\mathbf{A}_1\|_2^2 = \|\mathbf{A}_1\|_2^2 + \|\mathbf{w}\|_2^2 \quad \Rightarrow \quad \mathbf{w} = \mathbf{0} \,.$$

$$\blacktriangleright \quad \mathbf{A}_1 = \left[\begin{array}{c|c} \sigma & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{B} \end{array}\right] .$$

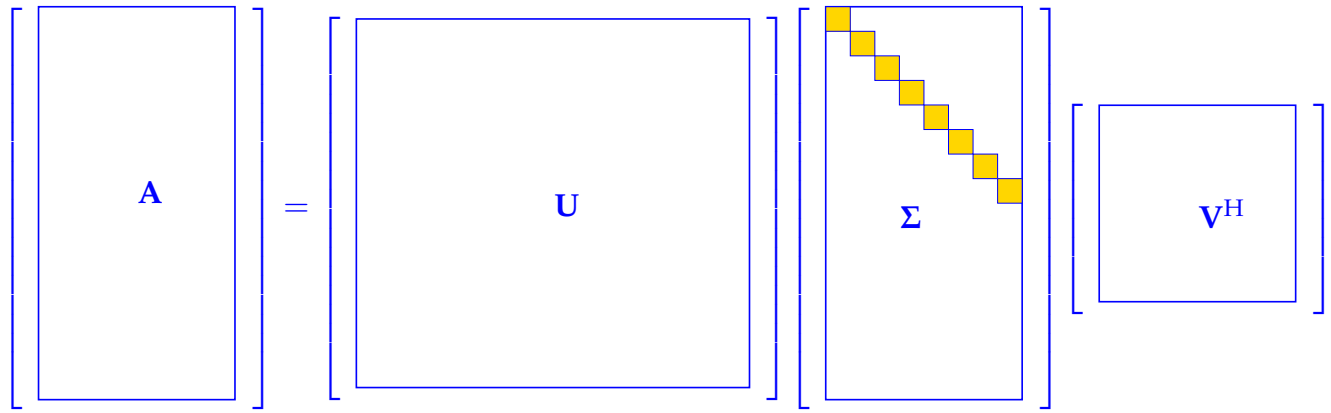
Then apply the induction argument to $\mathbf{B}$.

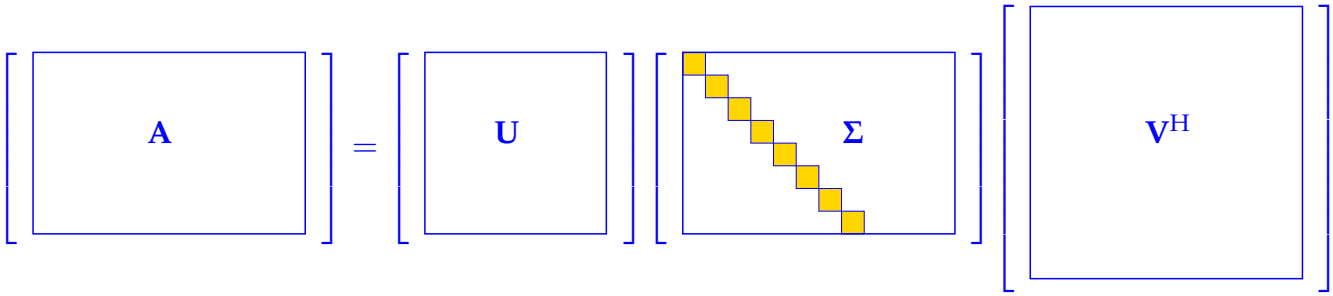
□

Definition 3.4.1.3. Singular value decomposition (SVD)

The decomposition $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{H}}$ of Thm. 3.4.1.1 is called singular value decomposition (SVD) of $\mathbf{A}$. The diagonal entries σ_i of $\mathbf{\Sigma}$ are the singular values of $\mathbf{A}$. The columns of $\mathbf{U}/\mathbf{V}$ are the left/right **singular vectors** of $\mathbf{A}$.

Next, we visualize the structure of the singular value decomposition of a matrix $\mathbf{A} = \mathbb{K}^{m,n}$.



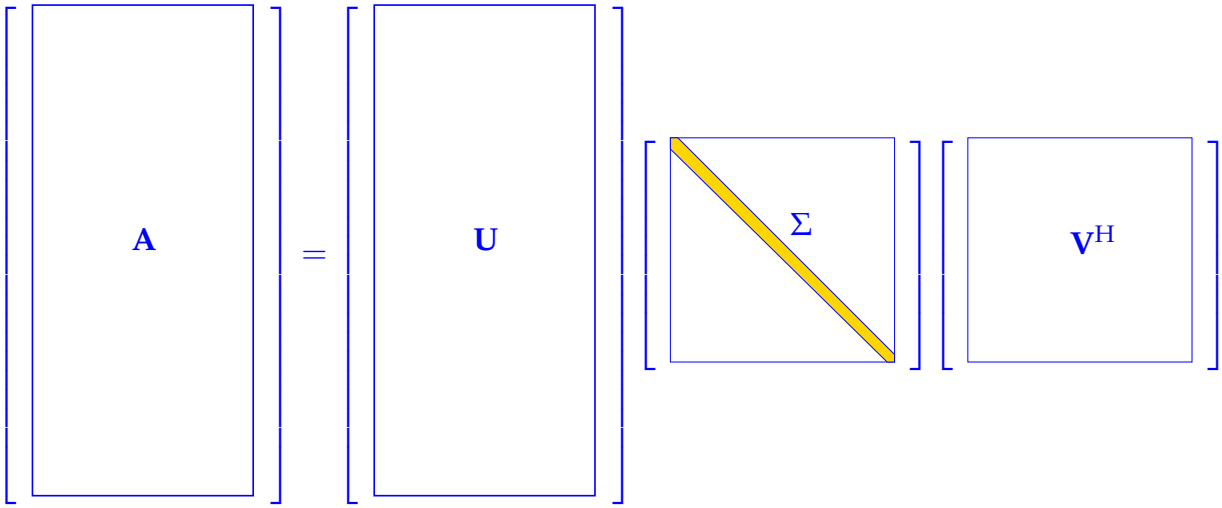


§3.4.1.4 (Economical singular value decomposition) As in the case of the QR-decomposition, compare (3.3.3.1) and (3.3.3.1), we can also drop the bottom zero rows of $\mathbf{\Sigma}$ and the corresponding columns of $\mathbf{U}$ in the case of $m > n$. Thus we end up with an "economical" singular value decomposition of $\mathbf{A} \in \mathbb{K}^{m,n}$:

$$\begin{aligned} m \geq n\text{:}\quad & \mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{H}}\,,\quad \mathbf{U} \in \mathbb{K}^{m,n}\,,\quad \mathbf{\Sigma} \in \mathbb{K}^{n,n}\,,\quad \mathbf{V} \in \mathbb{K}^{n,n}\quad , \quad \mathbf{U}^{\mathrm{H}}\mathbf{U} = \mathbf{I}_n \text{ ,}\mathbf{V} \text{ unitary ,} \\ & & & & & \text{(}\mathbf{U}\text{ orthonormal columns)} \\ m < n\text{:}\quad & \mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{H}}\,,\quad \mathbf{U} \in \mathbb{K}^{m,m}\,,\quad \mathbf{\Sigma} \in \mathbb{K}^{m,m}\,,\quad \mathbf{V} \in \mathbb{K}^{n,m}\quad , \quad \mathbf{U} \text{ unitary ,}\mathbf{V}^{\mathrm{H}}\mathbf{V} = \mathbf{I}_m\,. \\ & & & & & \text{(}\mathbf{V}\text{ orthonormal columns)} \end{aligned} \tag{3.4.1.5}$$

with true diagonal matrices $\mathbf{\Sigma}$, whose diagonals contain the singular values of $\mathbf{A}$.

Visualization of economimcal SVD for $m > 0$:



The economical SVD is also called **thin SVD** in literature [GV13, Sect. 2.3.4]. ⌟

An alternative motivation and derivation of the SVD is based on diagonalizing the Hermitian matrices $\mathbf{A}\mathbf{A}^{\mathrm{H}} \in \mathbb{R}^{m,m}$ and $\mathbf{A}^{\mathrm{H}}\mathbf{A} \in \mathbb{R}^{n,n}$. The relationship is made explicit in the next lemma.

Lemma 3.4.1.6.

The squares σ_i^2 of the non-zero singular values of $\mathbf{A}$ are the non-zero eigenvalues of $\mathbf{A}^{\mathrm{H}}\mathbf{A}$, $\mathbf{A}\mathbf{A}^{\mathrm{H}}$ with associated eigenvectors $(\mathbf{V})_{:,1}, \ldots, (\mathbf{V})_{:,p}$, $(\mathbf{U})_{:,1}, \ldots, (\mathbf{U})_{:,p}$, respectively.

Proof. $\mathbf{A}\mathbf{A}^{\mathrm{H}}$ and $\mathbf{A}^{\mathrm{H}}\mathbf{A}$ are similar ($\rightarrow$ Lemma 9.1.0.6) to diagonal matrices with non-zero diagonal entries σ_i^2 ($\sigma_i \neq 0$), e.g.,

$$\mathbf{A}\mathbf{A}^{\mathrm{H}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{H}}\mathbf{V}\mathbf{\Sigma}^{\mathrm{H}}\mathbf{U}^{\mathrm{H}} = \mathbf{U}\underbrace{\mathbf{\Sigma}\mathbf{\Sigma}^{\mathrm{H}}}_{\text{diagonal matrix}}\mathbf{U}^{\mathrm{H}}\,.$$

□

□

Remark 3.4.1.7 (SVD and additive rank-1 decomposition $\rightarrow$ [Gut09, Cor. 11.2], [NS02, Thm. 9.8]) Recall from linear algebra that rank-1 matrices coincide with tensor products of vectors:

$$\mathbf{A} \in \mathbb{K}^{m,n} \quad \text{and} \quad \operatorname{rank}(\mathbf{A}) = 1 \quad \Leftrightarrow \quad \exists \mathbf{u} \in \mathbb{K}^m, \mathbf{v} \in \mathbb{K}^n\text{:}\quad \mathbf{A} = \mathbf{u}\mathbf{v}^{\mathrm{H}}\,, \tag{3.4.1.8}$$

because $\operatorname{rank}(\mathbf{A}) = 1$ means that $\mathbf{Ax} = \mu(\mathbf{x})\mathbf{u}$ for some $\mathbf{u} \in \mathbb{K}^m$ and a linear form $\mathbf{x} \in \mathbb{K}^n \mapsto \mu(\mathbf{x}) \in \mathbb{K}$. By the Riesz representation theorem the latter can be written as $\mu(\mathbf{x}) = \mathbf{v}^{\mathrm{H}}\mathbf{x}$.

The singular value decomposition provides an additive decomposition into rank-1 matrices:

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{H}} = \sum_{j=1}^{p} \sigma_j \, (\mathbf{U})_{:,j}(\mathbf{V})_{:,j}^{\mathrm{H}} \,. \tag{3.4.1.9}$$

Since the columns of $\mathbf{U}$ and $\mathbf{V}$ are orthonormal, we immediately conclude:

$$\mathbf{A}(\mathbf{V})_{:,j} = \sigma_j(\mathbf{U})_j \quad , \quad \mathbf{A}^{\mathrm{H}}(\mathbf{U})_j = \sigma_j(\mathbf{V})_{:,j} \, , \quad j \in \{1, \ldots, p\} \, . \tag{3.4.1.10}$$

Remark 3.4.1.11 (Uniqueness of SVD)

The SVD from Def. 3.4.1.3 is not (necessarily) unique, but the singular values are.

Proof. Proof by contradiction: assume that $\mathbf{A}$ has two singular value decompositions

$$\mathbf{A} = \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^{\mathrm{H}} = \mathbf{U}_2\mathbf{\Sigma}_2\mathbf{V}_2^{\mathrm{H}} \;\Rightarrow\; \mathbf{U}_1 \underbrace{\mathbf{\Sigma}_1\mathbf{\Sigma}_1^{\mathrm{H}}}_{=\operatorname{diag}(\sigma_1^2,\ldots,\sigma_m^2)} \mathbf{U}_1^{\mathrm{H}} = \mathbf{A}\mathbf{A}^{\mathrm{H}} = \mathbf{U}_2 \underbrace{\mathbf{\Sigma}_2\mathbf{\Sigma}_2^{\mathrm{H}}}_{=\operatorname{diag}(\sigma_1^2,\ldots,\sigma_m^2)} \mathbf{U}_2^{\mathrm{H}} \, .$$

The two diagonal matrices are similar, which implies that they have the same eigenvalues, which agree with their diagonal entries. Since the latter are sorted, the diagonals must agree. □

§3.4.1.12 (SVD, nullspace, and image space) The SVD give complete information about all crucial subspaces associated with a matrix:

Lemma 3.4.1.13. SVD and rank of a matrix → [NS02, Cor. 9.7]

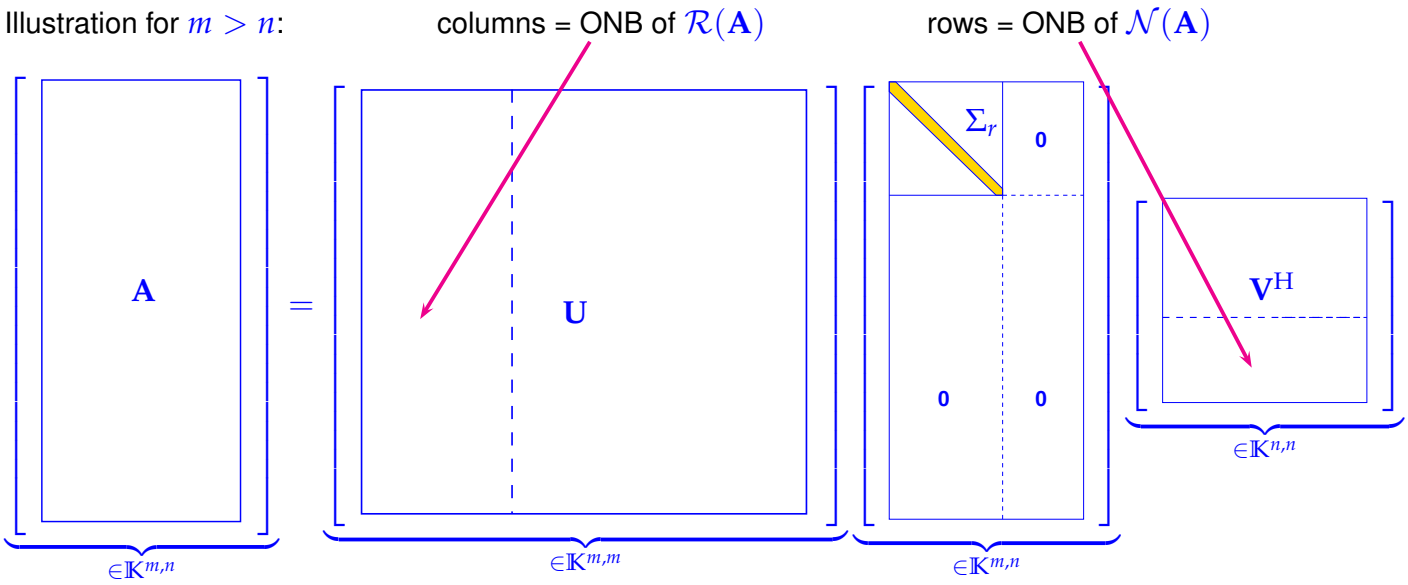
Let $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{H}}$ be the SVD of $\mathbf{A} \in \mathbb{K}^{m,n}$ according to Thm. 3.4.1.1. If, for some $1 \le r \le p := \min\{m,n\}$, the singular values of $\mathbf{A} \in \mathbb{K}^{m,n}$ satisfy

$$\sigma_1 \ge \cdots \ge \sigma_r > \sigma_{r+1} = \cdots \sigma_p = 0 \, ,$$

then

- $\operatorname{rank}(\mathbf{A}) = r$ *(no. of non-zero singular values)* ,
- $\mathcal{N}(\mathbf{A}) = \operatorname{Span}\{(\mathbf{V})_{:,r+1}, \ldots, (\mathbf{V})_{:,n}\}$,
- $\mathcal{R}(\mathbf{A}) = \operatorname{Span}\{(\mathbf{U})_{:,1}, \ldots, (\mathbf{U})_{:,r}\}$.

Illustration for $m > n$:



(3.4.1.14)

⌟

Review question(s) 3.4.1.15 (SVD: Definition and theory)

(Q3.4.1.15.A) If a square matrix $\mathbf{A} \in \mathbb{R}^{n,n}$ is given as $\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}^\top$ with an orthogonal matrix $\mathbf{Q} \in \mathbb{R}^{n,n}$ and a diagonal matrix $\mathbf{D} \in \mathbb{R}^{n,n}$, then what is a singular value decomposition of $\mathbf{A}$?

(Q3.4.1.15.B) What is a full singular value decomposition of $\mathbf{A} = \mathbf{u}\mathbf{v}^\top$, $\mathbf{u} \in \mathbb{R}^m$, $\mathbf{v} \in \mathbb{R}^n$?

(Q3.4.1.15.C) Based on the SVD give a proof of the fundamental dimension theorem from linear algebra:

$$\operatorname{rank}(\mathbf{A}) + \dim \mathcal{N}(\mathbf{A}) = n \quad \forall \mathbf{A} \in \mathbb{K}^{m,n} \; .$$

(Q3.4.1.15.D) Use the SVD of $\mathbf{A} \in \mathbb{K}^{m,n}$ to prove the fundamental relationships

$$\mathcal{R}(\mathbf{A}^{\mathrm{H}}) = \mathcal{N}(\mathbf{A})^\perp \quad , \quad \mathcal{N}(\mathbf{A}^{\mathrm{H}}) = \mathcal{R}(\mathbf{A})^\perp \; .$$

Here $X^\perp$ designates the orthogonal complement of a subspace $X \subset \mathbb{K}^d$ with respect to the Euclidean inner product:

$$X^\perp := \{\mathbf{v} \in \mathbb{K}^d : \; \mathbf{x}^{\mathrm{H}}\mathbf{v} = 0 \; \forall \mathbf{x} \in X\} \; .$$

(Q3.4.1.15.E) Use the SVD to show that every regular square matrix $\mathbf{A} \in \mathbb{R}^{n,n}$ can be factorized as

$$\mathbf{A} = \mathbf{Q}\mathbf{S} \; , \quad \mathbf{Q} \textit{ orthogonal} \; , \quad \mathbf{S} \textit{ symmetric, positive definite} \; ,$$

which is the so-called **polar decomposition** of $\mathbf{A}$.

(Q3.4.1.15.F) [SVD of a matrix with vanishing bottom rows] Assume that the bottom $m - n$ rows of $\mathbf{A} \in \mathbb{R}^{m,n}$, $m > n$, vanish, that is, $\mathbf{A}$ is of the form

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_* \\ \mathbf{O}_{m-n,n} \end{bmatrix} \quad \text{with} \quad \mathbf{A}_* \in \mathbb{R}^{n,n} \; .$$

Based on a known full singular-value decomposition $\mathbf{A}_* = \mathbf{U}_*\mathbf{\Sigma}_*\mathbf{V}_*^\top$ of $\mathbf{A}_*$, state a full SVD of $\mathbf{A}$.

(Q3.4.1.15.G) [Completion to regular matrix] Given $m, n \in \mathbb{N}$, $m < n$, and a matrix $\mathbf{T} \in \mathbb{R}^{n,m}$ with *full rank* m, sketch an algorithm that computes a matrix $\mathbf{X} \in \mathbb{R}^{n,n-m}$ such that $[\mathbf{T}\ \mathbf{X}] \in \mathbb{R}^{n,n}$ is regular/invertible.

The following result can be a starting point:

Lemma 3.4.1.13. SVD and rank of a matrix

Let $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{H}}$ be the SVD of $\mathbf{A} \in \mathbb{K}^{m,n}$ according to Thm. 3.4.1.1. If, for some $1 \le r \le p := \min\{m,n\}$, the singular values of $\mathbf{A} \in \mathbb{K}^{m,n}$ satisfy

$$\sigma_1 \ge \cdots \ge \sigma_r > \sigma_{r+1} = \cdots \sigma_p = 0\,,$$

then

- $\operatorname{rank}(\mathbf{A}) = r$ *(no. of non-zero singular values)*,
- $\mathcal{N}(\mathbf{A}) = \operatorname{Span}\{(\mathbf{V})_{:,r+1}, \ldots, (\mathbf{V})_{:,n}\}$,
- $\mathcal{R}(\mathbf{A}) = \operatorname{Span}\{(\mathbf{U})_{:,1}, \ldots, (\mathbf{U})_{:,r}\}$.

△

3.4.2 SVD in EIGEN

Video tutorial for Section 3.4.2 "SVD in EIGEN ": (9 minutes) Download link, tablet notes

→ review questions 3.4.2.10

The EIGEN class **JacobiSVD** is constructed from a matrix data type, computes the SVD of its argument during construction and offers access methods `MatrixU()`, `singularValues()`, and `MatrixV()` to request the SVD-factors and singular values.

C++-code 3.4.2.1: Computing SVDs in EIGEN ⇝ GITLAB

```
2  #include <Eigen/SVD>
3  using MatrixXd = Eigen::MatrixXd;
4  // Computation of (full) SVD A = UΣV^H → Thm. 3.4.1.1
5  // SVD factors are returned as dense matrices in natural order
6  inline std::tuple<MatrixXd,MatrixXd,MatrixXd> svd_full(const MatrixXd& A) {
7    const Eigen::JacobiSVD<MatrixXd> svd(A, Eigen::ComputeFullU | Eigen::ComputeFullV);
8    const MatrixXd & U = svd.matrixU(); // get unitary (square) matrix U
9    const MatrixXd & V = svd.matrixV(); // get unitary (square) matrix V
10   const VectorXd & sv = svd.singularValues(); // get singular values as vector
11   MatrixXd Sigma = MatrixXd::Zero(A.rows(), A.cols());
12   const unsigned p = sv.size(); // no. of singular values
13   Sigma.block(0,0,p,p) = sv.asDiagonal(); // set diagonal block of Σ
14   return {U,Sigma,V};
15 }
16
17 // Computation of economical (thin) SVD A = UΣV^H, see (3.4.1.5)
18 // SVD factors are returned as dense matrices in natural order
19 inline std::tuple<MatrixXd,MatrixXd,MatrixXd> svd_eco(const MatrixXd& A) {
20   const Eigen::JacobiSVD<MatrixXd> svd(A, Eigen::ComputeThinU | Eigen::ComputeThinV);
21   const MatrixXd & U = svd.matrixU(); // get matrix U with orthonormal
       columns
22   const MatrixXd & V = svd.matrixV(); // get matrix V with orthonormal
       columns
23   const VectorXd & sv = svd.singularValues(); // get singular values as vector
```

```
24   const MatrixXd Sigma = sv.asDiagonal(); // build diagonal matrix Σ
25   return {U,Sigma,V};
26 }
```

The second argument in the constructor of **JacobiSVD** determines, whether the methods `matrixU()` and `matrixV()` return the factor for the full SVD of Def. 3.4.1.3 or of the economical (thin) SVD (3.4.1.5): `Eigen::ComputeFull*` will select the full versions, whereas `Eigen::ComputeThin*` picks the economical versions → documentation.

Internally, the computation of the SVD is done by a sophisticated algorithm, for which key steps rely on orthogonal/unitary transformations. Also there we reap the benefit of the exceptional stability brought about by norm-preserving transformations → § 3.3.3.30.

EIGEN's algorithm for computing SVD is (numerically) stable → Def. 1.5.5.19

§3.4.2.2 (Computational cost of computing the SVD) According to EIGEN's documentation the SVD of a general dense matrix involves the following asymptotic complexity:

$$\text{cost(economical SVD of } \mathbf{A} \in \mathbb{K}^{m,n}) = O(\min\{m,n\}^2 \max\{m,n\})$$

The computational effort is (asymptotically) *linear in the larger matrix dimension*. ⌟

EXAMPLE 3.4.2.3 (SVD-based computation of the rank of a matrix) Based on Lemma 3.4.1.13, the SVD is the main tool for the stable computation of the rank of a matrix (→ Def. 2.2.1.3)

However, theory as reflected in Lemma 3.4.1.13 entails identifying zero singular values, which must rely on a threshold condition in a numerical code, recall Rem. 1.5.3.15. Given the SVD $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{H}}$, $\mathbf{\Sigma} = \operatorname{diag}(\sigma_1,\ldots,\sigma_{\min\{m,n\}})$, of a matrix $\mathbf{A} \in \mathbb{K}^{m,n}$, $\mathbf{A} \neq \mathbf{0}$ and a tolerance $\mathrm{tol} > 0$, we define the numerical rank

$$r := \sharp\left\{\sigma_i : |\sigma_i| \geq \mathrm{tol} \max_j\{|\sigma_j|\}\right\}. \tag{3.4.2.4}$$

The following code implements this rule.

C++-code 3.4.2.5: Computing rank of a matrix through SVD ↬ GITLAB

```
2  // Computation of the numerical rank of a non-zero matrix by means of
3  // singular value decomposition, cf. (3.4.2.4).
4  Eigen::Index rank_by_svd(const Eigen::MatrixXd &A, double tol = EPS) {
5    if (A.norm() == 0) {
6      return static_cast<Eigen::Index>(0);
7    }
8    const Eigen::JacobiSVD<Eigen::MatrixXd> svd(A);
9    const Eigen::VectorXd & sv =
10       svd.singularValues(); // Get sorted singular values as vector
11   const Eigen::Index n = sv.size();
12   Eigen::Index r = 0;
13   // Test relative size of singular values
14   while ((r < n) && (sv(r) >= sv(0) * tol)) {
15     r++;
```

```
16    }
17    return r;
18  }
```

EIGEN offers an equivalent built-in method `rank()` for objects representing singular value decompositions:

C++-code 3.4.2.6: Using `rank()` in EIGEN ↪ GITLAB

```
2  // Computation of the numerical rank of a matrix by means of SVD
3  Eigen::Index rank_eigen(const Eigen::MatrixXd &A, double tol = EPS) {
4    return A.jacobiSvd().setThreshold(tol).rank();
5  }
```

The method `setThreshold()` passes tol from (3.4.2.4) to `rank()`. ⌟

EXAMPLE 3.4.2.7 (Computation of nullspace and image space of matrices) "Computing" a subspace of $\mathbb{R}^k$ amounts to making available a (stable) *basis* of that subspace, ideally an *orthonormal basis*.

Lemma 3.4.1.13 taught us how to glean orthonormal bases of $\mathcal{N}(\mathbf{A})$ and $\mathcal{R}(\mathbf{A})$ from the SVD of a matrix $\mathbf{A}$. This immediately gives a numerical method and its implementation is given in the next two codes.

C++-code 3.4.2.8: ONB of $\mathcal{N}(\mathbf{A})$ through SVD ↪ GITLAB

```
2  // Computation of an ONB of the kernel of a matrix
3  Eigen::MatrixXd nullspace(const Eigen::MatrixXd &A, double tol = EPS) {
4    using index_t = Eigen::Index;
5    Eigen::JacobiSVD<Eigen::MatrixXd> svd(A, Eigen::ComputeFullV);
6    const index_t r = svd.setThreshold(tol).rank();
7    // Rightmost columns of V provide ONB of N(A)
8    Eigen::MatrixXd Z = svd.matrixV().rightCols(A.cols() - r);
9    return Z;
10 }
```

C++-code 3.4.2.9: ONB of $\mathcal{R}(\mathbf{A})$ through SVD ↪ GITLAB

```
2  // Computation of an ONB of the image space of a matrix
3  Eigen::MatrixXd rangespace(const Eigen::MatrixXd &A, double tol = EPS) {
4    using index_t = Eigen::Index;
5    Eigen::JacobiSVD<Eigen::MatrixXd> svd(A, Eigen::ComputeThinU);
6    const index_t r = svd.setThreshold(tol).rank();
7    // r left columns of U provide ONB of R(A)
8    return svd.matrixU().leftCols(r);
9  }
```

⌟

Review question(s) 3.4.2.10 (SVD in EIGEN)

(Q3.4.2.10.A) Please examine Code 3.4.2.9 and detect a potentially serious loss of efficiency. In which situations will this have an impact?

△

3.4.3 Solving General Least-Squares Problems by SVD

Video tutorial for Section 3.4.3 "Solving General Least-Squares Problems by SVD": (14 minutes) Download link, tablet notes

→ review questions 3.4.3.17

In a similar fashion as explained for QR-decomposition in Section 3.3.4, the singular valued decompisition (SVD, → Def. 3.4.1.3) can be used to transform general linear least squares problems (3.1.3.7) into a simpler form. In the case of SVD-based orthogonal transformation methods this simpler form involves merely a diagonal matrix.

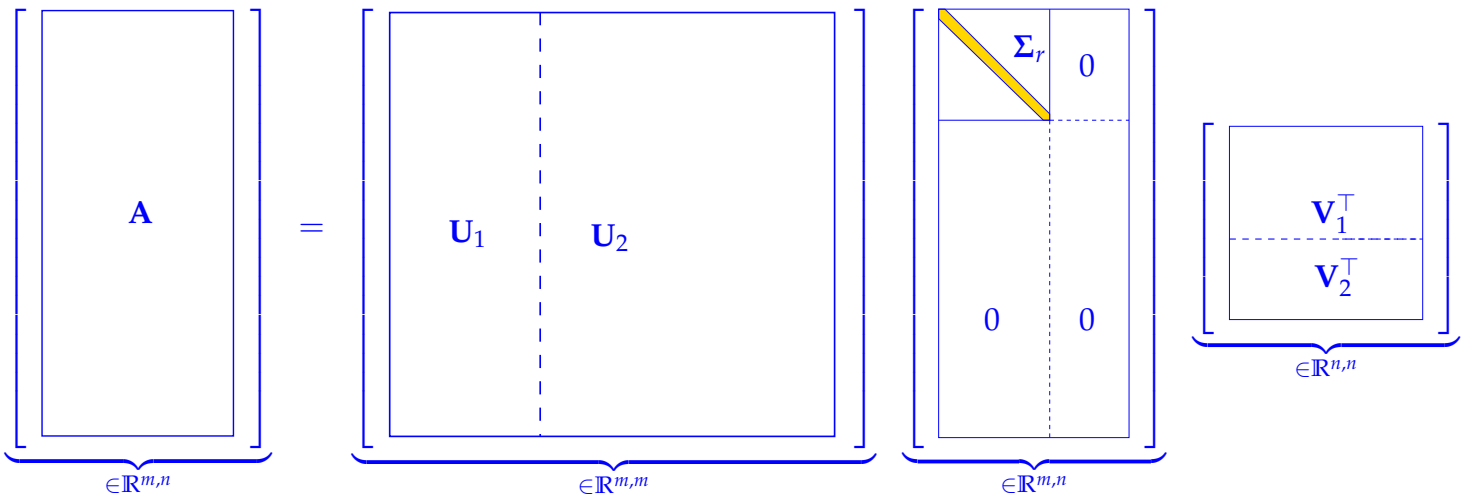
Here we consider the most general setting

$$\mathbf{A}\mathbf{x} = \mathbf{b} \in \mathbb{R}^m \quad \text{with} \quad \mathbf{A} \in \mathbb{R}^{m,n} \quad , \quad \operatorname{rank}(\mathbf{A}) = r \leq \min\{m,n\} \; .$$

In particular, we drop the assumption of full rank of $\mathbf{A}$. This means that the minimum norm condition (ii) in the definition (3.1.3.7) of a linear least squares problem may be required for singling out a unique solution.

We recall the (full) SVD of $\mathbf{A} \in \mathbb{R}^{m,n}$:

$$\mathbf{A} = \begin{bmatrix} \mathbf{U}_1 & \mathbf{U}_2 \end{bmatrix} \begin{bmatrix} \mathbf{\Sigma}_r & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{V}_1^\top \\ \mathbf{V}_2^\top \end{bmatrix}$$



(3.4.3.1)

with $\mathbf{U}_1 \in \mathbb{R}^{m,r}$, $\mathbf{U}_2 \in \mathbb{R}^{m,m-r}$ with *orthonormal* columns, $\mathbf{U} := [\mathbf{U}_1, \mathbf{U}_2]$ unitary,
$\mathbf{\Sigma}_r = \operatorname{diag}(\sigma_1, \ldots, \sigma_r) \in \mathbb{R}^{r,r}$ (singular values, Def. 3.4.1.3),
$\mathbf{V}_1 \in \mathbb{R}^{n,r}$, $\mathbf{V}_2 \in \mathbb{R}^{n,n-r}$ with *orthonormal* columns, $\mathbf{V} := [\mathbf{V}_1, \mathbf{V}_2]$ unitary.

We can proceed in two different ways, both of which we elaborate in the sequel:

Approach ❶: We can use the invariance of the 2-norm of a vector with respect to multiplication with $\mathbf{U} := [\mathbf{U}_1, \mathbf{U}_2]$, see Thm. 3.3.2.2, together with the fact that $\mathbf{U}$ is unitary, $\mathbf{U}^{-1} = \mathbf{U}^\top$, see Def. 6.3.1.2:

$$[\mathbf{U}_1, \mathbf{U}_2] \cdot \left[\begin{pmatrix} \mathbf{U}_1^\top \\ \mathbf{U}_2^\top \end{pmatrix} \right] = \mathbf{I} \; .$$

$$\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2 = \left\| [\mathbf{U}_1 \; \mathbf{U}_2] \begin{bmatrix} \mathbf{\Sigma}_r & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{V}_1^\top \\ \mathbf{V}_2^\top \end{bmatrix} \mathbf{x} - \mathbf{b} \right\|_2 = \left\| \begin{bmatrix} \mathbf{\Sigma}_r \mathbf{V}_1^\top \mathbf{x} \\ 0 \end{bmatrix} - \begin{bmatrix} \mathbf{U}_1^\top \mathbf{b} \\ \mathbf{U}_2^\top \mathbf{b} \end{bmatrix} \right\|_2 \tag{3.4.3.2}$$

We follow the same strategy as in the case of QR-based solvers for full-rank linear least squares problems. We choose $\mathbf{x}$ such that the first r components of $\begin{bmatrix}\boldsymbol{\Sigma}_r\mathbf{V}_1^\top\mathbf{x}\\ 0\end{bmatrix} - \begin{bmatrix}\mathbf{U}_1^\top\mathbf{b}\\ \mathbf{U}_2^\top\mathbf{b}\end{bmatrix}$ vanish:

$$\succ \quad \text{(possibly } \textit{underdetermined}\text{) } r\times n \text{ linear system} \qquad \boldsymbol{\Sigma}_r\mathbf{V}_1^\top\mathbf{x} = \mathbf{U}_1^\top\mathbf{b}\;. \tag{3.4.3.3}$$

To fix a unique solution in the case $r < n$ we appeal to the minimal norm condition in (3.1.3.7): appealing to the considerations of § 3.1.3.3, the solution $\mathbf{x}$ of (3.4.3.3) is unique up to contributions from

$$\mathcal{N}(\mathbf{V}_1^\top) \overset{\text{Lemma 3.1.2.12}}{=} \mathcal{R}(\mathbf{V}_1)^\perp \overset{\text{orthonormality}}{=} \mathcal{R}(\mathbf{V}_2)\;. \tag{3.4.3.4}$$

Since $\mathbf{V}$ is unitary, the minimal norm solution is obtained by setting contributions from $\mathcal{R}(\mathbf{V}_2)$ to zero, which amounts to choosing $\mathbf{x}\in\mathcal{R}(\mathbf{V}_1)$. This converts (3.4.3.3) into

$$\boldsymbol{\Sigma}_r\underbrace{\mathbf{V}_1^\top\mathbf{V}_1}_{=\mathbf{I}}\mathbf{z} = \mathbf{U}_1^\top\mathbf{b} \quad\Rightarrow\quad \mathbf{z} = \boldsymbol{\Sigma}_r^{-1}\mathbf{U}_1^\top\mathbf{b}\;.$$

$$\blacktriangleright \quad \text{generalized solution} \to \text{Def. 3.1.3.1} \qquad \mathbf{x}^\dagger = \mathbf{V}_1\boldsymbol{\Sigma}_r^{-1}\mathbf{U}_1^\top\mathbf{b} \quad , \quad \|\mathbf{r}\|_2 = \left\|\mathbf{U}_2^\top\mathbf{b}\right\|_2\;. \tag{3.4.3.5}$$

Approach ❷: From Thm. 3.1.2.1 we know that the generalized least-squares solution $\mathbf{x}^\dagger$ of $\mathbf{A}\mathbf{x}=\mathbf{b}$ solves the normal equations (3.1.2.2), and in § 3.1.3.3 we saw that $\mathbf{x}^\dagger$ lies in the orthogonal complement of $\mathcal{N}(\mathbf{A})$:

$$\mathbf{A}^\top\mathbf{A}\mathbf{x}^\dagger = \mathbf{A}^\top\mathbf{b} \quad , \quad \mathbf{x}^\dagger\in\mathcal{N}(\mathbf{A})^\perp\;. \tag{3.4.3.6}$$

By Lemma 3.4.1.13 and using the notations from (3.4.3.1) together with the fact that the columns of $\mathbf{V}$ form an orthonormal basis of $\mathbb{R}^n$:

$$\mathcal{N}(\mathbf{A}) = \mathcal{R}(\mathbf{V}_2) \quad\Leftrightarrow\quad \mathcal{N}(\mathbf{A})^\perp = \mathcal{R}(\mathbf{V}_1)\;. \tag{3.4.3.7}$$

Hence, we can write $\mathbf{x}^\dagger = \mathbf{V}_1\mathbf{y}$ for some $\mathbf{y}\in\mathbb{R}^r$. We plug this representation into the normal equations and also multiply with $\mathbf{V}_1^\top$, similar to what we did in § 3.1.3.3:

$$\blacktriangleright \qquad \mathbf{V}_1^\top\mathbf{A}^\top\mathbf{A}\mathbf{V}_1\mathbf{y} = \mathbf{V}_1^\top\mathbf{A}^\top\mathbf{b}\;. \tag{3.4.3.8}$$

Next, insert the SVD of $\mathbf{A}$:

$$\mathbf{V}_1^\top\mathbf{V}\boldsymbol{\Sigma}\underbrace{\mathbf{U}^\top\mathbf{U}}_{=\mathbf{I}}\boldsymbol{\Sigma}\mathbf{V}^\top\mathbf{V}_1^\top\mathbf{y} = \mathbf{V}_1^\top\mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^\top\mathbf{b}\;. \tag{3.4.3.9}$$

Then we switch to the block-partitioned form as in (3.4.3.1):

$$\underbrace{\mathbf{V}_1^\top[\mathbf{V}_1,\mathbf{V}_2]}_{\in\mathbb{R}^{r,n}}\underbrace{\left[\begin{array}{c|c}\boldsymbol{\Sigma}_r & \mathbf{O}\\ \hline \mathbf{O} & \mathbf{O}\end{array}\right]^\top\left[\begin{array}{c|c}\boldsymbol{\Sigma}_r & \mathbf{O}\\ \hline \mathbf{O} & \mathbf{O}\end{array}\right]}_{\in\mathbb{R}^{n,n}}\underbrace{\begin{bmatrix}\mathbf{V}_1^\top\\ \mathbf{V}_2^\top\end{bmatrix}\mathbf{V}_1}_{\in\mathbb{R}^{n,r}}\mathbf{y} = \mathbf{V}_1^\top[\mathbf{V}_1,\mathbf{V}_2]\left[\begin{array}{c|c}\boldsymbol{\Sigma}_r & \mathbf{O}\\ \hline \mathbf{O} & \mathbf{O}\end{array}\right]^\top\begin{bmatrix}\mathbf{U}_1^\top\\ \mathbf{U}_2^\top\end{bmatrix}\mathbf{b} \tag{3.4.3.10}$$

$$\Updownarrow$$

$$[\mathbf{I},\mathbf{O}]\left[\begin{array}{c|c}\boldsymbol{\Sigma}_r^2 & \mathbf{O}\\ \hline \mathbf{O} & \mathbf{O}\end{array}\right]\begin{bmatrix}\mathbf{I}\\ \mathbf{O}\end{bmatrix}\mathbf{y} = [\mathbf{I},\mathbf{O}]\left[\begin{array}{c|c}\boldsymbol{\Sigma}_r & \mathbf{O}\\ \hline \mathbf{O} & \mathbf{O}\end{array}\right]^\top\begin{bmatrix}\mathbf{U}_1^\top\\ \mathbf{U}_2^\top\end{bmatrix}\mathbf{b} \tag{3.4.3.11}$$

$$\Updownarrow$$

$$\Sigma_r^2 \mathbf{y} = \Sigma_r \mathbf{U}_1^\top \mathbf{b} \,. \tag{3.4.3.12}$$

Cancelling the invertible matrix Σ_r on both sides yields formula (3.4.3.5).

We remind that In a practical implementation, as in Code 3.4.2.5, one has to resort to the numerical rank from (3.4.2.4):

$$r = \max\{i\colon \sigma_i/\sigma_1 > \texttt{tol}\} \,,$$

where we have assumed that the singular values σ_j are sorted according to decreasing modulus.

C++-code 3.4.3.13: Computing generalized solution of $\mathbf{Ax} = \mathbf{b}$ via SVD ↪ GITLAB

```
Eigen::VectorXd lsqsvd(const Eigen::MatrixXd &A, const Eigen::VectorXd &b) {
  // Compute economical SVD, compare Code 3.4.2.1
  const Eigen::JacobiSVD<MatrixXd> svd(A, Eigen::ComputeThinU | Eigen::ComputeThinV);
  const Eigen::VectorXd & sv = svd.singularValues();
  const unsigned int r = svd.rank();  // Numerical rank, default tolerance
  // x^+ = V_1 Σ_r^{-1} U_1^H b, see (3.4.3.5)
  const MatrixXd & U = svd.matrixU();
  const MatrixXd & V = svd.matrixV();

  return V.leftCols(r) * (sv.head(r).cwiseInverse().asDiagonal() *
                          (U.leftCols(r).adjoint() * b));
}
```

The `solve()` method directly returns the generalized solution

C++-code 3.4.3.14: Computing generalized solution of $\mathbf{Ax} = \mathbf{b}$ via SVD ↪ GITLAB

```
Eigen::VectorXd lsqsvd_eigen(const Eigen::MatrixXd &A, const Eigen::VectorXd &b) {
  const Eigen::JacobiSVD<MatrixXd> svd(A, Eigen::ComputeThinU | Eigen::ComputeThinV);
  return svd.solve(b);
}
```

Remark 3.4.3.15 (Pseudoinverse and SVD → [Han02, Ch. 12], [DR08, Sect. 4.7]) From Thm. 3.1.3.6 we could conclude a general formula for the Moore-Penrose pseudoinverse of any matrix $\mathbf{A} \in \mathbb{R}^{m,n}$. Now, the solution formula (3.4.3.5) directly yields a concrete incarnation of the pseudoinverse $\mathbf{A}^+$.

Theorem 3.4.3.16. Pseudoinverse and SVD

If $\mathbf{A} \in \mathbb{K}^{m,n}$ with $\operatorname{rank}(\mathbf{A}) = r$ has the full singular value decomposition $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^{\mathrm{H}}$ (→ Thm. 3.4.1.1) partitioned as in (3.4.3.1)*, then its Moore-Penrose pseudoinverse (→ Thm. 3.1.3.6) is given by $\mathbf{A}^+ = \mathbf{V}_1\Sigma_r^{-1}\mathbf{U}_1^{\mathrm{H}}$.*

⌟

Review question(s) 3.4.3.17 (Solving general least-squares problems by SVD)

(Q3.4.3.17.A) Discuss the efficient implementation of a C++ function

```
Eigen::VectorXd solveRankOneLsq(const Eigen::VectorXd &u,
  const Eigen::VectorXd &v, const Eigen::VectorXd &b);
```

that returns the general least squares solution of $\mathbf{Ax} = \mathbf{b}$ for the rank-1 matrix $\mathbf{A} := \mathbf{uv}^\top$, $\mathbf{u} \in \mathbb{R}^m$, $\mathbf{v} \in \mathbb{R}^n$, $m \geq n$.

△

3.4.4 SVD-Based Optimization and Approximation

For the general least squares problem (3.1.3.7) we have seen the use of SVD for its numerical solution in Section 3.4.3. The the SVD was a powerful tool for solving a minimization problem for a 2-norm. In many other contexts the SVD is also a key component in numerical optimization.

3.4.4.1 Norm-Constrained Extrema of Quadratic Forms

Video tutorial for Section 3.4.4.1 "Norm-Constrained Extrema of Quadratic Forms": (11 minutes) Download link, tablet notes

→ review questions 3.4.4.13

We consider the following problem of finding the extrema of quadratic forms on the Euclidean unit sphere $\{\mathbf{x} \in \mathbb{K}^n : \|\mathbf{x}\|_2 = 1\}$:

$$\text{given} \quad \mathbf{A} \in \mathbb{K}^{m,n}\,,\ m \geq n, \quad \text{find } \mathbf{x} \in \mathbb{K}^n,\ \|\mathbf{x}\|_2 = 1\,, \quad \|\mathbf{A}\mathbf{x}\|_2 \to \min\,. \tag{3.4.4.1}$$

Use that multiplication with orthogonal/unitary matrices preserves the 2-norm (→ Thm. 3.3.2.2) and resort to the (full) singular value decomposition $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{H}}$ (→ Def. 3.4.1.3):

$$\begin{aligned}\min_{\|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2^2 &= \min_{\|\mathbf{x}\|_2=1} \left\|\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{H}}\mathbf{x}\right\|_2^2 = \min_{\|\mathbf{V}^{\mathrm{H}}\mathbf{x}\|_2=1} \left\|\mathbf{U}\mathbf{\Sigma}(\mathbf{V}^{\mathrm{H}}\mathbf{x})\right\|_2^2 \\ [\,\mathbf{y} = \mathbf{V}^{\mathrm{H}}\mathbf{x}\,] \quad &= \min_{\|\mathbf{y}\|_2=1} \|\mathbf{\Sigma}\mathbf{y}\|_2^2 = \min_{\|\mathbf{y}\|_2=1} (\sigma_1^2 y_1^2 + \cdots + \sigma_n^2 y_n^2) \geq \sigma_n^2\,.\end{aligned}$$

Since the singular values are assumed to be sorted as $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$, the minimum with value σ_n^2 is attained for $y_n^2 = 1$ and $y_1 = \cdots = y_{n-1} = 0$, that is, $\mathbf{V}^{\mathrm{H}}\mathbf{x} = \mathbf{y} = \mathbf{e}_n$ ($\hat{=}$ n-th Cartesian basis vector in $\mathbb{R}^n$). ⇒ minimizer $\mathbf{x}^* = \mathbf{V}\mathbf{e}_n = (\mathbf{V})_{:,n}$, minimal value $\|\mathbf{A}\mathbf{x}^*\|_2 = \sigma_n$.

C++ code 3.4.4.2: Solving (3.4.4.1) with EIGEN ↪ GITLAB

```
// EIGEN based function for solving (3.4.4.1);
// minimizer returned nin x, mininum as return value
double minconst(VectorXd &x,const MatrixXd &A) {
  const Eigen::Index m = A.rows();
  const Eigen::Index n = A.cols();
  if (m < n) {
    throw std::runtime_error("A must be tall matrix");
  }
  // SVD factor U is not computed!
  const Eigen::JacobiSVD<MatrixXd> svd(A,Eigen::ComputeThinV);
  x.resize(n); x.setZero(); x(n-1) = 1.0; // e_n
  x = svd.matrixV()*x;
  return (svd.singularValues())(n-1);
}
```

By similar arguments we can solve the corresponding norm constrained maximization problem

$$\text{given} \quad \mathbf{A} \in \mathbb{K}^{m,n}\,,\ m \geq n, \quad \text{find } \mathbf{x} \in \mathbb{K}^n,\ \|\mathbf{x}\|_2 = 1\,, \quad \|\mathbf{A}\mathbf{x}\|_2 \to \max\,,$$

and obtain the solution based on the SVD $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{H}}$ of $\mathbf{A}$:

$$\sigma_1 = \max_{\|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2 \quad , \quad (\mathbf{V})_{:,1} = \operatorname*{argmax}_{\|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2\,. \tag{3.4.4.3}$$

Recall: The Euclidean matrix norm (2-norm) of the matrix $\mathbf{A}$ ($\rightarrow$ Def. 1.5.5.10) is defined as the maximum in (3.4.4.3). Thus we have proved the following theorem:

Lemma 3.4.4.4. SVD and Euclidean matrix norm

If $\mathbf{A} \in \mathbb{K}^{m,n}$ has singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$, $p := \min\{m,n\}$, then its Euclidean matrix norm is given by $\|\mathbf{A}\|_2 = \sigma_1(\mathbf{A})$.
If $m = n$ and $\mathbf{A}$ is regular/invertible, then its 2-norm condition number is $\operatorname{cond}_2(\mathbf{A}) = \sigma_1/\sigma_n$.

EXAMPLE 3.4.4.5 (Fit of hyperplanes) For an important application from computational geometry, this example studies the power and versatility of orthogonal transformations in the context of (generalized) least squares minimization problems.

From school recall the Hesse normal form of a hyperplane $\mathcal{H}$ (= affine subspace of dimension $d-1$) in $\mathbb{R}^d$:

$$\mathcal{H} = \{\mathbf{x} \in \mathbb{R}^d : c + \mathbf{n}^\top \mathbf{x} = 0\}\ , \quad \mathbf{n} \in \mathbb{R}^d\ , \quad \|\mathbf{n}\|_2 = 1\ . \tag{3.4.4.6}$$

where $\mathbf{n}$ is the unit normal to $\mathcal{H}$ and $|c|$ gives the distance of $\mathcal{H}$ from $\mathbf{0}$. The Hesse normal form is convenient for computing the distance of points from $\mathcal{H}$, because the

$$\text{Euclidean distance of } \mathbf{y} \in \mathbb{R}^d \text{ from the plane is } \quad \operatorname{dist}(\mathcal{H}, \mathbf{y}) = |c + \mathbf{n}^\top \mathbf{y}|\ , \tag{3.4.4.7}$$

Goal: given the point coordinate vectors $\mathbf{y}_1, \ldots, \mathbf{y}_m \in \mathbb{R}^d$, $m > d$, find $\mathcal{H} \leftrightarrow \{c \in \mathbb{R}, \mathbf{n} \in \mathbb{R}^d, \|\mathbf{n}\|_2 = 1\}$, such that

$$\sum_{j=1}^{m} \operatorname{dist}(\mathcal{H}, \mathbf{y}_j)^2 = \sum_{j=1}^{m} |c + \mathbf{n}^\top \mathbf{y}_j|^2 \rightarrow \min\ . \tag{3.4.4.8}$$

Note that (3.4.4.8) is **not** a linear least squares problem due to the constraint $\|\mathbf{n}\|_2 = 1$. However, it turns out to be a minimization problem with *almost* the structure of (3.4.4.1) ($y_{k,\ell} := (\mathbf{y}_k)_\ell$):

$$(3.4.4.8) \quad \Leftrightarrow \quad \left\| \underbrace{\begin{bmatrix} 1 & y_{1,1} & \cdots & y_{1,d} \\ 1 & y_{2,1} & \cdots & y_{2,d} \\ \vdots & \vdots & & \vdots \\ 1 & y_{m,1} & \cdots & y_{m,d} \end{bmatrix}}_{=:\mathbf{A}} \underbrace{\begin{bmatrix} c \\ n_1 \\ \vdots \\ n_d \end{bmatrix}}_{=:\mathbf{x}} \right\|_2 \rightarrow \min \quad \text{under the constraint} \quad \|\mathbf{n}\|_2 = 1\ .$$

Note that the solution component c is not subject to the constraint. One is tempted to use this freedom to make one component of $\mathbf{Ax}$ vanish, but which one is not clear. This is why we need another preparatory step.

Step ❶: To convert the minimization problem into the form (3.4.4.1) we start with a QR-decomposition ($\rightarrow$ Section 3.3.3)

$$\mathbf{A} := \begin{bmatrix} 1 & y_{1,1} & \cdots & y_{1,d} \\ 1 & y_{2,1} & \cdots & y_{2,d} \\ \vdots & \vdots & & \vdots \\ 1 & y_{m,1} & \cdots & y_{m,d} \end{bmatrix} = \mathbf{QR}\ , \quad \mathbf{R} := \begin{bmatrix} r_{11} & r_{12} & \cdots & \cdots & r_{1,d+1} \\ 0 & r_{22} & \cdots & \cdots & r_{2,d+1} \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \vdots \\ 0 & & & & r_{d+1,d+1} \\ 0 & \cdots & & \cdots & 0 \\ \vdots & & & & \vdots \\ 0 & \cdots & & \cdots & 0 \end{bmatrix} \in \mathbb{R}^{m,d+1}\ .$$

$$\|\mathbf{Ax}\|_2 \to \min \quad \Leftrightarrow \quad \|\mathbf{Rx}\|_2 = \left\| \begin{bmatrix} r_{11} & r_{12} & \cdots & \cdots & r_{1,d+1} \\ 0 & r_{22} & \cdots & \cdots & r_{2,d+1} \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \vdots \\ 0 & & & & r_{d+1,d+1} \\ 0 & \cdots & & \cdots & 0 \\ \vdots & & & & \vdots \\ 0 & \cdots & & \cdots & 0 \end{bmatrix} \begin{bmatrix} c \\ n_1 \\ \vdots \\ \vdots \\ n_d \end{bmatrix} \right\|_2 \to \min \ . \tag{3.4.4.9}$$

Step ❷ Note that, if $\mathbf{n}$ is the solution of (3.4.4.8), then necessarily (why?)

$$c \cdot r_{11} + n_1 \cdot r_{12} + \cdots + r_{1,d+1} \cdot n_d = 0 \ .$$

This insight converts (3.4.4.9) to

$$\left\| \begin{bmatrix} r_{22} & r_{23} & \cdots & \cdots & r_{2,d+1} \\ 0 & r_{33} & \cdots & \cdots & r_{3,d+1} \\ \vdots & & \ddots & & \vdots \\ 0 & & & & r_{d+1,d+1} \end{bmatrix} \begin{bmatrix} n_1 \\ \vdots \\ \vdots \\ n_d \end{bmatrix} \right\|_2 \to \min \quad , \quad \|\mathbf{n}\|_2 = 1 \ . \tag{3.4.4.10}$$

(3.4.4.10) is now a problem of type (3.4.4.1), minimization on the Euclidean sphere. Hence, (3.4.4.10) can be solved using the SVD-based algorithm implemented in Code 3.4.4.2.

Note: Since $r_{11} = \|(\mathbf{A})_{:,1}\|_2 = \sqrt{m} \neq 0$, $c = -r_{11}^{-1} \sum_{j=1}^{d} r_{1,j+1} n_j$ can always be computed.

This algorithm is implemented as case `p==dim+1` in the following code, making heavy use of EIGEN's block access operations and the built-in QR-decomposition and SVD factorization.

C++-code 3.4.4.11: (Generalized) distance fitting of a hyperplane: solution of (3.4.4.12)
➥ GITLAB

```
// Solves constrained linear least squares problem
// (3.4.4.12) with dim passing d
std::pair<Eigen::VectorXd, Eigen::VectorXd> clsq(const MatrixXd& A,
                                                 const unsigned dim) {
  const unsigned p = A.cols();
  unsigned m = A.rows();
  if (p < dim + 1) {
    throw runtime_error("not enough unknowns");
  }
  if (m < dim) {
   throw runtime_error("not enough equations");
  }
  m = std::min(m, p); // Number of variables
  // First step:  orthogonal transformation, see Code 3.3.4.1
  MatrixXd R = A.householderQr().matrixQR().template triangularView<Eigen::Upper>();
  // compute matrix V from SVD composition of R, solve (3.4.4.10)
  MatrixXd V = R.block(p - dim, p - dim, m + dim - p, dim)
                .jacobiSvd(Eigen::ComputeFullV).matrixV();
  const VectorXd n = V.col(dim - 1); // Norm-constrained part of solution
     vector
  // Compute free part of solution vector
  const auto R_topleft = R.topLeftCorner(p - dim, p - dim);
```

```
  // Check for singular matrix
  const auto R_diag = R_topleft.diagonal().cwiseAbs();
  if (R_diag.minCoeff() < (numeric_limits<double>::epsilon())*R_diag.maxCoeff()) {
    throw runtime_error("Upper left block of R not regular");
  }
  const VectorXd c = -(R_topleft.template triangularView<Eigen::Upper>()).
                   solve(R.block(0, p - dim, p - dim, dim) * n);
  return {c,n};
}
```

Note that Code 3.4.4.11 solves the general problem: For $\mathbf{A} \in \mathbb{K}^{m,n}$ find $\mathbf{n} \in \mathbb{R}^d$, $\mathbf{c} \in \mathbb{R}^{n-d}$ such that

$$\left\| \mathbf{A} \begin{bmatrix} \mathbf{c} \\ \mathbf{n} \end{bmatrix} \right\|_2 \to \min \quad \text{with constraint} \quad \|\mathbf{n}\|_2 = 1 \,. \tag{3.4.4.12}$$

⌟

Review question(s) 3.4.4.13 (Norm-Constrained Extrema of Quadratic Forms)

(Q3.4.4.13.A) Let $\mathbf{M} \in \mathbb{R}^{n,n}$ be symmetric and positive definite (*s.p.d.*) and $\mathbf{A} \in \mathbb{R}^{m,n}$. Devise an algorithm for computing

$$\operatorname*{argmax}_{\mathbf{x} \in B} \|\mathbf{A}\mathbf{x}\| \quad , \quad B := \{\mathbf{x} \in \mathbb{R}^n,\ \mathbf{x}^\top \mathbf{M}\mathbf{x} = 1\} \,,$$

also based on the SVD of $\mathbf{M}$.

△

3.4.4.2 Best Low-Rank Approximation

Video tutorial for Section 3.4.4.2 "Best Low-Rank Approximation": (13 minutes) Download link, tablet notes

→ review questions 3.4.4.25

§3.4.4.14 (Low-Rank matrix compression) Matrix compression addresses the problem of approximating a given "generic" matrix (of a certain class) by means of matrix, whose "information content", that is, the number of reals needed to store it, is significantly lower than the information content of the original matrix.

Sparse matrices (→ Notion 2.7.0.1) are a prominent class of matrices with "low information content". Unfortunately, they cannot approximate dense matrices very well. Another type of matrices that enjoy "low information content", also called data sparse, are low-rank matrices.

Lemma 3.4.4.15.

If $\mathbf{A} \in \mathbb{R}^{m,n}$ has rank $r \leq \min\{m,n\}$ (→ Def. 2.2.1.3), then there exist $\mathbf{X} \in \mathbb{R}^{m,r}$ and $\mathbf{Y} \in \mathbb{R}^{n,r}$, such that $\mathbf{A} = \mathbf{X}\mathbf{Y}^\top$.

Proof. The lemma is a straightforward consequence of Lemma 3.4.1.13 and (3.4.1.14): If $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ is the SVD of $\mathbf{A}$, then choose

$$\mathbf{X} := (\mathbf{U})_{:,1:r}(\mathbf{\Sigma})_{1:r,1:r} \quad , \quad \mathbf{Y} = (\mathbf{V})_{:,1:r} \,.$$

□

None of the columns of $\mathbf{U}$ and $\mathbf{V}$ can vanish. Hence, in addition, we may assume that the columns of $\mathbf{U}$ are normalized: $\left\|(\mathbf{U})_{:,j}\right\|_2 = 1$, $j = 1,\ldots,r$.

It takes only $r(m+n-r)$ real numbers to store $\mathbf{A} \in \mathbb{R}^{m,n}$ with $\operatorname{rank}(\mathbf{A}) = r$.

Thus approximating a given matrix $\mathbf{A} \in \mathbb{R}^{m,n}$ with a rank-r matrix, $r \ll \min\{m,n\}$, can be regarded as an instance of matrix compression. The approximation error with respect to some matrix norm $\|\cdot\|$ will be minimal if we choose the best approximation

$$\mathbf{A}_r := \operatorname{argmin}\{\|\mathbf{A}-\mathbf{B}\| : \mathbf{B} \in \mathbb{R}^{m,n},\ \operatorname{rank}(\mathbf{B}) = r\}\,, \quad 1 \le r \le \min\{m,n\}\,. \tag{3.4.4.16}$$

⌟

Here we explore low-rank best approximation of general matrices with respect to the Euclidean matrix norm $\|\cdot\|_2$ induced by the 2-norm for vectors (→ Def. 1.5.5.10), and the Frobenius norm $\|\cdot\|_F$.

Definition 3.4.4.17. Frobenius norm

The Frobenius norm of $\mathbf{A} \in \mathbb{K}^{m,n}$ is defined as

$$\|\mathbf{A}\|_F^2 := \sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2\,.$$

It should be obvious that $\|\mathbf{A}\|_F$ invariant under orthogonal/unitary transformations of $\mathbf{A}$. Thus the Frobenius norm of a matrix $\mathbf{A}$, $\operatorname{rank}(\mathbf{A}) = r$, can be expressed through its singular values σ_j:

Frobenius norm and SVD: $\boxed{\|\mathbf{A}\|_F^2 = \sum_{j=1}^{r} \sigma_j^2}$ (3.4.4.18)

✎ notation: $\mathcal{R}_r(m,n) := \{\mathbf{A} \in \mathbb{K}^{m,n}\colon \operatorname{rank}(\mathbf{A}) \le r\}$, $m,n,r \in \mathbb{N}$

The next profound result links best approximation in $\mathcal{R}_r(m,n)$ and the singular value decomposition (→ Def. 3.4.1.3).

Theorem 3.4.4.19. Best low rank approximation → [Gut09, Thm. 11.6]

Let $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{H}}$ be the SVD of $\mathbf{A} \in \mathbb{K}^{m.n}$ (→ Thm. 3.4.1.1). For $1 \le k \le \operatorname{rank}(\mathbf{A})$ set

$$\mathbf{A}_k := \mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^{\mathrm{H}} = \sum_{\ell=1}^{k} \sigma_\ell (\mathbf{U})_{:,\ell}(\mathbf{V})_{:,\ell}^{\mathrm{H}} \quad \text{with} \quad \begin{aligned} \mathbf{U}_k &:= \left[(\mathbf{U})_{:,1},\ldots,(\mathbf{U})_{:,k}\right] \in \mathbb{K}^{m,k}\,,\\ \mathbf{V}_k &:= \left[(\mathbf{V})_{:,1},\ldots,(\mathbf{V})_{:,k}\right] \in \mathbb{K}^{n,k}\,,\\ \mathbf{\Sigma}_k &:= \operatorname{diag}(\sigma_1,\ldots,\sigma_k) \in \mathbb{K}^{k,k}\,. \end{aligned}$$

Then, for $\|\cdot\| = \|\cdot\|_F$ and $\|\cdot\| = \|\cdot\|_2$, holds true

$$\|\mathbf{A}-\mathbf{A}_k\| \le \|\mathbf{A}-\mathbf{F}\| \quad \forall \mathbf{F} \in \mathcal{R}_k(m,n)\,,$$

that is, $\mathbf{A}_k$ is the rank-k best approximation of $\mathbf{A}$ in the matrix norms $\|\cdot\|_F$ and $\|\cdot\|_2$.

This theorem teaches us that the rank-k-matrix that is *closest* to $\mathbf{A}$ (rank-k best approximation) in both the Euclidean matrix norm and the Frobenniusnorm (→ Def. 3.4.4.17) can be obtained by truncating the rank-1 sum expansion (3.4.1.9) obtained from the SVD of $\mathbf{A}$ after k terms.

Proof. (of Thm. 3.4.4.19) As in the statement of the theorem write $\mathbf{A}_k = \mathbf{U}_k\Sigma_k\mathbf{V}_k^{\mathrm{H}}$ for the truncated SVD. Obviously, since (here shown for $m \geq n$)

$$\mathbf{A} - \mathbf{A}_k = \mathbf{U}\left[\begin{array}{c|c} \mathbf{O}_{k,k} & \mathbf{O}_{k,n-k} \\ \hline \mathbf{O}_{n-k} & \begin{array}{ccc} \sigma_{k+1} & & \\ & \ddots & \\ & & \sigma_n \end{array} \\ \hline \mathbf{O}_{m-n,k} & \mathbf{O}_{m-n,n-k} \end{array}\right]\mathbf{V}^{\mathrm{H}} ,$$

and both matrix norms are *invariant* under multiplication with orthogonal matrices, we conclude

$$\operatorname{rank}\mathbf{A}_k = k \quad \text{and} \quad \|\mathbf{A} - \mathbf{A}_k\| = \|\mathbf{\Sigma} - \mathbf{\Sigma}_k\| = \begin{cases} \sigma_{k+1} & \text{, for } \|\cdot\| = \|\cdot\|_2 , \\ \sqrt{\sigma_{k+1}^2 + \cdots + \sigma_r^2} & \text{, for } \|\cdot\| = \|\cdot\|_F . \end{cases}$$

❶ First we tackle the Euclidean matrix norm $\|\cdot\| = \|\cdot\|_2$. For the sake of brevity we write $\mathbf{v}_j := (\mathbf{V})_{:,j}$, $\mathbf{u}_j := (\mathbf{U})_{:,j}$ for the columns of the SVD-factors $\mathbf{V}$ and $\mathbf{U}$, respectively. Pick $\mathbf{B} \in \mathbb{K}^{n,n}$, $\operatorname{rank}\mathbf{B} = k$.

$$\blacktriangleright \quad \dim\mathcal{N}(\mathbf{B}) = n - k \;\Rightarrow\; \mathcal{N}(\mathbf{B}) \cap \operatorname{Span}\{\mathbf{v}_1, \ldots, \mathbf{v}_{k+1}\} \neq \{0\} .$$

For $\mathbf{x} \in \mathcal{N}(\mathbf{B}) \cap \operatorname{Span}\{\mathbf{v}_1, \ldots, \mathbf{v}_{k+1}\}$, $\|\mathbf{x}\|_2 = 1$, we have an expansion into columns of $\mathbf{V}$: $\mathbf{x} = \sum_{j=1}^{k+1}(\mathbf{v}_j^{\mathrm{H}}\mathbf{x})\mathbf{v}_j$. We use this, the fact that $\mathbf{x}$ is a unit vector and the definition of the Euclidean matrix norm.

$$\|\mathbf{A} - \mathbf{B}\|_2^2 \geq \|(\mathbf{A} - \mathbf{B})\mathbf{x}\|_2^2 = \|\mathbf{A}\mathbf{x}\|_2^2 = \left\|\sum_{j=1}^{k+1}\sigma_j(\mathbf{v}_j^{\mathrm{H}}\mathbf{x})\mathbf{u}_j\right\|_2^2 = \sum_{j=1}^{k+1}\sigma_j^2(\mathbf{v}_j^{\mathrm{H}}\mathbf{x})^2 \geq \sigma_{k+1}^2 ,$$

because $\sum_{j=1}^{k+1}(\mathbf{v}_j^{\mathrm{H}}\mathbf{x})^2 = \|\mathbf{x}\|_2^2 = 1$.

❷ Now we turn to the Frobenius norm $\|\cdot\|_F$. We assume that $\mathbf{B} \in \mathbb{K}^{m,n}$, $\operatorname{rank}(\mathbf{B}) = k < \min\{m,n\}$, minimizes $\mathbf{A} - \mathbf{F}$ among all rank-k-matrices $\mathbf{F} \in \mathbb{K}^{m,n}$. We have to show that $\mathbf{B}$ coincides with the truncated SVD of $\mathbf{A}$: $\mathbf{B} = \mathbf{A}_k$.

The trick is to consider the full SVD of $\mathbf{B}$:

$$\mathbf{B} = \mathbf{U}_B\begin{bmatrix}\mathbf{\Sigma}_B & \mathbf{O} \\ \mathbf{O} & \mathbf{O}\end{bmatrix}\mathbf{V}_B^{\mathrm{H}} , \quad \begin{array}{l} \mathbf{U}_B \in \mathbb{K}^{m,m} \text{ unitary} , \\ \mathbf{\Sigma}_B \in \mathbb{R}^{k,k} \text{ diagonal} , \\ \mathbf{V}_B \in \mathbb{K}^{n,n} \text{ unitary} . \end{array}$$

Generically, we can write

$$\mathbf{U}_B^{\mathrm{H}}\mathbf{A}\mathbf{V}_B = \begin{bmatrix}\mathbf{L} + \mathbf{D} + \mathbf{R} & \mathbf{X}_{12} \\ \mathbf{X}_{21} & \mathbf{X}_{22}\end{bmatrix} , \quad \begin{array}{l} \mathbf{L} \in \mathbb{K}^{k,k} \text{ strictly lower triangular} , \\ \mathbf{D} \in \mathbb{K}^{k,k} \text{ diagonal} , \\ \mathbf{R} \in \mathbb{K}^{k,k} \text{ strictly upper triangular} , \end{array}$$

with some matrices $\mathbf{X}_{12} \in \mathbb{K}^{k,n-k}$, $\mathbf{X}_{21} \in \mathbb{K}^{m-k,k}$, $\mathbf{X}_{22} \in \mathbb{K}^{m-k,n-k}$. Then we introduce two $m \times n$ *rank-k*-matrices:

$$\mathbf{C}_1 := \mathbf{U}_B\begin{bmatrix}\mathbf{L} + \mathbf{\Sigma}_B + \mathbf{R} & \mathbf{X}_{12} \\ \mathbf{O} & \mathbf{O}\end{bmatrix}\mathbf{V}_B^{\mathrm{H}} ,$$

$$\mathbf{C}_2 := \mathbf{U}_B\begin{bmatrix}\mathbf{L} + \mathbf{\Sigma}_B + \mathbf{R} & \mathbf{O} \\ \mathbf{X}_{21} & \mathbf{O}\end{bmatrix}\mathbf{V}_B^{\mathrm{H}} .$$

Since $\|\mathbf{A}-\mathbf{B}\|_F$ is minimal, we conclude from the invariance of the Frobenius norm under orthogonal transformations

$$\|\mathbf{A}-\mathbf{C}_1\|_F^2 \geq \|\mathbf{A}-\mathbf{B}\|_F^2 = \|\mathbf{A}-\mathbf{C}_1\|_F^2 + \|\mathbf{L}\|_F^2 + \|\mathbf{R}\|_F^2 + \|\mathbf{X}_{12}\|_F^2 ,$$
$$\|\mathbf{A}-\mathbf{C}_2\|_F^2 \geq \|\mathbf{A}-\mathbf{B}\|_F^2 = \|\mathbf{A}-\mathbf{C}_2\|_F^2 + \|\mathbf{L}\|_F^2 + \|\mathbf{R}\|_F^2 + \|\mathbf{X}_{21}\|_F^2 .$$

Obviously, this implies $\mathbf{L}=\mathbf{O}$, $\mathbf{R}=\mathbf{O}$, $\mathbf{X}_{12}=\mathbf{O}$, and $\mathbf{X}_{21}=\mathbf{O}$, which means that

$$\mathbf{A} = \mathbf{U}\begin{bmatrix}\mathbf{D} & \mathbf{O}\\ \mathbf{O} & \mathbf{X}_{22}\end{bmatrix}\mathbf{V}_b^{\mathrm{H}} \quad \mathbf{d}\in\mathbb{K}^{k,k} \text{ diagonal} .$$

Write $\mathbf{D}=\operatorname{diag}(d_1,\ldots,d_k)$. Then we have with $r:=\operatorname{rank}(\mathbf{A})$

$$\|\mathbf{A}\|_F^2 = \sigma_1^2+\cdots+\sigma_r^2 = d_1^2+\cdots+d_k^2+\|\mathbf{X}_{22}\|_F^2 ,$$
$$\|\mathbf{A}-\mathbf{B}\|_F^2 = \|\mathbf{D}-\mathbf{\Sigma}_B\|_F^2+\|\mathbf{X}_{22}\|_F^2 = \sigma_{k+1}^2+\cdots+\sigma_r^2 .$$

Hence, by the minimizer property of $\mathbf{B}$

1. $\mathbf{\Sigma}_B=\mathbf{D}$, because any other choice makes $\|\mathbf{A}-\mathbf{B}\|_F$ bigger,
2. $\|\mathbf{X}_{22}\|$ must be minimal, which entails choosing $d_j=\sigma_j$, $j=1,\ldots,k$.

$$\blacktriangleright \quad \mathbf{U}_B^{\mathrm{H}}\mathbf{A}\mathbf{V}_B = \begin{bmatrix}\sigma_1 & & & \\ & \ddots & & \mathbf{O}\\ & & \sigma_k & \\ & \mathbf{O} & & \mathbf{X}_{22}\end{bmatrix} .$$

This is possible only, if the k leftmost columns of both $\mathbf{U}_B$ and $\mathbf{V}_B$ agree with those of the corresponding SVD-factors of $\mathbf{A}$, which means $\mathbf{B}=\mathbf{A}_k$.

□

The following code computes the low-rank best approximation of a dense matrix in EIGEN.

C++ code 3.4.4.20: SVD-based low-rank matrix compression, ⇝ GITLAB

```
Eigen::MatrixXd lowrankbestapprox(const Eigen::MatrixXd &A, unsigned int k) {
  // Compute economical SVD, compare Code 3.4.2.1
  const Eigen::JacobiSVD<MatrixXd> svd(
      A, Eigen::ComputeThinU | Eigen::ComputeThinV);
  // Form matrix product U_k Σ_k V_k.
  // Extract Σ_k as diagonal matrix of largest k singular
  // values.  EIGEN provides singular values in decreasing order!
  return (svd.matrixU().leftCols(k)) *
         (svd.singularValues().head(k).asDiagonal()) *
         (svd.matrixV().leftCols(k).transpose());
}
```

§3.4.4.21 (Error of low-rank best approxmation of a matrix) Since the matrix norms $\|\cdot\|_2$ and $\|\cdot\|_F$ are invariant under multiplication with orthogonal (unitary) matrices, we immediately obtain expressions for the norms of the best approximation error:

$$\text{In Euclidean matrix norm:} \quad \left\|\mathbf{A}-\mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^{\mathrm{H}}\right\|_2 = \sigma_{k+1} , \tag{3.4.4.22}$$

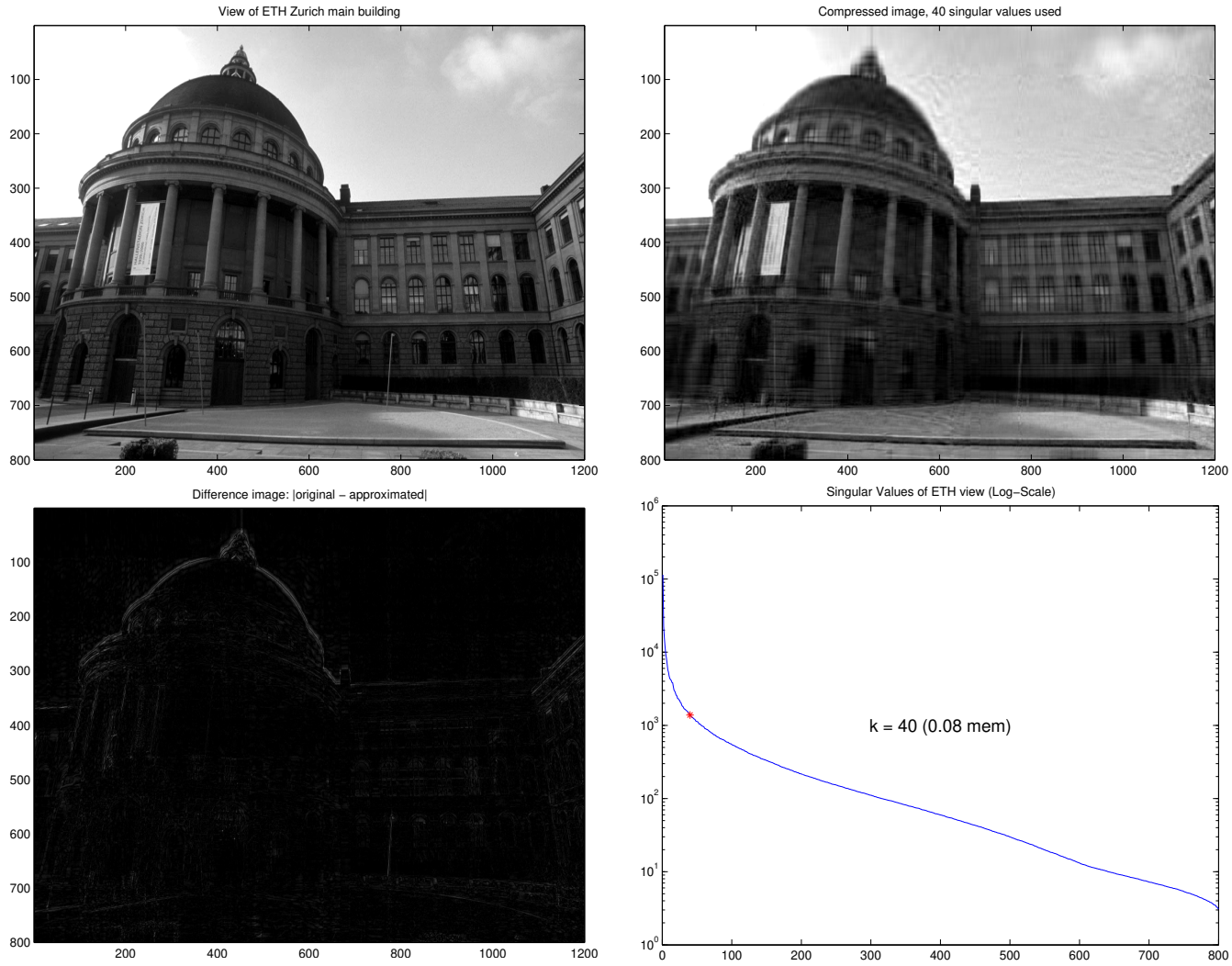
$$\text{in Frobenius norm:} \quad \left\|\mathbf{A}-\mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^{\mathrm{H}}\right\|_F^2 = \sum_{j=k+1}^{\min\{m,n\}}\sigma_j^2 . \tag{3.4.4.23}$$

This provides precise information about the best approximation error for rank-k matrices. In particular, the decay of the singular values of the matrix governs the convergence of the rank-k best approximation error as k increases. ⌟

EXAMPLE 3.4.4.24 (Image compression) A rectangular greyscale image composed of $m \times n$ pixels (greyscale, BMP format) can be regarded as a matrix $\mathbf{A} \in \mathbb{R}^{m,n}$, $(\mathbf{A})_{i,j} \in \{0,\ldots,255\}$, *cf.* Ex. 9.3.2.1. Thus low-rank approximation of the image matrix is a way to compress the image.

Thm. 3.4.4.19 ➢ best rank-k approximation of image: $\widetilde{\mathbf{A}} = \mathbf{U}_k \Sigma_k \mathbf{V}^\top$

Of course, the matrices $\mathbf{U}_l$, $\mathbf{V}_k$, and $\mathbf{\Sigma}_k$ are available from the economical (thin) SVD (3.4.1.5) of $\mathbf{A}$.



Note that there are better and faster ways to compress images than SVD (JPEG, Wavelets, etc.) ⌟

Review question(s) 3.4.4.25 (Best low-rank approximation)

(Q3.4.4.25.A) Show that for $\mathbf{A} \in \mathbb{R}^{m,n}$ and any *orthogonal* $\mathbf{Q} \in \mathbb{R}^{m,m}$

$$\|\mathbf{QA}\|_F = \|\mathbf{A}\|_F \,,$$

where $\|\cdot\|_F$ is the Frobenius norm of a matrix.

Definition 3.4.4.17. Frobenius norm

The **Frobenius norm** of $\mathbf{A} \in \mathbb{K}^{m,n}$ is defined as

$$\|\mathbf{A}\|_F^2 := \sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2 .$$

(Q3.4.4.25.B) Show that for any $\mathbf{A} \in \mathbb{R}^{m,n}$, with singular values σ_j, $j = 1, p := \ldots \min\{m,n\}$, holds

$$\|\mathbf{A}\|_F^2 = \sum_{j=1}^{p} \sigma_j^2 .$$

(Q3.4.4.25.C) Sketch the implementation of a C++ function

```
std::pair<Eigen::MatrixXd,Eigen::MatrixXd>
  lowRankApprox(const Eigen::MatrixXd &A, double tol);
```

that computes a matrix $\widetilde{A} \in \mathbb{R}^{m,n}$ of *minimal rank* $r \in \{1, \ldots, \min\{m,n\}\}$ such that

$$\left\| \mathbf{A} - \widetilde{\mathbf{A}} \right\|_2 \leq \texttt{tol} \cdot \|\mathbf{A}\|_2 .$$

The function should return $\widetilde{\mathbf{A}}$ in factorized form $\widetilde{\mathbf{A}} = \mathbf{X}\mathbf{Y}^T$ as a tuple of matrix factors $\mathbf{X} \in \mathbb{R}^{m,r}$, $\mathbf{Y} \in \mathbb{R}^{n,r}$.

(Q3.4.4.25.D) [Sum of squares of singular values] Given a matrix $\mathbf{M} \in \mathbb{R}^{n,m}$, $m,n \in \mathbb{R}$, represented by a C++ object `M` with an entry access operator

```
double operator () (unsigned int i, unsigned int j) const;
```

write a C++ code snippet that computes the sum of the squares of the singular values of $\mathbf{M}$.

Hint. Remember

Frobenius norm and SVD: $\boxed{\|\mathbf{A}\|_F^2 = \sum_{j=1}^{r} \sigma_j^2}$ (3.4.4.18)

(Q3.4.4.25.E) [Question (**Q**3.4.4.25.D) cnt'd: Sum of σ_i^4] Given a matrix $\mathbf{M} \in \mathbb{R}^{n,m}$, $m,n \in \mathbb{R}$, represented by a C++ object `M` with an entry access operator

```
double operator () (unsigned int i, unsigned int j) const;
```

draft a C++ code snippet that computes

$$s := \sum_{i=1}^{p} \sigma_i^4 , \quad p := \min\{m,n\} , \quad \{\sigma_i\}_{i=1}^{p} \hat{=} \text{ singular values of } \mathbf{M} .$$

△

3.4.4.3 Principal Component Data Analysis (PCA)

Video tutorial for Section 3.4.4.3 "Principal Component Data Analysis (PCA)": (28 minutes)
Download link, tablet notes

→ review questions 3.4.4.51

EXAMPLE 3.4.4.26 (Trend analysis) The objective is to extract information in the form of hidden "trends" from data.

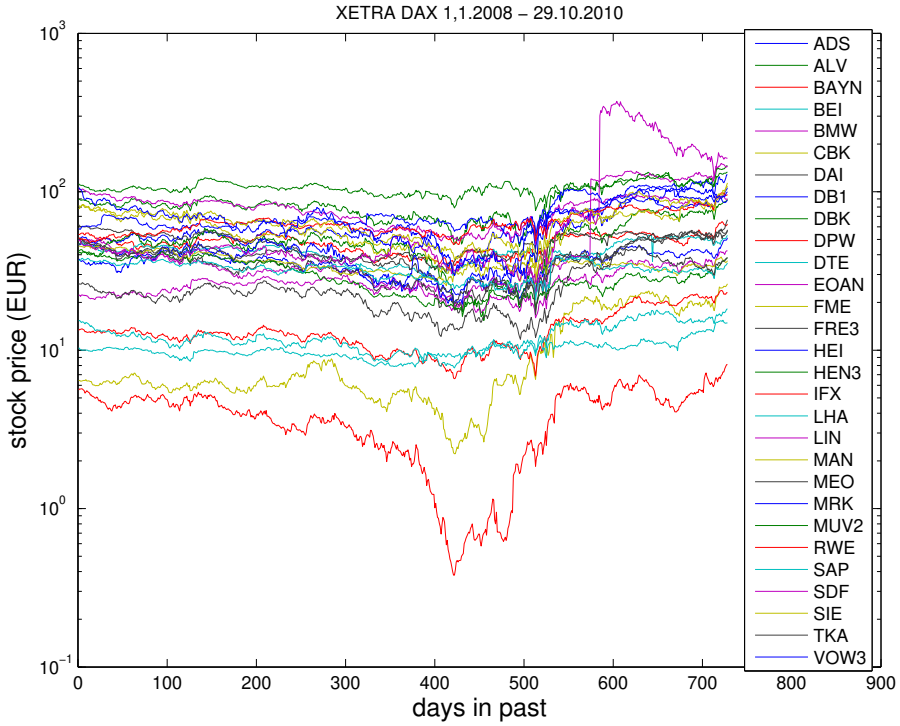


Fig. 86

We are given time series data:

◁ (end of day) stock prizes $\hat{=}$ n data vectors $\in \mathbb{R}^m$

Rephrased in the language of linear algebra:

Are there underlying governing trends ?

$\updownarrow$

Are there a few vectors $\mathbf{u}_1, \ldots, \mathbf{u}_p$, $p \ll n$, such that, *approximately*, all other data vectors $\in \mathrm{Span}\{\mathbf{u}_1, \ldots, \mathbf{u}_p\}$?

EXAMPLE 3.4.4.27 (Classification from measured data) Data vectors belong to different classes, where those in the same class are "qualitatively similar" in the sense that they are small (random) perturbations of a typical data vector. The task is to tease out the typical data patterns and tell which class every data vector belongs to.

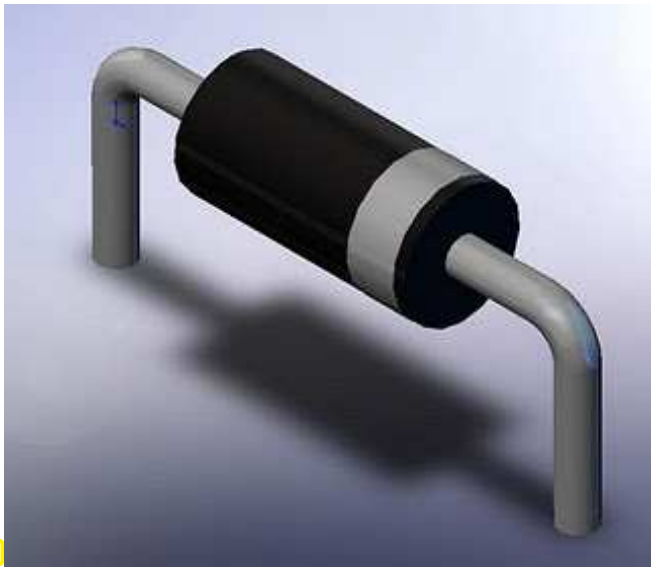


Fig. 87

Given: measured U-I characteristics of n diodes in a box
(data points $(U_j, I_j^{(k)})$, $j = 1, \ldots, m$, $k = 1, \ldots, n$)

Classification problem: find out
- how many different types of diodes in box,
- the U-I characteristic of each type.

Measurement errors !
Manufacturing tolerances !

The following plots display possible ("synthetic") measured data for two types of diodes; measurement errors and manufacturing tolerances taken into account by additive (Gaussian) random perturbations (noise).

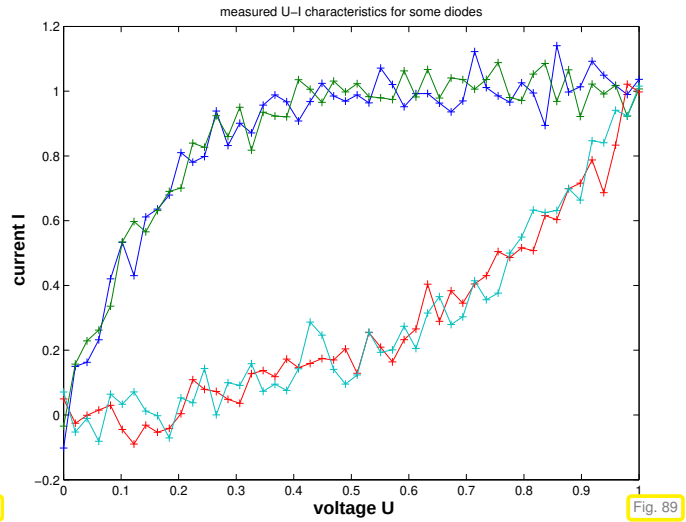


Fig. 88

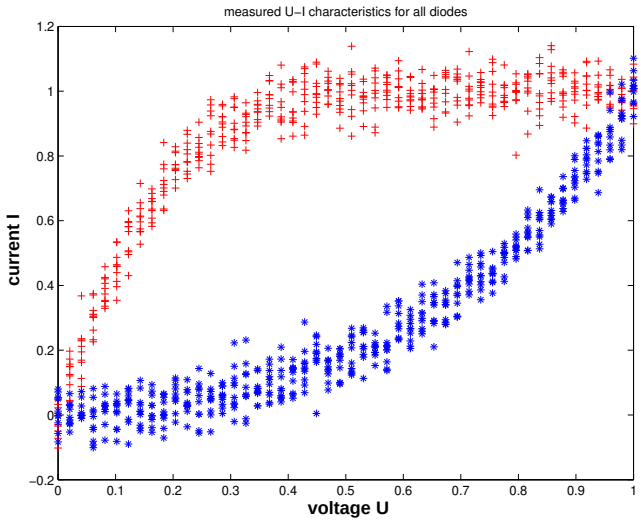


Fig. 89

Ex. 3.4.4.26 and Ex. 3.4.4.27 present typical tasks that can be tackled by principal component analysis. Now we give an abstract description as a problem of linear algebra.

Given: n data points $\mathbf{a}_j \in \mathbb{R}^m$, $j = 1, \ldots, n$, in m-dimensional (feature) space
(e.g., $\mathbf{a}_j$ may represent a finite *time series* or a *measured relationship* of physical quantities)

In Ex. 3.4.4.26: $n \mathrel{\hat{=}}$ number of stocks,
$m \mathrel{\hat{=}}$ number of days, for which stock prices are recorded

✦ Extreme case: all stocks follow exactly *one* trend

$$\leftrightarrow \quad \mathbf{a}_j \in \operatorname{Span}\{\mathbf{u}\} \quad \forall j = 1, \ldots, n \,,$$

for a trend vector $\mathbf{u} \in \mathbb{R}^m$, $\|\mathbf{u}\|_2 = 1$.

✦ Unlikely case: all stocks prices are governed by $p < n$ trends:

$$\leftrightarrow \quad \mathbf{a}_j \in \operatorname{Span}\{\mathbf{u}_1, \ldots, \mathbf{u}_p\} \quad \forall j = 1, \ldots, n \,, \tag{3.4.4.28}$$

with orthonormal trend vectors $\mathbf{u}_i \in \mathbb{R}^m$, $i = 1, \ldots, p$.

Why unlikely **?** Small *random fluctuations* will be present in each stock prize
Why orthonormal **?** Trends should be as "independent as possible" (minimally correlated)

Expressed using the terminology linear algebra:

$$\text{(3.4.4.28)} \quad \Leftrightarrow \quad \begin{array}{c} \operatorname{rank}(\mathbf{A}) = p \quad \text{for} \quad \mathbf{A} := [\mathbf{a}_1, \ldots, \mathbf{a}_n] \in \mathbb{R}^{m,n} \,, \\ \mathcal{R}(\mathbf{A}) = \operatorname{Span}\{\mathbf{u}_1, \ldots, \mathbf{u}_p\} \end{array} \tag{3.4.4.29}$$

✦ Realistic: stock prizes *approximately* follow a few trends

$$\mathbf{a}_j \in \operatorname{Span}\{\mathbf{u}_1, \ldots, \mathbf{u}_p\} + \text{"small perturbations"} \quad \forall j = 1, \ldots, m \,,$$

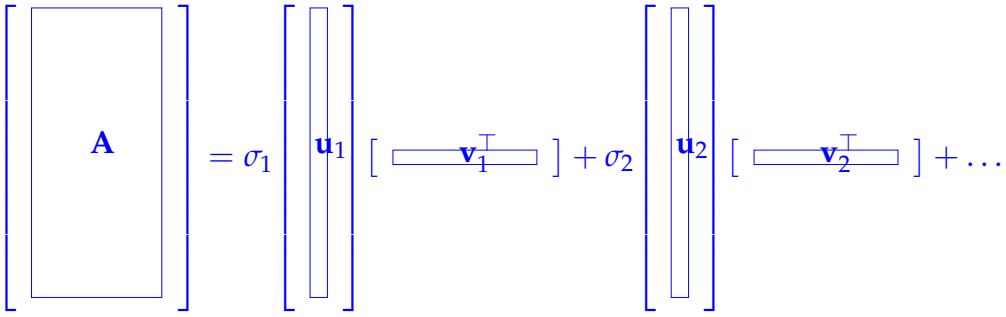
with orthonormal trend vectors $\mathbf{u}_i$, $i = 1, \ldots, p$.

Task (PCA): determine (minimal) p and orthonormal trend vectors $\mathbf{u}_i$, $i = 1, \ldots, p$

Now singular value decomposition (SVD) according to Def. 3.4.1.3 comes into play, because Lemma 3.4.1.13 tells us that it can supply an orthonormal basis of the image space of a matrix, *cf.* Code 3.4.2.9.

Issue: how to deal with (small, random) perturbations **?**

Recall Rem. 3.4.1.7, (3.4.1.9): If $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{H}^\top$ is the SVD of $\mathbf{A} \in \mathbb{R}^{m,n}$, then ($\mathbf{u}_j \hat{=}$ columns of $\mathbf{U}$, $\mathbf{v}_j \hat{=}$ columns of $\mathbf{V}$)



This already captures the case (3.4.4.28) and we see that the columns of $\mathbf{U}$ supply the trend vectors we are looking for!

❶ no perturbations:

$$\text{SVD:}\quad \mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{H}} \quad \text{satisfies} \quad \sigma_1 \geq \sigma_2 \geq \ldots \sigma_p > \sigma_{p+1} = \cdots = \sigma_{\min\{m,n\}} = 0\,,$$
$$\text{orthonormal trend vectors} \quad (\mathbf{U})_{:,1}, \ldots, (\mathbf{U})_{:,p}\,.$$

❷ with perturbations:

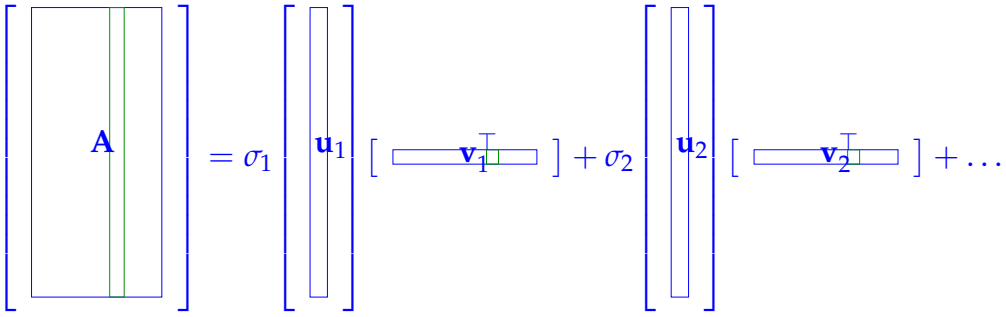
$$\text{SVD:}\quad \mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{H}} \quad \text{satisfies} \quad \sigma_1 \geq \sigma_2 \geq \ldots \sigma_p \gg \sigma_{p+1} \approx \cdots \approx \sigma_{\min\{m,n\}} \approx 0\,,$$
$$\text{orthonormal trend vectors} \quad (\mathbf{U})_{:,1}, \ldots, (\mathbf{U})_{:,p}\,.$$

If there is a pronounced gap in distribution of the singular values, which separates p large from $\min\{m,n\} - p$ relatively small singular values, this hints that $\mathcal{R}(\mathbf{A})$ has essentially dimension p. It depends on the application what one accepts as a "pronounced gap".

Frequently used criterion:

$$p = \min\left\{ q\colon \sum_{j=1}^{q} \sigma_j^2 \geq (1-\tau) \sum_{j=1}^{\min\{m,n\}} \sigma_j^2 \right\} \quad \text{for} \quad \tau \ll 1\,. \tag{3.4.4.30}$$

What is the Information carried by $\mathbf{V}$ in PCA context **?**



j-th data set ($\leftrightarrow$ time series #j) in j-th column of $\mathbf{A}$

$$(3.4.1.9) \quad \Rightarrow \quad (\mathbf{A})_{:,j} = \sigma_1 \mathbf{u}_1 (\mathbf{v}_1)_j + \sigma_2 \mathbf{u}_2 (\mathbf{v}_2)_j + \ldots$$

The j-th row of $\mathbf{V}$ (up to the p-th component) gives the weights with which the p identified trends contribute to data set j.

EXAMPLE 3.4.4.31 (PCA of stock prices → Ex. 3.4.4.26cnt'd) Stock prices are given as a large matrix $\mathbf{A} \in \mathbb{R}^{m,n}$:

columns of $\mathbf{A}$ $\rightarrow$ time series of end of day stock prices of individual stocks
rows of $\mathbf{A}$ $\rightarrow$ closing prices of DAX stocks on a particular day

The data were obtained from Yahoo Finance in 2016:

```
!/bin/csh
foreach i (ADS  ALV BAYN BEI BMW CBK DAI DBK DB1 LHA DPW DTE EOAN FRE3 \
           FME HEI HEN3 IFX SDF LIN MAN MRK MEO MUV2 RWE SAP SIE TKA VOW3)
wget -O "i".csv "http://ichart.finance.yahoo.com/table.csv?s=i.DE&a=00&b=1&
␣␣␣␣␣c=2008&d=09&e=30&f=2010&g=d&ignore=.csv"
sed -i -e 's/-/,/g' "i".csv
end
```

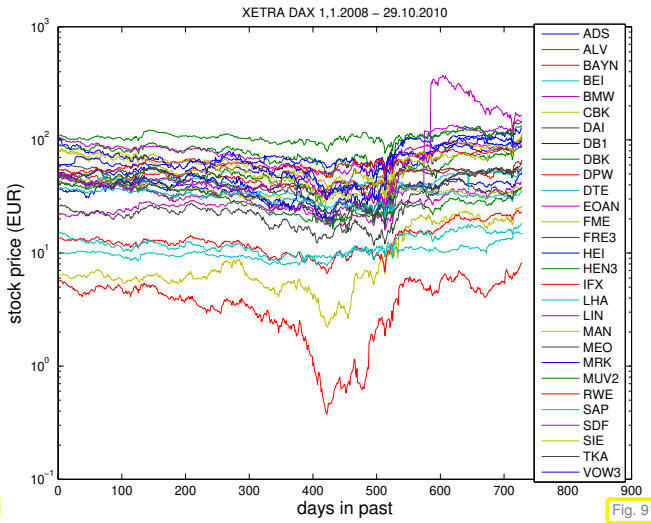


Fig. 90

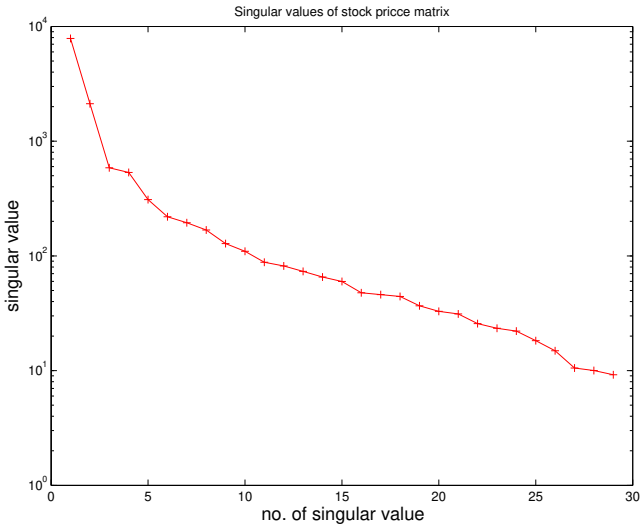


Fig. 91

We observe a pronounced decay of the singular values of $\mathbf{A}$. The plot of Fig. 91 is given in linear-logarithmic scale. The neat alignment of larger singular values indicates approximate *exponential decay* of the singular values.

➢ a few trends (corresponding to a few of the largest singular values) govern the time series.

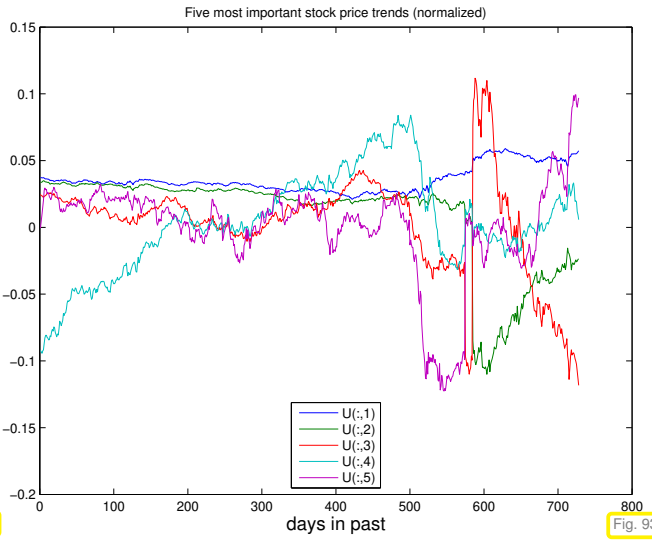


Fig. 92

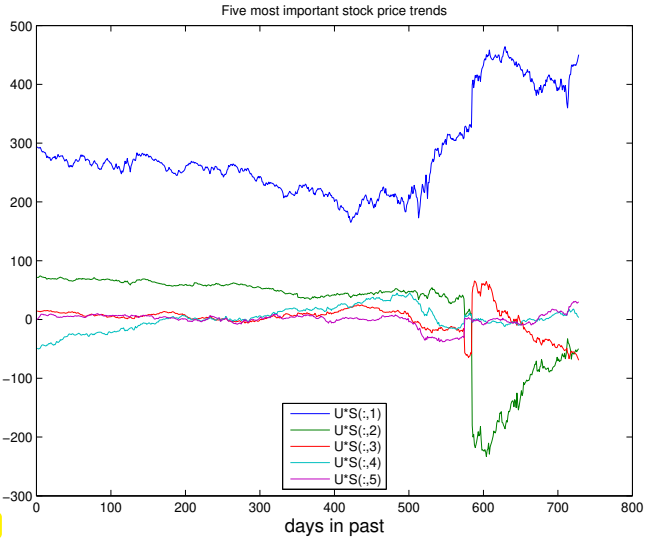


Fig. 93

Columns of $\mathbf{U}$ ($\rightarrow$ Fig. 92) in SVD $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ provide trend vectors, *cf.* Ex. 3.4.4.26 & Ex. 3.4.4.32.

When weighted with the corresponding singular value, the importance of a trend contribution emerges, see Fig. 93

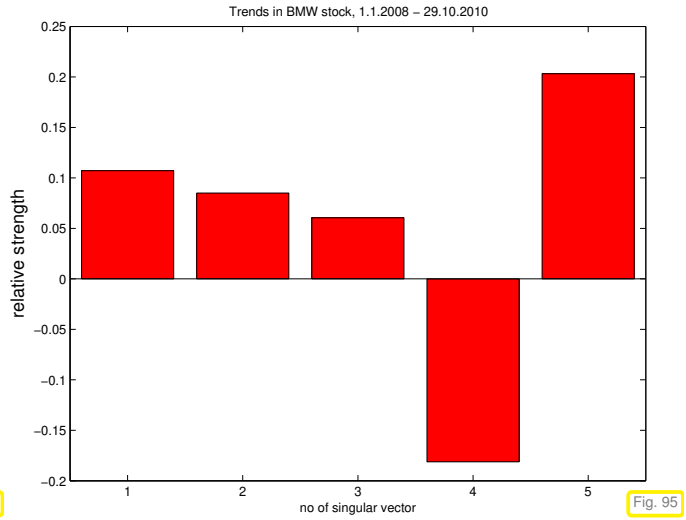


Fig. 94

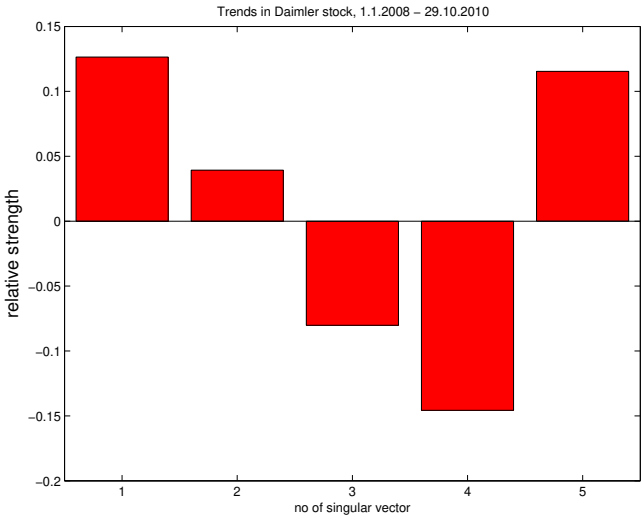


Fig. 95

Stocks of companies from the same sector of the economy should display similar contributions of major trend vectors, because their prices can be expected to be more closely correlated than stock prices in general. This is evident in 94 and Fig. 95 for two car makers.

⌟

EXAMPLE 3.4.4.32 (Principal component analysis for data classification → Ex. 3.4.4.27 cnt'd)

Given: measured U-I characteristics of $n = 20$ unknown diodes, $I(U)$ available for $m = 50$ voltages.

Sought: Number of different types of diodes in batch and reconstructed U-I characteristic for each type.

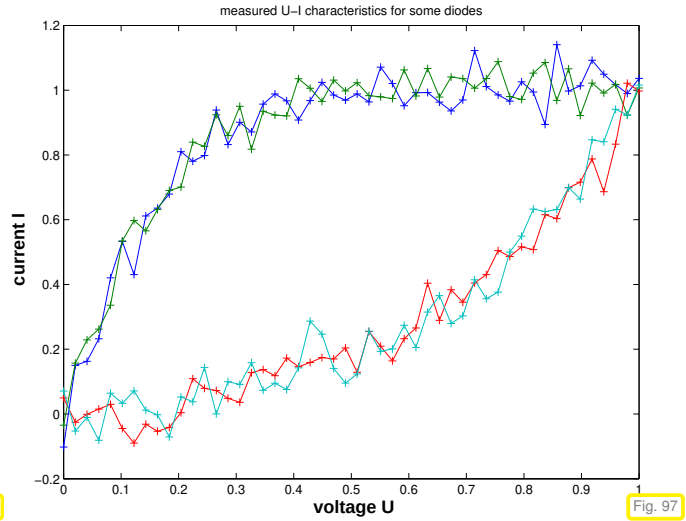


Fig. 96

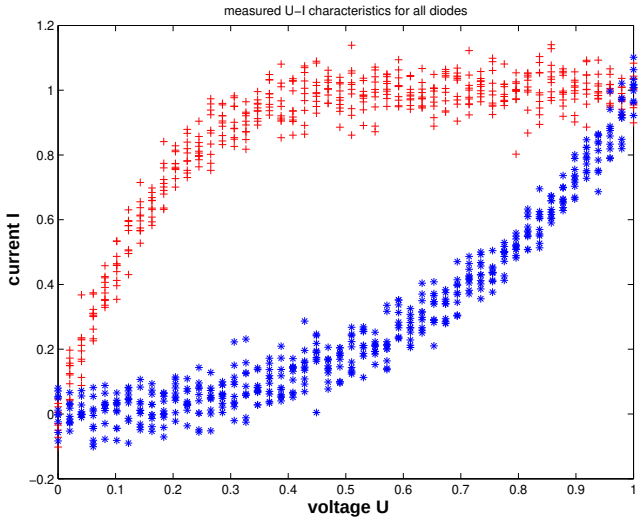


Fig. 97

Data matrix $\mathbf{A} \in \mathbb{R}^{m,n}$, $m \gg n$:

Rows $\mathbf{A}$ → series of measurements for different diodes (times/locations etc.),
Columns of $\mathbf{A}$ → measured values corresponding to one diode (time/location etc.).

Goal of PCA: detect linear correlations between columns of $\mathbf{A}$

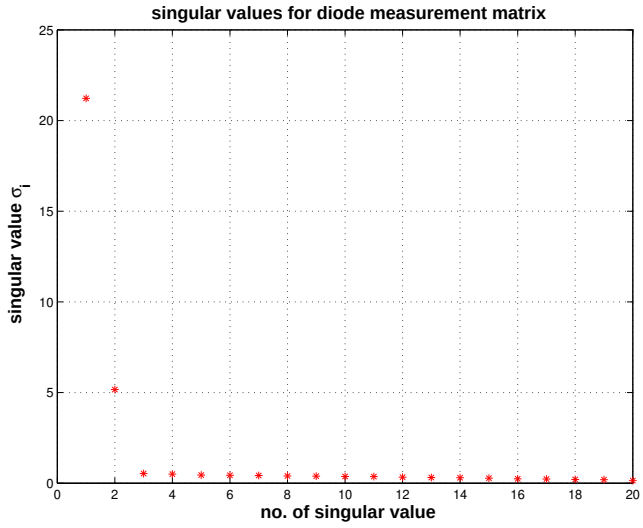


Fig. 98

← distribution of singular values of matrix
two dominant singular values !

measurements display linear correlation with **two** principal components

two types of diodes in batch

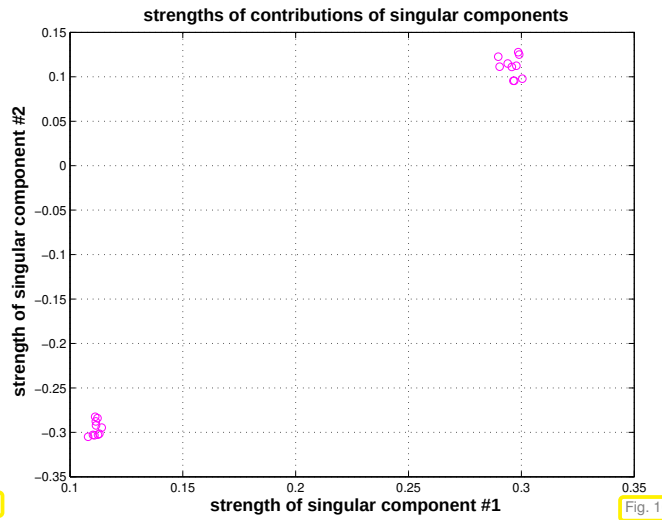


Fig. 99

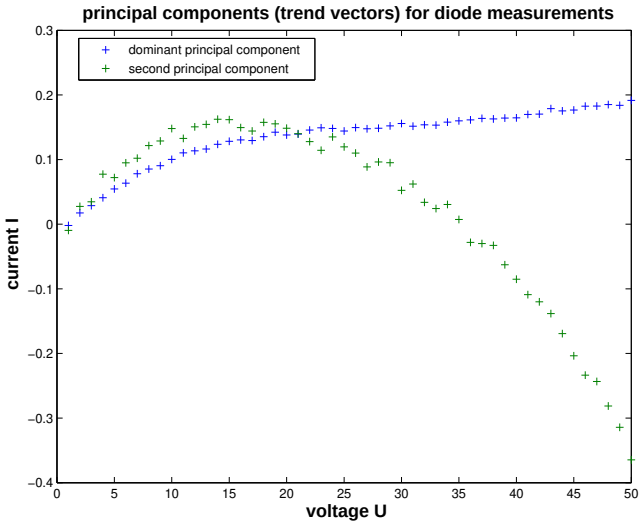


Fig. 100

Observations:

- First two rows of $\mathbf{V}$-matrix specify strength of contribution of the two leading principal components to each measurement

  ➢ Points $(\mathbf{V})_{:,1:2}$, which correspond to different diodes are neatly clustered in $\mathbb{R}^2$. To determine the type of diode i, we have to identify the cluster to which the point $((\mathbf{V})_{i,1}, \mathbf{V}_{i,2})$ belongs (→ cluster analysis, course "machine learning", see Rem. 3.4.4.43 below).

- The principal components themselves do not carry much useful information in this example.

⌟

EXAMPLE 3.4.4.33 (Data points (almost) confined to a subspace) More abstractly, above we tried to identify a subspace to which all data points $\mathbf{a}_i$ were "close". We saw that the SVD of

Data points • $\leftrightarrow \mathbf{a}_i \in \mathbb{R}^3$ "almost" located on on a plane ▷

Non-zero singular values of $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_n]$:

$$\begin{array}{c} 3.1378 \\ 1.8092 \\ \hline 0.1792 \end{array}$$

The third singular value is much smaller, which hints that the data points approximately lie in a 2D subspace spanned by the two first singular vectors of $\mathbf{A}$. ⌟

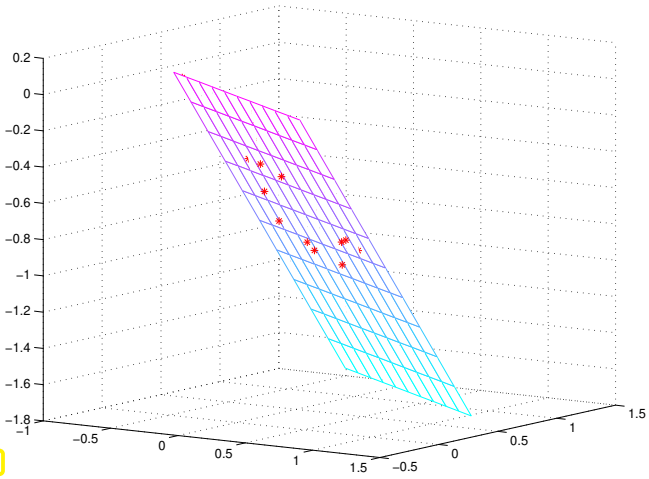


Fig. 101

§3.4.4.34 (Proper orthogonal decomposition (POD)) In the previous Ex. 3.4.4.33 we saw that the singular values of a matrix whose columns represent data points $\in \mathbb{R}^m$ tell us whether these points are all "approximately located" in a lower-dimensional subspace $V \subset \mathbb{R}^m$. This is linked to the following problem:

Problem of proper orthogonal decomposition (POD):

Given: Data points $\mathbf{a}_1, \ldots, \mathbf{a}_n \in \mathbb{R}^m$, $m, n \in \mathbb{N}$
Sought: For $k \leq \min\{m, n\}$, find a *subspace* $U_k \subset \mathbb{R}^m$ such that

$$U_k = \operatorname*{argmin}_{W \subset \mathbb{R}^m, \dim W = k} \sum_{j=1}^{n} \inf_{\mathbf{w} \in W} \|\mathbf{a}_j - \mathbf{w}\|_2^2 , \tag{3.4.4.35}$$

that is, we seek that k-dimensional subspace U_k of $\mathbb{R}^m$ for which the sum of squared distances of the data points to U_k is minimal.

We have already seen a similar problem in Ex. 3.4.4.5. For $m = 2$ we want to point out the difference to linear regression:

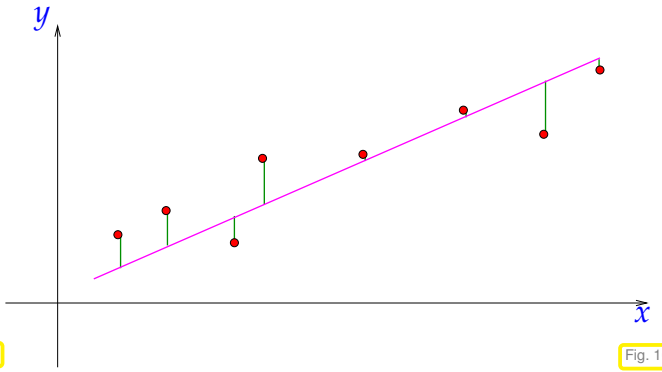


Fig. 102

Linear regression: Minimize the sum of squares of *vertical distances*.

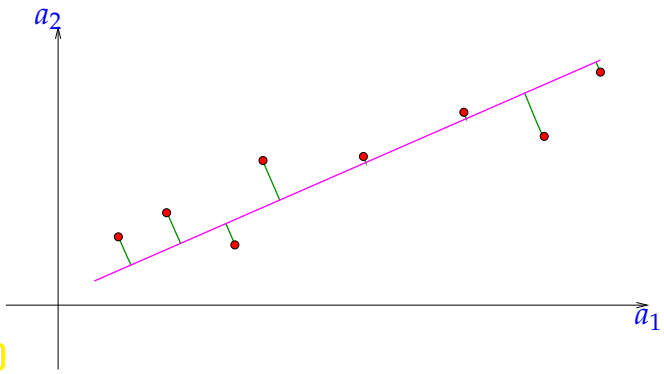


Fig. 103

POD: Minimize the sum of squares of *(minimal) distances*.

By finding a k-dimensional subspace we mean finding a, preferably orthonormal, basis of that subspace. Let us assume that $\{\mathbf{w}_1, \ldots, \mathbf{w}_k\}$ is an orthonormal basis (ONB) of a k-dimensional subspace $W \subset \mathbb{R}^m$. Then the orthogonal projection $\mathrm{P}_W\mathbf{x}$ of a point $\mathbf{x} \in \mathbb{R}^m$ onto W is given by

$$\mathrm{P}_W\mathbf{x} = \sum_{j=1}^{k} (\mathbf{w}_j^\top \mathbf{x})\mathbf{w}_j = \mathbf{W}\mathbf{W}^\top \mathbf{x} , \tag{3.4.4.36}$$

where $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_k] \in \mathbb{R}^{m,k}$. This formula is closely related to the normal equations for a linear

least squares problem, see Thm. 3.1.2.1 and § 3.1.1.8 for a visualization of an orthogonal projection. $\|\mathbf{x} - \mathrm{P}_W\mathbf{x}\|_2$ is the (minimal) distance of $\mathbf{x}$ to W.

Hence, again writing $\mathbf{W} \in \mathbb{R}^{m,k}$ for the matrix whose columns form an ONB ($\Rightarrow \mathbf{W}^\top\mathbf{W} = \mathbf{I}$) of $W \subset \mathbb{R}^m$, we have

$$\sum_{j=1}^{n} \inf_{w\in W} \|\mathbf{a}_j - \mathbf{w}\|_2^2 = \sum_{j=1}^{n} \left\|\mathbf{a}_j - \mathbf{W}\mathbf{W}^\top\mathbf{a}_j\right\|_2^2 = \left\|\mathbf{A} - \mathbf{W}\mathbf{W}^\top\mathbf{A}\right\|_F^2 , \tag{3.4.4.37}$$

where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix, see Def. 3.4.4.17, and $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_n] \in \mathbb{R}^{m,n}$.

Note that $\operatorname{rank}(\mathbf{W}\mathbf{W}^\top\mathbf{A}) \le k$. Let us write $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ for the SVD of $\mathbf{A}$, and $\mathbf{U}_k \in \mathbb{R}^{m,k}$, $\mathbf{V}_k \in \mathbb{R}^{n,k}$, and $\mathbf{\Sigma}_k \in \mathbb{R}^{k,k}$ for the truncated SVD-factors of $\mathbf{A}$ as introduced in Thm. 3.4.4.19. Then the rank-k best approximation result of that theorem implies

$$\left\|\mathbf{A} - \mathbf{W}\mathbf{W}^\top\mathbf{A}\right\|_F \ge \|\mathbf{A} - \mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k\|_F \quad \forall \mathbf{W} \in \mathbb{R}^{m,k} , \quad \mathbf{W}^\top\mathbf{W} = \mathbf{I} .$$

In fact, we can find $\mathbf{W} \in \mathbb{R}^{m,k}$ with orthonormal columns that realizes the minimum: just choose $\mathbf{W} := \mathbf{U}_k$ and verify

$$\mathbf{W}\mathbf{W}^\top\mathbf{A} = \mathbf{U}_k\mathbf{U}_k^\top\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = \mathbf{U}_k\begin{bmatrix}\mathbf{I}_k & \mathbf{O}\end{bmatrix}\mathbf{\Sigma}\mathbf{V}^\top = \mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^\top .$$

Theorem 3.4.4.38. Solution of POD problem

The subspace U_k spanned by the first k left singular vectors of $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_n] \in \mathbb{R}^{m,n}$ solves the POD problem (3.4.4.35)*:*

$$U_k = \mathcal{R}\big((\mathbf{U})_{:,1:k}\big) \quad \textit{with} \quad \mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top \quad \textit{the SVD of } \mathbf{A}.$$

Appealing to (3.4.4.23), the sum of the squared distances can be obtained as the sum of the squares of the remaining singular values $\sigma_{k=1}, \ldots, \sigma_p$, $p := \min\{m, n\}$, of $\mathbf{A}$:

$$\sum_{j=1}^{n} \inf_{w\in U_k} \|\mathbf{a}_j - \mathbf{w}\|_2^2 = \sum_{\ell=k+1}^{p} \sigma_\ell^2 . \tag{3.4.4.39}$$

As a consequence, the decay of the singular values again predicts how close the data points are to the POD subspaces U_k, $k = 1, \ldots, p-1$. ⌟

EXAMPLE 3.4.4.40 (Principal axis of a point cloud) Given $m > 2$ points $\mathbf{x}_j \in \mathbb{R}^k$, $j = 1, \ldots, m$, in k-dimensional space, we ask what is the "longest" and "shortest" diameter $\mathbf{d}_+$ and $\mathbf{d}_-$. This question can be stated rigorously in several different ways: here we ask for directions for which the point cloud will have maximal/minimal variance, when projected onto that direction:

$$\mathbf{d}_+ := \operatorname*{argmax}_{\|\mathbf{v}\|=1} Q(\mathbf{v}) ,$$
$$\mathbf{d}_- := \operatorname*{argmin}_{\|\mathbf{v}\|=1} Q(\mathbf{v}) , \qquad Q(\mathbf{v}) := \sum_{j=1}^{m} |(\mathbf{x}_i - \mathbf{c})^\top\mathbf{v}|^2 , \quad \mathbf{c} = \frac{1}{m}\sum_{j=1}^{m}\mathbf{x}_j . \tag{3.4.4.41}$$

The directions $\mathbf{d}_+$, $\mathbf{d}_-$ are called the principal axes of the point cloud, a term borrowed from mechanics and connected with the axes of inertia of an assembly of point masses.

Principal axes of a point cloud in 2D ▷

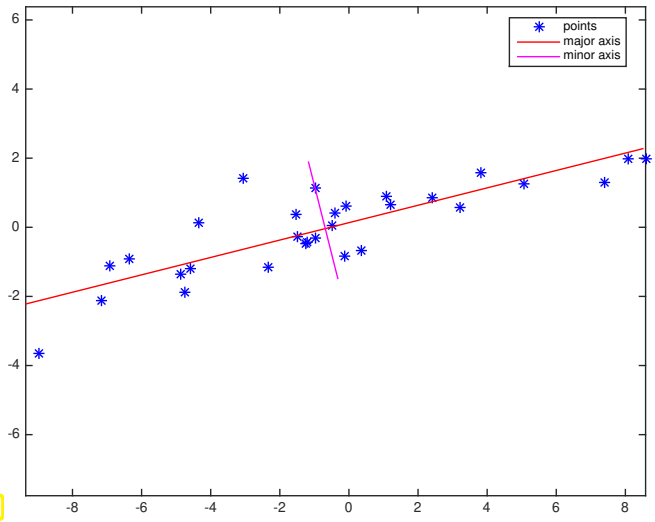


Fig. 104

$\mathbf{d}_+$, $\mathbf{d}_-$ can be computed by computing the extremizers of $\mathbf{x} \mapsto \|\mathbf{A}\mathbf{x}\|_2$ with

$$\mathbf{A} = \begin{bmatrix} (\mathbf{x}_1 - \mathbf{c})^\top \\ \vdots \\ (\mathbf{x}^m - \mathbf{c})^\top \end{bmatrix} \in \mathbb{R}^{m,k} \quad \Rightarrow \quad \begin{aligned} \mathbf{d}_+ &= \operatorname*{argmax}_{\|\mathbf{v}\|=1} \|\mathbf{A}\mathbf{x}\|_2 \,, \\ \mathbf{d}_- &= \operatorname*{argmin}_{\|\mathbf{v}\|=1} \|\mathbf{A}\mathbf{x}\|_2 \,, \end{aligned} \tag{3.4.4.42}$$

on $\{\mathbf{x} \in \mathbb{R}^k : \|\mathbf{x}\|_2 = 1\}$, using the SVD-based method presented in Section 3.4.4.1. ⌟

Remark 3.4.4.43 (Algorithm for cluster analysis) In data classification as presented in Ex. 3.4.4.32, after we have identified the p main trends ($\leftrightarrow$ singular values and left singular vectors) and how much they contribute to every data point ($\leftrightarrow$ right singular vectors), the last step is to perform a cluster analysis based on the right singular vectors $\mathbf{v}_1, \ldots, \mathbf{v}_p$.

Now we study the abstract problem of cluster analysis:

Given: ✦ N data points $\mathbf{x}_i \in \mathbb{R}^k$, $i = 1, \ldots, N$,

✦ Assume: number n of desired clusters is known in advance.

Sought: Partitioning of index set $\{1, \ldots, N\} = I_1 \cup \cdots \cup I_n$, achieving minimal mean least squares error

$$\mathrm{mlse} := \sum_{l=1}^{n} \sum_{i \in I_l} \|\mathbf{x}_i - \mathbf{m}_l\|_2^2 \,, \quad \mathbf{m}_l = \frac{1}{\sharp I_l} \sum_{i \in I_l} \mathbf{x}_i \,. \tag{3.4.4.44}$$

The subsets $\{\mathbf{x}_i : i \in I_l\}$ are called the clusters. The points $\mathbf{m}_l$ are their centers of gravity.

The Algorithm involves two components:

❶ Splitting of a cluster by separation along its principal axis, see Ex. 3.4.4.40 and Code 3.4.4.48:

$$\mathbf{a}_l := \operatorname*{argmax}_{\|\mathbf{v}\|_2=1} \{ \sum_{i \in I_l} |(\mathbf{x}_i - \mathbf{m}_l)^\top \mathbf{v}|^2 \} \,. \tag{3.4.4.45}$$

Relies on the algorithm from Ex. 3.4.4.40, see Code 3.4.4.48.

❷ Improvement of clusters using the Lloyd-Max algorithm, see Code 3.4.4.49. It involves two steps in turns:

(a) Given centers of gravity $\mathbf{m}_l$ redistribute points according to

$$I_l := \{ i \in \{1, \ldots, N\} \colon \|\mathbf{x}_i - \mathbf{m}_l\|_2 \leq \|\mathbf{x}_i - \mathbf{m}_k\|_2 \; \forall k \neq l \} \,, \tag{3.4.4.46}$$

that is, we assign each point to the nearest center of gravity, see Code 3.4.4.49.

(b) Recompute centers of gravity

$$\mathbf{m}_l = \frac{1}{\sharp I_l} \sum_{i \in I_l} \mathbf{x}_i \,. \tag{3.4.4.47}$$

We start with a single cluster, and then do repeated splitting (❶) and cluster rearrangement (❷) until we have reached the desired final number n of clusters, see Code 3.4.4.50.

C++-code 3.4.4.48: Principal axis point set separation ↪ GITLAB

```
// Separation of a set of points whose coordinates are stored in the
// columns of X according to their location w.r.t.  the principal axis
std::pair<VectorXi, VectorXi> princaxissep(const MatrixXd & X){
  const Eigen::Index N = X.cols(); // no.  of points
  const VectorXd g = X.rowwise().sum() / N; // Center of gravity, cf.  (3.4.4.47)
  MatrixXd Y = X - g.replicate(1,N); // Normalize point coordinates.
  // Compute principal axes, cf.  (3.4.4.45) and (3.4.4.3).  Note that the
  // SVD of a symmetric matrix is available through an orthonormal
  // basis of eigenvectors.
  const Eigen::SelfAdjointEigenSolver<MatrixXd> es(Y*Y.transpose());
  // Major principal axis
  Eigen::VectorXd a = es.eigenvectors().rightCols<1>();
  // Coordinates of points w.r.t.  to major principal axis
  Eigen::VectorXd c = a.transpose()*Y;
  // Split point set according to locations of projections on principal
     axis
  // std::vector with indices to prevent resizing of matrices
  std::vector<int> i1;
  std::vector<int> i2;
  for(int i = 0; i < c.size(); ++i){
    if(c(i) >= 0) {
      i1.push_back(i);
    }
    else {
      i2.push_back(i);
    }
  }
  // return the mapped std::vector as Eigen::VectorXd
  return {
    VectorXi::Map(i1.data(), static_cast<Eigen::Index>(i1.size())),
    VectorXi::Map(i2.data(), static_cast<Eigen::Index>(i2.size()))
  };
}
```

C++-code 3.4.4.49: Lloyd-Max algorithm for cluster indentification ↪ GITLAB

```
template <class Derived>
std::tuple<double, VectorXi, VectorXd> distcomp(const MatrixXd & X, const
    MatrixBase<Derived> & C){
  // Compute squared distances
  // d.row(j) = squared distances from all points in X to cluster j
  MatrixXd d(C.cols(), X.cols());
  for(int j = 0; j < C.cols(); ++j){
       MatrixXd Dv = X - C.col(j).replicate(1, X.cols());
       d.row(j) = Dv.array().square().colwise().sum();
  }
  // Compute minimum distance point association and sum of minimal
     squared distances
  VectorXi idx(d.cols());  VectorXd mx(d.cols());
```

```
13    for(int j = 0; j < d.cols(); ++j){
14        // mx(j) tells the minimal squared distance of point j to the
              nearest cluster
15        // idx(j) tells to which cluster point j belongs
16        mx(j) = d.col(j).minCoeff(&idx(j));
17    }
18    const double sumd = mx.sum();    // sum of all squared distances
19    // Computer sum of squared distances within each cluster
20    VectorXd cds(C.cols()); cds.setZero();
21    for(int j = 0; j < idx.size(); ++j){ // loop over all points
22        cds(idx(j)) += mx(j);
23      }
24    return std::make_tuple(sumd, idx, cds);
25  }
26
27  // Lloyd-Max iterative vector quantization algorithm for discrete point
28  // sets; the columns of X contain the points x_i, the columns of
29  // C initial approximations for the centers of the clusters.  The final
30  // centers are returned in C, the index vector idx specifies
31  // the association of points with centers.
32  template <class Derived>
33  void lloydmax(const MatrixXd & X, MatrixBase<Derived> & C, VectorXi & idx, VectorXd
        & cds, const double tol = 0.0001){
34    const Eigen::Index k = X.rows(); // dimension of space
35    const Eigen::Index N = X.cols(); // no.  of points
36    const Eigen::Index n = C.cols(); // no.  of clusters
37    if(k != C.rows()) {
38          throw std::logic_error("dimension mismatch");
39      }
40    double sd_old = std::numeric_limits<double>::max();
41    double sd = NAN;
42    std::tie(sd, idx, cds) = distcomp(X,C);
43    // Terminate, if sum of squared minimal distances has not changed much
44    while( (sd_old-sd)/sd > tol ){
45        // Compute new centers of gravity according to (3.4.4.47)
46        MatrixXd Ctmp(C.rows(),C.cols()); Ctmp.setZero();
47        // number of points in cluster for normalization
48        VectorXi nj(n); nj.setZero();
49        for(int j = 0; j < N; ++j){ // loop over all points
50            Ctmp.col(idx(j)) += X.col(j);
51            ++nj(idx(j));     // count associated points for normalization
52        }
53        for(int i = 0; i < Ctmp.cols(); ++i){
54            if(nj(i) > 0) {
55                C.col(i) = Ctmp.col(i)/nj(i); // normalization
56              }
57        }
58        sd_old = sd;
59        // Get new minimum association of the points to cluster points
60        // for next iteration
61        std::tie(sd, idx, cds) = distcomp(X,C);
62    }
63  }
64  // Note:  this function is needed to allow a call with an rvalue
65  // && stands for an rvalue reference and allows rvalue arguments
66  // such as C.leftCols(nc) to be passed by reference (C++ 11 feature)
67  template <class Derived>
68  void lloydmax(const MatrixXd & X, MatrixBase<Derived> && C, VectorXi & idx, VectorXd
        & cds, const double tol = 0.0001){
69      lloydmax(X, C, idx, cds, tol);
70  }
```

C++-code 3.4.4.50: Clustering of point set ↬ GITLAB

```
// n-quantization of point set in k-dimensional space based on
// minimizing the mean square error of Euclidean distances.  The
// columns of the matrix X contain the point coordinates, n specifies
// the desired number of clusters.
std::pair<MatrixXd, VectorXi> pointcluster(const MatrixXd & X, const int n){
  const Eigen::Index N = X.cols(); // no.  of points
  const Eigen::Index k = X.rows(); // dimension of space
  // Start with two clusters obtained by principal axis separation
  int nc = 1; // Current number of clusters
  // Initial single cluster encompassing all points
  VectorXi Ibig = VectorXi::LinSpaced(N,0,static_cast<int>(N-1));
  int nbig = 0; // Index of largest cluster
  MatrixXd C(X.rows(), n);    // matrix for cluster midpoints
  C.col(0) = X.rowwise().sum()/N; // center of gravity
  VectorXi idx(N); idx.setOnes();
  // Split largest cluster into two using the principal axis separation
  // algorithm
  while(nc < n){
       VectorXi i1;
       VectorXi i2;
       MatrixXd Xbig(k,Ibig.size());
       for(int i = 0; i < Ibig.size(); ++i) {     // slicing
              Xbig.col(i) = X.col(Ibig(i));
          }
       // separete Xbig into two clusters, i1 and i2 are index vectors
       std::tie(i1, i2) = princaxissep::princaxissep(Xbig);
       // new cluster centers of gravity
       VectorXd c1(k); c1.setZero();
       VectorXd c2(k); c2.setZero();
       for(int i = 0; i < i1.size(); ++i) {
            c1 += X.col(Ibig(i1(i)));
       }
       for(int i = 0; i < i2.size(); ++i){
            c2 += X.col(Ibig(i2(i)));
       }
       c1 /= static_cast<double>(i1.size()); // normalization
       c2 /= static_cast<double>(i2.size());
       C.col(nbig) = c1;
       C.col(nbig+1) = c2;
       ++nc; // Increase number of clusters
       // Improve clusters by Lloyd-Max iteration
       VectorXd cds;     // saves mean square error of clusters
       // Note C.leftCols(nc) is passed as rvalue reference (C++ 11)
       lloydmax::lloydmax(X,C.leftCols(nc),idx,cds);
       // Identify cluster with biggest contribution to mean square error
       cds.maxCoeff(&nbig);
       int counter = 0;
       // update Ibig with indices of points in cluster with biggest
          contribution
       for(int i = 0; i < idx.size(); ++i){
            if(idx(i) == nbig){
                 Ibig(counter) = i;
                 ++counter;
            }
       }
       Ibig.conservativeResize(counter);
```

```
57    }
58    return std::make_pair(C, idx);
59  }
```

⌟

Review question(s) 3.4.4.51 (Principal Component Data Analysis)

(Q3.4.4.51.A) Assume that the (sorted!) singular values σ_j, $j \in \{1,\ldots,\min\{m,n\}\}$, of $\mathbf{A} \in \mathbb{R}^{m,n}$, $m,n \gg 1$, obey the "asymptotic" decay law

$$\sigma_j \approx C\exp(-\alpha j) \quad \text{for some} \quad C > 0,\ \alpha > 0\,.$$

How much do you have to increase the rank of the best low-rank approximation of $\mathbf{A}$ with respect to the *Euclidean matrix norm* in order to reduce the approximation error by a factor of 2?

Can you also answer this question for the Frobenius matrix norm?

△

3.5 Total Least Squares

In the examples of Section 3.0.1 we generally considered overdetermined linear systems of equations $\mathbf{A}\mathbf{x} = \mathbf{b}$, for which only the right hand side vector $\mathbf{b}$ was affected by measurement errors. However, also the entries of the coefficient matrix $\mathbf{A}$ may have been obtained by measurement. This is the case, for instance, in the nodal analysis of electric circuits → Ex. 2.1.0.3. Then, it may be legitimate to seek a "better" matrix based on information contained in the whole linear system. This is the gist of the total least squares approach.

Given: overdetermined linear system of equations $\mathbf{A}\mathbf{x} = \mathbf{b}$, $\mathbf{A} \in \mathbb{R}^{m,n}$, $\mathbf{b} \in \mathbb{R}^m$, $m > n$.

Known: LSE solvable $\Leftrightarrow$ $\mathbf{b} \in \mathrm{Im}(\mathbf{A})$, *if* $\mathbf{A}$, $\mathbf{b}$ *were not perturbed*,

but $\mathbf{A}$, $\mathbf{b}$ are perturbed (measurement errors).

Sought: *Solvable* overdetermined system of equations $\widehat{\mathbf{A}}\mathbf{x} = \widehat{\mathbf{b}}$, $\widehat{\mathbf{A}} \in \mathbb{R}^{m,n}$, $\widehat{\mathbf{b}} \in \mathbb{R}^m$, "nearest" to $\mathbf{A}\mathbf{x} = \mathbf{b}$.

☞ least squares problem "turned upside down": now we are allowed to tamper with system matrix and right hand side vector!

Total least squares problem:

Given: $\mathbf{A} \in \mathbb{R}^{m,n}$, $m > n$, $\mathrm{rank}(\mathbf{A}) = n$, $\mathbf{b} \in \mathbb{R}^m$,

find: $\widehat{\mathbf{A}} \in \mathbb{R}^{m,n}$, $\widehat{\mathbf{b}} \in \mathbb{R}^m$ with

$$\left\| [\mathbf{A} \quad \mathbf{b}] - \left[\widehat{\mathbf{A}} \quad \widehat{\mathbf{b}}\right] \right\|_F \to \min \quad , \quad \widehat{\mathbf{b}} \in \mathcal{R}(\widehat{\mathbf{A}})\,. \tag{3.5.0.1}$$

$\widehat{\mathbf{b}} \in \mathcal{R}(\widehat{\mathbf{A}}) \;\Rightarrow\; \mathrm{rank}(\left[\widehat{\mathbf{A}} \quad \widehat{\mathbf{b}}\right]) = n$ ► (3.5.0.1) $\Rightarrow$ $\left[\widehat{\mathbf{A}} \quad \widehat{\mathbf{b}}\right] = \underset{\mathrm{rank}(\widehat{\mathbf{X}})=n}{\mathrm{argmin}} \left\| [\mathbf{A} \quad \mathbf{b}] - \widehat{\mathbf{X}} \right\|_F$.

☞ $\left[\widehat{\mathbf{A}} \quad \widehat{\mathbf{b}}\right]$ is the rank-n best approximation of $[\mathbf{A} \quad \mathbf{b}]$!

We face the problem to compute the best rank-n approximation of the given matrix $[\mathbf{A} \quad \mathbf{b}]$, a problem already treated in Section 3.4.4.2: Thm. 3.4.4.19 tells us how to use the SVD of $[\mathbf{A} \quad \mathbf{b}]$

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top\,, \quad \mathbf{U} \in \mathbb{R}^{m,n+1},\ \boldsymbol{\Sigma} \in \mathbb{R}^{n+1,n+1},\ \mathbf{V} \in \mathbb{R}^{n+1,n+1}\,, \tag{3.5.0.2}$$

to construct $\left[\widehat{\mathbf{A}} \quad \widehat{\mathbf{b}}\right]$:

$$[\mathbf{A} \quad \mathbf{b}] = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top = \sum_{j=1}^{n+1} \sigma_j (\mathbf{U})_{:,j} (\mathbf{V})_{:,j}^\top \overset{\text{Thm. 3.4.4.19}}{\Longrightarrow} \left[\widehat{\mathbf{A}} \quad \widehat{\mathbf{b}}\right] = \sum_{j=1}^{n} \sigma_j (\mathbf{U})_{:,j} (\mathbf{V})_{:,j}^\top\,. \tag{3.5.0.3}$$

$$\overset{\mathbf{V}\text{ orthogonal}}{\Longrightarrow} \left[\widehat{\mathbf{A}} \quad \widehat{\mathbf{b}}\right] (\mathbf{V})_{:,n+1} = \widehat{\mathbf{A}} (\mathbf{V})_{1:n,n+1} + \widehat{\mathbf{b}} (\mathbf{V})_{n+1,n+1} = 0\,. \tag{3.5.0.4}$$

(3.5.0.4) also provides the solution $\mathbf{x}$ of $\widehat{\mathbf{S}}\mathbf{x} = \widehat{\mathbf{b}}$,

$$\mathbf{x} := -\widehat{\mathbf{A}}^{-1}\widehat{\mathbf{b}} = -(\mathbf{V})_{1:n,n+1} / (\mathbf{V})_{n+1,n+1}\,, \tag{3.5.0.5}$$

if $(\mathbf{V})_{n+1,n+1} \neq 0$ (in numerical sense, of course).

C++-code 3.5.0.6: Total least squares via SVD ↪ GITLAB

```
// computes only solution x of fitted consistent LSE
VectorXd lsqtotal(const MatrixXd& A, const VectorXd& b) {
  const unsigned m = A.rows();
  const unsigned n = A.cols();
  MatrixXd C(m, n + 1);  C << A, b; // C = [A,b]
  // We need only the SVD-factor V, see (3.5.0.3)
  MatrixXd V = C.jacobiSvd(Eigen::ComputeThinU | Eigen::ComputeThinV).matrixV();

  // Compute solution according to (3.5.0.5);
  const double s = V(n, n);
  if (std::abs(s) < 1.0E-15) { cerr << "No solution!\n"; return {}; }
  return (-V.col(n).head(n) / s);
}
```

3.6 Constrained Least Squares

Video tutorial for Section 3.6 "Constrained Least Squares": (23 minutes) Download link, tablet notes

→ review questions 3.6.2.1

In the examples of Section 3.0.1 we expected all components of the right hand side vectors to be possibly affected by measurement errors. However, it might happen that some data are very reliable and in this case we would like the corresponding equation to be satisfied exactly.

linear least squares problem with linear constraint defined as follows:

Linear least squares problem with linear constraint:

Given: $\mathbf{A} \in \mathbb{R}^{m,n}$, $m \geq n$, $\mathrm{rank}(\mathbf{A}) = n$, $\mathbf{b} \in \mathbb{R}^m$,
$\mathbf{C} \in \mathbb{R}^{p,n}$, $p < n$, $\mathrm{rank}(\mathbf{C}) = p$, $\mathbf{d} \in \mathbb{R}^p$

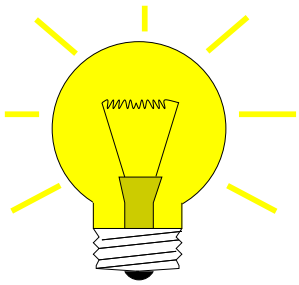
Find: $\mathbf{x} \in \mathbb{R}^n$ such that $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2 \to \min$ and $\boxed{\mathbf{C}\mathbf{x} = \mathbf{d}}$. (3.6.0.1)

Linear constraint

Here the constraint matrix $\mathbf{C}$ collects all the coefficients of those p equations that are to be satisfied exactly, and the vector $\mathbf{d}$ the corresponding components of the right hand side vector. Conversely, the m equations of the (overdetermined) LSE $\mathbf{A}\mathbf{x} = \mathbf{b}$ cannot be satisfied and are treated in a least squares sense.

3.6.1 Solution via Lagrangian Multipliers

§3.6.1.1 (A saddle point problem) Recall important technique from multidimensional calculus for tackling constrained minimization problems: Lagrange multipliers, see [Str09, Sect. 7.9].



Idea: coupling the constraint using the Lagrange multiplier $\mathbf{m} \in \mathbb{R}^p$

$$\mathbf{x} = \operatorname*{argmin}_{\mathbf{y} \in \mathbb{R}^n} \sup_{\mathbf{m} \in \mathbb{R}^p} L(\mathbf{y}, \mathbf{m}) \,, \tag{3.6.1.2}$$

$$L(\mathbf{y}, \mathbf{m}) := \frac{1}{2}\|\mathbf{A}\mathbf{y} - \mathbf{b}\|_2^2 + \mathbf{m}^\top(\mathbf{C}\mathbf{y} - \mathbf{d}) \,. \tag{3.6.1.3}$$

L as defined in (3.6.1.3) is called a **Lagrange function** or, in short, **Lagrangian**. The simple heuristics behind Lagrange multipliers is the observation:

$$\sup_{\mathbf{m} \in \mathbb{R}^p} L(\mathbf{y}, \mathbf{m}) = \infty, \quad \text{in case } \mathbf{C}\mathbf{x} \neq \mathbf{d}!$$

➥ A minimum in (3.6.1.2) can only attained, if the constraint is satisfied!

(3.6.1.2) is called a saddle point problem.

Solution of min-max problem like (3.6.1.2) is called a saddle point.

Saddle point of $F(x, m) = x^2 - 2xm$ ▷

Note that the function is "flat" in the saddle point •, that is, both the derivative with respect to $\mathbf{x}$ and with respect to $\mathbf{m}$ has to vanish.

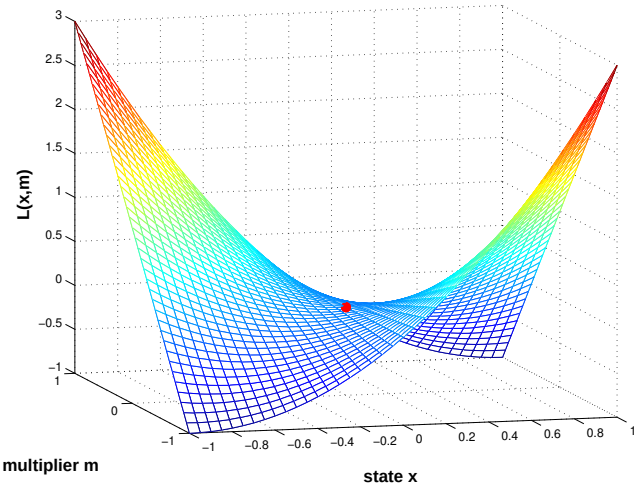


Fig. 105

⌟

§3.6.1.4 (Augmented normal equations) In a saddle point the Lagrange function is "flat", that is, all its partial derivatives have to vanish there. This yields the following necessary (and sufficient) conditions for the solution $\mathbf{x}$ of (3.6.1.2) and a saddle point in $\mathbf{x}, \mathbf{q}$: (For a similar technique employing multi-dimensional calculus see Rem. 3.1.2.5)

$$\frac{\partial L}{\partial \mathbf{x}}(\mathbf{x},\mathbf{q}) = \mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{b}) + \mathbf{C}^\top\mathbf{q} \stackrel{!}{=} 0\,, \tag{3.6.1.5a}$$

$$\frac{\partial L}{\partial \mathbf{m}}(\mathbf{x},\mathbf{q}) = \mathbf{C}\mathbf{x}-\mathbf{d} \stackrel{!}{=} 0\,. \tag{3.6.1.5b}$$

This is an $(n+p)\times(n+p)$ square linear system of equations, known as **augmented normal equations**:

$$\begin{bmatrix} \mathbf{A}^\top\mathbf{A} & \mathbf{C}^\top \\ \mathbf{C} & 0 \end{bmatrix}\begin{bmatrix} \mathbf{x} \\ \mathbf{q} \end{bmatrix} = \begin{bmatrix} \mathbf{A}^\top\mathbf{b} \\ \mathbf{d} \end{bmatrix}. \tag{3.6.1.6}$$

It belongs to the class of saddle-point type LSEs, that is, LSEs with a symmetric coefficient matrix with a zero right-lower square block. In the case $p = 0$, in the absence of a linear constraint, (3.6.1.6) collapses to the usual normal equations (3.1.2.2), $\mathbf{A}^\top\mathbf{A}\mathbf{x} = \mathbf{A}^\top\mathbf{b}$ for the overdetermined linear system of equations $\mathbf{A}\mathbf{x} = \mathbf{b}$.

As we know, a direct elimination solution algorithm for (3.6.1.6) amounts to finding an LU-decomposition of the coefficient matrix. Here we opt for its symmetric variant, the Cholesky decomposition, see Section 2.8. On the block-matrix level it can be found by considering the equation

$$\begin{bmatrix} \mathbf{A}^\top\mathbf{A} & \mathbf{C}^\top \\ \mathbf{C} & 0 \end{bmatrix} = \begin{bmatrix} \mathbf{R}^\top & 0 \\ \mathbf{G} & -\mathbf{S}^\top \end{bmatrix}\begin{bmatrix} \mathbf{R} & \mathbf{G}^\top \\ 0 & \mathbf{S} \end{bmatrix}, \quad \begin{array}{l} \mathbf{R},\mathbf{S} \in \mathbb{R}^{n,n} \text{ upper triangular matrices,} \\ \mathbf{G} \in \mathbb{R}^{p,n}\,. \end{array}$$

Thus the blocks of the Cholesky factors of the coefficient matrix of the linear system (3.6.1.6) can be determined in four steps.

① Compute $\mathbf{R}$ from $\mathbf{R}^\top\mathbf{R} = \mathbf{A}^\top\mathbf{A}$ → Cholesky decomposition → Section 2.8,
② Compute $\mathbf{G}$ from $\mathbf{R}^\top\mathbf{G}^\top = \mathbf{C}^\top$ → n forward substitutions → Section 2.3.2,
③ Compute $\mathbf{S}$ from $\mathbf{S}^\top\mathbf{S} = \mathbf{G}\mathbf{G}^\top$ → Cholesky decomposition → Section 2.8.

⌟

§3.6.1.7 (Extended augmented normal equations) The same caveats as those discussed for the regular normal equations in Rem. 3.2.0.3, Ex. 3.2.0.4, and Rem. 3.2.0.6, apply to the direct use of the augmented normal equations (3.6.1.6):

1. their condition number can be much bigger than that of the matrix $\mathbf{A}$,
2. forming $\mathbf{A}^\top\mathbf{A}$ may be vulnerable to roundoff,
3. the matrix $\mathbf{A}^\top\mathbf{A}$ may not be sparse, though $\mathbf{A}$ is.

As in § 3.2.0.7 also in the case of the augmented normal equations (3.6.1.6) switching to an extended version by introducing the residual $\mathbf{r} = \mathbf{A}\mathbf{x}-\mathbf{b}$ as a new unknown is a remedy, *cf.* (3.2.0.8). This leads to the following linear system of equations.

$$\begin{bmatrix} -\mathbf{I} & \mathbf{A} & 0 \\ \mathbf{A}^\top & 0 & \mathbf{C}^\top \\ 0 & \mathbf{C} & 0 \end{bmatrix}\begin{bmatrix} \mathbf{r} \\ \mathbf{x} \\ \mathbf{m} \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ 0 \\ \mathbf{d} \end{bmatrix} \quad \widehat{=} \quad \begin{array}{c}\text{Extended augmented} \\ \text{normal equations}\end{array}. \tag{3.6.1.8}$$

⌟

3.6.2 Solution via SVD

Idea: Identify the subspace in which the solution can vary without violating the constraint. Since $\mathbf{C}$ has *full rank*, this subspace agrees with the nullspace/kernel of $\mathbf{C}$.

From Lemma 3.4.1.13 and Ex. 3.4.2.7 we have learned that the SVD can be used to compute (an orthonormal basis of) the nullspace $\mathcal{N}(\mathbf{C})$. The suggests the following method for solving the constrained linear least squares problem (3.6.0.1).

① Compute an orthonormal basis of $\mathcal{N}(\mathbf{C})$ using SVD ($\rightarrow$ Lemma 3.4.1.13, (3.4.3.1)):

$$\mathbf{C} = \mathbf{U}[\Sigma\ 0]\begin{bmatrix}\mathbf{V}_1^\top \\ \mathbf{V}_2^\top\end{bmatrix} \quad , \quad \mathbf{U} \in \mathbb{R}^{p,p}, \Sigma \in \mathbb{R}^{p,p}, \mathbf{V}_1 \in \mathbb{R}^{n,p}, \mathbf{V}_2 \in \mathbb{R}^{n,n-p}$$

$$\blacktriangleright \quad \mathcal{N}(\mathbf{C}) = \mathcal{R}(\mathbf{V}_2)\ .$$

and the particular solution $\mathbf{x}_0 \in \mathcal{N}(\mathbf{C})^\top = \mathcal{R}(\mathbf{V}_1)$ of the constraint equation

$$\mathbf{x}_0 := \mathbf{V}_1\Sigma^{-1}\mathbf{U}^\top\mathbf{d}\ .$$

This gives us a representation of the solution $\mathbf{x}$ of (3.6.0.1) of the form

$$\mathbf{x} = \mathbf{x}_0 + \mathbf{V}_2\mathbf{y}\ , \quad \mathbf{y} \in \mathbb{R}^{n-p}\ .$$

② Insert this representation into (3.6.0.1). This yields a standard linear least squares problem with coefficient matrix $\mathbf{A}\mathbf{V}_2 \in \mathbb{R}^{m,n-p}$ and right hand side vector $\mathbf{b} - \mathbf{A}\mathbf{x}_0 \in \mathbb{R}^m$:

$$\|\mathbf{A}(\mathbf{x}_0 + \mathbf{V}_2\mathbf{y}) - \mathbf{b}\|_2 \rightarrow \min \quad \Leftrightarrow \quad \|\mathbf{A}\mathbf{V}_2\mathbf{y} - (\mathbf{b} - \mathbf{A}\mathbf{x}_0)\| \rightarrow \min\ .$$

Review question(s) 3.6.2.1 (Constrained least-squares problems)

(Q3.6.2.1.A) The angles of a flat triangle can be estimated by solving the linear system of equations

$$\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\\ 1 & 1 & 1\end{bmatrix}\begin{bmatrix}\alpha\\ \beta\\ \gamma\end{bmatrix} = \begin{bmatrix}\widetilde{\alpha}\\ \widetilde{\beta}\\ \widetilde{\gamma}\\ \pi\end{bmatrix}\ .$$

for given measured values $\widetilde{\alpha}$, $\widetilde{\beta}$, and $\widetilde{\gamma}$.

Find the least-squares solution, if *the bottom equation has to be satisfied exactly*. First recast into a linearly constrained least-squares problem

$$\|\mathbf{A}\mathbf{x} - \mathbf{b}\| \rightarrow \min \quad , \quad \mathbf{C}\mathbf{x} = \mathbf{d}\ ,$$

with $\mathbf{A} \in \mathbb{R}^{m,n}$, $\mathbf{C} \in \mathbb{R}^{p,n}$, $\mathbf{b} \in \mathbb{R}^m$, $\mathbf{d} \in \mathbb{R}^p$.

(Q3.6.2.1.B) Given $\mathbf{A} \in \mathbb{R}^{m,n}$, $\mathbf{C} \in \mathbb{R}^{p,n}$, $m, n, p \in \mathbb{N}$, $p < n$, we want to solve

$$\mathbf{x}^* = \operatorname*{argmax}_{\mathbf{x}\in\mathcal{C}} \|\mathbf{A}\mathbf{x}\|_2 \quad , \quad \mathcal{C} := \{\mathbf{x} \in \mathbb{R}^n : \ \|\mathbf{x}\|_2 = 1,\ \mathbf{C}\mathbf{x} = 0\}\ .$$

Sketch an SVD-based algorithm for computing $\mathbf{x}^*$.

△

Learning Outcomes

After having studied the contents of this chapter you should be able to

- give a rigorous definition of the least squares solution of an (overdetermined) linear system of equations,
- state the (extended) normal equations for any overdetermined linear system of equations,
- tell conditions for uniqueness and existence of solutions of the normal equations,
- define (economical) QR-decomposition and SVD of a matrix,
- know the asymptotic computational effort of computing economical QR and SVD factorizations,
- explain the use of QR-decomposition and, in particular, Givens rotations, for solving (overdetermined) linear systems of equations (in least squares sense),
- use SVD to solve least squares, (constrained) optimization, and low-rank best approximation problems
- explain the ideas underlying principal component analysis (PCA) and proper orthogonal decomposition (POD),
- formulate the augmented (extended) normal equations for a linearly constrained least squares problem.

Bibliography

[Bra06] Matthew Brand. “Fast low-rank modifications of the thin singular value decomposition”. In: *Linear Algebra Appl.* 415.1 (2006), pp. 20–30. DOI: 10.1016/j.laa.2005.07.021.

[DR08] W. Dahmen and A. Reusken. *Numerik für Ingenieure und Naturwissenschaftler*. Heidelberg: Springer, 2008 (cit. on pp. 219, 220, 227, 232–265, 276).

[GGK14] W. Gander, M.J. Gander, and F. Kwok. *Scientific Computing*. Vol. 11. Texts in Computational Science and Engineering. Heidelberg: Springer, 2014 (cit. on p. 217).

[GV13] Gene H. Golub and Charles F. Van Loan. *Matrix computations*. Fourth. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, 2013, pp. xiv+756 (cit. on pp. 243, 245, 249, 268).

[Gut07] M. Gutknecht. “Linear Algebra”. 2007 (cit. on p. 241).

[Gut09] M.H. Gutknecht. *Lineare Algebra*. Lecture Notes. SAM, ETH Zürich, 2009 (cit. on pp. 238, 266, 268, 281).

[Han02] M. Hanke-Bourgeois. *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens*. Mathematische Leitfäden. Stuttgart: B.G. Teubner, 2002 (cit. on pp. 232–235, 238, 245, 276).

[HRS16] J.S. Hesthaven, G. Rozza, and B. Stamm. *Certified Reduced Basis Methods for Parametrized Partial Differential Equations*. BCAM Springer Briefs. Cham: Springer, 2016.

[Hig02] N.J. Higham. *Accuracy and Stability of Numerical Algorithms*. 2nd ed. Philadelphia, PA: SIAM, 2002 (cit. on pp. 245, 252).

[Kal96] D. Kalman. “A singularly valuable decomposition: The SVD of a matrix”. In: *The College Mathematics Journal* 27 (1996), pp. 2–23 (cit. on p. 266).

[NS02] K. Nipp and D. Stoffer. *Lineare Algebra*. 5th ed. Zürich: vdf Hochschulverlag, 2002 (cit. on pp. 216, 241, 266, 268, 269).

[QSS00] A. Quarteroni, R. Sacco, and F. Saleri. *Numerical mathematics*. Vol. 37. Texts in Applied Mathematics. New York: Springer, 2000 (cit. on p. 233).

[QMN16] Alfio Quarteroni, Andrea Manzoni, and Federico Negri. *Reduced basis methods for partial differential equations*. Vol. 92. Unitext. Springer, Cham, 2016, pp. xi+296.

[Ste76] G. W. Stewart. “The economical storage of plane rotations”. In: *Numer. Math.* 25.2 (1976), pp. 137–138. DOI: 10.1007/BF01462266 (cit. on p. 249).

[Str19] D. Strang. *Linear Algebra and Learning from Data*. Cambridge University Press, 2019.

[Str09] M. Struwe. *Analysis für Informatiker*. Lecture notes, ETH Zürich. 2009 (cit. on pp. 224, 226, 266, 300).

[SCF18] Jan Svoboda, Thomas Cashman, and Andrew Fitzgibbon. *QRkit: Sparse, Composable QR Decompositions for Efficient and Stable Solutions to Problems in Computer Vision*. 2018.

[Vol08] S. Volkwein. *Model reduction using proper orthogonal decomposition*. Lecture notes. Graz, Austria: TU Graz, 2008.

Chapter 4

Filtering Algorithms

This chapter continues the theme of *numerical linear algebra*, also covered in Chapter 1, 2, 10. We will come across very special linear transformations ($\leftrightarrow$ matrices) and related algorithms. Surprisingly, these form the basis of a host of very important numerical methods for signal processing.

§4.0.0.1 (Time-discrete signals and sampling) From the perspective of signal processing we can identify

$$\text{vector} \quad \mathbf{x} \in \mathbb{R}^n \quad \leftrightarrow \quad \text{finite discrete (= sampled) signal.}$$

Sampling converts a time-continuous signal, represented by some real-valued physical quantity (pressure, voltage, power, etc.) into a time-discrete signal:

$X = X(t) \triangleq$ *time-continuous* signal, $0 \le t \le T$,

"sampling": $x_j = X(j\Delta t)$, $j = 0, \ldots, n-1$, $n \in \mathbb{N}$, $n\Delta t \le T$.

$\Delta t > 0 \triangleq$ time between samples.

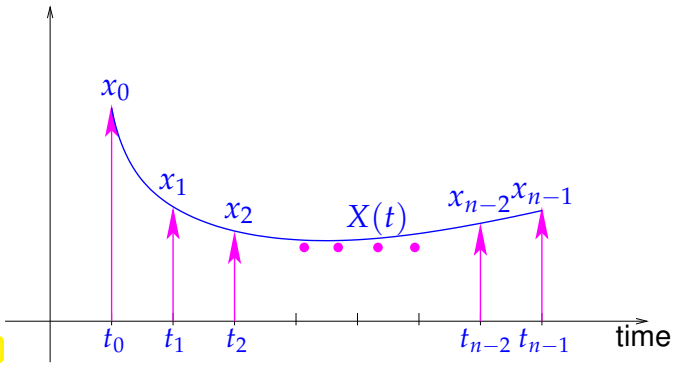


Fig. 106

As already indicated by the indexing the sampled values can be arranged in a vector $\mathbf{x} = [x_0, \ldots, x_{n-1}]^\top \in \mathbb{R}^n$.

Note that in this chapter, as is customary in signal processing, we adopt a C++-style indexing from 0: the components of a vector with length n carry indices $\in \{0, \ldots, n-1\}$.

As an idealization one sometimes considers a signal of infinite duration $X = X(t)$, $-\infty < t < \infty$. In this case sampling yields a *bi-infinite* time-discrete signal, represented by a *sequence* $(x_k)_{k\in\mathbb{Z}} \in \mathbb{R}^{\mathbb{Z}}$. If this sequence has a finite number of non-zero terms only, then we write $(0, \ldots, x_\ell, x_{\ell+1}, \ldots, x_{n-1}, x_n, 0, \ldots)$. ⌟

EXAMPLE 4.0.0.2 (Sampled audio signals) An important class of time-discrete signals are digital audio signals. Those may be obtained from sampling and analog-to-digital conversion of air pressure recorded by a microphone.

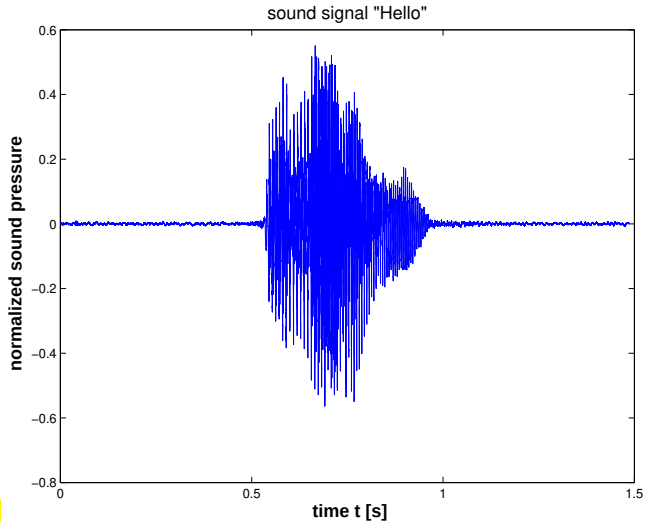


Fig. 107

A simple **file format** for storing digital audio signals is **WAV**. The audio signal is stored as a sequence of 16-bit integer values with a standard sampling rate of 44.1 kHz (44100 samples/second, sound sampled every $\Delta t = 2.2676 \cdot 10^{-5}$s).

◁ Sampled speech audio data, file `hello.wav`, ➙ GITLAB

```
Filtering > file hello.wav
hello.wav: RIFF (little-endian)
    data, WAVE audio, Microsoft PCM,
    16 bit, mono 44100 Hz
```

- In this chapter we neglect the additional "quantization", that is, the fact that, in practice, the values of a time-discrete signal $(x_k)_{k\in\mathbb{Z}}$ are again discrete, e.g., 16-bit integers for the WAV file format. Throughout, we consider only $x_k \in \mathbb{R}$.
- C++ codes handling standard file formats usually rely on *dedicated libraries*, for instance the libraries AudioFile, libsndfile, and Libwave for the WAV format.

Contents

4.1 Filters and Convolutions . . . 306
4.1.1 Discrete Finite Linear Time-Invariant Causal Channels/Filters . . . 306
4.1.2 LT-FIR Linear Mappings . . . 309
4.1.3 Discrete Convolutions . . . 312
4.1.4 Periodic Convolutions . . . 316
4.2 Discrete Fourier Transform (DFT) . . . 321
4.2.1 Diagonalizing Circulant Matrices . . . 321
4.2.2 Discrete Convolution via Discrete Fourier Transform . . . 329
4.2.3 Frequency filtering via DFT . . . 331
4.2.4 Real DFT . . . 338
4.2.5 Two-dimensional DFT . . . 340
4.2.6 Semi-discrete Fourier Transform [QSS00, Sect. 10.11] . . . 349
4.3 Fast Fourier Transform (FFT) . . . 359
4.4 Trigonometric Transformations . . . 368
4.4.1 Sine transform . . . 368
4.4.2 Cosine transform . . . 374
4.5 Toeplitz Matrix Techniques . . . 376
4.5.1 Matrices with Constant Diagonals . . . 376
4.5.2 Toeplitz Matrix Arithmetic . . . 379
4.5.3 The Levinson Algorithm . . . 380

4.1 Filters and Convolutions

4.1.1 Discrete Finite Linear Time-Invariant Causal Channels/Filters

Video tutorial for Section 4.1.1 "Discrete Finite Linear Time-Invariant Causal Channels/Filters": (11 minutes) Download link, tablet notes

In this section we study a finite linear time-invariant causal channel/filter, which is a widely used model for digital communication channels, e.g. in wireless communication theory. We adopt a mathematical perspective harnessing the toolbox of linear algebra as is common in modern engineering.

Mathematically speaking, a (discrete) channel/filter is a function/mapping $F : \ell^\infty(\mathbb{Z}) \to \ell^\infty(\mathbb{Z})$ from the vector space $\ell^\infty(\mathbb{Z})$ of bounded input sequences $\{x_j\}_{j\in\mathbb{Z}}$,

$$\ell^\infty(\mathbb{Z}) := \left\{ (x_j)_{j\in\mathbb{Z}} : \sup|x_j| < \infty \right\} ,$$

to bounded output sequences $(y_j)_{j\in\mathbb{Z}}$.

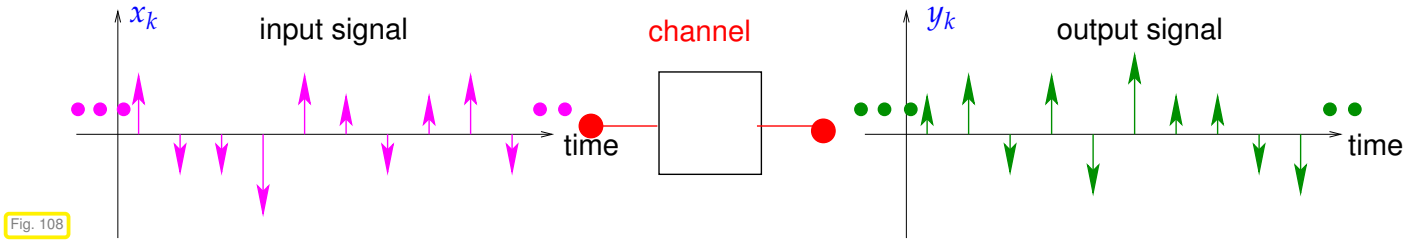


Fig. 108

Channel/filter: $F : \ell^\infty(\mathbb{Z}) \to \ell^\infty(\mathbb{Z})$, $(y_j)_{j\in\mathbb{Z}} = F((x_j)_{j\in\mathbb{Z}})$. (4.1.1.1)

In order to link (discrete) filters to linear algebra, we have to assume certain properties that are indicated by the attributes "finite ", "linear", "time-invariant" and "causal":

Definition 4.1.1.2. Finite channel/filter

A channel/filter $F : \ell^\infty(\mathbb{Z}) \to \ell^\infty(\mathbb{Z})$ is called **finite**, if every input signal of finite duration produces an output signal of finite duration,

$$\{ \exists M \in \mathbb{N}: \ |j| > M \ \Rightarrow \ x_j = 0 \} \ \Rightarrow \ \exists N \in \mathbb{N}: \ |k| > N \ \Rightarrow \ (F((x_j)_{j\in\mathbb{Z}}))_k = 0 ,$$ (4.1.1.3)

It is natural to assume that it should not matter when exactly signal is fed into the channel. To express this intuition more rigorously we introduce the time shift operator for signals: for $m \in \mathbb{Z}$

$$\mathsf{S}_m : \ell^\infty(\mathbb{Z}) \to \ell^\infty(\mathbb{Z}) \quad , \quad \mathsf{S}_m((x_j)_{j\in\mathbb{Z}}) = (x_{j-m})_{j\in\mathbb{Z}} .$$ (4.1.1.4)

Hence, by applying S_m we advance ($m < 0$) or delay ($m > 0$) a signal by $|m|\Delta t$. For a time-invariant filter time-shifts of the input propagate to the output unchanged.

Definition 4.1.1.5. Time-invariant channel/filter

A filter $F : \ell^\infty(\mathbb{Z}) \to \ell^\infty(\mathbb{Z})$ is called **time-invariant (TI)**, if shifting the input in time leads to the same output shifted in time by the same amount; it *commutes with the time shift operator* from (4.1.1.4):

$$\forall (x_j)_{j\in\mathbb{Z}} \in \ell^\infty(\mathbb{Z}), \ \forall m \in \mathbb{Z}: \ F(\mathsf{S}_m((x_j)_{j\in\mathbb{Z}})) = \mathsf{S}_m(F((x_j)_{j\in\mathbb{Z}})) .$$ (4.1.1.6)

Since a channer/filter is a mapping between vector spaces, it makes sense to talk about "linearity of F".

Definition 4.1.1.7. Linear channel/filter

A filter $F : \ell^\infty(\mathbb{Z}) \to \ell^\infty(\mathbb{Z})$ is called **linear**, if F is a linear mapping:

$$F(\alpha(x_j)_{j\in\mathbb{Z}} + \beta(y_j)_{j\in\mathbb{Z}}) = \alpha F((x_j)_{j\in\mathbb{Z}}) + \beta F((y_j)_{j\in\mathbb{Z}}) \tag{4.1.1.8}$$

for all sequences $(x_j)_{j\in\mathbb{Z}}$, $(y_j)_{j\in\mathbb{Z}} \in \ell^\infty(\mathbb{Z})$ and real numbers $\alpha, \beta \in \mathbb{R}$.

Slightly rewritten, this means that for all scaling factors $\alpha, \beta \in \mathbb{R}$

$$\text{output}(\alpha \cdot \text{signal 1} + \beta \cdot \text{signal 2}) = \alpha \cdot \text{output}(\text{signal 1}) + \beta \cdot \text{output}(\text{signal 2}) .$$

Of course, a signal should not trigger an output before it arrives at the filter; output may depend only on past and present inputs, not on the future.

Definition 4.1.1.9. Causal channel/filter

A filter $F : \ell^\infty(\mathbb{Z}) \to \ell^\infty(\mathbb{Z})$ is called **causal** (or physical, or nonanticipative), if the output does not start before the input

$$\forall M \in \mathbb{N}: \quad (x_j)_{j\in\mathbb{Z}} \in \ell^\infty(\mathbb{Z}),\ x_j = 0 \quad \forall j \le M \quad \Rightarrow \quad F((x_j)_{j\in\mathbb{Z}})_k = 0 \quad \forall k \le M . \tag{4.1.1.10}$$

Now we have collected all the properties of the class of filters in the focus of this section, called LT-FIR filters.

Acronym: **LT-FIR** $\hat{=}$ **f**inite (→ Def. 4.1.1.2), **l**inear (→ Def. 4.1.1.7), **t**ime-invariant (→ Def. 4.1.1.5), and causal (→ Def. 4.1.1.9) filter $F : \ell^\infty(\mathbb{Z}) \to \ell^\infty(\mathbb{Z})$

§4.1.1.11 (Impulse response) For the description of filters we rely on special input signals, analogous to the description of a linear mapping $\mathbb{R}^n \mapsto \mathbb{R}^m$ through a matrix, that is, its action on "coordinate vectors". The "coordinate vectors" in signal space $\ell^\infty(\mathbb{Z})$ are so-called impulses, signals that attain the value $+1$ for a single sampling point in time and are "mute" for all other times.

Definition 4.1.1.12. Impulse response

The **impulse response** (IR) of a channel/filter is the output for a single unit impulse at $t = 0$ as input, that is, the input signal is $x_j = \delta_{j,0} := \begin{cases} 1 & \text{, if } j = 0 \\ 0 & \text{else} \end{cases}$ (Kronecker symbol).

The impulse response of a finite filter can be described by a vector $\mathbf{h}$ of finite length n. In particular, the impulse response of a finite and causal filter is a sequence of the form $(\ldots, 0, h_0, h_1, \ldots, h_{n-1}, 0, \ldots)$, $n \in \mathbb{N}$. Such an impulse response is depicted in Fig. 110.

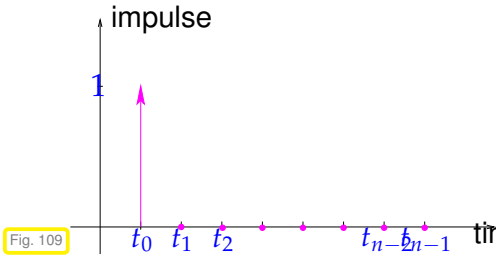


Fig. 109

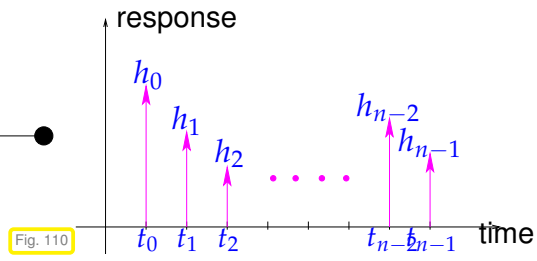


Fig. 110

Thanks to the special properties of LT-FIR filters, their impulse response offers a complete characterization. This will be explored in the next section. ⌟

Review question(s) 4.1.1.13 (Discrete finite linear time-invariant causal channels/filters)

(Q4.1.1.13.A) What is the output of a LT-FIR filter with impulse response $(\ldots,0,h_0,h_1,\ldots,h_{n-1},0,\ldots)$, $n \in \mathbb{N}$, if the input is a constant signal $(x_j)_{j\in\mathbb{Z}}$, $x_j = a$, $a \in \mathbb{R}$?

(Q4.1.1.13.B) [Filter with delayed feedback]

A filter setup is defined by feeding back the output signal with a delay Δt (effected by the shift operator S_1) and damped by a factor of 2 ▷

Judge whether this is a linear, time-invariant and causal filter, and determine its impulse response.

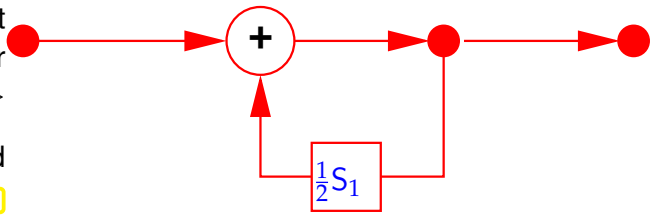


Fig. 111

Hint. First determine the output signal, when the impulse $(\delta_{0,j})_{j\in\mathbb{Z}}$ is sent through the filter.

△

4.1.2 LT-FIR Linear Mappings

Video tutorial for Section 4.1.2 "LT-FIR Linear Mappings": (12 minutes) Download link, tablet notes

We aim for a precise mathematical description of the impact of a finite, time-invariant, linear, causal filter on an input signal: Let $(\ldots,0,h_0,h_1,\ldots,h_{n-1},0,\ldots)$, $n \in \mathbb{N}$, be the **impulse response** (→ Def. 4.1.1.12) of that finite (→ Def. 4.1.1.2), linear (→ Def. 4.1.1.7), time-invariant (→ Def. 4.1.1.5), and causal (→ Def. 4.1.1.9) filter (LT-FIR) $F : \ell^\infty(\mathbb{Z}) \to \ell^\infty(\mathbb{Z})$:

$$F((\delta_{j,0})_{j\in\mathbb{Z}}) = (\ldots,0,h_0,h_1,\ldots,h_{n-1},0,\ldots)\,.$$

Owing to time-invariance we already know the response to a shifted unit pulse

$$F((\delta_{j,k})_{j\in\mathbb{Z}}) = (h_{j-k})_{j\in\mathbb{Z}} = \Big(\ \cdots\ \ 0\ \ \underset{\substack{\uparrow\\ t=k\Delta t}}{h_0}\ \ h_1\ \ \cdots\ \ \underset{\substack{\uparrow\\ t=(k+n-1)\Delta t}}{h_{n-1}}\ \ 0\ \ \cdots\ \Big)\,.$$

Every finite input signal $(\ldots,0,x_0,x_1,\ldots,x_{m-1},0,\ldots) \in \ell^\infty(\mathbb{Z})$ can be written as the superposition of scaled unit impulses, which, in turns, are time-shifted copies of a unit pulse at $t = 0$:

$$(x_j)_{j\in\mathbb{Z}} = \sum_{k=0}^{m-1} x_k (\delta_{j,k})_{j\in\mathbb{Z}} = \sum_{k=0}^{m-1} x_k S_k\Big((\delta_{j,0})_{j\in\mathbb{Z}}\Big)\,, \tag{4.1.2.1}$$

where S_k is the time-shift operator from (4.1.1.4). Applying the filter on both sides of this equation and using linearity and time-invariance we obtain

$$F\Big((x_j)_{j\in\mathbb{Z}}\Big) \overset{\text{linearity}}{=} \sum_{k=0}^{m-1} x_k F\Big(S_k\Big((\delta_{j,0})_{j\in\mathbb{Z}}\Big)\Big) \overset{\text{time-invariance}}{=} \sum_{k=0}^{m-1} x_k S_k\Big(F\Big((\delta_{j,0})_{j\in\mathbb{Z}}\Big)\Big)\,. \tag{4.1.2.2}$$

This leads to a fairly explicit formula for the output signal $\left(y_j\right)_{j\in\mathbb{Z}} := F\Big(\left(x_j\right)_{j\in\mathbb{Z}}\Big)$:

$$\begin{bmatrix} \vdots \\ 0 \\ y_0 \\ y_1 \\ \vdots \\ y_n \\ \vdots \\ \vdots \\ y_{m+n-3} \\ y_{m+n-2} \\ 0 \\ \vdots \end{bmatrix} = x_0 \begin{bmatrix} \vdots \\ 0 \\ h_0 \\ \vdots \\ h_{n-1} \\ 0 \\ 0 \\ \vdots \\ \vdots \\ 0 \\ 0 \\ \vdots \end{bmatrix} + x_1 \begin{bmatrix} \vdots \\ 0 \\ 0 \\ h_0 \\ \vdots \\ h_{n-1} \\ 0 \\ \vdots \\ \vdots \\ 0 \\ 0 \\ \vdots \end{bmatrix} + x_2 \begin{bmatrix} \vdots \\ 0 \\ 0 \\ 0 \\ h_0 \\ \vdots \\ h_{n-1} \\ 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \end{bmatrix} + \cdots + x_{m-1} \begin{bmatrix} \vdots \\ 0 \\ 0 \\ \vdots \\ \vdots \\ \vdots \\ 0 \\ h_0 \\ \vdots \\ h_{n-1} \\ 0 \\ \vdots \end{bmatrix} . \tag{4.1.2.3}$$

Thus, in compact notation we can write the non-zero components of the output signal $\left(y_j\right)_{j\in\mathbb{Z}}$ as

channel is causal and finite!

$$y_k = F\Big(\left(x_j\right)_{j\in\mathbb{Z}}\Big)_k = \sum_{j=0}^{m-1} h_{k-j}x_j \quad , \quad k = 0,\ldots,m+n-2 \quad (h_j := 0 \text{ for } j<0 \text{ and } j \geq n) \, . \tag{4.1.2.4}$$

Summary of the above considerations:

Superposition of impulse responses

The output $(\ldots,0,y_0,y_1,y_2,\ldots)$ of a finite, time-invariant, linear, and causal channel for finite length input $\mathbf{x} = (\ldots,0,x_0,\ldots,x_{m-1},0,\ldots) \in \ell^\infty(\mathbb{Z})$ is a *superposition* of x_j-weighted $j\Delta t$ time-shifted impulse responses.

EXAMPLE 4.1.2.6 (Visualization: superposition of impulse responses) The following diagrams give a visual display of the above considerations, namely of the superposition of impulse responses for a particular finite, time-invariant, linear, and causal filter (LT-FIR), and an input signal of duration $3\Delta t$, $\Delta t \mathrel{\hat{=}}$ time between samples. We see the case $m = 4$, $n = 5$.

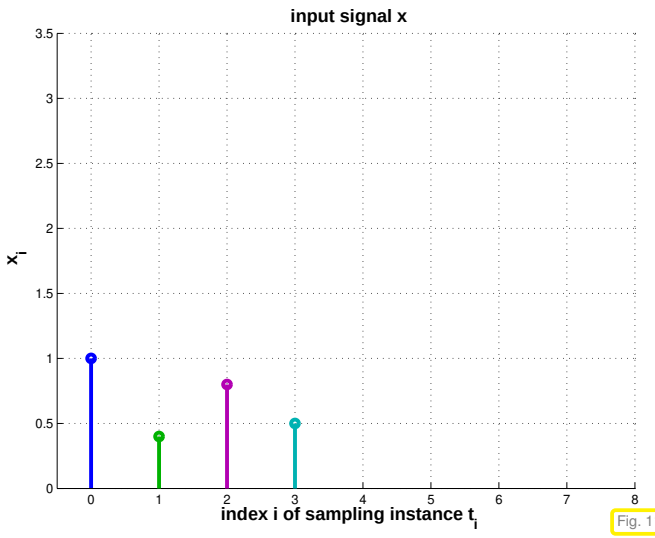


Fig. 112

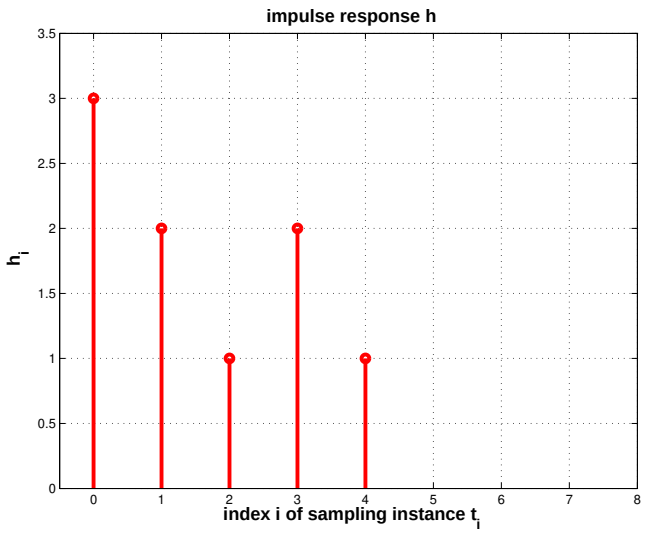


Fig. 113

In this special case the formula (4.1.2.3) becomes

$$\begin{bmatrix} \vdots \\ 0 \\ y_0 \\ y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ 0 \\ \vdots \end{bmatrix} = x_0 \begin{bmatrix} \vdots \\ 0 \\ h_0 \\ h_1 \\ h_2 \\ h_3 \\ h_4 \\ 0 \\ 0 \\ 0 \\ \vdots \end{bmatrix} + x_1 \begin{bmatrix} \vdots \\ 0 \\ 0 \\ h_0 \\ h_1 \\ h_2 \\ h_3 \\ h_4 \\ 0 \\ 0 \\ \vdots \end{bmatrix} + x_2 \begin{bmatrix} \vdots \\ 0 \\ 0 \\ 0 \\ h_0 \\ h_1 \\ h_2 \\ h_3 \\ h_4 \\ 0 \\ \vdots \end{bmatrix} + x_3 \begin{bmatrix} \vdots \\ 0 \\ 0 \\ 0 \\ 0 \\ h_0 \\ h_1 \\ h_2 \\ h_3 \\ h_4 \\ \vdots \end{bmatrix} .$$

This reflects the fact that the is a *linear* superposition of impulse responses:

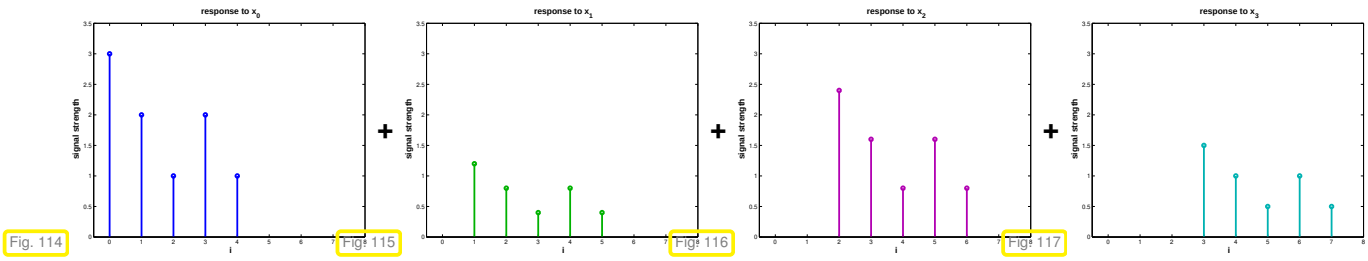


Fig. 114 Fig. 115 Fig. 116 Fig. 117

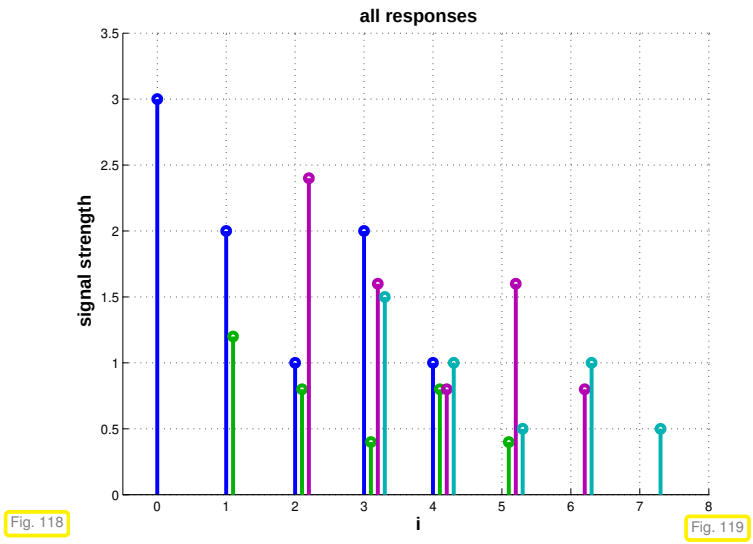


Fig. 118

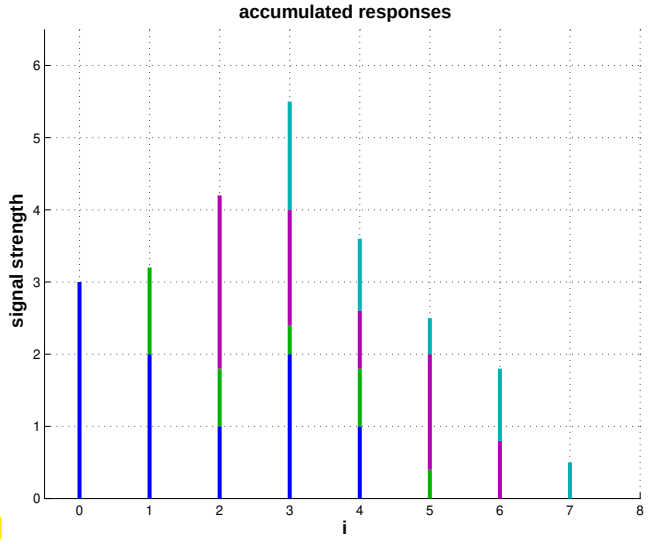


Fig. 119

The formula (4.1.2.4) characterizing the output sequence $(y_k)_{k\in\mathbb{Z}}$ is a special case of a fundamental bilinear operation on pairs of sequences, not necessarily finite.

Definition 4.1.2.7. Convolution of sequences

Given two sequences $(h_k)_{k\in\mathbb{Z}}, (x_k)_{k\in\mathbb{Z}}$, at least one of which is finite or decays sufficiently fast, their **convolution** is another sequence $(y_k)_{k\in\mathbb{Z}}$, defined as

$$y_k = \sum_{j\in\mathbb{Z}} h_{k-j} x_j \,, \quad k \in \mathbb{Z} \,. \tag{4.1.2.8}$$

✎ Notation: For the sequence arising from convolving two sequences $(h_k)_{k\in\mathbb{Z}}$ and $(x_k)_{k\in\mathbb{Z}}$ we write $(x_k) * (h_k)$.

Note that convolution is not well-defined on $\ell^\infty(\mathbb{Z}) \times \ell^\infty(\mathbb{Z})$. A counterexample is provided by constant, non-zero sequences. However, convolution has another interesting property, which can easily be established by re-indexing $j \leftarrow k - j$ in (4.1.2.8):

Theorem 4.1.2.9. Convolution of sequences commutes

If well-defined, the convolution of sequences commutes

$$(x_k) * (h_k) = (h_k) * (x_k) \,.$$

Review question(s) 4.1.2.10 (LT-FIR Linear Mappings)

(Q4.1.2.10.A) [Composition of LT-FIR filters] Given two LT-FIR channels with impulse responses $(\ldots, 0, h_0, h_1, \ldots, h_{n-1}, 0, \ldots)$, $n \in \mathbb{N}$ and $(\ldots, 0, g_0, g_1, \ldots, g_l, 0, \ldots, 0)$, $l \in \mathbb{N}$ we build another channel by composing them. This means we send the output of the first into the second. What is the impulse response of the composition?

(Q4.1.2.10.B) An simple LT-FIR channel has impulse response $(\ldots, 0, 0, h_0 := 1, h_1 := 1, 0, 0, \ldots)$. What is the impulse response of the channel that is constructed by composing N of the simple channels. You should see a familiar pattern!

(Q4.1.2.10.C) [] An LT-FIR channel has known impulse response $(\ldots, 0, h_0, h_1, \ldots, h_{n-1}, 0, \ldots)$, $n \in \mathbb{N}$. We know that it received a finite input signal (x_k) of duration $(m-1)\Delta t$ and we measure the output signal (y_k) in exactly the same timespan.

Outline how one can compute the input signal. When is this possible?

△

4.1.3 Discrete Convolutions

Video tutorial for Section 4.1.3 "Discrete Convolutions": (9 minutes) Download link, tablet notes

☞ You may also watch the explanations of 3Blue1Brown on convolutions here.

Computers can deal only with finite amounts of data. So algorithms can operate only on finite signals, which will be in the focus of this section. This continues the considerations undertaken in the beginning of Section 4.1.2, now with an emphasis on recasting operations in the language of linear algebra.

Remark 4.1.3.1 (The case of finite signals and filters) Again we consider a finite (→ Def. 4.1.1.2), linear (→ Def. 4.1.1.7), time-invariant (→ Def. 4.1.1.5), and causal (→ Def. 4.1.1.9) filter (LT-FIR) with impulse response $(\ldots, 0, h_0, h_1, \ldots, h_{n-1}, 0, \ldots)$, $n \in \mathbb{N}$. From (4.1.2.4) we learn that

$$\boxed{\text{duration(output signal)} \leq \text{duration(input signal)} + \text{duration(impulse response)}} \,.$$

We have seen this in (4.1.2.4), where an input signal with m pulses (duration $(m-1)\Delta t$) and an impulse response with n pulses (duration $(n-1)\Delta t$) spawned an output signal with $m+n-1$ pulses (duration $(m-1+n-1)\Delta t$).

Therefore, if we know that all input signals have a duration of at most $(m-1)\Delta t$, which means they are of the form $(\ldots, x_0, x_1, \ldots, x_{m-1}, 0, \ldots)$, we can model them as vectors $\mathbf{x} = [x_0, \ldots, x_{m-1}]^\top \in \mathbb{R}^m$, *cf.* § 4.0.0.1, and the filter can be viewed as a linear mapping $F : \mathbb{R}^n \to \mathbb{R}^{m+n-1}$, which takes us to the realm of linear algebra.

Thus, for the linear filter we have a matrix representation of (4.1.2.4). Let us first look at the special case $m = 4, n = 5$ presented in Ex. 4.1.2.6:

$$\begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \end{bmatrix} = x_0 \begin{bmatrix} h_0 \\ h_1 \\ h_2 \\ h_3 \\ h_4 \\ 0 \\ 0 \\ 0 \end{bmatrix} + x_1 \begin{bmatrix} 0 \\ h_0 \\ h_1 \\ h_2 \\ h_3 \\ h_4 \\ 0 \\ 0 \end{bmatrix} + x_2 \begin{bmatrix} 0 \\ 0 \\ h_0 \\ h_1 \\ h_2 \\ h_3 \\ h_4 \\ 0 \end{bmatrix} + x_3 \begin{bmatrix} 0 \\ 0 \\ 0 \\ h_0 \\ h_1 \\ h_2 \\ h_3 \\ h_4 \end{bmatrix} .$$

Here, we have already replaced the sequences with finite-length vectors. Translating this relationship into matrix-vector notation is easy:

$$\begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \end{bmatrix} = \begin{bmatrix} h_0 & 0 & 0 & 0 \\ h_1 & h_0 & 0 & 0 \\ h_2 & h_1 & h_0 & 0 \\ h_3 & h_2 & h_1 & h_0 \\ h_4 & h_3 & h_2 & h_1 \\ 0 & h_4 & h_3 & h_2 \\ 0 & 0 & h_4 & h_3 \\ 0 & 0 & 0 & h_4 \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \end{bmatrix} .$$

Writing $\mathbf{y} = [y_0, \ldots, y_{m+n-2}]^\top \in \mathbb{R}^{m+n-1}$ for the vector of the output signal we find for the general case the following matrix×vector representation of the action of the filter on the signal:

$$\mathbf{y} := \begin{bmatrix} y_0 \\ \vdots \\ \\ \\ \\ \\ \\ y_{m+n-2} \end{bmatrix} = \begin{bmatrix} h_0 & 0 & \cdots & 0 \\ h_1 & h_0 & & \vdots \\ & h_1 & \ddots & 0 \\ h_{n-1} & & \ddots & h_0 \\ 0 & h_{n-1} & & h_1 \\ & \ddots & \ddots & \\ 0 & \cdots & 0 & h_{n-1} \end{bmatrix} \begin{bmatrix} x_0 \\ \vdots \\ \\ \\ \\ \\ \\ x_{m-1} \end{bmatrix} =: \mathbf{Cx} . \tag{4.1.3.2}$$

Note that the $i+1$-th column of the matrix $\mathbf{C} \in \mathbb{R}^{m+n-1,m}$ is obtained by *cyclically permuting* column i, $i = 1, \ldots, m-1$. ⌟

Recall the formula

$$y_k = \sum_{j=0}^{m-1} h_{k-j} x_j , \quad k = 0, \ldots, m+n-2 \quad (h_j := 0 \text{ for } j < 0 \text{ and } j \geq n) . \tag{4.1.2.4}$$

supplying the non-zero terms of the convolution of two finite sequences $(h_k)_k$ and $(x_k)_k$ of length n and m, respectively. Both can be identified with vectors $[x_0, \ldots, x_{m-1}]^\top \in \mathbb{K}^m$, $[h_0, \ldots, h_{n-1}]^\top \in \mathbb{K}^n$, and, since (4.1.2.4) is a special case of the convolution of sequences introduced in Def. 4.1.2.7, we might call it a convolution of vectors. It represents a fundamental operation in signal theory.

Definition 4.1.3.3. Discrete convolution

Given $\mathbf{x} = [x_0,\ldots,x_{m-1}]^\top \in \mathbb{K}^m$, $\mathbf{h} = [h_0,\ldots,h_{n-1}]^\top \in \mathbb{K}^n$ their **discrete convolution** (DCONV) is the vector $\mathbf{y} = [y_0,\ldots,y_{m+n-2}]^\top \in \mathbb{K}^{m+n-1}$ with components

$$y_k = \sum_{j=0}^{m-1} h_{k-j}x_j\,, \quad k = 0,\ldots,m+n-2\,, \tag{4.1.3.4}$$

where we have adopted the convention $h_j := 0$ for $j < 0$ or $j \geq n$.

✎ Notation for discrete convolution (4.1.3.4): $\mathbf{y} = \mathbf{h} * \mathbf{x}$.

Remark 4.1.3.5 (Commutativity of discrete convolution) Discrete convolution (4.1.3.4) of two vectors is a commutative operation mirroring the result from Thm. 4.1.2.9 without the implicit assumptions required there.

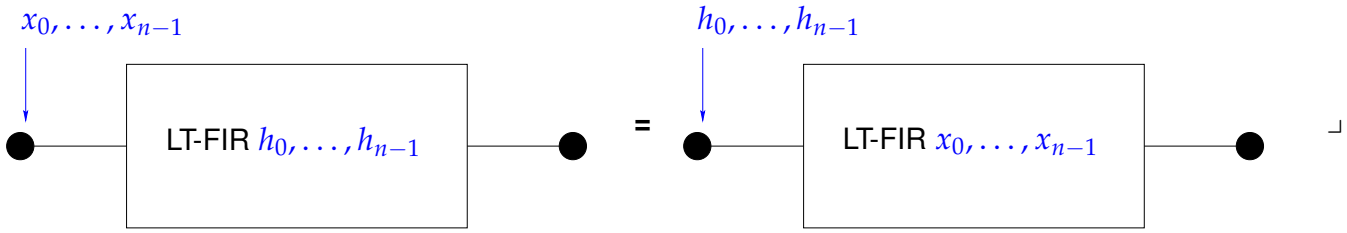
Using the notations of Def. 4.1.3.3 we embed the two vectors into bi-infinite sequences $\left(\widetilde{x}_j\right)_{j\in\mathbb{Z}}$, $\left(\widetilde{h}_j\right)_{j\in\mathbb{Z}}$ by zero-padding:

$$\widetilde{x}_j := \begin{cases} x_j & \text{for} \quad j \in \{0,\ldots,m-1\}\,, \\ 0 & \text{else} \end{cases}\,, \quad \widetilde{h}_j := \begin{cases} h_j & \text{for} \quad j \in \{0,\ldots,n-1\}\,, \\ 0 & \text{else} \end{cases}\,, \quad j \in \mathbb{Z}\,.$$

Thm. 4.1.2.9 confirms $\left(\widetilde{x}_j\right) * \left(\widetilde{h}_j\right) = \left(\widetilde{h}_j\right) * \left(\widetilde{x}_j\right)$, that is, the two sequences have exactly the same terms. By definition of $\widetilde{x}_j$ and $\widetilde{h}_j$ we see

$$\begin{aligned}(\mathbf{h}*\mathbf{x})_k &= \sum_{j=0}^{m-1} \widetilde{h}_{k-j}x_j = \sum_{j\in\mathbb{Z}} \widetilde{h}_{k-j}\widetilde{x}_j \\ &= \left(\left(\widetilde{x}_j\right) * \left(\widetilde{h}_j\right)\right)_k = \left(\left(\widetilde{h}_j\right) * \left(\widetilde{x}_j\right)\right)_k = \sum_{j\in\mathbb{Z}}^{n-1} \widetilde{x}_{k-j}\widetilde{h}_j = (\mathbf{x}*\mathbf{h})_k\,:\end{aligned}$$

The discrete convolution of vectors is *commutative*: Phrased in signal theory terminology, we have found that filter and signal can be "swapped":



§4.1.3.6 (Multiplication of polynomials) The formula (4.1.3.4) for the discrete convolution also occurs in a context completely detached from signal processing. "Surprisingly" the bilinear operation (4.1.2.4) (for $m = n$) that takes two input n-vectors and produces an output $2n-1$-vector also provides the coefficients of the product polynomial.

Concretely, consider two polynomials in t of degree $n-1$, $n \in \mathbb{N}$, with real or complex coefficients,

$$p(t) := \sum_{j=0}^{n-1} a_j t^j\,, \quad q(t) := \sum_{j=0}^{n-1} b_j t^j\,, \qquad a_j, b_j \in \mathbb{K}\,.$$

Their product pq will be a polynomial of degree $2n-2$:

$$(pq)(t) = \sum_{k=0}^{2n-2} c_k t^k \,, \quad c_k := \sum_{\ell=\max\{0,k-(n-1)\}}^{\min\{k,n-1\}} a_\ell b_{k-\ell} \,, \quad k = 0,\ldots,2n-2 \,. \tag{4.1.3.7}$$

Let us introduce dummy coefficients for $p(t)$ and $q(t)$, $a_j, b_j,\ j = 2n,\ldots,2n-2$, all set to 0. This can be easily done in a computer code by resizing the coefficient vectors of p and q and filling the new entries with zeros ("zero padding"). The above formula for c_j can then be rewritten as

$$c_j = \sum_{\ell=0}^{j} a_\ell b_{j-\ell} \,, \quad j = 0,\ldots,2n-2 \,. \tag{4.1.3.8}$$

Hence, the coefficients of the product polynomial can be obtained as the discrete convolution of the coefficient vectors of p and q:

$$\left[c_0\ c_1\ \ldots\ c_{2n-2}\right]^\top = \mathbf{a} * \mathbf{b} \quad ! \tag{4.1.3.9}$$

Moreover, this provides another proof for the commutativity of discrete convolution. ⌟

Remark 4.1.3.10 (Convolution of causal sequences) The notion of a discrete convolution of Def. 4.1.3.3 naturally extends to so-called causal sequences $\in \ell^\infty(\mathbb{N}_0)$, that is, bounded mappings $\mathbb{N}_0 \mapsto \mathbb{K}$: the (discrete) convolution of two sequences $(x_j)_{j\in\mathbb{N}_0}$, $(y_j)_{j\in\mathbb{N}_0}$ is the sequence $(z_j)_{j\in\mathbb{N}_0}$ defined by

$$z_k := \sum_{j=0}^{k} x_{k-j} y_j = \sum_{j=0}^{k} x_j y_{k-j} \,, \quad k \in \mathbb{N}_0 \,. \tag{4.1.2.8}$$

In this context recall the product formula for power series, **Cauchy product**, which can be viewed as a multiplication rule for "infinite polynomials" = power series. ⌟

Review question(s) 4.1.3.11 (Discrete convolutions)

(Q4.1.3.11.A) [Calculus of discrete convolutions] Let $* : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^{n+m-1}$ denote the discrete convolution of two vectors in Def. 4.1.3.3 defined as (C++ indexing, $\mathbf{x} = [x_0,\ldots,x_{n-1}]^\top$, $Vh = [h_0,\ldots,h_{m-1}]^\top$)

$$(\mathbf{h} * \mathbf{x})_k := \sum_{j=0}^{m-1} h_{k-j} x_j \,, \quad k = 0,\ldots,m+n-2 \,, \tag{4.1.3.4}$$

where we set non-existent components of $\mathbf{h}$ to zero in the sum.

Which of the following statements are true?

1. $\mathbf{h} * \mathbf{x} = \mathbf{x} * \mathbf{h}$ for all $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{h} \in \mathbb{R}^m$, $m,n \in \mathbb{N}$,
2. $\mathbf{x} * \mathbf{x} = \mathbf{0} \quad \Rightarrow \quad \mathbf{x} = \mathbf{0}$ for all $\mathbf{x} \in \mathbb{R}^n$
3. $(\mathbf{x}+\mathbf{y}) * \mathbf{h} = \mathbf{x} * \mathbf{h} + \mathbf{y} * \mathbf{h}$ for all $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$, $\mathbf{h} \in \mathbb{R}^m$.

(Q4.1.3.11.B) [Repeated convolution] Write $*$ for the convolution of two vectors according to Def. 4.1.3.3. Give a formula for the entries of the vector

$$\mathbf{b}^n := \underbrace{\begin{bmatrix}1\\1\end{bmatrix} * \cdots * \begin{bmatrix}1\\1\end{bmatrix}}_{n \text{ times}} \,.$$

(Q4.1.3.11.C) [Sum of discrete random variables] Consider two discrete *independent* random variables $X, Y : \Omega \to \mathbb{Z}$, Ω a probability space, and write for the probabilities

$$x_k := \mathbb{P}(X = k) \in [0,1] \quad , \quad y_k := \mathbb{P}(Y = k) \in [0,1] \, , \quad k \in \mathbb{Z} \, .$$

Give a formula for $\mathbb{P}(X + Y = k)$, $k \in \mathbb{Z}$, and relate it to the convolution of sequences.

Definition 4.1.2.7. Convolution of sequences

Given two sequences $(h_k)_{k\in\mathbb{Z}}, (x_k)_{k\in\mathbb{Z}}$, at least one of which is finite or decays sufficiently fast, their **convolution** is another sequence $(y_k)_{k\in\mathbb{Z}}$, defined as

$$y_k = \sum_{j\in\mathbb{Z}} h_{k-j} x_j \, , \quad k \in \mathbb{Z} \, . \tag{4.1.2.8}$$

△

4.1.4 Periodic Convolutions

Video tutorial for Section 4.1.4 "Periodic Convolutions": (12 minutes) Download link, tablet notes

Understanding how periodic signals interact with finite, linear, time-invariant, causal (FT-FIR) filters is an important stepping stone for developing algorithms for more general situations.

Definition 4.1.4.1. Periodic time-discrete signal

An n-**periodic** signal, $n \in \mathbb{N}$, is a sequence $(x_j)_{j\in\mathbb{Z}} \in \ell^\infty(\mathbb{Z})$ satisfying

$$x_{j+n} = x_j \quad \forall j \in \mathbb{Z} \, .$$

➢ Though infinite, an n-periodic signal $(x_j)_{j\in\mathbb{Z}}$ is uniquely determined by the finitely many values $x_0, \ldots, x_{n-1}$ and can be associated with a vector $\mathbf{x} = [x_0, \ldots, x_{n-1}]^\top \in \mathbb{R}^n$.

§4.1.4.2 (Linear filtering of periodic signals) Whenever the input signal of a linear and time-invariant (LTI) filter $F : \ell^\infty(\mathbb{Z}) \to \ell^\infty(\mathbb{Z})$ with impulse response $(h_j)_{j\in\mathbb{Z}}$ is n-periodic, so will be the output signal. To elaborate this we start from the convolution formula for sequences from Def. 4.1.2.7 and take into account the n-periodicity to compute the output $(y_k)_{k\in\mathbb{Z}} := F((x_k)_{k\in\mathbb{Z}})$:

$$\begin{aligned} y_k = \sum_{j\in\mathbb{Z}} h_{k-j} x_j &\overset{\text{Thm. 4.1.2.9}}{=} \sum_{j\in\mathbb{Z}} x_{k-j} h_j \overset{j\leftarrow\nu+\ell n}{=} \sum_{\nu=0}^{n-1} \sum_{\ell\in\mathbb{Z}} x_{k-\nu-\ell n} \, h_{\nu+\ell n} \\ &\overset{\text{periodicity}}{=} \sum_{\nu=0}^{n-1} \Big(\sum_{\ell\in\mathbb{Z}} h_{\nu+\ell n} \Big) x_{k-\nu} \, , \quad k \in \mathbb{Z} \, . \end{aligned} \tag{4.1.4.3}$$

From the n-periodicity of $(x_j)_{j\in\mathbb{Z}}$ we immediately conclude that $y_k = y_{k+n}$ for all $k \in \mathbb{Z}$. Thus, in the n-periodic setting, a *linear* and time-invariant (LTI) filter will give rise to a *linear* mapping $\mathbb{R}^n \mapsto \mathbb{R}^n$ according to

$$y_k = \sum_{j=0}^{n-1} p_j x_{k-j} = \sum_{j=0}^{n-1} p_{k-j} x_j \tag{4.1.4.4}$$

$$\text{for some } p_0,\dots,p_{n-1} \in \mathbb{R} \text{ satisfying } p_k := p_{k-n} \text{ for all } k \in \mathbb{Z}\,.$$

From (4.1.4.3) we see that the defining terms of the *n-periodic* sequence $(p_k)_{k\in\mathbb{Z}}$ can be computed according to

$$p_j = \sum_{\ell\in\mathbb{Z}} h_{j+\ell n}\,, \quad j \in \{0,\dots,n-1\}\,. \tag{4.1.4.5}$$

This sequence can be regarded as periodic impulse response, the output generated by the input sequence $\sum_{k\in\mathbb{Z}}(\delta_{nk,j})_{j\in\mathbb{Z}}$. It must not be mixed up with the impulse response (→ Def. 4.1.1.12) of the filter.

In matrix notation (4.1.4.4) reads

$$\begin{bmatrix} y_0 \\ \vdots \\ \\ \\ \\ \vdots \\ y_{n-1} \end{bmatrix} = \underbrace{\begin{bmatrix} p_0 & p_{n-1} & p_{n-2} & \cdots & & \cdots & p_1 \\ p_1 & p_0 & p_{n-1} & & & & \vdots \\ p_2 & p_1 & p_0 & \ddots & & & \\ \vdots & & \ddots & \ddots & \ddots & & \\ & & & \ddots & \ddots & \ddots & \\ \vdots & & & & \ddots & \ddots & p_{n-1} \\ p_{n-1} & \cdots & & & & p_1 & p_0 \end{bmatrix}}_{=:\mathbf{P}} \begin{bmatrix} x_0 \\ \vdots \\ \\ \\ \\ \vdots \\ x_{n-1} \end{bmatrix}. \tag{4.1.4.6}$$

where $(\mathbf{P})_{ij} = p_{i-j}$, $1 \le i,j \le n$, with $p_j := p_{j+n}$ for $1-n \le j < 0$.

⌟

The following special variant of a discrete convolution operation is motivated by the preceding § 4.1.4.2.

Definition 4.1.4.7. Discrete periodic convolution

The **discrete periodic convolution** of two n-periodic sequences $(p_k)_{k\in\mathbb{Z}}$, $(x_k)_{k\in\mathbb{Z}}$ yields the n-periodic sequence

$$(y_k) := (p_k) *_n (x_k) \quad , \quad y_k := \sum_{j=0}^{n-1} p_{k-j}x_j = \sum_{j=0}^{n-1} x_{k-j}p_j\,, \quad k \in \mathbb{Z}\,.$$

✎ notation for discrete periodic convolution: $(p_k) *_n (x_k)$

The identity claimed in (4.1.4.4) and in Def. 4.1.4.7 can be established by a simple index transformation $\ell := k - j$ and subsequent shifting of the sum, which does not change the value thanks to periodicity.

$$\sum_{j=0}^{n-1} p_{k-j}x_j = \sum_{\ell=k-n-1}^{k} \underbrace{p_\ell x_{k-\ell}}_{n\text{-periodic in } \ell} = \sum_{\ell=0}^{n-1} p_\ell x_{k-\ell}\,.$$

This means that the discrete periodic convolution of two sequences *commutes*.

Since n-periodic sequences can be identified with vectors in $\mathbb{K}^n$ (see above), we can also introduce the discrete periodic convolution of vectors:

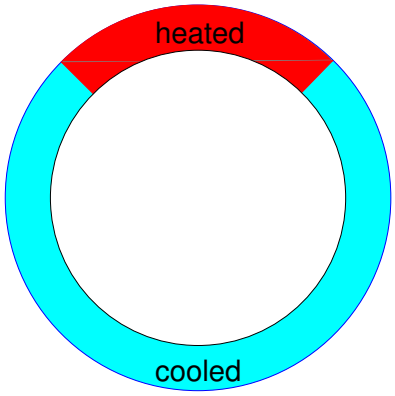
Def. 4.1.4.7 ➢ discrete periodic convolution of vectors: $\mathbf{y} = \mathbf{p} *_n \mathbf{x} \in \mathbb{K}^n$, $\mathbf{p},\mathbf{x} \in \mathbb{K}^n$.

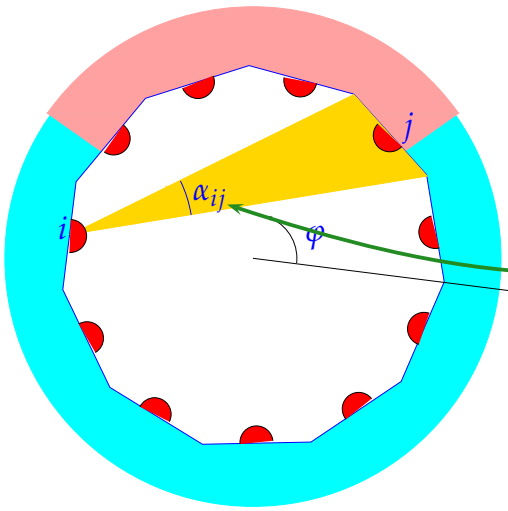
EXAMPLE 4.1.4.8 (Radiative heat transfer) Beyond signal processing discrete periodic convolutions occur in many mathematical models:

An engineering problem:

- cylindrical pipe,
- heated on part Γ_H of its perimeter ($\to$ prescribed heat flux),
- cooled on remaining perimeter Γ_K ($\to$ constant heat flux).

Task: compute local heat fluxes.





Modeling (discretization):

- approximation by regular n-polygon, edges Γ_j,
- isotropic radiation of each edge Γ_j (power I_j),

radiative heat flow $\Gamma_j \to \Gamma_i$: $P_{ji} := \frac{\alpha_{ij}}{\pi} I_j$,

opening angle: $\alpha_{ij} = \pi\,\gamma_{|i-j|}$, $1 \le i,j \le n$,

power balance: $\underbrace{\sum_{i=1,i\neq j}^{n} P_{ji}}_{=I_j} - \sum_{i=1,i\neq j}^{n} P_{ij} = Q_j\,.$ (4.1.4.9)

$Q_j \mathrel{\hat{=}}$ heat flux through Γ_j, satisfies

$$Q_j := \int_{\frac{2\pi}{n}(j-1)}^{\frac{2\pi}{n}j} q(\varphi)\,\mathrm{d}\varphi\,, \quad q(\varphi) := \begin{cases} \text{local heating} & \text{, if } \varphi \in \Gamma_H\,, \\ -\frac{1}{|\Gamma_K|}\int_{\Gamma_H} q(\varphi)\,\mathrm{d}\varphi & \text{(const.), if } \varphi \in \Gamma_K\,. \end{cases}$$

$$\text{(4.1.4.9)} \quad \Rightarrow \quad \text{LSE:} \quad I_j - \sum_{i=1,i\neq j}^{n} \frac{\alpha_{ij}}{\pi} I_i = Q_j\,, \quad j = 1,\ldots,n\,.$$

e.g. $n = 8$:

$$\begin{bmatrix} 1 & -\gamma_1 & -\gamma_2 & -\gamma_3 & -\gamma_4 & -\gamma_3 & -\gamma_2 & -\gamma_1 \\ -\gamma_1 & 1 & -\gamma_1 & -\gamma_2 & -\gamma_3 & -\gamma_4 & -\gamma_3 & -\gamma_2 \\ -\gamma_2 & -\gamma_1 & 1 & -\gamma_1 & -\gamma_2 & -\gamma_3 & -\gamma_4 & -\gamma_3 \\ -\gamma_3 & -\gamma_2 & -\gamma_1 & 1 & -\gamma_1 & -\gamma_2 & -\gamma_2 & -\gamma_4 \\ -\gamma_4 & -\gamma_3 & -\gamma_2 & -\gamma_1 & 1 & -\gamma_1 & -\gamma_2 & -\gamma_3 \\ -\gamma_3 & -\gamma_4 & -\gamma_3 & -\gamma_2 & -\gamma_1 & 1 & -\gamma_1 & -\gamma_2 \\ -\gamma_2 & -\gamma_3 & -\gamma_4 & -\gamma_3 & -\gamma_2 & -\gamma_1 & 1 & -\gamma_1 \\ -\gamma_1 & -\gamma_2 & -\gamma_3 & -\gamma_4 & -\gamma_3 & -\gamma_2 & -\gamma_1 & 1 \end{bmatrix} \begin{bmatrix} I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \\ I_6 \\ I_7 \\ I_8 \end{bmatrix} = \begin{bmatrix} Q_1 \\ Q_2 \\ Q_3 \\ Q_4 \\ Q_5 \\ Q_6 \\ Q_7 \\ Q_8 \end{bmatrix}. \qquad \text{(4.1.4.10)}$$

This is a linear system of equations with symmetric, singular, and (by Lemma 9.1.0.5, $\sum \gamma_i \le 1$) positive semidefinite ($\to$ Def. 1.1.2.6) system matrix.

Note that the matrices from (4.1.4.6) and (4.1.4.10) have the same structure!

Also observe that the LSE from (4.1.4.10) can be written by means of the discrete periodic convolution ($\to$ Def. 4.1.4.7) of vectors $\mathbf{y} = (1, -\gamma_1, -\gamma_2, -\gamma_3, -\gamma_4, -\gamma_3, -\gamma_2, -\gamma_1)$, $\mathbf{x} = (I_1,\ldots,I_8)$

$$\text{(4.1.4.10)} \quad \leftrightarrow \quad \mathbf{y} *_8 \mathbf{x} = [Q_1,\ldots,Q_8]^\top\,.$$

§4.1.4.11 (Circulant matrices) In Ex. 4.1.4.8 we have already seen a matrix of a special form, the matrix $\mathbf{P}$ in

$$\begin{bmatrix} y_0 \\ \vdots \\ \\ \\ \\ \vdots \\ y_{n-1} \end{bmatrix} = \underbrace{\begin{bmatrix} p_0 & p_{n-1} & p_{n-2} & \cdots & & \cdots & p_1 \\ p_1 & p_0 & p_{n-1} & & & & \vdots \\ p_2 & p_1 & p_0 & \ddots & & & \\ \vdots & & \ddots & \ddots & \ddots & & \\ & & & \ddots & \ddots & \ddots & \\ \vdots & & & & \ddots & \ddots & p_{n-1} \\ p_{n-1} & \cdots & & & & p_1 & p_0 \end{bmatrix}}_{=:\mathbf{P}} \begin{bmatrix} x_0 \\ \vdots \\ \\ \\ \\ \vdots \\ x_{n-1} \end{bmatrix}. \tag{4.1.4.6}$$

Matrices with this particular structure are so common that they have been given a special name.

Definition 4.1.4.12. Circulant matrix → [Han02, Sect. 54]

A matrix $\mathbf{C} = [c_{ij}]_{i,j=1}^{n} \in \mathbb{K}^{n,n}$ is **circulant**

$:\Leftrightarrow$ $\exists\, (p_k)_{k\in\mathbb{Z}}$ n-periodic sequence: $c_{ij} = p_{i-j}$, $1 \le i,j \le n$.

✎ Notation: We write $\operatorname{circul}(\mathbf{p}) \in \mathbb{K}^{n,n}$ for the circulant matrix generated by the periodic sequence/vector $\mathbf{p} = [p_0,\ldots,p_{n-1}]^\top \in \mathbb{K}^n$

The structure of a generic circulant matrix ("constant diagonals") can be visualized as

$$\operatorname{circul}(\mathbf{p}) = \begin{bmatrix} p_0 & p_{n-1} & p_{n-2} & \cdots & & \cdots & p_1 \\ p_1 & p_0 & & & & & p_2 \\ p_2 & & & & & & \vdots \\ \vdots & & & & & & \\ & & & & & & \vdots \\ p_{n-2} & & & & & & p_{n-1} \\ p_{n-1} & p_{n-2} & \cdots & & \cdots & p_1 & p_0 \end{bmatrix}$$

☞ A circulant matrix has constant (main, sub- and super-) diagonals (for which indices $j - i = \text{const.}$).

☞ columns/rows arise by *cyclic permutation* of the first column/row.

Similar to the case of banded matrices (→ Section 2.7.5) we note that the

"information content" of circulant matrix $\mathbf{C} \in \mathbb{K}^{n,n}$ is just n numbers $\in \mathbb{K}$.
(obviously, one vector $\mathbf{u} \in \mathbb{K}^n$ enough to define circulant matrix $\mathbf{C} \in \mathbb{K}^{n,n}$)

Supplement 4.1.4.13. Write $\mathbf{Z}((u_k)) \in \mathbb{K}^{n,n}$ for the circulant matrix generated by the n-periodic sequence $(u_k)_{k\in\mathbb{Z}}$. Denote by $\mathbf{y} := [y_0,\ldots,y_{n-1}]^\top$, $\mathbf{x} = [x_0,\ldots,x_{n-1}]^\top$ the vectors associated to n-periodic sequences. Then the commutativity of the discrete periodic convolution (→ Def. 4.1.4.7) involves

$$\operatorname{circul}(\mathbf{x})\mathbf{y} = \operatorname{circul}(\mathbf{y})\mathbf{x}\,. \tag{4.1.4.14}$$

⌟

Remark 4.1.4.15 (Reduction of discrete convolution to periodic convolution) Recall the discrete convolution (→ Def. 4.1.3.3) of two vectors $\mathbf{a} = [a_0,\ldots,a_{n-1}]^\top \in \mathbb{K}^n$, $\mathbf{b} = [b_0,\ldots,b_{n-1}]^\top \in \mathbb{K}^n$.

$$z_k := (\mathbf{a} * \mathbf{b})_k = \sum_{j=0}^{n-1} a_j b_{k-j}\,, \quad k = 0,\ldots,2n-2 \qquad (b_k := 0 \text{ for } k<0,\, k \geq n)\,.$$

Now expand $a_0,\ldots,a_{n-1}$ and $b_0,\ldots,b_{n-1}$ to $2n-1$-periodic sequences by zero padding.

$$x_k := \begin{cases} a_k & \text{, if } 0 \leq k < n\,, \\ 0 & \text{, if } n \leq k < 2n-1 \end{cases}\,, \quad y_k := \begin{cases} b_k & \text{, if } 0 \leq k < n\,, \\ 0 & \text{, if } n \leq k < 2n-1\,, \end{cases} \tag{4.1.4.16}$$

and periodic extension: $x_k = x_{2n-1+k}$, $y_k = y_{2n-1+k}$ for all $k \in \mathbb{Z}$.

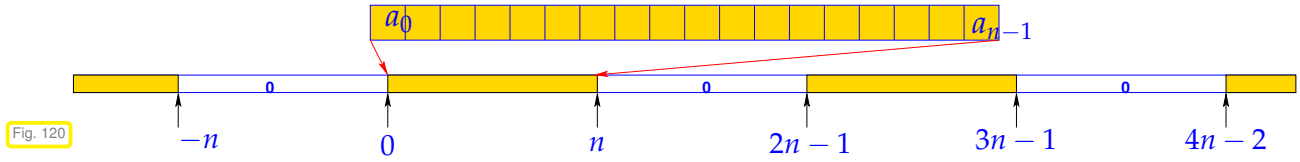


Fig. 120

The zero components prevent interaction of different periods:

$$z_k = \sum_{j=0}^{k} a_j b_{k-j} = \sum_{j=0}^{k} x_j y_{k-j} + \sum_{j=k+1}^{n-1} x_j \underbrace{y_{2n-1+k-j}}_{=0} + \sum_{j=n}^{2n-2} \underbrace{x_j}_{=0} y_{2n-1+k-j}\,, \quad k = 0,\ldots,n-1\,,$$

$$z_k = \sum_{j=k-n+1}^{n-1} a_j b_{k-j} = \sum_{j=0}^{k-n} x_j \underbrace{y_{k-j}}_{=0} + \sum_{j=k-n+1}^{n-1} x_j y_{k-j} + \sum_{j=n}^{2n-2} \underbrace{x_j}_{=0} y_{k-j}\,, \quad k = n,\ldots,2n-1\,.$$

This makes periodic and non-periodic discrete convolutions coincide. Writing $\mathbf{x},\mathbf{y} \in \mathbb{K}^n$ for the defining vectors of $(x_k)_{k\in\mathbb{Z}}$ and $(y_k)_{k\in\mathbb{Z}}$ we find

$$(\mathbf{a} * \mathbf{b})_k = (\mathbf{x} *_{2n-1} \mathbf{y})_k\,, \quad k = 0,\ldots,2n-2\,. \tag{4.1.4.17}$$

In the spirit of (4.1.3.2) we can switch to a matrix view of the reduction to periodic convolution:

$$\begin{bmatrix} z_0 \\ \vdots \\ z_{2n-2} \end{bmatrix} = \underbrace{\begin{bmatrix} b_0 & 0 & \cdots & 0 & b_{n-1} & \cdots & b_1 \\ b_1 & b_0 & 0 & \cdots & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ b_{n-1} & \cdots & b_1 & b_0 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & b_1 & b_0 & \ddots & \vdots \\ \vdots & \ddots & 0 & \ddots & \ddots & \ddots & b_{n-1} \\ 0 & \cdots & 0 & b_{n-1} & \cdots & b_1 & b_0 \end{bmatrix}}_{\text{a } (2n-1)\times(2n-1) \text{ circulant matrix!}} \begin{bmatrix} a_0 \\ \vdots \\ a_{n-1} \\ 0 \\ \vdots \\ 0 \end{bmatrix}\,. \tag{4.1.4.18}$$

Discrete convolution can be realized by multiplication with a circulant matrix (→ 4.1.4.12)

⌟

Review question(s) 4.1.4.19 (Periodic Convolutions)

(Q4.1.4.19.A) Let (y_k) be the (finite) output signal obtained from an LT-FIR channel F with impulse response $(\ldots,0,h_0,h_1,\ldots,h_{n-1},0,\ldots)$ for a finite input signal (x_k) with duration $(m-1)\Delta t$. For what $p \in \mathbb{N}$ do we get

$$\sum_{\ell\in\mathbb{Z}} \mathsf{S}_{\ell p}((y_k)) = F\left(\sum_{\ell\in\mathbb{Z}} \mathsf{S}_{\ell p}((x_k))\right),$$

where S_ν is the time-shift operator: $\mathsf{S}_\nu((z_k)) := (z_{k-\nu})_{k\in\mathbb{Z}}$?

(Q4.1.4.19.B) [Zero-padding] As in Question (**Q**4.1.4.19.A) let $(y_k)_{k\in\mathbb{Z}}$ be the (finite) output signal obtained from an LT-FIR channel F with impulse response $(\ldots,0,h_0,h_1,\ldots,h_{n-1},0,\ldots)$, $n \in \mathbb{N}$, for a finite input signal $(x_k)_{k\in\mathbb{Z}}$ with duration $(m-1)\Delta t$, $m \in \mathbb{N}$. For what $p \in \mathbb{N}$ can we count on

$$y_k = \left(F\Big(\sum_{\ell\in\mathbb{Z}} \mathsf{S}_{\ell p}((x_k))\Big)\right)_k, \quad k = 0,\ldots,m-1\ ?$$

(Q4.1.4.19.C) [Circulant matrices] Complete the following equivalent definition of a circulant matrix:

Definition [Ros21, SDef. 1]. Circulant matrix

A matrix $\mathbf{C} = [c_{ij}]_{i,j=1}^n \in \mathbb{K}^{n,n}$ is **circulant**, if for every entry c_{ij} with $i,j \in \{1,\ldots,n\}$ the following property holds

$$c_{i,j} = c_{k,\ell}\,, \quad k := \boxed{}\,, \quad \ell := \boxed{}\,.$$

(Linear algebra indexing applied)

Hint. Recall

Definition 4.1.4.12. Circulant matrix → [Han02, Sect. 54]

A matrix $\mathbf{C} = [c_{ij}]_{i,j=1}^n \in \mathbb{K}^{n,n}$ is **circulant**
$:\Leftrightarrow$ $\exists\,(p_k)_{k\in\mathbb{Z}}$ n-periodic sequence: $c_{ij} = p_{i-j}$, $1 \le i,j \le n$.

and use the "modulo operator" %.

△

4.2 Discrete Fourier Transform (DFT)

4.2.1 Diagonalizing Circulant Matrices

Video tutorial for Section 4.2.1 "Diagonalizing Circulant Matrices": (17 minutes)
Download link, tablet notes

Algorithms dealing with circulant matrices make use of their very special spectral properties. Full understanding requires familiarity with the theory of eigenvalues and eigenvectors of matrices from linear algebra, see [NS02, Ch. 7], [Gut09, Ch. 9].

EXPERIMENT 4.2.1.1 (Eigenvectors of circulant matrices) Now we are about to discover a very deep truth ...

Experimentally, we examine the eigenvalues and eigenvectors of two random 8×8 circulant matrices $\mathbf{C}_1$, $\mathbf{C}_2$ ($\rightarrow$ Def. 4.1.4.12), generated from random vectors with entries even distributed in $[0,1]$, `VectorXd::Random(n)`.

eigenvalues (real part) $\triangleright$

Little relationship between (complex!) eigenvalues can be observed, as can be expected from random matrices with entries $\in [0,1]$.

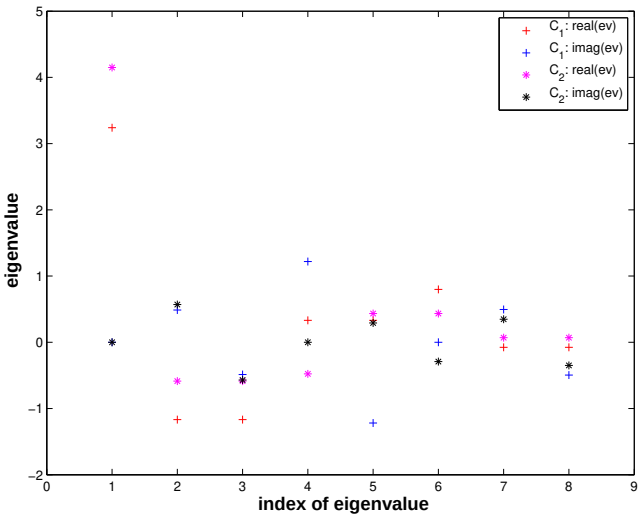
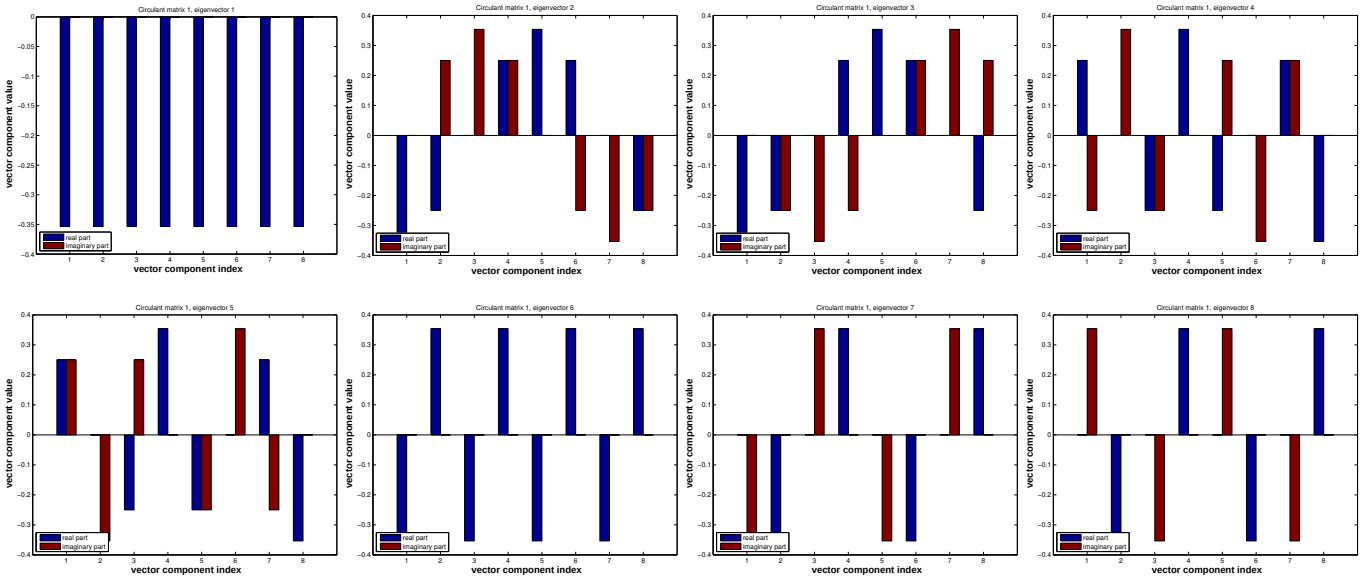


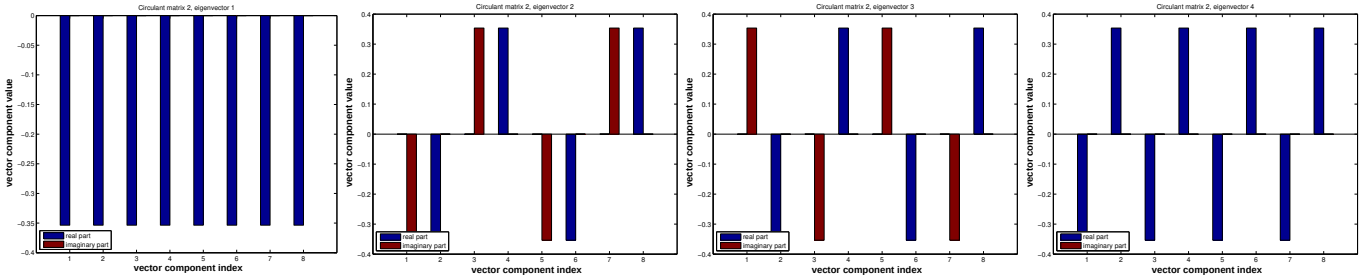
Fig. 121

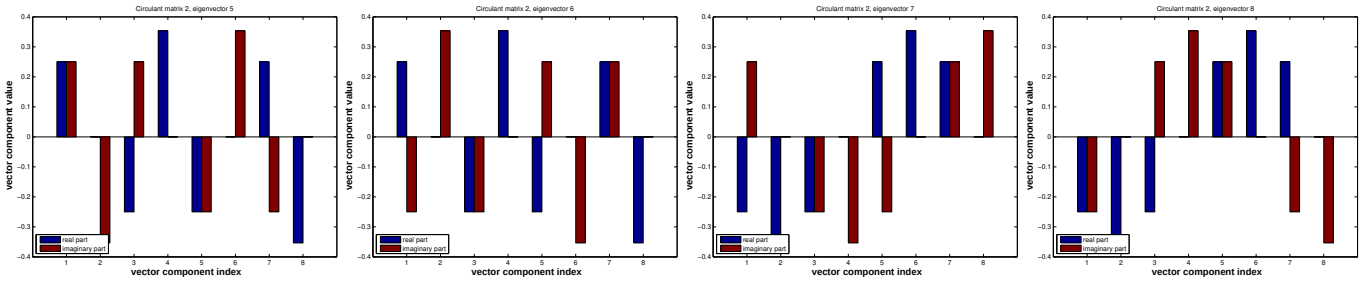
Now: the **surprise** ...

Eigenvectors of matrix $\mathbf{C}_1$, visualized through the size of the real and imaginary parts of their components.



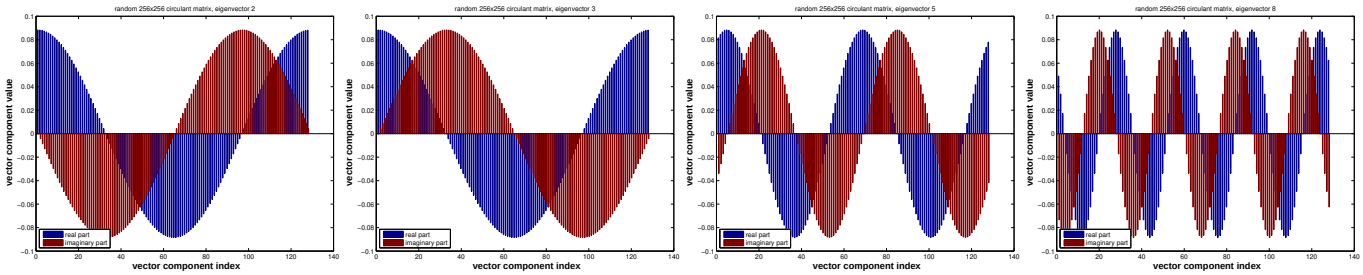
Eigenvectors of matrix $\mathbf{C}_2$





Observation: different random circulant matrices have the same eigenvectors!

Eigenvectors of circulant matrix $\mathbf{C} = \operatorname{circul}([1,2,\ldots,128]^\top)$:



The eigenvectors remind us of sampled *trigonometric functions* $\cos(k/n)$, $\sin(k/n)$, $k = 0,\ldots,n-1$! ⌟

Remark 4.2.1.2 (Eigenvectors of commuting matrices) An abstract result from linear algebra puts the surprising observation made in Exp. 4.2.1.1 in a wider context.

Theorem 4.2.1.3. Commuting matrices have the same eigenvectors

If $\mathbf{A},\mathbf{B} \in \mathbb{K}^{n,n}$ commute, that is, $\mathbf{AB} = \mathbf{BA}$, and $\mathbf{A}$ has n distinct eigenvalues, then the eigenspaces of $\mathbf{A}$ and $\mathbf{B}$ coincide.

Proof. Let $\mathbf{v} \in \mathbb{K}^n \setminus \{\mathbf{0}\}$ be an eigenvector of $\mathbf{A}$ with eigenvalue λ. Then

$$(\mathbf{A}-\lambda\mathbf{I})\mathbf{v} = 0 \quad \Rightarrow \quad \mathbf{B}(\mathbf{A}-\lambda\mathbf{I})\mathbf{v} = 0 \quad \overset{\mathbf{BA}=\mathbf{AB}}{\Rightarrow} \quad (\mathbf{A}-\lambda\mathbf{I})\mathbf{Bv} = 0\,.$$

Since in the case of n distinct eigenvalues $\dim \mathcal{N}(\mathbf{A}-\lambda\mathbf{I}) = 1$, we conclude that there is $\xi \in \mathbb{K}$: $\mathbf{Bv} = \xi\mathbf{v}$, $\mathbf{v}$ is an eigenvector of $\mathbf{B}$. Since the eigenvectors of $\mathbf{A}$ span $\mathbb{K}^n$, there cannot be eigenvectors of $\mathbf{B}$ that are not eigenvectors of $\mathbf{A}$.

Moreover, there is a basis of $\mathbb{K}^n$ consisting of eigenvectors of $\mathbf{B}$; $\mathbf{B}$ can be diagonalized. □

Next, by straightforward calculation one verifies that every circulant matrix commutes with the unitary and circulant cyclic permutation matrix

$$\mathbf{S} = \begin{bmatrix} 0 & 0 & 0 & \cdots & \cdots & 0 & 1 \\ 1 & 0 & 0 & & & \vdots & 0 \\ 0 & 1 & 0 & \ddots & & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & & \\ & & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & & \ddots & 0 & 0 \\ 0 & \cdots & & \cdots & 0 & 1 & 0 \end{bmatrix}. \tag{4.2.1.4}$$

As a unitary matrix $\mathbf{S}$ can be diagonalized. Observe that $\mathbf{S}^n - \mathbf{I} = \mathbf{O}$, that is the minimal polynomial of $\mathbf{S}$ is $\xi \mapsto \xi^n - 1$, which is irreducible, because it has n distinct roots (of unity). Therefore, by Thm. 4.2.1.3, $\mathbf{S}$ has n different eigenvalues and every eigenvector of $\mathbf{S}$ is also eigenvector of any circulant matrix.

By elementary means we can compute the eigenvectors of $\mathbf{S}$: Assume that $\mathbf{v} = [v_0, \ldots, v_{n-1}] \in \mathbb{C}^n \setminus \{\mathbf{0}\}$ satisfies $\mathbf{S}\mathbf{v} = \lambda\mathbf{v}$ for some $\lambda \in \mathbb{C}$. This implies for its components the following relationships:

$$\begin{aligned} & v_{j-1} = \lambda v_j \,, \quad j = 1, \ldots, n-1 \quad , \quad v_{n-1} = \lambda v_0 \,, \\ \Rightarrow \quad & v_j = \lambda^{n-j-i} v_{n-1} = \lambda^{n-j} v_0 \quad \Rightarrow \quad \lambda^n = 1 \,. \end{aligned}$$

Hence λ is an n-th root of unity and has the representation $\lambda = \exp\big(2\pi\imath\frac{k}{n}\big)$, $k \in \{0, \ldots, n-1\}$. Setting $v_0 := 1$ we find

$$v_j = \exp\big(2\pi\imath\frac{k}{n}\big)^j = \exp\big(2\pi\imath\frac{kj}{n}\big) \,, \quad j = 0, \ldots, n-1 \,.$$

⌟

Remark 4.2.1.5 (Why using $\mathbb{K} = \mathbb{C}$?) In Exp. 4.2.1.1 we saw that we get *complex* eigenvalues/eigenvectors for general circulant matrices. More generally, in many cases real matrices can be diagonalized only in $\mathbb{C}$, which is the ultimate reason for the importance of complex numbers.

Complex numbers also allow an elegant handling of trigonometric functions: recall from analysis the unified treatment of trigonometric functions via the *complex exponential function*

$$\exp(it) = \cos(t) + \imath\sin(t) \,, \quad t \in \mathbb{R} \,.$$

$\mathbb{C}$! The field of complex numbers $\mathbb{C}$ is the *natural framework* for the analysis of linear, time-invariant filters, and the development of algorithms for circulant matrices.

⌟

§4.2.1.6 (Eigenvectors of circulant matrices) Now we verify by direct computations that circulant matrices all have a particular set of eigenvectors. This will entail computing in $\mathbb{C}$, *cf.* Rem. 4.2.1.15.

✎ notation: nth **root of unity** $\omega_n := \exp(-2\pi\imath/n) = \cos(2\pi/n) - \imath\sin(2\pi/n)$, $n \in \mathbb{N}$

satisfies $\overline{\omega}_n = \omega_n^{-1}$, $\boxed{\omega_n^n = 1}$, $\omega_n^{n/2} = -1$, $\omega_n^k = \omega_n^{k+n}$ $\forall\, k \in \mathbb{Z}$, (4.2.1.7)

$$\sum_{k=0}^{n-1} \omega_n^{kj} = \sum_{k=0}^{n-1} \big(\omega_n^j\big)^k = \begin{cases} n & \text{, if } j = 0 \mod n \,, \\ 0 & \text{, if } j \neq 0 \mod n \,. \end{cases} \tag{4.2.1.8}$$

(4.2.1.8) is a simple consequence of the geometric sum formula

$$\begin{aligned} & \sum_{k=0}^{n-1} q^k = \frac{1-q^n}{1-q} \quad \forall\, q \in \mathbb{C} \setminus \{1\} \,, \quad n \in \mathbb{N} \,. \qquad (4.2.1.9) \\ \Rightarrow \quad & \sum_{k=0}^{n-1} \omega_n^{kj} = \frac{1-\omega_n^{nj}}{1-\omega_n^j} = \frac{1-\exp(-2\pi\imath j)}{1-\exp(-2\pi\imath j/n)} = 0 \,, \end{aligned}$$

because $\exp(-2\pi\imath j) = \omega_n^{nj} = (\omega_n^n)^j = 1$ for all $j \in \mathbb{Z}$.

In expressions like ω_n^{kl} the term "kl" will always designate an exponent and will never play the role of a superscript.

Now we want to confirm the conjecture gleaned from Exp. 4.2.1.1 that vectors with powers of roots of unity are eigenvectors for any circulant matrix. We do this by simple and straightforward computations:

We consider a general circulant matrix $\mathbf{C} \in \mathbb{C}^{n,n}$ ($\rightarrow$ Def. 4.1.4.12), with $c_{ij} := (\mathbf{C})_{i,j} = u_{i-j}$, for an n-periodic sequence $(u_k)_{k\in\mathbb{Z}}$, $u_k \in \mathbb{C}$. We "guess" an eigenvector,

$$\mathbf{v}_k \in \mathbb{C}^n: \quad \mathbf{v}_k := \left[\omega_n^{-jk}\right]_{j=0}^{n-1}, \quad k \in \{0,\ldots,n-1\},$$

and verify the eigenvector property by direct computation:

$$(\mathbf{C}\mathbf{v}_k)_j = \sum_{l=0}^{n-1} u_{j-l}\omega_n^{-lk} = \sum_{l=0}^{n-1} u_l \omega_n^{-(j-l)k} = \omega_n^{-jk}\sum_{l=0}^{n-1} u_l\omega_n^{lk} = \lambda_k \cdot \omega_n^{-jk} = \lambda_k \cdot (\mathbf{v}_k)_j . \tag{4.2.1.10}$$

change of summation index — independent of j !

$\mathbf{v}_k$ is eigenvector of $\mathbf{C}$ for eigenvalue $\lambda_k = \sum_{l=0}^{n-1} u_l\omega_n^{lk}$.

The set $\{\mathbf{v}_0,\ldots,\mathbf{v}_{n-1}\} \subset \mathbb{C}^n$ provides the so-called *orthogonal* trigonometric basis of $\mathbb{C}^n$ = *eigenvector basis* for circulant matrices

$$\{\mathbf{v}_0,\ldots,\mathbf{v}_{n-1}\} = \left\{ \begin{bmatrix} \omega_n^0 \\ \vdots \\ \omega_n^0 \end{bmatrix}, \begin{bmatrix} \omega_n^0 \\ \omega_n^{-1} \\ \vdots \\ \omega_n^{1-n} \end{bmatrix}, \cdots, \begin{bmatrix} \omega_n^0 \\ \omega_n^{2-n} \\ \omega_n^{2(2-n)} \\ \vdots \\ \omega_n^{-(n-1)(n-2)} \end{bmatrix}, \begin{bmatrix} \omega_n^0 \\ \omega_n^{1-n} \\ \omega_n^{2(1-n)} \\ \vdots \\ \omega_n^{-(n-1)^2} \end{bmatrix} \right\}. \tag{4.2.1.11}$$

From (4.2.1.8) we can conclude orthogonality of the basis vectors by straightforward computations:

$$\mathbf{v}_k := \left[\omega_n^{-jk}\right]_{j=0}^{n-1} \in \mathbb{C}^n: \quad \mathbf{v}_k^{\mathrm{H}}\mathbf{v}_m = \sum_{j=0}^{n-1} \omega_n^{jk}\omega_n^{-jm} = \sum_{j=0}^{n-1} \omega_n^{(k-m)j} \overset{(4.2.1.8)}{=} 0 \text{ , if } k \neq m . \tag{4.2.1.12}$$

The matrix effecting the change of basis trigonometrical basis $\rightarrow$ standard basis is called the **Fourier-matrix**

$$\mathbf{F}_n = \begin{bmatrix} \omega_n^0 & \omega_n^0 & \cdots & \omega_n^0 \\ \omega_n^0 & \omega_n^1 & \cdots & \omega_n^{n-1} \\ \omega_n^0 & \omega_n^2 & \cdots & \omega_n^{2n-2} \\ \vdots & \vdots & & \vdots \\ \omega_n^0 & \omega_n^{n-1} & \cdots & \omega_n^{(n-1)^2} \end{bmatrix} = \left[\omega_n^{\ell j}\right]_{\ell,j=0}^{n-1} = \left[\exp(-2\pi\imath\ell j/n)\right]_{\ell,j=0}^{n-1} \in \mathbb{C}^{n,n} . \tag{4.2.1.13}$$

Lemma 4.2.1.14. Properties of Fourier matrices

The scaled Fourier-matrix $\frac{1}{\sqrt{n}}\mathbf{F}_n$ is unitary ($\rightarrow$ Def. 6.3.1.2) : $\mathbf{F}_n^{-1} = \frac{1}{n}\mathbf{F}_n^{\mathrm{H}} = \frac{1}{n}\overline{\mathbf{F}}_n$.

Proof. The lemma is immediate from (4.2.1.12) and (4.2.1.8),because

$$\left(\mathbf{F}_n\mathbf{F}_n^{\mathrm{H}}\right)_{l,j} = \sum_{k=0}^{n-1} \omega_n^{(l-1)k}\overline{\omega_n^{(j-1)k}} = \sum_{k=0}^{n-1} \omega_n^{(l-1)k}\omega_n^{-(j-1)k} = \sum_{k=0}^{n-1} \omega_n^{k(l-j)} , \quad 1 \le l,j \le n .$$

□ ⌟

Remark 4.2.1.15 (Spectrum of Fourier matrix) We draw a conclusion from the properties stated in Lemma 4.2.1.14:

$$\tfrac{1}{n^2}\mathbf{F}_n^4 = I \quad \Rightarrow \quad \sigma(\tfrac{1}{\sqrt{n}}\mathbf{F}_n) \subset \{1,-1,i,-i\}\,,$$

because, if $\lambda \in \mathbb{C}$ is an eigenvalue of $\mathbf{F}_n$, then there is an eigenvector $\mathbf{x} \in \mathbb{C}^n \setminus \{0\}$ such that $\mathbf{F}_n\mathbf{x} = \lambda\mathbf{x}$, see Def. 9.1.0.1. ⌟

The next lemma summarizes what we have found in § 4.2.1.6.

Lemma 4.2.1.16. Diagonalization of circulant matrices (→ Def. 4.1.4.12)

For any circulant matrix $\mathbf{C} \in \mathbb{K}^{n,n}$, $c_{ij} = u_{i-j}$, $(u_k)_{k\in\mathbb{Z}}$ an n-periodic sequence, holds true

$$\mathbf{C}\overline{\mathbf{F}}_n = \overline{\mathbf{F}}_n \operatorname{diag}(d_0,\ldots,d_{n-1}) \quad , \quad [d_0,\ldots,d_{n-1}]^\top = \mathbf{F}_n[u_0,\ldots,u_{n-1}]^\top\,,$$

where $\mathbf{F}_n \in \mathbb{C}^{n,n}$ is the Fourier matrix: $\mathbf{F}_n = \left[\exp(-\frac{2\pi\imath}{n}\ell k)\right]_{\ell,k=0}^{n-1}$ from (4.2.1.13).

Proof. The computations from (4.2.1.10) established:

$$\mathbf{C}\mathbf{v}_k = \mathbf{C}\big(\overline{\mathbf{F}}_n\big)_{:,k} = (\mathbf{F})_{:,k}\sum_{\ell=0}^{n-1} u_\ell \omega_n^{\ell k} = \big(\overline{\mathbf{F}}_n\big)_{:,k}\sum_{\ell=0}^{n-1} u_\ell (\mathbf{F}_n)_{k,\ell} = \big(\overline{\mathbf{F}}_n\big)_{:,k}(\mathbf{F}_n\mathbf{u})_k\,.$$

Then invoke the rules of matrix×matrix multiplication. □

From this lemma and the fact $\overline{\mathbf{F}}_n = n\mathbf{F}_n^{-1}$ we conclude

$$\boxed{\mathbf{C} = \mathbf{F}_n^{-1}\operatorname{diag}(d_0,\ldots,d_{n-1})\mathbf{F}_n} \quad , \quad [d_0,\ldots,d_{n-1}]^\top = \mathbf{F}_n[u_0,\ldots,u_{n-1}]^\top\,. \tag{4.2.1.17}$$

As a consequence of Lemma 4.2.1.16, (4.2.1.17) the multiplication with Fourier-matrices will be crucial operation in algorithms for circulant matrices and discrete convolutions. Therefore this operation has been given a special name:

Definition 4.2.1.18. Discrete Fourier transform (DFT)

The linear map $\mathrm{DFT}_n : \mathbb{C}^n \mapsto \mathbb{C}^n$, $\mathrm{DFT}_n(\mathbf{y}) := \mathbf{F}_n\mathbf{y}$, $\mathbf{y} \in \mathbb{C}^n$, is called **discrete Fourier transform** (DFT), i.e. for $[c_0,\ldots,c_{n-1}] := \mathrm{DFT}_n(\mathbf{y})$

$$c_k = \sum_{j=0}^{n-1} y_j\,\omega_n^{kj} = \sum_{j=0}^{n-1} y_j\,\exp\big(-2\pi\imath\frac{kj}{n}\big) \quad , \quad k = 0,\ldots,n-1\,. \tag{4.2.1.19}$$

Recall the convention also adopted for the discussion of the DFT: vector indexes range from 0 to $n-1$!

Terminology: The result of DFT, $\mathbf{c} = \mathrm{DFT}_n(\mathbf{y}) = \mathbf{F}_n\mathbf{y}$, is also called the (discrete) Fourier transform of $\mathbf{y}$.

From $\mathbf{F}_n^{-1} = \frac{1}{n}\overline{\mathbf{F}}_n$ (→ Lemma 4.2.1.14) we find the *inverse discrete Fourier transform*:

$$c_k = (\mathrm{DFT}_n(\mathbf{y}))_k = \sum_{j=0}^{n-1} y_j \omega_n^{kj} \quad \Leftrightarrow \quad y_j = \left(\mathrm{DFT}_n^{-1}(\mathbf{c})\right)_j = \frac{1}{n}\sum_{k=0}^{n-1} c_k \omega_n^{-kj} \,. \tag{4.2.1.20}$$

§4.2.1.21 (Discrete Fourier transform in EIGEN and PYTHON)

- EIGEN-functions for discrete Fourier transform (and its inverse), supplied both as unary and binary versions, see usage guide.

  DFT: `c=fft.fwd(y)` or `fft.fwd(c,y)` ↔ inverse DFT: `y=fft.inv(c);` or `fft.inv(y, c);`

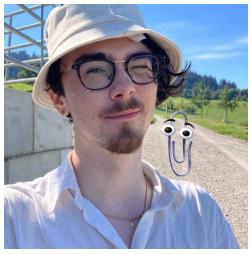
  Before caling these functions remember to

  1. `# include <unsupported/Eigen/FFT>`,
  2. instantiate an object of the helper class **Eigen::FFT**.

C++ code 4.2.1.22: Demo: discrete Fourier transform in EIGEN ⇝ GITLAB

```
2  int main(int /*argc*/, char ** /*argv*/) {
3    using Comp = std::complex<double>;
4    const Eigen::VectorXcd::Index n = 5;
5    Eigen::VectorXcd y(n);
6    Eigen::VectorXcd c(n);
7    Eigen::VectorXcd x(n);
8    y << Comp(1, 0), Comp(2, 1), Comp(3, 2), Comp(4, 3), Comp(5, 4);
9    Eigen::FFT fft; // DFT transform object
10   c = fft.fwd(y); // DTF of y, see Def. 4.2.1.18
11   x = fft.inv(c); // inverse DFT of c, see (4.2.1.20)
12
13   std::cout << "y = " << y.transpose() << std::endl
14             << "c = " << c.transpose() << std::endl
15             << "x = " << x.transpose() << std::endl;
16   return 0;
17 }
```

Some options can be given during the initialization of an **Eigen::FFT** object, see wiki and DOXYGEN documentation. The default options cover all use cases of DFT in this course.



In this course we use an EIGEN DFT extension implemented by Manuel Saladin, MSc CSE student at ETH Zurich, as part of a BSc thesis project, link. His code both substantially extends the DFT module originally shipped with EIGEN and aims for compatibility with the previous DFT interface.

◁ Manuel Saladin

- PYTHON-functions for discrete Fourier transform (and its inverse) are provided by the package `scipy.fft`

  DFT: `c=scipy.fft(y)` ↔ inverse DFT: `y=scipy.ifft(c)`,

  where `y` and `c` are `numpy`-arrays.

⌟

Review question(s) 4.2.1.23 (Diagonalizing circulant matrices and DFT)

(Q4.2.1.23.A) To practice complex arithmetic compute the discrete Fourier transform of $\mathbf{x} = [1, 2-\imath, -\imath, -1+2\imath]^\top$.

(Q4.2.1.23.B) Denote by $\mathrm{DFT}_n : \mathbb{C}^n \to \mathbb{C}^n$ the discrete Fourier transform,

$$(\mathrm{DFT}_n \mathbf{y})_k := \sum_{j=0}^{n-1} y_j \omega_n^{kj} , \quad k = 0, \ldots, n-1 , \quad \omega_n = \exp(-\tfrac{2\pi\imath}{n}) . \tag{4.2.1.19}$$

Show that

$$\mathbf{x}^{\mathrm{H}}\mathbf{y} = \tfrac{1}{n}\mathrm{DFT}_n(\mathbf{x})^{\mathrm{H}}\mathrm{DFT}_n(\mathbf{y}) .$$

Use the following lemma:

Lemma 4.2.1.14. Properties of Fourier matrix

The scaled Fourier-matrix $\frac{1}{\sqrt{n}}\mathbf{F}_n$ is unitary (→ Def. 6.3.1.2) : $\mathbf{F}_n^{-1} = \frac{1}{n}\mathbf{F}_n^{\mathrm{H}} = \frac{1}{n}\overline{\mathbf{F}}_n$.

(Q4.2.1.23.C) Explain, why the identity

$$\mathbf{C} = \mathbf{F}_n^{-1}\,\mathrm{diag}(d_0, \ldots, d_{n-1})\mathbf{F}_n , \quad [d_0, \ldots, d_{n-1}]^\top = \mathbf{F}_n[u_0, \ldots, u_{n-1}]^\top , \tag{4.2.1.17}$$

where $\mathbf{u} := [u_0, \ldots, u_{n-1}]^\top \in \mathbb{C}^n$ is the generating vector of the circulant matrix $\mathbf{C}$ according to the formula

$$(\mathbf{C})_{k,\ell} = (\mathbf{u})_{\ell-k \mod n} ,$$

still makes sense for the very special "circulant" matrix $\mathbf{C} = \mathbf{I}_n$.

(Q4.2.1.23.D) What are the eigenvalues and eigenvectors of the permutation matrix

$$\mathbf{P} = \begin{bmatrix} 0 & \ldots & & 0 & 1 \\ 1 & \ddots & & & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & 1 & 0 \end{bmatrix} \in \mathbb{C}^{n,n}$$

that effects a cyclic permutation of the entries of an n-vector?

(Q4.2.1.23.E) [SVD of circulant matrices] Based on the results

Lemma 4.2.1.16. Diagonalization of circulant matrices

For any circulant matrix $\mathbf{C} \in \mathbb{K}^{n,n}$, $c_{ij} = u_{i-j}$, $(u_k)_{k\in\mathbb{Z}}$ n-periodic sequence, holds true

$$\mathbf{C}\overline{\mathbf{F}}_n = \overline{\mathbf{F}}_n\,\mathrm{diag}(d_1, \ldots, d_n) \quad , \quad [d_0, \ldots, d_{n-1}]^\top = \mathbf{F}_n[u_0, \ldots, u_{n-1}]^\top .$$

and

Lemma 4.2.1.14. Properties of Fourier matrices

The scaled Fourier-matrix $\frac{1}{\sqrt{n}}\mathbf{F}_n$ is unitary (→ Def. 6.3.1.2) : $\mathbf{F}_n^{-1} = \frac{1}{n}\mathbf{F}_n^{\mathrm{H}} = \frac{1}{n}\overline{\mathbf{F}}_n$.

outline an efficient algorithm for computing the singular-value decomposition (SVD) of a circulant matrix.

Hint.

- Every $z \in \mathbb{C}$ can be written as $z = rz_0$ with $r \geq 0$ and $|z_0| = 1$.
- For matrices $\mathbf{A} \in \mathbb{C}^{m,n}$ the full SVD reads $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{H}}$ and involves *unitary* factors $\mathbf{U} \in \mathbb{C}^{m,m}$ and $\mathbf{V} \in \mathbb{C}^{n,n}$.

Also recall the definition of the (full) singular value decomposition:

Theorem 3.4.1.1. Singular value decomposition (SVD)

For any $\mathbf{A} \in \mathbb{K}^{m,n}$ there are unitary/orthogonal *matrices $\mathbf{U} \in \mathbb{K}^{m,m}$, $\mathbf{V} \in \mathbb{K}^{n,n}$ and a (generalized) diagonal matrix $\mathbf{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_p) \in \mathbb{R}^{m,n}$, $p := \min\{m,n\}$, $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$ such that*

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{H}} .$$

Definition 3.4.1.3. Singular value decomposition (SVD)

The decomposition $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{H}}$ of Thm. 3.4.1.1 is called **singular value decomposition (SVD)** of $\mathbf{A}$. The diagonal entries σ_i of $\mathbf{\Sigma}$ are the singular values of $\mathbf{A}$. The columns of $\mathbf{U}/\mathbf{V}$ are the left/right **singular vectors** of $\mathbf{A}$.

(Q4.2.1.23.F) [Diagonal circulant matrices] Characterize the set of all complex *diagonal circulant* $n \times n$-matrices, $n \in \mathbb{N}$. Is this set a subspace of $\mathbb{R}^{n,n}$?

(Q4.2.1.23.G) [Sum of discrete Fourier coefficients] Outline an efficient algorithm for computing

$$\sum_{j=1}^{n} (\mathrm{DFT}_n(\mathbf{y}))_j \quad \text{for given} \quad \mathbf{y} \in \mathbb{C}^n .$$

Hint.

$$\sum_{j=1}^{n} (\mathrm{DFT}_n(\mathbf{y}))_j = \mathbf{1}^{\top}\mathbf{F}_n\mathbf{y} = (\mathbf{F}_n\mathbf{1})^{\top}\mathbf{y} .$$

△

4.2.2 Discrete Convolution via Discrete Fourier Transform

Video tutorial for Section 4.2.2 "Discrete Convolution via DFT": (7 minutes) Download link, tablet notes

Coding the formula for the discrete periodic convolution of two periodic sequences from Def. 4.1.4.7,

$$(y_k) := (u_k) *_n (x_k) \quad , \quad y_k := \sum_{j=0}^{n-1} u_{k-j}x_j = \sum_{j=0}^{n-1} x_{k-j}u_j \, , \quad k \in \{0, \ldots, n-1\} ,$$

one could do this in a straightforward manner using two nested loops as in the following code, and with an asymptotic computational effort $O(n^2)$ for $n \to \infty$.

C++ code 4.2.2.1: Discrete periodic convolution: straightforward implementation ↪ GITLAB

```
2 Eigen::VectorXcd pconv(const Eigen::VectorXcd &u, const Eigen::VectorXcd &x) {
3   const Eigen::Index n = x.size();
4   Eigen::VectorXcd z = Eigen::VectorXcd::Zero(n);
```

```
5   // "naive" two-loop implementation of discrete periodic convolution
6   for (Eigen::Index k = 0; k < n; ++k) {
7     for (Eigen::Index j = 0, l = k; j <= k; ++j, --l) {
8       z[k] += u[l] * x[j];
9     }
10    for (Eigen::Index j = k + 1, l = n - 1; j < n; ++j, --l) {
11      z[k] += u[l] * x[j];
12    }
13  }
14  return z;
15 }
```

This codes relies on the associated vectors $\mathbf{u} = [u_0, \ldots, u_{n-1}]^\top \in \mathbb{C}^n$ and $\mathbf{x} = [x_0, \ldots, x_{n-1}]^\top \in \mathbb{C}^{n-1}$ for the sequences (u_k) and (x_k), respectively. Using this vectors, indexed from 0, the periodic convolution formula becomes

$$y_k = \sum_{j=0}^{k} (\mathbf{u})_{k-j} (\mathbf{x})_j + \sum_{j=k+1}^{n-1} (\mathbf{u})_{n+k-j} (\mathbf{x})_j \,.$$

Let us assume that a "magic" very efficient implementation of the discrete Fourier transform (DFT) is available ($\rightarrow$ Section 4.3). Then a much faster implementation of `pconv()` is possible and it is based on the link with the periodic discrete convolution of Def. 4.1.4.7. In § 4.1.4.11 we have seen that periodic convolution amounts to multiplication with a circulant matrix. In addition, (4.2.1.17) reduces multiplication with a circulant matrix to two multiplications with the Fourier matrix $\mathbf{F}_n$ (= DFT) and (componentwise) scaling operations. This suggests how to exploit the equivalence

$$\text{discrete periodic convolution} \quad z_k = \sum_{j=0}^{n-1} u_{k-j} x_j \quad (\rightarrow \text{Def. 4.1.4.7}), \quad k = 0, \ldots, n-1$$

$$\Updownarrow$$

$$\text{multiplication with circulant matrix } (\rightarrow \text{Def. 4.1.4.12}) \quad \mathbf{z} = \mathbf{C}\mathbf{x}, \quad \mathbf{C} := \left[u_{i-j}\right]_{i,j=1}^{n}.$$

Idea: (4.2.1.17) $\succ$ $\mathbf{z} = \mathbf{F}_n^{-1} \operatorname{diag}(\mathbf{F}_n \mathbf{u}) \mathbf{F}_n \mathbf{x}$

This formula is usually referred to as *convolution theorem*:

Theorem 4.2.2.2. Convolution theorem

The discrete periodic convolution $*_n$ between n-dimensional vectors $\mathbf{u} \in \mathbb{C}^n$ and $\mathbf{x} \in \mathbb{C}^n$ is equal to the inverse DFT of the component-wise product between the DFTs of $\mathbf{u}$ and $\mathbf{x}$; i.e.:

$$(\mathbf{u}) *_n (\mathbf{x}) := \left[\sum_{j=0}^{n-1} u_{(k-j) \mod n} x_j \right]_{k=0}^{n-1} = \mathbf{F}_n^{-1} \left[(\mathbf{F}_n \mathbf{u})_j (\mathbf{F}_n \mathbf{x})_j \right]_{j=1}^{n} . \tag{4.2.2.3}$$

Cast in a C++ function computing the periodic discrete convolution of two vectors the convolution theorem reads:

C++ code 4.2.2.4: Discrete periodic convolution: DFT implementation ⇝ GITLAB

```
2 inline Eigen::VectorXcd pconvfft(const Eigen::VectorXcd &u,
```

```
                                const Eigen::VectorXcd &x) {
  Eigen::FFT fft; // Helper object, see § 4.2.1.21
  // One-line implementation of (4.2.2.3)
  return fft.inv((fft.fwd(u)).cwiseProduct(fft.fwd(x)));
}
```

In Rem. 4.1.4.15 we learned that the discrete convolution of n-vectors ($\rightarrow$ Def. 4.1.3.3) can be accomplished by the *periodic* discrete convolution of $2n-1$-vectors (obtained by *zero padding*, see Rem. 4.1.4.15):

$$\mathbf{a},\mathbf{b} \in \mathbb{C}^n: \quad \mathbf{a} * \mathbf{b} = \begin{bmatrix} \mathbf{a} \\ \mathbf{0} \end{bmatrix} *_{2n-1} \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \end{bmatrix} \in \mathbb{C}^{2n-1} .$$

This idea underlies the following C++ implementation of the discrete convolution of two vectors.

C++ code 4.2.2.5: Implementation of discrete convolution ($\rightarrow$ Def. 4.1.3.3) based on periodic discrete convolution ➪ GITLAB

```
inline Eigen::VectorXcd fastconv(const Eigen::VectorXcd &h,
                                 const Eigen::VectorXcd &x) {
  assert(x.size() == h.size());
  const Eigen::Index n = h.size();
  // Zero padding, cf. (4.1.4.16), and periodic discrete convolution
  // of length 2n-1, Code 4.2.2.4
  return pconvfft(
      (Eigen::VectorXcd(2 * n - 1) << h, Eigen::VectorXcd::Zero(n - 1))
          .finished(),
      (Eigen::VectorXcd(2 * n - 1) << x, Eigen::VectorXcd::Zero(n - 1))
          .finished());
}
```

Review question(s) 4.2.2.6 (Discrete convolution via DFT)

(Q4.2.2.6.A) We saw that the discrete convolution of two vectors `h` and `x` of length n can be accomplished by

```
y = pconvfft(
(Eigen::VectorXcd(2 * n - 1) << h, Eigen::VectorXcd::Zero(n - 1))
.finished(),
(Eigen::VectorXcd(2 * n - 1) << x, Eigen::VectorXcd::Zero(n - 1))
.finished());
```

Here, `pconv()` implements periodic discrete convolution. How do you have to change the code so that it can compute the discrete convolution of two vectors $\mathbf{h} \in \mathbb{R}^n$, $\mathbf{x} \in \mathbb{R}^m$ for general $n, m \in \mathbb{N}$?

△

4.2.3 Frequency filtering via DFT

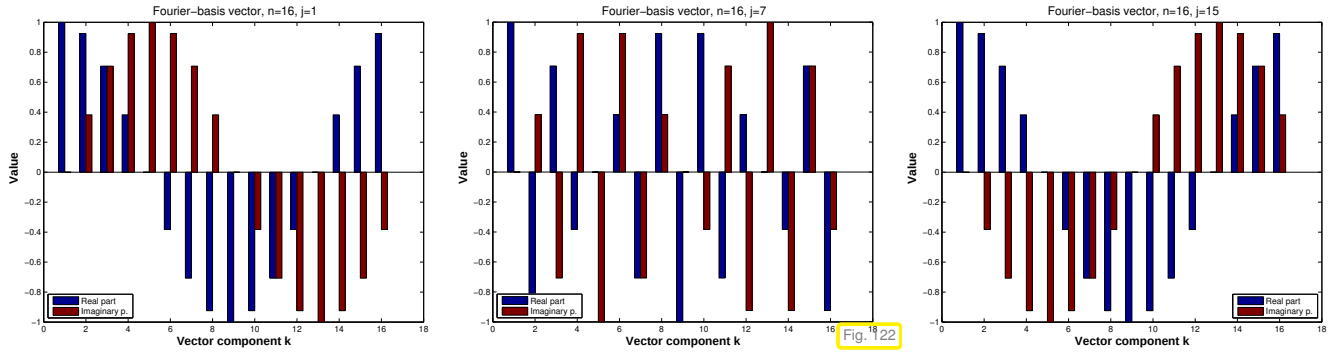
Video tutorial for Section 4.2.3 "Frequency filtering via DFT": (20 minutes) Download link, tablet notes

☞ A nice introduction by 3Blue1Brown can be found here.

The trigonometric basis vectors,

$$\mathbf{v}_k := [\exp(2\pi\imath jk/n)]_{j=0}^{n-1} = \left[\cos\left(\frac{2\pi jk}{n}\right) + \imath \sin\left(\frac{2\pi jk}{n}\right)\right]_{j=0}^{n-1} \in \mathbb{C}^n \,, \tag{4.2.3.1}$$

when interpreted as time-periodic signals, represent harmonic oscillations. This is illustrated when plotting some vectors of the trigonometric basis ($n = 16$):



"slow oscillation/low frequency" "fast oscillation/high frequency" "slow oscillation/low frequency"

Fig. 122

Dominant coefficients of a signal after transformation to trigonometric basis indicate dominant frequency components.

We say that the coefficients of a signal w.r.t. the trigonometric basis represent the signal in **frequency domain**, whereas original signal represented in the "pulse basis" of coordinate vectors is given in **time domain**.

§4.2.3.2 (Frequency decomposition of a signal) Since the trigonometric basis vectors form the columns of the (complex conjugate) Fourier matrix DFT (4.2.1.19) and inverse DFT (4.2.1.20),

$$c_k = \sum_{j=0}^{n-1} y_j\, \omega_n^{kj} \quad \Leftrightarrow \quad y_j = \frac{1}{n}\sum_{k=0}^{n-1} c_k\, \omega_n^{-kj} \tag{4.2.1.20}$$

effect the transformation from the "pulse basis" into a (scaled) trigonometric basis and vice versa. Thus, they convert time-domain and frequeny-domain representation of a signal into each other. To see this more clearly, we examine a real-valued signal of length $n = 2m+1$, $m \in \mathbb{N}$: $y_k \in \mathbb{R}$. Its DFT yields $c_0, \ldots, c_{n-1}$ and theses coefficients satisfy $c_k = \overline{c}_{n-k}$, because $\omega_n^{kj} = \overline{\omega}_n^{(n-k)j}$. Using this relationship we can write the original signal as a linear combination of sampled trigonometric functions with "frequencies" $k = 0, \ldots, m$:

$$\begin{aligned} ny_j &= c_0 + \sum_{k=1}^{m} c_k \omega_n^{-kj} + \sum_{k=m+1}^{2m} c_k \omega_n^{-kj} = c_0 + \sum_{k=1}^{m} c_k \omega_n^{-kj} + c_{n-k}\omega_n^{(k-n)j} \\ &= c_0 + 2\sum_{k=1}^{m} \mathrm{Re}(c_k)\cos(2\pi kj/n) + \mathrm{Im}(c_k)\sin(2\pi kj/n) \,, \quad j = 0, \ldots, n-1 \,, \end{aligned}$$

since $\omega_n^{\ell} = \cos(2\pi\ell/n) - i\sin(2\pi\ell/n)$.

The moduli $|c_k|, |c_{n-k}|$ of the coefficients obtained by DFT measure the strength with which an oscillation with frequency k is represented in the signal, $0 \le k \le \lfloor \frac{n}{2} \rfloor$.

EXAMPLE 4.2.3.3 (Frequency identification with DFT) The following C++ code generates a periodic signal composed of two base frequencies and distorts its by adding large random noise:

C++ code 4.2.3.4: Generation of noisy sinusoidal signal ⇝ GITLAB

```
VectorXd signalgen() {
  const int N = 64;
  const ArrayXd t = ArrayXd::LinSpaced(N, 0, N);
  const VectorXd x = ((2*M_PI/N*t).sin() + (14*M_PI/N*t).sin()).matrix();
  return x + VectorXd::Random(N);
}
```

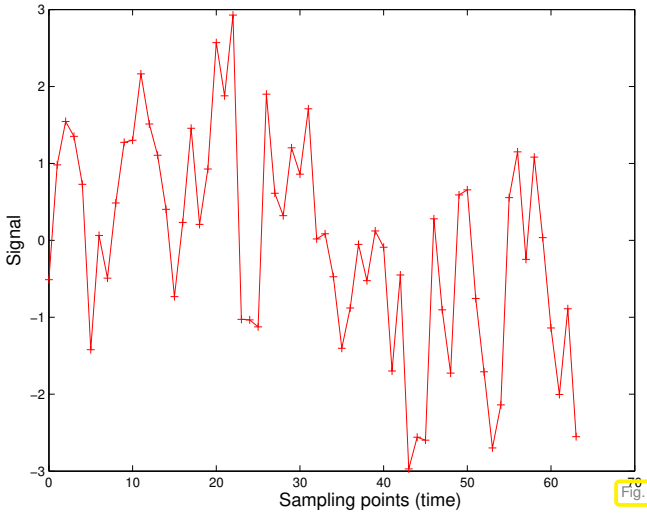


Fig. 123

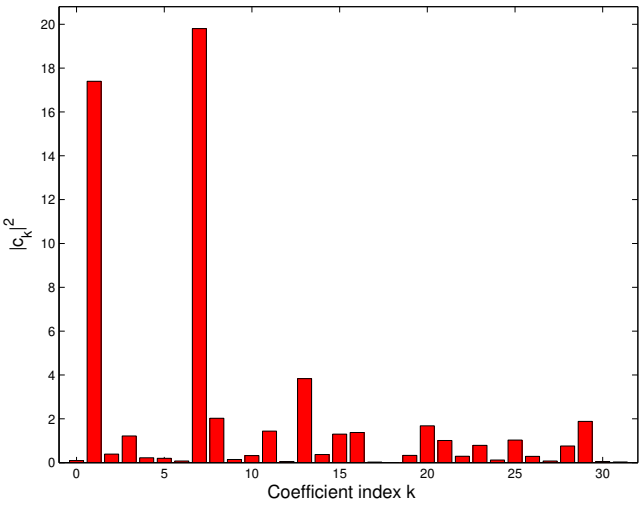


Fig. 124

Looking at the time-domain plot of the signal given in Fig. 123 (C++ code ⇝ GITLAB) it is hard to discern the underlying base frequencies. However, we observe that frequencies present in unperturbed signal are clearly evident in the frequency-domain representation after DFT. ⌟

EXAMPLE 4.2.3.5 (Detecting periodicity in data)

Google provides information about the frequency of certain words in web searches. In this example we study the data of for the word "Vorlesungsverzeichnis".

Raw data: weekly occurrences of "Vorlesungsverzeichnis" in Google web searches ▷

Some periodic pattern in the data is conspicuous to the "naked eye". How can an algorithm detect inherent periodicity of the data and find out the periods? Of course, by means of DFT!

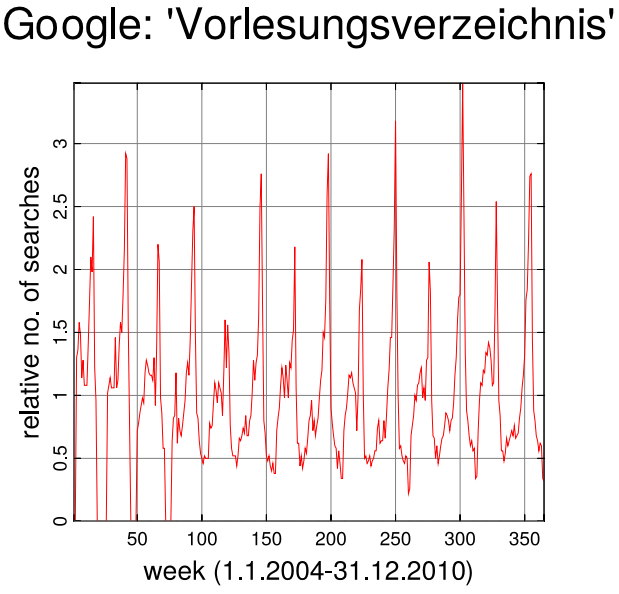


Fig. 125

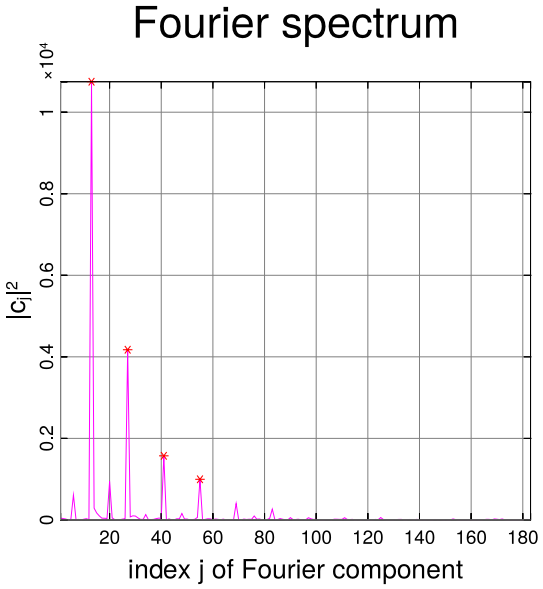


Fig. 126

DFT of the signal $(y_0, \ldots, y_{n-1})$, corresponding to the number of searches containing "Vorlesungsverzeichnis" in a large number n of consecutive weeks, yields the coefficients c_j, $j = 0, \ldots, n-1$.

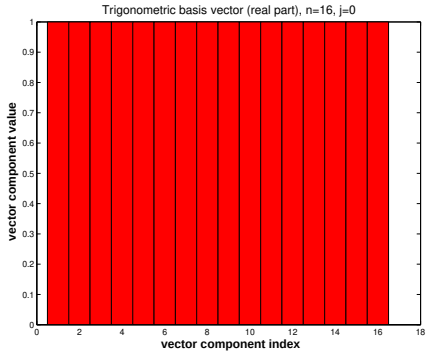
The plot beside (C++ code ➙ GITLAB) displays the "(Fourier) power spectrum" of the signal, the values $|c_j|^2$, $0 \leq j \leq \lfloor \frac{n}{2} \rfloor$.

We see that pronounced peaks in the power spectrum point to periodic structure of the data. The coefficient indices of the peaks of the power spectrum tell us dominant frequencies present in the data and, after inversion, the lengths of dominant periods.

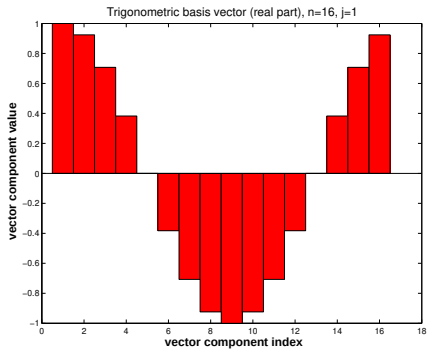
We can state the main message of this example as follows:

DFT is a computer's eye for periodic patterns in data

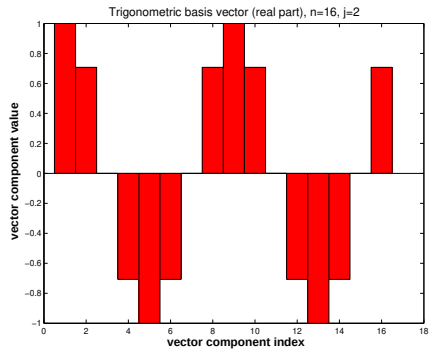
§4.2.3.6 ("Low" and "high" frequencies) Again, look at plots of real parts of trigonometric basis vectors $(\mathbf{F}_n)_{:,j}$ (= columns of Fourier matrix), $n = 16$.



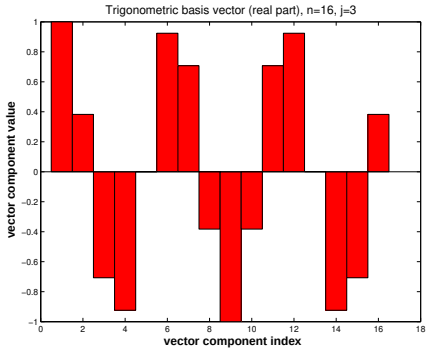
$\mathrm{Re}(\mathbf{F}_{16})_{:,0}$



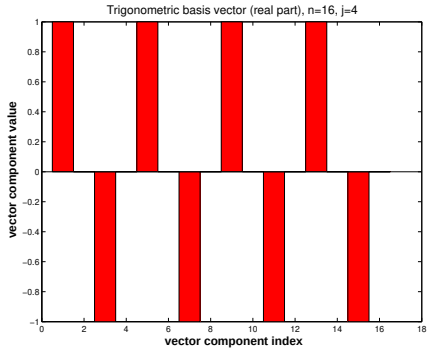
$\mathrm{Re}(\mathbf{F}_{16})_{:,1}$



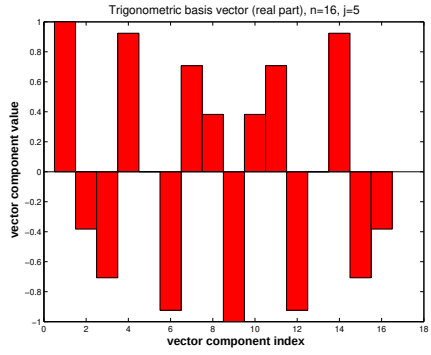
$\mathrm{Re}(\mathbf{F}_{16})_{:,2}$



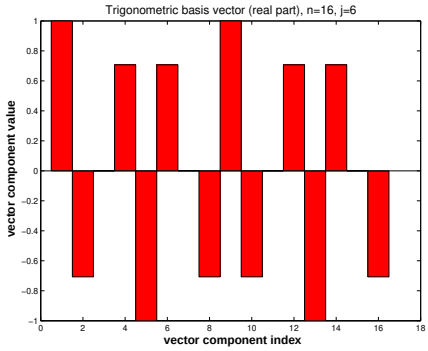
$\mathrm{Re}(\mathbf{F}_{16})_{:,3}$



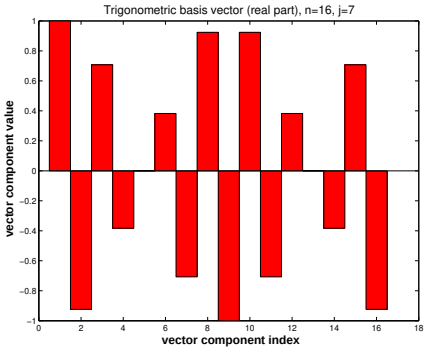
$\mathrm{Re}(\mathbf{F}_{16})_{:,4}$



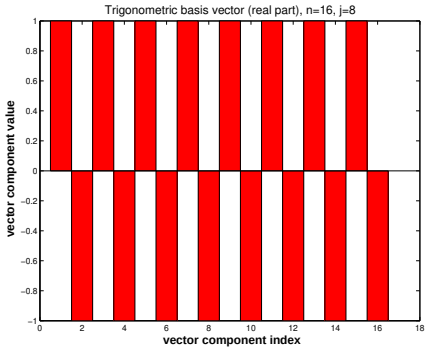
$\mathrm{Re}(\mathbf{F}_{16})_{:,5}$



$\mathrm{Re}(\mathbf{F}_{16})_{:,6}$



$\mathrm{Re}(\mathbf{F}_{16})_{:,7}$



$\mathrm{Re}(\mathbf{F}_{16})_{:,8}$

What about the remaining columns? They are just the first ones "wrapped around":

$$(\mathbf{F}_n)_{:,n-k} = \left[\omega_n^{j(n-k)}\right]_{j=0}^{n-1} \overset{(4.2.1.7)}{=} \left[\omega_n^{-jk}\right]_{j=0}^{n-1} = (\overline{\mathbf{F}}_n)_{:,k}\,, \quad k = 0,\ldots,n-1\,.$$

(Here we adopt C++ indexing and count the columns of the matrix from 0.)

Visually, the different basis vectors represent oscillatory signals with different frequencies.

This is also revealed by elementary trigonometric identities:

$$\begin{aligned}\mathrm{Re}(\mathbf{F}_n)_{:,j} &= \left(\mathrm{Re}\,\omega_n^{(j-1)k}\right)_{k=0}^{n-1} \\ &= \left(\mathrm{Re}\exp(-2\pi i(j-1)k/n)\right)_{k=0}^{n-1} \\ &= \left(\cos(2\pi(j-1)x)\right)_{x=0,\frac{1}{n},\ldots,1-\frac{1}{n}}\,.\end{aligned}$$

- Slow oscillations/low frequencies correspond to indices $j \approx 1$ and $j \approx n$.
- Fast oscillations/high frequencies correspond to indices $j \approx n/2$.

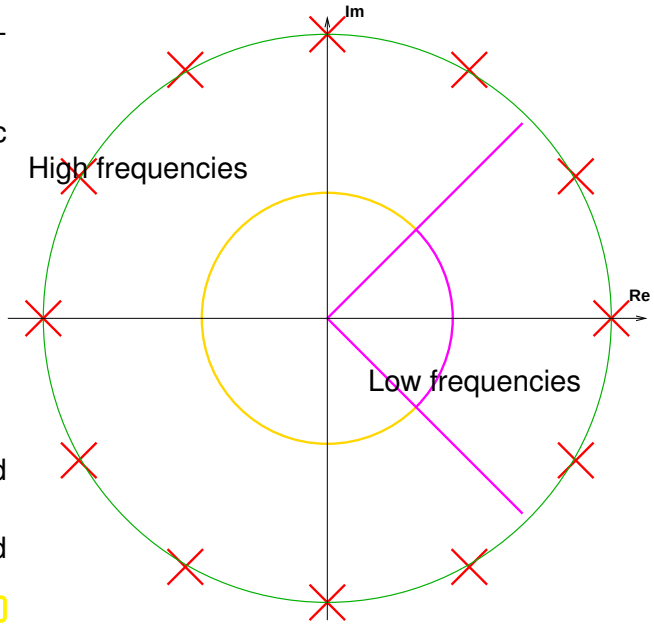


Fig. 127

The task of frequency filtering is to suppress or enhance a predefined range of frequencies contained in a signal.

❶ Perform DFT of the signal: $(y_0,\ldots,y_{n-1}) \overset{\text{DFT}}{\longmapsto} (c_0,\ldots,c_{n-1})$

❷ Operate on Fourier coefficients: $(c_0,\ldots,c_{n-1}) \mapsto (\widetilde{c}_0,\ldots,\widetilde{c}_{n-1})$

❸ Filtered signal by inverse DFT: $(\widetilde{c}_0,\ldots,\widetilde{c}_{n-1}) \overset{\text{DFT}^{-1}}{\longmapsto} (\widetilde{y}_0,\ldots,\widetilde{y}_{n-1})$

The following code does digital low-pass and high-pass filtering of a signal based on DFT and inverse DFT. It sets the obtained Fourier coefficients corresponding to high/low frequencies to zero and afterwards transforms back to time domain.

C++ code 4.2.3.7: DFT-based frequency filtering ➙ GITLAB

```
2 inline void freqfilter(const VectorXd &y, int k, VectorXd &low,
3                        VectorXd &high) {
4   const VectorXd::Index n = y.size();
5   if (n % 2 != 0) {
6     throw std::runtime_error("Even vector length required!");
```

```
  }
  const VectorXd::Index m = y.size() / 2;

  Eigen::FFT fft;                  // DFT helper object
  const VectorXcd c = fft.fwd(y); // Perform DFT of input vector

  VectorXcd clow = c;
  // Set high frequency coefficients to zero, Fig. 127
  for (int j = -k; j <= +k; ++j) {
    clow(m + j) = 0;
  }
  // (Complementary) vector of high frequency coefficients
  const VectorXcd chigh = c - clow;

  // Recover filtered time-domain signals
  low = fft.inv(clow).real();
  high = fft.inv(chigh).real();
}
```

The code could be optimised by exploiting $y_j \in \mathbb{R}$ and $c_{n/2-k} = \overline{c}_{n/2+k}$.

Summary: Map `y` $\mapsto$ `low` (in Code 4.2.3.7) $\hat{=}$ low pass filter .

Map `y` $\mapsto$ `high` (in Code 4.2.3.7) $\hat{=}$ high pass filter .

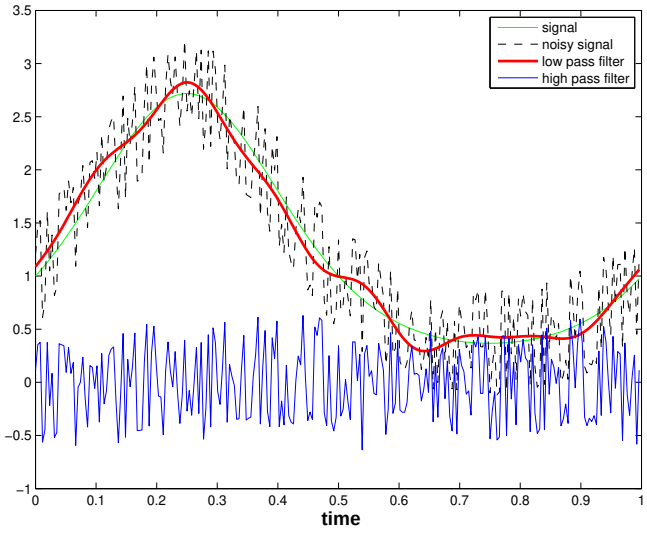
⌟

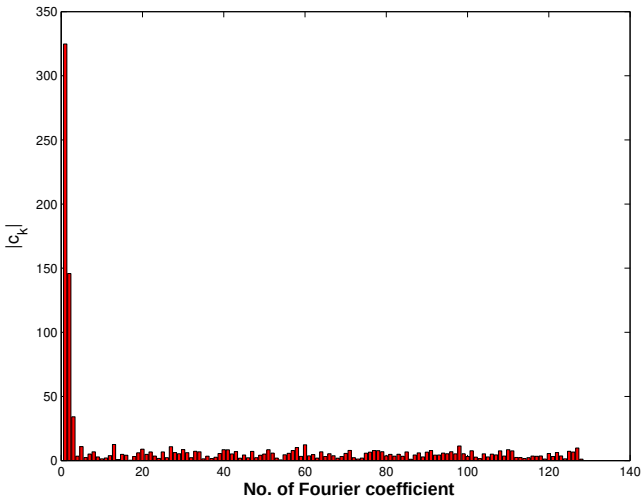
EXAMPLE 4.2.3.8 (Denoising by frequency filtering)

Signal perturbed by "deterministic noise":

```
n = 256; y = exp(sin(2*pi*((0:n-1)')/n)) + 0.5*sin(exp(1:n)');
```

We performed frequency filtering by Code 4.2.3.7 with $k = 120$.





Low pass filtering can be used for *denoising*, that is, the removal of high frequency perturbations of a signal. ⌟

EXAMPLE 4.2.3.9 (Sound filtering by DFT) Frequency filtering is ubiquitous in sound processing. Here we demonstrate it in PYTHON ⇝ GITLAB, which offers tools for audio processing through the `sounddevice` module.

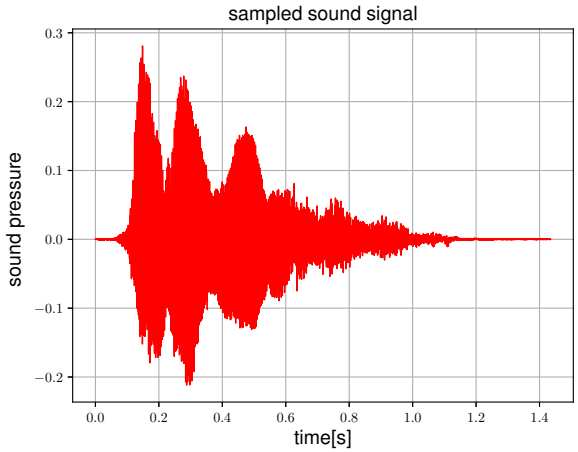


Fig. 128

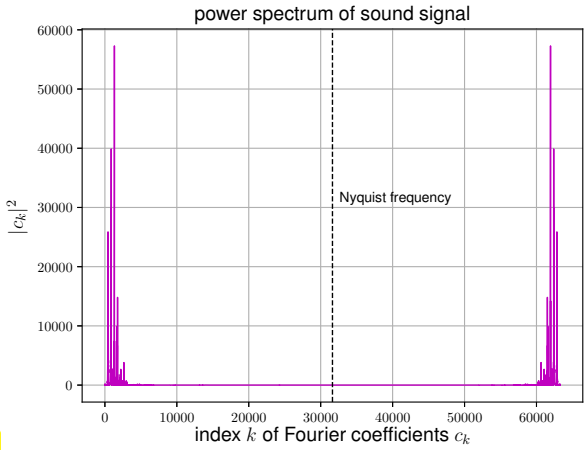


Fig. 129

The audio signal (duration ≈ 1.5s) of a human voice is plotted in time domain (vector $\mathbf{y} \in \mathbb{R}^n$, $n = 63274$) and as a power spectrum in frequency domain. The power spectrum of a signal $\mathbf{y} \in \mathbb{C}^n$ is the vector $\left(|c_j|^2\right)_{j=0}^{n-1}$, where $\mathbf{c} = \mathrm{DFT}_n \mathbf{y} = \mathbf{F}_n \mathbf{y}$ is the discrete Fourier transform of $\mathbf{y}$.

We see that the bulk of the signal's power $\|\mathbf{y}\|_2^2$ is contained in the low-frequency components. The paves the way for compressing the signal by low-pass filtering, that is, by dropping its high-frequency components and storing or transmitting only the discrete Fourier coefficients belonging to low frequencies. Refer to § 4.2.3.6 for precise information about the association of Fourier coefficients c_j with low and high frequencies.

Below we plot the squared moduli $|c_j|^2$ of the Fourier coefficients belonging to low frequencies and the low-pass filtered sound signals for different cut-off frequencies. Taking into account only low-frequency discrete Fourier coefficients does not severely distort the sound signal.

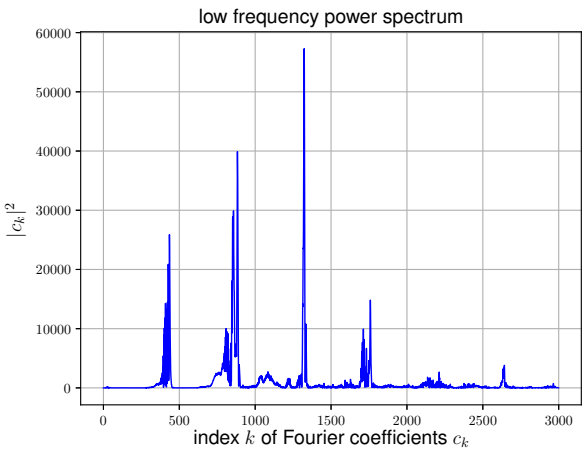


Fig. 130

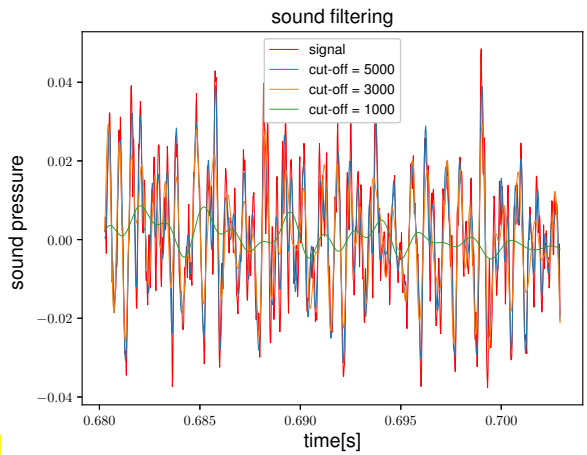


Fig. 131

Remark 4.2.3.10 (Linear Filtering) Low-pass and high-pass filtering via DFT implement a very general policy underlying general linear filtering of finite, time-discrete signals, represented by vectors $\mathbf{y} \in \mathbb{C}^n$

❶ Compute the coefficients of an *alternative basis representation* of $\mathbf{y}$.

❷ Apply some linear mapping to the obtain coefficient vector $\mathbf{c} \in \mathbb{C}^n$, yielding $\widetilde{\mathbf{c}}$.

❸ Recover the representation of $\widetilde{\mathbf{c}}$ in the standard basis of $\mathbb{C}^n$.

Review question(s) 4.2.3.11 (Frequency Filtering via DFT)

(Q4.2.3.11.A) Let $\mathbf{y} \in \mathbb{R}^n$, $n = 2^k$, $k \in \mathbb{N}$, be a vector describing an analog time-discrete finite signal.

Denote by $\mathbf{c} \in \mathbb{C}^n$ its discrete Fourier transform. Which components of $\mathbf{c}$ are related to "low-frequency content", and which to "high-frequency content"?

△

4.2.4 Real DFT

Every time-discrete signal obtained from sampling a time-dependent physical quantity will yield a real vector. Of couse, a real vector contains only half the information compared to complex vector of the same length. We aim to exploit this for a more efficient implementation of DFT.

Task: Efficient implementation of DFT (Def. 4.2.1.18) $(c_0, \ldots, c_{n-1})$ for *real* coefficients $(y_0, \ldots, y_{n-1})^\top \in \mathbb{R}^n$, $n = 2m$, $m \in \mathbb{N}$.

If $y_j \in \mathbb{R}$ in the DFT formula (4.2.1.19), we obtain redundant output: since $\omega_n^{(n-k)j} = \overline{\omega}_n^{kj}$, $k = 0, \ldots, n-1$, we conclude the following relationship between discrete Fourier coefficients of a real-valued signal.

$$\overline{c}_k = \sum_{j=0}^{n-1} y_j \overline{\omega}_n^{kj} = \sum_{j=0}^{n-1} y_j \omega_n^{(n-k)j} = c_{n-k} \;, \quad k = 1, \ldots, n-1 \;.$$

Idea: Map $\mathbf{y} \in \mathbb{R}^n$ to a vector $\mathbb{C}^m$ and use DFT of length m on it.

$$h_k = \sum_{j=0}^{m-1} (y_{2j} + i y_{2j+1})\, \omega_m^{jk} = \sum_{j=0}^{m-1} y_{2j}\, \omega_m^{jk} + i \cdot \sum_{j=0}^{m-1} y_{2j+1}\, \omega_m^{jk} \;, \tag{4.2.4.1}$$

$$\overline{h}_{m-k} = \sum_{j=0}^{m-1} \overline{y_{2j} + i y_{2j+1}}\, \overline{\omega}_m^{j(m-k)} = \sum_{j=0}^{m-1} y_{2j}\, \omega_m^{jk} - i \cdot \sum_{j=0}^{m-1} y_{2j+1}\, \omega_m^{jk} \;. \tag{4.2.4.2}$$

Thus, we can recover the framed sums from suitable combinations of the discrete Fourier coefficients $h_k \in \mathbb{C}$, $k = 0, \ldots, m-1$

$$\Rightarrow \quad \sum_{j=0}^{m-1} y_{2j}\, \omega_m^{jk} = \tfrac{1}{2}(h_k + \overline{h}_{m-k}) \;, \quad \sum_{j=0}^{m-1} y_{2j+1}\, \omega_m^{jk} = -\tfrac{1}{2} i (h_k - \overline{h}_{m-k}) \;.$$

Use simple identities for roots of unity to split the DFT of $\mathbf{y}$ into two sums:

$$c_k = \sum_{j=0}^{n-1} y_j\, \omega_n^{jk} = \sum_{j=0}^{m-1} y_{2j}\, \omega_m^{jk} + \omega_n^k \cdot \sum_{j=0}^{m-1} y_{2j+1}\, \omega_m^{jk} \;. \tag{4.2.4.3}$$

$$\blacktriangleright \quad \begin{cases} c_k = \tfrac{1}{2}(h_k + \overline{h}_{m-k}) - \tfrac{1}{2} i \omega_n^k (h_k - \overline{h}_{m-k}) \;, \quad k = 0, \ldots, m-1 \;, \\ c_m = \mathrm{Re}\{h_0\} - \mathrm{Im}\{h_0\} \;, \\ c_k = \overline{c}_{n-k} \;, \quad k = m+1, \ldots, n-1 \;. \end{cases} \tag{4.2.4.4}$$

C++ code 4.2.4.5: DFT of real vectors of length $n/2$ ➽ GITLAB

```
2 // Perform fft on a real vector y of even
3 // length and return (complex) coefficients in c
4 // Note:  The new EIGEN DFT method fwd() has this already implemented
      and
```

```
// we could also just call:  c = fft.fwd(y);
inline void fftreal(const VectorXd &y, VectorXcd &c) {
  const Eigen::Index n = y.size();
  const Eigen::Index m = n / 2;
  if (n % 2 != 0) {
    std::cout << "n must be even!\n";
    return;
  }

  // Step I: compute h from (4.2.4.1), (4.2.4.2)
  const std::complex<double> i(0, 1);  // Imaginary unit
  VectorXcd yc(m);
  for (Eigen::Index j = 0; j < m; ++j) {
    yc(j) = y(2 * j) + i * y(2 * j + 1);
  }

  Eigen::FFT fft;
  VectorXcd d = fft.fwd(yc);
  VectorXcd h(m + 1);
  h << d, d(0);

  c.resize(n);
  // Step II: implementation of (4.2.4.4)
  for (Eigen::Index k = 0; k < m; ++k) {
    c(k) = (h(k) + std::conj(h(m - k))) / 2. -
           i / 2. *
               std::exp(-2. * static_cast<double>(k) / static_cast<double>(n) *
                        M_PI * i) *
               (h(k) - std::conj(h(m - k)));
  }
  c(m) = std::real(h(0)) - std::imag(h(0));
  for (Eigen::Index k = m + 1; k < n; ++k) {
    c(k) = std::conj(c(n - k));
  }
}
```

Review question(s) 4.2.4.6 (Frequency Filtering via DFT and real DFT)

(Q4.2.4.6.A) For $\mathbf{y} \in \mathbb{R}^n$, what is the result of the linear mapping

$$\mathbf{y} \mapsto \mathrm{Re}\Big\{\mathrm{DFT}_n^{-1}\Big([(\mathrm{DFT}_n\mathbf{y})_1, 0, \ldots, 0]^\top\Big)\Big\} \quad ?$$

Here Re extracts the real parts of the components of a complex vector.

(Q4.2.4.6.B) Outline the implementation of a C++ function

```
std::vector<std::pair<int, std::complex<double>>>
selectDominantFrequencies(const Eigen::VectorXd &y, double tol);
```

that returns a sequence of pairs

$$(j, c_j) \in \mathbb{N}_0 \times \mathbb{C}\,, \quad j \in J^* := \operatorname*{argmin}_{J \subset \{0,\ldots,\lfloor n/2-1 \rfloor\}} \left\{ \sharp J : \sum_{j \in J} |c_j|^2 \geq (1 - \mathtt{tol}) \sum_{j=0}^{\lceil n/2-1 \rceil} |c_j|^2 \right\},$$

for $0 \leq \mathtt{tol} < 1$. Here $\left[c_j\right]_{j=0}^{n-1} \in \mathbb{C}^n$ is the discrete Fourier transform of the real argument vector $\mathbf{y} \in \mathbb{R}^n$: $\mathbf{c} := \mathrm{DFT}_n(\mathbf{y})$.

Discuss to what extent this function can be used for the compression of a sound signal.

(Q4.2.4.6.C) How would you implement a C++ function

```
Eigen::VectorXd reconstructFromFrequencies(
const std::vector<std::pair<int, std::complex<double>> &f);
```

that takes the output of `selectDominantFrequencies()` from Question (**Q**4.2.4.6.B) and returns the compressed signal in time domain?

Take into account that `selectDominantFrequencies()` merely looks at the first half of the discrete Fourier coefficients.

△

4.2.5 Two-dimensional DFT

Video tutorial for Section 4.2.5 "Two-Dimensional DFT": (20 minutes) Download link, tablet notes

Finite time-discrete signals are naturally described by vectors, recall § 4.0.0.1. They can be regarded as one-dimensional, and typical specimens are audio data given in WAV (Waveform Audio) format. Other types of data also have to be sent through channels, most importantly, images that can be viewed as two-dimensional data. The natural linear-algebra style representation of an image is a matrix, see Ex. 3.4.4.24.

In this we study the frequency decomposition of matrices. Due to the natural analogy

$$\text{one-dimensional data ("audio signal")} \quad \longleftrightarrow \quad \text{vector } \mathbf{y} \in \mathbb{R}^n,$$
$$\text{two-dimensional data ("image")} \quad \longleftrightarrow \quad \text{matrix. } \mathbf{Y} \in \mathbb{R}^{m,n},$$

these techniques are of fundamental importance for **image processing**.

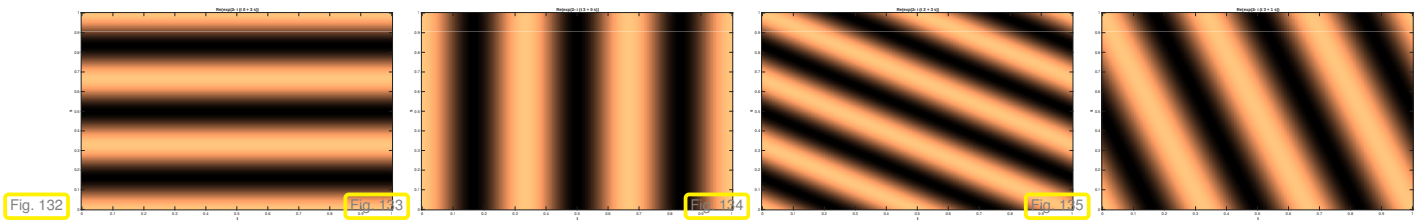
§4.2.5.1 (Matrix Fourier modes) The (inverse) discrete Fourier transform of a vector computes its coefficient of the representation in a basis provided by the columns of the Fourier matrix $\mathbf{F}_n$. The k-th column can be obtained by sampling harmonic oscillations of frequency k, $k = 0, \ldots, n-1$:

$$(\mathbf{F}_n)_{:,k} = \left[\exp(-2\pi\imath k t_j)\right]_{j=0}^{n-1} \quad t_j := \tfrac{j}{n}\,, \quad k = 0, \ldots, n-1\,.$$

What are the 2D counterpart of these vectors? The matrices obtained by sampling products of complex exponential (= trigonometric) functions, e.g., for $m, n \in \mathbb{N}$,

$$(t,s) \mapsto \exp(2\pi\imath k t) \cdot \exp(2\pi\imath \ell s) = \exp(2\pi\imath(kt + \ell s))\,, \quad \begin{matrix} k \in \{0, \ldots, m-1\}, \\ \ell \in \{0, \ldots, n-1\}\,, \end{matrix} \tag{4.2.5.2}$$

at the points $(t_j := \frac{j}{m}, s_r := \frac{r}{n})$, $j \in \{0, \ldots, m-1\}$, $r \in \{0, \ldots, n-1\}$. The functions from (4.2.5.2) are two-dimensional **plane waves** and their wavelength and direction is controlled by the parameters k and ℓ, as can be seen from the following plots of their real parts.



$k = 0, \ell = 3$ $k = 0, \ell = 3$ $k = 0, \ell = 3$ $k = 0, \ell = 3$

The matrices obtained by 2D sampling provide a two-dimensional trigonometric basis of $\mathbb{C}^{m,n}$, whose element are given by the *tensor product matrices*

$$\left\{(\mathbf{F}_m)_{:,j}(\mathbf{F}_n)_{:,\ell}^{\top}, 1 \le j \le m, 1 \le \ell \le n\right\} \subset \mathbb{C}^{m,n} . \tag{4.2.5.3}$$

Let a matrix $\mathbf{C} \in \mathbb{C}^{m,n}$ be given as a linear combination of these basis matrices with coefficients $y_{j_1,j_2} \in \mathbb{C}$, $0 \le j_1 < m$, $0 \le j_2 < n$:

$$\mathbf{C} = \sum_{j_1=0}^{m-1}\sum_{j_2=0}^{n-1} y_{j_1,j_2}\,(\mathbf{F}_m)_{:,j_1}(\mathbf{F}_n)_{:,j_2}^{\top} . \tag{4.2.5.4}$$

Then the entries of $\mathbf{C}$ can be computed by two *nested* discrete Fourier transforms:

$$(\mathbf{C})_{k_1,k_2} = \sum_{j_1=0}^{m-1}\sum_{j_2=0}^{n-1} y_{j_1,j_2}\,\omega_m^{j_1k_1}\omega_n^{j_2k_2} = \sum_{j_1=0}^{m-1}\omega_m^{j_1k_1}\left(\sum_{j_2=0}^{n-1}\omega_n^{j_2k_2}y_{j_1,j_2}\right) , \quad 0 \le k_1 < m , 0 \le k_2 < n .$$

Note that C++ indexing is applied throughout.

⌟

The coefficients $y_{j_1,j_2} \in \mathbb{C}$, $0 \le j_1 < m$, $0 \le j_2 < n$, can also be regarded as entries of a matrix $\mathbf{Y} \in \mathbb{C}^{m,n}$. Thus we can rewrite the above expressions: for all $0 \le k_1 < m$, $0 \le k_2 < n$,

$$(\mathbf{C})_{k_1,k_2} = \sum_{j_1=0}^{m-1}\sum_{j_2=0}^{n-1}(\mathbf{Y})_{j_1,j_2}\,\omega_m^{j_1k_1}\omega_n^{j_2k_2} = \sum_{j_1=0}^{m-1}\left(\mathbf{F}_n(\mathbf{Y})_{j_1,:}\right)_{k_2}\omega_m^{j_1k_1} , \quad 0 \le k_1 < m , 0 \le k_2 < n \tag{4.2.5.5}$$

$$\blacktriangleright \quad \boxed{\mathbf{C} = \mathbf{F}_m(\mathbf{F}_n\mathbf{Y}^{\top})^{\top} = \mathbf{F}_m\mathbf{Y}\mathbf{F}_n} ,$$

because $\mathbf{F}_n^{\top} = \mathbf{F}_n$. These formulas define the **two-dimensional discrete Fourier transform** of the matrix $\mathbf{Y} \in \mathbb{C}^{m,n}$. We abbreviate it by $\mathrm{DFT}_{m,n} : \mathbb{C}^{m,n} \to \mathbb{C}^{m,n}$.

From Lemma 4.2.1.14 we immediately get the inversion formula:

$$\mathbf{C} := \mathrm{DFT}_{m,n}(\mathbf{Y}) = \sum_{j_1=0}^{m-1}\sum_{j_2=0}^{n-1} y_{j_1,j_2}\,(\mathbf{F}_m)_{:,j_1}(\mathbf{F}_n)_{:,j_2}^{\top} \quad \Rightarrow \quad \boxed{\mathbf{Y} = \mathbf{F}_m^{-1}\mathbf{C}\mathbf{F}_n^{-1} = \tfrac{1}{mn}\overline{\mathbf{F}}_m\mathbf{C}\overline{\mathbf{F}}_n} . \tag{4.2.5.6}$$

The following two codes implement (4.2.5.5) and (4.2.5.6) using the DFT facilities of EIGEN.

C++ code 4.2.5.7: Two-dimensional discrete Fourier transform ➙ GITLAB

```
2  // 2D DFT implementation using 1D DFTs
3  template <typename Scalar>
4  void fft2(Eigen::MatrixXcd &C, const Eigen::MatrixBase<Scalar> &Y) {
5    using idx_t = Eigen::MatrixXcd::Index;
6    const idx_t m = Y.rows();
7    const idx_t n = Y.cols();
8    C.resize(m, n);
9    Eigen::MatrixXcd tmp(m, n);
10
11   Eigen::FFT fft; // Helper class for DFT
12   // Transform rows of matrix Y
13   for (idx_t k = 0; k < m; k++) {
14     tmp.row(k) = fft.fwd(Y.row(k));
15   }
```

```
16
17   // Transform columns of temporary matrix
18   for (idx_t k = 0; k < n; k++) {
19     C.col(k) = fft.fwd(tmp.col(k));
20   }
21 }
```

C++ code 4.2.5.8: Inverse two-dimensional discrete Fourier transform ⇢ GITLAB

```
2  // 2D Inverse DFT Implementation using a 2D DFT
3  template <typename Scalar>
4  void ifft2(Eigen::MatrixXcd &C, const Eigen::MatrixBase<Scalar> &Y) {
5    using idx_t = Eigen::MatrixXcd::Index;
6    const idx_t m = Y.rows();
7    const idx_t n = Y.cols();
8    fft2(C, Y.conjugate());
9    C = C.conjugate() / (m * n);
10 }
```

Remark 4.2.5.9 (Two-dimensional DFT in EIGEN and PYTHON)

EIGEN: The new EIGEN DFT interface by M. Saladin directly provides the functionality of `fft2()` from Code 4.2.5.7 and `ifft2()` from Code 4.2.5.8 through the `fwd()` and `inv()` member functions when supplied with matrix arguments instead of vector arguments!

PYTHON: The two-dimensional DFT is provided by the PYTHON function:

```
numpy.fft2(Y) .
```

⌟

§4.2.5.10 (Periodic convolution of matrices) In Section 4.2.2 we linked (periodic) convolutions of vectors $\in \mathbb{C}^n$

$$\mathbf{y} = \mathbf{p} *_n \mathbf{x} := \left[\sum_{j=0}^{n-1} x_j \, p_{(k-j) \mod n} \right]_{k=0}^{n-1}, \quad \mathbf{p}, \mathbf{x} \in \mathbb{C}^n , \tag{4.2.5.11}$$

and discrete Fourier transforms and found a convolution theorem.

Theorem 4.2.2.2. Convolution theorem

The discrete periodic convolution $*_n$ between n-dimensional vectors $\mathbf{u} \in \mathbb{C}^n$ and $\mathbf{x} \in \mathbb{C}^n$ is equal to the inverse DFT of the component-wise product between the DFTs of $\mathbf{u}$ and $\mathbf{x}$; i.e.:

$$(\mathbf{u}) *_n (\mathbf{x}) := \left[\sum_{j=0}^{n-1} u_{(k-j) \mod n} x_j \right]_{k=0}^{n-1} = \mathbf{F}_n^{-1} \left[(\mathbf{F}_n \mathbf{u})_j (\mathbf{F}_n \mathbf{x})_j \right]_{j=1}^{n} . \tag{4.2.2.3}$$

A similar result can also be achieved in two dimensions. We consider the following bilinear mapping $\mathsf{B} : \mathbb{C}^{m,n} \times \mathbb{C}^{m,n} \to \mathbb{C}^{m,n}$ (C++ indexing)

$$(\mathsf{B}(\mathbf{X}, \mathbf{Y}))_{k,\ell} = \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} (\mathbf{X})_{i,j} (\mathbf{Y})_{(k-i) \mod m, (\ell-j) \mod n}, \quad \begin{array}{l} k = 0, \ldots, m-1 , \\ \ell = 0, \ldots, n-1 , \end{array} \tag{4.2.5.12}$$

Here, as in (4.2.5.11), $\mod *$ designates the remainder of integer division like the % operator in C++ and is applied to indices of matrix entries. The formula (4.2.5.12) defines the **two-dimensional discrete periodic convolution**, *cf.* Def. 4.1.4.7. Generalizing the notation for the 1D discrete periodic convolution (4.2.5.11) we also write

$$\mathbf{X} *_{m,n} \mathbf{Y} := \mathrm{B}(\mathbf{X},\mathbf{Y}) = \left[\sum_{i=0}^{m-1} \sum_{j=0}^{n-1} (\mathbf{X})_{i,j} (\mathbf{Y})_{(k-i) \mod m,(\ell-j) \mod n} \right]_{\substack{k=0,\ldots,m-1 \\ \ell=0,\ldots,n-1}} , \quad \mathbf{X},\mathbf{Y} \in \mathbb{C}^{m,n} .$$

A direct loop-based implementation of the formula (4.2.5.12) involves an asymptotic computational effort of $O(m^2n^2)$ for $m,n \to \infty$.

C++ code 4.2.5.13: Straightforward implementation of 2D discrete periodic convolution
➪ GITLAB

```
// Straightforward implementation of 2D periodic convolution
template <typename Scalar1, typename Scalar2, class EigenMatrix>
void pmconv_basic(const Eigen::MatrixBase<Scalar1> &X,
                  const Eigen::MatrixBase<Scalar2> &Y, EigenMatrix &Z) {
  using idx_t = typename EigenMatrix::Index;
  using val_t = typename EigenMatrix::Scalar;
  const idx_t n = X.cols();
  const idx_t m = X.rows();
  if ((m != Y.rows()) || (n != Y.cols())) {
    throw std::runtime_error("pmconv: size mismatch");
  }
  Z.resize(m, n); // Ensure right size of output matrix
  // Normalization of indices
  auto idxwrap = [](const idx_t L, int i) {
    return ((i >= L) ? i - L : ((i < 0)?i + L : i));
  };
  // Implementation of (4.2.5.12)
  for (int i = 0; i < m; i++) {
    for (int j = 0; j < n; j++) {
      val_t s = 0;
      for (int k = 0; k < m; k++) {
        for (int l = 0; l < n; l++) {
          s += X(k, l) * Y(idxwrap(m, i - k), idxwrap(n, j - l));
        }
      }
      Z(i, j) = s;
    }
  }
}
```

The key discovery of Section 4.2.1 about the diagonalization of the discrete periodic convolution operation in the Fourier basis carries over to two dimensions, because 2D discrete periodic convolution admits a diagonalization by switching to the trigonometric basis of $\mathbb{C}^{m,n}$, analogous to (4.2.1.17).

In (4.2.5.12) for $r \in \{0,\ldots,m-1\}$, $s \in \{0,\ldots,n-1\}$, set

$$\mathbf{Y} = (\mathbf{F}_m)_{:,r} (\mathbf{F}_n)_{s,:} \in \mathbb{C}^{m,n} \quad \leftrightarrow \quad (\mathbf{Y})_{i,j} = \omega_m^{ri} \omega_n^{sj} , \quad 0 \le i < m,\ 0 \le j < n , \quad \omega_k := \exp(-2\pi\imath/k) .$$

Remember the peridicities (P) $\omega_m^{m+\ell} = \omega_m^\ell$, $\omega_n^{n+\ell} = \omega_n^\ell$, $\ell \in \mathbb{Z}$, which yield

$$\begin{aligned}(\mathsf{B}(\mathbf{X},\mathbf{Y}))_{k,\ell} &= \sum_{i=0}^{m-1}\sum_{j=0}^{n-1}(\mathbf{X})_{i,j}(\mathbf{Y})_{(k-i) \mod m,(\ell-j) \mod n} \overset{\text{(P)}}{=} \sum_{i=0}^{m-1}\sum_{j=0}^{n-1}(\mathbf{X})_{i,j}\omega_m^{r(k-i)}\omega_n^{s(\ell-j)} \\ &= \left(\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}(\mathbf{X})_{i,j}\overline{\omega_m^{ri}\omega_n^{sj}}\right)\cdot\omega_m^{rk}\omega_n^{s\ell} \quad \forall k \in \{0,\ldots,m-1\},\, \ell \in \{0,\ldots,n-1\}\,.\end{aligned}$$

$$\blacktriangleright \quad \mathsf{B}(\mathbf{X}, \underbrace{(\mathbf{F}_m)_{:,r}(\mathbf{F}_n)_{s,:}}_{\text{"eigenvector"}}) = \underbrace{\left(\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}(\mathbf{X})_{i,j}\overline{\omega_m^{ri}\omega_n^{sj}}\right)}_{\text{"eigenvalue", see (4.2.5.4)}}(\mathbf{F}_m)_{:,r}(\mathbf{F}_n)_{s,:}\,. \tag{4.2.5.14}$$

Hence, the (complex conjugated) two-dimensional discrete Fourier transform of $\overline{\mathbf{X}}$ according to (4.2.5.4) provides the eigenvalues of the linear mapping $\mathbf{Y} \mapsto \mathsf{B}(\mathbf{X},\mathbf{Y})$, $\mathbf{X} \in \mathbb{C}^{m,n}$ fixed. Thus we have arrived at a 2D version of the convolution theorem Thm. 4.2.2.2.

Theorem 4.2.5.15. 2D convolution theorem

For any $\mathbf{X},\mathbf{Y} \in \mathbb{C}^{m,n}$, we have

$$\mathbf{X} *_{m,n} \mathbf{Y} = \mathsf{DFT}_{m,n}^{-1}\left(\overline{\mathsf{DFT}_{m,n}(\overline{\mathbf{X}})} \odot \mathsf{DFT}_{m,n}(\mathbf{Y})\right)\,,$$

where $\odot$ stands for the entrywise multiplication of matrices of equal size.

This suggests the following DFT-based algorithm for evaluating the periodic convolution of matrices:

❶ Compute $\hat{\mathbf{Y}}$ by 2D DFT of $\overline{\mathbf{Y}}$, see Code 4.2.5.8

❷ Compute $\hat{\mathbf{X}}$ as complex conjugate of 2D DFT of $\overline{\mathbf{X}}$, see Code 4.2.5.7.

❸ Component-wise multiplication of $\hat{\mathbf{X}}$ and $\hat{\mathbf{Y}}$: $\hat{\mathbf{Z}} = \hat{\mathbf{X}} .* \hat{\mathbf{Y}}$.

❹ Compute $\mathbf{Z}$ through inverse 2D DFT of $\hat{\mathbf{Z}}$.

C++ code 4.2.5.16: DFT-based 2D discrete periodic convolution ↪ GITLAB

```
// DFT based implementation of 2D periodic convolution
template <typename Scalar1, typename Scalar2, class EigenMatrix>
void pmconv(const Eigen::MatrixBase<Scalar1> &X,
            const Eigen::MatrixBase<Scalar2> &Y, EigenMatrix &Z) {
  using Comp = std::complex<double>;
  using idx_t = typename EigenMatrix::Index;
  const idx_t n = X.cols();
  const idx_t m = X.rows();
  if ((m != Y.rows()) || (n != Y.cols())) {
    throw std::runtime_error("pmconv: size mismatch");
  }
  Z.resize(m, n);
  Eigen::MatrixXcd Xh(m, n);
  Eigen::MatrixXcd Yh(m, n);
  // Step ❶: 2D DFT of Y
  fft2(Yh, (Y.template cast<Comp>()));
  // Step ❷: 2D DFT of X
  fft2(Xh, (X.template cast<Comp>()));
  // Steps ❸, ❹: inverse DFT of component-wise product
  ifft2(Z, Xh.cwiseProduct(Yh));
}
```

⌟

EXAMPLE 4.2.5.17 (Deblurring by DFT) 2D discrete convolutions are important for image processing. Let a gray-scale pixel image be stored in the matrix $\mathbf{P} \in \mathbb{R}^{m,n}$, actually $\mathbf{P} \in \{0,\ldots,255\}^{m,n}$, see also Ex. 3.4.4.24.

Write $(p_{l,k})_{l,k\in\mathbb{Z}}$ for the periodically extended image:

$$p_{l,j} = (\mathbf{P})_{l+1,j+1} \quad \text{for} \quad 0 \le l \le m-1, 0 \le j \le n-1\,, \quad p_{l,j} = p_{l+m,j+n} \quad \forall l,k \in \mathbb{Z}\,.$$

Blurring is a technical term for undesirable cross-talk between neighboring pixels: pixel values get replaced by a linear combination of near-by pixel values. This is a good approximate model for the effect of distortion in optical transmission systems like lenses. Blurring can be described by a small matrix called the point-spread function (PSF):

$$c_{l,j} = \sum_{k=-L}^{L}\sum_{q=-L}^{L} s_{k,q}\, p_{l+k,j+q}\,, \quad \begin{array}{l} 0 \le l < m\,, \\ 0 \le j < n\,, \end{array} \quad L \in \{1,\ldots,\min\{m,n\}\}\,. \tag{4.2.5.18}$$

$c_{l,j}$: blurred image; $s_{k,q}$: point spread function

Here the entries of the PSF are referenced as $s_{k,q}$ also with negative indices. We also point out that usually L will be small compared to m and n, we have $s_{k,m} \ge 0$, and $\sum_{k=-L}^{L}\sum_{q=-L}^{L} s_{k,q} = 1$. Hence blurring amounts to *averaging of local pixel values*. You may also want to look at this YouTube Video about "Convolution in Image Processing".

In the experiments reported below we used: $L = 2$ and the PSF $s_{k,q} = \dfrac{1}{1+k^2+q^2}$, $-5 \le k,q \le 5$, normalized to entry sum $= 1$.

C++ code 4.2.5.19: Set-up of point spread function (PSF) $(s_{k,q})$ ➙ GITLAB

```
2 void psf(const Eigen::Index L, MatrixXd& S) {
3   const VectorXd x = VectorXd::LinSpaced(2*L+1, -static_cast<double>(L),
        static_cast<double>(L));
4   const MatrixXd X = x.replicate(1,x.size());
5   const MatrixXd Y = (x.transpose()).replicate(x.size(),1);
6   const MatrixXd E = MatrixXd::Ones(2*L+1, 2*L+1);
7   S = E.cwiseQuotient(E + X.cwiseProduct(X) + Y.cwiseProduct(Y));
8   S /= S.sum();
9 }
```

This is how this PSF acts on an image according to (4.2.5.18):

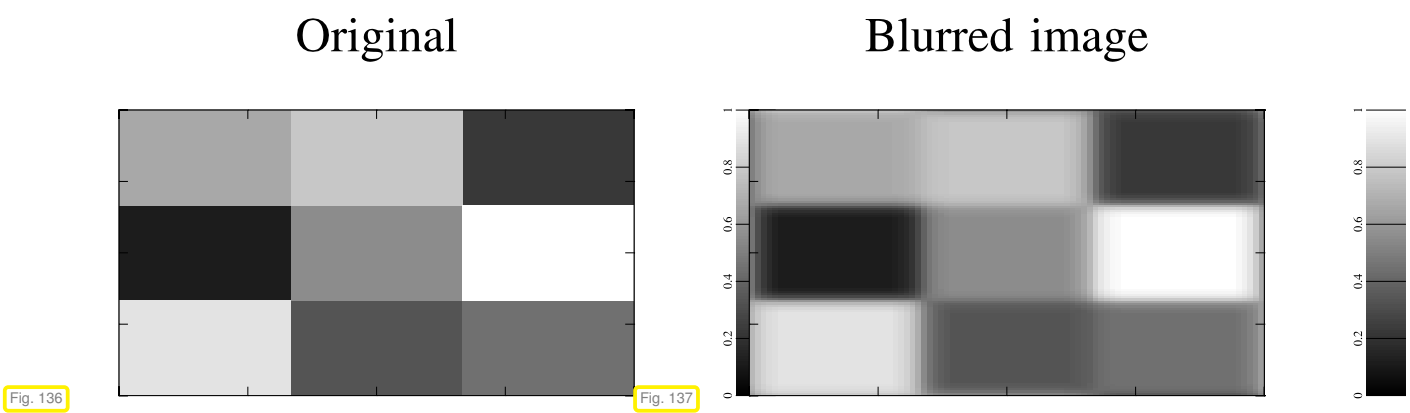


Fig. 136 Fig. 137

Of course, (4.2.5.18) defines a linear operator $\mathcal{B} : \mathbb{R}^{m,n} \mapsto \mathbb{R}^{m,n}$ ("blurring operator").

C++ code 4.2.5.20: Blurring operator ↪ GITLAB

```
2  inline MatrixXd blur(const MatrixXd &P, const MatrixXd &S) {
3    typedef Eigen::Index index_t;
4    auto dimensions = std::make_tuple(P.rows(), P.cols(), S.rows(), S.cols());
5    const auto [ m, n, M, N ] = dimensions;
6    const index_t L = (M - 1) / 2;
7
8    if (M != N) {
9      std::cout << "Error: S not quadratic!\n";
10   }
11
12   MatrixXd C(m, n);
13   for (index_t l = 1; l <= m; ++l) {
14     for (index_t j = 1; j <= n; ++j) {
15       double s = 0;
16       for (index_t k = 1; k <= (2 * L + 1); ++k) {
17         for (index_t q = 1; q <= (2 * L + 1); ++q) {
18           index_t kl = l + k - L - 1;
19           if (kl < 1) {
20             kl += m;
21           } else if (kl > m) {
22             kl -= m;
23           }
24           index_t jm = j + q - L - 1;
25           if (jm < 1) {
26             jm += n;
27           } else if (jm > n) {
28             jm -= n;
29           }
30           s += P(kl - 1, jm - 1) * S(k - 1, q - 1);
31         }
32       }
33       C(l - 1, j - 1) = s;
34     }
35   }
36   return C;
37 }
```

Yet, does (4.2.5.18) ring a bell? Hidden in (4.2.5.18) is a 2D discrete periodic convolution, see

(4.2.5.12)!

$$
\begin{aligned}
c_{l,j} &= \sum_{k=-L}^{L}\sum_{q=-L}^{L} s_{k,q}(\mathbf{P})_{(l+k) \mod m,(j+q) \mod n} \\
&= \sum_{k=-L}^{L}\sum_{q=-L}^{L} s_{-k,-q}(\mathbf{P})_{(l-k) \mod m,(j-q) \mod n} \\
&= \sum_{k=0}^{L}\sum_{q=0}^{L} s_{-k,-q}(\mathbf{P})_{(l-k) \mod m,(j-q) \mod n} + \\
&\quad \sum_{k=0}^{L}\sum_{q=-L}^{-1} s_{-k,-q}(\mathbf{P})_{(l-k) \mod m,(j-(q+n)) \mod n} + \\
&\quad \sum_{k=-L}^{-1}\sum_{q=0}^{L} s_{-k,-q}(\mathbf{P})_{(l-(k+m)) \mod m,(j-q) \mod n} + \\
&\quad \sum_{k=-L}^{-1}\sum_{q=-L}^{-1} s_{-k,-q}(\mathbf{P})_{(l-(k+m)) \mod m,(j-(q+n)) \mod n} \\
&= \sum_{k=0}^{L}\sum_{q=0}^{L} s_{-k,-q}(\mathbf{P})_{(l-k) \mod m,(j-q) \mod n} + \\
&\quad \sum_{k=0}^{L}\sum_{q=n-L}^{n-1} s_{-k,-q+n}(\mathbf{P})_{(l-k) \mod m,(j-q) \mod n} + \\
&\quad \sum_{k=m-L}^{m-1}\sum_{q=0}^{L} s_{-k+m,-q}(\mathbf{P})_{(l-k) \mod m,(j-q) \mod n} + \\
&\quad \sum_{k=m-L}^{m-1}\sum_{q=n-L}^{n-1} s_{-k+m,-q+n}(\mathbf{P})_{(l-k) \mod m,(j-q) \mod n} \,.
\end{aligned}
$$

Hence we have that the blured image is given by the matrix

$$
\mathbf{C} = \mathbf{S} *_{m,n} \mathbf{P} \quad \text{with} \quad (\mathbf{S})_{k,q} = \begin{cases} s_{-k,-q} & \text{, for } 0 \le k,q \le L \,, \\ s_{-k+m,-q} & \text{, for } m-L \le k < m \,, \quad 0 \le q \le L \,, \\ s_{-k,q+n} & \text{, for } 0 \le k \le L \,, \quad n-L \le q < n \,, \\ s_{-k+m,-q+m} & \text{, for } m-L \le k < m \,, \quad n-L \le q < n \,, \\ 0 & \text{else.} \end{cases} \tag{4.2.5.21}
$$

The weights of the PSF are placed in the corners of the matrix $\mathbf{S}$. For instance, in the case of $L = 2$ we arrive at the $m \times n$-matrix

$$
\mathbf{S} = \begin{bmatrix} s_{0,0} & s_{0,-1} & s_{0,-2} & 0 & \dots & 0 & s_{0,1} & s_{0,2} \\ s_{-1,0} & s_{-1,-1} & s_{-1,-2} & 0 & \dots & 0 & s_{-1,1} & s_{-1,2} \\ s_{-2,0} & s_{-2,-1} & s_{-2,-2} & 0 & \dots & 0 & s_{-2,1} & s_{-2,2} \\ 0 & 0 & 0 & 0 & & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ s_{1,0} & s_{1,-1} & s_{1,-2} & 0 & \dots & 0 & s_{1,1} & s_{1,2} \\ s_{2,0} & s_{2,-1} & s_{2,-2} & 0 & \dots & 0 & s_{2,1} & s_{2,2} \end{bmatrix} .
$$

In light of Thm. 4.2.5.15 and of the algorithm implemented in Code 4.2.5.16, now it hardly comes as a surprise that DFT comes handy for reversing the effect of the blurring!

We supplement an elementary derivation and revisit the considerations of § 4.2.5.1 and recall the derivation of (4.2.1.10) and Lemma 4.2.1.16. Let us blur an "image" $\left[\omega_m^{\nu k}\omega_n^{\mu q}\right]_{\substack{0\le k<n\\0\le q<m}} \in \mathbb{C}^{m,n}$, $0 \le \nu < m$, $0 \le \mu < n$, showing a plane wave. We send it through the blurring operator $\mathcal{B}$:

$$\left(\mathcal{B}(\left(\omega_m^{\nu k}\omega_n^{\mu q}\right)_{k,q\in\mathbb{Z}}\right)_{l,j} = \sum_{k=-L}^{L}\sum_{q=-L}^{L} s_{k,q}\omega_m^{\nu(l+k)}\omega_n^{\mu(j+q)} = \omega_m^{\nu l}\omega_n^{\mu j}\sum_{k=-L}^{L}\sum_{q=-L}^{L} s_{k,q}\omega_m^{\nu k}\omega_n^{\mu q}\,.$$

The two-dimensional periodic sequences $\mathbf{V}_{\nu,\mu} := \left(\omega_m^{\nu k}\omega_n^{\mu q}\right)_{k,q\in\mathbb{Z}}$, $0 \le \mu < m$, $0 \le \nu < n$ are the eigenvectors of the blurring operator $\mathcal{B}$:

$$\mathcal{B}\mathbf{V}_{\nu,\mu} = \lambda_{\nu,\mu}\mathbf{V}_{\nu,\mu}\,, \quad \text{with eigenvalue} \quad \lambda_{\nu,\mu} = \underbrace{\sum_{k=-L}^{L}\sum_{q=-L}^{L} s_{k,q}\omega_m^{\nu k}\omega_n^{\mu q}}_{\text{2-dimensional DFT of point spread function !}} \tag{4.2.5.22}$$

Thus the inversion of the blurring operator boils down to componentwise scaling in "Fourier domain", see See also Code 4.2.5.16 for the same idea.

C++ code 4.2.5.23: DFT based deblurring ↪ GITLAB

```
2     Eigen::MatrixXd
3     deblur(const Eigen::MatrixXd &C, const Eigen::MatrixXd &S,
4            const double tol = 1e-3) {
5   typedef Eigen::Index index_t;
6   auto dimensions = std::make_tuple(C.rows(), C.cols(), S.rows(), S.cols());
7   const auto [m, n, M, N] = dimensions;
8   const index_t L = (M - 1) / 2;
9   if (M != N) {
10    throw std::runtime_error("Error: S not quadratic!");
11  }
12  Eigen::MatrixXd Spad = Eigen::MatrixXd::Zero(m, n);
13  // Zero padding, see (4.2.5.21)
14  Spad.block(0, 0, L + 1, L + 1) = S.block(L, L, L + 1, L + 1);
15  Spad.block(m - L, n - L, L, L) = S.block(0, 0, L, L);
16  Spad.block(0, n - L, L + 1, L) = S.block(L, 0, L + 1, L);
17  Spad.block(m - L, 0, L, L + 1) = S.block(0, L, L, L + 1);
18  // Inverse of blurring operator
19  Eigen::FFT fft;
20  const Eigen::MatrixXcd SF = fft.fwd(Spad);
21  // Test for invertibility
22  if (SF.cwiseAbs().minCoeff() < tol * SF.cwiseAbs().maxCoeff()) {
23    std::cerr << "Error: Deblurring impossible!\n";
24  }
25  // DFT based deblurring
26  return fft.inv(fft.fwd(C).cwiseQuotient(SF));
27 }
```

Note that this code checks whether deblurring is possible, that is, whether the blurring operator is really invertible. A near singular blurring operator manifests itself through entries of its DFT close to zero. ⌟

Review question(s) 4.2.5.24 (Two-dimensional DFT)

(Q4.2.5.24.A) Describe a function $(t,s) \mapsto g_n(t,s)$, such that for any $n \in \mathbb{N}$

$$\left[\mathrm{Re}\left\{(\mathbf{F}_n)_{:,j}^{\top}(\mathbf{F}_n)_{:,\ell}\right\}\right]_{0\le j,\ell<n} = \left[g_n(t_j,s_\ell)\right]_{\substack{j=0,\ldots,n-1\\\ell=0,\ldots,n-1}} \quad t_j := \frac{j}{n}, s_\ell := \frac{\ell}{n}\,, \quad n \in \mathbb{N}\,.$$

Here $\mathbf{F}_n$, $n \in \mathbb{N}$, is the Fourier matrix

$$\mathbf{F}_n = \begin{bmatrix} \omega_n^0 & \omega_n^0 & \cdots & \omega_n^0 \\ \omega_n^0 & \omega_n^1 & \cdots & \omega_n^{n-1} \\ \omega_n^0 & \omega_n^2 & \cdots & \omega_n^{2n-2} \\ \vdots & \vdots & & \vdots \\ \omega_n^0 & \omega_n^{n-1} & \cdots & \omega_n^{(n-1)^2} \end{bmatrix} = \left[\omega_n^{lj}\right]_{l,j=0}^{n-1} \in \mathbb{C}^{n,n} . \tag{4.2.1.13}$$

(Q4.2.5.24.B) How would you compute the discrete Fourier transform of a tensor product matrix $\mathbf{X} = \mathbf{u}\mathbf{v}^{\mathrm{H}}$, $\mathbf{u}, \mathbf{v} \in \mathbb{C}^n$.

(Q4.2.5.24.C) [2D DFT of circulant matrix] Let $\mathbf{C} \in \mathbb{C}^{n,n}$ be the circulant matrix generated by the vector $\mathbf{p} = [p_0, \ldots, p_{n-1}] \in \mathbb{C}^n$,

$$\mathbf{C} := \begin{bmatrix} p_0 & p_{n-1} & p_{n-2} & \cdots & & \cdots & p_1 \\ p_1 & p_0 & & & & & p_2 \\ p_2 & & & & & & \vdots \\ \vdots & & & & & & \\ & & & & & & \vdots \\ p_{n-2} & & & & & & p_{n-1} \\ p_{n-1} & p_{n-2} & \cdots & & \cdots & p_1 & p_0 \end{bmatrix} .$$

Sketch an algorithm for computing the two-dimensional DFT of $\mathbf{C}$, $\mathrm{DFT}_{n,n}(\mathbf{X})$, efficiently.

Hint. Use the definition of the 2D DFT,

$$\mathrm{DFT}_{m,n}(\mathbf{Y}) = \mathbf{F}_m \mathbf{Y} \mathbf{F}_n \,, \quad \mathbf{F}_k \in \mathbb{C}^{k,k} \quad \text{the Fourier matix,}$$

and

Lemma 4.2.1.16. Diagonalization of circulant matrices

For any circulant matrix $\mathbf{C} \in \mathbb{K}^{n,n}$, $c_{ij} = u_{i-j}$, $(u_k)_{k\in\mathbb{Z}}$ an n-periodic sequence, holds true

$$\mathbf{C}\overline{\mathbf{F}}_n = \overline{\mathbf{F}}_n \operatorname{diag}(d_0, \ldots, d_{n-1}) \quad , \quad [d_0, \ldots, d_{n-1}]^\top = \mathbf{F}_n [u_0, \ldots, u_{n-1}]^\top ,$$

where $\mathbf{F}_n \in \mathbb{C}^{n,n}$ is the Fourier matrix: $\mathbf{F}_n = \left[\exp(-\frac{2\pi\imath}{n}\ell k)\right]_{\ell,k=0}^{n-1}$ from (4.2.1.13).

△

4.2.6 Semi-discrete Fourier Transform [QSS00, Sect. 10.11]

Starting from § 4.1.4.2 we mainly looked at time-discrete n-*periodic* signals, which can be mapped to vectors $\in \mathbb{R}^n$. This led to discrete periodic convolution (→ Def. 4.1.4.7) and the discrete Fourier transform (DFT) (→ Def. 4.2.1.18) as (bi-)linear mappings in $\mathbb{C}^n$.

§4.2.6.1 ("Squeezing" the DFT) In this section we are concerned with non-periodic signals of infinite duration as introduced in § 4.0.0.1.

Idea: Study the limit $n \to \infty$ for DFT in the n-periodic setting

Let $(y_k)_{k\in\mathbb{Z}}$ be an n-periodic sequence (signal), $n = 2m+1$, $m \in \mathbb{N}$, with generating vector $\mathbf{y} := [y_0, \ldots, y_{n-1}]^\top$. Thanks to periodicity we can rewrite the periodic DFT $\mathbf{c} = \mathrm{DFT}_n\mathbf{y}$ with a simple change of indexing:

$$\text{DFT} \to \text{Def. 4.2.1.18:} \qquad (\mathbf{c})_k = c_k := \sum_{j=-m}^{m} y_j \exp(-2\pi i \frac{kj}{n}) , \quad k = 0, \ldots, n-1 . \tag{4.2.6.2}$$

Next, we associate a point $t_k \in [0,1[$ with each index k of the DFT $(c_k)_{k=0}^{n-1}$:

$$k \in \{0, \ldots, n-1\} \quad \longleftrightarrow \quad t_k := \frac{k}{n} . \tag{4.2.6.3}$$

Thus we can view $(c_k)_{k=0}^{n-1}$ as the heights of n pulses evenly spaced in the interval $[0,1[$.

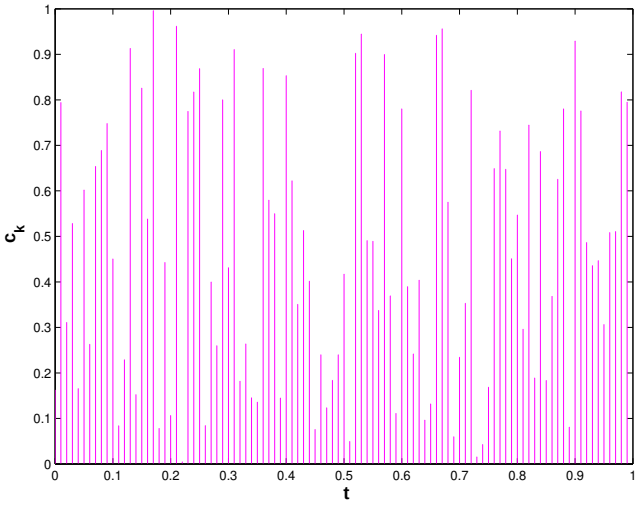


Fig. 138

◁ "Squeezing" a vector $\in \mathbb{R}^n$ into $[0,1[$.

We can interpret the values c_k as sampled values of a function defined on $[0,1]$

$$c_k \quad \leftrightarrow \quad c(t_k);,$$

$$t_k = \frac{k}{n} , \quad k = 0, \ldots, n-1 .$$

This makes it possible to pass from a discrete finite signal to a continuous signal.

In a sense, formally, we can rewrite (4.2.6.2) as

$$\text{DFT:} \qquad c(t_k) := c_k = \sum_{j=-m}^{m} y_j \exp(-2\pi \imath j t_k) , \quad k = 0, \ldots, n-1 . \tag{4.2.6.4}$$

The notation indicates that we read c_k as the value of a *function* $c : [0,1[\mapsto \mathbb{C}$ for argument t_k. ⌟

EXAMPLE 4.2.6.5 ("Squeezed" DFT of a periodically truncated signal) We consider the bi-infinite discrete signal $(y_j)_{j\in\mathbb{Z}}$, "concentrated around 0"

$$y_j = \frac{1}{1+j^2} , \quad j \in \mathbb{Z} .$$

We examine the DFT of the $2m+1$-periodic signal obtained by periodic extension of $(y_k)_{k=-m}^{m}$, C++ code ➙ GITLAB.

Period of signal $y_i = 33$

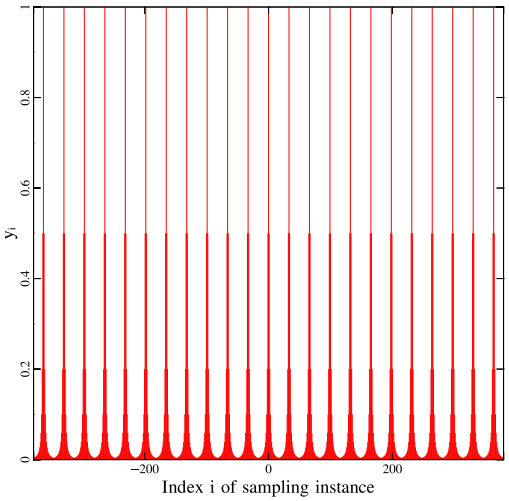


Fig. 139

DFT of period 33 signal

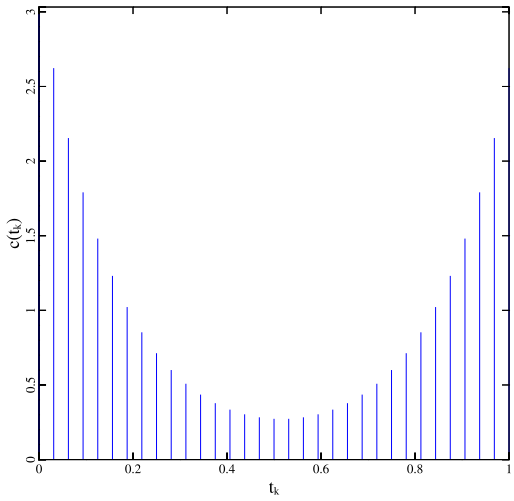


Fig. 140

Period of signal $y_i = 65$

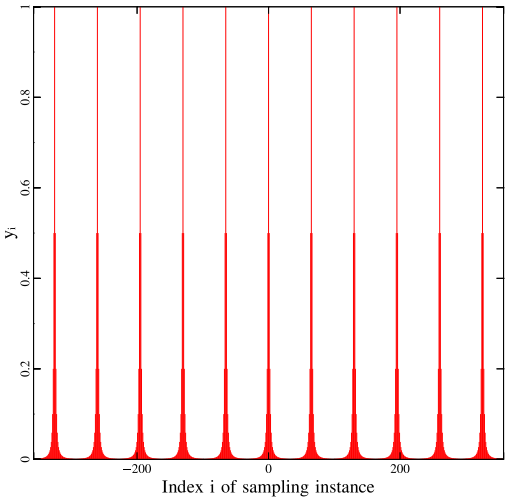


Fig. 141

DFT of period 65 signal

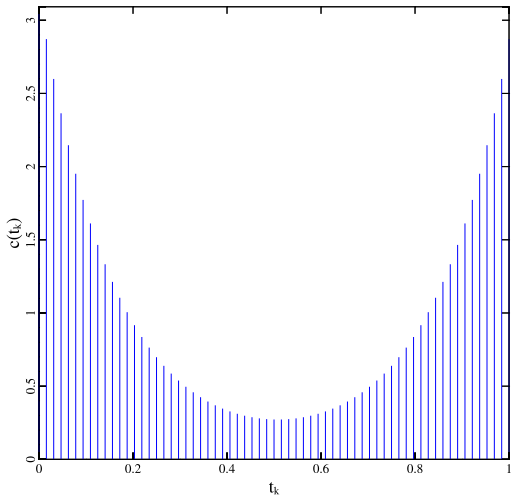


Fig. 142

Period of signal $y_i = 129$

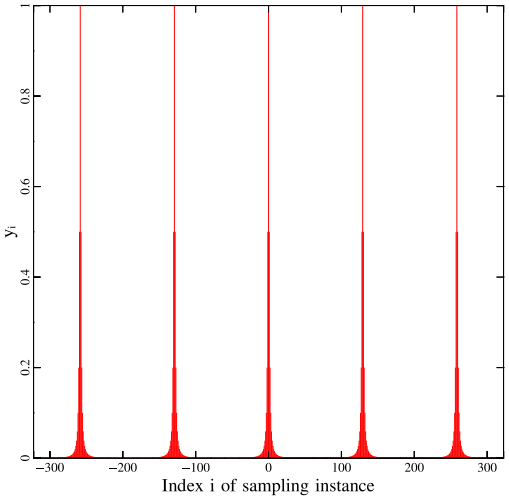


Fig. 143

DFT of period 129 signal

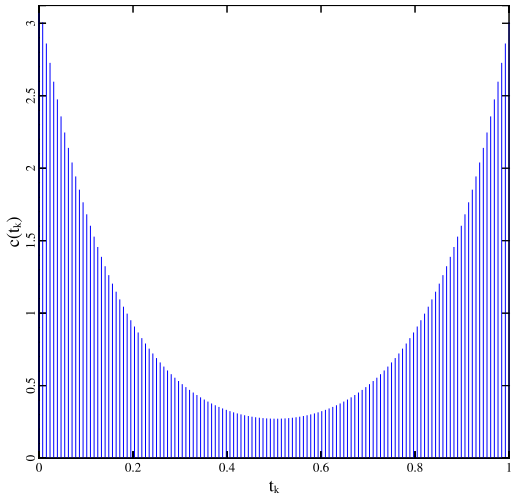


Fig. 144

Period of signal $y_i = 257$

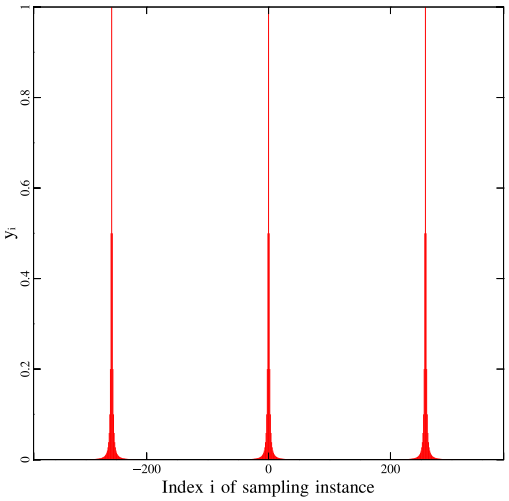


Fig. 145

DFT of period 257 signal

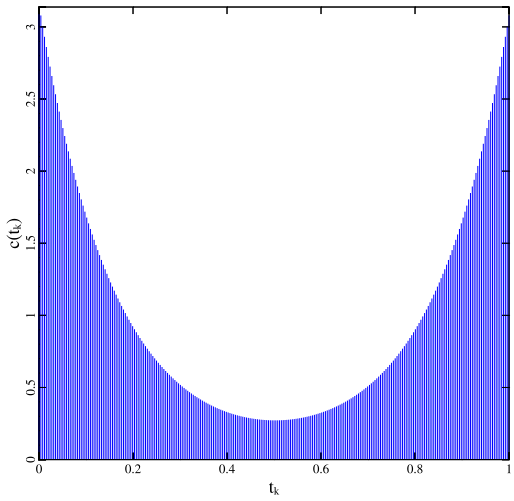


Fig. 146

The visual impression is that the values $c(t_k)$ "converge" to those of a function $c : [0,1[\mapsto \mathbb{R}$ in the sampling points t_k.

⌟

§4.2.6.6 (Fourier series) Now we pass to the limit $m \to \infty$ in (4.2.6.4) and keep the "sampling a function" perspective: $c_k = c(t_k)$. Note that passing to the limit amounts to dropping the assumption of periodicity!

$$\blacktriangleright \quad \boxed{c(t) = \sum_{k \in \mathbb{Z}} y_k \exp(-2\pi \imath k t)} \,. \tag{4.2.6.7}$$

Terminology: The series (= infinite sum) on the right hand side of (4.2.6.7) is called a **Fourier series** (link).

The function $c : [0,1[\mapsto \mathbb{C}$ defined by (4.2.6.7) is called the **Fourier transform** of the sequence $(y_k)_{k \in \mathbb{Z}}$ (, if the series converges).

Corollary 4.2.6.8. Periodicity of Fourier transforms

Fourier transforms $t \mapsto c(t)$ are 1-periodic functions $\mathbb{R} \to \mathbb{C}$.

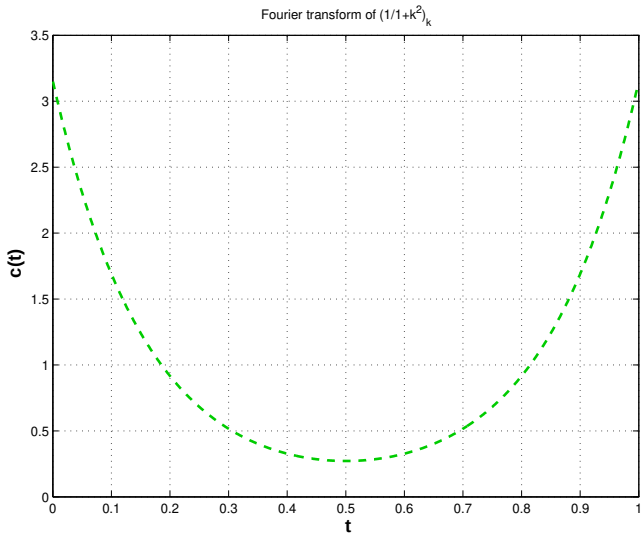


Fig. 147

Thus, the limit we "saw" in Ex. 4.2.6.5 is actually the Fourier transform of the sequence $(y_k)_{k\in\mathbb{Z}}$!

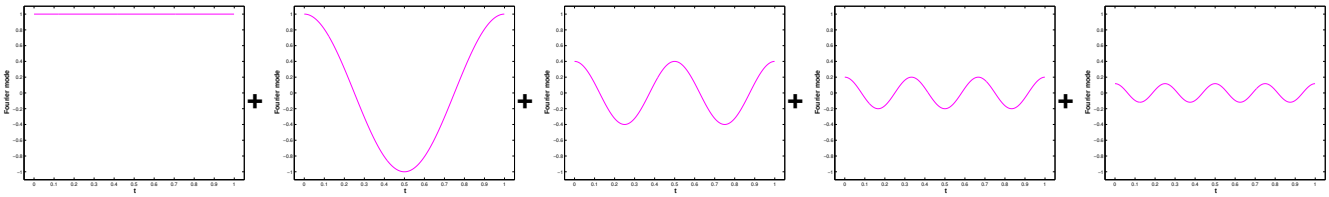
◁ Fourier transform of $y_k := \frac{1}{1+k^2}$

From (4.2.6.7) we conclude:

Fourier transform

=

weighted sum of Fourier modes $t \mapsto \exp(-2\pi\imath kt)$, $k \in \mathbb{Z}$



→ related animation on Wikipedia.

It is possible to derive a closed-form expression for the function displayed in Fig. 148:

$$c(t) = \sum_{k\in\mathbb{Z}} \frac{1}{1+k^2}\exp(-2\pi\imath kt) = \frac{\pi}{e^{\pi}-e^{-\pi}}\cdot\left(e^{\pi-2\pi t}+e^{2\pi t-\pi}\right) \in C^{\infty}([0,1])\,.$$

Note that when considered as a 1-periodic function on $\mathbb{R}$, this $c(t)$ is merely continuous. ⌟

Remark 4.2.6.9 (Decay conditions for bi-infinite signals) The considerations above were based on

- truncation of $(y_k)_{k\in\mathbb{Z}}$ to $(y_k)_{k=-m}^{m}$ and
- periodic continuation to an $2m+1$-periodic signal.

Obviously, only if the signal is *concentrated around* $k = 0$ this procedure will not lose essential information contained in the signal, which suggests decay conditions for the coefficients of Fourier series.

$$\text{Minimal requirement:} \quad \lim_{k\to\infty} |y_k| = 0\,, \tag{4.2.6.10}$$

$$\text{Stronger requirement:} \quad \sum_{k\in\mathbb{Z}} |y_k| < \infty\,. \tag{4.2.6.11}$$

The summability condition (4.2.6.11) implies (4.2.6.10). Moreover, (4.2.6.11) ensures that the Fourier series (4.2.6.7) converges uniformly [Str09, Def. 4.8.1] because the exponentials are all bounded by 1 in modulus. From [Str09, Thm. 4.8.1] we learn that limits of uniformly convergent series of continuous functions posses a continuous limit. As a consequence $c : [0,1[\mapsto \mathbb{C}$ is continuous, if (4.2.6.11) holds. ⌟

EXAMPLE 4.2.6.12 (Convergence of Fourier sums) We consider the following infinite signal, satisfying the summation condition (4.2.6.11): $y_k = \frac{1}{1+k^2}$, $k \in \mathbb{Z}$, see Ex. 4.2.6.5. We monitored: approximation of the Fourier transform $c(t)$ by Fourier sums $c_m(t)$, see (4.2.6.14).

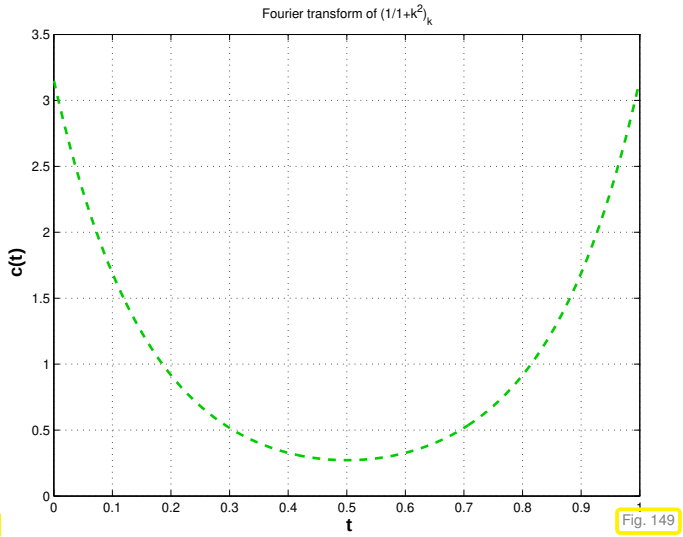


Fig. 148

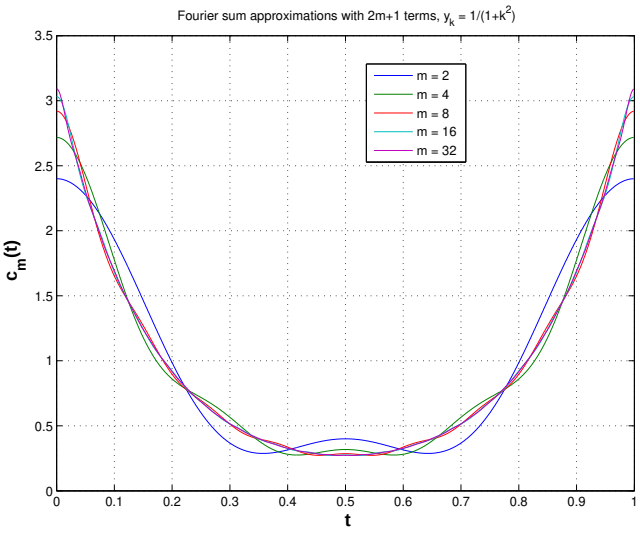


Fig. 149

We observe convergence of Fourier sums in "eyeball norm". Quantitative estimates can be deduced from decay properties of the sequence $(y_k)_{k\in\mathbb{Z}}$. If it is summable according to (4.2.6.11), then

$$\left|\sum_{k\in\mathbb{Z}} y_k \exp(-2\pi\imath kt) - \sum_{k=-M}^{M} y_k \exp(-2\pi\imath kt)\right| \leq \sum_{|k|>M} |y_k| \to 0 \quad \text{for} \quad M\to\infty$$

Further quantitative statements about convergence can be deduced from Thm. 4.2.6.33 below. ⌟

Remark 4.2.6.13 (Numerical summation of Fourier series)

Assuming sufficiently fast decay of the signal $(y_k)_{k\in\mathbb{Z}}$ for $k\to\infty$ ($\to$ Rem. 4.2.6.9), we can *approximate* the Fourier series (4.2.6.7) by a Fourier sum

$$c(t) \approx c_M(t) := \sum_{k=-M}^{M} y_k \exp(-2\pi ikt)\,, \quad M \gg 1\,. \tag{4.2.6.14}$$

Task: Approximate evaluation of $c(t)$ at N equidistant points $t_j := \frac{j}{N}$, $j = 0,\ldots,N$ (e.g., for plotting it).

$$c(t_j) = \lim_{M\to\infty} \sum_{k=-M}^{M} y_k \exp(-2\pi ikt_j) \approx \sum_{k=-M}^{M} y_k \exp(-2\pi i\frac{kj}{N})\,, \tag{4.2.6.15}$$

for $j = 0,\ldots,N-1$.

Note that in the case $N = M$ (4.2.6.15) coincides with a *discrete Fourier transform* (DFT, $\to$ Def. 4.2.1.18). The following code demonstrates the evaluation of a Fourier series at equidistant points using DFT.

C++ code 4.2.6.16: DFT-based evaluation of Fourier sum at equidistant points ↪ GITLAB

```
// evaluate scalar function with a vector
// DFT based approximate evaluation of Fourier series
// signal is a functor providing the y_k
// M specifies truncation of series according to (4.2.6.14)
// N is the number of equidistant evaluation points for c in
// [0,1[.
template <class Function>
VectorXcd foursum(const Function &signal, int M, int N) {
  using index_t = Eigen::Index;
  const int m = 2 * M + 1; // length of the signal
  // sample signal
```

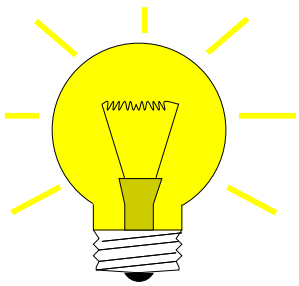
```
13   // VectorXd y = feval(signal, VectorXd::LinSpaced(m, -M, M));
14   VectorXd y = VectorXd::LinSpaced(m, -M, M).unaryExpr(signal);
15   // Ensure that there are more sampling points than terms in series
16   const int l = static_cast<int>(m > N ? ceil(static_cast<double>(m) / N) : 1);
17   N *= l;
18   // Zero padding and wrapping of signal, see
19   // Code 4.2.3.7
20   VectorXd y_ext = VectorXd::Zero(N);
21   y_ext.head(M + 1) = y.tail(M + 1);
22   y_ext.tail(M) = y.head(M);
23   // Perform DFT and decimate output vector
24   Eigen::FFT fft;
25   const Eigen::VectorXcd k = fft.fwd(y_ext);
26   Eigen::VectorXcd c(N / l);
27   for (int i = 0; i < N / l; ++i) {
28     c(i) = k(static_cast<index_t>(i) * static_cast<index_t>(l));
29   }
30   return c;
31 }
```

§4.2.6.17 (Inverting the Fourier transform) Now we perform a similar passage to the limit as above for the inverse DFT (4.2.1.20), $n = 2m+1$,

$$y_j = \frac{1}{n}\sum_{k=0}^{n-1} c_k \exp(2\pi i \frac{jk}{n}) , \quad j = -m, \dots, m . \tag{4.2.6.18}$$

We adopt a function perspective as before: $c_k \leftrightarrow c(t_k)$, $t_k = \frac{k}{n}$, *cf.* (4.2.6.3), and rewrite

$$y_j = \frac{1}{n}\sum_{k=0}^{n-1} c(t_k) \exp(2\pi i j t_k) , \quad j = -m, \dots, m . \tag{4.2.6.19}$$



Then pass to the limit $m \to \infty$ in (4.2.6.19)

Insight: The right hand side of (4.2.6.19) is a Riemann sum, *cf.* [Str09, Sect. 6.2]

▶ In the limit $m \to \infty$ the sum becomes an integral!

$$\boxed{y_j = \int_0^1 c(t) \exp(2\pi i j t)\, \mathrm{d}t , \quad j \in \mathbb{Z}} . \tag{4.2.6.20}$$

This formula is the inversion of the summation of a Fourier series (4.2.6.7)!

In fact, this is not surprising, because for a Fourier series

$$c(t) = \sum_{k\in\mathbb{Z}} y_j \exp(-2\pi \imath k t) , \quad t \in \mathbb{R} ,$$

satisfying the summabiliy condition (4.2.6.11) we can swap integration and summation and directly compute

$$\int_0^1 c(t)\exp(2\pi ijt)\,\mathrm{d}t = \int_0^1 \sum_{k\in\mathbb{Z}} y_j \exp(-2\pi\imath kt))\exp(2\pi ijt)\,\mathrm{d}t$$
$$= \sum_{k\in\mathbb{Z}} y_j \int_0^1 \exp(2\pi\imath(j-k)t)\,\mathrm{d}t = y_j\,,$$

because of the "orthogonality relation"

$$\int_0^1 \exp(2\pi\imath nt)\,\mathrm{d}t = \begin{cases} 1 & \text{, if } n = 0\,, \\ 0 & \text{, if } n \neq 0\,. \end{cases}$$

The formula (4.2.6.20) allows to recover the signal $(y_k)_{k\in\mathbb{Z}}$ from its Fourier transform $c(t)$.

⌟

§4.2.6.21 (Fourier transform as linear mapping) Assuming sufficiently fast decay of the infinite sequence $(y_k)_{k\in\mathbb{Z}} \in \mathbb{C}^{\mathbb{Z}}$, combining (4.2.6.7) and (4.2.6.20) we have found the relationship

$$\text{(4.2.6.7):}\quad c(t) = \sum_{k\in\mathbb{Z}} y_k \exp(-2\pi\imath kt) \quad\leftrightarrow\quad \text{(4.2.6.20):}\quad y_k = \int_0^1 c(t)\exp(2\pi\imath kt)\,\mathrm{d}t\,.$$

Terminology: y_j from (4.2.6.20) is called the j-th Fourier coefficient of the function c.

✎ Notation: $\widehat{c}_j := y_j$ with y_j defined by (4.2.6.20) $\hat{=}$ j-th Fourier coefficient of $c : [0,1[\to \mathbb{C}$

In a sense, Fourier series summation maps to sequence to a 1-periodic function, Fourier coefficient extraction a 1-periodic function to a sequence

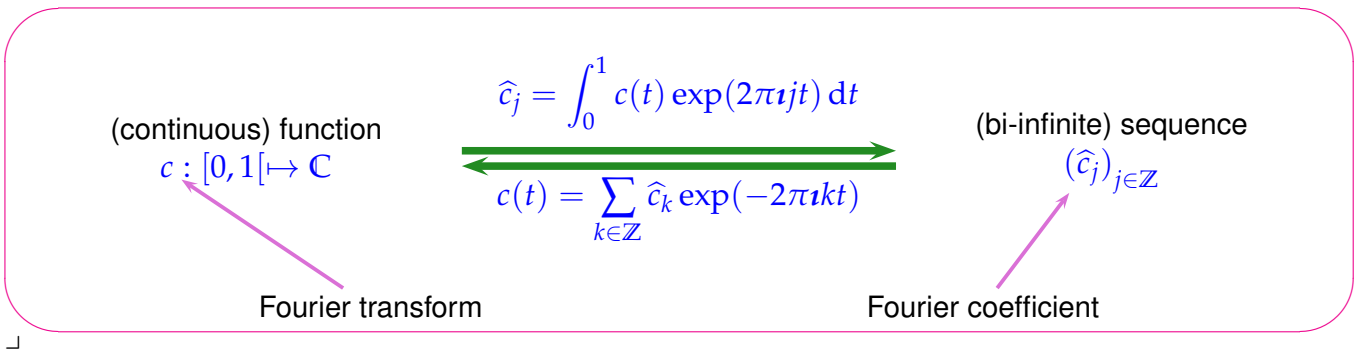
$$\text{sequence}\quad \in \mathbb{C}^{\mathbb{Z}} \xrightarrow[\text{Fourier coefficients}]{\text{Fourier series}} \text{function}\quad [0,1[\to \mathbb{C}\,.$$

Both the space $\mathbb{C}^{\mathbb{Z}}$ of bi-infinite sequences and the space of functions $[0,1[\to \mathbb{C}$ are vector spaces equipped with "termwise/pointwise" addition and scalar multiplication. Then it is clear that

- the series summation mapping $(\widehat{c}_k)_{k\in\mathbb{Z}} \mapsto c$ from (4.2.6.7),
- and the Fourier coeffiicient extraction mapping $c \mapsto (\widehat{c}_k)_{k\in\mathbb{Z}}$ from (4.2.6.20)

are *linear*! (Recall the concept of a linear mapping as explained in [NS02, Ch. 6].)

Let us summarize the fundamental correspondences:



⌟

Remark 4.2.6.22 (Filtering in Fourier domain) What happens to the Fourier transform of a bi-infinite signal, if it passes through a channel?

Consider a (bi-)infinite signal $(x_k)_{k\in\mathbb{Z}}$ sent through a finite (→ Def. 4.1.1.2, linear (→ Def. 4.1.1.7) time-invariant (→ Def. 4.1.1.5) causal (→ Def. 4.1.1.9) channel with impulse response $(\ldots,0,h_0,\ldots,h_{n-1}0,\ldots)$ (→ Def. 4.1.1.12). By (4.1.2.4) this results in the output signal

$$y_k = \sum_{j=0}^{n-1} h_j x_{k-j}\,, \quad k\in\mathbb{Z}\,. \tag{4.2.6.23}$$

We introduce the Fourier transforms of the infinite signals:

$$(y_k)_{k\in\mathbb{Z}} \quad \leftrightarrow \quad t\mapsto c(t) \quad , \quad (x_j)_{j\in\mathbb{Z}} \quad \leftrightarrow t\mapsto b(t)\,.$$

We also assume that $(x_k)_{k\in\mathbb{Z}}$ satisfies the summability condition (4.2.6.11). Then elementary computations establish a relationship between the Fourier transforms:

$$\begin{aligned} c(t) = \sum_{k\in\mathbb{Z}} y_k \exp(-2\pi\imath kt) &= \sum_{k\in\mathbb{Z}}\sum_{j=0}^{n-1} h_j x_{k-j}\exp(-2\pi\imath kt) \\ \text{[shift summation index } k\text{]} \quad &= \sum_{j=0}^{n-1}\sum_{k\in\mathbb{Z}} h_j x_k \exp(-2\pi\imath jt)\exp(-2\pi\imath kt) \\ &= \underbrace{\left(\sum_{j=0}^{n-1} h_j \exp(-2\pi\imath jt)\right)}_{\text{trigonometric polynomial of degree } n-1} b(t)\,. \end{aligned} \tag{4.2.6.24}$$

Definition 4.2.6.25. Trigonometric polynomial

A trigonometric polynomial is a function $\mathbb{R}\to\mathbb{C}$ that is a weighted sum of finitely many terms $t\to\exp(-2\pi\imath kt)$, $k\in\mathbb{Z}$.

We summarize the insight gleaned from (4.2.6.24):

Discrete convolution in Fourier domain

The discrete convolution of a signal with finite impulse response amounts to a multiplication of its Fourier transform with a trigonometric polynomial whose coefficients are given by the impulse response.

⌟

§4.2.6.27 (Fourier transform and convolution) In fact, the observation made in Rem. 4.2.6.22 is a special case of a more general result that provides a version of the convolution theorem Thm. 4.2.2.2 for the Fourier transform.

Theorem 4.2.6.28. Convolution theorem

Let $t\mapsto c(t)$ and $b\mapsto b(t)$ be the Fourier transforms of the two summable bi-infinite sequences $(y_k)_{k\in\mathbb{Z}}$ and $(x_k)_{k\in\mathbb{Z}}$, respectively. Then the pointwise product $t\mapsto c(t)b(t)$ is the Fourier transform of the convolution (→ Def. 4.1.2.7)

$$(x_k)*(y_k) := \left\{\ell\in\mathbb{Z}\mapsto\sum_{k\in\mathbb{Z}} x_k y_{\ell-k}\right\}\in\mathbb{C}^{\mathbb{Z}}\,.$$

Proof. (formal) Ignoring issues of convergence, we may just multiply the two Fourier sequences and sort the resulting terms:

$$\left(\sum_{k\in\mathbb{Z}} y_k \exp(-2\pi\imath kt)\right)\cdot\left(\sum_{j\in\mathbb{Z}} x_j \exp(-2\pi\imath jt)\right) = \sum_{\ell\in\mathbb{Z}} \underbrace{\left(\sum_{k\in\mathbb{Z}} y_k x_{\ell-k}\right)}_{=((y_k)*(x_k))_\ell} \exp(-2\pi\imath\ell t)\,.$$

□ ⌟

§4.2.6.29 (Isometry property of Fourier transform) We will find a *conservation of power* through Fourier transform. This is related to the assertion of Lemma 4.2.1.14 for the Fourier matrix $\mathbf{F}_n$, see (4.2.1.13), namely that $\frac{1}{\sqrt{n}}\mathbf{F}_n$ is *unitary* (→ Def. 6.3.1.2), which implies

$$\frac{1}{\sqrt{n}}\mathbf{F}_n \quad \text{unitary} \quad \overset{\text{Thm. 3.3.2.2}}{\blacktriangleright} \quad \left\|\frac{1}{\sqrt{n}}\mathbf{F}_n\mathbf{y}\right\|_2 = \|\mathbf{y}\|_2\,. \tag{4.2.6.30}$$

Since DFT boils down to multiplication with $\mathbf{F}_n$ (→ Def. 4.2.1.18), we conclude from (4.2.6.30)

$$c_k \text{ from (4.2.6.2)} \quad \Rightarrow \quad \frac{1}{n}\sum_{k=0}^{n-1}|c_k|^2 = \sum_{j=-m}^{m}|y_j|^2\,. \tag{4.2.6.31}$$

Now we adopt the function perspective again and associated $c_k \leftrightarrow c(t_k)$. Then we pass to the limit $m\to\infty$, appeal to Riemann summation (see above), and conclude

$$(4.2.6.31) \quad \overset{m\to\infty}{\Longrightarrow} \quad \int_0^1 |c(t)|^2\,\mathrm{d}t = \sum_{j\in\mathbb{Z}}|y_j|^2\,. \tag{4.2.6.32}$$

Theorem 4.2.6.33. Isometry property of the Fourier transform

If the Fourier coefficients satisfy $\sum_{k\in\mathbb{Z}}|\widehat{c}_j|^2<\infty$, *then the Fourier series*

$$c(t) = \sum_{k\in\mathbb{Z}} \widehat{c}_k \exp(-2\pi\imath kt)$$

yields a function $c\in L^2(]0,1[)$ *that satisfies*

$$\|c\|^2_{L^2(]0,1[)} := \int_0^1 |c(t)|^2\,\mathrm{d}t = \sum_{k\in\mathbb{Z}}|\widehat{c}_j|^2\,.$$

Recalling the concept of the L^2-norm of a function, see (5.2.4.6), the theorem can be stated as follows:

> Thm. 4.2.6.33 ↔ The L^2-norm of a Fourier transform agrees with the Euclidean norm of the corresponding sequence.

Here the Euclidean norm of a sequence is understood as $\left\|(y_k)_{k\in\mathbb{Z}}\right\|_2^2 := \sum_{k\in\mathbb{Z}}|y_j|^2$.

From Thm. 4.2.6.33 we can also conclude that the Fourier transform is *injective*: If and only if $c(t)=0$, all its Fourier coefficients will be zero. ⌟

Review question(s) 4.2.6.34 (Semi-discrete Fourier transform)

(Q4.2.6.34.A) Section 4.1.1 introduced the shift operator

$$S_m : \ell^\infty(\mathbb{Z}) \to \ell^\infty(\mathbb{Z}) \quad , \quad S_m\big((x_j)_{j\in\mathbb{Z}}\big) = \big(x_{j-m}\big)_{j\in\mathbb{Z}} . \tag{4.1.1.4}$$

Let $t \mapsto c(t)$ be the Fourier transform of the sequence $(y_k)_{k\in\mathbb{Z}} \in \ell^\infty(\mathbb{Z})$,

$$c(t) = \sum_{k\in\mathbb{Z}} y_k \exp(-2\pi\imath kt) ,$$

What is the relationship of $t \mapsto c(t)$ is the Fourier transform of the shifted sequence $S_m(y_k)_{k\in\mathbb{Z}}$, $m \in \mathbb{Z}$?

△

4.3 Fast Fourier Transform (FFT)

Video tutorial for Section 4.3 "Fast Fourier Transform (FFT)": (16 minutes) Download link, tablet notes

You might have been wondering why the reduction to DFTs has received so much attention in Section 4.2.2. An explanation is given now.

At first glance, DFT in $\mathbb{C}^n$,

$$c_k = \sum_{j=0}^{n-1} y_j \omega_n^{kj} \quad , \quad k = 0,\dots,n-1 , \tag{4.2.1.19}$$

seems to require an asymptotic computational effort of $O(n^2)$ (matrix×vector multiplication the dense Fourier matrix).

C++-code 4.3.0.1: Double-loop implementation of DFT

```
// DFT (4.2.1.19) of vector y returned in c
void naivedft(const VectorXcd& y, VectorXcd& c) {
  using idx_t = VectorXcd::Index;
  const idx_t n = y.size();
  const std::complex<double> i(0,1);
  c.resize(n);
  // root of unity ω_n, w holds its powers
  const std::complex<double> w = std::exp(-2*M_PI/static_cast<double>(n)*i);
  std::complex<double> s = w;
  c(0) = y.sum();
  for (idx_t j = 1; j < n; ++j) {
    c(j) = y(n-1);
    for (idx_t k = n-2; k >= 0; --k) {
      c(j) = c(j)*s + y(k);
    }
    s *= w;
  }
}
```

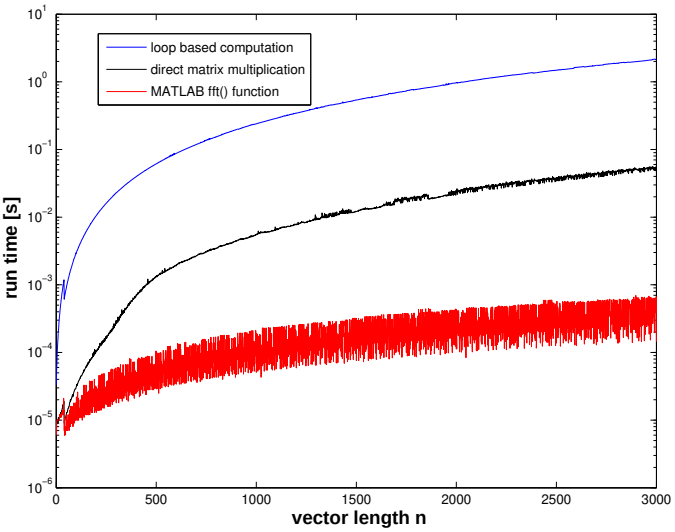
EXPERIMENT 4.3.0.2 (Runtimes of DFT implementations) We examine the runtimes of calls to built-in DFT functions in both MATLAB and EIGEN ↬ GITLAB.

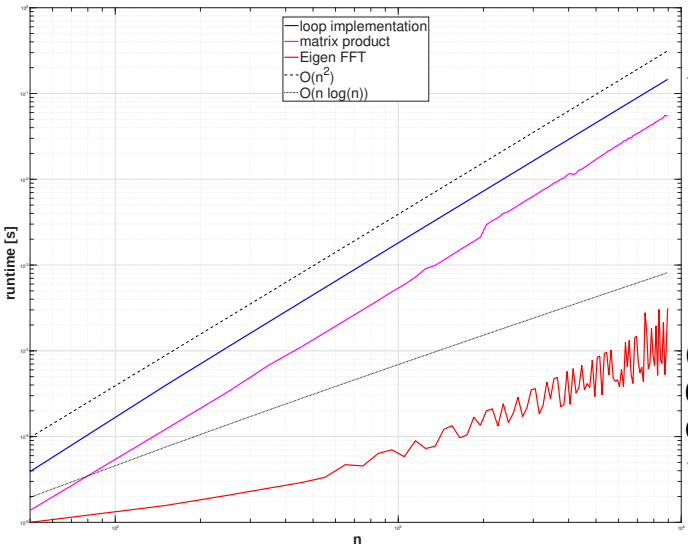
Timings in MATLAB ▷:

1. Straightforward implementation involving MATLAB loops
2. Multiplication with Fourier matrix (4.2.1.13)
3. MATLAB's built-in function `fft()`

(MATLAB V6.5, Linux, Mobile Intel Pentium 4 - M CPU 2.40GHz, minimum over 5 runs)

Similar runtimes would be obtained for PYTHON's `numpy.fft()`.





◁ DFT runtimes in EIGEN:

1. Straightforward implementation involving C++ loops, see Code 4.3.0.1
2. Multiplication with Fourier matrix (4.2.1.13)
3. M. Saladin's **FFT** class, method `fwd()`, PocketFFT backend.

(Intel® Core i9-13900k CPU0-@3GHz, no boost, 64bit, 32GB RAM, OS: Linux, Pop_OS 22.04 LTS, Compiler: gcc (Ubuntu 12.3.0-1ubuntu1 22.04.2) 12.3.0, CXX_FLAGS = -O3 -DNDEBUG)

The secret of MATLAB's/EIGEN's/PYTHON's `fft()`:

the Fast Fourier Transform (FFT) algorithm [DV90]

(discovered by C.F. Gauss in 1805, rediscovered by Cooley & Tuckey in 1965, one of the "top ten algorithms of the century").

§4.3.0.3 (FFT algorithm: derivation and complexity) To understand how the discrete Fourier transform of n-vectors can be implemented with an asymptotic computational effort smaller than $O(n^2)$ we start with an elementary manipulation of (4.2.1.19) for $n = 2m$, $m \in \mathbb{N}$:

$$\begin{aligned} c_k &= \sum_{j=0}^{n-1} y_j e^{-\frac{2\pi i}{n}jk} = \sum_{j=0}^{m-1} y_{2j} e^{-\frac{2\pi i}{n}2jk} + \sum_{j=0}^{m-1} y_{2j+1} e^{-\frac{2\pi i}{n}(2j+1)k} \\ &= \underbrace{\sum_{j=0}^{m-1} y_{2j} \underbrace{e^{-\frac{2\pi i}{m}jk}}_{=\omega_m^{jk}}}_{=:\widetilde{c}_k^{\mathtt{even}}} + e^{-\frac{2\pi i}{n}k} \cdot \underbrace{\sum_{j=0}^{m-1} y_{2k+1} \underbrace{e^{-\frac{2\pi i}{m}jk}}_{=\omega_m^{jk}}}_{=:\widetilde{c}_k^{\mathtt{odd}}}, \quad k \in \mathbb{Z}\,. \end{aligned} \tag{4.3.0.4}$$

and note the m-periodicity: $\widetilde{c}_k^{\mathtt{even}} = \widetilde{c}_{k+m}^{\mathtt{even}}$, $\widetilde{c}_k^{\mathtt{odd}} = \widetilde{c}_{k+m}^{\mathtt{odd}}$ for all $k \in \mathbb{Z}$.

The key observation is that the sequences/m-vectors $\widetilde{c}_k^{\mathtt{even}}$ and $\widetilde{c}_k^{\mathtt{odd}}$ can be computed with DFTs of length

m!

$$\text{with} \quad \mathbf{y}_{\text{even}} := [y_0, y_2, \ldots, y_{n-2}]^\top \in \mathbb{C}^m: \quad [\widetilde{c}_k^{\text{even}}]_{k=0}^{m-1} = \text{DFT}_m(\mathbf{y}_{\text{even}}) ,$$
$$\text{with} \quad \mathbf{y}_{\text{odd}} := [y_1, y_3, \ldots, y_{n-1}]^\top \in \mathbb{C}^m: \quad [\widetilde{c}_k^{\text{odd}}]_{k=0}^{m-1} = \text{DFT}_m(\mathbf{y}_{\text{odd}}) .$$

This means that for even n we can compute $\text{DFT}_n(\mathbf{y})$ from two DFTs of half the length plus $\sim n$ additions and multiplications.

(4.3.0.4): ► DFT of length $2m$ = 2× DFT of length m + 2m additions & multiplications

Idea for $n = 2^L$: Divide & conquer recursion

► FFT-algorithm

The following code shows an EIGEN-based Recursive FFT implementation for DFT of length $n = 2^L$.

C++ code 4.3.0.5: Recursive FFT ↪ GITLAB

```
2    // Recursive DFT for vectors of length n = 2^L
3    VectorXcd
4    fftrec(const VectorXcd &y) {
5  const VectorXcd::Index n = y.size();
6
7  // Nothing to do for DFT of length 1
8  if (n == 1) {
9    return y;
10 }
11 if (n % 2 != 0) {
12   throw std::runtime_error("size(y) must be even!");
13 }
14
15 // Even/odd splitting by rearranging the vector components into a
16 // n/2 × 2 matrix!  See Rem. 1.2.3.6 for use of Eigen::Map
17 const Eigen::Map<const Eigen::Matrix<std::complex<double>, Eigen::Dynamic,
18                                      Eigen::Dynamic, Eigen::RowMajor>>
19     Y(y.data(), n / 2, 2);
20 const VectorXcd c1 = fftrec(Y.col(0));
21 const VectorXcd c2 = fftrec(Y.col(1));
22 // Root of unity ω_n
23 const std::complex<double> omega =
24     std::exp(-2 * M_PI / static_cast<double>(n) * std::complex<double>(0, 1));
25 // Factor in (4.3.0.4)
26 std::complex<double> s(1.0, 0.0);
27 VectorXcd c(n);
28 // Scaling of DFT of odd components plus periodic continuation of
29 // the vectors c1, c2
30 for (Eigen::Index k = 0; k < n; ++k) {
31   c(k) = c1(k % (n / 2)) + c2(k % (n / 2)) * s;
32   s *= omega;
33 }
34 return c;
35 }
```

Visualization of computational cost of `fftrec()` from Code 4.3.0.5:

We see that in Code 4.3.0.5 each level of the recursion requires $O(2^L)$ elementary operations.

Asymptotic complexity of FFT

Asymptotic complexity of FFT algorithm for $n = 2^L$: $O(L2^L) = O(n \log_2 n)$

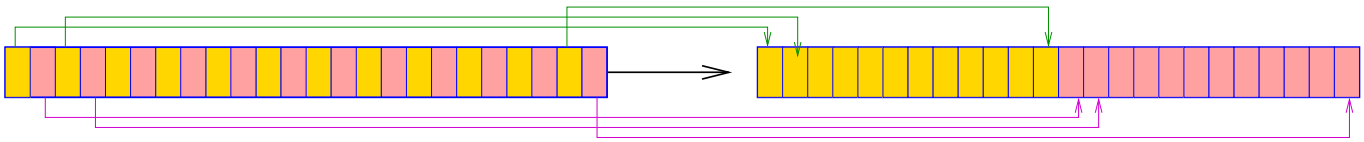
(`fft.fwd()`/`fft.inv()`-function calls in EIGEN: computational cost is $\approx 5n \log_2 n$).

⌟

Remark 4.3.0.7 (FFT algorithm by matrix factorization) The FFT algorithm can also be analyzed on the level of matrix-vector calculus:

For $n = 2m$, $m \in \mathbb{N}$,

consider even-odd sorting $\mathrm{P}_m^{\mathrm{OE}}(1,\ldots,n) = (1,3,\ldots,n-1,2,4,\ldots,n)$.



Also use the notation $\mathrm{P}_m^{\mathrm{OE}}$ for the corresponding permutation of the rows of a matrix $\in \mathbb{C}^{n,n}$, $n = 2m$.

As $\omega_n^{2j} = \omega_m^j$ we conclude

$$\mathrm{P}_m^{\mathrm{OE}}\mathbf{F}_n = \left[\begin{array}{c|c} \mathbf{F}_m & \mathbf{F}_m \\ \hline \mathbf{F}_m \begin{bmatrix} \omega_n^0 & & & \\ & \omega_n^1 & & \\ & & \ddots & \\ & & & \omega_n^{n/2-1} \end{bmatrix} & \mathbf{F}_m \begin{bmatrix} \omega_n^{n/2} & & & \\ & \omega_n^{n/2+1} & & \\ & & \ddots & \\ & & & \omega_n^{n-1} \end{bmatrix} \end{array}\right] =$$

$$\left[\begin{array}{c|c} \mathbf{F}_m & \\ \hline & \mathbf{F}_m \end{array}\right] \left[\begin{array}{c|c} \mathbf{I} & \mathbf{I} \\ \hline \begin{bmatrix} \omega_n^0 & & & \\ & \omega_n^1 & & \\ & & \ddots & \\ & & & \omega_n^{n/2-1} \end{bmatrix} & \begin{bmatrix} -\omega_n^0 & & & \\ & -\omega_n^1 & & \\ & & \ddots & \\ & & & -\omega_n^{n/2-1} \end{bmatrix} \end{array}\right]$$

This reveals how to apply a divide-and-conquer idea when evaluating $\mathbf{F}_n\mathbf{x}$.

Example: partitioning of Fourier matrix for $n = 10$

$$P_5^{\mathrm{OE}}\mathbf{F}_{10} = \left[\begin{array}{ccccc|ccccc} \omega^0 & \omega^0 & \omega^0 & \omega^0 & \omega^0 & \omega^0 & \omega^0 & \omega^0 & \omega^0 & \omega^0 \\ \omega^0 & \omega^2 & \omega^4 & \omega^6 & \omega^8 & \omega^0 & \omega^2 & \omega^4 & \omega^6 & \omega^8 \\ \omega^0 & \omega^4 & \omega^8 & \omega^2 & \omega^6 & \omega^0 & \omega^4 & \omega^8 & \omega^2 & \omega^6 \\ \omega^0 & \omega^6 & \omega^2 & \omega^8 & \omega^4 & \omega^0 & \omega^6 & \omega^2 & \omega^8 & \omega^4 \\ \omega^0 & \omega^8 & \omega^6 & \omega^4 & \omega^2 & \omega^0 & \omega^8 & \omega^6 & \omega^4 & \omega^2 \\ \hline \omega^0 & \omega^1 & \omega^2 & \omega^3 & \omega^4 & \omega^5 & \omega^6 & \omega^7 & \omega^8 & \omega^9 \\ \omega^0 & \omega^3 & \omega^6 & \omega^9 & \omega^2 & \omega^5 & \omega^8 & \omega^1 & \omega^4 & \omega^7 \\ \omega^0 & \omega^5 & \omega^0 & \omega^5 & \omega^0 & \omega^5 & \omega^0 & \omega^5 & \omega^0 & \omega^5 \\ \omega^0 & \omega^7 & \omega^4 & \omega^1 & \omega^8 & \omega^5 & \omega^2 & \omega^9 & \omega^6 & \omega^3 \\ \omega^0 & \omega^9 & \omega^8 & \omega^7 & \omega^6 & \omega^5 & \omega^4 & \omega^3 & \omega^2 & \omega^1 \end{array}\right], \quad \omega := \omega_{10}\,.$$

What if $n \neq 2^L$? Quoted from MATLAB manual:

To compute an n-point DFT when n is composite (that is, when $n = pq$), the FFTW library decomposes the problem using the Cooley-Tukey algorithm, which first computes p transforms of size q, and then computes q transforms of size p. The decomposition is applied recursively to both the p- and q-point DFTs until the problem can be solved using one of several machine-generated fixed-size "codelets." The codelets in turn use several algorithms in combination, including a variation of Cooley-Tukey, a prime factor algorithm, and a split-radix algorithm. The particular factorization of n is chosen heuristically.

The execution time for fft depends on the length of the transform. It is fastest for powers of two. It is almost as fast for lengths that have only small prime factors. It is typically several times slower for lengths that are prime or which have large prime factors → Ex. 4.3.0.13.

Remark 4.3.0.8 (FFT based on general factorization) We motivate the Fast Fourier transform algorithm for DFT of length $n = pq$, $p,q \in \mathbb{N}$ (Cooley-Tuckey algorithm). Again, we start with re-indexing in the DFT formula for a vector $\mathbf{y} = [y_0,\ldots,y_{n-1}] \in \mathbb{C}^n$.

$$c_k = \sum_{j=0}^{n-1} y_j \omega_n^{jk} \overset{[j=:lp+m]}{=} \sum_{m=0}^{p-1}\sum_{l=0}^{q-1} y_{lp+m} e^{-\frac{2\pi i}{pq}(lp+m)k} = \sum_{m=0}^{p-1} \omega_n^{mk} \sum_{l=0}^{q-1} y_{lp+m}\, \omega_q^{l(k \mod q)} . \tag{4.3.0.9}$$

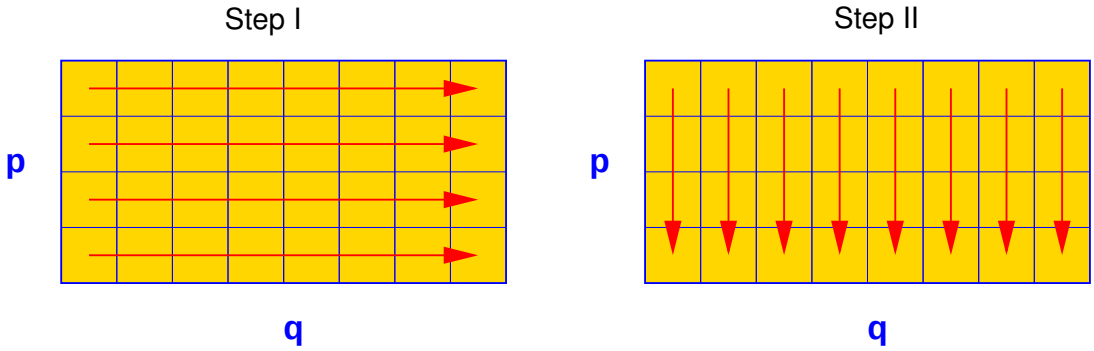
Step I: perform p DFTs of length q, $\mathbf{z}_m = \mathrm{DFT}_q\left([y_{lp+m}]_{l=0}^{q-1} \right)$:

$$(\mathbf{z}_m)_k = z_{m,k} := \sum_{l=0}^{q-1} y_{lp+m}\, \omega_q^{lk} , \quad 0 \le m < p , \quad 0 \le k < q .$$

Step II: for $k =: rq + s$, $0 \le r < p$, $0 \le s < q$ compute

$$c_{rq+s} = \sum_{m=0}^{p-1} e^{-\frac{2\pi i}{pq}(rq+s)m} z_{m,s} = \sum_{m=0}^{p-1} (\omega_n^{ms} z_{m,s}) \omega_p^{mr} ,$$

which is amounts to q DFTs of length p after n multiplications. This gives all components c_k of $\mathrm{DFT}_n\mathbf{y}$.



In fact, the above considerations are the same as those elaborated in Section 4.2.5 that showed that a two-dimensional DFT of $\mathbf{Y} \in \mathbb{C}^{m,n}$ can be done by carrying out m one-dimensional DFTs of length n plus n one-dimensional DFTs of length m, see (4.2.5.5) and Code 4.2.5.7. ⌟

Remark 4.3.0.10 (Rader's algorithm: FFT for prime n) When $n \neq 2^L$, even the Cooley-Tuckey algorithm of Rem. 4.3.0.8 will eventually lead to a DFT for a vector with prime length.

Quoted from the MATLAB manual:

When n is a prime number, the FFTW library first decomposes an n-point problem into three $(n-1)$-point problems using Rader's algorithm [Rad68]. It then uses the Cooley-Tukey decomposition described above to compute the $(n-1)$-point DFTs.

Details of Rader's algorithm: starting point is a theorem from number theory:

$$\forall p \in \mathbb{N} \text{ prime} \quad \exists g \in \{1,\ldots,p-1\}: \quad \{g^k \mod p : k = 1,\ldots,p-1\} = \{1,\ldots,p-1\} ,$$

$$\blacktriangleright \quad \text{permutation} \quad \mathrm{P}_{p,g} : \{1,\ldots,p-1\} \mapsto \{1,\ldots,p-1\} , \quad \mathrm{P}_{p,g}(k) = g^k \mod p ,$$

reversing permutation $\mathrm{P}_k : \{1,\ldots,k\} \mapsto \{1,\ldots,k\}$, $\mathrm{P}_k(i) = k-i+1$.

With these two permutations we can achieve something amazing:

For the Fourier matrix $\mathbf{F}_p = (f_{ij})_{i,j=1}^{p}$ the permuted block $\mathrm{P}_{p-1}\mathrm{P}_{p,g}(f_{ij})_{i,j=2}^{p-1}\mathrm{P}_{p,g}^{\top}$ is circulant.

Example for $p = 13$: $g = 2$, permutation $(2\ 4\ 8\ 3\ 6\ 12\ 11\ 9\ 5\ 10\ 7\ 1)$

$$\mathbf{F}_{13} \longrightarrow \left[\begin{array}{c|cccccccccccc}
\omega^0 & \omega^0 & \omega^0 & \omega^0 & \omega^0 & \omega^0 & \omega^0 & \omega^0 & \omega^0 & \omega^0 & \omega^0 & \omega^0 & \omega^0 \\ \hline
\omega^0 & \omega^2 & \omega^4 & \omega^8 & \omega^3 & \omega^6 & \omega^{12} & \omega^{11} & \omega^9 & \omega^5 & \omega^{10} & \omega^7 & \omega^1 \\
\omega^0 & \omega^1 & \omega^2 & \omega^4 & \omega^8 & \omega^3 & \omega^6 & \omega^{12} & \omega^{11} & \omega^9 & \omega^5 & \omega^{10} & \omega^7 \\
\omega^0 & \omega^7 & \omega^1 & \omega^2 & \omega^4 & \omega^8 & \omega^3 & \omega^6 & \omega^{12} & \omega^{11} & \omega^9 & \omega^5 & \omega^{10} \\
\omega^0 & \omega^{10} & \omega^7 & \omega^1 & \omega^2 & \omega^4 & \omega^8 & \omega^3 & \omega^6 & \omega^{12} & \omega^{11} & \omega^9 & \omega^5 \\
\omega^0 & \omega^5 & \omega^{10} & \omega^7 & \omega^1 & \omega^2 & \omega^4 & \omega^8 & \omega^3 & \omega^6 & \omega^{12} & \omega^{11} & \omega^9 \\
\omega^0 & \omega^9 & \omega^5 & \omega^{10} & \omega^7 & \omega^1 & \omega^2 & \omega^4 & \omega^8 & \omega^3 & \omega^6 & \omega^{12} & \omega^{11} \\
\omega^0 & \omega^{11} & \omega^9 & \omega^5 & \omega^{10} & \omega^7 & \omega^1 & \omega^2 & \omega^4 & \omega^8 & \omega^3 & \omega^6 & \omega^{12} \\
\omega^0 & \omega^{12} & \omega^{11} & \omega^9 & \omega^5 & \omega^{10} & \omega^7 & \omega^1 & \omega^2 & \omega^4 & \omega^8 & \omega^3 & \omega^6 \\
\omega^0 & \omega^6 & \omega^{12} & \omega^{11} & \omega^9 & \omega^5 & \omega^{10} & \omega^7 & \omega^1 & \omega^2 & \omega^4 & \omega^8 & \omega^3 \\
\omega^0 & \omega^3 & \omega^6 & \omega^{12} & \omega^{11} & \omega^9 & \omega^5 & \omega^{10} & \omega^7 & \omega^1 & \omega^2 & \omega^4 & \omega^8 \\
\omega^0 & \omega^8 & \omega^3 & \omega^6 & \omega^{12} & \omega^{11} & \omega^9 & \omega^5 & \omega^{10} & \omega^7 & \omega^1 & \omega^2 & \omega^4 \\
\omega^0 & \omega^4 & \omega^8 & \omega^3 & \omega^6 & \omega^{12} & \omega^{11} & \omega^9 & \omega^5 & \omega^{10} & \omega^7 & \omega^1 & \omega^2
\end{array}\right]$$

Then apply fast algorithms for multiplication with circulant matrices (= discrete periodic convolution, see § 4.1.4.11) to right lower $(n-1)\times(n-1)$ block of permuted Fourier matrix. These fast algorithms rely on DFTs of length $n-1$, see Code 4.2.2.4. ⌟

Since in Section 4.2 we could implement important operations based on the discrete Fourier transform, we can now reap the fruits of the availability of a fast implementation of DFT:

Asymptotic complexity of `fft.fwd()`/`fft.inv()` for $\mathbf{y} \in \mathbb{C}^n$ = $O(n \log n)$.

← Section 4.2.2

Asymptotic complexity of discrete periodic convolution,see Code 4.2.2.4:
Cost(`pconvfft(u,x)`, $\mathbf{u},\mathbf{x} \in \mathbb{C}^n$) = $O(n \log n)$.

Asymptotic complexity of discrete convolution, see Code 4.2.2.5:
Cost(`myconv(h,x)`, $\mathbf{h},\mathbf{x} \in \mathbb{C}^n$) = $O(n \log n)$.

The warning issued in Exp. 2.3.1.7 carries over to numerical methods for signal processing:

Never implement DFT/FFT by yourself!

Under all circumstances use high-quality numerical libraries!

§4.3.0.11 (PocketFFT: Default backend for Eigen::FFT) Unless other backends are specified in the constructor of **Eigen::FFT**, the new implementation of EIGEN's FFT module uses the header-only C++ FTT library PocketFFT. Note that also the NUMPY (from 1.17) and SCIPY (from 1.4) PYTHON libraries rely on PocketFFT as the underlying default FFT implementation. In addition, SCIPY allows the user to select from a range of other FFT backends. ⌟

§4.3.0.12 (FFTW - A highly-performing self-tuning library for FFT)

From FFTW homepage: **FFTW** is a C subroutine library for computing the discrete Fourier transform (DFT) in one or more dimensions, of arbitrary input size, and of both real and complex data.

FFTW will perform well on most architectures without modification. Hence the name, "FFTW," which stands for the somewhat whimsical title of "Fastest Fourier Transform in the West."

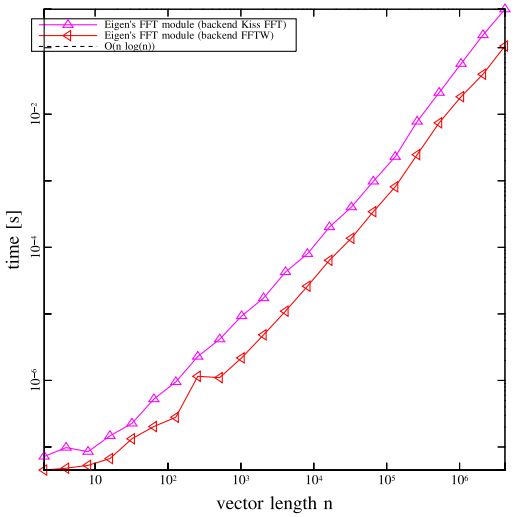
Supplementary literature. [FJ05] offers a comprehensive presentation of the design and implementation of the FFTW library (version 3.x). This paper also conveys the many tricks it takes to achieve satisfactory performance for DFTs of arbitrary length.

FFTW can be installed from source following the instructions from the installation page after downloading the source code of FFTW 3.3.8 from the download page. Precompiled binaries for various linux distributions are available in their main package repositories:

- Ubuntu/Debian: `apt-get install fftw3 fftw3-dev`
- Fedora: `dnf install fftw fftw-devel`

EIGEN's FFT module can use different backend implementations, one of which is the **FFTW** library. The backend may be enabled by defining the preprocessor directive **Eigen_FFTW_DEFAULT** (prior to inclusion of `unsupported/Eigen/FFT`) and linking with the FFTW library (`-lfftw3`). This setup precedure may be handled automatically by a build system like CMake (see `set_eigen_fft_backend` macro on ↪ GITLAB). ⌟

EXAMPLE 4.3.0.13 (Efficiency of FFT for different backend implementations) We measure the runtimes of FTT in EIGEN linking with different libraries, vector lengths $n = 2^L$.



Platform:

- Linux (Ubuntu 16.04 64bit)
- Intel(R) Core(TM) i7-4600U CPU @ 2.10GHz
- L2 256KB, L3 4 MB, 8 GB DDR3 @ 1.60GHz
- Clang 3.8.0, -O3 ⌟

For reasonably high input sizes the FFTW backend gives, compared to EIGEN's default backend (Kiss FFT), a speedup of 2-4x.

Supplementary literature. FTT is covered in almost every textbook on elementary numerical methods, see, for instance [DR08, Sect. 8.7.3], [Han02, Sect. 53], [QSS00, Sect. 10.9.2].

There thousands of online tutorials on FFT, for instance

- The Fast Fourier Transform (FFT): Most Ingenious Algorithm Ever? (Offers an unconventional perspective based on polynomial multiplication.)

Review question(s) 4.3.0.14 (The Fast Fourier Transform (FFT))

(Q4.3.0.14.A) What is the asymptotic complexity for $m, n \to \infty$ of the two-dimensional DFT of a matrix $\mathbf{Y} \in \mathbb{C}^{m,n}$ carried out with the following code:

C++ code 4.2.5.7: Two-dimensional discrete Fourier transform ⇝ GITLAB

```
// 2D DFT implementation using 1D DFTs
template <typename Scalar>
void fft2(Eigen::MatrixXcd &C, const Eigen::MatrixBase<Scalar> &Y) {
  using idx_t = Eigen::MatrixXcd::Index;
  const idx_t m = Y.rows();
  const idx_t n = Y.cols();
  C.resize(m, n);
  Eigen::MatrixXcd tmp(m, n);

  Eigen::FFT fft; // Helper class for DFT
  // Transform rows of matrix Y
  for (idx_t k = 0; k < m; k++) {
    tmp.row(k) = fft.fwd(Y.row(k));
  }

  // Transform columns of temporary matrix
  for (idx_t k = 0; k < n; k++) {
    C.col(k) = fft.fwd(tmp.col(k));
  }
}
```

(Q4.3.0.14.B) Assume that an FFT implementation is available only for vectors of length $n = 2^L$, $L \in \mathbb{N}$. How do you have to modify the following C++ function for the discrete convolution of two vectors $\mathbf{h}, \mathbf{x} \in \mathbb{C}^n$ to ensure that it still enjoys an asymptotic complexity of $O(n \log n)$ for $n \to \infty$?

C++ code 4.2.2.5: Implementation of discrete convolution (→ Def. 4.1.3.3) based on periodic discrete convolution ⇝ GITLAB

```
inline Eigen::VectorXcd fastconv(const Eigen::VectorXcd &h,
                                 const Eigen::VectorXcd &x) {
  assert(x.size() == h.size());
  const Eigen::Index n = h.size();
  // Zero padding, cf. (4.1.4.16), and periodic discrete convolution
  // of length 2n-1, Code 4.2.2.4
  return pconvfft(
      (Eigen::VectorXcd(2 * n - 1) << h, Eigen::VectorXcd::Zero(n - 1))
          .finished(),
      (Eigen::VectorXcd(2 * n - 1) << x, Eigen::VectorXcd::Zero(n - 1))
          .finished());
}
```

(Q4.3.0.14.C) Again assume that the FFT implementation of EIGEN is available only for vectors of length $n = 2^L$, $L \in \mathbb{N}$. Propose changes to the following C++ function for the discrete periodic convolution of two vectors $\mathbf{u}, \mathbf{x} \in \mathbb{C}^n$ that preserve the asymptotic complexity of $O(n \log n)$ for $n \to \infty$.

C++ code cpp:pconvfft: Discrete periodic convolution: DFT implementation ⇝ GITLAB

```
inline Eigen::VectorXcd pconvfft(const Eigen::VectorXcd &u,
                                 const Eigen::VectorXcd &x) {
  Eigen::FFT fft; // Helper object, see § 4.2.1.21
```

```
  // One-line implementation of (4.2.2.3)
  return fft.inv(((fft.fwd(u)).cwiseProduct(fft.fwd(x))));
}
```

(Q4.3.0.14.D) A family of square matrices $\mathbf{H}_m \in \mathbb{R}^{n,n}$, $m \in \mathbb{N}_0$, $n := 2^m$, is recursively defined as

$$\mathbf{H}_0 := [1] \quad , \quad \mathbf{H}_m := \begin{bmatrix} \mathbf{H}_{m-1} & \mathbf{H}_{m-1} \\ \mathbf{H}_{m-1} & -\mathbf{H}_{m-1} \end{bmatrix} , \quad m \in \mathbb{N} .$$

Devise a recursive algorithm for computing the matrix×vector product $\mathbf{H}_m\mathbf{x}$, $\mathbf{x} \in \mathbb{R}^n$ and determine its asymptotic complexity in terms of $n := 2^m \to \infty$.

△

4.4 Trigonometric Transformations

Supplementary literature. [Han02, Sect. 55], see also [Str99] for an excellent presentation of various variants of the cosine transform.

Keeping in mind $\exp(2\pi ix) = \cos(2\pi x) + \imath \sin(2\pi x)$ we may also consider the real/imaginary parts of the Fourier basis vectors $(\mathbf{F}_n)_{:,j}$ as bases of $\mathbb{R}^n$ and define the corresponding basis transformation. They can all be realized by means of `fft` with an asymptotic computational effort of $O(n \log n)$. These transformations *avoid the use of complex numbers*. Details are given in the sequel.

4.4.1 Sine transform

This is a purely real trigonometric basis transform in $\mathbb{R}^{n-1}$, $n \in \mathbb{N}$, and the "sine basis vectors" are given by the following formulas:

Standard basis of $\mathbb{R}^{n-1}$ "Sine basis"

$$\left\{ \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \cdots \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \right\} \longleftarrow \left\{ \begin{bmatrix} \sin(\frac{\pi}{n}) \\ \sin(\frac{2\pi}{n}) \\ \vdots \\ \sin(\frac{(n-1)\pi}{n}) \end{bmatrix} \begin{bmatrix} \sin(\frac{2\pi}{n}) \\ \sin(\frac{4\pi}{n}) \\ \vdots \\ \sin(\frac{2(n-1)\pi}{n}) \end{bmatrix} \cdots \begin{bmatrix} \sin(\frac{(n-1)\pi}{n}) \\ \sin(\frac{2(n-1)\pi}{n}) \\ \vdots \\ \sin(\frac{(n-1)^2\pi}{n}) \end{bmatrix} \right\} \tag{4.4.1.1}$$

Basis transform matrix (sine basis → standard basis): $\mathbf{S}_n := (\sin(jk\pi/n))_{j,k=1}^{n-1} \in \mathbb{R}^{n-1,n-1}$.

Lemma 4.4.1.2. Properties of the sine matrix

$\sqrt{2/n}\,\mathbf{S}_n \in \mathbb{R}^{n,n}$ *is real, symmetric and orthogonal (→ Def. 6.3.1.2)*

Proof. Make a detour via the complex plane, rewrite $\sin(jk\pi/n) = \frac{1}{2\imath}(\exp(\imath\pi jk/n) + \exp(-iu\pi jk/n))$, follow the proof of Lemma 4.2.1.14 and, finally, use the geometric sum formula (4.2.1.8). □

The **sine transform** of a vector $\mathbf{y} = [y_1, \ldots, y_{n-1}]^\top \in \mathbb{R}^{n-1}$ yields another vector $\mathbf{s} = [s_1, \ldots, s_{n-1}]^\top$ according to

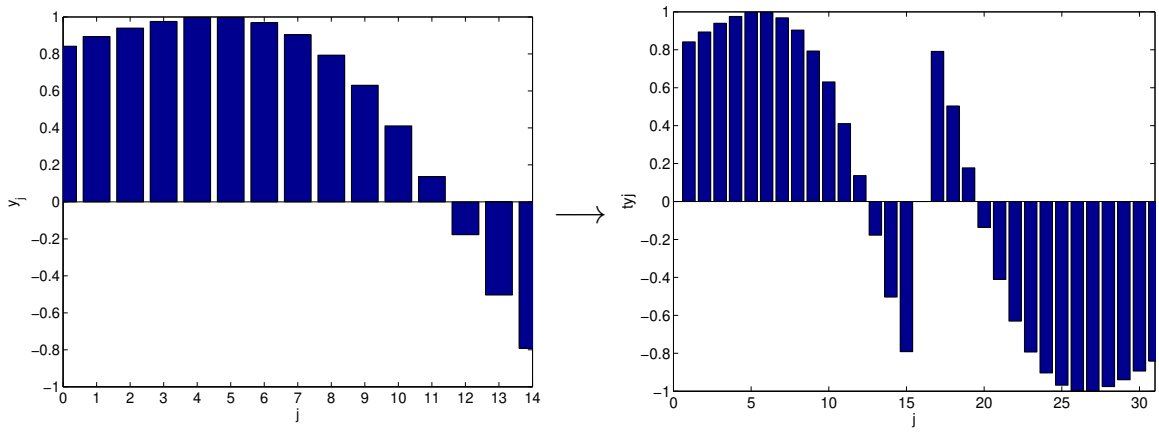
$$s_k = \sum_{j=1}^{n-1} y_j \sin(\pi jk/n) \quad , \quad k = 1, \ldots, n-1 \, . \tag{4.4.1.3}$$

It defines a linear mapping $\mathbb{R}^{n-1} \to \mathbb{R}^{n-1}$.

By elementary consideration we can devise a DFT-based algorithm for the sine transform ($\hat{=}$ $\mathbf{S}_n \times$vector):

tool: "wrap around": $\widetilde{\mathbf{y}} \in \mathbb{R}^{2n}$: $\widetilde{y}_j = \begin{cases} y_j & \text{, if } j = 1, \ldots, n-1 \, , \\ 0 & \text{, if } j = 0, n \, , \\ -y_{2n-j} & \text{, if } j = n+1, \ldots, 2n-1 \, . \end{cases}$ ($\widetilde{\mathbf{y}}$ "odd")

This "wrap around" transformation can be visualized as follows:



Next we use $\sin(x) = \frac{1}{2\imath}(\exp(\imath x) - \exp(-\imath x)$ to identify the DFT of a wrapped around vector as a sine transform:

$$\begin{aligned} (\mathbf{F}_{2n}\widetilde{\mathbf{y}})_k &\overset{(4.2.1.19)}{=} \sum_{j=1}^{2n-1} \widetilde{y}_j e^{-\frac{2\pi\imath}{2n}kj} = \sum_{j=1}^{n-1} y_j e^{-\frac{\pi\imath}{n}kj} - \sum_{j=n+1}^{2n-1} y_{2n-j} e^{-\frac{\pi\imath}{n}kj} \\ &= \sum_{j=1}^{n-1} y_j (e^{-\frac{\pi\imath}{n}kj} - e^{\frac{\pi\imath}{n}kj}) = -2i\,(\mathbf{S}_n\mathbf{y})_k \quad , k = 1, \ldots, n-1 \, . \end{aligned}$$

C++ code 4.4.1.4: Wrap-around implementation of sine transform ↪ GITLAB

```
// Simple sine transform of y ∈ R^n − 1 into
// c ∈ R^{n−1} by (4.4.1.3)
void sinetrfwrap(const VectorXd &y, VectorXd &c) {
  VectorXd::Index n = y.size() + 1;
  // Create wrapped vector ỹ
  VectorXd yt(2 * n);
  yt << 0, y, 0, -y.reverse();

  Eigen::VectorXcd ct;
  Eigen::FFT fft; // DFT helper class
  fft.fwd(ct, yt);

  const std::complex<double> v(0, 2); // factor 2ı
```

```
  c = (-ct.middleRows(1, n - 1) / v).real();
}
```

Remark 4.4.1.5 (Sine transform via DFT of half length) The simple Code 4.4.1.4 relies on a DFT for vectors of length $2n$, which may be a waste of computational resources in some applications. A DFT of length n is sufficient as demonstrated by the following manipulations.

Step ①: transform of the coefficients

$$\widetilde{y}_j = \sin(j\pi/n)(y_j + y_{n-j}) + \tfrac{1}{2}(y_j - y_{n-j}) \;,\quad j = 1,\ldots,n-1 \quad,\quad \widetilde{y}_0 = 0 \;. \tag{4.4.1.6}$$

Step ②: real DFT (→ Section 4.2.4) of $(\widetilde{y}_0,\ldots,\widetilde{y}_{n-1}) \in \mathbb{R}^n$: $\quad c_k := \sum_{j=0}^{n-1} \widetilde{y}_j\, e^{-\frac{2\pi i}{n}jk}$

$$\begin{aligned}
\text{Hence}\quad \operatorname{Re}\{c_k\} &= \sum_{j=0}^{n-1} \widetilde{y}_j \cos(-\tfrac{2\pi i}{n}jk) = \sum_{j=1}^{n-1} (y_j + y_{n-j}) \sin(\tfrac{\pi j}{n}) \cos(\tfrac{2\pi i}{n}jk) \\
&= \sum_{j=0}^{n-1} 2y_j \sin(\tfrac{\pi j}{n}) \cos(\tfrac{2\pi i}{n}jk) = \sum_{j=0}^{n-1} y_j \left(\sin(\frac{2k+1}{n}\pi j) - \sin(\frac{2k-1}{n}\pi j) \right) \\
&= s_{2k+1} - s_{2k-1} \;. \\
\operatorname{Im}\{c_k\} &= \sum_{j=0}^{n-1} \widetilde{y}_j \sin(-\tfrac{2\pi i}{n}jk) = -\sum_{j=1}^{n-1} \tfrac{1}{2}(y_j - y_{n-j}) \sin(\tfrac{2\pi i}{n}jk) = -\sum_{j=1}^{n-1} y_j \sin(\tfrac{2\pi i}{n}jk) \\
&= -s_{2k} \;.
\end{aligned}$$

Step ③: extraction of s_k

$$s_{2k+1} \;,\quad k = 0,\ldots,\tfrac{n}{2}-1 \quad\blacktriangleright\quad \text{from recursion } s_{2k+1} - s_{2k-1} = \operatorname{Re}\{c_k\} \;,\quad s_1 = \sum_{j=1}^{n-1} y_j \sin(\pi j/n) \;,$$

$$s_{2k} \;,\quad k = 1,\ldots,\tfrac{n}{2}-2 \quad\blacktriangleright\quad s_{2k} = -\operatorname{Im}\{c_k\} \;.$$

Implementation (via a `fft` of length $n/2$):

C++ code 4.4.1.7: Sine transform ↬ GITLAB

```
inline void sinetransform(const Eigen::VectorXd &y, Eigen::VectorXd &s) {
  const Eigen::Index n = y.rows() + 1;
  constexpr std::complex<double> i(0, 1);

  // Step ①: Precompute sine factors
  const Eigen::VectorXd x =
      Eigen::VectorXd::LinSpaced(n - 1, 1, static_cast<double>(n - 1));
  const Eigen::VectorXd sinevals = x.unaryExpr([&](double z) {
    return imag(std::pow(std::exp(i * M_PI / static_cast<double>(n)), z));
  });
  // Transform data vector according to (4.4.1.6)
  Eigen::VectorXd yt(n);
  yt(0) = 0;
  yt.tail(n - 1) = sinevals.array() * (y + y.reverse()).array() +
                   0.5 * (y - y.reverse()).array();
  // Step ②: Perform DFT of a real vector, using the complex result.
  Eigen::VectorXcd c;
```

```
19    Eigen::FFT fft;
20    fft.fwd(c, yt);
21    // Step ③: Extract sine transform from complex result of DFT
22    s.resize(n);
23    s(0) = sinevals.dot(y);
24
25    for (Eigen::Index k = 2; k <= n - 1; ++k) {
26      const Eigen::Index j =
27          k - 1;  // Shift index to consider indices starting from 0
28      if (k % 2 == 0) {
29        s(j) = -c(k / 2).imag();
30      } else {
31        s(j) = s(j - 2) + c((k - 1) / 2).real();
32      }
33    }
34  }
```

⌟

EXAMPLE 4.4.1.8 (Diagonalization of local translation invariant linear grid operators) We consider a so-called **5-point stencil**-operator on $\mathbb{R}^{n,n}$, $n \in \mathbb{N}$, defined as follows

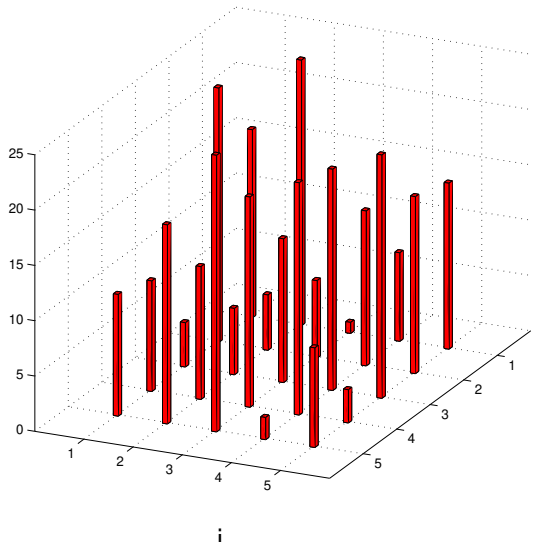
$$\mathsf{T} : \begin{cases} \mathbb{R}^{n,n} & \to & \mathbb{R}^{n,n}, \\ \mathbf{X} & \mapsto & \mathsf{T}(\mathbf{X}), \end{cases} \quad (\mathsf{T}(\mathbf{X}))_{ij} := cx_{ij} + c_y x_{i,j+1} + c_y x_{i,j-1} + c_x x_{i+1,j} + c_x x_{i-1,j} \tag{4.4.1.9}$$

with coefficients $c, c_y, c_x \in \mathbb{R}$, convention: $x_{ij} := 0$ for $(i,j) \notin \{1,\ldots,n\}^2$.

A matrix can be regarded as a function that assigns values (= matrix entries) to the points of a 2D lattice/-grid:

$$\begin{array}{c} \text{Matrix } \mathbf{X} \in \mathbb{R}^{n,n} \\ \updownarrow \\ \text{grid function} \in \{1,\ldots,n\}^2 \mapsto \mathbb{R} \end{array}$$

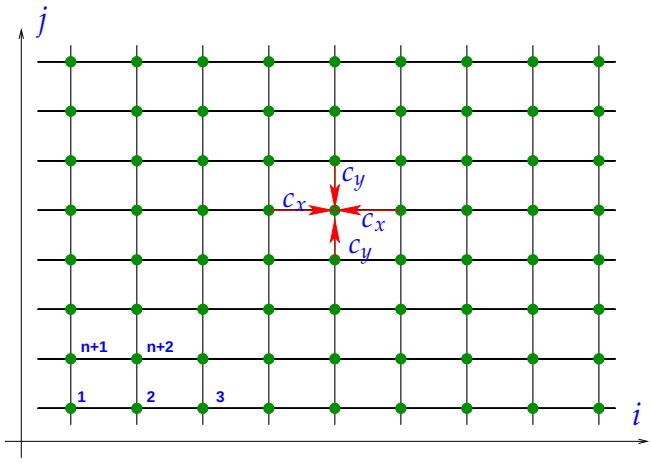
Visualization of a grid function ▷



Identification $\mathbb{R}^{n,n} \cong \mathbb{R}^{n^2}$, $x_{ij} \sim \widetilde{x}_{(j-1)n+i}$ (row-wise numbering) gives a matrix representation $\mathbf{T} \in \mathbb{R}^{n^2,n^2}$ of T:

$$\mathbf{T} = \begin{bmatrix} \mathbf{C} & c_y\mathbf{I} & 0 & \cdots & \cdots & 0 \\ c_y\mathbf{I} & \mathbf{C} & c_y\mathbf{I} & & & \vdots \\ 0 & \ddots & \ddots & \ddots & & \\ \vdots & & & c_y\mathbf{I} & \mathbf{C} & c_y\mathbf{I} \\ 0 & \cdots & \cdots & 0 & c_y\mathbf{I} & \mathbf{C} \end{bmatrix} \in \mathbb{R}^{n^2,n^2} ,$$

$$\mathbf{C} = \begin{bmatrix} c & c_x & 0 & \cdots & \cdots & 0 \\ c_x & c & c_x & & & \vdots \\ 0 & \ddots & \ddots & \ddots & & \\ \vdots & & & c_x & c & c_x \\ 0 & \cdots & \cdots & 0 & c_x & c \end{bmatrix} \in \mathbb{R}^{n,n} .$$

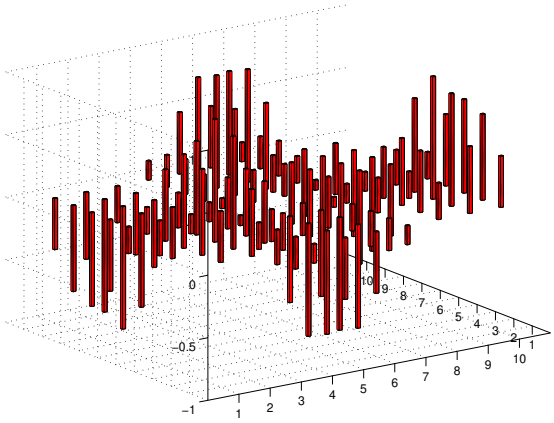


We already know the sine basis of $\mathbb{R}^{n,n}$:

$$\mathbf{B}^{kl} = \left(\sin(\tfrac{\pi}{n+1}ki)\sin(\tfrac{\pi}{n+1}lj)\right)_{i,j=1}^{n} . \quad (4.4.1.10)$$

These matrices will also provide a basis of the vector space of grid functions $\{1,\ldots,n\}^2 \mapsto \mathbb{R}$.

$n = 10$: grid function $\mathbf{B}^{2,3}$ ➢



The key observation is that elements of the sine basis are eigenvectors of T:

$$\begin{aligned}(T(\mathbf{B}^{kl}))_{ij} &= c\sin(\tfrac{\pi}{n+1}ki)\sin(\tfrac{\pi}{n+1}lj) + c_y\sin(\tfrac{\pi}{n+1}ki)\left(\sin(\tfrac{\pi}{n+1}l(j-1)) + \sin(\tfrac{\pi}{n+1}l(j+1))\right) + \\ &\qquad c_x\sin(\tfrac{\pi}{n+1}lj)\left(\sin(\tfrac{\pi}{n+1}k(i-1)) + \sin(\tfrac{\pi}{n+1}k(i+1))\right) \\ &= \sin(\tfrac{\pi}{n+1}ki)\sin(\tfrac{\pi}{n+1}lj)(c + 2c_y\cos(\tfrac{\pi}{n+1}l) + 2c_x\cos(\tfrac{\pi}{n+1}k))\end{aligned}$$

Hence $\mathbf{B}^{kl}$ is eigenvector of T (or $\mathbf{T}$ after row-wise numbering) and the corresponding eigenvalue is given by $c + 2c_y\cos(\frac{\pi}{n+1}l) + 2c_x\cos(\frac{\pi}{n+1}k)$. Recall very similar considerations for discrete (periodic) convolutions in 1D (→ § 4.2.1.6) and 2D (→ § 4.2.5.10)

The basis transform can be implemented efficiently based on the 1D sine transform:

$$\mathbf{X} = \sum_{k=1}^{n}\sum_{l=1}^{n} y_{kl}\mathbf{B}^{kl} \quad \Rightarrow \quad x_{ij} = \sum_{k=1}^{n}\sin(\tfrac{\pi}{n+1}ki)\sum_{l=1}^{n} y_{kl}\sin(\tfrac{\pi}{n+1}lj) .$$

Hence nested sine transforms (→ Section 4.2.5) for rows/columns of $\mathbf{Y} = (y_{kl})_{k,l=1}^{n}$.

Here: implementation of sine transform (4.4.1.3) with "wrapping"-technique.

C++ code 4.4.1.11: 2D sine transform ⇝ GITLAB

```
2 inline void sinetransform2d(const Eigen::MatrixXd &Y, Eigen::MatrixXd &S) {
3   const Eigen::Index m = Y.rows();
4   const Eigen::Index n = Y.cols();
5
6   Eigen::VectorXcd c;
```

```
  Eigen::FFT fft;
  const std::complex<double> i(0, 1);

  // Implementation of sine transforms as in Code 4.4.1.4
  Eigen::MatrixXcd C(2 * m + 2, n);
  C.row(0) = Eigen::VectorXcd::Zero(n);
  C.middleRows(1, m) = Y.cast<std::complex<double>>();
  C.row(m + 1) = Eigen::VectorXcd::Zero(n);
  C.middleRows(m + 2, m) = -Y.colwise().reverse().cast<std::complex<double>>();

  // FFT on each column of C
  for (int i = 0; i < n; ++i) {
    fft.fwd(c, C.col(i));
    C.col(i) = c;
  }

  // Another sine transform implemented by wrap around
  C.middleRows(1, m) = i * C.middleRows(1, m) / 2.;
  Eigen::MatrixXcd C2(2 * n + 2, m);
  C2.row(0) = Eigen::VectorXcd::Zero(m);
  C2.middleRows(1, n) = C.middleRows(1, m).transpose();
  C2.row(n + 1) = Eigen::VectorXcd::Zero(m);
  C2.middleRows(n + 2, n) = -C.middleRows(1, m).transpose().colwise().reverse();

  // FFT on each column of C2
  for (int i = 0; i < m; ++i) {
    fft.fwd(c, C2.col(i));
    C2.col(i) = c;
  }

  S = (i * C2.middleRows(1, n).transpose() / 2.).real();
}
```

C++ code 4.4.1.12: FFT-based solution of local translation invariant linear operators ➙ GITLAB

```
inline void fftbasedsolutionlocal(const Eigen::MatrixXd& B, double c, double cx,
                                  double cy, Eigen::MatrixXd& X) {
  const Eigen::Index m = B.rows();
  const Eigen::Index n = B.cols();

  // Eigen's meshgrid
  const Eigen::MatrixXd I =
      Eigen::RowVectorXd::LinSpaced(n, 1, static_cast<double>(n))
          .replicate(m, 1);
  const Eigen::MatrixXd J =
      Eigen::VectorXd::LinSpaced(m, 1, static_cast<double>(m)).replicate(1, n);
  // Perform 2D sine transform of right-hand side matrix data.
  Eigen::MatrixXd X_;
  sinetransform2d(B, X_);
  // Carry out entrywise scaling with trigonometric polynomial defined
  // by the 5-point stencil
  Eigen::MatrixXd T;
  T = c + 2 * cx * (M_PI / (static_cast<double>(n) + 1) * I).array().cos() +
      2 * cy * (M_PI / (static_cast<double>(m) + 1) * J).array().cos();
  X_ = X_.cwiseQuotient(T);
  // Transform back from "sine domain" to "matrix domain".
  sinetransform2d(X_, X);
  X = 4 * X / ((m + 1) * (n + 1));
```

```
25 }
```

Thus the diagonalization of $\mathbf{T}$ via 2D sine transform yields an efficient algorithm for solving linear system of equations $\mathsf{T}(\mathbf{X}) = \mathbf{B}$: computational cost $O(n^2 \log n)$. ⌟

EXPERIMENT 4.4.1.13 (Efficiency of FFT-based solver) In the experiment we test the gain in runtime obtained by using DFT-based algorithms for solving linear systems of equations with coefficient matrix $\mathbf{T}$ induced by the operator T from (4.4.1.9) with the values

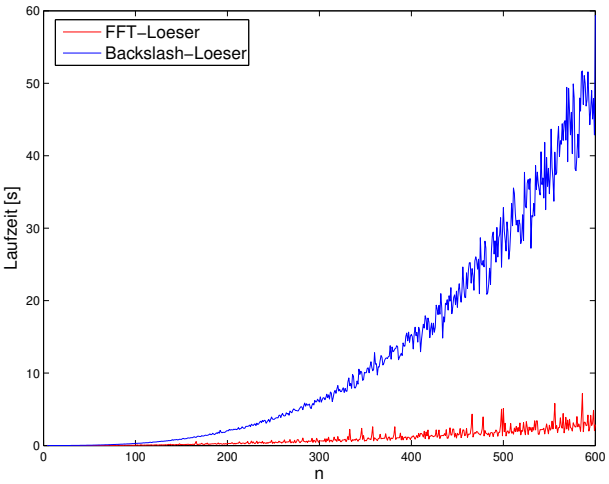
$$c = 4 \quad , \quad c_x = c_y = -1 \, .$$

This means

$$\mathbf{T} := \begin{bmatrix} \mathbf{C} & -\mathbf{I} & 0 & \cdots & \cdots & 0 \\ -\mathbf{I} & \mathbf{C} & -\mathbf{I} & & & \vdots \\ 0 & \ddots & \ddots & \ddots & & \\ \vdots & & & \mathbf{I} & \mathbf{C} & -\mathbf{I} \\ 0 & \cdots & \cdots & 0 & -\mathbf{I} & \mathbf{C} \end{bmatrix} \in \mathbb{R}^{n^2,n^2} \, , \quad \mathbf{C} := \begin{bmatrix} 4 & -1 & 0 & \cdots & \cdots & 0 \\ -1 & 4 & -1 & & & \vdots \\ 0 & \ddots & \ddots & \ddots & & \\ \vdots & & & -1 & 4 & -1 \\ 0 & \cdots & \cdots & 0 & -1 & 4 \end{bmatrix} \in \mathbb{R}^{n,n} \, .$$

`tic-toc`-timing (MATLAB) V7, Linux, Intel Pentium 4 Mobile CPU 1.80GHz)

Similar results would be obtained by an implementation in PYTHON.



⌟

4.4.2 Cosine transform

Another trigonometric basis transform in $\mathbb{R}^n$, $n \in \mathbb{N}$:

standard basis of $\mathbb{R}^n$ "cosine basis"

$$\left\{ \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \cdots \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \right\} \leftarrow \left\{ \begin{bmatrix} 2^{-1/2} \\ \cos(\frac{\pi}{2n}) \\ \cos(\frac{2\pi}{2n}) \\ \vdots \\ \cos(\frac{(n-1)\pi}{2n}) \end{bmatrix} \begin{bmatrix} 2^{-1/2} \\ \cos(\frac{3\pi}{2n}) \\ \cos(\frac{6\pi}{2n}) \\ \vdots \\ \cos(\frac{3(n-1)\pi}{2n}) \end{bmatrix} \cdots \begin{bmatrix} 2^{-1/2} \\ \cos(\frac{(2n-1)\pi}{2n}) \\ \cos(\frac{2(2n-1)\pi}{2n}) \\ \vdots \\ \cos(\frac{(n-1)(2n-1)\pi}{2n}) \end{bmatrix} \right\} .$$

Basis transform matrix (cosine basis $\to$ standard basis):

$$\mathbf{C}_n = \left[c_{ij}\right]_{i,j=0}^{n-1} \in \mathbb{R}^{n,n} \quad \text{with} \quad c_{ij} = \begin{cases} 2^{-1/2} & \text{, if } i = 0 \, , \\ \cos(i\frac{2j+1}{2n}\pi) & \text{, if } i > 0 \, . \end{cases}$$

Lemma 4.4.2.1. Properties of cosine matrix

The matrix $\sqrt{2/n}\,\mathbf{C}_n \in \mathbb{R}^{n,n}$ is real and orthogonal (→ Def. 6.3.1.2) .

Note: $\mathbf{C}_n$ is not symmetric.

cosine transform of $\mathbf{y} = [y_0, \ldots, y_{n-1}]^\top$:

$$c_k = \sum_{j=0}^{n-1} y_j \cos(k\tfrac{2j+1}{2n}\pi) \quad , \quad k = 1, \ldots, n-1 \, , \tag{4.4.2.2}$$

$$c_0 = \frac{1}{\sqrt{2}} \sum_{j=0}^{n-1} y_j \, .$$

Implementation of $\mathbf{Cy}$ using the "wrapping"-technique as in Code 4.4.1.4:

C++ code 4.4.2.3: Cosine transform ↪ GITLAB

```
2  inline void cosinetransform(const Eigen::VectorXd &y, Eigen::VectorXd &c) {
3    const Eigen::Index n = y.size();
4
5    Eigen::VectorXd y_(2 * n);
6    y_.head(n) = y;
7    y_.tail(n) = y.reverse();
8
9    // FFT
10   Eigen::VectorXcd z;
11   Eigen::FFT fft;
12   fft.fwd(z, y_);
13
14   const std::complex<double> i(0, 1);
15   c.resize(n);
16   c(0) = z(0).real() / (2 * sqrt(2));
17   for (Eigen::Index j = 1; j < n; ++j) {
18     c(j) = (0.5 * pow(exp(-i * M_PI / (2 * static_cast<double>(n))), j) * z(j))
19                .real();
20   }
21 }
```

Implementation of $\mathbf{C}_n^{-1}\mathbf{y}$ ("Wrapping"-technique):

C++-code 4.4.2.4: Inverse cosine transform ↪ GITLAB

```
2  inline void icosinetransform(const Eigen::VectorXd &c, Eigen::VectorXd &y) {
3    const Eigen::Index n = c.size();
4
5    const std::complex<double> i(0, 1);
6    Eigen::VectorXcd c_1(n + 1);
7    c_1(0) = sqrt(2) * c(0);
8    for (Eigen::Index j = 1; j < n; ++j) {
9      c_1(j) = pow(exp(-i * M_PI / (2 * static_cast<double>(n))), j) * c(j);
10   }
11   c_1(n) = 0;
12
13   // FFT
14   Eigen::VectorXd z;
```

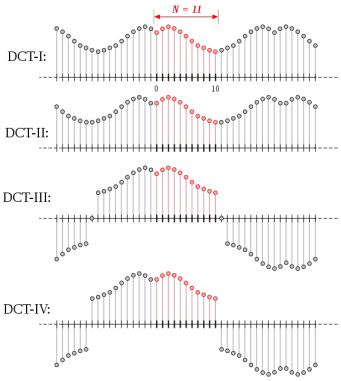
```
  // using HalfSpectrum mode, c_1 is interpreted as a symmetric
     conjugate
  // vector half, i.e.  the inverse FFT is actually computed on
  // [c_1, c_1.tail(n-1).reverse().conjugate()].
  Eigen::FFT<Eigen::FFTOption::HalfSpectrum | Eigen::FFTOption::Scaled> fft;
  fft.inv(z, c_1, 2 * n);

  // To obtain the same result as Matlab,
  // shift the inverse FFT result by 1.
  Eigen::VectorXd y_(2 * n);
  y_.head(2 * n - 1) = z.tail(2 * n - 1);
  y_(2 * n - 1) = z(0);

  y = 2 * y_.head(n);
}
```

Remark 4.4.2.5 (Cosine transforms for compression)

The cosine transforms discussed above are named DCT-II and DCT-III.
Various cosine transforms arise by imposing various boundary conditions:

- DCT-II: even around $-1/2$ and $N-1/2$
- DCT-III: even around 0 and odd around N

DCT-II is used in JPEG-compression while a slightly modified DCT-IV makes the main component of MP3, AAC and WMA formats.



⌟

Review question(s) 4.4.2.6 (Trigonometric Transformations)

△

4.5 Toeplitz Matrix Techniques

Video tutorial for Section 4.5 "Toeplitz Matrix Techniques": (20 minutes) Download link, tablet notes

This section examines FFT-based algorithms for more general problems in numerical linear algebra. It connects to the matrix perspective of DFT and linear filters that was adopted occasionally in Section 4.1 and Section 4.2.

4.5.1 Matrices with Constant Diagonals

EXAMPLE 4.5.1.1 (Parameter identification for linear time-invariant filters) We want to determine the impulse response of an LT-FIR channel from (noisy) measurements:

- Given: $(x_k)_{k\in\mathbb{Z}}$ m-periodic discrete signal = *known* input
- Given: $(y_k)_{k\in\mathbb{Z}}$ m-periodic *measured*$^{(*)}$ output signal of a linear time-invariant filter, see Section 4.1.1.
  $(*)$ ➔ inevitably affected by measurement errors!

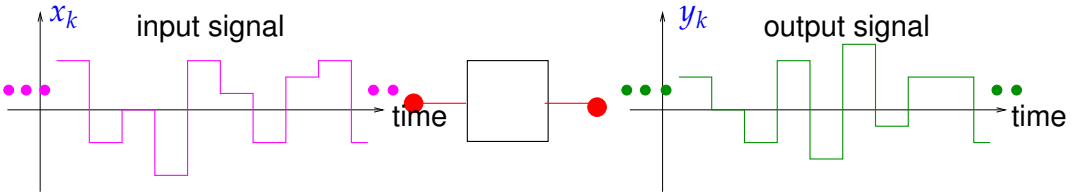
• Sought: Estimate for the impulse response (→ Def. 4.1.1.12) of the filter

This task reminds us of the parameter estimation problem from Ex. 3.0.1.4, which we tackled with least squares techniques. We employ similar ideas for the current problem

• Known: impulse response of filter has maximal duration $n\Delta t$, $n \in \mathbb{N}$, $n \leq m$

cf. (4.1.2.4)

$$\exists \mathbf{h} = [h_0, \ldots, h_{n-1}]^\top \in \mathbb{R}^n, \quad n \leq m : \quad y_k = \sum_{j=0}^{n-1} h_j x_{k-j} \,. \tag{4.5.1.2}$$



If the y_k were exact, we could retrieve $h_0, \ldots, h_{n-1}$ by examining only $y_0, \ldots, y_{n-1}$ and inverting the discrete periodic convolution (→ Def. 4.1.4.7) using (4.2.1.17).

However, in case the y_k are affected by measurements errors it is advisable to use all available y_k for a least squares estimate of the impulse response.

We can now formulate the least squares parameter identification problem: seek $\mathbf{h} = [h_0, \ldots, h_{n-1}]^\top \in \mathbb{R}^n$ with

$$\|\mathbf{A}\mathbf{h} - \mathbf{y}\|_2 = \left\| \begin{bmatrix} x_0 & x_{-1} & \cdots & & \cdots & x_{1-n} \\ x_1 & x_0 & x_{-1} & & & \vdots \\ \vdots & x_1 & x_0 & \ddots & & \\ & & \ddots & \ddots & & \vdots \\ \vdots & & & \ddots & \ddots & x_{-1} \\ x_{n-1} & & & & x_1 & x_0 \\ x_n & x_{n-1} & & & & x_1 \\ \vdots & & & & & \vdots \\ x_{m-1} & \cdots & & & \cdots & x_{m-n} \end{bmatrix} \begin{bmatrix} h_0 \\ \vdots \\ \\ \\ \vdots \\ h_{n-1} \end{bmatrix} - \begin{bmatrix} y_0 \\ \vdots \\ \\ \\ \vdots \\ y_{m-1} \end{bmatrix} \right\|_2 \to \min \,.$$

This is a linear least squares problem as introduced in Chapter 3 with a coefficient matrix $\mathbf{A}$ that enjoys the property that $(\mathbf{A})_{ij} = x_{i-j}$, which means that all its diagonals have constant entries.

The coefficient matrix for the normal equations (→ Section 3.1.2, Thm. 3.1.2.1) corresponding to the above linear least squares problem is

$$\mathbf{M} := \mathbf{A}^\top \mathbf{A} \quad , \quad (\mathbf{M})_{ij} = \sum_{k=0}^{m-1} x_{k-i} x_{k-j} =: z_{i-j}$$

for some m-periodic sequence $(z_k)_{k\in\mathbb{Z}}$, due to the m- periodicity of $(x_k)_{k\in\mathbb{Z}}$.

➢ $\mathbf{M} \in \mathbb{R}^{n,n}$ is a *matrix with constant diagonals* & symmetric positive semi-definite (→ Def. 1.1.2.6)
("constant diagonals" ⇔ $(\mathbf{M})_{i,j}$ depends only on $i - j$)

⌟

EXAMPLE 4.5.1.3 (Linear regression for stationary Markov chains) We consider a sequence of scalar random variables: $(Y_k)_{k\in\mathbb{Z}}$, a so-called Markov chain. These can be thought of as values for a random quantity sampled at equidistant points in time.

We assume *stationary* (time-independent) correlations, that is, with $(\mathcal{A}, \Omega, dP)$ denoting the underlying probability space,

$$\mathcal{E}(Y_{i-j}Y_{i-k}) = \int_\Omega Y_{i-j}(\omega)Y_{i-k}(\omega)\,\mathrm{d}\mathbb{P}(\omega) = u_{k-j} \quad \forall i,j,k \in \mathbb{Z}\,, \quad u_i = u_{-i}\,.$$

Here $\mathcal{E}$ stands for the expectation of a random variable.

Model: We expect a finite linear dependency of the form

$$\exists \mathbf{x} = [x_1,\ldots,x_n]^\top \in \mathbb{R}^n\colon \quad Y_k = \sum_{j=1}^n x_j Y_{k-j} \quad \forall k \in \mathbb{Z}\,.$$

with *unknown* parameters x_j, $j = 1,\ldots,n$. Our task is to estimate the parameters $x_1,\ldots,x_n$ based on the *known correlations* u_ℓ. We try to minimize the expectation of the square of the expectation of the residual. This means that for some fixed $i \in \mathbb{Z}$ we use the

$$\text{estimator:} \qquad \mathbf{x} = \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^n} \mathcal{E}\left|Y_i - \sum_{j=1}^n x_j Y_{i-j}\right|^2\,. \tag{4.5.1.4}$$

The trick ist to use the linearity of the expectation, which makes it possible to convert (4.5.1.4) into

$$\mathbf{x} = [x_1,\ldots,x_n]^\top \in \mathbb{R}^n\colon \qquad E|Y_i|^2 - 2\sum_{j=1}^n x_j u_k + \sum_{k,j=1}^n x_k x_j u_{k-j} \to \min\,.$$

$$\blacktriangleright \qquad \mathbf{x}^\top\mathbf{A}\mathbf{x} - 2\mathbf{b}^\top\mathbf{x} \to \min \quad \text{with} \quad \mathbf{b} = [u_k]_{k=1}^n\,, \quad \mathbf{A} = \left[u_{i-j}\right]_{i,j=1}^n\,. \tag{4.5.1.5}$$

By definition $\mathbf{A}$ is a so-called covariance matrix and, as such, has to be symmetric and positive definite ($\to$ Def. 1.1.2.6). By its very definition it has constant diagonals. Also note that

$$\mathbf{x}^\top\mathbf{A}\mathbf{x} - 2\mathbf{b}^\top\mathbf{x} = (\mathbf{x}-\mathbf{x}^*)^\top\mathbf{A}(\mathbf{x}-\mathbf{x}^*) - \mathbf{x}^*\mathbf{A}\mathbf{x}^*\,, \tag{4.5.1.6}$$

with $\mathbf{x}^* = \mathbf{A}^{-1}\mathbf{b}$. Therefore $\mathbf{x}^*$ is the unique minimizer of $\mathbf{x}^\top\mathbf{A}\mathbf{x} - 2\mathbf{b}^\top\mathbf{x}$. The problem is reduced to solving the linear system of equations $\mathbf{A}\mathbf{x} = \mathbf{b}$ (Yule-Walker-equation, see below). ⌟

Matrices with constant diagonals occur frequently in mathematical models, see Ex. 4.5.1.1, Ex. 4.5.1.3. They generalize circulant matrices ($\to$ Def. 4.1.4.12).

Definition 4.5.1.7. Toeplitz matrix

$\mathbf{T} = (t_{ij})_{i,j=1}^n \in \mathbb{K}^{m,n}$ is a Toeplitz matrix, if there is a vector $\mathbf{u} = [u_{-m+1},\ldots,u_{n-1}] \in \mathbb{K}^{m+n-1}$ such that $t_{ij} = u_{j-i}$, $1 \le i \le m$, $1 \le j \le n$.

$$\mathbf{T} = \begin{bmatrix} u_0 & u_1 & \cdots & & \cdots & u_{n-1} \\ u_{-1} & u_0 & u_1 & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & u_1 \\ u_{1-m} & \cdots & & \cdots & u_{-1} & u_0 \end{bmatrix}$$

Note: The "information content" of a matrix $\mathbf{M} \in \mathbb{K}^{m,n}$ with constant diagonals, that is, $(\mathbf{M})_{i,j} = m_{i-j}$, is $m+n-1$ numbers $\in \mathbb{K}$.
Hence, though potentially densely populated, $m \times n$ Toeplitz matrices are data-sparse with information content $\ll mn$.

4.5.2 Toeplitz Matrix Arithmetic

Given: $\mathbf{T} = [u_{j-i}] \in \mathbb{K}^{m,n}$, a Toeplitz matrix with generating vector $\mathbf{u} = [u_{-m+1}, \ldots, u_{n-1}]^\top \in \mathbb{K}^{m+n-1}$, see Def. 4.5.1.7.

Task: Efficient evaluation of matrix×vector product $\mathbf{Tx}$, $\mathbf{x} \in \mathbb{K}^n$

To motivate the approach we realize that we have already encountered Toeplitz matrices in the convolution of finite signals discussed in Rem. 4.1.3.1, see (4.1.3.2). The trick introduced in Rem. 4.1.4.15 was to extend the matrix to a circulant matrix by zero padding, compare (4.1.4.18).

Idea: Extend $\mathbf{T} \in \mathbb{K}^{m,n}$ to a circulant matrix (→ Def. 4.1.4.12) $\mathbf{C} \in \mathbb{K}^{m+n,m+n}$ generated by the $m+n$-periodic sequence $(c_j)_{j\in\mathbb{Z}}$ given by

$$c_j = \begin{cases} u_j & \text{for } j = -m+1, \ldots, n-1\,, \\ 0 & \text{for } j = n\,, \end{cases} \quad + \text{ periodic extension.}$$

The upper left $m \times n$ block of $\mathbf{C}$ contains $\mathbf{T}$:

$$(\mathbf{C})_{ij} = c_{i-j}\,, \quad 1 \le i,j \le m+n \quad \Rightarrow \quad (\mathbf{C})_{1:m,1:n} = \mathbf{T}\,. \tag{4.5.2.1}$$

The following formula demonstrates the structure of $\mathbf{C}$ in the case $m = n$.

$$\mathbf{C} = \left[\begin{array}{ccccccc|ccccccc} u_0 & u_1 & \cdots & & \cdots & u_{n-1} & & 0 & u_{1-n} & \cdots & & \cdots & u_{-1} \\ u_{-1} & u_0 & u_1 & & & \vdots & & u_{n-1} & 0 & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \vdots & & \vdots & \ddots & \ddots & & & \\ \vdots & & \ddots & \ddots & \ddots & \vdots & & & & & & \ddots & \\ \vdots & & & \ddots & \ddots & u_1 & & \vdots & & & \ddots & \ddots & u_{1-n} \\ u_{1-n} & \cdots & & \cdots & u_{-1} & u_0 & & u_1 & & & & u_{n-1} & 0 \\ \hline 0 & u_{1-n} & \cdots & & \cdots & u_{-1} & & u_0 & u_1 & \cdots & & \cdots & u_{n-1} \\ u_{n-1} & 0 & \ddots & & & \vdots & & u_{-1} & u_0 & u_1 & & & \vdots \\ \vdots & \ddots & \ddots & & & & & \vdots & \ddots & \ddots & \ddots & & \vdots \\ & & & & \ddots & & & \vdots & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & u_{1-n} & & \vdots & & & \ddots & \ddots & u_1 \\ u_1 & & & & u_{n-1} & 0 & & u_{1-n} & \cdots & & \cdots & u_{-1} & u_0 \end{array}\right]$$

Recall from 4.3 that the multiplication with a circulant $(m+n)\times(m+n)$-matrix (= discrete periodic convolution → Def. 4.1.4.7) can be carried out by means of DFT/FFT with an asymptotic computational effort of $O((m+n)\log(m+n))$ for $m,n \to \infty$, see Code 4.2.2.4.

From (4.5.2.1) it is clear how to implement matrix×vector for the Toeplitz matrix $\mathbf{T}$

$$\text{zero padding} \longrightarrow \quad \mathbf{C}\begin{bmatrix}\mathbf{x}\\ 0\end{bmatrix} = \begin{bmatrix}\mathbf{Tx}\\ *\end{bmatrix}$$

Therefore the asymptotic computational effort for computing $\mathbf{Tx}$ is $O((n+m)\log(m+n))$ for $m,n \to \infty$, provided that an FFT-based algorithm for discrete periodic convolution is used, see Code 4.2.2.4. This complexity is almost optimal in light of the data complexity $O(m+n)$ of the Toeplitz matrix.

4.5.3 The Levinson Algorithm

Given: Symmetric positive definite (s.p.d.) (→ Def. 1.1.2.6) Toeplitz matrix $\mathbf{T} = (u_{j-i})_{i,j=1}^{n} \in \mathbb{R}^{n,n}$ with generating vector $\mathbf{u} = [u_{-n+1}, \ldots, u_{n-1}] \in \mathbb{R}^{2n-1}$, $u_{-k} = u_k$.

$$\mathbf{T} = \begin{bmatrix} 1 & u_1 & u_2 & \ldots & \ldots & u_{n-1} \\ u_1 & 1 & u_1 & u_2 & & u_{n-2} \\ u_2 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & u_2 \\ u_{n-2} & & u_2 & u_1 & 1 & u_1 \\ u_{n-1} & u_{n-2} & \ldots & u_2 & u_1 & 1 \end{bmatrix}$$

Note that the symmetry of a Toeplitz matrix is induced by the property $u_{-k} = u_k$ of its generating vector. Without loss of generality we assume that $\mathbf{T}$ has unit diagonal, $u_0 = 1$.

Task: Find an efficient solution algorithm for the LSE $\mathbf{T}\mathbf{x} = \mathbf{b} = [b_1, \ldots, b_n]^\top$, $\mathbf{b} \in \mathbb{R}^n$, the Yule-Walker problem from 4.5.1.3.

Employ a *recursive* (inductive) solution strategy.

Define:
- $\mathbf{T}_k := \left[u_{j-i}\right]_{i,j=1}^{k} \in \mathbb{K}^{k,k}$ (left upper block of $\mathbf{T}$) ➢ $\mathbf{T}_k$ is s.p.d. Toeplitz matrix,
- $\mathbf{x}^k \in \mathbb{K}^k$: $\mathbf{T}_k\mathbf{x}^k = \mathbf{b}^k := [b_1, \ldots, b_k]^\top \Leftrightarrow \mathbf{x}^k = \mathbf{T}_k^{-1}\mathbf{b}^k$,
- $\mathbf{u}^k := [u_1, \ldots, u_k]^\top \in \mathbb{R}^k$

We block partition the linear system of equations $\mathbf{T}_{k+1}\mathbf{x}^{k+1} = \mathbf{b}^{k+1}$, $k < n$:

$$\mathbf{T}_{k+1}\mathbf{x}^{k+1} = \left[\begin{array}{ccc|c} & \mathbf{T}_k & & \begin{matrix} u_k \\ \vdots \\ u_1 \end{matrix} \\ \hline u_k & \cdots & u_1 & 1 \end{array}\right] \left[\begin{array}{c} \widetilde{\mathbf{x}}^{k+1} \\ \hline x_{k+1}^{k+1} \end{array}\right] = \left[\begin{array}{c} b_1 \\ \vdots \\ b_k \\ \hline b_{k+1} \end{array}\right] = \begin{bmatrix} \mathbf{b}^k \\ b_{k+1} \end{bmatrix} \tag{4.5.3.1}$$

Now recall block Gaussian elimination/block-LU decomposition from Rem. 2.3.1.14, Rem. 2.3.2.19. For a regular matrix $\mathbf{A} \in \mathbb{R}^{k,k}$, $\mathbf{c} \in \mathbb{R}^k$, $\widetilde{\mathbf{b}} \in \mathbb{R}^k$, $\beta \in \mathbb{R}$, we obtain (see also (2.6.0.7) and (2.6.0.8))

$$\begin{bmatrix} \mathbf{A} & \mathbf{c} \\ \mathbf{c}^\top & 1 \end{bmatrix} \begin{bmatrix} \widetilde{\mathbf{x}} \\ \xi \end{bmatrix} = \begin{bmatrix} \widetilde{\mathbf{b}} \\ \beta \end{bmatrix} \quad \Rightarrow \quad \begin{aligned} \xi &= \frac{\beta - \mathbf{c}^\top \mathbf{A}^{-1}\widetilde{\mathbf{b}}}{1 - \mathbf{c}^\top \mathbf{A}^{-1}\mathbf{c}}\,, \\ \widetilde{\mathbf{x}} &= \mathbf{A}^{-1}\left(\widetilde{\mathbf{b}} - \xi\mathbf{c}\right)\,, \end{aligned} \tag{4.5.3.2}$$

provided that $1 - \mathbf{c}^\top \mathbf{A}^{-1}\mathbf{c} \neq 0$. This condition is satisfied, if the coefficient matrix from (4.5.3.2) is s.p.d., see (2.8.0.12). The formulas (4.5.3.2) applied to (4.5.3.1) teach us how to eliminate $\widetilde{\mathbf{x}}^{k+1}$ and obtain an expression for x_{k+1}^{k+1}.

To state the formulas concisely, we introduce reversing permutations. For a vector they can be realized by EIGEN's `reverse()` method.

$$\mathsf{P}_k : \{1, \ldots, k\} \mapsto \{1, \ldots, k\} \quad , \quad \mathsf{P}_k(i) := k - i + 1\,. \tag{4.5.3.3}$$

Then we carry out block elimination using $\mathbf{T}_k\mathbf{x}_k = \mathbf{b}^k$:

$$\begin{cases} \widetilde{\mathbf{x}}_{k+1} = \mathbf{T}_k^{-1}(\mathbf{b}^k - x_{k+1}^{k+1}\mathsf{P}_k\mathbf{u}^k) = \mathbf{x}^k - x_{k+1}^{k+1}\mathbf{T}_k^{-1}\mathsf{P}_k\mathbf{u}^k\,, \\ x_{k+1}^{k+1} = b_{k+1} - \mathsf{P}_k\mathbf{u}^k \cdot \widetilde{\mathbf{x}}^{k+1} = b_{k+1} - \left(\mathsf{P}_k\mathbf{u}^k\right)^\top \mathbf{x}^k + x_{k+1}^{k+1}\left(\mathsf{P}_k\mathbf{u}^k\right)^\top \mathbf{T}_k^{-1}\mathsf{P}_k\mathbf{u}^k\,. \end{cases} \tag{4.5.3.4}$$

The recursive idea is clear after introducing the *auxiliary vectors* $\mathbf{y}^k := \mathbf{T}_k^{-1}\mathsf{P}_k\mathbf{u}^k$, which converts (4.5.3.4) into

$$\mathbf{x}^{k+1} = \begin{bmatrix} \widetilde{\mathbf{x}}^{k+1} \\ x_{k+1}^{k+1} \end{bmatrix} \quad \text{with} \quad \begin{aligned} x_{k+1}^{k+1} &= (b_{k+1} - \left(\mathsf{P}_k\mathbf{u}^k\right)^\top \mathbf{x}^k)/\sigma_k \\ \widetilde{\mathbf{x}}^{k+1} &= \mathbf{x}^k - x_{k+1}^{k+1}\mathbf{y}^k \end{aligned} \quad , \quad \sigma_k := 1 - \left(\mathsf{P}_k\mathbf{u}^k\right)^\top \mathbf{y}^k \, . \tag{4.5.3.5}$$

Here, the assumption that $\mathbf{T}$ is s.p.d. ensures $\sigma_k \geq 1$.

Therefore, solving $\mathbf{T}_{k+1}\mathbf{x}^{k+1} = \mathbf{b}^{k+1}$ seems to entail the following steps:

❶ Solve $\mathbf{T}_k\mathbf{y}^k = \mathsf{P}_k\mathbf{u}^k$ ($k \times k$ s.p.d. Toeplitz LSE, recursion).

❷ Solve $\mathbf{T}_k\mathbf{x}^k = \mathbf{b}^k$ ($k \times k$ s.p.d. Toeplitz LSE, recursion).

❸ Compute $\mathbf{x}^{k+1}$ according to (4.5.3.5).

§4.5.3.6 (asymptotic complexity) Obviously, given $\mathbf{x}^k$ and $\mathbf{y}^k$, the evaluations involved in (4.5.3.5) take $O(k)$ operations for $k \to \infty$, in order to get $\mathbf{x}^{k+1}$.

It seems that two recursive calls are necessary in order to obtain $\mathbf{y}^k$ and $\mathbf{x}^k$, which enter (4.5.3.5): this is too expensive!

If $\mathbf{b}^k = \mathsf{P}_k\mathbf{u}^k$, then $\mathbf{x}^k = \mathbf{y}^k$

Simple linear recursion sufficient to compute $\mathbf{y}^k$

Hence, $\mathbf{y}^k$ can be computed with an asymptotic cost of $O(k^2)$ for $k \to \infty$. Once the $\mathbf{y}^k$ are available, another simple linear recursion gives us $\mathbf{x}^k$ with a cost of $O(k^2)$ for $k \to \infty$.

Cost for solving $\mathbf{Tx} = \mathbf{b}$ = $O(k^2)$ for $k \to \infty$. ⌟

Below we give a C++ implementation of the Levinson algorithm for the solution of the Yule-Walker problem $\mathbf{Tx} = \mathbf{b}$ with an s.p.d. Toeplitz matrix described by its generating vector $\mathbf{u}$ (recursive implementation, $\mathbf{x}^k$, $\mathbf{y}^k$ computed simultaneously, u_{n+1} not used!)

C++ code 4.5.3.7: Levinson algorithm ↪ GITLAB

```
2  void levinson(const Eigen::VectorXd &u, const Eigen::VectorXd &b,
3                Eigen::VectorXd &x, Eigen::VectorXd &y) {
4    const Eigen::Index k = u.size() - 1; // Matrix size - 1
5    // Trivial case of 1×1 linear sysrtem
6    if (k == 0) {
7      x.resize(1); x(0) = b(0);
8      y.resize(1); y(0) = u(0);
9      return;
10   }
11   // Vectors holding result of recursive call
12   Eigen::VectorXd xk;
13   Eigen::VectorXd yk;
14   // Recursive call for computing x^k and y^k
15   levinson(u.head(k), b.head(k), xk, yk);
16   // Coefficient σ_k from (4.5.3.5)
17   const double sigma = 1 - u.head(k).dot(yk);
18   // Update of x according to (4.5.3.5)
19   const double t = (b(k) - u.head(k).reverse().dot(xk)) / sigma;
20   x = xk - t * yk.head(k).reverse();
21   x.conservativeResize(x.size() + 1);
22   x(x.size() - 1) = t;
```

```
23    // Update of vectors y^k
24    const double s = (u(k) - u.head(k).reverse().dot(yk)) / sigma;
25    y = yk - s * yk.head(k).reverse();
26    y.conservativeResize(y.size() + 1);
27    y(y.size() - 1) = s;
28  }
```

Note that this implementation of the Levinson algorithm employs a simple linear recursion with computational cost $\sim (n-k)$ on level k, $k = 0, \ldots, n-1$, which results in an overall asymptotic complexity of $O(n^2)$ for $n \to \infty$, as already discussed in § 4.5.3.6.

Remark 4.5.3.8 (Fast Toeplitz solvers) Meanwhile researchers have found better methods [Ste03]: now there are FFT-based algorithms for solving $\mathbf{T}\mathbf{x} = \mathbf{b}$, $\mathbf{T}$ a Toeplitz matrix, with asymptotic complexity $O(n \log^3 n)$**!** ⌟

Supplementary literature. [DR08, Sect. 8.5]: Very detailed and elementary presentation, but the discrete Fourier transform through trigonometric interpolation, which is not covered in this chapter. Hardly addresses discrete convolution.

[Han02, Ch. IX] presents the topic from a mathematical point of view stressing approximation and trigonometric interpolation. Good reference for algorithms for circulant and Toeplitz matrices.

[Sau06, Ch. 10] also discusses the discrete Fourier transform with emphasis on interpolation and (least squares) approximation. The presentation of signal processing differs from that of the course.

There is a vast number of books and survey papers dedicated to discrete Fourier transforms, see, for instance, [Bri88; DV90]. Issues and technical details way beyond the scope of the course are discussed in these monographs.

Review question(s) 4.5.3.9 (Toeplitz matrix techniques)

(Q4.5.3.9.A) Give an example of a Toeplitz matrix $\mathbf{T} \in \mathbb{R}^{n,n}$, $n > 2$, with $\operatorname{rank}(\mathbf{T}) = 1$.

(Q4.5.3.9.B) Show that the product of two lower triangular Toeplitz matrices is a Toeplitz matrix again.

△

Bibliography

[Akv+25] M. Akveld, A. Iozzi, P. Jossen, and F. da Lio. “Komplexe Analysis”. 2025.

[Bri88] E.O. Brigham. *The Fast Fourier Transform and Its Applications*. Englewood Cliffs, NJ: Prentice-Hall, 1988 (cit. on p. 382).

[DR08] W. Dahmen and A. Reusken. *Numerik für Ingenieure und Naturwissenschaftler*. Heidelberg: Springer, 2008 (cit. on pp. 366, 382).

[DV90] P. Duhamel and M. Vetterli. “Fast Fourier transforms: a tutorial review and a state of the art”. In: *Signal Processing* 19 (1990), pp. 259–299 (cit. on pp. 360, 382).

[FJ05] M. Frigo and S. G. Johnson. “The Design and Implementation of FFTW3”. In: *Proceedings of the IEEE* 93.2 (Feb. 2005), pp. 216–231. DOI: 10.1109/JPROC.2004.840301 (cit. on p. 366).

[Gut09] M.H. Gutknecht. *Lineare Algebra*. Lecture Notes. SAM, ETH Zürich, 2009 (cit. on p. 322).

[Han02] M. Hanke-Bourgeois. *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens*. Mathematische Leitfäden. Stuttgart: B.G. Teubner, 2002 (cit. on pp. 319, 321, 366, 368, 382).

[HR11] Georg Heinig and Karla Rost. “Fast algorithms for Toeplitz and Hankel matrices”. In: *Linear Algebra Appl.* 435.1 (2011), pp. 1–59.

[NS02] K. Nipp and D. Stoffer. *Lineare Algebra*. 5th ed. Zürich: vdf Hochschulverlag, 2002 (cit. on pp. 322, 356).

[QSS00] A. Quarteroni, R. Sacco, and F. Saleri. *Numerical mathematics*. Vol. 37. Texts in Applied Mathematics. New York: Springer, 2000 (cit. on pp. 349, 366).

[Rad68] C.M. Rader. “Discrete Fourier Transforms when the Number of Data Samples Is Prime”. In: *Proceedings of the IEEE* 56 (1968), pp. 1107–1108 (cit. on p. 364).

[Ros21] Andreas Rosowski. *On Fast Computation of a Circulant Matrix-Vector Product*. 2021 (cit. on p. 321).

[Sau06] T. Sauer. *Numerical analysis*. Boston: Addison Wesley, 2006 (cit. on p. 382).

[Ste03] M. Stewart. “A Superfast Toeplitz Solver with Improved Numerical Stability”. In: *SIAM J. Matrix Analysis Appl.* 25.3 (2003), pp. 669–693 (cit. on p. 382).

[Str99] Gilbert Strang. “The Discrete Cosine Transform”. In: *SIAM Review* 41.1 (1999), pp. 135–147. DOI: 10.1137/S0036144598336745 (cit. on p. 368).

[Str09] M. Struwe. *Analysis für Informatiker*. Lecture notes, ETH Zürich. 2009 (cit. on pp. 353, 355).

Chapter 5

Machine Learning of One-Dimensional Data (Data Interpolation and Data Fitting in 1D)

Contents

5.1 Abstract Interpolation (AI) . . . 384
5.2 Global Polynomial Interpolation . . . 392
5.2.1 Uni-Variate Polynomials . . . 392
5.2.2 Polynomial Interpolation: Theory . . . 394
5.2.3 Polynomial Interpolation: Algorithms . . . 398
5.2.4 Polynomial Interpolation: Sensitivity . . . 414
5.3 Shape-Preserving Interpolation . . . 418
5.3.1 Shape Properties of Functions and Data . . . 419
5.3.2 Piecewise Linear Interpolation . . . 421
5.3.3 Cubic Hermite Interpolation . . . 423
5.4 Splines . . . 430
5.4.1 Spline Function Spaces . . . 430
5.4.2 Cubic-Spline Interpolation . . . 431
5.4.3 Structural Properties of Cubic Spline Interpolants . . . 436
5.4.4 Shape Preserving Spline Interpolation . . . 440
5.5 Algorithms for Curve Design . . . 445
5.5.1 CAD Task: Curves from Control Points . . . 445
5.5.2 Bezier Curves . . . 448
5.5.3 Spline Curves . . . 452
5.6 Trigonometric Interpolation . . . 456
5.6.1 Trigonometric Polynomials . . . 456
5.6.2 Reduction to Lagrange Interpolation . . . 457
5.6.3 Equidistant Trigonometric Interpolation . . . 459
5.7 Least Squares Data Fitting . . . 464

5.1 Abstract Interpolation (AI)

The task of (one-dimensional, scalar) **data interpolation** (point interpolation) can be described as follows:

One-dimensional interpolation

Given: data points (t_i, y_i), $i = 0, \ldots, n$, $n \in \mathbb{N}$, $t_i \in I \subset \mathbb{R}$, $y_i \in \mathbb{R}$

Objective: *Reconstruction of a function* $f : I \mapsto \mathbb{R}$

- satisfying the $n+1$ **interpolation conditions** (IC)

$$f(t_i) = y_i, \quad i = 0, \ldots, n \,. \tag{5.1.0.2}$$

- and belonging to a set V of eligible functions.

The function f we find is called the **interpolant** of the given data set $\{(t_i, y_i)\}_{i=0}^{n}$.

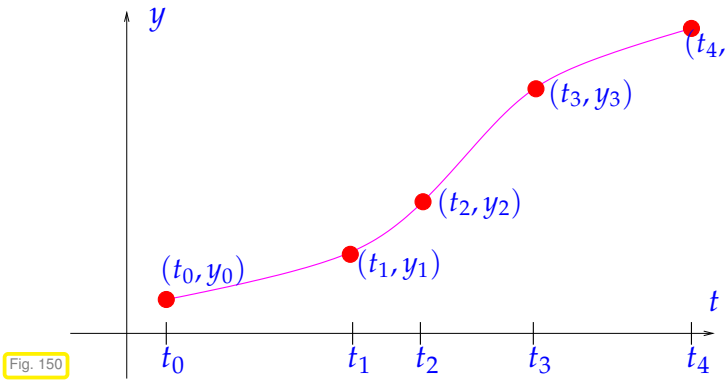


Fig. 150

Parlance: The numbers $t_i \in \mathbb{R}$ are called **nodes**, the $y_i \in \mathbb{R}$ are the (data) **values**.

◁ • $\hat{=}$ data points $(t_i, y_i) \in \mathbb{R}^2$

$t_i \hat{=}$ nodes

$y_i \hat{=}$ values

Graph of interpolant f passes through data points.

Of course, a necessary requirement on the data is that the t_i are pairwise distinct :

$$t_i \neq t_i \text{ , if } \quad i \neq j \quad \text{for all} \quad i, j \in \{0, \ldots, n\} \,.$$

Remark 5.1.0.3 (Generalization of data) In (supervised) machine learning this task is called the **generalization** of the data, because we aim for the creation of a model in the form of the function $f : I \to \mathbb{R}$ that permits us to generate new data points based on what we have "learned" from the provided data. ⌟

For ease of presentation we will usually assume that the nodes are ordered: $t_0 < t_1 < \cdots < t_n$ and $[t_0, t_n] \subset I$. However, algorithms often must not take for granted sorted nodes.

Remark 5.1.0.4 (Interpolation of vector-valued data) A natural generalization is data interpolation with vector-valued data values, seeking a function $\mathbf{f} : I \to \mathbb{R}^d$, $d \in \mathbb{N}$, such that, for given data points $(t_i, \mathbf{y}_i)$, $t_i \in I$ mutually different, $\mathbf{y}_i \in \mathbb{R}^d$, it satisfies the interpolation conditions $\mathbf{f}(t_i) = \mathbf{y}_i$, $i = 0, \ldots, n$.

In this case all methods available for scalar data can be applied component-wise.

An important application is curve reconstruction, that is the interpolation of points $\mathbf{y}_0, \ldots, \mathbf{y}_n \in \mathbb{R}^2$ in the plane.

A particular aspect of this problem is that the nodes t_i also have to be found, usually from the location of the $\mathbf{y}_i$ in a preprocessing step.

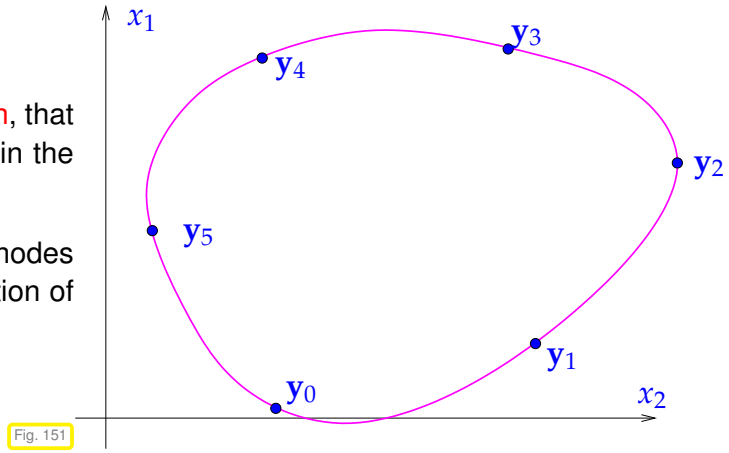


Fig. 151

⌟

Remark 5.1.0.5 (Multi-dimensional data interpolation) In many applications (computer graphics, computer vision, numerical method for partial differential equations, remote sensing, geodesy, etc.) one has to reconstruct functions of several variables.

This leads to task of multi-dimensional data interpolation:

Given:	data points $(\mathbf{x}_i, \mathbf{y}_i)$, $i = 0, \ldots, n$, $n \in \mathbb{N}$, $\mathbf{x}_i \in D \subset \mathbb{R}^m$, $m > 1$, $\mathbf{y}_i \in \mathbb{R}^d$
Objective:	*reconstruction of a (continuous) function* $\mathbf{f} : D \mapsto \mathbb{R}^d$ satisfying the $n+1$ interpolation conditions $\mathbf{f}(\mathbf{x}_i) = \mathbf{y}_i$, $i = 0, \ldots, n$.

Significant additional challenges arise in a genuine multidimensional setting. A treatment is beyond the scope of this course. However, the one-dimensional techniques presented in this chapter are relevant even for multi-dimensional data interpolation, if the points $\mathbf{x}_i \in \mathbb{R}^m$ are points of a finite lattice also called tensor product grid.

For instance, for $m = 2$ this is the case, if

$$\{\mathbf{x}_i\}_i = \left\{ [t_k, s_l]^\top \in \mathbb{R}^2 : k \in \{0, \ldots, K\},\ l \in \{0, \ldots, L\} \right\}, \tag{5.1.0.6}$$

where $t_k \in \mathbb{R}$, $k = 0, \ldots, K$, and s_l, $l = 0, \ldots, L$, $K, L \in \mathbb{N}$, are pairwise distinct nodes. ⌟

§5.1.0.7 (Interpolation schemes) When we talk about "interpolation schemes" in 1D, we mean a mapping

$$\mathrm{I} : \begin{cases} \mathbb{R}^{n+1} \times \mathbb{R}^{n+1} & \to \quad \{f : I \to \mathbb{R}\} \\ \left([t_i]_{i=0}^n, [y_i]_{i=0}^n\right) & \mapsto \quad \text{interpolant} \end{cases} .$$

Once the function space to which the interpolant belongs is specified, then an interpolation scheme defines an "interpolation problem" in the sense of § 1.5.5.1. Sometimes, only the data values y_i are consider input data, whereas the dependence of the interpolant on the nodes t_i is suppressed, see Section 5.2.4. ⌟

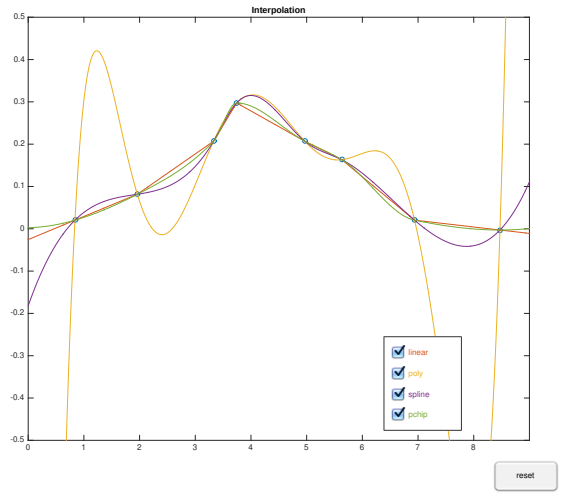


Fig. 152

◁ There are infinitely many ways to fix an interpolant for given data points.

Interpolants can have vastly different properties.

In this chapter we will discuss a few widely used methods to build interpolants and their different properties will become apparent.

We may (have to!) impose additional requirements on the interpolant:

- minimal smoothness of $\mathbf{f}$, e.g. $\mathbf{f} \in C^1$, etc.
- special shape of $\mathbf{f}$ (positivity, monotonicity, convexity → Section 5.3 below)

EXAMPLE 5.1.0.8 (Constitutive relations from measurements) This example addresses an important application of data interpolation in 1D.

In this context: t, y $\hat{=}$ two state variables of a physical system, where t determines y: a functional dependence $y = y(t)$ is assumed.

Examples: t and y could be

t	y
voltage U	current I
pressure p	density ρ
magnetic field H	magnetic flux B
...	...

Known: several *accurate* $(*)$ measurements

$$(t_i, y_i)\,, \quad i = 1, \ldots, m$$

Why do we need to extract the constitutive relations as a function? Imagine that t, y correspond to the voltage U and current I measured for a 2-port non-linear circuit element (like a diode). This element will be part of a circuit, which we want to simulate based on nodal analysis as in Ex. 8.1.0.1. In order to solve the resulting non-linear system of equations $F(\mathbf{u}) = 0$ for the nodal potentials (collected in the vector $\mathbf{u}$) by means of Newton's method ($\rightarrow$ Section 8.5) we need the voltage-current relationship for the circuit element as a continuously differentiable function $I = f(U)$.

$(*)$ Meaning of attribute "accurate": justification for interpolation. If measured values y_i were affected by considerable errors, one would not impose the interpolation conditions (5.1.0.2), but opt for data fitting ($\rightarrow$ Section 5.7). ⌟

We can distinguish two aspects of the interpolation problem:

❶ Find interpolant $f : I \subset \mathbb{R} \to \mathbb{R}$ and store/represent it (internally).

❷ Evaluate f at a few or many evaluation points $x \in I$

Remark 5.1.0.9 (Mathematical functions in a numerical code) What does it mean to "represent" or "make available" a function $f : I \subset \mathbb{R} \mapsto \mathbb{R}$ in a computer code?

! A general "mathematical" function $f : I \subset \mathbb{R} \mapsto \mathbb{R}^d$, I an interval, contains an "infinite amount of information".

Rather, in the context of numerical methods, "function" should be read as "subroutine", a piece of code that can, for any $x \in I$, compute $f(x)$ in finite time. Even this has to be qualified, because we can only pass machine numbers $x \in I \cap \mathbb{M}$ ($\rightarrow$ § 1.5.2.1) and, of course, in most cases, $f(x)$ will be an approximation. In a C++ code a simple real valued function can be incarnated through a function object of a type as given in Code 5.1.0.10, see also Section 0.3.3.

C++-code 5.1.0.10: C++ data type representing a real-valued function ↬ GITLAB

```
class Function {
  private:
    // various internal data describing f
  public:
    // Constructor:  expects information for specifying the function
    Function(/* ... */);
    // Evaluation operator
    double operator () (double t) const;
};
```

⌟

Remark 5.1.0.11 (A data type designed for interpolation problems) If a constitutive relationship for a circuit element is needed in a C++ simulation code (→ Ex. 5.1.0.8), the following specialized **Function** class could be used to represent it. It demonstrates the concrete object oriented implementation of an interpolant.

C++-code 5.1.0.12: C++ class representing an interpolant in 1D ↪ GITLAB

```
1
2  class Interpolant {
3    private:
4      // Various internal data describing f
5      // Can be the coefficients of a basis representation (5.1.0.14)
6    public:
7      // Constructor:  computation of coefficients c_j of representation
           (5.1.0.14)
8      Interpolant(const vector<double>& t, const vector<double>& y);
9      // Evaluation operator for interpolant f
10     double operator() (double t) const;
11 };
```

Two main components have to be designed and implemented:

- The constructor, which is in charge of "setup", e.g. building and solving a linear system of equations, see (5.1.0.23) below.
- The evaluation operator **operator** `()`, e.g., implemented as evaluation of a linear combination, refer to (5.1.0.14) below.

Crucial issue: computational effort for evaluation of interpolant at single point: $O(1)$ or $O(n)$ (or in between)?

⌟

§5.1.0.13 (Internal representation of classes of mathematical functions)

➡ Idea: parametrization, a finite number of parameters $c_0, \ldots, c_m$, $m \in \mathbb{N}$, characterizes f.

Special case: Representation with *finite linear combination* of **basis functions** $b_j : I \subset \mathbb{R} \mapsto \mathbb{R}$, $j = 0, \ldots, m$:

$$f = \sum_{j=0}^{m} c_j b_j \quad \Leftrightarrow \quad f(t) = \sum_{j=0}^{m} c_j b_j(t),\ t \in I \quad , \quad c_j \in \mathbb{R}^d \, . \tag{5.1.0.14}$$

↣ f belongs to a finite dimensional function space

$$V_m := \operatorname{Span}\{\{t \mapsto b_0(t)\}, \ldots, \{t \mapsto b_m(t)\}\} \quad [\ = \operatorname{Span}\{b_0, \ldots, b_m\}\ ] \, ,$$

with $\dim V_m = m + 1$, provided that $\{\{t \mapsto b_i(t)\}\}_{i=0}^{m}$ is linearly independent, which is already implied by the term "basis functions".

Of course, the basis functions b_j should be "simple" in the sense that $b_j(x)$ can be computed efficiently for every $x \in I$ and every $j = 0, \ldots, m$.

Note that the basis functions *may depend on the nodes* t_i, but they *must not depend on the values* y_i.

➙ The internal representation of f (in the data member section of the class **Function** from Code 5.1.0.10) will then boil down to storing the coefficients/parameters c_j, $j = 0, \ldots, m$.

Note: The focus in this chapter will be on the special case that the data interpolants belong to a finite-dimensional space of functions spanned by "simple" basis functions.

⌟

EXAMPLE 5.1.0.15 (Piecewise linear interpolation, see also Section 5.3.2) Recall: A linear function in 1D is a function of the form $x \mapsto a + bx$, $a, b \in \mathbb{R}$ (polynomial of degree 1).

Piecewise linear interpolation

= connect data points (t_i, y_i), $i = 0, \ldots, n$, $t_{i-1} < t_i$, by line segments

➢ interpolating polygonal line

Piecewise linear interpolant of data ▷

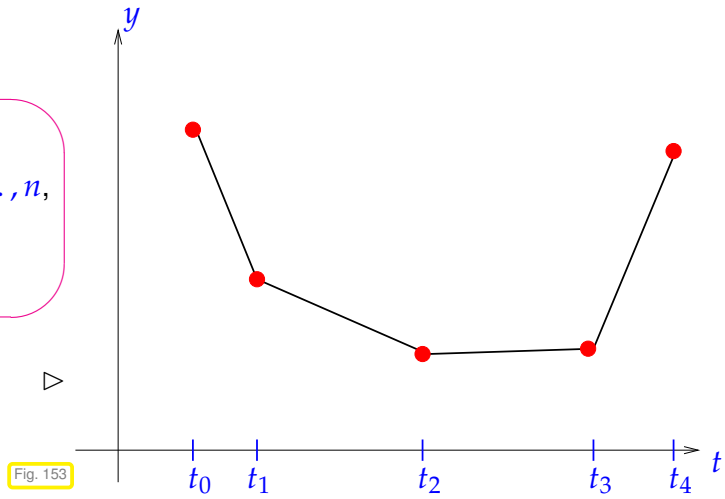


Fig. 153

What is the space V of functions from which we select the interpolant? Remember that a linear function $\mathbb{R} \to \mathbb{R}$ always can be written as $t \mapsto \alpha + \beta t$ with suitable coefficients $\alpha, \beta \in \mathbb{R}$. We can use this formula locally on every interval between two nodes. Assuming sorted nodes, $t_0 < t_1 < \cdots < t_n$, this leads to the mathematical definition

$$V := \left\{ f \in C^0(I) : \ f(t) = \alpha_i + \beta_i t \text{ for } t \in [t_i, t_{i+1}], \ i = 0, \ldots, n-1 \right\} . \tag{5.1.0.16}$$

Here, $C^0(I)$ designates the space of continuous functions $I \to \mathbb{R}$, $I := [t_0, t_n]$. Note that "$f \in C^0(I)$" is necessary to render (5.1.0.16) non-ambiguous.

Now, what could be a convenient set of basis functions $\{b_j\}_{j=0}^n$ for representing the piecewise linear interpolant through $n+1$ data points? A possible choice is the "Tent function" ("hat function") basis:

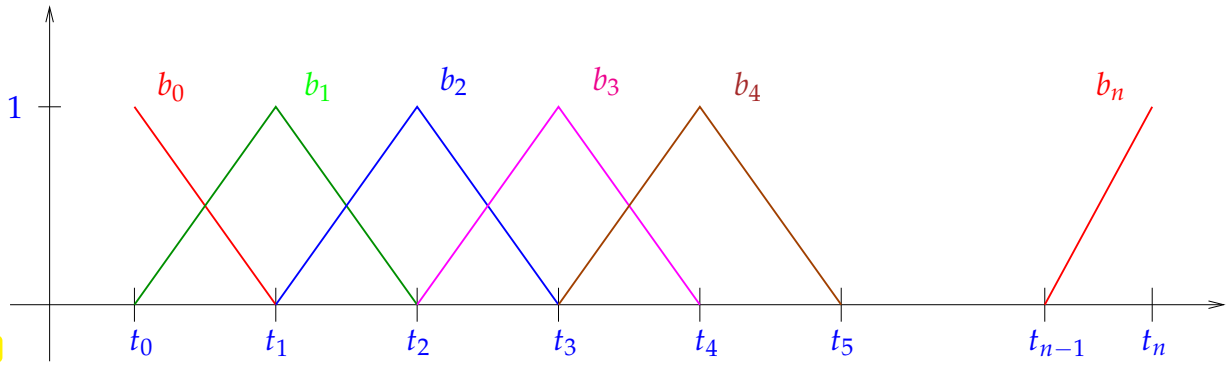


Fig. 154

Note: in Fig. 154 the basis functions have to be extended by zero outside the t-range where they are drawn.

Explicit formulas for these basis functions can be given and bear out that they are really "simple":

$$b_0(t) = \begin{cases} 1 - \frac{t-t_0}{t_1-t_0} & \text{for } t_0 \le t < t_1 \,, \\ 0 & \text{for } t \ge t_1 \,. \end{cases}$$

$$b_j(t) = \begin{cases} 1 - \frac{t_j - t}{t_j - t_{j-1}} & \text{for } t_{j-1} \le t < t_j \,, \\ 1 - \frac{t-t_j}{t_{j+1}-t_j} & \text{for } t_j \le t < t_{j+1} \,, \\ 0 & \text{elsewhere in } [t_0, t_n] \,. \end{cases} \quad , \quad j = 1, \ldots, n-1 \,, \tag{5.1.0.17}$$

$$b_n(t) = \begin{cases} 1 - \frac{t_n - t}{t_n - t_{n-1}} & \text{for } t_{n-1} \le t < t_n \,, \\ 0 & \text{for } t < t_{n-1} \,. \end{cases}$$

Moreover, these basis functions are *uniquely* determined by the conditions

- b_j is continuous on $[t_0, t_n]$,
- b_j is linear on each interval $[t_{i-1}, t_i]$, $i = 1, \ldots, n$,
- $b_j(t_i) = \delta_{ij} := \begin{cases} 1 & \text{, if } i = j \,, \\ 0 & \text{else.} \end{cases}$ ➢ a so-called cardinal basis for the node set $\{t_i\}_{i=0}^n$.

This last condition implies a simple basis representation of a (the ?) piecewise linear interpolant of the data points (t_i, y_i), $i = 0, \ldots, n$:

$$f(t) = \sum_{j=0}^{n} y_j b_j(t) \,, \quad t_0 \le t \le t_n \,, \tag{5.1.0.18}$$

where the b_j are given by (5.1.0.17). ⌟

The property $b_j(t_i) = \delta_{ij}$, $i, j = 1, \ldots, n$, of the tent function basis is so important that it has been given a special name:

Definition 5.1.0.19. Cardinal basis

A basis $\{b_0, \ldots, b_n\}$ of an $n+1$-dimensional vector space of functions $f : I \subset \mathbb{R} \to \mathbb{R}$ is a **cardinal basis** with respect to the set $\{t_0, \ldots, t_n\} \subset I$ of nodes, if

$$b_j(t_i) = \delta_{ij} := \begin{cases} 1 & \text{, if } i = j \,, \\ 0 & \text{else,} \end{cases} \quad i, j \in \{0, \ldots, n\} \,. \tag{5.1.0.20}$$

§5.1.0.21 (Interpolation as a linear mapping) We consider the setting for interpolation that the interpolant belongs to a finite-dimension space V_m of functions spanned by basis functions $b_0, \ldots, b_m$, see Rem. 5.1.0.9. Then the interpolation conditions imply that the basis expansion coefficients satisfy a linear system of equations:

$$\text{(5.1.0.2) \& (5.1.0.14)} \quad \Rightarrow \quad f(t_i) = \sum_{j=0}^{m} c_j b_j(t_i) = y_i \,, \quad i = 0, \ldots, n \,, \tag{5.1.0.22}$$

$$\Updownarrow$$

$$\mathbf{A}\mathbf{c} := \begin{bmatrix} b_0(t_0) & \ldots & b_m(t_0) \\ \vdots & & \vdots \\ b_0(t_n) & \ldots & b_m(t_n) \end{bmatrix} \begin{bmatrix} c_0 \\ \vdots \\ c_m \end{bmatrix} = \begin{bmatrix} y_0 \\ \vdots \\ y_n \end{bmatrix} =: \mathbf{y} \,. \tag{5.1.0.23}$$

This is an $(n+1)\times(m+1)$ linear system of equations !

The interpolation problem in V_m and the linear system (5.1.0.23) are really equivalent in the sense that (unique) solvability of one implies (unique) solvability of the other.

Necessary condition for unique solvability of interpolation problem (5.1.0.22) $\forall \mathbf{y}$: $m = n$

If $m = n$ and $\mathbf{A}$ from (5.1.0.23) regular ($\rightarrow$ Def. 2.2.1.1),then for *any* values y_j, $j = 0,\ldots,n$ we can find coefficients c_j, $j = 0,\ldots,n$, and, from them build the interpolant according to (5.1.0.14):

$$f = \sum_{j=0}^{n} (\mathbf{A}^{-1}\mathbf{y})_j b_j \,. \tag{5.1.0.24}$$

For *fixed* nodes t_i the interpolation problem (5.1.0.22) defines linear mapping

$$\mathsf{I} : \begin{cases} \mathbb{R}^{n+1} & \mapsto & V_n \\ \mathbf{y} & \mapsto & f \end{cases}$$

data space function space

Beware, "linear" in the statement above has nothing to do with a linear function or piecewise linear interpolation discussed in Ex. 5.1.0.15!

Definition 5.1.0.25. Linear interpolation operator

An interpolation operator $\mathsf{I} : \mathbb{R}^{n+1} \mapsto C^0([t_0,t_m])$ for the given nodes $t_0 < t_1 < \cdots < t_n$ is called linear, if

$$\mathsf{I}(\alpha\mathbf{y} + \beta\mathbf{z}) = \alpha\mathsf{I}(\mathbf{y}) + \beta\mathsf{I}(\mathbf{z}) \quad \forall \mathbf{y},\mathbf{z} \in \mathbb{R}^{n+1},\ \alpha,\beta \in \mathbb{R}\,. \tag{5.1.0.26}$$

✎ Notation: $C^0([t_0,t_m]) \hat{=}$ vector space of continuous functions on $[t_0,t_m]$ ⌟

Review question(s) 5.1.0.27 (Abstract Interpolation)

(Q5.1.0.27.A) Let $\{b_0,\ldots,b_n\}$ be a basis of a subspace V of the space $C^0(I)$ of continuous functions $I \subset \mathbb{R} \to \mathbb{R}$. Which linear systems has to be solved to determined the basis expansion coefficients for the interpolant $f \in V$ satisfying the interpolation conditions $f(t_i) = y_i$ for given node set $\{t_0,t_1,\ldots,t_n\}$ and values $y_i \in \mathbb{R}$?

How does reordering the nodes affect the coefficient matrix of that linear system?

(Q5.1.0.27.B) Given $I \subset \mathbb{R}$ and the node set $\{t_0,t_1,\ldots,t_n\} \subset I$, the ReLU basis of the space V of piecewise linear continuous functions on that node set is comprised of the functions

$$r_0(t) := 1 \quad , \quad r_i(t) := \begin{cases} 0 & \text{for} \quad t < t_{i-1}\,, \\ t - t_{i-1} & \text{for} \quad t \geq t_{i-1}\,, \end{cases} \quad i \in \{1,\ldots,n\}\,, \quad t \in I\,.$$

- Show that this set of functions $\{r_0,r_1,\ldots,r_n\}$ is really a basis of V.
- Assuming that the nodes are sorted, $t_0 < t_1 < \cdots < t_n$, describe the structure of the coefficient matrix of that linear system that has to be solved to determine the ReLU basis coefficients of an interpolant.

△

5.2 Global Polynomial Interpolation

(Global) polynomial interpolation, that is, interpolation into spaces of functions spanned by polynomials up to a certain degree, is the simplest interpolation scheme and of great importance as building block for more complex algorithms.

5.2.1 Uni-Variate Polynomials

Polynomials in a single variable are familiar and simple objects:

Notation: Vector space of the **(uni-variate) polynomials** of **degree** $\leq k$, $k \in \mathbb{N}$:

$$\mathcal{P}_k := \{t \mapsto \alpha_k t^k + \alpha_{k-1} t^{k-1} + \cdots + \alpha_1 t + \alpha_0 \cdot 1 \,,\, \alpha_j \in \mathbb{R}\} \,. \tag{5.2.1.1}$$

leading coefficient

Terminology: The functions $t \mapsto t^k$, $k \in \mathbb{N}_0$, are called monomials and the formula $t \mapsto \alpha_k t^k + \alpha_{k-1} t^{k-1} + \cdots + \alpha_0$ is the monomial representation of a polynomial.

Obviously, $\mathcal{P}_k$ is a vector space, see [NS02, Sect. 4.2, Bsp. 4]. What is its dimension?

Theorem 5.2.1.2. Dimension of space of polynomials

$$\dim \mathcal{P}_k = k+1 \quad \textit{and} \quad \mathcal{P}_k \subset C^\infty(\mathbb{R}).$$

In fact, $\mathcal{P}_k$ can be regarded as a finite-dimensional subspace of the space $C^0(\mathbb{R})$ of continuous functions $\mathbb{R} \to \mathbb{R}$. The monomial representation introduced above is a way to write a polynomials as a linear combination of the special basis functions $t \mapsto t^k$, see Rem. 5.1.0.9.

Proof. (of Thm. 5.2.1.2) Dimension formula by linear independence of monomials. □

As a consequence of Thm. 5.2.1.2 the monomial representation of a polynomial is unique.

§5.2.1.3 (The charms of polynomials) Why are polynomials important in computational mathematics?

- ➙ Easy to compute (only elementary operations required), integrate and differentiate
- ➙ Vector space & algebra
- ➙ Analysis: Taylor polynomials & power series ⌟

Remark 5.2.1.4 (Monomial representation) Polynomials (of degree k) in monomial representation are stored as a vector of their coefficients a_j, $j = 0, \ldots, k$. A convention for the ordering has to be fixed. For instance, the NUMPY module of PYTHON stores the coefficients of the monomial representation in an array in *descending order*:

PYTHON: $p(t) := \alpha_k t^k + \alpha_{k-1} t^{k-1} + \cdots + \alpha_0$ ➙ array $(\alpha_k, \alpha_{k-1}, \ldots, \alpha_0)$ (ordered!).

Thus the evaluation of a polynomial given through an array of monomial coefficients reads as:

```
In [8]: numpy.polyval([3,0,1], 5)   # 3*5^2 + 0*5^1 + 1
Out[8]: 76
```

⌟

§5.2.1.5 (Horner scheme → [DR08, Bem. 8.11]) *Efficient* evaluation of a polynomial in monomial representation through Horner scheme as indicated by the following representation:

$$p(t) = t(\cdots t(t(\alpha_n t + \alpha_{n-1}) + \alpha_{n-2}) + \cdots + \alpha_1) + \alpha_0 \,. \tag{5.2.1.6}$$

The following code gives an implementation based on vector data types of EIGEN. The function is vectorized in the sense that many evaluation points are processed in parallel.

C++-code 5.2.1.7: Horner scheme (vectorized version) ➙ GITLAB

```
// Efficient evaluation of a polynomial in monomial representation
// using the Horner scheme (5.2.1.6)
// IN: p = vector of monomial coefficients, length = degree + 1
// (leading coefficient in p(0), PYTHON convention Rem. 5.2.1.4)
// t = vector of evaluation points t_i
// OUT: vector of values: polynomial evaluated at t_i
Eigen::VectorXd horner(const Eigen::VectorXd &p, const Eigen::VectorXd &t) {
  const VectorXd::Index n = t.size();
  Eigen::VectorXd y{p[0] * VectorXd::Ones(n)};
  for (unsigned i = 1; i < p.size(); ++i)
    y = t.cwiseProduct(y) + p[i] * VectorXd::Ones(n);
  return y;
}
```

Optimal asymptotic complexity: $O(n)$

The Horner scheme is implemented in PYTHON's "built-in"-function `numpy.polyval(p,x)`. The argument `x` can be a matrix or a vector. In this case the function evaluates the polynomial described by `p` for each entry/component. Heed Rem. 5.2.1.4. ⌟

Review question(s) 5.2.1.8 (Polynomials)

(Q5.2.1.8.A) Why are polynomials the most widely used class of functions in numerical computations?

(Q5.2.1.8.B) What are the dimensions of the following two subspaces of the space $\mathcal{P}_k$ of polynomials of degree $\leq k$,

- $\mathcal{P}_k^{\text{even}} := \{p \in \mathcal{P}_k : \ p(t) = p(-t) \ \forall t \in \mathbb{R}\}$,
- $\mathcal{P}_k^{\text{odd}} := \{p \in \mathcal{P}_k : \ p(t) = -p(-t) \ \forall t \in \mathbb{R}\}$?

(Q5.2.1.8.C) For given $k \in \mathbb{N}$ we store the monomial coefficients of the polynomial $p(t) := \alpha_k t^k + \alpha_{k-1} t^{k-1} + \cdots + \alpha_0$ in a vector $\mathbf{a} := [\alpha_k, \ldots, \alpha_0] \in \mathbb{R}^{k+1}$. Find a matrix $\mathbf{D} \in \mathbb{R}^{k,k+1}$ such that $\mathbf{D}\mathbf{a} \in \mathbb{R}^k$ provides the monomial coefficients of the derivative p'.

(Q5.2.1.8.D) The mapping

$$\Phi : \begin{cases} \mathcal{P}_k & \to & \mathbb{R} \\ p & \mapsto & \int_0^1 p(t)\,\mathrm{d}t \end{cases}$$

is obviously linear and, therefore, has a matrix representation with respect to the monomial basis $\{t \mapsto 1, t \mapsto t, t \mapsto t^2, \ldots, t \mapsto t^k\}$ of $\mathcal{P}_k$. Find that matrix.

(Q5.2.1.8.E) A problem from linear algebra: Prove that the functions of the monomial basis

$$\left\{t \mapsto t^\ell\right\}_{\ell=0}^{n} \subset \mathcal{P}_n \,, \quad n \in \mathbb{N} \,,$$

are linearly independent and, thus, form a basis of $\mathcal{P}_n$.

Hint. Differentiate several times!

(Q5.2.1.8.F) The factorized representation of a polynomial $p \in \mathcal{P}_n$, $n \in \mathbb{N}$, writes it in the form

$$p(t) = \gamma_0(t-\gamma_1)\cdot\cdots\cdot(t-\gamma_n)\,, \quad \gamma_i \in \mathbb{R}\,, \quad i = 0,\ldots,n\,. \tag{5.2.1.9}$$

Somebody proposes to represent generic polynomials used in a numerical code in factorized form through the vectors $[\gamma_0,\gamma_1,\ldots,.\gamma_n] \in \mathbb{R}^{n+1}$ of coefficients. Discuss the pros and cons.

△

5.2.2 Polynomial Interpolation: Theory

Supplementary literature. This topic is also presented in [DR08, Sect. 8.2.1], [QSS00, Sect. 8.1], [AG11, Ch. 10].

Now we consider the interpolation problem introduced in Section 5.1 for the special case that the sought interpolant belongs to the polynomial space $\mathcal{P}_k$ (with suitable degree k).

***Lagrange polynomial interpolation problem* (LIP)**

Given the set of interpolation nodes $\{t_0,\ldots,t_n\} \subset \mathbb{R}$, $n \in \mathbb{N}$, and the values $y_0,\ldots,y_n \in \mathbb{R}$ compute $p \in \mathcal{P}_n$ such that it satisfies the interpolation conditions (IC)

$$p(t_j) = y_j \quad \text{for} \quad j = 0,\ldots,n\,. \tag{5.2.2.2}$$

Is this a well-defined problem? Obviously, it fits the framework developed in Rem. 5.1.0.9 and § 5.1.0.21, because $\mathcal{P}_n$ is a finite-dimensional space of functions, for which we already know a basis, the monomials. Thus, in principle, we could examine the matrix $\mathbf{A}$ from (5.1.0.23) to decide, whether the polynomial interpolant exists and is unique. However, there is a shorter way.

§5.2.2.3 (Lagrange polynomials) For a given set $\{t_0,t_1,\ldots,t_n\} \subset \mathbb{R}$ of nodes consider the

$$\text{Lagrange polynomials} \qquad L_i(t) := \prod_{\substack{j=0\\ j\neq i}}^{n} \frac{t-t_j}{t_i-t_j}\,, \quad i = 0,\ldots,n\,. \tag{5.2.2.4}$$

➙ Evidently, the Lagrange polynomials satisfy $L_i \in \mathcal{P}_n$ and $L_i(t_j) = \delta_{ij} := \begin{cases} 1 & \text{if } i = j\,,\\ 0 & \text{else.}\end{cases}$

From this relationship we infer that the Lagrange polynomials are linearly independent. Since there are $n+1 = \dim \mathcal{P}_n$ different Lagrange polynomials, we conclude that they form a basis of $\mathcal{P}_n$, which is a cardinal basis for the node set $\{t_i\}_{i=0}^n$. ⌟

EXAMPLE 5.2.2.5 (Lagrange polynomials for uniformly spaced nodes)

Consider the equidistant nodes in $[-1,1]$:

$$\mathcal{T} := \{t_j = -1 + \tfrac{2}{n}\, j\}\,, \quad j = 0,\ldots,n\,.$$

The plot shows the Lagrange polynomials for this set of nodes that do not vanish in the nodes t_0, t_2, and t_5, respectively.

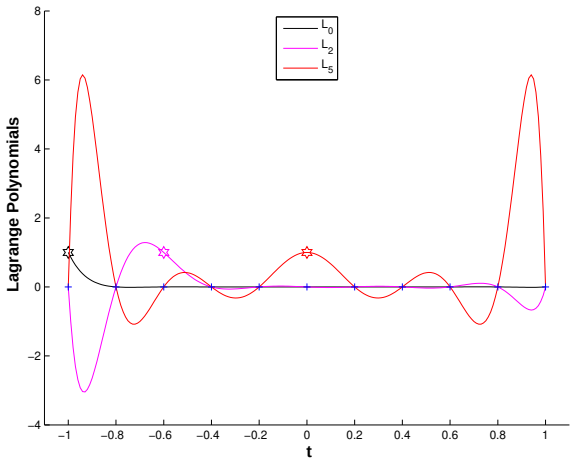


Fig. 155

The Lagrange polynomial interpolant p for data points $(t_i, y_i)_{i=0}^n$ allows a straightforward representation with respect to the basis of Lagrange polynomials for the node set $\{t_i\}_{i=0}^n$:

$$p(t) = \sum_{i=0}^{n} y_i L_i(t) \quad \Leftrightarrow \quad p \in \mathcal{P}_n \quad \text{and} \quad p(t_i) = y_i\,. \tag{5.2.2.6}$$

Theorem 5.2.2.7. Existence & uniqueness of Lagrange interpolation polynomial → **[QSS00, Thm. 8.1], [DR08, Satz 8.3]**

The general Lagrange polynomial interpolation problem admits a unique solution $p \in \mathcal{P}_n$ for any set of data points $\{(t_i, y_i)\}_{i=0}^n$, $n \in \mathbb{N}$, with pairwise distinct interpolation nodes $t_i \in \mathbb{R}$ ($i \neq j \Rightarrow t_i \neq t_j$).

Proof. Consider the *linear* evaluation operator

$$\mathrm{eval}_{\mathcal{T}} : \begin{cases} \mathcal{P}_n & \mapsto \mathbb{R}^{n+1}\,, \\ p & \mapsto (p(t_i))_{i=0}^n\,, \end{cases}$$

which maps between finite-dimensional vector spaces of the same dimension, see Thm. 5.2.1.2.

Representation (5.2.2.6) $\Rightarrow$ existence of interpolating polynomial
$\Rightarrow$ $\mathrm{eval}_{\mathcal{T}}$ is surjective ("onto")

Known from linear algebra: for a linear mapping $T : V \mapsto W$ between finite-dimensional vector spaces with $\dim V = \dim W$ holds the equivalence

$$T \text{ surjective} \quad \Leftrightarrow \quad T \text{ bijective} \quad \Leftrightarrow \quad T \text{ injective.}$$

Applying this equivalence to $\mathrm{eval}_{\mathcal{T}}$ yields the assertion of the theorem □

Corollary 5.2.2.8. Lagrange interpolation as linear mapping → **§ 5.1.0.21**

The polynomial interpolation in the nodes $\mathcal{T} := \{t_j\}_{j=0}^n$ defines a linear operator

$$\mathsf{I}_{\mathcal{T}} : \begin{cases} \mathbb{R}^{n+1} & \to \mathcal{P}_n\,, \\ (y_0,\ldots,y_n)^T & \mapsto \textit{interpolating polynomial } p\,. \end{cases} \tag{5.2.2.9}$$

Remark 5.2.2.10 (Vandermonde matrix) Lagrangian polynomial interpolation leads to linear systems of equations also for the representation coefficients of the polynomial interpolant in monomial basis, see § 5.1.0.21:

$$p(t_j) = y_j \iff \sum_{i=0}^{n} a_i t_j^i = y_j\,, \quad j = 0,\ldots,n$$
$$\iff \text{solution of } (n+1)\times(n+1) \text{ linear system } \mathbf{Va} = \mathbf{y} \text{ with matrix}$$

$$\mathbf{V} = \begin{bmatrix} 1 & t_0 & t_0^2 & \cdots & t_0^n \\ 1 & t_1 & t_1^2 & \cdots & t_1^n \\ 1 & t_2 & t_2^2 & \cdots & t_2^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_n & t_n^2 & \cdots & t_n^n \end{bmatrix} \in \mathbb{R}^{n+1,n+1}\,. \tag{5.2.2.11}$$

A matrix in the form of $\mathbf{V}$ is called Vandermonde matrix.

The following code initializes a Vandermonde matrix in EIGEN:

C++ code 5.2.2.12: Initialization of Vandermonde matrix ⇝ GITLAB

```
// Initialization of a Vandermonde matrix (5.2.2.11)
// from interpolation points t_i.
MatrixXd vander(const VectorXd &t) {
  const VectorXd::Index n = t.size();
  MatrixXd V(n,n); V.col(0) = VectorXd::Ones(n); V.col(1) = t;
  // Store componentwise integer powers of point coordinate vector
  // into the columns of the Vandermonde matrix
  for (int j=2;j<n;j++) V.col(j) = (t.array().pow(j)).matrix();
  return V;
}
```

⌟

Remark 5.2.2.13 (Matrix representation of interpolation operator) In the case of Lagrange interpolation:

- if Lagrange polynomials are chosen as basis for $\mathcal{P}_n$, then $\mathrm{I}_{\mathcal{T}}$ is represented by the identity matrix;
- if monomials are chosen as basis for $\mathcal{P}_n$, then $\mathrm{I}_{\mathcal{T}}$ is represented by the inverse of the Vandermonde matrix $\mathbf{V}$, see (5.2.2.11).

⌟

Remark 5.2.2.14 (Generalized polynomial interpolation → [DR08, Sect. 8.2.7], [QSS00, Sect. 8.4]) The following generalization of Lagrange interpolation is possible: We still seek a polynomial interpolant, but beside function values also prescribe derivatives up to a certain order for interpolating polynomial at given nodes.

Convention: indicate occurrence of derivatives as interpolation conditions by *multiple nodes*.

Generalized polynomial interpolation problem

Given the (possibly multiple) nodes $t_0, \ldots, t_n$, $n \in \mathbb{N}$, $-\infty < t_0 \le t_1 \le \cdots \le t_n < \infty$ and the values $y_0, \ldots, y_n \in \mathbb{R}$ compute $p \in \mathcal{P}_n$ such that

$$\frac{\mathrm{d}^k}{\mathrm{d}t^k} p(t_j) = y_j \quad \text{for} \quad k = 0, \ldots, \ell_j \quad \text{and} \quad j = 0, \ldots, n \,, \tag{5.2.2.15}$$

where $\ell_j := \max\{i - i' : t_j = t_i = t_{i'},\, i, i' = 0, \ldots, n\}$, and $\ell_j + 1$ is the multiplicity of the node t_j.

Admittedly, the statement of the generalized polynomial interpolation problem is hard to decipher. Let us look at a simple special case, which is also the most important case of generalized Lagrange interpolation. It is the case when all the multiplicities are equal to 2. It is called **Hermite interpolation** (or osculatory interpolation) and the generalized interpolation conditions read for nodes $t_0 = t_1 < t_2 = t_3 < \cdots < t_{n-1} = t_n$ (note the double nodes!) [QSS00, Ex. 8.6]:

$$p(t_{2j}) = y_{2j} \quad , \quad p'(t_{2j}) = y_{2j+1} \,, \quad j = 0, \ldots, n/2 \,.$$

Theorem 5.2.2.16. Existence & uniqueness of generalized Lagrange interpolation polynomials

The generalized polynomial interpolation problem (5.2.2.15) *admits a unique solution* $p \in \mathcal{P}_n$ *for any* $(t_i, y)i)$

Definition 5.2.2.17. Generalized Lagrange polynomials

The generalized Lagrange polynomials for the nodes $\mathcal{T} = \{t_j\}_{j=0}^n \subset \mathbb{R}$ (multiple nodes allowed) are defined as $L_i := \mathrm{I}_{\mathcal{T}}(\mathbf{e}_{i+1})$, $i = 0, \ldots, n$, where $\mathbf{e}_i = (0, \ldots, 0, 1, 0, \ldots, 0)^T \in \mathbb{R}^{n+1}$ are the unit vectors.

Note: The linear interpolation operator $\mathrm{I}_{\mathcal{T}}$ in this definition refers to generalized Lagrangian interpolation. Its existence is guaranteed by Thm. 5.2.2.16.

⌟

EXAMPLE 5.2.2.18 (Generalized Lagrange polynomials for Hermite Interpolation)

Consider the node set

$$\mathcal{T} = \{t_0 = 0, t_1 = 0, t_2 = 1, t_3 = 1\} \,.$$

The plot shows the four unique generalized Lagrange polynomials of degree $n = 3$ for these nodes. They satisfy

$$\begin{aligned} &p_0(0) = 1, p_0(1) = p_0'(0) = p_0'(1) = 0 \,, \\ &p_1(1) = 1, p_1(0) = p_1'(0) = p_1'(1) = 0 \,, \\ &p_2'(0) = 1, p_2(1) = p_2(0) = p_2'(1) = 0 \,, \\ &p_3'(1) = 1, p_3(1) = p_3(0) = p_3'(0) = 0 \,. \end{aligned}$$

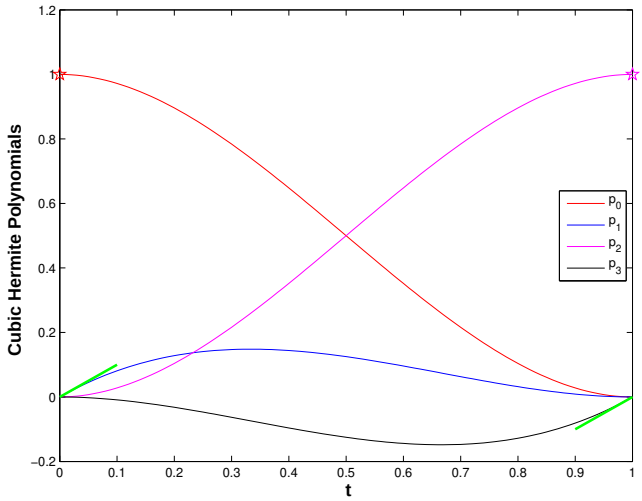


Fig. 156

More details are given in Section 5.3.3. For explicit formulas for the polynomials see (5.3.3.5). ⌟

Review question(s) 5.2.2.19 (Polynomial interpolation: theory)

(Q5.2.2.19.A) For a set $\{t_0, t_1, \ldots, t_n\} \subset \mathbb{R}$ of nodes the associated Lagrange polynomials are

$$L_i(t) := \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{t - t_j}{t_i - t_j}, \quad i = 0, \ldots, n .$$

Write down the Lagrange polynomials L_0, L_1, L_2, L_3 for the node set $\{0, 1, 2, 3\}$.

(Q5.2.2.19.B) Denote by L_i, $i = 0, \ldots, n$, $n \in \mathbb{N}$, the Lagrange polynomials for the node set $\{t_0, t_1, \ldots, t_n\} \subset \mathbb{R}$ that is assumed to be sorted $t_0 < t_1 < \cdots < t_n$.

What can you say about the sign of $p(t)$, where $p(t) = L_k(t)L_m(t)$, $t \in \mathbb{R}$?

(Q5.2.2.19.C) For a given node set $\{t_0, t_1, \ldots, t_n\} \subset \mathbb{R}$ the associated Vandermonde matrix reads

$$\mathbf{V} = \begin{bmatrix} 1 & t_0 & t_0^2 & \cdots & t_0^n \\ 1 & t_1 & t_1^2 & \cdots & t_1^n \\ 1 & t_2 & t_2^2 & \cdots & t_2^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_n & t_n^2 & \cdots & t_n^n \end{bmatrix} \in \mathbb{R}^{n+1,n+1} .$$

Sketch an efficient implementation of the C++ function

```
Eigen::VectorXd vanderMult(const Eigen::VectorXd &t,
const Eigen::VectorXd &x);
```

for computing $\mathbf{V}\mathbf{x}$ for some $\mathbf{x} \in \mathbb{R}^{n+1}$.

△

5.2.3 Polynomial Interpolation: Algorithms

Now we consider the algorithmic realization of Lagrange interpolation as introduced in Section 5.2.2. The goal is to achieve an efficient implementation of the class **Interpolant** introduced in Code 5.1.0.12 specialized for Lagrange polynomial interpolation. The setting is a follows:

Given: nodes $\mathcal{T} := \{-\infty < t_0 < t_1 < \ldots < t_n < \infty\}$,
values $\boldsymbol{y} := \{y_0, y_1, \ldots, y_n\}$,

Notation: we write $p := \mathrm{I}_{\mathcal{T}}(\boldsymbol{y})$ for the unique Lagrange polynomial interpolant, whose existence is asserted by Thm. 5.2.2.7.

When used in a numerical code, different demands can be made for a class that implements Lagrange interpolants. These demands determine, which algorithm is most suitable for the constructors and the evaluation operators.

5.2.3.1 Multiple evaluations

Task: For ❶ a *fixed* set $\{t_0, \ldots, t_n\}$ of nodes,
and ❷ *many* different given data values y_i, $i = 0, \ldots, n$
and ❸ *many* arguments x_k, $k = 1, \ldots, N$, $N \gg 1$,
efficiently compute all $p(x_k)$ for $p \in \mathcal{P}_n$ interpolating in (t_i, y_i), $i = 0, \ldots, n$.
See '

The definition of a possible interpolator data type could be as follows:

C++ code 5.2.3.1: Polynomial Interpolation ➽ GITLAB

```
class PolyInterp {
 public:
  // Constructors taking node vector [t_0,...,t_n]^T as argument
  PolyInterp(const Eigen::VectorXd &t);
  template <typename SeqContainer>
  PolyInterp(const SeqContainer &v);
  // Evaluation operator for data (y_0,...,y_n); computes
  // p(x_k) for x_ks passed in x
  Eigen::VectorXd eval(const Eigen::VectorXd &y,
                       const Eigen::VectorXd &x) const;
 private:
  // various internal data describing p
  Eigen::VectorXd _t;
};
```

The member function `eval(y,x)` expects n data values in `y` and (any number of) evaluation points in `x` ($\leftrightarrow [x_1,\ldots,x_N]^\top$) and returns the vector $[p(x_1),\ldots,p(x_N)]^\top$, where p is the Lagrange polynomial interpolant.

An implementation directly based on the evaluation of Lagrange polynomials (5.2.2.4) and (5.2.2.6) would incur an asymptotic computational effort of $O(n^2N)$ for every single invocation of `eval` and large n, N.

§5.2.3.2 (Barycentric interpolation formula)

By means of *pre-computing* parts of the Lagrange polynomials L_i the asymptotic effort for `eval` can be reduced substantially.

Simple manipulations starting from (5.2.2.6) give an altenative representation of p:

$$p(t) = \sum_{i=0}^{n} L_i(t)\, y_i = \sum_{i=0}^{n} \prod_{\substack{j=0\\ j\neq i}}^{n} \frac{t-t_j}{t_i-t_j}\, y_i = \sum_{i=0}^{n} \lambda_i \prod_{\substack{j=0\\ j\neq i}}^{n} (t-t_j)\, y_i = \prod_{j=0}^{n}(t-t_j) \cdot \sum_{i=0}^{n} \frac{\lambda_i}{t-t_i}\, y_i\,,$$

where $\lambda_i = \dfrac{1}{(t_i-t_0)\cdots(t_i-t_{i-1})(t_i-t_{i+1})\cdots(t_i-t_n)}$, $i=0,\ldots,n$: independent of y_i!

From the above formula, with $p(t)\equiv 1$, $y_i = 1$:

$$1 = \prod_{j=0}^{n}(t-t_j)\ \sum_{i=0}^{n} \frac{\lambda_i}{t-t_i} \quad \Rightarrow \quad \prod_{j=0}^{n}(t-t_j) = \frac{1}{\sum_{i=0}^{n} \frac{\lambda_i}{t-t_i}}$$

$$\blacktriangleright \quad \text{Barycentric interpolation formula} \quad p(t) = \frac{\displaystyle\sum_{i=0}^{n} \frac{\lambda_i}{t-t_i}\, y_i}{\displaystyle\sum_{i=0}^{n} \frac{\lambda_i}{t-t_i}}\,. \tag{5.2.3.3}$$

with $\lambda_i = \dfrac{1}{(t_i-t_0)\cdots(t_i-t_{i-1})(t_i-t_{i+1})\cdots(t_i-t_n)}$, $i=0,\ldots,n$, independent of t and y_i [Tre13, Thm. 5.1]. Hence, the values λ_i can be *precomputed*!

The use of (5.2.3.3) involves

- computation of weights λ_i, $i=0,\ldots,n$: cost $O(n^2)$ (only once!),

✦ cost $O(n)$ for every subsequent evaluation of p.

$\Rightarrow$ total asymptotic complexity $O(Nn) + O(n^2)$ ⌟

The following C++ class demonstrated the use of the barycentric interpolation formula for efficient multiple point evaluation of a Lagrange interpolation polynomial:

C++-code 5.2.3.4: Class for multiple data/multiple point evaluations ↪ GITLAB

```
2  template <typename NODESCALAR = double> class BarycPolyInterp {
3  private:
4    using nodeVec_t = Eigen::Matrix<NODESCALAR, Eigen::Dynamic, 1>;
5    using idx_t = typename nodeVec_t::Index;
6    // Number n of interpolation points, deg polynomial +1
7    const idx_t n;
8    // Locations of n interpolation points
9    nodeVec_t t;
10   // Precomputed values λ_i, i = 0,...,n-1
11   nodeVec_t lambda;
12
13 public:
14   // Constructors taking node vector [t_0,...,t_n]^T as
15   // argument
16   explicit BarycPolyInterp(const nodeVec_t &_t);
17   // The interpolation points may also be passed in an STL container
18   template <typename SeqContainer> explicit BarycPolyInterp(const SeqContainer &v);
19   // Computation of p(x_k) for data values
20   // (y_0,...,y_n) and evaluation points x_k
21   template <typename RESVEC, typename DATAVEC>
22   RESVEC eval(const DATAVEC &y, const nodeVec_t &x) const;
23
24 private:
25   void init_lambda();
26 };
```

C++-code 5.2.3.5: Interpolation class: constructors ↪ GITLAB

```
2  template <typename NODESCALAR>
3  BarycPolyInterp<NODESCALAR>::BarycPolyInterp(const nodeVec_t &_t)
4      : n(_t.size()), t(_t), lambda(n) {
5    init_lambda();
6  }
7
8  template <typename NODESCALAR>
9  template <typename SeqContainer>
10 BarycPolyInterp<NODESCALAR>::BarycPolyInterp(const SeqContainer &v)
11     : n(v.size()), t(n), lambda(n) {
12   idx_t ti = 0;
13   for (auto tp : v) {
14     t(ti++) = tp;
15   }
16   init_lambda();
17 }
```

C++-code 5.2.3.6: Interpolation class: precomputations ↪ GITLAB

```
2  template <typename NODESCALAR>
3  void BarycPolyInterp<NODESCALAR>::init_lambda() {
```

```
  // Precompute the weights λ_i with effort O(n^2)
  for (unsigned k = 0; k < n; ++k) {
    // little workaround:  in EIGEN cannot subtract a vector
    // from a scalar; multiply scalar by vector of ones
    lambda(k) =
        1. / ((t(k) * nodeVec_t::Ones(k) - t.head(k)).prod() *
              (t(k) * nodeVec_t::Ones(n - k - 1) - t.tail(n - k - 1)).prod());
  }
}
```

C++-code 5.2.3.7: Interpolation class: multiple point evaluations ↪ GITLAB

```
template <typename NODESCALAR>
template <typename RESVEC, typename DATAVEC>
[[nodiscard]] RESVEC BarycPolyInterp<NODESCALAR>::eval(const DATAVEC &y,
                                       const nodeVec_t &x) const {
  const idx_t N = x.size(); // No.  of evaluation points
  RESVEC p(N);              // Ouput vector
  // Compute quotient of weighted sums of λ_i/(t-t_i),
  // effort O(n)
  for (int i = 0; i < N; ++i) {
    const nodeVec_t z = (x[i] * nodeVec_t::Ones(n) - t);

    // Check if we want to evaluate close to a node
    const double tref{z.cwiseAbs().maxCoeff()}; // reference size
    idx_t k;
    if (z.cwiseAbs().minCoeff(&k) <
        tref * std::abs(std::numeric_limits<NODESCALAR>::epsilon())) {
      // evaluation at node t_k
      p[i] = y[k];
    } else {
      const nodeVec_t mu = lambda.cwiseQuotient(z);
      p[i] = (mu.cwiseProduct(y)).sum() / mu.sum();
    }
  } // end for
  return p;
}
```

Runtime measurements of direct evaluation of a polynomial in monomial representation vs. barycentric formula are reported in Exp. 5.2.3.13.

5.2.3.2 Single evaluation

Supplementary literature. This topic is also discussed in [DR08, Sect. 8.2.2].

Task: Given a set of interpolation points (t_j, y_j), $j = 0, \ldots, n$, with pairwise different interpolation nodes t_j, perform *a single* point evaluation of the Lagrange polynomial interpolant p at $x \in \mathbb{R}$.

We discuss the efficient implementation of the following function for $n \gg 1$. It is meant for a single evaluation of a Lagrange interpolant.

```
double eval(const Eigen::VectorXd &t,const Eigen::VectorXd &y,
double x);
```

§5.2.3.8 (Aitken-Neville scheme) The starting point is a recursion formula for partial Lagrange interpolants: For $0 \le k \le \ell \le n$ define

$$p_{k,\ell} := \text{unique interpolating polynomial of degree } \ell - k \text{ through } (t_k, y_k), \ldots, (t_\ell, y_\ell),$$

From the uniqueness of polynomial interpolants ($\rightarrow$ Thm. 5.2.2.7) we find

$$\begin{aligned} p_{k,k}(x) &\equiv y_k \quad \text{(“constant polynomial”)}\,, \quad k = 0, \ldots, n\,, \\ p_{k,\ell}(x) &= \frac{(x - t_k)p_{k+1,\ell}(x) - (x - t_\ell)p_{k,\ell-1}(x)}{t_\ell - t_k} \\ &= p_{k+1,\ell}(x) + \frac{x - t_\ell}{t_\ell - t_k}\left(p_{k+1,\ell}(x) - p_{k,\ell-1}(x)\right)\,, \quad 0 \le k < \ell \le n\,, \end{aligned} \tag{5.2.3.9}$$

because the left and right hand sides represent polynomials of degree $\ell - k$ through the points (t_j, y_j), $j = k, \ldots, \ell$:

$$\frac{(x - t_k)p_{k+1,\ell}(x) - (x - t_\ell)p_{k,\ell-1}(x)}{t_\ell - t_k} = \begin{cases} y_k & \text{for } x = t_k \quad [\, p_{k,\ell-1}(t_k) = y_k \,]\,, \\ y_j & \text{for } x = t_j\,, \quad k < j < \ell\,, \\ y_\ell & \text{for } x = t_\ell \quad [\, p_{k+1,\ell}(t_\ell) = y_\ell \,]\,. \end{cases}$$

Thus the values of the partial Lagrange interpolants can be computed sequentially and their dependencies can be expressed by the following so-called Aitken-Neville scheme:

$\ell - k =$	0	1	2	3
t_0	$y_0 =: p_{0,0}(x)$	$\nearrow p_{0,1}(x)$	$\nearrow p_{0,2}(x)$	$\nearrow p_{0,3}(x)$
t_1	$y_1 =: p_{1,1}(x)$	$\nearrow p_{1,2}(x)$	$\nearrow p_{1,3}(x)$	
t_2	$y_2 =: p_{2,2}(x)$	$\nearrow p_{2,3}(x)$		
t_3	$y_3 =: p_{3,3}(x)$			

(ANS)

Here, the arrows indicate contributions to the convex linear combinations of (5.2.3.9). The computation can advance from left to right, which is done in following C++ code.

C++-code 5.2.3.10: Aitken-Neville algorithm ↪ GITLAB

```
2  // Aitken-Neville algorithm for evaluation of interpolating polynomial
3  // IN: t, y:  (vectors of) interpolation data points
4  // x:  (single) evaluation point
5  // OUT: value of interpolant in x
6  double ANipoleval(const Eigen::VectorXd& t, Eigen::VectorXd y, const double x) {
7    for (int i = 1; i < y.size(); ++i) {
8      // Careful:  int loop index required for comparison >= 0 !
9      for (int k = i - 1; k >= 0; --k) {
10       // Recursion (5.2.3.9)
11       y[k] = y[k + 1] + (y[k + 1] - y[k])*(x - t[i])/(t[i] - t[k]);
12     }
13   }
14   return y[0];
15 }
```

The vector `y` contains the diagonals (from bottom left to top right) of the above triangular tableaux:

Note that the algorithm merely needs to store that single vector, which translates into $O(n)$ required memory for $n \to \infty$.

i	y[0]	y[1]	y[2]	y[3]
0	y_0	y_1	y_2	y_3
1	$p_{0,1}(x)$	y_1	y_2	y_3
2	$p_{0,2}(x)$	$p_{1,2}(x)$	y_2	y_3
3	$p_{0,3}(x)$	$p_{1,3}(x)$	$p_{2,3}(x)$	y_3

Asymptotic complexity of `ANipoeval` in terms of number of data points is $O(n^2)$ (two nested loops). This is the same as for evaluation based on the barycentric formula, but the Aitken-Neville has a key advantage discussed in the next §. ⌟

§5.2.3.11 (Polynomial interpolation with data updates) The Aitken-Neville algorithm has another interesting feature, when we run through the Aitken-Neville scheme from the top left corner:

$n =$	0	1	2	3
t_0	$y_0 =: p_{0,0}(x)$	$\to p_{0,1}(x)$	$\to p_{0,2}(x)$	$\to p_{0,3}(x)$
t_1	$y_1 =: p_{1,1}(x)$	$\to p_{1,2}(x)$	$\to p_{1,3}(x)$	
t_2	$y_2 =: p_{2,2}(x)$	$\to p_{2,3}(x)$		
t_3	$y_3 =: p_{3,3}(x)$			

Thus, the values of partial polynomial interpolants at x can be computed before all data points are even processed. This results in an "update-friendly" algorithm that can efficiently supply the point values $p_{0,k}(x)$, $k = 0, \ldots, n$, while being supplied with the data points (t_i, y_i). It can be used for the efficient implementation of the following interpolator class:

C++-code 5.2.3.12: Single point evaluation with data updates ↪ GITLAB

```
class PolyEval {
private:
  // evaluation point and various internal data describing the
     polynomials
public:
  // Constructor taking the evaluation point as argument
  PolyEval(double x);
  // Add another data point and update internal information
  void addPoint(double t, double y);
  // Value of current interpolating polynomial at x
  double eval(void) const;
};
```

⌟

EXPERIMENT 5.2.3.13 (Timing polynomial evaluations)

Timing for single-point evaluation

Comparison of the computational time needed for polynomial interpolation of

$$\{t_i = i\}_{i=1,\ldots,n}, \qquad \{y_i = \sqrt{i}\}_{i=1,\ldots,n},$$

$n = 3, \ldots, 200$, and evaluation in a single point $x \in [0, n]$.

Minimum computational time over 100 runs ➙

The measurements were carried out with the code `polytiming.cpp` ➙ GITLAB, gcc with -O3.

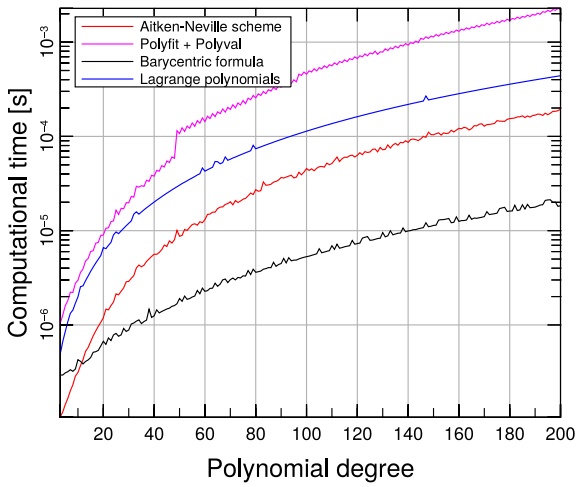


Fig. 157

This uses functions given in Code 5.2.3.7, Code 5.2.3.10 and the function `polyfit()` (with a clearly greater computational effort !). `polyfit()` is the equivalent to PYTHON's/MATLAB's built-in `polyfit`. The implementation can be found on GitLab.

⌟

Review question(s) 5.2.3.14 (Polynomial Interpolation: Algorithms)

(Q5.2.3.14.A) The Aitken-Neville scheme was introduced as

$n =$	0	1	2	3
t_0	$y_0 =: p_{0,0}(x)$	$\nearrow p_{0,1}(x)$	$\nearrow p_{0,2}(x)$	$\nearrow p_{0,3}(x)$
t_1	$y_1 =: p_{1,1}(x)$	$\nearrow p_{1,2}(x)$	$\nearrow p_{1,3}(x)$	
t_2	$y_2 =: p_{2,2}(x)$	$\nearrow p_{2,3}(x)$		
t_3	$y_3 =: p_{3,3}(x)$			

(ANS)

Give an interpretation of the quantities $p_{k,\ell}$ occurring in (ANS).

(Q5.2.3.14.B) Describe a scenario for the evaluation of degree-n Lagrange polynomial interpolants in a *single* point $x \in \mathbb{R}$ where the use of the barycentric interpolation formula

$$p(t) = \sum_{i=0}^{n} \frac{\lambda_i}{t - t_i} y_i \; : \; \sum_{i=0}^{n} \frac{\lambda_i}{t - t_i}, \quad \lambda_i := \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{1}{t_i - t_j}, \tag{5.2.3.3}$$

is more efficient than computations based on the Aitken-Neville scheme.

△

5.2.3.3 Extrapolation to Zero

Extrapolation is interpolation with the evaluation point t outside the interval $[\inf_{j=0,\ldots,n} t_j, \sup_{j=0,\ldots,n} t_j]$. In the sequel we assume $t = 0$, $t_i > 0$. Of course, Lagrangian polynomial interpolation can also be used for extrapolation. In this section we give a very important application of this "Lagrangian extrapolation".

Task: compute the limit $\lim_{h \to 0} \psi(h)$ with prescribed accuracy, though the evaluation of the function $\psi = \psi(h)$ (maybe given in procedural form only) for very small arguments $|h| \ll 1$ is difficult, usually because of numerically instability (→ Section 1.5.5).

The extrapolation technique introduced below works well, if

- ψ is an *even function* of its argument: $\psi(t) = \psi(-t)$,
- $\psi = \psi(h)$ behaves "nicely" around $h = 0$.

Theory: The analysis of extrapolation techniques usually relies on the existence of an asymptotic expansion in h^2

$$f(h) = f(0) + A_1 h^2 + A_2 h^4 + \cdots + A_n h^{2n} + R(h) \quad , \quad A_k \in \mathbb{R} \, ,$$

with remainder estimate $|R(h)| = O(h^{2n+2})$ for $h \to 0$.

Idea: approximating an inaccessible limit by extrapolation to zero

① Pick $h_0, \ldots, h_n$ for which ψ can be evaluated "safely".

② evaluation of $\psi(h_i)$ for different h_i, $i = 0, \ldots, n$, $|h_i| > 0$.

③ Appeoximate $\psi(0) \approx p(0)$, $p \triangleq$ interpolating polynomial $p \in \mathcal{P}_n$, $p(h_i) = \psi(h_i)$.

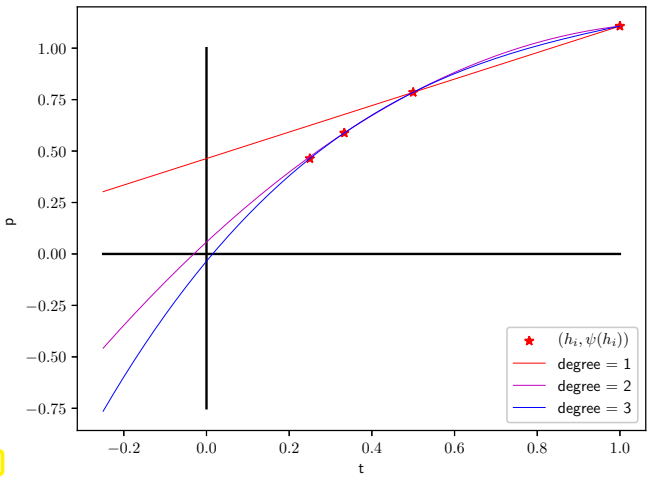


Fig. 158

◁ extrapolating polynomials

In this manufactured example we have $\psi(t) = \arctan(2t)$, which means $\psi(0) = 0$. The higher the degree of the extrapolating polynomial p, the closer is $p(0)$ to 0.

§5.2.3.16 (Numerical differentiation through extrapolation) In Ex. 1.5.4.7 we have already seen a situation, where we wanted to compute the limit of a function $\psi(h)$ for $h \to 0$, but could not do it with sufficient accuracy. In this case $\psi(h)$ was a one-sided difference quotient with span h, meant to approximate $f'(x)$ for a differentiable function f. The cause of numerical difficulties was cancellation → § 1.5.4.5.

Now we will see how to dodge cancellation in difference quotients and how to use extrapolation to zero to computes derivatives with high accuracy:

Given: smooth function $f : I \subset \mathbb{R} \mapsto \mathbb{R}$ in procedural form: `function y = f(x)`

Sought: (approximation of) $f'(x)$, $x \in I$.

Natural idea: approximation of derivative by (symmetric) difference quotient

$$\frac{df}{dx}(x) \approx \frac{f(x+h) - f(x-h)}{2h} \, . \tag{5.2.3.17}$$

▶ straightforward implementation fails due to cancellation in the numerator, see also Ex. 1.5.4.7.

C++-code 5.2.3.18: Numeric differentiation through difference quotients ➙ GITLAB

```
// numerical differentiation using difference quotient
// f'(x) = lim_{h→0} (f(x+h)−f(x))/h
// IN: f (function object) = function to derive
// df = exact derivative (to compute error),
// name = string of function name (for plot filename)
// OUT: plot of error will be saved as "<name>numdiff.eps"
template <class Function, class Derivative>
void diff(const double x, Function& f, Derivative& df, const std::string name) {
  std::vector<long double> error, h;
  // build vector of widths of difference quotients
  for (int i = -1; i >= -61; i -= 5) h.push_back(std::pow(2,i));
  for (unsigned j = 0; j < h.size(); ++j) {
    // compute approximate solution using difference quotient
    double df_approx = (f(x + h[j]) - f(x))/h[j];
    // compute relative error
    double rel_err = std::abs((df_approx - df(x))/df(x));
    error.push_back(rel_err);
  }
```

This is apparent in the following approximation error tables for three simple functions and $x = 1.1$.

$f(x) = \arctan(x)$

h	Relative error
2^{-1}	0.20786640808609
2^{-6}	0.00773341103991
2^{-11}	0.00024299312415
2^{-16}	0.00000759482296
2^{-21}	0.00000023712637
2^{-26}	0.00000001020730
2^{-31}	0.00000005960464
2^{-36}	0.00000679016113

$f(x) = \sqrt{x}$

h	Relative error
2^{-1}	0.09340033543136
2^{-6}	0.00352613693103
2^{-11}	0.00011094838842
2^{-16}	0.00000346787667
2^{-21}	0.00000010812198
2^{-26}	0.00000001923506
2^{-31}	0.00000001202188
2^{-36}	0.00000198842224

$f(x) = \exp(x)$

h	Relative error
2^{-1}	0.29744254140026
2^{-6}	0.00785334954789
2^{-11}	0.00024418036620
2^{-16}	0.00000762943394
2^{-21}	0.00000023835113
2^{-26}	0.00000000429331
2^{-31}	0.00000012467100
2^{-36}	0.00000495453865

Recall the considerations elaborated in Ex. 1.5.4.7. Owing to the impact of roundoff errors amplified by cancellation, $h \to 0$ does not achieve arbitrarily high accuracy. Rather, we observe fewer correct digits for very small h!

Extrapolation offers a numerically stable (→ Def. 1.5.5.19) alternative, because for a $2(n+1)$-times continuously differentiable function $f : I \subset \mathbb{R} \mapsto \mathbb{R}$, $x \in I$ we find that the symmetric difference quotient behaves like a polynomial in h^2 in the vicinity of $h = 0$. Consider Taylor sum of f in x with Lagrange remainder term:

$$\psi(h) := \frac{f(x+h) - f(x-h)}{2h} \sim f'(x) + \sum_{k=1}^{n} \frac{1}{(2k)!} \frac{\mathrm{d}^{2k} f}{\mathrm{d}x^{2k}}(x)\, h^{2k} + \frac{1}{(2n+2)!} f^{(2n+2)}(\xi(x)) \,.$$

Since $\lim_{h\to 0} \psi(h) = f'(x)$

➙ approximate $f'(x)$ by interpolation of ψ in points h_i.

The following C++ function `diffex()` implements extrapolation to zero of symmetric difference quotients relying on the update-friendly version of the Aitken-Neville algorithm as presented in § 5.2.3.11, Code 5.2.3.12. Note that the extrapolated value taking into account all available difference quotients always resides in `y[0]`.

C++-code 5.2.3.19: Numerical differentiation by adaptive extrapolation to zero ⇝ GITLAB

```
2  // Extrapolation based numerical differentation
3  // with a posteriori error control
4  // f:  handle of a function defined in a neighbourhood of x ∈ ℝ
5  // x:  point at which approximate derivative is desired
6  // h0:  initial distance from x
7  // rtol:  relative target tolerance, atol:  absolute tolerance
8  template <class Function>
9  double diffex(Function& f, const double x, const double h0,
10           const double rtol, const double atol) {
11   const unsigned nit = 10; // Maximum depth of extrapolation
12   VectorXd h(nit); h[0] = h0; // Widths of difference quotients
13   VectorXd y(nit); // Approximations returned by difference quotients
14   y[0] = (f(x + h0) - f(x - h0))/(2*h0); // Widest difference quotient
15
16   // using Aitken-Neville scheme with x = 0, see Code 5.2.3.10
17   for (unsigned i = 1; i < nit; ++i) {
18     // create data points for extrapolation
19     h[i] = h[i-1]/2; // Next width half as big
20     y[i] = ( f(x + h[i]) - f(x - h[i]) )/(2.0*h[i]); // difference quotient
21     // Aitken-Neville update
22     for (int k = static_cast<int>(i - 1); k >= 0; --k) {
23       y[k] = y[k+1] - (y[k+1]-y[k])*h[i]/(h[i]-h[k]);
24     }
25     // termination of extrapolation when desired tolerance is reached
26     const double errest = std::abs(y[1]-y[0]); // error indicator
27     if ( errest < rtol*std::abs(y[0]) || errest < atol ) { //
28       break;
29     }
30   }
31   return y[0]; // Return value extrapolated from largest number of
       difference quotients
32 }
```

While the extrapolation table (→ § 5.2.3.11) is computed, more and more accurate approximations of $f'(x)$ become available. Thus, the difference between the two last approximations (stored in `y[0]` and `y[1]` in Code 5.2.3.19) can be used to gauge the error of the current approximation, it provides an **error indicator**, which can be used to decide when the level of extrapolation is sufficient, see Line 27.

`auto f = [](double x) {return std::atan(x);}`
`diffex(f,1.1,0.5)`

Degree	Relative error
0	0.04262829970946
1	0.02044767428982
2	0.00051308519253
3	0.00004087236665
4	0.00000048930018
5	0.00000000746031
6	0.00000000001224

`auto g = [](double x) {return std::sqrt(x);}`
`diffex(g,1.1,0.5)`

Degree	Relative error
0	0.02849215135713
1	0.01527790811946
2	0.00061205284652
3	0.00004936258481
4	0.00000067201034
5	0.00000001253250
6	0.00000000004816
7	0.00000000000021

`auto h = [](double x) {return std::exp(x);}`
`diffex(h,1.1,0.5)`

Degree	Relative error
0	0.04219061098749
1	0.02129207652215
2	0.00011487434095
3	0.00000825582406
4	0.00000000589624
5	0.00000000009546
6	0.00000000000002

Advantage: guaranteed accuracy ➙ efficiency

Review question(s) 5.2.3.20 (Extrapolation to zero)

(Q5.2.3.20.A) We consider a convergent sequence $(\alpha_n)_{n\in\mathbb{N}}$ of real numbers, whose terms can be obtained as ouput of a black-box function **double** `alpha(unsigned int n)`. Function calls become the more expensive the larger n. How might extrapolation to zero be employed to compute the limit $\lim_{n\to\infty} \alpha_n$?

Hint. Consider a function $\psi :]0,1] \to]\mathbb{R}$ with $\psi(h) := \alpha_n$, if $h = n^{-1}$.

(Q5.2.3.20.B) We consider the space of even polynomials

$$\mathcal{P}_{2n}^{\text{even}} := \left\{ t \mapsto \gamma_0 + \gamma_2 t^2 + \gamma_4 t^4 + \cdots + \gamma_{2n} t^{2n},\ \gamma_{2j} \in \mathbb{R},\ j = 0,\ldots,n \right\}, \quad n \in \mathbb{N}\,.$$

- Given data points (t_j, y_j), $j = 0,\ldots,n$, $t_j > 0$, show that there is a unique $p \in \mathcal{P}_{2n}^{\text{even}}$ satisfying the interpolation conditions $p(t_j) = y_j$ for all $\in \{0,\ldots,n\}$.
- You have an implementation of the functor class for the evaluation of polynomial interpolants at your disposal, see Code 5.2.3.1:

C++ code Code 5.2.3.1: Polynomial Interpolation ↪ GITLAB

```
class PolyInterp {
 public:
  // Constructors taking node vector [t_0,...,t_n]^T as argument
  PolyInterp(const Eigen::VectorXd &t);
  template <typename SeqContainer>
  PolyInterp(const SeqContainer &v);
  // Evaluation operator for data (y_0,...,y_n); computes
  // p(x_k) for x_ks passed in x
  Eigen::VectorXd eval(const Eigen::VectorXd &y,
                       const Eigen::VectorXd &x) const;
 private:
  // various internal data describing p
  Eigen::VectorXd _t;
};
```

Explain how you can use an object of this type to perform multiple evaluations of the unique even polynomial interpolant $p \in \mathcal{P}_{2n}^{\text{even}}$ of the data points (t_j, y_j), $j = 0,\ldots,n$, $t_j > 0$.

△

5.2.3.4 Newton Basis and Divided Differences

Supplementary literature. We also refer to [DR08, Sect. 8.2.4], [QSS00, Sect. 8.2].

In § 5.2.3.8 we have seen a method to evaluate partial polynomial interpolants for a single or a few evaluation points efficiently. Now we want to do this for *many* evaluation points that may not be known when we receive information about the first interpolation points.

C++ code 5.2.3.21: Polynomial evaluation ↪ GITLAB

```
class PolyEval {
private:
  // evaluation point and various internal data describing the
     polynomials
public:
```

```
  // Idle Constructor
  PolyEval();
  // Add another data point and update internal information
  void addPoint(double t, double y);
  // Evaluation of current interpolating polynomial at many points
  Eigen::VectorXd operator()(const Eigen::VectorXd &x) const;
};
```

The challenge: Both `addPoint()` and the evaluation operator **operator** `()` may be called many times and the implementation has to remain efficient under these circumstances.

Why not use the techniques from § 5.2.3.2? Drawback of the Lagrange basis or barycentric formula: adding another data point affects *all* basis polynomials/all precomputed values!

§5.2.3.22 (Newton basis for $\mathcal{P}_n$)

Our tool now is an "update friendly" representation of the polynomial interpolants in terms of the **Newton basis** for $\mathcal{P}_n$

$$N_0(t) := 1\,, \quad N_1(t) := (t-t_0)\,, \quad \ldots \quad , \quad N_n(t) := \prod_{i=0}^{n-1}(t-t_i)\,. \tag{5.2.3.23}$$

Note that, clearly, $N_n \in \mathcal{P}_n$ with *leading coefficient* 1. This implies the linear independence of $\{N_0,\ldots,N_n\}$ and, in light of $\dim \mathcal{P}_n = n+1$ by Thm. 5.2.1.2, gives us the basis property of that subset of $\mathcal{P}_n$.

The abstract considerations of § 5.1.0.21 still apply and we get an $(n+1)\times(n+1)$ linear system of equations for the coefficients a_j, $j=0,\ldots,n$, of the polynomial interpolant in Newton basis:

$$a_j \in \mathbb{R}\colon \quad a_0N_0(t_j) + a_1N_1(t_j) + \cdots + a_nN_n(t_j) = y_j, \quad j=0,\ldots,n\,. \tag{5.2.3.24}$$

$\Leftrightarrow$ *triangular* linear system

$$\begin{bmatrix} N_0(t_0) & N_1(t_0) & \cdots & N_n(t_0) \\ N_0(t_1) & N_1(t_1) & \cdots & N_n(t_1) \\ \vdots & \vdots & & \vdots \\ N_0(t_n) & N_1(t_n) & \cdots & N_n(t_n) \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix}$$

$$\Updownarrow$$

$$\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & (t_1-t_0) & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 1 & (t_n-t_0) & \cdots & \prod_{i=0}^{n-1}(t_n-t_i) \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix}. \tag{5.2.3.25}$$

This triangular linear system can be solved by simple forward substitution:

$$\begin{aligned} a_0 &= y_0\,, \\ a_1 &= \frac{y_1-a_0}{t_1-t_0} = \frac{y_1-y_0}{t_1-t_0}\,, \\ a_2 &= \frac{y_2-a_0-(t_2-t_0)a_1}{(t_2-t_0)(t_2-t_1)} = \frac{y_2-y_0-(t_2-t_0)\frac{y_1-y_0}{t_1-t_0}}{(t_2-t_0)(t_2-t_1)} = \frac{\frac{y_2-y_0}{t_2-t_0}-\frac{y_1-y_0}{t_1-t_0}}{t_2-t_1}\,, \end{aligned}$$

$\vdots$

We observe that in the course of forward substitution the same quantities computed again and again. This suggests that a more efficient implementation is possible. ⌟

§5.2.3.26 (Divided differences) In order to reveal the pattern, we turn to a new interpretation of the coefficients $a_j \in \mathbb{R}$ of the interpolating polynomials

$$p(t) = a_0 N_0(t) + a_1 N_1(t) + \cdots + a_n N_n(t) , \quad t \in \mathbb{R} ,$$

represented in the Newton basis $\{N_0, \ldots, N_n\}$ from (5.2.3.23). We start with the following observation.

> The Newton basis polynomial $N_j(t)$ has degree j and leading coefficient 1
>
> ⇚ a_j is the leading coefficient of the interpolating polynomial $p_{0,j}$.

Using the notation $p_{\ell,m}$ for partial polynomial interpolants through the data points $(t_\ell, y_\ell), \ldots, (t_m, y_m)$, which was introduced in Section 5.2.3.2, see (5.2.3.9) we can state the recursion

$$\begin{aligned} p_{k,k}(x) &\equiv y_k \quad \text{(“constant polynomial”)} , && k = 0, \ldots, n , \\ p_{k,\ell}(x) &= \frac{(x - t_k) p_{k+1,\ell}(x) - (x - t_\ell) p_{k,\ell-1}(x)}{t_\ell - t_k} \\ &= p_{k+1,\ell}(x) + \frac{x - t_\ell}{t_\ell - t_k} \left(p_{k+1,\ell}(x) - p_{k,\ell-1}(x) \right) , && 0 \le k \le \ell \le n , \end{aligned} \tag{5.2.3.9}$$

Since both $p_{k+1,\ell}$ and $p_{k,\ell-1}$ are of the same degree and multiplication with the monomials $(x - t_k)$ and $(x - t_\ell)$ does not affect the leading coefficients, (**??**) implies a recursion for the leading coefficients $a_{\ell,m}$ of the interpolating polynomials $p_{\ell,m}$,

$$a_{\ell,m} = \frac{a_{\ell+1,m} - a_{\ell,m-1}}{t_m - t_\ell} , \quad 0 \le \ell \le m \le n . \tag{5.2.3.27}$$

Hence, instead of using elimination for a triangular linear system, we find a simpler and more efficient algorithm using the so-called **divided differences**, which are defined by the recursion

$$\begin{aligned} y[t_i] &= y_i \\ y[t_i, \ldots, t_{i+k}] &= \frac{y[t_{i+1}, \ldots, t_{i+k}] - y[t_i, \ldots, t_{i+k-1}]}{t_{i+k} - t_i} \quad \text{(recursion)} . \end{aligned} \tag{5.2.3.28}$$

We adopt this strange notation for the sake of compatibility with literature. ⌟

§5.2.3.29 (Efficient computation of divided differences) Divided differences can be computed by the divided differences scheme, which is closely related to the Aitken-Neville scheme from Code 5.2.3.10:

$$\begin{array}{c|cccc} t_0 & y[t_0] & & & \\ & & > y[t_0,t_1] & & \\ t_1 & y[t_1] & & > y[t_0,t_1,t_2] & \\ & & > y[t_1,t_2] & & > y[t_0,t_1,t_2,t_3], \\ t_2 & y[t_2] & & > y[t_1,t_2,t_3] & \\ & & > y[t_2,t_3] & & \\ t_3 & y[t_3] & & & \end{array} \tag{5.2.3.30}$$

The elements can be computed from left to right, every "$>$" indicates the evaluation of the recursion formula (5.2.3.28).

However, we can again resort to the idea of § 5.2.3.11 and traverse (5.2.3.30) along the diagonals from left bottom to right top: If a new datum (t_0, y_0) is added, it is enough to compute the $n+1$ new divided differences

$$y[t_0] \quad , \quad y[t_0,t_1] \quad , \quad y[t_0,t_1,t_2] \, , \ldots \, , y[t_0,\ldots,t_n] \, .$$

The C++ function `divdiff()` listed in Code 5.2.3.31 computes divided differences for data points (t_i, y_i), $i = 0,\ldots,n$, in this fashion. For $n = 3$ the values of the outer loop variable `l` in the different combinations are as follows:

$$\begin{array}{c|c|c|c|c|c|c} t_0 & y[t_0] & & & & & \\ & & l=3> & y[t_0,t_1] & & & \\ t_1 & y[t_1] & & & l=3> & y[t_0,t_1,t_2] & \\ & & l=2> & y[t_1,t_2] & & & l=3> \quad y[t_0,t_1,t_2,t_3], \\ t_2 & y[t_2] & & & l=2> & y[t_1,t_2,t_3] & \\ & & l=1> & y[t_2,t_3] & & & \\ t_3 & y[t_3] & & & & & \end{array} \tag{5.2.3.30}$$

In `divdiff()` the divided differences $y[t_0], y[t_0,t_1], \ldots, y[t_0,\ldots,t_n]$ overwrite the original data values y_j in the vector `y` (in-situ computation).

C++ code 5.2.3.31: In-situ computation of divided differences ➽ GITLAB

```
// IN: t = node set (mutually different)
// y = nodal values
// OUT: y = coefficients of polynomial in Newton basis
void divdiff(const Eigen::VectorXd &t, Eigen::VectorXd &y) {
  const Eigen::Index n = y.size();
  // Follow scheme (5.2.3.30), recursion (5.2.3.27)
  for (Eigen::Index l = 1; l < n; ++l) {
    for (Eigen::Index j = n - l; j < n; ++j) {
      y[j] = (y[j] - y[j - 1]) / (t[j] - t[n - 1 - l]);
    }
  }
}
```

Computational effort: $O(n^2)$ for no. of data points $n \to \infty$

By derivation, the computed finite differences are the coefficients of interpolating polynomials in the Newton basis

$$p(t) = a_0 + a_1(t-t_0) + a_2(t-t_0)(t-t_1) + \cdots + a_n \prod_{j=0}^{n-1}(t-t_j) \tag{5.2.3.32}$$
$$a_0 = y[t_0], \; a_1 = y[t_0,t_1], \; a_2 = y[t_0,t_1,t_2], \; \ldots .$$

Thus, `divdiff()` from Code 5.2.3.31 computes the coefficients a_j, $j = 0,\ldots,n$, of the polynomial interpolant with respect to the Newton basis. It uses only the first $j+1$ data points to find a_j. ⌟

§5.2.3.33 (Efficient evaluation of polynomial in Newton form) Let a polynomial be given in "Newton

form", that is, as a linear combination of Newton basis polynomials as introduced in (5.2.3.23):

$$\begin{aligned} p(t) &= a_0 N_0(t) + a_1 N_1(t) + \cdots + a_n N_n(t) \\ &= a_0 \cdot 1 + a_1 \underbrace{(t-t_0)}_{=N_1(t)} + a_2 \underbrace{(t-t_0)(t-t_1)}_{=N_2(t)} + \cdots + a_n \underbrace{\prod_{j=0}^{n-1}(t-t_j)}_{=N_n(t)}, \quad t \in \mathbb{R}, \end{aligned}$$

with known coefficients a_j, $j = 0, \ldots, n$, e.g., available as the components of a vector. Embark on "associative rewriting",

$$p(t) = (\ldots((a_n(t-t_{n-1}) + a_{n-1})(t-t_{n-2}) + a_{n-2}) \cdot \ldots \cdot + a_1)(t-t_0) + a_0 ,$$

which reveals how we can perform the "backward evaluation" of $p(t)$ in the spirit of Horner's scheme (→ § 5.2.1.5, [DR08, Alg. 8.20]):

$$p \leftarrow a_n, \quad p \leftarrow (t-t_{n-1})p + a_{n-1}, \quad p \leftarrow (t-t_{n-2})p + a_{n-2}, \quad \ldots .$$

A C++ implementation of this idea is given next.

C++ code 5.2.3.34: Divided differences evaluation by modified Horner scheme ↬ GITLAB

```
2  // Evaluation of a polynomial in Newton form, that is, represented
       through the
3  // vector of its basis expansion coefficients with respect to the Newton
       basis
4  // (5.2.3.23).
5  Eigen::VectorXd evalNewtonForm(const Eigen::VectorXd &t,
6                                 const Eigen::VectorXd &a,
7                                 const Eigen::VectorXd &x) {
8    const Eigen::Index n = a.size() - 1;
9    const Eigen::VectorXd ones = Eigen::VectorXd::Ones(x.size());
10   Eigen::VectorXd p{a[n] * ones};
11   for (Eigen::Index j = n - 1; j >= 0; --j) {
12     p = (x - t[j] * ones).cwiseProduct(p) + a[j] * ones;
13   }
14   return p;
15 }
```

Computational effort: Asymptotically $O(n)$ for a single evaluation of $p(t)$.
(Can be interleaved with the computation of the a_js, see Code 5.2.3.21.)

EXAMPLE 5.2.3.35 (Class `PolyEval`) We show the implementation of a C++ class supporting the efficient update and evaluation of an interpolating polynomial making use of

- the representation of the Lagrange polynomial interpolants in the Newton basis (5.2.3.23),
- the computation of representation coefficients through a divided difference scheme (5.2.3.30), see Code 5.2.3.31,
- and point evaluations of the polynomial interpolants by means of Horner-like scheme as introduced in Code 5.2.3.34.

To understand the code return to the triangular linear system for the Newton basis expansion coefficients

a_j of a Lagrange polynomial interpolant of degree n through (t_i, y_i), $i = 0, \ldots, n$:

$$\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & (t_1 - t_0) & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 1 & (t_n - t_0) & \cdots & \prod_{i=0}^{n-1}(t_n - t_i) \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix} . \tag{5.2.3.25}$$

Given, $a_0, \ldots, a_{n-1}$ we can thus compute a_n from

$$\begin{aligned} a_n &= \prod_{i=0}^{n-1}(t_n - t_i)^{-1} \left(y_n - \sum_{k=0}^{n-1} \prod_{i=0}^{k-1} (t_n - t_i)\, a_k \right) \\ &= \prod_{i=0}^{n-1}(t_n - t_i)^{-1} y_n - \sum_{k=0}^{n-1} \prod_{i=k}^{n-1} (t_n - t_i)^{-1}\, a_k \\ &= (\ldots(((y_n - a_0)/(t_n - t_0) - a_1)/(t_n - t_1) - a_2)/\cdots - a_{n-1})/(t_n - t_{n-1}) . \end{aligned}$$

C++-code 5.2.3.36: Definition of a class for "update friendly" polynomial interpolant ➙ GITLAB

```
class PolyEval {
private:
  std::vector<double> t; // Interpolation nodes
  std::vector<double> y; // Coefficients in Newton representation
public:
  PolyEval(); // Idle constructor
  void addPoint(double t, double y); // Add another data point
  // evaluate value of current interpolating polynomial at x,
  double operator() (double x) const;
};
```

C++-code 5.2.3.37: Implementation of class PolyEval ➙ GITLAB

```
PolyEval::PolyEval() {}

void PolyEval::addPoint(double td, double yd) {
  t.push_back(td);
  y.push_back(yd);
  int n = t.size();
  for (int j = 0; j < n - 1; j++)
    y[n - 1] = ((y[n - 1] - y[j]) / (t[n - 1] - t[j]));
}

double PolyEval::operator()(double x) const {
  double s = y.back();
  for (int i = y.size() - 2; i >= 0; --i)
    s = s * (x - t[i]) + y[i];
  return s;
}
```

⌟

Remark 5.2.3.38 (Divided differences and derivatives) If $y_0, \ldots, y_n$ are the values of a smooth function

f in the points $t_0, \ldots, t_n$, that is, $y_j := f(t_j)$, then

$$y[t_i, \ldots, t_{i+k}] = \frac{f^{(k)}(\xi)}{k!}$$

for a certain $\xi \in [t_i, t_{i+k}]$, see [DR08, Thm. 8.21]. ⌟

Review question(s) 5.2.3.39 (Newton basis and divided differences)

(Q5.2.3.39.A) Given a node set $\{t_0, t_1, \ldots, t_n\}$ let $a_0, \ldots, a_n \in \mathbb{R}$ be the coefficients of a polynomial p in the associated Newton basis $\{N_0, \ldots, N_n\}$. Outline an efficient algorithm for computing the basis expansion coefficients of p with respect to the basis $\{L_0, \ldots, L_n\}$ of Lagrange polynomials for given node set.

(Q5.2.3.39.B) Given the value vector $\mathbf{y} \in \mathbb{R}^{n+1}$ and the node set $\{t_0, \ldots, t_n\} \subset \mathbb{R}$, remember the notation $\mathbf{y}[t_k, t_\ell]$, for divided differences: $\mathbf{y}[t_k, t_\ell]$ is the leading coefficient of the unique polynomial interpolating the data points $\left(t_j, (\mathbf{y})_j\right)_{j=k}^{\ell}$, $0 \le k, \ell \le n$ (C++ indexing).

What can you conclude from $\mathbf{y}[t_0, t_j] = 0$ for all $j \in \{m, \ldots, n\}$ for some $m \in \{1, \ldots, n\}$?

△

5.2.4 Polynomial Interpolation: Sensitivity

Supplementary literature. For related discussions see [QSS00, Sect. 8.1.3].

This section addresses a *major shortcoming of polynomial interpolation* in case the interpolation knots t_i are imposed, which is usually the case when given data points have to be interpolated, *cf.* Ex. 5.1.0.8.

This liability has to do with the sensitivity of the Lagrange polynomial interpolation problem. From Section 2.2.2 remember that the **sensitivity/conditioning** of a problem provides a measure for the propagation of perturbations in the data/inputs to the results/outputs.

§5.2.4.1 (The Lagrange polynomial interpolation problem) As explained in § 1.5.5.1 a "problem" in the sense of numerical analysis is a mapping/function from a data/input set X into a set $Y$of results/outputs.

Owing to the existence and uniqueness of the polynomial interpolant as asserted in Thm. 5.2.2.7, the Lagrange polynomial interpolation problem (LIP) as introduced in Section 5.2.2 describes a mapping

$$(\mathbb{R} \times \mathbb{R})^{n+1} \longrightarrow \mathcal{P}_n \quad , \quad ((t_i, y_i))_{i=0}^n \mapsto p \in \mathcal{P}_n \colon p(t_i) = y_i \, , \quad i = 0, \ldots, n \, . \tag{5.2.4.2}$$

from sets of $n+1$ data points, $n \in \mathbb{N}_0$ to polynomials of degree n. Hence, LIP maps a finite sequence of numbers to a function, and both the data/input set and result/output set have the structure of vector spaces.

A more restricted view considers the *linear* interpolation operator from Cor. 5.2.2.8

$$\mathsf{I}_{\mathcal{T}} : \begin{cases} \mathbb{R}^{n+1} & \to \mathcal{P}_n \, , \\ (y_0, \ldots, y_n)^T & \mapsto \text{interpolating polynomial } p \, . \end{cases} \tag{5.2.2.9}$$

and identifies the Lagrange polynomial interpolation problem with $\mathsf{I}_{\mathcal{T}}$, that is, with the mapping taking only data values to a polynomial. The interpolation nodes are treated as parameters and not considered data. For the sake of simplicity we adopt this view in the sequel. ⌟

EXAMPLE 5.2.4.3 (Oscillating polynomial interpolant (Runge's counterexample) → [DR08, Sect. 8.3], [QSS00, Ex. 8.1]) This example offers a glimpse of the problems haunting polynomial interpolation.

We examine the polynomial Lagrange interpolant (→ Section 5.2.2, (5.2.2.2)) for uniformly spaced nodes and the following data:

$$\mathcal{T} := \left\{ -5 + \tfrac{10}{n} j \right\}_{j=0}^{n},$$
$$y_j = \frac{1}{1+t_j^2}, \quad j = 0, \dots n.$$

Plotted is the interpolant for $n = 10$ and the interpolant for which the data value at $t = 0$ has been perturbed by 0.1
(See also Ex. 6.2.2.11 below.)

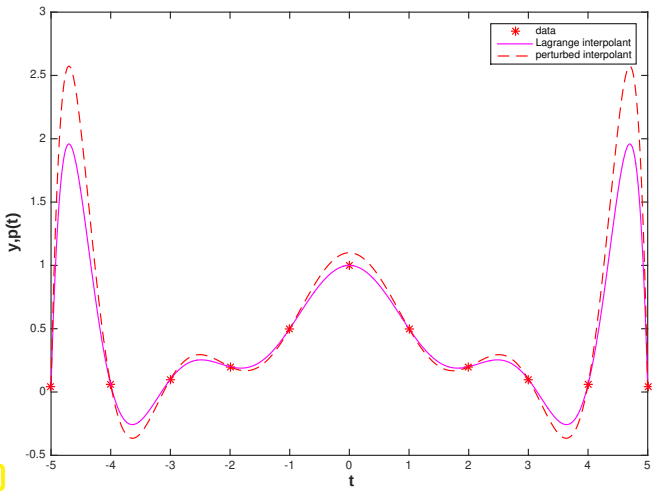


Fig. 159

☞ possible strong oscillations of interpolating polynomials of *high degree* on *uniformly spaced* nodes!

☞ Slight perturbations of data values can engender strong variations of a *high-degree* Lagrange interpolant "far away".

In fuzzy terms, what we have observed is "high sensitivity" of polynomial interpolation with respect to perturbations in the data values: small perturbations in the data can cause big variations of the polynomial interpolants in certain points, which is clearly undesirable. ⌟

§5.2.4.4 (Norms on spaces of functions) For measuring the size of perturbations we need norms (→ Def. 1.5.5.4) on data and result spaces. For the value vectors $\mathbf{y} := [y_0, \dots, y_n]^\top \in \mathbb{R}^{n+1}$ we can use any vector norm, see § 1.5.5.3, for instance the maximum norm $\|\mathbf{y}\|_\infty$.

However the result space is a vector space of functions $I \subset \mathbb{R} \to \mathbb{R}$ and so we also need norms on the vector space of continuous functions $C^0(I)$, $I \subset \mathbb{R}$. The following norms are the most relevant:

$$\text{supremum norm} \quad \|f\|_{L^\infty(I)} := \sup\{|f(t)| : t \in I\}, \tag{5.2.4.5}$$
$$L^2\text{-norm} \quad \|f\|_{L^2(I)}^2 := \int_I |f(t)|^2 \, \mathrm{d}t, \tag{5.2.4.6}$$
$$L^1\text{-norm} \quad \|f\|_{L^1(I)} := \int_I |f(t)| \, \mathrm{d}t. \tag{5.2.4.7}$$

Note the relationship with the vector norms introduced in § 1.5.5.3. ⌟

§5.2.4.8 (Sensitivity of linear problem maps) In § 5.1.0.21 we have learned that (polynomial) interpolation gives rise to a linear problem map, see Def. 5.1.0.25. For this class of problem maps the investigation of sensitivity has to study operator norms, a generalization of matrix norms (→ Def. 1.5.5.10).

Let $\mathsf{L} : X \to Y$ be a *linear* problem map between two normed spaces, the data space X (with norm $\|\cdot\|_X$) and the result space Y (with norm $\|\cdot\|_Y$). Thanks to linearity, perturbations of the result $\mathbf{y} := \mathsf{L}(\mathbf{x})$ for the input $\mathbf{x} \in X$ can be expressed as follows:

$$\mathsf{L}(\mathbf{x} + \delta\mathbf{x}) = \mathsf{L}(\mathbf{x}) + \mathsf{L}(\delta\mathbf{x}) = \mathbf{y} + \mathsf{L}(\delta\mathbf{x}).$$

Hence, the sensitivity (in terms of propagation of absolute errors) can be measured by the **operator norm**

$$\|\mathsf{L}\|_{X\to Y} := \sup_{\delta\mathbf{x}\in X\setminus\{\mathbf{0}\}} \frac{\|\mathsf{L}(\delta\mathbf{x})\|_Y}{\|\delta\mathbf{x}\|_X} \,. \tag{5.2.4.9}$$

This can be read as the "matrix norm of L", *cf.* Def. 1.5.5.10. ⌟

It seems challenging to compute the operator norm (5.2.4.9) for $\mathsf{L} = \mathsf{I}_{\mathcal{T}}$ ($\mathsf{I}_{\mathcal{T}}$ the Lagrange interpolation operator for node set $\mathcal{T} \subset I$), $X = \mathbb{R}^{n+1}$ (equipped with a vector norm), and $Y = C(I)$ (endowed with a norm from § 5.2.4.4). The next lemma will provide surprisingly simple concrete formulas.

Lemma 5.2.4.10. Absolute conditioning of polynomial interpolation

Given a mesh $\mathcal{T} \subset \mathbb{R}$ with generalized Lagrange polynomials L_i, $i = 0,\ldots,n$, and fixed $I \subset \mathbb{R}$, the norm of the interpolation operator satisfies

$$\|\mathsf{I}_{\mathcal{T}}\|_{\infty\to\infty} := \sup_{\mathbf{y}\in\mathbb{R}^{n+1}\setminus\{0\}} \frac{\|\mathsf{I}_{\mathcal{T}}(\mathbf{y})\|_{L^\infty(I)}}{\|\mathbf{y}\|_\infty} = \left\|\sum_{i=0}^n |L_i|\right\|_{L^\infty(I)} , \tag{5.2.4.11}$$

$$\|\mathsf{I}_{\mathcal{T}}\|_{2\to 2} := \sup_{\mathbf{y}\in\mathbb{R}^{n+1}\setminus\{0\}} \frac{\|\mathsf{I}_{\mathcal{T}}(\mathbf{y})\|_{L^2(I)}}{\|\mathbf{y}\|_2} \le \left(\sum_{i=0}^n \|L_i\|^2_{L^2(I)}\right)^{\frac{1}{2}} . \tag{5.2.4.12}$$

Proof. (for the L^∞-Norm) By $\triangle$-inequality

$$\|\mathsf{I}_{\mathcal{T}}(\mathbf{y})\|_{L^\infty(I)} = \left\|\sum_{j=0}^n y_j L_j\right\|_{L^\infty(I)} \le \sup_{t\in I} \sum_{j=0}^n |y_j||L_j(t)| \le \|\mathbf{y}\|_\infty \left\|\sum_{i=0}^n |L_i|\right\|_{L^\infty(I)} ,$$

equality in (5.2.4.11) for $\mathbf{y} := (\operatorname{sgn}(L_j(t^*)))_{j=0}^n$, $t^* := \operatorname{argmax}_{t\in I} \sum_{i=0}^n |L_i(t)|$. □

Proof. (for the L^2-Norm) By the $\triangle$-inequality and the Cauchy-Schwarz inequality in $\mathbb{R}^{n+1}$,

$$\sum_{j=0}^n a_j b_j \le \left(\sum_{j=0}^n |a_j|^2\right)^{\frac{1}{2}} \left(\sum_{j=0}^n |b_j|^2\right)^{\frac{1}{2}} \quad \forall a_j, b_j \in \mathbb{R} ,$$

we can estimate

$$\|\mathsf{I}_{\mathcal{T}}(\mathbf{y})\|_{L^2(I)} \le \sum_{j=0}^n |y_j| \|L_j\|_{L^2(I)} \le \left(\sum_{j=0}^n |y_j|^2\right)^{\frac{1}{2}} \left(\sum_{j=0}^n \|L_j\|^2_{L^2(I)}\right)^{\frac{1}{2}} .$$

□

Terminology: **Lebesgue constant** of $\mathcal{T}$: $\lambda_{\mathcal{T}} := \left\|\sum_{i=0}^n |L_i|\right\|_{L^\infty(I)} = \|\mathsf{I}_{\mathcal{T}}\|_{\infty\to\infty}$

Remark 5.2.4.13 (Lebesgue constant for equidistant nodes) We consider Lagrange interpolation for the special setting

$$I = [-1,1], \quad \mathcal{T} = \{-1 + \tfrac{2k}{n}\}_{k=0}^n \quad \text{(uniformly spaced nodes)}.$$

Asymptotic estimate (with (5.2.2.4) and Stirling formula): for $n = 2m$

$$|L_m(1-\tfrac{1}{n})| = \frac{\frac{1}{n}\cdot\frac{1}{n}\cdot\frac{3}{n}\cdot\ldots\cdot\frac{n-3}{n}\cdot\frac{n+1}{n}\cdot\ldots\cdot\frac{2n-1}{n}}{\left(\frac{2}{n}\cdot\frac{4}{n}\cdot\ldots\cdot\frac{n-2}{n}\cdot 1\right)^2} = \frac{(2n)!}{(n-1)2^{2n}((n/2)!)^2 n!} \sim \frac{2^{n+3/2}}{\pi\,(n-1)\,n}$$

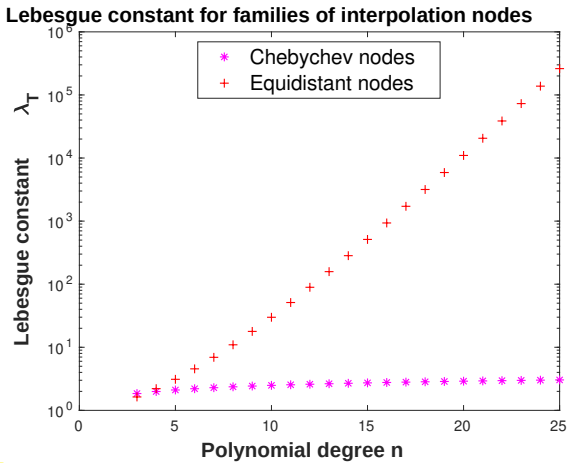


Fig. 160

◁ Numerically computed values for Lebesgue constants for different families of interpolation nodes. The $\|\cdot\|_\infty$-norm of the sum of Lagrange polynomials was approximated by sampling in 10^5 equidistant points, see Code 5.2.4.14.

Sophisticated theory [CR92] gives a lower bound for the Lebesgue constant for uniformly spaced nodes:

$$\lambda_{\mathcal{T}} \geq Ce^{n/2}$$

with $C > 0$ independent of n.

We can also perform a numerical evaluation of the expression

$$\lambda_{\mathcal{T}} = \left\| \sum_{i=0}^{n} |L_i| \right\|_{L^\infty(I)},$$

for the Lebesgue constant of polynomial interpolation, see Lemma 5.2.4.10. The following code demonstrates this:

C++ code 5.2.4.14: *Approximate* computation of Lebesgue constants ↪ GITLAB

```
2  // Computation of Lebesgue constant of polynomial interpolation
3  // with nodes t_i passed in the vector t based on (5.2.4.11).
4  // N specifies the number of sampling points for the approximate
5  // computation of the maximum norm of the Lagrange polynimial
6  // on the interval [-1,1].
7  double lebesgue(const VectorXd& t, const unsigned& N) {
8    const unsigned n = t.size();
9    // compute denominators of normalized Lagrange polynomials relative to
       the nodes t
10   VectorXd den(n);
11   for (unsigned i = 0; i < n; ++i) {
12     VectorXd tmp(n - 1);
13     // Note:  comma initializer can't be used with vectors of length 0
14     if (i == 0) {
15       tmp = t.tail(n - 1);
16     }
17     else if (i == n - 1) {
18       tmp = t.head(n - 1);
19     }
20     else {
21       tmp << t.head(i), t.tail(n - (i + 1));
22     }
23     den(i) = (t(i) - tmp.array()).prod();
24   }
25
26   double l = 0; // return value
27   for (unsigned j = 0; j < N; ++j) {
28     const double x = -1 + j*(2./N); // sampling point for ||·||_{L^∞([-1,1])}
29     double s = 0;
30     for (unsigned k = 0; k < n; ++k) {
31       // v provides value of normalized Lagrange polynomials
32       VectorXd tmp(n - 1);
33       if (k == 0) {
34         tmp = t.tail(n - 1);
35       }
36       else if (k == n - 1) {
```

```
37          tmp = t.head(n - 1);
38        }
39        else {
40          tmp << t.head(k), t.tail(n - (k + 1));
41        }
42        const double v = (x - tmp.array()).prod()/den(k);
43        s += std::abs(v); // sum over modulus of the polynomials
44      }
45      l = std::max(l, s); // maximum of sampled values
46    }
47    return l;
48  }
```

Note: In Code 5.2.4.14 the norm $\|L_i\|_{L^\infty(I)}$ can be computed only approximately by taking the maximum modulus of function values in many sampling points. ⌟

§5.2.4.15 (Importance of knowing the sensitivity of polynomial interpolation) In Ex. 5.1.0.8 we learned that interpolation is an important technique for obtaining a mathematical (and algorithmic) description of a constitutive relationship from measured data. If the interpolation operator is poorly conditioned, tiny measurement errors will lead to big (local) deviations of the interpolant from its "true" form.

Since measurement errors are inevitable, poorly conditioned interpolation procedures are useless for determining constitutive relationships from measurements.

Due to potentially "high sensitivity" interpolation with global polynomials of high degree is not suitable for data interpolation.

⌟

Review question(s) 5.2.4.16 (Polynomial Interpolation: Sensitivity)

(Q5.2.4.16.A) Consider the node set $\mathcal{T} := \{0, 1, \ldots, n\}$, $n \in \mathbb{N}$, and the associated linear polynomial interpolation operator $\mathsf{I}_\mathcal{T} : \mathbb{R}^{n+1} \to \mathcal{P}_n$. Quantify the sensitivity of the mapping

$$\mathbb{R} \to \mathbb{R} \quad , \quad y_n \mapsto \mathsf{I}_\mathcal{T} \left(\begin{bmatrix} * \\ \vdots \\ * \\ y_n \end{bmatrix} \right) \left(\frac{1}{2} \right) ,$$

Hint. The Lagrange polynomials for a node set $\{t_0, \ldots, t_n\} \subset \mathbb{R}$ are given by

$$L_i(t) := \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{t - t_j}{t_i - t_j} , \quad i = 0, \ldots, n . \tag{5.2.2.4}$$

△

5.3 Shape-Preserving Interpolation

When reconstructing a quantitative dependence of quantities from measurements, first principles from physics often stipulate qualitative constraints, which translate into *shape properties* of the function f, e.g., when modelling the material law for a gas:

$$t_i \text{ pressure values, } y_i \text{ densities} \quad \succ \quad f \text{ positive \& monotonic.}$$

Notation: given data: $(t_i, y_i) \in \mathbb{R}^2$, $i = 0, \ldots, n$, $n \in \mathbb{N}$, $t_0 < t_1 < \cdots < t_n$.

EXAMPLE 5.3.0.1 (Magnetization curves)

For many materials physics stipulates properties of the functional dependence of magnetic flux B from magnetic field strength H:

- $H \mapsto B(H)$ smooth (at least C^1),
- $H \mapsto B(H)$ monotonic (increasing),
- $H \mapsto B(H)$ concave

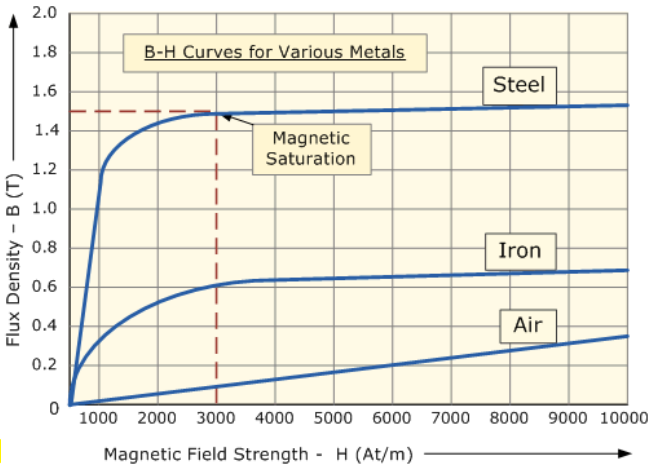


Fig. 161

⌟

5.3.1 Shape Properties of Functions and Data

§5.3.1.1 (The "shape" of data) The section is about "shape preservation". In the previous example we have already seen a few properties that constitute the "shape" of a function: sign, monotonicity and curvature. Now we have to identify analogous properties of data sets in the form of sequences of interpolation points (t_j, y_j), $j = 0, \ldots, n$, t_j pairwise distinct.

Definition 5.3.1.2. monotonic data

The data (t_i, y_i) are called **monotonic** when $y_i \geq y_{i-1}$ or $y_i \leq y_{i-1}$, $i = 1, \ldots, n$.

Definition 5.3.1.3. Convex/concave data

The data $\{(t_i, y_i)\}_{i=0}^n$ are called **convex** (**concave**) if

$$\Delta_j \overset{(\geq)}{\leq} \Delta_{j+1}\,, \quad j = 1, \ldots, n-1 \quad , \quad \Delta_j := \frac{y_j - y_{j-1}}{t_j - t_{j-1}}\,, \quad j = 1, \ldots, n\,.$$

Mathematical characterization of convex data:

$$y_i \leq \frac{(t_{i+1} - t_i)y_{i-1} + (t_i - t_{i-1})y_{i+1}}{t_{i+1} - t_{i-1}} \qquad \forall\, i = 1, \ldots, n-1,$$

i.e., each data point lies below the line segment connecting the other data, *cf.* definition of convexity of a function [Str09, Def. 5.5.2].

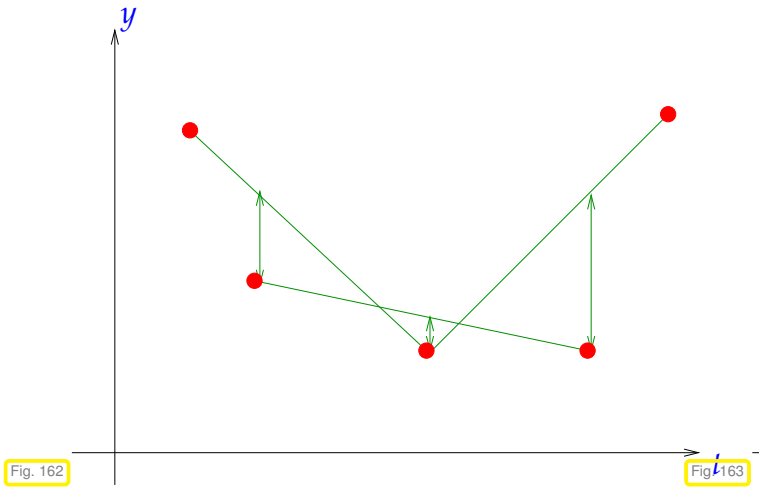


Fig. 162

Convex data

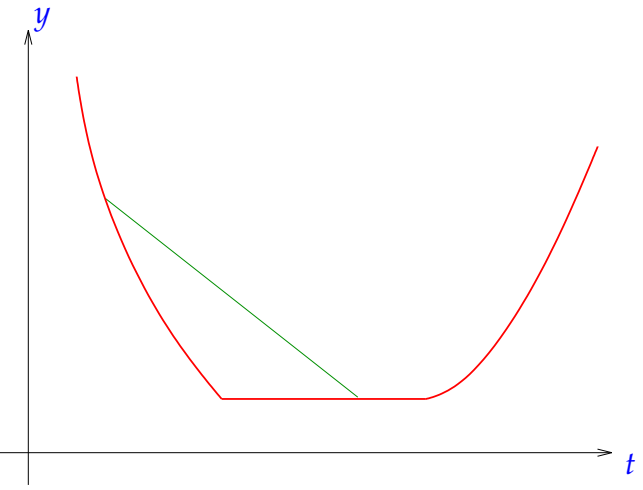


Fig. 163

Convex function

Definition 5.3.1.4. Convex/concave function → [Str09, Def. 5.5.2]

$$f : I \subset \mathbb{R} \mapsto \mathbb{R} \quad \begin{matrix}\textbf{convex}\\ \textbf{concave}\end{matrix} \quad :\Leftrightarrow \quad \begin{matrix} f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y) \\ f(\lambda x + (1-\lambda)y) \ge \lambda f(x) + (1-\lambda) f(y) \end{matrix} \quad \begin{matrix} \forall\, 0 \le \lambda \le 1\,, \\ \forall\, x, y \in I\,. \end{matrix}$$

§5.3.1.5 ((Local) shape preservation) Now we consider interpolation problem to build an interpolant f with special properties inherited from the given data (t_i, y_i), $i = 0, \ldots, n$.

Goal: shape preserving interpolation:

positive data	$\longrightarrow$	positive interpolant f,
monotonic data	$\longrightarrow$	monotonic interpolant f,
convex data	$\longrightarrow$	convex interpolant f.

More ambitious goal: **local shape preserving interpolation**: for each subinterval $I = [t_i, t_{i+j}]$

positive data in I	$\longrightarrow$	locally positive interpolant $f\|_I$,
monotonic data in I	$\longrightarrow$	locally monotonic interpolant $f\|_I$,
convex data in I	$\longrightarrow$	locally convex interpolant $f\|_I$.

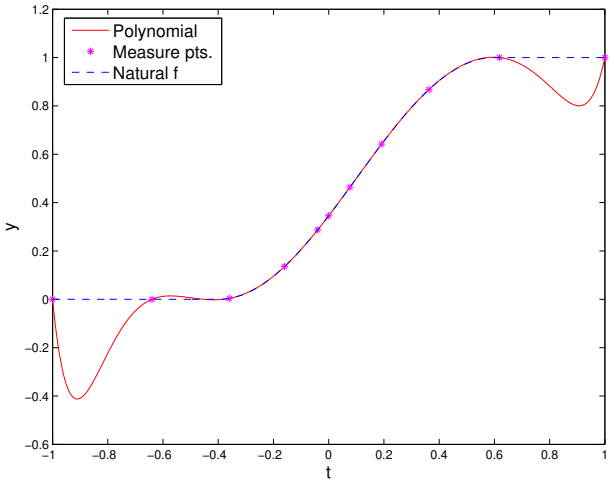
EXPERIMENT 5.3.1.6 (Bad behavior of global polynomial interpolants)

We perform Lagrange interpolation for the following positive and monotonic data:

t_i	-1.0	-0.640	-0.3600	-0.1600	-0.0400	0.0000	0.0770	0.1918	0.3631	0.6187	1.0
y_i	0.0	0.000	0.0039	0.1355	0.2871	0.3455	0.4639	0.6422	0.8678	1.0000	1.0

created by taking points on the graph of

$$f(t) = \begin{cases} 0 & \text{if } t < -\frac{2}{5}\,, \\ \frac{1}{2}(1 + \cos(\pi(t - \frac{3}{5}))) & \text{if } -\frac{2}{5} < t < \frac{3}{5}\,, \\ 1 & \text{otherwise.} \end{cases}$$



⌟

← Interpolating polynomial, degree = 10

We observe oscillations at the endpoints of the interval (see Fig. 159), which means that we encounter

- no locality,
- no positivity,
- no monotonicity,
- no local conservation of the curvature,

in the case of global polynomial interpolation.

5.3.2 Piecewise Linear Interpolation

There is a very simple method of achieving perfect shape preservation by means of a linear (→ § 5.1.0.21) interpolation operator into the space of continuous functions:

Data: $(t_i, y_i) \in \mathbb{R}^2$, $i = 0, \ldots, n$, $n \in \mathbb{N}$, $t_0 < t_1 < \cdots < t_n$.

Then the piecewise linear interpolant $s : [t_0, t_n] \to \mathbb{R}$ is defined as, *cf.* Ex. 5.1.0.15:

$$s(t) = \frac{(t_{i+1} - t)y_i + (t - t_i)y_{i+1}}{t_{i+1} - t_i} \quad \text{for} \quad t \in [t_i, t_{i+1}] \,. \tag{5.3.2.1}$$

The piecewise linear interpolant is also called a polygonal curve. It is continuous and consists of n line segments.

Piecewise linear interpolant of data from Fig. 162 ▷

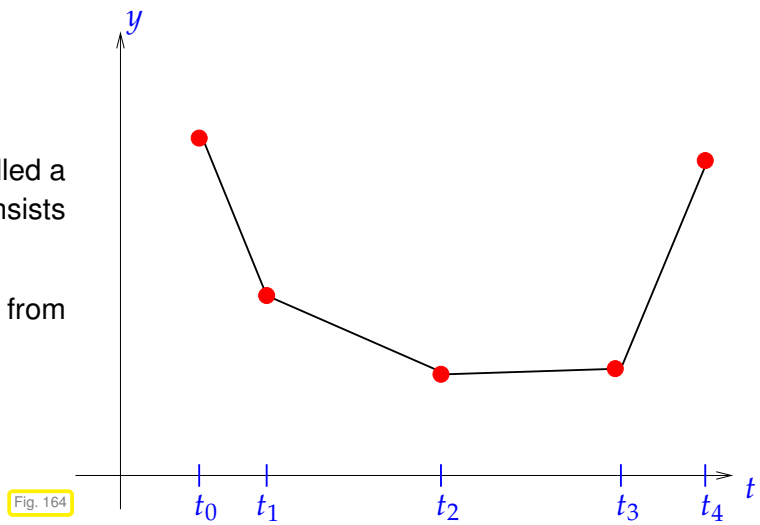


Fig. 164

Piecewise linear interpolation means simply "connect the data points in $\mathbb{R}^2$ using straight lines".

Obvious: linear interpolation is linear (as mapping $\mathbf{y} \mapsto s$, see Def. 5.1.0.25) and local in the following sense:

$$y_j = \delta_{ij} \,, \quad i, j = 0, \ldots, n \quad \Rightarrow \quad \operatorname{supp}(s) \subset [t_{i-1}, t_{i+1}] \,. \tag{5.3.2.2}$$

As obvious are the properties asserted in the following theorem. The local preservation of curvature is a

straightforward consequence of Def. 5.3.1.3.

Theorem 5.3.2.3. Local shape preservation by piecewise linear interpolation

Let $s \in C([t_0,t_n])$ be the piecewise linear interpolant of $(t_i,y_i) \in \mathbb{R}^2$, $i = 0,\ldots,n$, for every subinterval $I = [t_j,t_k] \subset [t_0,t_n]$:

if $(t_i,y_i)|_I$ are positive/negative	$\Rightarrow$	*$s|_I$ is positive/negative,*
if $(t_i,y_i)|_I$ are monotonic (increasing/decreasing)	$\Rightarrow$	*$s|_I$ is monotonic (increasing/decreasing),*
if $(t_i,y_i)|_I$ are convex/concave	$\Rightarrow$	*$s|_I$ is convex/concave.*

Local shape preservation = perfect shape preservation!

Bad news: none of this properties carries over to local polynomial interpolation of higher polynomial degree $d > 1$.

EXAMPLE 5.3.2.4 (Piecewise quadratic interpolation) We consider the following generalization of piecewise linear interpolation of data points $(t_j,y_j) \in \mathbb{R} \times \mathbb{R}$, $j = 0,\ldots,n$.

From Thm. 5.2.2.7 we know that a parabola (polynomial of degree 2) is uniquely determined by 3 data points. Thus, the idea is to form groups of three adjacent data points and interpolate each of these triplets by a 2nd-degree polynomial (parabola).

Assume: $n = 2m$ even

piecewise quadratic interpolant $q : [\min\{t_i\},\max\{t_i\}] \mapsto \mathbb{R}$ is defined by

$$q_j := q_{|[t_{2j-2},t_{2j}]} \in \mathcal{P}_2 \quad , \quad q_j(t_i) = y_i \, , \quad i = 2j-2, 2j-1, 2j \, , \quad j = 1,\ldots,m \, . \tag{5.3.2.5}$$

Nodes as in Exp. 5.3.1.6

Piecewise linear (blue) and quadratic (red) interpolants ▷

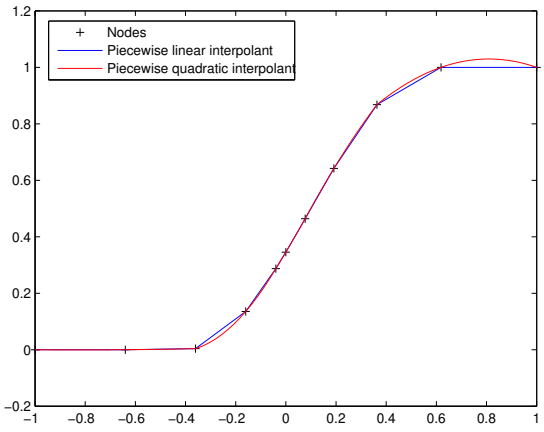


Fig. 165

No shape preservation for piecewise quadratic interpolant

The "only" drawback of piecewise linear interpolation:

interpolant is only C^0 but not C^1 (no continuous derivative).

However: Interpolant usually serves as input for other numerical methods like a Newton-method for solving non-linear systems of equations, see Section 8.5, which requires derivatives.

5.3.3 Cubic Hermite Interpolation

Aim: construct local shape-preserving (→ Section 5.3) (linear ?) interpolation operator that fixes shortcoming of piecewise linear interpolation by ensuring C^1-smoothness of the interpolant.

✎ notation: $C^1([a,b]) \triangleq$ space of *continuously differentiable* functions $[a,b] \mapsto \mathbb{R}$.

5.3.3.1 Definition and Algorithms

Given: mesh points $(t_i, y_i) \in \mathbb{R}^2$, $i = 0,\ldots,n$, $\quad t_0 < t_1 < \cdots < t_n$.

Goal: build function $f \in C^1([t_0,t_n])$ satisfying the interpolation conditions $f(t_i) = y_i$, $i = 0,\ldots,n$.

Definition 5.3.3.1. Cubic Hermite polynomial interpolant

Given data points $(t_j, y_j) \in \mathbb{R} \times \mathbb{R}$, $j = 0,\ldots,n$, with pairwise distinct ordered nodes t_j, and slopes $c_j \in \mathbb{R}$, the piecewise **cubic Hermite interpolant** $s : [t_0,t_n] \to \mathbb{R}$ is defined by the requirements

$$s_{|[t_{i-1},t_i]} \in \mathcal{P}_3\,, \quad i = 1,\ldots,n \quad , \quad \boxed{s(t_i) = y_i} \quad , \quad \boxed{s'(t_i) = c_i} \quad , \quad i = 0,\ldots,n\,.$$

Corollary 5.3.3.2. Smoothness of cubic Hermite polynomial interpolant

Piecewise cubic Hermite interpolants are continuously differentiable on their interval of definition.

Proof. The assertion of the corollary follows from the agreement of function values and first derivative values on nodes shared by two intervals, on each of which the piecewise cubic Hermite interpolant is a polynomial of degree 3. □

§5.3.3.3 (Local representation of piecewise cubic Hermite interpolant) Locally, we can write a piecewise cubic Hermite interpolant as a linear combination of generalized cardinal basis functions with coefficients supplied by the data values y_j and the slopes c_j:

$$s(t) = y_{i-1}H_1(t) + y_iH_2(t) + c_{i-1}H_3(t) + c_iH_4(t) \quad , \quad t \in [t_{i-1},t_i]\,, \tag{5.3.3.4}$$

where the basic functions H_k, $k = 1,2,3,4$, are as follows:

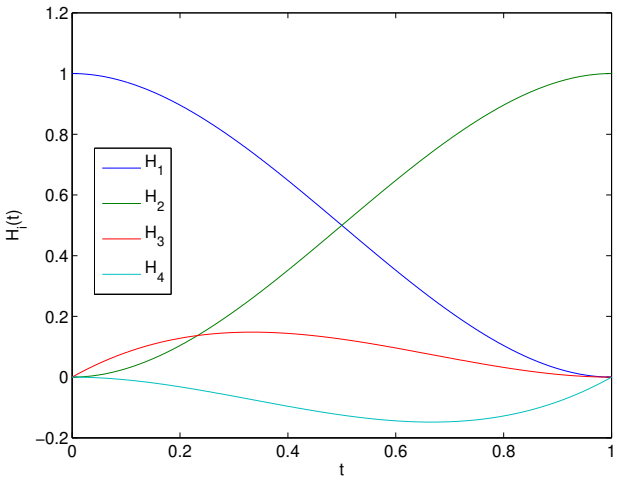


Fig. 166

Local basis polynomials on $[0,1]$

$$H_1(t) := \phi\left(\frac{t_i - t}{h_i}\right)\,, \tag{5.3.3.5a}$$

$$H_2(t) := \phi\left(\frac{t - t_{i-1}}{h_i}\right)\,, \tag{5.3.3.5b}$$

$$H_3(t) := -h_i\psi\left(\frac{t_i - t}{h_i}\right)\,, \tag{5.3.3.5c}$$

$$H_4(t) := h_i\psi\left(\frac{t - t_{i-1}}{h_i}\right)\,, \tag{5.3.3.5d}$$

$$h_i := t_i - t_{i-1}\,, \tag{5.3.3.5e}$$

$$\phi(\tau) := 3\tau^2 - 2\tau^3\,, \tag{5.3.3.5f}$$

$$\psi(\tau) := \tau^3 - \tau^2\,. \tag{5.3.3.5g}$$

By tedious, but straightforward computations using the chain rule we find the following values for H_k and H_k' at the endpoints of the interval $[t_{i-1}, t_i]$.

	$H(t_{i-1})$	$H(t_i)$	$H'(t_{i-1})$	$H'(t_i)$
H_1	1	0	0	0
H_2	0	1	0	0
H_3	0	0	1	0
H_4	0	0	0	1

This amounts to a proof for (5.3.3.4) (why?).

The formula (5.3.3.4) is handy for the local evaluation of piecewise cubic Hermit interpolants. The function `hermloceval()` in Code 5.3.3.6 performs the efficient evaluation (in multiple points) of a piecewise cubic polynomial s on t_1, t_2 uniquely defined by the constraints $s(t_1) = y_1$, $s(t_2) = y_2$, $s'(t_1) = c_1$, $s'(t_2) = c_2$:

C++ code 5.3.3.6: Local evaluation of cubic Hermite polynomial ⇝ GITLAB

```
// Multiple point evaluation of Hermite polynomial
// y1, y2:  data values
// c1, c2:  slopes
Eigen::VectorXd hermloceval(Eigen::VectorXd t, double t1, double t2,
                            double y1, double y2, double c1, double c2) {
  const double h = t2 - t1, a1 = y2 - y1, a2 = a1 - h * c1, a3 = h * c2 - a1 - a2;
  t = ((t.array() - t1) / h).matrix();
  return (y1 + (a1 + (a2 + a3 * t.array()) * (t.array() - 1)) * t.array()).matrix();
}
```

⌟

§5.3.3.7 (Linear Hermite interpolation) However, the data for an interpolation problem (→ Section 5.1) are merely the interpolation points (t_j, y_j), $j = 0, \ldots, n$, but not the slopes of the interpolant at the nodes. Thus, in order to define an interpolation operator into the space of piecewise cubic Hermite functions, we have supply a mapping $\mathbb{R}^{n+1} \times \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ computing the slopes c_j from the data points.

Since this mapping should be local it is natural to rely on (weighted) averages of the local slopes Δ_j (→ Def. 5.3.1.3) of the data, for instance

$$c_i = \begin{cases} \Delta_1 & \text{, for } i = 0 \text{ ,} \\ \Delta_n & \text{, for } i = n \text{ ,} \\ \frac{t_{i+1}-t_i}{t_{i+1}-t_{i-1}}\Delta_i + \frac{t_i - t_{i-1}}{t_{i+1}-t_{i-1}}\Delta_{i+1} & \text{, if } 1 \le i < n \text{ .} \end{cases} \quad , \quad \Delta_j := \frac{y_j - y_{j-1}}{t_j - t_{j-1}} \text{ ,} j = 1, \ldots, n \text{ .} \tag{5.3.3.8}$$

Leads to a linear (→ Def. 5.1.0.25), local Hermite interpolation operator

"Local" means, that, if the values y_j are non-zero for only a few adjacent data points with indices $j = k, \ldots, k+m$, $m \in \mathbb{N}$ small, then the Hermite interpolant s is supported on $[t_{k-\ell}, t_{k+m+\ell}]$ for small $\ell \in \mathbb{N}$ independent of k and m. ⌟

EXAMPLE 5.3.3.9 (Average-based pecewise cubic Hermite interpolation)

Hermite interpolation

Data points:

- 11 equispaced nodes

$$t_j = -1 + 0.2\,j, \quad j = 0, \ldots, 10.$$

in the interval $I = [-1, 1]$,

- $y_i = f(t_i)$ with

$$f(x) := \sin(5x)\, e^x \,.$$

Here we used weighted averages of slopes as in (5.3.3.8).

For details see code `hermintp1.hpp` ↬ GITLAB.

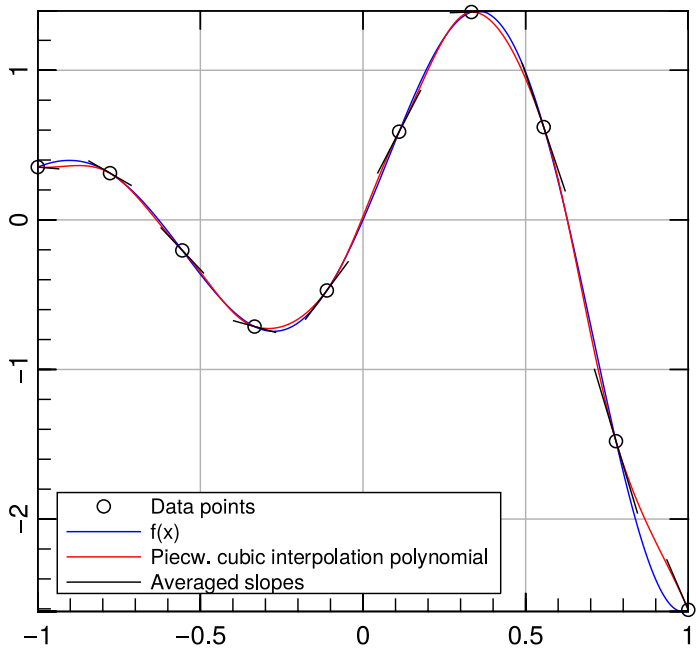


Fig. 167

No strict local/global preservation of monotonicity!

5.3.3.2 Local Monotonicity-Preserving Hermite Interpolation

From Ex. 5.3.3.9 we learn that, if the slopes are chosen according to (5.3.3.8), then the resulting Hermite interpolation does not preserve monotonicity.

Consider the situation sketched on the right ▷

The red circles (●) represent data points, the blue line (—) the piecewise linear interpolant → Section 5.3.2.

In the nodes marked with ↦ the first derivative of a monotonicity preserving C^1-smooth interpolant must vanish! Otherwise an "overshoot" occurs, see also Fig. 165.

Of course, this will be violated, when a (weighted) arithmetic average is used for the computation of slopes for cubic Hermite interpolation.

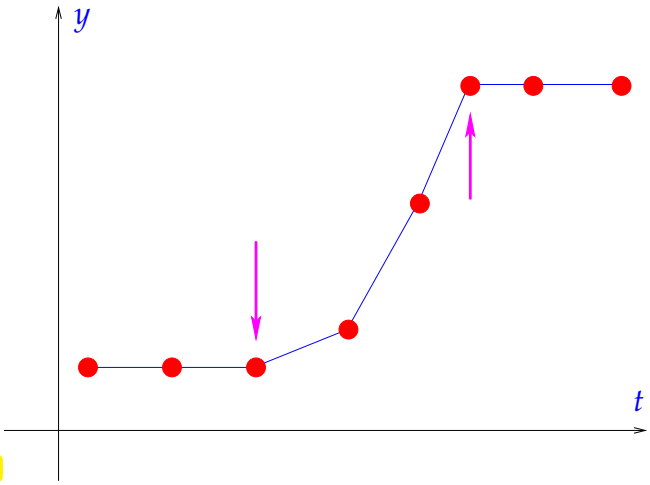


Fig. 168

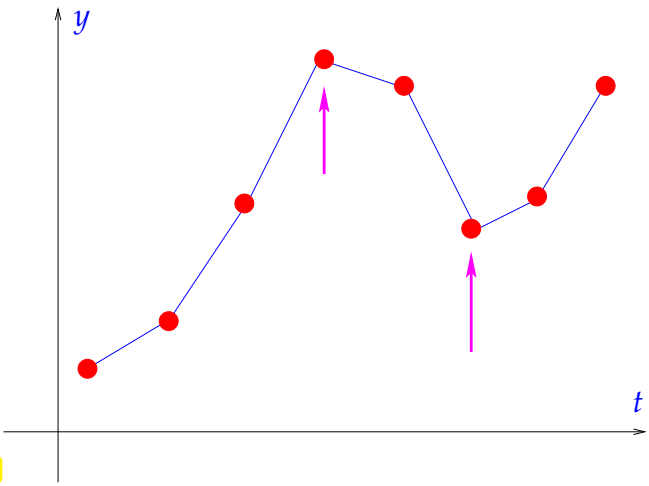


Fig. 169

◁ Consider the situation sketched on the left.

The red circles (●) represent data points, the blue line (—) the piecewise linear interpolant → Section 5.3.2.

A *local* monotonicity preserving C^1-smooth interpolant (→ § 5.3.1.5) s must be flat (= vanishing first derivative) in data points (t_j, y_j), for which

$$y_{j-1} < y_j \quad \text{and} \quad y_{j+1} < y_j \,, \\ y_{j-1} > y_j \quad \text{and} \quad y_{j+1} > y_j \,,$$

in "local extrema" of the data set.

Otherwise, overshoots or undershoots would destroy local monotonicity on one side of the extremum.

§5.3.3.10 (Limiting of local slopes) From the discussion of Fig. 168 and Fig. 169 it is clear that local monotonicity preservation entails that the local slopes c_i of a cubic Hermite interpolant (→ Def. 5.3.3.1) have to fulfill

$$c_i = \begin{cases} 0 & \text{, if } \operatorname{sgn}(\Delta_i) \neq \operatorname{sgn}(\Delta_{i+1}) \,, \\ \text{some "average" of } \Delta_i, \Delta_{i+1} & \text{otherwise} \end{cases} \,, \quad i = 1, \ldots, n-1 \,. \tag{5.3.3.11}$$

✎ notation: sign function $\operatorname{sgn}(\xi) = \begin{cases} 1 & \text{, if } \xi > 0 \,, \\ 0 & \text{, if } \xi = 0 \,, \\ -1 & \text{, if } \xi < 0 \,. \end{cases}$

A slope selection rule that enforces (5.3.3.11) is called a **limiter**.

Of course, testing for equality with zero does not make sense for data that may be affected by measurement or roundoff errors. Thus, the "average" in (5.3.3.11) must be close to zero already when either $\Delta_i \approx 0$ or $\Delta_{i+1} \approx 0$. This is satisfied by the **weighted harmonic mean**

$$c_i = \frac{1}{\frac{w_a}{\Delta_i} + \frac{w_b}{\Delta_{i+1}}} \,, \tag{5.3.3.12}$$

with weights $w_a > 0$, $w_b > 0$, $\quad (w_a + w_b = 1)$.

The harmonic mean = "smoothed $\min(\cdot,\cdot)$-function".

Obviously $\Delta_i \to 0$ or $\Delta_{i+1} \to 0$ in (5.3.3.12), then $c_i \to 0$.

Contour plot of the harmonic mean of a and b ➙ ($w_a = w_b = 1/2$).

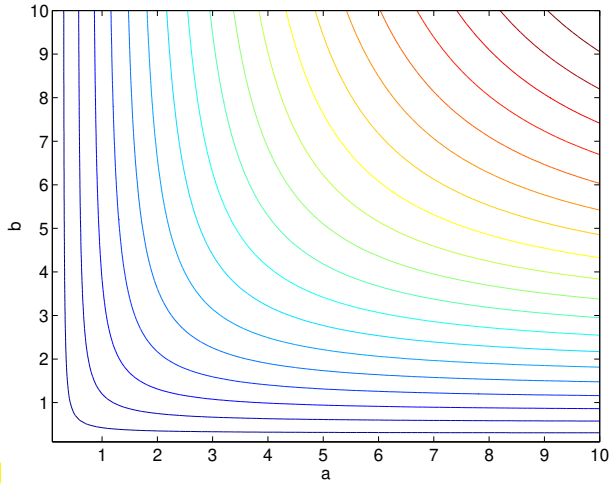


Fig. 170

A good choice of the weights is:

$$w_a = \frac{2h_{i+1} + h_i}{3(h_{i+1} + h_i)} \quad , \quad w_b = \frac{h_{i+1} + 2h_i}{3(h_{i+1} + h_i)} \, ,$$

This yields the following local slopes, unless (5.3.3.11) enforces $c_i = 0$:

$$\overset{\operatorname{sgn}(\Delta_1)=\operatorname{sgn}(\Delta_2)}{\to} \quad c_i = \begin{cases} \Delta_1 & \text{, if } i = 0 \, , \\ \dfrac{3(h_{i+1}+h_i)}{\frac{2h_{i+1}+h_i}{\Delta_i} + \frac{2h_i+h_{i+1}}{\Delta_{i+1}}} & \text{, for } i \in \{1, \ldots, n-1\} \, , \\ \Delta_n & \text{, if } i = n \, , \end{cases} \quad h_i := t_i - t_{i-1} \, . \qquad (5.3.3.13)$$

Piecewise cubic Hermite interpolation with local slopes chosen according to (5.3.3.11) and (5.3.3.13) is available through the MATLAB function/PYTHON class `v = pchip(t,y,x);`/`scipy.interpolate.PchipInterpolator`. The argument `t` passes the interpolation nodes, `y` the corresponding data values, and `x` is a vector of evaluation points. ⌟

EXAMPLE 5.3.3.14 (Monotonicity preserving piecewise cubic polynomial interpolation)

Data from Exp. 5.3.1.6

Plot created with MATLAB-function call

```
v = pchip(t,y,x);
```

- `t`: Data nodes t_j
- `y`: Data values y_j
- `x`: Evaluation points x_i
- `v`: Vector $s(x_i)$

We observe perfect local monotonicity preservation, no under- or overshoots at extrema. ▷

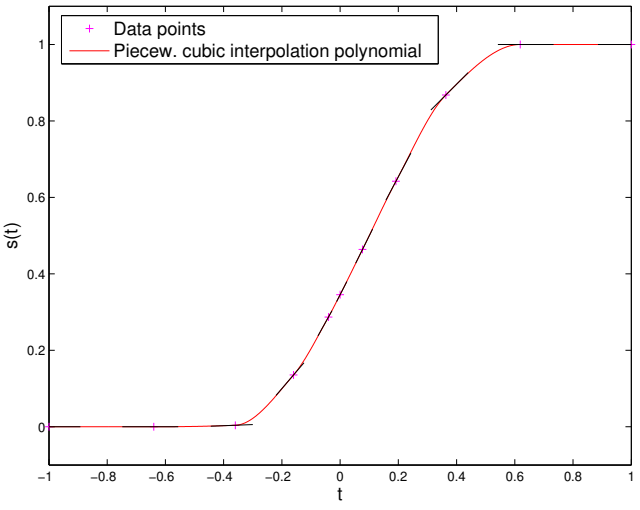


Fig. 171

⌟

Remark 5.3.3.15 (Non-linear cubic Hermite interpolation) Note that the mapping $\mathbf{y} := [y_0, \ldots, y_n] \to c_i$ defined by (5.3.3.11) and (5.3.3.13) is not linear.

➢ The "`pchip` interpolaton operator" does **not** provide a **linear** mapping from data space $\mathbb{R}^{n+1}$ into $C^1([t_0, t_n])$ (in the sense of Def. 5.1.0.25).

In fact, the non-linearity of the piecewise cubic Hermite interpolation operator is necessary for (only global) monotonicity preservation:

Theorem 5.3.3.16. Property of linear, monotonicity preserving interpolation into C^1

If, for fixed node set $\{t_j\}_{j=0}^n$, $n \geq 2$, an interpolation scheme $\mathsf{I} : \mathbb{R}^{n+1} \to C^1(I)$ is linear as a mapping from data values to continuous functions on the interval covered by the nodes ($\to$ Def. 5.1.0.25), and monotonicity preserving, then $\mathsf{I}(\mathbf{y})'(t_j) = 0$ for all $\mathbf{y} \in \mathbb{R}^{n+1}$ and $j = 1, \ldots, n-1$.

Of course, an interpolant that is flat in all data points, as stipulated by Thm. 5.3.3.16 for a lineaer, monotonicity preserving, C^1-smooth interpolation scheme, does not make much sense.

At least, the piecewise cubic Hermite interpolation operator is local (in the sense discussed in § 5.3.3.7).

⌟

Theorem 5.3.3.17. Monotonicity preservation of limited cubic Hermite interpolation

The cubic Hermite interpolation polynomial with slopes as in (5.3.3.13) *provides a* local monotonicity-preserving C^1*-interpolant.*

Proof. See F. Fritsch und R. Carlson, *Monotone piecewise cubic interpolation*, SIAM J. Numer. Anal., 17 (1980), S. 238–246. □

The next code demonstrates the calculation of the slopes c_i in MATLAB's `pchip` (details in [FC80]):

C++ code 5.3.3.18: Monotonicity preserving slopes in pchip ↪ GITLAB

```
1  # include <Eigen/Dense>
2
3  namespace pchipslopes {
4
5  using Eigen::VectorXd;
6
7  // using forward declaration of the function pchipend, implementation
       below
8  double pchipend(double h1, double h2, double del1, double del2);
9
10 inline void pchipslopes(const VectorXd& t, const VectorXd& y, VectorXd& c) {
11   // Calculation of local slopes c_i for shape preserving cubic Hermite
         interpolation, see (5.3.3.11), (5.3.3.13)
12   // t, y are vectors passing the data points
13   const unsigned n = t.size();
14   const VectorXd h = t.tail(n - 1) - t.head(n - 1);
15   const VectorXd delta = (y.tail(n - 1) - y.head(n - 1)).cwiseQuotient(h); // linear
         slopes
16   c = VectorXd::Zero(n);
17
18   // compute reconstruction slope according to (5.3.3.13)
19   for (unsigned i = 0; i < n - 2; ++i) {
20     if (delta(i)*delta(i + 1) > 0) {
21       const double w1 = 2*h(i + 1) + h(i);
22       const double w2 = h(i + 1) + 2*h(i);
23       c(i + 1) = (w1 + w2)/(w1/delta(i) + w2/delta(i + 1));
```

```
    }
  }
  // Special slopes at endpoints, beyond (5.3.3.13)
  c(0) = pchipend(h(0), h(1), delta(0), delta(1));
  c(n - 1) = pchipend(h(n - 2), h(n - 3), delta(n - 2), delta(n - 3));
}

inline double pchipend(const double h1, const double h2, const double del1, const
    double del2) {
  // Non-centered, shape-preserving, three-point formula
  double d = ((2*h1 + h2)*del1 - h1*del2)/(h1 + h2);
  if (d*del1 < 0) {
    d = 0;
  }
  else if (del1*del2 < 0 && std::abs(d) > std::abs(3*del1)) {
    d = 3*del1;
  }
  return d;
}

} //namespace pchipslopes
```

Review question(s) 5.3.3.19 (Shape-preserving interpolation)

(Q5.3.3.19.A) State three shortcomings of global polynomial interpolation, which are remedied by using piecewise linear interpolation.

(Q5.3.3.19.B) Given data points $(t_i, y_i) \in \mathbb{R}^2$ and a continuous interpolant f generated by a local monotonicity preserving interpolation scheme, show that

$$\left\{ f(t),\ t \in [\min_i\{t_i\}, \max_i\{t_i\}] \right\} = [\min_i\{y_i\}, \max_i\{y_i\}] \,.$$

(Q5.3.3.19.C) Show by counterexample that a locally convexity preserving interpolation scheme can generate an interpolant with negative function values even if the data values are all positive.

(Q5.3.3.19.D) Given data points (t_i, y_i), $i = 0, \ldots, n$, $t_{i-} < t_i$, $i = 1, \ldots, n$, we define

$$f(t) = y_0 + \int_{t_0}^{t} p(\tau)\,\mathrm{d}\tau \,,$$

where p is the piecewise linear interpolant of

$$\left(\frac{t_i + t_{i+1}}{2}, \frac{y_{i+1} - y_i}{t_{i+1} - t_i} \right), \quad i = 0, \ldots, n-1 \,.$$

Is the mapping $(t_i, y_i)_i \to f$ a globally monotonicity/convexity preserving?

(Q5.3.3.19.E) Given nodes $t_0 < t_1 < \cdots < t_n$ let $\mathcal{P}_3(\mathcal{T})$ stand for the function space

$$\mathcal{P}_3(\mathcal{T}) := \left\{ f \in C^1([t_0, t_n]) \colon f|_{[t_{i-1}, t_i]} \in \mathcal{P}_3,\ i = 1, \ldots n \right\} .$$

Propose a *linear* interpolation operator

$$\mathsf{F}_{\mathcal{T}} : \mathbb{R}^{n+1} \to \mathcal{P}_3(\mathcal{T}) \quad , \quad (\mathsf{F}_{\mathcal{T}}(\mathbf{y}))(t_j) = (\mathbf{y})_j \,, \quad j = 0, \ldots, n \,,$$

that is locally monotonicity preserving.

(Q5.3.3.19.F) Given an ordered node set $t_0 < t_1 < \cdots < t_n$ and associated data values y_i, $i = 0, \ldots, n$, the `pchip` interpolant is a piecewise cubic Hermite interpolant, with slopes chosen according to the formula

$$\overset{\operatorname{sgn}(\Delta_1)=\operatorname{sgn}(\Delta_2)}{\rightarrow} \quad c_i = \begin{cases} \Delta_1 & \text{, if } i = 0 \text{ ,} \\ \frac{3(h_{i+1}+h_i)}{\frac{2h_{i+1}+h_i}{\Delta_i} + \frac{2h_i+h_{i+1}}{\Delta_{i+1}}} & \text{, for } i \in \{1, \ldots, n-1\} \text{ ,} \\ \Delta_n & \text{, if } i = n \text{ ,} \end{cases} \quad \begin{array}{l} h_i := t_i - t_{i-1} \text{ ,} \\ \Delta_i := \frac{y_i - y_{i-1}}{t_i - t_{i-1}} \text{ .} \end{array} \tag{5.3.3.13}$$

(i) Determine the supports $\operatorname{supp} b_i \subset \mathbb{R}$ of the functions b_i, $i = 0, \ldots, n$, which is the `pchip` interpolant for the data values $y_0 = 0, \ldots, y_{i-1} = 0, y_i = 1, y_{i+1} = 0, \ldots, y_n = 0$.

(ii) Denote by $p(\mathbf{y})$ the `pchip` interpolant for the data vector $\mathbf{y} := [y_0, \ldots, y_n]^\top \in \mathbb{R}^{n+1}$. Can we write p as

$$p(\mathbf{y})(t) = \sum_{i=0}^{n} y_i b_i(t) \text{ ,} \quad t_0 \le t \le t_n \text{ ,}$$

with suitable functions b_i?

△

5.4 Splines

Piecewise cubic Hermite Interpolation presented in Section 5.3.3 entailed determining reconstruction slopes c_i. Now we learn about a way how to do piecewise polynomial interpolation, which results in C^k-interpolants, $k > 0$, and dispenses with auxiliary slopes. The idea is to obtain the missing conditions implicitly from extra continuity conditions, built into spaces of so-called splines. These are of fundamental importance for modern computer-aided geometric design (CAGD).

Supplementary literature. Splines are also presented in [DR08, Ch. 9].

5.4.1 Spline Function Spaces

Definition 5.4.1.1. Spline space → [QSS00, Def. 8.1]

Given an interval $I := [a, b] \subset \mathbb{R}$ and a **knot sequence** $\mathcal{M} := \{a = t_0 < t_1 < \ldots < t_{n-1} < t_n = b\}$, $n \in \mathbb{N}$, the vector space $\mathcal{S}_{d,\mathcal{M}}$ of the **spline functions** of degree d (or order $d+1$), $d \in \mathbb{N}_0$, is defined by

$$\mathcal{S}_{d,\mathcal{M}} := \{s \in C^{d-1}(I) \colon s_j := s_{|[t_{j-1},t_j]} \in \mathcal{P}_d \ \forall j = 1, \ldots, n\} \text{ .}$$

$d-1$-times continuously differentiable — *locally* polynomial of degree d

Do not mix up "knots" = "breakpoints" of a spline functions, and "nodes", the first values in data tuples (t_i, y_i) for 1D interpolation. In the case of spline interpolation, knots may serves as nodes, but not necessarily.

Let's make explicit the spline spaces of the lowest degrees:

- $d = 0$: $\mathcal{M}$-piecewise constant *discontinuous* functions
- $d = 1$: $\mathcal{M}$-piecewise linear *continuous* functions
- $d = 2$: *continuously differentiable* $\mathcal{M}$-piecewise quadratic functions

The dimension of spline space can be found by a counting argument (heuristic): We count the number of "degrees of freedom" (d.o.f.s) possessed by a $\mathcal{M}$-piecewise polynomial of degree d, and subtract the number of linear constraints implicitly contained in Def. 5.4.1.1:

$$\dim \mathcal{S}_{d,\mathcal{M}} = n \cdot \dim \mathcal{P}_d - \#\{C^{d-1} \text{ continuity constraints}\} = n \cdot (d+1) - (n-1) \cdot d = n + d \,.$$

Theorem 5.4.1.2. Dimension of spline space

The space $\mathcal{S}_{d,\mathcal{M}}$ from Def. 5.4.1.1 has dimension

$$\dim \mathcal{S}_{d,\mathcal{M}} = n + d \,.$$

Remark 5.4.1.3 (Differentiating and integrating splines) Obviously, spline spaces are mapped onto each other by differentiation & integration:

$$s \in \mathcal{S}_{d,\mathcal{M}} \quad \Rightarrow \quad s' \in \mathcal{S}_{d-1,\mathcal{M}} \quad \wedge \quad \left\{ t \mapsto \int_a^t s(\tau)\,\mathrm{d}\tau \right\} \in \mathcal{S}_{d+1,\mathcal{M}} \,. \tag{5.4.1.4}$$

⌟

Review question(s) 5.4.1.5 (Spline function spaces)

(Q5.4.1.5.A) Given an (ordered) knot set $\mathcal{M} := \{t_0 < t_1 < \cdots < t_n\} \subset \mathbb{R}$ use a counting argument to determine the dimension of the space of piecewise polynomials

$$\mathcal{S}^k_{d,\mathcal{M}} := \{ s \in C^k(I) \colon s|_{[t_{j-1},t_j]} \in \mathcal{P}_d \; \forall j = 1,\ldots,n \} \,.$$

(Q5.4.1.5.B) Consider the knot set $\mathcal{M} := \{0, \frac{1}{2}, 1\}$. For arbitrary numbers y_0, y_1, c_0, c_1 does there exist $s \in \mathcal{S}_{2,\mathcal{M}}$ such that

$$s(0) = y_0 \,, \quad s(1) = y_1 \,, \quad s'(0) = c_0 \,, \quad s'(1) = c_1 \;\textbf{?}$$

Is s unique?

△

5.4.2 Cubic-Spline Interpolation

We already know the special case of interpolation in $\mathcal{S}_{1,\mathcal{M}}$, when the interpolation nodes are the knots of $\mathcal{M}$, because this boils down to simple piecewise linear interpolation, see Section 5.3.2. Now we explore the interpolation by means of splines of degree $d = 3$.

Supplementary literature. More details can be found in [Han02, pp. XIII, 46], [QSS00, Sect. 8.6.1].

Remark 5.4.2.1 (Perceived smoothness of cubic splines) Cognitive psychology teaches us that the human eye perceives a C^2-function as "smooth", while it can still spot the abrupt change of curvature at the possible discontinuities of the second derivatives of a cube Hermite interpolant (→ Def. 5.3.3.1).

For this reason the simplest spline functions featuring C^2-smoothness are of great importance in computer aided design (CAD). They are the cubic splines, $\mathcal{M}$-piecewise polynomials of degree 3 contained in $\mathcal{S}_{3,\mathcal{M}}$ (→ Def. 5.4.1.1). ⌟

§5.4.2.2 (Cubic spline interpolants) The definition of a cubic spline interpolant is straightforward and matches the abstract concept of an interpolant introduced in Section 5.1. Also note the relationship with Hermite interpolation discussed in Section 5.3.3.

Definition 5.4.2.3. Cubic-spline interpolant

Given a node set/knot set $\mathcal{M} := \{t_0 < t_1 < \cdots < t_n\}$, $n \in \mathbb{N}$, and data values $y_0, \ldots, y_n \in \mathbb{R}$, an associated **cubic spline interpolant** is a function $s \in \mathcal{S}_{3,\mathcal{M}}$ that complies with the interpolation conditions

$$s(t_j) = y_j \quad , \quad j = 0, \ldots, n \, . \tag{5.4.2.4}$$

Note that in the case of cubic spline interpolation the spline knots and interpolation nodes coincide.

From *dimensional considerations* it is clear that the interpolation conditions will fail to fix the interpolating cubic spline uniquely:

$$\dim \mathcal{S}_{3,\mathcal{M}} - \#\{\text{ interpolation conditions}\} = (n+3) - (n+1) = 2 \text{ free d.o.f.}$$

Obviously, "two conditions are missing", which means that the interpolation problem for cubic splines is not well-defined by (5.4.2.4). We have to impose two additional conditions. Different ways to do this will lead to different cubic spline interpolants for the same set of data points. ⌟

§5.4.2.5 (Computing cubic spline interpolants) We opt for a linear interpolation scheme (→ Def. 5.1.0.25) into the spline space $\mathcal{S}_{3,\mathcal{M}}$, which means that the two additional conditions must depend linearly on the data values. As explained in § 5.1.0.21, a linear interpolation scheme will lead to a linear system of equations for expansion coefficients with respect to a suitable basis.

We reuse the local representation of a cubic spline through cubic Hermite cardinal basis polynomials from (5.3.3.5):

(5.3.3.4)

$$\begin{aligned} s_{|[t_{j-1},t_j]}(t) = \; & s(t_{j-1}) \cdot (1 - 3\tau^2 + 2\tau^3) + \\ & s(t_j) \cdot (3\tau^2 - 2\tau^3) + \\ & h_j s'(t_{j-1}) \cdot (\tau - 2\tau^2 + \tau^3) + \\ & h_j s'(t_j) \cdot (-\tau^2 + \tau^3) \, , \end{aligned} \tag{5.4.2.6}$$

with $h_j := t_j - t_{j-1}$, $\tau := (t - t_{j-1})/h_j$.

➢ The task of cubic spline interpolation boils down to finding slopes $s'(t_j)$ in the knots of the mesh $\mathcal{M}$.

Once these slopes are known, the efficient local evaluation of a cubic spline function can be done in the same way as for a cubic Hermite interpolant, see Section 5.3.3.1, Code 5.3.3.6.

Note: if $s(t_j)$, $s'(t_j)$, $j = 0, \ldots, n$, are fixed, then the representation (5.4.2.6) already guarantees $s \in C^1([t_0, t_n])$, *cf.* the discussion for cubic Hermite interpolation, Section 5.3.3.

➢ only the continuity of s''
- ✦ has to be enforced by the choice of slopes $s'(t_j)$
- ⇕
- ✦ will yield extra conditions to fix the slopes $s'(t_j)$

However, do the

✦ interpolation conditions (5.4.2.4) $s(t_j) = y_j$, $j = 0, \ldots, n$, and the
✦ smoothness constraint $s \in C^2([t_0, t_n])$

uniquely determine the unknown slopes $c_j := s'(t_j)$**?** In light of the discussion in § 5.4.2.2, probably not, because we have not yet described the two additional required conditions. Next, we work out the details:

From $s \in C^2([t_0, t_n])$ we obtain $n-1$ continuity constraints for $s''(t)$ at the internal nodes

$$s''_{|[t_{j-1},t_j]}(t_j) = s''_{|[t_j,t_{j+1}]}(t_j)\,, \quad j = 1, \ldots, n-1\,. \tag{5.4.2.7}$$

Based on (5.4.2.6), we express (5.4.2.7) in concrete terms, using

$$\begin{aligned} s''_{|[t_{j-1},t_j]}(t) = {} & s(t_{j-1})h_j^{-2}6(-1+2\tau) + s(t_j)h_j^{-2}6(1-2\tau) \\ & + h_j^{-1}s'(t_{j-1})(-4+6\tau) + h_j^{-1}s'(t_j)(-2+6\tau)\,, \quad \tau := (t-t_{j-1})/h_j\,, \end{aligned} \tag{5.4.2.8}$$

which can be obtained by the chain rule and from $\frac{d\tau}{dt} = h_j^{-1}$.

$$\begin{aligned} \overset{(5.4.2.8)}{\Rightarrow} \quad & s''_{|[t_{j-1},t_j]}(t_{j-1}) = -6\cdot s(t_{j-1})h_j^{-2} + 6\cdot s(t_j)h_j^{-2} - 4\cdot h_j^{-1}s'(t_{j-1}) - 2\cdot h_j^{-1}s'(t_j)\,, \\ & s''_{|[t_{j-1},t_j]}(t_j) = 6\cdot s(t_{j-1})h_j^{-2} + -6\cdot s(t_j)h_j^{-2} + 2\cdot h_j^{-1}s'(t_{j-1}) + 4\cdot h_j^{-1}s'(t_j)\,. \end{aligned}$$

In particular, replacing $j \leftarrow j+1$ in the top equation, yields

$$s''_{|[t_j,t_{j+1}]}(t_j) = -6\cdot s(t_j)h_{j+1}^{-2} + 6\cdot s(t_{j+1})h_{j+1}^{-2} - 4\cdot h_{j+1}^{-1}s'(t_j) - 2\cdot h_{j+1}^{-1}s'(t_{j+1})\,.$$

Inserting these formulas into (5.4.2.7) leads to $n-1$ linear equations for the $n+1$ unknown slopes $c_j := s'(t_j)$. Taking into account the (known) interpolation conditions $s(t_i) = y_i$, we get

$$\frac{1}{h_j}c_{j-1} + \left(\frac{2}{h_j} + \frac{2}{h_{j+1}}\right)c_j + \frac{1}{h_{j+1}}c_{j+1} = 3\left(\frac{y_j - y_{j-1}}{h_j^2} + \frac{y_{j+1}-y_j}{h_{j+1}^2}\right)\,, \tag{5.4.2.9}$$

for $j = 1, \ldots, n-1$.

Actually (5.4.2.9) amounts to an $(n-1)\times(n+1)$, that is, *underdetermined*, linear system of equations. The dimensions make perfect sense, because

$n-1$ equations $\hat{=}$ number of interpolation conditions
$n+1$ unknowns$\hat{=}$ dimension of cubic spline space on knot set $\{t_0 < t_1 < \cdots < t_n\}$

This linear system of equations can be written in matrix-vector form:

$$\begin{bmatrix} b_0 & a_1 & b_1 & 0 & \cdots & & \cdots & 0 \\ 0 & b_1 & a_2 & b_2 & & & & \\ & 0 & \ddots & \ddots & \ddots & & & \vdots \\ \vdots & & & \ddots & \ddots & \ddots & & \\ & & & & \ddots & a_{n-2} & b_{n-2} & 0 \\ 0 & \cdots & & \cdots & 0 & b_{n-2} & a_{n-1} & b_{n-1} \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ \\ \vdots \\ \\ c_{n-1} \\ c_n \end{bmatrix} = \begin{bmatrix} 3\left(\frac{y_1-y_0}{h_1^2} + \frac{y_2-y_1}{h_2^2}\right) \\ \\ \vdots \\ \\ 3\left(\frac{y_{n-1}-y_{n-2}}{h_{n-1}^2} + \frac{y_n-y_{n-1}}{h_n^2}\right) \end{bmatrix}\,. \tag{5.4.2.10}$$

with

$$b_i := \frac{1}{h_{i+1}}\,, \quad a_i := \frac{2}{h_i} + \frac{2}{h_{i+1}}\,, \quad \begin{array}{l} i = 0, 1, \ldots, n-1\,, \\ [\ b_i, a_i > 0 \ , \ a_i = 2(b_i + b_{i-1})\ ]\,. \end{array}$$

➙ two additional constraints are required, as already noted in § 5.4.2.2. ⌟

§5.4.2.11 (Types of cubic spline interpolants) To saturate the remaining two degrees of freedom the following three approaches are popular. All of them involve conditions *linear* in the data values.

① Complete cubic spline interpolation: $s'(t_0) = c_0$, $s'(t_n) = c_n$ prescribed according to

$$c_0 := \mathbf{q}^\top \mathbf{y} \quad , \quad c_1 := \mathbf{p}^\top \mathbf{y} \, ,$$

for given vectors $\mathbf{p}, \mathbf{q} \in \mathbb{R}^{n+1}$ and $\mathbf{y} := [y_0, \ldots, y_n]^\top \in \mathbb{R}^{n+1}$ the vector of data values. Then the first and last column can be removed from the system matrix of (5.4.2.10). Their products with c_0 and c_n, respectively, have to be subtracted from the right hand side of (5.4.2.10), which leads to the $(n-1) \times (n-1)$ linear system of equations:

$$\begin{bmatrix} a_1 & b_1 & 0 & \cdots & \cdots & 0 \\ b_1 & a_2 & b_2 & & & \\ 0 & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & 0 \\ & & & \ddots & a_{n-2} & b_{n-2} \\ 0 & \cdots & \cdots & 0 & b_{n-2} & a_{n-1} \end{bmatrix} \begin{bmatrix} c_1 \\ \vdots \\ c_{n-1} \end{bmatrix} = \begin{bmatrix} 3\left(\frac{y_1-y_0}{h_1^2} + \frac{y_2-y_1}{h_2^2}\right) - c_0 b_0 \\ \vdots \\ 3\left(\frac{y_{n-1}-y_{n-2}}{h_{n-1}^2} + \frac{y_n-y_{n-1}}{h_n^2}\right) - c_n b_{n-1} \end{bmatrix} . \quad (5.4.2.12)$$

② Natural cubic spline interpolation: $s''(t_0) = s''(t_n) = 0$

$$\tfrac{2}{h_1} c_0 + \tfrac{1}{h_1} c_1 = 3 \frac{y_1 - y_0}{h_1^2} \quad , \quad \tfrac{1}{h_n} c_{n-1} + \tfrac{2}{h_n} c_n = 3 \frac{y_n - y_{n-1}}{h_n^2} \, .$$

Combining these two extra equations with (5.4.2.10), we arrive at a linear system of equations with *tridiagonal s.p.d.* ($\to$ Def. 1.1.2.6, Lemma 2.8.0.13) system matrix and unknowns $c_0, \ldots, c_n$:

$$\begin{bmatrix} 2 & 1 & 0 & \cdots & & & \cdots & 0 \\ b_0 & a_1 & b_1 & 0 & \cdots & & \cdots & 0 \\ 0 & b_1 & a_2 & b_2 & & & & \\ & 0 & \ddots & \ddots & \ddots & & & \vdots \\ \vdots & & & \ddots & \ddots & \ddots & & \\ & & & & \ddots & a_{n-2} & b_{n-2} & 0 \\ 0 & \cdots & & \cdots & 0 & b_{n-2} & a_{n-1} & b_{n-1} \\ 0 & \cdots & & & \cdots & 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} c_0 \\ \vdots \\ c_n \end{bmatrix} = \begin{bmatrix} 3\frac{y_1-y_0}{h_1} \\ 3\left(\frac{y_1-y_0}{h_1^2} + \frac{y_2-y_1}{h_2^2}\right) \\ \vdots \\ 3\left(\frac{y_{n-1}-y_{n-2}}{h_{n-1}^2} + \frac{y_n-y_{n-1}}{h_n^2}\right) \\ 3\frac{y_n-y_{n-1}}{h_n} \end{bmatrix} . \quad (5.4.2.13)$$

Owing to Thm. 2.7.5.4 this linear system of equations can be solved with an asymptotic computational effort of $O(n)$ for $n \to \infty$.

③ Periodic cubic spline interpolation: $s'(t_0) = s'(t_n)$ ($\succ$ $c_0 = c_n$), $s''(t_0) = s''(t_n)$

This removes one unknown and adds another equations so that we end up with an $n \times n$-linear system with s.p.d. ($\to$ Def. 1.1.2.6) system matrix

$$\mathbf{A} := \begin{bmatrix} a_1 & b_1 & 0 & \cdots & 0 & b_0 \\ b_1 & a_2 & b_2 & & & 0 \\ 0 & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & 0 \\ 0 & & & \ddots & a_{n-1} & b_{n-1} \\ b_0 & 0 & \cdots & 0 & b_{n-1} & a_0 \end{bmatrix} , \quad \begin{aligned} b_i &:= \tfrac{1}{h_{i+1}}, \quad i = 0, 1, \ldots, n-1 \, , \\ a_i &:= \tfrac{2}{h_i} + \tfrac{2}{h_{i+1}}, \quad i = 0, 1, \ldots, n-1 \, . \end{aligned}$$

This linear system can be solved with rank-1-modifications techniques (see § 2.6.0.12, Lemma 2.6.0.21) + tridiagonal elimination: asymptotic computational effort $O(n)$.

Remark 5.4.2.14 (Piecewise cubic interpolation schemes) Let us review three different classes of interpolation schemes relying on piecewise cubic polynomials with respect to a prescribed node set:

- Piecewise cubic local Lagrange interpolation
  - ➢ Extra degrees of freedom fixed by putting four nodes in one interval
  - ➥ yields merely C^0-interpolant; perfectly local.
- Cubic Hermite interpolation
  - ➢ Extra degrees of freedom fixed by locally reconstructed slopes, e.g. (5.3.3.13)
  - ➥ yields C^1-interpolant; still local.
- Cubic spline interpolation
  - ➢ Extra degrees of freedom fixed by C^2-smoothness, complete/natural/periodic constraint.
  - ➥ yields C^2-interpolant; non-local.

EXAMPLE 5.4.2.15 (Control of a robotic arm [Han02, Sect. 46])

Given times $\mathcal{M} := \{t_0 < t_1 < \cdots < t_n\}$, $n \in \mathbb{N}$, the tip of a robotic arm is to visit the point in space $\mathbf{y}_i \in \mathbb{R}^3$ at time t_i. At initial time t_0 and final time t_n the arm is to be at rest.

Fig. 172

The path of the tip of the robotic arm can conveniently be described by the the componentwise complete cubic spline interpolant $\mathbf{s} : [t_0, t_n] \to \mathbb{R}^3$ satisfying

$$(\mathbf{s})_j \in \mathcal{S}_{3,\mathcal{M}} \,, \quad j = 1,2,3 \quad , \quad \mathbf{s}(t_i) = \mathbf{y}_i \,, \quad i = 0,\ldots,n \quad , \quad \mathbf{s}'(t_0) = \mathbf{s}'(t_n) = \mathbf{0} \,.$$

Note that the first derivative $t \mapsto \mathbf{s}'(t)$ is the velocity of the tip, while the second derivative $t \mapsto \mathbf{s}''$ describes its acceleration. Hence, $\mathbf{s} \in C^2([t_0, t_n], \mathbb{R}^3)$ ensures that the acceleration does not change abruptly, which is important for curbing wear on the joints of the robotic arm.

Review question(s) 5.4.2.16 (Cubic spline interpolation)

(Q5.4.2.16.A) What do cubic spline interpolants and piecewise cubic Hermite interpolants have in common and in which respects are they different?

(Q5.4.2.16.B) Given the node set $\{t_0 < t_1 < \cdots < t_n\} \subset \mathbb{R}$ assume that the data values are obtained by sampling a polynomial $p : \mathbb{R} \to \mathbb{R}$, $I := [t_0, t_n]$: $y_i := p(t_i)$, $i = 0,\ldots,n$. Characterize the maximal vector space of polynomials p such that $p \equiv s$, where s is the natural cubic spline interpolant of $(t_i, y_i)_{i=0}^n$. What is its dimension?

Remember that, beside the interpolation conditions $s(t_i) = y_i$ the natural cubic spline interpolant also satisfies $s''(t_0) = s''(t_n) = 0$.

(Q5.4.2.16.C) Characterize the rank-1 modification of

$$\mathbf{A} := \begin{bmatrix} a_1 & b_1 & 0 & \cdots & 0 & b_0 \\ b_1 & a_2 & b_2 & & & 0 \\ 0 & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & 0 \\ 0 & & & \ddots & a_{n-1} & b_{n-1} \\ b_0 & 0 & \cdots & 0 & b_{n-1} & a_0 \end{bmatrix} \in \mathbb{R}^{n,n}$$

that turns it into a genuine tri-diagonal matrix.

(Q5.4.2.16.D) The node/knot set $\mathcal{M} := \{t_0 < t_1 < \cdots < t_n\}$, $n \geq 3$, is given. Beside the interpolation conditions $s(t_i) = y_i$ for data values y_i, $i = 0, \ldots, n$, a complete cubic spline interpolant $s \in \mathcal{S}_{3,\mathcal{M}}$ is determined by the additional two conditions

$$s'(t_0) = \alpha_0 y_0 + \alpha_1 y_1 + \alpha_2 y_2 + \alpha_3 y_3 \quad , \quad s'(t_n) = \beta_0 y_n + \beta_1 y_{n-1} + \beta_2 y_{n-2} + \beta_3 y_{n-3} \, ,$$

for given weights $\alpha_j, \beta_j \in \mathbb{R}$.

Find weights $\alpha_0, \ldots, \alpha_4$ and $\beta_1, \ldots, \beta_4$ such that in the case $y_i = p(t_i)$ for some cubic polynomial $p \in \mathcal{P}_3$ one always finds $s \equiv p$ on $[t_0, t_n]$.

(Q5.4.2.16.E) Find cubic polynomials b_1, b_2, b_3, b_4 such that

$$\begin{array}{llll} b_1(0) = 1 \, , & b_1(1) = 0; \, , & b_1''(0) = 0 \, , & b_1''(1) = 0 \, , \\ b_2(0) = 0 \, , & b_1(2) = 1; \, , & b_2''(0) = 0 \, , & b_2''(1) = 0 \, , \\ b_3(0) = 0 \, , & b_1(3) = 0; \, , & b_3''(0) = 1 \, , & b_3''(1) = 0 \, , \\ b_4(0) = 0 \, , & b_1(4) = 0; \, , & b_4''(0) = 0 \, , & b_4''(1) = 1 \, . \end{array}$$

(Q5.4.2.16.F) For the computation of a cubic spline interpolant in § 5.4.2.5 we used a local, that is, inside the knot intervals, representation of the spline function by means of cardinal basis polynomials for Hermite interpolation

$$s_{|[t_{j-1},t_j]}(t) = a_j(1 - 3\tau^2 + 2\tau^3) + b_j(3\tau^2 - 2\tau^3) + c_j(\tau - 2\tau^2 + \tau^3) + d_j(-\tau^2 + \tau^3) \, , \quad (5.4.2.17)$$

($h_j := t_j - t_{j-1}$, $\tau := (t - t_{j-1})/h_j$) with coefficients $a_j, b_j, c_j, d_j \in \mathbb{R}$.

Explain the big benefits derived from using (5.4.2.17).

(Q5.4.2.16.G) This question continues Question (**Q**5.4.2.16.F). Why is using the local representation

$$s_{|[t_{j-1},t_j]}(t) = a_j t^3 + b_j t^2 + c_j t + d_j \, , \quad a_j, b_j, c_j, d_j \in \mathbb{R} \, ,$$

instead of (5.4.2.17) as starting point for the computation of a cubic spline interpolant not a good idea?

△

5.4.3 Structural Properties of Cubic Spline Interpolants

§5.4.3.1 (Extremal properties of natural cubic spline interpolants → [QSS00, Sect. 8.6.1, Property 8.2]) Splines are special! For a function $f : [a, b] \mapsto \mathbb{R}$, $f \in C^2([a, b])$, the term

$$E_{\mathrm{bd}}(f) := \tfrac{1}{2} \int_a^b |f''(t)|^2 \, \mathrm{d}t \, , \quad (5.4.3.2)$$

models the elastic **bending energy** of a rod, whose shape is described by the graph of f (Soundness check: zero bending energy for straight rod). We will show that cubic spline interpolants have minimal bending energy among all C^2-smooth interpolating functions.

Theorem 5.4.3.3. Optimality of natural cubic spline interpolant

Given a knot/nodes set $\mathcal{M} := \{a = t_0 < t_1 < \cdots < t_n = b\}$ in the interval $[a,b]$, let $s \in \mathcal{S}_{3,\mathcal{M}}$ be the natural cubic spline interpolant of the data points $(t_i, y_i) \in \mathbb{R}^2$, $i = 0,\ldots,n$.
Then s minimizes the elastic bending energy among all interpolating functions in $C^2([a,b])$, that is

$$E_{\mathrm{bd}}(s) \le E_{\mathrm{bd}}(f) \quad \forall f \in C^2([a,b]),\ f(t_i) = y_i,\ i = 0,\ldots,n\,.$$

Idea of proof: variational calculus

We show that any small perturbation of s such that the perturbed spline still satisfies the interpolation conditions leads to an increase in elastic energy.

Pick *perturbation direction* $k \in C^2([t_0,t_n])$ satisfying $k(t_i) = 0$, $i = 0,\ldots,n$:

$$\begin{aligned} E_{\mathrm{bd}}(s+k) &= \tfrac{1}{2}\int_a^b |s'' + k''|^2 \,\mathrm{d}t \\ &= E_{\mathrm{bd}}(s) + \underbrace{\int_a^b s''(t)k''(t)\,\mathrm{d}t}_{:=I} + \underbrace{\tfrac{1}{2}\int_a^b |k''|^2\,\mathrm{d}t}_{\ge 0}\,. \end{aligned} \tag{5.4.3.4}$$

We scrutinize I, split it into the contributions of individual knot intervals, integrate by parts twice, and use $s^{(4)} \equiv 0$. By $s'''(t_j^{\pm})$ we denote the limits of the discontinuous third derivative $t \mapsto s'''(t)$ from the left ($-$) and right ($+$) of the knot t_j.

$$\begin{aligned} I &= \sum_{j=1}^n \int_{t_{j-1}}^{t_j} s''(t)k''(t)\,\mathrm{d}t \\ &= \sum_{j=1}^n \left\{ -\int_{t_{j-1}}^{t_j} s'''(t)k'(t)\,\mathrm{d}t + s''(t_j)k'(t_j) - s''(t_{j-1})k'(t_{j-1}) \right\} \\ &= \sum_{j=1}^n \left\{ \int_{t_{j-1}}^{t_j} s^{(4)}(t)k(t)\,\mathrm{d}t + s''(t_j)k'(t_j) - s''(t_{j-1})k'(t_j) - s'''(t_j^-)k(t_j) + s'''(t_{j-1}^+)k(t_{j-1}) \right\} \\ &= -\sum_{j=1}^n \Big(s'''(t_j^-)\underbrace{k(t_j)}_{=0} - s'''(t_{j-1}^+)\underbrace{k(t_{j-1})}_{=0} \Big) + \underbrace{s''(t_n)}_{=0}k'(t_n) - \underbrace{s''(t_0)}_{=0}k'(t_0) = 0\,. \end{aligned}$$

In light of (5.4.3.4): no perturbation k compatible with interpolation conditions can make the bending energy of s decrease! ⌟

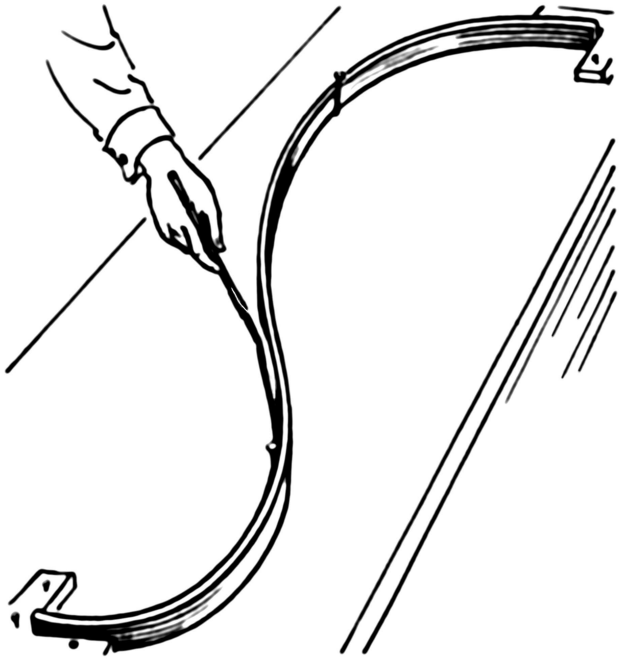
Remark 5.4.3.5 (Origin of the term "Spline")

§ 5.4.3.1: (Natural) cubic spline interpolant provides C^2-curve of minimal elastic bending energy that travels through prescribed points.

Nature: A thin elastic rod fixed a certain points attains a shape that minimizes its potential bending energy (virtual work principle of statics).

Cubic spline interpolation approximates shape of elastic rods. Such rods were in fact used in the manufacturing of ship hulls as "analog computers" for "interpolating points" that were specified by the designer of the ship.

Cubic spline interpolation before the age of computers ▷



⌟

Remark 5.4.3.6 (Shape preservation of cubic spline interpolation)

Data $s(t_j) = y_j$ from Exp. 5.3.1.6 and

$$s'(t_0) = c_0 := \frac{y_1 - y_0}{t_1 - t_0}\,,$$
$$s'(t_n) = c_n := \frac{y_n - y_{n-1}}{t_n - t_{n-1}}\,.$$

The cubic spline interpolant is nor monotonicity nor curvature preserving

This is not surprising in light of Thm. 5.3.3.16, because cubic spline interpolation is a linear interpolation scheme.

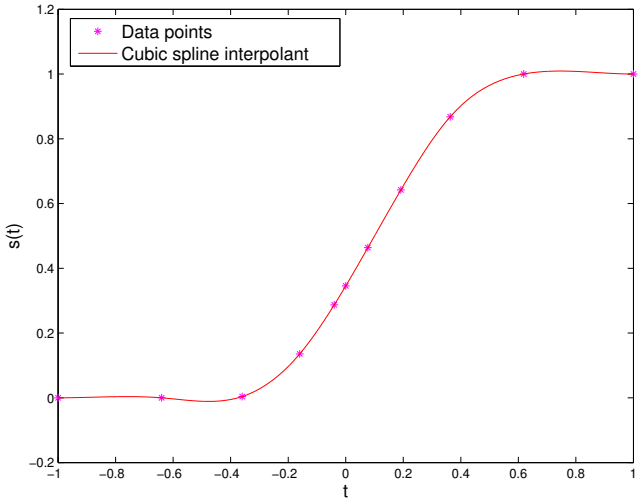


Fig. 173

⌟

§5.4.3.7 (Weak locality of an interpolation scheme) In terms of locality of interpolation schemes, in the sense of § 5.3.3.7, we habe seen:

- Piecewise linear interpolation (→ Section 5.3.2) is strictly local: Changing a single data value y_j affects the interpolant only on the interval $]t_{j-1}, t_{j+1}[$.
- Monotonicity preserving piecewise cubic Hermite interpolation (→ Section 5.3.3.2) is still local, because changing y_j will lead to a change in the interpolant only in $]t_{j-2}, t_{j+2}[$ (the remote intervals are affected through the averaging of local slopes).
- Polynomial Lagrange interpolation is highly non-local, see Ex. 5.2.4.3.

We can weaken the notion of locality of an interpolation scheme on an ordered node set $\{t_i\}_{i=0}^n$:

➢ (weak) locality measures the impact of a perturbation of a data value y_j at points $t \in [t_0, t_n]$ as a function of $|t - t_i|$.

➢ an interpolation scheme is weakly local, if the impact of the perturbation of y_i displays a rapid (e.g. exponential) decay as $|t - t_i|$ increases.

For a *linear* interpolation scheme (→ § 5.1.0.21) locality can be deduced from the decay of the cardinal interpolants/cardinal basis functions (→ Lagrange polynomials of § 5.2.2.3), that is, the functions $b_j := \mathsf{I}(\mathbf{e}_j)$, where $\mathbf{e}_j$ is the j-th unit vector, and I the interpolation operator. Then weak locality can be quantified as

$$\exists \lambda > 0 | : \quad |b_j(t)| \leq b_j(t_j) \cdot \exp(-\lambda |t - t_j|) \, , \quad t \in [t_0, t_n] \, . \tag{5.4.3.8}$$

Remember:

- Lagrange polynomials satisfying (5.2.2.4) provide cardinal interpolants for polynomial interpolation → § 5.2.2.3. As is clear from Fig. 155, they do not display any decay away from their "base node". Rather, they grow strongly. Hence, there is no locality in global polynomial interpolation.
- Tent functions (→ Fig. 154) are the cardinal basis functions for piecewise linear interpolation, see Ex. 5.1.0.15. Hence, this scheme is perfectly local, see (5.3.2.2).

Given a knot/node set $\mathcal{M} := \{t_0 < t_1 < \cdots < t_n\}$ the ith natural cubic cardinal spline is defined by the conditions

$$L_i \in \mathcal{S}_{3,\mathcal{M}}, \quad L_i(t_j) = \delta_{ij} \, , \quad L_i''(t_0) = L_i''(t_n) = 0 \, . \tag{5.4.3.9}$$

These functions will supply a cardinal basis of $\mathcal{S}_{3,\mathcal{M}}$ and, according to (5.1.0.18), for natural cubic spline interpolants we have the formula

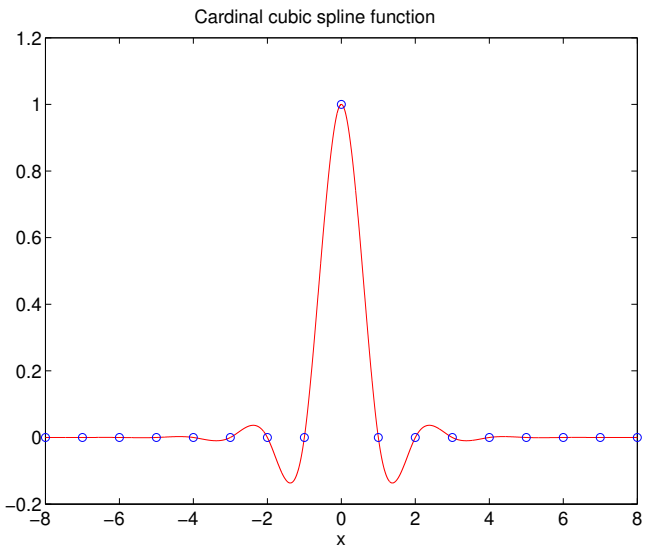
$$\blacktriangleright \quad \text{Natural spline interpolant:} \quad s(t) = \sum_{j=0}^{n} y_j L_j(t) \, .$$

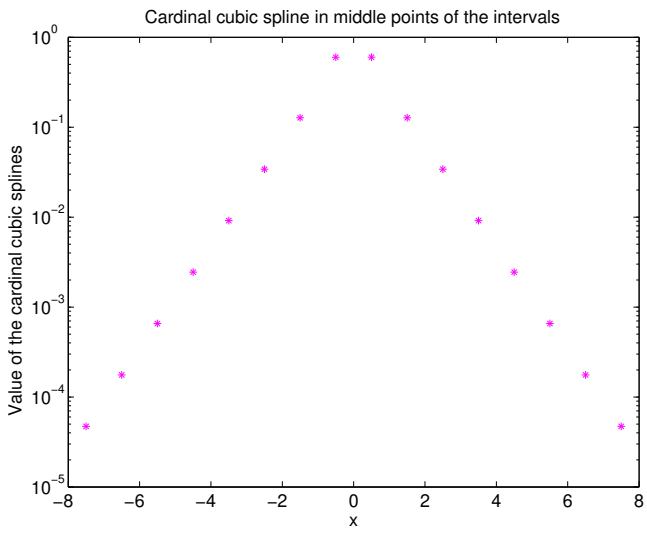
This means that $t \mapsto L_i(t)$ completely characterizes the impact that a change of the value y_i has on s. A very rapid decay of L_i means that the value y_i does not influence s much outside a small neighborhood of t_i.

Fast (exponential) decay of L_i ↔ weak locality of natural cubic spline interpolation.

⌟

EXPERIMENT 5.4.3.10 (Decay of cardinal basis functions for natural cubic spline interpolation) We examine the cardinal basis of $\mathcal{S}_{3,\mathcal{M}}$ ("cardinal splines") associated with natural cubic spline interpolation on an equidistant knot set $\mathcal{M} := \{t_j = j - 8\}_{j=0}^{16}$:





We observe a rapid *exponential decay* of the cardinal splines, which is expressed by the statement that "cubic spline interpolation is weakly local". ⌟

Review question(s) 5.4.3.11 (Structural properties of cubic spline interpolants)

(Q5.4.3.11.A) Given data points $(t_i, y_i) \in \mathbb{R}^2$, $t_0 < t_1 < \cdots < t_n$, $n \in \mathbb{N}$, show that the piecewise linear interpolant $p \in C^0([t_0, t_1])$ minimizes the elastic energy

$$E_{\mathtt{el}}(f) := \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} |f'(t)|^2 \,\mathrm{d}t$$

among all continuous, piecewise continuously differentiable interpolants.

(Q5.4.3.11.B) Given data points (t_i, y_i), $i = 0, \ldots, n$, $t_{i-1} < t_i$, $y_i \in \mathbb{R}$, show that the solution f^* of the minimization problem

$$f^* := \operatorname*{argmin}_{f \in C^2[t_0, t_n]} E_m(f) \quad , \quad E_m(f) := \sum_{i=0}^{n} |f(t_i) - y_i|^2 + \int_{t_0}^{t_n} |f''(t)|^2 \,\mathrm{d}t \,.$$

is a cubic spline (with respect to the knot set $\mathcal{M} := \{t_0 < t_1 < \cdots < t_n\}$), which satisfies $s''(t_0) = s''(t_n) = 0$.

Of course, you should invoke the result about the optimality of the natural cubic spline interpolant.

Theorem Thm. 5.4.3.3. Optimality of natural cubic spline interpolant

Given a knot/nodes set $\mathcal{M} := \{a = t_0 < t_1 < \cdots < t_n = b\}$ in the interval $[a, b]$, let $s \in \mathcal{S}_{3,\mathcal{M}}$ be the natural cubic spline interpolant of the data points $(t_i, y_i) \in \mathbb{R}^2$, $i = 0, \ldots, n$.
Then s minimizes the elastic bending energy among all interpolating functions in $C^2([a, b])$, that is

$$E_{\mathrm{bd}}(s) \leq E_{\mathrm{bd}}(f) \quad \forall f \in C^2([a, b]),\ f(t_i) = y_i,\ i = 0, \ldots, n \,.$$

△

5.4.4 Shape Preserving Spline Interpolation

According to Rem. 5.4.3.6 cubic spline interpolation is neither monotonicity preserving nor curvature preserving. Necessarily so according to Thm. 5.3.3.16, because it is a *linear* interpolation scheme producing a continuously differentiable interpolant with slopes $\neq 0$ in the nodes.

This section presents a non-linear *quadratic spline* (→ Def. 5.4.1.1, C^1-functions) based interpolation scheme that preserves both monotonicity and curvature of data even in a local sense, *cf.* Section 5.3, § 5.3.1.5. The construction was presented in [Sch83], is fairly intricate and will be presented step by step.

As with all 1D data interpolation tasks we are given data points $(t_i, y_i) \in \mathbb{R}^2$, $i = 0, \ldots, n$, and we assume that their first coordinates are ordered: $t_0 < t_1 < \cdots < t_n$.

In order to obtain extra flexibility required for shape preservation, the key idea of [Sch83] is to use an *extended* knot set $\mathcal{M} \subset [t_0, t_n]$ containing n additional knots beside $\{t_j\}_{j=0}^{n}$.

Then we construct an interpolating quadratic spline function that satisfies $s \in \mathcal{S}_{2,\mathcal{M}}$, $s(t_i) = y_i$, $i = 0, \ldots, n$ and locally preserves the "shape" of the data in the sense of § 5.3.1.1.

We stress that unlike in the case of cubic spline interpolation the knot set will not agree with the node set in this case $\mathcal{M} \neq \{t_j\}_{j=0}^n$: the interpolant s interpolates the data in the points t_i but is piecewise polynomial with respect to $\mathcal{M}$!

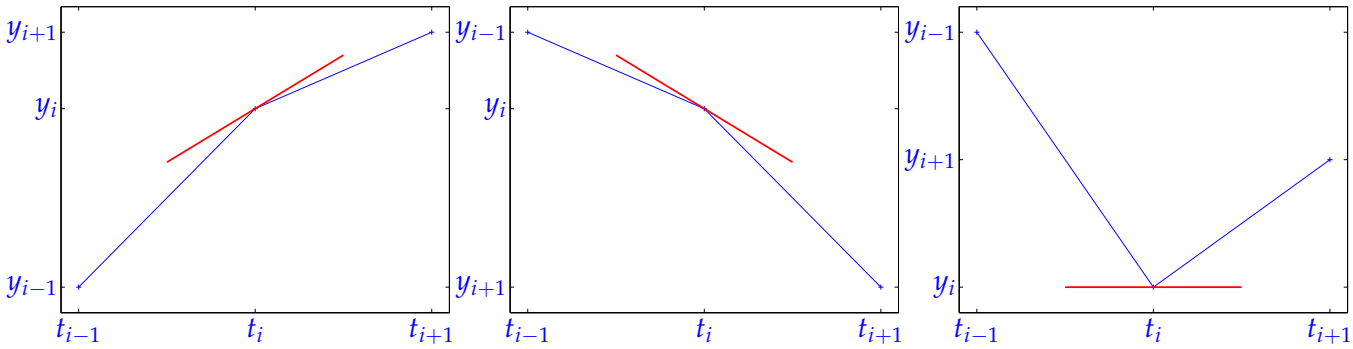
We proceed in four steps:

① Shape preserving choice of slopes c_i, $i = 0, \ldots, n$, as in Section 5.3.3.2, [MR81; QQY01]:

Recall (5.3.3.11) and (5.3.3.13): we fix the slopes c_i in the nodes using the harmonic mean of data slopes Δ_j, the final interpolant will be tangent to these segments in the points (t_i, y_i). If (t_i, y_i) is a local maximum or minimum of the data, then apply slope limiting: c_j is set to zero ($\to$ § 5.3.3.10):

$$\text{Limiter} \quad c_i := \begin{cases} \dfrac{2}{\Delta_i^{-1} + \Delta_{i+1}^{-1}} & \text{, if } \operatorname{sign}(\Delta_i) = \operatorname{sign}(\Delta_{i+1}) \text{ ,} \\ 0 & \text{otherwise,} \end{cases} \qquad i = 1, \ldots, n-1 \text{ .}$$
$$c_0 := 2\Delta_1 - c_1 \text{ ,} \qquad c_n := 2\Delta_n - c_{n-1} \text{ ,} \tag{5.4.4.1}$$

where $\Delta_j := \frac{y_j - y_{j-1}}{t_j - t_{j-1}}$ designates the local data slopes.



Figures: slopes according to limited harmonic mean formula

② Choice of "extra knots" $p_i \in]t_{i-1}, t_i]$, $i = 1, \ldots, n$:

Rule:

Let T_i be the unique straight line through (t_i, y_i) with slope c_i; — in figure ▷

If the intersection of T_{i-1} and T_i is non-empty and has a t-coordinate $\in]t_{i-1}, t_i]$,

☞ then $p_i :=$ t-coordinate of $T_{i-1} \cap T_i$,

☞ otherwise $p_i = \frac{1}{2}(t_{i-1} + t_i)$.

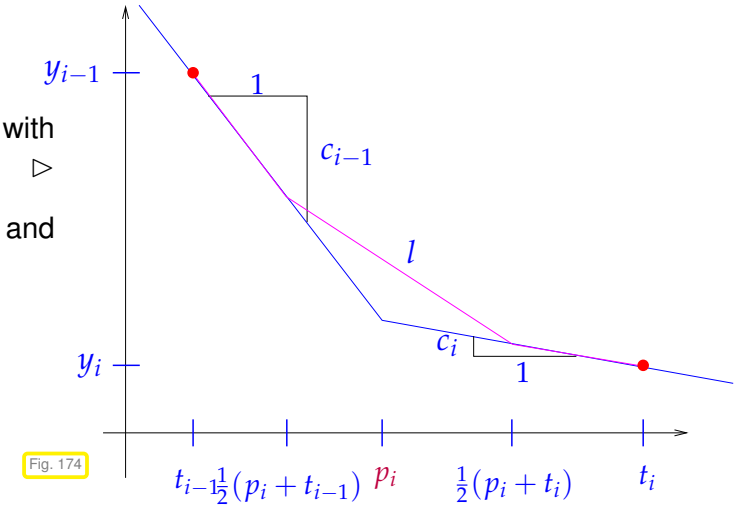


Fig. 174

These points will be used to build the knot set for the final quadratic spline:

$$\mathcal{M} = \{t_0 < p_1 \leq t_1 < p_2 \leq \cdots < p_n \leq t_n\} \text{ .} \tag{5.4.4.2}$$

The following code implements the construction of the auxiliary knots

C++ code 5.4.4.3: Selection of auxiliary knots p_j ↪ GITLAB

```
Eigen::VectorXd extra_knots(const Eigen::VectorXd& t, const Eigen::VectorXd &y,
                            const Eigen::VectorXd& c) {
  const unsigned n = t.size();
  assert((n == y.size()) && (n == c.size()));
  Eigen::VectorXd p = (t[n - 1] + 1.0)*Eigen::VectorXd::Ones(n - 1);
  for (unsigned j = 0; j < n - 1; ++j) {
    // Safe comparison, just to avoid division by zero exception
    // If both slopes a close p_j will certainly lie outside the node interval
    if (c[j] != c[j + 1]) {
      p[j] = (y[j+1] - y[j] + t[j]*c[j] - t[j+1]*c[j+1]) / (c[j+1]- c[j]);
    }
    if (p[j] < t[j] || p[j] > t[j+1]) {
      p[j] = 0.5*(t[j] + t[j+1]);
    }
  }
  return p;
}
```

③ Construction of auxiliary polygonal line:

We chose L to be the linear ($d = 1$) spline (polygonal line) on the knot set $\mathcal{M}'$ of midpoints of the knot intervals of $\mathcal{M}$ from (5.4.4.2)

$$\mathcal{M}' = \{t_0 < \tfrac{1}{2}(t_0 + p_1) < \tfrac{1}{2}(p_1 + t_1) < \tfrac{1}{2}(t_1 + p_2) < \cdots < \tfrac{1}{2}(t_{n-1} + p_n) < \tfrac{1}{2}(p_n + t_n) < t_n\} ,$$

satisfying the conditions

$$L(t_i) = y_i \quad , \quad L'(t_i) = c_i . \tag{5.4.4.4}$$

- In each interval $(\frac{1}{2}(p_j + t_j), \frac{1}{2}(t_j + p_{j+1}))$ the linear spline L corresponds to the line segment of slope c_j passing through the data point (t_j, y_j).
- In each interval $(\frac{1}{2}(t_j + p_{j+1}), \frac{1}{2}(p_{j+1} + t_{j+1}))$ the linear spline L corresponds to the line segment connecting the ones on the other knot intervals of $\mathcal{M}'$, see Fig. 174.

Given the choice of slopes according to (5.4.4.1), a detailed analysis shows the following shpae-preserving property of L.

L "inherits" local monotonicity and curvature from the data.

EXAMPLE 5.4.4.5 (Auxiliary construction for shape preserving quadratic spline interpolation) We verify the above statements about the polygonal line L for the data points $(j, \cos(j))_{j=0}^{12}$, which are marked as • in the left plot.

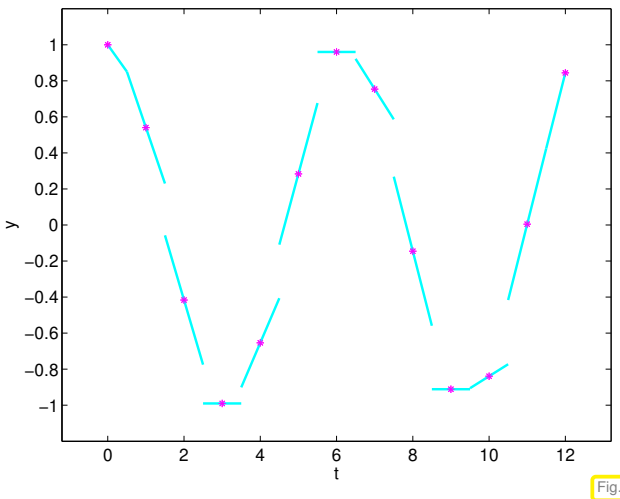


Fig. 175

Local slopes c_i, $i = 0, \ldots, n$

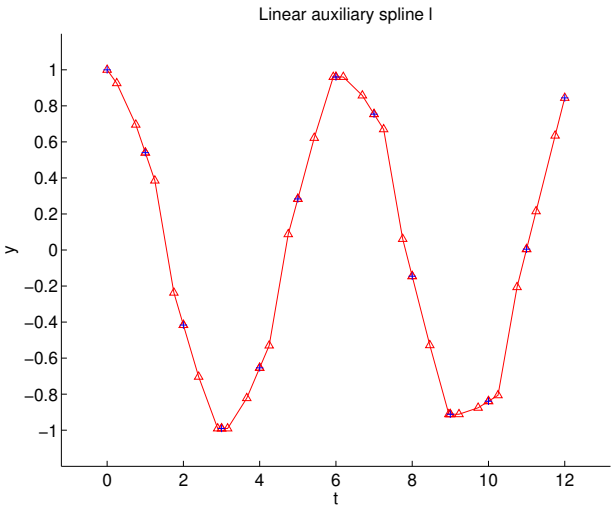


Fig. 176

Auxiliary linear spline L

The reader is encouraged to check by inspection of the plots that the piecewise linear function L locally preserves monotonicity and curvature. ⌟

④ Local quadratic approximation / interpolation of L:

Tedious, but elementary calculus, confirms the following fact:

Lemma 5.4.4.6.

If g is a linear spline (polygonal line) through the three points

$$(a, y_a)\ ,\ (\frac{1}{2}(a+b), w)\ ,\ (b, y_b) \quad \textit{with} \quad a < b\ ,\quad y_a, y_b, w \in \mathbb{R}\ ,$$

then the parabola

$$p(t) := (y_a(b-t)^2 + 2w(t-a)(b-t) + y_b(t-a)^2)/(b-a)^2\ ,\quad a \le t \le b\ ,$$

satisfies

1. $p(a) = y_a \quad , \quad p(b) = y_b \quad , \quad p'(a) = g'(a) \quad , \quad p'(b) = g'(b)$,
2. *g monotonic increasing / decreasing* $\Rightarrow$ *p monotonic increasing / decreasing,*
3. *g convex / concave* $\Rightarrow$ *p convex / concave.*

The proof boils down to discussing many cases as indiated in the following plots:

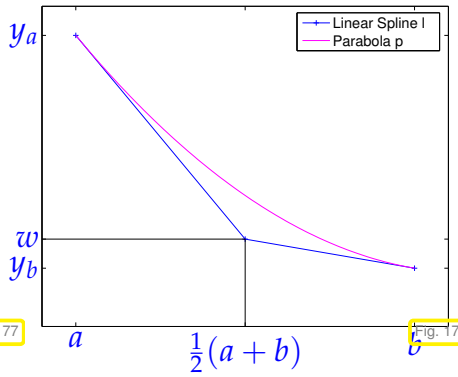


Fig. 177

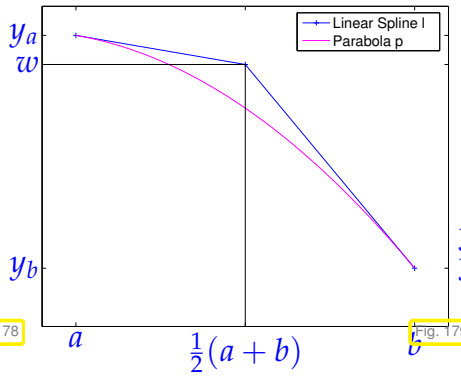


Fig. 178

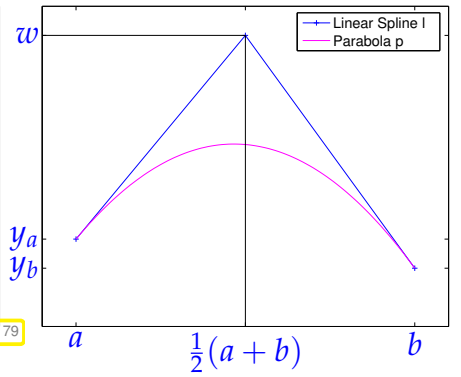


Fig. 179

Lemma 5.4.4.6 implies that the final quadratic spline that passes through the points (t_j, y_j) with slopes c_j can be built locally as the parabola p using the linear spline L that plays the role of g in the lemma.

EXAMPLE 5.4.4.7 (Ex. 5.4.4.5 cnt'd)

We use the same data points as in Ex. 5.4.4.5 and build a quadratic spline $s \in \mathcal{S}_{2,\mathcal{M}}$, $\mathcal{M}$ the knot set from (5.4.4.2), based on the formula suggested by Lemma 5.4.4.6.

— $\hat{=}$ interpolating quadratic spline s ➞

Local shape preservation is evident.

However, since $s \notin C^2([t_0, t_n])$ in general, we see "kinks", *cf.* Rem. 5.4.2.1.

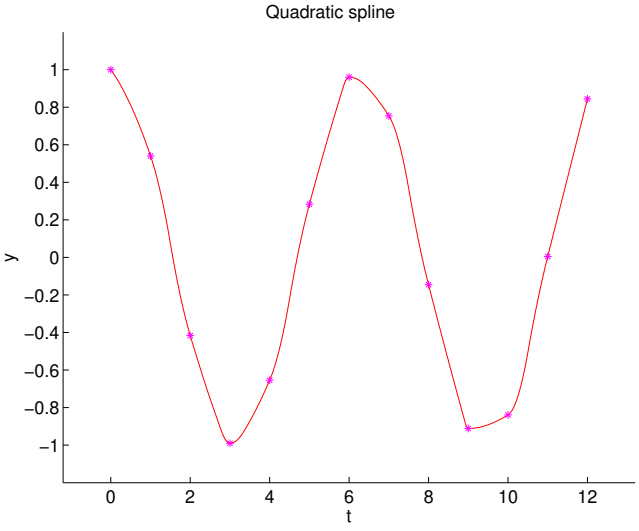


Fig. 180

EXAMPLE 5.4.4.8 (Cardinal shape preserving quadratic spline) We examine the shape preserving quadratic spline that interpolates data values $y_j = 0$, $j \neq i$, $y_i = 1$, $i \in \{0, \ldots, n\}$, on an equidistant node set.

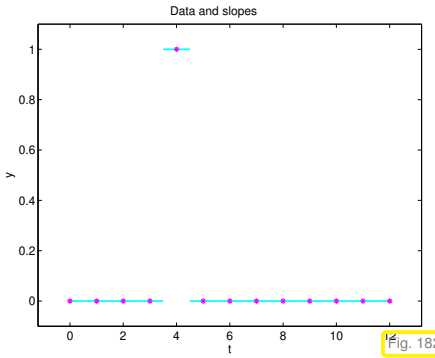


Fig. 181

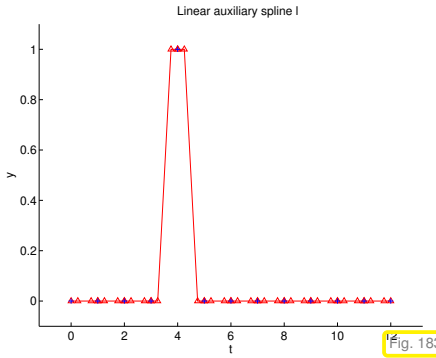


Fig. 182

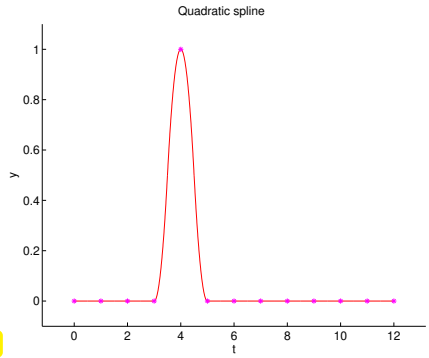


Fig. 183

Shape preserving quadratic spline interpolation is a local, **but** not a linear interpolation scheme.

EXAMPLE 5.4.4.9 (Shape preserving quadratic spline interpolation)

We reuse the data from Exp. 5.3.1.6:

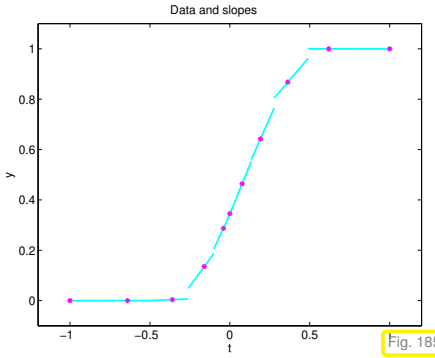


Fig. 184

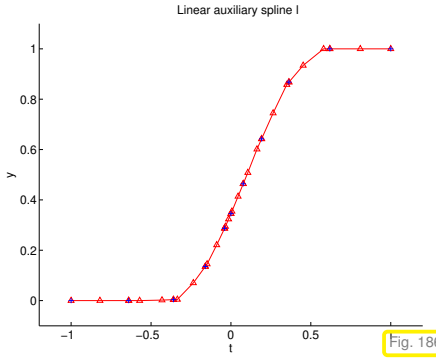


Fig. 185

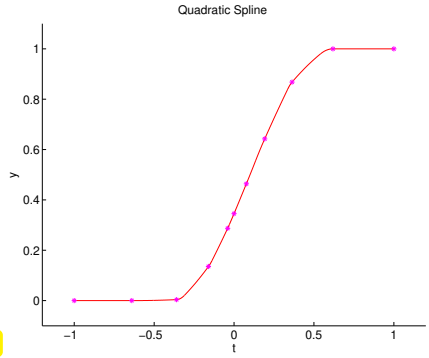


Fig. 186

Data from [MR81]:

t_i	0	1	2	3	4	5	6	7	8	9	10	11	12
y_i	0	0.3	0.5	0.2	0.6	1.2	1.3	1	1	1	1	0	-1

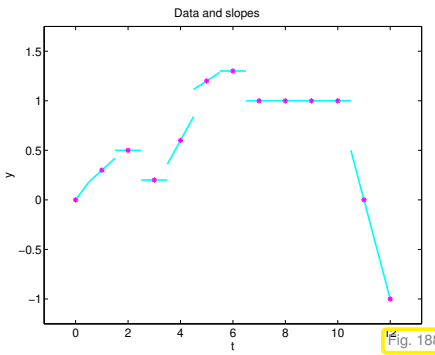


Fig. 187

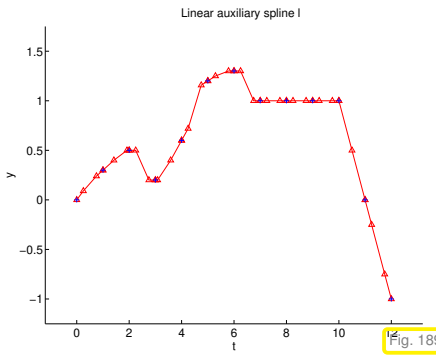


Fig. 188

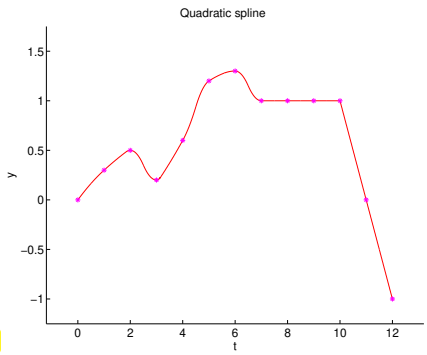


Fig. 189

In all cases we observe excellent local shape preservation. The implementation can be found in `shapepresintp.hpp` ➥ GITLAB. ⌟

Review question(s) 5.4.4.10 (Shape-preserving spline interpolation)

(Q5.4.4.10.A) Argue why an interpolant of the three data points $(0,0),(1,0),(2,1),(3,1)$ in $\mathcal{S}_{2,\mathcal{M}}$, $\mathcal{M} := \{0,1,2,3\}$, *cannot* be monotonicity preserving.

(Q5.4.4.10.B) Let $\mathsf{S} : I^{n+1} \times \mathbb{R}^{n+1} \to C^1(I)$, $I \subset \mathbb{R}$ a bounded interval, be an interpolation scheme, that is, given data points (t_i, y_i), $t_i \in I$, $t_i \neq t_j$ for $i \neq j$, $i,j = 0,\ldots,n$, $f := \mathsf{S}([t_i]_{i=0}^n, [y_i]_{i=0}^n)$ is a continuously differentiable function $f : I \to \mathbb{R}$ satisfying the interpolation conditions $f(t_i) = y_i$, $i = 0,\ldots,n$. Assume that S is *locally monotonicity-preserving*.

What is the support of $\mathsf{S}([t_i]_{i=0}^n, \mathbf{e}_k)$, if the nodes t_i are sorted, $t_0 < t_1 < t_2 < \cdots < t_n$ and $\mathbf{e}_k$ stands for the Cartesian coordinate vector $\left[\delta_{kj}\right]_{j=0}^n \in \mathbb{R}^{n+1}$?

△

5.5 Algorithms for Curve Design

In this section we study a peculiar task of "data fitting in 1D", which arises in **C**omputer **A**ided **D**esign (**CAD**).

5.5.1 CAD Task: Curves from Control Points

Data: Using tablet and pen a designer jots down $n+1$, $n \in \mathbb{N}$, ordered **control points** in the plane, whose coordinate vectors we denote by $\mathbf{p}_\ell \in \mathbb{R}^2$, $\ell \in \{0,\ldots,n\}$.

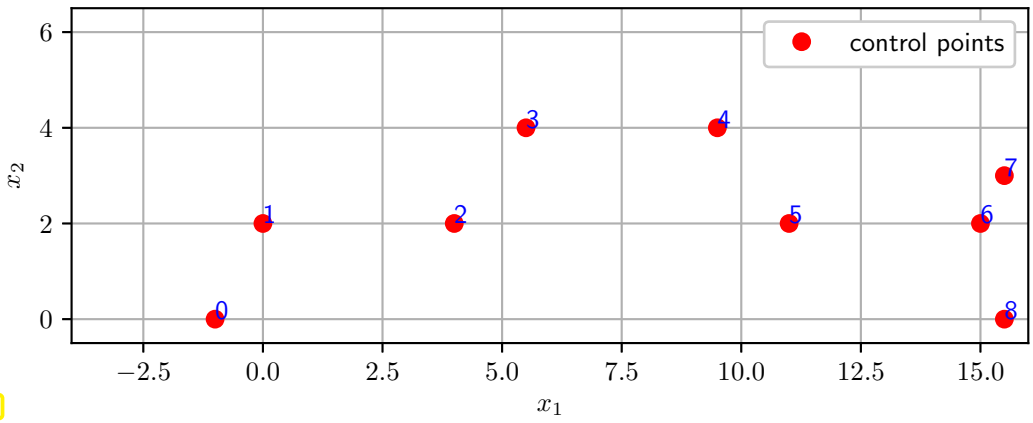


Fig. 190

◁ Ordered control points

Task: Find an algorithm that "builds" a **smooth curve**, whose shape "is inspired" by the locations of the control points.

Let us give a more concrete meaning to some aspects of this rather vague specification.

§5.5.1.2 (Parameterized curves)

Definition 5.5.1.3. Planar curve

A **planar curve** $\mathcal{C}$ is a subset of $\mathbb{R}^2$ such that

$$\mathcal{C} = \{\mathbf{c}(t) : t \in [a,b]\} ,$$

with a *continuous* function $\mathbf{c} : [a,b] \to \mathbb{R}^2$ ($a,b \in \mathbb{R}$, $b > a$), called a **parameterization**.

Note that the parameterization of a curve is *not unique*.

EXAMPLE 5.5.1.4 (A parameterized curve)

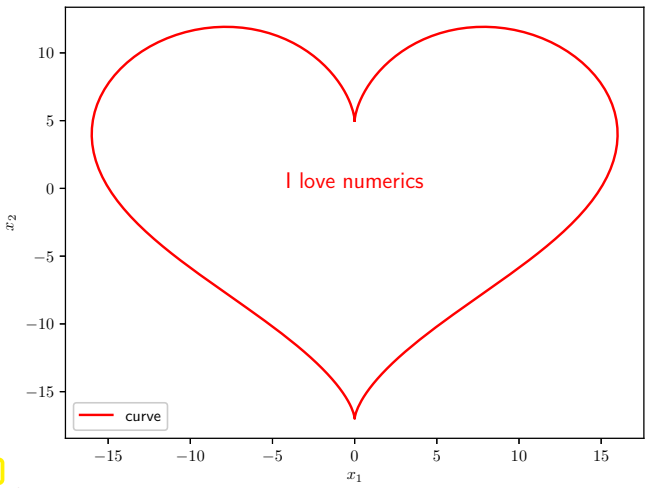


Fig. 191

The heart-shaped curve is described by the parameterization

$$\mathbf{c}(t) = \begin{bmatrix} 16\sin^3 t \\ 13\cos t - 5\cos(2t) - 2\cos(3t) - \cos(4t) \end{bmatrix}, \quad t \in [0, 2\pi] .$$

This is an example of a parameterization by a trigonometric polynomial.

Thus, "building" a curve amounts to finding a parameterization. In algorithms, we have to choose the paramterization from sets of *simple functions*, whose elements can be characterized by finitely many, more precisely, a few, real numbers (unfortunately, often also called "parameters").

EXAMPLE 5.5.1.5 (Polynomial planar curves) We call a planar curve (→ Def. 5.5.1.3) **polynomial** of degree $d \in \mathbb{N}_0$, if it has a parameterization $\mathbf{c} : [a,b] \to \mathbb{R}^2$of the form

$$\mathbf{c}(t) = \sum_{k=0}^{d} \mathbf{a}_k t^k \quad \text{with vectors} \quad \mathbf{a}_k \in \mathbb{R}^2,\ k \in \{0,\dots,d\} . \tag{5.5.1.6}$$

We also write $\mathbf{c} \in (\mathcal{P}_d)^2$ to indicate that $\mathbf{c}$ is polynomial of degree d. For $d = 1$ we recover a line segment with endpoints $\mathbf{c}(a)$ and $\mathbf{c}(b)$.

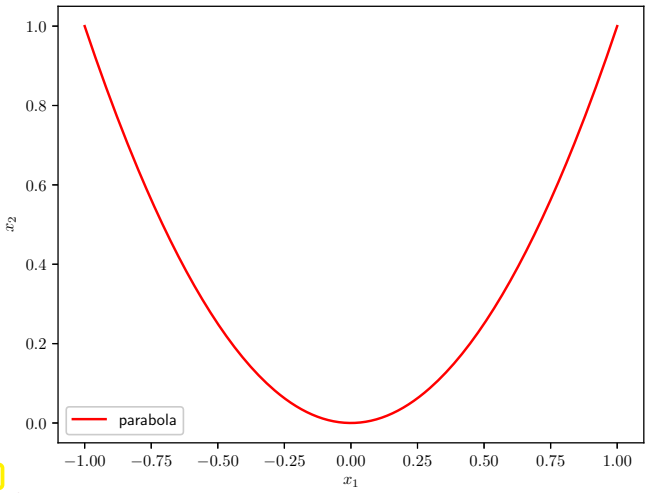


Fig. 192

◁ a parabola over $[-1,1]$.

Every finite section of a graph of a polynomial can be regarded as a polynomial curve.

§5.5.1.7 (Smooth curves) The "smoothness" of a curve is directly related to the *differentiability class* of its parameterization. As remarked in the beginning of Section 5.4.2, a curve is "visually smooth", if its parameterization $\mathbf{c} : [a,b] \to \mathbb{R}^2$ is twice continuously differentiable: $\mathbf{c} \in C^2([a,b]) \times C^2([a,b]) =: (C^2([a,b]))^2$. Hence, the polygon connecting the control points is not an acceptable curve, because it is merely continuous and has kinks.

The meaning of "shape fidelity" cannot be captured rigorously, but some aspects of it will be discussed below.

EXAMPLE 5.5.1.8 (Curves from interpolation) After what we have learned in Section 5.2, Section 5.3, and Section 5.3.3, it is a natural idea to try and construct the parameterization of a curve by interpolating the control points $\mathbf{p}_\ell$, $\ell \in \{0,\ldots,n\}$ using an established scheme for 1D interpolation of the data points $(t_\ell, \mathbf{p}_\ell)$ with mutually distinct nodes $t_\ell \in \mathbb{R}$. However how should we choose the t_ℓ?

The idea is to extract the nodes t_ℓ from the accumulated length of the segments of the polygon with corners $\mathbf{p}_\ell$:

$$t_0 := 0 \quad , \quad t_\ell := \sum_{k=1}^{\ell} \|\mathbf{p}_k - \mathbf{p}_{k-1}\| \, , \quad \ell \in \{1,\ldots,n\} \, . \tag{5.5.1.9}$$

This choice of the nodes together with three different interpolation schemes was used to generate curves based on a "car-shaped" sequence of control points displayed in Fig. 190.

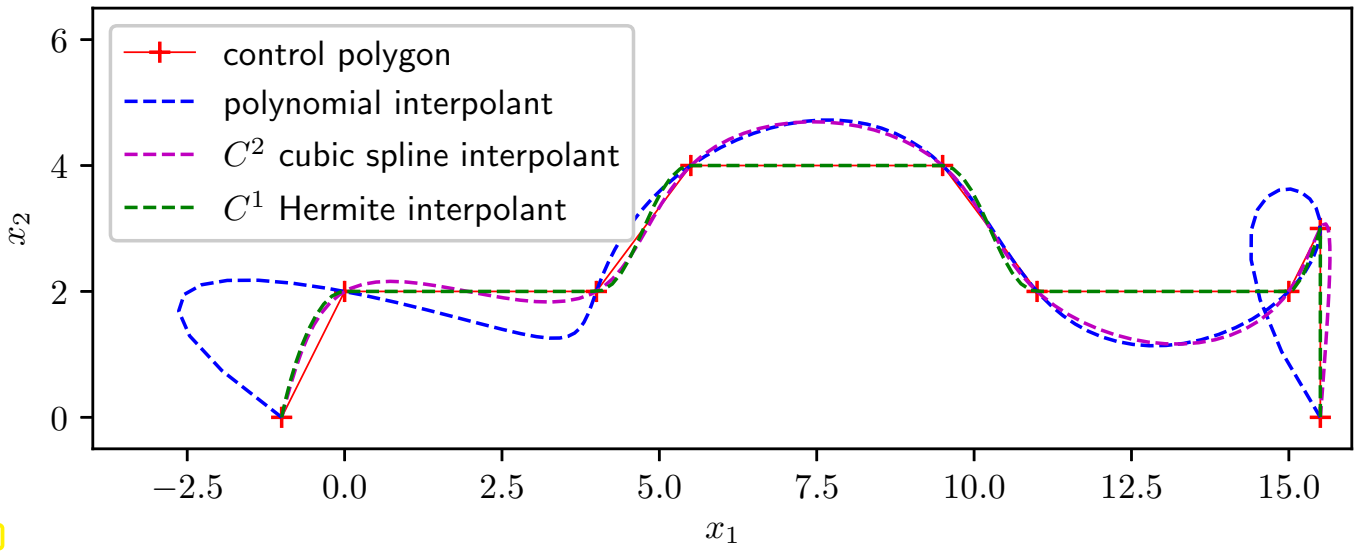


Fig. 193

Obviously, this is not what a designer wants to obtain:

- The polynomial interpolant oscillates wildly, a typical observation with high-degree polynomial interpolants, see Exp. 5.3.1.6.
- The cubic spline interpolant's "shape fidelity" is found wanting.
- The Hermite interpolant lacks smoothness; it is merely of class C^1.

5.5.2 Bezier Curves

We abandon the idea of interpolating the control points $\mathbf{p}_\ell$, $\ell \in \{0,\ldots,n\}$, but restrict ourselves to the class of polynomial curves as introduced in Ex. 5.5.1.5. More precisely, we aim for a polynomial curve *of degree* n, the same degree as required for a polynomial interpolating $n+1$ data points, recall Thm. 5.2.2.7. For the polynomial curve to be constructed, the so-called **Bezier curve**, we write $\mathbf{b}^k : [0,1] \to \mathbb{R}^2$. As parameter interval we use $[0,1]$ throughout in this section.

Idea: Construct $\mathbf{b}^n$ *recursively* by *convex combination*.

Definition 5.5.2.1. Convex combination

A **convex combination** of elements $\mathbf{v}_1,\ldots,\mathbf{v}_m$, $m \in \mathbb{N}$, of a real vector space V is a linear combination

$$\xi_1\mathbf{v}_1 + \xi_2\mathbf{v}_2 + \cdots + \xi_m\mathbf{v}_m \in V \;,$$

with coefficients satisfying

$$0 \le \xi_k \le 1 \;, \quad k = 1,\ldots,m \quad , \quad \xi_1 + \xi_2 + \cdots + \xi_m = 1 \;.$$

To explain the construction we write $\mathbf{b}_i^k : [0,1] \to \mathbb{R}^2$, $k = 1,\ldots,n$, $i = 0,\ldots,n-k$ for the polynomial curves of degree k that are Bezier curves for the sub-sequence $(\mathbf{p}_i,\ldots,\mathbf{p}_{i+k})$ of control points. We define *recursively* (w.r.t. to the degree k) by two-term convex combinations

$$\begin{aligned} &\mathbf{b}_i^1(t) := \mathbf{p}_i(1-t) + \mathbf{p}_{i+1}t \;, && i = 0,\ldots,n-1 \;, \\ &\mathbf{b}_i^k(t) := t\,\mathbf{b}_{i+1}^{k-1}(t) + (1-t)\mathbf{b}_i^{k-1}(t) \;, && \begin{array}{l} i = 0,\ldots,n-k \;, \\ k = 2,\ldots,n \;, \end{array} \end{aligned} \qquad t \in [0,1] \;. \tag{5.5.2.2}$$

Finally, we set

$$\mathbf{b}^n(t) := \mathbf{b}_0^n(t) \;.$$

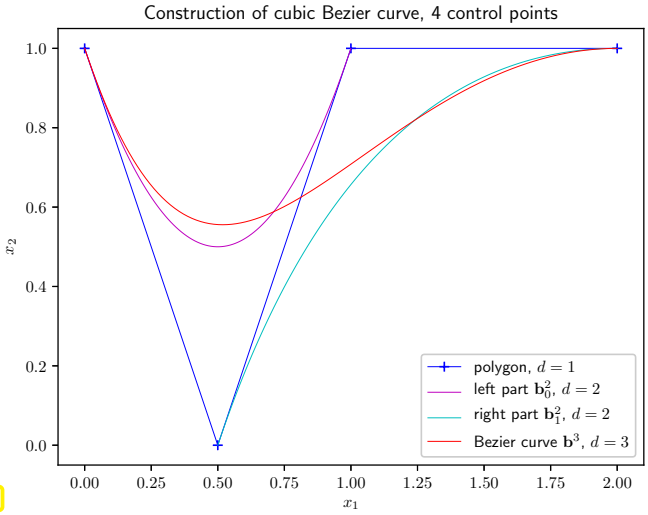


Fig. 194

The constuction is illustrated for $n = 3$:

$$\begin{aligned}
\mathbf{b}_0^1(t) &= (1-t)\mathbf{p}_0 + t\mathbf{p}_1 \,, \\
\mathbf{b}_1^1(t) &= (1-t)\mathbf{p}_1 + t\mathbf{p}_2 \,, \\
\mathbf{b}_2^1(t) &= (1-t)\mathbf{p}_2 + t\mathbf{p}_3 \,, \\
\mathbf{b}_0^2(t) &= (1-t)\mathbf{b}_0^1(t) + t\mathbf{b}_1^1(t) \,, \\
\mathbf{b}_1^2(t) &= (1-t)\mathbf{b}_1^1(t) + t\mathbf{b}_2^1(t) \,, \\
\mathbf{b}^3(t) &= (1-t)\mathbf{b}_0^2(t) + t\mathbf{b}_1^2(t) \,.
\end{aligned}$$

A recursive definition suggests proofs by induction and a simple one yields the insight that Bezier curves connect the first and last control point.

Lemma 5.5.2.3.

Given mutually distinct points $\mathbf{p}_0, \ldots, \mathbf{p}_n$, the functions $\mathbf{b}_i^k$, $k \in \{1, \ldots, n\}$, $i \in \{0, \ldots, n-k\}$, defined by (5.5.2.2) *satisfy*

$$\mathbf{b}_i^k(0) = \mathbf{p}_i \quad , \quad \mathbf{b}_i^k(1) = \mathbf{p}_{i+k} \,.$$

Concerning shape-fidelity the following concept is important:

Definition 5.5.2.4. Convex hull

The **convex hull** of $m \in \mathbb{N}$ elements $\mathbf{v}_1, \ldots, \mathbf{v}_m$ of a real vector space V is the set of all of their convex combinations (→ Def. 5.5.2.1):

$$\operatorname{convex}\{\mathbf{v}_1, \ldots, \mathbf{v}_m\} := \left\{ \sum_{k=1}^{m} \xi_k \mathbf{v}_k : \; 0 \le \xi_k \le 1, \; \sum_{k=1}^{m} \xi_k = 1 \right\} .$$

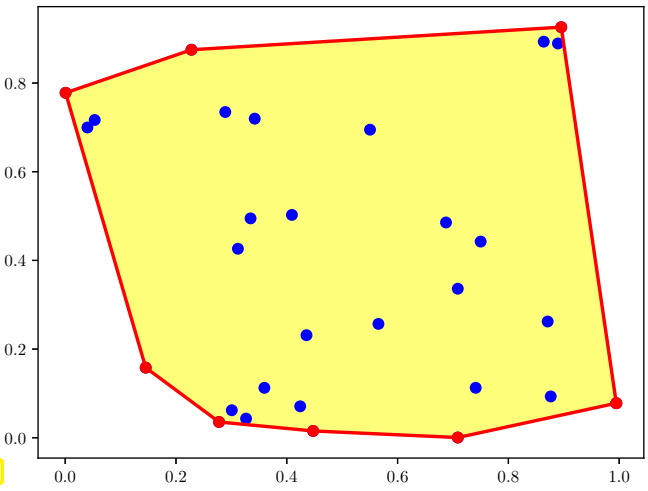


Fig. 195

◁ ■ ≙ convex hull of some points in the plane

— ≙ boundary of convex hull

Theorem 5.5.2.5. Bezier curves stay in convex hull

The Bezier curves $\mathbf{b}_i^k$, $k \in \{1,\ldots,n\}$, $i \in \{0,\ldots,n-k\}$, *defined by* (5.5.2.2) *are confined to the convex hulls of their underlying control points* $\mathbf{p}_i,\ldots,\mathbf{p}_{i+k}$*:*

$$\mathbf{b}_i^k(t) \in \operatorname{convex}\{\mathbf{p}_i,\ldots,\mathbf{p}_{i+k}\} \quad \forall t \in [0,1]\,.$$

Proof. The result follows by straightforward induction since every convex combination of two points in the convex hull again belongs to it. □

The next theorem states a deep result also connected with shape-fidelity. Its proof is beyond the scope of this course:

Theorem 5.5.2.6. Variation -diminishing property of Bezier curves

No straight line intersects a Bezier curve $\mathbf{b}_i^k$ *more times than it intersects the* ***control polygon*** *formed by its underlying control points* $\mathbf{p}_i,\ldots,\mathbf{p}_{i+k}$.

EXPERIMENT 5.5.2.7 (Poor shape-fidelity of high-degree Bezier curves)

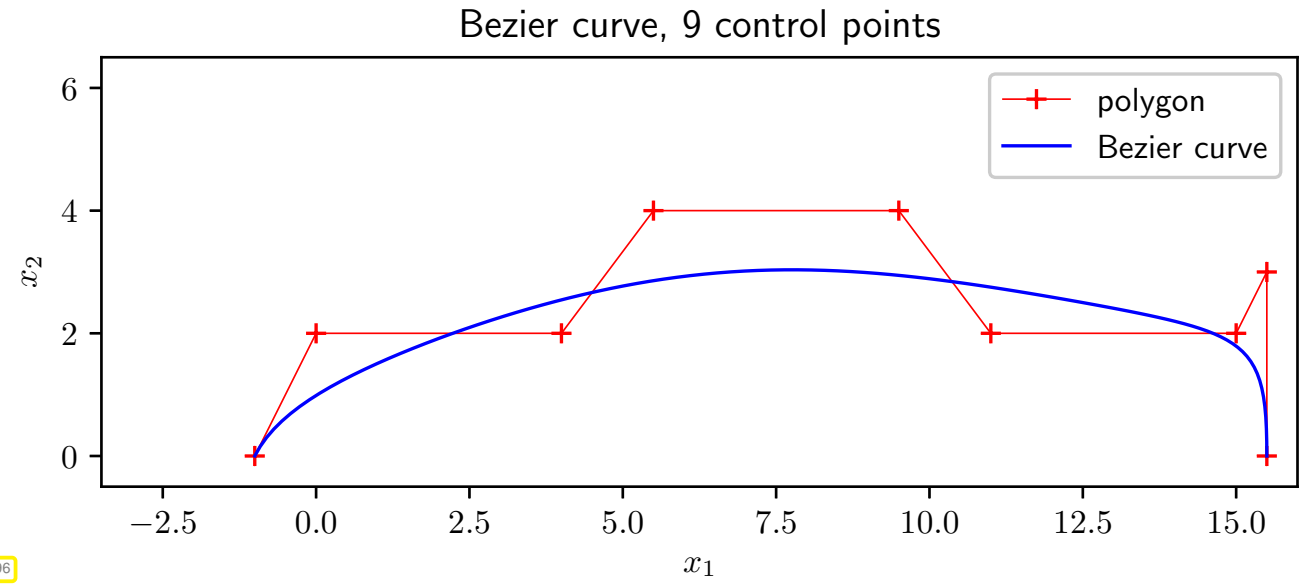


Fig. 196

Higher-degree Bezier curves are oblivious of local features!

Nevertheless, in order to motivate the construction of a better method, we continue examining Bezier curves and study them from a different perspective. From Thm. 5.5.2.5 we know that $\mathbf{b}^n(t) \in \operatorname{convex}\{\mathbf{p}_0,\ldots,\mathbf{p}_n\}$ for all $t \in [0,1]$. Thus, there must exists t-dependent weights $0 \le \xi_i(t) \le 1$ such that

$$\mathbf{b}^n(t) = \sum_{i=0}^{n} \xi_i(t)\mathbf{p}_i \quad \forall t \in [0,1]\,.$$

Theorem 5.5.2.8. Bernstein basis representation of Bezier curves

The Bezier curve $\mathbf{b}^n : [0,1] \to \mathbb{R}^2$ *as defined in* (5.5.2.2) *based on the control points* $\mathbf{p}_0, \ldots, \mathbf{p}_n$, $n \in \mathbb{N}$, *can be written as*

$$\mathbf{b}^n(t) = \sum_{i=0}^{n} B_i^n(t)\, \mathbf{p}_i \,, \tag{5.5.2.9}$$

with the ***Bernstein polynomials***

$$B_i^n(t) := \binom{n}{i} t^i (1-t)^{n-i} \,, \quad i \in \{0, \ldots, n\} \,. \tag{5.5.2.10}$$

Proof. Of course, we use induction in the polynomial degree based on (5.5.2.2) and the Pascal triangle for binomial coefficients. □

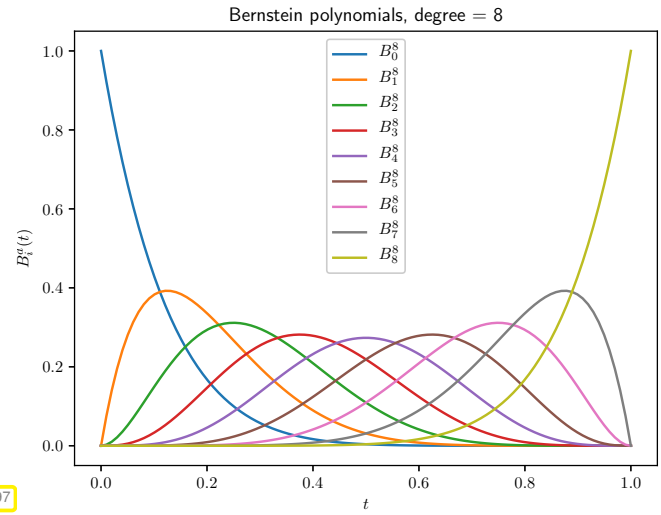


Fig. 197

◁ All Bernstein polynomials of degree $n = 8$.

Lemma 5.5.2.11. Basis property of Bernstein polynomials

The Bernstein polynomials B_i^n, $i = 0, \ldots, n$, *form a basis of the space* $\mathcal{P}_n$ *of polynomials of degree* $\leq n$.

We can verify by direct computation based on the multinomial theorem that the Bernstein polynomials satisfy

$$0 \leq B_i^n(t) \leq 1 \quad , \quad \sum_{i=0}^{n} B_i^n(t) = 1 \quad \forall t \in [0,1] \,. \tag{5.5.2.12}$$

Remark 5.5.2.13 (Modified Horner scheme for evaluation of Bezier polynomials) The representation (5.5.2.9) together with (6.2.1.3) paves the way for an efficient evaluation of Bezier curves for many parameter values. For example, this is important for fast graphical rendering of Bezier curves.

The algorithm is similar to that from § 5.2.3.33 and based on rewriting

$$\begin{aligned}
\mathbf{b}^n(t) &= \sum_{i=0}^{n} B_i^n(t)\, \mathbf{p}_i \\
&= \Bigg(\Big(\Big(\big((1-t)\,\mathbf{p}_0 + \frac{n}{1}\, t\, \mathbf{p}_1 \big)(1-t) + \frac{n(n-1)}{1\cdot 2}\, t^2\, \mathbf{p}_2 \Big)(1-t) + \\
&\qquad \frac{n(n-1)(n-2)}{1\cdot 2\cdot 3}\, t^3\, \mathbf{p}_3 \Big)(1-t) + \\
&\qquad\qquad \vdots \qquad\qquad \Bigg)(1-t) + t^n\, \mathbf{p}_n \,.
\end{aligned}$$

and evaluating the terms in brackets from the innermost to the outer.

C++ code 5.5.2.14: Evaluation of Bezier polynomial curve for many parameter arguments
➙ GITLAB

```
Eigen::MatrixXd evalBezier(const Eigen::MatrixXd &nodes,
                           const Eigen::RowVectorXd &t) {
  assert((nodes.rows() == 2) && "Nodes must have two coordinates");
  const Eigen::Index n = nodes.cols();  // Number of control points
  const Eigen::Index d = n - 1;         // Polynomial degree
  const Eigen::Index N = t.size(); // No.  of evaluation points
  // Vector containing 1-t ("one minus t")
  const auto oml{Eigen::RowVectorXd::Constant(N, 1.0) - t};
  // Modified Horner scheme for polynomial in Bezier form
  // Vector for returning result, initialized with p[0]*(1-t)
  Eigen::MatrixXd res = nodes.col(0) * oml;
  double binom_val = 1.0; // Accumulate binomial coefficients
  // Powers of argument values
  Eigen::RowVectorXd t_pow{Eigen::RowVectorXd::Constant(N, 1.0)};
  for (int i = 1; i < d; ++i) {
    t_pow.array() *= t.array();
    binom_val *= (static_cast<double>(d - i) + 1.0) / i;
    res += binom_val * nodes.col(i) * t_pow;
    res.row(0).array() *= oml.array();
    res.row(1).array() *= oml.array();
  }
  res += nodes.col(d) * (t.array() * t_pow.array()).matrix();
  return res;
}
```

The asymptotic computational effort for the evaluation for $N \in \mathbb{N}$ parameter values is $O(Nn)$ for $n, N \to \infty$. ⌟

5.5.3 Spline Curves

As we have seen in Exp. 5.5.2.7, Bezier curves based on global polynomials cannot represent shapes well, because they lack locality and flexibility. This agrees with the observations made in Section 5.3. There we learned that superior shape preservation could be achieved by using *piecewise polynomial* functions. The same policy will also be successful for curve design.

Again, without loss of generality we restrict ourselves to the parameter interval $[0,1]$, *cf.* Def. 5.5.1.3. We can conveniently reuse spaces of piecewise polynomials introduced earlier in Section 5.4:

Definition 5.4.1.1. Spline space

Given an interval $I := [a,b] \subset \mathbb{R}$ and a **knot sequence** $\mathcal{M} := \{a = t_0 < t_1 < \ldots < t_{m-1} < t_m = b\}$, $m \in \mathbb{N}$, the vector space $\mathcal{S}_{d,\mathcal{M}}$ of the **spline functions** of degree d (or order $d+1$) is defined by

$$\mathcal{S}_{d,\mathcal{M}} := \{s \in C^{d-1}(I) \colon s_j := s_{|[t_{j-1},t_j]} \in \mathcal{P}_d \; \forall j = 1,\ldots,m\} \,.$$

$d-1$-times continuously differentiable — *locally* polynomial of degree d

It goes without saying that a **spline curve** of degree $d \in \mathbb{N}$ is a curve $\mathbf{s} : [0,1] \to \mathbb{R}^2$ that possesses a parameterization that is component-wise a spline function of degree d with respect to a knot sequence $\mathcal{M} := \{t_0 = 0 < t_1 < \ldots < t_{m-1} < t_m = 1\}$ for the interval $[0,1]$: $\mathbf{s} \in (\mathcal{S}_{d,\mathcal{M}})^2$!

This does not yet bring us closer to the goal of building a shape-aware spline curve from $n+1$ given control points $\mathbf{p}_0,\ldots,\mathbf{p}_n$, $\mathcal{P}_\ell \in \mathbb{R}^2$, $n \in \mathbb{N}$. To begin with, recall

$$\dim(\mathcal{S}_{d,\mathcal{M}})^2 = 2(m+d)\ . \tag{5.5.3.1}$$

A first consideration realizes that $n+1$ control points $\in \mathbb{R}^2$ correspond to $2(n+1)$ degrees of freedom and matching that with the dimension of the space of spline curves we see that we should choose

$$m := n+1-d$$

knots in order to facilitate the construction of a unique curve from $\mathbf{p}_0,\ldots,\mathbf{p}_n$.

§5.5.3.2 (B-splines) Now we take he cue from the representation of Bezier curves by means of Bernstein polynomials as established Thm. 5.5.2.8 and aim for the construction of spline counterparts of the Bezier polynomials.

For an elegant presentation we admit **generalized knot sequences** that may contain *multiple knots*

$$\mathcal{U} := \{0 =: u_0 \le u_1 \le u_2 \le \cdots \le u_{m-1} \le u_m := 1\}\ , \quad m \in \mathbb{N}\ . \tag{5.5.3.3}$$

The number of times a u_k-value occurs in the sequence is called its multiplicity

The following definition can be motivated by a recursion satisfied by the Bernstein polynomials from (6.2.1.3):

$$B_i^k(t) := \binom{k}{i} t^i (1-t)^{k-i} \quad \Rightarrow \quad \begin{aligned} B_0^k(t) &= (1-t)B_0^{k-1}(t)\ , \\ B_i^k(t) &= tB_{i-1}^{k-1}(t) + (1-t)B_i^k(t)\ , \quad i=1,\ldots,k-1 \qquad t \in \mathbb{R}\ . \\ B_k^k(t) &= tB_{k-1}^{k-1}(t)\ , \end{aligned} \tag{5.5.3.4}$$

In particular, we notice that B_i^k is a two-term convex combination of B_{i-1}^k and B_i^k. We pursue something similar with splines:

Definition 5.5.3.5. B-splines

Given a generalized knot sequence $\mathcal{U} := \{0 =: u_0 \le u_1 \le u_2 \le \cdots \le u_{\ell-1} \le u_\ell := 1\}$ the B-splines N_i^k, $k \in \{0,\ldots,\ell-1\}$, $i \in \{0,\ldots,\ell-k-1\}$, are defined via the recursion

$$\begin{aligned} N_i^0(t) &:= \begin{cases} 1 & \text{for} \quad u_i \le t < u_{i+1}\ , \\ 0 & \text{elsewhere}\ , \end{cases} \quad i = 0,\ldots,\ell-1\ , \\ N_i^k(t) &:= \frac{t-u_i}{u_{i+k}-u_i} N_i^{k-1}(t) + \frac{u_{i+k+1}-t}{u_{i+k+1}-u_{i+1}} N_{i+1}^{k-1}(t)\ , \quad i = 0,\ldots,\ell-k-1\ , \end{aligned} \tag{5.5.3.6}$$

where we adopt the convention that "$\frac{0}{0} = 0$".

Note that also N_i^k is a two-term *convex combination* of N_i^{k-1} and N_{i+1}^k. From this we can deduce the following properties by simple induction:

Theorem 5.5.3.7. Elementary properties of B-splines

The B-splines based on the generalized knot sequence $\mathcal{U}$ *and defined by* (5.5.3.6) *satisfy*

- N_i^k *is a* $\mathcal{U}$*-piecewise polynomial of degree* $\leq k$,
- $0 \leq N_i^k(t) \leq 1$ *for all* $0 \leq t \leq 1$,
- N_i^k *vanishes outside the interval* $[u_i, \ldots, u_{i+k+1}]$: $\operatorname{supp}(N_i^k) = [u_i, u_{i+k+1}]$,
- $\sum_{j=i-k}^{i} N_j^k(t) = 1$ *for all* $t \in [u_i, u_{i+1}[$, $i = k, \ldots, \ell - k - 1$

The main "magic" result about B-splines concerns their smoothness properties that are by no means obvious from the recursive definition (5.5.3.6). To make a concise statement we build special generalized knot sequences from a (standard) knot sequence by copying endpoints d times: Given a degree $d \in \mathbb{N}_0$ and

$$\mathcal{M} := \{0 =: t_0 < t_1 < \ldots < t_{m-1} < t_m := 1\}\,, \quad m \in \mathbb{N}\,,$$

we construct a generalized knot sequence with $\ell + 1 := 2d + m + 1$ members

$$\mathcal{U}_{\mathcal{M}}^d := \{0 =: \underbrace{u_0 = \cdots = u_d}_{d+1 \text{ times}} < u_{d+1} := t_1 < \ldots < u_{d+m-1} := t_{m-1} < \underbrace{u_{d+m} = \cdots = u_{2d+m}}_{d+1 \text{ times}} := 1\}\,.$$

These plots show B-splines for special generalized knot sequences of the type $\mathcal{U}_{\mathcal{M}}^d$:

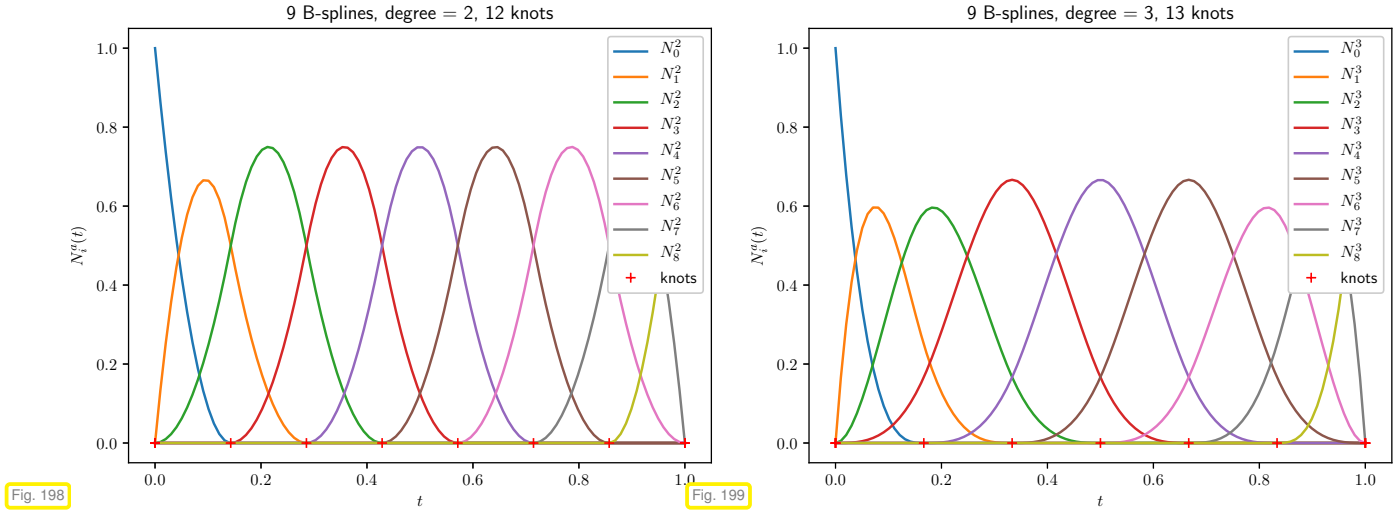


Fig. 198 Fig. 199

We notice that on that generalized knot sequence (5.5.3.6) provides $\ell - d = m + d$ B-spline functions N_i^d, $i = 0, \ldots, m + d - 1$. In light of Thm. 5.4.1.2 this is necessary for the following much more general assertion to hold, whose proof, unfortunately, is very technical and will be skipped.

Theorem 5.5.3.8. Basis property of B-splines

The B-splines of degree d *on the generalized knot set* $\mathcal{U}_{\mathcal{M}}^d$ *according to Def. 5.5.3.5 satisfy*

- $N_i^d \in \mathcal{S}_{d,\mathcal{M}}$ *for all* $i = 0, \ldots, m + d - 1$,
- $\left\{N_0^d, \ldots, N_{m+d-1}^d\right\}$ *is a basis of the spline space* $\mathcal{S}_{d,\mathcal{M}}$.

In particular, we conclude that N_i^d is $d-1$-times continuously differentiable: $N_i^d \in C^{d-1}([0,1])$! So, already for $d = 3$, the case of cubic splines we obtain curves that look perfectly smooth.

Inspired by Thm. 5.5.2.8 we propose the following recipe for constructing a degree-d spline curve from control points $\mathbf{p}_0, \ldots, \mathbf{p}_n$:

❶ Setting $m := n+1-d$ to match the numbers of degrees of freedom choose a knot sequence

$$\mathcal{M} := \{0 =: t_0 < t_1 < \ldots < t_{m-1} < t_m := 1\}\,, \quad m \in \mathbb{N}\,,$$

❷ Parameterize the curve by a linear combination of B-splines based on $\mathcal{U}_{\mathcal{M}}^d$:

$$\mathbf{s}(t) := \sum_{j=0}^{n} N_j^d(t)\,\mathbf{p}_j\,, \quad t \in [0,1]\,. \tag{5.5.3.9}$$

EXPERIMENT 5.5.3.10 (Curve design based on B-splines) We display the spline curves induced by the "car-shaped" set of control points using equidistant knots in $[0,1]$:

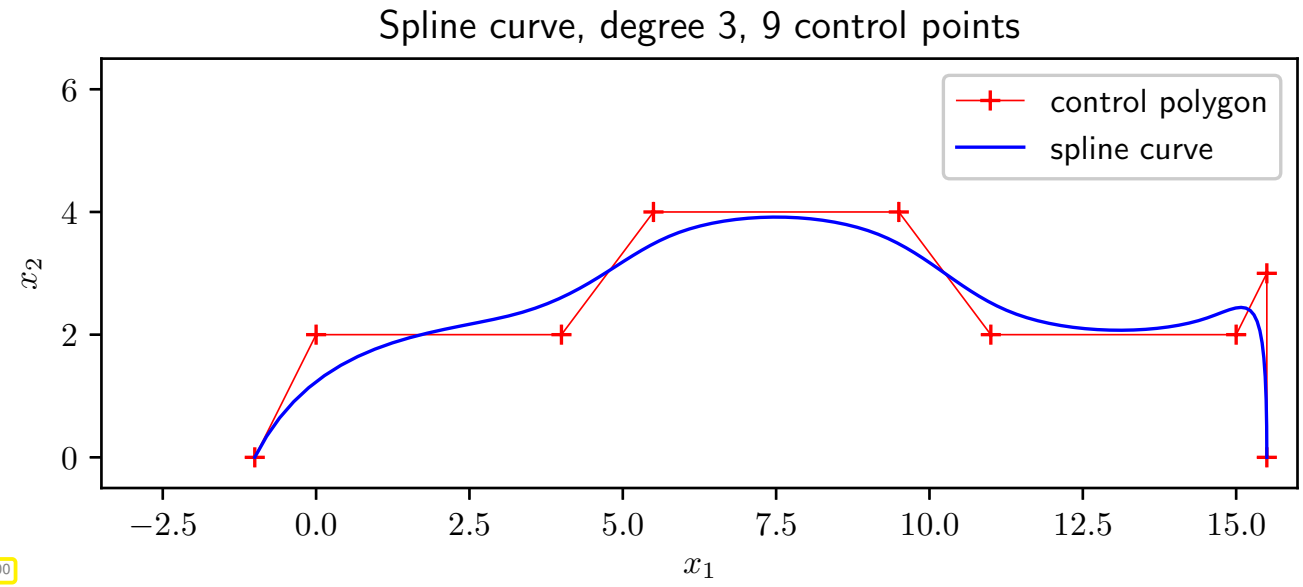


Fig. 200

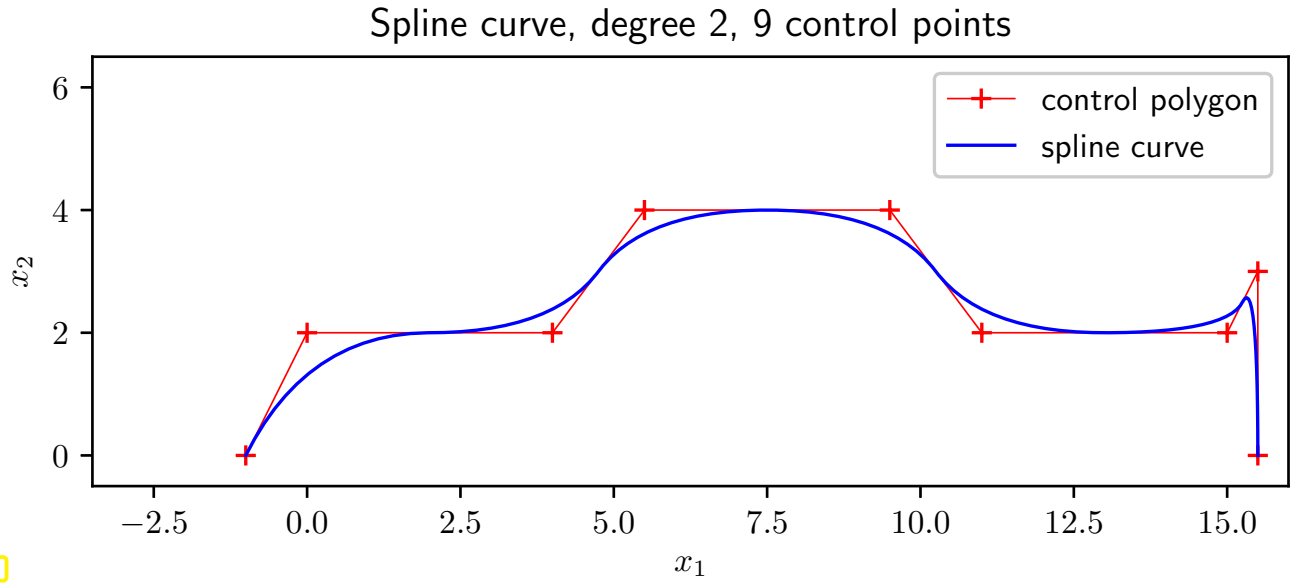


Fig. 201

Superior shape fidelity compared to Bezier curves!

BUT, considerable room for improvement still remains.

§5.5.3.11 (Node insertion)

5.6 Trigonometric Interpolation

We consider time series data (t_i, y_i), $i = 0, \ldots, n$, $t_i, y_i \in \mathbb{R}$, obtained by sampling a time-dependent scalar physical quantity $t \mapsto \varphi(t)$. We **know** that φ is a T-periodic function with period $T > 0$, that is $\varphi(t) = \varphi(t+T)$ for all $t \in \mathbb{R}$. In the spirit of shape preservation (→ Section 5.3) an interpolant f of the time series should also be T-periodic: $f(t+T) = f(t)$ for all $t \in \mathbb{R}$.

Assumption 5.6.0.1. Sampling in a period

We assume the period $T > 0$ to be known and $t_i \in [0, T[$ for all interpolation nodes t_i, $i = 0, \ldots, n$.

In the sequel, for the case of simplicity, we consider only $T = 1$.

Task: Given $T > 0$ and data points (t_i, y_i), $y_i \in \mathbb{K}$, $t_i \in [0, T[$, find a T-periodic function $f : \mathbb{R} \to \mathbb{K}$ (the interpolant), $f(t+T) = f(t)$ $\forall t \in \mathbb{R}$, that satisfies the interpolation conditions

$$f(t_i) = y_i \,, \quad i = 0, \ldots, n \,. \tag{5.6.0.2}$$

Supplementary literature. This topic is also presented in [DR08, Sect. 8.5].

5.6.1 Trigonometric Polynomials

The most fundamental periodic functions are derived from the trigonometric functions sin and cos and dilations of them (A dilation of a function $t \mapsto \psi(t)$ is a function of the form $t \mapsto \psi(ct)$ with some $c > 0$).

Definition 5.6.1.1. Space of trigonometric polynomials

The vector space of 1-periodic **trigonometric polynomials** of degree $2n$, $n \in \mathbb{N}$, is given by

$$\mathcal{P}_{2n}^T := \operatorname{Span}\{t \mapsto \cos(2\pi jt), t \mapsto \sin(2\pi jt)\}_{j=0}^{n} \subset C^\infty(\mathbb{R}) \,.$$

The terminology is natural after recalling expressions for trigonometric functions via complex exponentials ("Euler's formula")

$$e^{it} = \cos t + \imath \sin t \quad \Rightarrow \quad \begin{cases} \cos t = \frac{1}{2}(e^{\imath t} + e^{-\imath t}) \\ \sin t = \frac{1}{2\imath}(e^{\imath t} - e^{-\imath t}) \end{cases} . \tag{5.6.1.2}$$

Thus we can rewrite $q \in \mathcal{P}_{2n}^T$ given in the form

$$q(t) = \alpha_0 + \sum_{j=1} \alpha_j \cos(2\pi jt) + \beta_j \sin(2\pi jt) \,, \quad \alpha_j, \beta_j \in \mathbb{R} \,, \tag{5.6.1.3}$$

by means of complex exponentials making use of Euler's formula (5.6.1.2):

$$\begin{aligned} q(t) &= \alpha_0 + \sum_{j=1}^{n} \alpha_j \cos(2\pi jt) + \beta_j \sin(2\pi jt) \\ &= \alpha_0 + \frac{1}{2}\Big\{ \sum_{j=1}^{n} (\alpha_j - \imath\beta_j) e^{2\pi\imath jt} + (\alpha_j + \imath\beta_j) e^{-2\pi\imath jt} \Big\} \\ &= \alpha_0 + \tfrac{1}{2} \sum_{j=-n}^{-1} (\alpha_{-j} + \imath\beta_{-j}) e^{2\pi\imath jt} + \tfrac{1}{2} \sum_{j=1}^{n} (\alpha_j - \imath\beta_j) e^{2\pi\imath jt} \end{aligned}$$

$$= e^{-2\pi\imath nt}\sum_{j=0}^{2n}\gamma_j e^{2\pi\imath jt}\,, \quad \text{with} \quad \gamma_j = \begin{cases} \frac{1}{2}(\alpha_{n-j} + \imath\beta_{n-j}) & \text{for } j = 0,\ldots,n-1\,, \\ \alpha_0 & \text{for } j = n\,, \\ \frac{1}{2}(\alpha_{j-n} - \imath\beta_{j-n}) & \text{for } j = n+1,\ldots,2n\,. \end{cases} \tag{5.6.1.4}$$

Note that $\gamma_j \in \mathbb{C}$ even if $\alpha_j, \beta_j \in \mathbb{R}$ ➢ We mainly work in $\mathbb{C}$ in the context of trigonometric interpolation! Admit $y_k \in \mathbb{C}$.

From the above manipulations we conclude

$$q \in \mathcal{P}_{2n}^T \;\Rightarrow\; q(t) = e^{-2\pi\imath nt}\cdot p(e^{2\pi\imath t}) \quad \text{with} \quad p(z) = \sum_{j=0}^{2n}\gamma_j z^j \in \mathcal{P}_{2n}\,, \tag{5.6.1.5}$$

(a polynomial !)

and γ_j from (5.6.1.4).

(After scaling) a trigonometric polynomial of degree $2n$ is a regular polynomial $\in \mathcal{P}_{2n}$ (in $\mathbb{C}$) restricted to the unit circle $\mathbb{S}^1 \subset \mathbb{C}$.

✎ notation: $\mathbb{S}^1 := \{z \in \mathbb{C}\colon |z| = 1\}$ is the unit circle in the complex plane.

The relationship (5.6.1.5) justifies calling $\mathcal{P}_{2n}^T$ a space of trigonometric *polynomials*. It also defines a bijective linear mapping of vector spaces $\mathcal{P}_{2n} \to \mathcal{P}_{2n}^T$. This immediately reveals the dimension of $\mathcal{P}_{2n}^T$:

Corollary 5.6.1.6. Dimension of $\mathcal{P}_{2n}^T$

The vector space $\mathcal{P}_{2n}^T$ has dimension $\dim \mathcal{P}_{2n}^T = 2n+1$.

5.6.2 Reduction to Lagrange Interpolation

We observed that trigonometric polynomials are standard (complex) polynomials in disguise. Thus we can relate trigonometric interpolation to well-known standard Lagrangian interpolation discussed in Section 5.2.2. In fact, we slightly extend the method, because now we admit interpolation nodes $\in \mathbb{C}$. All results obtained earlier carry over to this setting.

The key tool is a smooth bijective mapping between $I := [0,1[$ and $\mathbb{S}^1$ defined as

$$\Phi_{\mathbb{S}^1} : I \to \mathbb{S}^1 \quad , \quad t \mapsto z := \exp(2\pi\imath t)\,, \tag{5.6.2.1}$$

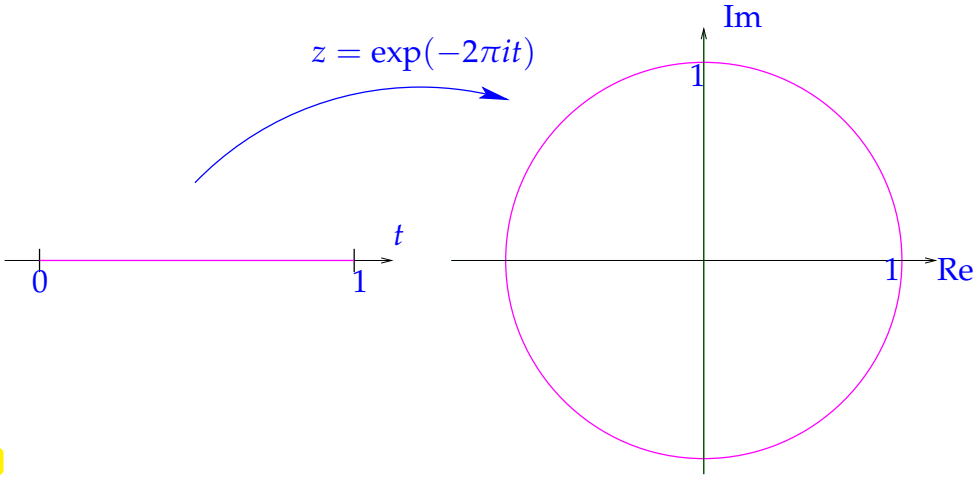


Fig. 202

and the relationship from (5.6.1.5):

$$q \in \mathcal{P}_{2n}^T \;\Leftrightarrow\; q(t) = e^{-2\pi\imath nt}p(e^{2\pi\imath t}) \quad \text{for some} \quad p \in \mathcal{P}_{2n}\,. \tag{5.6.2.2}$$

Trigonometric interpolation through data points (t_k, y_k), $k = 0, \ldots, 2n$	$\Longleftrightarrow$	Lagrange polynomial interpolation through data points $(e^{2\pi\imath t_k}, e^{2\pi\imath n t_k} y_k)$, $k = 0, \ldots, 2n$

Trigonometric interpolation $\longleftrightarrow$ polynomial interpolation on $\mathbb{S}^1$

All theoretical results and algorithms from polynomial interpolation carry over to trigonometric interpolation

- Existence and uniqueness of trigonometric interpolation polynomial, see Thm. 5.2.2.7,
- Concept of Lagrange polynomials, see (5.2.2.4),
- the algorithms and representations discussed in Section 5.2.3.

The next code demonstrates the use of standard routines for polynomial interpolation provided by **BarycPolyInterp** ($\rightarrow$ Code 5.2.3.7) for trigonometric interpolation.

C++-code 5.6.2.3: Evaluation of trigonometric interpolation polynomial in many points

```
// Evaluation of trigonometric interpolant at numerous points
// IN : t = vector of nodes t_0,...,t_n ∈ [0,1[
// y = vector of data y_0,...,y_n
// x = vector of evaluation points x_1,...,x_N
// OUT : vector of values of the interpolant at points in x
VectorXd trigpolyval(const VectorXd& t, const VectorXd& y, const VectorXd& x) {
  using idx_t = VectorXd::Index;
  using comp = std::complex<double>;
  const idx_t N = y.size(); // Number of data points
  if (N % 2 == 0) {
    throw std::runtime_error("Number of points must be odd!");
  }
  const auto n = static_cast<double>(N - 1)/2;
  const std::complex<double> M_I(0,1); // imaginary unit
  // interpolation nodes and evalutation points on unit circle
  const VectorXcd tc = ( 2*M_PI*M_I*t ).array().exp().matrix();
  const VectorXcd xc = ( 2*M_PI*M_I*x ).array().exp().matrix();
  // Rescaled values, according to q(t) = e^{-2πint} · p(e^{2πit}), see (5.6.1.4)
  const VectorXcd z = ((2*n*M_PI*M_I*t).array().exp() * y.array()).matrix();
  // Evaluation of interpolating polynomial on unit circle using the
  // barycentric interpolation formula in ℂ, see Code 5.2.3.4
  const BarycPolyInterp<comp> Interpol(tc);
  auto p = Interpol.eval<VectorXcd>(z,xc);
  // Undo the scaling, see (5.6.1.5)
  VectorXcd qc = ((-2*n*M_PI*M_I*x).array().exp() * p.array()).matrix();
  return (qc.real()); // imaginary part is zero, cut it off
}
```

The computational effort is $O(n^2 + Nn)$, $N \triangleq$ no. of evaluation points, as for barycentric polynomial interpolation studied in § 5.2.3.2.

The next code finds the coefficients $\alpha_j, \beta_j \in \mathbb{R}$ of a trigonometric interpolation polynomial in the real-valued representation (5.6.1.3) for real-valued data $y_j \in \mathbb{R}$ by simply solving the linear system of equations arising from the interpolation conditions (5.6.0.2).

C++-code 5.6.2.4: Computation of coefficients of trigonometric interpolation polynomial, general nodes ↬ GITLAB

```
//Computes expansion coefficients of trigonometric polyonomials (5.6.1.3)
// IN : t = vector of nodes t_0,...,t_n ∈ [0,1[
// y = vector of data y_0,...,y_n
// OUT: pair of vectors storing the basis expansion coefficients α_j, β_j,
    see Def. 5.6.1.1
std::pair<VectorXd,VectorXd>
trigpolycoeff(const VectorXd& t, const VectorXd& y) {
  const unsigned N = y.size();
  const unsigned n = (N-1)/2;
  if (N % 2 == 0) {
    throw std::logic_error("Number of points must be odd!");
  }

  // build system matrix M
  MatrixXd M(N, N);
  M.col(0) = VectorXd::Ones(N);
  for (unsigned c = 1; c <= n; ++c) {
    M.col(c) = ( 2*M_PI*c*t ).array().cos().matrix();
    M.col(n + c) = ( 2*M_PI*c*t ).array().sin().matrix();
  }
  // solve LSE and extract coefficients α_j and β_j
  VectorXd c = M.lu().solve(y);
  return {c.head(n + 1),c.tail(n)};
}
```

The asymptotic computational effort of this implementation is dominated by the cost for Gaussian elimination applied to a fully populated (dense) matrix, see Thm. 2.5.0.2: $O(n^3)$ for $n \to \infty$.

5.6.3 Equidistant Trigonometric Interpolation

Often timeseries data for a time-periodic quantity are measured with a constant rhythm over the entire (known) period of duration $T > 0$, that is, $t_j = j\Delta t$, $\Delta t = \frac{T}{n+1}$, $j = 0,\dots,n$. In this case, the formulas for computing the coefficients of the interpolating trigonometric polynomial ($\to$ Def. 5.6.1.1) become special versions of the discrete Fourier transform (DFT, see Def. 4.2.1.18) studied in Section 4.2. Efficient implementation can thus harness the speed of FFT introduced in Section 4.3.

§5.6.3.1 (Computation of equidistant trigonometric interpolants) Now we consider trigonometric interpolation in the 1-periodic setting with $2n+1$ *uniformly distributed* interpolation nodes $t_k = \dfrac{k}{2n+1}$, $k = 0,\dots,2n$, and associated data values y_k. Existence and uniqueness of an interpolating trigonometric polynomial $q \in \mathcal{P}_{2n}^T$, $q(t_k) = y_k$, was established in Section 5.6.2. We rely on the relationship

$$q \in \mathcal{P}_{2n}^T \quad \Rightarrow \quad q(t) = e^{-2\pi\imath nt} \cdot p(e^{2\pi\imath t}) \quad \text{with} \quad p(z) = \sum_{j=0}^{2n} \gamma_j z^j \in \mathcal{P}_{2n}\,, \tag{5.6.1.5}$$

to arrive at the following $(2n+1)\times(2n+1)$ linear system of equations for computing the unknown coeffi-

$$\sum_{j=0}^{2n}\gamma_j\exp\big(2\pi\imath\frac{jk}{2n+1}\big) = (\mathbf{b})_k := \exp\big(2\pi\imath\frac{nk}{2n+1}\big)y_k\,,\quad k=0,\ldots,2n\,. \tag{5.6.3.2}$$

$$\Updownarrow$$

cients γ_j:

$$\overline{\mathbf{F}}_{2n+1}\mathbf{c}=\mathbf{b}\,,\quad \mathbf{c}=[\gamma_0,\ldots,\gamma_{2n}]^\top \overset{\text{Lemma 4.2.1.14}}{\Rightarrow} \boxed{\mathbf{c}=\frac{1}{2n+1}\mathbf{F}_{2n+1}\mathbf{b}}\,. \tag{5.6.3.3}$$

$(2n+1)\times(2n+1)$ (conjugate) Fourier matrix, see (4.2.1.13)!

The linear system of equations (5.6.3.2) is amenable to fast solution by means of FFT with $O(n\log n)$ asymptotic complexity for $n\to\infty$, see Section 4.3.

The following C++ code demonstrates the efficient computation of the coefficients $\alpha_j,\beta_j\in\mathbb{R}$ of the trigonometric polynomial

$$q(t)=\alpha_0+\sum_{j=1}^{n}\alpha_j\cos(2\pi jt)+\beta_j\sin(2\pi jt)\,,\quad \alpha_j,\beta_j\in\mathbb{R}\,, \tag{5.6.1.3}$$

interpolating the data points $\left(\frac{k}{2n+1},y_k\right)_{k=0}^{2n}$. Only the data values y_j need to be passed.

C++-code 5.6.3.4: Efficient computation of coefficient of trigonometric interpolation polynomial (*equidistant nodes*) ↪ GITLAB

```
2   // Efficient FFT-based computation of coefficients in expansion
3   // (5.6.1.3) for a trigonometric interpolation polynomial in
4   // equidistant points (j/(2n+1), y_j), j = 0,...,2n.
5   // IN : y has to be a row vector of odd length, return values are
6   // column vectors OUT: vectors [alpha_j]_j,
7   // [alpha_j]_j of expansion coefficients with respect
8   // to trigonometric basis from Def. 5.6.1.1
9   std::pair<VectorXd, VectorXd>
10  trigipequid(const VectorXd &y) {
11  using index_t = VectorXcd::Index;
12  const index_t N = y.size();
13  const index_t n = (N - 1) / 2;
14  if (N % 2 != 1) {
15    throw std::logic_error("Number of points must be odd!");
16  }
17  // prepare data for fft
18  const std::complex<double> M_I(0, 1); // imaginary unit
19  // right hand side vector b from (5.6.3.3)
20  VectorXcd b(N);
21  for (index_t k = 0; k < N; ++k) {
22    b(k) = y(k) * std::exp(2 * M_PI * M_I *
23                           (static_cast<double>(n) / static_cast<double>(N) *
24                            static_cast<double>(k)));
25  }
26  Eigen::FFT fft;              // DFT helper class
27  VectorXcd c = fft.fwd(b); // means that "c = fft(b)"
28
29  // By (5.6.1.4) we can recover
30  // alpha_j = 1/2(gamma_{n-j} + gamma_{n+j}) and
31  // beta_j = 1/(2i)(gamma_{n-j} - gamma_{n+j}),
32  // j = 1,...,n, alpha_0 = gamma_n.
33  VectorXcd alpha(n + 1);
```

```
34    VectorXcd beta(n);
35    alpha(0) = c(n) / static_cast<double>(N);
36    for (index_t l = 1; l <= n; ++l) {
37      alpha(l) = (c(n - l) + c(n + l)) / static_cast<double>(N);
38      beta(l - 1) = -M_I * (c(n - l) - c(n + l)) / static_cast<double>(N);
39    }
40    return {alpha.real(), beta.real()};
41  }
```

EXAMPLE 5.6.3.5 (Runtime comparison for computation of coefficient of trigonometric interpolation polynomials)

Runtime measurements for MATLAB equivalents of codes Code 5.6.2.4 and Code 5.6.3.4

`tic-toc`-timings ▷

MATLAB 7.10.0.499 (R2010a)

CPU: Intel Core i7, 2.66 GHz, 2 cores, L2 256 KB, L3 4 MB

OS: Mac OS X, Darwin Kernel Version 10.5.0

(Similar runtimes would also be measured for C++ or PYTHON codes.)

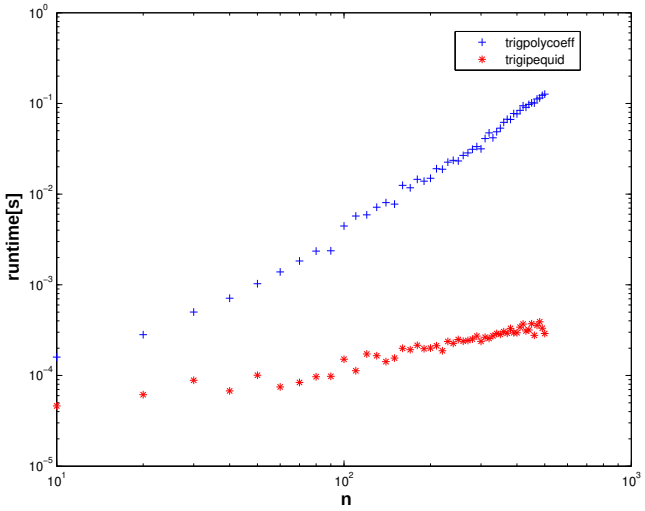


Fig. 203

We make the same observation as in Ex. 4.3.0.13: massive gain in efficiency through relying on FFT. ⌟

§5.6.3.6 (Efficient evaluation of trigonometric interpolation polynomials) Imagine we want to plot a 1-periodic real-valued trigonometric polynomial $q \in \mathcal{P}_{2n}^T$ given through its coefficients α_j, β_j. To that end we have to evaluate it in many points, which are naturally chosen as uniformly spaced in $[0,1]$.

Task: Find an efficient way to evaluate the trigonometric polynomial

$$q(t) = \alpha_0 + \sum_{j=1} \alpha_j \cos(2\pi jt) + \beta_j \sin(2\pi jt) \;, \quad \alpha_j, \beta_j \in \mathbb{R} \;, \tag{5.6.1.3}$$

at *equidistant* points $\frac{k}{N}$, $N > 2n$. $k = 0, \ldots, N-1$.

As in (5.6.1.4) we can rewrite

$$q(t) = e^{-2\pi \imath nt} \sum_{j=0}^{2n} \gamma_j e^{2\pi \imath jt} \;, \quad \text{with} \quad \gamma_j = \begin{cases} \frac{1}{2}(\alpha_{n-j} + \imath \beta_{n-j}) & \text{for } j = 0, \ldots, n-1 \;, \\ \alpha_0 & \text{for } j = n \;, \\ \frac{1}{2}(\alpha_{j-n} - \imath \beta_{j-n}) & \text{for } j = n+1, \ldots, 2n \;. \end{cases}$$

(5.6.1.4) ⟹ $q(k/N) = e^{-2\pi \imath nk/N} \sum_{j=0}^{2n} \gamma_j \exp(2\pi \imath \frac{kj}{N}) \;, \quad k = 0, \ldots, N-1 \;.$

⟹ $q(k/N) = e^{-2\pi \imath kn/N} v_j \quad \text{with} \quad \mathbf{v} = \overline{\mathbf{F}}_N \widetilde{\mathbf{c}} \;,$ (5.6.3.7)

Fourier matrix, see (4.2.1.13).

where $\widetilde{\mathbf{c}} \in \mathbb{C}^N$ is obtained by zero padding of $\mathbf{c} := (\gamma_0, \ldots, \gamma_{2n})^T$:

$$(\widetilde{\mathbf{c}})_k = \begin{cases} \gamma_j & \text{, for } k = 0, \ldots, 2n \text{ ,} \\ 0 & \text{, for } k = 2n+1, \ldots, N-1 \text{ .} \end{cases}$$

The FFT-based implementation is realized in the following code, which accepts the coefficients α_j, β_j as components of the vectors `a` and `b`, and returns $q(k/N)$, $k = 0, \ldots, N-1$, in the vector `y`.

C++-code 5.6.3.8: Fast evaluation of trigonometric polynomial at *equidistant* points ➙ GITLAB

```
2     void
3     trigipequidcomp(const VectorXcd &a, const VectorXcd &b, const unsigned N,
4                     VectorXcd &y) {
5   const unsigned n = a.size() - 1;
6   if (N < (2 * n - 1)) {
7     std::cerr << "N is too small! Must be larger than 2*n";
8     return;
9   }
10  const std::complex<double> iu(0, 1); // imaginary unit
11  // build vector γ
12  VectorXcd gamma(2 * n + 1);
13  gamma(n) = a(0);
14  for (unsigned k = 0; k < n; ++k) {
15    gamma(k) = 0.5 * (a(n - k) + iu * b(n - k - 1));
16    gamma(n + k + 1) = 0.5 * (a(k + 1) - iu * b(k));
17  }
18  // zero padding to obtain c̃
19  VectorXcd ch(N);
20  ch << gamma, VectorXcd::Zero(N - (2 * n + 1));
21
22  // realize multiplication with conjugate fourier matrix
23  Eigen::FFT fft;
24  const VectorXcd chCon = ch.conjugate();
25  VectorXcd v = fft.fwd(chCon).conjugate();
26
27  // final scaling, implemented without efficiency considerations
28  y = VectorXcd(N);
29  for (unsigned k = 0; k < N; ++k) {
30    y(k) = v(k) * std::exp(-2. * k * n * M_PI / N * iu);
31  }
32 }
```

The next code merges the steps of computing the coefficient of the trigonometric interpolation polynomial in equidistant points and its evaluation in another set of M equidistant points.

C++ code 5.6.3.9: *Equidistant* points: fast on the fly evaluation of trigonometric interpolation polynomial ➙ GITLAB

```
2 // Evaluation of trigonometric interpolation polynomial through
3 // (j/(2n+1), y_j), j = 0,...,2n in equidistant points
4 // k/M, k = 0,M-1 IN : y = vector of values to
5 // be interpolated
6 // q (COMPLEX!) will be used to save the return values
7 void trigpolyvalequid(const VectorXd y, const int M, VectorXd &q) {
8   const int N = y.size();
9   if (N % 2 == 0) {
```

```
10      std::cerr << "Number of points must be odd!\n";
11      return;
12    }
13    const int n = (N - 1) / 2;
14    // computing coefficient γ_j, see (5.6.3.3)
15    VectorXcd a, b;
16    trigipequid(y, a, b);
17
18    std::complex<double> i(0, 1);
19    VectorXcd gamma(2 * n + 1);
20    gamma(n) = a(0);
21    for (int k = 0; k < n; ++k) {
22      gamma(k) = 0.5 * (a(n - k) + i * b(n - k - 1));
23      gamma(n + k + 1) = 0.5 * (a(k + 1) - i * b(k));
24    }
25
26    // zero padding
27    VectorXcd ch(M);
28    ch << gamma, VectorXcd::Zero(M - (2 * n + 1));
29
30    // build conjugate fourier matrix
31    Eigen::FFT fft;
32    const VectorXcd chCon = ch.conjugate();
33    const VectorXcd v = fft.fwd(chCon).conjugate();
34
35    // multiplicate with conjugate fourier matrix
36    VectorXcd q_complex = VectorXcd(M);
37    for (int k = 0; k < M; ++k) {
38      q_complex(k) = v(k) * std::exp(-2. * k * n * M_PI / M * i);
39    }
40    // complex part is zero up to machine precision, cut off!
41    q = q_complex.real();
42  }
```

Review question(s) 5.6.3.10 (Trigonomentic interpolation)

(Q5.6.3.10.A) You have at your disposal the function

```
std::pair<Eigen::VectorXd,Eigen::VectorXd>
trigpolycoeff(comnst Eigen::VectorXd &t, const Eigen::VectorXd &y);
```

that computes the coefficients α_j, β_j, of the 1-*periodic* interpolating trigonometric polynomial $q \in \mathcal{P}_{2n}^T$

$$q(t) = \alpha_0 + \sum_{j=1}^{n} \{\alpha_j \cos(2\pi jt) + \beta_j \sin(2\pi jt)\} .$$

for the data points $(t_k, y_k) \in \mathbb{R}^2$, $k = 0, \ldots, 2n$, passed in the vectors `t` and `y` of equal length.

However, now we have to process data points that are obtained by sampling a T-periodic function, $T > 0$ known. Describe, how you can use `trigpolycoeff()` to obtain a T-periodic interpolant $p : \mathbb{R} \to \mathbb{R}$.

(Q5.6.3.10.B) Let $q \in \mathcal{P}_{2n}^T$ be the 1-periodic trigonometric interpolant of (t_k, y_k), $k = 0, \ldots, 2n$, $0 < t_k < 1$:

$$q(t) = \alpha_0 + \sum_{j=1}^{n} \{\alpha_j \cos(2\pi jt) + \beta_j \sin(2\pi jt)\} .$$

What are the basis expansion coefficients $\widetilde{\alpha}_j, \widetilde{\beta}_j$ for the trigonometric interpolant $\widetilde{q} \in \mathcal{P}_{2n}^T$ through $(\widetilde{t}_k, y_k)$, $k = 0, \ldots, 2n$, with $\widetilde{t}_k = t_k + \Delta t$, $\Delta t > 0$ given?

(Q5.6.3.10.C) Outline an implementation of a C++ function

```
double evalTrigPol(const Eigen::VectorXd &alpha, const
   Eigen::VectorXd &beta, double t);
```

that computes $q(t)$ for

$$q(t) = \alpha_0 + \sum_{j=1}^{n} \{\alpha_j \cos(2\pi jt) + \beta_j \sin(2\pi jt)\} ,$$

using only a *single call* of a mathematical function of the C++ standard library. The coefficients are passed through the vectors `alpha` and `beta`, the evaluation point t in `t`.

△

5.7 Least Squares Data Fitting

Video tutorial for Section 5.7 "Least Squares Data Fitting": (13 minutes) Download link, tablet notes

→ review questions 5.7.0.31

As remarked in Ex. 5.1.0.8 the basic assumption underlying the reconstructing of the functional dependence of two quantities by means of interpolation is that of accurate data. In case of data uncertainty or measurement errors the exact satisfaction of interpolation conditions ceases to make sense and we are better off reconstructing a fitting function that is merely "close to the data" in a sense to be made precise next.

The most general task of (multidimensional, vector-valued) **least squares data fitting** can be described as follows:

Least square data fitting

Given: data points $(\mathbf{t}_i, \mathbf{y}_i)$, $i \in \{1, \ldots, m\}$, $m \in \mathbb{N}$, $\mathbf{t}_i \in D \subset \mathbb{R}^k$, $\mathbf{y}_i \in \mathbb{R}^d$, $d \in \mathbb{N}$

Objective: Find a (continuous) function $\mathbf{f} : D \mapsto \mathbb{R}^d$ in some set $S \subset C^0(D)$ of admissible functions satisfying

$$\mathbf{f} \in \operatorname*{argmin}_{\mathbf{g} \in S} \sum_{i=1}^{m} \|\mathbf{g}(\mathbf{t}_i) - \mathbf{y}_i\|_2^2 . \tag{5.7.0.2}$$

Such a function $\mathbf{f}$ is called a (best) least squares fit for the data in S.

Focus in this section: $k = 1$, $d = 1$ (one-dimensional scalar setting)

↔ fitting of *scalar* data depending on *one* parameter ($t \in \mathbb{R}$),

↔ Data points (t_i, y_i), $t_i \in I \subset \mathbb{R}$, $y_i \in \mathbb{R}$ ➢ $S \subset C^0(I)$

↔ (best) least squares fit $f : I \subset \mathbb{R} \to \mathbb{R}$ is a function of one real variable.

$$k = 1\,,\quad d = 1\text{:}\qquad (5.7.0.2) \quad\Leftrightarrow\quad f \in \operatorname*{argmin}_{g\in S} \sum_{i=1}^{m} |g(t_i) - y_i|^2\,. \tag{5.7.0.3}$$

§5.7.0.4 (Linear spaces of admissible functions) Consider a special variant of the general least squares data fitting problem: The set S of admissible continuous functions is now chosen as a finite-dimensional vector space $V_n \subset C^0(D)$, $\dim V_n = n \in \mathbb{N}$, *cf.* the discussion in § 5.1.0.21 for interpolation.

Choose basis of V_n: $\quad V_n = \operatorname{Span}\{\mathbf{b}_1, \ldots, \mathbf{b}_n\}$, $\quad \mathbf{b}_j : D \to \mathbb{R}^d$ continuous

The best least squares fit $\mathbf{f} \in V_n$ can be represented by a *finite linear combination* of the basis functions $\mathbf{b}_j$:

$$\mathbf{f}(t) = \sum_{j=1}^{n} x_j \mathbf{b}_j(t) \quad,\quad x_j \in \mathbb{R}\,. \tag{5.7.0.5}$$

Often, in the case $d > 1$, V_n is chosen as a product space

$$V_n = \underbrace{W \times \cdots \times W}_{d \text{ factors}}\,, \tag{5.7.0.6}$$

of a space $W \subset C^0(D)$ of $\mathbb{R}$-valued functions $D \mapsto \mathbb{R}$. In this case

$$\dim V_n = d \cdot \dim W\,.$$

If $\ell := \dim W$ and $\{q_1, \ldots, q_\ell\}$ is a basis of W, then

$$\{\mathbf{e}_i q_j\}_{\substack{i=1,\ldots,d\\ j=1,\ldots,\ell}} = \{\mathbf{e}_1 q_1, \mathbf{e}_2 q_1, \ldots, \mathbf{e}_d q_1, \mathbf{e}_1 q_2, \ldots, \mathbf{e}_d q_2, \mathbf{e}_1 q_3, \ldots, \mathbf{e}_1 q_\ell, \ldots \mathbf{e}_d q_\ell\} \tag{5.7.0.7}$$

is a basis of V_n ($\mathbf{e}_i \mathrel{\hat{=}} i$-th unit vector). ⌟

§5.7.0.8 (Linear data fitting → [DR08, Sect. 4.1]) We adopt the setting of § 5.7.0.4 of an n-dimensional space V_n of admissible functions with basis $\{\mathbf{b}_1, \ldots, \mathbf{b}_n\}$. Then the least squares data fitting problem can be recast as follows.

General Linear least squares fitting problem:

Given:
- data points $(\mathbf{t}_i, \mathbf{y}_i) \in \mathbb{R}^k \times \mathbb{R}^d$, $\quad i = 1, \ldots, m$
- basis functions $\mathbf{b}_j : D \subset \mathbb{R}^k \mapsto \mathbb{R}^d$, $j = 1, \ldots, n$, $n < m$

Sought: coefficients $x_j \in \mathbb{R}$, $j = 1, \ldots, n$, such that

$$\mathbf{x} := [x_1, \ldots, x_n]^\top := \operatorname*{argmin}_{z_j \in \mathbb{R}} \sum_{i=1}^{m} \left\| \sum_{j=1}^{n} z_j \mathbf{b}_j(\mathbf{t}_i) - \mathbf{y}_i \right\|_2^2\,. \tag{5.7.0.9}$$

Special cases:

- For $k = 1$, $d = 1$, data points $(t_i, y_i) \in \mathbb{R} \times \mathbb{R}$ (scalar, one-dimensional setting), and $V_n = \operatorname{Span}\{b_1, \ldots, b_n\}$, we seek coefficients $x_j \in \mathbb{R}$, $j = 1, \ldots, n$, as the components of a vector $\mathbf{x} = [x_1, \ldots, x_n]^\top \in \mathbb{R}^n$ satisfying

$$\mathbf{x} = \operatorname*{argmin}_{\mathbf{z}\in\mathbb{R}^n} \sum_{i=1}^{m} \left| \sum_{j=1}^{n} (\mathbf{z})_j b_j(t_i) - y_i \right|^2\,. \tag{5.7.0.10}$$

- If V_n is a product space according to (5.7.0.6) with basis (5.7.0.7), then (5.7.0.9) amounts to finding vectors $\mathbf{x}_j \in \mathbb{R}^d$, $j = 1, \ldots, \ell$, with

$$(\mathbf{x}_1, \ldots, \mathbf{x}_\ell) = \operatorname*{argmin}_{\mathbf{z}_j \in \mathbb{R}^d} \sum_{i=1}^{m} \left\| \sum_{j=1}^{\ell} \mathbf{z}_j q_j(\mathbf{t}_i) - \mathbf{y}_i \right\|_2^2 . \tag{5.7.0.11}$$

⌟

EXAMPLE 5.7.0.12 (Linear parameter estimation = linear data fitting → Ex. 3.0.1.4, Ex. 3.1.1.5) The linear parameter estimation/linear regression problem presented in Ex. 3.0.1.4 can be recast as a linear data fitting problem with

- $k = n$, $d = 1$, data points $(\mathbf{x}_i, y_i) \in \mathbb{R}^k \times \mathbb{R}$,
- an $k+1$-dimensional space $V_n = \{\mathbf{x} \mapsto \mathbf{a}^\top \mathbf{x} + \beta,\ \mathbf{a} \in \mathbb{R}^k, \beta \in \mathbb{R}\}$ of affine linear admissible functions,
- the choice of basis $\{\mathbf{x} \mapsto (\mathbf{x})_1, \ldots, \mathbf{x} \mapsto (\mathbf{x})_k, \mathbf{x} \mapsto 1\}$.

⌟

§5.7.0.13 (Linear data fitting as a lineare least squares problem) Linear (least squares) data fitting leads to an overdetermined linear system of equations for which we seek a least squares solution (→ Def. 3.1.1.1) as in Section 3.1.1. To see this rewrite

$$\sum_{i=1}^{m} \left\| \sum_{j=1}^{n} z_j \mathbf{b}_j(\mathbf{t}_i) - \mathbf{y}_i \right\|_2^2 = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} \sum_{r=1}^{d} (\mathbf{b}_j(\mathbf{t}_i))_r z_j - (\mathbf{y}_i)_r \right)^2 .$$

Theorem 5.7.0.14. Least squares solution of data fitting problem

The solution $\mathbf{x} = [x_1, \ldots, x_n]^\top \in \mathbb{R}^n$ of the linear least squares fitting problem (5.7.0.9) *is the least squares solution of the overdetermined linear system of equations*

$$\begin{bmatrix} \mathbf{A}_1 \\ \vdots \\ \mathbf{A}_d \end{bmatrix} \mathbf{x} = \begin{bmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_d \end{bmatrix} , \tag{5.7.0.15}$$

with $\mathbf{A}_r := \begin{bmatrix} (\mathbf{b}_1(\mathbf{t}_1))_r & \ldots & (\mathbf{b}_n(\mathbf{t}_1))_r \\ \vdots & & \vdots \\ (\mathbf{b}_1(\mathbf{t}_m))_r & \ldots & (\mathbf{b}_n(\mathbf{t}_m))_r \end{bmatrix} \in \mathbb{R}^{m,n}$, $\mathbf{b}_r := \begin{bmatrix} (\mathbf{y}_1)_r \\ \vdots \\ (\mathbf{y}_m)_r \end{bmatrix} \in \mathbb{R}^m$, $r = 1, \ldots, d$.

In the one-dimensional, scalar case ($d = 1$) of (5.7.0.10) the related overdetermined linear system of equations is

$$\begin{bmatrix} b_1(t_1) & \ldots & b_n(t_1) \\ \vdots & & \vdots \\ b_1(t_m) & \ldots & b_n(t_m) \end{bmatrix} \mathbf{x} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} . \tag{5.7.0.16}$$

Obviously, for $m = n$ we recover the 1D data interpolation problem from § 5.1.0.21. For $d > 1$ the overdetermined linear system of equations (5.7.0.15) can be thought of as d systems of the type (5.7.0.16) stacked on top of each other, one for every component of the data vectors.

Having reduced the linear least squares data fitting problem to finding the least squares solution of an overdetermined linear system of equations, we can now apply theoretical results about least squares solutions, for instance, Cor. 3.1.2.13. The key issue is, whether the coefficient matrix of (5.7.0.16) has full rank n. Of course, this will depend on the location of the t_i.

Lemma 5.7.0.17. Unique solvability of linear least squares fitting problem

The scalar one-dimensional linear least squares fitting problem (5.7.0.10) *with* $\dim V_n = n$, V_n *the vector space of admissible functions, has a unique solution,* if and only if *there are* $t_{i_1}, \ldots, t_{i_n}$ *such that*

$$\begin{bmatrix} b_1(t_{i_1}) & \ldots & b_n(t_{i_1}) \\ \vdots & & \vdots \\ b_1(t_{i_n}) & \ldots & b_n(t_{i_n}) \end{bmatrix} \in \mathbb{R}^{n,n} \quad \textit{is invertible,} \tag{5.7.0.18}$$

which is independent of the choice of basis of V_n.

Equivalent to (5.7.0.18) is the requirement, that there is an n-subset of $\{t_1, \ldots, t_n\}$ such that the corresponding interpolation problem for V_n has a unique solution for any data values y_i. ⌟

EXAMPLE 5.7.0.19 (Polynomial fitting)

Special variant of scalar ($d = 1$), one-dimensional $k = 1$ linear data fitting (→ § 5.7.0.8): we choose the space of admissible functions as polynomials of degree $n - 1$,

$$V_n = \mathcal{P}_{n-1} \quad , \quad \textit{e.g.} \text{ with basis } \quad b_j(t) = t^{j-1} \quad \text{(monomial basis, Section 5.2.1)} .$$

The corresponding overdetermined linear system of equations (5.7.0.16) now reads:

$$\begin{bmatrix} 1 & t_1 & t_1^2 & \ldots & t_1^{n-1} \\ 1 & t_2 & t_2^2 & \ldots & t_2^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & t_m & t_m^2 & \ldots & t_m^{n-1} \end{bmatrix} \mathbf{x} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} , \tag{5.7.0.20}$$

which, for $M \geq n$, has full rank, because it contains invertible Vandermonde matrices (5.2.2.11), Rem. 5.2.2.10.

The next code demonstrates the computation of the fitting polynomial with respect to the monomial basis of $\mathcal{P}_{n-1}$:

C++ code 5.7.0.21: Polynomial fitting ↪ GITLAB

```
2  // Solver for polynomial (degree passed in 'degree') linear least
       squares data
3  // fitting problem,data points passed in t and y
4  VectorXd polyfit(const VectorXd& t, const VectorXd& y, unsigned int degree) {
5    // Initialize the coefficient matrix of (5.7.0.20)
6    Eigen::MatrixXd A = Eigen::MatrixXd::Ones(t.size(), degree + 1);
7    for (unsigned int j = 1; j < degree + 1; ++j)
8      A.col(j) = A.col(j - 1).cwiseProduct(t);
9    // Use EIGEN's built-in least squares solver, see Code 3.3.4.2
10   Eigen::VectorXd coeffs = A.householderQr().solve(y);
```

```
11    // leading coefficients have low indices.
12    return coeffs.reverse();
13  }
```

The function `polyfit` returns a vector $[x_1, x_2, \ldots, x_n]^\top$ describing the fitting polynomial according to the convention

$$p(t) = x_1 t^{n-1} + x_2 t^{n-2} + \cdots + x_{n-1} t + x_n \,. \tag{5.7.0.22}$$

⌟

EXPERIMENT 5.7.0.23 (Polynomial interpolation vs. polynomial fitting)

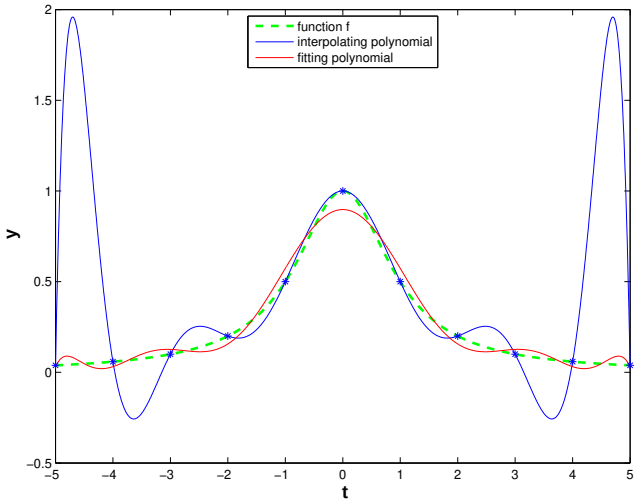


Fig. 204

Data from function $f(t) = \dfrac{1}{1+t^2}$,

- polynomial degree $d = 10$,
- interpolation through data points $(t_j, f(t_j))$, $j = 0, \ldots, d$, $t_j = -5 + j$, see Ex. 5.2.4.3,
- fitting to data points $(t_j, f(t_j))$, $j = 0, \ldots, 3d$, $t_j = -5 + j/3$.

Experiment carried out with the code `interpfit.cpp` ↬ GITLAB.

Fitting helps curb oscillations that haunt polynomial interpolation: in terms of "shape preservation" the fitted polynomial is clearly superior to the interpolating polynomial. ⌟

Remark 5.7.0.24 ("Overfitting") In the previous example we saw that the degree-10 polynomial fitted to the data in 31 points rather well reflects the "shape of the data", see Fig. 204. If we had fit a polynomial of degree 30 to the data instead, the result would have been a function with humongous oscillations, because we would have recovered the interpolating polynomial at equidistant nodes, *cf.* Ex. 5.2.4.3.

Thus, in Exp. 5.7.0.23 we see a manifestation of a widely observed phenomenon, which plays a big role in machine learning:

Overfitting

Fitting data with functions from a large space often produces poorer results (in terms of "shape preservation" and "generalization error") than the use of a smaller subspace for fitting.

⌟

§5.7.0.26 (Data fitting with regularization) Let us recall two observations made in the context of one-dimensional interpolation of scalar data:

- High-degree global polynomial interpolants usually suffer from massive oscillations.
- The "nice" cubic-spline interpolant s on $[t_0, t_n]$ could also be characterized as the C^2-interpolant with minimal bending energy $\int_{t_0}^{t_n} |s''(t)|\, \mathrm{d}t$, see Thm. 5.4.3.3.

This motivates augmenting the fitting problem with another term depending on inner product norms of derivatives of the fitting functions.

Concretely, given data points $(t_i, y_i) \in \mathbb{R}^2$, $i = 0, \ldots, m$, $t_i \in [a,b]$, and seeking the fitting function f in the n-dimensional space

$$V_n := \operatorname{Span}\{b_1, \ldots, b_n\} \subset C^2([a,b]) \,,$$

we determine it as

$$f \in \operatorname*{argmin}_{g \in V} \left\{ \sum_{i=0}^{n} |g(t_i) - y_i|^2 + \alpha \int_a^b |g''(t)|^2 \,\mathrm{d}t \right\} \,, \tag{5.7.0.27}$$

with a regularization parameter $\alpha > 0$. Since $\int_a^b |g''(t)| \,\mathrm{d}t$ will be large, if g oscillates wildly, for large α the extra regularization term will serve to suppress oscillations.

Plugging a basis expansion $g = \sum_{j=1}^{n} x_j b_j$ into (5.7.0.27), we arrive at the quadratic minimization problem

$$\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 + \alpha \mathbf{x}^\top \mathbf{M} \mathbf{x} \to \min \,, \tag{5.7.0.28}$$

$$\mathbf{A} := \left[b_j(t_i) \right]_{\substack{i=0,\ldots,m \\ j=1,\ldots,n}} \in \mathbb{R}^{m+1,n} \quad , \quad \mathbf{M} := \left[\int_a^b b_j''(t)\, b_k''(t) \,\mathrm{d}t \right]_{j,k=1}^{n} \in \mathbb{R}^{n,n} \,. \tag{5.7.0.29}$$

For this minimization problem we can find a linear system of equations for minimizers by setting $\mathbf{grad}\, \phi(\mathbf{x}) \stackrel{!}{=} \mathbf{0}$ for $\phi(\mathbf{x}) := \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 + \alpha \mathbf{x}^\top \mathbf{M} \mathbf{x}$. As in § 3.6.1.4 we find generalized normal equations

$$\begin{aligned} &\mathbf{grad}\, \phi(\mathbf{x}) = 2\mathbf{A}^\top \mathbf{A}\mathbf{x} - 2\mathbf{A}^\top \mathbf{y} + 2\alpha \mathbf{M}\mathbf{x} = \mathbf{0} \,, \\ &\Updownarrow \\ &\boxed{(\mathbf{A}^\top \mathbf{A} + \alpha \mathbf{M})\mathbf{x} = \mathbf{A}^\top \mathbf{y}} \,. \end{aligned} \tag{5.7.0.30}$$

This is an $n \times n$ linear system of equations with symmetric positive semi-definite coefficient matrix. ⌟

Review question(s) 5.7.0.31 (Least squares data fitting)

(Q5.7.0.31.A) [] Given data points $(t_i, y_i) \in \mathbb{R}^2$, $i = 0, \ldots, m$, $t_i \in I \subset \mathbb{R}$, let $\{b_1, \ldots, b_n\} \subset C^0(I)$ be a basis of the space V of admissible fitting functions. A least squares fit can be computed as

$$f = \sum_{j=1}^{n} (\mathbf{x})_j b_j \quad \text{where} \quad \mathbf{x} \in \operatorname*{argmin}_{\mathbf{z} \in \mathbb{R}^n} \sum_{i=0}^{m} \left| \sum_{j=1}^{n} (\mathbf{z})_j b_j(t_i) - y_i \right|^2 .$$

Show that f does not depend on the choice of basis for V.

(Q5.7.0.31.B) Given $n+1$ data points $(t_i, y_i) \in \mathbb{R}^2$, $t_0 < t_1 < \cdots < t_n$, we can define a regularized piecewise linear fit as

$$f \in \operatorname*{argmin}_{g \in \mathcal{S}_{1,\mathcal{M}}} \left\{ \sum_{i=0}^{n} |g(t_i) - y_i|^2 + \sum_{i=1}^{n} \int_{t_{i-1}}^{t_i} |g'(t)|^2 \,\mathrm{d}t \right\} \,,$$

where the knot set $\mathcal{M}$ is given by the node set: $\mathcal{M} := \{t_0 < t_1 < \cdots < t_n\}$. Using the "cardinal basis" of $\mathcal{S}_{1,\mathcal{M}}$ comprised of "tent functions", derive the linear system that has to be solved to find f.

△

Learning outcomes

After you have studied this chapter you should

- understand the use of basis functions for representing functions on a computer,
- know the concept of a interpolation operator and what its linearity means,
- know the connection between linear interpolation operators and linear systems of equations,
- be familiar with efficient algorithms for polynomial interpolation in different settings,
- know the meaning and significance of "sensitivity" in the context of interpolation,
- Be familiar with the notions of "shape preservation" for an interpolation scheme and its different aspects (monotonicity, curvature),
- know the details of cubic Hermite interpolation and how to ensure that it is monotonicity preserving.
- know what splines are and how cubic spline interpolation with different endpoint constraints works.

Bibliography

[Aki70] H. Akima. “A New Method of Interpolation and Smooth Curve Fitting Based on Local Procedures”. In: *J. ACM* 17.4 (1970), pp. 589–602.

[AG11] Uri M. Ascher and Chen Greif. *A first course in numerical methods*. Vol. 7. Computational Science & Engineering. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2011, pp. xxii+552. DOI: 10.1137/1.9780898719987 (cit. on p. 394).

[Bei+16] Lourenço Beirão da Veiga, Annalisa Buffa, Giancarlo Sangalli, and Rafael Vázquez. “An introduction to the numerical analysis of isogeometric methods”. In: *Numerical simulation in physics and engineering*. Vol. 9. SEMA SIMAI Springer Ser. Springer, [Cham], 2016, pp. 3–69.

[BT04] Jean-Paul Berrut and Lloyd N. Trefethen. “Barycentric Lagrange interpolation”. In: *SIAM Rev.* 46.3 (2004), 501–517 (electronic). DOI: 10.1137/S0036144502417715.

[CR92] D. Coppersmith and T.J. Rivlin. “The growth of polynomials bounded at equally spaced points”. In: *SIAM J. Math. Anal.* 23.4 (1992), pp. 970–983 (cit. on p. 417).

[DR08] W. Dahmen and A. Reusken. *Numerik für Ingenieure und Naturwissenschaftler*. Heidelberg: Springer, 2008 (cit. on pp. 393–396, 401, 408, 412, 414, 415, 430, 456, 465).

[FH07] Michael S. Floater and Kai Hormann. “Barycentric rational interpolation with no poles and high rates of approximation”. In: *Numer. Math.* 107.2 (2007), pp. 315–331. DOI: 10.1007/s00211-007-0093-y.

[FC80] F.N. Fritsch and R.E. Carlson. “Monotone Piecewise Cubic Interpolation”. In: *SIAM J. Numer. Anal.* 17.2 (1980), pp. 238–246 (cit. on p. 428).

[Gou05] E. Gourgoulhon. *An introduction to polynomial interpolation*. Slides for School on spectral methods. 2005.

[Han02] M. Hanke-Bourgeois. *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens*. Mathematische Leitfäden. Stuttgart: B.G. Teubner, 2002 (cit. on pp. 431, 435).

[Kva14] Boris I. Kvasov. “Monotone and convex interpolation by weighted quadratic splines”. In: *Adv. Comput. Math.* 40.1 (2014), pp. 91–116. DOI: 10.1007/s10444-013-9300-9.

[MR81] D. McAllister and J. Roulier. “An algorithm for computing a shape-preserving osculatory quadratic spline”. In: *ACM Trans. Math. Software* 7.3 (1981), pp. 331–347 (cit. on pp. 441, 444).

[Moh15] A. Mohsen. “CORRECTING THE FUNCTION DERIVATIVE ESTIMATION USING LAGRANGIAN INTERPOLATION”. In: *ZAMP* (2015).

[NS02] K. Nipp and D. Stoffer. *Lineare Algebra*. 5th ed. Zürich: vdf Hochschulverlag, 2002 (cit. on p. 392).

[QQY01] A.L. Dontchevand H.-D. Qi, L.-Q. Qi, and H.-X. Yin. “Convergence of Newton’s method for convex best interpolation”. In: *Numer. Math.* 87.3 (2001), pp. 435–456 (cit. on p. 441).

[QSS00] A. Quarteroni, R. Sacco, and F. Saleri. *Numerical mathematics*. Vol. 37. Texts in Applied Mathematics. New York: Springer, 2000 (cit. on pp. 394–397, 408, 414, 415, 430, 431, 436).

[Sch83] Larry L. Schumaker. “On shape preserving quadratic spline interpolation”. In: *SIAM J. Numer. Anal.* 20.4 (1983), pp. 854–864. DOI: 10.1137/0720057 (cit. on p. 440).

[SS25] Zewen Shen and Kirill Serkh. “On polynomial interpolation in the monomial basis”. In: *SIAM J. Numer. Anal.* 63.2 (2025), pp. 469–494. DOI: 10.1137/23M1623215.

[Str09] M. Struwe. *Analysis für Informatiker*. Lecture notes, ETH Zürich. 2009 (cit. on pp. 419, 420).

[Tre13] Lloyd N. Trefethen. *Approximation theory and approximation practice*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2013, viii+305 pp.+back matter (cit. on p. 399).

[Tre14] N. Trefethen. *SIX MYTHS OF POLYNOMIAL INTERPOLATION AND QUADRATURE*. Slides. 2014.

[WX12] Haiyong Wang and Shuhuang Xiang. "On the convergence rates of Legendre approximation". In: *Math. Comp.* 81.278 (2012), pp. 861–877. DOI: 10.1090/S0025-5718-2011-02549-4.

Chapter 6

Approximation of Functions in 1D

6.1 Introduction

Video tutorial for Section 6.1 "Approximation of Functions in 1D: Introduction": (7 minutes) Download link, tablet notes

In Chapter 5 we aimed to fill the gaps between given data points by constructing a function connecting them. In this chapter the function is available already, at least in principle, but its evaluation is too costly, which forces us to replace it with a "cheaper" or "simpler" function, sometimes called a surrogate function.

§6.1.0.1 (General function approximation problem)

Approximation of functions: Generic view

Given: a **function** $\mathbf{f} : D \subset \mathbb{R}^n \mapsto \mathbb{R}^d$, $n, d \in \mathbb{N}$, often in procedural form, e.g. for $n = d = 1$, as **double** `f(double)` ($\rightarrow$ Rem. 5.1.0.9).

Goal: Find a *"simple"* function $\widetilde{\mathbf{f}} : D \mapsto \mathbb{R}^d$ such that the approximation error $\mathbf{f} - \widetilde{\mathbf{f}}$ is *"small"* .

Let us clarify the meaning of some terms:

Made more precise: *"simple"*

The function $\widetilde{\mathbf{f}}$ can be encoded by a small amount of information and is easy to evaluate. For instance, this is the case for polynomial or piecewise polynomial $\widetilde{\mathbf{f}}$.

Made more precise: *"small"* approximation error

$\left\|\mathbf{f} - \widetilde{\mathbf{f}}\right\|$ is small for some norm $\|\cdot\|$ on the space $C^0(\overline{D})$ of (piecewise) continous functions.

The most commonly used norms are

- the supremum norm $\|\mathbf{g}\|_\infty := \|\mathbf{g}\|_{L^\infty(D)} := \max_{x \in D} |\mathbf{g}(x)|$, see (5.2.4.5).

  If the approximation error is small with respect to the supremum norm, $\widetilde{\mathbf{f}}$ is also called a good uniform approximant of $\mathbf{f}$.
- the L^2-norm $\|\mathbf{g}\|_2^2 := \|\mathbf{g}\|_{L^2(D)}^2 = \int_D |\mathbf{g}(x)|^2 dx$, see (5.2.4.6).

Below we consider only the case $n = d = 1$: approximation of scalar valued functions defined on an

interval. The techniques can be applied componentwise in order to cope with the case of vector valued function ($d > 1$). ⌟

EXAMPLE 6.1.0.3 (Model reduction in circuit simulation)

The non-linear circuit ($\triangleright|$ stands for a diode, a circuit element, whose resistance strongly depends on the voltage) sketched beside has two ports.

For the sake of circuit simulation it should be replaced by a non-linear lumped circuit element characterized by a single voltage-current constitutive relationship $I = I(U)$. For any value of the voltage U the current I can be computed by solving a (large) non-linear system of equations, see Ex. 8.1.0.1.

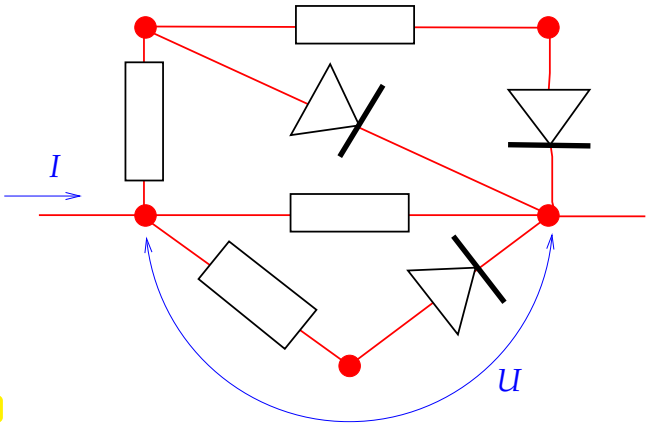


Fig. 205

A faster alternative is the advance approximation of the function $U \mapsto I(U)$ based on a few computed values $I(U_i)$, $i = 0,\ldots,n$, followed by the fast evaluation of the approximant $U \mapsto \widetilde{I}(U)$ during actual circuit simulations. This is an example of model reduction by approximation of functions: a complex subsystem in a mathematical model is replaced by a surrogate function.

In this example we also encounter a typical situation: we have nothing at our disposal but, possibly expensive, point evaluations of the function $U \mapsto I(U)$ ($U \mapsto I(U)$ in "procedural form", see Rem. 5.1.0.9). The number of evaluations of $I(U)$ will largely determine the cost of building $\widetilde{I}$.

This application displays a fundamental difference compared to the reconstruction of constitutive relationships from a priori measurements → Ex. 5.1.0.8: Now we are free to choose the number and location of the data points, because we can simply evaluate the function $U \mapsto I(U)$ for any U and as often as needed.

C++ code 6.1.0.4: Class describing a 2-port circuit element for circuit simulation

```
class CircuitElement {
  private:
  // internal data describing U ↦ Ĩ(U).
  public:
  // Constructor taking some parameters and building Ĩ
  CircuitElement(const Parameters &P);
  // Point evaluation operators for Ĩ and d/dU Ĩ
  double I(double U) const;
  double dIdU(double U) const;
};
```

⌟

§6.1.0.5 (Approximation schemes) We define an abstract concept for the sake of clarity: When in this chapter we talk about an "approximation scheme" (in 1D) we refer to a mapping $A : X \mapsto V$, where X and V are spaces of functions $I \mapsto \mathbb{K}$, $I \subset \mathbb{R}$ an interval.

Examples are

- $X = C^k(I)$, the spaces of functions $I \mapsto \mathbb{K}$ that are k times continuously differentiable, $k \in \mathbb{N}$.
- $V = \mathcal{P}_m(I)$, the space of polynomials of degree $\leq k$, see Section 5.2.1

- $V = \mathcal{S}_{d,\mathcal{M}}$, the space of splines of degree d on the knot set $\mathcal{M} \subset I$, see Def. 5.4.1.1.
- $V = \mathcal{P}_{2n}^T$, the space of trigonometric polynomials of degree $2n$, see Def. 5.6.1.1.

⌟

§6.1.0.6 (Approximation by interpolation) In Chapter 5 we discussed ways to construct functions whos graph runs through given data points, see 5.1. We can hope that the interpolant will approximate the function, if the data points are also located on the graph of that function. Thus every interpolation scheme, see § 5.1.0.7, spawns a corresponding approximation scheme.

Interpolation scheme + sampling → approximation scheme

$$f : I \subset \mathbb{R} \to \mathbb{K} \xrightarrow{\text{sampling}} (t_i, y_i := f(t_i))_{i=0}^m \xrightarrow{\text{interpolation}} \widetilde{f} := \mathrm{I}_{\mathcal{T}}\mathbf{y} \quad (\widetilde{f}(t_i) = y_i)\,.$$

free choice of nodes $t_i \in I$

In this chapter we will mainly study approximation by interpolation relying on the interpolation schemes (→ § 5.1.0.7) introduced in Section 5.2, Section 5.3.3, and Section 5.4.

There is *additional freedom* compared to data interpolation: we can choose the interpolation nodes in a smart way in order to obtain an accurate interpolant $\widetilde{f}$.

⌟

Remark 6.1.0.8 (Interpolation and approximation: enabling technologies) Approximation and interpolation (→ Chapter 5) are key components of many numerical methods, like for integration, differentiation and computation of the solutions of differential equations, as well as for computer graphics and generation of smooth curves and surfaces.

This chapter is a "foundations" part of the course

⌟

Contents

6.1 Introduction . . . 473
6.2 Approximation by Global Polynomials . . . 476
6.2.1 Polynomial Approximation: Theory . . . 477
6.2.2 Error Estimates for Polynomial Interpolation . . . 483
6.2.3 Chebychev Interpolation . . . 499
6.3 Mean Square Best Approximation . . . 515
6.3.1 Abstract Theory . . . 515
6.3.2 Polynomial Mean Square Best Approximation . . . 520
6.4 Uniform Best Approximation . . . 526
6.5 Approximation by Trigonometric Polynomials . . . 530
6.5.1 Approximation by Trigonometric Interpolation . . . 531
6.5.2 Trigonometric Interpolation Error Estimates . . . 532
6.5.3 Trigonometric Interpolation of Analytic Periodic Functions . . . 539
6.6 Approximation by Piecewise Polynomials . . . 545
6.6.1 Piecewise Polynomial Lagrange Interpolation . . . 546
6.6.2 Cubic Hermite Interpolation: Error Estimates . . . 549
6.6.3 Cubic Spline Interpolation: Error Estimates [Han02, Ch. 47] . . . 551

Review question(s) 6.1.0.9 (Approximation of Functions in 1D: Introduction)

(Q6.1.0.9.A) Describe an interpolation-based approximation scheme $\mathcal{A} : C^0(I) \to C^0(I)$ for the approximation of functions in $C^0(I)$, $I \subset \mathbb{R}$ a closed interval, such that

- $\mathcal{A}(f)$ is increasing whenever f is increasing,
- for convex f also $\mathcal{A}(f)$ is convex.

(Q6.1.0.9.B) Let $\mathcal{A} : C^0(I) \to C^0(I)$, $I \subset \mathbb{R}$ a closed interval, be an approximation scheme. How can you ensure that

$$f \not\equiv 0 \quad \Rightarrow \quad \mathcal{A}(f) \not\equiv 0 \,\textbf{?}$$

△

6.2 Approximation by Global Polynomials

Video tutorial for Section 6.2 "Polynomial Approximation: Theory": (13 minutes) Download link, tablet notes

The space $\mathcal{P}_k$ of polynomials of degree $\leq k$ has been introduced in Section 5.2.1. For reasons listed in § 5.2.1.3 polynomials are the most important theoretical and practical tool for the approximation of functions. The next example presents an important case of approximation by polynomials.

EXAMPLE 6.2.0.1 (Taylor approximation → [Str09, Sect. 5.5]) The local approximation of *sufficiently smooth* functions by polynomials is a key idea in calculus, which manifests itself in the importance of approximation by Taylor polynomials: For $f \in C^k(I)$, $k \in \mathbb{N}$, $I \subset \mathbb{R}$ an interval, we approximate

$$f(t) \approx \underbrace{f(t_0) + f'(t_0)(t-t_0) + \frac{f^{(2)}(t_0)}{2}(t-t_0)^2 + \cdots + \frac{f^{(k)}(t_0)}{k!}(t-t_0)^k}_{=:T_k(t) \in \mathcal{P}_k} \quad , \quad \text{for some } t_0 \in I \, .$$

✎ Notation: $f^{(k)} \triangleq k$-th derivative of function $f : I \subset \mathbb{R} \to \mathbb{K}$

The Taylor polynomial T_k of degree k approximates f in a neighbourhood $J \subset I$ of t_0 (J can be small!); it supplies a *local approximation* of f. The approximation error can be expressed through remainder formulas [Str09, Bem. 5.5.1]

$$f(t) - T_k(t) = \int_{t_0}^{t} f^{(k+1)}(\tau) \frac{(t-\tau)^k}{k!} \, \mathrm{d}\tau \tag{6.2.0.2a}$$

$$= f^{(k+1)}(\xi) \frac{(t-t_0)^{k+1}}{(k+1)!} \quad , \quad \xi = \xi(t,t_0) \in]\min(t,t_0), \max(t,t_0)[\, , \tag{6.2.0.2b}$$

which shows that for $f \in C^{k+1}(I)$ the Taylor polynomial T_k is pointwise close to $f \in C^{k+1}(I)$, if the interval I is small and $f^{(k+1)}$ is bounded pointwise.

Approximation by Taylor polynomials is easy and direct but inefficient: a polynomial of lower degree often gives the same accuracy. Moreover, when f is available only in procedural form as **double** `f(double)`, (approximations of) higher order derivatives are difficult to obtain. ⌟

§6.2.0.3 (Nested approximation spaces of polynomials) Obviously, for every interval $I \subset \mathbb{R}$, the spaces of polynomials are nested in the following sense:

$$\mathcal{P}_0 \subset \mathcal{P}_1 \subset \cdots \subset \mathcal{P}_m \subset \mathcal{P}_{m+1} \subset \cdots \subset C^\infty(I)\,, \tag{6.2.0.4}$$

with finite, but increasing dimensions $\dim \mathcal{P}_m = m+1$ according to Thm. 5.2.1.2.

With this family of nested spaces of polynomials at our disposal, it is natural to study *associated families of approximation schemes*, one for each degree, mapping into $\mathcal{P}_m$, $m \in \mathbb{N}_0$. ⌟

6.2.1 Polynomial Approximation: Theory

§6.2.1.1 (Scope of polynomial approximation) Sloppily speaking, according to (6.2.0.2b) the Taylor polynomials from Ex. 6.2.0.1 provide *uniform* (→ § 6.1.0.1) approximation of a *smooth* function f in (small) intervals, provided that its derivatives do not blow up "too fast" (We do not want to make this precise here).

The question is, whether polynomials still offer uniform approximation on arbitrary bounded closed intervals and for functions that are merely continuous, but not any smoother. The answer is YES and this profound result is known as the Weierstrass Approximation Theorem. Here we give an extended version with a concrete formula due to Bernstein, see [Dav75, Section 6.2].

Theorem 6.2.1.2. Uniform approximation by polynomials

For $f \in C^0([0,1])$, define the n-th Bernstein approximant as

$$p_n(t) = \sum_{j=0}^n f(j/n)\binom{n}{j} t^j (1-t)^{n-j}\,, \quad p_n \in \mathcal{P}_n\,. \tag{6.2.1.3}$$

It satisfies $\|f - p_n\|_\infty \to 0$ for $n \to \infty$. If $f \in C^m([0,1])$, then even $\left\|f^{(k)} - p_n^{(k)}\right\|_\infty \to 0$ for $n \to \infty$ and all $0 \le k \le m$.

✎ Notation: $g^{(k)} \triangleq k$-th derivative of a function $g : I \subset \mathbb{R} \to \mathbb{K}$.

In (6.2.1.3) the function f is approximated by a linear combination of Bernstein polynomials of degree n

$$B_j^n(t) := \binom{n}{j} t^j (1-t)^{n-j} \quad , \quad B_j^n \in \mathcal{P}_n\,. \tag{6.2.1.4}$$

Plots of Bernstein polynomials of degree $n = 7$ ▷

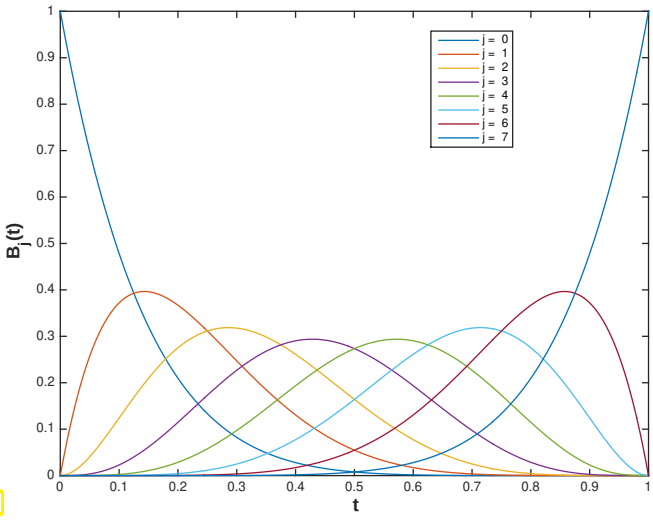


Fig. 206

Bernstein polynomials provide a positive **partition of unity** over $[0,1]$:

$$\sum_{j=0}^n B_j^n(t) = 1 \quad \forall t \in \mathbb{R}\,, \tag{6.2.1.5}$$

$$0 \le B_j^n(t) \le 1 \quad \forall 0 \le t \le 1\,. \tag{6.2.1.6}$$

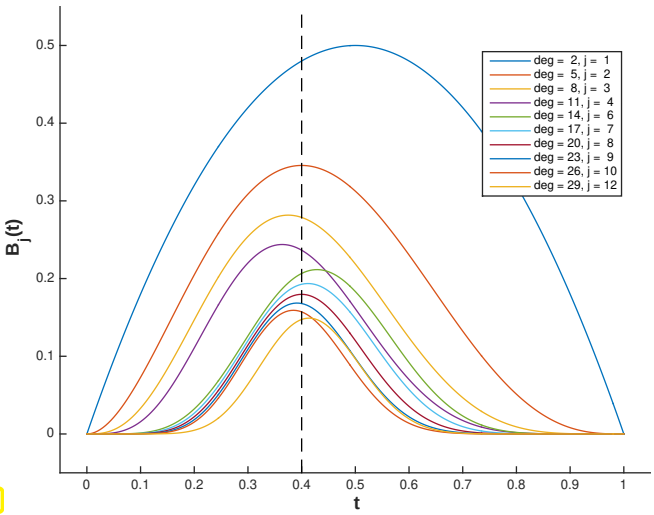


Fig. 207

◁ Since $\frac{d}{dt}B_j^n(t) = B_j^n(t)(\frac{j}{t} - \frac{n-j}{1-t})$, B_j^n has its unique local maximum in $[0,1]$ at the site $t_{\max} := \frac{j}{n}$. As $n \to \infty$ the Bernstein polynomials become more and more concentrated around the maximum.

Proof. (of Thm. 6.2.1.2, first part) Fix $t \in [0,1]$. Using the notations from (6.2.1.3) and the identity (6.2.1.5) we find

$$f(t) - p_n(t) = \sum_{k=0}^{n} (f(t) - f(j/n)) B_j^n(t) \,. \tag{6.2.1.7}$$

As we see from Fig. 207, for large n the bulk of sum will be contributed by Bernstein polynomial with index $j/n \approx t$, because for every $\delta > 0$

$$\sum_{|j/n-t|>\delta} B_j^n(t) \le \frac{1}{\delta^2} \sum_{|j/n-t|>\delta} (j/n-t)^2 B_j^n(t) \le \frac{1}{\delta^2} \sum_{j=0}^{n} (j/n-t)^2 B_j^n(t) \overset{(*)}{=} \frac{nt(1-t)}{\delta^2 n^2} \le \frac{1}{4n\delta^2} \,.$$

$\sum_{|j/n-t|>\delta}$ means summation over $j \in \mathbb{N}_0$ with summation indices confined to the set $\{j : |j/n - t| > \delta\}$.
The identity $(*)$ can be established by direct but tedious computations [Dav75, Formulas (6.2.8)].

Combining this estimate with (6.2.1.6) and (6.2.1.7) we arrive at

$$|f(t) - p_n(t)| \le \sum_{|j/n-t|>\delta} \frac{1}{4n\delta^2} |f(t) - f(j/n)| + \sum_{|j/n-t|\le\delta} |f(t) - f(j/n)| \,.$$

Since, f is uniformly continuous on $[0,1]$, given $\epsilon > 0$ we can choose $\delta > 0$ independently of t such that $|f(s) - f(t)| < \epsilon$, if $|s - t| < \delta$. The, if we choose $n > (\epsilon\delta^2)^{-1}$, we can bound

$$|f(t) - p_n(t)| \le (\|f\|_\infty + 1)\epsilon \quad \forall t \in [0,1] \,.$$

This means that p_n is arbitrarily close to f for sufficiently large n. □ ⌟

EXPERIMENT 6.2.1.8 (Bernstein approximants) We compute and plot p_n, $n = 1, \ldots, 25$, for two functions

$$f_1(t) := \begin{cases} 0 & \text{, if } |2t-1| > \frac{1}{2} \,, \\ \frac{1}{2}(1 + \cos(2\pi(2t-1))) & \text{else,} \end{cases} \quad , \quad f_2(t) := \frac{1}{1 + e^{-12(x-1/2)}} \,.$$

The following plots display the sequences of the polynomials p_n for $n = 2, \ldots, 25$.

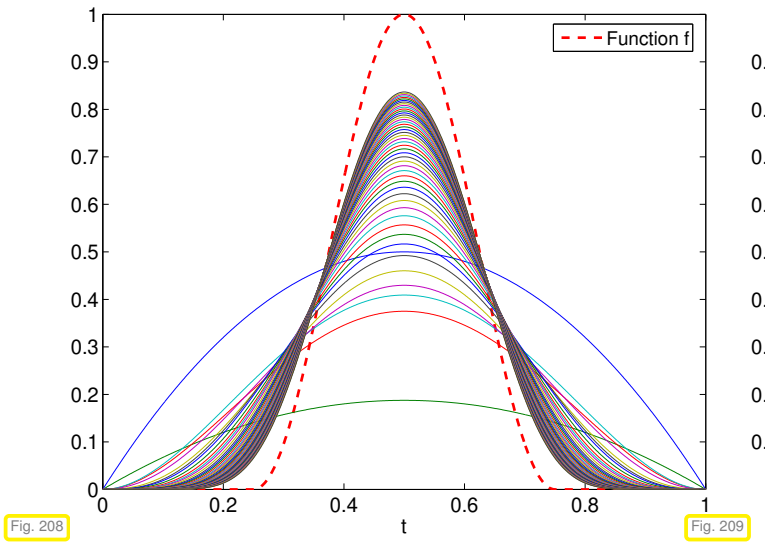


Fig. 208

$f = f_1$

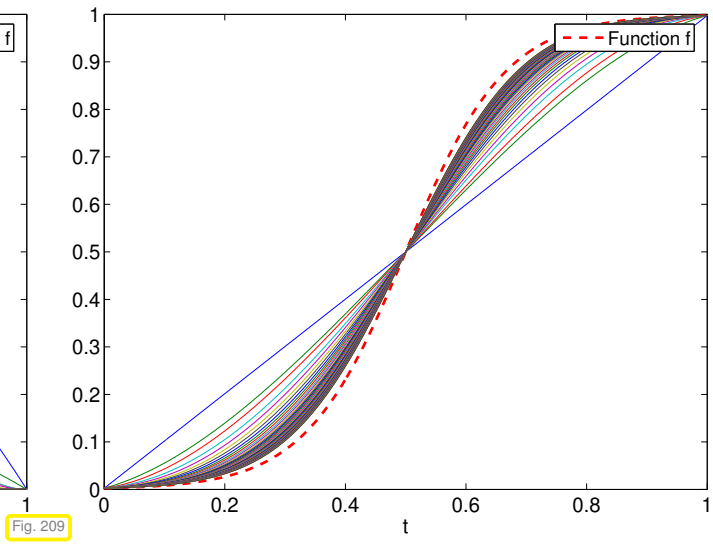


Fig. 209

$f = f_2$

We see that the Bernstein approximants "slowly" edge closer and closer to f. Apparently it takes a very large degree to get really close to f. ⌟

§6.2.1.9 (Best approximation) Now we introduce a concept needed to gauge how close an approximation scheme gets to the best possible performance.

Definition 6.2.1.10. (Size of) best approximaton error

Let $\|\cdot\|$ be a (semi-)norm on a space X of functions $I \mapsto \mathbb{K}$, $I \subset \mathbb{R}$ an interval. The (size of the) best approximation error of $f \in X$ in the space $\mathcal{P}_k$ of polynomials of degree $\leq k$ with respect to $\|\cdot\|$ is

$$\operatorname{dist}_{\|\cdot\|}(f, \mathcal{P}_k) := \inf_{p \in \mathcal{P}_k} \|f - p\| \,.$$

The notation $\operatorname{dist}_{\|\cdot\|}$ is motivated by the notation of "distance" as distance to the nearest point in a set.

For the L^2-norm $\|\cdot\|_2$ and the supremum norm $\|\cdot\|_\infty$ the best approximation error is well defined for $C = C^0(I)$.

The polynomial realizing best approximation w.r.t. $\|\cdot\|$ may neither be unique nor computable with reasonable effort. Often one is content with rather sharp upper bounds like those asserted in the next theorem, due to Jackson [Dav75, Thm. 13.3.7].

Theorem 6.2.1.11. L^∞ polynomial best approximation estimate

If $f \in C^r([-1,1])$ (r times continuously differentiable), $r \in \mathbb{N}$, then, for any polynomial degree $n \geq r$,

$$\inf_{p \in \mathcal{P}_n} \|f - p\|_{L^\infty([-1,1])} \leq (1 + \pi^2/2)^r \frac{(n-r)!}{n!} \left\| f^{(r)} \right\|_{L^\infty([-1,1])} ,$$

where

$$\left\| f^{(r)} \right\|_{L^\infty([-1,1])} := \max_{x \in [-1,1]} |f^{(r)}(x)| .$$

As above, $f^{(r)}$ stands for the r-th derivative of f. Using Stirling's formula

$$\sqrt{2\pi}\, n^{n+1/2}e^{-n} \leq n! \leq e\, n^{n+1/2}e^{-n} \quad \forall n \in \mathbb{N}\,, \tag{6.2.1.12}$$

we can get a looser bound of the form

$$\inf_{p\in\mathcal{P}_n} \|f-p\|_{L^\infty([-1,1])} \leq C(r)n^{-r}\left\|f^{(r)}\right\|_{L^\infty([-1,1])}\,, \tag{6.2.1.13}$$

with $C(r)$ dependent on r, but *independent of* f and, in particular, the polynomial *degree* n. Using the Landau symbol from Def. 1.4.1.2 we can rewrite the statement of (6.2.1.13) in asymptotic form

$$\text{(6.2.1.13)} \quad \Rightarrow \quad \inf_{p\in\mathcal{P}_n} \|f-p\|_{L^\infty([-1,1])} = O(n^{-r}) \quad \text{for} \quad n\to\infty\,.$$

⌟

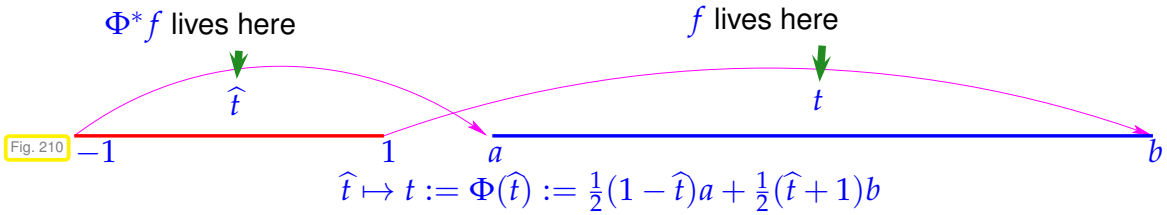
§6.2.1.14 (Transformation of polynomial approximation schemes) What if a polynomial approximation scheme is defined only on a special interval, say $[-1,1]$. Then by the following trick it can be transferred to any interval $[a,b] \subset \mathbb{R}$.

Assume that an interval $[a,b] \subset \mathbb{R}$, $a<b$, and a polynomial approximation scheme $\widehat{\mathrm{A}} : C^0([-1,1]) \to \mathcal{P}_n$ are given. Based on the affine linear mapping

$$\Phi : [-1,1] \to [a,b] \quad , \quad \Phi(\hat{t}) := a + \tfrac{1}{2}(\hat{t}+1)(b-a)\,, \quad -1 \leq \hat{t} \leq 1\,, \tag{6.2.1.15}$$

we can introduce the affine pullback of functions:

$$\boxed{\Phi^* : C^0([a,b]) \to C^0([-1,1]) \quad , \quad \Phi^*(f)(\hat{t}) := (f\circ\Phi)(\hat{t}) := f(\Phi(\hat{t}))\,, \quad -1 \leq \hat{t} \leq 1}\,. \tag{6.2.1.16}$$



We add the important observations that affine pullbacks are linear and bijective, they are isomorphisms of the involved vector spaces of functions (what is the inverse?).

Lemma 6.2.1.17. Affine pullbacks preserve polynomials

If $\Phi^* : C^0([a,b]) \to C^0([-1,1])$ is an affine pullback according to (6.2.1.15) *and* (6.2.1.16)*, then $\Phi^* : \mathcal{P}_n \to \mathcal{P}_n$ is a bijective linear mapping for any $n \in \mathbb{N}_0$.*

Proof. This is a consequence of the fact that translations and dilations take polynomials to polynomials of the *same* degree: for monomials we find

$$\Phi^*\{t \to t^n\} = \{\hat{t} \to (a + \tfrac{1}{2}(\hat{t}+1)(b-a))^n\} \in \mathcal{P}_n\,.$$

□

The lemma tells us that the spaces of polynomials of some maximal degree are *invariant under affine pullback*. Thus, we can *define* a *polynomial* approximation scheme A on $C^0([a,b])$ by

$$\mathsf{A} : C^0([a,b]) \to \mathcal{P}_n \quad , \quad \boxed{\mathsf{A} := (\Phi^*)^{-1} \circ \widehat{\mathsf{A}} \circ \Phi^*} \, , \tag{6.2.1.18}$$

whenever $\widehat{\mathsf{A}}$ is a polynomial approximation scheme on $[-1,1]$. ⌟

§6.2.1.19 (Transforming approximation error estimates) Thm. 6.2.1.11 targets only the special interval $[-1,1]$. What does it imply for polynomial best approximation on a general interval $[a,b]$? To answer this question we apply techniques from § 6.2.1.14, in particular the pullback (6.2.1.16).

We first have to study the change of norms of functions under the action of affine pullbacks:

Lemma 6.2.1.20. Transformation of norms under affine pullbacks

For every $f \in C^0([a,b])$ we have

$$\|f\|_{L^\infty([a,b])} = \|\Phi^* f\|_{L^\infty([-1,1])} \quad , \quad \|f\|_{L^2([a,b])} = \sqrt{\frac{|b-a|}{2}} \, \|\Phi^* f\|_{L^2([-1,1])} \, . \tag{6.2.1.21}$$

Proof. The first estimate should be evident, and the second is a consequence of the transformation formula for integrals [Str09, Satz 6.1.5]

$$\int_{-1}^{1} \Phi^* f(\hat{t}) \, \mathrm{d}\hat{t} = \frac{b-a}{2} \int_a^b f(t) \, \mathrm{d}t \, , \tag{6.2.1.22}$$

and the definition of the L^2-norm from (5.2.4.6). □

Thus, for norms of the approximation errors of polynomial approximation schemes defined by affine transformation (6.2.1.18) we get

$$\begin{aligned} \|f - \mathsf{A}f\|_{L^\infty([a,b])} &= \left\| \Phi^* f - \widehat{\mathsf{A}}(\Phi^* f) \right\|_{L^\infty([-1,1])} \, , \\ \|f - \mathsf{A}f\|_{L^2([a,b])} &= \sqrt{\frac{|b-a|}{2}} \left\| \Phi^* f - \widehat{\mathsf{A}}(\Phi^* f) \right\|_{L^2([-1,1])} \, , \end{aligned} \quad \forall f \in C^0([a,b]) \, . \tag{6.2.1.23}$$

Equipped with approximation error estimates for A, we can infer corresponding estimates for $\widehat{\mathsf{A}}$.

The bounds for approximation errors often involve norms of derivatives as in Thm. 6.2.1.11. Hence, it is important to understand the interplay of pullback and differentiation: By the 1D chain rule

$$\frac{d}{d\hat{t}}(\Phi^* f)(\hat{t}) = \frac{df}{dt}(\Phi(\hat{t}))\frac{d\Phi}{d\hat{t}} = \frac{df}{dt}(\Phi(\hat{t})) \cdot \tfrac{1}{2}(b-a) \, ,$$

which implies a simple scaling rule for derivatives of arbitrary order $r \in \mathbb{N}_0$:

$$(\Phi^* f)^{(r)} = \left(\frac{b-a}{2}\right)^r \Phi^*(f^{(r)}) \, . \tag{6.2.1.24}$$

Lemma 6.2.1.20

$$\left\| (\Phi^* f)^{(r)} \right\|_{L^\infty([-1,1])} = \left(\frac{b-a}{2}\right)^r \left\| f^{(r)} \right\|_{L^\infty([a,b])} \, , \quad f \in C^r([a,b]), r \in \mathbb{N}_0 \, . \tag{6.2.1.25}$$

⌟

§6.2.1.26 (Polynomial best approximation on general intervals) The estimate (6.2.1.24) together with Thm. 6.2.1.11 paves the way for bounding the polynomial best approximation error on arbitrary intervals $[a,b]$, $a,b \in \mathbb{R}$. Based on the affine mapping $\Phi : [-1,1] \to [a,b]$ from (6.2.1.15) and writing Φ^* for the pullback according to (6.2.1.16) we can chain estimates. If $f \in C^r([a,b])$ and $n \geq r$, then

$$\begin{aligned}\inf_{p\in\mathcal{P}_n} \|f-p\|_{L^\infty([a,b])} &\overset{(*)}{=} \inf_{p\in\mathcal{P}_n} \|\Phi^* f - p\|_{L^\infty([-1,1])} \\ &\overset{\text{Thm. 6.2.1.11}}{\leq} (1+\pi^2/2)^r \frac{(n-r)!}{n!} \left\| (\Phi^* f)^{(r)} \right\|_{L^\infty([-1,1])} \\ &\overset{(6.2.1.24)}{=} (1+\pi^2/2)^r \frac{(n-r)!}{n!} \left(\frac{b-a}{2}\right)^r \left\| f^{(r)} \right\|_{L^\infty([a,b])} .\end{aligned}$$

In step $(*)$ we used the result of Lemma 6.2.1.17 that $\Phi^* p \in \mathcal{P}_n$ for all $p \in \mathcal{P}_n$. Invoking the arguments that gave us (6.2.1.13), we end up with the simpler bound

$$f \in C^r([a,b]) \quad \Rightarrow \quad \inf_{p\in\mathcal{P}_n} \|f-p\|_{L^\infty([a,b])} \leq C(r) \left(\frac{b-a}{n}\right)^r \left\| f^{(r)} \right\|_{L^\infty([a,b])} . \tag{6.2.1.27}$$

Observe that the length of the interval enters the bound in r-th power. ⌟

Review question(s) 6.2.1.28 (Polynomial Approximation: Theory)

(Q6.2.1.28.A) What does Jackson's theorem Thm. 6.2.1.11 tell us about the polynomial best approximation error

$$\inf_{p\in\mathcal{P}_n} \|f-p\|_{L^\infty([-1,1])} \quad \text{for} \quad f(t) := sin(t) \ \textbf{?}$$

What does it predict for the case $r = n$?

Theorem 6.2.1.11. L^∞ polynomial best approximation estimate

If $f \in C^r([-1,1])$ (r times continuously differentiable), $r \in \mathbb{N}$, then, for any polynomial degree $n \geq r$,

$$\inf_{p\in\mathcal{P}_n} \|f-p\|_{L^\infty([-1,1])} \leq (1+\pi^2/2)^r \frac{(n-r)!}{n!} \left\| f^{(r)} \right\|_{L^\infty([-1,1])} ,$$

where $$\left\| f^{(r)} \right\|_{L^\infty([-1,1])} := \max_{x\in[-1,1]} |f^{(r)}(x)| .$$

(Q6.2.1.28.B) Consider the discontinuous step function on $[-1,1]$:

$$s(t) = \begin{cases} 0 & \text{for} \quad -1 \leq t < 0 , \\ 1 & \text{for} \quad 0 \leq t \leq 1 . \end{cases}$$

What can you say about $\inf_{p\in\mathcal{P}_n} \|s-p\|_{L^\infty([-1,1])}$?

△

6.2.2 Error Estimates for Polynomial Interpolation

Video tutorial for Section 6.2.2 "Error Estimates for Polynomial Interpolation": (12 minutes)
Download link, tablet notes

In Section 5.2.2, Cor. 5.2.2.8, we introduced the Lagrangian polynomial interpolation operator $\mathsf{I}_{\mathcal{T}} : \mathbb{K}^{n+1} \to \mathcal{P}_n$ belonging to a node set $\mathcal{T} = \{t_j\}_{j=0}^n$. In the spirit of § 6.1.0.6 it induces an approximation scheme on $C^0(I)$, $I \subset \mathbb{R}$ an interval, if $\mathcal{T} \subset I$.

Definition 6.2.2.1. Lagrangian (interpolation polynomial) approximation scheme

Given an interval $I \subset \mathbb{R}$, $n \in \mathbb{N}$, a node set $\mathcal{T} = \{t_0, \ldots, t_n\} \subset I$, the Lagrangian (interpolation polynomial) approximation scheme $\mathsf{L}_{\mathcal{T}} : C^0(I) \to \mathcal{P}_n$ is defined by

$$\mathsf{L}_{\mathcal{T}}(f) := \mathsf{I}_{\mathcal{T}}(\mathbf{y}) \in \mathcal{P}_n \quad \text{with} \quad \begin{aligned} \mathbf{y} &:= [f(t_0), \ldots, f(t_n)]^T \in \mathbb{R}^{n+1} , \\ \mathsf{I}_{\mathcal{T}}(\mathbf{y})(t_j) &= (\mathbf{y})_j , \quad j = 0, \ldots, n . \end{aligned}$$

Our goal in this section will be

to estimate the norm of the interpolation error $\|f - \mathsf{I}_{\mathcal{T}} f\|$ (for relevant norm on $C(I)$).

6.2.2.1 Convergence of Interpolation Errors

We start with an abstract discussion of "convergence of errors/error norms" in order to create awareness of what behaviors or phenomena we should be looking for and how we can detect them. You may read Section 6.2.2.2 first, if you prefer to see concrete cases and examples before adopting a higher-level perspective.

§6.2.2.2 (Families of Lagrangian interpolation polynomial approximation schemes) Already Thm. 6.2.1.11 considered the size of the best approximation error in $\mathcal{P}_n$ as a function of the polynomial degree n. In the same vein, we may study a *family* of Lagrange interpolation schemes $\{\mathsf{L}_{\mathcal{T}_n}\}_{n\in\mathbb{N}_0}$ on $I \subset \mathbb{R}$ induced by a *family of node sets* $\{\mathcal{T}_n\}_{n\in\mathbb{N}_0}$, $\mathcal{T}_n \subset I$, according to Def. 6.2.2.1.

An example for such a family of node sets on $I := [a, b]$ are the equidistant or equispaced nodes

$$\mathcal{T}_n := \{t_j^{(n)} := a + (b-a)\frac{j}{n} : j = 0, \ldots, n\} \subset I . \tag{6.2.2.3}$$

For families of Lagrange interpolation schemes $\{\mathsf{L}_{\mathcal{T}_n}\}_{n\in\mathbb{N}_0}$ we can shift the focus onto estimating the asymptotic behavior of the norm of the interpolation error for $n \to \infty$. ⌟

EXPERIMENT 6.2.2.4 (Asymptotic behavior of Lagrange interpolation error) We perform polynomial interpolation of $f(t) = \sin t$ on equispaced nodes in $I = [0, \pi]$: $\mathcal{T}_n = \{j\pi/n\}_{j=0}^n$. Write p_n for the polynomial interpolants: $p_n := \mathsf{L}_{\mathcal{T}_n} f \in \mathcal{P}_n$.

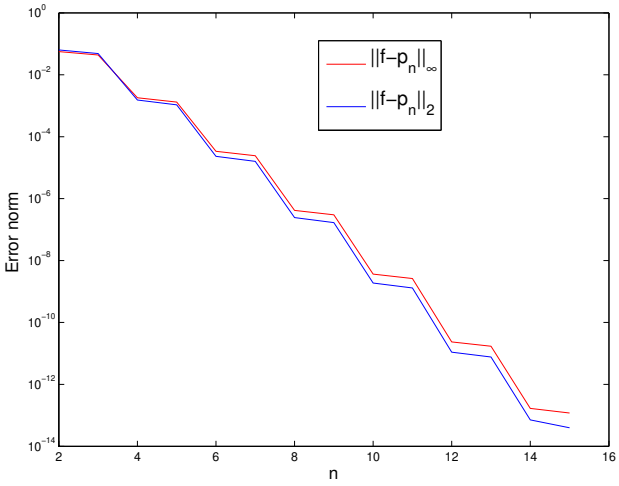


Fig. 211

In the numerical experiment the norms of the interpolation errors can be computed only approximately as follows.

- L^∞-norm: approximated by sampling on a grid of meshsize $\pi/1000$.
- L^2-norm: numerical quadrature (→ Chapter 7) with trapezoidal rule (7.5.0.4) on a grid of meshsize $\pi/1000$.

◁ approximate norms $\|f - \mathsf{L}_{\mathcal{T}_n} f\|_*$, $* = 2, \infty$.

§6.2.2.5 (Classification of asymptotic behavior of norms of the interpolation error) In the previous experiment we observed a clearly visible regular behavior of $\|f - \mathsf{L}_{\mathcal{T}_n} f\|$ as we increased the polynomial degree n. The prediction of the decay law for $\|f - \mathsf{L}_{\mathcal{T}_n} f\|$ for $n \to \infty$ is one goal in the study of interpolation errors.

Often this goal can be achieved, even if a rigorous quantitative bound for a norm of the interpolation error remains elusive. In other words, in many cases

> no quantitative bound for $\|f - \mathsf{L}_{\mathcal{T}_n} f\|$ can usually be given, but the decay of this norm of the interpolation error for increasing n can often be described precisely.

Now we introduce some important terminology for the qualitative description of the behavior of $\|f - \mathsf{L}_{\mathcal{T}_n} f\|$ as a function of the polynomial degree n. We assume that

$$\exists\, C \neq C(n) > 0: \quad \|f - \mathsf{L}_{\mathcal{T}_n} f\| \leq C\, T(n) \quad \text{for } n \to \infty\,. \tag{6.2.2.6}$$

Definition 6.2.2.7. Types of asymptotic convergence of approximation schemes

Writing $T(n)$ for the bound of the norm of the interpolation error according to (6.2.2.6) we distinguish the following *types of asymptotic behavior*:

$$\begin{array}{lll} \exists\, p > 0: & T(n) \leq n^{-p} & : \text{ algebraic convergence, with rate } p > 0\,, \\ \exists\, 0 < q < 1: & T(n) \leq q^n & : \text{ exponential convergence}\,, \end{array} \qquad \forall n \in \mathbb{N}\,.$$

The bounds are assumed to be sharp in the sense, that no bounds with larger rate p (for algebraic convergence) or smaller q (for exponential convergence) can be found.

Convergence behavior of norms of the interpolation error is often expressed by means of the Landau-O-notation, *cf.* Def. 1.4.1.2:

$$\begin{array}{ll} \text{Algebraic convergence:} & \|f - \mathsf{I}_{\mathcal{T}} f\| = O(n^{-p}) \\ \text{Exponential convergence:} & \|f - \mathsf{I}_{\mathcal{T}} f\| = O(q^n) \end{array} \qquad \text{for } n \to \infty \text{ (“asymptotic!”)}$$

Remark 6.2.2.8 (Different meanings of “convergence”) Unfortunately, as in many other fields of mathematics and beyond, also in numerical analysis the meaning of terms is context-dependent:

Beware: same concept $\leftrightarrow$ different meanings:

- convergence of a sequence (e.g. of iterates $\boldsymbol{x}^{(k)}$ $\rightarrow$ Section 8.2)
- convergence of an approximation (dependent on an approximation parameter, e.g. n)

⌟

§6.2.2.9 (Determining the type of convergence in numerical experiments $\rightarrow$ § 1.4.1.6) Given pairs (n_i, ϵ_i), $i = 1,2,3,\ldots$, $n_i \hat{=}$ polynomial degrees, $\epsilon_i \hat{=}$ (measured) norms of interpolation errors, how can we tease out the likely type of convergence according to Def. 6.2.2.7? A similar task was already encountered in § 1.4.1.6, where we had to extract information about asymptotic complexity from runtime measurements.

Assumption 6.2.2.10. Sharpness of error bounds

We assume that the error bound is (asymptotically) sharp:

$$\exists C \neq C(n): \quad \epsilon_i \approx CT(n_i) \quad \forall i \,.$$

❶ Conjectured: algebraic convergence: $\epsilon_i \approx Cn_i^{-p}$

$$\log(\epsilon_i) \approx \log(C) - p\log n_i \qquad \text{(affine linear in log-log scale)} \,.$$

Slope of line approximating $(\log n_i, log(\epsilon_i))$ predicts rate of algebraic convergence: Apply linear regression as explained in Ex. 3.1.1.5 for data points $(\log n_i, \log \epsilon_i)$ $\succ$ least squares estimate for rate p.

❶ Conjectured: exponential convergence: $\epsilon_i \approx C\exp(-\beta n_i)$

$$\log \epsilon_i \approx \log(C) - \beta n_i \qquad \text{(affine linear in lin-log scale)} \,.$$

Apply linear regression ($\rightarrow$ Ex. 3.1.1.5)to points $(n_i, \log \epsilon_i)$ $\succ$ estimate for $q := \exp(-\beta)$.

☞ Fig. 211: we suspect exponential convergence in Exp. 6.2.2.4. ⌟

EXAMPLE 6.2.2.11 (Runge's example $\rightarrow$ Ex. 5.2.4.3) We examine the polynomial interpolant of $f(t) = \frac{1}{1+t^2}$ for equispaced nodes:

$$\mathcal{T}_n := \left\{ t_j := -5 + \tfrac{10}{n} j \right\}_{j=0}^{n} , \quad j = 0,\ldots,n \quad \succ \quad y_j = \frac{1}{1+t_j^2} \,.$$

We rely on an approximate computation of the supremum norm of the interpolation error by means of sampling as in Exp. 6.2.2.4; here we used 1000 equidistant sampling points, see Code 6.2.2.12.

C++ code 6.2.2.12: Computing the interpolation error for Runge's example ↬ GITLAB

```
// Note:  "quick & dirty" implementation !
// Lambda function representing x ↦ (1+x^2)^-1
auto f = [](double x) { return 1. / (1 + x * x); };
// 1000 sampling points for approximate maximum norm
const VectorXd x = VectorXd::LinSpaced(1000, -5, 5);
// Sample function
const VectorXd fx = x.unaryExpr(f); // evaluate f at x

std::vector<double> err; // Accumulate error norms here
for (int d = 1; d <= 20; ++d) {
```

```
12      // Interpolation nodes
13      const VectorXd t = Eigen::VectorXd::LinSpaced(d + 1, -5, 5);
14      // Interpolation data values
15      const VectorXd ft = feval(f, t);
16      // Compute interpolating polynomial in monomial representation
17      const VectorXd p = polyfit(t, ft, d);
18      // Evaluate polynomial interpolant in sampling points
19      const VectorXd y = polyval(p, x);
20      // Approximate supremum norm of interpolation error
21      err.push_back((y - fx).cwiseAbs().maxCoeff());
22    }
```

Here, `polyfit()` computes the monomial coefficients of a polynomial interpolant, while `polyval()` uses the vectorized Horner scheme of Code 5.2.1.7 to evaluate the polynomial in given points. The names of the functions are borrowed from PYTHON, see numpy.poly1d.

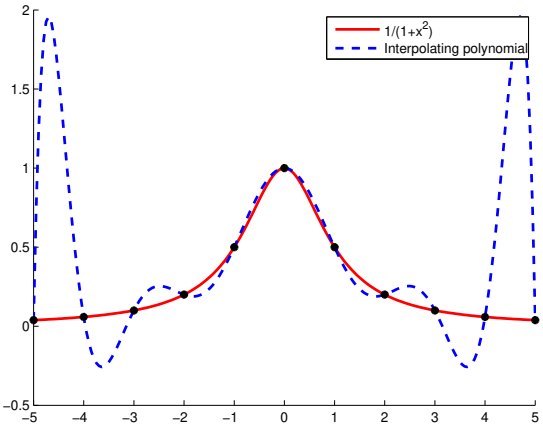


Fig. 212

Interpolating polynomial, $n = 10$

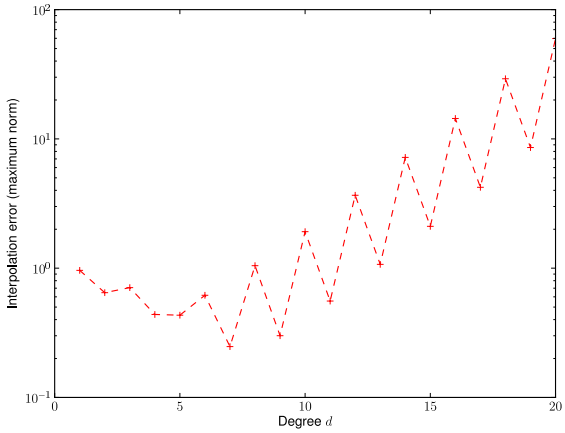


Fig. 213

Approximate $\left\|f - \mathsf{L}_{\mathcal{T}_n} f\right\|_\infty$ on $[-5,5]$

Note: approximation of $\|f - \mathsf{L}_{\mathcal{T}_n} f\|_\infty$ by *sampling* in 1000 equidistant points.

Observation: Strong oscillations of $\mathsf{I}_{\mathcal{T}} f$ near the endpoints of the interval, which seem to cause

$$\|f - \mathsf{L}_{\mathcal{T}} f\|_{L^\infty(]-5,5[)} \xrightarrow{n\to\infty} \infty \,.$$

Though polynomials possess great power to approximate functions, see Thm. 6.2.1.11 and Thm. 6.2.1.2, here polynomial interpolants fail completely. Approximation theorists even discovered the following "negative result":

Theorem 6.2.2.13. Divergent polynomial interpolants

Given a sequence of meshes of increasing size $\{\mathcal{T}_n\}_{n=1}^\infty$, $\mathcal{T}_j = \{t_0^{(n)}, \ldots, t_n^{(n)}\} \subset [a,b]$, $a \le t_0^{(n)} < t_2^{(j)} < \cdots < t_n^{(n)} \le b$, there exists a continuous function f such that the sequence of interpolating polynomials $(\mathsf{L}_{\mathcal{T}_n} f)_{n=1}^\infty$ does not converge to f uniformly as $n \to \infty$.

Review question(s) 6.2.2.14 (Convergence of interpolation errors)

(Q6.2.2.14.A) Assume that the interpolation error for some family $(\mathsf{L}_{\mathcal{T}_n})_{n\in\mathbb{N}}$, $\sharp\mathcal{T}_n = n+1$, of Lagrangian polynomial interpolation schemes and some function $f \in C^0(I)$, $I \subset \mathbb{R}$, converges algebraically ac-

cording to

$$\exists p > 0,\ C \neq C(n): \quad \|f - \mathsf{L}_{\mathcal{T}_n}\|_{L^\infty(I)} \approx Cn^{-p} \quad \forall n \in \mathbb{N}\ .$$

How do you have to raise the polynomial degree n in order to reduce the maximum norm of the interpolation error approximately by a factor of 2?

Definition 6.2.2.7. Types of asymptotic convergence of approximation schemes

Writing $T(n)$ for the bound of the norm of the interpolation error according to (6.2.2.6) we distinguish the following *types of asymptotic behavior*:

$$\begin{array}{lll} \exists\ p > 0: & T(n) \leq n^{-p} & : \text{ algebraic convergence, with rate } p > 0\ , \\ \exists\ 0 < q < 1: & T(n) \leq q^n & : \text{ exponential convergence }, \end{array} \quad \forall n \in \mathbb{N}\ .$$

The bounds are assumed to be sharp in the sense, that no bounds with larger rate p (for algebraic convergence) or smaller q (for exponential convergence) can be found.

(Q6.2.2.14.B) Let $f \in C^0(I)$, $I \subset \mathbb{R}$, be given along with a family $(\mathsf{L}_{\mathcal{T}_n})_{n\in\mathbb{N}}$, $\sharp\mathcal{T}_n = n+1$, of Lagrangian polynomial interpolation schemes. You know that the maximum norm of the interpolation error for f enjoys exponential convergence of the form

$$\exists \beta > 0\ , \quad C \neq C(n): \quad \|f - \mathsf{L}_{\mathcal{T}_n}\|_{L^\infty(I)} \approx C\exp(-\beta n) \quad \forall n \in \mathbb{N}\ .$$

How do you have to increase n in order to halve the maximum norm of the interpolation error?

(Q6.2.2.14.C) Discuss the statement:

Exponential convergence is always faster than algebraic convergence.

△

6.2.2.2 Interpolands of Finite Smoothness

Video tutorial for Section 6.2.2.2 "Error Estimates for Polynomial Interpolation: Interpolands of Finite Smoothness": (17 minutes) Download link, tablet notes

Now we aim to establish bounds for the supremum norm of the interpolation error of Lagrangian interpolation similar to the result of Jackson's best approximation theorem.

Theorem 6.2.1.11. L^∞ polynomial best approximation estimate

If $f \in C^r([-1,1])$ (r times continuously differentiable), $r \in \mathbb{N}$, then, for any polynomial degree $n \geq r$,

$$\inf_{p\in\mathcal{P}_n} \|f - p\|_{L^\infty([-1,1])} \leq (1 + \pi^2/2)^r \frac{(n-r)!}{n!} \left\|f^{(r)}\right\|_{L^\infty([-1,1])}\ ,$$

where

$$\left\|f^{(r)}\right\|_{L^\infty([-1,1])} := \max_{x\in[-1,1]} |f^{(r)}(x)|.$$

It states a result for at least continuously differentiable functions and its bound for the polynomial best approximation error involves norms of certain derivatives of the function f to be approximated. Thus some smoothness of f is required, but only a few of derivatives need exist. Thus, we say, that Thm. 6.2.1.11

deals with functions of finite smoothnes, that is, $f \in C^k$ for some $k \in \mathbb{N}$. In this section we aim to bound polynomial interpolation errors for such functions.

Theorem 6.2.2.15. Representation of interpolation error [DR08, Thm. 8.22], [Han02, Thm. 37.4]

We consider $f \in C^{n+1}(I)$ and the Lagrangian interpolation approximation scheme ($\rightarrow$ Def. 6.2.2.1) for a node set $\mathcal{T} := \{t_0, \ldots, t_n\} \subset I$. Then,

for every $t \in I$ there exists a $\tau_t \in]\min\{t, t_0, \ldots, t_n\}, \max\{t, t_0, \ldots, t_n\}[$ such that

$$f(t) - \mathsf{L}_{\mathcal{T}}(f)(t) = \frac{f^{(n+1)}(\tau_t)}{(n+1)!} \cdot \prod_{j=0}^{n}(t - t_j)\,. \tag{6.2.2.16}$$

Proof. Write $w_{\mathcal{T}}(t) := \prod_{j=0}^{n}(t - t_j) \in \mathcal{P}_{n+1}$ and *fix* $t \in I \setminus \mathcal{T}$.

$$t \neq t_j \;\Rightarrow\; w_{\mathcal{T}}(t) \neq 0 \;\Rightarrow\; \exists c = c(t) \in \mathbb{R}\colon\; f(t) - \mathsf{L}_{\mathcal{T}}(f)(t) = c(t) w_{\mathcal{T}}(t) \tag{6.2.2.17}$$

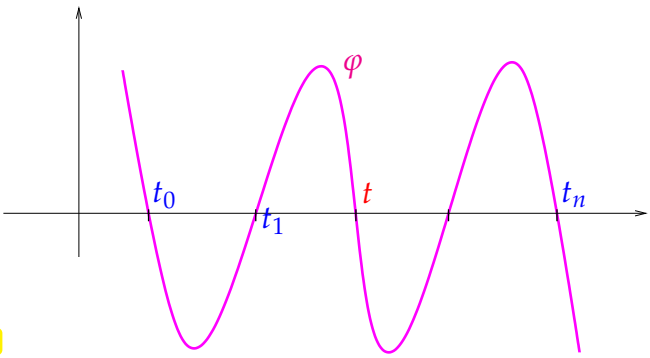


Fig. 214

Consider the *auxiliary function*

$$\varphi(x) := f(x) - \mathsf{L}_{\mathcal{T}}(f)(x) - c w_{\mathcal{T}}(x)$$

that belongs to $C^{n+1}(I)$ and has $n+2$ *distinct* zeros $t_0, \ldots, t_n, t$.

By iterated application of the mean value theorem [Str09, Thm .5.2.1]/Rolle's theorem

$$f \in C^1([a,b]),\; f(a) = f(b) = 0 \;\Rightarrow\; \exists \xi \in]a,b[\colon\; f'(\xi) = 0\,, \tag{6.2.2.18}$$

to higher and higher derivatives, we conclude that

$$\varphi^{(m)} \text{ has } n + 2 - m \text{ distinct zeros in } I.$$

$$\overset{m:=n+1}{\Rightarrow} \quad \exists \tau_t \in I\colon\; \varphi^{(n+1)}(\tau_t) = f^{(n+1)}(\tau_t) - c(n+1)! = 0\,.$$

This fixes the value of $c = \frac{f^{(n+1)}(\tau_t)}{(n+1)!}$ and by (6.2.2.17) this amounts to the assertion of the theorem. □

Remark 6.2.2.19 (Explicit representation of error of polynomial interpolation) The previous theorem can be refined:

Lemma 6.2.2.20. Error representation for polynomial Lagrange interpolation

For $f \in C^{n+1}(I)$ let $\mathsf{I}_{\mathcal{T}} \in \mathcal{P}_n$ stand for the unique Lagrange interpolant ($\rightarrow$ Thm. 5.2.2.7) of f in the node set $\mathcal{T} := \{t_0, \ldots, t_n\} \subset I$. Then for all $t \in I$ the interpolation error is

$$f(t) - \mathsf{I}_{\mathcal{T}}(f)(t) = \int_0^1 \int_0^{\tau_1} \cdots \int_0^{\tau_{n-1}} \int_0^{\tau_n} f^{(n+1)}(t_0 + \tau_1(t_1 - t_0) + \cdots + \tau_n(t_n - t_{n-1}) + \tau(t - t_n))\, \mathrm{d}\tau \mathrm{d}\tau_n \cdots \mathrm{d}\tau_1 \cdot \prod_{j=0}^{n}(t - t_j)\,.$$

The proof relies on induction on n, use (5.2.3.9) and the fundamental theorem of calculus, see [Ran00, Sect. 3.1]. ⌟

Remark 6.2.2.21 (Error representation for generalized Lagrangian interpolation) A result analogous to Lemma 6.2.2.20 holds also for general polynomial interpolation with multiple nodes as defined in (5.2.2.15). ⌟

Lemma 6.2.2.20 provides an *exact* formula (6.5.2.27) for the interpolation error. From it and also from Thm. 6.2.2.15 we can derive *estimates* for the supremum norm of the interpolation error on the interval I as follows:

❶ first bound the right hand side via $f^{(n+1)}(\tau_t) \leq \left\| f^{(n+1)} \right\|_{L^\infty(I)}$,

❷ then increase the right hand side further by switching to the maximum (in modulus) w.r.t. t (the resulting bound does no longer depend on t!),

❸ and, finally, take the maximum w.r.t. t on the left of $\leq$.

This yields the following interpolation error estimate for degree-n Lagrange interpolation on the node set $\{t_0, \ldots, t_n\}$:

$$\text{Thm. 6.2.2.15} \quad \Rightarrow \quad \boxed{\|f - \mathsf{I}_{\mathcal{T}} f\|_{L^\infty(I)} \leq \frac{\|f^{(n+1)}\|_{L^\infty(I)}}{(n+1)!} \max_{t \in I} |(t - t_0) \cdot \cdots \cdot (t - t_n)|}\,. \qquad (6.2.2.22)$$

Remark 6.2.2.23 (Significance of smoothness of interpoland) The estimate (6.2.2.22) hinges on bounds for (higher) derivatives of the interpoland f, which, essentially, should belong to $C^{n+1}(I)$. The same can be said about the estimate of Thm. 6.2.1.11.

This reflects a general truth about estimates of norms of the interpolation error:

Quantitative interpolation error estimates rely on smoothness!

⌟

EXAMPLE 6.2.2.24 (Error of polynomial interpolation Exp. 6.2.2.4 cnt'd) Now we are in a position to give a theoretical explanation for exponential convergence observed for polynomial interpolation of $f(t) =$

$\sin(t)$ on equidistant nodes: by Lemma 6.2.2.20 and (6.2.2.22)

$$\left\|f^{(k)}\right\|_{L^\infty(I)} \le 1\,, \quad \forall k \in \mathbb{N}_0 \quad \Rightarrow \quad \|f-p\|_{L^\infty(I)} \le \frac{1}{(1+n)!}\max_{t\in I}\left|(t-0)(t-\tfrac{\pi}{n})(t-\tfrac{2\pi}{n})\cdot\cdots\cdot(t-\pi)\right| \le \frac{1}{n+1}\left(\frac{\pi}{n}\right)^{n+1}.$$

➙ Uniform asymptotic (even more than) exponential convergence of the interpolation polynomials (independently of the set of nodes $\mathcal{T}$. In fact, $\|f-p\|_{L^\infty(I)}$ decays even faster than exponential!) ⌟

EXAMPLE 6.2.2.25 (Runge's example Ex. 6.2.2.11 cnt'd) How can the blow-up of the interpolation error observed in Ex. 6.2.2.11 be reconciled with Lemma 6.2.2.20 ?

Here $f(t) = \frac{1}{1+t^2}$ allows only to conclude $|f^{(n)}(t)| = 2^n n!\cdot O(|t|^{-2-n})$ for $n\to\infty$.

➙ Possible blow-up of error bound from Thm. 6.2.2.15 $\to\infty$ for $n\to\infty$. ⌟

Remark 6.2.2.26 (L^2-error estimates for polynomial interpolation) Thm. 6.2.2.15 gives error estimates for the L^∞-norm. What about other norms?

From Lemma 6.2.2.20 we know the error representation

$$f(t)-\mathrm{I}_{\mathcal{T}}(f)(t) = \int_0^1\int_0^{\tau_1}\cdots\int_0^{\tau_{n-1}}\int_0^{\tau_n} f^{(n+1)}(t_0+\tau_1(t_1-t_0)+\cdots + \tau_n(t_n-t_{n-1})+\tau(t-t_n))\,\mathrm{d}\tau\mathrm{d}\tau_n\cdots\mathrm{d}\tau_1\cdot\prod_{j=0}^n(t-t_j)\,.$$

We also repeatedly use the Cauchy-Schwarz inequality for integrals:

$$\left|\int_a^b f(t)g(t)\,\mathrm{d}t\right|^2 \le \int_a^b |f(t)|^2\,\mathrm{d}t\cdot\int_a^b |g(t)|^2\,\mathrm{d}t\,,\quad \forall f,g\in C^0([a,b])\,. \tag{6.2.2.27}$$

Thus we can estimate, $(\ldots) = t_0+\tau_1(t_1-t_0)+\cdots+\tau_n(t_n-t_{n-1})+\tau(t-t_n)$,

$$\|f-\mathrm{I}_{\mathcal{T}}(f)\|^2_{L^2(I)} = \int_I\left|\int_0^1\int_0^{\tau_1}\cdots\int_0^{\tau_{n-1}}\int_0^{\tau_n} f^{(n+1)}(\ldots)\,\mathrm{d}\tau\mathrm{d}\tau_n\cdots\mathrm{d}\tau_1\cdot\underbrace{\prod_{j=0}^n(t-t_j)}_{|t-t_j|\le|I|}\right|^2\mathrm{d}t$$

$$\le \int_I |I|^{2n+2}\underbrace{\mathrm{vol}_{(n+1)}(S_{n+1})}_{=1/(n+1)!}\int_{S_{n+1}}|f^{(n+1)}(\ldots)|^2\,\mathrm{d}\boldsymbol{\tau}\,\mathrm{d}t$$

$$= \int_I\frac{|I|^{2n+2}}{(n+1)!}\int_I\underbrace{\mathrm{vol}_{(n)}(C_{t,\tau})}_{\le 2^{(n-1)/2}/n!}|f^{(n+1)}(\tau)|^2\,\mathrm{d}\tau\mathrm{d}t\,,$$

where

$$S_{n+1} := \{\mathbf{x}\in\mathbb{R}^{n+1}\colon 0\le x_n\le x_{n-1}\le\cdots\le x_1\le 1\}\quad\text{(unit simplex)}\,,$$
$$C_{t,\tau} := \{\mathbf{x}\in S_{n+1}\colon t_0+x_1(t_1-t_0)+\cdots+x_n(t_n-t_{n-1})+x_{n+1}(t-t_n)=\tau\}\,.$$

This gives the bound for the L^2-norm of the error:

$$\Rightarrow \qquad \|f - \mathsf{L}_{\mathcal{T}}(f)\|_{L^2(I)} \leq \frac{2^{(n-1)/4}|I|^{n+1}}{\sqrt{(n+1)!n!}} \left(\int_I |f^{(n+1)}(\tau)|^2 \,\mathrm{d}\tau \right)^{1/2} . \tag{6.2.2.28}$$

Notice: $f \mapsto \left\| f^{(n)} \right\|_{L^2(I)}$ defines a seminorm on $C^{n+1}(I)$
(Sobolev-seminorm, measure of the smoothness of a function).

Estimates like (6.2.2.28) play a key role in the analysis of numerical methods for solving partial differential equations ($\rightarrow$ course "Numerical methods for partial differential equations"). ⌟

Remark 6.2.2.29 (Interpolation error estimates and the Lebesgue constant [Tre13, Thm. 15.1]) The sensitivity of a (polynomial) interpolation scheme $\mathsf{I}_{\mathcal{T}} : \mathbb{K}^{n+1} \to C^0(I)$, $\mathcal{T} \subset I$ a node set, as introduced in Section 5.2.4 and expressed by the Lebesgue constant ($\rightarrow$ Lemma 5.2.4.10)

$$\lambda_{\mathcal{T}} := \|\mathsf{I}_{\mathcal{T}}\|_{\infty\to\infty} := \sup_{\mathbf{y}\in\mathbb{R}^{n+1}\setminus\{0\}} \frac{\|\mathsf{I}_{\mathcal{T}}(\mathbf{y})\|_{L^\infty(I)}}{\|\mathbf{y}\|_\infty} ,$$

establishes an important connection between the norms of the interpolation error and of the best approximation error.

We first observe that the polynomial approximation scheme $\mathsf{L}_{\mathcal{T}}$ induced by $\mathsf{I}_{\mathcal{T}}$ *preserves polynomials* of degree $\leq n := \sharp\mathcal{T} - 1$:

$$\mathsf{L}_{\mathcal{T}} p = \mathsf{I}_{\mathcal{T}}\big([p(t)]_{t\in\mathcal{T}}\big) = p \quad \forall p \in \mathcal{P}_n . \tag{6.2.2.30}$$

Thus, by the triangle inequality, for a generic norm on $C^0(I)$ and $\|\mathsf{L}_{\mathcal{T}}\|$ designating the associated operator norm of the linear mapping $\mathsf{L}_{\mathcal{T}}$, *cf.* (5.2.4.9),

$$\|f - \mathsf{L}_{\mathcal{T}} f\| \overset{(6.2.2.30)}{=} \|(f-p) - \mathsf{L}_{\mathcal{T}}(f-p)\| \leq (1 + \|\mathsf{L}_{\mathcal{T}}\|)\|f-p\| \quad \forall p \in \mathcal{P}_n ,$$

$$\blacktriangleright \qquad \|f - \mathsf{L}_{\mathcal{T}} f\| \leq (1 + \|\mathsf{L}_{\mathcal{T}}\|) \underbrace{\inf_{p\in\mathcal{P}_n} \|f-p\|}_{\text{best approximation error}} . \tag{6.2.2.31}$$

Note that for $\|\cdot\| = \|\cdot\|_{L^\infty(I)}$, since $\left\|[f(t)]_{t\in\mathcal{T}}\right\|_\infty \leq \|f\|_{L^\infty(I)}$, we can estimate the operator norm, *cf.* (5.2.4.9),

$$\|\mathsf{L}_{\mathcal{T}}\|_{L^\infty(I)\to L^\infty(I)} \leq \|\mathsf{I}_{\mathcal{T}}\|_{\mathbb{R}^{n+1}\to L^\infty(I)} = \lambda_{\mathcal{T}} , \tag{6.2.2.32}$$

$$\blacktriangleright \qquad \boxed{\|f - \mathsf{L}_{\mathcal{T}} f\|_{L^\infty(I)} \leq (1 + \lambda_{\mathcal{T}}) \inf_{p\in\mathcal{P}_n} \|f-p\|_{L^\infty(I)} \quad \forall f \in C^0(I)} . \tag{6.2.2.33}$$

Hence, if a bound for $\lambda_{\mathcal{T}}$ is available, the best approximation error estimate of Thm. 6.2.1.11 immediately yields interpolation error estimates. ⌟

Review question(s) 6.2.2.34 (Interpolands of finite smoothness)

(Q6.2.2.34.A) For the interpolation error observed in Exp. 6.2.2.4 we found the bound

$$\|f - \mathsf{L}_{\mathcal{T}_n} f\|_{L^\infty(I)} \leq \frac{1}{n+1}\left(\frac{\pi}{n}\right)^{n+1} \quad \forall n \in \mathbb{N} .$$

Explain why this behavior of the maximum norm of the interpolation error is also called "superexponential convergence".

(Q6.2.2.34.B) For Lagrange polynomial interpolation we have seen the following error representation.

Theorem 6.2.2.15. Representation of interpolation error

We consider $f \in C^{n+1}(I)$ and the Lagrangian interpolation approximation scheme (→ Def. 6.2.2.1) for a node set $\mathcal{T} := \{t_0, \ldots, t_n\} \subset I$. Then,

for every $t \in I$ there exists a $\tau_t \in]\min\{t, t_0, \ldots, t_n\}, \max\{t, t_0, \ldots, t_n\}[$ such that

$$f(t) - \mathsf{L}_{\mathcal{T}}(f)(t) = \frac{f^{(n+1)}(\tau_t)}{(n+1)!} \cdot \prod_{j=0}^{n}(t - t_j) \,. \tag{6.5.2.27}$$

This was obtained by "counting the zeros" of derivatives of the auxiliary function

$$\varphi(x) := f(x) - \mathsf{L}_{\mathcal{T}}(f)(x) - c w_{\mathcal{T}}(x) \,, \quad a \le x \le b \,,$$

where $w_{\mathcal{T}}$ was the nodal polynomial belonging to the node set $\mathcal{T}$ and $c \in \mathbb{R}$ was chosen to ensure $\varphi(t) = 0$ for a fixed $t \in [a,b] \setminus \mathcal{T}$.

Now we consider the cubic Hermite interpolation operator

$$\mathsf{H} : C^1([a,b]) \to \mathcal{P}_3 \quad , \quad p := \mathsf{H}(f) \quad \text{such that} \quad \begin{array}{l} p(a) = f(a) \quad , \quad p(b) = f(b) \,, \\ p'(a) = f'(a) \quad , \quad p'(b) = f'(b) \,. \end{array}$$

Here p', f' stands for the derivative. Fixing $t \in]a,b]$ and using the auxiliary function

$$\psi(x) := f(x) - \mathsf{H}(f)(x) - c(x-a)^2(x-b)^2 \,, \quad a \le x \le b \,,$$

derive an error representation formula for $f(t) = \mathsf{H}(t)(t)$.

(Q6.2.2.34.C) Assume that for a sequence of node sets $(\mathcal{T}_n)_{n\in\mathbb{N}}$, the associated Lebesgue constants of Lagrange polynomial interpolation with respect to the maximum norm behave like $\lambda_{\mathcal{T}_n} = O(n^q)$ for $n \to \infty$ and some $q \in \mathbb{N}$.

What conclusions about the maximum norm of the interpolation error of Lagrange interpolation can then be drawn from Jackson's theorem?

Theorem 6.2.1.11. L^∞ polynomial best approximation estimate

If $f \in C^r([-1,1])$ (r times continuously differentiable), $r \in \mathbb{N}$, then, for any polynomial degree $n \ge r$,

$$\inf_{p\in\mathcal{P}_n} \|f - p\|_{L^\infty([-1,1])} \le (1 + \pi^2/2)^r \frac{(n-r)!}{n!} \left\| f^{(r)} \right\|_{L^\infty([-1,1])} \,,$$

where $$\left\| f^{(r)} \right\|_{L^\infty([-1,1])} := \max_{x\in[-1,1]} |f^{(r)}(x)| .$$

△

6.2.2.3 Analytic Interpolands

Video tutorial for Section 6.2.2.3 "Error Estimates for Polynomial Interpolation: Analytic Interpolands": (27 minutes) Download link, tablet notes

We have seen that for some Lagrangian approximation schemes applied to certain functions we can observe exponential convergence (→ Def. 6.2.2.7) of the approximation error for increasing polynomial degree. This section presents a class of interpolands, which often enable this convergence.

Definition 6.2.2.35. Real-analytic functions

A function $f \in C^\infty(I)$ defined on an open interval $I \subset \mathbb{R}$ is called **(real-)analytic** on I, if it possesses a convergent Taylor series at every point $t_0 \in I$:

$$\forall t_0 \in I: \quad \exists \rho = \rho(t_0) > 0: \quad f(t) = \sum_{k=0}^{\infty} \frac{f^{(k)}(t_0)}{k!}(t-t_0)^k \quad \forall t \in I: \ |t-t_0| < \rho(t_0) \ .$$

$\rho(t_0)$ is called the **radius of convergence** of the Taylor series.

We may say that an analytic function locally agrees with a "polynomial of degree ∞", because this is exactly what a convergent **power series**

$$f(t) = \sum_{k=0}^{\infty} a_k (t-t_0)^k \ , \quad a_k \in \mathbb{R} \ , \tag{6.2.2.36}$$

represents. Note that f need not be given by its Taylor power series on the whole interval I. Those may converge only on small sub-intervals. Def. 6.2.2.35 merely tells us that I can be covered by such sub-intevals.

EXAMPLE 6.2.2.37 (Hump function) We consider the C^∞-function

$$f(t) := \frac{1}{1+t^2} \ , \quad t \in \mathbb{R} \ ,$$

on the interval $I = [-5,5]$. By the geometric sum formula we have as Taylor series at $t_0 = 0$:

$$f(t) = \sum_{k=0}^{\infty} (-1)^k t^{2k} \ , \quad |t| < 1 \ ,$$

whose radius of convergence $= 1$. More generally, deeper theory tells us that the Taylor series at t_0 has radius of convergence $= \sqrt{t_0^2+1}$. Thus, in $[-5,5]$ cannot be represented by a single power series, though it is a perfectly smooth function. ⌟

EXAMPLE 6.2.2.38 (Square root function) We consider the the function $f(t) := \sqrt{t}$ on $I :=]0,1[$. From calculus we know the power series

$$\sqrt{1+x} = 1 + \sum_{k=1}^{\infty} (-1)^k \prod_{j=0}^{k-1} \frac{j-\frac{1}{2}}{j+1} x^k \ , \quad |x| < 1 \ . \tag{6.2.2.39}$$

It converges for a all x with $|x| < 1$ [Str09, Satz 3.7.2]. Using, (6.2.2.39) we get the Taylor series for f at $t_0 \in I$

$$\sqrt{t} = \sqrt{t_0}\sqrt{1+\frac{t-t_0}{t_0}} = \sqrt{t_0} + \sum_{k=1}^{\infty} (-1)^k \prod_{j=0}^{k-1} \frac{j-\frac{1}{2}}{j+1} \left(\frac{t-t_0}{t_0}\right)^k \ , \quad \left|\frac{t-t_0}{t_0}\right| < 1 \ .$$

This series converges only in the open ball around t_0 of radius $|t_0|$. The closer we get to the "singular point" $t = 0$, the smaller the radius of convergence. ⌟

Remark 6.2.2.40 (Analytic functions everywhere) Analyticity of a function seems to be a very special property confined the functions that are given through simple formulas. Yet, this is not true:

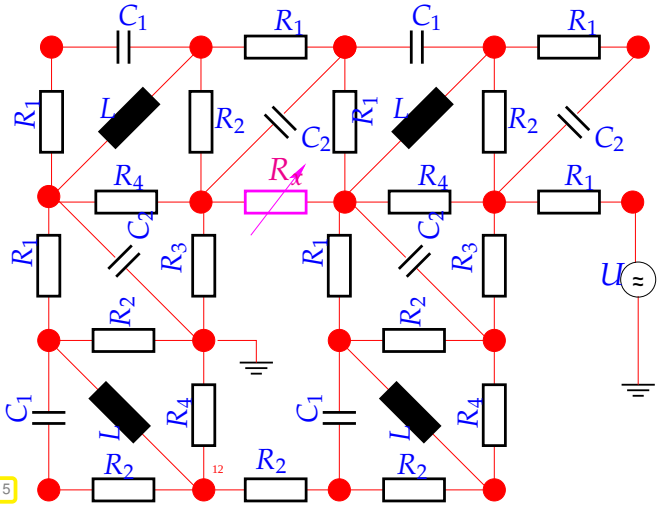


Fig. 215

◁ Linear RLC-circuit with variable resistor, see also Ex. 2.6.0.24.

For the shown linear electric circuit all branch currents and voltages will depend analytically on the resistance R_x of the variable resistor. This a consequence of the fact that

$$t \mapsto \mathbf{v}(t)^\top \mathbf{A}(t)^{-1}\mathbf{u}(t)\,, \quad t \in I \subset \mathbb{R}\,,$$

is an analytic function (where well-defined), provided that the components of $\mathbf{u}(t), \mathbf{v}(t) \in \mathbb{R}^n$ and the entries of $\mathbf{A}(t) \in \mathbb{R}^{n,n}$ are analytic functions of t.

This remark remains true for many output quantities of physical models considered as functions of model parameters or input quantities. ⌟

§6.2.2.41 (Approximation by truncated power series) A first glimpse of the relevance of analyticity for polynomial approximation: Let $I \subset \mathbb{R}$ be an *closed* interval and f real-analytic on I according to Def. 6.2.2.35.

We add a stronger assumption: There is a $t_0 \in I$ and $\rho > 0$ such that

- the Taylor series of f at t_0

$$f(t) = \sum_{k=0}^{\infty} a_k (t - t_0)^k\,, \quad a_k \in \mathbb{R}\,, \tag{6.2.2.42}$$

  has radius of convergence ρ,
- I is contained in the ρ-ball around t_0

$$I \subset \{t \in \mathbb{R} : |t - t_0| < \rho\}\,.$$

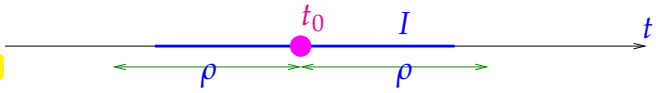


Fig. 216

We approximate f by its truncated Taylor series

$$p_n(t) := \sum_{k=0}^{n} a_k (t - t_0)^k \in \mathcal{P}_n\,. \tag{6.2.2.43}$$

Since I is closed we can find $0 < r < \rho$ such that

$$I \subset \{t \in \mathbb{R} : |t - t_0| \le r\}\,. \tag{6.2.2.44}$$

The convergence theory of power series [Str09, Sect. 3.7] ensures that for any $r < R < \rho$

$$\sum_{k=0}^{\infty} |a_k| R^k =: C \le \infty\,. \tag{6.2.2.45}$$

Now we combine (6.2.2.42), (6.2.2.43), (6.2.2.45): for arbitrary $t \in I$ we arrive at

$$|f(t) - p_n(t)| = \left| \sum_{k=n+1}^{\infty} a_k (t - t_0)^k \right| \le \sum_{k=n+1}^{\infty} |a_k| R^k \left(\frac{t - t_0}{R}\right)^k \le C \left(\frac{r}{R}\right)^{n+1}\,.$$

$$\blacktriangleright \quad \|f - p_n\|_{L^\infty(I)} \le C\, q^{n+1} \quad \forall n \in \mathbb{N} \quad \text{with} \quad q := r/R < 1\,. \tag{6.2.2.46}$$

This confirms exponential convergence of the approximation error incurred by truncating the Taylor series.

⌟

The previous § heavily relied on the assumption that f possesses a power series representation that converges on the entire interval I and *even beyond*. As we see from Ex. 6.2.2.37 and Ex. 6.2.2.38 this will usually not be the case.

This is why we continue with a key observation: a power series makes perfect sense for complex arguments. Any function $f \in C^\infty(I)$ that, locally, for $t_0 \in I$ can be written as

$$f(t) = \sum_{n=0}^{\infty} a_n (t - t_0)^n \,, \quad a_n \in \mathbb{R} \,, \quad \forall t \in \mathbb{R} : \ |t - t_0| < \rho(t_0) \,,$$

can be extended to a complex-valued function defined on a $\mathbb{C}$-disk around t_0 by

$$f(z) = \sum_{n=0}^{\infty} a_n (z - t_0)^n \,, \quad \forall z \in \mathbb{C} : \ |z - t_0| < \rho(t_0) \,.$$

Real- and complex-analytic functions

Every real-analytic function on $I \subset \mathbb{R}$ can be extended to a complex-analytic function on some open set $D \subset \mathbb{C}$ with $I \subset D$.

For, this reason no distinction between real and complex analyticity has to be made. For the sake of completeness, the definition of an analytic function on $D \subset \mathbb{C}$ is given, nevertheless.

Definition 6.2.2.48. Analyticity of a complex valued function

let $D \subset \mathbb{C}$ be an *open* set in the complex plane. A function $f : D \to \mathbb{C}$ is called (complex)-analytic/holomorphic in D, if for *every* point $z \in D$ one can find $\rho(z) > 0$ and $a_k = a_k(z) \in \mathbb{C}$, $k \in \mathbb{N}_0$, such that

$$f(w) = \sum_{k=0}^{\infty} a_k (w - z)^k \quad \forall w \in D : \ |z - w| < \rho(z) \,.$$

Be aware, that also this definition asserts the existence of a *local* power series representation of f only.

§6.2.2.49 (Residue calculus) Why go $\mathbb{C}$? The reason is that this permits us to harness powerful tool from complex analysis ($\to$ course 401-0302-10L Mathematische Methoden in the BSc program CSE), a field of mathematics, which studies analytic functions. One of theses tools is the residue theorem.

Theorem 6.2.2.50. Residue theorem [Rem84, Ch. 13], [Akv+25, Satz 4.24]

Let $D \subset \mathbb{C}$ be an open set, $G \subset D$ a closed set contained in D, $\gamma := \partial G$ its (piecewise smooth and oriented) boundary, and Π a finite *set contained in the interior of G.*

Then for each function f that is analytic in $D \setminus \Pi$ holds

$$\frac{1}{2\pi\imath} \int_\gamma f(z)\, dz = \sum_{p \in \Pi} \operatorname{res}_p f \,,$$

where $\operatorname{res}_p f$ is the residual of f in $p \in \mathbb{C}$.

- Note that the integral $\int_\gamma$ in Thm. 6.2.2.50 is a path integral in the complex plane ("contour integral", see [Akv+25, Sect. 3.3]): If the path of integration γ is described by a parameterization $\tau \in J \mapsto \gamma(\tau) \in \mathbb{C}$, $J \subset \mathbb{R}$, then

$$\int_\gamma f(z)\,\mathrm{d}z := \int_J f(\gamma(\tau)) \cdot \dot{\gamma}(\tau)\,\mathrm{d}\tau\,, \tag{6.2.2.51}$$

  where $\dot{\gamma}$ designates the derivative of γ with respect to the parameter, and $\cdot$ indicates multiplication in $\mathbb{C}$. For contour integrals we have the estimate

$$\left|\int_\gamma f(z)\,\mathrm{d}z\right| \le |\gamma| \max_{z\in\gamma} |f(z)|\,. \tag{6.2.2.52}$$

- Π often stands for the set of poles of f, that is, points where "f attains the value ∞".

The residue theorem is very useful, because there are simple formulas for $\operatorname{res}_p f$:

Lemma 6.2.2.53. Residue formula for quotients

let g and h be complex valued functions that are both analytic in a neighborhood of $p \in \mathbb{C}$, and satisfy $h(p) = 0$, $h'(p) \neq 0$. Then

$$\operatorname{res}_p \frac{g}{h} = \frac{g(p)}{h'(p)}\,.$$

⌟

§6.2.2.54 (Residue remainder formula for Lagrange interpolation) Now we consider a polynomial Lagrangian approximation scheme on the interval $I := [a,b] \subset \mathbb{R}$, based on the node set $\mathcal{T} := \{t_0,\ldots,t_n\} \subset I$.

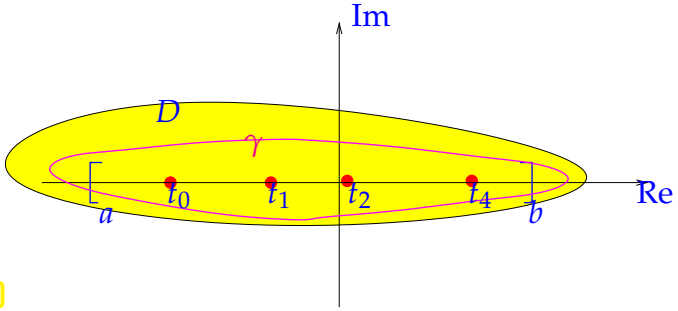


Fig. 217

Assumption 6.2.2.55. Analyticity of interpoland

We assume that the interpoland $f : [a,b] \to \mathbb{C}$ *can be extended* to a function $f : D \subset \mathbb{C} \to \mathbb{C}$, which is analytic (→ Def. 6.2.2.48) on the open set $D \subset \mathbb{C}$ with $[a,b] \subset D$.

Key is the following representation of the Lagrange polynomials (5.2.2.4) for node set $\mathcal{T} = \{t_0,\ldots,t_n\}$:

$$L_j(t) = \prod_{k=0,k\neq j}^{n} \frac{t-t_k}{t_j-t_k} = \frac{w(t)}{(t-t_j)\prod_{k=0,k\neq j}^{n}(t_j-t_k)} = \frac{w(t)}{(t-t_j)w'(t_j)}\,, \tag{6.2.2.56}$$

$$\text{where} \quad w(t) = (t-t_0)\cdot\cdots\cdot(t-t_n) \in \mathcal{P}_{n+1}\,.$$

Consider the following parameter dependent function g_t, whose set of poles in D is $\Pi = \{t, t_0,\ldots,t_n\}$

$$g_t(z) := \frac{f(z)}{(z-t)w(z)}\,,\quad z \in \mathbb{C}\setminus\Pi\,,\quad t \in [a,b]\setminus\{t_0,\ldots,t_n\}\,.$$

➢ g_t is analytic on $D \setminus \{t, t_0, \ldots, t_n\}$ (t must be regarded as parameter!)

Apply residue theorem Thm. 6.2.2.50 to g_t and a closed path of integration $\gamma \subset D$ winding once around $[a,b]$, such that its interior is simply connected, see the magenta curve in Fig. 217:

$$\frac{1}{2\pi\imath}\int_\gamma g_t(z)\,\mathrm{d}z = \operatorname{res}_t g_t + \sum_{j=0}^n \operatorname{res}_{t_j} g_t \overset{\text{Lemma 6.2.2.53}}{=} \frac{f(t)}{w(t)} + \sum_{j=0}^n \frac{f(t_j)}{(t_j - t)w'(t_j)}$$

Possible, because all zeros of w are single zeros!

$$f(t) = \underbrace{-\sum_{j=0}^n f(t_j)\underbrace{\frac{w(t)}{(t_j-t)w'(t_j)}}_{-\text{Lagrange polynomial !}}}_{\text{polynomial interpolant !}} + \underbrace{\frac{w(t)}{2\pi\imath}\int_\gamma g_t(z)\,\mathrm{d}z}_{\text{interpolation error !}} . \tag{6.2.2.57}$$

This is the famous Hermite integral formula [Tre13, Ch. 11], a representation formula for the interpolation error, an alternative to that of Thm. 6.2.2.15 and Lemma 6.2.2.20. We conclude that for all $t \in [a,b]$

$$|f(t) - \mathsf{L}_\mathcal{T} f(t)| = \left|\frac{w(t)}{2\pi\imath}\int_\gamma \frac{f(z)}{(z-t)w(z)}\,\mathrm{d}z\right| \le \frac{|\gamma|}{2\pi}\frac{\max\limits_{a\le\tau\le b}|w(\tau)|}{\min\limits_{z\in\gamma}|w(z)|}\cdot\frac{\max\limits_{z\in\gamma}|f(z)|}{\operatorname{dist}([a,b],\gamma)} . \tag{6.2.2.58}$$

In concrete setting, in order to exploit the estimate (6.2.2.58) to study the n-dependence of the supremum norm of the interpolation error, we need to know

- an upper bound for $|w(t)|$ for $a \le t \le b$,
- an a lower bound for $|w(z)|$, $z \in \gamma$, for a suitable path of integration $\gamma \subset D$,
- a lower bound for the distance of the path γ and the interval $[a,b]$ in the complex plane.

⌟

Remark 6.2.2.59 (Frobenius' derivation of the Hermite integral formula [Boo05, Sect. 9]) We give a more elementary alternative derivation of the interpolation error formula in (6.2.2.57), which does not rely on the residual theorem. We retain the node set $\mathcal{T} := \{t_0, \ldots, t_n\} \subset [a,b]$ and Ass. 6.2.2.55. We write

$$w_j(t) := (t-t_0)\cdot\cdots\cdot(t-t_{j-1}) , \quad j \in \{1,\ldots,n+1\} \quad , \quad w_0 := 1 ,$$

for nodal polynomials: $w_j \in \mathcal{P}_j$. They obviously satisfy

$$tw_{j-1}(t) = w_j(t) + t_{j-1}w_{j-1}(t) , \quad t \in \mathbb{R} , \quad j = 1,\ldots,n+1 . \tag{6.2.2.60}$$

We pick $z \in \mathbb{C} \setminus \mathcal{T}$, replace $t \to z$ in (6.2.2.60), and divide by $w_{j-1}(z)w_j(z)$, which yields

$$\frac{z}{w_j(z)} = \frac{1}{w_{j-1}(z)} + \frac{t_{j-1}}{w_j(z)} , \quad z \in \mathbb{C} \setminus \mathcal{T} , \quad j = 1,\ldots,n+1 . \tag{6.2.2.61}$$

From the two identities (6.2.2.60) and (6.2.2.61) we obtain by summation

$$z\cdot\sum_{j=1}^{n+1}\frac{w_{j-1}(t)}{w_j(z)} = \sum_{j=1}^{n+1}\left\{\frac{w_{j-1}(t)}{w_{j-1}(z)} + \frac{t_{j-1}w_{j-1}(t)}{w_j(z)}\right\} , \tag{6.2.2.62}$$

$$t \cdot \sum_{j=1}^{n+1} \frac{w_{j-1}(t)}{w_j(z)} = \sum_{j=1}^{n+1} \left\{ \frac{w_j(t)}{w_j(z)} + \frac{t_{j-1} w_{j-1}(t)}{w_j(z)} \right\} . \tag{6.2.2.63}$$

Subtracting the last two formulas we get a telescopic sum:

$$(z-t) \sum_{j=1}^{n+1} \frac{w_{j-1}(t)}{w_j(z)} = \sum_{j=1}^{n+1} \left\{ \frac{w_{j-1}(t)}{w_{j-1}(z)} - \frac{w_j(t)}{w_j(z)} \right\} = 1 - \frac{w_{n+1}(t)}{w_{n+1}(z)} , \quad z \notin \mathcal{T} , \quad t \in \mathbb{R} .$$

$$\frac{1}{z-t} = \sum_{j=1}^{n+1} \frac{w_{j-1}(t)}{w_j(z)} + \frac{w_{n+1}(t)}{(z-t) w_{n+1}(z)} , \quad t \in \mathbb{R} , \quad z \notin \mathcal{T} \cup \{t\} . \tag{6.2.2.64}$$

Recall the Cauchy integral representation formula

$$f(t) = \frac{1}{2\pi\imath} \int_\gamma \frac{f(z)}{z-t} \, \mathrm{d}z , \quad t \in [a,b] , \tag{6.2.2.65}$$

where $\gamma \subset D$ is a curve enclosing $[a,b]$ as drawn in Fig. 217. We can rewrite (6.2.2.65) using (6.2.2.64),

$$f(t) = \underbrace{\sum_{j=1}^{n+1} \left\{ \frac{1}{2\pi\imath} \int_\gamma \frac{f(z)}{w_j(z)} \, \mathrm{d}z \cdot w_{j-1}(t) \right\}}_{=:p(t)} + \frac{1}{2\pi\imath} \int_\gamma \frac{f(z)}{(z-t) w_{n+1}(z)} \, \mathrm{d}z \cdot w_{n+1}(t) . \tag{6.2.2.66}$$

By the definition of w_j the function $t \mapsto p(t)$ is a polynomial of degree n that interpolates f in $\mathcal{T}$, the set of zeros of w_{n+1}. Hence p is the unique Lagrange polynomial interpolant of f with respect to the node set $\mathcal{T}$. Thus the second term in (6.2.2.66) represents the interpolation error and obviously agrees with the formula found in (6.2.2.57). ⌟

Remark 6.2.2.67 (Determining the domain of analyticity) The subset of $\mathbb{C}$, where a function f given by a formula, is analytic can often be determined without computing derivatives using the following consequence of the chain rule:

Theorem 6.2.2.68. Composition and products of analytic functions

If $f, h : D \subset \mathbb{C} \to \mathbb{C}$ and $g : U \subset \mathbb{C} \to \mathbb{C}$ are analytic in the open sets D and U, respectively, then
(i) the composition $f \circ g$ is analytic in $\{z \in U : g(z) \in D\}$,
(ii) the product $f \cdot h$ is analytic on D.

This can be combined with the following facts:

- Polynomials, $\exp(z)$, $\sin(z)$, $\cos(z)$, $\sinh(z)$, $\cosh(z)$ are analytic on $\mathbb{C}$ (entire functions).
- Rational functions (quotients of polynomials) are analytic everywhere except in the zeros of their denominator.
- The square root $z \to \sqrt{z}$ is analytic in $\mathbb{C} \setminus]-\infty, 0]$.

For example, according to these rules the function $f(t) = (1+t^2)^{-1}$ can be extended to an analytic function on $\mathbb{C} \setminus \{-\imath, \imath\}$. If $\mathbf{A} \in \mathbb{C}^{n,n}$, $\mathbf{b} \in \mathbb{R}^n$, $n \in \mathbb{N}$, then $z \mapsto (\mathbf{A} + z\mathbf{I})^{-1}\mathbf{b} \in \mathbb{C}^n$ is (componentwise) analytic in $\mathbb{C} \setminus \sigma(\mathbf{A})$, where $\sigma(\mathbf{A})$ denotes the set of eigenvalues of $\mathbf{A}$ (the spectrum). ⌟

Review question(s) 6.2.2.69 (Analytic interpolants)

(Q6.2.2.69.A) We consider the Lagrange polynomial interpolation of the entire function $f(t) = e^t$ on $I = [0,1]$ with equidistant nodes $\mathcal{T} := \left\{t_j = \frac{j}{n}\right\}_{j=0}^{n}$, $n \in \mathbb{N}$. As integration path in the residual remainder estimate

$$|f(t) - \mathsf{L}_{\mathcal{T}} f(t)| = \left| \frac{w(t)}{2\pi\imath} \int_\gamma \frac{f(z)}{(z-t)w(z)} \,\mathrm{d}z \right| \leq \frac{|\gamma|}{2\pi} \frac{\max\limits_{a\leq\tau\leq b} |w(\tau)|}{\min\limits_{z\in\gamma} |w(z)|} \cdot \frac{\max\limits_{z\in\gamma} |f(z)|}{\operatorname{dist}([a,b],\gamma)} \,. \tag{6.2.2.58}$$

we choose the square path (endowed with an arbitrary orientation)

$$\gamma := \{-1 + t\imath\}_{t=-3}^{3} \cup \{t + 3\imath\}_{t=-1}^{2} \cup \{2 + t\imath\}_{t=3}^{-3} \cup \{t - 3\imath\}_{t=2}^{-1} \,.$$

Work out the bound from (6.2.2.58) in this case.

(Q6.2.2.69.B) [Interpolands analytic on stadium domains] Assume that $f : [-1,1] \to \mathbb{C}$ possesses an analytic extension to an open set $D \subset \mathbb{C}$ containing the stadium domain

$$S := \left\{ z \in \mathbb{C} \colon \min_{-1\leq t\leq 1} |z - t| \leq 1 \right\} .$$

Let $\mathcal{T}_n := \{t_0, \ldots, t_n\} \subset [-1,1]$ a sequence of sets of interpolation nodes and $\mathsf{I}_n : C^0([-1,1]) \to \mathcal{P}_n$ the associated family of Lagrangian polynomial interpolation operators. Based on (6.2.2.58) show that the supremum norm of the interpolation error $\|f - \mathsf{I}_n f\|_{\infty,[-1,1]}$ *converges* to zero *exponentially* in the degree n as $n \to \infty$.

(Q6.2.2.69.C) For Lagrange polynomial interpolation of $f(t) = \frac{1}{1+t^2}$ in the nodes $\mathcal{T} := \left\{-5 + \frac{10}{n} j\right\}_{j=0}^{n}$, sketch a valid integration path $\gamma \subset \mathbb{C}$ for the estimate (6.2.2.58).

(Q6.2.2.69.D) What is the problem, if you want to apply (6.2.2.58) to estimate the error of Lagrange polynomial interpolation of $t \mapsto \sqrt{t}$ in equidistant nodes in $[0,1]$?

(Q6.2.2.69.E) Find largest subset of the complex plane $\mathbb{C}$, to which the logistic curve function

$$f(t) := \frac{1}{1 + \exp(-t)} \,, \quad t \in \mathbb{R} \,,$$

possesses an analytic extension.

Note that from Euler's formula

$$\exp(x + \imath y) = \exp(x)(\cos(y) + \imath \sin(y)) \,, \quad x, y \in \mathbb{R} \,.$$

△

6.2.3 Chebychev Interpolation

As pointed out in § 6.1.0.6, when we build approximation schemes from interpolation schemes, we have the extra freedom to choose the sampling points (= interpolation nodes). Now, based on the insight into the structure of the interpolation error gained from Thm. 6.2.2.15, we seek to choose "optimal" sampling points. They will give rise to the so-called Chebychev polynomial approximation schemes, also known as Chebychev interpolation.

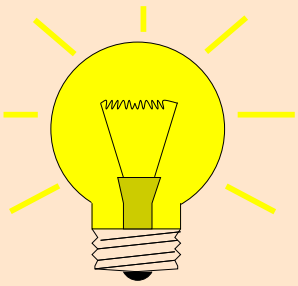
6.2.3.1 Motivation and Definition

Video tutorial for Section 6.2.3.1 "Chebychev Interpolation: Motivation and Definition": (21 minutes) Download link, tablet notes

Setting:
- Without loss of generality (→ § 6.2.1.14): $I = [-1,1]$,
- interpoland $f : I \to \mathbb{R}$ at least continuous, $f \in C^0(I)$,
- set of interpolation nodes $\mathcal{T} := \{-1 \le t_0 < t_1 < \cdots < t_{n-1} < t_n \le 1\}$, $n \in \mathbb{N}$.

Recall Thm. 6.2.2.15:
$$\|f - \mathsf{L}_{\mathcal{T}} f\|_{L^\infty(I)} \le \frac{1}{(n+1)!} \left\| f^{(n+1)} \right\|_{L^\infty(I)} \|w\|_{L^\infty(I)} ,$$
with nodal polynomial $w(t) := (t - t_0) \cdot \cdots \cdot (t - t_n)$.

Optimal choice of interpolation nodes independent of interpoland



Idea: choose nodes $t_0, \ldots, t_n$ such that $\|w\|_{L^\infty(I)}$ is minimal!

This is equivalent to finding a polynomial $q \in \mathcal{P}_{n+1}$
- with leading coefficient $= 1$,
- such that it minimizes the norm $\|q\|_{L^\infty(I)}$.

➤ Then choose nodes $t_0, \ldots, t_n$ as zeros of q.
(caution: all t_j must lie in I!)

Remark 6.2.3.2 (A priori and a posteriori choice of optimal interpolation nodes) We stress that we aim for an "optimal" *a priori* choice of interpolation nodes, a choice that is made **before** any information about the interpoland becomes available.

Of course, an *a posteriori* choice based on information gleaned from evaluations of the interpoland f may yield much better interpolants (in the sense of smaller norm of the interpolation error). Many modern algorithms employ this a posteriori adaptive approximation policy, but this chapter will not cover them.

However, see Section 7.6 for the discussion of an a posteriori adaptive approach for the numerical approximation of definite integrals. ⌟

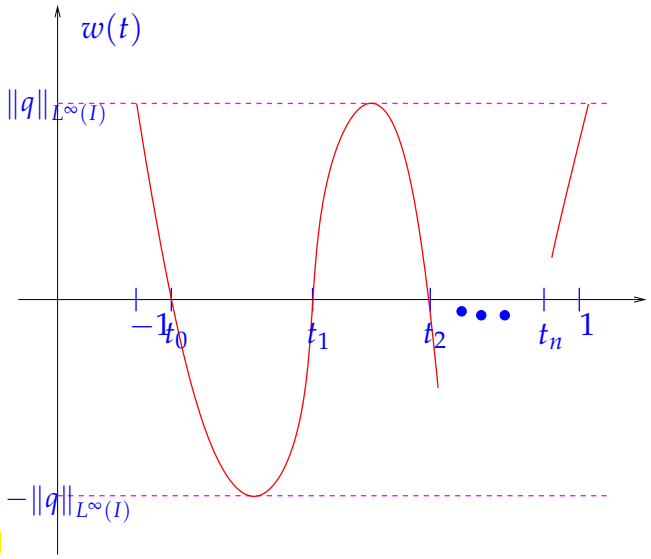


Fig. 218

Requirements on q (by heuristic reasoning)

Optimal polynomials q will exist, but they seem to be elusive. First, we develop some insights in how they must "look like":
- If t^* is an extremal point of q ➙ $|q(t^*)| = \|q\|_{L^\infty(I)}$,
- q has $n+1$ zeros in I $(*)$,
- $|q(-1)| = |q(1)| = \|q\|_{L^\infty(I)}$.

➢ q has $n+2$ extrema in $[-1,1]$

$(*)$ is motivated by an indirect argument:

If $q(t) = (t-t_0)\cdot\cdot\cdot\cdot\cdot(t-t_{n+1})$ with $t_0 < -1$, then

$$|p(t)| := |(t+1)(t-t_1)\cdot\cdot\cdot\cdot\cdot(t-t_{n+1})| < |q(t)| \quad \forall t \in I \quad \text{(why ?)},$$

which contradicts the minimality property of q. Same argument for $t_0 > 1$. The reasonings leading to the above heuristic demands will be elaborated in the proof of Thm. 6.2.3.8.

Are there polynomials satisfying these requirements? If so, do they allow a simple characterization?

Definition 6.2.3.3. Chebychev polynomials → [Han02, Ch. 32]

The n^{th} Chebychev polynomial is $T_n(t) := \cos(n \arccos t), \quad -1 \le t \le 1, n \in \mathbb{N}_0.$

The next result confirms that the T_n are polynomials, indeed.

Theorem 6.2.3.4. 3-term recursion for Chebychev polynomials → [Han02, (32.2)]

*The function T_n defined in Def. 6.2.3.3 satisfy the **3-term recursion***

$$T_{n+1}(t) = 2t\,T_n(t) - T_{n-1}(t) \quad , \quad T_0 \equiv 1\,, \quad T_1(t) = t\,, \quad n \in \mathbb{N}\,. \tag{6.2.3.5}$$

Proof. Just use the trigonometric identity $\cos(n+1)x = 2\cos nx \cos x - \cos(n-1)x$ with $\cos x = t$. □

The theorem implies:
- $T_n \in \mathcal{P}_n$,
- their leading coefficients are equal to 2^{n-1},
- the T_n are linearly independent,
- $\{T_j\}_{j=0}^n$ is a basis of $\mathcal{P}_n = \mathrm{Span}\{T_0, \ldots, T_n\}$, $n \in \mathbb{N}_0$.

See Code 6.2.3.6 for algorithmic use of the 3-term recursion (6.2.3.5).

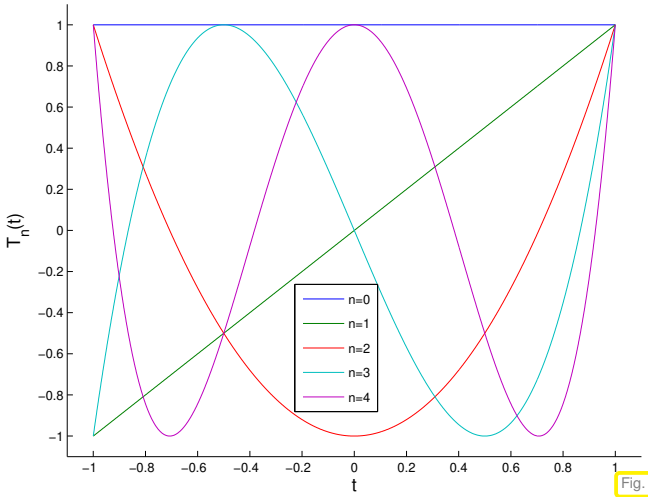


Fig. 219

Chebychev polynomials $T_0, \ldots, T_4$

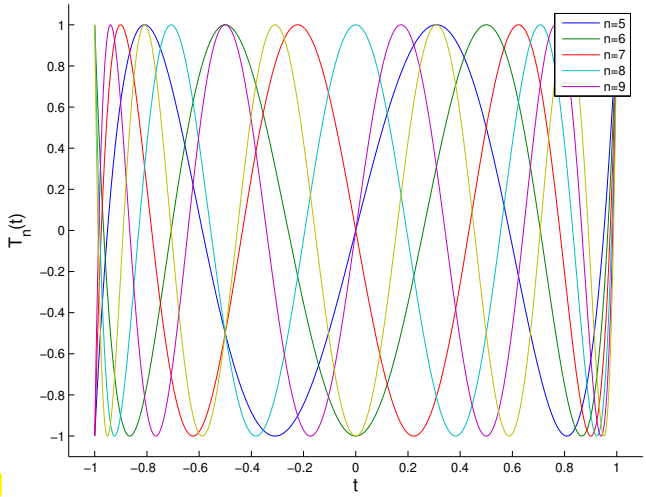


Fig. 220

Chebychev polynomials $T_5, \ldots, T_9$

C++ code 6.2.3.6: Efficient evaluation of Chebychev polynomials up to a certain degree ➽ GITLAB

```
// Computes the values of the Chebychev polynomials T_0,...,T_d
// at points passed in x using the 3-term recursion
// (6.2.3.5). The values T_k(x_j), are returned in
// row k+1 of V.
void chebpolmult(const unsigned int d, const RowVectorXd& x, MatrixXd& V) {
  const unsigned int n = x.size();
```

```
  V = MatrixXd::Ones(d + 1, n);  // T_0 ≡ 1
  if (d == 0) return;
  V.block(1, 0, 1, n) = x;  // T_1(x) = x
  if (d == 1) return;
  for (unsigned int k = 1; k < d; ++k) {
    const RowVectorXd p = V.block(k, 0, 1, n);       // p = T_k
    const RowVectorXd q = V.block(k - 1, 0, 1, n);   // q = T_{k-1}
    V.block(k + 1, 0, 1, n) =
        2 * x.cwiseProduct(p) - q;  // 3-term recursion
  }
}
```

From Def. 6.2.3.3 we conclude that T_n attains the values ± 1 in its extrema with alternating signs, thus matching our heuristic demands:

$$|T_n(\bar{t}_k)| = 1 \Leftrightarrow \exists\, k = 0,\ldots,n\colon \quad \bar{t}_k = \cos\frac{k\pi}{n}\,, \qquad \|T_n\|_{L^\infty([-1,1])} = 1\,. \tag{6.2.3.7}$$

What is still open is the validity of the heuristics guiding the choice of the optimal nodes. The next fundamental theorem will demonstrate that, after scaling, the T_n really supply polynomials on $[-1,1]$ with fixed leading coefficient and minimal supremum norm.

Theorem 6.2.3.8. Minimax property of the Chebychev polynomials [DH03, Section 7.1.4.], [Han02, Thm. 32.2]

The polynomials T_n from Def. 6.2.3.3 minimize the supremum norm in the following sense:

$$\|T_n\|_{L^\infty([-1,1])} = \inf\{\|p\|_{L^\infty([-1,1])}\colon p \in \mathcal{P}_n, p(t) = 2^{n-1}t^n + \cdots\}\,, \quad \forall n \in \mathbb{N}\,.$$

Proof. (indirect) Assume

$$\exists q \in \mathcal{P}_n\,, \text{ leading coefficient } = 2^{n-1}\colon \quad \|q\|_{L^\infty([-1,1])} < \|T_n\|_{L^\infty([-1,1])}\,. \tag{6.2.3.9}$$

$(T_n - q)(x) > 0$ in the local maxima of T_n, and $(T_n - q)(x) < 0$ in all local minima of T_n.

From our knowledge about the $n+1$ local extrema of T_n in $[-1,1]$ (They have alternating signs!), see (6.2.3.7), we conclude that $T_n - q$ changes sign at least $n+1$ times. This implies that $T_n - q$ has at least n zeros. As a consequence, $T_n - q \equiv 0$, because $T_n - q \in \mathcal{P}_{n-1}$ (same leading coefficient!).

This cannot be reconciled with the properties (6.2.3.9) of q and, thus, leads to a contradiction. □

The zeros of T_n are $\quad t_k = \cos\left(\frac{2k+1}{2n}\pi\right)\,, \quad k = 0,\ldots,n-1\,.$ (6.2.3.10)

Too see this, notice

$$T_n(t) = 0 \overset{\text{zeros of } \cos}{\Leftrightarrow} n\arccos t \in (2\mathbb{Z}+1)\frac{\pi}{2}$$
$$\overset{\arccos \in [0,\pi]}{\Leftrightarrow} t \in \left\{\cos\left(\frac{2k+1}{n}\frac{\pi}{2}\right),\ k = 0,\ldots,n-1\right\}\,.$$

Thus, we have identified the t_k from (6.2.3.10) as optimal interpolation nodes for a Lagrangian approximation scheme. The t_k are known as Chebychev nodes. Their distribution in $[-1,1]$ is plotted in Fig. 221, a geometric construction is indicated in Fig. 222.

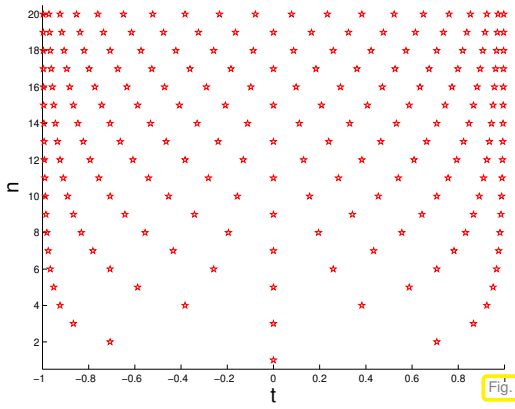


Fig. 221

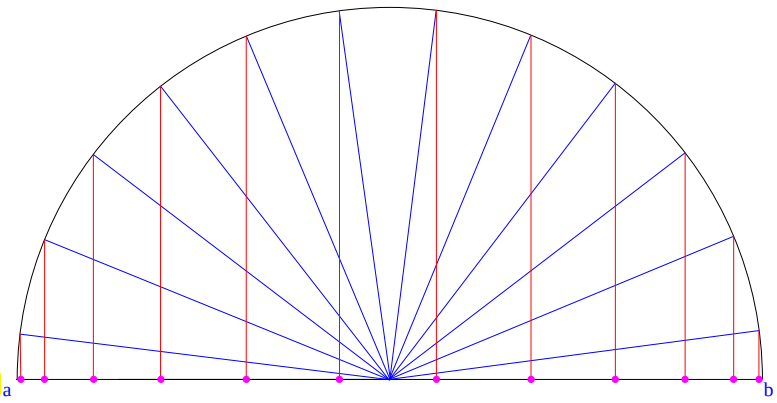
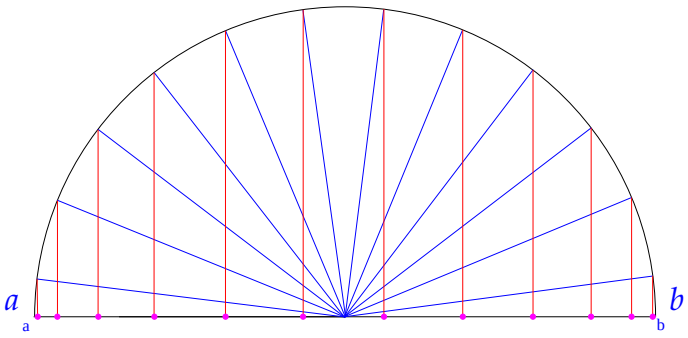


Fig. 222

Remark 6.2.3.11 (Chebychev nodes on arbitrary interval) Following the recipe of § 6.2.1.14 Chebychev interpolation on an arbitrary interval $[a,b]$ can immediately be defined. The same polynomial Lagrangian approximation scheme is obtained by transforming the Chebychev nodes (6.2.3.10) from $[-1,1]$ to $[a,b]$ using the unique affine transformation (6.2.1.15):

$$\widehat{t} \in [-1,1] \mapsto t := a + \tfrac{1}{2}(\widehat{t}+1)(b-a) \in [a,b]\ ].$$



The Chebychev nodes in the interval $I = [a,b]$ are

$$t_k := a + \tfrac{1}{2}(b-a)\left(\cos\left(\frac{2k+1}{2(n+1)}\,\pi\right)+1\right), \tag{6.2.3.12}$$
$$k = 0,\ldots,n\,.$$

Parlance: When we use Chebychev nodes for polynomial interpolation we call the resulting Lagrangian approximation scheme **Chebychev interpolation**.

Review question(s) 6.2.3.13 (Chebychev interpolation: motivation and definition)

(Q6.2.3.13.A) We write $T_n \in \mathcal{P}_n$ for the n-th Chebychev polynomial:

Definition 6.2.3.3. Chebychev polynomials

The n^{th} Chebychev polynomial is $T_n(t) := \cos(n \arccos t), \quad -1 \le t \le 1, n \in \mathbb{N}.$

What is the composition $T_n \circ T_m$, $T_n \circ T_m = T_n(T_m(t))$ of two Chebychev polynomials?

(Q6.2.3.13.B) Writing $T_n \in \mathcal{P}_n$ for the n-th Chebychev polynomial, using a trigonometric identity show that

$$2T_m(t)T_n(t) = T_{m+n}(t) + T_{m-n}(t) \quad \forall m,n \in \mathbb{N}_0\,, \quad M \ge n\,.$$

△

6.2.3.2 Chebychev Interpolation Error Estimates

Video tutorial for Section 6.2.3.2 "Chebychev Interpolation Error Estimates": (14 minutes) Download link, tablet notes

EXAMPLE 6.2.3.14 (Polynomial interpolation: Chebychev nodes versus equidistant nodes) We consider Runge's function $f(t) = \frac{1}{1+t^2}$, see Ex. 6.2.2.11, and compare polynomial interpolation based on uniformly spaced nodes and Chebychev nodes in terms of behavior of interpolants.

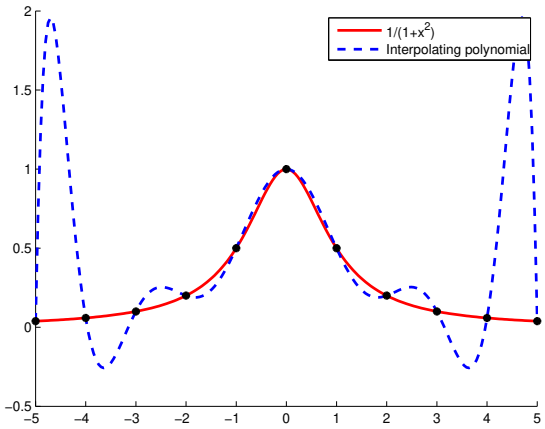


Fig. 223

Equidistant nodes

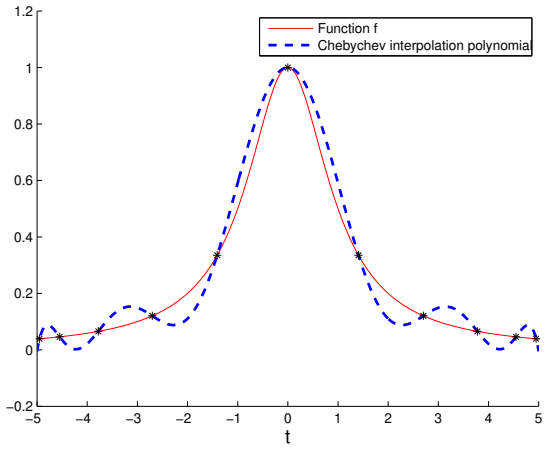


Fig. 224

Chebychev nodes

We observe that the Chebychev nodes cluster at the endpoints of the interval, which successfully suppresses the huge oscillations haunting equidistant interpolation there. ⌟

§6.2.3.15 (Finite-smoothness error estimates for Chebychev interpolation) Note the following features of Chebychev interpolation on the interval $[-1,1]$:

- Use of "optimal" interpolation nodes $\mathcal{T} = \left\{\widehat{t}_k := \cos\left(\frac{2k+1}{2(n+1)}\,\pi\right),\ k = 0,\ldots,n\right\}$,
- corresponding to the nodal polynomial $w(t) = (t-t_0)\cdots(t-t_n) = 2^{-n}T_{n+1}(t)$, $\|w\|_{L^\infty(I)} = 2^{-n}$, with leading coefficient 1.

Then, by

Theorem 6.2.2.15. Representation of interpolation error

We consider $f \in C^{n+1}(I)$ and the Lagrangian interpolation approximation scheme (→ Def. 6.2.2.1) for a node set $\mathcal{T} := \{t_0,\ldots,t_n\} \subset I$. Then,

for every $t \in I$ there exists a $\tau_t \in]\min\{t,t_0,\ldots,t_n\}, \max\{t,t_0,\ldots,t_n\}[$ such that

$$f(t) - \mathrm{I}_{\mathcal{T}}(f)(t) = \frac{f^{(n+1)}(\tau_t)}{(n+1)!}\cdot\prod_{j=0}^{n}(t-t_j)\,. \tag{6.2.3.16}$$

we immediately get an interpolation error estimate for Chebychev interpolation of $f \in C^{n+1}([-1,1])$:

$$\|f - \mathrm{I}_{\mathcal{T}}(f)\|_{L^\infty([-1,1])} \le \frac{2^{-n}}{(n+1)!}\left\|f^{(n+1)}\right\|_{L^\infty([-1,1])}\,. \tag{6.2.3.17}$$

Estimates for the Chebychev interpolation error on $[a,b]$ are easily derived from (6.2.3.17):

$$p \in \mathcal{P}_n \quad\wedge\quad p(t_j) = f(t_j) \quad\Leftrightarrow\quad \widehat{p} \in \mathcal{P}_n \quad\wedge\quad \widehat{p}(\widehat{t}_j) = \widehat{f}(\widehat{t}_j)\,,$$

an the affine pullback introduced and discussed in in § 6.2.1.14. For instance, repeated application of the chain rule yields the formula $\frac{\mathrm{d}^n\widehat{f}}{\mathrm{d}\widehat{t}^n}(\widehat{t}) = (\frac{1}{2}|I|)^n\,\frac{\mathrm{d}^n f}{\mathrm{d}t^n}(t)$.

$$\|f - \mathrm{I}_{\mathcal{T}}(f)\|_{L^\infty(I)} = \left\|\widehat{f} - \mathrm{I}_{\widehat{\mathcal{T}}}(\widehat{f})\right\|_{L^\infty([-1,1])} \leq \frac{2^{-n}}{(n+1)!}\left\|\frac{\mathrm{d}^{n+1}\widehat{f}}{\mathrm{d}\widehat{t}^{n+1}}\right\|_{L^\infty([-1,1])}$$
$$\leq \frac{2^{-2n-1}}{(n+1)!}|I|^{n+1}\left\|f^{(n+1)}\right\|_{L^\infty(I)} . \tag{6.2.3.18}$$

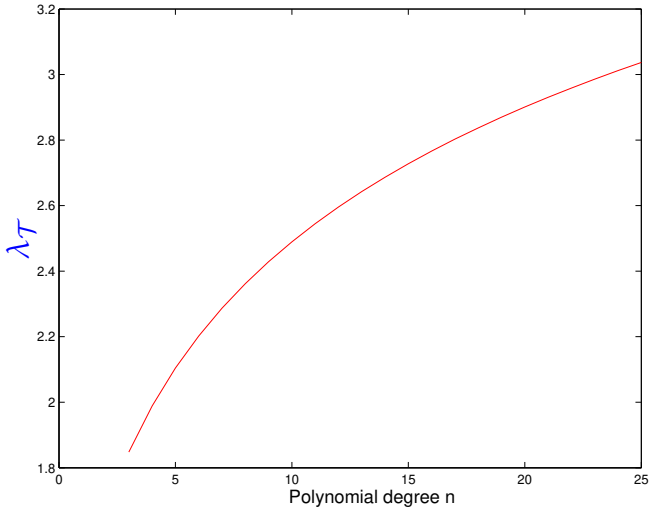
⌟

Remark 6.2.3.19 (Lebesgue Constant for Chebychev nodes [Tre13, Thm 15.2]) We saw in Section 5.2.4 and, in particular, in Rem. 5.2.4.13 that the Lebesgue constant $\lambda_{\mathcal{T}}$ that measures the sensitivity of a polynomial interpolation scheme, blows up exponentially with increasing number of *equispaced* interpolation nodes. In stark contrast $\lambda_{\mathcal{T}}$ grows only logarithmically in the number of Chebychev nodes.

More precisely, sophisticated theory [CB95; Ver86; Ver90] supplies the bound

$$\boxed{\lambda_{\mathcal{T}} \leq \tfrac{2}{\pi}\log(1+n)+1} . \tag{6.2.3.20}$$

Measured Lebesgue constant for Chebychev nodes based on approximate evaluation of (5.2.4.11) by sampling. ▷



Combining (6.2.3.20) with the general estimate from Rem. 6.2.2.29

$$\|f - \mathrm{L}_{\mathcal{T}} f\|_{L^\infty(I)} \leq (1+\lambda_{\mathcal{T}}) \inf_{p\in\mathcal{P}_n} \|f-p\|_{L^\infty(I)} \quad \forall f \in C^0(I) , \tag{6.2.2.33}$$

and the bound for the best approximation error by polynomials for $f \in C^r([-1,1])$ from Thm. 6.2.1.11,

$$\inf_{p\in\mathcal{P}_n} \|f-p\|_{L^\infty([-1,1])} \leq (1+\pi^2/2)^r \frac{(n-r)!}{n!}\left\|f^{(r)}\right\|_{L^\infty([-1,1])} ,$$

we end up with a bound for the supremum norm of the interpolation error in the case of Chebychev interpolation on $[-1,1]$

$$\|f - \mathrm{L}_{\mathcal{T}} f\|_{L^\infty([-1,1])} \leq (2/\pi\log(1+n)+2)(1+\pi^2/2)^r \frac{(n-r)!}{n!}\left\|f^{(r)}\right\|_{L^\infty([-1,1])} . \tag{6.2.3.21}$$

Emphasizing the asymptotic behavior of the maximum norm of the interpolation error, we can infer for Chebychev interpolation

$$f \in C^r([-1,1]) \quad \Rightarrow \quad \|f - \mathrm{L}_{\mathcal{T}} f\|_{L^\infty([-1,1])} = O\left(\frac{\log n}{n^r}\right) \quad \text{for} \quad n \to \infty , \tag{6.2.3.22}$$

which could be dubbed "almost algebraic convergence" with rate r. This guarantees convergence, if only $f \in C^1([-1,1])$! ⌟

EXPERIMENT 6.2.3.23 (Chebychev interpolation errors) Now we empirically investigate the behavior of norms of the interpolation error for Chebychev interpolation and functions with different (smoothness) properties as we increase the number of interpolation nodes.

In the experiments, for $I = [a,b]$ we set $x_l := a + \frac{b-a}{N} l,\ l = 0,...,N,\ N = 1000$, and we approximate the norms of the interpolation error as follows ($p \hat{=}$ interpolating polynomial):

$$\|f - p\|_\infty \approx \max_{0 \le l \le N} |f(x_l) - p(x_l)| \tag{6.2.3.24}$$

$$\|f - p\|_2^2 \approx \frac{b-a}{2N} \sum_{0 \le l < N} \left(|f(x_l) - p(x_l)|^2 + |f(x_{l+1}) - p(x_{l+1})|^2 \right) \tag{6.2.3.25}$$

① $f(t) = (1+t^2)^{-1}$, $I = [-5,5]$ (see Ex. 6.2.2.11): analytic in a neighborhood of I.
Interpolation with $n = 10$ Chebychev nodes (plot on the left).

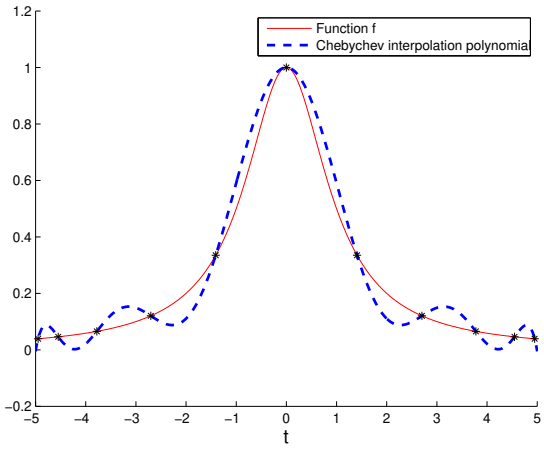
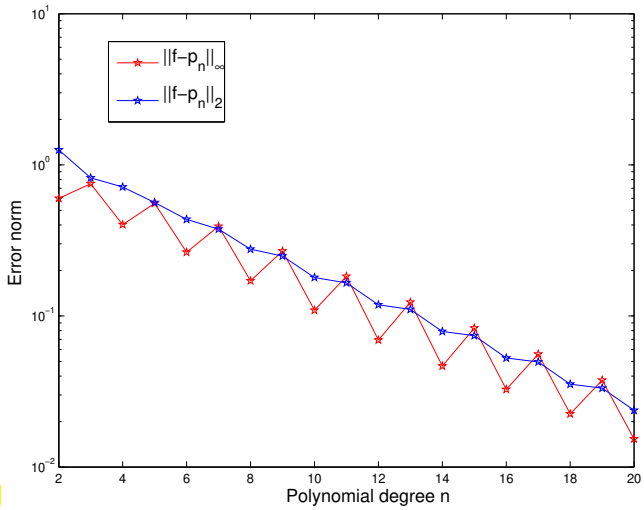


Fig. 225



Notice: exponential convergence ($\to$ Def. 6.2.2.7) of the Chebychev interpolation:

$$p_n \to f, \quad \|f - \mathsf{I}_n f\|_{L^\infty([-5,5])} \approx 0.8^n$$

.

② $f(t) = \max\{1 - |t|, 0\}$, $I = [-2,2]$, $n = 10$ nodes (plot on the left).

Now $f \in C^0(I)$ but $f \notin C^1(I)$.

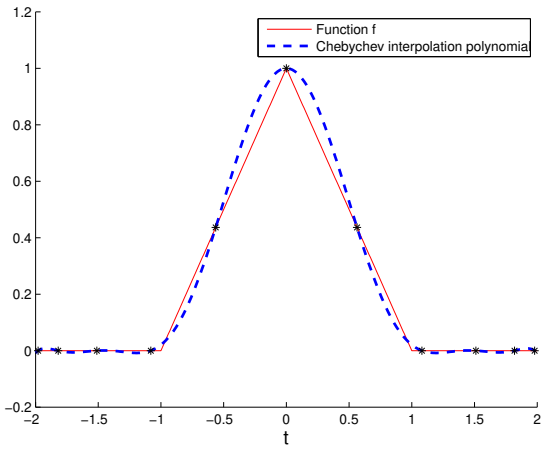


Fig. 226

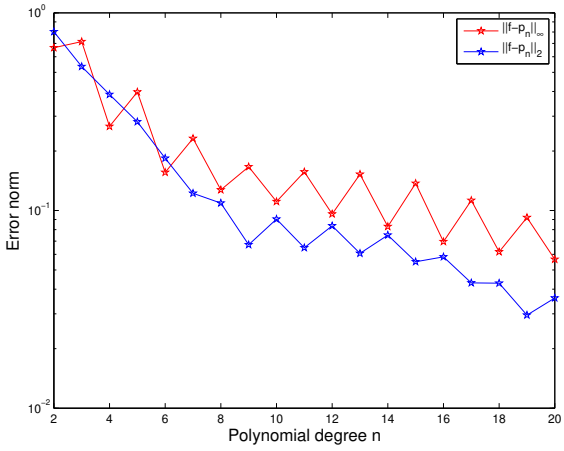


Fig. 227

From the doubly logarithmic plot we conclude ➙

- no exponential convergence
- algebraic convergence (**?**)

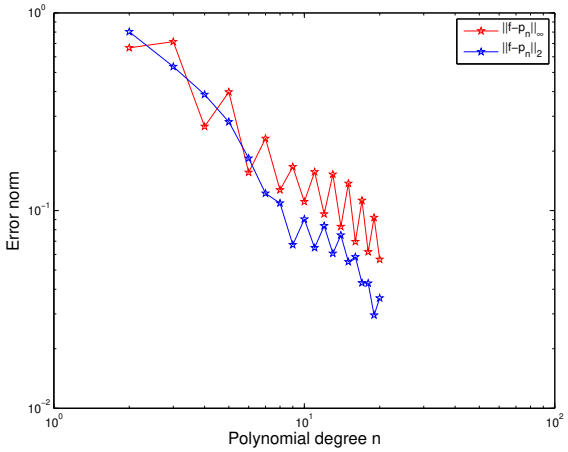


Fig. 228

③ $f(t) = \begin{cases} \frac{1}{2}(1+\cos \pi t) & |t| < 1 \\ 0 & 1 \le |t| \le 2 \end{cases}$ $I = [-2,2]$, $n = 10$ (plot on the left).

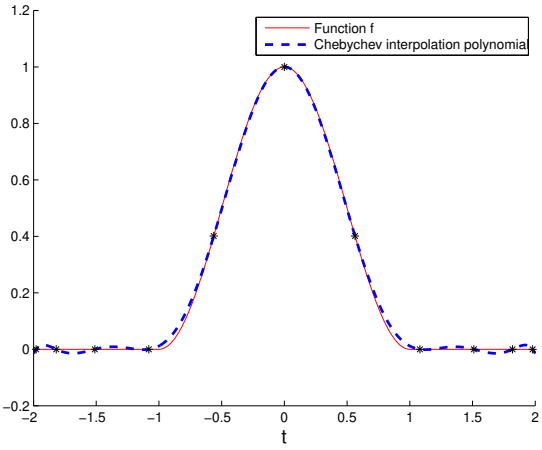


Fig. 229

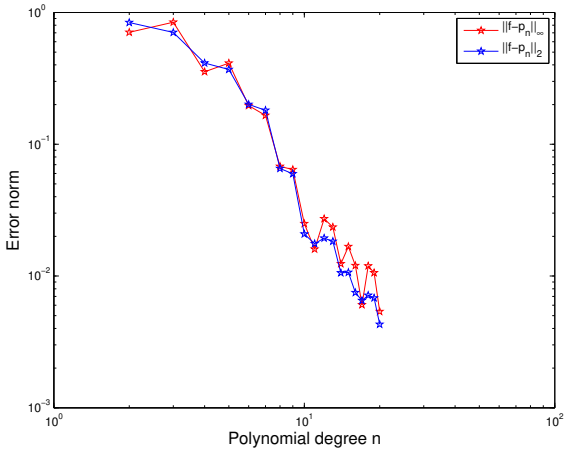


Fig. 230

Notice: only (vaguely) algebraic convergence.

Summary of observations, *cf.* § 6.2.2.9:

- ✦ Essential role of smoothness of f: slow convergence of approximation error of the Cheychev interpolant if f enjoys little smoothness, *cf.* also (6.2.2.22),
- ✦ for analytic $f \in C^\infty$ ($\rightarrow$ Def. 6.2.2.48) the approximation error of the Cheychev interpolant seems to decay to zero exponentially in the polynomial degree n.

⌟

Remark 6.2.3.26 (Chebychev interpolation of analytic functions [Tre13, Ch. 8]) Assuming that the interpoland f possesses an analytic extension to a complex neighborhood D of $[-1,1]$, we now apply the theory of Section 6.2.2.3 to bound the supremum norm of the Chebychev interpolation error of f on $[-1,1]$.

To convert

$$|f(t) - \mathsf{L}_\mathcal{T} f(t)| \le \left| \frac{w(t)}{2\pi\imath} \int_\gamma \frac{f(z)}{(z-t)w(z)} \,\mathrm{d}z \right| \le \frac{|\gamma|}{2\pi} \frac{\max\limits_{a\le\tau\le b} |w(\tau)|}{\min\limits_{z\in\gamma} |w(z)|} \cdot \frac{\max\limits_{z\in\gamma} |f(z)|}{\operatorname{dist}([a,b],\gamma)} \,, \tag{6.2.2.58}$$

as obtained in Section 6.2.2.3, into a more concrete estimate, we have to study the behavior of

$$w_n(t) = (t-t_0)(t-t_1)\cdot\cdots\cdot(t-t_n)\,, \quad t_k = \cos\left(\frac{2k+1}{2n+2}\,\pi\right), k = 0,\ldots,n\,,$$

where the t_k are the Chebychev nodes according to (6.2.3.12). They are the zeros of the Chebychev polynomial ($\to$ Def. 6.2.3.3) of degree $n+1$. Since w has leading coefficient 1, we conclude $w = 2^{-n}T_{n+1}$, and

$$\max_{-1\le t\le 1} |w(t)| \le 2^{-n}\,. \tag{6.2.3.27}$$

Next, we fix a suitable path $\gamma \subset D$ for integration:For a constant $\rho > 1$ we set

$$\begin{aligned}\gamma &:= \{z = \cos(\theta - \imath\log\rho),\, 0\le\theta\le 2\pi\}\\ &= \left\{z = \frac{1}{2}(\exp(\imath(\theta-\imath\log\rho)) + \exp(-\imath(\theta-\imath\log\rho))),\, 0\le\theta\le 2\pi\right\}\\ &= \left\{z = \frac{1}{2}\left(\rho e^{\imath\theta} + \rho^{-1}e^{-iu\theta}\right),\, 0\le\theta\le 2\pi\right\}\\ &= \left\{z = \tfrac{1}{2}(\rho+\rho^{-1})\cos\theta + \imath\tfrac{1}{2}(\rho-\rho^{-1})\sin\theta,\, 0\le\theta\le 2\pi\right\}\end{aligned}$$

Thus, we see that γ is an ellipse with foci ± 1, large semi-axis $\frac{1}{2}(\rho+\rho^{-1}) > 1$ and small semi-axis $\frac{1}{2}(\rho-\rho^{-1}) > 0$.

The figure shows elliptical integration contours for different values of ρ ▷

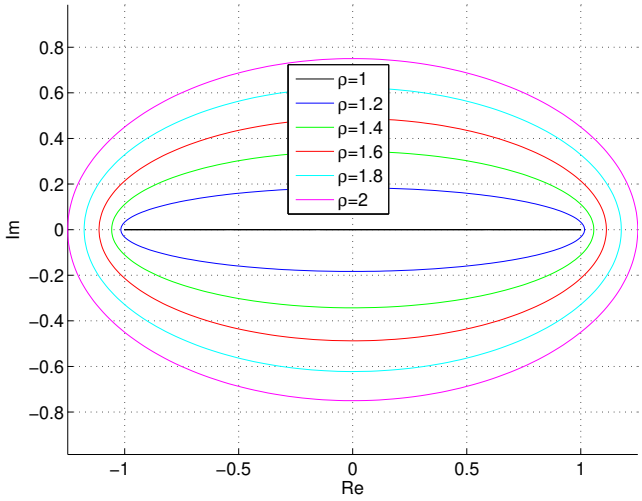


Fig. 231

Appealing to geometric evidence, we find $\operatorname{dist}(\gamma,[-1,1]) = \frac{1}{2}(\rho+\rho^{-1}) - 1$, which gives another term in (6.2.2.58).

The rationale for choosing this particular integration contour is that the cos in its defintion nicely cancels the arccos in the formula for the Chebychev polynomials. This lets us compute

$$\begin{aligned}|2^n w(\gamma(\theta))|^2 &= |T_{n+1}(\cos(\theta-\imath\log\rho))|^2\\ &= |\cos((n+1)(\theta-\imath\log\rho))|^2\\ &= \cos((n+1)(\theta-\imath\log\rho))\cdot\overline{\cos((n+1)(\theta-\imath\log\rho))}\\ &= \frac{1}{4}(\rho^s e^{\imath(n+1)\theta} + \rho^{-(n+1)}e^{-\imath(n+1)\theta})(\rho^{(n+1)}e^{-\imath(n+1)\theta} + \rho^{-(n+1)}e^{\imath(n+1)\theta})\\ &= \frac{1}{4}(\rho^{2(n+1)} + \rho^{-2(n+1)} + e^{2\imath(n+1)\theta} + e^{-2\imath(n+1)\theta})\\ &= \frac{1}{4}(\rho^{n+1} - \underbrace{\rho^{-(n+1)}}_{<1})^2 + \underbrace{\frac{1}{4}(e^{\imath(n+1)\theta} + e^{-\imath(n+1)\theta})^2}_{=\cos^2((n+1)\theta)\ge 0} \ge \frac{1}{4}(\rho^{n+1}-1)^2\,,\end{aligned}$$

for all $0 \leq \theta \leq 2\pi$, which provides a lower bound for $|w_n|$ on γ. Plugging all these estimates into (6.2.2.58) we arrive at

$$\|f - \mathsf{L}_\mathcal{T} f\|_{L^\infty([-1,1])} \leq \frac{2|\gamma|}{\pi} \frac{1}{(\rho^{n+1}-1)(\rho+\rho^{-1}-2)} \cdot \max_{z\in\gamma} |f(z)| \quad . \tag{6.2.3.28}$$

Note that instead on the nodal polynomial w we have inserted T_{n+1} into (6.2.2.58), which is a simple multiple. The factor will cancel.

The supremum norm of the interpolation error converges exponentially ($\rho > 1$!):

$$\|f - \mathsf{L}_\mathcal{T} f\|_{L^\infty([-1,1])} = O(\rho^{-n}) \quad \text{for} \quad n \to \infty \, .$$

EXPERIMENT 6.2.3.29 (Chebychev interpolation of analytic function → Exp. 6.2.3.23 cnt'd)

Modification: the same function $f(t) = (1+t^2)^{-1}$ on a smaller interval $I = [-1,1]$.

(Faster) exponential convergence than on the interval $I =]-5,5[$:

$$\|f - \mathsf{I}_n f\|_{L^2([-1,1])} \approx 0.42^n.$$

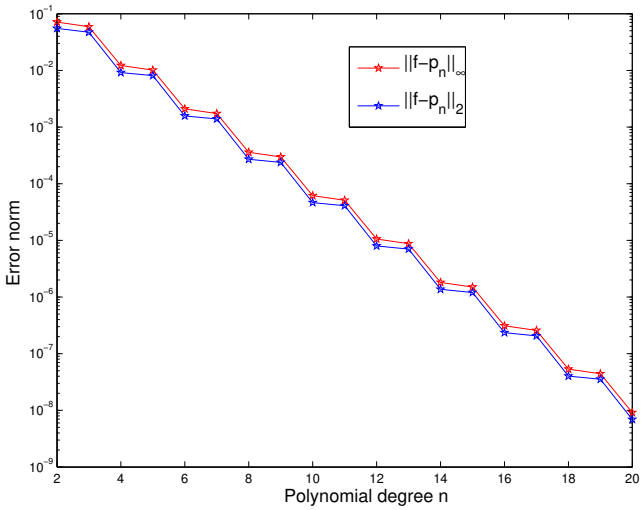


Fig. 232

Explanation, *cf.* Rem. 6.2.3.26: for $I = [-1,1]$ the poles $\pm i$ of f are farther away relative to the size of the interval than for $I = [-5,5]$.

Review question(s) 6.2.3.30 (Chebychev interpolation error estimates)

(Q6.2.3.30.A)

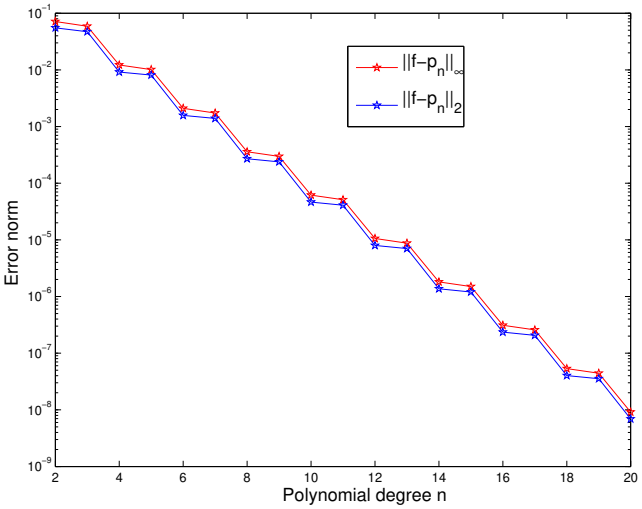


Fig. 233

Plotting norms of the Chebychev interpolation
for the "Runge function"

$$f(t) := \frac{1}{1+t^2}$$

on $[-1,1]$ we observe a strange staircase pa

Guess what could be the cause of its emerg

△

6.2.3.3 Chebychev Interpolation: Computational Aspects

Video tutorial for Section 6.2.3.3 "Chebychev Interpolation: Computational Aspects": (11 minutes) Download link, tablet notes

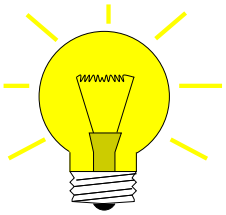
Task: Given: polynomial degree $n \in \mathbb{N}$ and a continuous function $f : [-1,1] \mapsto \mathbb{R}$

Sought: efficient representation/evaluation of Chebychev interpolant $p \in \mathcal{P}_n$ (= polynomial Lagrange interpolant of degree $\leq n$ in Chebychev nodes (6.2.3.12) on $[-1,1]$).

More concretely, this boils down to a implementation of the following class:

C++ code 6.2.3.31: Definition of class for Chebychev interpolation

```
class ChebInterp {
  private:
  // various internal data describing Chebychev interpolating polynomial p
  public:
  // Constructor taking function f and degree n as arguments
  template <typename Function>
  PolyInterp(const Function &f,unsigned int n);
  // Evaluation operator:  y_j = p(x_j), j = 1,...,m (m "large")
  template <typename Vector>
  void eval(const Vector &x, Vector &y) const;
};
```



Idea: internally represent p as a linear combination of Chebychev polynomials, a **Chebychev expansion**:

$$p(t) = \sum_{j=0}^{n} \alpha_j T_j(t) , \quad t \in \mathbb{R} , \quad \alpha_j \in \mathbb{R} ,$$

where T_j is the Chebychev polynomial of degree j, see Def. 6.2.3.3.

The representation (6.2.3.3) is always possible, because $\{T_0, \ldots, T_n\}$ is a *basis* of $\mathcal{P}_n$, owing to $\deg T_n = n$. The representation is amenable to efficient evaluation and computation by means of special algorithms.

§6.2.3.32 (Fast evaluation of Chebychev expansion → [Han02, Alg. 32.1]) Let us assume that the Chebychev expansion coefficients α_j in (6.2.3.3) are given and wonder, how we can efficiently compute $p(x)$ for some $x \in \mathbb{R}$:

Task: Given $n \in \mathbb{N}$, $x \in \mathbb{R}$, and the Chebychev expansion coefficients $\alpha_j \in \mathbb{R}$, $j = 0, \ldots, n$, compute $p(x)$ with

$$p(x) = \sum_{j=0}^{n} \alpha_j T_j(x) , \quad \alpha_j \in \mathbb{R} . \tag{6.2.3.3}$$

Idea: Use the 3-term recurrence (6.2.3.5)

$$T_j(x) = 2xT_{j-1}(x) - T_{j-2}(x) , \quad j = 2,3,\ldots , \tag{6.2.3.5}$$

to design a recursive evaluation scheme.

By means of (6.2.3.5) rewrite (6.2.3.3) as

$$p(x) = \sum_{j=0}^{n-1} \alpha_j T_j(x) + \alpha_n T_n(x) \overset{(6.2.3.5)}{=} \sum_{j=0}^{n-1} \alpha_j T_j(x) + \alpha_n (2xT_{n-1}(x) - T_{n-2}(x))$$
$$= \sum_{j=0}^{n-3} \alpha_j T_j(x) + (\alpha_{n-2} - \alpha_n) T_{n-2}(x) + (\alpha_{n-1} + 2x\alpha_n) T_{n-1}(x) \,.$$

We recover the point value $p(x)$ as the point value of another polynomial of degree $n-1$ with known Chebychev expansion:

$$p(x) = \sum_{j=0}^{n-1} \widetilde{\alpha}_j T_j(x) \quad \text{with} \quad \widetilde{\alpha}_j = \begin{cases} \alpha_j + 2x\alpha_{j+1} & \text{, if } j = n-1 \,, \\ \alpha_j - \alpha_{j+2} & \text{, if } j = n-2 \,, \\ \alpha_j & \text{else.} \end{cases} \tag{6.2.3.33}$$

This inspires the recursive algorithm of Code 6.2.3.34. A loop-based implementation without recursive function calls is also possible and given in Code 6.2.3.35.

C++ code 6.2.3.34: Recursive evaluation of Chebychev expansion (6.2.3.3) ↬ **GITLAB**

```
// Recursive evaluation of a polynomial p = sum_{j=1}^{n+1} a_j T_{j-1} at point x
// based on (6.2.3.33)
// IN : Vector of coefficients a
// evaluation point x
// OUT: Value at point x
double recclenshaw(const VectorXd& a, const double x) {
  const VectorXd::Index n = a.size() - 1;
  if (n == 0) { return a(0); }            // Constant polynomial
  if (n == 1) { return (x*a(1) + a(0)); } // Value a_1 * x + a_0
  VectorXd new_a(n);
  new_a << a.head(n - 2), a(n - 2) - a(n), a(n - 1)+ 2*x*a(n);
  return recclenshaw(new_a,x); // recursion
}
```

Non-recursive version: Clenshaw algorithm

C++ code 6.2.3.35: Clenshaw algorithm for evalation of Chebychev expansion (6.2.3.3) ↬ **GITLAB**

```
// Clenshaw algorithm for evaluating p = sum_{j=1}^{n+1} a_j T_{j-1}
// at points passed in vector x
// IN : a = [alpha_j], coefficients for p = sum_{j=1}^{n+1} alpha_j T_{j-1}
// x = (many) evaluation points
// OUT: values p(x_j) for all j
VectorXd clenshaw(const VectorXd& a, const VectorXd& x) {
  const int n = a.size() - 1;  // degree of polynomial
  MatrixXd d(n + 1, x.size()); // temporary storage for intermediate values
  for (int c = 0; c < x.size(); ++c) d.col(c) = a;
  for (int j = n - 1; j > 0; --j) {
    d.row(j) += 2*x.transpose().cwiseProduct(d.row(j+1)); // see (6.2.3.33)
    d.row(j-1) -= d.row(j + 1);
  }
  return d.row(0) + x.transpose().cwiseProduct(d.row(1));
}
```

Computational effort : $O(nm)$ for evaluation at m points, $m, n \to \infty$. ⌟

§6.2.3.36 (Computation of Chebychev expansions of interpolants) Chebychev interpolation is a linear interpolation scheme, see § 5.1.0.21. Thus, the expansion α_j in (6.2.3.3) can be computed by solving a linear system of equations of the form (5.1.0.23). However, for Chebychev interpolation this linear system can be cast into a very special form, which paves the way for its fast direct solution:

Task: Efficiently compute the Chebychev expansion coefficients α_j in (6.2.3.3) from the interpolation conditions

$$p(t_k) = f(t_k)\,, \quad k = 0,\ldots,n\,, \quad \text{for} \quad t_k := \cos\left(\frac{2k+1}{2(n+1)}\pi\right)\,. \tag{6.2.3.37}$$

Chebychev nodes

Trick: Ttransform of p into a 1-periodic function q, which turns out to be a trigonometric polynomial according to Def. 4.2.6.25. Using the definition of the Chebychev polynomials and Euler's formula we get

$$\begin{aligned} q(s) &:= p(\cos 2\pi s) = \sum_{j=0}^{n} \alpha_j T_j(\cos 2\pi s) \overset{\text{Def. 6.2.3.3}}{=} \sum_{j=0}^{n} \alpha_j \cos(2\pi j s) \\ &= \sum_{j=0}^{n} \tfrac{1}{2}\alpha_j(\exp(2\pi\imath j s) + \exp(-2\pi\imath j s)) \quad [\text{ by } \cos z = \tfrac{1}{2}(e^z + e^{-z})\,] \\ &= \sum_{j=-n}^{n+1} \beta_j \exp(-2\pi\imath j s)\,, \quad \text{with} \quad \beta_j := \begin{cases} 0 & \text{, for } j = n+1\,, \\ \frac{1}{2}\alpha_j & \text{, for } j = 1,\ldots,n\,, \\ \alpha_0 & \text{, for } j = 0\,, \\ \frac{1}{2}\alpha_{n-j} & \text{, for } j = -n,\ldots,-1\,. \end{cases} \end{aligned} \tag{6.2.3.38}$$

The interpolation conditions (6.2.3.37) for p become interpolation conditions for q in transformed nodes, which turn out to be *equidistant*:

$$t = \cos(2\pi s) \overset{(6.2.3.37)}{\Longrightarrow} q\left(\frac{2k+1}{4(n+1)}\right) = y_k := f(t_k)\,, \quad k = 0,\ldots,n\,. \tag{6.2.3.39}$$

This amounts to a Lagrange polynomial interpolation problem for *equidistant points on the unit circle* as we have seen them in Section 5.6.3.

Also observe the following even symmetry with respect to $s = \frac{1}{2}$:

$$\begin{aligned} & q(s) = q(1-s) \\ & \Downarrow\leftarrow \text{(6.2.3.39)} \\ & q(1 - \frac{2k+1}{4(n+1)}) = y_k\,, \quad k = 0,\ldots,n\,. \end{aligned}$$

It ensures that the coefficients β_j actually satisfy the constraints implied by their relationship with α_j.

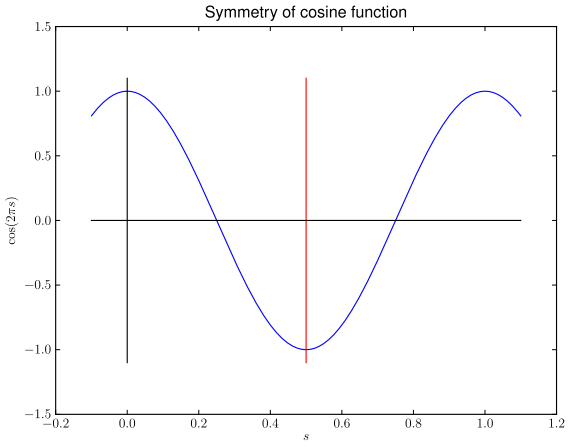


Fig. 234

Thanks to the symmetry of q, see Fig. 235, we can augment the interpolation conditions (6.2.3.39) and demand

$$q\left(\frac{k}{2(n+1)}+\frac{1}{4(n+1)}\right) = z_k := \begin{cases} y_k & \text{, for } k=0,\ldots,n\,, \\ y_{2n+1-k} & \text{, for } k=n+1,\ldots,2n+1\,. \end{cases} \tag{6.2.3.40}$$

In a sense, we have just reflected the interpolation conditions at $s=\frac{1}{2}$:

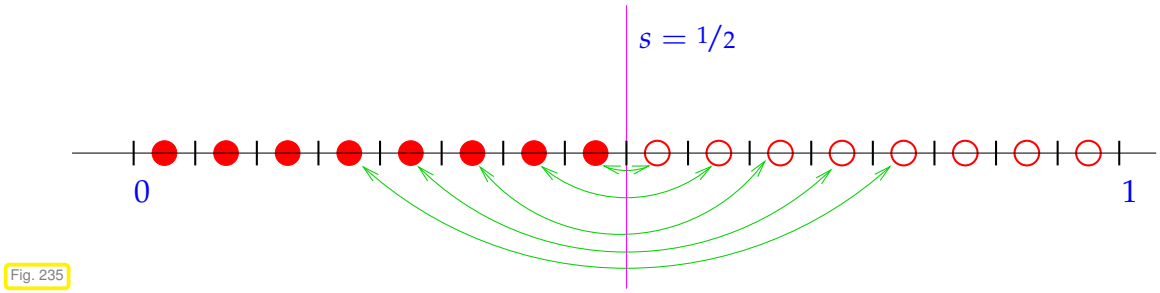


Fig. 235

Let us summarize our insights gained from switching from p to q:

> Chebychev expansion for Chebychev interpolants
> ⇕
> Trigonometric interpolation at equidistant points → Section 5.6.3

Trigonometric interpolation at equidistant points can be done very efficiently by means of FFT-based algorithms, see Code 5.6.3.4. We can also apply these for the computation of Chebychev expansion coefficients.

From (6.2.3.40) we can derive an $(2n+1)\times(2n+1)$ square linear system of equations for the unknown coefficients β_j.

$$q\left(\frac{k}{2(n+1)}+\frac{1}{4(n+1)}\right) = \sum_{j=-n}^{n+1}\left(\beta_j \exp\left(-\frac{2\pi\imath j}{4(n+1)}\right)\right)\exp\left(-\frac{2\pi\imath}{2(n+1)}kj\right) = z_k\,.$$

$$\Updownarrow$$

$$\sum_{j=0}^{2n+1}\left(\beta_{j-n}\exp\left(-\frac{2\pi\imath(j-n)}{4(n+1)}\right)\right)\underbrace{\exp\left(-\frac{2\pi\imath}{2(n+1)}kj\right)}_{=\omega_{2(n+1)}^{kj}\,!} = \exp\left(-\pi\imath\frac{nk}{n+1}\right)z_k\,,\quad k=0,\ldots,2n+1\,.$$

$$\Updownarrow$$

$$\mathbf{F}_{2(n+1)}\mathbf{c}=\mathbf{b} \quad \text{with} \quad \begin{aligned} \mathbf{c} &= \left[\beta_{j-n}\exp\left(-\tfrac{2\pi\imath(j-n)}{4(n+1)}\right)\right]_{j=0}^{2n+1}\,, \\ \mathbf{b} &= \left[\exp\left(-\pi\imath\tfrac{nk}{n+1}\right)z_k\right]_{k=0}^{2n+1}\,. \end{aligned} \tag{6.2.3.41}$$

$(2n+2)\times(2n+2)$ Fourier matrix, see (4.2.1.13)

Thus we can solve (6.2.3.41) by means of the inverse discrete Fourier transform of length $2(n+1)$, see 4.2. Using the FTT algorithm we can do this with an asymptotic complexity of $O(n\log n)$ for $n\to\infty$ as explained in Section 4.3.

Note that by the by the symmetry of that data vector $\mathbf{z}$ implied by (6.2.3.40) we find $\beta_{2n+1}=0$, which complies with (6.2.3.38). The following code demonstrates the FFT-based computation of the coefficients of the Chebychev expansion of the Chebychev interpolant on $[-1,1]$.

C++ code 6.2.3.42: Efficient computation of Chebychev expansion coefficient of Chebychev interpolant ↬ GITLAB

```
2   // efficiently compute coefficients α_j in the Chebychev
3   // expansion p = Σ_{j=0}^{n} α_j T_j of
4   // p ∈ P_n based on values y_k, k = 0,...,n,
5   // in Chebychev nodes t_k, k = 0,...,n IN: values
6   // y_k passed in y OUT: coefficients α_j
7   VectorXd
8   chebexp(const VectorXd &y) {
9   const Index n = y.size() - 1;          // degree of polynomial
10  const std::complex<double> M_I(0, 1); // imaginary unit
11  // create vector z, see (6.2.3.40)
12  VectorXcd b(2 * (n + 1));
13  const std::complex<double> om =
14      -M_I * (M_PI * static_cast<double>(n)) / static_cast<double>(n + 1);
15  for (int j = 0; j <= n; ++j) {
16    b(j) = std::exp(om * static_cast<double>(j)) *
17           y(j); // this cast to double is necessary!!
18    b(2 * n + 1 - j) = std::exp(om * static_cast<double>(2 * n + 1 - j)) * y(j);
19  }
20
21  // Solve linear system (6.2.3.41) with effort O(n log n)
22  Eigen::FFT fft;          // EIGEN's helper class for DFT
23  VectorXcd c = fft.inv(b); // -> c = ifft(z), inverse fourier transform
24  // recover β_j, see (6.2.3.41)
25  VectorXd beta(c.size());
26  const std::complex<double> sc = M_PI_2 / static_cast<double>(n + 1) * M_I;
27  for (unsigned j = 0; j < c.size(); ++j) {
28    beta(j) = (std::exp(sc * static_cast<double>(-n + j)) * c[j]).real();
29  }
30  // recover α_j, see (6.2.3.38)
31  VectorXd alpha = 2 * beta.segment(n, n);
32  alpha(0) = beta(n);
33  return alpha;
34 }
```

⌟

Remark 6.2.3.43 (Chebychev representation of built-in functions) Computers use approximation by sums of Chebychev polynomials in the computation of functions like $\log, \exp, \sin, \cos, \ldots$. The evaluation by means of Clenshaw algorithm according to Code 6.2.3.35 is more efficient and stable than the approximation by Taylor polynomials. ⌟

Review question(s) 6.2.3.44 (Chebychev Interpolation: Computational Aspects)

(Q6.2.3.44.A) Outline an efficient algorithm for computing the Chebychev expansion

$$p(t) = \sum_{k=0}^{n} a_k T_k(t) , \quad t \in \mathbb{R} , \quad a_k \in \mathbb{R} ,$$

of the Lagrange polynomial interpolant of $f \in C^0([-1,1])$ for an *arbitrary node set* $\mathcal{T} = \{t_0, \ldots, t_n\}$, $n \in \mathbb{N}$.

(Q6.2.3.44.B) Devise an efficient algorithm for evaluating a polynomial of degree $n \in \mathbb{N}$ given through its

Chebychev expansion

$$p(t) = \sum_{k=0}^{n} a_k T_k(t) , \quad t \in \mathbb{R} , \quad a_k \in \mathbb{R} ,$$

at all the points

$$x_k := \cos\left(\frac{k}{m}\pi\right) , \quad k = 0, \ldots, m , \quad m \in \mathbb{N} .$$

△

6.3 Mean Square Best Approximation

There is a particular family of norms for which the best approximant of a function f in a finite dimensional function space V_N, that is, the element of V_N that is closest to f with respect to that particular norm can actually be computed. It turns out that this computation boils down to solving a kind of least squares problem, similar to the least squares problems in $\mathbb{K}^n$ discussed in Chapter 3.

6.3.1 Abstract Theory

Concerning mean square best approximation it is useful to learn an abstract framework first into which the concrete examples can be fit later.

6.3.1.1 Mean Square Norms

Mean square norms generalize the Euclidean norm on $\mathbb{K}^n$, see [NS02, Sect. 4.4]. In a sense, they endow a vector space with a geometry and give a meaning to concepts like "orthogonality".

Definition 6.3.1.1. (Semi-)inner product [NS02, Sect. 4.4]

Let V be a vector space over the field $\mathbb{K}$. A mapping $b : V \times V \to \mathbb{K}$ is called an inner product on V, if it satisfies

(i) b is linear in the first argument: $b(\alpha v + \beta w, u) = \alpha b(v, u) + \beta b(w, u)$ for all $\alpha, \beta \in \mathbb{K}$, $u, v, w \in V$,

(ii) b is (anti-)symmetric: $b(v, w) = \overline{b(w, v)}$ ($\overline{} \hat{=}$ complex conjugation),

(iii) b is positive definite: $v \neq 0 \quad \Leftrightarrow \quad b(v, v) > 0$.

b is a semi-inner product, if it still complies with (i) and (ii), but is only positive semi-definite: $b(v, v) \geq 0$ for all $v \in V$.

✎ notation: usually we write $(\cdot, \cdot)_V$ for an inner product on the vector space V.

Definition 6.3.1.2. Orthogonality

Let V be a vector space equipped with a (semi-)inner product $(\cdot, \cdot)_V$. Any two elements v and w of V are called orthogonal, if $(v, w)_V = 0$. We write $v \perp w$.

✎ notation: If $W \subset V$ is a (closed) subsspace: $v \perp W \quad :\Leftrightarrow \quad (v, w)_V = 0 \; \forall \, w \in W$.

Theorem 6.3.1.3. Mean square (semi-)norm/Inner product (semi-)norm

If $(\cdot,\cdot)_V$ is a (semi-)inner product (→ Def. 6.3.1.1) on the vector space V, then

$$\|v\|_V := \sqrt{(v,v)_V}$$

defines a (semi-)norm (→ Def. 1.5.5.4) on V, the mean square (semi-)norm/ inner product (semi-)norm induced by $(\cdot,\cdot)_V$.

§6.3.1.4 (Examples for mean square norms)

- The Euclidean norm on $\mathbb{K}^n$ induced by the dot product (Euclidean inner product).

$$(\mathbf{x},\mathbf{y})_{\mathbb{K}^n} := \sum_{j=1}^{n} (\mathbf{x})_j \overline{(\mathbf{y})_j} \quad \text{["Mathematical indexing" !]} \quad \mathbf{x},\mathbf{y} \in \mathbb{K}^n \,.$$

- The L^2-norm (5.2.4.6) on $C^0([a,b])$ induced by the $L^2([a,b])$ inner product

$$(f,g)_{L^2([a,b])} := \int_a^b f(\tau)\overline{g}(\tau)\,\mathrm{d}\tau \,, \quad f,g, \in C^0([a,b]) \,. \tag{6.3.1.5}$$

⌟

6.3.1.2 Normal Equations

Mean square best approximation = best approximation in a mean square norm

From § 3.1.1.8 we know that in Euclidean space $\mathbb{K}^n$ the best approximation of vector $\mathbf{x} \in \mathbb{K}^n$ in a subspace $V \subset \mathbb{K}^n$ is unique and given by the orthogonal projection of $\mathbf{x}$ onto V. Now we generalize this to vector spaces equipped with inner products.

$X \mathrel{\hat{=}}$ a vector space over $\mathbb{K} = \mathbb{R}$, equipped with an mean square semi-norm $\|\cdot\|_X$ induced by a semi-inner product $(\cdot,\cdot)_X$, see Thm. 6.3.1.3.
It can be an infinite dimensional function space, e.g., $X = C^0([a,b])$.

$V \mathrel{\hat{=}}$ a finite-dimensional subspace of X, with basis $\mathfrak{B}_V := \{b_1,\ldots,b_N\} \subset V$, $N := \dim V$.

Assumption 6.3.1.6.

The semi-inner product $(\cdot,\cdot)_X$ is a genuine inner product (→ Def. 6.3.1.1) on V, that is, it is positive definite: $(v,v)_X > 0 \;\forall v \in V \setminus \{0\}$.

Now we give a *formula* for the element q of V, which is nearest to a given element f of X with respect to

the norm $\|\cdot\|_X$. This is a genuine generalization of Thm. 3.1.2.1.

Theorem 6.3.1.7. Mean square norm best approximation through normal equations

Given any $f \in X$ there is a unique $q \in V$ *such that*

$$\|f - q\|_X = \inf_{p \in V} \|f - p\|_X \,.$$

Its coefficients γ_j, $j = 1, \ldots, N$, with respect to the basis $\mathfrak{B}_V := \{b_1, \ldots, b_N\}$ of V ($q = \sum_{j=1}^N \gamma_j b_j$) are the unique solution of the normal equations

$$\mathbf{M}\left[\gamma_j\right]_{j=1}^N = \left[\left(f, b_j\right)_X\right]_{j=1}^N , \quad \mathbf{M} := \begin{bmatrix} (b_1,b_1)_X & \ldots & (b_1,b_N)_X \\ \vdots & & \vdots \\ (b_N,b_1)_X & \ldots & (b_N,b_N)_X \end{bmatrix} \in \mathbb{K}^{N,N} . \tag{6.3.1.8}$$

Proof. (inspired by Rem. 3.1.2.5) We first show that $\mathbf{M}$ is s.p.d. ($\to$ Def. 1.1.2.6). Symmetry is clear from the definition and the symmetry of $(\cdot,\cdot)_X$. That $\mathbf{M}$ is even positive definite follows from

$$\mathbf{x}^{\mathrm{H}}\mathbf{M}\mathbf{x} = \sum_{k=1}^N \sum_{j=1}^N \xi_k \overline{\xi}_j \left(b_k, b_j\right)_X = \left\| \sum_{j=1}^N \xi_j b_j \right\|_X^2 > 0 \,, \tag{6.3.1.9}$$

if $\mathbf{x} := \left[\xi_j\right]_{j=1}^N \neq \mathbf{0} \;\Leftrightarrow\; \sum_{j=1}^N \xi_j b_j \neq 0$, since $\|\cdot\|_X$ is a *norm* on V by Ass. 6.3.1.6.

Now, writing $\mathbf{c} := \left[\gamma_j\right]_{j=1}^N \in \mathbb{K}^N$, $\mathbf{b} := \left[\left(f, b_j\right)_X\right]_{j=1}^N \in \mathbb{K}^N$, and using the basis representation

$$q = \sum_{j=1}^N \gamma_j b_j \,,$$

we find

$$\Phi(\mathbf{c}) := \|f - q\|_X^2 = \|f\|_X^2 - 2\mathbf{b}^\top \mathbf{c} + \mathbf{c}^\top \mathbf{M}\mathbf{c} \,.$$

Applying the differentiation rules from Ex. 8.5.1.19 to $\Phi : \mathbb{K}^N \to \mathbb{R}$, $\mathbf{c} \mapsto \Phi(\mathbf{c})$, we obtain

$$\mathbf{grad}\,\Phi(\mathbf{c}) = 2(\mathbf{M}\mathbf{c} - \mathbf{b}) \,, \tag{6.3.1.10}$$

$$\mathrm{H}\,\Phi(\mathbf{c}) = 2\mathbf{M} \,. \quad \text{(independent of } \mathbf{c}\text{!)} \tag{6.3.1.11}$$

Since $\mathbf{M}$ is s.p.d., the unique solution of $\mathbf{grad}\,\Phi(\mathbf{c}) = \mathbf{M}\mathbf{c} - \mathbf{b} = 0$ yields the unique global minimizer of Φ; the Hessian $2\mathbf{M}$ is s.p.d. everywhere!

□

The unique q from Thm. 6.3.1.7 is called the best approximant of f in V.

Corollary 6.3.1.12. Best approximant by orthogonal projection

If q is the best approximant of f in V, then $f - q$ is orthogonal to every $p \in V$:

$$(f - q, p)_X = 0 \quad \forall p \in V \quad \Leftrightarrow \quad f - q \perp V \,.$$

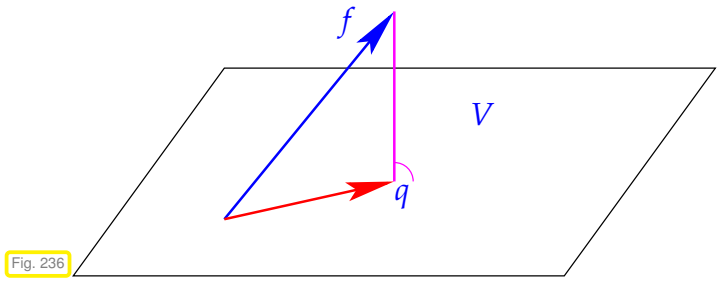


Fig. 236

The message of Cor. 6.3.1.12:

◁ the best approximation error $f - q$ for $f \in X$ in V is orthogonal to the subspace V.

See § 3.1.1.8 for related discussion in Euclidean space $\mathbb{K}^n$.

Remark 6.3.1.13 (Connetion with linear least squares problems Chapter 3) In Section 3.1.1 we introduced the concept of least squares solutions of overdetermined linear systems of equations $\mathbf{Ax} = \mathbf{b}$, $\mathbf{A} \in \mathbb{R}^{m,n}$, $m > n$, see Def. 3.1.1.1. Thm. 3.1.2.1 taught that the normal equations $\mathbf{A}^\top \mathbf{AX} = \mathbf{A}^\top \mathbf{b}$ give the least squares solution, if $\mathrm{rank}(\mathbf{A}) = n$.

In fact, Thm. 3.1.2.1 and the above Thm. 6.3.1.7 agree if $X = \mathbb{K}^n$ (Euclidean space) and $V = \mathrm{Span}\{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$, where $\mathbf{a}_j \in \mathbb{R}^m$ are the columns of $\mathbf{A}$ and $N = n$. ⌟

6.3.1.3 Orthonormal Bases

In the setting of Section 6.3.1.2 we may ask: Which choice of basis $\mathfrak{B} = \{b_1, \ldots, b_N\}$ of $V \subset X$ renders the normal equations (6.3.1.8) particularly simple? Answer: A basis $\mathfrak{B}$, for which $\left(b_k, b_j\right)_X = \delta_{kj}$ (δ_{kj} the Kronecker symbol), because this will imply $\mathbf{M} = \mathbf{I}$ for the coefficient matrix of the normal equations.

Definition 6.3.1.14. Orthonormal basis

A subset $\{b_1, \ldots, b_N\}$ of an N-dimensional vector space V with inner product ($\to$ Def. 6.3.1.1) $(\cdot,\cdot)_V$ is an orthonormal basis (ONB), if $\left(b_k, b_j\right)_V = \delta_{kj}$.

A basis of $\{b_1, \ldots, b_N\}$ of V is called orthogonal, if $\left(b_k, b_j\right)_V = 0$ for $k \neq j$.

Corollary 6.3.1.15. ONB representation of best approximant

If $\{b_1, \ldots, b_N\}$ is an orthonormal basis ($\to$ Def. 6.3.1.14) of $V \subset X$, then the best approximant $q := \mathrm{argmin}_{p \in V} \|f - p\|_X$ of $f \in X$ has the representation

$$q = \sum_{j=1}^{N} \left(f, b_j\right)_X b_j \,. \tag{6.3.1.16}$$

§6.3.1.17 (Gram-Schmidt orthonormalization) From Section 1.5.1 we already know how to compute orthonormal bases: The algorithm from § 1.5.1.1 can be run in the framework of any vector space V

endowed with an inner product $(\cdot,\cdot)_V$ and induced mean square norm $\|\cdot\|_V$.

```
1:  b_1 := p_1/‖p_1‖_V    % 1st output vector
2:  for j = 2,...,k do
      { % Orthogonal projection
3:      b_j := p_j
4:      for ℓ = 1,2,...,j−1 do
5:        { b_j ← b_j − (p_j,b_ℓ)_V b_ℓ }
6:      if ( b_j = 0 ) then STOP
7:      else { b_j ← b_j/‖b_j‖_V }
      }
```

(6.3.1.18)

Theorem 6.3.1.19. Gram-Schmidt orthonormalization

When supplied with $k \in \mathbb{N}$ linearly independent *vectors $p_1,\ldots,p_k \in V$ in a vector space with inner product $(\cdot,\cdot)_V$, Algorithm* (6.3.1.18) *computes vectors $b_1,\ldots,b_k$ with*

$$(b_\ell,b_j)_V = \delta_{\ell j}\,, \quad \ell,j \in \{1,\ldots,k\}\,,$$
$$\mathrm{Span}\{b_1,\ldots,b_\ell\} = \mathrm{Span}\{p_1,\ldots,p_\ell\}$$

for all $\ell \in \{1,\ldots,k\}$.

⌟

This suggests the following alternative approach to the computation of the mean square best approximant q in V of $f \in X$:

❶ Orthonormalize a basis $\{b_1,\ldots,b_N\}$ of V, $N := \dim V$, using Gram-Schmidt algorithm (6.3.1.18).

❷ Compute q according to (6.3.1.16).

Number of inner products to be evaluated: $O(N^2)$ for $N \to \infty$.

Review question(s) 6.3.1.20 (Mean-square best approximation: abstract theory)

(Q6.3.1.20.A) Let V be a finite-dimensional vector space with inner product $(\cdot,\cdot)_V$. What is an orthonormal basis (ONB) of V?

(Q6.3.1.20.B) Let X be a finite-dimensional real vector space with inner product $(\cdot,\cdot)_X$ and equipped with a basis $\{b_1,\ldots,b_N\}$, $N := \dim X$.

Show that the coefficient matrix

$$\mathbf{M} := \begin{bmatrix} (b_1,b_1)_X & \ldots & (b_1,b_N)_X \\ \vdots & & \vdots \\ (b_N,b_1)_X & \ldots & (b_N,b_N)_X \end{bmatrix} \in \mathbb{R}^{N,N}$$

is *symmetric positive definite*.

Definition 1.1.2.6. Symmetric positive definite (s.p.d.) matrices

$\mathbf{M} \in \mathbb{K}^{n,n}$, $n \in \mathbb{N}$, is **symmetric (Hermitian) positive definite** (s.p.d.), if

$$\mathbf{M} = \mathbf{M}^{\mathrm{H}} \quad \text{and} \quad \forall \mathbf{x} \in \mathbb{K}^n\colon \ \mathbf{x}^{\mathrm{H}}\mathbf{M}\mathbf{x} > 0 \ \Leftrightarrow \ \mathbf{x} \neq 0\,.$$

If $\mathbf{x}^{\mathrm{H}}\mathbf{M}\mathbf{x} \geq 0$ for all $\mathbf{x} \in \mathbb{K}^n$ ▷ **$\mathbf{M}$ positive semi-definite**.

△

6.3.2 Polynomial Mean Square Best Approximation

Now we apply the results of Section 6.3.1 in the following setting:

X: function space $C^0([a,b])$, $-\infty < a < b < \infty$, of $\mathbb{R}$-values continuous functions,
V: space $\mathcal{P}_m$ of polynomials of degree $\leq m$.

Remark 6.3.2.1 (Inner products on spaces $\mathcal{P}_m$ of polynomials) To match the abstract framework of Section 6.3.1 we need to find (semi-)inner products on $C^0([a,b])$ that supply positive definite inner products on $\mathcal{P}_m$. The following options are commonly considered:

- On any interval $[a,b]$ we can use the $L^2([a,b])$-inner product $(\cdot,\cdot)_{L^2([a,b)]}$, defined in (6.3.1.5).
- Given a positive integrable **weight function**

$$w : [a,b] \to \mathbb{R}\,, \quad w(t) > 0 \quad \text{for all} \quad t \in [a,b]\,, \quad \int_a^b |w(t)|\,\mathrm{d}t < \infty\,, \tag{6.3.2.2}$$

  we can consider the weighted L^2-inner product on the interval $[a,b]$

$$(f,g)_{w,[a,b]} := \int_a^b w(\tau) f(\tau) \overline{g}(\tau)\,\mathrm{d}\tau\,. \tag{6.3.2.3}$$

- For $n \geq m$ and $n+1$ distinct points collected in the set $\mathcal{T} := \{t_0, t_1, \ldots, t_n\} \subset [a,b]$ we can use the discrete L^2-inner product

$$(f,g)_{\mathcal{T}} := \sum_{j=0}^{n} f(t_j)\overline{g(t_j)}\,. \tag{6.3.2.4}$$

  Since a polynomial of degree $\leq m$ must be zero everywhere, if it vanishes in at least $m+1$ distinct points, $(f,g)_{\mathcal{T}}$ is positive definite on $\mathcal{P}_m$.

For all these inner products on $\mathcal{P}_m$ holds

$$(\{t \mapsto tf(t)\}, g)_X = (f, \{t \mapsto tg(t)\})_X\,, \quad f,g \in C^0([a,b])\,, \tag{6.3.2.5}$$

that is, multiplication with the independent variable can be shifted to the other function inside the inner product.

✎ notation: Note that we have to plug a function into the slots of the inner products; this is indicated by the notation $\{t \mapsto \ldots\}$.

Assumption 6.3.2.6. Self-adjointness of multiplication operator

We assume the inner product $(\cdot,\cdot)_X$ to satisfy (6.3.2.5).

⌟

The ideas of Section 6.3.1.3 that center around the use of orthonormal bases can also be applied to

polynomials.

Definition 6.3.2.7. Orthonormal polynomials → **Def. 6.3.1.14**

Let $(\cdot,\cdot)_X$ be an inner product on $\mathcal{P}_m$. A sequence $r_0, r_1, \ldots, r_m$ provides orthonormal polynomials (ONPs) with respect to $(\cdot,\cdot)_X$, if

$$r_\ell \in \mathcal{P}_\ell \quad , \quad (r_k, r_\ell)_X = \delta_{k\ell} \; , \quad \ell, k \in \{0, \ldots, m\} \; . \tag{6.3.2.8}$$

The polynomials are just orthogonal, if $(r_k, r_\ell)_X = 0$ for $k \neq \ell$.

By virtue of Thm. 6.3.1.19 orthonormal polynomials can be generated by applying Gram-Schmidt orthonormalization from § 6.3.1.17 to the ordered basis of monomials $\{t \mapsto t^j\}_{j=0}^m$.

Lemma 6.3.2.9. Uniqueness of orthonormal polynomials

The sequence of orthonormal polynomials from Def. 6.3.2.7 is unique up to signs, supplies an $(\cdot,\cdot)_X$-orthonormal basis (→ Def. 6.3.1.14) of $\mathcal{P}_m$, and satisfies

$$\operatorname{Span}\{r_0, \ldots, r_k\} = \mathcal{P}_k \; , \quad k \in \{0, \ldots, m\} \; . \tag{6.3.2.10}$$

Proof. Comparing Def. 6.3.1.14 and (6.3.2.8) the ONB-property of $\{r_0, \ldots, r_m\}$ is immediate. Then (6.3.2.8) follows from dimensional considerations.

r_0 must be a constant, which, up to sign, is fixed by the normalization condition $\|r_0\|_X = 1$.

$\mathcal{P}_{k-1} \subset \mathcal{P}_k$ has co-dimension 1 so that there a unit "vector" in $\mathcal{P}_k$, which is orthogonal to $\mathcal{P}_{k-1}$ and unique up to sign.

□

§6.3.2.11 (Orthonormal polynomials by orthogonal projection) Let $r_0, \ldots, r_m \in \mathcal{T}_p$ be a sequence of orthonormal polynomials according to Def. 6.3.2.7. From (6.3.2.8) we conclude that $r_k \in \mathcal{P}_k$ has leading coefficient $\neq 0$.

Hence $s_k(t) := t \cdot r_k(t)$ is a polynomial of degree $k+1$ with leading coefficient $\neq 0$, that is $s_k \in \mathcal{P}_{k+1} \setminus \mathcal{P}_k$. Therefore, r_{k+1} can be obtain by *orthogonally projecting* s_k onto $\mathcal{P}_k$ plus normalization, *cf.* Lines 4-5 of Algorithm (6.3.1.18):

$$r_{k+1} = \pm \frac{\widetilde{r}_{k+1}}{\|\widetilde{r}_{k+1}\|_X} \; , \quad \widetilde{r}_{k+1} = s_k - \sum_{j=0}^{k} (s_k, r_j)_X r_j \; . \tag{6.3.2.12}$$

Straightforward computations confirm that $r_{k+1} \perp \mathcal{P}_k$.

The sum in (6.3.2.12) collapses to two terms! In fact, since $(r_k, q)_X = 0$ for all $q \in \mathcal{P}_{k-1}$, by Ass. 6.3.2.6

$$(s_k, r_j)_X = (\{t \mapsto t r_k(t)\}, r_j)_X \overset{(6.3.2.5)}{=} (r_k, \{t \mapsto t r_j\})_X = 0 \; , \quad \text{if } j < k-1 \; ,$$

because in this case $\{t \mapsto t r_j\} \in \mathcal{P}_{k-1}$. As a consequence (6.3.2.12) reduces to the 3-term recursion

$$\begin{aligned} r_{k+1} &= \pm \frac{\widetilde{r}_{k+1}}{\|\widetilde{r}_{k+1}\|_X} \; , \\ \widetilde{r}_{k+1} &= s_k - (\{t \mapsto t r_k\}, r_k)_X r_k - (\{t \mapsto t r_k\}, r_{k-1})_X r_{k-1} \end{aligned} \quad , \quad k = 1, \ldots, m-1 \; . \tag{6.3.2.13}$$

The recursion starts with $r_{-1} := 0$, $r_0 = \{t \mapsto 1/\|1\|_X\}$. ⌟

The 3-term recursion (6.3.2.13) can be recast in various ways. Forgoing normalization the next theorem presents one of them.

Theorem 6.3.2.14. 3-term recursion for orthogonal polynomials

*Given **any** inner product $(\cdot,\cdot)_X$ on $\mathcal{P}_m$, $m \in \mathbb{N}$, define $p_{-1} := 0$, $p_0 = 1$, and*

$$p_{k+1}(t) := (t - \alpha_{k+1})p_k(t) - \beta_k p_{k-1}(t)\,, \quad k = 0,1,\ldots,m-1\,,$$
$$\text{with} \quad \alpha_{k+1} := \frac{(\{t \mapsto tp_k(t)\}, p_k)_X}{\|p_k\|_X^2}\,, \quad \beta_k := \frac{\|p_k\|_X^2}{\|p_{k-1}\|_X^2}\,. \tag{6.3.2.15}$$

Then
- $p_k \in \mathcal{P}_k$ *has leading coefficient* $= 1$, *and*
- $\{p_0, p_1, \ldots, p_m\}$ *is an orthogonal basis of* $\mathcal{P}_m$.

Proof. (by rather straightforward induction) We first confirm, thanks to the definition of α_1,

$$(p_0, p_1)_X = (p_0, \{t \mapsto (t-\alpha_1)p_0(t)\})_X = (p_0, \{t \mapsto tp_0(t)\})_X - \alpha_1(p_0,p_0)_X = 0\,.$$

For the induction step we assume that the assertion is true for $p_0, \ldots, p_k$ and observe that for p_{k+1} according to (6.3.2.15) we have

$$\begin{aligned}
(p_k, p_{k+1})_X &= (p_k, \{t \mapsto (t-\alpha_{k+1})p_k(t)\} - \beta_k p_{k-1})_X \\
&= (p_k, \{t \mapsto tp_k(t)\})_X - \underbrace{\alpha_{k+1}(p_k,p_k)_X}_{=(\{t\mapsto tp_k(t)\},p_k)_X} - \beta_k \underbrace{(p_k, p_{k-1})_X}_{=0} = 0\,,\\
(p_{k-1}, p_{k+1})_X &= (p_{k-1}, \{t \mapsto (t-\alpha_{k+1})p_k(t)\} - \beta_k p_{k-1})_X \\
&= (p_{k-1}, \{t \mapsto tp_k(t)\})_X - \alpha_{k+1}(p_{k-1},p_k)_X - \beta_k(p_k,p_{k-1})_X = 0\,,\\
(p_\ell, p_{k+1})_X &= (p_\ell, \{t \mapsto (t-\alpha_{k+1})p_k(t)\} - \beta_k p_{k-1})_X \\
&= \underbrace{(p_\ell, \{t \mapsto tp_k(t)\})_X}_{=0} - \alpha_{k+1}\underbrace{(p_\ell,p_k)_X}_{0} - \beta_k\underbrace{(p_\ell,p_{k-1})_X}_{0} = 0\,, \quad \ell = 0,\ldots,k-2\,.
\end{aligned}$$

This amounts to the assertion of orthogonality for $k+1$. Above, several inner product vanish because of the induction hypothesis!

□

Remark 6.3.2.16 ($L^2([-1,1])$-orthogonal polynomials) An important inner product on $C^0[-1,1]$ is the L^2-inner product, see (6.3.1.5)

$$(f,g)_{L^2([-1,1])} := \int_{-1}^{1} f(\tau)\overline{g}(\tau)\,\mathrm{d}\tau\,, \quad f,g, \in C^0([-1,1])\,.$$

It is a natural question what is the unique sequence of $L^2([-1,1])$-orthonormal polynomials. Their rather simple characterization will be discussed in the sequel.

The Legendre polynomials P_n can be defined by the 3-term recursion

$$P_{n+1}(t) := \frac{2n+1}{n+1} t P_n(t) - \frac{n}{n+1} P_{n-1}(t) , \tag{7.4.2.21}$$

$$P_0 := 1 , \quad P_1(t) := t .$$

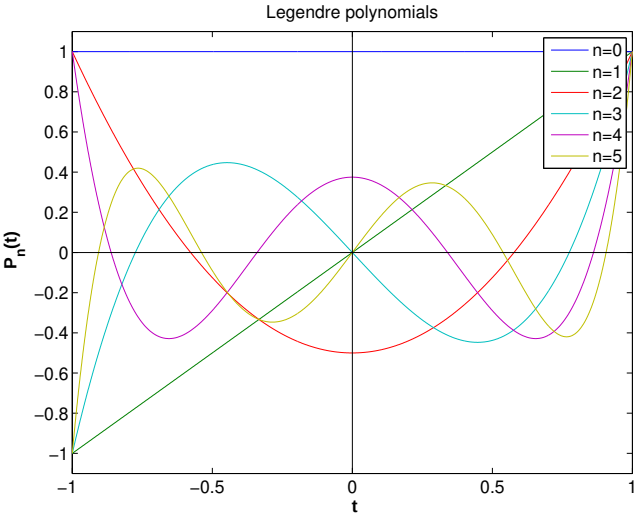


Fig. 237

$P_k \in \mathcal{P}_k$ is immediate and so is the parity

$$P_k(t) = P_k(-t) \quad \text{if } k \text{ is even}, \quad P_k(t) = -P_k(-t) \quad \text{if } k \text{ is odd} . \tag{6.3.2.17}$$

Orthogonality, $(P_k, P_m)_{L^2(\Omega)} = 0$ or $m \neq k$, as well as $\|\|_{L^{P_k}([-1,1)} 2^2 = {}^2/_{2n+1}$ can be proved by induction based on (6.3.2.17).This implies that the $L^{]}([-1,1)2$-orthonormal polynomials are $r_k = \sqrt{k+{}^1/_2} P_k$ and their 3-term recursion from Thm. 6.3.2.14 reads

$$r_{n+1}(t) = \frac{\sqrt{4(n+1)^2-1}}{n+1} t r_n(t) - \frac{n}{n+1}\sqrt{\frac{4(n+1)^2-1}{4n^2-1}}\, u_{n-1}(t) , \tag{6.3.2.18}$$

$$r_{-1} := 0 , \quad r_0 = \tfrac{1}{2}\sqrt{2} . \tag{6.3.2.19}$$

⌟

§6.3.2.20 (Discrete orthogonal polynomials) Since they involve integrals, weighted L^2-inner products (6.3.2.3) are not accessible computationally, unless one resigns to approximation, see Chapter 7 for corresponding theory and techniques.

Therefore, given a point set $\mathcal{T} := \{t_0, t_1, \ldots, t_n\}$, we focus on the associated discrete L^2-inner product

$$(f,g)_X := (f,g)_{\mathcal{T}} := \sum_{j=0}^{n} f(t_j)\overline{g(t_j)} , \quad f,g \in C^0([a,b]) , \tag{6.3.2.4}$$

which is positive definite on $\mathcal{P}_n$ and satisfies Ass. 6.3.2.6.

The polynomials p_k generated by the 3-term recursion (6.3.2.15) from Thm. 6.3.2.14 are then called discrete orthogonal polynomials. The following C++ code computes the recursion coefficients α_k and β_k, $k = 1, \ldots, n-1$.

C++ code 6.3.2.21: Computation of weights in 3-term recursion for discrete orthogonal polynomials ➙ GITLAB

```
2 // Computation of coefficients α, β from 6.3.2.14
3 // IN : t = points in the definition of the discrete L^2-inner product
4 // n = maximal index desired
5 // alpha, beta are used to save coefficients of recursion
6 void coeffortho(const VectorXd& t, const Index n, VectorXd& alpha, VectorXd& beta) {
7   const Index m = t.size(); // maximal degree of orthogonal polynomial
```

```
8    alpha = VectorXd( std::min(n-1, m-2) + 1 );
9    beta = VectorXd( std::min(n-1, m-2) + 1 );
10   alpha(0) = t.sum()/static_cast<double>(m);
11   // initialization of recursion; we store only the values of
12   // the polynomials a the points in T
13   VectorXd p0;
14   VectorXd p1 = VectorXd::Ones(m);
15   VectorXd p2 = t - alpha(0)*VectorXd::Ones(m);
16   for (Index k = 0; k < std::min(n-1, m-2); ++k) {
17     p0 = p1; p1 = p2;
18     // 3-term recursion (6.3.2.15)
19     alpha(k+1) = p1.dot(t.cwiseProduct(p1))/p1.squaredNorm();
20     beta(k) = p1.squaredNorm()/p0.squaredNorm();
21     p2 = (t-alpha(k+1)*VectorXd::Ones(m)).cwiseProduct(p1)-beta(k)*p0;
22   }
23 }
```

⌟

§6.3.2.22 (Polynomial fitting) Given a point set $\mathcal{T} := \{t_0, t_1, \ldots, t_n\} \subset [a,b]$, and a function $f : [a,b] \to \mathbb{K}$, we may seek to approximate f by its polynomial best approximant with respect to the discrete L^2-norm $\|\cdot\|_{\mathcal{T}}$ induced by the discrete L^2-inner product (6.3.2.4).

Definition 6.3.2.23. Fitted polynomial

Given a point set $\mathcal{T} := \{t_0, t_1, \ldots, t_n\} \subset [a,b]$, and a function $f : [a,b] \to \mathbb{K}$ we call

$$q_k := \operatorname*{argmin}_{p \in \mathcal{P}_k} \|f - p\|_{\mathcal{T}} , \quad k \in \{0, \ldots, n\} ,$$

the **fitted polynomial** to f on $\mathcal{T}$ of degree k, $k \leq n$.

The stable and efficient computation of fitting polynomials can rely on combining Thm. 6.3.2.14 with Cor. 6.3.1.15:

❶ (Pre-)compute the weights α_ℓ and β_ℓ for the 3-term recursion (6.3.2.15).

❷ (Pre-)compute the values of the orthogonal polynomials p_k at desired evaluation points $x_i \in \mathbb{R}$, $i = 1, \ldots, N$.

❸ Compute the inner product $(f, p_\ell)_X$, $\ell = 0, \ldots, k$, and use (6.3.1.16) to linearly combine the vectors $[p_\ell(x_i)]_{i=1}^N$, $\ell = 0, \ldots, k$.

This yields $[q(x_i)]_{i=1}^N$, $q \in \mathcal{P}_k$ the fitting polynomial. ⌟

EXAMPLE 6.3.2.24 (Approximation by discrete polynomial fitting) We use equidistant points $\mathcal{T} := \{t_k = -1 + k\frac{2}{m},\ k = 0, \ldots, m\} \subset [-1,1]$, $m \in \mathbb{N}$ to compute fitting polynomials ($\to$ Def. 6.3.2.23) for two different functions.

We monitor the L^2-norm and L^∞-norm of the approximation error, both norms approximated by sampling in $\xi_j = -1 + \frac{j}{500}$, $j = 0, \ldots, 1000$.

① $f(t) = (1 + (5t)^2)^{-1}$, $I = [-1,1]$ $\to$ Ex. 6.2.2.11, analytic in complex neighborhood of $[-1,1]$:

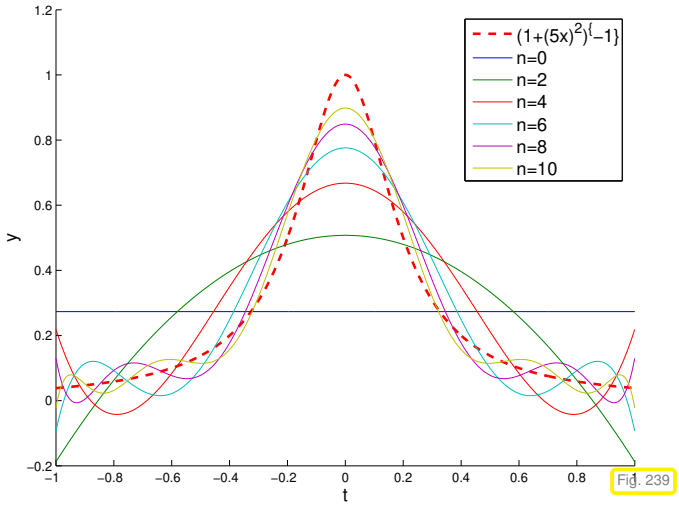


Fig. 238

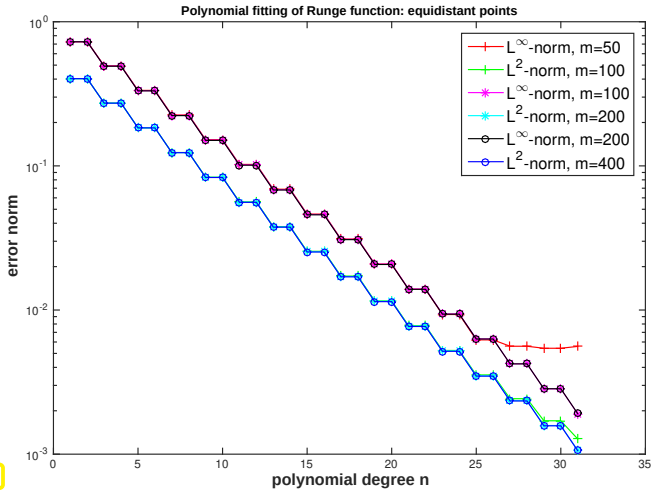


Fig. 239

➢ We observe exponential convergence (→ Def. 6.2.2.7) in the polynomial degree n.

② $f(t) = \max\{0, 1 - 2 * |x + \frac{1}{4}|\}$, f only in $C^0([-1,1])$:

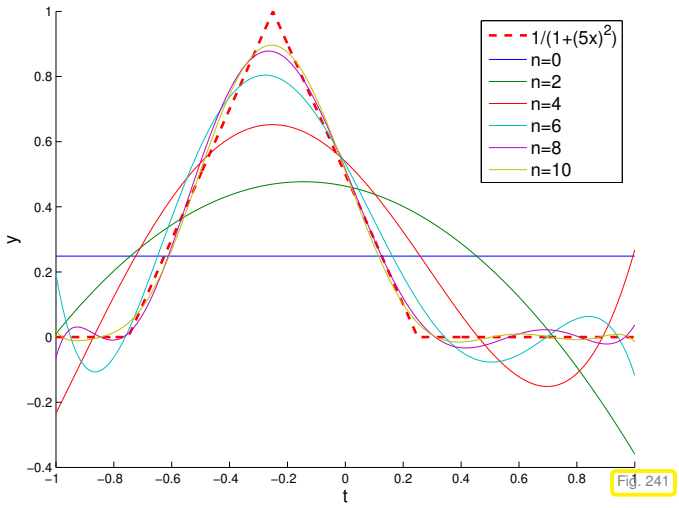


Fig. 240

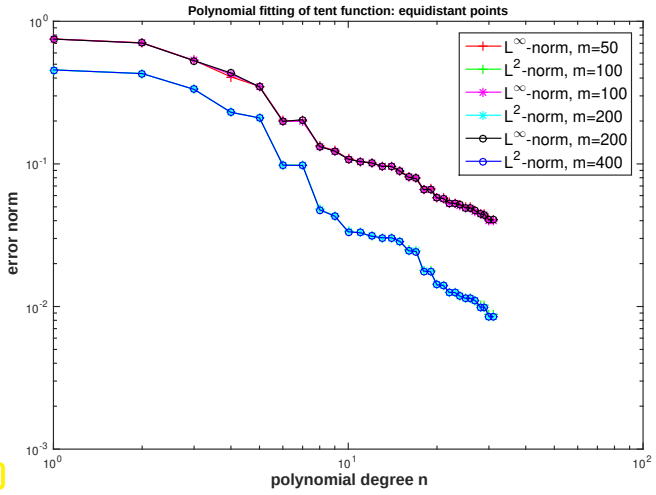


Fig. 241

➢ We observe only algebraic convergence (→ Def. 6.2.2.7 in the polynomial degree n (for $n \ll m$!).

③ "bump function"

$$f(t) = \max\{\cos(4\pi|t + \frac{1}{4}|), 0\} \ .$$

➢ Merely $f \in C^1([-1,1])$

Doubly logarithmic plot suggests "asymptotic" algebraic convergence.

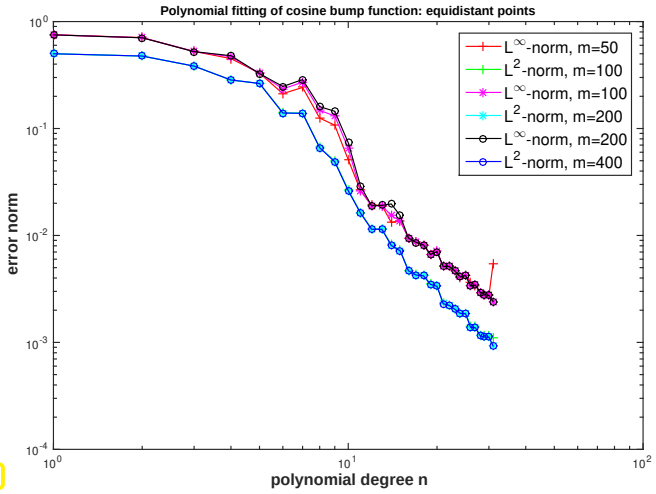


Fig. 242

Review question(s) 6.3.2.25 (Polynomial mean-square best approximation)

(Q6.3.2.25.A) Given $a, b \in \mathbb{R}$ and a positive continuous weight function $w \in C^0([a,b])$, $w(t) > 0$ for all $t \in [a,b]$, write $\{P_0, P_1, P_2\}$, $P_j \in \mathcal{P}_j$, $j \in \mathbb{N}_0$, for the sequence of orthonormal polynomials with respect

to the weighted L^2-inner product on the interval $[a,b]$

$$(f,g)_{w,[a,b]} := \int_a^b w(\tau) f(\tau) \overline{g}(\tau) \,\mathrm{d}\tau \,. \tag{6.3.2.3}$$

Give an indirect proof that P_j must have j distinct zeros in $]a,b[$. To that end assume that P_j has only $\ell < j$ zeros $z_1, \ldots, z_\ell$ in $]a,b[$, at which it changes sign and consider the polynomial

$$q(t) := (t - z_1) \cdot \cdots \cdot (t - z_\ell) \quad , \quad q \in \mathcal{P}_\ell \,,$$

in order to arrive at a contradiction.

△

6.4 Uniform Best Approximation

§6.4.0.1 (The alternation theorem) Given an interval $[a,b]$ we seek a best approximant of a function $f \in C^0([a,b])$ in the space $\mathcal{P}_n$ of polynomials of degree $\leq n$ with respect to the supremum norm $\|\cdot\|_{L^\infty([a,b])}$:

$$q \in \operatorname*{argmin}_{p \in \mathcal{P}_n} \|f - p\|_{L^\infty(I)} \,.$$

The results of Section 6.3.1 cannot be applied because the supremum norm is not induced by an inner product on $\mathcal{P}_n$.

Theory provides us with surprisingly precise necessary and sufficient conditions to be satisfied by the polynomial $L^\infty([a,b])$-best approximant q.

Theorem 6.4.0.2. Chebychev alternation theorem

Given $f \in C^0[a,b]$, $a < b$, and a polynomial degree $n \in \mathbb{N}$, a polynomial $q \in \mathcal{P}_n$ satisfies

$$q = \operatorname*{argmin}_{p \in \mathcal{P}_n} \|f - p\|_{L^\infty(I)}$$

if and only if *there exist $n+2$ points $a \leq \xi_0 < \xi_1 < \cdots < \xi_{n+1} \leq b$ such that*

$$|e(\xi_j)| = \|e\|_{L^\infty([a,b])} \,, \quad j = 0, \ldots, n+1 \,,$$
$$e(\xi_j) = -e(\xi_{j+1}) \,, \quad j = 0, \ldots, n \,,$$

where $e := f - q$ denotes the approximation error.

Visualization of the behavior of the $L^\infty([a,b])$-best approximation error $e := f - q$ according to the Chebychev alternation theorem Thm. 6.4.0.2. ▷

The extrema of the approximation error are sometimes called alternants.

Compare with the shape of the Chebychev polynomials → Def. 6.2.3.3.

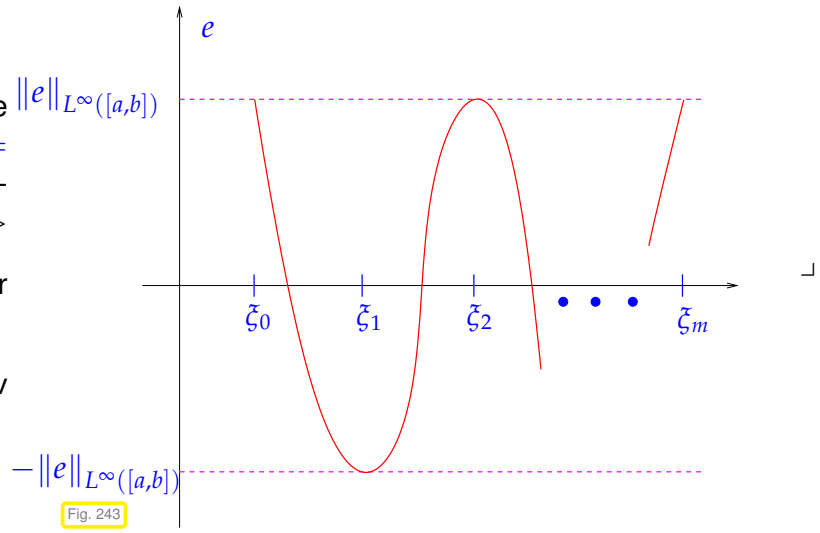


Fig. 243

§6.4.0.3 (Remez algorithm) The widely used iterative algorithm (Remez algorithm) for finding an L^∞-best approximant is motivated by the alternation theorem. The idea is to determine successively better approximations of the set of alternants: $\mathcal{A}^{(0)} \to \mathcal{A}^{(1)} \to \ldots$, $\sharp\mathcal{A}^{(l)} = n+2$.

Key is the observation that, due to the alternation theorem, the polynomial $L^\infty([a,b])$-best approximant q will satisfy (one of the) interpolation conditions

$$q(\xi_k) \pm (-1)^k \delta = f(\xi_k)\,, \quad k - 0, \ldots, n+1\,, \quad \delta := \|f-q\|_{L^\infty([a,b])}\,. \tag{6.4.0.4}$$

① Initial guess $\mathcal{A}^{(0)} := \{\xi_0^{(0)} < \xi_1^{(0)} < \cdots < \xi_n^{(0)} < \xi_{n+1}^{(0)}\} \subset [a,b]$ "arbitrary", for instance extremal points of the Chebychev polynomial T_{n+1}, → Def. 6.2.3.3, so so-called Chebychev alternants

$$\xi_j^{(0)} = \tfrac{1}{2}(a+b) + \tfrac{1}{2}(b-a)\cos\left(\frac{j}{n+1}\pi\right)\,, \quad j = 0, \ldots, n+1\,. \tag{6.4.0.5}$$

② Given approximate alternants $\mathcal{A}^{(l)} := \{\xi_0^{(l)} < \xi_1^{(l)} < \cdots < \xi_n^{(l)} < \xi_{n+1}^{(l)}\} \subset [a,b]$ determine $q \in \mathcal{P}_n$ and a deviation $\delta \in \mathbb{R}$ satisfying the extended interpolation condition

$$q(\xi_k^{(l)}) + (-1)^k \delta = f(\xi_k^{(l)})\,, \quad k = 0, \ldots, n+1\,. \tag{6.4.0.6}$$

After choosing a basis for $\mathcal{P}_n$, this is $(n+2) \times (n+2)$ linear system of equations, *cf.* § 5.1.0.21.

③ Choose $\mathcal{A}^{(l+1)}$ as the set of extremal points of $f-q$, truncated in case more than $n+2$ of these exist.

These extreme can be located approximately by sampling on a fine grid covering $[a,b]$). If the derivative of $f \in C^1([a,b])$ is available, too, then search for zeros of $(f-p)'$ using the secant method from § 8.4.2.28.

④ If $\|f-q\|_{L^\infty([a,b])}| \le \texttt{TOL} \cdot \|d\|_{L^\infty([a,b])}$ STOP, else GOTO ②.
(TOL is a prescribed relative tolerance.)

C++ code 6.4.0.7: Remez algorithm for uniform polynomial approximation on an interval
➙ GITLAB

```
2 // IN : f = handle to the Function, point evaluation
```

```
// df = handle to the Derivative of f, point evaluation
// a, b = interval boundaries
// d = degree of polynomial
// c will be saved to save the coefficients of the interpolant in
    monomial
// basis
template <class Function, class Derivative>
void remez(const Function &f, const Derivative &df, const double a,
           const double b, const unsigned d, const double tol, VectorXd &c) {
  const unsigned n = 8 * d;                    // number of sampling points
  const VectorXd xtab = VectorXd::LinSpaced(n, a, b); // points of sampling grid
  const VectorXd ftab = feval(f, xtab); // function values at sampling grid
  const double fsupn =
    ftab.cwiseAbs().maxCoeff();   // approximate supremum norm of f
  const VectorXd dftab = feval(df, xtab); // derivative values at sampling
     grid

  // The vector xe stores the current guess for the alternants
  // initial guess is Chebychev alternant (6.4.0.5)
  const double h = M_PI / (d + 1);
  VectorXd xe(d + 2);
  for (unsigned i = 0; i < d + 2; ++i) {
    xe(i) = (a + b) / 2. + (a - b) / 2. * std::cos(h * i);
  }

  VectorXd fxe = feval(f, xe); // f evaluated at alternants
  const unsigned maxit = 10;
  // Main iteration loop of Remez algorithm
  for (unsigned k = 0; k < maxit; ++k) {
    // Interpolation at d+2 points xe with deviations
    // ±δ Algorithm uses monomial basis, which is not
    // optimal
    MatrixXd V = vander(xe);
    MatrixXd A(d + 2, d + 2);
    // build Matrix A, LSE
    A.block(0, 0, d + 2, d + 1) = V.block(0, 1, d + 2, d + 1);
    for (unsigned r = 0; r < d + 2; ++r) {
      A(r, d + 1) = std::pow(-1, r);
    }

    c = A.lu().solve(fxe); // solve for coefficients of polynomial q
    VectorXd cd(
        d); // to compute monomial coefficients of derivative q'
    for (unsigned i = 0; i < d; ++i) {
      cd(i) = (d - i) * c(i);
    }

    // Find initial guesses for the inner extremes by sampling
    // track sign changes of the derivative of the approximation error
    VectorXd deltab = polyval(cd, xtab) - dftab;
    const VectorXd s = deltab.head(n - 1).cwiseProduct(deltab.tail(n - 1));
    const VectorXd ind = findNegative(s); // ind = find(s < 0)
    VectorXd xx0 = select(xtab, ind);    // approximate zeros of e'
    const unsigned nx = ind.size(); // number of approximate zeros

    if (nx < d) { // too few extrema; bail out
      std::cerr << "Too few extrema!\n";
      return;
    }

    // Secant method to determin zeros of derivative of approximation
```

```
    // error
    VectorXd F0 = polyval(cd, xx0) - feval(df, xx0);
    // initial guesses from shifting sampling points
    VectorXd xx1 = xx0 + (b - a) / (2 * n) * VectorXd::Ones(xx0.size());
    VectorXd F1 = polyval(cd, xx1) - feval(df, xx1);
    // Main loop of the secant method
    while (F1.cwiseAbs().minCoeff() > 1e-12) {
      const VectorXd xx2 = xx1 - (F1.cwiseQuotient(F1 - F0)).cwiseProduct(xx1 - xx0);
      xx0 = xx1;
      xx1 = xx2;
      F0 = F1;
      F1 = polyval(cd, xx1) - feval(df, xx1);
    }

    // Determine new approximation for alternants; store in xe
    // If too many zeros of the derivative (f - p)'
    // have been found, select those, where the deviation is maximal
    if (nx == d) {
      xe = VectorXd(xx0.size() + 2);
      xe << a, xx0, b;
    } else if (nx == d + 1) {
      xe = VectorXd(xx0.size() + 1);
      if (xx0.minCoeff() - a > b - xx0.maxCoeff()) {
        xe << a, xx0;
      }
      else {
        xe << xx0, b;
      }
    } else if (nx == d + 2) {
      xe = xx0;
    } else {
      const VectorXd del = (polyval(c.head(d + 1), xx0) - feval(f, xx0)).cwiseAbs();
      VectorXd ind = sort_indices(del);
      xe = select(xx0, ind.tail(d + 2));
    }

    // Deviation in sampling points and approximate alternants
    fxe = feval(f, xe);
    VectorXd del(xe.size() + 2);
    del << polyval(c.head(d + 1), a * VectorXd::Ones(1)) -
      ftab(0) * VectorXd::Ones(1),
      polyval(c.head(d + 1), xe) - fxe,
      polyval(c.head(d + 1), b * VectorXd::Ones(1)) -
      ftab(ftab.size() - 1) * VectorXd::Ones(1);
    // Approximation of supremum norm of approximation error
    const double dev = del.cwiseAbs().maxCoeff();
    // Termination of Remez iteration
    if (dev < tol * fsupn) {
      break;
    }
  }
  const VectorXd tmp = c.head(d + 1);
  c = tmp;
}
```

EXPERIMENT 6.4.0.8 (Convergence of Remez algorithm) We examine the convergence of the Remez algorithm from Code 6.4.0.7 for two different functions:

- $f(t) = (1+t^2)^{-1}$, $I = [-5,5]$ → Bsp. 6.2.2.11
- $f(t) = \begin{cases} \frac{1}{2}(1+\cos(2\pi t)) & \text{, if } |t| < \frac{1}{2} \\ 0 & \text{else.} \end{cases}$, $I = [-1,1]$

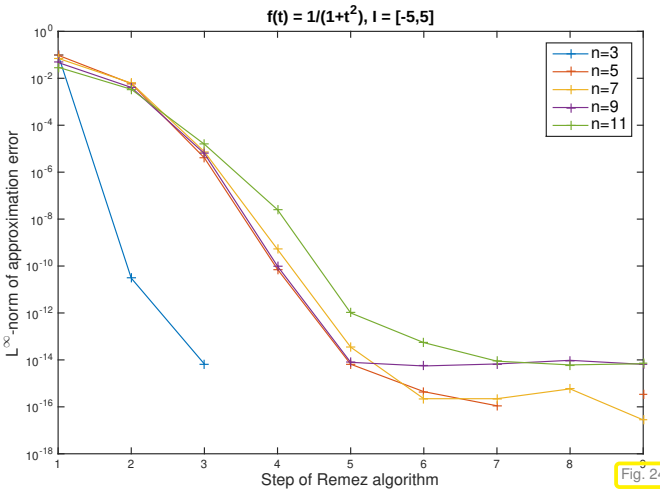


Fig. 244

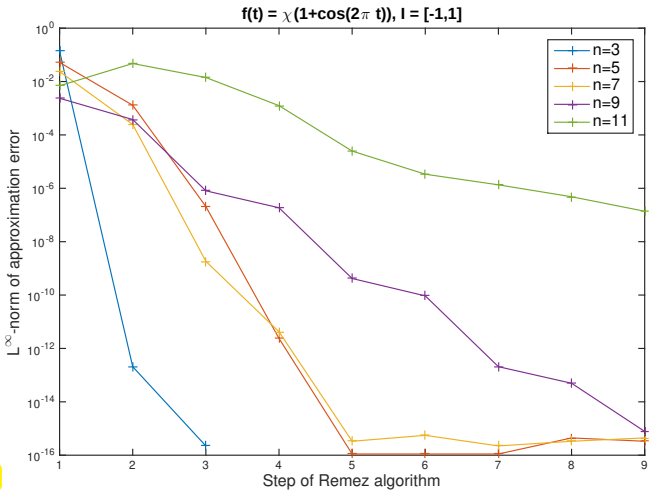


Fig. 245

Convergence in both cases; faster convergence observed for smooth function, for which machine precision is reached after a few steps. ⌟

Review question(s) 6.4.0.9 (Uniform best approximation)

△

6.5 Approximation by Trigonometric Polynomials

Now we address the approximation of a continuous *1-periodic* function

$$f \in C^0(\mathbb{R}) \quad , \quad f(t+1) = f(t) \quad \forall t \in \mathbb{R} \, .$$

Policy: In the interest of "structure preservation" approximate f in space of functions with "built-in" 1-periodicity. This already rules out approximation on $[0,1]$ by global polynomials, because those can never the extended to globally 1-periodic functions.

The natural space for approximating generic periodic functions is a space of trigonometric polynomials with the same period.

Remember from Def. 5.6.1.1 different ways to represent the space $\mathcal{P}_{2n}^T$ of 1-periodic trigonometric polynomials of degree $2n$, $n \in \mathbb{N}$. We can use either a real-valued or a complex-valued basis.

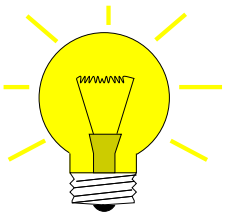
$$\mathcal{P}_{2n}^T = \operatorname{Span}\Big\{ t \mapsto 1, t \mapsto \sin(2\pi t), t \mapsto \cos(2\pi t), t \mapsto \sin(4\pi t), t \mapsto \cos(4\pi t), \ldots \\ t \mapsto \sin(2\pi n t), t \mapsto \cos(2\pi n t) \Big\} \tag{6.5.0.1a}$$

$$\mathcal{P}_{2n}^T = \operatorname{Span}\{ t \mapsto \exp(-2\pi \imath k t) : k = -n \ldots, n \} \, . \tag{6.5.0.1b}$$

Both sets of functions provide a basis for the same space $\mathcal{P}_{2n}^T$, when considered as a vector space over $\mathbb{C}$ of $\mathbb{C}$-valued functions on $\mathbb{R}$. The complex-valued basis allows simpler manipulations and will mainly be used in the sequel.

6.5.1 Approximation by Trigonometric Interpolation

Video tutorial for Section 6.5.1 "Approximation by Trigonometric Interpolation": (5 minutes)
Download link, tablet notes



Idea: Adapt the policy of approximation by interpolation from § 6.1.0.6

Here: Employ trigonometric interpolation from Section 5.6
into space $\mathcal{P}_{2n}^T$ of 1-periodic trigonometric polynomials → Def. 5.6.1.1.

Recall: Trigonometric interpolation → Section 5.6

Given nodes $t_0 < t_1 < \cdots < t_{2n}$, $t_k \in [0,1[$, and values $y_k \in \mathbb{R}$, $k = 0,\ldots,2n$ find

$$q \in \mathcal{P}_{2n}^T := \operatorname{Span}\{t \mapsto \cos(2\pi jt), t \mapsto \sin(2\pi jt)\}_{j=0}^{n}\,, \tag{6.5.1.2}$$

$$\text{with} \quad q(t_k) = y_k \quad \text{for all} \quad k = 0,\ldots,2n\,. \tag{6.5.1.3}$$

Terminology: $\mathcal{P}_{2n}^T \hat{=}$ space of trigonometric polynomials of degree $2n$.

From Section 5.6 remember a few more facts about trigonometric polynomials and trigonometric interpolation:

- Cor. 5.6.1.6: Dimension of the space of trigonometric polynomials: $\dim \mathcal{P}_{2n}^T = 2n+1$
- Trigonometric interpolation can be reduced to polynomial interpolation on the unit circle $\mathbb{S}^1 \subset \mathbb{C}$ in the complex plane, see (5.6.1.5).
  ⇒ existence & uniqueness of trigonometric interpolant q satisfying (6.5.1.2) and (6.5.1.3)
- There are very efficient FFT-based algorithms for trigonometric interpolation in equidistant nodes $t_k = \frac{k}{2n+1}$, $k = 0,\ldots,2n$, see Code 5.6.3.4.

The relationship of trigonometric interpolation and polynomial interpolation on the unit circle suggests a uniform distribution of nodes for general trigonometric interpolation.

Nodes for function approximation by trigonometric interpolation

Trigonometric approximation of generic *1-periodic* continuous functions $\in C^0(\mathbb{R})$ in $\mathcal{P}_{2n}^T$ usually relies on equidistant interpolation nodes $t_k = \dfrac{k}{2n+1}$, $k = 0,\ldots,2n$.

✎ notation: trigonometric interpolation operator in $2n+1$ equidistant nodes $t_k = \frac{k}{2n+1}$, $k = 0,\ldots,2n$

$$\mathsf{T}_n : C^0([0,1[) \to \mathcal{P}_{2n}^T \quad , \quad \mathsf{T}_n(f)(t_k) = f(t_k) \quad \forall k \in \{0,\ldots,2n\}\,. \tag{6.5.1.5}$$

Note that a function $f \in C^0([0,1[)$ can spawn a discontinuous 1-periodic function on $\mathbb{R}$

A prominent example is the "sawtooth function" ▷

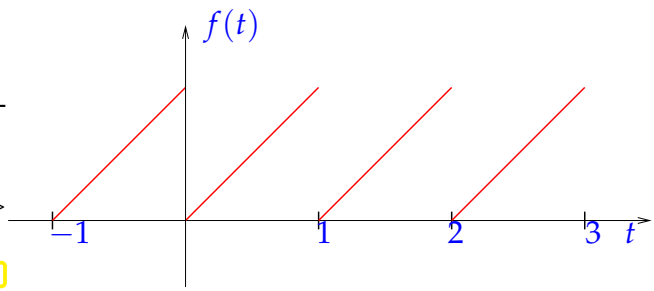


Fig. 246

Review question(s) 6.5.1.6 (Approximation by trigonometric interpolation)

(Q6.5.1.6.A) Two different (ordered) sets of basis functions for the space $\mathcal{P}_{2n}^T$ of trigonometric polynomials of degree $2n$ are widely used:

1. The real-valued trigonometric basis

$$\mathcal{P}_{2n}^T = \operatorname{Span}\Big\{ t \mapsto 1, t \mapsto \sin(2\pi t), t \mapsto \cos(2\pi t), t \mapsto \sin(4\pi t), t \mapsto \cos(4\pi t), \dots \\ t \mapsto \sin(2\pi n t), t \mapsto \cos(2\pi n t) \Big\} .$$

2. The complex-valued exponential basis

$$\mathcal{P}_{2n}^T = \operatorname{Span}\{ t \mapsto \exp(-2\pi \imath k t) : k = -n \dots, n \} .$$

State the matrix $\mathbf{C} \in \mathbb{C}^{2n+1,2n+1}$ converting a representation in trigonometric basis into a representation with respect to exponential basis:

(Q6.5.1.6.B) When can a function $f_0 \in C^m([0,1])$ be extended to a 1-periodic function $f \in C^m(\mathbb{R})$?

That f is an **extension** of f_0 means that $f|_{[0,1]} \equiv f_0$.

(Q6.5.1.6.C) Write $\mathsf{L}_{\mathcal{T}}^T : C_{\mathrm{per}}^0([0,1]) \to \mathcal{P}_{2n}^T$ for the approximation scheme based on trigonometric interpolation in the node set $\mathcal{T} := \{t_0, t_1, \dots, t_{2n}\} \subset \mathbb{R}$, $n \in \mathbb{N}$.

Show that $(\mathsf{L}_{\mathcal{T}}^T f)(t) \in \mathbb{R}$ for all $t \in \mathbb{R}$, if f is 1-periodic and real-valued: $f(t) \in \mathbb{R}$ for all $t \in \mathbb{R}$.

(Q6.5.1.6.D) Let

- $\mathsf{L}_n : C^0([-1,1]) \to \mathcal{P}_n$ denote the family of approximation schemes spawned by Chebychev interpolation on $[-1,1]$.
- $\mathsf{T}_n : C^0([0,1]) \to \mathcal{P}_{2n}^T$ stand for the family of approximation schemes related to trigonometric interpolation in equidistant nodes $t_k := \frac{k}{2n+1}$, $k = 0, \dots, 2n$.

Express L_n in terms of T_n using

- the Chebychev polynomials $T_0, \dots, T_n$ as a basis of $\mathcal{P}_n$,
- the complex exponentials $t \mapsto \exp(-2\pi \imath k t)$, $k = -n, \dots, n$, as a basis of $\mathcal{P}_{2n}^T$.

△

6.5.2 Trigonometric Interpolation Error Estimates

Video tutorial for Section 6.5.2 "Trigonometric Interpolation Error Estimates": (14 minutes) Download link, tablet notes

From (6.5.1.5) use the notation T_n for trigonometric interpolation in the $2n+1$ equidistant nodes $t_k := \dfrac{k}{2n+1}$. Our focus will be on the asymptotic behavior of

$$\|f - \mathsf{T}_n f\|_{L^\infty([0,1[)} \quad \text{and} \quad \|f - \mathsf{T}_n f\|_{L^2([0,1[)} \quad \text{as} \quad n \to \infty ,$$

for functions $f : [0,1[\to \mathbb{C}$ with different smoothness properties. To begin with we report an empiric study.

EXPERIMENT 6.5.2.1 (Interpolation error: trigonometric interpolation) Now we study the asymptotic behavior of the error of equidistant trigonometric interpolation as $n \to \infty$ in a numerical experiment for functions with different smoothness properties.

#1 Step function: $f(t) = 0$ for $|t - \frac{1}{2}| > \frac{1}{4}$, $f(t) = 1$ for $|t - \frac{1}{2}| \leq \frac{1}{4}$

#2 C^∞ periodic function: $f(t) = \dfrac{1}{\sqrt{1 + \frac{1}{2}\sin(2\pi t)}}$.

#3 "wedge function": $f(t) = |t - \frac{1}{2}|$

Approximate computation of norms of interpolation errors on equidistant grid with 4096 points.

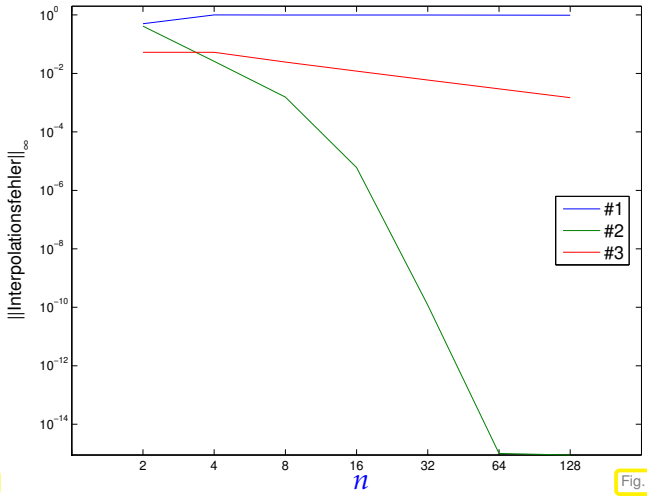


Fig. 247

Maximum norm of interpolation error

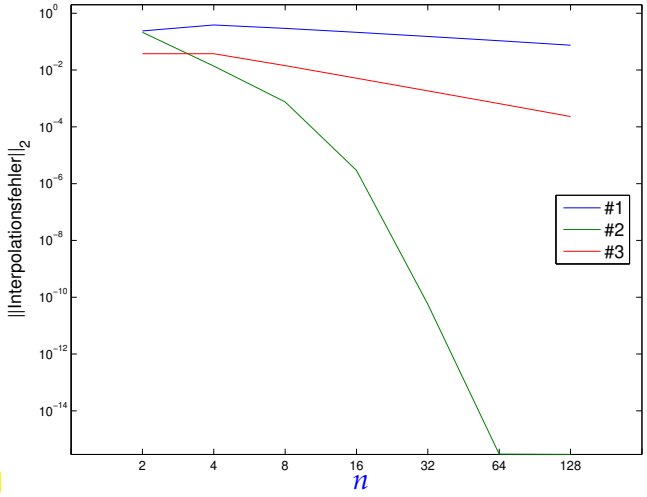


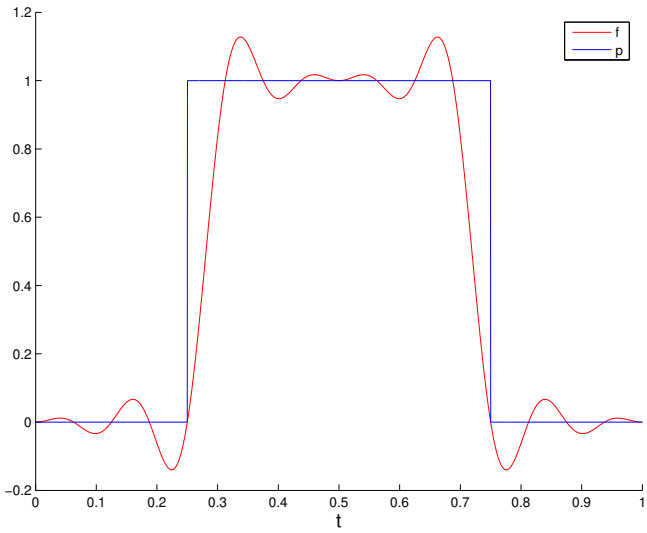
Fig. 248

$L^2([0,1])$-norm of interpolation error

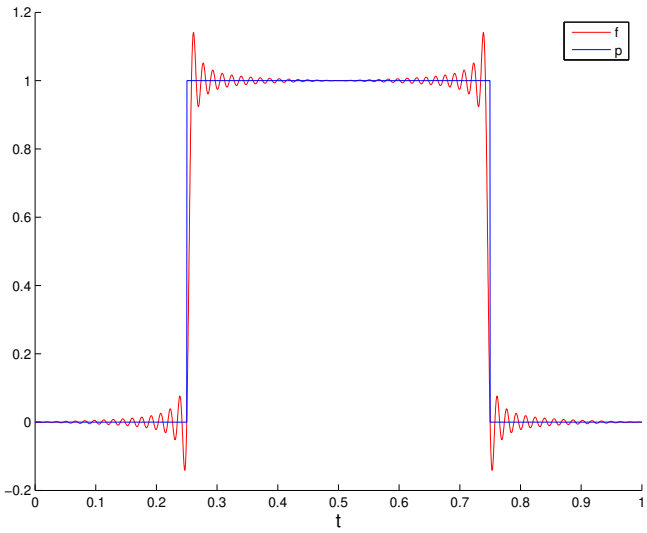
Observations: Function #1: no convergence in L^∞-norm, algebraic convergence in L^2-norm
Function #3: algebraic convergence in both norms
Function #2: exponential convergence in both norms

We conclude that in this experiment higher smoothness of f leads to faster convergence of the trigonometric interplant. ⌟

EXPERIMENT 6.5.2.2 (Gibbs phenomenon) Of course the smooth trigonometric interpolants of the step function must fail to converge in the L^∞-norm in Exp. 6.5.2.1. Moreover, they will not even converge "visually" to the step function, which becomes manifest by a closer inspection of the interpolants.



$n = 16$



$n = 128$

We observe massive "overshooting oscillations" in a neighborhood of the discontinuity. This is the notorious **Gibbs phenomenon** affecting approximation schemes relying on trigonometric polynomials. ⌟

§6.5.2.3 (Fourier series → (4.2.6.7)**)** From (6.5.0.1b) we know that the complex vector space space of trigonometric polynomials $\mathcal{P}_{2n}^T$ is spanned by the $2n+1$ Fourier modes $t \mapsto \exp(2\pi\imath kt)$ of "lowest frequency", that is, for $-n \le k \le n$

$$p \in \mathcal{P}_{2n}^T \quad \Rightarrow \quad p(t) = \sum_{k=-n}^{n} a_k \exp(-2\pi\imath kt) \quad \text{for some} \quad a_k \in \mathbb{C}\,. \tag{6.5.2.4}$$

Now let us make a connection: In Section 4.2.6 we learned that every function $f : [0,1[\to \mathbb{C}$ with finite $L^2([0,1])$-norm

$$\|f\|_{L^2([0,1])}^2 := \int_0^1 |f(t)|^2 \,\mathrm{d}t < \infty\,,$$

can be expanded in a **Fourier series** (→ Thm. 4.2.6.33)

$$f(t) = \sum_{k=-\infty}^{\infty} \widehat{f}_k \exp(-2\pi\imath kt) \quad \text{in } L^2([0,1]) \quad , \quad \widehat{f}_k := \int_0^1 f(t)\exp(2\pi\imath kt)\,\mathrm{d}t\,. \tag{6.5.2.5}$$

We add, that a limit in $L^2([0,1])$ means that $\left\| f - \sum_{k=-M}^{M} \widehat{f}_k \exp(-2\pi\imath k\cdot) \right\|_{L^2([0,1])} \to 0$ for $M \to \infty$. Also note the customary the notation $\widehat{f}_k$ for the **Fourier coefficients**.

Seeing (6.5.2.4) and (6.5.2.5) side-by-side and understanding that trigonometric polynomials are finite Fourier series suggests that we investigate the approximation of Fourier series by trigonometric interpolants.

Idea: Study trigonometric interpolation error for interpolands given in Fourier series representation (6.5.2.5)

⌟

Remark 6.5.2.6 ($L^2(]0,1[)$: Natural setting for trigonometric interpolation) A fundamental result about functions given through Fourier series was the following fundamental isometry property of the mapping taking a function to the sequence of its Fourier coefficients.

Theorem 4.2.6.33. Isometry property of the Fourier transform

If the Fourier coefficients satisfy $\sum_{k\in\mathbb{Z}} |\widehat{c}_j|^2 < \infty$, *then the Fourier series*

$$c(t) = \sum_{k\in\mathbb{Z}} \widehat{c}_k \exp(-2\pi\imath kt)$$

yields a function $c \in L^2([0,1])$ *that satisfies*

$$\|c\|^2_{L^2([0,1])} := \int_0^1 |c(t)|^2 \,\mathrm{d}t = \sum_{k\in\mathbb{Z}} |\widehat{c}_j|^2 \,.$$

This paves the way for estimating the $L^2([0,1])$-norm of interpolation/approximation errors, once we have information about the decay of their Fourier coefficients.

The $L^2([0,1])$ is also a highly relevant quantity in engineering application, when $t \mapsto c(t)$ is regarded as a time-dependent signal. In this case $\|c\|_{L^2([0,1])}$ is the root mean square (RMS) power of the signal. ⌟

§6.5.2.7 (Aliasing) Guided by the insights from § 6.5.2.3, we study the action of the trigonometric interpolation operator T_n from (6.5.1.5) on individual Fourier modes

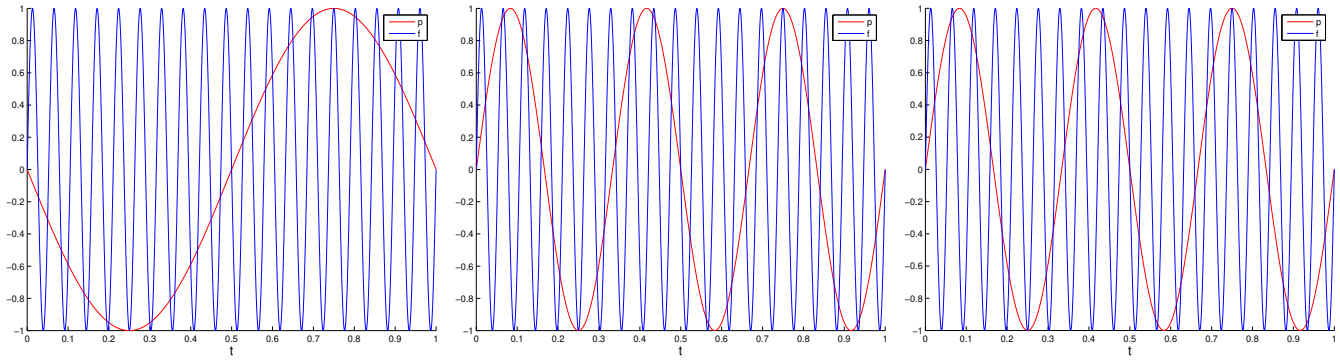
$$\mu_k(t) := \exp(-2\pi k\imath t)\,, \quad t \in \mathbb{R}\,, \quad k \in \mathbb{Z}\,.$$

Due to the 1-periodicity of $t \mapsto \exp(-2\pi\imath t)$ we find for every node $t_j := \frac{j}{2n+1}$, $j = 0,\ldots,2n$,

$$\mu_k(t_j) = \exp(-2\pi\imath k\tfrac{j}{2n+1}) = \exp(-2\pi\imath(k-\ell(2n+1))\tfrac{j}{2n+1}) = \mu_{k-\ell(2n+1)}(t_j) \quad \forall \ell \in \mathbb{Z}\,.$$

When sampled on the node set $\mathcal{T}_n := \{t_0,\ldots,t_{2n}\}$ all the Fourier modes $\mu_{k-\ell(2n+1)}$, $\ell \in \mathbb{Z}$, yield the same values. Thus trigonometric interpolation cannot distinguish them! This phenomenon is called aliasing.

Aliasing demonstrated for $f(t) = \sin(2\pi\cdot 19t) = \mathrm{Im}(\exp(2\pi\imath 19t))$ for different node sets.



$n = 2$ $n = 4$ $n = 8$

The "low-frequency" sine waves plotted in red coincide with f on the node set $\mathcal{T}_n$.

Since $\mathrm{T}_n\mu_k = \mu_k$ for $k = -n,\ldots,n$, that is, for $\mu_k \in \mathcal{P}^T_{2n}$, the aliasing effect yields

$$\mathrm{T}_n\mu_k = \mu_{\widetilde{k}} \quad , \quad \widetilde{k} \in \{-n,\ldots,n\}\,, \quad k-\widetilde{k} \in (2n+1)\mathbb{Z} \quad [\,\widetilde{k} := k \mod (2n+1)\,]\,. \tag{6.5.2.8}$$

For instance, we have $\widetilde{n} = n$, $\widetilde{n+1} = -n$, $\widetilde{-n} = -n$, $\widetilde{-n-1} = n$, $\widetilde{2n} = -1$, etc.

► Trigonometric interpolation by T_n maps all Fourier modes ("frequencies") to another single Fourier mode in the finite range $\{-n,\ldots,n\}$.

⌟

§6.5.2.9 (Fourier representation of trigonometric interpolation error) From (6.5.2.8), by linearity of T_n, we obtain for $f : [0,1[\to \mathbb{C}$ in Fourier series representation

$$f(t) = \sum_{j=-\infty}^{\infty} \widehat{f}_j \mu_j(t) \quad \blacktriangleright \quad \mathsf{T}_n(f)(t) = \sum_{j=-n}^{n} \gamma_j \mu_j(t)\,, \quad \gamma_j = \sum_{\ell=-\infty}^{\infty} \widehat{f}_{j+\ell(2n+1)}\,. \tag{6.5.2.10}$$

We can read the trigonometric polynomial $\mathsf{T}_n f \in \mathcal{P}_{2n}^T$ as a Fourier series with non-zero coefficients only in the index range $\{-n,\ldots,n\}$. Thus, for the Fourier coefficients $\widehat{E}_j$ of the trigonometric interpolation error $E(t) := f(t) - \mathsf{T}_n f(t)$ we find from (6.5.2.10)

$$\widehat{E}_j = \begin{cases} -\sum\limits_{\ell\in\mathbb{Z}\setminus\{0\}} \widehat{f}_{j+\ell(2n+1)} & \text{, if } j \in \{-n,\ldots,n\}\,, \\ \widehat{f}_j & \text{, if } |j| > n\,, \end{cases} \qquad j \in \mathbb{Z}\,. \tag{6.5.2.11}$$

Since we have (sufficient smoothness of f assumed)

$$E(t) := f(t) - \mathsf{T}_n f(t) = \sum_{j=-\infty}^{\infty} \widehat{E}_j e^{-2\pi\imath jt}\,, \tag{6.5.2.12}$$

we conclude from the isometry property asserted in Thm. 4.2.6.33 and the triangle inequality

$$\|f - \mathsf{T}_n f\|_{L^2(]0,1[)}^2 = \sum_{j=-n}^{n} \left| \sum_{\ell\in\mathbb{Z}\setminus\{0\}} \widehat{f}_{j+\ell(2n+1)} \right|^2 + \sum_{|j|>n} |\widehat{f}_j|^2\,, \tag{6.5.2.13}$$

$$\|f - \mathsf{T}_n f\|_{L^\infty(]0,1[)} \le \sum_{j=-n}^{n} \sum_{\ell\in\mathbb{Z}\setminus\{0\}} |\widehat{f}_{j+\ell(2n+1)}| + \sum_{|j|>n} |\widehat{f}_j|\,. \tag{6.5.2.14}$$

In order to estimate these norms of the trigonometric interpolation error we need quantitative information about the decay of the Fourier coefficients $\widehat{f}_j$ as $|j| \to \infty$. ⌟

§6.5.2.15 (Fourier expansions of derivatives) For 1-periodic $c \in C^0(\mathbb{R})$ with integrable derivative $\dot{c} := \frac{d}{dt}c$ we find by integration by parts (boundary terms cancel due to periodicity)

$$\widehat{\dot{c}}_j = \int_0^1 \dot{c}(t) e^{2\pi\imath jt}\,\mathrm{d}t = -2\pi\imath j \cdot \int_0^1 c(t) e^{2\pi\imath jt}\,\mathrm{d}t = (-2\pi\imath j)\,\widehat{c}_j\,, \quad j \in \mathbb{Z}\,.$$

We can also arrive at this formula by (formal) term-wise differentiation of the Fourier series:

$$c(t) = \sum_{j=-\infty}^{\infty} \widehat{c}_j e^{-2\pi\imath jt} \quad \Longrightarrow \quad \dot{c}(t) = \sum_{j=-\infty}^{\infty} \underbrace{(-2\pi\imath j)\widehat{c}_j}_{=\widehat{\dot{c}}_j} e^{-2\pi\imath jt}\,. \tag{6.5.2.16}$$

These considerations essentially provide a formal proof of the following result.

Lemma 6.5.2.17. Fourier coefficients of derivatives

For the Fourier coefficients of the derivatives a 1-periodic function $f \in C^{k-1}(\mathbb{R})$, $k \in \mathbb{N}$, with integrable k-th derivative $f^{(k)}$ holds

$$\widehat{(f^{(k)})}_j = (-2\pi\imath j)^k\,\widehat{f}_j\,, \quad j \in \mathbb{Z}\,.$$

⌟

§6.5.2.18 (Fourier coefficients and smoothness) From Lemma 6.5.2.17 and the trivial estimates ($|\exp(2\pi\imath t)| = 1$)

$$|\widehat{f}_j| \le \int_0^1 |f(t)|\,\mathrm{d}t \le \|f\|_{L^1(]0,1[)} \quad \forall j \in \mathbb{Z}\,, \tag{6.5.2.19}$$

we conclude that $\left((2\pi|j|)^m \widehat{f}_j\right)_{j\in\mathbb{Z}}$, $m \in \mathbb{N}$, is bounded, provided that $f \in C^{m-1}(\mathbb{R})$ with integrable m-th derivative

Lemma 6.5.2.20. Decay of Fourier coefficients

If $f \in C^{k-1}(\mathbb{R})$ with integrable k-th derivative, then $\widehat{f}_j = O(|j|^{-k})$ for $|j| \to \infty$

Decay of Fourier coefficients and smoothness

The smoother a periodic function the faster the decay of its Fourier coefficients

The isometry property of Thm. 4.2.6.33 also yields for $f \in C^{k-1}(\mathbb{R})$ with $f^{(k)} \in L^2(]0,1[)$ that

$$\left\|f^{(k)}\right\|^2_{L^2(]0,1[)} = (2\pi)^{2k} \sum_{j=-\infty}^{\infty} |j|^{2k} |\widehat{f}_j|^2\,. \tag{6.5.2.22}$$

⌟

We can now combine the identity (6.5.2.22) with (6.5.2.13) and obtain an interpolation error estimate in $L^2(]0,1[)$-norm.

Theorem 6.5.2.23. Finite-smoothness L^2-error estimate for trigonometric interpolation

If $f \in C^{k-1}(\mathbb{R})$, $k \in \mathbb{N}$, with square-integrable k-th derivative ($f^{(k)} \in L^2(]0,1[)$), then

$$\|f - \mathsf{T}_n f\|_{L^2(]0,1[)} \le \sqrt{1+c_k}\,(2\pi n)^{-k} \left\|f^{(k)}\right\|_{L^2(]0,1[)}\,, \tag{6.5.2.24}$$

with $c_k = 2\sum_{\ell=1}^{\infty}(2\ell-1)^{-2k} < \infty$.

Proof. From Thm. 4.2.6.33 and Lemma 6.5.2.17 we infer

$$\sum_{j\in\mathbb{Z}} |2\pi j|^{2k} |\widehat{f}_j|^2 = \left\|f^{(k)}\right\|^2_{L^2([0,1])}\,.$$

As a tool we will need the Cauchy-Schwarz inequality for quadratically convergent sequences:

$$\left|\sum_{\ell=1}^{\infty} a_\ell b_\ell\right|^2 \le \sum_{\ell=1}^{\infty} |a_\ell|^2 \cdot \sum_{\ell=1}^{\infty} |b_\ell|^2 \quad \forall (a_\ell), (b_\ell) \in \mathbb{C}^{\mathbb{N}}\,. \tag{6.5.2.25}$$

We start from

$$\|f - \mathsf{T}_n f\|^2_{L^2(]0,1[)} = \sum_{j=-n}^{n} \left|\sum_{\ell\in\mathbb{Z}\setminus\{0\}} \widehat{f}_{j+\ell(2n+1)}\right|^2 + \sum_{|j|>n} |\widehat{f}_j|^2\,, \tag{6.5.2.13}$$

and then use (6.5.2.25),

$$\begin{aligned}
\|f - \mathsf{T}_n f\|^2_{L^2(]0,1[)} &= \sum_{j=-n}^{n} \left| \sum_{|\ell|\geq 1} \widehat{f}_{j+\ell(2n+1)} \right|^2 + \sum_{|j|>n} |\widehat{f}_j|^2 \\
&= \sum_{j=-n}^{n} \left| \sum_{|\ell|\geq 1} |2\pi(j+\ell(2n+1))|^{-k} \left(|2\pi(j+\ell(2n+1))|^k \widehat{f}_{j+\ell(2n+1)} \right) \right|^2 + \\
&\quad \sum_{|j|>n} |2\pi j|^{-2k} |2\pi j|^{2k} |\widehat{f}_j|^2 \\
&\leq \sum_{j=-n}^{n} \left\{ \sum_{|\ell|\geq 1} |2\pi(j+\ell(2n+1))|^{-2k} \cdot \sum_{|\ell|\geq 1} |2\pi(j+\ell(2n+1))|^{2k} |\widehat{f}_{j+\ell(2n+1)}|^2 \right\} + \\
&\quad \sum_{|j|>n} |2\pi j|^{-2k} |2\pi j|^{2k} |\widehat{f}_j|^2 \, .
\end{aligned}$$

We plug in the estimate

$$\sum_{|\ell|\geq 1} |2\pi(j+\ell(2n+1))|^{-2k} \leq 2 \sum_{\ell\geq 1} |2\pi(\ell(2n+1)-n)|^{-2k} \leq (2\pi n)^{-2k} c_k \, ,$$

which yields

$$\begin{aligned}
\|f - \mathsf{T}_n f\|^2_{L^2(]0,1[)} &\leq c_k \, (2\pi n)^{-2k} \sum_{j=-n}^{n} \sum_{|\ell|\geq 1} |2\pi(j+\ell(2n+1))|^{2k} |\widehat{f}_{j+\ell(2n+1)}|^2 + \\
&\quad (2\pi n)^{-2n} \sum_{|j|>n} |2\pi j|^{2k} |\widehat{f}_j|^2 \\
&\leq (1+c_k)(2\pi n)^{-2k} \left\| f^{(k)} \right\|^2_{L^2([0,1])} \, .
\end{aligned}$$

□

Thm. 6.5.2.23 confirms *algebraic convergence* of the L^2-norm of the trigonometric interpolation error for functions with *limited smoothness*. Higher rates can be expected for smoother functions, which we have also found in cases #1 and #3 in Exp. 6.5.2.1.

Review question(s) 6.5.2.26 (Trigonometric Interpolation Error Estimates)

(Q6.5.2.26.A) We know that trigonometric interpolation can be regarded as standard polynomial interpolation with nodes located on the unit circle $\mathbb{S}^1 \subset \mathbb{C}$.

Nevertheless, it is not possible to apply the following theorem to obtain error estimates for trigonometric interpolation. Why?

Theorem 6.2.2.15. Representation of interpolation error

We consider $f \in C^{n+1}(I)$ and the Lagrangian interpolation approximation scheme (→ Def. 6.2.2.1) for a node set $\mathcal{T} := \{t_0, \ldots, t_n\} \subset I$. Then,

for every $t \in I$ there exists a $\tau_t \in]\min\{t, t_0, \ldots, t_n\}, \max\{t, t_0, \ldots, t_n\}[$ such that

$$f(t) - \mathsf{L}_{\mathcal{T}}(f)(t) = \frac{f^{(n+1)}(\tau_t)}{(n+1)!} \cdot \prod_{j=0}^{n} (t - t_j) \, . \tag{6.5.2.27}$$

△

6.5.3 Trigonometric Interpolation of Analytic Periodic Functions

Video tutorial for Section 6.5.3 "Trigonometric Interpolation of Analytic Periodic Functions": (16 minutes) Download link, tablet notes

In Section 6.2.2.3 we saw that we can expect exponential decay of the maximum norm of polynomial interpolation errors in the case of "very smooth" interpolands. To capture this property of functions we resorted to the notion of analytic functions, as defined in Def. 6.2.2.48. Since trigonometric interpolation is closely connected to polynomial interpolation (on the unit circle $\mathbb{S}^1$, see Section 5.6.2), it is not surprising that analyticity of interpolands will also involve exponential convergence of trigonometric interpolants. This result will be established in this section.

In case #2 of Exp. 6.5.2.1 we already say an instance of exponential convergence for an analytic interpoland. A more detailed study follows.

EXPERIMENT 6.5.3.1 (Trigonometric interpolation of 1-periodic analytic functions)

We study the convergence of equidistant trigonometric interpolation for the 1-periodic interpoland tr

$$f(t) = \frac{1}{\sqrt{1-\alpha\sin(2\pi t)}} \quad \text{on } I = [0,1]\,. \tag{6.5.3.2}$$

For $0 \le \alpha < 1$ we have $f \in C^\infty(\mathbb{R})$, and f is even analytic ($\rightarrow$ Def. 6.2.2.48).

Approximative computations of error norms by "oversampling" in 4096 points. ▷

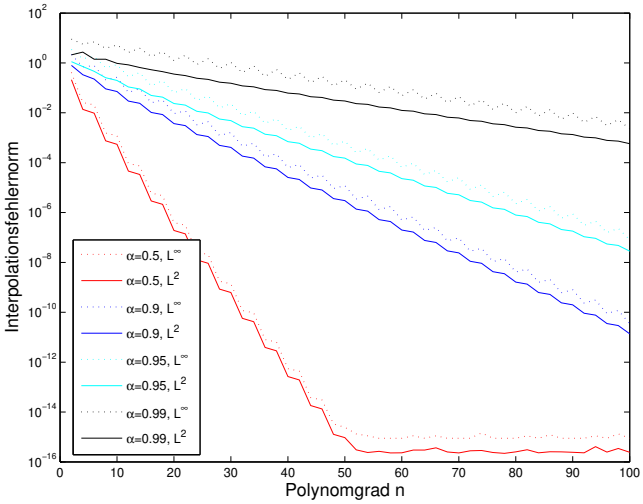


Fig. 249

➢ Observation: exponential convergence in n, faster for smaller α ⌟

§6.5.3.3 (Analytic periodic functions) Assume that a 1-periodic function $f : \mathbb{R} \to \mathbb{R}$ possesses an analytic extension to an open domain $D \subset \mathbb{C}$ beyond the interval $[0,1]$: $[0,1] \subset D$.

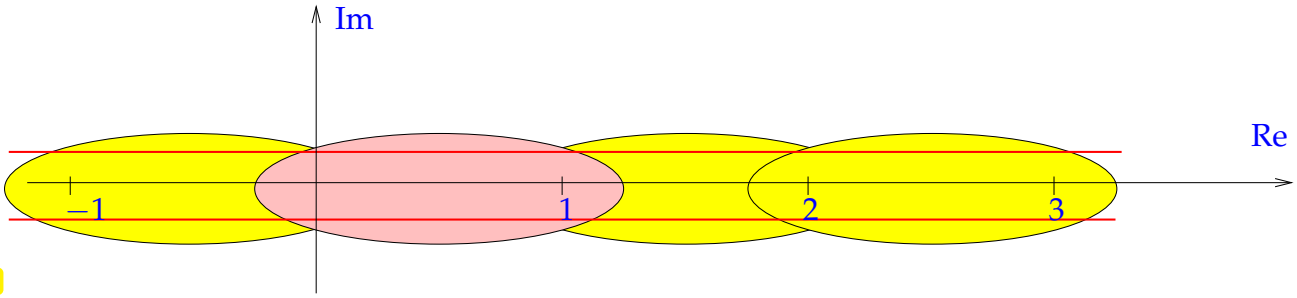


Fig. 250

Then, thanks to 1-peridicity f will also have an analytic extension to $D_{\rm per} := \bigcup_{k\in\mathbb{Z}} (D+k)$. That domain will contain a strip parallel to the real axis, see Fig. 250:

$$\exists \eta^- < 0 < \eta^+ : \quad \{z \in \mathbb{C} : \ \eta^- \le \mathrm{Im}(z) \le \eta^+\} \subset D_{\rm per}\,.$$

⌟

§6.5.3.4 (Decay of Fourier coefficients of analytic functions) Lemma 6.5.2.20 asserts algebraic decay of the Fourier coefficients of functions with limited smoothness. As analytic 1-periodic functions are

"infinitely smooth", the will always belong to $C^\infty(\mathbb{R})$, we expect a stronger result in this case. In fact, we can conclude *exponential decay* of the Fourier coefficients.

Theorem 6.5.3.5. Exponential decay of Fourier coefficients of analytic functions

If $f : \mathbb{R} \to \mathbb{C}$ is 1-periodic and has an analytic extension to the strip

$$\bar{S} := \{z \in \mathbb{C}: \ -\eta \le \operatorname{Im} z \le \eta\} \ , \quad \text{for some } \eta > 0 \ ,$$

then its Fourier coefficients decay according to

$$|\widehat{f}_j| \le q^{|k|} \cdot \|f\|_{L^\infty(\bar{S})} \quad \forall k \in \mathbb{Z} \quad \text{with} \quad q := \exp(-2\pi\eta) \in]0,1[\ . \tag{6.5.3.6}$$

Proof. ❶: A first variant of the proof uses techniques from complex analysis:

Let $f : \mathbb{R} \mapsto \mathbb{C}$ be 1-periodic with analytic extension to the (closed) strip

$$S := \{z \in \mathbb{C}: -\eta \le \operatorname{Im}\{z\} \le \eta\}, \ \eta > 0 \ .$$

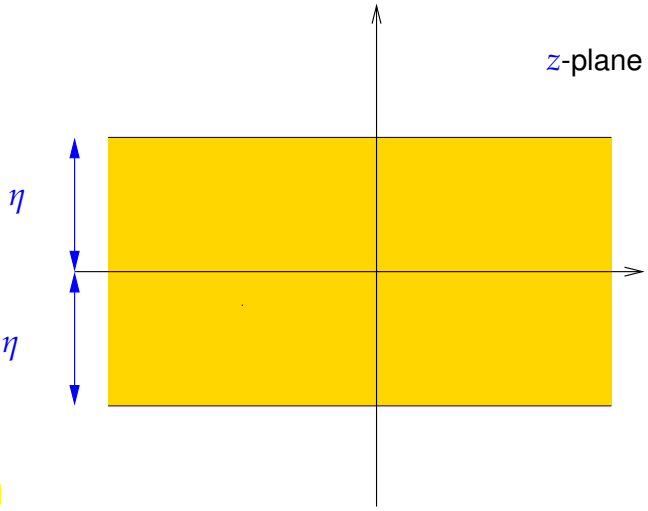


Fig. 251

We recall the fundamental Cauchy integral theorem from complex analysis.

Theorem 6.5.3.7. Cauchy integral theorem [Rem02, Satz 7.1.2]

Let $f : D \to \mathbb{C}$ be analytic in $D \subset \mathbb{C}$ and $U \subset D$ be simply connected and strictly contained in D. Then

$$\int_{\partial U} f(z)\, dz = 0 \ .$$

We apply this theorem to either rectangle

$$U^+ := \{z = \xi + \imath\eta, \ 0 \le \xi \le 1, 0 \le \eta \le r\} \quad \text{for} \quad k \ge 0 \ ,$$
$$U^- := \{z = \xi + \imath\eta, \ 0 \le \xi \le 1, 0 \le \eta \le r\} \quad \text{for} \quad k < 0 \ ,$$

and note that the contributions of the sides parallel to the imaginary axis cancel thanks to 1-periodicity (and their opposite orientation).

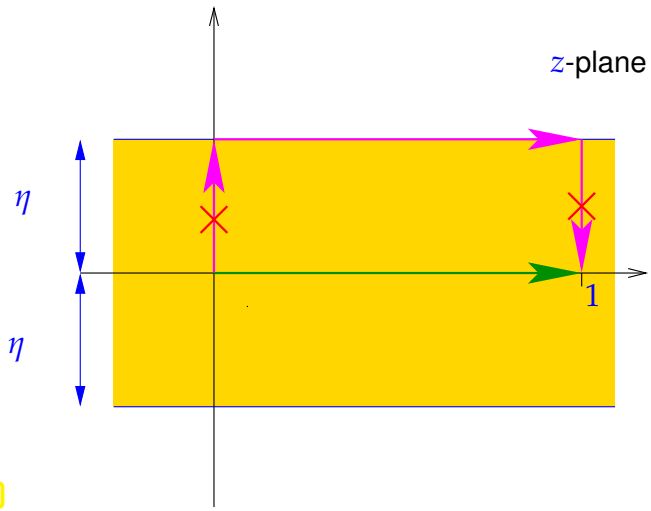


Fig. 252

◁ Equivalent integration paths for the computation of Fourier coefficients $\widehat{f}_k$, $k > 0$.

⟶ integration path $0 \to 1$ on the real line.

⟶ integration path through the upper half complex plane. Contributions of sections parallel to the imaginary axis cancel.

Thus we compute the Fourier coefficients $\widehat{f}_k$ of f by a different integral. For $k > 0$ we get

$$\widehat{f}_k = \int_0^1 f(t)e^{2\pi\imath kt}\,\mathrm{d}t = \int_{\partial U^+\setminus\mathbb{R}} f(z)e^{2\pi\imath kz}\,\mathrm{d}z = \int_0^1 f(s+\imath\eta)e^{2\pi\imath k(s+\imath\eta)}\,\mathrm{d}s = e^{-2\pi\eta k}\int_0^1 f(s+\imath\eta)e^{2\pi\imath ks}\,\mathrm{d}s\,,$$

which leads to the estimate

$$|\widehat{f}_k| \le e^{-2\pi\eta k}\max\{|f(t+\imath\eta)|, 0 \le t \le 1\}\,.$$

Analogously, for $k < 1$ we compute

$$\widehat{f}_k = \int_0^1 f(t)e^{2\pi\imath kt}\,\mathrm{d}t = \int_{\partial U^-\setminus\mathbb{R}} f(z)e^{2\pi\imath kz}\,\mathrm{d}z = \int_0^1 f(s-\imath\eta)e^{2\pi\imath k(s-\imath\eta)}\,\mathrm{d}s = e^{2\pi\eta k}\int_0^1 f(s+\imath\eta)e^{2\pi\imath ks}\,\mathrm{d}s\,,$$

and obtain the bound

$$|\widehat{f}_k| \le e^{2\pi\eta k}\max\{|f(t-\imath\eta)|, 0 \le t \le 1\}\,.$$

The assertion of the theorem follows directly.

❷: A second variant of the proof merely relies on classical calculus.

By the assumptions of the theorem the function $f : \mathbb{R} \to \mathbb{C}$ can be expanded into a power series with radius of convergence η around at every $y \in \mathbb{R}$

$$\forall y \in \mathbb{R}\colon \exists c_n(y) \in \mathbb{C}\colon \quad f(x) = \sum_{n=0}^{\infty} c_n(y)(x-y)^n \quad \forall x : |x-y| \le \eta\,. \tag{6.5.3.8}$$

This implies $|c_n(y)|\eta^n \le C(y)$ and, since $[0,1]$ is compact and $y \mapsto C(y)$ is continuous,

$$\exists C > 0\colon \quad |c_n(y)|\eta^n \le C \quad \forall n \in \mathbb{N}_0,\ \forall y \in \mathbb{R}\,. \tag{6.5.3.9}$$

The power series (6.5.3.8) is a Taylor series, which means

$$c_n(y) = \frac{1}{n!}f^{(n)}(y) \overset{(6.5.3.9)}{\Longrightarrow} |f^{(n)}(y)| \le Cn!\eta^{-n} \quad \forall n \in \mathbb{N}_0\,, \quad y \in \mathbb{R}\,. \tag{6.5.3.10}$$

We use this estimate to bound the Fourier coefficients of f. Starting from the defining formula (4.2.6.20), we continue with n-fold integration by parts

$$\widehat{f}_k := \int_0^1 f(t)\exp(2\pi\imath kt)\,\mathrm{d}t = \frac{(-1)^n}{(2\pi\imath k)^n}\int_0^1 f^{(n)}(t)\exp(2\pi\imath kt)\,\mathrm{d}t\,. \tag{6.5.3.11}$$

We combine this formula with (6.5.3.10) and obtain

$$|\widehat{f}_k| \leq C\frac{n!}{(2\pi|k|\eta)^n} \quad \forall n \in \mathbb{N},\ k \in \mathbb{Z}\,. \tag{6.5.3.12}$$

Next, we use Stirling's formula (6.2.1.12) in the form

$$n! \leq en^{n+1/2}e^{-n}\,, \quad n \in \mathbb{N}\,,$$

which gives

$$|\widehat{f}_k| \leq Ce\frac{n^{n+1/2}}{(2\pi|k|\eta)^n}e^{-n} \quad \forall n \in \mathbb{N},\ k \in \mathbb{Z}\,.$$

We can also "interpolate" and replace n with a real number.

$$\blacktriangleright \quad |\widehat{f}_k| \leq Ce\frac{r^{r+1/2}}{(2\pi|k|\eta)^r}e^{-r} \quad \forall r \geq 1,\ k \in \mathbb{Z}\,.$$

Finally, we set $r := 2\pi|k|\eta$ and arrive at

$$|\widehat{f}_k| \leq Ce\sqrt{2\pi k\eta}\underbrace{\exp(-2\pi\eta)}_{=:q}^{|k|}\,, \quad k \in \mathbb{Z}\,.$$

This bound is slightly less explicit than that obtain in ❶. □ ⌟

Knowing exponential decay of the Fourier coefficients, the geometric sum formula can be used to extract estimates for the trigonometric interpolation operator T_n (for equispaced nodes) from (6.5.2.13) and (6.5.2.14):

Lemma 6.5.3.13. Interpolation error estimates for exponentially decaying Fourier coefficients

If $f : \mathbb{R} \to \mathbb{C}$ is 1-periodic and $f \in L^2(]0,1[)$, then

$$\exists M > 0, \rho \in]0,1[:\ |\widehat{f}(k)| \leq M\rho^{|k|} \quad \Rightarrow \quad \begin{aligned} \|f - \mathsf{T}_n(f)\|_{L^2(]0,1[)} &\leq M\frac{\rho^{n/2}}{1-\rho^n}\frac{2\sqrt{2}}{\sqrt{1-\rho^2}}\,, \\ \|f - \mathsf{T}_n(f)\|_{L^\infty(]0,1[)} &\leq 4M\frac{\rho^{n/2}}{1-\rho}\,. \end{aligned}$$

This estimate can be combined with the result of Thm. 6.5.3.5 and gives the main result of this section:

Theorem 6.5.3.14. Exponential convergence of trigonometric interpolation for analytic interpolands

If $f : \mathbb{R} \to \mathbb{C}$ is 1-periodic and possesses an analytic extension to the strip

$$\bar{S} := \{z \in \mathbb{C}:\ -\eta \leq \operatorname{Im} z \leq \eta\}\,, \quad \text{for some } \eta > 0\,,$$

then there is $C_\eta > 0$ depending only on η such that

$$\|f - \mathsf{T}_n f\|_* \leq C_\eta e^{-\pi\eta n}\|f\|_{L^\infty(\bar{S})}\,, \quad n \in \mathbb{N}\,, \quad (\,* = L^2(]0,1[), L^\infty(]0,1[)\,)\,. \tag{6.5.3.15}$$

The speed of exponential convergence clearly depends on the width η of the "strip of analyticity" $\bar{S}$.

§6.5.3.16 (Convergence of trigonometric interpolation for analytic interpolands) We can now give a precise explanation of the observations made in Exp. 6.5.3.1, *cf.* Rem. 6.2.3.26. Similar to Chebychev interpolants, also trigonometric interpolants converge exponentially fast, if the interpoland f is 1-periodic *analytic* ($\rightarrow$ Def. 6.2.2.48) in a strip around the real axis in $\mathbb{C}$, see Thm. 6.5.3.14 for details.

Thus we have to determine the maximal open subset D of $\mathbb{C}$ to which the function

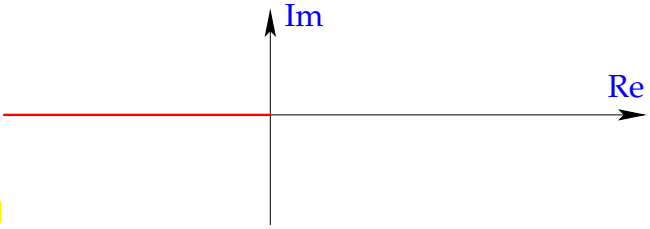
$$f(t) = \frac{1}{\sqrt{1-\alpha\sin(2\pi t)}}\,, \quad t \in [0,1]\,, \quad 0 < \alpha < 1\,, \tag{6.5.3.2}$$

possesses an analytic extension. Usually it is easier to determine the complement $P := \mathbb{C} \setminus D$, the "domain of singularity". We start with a result from complex analysis.

Lemma 6.5.3.17. Principal branch of the square root

The square root function $t \mapsto \sqrt{t}$, $t \geq 0$, can be extended to an analytic function on

$$\mathbb{C} \setminus \mathbb{R}_0^- := \{z \in \mathbb{C} : \ \mathrm{Re}(z) > 0 \quad \textit{or} \quad \mathrm{Im}(z) \neq 0\}\,.$$



◁ domain of singularity of $z \mapsto \sqrt{z}$, principal branch

Fig. 253

As a consequence of this and of

Theorem 6.2.2.68. Composition and products of analytic functions

If $f, h : D \subset \mathbb{C} \to \mathbb{C}$ and $g : U \subset \mathbb{C} \to \mathbb{C}$ are analytic in the open sets D and U, respectively, then
(i) the composition $f \circ g$ is analytic in $\{z \in U : \ g(z) \in D\}$,
(ii) the product $f \cdot h$ is analytic on D.

we find for the domain of singularity of f

$$P = \left\{z \in \mathbb{C} : \ 1 - \alpha\sin(2\pi z) \in \mathbb{R}_0^-\right\}\,.$$

Based on the identities

$$\sin(\imath y) = \frac{1}{2\imath}(\exp(-y) - \exp(y)) = \imath\sinh(y)\,,$$
$$\cos(\imath y) = \frac{1}{2}(\exp(-y) + \exp(y)) = \cosh(y)$$

for $y \in \mathbb{R}$, and using addition theorems for trigonometric functions, we find with $z = x + \imath y$, $x, y \in \mathbb{R}$,

$$\begin{aligned} & 1 + \alpha\sin(2\pi z) \in \mathbb{R}_0^- \\ & \Updownarrow \\ & \sin(2\pi z) = \sin(2\pi x)\cosh(2\pi y) + \imath\cos(2\pi x)\sinh(2\pi y) \in]-\infty, -1 - \tfrac{1}{\alpha}] \\ & \Updownarrow \\ & \sin(2\pi x)\cosh(2\pi y) \leq -1 - \tfrac{1}{\alpha} \quad \text{and} \quad \cos(2\pi x)\sinh(2\pi y) = 0\,. \end{aligned}$$

Note that $y = 0$ is not possible, because this would imply $|\sin(2\pi z)| \leq 1$. Hence, the term $x \mapsto \cos(2\pi x)$ must make the imaginary part vanish, which means

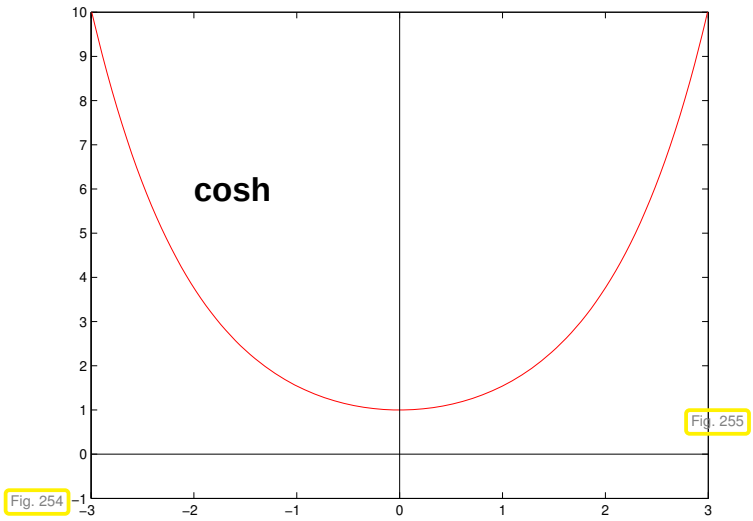
$$2\pi x \in (2\mathbb{Z}+1)\frac{\pi}{2} \quad \Leftrightarrow \quad x \in \tfrac{1}{2}\mathbb{Z} + \tfrac{1}{4}\,.$$

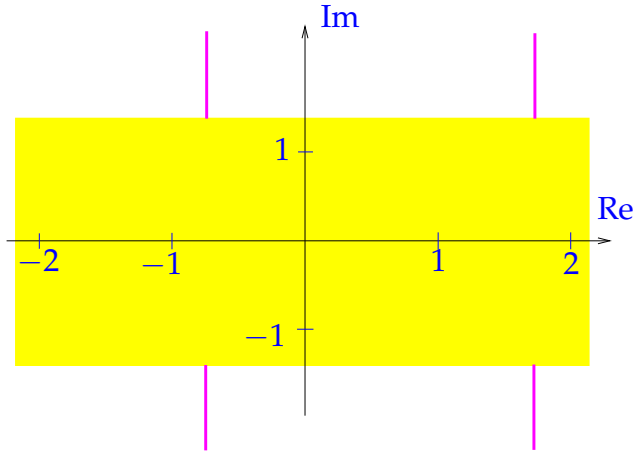
Thus we have $\sin(2\pi x) = \pm 1$. As $\cosh(2\pi y) > 0$, the sine has be negative, which leaves as only remaining choices for the real part

$$x \in \mathbb{Z} + \tfrac{3}{4} \quad \Leftrightarrow \quad \sin(2\pi x) = -1 \,.$$

As $\xi \mapsto \cosh(\xi)$ is a positive even function, we find the following domain of analyticity of f:

$$\mathbb{C} \setminus \bigcup_{k\in\mathbb{Z}} (k + \tfrac{3}{4} + i(\mathbb{R}\setminus]-\zeta,\zeta[)) \,, \quad \zeta \in \mathbb{R}^+ \,, \; \cosh(2\pi\zeta) = 1 + \tfrac{1}{\alpha} \,.$$





➢ f analytic in strip $S := \{z \in \mathbb{C} : : -\zeta < \mathrm{Im}(z) < \zeta\}$.

➢ As α decreases the strip of analyticity becomes wider, since $x \to \cosh(x)$ is increasing for $x > 0$. ⌟

Review question(s) 6.5.3.18 (Trigonometric interpolation of analytic periodic functions)

(Q6.5.3.18.A) From complex analysis we know the following result.

Theorem 6.5.3.19. Series of analytic functions

Assume that all the functions $f_k : D \to \mathbb{C}$, $k \in \mathbb{N}$, are analytic on the open set $D \subset \mathbb{C}$ and

$$\lim_{n\to\infty} \sup_{z\in D} \sum_{k=n}^{\infty} |f_k(z)| = 0 \,.$$

Then the series

$$F(z) := \sum_{k=1}^{\infty} f_k(z)$$

converges for all $z \in D$ and defines an analytic function $F : D \to \mathbb{C}$.

Show that

$$f(t) := \sum_{k\in\mathbb{Z}} a_k \exp(2\pi\imath t) \,, \quad t \in \mathbb{R} \,,$$

can be extended to a 1-periodic *analytic* function defined on a $\mathbb{C}$-neighborhood of $\mathbb{R}$, if

$$\exists C > 0,\ 0 \le q < 1 : \quad |a_k| \le C q^{|k|} \quad \forall k \in \mathbb{Z} \,.$$

(Q6.5.3.18.B) Determine the maximal open set $D \subset \mathbb{C}$, to which the function

$$f(t) = \frac{1}{1 + \sin^2(2\pi t)} \,, \quad t \in \mathbb{R} \,,$$

can be extended analytically. What is the maximal width of a strip

$$S := \{z \in \mathbb{C} : |\operatorname{Im}(z)| < \eta\} \quad \text{for some} \quad \eta > 0\,,$$

such that f has an analytic extension to S.

△

6.6 Approximation by Piecewise Polynomials

Recall some alternatives to interpolation by global polynomials discussed in Chapter 5:

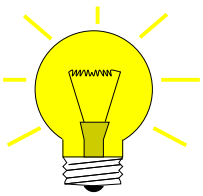
- piecewise linear/quadratic interpolation → Section 5.3.2, Ex. 5.3.2.4,
- cubic Hermite interpolation → Section 5.3.3,
- (cubic) spline interpolation → Section 5.4.2.

☞ All these interpolation schemes rely on piecewise polynomials (of different global smoothness)

Focus in this section: function approximation by *piecewise polynomial* interpolants

§6.6.0.1 (Grid/mesh) The attribute "piecewise" refers to *partitioning of the interval* on which we aim to approximate. In the case of data interpolation the natural choice was to use intervals defined by interpolation nodes. Yet we already saw exceptions in the case of shape-preserving interpolation by means of quadratic splines, see Section 5.4.4.

In the case of function approximation based on an interpolation scheme the additional freedom to choose the interpolation nodes suggests that those be decoupled from the partitioning.



Idea: use piecewise polynomials with respect to a grid/mesh

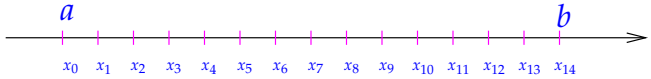
$$\mathcal{M} := \{a = x_0 < x_1 < \ldots < x_{m-1} < x_m = b\} \tag{6.6.0.2}$$

to approximate function $f : [a,b] \mapsto \mathbb{R}$, $a < b$.

Borrowing from terminology for splines, *cf.* Def. 5.4.1.1, the underlying mesh for piecewise polynomial approximation is sometimes called the "knot set".

Terminology:

- $x_j \mathrel{\hat{=}}$ nodes of the mesh $\mathcal{M}$,
- $[x_{j-1}, x_j[\mathrel{\hat{=}}$ intervals/cells of the mesh,
- $h_{\mathcal{M}} := \max_j |x_j - x_{j-1}| \mathrel{\hat{=}}$ mesh width,
- If $x_j = a + jh \mathrel{\hat{=}}$ equidistant (uniform) mesh with meshwidth $h > 0$



⌟

Remark 6.6.0.3 (Local approximation by piecewise polynomials) We will see that most approximation schemes relying on piecewise polynomials are local in the sense that finding the approximant on a cell of the mesh relies only on a fixed number of function evaluations in a neighborhood of the cell.

➢ $O(1)$ computational effort to find interpolant on $[x_{j-1}, x_j]$ (independent of m)

➢ $O(m)$ computational effort to determine piecewise polynomial approximant for $m \to \infty$ ("fine meshes")

Contrast this with the computational cost of computing global polynomial interpolants, which will usually be $O(n^2)$ for polynomial degree $n \to \infty$. ⌟

6.6.1 Piecewise Polynomial Lagrange Interpolation

Video tutorial for Section 6.6.1 "Piecewise Polynomial Lagrange Interpolation": (17 minutes)
Download link, tablet notes

Given: interval $[a,b] \subset \mathbb{R}$ endowed with mesh $\mathcal{M} := \{a = x_0 < x_1 < \ldots < x_{m-1} < x_m = b\}$.

Recall theory of polynomial interpolation → Section 5.2.2: $n+1$ data points needed to fix interpolating polynomial, see Thm. 5.2.2.7.

Approach to *local* Lagrange interpolation (→ (5.2.2.2)) of $f \in C(I)$ on mesh $\mathcal{M}$

General local Lagrange interpolation on a mesh (PPLIP)

❶ Choose local degree $n_j \in \mathbb{N}_0$ for each cell of the mesh, $j = 1,\ldots,m$.
❷ Choose set of *local* interpolation nodes

$$\mathcal{T}^j := \{t_0^j,\ldots,t_{n_j}^j\} \subset I_j := [x_{j-1},x_j]\,, \quad j = 1,\ldots,m\,,$$

for each mesh cell/grid interval I_j.
❸ Define piecewise polynomial (PP) interpolant $s : [x_0,x_m] \to \mathbb{K}$:

$$s_j := s_{|I_j} \in \mathcal{P}_{n_j} \quad \text{and} \quad s_j(t_i^j) = f(t_i^j) \quad i = 0,\ldots,n_j\,, \quad j = 1,\ldots,m\,. \tag{6.6.1.2}$$

Owing to Thm. 5.2.2.7, s_j is well defined.

Corollary 6.6.1.3. Piecewise polynomials Lagrange interpolation operator

The mapping $f \mapsto s$ defines a linear operator $\mathsf{I}_{\mathcal{M}} : C^0([a,b]) \mapsto C^0_{\mathcal{M},\mathrm{pw}}([a,b])$ in the sense of Def. 5.1.0.25.

Obviously, $\mathsf{I}_{\mathcal{M}}$ depends on $\mathcal{M}$, the local degrees n_j, and the sets $\mathcal{T}_j$ of local interpolation points (the latter two are suppressed in notation).

Corollary 6.6.1.4. Continuous local Lagrange interpolants

If the local degrees n_j are at least 1 and the local interpolation nodes t_k^j, $j = 1,\ldots,m$, $k = 0,\ldots,n_j$, for local Lagrange interpolation satisfy

$$t_{n_j}^j = t_0^{j+1} \quad \forall j = 1,\ldots,m-1 \quad \Rightarrow \quad s \in C^0([a,b])\,, \tag{6.6.1.5}$$

then the piecewise polynomial Lagrange interpolant according to (6.6.1.2) *is continuous on $[a,b]$: $s \in C^0([a,b])$.*

Focus: asymptotic behavior of (some norm of) interpolation error

$$\|f - \mathrm{I}_{\mathcal{M}} f\| \leq CT(N) \quad \text{for} \quad N \to \infty\,, \tag{6.6.1.6}$$

where $N := \sum_{j=1}^{m} (n_j + 1)$.

The decay of the bound $T(N)$ will characterize the type of convergence:

☛ algebraic convergence or exponential convergence, see Section 6.2.2, Def. 6.2.2.7.

But why do we choose this strange number N as parameter when investigating the approximation error?

Because, by Thm. 5.2.1.2, it agrees with the dimension of the space of discontinuous, piecewise polynomials functions

$$\{q : [a,b] \to \mathbb{R}\colon q_{|I_j} \in \mathcal{P}_{n_j}\ \forall j = 1,\ldots,m\} \quad !$$

This dimension tells us the number of real parameters we need to describe the interpolant s, that is, the "information cost" of s. N is also proportional to the number of interpolation conditions, which agrees with the number of f-evaluations needed to compute s (why only proportional in general?).

Special case: uniform polynomial degree $n_j = n$ for all $j = 1,\ldots,m$.

► Then we may aim for estimates $\|f - \mathrm{I}_{\mathcal{M}} f\| \leq CT(h_{\mathcal{M}})$ for $h_{\mathcal{M}} \to 0$

in terms of meshwidth $h_{\mathcal{M}}$.

Terminology: investigations of this kind are called the study of h-convergence.

EXAMPLE 6.6.1.7 (h-convergence of piecewise polynomial interpolation)

Compare Exp. 5.3.1.6:

$f(t) = \arctan t$, $I = [-5,5]$

Grid $\mathcal{M} := \{-5, -\frac{5}{2}, 0, \frac{5}{2}, 5\}$

Local interpolation nodes equidistant in I_j, endpoints included, (6.6.1.5) satisfied.

Plots of the piecewise linear, quadratic and cubic polynomial interpolants ▷

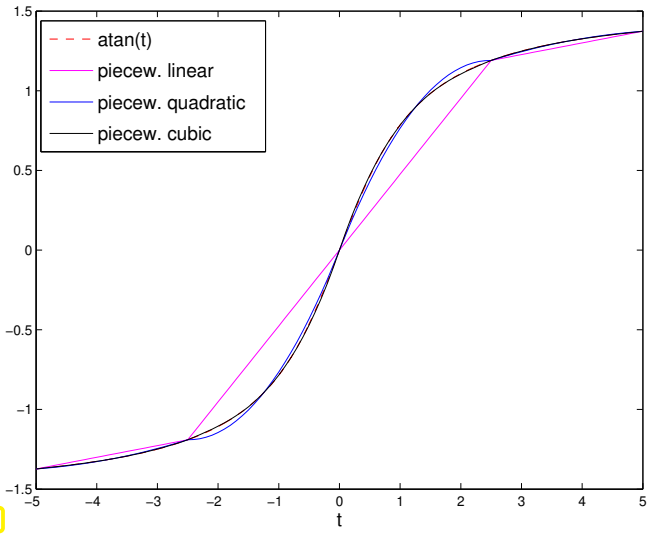


Fig. 256

- *Sequence* of (equidistant) meshes: $\mathcal{M}_i := \{-5 + j\,2^{-i}10\}_{j=0}^{2^i}$, $i = 1,\ldots,6$.
- *Equidistant* local interpolation nodes (endpoints of grid intervals included).

Monitored: interpolation error in (approximate) L^∞- and L^2-norms, see (6.2.3.25), (6.2.3.24)

$$\|g\|_{L^\infty([-5,5])} \approx \max_{j=0,\ldots,1000} |g(-5 + j/100)|\,,$$

$$\|g\|_{L^2([-5,5])} \approx \sqrt{\frac{1}{1000}} \cdot \left(\frac{1}{2} g(-5)^2 + \sum_{j=1}^{999} |g(-5+j/100)|^2 + \frac{1}{2} g(5)^2 \right)^{1/2} .$$

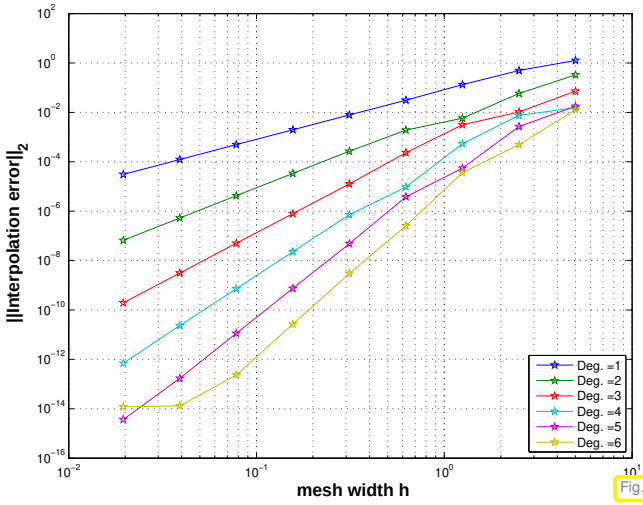


Fig. 257

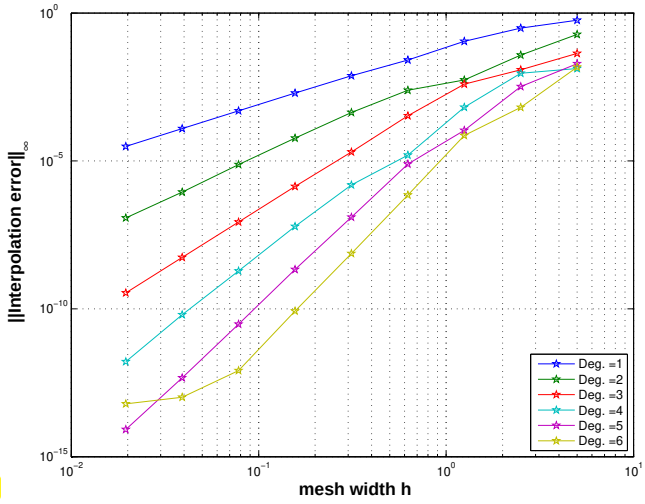


Fig. 258

Observation: Algebraic convergence (→ Def. 6.2.2.7) for meshwidth $h_{\mathcal{M}} \to 0$

(nearly linear error norm graphs in doubly logarithmic scale, see § 6.2.2.9)

Observation: rate of algebraic convergence increases with polynomial degree n

Rates α of algebraic convergence $O(h_{\mathcal{M}}^{\alpha})$ of norms of interpolation error:

n	1	2	3	4	5	6
w.r.t. L^2-norm	1.9957	2.9747	4.0256	4.8070	6.0013	5.2012
w.r.t. L^∞-norm	1.9529	2.8989	3.9712	4.7057	5.9801	4.9228

➢ Higher polynomial degree provides faster algebraic decrease of interpolation error norms. Empiric evidence for rates $\alpha = p+1$

Here: rates estimated by linear regression (→ Ex. 3.1.1.5) based on PYTHON's `polyfit` and the interpolation errors for meshwidth $h \leq 10 \cdot 2^{-5}$. This was done in order to avoid erratic "preasymptotic", that is, for large meshwidth h, behavior of the error.

The bad rates for $n = 6$ are probably due to the impact of roundoff, because the norms of the interpolation error had dropped below machine precision, see Fig. 257, 258. ⌟

§6.6.1.8 (Approximation error estimates for piecewise polynomial Lagrange interpolation) The observations made in Ex. 6.6.1.7 are easily explained by applying the polynomial interpolation error estimates of Section 6.2.2, for instance

$$\|f - \mathsf{I}_{\mathcal{T}} f\|_{L^\infty(I)} \leq \frac{\left\| f^{(n+1)} \right\|_{L^\infty(I)}}{(n+1)!} \max_{t \in I} |(t-t_0) \cdot \cdots \cdot (t-t_n)| . \tag{6.2.2.22}$$

locally on the mesh intervals $[x_{j-1}, x_j]$, $j = 1, \ldots, m$: for constant polynomial degree $n = n_j$, $j = 1, \ldots, m$, we get for $f \in C^{n+1}([x_0, x_m])$ (smoothness requirement)

$$(6.2.2.22) \quad \Rightarrow \quad \|f - s\|_{L^\infty([x_0,x_m])} \leq \frac{h_{\mathcal{M}}^{n+1}}{(n+1)!} \left\| f^{(n+1)} \right\|_{L^\infty([x_0,x_m])} , \tag{6.6.1.9}$$

with mesh width $h_{\mathcal{M}} := \max\{|x_j - x_{j-1}| : j = 1, \ldots, m\}$. ⌟

Another special case: fixed mesh $\mathcal{M}$, uniform polynomial degree n

Study estimates $\|f - \mathrm{I}_{\mathcal{M}} f\| \leq CT(n)$ for $n \to \infty$.

Terminology: investigation of p-convergence

EXAMPLE 6.6.1.10 (p-convergence of piecewise polynomial interpolation) We study p-convergence in the setting of Ex. 6.6.1.7.

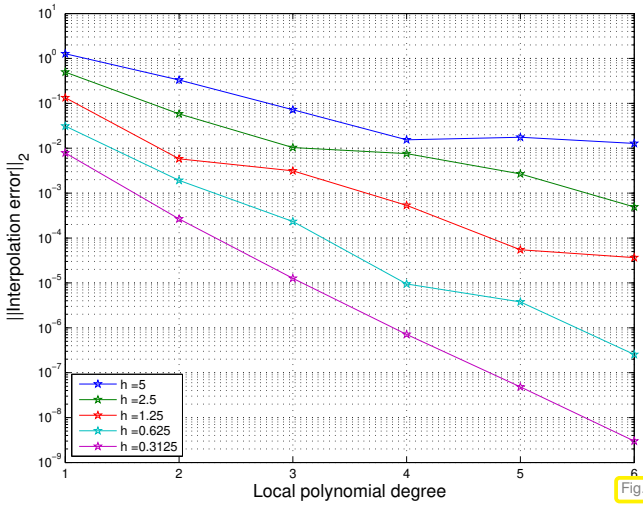


Fig. 259

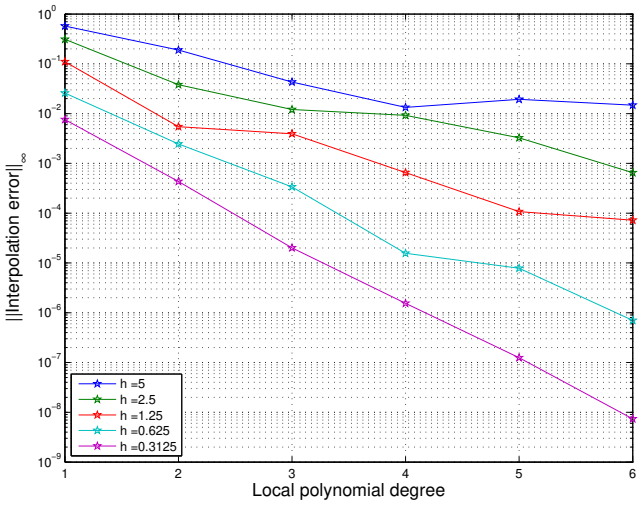


Fig. 260

Observation: (apparent) exponential convergence in polynomial degree

Note: in the case of p-convergence the situation is the same as for standard polynomial interpolation, see 6.2.2.

In this example we deal with an analytic function, see Rem. 6.2.3.26. Though equidistant local interpolation nodes are used *cf.* Ex. 6.2.2.11, the mesh intervals seems to be small enough that even in this case exponential convergence prevails. ⌟

6.6.2 Cubic Hermite Interpolation: Error Estimates

Video tutorial for Section 6.6.2 "Cubic Hermite and Spline Interpolation: Error Estimates": (10 minutes) Download link, tablet notes

See Section 5.3.3 for definition and algorithms for cubic Hermite interpolation of data points, with a focus on shape preservation, however. If the derivative f' of the interpoland f is available (in procedural form), then it can be used to fix local cubic polynomials by prescribing point values and derivative values in the endpoints of grid intervals.

Definition 6.6.2.1. Piecewise cubic Hermite interpolant (with exact slopes) → **Def. 5.3.3.1**

Given $f \in C^1([a,b])$ and a mesh $\mathcal{M} := \{a = x_0 < x_1 < \ldots < x_{m-1} < x_m = b\}$ the piecewise cubic Hermite interpolant (with exact slopes) $s : [a,b] \to \mathbb{R}$ is defined as

$$s_{|[x_{j-1},x_j]} \in \mathcal{P}_3 \,, \quad j = 1,\ldots,m \quad , \quad s(x_j) = f(x_j) \,, \quad s'(x_j) = f'(x_j) \,, \quad j = 0,\ldots,m \,.$$

Clearly, the piecewise cubic Hermite interpolant is continuously differentiable: $s \in C^1([a,b])$, *cf.* Cor. 5.3.3.2.

EXPERIMENT 6.6.2.2 (Convergence of Hermite interpolation with exact slopes) In this experiment

we study the h-convergence of Cubic Hermite interpolation for a smooth function.

Piecewise cubic Hermite interpolation of

$$f(x) = \arctan(x) \, .$$

- ✦ domain: $I = (-5,5)$
- ✦ mesh $\mathcal{T} = \{-5 + hj\}_{j=0}^{n} \subset I$, $h = \frac{10}{n}$,
- ✦ exact slopes $c_i = f'(t_i)$, $i = 0, \ldots, n$

algebraic convergence $O(h^4)$

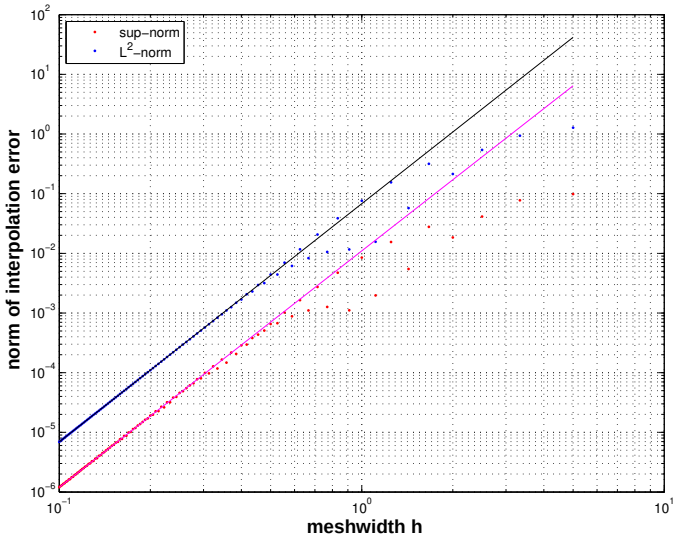


Fig. 261

Approximate computation of error norms analoguous to Ex. 6.6.1.7.

The observation made in Exp. 6.6.2.2 matches the theoretical prediction of the rate of algebraic convergence for cubic Hermite interpolation with exact slopes for a smooth function.

Theorem 6.6.2.3. Convergence of approximation by cubic Hermite interpolation

Let s be the cubic Hermite interpolant of $f \in C^4([a,b])$ on a mesh $\mathcal{M} := \{a = x_0 < x_1 < \ldots < x_{m-1} < x_m = b\}$ according to Def. 6.6.2.1. Then

$$\|f - s\|_{L^\infty([a,b])} \leq \frac{1}{4!} h_{\mathcal{M}}^4 \left\| f^{(4)} \right\|_{L^\infty([a,b])} ,$$

with the meshwidth $h_{\mathcal{M}} := \max_j |x_j - x_{j-1}|$.

In Section 5.3.3.2 we saw variants of cubic Hermite interpolation, for which the slopes $c_j = s'(x_j)$ were computed from the values y_j in preprocessing step. Now we study the use of such a scheme for approximation.

EXPERIMENT 6.6.2.4 (Convergence of Hermite interpolation with averaged slopes)

Piecewise cubic Hermite interpolation of

$$f(x) = \arctan(x) \ .$$

- ✦ domain: $I = (-5,5)$
- ✦ equidistant mesh $\mathcal{T}$ in I, see Exp. 6.6.2.2,
- ✦ averaged local slopes, see (5.3.3.8)

➤ algebraic convergence $O(h^3)$ in meshwidth

Code ⇢ GITLAB

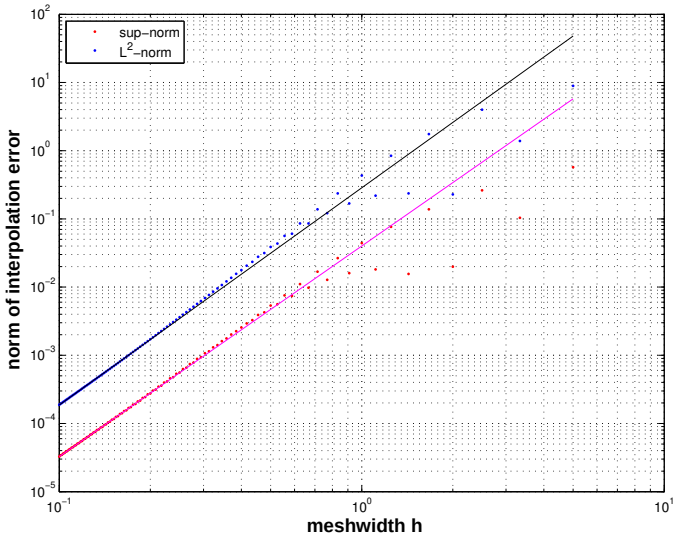


Fig. 262

We observe lower rate of algebraic convergence compared to the use of exact slopes due to averaging (5.3.3.8). From the plot we deduce $O(h^3)$ asymptotic decay of the L^2- and L^∞-norms of the approximation error for meshwidth $h \to 0$.

6.6.3 Cubic Spline Interpolation: Error Estimates [Han02, Ch. 47]

Recall concept and algorithms for cubic spline interpolation from Section 5.4.2. As an interpolation scheme it can also serve as the foundation for an approximation scheme according to § 6.1.0.6: the mesh will double as knot set, see Def. 5.4.1.1. Cubic spline interpolation is not local as we saw in § 5.4.3.7. Nevertheless, cubic spline interpolants can be computed with an effort of $O(m)$ as elaborated in § 5.4.2.5.

We have seen three main classes of cubic spline interpolants $s \in \mathcal{S}_{3,\mathcal{M}}$ of data points with node set $\mathcal{M} = \{a = t_0 < t_1 < \cdots < t_n = b\}$ § 5.4.2.11: the complete cubic spline (s' prescribed at endpoints), the natural cubic spline ($s''(a) = s''(b) = 0$), and the periodic cubic spline ($s'(a) = s'(b), s''(a) = s''(b)$). Obviously, both the natural and periodic cubic spline do not make much sense for approximating a generic continuous function $f \in C^0([a,b])$. So we focus on complete cubic spline interpolants with endpoint slopes inherited from the interpoland:

Definition 6.6.3.1. Complete cubic spline interpolant

Given $f \in C^1([a,b])$ and a mesh (= knot set) $\mathcal{M} := \{a = x_0 < x_1 < \ldots < x_{m-1} < x_m = b\}$ the **complete cubic spline Hermite interpolant** s is defined by the conditions

$$s \in \mathcal{S}_{3,\mathcal{M}} \ , \quad s(x_j) = f(x_j) \ , \quad j = 0,\ldots,m \ , \quad s'(a) = f'(a) \ , \quad s'(b) = f'(b) \ .$$

In § 5.4.2.5 and § 5.4.2.11 we found that interpolation condition at the knots plus fixing the derivatives at the endpoints uniquely determine a cubic spline function. Hence, the above definition is valid.

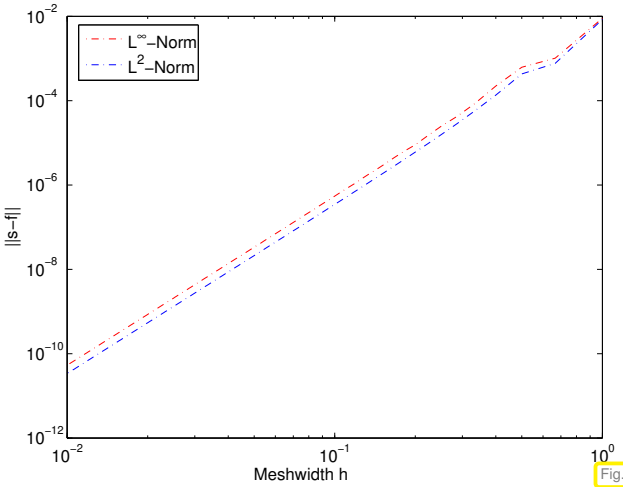
EXPERIMENT 6.6.3.2 (Approximation by complete cubic spline interpolants) We take $I = [-1,1]$ and rely on an equidistant mesh (knot set) $\mathcal{M} := \{-1 + \frac{2}{n}j\}_{j=0}^{n}$, $n \in \mathbb{N}$ ➤ meshwidth $h = 2/n$.

We study h-convergence of complete ($\to$ § 5.4.2.11) cubic spline interpolation according to Def. 6.6.3.1, where the slopes at the endpoints of the interval are made to agree with the derivatives of the interpoland

at these points. As interpolands we consider

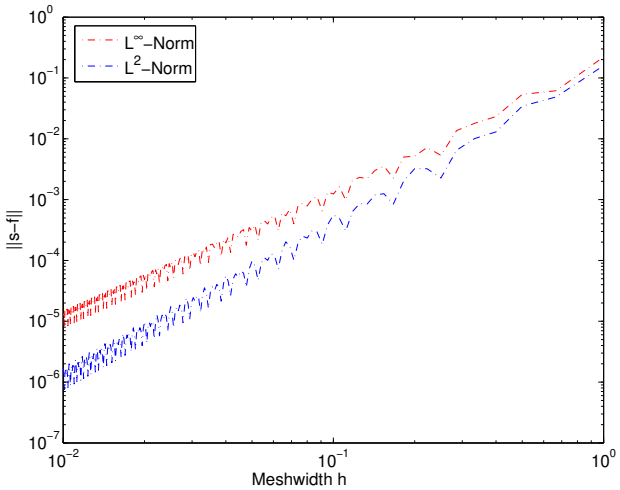
$$f_1(t) = \frac{1}{1+e^{-2t}} \in C^\infty(I) \quad , \quad f_2(t) = \begin{cases} 0 & \text{, if } t < -\frac{2}{5} \text{ ,} \\ \frac{1}{2}(1+\cos(\pi(t-\frac{3}{5}))) & \text{, if } -\frac{2}{5} < t < \frac{3}{5} \text{ ,} \\ 1 & \text{otherwise.} \end{cases} \in C^1(I) \text{ .}$$

Fig. 263



$$\|f_1 - s\|_{L^\infty([-1,1])} = O(h^4)$$

Fig. 264



$$\|f_2 - s\|_{L^\infty([-1,1])} = O(h^2)$$

The codes used to run this experiment can be accessed through ➙ GITLAB.

We observe algebraic order of convergence in h with empiric rate approximately given by $\min\{1 +$ regularity of $f, 4\}$. ⌟

We remark that there is the following theoretical result [HM76], [DR08, Rem. 9.2]:

$$f \in C^4([t_0,t_n]) \qquad \|f - s\|_{L^\infty([t_0,t_n])} \leq \frac{5}{384} h^4 \left\| f^{(4)} \right\|_{L^\infty([t_0,t_n])} \text{ .} \tag{6.6.3.3}$$

Summary and Learning Outcomes

This chapter is meant to impart the following knowledge and skills.

- You should be able to extract the (asymptotic) convergence of approximation errors from empiric data.
- You should know a few relevant norms on spaces of functions.
- You should be able to construct and approximation scheme on an arbitrary interval from an interpolatioon scheme on a fixed interval.
- You should recall bounds for the pointwise approximation error of polynomial interpolation for functions in C^r and analytic functions.
- You should be familiar with Chebychev interpolation: rationale, definition, and algorithms.
- You should know about trigonometric interpolation and the behavior of the associated pointwise approximation errors.
- You should be able to predict the convergence of piecewise polynomial interpolation in terms of meshwidth $h \to 0$.

Bibliography

[Akv+25] M. Akveld, A. Iozzi, P. Jossen, and F. da Lio. “Komplexe Analysis”. 2025 (cit. on pp. 495, 496).

[Boo05] Carl de Boor. “Divided differences”. In: *Surv. Approx. Theory* 1 (2005), pp. 46–69 (cit. on p. 497).

[Bör21] Steffen Börm. *On iterated interpolation*. 2021.

[CB95] Q. Chen and I. Babuska. “Approximate optimal points for polynomial interpolation of real functions in an interval and in a triangle”. In: *Comp. Meth. Appl. Mech. Engr.* 128 (1995), pp. 405–417 (cit. on p. 505).

[DR08] W. Dahmen and A. Reusken. *Numerik für Ingenieure und Naturwissenschaftler*. Heidelberg: Springer, 2008 (cit. on pp. 488, 552).

[Dav75] P.J. Davis. *Interpolation and Approximation*. New York: Dover, 1975 (cit. on pp. 477–479).

[DY10] L. Demanet and L. Ying. *On Chebyshev interpolation of analytic functions*. Online notes. 2010.

[DH03] P. Deuflhard and A. Hohmann. *Numerical Analysis in Modern Scientific Computing*. Vol. 43. Texts in Applied Mathematics. Springer, 2003 (cit. on p. 502).

[EZ66] H. Ehlich and K. Zeller. “Auswertung der Normen von Interpolationsoperatoren”. In: *Math. Ann.* 164 (1966), pp. 105–112. DOI: 10.1007/BF01429047.

[HM76] C.A. Hall and W.W. Meyer. “Optimal error bounds for cubic spline interpolation”. In: *J. Approx. Theory* 16 (1976), pp. 105–122 (cit. on p. 552).

[Han02] M. Hanke-Bourgeois. *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens*. Mathematische Leitfäden. Stuttgart: B.G. Teubner, 2002 (cit. on pp. 488, 501, 502, 510, 551).

[JWZ19] Peter Jantsch, Clayton G. Webster, and Guannan Zhang. “On the Lebesgue constant of weighted Leja points for Lagrange interpolation on unbounded domains”. In: *IMA J. Numer. Anal.* 39.2 (2019), pp. 1039–1057. DOI: 10.1093/imanum/dry002.

[NS02] K. Nipp and D. Stoffer. *Lineare Algebra*. 5th ed. Zürich: vdf Hochschulverlag, 2002 (cit. on p. 515).

[Ran00] R. Rannacher. *Einführung in die Numerische Mathematik*. Vorlesungsskriptum Universität Heidelberg. 2000 (cit. on p. 489).

[Rem84] R. Remmert. *Funktionentheorie I*. Grundwissen Mathematik 5. Berlin: Springer, 1984 (cit. on p. 495).

[Rem02] R. Remmert. *Funktionentheorie I*. Grundwissen Mathematik 5. Berlin: Springer, 2002 (cit. on p. 540).

[Str09] M. Struwe. *Analysis für Informatiker*. Lecture notes, ETH Zürich. 2009 (cit. on pp. 476, 481, 488, 493, 494).

[Tad86] Eitan Tadmor. “The Exponential Accuracy of Fourier and Chebyshev Differencing Methods”. In: *SIAM Journal on Numerical Analysis* 23.1 (1986), pp. 1–10. DOI: 10.1137/0723001.

[TT10] Rodney Taylor and Vilmos Totik. “Lebesgue constants for Leja points”. In: *IMA J. Numer. Anal.* 30.2 (2010), pp. 462–486. DOI: 10.1093/imanum/drn082.

[Tre13] Lloyd N. Trefethen. *Approximation theory and approximation practice*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2013, viii+305 pp.+back matter (cit. on pp. 491, 497, 505, 507).

[Tre] N. Trefethen. *Six myths of polynomial interpolation and quadrature*. Slides, University of Oxford.

[Ver86] P. Vertesi. "On the optimal Lebesgue constants for polynomial interpolation". In: *Acta Math. Hungaria* 47.1-2 (1986), pp. 165–178 (cit. on p. 505).

[Ver90] P. Vertesi. "Optimal Lebesgue constant for Lagrange interpolation". In: *SIAM J. Numer. Anal.* 27.5 (1990), pp. 1322–1331 (cit. on p. 505).

Chapter 7

Numerical Quadrature

7.1 Introduction

Video tutorial for Section 7.1 "Numerical Quadrature: Introduction": (4 minutes) Download link, tablet notes

Numerical quadrature deals with the *approximate* numerical evaluation of integrals $\int_\Omega f(\mathbf{x})\,\mathrm{d}\mathbf{x}$ for a given (closed) integration domain $\Omega \subset \mathbb{R}^d$. Thus, the underlying problem in the sense of § 1.5.5.1 is the mapping

$$I : \begin{cases} C^0(\Omega) & \to & \mathbb{R} \\ f & \mapsto & \int_\Omega f(\mathbf{x})\,\mathrm{d}\mathbf{x} \end{cases}, \tag{7.1.0.1}$$

with data space $X := C^0(\Omega)$ and result space $Y := \mathbb{R}$.

If f is complex-valued or vector-valued, then so is the integral. The methods presented in this chapter can immediately be generalized to this case by componentwise application.

§7.1.0.2 (Integrands in procedural form) The integrand f, a continuos function $f : \Omega \subset \mathbb{R}^d \mapsto \mathbb{R}$ should not be thought of as given by an analytic expression, but as given in procedural form, *cf.* Rem. 5.1.0.9.

For instance, in C++ the integrand is provided through a "functor" data type with an evaluation operator **double operator** `(Point &)` or a corresponding member function, see Code 5.1.0.10 for an example, or a lambda function (*to* Section 0.3.3).

General methods for numerical quadrature should rely only on finitely many *point evaluations* of the integrand. ⌟

In this chapter the focus is on the special case $d = 1$: $\Omega = [a, b]$ (an interval).

(Multidimensional numerical quadrature is substantially more difficult, unless Ω is tensor-product domain, a multi-dimensional box. Multidimensional numerical quadrature will be treated in the course "Numerical methods for partial differential equations".)

Remark 7.1.0.3 (Importance of numerical quadrature)

☞ Numerical quadrature methods are key building blocks for so-called variational methods for the numerical treatment of partial differential equations. A prominent example is the finite element method. They are also a pivotal device for the numerical solution of integral equations and in computational statistics. ⌟

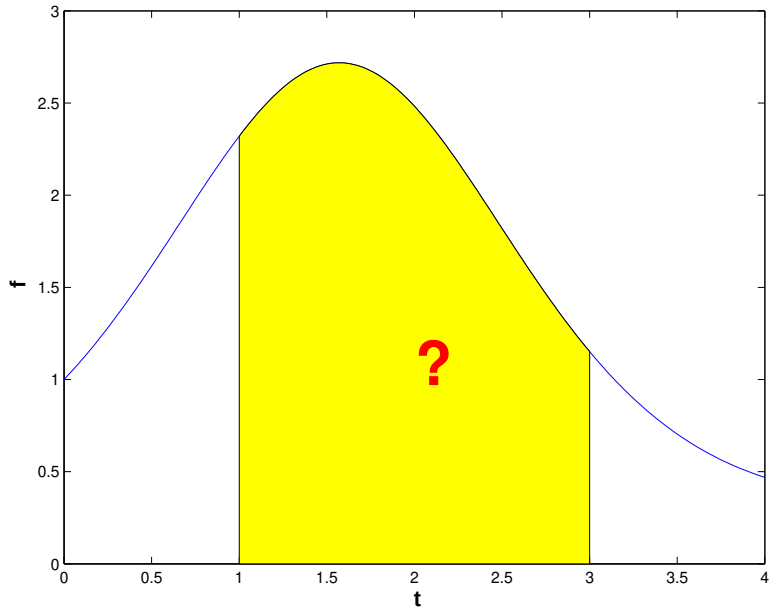


Fig. 265

For $d = 1$, from a geometric point of view methods for numerical quadrature aimed at computing

$$\int_a^b f(t)\,\mathrm{d}t$$

seek to *approximate* an area under the graph of the function f.

◁ area corresponding to value of an integral

EXAMPLE 7.1.0.4 (Heating production in electrical circuits) In Ex. 2.1.0.3 we learned about the nodal analysis of electrical circuits. Its application to a non-linear circuit will be discussed in Ex. 8.1.0.1, which will reveal that every computation of currents and voltages can be rather time-consuming. In this example we consider a non-linear circuit in quasi-stationary operation (capacities and inductances are ignored). Then the computation of branch currents and nodal voltages entails solving a non-linear system of equations.

Now assume time-harmonic periodic excitation with period $T > 0$.

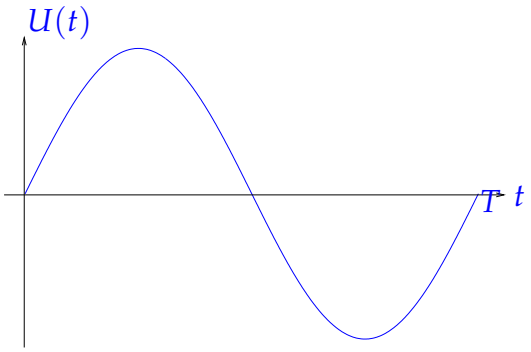


Fig. 266

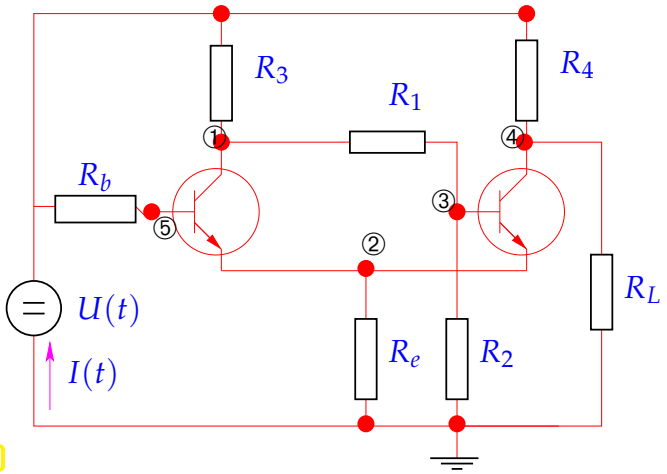


Fig. 267

The goal is to compute the energy dissipated by the circuit, which is equal to the energy injected by the voltage source. This energy can be obtained by integrating the power $P(t) = U(t)I(t)$ over period $[0, T]$:

$$W_{\text{therm}} = \int_0^T U(t)I(t)\,\mathrm{d}t\,, \quad \text{where} \quad I = I(U)\,.$$

double `I(double U)` involves solving non-linear system of equations, see Ex. 8.1.0.1!

This is a typical example where "point evaluation" by solving the non-linear circuit equations is the only way to gather information about the integrand. ⌟

Contents

7.1 Introduction 555
7.2 Quadrature Formulas – Quadrature Rules 557
7.3 Polynomial Quadrature Formulas 561
7.4 Gauss Quadrature 564
7.4.1 Order of a Quadrature Rule 564
7.4.2 Maximal-Order Quadrature Rules 567
7.4.3 Quadrature Error Estimates 575
7.5 Composite Quadrature 580
7.6 Adaptive Quadrature 588

Supplementary literature. Numerical quadrature is covered in [Han02, p. VII] and [DR08, Ch. 10].

Review question(s) 7.1.0.5 (Numerical quadrature: introduction)

(Q7.1.0.5.A) Let $\mathbf{A} \in \mathbb{R}^{n,n}$ a symmetric positive definite (s.p.d.) matrix and define

$$\widetilde{\mathbf{A}}(\xi) \in \mathbb{R}^{n,n}: \quad \left(\widetilde{\mathbf{A}}(\xi)\right)_{i,j} := \begin{cases} (\mathbf{A})_{i,j} & \text{for} \quad (i,j) \neq (1,1)\,, \\ (\mathbf{A})_{1,1} + \xi & \text{for} \quad (i,j) = (1,1)\,. \end{cases}$$

We consider the linear system of equations $\widetilde{\mathbf{A}}(\xi)\mathbf{x} = \mathbf{b}$ for given right-hand side vector $\mathbf{b} \in \mathbb{R}^n$.

We assume that $\xi : \Omega \to \mathbb{R}$ is a real-valued continuous random variable that is uniformly distributed in $[0,1]$. Derive integral expressions for the expectation and variance of the random variable

$$\eta := (\mathbf{x})_1 + \cdots + (\mathbf{x})_n\,.$$

△

7.2 Quadrature Formulas – Quadrature Rules

Video tutorial for Section 7.2 "Quadrature Formulas/Rules": (13 minutes) Download link, tablet notes

Quadrature formulas realize the approximation of an integral through finitely many point evaluations of the integrand.

Definition 7.2.0.1. Quadrature formula/quadrature rule

An n-point quadrature formula/quadrature rule (QR) on $[a,b]$ provides an approximation of the value of an integral through a *weighted sum* of point values of the integrand:

$$\int_a^b f(t)\,\mathrm{d}t \;\approx\; Q_n(f) := \sum_{j=1}^{n} w_j^{(n)} f(c_j^{(n)})\,. \tag{7.2.0.2}$$

Terminology:
$w_j^{(n)}$: quadrature weights $\in \mathbb{R}$
$c_j^{(n)}$: quadrature nodes $\in [a,b]$ (also called quadrature points)

Obviously (7.2.0.2) is compatible with integrands f given in procedural form as `double f(double t)`, compare § 7.1.0.2.

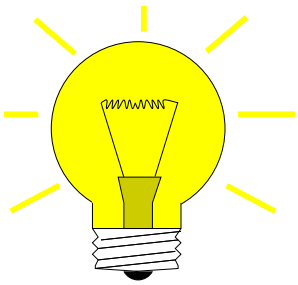
C++-code 7.2.0.3: C++ template implementing generic quadrature formula ➙ GITLAB

```
// Generic numerical quadrature routine implementing (7.2.0.2):
// f is a handle to a function, e.g. as lambda function
// c, w pass quadrature nodes c_j ∈ [a,b], and weights w_j ∈ ℝ
// in a Eigen::VectorXd
template <class Function>
double quadformula(Function &&f, const Eigen::VectorXd& c,const Eigen::VectorXd& w) {
  const Eigen::Index n = c.size();
  double I = 0;
  for (Eigen::Index i = 0; i < n; ++i) { I += w(i)*f(c(i)); }
  return I;
}
```

A single invocation costs n point evaluations of the integrand plus n additions and multiplications.

Remark 7.2.0.4 (Transformation of quadrature rules) In the setting of function approximation by polynomials we learned in § 6.2.1.14 that an approximation schemes for any interval could be obtained from an approximation scheme on a single reference interval ($[-1,1]$ in § 6.2.1.14) by means of affine pullback, see (6.2.1.18). A similar affine transformation technique makes it possible to derive quadrature formula for an arbitrary interval from a single quadrature formula on a reference interval.

Given: quadrature formula $(\widehat{c}_j, \widehat{w}_j)_{j=1}^n$ on reference interval $[-1,1]$



Idea: transformation formula for integrals

$$\int_a^b f(t)\,\mathrm{d}t = \tfrac{1}{2}(b-a)\int_{-1}^1 \widehat{f}(\tau)\,\mathrm{d}\tau\,, \qquad \widehat{f}(\tau) := f(\tfrac{1}{2}(1-\tau)a + \tfrac{1}{2}(\tau+1)b)\,. \tag{7.2.0.5}$$

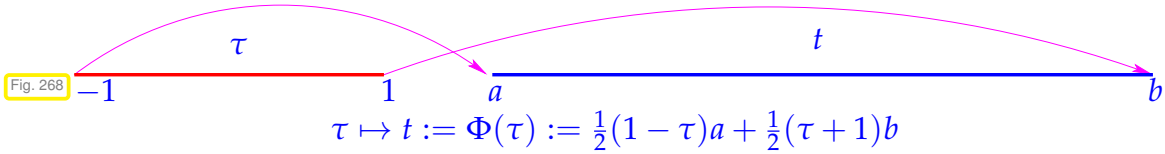


Fig. 268

$$\tau \mapsto t := \Phi(\tau) := \tfrac{1}{2}(1-\tau)a + \tfrac{1}{2}(\tau+1)b$$

Note that $\widehat{f}$ is the affine pullback $\Phi^* f$ of f to $[-1,1]$ as defined in (6.2.1.16).

quadrature formula for general interval $[a,b]$, $a,b \in \mathbb{R}$:

$$\int_a^b f(t)\,\mathrm{d}t \approx \tfrac{1}{2}(b-a)\sum_{j=1}^n \widehat{w}_j \widehat{f}(\widehat{c}_j) = \sum_{j=1}^n w_j f(c_j) \quad \text{with} \quad \begin{aligned} c_j &= \tfrac{1}{2}(1-\widehat{c}_j)a + \tfrac{1}{2}(1+\widehat{c}_j)b\,, \\ w_j &= \tfrac{1}{2}(b-a)\widehat{w}_j\,. \end{aligned}$$

In words, the nodes are just mapped through the affine transformation $c_j = \Phi(\widehat{c}_j)$, the weights are scaled by the ratio of lengths of $[a,b]$ and $[-1,1]$.

A 1D quadrature formula on arbitrary intervals can be specified by providing its weights $\widehat{w}_j$ /nodes $\widehat{c}_j$ for the integration domain $[-1,1]$ (reference interval). Then the above transformation is assumed.

Another common choice for the reference interval: $[0,1]$, pay attention! ⌟

Remark 7.2.0.6 (Tables of quadrature rules) In many codes families of quadrature rules are used to control the quadrature error. Usually, suitable sequences of weights $w_j^{(n)}$ and nodes $c_j^{(n)}$ are precomputed and stored in tables up to sufficiently large values of n. A possible interface could be the following:

```
struct QuadTab {
  template <typename VecType>
    static void getrule(int n,VecType &c,VecType &w,
                        double a=-1.0,double b=1.0);
}
```

Calling the method `getrule()` fills the vectors `c` and `w` with the nodes and the weights for a desired n-point quadrature on $[a,b]$ with $[-1,1]$ being the default reference interval. For **VecType** we may assume the basic functionality of **Eigen::VectorXd**. ⌟

§7.2.0.7 (Quadrature rules by approximation schemes) Every approximation scheme $\mathsf{A} : C^0([a,b]) \to V$, V a space of "simple functions" on $[a,b]$, see § 6.1.0.5, gives rise to a method for numerical quadrature according to

$$\int_a^b f(t)\,\mathrm{d}t \approx Q_{\mathsf{A}}(f) := \int_a^b (\mathsf{A}f)(t)\,\mathrm{d}t\,. \tag{7.2.0.8}$$

As explained in § 6.1.0.6 every interpolation scheme $\mathsf{I}_{\mathcal{T}} : \mathbb{R}^{n+1} \to V$ based on the node set $\mathcal{T} = \{t_0, t_1, \ldots, t_n\} \subset [a,b]$ (→ § 5.1.0.7) induces an approximation scheme, and, hence, also a quadrature scheme on $[a,b]$:

$$\int_a^b f(t)\,\mathrm{d}t \approx \int_a^b \mathsf{I}_{\mathcal{T}}[f(t_0),\ldots,f(t_n)]^\top(t)\,\mathrm{d}t\,. \tag{7.2.0.9}$$

Lemma 7.2.0.10. Quadrature formulas from linear interpolation schemes

Every linear *interpolation operator* $\mathsf{I}_{\mathcal{T}}$ *according to Def. 5.1.0.25 spawns a quadrature formula (→ Def. 7.2.0.1) by* (7.2.0.9).

Proof. Writing $\mathbf{e}_j$ for the j-th unit vector of $\mathbb{R}^{n+1}$, $j = 0,\ldots,n$, we have by linearity

$$\int_a^b \mathsf{I}_{\mathcal{T}}[f(t_0),\ldots,f(t_n)]^\top\,\mathrm{d}t = \sum_{j=0}^n f(t_j) \underbrace{\int_a^b (\mathsf{I}_{\mathcal{T}}(\mathbf{e}_j))(t)\,\mathrm{d}t}_{\text{weight } w_j}\,. \tag{7.2.0.11}$$

Hence, we have arrived at an $n+1$-point quadrature formula with nodes t_j, whose weights are the integrals of the cardinal interpolants for the interpolation scheme $\mathcal{T}$. □

Summing up, we have found

interpolation schemes ⟶ approximation schemes ⟶ quadrature schemes

⌟

§7.2.0.12 (Convergence of numerical quadrature) In general the quadrature formula (7.2.0.2) will only provide an approximate value for the integral.

➢ For a generic integrand we will encounter a non-vanishing

$$\text{quadrature error} \qquad E_n(f) := \left| \int_a^b f(t)\,\mathrm{d}t - Q_n(f) \right|$$

As in the case of function approximation by interpolation Section 6.2.2, our focus will on the asymptotic behavior of the quadrature error as a function of the number n of point evaluations of the integrand.

Therefore consider families of quadrature rules $\{Q_n\}_n$ (→ Def. 7.2.0.1) described by

- ✦ quadrature weights $\left\{w_j^n,\ j=1,\ldots,n\right\}_{n\in\mathbb{N}}$ and
- ✦ quadrature nodes $\left\{c_j^n,\ j=1,\ldots,n\right\}_{n\in\mathbb{N}}$.

We study the *asymptotic* behavior of the quadrature error $E(n)$ for $n \to \infty$

As in the case of interpolation errors in § 6.2.2.5 we make the usual qualitative distinction, see Def. 6.2.2.7:

- ▷ algebraic convergence $E(n) = O(n^{-p})$, rate $p > 0$
- ▷ exponential convergence $E(n) = O(q^n)$, $0 \le q < 1$

Note that the number n of nodes agrees with the number of f-evaluations required for the evaluation of the quadrature formula. This is usually used as a *measure for the cost* of computing $Q_n(f)$.

Therefore, in the sequel, we consider the quadrature error as a function of n. ⌟

§7.2.0.13 (Quadrature error from approximation error) Bounds for the maximum norm of the approximation error of an approximation scheme directly translate into estimates of the quadrature error of the induced quadrature scheme (7.2.0.8):

$$\left| \int_a^b f(t)\,\mathrm{d}t - Q\mathrm{A}(f) \right| \le \int_a^b |f(t) - \mathrm{A}(f)(t)|\,\mathrm{d}t \le |b-a| \|f - \mathrm{A}(f)\|_{L^\infty([a,b])} \,. \tag{7.2.0.14}$$

Hence, the various estimates derived in Section 6.2.2 and Section 6.2.3.2 give us quadrature error estimates "for free". More details will be given in the next section. ⌟

Review question(s) 7.2.0.15 (Quadrature formulas)

(Q7.2.0.15.A) Explain the structure of a quadrature formula/rule for the approximation of the integral $\int_a^b f(t)\,\mathrm{d}t$, $a, b \in \mathbb{R}$.

(Q7.2.0.15.B) The integral satisfies

$$g, f \in C^0([a,b]) \quad , \quad g \le f \quad \text{on} \quad [a,b] \quad \Rightarrow \quad \int_a^b g(t)\,\mathrm{d}t \le \int_a^b f(t)\,\mathrm{d}t \,.$$

Formulate necessary and sufficient conditions on an n-point quadrature rule Q_n such that

$$g, f \in C^0([a,b]) \quad , \quad g \le f \quad \text{on} \quad [a,b] \quad \Rightarrow \quad Q_n(g) \le Q_n(f) \,.$$

(Q7.2.0.15.C) The documentation of the C++ function

```
Eigen::Matrix<double, Eigen::Dynamic, 2>
  getQuadRule(unsigned int n);
```

claims that it provides a family of quadrature rules on the interval $[-1,1]$, and that `n` passes the number of quadrature nodes/points.

Why does this claim make sense and which piece of information is missing? How could you retrieve that missing piece, if you can call the function `getQuadRule()`.

(Q7.2.0.15.D) Consider the following C++ function for numerical quadrature.

```
template <typename QUADRULE, typename FUNCTION>
auto integrate(const QUADRULE &qr, FUNCTION &&f)
    -> decltype(qr[0].second * f(qr[0].first)) {
  using Scalar = decltype(qr[0].second * f(qr[0].first));
  Scalar s{0};
  for (auto nw : qr) {
    s += nw.second * f(nw.first);
  }
  return s;
}
```

Explain the requirements on the types **QUADRULE** and **FUNCTION**.

Hint. The C++ class **std::pair** is an abstract container for two objects of different types that can be accessed via data members `first` and `second`.

(Q7.2.0.15.E) A integration path γ in the complex plane $\mathbb{C}$ is usually given by its parameterization $\gamma : [a,b] \to \mathbb{C}$, $a,b \in \mathbb{R}$. If $f : D \subset \mathbb{C} \to \mathbb{C}$ is continuous and $\gamma([a,b]) \subset D$, then the path integral of f along γ is defined as

$$\int_\gamma f(z)\,\mathrm{d}z := \int_I f(\gamma(\tau)) \cdot \dot{\gamma}(\tau)\,\mathrm{d}\tau\,, \tag{6.2.2.51}$$

where $\dot{\gamma}$ designates the derivative of γ with respect to the parameter, and $\cdot$ indicates multiplication in $\mathbb{C}$.

How do quadrature formulas look like that can be used for the approximate computation of

$$\int_{\partial\mathbb{D}} f(z)\,\mathrm{d}z\,, \quad \mathbb{D} := \{z \in \mathbb{C} : |z| \le 1\}\,.$$

Hint. A natural parameterization of $\partial\mathbb{D}$ is provided by the complex exponential.

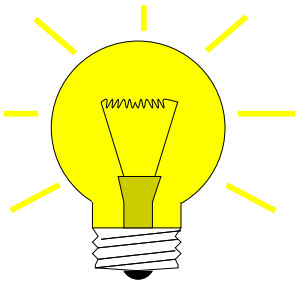
△

7.3 Polynomial Quadrature Formulas

Video tutorial for Section 7.3 "Polynomial Quadrature Formulas": (9 minutes) Download link, tablet notes

Now we specialize the general recipe of § 7.2.0.7 for approximation schemes based on global polynomials, the Lagrange approximation scheme as introduced in Section 6.2, Def. 6.2.2.1.

Supplementary literature. This topic is discussed in [DR08, Sect. 10.2].



Idea: replace integrand f with $p_{n-1} := \mathrm{I}_{\mathcal{T}} \in \mathcal{P}_{n-1}$ = polynomial Lagrange interpolant of f ($\to$ Cor. 5.2.2.8) for given node set $\mathcal{T} := \{t_0,\ldots,t_{n-1}\} \subset [a,b]$

$$\int_a^b f(t)\,\mathrm{d}t \approx Q_n(f) := \int_a^b p_{n-1}(t)\,\mathrm{d}t\,. \tag{7.3.0.1}$$

The cardinal interpolants for Lagrange interpolation are the Lagrange polynomials (5.2.2.4)

$$L_i(t) := \prod_{\substack{j=0 \\ j\neq i}}^{n-1} \frac{t-t_j}{t_i-t_j}, \quad i=0,\ldots,n-1 \quad \overset{(5.2.2.6)}{\blacktriangleright} \quad p_{n-1}(t) = \sum_{i=0}^{n-1} f(t_i)L_i(t) \,.$$

Then (7.2.0.11) amounts to the n-point quadrature formula

$$\int_a^b p_{n-1}(t)\,\mathrm{d}t = \sum_{i=0}^{n-1} f(t_i) \int_a^b L_i(t)\,\mathrm{d}t \quad \blacktriangleright \quad \begin{array}{l} \text{nodes} \quad c_i = t_{i-1}\,, \\ \text{weights} \quad w_i := \int_a^b L_{i-1}(t)\,\mathrm{d}t\,. \end{array} \tag{7.3.0.2}$$

EXAMPLE 7.3.0.3 (Midpoint rule)

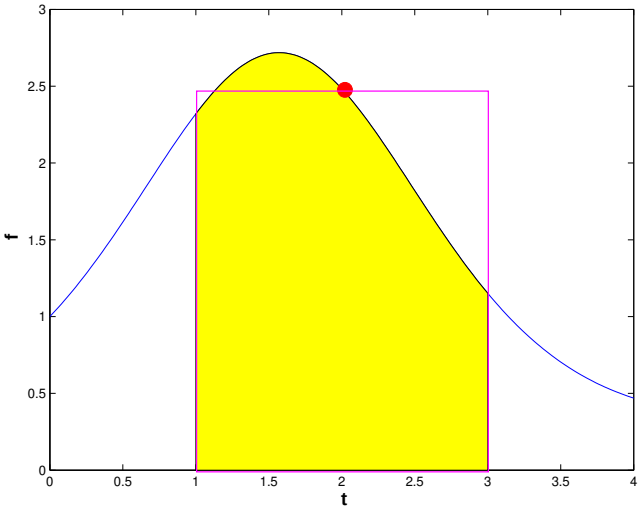


Fig. 269

The midpoint rule is (7.3.0.2) for $n = 1$ and $t_0 = \frac{1}{2}(a+b)$. It leads to the 1-point quadrature formula

$$\int_a^b f(t)\,\mathrm{d}t \approx Q_{\mathrm{mp}}(f) = (b-a)f(\tfrac{1}{2}(a+b)) \,.$$

"midpoint"

◁ the area under the graph of f is approximated by the area of a rectangle.

EXAMPLE 7.3.0.4 (Newton-Cotes formulas → [Han02, Ch. 38]) The $n := m+1$-point Newton-Cotes formulas arise from Lagrange interpolation in equidistant nodes (6.2.2.3) in the integration interval $[a,b]$:

▶ Equidistant quadrature nodes $t_j := a + hj, \quad h := \dfrac{b-a}{m}, \quad j = 0,\ldots,m$:

The weights for the interval $[0,1]$ can be found, e.g., by symbolic computation using MAPLE: the following MAPLE function expects the polynomial degree as input argument and computes the weights for the interval $[0,1]$:

```
> newtoncotes := m -> factor(int(interp([seq(i/n, i=0..m)],
                    [seq(f(i/n), i=0..m)], z),z=0..1)):
```

Weights on general intervals $[a,b]$ can then be deduced by the affine transformation rule as explained in Rem. 7.2.0.4.

• $n = 2$: Trapezoidal rule (integrate linear interpolant of integrand in endpoints)

```
> trapez := newtoncotes(1);
```

$$\widehat{Q}_{\mathrm{trp}}(f) := \frac{1}{2}(f(0)+f(1)) \qquad (7.3.0.5)$$

$$\left(\int_a^b f(t)\,\mathrm{d}t \approx \frac{b-a}{2}(f(a)+f(b)) \right)$$

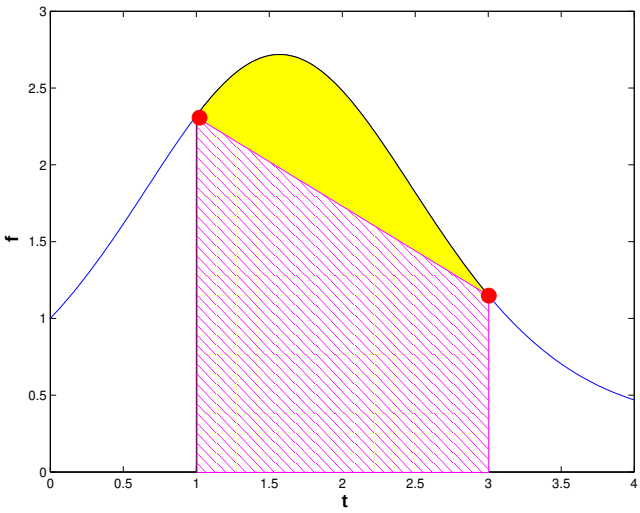


Fig. 270

- $n = 3$: Simpson rule

```
> simpson := newtoncotes(2);
```

$$\frac{1}{6}\left(f(0)+4f(\tfrac{1}{2})+f(1)\right) \quad \left(\int_a^b f(t)\,\mathrm{d}t \approx \frac{b-a}{6}\left(f(a)+4f\left(\frac{a+b}{2}\right)+f(b)\right) \right) \qquad (7.3.0.6)$$

- $n = 5$: Milne rule

```
> milne := newtoncotes(4);
```

$$\frac{1}{90}\left(7f(0)+32f(\tfrac{1}{4})+12f(\tfrac{1}{2})+32f(\tfrac{3}{4})+7f(1)\right)$$

$$\left(\frac{b-a}{90}(7f(a)+32f(a+(b-a)/4)+12f(a+(b-a)/2)+32f(a+3(b-a)/4)+7f(b))\right)$$

- $n = 7$: Weddle rule

```
> weddle := newtoncotes(6);
```

$$\frac{1}{840}\left(41f(0)+216f(\tfrac{1}{6})+27f(\tfrac{1}{3})+272f(\tfrac{1}{2})+27f(\tfrac{2}{3})+216f(\tfrac{5}{6})+41f(1)\right).$$

- $n \geq 8$: quadrature formulas with *negative* weights

```
> newtoncotes(8);
```

$$\frac{1}{28350}\,(989f(0)+5888f(\tfrac{1}{8})-928f(\tfrac{1}{4})+10496f(\tfrac{3}{8})$$
$$-4540f(\tfrac{1}{2})+10496f(\tfrac{5}{8})-928f(\tfrac{3}{4})+5888f(\tfrac{7}{8})+989f(1))$$

From Ex. 6.2.2.11 we know that the approximation error incurred by Lagrange interpolation in equidistant nodes can blow up even for analytic functions. This blow-up can also infect the quadrature error of Newton-Cotes formulas for large n, which renders them essentially useless. In addition they will be marred by large (in modulus) and negative weights, wich compromises numerical stability ($\rightarrow$ Def. 1.5.5.19)

No negative weights!

Quadrature formulas with negative weights should not be used, not even considered!

Remark 7.3.0.8 (Clenshaw-Curtis quadrature rules [Tre08]) The considerations of Section 6.2.3 confirmed the superiority of the "optimal" Chebychev nodes (6.2.3.10) for globally polynomial Lagrange interpolation. This suggests that we use these nodes also for numerical quadrature with weights given by (7.3.0.2). This yields the so-called Clenshaw-Curtis rules with the following rather desirable property:

Theorem 7.3.0.9. Positivity of Clenshaw-Curtis weights [Fej33]

The weights w_j^n, $j = 1, \ldots, n$, for every n-point Clenshaw-Curtis rule are positive.

The weights of any n-point Clenshaw-Curtis rule can be computed with a computational effort of $O(n \log n)$ using FFT [Wal06], [Tre08, Sect. 2]. ⌟

§7.3.0.10 (Error estimates for polynomial quadrature) As a concrete application of § 7.2.0.13, (7.2.0.14) we use the L^∞-bound (6.2.2.22) for Lagrange interpolation

$$\|f - \mathsf{L}_\mathcal{T} f\|_{L^\infty(I)} \leq \frac{\left\|f^{(n+1)}\right\|_{L^\infty(I)}}{(n+1)!} \max_{t \in I} |(t - t_0) \cdot \cdots \cdot (t - t_n)| \,. \tag{6.2.2.22}$$

to conclude for any n-point quadrature rule based on polynomial interpolation:

$$f \in C^n([a,b]) \quad \Rightarrow \quad \left| \int_a^b f(t)\,\mathrm{d}t - Q_n(f) \right| \leq \frac{1}{n!}(b-a)^{n+1} \left\|f^{(n)}\right\|_{L^\infty([a,b])} \,. \tag{7.3.0.11}$$

Much sharper estimates for Clenshaw-Curtis rules (→ Rem. 7.3.0.8) can be inferred from the interpolation error estimate (6.2.3.18) for Chebychev interpolation. For functions with *limited smoothness* algebraic convergence of the quadrature error for Clenshaw-Curtis quadrature follows from (6.2.3.21). For integrands that possess an *analytic* extension to the complex plane in a neighborhood of $[a,b]$, we can conclude exponential convergence from (6.2.3.28). ⌟

Review question(s) 7.3.0.12 (Polynomial quadrature formulas)

(Q7.3.0.12.A) Given an n-point quadrature formula on $[-1,1]$ in terms of pairs of weights and nodes (w_j, c_j), $j = 1, \ldots, n$, how can you decide whether it is a "polynomial quadrature formula", that is, a quadrature formula that is based on Lagrange polynomial interpolation of the integrand.

△

7.4 Gauss Quadrature

Supplementary literature. Gauss quadrature is discussed in detail in [Han02, Ch. 40-41], [DR08, Sect.10.3]

7.4.1 Order of a Quadrature Rule

Video tutorial for Section 7.4.1 "Order of a Quadrature Rule": (9 minutes) Download link, tablet notes

How to gauge the "quality" of an n-point quadrature formula Q_n without testing it for specific integrands? The next definition gives a classical answer.

Definition 7.4.1.1. Order of a quadrature rule

The **order** of a quadrature rule $Q_n : C^0([a,b]) \to \mathbb{R}$ is defined as

$$\text{order}(Q_n) := \max\{m \in \mathbb{N}_0: \quad Q_n(p) = \int_a^b p(t)\,\mathrm{d}t \quad \forall p \in \mathcal{P}_m\} + 1\,, \tag{7.4.1.2}$$

that is, as the maximal degree $+1$ of polynomials for which the quadrature rule is guaranteed to be exact.

§7.4.1.3 (Invariance of order under (affine) transformation) First we note a simple consequence of the invariance of the polynomial space $\mathcal{P}_n$ under affine pullback, see Lemma 6.2.1.17.

Corollary 7.4.1.4. Invariance of order under affine transformation

An affine transformation of a quadrature rule according to Rem. 7.2.0.4 does not change its order.

⌟

§7.4.1.5 (Order of polynomial quadrature rules) Further, by construction all polynomial n-point quadrature rules possess order at least n.

Theorem 7.4.1.6. Sufficient order conditions for quadrature rules

An n-point quadrature rule on $[a,b]$ ($\to$ Def. 7.2.0.1)

$$Q_n(f) := \sum_{j=1}^n w_j f(c_j)\,, \quad f \in C^0([a,b])\,,$$

with nodes $t_j \in [a,b]$ and weights $w_j \in \mathbb{R}$, $j = 1,\ldots,n$, has order $\geq n$, if and only if

$$w_j = \int_a^b L_{j-1}(t)\,dt\,, \quad j = 1,\ldots,n\,,$$

where L_k, $k = 0,\ldots,n-1$, is the k-th Lagrange polynomial (5.2.2.4) associated with the ordered node set $\{c_1, c_2, \ldots, c_n\}$.

Proof. The conclusion of the theorem is a direct consequence of the facts that

$$\mathcal{P}_{n-1} = \text{Span}\{L_0,\ldots,L_{n-1}\} \quad \text{and} \quad L_k(t_j) = \delta_{k+1,j}\,, \quad k+1, j \in \{1,\ldots,n\}\,:$$

just plug a Lagrange polynomial into the quadrature formula. □

By construction (7.3.0.2) polynomial n-point quadrature formulas (7.3.0.1) exact for $f \in \mathcal{P}_{n-1} \Rightarrow$ n-point polynomial quadrature formula has at least order n. ⌟

Remark 7.4.1.7 (Linear system for quadrature weights) Thm. 7.4.1.6 provides a concrete formula for quadrature weights, which guaranteed order n for an n-point quadrature formula. Yet evaluating integrals

of Lagrange polynomials may be cumbersome. Here we give a general recipe for finding the weights w_j according to Thm. 7.4.1.6 without dealing with Lagrange polynomials.

Given: arbitrary nodes $c_1, \ldots, c_n$ for n-point (local) quadrature formula on $[a,b]$

From Def. 7.4.1.1 we immediately conclude the following procedure: If $p_1, \ldots, p_n$ is *any basis* of $\mathcal{P}_{n-1}$, then, thanks to the linearity of the integral and quadrature formulas,

$$Q_n(p_j) = \int_a^b p_j(t)\,\mathrm{d}t \quad \forall j = 1,\ldots,n \quad \Leftrightarrow \quad Q_n \text{ has order } \geq n\,. \tag{7.4.1.8}$$

➢ $n \times n$ linear system of equations, see (7.4.2.3) for an example:

$$\begin{bmatrix} p_1(c_1) & \ldots & p_1(c_n) \\ \vdots & & \vdots \\ p_n(c_1) & \ldots & p_n(c_n) \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix} = \begin{bmatrix} \int_a^b p_1(t)\,\mathrm{d}t \\ \vdots \\ \int_a^b p_n(t)\,\mathrm{d}t \end{bmatrix}. \tag{7.4.1.9}$$

For instance, for the computation of quadrature weights, one may choose the monomial basis $p_j(t) = t^j$. ⌟

EXAMPLE 7.4.1.10 (Orders of simple polynomial quadrature formulas) From the order rule for polynomial quadrature rule we immediately conclude the orders of simple representatives.

n		Order
1	midpoint rule	2
2	trapezoidal rule (7.3.0.5)	2
3	Simpson rule (7.3.0.6)	4
4	$\frac{3}{8}$-rule	4
5	Milne rule	6

The orders for odd n surpass the predictions of Thm. 7.4.1.6 by 1, which can be verified by straightforward computations; following Def. 7.4.1.1 check the exactness of the quadrature rule on $[0,1]$ (this is sufficient $\rightarrow$ Cor. 7.4.1.4) for monomials $\{t \mapsto t^k\}$, $k = 0,\ldots,q-1$, which form a basis of $\mathcal{P}_q$, where q is the order that is to be confirmed: essentially one has to show

$$Q(\{t \mapsto t^k\}) = \sum_{j=1}^{n} w_j c_j^k = \frac{1}{k+1}\,, \quad k = 0,\ldots,q-1\,, \tag{7.4.1.11}$$

where $Q \mathrel{\hat{=}}$ quadrature rule on $[0,1]$ given by (7.2.0.2).

For the Simpson rule (7.3.0.6) we can also confirm order 4 with symbolic calculations in MAPLE:

```
> rule := 1/3*h*(f(2*h)+4*f(h)+f(0))
> err := taylor(rule - int(f(x),x=0..2*h),h=0,6);
```

$$err := \left(\frac{1}{90}\left(D^{(4)}\right)(f)(0)h^5 + O\left(h^6\right), h, 6\right)$$

➢ Composite Simpson rule possesses order 4, indeed **!** ⌟

Review question(s) 7.4.1.12 (Order of a quadrature rule)

(Q7.4.1.12.A) What is the minimal order of an n-point polynomial quadrature rule?

(Q7.4.1.12.B) What is meant by the statement that a linear mapping $X \to Y$, X, Y finite-dimensional vector spaces, is uniquely determined by its action on the elements of a basis of X?

(Q7.4.1.12.C) In order to determine the weights of an order-n n-point quadrature rule on $[-1,1]$ we rely on the monomials $t \to t^k$, $k = 0, \ldots, n-1$, as basis polynomials. Write down the linear system of equations for the weights w_j, $j = 1, \ldots, n$.

(Q7.4.1.12.D) Discuss the statement:

"A quadrature rule is good, if it yields a small quadrature error."

△

7.4.2 Maximal-Order Quadrature Rules

Video tutorial for Section 7.4.2 "Maximal-Order Quadrature Rules": (16 minutes)
Download link, tablet notes

A natural question is whether an n-point quadrature formula achieve an order $> n$. A negative result limits the maximal order that can be achieved:

Theorem 7.4.2.1. Maximal order of n-point quadrature rule

The maximal order of an n-point quadrature rule is $2n$.

Proof. Consider a generic n-point quadrature rule according to Def. 7.2.0.1

$$Q_n(f) := \sum_{j=1}^{n} w_j^n f(c_j^n) , \tag{7.2.0.2}$$

We build a polynomial of degree $2n$ that cannot be integrated exactly by Q_n. We choose polynomial

$$q(t) := (t - c_1)^2 \cdot \cdots \cdot (t - c_n)^2 \in \mathcal{P}_{2n} .$$

On the one hand, q is strictly positive almost everywhere, which means

$$\int_a^b q(t)\,\mathrm{d}t > 0 .$$

On the other hand, we find a different value

$$Q_n(q) = \sum_{j=1}^{n} w_j^n \underbrace{q(c_j^n)}_{=0} = 0 .$$

□

Can we at least find n-point rules with maximal order $2n$?

Heuristics: A quadrature formula has order $m \in \mathbb{N}$ already, if it is exact for m polynomials $\in \mathcal{P}_{m-1}$ that form a basis of $\mathcal{P}_{m-1}$ (recall Thm. 5.2.1.2).

$\Updownarrow$

An n-point quadrature formula has $2n$ "degrees of freedom" (n node positions, n weights).

$\Downarrow$

It might be possible to achieve order $2n = \dim \mathcal{P}_{2n-1}$
("No. of equations = No. of unknowns")

EXAMPLE 7.4.2.2 (2-point quadrature rule of order 4) Necessary & sufficient conditions for order 4, *cf.* (7.4.1.9), integrate the functions of the monomial basis of $\mathcal{P}_3$ exactly:

$$Q_n(p) = \int_a^b p(t)\,\mathrm{d}t \;\forall p \in \mathcal{P}_3 \quad \Leftrightarrow \quad Q_n(\{t \mapsto t^q\}) = \frac{1}{q+1}(b^{q+1} - a^{q+1})\,, \quad q = 0,1,2,3\,.$$

4 equations for weights w_j and nodes c_j, $j = 1,2$ ($a = -1, b = 1$), *cf.* Rem. 7.4.1.7

$$\begin{aligned} \int_{-1}^{1} 1\,\mathrm{d}t = 2 = 1w_1 + 1w_2 \quad , \quad \int_{-1}^{1} t\,\mathrm{d}t = 0 = c_1w_1 + c_2w_2 \\ \int_{-1}^{1} t^2\,\mathrm{d}t = \frac{2}{3} = c_1^2w_1 + c_2^2w_2 \quad , \quad \int_{-1}^{1} t^3\,\mathrm{d}t = 0 = c_1^3w_1 + c_2^3w_2\,. \end{aligned} \tag{7.4.2.3}$$

Solve using MAPLE:

```
> eqns := {seq(int(x^k, x=-1..1) = w[1]*xi[1]^k+w[2]*xi[2]^k,k=0..3)};
> sols := solve(eqns, indets(eqns, name)):
> convert(sols, radical);
```

➢ weights & nodes: $\left\{w_2 = 1, w_1 = 1, c_1 = 1/3\sqrt{3}, c_2 = -1/3\sqrt{3}\right\}$

quadrature formula (order 4): $\int_{-1}^{1} f(x)\,dx \approx f\left(\frac{1}{\sqrt{3}}\right) + f\left(-\frac{1}{\sqrt{3}}\right)$ (7.4.2.4)

§7.4.2.5 (Construction of n-point quadrature rules with maximal order $2n$) First we search for *necessary conditions* that have to be met by the nodes, if an n-point quadrature rule has order $2n$.

Optimist's **assumption**: $\exists$ family of n-point quadrature formulas on $[-1,1]$

$$Q_n(f) := \sum_{j=1}^{n} w_j^{(n)} f(c_j^{(n)}) \approx \int_{-1}^{1} f(t)\,\mathrm{d}t\,, \quad w_j^{(n)} \in \mathbb{R}\,, n \in \mathbb{N}\,,$$

of order $2n$ $\Leftrightarrow$ exact for polynomials $\in \mathcal{P}_{2n-1}$. (7.4.2.6)

Define $\bar{P}_n(t) := (t - c_1^{(n)}) \cdot \cdots \cdot (t - c_n^{(n)})\,, \quad t \in \mathbb{R} \quad \Rightarrow \quad \bar{P}_n \in \mathcal{P}_n\,.$

Note: $\bar{P}_n$ has leading coefficient $= 1$.

By assumption on the order of Q_n we know that for any $q \in \mathcal{P}_{n-1}$

$$\int_{-1}^{1} \underbrace{q(t)\bar{P}_n(t)}_{\in\mathcal{P}_{2n-1}}\,\mathrm{d}t \overset{(7.4.2.6)}{=} \sum_{j=1}^{n} w_j^{(n)} q(c_j^{(n)}) \underbrace{\bar{P}_n(c_j^{(n)})}_{=0} = 0\,.$$

We conclude $L^2([-1,1])$-orthogonality: $\int_{-1}^{1} q(t)\bar{P}_n(t)\,\mathrm{d}t = 0 \quad \forall q \in \mathcal{P}_{n-1}\,.$ (7.4.2.7)

$L^2(]-1,1[)$-inner product of q and $\bar{P}_n$, see (6.3.1.5).

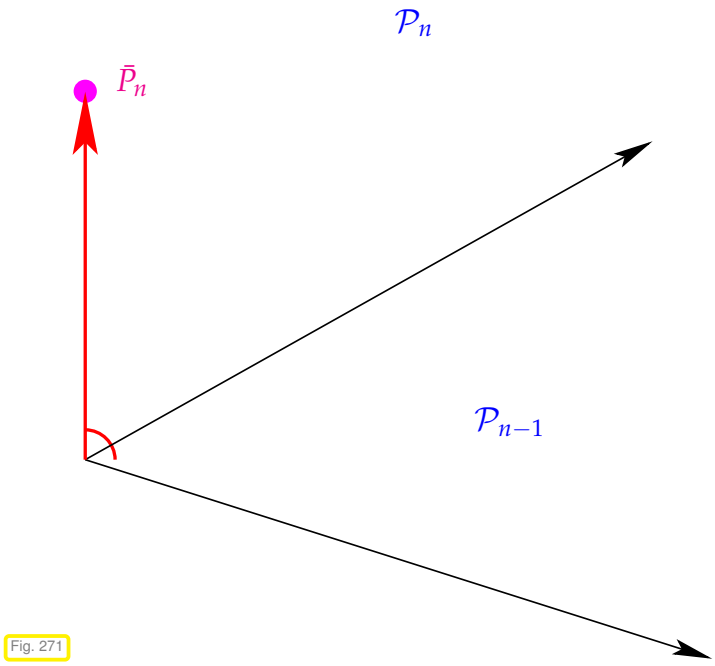


Fig. 271

$\mathcal{P}_n$ equipped with the inner product $(p,q) \mapsto \int_{-1}^{1} p(t)q(t)\,\mathrm{d}t$ can be viewed as an Euclidean space:

$\triangleleft$ $\quad \bar{P}_n \perp \mathcal{P}_{n-1}$

$\mathcal{P}_{n-1} \subset \mathcal{P}_n$ is a subspace of co-dimension 1. Hence, $\mathcal{P}_{n-1}$ has a 1-dimensional orthogonal complement in $\mathcal{P}_n$. By (7.4.2.7) $\bar{P}_n$ belongs to that complement. It takes one additional condition to fix $\bar{P}_n$ and the requirement that its leading coefficient be $= 1$ is that condition.

We can also give algebraic arguments for existence and uniqueness of $\bar{P}_n$. Switching to a monomial representation of $\bar{P}_n$

$$\bar{P}_n(t) = t^n + \alpha_{n-1}t^{n-1} + \cdots + \alpha_1 t + \alpha_0 \,,$$

by linearity of the integral, (7.4.2.7) is equivalent to

$$\begin{aligned} (7.4.2.7) \quad &\Leftrightarrow \quad \int_{-1}^{1} \bigl(t^n + \alpha_{n-1}t^{n-1} + \cdots + \alpha_1 t + \alpha_0\bigr) q(t)\,\mathrm{d}t = 0 \quad \forall q \in \mathcal{P}_{n-1} \\ &\Leftrightarrow \quad \sum_{j=0}^{n-1} \alpha_j \int_{-1}^{1} t^j t^\ell \,\mathrm{d}t = -\int_{-1}^{1} t^n t^\ell \,\mathrm{d}t \,, \quad \ell = 0,\ldots,n-1 \,. \end{aligned}$$

This is a linear system of equations $\mathbf{A}\bigl[\alpha_j\bigr]_{j=0}^{n-1} = \mathbf{b}$ with a symmetric, positive definite ($\rightarrow$ Def. 1.1.2.6) coefficient matrix $\mathbf{A} \in \mathbb{R}^{n,n}$. The $\mathbf{A}$ is positive definite can be concluded from

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} = \int_{-1}^{1} \bigl(\sum_{j=0}^{n-1} (\mathbf{x})_j t^j\bigr)^2 \,\mathrm{d}t > 0 \quad \text{, if} \quad \mathbf{x} \neq 0 \,.$$

Hence, $\mathbf{A}$ is regular and the coefficients α_j are uniquely determined. Thus there is *only one* n-point quadrature rule of order $2n$.

The nodes of an n-point quadrature formula of order $2n$, if it exists, must coincide with the unique zeros of the polynomials $\bar{P}_n \in \mathcal{P}_n \setminus \{0\}$ satisfying (7.4.2.7).

Remark 7.4.2.8 (Gram-Schmidt orthogonalization of polynomials)

Recall: $(f,g) \mapsto \int_a^b f(t)g(t)\,\mathrm{d}t$ is an inner product on $C^0([a,b])$, the L^2-inner product, see Rem. 6.3.2.1, [NS02, Sect. 4.4, Ex. 2], [Gut09, Ex. 6.5]

➢ Treat space of polynomials $\mathcal{P}_n$ as a vector space equipped with an inner product.

➢ As we have seen in Section 6.3.2, abstract techniques for vector spaces with inner product can be applied to polynomials, for instance Gram-Schmidt orthogonalization, *cf.* § 6.3.1.17, [NS02, Thm. 4.8], [Gut09, Alg. 6.1].

Now carry out the abstract Gram-Schmidt orthogonalization according to Algorithm (6.3.1.18) and recall Thm. 6.3.1.19: in a vector space V with inner product $(\cdot,\cdot)_V$ orthogonal vectors $q_0, q_1, \ldots$ spanning the same subspaces as the linearly independent vectors $v_0, v_1, \ldots$ are constructed recursively via

$$q_n := v_n - \sum_{k=0}^{n-1} \frac{(v_n, q_k)_V}{(q_k, q_k)_V} q_k \,, \quad n = 1, 2, \ldots \quad , \quad q_0 := v_0 \,. \tag{7.4.2.9}$$

➢ Construction of $\bar{P}_n$ by Gram-Schmidt orthogonalization of the monomial basis $\{1, t, t^2, \ldots, t^{n-1}\}$ (the v_ks in (7.4.2.9)!) of $\mathcal{P}_{n-1}$ w.r.t. $L^2([-1,1])$-inner product through the recursion

$$\bar{P}_0(t) := 1 \quad , \quad \bar{P}_n(t) = t^n - \sum_{k=0}^{n-1} \frac{\int_{-1}^{1} t^n \bar{P}_k(t)\,\mathrm{d}t}{\int_{-1}^{1} \bar{P}_k^2(t)\,\mathrm{d}t} \cdot \bar{P}_k(t) \tag{7.4.2.10}$$

Note: $\bar{P}_n$ has leading coefficient $= 1 \quad \Rightarrow \quad \bar{P}_n$ uniquely defined (up to sign) by (7.4.2.10). ⌟

The considerations so far only reveal *necessary conditions* on the nodes of an n-point quadrature rule of order $2n$:

They do by no means confirm the existence of such rules, but offer a clear hint on how to construct them:

Theorem 7.4.2.11. Existence of n-point quadrature formulas of order $2n$

Let $\{\bar{P}_n\}_{n\in\mathbb{N}_0}$ be a family of non-zero polynomials that satisfies

- $\bar{P}_n \in \mathcal{P}_n$,
- $\int_{-1}^{1} q(t)\bar{P}_n(t)\,\mathrm{d}t = 0$ *for all* $q \in \mathcal{P}_{n-1}$ *($L^2([-1,1])$-orthogonality),*
- *The set $\{c_j^{(n)}\}_{j=1}^{m}$, $m \le n$, of real zeros of $\bar{P}_n$ is contained in $[-1,1]$.*

Then the quadrature rule (→ Def. 7.2.0.1) $Q_n(f) := \sum_{j=1}^{m} w_j^{(n)} f(c_j^{(n)})$

with weights chosen according to Thm. 7.4.1.6 provides a quadrature formula of order $2n$ on $[-1,1]$.

Proof. Conclude from the orthogonality of the $\bar{P}_n$ that $\{\bar{P}_k\}_{k=0}^{n}$ is a basis of $\mathcal{P}_n$ and

$$\int_{-1}^{1} h(t)\bar{P}_n(t)\,\mathrm{d}t = 0 \quad \forall h \in \mathcal{P}_{n-1} \,. \tag{7.4.2.12}$$

Recall division of polynomials with remainder (Euclid's algorithm → Course "Diskrete Mathematik"): for any $p \in \mathcal{P}_{2n-1}$

$$p(t) = h(t)\bar{P}_n(t) + r(t) \,, \quad \text{for some} \quad h \in \mathcal{P}_{n-1}, r \in \mathcal{P}_{n-1} \,. \tag{7.4.2.13}$$

Apply this representation to the integral:

$$\int_{-1}^{1} p(t)\,\mathrm{d}t = \underbrace{\int_{-1}^{1} h(t)\bar{P}_n(t)\,\mathrm{d}t}_{=0 \text{ by (7.4.2.12)}} + \int_{-1}^{1} r(t)\,\mathrm{d}t \overset{(*)}{=} \sum_{j=1}^{m} w_j^{(n)} r(c_j^{(n)}) \,, \tag{7.4.2.14}$$

$(*)$: by choice of weights according to Rem. 7.4.1.7 Q_n is exact for polynomials of degree $\leq n-1$!

By choice of nodes as zeros of $\bar{P}_n$ using (7.4.2.12):

$$\sum_{j=1}^{m} w_j^{(n)} p(c_j^{(n)}) \overset{(7.4.2.13)}{=} \sum_{j=1}^{m} w_j^{(n)} h(c_j^{(n)}) \underbrace{\bar{P}_n(c_j^{(n)})}_{=0} + \sum_{j=1}^{m} w_j^{(n)} r(c_j^{(n)}) \overset{(7.4.2.14)}{=} \int_{-1}^{1} p(t)\,\mathrm{d}t\,.$$

□

§7.4.2.15 (Legendre polynomials and Gauss-Legendre quadrature) The family of polynomials $\{\bar{P}_n\}_{n\in\mathbb{N}_0}$ are so-called orthogonal polynomials w.r.t. the $L^2(]-1,1[)$-inner product, see Def. 6.3.2.7. We have made use of orthogonal polynomials already in Section 6.3.2. $L^2([-1,1])$-orthogonal polynomials play a key role in analysis.

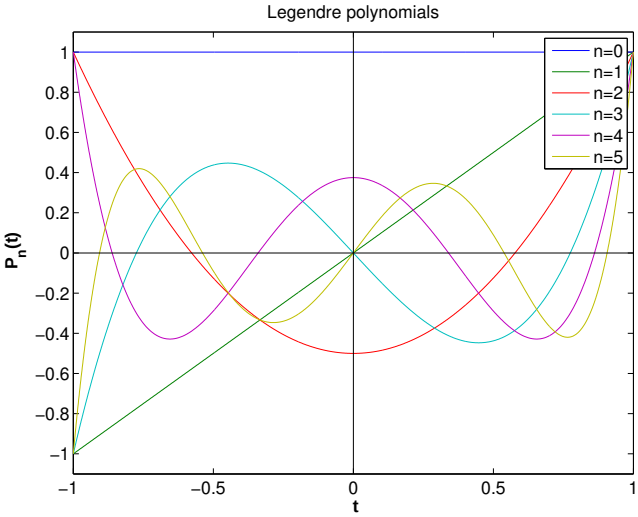
The $L^2(]-1,1[)$-orthogonal are those already discussed in Rem. 6.3.2.16:

Definition 7.4.2.16. Legendre polynomials

The n-th Legendre polynomial P_n is defined by

- $P_n \in \mathcal{P}_n$,
- $\int_{-1}^{1} P_n(t)q(t)\,\mathrm{d}t = 0 \quad \forall q \in \mathcal{P}_{n-1}$,
- $P_n(1) = 1$.

Legendre polynomials $P_0,\ldots,P_5$ ➢ Fig. 272



Notice: the polynomials $\bar{P}_n$ defined by (7.4.2.10) and the Legendre polynomials P_n of Def. 7.4.2.16 (merely) *differ* by a constant factor!

Gauss points $\xi_j^{(n)}$ = zeros of Legendre polynomial P_n

Note: the above considerations, recall (7.4.2.7), show that the nodes of an n-point quadrature formula of order $2n$ on $[-1,1]$ must agree with the zeros of $L^2(]-1,1[)$-orthogonal polynomials.

n-point quadrature formulas of order $2n$ are unique

This is not surprising in light of "$2n$ equations for $2n$ degrees of freedom".

We are not done yet: the zeros of $\bar{P}_n$ from (7.4.2.10) may lie outside $[-1,1]$.
In principle $\bar{P}_n$ could also have less than n real zeros.

The next lemma shows that all this cannot happen.

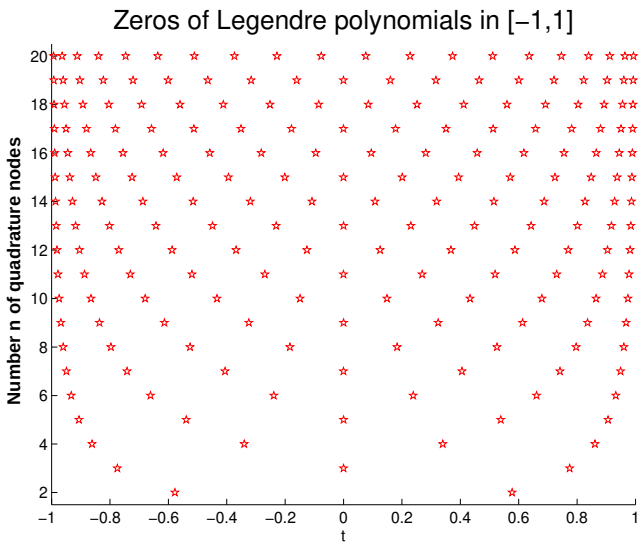


Fig. 273

◁ Obviously:

Lemma 7.4.2.17. Zeros of Legendre polynomials

P_n *has* n *distinct zeros in* $]-1,1[$.

Zeros of Legendre polynomials = Gauss points

Proof. *(indirect)* Assume that P_n has only $m < n$ zeros $\zeta_1, \ldots, \zeta_m$ in $]-1,1[$ *at which it changes sign.* Define

$$q(t) := \prod_{j=1}^{m}(t-\zeta_j) \quad \Rightarrow \quad qP_n \geq 0 \quad \text{or} \quad qP_n \leq 0\,.$$

$$\Rightarrow \quad \int_{-1}^{1} q(t)P_n(t)\,\mathrm{d}t \neq 0\,.$$

As $q \in \mathcal{P}_{n-1}$, this contradicts (7.4.2.12).

□

Definition 7.4.2.18. Gauss-Legendre quadrature formulas

The n-point Quadrature formulas whose nodes, the Gauss points, are given by the zeros of the n-th Legendre polynomial (→ Def. 7.4.2.16), and whose weights are chosen according to Thm. 7.4.1.6, are called **Gauss-Legendre quadrature formulas**.

⌟

The last part of this section examines the non-trivial question of how to compute the Gauss points given as the zeros of Legendre polynomials. Many different algorithms have been devised for this purpose and we focus on one that employs tools from numerical linear algebra and relies on particular algebraic properties of the Legendre polynomials.

Remark 7.4.2.19 (3-Term recursion for Legendre polynomials) From Thm. 6.3.2.14 we learn the orthogonal polynomials satisfy the 3-term recursion (6.3.2.15), see also (7.4.2.21). To keep this chapter self-contained we derive it independently for Legendre polynomials.

Note: the polynomials $\bar{P}_n$ from (7.4.2.10) are uniquely characterized by the two properties (try a proof!)

(i) $\bar{P}_n \in \mathcal{P}_n$ with leading coefficient 1: $\bar{P}(t) = t^n + \ldots$,

(ii) $\int_{-1}^{1} \bar{P}_k(t)\bar{P}_j(t)\,\mathrm{d}t = 0$, if $j \neq k$ ($L^2(]-1,1[)$-orthogonality).

➢ we get the same polynomials $\bar{P}_n$ by another Gram-Schmidt orthogonalization procedure, *cf.* (7.4.2.9) and § 6.3.2.11:

$$\bar{P}_{n+1}(t) = t\bar{P}_n(t) - \sum_{k=0}^{n} \frac{\int_{-1}^{1} \tau\bar{P}_n(\tau)\bar{P}_k(\tau)\,\mathrm{d}\tau}{\int_{-1}^{1} \bar{P}_k^2(\tau)\,\mathrm{d}\tau} \cdot \bar{P}_k(t)$$

By the orthogonality property (7.4.2.12) the sum collapses, since

$$\int_{-1}^{1} \tau\bar{P}_n(\tau)\bar{P}k(\tau)\,\mathrm{d}\tau = \int_{-1}^{1} \bar{P}_n(\tau)\underbrace{(\tau\bar{P}k(\tau))}_{\in\mathcal{P}_{k+1}}\,\mathrm{d}\tau = 0\,,$$

if $k+1 < n$:

$$\bar{P}_{n+1}(t) = t\bar{P}_n(t) - \frac{\int_{-1}^{1} \tau\bar{P}_n(\tau)\bar{P}_n(\tau)\,\mathrm{d}\tau}{\int_{-1}^{1} \bar{P}_n^2(\tau)\,\mathrm{d}\tau} \cdot \bar{P}_n(t) - \frac{\int_{-1}^{1} \tau\bar{P}_n(\tau)\bar{P}_{n-1}(\tau)\,\mathrm{d}\tau}{\int_{-1}^{1} \bar{P}_{n-1}^2(\tau)\,\mathrm{d}\tau} \cdot \bar{P}_{n-1}(t)\,. \tag{7.4.2.20}$$

After rescaling (tedious!) we obtain the famous **3-term recursion** for Legendre polynomials

$$P_{n+1}(t) := \frac{2n+1}{n+1}\,tP_n(t) - \frac{n}{n+1}P_{n-1}(t) \quad , \quad P_0 := 1\,, \quad P_1(t) := t \quad . \tag{7.4.2.21}$$

In Section 6.2.3.1 we discovered a similar 3-term recursion (6.2.3.5) for Chebychev polynomials. Coincidence? Of course not, nothing in mathematics holds "by accident". By Thm. 6.3.2.14 3-term recursions are a distinguishing feature of so-called families of orthogonal polynomials, to which the Chebychev polynomials belong as well, spawned by Gram-Schmidt orthogonalization with respect to a weighted L^2-inner product, however, see [Han02, p. VI].

➤ Efficient and *stable* evaluation of Legendre polynomials by means of 3-term recursion (7.4.2.21), *cf.* the analoguous algorithm for Chebychev polynomials given in Code 6.2.3.6.

C++-code 7.4.2.22: computing Legende polynomials

```
2  // returns the values of the first n - 1 legendre polynomials
3  // in point x as columns of the matrix L
4  void legendre(const unsigned n, const VectorXd& x, MatrixXd& L) {
5    L = MatrixXd::Ones(n,n); // p_0(x) = 1
6    L.col(1) = x;            // p_1(x) = x
7    for (unsigned j = 1; j < n - 1; ++j) {
8      // p_{j+1}(x) = (2j+1)/(j+1) x p_j(x) - j/(j+1) p_{j-1}(x)  (7.4.2.21)
9      L.col(j+1) = (2.*j+1)/(j+1.)*L.col(j-1).cwiseProduct(x)
10       -j/(j+1.)*L.col(j-1);
11   }
12 }
```

Comments on Code 7.4.2.22:

- return value: matrix $\mathbf{L}$ with $(\mathbf{L})_{ij} = P_i(x_j)$
- lines 5-6: take into account initialization of Legendre 3-term recursion (7.4.2.21)

⌟

Remark 7.4.2.23 (Computing Gauss nodes and weights) There are several efficient ways to find the Gauss points. Here we discuss an intriguing connection with an eigenvalue problem.

Compute nodes/weights of Gaussian quadrature by solving an eigenvalue problem!

(Golub-Welsch algorithm [Gan+05, Sect. 3.5.4], [Tre08, Sect. 1])

In codes Gauss nodes and weights are usually retrieved from tables, *cf.* Rem. 7.2.0.6.

C++-code 7.4.2.24: Golub-Welsch algorithm ⇝ GITLAB

```
2  struct QuadRule {
3    Eigen::VectorXd nodes_, weights_;
4  } __attribute__((aligned(32)));
5
6  inline QuadRule gaussquad(const unsigned int n) {
7    QuadRule qr;
8    // Symmetric matrix whose eigenvalues provide Gauss points
9    Eigen::MatrixXd M = Eigen::MatrixXd::Zero(n, n);
10   for (unsigned int i = 1; i < n; ++i) {  //
11     const double b = i / std::sqrt(4. * i * i - 1.);
12     M(i, i - 1) = M(i - 1, i) = b;
13   }  //
14   // using Eigen's built-in solver for symmetric eigenvalue problems
15   const Eigen::SelfAdjointEigenSolver<Eigen::MatrixXd> eig(M);
16
17   qr.nodes_ = eig.eigenvalues();  // Gauss quadrature nodes as eigenvalues!
18   qr.weights_ = 2 * eig.eigenvectors().topRows<1>().array().pow(2);  //
19   return qr;
20 }
```

Justification: rewrite 3-term recurrence (7.4.2.21) for scaled Legendre polynomials $\widetilde{P}_n = \frac{1}{\sqrt{n+1/2}} P_n$

$$t\widetilde{P}_n(t) = \underbrace{\frac{n}{\sqrt{4n^2-1}}}_{=:\beta_n}\widetilde{P}_{n-1}(t) + \underbrace{\frac{n+1}{\sqrt{4(n+1)^2-1}}}_{=:\beta_{n+1}}\widetilde{P}_{n+1}(t)\,. \tag{7.4.2.25}$$

For fixed $t \in \mathbb{R}$ (7.4.2.25) can be expressed as

$$t\underbrace{\begin{bmatrix}\widetilde{P}_0(t)\\ \widetilde{P}_1(t)\\ \vdots\\ \widetilde{P}_{n-1}(t)\end{bmatrix}}_{=:\mathbf{p}(t)\in\mathbb{R}^n} = \underbrace{\begin{bmatrix} 0 & \beta_1 & & & & \\ \beta_1 & 0 & \beta_2 & & & \\ & \beta_2 & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & & 0 & \beta_{n-1}\\ & & & & \beta_{n-1} & 0\end{bmatrix}}_{=:\mathbf{J}_n\in\mathbb{R}^{n,n}} \begin{bmatrix}\widetilde{P}_0(t)\\ \widetilde{P}_1(t)\\ \vdots\\ \widetilde{P}_{n-1}(t)\end{bmatrix} + \begin{bmatrix}0\\ \vdots\\ 0\\ \beta_n\widetilde{P}_n(t)\end{bmatrix}$$

$$\blacktriangleright \quad \widetilde{P}_n(\xi) = 0 \;\Leftrightarrow\; \xi\mathbf{p}(\xi) = \mathbf{J}_n\mathbf{p}(\xi) \quad \text{(an eigenvalue problem!)}\,.$$

$\blacktriangleright$ The zeros of P_n can be obtained as the n real eigenvalues of the symmetric tridiagonal matrix $\mathbf{J}_n \in \mathbb{R}^{n,n}$!

This matrix $\mathbf{J}_n$ is initialized in Line 10–Line 13 of Code 7.4.2.24. The computation of the weights in Line 18 of Code 7.4.2.24 is explained in [Gan+05, Sect. 3.5.4]. ⌟

Remark 7.4.2.26 (Asymptotic methods for the computation of Gauss-Legendre quadrature rules) The fastest methods for the computation of nodes for p-point Gauss-Legendre quadrature rules make use

of asymptotic formulas for the zeros of special functions and achieve an asymptotic complexity of $O(p)$ for $p \to \infty$, see article and [Bog14]. ⌟

Review question(s) 7.4.2.27 (Gauss-Legendre quadrature)

(Q7.4.2.27.A) [Gauss-Radau quadrature] You insist that for a family of p-point quadrature rules on $[-1,1]$ the endpoints $x = -1$ and $x = 1$ belong to the set of quadrature nodes. What order can you achieve?

△

7.4.3 Quadrature Error Estimates

Video tutorial for Section 7.4.3 "(Gauss-Legendre) Quadrature Error Estimates": (18 minutes) Download link, tablet notes

The Gauss-Legendre quadrature formulas do not only enjoy maximal order, but another key property that can be regarded as essential for viable families of quadrature rules.

Obviously ▷

Lemma 7.4.3.1. Positivity of Gauss-Legendre quadrature weights

The weights of the Gauss-Legendre quadrature formulas are positive.

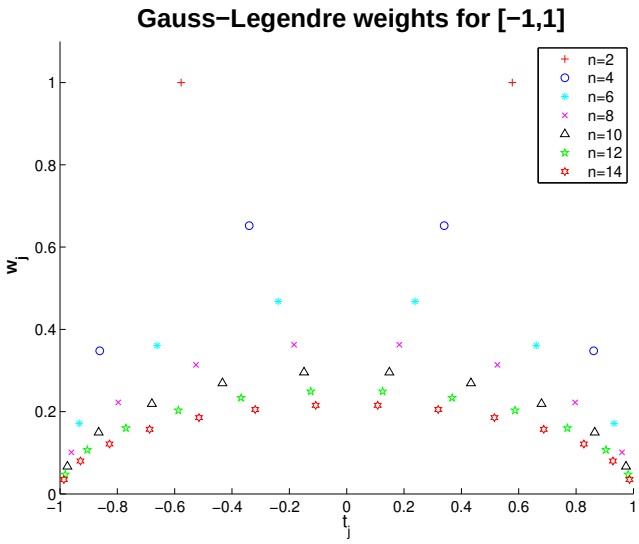


Fig. 274

Proof. Writing $\xi_j^{(n)}$, $j = 1, \ldots, n$, for the nodes (Gauss points) of the n-point Gauss-Legendre quadrature formula, $n \in \mathbb{N}$, we define

$$q_k(t) = \prod_{\substack{j=1 \\ j \neq k}}^{n} (t - \xi_j^{(n)})^2 \quad \Rightarrow \quad q_k \in \mathcal{P}_{2n-2} \, .$$

This polynomial is integrated exactly by the quadrature rule: since $q_k(\xi_j^{(n)}) = 0$ for $j \neq k$

$$0 < \int_{-1}^{1} q(t) \, \mathrm{d}t = w_k^{(n)} \underbrace{q(\xi_k^{(n)})}_{>0} \, ,$$

where $w_j^{(n)}$ are the quadrature weights. □

§7.4.3.2 (Quadrature error and best approximation error) The positivity of the weights $w_j^{(n)}$ for all n-point Gauss-Legendre and Clenshaw-Curtis quadrature rules has important consequences.

Theorem 7.4.3.3. Quadrature error estimate for quadrature rules with positive weights

For every n-point quadrature rule Q_n as in (7.2.0.2) of order $q \in \mathbb{N}$ with weights $w_j \geq 0$, $j = 1,\ldots,n$ the quadrature error satisfies

$$E_n(f) := \left| \int_a^b f(t)\,dt - Q_n(f) \right| \leq 2|b-a| \underbrace{\inf_{p\in\mathcal{P}_{q-1}} \|f-p\|_{L^\infty([a,b])}}_{\text{best approximation error}} \quad \forall f \in C^0([a,b])\,. \tag{7.4.3.4}$$

Proof. The proof runs parallel to the derivation of (6.2.2.33). Writing $E_n(f)$ for the quadrature error, the left hand side of (7.4.3.4), we find by the definition Def. 7.4.1.1 of the order of a quadrature rule that *for all* $p \in \mathcal{P}_{q-1}$

$$\begin{aligned} E_n(f) = E_n(f-p) &\leq \left| \int_a^b (f-p)(t)\,\mathrm{d}t \right| + \left| \sum_{j=1}^n w_j (f-p)(c_j) \right| \\ &\leq |b-a| \|f-p\|_{L^\infty([a,b])} + \Big(\sum_{j=1}^n |w_j|\Big) \|f-p\|_{L^\infty([a,b])}\,. \end{aligned} \tag{7.4.3.5}$$

Since the quadrature rule is exact for constants and $w_j \geq 0$

$$\sum_{j=1}^n |w_j| = \sum_{j=1}^n w_j = |b-a|\,,$$

which finishes the proof. □

Drawing on best approximation estimates from Section 6.2.1 and Rem. 6.2.3.26, we immediately get results about the asymptotic decay of the quadrature error for n-point Gauss-Legendre and Clenshaw-Curtis quadrature as $n \to \infty$:

$$\begin{aligned} &f \in C^r([a,b]) && \Rightarrow \quad E_n(f) \to 0 \text{ algebraically with rate } r, \\ &f \in C^\infty([a,b]) \text{ "analytically extensible"} && \Rightarrow \quad E_n(f) \to 0 \text{ exponentially}, \end{aligned}$$

as $n \to \infty$, see Def. 6.2.2.7 for type of convergence.

Appealing to (6.2.1.27) and (6.2.2.22), the dependence of the constants on the length of the integration interval can be quantified for integrands with limited smoothness.

Lemma 7.4.3.6. Quadrature error estimates for C^r-integrands

For every n-point quadrature rule Q_n as in (7.2.0.2) of order $q \in \mathbb{N}$ with weights $w_j \geq 0$, $j = 1,\ldots,n$ we find that the quadrature error $E_n(f)$ for and integrand $f \in C^r([a,b])$, $r \in \mathbb{N}_0$, satisfies

$$\text{in the case } q \geq r: \quad E_n(f) \leq C\,(q-1)^{-r} |b-a|^{r+1} \left\| f^{(r)} \right\|_{L^\infty([a,b])}\,, \tag{7.4.3.7}$$

$$\text{in the case } q < r: \quad E_n(f) \leq \frac{|b-a|^{q+1}}{q!} \left\| f^{(q)} \right\|_{L^\infty([a,b])}\,, \tag{7.4.3.8}$$

with a constant $C > 0$ independent of n, f, and $[a,b]$.

Proof. The first estimate (7.4.3.7) is immediate from (6.2.1.27). The second bound (7.4.3.8) is obtained by combining (7.4.3.4) and (6.2.2.22).

□

Please note the different estimates depending on whether the smoothness of f (as described by r) or the order of the quadrature rule is the "limiting factor". ⌟

EXAMPLE 7.4.3.9 (Convergence of global quadrature rules)

We examine three families of global polynomial (→ Thm. 7.4.1.6) quadrature rules: Newton-Cotes formulas, Gauss-Legendre rules, and Clenshaw-Curtis rules. We record the convergence of the quadrature errors for the interval $[0,1]$ and two different functions

1. $f_1(t) = \frac{1}{1+(5t)^2}$, an analytic function, see Rem. 6.2.2.67,
2. $f_2(t) = \sqrt{t}$, merely continuous, derivatives singular in $t = 0$.

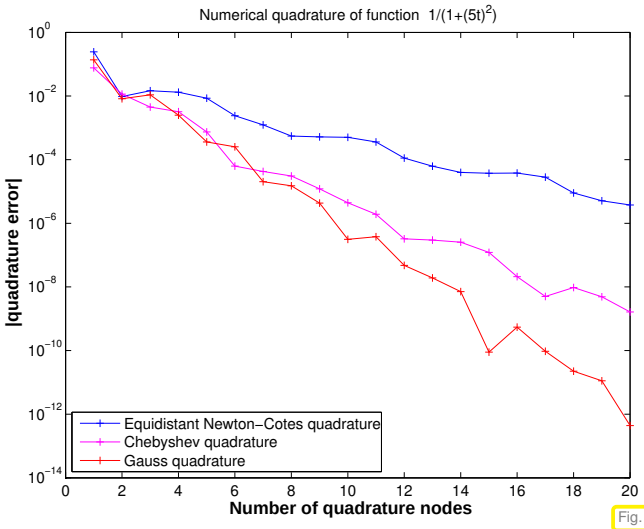


Fig. 275

quadrature error, $f_1(t) := \frac{1}{1+(5t)^2}$ on $[0,1]$

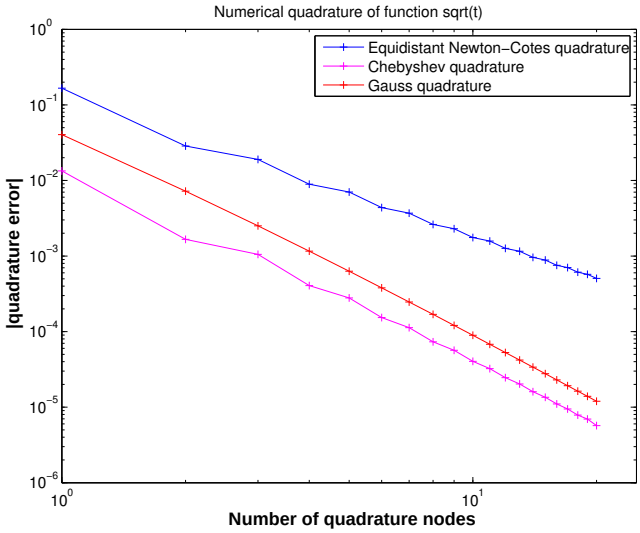


Fig. 276

quadrature error, $f_2(t) := \sqrt{t}$ on $[0,1]$

Asymptotic behavior of quadrature error $\epsilon_n := \left| \int_0^1 f(t)\,\mathrm{d}t - Q_n(f) \right|$ for "$n \to \infty$":

➢ exponential convergence $\epsilon_n \approx O(q^n)$, $0 < q < 1$, for C^∞-integrand f_1 ⇝ : Newton-Cotes quadrature : $q \approx 0.61$, Clenshaw-Curtis quadrature : $q \approx 0.40$, Gauss-Legendre quadrature : $q \approx 0.27$

➢ algebraic convergence $\epsilon_n \approx O(n^{-\alpha})$, $\alpha > 0$, for integrand f_2 with singularity at $t = 0$ ⇝ Newton-Cotes quadrature : $\alpha \approx 1.8$, Clenshaw-Curtis quadrature : $\alpha \approx 2.5$, Gauss-Legendre quadrature : $\alpha \approx 2.7$

⌟

Remark 7.4.3.10 (Removing a singularity by transformation) From Ex. 7.4.3.9 teaches us that a lack of smoothness of the integrand can thwart exponential convergence and severely limits the rate of algebraic convergence of a global quadrature rule for $n \to \infty$.

Idea: recover integral with smooth integrand by "analytic preprocessing"

Here is an example:

For a general but smooth $f \in C^\infty([0,b])$ compute $\int_0^b \sqrt{t} f(t)\,\mathrm{d}t$ via a quadrature rule, e.g., n-point Gauss-Legendre quadrature on $[0,b]$. Due to the presence of a square-root singularity at $t = 0$ the direct

application of n-point Gauss-Legendre quadrature will result in a rather slow algebraic convergence of the quadrature error as $n \to \infty$, see Ex. 7.4.3.9.

Trick: Transformation of integrand by substitution rule:

$$\text{substitution } s = \sqrt{t}: \quad \int_0^b \sqrt{t} f(t)\,\mathrm{d}t = \int_0^{\sqrt{b}} \boxed{2s^2 f(s^2)}\,\mathrm{d}s\,. \tag{7.4.3.11}$$

Then: Apply Gauss-Legendre quadrature rule to smooth integrand

Remark 7.4.3.12 (The message of asymptotic estimates) There is one blot on most n-asymptotic estimates obtained from Thm. 7.4.3.3: the bounds usually involve quantities like norms of higher derivatives of the interpoland that are elusive in general, in particular for integrands given only in procedural form, see § 7.1.0.2. Such unknown quantities are often hidden in "generic constants C". Can we extract useful information from estimates marred by the presence of such constants?

For fixed integrand f let us assume *sharp* algebraic convergence (in n) with rate $r \in \mathbb{N}$ of the quadrature error $E_n(f)$ for a family of n-point quadrature rules:

$$E_n(f) = O(n^{-r}) \overset{\text{sharp}}{\Longrightarrow} E_n(f) \approx Cn^{-r}\,, \tag{7.4.3.13}$$

with a "generic constant $C > 0$" independent of n.

Goal: Reduction of the quadrature error by a factor of $\rho > 1$

Which (minimal) increase in the number n of quadrature points accomplishes this?

$$\frac{Cn_{\text{old}}^{-r}}{Cn_{\text{new}}^{-r}} \overset{!}{=} \rho \quad \Leftrightarrow \quad \boxed{n_{\text{new}} : n_{\text{old}} = \sqrt[r]{\rho}}\,. \tag{7.4.3.14}$$

In the case of *algebraic convergence* with rate $r \in \mathbb{R}$ a reduction of the quadrature error by a factor of ρ is bought by an increase of the number of quadrature points by a factor of $\rho^{1/r}$.

(7.4.3.7) ➢ gains in accuracy are "cheaper" for smoother integrands!

Now assume *sharp* exponential convergence (in n) of the quadrature error $E_n(f)$ for a family of n-point quadrature rules, $0 \le q < 1$:

$$E_n(f) = O(q^n) \overset{\text{sharp}}{\Longrightarrow} E_n(f) \approx Cq^n\,, \tag{7.4.3.15}$$

with a "generic constant $C > 0$" independent of n.

Error reduction by a factor $\rho > 1$ results from

$$\frac{Cq^{n_{\text{old}}}}{Cq^{n_{\text{new}}}} \overset{!}{=} \rho \quad \Leftrightarrow \quad \boxed{n_{\text{new}} - n_{\text{old}} = -\frac{\log \rho}{\log q}}\,.$$

In the case of *exponential convergence* (7.4.3.15) a fixed increase of the number of quadrature points by $-\log\rho : \log q$ results in a reduction of the quadrature error by a factor of $\rho > 1$.

⌟

Review question(s) 7.4.3.16 (Quadrature error estimates)

(Q7.4.3.16.A) By the substitution $s = \sqrt{t}$ we could transform

$$\int_0^b \sqrt{t} f(t)\,\mathrm{d}t = 2\int_0^{\sqrt{b}} s^2 f(s^2)\,\mathrm{d}s\,.$$

Describe the weights $w_j^{(n)}$ and nodes $c_j^{(n)}$, $j = 1,\ldots,n$, $n \in \mathbb{N}$, of a family of quadrature rules satisfying

$$\int_0^b \sqrt{t} f(t)\,\mathrm{d}t \approx \sum_{j=1}^n w_j^{(n)} f(c_j^{(n)})\,,$$

and enjoying n-asymptotic exponential convergence, if f possess an analytic extension beyond $[0,b]$.

(Q7.4.3.16.B) Let $(Q_n)_{n\in\mathbb{N}}$ denote a family of quadrature rules with n the number of quadrature points. For the approximate evaluation of $\int_a^b f(t)\,\mathrm{d}t$ the following adaptive algorithm is employed

```
n := 0;
do
  n := n+1;
while (|Q_{n+1}(f) − Q_n(f)| ≤ tol · |Q_{n+1}(f)|);
return Q_{n+1}(f);
```

Assuming a sharp asymptotic behavior like $E_n(f) = O(n^{-r})$, $r \in \mathbb{N}$, (algebraic convergence with rate r) for $n \to \infty$ of the quadrature error $E_n(f)$, how much extra work may be incurred when reducing the tolerance by a factor of 10.

(Q7.4.3.16.C) [Improper integral with logarithmic weight] We consider the improper integral

$$I(f) := \int_0^1 \log(t) f(t)\,\mathrm{d}t\,, \tag{7.4.3.17}$$

where $f : [0,1] \to \mathbb{R}$ is supposed to possess an analytic extension to a $\mathbb{C}$-neighborhood of $[0,1]$.

- Is it possible to apply the n-point Gauss-Legendre quadrature formula on $[0,1]$ to (7.4.3.17)?
- What is your prediction for the behavior of the quadrature error of the n-point Gauss-Legendre quadrature formula on $[0,1]$ applied to (7.4.3.17) for $n \to \infty$?

(Q7.4.3.16.D) ["Regularization" of integrands by transformation] Transform (7.4.3.17) by the change of variables ("substitution")

$$t = \varphi(\tau)\,,\quad \varphi(\tau) := \sin^4(\tfrac{\pi}{2}\tau)\,,\quad \tau \in [0,1]\,.$$

When applying the n-point Gauss-Legendre quadrature formula on $[0,1]$ to the transformed integral, how will the quadrature error behave for $n \to \infty$?

Hint. Examine the differentiability properties of the transformed integrand in $\tau = 0$.

Hint. Remember the product rule for the k-th derivative

$$(fg)^{(k)}(\tau) = \sum_{j=0}^{k} \binom{k}{j} f^{(j)}(\tau) g^{(k-j)}(\tau) , \quad f,g \in C^k . \tag{7.4.3.18}$$

Hint. How does $\tau \mapsto \varphi(\tau)$ and its derivatives behave as $\tau \to 0$: $\varphi^{(k)}(\tau) = O(\tau^?)$ for $\tau \to 0$.

△

7.5 Composite Quadrature

Video tutorial for Section 7.5 "Composite Quadrature": (18 minutes) Download link, tablet notes

In Chapter 6, Section 6.6.1 we studied approximation by piecewise polynomial interpolants. A similar idea underlies the so-called composite quadrature rules on an interval $[a,b]$. Analogously to piecewise polynomial techniques they start from a grid/mesh

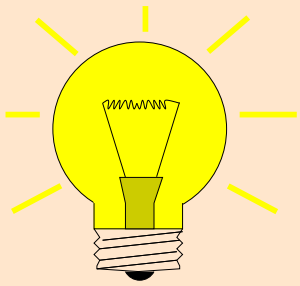
$$\mathcal{M} := \{a = x_0 < x_1 < \ldots < x_{m-1} < x_m = b\} \tag{6.6.0.2}$$

and appeal to the trivial identity

$$\int_a^b f(t)\,\mathrm{d}t = \sum_{j=1}^{m} \int_{x_{j-1}}^{x_j} f(t)\,\mathrm{d}t . \tag{7.5.0.1}$$

On each mesh interval $[x_{j-1}, x_j]$ we then use a local quadrature rule, which may be one of the polynomial quadrature formulas from 7.3.

General construction of composite quadrature rules



Idea:
- Partition integration domain $[a,b]$ by a mesh/grid (→ Section 6.6)
  $$\mathcal{M} := \{a = x_0 < x_1 < \ldots < x_m = b\}$$
- Apply quadrature formulas from Section 7.3, Section 7.4 locally on mesh intervals $I_j := [x_{j-1}, x_j]$, $j = 1, \ldots, m$, and sum up.

composite quadrature rule

EXAMPLE 7.5.0.3 (Simple composite polynomial quadrature rules)

Composite trapezoidal rule, *cf.* (7.3.0.5)

$$\int_a^b f(t)\mathrm{d}t = \tfrac{1}{2}(x_1 - x_0)f(a) + \sum_{j=1}^{m-1} \tfrac{1}{2}(x_{j+1} - x_{j-1})f(x_j) + \tfrac{1}{2}(x_m - x_{m-1})f(b) . \tag{7.5.0.4}$$

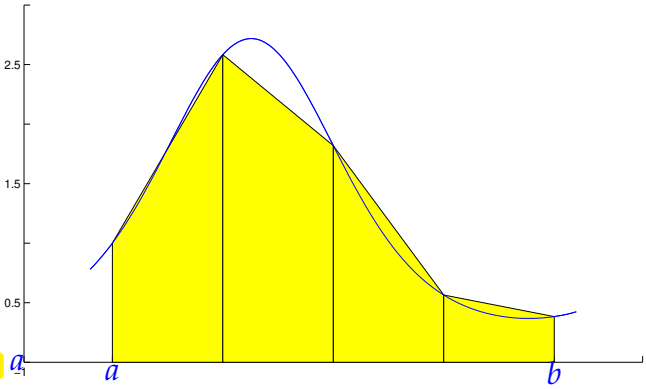


Fig. 277

➢ arising from piecewise linear interpolation of f.

Composite Simpson rule, *cf.* (7.3.0.6)

$$\int_a^b f(t)\mathrm{d}t = \frac{1}{6}(x_1 - x_0)f(a) + \sum_{j=1}^{m-1} \frac{1}{6}(x_{j+1} - x_{j-1})f(x_j) + \sum_{j=1}^{m} \frac{2}{3}(x_j - x_{j-1})f(\tfrac{1}{2}(x_j + x_{j-1})) + \frac{1}{6}(x_m - x_{m-1})f(b)\ . \tag{7.5.0.5}$$

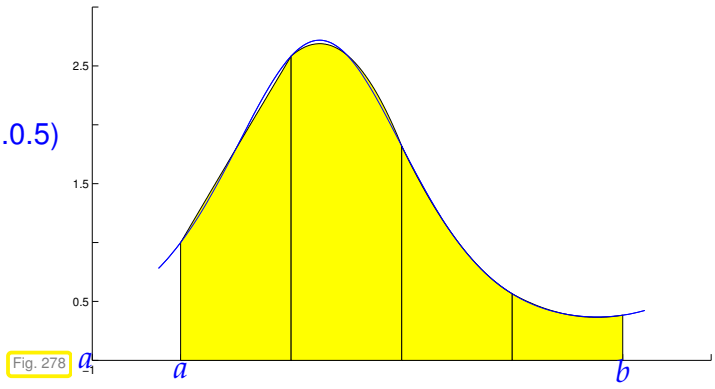


Fig. 278

➢

related to piecewise quadratic Lagrangian interpolation.

Formulas (7.5.0.4), (7.5.0.5) directly suggest efficient implementation with minimal number of f-evaluations.

C++-code 7.5.0.6: Equidistant composite trapezoidal rule (7.5.0.4)

```
#include <cassert>

// N-interval equidistant trapezoidal rule
template <class Function>
double trapezoidal(Function &f, const double a, const double b,
                   const unsigned N) {
  double I = 0;
  const double h = (b - a) / N; // interval length

  for (unsigned i = 0; i < N; ++i) {
    // rule:  T = (b - a)/2 * (f(a) + f(b)),
    // apply on N intervals:  [a + i*h, a + (i+1)*h], i=0..(N-1)
    I += h / 2 * (f(a + i * h) + f(a + (i + 1) * h));
  }
  return I;
}

// Alternative implementation of n-point equidistant trapezoidal rule
template <typename Functor>
double equidTrapezoidalRule(Functor &&f, double a, double b, unsigned int n) {
  assert(n>=2);
  const double h = (b - a)/(n-1);
  double t = a + h;
  double s = 0.0;
  for (unsigned int i = 1; i < n-1; t += h, ++i) {
    s += f(t);
  }
  return (0.5 * h * f(a) + 0.5 * h * f(b) + h * s);
}
```

C++-code 7.5.0.7: Equidistant composite Simpson rule (7.5.0.5)

```
template <class Function>
```

```
double simpson(Function& f, const double a, const double b, const unsigned N) {
  double I = 0;
  const double h = (b - a)/N; // intervall length

  for (unsigned i = 0; i < N; ++i) {
    // rule:  S = (b - a)/6*( f(a) + 4*f(0.5*(a + b)) + f(b) )
    // apply on [a + i*h, a + (i+1)*h]
    I += h/6*( f(a + i*h) + 4*f(a + (i+0.5)*h) + f(a + (i+1)*h) );
  }

  return I;
}
```

In both cases the function object passed in `f` must provide an evaluation operator **double operator** (**double**) **const**. ⌟

Remark 7.5.0.8 (Composite quadrature and piecewise polynomial interpolation) Composite quadrature scheme based on local polynomial quadrature can usually be understood as "quadrature by approximation schemes" as explained in § 7.2.0.7. The underlying approximation schemes belong to the class of general local Lagrangian interpolation schemes introduced in Section 6.6.1.

In other words, many composite quadrature schemes arise from replacing the integrand by a piecewise interpolating polynomial, see Fig. 277 and Fig. 278 and compare with Fig. 256. ⌟

To see the main rationale behind the use of composite quadrature rules recall Lemma 7.4.3.6: for a polynomial quadrature rule (7.3.0.1) of order q with positive weights and $f \in C^r([a,b])$ the quadrature error shrinks with the $\min\{r,q\}+1$-st power of the length $|b-a|$ of the integration domain! Hence, applying polynomial quadrature rules to small mesh intervals should lead to a small overall quadrature error.

§7.5.0.9 (Quadrature error estimate for composite polynomial quadrature rules) Assume a composite quadrature rule Q on $[x_0, x_m] = [a,b]$, $b > a$, based on n_j-point local quadrature rules $Q^j_{n_j}$ with *positive weights* (e.g. local Gauss-Legendre quadrature rules or local Clenshaw-Curtis quadrature rules) and of *fixed* orders $q_j \in \mathbb{N}$ on each mesh interval $[x_{j-1}, x_j]$. From Lemma 7.4.3.6 recall the estimate for $f \in C^r([x_{j-1}, x_j])$

$$\left| \int_{x_{j-1}}^{x_j} f(t)\,\mathrm{d}t - Q^j_{n_j}(f) \right| \leq C|x_j - x_{j-1}|^{\min\{r,q_j\}+1} \left\| f^{(\min\{r,q_j\})} \right\|_{L^\infty([x_{j-1},x_j])} . \tag{7.3.0.11}$$

with $C > 0$ independent of f and j. For $f \in C^r([a,b])$, summing up these bounds we get for the global quadrature error

$$\left| \int_{x_0}^{x_m} f(t)\,\mathrm{d}t - Q(f) \right| \leq C \sum_{j=1}^{m} h_j^{\min\{r,q_j\}+1} \left\| f^{(\min\{r,q_j\})} \right\|_{L^\infty([x_{j-1},x_j])} ,$$

with local meshwidths $h_j = x_j - x_{j-1}$. If $q_j = q$, $q \in \mathbb{N}$, for all $j = 1,\ldots,m$, then, as $\sum_j h_j = b-a$,

$$\left| \int_{x_0}^{x_m} f(t)\,\mathrm{d}t - Q(f) \right| \leq C\, h_{\mathcal{M}}^{\min\{q,r\}}\, |b-a| \left\| f^{(\min\{q,r\})} \right\|_{L^\infty([a,b])} , \tag{7.5.0.10}$$

with (global) meshwidth $h_{\mathcal{M}} := \max_j h_j$.

(7.5.0.10) $\longleftrightarrow$ Algebraic convergence in no. of f-evaluations for $n \to \infty$

⌟

§7.5.0.11 (Constructing families of composite quadrature rules) As with polynomial quadrature rules, we study the asymptotic behavior of the quadrature error for families of composite quadrature rules as a function on the total number n of function evaluations.

As in the case of $\mathcal{M}$-piecewise polynomial approximation of function ($\to$ Section 6.6.1) families of composite quadrature rules can be generated in two different ways:

(I) use a sequence of successively refined meshes $\left(\mathcal{M}_k = \{x_j^k\}_j\right)_{k\in\mathbb{N}}$ with $\sharp\mathcal{M} = m(k)+1$, $m(k) \to \infty$ for $k \to \infty$, , combined with the *same* (transformed, $\to$ Rem. 7.2.0.4) local quadrature rule on all mesh intervals $[x_{j-1}^k, x_j^k]$. Examples are the composite trapezoidal rule and composite Simpson rule from Ex. 7.5.0.3 on sequences of equidistant meshes.
➢ h-convergence

(II) On a *fixed* mesh $\mathcal{M} = \{x_j\}_{j=0}^m$, on each cell use the same (transformed) local quadrature rule taken from a *sequence* of polynomial quadrature rules of increasing order.
➢ p-convergence

⌟

EXPERIMENT 7.5.0.12 (Quadrature errors for composite quadrature rules) Composite quadrature rules based on

- trapezoidal rule (7.3.0.5) ➢ local order 2 (exact for linear functions, see Ex. 7.4.1.10),
- Simpson rule (7.3.0.6) ➢ local order 4 (exact for cubic polynomials, see Ex. 7.4.1.10)

on equidistant mesh $\mathcal{M} := \{jh\}_{j=0}^n$, $h = 1/n$, $n \in \mathbb{N}$.

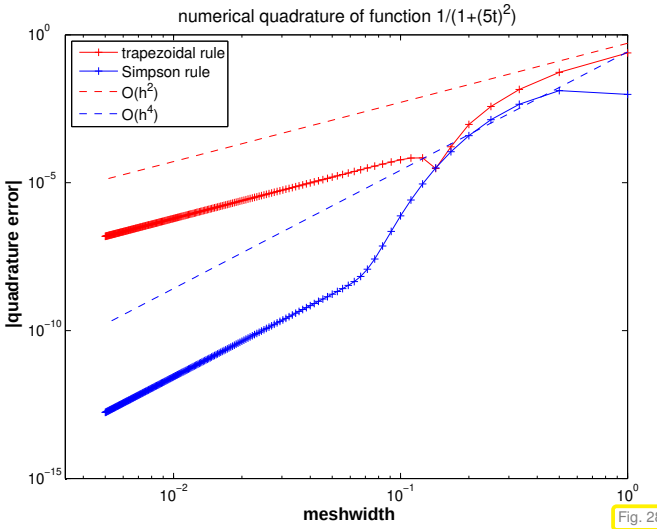


Fig. 279

quadrature error, $f_1(t) := \frac{1}{1+(5t)^2}$ on $[0,1]$

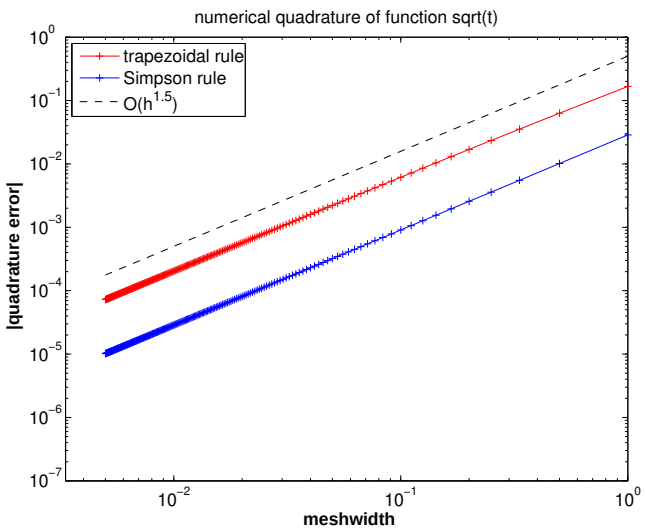


Fig. 280

quadrature error, $f_2(t) := \sqrt{t}$ on $[0,1]$

Asymptotic behavior of quadrature error $E(n) := \left|\int_0^1 f(t)\,\mathrm{d}t - Q_n(f)\right|$ for meshwidth "$h \to 0$":

☛ Throughout we observe algebraic convergence $E(n) = O(h^\alpha)$ of with rate $\alpha > 0$ for $h = n^{-1} \to 0$

➢ Sufficiently smooth integrand f_1: trapezoidal rule $\to \alpha = 2$, Simpson rule $\to \alpha = 4$!?

➢ singular integrand f_2: $\alpha = 3/2$ for trapezoidal rule **&** Simpson rule **!**

(lack of) smoothness of integrand limits convergence!

⌟

Remark 7.5.0.13 (Composite quadrature rules vs. global quadrature rules) For a fixed integrand $f \in C^r([a,b])$ of limited smoothness on an interval $[a,b]$ we compare

- a family of composite quadrature rules based on single local ℓ-point rule (with positive weights) of order q on a sequence of equidistant meshes $\left(\mathcal{M}_k = \{x_j^k\}_j\right)_{k\in\mathbb{N}}$,
- the family of Gauss-Legendre quadrature rules from Def. 7.4.2.18.

We study the asymptotic dependence of the quadrature error on the number n of function evaluations.

For the composite quadrature rules we have $n \approx \ell \sharp \mathcal{M}_k \approx \ell h_{\mathcal{M}}^{-1}$. Combined with (7.5.0.10), we find for quadrature error $E_n^{\mathrm{comp}}(f)$ of the composite quadrature rules

$$E_n^{\mathrm{comp}}(f) \le C_1\, n^{-\min\{q,r\}}\,, \tag{7.5.0.14}$$

with $C_1 > 0$ independent of $\mathcal{M} = \mathcal{M}_k$.

The quadrature errors $E_n^{\mathrm{GL}}(f)$ of the n-point Gauss-Legendre quadrature rules are given in Lemma 7.4.3.6, (7.4.3.7):

$$E_n^{\mathrm{GL}}(f) \le C_2\, n^{-r}\,, \tag{7.5.0.15}$$

with $C_2 > 0$ independent of n.

Gauss-Legendre quadrature converges *at least as fast* fixed order composite quadrature on equidistant meshes.

Moreover, Gauss-Legendre quadrature "automatically detects" the smoothness of the integrand, and enjoys fast exponential convergence for analytic integrands.

Use Gauss-Legendre quadrature instead of fixed order composite quadrature on equidistant meshes.

⌟

EXPERIMENT 7.5.0.16 (Empiric convergence of equidistant trapezoidal rule) Sometimes there are surprises: Now we will witness a convergence behavior of a composite quadrature rule that is much better than predicted by the order of the local quadrature formula.

We consider the equidistant trapezoidal rule (order 2), see (7.5.0.4), Code 7.5.0.6

$$\int_a^b f(t)\,\mathrm{d}t \approx T_m(f) := h\left(\tfrac{1}{2}f(a) + \sum_{k=1}^{m-1} f(a+kh) + \tfrac{1}{2}f(b)\right)\,,\quad h := \frac{b-a}{m}\,. \tag{7.5.0.17}$$

and the 1-periodic smooth (analytic) integrand

$$f(t) = \frac{1}{\sqrt{1 - a\sin(2\pi t - 1)}}\,,\quad 0 < a < 1\,.$$

(As "exact value of integral" we use T_{500} in the computation or quadrature errors.)

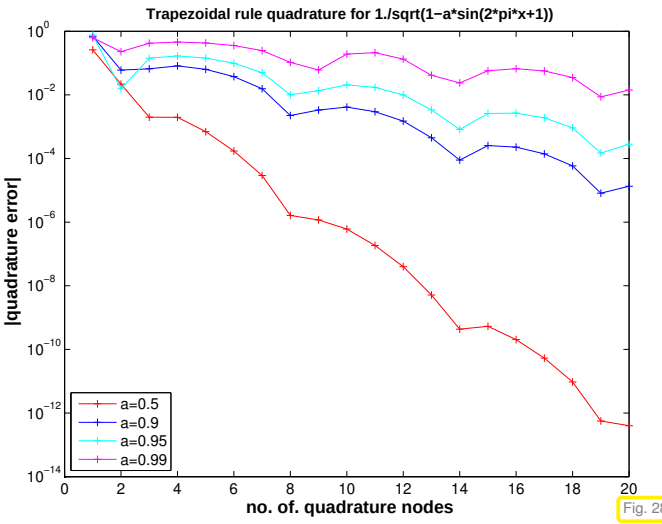


Fig. 281

quadrature error for $T_n(f)$ on $[0,1]$

exponential convergence !!

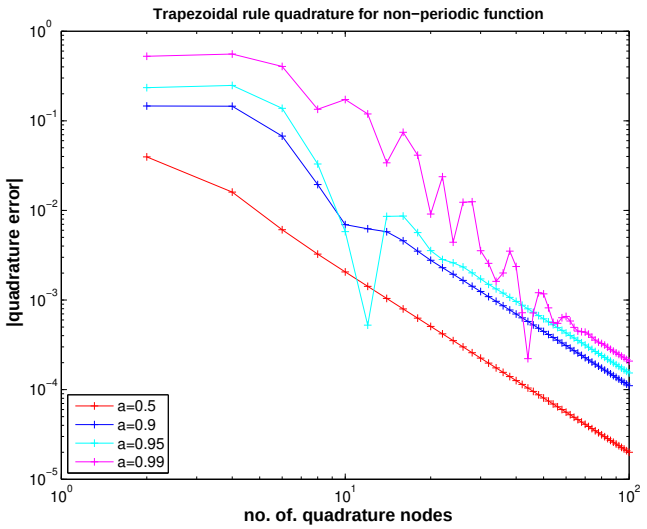


Fig. 282

quadrature error for $T_n(f)$ on $[0,\frac{1}{2}]$

merely algebraic convergence

§7.5.0.18 (The magic of the equidistant trapezoidal rule (for periodic integrands))

In this § we use $I := [0,1[$ as a reference interval, *cf.* Exp. 7.5.0.16. We rely on similar techniques as in Section 5.6, Section 5.6.2. Again, a key tool will be the bijective mapping, see Fig. 202,

$$\Phi_{\mathbb{S}^1} : I \to \mathbb{S}^1 := \{z \in \mathbb{C} : |z| = 1\} \quad , \quad t \mapsto z := \exp(2\pi\imath t) \, , \tag{5.6.2.1}$$

which induces the general pullback, *c.f.* (6.2.1.16),

$$(\Phi_{\mathbb{S}^1}^{-1})^* : C^0([0,1[) \to C^0(\mathbb{S}^1) \quad , \quad ((\Phi_{\mathbb{S}^1}^{-1})^* f)(z) := f(\Phi_{\mathbb{S}^1}^{-1}(z)) \, , \quad z \in \mathbb{S}^1 \, .$$

If $f \in C^r(\mathbb{R})$ and 1-periodic, then $(\Phi_{\mathbb{S}^1}^{-1})^* f \in C^r(\mathbb{S}^1)$. Further, $\Phi_{\mathbb{S}^1}$ maps equidistant nodes on $I := [0,1]$ to equispaced nodes on $\mathbb{S}^1$, which are the roots of unity:

$$\Phi_{\mathbb{S}^1}(\tfrac{j}{n}) = \exp(2\pi\imath\tfrac{j}{n}) \quad [\, \exp(2\pi\imath\tfrac{j}{n})^n = 1; \,] \, . \tag{7.5.0.19}$$

Now consider an n-point polynomial quadrature rule on $\mathbb{S}^1$ based on the set of equidistant nodes $\mathcal{Z} := \{z_j := \exp(2\pi\imath\frac{j-1}{n}), j = 1,\ldots,n\}$ and defined as

$$Q_n^{\mathbb{S}^1}(g) := \int_{\mathbb{S}^1} (\mathsf{L}_{\mathcal{Z}} g)(\tau) \, \mathrm{d}S(\tau) = \sum_{j=1}^n w_j^{\mathbb{S}^1} g(z_j) \, , \tag{7.5.0.20}$$

where $\mathsf{L}_{\mathcal{Z}}$ is the Lagrange interpolation operator ($\to$ Def. 6.2.2.1). This means that the weights obey Thm. 7.4.1.6, where the definition (5.2.2.4) of Lagrange polynomials remains the same for complex nodes. By sheer *symmetry*, all the weights have to be the same, which, since the rule will be at least of order 1, means

$$w_j^{\mathbb{S}^1} = \frac{2\pi}{n} \, , \quad j = 1,\ldots,n \, .$$

Moreover, the quadrature rule $Q_n^{\mathbb{S}^1}$ will be of order n, see Def. 7.4.1.1, that is, it will *integrate polynomials of degree* $\leq n-1$, *exactly*.

By transformation ($\to$ Rem. 7.2.0.4 and pullback (7.5.0.18), $Q_n^{\mathbb{S}^1}$ induces a quadrature rule on $I := [0,1]$ by

$$Q_n^I(f) := \frac{1}{2\pi} Q_n^{\mathbb{S}^1}((\Phi_{\mathbb{S}^1}^{-1})^* f) = \frac{1}{2\pi} \sum_{j=1}^n w_j^{\mathbb{S}^1} f(\Phi^{-1}(z_j)) = \sum_{j=1}^n \frac{1}{n} f(\tfrac{j-1}{n}) \, . \tag{7.5.0.21}$$

This is exactly the equidistant trapezoidal rule(7.5.0.17), if f is 1-periodic, $f(0) = f(1)$: $Q_n^I = T_n$. Hence we arrive at the following estimate for the quadrature error

$$E_n(f) := \left| \int_0^1 f(t)\,\mathrm{d}t - T_n(f) \right| \leq 2\pi \max_{z\in S^1} \left| \left((\Phi_{S^1}^{-1})^* f\right)(z) - \left(\mathsf{L}_{\mathcal{Z}}(\Phi_{S^1}^{-1})^* f\right)(z) \right| .$$

Equivalently, one can show that T_n integrates trigonometric polynomials up to degree $2n-1$ exactly. Remember from Section 5.6.1 that the $2n+1$-dimensional space of 1-periodic trigonometric polynomials of degree $2n$ can be defined as

$$\mathcal{P}_{2n}^T := \mathrm{Span}\{t \mapsto \exp(2\pi\imath j t) : \ j = -n, \ldots, n\} .$$

By elementary computations we find:

$$f(t) = e^{2\pi\imath k t} \quad \blacktriangleright \quad \begin{cases} \int_0^1 f(t)\,\mathrm{d}t = \begin{cases} 0 & \text{, if } k \neq 0 , \\ 1 & \text{, if } k = 0 . \end{cases} \\ T_n(f) = \frac{1}{n} \sum_{l=0}^{n-1} e^{\frac{2\pi\imath}{n} lk} \overset{(4.2.1.8)}{=} \begin{cases} 0 & \text{, if } k \notin n\mathbb{Z} , \\ 1 & \text{, if } k \in n\mathbb{Z} . \end{cases} \end{cases}$$

The second identity is a consequence of the geometric sum formula (4.2.1.9).

Lemma 7.5.0.22. Exact quadrature by equidistant trapezoidal rule

The n-point equidistant trapezoidal quadrature rule on $[0,1]$

$$\int_0^1 f(t)\,\mathrm{d}t \approx T_n(f) := h\left(\tfrac{1}{2}f(0) + \sum_{k=1}^{n-1} f(kh) + \tfrac{1}{2}f(1)\right) , \quad h := \frac{1}{n} . \tag{7.5.0.17}$$

is exact for trigonometric polynomials ($\rightarrow$ Section 6.5.1) of degree $\leq 2n-2$.

$\blacktriangleright$ Since the weights of the equidistant trapezoidal rule are clearly positive, by Thm. 7.4.3.3 the asymptotic behavior of the quadrature error can directly be inferred from estimates for equidistant trigonometric interpolation. Such estimates are given, e.g., in Thm. 6.5.3.14 and they confirm exponential convergence for periodic integrands that allow analytic extension, in agreement with the observations made in Exp. 7.5.0.16.

Numerical Quadrature of *periodic integrands*

Use the equidistant trapezoidal rule for the numerical quadrature of a periodic integrand over its interval of periodicity.

$\lrcorner$

Remark 7.5.0.24 (Approximate computation of Fourier coefficients)

Recall from Section 4.2.6: recovery of signal $(y_k)_{k\in\mathbb{Z}}$ from its Fourier transform $c(t)$

$$y_j = \int_0^1 c(t)\exp(2\pi i j t)\,\mathrm{d}t . \tag{4.2.6.20}$$

Task: approximate computation of y_j

Recall: $c(t)$ obtained from $(y_k)_{k\in\mathbb{Z}}$ through Fourier series

$$c(t) = \sum_{k\in\mathbb{Z}} y_k \exp(-2\pi ikt)\,. \tag{4.2.6.7}$$

➢ $c(t)$ smooth & 1-periodic for finite/rapidly decaying $(y_k)_{k\in\mathbb{Z}}$.

Exp. 7.5.0.16 ► use equidistant trapezoidal rule (7.5.0.17) for approximate evaluation of integral in (4.2.6.20).

☞ Boils down to inverse DFT (4.2.1.20); hardly surprising in light of the derivation of (4.2.6.20) in Section 4.2.6.

C++-code 7.5.0.25: DFT-based approximate computation of Fourier coefficients

```
#include <complex>
#include <iostream>
#include <unsupported/Eigen/FFT>
#include <vector>

template <class Function>
void fourcoeffcomp(std::vector<std::complex<double>> &y, Function &c,
                   const unsigned m, const unsigned ovsmpl = 2) {
  // Compute the Fourier coefficients y_{-m},...,y_m of the
  // function c:[0,1[→ℂ using an oersampling
  // factor ovsmpl.  c must be a handle to a function
  // @(t), e.g.  a lambda function
  const unsigned N = (2 * m + 1) * ovsmpl; // number of quadrature points
  const double h = 1. / N;

  // evaluate function in N points
  Eigen::VectorXcd c_eval(N);
  for (unsigned i = 0; i < N; ++i) {
    c_eval(i) = c(i * h);
  }

  // inverse discrete fourier transformation
  Eigen::FFT fft;
  Eigen::VectorXcd z;
  fft.inv(z, c_eval);

  // Undo oversampling and wrapping of Fourier coefficient array
  // -> y contains same values as z but in a different order:
  // y = [z(N-m+1:N), z(1:m+1)]
  y = std::vector<std::complex<double>>();
  y.reserve(N);
  for (unsigned i = N - m; i < N; ++i) {
    y.push_back(z[i]);
  }
  for (unsigned j = 0; j < m + 1; ++j) {
    y.push_back(z[j]);
  }
}
```

⌟

Review question(s) 7.5.0.26 (Composite quadrature rules)

(Q7.5.0.26.A) Assume that $f : D \subset \mathbb{C} \to \mathbb{C}$ is analytic on the open subset D of the complex plane. By

the Cauchy integral formula we have for $z \in D$

$$f^{(k)}(z) = \frac{k!}{2\pi\imath} \int_{\partial B_r(z)} \frac{f(w)}{(w-z)^{k+1}} \,\mathrm{d}w\,, \quad k \in \mathbb{N}_0\,, \tag{7.5.0.27}$$

where

$$B_r(z) := \{w \in \mathbb{C} : \ |w-z| \le r\},\,, \quad r > 0\,,$$

is the closed disk with radius $r > 0$ around z and r is so small that $B_r(z) \subset D$.

Using the parameterization

$$\partial B_r(z) = \{z + r\exp(2\pi\imath t),\ t \in [0,1]\}$$

outline how (7.5.0.27) can be approximated by means of numerical quadrature.

Hint. Recall the definition of a path integral in the complex plane ("contour integral"): If the path of integration γ is described by a parameterization $\tau \in J \mapsto \gamma(\tau) \in \mathbb{C}$, $J \subset \mathbb{R}$, then

$$\int_\gamma f(z)\,\mathrm{d}z := \int_J f(\gamma(\tau)) \cdot \dot\gamma(\tau)\,\mathrm{d}\tau\,, \tag{6.2.2.51}$$

△

7.6 Adaptive Quadrature

Video tutorial for Section 7.6 "Adaptive Quadrature": (13 minutes) Download link, tablet notes

Hitherto, we have just "blindly" applied quadrature rules for the approximate evaluation of $\int_a^b f(t)\,\mathrm{d}t$, oblivious of any properties of the integrand f. This led us to the conclusion of Rem. 7.5.0.13 that Gauss-Legendre quadrature (→ Def. 7.4.2.18) should be preferred to composite quadrature rules (→ Section 7.5) in general. Now the composite quadrature rule will partly be rehabilitated, because they offer the flexibility to *adjust the quadrature rule to the integrand*, a policy known as adaptive quadrature.

Adaptive numerical quadrature

The policy of adaptive quadrature approximates $\int_a^b f(t)\,\mathrm{d}t$ by a quadrature formula (7.2.0.2), whose nodes c_j^n are chosen depending on the integrand f.

We distinguish

(I) a priori adaptive quadrature: the nodes are *fixed* before the evaluation of the quadrature formula, taking into account external information about f, and

(II) a posteriori adaptive quadrature: the node positions are chosen or improved based on information gleaned *during the computation* inside a loop. It terminates when sufficient accuracy has been reached.

In this section we will chiefly discuss a posteriori adaptive quadrature for composite quadrature rules (→ Section 7.5) based on a single local quadrature rule (and its transformation).

Supplementary literature. [DH03, Sect. 9.7]

EXAMPLE 7.6.0.2 (Rationale for adaptive quadrature) This example presents an extreme case. We consider the composite trapezoidal rule (7.5.0.4) on a mesh $\mathcal{M} := \{a = x_0 < x_1 < \cdots < x_m = b\}$ and for the integrand $f(t) = \frac{1}{10^{-4}+t^2}$ on $[-1,1]$.

f is a spike-like function ▷

Intuition: quadrature nodes should cluster around 0, whereas hardly any are needed close to the endpoints of the integration interval, where the function has very small (in modulus) values.

➢ Use locally refined mesh !

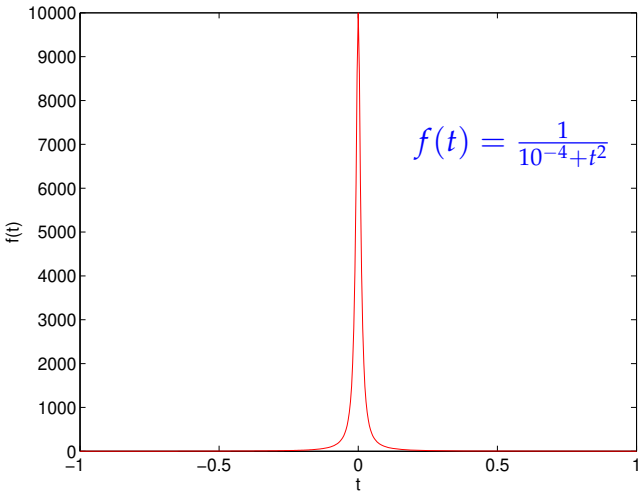


Fig. 283

A quantitative justification can appeal to (7.3.0.11) and the resulting bound for the local quadrature error (for $f \in C^2([a,b])$):

$$\int_{x_{k-1}}^{x_k} f(t)\,\mathrm{d}t - \frac{1}{2}(f(x_{k-1}) + f(x_k)) \leq \tfrac{1}{8} h_k^3 \|f''\|_{L^\infty([x_{k-1},x_k])}\,, \quad h_k := x_k - x_{k-1}\,. \tag{7.6.0.3}$$

➢ Suggests the use of small mesh intervals, where $|f''|$ is large **!**

§7.6.0.4 (Goal: equidistribution of errors) The ultimate but elusive goal is to find a mesh with a minimal number of cells that just delivers a quadrature error below a prescribed threshold. A more practical goal is to adjust the local meshwidths $h_k := x_k - x_{k-1}$ in order to *achieve a minimal sum of local error bounds*. This leads to the constrained minimization problem:

$$\sum_{k=1}^{m} h_k^3 \|f''\|_{L^\infty([x_{k-1},x_k])} \to \min \quad \text{s.t.} \quad \sum_{k=1}^{m} h_k = b - a\,. \tag{7.6.0.5}$$

Lemma 7.6.0.6.

Let $f : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ be a convex function with $f(0) = 0$ and $x > 0$. Then the constrained minimization problem: seek $\zeta_1, \ldots, \zeta_m \in \mathbb{R}_0^+$ such that

$$\sum_{k=1}^{m} f(\zeta_k) \to \min \quad \textit{and} \quad \sum_{k=1}^{m} \zeta_k = x\,, \tag{7.6.0.7}$$

has the solution $\zeta_1 = \zeta_2 = \cdots = \zeta_m = \frac{x}{m}$.

This means that we should strive for equal bounds $h_k^3 \|f''\|_{L^\infty([x_{k-1},x_k])}$ for all mesh cells.

Error equidistribution principle

The mesh for *a posteriori adaptive composite numerical quadrature* should be chosen to achieve equal contributions of all mesh intervals to the quadrature error

⌟

A indicated above, guided by the equidistribution principle, the improvement of the mesh will be done gradually in an iteration. The change of the mesh in each step is called mesh adaptation and there are two fundamentally different ways to do it:

(I) by moving nodes, keeping their total number, but making them cluster where mesh intervals should be small, or

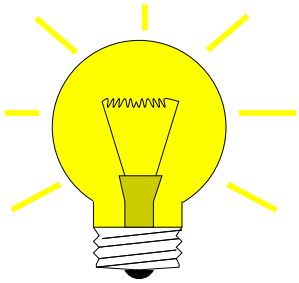
(II) by adding nodes, where mesh intervals should be small (mesh refinement).

Algorithms for a posteriori adaptive quadrature based on mesh refinement usually have the following structure:

Adaptation loop for numerical quadrature

(1) ESTIMATE: based on available information compute an approximation for the quadrature error on every mesh interval.
(2) CHECK TERMINATION: if total error sufficient small → **STOP**
(3) MARK: single out mesh intervals with the largest or above average error contributions.
(4) REFINE: add node(s) inside the marked mesh intervals. GOTO (1)

§7.6.0.10 (Adaptive multilevel quadrature) We now see a concrete algorithm based on the two composite quadrature rules introduced in Ex. 7.5.0.3.



Idea: local error estimation by comparing local results of two quadrature formulas Q_1, Q_2 of *different* order → local error estimates

heuristics: $\text{error}(Q_2) \ll \text{error}(Q_1) \quad \Rightarrow \quad \text{error}(Q_1) \approx Q_2(f) - Q_1(f)$.

Here: Q_1 = trapezoidal rule (order 2) ↔ Q_2 = Simpson rule (order 4)

Given: initial mesh $\mathcal{M} := \{a = x_0 < x_1 < \cdots < x_m = b\}$

❶ (Error estimation)

For $I_k = [x_{k-1}, x_k]$, $k = 1, \ldots, m$ (midpoints $p_k := \frac{1}{2}(x_{k-1} + x_k)$)

$$\text{EST}_k := \Big| \underbrace{\frac{h_k}{6}(f(x_{k-1}) + 4f(p_k) + f(x_k))}_{\text{Simpson rule}} - \underbrace{\frac{h_k}{4}(f(x_{k-1}) + 2f(p_k) + f(x_k))}_{\text{trapezoidal rule on split mesh interval}} \Big| \,. \tag{7.6.0.11}$$

❷ (Check termination)

Simpson rule on $\mathcal{M}$ $\Rightarrow$ intermediate approximation $I \approx \int_a^b f(t)\,\mathrm{d}t$

If $\sum_{k=1}^{m} \mathtt{EST}_k \leq \mathtt{RTOL} \cdot I$ ($RTOL$:= prescribed relative tolerance) $\Rightarrow$ **STOP** (7.6.0.12)

❸ (Marking)

Marked intervals: $\mathcal{S} := \{k \in \{1,\ldots,m\}\colon \mathtt{EST}_k \geq \eta \cdot \frac{1}{m}\sum_{j=1}^{m} \mathtt{EST}_j\}$, $\eta \approx 0.9$. (7.6.0.13)

❹ (Local mesh refinement)

new mesh: $\mathcal{M}^* := \mathcal{M} \cup \{p_k := \frac{1}{2}(x_{k-1} + x_k)\colon k \in \mathcal{S}\}$. (7.6.0.14)

Then continue with step ❶ and mesh $\mathcal{M} \leftarrow \mathcal{M}^*$.

The following C++ code give a (non-optimal) recursive implementation

C++-code 7.6.0.15: h-adaptive numerical quadrature ↪ GITLAB

```
// Adaptive multilevel quadrature of a function passed in f.
// The vector M passes the positions of current quadrature nodes
template <class Function>
double adaptquad(Function& f, VectorXd& M,double rtol,double atol) {
  const std::size_t n = M.size(); // number of nodes
  // distance of quadature nodes
  const VectorXd h = M.tail(n-1)-M.head(n-1);
  const VectorXd mp = 0.5*(M.head(n-1)+M.tail(n-1)); // midpoints
  // Values of integrand at nodes and midpoints
  VectorXd fx(n);
  VectorXd fm(n - 1);
  for (unsigned i = 0; i < n; ++i) {
    fx(i) = f(M(i));
  }
  for (unsigned j = 0; j < n - 1; ++j) {
    fm(j) = f(mp(j));
  }
  // trapezoidal rule (7.5.0.4)
  const VectorXd trp_loc = 1./4*h.cwiseProduct(fx.head(n-1)+2*fm+fx.tail(n-1));
  // Simpson rule (7.5.0.5)
  const VectorXd simp_loc = 1./6*h.cwiseProduct(fx.head(n-1)+4*fm+fx.tail(n-1));

  // Simpson approximation for the integral value
  double I = simp_loc.sum();
  // local error estimate (7.6.0.11)
  const VectorXd est_loc = (simp_loc-trp_loc).cwiseAbs();
  // estimate for quadrature error
  const double err_tot = est_loc.sum();

  // STOP: Termination based on (7.6.0.12)
  if ( err_tot > rtol*std::abs(I) && err_tot > atol ) { //
    // find cells where error is large
    std::vector<double> new_cells;
    for (unsigned i = 0; i < est_loc.size(); ++i) {
      // MARK by criterion (7.6.0.13) & REFINE by (7.6.0.14)
      if (est_loc(i) > 0.9/static_cast<double>(n-1)*err_tot) {
        // new quadrature point = midpoint of interval with large error
        new_cells.push_back(mp(i));
      }}

    // create new set of quadrature nodes
    // (necessary to convert std::vector to Eigen vector)
    const Eigen::Map<VectorXd> tmp(new_cells.data(),
      static_cast<Eigen::Index>(new_cells.size()));
    VectorXd new_M(M.size() + tmp.size());
    new_M << M, tmp; // concatenate old cells and new cells
    // nodes of a mesh are supposed to be sorted
    std::sort(new_M.begin(),new_M.end());
    I = adaptquad(f, new_M, rtol, atol); // recursion
  }
  return I;
}
```

Comments on Code 7.6.0.15:

- Arguments: `f` ≙ *handle* to function f, `M` ≙ initial mesh, `rtol` ≙ relative tolerance for termination, `atol` ≙ absolute tolerance for termination, necessary in case the exact integral value $= 0$, which renders a relative tolerance meaningless.

- Line 7: compute lengths of mesh-intervals $[x_{j-1}, x_j]$,
- Line 9: store positions of midpoints p_j,
- Line 10: evaluate function (vector arguments!),
- Line 19: local composite trapezoidal rule (7.5.0.4),
- Line 21: local simpson rule (7.3.0.6),
- Line 24: value obtained from composite simpson rule is used as intermediate approximation for integral value,
- Line 26: difference of values obtained from local composite trapezoidal rule ($\sim Q_1$) and local simpson rule ($\sim Q_2$) is used as an estimate for the local quadrature error.
- Line 28: estimate for global error by summing up moduli of local error contributions,
- Line 32: terminate, once the estimated total error is below the relative or absolute error threshold,
- Line 50 otherwise, add midpoints of mesh intervals with large error contributions according to (7.6.0.14) to the mesh and continue.

C++-code 7.6.0.16: Call of `adaptquad()`:

```
1  # include "./adaptquad.hpp"
2  # include <Eigen/Dense>
3  # include <cmath>
4  # include <iostream>
5
6  using Eigen::VectorXd;
7
8  int main () {
9    auto f =[](double x) { return std::exp(-x*x); };
10   VectorXd M(4);
11   M << -100, 0.1, 0.5, 100;
12   std::cout << "Sqrt(Pi) - Int_{-100}^{100} exp(-x*x) dx = ";
13   std::cout << adaptquad::adaptquad(f, M, 1e-10, 1e-12) - std::sqrt(M_PI) << "\n";
14   return 0;
15 }
```

⌟

Remark 7.6.0.17 (Estimation of "wrong quadrature error"?) In Code 7.6.0.15 we use the higher order quadrature rule, the Simpson rule of order 4, to compute an approximate value for the integral. This is reasonable, because it would be foolish not to use this information after we have collected it for the sake of error estimation.

Yet, according to our heuristics, what `est_loc` and `est_tot` give us are estimates for the error of the second-order trapezoidal rule, which we do not use for the actual computations.

However, experience teaches that

> `est_loc` gives useful (for the sake of mesh refinement) information about the *distribution* of the error of the Simpson rule, though it fails to capture its *size*.

Therefore, the termination criterion of Line 32 may not be appropriate! ⌟

EXPERIMENT 7.6.0.18 (h-adaptive numerical quadrature) In this numerical test we investigate whether the adaptive technique from § 7.6.0.10 produces an appropriate distribution of integration nodes. We do

this for different functions.

✦ approximate $\int_0^1 \exp(6\sin(2\pi t))\,\mathrm{d}t$, initial mesh $\mathcal{M}_0 = \{j/10\}_{j=0}^{10}$

Algorithm: adaptive quadrature, Code 7.6.0.15 with tolerances `rtol` $= 10^{-6}$, `abstol` $= 10^{-10}$

We monitor the distribution of quadrature points during the adaptive quadrature and the true and estimated quadrature errors. The "exact" value for the integral is computed by composite Simpson rule on an equidistant mesh with 10^7 intervals.

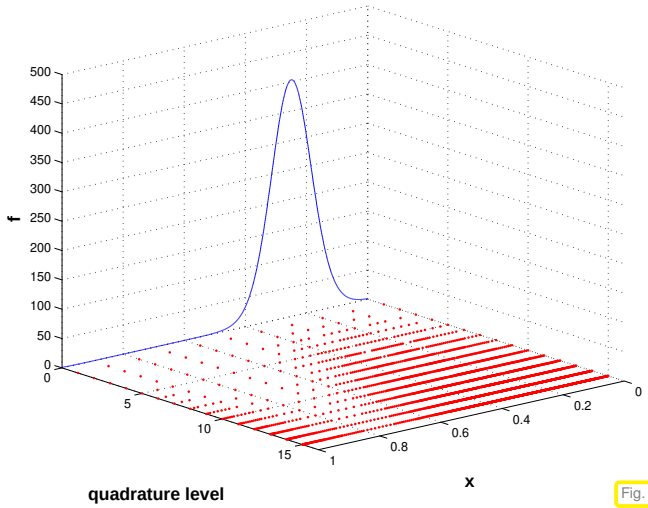


Fig. 284

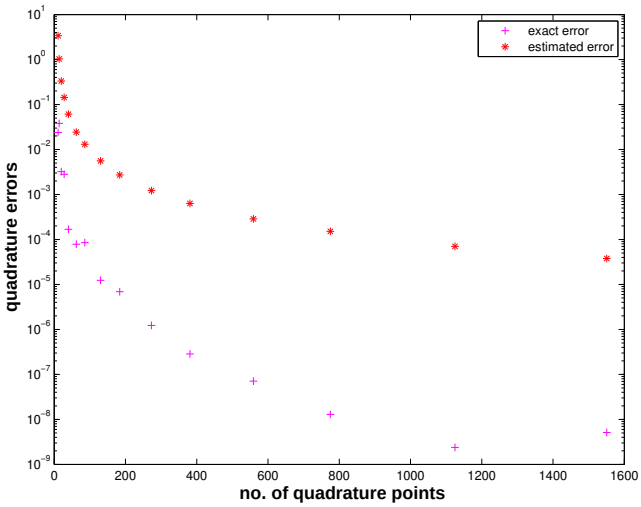


Fig. 285

✦ approximate $\int_0^1 \min\{\exp(6\sin(2\pi t)), 100\}\,\mathrm{d}t$, initial mesh as above

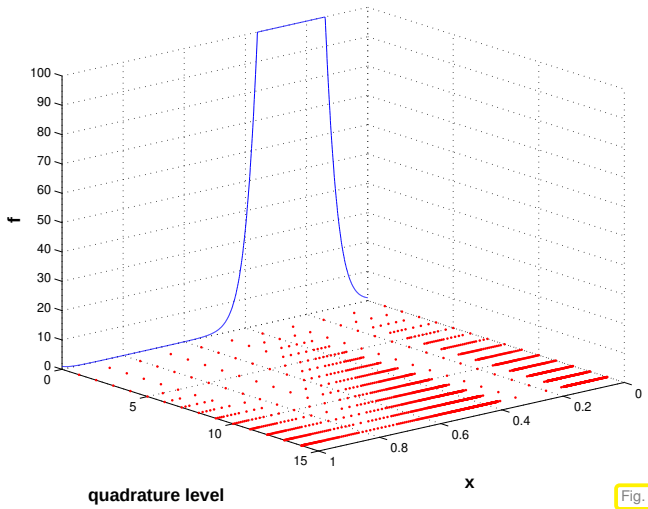


Fig. 286

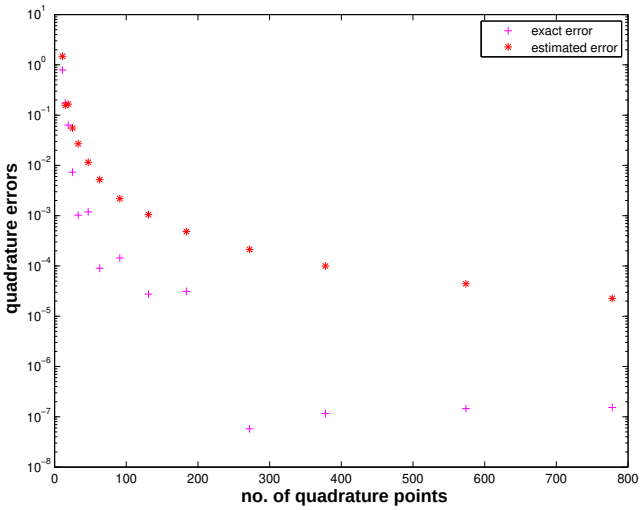


Fig. 287

Observation:

- Adaptive quadrature locally decreases meshwidth where integrand features variations or kinks.
- Trend for estimated error mirrors behavior of true error.
- Overestimation may be due to taking the modulus in (7.6.0.11)

However, the important piece of information we want to extract from EST_k is about the *distribution* of the quadrature error.

⌟

Remark 7.6.0.19 (Adaptive quadrature in PYTHON)

`q = scipy.integrate.quad(fun,a,b,tol)`:	adaptive multigrid quadrature (local low order quadrature formulas)
`q = scipy.integrate.quadrature(fun,a,b,tol)`:	adaptive Gauss-Lobatto quadrature

⌟

Review question(s) 7.6.0.20 (Adaptive quadrature)

(Q7.6.0.20.A) For the composite trapezoidal rule applied on a mesh $\mathcal{M} := \{a = x_0 < x_1 < \cdots < x_m := b\}$ we have the following estimate for the local quadrature error for an integrand $f \in C^2([a,b])$:

$$\int_{x_{k-1}}^{x_k} f(t)\,\mathrm{d}t - \frac{1}{2}(f(x_{k-1}) + f(x_k)) \leq \tfrac{1}{8} h_k^3 \|f''\|_{L^\infty([x_{k-1},x_k])}\,, \quad h_k := x_k - x_{k-1}\,. \tag{7.6.0.3}$$

We consider the singular integrand $f(t) = \sqrt{t}$ on $[0,1]$. What mesh $\mathcal{M}$ has to be chosen to ensure the equidistribution of the error bounds from (7.6.0.3).

(Q7.6.0.20.B) For a posteriori adaptive mesh refinement for the approximation of $\int_a^b f(t)\,\mathrm{d}t$ we employed the following estimate of the local quadrature error on the mesh $\mathcal{M} := \{a = x_0 < x_1 < \cdots < x_m := b\}$

$$\mathrm{EST}_k := \Big| \underbrace{\frac{h_k}{6}(f(x_{k-1}) + 4f(p_k) + f(x_k))}_{\text{Simpson rule}} - \underbrace{\frac{h_k}{4}(f(x_{k-1}) + 2f(p_k) + f(x_k))}_{\text{trapezoidal rule on split mesh interval}} \Big|\,. \tag{7.6.0.11}$$

We could also have use the two lowest order Gauss-Legendre quadrature rules for that purpose,

- The 1-point midpoint rule, on $[-1,1]$ defined by the node $c_1 := 0$ and $w_1 := 2$,
- the 2-point Gauss-Legendre rule from Ex. 7.4.2.2, on $[-1,1]$ by the weights/nodes $\left\{w_2 = 1, w_1 = 1, c_1 = 1/3\sqrt{3}, c_2 = -1/3\sqrt{3}\right\}$.

Write down the formula for the resulting estimator EST_k and compare it with the choice (7.6.0.11) in terms of number of required f-evaluations.

△

Learning Outcomes

- You should know what is a quadrature formula and terminology connected with it,
- You should be able to transform quadrature formulas to arbitrary intervals.
- You should understand how a interpolation and approximation schemes spawn quadrature formulas and how quadrature errors are connected to interpolation/approximation errors.
- You should be able to compute the weights of polynomial quadrature formulas.
- You should know the concept of order of a quadrature rule and why it is invariant under (affine) transformation
- You should remember the maximal and minimal order of polynomial quadrature rules.
- You should know the order of the n-point Gauss-Legendre quadrature rule.
- You should understand why Gauss-Legendre quadrature converges exponentially for integrands that can be extended analytically and algebraically for integrands with limited smoothness.

- You should be apply to apply regularizing transformations to integrals with non-smooth integrands.
- You should know about asymptotic convergence of the h-version of composite quadrature.
- You should know the principles of adaptive composite quadrature.

Bibliography

[BL19] L. Banjai and M. López-Fernández. “Efficient high order algorithms for fractional integrals and fractional differential equations”. In: *Numer. Math.* 141.2 (2019), pp. 289–317. DOI: 10.1007/s00211-018-1004-0.

[Bog14] I. Bogaert. “Iteration-free computation of Gauss-Legendre quadrature nodes and weights”. In: *SIAM J. Sci. Comput.* 36.3 (2014), A1008–A1026. DOI: 10.1137/140954969 (cit. on p. 575).

[DR08] W. Dahmen and A. Reusken. *Numerik für Ingenieure und Naturwissenschaftler*. Heidelberg: Springer, 2008 (cit. on pp. 557, 561, 564).

[DH03] P. Deuflhard and A. Hohmann. *Numerical Analysis in Modern Scientific Computing*. Vol. 43. Texts in Applied Mathematics. Springer, 2003 (cit. on p. 589).

[Fej33] L. Fejér. “Mechanische Quadraturen mit positiven Cotesschen Zahlen”. In: *Math. Z.* 37.1 (1933), pp. 287–309. DOI: 10.1007/BF01474575 (cit. on p. 564).

[Gan+05] M. Gander, W. Gander, G. Golub, and D. Gruntz. *Scientific Computing: An introduction using MATLAB*. Springer, 2005 (cit. on p. 574).

[GLR07] Andreas Glaser, Xiangtao Liu, and Vladimir Rokhlin. “A fast algorithm for the calculation of the roots of special functions”. In: *SIAM J. Sci. Comput.* 29.4 (2007), pp. 1420–1438. DOI: 10.1137/06067016X.

[Gut09] M.H. Gutknecht. *Lineare Algebra*. Lecture Notes. SAM, ETH Zürich, 2009 (cit. on pp. 569, 570).

[Han02] M. Hanke-Bourgeois. *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens*. Mathematische Leitfäden. Stuttgart: B.G. Teubner, 2002 (cit. on pp. 557, 562, 564, 573).

[Joh08] S.G. Johnson. *Notes on the convergence of trapezoidal-rule quadrature*. MIT online course notes, http://math.mit.edu/ stevenj/trapezoidal.pdf. 2008.

[NS02] K. Nipp and D. Stoffer. *Lineare Algebra*. 5th ed. Zürich: vdf Hochschulverlag, 2002 (cit. on pp. 569, 570).

[Tre08] Lloyd N. Trefethen. “Is Gauss quadrature better than Clenshaw-Curtis?” In: *SIAM Rev.* 50.1 (2008), pp. 67–87. DOI: 10.1137/060659831 (cit. on pp. 564, 574).

[TWXX] Lloyd N. Trefethen and J. A. C. Weideman. *THE EXPONENTIALLY CONVERGENT TRAPEZOIDAL RULE*. XX.

[Wal06] Jörg Waldvogel. “Fast construction of the Fejér and Clenshaw-Curtis quadrature rules”. In: *BIT* 46.1 (2006), pp. 195–202. DOI: 10.1007/s10543-006-0045-4 (cit. on p. 564).

[Wal11] Jörg Waldvogel. “Towards a general error theory of the trapezoidal rule”. In: *Approximation and computation*. Vol. 42. Springer Optim. Appl. Springer, New York, 2011, pp. 267–282. DOI: 10.1007/978-1-4419-6594-3_17.

Chapter 8

Iterative Methods for Non-Linear Systems of Equations

8.1 Introduction

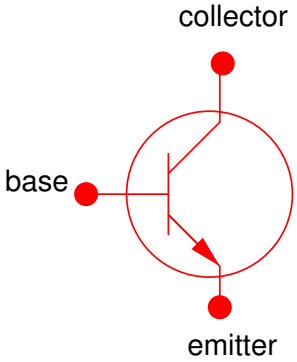
Video tutorial for Section 8.1 "Iterative Methods for Non-Linear Systems of Equations: Introduction": (6 minutes) Download link, tablet notes

EXAMPLE 8.1.0.1 (Non-linear electric circuit)

Non-linear systems naturally arise in mathematical models of electrical circuits, once non-linear circuit elements are introduced. This generalizes Ex. 2.1.0.3, where the current-voltage relationship for all circuit elements was the simple proportionality (2.1.0.5) (of the complex amplitudes U and I).

As an example we consider the Schmitt trigger circuit ▷

Its key non-linear circuit element is the NPN bipolar junction transistor:



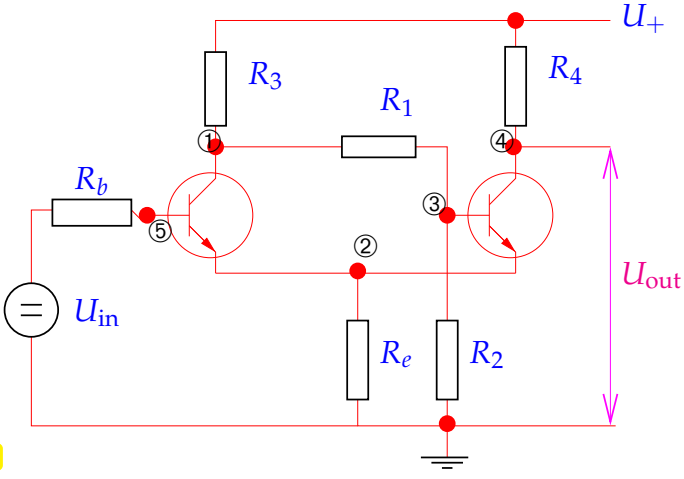


Fig. 288

A transistor has three ports: emitter, collector, and base. Transistor models give the port currents as functions of the applied voltages, for instance the Ebers-Moll model (large signal approximation):

$$\begin{aligned} I_{\rm C} &= I_{\rm S}\left(e^{\frac{U_{\rm BE}}{U_{\rm T}}} - e^{\frac{U_{\rm BC}}{U_{\rm T}}}\right) - \frac{I_{\rm S}}{\beta_R}\left(e^{\frac{U_{\rm BC}}{U_{\rm T}}} - 1\right) = I_{\rm C}(U_{\rm BE}, U_{\rm BC})\,, \\ I_{\rm B} &= \frac{I_{\rm S}}{\beta_F}\left(e^{\frac{U_{\rm BE}}{U_{\rm T}}} - 1\right) + \frac{I_{\rm S}}{\beta_R}\left(e^{\frac{U_{\rm BC}}{U_{\rm T}}} - 1\right) = I_{\rm B}(U_{\rm BE}, U_{\rm BC})\,, \\ I_{\rm E} &= I_{\rm S}\left(e^{\frac{U_{\rm BE}}{U_{\rm T}}} - e^{\frac{U_{\rm BC}}{U_{\rm T}}}\right) + \frac{I_{\rm S}}{\beta_F}\left(e^{\frac{U_{\rm BE}}{U_{\rm T}}} - 1\right) = I_{\rm E}(U_{\rm BE}, U_{\rm BC})\,. \end{aligned} \tag{8.1.0.2}$$

$I_{\rm C}$, $I_{\rm B}$, $I_{\rm E}$: current in collector/base/emitter,
$U_{\rm BE}$, $U_{\rm BC}$: potential drop between base-emitter, base-collector.

The parameters have the following meanings: β_F is the forward common emitter current gain (20 to 500), β_R is the reverse common emitter current gain (0 to 20), I_S is the reverse saturation current (on the order of 10^{-15} to 10^{-12} amperes), U_T is the thermal voltage (approximately 26 mV at 300 K).

The circuit of Fig. 288 has 5 nodes ①–⑤ with unknown nodal potentials. Kirchhoffs law (2.1.0.4) plus the constitutive relations gives an equation for each of them.

Non-linear system of equations from nodal analysis, static case (→ Ex. 2.1.0.3):

$$\begin{aligned}
&① : && R_3^{-1}(U_1 - U_+) + R_1^{-1}(U_1 - U_3) + I_C(U_5 - U_1, U_5 - U_2) &&= 0\,,\\
&② : && R_e^{-1}U_2 + I_E(U_5 - U_1, U_5 - U_2) + I_E(U_3 - U_4, U_3 - U_2) &&= 0\,,\\
&③ : && R_1^{-1}(U_3 - U_1) + I_B(U_3 - U_4, U_3 - U_2) &&= 0\,,\\
&④ : && R_4^{-1}(U_4 - U_+) + I_C(U_3 - U_4, U_3 - U_2) &&= 0\,,\\
&⑤ : && R_b^{-1}(U_5 - U_{\text{in}}) + I_B(U_5 - U_1, U_5 - U_2) &&= 0\,.
\end{aligned} \tag{8.1.0.3}$$

5 equations ↔ 5 unknowns U_1, U_2, U_3, U_4, U_5

Formally: (8.1.0.3) ⟷ $F(\mathbf{u}) = \mathbf{0}$ with a function $F : \mathbb{R}^5 \to \mathbb{R}^5$ ⌟

Remark 8.1.0.4 (*General* non-linear systems of equations) A non-linear system of equations is a concept almost *too abstract to be useful*, because it covers an extremely wide variety of problems . Nevertheless in this chapter we will mainly look at "generic" methods for such systems. This means that every method discussed may take a good deal of fine-tuning before it will really perform satisfactorily for a given non-linear system of equations. ⌟

§8.1.0.5 (Generic/general non-linear system of equations) Let us try to describe the "problem" of having to solve a non-linear system of equations, where the concept of a "problem" was first introduced in § 1.5.5.1.

Given: function $F : D \subset \mathbb{R}^n \mapsto \mathbb{R}^n$, $n \in \mathbb{N}$

⇕

Possible meanings:
- ☞ F is known as an analytic expression.
- ☞ F is merely available in procedural form allowing point evaluations.

Here, D is the domain of definition of the function F, which cannot be evaluated for $\mathbf{x} \notin D$.

Sought: solution(s) $\mathbf{x} \in D$ of non-linear equation $F(\mathbf{x}) = 0$

Note: $F : D \subset \mathbb{R}^n \mapsto \mathbb{R}^n$ ↔ "same number n of equations and unknowns"

In contrast to the situation for linear systems of equations (→ Thm. 2.2.1.4), the class of non-linear systems is far too big to allow a general theory:

> There are no general results on the existence & uniqueness of solutions of a "generic" non-linear system of equations $F(\mathbf{x}) = \mathbf{0}$.

⌟

Contents

8.1 Introduction **598**

8.2 Iterative Methods . . . 601
8.2.1 Fundamental Concepts . . . 601
8.2.2 Speed of Convergence . . . 605
8.2.3 Termination Criteria/Stopping Rules . . . 610
8.3 Fixed-Point Iterations . . . 614
8.3.1 Consistent Fixed-Point Iterations . . . 615
8.3.2 Convergence of Fixed-Point Iterations . . . 616
8.4 Finding Zeros of Scalar Functions . . . 623
8.4.1 Bisection . . . 623
8.4.2 Model Function Methods . . . 625
8.4.3 Asymptotic Efficiency of Iterative Methods for Zero Finding . . . 638
8.5 Newton's Method in $\mathbb{R}^n$ . . . 642
8.5.1 The Newton Iteration . . . 642
8.5.2 Convergence of Newton's Method . . . 654
8.5.3 Termination of Newton Iteration . . . 657
8.5.4 Damped Newton Method . . . 659
8.6 Quasi-Newton Method . . . 663
8.7 Non-linear Least Squares [DR08, Ch. 6] . . . 670
8.7.1 (Damped) Newton Method . . . 673
8.7.2 Gauss-Newton Method . . . 675
8.7.3 Trust Region Method (Levenberg-Marquardt Method) . . . 679

Review question(s) 8.1.0.6 (Iterative Methods for Non-Linear Systems of Equations: Introduction)

(Q8.1.0.6.A) State that nonlinear system of equations in the form $F(\mathbf{x}) = \mathbf{0}$ with a function $F : \mathbb{R}^n \to \mathbb{R}^n$ whose solution answers the following question:

> How does the diagonal of the given matrix $\mathbf{A} \in \mathbb{R}^{n,n}$ have to be modified (yielding a matrix $\widetilde{\mathbf{A}}$) so that the linear system of equations $\widetilde{\mathbf{A}}\mathbf{x} = \mathbf{b}$, $\mathbf{b} \in \mathbb{R}^n$ given, has a prescribed solution $\mathbf{x}^* \in \mathbb{R}^n$?

When does the non-linear system of equations have a unique solution and what is it?

(Q8.1.0.6.B) Which non-linear system of equations is solved by every

$$\mathbf{x}^* \in \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^n} \Phi(\mathbf{x}) \,, \quad \Phi : \mathbb{R}^n \to \mathbb{R} \text{ continuously differentiable?}$$

Hint. From your analysis course remember *necessary* conditions for a global minimum of a continuously differentiable function.

(Q8.1.0.6.C) A diode is a non-linear circuit element that yields vastly different currents depending on the polarity of the applied voltage. Quantitatively, its current voltage relationship is described by the Shockley diode equation

$$I(U) = I_S\left(\exp\left(\frac{U}{U_T}\right) - 1\right) \,,$$

where I_S is the reverse bias saturation current, and U_T so-called thermal voltage, both known parameters.

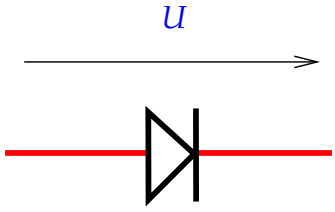


Fig. 289

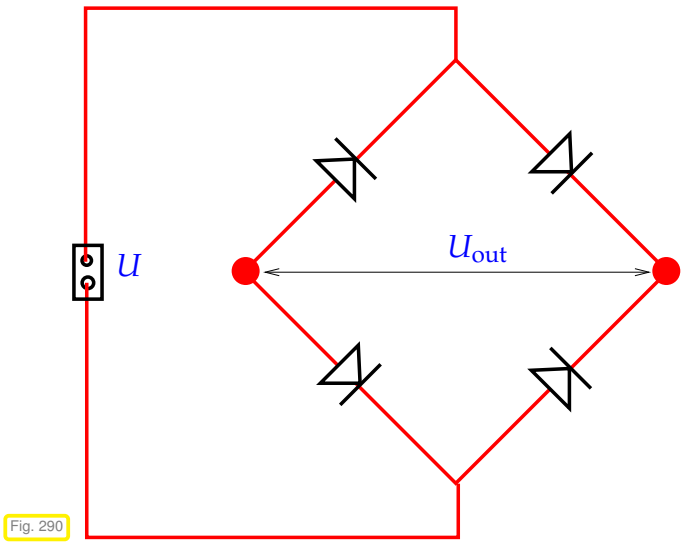


Fig. 290

◁ The circuit drawn beside is a rectifier.

Using nodal circuit analysis find the non-linear system of equations that permits us to compute the output voltage U_{out} between the two nodes •, when an input voltage U is applied?

Hint. Nodal analysis of electric circuits is explained in Ex. 2.1.0.3. You may look up that example.

(Q8.1.0.6.D) [Inverse function] From analysis you know that every monotonic function $f : I \subset \mathbb{R} \to \mathbb{R}$ can be inverted on its range $f(I)$. Reformulate the task of evaluating $f^{-1}(y)$ for $y \in f(I)$ as a non-linear equation in the standard form $F(x) = 0$ for a suitable function F.

△

8.2 Iterative Methods

Remark 8.2.0.1 (Necessity of iterative approximation) Gaussian elimination (→ Section 2.3) provides an algorithm that, if carried out in exact arithmetic (no roundoff errors), computes the solution of a linear system of equations with a *finite* number of elementary operations. However, linear systems of equations represent an exceptional case, because it is hardly ever possible to solve general systems of non-linear equations using only finitely many elementary operations. ⌟

8.2.1 Fundamental Concepts

Video tutorial for Section 8.2.1 "Iterative Methods: Fundamental Concepts": (6 minutes) Download link, tablet notes

§8.2.1.1 (Generic iterations)

All methods for general non-linear systems of equations are iterative in the sense that they will usually

yield only approximate solutions whenever they terminate after finite time.

An iterative method for (approximately) solving the non-linear equation $F(\mathbf{x}) = 0$ is an algorithm generating an arbitrarily long sequence $\left(\mathbf{x}^{(k)}\right)_k$ of approximate solutions.

$\mathbf{x}^{(k)} \hat{=} k$-th iterate

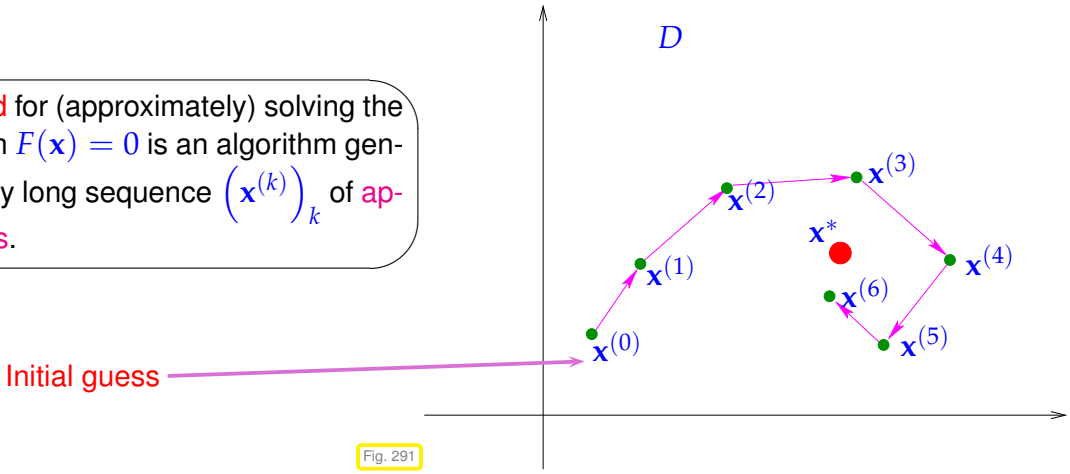


Fig. 291

⌟

§8.2.1.2 (Key issues with iterative methods) When applying an iterative method to solve a non-linear system of equations $F(\mathbf{x}) = 0$, the following issues arise:

- Convergence: Does the sequence $(\mathbf{x}^{(k)})_k$ converge to a limit: $\lim_{k\to\infty} \mathbf{x}^{(k)} = \mathbf{x}^*$?
- Consistency: Does the limit, if it exists, provide a solution of the non-linear system of equations: $F(\mathbf{x}^*) = 0$?
- Speed of convergence: How "fast" does $\left\|\mathbf{x}^{(k)} - \mathbf{x}^*\right\|$ ($\|\cdot\|$ a suitable norm on $\mathbb{R}^N$) decrease for increasing k?

More formal definitions can be given:

Definition 8.2.1.3. Convergence of iterative methods

An iterative method converges (for fixed initial guess(es)) $:\Leftrightarrow$ $\mathbf{x}^{(k)} \overset{k\to\infty}{\to} \mathbf{x}^*$ and $F(\mathbf{x}^*) = \mathbf{0}$.

§8.2.1.4 ((Stationary) m-point iterative method) All the iterative methods discussed below fall in the class of (stationary) m-point, $m \in \mathbb{N}$, iterative methods, for which the iterate $\mathbf{x}^{(k+1)}$ depends on F and the m most recent iterates $\mathbf{x}^{(k)}, \ldots, \mathbf{x}^{(k-m+1)}$, e.g.,

$$\mathbf{x}^{(k+1)} = \underbrace{\Phi_F(\mathbf{x}^{(k)}, \ldots, \mathbf{x}^{(k-m+1)})}_{\text{iteration function for } m\text{-point method}} \tag{8.2.1.5}$$

Terminology: Φ_F is called the **iteration function**.

Note: The initial guess(es) $\mathbf{x}^{(0)}, \ldots, \mathbf{x}^{(m-1)} \in \mathbb{R}^n$ have to be provided.

Visualization of a 1-point iteration:

$$\mathbf{x}^{(k+1)} = \Phi(\mathbf{x}^{(k)}) \,, \qquad (8.2.1.6)$$

with an iteration function

$$\Phi : D \subset \mathbb{R}^n \to \mathbb{R}^n \,.$$

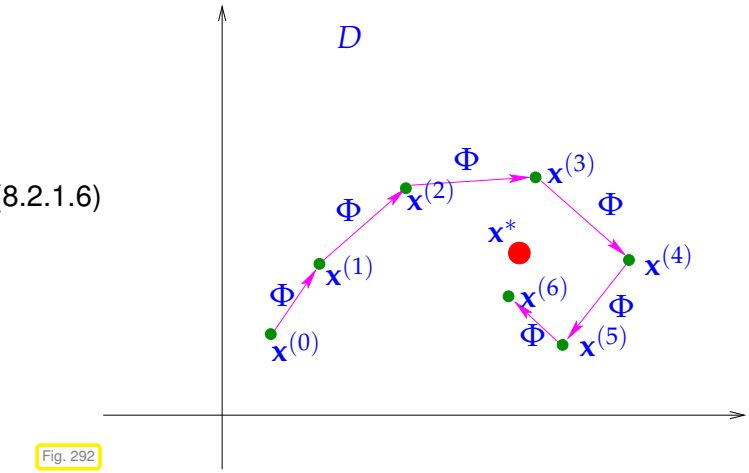


Fig. 292

⌟

Definition 8.2.1.7. Consistency of iterative methods

A stationanry m-point iterative method

$$\mathbf{x}^{(k+1)} = \Phi_F(\mathbf{x}^{(k)}, \ldots, \mathbf{x}^{(k-m+1)}) \,, \quad m \in \mathbb{N} \,, \qquad (8.2.1.5)$$

is **consistent** with the non-linear system of equations $F(\mathbf{x}) = \mathbf{0}$, if and only if

$$\Phi_F(\mathbf{x}^*, \ldots, \mathbf{x}^*) = \mathbf{x}^* \quad \Longleftrightarrow \quad F(\mathbf{x}^*) = \mathbf{0} \,.$$

Theorem 8.2.1.8. Consistency and convergence

The limit of a convergent sequence $\left(\mathbf{x}^{(k)}\right)_{k\in\mathbb{N}_0}$ generated by an m-point stationary iterative method with a continuous iteration function Φ_F that is consistent with $F(\mathbf{x}) = \mathbf{0}$ is a solution:

$$\left.\begin{aligned} &\mathbf{x}^{(k+1)} := \Phi_F(\mathbf{x}^{(k)}, \ldots, \mathbf{x}^{(k-m+1)}) \,, \\ &\mathbf{x}^* := \lim_{k\to\infty} \mathbf{x}^{(k)} \end{aligned}\right\} \quad \Rightarrow \quad F(\mathbf{x}^*) = \mathbf{0} \,.$$

Proof. The very definition of continuity means that limits can be "pulled into a function":

$$\begin{aligned} \mathbf{x}^* &= \lim_{k\to\infty} \mathbf{x}^{(k+1)} = \lim_{k\to\infty} \Phi_F(\mathbf{x}^{(k)}, \ldots, \mathbf{x}^{(k-m+1)}) \\ &= \Phi_F\left(\lim_{k\to\infty} \mathbf{x}^{(k)}, \ldots, \lim_{k\to\infty} \mathbf{x}^{(k-m+1)}\right) = \Phi_F(\mathbf{x}^*, \ldots, \mathbf{x}^*) \,. \end{aligned}$$

Appealing to Def. 8.2.1.7 finishes the proof. □

For a consistent stationary iterative method we can study the error of the iterates $\mathbf{x}^{(k)}$ defined as:

$$\mathbf{e}^{(k)} := \mathbf{x}^{(k)} - \mathbf{x}^* \,.$$

⌟

§8.2.1.9 (Local convergence of iterative methods) Unfortunately, convergence may critically depend on

the choice of initial guesses. The property defined next weakens this dependence:

Definition 8.2.1.10. Local and global convergence → [Han02, Def. 17.1]

As stationary m-point iterative method converges locally to $\mathbf{x}^* \in \mathbb{R}^n$, if there is a neighborhood $U \subset D$ of $\mathbf{x}^*$, such that

$$\mathbf{x}^{(0)}, \ldots, \mathbf{x}^{(m-1)} \in U \quad \Rightarrow \quad \mathbf{x}^{(k)} \text{ well defined} \quad \wedge \quad \lim_{k\to\infty} \mathbf{x}^{(k)} = \mathbf{x}^*$$

where $(\mathbf{x}^{(k)})_{k\in\mathbb{N}_0}$ is the (infinite) sequence of iterates.
If $U = D$, the iterative method is globally convergent.

Illustration of local convergence ▷

(Only initial guesses "sufficiently close" to $\mathbf{x}^*$ guarantee convergence.)

Unfortunately, the neighborhood U is rarely known a priori. It may also be very small.

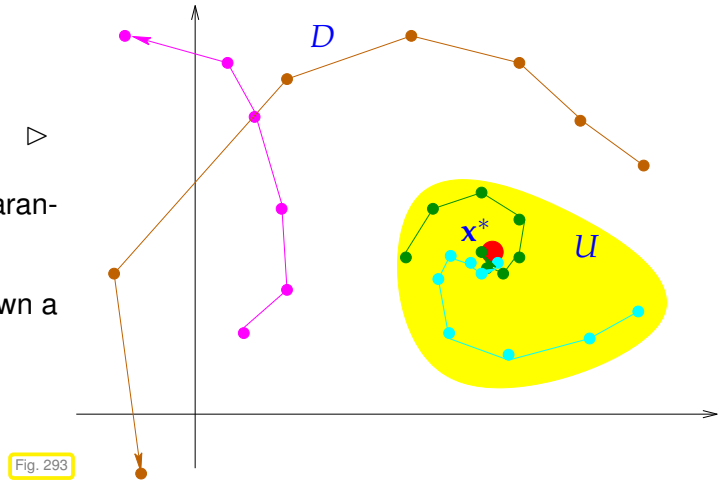


Fig. 293

⌟

Our goal: Given a non-linear system of equations, find iterative methods that converge (locally) to a solution of $F(\mathbf{x}) = 0$.

Two general questions: How to measure, describe, and predict the speed of convergence?
When to terminate the iteration?

Review question(s) 8.2.1.11 (Fundamentals of iterative methods)

(Q8.2.1.11.A) Rewrite an m-point iterative method for solving $F(\mathbf{x}) = 0$,

$$\mathbf{x}^{(k+1)} = \Phi_F(\mathbf{x}^{(k)}, \ldots, \mathbf{x}^{(k-m+1)}) ,$$

with iteration function

$$\Phi_F : \underbrace{\mathbb{R}^n \times \cdots \times \mathbb{R}^n}_{m \text{ times}} \to \mathbb{R}^n$$

as a 1-point iteration (also called a fixed-point iteration). What does consistency mean for that 1-point iteration.

(Q8.2.1.11.B) When is the following 1-point iterative method

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \mathbf{M}F(\mathbf{x}^{(k)}) , \quad \mathbf{M} \in \mathbb{R}^{n,n} ,$$

consistent with the non-linear system of equations $F(\mathbf{x}) = 0$, $F : D \subset \mathbb{R}^n \to \mathbb{R}^n$.

Definition 8.2.1.7. Consistency of iterative methods

A stationary m-point iterative method is consistent with the non-linear system of equations $F(\mathbf{x}) = 0$, if and only if

$$\Phi_F(\mathbf{x}^*, \ldots, \mathbf{x}^*) = \mathbf{x}^* \quad \Longleftrightarrow \quad F(\mathbf{x}^*) = 0 \,.$$

△

8.2.2 Speed of Convergence

Video tutorial for Section 8.2.2 "Iterative Methods: Speed of Convergence": (15 minutes) Download link, tablet notes

Here and in the sequel, $\|\cdot\|$ designates a generic vector norm on $\mathbb{R}^n$, see Def. 1.5.5.4. Any occurring matrix norm is induced by this vector norm, see Def. 1.5.5.10.

It is important to be aware which statements depend on the choice of norm and which do not!

"*Speed of convergence*" measures the decrease of a norm (see Def. 1.5.5.4) of the iteration error

Definition 8.2.2.1. Linear convergence

A sequence $\mathbf{x}^{(k)}$, $k = 0, 1, 2, \ldots$, in $\mathbb{R}^n$ **converges linearly** to $\mathbf{x}^* \in \mathbb{R}^n$,

$$\exists 0 < L < 1: \quad \left\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\right\| \leq L \left\|\mathbf{x}^{(k)} - \mathbf{x}^*\right\| \quad \forall k \in \mathbb{N}_0 \,.$$

Terminology: The least upper bound for L gives the **rate of convergence**:

$$\text{rate} = \sup_{k \in \mathbb{N}_0} \frac{\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\right\|}{\left\|\mathbf{x}^{(k)} - \mathbf{x}^*\right\|} \,, \quad \mathbf{x}^* := \lim_{k \to \infty} \mathbf{x}^{(k)} \,. \tag{8.2.2.2}$$

Remark 8.2.2.3 (Impact of choice of norm)

Fact of convergence of an iteration is	independent	of choice of norm
Fact of linear convergence	depends	on choice of norm
Rate of linear convergence	depends	on choice of norm

The first statement is a consequence of the equivalence of all norms on the finite dimensional vector space $\mathbb{K}^n$:

Definition 8.2.2.4. Equivalence of norms → [Str09, Def. 4.4.2]

Two norms $\|\cdot\|_a$ and $\|\cdot\|_b$ on a vector space V are equivalent if

$$\exists \underline{C}, \overline{C} > 0: \quad \underline{C}\|v\|_a \leq \|v\|_b \leq \overline{C}\|v\|_a \quad \forall v \in V \,.$$

Theorem 8.2.2.5. Equivalence of all norms on finite-dimensional vector spaces → **[Str09, Satz 4.4.1]**

If $\dim V < \infty$ all norms (→ Def. 1.5.5.4) on V are equivalent (→ Def. 8.2.2.4).

⌟

Remark 8.2.2.6 (Detecting linear convergence) Often we will study the behavior of a consistent iterative method for a model problem in a numerical experiments and measure the norms of the iteration errors $\mathbf{e}^{(k)} := \mathbf{x}^{(k)} - \mathbf{x}^*$. How can we tell that the method enjoys linear convergence?

norms of iteration errors

$\updownarrow$

$\sim$ on straight line in lin-log plot

$$\left\|\mathbf{e}^{(k)}\right\| \leq L^k \left\|\mathbf{e}^{(0)}\right\| ,$$
$$\log(\left\|\mathbf{e}^{(k)}\right\|) \leq k\log(L) + \log(\left\|\mathbf{e}^{(0)}\right\|) .$$

● $\hat{=}$ linear cvg., ● $\hat{=}$ faster than linear cvg.

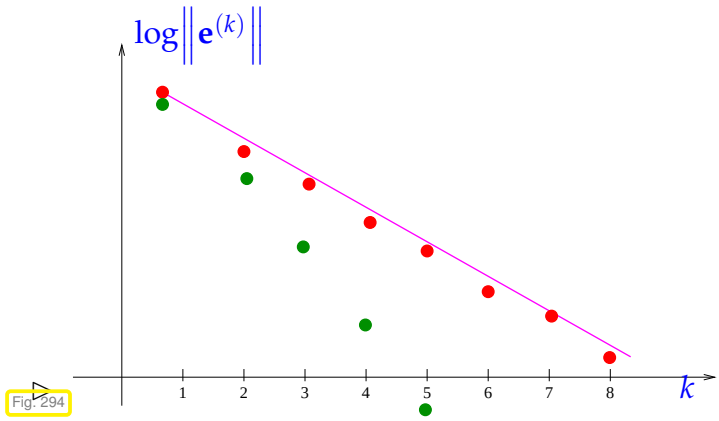


Fig. 294

Let us abbreviate the error norm in step k by $\epsilon_k := \left\|\mathbf{x}^{(k)} - \mathbf{x}^*\right\|$. In the case of linear convergence (see Def. 8.2.2.1) assume (with $0 < L < 1$)

$$\epsilon_{k+1} \approx L\epsilon_k \quad \Rightarrow \quad \log\epsilon_{k+1} \approx \log L + \log\epsilon_k \quad \Rightarrow \quad \log\epsilon_k \approx k\log L + \log\epsilon_0 . \tag{8.2.2.7}$$

We conclude that $\log L < 0$ determines the slope of the graph in lin-log error chart.

Related: guessing time complexity $O(n^\alpha)$ of an algorithm from measurements, see § 1.4.1.6.

Note the green dots ● in Fig. 294: Any "faster" convergence also qualifies as linear convergence in the strict sense of the definition. However, whenever this term is used, we tacitly imply, that no "faster convergence" prevails and that the estimates in (8.2.2.7) are sharp. ⌟

EXAMPLE 8.2.2.8 (Linearly convergent iteration) We consider the iteration ($n = 1$):

$$x^{(k+1)} = x^{(k)} + \frac{\cos(x^{(k)}) + 1}{\sin(x^{(k)})} .$$

In the C++ code Code 8.2.2.9 x has to be initialized with the different values for x_0.

Note: The final iterate $x^{(15)}$ replaces the exact solution x^* in the computation of the rate of convergence.

C++ code 8.2.2.9: Simple fixed point iteration in 1D ↪ GITLAB

```
2 void fpit(double x0, VectorXd &rates, VectorXd &err)
3 {
4   const Eigen::Index N = 15;
5   double x = x0; // initial guess
6   VectorXd y(N);
7
```

```
8     for (int i=0; i<N; ++i) {
9       x = x + (cos(x)+1)/sin(x);
10      y(i) = x;
11    }
12    err.resize(N); rates.resize(N);
13    err = y-VectorXd::Constant(N,x);
14    rates = err.bottomRows(N-1).cwiseQuotient(err.topRows(N-1));
15  }
```

k	$x^{(0)} = 0.4$		$x^{(0)} = 0.6$		$x^{(0)} = 1$	
	$x^{(k)}$	$\frac{\lvert x^{(k)}-x^{(15)}\rvert}{\lvert x^{(k-1)}-x^{(15)}\rvert}$	$x^{(k)}$	$\frac{\lvert x^{(k)}-x^{(15)}\rvert}{\lvert x^{(k-1)}-x^{(15)}\rvert}$	$x^{(k)}$	$\frac{\lvert x^{(k)}-x^{(15)}\rvert}{\lvert x^{(k-1)}-x^{(15)}\rvert}$
2	3.3887	0.1128	3.4727	0.4791	2.9873	0.4959
3	3.2645	0.4974	3.3056	0.4953	3.0646	0.4989
4	3.2030	0.4992	3.2234	0.4988	3.1031	0.4996
5	3.1723	0.4996	3.1825	0.4995	3.1224	0.4997
6	3.1569	0.4995	3.1620	0.4994	3.1320	0.4995
7	3.1493	0.4990	3.1518	0.4990	3.1368	0.4990
8	3.1454	0.4980	3.1467	0.4980	3.1392	0.4980

Rate of convergence ≈ 0.5

Plot of modulus of iteration errors for different initial guesses (see above table) ▷

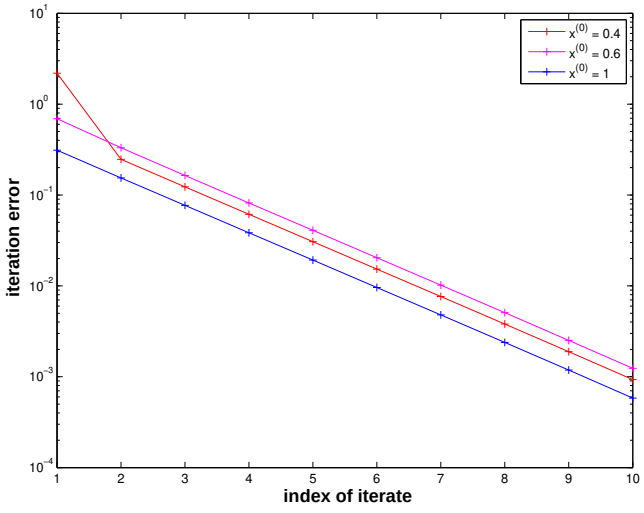


Fig. 295

Observation:

Linear convergence as in Def. 8.2.2.1

⇕

error graphs = straight lines in lin-log scale

→ Rem. 8.2.2.6

There are notions of convergence that guarantee a much faster (asymptotic) decay of the norms of the iteration errors than linear convergence from Def. 8.2.2.1.

Definition 8.2.2.10. Order of convergence → [Han02, Sect. 17.2], [DR08, Def. 5.14], [QSS00, Def. 6.1]

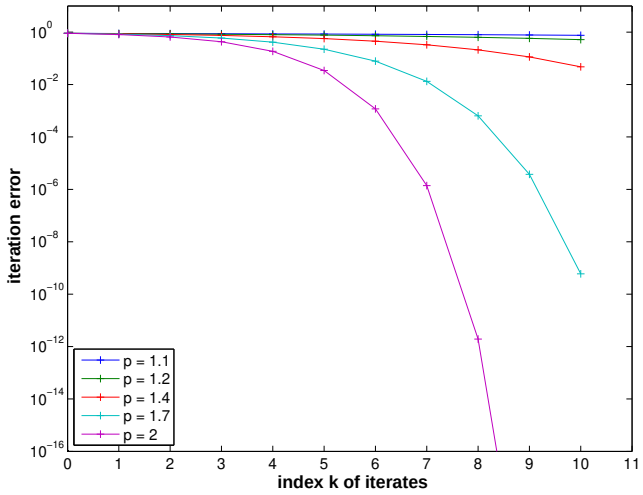
A *convergent* sequence $\mathbf{x}^{(k)}$, $k = 0,1,2,\ldots$, in $\mathbb{R}^n$ with limit $\mathbf{x}^* \in \mathbb{R}^n$ converges with **order** p, $p \geq 1$, if

$$\exists C > 0: \quad \left\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\right\| \leq C\left\|\mathbf{x}^{(k)} - \mathbf{x}^*\right\|^p \quad \forall k \in \mathbb{N}_0\,, \tag{8.2.2.11}$$

and, in addition, $C < 1$ in the case $p = 1$ (linear convergence → Def. 8.2.2.1).

Of course, the order p of convergence of an iterative method refers to the largest possible p in the def-

inition, that is, the error estimate will in general not hold, if p is replaced with $p+\epsilon$ for any $\epsilon > 0$, *cf.* Rem. 1.4.1.3.



◁ Qualitative error graphs for convergence of order p
(lin-log scale, k plotted versus $\log\left\|\mathbf{x}^{(k)}-\mathbf{x}^*\right\|$)

In the case of convergence of order p ($p>1$) according to Def. 8.2.2.10 and assuming sharpness of the error bound we obtain for the error norms $\epsilon_k := \left\|\mathbf{x}^{(k)}-\mathbf{x}^*\right\|$:

$$\epsilon_{k+1} \approx C\epsilon_k^p \quad \Rightarrow \quad \log\epsilon_{k+1} = \log C + p\log\epsilon_k \quad \Rightarrow \quad \log\epsilon_{k+1} = \log C\sum_{l=0}^{k} p^l + p^{k+1}\log\epsilon_0$$

$$\Rightarrow \quad \log\epsilon_{k+1} = -\frac{\log C}{p-1} + \left(\frac{\log C}{p-1} + \log\epsilon_0\right)p^{k+1} .$$

In this case, the error graph of the function $k \mapsto \log\epsilon_k$ is a concave ("downward bent") power curve (for sufficiently small ϵ_0 !)

Remark 8.2.2.12 (Detecting order $p>1$ of convergence) How can we guess the order of convergence ($\rightarrow$ Def. 8.2.2.10) from tabulated error norms measured in a numerical experiment?

Abbreviate by $\epsilon_k := \left\|\mathbf{x}^{(k)}-\mathbf{x}^*\right\|$ the norm of the iteration error.

$$\text{assume} \quad \epsilon_{k+1} \approx C\epsilon_k^p \quad \Rightarrow \quad \log\epsilon_{k+1} \approx \log C + p\log\epsilon_k \quad \Rightarrow \quad \frac{\log\epsilon_{k+1}-\log\epsilon_k}{\log\epsilon_k - \log\epsilon_{k-1}} \approx p .$$

➢ monitor the quotients $(\log\epsilon_{k+1}-\log\epsilon_k)/(\log\epsilon_k-\log\epsilon_{k-1})$ over several steps of the iteration. ⌟

EXAMPLE 8.2.2.13 (quadratic convergence = convergence of order 2) From your analysis course [Str09, Bsp. 3.3.2(iii)] recall the famous iteration for computing $\sqrt{a}$, $a>0$:

$$x^{(k+1)} = \frac{1}{2}(x^{(k)} + \frac{a}{x^{(k)}}) \quad \Rightarrow \quad |x^{(k+1)}-\sqrt{a}| = \frac{1}{2x^{(k)}}|x^{(k)}-\sqrt{a}|^2 . \tag{8.2.2.14}$$

By the arithmetic-geometric mean inequality (AGM) $\sqrt{ab} \le \frac{1}{2}(a+b)$ we conclude: $x^{(k)} > \sqrt{a}$ for $k \ge 1$. Therefore estimate from (8.2.2.14) means that the sequence from (8.2.2.14) converges with order 2 to $\sqrt{a}$.

Note: $x^{(k+1)} < x^{(k)}$ for all $k \ge 2$ ➢ $(x^{(k)})_{k\in\mathbb{N}_0}$ converges as a decreasing sequence that is bounded from below ($\rightarrow$ analysis course)

Numerical experiment: iterates for $a = 2$:

k	$x^{(k)}$	$e^{(k)} := x^{(k)} - \sqrt{2}$	$\log \frac{\lvert e^{(k)}\rvert}{\lvert e^{(k-1)}\rvert} : \log \frac{\lvert e^{(k-1)}\rvert}{\lvert e^{(k-2)}\rvert}$
0	2.00000000000000000	0.58578643762690485	
1	1.50000000000000000	0.08578643762690485	
2	1.41666666666666652	0.00245310429357137	1.850
3	1.41421568627450966	0.00000212390141452	1.984
4	1.41421356237468987	0.00000000000159472	2.000
5	1.41421356237309492	0.00000000000000022	0.630

Note the doubling of the number of correct digits in each step ! [impact of roundoff !]

The doubling of the number of significant digits for the iterates holds true for any quadratically convergent iteration:

Recall from Rem. 1.5.3.4 that the relative error ($\to$ Def. 1.5.3.3) tells the number of significant digits. Indeed, denoting the relative error in step k by δ_k, we have in the case of quadratic convergence.

$$\begin{aligned} x^{(k)} = x^*(1+\delta_k) \quad &\Rightarrow \quad x^{(k)} - x^* = \delta_k x^* \,. \\ \Rightarrow |x^*\delta_{k+1}| = |x^{(k+1)} - x^*| \le C|x^{(k)} - x^*|^2 &= C|x^*\delta_k|^2 \\ \Rightarrow \qquad |\delta_{k+1}| &\le C|x^*|\delta_k^2 \,. \end{aligned} \tag{8.2.2.15}$$

Note: $\delta_k \approx 10^{-\ell}$ means that $\mathbf{x}^{(k)}$ has ℓ significant digits.

Also note that if $C \approx 1$, then $\delta_k = 10^{-\ell}$ and (8.2.2.13) implies $\delta_{k+1} \approx 10^{-2\ell}$. ⌟

Review question(s) 8.2.2.16 (Iterative methods: speed of convergence)

(Q8.2.2.16.A) You have a table of the iterates $x^{(k)} \in \mathbb{R}$, $k = 1, \ldots, N$, $N \gg 1$, produced by some *convergent* iterative method. How do you proceed to produce evidence

1. for linear convergence,
2. for convergence with order p, $p > 1$.

Definition 8.2.2.1. Linear convergence

A sequence $\mathbf{x}^{(k)}$, $k = 0, 1, 2, \ldots$, in $\mathbb{R}^n$ **converges linearly** to $\mathbf{x}^* \in \mathbb{R}^n$,

$$\exists 0 < L < 1: \quad \left\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\right\| \le L \left\|\mathbf{x}^{(k)} - \mathbf{x}^*\right\| \quad \forall k \in \mathbb{N}_0 \,.$$

Definition 8.2.2.10. Order of convergence

A **convergent** sequence $\mathbf{x}^{(k)}$, $k = 0, 1, 2, \ldots$, in $\mathbb{R}^n$ with limit $\mathbf{x}^* \in \mathbb{R}^n$ converges with order p, if

$$\exists C > 0: \quad \left\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\right\| \le C \left\|\mathbf{x}^{(k)} - \mathbf{x}^*\right\|^p \quad \forall k \in \mathbb{N}_0 \,, \tag{8.2.2.17}$$

and, in addition, $C < 1$ in the case $p = 1$ (linear convergence $\to$ Def. 8.2.2.1).

(Q8.2.2.16.B) Consider the following iteration in $\mathbb{R}^2$

$$\mathbf{x}^{(k+1)} = \begin{bmatrix} 1/3 & 1 \\ 0 & 1/2 \end{bmatrix} \mathbf{x}^{(k)} \,. \tag{8.2.2.18}$$

What is $\lim_{k\to\infty} \mathbf{x}^{(k)}$? Does this iteration generate linearly convergent sequences?

Hint. The matrix in (8.2.2.18) can be diagonalized.

(Q8.2.2.16.C) Assume that all sequences $\left(\mathbf{x}^{(k)}\right)_{k\in\mathbb{N}_0}$ produced by a 1-point iteration $\mathbf{x}^{(k+1)} := \Phi(\mathbf{x}^{(k)})$, with iteration function $\Phi : \mathbb{R}^n \to \mathbb{R}^n$, satisfy

$$\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\right\| \leq C\left\|\mathbf{x}^{(k)} - \mathbf{x}^*\right\|^p \quad \forall k \in \mathbb{N}_0 \,,$$

for some $p > 1$, $C > 0$ and $\mathbf{x}^* \in \mathbb{R}^n$.

Give a sharp criterion for the initial guess $\mathbf{x}^{(0)} \in \mathbb{R}^n$ that guarantees convergence of the resulting sequence.

Hint. When will the sequence $\left(\left\|x^{(k)} - x^*\right\|\right)_{k\in\mathbb{N}_0}$ be decreasing?

△

8.2.3 Termination Criteria/Stopping Rules

Video tutorial for Section 8.2.3 "Iterative Methods: Termination Criteria/Stopping Rules": (14 minutes) Download link, tablet notes

Supplementary literature. Also discussed in [AG11, Sect. 3.1, p. 42].

As remarked above, usually (even without roundoff errors) an iteration will never arrive at an/the exact solution $\mathbf{x}^*$ after finitely many steps. Thus, we can only hope to compute an *approximate* solution by accepting $\mathbf{x}^{(K)}$ as result for some $K \in \mathbb{N}_0$. Termination criteria (stopping rules) are used to determine a suitable value for K.

For the sake of efficiency ▷ stop iteration when iteration error is just "small enough"
("small enough" depends on the concrete problem and user demands.)

§8.2.3.1 (Classification of termination criteria (stopping rules) for iterative solvers for non-linear systems of equations)

A termination criterion (stopping rule) is an algorithm deciding in each step of an iterative method whether to **STOP** or to **CONTINUE**.

We can distinguish two types of stopping rules:

A priori termination	A posteriori termination
Decision to stop based on information about F and $\mathbf{x}^{(0)}$, made before starting the iteration.	Beside $\mathbf{x}^{(0)}$ and F, also current and past iterates are used to decide about termination.

A termination criterion for a convergent iteration is deemed reliable, if it lets the iteration **CONTINUE**, until the iteration error $\mathbf{e}^{(k)} := \mathbf{x}^{(k)} - \mathbf{x}^*$, $\mathbf{x}^*$ the limit value, satisfies certain conditions (usually imposed before the start of the iteration). ⌟

§8.2.3.2 (Ideal termination) Writing $\mathbf{x}^*$ for the desired solution, termination criteria are usually meant to ensure accuracy of the final iterate $\mathbf{x}^{(K)}$ in the following sense:

$$\left\|\mathbf{x}^{(K)}-\mathbf{x}^*\right\| \le \tau_{\text{abs}}, \quad \tau_{\text{abs}} \;\hat{=}\; \text{prescribed (absolute) tolerance.}$$
or
$$\left\|\mathbf{x}^{(K)}-\mathbf{x}^*\right\| \le \tau_{\text{rel}}\|\mathbf{x}^*\|, \quad \tau_{\text{rel}} \;\hat{=}\; \text{prescribed (relative) tolerance.}$$

it seems that the second criterion, asking that the relative (→ Def. 1.5.3.3) iteration error be below a prescribed threshold, alone would suffice, but the absolute tolerance should be checked, if, by "accident", $\|\mathbf{x}^*\| = 0$ is possible. Otherwise, the iteration might fail to terminate at all.

Both criteria enter the "ideal (a posteriori) termination rule":

$$\text{STOP at step} \quad K = \min\left\{k \in \mathbb{N}_0 \colon \left\|\mathbf{x}^{(k)}-\mathbf{x}^*\right\| \le \left\{\begin{array}{c}\tau_{\text{abs}}\\ \textbf{or}\\ \tau_{\text{rel}}\|\mathbf{x}^*\|\end{array}\right.\right\}. \tag{8.2.3.3}$$

As pointed out before, the comparison $\left\|\mathbf{x}^{(k)}-\mathbf{x}^*\right\| \le \tau_{\text{abs}}$ is necessary to ensure termination when $\mathbf{x}^* = \mathbf{0}$ can happen.

Obviously, (8.2.3.3) achieves the optimum in terms of efficiency and reliability. Obviously, this termination criterion is not practical, because $\mathbf{x}^*$ is not known. Algorithmic feasible stopping rules have to replace $\left\|\mathbf{x}^{(k)}-\mathbf{x}^*\right\|$ and $\|\mathbf{x}^*\|$ with (upper/lower) bounds or estimates.

⌟

§8.2.3.4 (Practical termination criteria for iterations) The following termination criteria are commonly used in numerical codes:

① A priori termination: stop iteration after fixed number of steps (possibly depending on $\mathbf{x}^{(0)}$).

Drawback: hard to ensure prescribed accuracy!

(A priori $\hat{=}$ without actually taking into account the computed iterates, see § 8.2.3.1)

Invoking additional properties of either the non-linear system of equations $F(\mathbf{x}) = 0$ or the iteration it is sometimes possible to tell that for sure $\left\|\mathbf{x}^{(k)}-\mathbf{x}^*\right\| \le \tau$ for all $k \ge K$, though this K may be (significantly) larger than the optimal termination index from (8.2.3.3), see § 8.2.3.7.

② Residual based termination: STOP convergent iteration $\{\mathbf{x}^{(k)}\}_{k\in\mathbb{N}_0}$, when

$$\left\|F(\mathbf{x}^{(k)})\right\| \le \tau\,, \qquad \tau \hat{=} \text{ prescribed tolerance } > 0\,.$$

no guaranteed accuracy

Consider the case $n = 1$. If $F : D \subset \mathbb{R} \to \mathbb{R}$ is "flat" in the neighborhood of a zero x^*, then a small value of $|F(x)|$ does not mean that x is close to x^*.

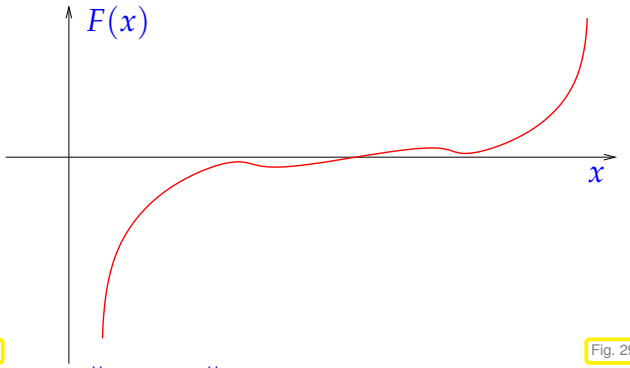


Fig. 296

$\left\|F(\mathbf{x}^{(k)})\right\|$ small $\not\Rightarrow$ $|x - x^*|$ small

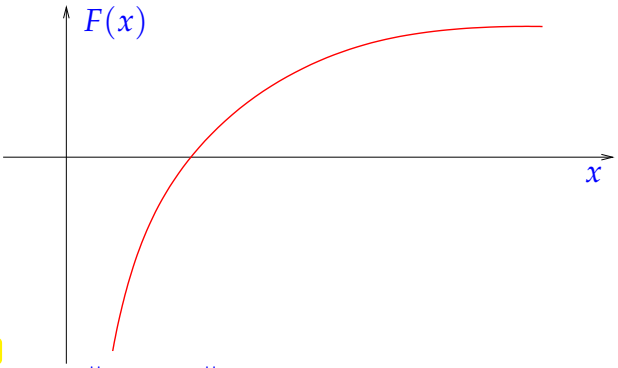


Fig. 297

$\left\|F(\mathbf{x}^{(k)})\right\|$ small $\Rightarrow$ $|x - x^*|$ small

③ Correction based termination:

STOP convergent iteration $\{\mathbf{x}^{(k)}\}_{k\in\mathbb{N}_0}$, when

$$\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\| \leq \begin{cases} \tau_{\text{abs}} \\ \textbf{or} \\ \tau_{\text{rel}}\left\|\mathbf{x}^{(k+1)}\right\| \end{cases}, \qquad \begin{matrix}\tau_{\text{abs}} \\ \tau_{\text{rel}}\end{matrix} \text{ prescribed } \begin{matrix}\text{absolute} \\ \text{relative}\end{matrix} \text{ tolerances } > 0\,.$$

Also for this criterion, we have no guarantee that (8.2.3.3) will be satisfied only remotely.

⌟

Remark 8.2.3.5 (STOP, when stationary in $\mathbb{M}$) A special variant of correction based termination exploits that $\mathbb{M}$ is finite! (→ Section 1.5.3)

Wait until (convergent) iteration becomes stationary in the discrete set $\mathbb{M}$ of machine numbers!

possibly grossly inefficient ! (always computes "up to machine precision")

C++ code 8.2.3.6: Square root iteration → Ex. 8.2.2.13 ↬ GITLAB

```
2  double sqrtit(double a)
3  {
4    double x_old = -1;
5    double x = a;
6    while (x_old != x) {
7      x_old = x;
8      x = 0.5*(x+a/x);
9    }
10   return x;
11 }
```

⌟

§8.2.3.7 (A posteriori termination criterion for linearly convergent iterations → [DR08, Lemma 5.17, 5.19]) Let us assume that we know that an iteration is linearly convergent (→ Def. 8.2.2.1) with rate of convergence $0 < L < 1$:

Definition 8.2.2.1. Linear convergence

A sequence $\mathbf{x}^{(k)}$, $k = 0, 1, 2, \ldots$, in $\mathbb{R}^n$ **converges linearly** to $\mathbf{x}^* \in \mathbb{R}^n$,

$$\exists 0 < L < 1: \quad \left\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\right\| \leq L\left\|\mathbf{x}^{(k)} - \mathbf{x}^*\right\| \quad \forall k \in \mathbb{N}_0\,.$$

The following simple manipulations give an a posteriori termination criterion (for linearly convergent itera-

tions with rate of convergence $0 < L < 1$):

$$\left\|\mathbf{x}^{(k)} - \mathbf{x}^*\right\| \overset{\triangle\text{-inequ.}}{\leq} \left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\| + \left\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\right\| \leq \left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\| + L\left\|\mathbf{x}^{(k)} - \mathbf{x}^*\right\| .$$

Iterates satisfy: $$\boxed{\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\right\| \leq \frac{L}{1-L}\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\|} . \tag{8.2.3.8}$$

This suggests that we take the right hand side of (8.2.3.8) as a posteriori error bound and use it instead of the inaccessible $\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\right\|$ for checking absolute and relative accuracy in (8.2.3.3). The resulting termination criterium will be reliable (→ § 8.2.3.1), since we will certainly have achieved the desired accuracy when we stop the iteration.

Estimating the rate of convergence L might be difficult.

Pessimistic estimate for L will not compromise reliability.

(Using $\widetilde{L} > L$ in (8.2.3.8) still yields a valid *upper bound* for $\left\|\mathbf{x}^{(k)} - \mathbf{x}^*\right\|$. Hence, the result can be trusted, though we might have wasted computational resources by needlessly carrying on with the iteration.) ⌟

EXAMPLE 8.2.3.9 (A posteriori error bound for linearly convergent iteration) We revisit the iteration of Ex. 8.2.2.8:

$$x^{(k+1)} = x^{(k)} + \frac{\cos x^{(k)} + 1}{\sin x^{(k)}} \quad \Rightarrow x^{(k)} \to \pi \quad \text{for } x^{(0)} \text{ close to } \pi .$$

Observed rate of convergence: $L = 1/2$

Error and error bound for $x^{(0)} = 0.4$:

k	$\lvert x^{(k)} - \pi\rvert$	$\frac{L}{1-L}\lvert x^{(k)} - x^{(k-1)}\rvert$	slack of bound
1	2.191562221997101	4.933154875586894	2.741592653589793
2	0.247139097781070	1.944423124216031	1.697284026434961
3	0.122936737876834	0.124202359904236	0.001265622027401
4	0.061390835206217	0.061545902670618	0.000155067464401
5	0.030685773472263	0.030705061733954	0.000019288261691
6	0.015341682696235	0.015344090776028	0.000002408079792
7	0.007670690889185	0.007670991807050	0.000000300917864
8	0.003835326638666	0.003835364250520	0.000000037611854
9	0.001917660968637	0.001917665670029	0.000000004701392
10	0.000958830190489	0.000958830778147	0.000000000587658
11	0.000479415058549	0.000479415131941	0.000000000073392
12	0.000239707524646	0.000239707533903	0.000000000009257
13	0.000119853761949	0.000119853762696	0.000000000000747
14	0.000059926881308	0.000059926880641	0.000000000000667
15	0.000029963440745	0.000029963440563	0.000000000000181

Hence: the a posteriori error bound is highly accurate in this case!

⌟

Review question(s) 8.2.3.10 (Iterative methods: termination criteria)

(Q8.2.3.10.A) Let $x_0 \in \mathbb{R}$ be an approximation for a zero of a continuously differentiable function $f : \mathbb{R} \to \mathbb{R}$, that is, $f(x_0) \approx 0$. Argue, why

$$\Delta s := \left| \frac{f(x_0)}{f'(x_0)} \right|$$

can be used to estimate the error $x_0 - x^*$, where x^* is a zero of f close to x_0.

(Q8.2.3.10.B) For a *linearly convergent* sequence $\left(\mathbf{x}^{(k)}\right)_{k\in\mathbb{N}_0} \subset \mathbb{R}^n$ with limit $\mathbf{x}^*$ we found the a posteriori error estimate

$$\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\right\| \leq \frac{L}{1-L}\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\| , \tag{8.2.3.8}$$

where $0 \leq L < 1$ is the rate of (linear) convergence. Based on (8.2.3.8), the following stopping rule is employed

$$\text{STOP, as soon as} \quad \frac{L}{1-L}\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\| \leq \tau_{\text{abs}} ,$$

where $\tau_{\text{abs}} > 0$ is a user-supplied threshold. Assume that the true ("sharp") rate of linear convergence is $L \in [0, 1[$, that is,

$$\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\right\| \approx L\left\|\mathbf{x}^{(k)} - \mathbf{x}^*\right\| \quad \forall k \in \mathbb{N}_0 ,$$

but for stopping rule a larger value $\overline{L}$, $L < \overline{L} < 1$ is used. Discuss what this means for the number of steps until termination and the absolute accuracy of the returned approximation.

△

8.3 Fixed-Point Iterations

Video tutorial for Section 8.3 "Fixed-Point Iterations": (12 minutes) Download link, tablet notes

Supplementary literature. The contents of this section are also treated in [DR08, Sect. 5.3], [QSS00, Sect. 6.3], [AG11, Sect .3.3]

As before we consider a non-linear system of equations $F(\mathbf{x}) = \mathbf{0}$, $F : D \subset \mathbb{R}^n \mapsto \mathbb{R}^n$.

1-point stationary iterative methods, see (8.2.1.5), for $F(\mathbf{x}) = \mathbf{0}$ are also called fixed point iterations.

A fixed point iteration is defined by iteration function $\Phi : U \subset \mathbb{R}^n \mapsto \mathbb{R}^n$:

iteration function $\Phi : U \subset \mathbb{R}^n \mapsto \mathbb{R}^n$
initial guess $\mathbf{x}^{(0)} \in U$

$$\succ \quad \underbrace{\boxed{\text{iterates } (\mathbf{x}^{(k)})_{k\in\mathbb{N}_0}: \quad \mathbf{x}^{(k+1)} := \Phi(\mathbf{x}^{(k)})}}_{\to \text{ 1-point method, } \textit{cf.} \text{ (8.2.1.5)}} .$$

Here, U designates the domain of definition of the iteration function Φ.

Note that the sequence of iterates need not be well defined: $\mathbf{x}^{(k)} \notin U$ possible !

8.3.1 Consistent Fixed-Point Iterations

Next, we specialize Def. 8.2.1.7 for fixed point iterations:

Definition 8.3.1.1. Consistency of fixed point iterations, *c.f.* Def. 8.2.1.7

A fixed point iteration $\mathbf{x}^{(k+1)} = \Phi(\mathbf{x}^{(k)})$ is consistent with $F(\mathbf{x}) = \mathbf{0}$, if, for $\mathbf{x} \in U \cap D$,

$$F(\mathbf{x}) = \mathbf{0} \quad \Leftrightarrow \quad \Phi(\mathbf{x}) = \mathbf{x} \,.$$

Fig. 298

Note: iteration function Φ continuous **and** fixed point iteration (locally) convergent to $\mathbf{x}^* \in U$

$$\Downarrow$$

$\mathbf{x}^*$ is a **fixed point** of Φ: $\Phi(\mathbf{x}^*) = \mathbf{x}^*$.

This is an immediate consequence of the fact that for a continuous function limits and function evaluations commute [Str09, Sect. 4.1].

General construction of fixed point iterations that is consistent with $F(\mathbf{x}) = \mathbf{0}$:

❶ Rewrite equivalently $F(\mathbf{x}) = \mathbf{0} \quad \Leftrightarrow \quad \Phi(\mathbf{x}) = \mathbf{x}$ and then
❷ use the fixed point iteration

$$\mathbf{x}^{(k+1)} := \Phi(\mathbf{x}^{(k)}) \,. \tag{8.3.1.2}$$

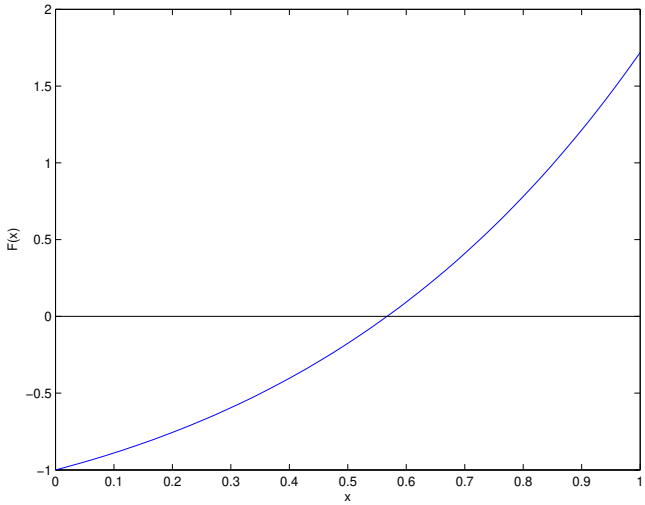
Note: There are *many ways* to transform $F(\mathbf{x}) = 0$ into a fixed point form **!**

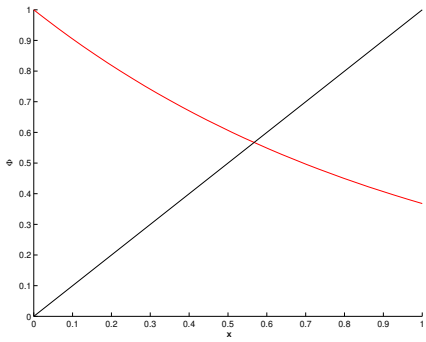
EXPERIMENT 8.3.1.3 (Many choices for consistent fixed point iterations) In this example we construct three different consistent fixed point iteration for a single scalar ($n = 1$) non-linear equation $F(x) = 0$. In numerical experiments we will see that they behave very differently.

$$F(x) = xe^x - 1 \,, \quad x \in [0,1] \,.$$

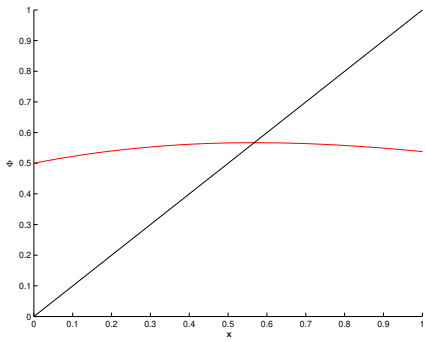
Different fixed-point forms:

$$\begin{aligned} \Phi_1(x) &= e^{-x} \,, \\ \Phi_2(x) &= \frac{1+x}{1+e^x} \,, \\ \Phi_3(x) &= x + 1 - xe^x \,. \end{aligned}$$

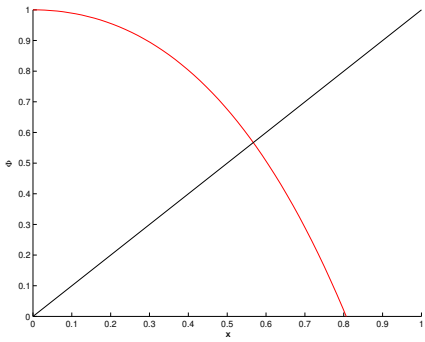




function Φ_1



function Φ_2



function Φ_3

With the same intial guess $x^{(0)} = 0.5$ for all three fixed point iterations we obtain the following iterates:

k	$x^{(k+1)} := \Phi_1(x^{(k)})$	$x^{(k+1)} := \Phi_2(x^{(k)})$	$x^{(k+1)} := \Phi_3(x^{(k)})$
0	0.500000000000000	0.500000000000000	0.500000000000000
1	0.606530659712633	0.566311003197218	0.675639364649936
2	0.545239211892605	0.567143165034862	0.347812678511202
3	0.579703094878068	0.567143290409781	0.855321409174107
4	0.560064627938902	0.567143290409784	-0.156505955383169
5	0.571172148977215	0.567143290409784	0.977326422747719
6	0.564862946980323	0.567143290409784	-0.619764251895580
7	0.568438047570066	0.567143290409784	0.713713087416146
8	0.566409452746921	0.567143290409784	0.256626649129847
9	0.567559634262242	0.567143290409784	0.924920676910549
10	0.566907212935471	0.567143290409784	-0.407422405542253

We can also tabulate the modulus of the iteration error and mark correct digits with red:

k	$\lvert x_1^{(k+1)} - x^*\rvert$	$\lvert x_2^{(k+1)} - x^*\rvert$	$\lvert x_3^{(k+1)} - x^*\rvert$
0	0.067143290409784	0.067143290409784	0.067143290409784
1	0.039387369302849	0.000832287212566	0.108496074240152
2	0.021904078517179	0.000000125374922	0.219330611898582
3	0.012559804468284	0.000000000000003	0.288178118764323
4	0.007078662470882	0.000000000000000	0.723649245792953
5	0.004028858567431	0.000000000000000	0.410183132337935
6	0.002280343429460	0.000000000000000	1.186907542305364
7	0.001294757160282	0.000000000000000	0.146569797006362
8	0.000733837662863	0.000000000000000	0.310516641279937
9	0.000416343852458	0.000000000000000	0.357777386500765
10	0.000236077474313	0.000000000000000	0.974565695952037

Observed: linear convergence of $x_1^{(k)}$, quadratic convergence of $x_2^{(k)}$,
no convergence (erratic behavior of $x_3^{(k)}$) ($x_i^{(0)} = 0.5$ in all cases).

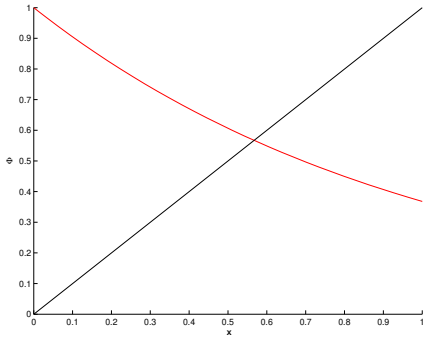
Question: Can we explain/forecast the behaviour of a fixed point iteration?

8.3.2 Convergence of Fixed-Point Iterations

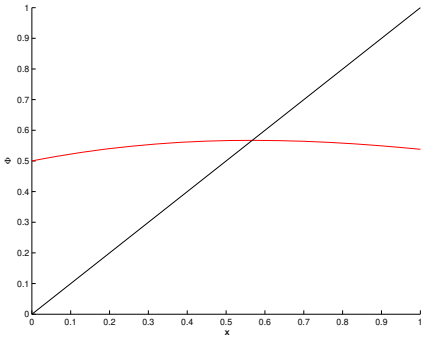
In this section we will try to find easily verifiable conditions that ensure convergence (of a certain order) of fixed point iterations. It will turn out that these conditions are surprisingly simple and general.

EXPERIMENT 8.3.2.1 (Exp. 8.3.1.3 revisited)

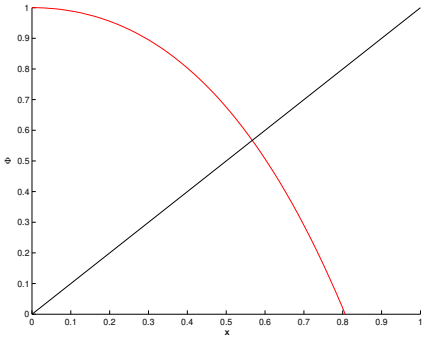
In Exp. 8.3.1.3 we observed vastly different behavior of different fixed point iterations for $n = 1$. Is it possible to predict this from the shape of the graph of the iteration functions?



Φ_1: linear convergence ?



Φ_2: quadratic convergence ?



function Φ_3: no convergence

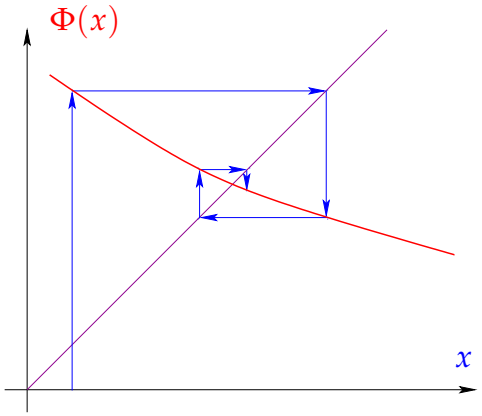
⌟

Remark 8.3.2.2 (Visualization of fixed point iterations in 1D)

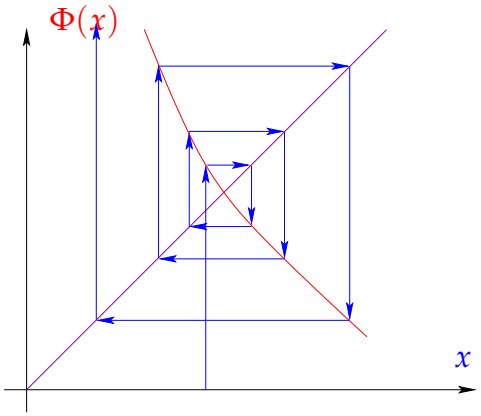
1D setting ($n = 1$): $\Phi : \mathbb{R} \mapsto \mathbb{R}$ continuously differentiable, $\Phi(x^*) = x^*$

fixed point iteration: $x^{(k+1)} = \Phi(x^{(k)})$

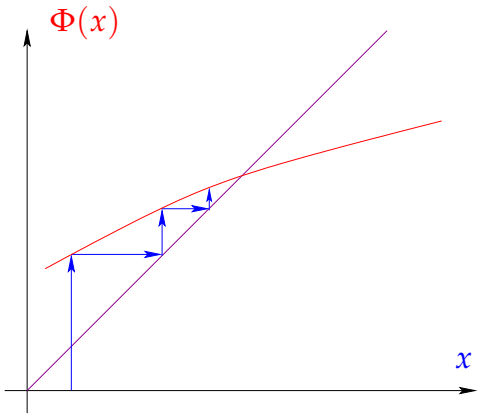
In 1D it is possible to visualize the different convergence behavior of fixed point iterations: In order to construct $x^{(k+1)}$ from $x^{(k)}$ one moves vertically to $(x^{(k)}, x^{(k+1)} = \Phi(x^{(k)}))$, then horizontally to the angular bisector of the first/third quadrant, that is, to the point $(x^{(k+1)}, x^{(k+1)})$. Returning vertically to the abscissa gives $x^{(k+1)}$.



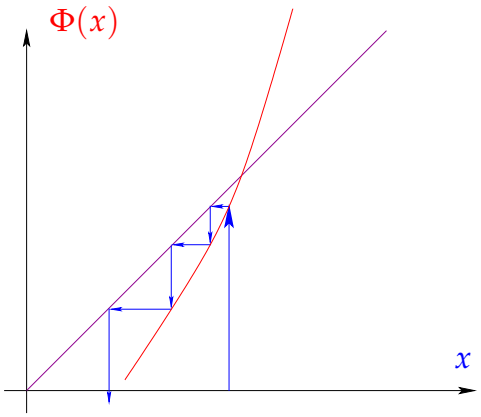
$-1 < \Phi'(x^*) \leq 0$ ➢ convergence



$\Phi'(x^*) < -1$ ➢ divergence



$0 \leq \Phi'(x^*) < 1$ ➢ convergence



$1 < \Phi'(x^*)$ ➢ divergence

Numerical examples for iteration functions ➢ Exp. 8.3.1.3, iteration functions Φ_1 and Φ_3

It seems that the slope of the iteration function Φ in the fixed point, that is, in the point where it intersects

the bisector of the first/third quadrant, is crucial. ⌟

Now we investigate rigorously, when a fixed point iteration will lead to a convergent iteration with a particular qualitative kind of convergence according to Def. 8.2.2.10.

Definition 8.3.2.3. Contractive mapping

$\Phi : U \subset \mathbb{R}^n \mapsto \mathbb{R}^n$ is contractive (w.r.t. norm $\|\cdot\|$ on $\mathbb{R}^n$), if

$$\exists L < 1: \quad \|\Phi(\mathbf{x}) - \Phi(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\| \quad \forall \mathbf{x}, \mathbf{y} \in U \, . \tag{8.3.2.4}$$

A simple consideration: if $\Phi(\mathbf{x}^*) = \mathbf{x}^*$ (fixed point), then a fixed point iteration induced by a contractive mapping Φ satisfies

$$\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\right\| = \left\|\Phi(\mathbf{x}^{(k)}) - \Phi(\mathbf{x}^*)\right\| \overset{(8.3.2.4)}{\leq} L\left\|\mathbf{x}^{(k)} - \mathbf{x}^*\right\| ,$$

that is, the iteration converges (at least) linearly (→ Def. 8.2.2.1).

Note that Φ contractive $\Rightarrow$ Φ has *at most one* fixed point.

Remark 8.3.2.5 (Banach's fixed point theorem → [Str09, Satz 6.5.2],[DR08, Satz 5.8]) A key theorem in calculus (also functional analysis):

Theorem 8.3.2.6. Banach's fixed point theorem

If $D \subset \mathbb{K}^n$ ($\mathbb{K} = \mathbb{R}, \mathbb{C}$) closed and bounded and $\Phi : D \mapsto D$ satisfies

$$\exists L < 1: \quad \|\Phi(\mathbf{x}) - \Phi(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\| \quad \forall \mathbf{x}, \mathbf{y} \in D \, ,$$

then there is a unique fixed point $\mathbf{x}^* \in D$, $\Phi(\mathbf{x}^*) = \mathbf{x}^*$, which is the limit of the sequence of iterates $\mathbf{x}^{(k+1)} := \Phi(x^{(k)})$ for any $\mathbf{x}^{(0)} \in D$.

Proof. Proof based on 1-point iteration $\mathbf{x}^{(k)} = \Phi(\mathbf{x}^{(k-1)})$, $\mathbf{x}^{(0)} \in D$:

$$\left\|\mathbf{x}^{(k+N)} - \mathbf{x}^{(k)}\right\| \leq \sum_{j=k}^{k+N-1} \left\|\mathbf{x}^{(j+1)} - \mathbf{x}^{(j)}\right\| \leq \sum_{j=k}^{k+N-1} L^j \left\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\right\|$$

$$\leq \frac{L^k}{1-L}\left\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\right\| \xrightarrow{k\to\infty} 0 \, .$$

$(\mathbf{x}^{(k)})_{k\in\mathbb{N}_0}$ Cauchy sequence ➤ convergent $\mathbf{x}^{(k)} \xrightarrow{k\to\infty} \mathbf{x}^*$.
Continuity of Φ ➤ $\Phi(\mathbf{x}^*) = \mathbf{x}^*$. Uniqueness of fixed point is evident.

□ ⌟

A simple criterion for a differentiable Φ to be contractive:

Lemma 8.3.2.7. Sufficient condition for local linear convergence of fixed point iteration → [Han02, Thm. 17.2], [DR08, Cor. 5.12]

If $\Phi : U \subset \mathbb{R}^n \mapsto \mathbb{R}^n$, $\Phi(\mathbf{x}^*) = \mathbf{x}^*$,$\Phi$ differentiable in $\mathbf{x}^*$, and $\|\mathrm{D}\,\Phi(\mathbf{x}^*)\| < 1$, then the fixed point iteration

$$\mathbf{x}^{(k+1)} := \Phi(\mathbf{x}^{(k)}) \,, \tag{8.3.1.2}$$

converges locally and at least linearly that is *matrix norm, Def. 1.5.5.10 !*

$$\exists 0 \le L < 1\colon \quad \left\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\right\| \le L\left\|\mathbf{x}^{(k)} - \mathbf{x}^*\right\| \quad \forall k \in \mathbb{N}_0 \,,$$

provided that the initial guess $\mathbf{x}^{(0)}$ belongs to some neighborhood of $\mathbf{x}^*$.

✎ notation: $\mathrm{D}\,\Phi(\mathbf{x}) \hat{=}$ Jacobian (*ger.:* Jacobi-Matrix) of Φ at $\mathbf{x} \in D$ → [Str09, Sect. 7.6]

$$\mathrm{D}\,\Phi(\mathbf{x}) = \left[\frac{\partial \Phi_i}{\partial x_j}(\mathbf{x})\right]_{i,j=1}^{n} = \begin{bmatrix} \frac{\partial \Phi_1}{\partial x_1}(\mathbf{x}) & \frac{\partial \Phi_1}{\partial x_2}(\mathbf{x}) & \cdots & \cdots & \frac{\partial \Phi_1}{\partial x_n}(\mathbf{x}) \\ \frac{\partial \Phi_2}{\partial x_1}(\mathbf{x}) & & & & \frac{\partial \Phi_2}{\partial x_n}(\mathbf{x}) \\ \vdots & & & & \vdots \\ \frac{\partial \Phi_n}{\partial x_1}(\mathbf{x}) & \frac{\partial \Phi_n}{\partial x_2}(\mathbf{x}) & \cdots & \cdots & \frac{\partial \Phi_n}{\partial x_n}(\mathbf{x}) \end{bmatrix} . \tag{8.3.2.8}$$

A "visualization" of the statement of Lemma 8.3.2.7 has been provided in Rem. 8.3.2.2: The iteration converges *locally*, if Φ is flat in a neighborhood of x^*, it will diverge, if Φ is steep there.

Proof. (of Lemma 8.3.2.7) By the definition of the derivative

$$\|\Phi(\mathbf{y}) - \Phi(\mathbf{x}^*) - D\Phi(\mathbf{x}^*)(\mathbf{y} - \mathbf{x}^*)\| \le \psi(\|\mathbf{y} - \mathbf{x}^*\|)\|\mathbf{y} - \mathbf{x}^*\| \,,$$

with $\psi : \mathbb{R}_0^+ \mapsto \mathbb{R}_0^+$ satisfying $\lim_{t\to 0} \psi(t) = 0$.

Choose $\delta > 0$ such that

$$L := \psi(t) + \|D\Phi(\mathbf{x}^*)\| \le \tfrac{1}{2}(1 + \|D\Phi(\mathbf{x}^*)\|) < 1 \quad \forall 0 \le t < \delta \,.$$

By inverse triangle inequality

$$|\|\mathbf{a}\| - \|\mathbf{b}\|| \le \|\mathbf{a} + \mathbf{b}\| \quad \forall \mathbf{a}, \mathbf{b} \in \mathbb{R}^n \,,$$

and thanks to $\Phi(\mathbf{x}^*) = \mathbf{x}^*$ we obtain for the fixed-point iteration

$$\begin{aligned} &\|\Phi(\mathbf{x}) - \mathbf{x}^*\| - \|D\Phi(\mathbf{x}^*)(\mathbf{x} - \mathbf{x}^*)\| \le \psi(\|\mathbf{x} - \mathbf{x}^*\|)\|\mathbf{x} - \mathbf{x}^*\| \\ \blacktriangleright \quad &\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\right\| \le (\psi(t) + \|D\Phi(\mathbf{x}^*)\|)\left\|\mathbf{x}^{(k)} - \mathbf{x}^*\right\| \le L\left\|\mathbf{x}^{(k)} - \mathbf{x}^*\right\| \,, \end{aligned}$$

if $\left\|\mathbf{x}^{(k)} - \mathbf{x}^*\right\| < \delta$.

□

Lemma 8.3.2.9. Sufficient condition for linear convergence of fixed point iteration

Let U be convex and $\Phi : U \subset \mathbb{R}^n \mapsto \mathbb{R}^n$ be continuously differentiable with

$$L := \sup_{\mathbf{x}\in U} \|D\Phi(\mathbf{x})\| < 1 \, .$$

If $\Phi(\mathbf{x}^*) = \mathbf{x}^*$ for some interior point $\mathbf{x}^* \in U$, then the fixed point iteration $\mathbf{x}^{(k+1)} = \Phi(\mathbf{x}^{(k)})$ with $\mathbf{x}^{(0)} \in U$ converges to $\mathbf{x}^*$ at least linearly with rate L.

Recall: $U \subset \mathbb{R}^n$ convex :$\Leftrightarrow$ $(t\mathbf{x} + (1-t)\mathbf{y}) \in U$ for all $\mathbf{x}, \mathbf{y} \in U$, $0 \le t \le 1$

Proof. (of Lemma 8.3.2.9) By the mean value theorem

$$\begin{aligned} \Phi(\mathbf{x}) - \Phi(\mathbf{y}) = \int_0^1 D\Phi(\mathbf{x} + \tau(\mathbf{y}-\mathbf{x}))(\mathbf{y}-\mathbf{x}) \, \mathrm{d}\tau \quad \forall \mathbf{x}, \mathbf{y} \in \mathrm{dom}(\Phi) \, . \\ \Rightarrow \quad \|\Phi(\mathbf{x}) - \Phi(\mathbf{y})\| \le L\|\mathbf{y} - \mathbf{x}\| \, , \\ \Rightarrow \quad \left\|(\mathbf{x})^{(k+1)} - \mathbf{x}^*\right\| \le L\left\|\mathbf{x}^{(k)} - \mathbf{x}^*\right\| \, . \end{aligned}$$

We find that Φ is contractive on U with unique fixed point $\mathbf{x}^*$, to which $\mathbf{x}^{(k)}$ converges linearly for $k \to \infty$. □

Remark 8.3.2.10 (Bound for *asymptotic* rate of linear convergence) By asymptotic rate of a linearly converging iteration we mean the contraction factor for the norm of the iteration error that we can expect, when we are already very close to the limit $\mathbf{x}^*$.

If $0 < \|D\Phi(\mathbf{x}^*)\| < 1$, $\mathbf{x}^{(k)} \approx \mathbf{x}^*$ then the (worst) asymptotic rate of linear convergence is $L = \|D\Phi(x^*)\|$ ⌟

EXAMPLE 8.3.2.11 (Multidimensional fixed point iteration) In this example we encounter the first genuine *system* of non-linear equations and apply Lemma 8.3.2.9 to it.

$$\begin{array}{ccc} \text{System of equations} & \text{in} & \text{fixed point form:} \\ \begin{cases} x_1 - c(\cos x_1 - \sin x_2) = 0 \\ (x_1 - x_2) - c\sin x_2 = 0 \end{cases} & \Rightarrow & \begin{cases} c(\cos x_1 - \sin x_2) = x_1 \\ c(\cos x_1 - 2\sin x_2) = x_2 \end{cases} . \end{array}$$

$$\text{Define:} \quad \Phi\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = c\begin{bmatrix} \cos x_1 - \sin x_2 \\ \cos x_1 - 2\sin x_2 \end{bmatrix} \quad \Rightarrow \quad D\Phi\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = -c\begin{bmatrix} \sin x_1 & \cos x_2 \\ \sin x_1 & 2\cos x_2 \end{bmatrix} .$$

Choose *appropriate* norm: $\|\cdot\|$ = ∞-norm $\|\cdot\|_\infty$ ($\rightarrow$ Example 1.5.5.12) ;

$$\text{if } c < \frac{1}{3} \quad \Rightarrow \quad \|D\Phi(\mathbf{x})\|_\infty < 1 \quad \forall \mathbf{x} \in \mathbb{R}^2 \, ,$$

➢ (at least) linear convergence of the fixed point iteration. The existence of a fixed point is also guaranteed, because Φ maps into the closed set $[-3,3]^2$. Thus, the Banach fixed point theorem, Thm. 8.3.2.6, can be applied. ⌟

What about higher order convergence ($\rightarrow$ Def. 8.2.2.10, *cf.* Φ_2 in Ex. 8.3.1.3)? Also in this case we should study the derivatives of the iteration functions in the fixed point (limit point).

We give a refined convergence result only for $n = 1$ (scalar case, $\Phi : \mathrm{dom}(\Phi) \subset \mathbb{R} \mapsto \mathbb{R}$):

Theorem 8.3.2.12. Taylor's formula → **[Str09, Sect. 5.5]**

If $\Phi : U \subset \mathbb{R} \mapsto \mathbb{R}$, U interval, is $m+1$ times continuously differentiable, $x \in U$

$$\Phi(y) - \Phi(x) = \sum_{k=1}^{m} \frac{1}{k!} \Phi^{(k)}(x)(y-x)^k + O(|y-x|^{m+1}) \quad \forall y \in U \,. \tag{8.3.2.13}$$

Now apply Taylor expansion (8.3.2.13) to iteration function Φ:

If $\Phi(x^*) = x^*$ and $\Phi : \mathrm{dom}(\Phi) \subset \mathbb{R} \mapsto \mathbb{R}$ is "sufficiently smooth", it tells us that

$$x^{(k+1)} - x^* = \Phi(x^{(k)}) - \Phi(x^*) = \sum_{l=1}^{m} \frac{1}{l!} \Phi^{(l)}(x^*)(x^{(k)} - x^*)^l + O(|x^{(k)} - x^*|^{m+1}) \,. \tag{8.3.2.14}$$

Here we used the Landau symbol $O(\cdot)$ to describe the local behavior of a remainder term in the vicinity of x^*

Lemma 8.3.2.15. Higher order local convergence of fixed point iterations

If $\Phi : U \subset \mathbb{R} \mapsto \mathbb{R}$ is $m+1$ times continuously differentiable, $\Phi(x^*) = x^*$ for some x^* in the interior of U, and $\Phi^{(l)}(x^*) = 0$ for $l = 1, \ldots, m$, $m \geq 1$, then the fixed point iteration (8.3.1.2) *converges locally to x^* with order $\geq m+1$ (→ Def. 8.2.2.10).*

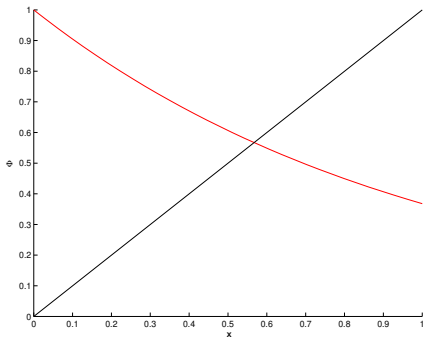
Proof. For neighborhood $\mathcal{U}$ of x^*

$$\text{(8.3.2.14)} \quad \Rightarrow \quad \exists C > 0\text{:} \quad |\Phi(y) - \Phi(x^*)| \leq C\,|y - x^*|^{m+1} \quad \forall y \in \mathcal{U} \,.$$
$$\delta^m C < 1/2 : \quad |x^{(0)} - x^*| < \delta \quad \Rightarrow \quad |x^{(k)} - x^*| < 2^{-k}\delta \quad \succ \text{ local convergence .}$$

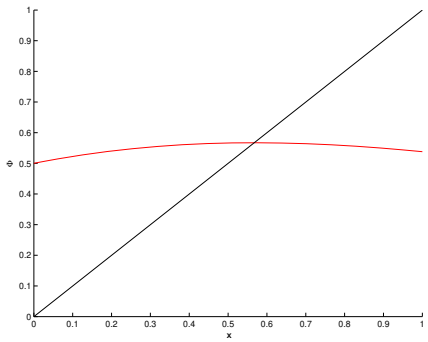
Then appeal to (8.3.2.14)

□

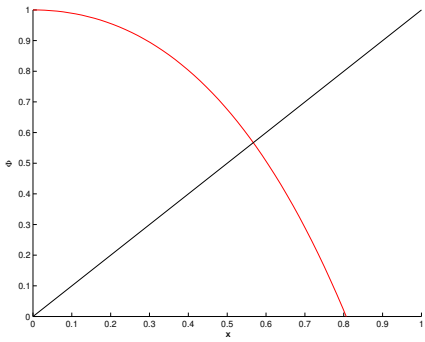
EXPERIMENT 8.3.2.16 (Exp. 8.3.2.1 continued) Now, Lemma 8.3.2.9 and Lemma 8.3.2.15 permit us a precise prediction of the (asymptotic) convergence we can expect from the different fixed point iterations studied in Exp. 8.3.1.3.



function Φ_1



function Φ_2



function Φ_3

$$\Phi_2'(x) = \frac{1 - xe^x}{(1+e^x)^2} = 0 \quad \text{, if} \quad xe^x - 1 = 0 \quad \text{hence quadratic convergence ! .}$$

Since $x^* e^{x^*} - 1 = 0$, simple computations yield

$$\Phi_1'(x) = -e^{-x} \quad \Rightarrow \quad \Phi_1'(x^*) = -x^* \approx -0.56 \quad \text{hence local linear convergence .}$$

$$\Phi_3'(x) = 1 - xe^x - e^x \quad \Rightarrow \quad \Phi_3'(x^*) = -\frac{1}{x^*} \approx -1.79 \quad \text{hence no convergence .}$$

⌟

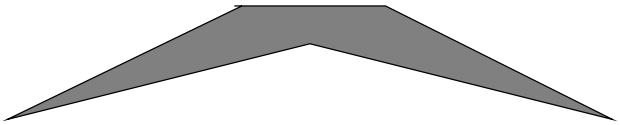
Remark 8.3.2.17 (Termination criterion for contractive fixed point iteration)

We recall the considerations of § 8.2.3.7 about a termination criterion for contractive fixed point iteration (= linearly convergence fixed point iteration → Def. 8.2.2.1), *c.f.* (8.3.2.4), with contraction factor (= rate of convergence) $0 \leq L < 1$:

$$\begin{aligned}\left\|\mathbf{x}^{(k+m)} - \mathbf{x}^{(k)}\right\| &\overset{\triangle\text{-ineq.}}{\leq} \sum_{j=k}^{k+m-1} \left\|\mathbf{x}^{(j+1)} - \mathbf{x}^{(j)}\right\| \leq \sum_{j=k}^{k+m-1} L^{j-k}\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\| \\ &= \frac{1-L^m}{1-L}\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\| \leq \frac{1-L^m}{1-L}L^{k-l}\left\|\mathbf{x}^{(l+1)} - \mathbf{x}^{(l)}\right\| .\end{aligned}$$

Hence, for $m \to \infty$, with $\mathbf{x}^* := \lim_{k\to\infty} \mathbf{x}^{(k)}$ we find the estimate

$$\left\|\mathbf{x}^* - \mathbf{x}^{(k)}\right\| \leq \frac{L^{k-l}}{1-L}\left\|\mathbf{x}^{(l+1)} - \mathbf{x}^{(l)}\right\| . \tag{8.3.2.18}$$



Set $l = 0$ in (8.3.2.18)	Set $l = k-1$ in (8.3.2.18)
a priori termination criterion	a posteriori termination criterion
$\left\|\mathbf{x}^* - \mathbf{x}^{(k)}\right\| \leq \frac{L^k}{1-L}\left\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\right\|$ (8.3.2.19)	$\left\|\mathbf{x}^* - \mathbf{x}^{(k)}\right\| \leq \frac{L}{1-L}\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right\|$ (8.3.2.20)

With the same arguments as in § 8.2.3.7 we see that overestimating L, that is, using a value for L that is larger than the true value, still gives reliable termination criteria.

However, whereas overestimating L in (8.3.2.20) will not lead to a severe deterioration of the bound, unless $L \approx 1$, using a pessimistic value for L in (8.3.2.19) will result in a bound way bigger than the true bound, if $k \gg 1$. Then the a priori termination criterion (8.3.2.19) will recommend termination many iterations after the accuracy requirements have already been met. This will thwart the efficiency of the method. ⌟

Review question(s) 8.3.2.21 (Fixed-point iterations)

(Q8.3.2.21.A) Let $\mathbf{x}^{(k)}$, $k \in \mathbb{N}_0$, be the iterates produced by a fixed-point iteration $\mathbf{x}^{(k+1)} = \Phi(\mathbf{x}^{(k)})$, $\Phi : \mathbb{R}^n \to \mathbb{R}^n$. Formulate a 2-point iteration

$$\mathbf{s}^{(k+1)} = \Psi(k, \mathbf{s}^{(k)}, \mathbf{s}^{(k-1)}) , \quad k \in \mathbb{N}_0 ,$$

that produces the sequence

$$\mathbf{s}^{(k)} := \frac{1}{k+1}\sum_{j=0}^{k} \mathbf{x}^{(j)} .$$

Is this a stationary 2-point iteration?

Definition § 8.2.1.4. Stationary m-point iterative method

A **stationary m-point iterative method** produces the sequences $\left\{\mathbf{x}^{(k)}\right\}_k \subset \mathbb{R}^n$ of iterates according to the formula

$$\mathbf{x}^{(k+1)} = \Phi_F(\mathbf{x}^{(k)}, \ldots, \mathbf{x}^{(k-m+1)}) ,$$

where $\Phi_F : \mathbb{R}^n \times \cdots \times \mathbb{R}^n \to \mathbb{R}^n$ is a given **iteration function**.

(Q8.3.2.21.B) Given $a > 0$ the following iteration functions spawn fixed-point iterations for the computation of $\sqrt{a}$:

$$\begin{aligned}
\varphi_1(x) &:= a + x - x^2 , \\
\varphi_2(x) &:= a/x , \\
\varphi_3(x) &:= 1 + x - \frac{1}{a}x^2 , \\
\varphi_4(x) &:= \frac{1}{2}(x + a/x) .
\end{aligned}$$

Predict the behavior and the type of convergence of the induced fixed-point iterations when started with an initial guess "sufficiently close" to $\sqrt{a}$.

△

8.4 Finding Zeros of Scalar Functions

Supplementary literature. [AG11, Ch. 3] is also devoted to this topic. The algorithm of "bisection" discussed in the next subsection, is treated in [DR08, Sect. 5.5.1] and [AG11, Sect. 3.2].

Now, we focus on scalar case $n = 1$: $F : I \subset \mathbb{R} \mapsto \mathbb{R}$ **continuous**, I interval

Sought: $x^* \in I$: $F(x^*) = 0$

8.4.1 Bisection

Video tutorial for Section 8.4.1 "Finding Zeros of Scalar Functions: Bisection": (7 minutes) Download link, tablet notes

Idea: use ordering of real numbers & intermediate value theorem [Str09, Sect. 4.6]

Input: $a, b \in I$ such that

$$F(a)F(b) < 0 . \quad \text{(different signs!)}$$

$$\exists\, x^* \in]\min\{a,b\}, \max\{a,b\}[: \quad F(x^*) = 0 ,$$

as we conclude from the intermediate value theorem.

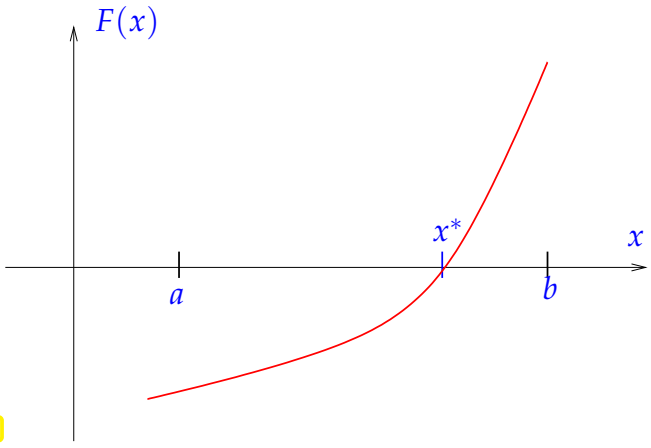


Fig. 299

► Find a sequence of intervals with geometrically decreasing lengths, in each of which F will change sign.

Such a sequence can easily be found by testing the sign of F at the midpoint of the current interval, see Code 8.4.1.2.

§8.4.1.1 (Bisection method) The following C++ code implements the bisection method for finding the zeros of a function passed through the function handle `F` in the interval $[a,b]$ with absolute tolerance `tol`.

C++ code 8.4.1.2: Bisection method for solving $F(x) = 0$ on $[a,b]$ ➽ GITLAB

```
2  // Searching zero of F in [a,b] by bisection
3  template <typename Func, typename Scalar>
4  Scalar bisect(Func&& F, Scalar a, Scalar b, Scalar tol)
5  {
6    if (a > b) {
7      std::swap(a,b); // sort interval bounds
8    }
9    if (F(a)*F(b) > 0) {
10     throw std::logic_error("f(a) and f(b) have same sign");
11   }
12   static_assert(std::is_floating_point<Scalar>::value,
13                 "Scalar must be a floating point type");
14   const int v=F(a) < 0 ? 1 : -1;
15   Scalar x = (a+b)/2; // determine midpoint
16   // termination, relies on machine arithmetic if tol = 0
17   while (b-a > tol) {
18     assert(a<=x && x<=b); // assert invariant
19     // sgn(f(x)) = sgn(f(b)), then use x as next right boundary
20     if (v*F(x) > 0) {
21       b=x;
22     }
23     // sgn(f(x)) = sgn(f(a)), then use x as next left boundary
24     else {
25       a=x;
26     }
27     x = (a+b)/2; // determine next midpoint
28   }
29   return x;
30 }
```

Line 18: the test `((a<x)&& (x<b))` offers a safeguard against an infinite loop in case `tol` $<$ resolution of $\mathbb{M}$ at zero x^* (*cf.* "$\mathbb{M}$-based termination criterion").

This is also an example for an algorithm that (in the case of `tol=0`) uses the properties of machine arithmetic to define an a posteriori termination criterion, see Section 8.2.3. The iteration will terminate, when, e.g., $a \widetilde{+} \frac{1}{2}(b-a) = a$ ($\widetilde{+}$ is the floating point realization of addition), which, by the Ass. 1.5.3.11 can only happen, when

$$\left|\tfrac{1}{2}(b-a)\right| \leq \mathrm{EPS} \cdot |a| \,.$$

Since the exact zero is located between a and b, this condition implies a relative error $\leq$ EPS of the computed zero.

Advantages:
- "foolproof", robust: will always terminate with a zero of requested accuracy,
- requires only point evaluations of F,
- works with *any* continuous function F, no derivatives needed.

Drawbacks: Merely "linear-type" $(*)$ convergence: $|x^{(k)} - x^*| \leq 2^{-k}|b-a|$

$\log_2\left(\frac{|b-a|}{\texttt{tol}}\right)$ steps necessary

$(*)$: the convergence of a bisection algorithm is not linear in the sense of Def. 8.2.2.1, because the condition $\left\|x^{(k+1)} - x^*\right\| \leq L\left\|x^{(k)} - x^*\right\|$ might be violated at any step of the iteration.

Remark 8.4.1.3 (Generalized bisection methods) It is straightforward to combine the bisection idea with more elaborate "model function methods" as they will be discussed in the next section: Instead of stubbornly choosing the midpoint of the probing interval $[a,b]$ ($\rightarrow$ Code 8.4.1.2) as next iterate, one may use a refined guess for the location of a zero of F in $[a,b]$.

A method of this type is used by MATLAB's `fzero` function for root finding in 1D [QSS00, Sect. 6.2.3]. ⌟

Review question(s) 8.4.1.4 (Finding Zeros of Scalar Functions: Bisection)

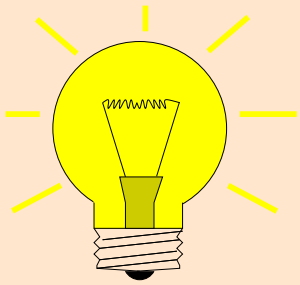
(Q8.4.1.4.A) We use the bisection method to find a zero of $f : [0.5, 2] \to \mathbb{R}$ with $f(1) < 0$ and $f(2) > 0$. Find an a priori bound for the number of steps needed to determine a zero with a guaranteed relative error of 10^{-6}.

(Q8.4.1.4.B) What prevents us from using bisection to find zeros of a function $f : D \subset \mathbb{C} \to \mathbb{C}$?

△

8.4.2 Model Function Methods

$\hat{=}$ class of iterative methods for finding zeroes of F: iterate in step $k+1$ is computed according to the following idea:



Idea: Given recent iterates (approximate zeroes) $x^{(k)}, x^{(k-1)}, \ldots, x^{(k-m+1)}$, $m \in \mathbb{N}$

❶ replace F with a k-dependent model function $\widetilde{F}_k$

(based on function values $F(x^{(k)}), F(x^{(k-1)}), \ldots, F(x^{(k-m+1)})$ and, possibly, derivative values $F'(x^{(k)}), F'(x^{(k-1)}), \ldots, F'(x^{(k-m+1)})$)

❷ $x^{(k+1)} :=$ zero of $\widetilde{F}_k$: $\widetilde{F}_k(x^{(k+1)}) = 0$

(has to be readily available $\leftrightarrow$ analytic formula)

Distinguish (see § 8.2.1.1 and (8.2.1.5)):

one-point methods : $x^{(k+1)} = \Phi_F(x^{(k)})$, $k \in \mathbb{N}$ (e.g., fixed point iteration $\rightarrow$ Section 8.3)

multi-point methods : $x^{(k+1)} = \Phi_F(x^{(k)}, x^{(k-1)}, \ldots, x^{(k-m)})$, $k \in \mathbb{N}$, $m = 2, 3, \ldots$.

8.4.2.1 Newton Method in the Scalar Case

Video tutorial for Section 8.4.2.1 "Newton Method in the Scalar Case": (20 minutes) Download link, tablet notes

Again we consider the problem of finding zeros of the function $F : I \subset \mathbb{R} \to \mathbb{R}$ defined on an interval I: we seek $x^* \in I$ such that $F(x^*) = 0$. Now we impose stricter smoothness requirements and we assume that $F : I \subset \mathbb{R} \mapsto \mathbb{R}$ is *continuously differentiable*, which means that both F and its derivative F' have to be continuous on I.

Now model function := tangent at F in $x^{(k)}$:

$$\widetilde{F}_k(x) := F(x^{(k)}) + F'(x^{(k)})(x - x^{(k)})$$

take $x^{(k+1)}$:= zero of tangent

We obtain the Newton iteration (N.I.)

$$x^{(k+1)} := x^{(k)} - \frac{F(x^{(k)})}{F'(x^{(k)})}\,, \qquad (8.4.2.1)$$

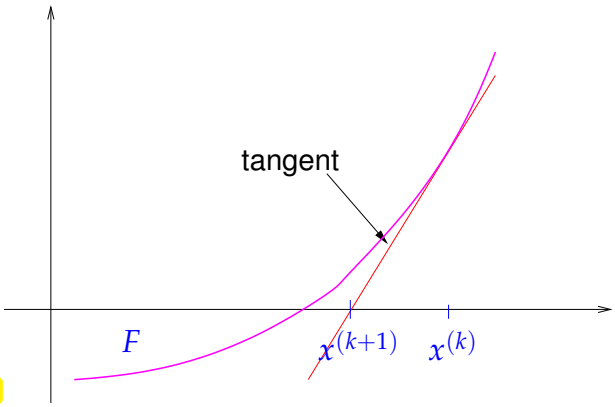


Fig. 300

that requires $F'(x^{(k)}) \neq 0$.

The following C++ code snippet implements a generic Newton method for zero finding.

- The types **FuncType** and **DervType** must be functor types and provide an evaluation operator `Scalar operator (Scalar)const`.
- The arguments `F` and `DF` must provide functors for F and F'.

C++ code 8.4.2.2: Newton method in the scalar case $n = 1$

```
template <typename FuncType,typename DervType,typename Scalar>
Scalar newton1D(FuncType &&F, DervType &&DF,
          const Scalar &x0, double rtol, double atol)
{
  Scalar s, x = x0;
  do {
    s = F(x)/DF(x);  // compute Newton correction
    x -= s;          // compute next iterate
  }
  // correction based termination (relative and absolute)
  while ((std::abs(s) > rtol*std::abs(x)) && (std::abs(s) > atol));
  return (x);
}
```

This code implements a correction-based termination criterion as introduced in § 8.2.3.4, see also § 8.2.3.2 for a discussion of absolute and relative tolerances.

EXAMPLE 8.4.2.3 (Square root iteration as a Newton iteration) In Ex. 8.2.2.13 we learned about the *quadratically convergent* fixed point iteration (8.2.2.14) for the approximate computation of the square root of a positive number. It can be derived as a Newton iteration (8.4.2.1)!

For $F(x) = x^2 - a$, $a > 0$, we find $F'(x) = 2x$, and, thus, the Newton iteration for finding zeros of F reads:

$$x^{(k+1)} = x^{(k)} - \frac{(x^{(k)})^2 - a}{2x^{(k)}} = \frac{1}{2}\left(x^{(k)} + \frac{a}{x^{(k)}}\right)\,,$$

which is exactly (8.2.2.14). Thus, for this F Newton's method converges globally with order $p = 2$. ⌟

EXAMPLE 8.4.2.4 (Newton method in 1D (→ Exp. 8.3.1.3)) Newton iterations for two different scalar non-linear equations $F(x) = 0$ with the same solution sets:

$$F(x) = xe^x - 1 \Rightarrow \quad F'(x) = e^x(1+x) \quad \Rightarrow \quad x^{(k+1)} = x^{(k)} - \frac{x^{(k)}e^{x^{(k)}} - 1}{e^{x^{(k)}}(1+x^{(k)})} = \frac{(x^{(k)})^2 + e^{-x^{(k)}}}{1+x^{(k)}}$$

$$F(x) = x - e^{-x} \quad \Rightarrow \quad F'(x) = 1 + e^{-x} \quad \Rightarrow \quad x^{(k+1)} = x^{(k)} - \frac{x^{(k)} - e^{-x^{(k)}}}{1 + e^{-x^{(k)}}} = \frac{1 + x^{(k)}}{1 + e^{x^{(k)}}} \,.$$

Exp. 8.3.1.3 confirms quadratic convergence in both cases! (→ Def. 8.2.2.10)

Note that for the computation of its zeros, the function F in this example can be recast in different forms!
⌟

§8.4.2.5 (Convergence of Newton's method in 1D) In fact, based on Lemma 8.3.2.15,

Lemma 8.3.2.15. Higher order local convergence of fixed point iterations

If $\Phi : U \subset \mathbb{R} \mapsto \mathbb{R}$ is $m+1$ times continuously differentiable, $\Phi(x^*) = x^*$ for some x^* in the interior of U, and $\Phi^{(l)}(x^*) = 0$ for $l = 1, \ldots, m$, $m \geq 1$, then the fixed point iteration (8.3.1.2) *converges locally to x^* with order $\geq m+1$ (→ Def. 8.2.2.10).*

it is straightforward to show local quadratic convergence of Newton's method to a zero x^* of F, provided that $F'(x^*) \neq 0$. We easily see that the Newton iteration

$$x^{(k+1)} := x^{(k)} - \frac{F(x^{(k)})}{F'(x^{(k)})} \,, \tag{8.4.2.1}$$

is a fixed fixed point iteration (→ Section 8.3) with iteration function

$$\Phi(x) := x - \frac{F(x)}{F'(x)} \quad \blacktriangleright \quad \text{(8.4.2.1)} \quad \Leftrightarrow \quad x^{(k+1)} = \Phi(x^{(k)}) \,. \tag{8.4.2.6}$$

Invoking the quotient rule for differentiation we get

$$\Phi'(x) = \frac{F(x)F''(x)}{(F'(x))^2} \quad \Rightarrow \quad \Phi'(x^*) = 0 \,, \quad \text{if } F(x^*) = 0,\ F'(x^*) \neq 0 \,. \tag{8.4.2.7}$$

Thus from Lemma 8.3.2.15 we conclude the following result:

Convergence of Newton's method in 1D

Newton's method locally converges quadratically (→ Def. 8.2.2.10) to a zero x^* of F, if $F'(x^*) \neq 0$ and F is three times continuously differentiable in a neighborhood of x^*.

⌟

EXAMPLE 8.4.2.9 (Newton iteration by implicit differentiation of F) This is a circuit model for an electric transmission line with voltage $U > 0$ imposed on one end, and grounded at the other:

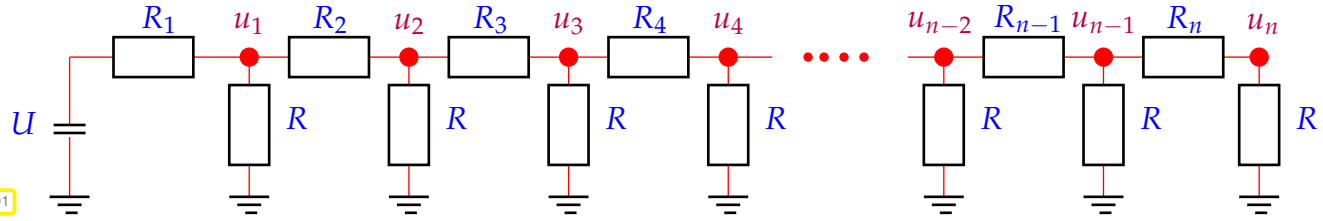


Fig. 301

How do we have to choose the leak resistance $R > 0$ in the linear circuit displayed in Fig. 301 in order to achieve a prescribed potential $u^* \in]0, U[$ at one of the nodes?

The circuit displayed in Fig. 301 is composed of linear resistors only. Thus we can use the nodal analysis of the circuit introduced in Ex. 2.1.0.3 in order to derive a linear system of equations for the nodal potentials u_j

in the nodes represented by • in Fig. 301. Kirchhoff's current law (2.1.0.4) plus the constitutive relationship $I = U/R$ for a resistor with resistance R give

$$\begin{aligned}
&\text{Node } 1: && \frac{1}{R_1}(u_1 - U) + \frac{1}{R}u_1 + \frac{1}{R_2}(u_1 - u_2) = 0\,, \\
&\text{Node } j: && \frac{1}{R_{j-1}}(u_j - u_{j-1}) + \frac{1}{R}u_j + \frac{1}{R_j}(u_j - u_{j+1}) = 0\,, \quad j = 2,\ldots,n-1\,, \\
&\text{Node } n: && \frac{1}{R_n}(u_n - u_{n-1}) + \frac{1}{R}u_n = 0\,.
\end{aligned} \tag{8.4.2.10}$$

These n equations are equivalent to a linear system of equations for the vector $\mathbf{u} = \left[u_j\right]_{j=1}^{n} \in \mathbb{R}^n$, which reads in compact notation

$$\left(\mathbf{A} + \frac{1}{R}\cdot\mathbf{I}\right)\mathbf{u} = \mathbf{b}\,, \tag{8.4.2.11}$$

$$\mathbf{A} = \begin{bmatrix}
\frac{1}{R_1}+\frac{1}{R_2} & -\frac{1}{R_2} & & & & & \\
-\frac{1}{R_2} & \frac{1}{R_2}+\frac{1}{R_3} & -\frac{1}{R_3} & & & & \\
 & -\frac{1}{R_3} & \frac{1}{R_3}+\frac{1}{R_4} & -\frac{1}{R_4} & & & \\
 & & \ddots & \ddots & \ddots & & \\
 & & & -\frac{1}{R_{n-2}} & \frac{1}{R_{n-2}}+\frac{1}{R_{n-1}} & -\frac{1}{R_{n-1}} & \\
 & & & & -\frac{1}{R_{n-1}} & \frac{1}{R_n}+\frac{1}{R_{n-1}} & -\frac{1}{R_n} \\
 & & & & & -\frac{1}{R_n} & \frac{1}{R_n}
\end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} \frac{U}{R_1} \\ 0 \\ \vdots \\ \\ \vdots \\ 0 \end{bmatrix}.$$

Thus the current problem can be formulated as: find $x \in \mathbb{R}$, $x := R^{-1} > 0$, such that

$$F(x) = 0 \quad \text{with} \quad F : \begin{cases} \mathbb{R} & \to & \mathbb{R} \\ x & \mapsto & \mathbf{w}^\top(\mathbf{A} + x\mathbf{I})^{-1}\mathbf{b} - u^* \end{cases}\,, \tag{8.4.2.12}$$

where $\mathbf{A} \in \mathbb{R}^{n,n}$ is a symmetric, tridiagonal, diagonally dominant matrix, $\mathbf{w} \in \mathbb{R}^n$ is a Cartesian coordinate vector singling out the node of interest, and $\mathbf{b}$ takes into account the exciting voltage U.

To begin with, we reason that for $R \to 0$ and $R \to \infty$ we expect $u_j \to 0$ and $u_j \to U$, respectively. Therefore, from the intermediate value theorem we conclude the existence of a solution.

In order to apply Newton's method to (8.4.2.12), we have to determine the derivative $F'(x)$ and so by implicit differentiation [Str09, Sect. 7.8], first rewriting ($\mathbf{u}(x) \hat{=}$ vector of nodal potentials as a function of $x = R^{-1}$)

$$F(x) = \mathbf{w}^\top\mathbf{u}(x) - u^* \quad , \quad (\mathbf{A} + x\mathbf{I})\mathbf{u}(x) = \mathbf{b}\,.$$

Then we differentiate the linear system of equations defining $\mathbf{u}(x)$ on both sides with respect to x using the product rule (8.5.1.17)

$$(\mathbf{A} + x\mathbf{I})\mathbf{u}(x) = \mathbf{b} \quad \overset{\frac{d}{dx}}{\Longrightarrow} \quad (\mathbf{A} + x\mathbf{I})\mathbf{u}'(x) + \mathbf{u}(x) = \mathbf{0}\,.$$

$$\blacktriangleright \quad \mathbf{u}'(x) = -(\mathbf{A} + x\mathbf{I})^{-1}\mathbf{u}(x)\,. \tag{8.4.2.13}$$

$$\blacktriangleright \quad F'(x) = \mathbf{w}^\top\mathbf{u}'(x) = -\mathbf{w}^\top(\mathbf{A} + x\mathbf{I})^{-1}\mathbf{u}(x)\,. \tag{8.4.2.14}$$

Thus, the Newton iteration for (8.4.2.12) reads:

$$x^{(k+1)} = x^{(k)} + F'(x^{(k)})^{-1}F(x^{(k)}) = \frac{\mathbf{w}^\top\mathbf{u}(x^{(k)}) - 1}{\mathbf{w}^\top\mathbf{z}(\mathbf{x}^{(k)})}\,, \tag{8.4.2.15}$$

$$\text{with} \quad (\mathbf{A} + x^{(k)}\mathbf{I})\mathbf{u}(x^{(k)}) = \mathbf{b}\,,$$
$$(\mathbf{A} + x^{(k)}\mathbf{I})\mathbf{z}(\mathbf{x}^{(k)}) = \mathbf{u}(x^{(k)})\,.$$

In each step of the iteration we have to solve two linear systems of equations, which can be done with asymptotic effort $O(n)$ in this case, because $\mathbf{A} + x^{(k)}\mathbf{I}$ is tridiagonal.

Note that in a practical application one must demand $x > 0$, in addition, because the solution must provide a meaningful conductance (= inverse resistance.)

Also note that bisection (→ 8.4.1) is a viable alternative to using Newton's method in this case. ⌟

Supplementary literature. Newton's method in 1D is discussed in [Han02, Sect. 18.1], [DR08, Sect. 5.5.2], [AG11, Sect. 3.4].

Review question(s) 8.4.2.16 (Newton's method in one dimension)

(Q8.4.2.16.A) State Newton's iteration for finding the zero of $f(x) := \alpha x + \beta$, $\alpha, \beta \in \mathbb{R}$, $x \in \mathbb{R}$. What can you say about its convergence?

(Q8.4.2.16.B) The Lambert W-function $W : I \subset \mathbb{R} \to \mathbb{R}$ is implicitly defined through

$$W(x)\exp(W(x)) = x \quad , \quad W(x) > 0 \quad \text{for} \quad x > 0\,.$$

1. What is the maximal domain of definition I?
2. State the Newton iteration for computing $W(x)$, $x \in I$.
3. Sketch a C++ function that can compute $W(x)$ with a relative error equal to the machine precision EPS.

(Q8.4.2.16.C) Planck's radiation law describes the spectral emissive power of a black body by the formula

$$B(\nu, T) = \frac{2h\nu^3}{c^2}\frac{1}{\exp\left(\frac{h\nu}{k_B T}\right) - 1}\,, \qquad \begin{array}{ll} k_B & \hat{=} \text{ Boltzmann constant,} \\ h & \hat{=} \text{ Planck constant,} \\ c & \hat{=} \text{ speed of light ,} \end{array}$$

as function of the frequency $\nu > 0$ and of the temperature $T > 0$.

Outline an algorithm based on Newton's method that can determine the frequency $\nu_{\max}$ of the emission maximum as a function of T.

(Q8.4.2.16.D) What will happen, if one applies Newton's method to $F(x) := x^2 + 1$ with $x^{(0)} = 1$ or $x^{(0)} = 2$?

△

8.4.2.2 Special One-Point Methods

Idea underlying other one-point methods: non-linear local approximation

Useful, if *a priori knowledge* about the structure of F (e.g. about F being a rational function, see below) is available. This is often the case, because many problems of 1D zero finding are posed for functions given in analytic form with a few parameters.

Prerequisite: Smoothness of F: $F \in C^m(I)$ for some $m > 1$

EXAMPLE 8.4.2.17 (Halley's iteration → [Han02, Sect. 18.3]) This example demonstrates that non-polynomial model functions can offer excellent approximation of F. In this example the model function is chosen as a quotient of two linear function, that is, from the simplest class of true rational functions.

Of course, that this function provides a good model function is merely "a matter of luck", unless you have some more information about F. Such information might be available from the application context.

Given $x^{(k)} \in I$, next iterate := zero of model function: $h(x^{(k+1)}) = 0$, where

$$h(x) := \frac{a}{x+b} + c \quad \text{(rational function) such that } F^{(j)}(x^{(k)}) = h^{(j)}(x^{(k)}) \,, \quad j = 0,1,2 \,.$$

$$\frac{a}{x^{(k)}+b} + c = F(x^{(k)}) \,, \; -\frac{a}{(x^{(k)}+b)^2} = F'(x^{(k)}) \,, \; \frac{2a}{(x^{(k)}+b)^3} = F''(x^{(k)}) \,.$$

$$x^{(k+1)} = x^{(k)} - \frac{F(x^{(k)})}{F'(x^{(k)})} \cdot \frac{1}{1 - \frac{1}{2}\frac{F(x^{(k)})F''(x^{(k)})}{F'(x^{(k)})^2}} \,.$$

Halley's iteration for $F(x) = \frac{1}{(x+1)^2} + \frac{1}{(x+0.1)^2} - 1 \,, \quad x > 0 :$ and $x^{(0)} = 0$

k	$x^{(k)}$	$F(x^{(k)})$	$x^{(k)} - x^{(k-1)}$	$x^{(k)} - x^*$
1	0.19865959351191	10.90706835180178	-0.19865959351191	-0.84754290138257
2	0.69096314049024	0.94813655914799	-0.49230354697833	-0.35523935440424
3	1.02335017694603	0.03670912956750	-0.33238703645579	-0.02285231794846
4	1.04604398836483	0.00024757037430	-0.02269381141880	-0.00015850652965
5	1.04620248685303	0.00000001255745	-0.00015849848821	-0.00000000804145

Compare with Newton method (8.4.2.1) for the same problem:

k	$x^{(k)}$	$F(x^{(k)})$	$x^{(k)} - x^{(k-1)}$	$x^{(k)} - x^*$
1	0.04995004995005	44.38117504792020	-0.04995004995005	-0.99625244494443
2	0.12455117953073	19.62288236082625	-0.07460112958068	-0.92165131536375
3	0.23476467495811	8.57909346342925	-0.11021349542738	-0.81143781993637
4	0.39254785728080	3.63763326452917	-0.15778318232269	-0.65365463761368
5	0.60067545233191	1.42717892023773	-0.20812759505112	-0.44552704256257
6	0.82714994286833	0.46286007749125	-0.22647449053641	-0.21905255202615
7	0.99028203077844	0.09369191826377	-0.16313208791011	-0.05592046411604
8	1.04242438221432	0.00592723560279	-0.05214235143588	-0.00377811268016
9	1.04618505691071	0.00002723158211	-0.00376067469639	-0.00001743798377
10	1.04620249452271	0.00000000058056	-0.00001743761199	-0.00000000037178

Note that Halley's iteration is superior in this case, since F is a rational function.

! Newton method converges more slowly, but also needs less effort per step (→ Section 8.4.3) ⌟

§8.4.2.18 (Preconditioning of Newton's method) In the previous example Newton's method performed rather poorly. Often its convergence can be boosted by converting the non-linear equation to an equivalent one (that is, one with the same solutions) for another function g, which is "closer to a linear function":

Assume that (locally) $F \approx \widehat{F}$, where $\widehat{F}$ is (locally) invertible with an inverse $\widehat{F}^{-1}$ that can be evaluated with little effort.

$$\blacktriangleright \quad g(x) := \widehat{F}^{-1}(F(x)) \approx x \,.$$

Then apply Newton's method to $G(x) := g(x) - \widehat{F}^{-1}(0)$, using the formula for the derivative of the inverse of a function

$$\frac{d}{dy}(\widehat{F}^{-1})(y) = \frac{1}{\widehat{F}'(\widehat{F}^{-1}(y))} \quad \Rightarrow \quad g'(x) = \frac{1}{\widehat{F}'(g(x))} \cdot F'(x) \,.$$

This results in the Newton iteration

$$x^{(k+1)} = x^{(k)} - \frac{G(x^{(k)})\widehat{F}'(g(x^{(k)}))}{F'(x^{(k)})} \,, \tag{8.4.2.19}$$
$$g(x^{(k)}) = \widehat{F}^{-1}(F(x^{(k)})) \quad , \quad G(x^{(k)}) = g(x^{(k)}) - y_0 \,, \quad y_0 := \widehat{F}^{-1}(0) \,.$$

Since G is "almost linear" this Newton iteration can be expected to enjoy quadratic convergence for initial guesses $x^{(0)}$ from a large set. A good initial guess is $x^{(0)} := \widehat{F}^{-1}(0)$. ⌟

EXAMPLE 8.4.2.20 (Preconditioned Newton method) As in Ex. 8.4.2.17 we consider

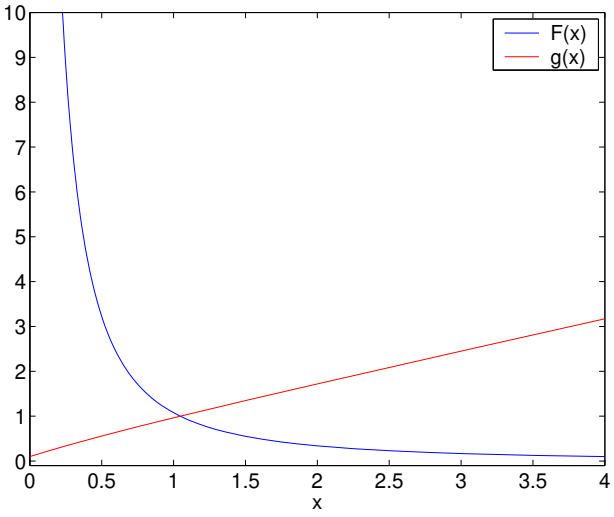
$$F(x) = \frac{1}{(x+1)^2} + \frac{1}{(x+0.1)^2} - 1 \,, \quad x > 0 \,,$$

and try to find its zeros.

Observation:

$$F(x) + 1 \approx 2x^{-2} \text{ for } x \gg 1$$

and so $g(x) := \dfrac{1}{\sqrt{F(x)+1}}$ is "almost" linear for $x \gg 1$.



Idea: instead of $F(x) \stackrel{!}{=} 0$ tackle $g(x) \stackrel{!}{=} 1$ with Newton's method (8.4.2.1).

$$x^{(k+1)} = x^{(k)} - \frac{g(x^{(k)}) - 1}{g'(x^{(k)})} = x^{(k)} + \left(\frac{1}{\sqrt{F(x^{(k)}) + 1}} - 1 \right) \frac{2(F(x^{(k)}) + 1)^{3/2}}{F'(x^{(k)})}$$
$$= x^{(k)} + \frac{2(F(x^{(k)}) + 1)(1 - \sqrt{F(x^{(k)}) + 1})}{F'(x^{(k)})} \,.$$

Convergence recorded for $x^{(0)} = 0$:

k	$x^{(k)}$	$F(x^{(k)})$	$x^{(k)} - x^{(k-1)}$	$x^{(k)} - x^*$
1	0.91312431341979	0.24747993091128	0.91312431341979	-0.13307818147469
2	1.04517022155323	0.00161402574513	0.13204590813344	-0.00103227334125
3	1.04620244004116	0.00000008565847	0.00103221848793	-0.00000005485332
4	1.04620249489448	0.00000000000000	0.00000005485332	-0.00000000000000

For zero finding there is wealth of iterative methods that offer higher order of convergence. One class is discussed next.

§8.4.2.21 (Modified Newton methods) Taking the cue from the iteration function of Newton's method (8.4.2.1), we extend it by introducing an extra function H:

$$\blacktriangleright \quad \text{new fixed point iteration :} \quad \Phi(x) = x - \frac{F(x)}{F'(x)} H(x) \quad \text{with "proper" } H : I \mapsto \mathbb{R} .$$

Still, every zero of F is a fixed point of this Φ,that is, the fixed point iteration is still consistent ($\rightarrow$ Def. 8.3.1.1).

Aim: find H such that the method is of p-th order. The main tool is Lemma 8.3.2.15, which tells us that we have to ensure $\Phi^{(\ell)}(x^*) = 0$, $1 \leq \ell \leq p-1$, guarantees local convergence of order p.

Assume: F smooth "enough" and $\exists\, x^* \in I$: $F(x^*) = 0$, $F'(x^*) \neq 0$. Then we can compute the derivatives of Φ appealing to the product rule and quotient rule for derivatives.

$$\Phi = x - uH \quad , \quad \Phi' = 1 - u'H - uH' \quad , \quad \Phi'' = -u''H - 2u'H - uH'' ,$$
$$\text{with } u = \frac{F}{F'} \quad \Rightarrow \quad u' = 1 - \frac{FF''}{(F')^2} \quad , \quad u'' = -\frac{F''}{F'} + 2\frac{F(F'')^2}{(F')^3} - \frac{FF'''}{(F')^2} .$$

$F(x^*) = 0 \quad \blacktriangleright \quad u(x^*) = 0, u'(x^*) = 1, u''(x^*) = -\frac{F''(x^*)}{F'(x^*)}$.

$$\blacktriangleright \quad \Phi'(x^*) = 1 - H(x^*) \quad , \quad \Phi''(x^*) = \frac{F''(x^*)}{F'(x^*)} H(x^*) - 2H'(x^*) . \tag{8.4.2.22}$$

Lemma 8.3.2.15 $\blacktriangleright$ **Necessary** conditions for local convergence of order p:

$$p = 2 \text{ (quadratical convergence):} \quad H(x^*) = 1 ,$$
$$p = 3 \text{ (cubic convergence):} \quad H(x^*) = 1 \quad \wedge \quad H'(x^*) = \frac{1}{2}\frac{F''(x^*)}{F'(x^*)} .$$

Trial expression: $H(x) = G(1 - u'(x))$ with "appropriate" G

$$\blacktriangleright \quad \text{fixed point iteration} \quad x^{(k+1)} = x^{(k)} - \frac{F(x^{(k)})}{F'(x^{(k)})} G\left(\frac{F(x^{(k)})F''(x^{(k)})}{(F'(x^{(k)}))^2} \right) . \tag{8.4.2.23}$$

Lemma 8.4.2.24. Cubic convergence of modified Newton methods

If $F \in C^2(I)$, $F(x^*) = 0$, $F'(x^*) \neq 0$, $G \in C^2(U)$ in a neighbourhood U of 0, $G(0) = 1$, $G'(0) = \frac{1}{2}$, then the fixed point iteration (8.4.2.23) *converge locally cubically to x^*.*

Proof. We apply Lemma 8.3.2.15, which tells us that both derivatives from (8.4.2.22) have to vanish. Using the definition of H we find.

$$H(x^*) = G(0) \quad , \quad H'(x^*) = -G'(0)u''(x^*) = G'(0)\frac{F''(x^*)}{F'(x^*)} .$$

Plugging these expressions into (8.4.2.22) finishes the proof. □

EXPERIMENT 8.4.2.25 (Application of modified Newton methods)

- $G(t) = \dfrac{1}{1-\frac{1}{2}t}$ ➡ Halley's iteration (→ Ex. 8.4.2.17)
- $G(t) = \dfrac{2}{1+\sqrt{1-2t}}$ ➡ Euler's iteration
- $G(t) = 1+\frac{1}{2}t$ ➡ quadratic inverse interpolation

Numerical experiment:

$$F(x) = xe^x - 1\,,$$
$$x^{(0)} = 5$$

k	$e^{(k)} := x^{(k)} - x^*$		
	Halley	Euler	Quad. Inv.
1	2.81548211105635	3.57571385244736	2.03843730027891
2	1.37597082614957	2.76924150041340	1.02137913293045
3	0.34002908011728	1.95675490333756	0.28835890388161
4	0.00951600547085	1.25252187565405	0.01497518178983
5	0.00000024995484	0.51609312477451	0.00000315361454
6		0.14709716035310	
7		0.00109463314926	
8		0.00000000107549	

Review question(s) 8.4.2.26 (Special 1-point iterative methods for root finding)

(Q8.4.2.26.A) The generic iteration for a modified Newton method for solving the scalar zero-finding problem $F(x) = 0$ is

$$x^{(k+1)} = x^{(k)} - \frac{F(x^{(k)})}{F'(x^{(k)})}\, G\left(\frac{F(x^{(k)})F''(x^{(k)})}{(F'(x^{(k)}))^2}\right)\,, \tag{8.4.2.23}$$

with a C^2-function $G : U \subset \mathbb{R} \to \mathbb{R}$ defined in a neighborhood of 0.

Give the explicit formula for the iteration (8.4.2.23) for $F(x) = xe^x - 1$ and $G(t) = (1-\frac{1}{2}t)^{-1}$ (Halley's iteration).

△

8.4.2.3 Multi-Point Methods

Video tutorial for Section 8.4.2.3 "Multi-Point Methods": (12 minutes) Download link, tablet notes

Supplementary literature. The secant method is presented in [Han02, Sect. 18.2], [DR08, Sect. 5.5.3], [AG11, Sect. 3.4].

Construction of multi-point iterations in 1D

Idea: Replace F with interpolating polynomial

producing interpolatory model function methods

§8.4.2.28 (The secant method)

The secant method is the simplest representative of model function multi-point methods:

The figure illustrates the geometric idea underlying this 2-point method for zero finding:

$x^{(k+1)}$ is the zero of the secant (red line ▷) connecting the points $(x^{(k-1)}, F(x^{(k-1)}))$ and $(x^k, F(x^k))$ on the graph of F.

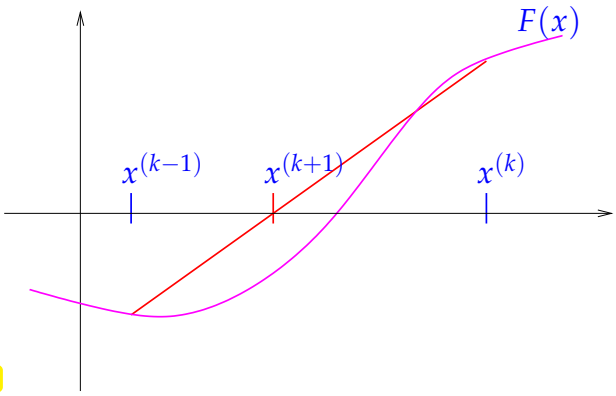


Fig. 302

The secant line is the graph of the function

$$s(x) = F(x^{(k)}) + \frac{F(x^{(k)}) - F(x^{(k-1)})}{x^{(k)} - x^{(k-1)}}(x - x^{(k)}) , \tag{8.4.2.29}$$

$$x^{(k+1)} = x^{(k)} - \frac{F(x^{(k)})(x^{(k)} - x^{(k-1)})}{F(x^{(k)}) - F(x^{(k-1)})} . \tag{8.4.2.30}$$

The following C++ code snippet demonstrates the implementation of the abstract secant method for finding the zeros of a function passed through the functor `F`.

C++ code 8.4.2.31: Secant method for 1D non-linear equation ↪ GITLAB

```
// Secand method for solving F(x) = 0 for F : D ⊂ ℝ → ℝ,
// initial guesses x_0, x_1,
// tolerances atol (absolute), rtol (relative)
template <typename Func>
double secant(double x0, double x1, Func &&F, double rtol, double atol,
              unsigned int maxIt) {
  double fo = F(x0);
  for (unsigned int i = 0; i < maxIt; ++i) {
    const double fn = F(x1);
    const double s = fn * (x1 - x0) / (fn - fo); // secant correction
    x0 = x1;
    x1 = x1 - s;
    // correction based termination (relative and absolute)
    if (abs(s) < max(atol, rtol * min(abs(x0), abs(x1)))) {
      return x1;
    }
    fo = fn;
  }
  return x1;
}
```

This code demonstrates several important features of the secant method:

- Only *one* function evaluation per step
- no derivatives required!
- 2-point method: two initial guesses needed

Remember: $F(x)$ may only be available as output of a (complicated) procedure. In this case it is difficult to find a procedure that evaluates $F'(x)$. Thus the significance of methods that do not involve evaluations of derivatives. ⌟

EXPERIMENT 8.4.2.32 (Convergence of secant method) We empirically examine the convergence of the secant method from Code 8.4.2.31 for a model problem.

$F(x) = xe^x - 1$, using secant method with initial guesses $x^{(0)} = 0$, $x^{(1)} = 5$.

k	$x^{(k)}$	$F(x^{(k)})$	$e^{(k)} := x^{(k)} - x^*$	$\frac{\log\lvert e^{(k+1)}\rvert - \log\lvert e^{(k)}\rvert}{\log\lvert e^{(k)}\rvert - \log\lvert e^{(k-1)}\rvert}$
2	0.00673794699909	-0.99321649977589	-0.56040534341070	
3	0.01342122983571	-0.98639742654892	-0.55372206057408	24.43308649757745
4	0.98017620833821	1.61209684919288	0.41303291792843	2.70802321457994
5	0.38040476787948	-0.44351476841567	-0.18673852253030	1.48753625853887
6	0.50981028847430	-0.15117846201565	-0.05733300193548	1.51452723840131
7	0.57673091089295	0.02670169957932	0.00958762048317	1.70075240166256
8	0.56668541543431	-0.00126473620459	-0.00045787497547	1.59458505614449
9	0.56713970649585	-0.00000990312376	-0.00000358391394	1.62641838319117
10	0.56714329175406	0.00000000371452	0.00000000134427	
11	0.56714329040978	-0.00000000000001	-0.00000000000000	

Rem. 8.2.2.12: the rightmost column of the table provides an estimate for the order of convergence → Def. 8.2.2.10. For further explanations see Rem. 8.2.2.12.

A startling observation: the method seems to have a *fractional* (!) order of convergence, see Def. 8.2.2.10. ⌟

Remark 8.4.2.33 (Fractional order of convergence of secant method) Indeed, a fractional order of convergence can be proved for the secant method, see [Han02, Sect. 18.2]. Here we give an asymptotic argument that holds, if the iterates are already very close to the zero x^* of F.

We can write the secant method in the form (8.2.1.5)

$$x^{(k+1)} = \Phi(x^{(k)}, x^{(k-1)}) \quad \text{with} \quad \Phi(x,y) = \Phi(x,y) := x - \frac{F(x)(x-y)}{F(x)-F(y)} \,. \tag{8.4.2.34}$$

Using Φ we find a recursion for the iteration error $e^{(k)} := x^{(k)} - x^*$:

$$e^{(k+1)} = \Phi(x^* + e^{(k)}, x^* + e^{(k-1)}) - x^* \,. \tag{8.4.2.35}$$

Thanks to the asymptotic perspective we may assume that $|e^{(k)}|, |e^{(k-1)}| \ll 1$ so that we can rely on two-dimensional Taylor expansion around $(x^*, x^{(*)})$, *cf.* [Str09, Satz 7.5.2]:

$$\begin{aligned} \Phi(x^*+h, x^*+k) = \Phi(x^*,x^*) + \frac{\partial\Phi}{\partial x}(x^*,x^*)h + \frac{\partial\Phi}{\partial y}(x^*,x^*)k + \\ \tfrac{1}{2}\frac{\partial^2\Phi}{\partial^2 x}(x^*,x^*)h^2 + \frac{\partial^2\Phi}{\partial x\partial y}(x^*,x^*)hk + \tfrac{1}{2}\frac{\partial^2\Phi}{\partial^2 y}(x^*,x^*)k^2 + R(x^*,h,k) \,, \\ \text{with} \quad |R| \le C(h^3 + h^2k + hk^2 + k^3) \,. \end{aligned} \tag{8.4.2.36}$$

Computations invoking the quotient rule and product rule and using $F(x^*) = 0$ show

$$\Phi(x^*,x^*) = x^* \,, \quad \frac{\partial\Phi}{\partial x}(x^*,x^*) = \frac{\partial\Phi}{\partial y}(x^*,x^*) = \tfrac{1}{2}\frac{\partial^2\Phi}{\partial^2 x}(x^*,x^*) = \tfrac{1}{2}\frac{\partial^2\Phi}{\partial^2 y}(x^*,x^*) = 0 \,.$$

We may also use MAPLE to find the Taylor expansion (assuming F sufficiently smooth):

```
> Phi := (x,y) -> x-F(x)*(x-y)/(F(x)-F(y));
> F(s) := 0;
```

```
> e2 = normal(mtaylor(Phi(s+e1,s+e0)-s,[e0,e1],4));
```

➢ truncated error propagation formula (products of three or more error terms ignored)

$$e^{(k+1)} \doteq \frac{1}{2}\frac{F''(x^*)}{F'(x^*)} e^{(k)} e^{(k-1)} = C e^{(k)} e^{(k-1)} \,. \tag{8.4.2.37}$$

How can we deduce the order of converge from this recursion formula? We try $e^{(k)} = K(e^{(k-1)})^p$ inspired by the estimate in Def. 8.2.2.10:

$$\Rightarrow \quad e^{(k+1)} = K^{p+1}(e^{(k-1)})^{p^2}$$
$$\Rightarrow \quad (e^{(k-1)})^{p^2-p-1} = K^{-p}C \quad \Rightarrow \quad p^2 - p - 1 = 0 \quad \Rightarrow \quad p = \tfrac{1}{2}(1 \pm \sqrt{5}) \,.$$

The second implication is clear after realizing that that the equation has to be satisfied for all k and that the right-hand side does not depend on k. As $e^{(k)} \to 0$ for $k \to \infty$ we get the order of convergence $p = \frac{1}{2}(1+\sqrt{5}) \approx 1.62$ (see Exp. 8.4.2.32 !) ⌟

EXAMPLE 8.4.2.38 (local convergence of the secant method)

Model problem: find zero of

$$F(x) = \arctan(x)$$

· $\hat{=}$ secant method converges for a pair $(x^{(0)}, x^{(1)}) \in \mathbb{R}^2_+$ of initial guesses. ▷

We observe that the secant method will converge only for initial guesses sufficiently close to 0 = local convergence → Def. 8.2.1.10

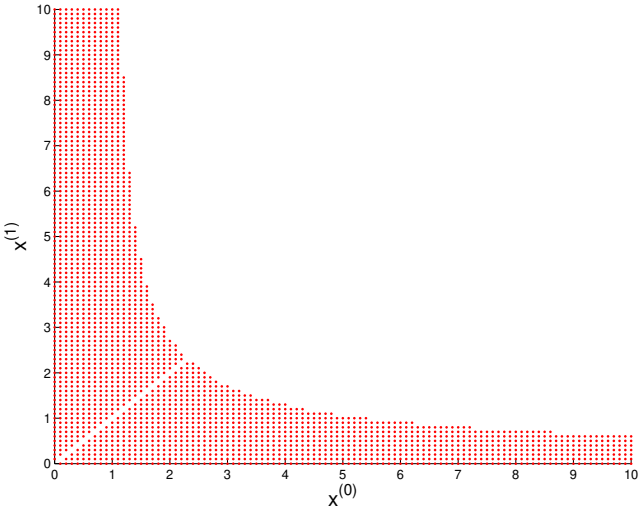
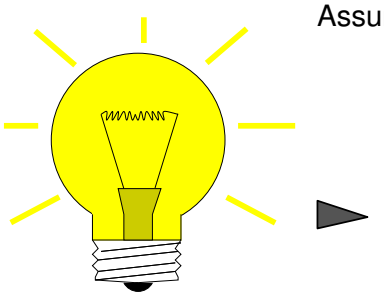


Fig. 303

⌟

§8.4.2.39 (Inverse interpolation) Another class of multi-point methods: *inverse interpolation*



Assume: $F : I \subset \mathbb{R} \mapsto \mathbb{R}$ one-to-one (monotone)

$$F(x^*) = 0 \quad \Rightarrow \quad F^{-1}(0) = x^* \,.$$

- Interpolate F^{-1} by polynomial p of degree $m-1$ determined by

$$p(F(x^{(k-j)})) = x^{(k-j)} \,, \quad j = 0, \ldots, m-1 \,.$$

- New approximate zero $x^{(k+1)} := p(0)$

The graph of F^{-1} can be obtained by reflecting the graph of F at the angular bisector. ▷

$$F(x^*) = 0 \quad \Leftrightarrow \quad F^{-1}(0) = x^*$$

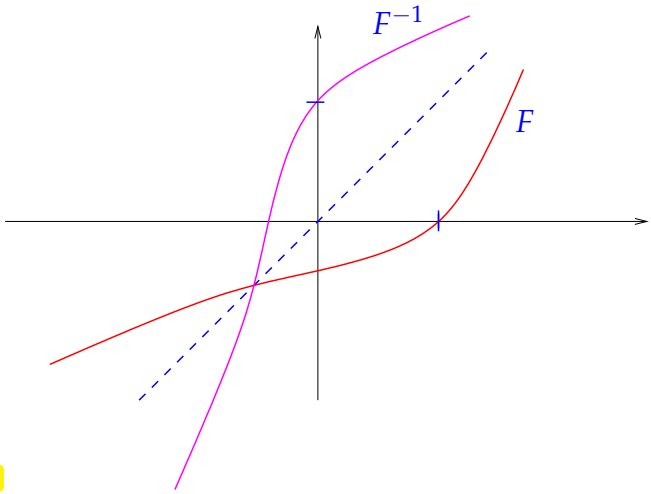


Fig. 304

Case $m = 2$ (2-point method)

➢ secant method

The interpolation polynomial is a line. In this case we do not get a new method, because the inverse function of a linear function (polynomial of degree 1) is again a polynomial of degree 1.

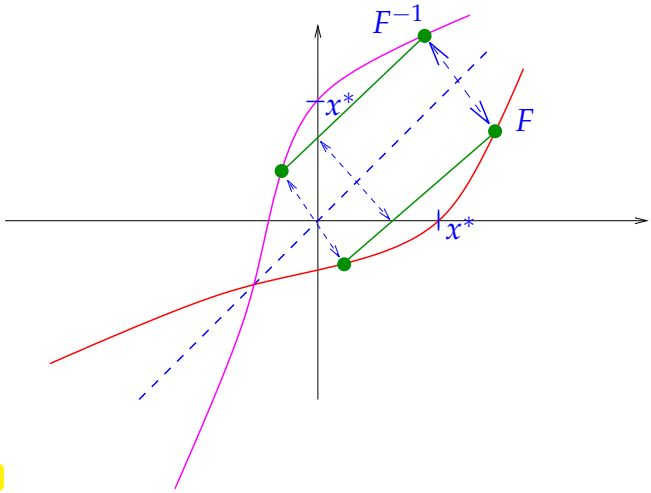


Fig. 305

Case $m = 3$: quadratic inverse interpolation, a 3-point method, see [Mol04, Sect. 4.5]

We interpolate the points $(F(x^{(k)}), x^{(k)}), (F(x^{(k-1)}), x^{(k-1)}), (F(x^{(k-2)}), x^{(k-2)})$ with a parabola (polynomial of degree 2). Note the importance of monotonicity of F, which ensures that $F(x^{(k)}), F(x^{(k-1)}), F(x^{(k-2)})$ are mutually different.

MAPLE code:
```
p := x-> a*x^2+b*x+c;
solve({p(f0)=x0,p(f1)=x1,p(f2)=x2},{a,b,c});
assign(%); p(0);
```

$$\blacktriangleright \quad x^{(k+1)} = \frac{F_0^2(F_1x_2 - F_2x_1) + F_1^2(F_2x_0 - F_0x_2) + F_2^2(F_0x_1 - F_1x_0)}{F_0^2(F_1 - F_2) + F_1^2(F_2 - F_0) + F_2^2(F_0 - F_1)} .$$

$$\left(F_0 := F(x^{(k-2)}), F_1 := F(x^{(k-1)}), F_2 := F(x^{(k)}), x_0 := x^{(k-2)}, x_1 := x^{(k-1)}, x_2 := x^{(k)} \right)$$

⌟

EXPERIMENT 8.4.2.40 (Convergence of quadratic inverse interpolation) We test the method for the model problem/initial guesses $F(x) := xe^x - 1 = 0$, $x^{(0)} = 0$, $x^{(1)} = 2.5$, $x^{(2)} = 5$.

k	$x^{(k)}$	$F(x^{(k)})$	$e^{(k)} := x^{(k)} - x^*$	$\frac{\log\lvert e^{(k+1)}\rvert - \log\lvert e^{(k)}\rvert}{\log\lvert e^{(k)}\rvert - \log\lvert e^{(k-1)}\rvert}$
3	0.08520390058175	-0.90721814294134	-0.48193938982803	
4	0.16009252622586	-0.81211229637354	-0.40705076418392	3.33791154378839
5	0.79879381816390	0.77560534067946	0.23165052775411	2.28740488912208
6	0.63094636752843	0.18579323999999	0.06380307711864	1.82494667289715
7	0.56107750991028	-0.01667806436181	-0.00606578049951	1.87323264214217
8	0.56706941033107	-0.00020413476766	-0.00007388007872	1.79832936980454
9	0.56714331707092	0.00000007367067	0.00000002666114	1.84841261527097
10	0.56714329040980	0.00000000000003	0.00000000000001	

Also in this case the numerical experiment hints at a fractional rate of convergence $p \approx 1.8$, as in the case of the secant method, see Rem. 8.4.2.33. ⌟

Review question(s) 8.4.2.41 (Multi-point methods)

(Q8.4.2.41.A) The inverse quadratic interpolation method for solving $F(x) = 0$ reads

$$x^{(k+1)} = \frac{F_0^2(F_1x_2 - F_2x_1) + F_1^2(F_2x_0 - F_0x_2) + F_2^2(F_0x_1 - F_1x_0)}{F_0^2(F_1 - F_2) + F_1^2(F_2 - F_0) + F_2^2(F_0 - F_1)} .$$

$$\left(F_0 := F(x^{(k-2)}), F_1 := F(x^{(k-1)}), F_2 := F(x^{(k)}), x_0 := x^{(k-2)}, x_1 := x^{(k-1)}, x_2 := x^{(k)} \right)$$

Which evaluations and computations have to be carried out in a single step?

(Q8.4.2.41.B) We apply the secant method with initial guesses $x^{(0)}, x^{(1)} \in [a,b]$ to find the zero of a function $f : [a,b] \to \mathbb{R}$ with the following properties:

- f is increasing,
- f is convex,
- $f(a) < 0$ and $f(b) > 0$.

Give a "geometric proof" that the iteration will converge.

△

8.4.3 Asymptotic Efficiency of Iterative Methods for Zero Finding

Video tutorial for Section 8.4.3 "Asymptotic Efficiency of Iterative Methods for Zero Finding": (10 minutes) Download link, tablet notes

§8.4.3.1 (Efficiency) Efficiency is measured by forming the ratio of gain and the effort required to achieve it:

$$\text{Efficiency} = \frac{\text{gain}}{\text{effort}} .$$

For iterative methods for solving $F(\mathbf{x}) = 0$, $F : D \subset \mathbb{R}^n \to \mathbb{R}^n$, this means the following:

Efficiency of an iterative method (for solving $F(\mathbf{x}) = 0$)	$\leftrightarrow$	computational effort to reach a prescribed number of correct digits in result.

Ingredient ❶: Computational effort W per step

$$\text{e.g,} \quad W \approx \frac{\#\{\text{evaluations of } \mathrm{D}F\}}{\text{step}} + n \cdot \frac{\#\{\text{evaluations of } F'\}}{\text{step}} + \cdots .$$

Ingredient ❷: Number of steps $k = k(\rho)$ to achieve a *relative reduction of the error* by a factor of ρ (= gain),

$$\left|e^{(k)}\right| \leq \rho \left|e^{(0)}\right| , \qquad 0 < \rho < 1 \text{ prescribed.} \tag{8.4.3.2}$$

Here, $e^{(k)}$ stands for the iteration error in the k-th step.

Notice: $|\log \rho|$ $\leftrightarrow$ Gain in no. of significant digits of $x^{(k)}$ $[\log = \log_{10}]$

Measure for efficiency:

$$\text{Efficiency} := \frac{\text{no. of digits gained}}{\text{total work required}} = \frac{|\log \rho|}{k(\rho) \cdot W} \tag{8.4.3.3}$$

⌟

§8.4.3.4 (Minimal number of iteration steps) Let us consider an iterative method generating a sequence of approximate solutions the converges with order $p \geq 1$ ($\rightarrow$ Def. 8.2.2.10). From its error recursion we want to derive an estimate for the minimal number $k(\rho) \in \mathbb{N}$ of iteration steps required to achieve (8.4.3.2).

❶ Case $p = 1$, linearly convergent iteration:

Definition 8.2.2.1. Linear convergence

A sequence $\mathbf{x}^{(k)}$, $k = 0, 1, 2, \ldots$, in $\mathbb{R}^n$ **converges linearly** to $\mathbf{x}^* \in \mathbb{R}^n$,

$$\exists 0 < L < 1: \quad \left\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\right\| \leq L \left\|\mathbf{x}^{(k)} - \mathbf{x}^*\right\| \quad \forall k \in \mathbb{N}_0 .$$

This implies the recursion and estimate for the error norms:

$$\left|e^{(k+1)}\right| \leq C \left|e^{(k)}\right| \quad \text{with a constant} \quad 0 \leq C < 1 , \tag{8.4.3.5}$$

$$\blacktriangleright \quad \left|e^{(k)}\right| \leq C^k \left|e^{(0)}\right| \quad \forall k \in \mathbb{N}_0 . \tag{8.4.3.6}$$

$$\blacktriangleright \quad \left|e^{(k)}\right| \overset{!}{\leq} \rho \left|e^{(0)}\right| \quad \text{takes} \quad k(\rho) \geq \frac{\log \rho}{\log C} \quad \text{steps} . \tag{8.4.3.7}$$

❷ Case $p > 1$, order-p convergence:

Definition 8.2.2.10. order of convergence

A *convergent* sequence $\mathbf{x}^{(k)}$, $k = 0, 1, 2, \ldots$, in $\mathbb{R}^n$ with limit $\mathbf{x}^* \in \mathbb{R}^n$ converges with order p, $p \geq 1$, if

$$\exists C > 0: \quad \left\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\right\| \leq C \left\|\mathbf{x}^{(k)} - \mathbf{x}^*\right\|^p \quad \forall k \in \mathbb{N}_0 , \tag{8.4.3.8}$$

and, in addition, $C < 1$ in the case $p = 1$ (linear convergence $\rightarrow$ Def. 8.2.2.1).

For $p > 1$ this results in the error estimate

$$\begin{aligned}\left|e^{(k)}\right| &\leq C\left|e^{(k-1)}\right|^p \leq C^{1+p}\left|e^{(k-2)}\right|^{p^2} \leq C^{1+p+p^2}\left|e^{(k-3)}\right|^{p^3} \leq \dots \\ &\leq C^{1+p+p^2+p^3+\cdots+p^{k-1}}\left|e^{(0)}\right|^{p^k} = C^{\frac{p^k-1}{p-1}}\left|e^{(0)}\right|^{p^k-1}\left|e^{(0)}\right| , \quad k \in \mathbb{N} ,\end{aligned}$$

for some constant $C > 0$. Here, we use the geometric sum formula

$$1 + p^2 + p^3 + \cdots + p^{k-1} = \frac{p^k - 1}{p - 1} , \quad k \in \mathbb{N} .$$

We make the assumption that this estimate predicts convergence

$$\left|e^{(0)}\right| \quad \text{so small that} \quad C^{\frac{1}{p-1}}\left|e^{(0)}\right| < 1 . \tag{8.4.3.9}$$

$$\blacktriangleright \quad \left|e^{(k)}\right| \overset{!}{\leq} \rho\left|e^{(0)}\right| \quad \text{requires} \quad p^k \geq 1 + \frac{\log \rho}{\log C/_{p-1} + \log(\left|e^{(0)}\right|)} .$$

This permits us to estimate the minimal number of steps we have to execute to guarantee a reduction of the error by a factor of ρ

$$k(\rho) \geq \frac{\log(1 + \frac{\log \rho}{\log L_0})}{\log p} \quad , \quad L_0 := C^{1/p-1}\left|e^{(0)}\right| < 1 \quad \text{by (8.4.3.9)} . \tag{8.4.3.10}$$

Summing up, (8.4.3.7) and (8.4.3.10) give us explicit formulas for $k(\rho)$ as a function of ρ. ⌟

§8.4.3.11 (Asymptotic efficiency) Now we adopt an asymptotic perspective and ask for a large reduction of the error, that is $\rho \ll 1$.

If $\rho \ll 1$, then ($\log \rho, \log L_0 < 0$!)

$$\log(1 + \frac{\log \rho}{\log L_0}) \approx \log|\log \rho| - \log|\log L_0| \approx \log|\log \rho| .$$

This simplification will be made in the context of asymptotic considerations $\rho \to 0$ below.

$\blacktriangleright$ **asymptotic** efficiency for $\rho \ll 1$ (➜ $|\log \rho| \to \infty$):

$$\text{Efficiency}_{|\rho \ll 1} = \begin{cases} -\dfrac{\log C}{W} & \text{, if } \quad p = 1 , \\ \dfrac{\log p}{W} \cdot \dfrac{|\log \rho|}{\log(|\log \rho|)} & \text{, if } \quad p > 1 . \end{cases} \tag{8.4.3.12}$$

We conclude that

- when requiring high accuracy, linearly convergent iterations should not be used, because their efficiency does not increase for $\rho \to 0$,
- for method of order $p > 1$, the factor $\frac{\log p}{W}$ offers a gauge for efficiency.

EXAMPLE 8.4.3.13 (Efficiency of iterative methods) We "simulate" iterations to explore the quantitative dependence of efficiency on the order of the methods and the target accuracy.

We choose $\left|e^{(0)}\right| = 0.1$, $\rho = 10^{-8}$.

The plot displays the number of iteration steps according to (8.4.3.10).

Higher-order method require substantially fewer steps compared to low-order methods.

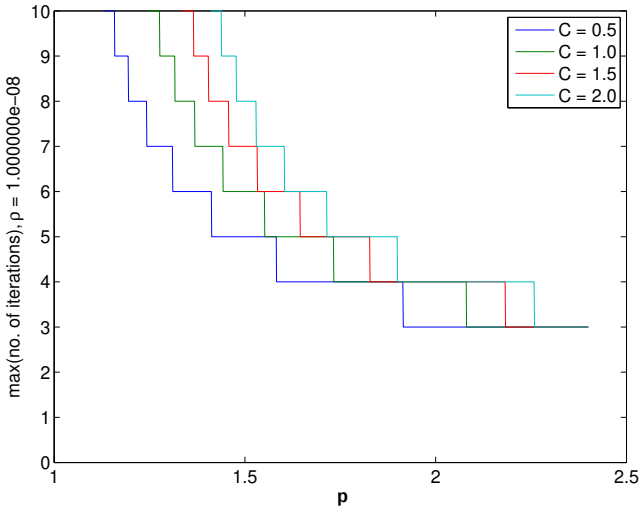


Fig. 306

We compare

- Newton's method from Section 8.4.2.1 and the
- secant method, see § 8.4.2.28,

in terms of number of steps required for a prescribed guaranteed error reduction, assuming $C = 1$ in both cases and for $\left|e^{(0)}\right| = 0.1$.

We observe that Newton's method requires only marginally fewer steps than the secant method.

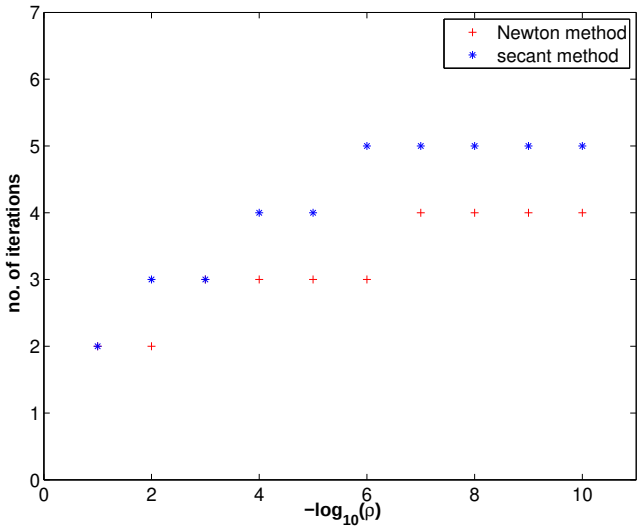


Fig. 307

§8.4.3.14 (Comparison of Newton's method and of the secant method) We draw conclusions from the discussion above and (8.4.3.12):

$$W_{\text{Newton}} = 2W_{\text{secant}}\,, \quad p_{\text{Newton}} = 2,\ p_{\text{secant}} = 1.62 \quad \succ \quad \frac{\log p_{\text{Newton}}}{W_{\text{Newton}}} : \frac{\log p_{\text{secant}}}{W_{\text{secant}}} = 0.71\,.$$

We set the effort for a step of Newton's method to twice that for a step of the secant method from Code 8.4.2.31, because we need an additional evaluation of the derivative F' in Newton's method.

$\succ$ The secant method is more efficient than Newton's method!

Review question(s) 8.4.3.15 (Asymptotic efficiency of iterative methods)

(Q8.4.3.15.A) How many step of an iteration that *converges linearly* with rate $L \in [0,1[$ are required to gain 6 digits of accuracy compared to the initial guess.

△

8.5 Newton's Method in $\mathbb{R}^n$

We return to non-linear systems of n equations with n unknowns, $n \in \mathbb{N}$:

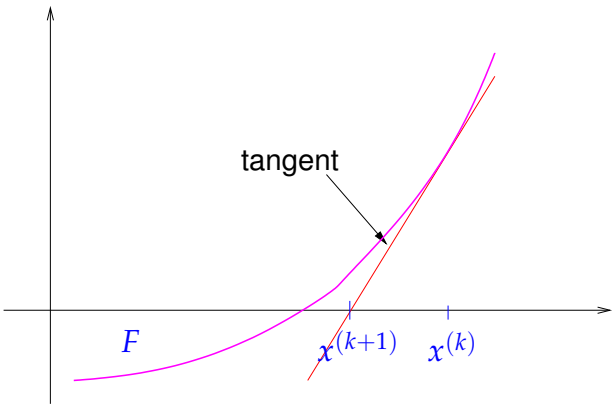
$$\text{for } F : D \subset \mathbb{R}^n \mapsto \mathbb{R}^n \quad \text{find} \quad \mathbf{x}^* \in D: \quad F(\mathbf{x}^*) = \mathbf{0}.$$

Throughout this section we assume that $F : D \subset \mathbb{R}^n \mapsto \mathbb{R}^n$ is *continuously differentiable*. Next we generalized the one-dimensional Newton method introduced in Section 8.4.2.1.

8.5.1 The Newton Iteration

Video tutorial for Section 8.5.1 "The Newton Iteration in $\mathbb{R}^n$ (I)": (10 minutes) Download link, tablet notes

Video tutorial for Section 8.5.1 "The Newton Iteration in $\mathbb{R}^n$ (II)": (15 minutes) Download link, tablet notes



Recall that the idea underlying the 1D Newton method for generating a sequence $\left(x^{(k)}\right)_{k\in\mathbb{N}_0}$ of approximate solutions of $F(x) = 0$ is

1. to replace F by its tangent in $(x^{(k)}, F(x^{(k)}))$,
2. to compute the next iterate $x^{(k+1)}$ as the zero of that tangent.

This led to the 1D Newton iteration

$$\boxed{x^{(k+1)} := x^{(k)} - \frac{F(x^{(k)})}{F'(x^{(k)})}}\,, \tag{8.4.2.1}$$

well defined provided that $F'(x^{(k)}) \neq 0$ for all k.

Remark 8.5.1.1 (Derivative-based local linear approximation of functions) The approximating tangent of the 1D Newton method for computing a zero of $F : I \subset \mathbb{R} \to \mathbb{R}$ is the graph of the function

$$\widetilde{F}(x) := F(x^{(k)}) + F'(x^{(k)})(x - x^{(k)})\,, \quad x \in \mathbb{R}\,. \tag{8.5.1.2}$$

From another perspective, $\widetilde{F}$ is just the Taylor expansion of F around $x^{(k)}$ truncated after the linear term. Thus, if F is twice continuous differentiable, $F \in C^2(I)$, by Taylor's formula [Str09, Satz 5.5.1] the tangent satisfies

$$|F(x) - \widetilde{F}(x)| = |F(x) - F(x^{(k)}) - F'(x^{(k)})(x - x^{(k)})| = O(|x - x^{(k)}|^2) \quad \text{for} \quad x \to x^{(k)}\,, \tag{8.5.1.3}$$

that is the tangent provides a local quadratic approximation of F around $x^{(k)}$. This directly gives us the local quadratic convergence of the 1D Newton method, recall § 8.4.2.5.

We know an analogous construction for general twice continuously differentiable $F : D \subset \mathbb{R}^n \mapsto \mathbb{R}^n$. We define the *affine linear* function

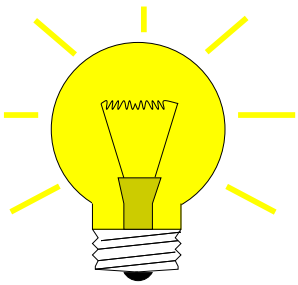
$$\widetilde{F}(\mathbf{x}) := F(\mathbf{x}^{(k)}) + \mathrm{D}F(\mathbf{x}^{(k)})(\mathbf{x} - \mathbf{x}^{(k)})\,, \quad \mathbf{x} \in \mathbb{R}^n\,, \tag{8.5.1.4}$$

where $\mathrm{D}F(\mathbf{z}) \in \mathbb{R}^{n,n}$ is the Jacobian of $F = [F_1, \ldots, F_n]^\top$ in $\mathbf{z} \in D$:

$$\mathrm{D}F(\mathbf{z}) = \left[\frac{\partial F_i}{\partial x_j}(\mathbf{z})\right]_{i,j=1}^{n} = \begin{bmatrix} \frac{\partial F_1}{\partial x_1}(\mathbf{z}) & \frac{\partial F_1}{\partial x_2}(\mathbf{z}) & \cdots & \cdots & \frac{\partial F_1}{\partial x_n}(\mathbf{z}) \\ \frac{\partial F_2}{\partial x_1}(\mathbf{z}) & & & & \frac{\partial F_2}{\partial x_n}(\mathbf{z}) \\ \vdots & & & & \vdots \\ \frac{\partial F_n}{\partial x_1}(\mathbf{z}) & \frac{\partial F_n}{\partial x_2}(\mathbf{z}) & \cdots & \cdots & \frac{\partial F_n}{\partial x_n}(\mathbf{z}) \end{bmatrix} . \tag{8.3.2.8}$$

This is the multi-dimensional generalization of a truncated Taylor expansion. From analysis we know that also in this case we have a local affine linear approximation by $\widetilde{F}$ with an error that quadratically depends on the distance to $\mathbf{x}^{(k)}$:

$$\left\| F(\mathbf{x}) - \widetilde{F}(\mathbf{x}) \right\| = \left\| F(\mathbf{x}) - F(\mathbf{x}^{(k)}) - \mathrm{D}F(\mathbf{x}^{(k)})(\mathbf{x} - \mathbf{x}^{(k)}) \right\| = O(\left\| \mathbf{x} - \mathbf{x}^{(k)} \right\|^2) \quad \text{for} \quad \mathbf{x} \to \mathbf{x}^{(k)} . \tag{8.5.1.5}$$



Idea (→ Section 8.4.2.1): local linearization:

Given $\mathbf{x}^{(k)} \in D$ ➢ $\mathbf{x}^{(k+1)}$ as zero of affine linear model function

$$F(\mathbf{x}) \approx \widetilde{F}_k(\mathbf{x}) := F(\mathbf{x}^{(k)}) + \mathrm{D}F(\mathbf{x}^{(k)})(\mathbf{x} - \mathbf{x}^{(k)}) ,$$

$\mathrm{D}F(\mathbf{x}) \in \mathbb{R}^{n,n}$ = Jacobian, $\mathrm{D}F(\mathbf{x}) := \left[\frac{\partial F_j}{\partial x_k}(\mathbf{x})\right]_{j,k=1}^{n}$.

Newton iteration: (generalizes (8.4.2.1) to $n > 1$)

$$\mathbf{x}^{(k+1)} := \mathbf{x}^{(k)} - \mathrm{D}F(\mathbf{x}^{(k)})^{-1} F(\mathbf{x}^{(k)}) \quad , \quad [\text{ if } \mathrm{D}F(\mathbf{x}^{(k)}) \text{ regular }] \tag{8.5.1.6}$$

Terminology: $-\mathrm{D}F(\mathbf{x}^{(k)})^{-1} F(\mathbf{x}^{(k)})$ is called the Newton correction

EXAMPLE 8.5.1.7 (Visualization: local affine linear approximation for $n = 2$) We give a graphical illustration of the idea of Newton's method for $n = 2$, where we seek $\mathbf{x}^* \in \mathbb{R}^2$ with $F(\mathbf{x}^*) = \mathbf{0}$, $F(\mathbf{x}) = [F_1(\mathbf{x}), F_2(\mathbf{x})]^\top$.

Sought: intersection point $\mathbf{x}^*$ of the curves $F_1(\mathbf{x}) = 0$ (solid red) and $F_2(\mathbf{x}) = 0$ (solid green).

Idea: Choose $\mathbf{x}^{(k+1)}$ as he intersection of two straight lines (dashed in Fig. 308)

$$L_1 := \{\mathbf{x} \in \mathbb{R}^2 : \widetilde{F}_{k,1}(\mathbf{x}) = 0\} ,$$
$$L_2 := \{\mathbf{x} \in \mathbb{R}^2 : \widetilde{F}_{k,2}(\mathbf{x}) = 0\} ,$$

which are the zero sets of the components of the affine linear model function $\widetilde{F}_k$. They are approximations of the zero curves of the components of F.

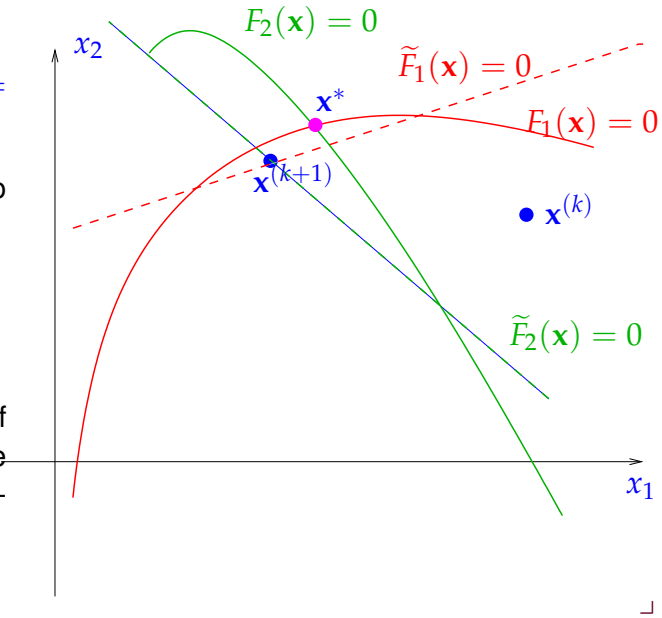


Fig. 308

§8.5.1.8 (Generic Newton method in C++) The following C++ implementation of the skeleton for Newton's method uses a *correction based* a posteriori termination criterion for the Newton iteration. It stops the

iteration if the *relative* size of the Newton correction drops below the prescribed relative tolerance `rtol`. If $\mathbf{x}^* \approx 0$ also the *absolute* size of the Newton correction has to be tested against an absolute tolerance `atol` in order to avoid non-termination despite convergence of the iteration,

C++ code 8.5.1.9: Newton's method in C++ ➙ GITLAB

```
template <typename FuncType, typename JacType, typename VecType>
VecType newton(FuncType &&F, JacType &&DFinv, VecType x, const double rtol,
               const double atol) {
  // Note that the vector x passes both the initial guess and also
  // contains the iterates
  VecType s(x.size()); // Vector for Newton corrections
  // Main loop
  do {
    s = DFinv(x, F(x)); // compute Newton correction
    x -= s;             // compute next iterate
  }
  // correction based termination (relative and absolute)
  while ((s.norm() > rtol * x.norm()) && (s.norm() > atol));
  return x;
}
```

☞ Objects of type **FuncType** must feature

```
Eigen::VectorXd operator(const Eigen::VectorXd &x);
```

that evaluates $F(\mathbf{x})$ ($\mathbf{x} \leftrightarrow$ `x`).

☞ Objects of type **JacType** must provide a method

```
Eigen::VectorXd operator (const Eigen::VectorXd &x, const
   Eigen::VectorXd &f);
```

that computes the Newton correction, that is it returns the solution of a linear system with system matrix $\mathrm{D}F(\mathbf{x})$ ($\mathbf{x} \leftrightarrow$ `x`) and right hand side $\mathbf{f} \leftrightarrow$ `f`.

☞ The function returns the computed approximate solution of the non-linear system.

The next code demonstrates the invocation of `newton` for a 2×2 non-linear system from a code relying on EIGEN. It also demonstrates the use of fixed size eigen matrices and vectors.

C++ code 8.5.1.10: Calling `newton` with EIGEN data types ➙ GITLAB

```
void newton2Ddriver() {
  // Function F defined through lambda function
  auto F = [](const Eigen::Vector2d &x) {
    Eigen::Vector2d z;
    const double x1 = x(0);
    const double x2 = x(1);
    z << x1 * x1 - 2 * x1 - x2 + 1, x1 * x1 + x2 * x2 - 1;
    return (z);
  };
  // Lambda function for the computation of the Newton correction
  auto DFinv = [](const Eigen::Vector2d &x, const Eigen::Vector2d &f) {
    Eigen::Matrix2d J;
    const double x1 = x(0);
    const double x2 = x(1);
    // Jacobian of F
    J << 2 * x1 - 2, -1, 2 * x1, 2 * x2;
```

```
    // Solve 2x2 linear system of equations
    Eigen::Vector2d s = J.lu().solve(f);
    return (s);
  };
  // initial guess
  const Eigen::Vector2d x0(2., 3.);

  // Invoke Newton's method
  const Eigen::Vector2d x = newton(F, DFinv, x0, 1E-6, 1E-8);
  std::cout << "||F(x)|| = " << F(x).norm() << std::endl;
}
```

New aspect for $n \gg 1$ (compared to $n = 1$-dimensional case, Section 8.4.2.1):
Computation of the Newton correction may be expensive!
(because it involves the solution of a LSE, *cf.* Thm. 2.5.0.2)

Remark 8.5.1.11 (Affine invariance of Newton method) An important property of the Newton iteration (8.5.1.6) is its affine invariance → [Deu11, Sect .1.2.2]

Consider $G_{\mathbf{A}}(\mathbf{x}) := \mathbf{A}F(\mathbf{x})$ with regular $\mathbf{A} \in \mathbb{R}^{n,n}$ so that $F(\mathbf{x}^*) = 0 \Leftrightarrow G_{\mathbf{A}}(\mathbf{x}^*) = 0$.

Affine invariance: The Newton iterations for $G_{\mathbf{A}}(\mathbf{x}) = \mathbf{0}$ are the same for all regular $\mathbf{A}$!

This is confirmed by a simple computation:

$$\mathrm{D}\,G_{\mathbf{A}}(\mathbf{x}) = \mathbf{A}\,\mathrm{D}\,F(\mathbf{x}) \quad \Rightarrow \quad \mathrm{D}\,G_{\mathbf{A}}(\mathbf{x})^{-1}G_{\mathbf{A}}(\mathbf{x}) = \mathrm{D}\,F(\mathbf{x})^{-1}\mathbf{A}^{-1}\mathbf{A}F(\mathbf{x}) = \mathrm{D}\,F(\mathbf{x})^{-1}F(\mathbf{x})\,.$$

Why is this an interesting property? Affine invariance should be used as a guideline for

- convergence theory for Newton's method: assumptions and results should be affine invariant, too.
- modifying and extending Newton's method: resulting schemes should preserve affine invariance.

In particular, termination criteria for Newton's method should also be affine invariant in the sense that, when applied for $G_{\mathbf{A}}$ they STOP the iteration at exactly the same step for any choice of the regular matrix $\mathbf{A}$.

The function $F : \mathbb{R}^n \to \mathbb{R}^n$ defining the non-linear system of equations may be given in various formats, as explicit expression or rather implicitly. In most cases, $\mathrm{D}\,F$ has to be computed *symbolically* in order to obtain concrete formulas for the Newton iteration. We now learn how these symbolic computations can be carried out harnessing advanced techniques of multi-variate calculus.

§8.5.1.12 (Derivatives and their role in Newton's method) The reader will probably agree that the derivative of a function $F : I \subset \mathbb{R} \to \mathbb{R}$ in $x \in I$ is a number $F'(x) \in \mathbb{R}$, the derivative of a function $F : D \subset \mathbb{R}^n \to \mathbb{R}^m$, in $\mathbf{x} \in D$ a matrix $\mathrm{D}\,F(\mathbf{x}) \in \mathbb{R}^{m,n}$. However, the nature of a derivative in a point is

that of a linear mapping that approximates F locally up to second order:

Definition 8.5.1.13. Derivative of functions between vector spaces

Let V, W be finite dimensional vector spaces and $F : D \subset V \mapsto W$ a sufficiently smooth mapping. The **derivative** (differential) $\mathrm{D}F(\mathbf{x})$ of F in $\mathbf{x} \in V$ is the *unique* linear mapping $\mathrm{D}F(\mathbf{x}) : V \mapsto W$ such that there is a $\delta > 0$ and a function $\epsilon : [0,\delta] \to \mathbb{R}^+$ satisfying $\lim_{\xi\to 0} \epsilon(\xi) = 0$ such that

$$\|F(\mathbf{x}+\mathbf{h}) - F(\mathbf{x}) - \mathrm{D}F(\mathbf{x})\mathbf{h}\| = \epsilon(\|\mathbf{h}\|) \quad \forall \mathbf{h} \in V\,,\ \|\mathbf{h}\| < \delta\,. \tag{8.5.1.14}$$

The vector $\mathbf{h}$ in (8.5.1.14) may be called "direction of differentiation".

- ☞ Note that $\mathrm{D}F(\mathbf{x})\mathbf{h} \in W$ is the vector returned by the linear mapping $\mathrm{D}F(\mathbf{x})$ when applied to $\mathbf{h} \in V$.
- ☞ In Def. 8.5.1.13 $\|\cdot\|$ can be any norm on V ($\rightarrow$ Def. 1.5.5.4).
- ☞ A common shorthand notation for (8.5.1.14) relies on the "little-o" Landau symbol:

  $$\|F(\mathbf{x}+\mathbf{h}) - F(\mathbf{x}) - \mathrm{D}F(\mathbf{h})\mathbf{h}\| = o(\mathbf{h}) \quad \text{for} \quad \mathbf{h} \to 0\,,$$

  which designates a remainder term tending to 0 as its arguments tends to 0.
- ☞ Choosing bases of V and W, $\mathrm{D}F(\boldsymbol{x})$ can be described by the Jacobian (matrix) (8.3.2.8), because every linear mapping between finite dimensional vector spaces has a matrix representation after bases have been fixed. Thus, the derivative is usually written as a matrix-valued function on D.

In the context of the Newton iteration (8.5.1.6) the computation of the Newton correction $\mathbf{s}$ in the $k+1$-th step amounts to solving a linear system of equations:

$$\mathbf{s} = -\mathrm{D}F(\mathbf{x}^{(k)})^{-1}F(\mathbf{x}^{(k)}) \quad \Leftrightarrow \quad \mathrm{D}F(\mathbf{x}^{(k)})\mathbf{s} = -F(\mathbf{x}^{(k)})\,.$$

Matching this with Def. 8.5.1.13 we see that we need only determine expressions for $\mathrm{D}F(\mathbf{x}^{(k)})\mathbf{h}$, $\mathbf{h} \in V$, in order to state the LSE yielding the Newton correction. This will become important when applying the "compact" differentiation rules discussed next. ⌟

§8.5.1.15 ("High-level" Differentiation rules $\rightarrow$ Repetition of basic analysis skills)

Video tutorial for § 8.5.1.15 "Multi-dimensional Differentiation": (20 minutes) Download link, tablet notes

Stating the Newton iteration (8.5.1.6) for $F : \mathbb{R}^n \mapsto \mathbb{R}^n$ through an analytic formula entails computing the Jacobian $\mathrm{D}F$. The safe, but tedious way is to use the definition (8.3.2.8) directly and compute the partial derivatives.

To avoid cumbersome component-oriented considerations, it is sometimes useful to know the *rules of multidimensional differentiation*:

Immediate from Def. 8.5.1.13 are the following differentiation rules (V, W, U, Z are finite-dimensional normed vector spaces, all functions assumed to be differentiable):

- For $F : V \mapsto W$ linear, we have $\mathrm{D}F(\mathbf{x}) = F$ for all $\mathbf{x} \in V$

  (For instance, if $F : \mathbb{R}^n \to \mathbb{R}^m$, $F(\mathbf{x}) = \mathbf{A}\mathbf{x}$, $\mathbf{A} \in \mathbb{R}^{m,n}$, then $\mathrm{D}F(\mathbf{x}) = \mathbf{A}$ for all $\mathbf{x} \in \mathbb{R}^n$.)
- Chain rule: For $F : V \mapsto W$, $G : U \mapsto V$ sufficiently smooth

  $$\mathrm{D}(F \circ G)(\mathbf{x})\mathbf{h} = \mathrm{D}F(G(\mathbf{x}))(\mathrm{D}G(\mathbf{x})\mathbf{h})\,, \quad \mathbf{h} \in V\,, \mathbf{x} \in D\,. \tag{8.5.1.16}$$

This can be justified by a formal computation

$$F(G(\mathbf{x}+\mathbf{h})) = F(G(\mathbf{x}) + \mathrm{D}\,G(\mathbf{x})\mathbf{h} + o(\|\mathbf{h}\|)) = F(G(\mathbf{x})) + \mathrm{D}\,F(G(\mathbf{x})) \circ \mathrm{D}\,G(\mathbf{x})\,\mathbf{h} + o(\|\mathbf{h}\|)) \,.$$

- Product rule: $F : D \subset V \mapsto W$, $G : D \subset V \mapsto U$ sufficiently smooth, $b : W \times U \mapsto Z$ bilinear, *ie.*, linear in each argument:

$$T(\mathbf{x}) := b(F(\mathbf{x}), G(\mathbf{x})) \quad \Rightarrow \quad \mathrm{D}\,T(\mathbf{x})\mathbf{h} = b(\mathrm{D}\,F(\mathbf{x})\mathbf{h}, G(\mathbf{x})) + b(F(\mathbf{x}), \mathrm{D}\,G(\mathbf{x})\mathbf{h}) \,, \quad \mathbf{h} \in V, \mathbf{x} \in D \,. \tag{8.5.1.17}$$

We see this by formal computations, making heavy use of the bilinearity of b in step $(*)$:

$$\begin{aligned} T(\mathbf{x}+\mathbf{h}) &= b(F(\mathbf{x}+\mathbf{h}), G(\mathbf{x}+\mathbf{h})) \\ &= b(F(\mathbf{x}) + \mathrm{D}\,F(\mathbf{x})\mathbf{h} + o(\|\mathbf{h}\|), G(\mathbf{x}) + \mathrm{D}\,G(\mathbf{x})\mathbf{h} + o(\|\mathbf{h}\|)) \\ &\overset{(*)}{=} b(F(\mathbf{x}), G(\mathbf{x})) + \underbrace{b(\mathrm{D}\,F(\mathbf{x})\mathbf{h}, G(\mathbf{x})) + b(F(\mathbf{x}), \mathrm{D}\,G(\mathbf{x})\mathbf{h})}_{=\mathrm{D}\,T(\mathbf{x})\mathbf{h}} + o(\|\mathbf{h}\|) \,, \end{aligned}$$

where the term $b(\mathrm{D}\,F(\mathbf{x})\mathbf{h}, \mathrm{D}\,G(\mathbf{x})\mathbf{h})$ is obviously $O(\|\mathbf{h}\|^2)$ for $\mathbf{h} \to \mathbf{0}$ and, therefore, can be "thrown into the garbage bin" of $o(\|\mathbf{h}\|)$.

Advice: If you do not feel comfortable with these rules of multidimensional differential calculus, please resort to detailed componentwise/entrywise calculations according to (8.3.2.8) ("pedestrian differentiation"), though they may be tedious.

The first and second derivatives of real-valued functions occur frequently and have special names, see [Str09, Def. 7.3.2] and [Str09, Satz 7.5.3].

Definition 8.5.1.18. Gradient and Hessian

For sufficiently smooth $F : D \subset \mathbb{R}^n \mapsto \mathbb{R}$ the **gradient** $\mathbf{grad}\,F : D \mapsto \mathbb{R}^n$, and the **Hessian** (matrix) $\mathrm{H}\,F(\mathbf{x}) : D \mapsto \mathbb{R}^{n,n}$ are defined as

$$\mathbf{grad}\,F(\mathbf{x})^T\mathbf{h} := \mathrm{D}\,F(\mathbf{x})\mathbf{h} \,, \qquad \mathbf{h}_1^T\,\mathrm{H}\,F(\mathbf{x})\mathbf{h}_2 := \mathrm{D}(\mathrm{D}\,F(\mathbf{x})(\mathbf{h}_1))(\mathbf{h}_2) \,, \quad \forall \mathbf{h}, \mathbf{h}_1, \mathbf{h}_2 \in \mathbb{R}^n \,.$$

In other words,

- the gradient $\mathbf{grad}\,F(\mathbf{x})$ is the vector representative of the linear mapping $\mathrm{D}\,F(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$,
- and the Hessian $\mathrm{H}\,F(\mathbf{x})$ is the matrix representative of the *bilinear* mapping $\mathrm{D}\{\mathbf{z} \mapsto \mathrm{D}\,F(\mathbf{z})\}(\mathbf{x}) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$.

⌟

EXAMPLE 8.5.1.19 (Derivative of a bilinear form) We start with a simple example:

$$\Psi : \mathbb{R}^n \mapsto \mathbb{R} \quad , \quad \Psi(\mathbf{x}) := \mathbf{x}^\top \mathbf{A}\mathbf{x} \,, \quad \mathbf{A} \in \mathbb{R}^{n,n} \,.$$

This is the general matrix representation of a bilinear form on $\mathbb{R}^n$. We want to compute the gradient of Ψ. We do this in two ways:

❶ "High level differentiation": We apply the product rule (8.5.1.17) with $D = V = W = U = \mathbb{R}^n$, $Z = \mathbb{R}$, $F, G = Id$, which means $\mathrm{D}F(\mathbf{x}) = \mathrm{D}G(\mathbf{x}) = \mathbf{I}$, and the *bilinear form* $b(\mathbf{x},\mathbf{y}) := \mathbf{x}^T\mathbf{A}\mathbf{y}$:

$$\blacktriangleright \quad \mathrm{D}\Psi(\mathbf{x})\mathbf{h} = \mathbf{h}^\top\mathbf{A}\mathbf{x} + \mathbf{x}^\top\mathbf{A}\mathbf{h} = (\mathbf{A}\mathbf{x})^\top\mathbf{h} = \mathbf{x}^\top\mathbf{A}\mathbf{h} = \underbrace{\left(\mathbf{x}^\top\mathbf{A}^\top + \mathbf{x}^\top\mathbf{A}\right)}_{=(\mathbf{grad}\,\Psi(\mathbf{x}))^\top}\mathbf{h}\,,$$

Hence, $\mathbf{grad}\,\Psi(\mathbf{x}) = (\mathbf{A} + \mathbf{A}^T)\mathbf{x}$ according to the definition of a gradient (→ Def. 8.5.1.18).

❷ "Low level/pedestrian differentiation": Using the rules of matrix×vector multiplication, Ψ can be written in terms of the vector components $x_i := (\mathbf{x})_i$, $i = 1,\ldots,n$: Fixing $i \in \{1,\ldots,n\}$ we find

$$\Psi(\mathbf{x}) = \sum_{k=1}^n\sum_{j=1}^n (\mathbf{A})_{k,j}x_kx_j = (\mathbf{A})_{i,i}x_i^2 + \sum_{\substack{j=1\\ j\neq i}}^n (\mathbf{A})_{i,j}x_ix_j + \sum_{\substack{k=1\\ k\neq i}}^n (\mathbf{A})_{k,i}x_kx_i + \sum_{\substack{k=1\\ k\neq i}}^n\sum_{\substack{j=1\\ j\neq i}}^n (\mathbf{A})_{k,j}x_kx_j\,.$$

Now computing the partial derivative with respect to x_i is easy:

$$\begin{aligned}\frac{\partial\Psi(\mathbf{x})}{\partial x_i}(\mathbf{x}) &= 2(\mathbf{A})_{i,i}x_i + \sum_{\substack{j=1\\ j\neq i}}^n (\mathbf{A})_{i,j}x_j + \sum_{\substack{k=1\\ k\neq i}}^n (\mathbf{A})_{k,i}x_k = \sum_{j=1}^n (\mathbf{A})_{i,j}x_j + \sum_{k=1}^n (\mathbf{A})_{k,i}x_k\\ &= (\mathbf{A}\mathbf{x} + \mathbf{A}^\top\mathbf{x})_i\,,\quad i = 1,\ldots,n\,.\end{aligned}$$

This provides the components of the gradient, since $i \in \{1,\ldots,n\}$ was arbitrary.

Of course, the results obtained by both methods must agree! ⌟

EXAMPLE 8.5.1.20 (Derivative of Euclidean norm) We seek the derivative of the Euclidean norm, that is, of the function $F(\mathbf{x}) := \|\mathbf{x}\|_2$, $\mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ (F is defined but not differentiable in $\mathbf{x} = \mathbf{0}$, just look at the case $n = 1$!)

❶ "High level differentiation": We can write F as the composition of two functions $F = G \circ H$ with

$$\begin{aligned}G &: \mathbb{R}_+ \to \mathbb{R}_+\,,\quad G(\xi) := \sqrt{\xi}\,,\\ H &: \mathbb{R}^n \to \mathbb{R}\,,\quad H(\mathbf{x}) := \mathbf{x}^\top\mathbf{x}\,.\end{aligned}$$

Using the rule for the differentiation of bilinear forms from Ex. 8.5.1.19 for the case $\mathbf{A} = \mathbf{I}$ and basic calculus, we find

$$\begin{aligned}\mathrm{D}H(\mathbf{x})\mathbf{h} &= 2\mathbf{x}^\top\mathbf{h}\,,\quad \mathbf{x},\mathbf{h} \in \mathbb{R}^n\,,\\ \mathrm{D}G(\xi)\zeta &= \frac{\zeta}{2\sqrt{\xi}}\,,\quad \xi > 0,\ \zeta \in \mathbb{R}\,.\end{aligned}$$

Finally, the chain rule (8.5.1.16) gives

$$\mathrm{D}F(\mathbf{x})\mathbf{h} = \mathrm{D}G(H(\mathbf{x}))(\mathrm{D}H(\mathbf{x})\mathbf{h}) = \frac{2\mathbf{x}^\top\mathbf{h}}{2\sqrt{\mathbf{x}^\top\mathbf{x}}} = \frac{\mathbf{x}^\top}{\|\mathbf{x}\|_2}\cdot\mathbf{h}\,. \tag{8.5.1.21}$$

$$\text{Def. 8.5.1.18} \quad\Rightarrow\quad \mathbf{grad}\,F(\mathbf{x}) = \frac{\mathbf{x}}{\|\mathbf{x}\|_2}\,.$$

❷ "Pedestrian differentiation": We can explicitly write $F(\mathbf{x})$ as

$$F(\mathbf{x}) = \sqrt{x_1^2 + x_2^2 + \cdots + x_i^2 + \cdots + x_n^2}\,,$$

$$\frac{\partial (F(\mathbf{x}))_i}{\partial x_i} = \frac{1}{2\sqrt{x_1^2 + x_2^2 + \cdots + x_i^2 + \cdots + x_n^2}} \cdot 2x_i\,, i = 1, \ldots, n \quad \text{[chain rule]}\,.$$

$$\mathbf{grad}\, F(\mathbf{x}) = \frac{1}{\|\mathbf{x}\|_2}[x_1, x_2, \ldots, x_n]^\top = \frac{\mathbf{x}}{\|\mathbf{x}\|_2}\,.$$

⌟

§8.5.1.22 (Newton iteration via product rule) This paragraph explains the use of the general product rule (8.5.1.17) to derive the linear system solved by the Newton correction. It implements the insights from § 8.5.1.12.

We seek solutions of $F(\mathbf{x}) = \mathbf{0}$ with $F(\mathbf{x}) := b(G(\mathbf{x}), H(\mathbf{x}))$, where

- V, W are some vector spaces (finite- or even infinite-dimensional),
- $G : D \to V$, $H : D \to W$, $D \subset \mathbb{R}^n$, are continuously differentiable in the sense of Def. 8.5.1.13,
- $b : V \times W \mapsto \mathbb{R}^n$ is bilinear (linear in each argument).

According to the general product rule (8.5.1.17) we have

$$\mathrm{D}\, F(\mathbf{x})\mathbf{h} = b(\mathrm{D}\, G(\mathbf{x})\mathbf{h}, H(\mathbf{x})) + b(G(\mathbf{x}), \mathrm{D}\, H(\mathbf{x})\mathbf{h})\,, \quad \mathbf{h} \in \mathbb{R}^n\,. \tag{8.5.1.23}$$

This already defines the linear system of equations to be solved to compute the Newton correction $\mathbf{s}$

$$b(\mathrm{D}\, G(\mathbf{x}^{(k)})\mathbf{s}, H(\mathbf{x}^{(k)})) + b(G(\mathbf{x}^{(k)}), \mathrm{D}\, H(\mathbf{x}^{(k)})\mathbf{s}) = -b(G(\mathbf{x}^{(k)}), H(\mathbf{x}^{(k)}))\,. \tag{8.5.1.24}$$

Since the left-hand side is linear in $\mathbf{s}$, this really represents a square linear system of n equations. The next example will present a concrete case. ⌟

§8.5.1.25 (A "quasi-linear" system of equations) We call a quasilinear system of equations a non-linear equation of the form

$$\mathbf{A}(\mathbf{x})\mathbf{x} = \mathbf{b} \quad \text{with } \mathbf{b} \in \mathbb{R}^n\,, \tag{8.5.1.26}$$

where $\mathbf{A} : D \subset \mathbb{R}^n \to \mathbb{R}^{n,n}$ is a matrix-valued function. On other words, a quasi-linear system is a "linear system of equations with solution-dependent system matrix". It incarnates a zero-finding problem for a function $F : \mathbb{R}^n \to \mathbb{R}^n$ with a bilinear structure as introduced in § 8.5.1.22.

For many quasi-linear systems, for which there exist solutions, the fixed point iteration (→ Section 8.3)

$$\mathbf{x}^{(k+1)} = \mathbf{A}(x^{(k)})^{-1}\mathbf{b} \quad \Leftrightarrow \quad \mathbf{A}(\mathbf{x}^{(k)})\mathbf{x}^{(k+1)} = \mathbf{b}\,, \tag{8.5.1.27}$$

provides a convergent iteration, provided that a good initial guess is available.

We can also reformulate

$$(8.5.1.26) \quad \Leftrightarrow \quad F(\mathbf{x}) = \mathbf{0} \quad \text{with} \quad F(\mathbf{x}) = \mathbf{A}(\mathbf{x})\mathbf{x} - \mathbf{b}\,.$$

If $\mathbf{x} \mapsto \mathbf{A}(\mathbf{x})$ is differentiable, the product rule (8.5.1.17) yields

$$\mathrm{D}\, F(\mathbf{x})\mathbf{h} = (\mathrm{D}\, \mathbf{A}(\mathbf{x})\mathbf{h})\mathbf{x} + \mathbf{A}(\mathbf{x})\mathbf{h}\,, \quad \mathbf{h} \in \mathbb{R}^n\,. \tag{8.5.1.28}$$

Note that $\mathrm{D}\,\mathbf{A}(\mathbf{x}^{(k)})$ is a mapping from $\mathbb{R}^n$ into $\mathbb{R}^{n,n}$, which gets $\mathbf{h}$ as an argument. Then the Newton iteration reads

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \mathbf{s}\,, \quad \mathrm{D}\,F(\mathbf{x}^{(k)})\mathbf{s} = (\mathrm{D}\,\mathbf{A}(\mathbf{x}^{(k)})\mathbf{s})\mathbf{x}^{(k)} + \mathbf{A}(\mathbf{x}^{(k)})\mathbf{s} = \mathbf{A}(\mathbf{x}^{(k)})\mathbf{x}^{(k)} - \mathbf{b}\,. \tag{8.5.1.29}$$

The next example will examine a concrete quasi-linear system of equations. ⌟

EXAMPLE 8.5.1.30 (A special quasi-linear system of equations) We consider the quasi-linear system of equations

$$\mathbf{A}(\mathbf{x})\mathbf{x} = \mathbf{b} \quad , \quad \mathbf{A}(\mathbf{x}) := \begin{bmatrix} \gamma(\mathbf{x}) & 1 & & & & \\ 1 & \gamma(\mathbf{x}) & 1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & \gamma(\mathbf{x}) & 1 \\ & & & & 1 & \gamma(\mathbf{x}) \end{bmatrix} \in \mathbb{R}^{n\times n}\,, \tag{8.5.1.31}$$

where $\gamma(\mathbf{x}) := 3 + \|\mathbf{x}\|_2$ (Euclidean vector norm), the right hand side vector $\mathbf{b} \in \mathbb{R}^n$ is given and $\mathbf{x} \in \mathbb{R}^n$ is unknown.

In order to compute the derivative of $F(\mathbf{x}) := \mathbf{A}(\mathbf{x})\mathbf{x} - \mathbf{b}$ it is advisable to rewrite

$$\mathbf{A}(\mathbf{x})\mathbf{x} = \mathbf{T}\mathbf{x} + \mathbf{x}\|\mathbf{x}\|_2\,, \quad \mathbf{T} := \begin{bmatrix} 3 & 1 & & & & \\ 1 & 3 & 1 & & & \\ & \ddots & 3 & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & 3 & 1 \\ & & & & 1 & 3 \end{bmatrix}.$$

The derivative of the first term is straightforward, because it is linear in $\mathbf{x}$, see the discussion following Def. 8.5.1.13.

The "pedestrian" approach to the second term starts with writing it explicitly in components as

$$(\mathbf{x}\|\mathbf{x}\|)_i = x_i\sqrt{x_1^2 + \cdots + x_n^2}\,, \quad i = 1,\ldots,n\,.$$

Then we can compute the Jacobian according to (8.3.2.8) by taking partial derivatives:

$$\frac{\partial}{\partial x_i}(\mathbf{x}\|\mathbf{x}\|)_i = \sqrt{x_1^2 + \cdots + x_n^2} + x_i\frac{x_i}{\sqrt{x_1^2 + \cdots + x_n^2}}\,,$$

$$\frac{\partial}{\partial x_j}(\mathbf{x}\|\mathbf{x}\|)_i = x_i\frac{x_j}{\sqrt{x_1^2 + \cdots + x_n^2}}\,, \quad j \neq i\,.$$

For the "high level" treatment of the second term $\mathbf{x} \mapsto \mathbf{x}\|\mathbf{x}\|_2$ we apply the product rule (8.5.1.17), together with (8.5.1.21):

$$\mathrm{D}\,F(\mathbf{x})\mathbf{h} = \mathbf{T}\mathbf{h} + \|\mathbf{x}\|_2\mathbf{h} + \mathbf{x}\frac{\mathbf{x}^\top\mathbf{h}}{\|\mathbf{x}\|_2} = \big(\mathbf{A}(\mathbf{x}) + \frac{\mathbf{x}\mathbf{x}^\top}{\|\mathbf{x}\|_2}\big)\mathbf{h}\,.$$

Thus, in concrete terms the Newton iteration (8.5.1.29) becomes

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \Big(\mathbf{A}(\mathbf{x}^{(k)}) + \frac{\mathbf{x}^{(k)}(\mathbf{x}^{(k)})^\top}{\|\mathbf{x}\|_2}\Big)^{-1}(\mathbf{A}(\mathbf{x}^{(k)})\mathbf{x}^{(k)} - \mathbf{b})\,.$$

Note that the coefficient matrix of the linear system to be solved in each step is a rank-1-modification (2.6.0.16) of the symmetric positive definite tridiagonal matrix $\mathbf{A}(\mathbf{x}^{(k)})$, *cf.* Lemma 2.8.0.13. Thus the Sherman-Morrison-Woodbury formula from Lemma 2.6.0.21 can be used to solve it efficiently. ⌟

§8.5.1.32 (Implicit differentiation of bilinear expressions) Given are

- a finite-dimensional vector space W,
- a continuously differentiable function $G : D \subset \mathbb{R}^n \to V$ into some vector space V,
- a bilinear mapping $b : V \times W \to W$.

Let $F : D \to W$ be *implicitly defined* by

$$F(\mathbf{x}): \qquad b(G(\mathbf{x}), F(\mathbf{x})) = \mathbf{b} \in W \,. \tag{8.5.1.33}$$

This relationship will provide a valid definition of F in a neighborhood of $\mathbf{x}_0 \in W$, if we assume that there is $\mathbf{x}_0, \mathbf{z}_0 \in W$ such that $b(G(\mathbf{x}_0), z_0) = \mathbf{b}$, and that the linear mapping $\mathbf{z} \mapsto b(G(\mathbf{x}_0), \mathbf{z})$ is *invertible*. Then, for $\mathbf{x}$ close to $\mathbf{x}_0$, $F(\mathbf{x})$ can be computed by solving a square linear system of equations in W. In Ex. 8.4.2.9 we already saw an example of an implicitly defined F for $W = \mathbb{R}$.

We want to solve $F(\mathbf{x}) = \mathbf{0}$ for this implicitly defined F by means of Newton's method. In order to determine the derivative of F we resort to implicit differentiation [Str09, Sect. 7.8] of the defining equation (8.5.1.33) by means of the general product rule (8.5.1.17). We formally differentiate both sides of (8.5.1.33):

$$b(\mathrm{D}\, G(\mathbf{x})\mathbf{h}, F(\mathbf{x})) + b(G(\mathbf{x}), \mathrm{D}\, F(\mathbf{x})\mathbf{h}) = \mathbf{0} \quad \forall \mathbf{h} \in W \,, \tag{8.5.1.34}$$

and find that the Newton correction $\mathbf{s}$ in the $k+1$-th Newton step can be computed as follows:

$$\begin{aligned} & \mathrm{D}\, F(\mathbf{x}^{(k)})\mathbf{s} = -F(\mathbf{x}^{(k)}) \quad \Rightarrow \quad b(G(\mathbf{x}^{(k)}), \mathrm{D}\, F(\mathbf{x}^{(k)})\mathbf{s}) = -b(G(\mathbf{x}^{(k)}), F(\mathbf{x}^{(k)})) \\ \overset{(8.5.1.34)}{\Rightarrow} \quad & b(\mathrm{D}\, G(\mathbf{x}^{(k)})\mathbf{s}, F(\mathbf{x}^{(k)})) = b(G(\mathbf{x}^{(k)}), F(\mathbf{x}^{(k)})) \,, \end{aligned}$$

which constitutes an $\dim W \times \dim W$ linear system of equations. The next example discusses a concrete application of implicit differentiation with $W = \mathbb{R}^{n,n}$. ⌟

EXAMPLE 8.5.1.35 (Derivative of matrix inversion) We consider matrix inversion as a mapping and (formally) compute its derivative, that is, the derivative of function

$$\mathrm{inv} : \begin{cases} \mathbb{R}^{n,n}_* & \to & \mathbb{R}^{n,n} \\ \mathbf{X} & \mapsto & \mathbf{X}^{-1} \end{cases} \,,$$

where $\mathbb{R}^{n,n}_*$ denotes the (open) set of invertible $n \times n$-matrices, $n \in \mathbb{N}$.

We apply the technique of implicit differentiation from § 8.5.1.32 to the equation

$$\mathrm{inv}(\mathbf{X}) \cdot \mathbf{X} = \mathbf{I} \,, \quad \mathbf{X} \in \mathbb{R}^{n,n}_* \,. \tag{8.5.1.36}$$

Differentiation on both sides of (8.5.1.36) by means of the product rule (8.5.1.17) yields

$$\begin{aligned} & \mathrm{D}\,\mathrm{inv}(\mathbf{X})\mathbf{H} \cdot \mathbf{X} + \mathrm{inv}(\mathbf{X}) \cdot \mathbf{H} = \mathbf{O} \,, \quad \mathbf{H} \in \mathbb{R}^{n,n} \,, \\ \blacktriangleright \quad & \boxed{\mathrm{D}\,\mathrm{inv}(\mathbf{X})\mathbf{H} = -\mathbf{X}^{-1}\mathbf{H}\mathbf{X}^{-1}} \,, \quad \mathbf{H} \in \mathbb{R}^{n,n} \,. \end{aligned} \tag{8.5.1.37}$$

For $n = 1$ we get $\mathrm{D}\,\mathrm{inv}(x)h = -\frac{h}{x^2}$, which recovers the well-known derivative of the function $x \to x^{-1}$. ⌟

EXAMPLE 8.5.1.38 (Matrix inversion by means of Newton's method → [Ale12; PS91]) Surprisingly, it is possible to obtain the inverse of a matrix as the solution of a non-linear system of equations. Thus it can be computed using Newton's method.

Given a regular matrix $\mathbf{A} \in \mathbb{R}^{n,n}$, its inverse can be defined as the unique zero of a function:

$$\mathbf{X} = \mathbf{A}^{-1} \iff F(\mathbf{X}) = \mathbf{O} \quad \text{for} \quad F : \begin{cases} \mathbb{R}^{n,n}_* & \to & \mathbb{R}^{n,n} \\ \mathbf{X} & \mapsto & \mathbf{A} - \mathbf{X}^{-1} \end{cases} .$$

Using (8.5.1.37) we find for the derivative of F in $\mathbf{X} \in \mathbb{R}^{n,n}_*$

$$\mathrm{D}F(\mathbf{X})\mathbf{H} = \mathbf{X}^{-1}\mathbf{H}\mathbf{X}^{-1} , \quad \mathbf{H} \in \mathbb{R}^{n,n} . \tag{8.5.1.39}$$

The abstract Newton iteration (8.5.1.6) for F reads

$$\mathbf{X}^{(k+1)} = \mathbf{X}^{(k)} - \mathbf{S} , \quad \mathbf{S} := \mathrm{D}F(\mathbf{X}^{(k)})^{-1}F(\mathbf{X}^{(k)}) . \tag{8.5.1.40}$$

The Newton correction $\mathbf{S}$ in the k-th step solves the *linear system of equations*

$$\mathrm{D}F(\mathbf{X}^{(k)})\mathbf{S} \overset{(8.5.1.39)}{=} (\mathbf{X}^{(k)})^{-1}\mathbf{S}(\mathbf{X}^{(k)})^{-1} = F(\mathbf{X}^{(k)}) = \mathbf{A} - (\mathbf{X}^{(k)})^{-1} .$$

$$\Rightarrow \quad \mathbf{S} = \mathbf{X}^{(k)}(\mathbf{A} - (\mathbf{X}^{(k)})^{-1})\mathbf{X}^{(k)} = \mathbf{X}^{(k)}\mathbf{A}\mathbf{X}^{(k)} - \mathbf{X}^{(k)} . \tag{8.5.1.41}$$

in (8.5.1.40) ⇒

$$\boxed{\mathbf{X}^{(k+1)} = \mathbf{X}^{(k)} - (\mathbf{X}^{(k)}\mathbf{A}\mathbf{X}^{(k)} - \mathbf{X}^{(k)}) = \mathbf{X}^{(k)}(2\mathbf{I} - \mathbf{A}\mathbf{X}^{(k)})} . \tag{8.5.1.42}$$

This is the Newton iteration (8.5.1.6) for $F(\mathbf{X}) = \mathbf{O}$ that we expect to converge locally to $\mathbf{X}^* := \mathbf{A}^{-1}$. ⌟

Remark 8.5.1.43 (Simplified Newton method [DR08, Sect. 5.6.2]) Computing the Newton correction can be expensive owing to the $O(n^3)$ asymptotic cost (→ § 2.5.0.4) of solving a different large $n \times n$ linear system of equations in every step of the Newton iteration (8.5.1.6).

We know that the cost of a linear solve can be reduced to $O(n^2)$ if the coefficient matrix is available in LU- or QR factorized form, see, e.g., § 2.3.2.15. This motivates the attempt to "samemp*freeze" the Jacobian in the Newton iteration and use $\mathrm{D}F(x^{(0)})$ throughout, which leads to the simplified Newton iteration:

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \mathrm{D}F(\mathbf{x}^{(0)})^{-1}F(\mathbf{x}^{(k)}) , \quad k = 0, 1, \ldots .$$

The following C++ function implements a template for this simplified Newton Method and uses the same Jacobian $\mathrm{D}F(\mathbf{x}^{(0)})$ for all steps, which makes it possible to reuse an LU-decomposition, *cf.* Rem. 2.5.0.10.

C++ code 8.5.1.44: Efficient implementation of simplified Newton method ↪ GITLAB

```
// C++ template for simplified Newton method
template<typename Func, typename Jac, typename Vec>
void simpnewton(Vec& x, Func F, Jac DF, double rtol, double atol)
{
  auto lu = DF(x).lu(); // do LU decomposition once!
  Vec s;                // Newton correction
  double ns = NAN;      // auxiliary variables for termination control
  double nx = NAN;
  do {
    s = lu.solve(F(x));
    x = x-s;            // new iterate
    ns = s.norm(); nx = x.norm();
  }
  // termination based on relative and absolute tolerance
  while((ns > rtol*nx) && (ns > atol));
```

```
17 }
```

Drawback: Switching to the simplified Newton method usually sacrifices the asymptotic quadratic convergence of the Newton method: merely linear convergence can be expected.

⌟

Remark 8.5.1.45 (Numerical Differentiation for computation of Jacobian) If $\mathrm{D}F(\mathbf{x})$ is not available (e.g. when $F(\mathbf{x})$ is given only as a procedure) we may resort to approximation by difference quotients:

$$\text{Numerical Differentiation:} \quad \frac{\partial F_i}{\partial x_j}(\mathbf{x}) \approx \frac{F_i(\mathbf{x}+h\vec{e}_j) - F_i(\mathbf{x})}{h} \,.$$

Caution: Roundoff errors wreak havoc for small h → Ex. 1.5.4.7 **!** Therefore use $h \approx \sqrt{\mathrm{EPS}}$. ⌟

Review question(s) 8.5.1.46 (Newton's iteration in $\mathbb{R}^n$)

(Q8.5.1.46.A) In Ex. 7.4.2.2 we found the following non-linear system of equations for the weights w_1, w_2 and nodes c_1, c_2 of a quadrature rule of order 4 on $[-1,1]$:

$$\begin{aligned} w_1 + w_2 &= 2 \,, \\ c_1 w_1 + c_2 w_2 &= 0 \,, \\ c_1^2 w_1 + c_2^2 w_2 &= \tfrac{2}{3} \,, \\ c_1^3 w_1 + c_2^3 w_2 &= 0 \,. \end{aligned}$$

Write down a function $F : \mathbb{R}^4 \to \mathbb{R}^4$ such that this non-linear system corresponds to $F(\mathbf{x}) = \mathbf{0}$, $\mathbf{x} = [w_1, w_2, c_1, c_2]^\top$ and then derive the corresponding Newton iteration.

(Q8.5.1.46.B) Consider the following non-linear interpolation problem for a twice continuously differentiable function $f : [-1,1] \to \mathbb{R}$. Seek a node set $\{t_0, \ldots, t_n\} \subset [-1,1]$ and a polynomial $p \in \mathcal{P}_n$ such that

$$p(t_k) = f(t_k) \quad , \quad p'(t_k) = f'(t_k) \quad \forall k \in \{0, \ldots, n\} \,.$$

Recast this problem as a non-linear system of equations for the nodes t_k, $k = 0, \ldots, n$, and the $n+1$ monomial coefficients of p and then state the corresponding Newton iteration.

(Q8.5.1.46.C) For a symmetric positive definite matrix $\mathbf{A} = \mathbf{A}^\top \in \mathbb{R}^{n,n}$ derive the Newton iteration for solving $F(\mathbf{X}) = \mathbf{O}$, where

$$F : \mathrm{Sym}(n) \to \mathrm{Sym}(n) \quad , \quad F(\mathbf{X}) := \mathbf{X}^2 - \mathbf{A} \,,$$

and $\mathrm{Sym}(n)$ stands for the vector space of symmetric $n \times n$-matrices.

(Q8.5.1.46.D) In order to find a global minimum of a twice differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, we could first try determine vectors $\mathbf{x}^* \in \mathbb{R}^n$ such that $\mathbf{grad}\, f(\mathbf{x}^*) = \mathbf{0}$. What is the Newton iteration for this $n \times n$ non-linear system of equations?

(Q8.5.1.46.E) Is the stopping rule implemented in the following code affine invariant?

C++ code 8.5.1.9: Newton's method in C++ ➽ GITLAB

```
2 template <typename FuncType, typename JacType, typename VecType>
```

```
3  VecType newton(FuncType &&F, JacType &&DFinv, VecType x, const double rtol,
4                 const double atol) {
5    // Note that the vector x passes both the initial guess and also
6    // contains the iterates
7    VecType s(x.size()); // Vector for Newton corrections
8    // Main loop
9    do {
10     s = DFinv(x, F(x)); // compute Newton correction
11     x -= s;             // compute next iterate
12   }
13   // correction based termination (relative and absolute)
14   while ((s.norm() > rtol * x.norm()) && (s.norm() > atol));
15   return x;
16 }
```

△

8.5.2 Convergence of Newton's Method

Video tutorial for Section 8.5.2 "Convergence of Newton's Method": (9 minutes)
Download link, tablet notes

Notice that the Newton iteration (8.5.1.6) is a fixed point iteration (→ Section 8.3) with iteration function

$$\Phi(\mathbf{x}) = \mathbf{x} - \mathrm{D}F(\mathbf{x})^{-1}F(\mathbf{x})\,.$$

["product rule" (8.5.1.17) : $\mathrm{D}\Phi(\mathbf{x}) = \mathbf{I} - \mathrm{D}(\{\mathbf{x}\mapsto \mathrm{D}F(\mathbf{x})^{-1}\})F(\mathbf{x}) - \mathrm{D}F(\mathbf{x})^{-1}\,\mathrm{D}F(\mathbf{x})$]

$$F(\mathbf{x}^*) = \mathbf{0} \quad\Rightarrow\quad \mathrm{D}\Phi(\mathbf{x}^*) = \mathbf{O}\,,$$

that is, the derivative (Jacobian) of the iteration function of the Newton fixed point iteration vanishes in the limit point. Thus from Lemma 8.3.2.15 we draw the same conclusion as in the scalar case $n = 1$, *cf.* Section 8.4.2.1.

Local quadratic convergence of Newton's method, if $\mathrm{D}F(\mathbf{x}^*)$ regular

This can easily be seen by the following formal argument valid for $F \in C^2(D)$ with $F(\mathbf{x}^*) \neq \mathbf{0}$.

$$\begin{aligned}\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\right\| &= \left\|\Phi(\mathbf{x}^{(k)}) - \Phi(\mathbf{x}^*)\right\| \\ &= \left\|\mathrm{D}\Phi(\mathbf{x}^*)(\mathbf{x}^{(k)} - \mathbf{x}^*) + O(\left\|\mathbf{x}^{(k)} - \mathbf{x}^*\right\|^2)\right\| = O(\left\|\mathbf{x}^{(k)} - \mathbf{x}^*\right\|^2) \quad \text{for} \quad \mathbf{x}^{(k)} \approx \mathbf{x}^*\,.\end{aligned}$$

EXPERIMENT 8.5.2.1 (Convergence of Newton's method in 2D) We study the convergence of Newton's method empirically for $n = 2$ for

$$F(\mathbf{x}) = \begin{bmatrix} x_1^2 - x_2^4 \\ x_1 - x_2^3 \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \in \mathbb{R}^2 \quad \text{with solution} \quad F(\begin{bmatrix} 1 \\ 1 \end{bmatrix}) = 0\,. \tag{8.5.2.2}$$

Jacobian (analytic computation): $\mathrm{D}F(\mathbf{x}) = \begin{bmatrix} \partial_{x_1}F_1(x) & \partial_{x_2}F_1(x) \\ \partial_{x_1}F_2(x) & \partial_{x_2}F_2(x) \end{bmatrix} = \begin{bmatrix} 2x_1 & -4x_2^3 \\ 1 & -3x_2^2 \end{bmatrix}$

Realization of Newton iteration (8.5.1.6):

1. Solve LSE
$$\begin{bmatrix} 2x_1 & -4x_2^3 \\ 1 & -3x_2^2 \end{bmatrix} \Delta\mathbf{x}^{(k)} = -\begin{bmatrix} x_1^2 - x_2^4 \\ x_1 - x_2^3 \end{bmatrix},$$
where $\mathbf{x}^{(k)} = [x_1, x_2]^T$.
2. Set $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \Delta\mathbf{x}^{(k)}$.

Monitoring the iteration we obtain the following iterates/error norms:

k	$\mathbf{x}^{(k)}$	$\epsilon_k := \|\mathbf{x}^* - \mathbf{x}^{(k)}\|_2$	$\frac{\log\epsilon_{k+1} - \log\epsilon_k}{\log\epsilon_k - \log\epsilon_{k-1}}$
0	$[0.7,\ 0.7]^T$	4.24e-01	
1	$[0.87850000000000,\ 1.064285714285714]^T$	1.37e-01	1.69
2	$[1.01815943274188,\ 1.00914882463936]^T$	2.03e-02	2.23
3	$[1.00023355916300,\ 1.00015913936075]^T$	2.83e-04	2.15
4	$[1.00000000583852,\ 1.00000002726552]^T$	2.79e-08	1.77
5	$[0.999999999999998,\ 1.000000000000000]^T$	2.11e-15	
6	$[1,\ 1]^T$		

☞ (Some) evidence of quadratic convergence, see Rem. 8.2.2.12. ⌟

EXAMPLE 8.5.2.3 (Convergence of Newton's method for matrix inversion → [Ale12; PS91]) in Ex. 8.5.1.38 we have derived the Newton iteration for
$$F(\mathbf{X}) = \mathbf{O} \quad \text{with} \quad F : \mathbb{R}^{n,n} \to \mathbb{R}^{n,n}, \quad F(\mathbf{X}) := \mathbf{A} - \mathbf{X}^{-1},$$
for a given regular matrix $\mathbf{A} \in \mathbb{R}^{n,n}$:
$$\mathbf{X}^{(k+1)} = \mathbf{X}^{(k)} - \left(\mathbf{X}^{(k)}\mathbf{A}\mathbf{X}^{(k)} - \mathbf{X}^{(k)}\right) = \mathbf{X}^{(k)}\left(2\mathbf{I} - \mathbf{A}\mathbf{X}^{(k)}\right). \tag{8.5.1.42}$$

Now we study the local convergence of this iteration by direct estimates. To that end we first derive a recursion for the iteration errors $\mathbf{E}^{(k)} := \mathbf{X}^{(k)} - \mathbf{A}^{-1}$:
$$\begin{aligned} \mathbf{E}^{(k+1)} &= \mathbf{X}^{(k+1)} - \mathbf{A}^{-1} \\ &\overset{(8.5.1.42)}{=} \mathbf{X}^{(k)}\left(2\mathbf{I} - \mathbf{A}\mathbf{X}^{(k)}\right) - \mathbf{A}^{-1} \\ &= (\mathbf{E}^{(k)} + \mathbf{A}^{-1})\left(2\mathbf{I} - \mathbf{A}(\mathbf{E}^{(k)} + \mathbf{A}^{-1})\right) - \mathbf{A}^{-1} \\ &= (\mathbf{E}^{(k)} + \mathbf{A}^{-1})(\mathbf{I} - \mathbf{A}\mathbf{E}^{(k)}) - \mathbf{A}^{-1} = -\mathbf{E}^{(k)}\mathbf{A}\mathbf{E}^{(k)}. \end{aligned}$$

For the norm of the iteration error (a matrix norm → Def. 1.5.5.10) we conclude from submultiplicativity (1.5.5.11) a recursive estimate

► $$\left\|\mathbf{E}^{(k+1)}\right\| \le \left\|\mathbf{E}^{(k)}\right\|^2 \|\mathbf{A}\|. \tag{8.5.2.4}$$

This holds for any matrix norm according to Def. 1.5.5.10, which is induced by a vector norm. For the relative iteration error we obtain
$$\underbrace{\frac{\left\|\mathbf{E}^{(k+1)}\right\|}{\|\mathbf{A}\|}}_{\text{relative error}} \le \left(\underbrace{\frac{\left\|\mathbf{E}^{(k)}\right\|}{\|\mathbf{A}\|}}_{\text{relative error}}\right)^2 \underbrace{\|\mathbf{A}\|\left\|\mathbf{A}^{-1}\right\|}_{=\mathrm{cond}(\mathbf{A})}, \tag{8.5.2.5}$$

where the condition number is defined in Def. 2.2.2.7.

From (8.5.2.4) we conclude that the iteration will converge ($\lim_{k\to\infty}\left\|\mathbf{E}^{(k)}\right\| = 0$), if

$$\left\|\mathbf{E}^{(0)}\mathbf{A}\right\| = \left\|\mathbf{X}^{(0)}\mathbf{A} - \mathbf{I}\right\| < 1\,, \tag{8.5.2.6}$$

which gives a condition on the initial guess $\mathbf{S}^{(0)}$. Now let us consider the Euclidean matrix norm $\|\cdot\|_2$, which can be expressed in terms of eigenvalues, see Cor. 1.5.5.16. Motivated by this relationship, we use the initial guess $\mathbf{X}^{(0)} = \alpha\mathbf{A}^\top$ with $a > 0$ still to be determined.

$$\left\|\mathbf{X}^{(0)}\mathbf{A} - \mathbf{I}\right\|_2 = \left\|\alpha\mathbf{A}^\top\mathbf{A} - \mathbf{I}\right\|_2 = \alpha\|\mathbf{A}\|_2^2 - 1 \overset{!}{<} 1 \quad\Leftrightarrow\quad \alpha < \frac{2}{\|\mathbf{A}\|_2^2}\,,$$

which is a sufficient condition for the initial guess $\mathbf{X}^{(0)} = \alpha\mathbf{A}^\top$, in order to make (8.5.1.42) converge. In this case we infer quadratic convergence from both (8.5.2.4) and (8.5.2.5). ⌟

There is a sophisticated theory about the convergence of Newton's method. For example one can find the following theorem in [DH03, Thm. 4.10], [Deu11, Sect. 2.1]):

Theorem 8.5.2.7. Local quadratic convergence of Newton's method

If:
- *(A)* $D \subset \mathbb{R}^n$ *open and convex,*
- *(B)* $F : D \mapsto \mathbb{R}^n$ *continuously differentiable,*
- *(C)* $\mathrm{D}F(\mathbf{x})$ *regular* $\forall \mathbf{x} \in D$,
- *(D)* $\exists L \geq 0$: $\left\|\mathrm{D}F(\mathbf{x})^{-1}(\mathrm{D}F(\mathbf{x}+\mathbf{v}) - \mathrm{D}F(\mathbf{x}))\right\|_2 \leq L\|\mathbf{v}\|_2 \quad \begin{array}{l}\forall \mathbf{v} \in \mathbb{R}^n, \mathbf{v}+\mathbf{x} \in D,\\ \forall \mathbf{x} \in D\end{array}$,
- *(E)* $\exists \mathbf{x}^*$: $F(\mathbf{x}^*) = 0$ *(existence of solution in* D*)*
- *(F) initial guess* $\mathbf{x}^{(0)} \in D$ *satisfies* $\rho := \left\|\mathbf{x}^* - \mathbf{x}^{(0)}\right\|_2 < \frac{2}{L} \;\wedge\; B_\rho(\mathbf{x}^*) \subset D$.

then *the Newton iteration* (8.5.1.6) *satisfies:*
- *(i)* $\mathbf{x}^{(k)} \in B_\rho(\mathbf{x}^*) := \{\mathbf{y} \in \mathbb{R}^n, \|\mathbf{y} - \mathbf{x}^*\| < \rho\}$ *for all* $k \in \mathbb{N}$,
- *(ii)* $\lim_{k\to\infty} \mathbf{x}^{(k)} = \mathbf{x}^*$,
- *(iii)* $\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\right\|_2 \leq \frac{L}{2}\left\|\mathbf{x}^{(k)} - \mathbf{x}^*\right\|_2^2$ *(local quadratic convergence) .*

✎ notation: ball $B_\rho(\mathbf{z}) := \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x} - \mathbf{z}\|_2 \leq \rho\}$

Terminology: (D) ≙ affine invariant Lipschitz condition

Usually, it is hardly possible to verify the assumptions of the theorem for a concrete non-linear system of equations, because neither L nor x^* are known.

In general: a priori estimates as in Thm. 8.5.2.7 are of little practical relevance.

Review question(s) 8.5.2.8 (Convergence of Newton's method)

(Q8.5.2.8.A) Outline how you would empirically investigate the *quadratic convergence* of a Newton iteration

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \mathrm{D}F(\mathbf{x}^{(k)})^{-1}F(\mathbf{x}^{(k)}) \tag{8.5.1.6}$$

for solving $F(\mathbf{x}) = \mathbf{0}$.

(Q8.5.2.8.B) Indentifying $\mathbb{C} \simeq \mathbb{R}^2$ we can consider the equation $z^2 = 1$ as a 2×2 non-linear system of equations.

Now please take a look at the convergence theorem Thm. 8.5.2.7 for Newton's method in $\mathbb{R}^n$. Verify the assumptions of this theorem for $D := \{z \in \mathbb{C} : |z - 1| < \frac{1}{2}\}$ and the function $F : \mathbb{R}^2 \to \mathbb{R}^2$ for which solving $F(\mathbf{x}) = \mathbf{0}$ in $\mathbb{R}^2$ is equivalent to solving $z^2 = 1$ in $\mathbb{C}$. Estimate the Lipschitz constant L.

△

8.5.3 Termination of Newton Iteration

Video tutorial for Section 8.5.3 "Termination of Newton Iteration": (7 minutes) Download link, tablet notes

An abstract discussion of ways to stop iterations for solving $F(\mathbf{x}) = \mathbf{0}$ was presented in Section 8.2.3, with "ideal termination" (→ § 8.2.3.2) as ultimate, but unfeasible, goal.

Yet, in 8.5.2 we saw that Newton's method enjoys (asymptotic) quadratic convergence, which means rapid decrease of the relative error of the iterates, once we are close to the solution, which is exactly the point, when we want to **STOP**. As a consequence, asymptotically, the Newton correction (difference of two consecutive iterates) yields rather precise information about the size of the error:

$$\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\right\| \ll \left\|\mathbf{x}^{(k)} - \mathbf{x}^*\right\| \quad \Rightarrow \quad \left\|\mathbf{x}^{(k)} - \mathbf{x}^*\right\| \approx \left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\| . \tag{8.5.3.1}$$

This suggests the following correction based termination criterion:

$$\begin{aligned} &\textbf{STOP}, \text{ as soon as } \quad \left\|\Delta\mathbf{x}^{(k)}\right\| \leq \tau_{\mathrm{rel}} \left\|\mathbf{x}^{(k)}\right\| \quad \textbf{or} \quad \left\|\Delta\mathbf{x}^{(k)}\right\| \leq \tau_{\mathrm{abs}} , \\ &\text{with} \quad \text{Newton correction} \quad \Delta\mathbf{x}^{(k)} := \mathrm{D}F(\mathbf{x}^{(k)})^{-1} F(\mathbf{x}^{(k)}). \end{aligned} \tag{8.5.3.2}$$

Here, $\|\cdot\|$ can be any suitable vector norm, $\tau_{\mathrm{rel}} \mathrel{\hat{=}}$ relative tolerance, $\tau_{\mathrm{abs}} \mathrel{\hat{=}}$ absolute tolerance, see § 8.2.3.2.

➢ quit iterating as soon as $\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\| = \left\|\mathrm{D}F(\mathbf{x}^{(k)})^{-1} F(\mathbf{x}^{(k)})\right\| < \tau \left\|\mathbf{x}^{(k)}\right\|$,
with τ = tolerance

→ uneconomical: one needless update, because $\mathbf{x}^{(k)}$ would already be accurate enough.

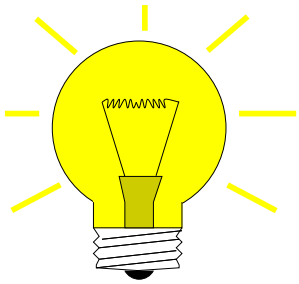
Remark 8.5.3.3 (Newton's iteration; computational effort and termination) Some facts about the Newton method for solving large ($n \gg 1$) non-linear systems of equations:

- ☛ Solving the linear system to compute the Newton correction may be expensive (asymptotic computational effort $O(n^3)$ for direct elimination → § 2.3.1.5) and accounts for the bulk of numerical cost of a single step of the iteration.
- ☛ In applications only very few steps of the iteration will be needed to achieve the desired accuracy due to fast quadratic convergence.

▷ The termination criterion (8.5.3.2) computes the last Newton correction $\Delta\mathbf{x}^{(k)}$ needlessly, because $\mathbf{x}^{(k)}$ already accurate enough**!**

Therefore we would like to use an a-posteriori termination criterion that dispenses with computing (and "inverting") another Jacobian $\mathrm{D}F(\mathbf{x}^{(k)})$ just to tell us that $\mathbf{x}^{(k)}$ is already accurate enough. ⌟

§8.5.3.4 (Termination of Newton iteration based on simplified Newton correction) Due to fast asymptotic quadratic convergence, we can expect $\mathrm{D}F(\mathbf{x}^{(k-1)}) \approx \mathrm{D}F(\mathbf{x}^{(k)})$ during the final steps of the iteration.



Idea: Replace $\mathrm{D}F(\mathbf{x}^{(k)})$ with $\mathrm{D}F(\mathbf{x}^{(k-1)})$ in any correction based termination criterion.

Rationale: LU-decomposition of $\mathrm{D}F(\mathbf{x}^{(k-1)})$ is already available
➤ reduced effort.

Terminology: $\Delta\bar{\mathbf{x}}^{(k)} := \mathrm{D}F(\mathbf{x}^{(k-1)})^{-1}F(\mathbf{x}^{(k)}) \triangleq$ simplified Newton correction

Economical correction based termination criterion for Newton's method:

STOP, as soon as $\left\|\Delta\bar{\mathbf{x}}^{(k)}\right\| \leq \tau_{\mathrm{rel}}\left\|\mathbf{x}^{(k)}\right\|$ **or** $\left\|\Delta\bar{\mathbf{x}}^{(k)}\right\| \leq \tau_{\mathrm{abs}}$,

with simplified Newton correction $\Delta\bar{\mathbf{x}}^{(k)} := \mathrm{D}F(\mathbf{x}^{(k-1)})^{-1}F(\mathbf{x}^{(k)})$. (8.5.3.5)

Note that (8.5.3.5) is affine invariant → Rem. 8.5.1.11.

Effort: Reuse of LU-factorization (→ Rem. 2.5.0.10) of $\mathrm{D}F(\mathbf{x}^{(k-1)})$ ➤ $\Delta\bar{\mathbf{x}}^{(k)}$ available with $O(n^2)$ operations

C++11 code 8.5.3.6: Generic Newton iteration with termination criterion (8.5.3.5) ↪ GITLAB

```
template <typename FuncType, typename JacType, typename VecType>
void newton_stc(const FuncType &F, const JacType &DF, VecType &x, double rtol,
                double atol) {
  using scalar_t = typename VecType::Scalar;
  scalar_t sn;
  do {
    auto jacfac = DF(x).lu(); // LU-factorize Jacobian ]
    x -= jacfac.solve(F(x)); // Compute next iterate
    // Compute norm of simplified Newton correction
    sn = jacfac.solve(F(x)).norm();
  }
  // Termination based on simplified Newton correction
  while ((sn > rtol * x.norm()) && (sn > atol));
}
```

⌟

Remark 8.5.3.7 (Residual based termination of Newton's method) If we used the residual based termination criterion

$$\left\|F(\mathbf{x}^{(k)})\right\| \leq \tau\,,$$

then the resulting algorithm would not be affine invariant, because for $F(\mathbf{x}) = \mathbf{0}$ and $\mathbf{A}F(\mathbf{x}) = \mathbf{0}$, $\mathbf{A} \in \mathbb{R}^{n,n}$ regular, the Newton iteration might terminate with different iterates. ⌟

Summary: Newton's method

converges *asymptotically* very fast: doubling of number of significant digits in each step

often a very small region of convergence, which requires an initial guess rather close to the solution.

Review question(s) 8.5.3.9 (Termination of Newton's iteration)

(Q8.5.3.9.A) Show that the following stopping rule for Newton's method,

$$\text{STOP, as soon as } \left\|\Delta\bar{\mathbf{x}}^{(k)}\right\| \leq \tau_{\text{rel}}\left\|\mathbf{x}^{(k)}\right\| \quad \text{or} \quad \left\|\Delta\bar{\mathbf{x}}^{(k)}\right\| \leq \tau_{\text{abs}}\,, \tag{8.5.3.5}$$

$$\text{with simplified Newton correction } \Delta\bar{\mathbf{x}}^{(k)} := \mathrm{D}\,F(\mathbf{x}^{(k-1)})^{-1}F(\mathbf{x}^{(k)}),$$

is **affine invariant**.

Recall that an implementation of Newton's method (including stopping rules) for solving $G_{\mathbf{A}}(\mathbf{x}) = \mathbf{0}$, $G_{\mathbf{A}}(\mathbf{x}) := \mathbf{A}F(\mathbf{x})$, $F : D \subset \mathbb{R}^n \to \mathbb{R}^n$, is called affine invariant, if the same sequence of iterates is produced for every regular matrix $\mathbf{A} \in \mathbb{R}^{n,n}$.

△

8.5.4 Damped Newton Method

Video tutorial for Section 8.5.4 "Damped Newton Method": (11 minutes) Download link, tablet notes

Potentially big problem: Newton method converges quadratically, but only *locally*, which may render it useless, if convergence is guaranteed only for initial guesses very close to exact solution, see also Ex. 8.4.2.38.

In this section we study a method to enlarge the region of convergence, at the expense of quadratic convergence, of course.

EXAMPLE 8.5.4.1 (Local convergence of Newton's method) The dark side of local convergence (→ Def. 8.2.1.10): for many initial guesses $\mathbf{x}^{(0)}$ Newton's method will not converge!

In 1D two main causes can be identified:

❶ "Wrong direction" of Newton correction:

$$F(x) = xe^x - 1 \quad \Rightarrow \quad F'(-1) = 0$$

$$x^{(0)} < -1 \quad \Rightarrow \quad x^{(k)} \to -\infty\,,$$
$$x^{(0)} > -1 \quad \Rightarrow \quad x^{(k)} \to x^*\,,$$

because all Newton corrections for $x^{(k)} < -1$ make the iterates decrease even further.

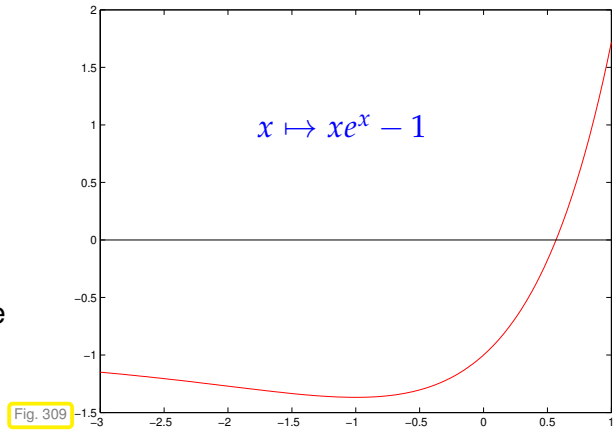


Fig. 309

❷ Newton correction is too large:

$$F(x) = \arctan(ax)\ , \quad a > 0, x \in \mathbb{R}$$

with zero $x^* = 0$.

If $x^{(k)}$ is located where the function is "flat", the intersection of the tangents with the x-axis is "far out", see Fig. 311.

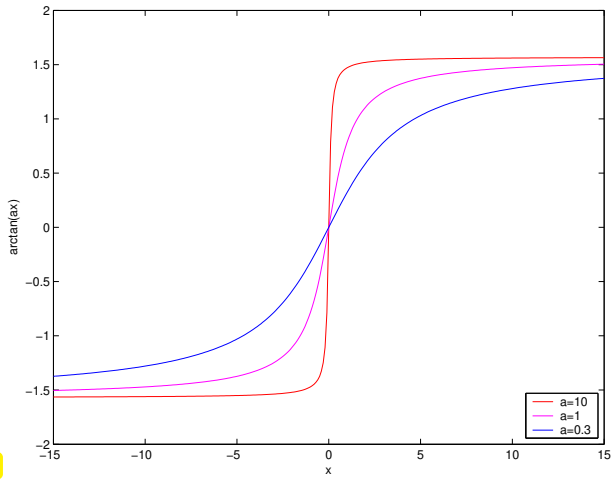


Fig. 310

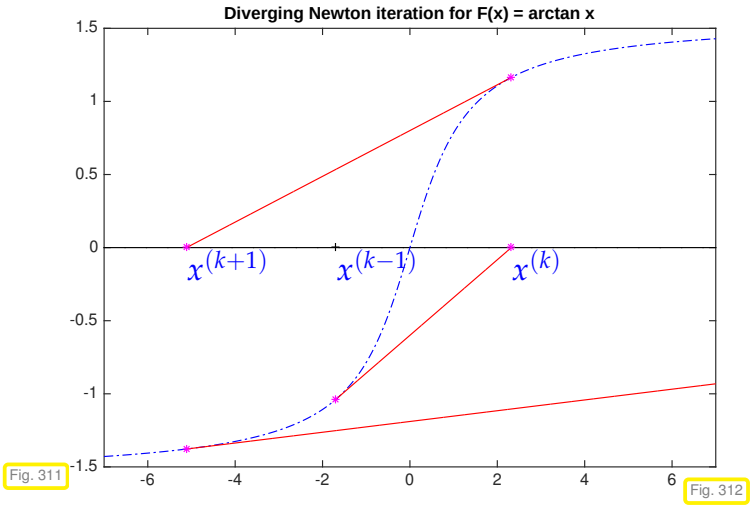


Fig. 311

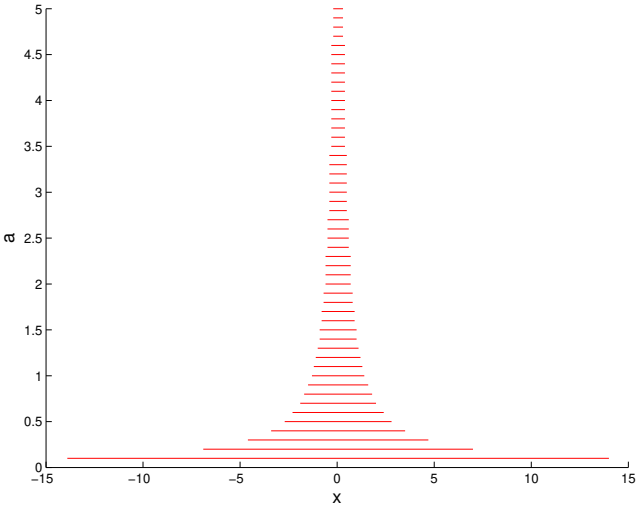
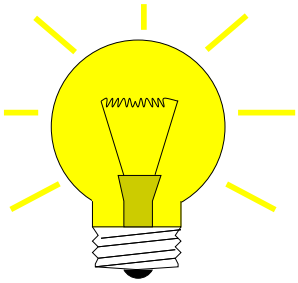


Fig. 312

In Fig. 312 the red zone = $\{x^{(0)} \in \mathbb{R}, x^{(k)} \to 0\}$, domain of initial guesses for which Newton's method converges.

⌟

If the Newton correction points in the wrong direction (Item ❶), no general remedy is available. If the Newton correction is too large (Item ❷), there is an effective cure:



we observe "overshooting" of Newton correction

Idea: damping of Newton correction:

With $\lambda^{(k)} > 0$: $\mathbf{x}^{(k+1)} := \mathbf{x}^{(k)} - \lambda^{(k)} \mathrm{D}F(\mathbf{x}^{(k)})^{-1} F(\mathbf{x}^{(k)})$. (8.5.4.2)

Terminology: $\lambda^{(k)}$ = damping factor, $\lambda \in]0,1]$.

Affine invariant damping strategy

Choice of damping factor: affine invariant natural monotonicity test (NMT) [Deu11, Ch. 3]:

choose "maximal" $0 < \lambda^{(k)} \leq 1$: $\left\| \Delta\bar{\mathbf{x}}(\lambda^{(k)}) \right\|_2 \leq (1 - \frac{\lambda^{(k)}}{2}) \left\| \Delta\mathbf{x}^{(k)} \right\|_2$ (8.5.4.4)

where
$$\Delta\mathbf{x}^{(k)} := \mathrm{D}F(\mathbf{x}^{(k)})^{-1}F(\mathbf{x}^{(k)}) \quad \rightarrow \text{current Newton correction ,}$$
$$\Delta\bar{\mathbf{x}}(\lambda^{(k)}) := \mathrm{D}F(\mathbf{x}^{(k)})^{-1}F(\mathbf{x}^{(k)} - \lambda^{(k)}\Delta\mathbf{x}^{(k)}) \quad \rightarrow \text{tentative simplified Newton correction .}$$

Heuristics behind control of damping:

- When the method converges $\Leftrightarrow$ size of Newton correction decreases $\Leftrightarrow$ (8.5.4.4) satisfied.
- In the case of strong damping ($\lambda^{(k)} \ll 1$) the size of the Newton correction cannot be expected to shrink significantly, since iterates do not change much $\succ$ factor $(1 - \frac{1}{2}\lambda^{(k)})$ in (8.5.4.4).

Note: As before, reuse of LU-factorization in the computation of $\Delta\mathbf{x}^{(k)}$ and $\Delta\bar{\mathbf{x}}(\lambda^{(k)})$.

Policy: Reduce damping factor by a factor $q \in]0,1[$ (usually $q = \frac{1}{2}$) until the affine invariant natural monotonicity test (8.5.4.4) passed, see Line 20 in the following C++ code.

C++ code 8.5.4.5: Generic damped Newton method based on natural monotonicity test ↪ GITLAB

```
1  template <typename FuncType,typename JacType,typename VecType>
2  void dampnewton(const FuncType &F,const JacType &DF,
3                  VecType &x,double rtol,double atol)
4  {
5    using index_t = typename VecType::Index;
6    using scalar_t = typename VecType::Scalar;
7    const index_t n = x.size(); // No.  of unknowns
8    const scalar_t lmin = 1E-3; // Minimal damping factor
9    scalar_t lambda = 1.0; // Initial and actual damping factor
10   VecType s(n),st(n); // Newton corrections
11   VecType xn(n);      // Tentative new iterate
12   scalar_t sn,stn;    // Norms of Newton corrections
13
14   do {
15     auto jacfac = DF(x).lu(); // LU-factorize Jacobian
16     s = jacfac.solve(F(x));   // Newton correction
17     sn = s.norm();            // Norm of Newton correction
18     lambda *= 2.0;
19     do {
20       lambda /= 2; // Reduce damping factor
21       if (lambda < lmin) throw "No convergence: lambda -> 0";
22       xn = x-lambda*s;           // Tentative next iterate
23       st = jacfac.solve(F(xn)); // Simplified Newton correction
24       stn = st.norm();
25     }
26     while (stn > (1-lambda/2)*sn); // Natural monotonicity test
27     x = xn; // Now:  xn accepted as new iterate
28     lambda = std::min(2.0*lambda,1.0); // Try to mitigate damping
29   }
30   // Termination based on simplified Newton correction
31   while ((stn > rtol*x.norm()) && (stn > atol));
32 }
```

The arguments for Code 8.5.4.5 are the same as for Code 8.5.3.6. As termination criterion is uses (8.5.3.5). Note that all calls to **solve** boil down to forward/backward elimination for triangular matrices and incur cost of $O(n^2)$ only.

Note: The LU-factorization of the Jacobi matrix $\mathrm{D}F(\mathbf{x}^{(k)})$ is done once per *successful* iteration step and

reused for the computation of the simplified Newton correction in Line 23 of the above C++ code.

EXPERIMENT 8.5.4.6 (Damped Newton method) We test the damped Newton method for Item ❷ of Ex. 8.5.4.1, where excessive Newton corrections made Newton's method fail.

$F(x) = \arctan(x)$,
- $x^{(0)} = 20$
- $q = \frac{1}{2}$
- `LMIN = 0.001`

We observe that damping is effective and asymptotic quadratic convergence is recovered.

k	$\lambda^{(k)}$	$x^{(k)}$	$F(x^{(k)})$
1	0.03125	0.94199967624205	0.75554074974604
2	0.06250	0.85287592931991	0.70616132170387
3	0.12500	0.70039827977515	0.61099321623952
4	0.25000	0.47271811131169	0.44158487422833
5	0.50000	0.20258686348037	0.19988168667351
6	1.00000	-0.00549825489514	-0.00549819949059
7	1.00000	0.00000011081045	0.00000011081045
8	1.00000	-0.00000000000001	-0.00000000000001

EXPERIMENT 8.5.4.7 (Failure of damped Newton method) We examine the effect of damping in the case of Item ❶ of Ex. 8.5.4.1.

- As in Ex. 8.5.4.1: $F(x) = xe^x - 1$,
- Initial guess for damped Newton method: $x^{(0)} = -1.5$

This time the initial guess is to the left of the global minimum of the function.

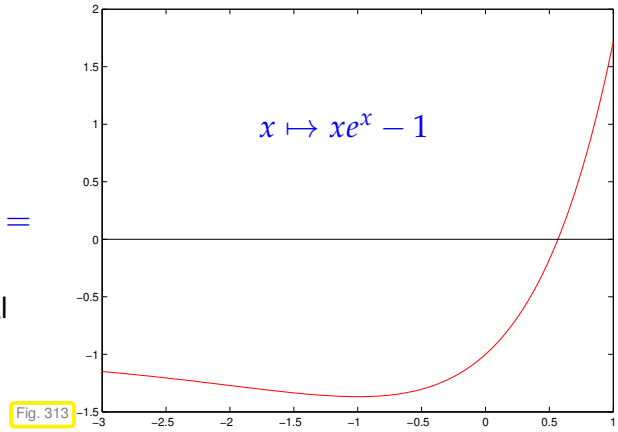


Fig. 313

Observation:

Newton correction pointing in "wrong direction"

➤ no convergence despite damping

k	$\lambda^{(k)}$	$x^{(k)}$	$F(x^{(k)})$
1	0.25000	-4.4908445351690	-1.0503476286303
2	0.06250	-6.1682249558799	-1.0129221310944
3	0.01562	-7.6300006580712	-1.0037055902301
4	0.00390	-8.8476436930246	-1.0012715832278
5	0.00195	-10.5815494437311	-1.0002685596314
Bailed out because of `lambda < LMIN` !			

Review question(s) 8.5.4.8 (Damped Newton method)

(Q8.5.4.8.A) Give an example of a one-dimensional zero-finding problem, where a suitable damping strategy can ensure global convergence of Newton's method.

(Q8.5.4.8.B) Show that the affine invariant damping strategy for Newton's method

Choice of damping factor: affine invariant natural monotonicity test (NMT) [Deu11, Ch. 3]:

choose "maximal" $0 < \lambda^{(k)} \leq 1$: $$\left\| \Delta\bar{\mathbf{x}}(\lambda^{(k)}) \right\|_2 \leq (1 - \frac{\lambda^{(k)}}{2}) \left\| \Delta\mathbf{x}^{(k)} \right\|_2 \tag{8.5.4.4}$$

where
$$\Delta\mathbf{x}^{(k)} := \mathrm{D}F(\mathbf{x}^{(k)})^{-1}F(\mathbf{x}^{(k)}) \quad \rightarrow \text{ current Newton correction ,}$$
$$\Delta\bar{\mathbf{x}}(\lambda^{(k)}) := \mathrm{D}F(\mathbf{x}^{(k)})^{-1}F(\mathbf{x}^{(k)} - \lambda^{(k)}\Delta\mathbf{x}^{(k)}) \quad \rightarrow \text{ tentative simplified Newton correction .}$$

leads to an **affine invariant** iteration.

△

8.6 Quasi-Newton Method

Video tutorial for Section 8.6 "Quasi-Newton Method": (15 minutes) Download link, tablet notes

We start with the following question: How can we solve the non-linear system of equations $F(\mathbf{x}) = \mathbf{0}$, $D : D \subset \mathbb{R}^n \to \mathbb{R}^n$, iteratively, in case $\mathrm{D}F(\mathbf{x})$ is not available and numerical differentiation (see Rem. 8.5.1.45) is too expensive**?**

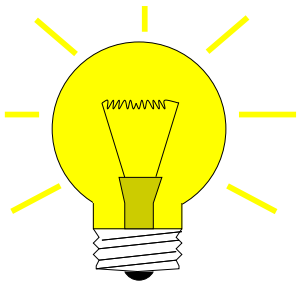
In 1D ($n = 1$) we can choose among many **derivative-free** methods that rely on F-evaluations alone, for instance the secant method (8.4.2.30) from Section 8.4.2.3:

$$x^{(k+1)} = x^{(k)} - \frac{F(x^{(k)})(x^{(k)} - x^{(k-1)})}{F(x^{(k)}) - F(x^{(k-1)})} = x^{(k)} - \left(\frac{F(x^{(k)}) - f(x^{(k-1)})}{x^{(k)} - x^{(k-1)}} \right)^{-1} F(x^{(k)}) . \tag{8.4.2.30}$$

Recall from Rem. 8.4.2.33 that the secant method converges locally with order $p \approx 1.6$ and beats Newton's method in terms of efficiency ($\rightarrow$ Section 8.4.3).

Compare (8.4.2.30) with Newton's method in 1D for solving $F(x) = 0$:

$$x^{(k+1)} = x^{(k)} - F'(x^{(k)})^{-1}F(x^{(k)}) . \tag{8.4.2.1}$$

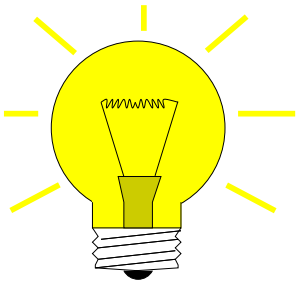


We realize that this iteration amounts to a "Newton-type iteration" based on the approximation

$$F'(x^{(k)}) \approx \frac{F(x^{(k)}) - F(x^{(k-1)})}{x^{(k)} - x^{(k-1)}} \quad \text{"difference quotient"} \tag{8.6.0.1}$$

already computed **!** $\rightarrow$ cheap

Unfortunately, it is not immediate how to generalize the secant method to $n > 1$.



Idea: Rewrite (8.6.0.1) as a **secant condition** for an approximation $\mathbf{J}_k \approx \mathrm{D}F(\mathbf{x}^{(k)})$, $\mathbf{x}^{(k)} \hat{=}$ the current iterate:

$$\mathbf{J}_k(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}) = F(\mathbf{x}^{(k)}) - F(\mathbf{x}^{(k-1)}) \,. \tag{8.6.0.2}$$

$\blacktriangleright$ Iteration: $\mathbf{x}^{(k+1)} := \mathbf{x}^{(k)} - \mathbf{J}_k^{-1} F(\mathbf{x}^{(k)})$. (8.6.0.3)

Embarrassment of choice: Many different matrices $\mathbf{J}_k$ fulfill (8.6.0.2)!

$\blacktriangleright$ We need extra conditions to fix $\mathbf{J}_k \in \mathbb{R}^{n,n}$.

Reasoning: If we assume that $\mathbf{J}_k$ is a good approximation of $\mathrm{D}F(\mathbf{x}^{(k)})$, then it would be foolish not to use the information contained in $\mathbf{J}_k$ for the construction of $\mathbf{J}_{k+1}$.

$\blacktriangleright$ Guideline: obtain $\mathbf{J}_k$ through a "small" modification of $\mathbf{J}_{k-1}$ compliant with (8.6.0.2)

What can "small modification" mean? Demand that $\mathbf{J}_k$ acts like $\mathbf{J}_{k-1}$ on the orthogonal complement of the one-dimensional subspace of $\mathbb{R}^n$ generated by the vector $\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}$! This is expressed by

$$\text{Broyden's conditions:} \quad \mathbf{J}_k \mathbf{z} = \mathbf{J}_{k-1}\mathbf{z} \quad \forall \mathbf{z}\colon \mathbf{z} \perp (\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}) \,. \tag{8.6.0.4}$$

Together with the secant condition (8.6.0.2) this uniquely determines $\mathbf{J}_k$:

(8.6.0.2), (8.6.0.4) $\blacktriangleright$

$$\mathbf{J}_k := \mathbf{J}_{k-1} + \frac{F(\mathbf{x}^{(k)})(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)})^\top}{\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|_2^2} \tag{8.6.0.5}$$

Note that the update formula (8.6.0.5) means that $\mathbf{J}_k$ is spawned by a rank-1-modification of $\mathbf{J}_{k-1}$. We have arrived at a well-defined iterative method.

Final form of **Broyden's quasi-Newton method** for solving $F(\mathbf{x}) = 0$:

$$\begin{aligned} &\mathbf{x}^{(k+1)} := \mathbf{x}^{(k)} + \Delta\mathbf{x}^{(k)} \quad , \quad \Delta\mathbf{x}^{(k)} := -\mathbf{J}_k^{-1} F(\mathbf{x}^{(k)}) \,, \\ &\mathbf{J}_{k+1} := \mathbf{J}_k + \frac{F(\mathbf{x}^{(k+1)})(\Delta\mathbf{x}^{(k)})^\top}{\|\Delta\mathbf{x}^{(k)}\|_2^2} \,, \qquad k \in \mathbb{N}_0 \,. \end{aligned} \tag{8.6.0.6}$$

To start the iteration we have to initialize $\mathbf{J}_0$, e.g. with the exact Jacobi matrix $\mathrm{D}F(\mathbf{x}^{(0)})$.

Remark 8.6.0.7 (Minimality property of Broyden's rank-1-modification) In another sense $\mathbf{J}_l$ is closest to $\mathbf{J}_{k-1}$ under the constraint of the secant condition (8.6.0.2):

Let $\mathbf{x}^{(k)}$ and $\mathbf{J}_k$ be the iterates and matrices, respectively, from Broyden's method (8.6.0.6), and let $\mathbf{J} \in \mathbb{R}^{n,n}$ satisfy the same secant condition (8.6.0.2) as $\mathbf{J}_{k+1}$:

$$\mathbf{J}(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}) = F(\mathbf{x}^{(k+1)}) - F(\mathbf{x}^{(k)}) \,. \tag{8.6.0.8}$$

Then from $\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} = -\mathbf{J}_k^{-1} F(\mathbf{x}^{(k)})$ we obtain

$$(\mathbf{I} - \mathbf{J}_k^{-1}\mathbf{J})(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}) = -\mathbf{J}_k^{-1} F(\mathbf{x}^{(k)}) - \mathbf{J}_k^{-1}(F(\mathbf{x}^{(k+1)}) - F(\mathbf{x}^{(k)})) = -\mathbf{J}_k^{-1} F(\mathbf{x}^{(k+1)}) \,. \tag{8.6.0.9}$$

From this we get the identity

$$\begin{aligned}\mathbf{I}-\mathbf{J}_k^{-1}\mathbf{J}_{k+1} &= \mathbf{I}-\mathbf{J}_k^{-1}\left(\mathbf{J}_k+\frac{F(\mathbf{x}^{(k+1)})(\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)})^\top}{\|\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}\|_2^2}\right)\\ &= -\mathbf{J}_k^{-1}F(\mathbf{x}^{(k+1)})\frac{(\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)})^\top}{\|\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}\|_2^2} =\\ &\overset{(8.6.0.9)}{=} (\mathbf{I}-\mathbf{J}_k^{-1}\mathbf{J})\frac{(\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)})(\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)})^\top}{\|\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}\|_2^2}\,.\end{aligned}$$

Using the submultiplicative property (1.5.5.11) of the Euclidean matrix norm, we conclude

$$\left\|\mathbf{I}-\mathbf{J}_k^{-1}\mathbf{J}_{k+1}\right\| \le \left\|\mathbf{I}-\mathbf{J}_k^{-1}\mathbf{J}\right\| \text{, because } \left\|\frac{(\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)})(\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)})^\top}{\|\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}\|_2^2}\right\|_2 \le 1\,,$$

which we saw in Ex. 1.5.5.20. This estimate holds for *all* matrices $\mathbf{J}$ satisfying (8.6.0.8).

We may read this as follows: (8.6.0.5) gives the $\|\cdot\|_2$-minimal relative correction of $\mathbf{J}_{k-1}$, such that the secant condition (8.6.0.2) holds. ⌟

EXPERIMENT 8.6.0.10 (Broydens quasi-Newton method: Convergence) We revisit the 2×2 non-linear system of the Exp. 8.5.2.1,

$$F(\mathbf{x}) = \begin{bmatrix} x_1^2-x_2^4 \\ x_1-x_2^3 \end{bmatrix}\,,\quad \mathbf{x}=\begin{bmatrix} x_1\\ x_2\end{bmatrix}\in\mathbb{R}^2 \quad \text{with solution} \quad F(\begin{bmatrix}1\\1\end{bmatrix})=0\,, \tag{8.5.2.2}$$

and take $\mathbf{x}^{(0)}=[0.7,0.7]^T$. As starting value for the matrix iteration we use $\mathbf{J}_0=\mathrm{D}F(\mathbf{x}^{(0)})$.

The numerical example shows that, in terms of convergence, the method is:

- slower than Newton method (8.5.1.6),
- faster than the simplified Newton method (see Rem. 8.5.1.43)

In particular, we cannot expect Broyden's quasi-Newton method to converge locally quadratically.

▷

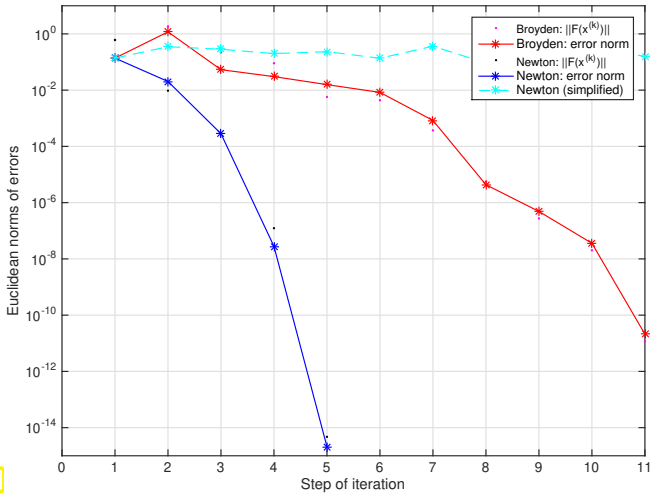


Fig. 314

⌟

Remark 8.6.0.11 (Convergence monitors) In general, any iterative methods for non-linear systems of equations convergence can fail, that is it may stall or even diverge.

Demand on good numerical software: Algorithms should warn users of impending failure. For iterative methods this is the task of convergence monitors, that is, conditions, *cheaply verifiable* a posteriori during the iteration, that indicate stalled convergence or divergence.

For the damped Newton's method this role can be played by the natural monotonicity test, see Code 8.5.4.5; if it fails repeatedly, then the iteration should terminate with an error status.

For Broyden's quasi-Newton method, a similar strategy can rely on the relative size of the "simplified Broyden corrections" $\mathbf{J}_{k-1}^{-1}F(\mathbf{x}^{(k)})$:

$$\text{Convergence monitor for (8.6.0.6) :} \qquad \mu := \frac{\left\|\mathbf{J}_{k-1}^{-1}F(\mathbf{x}^{(k)})\right\|}{\left\|\Delta\mathbf{x}^{(k-1)}\right\|} < 1 \ ? \tag{8.6.0.12}$$

⌟

EXPERIMENT 8.6.0.13 (Monitoring convergence for Broyden's quasi-Newton method)

Fig. 315

We rely on the setting of Exp. 8.6.0.10.

We track
1. the Euclidean norm of the iteration error,
2. and the value of the convergence monitor from (8.6.0.12).

◁ Decay of (norm of) iteration error and μ are well correlated.

⌟

Remark 8.6.0.14 (Damped Broyden method) Option to improve robustness (increase region of local convergence):

damped Broyden method (*cf.* same idea for Newton's method, Section 8.5.4)

⌟

§8.6.0.15 (Implementation of Broyden's quasi-Newton method) As remarked,

$$\mathbf{J}_{k+1} := \mathbf{J}_k + \frac{F(\mathbf{x}^{(k+1)})(\Delta\mathbf{x}^{(k)})^\top}{\left\|\Delta\mathbf{x}^{(k)}\right\|_2^2} \tag{8.6.0.6}$$

represents a rank-1-update of the approximate Jacobians $\mathbf{J}_k$, which are then used as coefficient matrices for $n \times n$ linear systems of equations. In § 2.6.0.12 we already discussed efficient algorithms for updating the solutions of LSEs in the case of rank-1-modifications of the coefficient matrices. Now these techniques come handy.

Idea: use the Sherman-Morrison-Woodbury update-formula from Lemma 2.6.0.21,

$$(\mathbf{A}+\mathbf{U}\mathbf{V}^{\mathrm{H}})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}(\mathbf{I}+\mathbf{V}^{\mathrm{H}}\mathbf{A}^{-1}\mathbf{U})^{-1}\mathbf{V}^{\mathrm{H}}\mathbf{A}^{-1}\,,$$

for the special case $k = 1$, $\mathbb{K} = \mathbb{R}$,

$$(\mathbf{A}+\mathbf{u}\mathbf{v}^\top)^{-1} = \left(\mathbf{I} - \frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}^\top}{1+\mathbf{v}^\top\mathbf{A}^{-1}\mathbf{u}}\right)\mathbf{A}^{-1}\,, \quad 1+\mathbf{v}^\top\mathbf{A}^{-1}\mathbf{u} \neq 0\,,$$

which yields, with $\mathbf{u} := F(\mathbf{x}^{(k+1)})$, $\mathbf{v} := \frac{\Delta\mathbf{x}^{(k)}}{\left\|\Delta\mathbf{x}^{(k)}\right\|^2}$, a *recursion* for the inverses of the approximate Jacobians,

$$\mathbf{J}_{k+1}^{-1} = \left(\mathbf{I} - \frac{\mathbf{J}_k^{-1}F(\mathbf{x}^{(k+1)})(\Delta\mathbf{x}^{(k)})^T}{\left\|\Delta\mathbf{x}^{(k)}\right\|_2^2 + (\Delta\mathbf{x}^{(k)})^\top\mathbf{J}_k^{-1}F(\mathbf{x}^{(k+1)})}\right)\mathbf{J}_k^{-1} = \left(\mathbf{I} - \frac{\Delta\overline{\mathbf{x}}^{(k+1)}(\Delta\mathbf{x}^{(k)})^T}{\left\|\Delta\mathbf{x}^{(k)}\right\|_2^2 + (\Delta\mathbf{x}^{(k)})^\top\Delta\overline{\mathbf{x}}^{(k+1)}}\right)\mathbf{J}_k^{-1}\,, \tag{8.6.0.16}$$

with the "simplified quasi-Newton correction" $\Delta\bar{\mathbf{x}}^{(k+1)} := \mathbf{J}_k^{-1}F(\mathbf{x}^{(k+1)})$. This gives a well defined $\mathbf{J}_{k+1}$, if

$$\left|(\Delta\mathbf{x}^{(k)})^\top \Delta\bar{\mathbf{x}}^{(k+1)}\right| < \left\|\Delta\mathbf{x}^{(k)}\right\|_2^2 , \tag{8.6.0.17}$$

which can be expected to hold, if the method converges and the initial guess is sufficiently close to $\mathbf{x}^*$. Note that the simplified quasi-Newton correction is also needed for the convergence monitor (8.6.0.12).

The iterated application of (8.6.0.16) pays off, if the iteration terminates after only a few steps. In particular, for large $n \gg 1$ it is not advisable to form the matrices $\mathbf{J}_k^{-1}$ (which will usually be dense in contrast to $\mathbf{J}_k$), because we can employ fast successive multiplications with rank-1-matrices ($\rightarrow$ Ex. 1.4.3.1) to apply $\mathbf{J}_k^{-1}$ to the vector $F(\mathbf{x}^{(k)})$: Let us assume that

- the quasi-Newton corrections $\Delta\mathbf{x}^{(\ell)}$, for $\ell = 0,\ldots,k-1$,
- and the simplified quasi-Newton corrections $\Delta\bar{\mathbf{x}}^{(\ell)} := \mathbf{J}_{\ell-1}^{-1}F(\mathbf{x}^{(\ell)})$ for $\ell = 1,\ldots,k$

have already been stored in the course of the iteration. Then we can compute

$$\Delta\mathbf{x}^{(k)} = -\mathbf{J}_k^{-1}F(\mathbf{x}^{(k)}) = -\prod_{\ell=0}^{k-1}\left(\mathbf{I} - \frac{\Delta\bar{\mathbf{x}}^{(\ell+1)}(\Delta\mathbf{x}^{(\ell)})^T}{\left\|\Delta\mathbf{x}^{(\ell)}\right\|_2^2 + (\Delta\mathbf{x}^{(\ell)})^\top\Delta\bar{\mathbf{x}}^{(\ell+1)}}\right)\mathbf{J}_0^{-1}F(\mathbf{x}^{(k)}) .$$

Except for a single solution of a linear system this can be implemented with simple vector arithmetic:

$$\begin{aligned}
&\mathbf{t}_0 \in \mathbb{R}^n\colon \quad \mathbf{J}_0\mathbf{t}_0 = F(\mathbf{x}^{(k)}) ,\\
&\mathbf{t}_{\ell+1} := \mathbf{t}_\ell - \Delta\bar{\mathbf{x}}^{(\ell+1)}\frac{(\Delta\mathbf{x}^{(\ell)})^T\mathbf{t}_\ell}{\left\|\mathbf{x}^{(\ell)}\right\|_2^2 + (\Delta\mathbf{x}^{(\ell)})^\top\Delta\bar{\mathbf{x}}^{(\ell+1)}} , \quad \ell = 0,\ldots,k-1 ,\\
&\Delta\mathbf{x}^{(k)} := -\mathbf{t}_k ,
\end{aligned} \tag{8.6.0.18}$$

with a computational effort $O(n^3 + nk)$ for $n \to \infty$, assuming that a standard elimination-based direct solver is used to get $\mathbf{t}_0$, recall Thm. 2.5.0.2. Based on the next iterate $\mathbf{x}^{(k+1)} := \mathbf{x}^{(k)} + \Delta\mathbf{x}^{(k)}$, we obtain $\Delta\bar{\mathbf{x}}^{(k+1)} := \mathbf{J}_k^{-1}F(\mathbf{x}^{(k+1)})$ in a similar fashion:

$$\Delta\bar{\mathbf{x}}^{(k+1)} = \prod_{\ell=0}^{k-1}\left(\mathbf{I} - \frac{\Delta\bar{\mathbf{x}}^{(\ell+1)}(\Delta\mathbf{x}^{(\ell)})^T}{\left\|\Delta\mathbf{x}^{(\ell)}\right\|_2^2 + (\Delta\mathbf{x}^{(\ell)})^\top\Delta\bar{\mathbf{x}}^{(\ell+1)}}\right)\mathbf{J}_0^{-1}F(\mathbf{x}^{(k+1)}) . \tag{8.6.0.19}$$

Thus, the cost for N steps of Broyden's quasi-Newton algorithm is asymptotically $O(n^3 + n^2N + nN^2)$ for $n, N \to \infty$, because the expensive LU-factorization of $\mathbf{J}_0$ will be carried out only once.

This is implemented in the following function `upbroyd()`, whose arguments are

- a functor object `F` implementing $F : \mathbb{R}^n \to \mathbb{R}^n$,
- the initial guess $\mathbf{x}^{(0)} \in \mathbb{R}^n$ in `x`,
- another functor object `J` providing the Jacobian $DF(\mathbf{x}) \in \mathbb{R}^{n,n}$,
- relative and absolute tolerance `reltol` and `abstol` for correction based termination as discussed in Section 8.5.3,
- the maximal number of iterations `maxit`,
- and an optional monitor object for tracking the progress of the iteration, see the explanations concerning recorder objects in § 0.3.3.4.

The implementation makes use of the following type definition:

```
template <typename T, int N> using Vector = Eigen::Matrix<T, N, 1>;
```

The function is templated to allow its use for both fixed-size and variable size vector types of EIGEN.

C++ code 8.6.0.20: Implementation of quasi-Newton method with recursive update of approximate Jacobians. ➔ GITLAB

```
template <typename FUNCTION, typename JACOBIAN, typename SCALAR,
          int N = Eigen::Dynamic,
          typename MONITOR =
              std::function<void(unsigned int, Vector<SCALAR, N>,
                                 Vector<SCALAR, N>, Vector<SCALAR, N>)>>
Vector<SCALAR, N> upbroyd(
    FUNCTION &&F, Vector<SCALAR, N> x, JACOBIAN &&J, SCALAR reltol,
    SCALAR abstol, unsigned int maxit = 20,
    MONITOR &&monitor = [](unsigned int /*itnum*/,
                           const Vector<SCALAR, N> & /*x*/,
                           const Vector<SCALAR, N> & /*fx*/,
                           const Vector<SCALAR, N> & /*dx*/) {}) {
  // Calculate LU factorization of initial Jacobian once, cf.
     Rem. 2.5.0.10
  auto fac = J.lu();
  // First quasi-Newton correction Δx^(0) := -J_0^{-1} F(x^(0))
  Vector<SCALAR, N> s = -fac.solve(F(x));
  // Store the first quasi-Newton correction Δx^(0)
  std::vector<Vector<SCALAR, N>> dx{s};
  x += s; // x^(1) := x^(0) + Δx^(0)
  auto f = F(x); // Here = F(x^(1))
  // Array storing simplified quasi-Newton corrections Δx̄^(ℓ)
  std::vector<Vector<SCALAR, N>> dxs{};
  // Array of denominators ||x^(ℓ)||_2^2 + (Δx^(ℓ))^T Δx̄^(ℓ+1)
  std::vector<SCALAR> den{};
  monitor(0, x, f, s); // Record start of iteration
  // Main loop with correction based termination control
  for (unsigned int k = 1;
       ((s.norm() >= reltol * x.norm()) && (s.norm() >= abstol) && (k < maxit));
       ++k) {
    // Compute J_0^{-1} F(x^(k)), needed for both recursions
    s = fac.solve(f);
    // (8.6.0.19): recursion for next simplified quasi-Newton correction
    Vector<SCALAR, N> ss = s;
    for (unsigned int l = 1; l < k; ++l) {
      ss -= dxs[l - 1] * (dx[l - 1].dot(ss)) / den[l - 1];
    }
    // Store next denominator ||x^(k-1)||_2^2 + (Δx^(k-1))^T Δx̄^(k)
    den.push_back(dx[k - 1].squaredNorm() + dx[k - 1].dot(ss));
    // Store current simplified quasi-Newton correction Δx̄^(k)
    dxs.push_back(ss);
    // (8.6.0.18): Compute next quasi-Newton correction recursively
    for (unsigned int l = 0; l < k; ++l) {
      s -= dxs[l] * (dx[l].dot(s)) / den[l];
    }
    s *= (-1.0); // Comply with sign convention
    dx.push_back(s);
    // Compute next iterate x^(k+1) and F(x^(k+1))
    x += s;
    f = F(x);
    monitor(k, x, f, s); // Record progress
```

```
52    }
53    return x;
54  }
```

Computational cost : N steps

- $O(N^2 \cdot n)$ operations with vectors, (Level I)
- 1 LU-decomposition of $\mathbf{J}_0$, $N\times$ solutions of LSEs, see Section 2.3.2
- N evaluations of F !

Memory cost : N steps

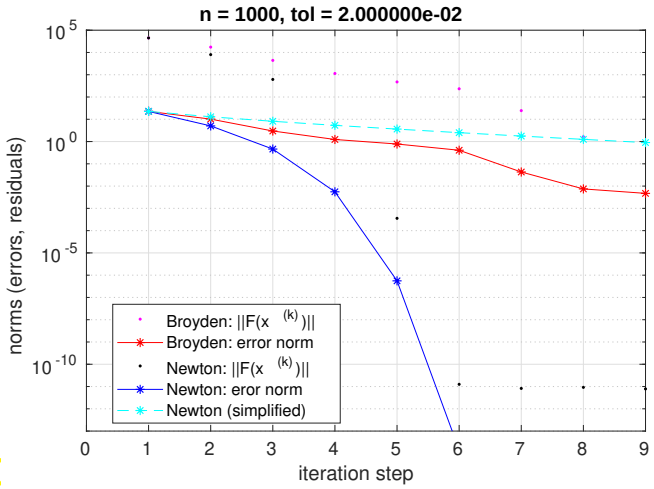
- LU-factors of $\mathtt{J}$ + auxiliary vectors $\in \mathbb{R}^n$,
- $2N$ vectors $\Delta\mathbf{x}^{(k)} \in \mathbb{R}^n$, $\Delta\overline{\mathbf{x}}^{(k)}$.

EXPERIMENT 8.6.0.21 (Broyden method for a large non-linear system)

$$F(\mathbf{x}) = \begin{cases} \mathbb{R}^n & \mapsto \mathbb{R}^n \\ \mathbf{x} & \mapsto \operatorname{diag}(\mathbf{x})\mathbf{A}\mathbf{x} - \mathbf{b} \end{cases},$$
$$\mathbf{b} = [1,2,\ldots,n] \in \mathbb{R}^n,$$
$$\mathbf{A} = \mathbf{I} + \mathbf{a}\mathbf{a}^T \in \mathbb{R}^{n,n},$$
$$\mathbf{a} = \frac{1}{\sqrt{\mathbf{1}\cdot\mathbf{b} - \mathbf{1}}}(\mathbf{b} - \mathbf{1}).$$

Initial guess: `h = 2/n; x0 = (2:h:4-h)';`

The results resemble those of Exp. 8.6.0.10 ▷ Fig. 316



Efficiency comparison: Broyden method ⟷ Newton method:
(in case of dimension n use tolerance $\mathtt{tol} = 2n \cdot 10^{-5}$, `h = 2/n; x0 = (2:h:4-h)';`)

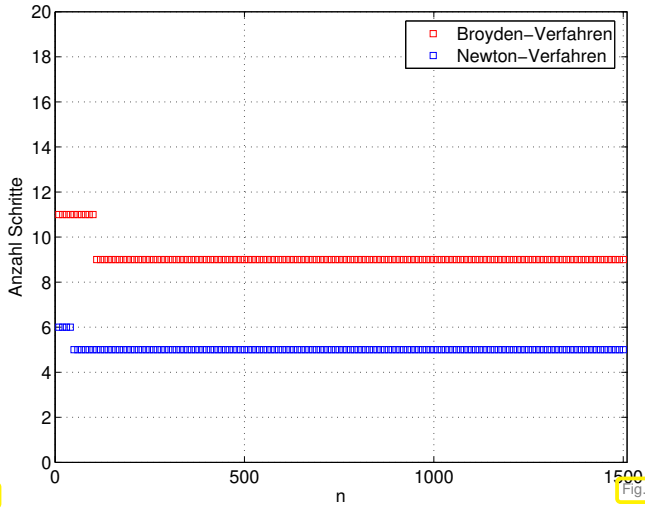


Fig. 317

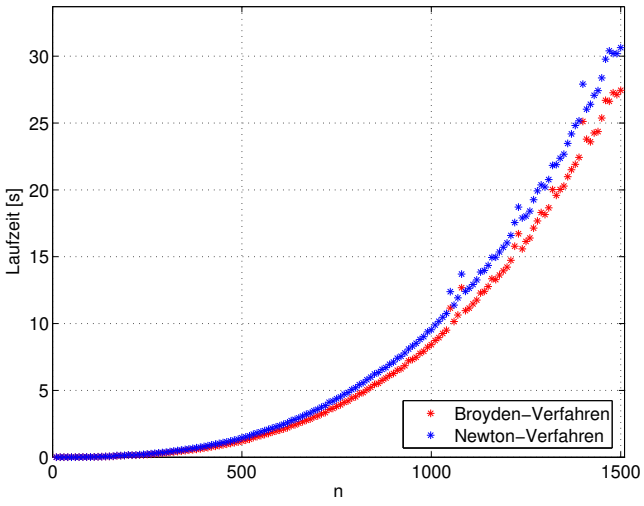


Fig. 318

☞ In conclusion,
the Broyden method is worthwhile for dimensions $n \gg 1$ and low accuracy requirements.

Supplementary literature. A comprehensive monograph about all aspects of Newton's methods and generalizations in [Deu11]. The multi-dimensional Newton method is also presented in

[Han02, Sect. 19], [DR08, Sect. 5.6], [AG11, Sect. 9.1].

For background information about quasi-Newton methods refer to [QSS00, Sect. 7.1.4], [Wer92, p. 2.3.2].

Review question(s) 8.6.0.22 (Quasi-Newton methods)

(Q8.6.0.22.A) Under what conditions is Broyden's quasi-Newton method for solving $F(\mathbf{x}) = \mathbf{0}$, $F : D \subset \mathbb{R}^n \to \mathbb{R}^n$, $n \in \mathbb{N}$,

$$\begin{aligned} \mathbf{x}^{(k+1)} &:= \mathbf{x}^{(k)} + \Delta\mathbf{x}^{(k)} \quad , \quad \Delta\mathbf{x}^{(k)} := -\mathbf{J}_k^{-1} F(\mathbf{x}^{(k)}) \, , \\ \mathbf{J}_{k+1} &:= \mathbf{J}_k + \frac{F(\mathbf{x}^{(k+1)})(\Delta\mathbf{x}^{(k)})^\top}{\|\Delta\mathbf{x}^{(k)}\|_2^2} \, , \qquad k \in \mathbb{N}_0 \, , \end{aligned} \tag{8.6.0.6}$$

affine invariant?

Remember that an iterative method for solving $F(\mathbf{x}) = \mathbf{0}$ is called affine invariant, if it produces the same sequence of iterates when applied (with the same initial guess) to $\mathbf{A}F(\mathbf{x}) = \mathbf{0}$ with any regular matrix $\mathbf{A} \in \mathbb{R}^{n,n}$.

(Q8.6.0.22.B) Show that the matrices $\mathbf{J}_k$ from Broyden's quasi-Newton method (8.6.0.6) satisfy the secant condition

$$\mathbf{J}_k(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}) = F(\mathbf{x}^{(k)}) - F(\mathbf{x}^{(k-1)}) \, . \tag{8.6.0.2}$$

△

8.7 Non-linear Least Squares [DR08, Ch. 6]

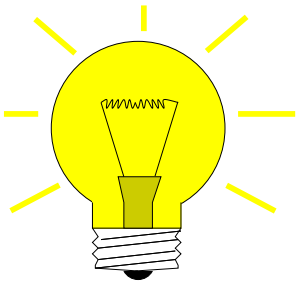
Video tutorial for Section 8.7 "Non-linear Least Squares": (7 minutes) Download link, tablet notes

So far we have studied non-linear systems of equations $F(\mathbf{x}) = \mathbf{0}$ with the same number $n \in \mathbb{N}$ of unknowns and equations: $F : D \subset \mathbb{R}^n \to \mathbb{R}^n$. This generalizes square linear systems of equations whose numerical treatment was the subjects of Chapter 2. Then, in Chapter 3, we turned our attention to *overdetermined* linear systems of equations $\mathbf{Ax} = \mathbf{b}$ with $\mathbf{A} \in \mathbb{R}^{m,n}$, $m > n$. Now we take the same step for non-linear systems of equations $F(\mathbf{x}) = \mathbf{0}$ and admit non-linear functions $F : D \subset \mathbb{R}^n \to \mathbb{R}^m$, $m > n$.

For overdetermined linear systems of equations we had to introduce the concept of a least-squares solution in Section 3.1, Def. 3.1.1.1. The same concept will apply in the non-linear case.

EXAMPLE 8.7.0.1 (Least squares data fitting) In Section 5.1 we discussed the reconstruction of a *parameterized function* $f(x_1, \ldots, x_n; \cdot) : D \subset \mathbb{R} \mapsto \mathbb{R}$ from data points (t_i, y_i), $i = 1, \ldots, n$, by imposing interpolation conditions. We demanded that the number n of parameters agreed with number of data points. Thus, in the case of a general depedence of f on the parameters, the interpolation conditions (5.1.0.2) yield a non-linear system of equations.

The interpolation approach is justified in the case of *highly accurate data*. However, we frequently encountered *inaccurate data*, for instance, due to *measurement errors*. As we discussed in Section 5.7 this renders the interpolation approach dubious, also in light of the impact of "outliers".



Idea: Mitigate the impact of data uncertainty by choosing *fewer parameters than data points*.

➢ Thus measurement errors can "average out".

Guided by this idea we arrive at a particular version of the least squares data fitting problem from Section 5.7, *cf.* (5.7.0.2).

Non-linear least squares fitting problem

Given:
- ✦ data points (t_i, y_i), $i = 1, \ldots, m$
- ✦ (symbolic formula) for parameterized function $f(x_1, \ldots, x_n; \cdot) : D \subset \mathbb{R} \mapsto \mathbb{R}$, $n < m$

Sought: parameter values $x_1^*, \ldots, x_n^* \in \mathbb{R}$ such that

$$(x_1^*, \ldots, x_n^*) = \operatorname*{argmin}_{\mathbf{x} \in \mathbb{R}^n} \sum_{i=1}^{m} |f(x_1, \ldots, x_n; t_i) - y_i|^2 \,. \tag{8.7.0.3}$$

As we did in § 5.7.0.13 for the linear case, we can rewrite (8.7.0.3) by introducing

$$F(\mathbf{x}) := \begin{bmatrix} f(x_1, \ldots, x_n, t_1) - y_1 \\ \vdots \\ f(x_1, \ldots, x_n, t_m) - y_m \end{bmatrix} \quad , \quad \mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \,. \tag{8.7.0.4}$$

With this notation we have the equivalence

$$(8.7.0.3) \quad \Leftrightarrow \quad \mathbf{x}^* = \operatorname*{argmin}_{\mathbf{x} \in \mathbb{R}^n} \sum_{i=1}^{m} |f(x_1, \ldots, x_n, t_i) - y_i|^2 = \operatorname*{argmin}_{\mathbf{x} \in \mathbb{R}^n} \|F(\mathbf{x})\|_2^2 \,. \tag{8.7.0.5}$$

The previous example motivates the following definition generalizing Def. 3.1.1.1.

Definition 8.7.0.6. Non-linear least-squares solution

Given $F : D \subset \mathbb{R}^n \mapsto \mathbb{R}^m$, $m, n \in \mathbb{N}$, $m > n$, we call $\mathbf{x}^*$ a **non-linear least squares solution** of $F(\mathbf{x}) = \mathbf{0}$, if

$$\mathbf{x}^* \in \operatorname{argmin}_{\mathbf{x} \in D} \|F(\mathbf{x})\|_2^2 \,.$$

The search for such non-linear least-squares solutions is our current concern.

Non-linear least squares problem

Given: $F : D \subset \mathbb{R}^n \mapsto \mathbb{R}^m$, $m, n \in \mathbb{N}$, $m > n$.

Find: $\mathbf{x}^* \in D$: $\mathbf{x}^* = \operatorname{argmin}_{\mathbf{x} \in D} \Phi(\mathbf{x})$, $\Phi(\mathbf{x}) := \frac{1}{2}\|F(\mathbf{x})\|_2^2$. (8.7.0.7)

Terminology: $D \triangleq$ parameter space, $x_1, \ldots, x_n \triangleq$ parameter.

As in the case of linear least squares problems (→ Section 3.1.1): a non-linear least squares problem is related to an overdetermined non-linear system of equations $F(\mathbf{x}) = \mathbf{0}$.

As for non-linear systems of equations discussed in the beginning of this chapter, existence and uniqueness of $\mathbf{x}^*$ in (8.7.0.7) has to be established in each concrete case!

Remark 8.7.0.8 ("Full-rank condition") Recall from Rem. 3.1.2.15, Ex. 3.1.2.17, Rem. 3.1.2.18 that for a linear least-squares problem $\|\mathbf{Ax} - \mathbf{b}\| \to \min$ full rank of the matrix $\mathbf{A} \in \mathbb{R}^{m,n}$ was linked to a "good model", in which every parameter had an influence independently of the others.

Also in the non-linear setting we we require "independence for each parameter":

$$\exists \text{ neighbourhood } \mathcal{U}(\mathbf{x}^*) \quad \text{such that} \quad \mathrm{D}F(\mathbf{x}) \text{ has } \textit{full rank } n \quad \forall\, \mathbf{x} \in \mathcal{U}(\mathbf{x}^*)\,. \tag{8.7.0.9}$$

This means that the columns of the Jacobi matrix $DF(\mathbf{x})$ must be linearly independent.

If (8.7.0.9) is not satisfied, then the parameters are redundant in the sense that fewer parameters would be enough to model the same dependence (locally at $\mathbf{x}^*$), *cf.* Rem. 3.1.2.18. ⌟

Review question(s) 8.7.0.10 (Non-linear least squares)

(Q8.7.0.10.A) The one-dimensional non-linear least-squares data fitting problem for a sequence $(t_i, y_i) \in \mathbb{R}^2$, $i = 1, \ldots, m$ of data points relies on families of parameterized functions

$$f(x_1, \ldots, x_n; \cdot) : D \subset \mathbb{R} \mapsto \mathbb{R}\,, \quad n < m\,,$$

and seeks to find parameter values $x_1^*, \ldots, x_n^* \in \mathbb{R}$ such that

$$(x_1^*, \ldots, x_n^*) = \operatorname*{argmin}_{\mathbf{x} \in \mathbb{R}^n} \sum_{i=1}^{m} |f(x_1, \ldots, x_n; t_i) - y_i|^2\,. \tag{8.7.0.3}$$

Explain, why this generalizes one-dimensional linear least-squares data fitting.

(Q8.7.0.10.B) Given data points $(t_i, y_i) \in \mathbb{R}^2$, $i = 1, \ldots, m$, are to be fitted in least-squares sense by *linear combinations* of two functions $t \mapsto e^{\lambda_i t}$, $\lambda_i \in \mathbb{R}$, $i = 1, 2$. Bring this problem in the form

1D non-linear least squares fitting problem

Given:
- ✦ data points (t_i, y_i), $\quad i = 1, \ldots, m$
- ✦ (symbolic formula) for parameterized function
$$f(x_1, \ldots, x_n; \cdot) : D \subset \mathbb{R} \mapsto \mathbb{R}, \quad n < m$$

Sought: parameter values $x_1^*, \ldots, x_n^* \in \mathbb{R}$ such that

$$(x_1^*, \ldots, x_n^*) = \operatorname*{argmin}_{\mathbf{x} \in \mathbb{R}^n} \sum_{i=1}^{m} |f(x_1, \ldots, x_n; t_i) - y_i|^2\,. \tag{8.7.0.3}$$

by describing a suitable parameterized function f.

(Q8.7.0.10.C) We expect a real-valued random variable X to be normally distributed $X \sim N(\sigma, \mu)$, that is, its density is given by the function

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\,, \quad x \in \mathbb{R}\,, \quad \sigma > 0\,.$$

We collect $m \gg 1$ independent samples $x_i \in \mathbb{R}$ of X, and want to use them to estimate σ. Outline how this can be done using non-linear least-squares fitting.

You may resort to the empiric cumulative distribution function

$$C_X : \{x_1, \ldots, x_m\} \to \mathbb{R} \quad , \quad C_X(x_i) := \frac{1}{m} \sharp\{j : \ x_j \leq x_i\} \ , \quad i \in \{1, \ldots, m\} \ ,$$

to formulate an overdetermined non-linear system of equations.

(Q8.7.0.10.D) A scientist proposes that you fit times series data $(t_i, y_i) \in \mathbb{R}^2$, $i = 1, \ldots, n$, by linear combinations of m shifted exponentials $t \mapsto \exp(\lambda(t - c_j))$, $j = 1, \ldots, m$, with unknown shifts $c_j \in \mathbb{R}$. Is this a good idea? Justify your judgment by examining the associated non-linear system of equations $F(\mathbf{x}) = \mathbf{0}$ and $\mathrm{D}\, F(\mathbf{x})$.

△

8.7.1 (Damped) Newton Method

Video tutorial for Section 8.7.1 "Non-linear Least Squares: (Damped) Newton Method": (13 minutes) Download link, tablet notes

We examine a first, natural approach to solving the non-linear least squares problem

$$\mathbf{x}^* \in D: \quad \mathbf{x}^* = \operatorname{argmin}_{\mathbf{x} \in D} \Phi(\mathbf{x}) \ , \quad \Phi(\mathbf{x}) := \tfrac{1}{2} \|F(\mathbf{x})\|_2^2 \ . \tag{8.7.0.7}$$

We assume that $F : D \subset \mathbb{R}^n \to \mathbb{R}^m$ is *twice continuously differentiable*. Then the non-linear least-squares solution $\mathbf{x}^*$ has to be a zero of the derivative of $\mathbf{x} \mapsto \Phi(\mathbf{x})$:

$$\Phi(\mathbf{x}^*) = \min \quad \Rightarrow \quad \mathbf{grad}\, \Phi(\mathbf{x}) = \mathbf{0} \ , \quad \text{where} \quad \mathbf{grad}\, \Phi(\mathbf{x}) := \left[\frac{\partial \Phi}{\partial x_1}(\mathbf{x}), \ldots, \frac{\partial \Phi}{\partial x_n}(\mathbf{x}) \right]^\top \in \mathbb{R}^n \ .$$

§8.7.1.1 (The Newton iteration) Note that $\mathbf{grad}\, \Phi : D \subset \mathbb{R}^n \mapsto \mathbb{R}^n$. The simple idea is to use Newton's method (→ Section 8.5) to solve the non-linear $n \times n$ system of equations $\mathbf{grad}\, \Phi(\mathbf{x}) = \mathbf{0}$.

The Newton iteration (8.5.1.6) for non-linear system of equations $\mathbf{grad}\, \Phi(\mathbf{x}) = \mathbf{0}$ is

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \mathrm{H}\, \Phi(\mathbf{x}^{(k)})^{-1} \mathbf{grad}\, \Phi(\mathbf{x}^{(k)}) \ , \tag{8.7.1.2}$$

where $\mathrm{H}\, \Phi(\mathbf{x}) \in \mathbb{R}^{n,n}$ is the **Hessian matrix** of Φ, see Def. 8.5.1.18:

$$\mathrm{H}\, \Phi(\mathbf{x}) = \left[\frac{\partial^2 \Phi}{\partial x_i \partial x_j} \right]_{i,j=1}^{n} \in \mathbb{R}^{n,n} \ .$$

Using the definition $\Phi(\mathbf{x}) := \frac{1}{2} \|F(\mathbf{x})\|^2$ we can express $\mathbf{grad}\, \Phi$ and $\mathrm{H}\, \Phi$ in terms of $F : \mathbb{R}^n \mapsto \mathbb{R}^n$. First, since $\Phi(\mathbf{x}) = (G \circ F)(\mathbf{x})$ with $G : \mathbb{R}^m \to \mathbb{R}$, $G(\mathbf{z}) := \frac{1}{2} \|\mathbf{z}\|^2$, the chain rule

$$\mathrm{D}(H \circ G)(\mathbf{x})\mathbf{h} = \mathrm{D}\, H(G(\mathbf{x}))(\mathrm{D}\, G(\mathbf{x})\mathbf{h}) \ , \quad \mathbf{h} \in V \, , \mathbf{x} \in D \ , \tag{8.5.1.16}$$

gives

$$\mathrm{D}\, \Phi(\mathbf{x})\mathbf{h} = \mathrm{D}\, G(F(\mathbf{x}))\, \mathrm{D}\, F(\mathbf{x})\mathbf{h} = F(\mathbf{x})^\top \mathrm{D}\, F(\mathbf{x})\mathbf{h} \ , \tag{8.7.1.3}$$

$$\blacktriangleright \quad \mathbf{grad}\, \Phi(\mathbf{x}) = \mathrm{D}\, F(\mathbf{x})^T F(\mathbf{x}) \ . \tag{8.7.1.4}$$

Alternatively, we could also have used the product rule

$$T(\mathbf{x}) := b(H(\mathbf{x}), G(\mathbf{x})) \quad \Rightarrow \quad \mathrm{D}\,T(\mathbf{x})\mathbf{h} = b(\mathrm{D}\,H(\mathbf{x})\mathbf{h}, G(\mathbf{x})) + b(H(\mathbf{x}), \mathrm{D}\,G(\mathbf{x})\mathbf{h}) \,, \quad \mathbf{h} \in V, \mathbf{x} \in D \,. \tag{8.5.1.17}$$

with $b : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, $b(\mathbf{x}, \mathbf{y}) := \frac{1}{2}\mathbf{x}^\top \mathbf{y}$, and $H = G := F$, which yields

$$\mathrm{D}\,\Phi(\mathbf{x})\mathbf{h} = \tfrac{1}{2}(\mathrm{D}\,F(\mathbf{x})\mathbf{h})^\top F(\mathbf{x}) + \tfrac{1}{2}F(\mathbf{x})^\top(\mathrm{D}\,F(\mathbf{x})\mathbf{h}) \,, \quad \mathbf{h} \in \mathbb{R}^n \,,$$

the same as before, of course.

A third option is direct "pedestrian-style" differentiation using the Taylor approximation

$$F(\mathbf{x}+\mathbf{h}) = F(\mathbf{x}) + \mathrm{D}\,F(\mathbf{x})\mathbf{h} + O(\|\mathbf{h}\|_2^2) \quad \text{for} \quad h \to \mathbf{0} :$$

$$\begin{aligned}
\Phi(\mathbf{x}+\mathbf{h}) &= \tfrac{1}{2}\|F(\mathbf{x}+\mathbf{h})\|_2^2 = \tfrac{1}{2}F(\mathbf{x}+\mathbf{h})^\top F(\mathbf{x}+\mathbf{h}) \\
&= \frac{1}{2}\Big(F(\mathbf{x}) + \mathrm{D}\,F(\mathbf{x})\mathbf{h} + O(\|\mathbf{h}\|_2^2)\Big)^\top \Big(F(\mathbf{x}) + \mathrm{D}\,F(\mathbf{x})\mathbf{h} + O(\|\mathbf{h}\|_2^2)\Big) \\
&= \tfrac{1}{2}F(\mathbf{x})^\top F(\mathbf{x}) + \tfrac{1}{2}F(\mathbf{x})^\top \mathrm{D}\,F(\mathbf{x})\mathbf{h} + \tfrac{1}{2}(\mathrm{D}\,F(\mathbf{x})\mathbf{h})^\top F(\mathbf{x}) + O(\|\mathbf{h}\|_2^2) \\
&= \Phi(\mathbf{x}) + \underbrace{F(\mathbf{x})^\top \mathrm{D}\,F(\mathbf{x})}_{=\mathbf{grad}\,\Phi(\mathbf{x})^\top}\mathbf{h} + O(\|\mathbf{h}\|_2^2) \quad \text{for} \quad h \to \mathbf{0} \,.
\end{aligned}$$

In a second step, we can apply the product rule (8.5.1.17) to $\mathbf{x} \mapsto \mathrm{D}\,F(\mathbf{x})^T F(\mathbf{x})$ and get

$$\begin{aligned}
&\mathrm{H}\,\Phi(\mathbf{x}) := \mathrm{D}(\mathbf{grad}\,\Phi)(\mathbf{x}) = \mathrm{D}\,F(\mathbf{x})^T \mathrm{D}\,F(\mathbf{x}) + \sum_{j=1}^m F_j(\mathbf{x})\,\mathrm{D}^2 F_j(\mathbf{x}) \,, \\
&\qquad\qquad \Updownarrow \\
&(\mathrm{H}\,\Phi(\mathbf{x}))_{i,k} = \sum_{j=1}^n \left\{ \frac{\partial^2 F_j}{\partial x_i \partial x_k}(\mathbf{x}) F_j(\mathbf{x}) + \frac{\partial F_j}{\partial x_k}(\mathbf{x}) \frac{\partial F_j}{\partial x_i}(\mathbf{x}) \right\} .
\end{aligned} \tag{8.7.1.5}$$

We make the recommendation, *cf.* § 8.5.1.15, that when in doubt, the reader should differentiate components of matrices and vectors! Let us pursue this "pedestrian option" also in this case. To begin with, we recall that the derivative of the i-th component of $\mathbf{grad}\,\Phi$ yields the i-th row of the Jacobian $(\mathrm{D}\,\mathbf{grad}\,\Phi)(\mathbf{x}) = \mathrm{H}\,\Phi(\mathbf{x}) \in \mathbb{R}^{n,n}$. So for some $i \in \{1, \ldots, n\}$ we abbreviate

$$g(\mathbf{x}) := (\mathbf{grad}\,\Phi(\mathbf{x}))_i \overset{(8.7.1.4)}{=} (\mathrm{D}\,F(\mathbf{x}))_{:,i}^\top F(\mathbf{x}) \,.$$

We compute the components of the gradient of g, which give entries of $\mathrm{H}\,\Phi$:

$$\begin{aligned}
(\mathrm{H}\,\Phi(\mathbf{x}))_{i,k} = \frac{\partial g}{\partial x_k}(\mathbf{x}) &= \frac{\partial}{\partial x_k}\left\{ \sum_{\ell=1}^m \frac{\partial F_\ell}{\partial x_i}(\mathbf{x}) F_\ell(\mathbf{x}) \right\} \\
&= \sum_{\ell=1}^m \left\{ \frac{\partial^2 F_\ell}{\partial x_i \partial x_k}(\mathbf{x}) + \frac{\partial F_\ell}{\partial x_i}(\mathbf{x}) \frac{\partial F_\ell}{\partial x_k}(\mathbf{x}) \right\} , \quad k = 1, \ldots, n \,.
\end{aligned}$$

Of course, we end up with the same formula as in (8.7.1.5).

The above derivative formulas permit us to rewrite (8.7.1.2) in concrete terms. We obtain the Newton correction $\mathbf{s} \in \mathbb{R}^n$ to the current Newton iterate $\mathbf{x}^{(k)}$ by solving the $n \times n$ linear system of equations

$$\underbrace{\left(\mathrm{D}\,F(\mathbf{x}^{(k)})^T \mathrm{D}\,F(\mathbf{x}^{(k)}) + \sum_{j=1}^m F_j(\mathbf{x}^{(k)})\,\mathrm{D}^2 F_j(\mathbf{x}^{(k)}) \right)}_{=\mathrm{H}\,\Phi(\mathbf{x}^{(k)})} \mathbf{s} = -\underbrace{\mathrm{D}\,F(\mathbf{x}^{(k)})^T F(\mathbf{x}^{(k)})}_{=\mathbf{grad}\,\Phi(\mathbf{x}^{(k)})} \,. \tag{8.7.1.6}$$

All the techniques presented in Section 8.5 (damping, termination) can now be applied to the particular Newton iteration for $\mathbf{grad}\,\Phi = \mathbf{0}$. We refer to that section. ⌟

Remark 8.7.1.7 (Newton method and minimization of quadratic functional) Newton's method (8.7.1.2) for (8.7.0.7) can be read as *successive minimization* of a local quadratic approximation of Φ:

$$\Phi(\mathbf{x}) \approx Q(\mathbf{s}) := \Phi(\mathbf{x}^{(k)}) + \mathbf{grad}\,\Phi(\mathbf{x}^{(k)})^T\mathbf{s} + \frac{1}{2}\mathbf{s}^T\,\mathrm{H}\,\Phi(\mathbf{x}^{(k)})\mathbf{s}\,, \tag{8.7.1.8}$$

$$\mathbf{grad}\,Q(\mathbf{s}) = 0 \quad\Leftrightarrow\quad \mathrm{H}\,\Phi(\mathbf{x}^{(k)})\mathbf{s} + \mathbf{grad}\,\Phi(\mathbf{x}^{(k)}) = 0 \quad\Leftrightarrow\quad (8.7.1.6)\,.$$

➢ So we deal with yet another model function method (→ Section 8.4.2) with quadratic model function Q for Φ.

⌟

Review question(s) 8.7.1.9 (Non-linear Least Squares: (Damped) Newton Method)

(Q8.7.1.9.A) For $\mathbf{grad}\,\Phi$ and $\mathrm{H}\,\Phi$ for $\Phi(\mathbf{x}) := \frac{1}{2}\|F(\mathbf{x})\|_2^2$, $F = [F_1,\ldots,F_m]^\top : \mathbb{R}^n \to \mathbb{R}^m$ twice continuously differentiable compute

$$\mathbf{grad}\,\Phi(\mathbf{x}) := \left[\frac{\partial\Phi}{\partial x_1}(\mathbf{x}),\ldots,\frac{\partial\Phi}{\partial x_n}(\mathbf{x})\right]^\top \in \mathbb{R}^n \quad,\quad \mathrm{H}\,\Phi(\mathbf{x}) = \left[\frac{\partial^2\Phi}{\partial x_i\partial x_j}\right]_{i,j=1}^n \in \mathbb{R}^{n,n}\,,$$

in the "pedestrian way" based on partial derivatives.

(Q8.7.1.9.B) Consider the case $F(\mathbf{x}) := \mathbf{A}\mathbf{x} - \mathbf{b}$, $\mathbf{A} \in \mathbb{R}^{m,n}$, $\mathbf{b} \in \mathbb{R}^m$. What does the Newton iteration

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \mathrm{H}\,\Phi(\mathbf{x}^{(k)})^{-1}\mathbf{grad}\,\Phi(\mathbf{x}^{(k)})\,,\quad k \in \mathbb{N}_0\,, \tag{8.7.1.2}$$

boil down to in this case ($\Phi(\mathbf{x}) := \frac{1}{2}\|F(\mathbf{x})\|_2^2$)?

△

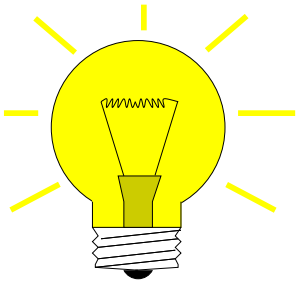
8.7.2 Gauss-Newton Method

Video tutorial for Section 8.7.2 "(Trust-region) Gauss-Newton Method": (13 minutes) Download link, tablet notes

The Newton method derived in Section 8.7.1 hinges on the availability of *second derivatives* of F. This compounds difficulties of implementation, in particular, if F is given only implicitly or in procedural form. Now we will learn about a method for the non-linear least-squares problem

$$\mathbf{x}^* \in D:\quad \mathbf{x}^* = \operatorname{argmin}_{\mathbf{x}\in D}\Phi(\mathbf{x})\,,\quad \Phi(\mathbf{x}) := \tfrac{1}{2}\|F(\mathbf{x})\|_2^2\,, \tag{8.7.0.7}$$

that relies on first derivatives of F only.



Idea: *Local linearization* of F (here at $\mathbf{y}$):

$$F(\mathbf{x}) \approx F(\mathbf{y}) + \mathrm{D}\,F(\mathbf{y})(\mathbf{x}-\mathbf{y})\,.$$

➢ Leads to a sequence of *linear* least squares problems

Details: Employing local linearization we find that the minimization problem

$$\operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^n}\|F(\mathbf{x})\|_2 \qquad \text{is approximated by} \qquad \underbrace{\operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^n}\|F(\mathbf{x}_0) + \mathrm{D}\,F(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0)\|_2}_{(\spadesuit)}\,,$$

where $\mathbf{x}_0$ is an approximation of the solution $\mathbf{x}^*$ of (8.7.0.7). This is a linear least squares problem in the standard form given in Def. 3.1.1.1.

$$(\spadesuit) \quad \Leftrightarrow \quad \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2 \quad \text{with} \quad \begin{aligned} \mathbf{A} &:= \mathrm{D}F(\mathbf{x}_0) \in \mathbb{R}^{m,n} , \\ \mathbf{b} &:= -F(\mathbf{x}_0) + \mathrm{D}F(\mathbf{x}_0)\mathbf{x}_0 \in \mathbb{R}^m . \end{aligned}$$

Note that by condition (8.7.0.9) $\mathbf{A}$ has full rank, if $\mathbf{x}_0$ sufficiently close to $\mathbf{x}^*$.

Also be aware that that this approach is different from local quadratic approximation of Φ underlying Newton's method for (8.7.0.7), see Section 8.7.1, Rem. 8.7.1.7.

The idea of local linearization leads to the **Gauss-Newton iteration**, if we make the full-rank assumption (8.7.0.9), which, thanks to Cor. 3.1.2.13 , guarantees uniqueness of solutions of the linear least-squares problem ♠. Making the substitutions $\mathbf{x}_0 := \mathbf{x}^{(k)}$ and $\mathbf{s} := \mathbf{x} - \mathbf{x}^{(k)}$, we recover at the following iterative method:

$$\begin{aligned} &\text{Initial guess} \quad \mathbf{x}^{(0)} \in D \\ &\mathbf{x}^{(k+1)} := \mathbf{x}^{(k)} + \mathbf{s} \quad , \quad \mathbf{s} := \operatorname{argmin}_{\mathbf{s}'\in\mathbb{R}^n} \left\| F(\mathbf{x}^{(k)}) + \mathrm{D}F(\mathbf{x}^{(k)})\mathbf{s}' \right\|_2 . \end{aligned} \tag{8.7.2.1}$$

a linear least squares problem!

C++ code 8.7.2.2: Generic algorithm: Gauss-Newton method ⇝ GITLAB

```
2  template <typename FUNCTION, typename JACOBIAN>
3  Eigen::VectorXd gn(const Eigen::VectorXd &init, FUNCTION &&F, JACOBIAN &&J,
4                     double rtol = 1.0E-6, double atol = 1.0E-8) {
5    Eigen::VectorXd x = init; // Vector for iterates x^(k)
6    // Vector for Gauss-Newton correction s
7    Eigen::VectorXd s = J(x).householderQr().solve(F(x)); //
8    x = x - s;
9    // A posteriori termination based on absolute and relative tolerances
10   while ((s.norm() > rtol * x.norm()) && (s.norm() > atol)) { //
11     s = J(x).householderQr().solve(F(x)); //
12     x = x - s;
13   }
14   return x;
15 }
```

Comments on Code 8.7.2.2:

- ☞ Recall EIGEN's built-in functions for solving linear least-squares problems used in Line 7 and Line 11, see Code 3.3.4.2.
- ☞ Argument `x` passes initial guess $\mathbf{x}^{(0)} \in \mathbb{R}^n$, argument `F` must be a *handle* to a function $F : \mathbb{R}^n \mapsto \mathbb{R}^m$, argument `J` provides the Jacobian of F, namely $\mathrm{D}F : \mathbb{R}^n \mapsto \mathbb{R}^{m,n}$, arguments `rtol` and `atol` specify the tolerances for termination
- ☞ Line 10 implements the same correction-based stopping rule as it was proposed for Newton's method in Section 8.5.3, (8.5.3.2): The iteration terminates, if the Euclidean norm of the Gauss-Newton correction $\mathbf{s}$ is either small relative to the norm of the current iterate or small in absolute terms.

Note that the function of Code 8.7.2.2 also implements Newton's method (→ Section 8.5.1) in the case $m = n$!

Remark 8.7.2.3 (Gauss-Newton versus Newton) Let us summarize the pros and cons of using the Gauss-Newton approach:

Advantage of the Gauss-Newton method : second derivative of F not needed.
Drawback of the Gauss-Newton method : no local quadratic convergence. ⌟

EXAMPLE 8.7.2.4 (Non-linear fitting of data (II) → Ex. 8.7.0.1) Given data points (t_i, y_i), $i = 1, \ldots, m$, we consider the non-linear data fitting problem (8.7.0.5) for the parameterized function

$$f(x_1, x_2, x_3; t) := x_1 + x_2 \exp(-x_3 t) .$$

This means that we face an (overdetermined) non-linear system $F(\mathbf{x}) = \mathbf{0}$ with

$$F : \mathbb{R}^3 \mapsto \mathbb{R}^m \quad , \quad F(\mathbf{x}) := \begin{bmatrix} x_1 + x_2 \exp(-x_3 t_1) - y_1 \\ \vdots \\ x_1 + x_2 \exp(-x_3 t_m) - y_m \end{bmatrix} .$$

Computing partial derivatives, the Jacobian of F is seen to be

$$\mathrm{D}F(\mathbf{x}) = \begin{bmatrix} 1 & e^{-x_3 t_1} & -x_2 t_1 e^{-x_3 t_1} \\ \vdots & \vdots & \vdots \\ 1 & e^{-x_3 t_m} & -x_2 t_m e^{-x_3 t_m} \end{bmatrix} .$$

In this experiment we use data points (t_j, y_j), $j = 1, \ldots, m$, $m = 21$, $t_j = 1 + 0.3j$, $j = 1, \ldots, 21$ and "random" y_j generated by the following C++ code

```
Eigen::VectorXd gnrandinit(const Eigen::VectorXd &x) {
  std::srand((unsigned int)time(0));
  auto t = Eigen::VectorXd::LinSpaced((7.0 - 1.0) / 0.3 - 1, 1.0,
    7.0);
  auto y = x(0) + x(1) * ((-x(2) * t).array().exp());
  return y + 0.1 * (Eigen::VectorXd::Random(y.size()).array() - 0.5);
}
```

We study the convergence of

- the Newton method from Section 8.7.1,
- its extension, damped Newton method (→ Section 8.5.4),
- and the Gauss-Newton method

for different initial guesses:

✦ initial value $(1.8, 1.8, 0.1)^T$ (red curves, blue curves)
✦ initial value $(1.5, 1.5, 0.1)^T$ (cyan curves, green curves)

The first experiment investigates the iterative solution of non-linear least squares data fitting problem by means of the Newton method (8.7.1.6) and the damped Newton method from Code 8.5.4.5.

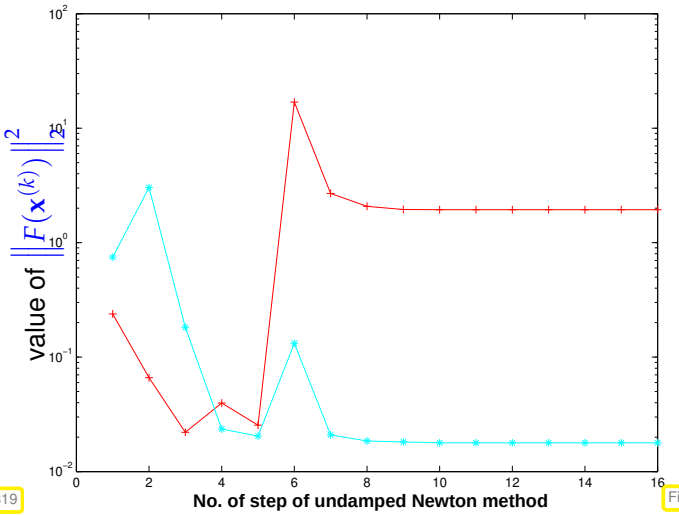


Fig. 319

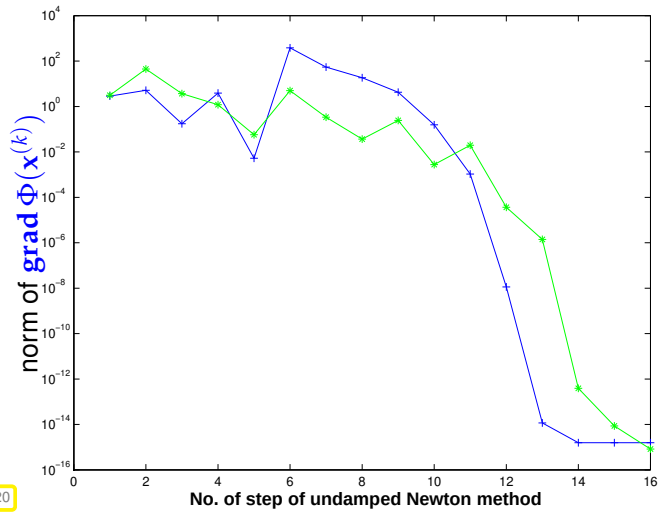


Fig. 320

Concerning the convergence behaviour of the plain Newton method we observe that

- for initial value $(1.8, 1.8, 0.1)^T$ (red curve) ➤ Newton method caught in local minimum,
- for initial value $(1.5, 1.5, 0.1)^T$ (cyan curve) ➤ fast (locally quadratic) convergence.

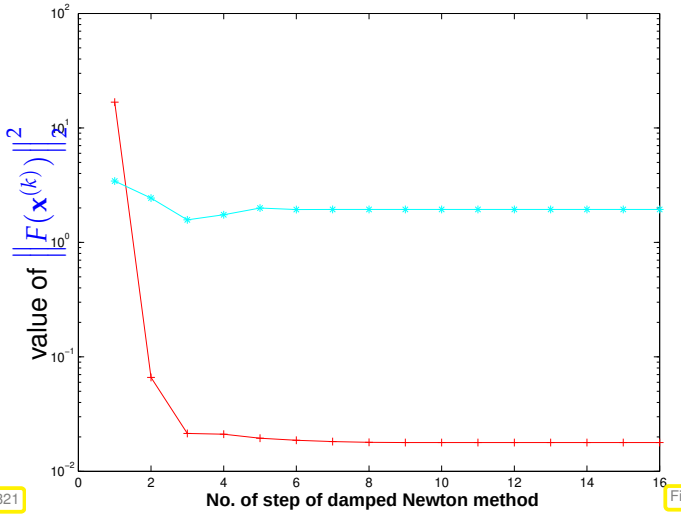


Fig. 321

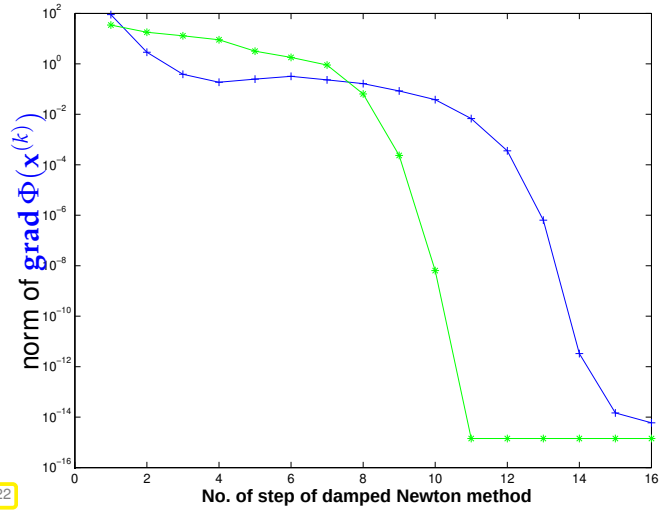


Fig. 322

The observed convergence behavior of the damped Newton method is as follows:

- For initial value $(1.8, 1.8, 0.1)^T$ (red curve) ➤ fast (locally quadratic) convergence,
- For initial value $(1.5, 1.5, 0.1)^T$ (cyan curve) ➤ Newton method caught in local minimum.

The second experiment studies iterative solution of non-linear least squares data fitting problem by means of the Gauss-Newton method (8.7.2.1), see Code 8.7.2.2.

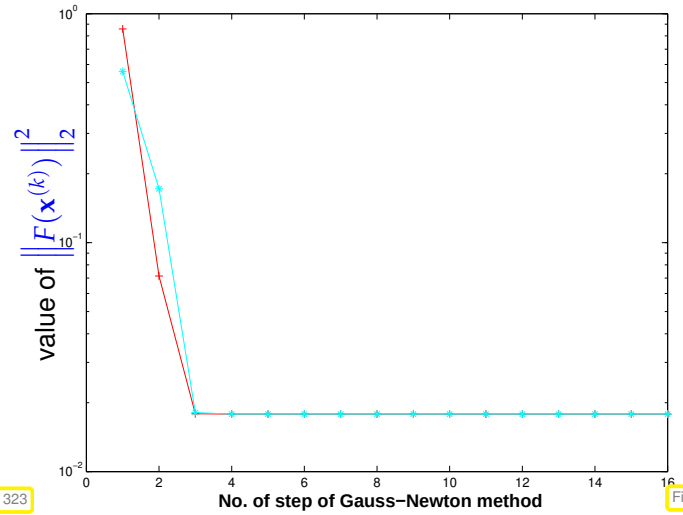


Fig. 323

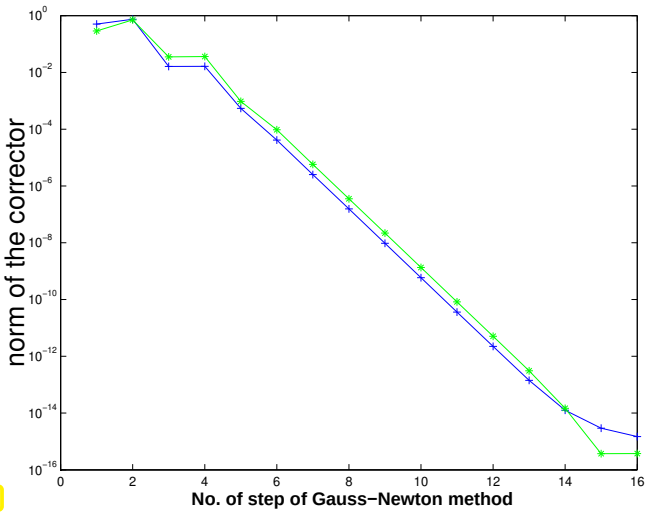


Fig. 324

For the Gauss-Newton method we observe linear convergence for *both* initial values (Refer to Def. 8.2.2.1, Rem. 8.2.2.6 for "linear convergence" and how to see it in error plots).

In this experiment the convergence of the Gauss-Newton method is asymptotically clearly slower than that of the Newton method, but less dependent on the choice of good initial guesses. This matches what is often observed in practical non-linear fitting. ⌟

8.7.3 Trust Region Method (Levenberg-Marquardt Method)

As in the case of Newton's method for non-linear systems of equations, see Section 8.5.4, often overshooting of Gauss-Newton corrections occurs. We can resort to a similar remedy as in the case of Newton's method: *damping* of Gauss-Newton corrections.

Idea: damping of the Gauss-Newton correction in (8.7.2.1) using a **penalty term**

$$\text{instead of} \quad \left\|F(\mathbf{x}^{(k)}) + \mathrm{D}F(\mathbf{x}^{(k)})\mathbf{s}\right\|_2^2 \quad \text{minimize} \quad \left\|F(\mathbf{x}^{(k)}) + \mathrm{D}F(\mathbf{x}^{(k)})\mathbf{s}\right\|_2^2 + \lambda\|\mathbf{s}\|_2^2 ,$$

where $\lambda > 0$ is a penalty parameter.

A central issue is the choice of the penalty parameter $\lambda > 0$. Since rigorous rules are elusive, heuristic strategies are used in practice, for instance

$$\lambda = \gamma\left\|F(\mathbf{x}^{(k)})\right\|_2 \quad , \quad \gamma := \begin{cases} 10 & \text{, if } \left\|F(\mathbf{x}^{(k)})\right\|_2 \geq 10 , \\ 1 & \text{, if } 1 < \left\|F(\mathbf{x}^{(k)})\right\|_2 < 10 , \\ 0.01 & \text{, if } \left\|F(\mathbf{x}^{(k)})\right\|_2 \leq 1 . \end{cases} \tag{8.7.3.1}$$

The minimization problem

$$\mathbf{s} := \operatorname*{argmin}_{\mathbf{z}\in\mathbb{R}^n} \left\|F(\mathbf{x}^{(k)}) + \mathrm{D}F(\mathbf{x}^{(k)})\mathbf{z}\right\|^2 + \lambda\|\mathbf{z}\|_2^2$$

is close to a standard linear least-squares problem and its solution $\mathbf{s}$ can be obtained by solving a linear system of equations, which we get by setting the gradient of

$$\begin{aligned} \Psi(\mathbf{z}) &:= \left\|F(\mathbf{x}^{(k)}) + \mathrm{D}F(\mathbf{x}^{(k)})\mathbf{z}\right\|^2 + \lambda\|\mathbf{z}\|_2^2 \\ &= \mathbf{z}^\top\left(\mathrm{D}F(\mathbf{x}^{(k)})^\top \mathrm{D}F(\mathbf{x}^{(k)}) + \lambda\mathbf{I}\right)\mathbf{z} + 2F(\mathbf{x}^{(k)})^\top \mathrm{D}F(\mathbf{x}^{(k)})\mathbf{z} + \left\|F(\mathbf{x}^{(k)})\right\|_2^2 , \\ \Rightarrow \quad & \mathbf{grad}\,\Psi(\mathbf{z}) = 2\left(\mathrm{D}F(\mathbf{x}^{(k)})^\top \mathrm{D}F(\mathbf{x}^{(k)})^\top + \lambda\mathbf{I}\right)\mathbf{z} + 2F(\mathbf{x}^{(k)})^\top\mathbf{z} , \end{aligned}$$

to zero. This leads to the following $n \times n$ linear system of normal equations for damped Gauss-Newton correction $\mathbf{s}$ in the k-th step, see Thm. 3.1.2.1:

$$\left(\mathrm{D}\,F(\mathbf{x}^{(k)})^T\,\mathrm{D}\,F(\mathbf{x}^{(k)}) + \lambda\mathbf{I}\right)\mathbf{s} = -\,\mathrm{D}\,F(\mathbf{x}^{(k)})^\top F(\mathbf{x}^{(k)})\;. \tag{8.7.3.2}$$

Review question(s) 8.7.3.3 (Gauss-Newton method)

(Q8.7.3.3.A) What becomes of the Gauss-Newton method for solving $F(\mathbf{x}) = \mathbf{0}$,

$$\mathbf{x}^{(k+1)} := \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^n} \left\| F(\mathbf{x}^{(k)}) + \mathrm{D}\,F(\mathbf{x}^{(k)})(\mathbf{x}-\mathbf{x}^{(k)}) \right\|_2^2 ,$$

in the linear case, when $F(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}$, $\mathbf{A} \in \mathbb{R}^{m,n}$, $\mathbf{b} \in \mathbb{R}^m$?

(Q8.7.3.3.B) What will you get when you apply the Gauss-Newton method

$$\mathbf{x}^{(k+1)} := \mathbf{x}^{(k)} - \mathbf{s} \quad , \quad \mathbf{s} := \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^n} \left\| -F(\mathbf{x}^{(k)}) + \mathrm{D}\,F(\mathbf{x}^{(k)})\mathbf{s} \right\|_2 .$$

to an $n \times n$ non-linear system of equations, that is, $F : D \subset \mathbb{R}^n \to \mathbb{R}^n$. You may assume that $\mathrm{D}\,F(\mathbf{x}^{(k)})$ always has full rank.

△

Learning Outcomes

- Knowledge about concepts related to the speed of convergence of an iteration for solving a non-linear system of equations.
- Ability to estimate type and orders of convergence from empiric data.
- Ability to predict asymptotic linear, quadratic and cubic convergence by inspection of the iteration function.
- Familiarity with (damped) Newton's method for general non-linear systems of equations and with the secant method in 1D.
- Ability to derive the Newton iteration for an (implicitly) given non-linear system of equations.
- Knowledge about quasi-Newton method as multi-dimensional generalizations of the secant method.

Bibliography

[Ale12] A. Alexanderian. *A basic note on iterative matrix inversion*. Onlie document. 2012 (cit. on pp. 652, 655).

[AG11] Uri M. Ascher and Chen Greif. *A first course in numerical methods*. Vol. 7. Computational Science & Engineering. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2011, pp. xxii+552. DOI: 10.1137/1.9780898719987 (cit. on pp. 610, 614, 623, 629, 633, 670).

[BC17] Heinz H. Bauschke and Patrick L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. Second. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. Springer, Cham, 2017, pp. xix+619. DOI: 10.1007/978-3-319-48311-5.

[DR08] W. Dahmen and A. Reusken. *Numerik für Ingenieure und Naturwissenschaftler*. Heidelberg: Springer, 2008 (cit. on pp. 607, 612, 614, 618, 619, 623, 629, 633, 652, 670–680).

[DH03] P. Deuflhard and A. Hohmann. *Numerical Analysis in Modern Scientific Computing*. Vol. 43. Texts in Applied Mathematics. Springer, 2003 (cit. on p. 656).

[Deu11] Peter Deuflhard. *Newton methods for nonlinear problems*. Vol. 35. Springer Series in Computational Mathematics. Heidelberg: Springer, 2011, pp. xii+424. DOI: 10.1007/978-3-642-23899-4 (cit. on pp. 645, 656, 660, 663, 669).

[Han02] M. Hanke-Bourgeois. *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens*. Mathematische Leitfäden. Stuttgart: B.G. Teubner, 2002 (cit. on pp. 604, 607, 619, 629, 630, 633, 635, 670).

[Mol04] C. Moler. *Numerical Computing with MATLAB*. Philadelphia, PA: SIAM, 2004 (cit. on p. 637).

[PS91] Victor Pan and Robert Schreiber. “An Improved Newton Iteration for the Generalized Inverse of a Matrix, with Applications”. In: *SIAM Journal on Scientific and Statistical Computing* 12.5 (1991), pp. 1109–1130. DOI: 10.1137/0912058 (cit. on pp. 652, 655).

[QSS00] A. Quarteroni, R. Sacco, and F. Saleri. *Numerical mathematics*. Vol. 37. Texts in Applied Mathematics. New York: Springer, 2000 (cit. on pp. 607, 614, 625, 670).

[Str19] D. Strang. *Linear Algebra and Learning from Data*. Cambridge University Press, 2019.

[Str09] M. Struwe. *Analysis für Informatiker*. Lecture notes, ETH Zürich. 2009 (cit. on pp. 605, 606, 608, 615, 618, 619, 621, 623, 628, 635, 642, 647, 651).

[Wer92] J. Werner. *Numerische Mathematik I. Lineare und nichtlineare Gleichungssysteme, Interpolation, numerische Integration*. vieweg studium. Aufbaukurs Mathematik. Braunschweig: Vieweg, 1992 (cit. on p. 670).

Chapter 9

Computation of Eigenvalues and Eigenvectors

Supplementary literature. [Bai+00] offers comprehensive presentation of numerical methods for the solution of eigenvalue problems from an algorithmic point of view.

EXAMPLE 9.0.0.1 (Resonances of linear electric circuits)

Simple electric circuit, *cf.* Ex. 2.1.0.3 ▷

- ✦ linear components (resistors, coils, capacitors) only,
- ✦ time-harmonic excitation (alternating voltage/current)
- ✦ "frequency domain" circuit model

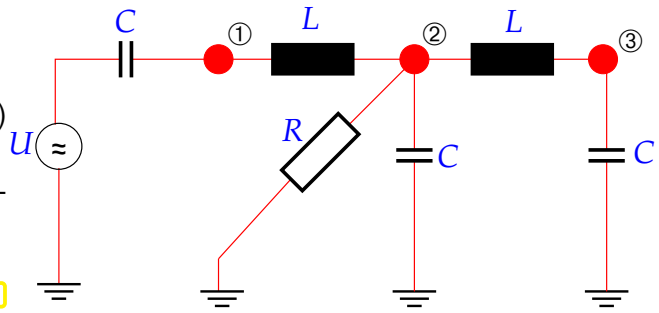


Fig. 325

Ex. 2.1.0.3: nodal analysis of linear ($\leftrightarrow$ composed of resistors, inductors, capacitors) electric circuit in frequency domain (at angular frequency $\omega > 0$) , see (2.1.0.6)

➢ linear system of equations for nodal potentials with complex system matrix $\mathbf{A}$

For circuit of Fig. 325: three unknown nodal potentials

➢ system matrix from nodal analysis at angular frequency $\omega > 0$:

$$\mathbf{A} = \begin{bmatrix} \imath\omega C + \frac{1}{\imath\omega L} & -\frac{1}{\imath\omega L} & 0 \\ -\frac{1}{\imath\omega L} & \imath\omega C + \frac{1}{R} + \frac{2}{\imath\omega L} & -\frac{1}{\imath\omega L} \\ 0 & -\frac{1}{\imath\omega L} & \imath\omega C + \frac{1}{\imath\omega L} \end{bmatrix}$$
$$= \begin{bmatrix} 0 & 0 & 0 \\ 0 & \frac{1}{R} & 0 \\ 0 & 0 & 0 \end{bmatrix} + \imath\omega \begin{bmatrix} C & 0 & 0 \\ 0 & C & 0 \\ 0 & 0 & C \end{bmatrix} + 1/\imath\omega \begin{bmatrix} \frac{1}{L} & -\frac{1}{L} & 0 \\ -\frac{1}{L} & \frac{2}{L} & -\frac{1}{L} \\ 0 & -\frac{1}{L} & \frac{1}{L} \end{bmatrix} .$$

$$\mathbf{A}(\omega) := \mathbf{W} + i\omega\mathbf{C} - i\omega^{-1}\mathbf{S} \quad , \quad \mathbf{W}, \mathbf{C}, \mathbf{S} \in \mathbb{R}^{n,n} \text{ symmetric} . \tag{9.0.0.2}$$

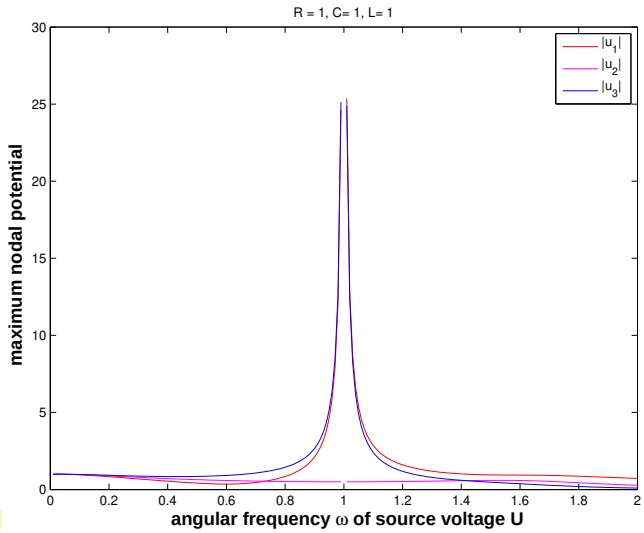


Fig. 326

◁ plot of $|u_i(U)|$, $i = 1,2,3$ for $R = L = C = 1$ (scaled model)

Blow-up of some nodal potentials for certain ω **!**

resonant frequencies = $\omega \in \{\omega \in \mathbb{R}: \mathbf{A}(\omega) \text{ singular}\}$

If the circuit is operated at a real resonant frequency, the circuit equations will not possess a solution. Of course, the real circuit will always behave in a well-defined way, but the linear model will break down due to extremely large currents and voltages. In an experiment this breakdown manifests itself as a rather explosive meltdown of circuits components. Hence, it is vital to determine resonant frequencies of circuits in order to avoid their destruction.

➥ relevance of numerical methods for solving:

$$\text{Find} \quad \omega \in \mathbb{C} \setminus \{0\}: \quad \mathbf{W} + \imath\omega\mathbf{C} + \frac{1}{\imath\omega}\mathbf{S} \quad \text{singular .}$$

This is a quadratic eigenvalue problem: find $\mathbf{x} \neq 0$, $\omega \in \mathbb{C} \setminus \{0\}$,

$$\mathbf{A}(\omega)\mathbf{x} = (\mathbf{W} + \imath\omega\mathbf{C} + \frac{1}{\imath\omega}\mathbf{S})\mathbf{x} = 0 \,. \tag{9.0.0.3}$$

Substitution: $\mathbf{y} = \frac{1}{\imath\omega}\mathbf{x} \leftrightarrow \mathbf{x} = \imath\omega\mathbf{y}$ [TM01, Sect. 3.4]:

$$(9.0.0.3) \quad \Leftrightarrow \quad \underbrace{\begin{bmatrix} \mathbf{W} & \mathbf{S} \\ \mathbf{I} & 0 \end{bmatrix}}_{:=\mathbf{M}} \underbrace{\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}}_{:=\mathbf{z}} = \omega \underbrace{\begin{bmatrix} -\imath\mathbf{C} & 0 \\ 0 & -\imath\mathbf{I} \end{bmatrix}}_{:=\mathbf{B}} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \tag{9.0.0.4}$$

➢ generalized linear eigenvalue problem of the form: find $\omega \in \mathbb{C}$, $\mathbf{z} \in \mathbb{C}^{2n} \setminus \{0\}$ such that

$$\mathbf{M}\mathbf{z} = \omega\mathbf{B}\mathbf{z} \,. \tag{9.0.0.5}$$

In this example one is mainly interested in the eigenvalues ω, whereas the eigenvectors $\mathbf{z}$ usually need not be computed.

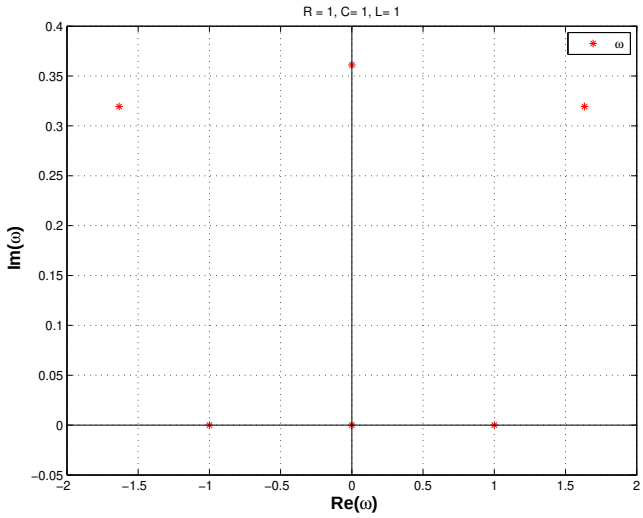


Fig. 327

◁ resonant frequencies for circuit from Fig. 325 (including decaying modes with $\mathrm{Im}(\omega) > 0$)

EXAMPLE 9.0.0.6 (Analytic solution of homogeneous linear ordinary differential equations → [Str09, Remark 5.6.1], [Gut09, Sect. 10.1],[NS02, Sect. 8.1], [DR08, Ex. 7.3])

Autonomous homogeneous linear ordinary differential equation (ODE):

$$\dot{\mathbf{y}} = \mathbf{A}\mathbf{y} \quad , \quad \mathbf{A} \in \mathbb{C}^{n,n} \, . \tag{9.0.0.7}$$

$$\mathbf{A} = \mathbf{S} \underbrace{\begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}}_{=:\mathbf{D}} \mathbf{S}^{-1} \, , \quad \mathbf{S} \in \mathbb{C}^{n,n} \text{ regular} \quad \Longrightarrow \quad \left(\dot{\mathbf{y}} = \mathbf{A}\mathbf{y} \overset{\mathbf{z}=\mathbf{S}^{-1}\mathbf{y}}{\longleftrightarrow} \dot{\mathbf{z}} = \mathbf{D}\mathbf{z} \right) .$$

➢ solution of initial value problem:

$$\dot{\mathbf{y}} = \mathbf{A}\mathbf{y} \, , \quad \mathbf{y}(0) = \mathbf{y}_0 \in \mathbb{C}^n \quad \Rightarrow \quad \mathbf{y}(t) = \mathbf{S}\mathbf{z}(t) \, , \quad \dot{\mathbf{z}} = \mathbf{D}\mathbf{z} \, , \quad \mathbf{z}(0) = \mathbf{S}^{-1}\mathbf{y}_0 \, .$$

The initial value problem for the *decoupled* homogeneous linear ODE $\dot{\mathbf{z}} = \mathbf{D}\mathbf{z}$ has a simple analytic solution

$$\mathbf{z}_i(t) = \exp(\lambda_i t)(\mathbf{z}_0)_i = \exp(\lambda_i t)\left((\mathbf{S}^{-1})^T_{i,:} \mathbf{y}_0 \right) .$$

In light of Rem. 1.3.1.3:

$$\mathbf{A} = \mathbf{S} \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} \mathbf{S}^{-1} \quad \Leftrightarrow \quad \mathbf{A}((\mathbf{S})_{:,i}) = \lambda_i((\mathbf{S})_{:,i}) \quad i = 1, \ldots, n \, . \tag{9.0.0.8}$$

In order to find the transformation matrix $\mathbf{S}$ all non-zero solution vectors (= eigenvectors) $\mathbf{x} \in \mathbb{C}^n$ of the linear eigenvalue problem

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$$

have to be found.

Contents

9.1 Theory of eigenvalue problems . . . 685
9.2 "Direct" Eigensolvers . . . 687
9.3 Power Methods . . . 690
9.3.1 Direct power method . . . 690
9.3.2 Inverse Iteration [DR08, Sect. 7.6], [QSS00, Sect. 5.3.2] . . . 697
9.3.3 Preconditioned inverse iteration (PINVIT) . . . 707
9.3.4 Subspace iterations . . . 710
9.4 Krylov Subspace Methods . . . 721

9.1 Theory of eigenvalue problems

Supplementary literature. [NS02, Ch. 7], [Gut09, Ch. 9], [QSS00, Sect. 1.7]

Definition 9.1.0.1. Eigenvalues and eigenvectors → [NS02, Sects. 7.1, 7.2], [Gut09, Sect. 9.1]

- $\lambda \in \mathbb{C}$ eigenvalue (*ger.:* Eigenwert) of $\mathbf{A} \in \mathbb{K}^{n,n}$ $:\Leftrightarrow$ $\underbrace{\det(\lambda\mathbf{I} - \mathbf{A})}_{\text{characteristic polynomial } \chi(\lambda)} = 0$
- spectrum of $\mathbf{A} \in \mathbb{K}^{n,n}$: $\sigma(\mathbf{A}) := \{\lambda \in \mathbb{C}: \lambda \text{ eigenvalue of } \mathbf{A}\}$
- eigenspace (*ger.:* Eigenraum) associated with eigenvalue $\lambda \in \sigma(\mathbf{A})$:
$$\mathbf{EigA}\lambda := \mathcal{N}\lambda\mathbf{I} - \mathbf{A}$$
- $\mathbf{x} \in \mathbf{EigA}\lambda \setminus \{0\}$ $\Rightarrow$ $\mathbf{x}$ is eigenvector
- Geometric multiplicity (*ger.:* Vielfachheit) of an eigenvalue $\lambda \in \sigma(\mathbf{A})$:
$$m(\lambda) := \dim \mathbf{EigA}\lambda$$

Two simple facts:

$$\lambda \in \sigma(\mathbf{A}) \quad \Rightarrow \quad \dim \mathbf{EigA}\lambda > 0\,, \tag{9.1.0.2}$$

$$\det(\mathbf{A}) = \det(\mathbf{A}^T) \quad \forall \mathbf{A} \in \mathbb{K}^{n,n} \quad \Rightarrow \quad \sigma(\mathbf{A}) = \sigma(\mathbf{A}^T)\,. \tag{9.1.0.3}$$

✎ notation: $\rho(\mathbf{A}) := \max\{|\lambda|: \lambda \in \sigma(\mathbf{A})\} \triangleq$ spectral radius of $\mathbf{A} \in \mathbb{K}^{n,n}$

Theorem 9.1.0.4. Bound for spectral radius

For any matrix norm $\|\cdot\|$ induced by a vector norm (→ Def. 1.5.5.10)

$$\rho(\mathbf{A}) \le \|\mathbf{A}\|\,.$$

Proof. Let $\mathbf{z} \in \mathbb{C}^n \setminus \{0\}$ be an eigenvector to the largest (in modulus) eigenvalue λ of $\mathbf{A} \in \mathbb{C}^{n,n}$. Then

$$\|\mathbf{A}\| := \sup_{\mathbf{x} \in \mathbb{C}^{n,n} \setminus \{0\}} \frac{\|\mathbf{A}\mathbf{x}\|}{\|\mathbf{x}\|} \ge \frac{\|\mathbf{A}\mathbf{z}\|}{\|\mathbf{z}\|} = |\lambda| = \rho(\mathbf{A})\,.$$

□

Lemma 9.1.0.5. Gershgorin circle theorem → [DR08, Thm. 7.13], [Han02, Thm. 32.1], [QSS00, Sect. 5.1]

For any $\mathbf{A} \in \mathbb{K}^{n,n}$ *holds true*

$$\sigma(\mathbf{A}) \subset \bigcup_{j=1}^{n} \left\{ z \in \mathbb{C} \colon |z - a_{jj}| \leq \sum_{i \neq j} |a_{ji}| \right\} .$$

Lemma 9.1.0.6. Similarity and spectrum → [Gut09, Thm. 9.7], [DR08, Lemma 7.6], [NS02, Thm. 7.2]

The spectrum of a matrix is invariant with respect to similarity transformations:

$$\forall \mathbf{A} \in \mathbb{K}^{n,n} \colon \quad \sigma(\mathbf{S}^{-1}\mathbf{A}\mathbf{S}) = \sigma(\mathbf{A}) \quad \forall \text{ regular } \mathbf{S} \in \mathbb{K}^{n,n} .$$

Lemma 9.1.0.7.

Existence of a one-dimensional invariant subspace

$$\forall \mathbf{C} \in \mathbb{C}^{n,n} \colon \quad \exists \mathbf{u} \in \mathbb{C}^{n} \colon \quad \mathbf{C}(\mathrm{Span}\{\mathbf{u}\}) \subset \mathrm{Span}\{\mathbf{u}\} .$$

Theorem 9.1.0.8. Schur normal form → [Hac94, Thm .2.8.1]

$$\forall \mathbf{A} \in \mathbb{K}^{n,n} \colon \quad \exists \mathbf{U} \in \mathbb{C}^{n,n} \text{ unitary} \colon \quad \mathbf{U}^H \mathbf{A} \mathbf{U} = \mathbf{T} \quad \text{with } \mathbf{T} \in \mathbb{C}^{n,n} \text{ upper triangular} .$$

Corollary 9.1.0.9. Principal axis transformation

$$\mathbf{A} \in \mathbb{K}^{n,n},\ \mathbf{A}\mathbf{A}^H = \mathbf{A}^H\mathbf{A} \colon \quad \exists \mathbf{U} \in \mathbb{C}^{n,n} \text{ unitary} \colon \quad \mathbf{U}^H \mathbf{A} \mathbf{U} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n) , \quad \lambda_i \in \mathbb{C} .$$

A matrix $\mathbf{A} \in \mathbb{K}^{n,n}$ with $\mathbf{A}\mathbf{A}^H = \mathbf{A}^H\mathbf{A}$ is called normal.

Examples of normal matrices are
- Hermitian matrices: $\mathbf{A}^H = \mathbf{A}$ ➤ $\sigma(\mathbf{A}) \subset \mathbb{R}$
- unitary matrices: $\mathbf{A}^H = \mathbf{A}^{-1}$ ➤ $|\sigma(\mathbf{A})| = 1$
- skew-Hermitian matrices: $\mathbf{A} = -\mathbf{A}^H$ ➤ $\sigma(\mathbf{A}) \subset i\mathbb{R}$

➤

Normal matrices can be diagonalized by *unitary* similarity transformations

Symmetric real matrices can be diagonalized by *orthogonal* similarity transformations

In Cor. 9.1.0.9:
- $\lambda_1, \ldots, \lambda_n$ = eigenvalues of $\mathbf{A}$
- Columns of $\mathbf{U}$ = orthonormal basis of eigenvectors of $\mathbf{A}$

Classes of relevant **eigenvalue problems (EVP)**:

❶ Given $\mathbf{A} \in \mathbb{K}^{n,n}$ find **all** eigenvalues (= spectrum of $\mathbf{A}$).
❷ Given $\mathbf{A} \in \mathbb{K}^{n,n}$ find $\sigma(\mathbf{A})$ plus **all** eigenvectors.
❸ Given $\mathbf{A} \in \mathbb{K}^{n,n}$ find **a few** eigenvalues and associated eigenvectors.

(Linear) generalized eigenvalue problem:

Given $\mathbf{A} \in \mathbb{C}^{n,n}$, regular $\mathbf{B} \in \mathbb{C}^{n,n}$, seek $\mathbf{x} \neq 0$, $\lambda \in \mathbb{C}$

$$\mathbf{A}\mathbf{x} = \lambda \mathbf{B}\mathbf{x} \quad \Leftrightarrow \quad \mathbf{B}^{-1}\mathbf{A}\mathbf{x} = \lambda \mathbf{x} \,. \tag{9.1.0.10}$$

$\mathbf{x} \hat{=}$ generalized eigenvector, $\lambda \hat{=}$ generalized eigenvalue

Obviously every generalized eigenvalue problem is equivalent to a standard eigenvalue problem

$$\mathbf{A}\mathbf{x} = \lambda \mathbf{B}\mathbf{x} \quad \Leftrightarrow \quad \mathbf{B}^{-1}\mathbf{A} = \lambda \mathbf{x} \,.$$

However, usually it is not advisable to use this equivalence for numerical purposes!

Remark 9.1.0.11 (Generalized eigenvalue problems and Cholesky factorization)

If $\mathbf{B} = \mathbf{B}^H$ s.p.d. (→ Def. 1.1.2.6) with Cholesky factorization $\mathbf{B} = \mathbf{R}^H\mathbf{R}$

$$\mathbf{A}\mathbf{x} = \lambda \mathbf{B}\mathbf{x} \quad \Leftrightarrow \quad \widetilde{\mathbf{A}}\mathbf{y} = \lambda \mathbf{y} \quad \text{where } \widetilde{\mathbf{A}} := \mathbf{R}^{-H}\mathbf{A}\mathbf{R}^{-1},\ \mathbf{y} := \mathbf{R}\mathbf{x} \,.$$

→ This transformation can be used for efficient computations. ⌟

9.2 "Direct" Eigensolvers

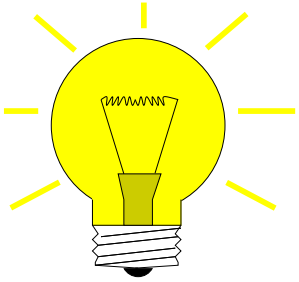
Purpose: solution of eigenvalue problems ❶, ❷ for dense matrices "up to machine precision"

MATLAB-function: `eig`

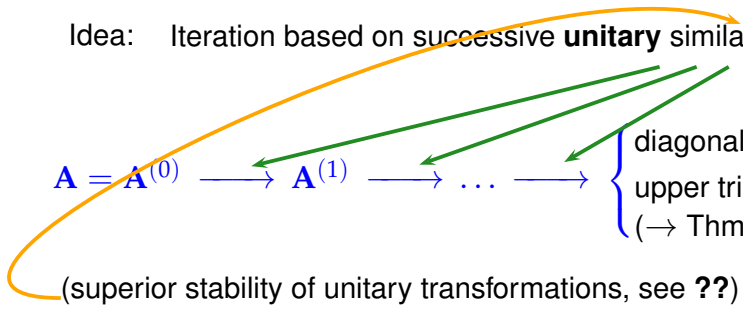
`d = eig(A)` : computes spectrum $\sigma(\mathbf{A}) = \{d_1, \ldots, d_n\}$ of $\mathbf{A} \in \mathbb{C}^{n,n}$
`[V,D] = eig(A)` : computes $\mathbf{V} \in \mathbb{C}^{n,n}$, *diagonal* $\mathbf{D} \in \mathbb{C}^{n,n}$ such that $\mathbf{A}\mathbf{V} = \mathbf{V}\mathbf{D}$

Remark 9.2.0.1 (QR-Algorithm → [GV89, Sect. 7.5], [NS02, Sect. 10.3],[Han02, Ch. 26],[QSS00, Sect. 5.5-5.7])

Note: All "direct" eigensolvers are iterative methods



Idea: Iteration based on successive **unitary** similarity transformations



(superior stability of unitary transformations, see **??**)

QR-algorithm (with shift)

✦ in general: quadratic convergence
✦ cubic convergence for normal matrices
(→ [GV89, Sect. 7.5, 8.2])

MATLAB-code 9.2.0.2: QR-algorithm with shift

Computational cost: $O(n^3)$ operations per step of the QR-algorithm

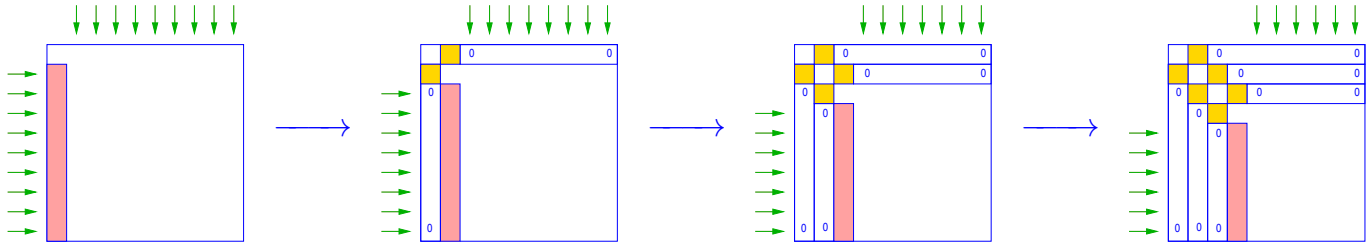
Library implementations of the QR-algorithm provide *numerically stable* eigensolvers (→ Def.1.5.5.19)

⌟

Remark 9.2.0.3 (Unitary similarity transformation to tridiagonal form)

Successive Householder similarity transformations of $\mathbf{A} = \mathbf{A}^H$:

(→ ≙ affected rows/columns, ▭ ≙ targeted vector)



transformation to tridiagonal form **!** (for general matrices a similar strategy can achieve a similarity transformation to upper Hessenberg form)

this transformation is used as a preprocessing step for QR-algorithm ➢ `eig`. ⌟

Similar functionality for generalized EVP $\mathbf{A}\mathbf{x} = \lambda\mathbf{B}\mathbf{x}$, $\mathbf{A},\mathbf{B} \in \mathbb{C}^{n,n}$

`d = eig(A,B)` : computes all generalized eigenvalues
`[V,D] = eig(A,B)` : computes $\mathbf{V} \in \mathbb{C}^{n,n}$, *diagonal* $\mathbf{D} \in \mathbb{C}^{n,n}$ such that $\mathbf{AV} = \mathbf{BVD}$

Note: (Generalized) eigenvectors can be recovered as columns of $\mathbf{V}$:

$$\mathbf{AV} = \mathbf{VD} \quad \Leftrightarrow \quad \mathbf{A}(\mathbf{V})_{:,i} = (\mathbf{D})_{i,i}\mathbf{V}_{:,i} \,,$$

if $\mathbf{D} = \operatorname{diag}(d_1,\ldots,d_n)$.

Remark 9.2.0.4 (Computational effort for eigenvalue computations)

Computational effort (#elementary operations) for `eig()`:

eigenvalues & eigenvectors of $\mathbf{A} \in \mathbb{K}^{n,n}$	$\sim 25n^3 + O(n^2)$
only eigenvalues of $\mathbf{A} \in \mathbb{K}^{n,n}$	$\sim 10n^3 + O(n^2)$
eigenvalues and eigenvectors $\mathbf{A} = \mathbf{A}^H \in \mathbb{K}^{n,n}$	$\sim 9n^3 + O(n^2)$
only eigenvalues of $\mathbf{A} = \mathbf{A}^H \in \mathbb{K}^{n,n}$	$\sim \frac{4}{3}n^3 + O(n^2)$
only eigenvalues of tridiagonal $\mathbf{A} = \mathbf{A}^H \in \mathbb{K}^{n,n}$	$\sim 30n^2 + O(n)$

$O(n^3)$!

Note: `eig` not available for sparse matrix arguments
Exception: `d=eig(A)` for sparse *Hermitian* matrices

EXAMPLE 9.2.0.5 (Runtimes of `eig`)

MATLAB-code 9.2.0.6:

```
1 A = rand(500,500); B = A'*A; C = gallery('tridiag',500,1,3,1);
```

➤
- ✦ **A** generic dense matrix
- ✦ **B** symmetric (s.p.d. → Def. 1.1.2.6) matrix
- ✦ **C** s.p.d. *tridiagonal* matrix

MATLAB-code 9.2.0.7: measuring runtimes of `eig`

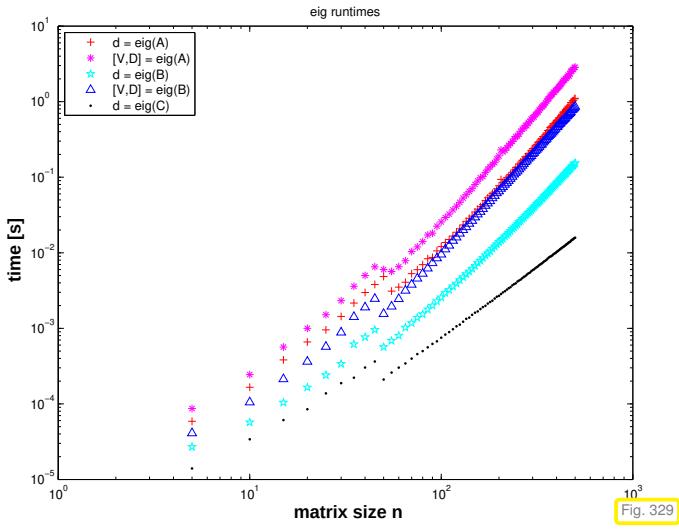


Fig. 328

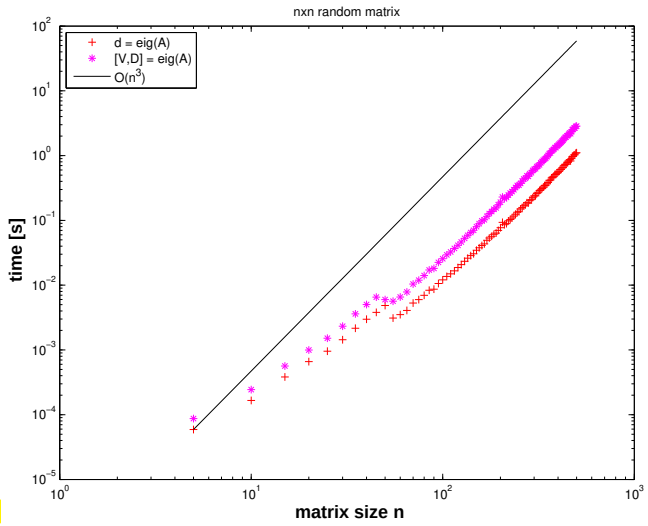


Fig. 329

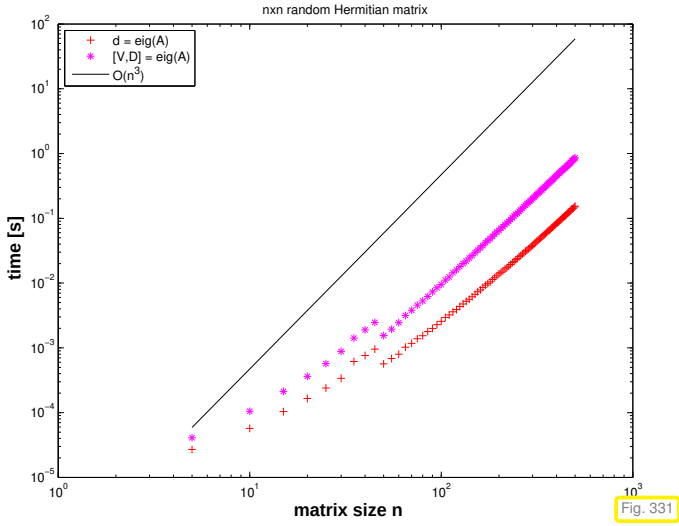


Fig. 330

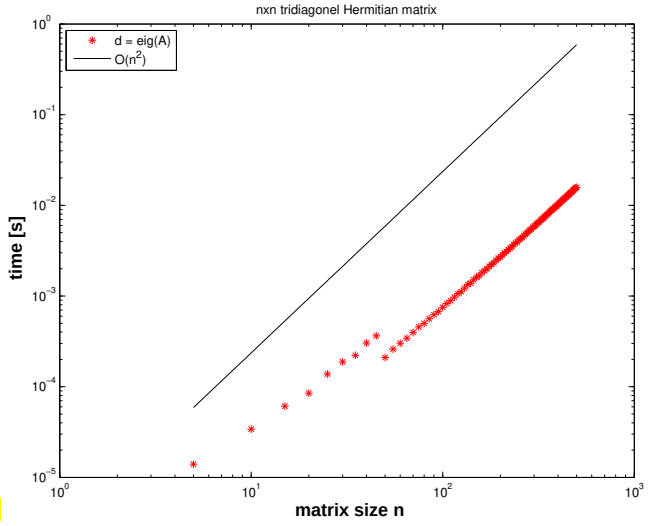


Fig. 331

☛ For the sake of efficiency: think which information you really need when computing eigenvalues/eigenvectors of dense matrices

Potentially more efficient methods for *sparse matrices* will be introduced below in Section 9.3, 9.4.

9.3 Power Methods

9.3.1 Direct power method

Supplementary literature. [DR08, Sect. 7.5], [QSS00, Sect. 5.3.1], [QSS00, Sect. 5.3]

EXAMPLE 9.3.1.1 ((Simplified) Page rank algorithm → [Gle15; LM06])

Model: Random surfer visits a web page, stays there for fixed time Δt, and then
❶ either follows each of ℓ links on a page with probabilty $1/\ell$.
❷ or resumes surfing at a randomly (with equal probability) selected page

Option ❷ is chosen with probability d, $0 \leq d \leq 1$, option ❶ with probability $1-d$.

Stationary Markov chain, state space $\hat{=}$ set of all web pages

Question: Fraction of time spent by random surfer on i-th page (= page rank $x_i \in [0,1]$)

This number $\in]0,1[$ can be used to gauge the "importance" of a web page, which, in turns, offers a way to sort the hits resulting from a keyword query: the GOOGLE idea.

Method: Stochastic simulation ▷

MATLAB-code 9.3.1.2: stochastic page rank simulation

Explanations Code 9.3.1.2:

- **??**: `rand` generates uniformly distributed pseudo-random numbers $\in [0,1[$
- Web graph encoded in $\mathbf{G} \in \{0,1\}^{N,N}$:

$$(\mathbf{G})_{ij} = 1 \quad \Rightarrow \quad \text{link } j \to i \,,$$

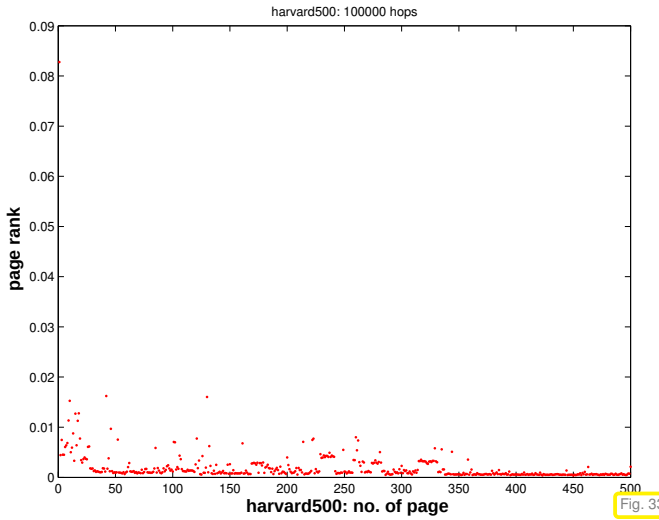


Fig. 332

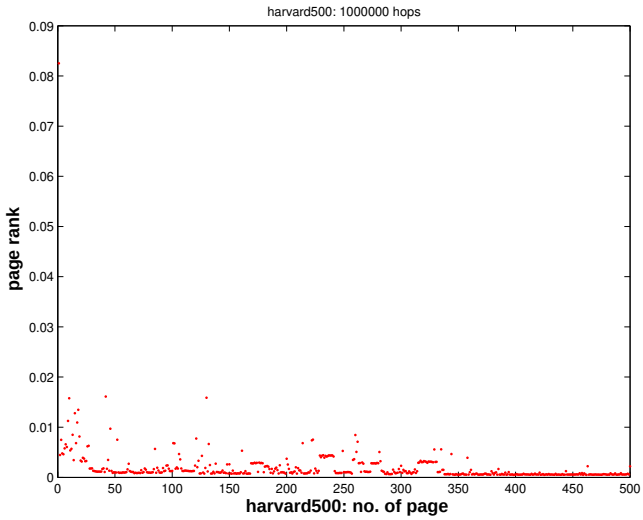


Fig. 333

Observation: relative visit times stabilize as the number of hops in the stochastic simulation $\to \infty$.

The limit distribution is called stationary distribution/invariant measure of the Markov chain. This is what we seek.

- Numbering of pages $1, \ldots, N$, $\ell_i \hat{=}$ number of links from page i

✦ $N \times N$-matrix of transition probabilities page $j \to$ page i: $\mathbf{A} = (a_{ij})_{i,j=1}^{N} \in \mathbb{R}^{N,N}$

$$a_{ij} \in [0,1] \quad \hat{=} \text{ probability to jump from page } j \text{ to page } i.$$
$$\Rightarrow \quad \sum_{i=1}^{N} a_{ij} = 1 \,. \tag{9.3.1.3}$$

A matrix $\mathbf{A} \in [0,1]^{N,N}$ with the property (9.3.1.3) is called a (column) stochastic matrix.

"Meaning" of $\mathbf{A}$: given $\mathbf{x} \in [0,1]^N$, $\|\mathbf{x}\|_1 = 1$, where x_i is the probability of the surfer to visit page i, $i = 1,\ldots,N$, at an instance t in time, $\mathbf{y} = \mathbf{A}\mathbf{x}$ satisfies

$$y_j \geq 0 \quad , \quad \sum_{j=1}^{N} y_j = \sum_{j=1}^{N}\sum_{i=1}^{N} a_{ji}x_i = \sum_{i=1}^{N} x_i \underbrace{\sum_{j=1}^{N} a_{ij}}_{=1} = \sum_{i=1}^{N} x_i = 1 \,.$$

$y_j \hat{=}$ probability for visiting page j at time $t + \Delta t$.

Transition probability matrix for page rank computation

$$(\mathbf{A})_{ij} = \begin{cases} \frac{1}{N} & \text{, if } (\mathbf{G})_{ij} = 0 \;\forall i = 1,\ldots,N \,, \\ d/N + (1-d)\dfrac{(\mathbf{G})_{ij}}{\ell_j} & \text{else.} \end{cases} \tag{9.3.1.4}$$

random jump to any other page — follow link

MATLAB-code 9.3.1.5: transition probability matrix for page rank

Note: special treatment of zero columns of $\mathbf{G}$, *cf.* (9.3.1.4)!

Stochastic simulation based on a single surfer is *slow*. Alternatives?

Thought experiment: Instead of a single random surfer we may consider $m \in \mathbb{N}$, $m \gg 1$, of them who visit pages independently. The fraction of time $m \cdot T$ they all together spend on page i will obviously be the same for $T \to \infty$ as that for a single random surfer.

Instead of counting the surfers we watch the proportions of them visiting particular web pages at an instance of time. Thus, after the k-th hop we can assign a number $x_i^{(k)} \in [0,1]$ to web page i, which gives the proportion of surfers currently on that page: $x_i^{(k)} := \frac{n_i^{(k)}}{m}$, where $n_i^{(k)} \in \mathbb{N}_0$ designates the number of surfers on page i after the k-th hop.

Now consider $m \to \infty$. The law of law of large numbers suggests that the ("infinitely many") surfers visiting page j will move on to other pages proportional to the transistion probabilities a_{ij}: in terms of proportions, for $m \to \infty$ the stochastic evolution becomes a deterministic discrete dynamical system and we find

$$x_i^{(k+1)} = \sum_{j=1}^{N} a_{ij} x_j^{(k)} \,, \tag{9.3.1.6}$$

that is, the proportion of surfers ending up on page i equals the sum of the proportions on the "source pages" weighted with the transition probabilities.

Notice that (9.3.1.6) amounts to matrix×vector. Thus, writing $\mathbf{x}^{(0)} \in [0,1]^N$, $\left\|\mathbf{x}^{(0)}\right\| = 1$ for the initial distribution of the surfers on the net we find

$$\mathbf{x}^{(k)} = \mathbf{A}^k \mathbf{x}^{(0)}$$

will be their mass distribution after k hops. If the limit exists, the i-th component of $\mathbf{x}^* := \lim_{k\to\infty} \mathbf{x}^{(k)}$ tells us which fraction of the (infinitely many) surfers will be visiting page i most of the time. Thus, $\mathbf{x}^*$ yields the stationary distribution of the Markov chain.

MATLAB-code 9.3.1.7: tracking fractions of many surfers

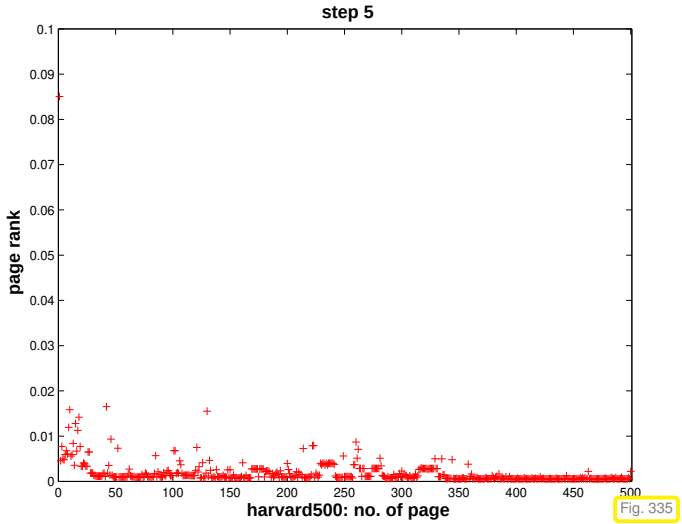


Fig. 334

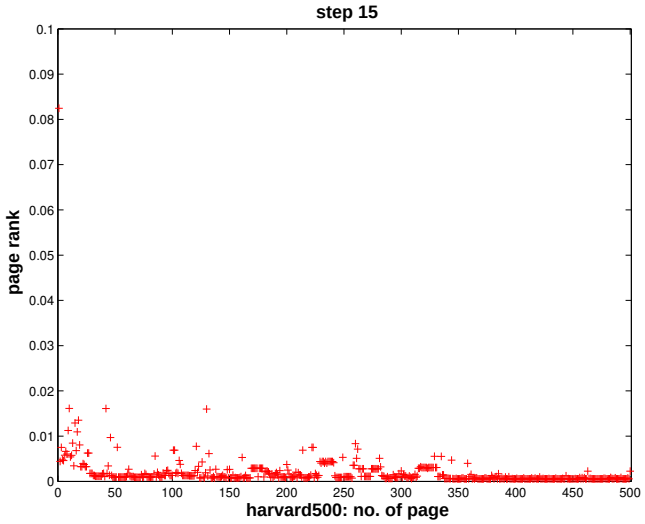


Fig. 335

Comparison:

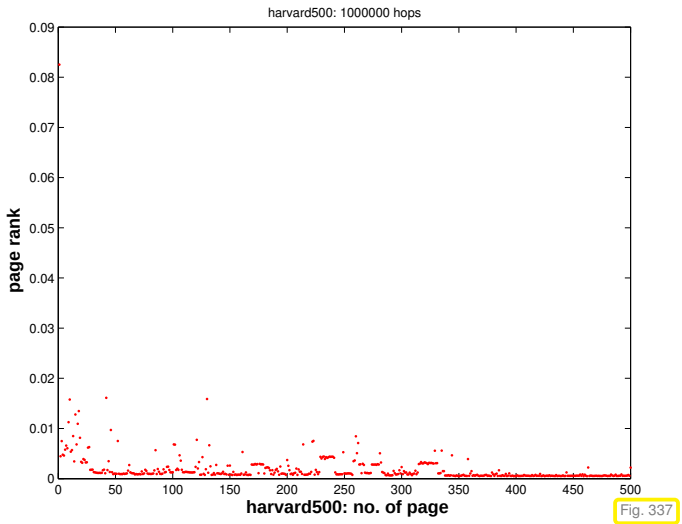


Fig. 336

Single surfer stochastic simulation

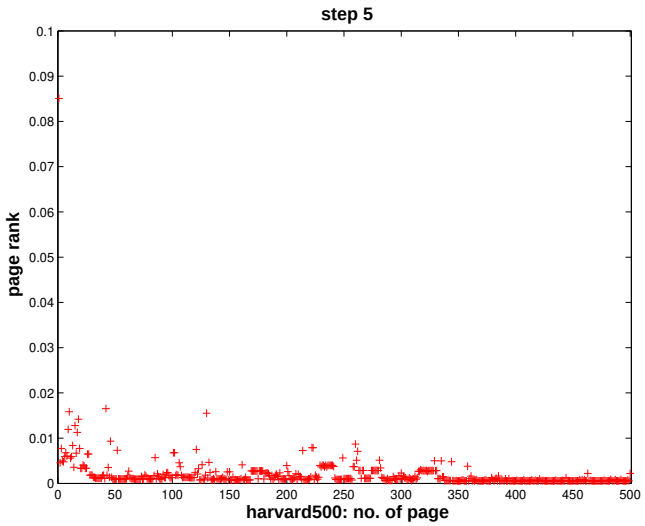


Fig. 337

Power method, Code 9.3.1.7

Observation: Convergence of the $\mathbf{x}^{(k)} \to \mathbf{x}^*$, and the limit must be a fixed point of the iteration function:

$$\succ \quad \mathbf{A}\mathbf{x}^* = \mathbf{x}^* \quad \Rightarrow \quad \mathbf{x}^* \in \mathbf{EigA}1\,.$$

Does $\mathbf{A}$ possess an eigenvalue $= 1$? Does the associated eigenvector really provide a probability distribution (after scaling), that is, are all of its entries non-negative? Is this probability distribution unique? To answer these questions we have to study the matrix $\mathbf{A}$:

For every stochastic matrix $\mathbf{A}$, by definition (9.3.1.3)

$$\mathbf{A}^T\mathbf{1} = \mathbf{1} \overset{(9.1.0.3)}{\Rightarrow} 1 \in \sigma(\mathbf{A}) ,$$
$$(1.5.5.14) \Rightarrow \|\mathbf{A}\|_1 = 1 \overset{\text{Thm. 9.1.0.4}}{\Rightarrow} \rho(\mathbf{A}) = 1 ,$$

where $\rho(\mathbf{A})$ is the spectral radius of the matrix $\mathbf{A}$, see Section 9.1.

For $\mathbf{r} \in \mathbf{Eig}\mathbf{A}1$, that is, $\mathbf{Ar} = \mathbf{r}$, denote by $|\mathbf{r}|$ the vector $(|r_i|)_{i=1}^N$. Since all entries of $\mathbf{A}$ are non-negative, we conclude by the triangle inequality that $\|\mathbf{Ar}\|_1 \leq \|\mathbf{A}|\mathbf{r}|\|_1$

$$\Rightarrow 1 = \|\mathbf{A}\|_1 = \sup_{\mathbf{x}\in\mathbb{R}^N} \frac{\|\mathbf{Ax}\|_1}{\|\mathbf{x}\|_1} \geq \frac{\|\mathbf{A}|\mathbf{r}|\|_1}{\||\mathbf{r}|\|_1} \geq \frac{\|\mathbf{Ar}\|_1}{\|\mathbf{r}\|_1} = 1 .$$
$$\Rightarrow \|\mathbf{A}|\mathbf{r}|\|_1 = \|\mathbf{Ar}\|_1 \overset{\text{if } a_{ij}>0}{\Rightarrow} |\mathbf{r}| = \pm\mathbf{r} .$$

Hence, different components of $\mathbf{r}$ cannot have opposite sign, which means, that $\mathbf{r}$ *can be chosen to have non-negative entries*, if the entries of $\mathbf{A}$ are strictly positive, which is the case for $\mathbf{A}$ from (9.3.1.4). After normalization $\|\mathbf{r}\|_1 = 1$ the eigenvector can be regarded as a probability distribution on $\{1,\ldots,N\}$.

If $\mathbf{Ar} = \mathbf{r}$ and $\mathbf{As} = \mathbf{s}$ with $(\mathbf{r})_i \geq 0$, $(\mathbf{s})_i \geq 0$, $\|\mathbf{r}\|_1 = \|\mathbf{s}\|_1 = 1$, then $\mathbf{A}(\mathbf{r}-\mathbf{s}) = \mathbf{r}-\mathbf{s}$. Hence, by the above considerations, also all the entries of $\mathbf{r}-\mathbf{s}$ are either non-negative or non-positive. By the assumptions on $\mathbf{r}$ and $\mathbf{s}$ this is only possible, if $\mathbf{r}-\mathbf{s} = 0$. We conclude that

$$\mathbf{A} \in]0,1]^{N,N} \text{ stochastic} \quad \Rightarrow \quad \dim \mathbf{Eig}\mathbf{A}1 = 1 . \tag{9.3.1.8}$$

Sorting the pages according to the size of the corresponding entries in $\mathbf{r}$ yields the famous "page rank".

Plot of entries of *unique* vector $\mathbf{r} \in \mathbb{R}^N$ with

$$0 \leq (\mathbf{r})_i \leq 1 ,$$
$$\|\mathbf{r}\|_1 = 1 ,$$
$$\boxed{\mathbf{Ar} = \mathbf{r}} .$$

Inefficient implementation!

MATLAB-code 9.3.1.9: computing page rank vector $\mathbf{r}$ via `eig`

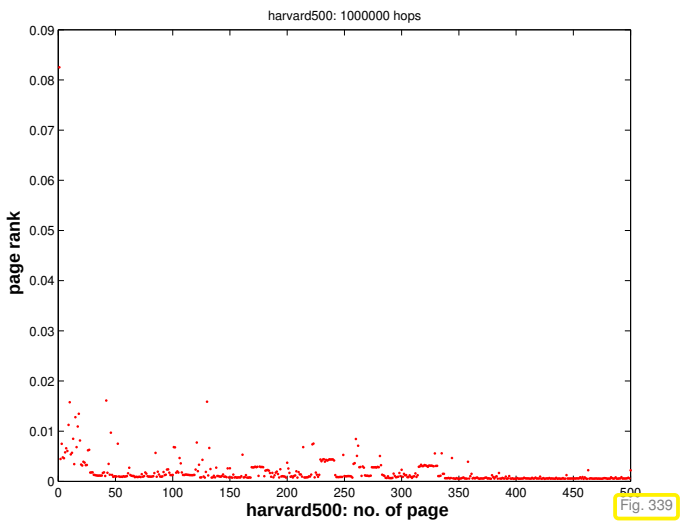


Fig. 338

stochastic simulation

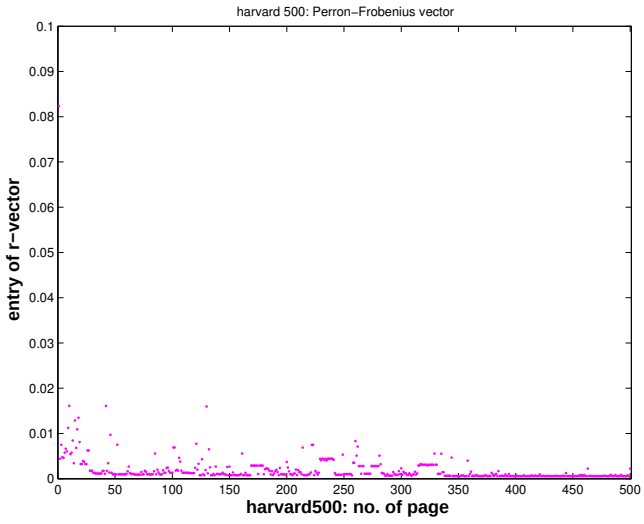


Fig. 339

eigenvector computation

The possibility to compute the stationary probability distribution of a Markov chain through an eigenvector of the transition probability matrix is due to a property of stationary Markov chains called ergodicity.

$\mathbf{A} \hat{=}$ page rank transition probability matrix, see Code 9.3.1.5, $d = 0.15$, harvard500 example.

Errors: ▷

$$\left\| \mathbf{A}^k \mathbf{x}_0 - \mathbf{r} \right\|_1 ,$$

with $\mathbf{x}_0 = \mathbf{1}/N$, $N = 500$.
We observe linear convergence! (→ Def. 8.2.2.1, iteration error vs. iteration count ≈ straight line lin-log plot)

⌟

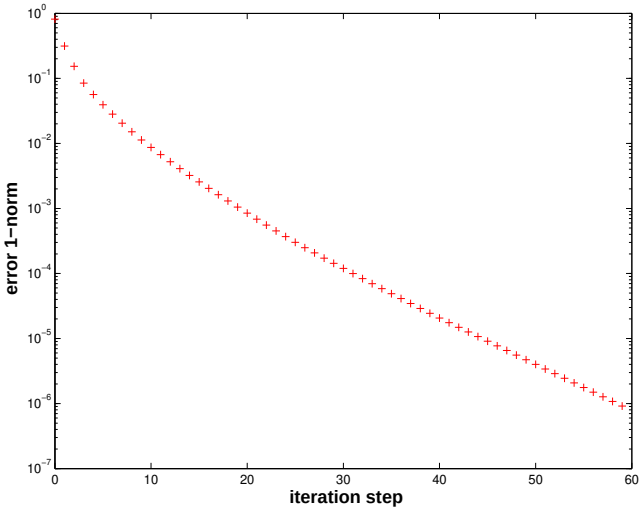
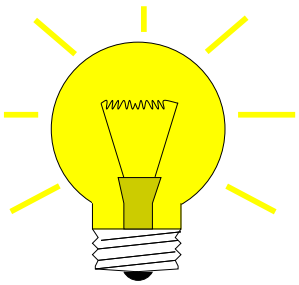


Fig. 340

The computation of page rank amounts to finding the eigenvector of the matrix $\mathbf{A}$ of transition probabilities that belongs to its *largest* eigenvalue 1. This is addressed by an important class of practical eigenvalue problems:

Task:	given $\mathbf{A} \in \mathbb{K}^{n,n}$, find largest (in modulus) eigenvalue of $\mathbf{A}$ and (an) associated eigenvector.



Idea: (suggested by page rank computation, Code 9.3.1.7)
Iteration: $\mathbf{z}^{(k+1)} = \mathbf{A}\mathbf{z}^{(k)}$, $\mathbf{z}^{(0)}$ arbitrary

EXAMPLE 9.3.1.10 (Power iteration → Ex. 9.3.1.1)

Try the above iteration for general 10×10-matrix, largest eigenvalue 10, algebraic multiplicity 1.

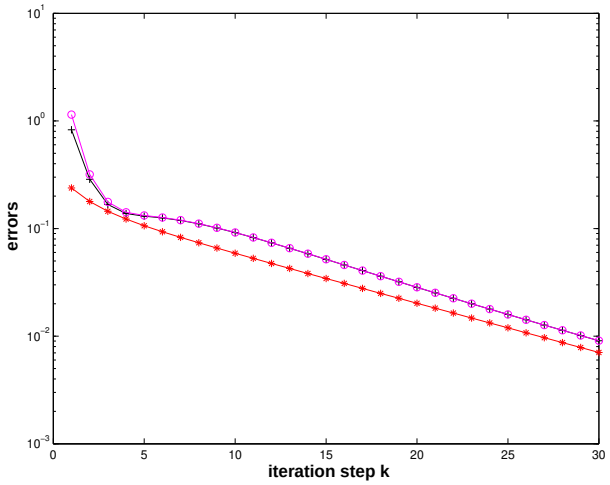


Fig. 341

MATLAB-code 9.3.1.11:

```
d = (1:10)'; n = length(d);
S = triu(diag(n:-1:1,0)+...
    ones(n,n));
A = S*diag(d,0)*inv(S);
```

◁ error norm $\left\| \frac{\mathbf{z}^{(k)}}{\|\mathbf{z}^{(k)}\|} - (\mathbf{S})_{:,10} \right\|$

(Note: $(\mathbf{S})_{:,10} \hat{=}$ eigenvector for eigenvalue 10)

$\mathbf{z}^{(0)}$ = random vector

Observation: linear convergence of (normalized) eigenvectors!

Suggests direct power method (*ger.:* Potenzmethode): iterative method (→ Section 8.2)

$$\begin{aligned} &\text{initial guess:} \quad \mathbf{z}^{(0)} \text{ "arbitrary"}\,, \\ &\text{next iterate:} \quad \mathbf{w} := \mathbf{A}\mathbf{z}^{(k-1)}\,, \quad \mathbf{z}^{(k)} := \frac{\mathbf{w}}{\|\mathbf{w}\|_2}\,, \quad k = 1,2,\ldots\,. \end{aligned} \tag{9.3.1.12}$$

Note: the "normalization" of the iterates in (9.3.1.12) does not change anything (in exact arithmetic) and helps *avoid overflow* in floating point arithmetic.

Computational effort: $1\times$ matrix$\times$vector per step ➢ inexpensive for sparse matrices

A persuasive theoretical justification for the direct power method:

Assume $\mathbf{A} \in \mathbb{K}^{n,n}$ to be diagonalizable:

$$\Leftrightarrow \quad \exists \text{ basis } \{\mathbf{v}_1,\ldots,\mathbf{v}_n\} \text{ of eigenvectors of } \mathbf{A}\colon \quad \mathbf{A}\mathbf{v}_j = \lambda_j \mathbf{v}_j,\ \lambda_j \in \mathbb{C}.$$

Assume

$$|\lambda_1| \le |\lambda_2| \le \cdots \le |\lambda_{n-1}| < |\lambda_n| \quad , \quad \|\mathbf{v}_j\|_2 = 1\,. \tag{9.3.1.13}$$

Key observations for power iteration (9.3.1.12)

$$\mathbf{z}^{(k)} = \mathbf{A}^k \mathbf{z}^{(0)} \qquad (\rightarrow \text{name "power method"}) \tag{9.3.1.14}$$

$$\mathbf{z}^{(0)} = \sum_{j=1}^{n} \zeta_j \mathbf{v}_j \quad \Rightarrow \quad \mathbf{z}^{(k)} = \sum_{j=1}^{n} \zeta_j \lambda_j^k \mathbf{v}_j\,. \tag{9.3.1.15}$$

Due to (9.3.1.13) for large $k \gg 1$ ($\Rightarrow |\lambda_n^k| \gg |\lambda_j^k|$ for $j \neq n$) the contribution of $\mathbf{v}_n$ (size $\zeta_n \lambda_n^k$) in the *eigenvector expansion* (9.3.1.15) will be much larger than the contribution (size $\zeta_n \lambda_j^k$) of any other eigenvector (, if $\zeta_n \neq 0$): the eigenvector for λ_n will swamp all other for $k \to \infty$.

Further (9.3.1.15) nutures expectation: $\mathbf{v}_n$ will become dominant in $\mathbf{z}^{(k)}$ the faster, the better $|\lambda_n|$ is separated from $|\lambda_{n-1}|$, see Thm. 9.3.1.21 for rigorous statement.

$\mathbf{z}^{(k)} \to$ eigenvector, but how do we get the associated eigenvalue λ_n **?**

When (9.3.1.12) has converged, two common ways to recover $\lambda_{\max}$ → [DR08, Alg. 7.20]

❶ $\mathbf{A}\mathbf{z}^{(k)} \approx \lambda_{\max}\mathbf{z}^{(k)}$ ➢ $|\lambda_n| \approx \frac{\|\mathbf{A}\mathbf{z}^{(k)}\|}{\|\mathbf{z}^{(k)}\|}$ (modulus only!)

❷ $\lambda_{\max} \approx \operatorname{argmin}_{\theta\in\mathbb{R}} \left\|\mathbf{A}\mathbf{z}^{(k)} - \theta\mathbf{z}^{(k)}\right\|_2^2$ ➤ $\lambda_{\max} \approx \frac{(\mathbf{z}^{(k)})^H \mathbf{A}\mathbf{z}^{(k)}}{\|\mathbf{z}^{(k)}\|_2^2}$.

This latter formula is extremely useful, which has earned it a special name:

Definition 9.3.1.16.

For $\mathbf{A} \in \mathbb{K}^{n,n}$, $\mathbf{u} \in \mathbb{K}^n$ the Rayleigh quotient is defined by

$$\rho_{\mathbf{A}}(\mathbf{u}) := \frac{\mathbf{u}^H \mathbf{A}\mathbf{u}}{\mathbf{u}^H \mathbf{u}} .$$

An immediate consequence of the definitions:

$$\lambda \in \sigma(\mathbf{A}) \quad , \quad \mathbf{z} \in \mathbf{Eig}\lambda\mathbf{A} \quad \Rightarrow \quad \rho_{\mathbf{A}}(\mathbf{z}) = \lambda . \tag{9.3.1.17}$$

EXAMPLE 9.3.1.18 (Direct power method → Ex. 9.3.1.18 cnt'd)

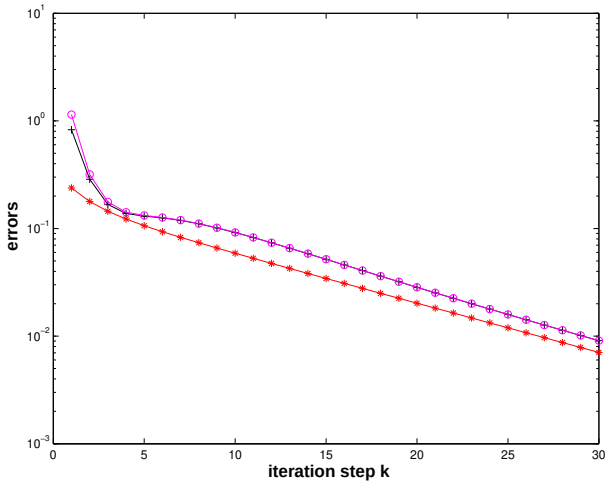


Fig. 342

MATLAB-code 9.3.1.19:
n = length(d); S = triu(diag(n:-1:1,0)+...
ones(n,n)); A = S*diag(d,0)*inv(S);

◁ `d = (1:10)';`
o : error $|\lambda_n - \rho_{\mathbf{A}}(\mathbf{z}^{(k)})|$
$*$: error norm $\left\|\mathbf{z}^{(k)} - \mathbf{s}_{\cdot,n}\right\|$
$+$: $\left|\lambda_n - \frac{\|\mathbf{A}\mathbf{z}^{(k-1)}\|_2}{\|\mathbf{z}^{(k-1)}\|_2}\right|$

$\mathbf{z}^{(0)}$ = random vector

Test matrices:

① `d=(1:10)';` ➢ $|\lambda_{n-1}| : |\lambda_n| = 0.9$
② `d = [ones(9,1); 2];` ➢ $|\lambda_{n-1}| : |\lambda_n| = 0.5$
③ `d = 1-2.^(-(1:0.5:5)');` ➢ $|\lambda_{n-1}| : |\lambda_n| = 0.9866$

MATLAB-code 9.3.1.20: Investigating convergence of direct power method

$$\rho_{\mathrm{EV}}^{(k)} := \frac{\left\|\mathbf{z}^{(k)} - \mathbf{s}_{\cdot,n}\right\|}{\left\|\mathbf{z}^{(k-1)} - \mathbf{s}_{\cdot,n}\right\|} ,$$

$$\rho_{\mathrm{EW}}^{(k)} := \frac{|\rho_{\mathbf{A}}(\mathbf{z}^{(k)}) - \lambda_n|}{|\rho_{\mathbf{A}}(\mathbf{z}^{(k-1)}) - \lambda_n|} .$$

	①		②		③	
k	$\rho_{\mathrm{EV}}^{(k)}$	$\rho_{\mathrm{EW}}^{(k)}$	$\rho_{\mathrm{EV}}^{(k)}$	$\rho_{\mathrm{EW}}^{(k)}$	$\rho_{\mathrm{EV}}^{(k)}$	$\rho_{\mathrm{EW}}^{(k)}$
22	0.9102	0.9007	0.5000	0.5000	0.9900	0.9781
23	0.9092	0.9004	0.5000	0.5000	0.9900	0.9791
24	0.9083	0.9001	0.5000	0.5000	0.9901	0.9800
25	0.9075	0.9000	0.5000	0.5000	0.9901	0.9809
26	0.9068	0.8998	0.5000	0.5000	0.9901	0.9817
27	0.9061	0.8997	0.5000	0.5000	0.9901	0.9825
28	0.9055	0.8997	0.5000	0.5000	0.9901	0.9832
29	0.9049	0.8996	0.5000	0.5000	0.9901	0.9839
30	0.9045	0.8996	0.5000	0.5000	0.9901	0.9844

Observation: linear convergence (→ **??**) ⌟

Theorem 9.3.1.21. Convergence of direct power method → [DR08, Thm. 25.1]

Let $\lambda_n > 0$ be the largest (in modulus) eigenvalue of $\mathbf{A} \in \mathbb{K}^{n,n}$ and have (algebraic) multiplicity 1. Let $\mathbf{v}, \mathbf{y}$ be the left and right eigenvectors of $\mathbf{A}$ for λ_n normalized according to $\|\mathbf{y}\|_2 = \|\mathbf{v}\|_2 = 1$. Then there is convergence

$$\left\|\mathbf{A}\mathbf{z}^{(k)}\right\|_2 \to \lambda_n \quad , \quad \mathbf{z}^{(k)} \to \pm\mathbf{v} \quad \textit{linearly with rate} \quad \frac{|\lambda_{n-1}|}{|\lambda_n|} \, ,$$

where $\mathbf{z}^{(k)}$ are the iterates of the direct power iteration and $\mathbf{y}^H\mathbf{z}^{(0)} \neq 0$ is assumed.

Remark 9.3.1.22 (Initial guess for power iteration)

roundoff errors ➤ $\mathbf{y}^H\mathbf{z}^{(0)} \neq 0$ always satisfied in practical computations

Usual (not the best!) choice for $\mathbf{x}^{(0)}$ = random vector ⌟

Remark 9.3.1.23 (Termination criterion for direct power iteration) (→ Section 8.2.3)

Adaptation of a posteriori termination criterion (8.3.2.20)

"relative change" $\leq$ `tol`:

$$\begin{cases} \min\left\|\mathbf{z}^{(k)} \pm \mathbf{z}^{(k-1)}\right\| \leq (1/L-1)\texttt{tol} \, , \\ \left| \frac{\|\mathbf{A}\mathbf{z}^{(k)}\|}{\|\mathbf{z}^{(k)}\|} - \frac{\|\mathbf{A}\mathbf{z}^{(k-1)}\|}{\|\mathbf{z}^{(k-1)}\|} \right| \leq (1/L-1)\texttt{tol} \quad \text{see (8.2.3.8)} \, . \end{cases}$$

Estimated rate of convergence ⌟

9.3.2 Inverse Iteration [DR08, Sect. 7.6], [QSS00, Sect. 5.3.2]

EXAMPLE 9.3.2.1 (Image segmentation)

Given: gray-scale image: intensity matrix $\mathbf{P} \in \{0,\ldots,255\}^{m,n}$, $m,n \in \mathbb{N}$
($(\mathbf{P})_{ij} \leftrightarrow$ pixel, $0 \mathrel{\hat{=}}$ black, $255 \mathrel{\hat{=}}$ white)

Loading and displaying images in MATLAB ▷

MATLAB-code 9.3.2.2: loading and displaying an image

(Fuzzy) task: Local segmentation
Find connected patches of image of the same shade/color

More general segmentation problem (non-local): identify parts of the image, not necessarily connected, with the same texture.

Next: Statement of (rigorously defined) problem, *cf.* **??**:

Preparation: Numbering of pixels $1 \ldots, mn$, e.g, lexicographic numbering:
- ✦ pixel set $\mathcal{V} := \{1\ldots, nm\}$
- ✦ indexing: index$(\text{pixel}_{i,j}) = (i-1)n + j$

✎ notation: $p_k := (\mathbf{P})_{ij}$, if $k = \text{index}(\text{pixel}_{i,j}) = (i-1)n + j$, $k = 1,\ldots,N := mn$

Local similarity matrix:

$$\mathbf{W} \in \mathbb{R}^{N,N}, \quad N := mn, \qquad (9.3.2.3)$$

$$(\mathbf{W})_{ij} = \begin{cases} 0 & \text{, if pixels } i, j \text{ not adjacent,} \\ 0 & \text{, if } i = j, \\ \sigma(p_i, p_j) & \text{, if pixels } i, j \text{ adjacent.} \end{cases}$$

$\leftrightarrow \hat{=}$ adjacent pixels ▷

Similarity function, e.g., with $\alpha > 0$

$$\sigma(x,y) := \exp(-\alpha(x-y)^2), \quad x, y \in \mathbb{R}.$$

Lexicographic numbering ▷

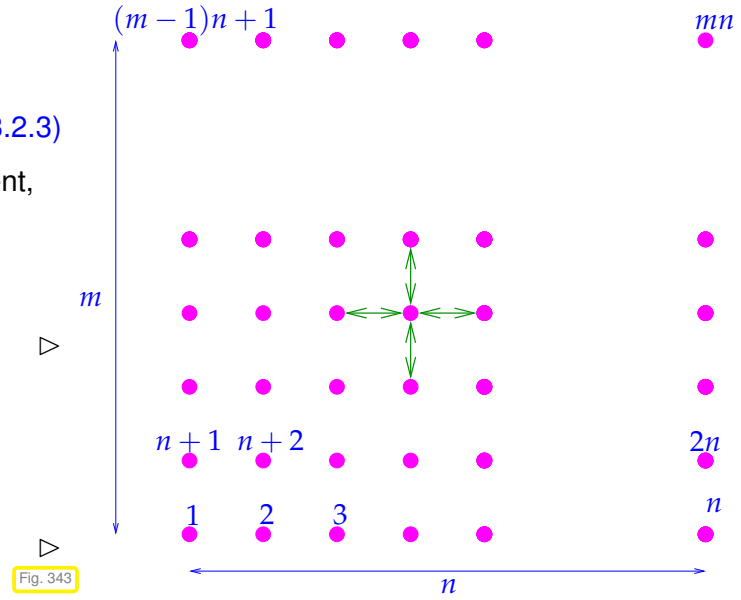


Fig. 343

The entries of the matrix $\mathbf{W}$ measure the "similarity" of neighboring pixels: if $(\mathbf{W})_{ij}$ is large, they encode (almost) the same intensity, if $(\mathbf{W})_{ij}$ is close to zero, then they belong to parts of the picture with very different brightness. In the latter case, the boundary of the segment may separate the two pixels.

Definition 9.3.2.4. Normalized cut (→ [SM00, Sect. 2])

For $\mathcal{X} \subset \mathcal{V}$ define the normalized cut as

$$\mathrm{Ncut}(\mathcal{X}) := \frac{\mathrm{cut}(\mathcal{X})}{\mathrm{weight}(\mathcal{X})} + \frac{\mathrm{cut}(\mathcal{X})}{\mathrm{weight}(\mathcal{V} \setminus \mathcal{X})},$$

$$\text{with} \quad \mathrm{cut}(\mathcal{X}) := \sum_{i \in \mathcal{X}, j \notin \mathcal{X}} w_{ij}, \quad \mathrm{weight}(\mathcal{X}) := \sum_{i \in \mathcal{X}, j \in \mathcal{X}} w_{ij}.$$

In light of local similarity relationship:

- $\mathrm{cut}(\mathcal{X})$ big ➢ substantial similarity of pixels across interface between $\mathcal{X}$ and $\mathcal{V} \setminus \mathcal{X}$.
- $\mathrm{weight}(\mathcal{X})$ big ➢ a lot of similarity of adjacent pixels inside $\mathcal{X}$.

➢ Segmentation problem (rigorous statement):

$$\text{find} \quad \mathcal{X}^* \subset \mathcal{V}: \quad \mathcal{X}^* = \operatorname*{argmin}_{\mathcal{X} \subset \mathcal{V}} \mathrm{Ncut}(\mathcal{X}). \qquad (9.3.2.5)$$

NP-hard combinatorial optimization problem !

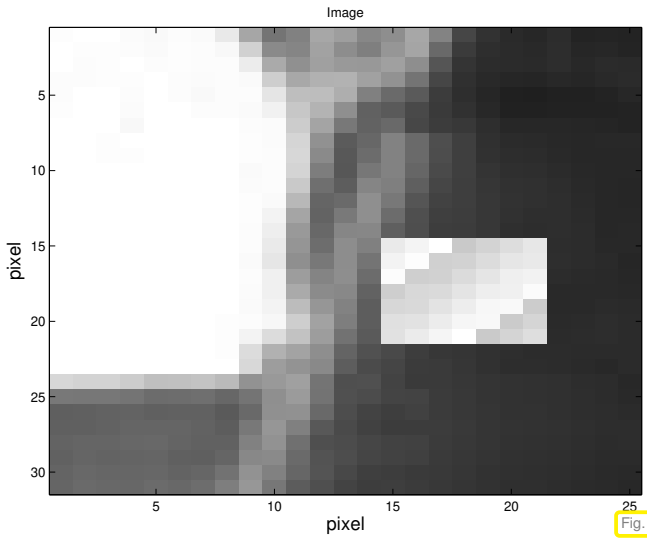


Fig. 344

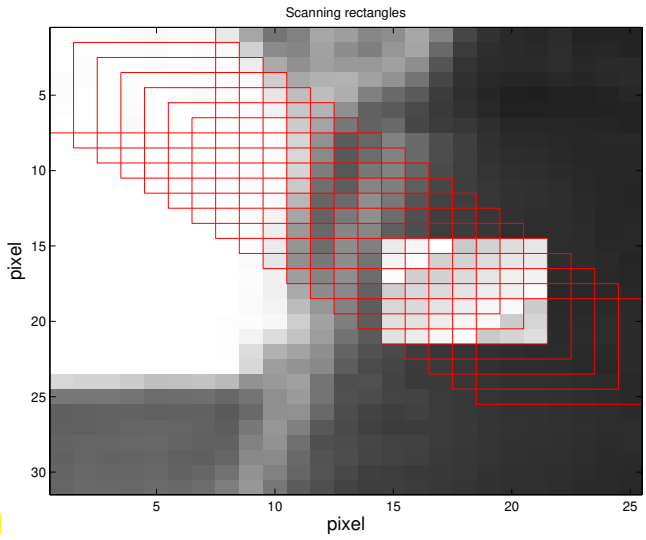


Fig. 345

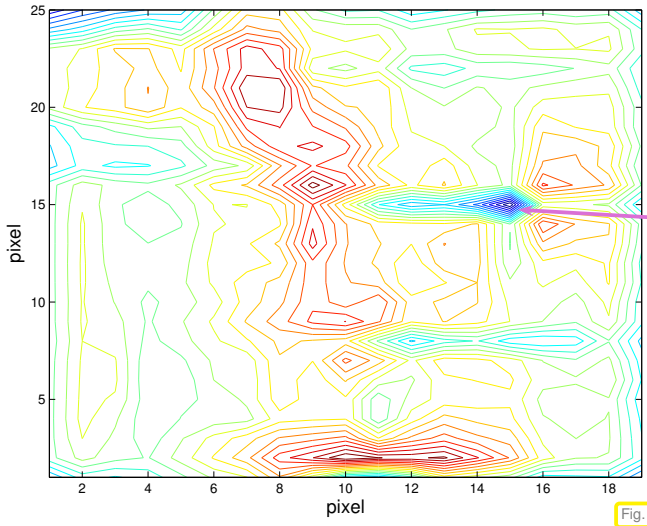


Fig. 346

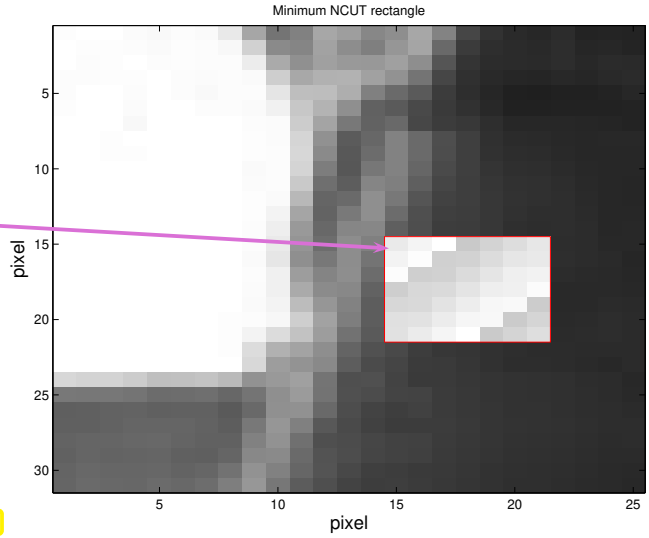


Fig. 347

△ $\mathrm{Ncut}(\mathcal{X})$ for pixel subsets $\mathcal{X}$ defined by sliding rectangles, see Fig. 345.

Equivalent reformulation:

indicator function: $$z : \{1,\ldots,N\} \mapsto \{-1,1\}\,, \quad z_i := z(i) = \begin{cases} 1 & \text{, if } i \in \mathcal{X}\,, \\ -1 & \text{, if } i \notin \mathcal{X}\,. \end{cases} \tag{9.3.2.6}$$

$$\mathrm{Ncut}(\mathcal{X}) = \frac{\sum\limits_{z_i>0, z_j<0} -w_{ij} z_i z_j}{\sum\limits_{z_i>0} d_i} + \frac{\sum\limits_{z_i>0, z_j<0} -w_{ij} z_i z_j}{\sum\limits_{z_i<0} d_i}\,, \tag{9.3.2.7}$$

$$d_i = \sum_{j\in\mathcal{V}} w_{ij} = \mathrm{weight}(\{i\})\,. \tag{9.3.2.8}$$

Sparse matrices:

$$\mathbf{D} := \mathrm{diag}(d_1,\ldots,d_N) \in \mathbb{R}^{N,N}\,, \quad \mathbf{A} := \mathbf{D} - \mathbf{W} = \mathbf{A}^\top\,. \tag{9.3.2.9}$$

Summary: (obvious) properties of these matrices

- $\mathbf{A}$ has positive diagonal and non-positive off-diagonal entries.
- $\mathbf{A}$ is diagonally dominant (→ Def. 2.8.0.8) ➢ $\mathbf{A}$ is positive semidefinite by Lemma 2.8.0.13.
- $\mathbf{A}$ has row sums $= 0$:

$$\mathbf{1}^\top \mathbf{A} = \mathbf{A}\mathbf{1} = 0\,. \tag{9.3.2.10}$$

MATLAB-code 9.3.2.11: assembly of $\mathbf{A}$, $\mathbf{D}$

Lemma 9.3.2.12. Ncut and Rayleigh quotient (→ [SM00, Sect. 2])

With $\mathbf{z} \in \{-1,1\}^N$ according to (9.3.2.6) *there holds*

$$\mathrm{Ncut}(\mathcal{X}) = \frac{\mathbf{y}^\top \mathbf{A}\mathbf{y}}{\mathbf{y}^\top \mathbf{D}\mathbf{y}} \quad , \quad \mathbf{y} := (\mathbf{1}+\mathbf{z}) - \beta(\mathbf{1}-\mathbf{z}) \, , \quad \beta := \frac{\sum\limits_{z_i>0} d_i}{\sum\limits_{z_i<0} d_i} \, .$$

generalized Rayleigh quotient $\rho_{\mathbf{A},\mathbf{D}}(\mathbf{y})$

Proof. Note that by (9.3.2.6) $(\mathbf{1}-\mathbf{z})_i = 0 \;\Leftrightarrow\; i \in \mathcal{X}$, $(\mathbf{1}+\mathbf{z})_i = 0 \;\Leftrightarrow\; i \notin \mathcal{X}$. Hence, since $(\mathbf{1}+\mathbf{z})^\top \mathbf{D}(\mathbf{1}-\mathbf{z}) = 0$,

$$4\,\mathrm{Ncut}(\mathcal{X}) = (\mathbf{1}+\mathbf{z})^\top \mathbf{A}(\mathbf{1}+\mathbf{z})\left(\frac{1}{\kappa \mathbf{1}^\top \mathbf{D}\mathbf{1}} + \frac{1}{(1-\kappa)\mathbf{1}^\top \mathbf{D}\mathbf{1}}\right) = \frac{\mathbf{y}^\top \mathbf{A}\mathbf{y}}{\beta \mathbf{1}^\top \mathbf{D}\mathbf{1}} \, ,$$

where $\kappa := \sum_{z_i>0} d_i / \sum_i d_i = \frac{\beta}{1+\beta}$. Also observe

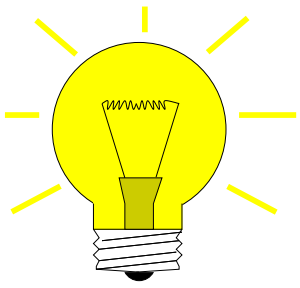
$$\mathbf{y}^\top \mathbf{D}\mathbf{y} = (\mathbf{1}+\mathbf{z})^\top \mathbf{D}(\mathbf{1}+\mathbf{z}) + \beta^2 (\mathbf{1}-\mathbf{z})^\top \mathbf{D}(\mathbf{1}-\mathbf{z}) = 4(\sum_{z_i>0} d_i + \beta^2 \sum_{z_i<0} d_i) = 4\beta \sum_i d_i = 4\beta \mathbf{1}^\top \mathbf{D}\mathbf{1} \, .$$

This finishes the proof. □

- (9.3.2.10) ⇒ $\mathbf{1} \in \mathbf{EigA}0$
- Lemma 2.8.0.13: $\mathbf{A}$ diagonally dominant ⟹ $\mathbf{A}$ is *positive semidefinite* (→ Def. 1.1.2.6)

➢ $\mathrm{Ncut}(\mathcal{X}) \geq 0$ and 0 is the smallest eigenvalue of $\mathbf{A}$.

However, we are by no means interested in a minimizer $\mathbf{y} \in \mathrm{Span}\{\mathbf{1}\}$ (with constant entries) that does not provide a meaningful segmentation.



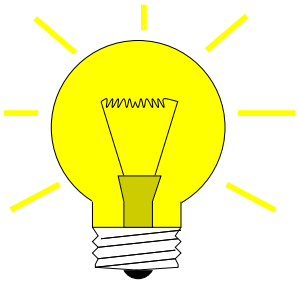
Idea: weed out undesirable constant minimizers by imposing orthogonality constraint (orthogonality w.r.t. inner product induced by $\mathbf{D}$, *cf.* Section 10.1)

$$\mathbf{y} \perp \mathbf{D}\mathbf{1} \;\Leftrightarrow\; \mathbf{1}^\top \mathbf{D}\mathbf{y} = 0\,. \tag{9.3.2.13}$$

segmentation problem (9.3.2.5) $\Leftrightarrow$ $\operatorname{argmin}_{\mathbf{y}\in\{2,-2\beta\}^N,\mathbf{1}^\top\mathbf{D}\mathbf{y}=0} \rho_{\mathbf{A},\mathbf{D}}(\mathbf{y})$. (9.3.2.14)

still NP-hard

➢ Minimizing $\mathrm{Ncut}(\mathcal{X})$ amounts to minimizing a (generalized) Rayleigh quotient (→ Def. 9.3.1.16) over a *discrete* set of vectors, which is still an NP-hard problem.



Idea: Relaxation

Discrete optimization problem → continuous optimization problem

(9.3.2.14) → $\operatorname{argmin}_{\mathbf{y}\in\mathbb{R}^N,\mathbf{y}\neq 0,\mathbf{1}^\top\mathbf{D}\mathbf{y}=0} \rho_{\mathbf{A},\mathbf{D}}(\mathbf{y})$. (9.3.2.15)

Task: (9.3.2.15) $\Leftrightarrow$ Find minimizer of (generalized) Rayleigh quotient under linear constraint

Here: linear constraint on $\mathbf{y}$: $\mathbf{1}^\top\mathbf{D}\mathbf{y} = 0$

The next theorem establishes a link between argument vectors that render the Rayleigh quotient extremal and eigenspace for extremal eigenvalues.

Theorem 9.3.2.16. Rayleigh quotient theorem

Let $\lambda_1 < \lambda_2 < \cdots < \lambda_m$, $m \le n$, be the sorted sequence of all (real!) eigenvalues of $\mathbf{A} = \mathbf{A}^{\mathrm{H}} \in \mathbb{C}^{n,n}$. Then

$$\mathbf{EigA}\lambda_1 = \operatorname*{argmin}_{\mathbf{y}\in\mathbb{C}^{n.n}\setminus\{0\}} \rho_{\mathbf{A}}(\mathbf{y}) \quad \textit{and} \quad \mathbf{EigA}\lambda_m = \operatorname*{argmax}_{\mathbf{y}\in\mathbb{C}^{n.n}\setminus\{0\}} \rho_{\mathbf{A}}(\mathbf{y}) .$$

Remark 9.3.2.17 (Min-max theorem)

Thm. 9.3.2.16 is a an immediate consequence of the following more general and fundamentally important result.

Theorem 9.3.2.18. Courant-Fischer min-max theorem → [GV89, Thm. 8.1.2]

Let $\lambda_1 < \lambda_2 < \cdots < \lambda_m$, $m \le n$, be the sorted sequence of the (real!) eigenvalues of $\mathbf{A} = \mathbf{A}^{\mathrm{H}} \in \mathbb{C}^{n,n}$. Write

$$U_0 = \{0\} , \quad U_\ell := \sum_{j=1}^{\ell} \mathbf{EigA}\lambda_j , \ell = 1,\ldots,m \quad \textit{and} \quad U_\ell^\perp := \{\mathbf{x}\in\mathbb{C}^n : \mathbf{u}^{\mathrm{H}}\mathbf{x} = 0 \ \forall \mathbf{u}\in U_\ell\} .$$

Then

$$\min_{\mathbf{y}\in U_{\ell-1}^\perp\setminus\{0\}} \rho_{\mathbf{A}}(\mathbf{y}) = \lambda_\ell , \quad 1 \le \ell \le m , \quad \operatorname*{argmin}_{\mathbf{y}\in U_{\ell-1}^\perp\setminus\{0\}} \rho_{\mathbf{A}}(\mathbf{y}) \subset \mathbf{EigA}\lambda_\ell .$$

Proof. For diagonal $\mathbf{A} \in \mathbb{R}^{n,n}$ the assertion of the theorem is obvious. Thus, Cor. 9.1.0.9 settles everything. □

A simple conclusion from Thm. 9.3.2.18: If $\mathbf{A} = \mathbf{A}^\top \in \mathbb{R}^{n,n}$ with eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$, then

$$\lambda_1 = \min_{\mathbf{z}\in\mathbb{R}^n} \rho_{\mathbf{A}}(\mathbf{z}) \quad , \quad \lambda_2 = \min_{\mathbf{z}\in\mathbb{R}^n, \mathbf{z}\perp\mathbf{v}_1} \rho_{\mathbf{A}}(\mathbf{z}) \, , \tag{9.3.2.19}$$

where $\mathbf{v}_1 \in \mathbf{Eig}\mathbf{A}\lambda_1 \setminus \{0\}$. ⌟

Well, in Lemma 9.3.2.12 we encounter a generalized Rayleigh quotient $\rho_{\mathbf{A},\mathbf{D}}(\mathbf{y})$! How can Thm. 9.3.2.16 be applied to it?

▶ Transformation idea: $\rho_{\mathbf{A},\mathbf{D}}(\mathbf{D}^{-1/2}\mathbf{z}) = \rho_{\mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}}(\mathbf{z}) \, , \quad \mathbf{y} \in \mathbb{R}^n \, .$ (9.3.2.20)

▶ Apply Thm. 9.3.2.18 to transformed matrix $\widetilde{\mathbf{A}} := \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$. *Elementary* manipulations show

$$\begin{aligned} \text{(9.3.2.15)} \quad \Leftrightarrow \quad & \operatorname*{argmin}_{\mathbf{1}^\top\mathbf{D}\mathbf{y}=0} \rho_{\mathbf{A},\mathbf{D}}(\mathbf{y}) \overset{\mathbf{z}=\mathbf{D}^{1/2}\mathbf{y}}{=} \operatorname*{argmin}_{\mathbf{1}^\top\mathbf{D}^{1/2}\mathbf{z}=0} \rho_{\mathbf{A},\mathbf{D}}(\mathbf{D}^{-1/2}\mathbf{z}) \\ & = \operatorname*{argmin}_{\mathbf{1}^\top\mathbf{D}^{1/2}\mathbf{z}=0} \rho_{\widetilde{\mathbf{A}}}(\mathbf{z}) \quad \text{with} \quad \widetilde{\mathbf{A}} := \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2} \, . \end{aligned} \tag{9.3.2.21}$$

Related: transformation of a generalized eigenvalue problem into a standard eigenvalue problem according to

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{B}\mathbf{x} \quad \overset{\mathbf{z}=\mathbf{B}^{1/2}\mathbf{x}}{\Longrightarrow} \quad \mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}\mathbf{z} = \lambda\mathbf{z} \, . \tag{9.3.2.22}$$

$\mathbf{B}^{1/2} \hat{=}$ square root of s.p.d. matrix $\mathbf{B}$ → Rem. 10.3.0.2.

For segmentation problem: $\mathbf{B} = \mathbf{D}$ diagonal with positive diagonal entries, see (9.3.2.9)

➥ $\mathbf{D}^{-1/2} = \operatorname{diag}(d_1^{-1/2}, \ldots, d_N^{-1/2})$ and $\widetilde{\mathbf{A}} := \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$ can easily be computed.

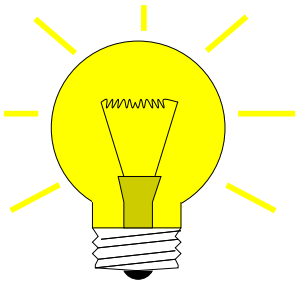
▶ In the sequel consider minimization problem/related eigenvalue problem

$$\boxed{\mathbf{z}^* = \operatorname*{argmin}_{\mathbf{1}^\top\mathbf{D}^{1/2}\mathbf{z}=0} \rho_{\widetilde{\mathbf{A}}}(\mathbf{z})} \quad \longleftrightarrow \quad \widetilde{\mathbf{A}}\mathbf{z} = \lambda\mathbf{z} \, . \tag{9.3.2.23}$$

Recover solution $\mathbf{y}^*$ of (9.3.2.15) as $\mathbf{y}^* = \mathbf{D}^{-1/2}\mathbf{z}^*$.

Still, Thm. 9.3.2.16 cannot be applied to (9.3.2.23):

How to deal with *constraint* $\mathbf{1}^\top\mathbf{D}^{1/2}\mathbf{z} = 0$?



Idea: Penalization

Add term $P(\mathbf{z})$ to $\rho_{\widetilde{\mathbf{A}}}(\mathbf{z})$ that becomes "sufficiently large" in case the constraint is violated.

$\blacktriangleright$ $\mathbf{z}^*$ can only be a minimizer of $\rho_{\widetilde{\mathbf{A}}}(\mathbf{z}) + P(\mathbf{z})$, if $P(\mathbf{z}) = 0$.

How to choose the penalty function $P(\mathbf{z})$ for the segmentation problem **?**

$$\left\{ \mathbf{1}^\top \mathbf{D}^{1/2}\mathbf{z} \neq 0 \quad \Rightarrow \quad P(\mathbf{z}) > 0 \right\} \text{ satisfied for } \boxed{P(\mathbf{z}) = \mu \frac{|\mathbf{1}^\top \mathbf{D}^{1/2}\mathbf{z}|^2}{\|\mathbf{z}\|_2^2}} ,$$

with penalty parameter $\mu > 0$.

$\blacktriangleright$ penalized minimization problem

dense rank-1 matrix

$$\begin{aligned} \mathbf{z}^* &= \operatorname*{argmin}_{\mathbf{z}\in\mathbb{R}^N\setminus\{0\}} \rho_{\widetilde{\mathbf{A}}}(\mathbf{z}) + P(\mathbf{z}) = \operatorname*{argmin}_{\mathbf{z}\in\mathbb{R}^N\setminus\{0\}} \rho_{\widetilde{\mathbf{A}}}(\mathbf{z}) + \frac{\mathbf{z}^\top(\mathbf{D}^{1/2}\mathbf{1}\mathbf{1}^\top\mathbf{D}^{1/2})\mathbf{z}}{\mathbf{z}^\top\mathbf{z}} \\ &= \operatorname*{argmin}_{\mathbf{z}\in\mathbb{R}^N\setminus\{0\}} \rho_{\widehat{A}}(\mathbf{z}) \quad \text{with} \quad \widehat{\mathbf{A}} := \widetilde{\mathbf{A}} + \mathbf{D}^{1/2}\mathbf{1}\mathbf{1}^\top\mathbf{D}^{1/2} . \end{aligned} \tag{9.3.2.24}$$

How to choose the penalty parameter μ **?**

In general: finding a "suitable" value for μ may be difficult or even impossible!

Here we are lucky:

$$\text{(9.3.2.10)} \quad \Rightarrow \quad \mathbf{A}\mathbf{1} = 0 \quad \Rightarrow \quad \widetilde{\mathbf{A}}(\mathbf{D}^{1/2}\mathbf{1}) = 0 \quad \Leftrightarrow \quad \mathbf{D}^{1/2}\mathbf{1} \in \mathbf{Eig}_{\widetilde{\mathbf{A}}}0 .$$

$\blacktriangleright$ Constraint in (9.3.2.23) means

Minimize over the orthogonal complement of an *eigenvector*. (9.3.2.25)

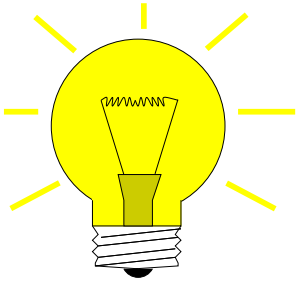
Cor. 9.1.0.9 $\blacktriangleright$ The orthogonal complement of an eigenvector of a *symmetric* matrix is spanned by the other eigenvectors (orthonormalization of eigenvectors belonging to the same eigenvalue is assumed).

(9.3.2.25) $\blacktriangleright$ The minimizer of (9.3.2.23) will be one of the other eigenvectors of $\widetilde{\mathbf{A}}$ that belongs to the smallest eigenvalue.

Note: This eigenvector $\mathbf{z}^*$ will be orthogonal to $\mathbf{D}^{1/2}\mathbf{1}$, it satisfies the constraint, and, thus, $P(\mathbf{z}^*) = 0$!

Note: eigenspaces of $\widetilde{\mathbf{A}}$ and $\widehat{\mathbf{A}}$ agree.

Note: Lemma 2.8.0.13 $\Longrightarrow$ $\widetilde{\mathbf{A}}$ is *positive semidefinite* ($\rightarrow$ Def. 1.1.2.6) with smallest eigenvalue 0.n



Idea: Choose penalization parameter μ in (9.3.2.24) such that $\mathbf{D}^{1/2}\mathbf{1}$ is guaranteed not to be an eigenvector belonging to the smallest eigenvalue of $\widehat{\mathbf{A}}$.

Safe choice: choose μ such that $\mathbf{D}^{1/2}\mathbf{1}$ will belong to the largest eigenvalue of $\widehat{\mathbf{A}}$: Thm. 9.1.0.4 suggests

$$\mu = \left\|\widetilde{\mathbf{A}}\right\|_\infty \overset{(1.5.5.13)}{=} 2 . \tag{9.3.2.26}$$

$$\mathbf{z}^* = \operatorname*{argmin}_{\mathbf{1}^\top \mathbf{D}^{1/2}\mathbf{z}=0} \rho_{\widetilde{\mathbf{A}}}(\mathbf{z}) = \operatorname*{argmin}_{\mathbf{z}\neq 0} \rho_{\widehat{\mathbf{A}}}(\mathbf{z}) \,. \tag{9.3.2.27}$$

By Thm. 9.3.2.16:

$$\begin{aligned} &\mathbf{z}^* \text{ = eigenvector belonging to minimal eigenvalue of } \widehat{\mathbf{A}}, \\ &\Updownarrow \\ &\mathbf{z}^* \text{ = eigenvector } \perp \mathbf{D}^{1/2}\mathbf{1} \text{ belonging to minimal eigenvalue of } \widetilde{\mathbf{A}}, \\ &\Updownarrow \\ &\mathbf{D}^{-1/2}\mathbf{z}^* \text{ = minimizer for (9.3.2.15).} \end{aligned}$$

§9.3.2.28 (Algorithm outline: Binary grayscale image segmentation)

❶ Given similarity function σ compute (sparse!) matrices $\mathbf{W}, \mathbf{D}, \mathbf{A} \in \mathbb{R}^{N,N}$, see (9.3.2.3), (9.3.2.9).

❷ Compute $\mathbf{y}^*$, $\|\mathbf{y}^*\|_2 = 1$, as eigenvector belonging to the *smallest eigenvalue* of $\widehat{\mathbf{A}} := \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2} + 2(\mathbf{D}^{1/2}\mathbf{1})(\mathbf{D}^{1/2}\mathbf{1})^\top$.

❸ Set $\mathbf{x}^* = \mathbf{D}^{-1/2}\mathbf{y}^*$ and define the image segment as pixel set

$$\mathcal{X} := \{i \in \{1,\ldots,N\} : x_i^* > \tfrac{1}{N}\sum_{i=1}^{N} x_i^*\} \,. \tag{9.3.2.29}$$

mean value of entries of $\mathbf{x}^*$

MATLAB-code 9.3.2.30: 1st stage of segmentation of grayscale image

```
% Read image and build matrices, see Code 9.3.2.11 and (9.3.2.9)
P = imread('image.pbm'); [m,n] = size(P); [A,D] = imgsegmat(P);
% Build scaling matrics
N = size(A,1); dv = sqrt(spdiags(A,0));
Dm = spdiags(1./dv,[0],N,N); % D^{-1/2}
Dp = spdiags(dv,[0],N,N);    % D^{-1/2}
% Build (densely populated !) matrix Â
c = Dp*ones(N,1); Ah = Dm*A*Dm + 2*c*c';
% Compute and sort eigenvalues; grossly inefficient !
[W,E] = eig(full(Ah)); [ev,idx] = sort(diag(E)); W(:,idx) = W;
% Obtain eigenvector x* belonging to 2nd smallest generalized
% eigenvalue of A and D
x = W(:,1); x = Dm*v;
% Extract segmented image
xs = reshape(x,m,n); Xidx = find(xs>(sum(sum(xs))/(n*m)));
```

1st-stage of segmentation of 31×25 grayscale pixel image (`root.pbm`, red pixels $\widehat{=}\ \mathcal{X}$, $\sigma(x,y) = \exp(-(x-y/10)^2)$)

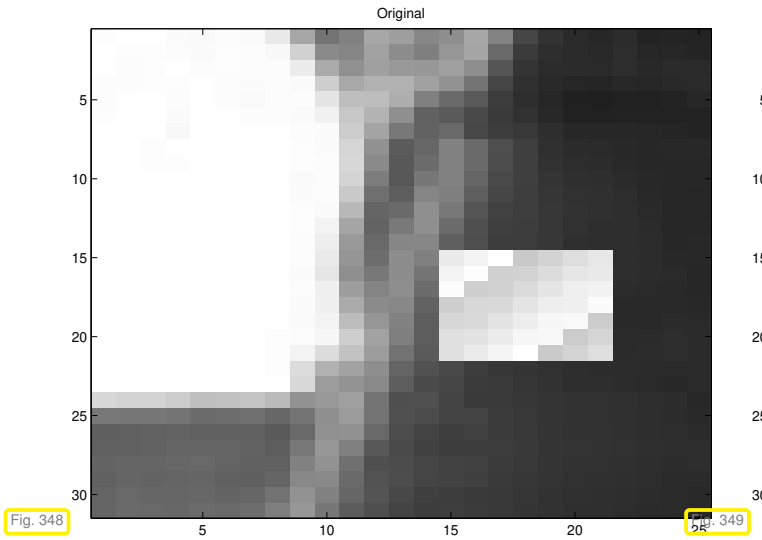


Fig. 348 Fig. 349

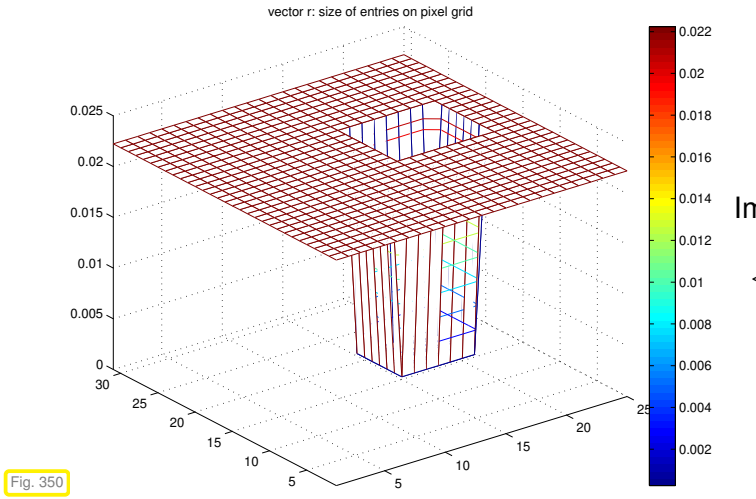


Fig. 350

Image from Fig. 348:

◁ eigenvector $\mathbf{x}^*$ plotted on pixel grid

To identify more segments, the same algorithm is *recursively applied* to segment parts of the image already determined.

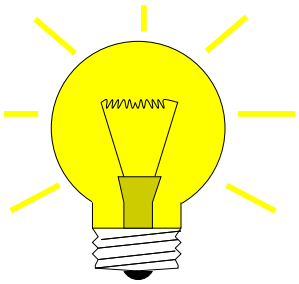
Practical segmentation algorithms rely on many more steps of which the above algorithm is only one, preceeded by substantial preprocessing. Moreover, they dispense with the strictly local perspective adopted above and take into account more distant connections between image parts, often in a randomized fashion [SM00].

The image segmentation problem falls into the wider class of graph partitioning problems. Methods based on (a few of) the eigenvector of the connectivity matrix belonging to the smallest eigenvalues are known as spectral partitioning methods. The eigenvector belonging to the smallest non-zero eigenvalue that we computed above is usually called the Fiedler vector of the graph, see [AKY99; ST96]. ⌟

The solution of the image segmentation problem by means of `eig` in Code 9.3.2.30 amounts a tremendous waste of computational resources: we compute *all* eigenvalues/eigenvectors of dense matrices, though only a single eigenvector associated with the smallest eigenvalue is of interest.

This motivates the quest to find *efficient* numerical methods for the following task.

Task:	given $\mathbf{A} \in \mathbb{K}^{n,n}$, find smallest (in modulus) eigenvalue of regular $\mathbf{A} \in \mathbb{K}^{n,n}$ and (an) associated eigenvector.



If $\mathbf{A} \in \mathbb{K}^{n,n}$ regular:

Smallest (in modulus) EV of $\mathbf{A}$ = $\left(\text{Largest (in modulus) EV of } \mathbf{A}^{-1}\right)^{-1}$

Direct power method (→ Section 9.3.1) for $\mathbf{A}^{-1}$ = inverse iteration

MATLAB-code 9.3.2.31: inverse iteration for computing $\lambda_{\min}(\mathbf{A})$ and associated eigenvector

Note: reuse of LU-factorization, see Rem. 2.5.0.10

Remark 9.3.2.32 (Shifted inverse iteration)

More general task:

For $\alpha \in \mathbb{C}$ find $\lambda \in \sigma(\mathbf{A})$ such that $|\alpha - \lambda| = \min\{|\alpha - \mu|, \mu \in \sigma(\mathbf{A})\}$

Shifted inverse iteration: [DR08, Alg .7.24]

$$\mathbf{z}^{(0)} \text{ arbitrary}, \quad \mathbf{w} = (\mathbf{A} - \alpha\mathbf{I})^{-1}\mathbf{z}^{(k-1)} \quad , \quad \mathbf{z}^{(k)} := \frac{\mathbf{w}}{\|\mathbf{w}\|_2}, \quad k = 1, 2, \ldots , \tag{9.3.2.33}$$

where: $(\mathbf{A} - \alpha\mathbf{I})^{-1}\mathbf{z}^{(k-1)} \hat{=}$ solve $(\mathbf{A} - \alpha\mathbf{I})\mathbf{w} = \mathbf{z}^{(k-1)}$ based on Gaussian elimination (↔ a single LU-factorization of $\mathbf{A} - \alpha\mathbf{I}$ as in Code 9.3.2.31).

Remark 9.3.2.34 ((Nearly) singular LSE in shifted inverse iteration)

What if "by accident" $\alpha \in \sigma(\mathbf{A})$ (⇔ $\mathbf{A} - \alpha\mathbf{I}$ singular) **?**

Stability of Gaussian elimination/LU-factorization (→ **??**) will ensure that "$\mathbf{w}$ from (9.3.2.33) points in the right direction"

In other words, roundoff errors may badly affect the length of the solution $\mathbf{w}$, but not its direction.

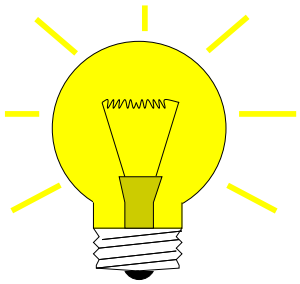
Practice [GT08]: If, in the course of Gaussian elimination/LU-factorization a zero pivot element is really encountered, then we just *replace it with* `eps`, in order to avoid `inf` values!

Thm. 9.3.1.21 ➢ Convergence of shifted inverse iteration for $\mathbf{A}^H = \mathbf{A}$:

Asymptotic linear convergence, Rayleigh quotient $\to \lambda_j$ with rate

$$\frac{|\lambda_j - \alpha|}{\min\{|\lambda_i - \alpha|, i \neq j\}} \quad \text{with} \quad \lambda_j \in \sigma(\mathbf{A}) \, , \quad |\alpha - \lambda_j| \leq |\alpha - \lambda| \quad \forall \lambda \in \sigma(\mathbf{A}) \, .$$

Extremely fast for $\alpha \approx \lambda_j$ **!**



Idea: A posteriori adaptation of shift

Use $\alpha := \rho_{\mathbf{A}}(\mathbf{z}^{(k-1)})$ in k-th step of inverse iteration.

§9.3.2.35 (Rayleigh quotient iteration → [Han02, Alg. 25.2])

MATLAB-code 9.3.2.36: Rayleigh quotient iteration (for *normal* $\mathbf{A} \in \mathbb{R}^{n,n}$)

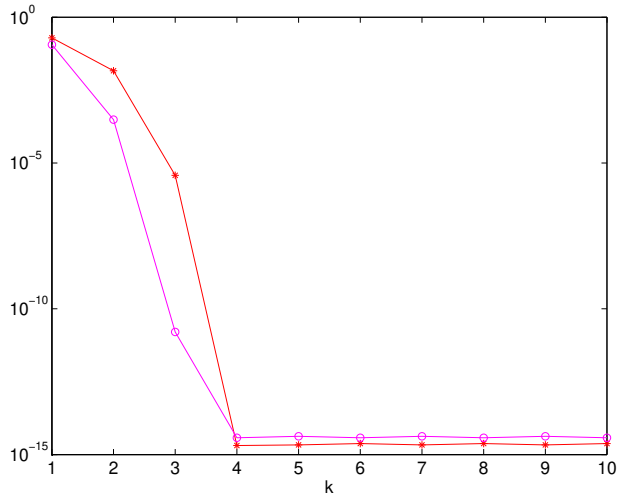
??: note use of `speye` to preserve spare matrix data format!

✦ Drawback compared with Code 9.3.2.31: reuse of LU-factorization no longer possible.

✦ Even if LSE nearly singular, stability of Gaussian elimination guarantees correct direction of $\mathbf{z}$, see discussion in Rem. 9.3.2.34.

EXAMPLE 9.3.2.37 (Rayleigh quotient iteration)

Monitored: iterates of Rayleigh quotient iteration (9.3.2.36) for s.p.d. $\mathbf{A} \in \mathbb{R}^{n,n}$



MATLAB-code 9.3.2.38:

```
d = (1:10)';
n = length(d);
Z = diag(sqrt(1:n),0)+ones(n,n);
[Q,R] = qr(Z);
A = Q*diag(d,0)*Q';
```

o : $|\lambda_{\min} - \rho_{\mathbf{A}}(\mathbf{z}^{(k)})|$

$*$: $\left\|\mathbf{z}^{(k)} - \mathbf{x}_j\right\|$, $\lambda_{\min} = \lambda_j$, $\mathbf{x}_j \in \mathbf{EigA}\lambda_j$, $\left\|\mathbf{x}_j\right\|_2 = 1$

k	$\lvert\lambda_{\min} - \rho_{\mathbf{A}}(\mathbf{z}^{(k)})\rvert$	$\left\|\mathbf{z}^{(k)} - \mathbf{x}_j\right\|$
1	0.09381702342056	0.20748822490698
2	0.00029035607981	0.01530829569530
3	0.00000000001783	0.00000411928759
4	0.00000000000000	0.00000000000000
5	0.00000000000000	0.00000000000000

Theorem 9.3.2.39. → [Han02, Thm. 25.4]

If $\mathbf{A} = \mathbf{A}^H$, then $\rho_{\mathbf{A}}(\mathbf{z}^{(k)})$ converges locally of order 3 *(→ Def. 8.2.2.10) to the smallest eigenvalue (in modulus), when $\mathbf{z}^{(k)}$ are generated by the Rayleigh quotient iteration* (9.3.2.36).

9.3.3 Preconditioned inverse iteration (PINVIT)

Task:	given $\mathbf{A} \in \mathbb{K}^{n,n}$, find smallest (in modulus) eigenvalue of regular $\mathbf{A} \in \mathbb{K}^{n,n}$ and (an) associated eigenvector.

➤ Options: inverse iteration (→ Code 9.3.2.31) and Rayleigh quotient iteration (9.3.2.36).

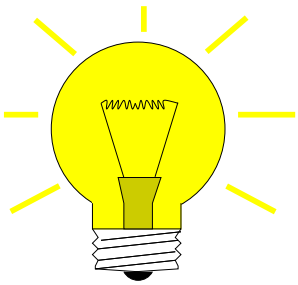
? What if direct solution of $\mathbf{Ax} = \mathbf{b}$ not feasible ?

This can happen, in case

- for large sparse $\mathbf{A}$ the amount of fill-in exhausts memory, despite sparse elimination techniques (→ Section 2.7.4),
- $\mathbf{A}$ is available only through a routine `evalA(x)` providing $\mathbf{A}\times$vector.

We expect that an approximate solution of the linear systems of equations encountered during inverse iteration should be sufficient, because we are dealing with approximate eigenvectors anyway.

Thus, iterative solvers for solving $\mathbf{Aw} = \mathbf{z}^{(k-1)}$ may be considered, see Chapter 10. However, the required accuracy is not clear a priori. Here we examine an approach that completely dispenses with an iterative solver and uses a *preconditioner* (→ Notion 10.3.0.3) instead.

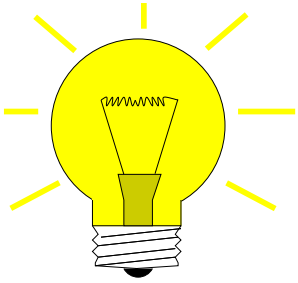


Idea: (for inverse iteration without shift, $\mathbf{A} = \mathbf{A}^H$ s.p.d.)
Instead of solving $\mathbf{Aw} = \mathbf{z}^{(k-1)}$ compute $\mathbf{w} = \mathbf{B}^{-1}\mathbf{z}^{(k-1)}$ with "inexpensive" s.p.d. approximate inverse $\mathbf{B}^{-1} \approx \mathbf{A}^{-1}$

➢ $\mathbf{B} \hat{=}$ Preconditioner for $\mathbf{A}$, see Notion 10.3.0.3

Possible to replace $\mathbf{A}^{-1}$ with $\mathbf{B}^{-1}$ in inverse iteration **?**

NO, because we are not interested in smallest eigenvalue of $\mathbf{B}$!



Replacement $\mathbf{A}^{-1} \to \mathbf{B}^{-1}$ possible only when applied to residual quantity

residual quantity = quantity that $\to 0$ in the case of convergence to exact solution

Natural residual quantity for eigenvalue problem $\mathbf{Ax} = \lambda\mathbf{x}$:

$$\mathbf{r} := \mathbf{Az} - \rho_{\mathbf{A}}(\mathbf{z})\mathbf{z} \quad , \quad \rho_{\mathbf{A}}(\mathbf{z}) = \text{Rayleigh quotient} \to \text{Def. 9.3.1.16 .}$$

Note: only *direction* of $\mathbf{A}^{-1}\mathbf{z}$ matters in inverse iteration (9.3.2.33)

$$(\mathbf{A}^{-1}\mathbf{z}) \parallel (\mathbf{z} - \mathbf{A}^{-1}(\mathbf{Az} - \rho_{\mathbf{A}}(\mathbf{z})\mathbf{z})) \quad \Rightarrow \quad \text{defines same next iterate!}$$

[Preconditioned inverse iteration (PINVIT) for s.p.d. $\mathbf{A}$]

$\mathbf{z}^{(0)}$ arbitrary,

$$\begin{aligned} \mathbf{w} &= \mathbf{z}^{(k-1)} - \mathbf{B}^{-1}(\mathbf{Az}^{(k-1)} - \rho_{\mathbf{A}}(\mathbf{z}^{(k-1)})\mathbf{z}^{(k-1)}) \,, \\ \mathbf{z}^{(k)} &= \frac{\mathbf{w}}{\|\mathbf{w}\|_2} \,, \end{aligned} \qquad k = 1, 2, \ldots \,. \tag{9.3.3.1}$$

Computational effort:

MATLAB-code 9.3.3.2: preconditioned inverse iteration (9.3.3.1)

1 matrix×vector
1 evaluation of preconditioner
A few AXPY-operations

EXAMPLE 9.3.3.3 (Convergence of PINVIT)

S.p.d. matrix $\mathbf{A} \in \mathbb{R}^{n,n}$, tridiagonal preconditioner, see Ex. 10.3.0.11

MATLAB-code 9.3.3.4:

```
A = spdiags(repmat([1/n,-1,2*(1+1/n),-1,1/n],n,1),
   [-n/2,-1,0,1,n/2],n,n);
evalA = @(x) A*x;
% inverse iteration
invB = @(x) A\x;
% tridiagonal preconditioning
B = spdiags(spdiags(A,[-1,0,1]),[-1,0,1],n,n); invB  = @(x) B\x;
```

Monitored: error decay during iteration of Code 9.3.3.2: $|\rho_{\mathbf{A}}(\mathbf{z}^{(k)}) - \lambda_{\min}(\mathbf{A})|$

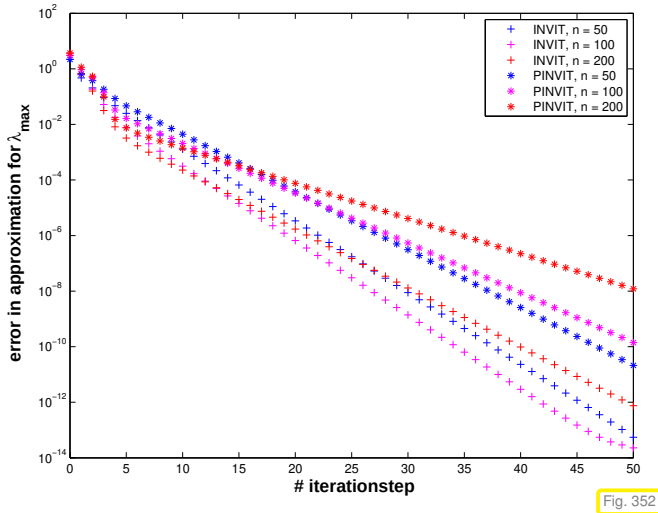


Fig. 351

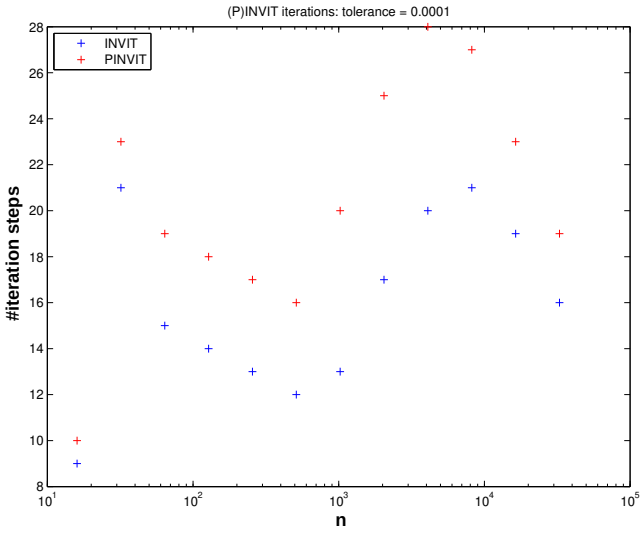


Fig. 352

Observation: linear convergence of eigenvectors also for PINVIT.

Theory [Ney99a; Ney99b]:

- linear convergence of (9.3.3.1)
- fast convergence, if spectral condition number $\kappa(\mathbf{B}^{-1}\mathbf{A})$ small

The theory of PINVIT [Ney99a; Ney99b] is based on the identity

$$\mathbf{w} = \rho_{\mathbf{A}}(\mathbf{z}^{(k-1)})\mathbf{A}^{-1}\mathbf{z}^{(k-1)} + (\mathbf{I} - \mathbf{B}^{-1}\mathbf{A})(\mathbf{z}^{(k-1)} - \rho_{\mathbf{A}}(\mathbf{z}^{(k-1)})\mathbf{A}^{-1}\mathbf{z}^{(k-1)}) \,. \tag{9.3.3.5}$$

For small residual $\mathbf{A}\mathbf{z}^{(k-1)} - \rho_{\mathbf{A}}(\mathbf{z}^{(k-1)})\mathbf{z}^{(k-1)}$ PINVIT almost agrees with the regular inverse iteration.

9.3.4 Subspace iterations

Remark 9.3.4.1 (Excited resonances)

Consider the non-autonomous ODE (excited harmonic oscillator)

$$\ddot{y} + \lambda^2 y = \cos(\omega t) \,, \tag{9.3.4.2}$$

with general solution

$$y(t) = \begin{cases} \dfrac{1}{\lambda^2 - \omega^2}\cos(\omega t) + A\cos(\lambda t) + B\sin(\lambda t) & \text{, if } \lambda \neq \omega \,, \\ \dfrac{t}{2\omega}\sin(\omega t) + A\cos(\lambda t) + B\sin(\lambda t) & \text{, if } \lambda = \omega \,. \end{cases} \tag{9.3.4.3}$$

▻ growing solutions possible in resonance case $\lambda = \omega$!

Now consider harmonically excited vibration modelled by ODE

$$\ddot{\mathbf{y}} + \mathbf{A}\mathbf{y} = \mathbf{b}\cos(\omega t) \,, \tag{9.3.4.4}$$

with *symmetric*, *positive (semi)definite* matrix $\mathbf{A} \in \mathbb{R}^{n,n}$, $\mathbf{b} \in \mathbb{R}^n$. By Cor. 9.1.0.9 there is an *orthogonal* matrix $\mathbf{Q} \in \mathbb{R}^{n,n}$ such that

$$\mathbf{Q}^\top \mathbf{A}\mathbf{Q} = \mathbf{D} := \operatorname{diag}(\lambda_1, \ldots, \lambda_n) \,.$$

where the $0 \leq \lambda_1 < \lambda_2 < \cdots < \lambda_n$ are the eigenvalues of $\mathbf{A}$.

▻ Transform ODE as in Ex. 9.0.0.6: with $\mathbf{z} = \mathbf{Q}^\top \mathbf{y}$

(9.3.4.4) ▻ $\ddot{\mathbf{z}} + \mathbf{D}\mathbf{z} = \mathbf{Q}^\top \mathbf{b}\,\cos(\omega t)$.

We have obtained decoupled linear 2nd-order scalar ODEs of the type (9.3.4.2).

▻ (9.3.4.4) can have growing (with time) solutions, if $\omega = \sqrt{\lambda_i}$ for some $i = 1, \ldots, n$

If $\omega = \sqrt{\lambda_j}$ for one $j \in \{1, \ldots, n\}$, then the solution for the initial value problem for (9.3.4.4) with $\mathbf{y}(0) = \dot{\mathbf{y}}(0) = 0$ ($\leftrightarrow \mathbf{z}(0) = \dot{\mathbf{z}}(0) = 0$) is

$$\mathbf{z}(t) \sim \frac{t}{2\omega}\sin(\omega t)\mathbf{e}_j + \text{bounded oscillations}$$

$$\Updownarrow$$

$$\mathbf{y}(t) \sim \frac{t}{2\omega}\sin(\omega t)(\mathbf{Q})_{:,j} + \text{bounded oscillations} \,.$$

j-th eigenvector of $\mathbf{A}$

▻ Eigenvectors of $\mathbf{A}$ $\leftrightarrow$ excitable states ⌟

EXAMPLE 9.3.4.5 (Vibrations of a truss structure ***cf.*** **[Han02, Sect. 3], MATLAB's `truss` demo)**

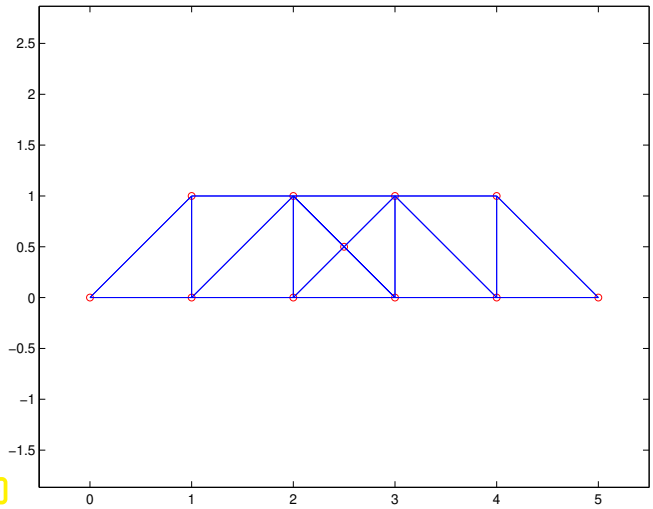


Fig. 353

◁ *a "bridge" truss]*
A truss is a structure composed of (massless) rods and point masses; we consider in-plane (2D) trusses.

Encoding: positions of masses + (sparse) connectivity matrix

MATLAB-code 9.3.4.6: Data for "bridge truss"

Assumptions:
- ✦ Truss in static equilibrium (perfect balance of forces at each point mass).
- ✦ Rods are *perfectly elastic* (*i.e.*, frictionless).

Hook's law holds for force in the direction of a rod:

$$F = \alpha \frac{\Delta l}{l} , \tag{9.3.4.7}$$

where
- ✦ l is the equilibrium length of the rod,
- ✦ Δl is the elongation of the rod effected by the force F in the direction of the rod
- ✦ α is a material coefficient (Young's modulus).

n point masses are numbered $1, \ldots, n$: $\mathbf{p}^i \in \mathbb{R}^2 \hat{=}$ position of i-th mass

We consider a swaying truss: description by time-dependent displacements $\mathbf{u}^i(t) \in \mathbb{R}^2$ of point masses:

$$\text{position of } i\text{-th mass at time } t \quad = \quad \mathbf{p}^i + \mathbf{u}^i(t); .$$

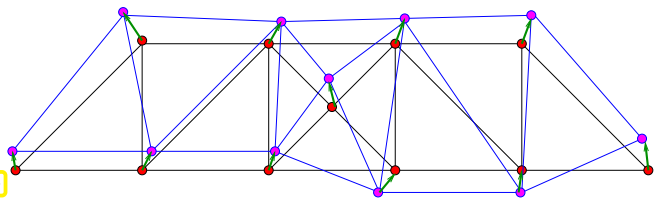


Fig. 354

◁ deformed truss:
- • $\hat{=}$ point masses at positions $\mathbf{p}^i$
- → $\hat{=}$ displacement vectors $\mathbf{u}^i$
- • $\hat{=}$ shifted masses at $\mathbf{p}^i + \mathbf{u}^i$

Equilibrium length and (time-dependent) elongation of rod connecting point masses i and j, $i \neq j$:

$$l_{ij} := \left\| \Delta \mathbf{p}^{ji} \right\|_2 , \quad \Delta \mathbf{p}^{ji} := \mathbf{p}^j - \mathbf{p}^i , \tag{9.3.4.8}$$

$$\Delta l_{ij}(t) := \left\| \Delta \mathbf{p}^{ji} + \Delta \mathbf{u}^{ji}(t) \right\|_2 - l_{ij} , \quad \Delta \mathbf{u}^{ji}(t) := \mathbf{u}^j(t) - \mathbf{u}^i(t) . \tag{9.3.4.9}$$

Extra (reaction) force on masses i and j:

$$F_{ij}(t) = -\alpha_{ij} \frac{\Delta l_{ij}}{l_{ij}} \cdot \frac{\Delta \mathbf{p}^{ji} + \Delta \mathbf{u}^{ji}(t)}{\left\| \Delta \mathbf{p}^{ji} + \Delta \mathbf{u}^{ji}(t) \right\|_2} . \tag{9.3.4.10}$$

Assumption:	Small displacements

Possibility of linearization by neglecting terms of order $\left\|\mathbf{u}^i\right\|_2^2$

(9.3.4.8)
(9.3.4.9)

$$F_{ij}(t) = \alpha_{ij}\left(\frac{1}{\left\|\Delta\mathbf{p}^{ji} + \Delta\mathbf{u}^{ji}(t)\right\|_2} - \frac{1}{\left\|\Delta\mathbf{p}^{ji}\right\|}\right)\cdot(\Delta\mathbf{p}^{ji} + \Delta\mathbf{u}^{ji}(t))\,. \tag{9.3.4.11}$$

Lemma 9.3.4.12. Taylor expansion of inverse distance function

For $\mathbf{x} \in \mathbb{R}^d \setminus \{0\}$, $\mathbf{y} \in \mathbb{R}^d$, $\|\mathbf{y}\|_2 < \|\mathbf{x}\|_2$ *holds for* $\mathbf{y} \to 0$

$$\frac{1}{\|\mathbf{x}+\mathbf{y}\|_2} = \frac{1}{\|\mathbf{x}\|_2} - \frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|_2^3} + O(\|\mathbf{y}\|_2^2)\,.$$

Proof. Simple Taylor expansion up to linear term for $f(\mathbf{x}) = (x_1^2 + \cdots + x_d^2)^{-1/2}$ and $f(\mathbf{x}+\mathbf{y}) = f(\mathbf{x}) + \mathbf{grad}\, f(\mathbf{x})\cdot\mathbf{y} + O(\|\mathbf{y}\|_2^2)$. □

Linearization of force: apply Lemma 9.3.4.12 to (9.3.4.11) and drop terms $O(\left\|\Delta\mathbf{u}^{ji}\right\|_2^2)$:

$$\begin{aligned} F_{ij}(t) &\approx -\alpha_{ij}\frac{\Delta\mathbf{p}^{ji}\cdot\Delta\mathbf{u}^{ji}(t)}{l_{ij}^3}\cdot(\Delta\mathbf{p}^{ji} + \Delta\mathbf{u}^{ji}(t)) \\ &\approx -\alpha_{ij}\frac{\Delta\mathbf{p}^{ji}\cdot\Delta\mathbf{u}^{ji}(t)}{l_{ij}^3}\cdot\Delta\mathbf{p}^{ji}\,. \end{aligned} \tag{9.3.4.13}$$

Newton's second law of motion: ($F_i \mathrel{\hat{=}}$ total force acting on i-th mass)

$$\boxed{m_i\frac{d^2}{dt^2}\mathbf{u}^i(t) = F_i = \sum_{\substack{j=1\\ j\neq i}}^{n} -F_{ij}(t)}\,, \tag{9.3.4.14}$$

$m_i \mathrel{\hat{=}}$ mass of point mass i.

$$m_i\frac{d^2}{dt^2}\mathbf{u}^i(t) = \sum_{\substack{j=1\\ j\neq i}}^{n}\alpha_{ij}\frac{1}{l_{ij}^3}\left(\Delta\mathbf{p}^{ji}(\Delta\mathbf{p}^{ji})^\top\right)(\mathbf{u}^j(t) - \mathbf{u}^i(t))\,. \tag{9.3.4.15}$$

Compact notation: collect all displacements into one vector $\mathbf{u}(t) = \left(\mathbf{u}^i(t)\right)_{i=1}^n \in \mathbb{R}^{2n}$

(9.3.4.15)

$$\boxed{\mathbf{M}\frac{d\mathbf{u}}{dt^2}(t) + \mathbf{A}\mathbf{u}(t) = \mathbf{f}(t)}\,. \tag{9.3.4.16}$$

with mass matrix $\mathbf{M} = \operatorname{diag}(m_1, m_1, \ldots, m_n, m_n)$

and stiffness matrix $\mathbf{A} \in \mathbb{R}^{2n,2n}$ with 2×2-blocks

$$\begin{aligned}(\mathbf{A})_{2i-1:2i,2i-1,2i} &= \sum_{\substack{j=1\\ j\neq i}}^{n} \alpha_{ij} \frac{1}{l_{ij}^3} \left(\Delta\mathbf{p}^{ji} (\Delta\mathbf{p}^{ji})^\top \right) , \quad i = 1, \ldots, n , \\ (\mathbf{A})_{2i-1:2i,2j-1:2j} &= -\alpha_{ij} \frac{1}{l_{ij}^3} \left(\Delta\mathbf{p}^{ji} (\Delta\mathbf{p}^{ji})^\top \right) , \quad i \neq j .\end{aligned} \tag{9.3.4.17}$$

and external forces $\mathbf{f}(t) = \left(\mathbf{f}^i(t) \right)_{i=1}^n$.

Note: stiffness matrix $\mathbf{A}$ is *symmetric, positive semidefinite* (→ Def. 1.1.2.6).

Rem. 9.3.4.1: if periodic external forces $\mathbf{f}(t) = \cos(\omega t)\mathbf{f}$, $\mathbf{f} \in \mathbb{R}^{2n}$, (wind, earthquake) act on the truss they can excite vibrations of (linearly in time) growing amplitude, if ω coincides with $\sqrt{\lambda_j}$ for an eigenvalue λ_j of $\mathbf{A}$.

Excited vibrations can lead to the collapse of a truss structure, *cf.* the notorious Tacoma-Narrows bridge disaster.

It is essential to know whether eigenvalues of a truss structure fall into a range that can be excited by external forces.

These will typically$^{(*)}$ be the low modes ↔ a few of the smallest eigenvalues.

($(*)$ Reason: fast oscillations will quickly be damped due to friction, which was neglected in our model.)

MATLAB-code 9.3.4.18: Computing resonant frequencies and modes of elastic truss

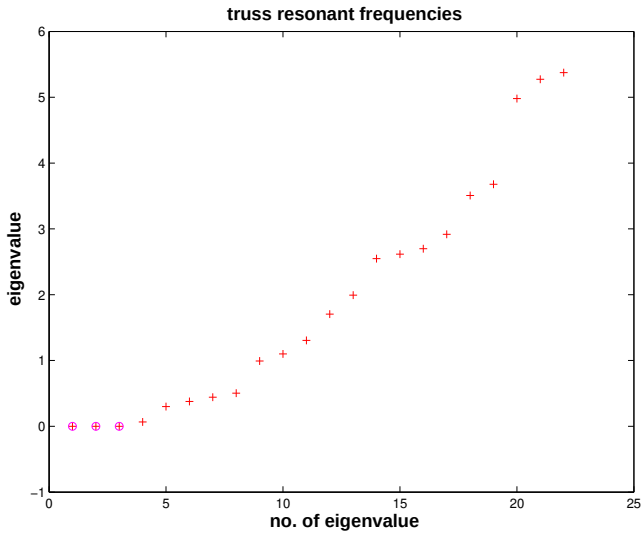
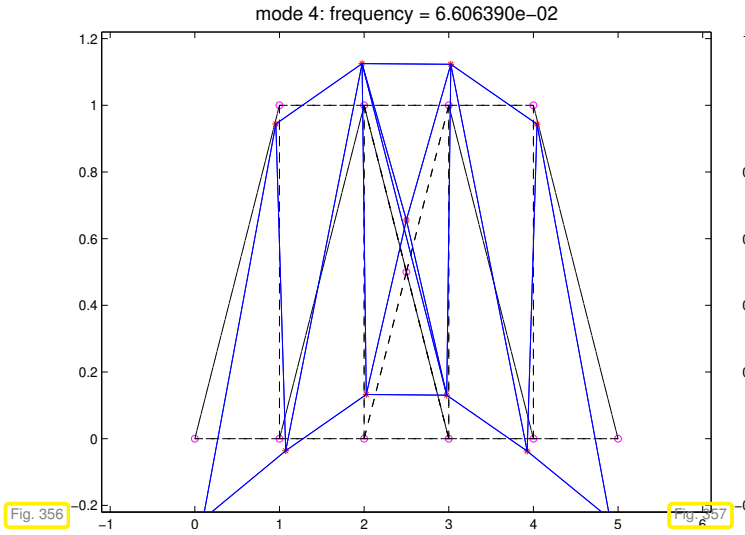
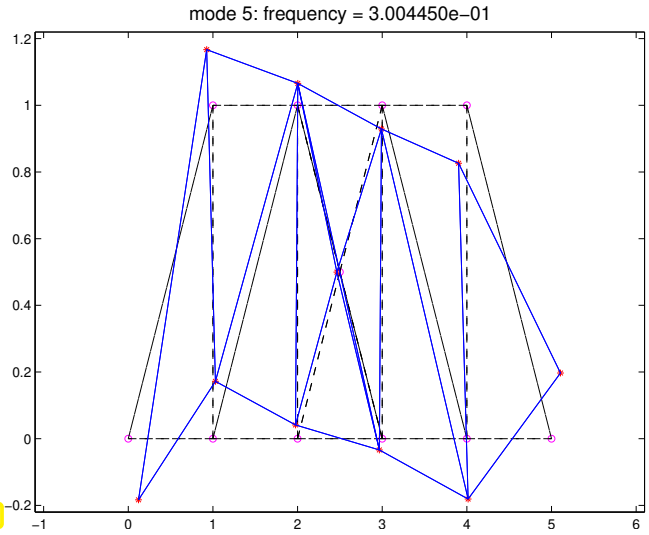
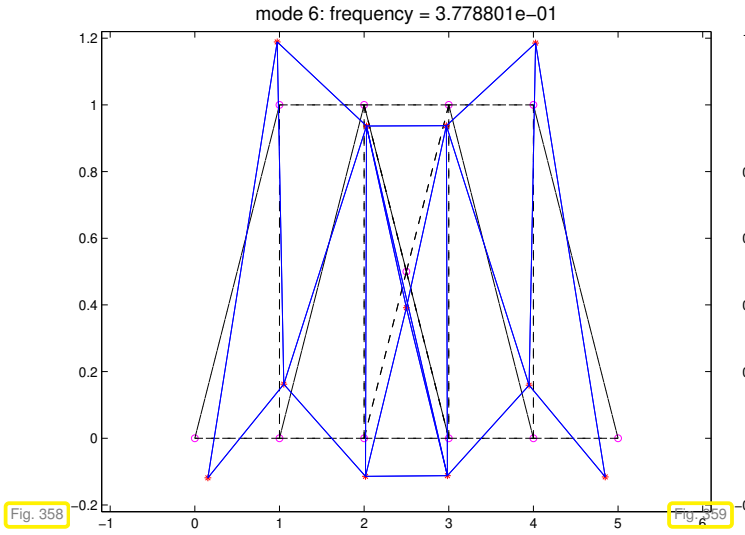


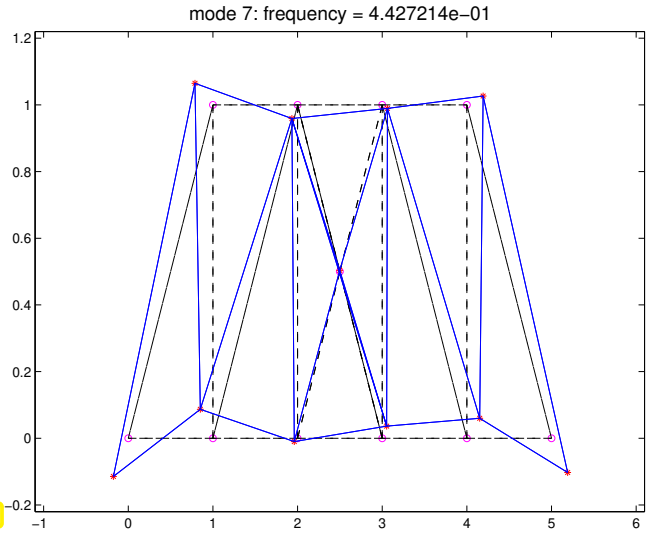
Fig. 355

◁ resonant frequencies of bridge truss from Fig. 353. The stiffness matrix will always possess three zero eigenvalues corresponding to rigid body modes (= displacements without change of length of the rods)









To compute *a few* of a truss's lowest resonant frequencies and excitable mode, we need efficient numerical methods for the following tasks. Obviously, Code 9.3.4.18 cannot be used for large trusses, because `eig` invariable operates on dense matrices and will be prohibitively slow and gobble up huge amounts of memory, also recall the discussion of Code 9.3.2.30.

Task:	Compute m, $m \ll n$, of the smallest/largest (in modulus) eigenvalues of $\mathbf{A} = \mathbf{A}^{\mathrm{H}} \in \mathbb{C}^{n,n}$ and associated eigenvectors.

Of course, we aim to tackle this task by iterative methods generalizing power iteration (→ Section 9.3.1) and inverse iteration (→ Section 9.3.2).

9.3.4.1 Orthogonalization

Preliminary considerations (in $\mathbb{R}$, $m = 2$):

According to Cor. 9.1.0.9: For $\mathbf{A} = \mathbf{A}^{\top} \in \mathbb{R}^{n,n}$ there is a factorization $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^{\top}$ with $\mathbf{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, $\lambda_j \in \mathbb{R}$, $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$, and $\mathbf{U}$ orthogonal. Thus, $\mathbf{u}_j := (\mathbf{U})_{:,j}$, $j = 1, \ldots, n$, are (mutually orthogonal) eigenvectors of $\mathbf{A}$.

Assume $0 \leq \lambda_1 \leq \cdots \leq \lambda_{n-2} < \lambda_{n-1} < \lambda_n$ (largest eigenvalues are simple).

If we just carry out the direct power iteration (9.3.1.12) for two vectors both sequences will converge to the largest (in modulus) eigenvector. However, we recall that all eigenvectors are mutually orthogonal. This

suggests that we orthogonalize the iterates of the second power iteration (that is to yield the eigenvector for the second largest eigenvalue) with respect to those of the first. This idea spawns the following iteration, *cf.* Gram-Schmidt orthogonalization in (10.2.2.4):

MATLAB-code 9.3.4.19: one step of subspace power iteration, $m = 2$

```
1 function [v,w] = sspowitstep(A,v,w)
2 v = A*v; w = A*w; % "power iteration", cf. (9.3.1.12)
3 % orthogonalization, cf. Gram-Schmidt orthogonalization (10.2.2.4)
4 v = v/norm(v); w = w - dot(v,w)*v; w = w/norm(w); % now w ⊥ v
```

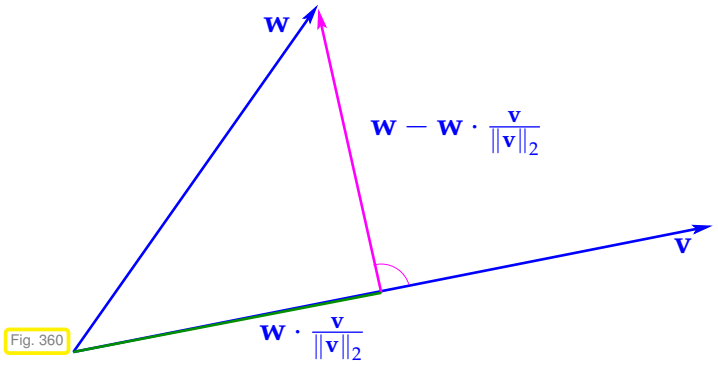


Fig. 360

◁ Orthogonalization of two vectors (see Line 4 of Code 9.3.4.19)

Analysis through eigenvector expansions ($\mathbf{v},\mathbf{w} \in \mathbb{R}^n$, $\|\mathbf{v}\|_2 = \|\mathbf{w}\|_2 = 1$)

$$\mathbf{v} = \sum_{i=1}^{n} \alpha_j \mathbf{u}_j \quad , \quad \mathbf{w} = \sum_{i=1}^{n} \beta_j \mathbf{u}_j \, ,$$

$$\Rightarrow \quad \mathbf{A}\mathbf{v} = \sum_{i=1}^{n} \lambda_j \alpha_j \mathbf{u}_j \quad , \quad \mathbf{A}\mathbf{w} = \sum_{i=1}^{n} \lambda_j \beta_j \mathbf{u}_j \, ,$$

$$\mathbf{v}_0 := \frac{\mathbf{v}}{\|\mathbf{v}\|_2} = \Big(\sum_{i=1}^{n} \lambda_j^2 \alpha_j^2\Big)^{-1/2} \sum_{i=1}^{n} \lambda_j \alpha_j \mathbf{u}_j \, ,$$

$$\mathbf{A}\mathbf{w} - (\mathbf{v}_0^\top \mathbf{A}\mathbf{w})\mathbf{v}_0 = \sum_{i=1}^{n} \left(\beta_j - \Big(\sum_{i=1}^{n} \lambda_j^2 \alpha_j \beta_j \Big/ \sum_{i=1}^{n} \lambda_j^2 \alpha_j^2 \Big) \alpha_j \right) \lambda_j \mathbf{u}_j \, .$$

We notice that $\mathbf{v}$ is just mapped to the next iterate in the regular direct power iteration (9.3.1.12). After many steps, it will be very close to $\mathbf{u}_n$, and, therefore, we may now assume $\mathbf{v} = \mathbf{u}_n \Leftrightarrow \alpha_j = \delta_{j,n}$ (Kronecker symbol).

$$\mathbf{z} := \mathbf{A}\mathbf{w} - (\mathbf{v}_0^\top \mathbf{A}\mathbf{w})\mathbf{v}_0 = 0 \cdot \mathbf{u}_n + \sum_{i=1}^{n-1} \lambda_j \beta_j \mathbf{u}_j \, ,$$

$$\mathbf{w}^{(\mathrm{new})} := \frac{\mathbf{z}}{\|\mathbf{z}\|_2} = \Big(\sum_{i=1}^{n-1} \lambda_j^2 \beta_j^2\Big)^{-1/2} \sum_{i=1}^{n-1} \lambda_j \beta_j \mathbf{u}_j \, .$$

The sequence $\mathbf{w}^{(k)}$ produced by repeated application of the mapping given by Code 9.3.4.19 asymptotically (that is, when $\mathbf{v}^{(k)}$ has already converged to $\mathbf{u}_n$) agrees with the sequence produced by the direct power method for $\widetilde{\mathbf{A}} := \mathbf{U}\,\mathrm{diag}(\lambda_1,\ldots,\lambda_{n-1},0)$. Its convergence will be governed by the relative gap $\lambda_{n-2}/\lambda_{n-1}$, see Thm. 9.3.1.21.

However: if $\mathbf{v}^{(k)}$ itself converges slowly, this reasoning does not apply.

EXAMPLE 9.3.4.20 (Subspace power iteration with orthogonal projection)

☞ construction of matrix $\mathbf{A} = \mathbf{A}^\top$ as in Ex. 9.3.2.37

MATLAB-code 9.3.4.21: power iteration with orthogonal projection for two vectors

$\sigma(\mathbf{A}) = \{1, 2, \ldots, 10\}$:

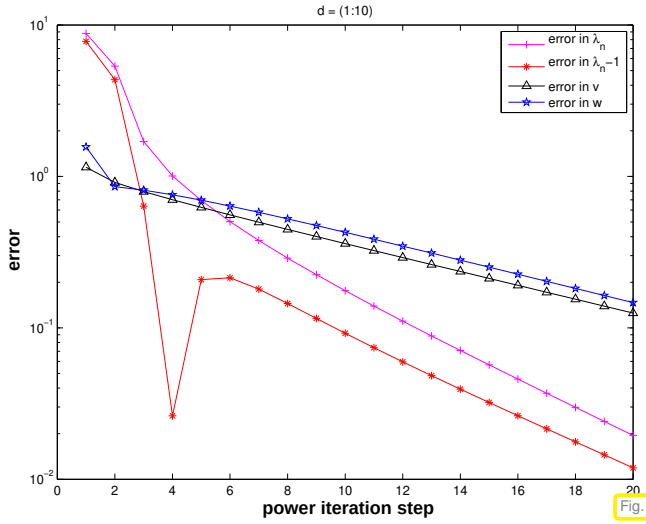


Fig. 361

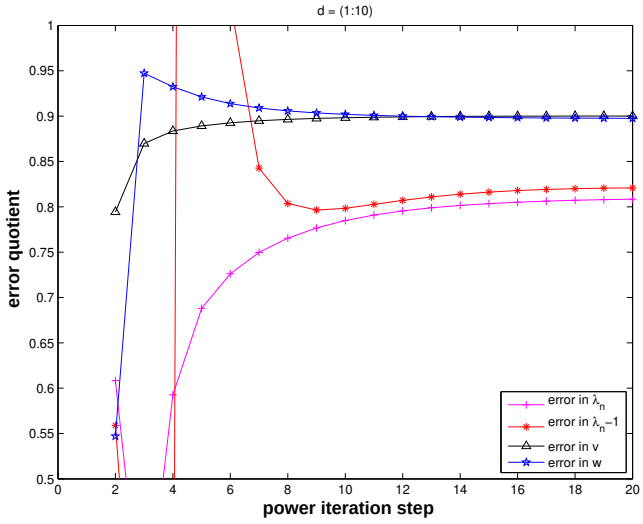


Fig. 362

$\sigma(\mathbf{A}) = \{0.5, 1, \ldots, 4, 9.5, 10\}$:

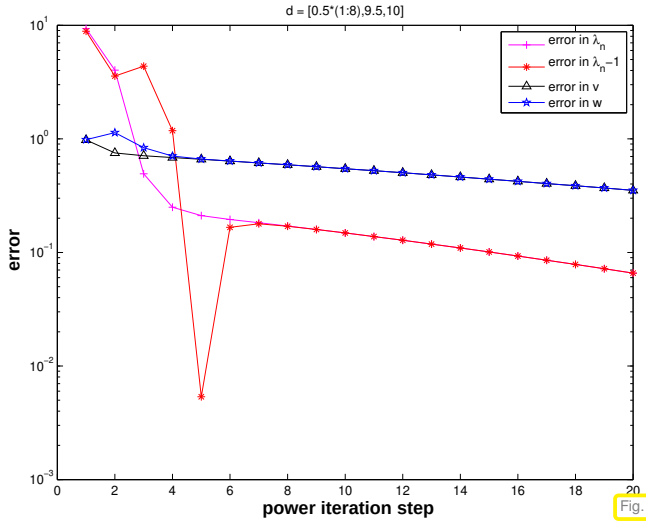


Fig. 363

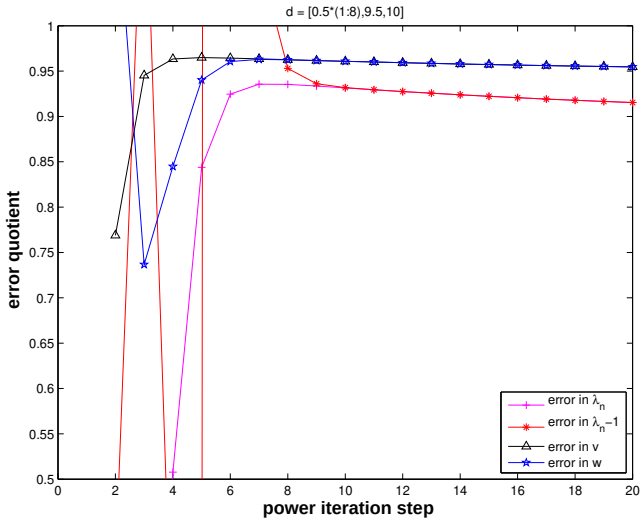


Fig. 364

Issue: generalization of orthogonalization idea to subspaces of dimension > 2

Nothing new:

Gram-Schmidt orthonormalization

($\rightarrow$ [NS02, Thm. 4.8], [Gut09, Alg. 6.1], [QSS00, Sect. 3.4.3])

Given: linearly independent vectors $\mathbf{v}_1, \ldots, \mathbf{v}_m \in \mathbb{R}^n$, $m \in \mathbb{N}$
Sought: vectors $\mathbf{q}_1, \ldots, \mathbf{q}_m \in \mathbb{R}^n$ such that

❷ $\mathbf{q}_l^\top \mathbf{q}_k = \delta_{lk}$ (orthonormality) , (9.3.4.22)

❷ $\mathrm{Span}\{\mathbf{q}_1, \ldots, \mathbf{q}_k\} = \mathrm{Span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ for all $k = 1, \ldots, m$. (9.3.4.23)

Constructive proof & algorithm for Gram-Schmidt orthonormalization:

$$\begin{aligned}
\mathbf{z}_1 &= \mathbf{v}_1 \,, \\
\mathbf{z}_2 &= \mathbf{v}_2 - \frac{\mathbf{v}_2^\top \mathbf{z}_1}{\mathbf{z}_1^\top \mathbf{z}_1}\mathbf{z}_1 \,, \\
\mathbf{z}_3 &= \mathbf{v}_3 - \frac{\mathbf{v}_3^\top \mathbf{z}_1}{\mathbf{z}_1^\top \mathbf{z}_1}\mathbf{z}_1 - \frac{\mathbf{v}_3^\top \mathbf{z}_2}{\mathbf{z}_2^\top \mathbf{z}_2}\mathbf{z}_2 \,, \\
&\vdots
\end{aligned} \tag{9.3.4.24}$$

+ normalization $$\mathbf{q}_k = \frac{\mathbf{z}_k}{\|\mathbf{z}_k\|_2} \,, \quad k = 1, \ldots, m \,. \tag{9.3.4.25}$$

Easy computation: the vectors $\mathbf{q}_1, \ldots, \mathbf{q}_m$ produced by (9.3.4.24) satisfy (9.3.4.22) and (9.3.4.23).

MATLAB-code 9.3.4.26: Gram-Schmidt orthonormalization (do not use, unstable algorithm!)

Warning! Code 9.3.4.26 provides an unstable implementation of Gram-Schmidt orthonormalization: for large n, m impact of round-off will destroy the orthogonality of the columns of $\mathbf{Q}$.

A stable implementation of Gram-Schmidt orthogonalization of the columns of a matrix $\mathbf{V} \in \mathbb{K}^{n,m}$, $m \leq n$, is provided by the following MATLAB command:

```
[Q,~] = qr(V,0)
```

(Asymptotic computational cost: $O(m^2 n)$)

dummy return value (for our purposes) — dummy argument

Detailed description of the algorithm behind `qr` and meaning of the return value `R` → Section 3.3.3.

EXAMPLE 9.3.4.27 (`qr` based orthogonalization, $m = 2$)

The following two MATLAB code snippets perform the same function, *cf.* Code 9.3.4.19:

MATLAB-code 9.3.4.28: sspowitstep1.m

MATLAB-code 9.3.4.29: sspowitstep2.m

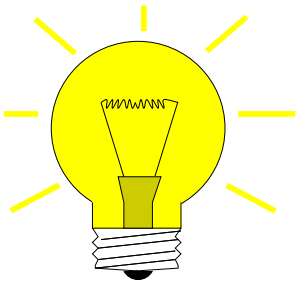
Explanation ➢ discussion of Gram-Schmidt orthonormalization. ⌟

MATLAB-code 9.3.4.30: General subspace power iteration step with `qr` based orthonormalization

9.3.4.2 Ritz projection

Observations on Code 9.3.4.19:

- the first column of $\mathbf{V}$, $(\mathbf{V})_{:,1}$, is a sequence of vectors created by the standard direct power method (9.3.1.12).
- reasoning: the other columns of $\mathbf{V}$, after each multiplication with $\mathbf{A}$ can be expected to contain a significant component in the direction of the eigenvector associated with the eigenvalue of largest modulus.



Idea: use information in $(\mathbf{V})_{:,2},\ldots,(\mathbf{V})_{:,\mathrm{end}}$ to accelerate convergence of $(\mathbf{V})_{:,1}$.

Since the columns of $\mathbf{V}$ span a *subspace* of $\mathbb{R}^n$, this idea can be recast as the following task:

Task:	given $\mathbf{v}_1,\ldots,\mathbf{v}_k \in \mathbb{K}^n$, $k \ll n$, extract (good approximations of) eigenvectors of $\mathbf{A} = \mathbf{A}^{\mathrm{H}} \in \mathbb{K}^{n,n}$ contained in $\mathrm{Span}\{\mathbf{v}_1,\ldots,\mathbf{v}_m\}$.

We take for granted that $\{\mathbf{v}_1,\ldots,\mathbf{v}_m\}$ is linearly independent.

Assumption: $\mathbf{EigA}\lambda \cap V := \mathrm{Span}\{\mathbf{v}_1,\ldots,\mathbf{v}_m\} \neq \{0\}$

$\Leftrightarrow$ V contains an eigenvector of $\mathbf{A}$

$$\begin{aligned} &\Leftrightarrow \quad \exists \mathbf{w} \in V \setminus \{0\}: \quad \mathbf{A}\mathbf{w} = \lambda \mathbf{w} \\ &\Leftrightarrow \quad \exists \mathbf{u} \in \mathbb{K}^m \setminus \{0\}: \quad \mathbf{A}\mathbf{V}\mathbf{u} = \lambda \mathbf{V}\mathbf{u} \\ &\Rightarrow \quad \exists \mathbf{u} \in \mathbb{K}^m \setminus \{0\}: \quad \mathbf{V}^{\mathrm{H}}\mathbf{A}\mathbf{V}\mathbf{u} = \lambda \mathbf{V}^{\mathrm{H}}\mathbf{V}\mathbf{u}\,, \end{aligned} \tag{9.3.4.31}$$

where $\mathbf{V} := (\mathbf{v}_1,\ldots,\mathbf{v}_m) \in \mathbb{K}^{n,m}$ and we used

$$V = \{\mathbf{V}\mathbf{u} : \mathbf{u} \in \mathbb{K}^m\} \quad \text{(linear combinations of the } \mathbf{v}_i\text{)}.$$

(9.3.4.31) ➢ $\mathbf{u} \in \mathbb{K}^k \setminus \{0\}$ solves $m \times m$ generalized eigenvalue problem

$$(\mathbf{V}^{\mathrm{H}}\mathbf{A}\mathbf{V})\mathbf{u} = \lambda(\mathbf{V}^{\mathrm{H}}\mathbf{V})\mathbf{u}\,. \tag{9.3.4.32}$$

Note:

$\{\mathbf{v}_1,\ldots,\mathbf{v}_m\}$ linearly independent

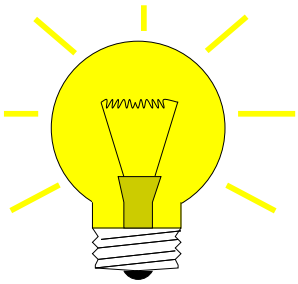
$\Updownarrow$

$\mathbf{V}$ has full rank m ($\to$ Def. 2.2.1.3)

$\Updownarrow$

$\mathbf{V}^{\mathrm{H}}\mathbf{V}$ is symmetric positive definite ($\to$ Def. 1.1.2.6)

If our initial assumption holds true and $\mathbf{u}$ solves (9.3.4.32) and is a simple eigenvalue, a corresponding $\mathbf{x} \in \mathbf{EigA}\lambda$ can be recovered as $\mathbf{x} = \mathbf{V}\mathbf{u}$.



Idea: Given a subspace $V = \mathrm{Im}(\mathbf{V}) \subset \mathbb{K}^n$, $\mathbf{V} \in \mathbb{K}^{n,m}$, $\dim(V) = m$, obtain *approximations* of (a few) eigenvalues and eigenvectors $\mathbf{x}_1,\ldots,\mathbf{x}_m$ of $\mathbf{A}$ by

❶ solving the generalized eigenvalue problem (9.3.4.32)

➔ eigenvectors $\mathbf{u}_1,\ldots,\mathbf{u}_k \in \mathbb{K}^m$ (linearly independent),

❷ and transforming them back according to $\mathbf{x}_k = \mathbf{V}\mathbf{u}_k$, $k = 1,\ldots,m$.

Terminology: (9.3.4.32) is called the Ritz projection of EVP $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ onto V

Terminology: $\sigma(\mathbf{V}^{\mathrm{H}}\mathbf{A}\mathbf{V}) \hat{=}$ Ritz values,
eigenvectors of $\mathbf{V}^{\mathrm{H}}\mathbf{A}\mathbf{V} \hat{=}$ Ritz vectors

Example: Ritz projection of $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ onto $\mathrm{Span}\{\mathbf{v},\mathbf{w}\}$:

$$(\mathbf{v},\mathbf{w})^{\mathrm{H}}\mathbf{A}(\mathbf{v},\mathbf{w})\begin{pmatrix}\alpha\\ \beta\end{pmatrix} = \lambda(\mathbf{v},\mathbf{w})^{\mathrm{H}}(\mathbf{v},\mathbf{w})\begin{pmatrix}\alpha\\ \beta\end{pmatrix} .$$

Note: If $\mathbf{V}$ is *unitary* ($\rightarrow$ Def. 6.3.1.2), then the generalized eigenvalue problem (9.3.4.32) will become a standard linear eigenvalue problem.

Remark 9.3.4.33 (Justification of Ritz projection by min-max theorem)

We revisit $m = 2$, see Code 9.3.4.19. Recall that by the min-max theorem Thm. 9.3.2.18

$$\mathbf{u}_n = \mathrm{argmax}_{\mathbf{x}\in\mathbb{R}^n}\rho_{\mathbf{A}}(\mathbf{x}) \quad , \quad \mathbf{u}_{n-1} = \mathrm{argmax}_{\mathbf{x}\in\mathbb{R}^n,\mathbf{x}\perp\mathbf{u}_n}\rho_{\mathbf{A}}(\mathbf{x}) . \tag{9.3.4.34}$$

Idea: maximize Rayleigh quotient over $\mathrm{Span}\{\mathbf{v},\mathbf{w}\}$, where $\mathbf{v}$, $\mathbf{w}$ are output by Code 9.3.4.19. This leads to the optimization problem

$$(\alpha^*,\beta^*) := \operatorname*{argmax}_{\alpha,\beta\in\mathbb{R},\,\alpha^2+\beta^2=1}\rho_{\mathbf{A}}(\alpha\mathbf{v}+\beta\mathbf{w}) = \operatorname*{argmax}_{\alpha,\beta\in\mathbb{R},\,\alpha^2+\beta^2=1}\rho_{(\mathbf{v},\mathbf{w})^\top\mathbf{A}(\mathbf{v},\mathbf{w})}\left(\begin{pmatrix}\alpha\\ \beta\end{pmatrix}\right) . \tag{9.3.4.35}$$

Then a better approximation for the eigenvector to the largest eigenvalue is

$$\mathbf{v}^* := \alpha^*\mathbf{v} + \beta^*\mathbf{w} .$$

Note that $\|\mathbf{v}^*\|_2 = 1$, if both $\mathbf{v}$ and $\mathbf{w}$ are normalized, which is guaranteed in Code 9.3.4.19.

Then, orthogonalizing $\mathbf{w}$ w.r.t $\mathbf{v}^*$ will produce a new iterate $\mathbf{w}^*$.

Again the min-max theorem Thm. 9.3.2.18 tells us that we can find $(\alpha^*,\beta^*)^\top$ as eigenvector to the largest eigenvalue of

$$(\mathbf{v},\mathbf{w})^\top\mathbf{A}(\mathbf{v},\mathbf{w})\begin{pmatrix}\alpha\\ \beta\end{pmatrix} = \lambda\begin{pmatrix}\alpha\\ \beta\end{pmatrix} . \tag{9.3.4.36}$$

Since eigenvectors of symmetric matrices are mutually orthogonal, we find $\mathbf{w}^* = \alpha_2\mathbf{v} + \beta_2\mathbf{w}$, where $(\alpha_2,\beta_2)^\top$ is the eigenvector of (9.3.4.36) belonging to the smallest eigenvalue. This assumes orthonormal vectors $\mathbf{v}$, $\mathbf{w}$. ⌟

MATLAB-code 9.3.4.37: one step of subspace power iteration with Ritz projection, matrix version

```
function V = sspowitsteprp(A,V)
V = A*V;              % power iteration applied to columns of V
[Q,R] = qr(V,0);      % orthonormalization, see Section 9.3.4.1
[U,D] = eig(Q'*A*Q);  % Solve Ritz projected m × m eigenvalue problem
V = Q*U;              % recover approximate eigenvectors
ev = diag(D);         % approximate eigenvalues
```

Note that he orthogonalization step in Code 9.3.4.37 is actually redundant, if exact arithmetic could be employed, because the Ritz projection could also be realized by solving the generalized eigenvalue problem.

However, prior orthogonalization is essential for numerical stability (→ Def. 1.5.5.19), *cf.* the discussion in Section 3.3.3.

EXAMPLE 9.3.4.38 (Power iteration with Ritz projection)

Matrix as in Ex. 9.3.4.20, $\sigma(\mathbf{A}) = \{0.5, 1, \ldots, 4, 9.5, 10\}$:

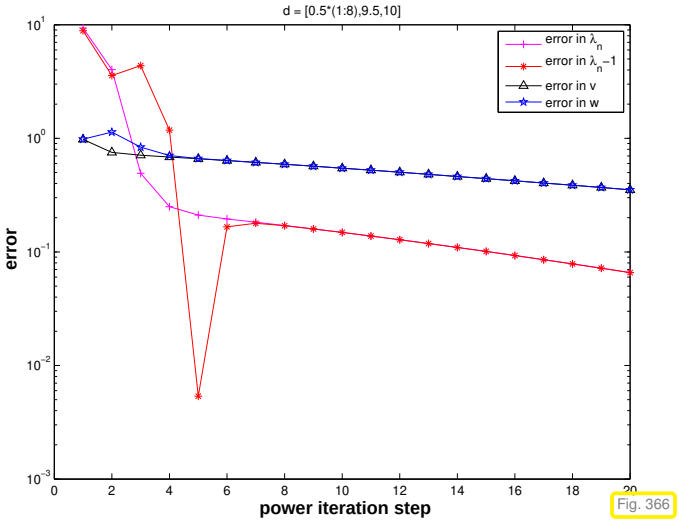


Fig. 365

simple orthonormalization, Ex. 9.3.4.20

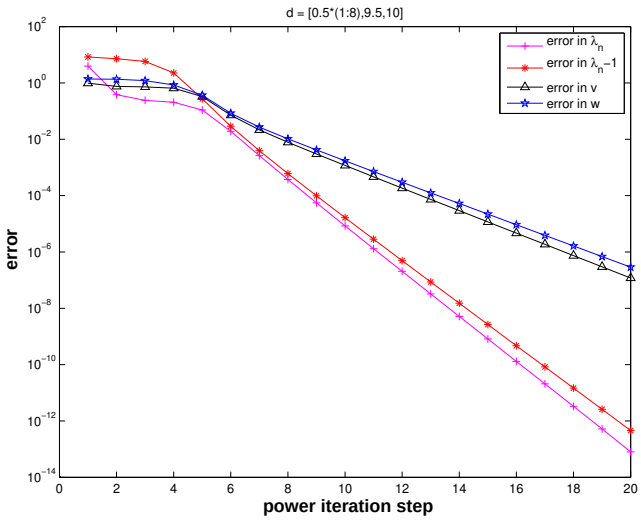


Fig. 366

with Ritz projection, Code 9.3.4.37

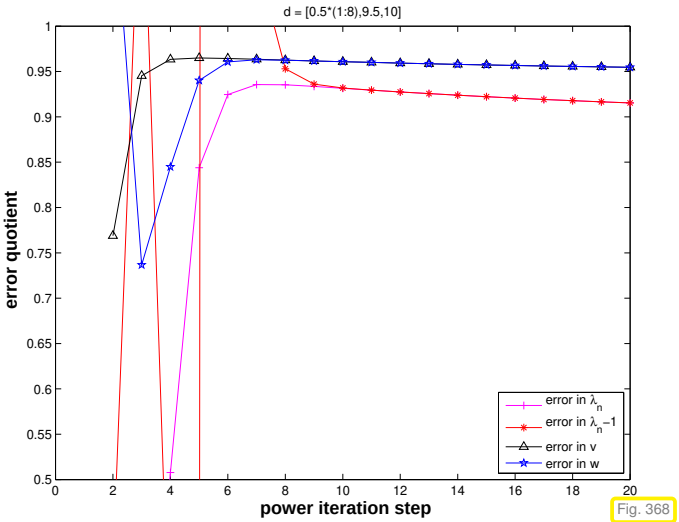


Fig. 367

simple orthonormalization, Ex. 9.3.4.20

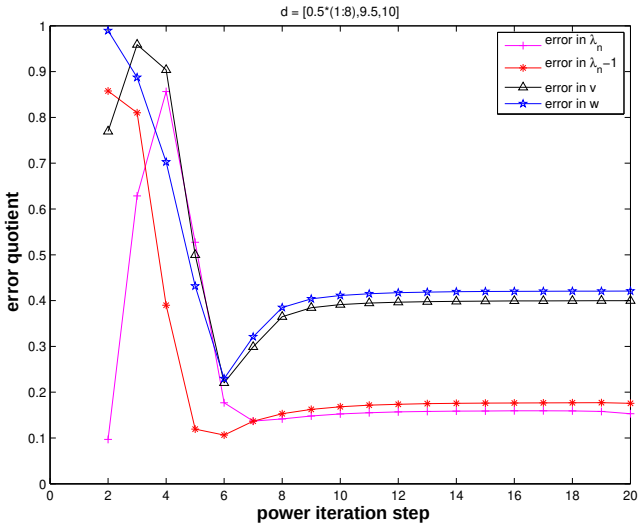


Fig. 368

with Ritz projection, Code 9.3.4.37

Observation: tremendous acceleration of power iteration through Ritz projection, convergence still linear but with much better rates.

⌟

In Code 9.3.4.37: diagonal entries of $\mathbf{D}$ provide approximations of eigenvalues. Their (relative) changes can be used as a termination criterion.

§9.3.4.39 (Subspace variant of direct power method with Ritz projection)

MATLAB-code 9.3.4.40: Subspace power iteration with Ritz projection

⌟

EXAMPLE 9.3.4.41 (Convergence of subspace variant of direct power method)

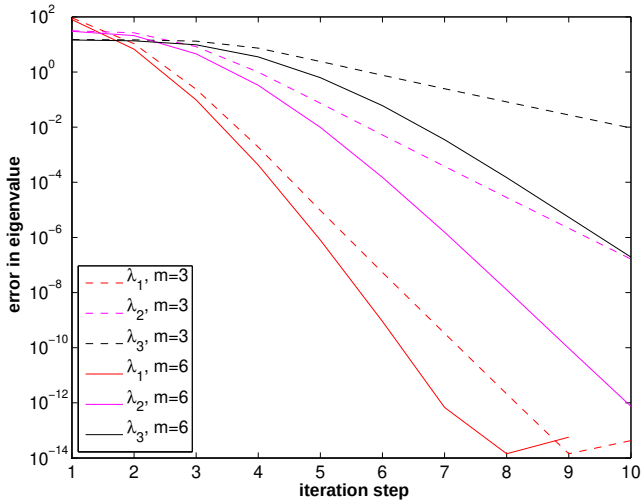


Fig. 369

S.p.d. test matrix: $a_{ij} := \min\{\frac{i}{j}, \frac{j}{i}\}$

```
n=200; A = gallery('lehmer',n);
```

"Initial eigenvector guesses":

```
V = eye(n,m);
```

- Observation: linear convergence of eigenvalues
- choice $m > k$ boosts convergence of eigenvalues

Remark 9.3.4.42 (Subspace power methods)

Analoguous to § 9.3.4.39: construction of subspace variants of inverse iteration (→ Code 9.3.2.31), PINVIT (9.3.3.1), and Rayleigh quotient iteration (9.3.2.36). ⌟

9.4 Krylov Subspace Methods

Supplementary literature. [Han02, Sect. 30]

All power methods (→ Section 9.3) for the eigenvalue problem (EVP) $\mathbf{Ax} = \lambda\mathbf{x}$ only rely on the last iterate to determine the next one (1-point methods, *cf.* (8.2.1.5))

➢ NO MEMORY, *cf.* discussion in the beginning of Section 10.2.

"Memory for power iterations": pursue same idea that led from the gradient method, § 10.1.3.3, to the conjugate gradient method, § 10.2.2.10: use information from previous iterates to achieve efficient minimization over larger and larger subspaces.

Min-max theorem, Thm. 9.3.2.18	:	$\mathbf{A} = \mathbf{A}^H$	⇒	EVPs	⇔	Finding extrema/stationary points of Rayleigh quotient (→ Def. 9.3.1.16)

Setting: EVP $\mathbf{Ax} = \lambda\mathbf{x}$ for real s.p.d. (→ Def. 1.1.2.6) matrix $\mathbf{A} = \mathbf{A}^T \in \mathbb{R}^{n,n}$

notations used below: $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$: eigenvalues of $\mathbf{A}$, counted with multiplicity, see Def. 9.1.0.1,

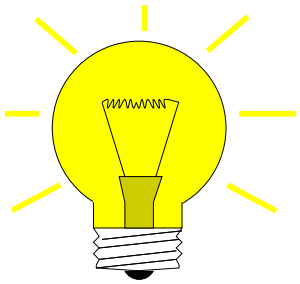
$\mathbf{u}_1, \ldots, \mathbf{u}_n \hat{=}$ corresponding orthonormal eigenvectors, *cf.* Cor. 9.1.0.9.

$$\mathbf{AU} = \mathbf{DU} \quad , \quad \mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_n) \in \mathbb{R}^{n,n} \, , \quad \mathbf{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) \, .$$

We recall

- the direct power method (9.3.1.12) from Section 9.3.1
- and the inverse iteration from Section 9.3.2

and how they produce sequences $(\mathbf{z}^{(k)})_{k\in\mathbb{N}_0}$ of vectors that are supposed to converge to a vector $\in \mathbf{EigA}\lambda_1$ or $\in \mathbf{EigA}\lambda_n$, respectively.



Idea: Better $\mathbf{z}^{(k)}$ from Ritz projection onto $V := \mathrm{Span}\{\mathbf{z}^{(0)},\ldots,\mathbf{z}^{(k)}\}$ (= space spanned by previous iterates)

Recall (→ Code 9.3.4.37) Ritz projection of an EVP $\mathbf{Ax} = \lambda\mathbf{x}$ onto a subspace $V := \mathrm{Span}\{\mathbf{v}_1,\ldots,\mathbf{v}_m\}$, $m < n$ ➡ smaller $m\times m$ generalized EVP

$$\underbrace{\mathbf{V}^T\mathbf{AV}}_{:=\mathbf{H}}\mathbf{x} = \lambda\mathbf{V}^T\mathbf{V}\mathbf{x} \quad , \quad \mathbf{V} := (\mathbf{v}_1,\ldots,\mathbf{v}_m) \in \mathbb{R}^{n,m} \, . \tag{9.4.0.1}$$

From Rayleigh quotient Thm. 9.3.2.16 and considerations in Section 9.3.4.2:

$$\begin{aligned} \mathbf{u}_n \in V \quad &\Rightarrow \quad \text{largest eigenvalue of (9.4.0.1)} = \lambda_{\max}(\mathbf{A}) \, , \\ \mathbf{u}_1 \in V \quad &\Rightarrow \quad \text{smallest eigenvalue of (9.4.0.1)} = \lambda_{\min}(\mathbf{A}) \, . \end{aligned}$$

Intuition: If $\mathbf{u}_n(\mathbf{u}_1)$ "well captured" by V (that is, the angle between the vector and the space V is small), then we can expect that the largest (smallest) eigenvalue of (9.4.0.1) is a good approximation for $\lambda_{\max}(\mathbf{A})(\lambda_{\min}(\mathbf{A}))$, and that, assuming normalization

$$\mathbf{Vw} \approx \mathbf{u}_1 \quad (\text{or } \mathbf{Vw} \approx \mathbf{u}_n) \, ,$$

where $\mathbf{w}$ is the corresponding eigenvector of (9.4.0.1).

For direct power method (9.3.1.12): $\mathbf{z}^{(k)} \| \mathbf{A}^k\mathbf{z}^{(0)}$

$$V = \mathrm{Span}\{\mathbf{z}^{(0)}, \mathbf{Az}^{(0)}, \ldots, \mathbf{A}^{(k)}\mathbf{z}^{(0)}\} = \mathcal{K}_{k+1}(\mathbf{A}, \mathbf{z}^{(0)}) \quad \text{a Krylov space, → Def. 10.2.1.1 .} \tag{9.4.0.2}$$

◁ direct power method with Ritz projection onto Krylov space from (9.4.0.2), *cf.* § 9.3.4.39.

MATLAB-code 9.4.0.3: Ritz projections onto Krylov space (9.4.0.2)

Note: implementation for demonstration purposes only (inefficient for sparse matrix $\mathbf{A}$!)

EXAMPLE 9.4.0.4 (Ritz projections onto Krylov space)

MATLAB-code 9.4.0.5:

```
n=100;
M=gallery('tridiag',-0.5*ones(n-1,1),2*ones(n,1),-1.5*ones(n-1,1));
[Q,R]=qr(M); A=Q'*diag(1:n)*Q; % synthetic matrix,
    σ(A) = {1,2,3,...,100}
```

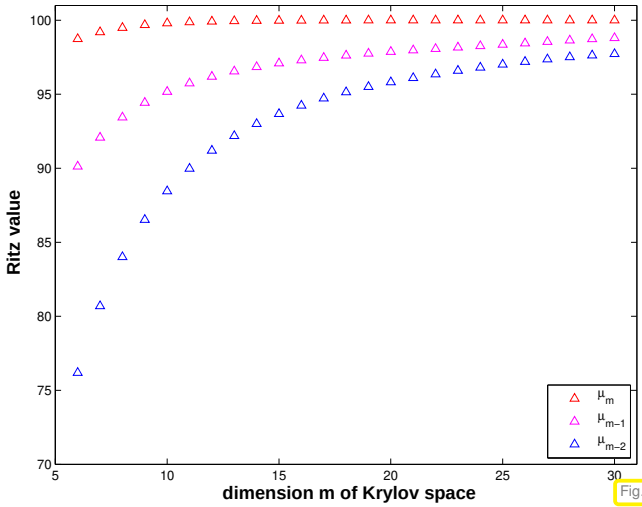


Fig. 370

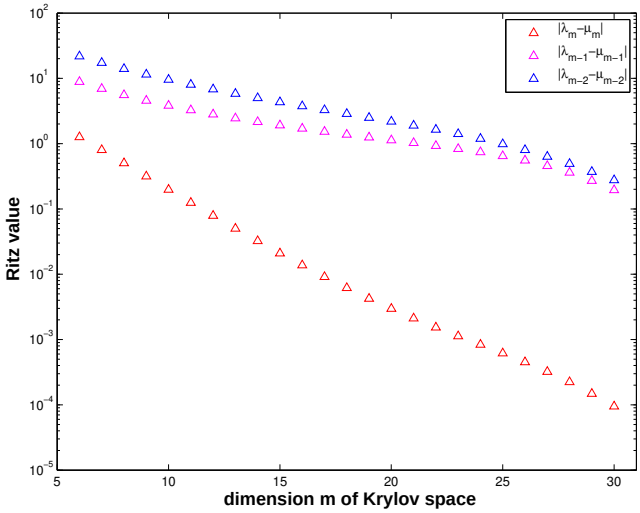


Fig. 371

Observation: "vaguely linear" convergence of largest Ritz values (notation μ_i) to largest eigenvalues. Fastest convergence of largest Ritz value $\rightarrow$ largest eigenvalue of $\mathbf{A}$

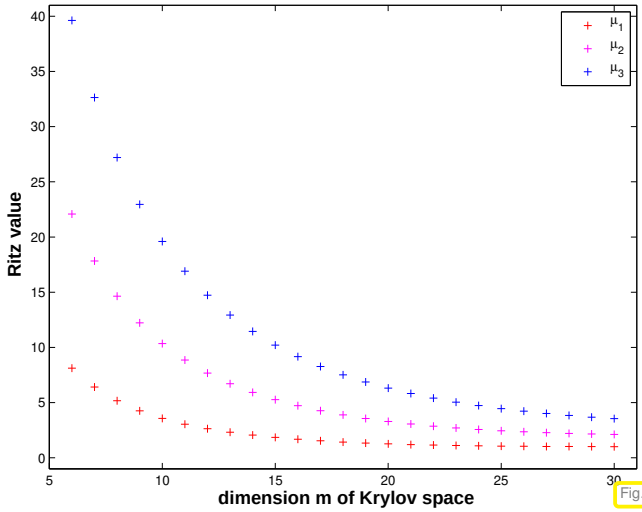


Fig. 372

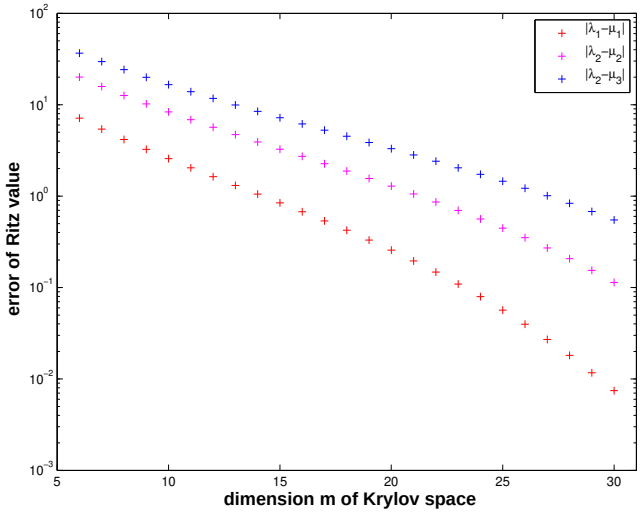


Fig. 373

Observation: *Also the smallest Ritz values converge* "vaguely linearly" to the smallest eigenvalues of $\mathbf{A}$. Fastest convergence of smallest Ritz value $\rightarrow$ smallest eigenvalue of $\mathbf{A}$. ⌟

? Why do smallest Ritz values converge to smallest eigenvalues of $\mathbf{A}$?

Consider direct power method (9.3.1.12) for $\widetilde{\mathbf{A}} := \nu\mathbf{I} - \mathbf{A}$, $\nu > \lambda_{\max}(\mathbf{A})$:

$$\mathbf{z}^{(0)} \text{ arbitrary}, \quad \widetilde{\mathbf{z}}^{(k+1)} = \frac{(\nu\mathbf{I} - \mathbf{A})\widetilde{\mathbf{z}}^{(k)}}{\left\|(\nu\mathbf{I} - \mathbf{A})\widetilde{\mathbf{z}}^{(k)}\right\|_2} \tag{9.4.0.6}$$

As $\sigma(\nu\mathbf{I} - \mathbf{A}) = \nu - \sigma(\mathbf{A})$ and eigenspaces agree, we infer from Thm. 9.3.1.21

$$\lambda_1 < \lambda_2 \quad \Rightarrow \quad \mathbf{z}^{(k)} \overset{k\to\infty}{\longrightarrow} \mathbf{u}_1 \quad \& \quad \rho_{\mathbf{A}}(\mathbf{z}^{(k)}) \overset{k\to\infty}{\longrightarrow} \lambda_1 \quad \text{linearly}. \tag{9.4.0.7}$$

By the binomial theorem (also applies to matrices, if they commute)

$$(\nu\mathbf{I} - \mathbf{A})^k = \sum_{j=0}^{k} \binom{k}{j} \nu^{k-j}\mathbf{A}^j \quad \Rightarrow \quad (\nu\mathbf{I} - \mathbf{A})^k\widetilde{\mathbf{z}}^{(0)} \in \mathcal{K}_k(\mathbf{A}, \mathbf{z}^{(0)}),$$

$$\boxed{\mathcal{K}_k(\nu\mathbf{I} - \mathbf{A}, \mathbf{x}) = \mathcal{K}_k(\mathbf{A}, \mathbf{x})}. \tag{9.4.0.8}$$

➢ $\mathbf{u}_1$ can also be expected to be "well captured" by $\mathcal{K}_k(\mathbf{A},\mathbf{x})$ and the smallest Ritz value should provide a good aproximation for $\lambda_{\min}(\mathbf{A})$.

Recall from Section 10.2.2 Lemma 10.2.2.5 :

Residuals $\mathbf{r}_0,\ldots,\mathbf{r}_{m-1}$ generated in CG iteration, § 10.2.2.10 applied to $\mathbf{Ax}=\mathbf{z}$ with $\mathbf{x}^{(0)}=0$, provide *orthogonal basis* for $\mathcal{K}_m(\mathbf{A},\mathbf{z})$ (, if $\mathbf{r}_k\neq 0$).

Inexpensive Ritz projection of $\mathbf{Ax}=\lambda\mathbf{x}$ onto $\mathcal{K}_m(\mathbf{A},\mathbf{z})$: orthogonal matrix

$$\mathbf{V}_m^T\mathbf{A}\mathbf{V}_m\mathbf{x}=\lambda\mathbf{x}\,,\quad \mathbf{V}_m:=\left(\frac{\mathbf{r}_0}{\|\mathbf{r}_0\|},\ldots,\frac{\mathbf{r}_{m-1}}{\|\mathbf{r}_{m-1}\|}\right)\in\mathbb{R}^{n,m}\,. \tag{9.4.0.9}$$

recall: residuals generated by *short recursions*, see § 10.2.2.10

Lemma 9.4.0.10. Tridiagonal Ritz projection from CG residuals

$\mathbf{V}_m^T\mathbf{A}\mathbf{V}_m$ *is a tridiagonal matrix.*

Proof. Lemma 10.2.2.5: $\{\mathbf{r}_0,\ldots,\mathbf{r}_{\ell-1}\}$ is an orthogonal basis of $\mathcal{K}_\ell(\mathbf{A},\mathbf{r}_0)$, if all the residuals are non-zero. As $\mathbf{A}\mathcal{K}_{\ell-1}(\mathbf{A},\mathbf{r}_0)\subset\mathcal{K}_\ell(\mathbf{A},\mathbf{r}_0)$, we conclude the orthogonality $\mathbf{r}_m^T\mathbf{A}\mathbf{r}_j$ for all $j=0,\ldots,m-2$. Since

$$\left(\mathbf{V}_m^T\mathbf{A}\mathbf{V}_m\right)_{ij}=\mathbf{r}_{i-1}^T\mathbf{A}\mathbf{r}_{j-1}\,,\quad 1\le i,j\le m\,,$$

the assertion of the theorem follows. □

$$\mathbf{V}_l^H\mathbf{A}\mathbf{V}_l=\begin{bmatrix}\alpha_1 & \beta_1 & & & & & \\ \beta_1 & \alpha_2 & \beta_2 & & & & \\ & \beta_2 & \alpha_3 & \ddots & & & \\ & & \ddots & \ddots & & & \\ & & & & \ddots & & \\ & & & & \ddots & \ddots & \beta_{k-1}\\ & & & & & \beta_{k-1} & \alpha_k\end{bmatrix}=:\mathbf{T}_l\in\mathbb{K}^{k,k}\quad\text{[tridiagonal matrix]} \tag{9.4.0.11}$$

Algorithm for computing $\mathbf{V}_l$ and $\mathbf{T}_l$:
Lanczos process
Computational effort/step:

1×	$\mathbf{A}\times$vector
2	dot products
2	AXPY-operations
1	division

Closely related to CG iteration, § 10.2.2.10, Code 10.2.2.11.

MATLAB-code 9.4.0.12: Lanczos process, *cf.* Code 10.2.2.11

Total computational effort for l steps of Lanczos process, if $\mathbf{A}$ has at most k non-zero entries per row: $O(nkl)$

Note: Code 9.4.0.12 assumes that no residual vanishes. This could happen, if $\mathbf{z}_0$ exactly belonged to the span of a few eigenvectors. However, in practical computations inevitable round-off errors will always ensure that the iterates do not stay in an invariant subspace of $\mathbf{A}$, *cf.* Rem. 9.3.1.22.

Convergence (what we expect from the above considerations) → [DH03, Sect. 8.5])

$$\text{In } l\text{-th step:} \quad \lambda_n \approx \mu_l^{(l)}\,,\ \lambda_{n-1} \approx \mu_{l-1}^{(l)}, \dots, \lambda_1 \approx \mu_1^{(l)}\,,$$
$$\sigma(\mathbf{T}_l) = \{\mu_1^{(l)}, \dots, \mu_l^{(l)}\}, \quad \mu_1^{(l)} \le \mu_2^{(l)} \le \cdots \le \mu_l^{(l)}\,.$$

EXAMPLE 9.4.0.13 (Lanczos process for eigenvalue computation)

$\mathbf{A}$ from Ex. 9.4.0.4

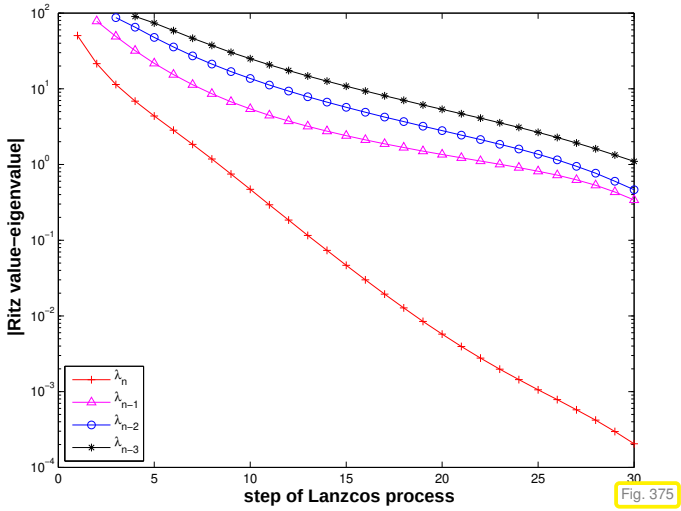


Fig. 374

```
A = gallery('minij',100);
```

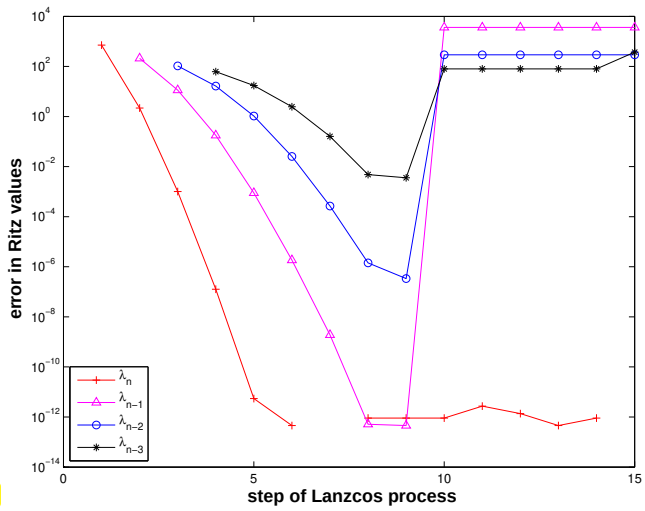


Fig. 375

Observation: same as in Ex. 9.4.0.4, linear convergence of Ritz values to eigenvalues.

However for $\mathbf{A} \in \mathbb{R}^{10,10}$, $a_{ij} = \min\{i,j\}$ good initial convergence, but sudden "jump" of Ritz values off eigenvalues!

Conjecture: Impact of roundoff errors, *cf.* Ex. 10.2.3.1

⌟

EXAMPLE 9.4.0.14 (Impact of roundoff on Lanczos process)

$$\mathbf{A} \in \mathbb{R}^{10,10} \quad , \quad a_{ij} = \min\{i,j\}\,.$$

```
A = gallery('minij',10);
```

Computed by `[V,alpha,beta] = lanczos(A,n,ones(n,1));`, see Code 9.4.0.12:

$$\mathbf{T} = \begin{pmatrix}
38.500000 & 14.813845 & & & & & & & & \\
14.813845 & 9.642857 & 2.062955 & & & & & & & \\
 & 2.062955 & 2.720779 & 0.776284 & & & & & & \\
 & & 0.776284 & 1.336364 & 0.385013 & & & & & \\
 & & & 0.385013 & 0.826316 & 0.215431 & & & & \\
 & & & & 0.215431 & 0.582380 & 0.126781 & & & \\
 & & & & & 0.126781 & 0.446860 & 0.074650 & & \\
 & & & & & & 0.074650 & 0.363803 & 0.043121 & \\
 & & & & & & & 0.043121 & 3.820888 & 11.991094 \\
 & & & & & & & & 11.991094 & 41.254286
\end{pmatrix}$$

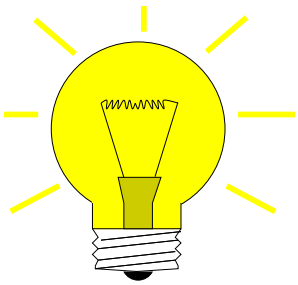
$$\sigma(\mathbf{A}) = \{0.255680,0.273787,0.307979,0.366209,0.465233,0.643104,1.000000,1.873023,5.048917,44.766069\}$$
$$\sigma(\mathbf{T}) = \{0.263867,0.303001,0.365376,0.465199,0.643104,1.000000,1.873023,5.048917,44.765976,44.766069\}$$

► Uncanny cluster of computed eigenvalues of $\mathbf{T}$ ("ghost eigenvalues", [GV89, Sect. 9.2.5])

$$\mathbf{V}^H\mathbf{V} = \begin{pmatrix}
1.000000 & 0.000000 & 0.000000 & 0.000000 & 0.000000 & 0.000000 & 0.000000 & 0.000251 & 0.258801 & 0.883711 \\
0.000000 & 1.000000 & -0.000000 & 0.000000 & 0.000000 & 0.000000 & 0.000000 & 0.000106 & 0.109470 & 0.373799 \\
0.000000 & -0.000000 & 1.000000 & 0.000000 & 0.000000 & 0.000000 & 0.000000 & 0.000005 & 0.005373 & 0.018347 \\
0.000000 & 0.000000 & 0.000000 & 1.000000 & -0.000000 & 0.000000 & 0.000000 & 0.000000 & 0.000096 & 0.000328 \\
0.000000 & 0.000000 & 0.000000 & -0.000000 & 1.000000 & 0.000000 & 0.000000 & 0.000000 & 0.000001 & 0.000003 \\
0.000000 & 0.000000 & 0.000000 & 0.000000 & 0.000000 & 1.000000 & -0.000000 & 0.000000 & 0.000000 & 0.000000 \\
0.000000 & 0.000000 & 0.000000 & 0.000000 & 0.000000 & -0.000000 & 1.000000 & -0.000000 & 0.000000 & 0.000000 \\
0.000251 & 0.000106 & 0.000005 & 0.000000 & 0.000000 & 0.000000 & -0.000000 & 1.000000 & -0.000000 & 0.000000 \\
0.258801 & 0.109470 & 0.005373 & 0.000096 & 0.000001 & 0.000000 & 0.000000 & -0.000000 & 1.000000 & 0.000000 \\
0.883711 & 0.373799 & 0.018347 & 0.000328 & 0.000003 & 0.000000 & 0.000000 & 0.000000 & 0.000000 & 1.000000
\end{pmatrix}$$

► Loss of orthogonality of residual vectors due to roundoff
(compare: impact of roundoff on CG iteration, Ex. 10.2.3.1

l	$\sigma(\mathbf{T}_l)$									
1										38.500000
2									3.392123	44.750734
3								1.117692	4.979881	44.766064
4							0.597664	1.788008	5.048259	44.766069
5						0.415715	0.925441	1.870175	5.048916	44.766069
6					0.336507	0.588906	0.995299	1.872997	5.048917	44.766069
7				0.297303	0.431779	0.638542	0.999922	1.873023	5.048917	44.766069
8			0.276160	0.349724	0.462449	0.643016	1.000000	1.873023	5.048917	44.766069
9		0.276035	0.349451	0.462320	0.643006	1.000000	1.873023	3.821426	5.048917	44.766069
10	0.263867	0.303001	0.365376	0.465199	0.643104	1.000000	1.873023	5.048917	44.765976	44.766069



Idea:

✦ do not rely on orthogonality relations of Lemma 10.2.2.5

✦ use explicit Gram-Schmidt orthogonalization [NS02, Thm. 4.8], [Gut09, Alg .6.1]

Details: inductive approach: given $\{\mathbf{v}_1,\ldots,\mathbf{v}_l\}$ ONB of $\mathcal{K}_l(\mathbf{A},\mathbf{z})$

► $$\widetilde{\mathbf{v}}_{l+1} := \mathbf{A}\mathbf{v}_l - \sum_{j=1}^{l}(\mathbf{v}_j^H\mathbf{A}\mathbf{v}_l)\,\mathbf{v}_j\,,\quad \mathbf{v}_{l+1} := \frac{\widetilde{\mathbf{v}}_{l+1}}{\|\widetilde{\mathbf{v}}_{l+1}\|_2} \;\Rightarrow\; \mathbf{v}_{l+1} \perp \mathcal{K}_l(\mathbf{A},\mathbf{z})\,. \qquad (9.4.0.15)$$

(Gram-Schmidt, *cf.* (10.2.2.4)) orthogonal

► Arnoldi process: In step l:

$1\times$	$\mathbf{A}\times$vector
$l+1$	dot products
l	AXPY-operations
n	divisions

➢ Computational cost for l steps, if at most k non-zero entries in each row of $\mathbf{A}$: $O(nkl^2)$

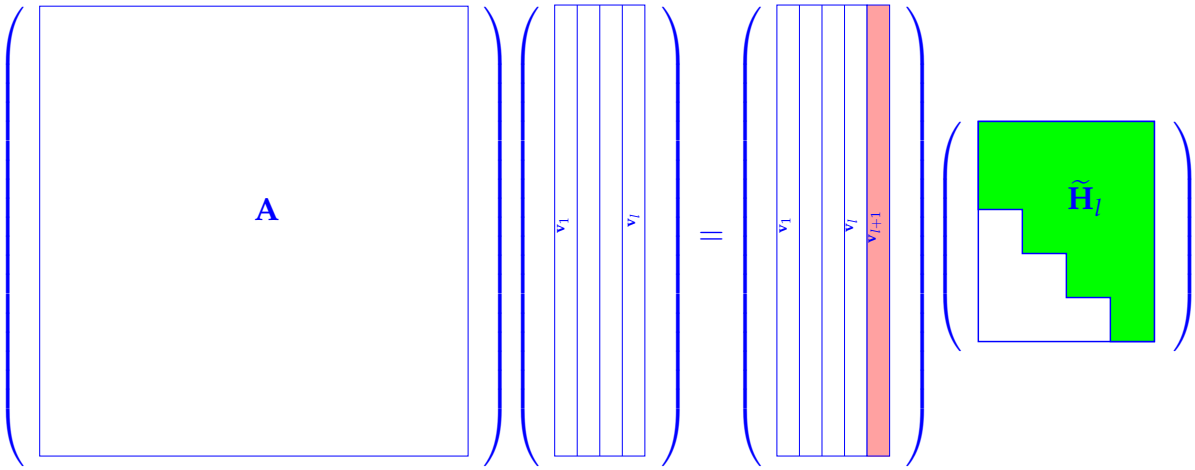
MATLAB-code 9.4.0.16: Arnoldi process

If it does not stop prematurely, the Arnoldi process of Code 9.4.0.16 will yield an *orthonormal basis* (ONB) of $\mathcal{K}_{k+1}(\mathbf{A},\mathbf{v}_0)$ for a **general** $\mathbf{A}\in\mathbb{C}^{n,n}$.

Algebraic view of the Arnoldi process of Code 9.4.0.16, meaning of output `H`:

$$\mathbf{V}_l=\left[\mathbf{v}_1,\ldots,\mathbf{v}_l\right]:\quad \mathbf{A}\mathbf{V}_l=\mathbf{V}_{l+1}\widetilde{\mathbf{H}}_l\quad,\quad \widetilde{\mathbf{H}}_l\in\mathbb{K}^{l+1,l}\text{ mit }\widetilde{h}_{ij}=\begin{cases}\mathbf{v}_i^H\mathbf{A}\mathbf{v}_j & \text{, if } i\le j\,,\\ \|\widetilde{\mathbf{v}}_i\|_2 & \text{, if } i=j+1\,,\\ 0 & \text{else.}\end{cases}$$

➡ $\widetilde{\mathbf{H}}_l$ = non-square upper Hessenberg matrices



Translate Code 9.4.0.16 to matrix calculus:

Lemma 9.4.0.17. Theory of Arnoldi process

For the matrices $\mathbf{V}_l\in\mathbb{K}^{n,l}$, $\widetilde{\mathbf{H}}_l\in\mathbb{K}^{l+1,l}$ arising in the l-th step, $l\le n$, of the Arnoldi process holds

(i) $\mathbf{V}_l^H\mathbf{V}_l=\mathbf{I}$ *(unitary matrix),*

(ii) $\mathbf{A}\mathbf{V}_l=\mathbf{V}_{l+1}\widetilde{\mathbf{H}}_l$, $\widetilde{\mathbf{H}}_l$ *is non-square upper Hessenberg matrix,*

(iii) $\mathbf{V}_l^H\mathbf{A}\mathbf{V}_l=\mathbf{H}_l\in\mathbb{K}^{l,l}$, $h_{ij}=\widetilde{h}_{ij}$ *for* $1\le i,j\le l$,

(iv) *If* $\mathbf{A}=\mathbf{A}^H$ *then* $\mathbf{H}_l$ *is tridiagonal (➢ Lanczos process)*

Proof. Direct from Gram-Schmidt orthogonalization and inspection of Code 9.4.0.16. □

Remark 9.4.0.18 (Arnoldi process and Ritz projection)

Interpretation of Lemma 9.4.0.17 (iii) & (i):

$\mathbf{H}_l\mathbf{x}=\lambda\mathbf{x}$ is a (generalized) Ritz projection of EVP $\mathbf{A}\mathbf{x}=\lambda\mathbf{x}$, *cf.* Section 9.3.4.2.

⌟

➢ Eigenvalue approximation for general EVP $\mathbf{A}\mathbf{x}=\lambda\mathbf{x}$ by Arnoldi process:

MATLAB-code 9.4.0.19: Arnoldi eigenvalue approximation

MATLAB-code 9.4.0.20: MATLAB-CODE Arnoldi eigenvalue approximation

```
[illegible]
    V = [V, vt/Ht(l+1,l)];
end
```

Heuristic termination criterion

Arnoldi process for computing the k largest (in modulus) eigenvalues of $\mathbf{A} \in \mathbb{C}^{n,n}$

1 $\mathbf{A}\times$vector per step
($\succ$ attractive for sparse matrices)

However: required storage increases with number of steps, *cf.* situation with GMRES, Section 10.4.1.

Heuristic termination criterion

EXAMPLE 9.4.0.21 (Stabilty of Arnoldi process)

$\mathbf{A} \in \mathbb{R}^{100,100}$, $a_{ij} = \min\{i,j\}$.

```
A = gallery('minij',100);
```

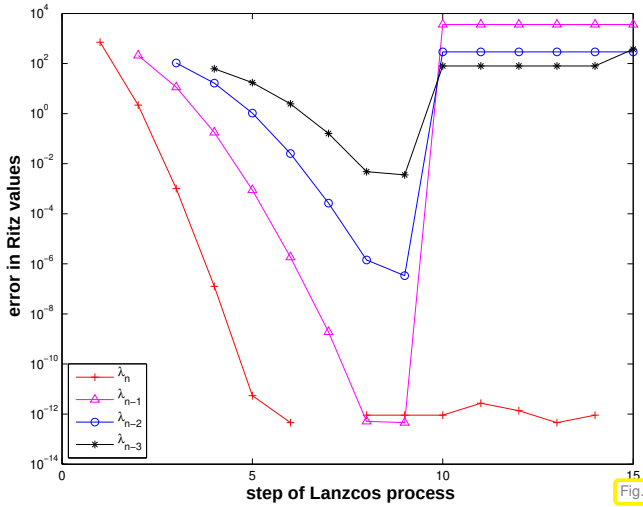


Fig. 376

Lanczos process: Ritz values

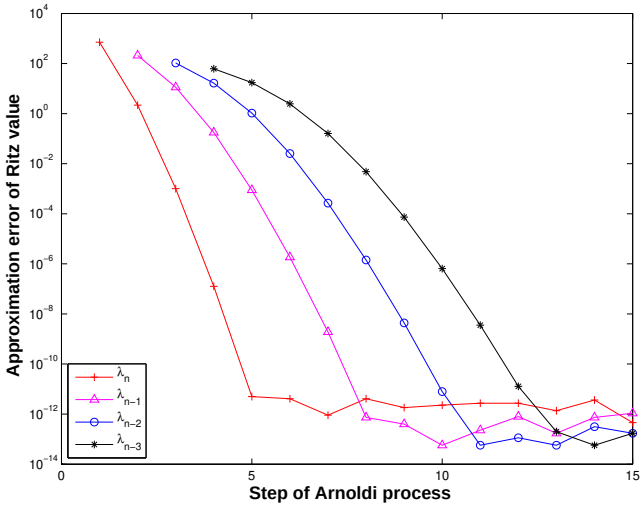


Fig. 377

Arnoldi process: Ritz values

Ritz values during Arnoldi process for `A = gallery('minij',10);` ↔ Ex. 9.4.0.13

l	$\sigma(\mathbf{H}_l)$									
1										38.500000
2									3.392123	44.750734
3								1.117692	4.979881	44.766064
4							0.597664	1.788008	5.048259	44.766069
5						0.415715	0.925441	1.870175	5.048916	44.766069
6					0.336507	0.588906	0.995299	1.872997	5.048917	44.766069
7				0.297303	0.431779	0.638542	0.999922	1.873023	5.048917	44.766069
8			0.276159	0.349722	0.462449	0.643016	1.000000	1.873023	5.048917	44.766069
9		0.263872	0.303009	0.365379	0.465199	0.643104	1.000000	1.873023	5.048917	44.766069
10	0.255680	0.273787	0.307979	0.366209	0.465233	0.643104	1.000000	1.873023	5.048917	44.766069

Observation: (almost perfect approximation of spectrum of $\mathbf{A}$)

For the above examples both the Arnoldi process and the Lanczos process are *algebraically equivalent*, because they are applied to a symmetric matrix $\mathbf{A} = \mathbf{A}^T$. However, they behave strikingly differently, which indicates that they are *not numerically equivalent*.

The Arnoldi process is much less affected by roundoff than the Lanczos process, because it does not take for granted orthogonality of the "residual vector sequence". Hence, the Arnoldi process enjoys superior numerical stability (→ **??**, Def. 1.5.5.19) compared to the Lanczos process. ⌟

EXAMPLE 9.4.0.22 (Eigenvalue computation with Arnoldi process)

Eigenvalue approximation from Arnoldi process for *non-symmetric* $\mathbf{A}$, initial vector `ones(100,1);`

MATLAB-code 9.4.0.23:

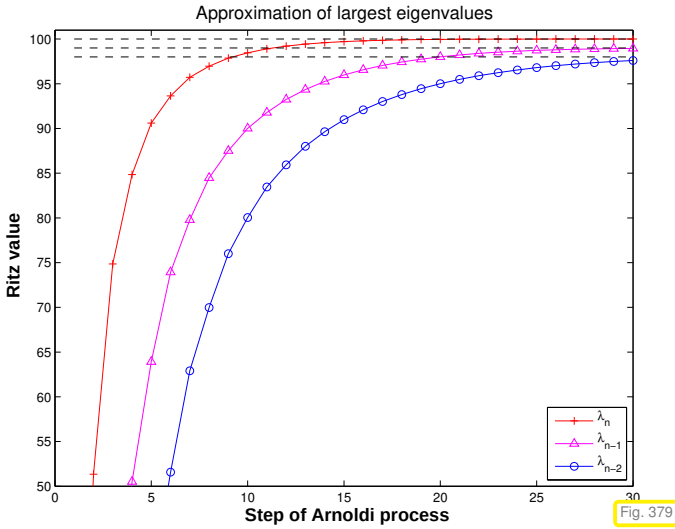


Fig. 378

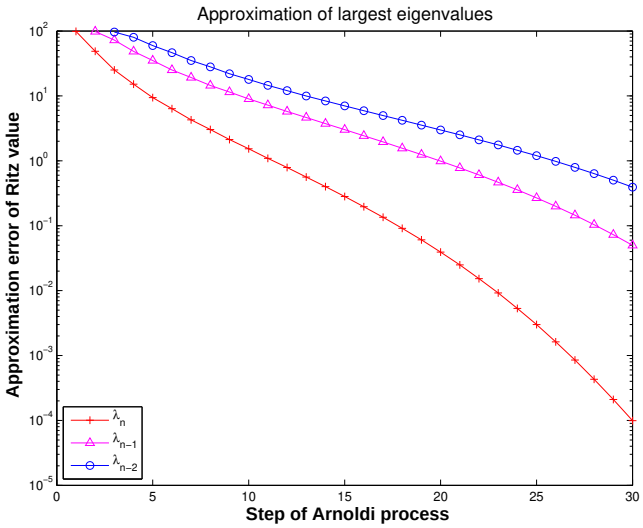


Fig. 379

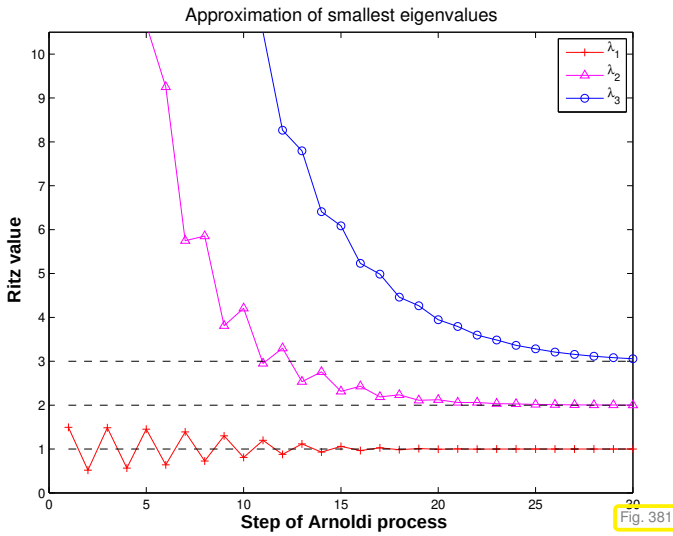


Fig. 380

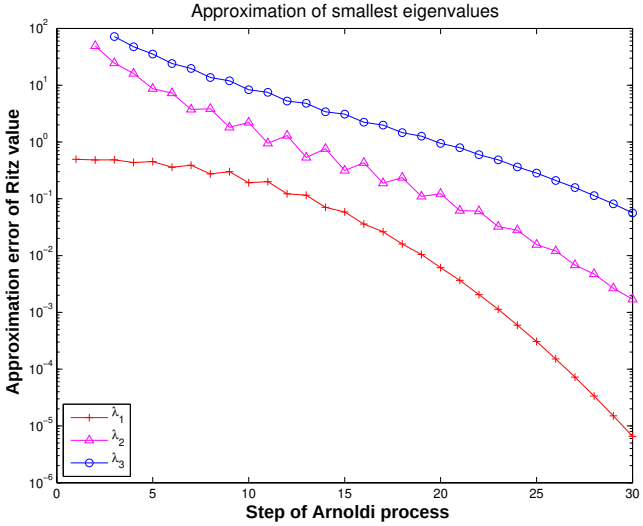


Fig. 381

Observation: "vaguely linear" convergence of largest and smallest eigenvalues, *cf.* Ex. 9.4.0.4. ⌟

Krylov subspace iteration methods (= Arnoldi process, Lanczos process) attractive for computing *a few* of the largest/smallest eigenvalues and associated eigenvectors of *large sparse matrices*.

Remark 9.4.0.24 (Krylov subspace methods for generalized EVP)

Adaptation of Krylov subspace iterative eigensolvers to generalized EVP: $\mathbf{A}\mathbf{x} = \lambda\mathbf{B}\mathbf{x}$, $\mathbf{B}$ s.p.d.: replace Euclidean inner product with "$\mathbf{B}$-inner product" $(\mathbf{x},\mathbf{y}) \mapsto \mathbf{x}^H\mathbf{B}\mathbf{y}$. ⌟

MATLAB-functions:

`d = eigs(A,k,sigma)` : k largest/smallest eigenvalues of $\mathbf{A}$

`d = eigs(A,B,k,sigma)` : k largest/smallest eigenvalues for generalized EVP $\mathbf{A}\mathbf{x} = \lambda\mathbf{B}\mathbf{x}$,$\mathbf{B}$ s.p.d.

`d = eigs(Afun,n,k)` : `Afun` = handle to function providing matrix×vector for $\mathbf{A}/\mathbf{A}^{-1}/\mathbf{A} - \alpha\mathbf{I}/(\mathbf{A} - \alpha\mathbf{B})^{-1}$. (Use flags to tell `eigs` about special properties of matrix behind `Afun`.)

`eigs` just calls routines of the open source ARPACK numerical library.

Bibliography

[AKY99] Charles J Alpert, Andrew B Kahng, and So-Zen Yao. “Spectral partitioning with multiple eigenvectors”. In: *Discrete Applied Mathematics* 90.1-3 (1999), pp. 3–26. DOI: 10.1016/S0166-218X(98)00083-3 (cit. on p. 705).

[Bai+00] Z.-J. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst. *Templates for the Solution of Algebraic Eigenvalue Problems*. Philadelphia, PA: SIAM, 2000 (cit. on p. 682).

[BF06] Yuri Boykov and Gareth Funka-Lea. “Graph Cuts and Efficient N-D Image Segmentation”. In: *International Journal of Computer Vision* 70.2 (Nov. 2006), pp. 109–131. DOI: 10.1007/s11263-006-7934-5.

[DR08] W. Dahmen and A. Reusken. *Numerik für Ingenieure und Naturwissenschaftler*. Heidelberg: Springer, 2008 (cit. on pp. 684, 686, 690, 695, 697, 706).

[DH03] P. Deuflhard and A. Hohmann. *Numerical Analysis in Modern Scientific Computing*. Vol. 43. Texts in Applied Mathematics. Springer, 2003 (cit. on p. 725).

[Gle15] David F. Gleich. “PageRank Beyond the Web”. In: *SIAM Review* 57.3 (2015), pp. 321–363. DOI: 10.1137/140976649 (cit. on p. 690).

[GV89] G.H. Golub and C.F. Van Loan. *Matrix computations*. 2nd. Baltimore, London: John Hopkins University Press, 1989 (cit. on pp. 687, 688, 701, 726).

[GT08] Craig Gotsman and Sivan Toledo. “On the computation of null spaces of sparse rectangular matrices”. In: *SIAM J. Matrix Anal. Appl.* 30.2 (2008), pp. 445–463. DOI: 10.1137/050638369 (cit. on p. 706).

[Gut09] M.H. Gutknecht. *Lineare Algebra*. Lecture Notes. SAM, ETH Zürich, 2009 (cit. on pp. 684–686, 716, 726).

[Hac94] Wolfgang Hackbusch. *Iterative solution of large sparse systems of equations*. Vol. 95. Applied Mathematical Sciences. New York: Springer-Verlag, 1994, pp. xxii+429 (cit. on p. 686).

[Hal70] K.M. Hall. “An r-dimensional quadratic placement algorithm”. In: *Management Science* 17.3 (1970), pp. 219–229.

[Han02] M. Hanke-Bourgeois. *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens*. Mathematische Leitfäden. Stuttgart: B.G. Teubner, 2002 (cit. on pp. 686, 687, 707, 710, 721).

[LMM19] C. Lackner, S. Meng, and P. Monk. “Determination of electromagnetic Bloch variety in a medium with frequency-dependent coefficients”. In: *J. Comput. Appl. Math.* 358 (2019), pp. 359–373. DOI: 10.1016/j.cam.2019.03.027.

[LM06] A.N. Lengville and C.D. Meyer. *Google’s PageRank and Beyond: The Science of Search Engine Rankings*. Princeton, NJ: Princeton University Press, 2006 (cit. on p. 690).

[Lux07] Ulrike von Luxburg. “A tutorial on spectral clustering”. In: *Stat. Comput.* 17.4 (2007), pp. 395–416. DOI: 10.1007/s11222-007-9033-z.

[Ney99a] K. Neymeyr. *A geometric theory for preconditioned inverse iteration applied to a subspace*. Tech. rep. 130. Tübingen, Germany: SFB 382, Universität Tübingen, Nov. 1999 (cit. on p. 709).

[Ney99b] K. Neymeyr. *A geometric theory for preconditioned inverse iteration: III. Sharp convergence estimates*. Tech. rep. 130. Tübingen, Germany: SFB 382, Universität Tübingen, Nov. 1999 (cit. on p. 709).

[NS02] K. Nipp and D. Stoffer. *Lineare Algebra*. 5th ed. Zürich: vdf Hochschulverlag, 2002 (cit. on pp. 684–687, 716, 726).

[QSS00] A. Quarteroni, R. Sacco, and F. Saleri. *Numerical mathematics*. Vol. 37. Texts in Applied Mathematics. New York: Springer, 2000 (cit. on pp. 685–687, 690, 697, 716).

[SM00] J.-B. Shi and J. Malik. "Normalized cuts and image segmentation". In: *IEEE Trans. Pattern Analysis and Machine Intelligence* 22.8 (2000), pp. 888–905 (cit. on pp. 698, 700, 705).

[ST96] D.A. Spielman and Shang-Hua Teng. "Spectral partitioning works: planar graphs and finite element meshes". In: *Foundations of Computer Science, 1996. Proceedings., 37th Annual Symposium on*. Oct. 1996, pp. 96–105. DOI: 10.1109/SFCS.1996.548468 (cit. on p. 705).

[Str09] M. Struwe. *Analysis für Informatiker*. Lecture notes, ETH Zürich. 2009 (cit. on p. 684).

[TM01] Françoise Tisseur and Karl Meerbergen. "The quadratic eigenvalue problem". In: *SIAM Rev.* 43.2 (2001), pp. 235–286. DOI: 10.1137/S0036144500381988 (cit. on p. 683).

Chapter 10

Krylov Methods for Linear Systems of Equations

Supplementary literature. There is a wealth of literature on iterative methods for the solution of linear systems of equations: The two books [Hac94] and [Saa03] offer a comprehensive treatment of the topic (the latter is available online for ETH students and staff).

Concise presentations can be found in [QSS00, Ch. 4] and [DR08, Ch. 13].

Learning outcomes:

- Understanding when and why iterative solution of linear systems of equations may be preferred to direct solvers based on Gaussian elimination.
-

= A class of iterative methods (→ Section 8.2) for approximate solution of large linear systems of equations $\mathbf{Ax} = \mathbf{b}$, $\mathbf{A} \in \mathbb{K}^{n,n}$.

BUT, we have reliable *direct* methods (Gauss elimination → Section 2.3, LU-factorization → § 2.3.2.15, QR-factorization → **??**) that provide an (apart from roundoff errors) exact solution with a *finite* number of elementary operations!

Alas, direct elimination may not be feasible, or may be grossly inefficient, because

- it may be too expensive (e.g. for $\mathbf{A}$ too large, sparse), → (2.3.2.10),
- inevitable fill-in may exhaust main memory,
- the system matrix may be available only as procedure `y=evalA(x)` $\leftrightarrow \mathbf{y} = \mathbf{Ax}$

Contents

10.1 Descent Methods [QSS00, Sect. 4.3.3] 734
10.1.1 Quadratic minimization context 734
10.1.2 Abstract steepest descent 735
10.1.3 Gradient method for s.p.d. linear system of equations 736
10.1.4 Convergence of the gradient method 737
10.2 Conjugate gradient method (CG) [Han02, Ch. 9], [DR08, Sect. 13.4], [QSS00, Sect. 4.3.4]741
10.2.1 Krylov spaces 742
10.2.2 Implementation of CG 743
10.2.3 Convergence of CG 746
10.3 Preconditioning [DR08, Sect. 13.5], [Han02, Ch. 10], [QSS00, Sect. 4.3.5] 750
10.4 Survey of Krylov Subspace Methods 756

10.4.1 Minimal residual methods 756
10.4.2 Iterations with short recursions [QSS00, Sect. 4.5] 757

10.1 Descent Methods [QSS00, Sect. 4.3.3]

Focus:

Linear system of equations $\mathbf{A}\mathbf{x} = \mathbf{b}$, $\mathbf{A} \in \mathbb{R}^{n,n}$, $\mathbf{b} \in \mathbb{R}^n$, $n \in \mathbb{N}$ given,

with symmetric positive definite (s.p.d., → Def. 1.1.2.6) system matrix $\mathbf{A}$

➡

$\mathbf{A}$-inner product $(\mathbf{x}, \mathbf{y}) \mapsto \mathbf{x}^\top \mathbf{A}\mathbf{y}$ ⇒ "$\mathbf{A}$-geometry"

Definition 10.1.0.1. Energy norm → [Han02, Def. 9.1]

A s.p.d. matrix $\mathbf{A} \in \mathbb{R}^{n,n}$ induces an energy norm

$$\|\mathbf{x}\|_A := (\mathbf{x}^\top \mathbf{A}\mathbf{x})^{1/2}, \quad \mathbf{x} \in \mathbb{R}^n .$$

Remark 10.1.0.2 (Krylov methods for complex s.p.d. system matrices) In this chapter, for the sake of simplicity, we restrict ourselves to $\mathbb{K} = \mathbb{R}$.

However, the (conjugate) gradient methods introduced below also work for LSE $\mathbf{A}\mathbf{x} = \mathbf{b}$ with $\mathbf{A} \in \mathbb{C}^{n,n}$, $\mathbf{A} = \mathbf{A}^H$ s.p.d. when $^\top$ is replaced with ^{H} (Hermitian transposed). Then, all theoretical statements remain valid unaltered for $\mathbb{K} = \mathbb{C}$. ⌟

10.1.1 Quadratic minimization context

Lemma 10.1.1.1. S.p.d. LSE and quadratic minimization problem [DR08, (13.37)]

A LSE with $\mathbf{A} \in \mathbb{R}^{n,n}$ s.p.d. and $\mathbf{b} \in \mathbb{R}^n$ is equivalent to a minimization problem:

$$\mathbf{A}\mathbf{x} = \mathbf{b} \quad \Leftrightarrow \quad \mathbf{x} = \arg\min_{\mathbf{y} \in \mathbb{R}^n} J(\mathbf{y}), \quad J(\mathbf{y}) := \tfrac{1}{2}\mathbf{y}^\top \mathbf{A}\mathbf{y} - \mathbf{b}^\top \mathbf{y} . \tag{10.1.1.2}$$

A quadratic functional

Proof. If $\mathbf{x}^* := \mathbf{A}^{-1}\mathbf{b}$ a straightforward computation using $\mathbf{A} = \mathbf{A}^T$ shows

$$\begin{aligned} J(\mathbf{x}) - J(\mathbf{x}^*) &= \tfrac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} - \mathbf{b}^T\mathbf{x} - \tfrac{1}{2}(\mathbf{x}^*)^T\mathbf{A}\mathbf{x}^* + \mathbf{b}^T\mathbf{x}^* \\ &\overset{\mathbf{b}=\mathbf{A}\mathbf{x}^*}{=} \tfrac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} - (\mathbf{x}^*)^T\mathbf{A}\mathbf{x} + \tfrac{1}{2}(\mathbf{x}^*)^T\mathbf{A}\mathbf{x}^* \\ &= \tfrac{1}{2}\|\mathbf{x} - \mathbf{x}^*\|_A^2 . \end{aligned} \tag{10.1.1.3}$$

Then the assertion follows from the properties of the energy norm. □

EXAMPLE 10.1.1.4 (Quadratic functional in 2D) Plot of J from (10.1.1.2) for $\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$, $\mathbf{b} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$.

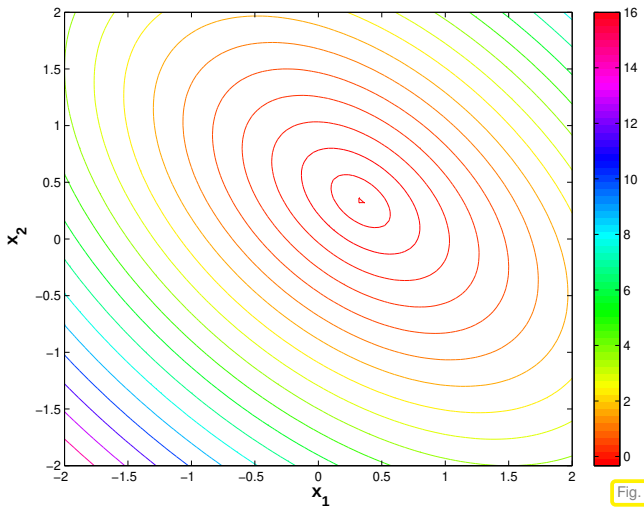


Fig. 382

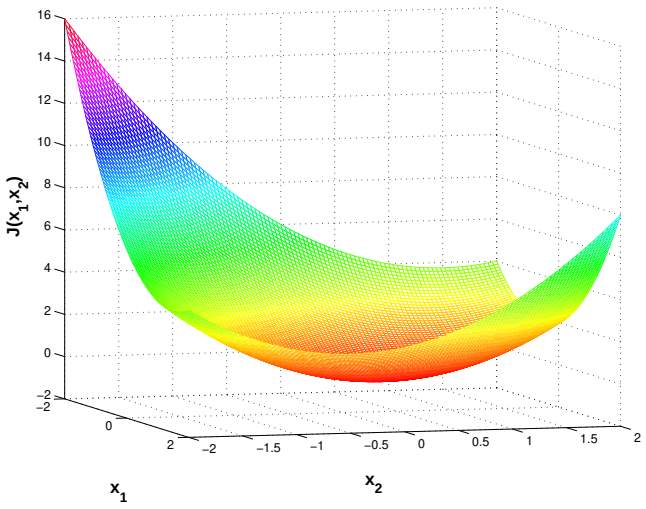
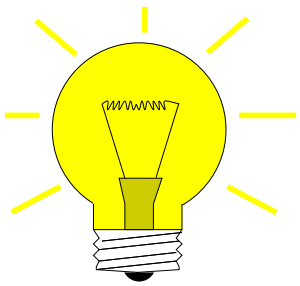


Fig. 383

Level lines of quadratic functionals with s.p.d. $\mathbf{A}$ are (hyper)ellipses



Algorithmic idea: (Lemma 10.1.1.1 ➢) Solve $\mathbf{Ax} = \mathbf{b}$ iteratively by successive solution of *simpler* minimization problems

10.1.2 Abstract steepest descent

Task:	Given continuously differentiable $F : D \subset \mathbb{R}^n \mapsto \mathbb{R}$, find minimizer $\mathbf{x}^* \in D$: $\mathbf{x}^* = \operatorname{argmin}_{\mathbf{x}\in D} F(\mathbf{x})$

Note that a minimizer need not exist, if F is not bounded from below (e.g., $F(x) = x^3$, $x \in \mathbb{R}$, or $F(x) = \log x$, $x > 0$), or if D is open (e.g., $F(x) = \sqrt{x}$, $x > 0$).

The existence of a minimizer is guaranteed if F is bounded from below and D is closed ($\rightarrow$ Analysis).

The most natural iteration:

§10.1.2.1 (Steepest descent (*ger.:* steilster Abstieg))

Initial guess $\mathbf{x}^{(0)} \in D$, $k = 0$
repeat
$\mathbf{d}_k := -\mathbf{grad}\, F(\mathbf{x}^{(k)})$
$t^* := \operatorname{argmin}_{t\in\mathbb{R}} F(\mathbf{x}^{(k)} + t\mathbf{d}_k)$ (line search)
$\mathbf{x}^{(k+1)} := \mathbf{x}^{(k)} + t^*\mathbf{d}_k$
$k := k + 1$
until $\left(\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right\| \leq \tau_{\text{rel}} \left\|\mathbf{x}^{(k)}\right\|$ **or** $\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right\| \leq \tau_{\text{abs}} \right)$

- $\mathbf{d}_k \triangleq$ *direction of steepest descent*
- linear search $\triangleq$ 1D minimization: use Newton's method ($\rightarrow$ Section 8.4.2.1) on derivative
- correction based a posteriori termination criterion, see Section 8.2.3 for a discussion.
($\tau \triangleq$ prescribed tolerance)

The gradient (→ [Str09, Kapitel 7])

$$\mathbf{grad}\, F(\mathbf{x}) = \begin{bmatrix} \frac{\partial F}{\partial x_i} \\ \vdots \\ \frac{\partial F}{\partial x_n} \end{bmatrix} \in \mathbb{R}^n \qquad (10.1.2.2)$$

provides the direction of local steepest ascent/descent of F

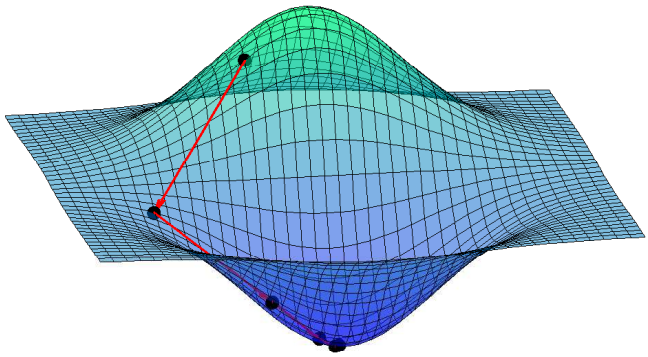


Fig. 384

Of course this very algorithm can encounter plenty of difficulties:

- iteration may get stuck in a *local minimum*,
- iteration may diverge or lead out of D,
- line search may not be feasible.

10.1.3 Gradient method for s.p.d. linear system of equations

However, for the quadratic minimization problem (10.1.1.2) § 10.1.2.1 will converge:

("Geometric intuition", see Fig. 382: quadratic functional J with s.p.d. $\mathbf{A}$ has unique global minimum, $\mathbf{grad}\, J \neq 0$ away from minimum, pointing towards it.)

Adaptation: steepest descent algorithm § 10.1.2.1 for quadratic minimization problem (10.1.1.2), see [QSS00, Sect. 7.2.4]:

$$F(\mathbf{x}) := J(\mathbf{x}) = \tfrac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{b}^\top \mathbf{x} \quad \Rightarrow \quad \mathbf{grad}\, J(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}\,. \qquad (10.1.3.1)$$

This follows from $\mathbf{A} = \mathbf{A}^\top$, the componentwise expression

$$J(\mathbf{x}) = \tfrac{1}{2}\sum_{i,j=1}^{n} a_{ij}x_i x_j - \sum_{i=1}^{n} b_i x_i$$

and the definition (10.1.2.2) of the gradient.

➢ For the descent direction in § 10.1.2.1 applied to the minimization of J from (10.1.1.2) holds

$$\mathbf{d}_k = \mathbf{b} - \mathbf{A}\mathbf{x}^{(k)} =: \mathbf{r}_k \quad \text{the residual (→ Def. 2.4.0.1) for } \mathbf{x}^{(k-1)}.$$

§ 10.1.2.1 for $F = J$ from (10.1.1.2): function to be minimized in line search step:

$$\varphi(t) := J(\mathbf{x}^{(k)} + t\mathbf{d}_k) = J(\mathbf{x}^{(k)}) + t\mathbf{d}_k^\top(\mathbf{A}\mathbf{x}^{(k)} - \mathbf{b}) + \tfrac{1}{2}t^2\mathbf{d}_k^\top\mathbf{A}\mathbf{d}_k \quad \rightarrowtail \text{ a parabola !}\,.$$

$$\frac{d\varphi}{dt}(t^*) = 0 \quad \Leftrightarrow \quad \boxed{t^* = \frac{\mathbf{d}_k^\top \mathbf{d}_k}{\mathbf{d}_k^\top \mathbf{A}\mathbf{d}_k}} \quad \text{(unique minimizer)}\,. \qquad (10.1.3.2)$$

Note: $\mathbf{d}_k = 0 \quad \Leftrightarrow \quad \mathbf{A}\mathbf{x}^{(k)} = \mathbf{b}$ (solution found !)

Note: $\mathbf{A}$ s.p.d. ($\to$ Def. 1.1.2.6) $\Rightarrow$ $\mathbf{d}_k^\top \mathbf{A}\mathbf{d}_k > 0$, if $\mathbf{d}_k \neq 0$

$\blacktriangleright$ $\varphi(t)$ is a parabola that is bounded from below (upward opening)

Based on (10.1.3.1) and (10.1.3.2) we obtain the following steepest descent method for the minimization problem (10.1.1.2):

Steepest descent iteration = gradient method for LSE $\mathbf{A}\mathbf{x} = \mathbf{b}$, $\mathbf{A} \in \mathbb{R}^{n,n}$ s.p.d., $\mathbf{b} \in \mathbb{R}^n$:

§10.1.3.3 (Gradient method for s.p.d. LSE)

Initial guess $\mathbf{x}^{(0)} \in \mathbb{R}^n$, $k = 0$
$\mathbf{r}_0 := \mathbf{b} - \mathbf{A}\mathbf{x}^{(0)}$
repeat
$t^* := \dfrac{\mathbf{r}_k^\top \mathbf{r}_k}{\mathbf{r}_k^\top \mathbf{A}\mathbf{r}_k}$
$\mathbf{x}^{(k+1)} := \mathbf{x}^{(k)} + t^*\mathbf{r}_k$
$\mathbf{r}_{k+1} := \mathbf{r}_k - t^*\mathbf{A}\mathbf{r}_k$
$k := k + 1$
until $\left(\left\| \mathbf{x}^{(k)} - \mathbf{x}^{(k-1)} \right\| \leq \tau_{\text{rel}} \left\| \mathbf{x}^{(k)} \right\|$ or $\left\| \mathbf{x}^{(k)} - \mathbf{x}^{(k-1)} \right\| \leq \tau_{\text{abs}} \right)$

MATLAB-code 10
$\mathbf{A}\mathbf{x} = \mathbf{b}$, $\mathbf{A}$ s.p.d.

⌟

Recursion for residuals, see **??** of Code 10.1.3.4:

$$\mathbf{r}_{k+1} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(k+1)} = \mathbf{b} - \mathbf{A}(\mathbf{x}^{(k)} + t^*\mathbf{r}_k) = \mathbf{r}_k - t^*\mathbf{A}\mathbf{r}_k \,. \tag{10.1.3.5}$$

One step of gradient method involves
- *A single* matrix×vector product with $\mathbf{A}$,
- 2 AXPY-operations ($\to$ Section 1.3.2) on vectors of length n,
- 2 dot products in $\mathbb{R}^n$.

Computational cost (per step) = cost(matrix×vector) + $O(n)$

➢ If $\mathbf{A} \in \mathbb{R}^{n,n}$ is a sparse matrix ($\to$ **??**) with "$O(n)$ nonzero entries", and the data structures allow to perform the matrix×vector product with a computational effort $O(n)$, then a single step of the gradient method costs $O(n)$ elementary operations.

➢ Gradient method of § 10.1.3.3 only needs $\mathbf{A}$×vector in procedural form `y = evalA(x)`.

10.1.4 Convergence of the gradient method

EXAMPLE 10.1.4.1 (Gradient method in 2D) S.p.d. matrices $\in \mathbb{R}^{2,2}$:

$$\mathbf{A}_1 = \begin{bmatrix} 1.9412 & -0.2353 \\ -0.2353 & 1.0588 \end{bmatrix} \,, \quad \mathbf{A}_2 = \begin{bmatrix} 7.5353 & -1.8588 \\ -1.8588 & 0.5647 \end{bmatrix}$$

Eigenvalues: $\sigma(\mathbf{A}_1) = \{1, 2\}$, $\sigma(\mathbf{A}_2) = \{0.1, 8\}$

✎ notation: spectrum of a matrix $\in \mathbb{K}^{n,n}$ $\sigma(\mathbf{M}) := \{\lambda \in \mathbb{C}: \lambda \text{ is eigenvalue of } \mathbf{M}\}$

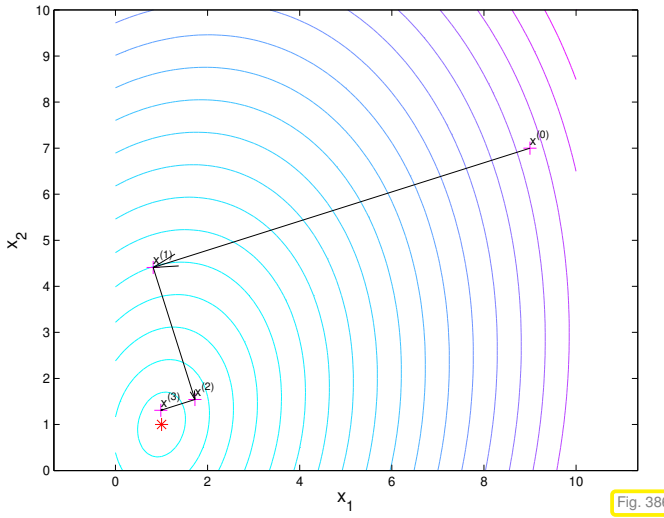


Fig. 385

iterates of § 10.1.3.3 for $\mathbf{A}_1$

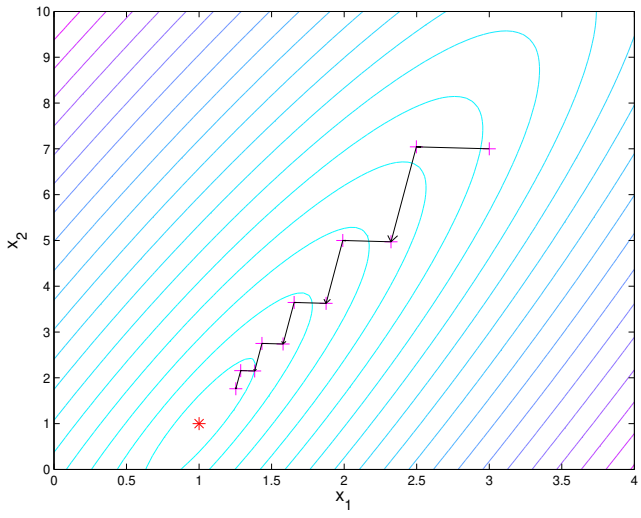


Fig. 386

iterates of § 10.1.3.3 for $\mathbf{A}_2$

Recall theorem on principal axis transformation: every real *symmetric* matrix can be diagonalized by *orthogonal* similarity transformations, see Cor. 9.1.0.9, [NS02, Thm. 7.8], [Gut09, Satz 9.15],

$$\mathbf{A} = \mathbf{A}^\top \in \mathbb{R}^{n,n} \quad \Rightarrow \quad \exists \mathbf{Q} \in \mathbb{R}^{n,n} \text{ orthogonal: } \quad \mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}^\top \quad , \quad \mathbf{D} = \operatorname{diag}(d_1, \ldots, d_n) \in \mathbb{R}^{n,n} \text{ diagonal .} \tag{10.1.4.2}$$

$$J(\mathbf{Q}\widehat{\mathbf{y}}) = \tfrac{1}{2}\widehat{\mathbf{y}}^\top \mathbf{D}\widehat{\mathbf{y}} - \underbrace{(\mathbf{Q}^\top \mathbf{b})^\top}_{=:\widehat{\mathbf{b}}^\top}\widehat{\mathbf{y}} = \tfrac{1}{2}\sum_{i=1}^{n} d_i \widehat{y}_i^2 - \widehat{b}_i \widehat{y}_i \, .$$

Hence, a rigid transformation (rotation, reflection) maps the level surfaces of J from (10.1.1.2) to ellipses with principal axes d_i. As $\mathbf{A}$ s.p.d. $d_i > 0$ is guaranteed.

Observations:

- Larger spread of spectrum leads to more elongated ellipses as level lines ➢ slower convergence of gradient method, see Fig. 386.
- Orthogonality of successive residuals $\mathbf{r}_k$, $\mathbf{r}_{k+1}$.

Clear from definition of § 10.1.3.3:

$$\mathbf{r}_k^\top \mathbf{r}_{k+1} = \mathbf{r}_k^\top \mathbf{r}_k - \mathbf{r}_k^\top \frac{\mathbf{r}_k^\top \mathbf{r}_k}{\mathbf{r}_k^\top \mathbf{A}\mathbf{r}_k} \mathbf{A}\mathbf{r}_k = 0 \, . \tag{10.1.4.3}$$

⌟

EXAMPLE 10.1.4.4 (Convergence of gradient method)

Convergence of gradient method for diagonal matrices, $\mathbf{x}^* = [1, \ldots, 1]^\top$, $\mathbf{x}^{(0)} = 0$:

```
d = 1:0.01:2;   A1 = diag(d);
d = 1:0.1:11;   A2 = diag(d);
d = 1:1:101;    A3 = diag(d);
```

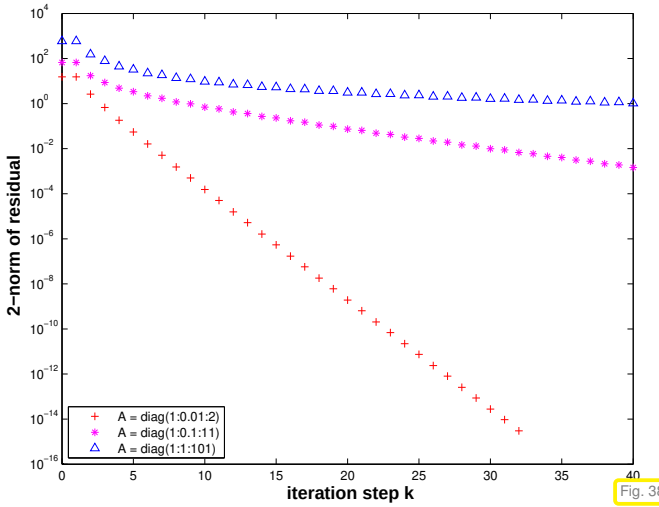


Fig. 387

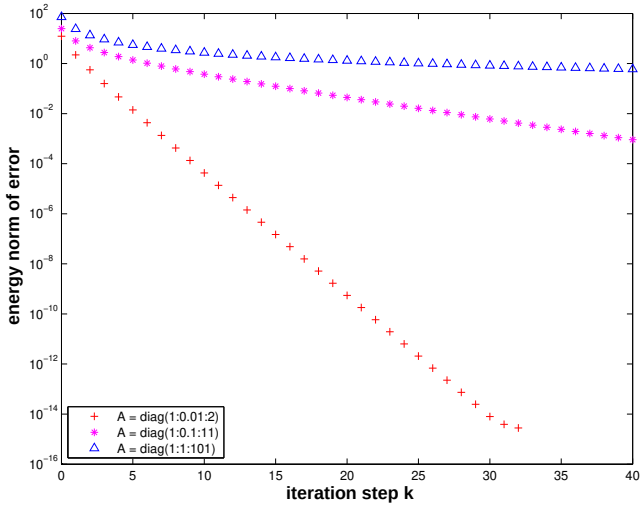


Fig. 388

Note: To study convergence it is *sufficient to consider diagonal matrices*, because

1. (10.1.4.2): for every $\mathbf{A} \in \mathbb{R}^{n,n}$ with $\mathbf{A}^\top = \mathbf{A}$ there is an orthogonal matrix $\mathbf{Q} \in \mathbb{R}^{n.n}$ such that $\mathbf{A} = \mathbf{Q}^\top \mathbf{D}\mathbf{Q}$ with a diagonal matrix $\mathbf{D}$ (principal axis transformation), $\rightarrow$ Cor. 9.1.0.9, [NS02, Thm. 7.8], [Gut09, Satz 9.15],
2. when applying the gradient method § 10.1.3.3 to both $\mathbf{A}\mathbf{x} = \mathbf{b}$ and $\mathbf{D}\widetilde{\mathbf{x}} = \widetilde{\mathbf{b}} := \mathbf{Q}\mathbf{b}$, then the iterates $\mathbf{x}^{(k)}$ and $\widetilde{\mathbf{x}}^{(k)}$ are related by $\mathbf{Q}\mathbf{x}^{(k)} = \widetilde{\mathbf{x}}^{(k)}$.

With $\widetilde{\mathbf{r}}_k := \mathbf{Q}\mathbf{r}_k$, $\widetilde{\mathbf{x}}^{(k)} := \mathbf{Q}\mathbf{x}^{(k)}$, using $\mathbf{Q}^\top\mathbf{Q} = \mathbf{I}$:

Initial guess $\mathbf{x}^{(0)} \in \mathbb{R}^n$, $k = 0$
$\mathbf{r}_0 := \mathbf{b} - \mathbf{Q}^\top\mathbf{D}\mathbf{Q}\mathbf{x}^{(0)}$
repeat
$$t^* := \frac{\mathbf{r}_k^\top \mathbf{Q}^\top \mathbf{Q}\mathbf{r}_k}{\mathbf{r}_k^\top \mathbf{Q}^\top \mathbf{D}\mathbf{Q}\mathbf{r}_k}$$
$\mathbf{x}^{(k+1)} := \mathbf{x}^{(k)} + t^*\mathbf{r}_k$
$\mathbf{r}_{k+1} := \mathbf{r}_k - t^*\mathbf{Q}^\top\mathbf{D}\mathbf{Q}\mathbf{r}_k$
$k := k+1$
until $\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right\| \leq \tau \left\|\mathbf{x}^{(k)}\right\|$

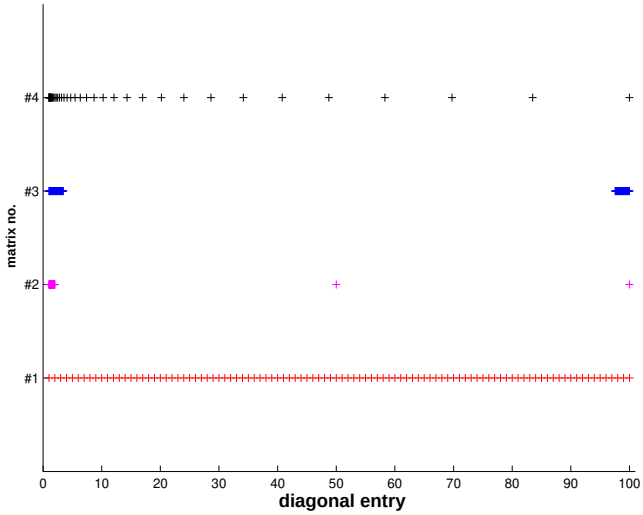
Initial guess $\widetilde{\mathbf{x}}^{(0)} \in \mathbb{R}^n$, $k = 0$
$\widetilde{\mathbf{r}}_0 := \widetilde{\mathbf{b}} - \mathbf{D}\widetilde{\mathbf{x}}^{(0)}$
repeat
$$t^* := \frac{\widetilde{\mathbf{r}}_k^\top \widetilde{\mathbf{r}}_k}{\widetilde{\mathbf{r}}_k^\top \mathbf{D}\widetilde{\mathbf{r}}_k}$$
$\widetilde{\mathbf{x}}^{(k+1)} := \widetilde{\mathbf{x}}^{(k)} + t^*\widetilde{\mathbf{r}}_k$
$\widetilde{\mathbf{r}}_{k+1} := \widetilde{\mathbf{r}}_k - t^*\mathbf{D}\widetilde{\mathbf{r}}_k$
$k := k+1$
until $\left\|\widetilde{\mathbf{x}}^{(k)} - \widetilde{\mathbf{x}}^{(k-1)}\right\| \leq \tau \left\|\widetilde{\mathbf{x}}^{(k)}\right\|$

Observation:

- ✦ linear convergence ($\rightarrow$ Def. 8.2.2.1), see also Rem. 8.2.2.6
- ✦ rate of convergence increases ($\leftrightarrow$ speed of convergence decreases) with spread of spectrum of $\mathbf{A}$

Impact of distribution of diagonal entries ($\leftrightarrow$ eigenvalues) of (diagonal matrix) $\mathbf{A}$
($\mathbf{b} = \mathbf{x}^* = 0$, `x0 = cos((1:n)');`)

Test matrix #1: `A=diag(d); d = (1:100);`
Test matrix #2: `A=diag(d); d = [1+(0:97)/97 , 50 , 100];`
Test matrix #3: `A=diag(d); d = [1+(0:49)*0.05, 100-(0:49)*0.05];`
Test matrix #4: eigenvalues exponentially dense at 1



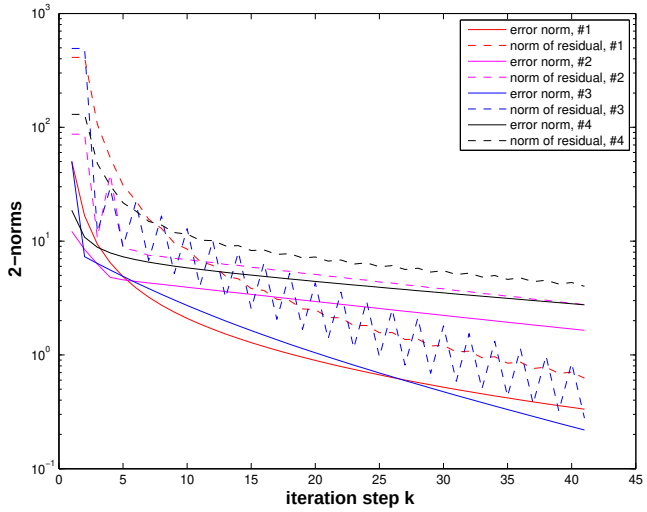


Fig. 389

Observation: Matrices #1, #2 & #4 ➢ little impact of distribution of eigenvalues on *asymptotic* convergence (exception: matrix #2)

⌟

Theory [Hac91, Sect. 9.2.2], [QSS00, Sect.7.2.4]:

Theorem 10.1.4.5. Convergence of gradient method/steepest descent

The iterates of the gradient method of § 10.1.3.3 satisfy

$$\left\|\mathbf{x}^{(k+1)}-\mathbf{x}^*\right\|_A \leq L\left\|\mathbf{x}^{(k)}-\mathbf{x}^*\right\|_A, \quad L := \frac{\operatorname{cond}_2(\mathbf{A})-1}{\operatorname{cond}_2(\mathbf{A})+1},$$

that is, the iteration converges at least linearly (→ Def. 8.2.2.1) w.r.t. energy norm (→ Def. 10.1.0.1).

✎ notation: $\operatorname{cond}_2(\mathbf{A})$ ≙ condition number (→ Def. 2.2.2.7) of $\mathbf{A}$ induced by 2-norm

Remark 10.1.4.6 (2-norm from eigenvalues → [Gut09, Sect. 10.6], [NS02, Sect. 7.4])

$$\mathbf{A}=\mathbf{A}^\top \quad \Rightarrow \quad \begin{aligned} \|\mathbf{A}\|_2 &= \max(|\sigma(\mathbf{A})|)\,, \\ \left\|\mathbf{A}^{-1}\right\|_2 &= \min(|\sigma(\mathbf{A})|)^{-1} \quad \text{, if } \mathbf{A} \text{ regular.} \end{aligned} \tag{10.1.4.7}$$

$$\mathbf{A}=\mathbf{A}^\top \quad \Rightarrow \quad \operatorname{cond}_2(\mathbf{A}) = \frac{\lambda_{\max}(\mathbf{A})}{\lambda_{\min}(\mathbf{A})}\,, \quad \text{where} \quad \begin{aligned} \lambda_{\max}(\mathbf{A}) &:= \max(|\sigma(\mathbf{A})|)\,, \\ \lambda_{\min}(\mathbf{A}) &:= \min(|\sigma(\mathbf{A})|)\,. \end{aligned} \tag{10.1.4.8}$$

✎ other notation $\kappa(\mathbf{A}) := \dfrac{\lambda_{\max}(\mathbf{A})}{\lambda_{\min}(\mathbf{A})}$ ≙ spectral condition number of $\mathbf{A}$

(for general $\mathbf{A}$: $\lambda_{\max}(\mathbf{A})/\lambda_{\min}(\mathbf{A})$ largest/smallest eigenvalue *in modulus*)

These results are an immediate consequence of the fact that

$$\forall \mathbf{A}\in\mathbb{R}^{n,n},\ \mathbf{A}^\top=\mathbf{A}\ \ \exists \mathbf{U}\in\mathbb{R}^{n,n},\ \mathbf{U}^{-1}=\mathbf{U}^\top:\ \ \mathbf{U}^\top\mathbf{A}\mathbf{U} \text{ is diagonal,}$$

see (10.1.4.2), Cor. 9.1.0.9, [NS02, Thm. 7.8], [Gut09, Satz 9.15].

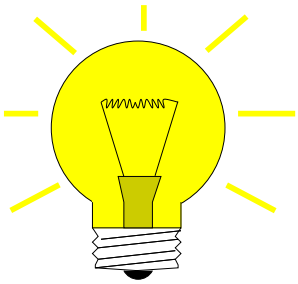
Please note that for general regular $\mathbf{M}\in\mathbb{R}^{n,n}$ we *cannot* expect $\operatorname{cond}_2(\mathbf{M})=\kappa(\mathbf{M})$. ⌟

10.2 Conjugate gradient method (CG) [Han02, Ch. 9], [DR08, Sect. 13.4], [QSS00, Sect. 4.3.4]

Again we consider a linear system of equations $\mathbf{Ax} = \mathbf{b}$ with s.p.d. ($\rightarrow$ Def. 1.1.2.6) system matrix $\mathbf{A} \in \mathbb{R}^{n,n}$ and given $\mathbf{b} \in \mathbb{R}^n$.

Liability of gradient method of Section 10.1.3:	NO MEMORY

1D line search in § 10.1.3.3 is oblivious of former line searches, which rules out reuse of information gained in previous steps of the iteration. This is a typical drawback of 1-point iterative methods.



Idea: Replace linear search with subspace correction

Given:
- initial guess $\mathbf{x}^{(0)}$
- nested subspaces $U_1 \subset U_2 \subset U_3 \subset \cdots \subset U_n = \mathbb{R}^n$, $\dim U_k = k$

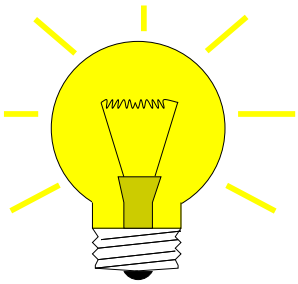
$$\mathbf{x}^{(k)} := \operatorname*{argmin}_{\mathbf{x} \in U_k + \mathbf{x}^{(0)}} J(x)\,, \qquad (10.2.0.1)$$

quadratic functional from (10.1.1.2)

Note: Once the subspaces U_k and $\mathbf{x}^{(0)}$ are fixed, the iteration (10.2.0.1) is well defined, because $J_{|U_k+\mathbf{x}^{(0)}}$ always possess a unique minimizer.

Obvious (from Lemma 10.1.1.1): $\mathbf{x}^{(n)} = \mathbf{x}^* = \mathbf{A}^{-1}\mathbf{b}$

Thanks to (10.1.1.3), definition (10.2.0.1) ensures: $\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\right\|_A \leq \left\|\mathbf{x}^{(k)} - \mathbf{x}^*\right\|_A$



How to find suitable subspaces U_k ?

Idea: $U_{k+1} \leftarrow U_k +$ " local steepest descent direction"

given by $-\mathbf{grad}\, J(\mathbf{x}^{(k)}) = \mathbf{b} - \mathbf{Ax}^{(k)} = \mathbf{r}_k$ (residual $\rightarrow$ Def. 2.4.0.1)

$$U_{k+1} = \operatorname{Span}\{U_k, \mathbf{r}_k\}\,, \quad \mathbf{x}^{(k)} \text{ from (10.2.0.1).} \qquad (10.2.0.2)$$

Obvious: $\mathbf{r}_k = 0 \;\Rightarrow\; \mathbf{x}^{(k)} = \mathbf{x}^* := \mathbf{A}^{-1}\mathbf{b}$ done ✔

Lemma 10.2.0.3. $\mathbf{r}_k \perp U_k$

With $\mathbf{x}^{(k)}$ according to (10.2.0.1), *U_k from* (10.2.0.2) *the residual $\mathbf{r}_k := \mathbf{b} - \mathbf{Ax}^{(k)}$ satisfies*

$$\mathbf{r}_k^\top \mathbf{u} = 0 \quad \forall \mathbf{u} \in U_k \quad (\text{"}\mathbf{r}_k \perp U_k\text{"}).$$

Geometric consideration: since $\mathbf{x}^{(k)}$ is the minimizer of J over the affine space $U_k + \mathbf{x}^{(0)}$, the projection of the steepest descent direction $\mathbf{grad}\, J(\mathbf{x}^{(k)})$ onto U_k has to vanish:

$$\mathbf{x}^{(k)} := \operatorname*{argmin}_{\mathbf{x}\in U_k+\mathbf{x}^{(0)}} J(x) \quad \Rightarrow \quad \mathbf{grad}\, J(\mathbf{x}^{(k)}) \perp U_k \,. \tag{10.2.0.4}$$

Proof. Consider

$$\psi(t) = J(\mathbf{x}^{(k)} + t\mathbf{u}) \,, \quad \mathbf{u} \in U_k, \quad t \in \mathbb{R} \,.$$

By (10.2.0.1), $t \mapsto \psi(t)$ has a global minimum in $t = 0$, which implies

$$\frac{d\psi}{dt}(0) = \mathbf{grad}\, J(\mathbf{x}^{(k)})^\top \mathbf{u} = (\mathbf{A}\mathbf{x}^{(k)} - \mathbf{b})^\top \mathbf{u} = 0 \,.$$

Since $\mathbf{u} \in U_k$ was arbitrary, the lemma is proved. □

Corollary 10.2.0.5.

If $\mathbf{r}_l \neq 0$ for $l = 0, \ldots, k$, $k \leq n$, then $\{\mathbf{r}_0, \ldots, \mathbf{r}_k\}$ is an orthogonal basis of U_k.

Lemma 10.2.0.3 also implies that, if $U_0 = \{0\}$, then $\dim U_k = k$ as long as $\mathbf{x}^{(k)} \neq \mathbf{x}^*$, that is, before we have converged to the exact solution.

(10.2.0.1) and (10.2.0.2) define the conjugate gradient method (CG) for the iterative solution of $\mathbf{Ax} = \mathbf{b}$
(hailed as a "top ten algorithm" of the 20th century, SIAM News, 33(4))

10.2.1 Krylov spaces

Definition 10.2.1.1. Krylov space

For $\mathbf{A} \in \mathbb{R}^{n,n}$, $\mathbf{z} \in \mathbb{R}^n$, $\mathbf{z} \neq 0$, the l-th Krylov space is defined as

$$\mathcal{K}_l(\mathbf{A}, \mathbf{z}) := \operatorname{Span}\{\mathbf{z}, \mathbf{A}\mathbf{z}, \ldots, \mathbf{A}^{l-1}\mathbf{z}\} \,.$$

Equivalent definition: $\mathcal{K}_l(\mathbf{A}, \mathbf{z}) = \{p(\mathbf{A})\mathbf{z}\colon p \text{ polynomial of degree } \leq l\}$

Lemma 10.2.1.2.

The subspaces $U_k \subset \mathbb{R}^n$, $k \geq 1$, defined by (10.2.0.1) *and* (10.2.0.2) *satisfy*

$$U_k = \operatorname{Span}\{\mathbf{r}_0, \mathbf{A}\mathbf{r}_0, \ldots, \mathbf{A}^{k-1}\mathbf{r}_0\} = \mathcal{K}_k(\mathbf{A}, \mathbf{r}_0) \,,$$

where $\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}^{(0)}$ is the initial residual.

Proof. (by induction) Obviously $\mathbf{A}\mathcal{K}_k(\mathbf{A}, \mathbf{r}_0) \subset \mathcal{K}_{k+1}(\mathbf{A}, \mathbf{r}_0)$. In addition

$$\mathbf{r}_k = \mathbf{b} - \mathbf{A}(\mathbf{x}^{(0)} + \mathbf{z}) \quad \text{for some } \mathbf{z} \in U_k \quad \Rightarrow \quad \mathbf{r}_k = \underbrace{\mathbf{r}_0}_{\in \mathcal{K}_{k+1}(\mathbf{A},\mathbf{r}_0)} - \underbrace{\mathbf{A}\mathbf{z}}_{\in \mathcal{K}_{k+1}(\mathbf{A},\mathbf{r}_0)} \,.$$

Since $U_{k+1} = \operatorname{Span}\{U_k, \mathbf{r}_k\}$, we obtain $U_{k+1} \subset \mathcal{K}_{k+1}(\mathbf{A}, \mathbf{r}_0)$. Dimensional considerations based on Lemma 10.2.0.3 finish the proof. □

10.2.2 Implementation of CG

Assume: basis $\{\mathbf{p}_1,\ldots,\mathbf{p}_l\}$, $l=1,\ldots,n$, of $\mathcal{K}_l(\mathbf{A},\mathbf{r})$ available

$$\mathbf{x}^{(l)} \in \mathbf{x}^{(0)} + \mathcal{K}_l(\mathbf{A},\mathbf{r}_0) \quad \succ \quad \text{set} \quad \mathbf{x}^{(l)} = \mathbf{x}^{(0)} + \gamma_1\mathbf{p}_1 + \cdots + \gamma_l\mathbf{p}_l \,.$$

For $\psi(\gamma_1,\ldots,\gamma_l) := J(\mathbf{x}^{(0)} + \gamma_1\mathbf{p}_1 + \cdots + \gamma_l\mathbf{p}_l)$ holds

$$\text{(10.2.0.1)} \quad \Leftrightarrow \quad \frac{\partial\psi}{\partial\gamma_j} = 0\,, \quad j=1,\ldots,l\,.$$

This leads to a linear system of equations by which the coefficients γ_j can be computed:

$$\begin{bmatrix} \mathbf{p}_1^\top\mathbf{A}\mathbf{p}_1 & \cdots & \mathbf{p}_1^\top\mathbf{A}\mathbf{p}_l \\ \vdots & & \vdots \\ \mathbf{p}_l^\top\mathbf{A}\mathbf{p}_1 & \cdots & \mathbf{p}_l^\top\mathbf{A}\mathbf{p}_l \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_l \end{bmatrix} = \begin{bmatrix} \mathbf{p}_1^\top\mathbf{r} \\ \vdots \\ \mathbf{p}_l^\top\mathbf{r} \end{bmatrix}, \quad \mathbf{r} := \mathbf{b} - \mathbf{A}\mathbf{x}^{(0)}\,. \tag{10.2.2.1}$$

Great simplification, if $\{\mathbf{p}_1,\ldots,\mathbf{p}_l\}$ $\mathbf{A}$-orthogonal basis of $\mathcal{K}_l(\mathbf{A},\mathbf{r})$: $\mathbf{p}_j^\top\mathbf{A}\mathbf{p}_i = 0$ for $i \neq j$.

Recall: s.p.d. $\mathbf{A}$ induces an inner product $\succ$ concept of orthogonality [NS02, Sect. 4.4], [Gut09, Sect. 6.2]. "$\mathbf{A}$-geometry" like standard Euclidean space.

Assume: $\mathbf{A}$-orthogonal basis $\{\mathbf{p}_1,\ldots,\mathbf{p}_n\}$ of $\mathbb{R}^n$ available, such that

$$\boxed{\operatorname{Span}\{\mathbf{p}_1,\ldots,\mathbf{p}_l\} = \mathcal{K}_l(\mathbf{A},\mathbf{r})}\,.$$

(Efficient) successive computation of $\mathbf{x}^{(l)}$ becomes possible, see [DR08, Lemma 13.24]
(LSE (10.2.2.1) becomes diagonal !)

Input: : initial guess $\mathbf{x}^{(0)} \in \mathbb{R}^n$
Given: : $\mathbf{A}$-orthogonal bases $\{\mathbf{p}_1,\ldots,\mathbf{p}_l\}$ of $\mathcal{K}_l(\mathbf{A},\mathbf{r}_0)$, $l=1,\ldots,n$
Output: : approximate solution $\mathbf{x}^{(l)} \in \mathbb{R}^n$ of $\mathbf{A}\mathbf{x} = \mathbf{b}$

$\mathbf{r}_0 := \mathbf{b} - \mathbf{A}\mathbf{x}^{(0)}$;

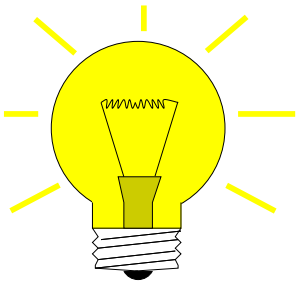
for $j=1$ **to** l **do** { $\mathbf{x}^{(j)} := \mathbf{x}^{(j-1)} + \dfrac{\mathbf{p}_j^\top\mathbf{r}_0}{\mathbf{p}_j^\top\mathbf{A}\mathbf{p}_j}\,\mathbf{p}_j$ } (10.2.2.2)

Task: Efficient computation of $\mathbf{A}$-orthogonal vectors $\{\mathbf{p}_1,\ldots,\mathbf{p}_l\}$ spanning $\mathcal{K}_l(\mathbf{A},\mathbf{r}_0)$ *during the CG iteration*.
$\mathbf{A}$-orthogonalities/orthogonalities ➤ short recursions

Lemma 10.2.0.3 implies orthogonality $\mathbf{p}_j \perp \mathbf{r}_m := \mathbf{b} - \mathbf{A}\mathbf{x}^{(m)}$, $1 \le j \le m \le l$. Also by $\mathbf{A}$-orthogonality of the $\mathbf{p}_k$

$$\mathbf{p}_j^T(\mathbf{b} - \mathbf{A}\mathbf{x}^{(m)}) = \mathbf{p}_j^\top\left(\underbrace{\mathbf{b} - \mathbf{A}\mathbf{x}^{(0)}}_{=\mathbf{r}_0} - \sum_{k=1}^{m}\frac{\mathbf{p}_k^\top\mathbf{r}_0}{\mathbf{p}_k^\top\mathbf{A}\mathbf{p}_k}\mathbf{A}\mathbf{p}_k\right) = 0\,. \tag{10.2.2.3}$$

From linear algebra we already know a way to construct orthogonal basis vectors:



(10.2.2.3) ⇒ Idea:

Gram-Schmidt orthogonalization [NS02, Thm. 4.8], [Gut09, Alg. 6
of residuals $\mathbf{r}_j := \mathbf{b} - \mathbf{A}\mathbf{x}^{(j)}$ w.r.t. $\mathbf{A}$-inner product:

$$\mathbf{p}_1 := \mathbf{r}_0 \,,\; \mathbf{p}_{j+1} := \underbrace{(\mathbf{b} - \mathbf{A}\mathbf{x}^{(j)})}_{=:\mathbf{r}_j} - \underbrace{\sum_{k=1}^{j} \frac{\mathbf{p}_k^\top \mathbf{A}\mathbf{r}_j}{\mathbf{p}_k^\top \mathbf{A}\mathbf{p}_k}\,\mathbf{p}_k}_{(*)} \,,\quad j = 1,\ldots,l-1 \quad . \tag{10.2.2.4}$$

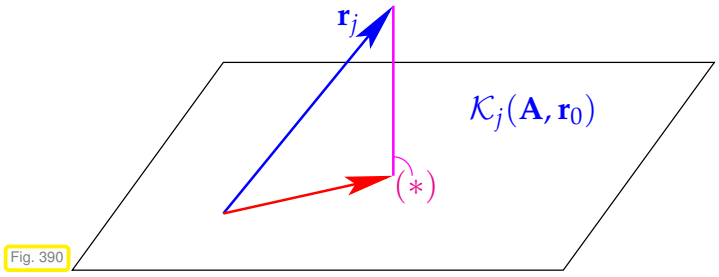


Fig. 390

Geometric interpretation of (10.2.2.4):

$(*)$ ≙ orthogonal projection of $\mathbf{r}_j$ on the subspace $\mathrm{Span}\{\mathbf{p}_1,\ldots,\mathbf{p}_j\}$

Lemma 10.2.2.5. Bases for Krylov spaces in CG

If they do not vanish, the vectors $\mathbf{p}_j$, $1 \le j \le l$, and $\mathbf{r}_j := \mathbf{b} - \mathbf{A}\mathbf{x}^{(j)}$, $0 \le j \le l$, from (10.2.2.2), (10.2.2.4) *satisfy*

(i) $\{\mathbf{p}_1,\ldots,\mathbf{p}_j\}$ is $\mathbf{A}$-orthogonal basis von $\mathcal{K}_j(\mathbf{A},\mathbf{r}_0)$,

(ii) $\{\mathbf{r}_0,\ldots,\mathbf{r}_{j-1}\}$ is orthogonal basis of $\mathcal{K}_j(\mathbf{A},\mathbf{r}_0)$, cf. Cor. 10.2.0.5

Proof. $\mathbf{A}$-orthogonality of $\mathbf{p}_j$ by construction, study (10.2.2.4).

$$\text{(10.2.2.2) \& (10.2.2.4)} \;\Rightarrow\; \mathbf{p}_{j+1} = \mathbf{r}_0 - \sum_{k=1}^{j} \frac{\mathbf{p}_k^\top \mathbf{r}_0}{\mathbf{p}_k^\top \mathbf{A}\mathbf{p}_k}\mathbf{A}\mathbf{p}_k - \sum_{k=1}^{j} \frac{\mathbf{p}_k^\top \mathbf{A}\mathbf{r}_j}{\mathbf{p}_k^\top \mathbf{A}\mathbf{p}_k}\mathbf{p}_k \,.$$
$$\Rightarrow\; \mathbf{p}_{j+1} \in \mathrm{Span}\{\mathbf{r}_0, \mathbf{p}_1,\ldots,\mathbf{p}_j, \mathbf{A}\mathbf{p}_1,\ldots,\mathbf{A}\mathbf{p}_j\} \,.$$

A simple induction argument confirms (i)

$$\text{(10.2.2.4)} \;\Rightarrow\; \mathbf{r}_j \in \mathrm{Span}\{\mathbf{p}_1,\ldots,\mathbf{p}_{j+1}\} \;\&\; \mathbf{p}_j \in \mathrm{Span}\{\mathbf{r}_0,\ldots,\mathbf{r}_{j-1}\} \,. \tag{10.2.2.6}$$

$$\blacktriangleright\quad \boxed{\mathrm{Span}\{\mathbf{p}_1,\ldots,\mathbf{p}_j\} = \mathrm{Span}\{\mathbf{r}_0,\ldots,\mathbf{r}_{j-1}\} = \mathcal{K}_l(\mathbf{A},\mathbf{r}_0)} \,. \tag{10.2.2.7}$$

$$\text{(10.2.2.3)} \;\Rightarrow\; \mathbf{r}_j \perp \mathrm{Span}\{\mathbf{p}_1,\ldots,\mathbf{p}_j\} = \mathrm{Span}\{\mathbf{r}_0,\ldots,\mathbf{r}_{j-1}\} \,. \tag{10.2.2.8}$$

□

Orthogonalities from Lemma 10.2.2.5 ➤ short recursions for $\mathbf{p}_k$, $\mathbf{r}_k$, $\mathbf{x}^{(k)}$!

$$\text{(10.2.2.3)} \;\Rightarrow\; \text{(10.2.2.4) collapses to}\quad \mathbf{p}_{j+1} := \mathbf{r}_j - \frac{\mathbf{p}_j^\top \mathbf{A}\mathbf{r}_j}{\mathbf{p}_j^\top \mathbf{A}\mathbf{p}_j}\mathbf{p}_j \,,\quad j = 1,\ldots,l \,.$$

recursion for residuals:

$$\text{(10.2.2.2)} \quad\blacktriangleright\quad \mathbf{r}_j = \mathbf{r}_{j-1} - \frac{\mathbf{p}_j^\top \mathbf{r}_0}{\mathbf{p}_j^\top \mathbf{A}\mathbf{p}_j}\mathbf{A}\mathbf{p}_j \,.$$

Lemma 10.2.2.5, (i) ➤ $$\mathbf{r}_{j-1}^H \mathbf{p}_j = \left(\mathbf{r}_0 + \sum_{k=1}^{m-1} \frac{\mathbf{r}_0^\top \mathbf{p}_k}{\mathbf{p}_k^T \mathbf{A}\mathbf{p}_k} \mathbf{A}\mathbf{p}_k \right)^T \mathbf{p}_j = \mathbf{r}_0^\top \mathbf{p}_j \,. \tag{10.2.2.9}$$

The orthogonality (10.2.2.9) together with (10.2.2.8) permits us to replace $\mathbf{r}_0$ with $\mathbf{r}_{j-1}$ in the actual implementation.

§10.2.2.10 (CG method for solving $\mathbf{A}\mathbf{x} = \mathbf{b}$, $\mathbf{A}$ s.p.d. → [DR08, Alg. 13.27])

Input : initial guess $\mathbf{x}^{(0)} \in \mathbb{R}^n$
Output : approximate solution $\mathbf{x}^{(l)} \in \mathbb{R}^n$

$\mathbf{p}_1 = \mathbf{r}_0 := \mathbf{b} - \mathbf{A}\mathbf{x}^{(0)}$;
for $j = 1$ **to** l **do** {

$$\mathbf{x}^{(j)} := \mathbf{x}^{(j-1)} + \frac{\mathbf{p}_j^T \mathbf{r}_{j-1}}{\mathbf{p}_j^T \mathbf{A}\mathbf{p}_j} \mathbf{p}_j;$$

$$\mathbf{r}_j = \mathbf{r}_{j-1} - \frac{\mathbf{p}_j^T \mathbf{r}_{j-1}}{\mathbf{p}_j^T \mathbf{A}\mathbf{p}_j} \mathbf{A}\mathbf{p}_j;$$

$$\mathbf{p}_{j+1} = \mathbf{r}_j - \frac{(\mathbf{A}\mathbf{p}_j)^T \mathbf{r}_j}{\mathbf{p}_j^T \mathbf{A}\mathbf{p}_j} \mathbf{p}_j;$$

}

Input: initial guess $\mathbf{x} \triangleq \mathbf{x}^{(0)} \in \mathbb{R}^n$
tolerance $\tau > 0$
Output: approximate solution $\mathbf{x} \triangleq \mathbf{x}^{(l)}$

$\mathbf{p} := \mathbf{r}_0 := \mathbf{r} := \mathbf{b} - \mathbf{A}\mathbf{x}$;
for $j = 1$ **to** $l_{\max}$ **do** {
$\beta := \mathbf{r}^T \mathbf{r}$;
$\mathbf{h} := \mathbf{A}\mathbf{p}$;
$\alpha := \frac{\beta}{\mathbf{p}^T \mathbf{h}}$;
$\mathbf{x} := \mathbf{x} + \alpha \mathbf{p}$;
$\mathbf{r} := \mathbf{r} - \alpha \mathbf{h}$;
if $\|\mathbf{r}\| \leq \tau \|\mathbf{r}_0\|$ **then stop**;
$\beta := \frac{\mathbf{r}^T \mathbf{r}}{\beta}$;
$\mathbf{p} := \mathbf{r} + \beta \mathbf{p}$;
}

In CG algorithm: $\mathbf{r}_j = \mathbf{b} - \mathbf{A}\mathbf{x}^{(k)}$ agrees with the residual associated with the current iterate (in exact arithmetic, *cf.* Ex. 10.2.3.1), but computation through short recursion is more efficient.

➢ We find that the CG method possesses all the algorithmic advantages of the gradient method, *cf.* the discussion in Section 10.1.3.

1 matrix×vector product, 3 dot products, 3 AXPY-operations per step:
If $\mathbf{A}$ sparse, $\mathrm{nnz}(\mathbf{A}) \sim n$ ➤ computational effort $O(n)$ per step

MATLAB-code 10.2.2.11: basic CG iteration for solving $\mathbf{A}\mathbf{x} = \mathbf{b}$, § 10.2.2.10

MATLAB-function:

```
x=pcg(A,b,tol,maxit,[],[],x0)      : Solve Ax = b with at most maxit CG steps: stop,
                                     when ||r_l|| : ||r_0|| < tol.
x=pcg(Afun,b,tol,maxit,[],[],x0): Afun = handle to function for computing A×vector.
[x,flag,relr,it,resv] = pcg(...) : diagnostic information about iteration
```

Remark 10.2.2.12 (A posteriori termination criterion for plain CG)

For any vector norm and associated matrix norm (→ Def. 1.5.5.10) hold (with residual $\mathbf{r}_l := \mathbf{b} - \mathbf{A}\mathbf{x}^{(l)}$)

$$\frac{1}{\operatorname{cond}(\mathbf{A})}\frac{\|\mathbf{r}_l\|}{\|\mathbf{r}_0\|} \leq \frac{\|\mathbf{x}^{(l)}-\mathbf{x}^*\|}{\|\mathbf{x}^{(0)}-\mathbf{x}^*\|} \leq \operatorname{cond}(\mathbf{A})\,\frac{\|\mathbf{r}_l\|}{\|\mathbf{r}_0\|}\,. \tag{10.2.2.13}$$

relative decrease of iteration error

(10.2.2.13) can easily be deduced from the error equation $\mathbf{A}(\mathbf{x}^{(k)} - \mathbf{x}^*) = \mathbf{r}_k$, see Def. 2.4.0.1 and (2.4.0.13). ⌟

10.2.3 Convergence of CG

Note: CG is a *direct solver*, because (in exact arithmetic) $\mathbf{x}^{(k)} = \mathbf{x}^*$ for some $k \leq n$

EXAMPLE 10.2.3.1 (Impact of roundoff errors on CG → [QSS00, Rem. 4.3])

Numerical experiment: `A=hilb(20)`, $\mathbf{x}^{(0)} = 0$, $\mathbf{b} = [1,\ldots,1]^\top$

Hilbert-Matrix: extremely ill-conditioned

residual norms during CG iteration $\mathbf{R} = \left[\mathbf{r}_0,\ldots,\mathbf{r}^{(10)}\right]$

▷:

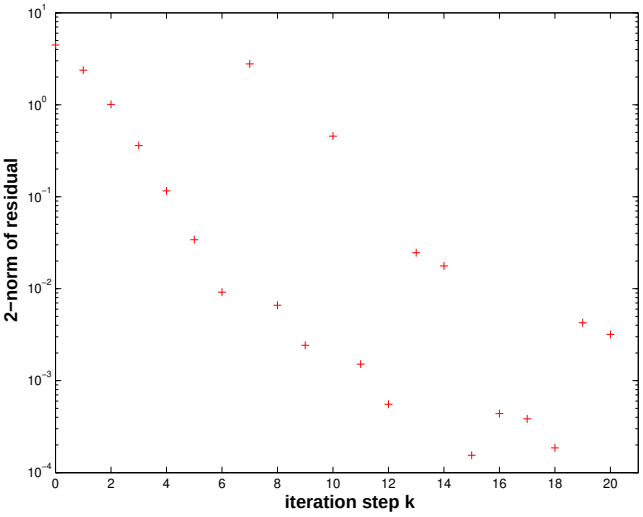


Fig. 391

$\mathbf{R}^\top\mathbf{R} =$

$$\begin{bmatrix}
1.000000 & -0.000000 & 0.000000 & -0.000000 & 0.000000 & -0.000000 & 0.016019 & -0.795816 & -0.430569 & 0.348133 \\
-0.000000 & 1.000000 & -0.000000 & 0.000000 & -0.000000 & 0.000000 & -0.012075 & 0.600068 & -0.520610 & 0.420903 \\
0.000000 & -0.000000 & 1.000000 & -0.000000 & 0.000000 & -0.000000 & 0.001582 & -0.078664 & 0.384453 & -0.310577 \\
-0.000000 & 0.000000 & -0.000000 & 1.000000 & -0.000000 & 0.000000 & -0.000024 & 0.001218 & -0.024115 & 0.019394 \\
0.000000 & -0.000000 & 0.000000 & -0.000000 & 1.000000 & -0.000000 & 0.000000 & -0.000002 & 0.000151 & -0.000118 \\
-0.000000 & 0.000000 & -0.000000 & 0.000000 & -0.000000 & 1.000000 & -0.000000 & 0.000000 & -0.000000 & 0.000000 \\
0.016019 & -0.012075 & 0.001582 & -0.000024 & 0.000000 & -0.000000 & 1.000000 & -0.000000 & -0.000000 & 0.000000 \\
-0.795816 & 0.600068 & -0.078664 & 0.001218 & -0.000002 & 0.000000 & -0.000000 & 1.000000 & 0.000000 & -0.000000 \\
-0.430569 & -0.520610 & 0.384453 & -0.024115 & 0.000151 & -0.000000 & -0.000000 & 0.000000 & 1.000000 & 0.000000 \\
0.348133 & 0.420903 & -0.310577 & 0.019394 & -0.000118 & 0.000000 & 0.000000 & -0.000000 & 0.000000 & 1.000000
\end{bmatrix}$$

➢ Roundoff
- ✦ destroys orthogonality of residuals
- ✦ prevents computation of exact solution after n steps.

Numerical instability (→ Def. 1.5.5.19) ➢ pointless to (try to) use CG as direct solver! ⌟

Practice: CG used for large n as *iterative solver*: $\mathbf{x}^{(k)}$ for some $k \ll n$ is expected to provide good approximation for $\mathbf{x}^*$

EXAMPLE 10.2.3.2 (Convergence of CG as iterative solver) CG (Code 10.2.2.11) & gradient method (Code 10.1.3.4) for LSE with sparse s.p.d. "Poisson matrix"

```
A = gallery('poisson',m); x0 = (1:n)'; b = zeros(n,1);
```

➢ $\mathbf{A} \in \mathbb{R}^{m^2,m^2}$

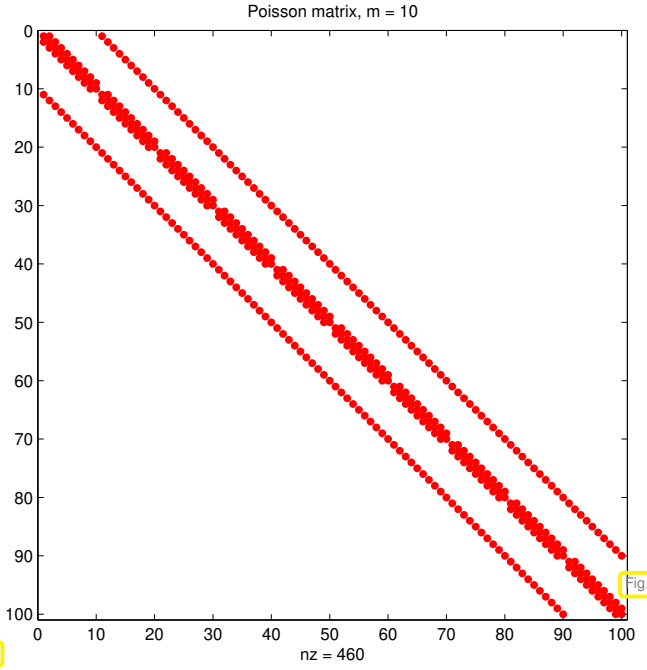


Fig. 392

poissonspy

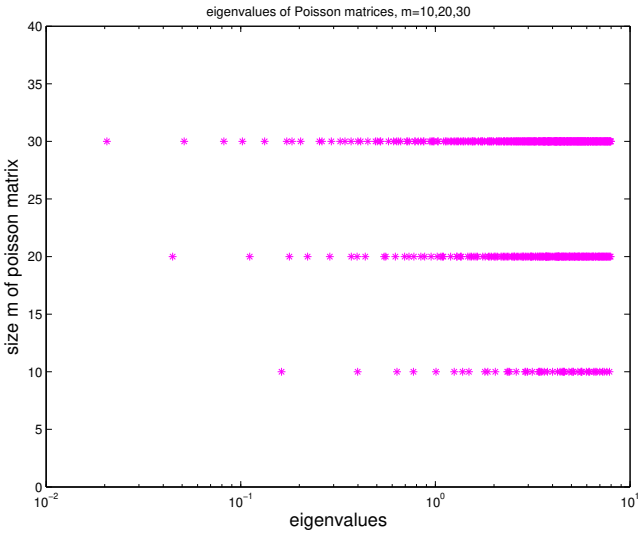


Fig. 393

poissoneig

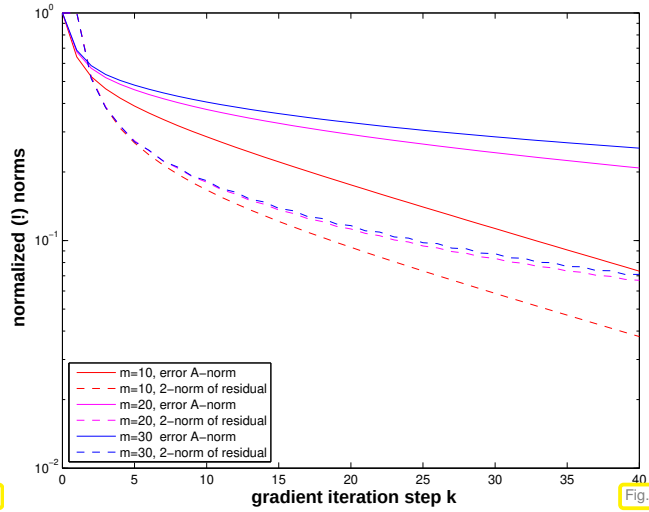


Fig. 394

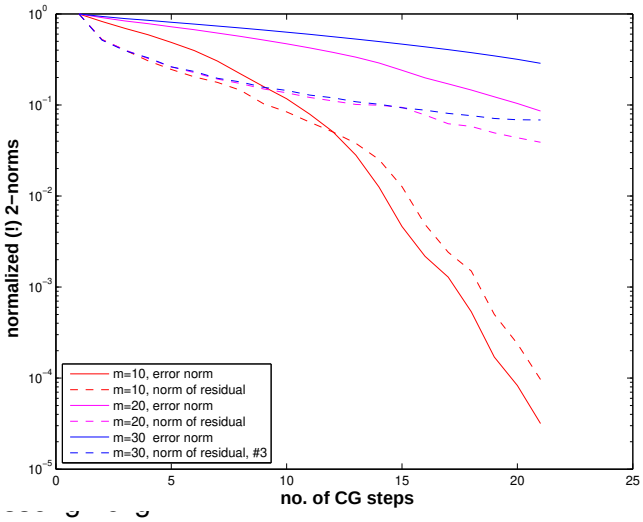


Fig. 395

cgcvgpoiss

Observations:

- CG much faster than gradient method (as expected, because it has "memory")
- Both, CG and gradient method converge more slowly for larger sizes of Poisson matrices.

Convergence theory: [Hac94, Sect. 9.4.3]

A simple consequence of (10.1.1.3) and (10.2.0.1):

Corollary 10.2.3.3. "Optimality" of CG iterates

Writing $\mathbf{x}^* \in \mathbb{R}^n$ for the exact solution of $\mathbf{A}\mathbf{x} = \mathbf{b}$ the CG iterates satisfy

$$\left\|\mathbf{x}^* - \mathbf{x}^{(l)}\right\|_A = \min\{\|\mathbf{y} - \mathbf{x}^*\|_A : \mathbf{y} \in \mathbf{x}^{(0)} + \mathcal{K}_l(\mathbf{A}, \mathbf{r}_0)\} \quad , \quad \mathbf{r}_0 := \mathbf{b} - \mathbf{A}\mathbf{x}^{(0)} .$$

This paves the way for a quantitative convergence estimate:

$$\mathbf{y} \in \mathbf{x}^{(0)} + \mathcal{K}_l(\mathbf{A}, \mathbf{r}) \quad \Leftrightarrow \quad \mathbf{y} = \mathbf{x}^{(0)} + \mathbf{A}\, p(\mathbf{A})(\mathbf{x} - \mathbf{x}^{(0)}) \ , \quad p = \text{polynomial of degree } \le l - 1 .$$

$$\blacktriangleright \quad \mathbf{x} - \mathbf{y} = q(\mathbf{A})(\mathbf{x} - \mathbf{x}^{(0)}), \quad q = \text{polynomial of degree } \le l \ , \quad q(0) = 1 .$$

$$\left\|\mathbf{x} - \mathbf{x}^{(l)}\right\|_A \le \boxed{\min\{\max_{\lambda \in \sigma(\mathbf{A})} |q(\lambda)| : q \text{ polynomial of degree } \le l \ , \quad q(0) = 1\}} \cdot \left\|\mathbf{x} - \mathbf{x}^{(0)}\right\|_A . \tag{10.2.3.4}$$

Bound this minimum for $\lambda \in [\lambda_{\min}(\mathbf{A}), \lambda_{\max}(\mathbf{A})]$ by using suitable "polynomial candidates"

Tool: Chebychev polynomials (→ Section 6.2.3.1) ➢ lead to the following estimate [Hac91, Satz 9.4.2], [DR08, Satz 13.29]

Theorem 10.2.3.5. Convergence of CG method

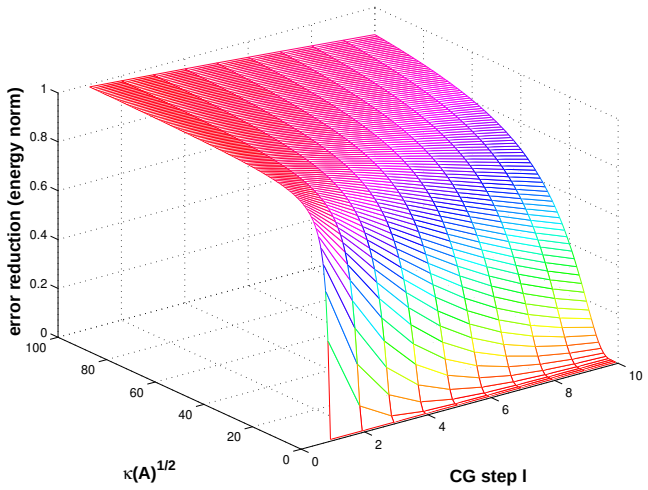
The iterates of the CG method for solving $\mathbf{A}\mathbf{x} = \mathbf{b}$ (see Code 10.2.2.11) with $\mathbf{A} = \mathbf{A}^\top$ s.p.d. satisfy

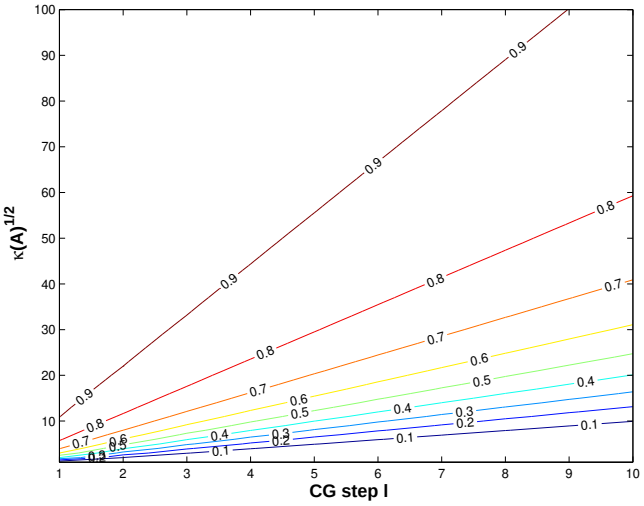
$$\begin{aligned}\left\|\mathbf{x} - \mathbf{x}^{(l)}\right\|_A &\le \frac{2\left(1 - \frac{1}{\sqrt{\kappa(\mathbf{A})}}\right)^l}{\left(1 + \frac{1}{\sqrt{\kappa(\mathbf{A})}}\right)^{2l} + \left(1 - \frac{1}{\sqrt{\kappa(\mathbf{A})}}\right)^{2l}} \left\|\mathbf{x} - \mathbf{x}^{(0)}\right\|_A \\ &\le 2\left(\frac{\sqrt{\kappa(\mathbf{A})} - 1}{\sqrt{\kappa(\mathbf{A})} + 1}\right)^l \left\|\mathbf{x} - \mathbf{x}^{(0)}\right\|_A .\end{aligned}$$

(recall: $\kappa(\mathbf{A})$ = spectral condition number of $\mathbf{A}$, $\kappa(\mathbf{A}) = \mathrm{cond}_2(\mathbf{A})$)

The estimate of this theorem confirms *asymptotic linear convergence* of the CG method (→ Def. 8.2.2.1) with a rate of $\dfrac{\sqrt{\kappa(\mathbf{A})} - 1}{\sqrt{\kappa(\mathbf{A})} + 1}$

Plots of bounds for error reduction (in energy norm) during CG iteration from Thm. 10.2.3.5:





MATLAB-code 10.2.3.6: plotting theoretical bounds for CG convergence rate

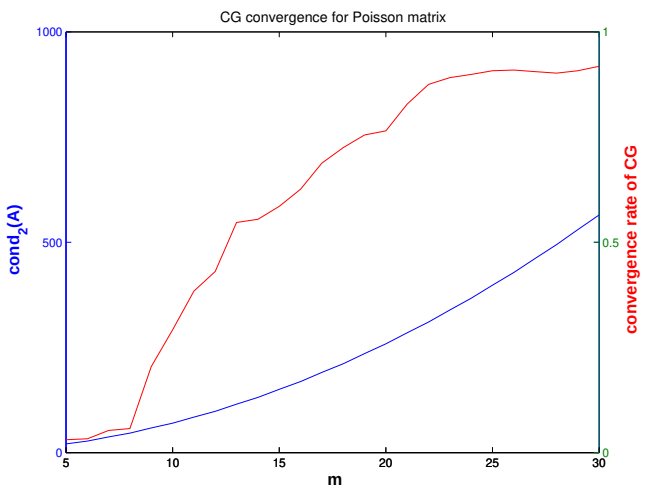
EXAMPLE 10.2.3.7 (Convergence rates for CG method)

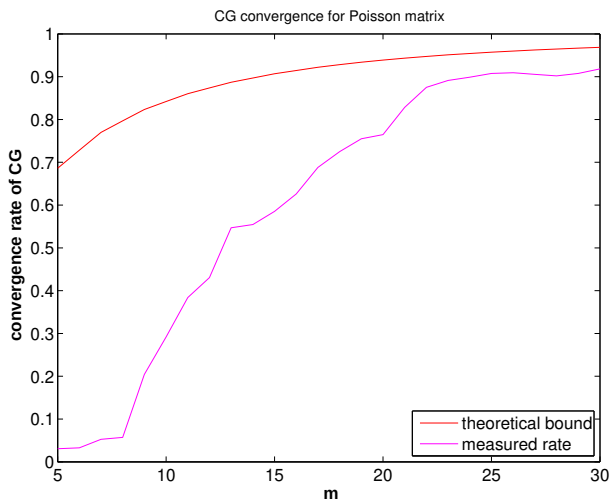
MATLAB-code 10.2.3.8: CG for Poisson matrix

```
1 A = gallery('poisson',m); n = size(A,1);
2 x0 = (1:n)'; b = ones(n,1); maxit = 30; tol =0;
3 [x,flag,relres,iter,resvec] = pcg(A,b,tol,maxit,[],[],x0);
```

Measurement of rate of (linear) convergence:

$$\text{rate} \approx \sqrt[10]{\frac{\|\mathbf{r}_{30}\|_2}{\|\mathbf{r}_{20}\|_2}}\,. \qquad (10.2.3.9)$$





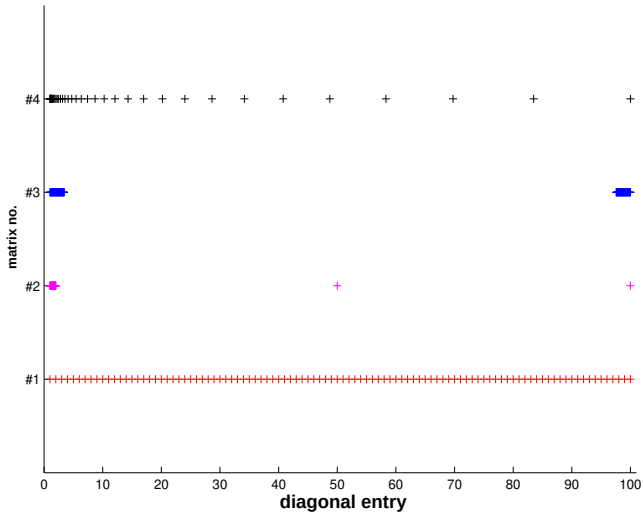
Justification for estimating the rate of linear convergence (→ Def. 8.2.2.1) of $\|\mathbf{r}_k\|_2 \to 0$:

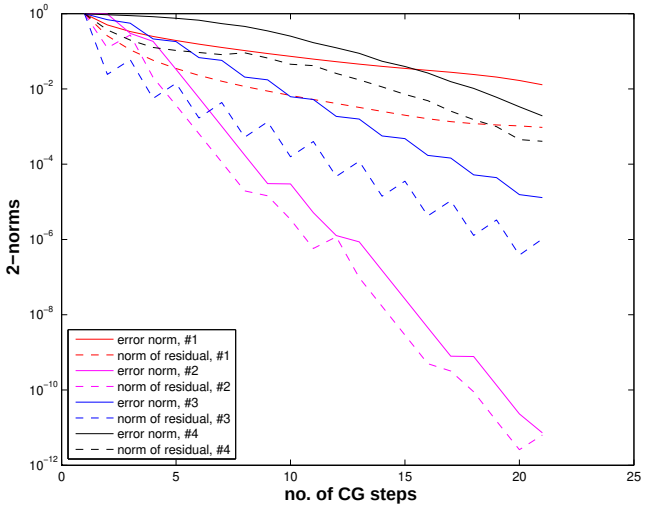
$$\|\mathbf{r}_{k+1}\|_2 \approx L\|\mathbf{r}_k\|_2 \quad \Rightarrow \quad \|\mathbf{r}_{k+m}\|_2 \approx L^m\|\mathbf{r}_k\|_2\,.$$

EXAMPLE 10.2.3.10 (CG convergence and spectrum → Ex. 10.1.4.4)

Test matrix #1: `A=diag(d); d = (1:100);`
Test matrix #2: `A=diag(d); d = [1+(0:97)/97 , 50 , 100];`
Test matrix #3: `A=diag(d); d = [1+(0:49)*0.05, 100-(0:49)*0.05];`
Test matrix #4: eigenvalues exponentially dense at 1

```
x0 = cos((1:n)'); b = zeros(n,1);
```





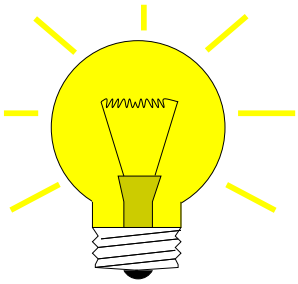
Observations: Distribution of eigenvalues has crucial impact on convergence of CG
(This is clear from the convergence theory, because detailed information about the spectrum allows a much better choice of "candidate polynomial" in (10.2.3.4) than merely using Chebychev polynomials)

➢ Clustering of eigenvalues leads to faster convergence of CG
(in stark contrast to the behavior of the gradient method, see Ex. 10.1.4.4)

CG convergence boosted by clustering of eigenvalues

10.3 Preconditioning [DR08, Sect. 13.5], [Han02, Ch. 10], [QSS00, Sect. 4.3.5]

Thm. 10.2.3.5 ➢ (Potentially) slow convergence of CG in case $\kappa(\mathbf{A}) \gg 1$.



Idea: Preconditioning

Apply CG method to transformed linear system

$$\widetilde{\mathbf{A}}\widetilde{\mathbf{x}} = \widetilde{\mathbf{b}} \quad , \quad \widetilde{\mathbf{A}} := \mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2} \, , \, \widetilde{\mathbf{x}} := \mathbf{B}^{1/2}\mathbf{x} \, , \, \widetilde{\mathbf{b}} := \mathbf{B}^{-1/2}\mathbf{b} \, , \qquad (10.3.0.1)$$

with "small" $\kappa(\widetilde{\mathbf{A}})$, $\mathbf{B} = \mathbf{B}^\top \in \mathbb{R}^{N,N}$ s.p.d. ≙ preconditioner.

Remark 10.3.0.2 (Square root of a s.p.d. matrix)

What is meant by the "square root" $\mathbf{B}^{1/2}$ of a s.p.d. matrix $\mathbf{B}$?

Recall (10.1.4.2) : for every $\mathbf{B} \in \mathbb{R}^{n,n}$ with $\mathbf{B}^\top = \mathbf{B}$ there is an orthogonal matrix $\mathbf{Q} \in \mathbb{R}^{n,n}$ such that $\mathbf{B} = \mathbf{Q}^\top \mathbf{D}\mathbf{Q}$ with a diagonal matrix $\mathbf{D}$ (→ Cor. 9.1.0.9, [NS02, Thm. 7.8], [LS09, Satz 9.15]). If $\mathbf{B}$ is s.p.d.

the (diagonal) entries of $\mathbf{D}$ are strictly positive and we can define

$$\mathbf{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n), \quad \lambda_i > 0 \quad \Rightarrow \quad \mathbf{D}^{1/2} := \operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}) \, .$$

This is generalized to

$$\mathbf{B}^{1/2} := \mathbf{Q}^\top \mathbf{D}^{1/2} \mathbf{Q} \, ,$$

and one easily verifies, using $\mathbf{Q}^\top = \mathbf{Q}^{-1}$, that $(\mathbf{B}^{1/2})^2 = \mathbf{B}$ and that $\mathbf{B}^{1/2}$ is s.p.d. In fact, these two requirements already determine $\mathbf{B}^{1/2}$ uniquely:

$\mathbf{B}^{1/2}$ is the unique s.p.d. matrix such that $(\mathbf{B}^{1/2})^2 = \mathbf{B}$.

⌟

Notion 10.3.0.3. Preconditioner

A s.p.d. matrix $\mathbf{B} \in \mathbb{R}^{n,n}$ is called a preconditioner (*ger.:* Vorkonditionierer) for the s.p.d. matrix $\mathbf{A} \in \mathbb{R}^{n,n}$, if

1. $\kappa(\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2})$ is "small" and
2. the evaluation of $\mathbf{B}^{-1}\mathbf{x}$ is about as expensive (in terms of elementary operations) as the matrix×vector multiplication $\mathbf{A}\mathbf{x}$, $\mathbf{x} \in \mathbb{R}^n$.

Recall: spectral condition number $\kappa(\mathbf{A}) := \dfrac{\lambda_{\max}(\mathbf{A})}{\lambda_{\min}(\mathbf{A})}$, see (10.1.4.8)

There are several equivalent ways to express that $\kappa(\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2})$ is "small":

- $\kappa(\mathbf{B}^{-1}\mathbf{A})$ is "small",

  because spectra agree $\sigma(\mathbf{B}^{-1}\mathbf{A}) = \sigma(\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2})$ due to similarity (→ Lemma 9.1.0.6)
- $\exists 0 < \gamma < \Gamma$, Γ/γ "small": $\gamma\,(\mathbf{x}^\top \mathbf{B}\mathbf{x}) \le \mathbf{x}^\top \mathbf{A}\mathbf{x} \le \Gamma\,(\mathbf{x}^\top \mathbf{B}\mathbf{x}) \quad \forall \mathbf{x} \in \mathbb{R}^n$,

  where equivalence is seen by transforming $\mathbf{y} := \mathbf{B}^{-1/2}\mathbf{x}$ and appealing to the min-max Thm. 9.3.2.18.

"Reader's digest" version of Notion 10.3.0.3:

S.p.d. $\mathbf{B}$ preconditioner $:\Leftrightarrow$ $\mathbf{B}^{-1}$ = cheap approximate inverse of $\mathbf{A}$

Problem: $\mathbf{B}^{1/2}$, which occurs prominently in (10.3.0.1) is usually not available with acceptable computational costs.

However, if one formally applies § 10.2.2.10 to the transformed system

$$\widetilde{\mathbf{A}}\widetilde{\mathbf{x}} := \left(\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}\right)(\mathbf{B}^{1/2}\mathbf{x}) = \widetilde{\mathbf{b}} := \mathbf{B}^{-1/2}\mathbf{b} \, ,$$

from (10.3.0.1), it becomes apparent that, after suitable transformation of the iteration variables $\mathbf{p}_j$ and $\mathbf{r}_j$, $\mathbf{B}^{1/2}$ and $\mathbf{B}^{-1/2}$ invariably occur in products $\mathbf{B}^{-1/2}\mathbf{B}^{-1/2} = \mathbf{B}^{-1}$ and $\mathbf{B}^{1/2}\mathbf{B}^{-1/2} = \mathbf{I}$. Thus, thanks to this intrinsic transformation square roots of $\mathbf{B}$ are not required for the implementation!

CG for $\widetilde{\mathbf{A}}\widetilde{\mathbf{x}} = \widetilde{\mathbf{b}}$

Input : initial guess $\widetilde{\mathbf{x}}^{(0)} \in \mathbb{R}^n$
Output : approximate solution $\widetilde{\mathbf{x}}^{(l)} \in \mathbb{R}^n$

$\widetilde{\mathbf{p}}_1 := \widetilde{\mathbf{r}}_0 := \widetilde{\mathbf{b}} - \mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}\widetilde{\mathbf{x}}^{(0)}$;
for $j = 1$ **to** l **do** {

$$\alpha := \frac{\widetilde{\mathbf{p}}_j^T \widetilde{\mathbf{r}}_{j-1}}{\widetilde{\mathbf{p}}_j^T \mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}\widetilde{\mathbf{p}}_j}$$
$$\widetilde{\mathbf{x}}^{(j)} := \widetilde{\mathbf{x}}^{(j-1)} + \alpha\, \widetilde{\mathbf{p}}_j;$$

$$\widetilde{\mathbf{r}}_j = \widetilde{\mathbf{r}}_{j-1} - \alpha \mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{1/2}\widetilde{\mathbf{p}}_j;$$

$$\widetilde{\mathbf{p}}_{j+1} = \widetilde{\mathbf{r}}_j - \frac{(\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}\widetilde{\mathbf{p}}_j)^T \widetilde{\mathbf{r}}_j}{\widetilde{\mathbf{p}}_j^T \mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}\widetilde{\mathbf{p}}_j}\, \widetilde{\mathbf{p}}_j;$$

}

Equivalent CG with transformed variables

Input : initial guess $\mathbf{x}^{(0)} \in \mathbb{R}^n$
Output : approximate solution $\mathbf{x}^{(l)} \in \mathbb{R}^n$

$\mathbf{B}^{1/2}\widetilde{\mathbf{r}}_0 := \mathbf{B}^{1/2}\widetilde{\mathbf{b}} - \mathbf{A}\mathbf{B}^{-1/2}\widetilde{\mathbf{x}}^{(0)}$;
$\mathbf{B}^{-1/2}\widetilde{\mathbf{p}}_1 := \mathbf{B}^{-1}(\mathbf{B}^{1/2}\widetilde{\mathbf{r}}_0)$;
for $j = 1$ **to** l **do** {

$$\alpha := \frac{(\mathbf{B}^{-1/2}\widetilde{\mathbf{p}}_j)^T \mathbf{B}^{1/2}\widetilde{\mathbf{r}}_{j-1}}{(\mathbf{B}^{-1/2}\widetilde{\mathbf{p}}_j)^T \mathbf{A}\mathbf{B}^{-1/2}\widetilde{\mathbf{p}}_j}$$
$$\mathbf{B}^{-1/2}\widetilde{\mathbf{x}}^{(j)} := \mathbf{B}^{-1/2}\widetilde{\mathbf{x}}^{(j-1)} + \alpha\, \mathbf{B}^{-1/2}\widetilde{\mathbf{p}}_j;$$

$$\mathbf{B}^{1/2}\widetilde{\mathbf{r}}_j = \mathbf{B}^{1/2}\widetilde{\mathbf{r}}_{j-1} - \alpha \mathbf{A}\mathbf{B}^{-1/2}\widetilde{\mathbf{p}}_j;$$

$$\mathbf{B}^{-1/2}\widetilde{\mathbf{p}}_{j+1} = \mathbf{B}^{-1}(\mathbf{B}^{-1/2}\widetilde{\mathbf{r}}_j) - \frac{(\mathbf{B}^{-1/2}\widetilde{\mathbf{p}}_j)^T \mathbf{A}\mathbf{B}^{-1}(\mathbf{B}^{1/2}\widetilde{\mathbf{r}}_j)}{(\mathbf{B}^{-1/2}\widetilde{\mathbf{p}}_j)^T \mathbf{A}\mathbf{B}^{-1/2}\widetilde{\mathbf{p}}_j}\, \mathbf{B}^{-1/2}\widetilde{\mathbf{p}}_j;$$

}

with the transformations:

$$\widetilde{\mathbf{x}}^{(k)} = \mathbf{B}^{1/2}\mathbf{x}^{(k)} \quad , \quad \widetilde{\mathbf{r}}_k = \mathbf{B}^{-1/2}\mathbf{r}_k \quad , \quad \widetilde{\mathbf{p}}_k = \mathbf{B}^{-1/2}\mathbf{r}_k \,. \tag{10.3.0.4}$$

§10.3.0.5 (Preconditioned CG method (PCG) [DR08, Alg. 13.32], [Han02, Alg. 10.1])

Input: initial guess $\mathbf{x} \in \mathbb{R}^n \triangleq \mathbf{x}^{(0)} \in \mathbb{R}^n$, tolerance $\tau > 0$
Output: approximate solution $\mathbf{x} \triangleq \mathbf{x}^{(l)}$

$\mathbf{p} := \mathbf{r} := \mathbf{b} - \mathbf{A}\mathbf{x}$; $\mathbf{p} := \mathbf{B}^{-1}\mathbf{r}$; $\mathbf{q} := \mathbf{p}$; $\tau_0 := \mathbf{p}^\top \mathbf{r}$;
for $l = 1$ **to** $l_{\max}$ **do** {
$\beta := \mathbf{r}^\top \mathbf{q}$; $\mathbf{h} := \mathbf{A}\mathbf{p}$; $\alpha := \frac{\beta}{\mathbf{p}^\top \mathbf{h}}$;
$\mathbf{x} := \mathbf{x} + \alpha\mathbf{p}$;
$\mathbf{r} := \mathbf{r} - \alpha\mathbf{h}$;
$\mathbf{q} := \mathbf{B}^{-1}\mathbf{r}$; $\beta := \frac{\mathbf{r}^\top \mathbf{q}}{\beta}$;
if $|\mathbf{q}^\top \mathbf{r}| \le \tau \cdot \tau_0$ **then stop**;
$\mathbf{p} := \mathbf{q} + \beta\mathbf{p}$;
} (10.3.0.6)

⌟

MATLAB-code 10.3.0.7: simple implementation of PCG algorithm § 10.3.0.5

Computational effort per step: 1 evaluation $\mathbf{A}\times$vector,
1 evaluation $\mathbf{B}^{-1}\times$vector,
3 dot products, 3 AXPY-operations

Remark 10.3.0.8 (Convergence theory for PCG) Assertions of Thm. 10.2.3.5 remain valid with $\kappa(\mathbf{A})$ replaced with $\kappa(\mathbf{B}^{-1}\mathbf{A})$ and energy norm based on $\widetilde{\mathbf{A}}$ instead of $\mathbf{A}$. ⌟

EXAMPLE 10.3.0.9 (Simple preconditioners)

$\mathbf{B}$ = easily invertible "part" of $\mathbf{A}$

- $\mathbf{B} = \mathtt{diag}(\mathbf{A})$: Jacobi preconditioner (diagonal scaling)
- $(\mathbf{B})_{ij} = \begin{cases} (\mathbf{A})_{ij} & \text{, if } |i-j| \le k \text{ ,} \\ 0 & \text{else,} \end{cases}$ for some $k \ll n$.
- Symmetric Gauss-Seidel preconditioner

Idea: Solve $\mathbf{A}\mathbf{x} = \mathbf{b}$ approximately in two stages:

① Approximation $\mathbf{A}^{-1} \approx \mathtt{tril}(\mathbf{A})$ (lower triangular part): $\widetilde{\mathbf{x}} = \mathtt{tril}(\mathbf{A})^{-1}\mathbf{b}$

② Approximation $\mathbf{A}^{-1} \approx \mathtt{triu}(\mathbf{A})$ (upper triangular part) and use this to approximately "solve" the error equation $\mathbf{A}(\mathbf{x}-\widetilde{\mathbf{x}}) = \mathbf{r}$, with residual $\mathbf{r} := \mathbf{b} - \mathbf{A}\widetilde{\mathbf{x}}$:

$$\mathbf{x} = \widetilde{\mathbf{x}} + \mathtt{triu}(\mathbf{A})^{-1}(\mathbf{b} - \mathbf{A}\widetilde{\mathbf{x}}) \,.$$

With $\mathbf{L}_A := \mathtt{tril}(\mathbf{A})$, $\mathbf{U}_A := \mathtt{triu}(\mathbf{A})$ one finds

$$\mathbf{x} = (\mathbf{L}_A^{-1} + \mathbf{U}_A^{-1} - \mathbf{U}_A^{-1}\mathbf{A}\mathbf{L}_A^{-1})\mathbf{b} \quad \blacktriangleright \quad \mathbf{B}^{-1} = \mathbf{L}_A^{-1} + \mathbf{U}_A^{-1} - \mathbf{U}_A^{-1}\mathbf{A}\mathbf{L}_A^{-1} \,. \tag{10.3.0.10}$$

For all these approaches the evaluation of $\mathbf{B}^{-1}\mathbf{r}$ can be done with effort of $O(n)$ in the case of a sparse matrix $\mathbf{A}$ (e.g. with $O(1)$ non-zero entries per row). However, there is absolutely no guarantee that $\kappa(\mathbf{B}^{-1}\mathbf{A})$ will be reasonably small. It will crucially depend on $\mathbf{A}$, if this can be expected. ⌟

More complicated preconditioning strategies:

- Incomplete Cholesky factorization, MATLAB-`ichol`, [DR08, Sect. 13.5]]
- Sparse approximate inverse preconditioner (SPAI)

EXAMPLE 10.3.0.11 (Tridiagonal preconditioning)

Efficacy of preconditioning of sparse LSE with tridiagonal part:

MATLAB-code 10.3.0.12: LSE for Ex. 10.3.0.11

```
A = spdiags(repmat([1/n,-1,2+2/n,-1,1/n],n,1),[-n/2,-1,0,1,n/2],n,n);
b = ones(n,1); x0 = ones(n,1); tol = 1.0E-4; maxit = 1000;
evalA = @(x) A*x;

% no preconditioning, see Code 10.3.0.7
invB = @(x) x; [x,rn] = pcgbase(evalA,b,tol,maxit,invB,x0);

% tridiagonal preconditioning, see Code 10.3.0.7
B = spdiags(spdiags(A,[-1,0,1]),[-1,0,1],n,n);
```

```
10 invB  = @(x) B\x; [x,rnpc] = pcgbase(evalA,b,tol,maxit,invB,x0);
     %
```

The Code 10.3.0.12 highlights the use of a preconditioner in the context of the PCG method; it only takes a function that realizes the application of $\mathbf{B}^{-1}$ to a vector. In Line 10 of the code this function is passed as function handle `invB`.

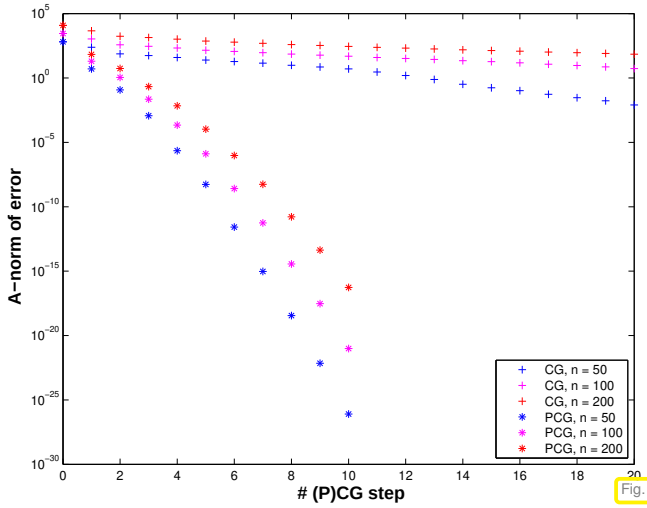


Fig. 396

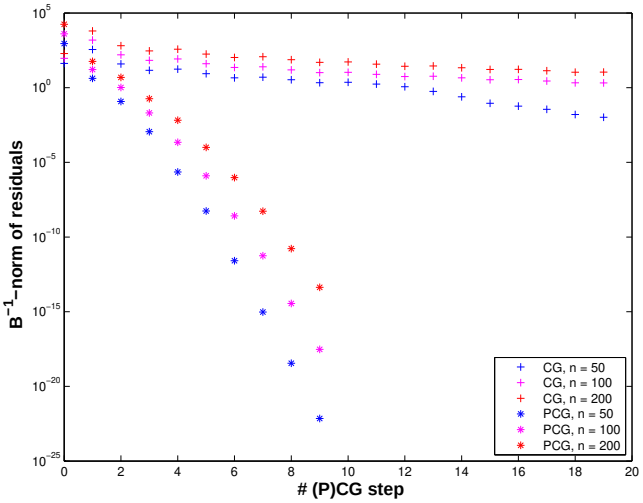


Fig. 397

n	# CG steps	# PCG steps
16	8	3
32	16	3
64	25	4
128	38	4
256	66	4
512	106	4
1024	149	4
2048	211	4
4096	298	3
8192	421	3
16384	595	3
32768	841	3

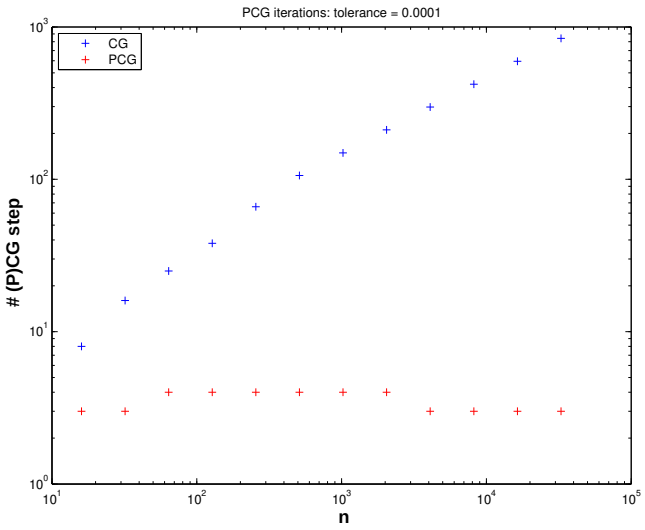


Fig. 398

Clearly in this example the tridiagonal part of the matrix is dominant for large n. In addition, its condition number grows $\sim n^2$ as is revealed by a closer inspection of the spectrum.

Preconditioning with the tridiagonal part manages to suppress this growth of the condition number of $\mathbf{B}^{-1}\mathbf{A}$ and ensures fast convergence of the preconditioned CG method ⌟

Remark 10.3.0.13 (Termination of PCG) Recall Rem. 10.2.2.12, (10.2.2.13):

$$\frac{1}{\operatorname{cond}(\mathbf{A})}\frac{\|\mathbf{r}_l\|}{\|\mathbf{r}_0\|} \leq \frac{\|\mathbf{x}^{(l)}-\mathbf{x}^*\|}{\|\mathbf{x}^{(0)}-\mathbf{x}^*\|} \leq \operatorname{cond}(\mathbf{A})\frac{\|\mathbf{r}_l\|}{\|\mathbf{r}_0\|}\,. \tag{10.2.2.13}$$

$\mathbf{B}$ good preconditioner ➤ $\operatorname{cond}_2(\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2})$ small (→ Notion 10.3.0.3)

Idea: consider (10.2.2.13) for

- Euclidean norm $\|\cdot\| = \|\cdot\|_2$ ↔ cond_2
- transformed quantities $\widetilde{\mathbf{x}}$, $\widetilde{\mathbf{r}}$, see (10.3.0.1), (10.3.0.4)

Monitor 2-norm of transformed residual:

$$\widetilde{\mathbf{r}} = \widetilde{\mathbf{b}} - \widetilde{\mathbf{A}}\widetilde{\mathbf{x}} = \mathbf{B}^{-1/2}\mathbf{r} \quad \Rightarrow \quad \|\widetilde{\mathbf{r}}\|_2^2 = \mathbf{r}^\top \mathbf{B}^{-1}\mathbf{r}\,.$$

(10.2.2.13) ➤ estimate for 2-norm of transformed iteration errors: $\left\|\widetilde{\mathbf{e}}^{(l)}\right\|_2^2 = (\mathbf{e}^{(l)})^\top \mathbf{B}\mathbf{e}^{(l)}$

Analogous to (10.2.2.13), estimates for energy norm (→ Def. 10.1.0.1) of error $\mathbf{e}^{(l)} := \mathbf{x} - \mathbf{x}^{(l)}$, $\mathbf{x}^* := \mathbf{A}^{-1}\mathbf{b}$:

Use error equation $\mathbf{A}\mathbf{e}^{(l)} = \mathbf{r}_l$:

$$\mathbf{r}_l^\top \mathbf{B}^{-1}\mathbf{r}_l = (\mathbf{B}^{-1}\mathbf{A}\mathbf{e}^{(l)})^\top \mathbf{A}\mathbf{e}^{(l)} \le \lambda_{\max}(\mathbf{B}^{-1}\mathbf{A})\left\|\mathbf{e}^{(l)}\right\|_A^2\,,$$

$$\left\|\mathbf{e}^{(l)}\right\|_A^2 = (\mathbf{A}\mathbf{e}^{(l)})^\top \mathbf{e}^{(l)} = \mathbf{r}_l^\top \mathbf{A}^{-1}\mathbf{r}_l = \mathbf{B}^{-1}\mathbf{r}^\top \mathbf{B}\mathbf{A}^{-1}\mathbf{r}_l \le \lambda_{\max}(\mathbf{B}\mathbf{A}^{-1})\,(\mathbf{B}^{-1}\mathbf{r}_l)^\top \mathbf{r}_l\,.$$

available during PCG iteration (10.3.0.6)

$$\frac{1}{\kappa(\mathbf{B}^{-1}\mathbf{A})}\frac{\left\|\mathbf{e}^{(l)}\right\|_A^2}{\left\|\mathbf{e}^{(0)}\right\|_A^2} \le \frac{(\mathbf{B}^{-1}\mathbf{r}_l)^\top \mathbf{r}_l}{(\mathbf{B}^{-1}\mathbf{r}_0)^\top \mathbf{r}_0} \le \kappa(\mathbf{B}^{-1}\mathbf{A})\frac{\left\|\mathbf{e}^{(l)}\right\|_A^2}{\left\|\mathbf{e}^{(0)}\right\|_A^2} \tag{10.3.0.14}$$

$\kappa(\mathbf{B}^{-1}\mathbf{A})$ "small" ➤ $\mathbf{B}^{-1}$-energy norm of residual $\approx$ $\mathbf{A}$-norm of error !
($\mathbf{r}_l \cdot \mathbf{B}^{-1}\mathbf{r}_l = \mathbf{q}^\top \mathbf{r}$ in Algorithm (10.3.0.6))

MATLAB-function: `[x,flag,relr,it,rv] = pcg(A,b,tol,maxit,B,[],x0);`
(`A`, `B` may be handles to functions providing $\mathbf{A}\mathbf{x}$ and $\mathbf{B}^{-1}\mathbf{x}$, resp.)

Remark 10.3.0.15 (Termination criterion in MATLAB-`pcg` → [QSS00, Sect. 4.6])

Implementation (skeleton) of MATLAB built-in `pcg`:

Listing 10.1: MATLAB-code : PCG algorithm

```
function x = pcg(Afun,b,tol,maxit,Binvfun,x0)
x = x0; r = b - feval(Afun,x); rho = 1;
for i = 1 : maxit
  y = feval(Binvfun,r);
  rho1 = rho; rho = r' * y;
  if (i == 1)
    p = y;
  else
    beta = rho / rho1;
    p = y + beta * p;
  end
  q = feval(Afun,p);
  alpha = rho /(p' * q);
  x = x + alpha * p;
  if (norm(b - evalf(Afun,x)) <= tolb*norm(b)), return; end
  r = r - alpha * q;
end
```

Dubious termination criterion !

10.4 Survey of Krylov Subspace Methods

10.4.1 Minimal residual methods

Idea: Replace Euclidean inner product in CG with $\mathbf{A}$-inner product

$$\left\|\mathbf{x}^{(l)}-\mathbf{x}\right\|_A \quad \text{replaced with} \quad \left\|\mathbf{A}(\mathbf{x}^{(l)}-\mathbf{x})\right\|_2 = \|\mathbf{r}_l\|_2$$

MINRES method [Hac91, Sect. 9.5.2] (for *any* symmetric matrix **!**)

Theorem 10.4.1.1.

For $\mathbf{A}=\mathbf{A}^H \in \mathbb{R}^{n,n}$ the residuals $\mathbf{r}_l$ generated in the MINRES iteration satisfy

$$\|\mathbf{r}_l\|_2 = \min\{\|\mathbf{A}\mathbf{y}-\mathbf{b}\|_2 : \mathbf{y} \in \mathbf{x}^{(0)} + \mathcal{K}_l(\mathbf{A},\mathbf{r}_0)\}$$

$$\|\mathbf{r}_l\|_2 \le \frac{2\left(1-\frac{1}{\kappa(\mathbf{A})}\right)^l}{\left(1+\frac{1}{\kappa(\mathbf{A})}\right)^{2l}+\left(1-\frac{1}{\kappa(\mathbf{A})}\right)^{2l}}\|\mathbf{r}_0\|_2 .$$

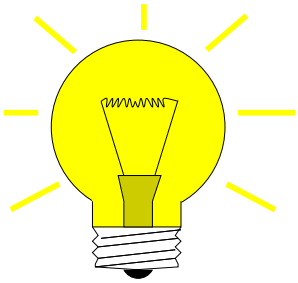
Note: similar formula for (linear) rate of convergence as for CG, see Thm. 10.2.3.5, but with $\sqrt{\kappa(\mathbf{A})}$ replaced with $\kappa(\mathbf{A})$ **!**

Iterative solver for $\mathbf{A}\mathbf{x}=\mathbf{b}$ with *symmetric* system matrix $\mathbf{A}$:

MATLAB-functions:
- `[x,flg,res,it,resv] = minres(A,b,tol,maxit,B,[],x0);`
- `[...] = minres(Afun,b,tol,maxit,Binvfun,[],x0);`

Computational costs : 1 $\mathbf{A}\times$vector, 1 $\mathbf{B}^{-1}\times$vector per step, a few dot products & SAXPYs
Memory requirement: a few vectors $\in \mathbb{R}^n$

Extension to general regular $\mathbf{A} \in \mathbb{R}^{n,n}$:



Idea: Solver overdetermined linear system of equations

$$\mathbf{x}^{(l)} \in \mathbf{x}^{(0)} + \mathcal{K}_l(\mathbf{A},\mathbf{r}_0): \quad \mathbf{A}\mathbf{x}^{(l)} = \mathbf{b}$$

in *least squares sense*, → Chapter 3.

$$\mathbf{x}^{(l)} = \operatorname{argmin}\{\|\mathbf{A}\mathbf{y}-\mathbf{b}\|_2 : \mathbf{y} \in \mathbf{x}^{(0)} + \mathcal{K}_l(\mathbf{A},\mathbf{r}_0)\} .$$

➤ GMRES method for general matrices $\mathbf{A} \in \mathbb{R}^{n,n}$ → [Han02, Ch. 16], [QSS00, Sect. 4.4.2]

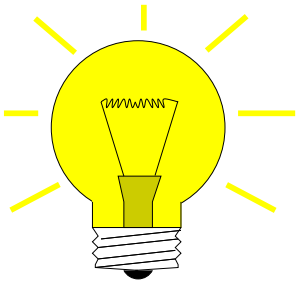
MATLAB-function:
- `[x,flag,relr,it,rv] = gmres(A,b,rs,tol,maxit,B,[],x0);`
- `[...] = gmres(Afun,b,rs,tol,maxit,Binvfun,[],x0);`

Computational costs : 1 $\mathbf{A}\times$vector, 1 $\mathbf{B}^{-1}\times$vector per step,
: $O(l)$ dot products & SAXPYs in l-th step
Memory requirements: $O(l)$ vectors $\in \mathbb{K}^n$ in l-th step

Remark 10.4.1.2 (Restarted GMRES) After many steps of GMRES we face considerable computational costs and memory requirements for every further step. Thus, the iteration may be *restarted* with the current iterate $\mathbf{x}^{(l)}$ as initial guess → `rs`-parameter triggers restart after every `rs` steps (Danger: failure to converge). ⌟

10.4.2 Iterations with short recursions [QSS00, Sect. 4.5]

Iterative methods for *general* regular system matrix $\mathbf{A}$:



Idea: Given $\mathbf{x}^{(0)} \in \mathbb{R}^n$ determine (better) approximation $\mathbf{x}^{(l)}$ through Petrov-Galerkin condition

$$\mathbf{x}^{(l)} \in \mathbf{x}^{(0)} + \mathcal{K}_l(\mathbf{A},\mathbf{r}_0): \quad \mathbf{p}^H(\mathbf{b}-\mathbf{A}\mathbf{x}^{(l)}) = 0 \quad \forall \mathbf{p} \in W_l ,$$

with suitable test space W_l, $\dim W_l = l$, e.g. $W_l := \mathcal{K}_l(\mathbf{A}^H,\mathbf{r}_0)$ (→ bi-conjugate gradients, BiCG)

➤ Zoo of methods with short recursions (i.e. constant effort per step)

MATLAB-function:
- `[x,flag,r,it,rv] = bicgstab(A,b,tol,maxit,B,[],x0)`
- `[...] = bicgstab(Afun,b,tol,maxit,Binvfun,[],x0);`

Computational costs : 2 $\mathbf{A}\times$vector, 2 $\mathbf{B}^{-1}\times$vector, 4 dot products, 6 SAXPYs per step
Memory requirements: 8 vectors $\in \mathbb{R}^n$

MATLAB-function:
- `[x,flag,r,it,rv] = qmr(A,b,tol,maxit,B,[],x0)`
- `[...] = qmr(Afun,b,tol,maxit,Binvfun,[],x0);`

Computational costs : 2 $\mathbf{A}\times$vector, 2 $\mathbf{B}^{-1}\times$vector, 2 dot products, 12 SAXPYs per step
Memory requirements: 10 vectors $\in \mathbb{R}^n$

- little (useful) convergence theory available
- stagnation & "breakdowns" commonly occur

EXAMPLE 10.4.2.1 (Failure of Krylov iterative solvers)

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & \cdots & & & \cdots & 0 \\ 0 & 0 & 1 & 0 & & & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & & & \\ & & & & & & \ddots & \vdots \\ \vdots & & & & & \ddots & \ddots & 0 \\ 0 & & & & & & 0 & 1 \\ 1 & 0 & \cdots & & & & \cdots & 0 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 0 \\ \vdots \\ \\ \\ \vdots \\ 0 \\ 1 \end{bmatrix} \quad \blacktriangleright \quad \mathbf{x} = \mathbf{e}_1 .$$

$$\mathbf{x}^{(0)} = 0 \quad \succ \quad \mathbf{r}_0 = \mathbf{e}_n \quad \succ \quad \mathcal{K}_l(\mathbf{A}, \mathbf{r}_0) = \mathrm{Span}\{\mathbf{e}_n, \mathbf{e}_{n-1}, \ldots, \mathbf{e}_{n-l+1}\}$$

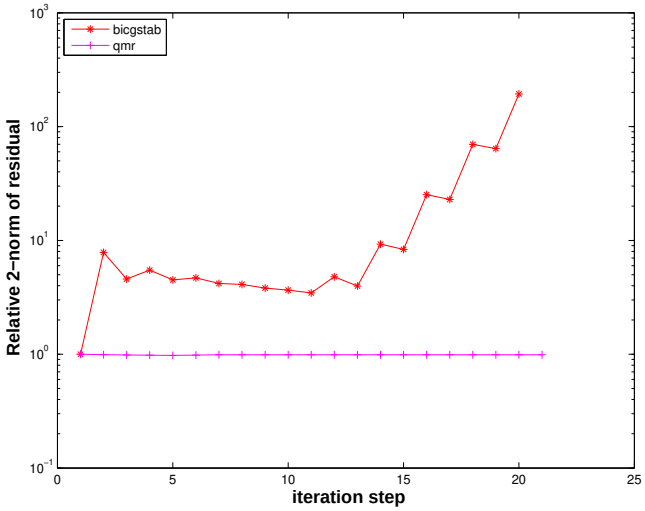
$$\blacktriangleright \quad \min\{\|\mathbf{y} - \mathbf{x}\|_2 : \mathbf{y} \in \mathcal{K}_l(\mathbf{A}, \mathbf{r}_0)\} = \begin{cases} 1 & \text{, if } l \leq n \text{ ,} \\ 0 & \text{, for } l = n \text{ .} \end{cases}$$

TRY & PRAY

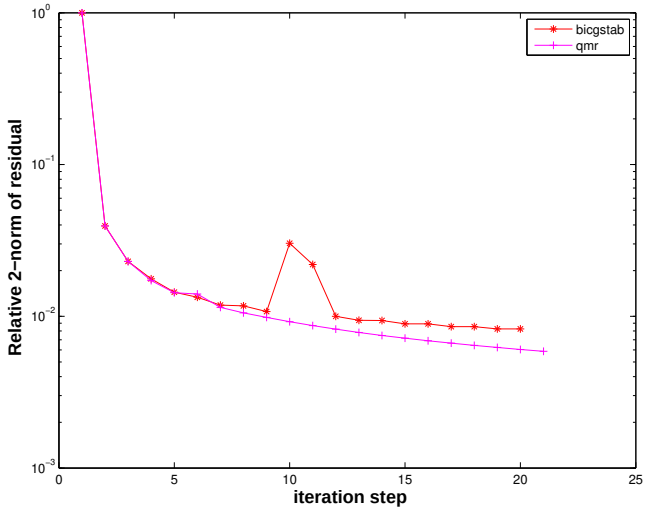
EXAMPLE 10.4.2.2 (Convergence of Krylov subspace methods for non-symmetric system matrix)

```
A = gallery('tridiag',-0.5*ones(n-1,1),2*ones(n,1),-1.5*ones(n-1,1));
B = gallery('tridiag',0.5*ones(n-1,1),2*ones(n,1),1.5*ones(n-1,1));
```

Plotted: $\|\mathbf{r}_l\|_2 : \|\mathbf{r}_0\|_2$:



tridiagonal matrix $\mathbf{A}$



tridiagonal matrix $\mathbf{B}$

Summary:

Advantages of Krylov methods vs. direct elimination (**IF** they converge at all/sufficiently fast).

- They require system matrix $\mathbf{A}$ in procedural form `y=evalA(x)` $\leftrightarrow \mathbf{y} = \mathbf{A}\mathbf{x}$ only.
- They can perfectly exploit sparsity of system matrix.
- They can cash in on low accuracy requirements (**IF** viable termination criterion available).
- They can benefit from a good initial guess.

Bibliography

[AK07] Owe Axelsson and János Karátson. “Mesh independent superlinear PCG rates via compact-equivalent operators”. In: *SIAM J. Numer. Anal.* 45.4 (2007), pp. 1495–1516. DOI: 10.1137/06066391X.

[DR08] W. Dahmen and A. Reusken. *Numerik für Ingenieure und Naturwissenschaftler*. Heidelberg: Springer, 2008 (cit. on pp. 733, 734, 741–755).

[Gut09] M.H. Gutknecht. *Lineare Algebra*. Lecture Notes. SAM, ETH Zürich, 2009 (cit. on pp. 738–740, 743, 744).

[Hac91] W. Hackbusch. *Iterative Lösung großer linearer Gleichungssysteme*. B.G. Teubner–Verlag, Stuttgart, 1991 (cit. on pp. 740, 748, 756).

[Hac94] Wolfgang Hackbusch. *Iterative solution of large sparse systems of equations*. Vol. 95. Applied Mathematical Sciences. New York: Springer-Verlag, 1994, pp. xxii+429 (cit. on pp. 733, 747).

[Han02] M. Hanke-Bourgeois. *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens*. Mathematische Leitfäden. Stuttgart: B.G. Teubner, 2002 (cit. on pp. 734, 741–755, 757).

[IM97] I.C.F. Ipsen and C.D. Meyer. *The idea behind Krylov methods*. Technical Report 97-3. Raleigh, NC: Math. Dep., North Carolina State University, Jan. 1997.

[LS09] A.R. Laliena and F.-J. Sayas. “Theoretical aspects of the application of convolution quadrature to scattering of acoustic waves”. In: *Numer. Math.* 112.4 (2009), pp. 637–678 (cit. on p. 750).

[NS02] K. Nipp and D. Stoffer. *Lineare Algebra*. 5th ed. Zürich: vdf Hochschulverlag, 2002 (cit. on pp. 738–740, 743, 744, 750).

[QSS00] A. Quarteroni, R. Sacco, and F. Saleri. *Numerical mathematics*. Vol. 37. Texts in Applied Mathematics. New York: Springer, 2000 (cit. on pp. 733–755, 757).

[Saa03] Yousef Saad. *Iterative methods for sparse linear systems*. Second. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2003, pp. xviii+528 (cit. on p. 733).

[Str09] M. Struwe. *Analysis für Informatiker*. Lecture notes, ETH Zürich. 2009 (cit. on p. 736).

[Win80] Ragnar Winther. “Some superlinear convergence results for the conjugate gradient method”. In: *SIAM J. Numer. Anal.* 17.1 (1980), pp. 14–17.

Chapter 11

Numerical Integration – Single Step Methods

For historical reasons the approximate solution of initial value problems for ordinary differential equations is called “Numerical Integration”. This chapter will introduce the most important class of numerical methods for that purpose.

Contents

11.1 Initial-Value Problems (IVPs) for Ordinary Differential Equations (ODEs) . . . 762
11.1.1 Ordinary Differential Equations (ODEs) . . . 762
11.1.2 Mathematical Modeling with Ordinary Differential Equations: Examples . 764
11.1.3 Theory of Initial-Value-Problems (IVPs) . . . 769
11.1.4 Evolution Operators . . . 773
11.2 Introduction: Polygonal Approximation Methods . . . 776
11.2.1 Explicit Euler method . . . 777
11.2.2 Implicit Euler method . . . 779
11.2.3 Implicit midpoint method . . . 781
11.3 General Single-Step Methods . . . 783
11.3.1 Definition . . . 783
11.3.2 (Asymptotic) Convergence of Single-Step Methods . . . 787
11.4 Explicit Runge-Kutta Single-Step Methods (RKSSMs) . . . 796
11.5 Adaptive Stepsize Control . . . 803
11.5.1 The Need for Timestep Adaptation . . . 803
11.5.2 Local-in-Time Stepsize Control . . . 805
11.5.3 Embedded Runge-Kutta Methods . . . 812

Supplementary literature. Some grasp of the meaning and theory of ordinary differential equations (ODEs) is indispensable for understanding the construction and properties of numerical methods. Relevant information can be found in [Str09, Sect. 5.6, 5.7, 6.5].

Books dedicated to numerical methods for ordinary differential equations:

- [DB02] excellent textbook, but geared to the needs of students of mathematics.
- [HNW93] and [HW11] : *the* standard reference.
- [HLW06]: wonderful book conveying deep insight, with emphasis on mathematical concepts.

11.1 Initial-Value Problems (IVPs) for Ordinary Differential Equations (ODEs)

Video tutorial for Section 11.1: Initial-Value Problems (IVPs) for Ordinary Differential Equations (ODEs): (35 minutes) Download link, tablet notes

☞ You may also watch the explanations of 3Blue1Brown here.

The title of this section contains a widely used acronym: **ODE** $\hat{=}$ **o**rdinary **d**ifferential **e**quation . In this section we present a few examples of models involving ODEs and briefly review the relevant mathematical theory.

11.1.1 Ordinary Differential Equations (ODEs)

§11.1.1.1 (Terminology and notations related to ordinary differential equations) In our parlance, a **(first-order) ordinary differential equation** (ODE) is an equation of the form

$$\dot{\mathbf{y}} := \frac{d\mathbf{y}}{dt}(t) = \mathbf{f}(t, \mathbf{y}(t)) \,, \tag{ODE}$$

with

- ☞ a (continuous) right-hand-side function (r.h.s) $\mathbf{f} : I \times D \to \mathbb{R}^N$ of time $t \in \mathbb{R}$ and state $\mathbf{y} \in \mathbb{R}^N$,
- ☞ defined on a (finite) time interval $I \subset \mathbb{R}$, and **state space** D, which is some sub-set of $\mathbb{R}^N$: $D \subset \mathbb{R}^N$, $N \in \mathbb{N}$.

✎ Notation (due to Newton): dot $\dot{}$ $\hat{=}$ (total) derivative with respect to time t

An ODE is called **autonomous**, if the right-hand-side function $\mathbf{f}$ does not depend on time: $\mathbf{f} = \mathbf{f}(\mathbf{y})$, see Def. 11.1.2.4 below.

In the context of mathematical modeling the state vector $\mathbf{y} \in \mathbb{R}^N$ is supposed to provide a complete (in the sense of the model) description of a system. Then (ODE) models a finite-dimensional dynamical system. Examples will be provided below, see Ex. 11.1.2.1, Ex. 11.1.2.5, and Ex. 11.1.2.7.

For $N > 1$ $\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y})$ can be viewed as a system of ordinary differential equations:

$$\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y}) \quad \Longleftrightarrow \quad \begin{bmatrix} \dot{y}_1 \\ \vdots \\ \dot{y}_N \end{bmatrix} = \begin{bmatrix} f_1(t, y_1, \ldots, y_N) \\ \vdots \\ f_N(t, y_1, \ldots, y_N) \end{bmatrix} .$$

Definition 11.1.1.2. Solution of an ordinary differential equation

A solution of the ODE $\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y})$ with continuous right hand side function $\mathbf{f}$ is a continuously differentiable *function* "of time t" $\mathbf{y} : J \subset I \to D$, defined on an open interval J, for which $\dot{\mathbf{y}}(t) = \mathbf{f}(t, \mathbf{y}(t))$ holds for all $t \in J$ ($\hat{=}$ "*pointwise*").

A solution describes a continuous trajectory in state space, a one-parameter family of states, parameterized by time.

It goes without saying that smoothness of the right hand side function $\mathbf{f}$ is inherited by solutions of the ODE:

Lemma 11.1.1.3. Smoothness of solutions of ODEs

Let $\mathbf{y} : I \subset \mathbb{R} \to D$ be a solution of the ODE $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$ on the time interval I.

If $\mathbf{f} : D \to \mathbb{R}^N$ is r-times continuously differentiable with respect to both arguments, $r \in \mathbb{N}_0$, then the trajectory $t \mapsto \mathbf{y}(t)$ is $r+1$-times continuously differentiable in the interior of I.

⌟

§11.1.1.4 (Scalar autonomous ODE: Solution by principals) We consider scalar ODEs, namely (ODE) in the case $N = 1$, and, in particular $\dot{y} = f(y)$ with $f : D \subset \mathbb{R} \to \mathbb{R}$, D an interval.

We embark on *formal calculations*. Assume that f is continuous and $f(t) \neq 0$ for all $t \in D$. Further, suppose that we know a principal $F : D \to \mathbb{R}$ of $\frac{1}{f}$, that is, a function $y \mapsto F(y)$ satisfying $\frac{dF}{dy} = \frac{1}{f}$ on D. Then, by the chain rule, every solution $y : I \subset \mathbb{R} \to \mathbb{R}$ of $\dot{y} = f(y)$ also solves

$$\frac{d}{dt}F(y(t)) = \frac{1}{f(y(t))}\dot{y}(t) = 1\,, \quad t \in D\,. \quad \Leftrightarrow \quad F(y(t)) = t - t_0 \quad \text{for some} \quad t_0 \in \mathbb{R}\,. \tag{11.1.1.5}$$

We also know that F is monotonic and, thus, possesses an inverse function F^{-1}. Integrating (11.1.1.5) and applying the fundamental theorem of calculus, we find

$$y(t) = F^{-1}(t - t_0) \quad \text{for some} \quad t_0 \in I\,. \tag{11.1.1.6}$$

This formula describes a one-parameter family of functions (t_0 is the parameter), all of which provide a solution of $\dot{y} = f(y)$ on a suitable interval.

A particularly simple case is $f(y) = \lambda y + c$, $\lambda, c \in \mathbb{R}$, the scalar ODE $\dot{y} = \lambda y + c$. Following the steps outlined above, we calculate the solution

$$\left[F(y) = \frac{1}{\lambda}\log(\lambda y + x) \quad \Rightarrow\right] \quad y(t) = \frac{1}{\lambda}\left(e^{\lambda(t-t_0)} - c\right)\,, \quad t \in \mathbb{R}\,. \tag{11.1.1.7}$$

⌟

§11.1.1.8 (Linear ordinary differential equations) Now we take a look at the simplest class of ODEs, which is also the most important.

Definition 11.1.1.9. Linear first-order ODE

A first-order ordinary differential equation $\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y})$, as introduced in § 11.1.1.1 is **linear**, if

$$\mathbf{f}(t, \mathbf{y}) = \mathbf{A}(t)\mathbf{y} \quad , \quad \mathbf{A} : I \to \mathbb{R}^{N,N} \quad \text{a continuous function}\,. \tag{11.1.1.10}$$

Lemma 11.1.1.11. Space of solutions of linear ODEs

The set of solutions $\mathbf{y} : I \to \mathbb{R}^N$ of (11.1.1.10) *is a vector space.*

Proof. We have to show that, if $\mathbf{y}, \mathbf{z} : I \to \mathbb{R}^N$ are two solutions of (11.1.1.10), then so are $\mathbf{y} + \mathbf{z}$ and $\alpha\mathbf{y}$ for all $\alpha \in \mathbb{R}$. This is an immediate consequence of the linearity of the operations of differentiation and matrix×vector multiplication. □

For the scalar case $N = 1$ (11.1.1.10) can be written as $\dot{y} = a(t)y$ with a continuous function $a : I \to \mathbb{R}$. In this case, the chain rule immediately verifies that for fixed $t_0 \in I$ every function

$$y(t) = C \exp\left(\int_{t_0}^{t} a(\tau)\, \mathrm{d}\tau \right), \quad C \in \mathbb{R}, \tag{11.1.1.12}$$

is a solution.

If the matrix $\mathbf{A} \in \mathbb{R}^{N,N}$ does not depend on time, (11.1.1.10) is known as a linear ODE with constant coefficients: $\dot{\mathbf{y}} = \mathbf{A}\mathbf{y}$. In this case we can choose $I = \mathbb{R}$, and the ODE can be solved by a **diagonalization technique** [Str09, Bemerkung 5.6.1], [NS02, Sect. 8.1]: If

$$\mathbf{A} = \mathbf{S}\mathbf{D}\mathbf{S}^{-1}, \quad \mathbf{S} \in \mathbb{R}^{N,N} \text{ regular}, \quad \mathbf{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_N) \in \mathbb{R}^{N,N}, \tag{11.1.1.13}$$

we can rewrite

$$\dot{\mathbf{y}} = \mathbf{S}\mathbf{D}\mathbf{S}^{-1}\mathbf{y} \quad \Rightarrow \quad \dot{\mathbf{z}} = \mathbf{D}\mathbf{z} \quad \text{with} \quad \mathbf{z}(t) := \mathbf{S}^{-1}\mathbf{y}(t) .$$

We get N *decoupled* scalar linear equations $\dot{z}_\ell = \lambda_\ell z_\ell$, $\ell = 1, \ldots, N$. Returning to $\mathbf{y}$ we find that *every solution* $\mathbf{y} : \mathbb{R} \to \mathbb{R}^N$ of $\dot{\mathbf{y}} = \mathbf{A}\mathbf{y}$ can be written as

$$\mathbf{y}(t) = \mathbf{S}\mathbf{z}(t) = \mathbf{S} \begin{bmatrix} e^{\lambda_1 t} & & \\ & \ddots & \\ & & e^{\lambda_N t} \end{bmatrix} \mathbf{S}^{-1}\mathbf{w} \quad \text{for some} \quad \mathbf{w} \in \mathbb{R}^N . \tag{11.1.1.14}$$

This is an N-parameter family of solutions. ⌟

11.1.2 Mathematical Modeling with Ordinary Differential Equations: Examples

Most models of physical systems and phenomena that are *continuously changing with time* involve ordinary differential equations.

EXAMPLE 11.1.2.1 (Growth with limited resources [Ama83, Sect. 1.1], [Han02, Ch. 60]) This is an example from population dynamics with a one-dimensional state space $D = \mathbb{R}_0^+$, $N = 1$. The interpretation of $y : [0, T] \mapsto \mathbb{R}$ is that of the population density of bacteria as a function of time. A scaled, non-dimensional model is assumed, that is by fixing reference units all quantities are regarded as "mere numbers".

ODE-based model: autonomous logistic differential equations [Str09, Ex. 5.7.2]

$$\dot{y} = f(y) := (\alpha - \beta y)\, y \tag{11.1.2.2}$$

- $y \mathrel{\hat{=}}$ population density, $[y] = \frac{1}{\mathrm{m}^2}$
  ➢ $\dot{y} \mathrel{\hat{=}}$ instantaneous change (growth/decay) of population density
- growth rate $\alpha - \beta y$ with growth coefficients $\alpha, \beta > 0$, $[\alpha] = \frac{1}{\mathrm{s}}$, $[\beta] = \frac{\mathrm{m}^2}{\mathrm{s}}$: decreases due to fiercer competition as population density increases.

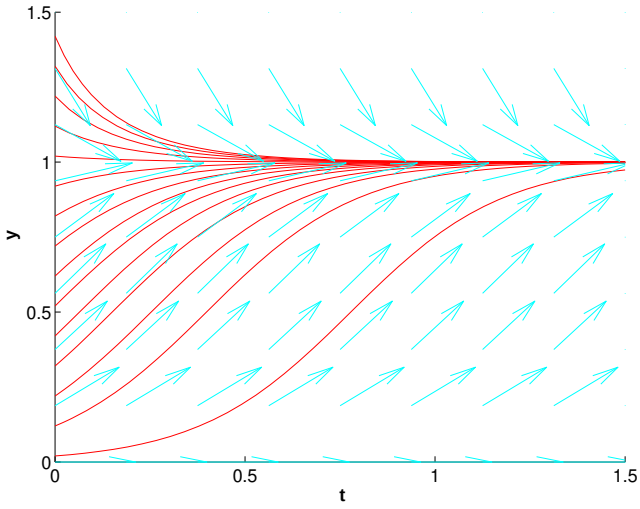


Fig. 399

Solution for different $y(0)$ ($\alpha, \beta = 5$)

By the technique from § 11.1.1.4 we can compute a *family of solutions* of (11.1.2.2) parameterized by the initial value $y(0) = y_0 > 0$:

$$y(t) = \frac{\alpha y_0}{\beta y_0 + (\alpha - \beta y_0)\exp(-\alpha t)}\,, \quad (11.1.2.3)$$

for all $t \in \mathbb{R}$.

Note: $f(y^*) = 0$ for $y^* \in \{0, \alpha/\beta\}$, which are the stationary points for the ODE (11.1.2.2). If $y(0) = y^*$ the solution will be constant in time.

Note that by fixing the initial value $y(0)$ we can single out a *unique* representative from the family of solutions. This will turn out to be a general principle, see Section 11.1.3. ⌟

Definition 11.1.2.4. Autonomous ODE

An ODE of the form $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$, that is, with a right hand side function that does not depend on time, but only on state, is called **autonomous**.

For an autonomous ODE the right hand side function defines a **vector field** ("velocity field") $\mathbf{y} \mapsto \mathbf{f}(\mathbf{y})$ on state space.

EXAMPLE 11.1.2.5 (Predator-prey model [Ama83, Sect. 1.1],[HLW06, Sect. 1.1.1],[Han02, Ch. 60], [DR08, Ex. 11.3]) We consider the following model from population dynamics:

Predators and prey coexist in an ecosystem. Without predators the population of prey would be governed by a simple exponential growth law. However, the growth rate of prey will decrease with increasing numbers of predators and, eventually, become negative. Similar considerations apply to the predator population and lead to an ODE model.

ODE-based model: autonomous Lotka-Volterra ODE:

$$\begin{aligned} \dot{u} &= (\alpha - \beta v)u \\ \dot{v} &= (\delta u - \gamma)v \end{aligned} \quad \leftrightarrow \quad \dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}) \quad \text{with} \quad \mathbf{y} = \begin{bmatrix} u \\ v \end{bmatrix}, \quad \mathbf{f}(\mathbf{y}) = \begin{bmatrix} (\alpha - \beta v)u \\ (\delta u - \gamma)v \end{bmatrix}, \quad (11.1.2.6)$$

with positive model parameters $\alpha, \beta, \gamma, \delta > 0$ and population densities

$$u(t) \mathrel{\hat{=}} \text{density of prey at time } t \quad , \quad v(t) \mathrel{\hat{=}} \text{density of predators at time } t.$$

Plot of right-hand side vector field $\mathbf{f}$ for Lotka-Volterra ODE ▷

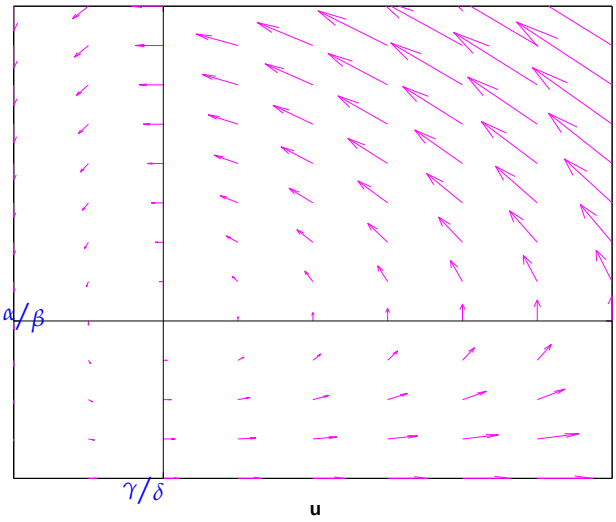


Fig. 400

Solution curves are trajectories of particles carried along by velocity field $\mathbf{f}$.

(Parameter values for Fig. 400: $\alpha = 2, \beta = 1, \delta = 1, \gamma = 1$.)

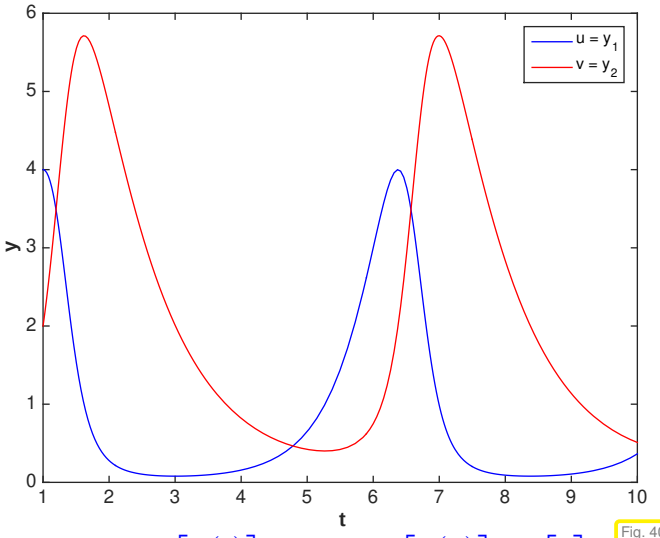


Fig. 401

Solution $\begin{bmatrix} u(t) \\ v(t) \end{bmatrix}$ for $\mathbf{y}_0 := \begin{bmatrix} u(0) \\ v(0) \end{bmatrix} = \begin{bmatrix} 4 \\ 2 \end{bmatrix}$

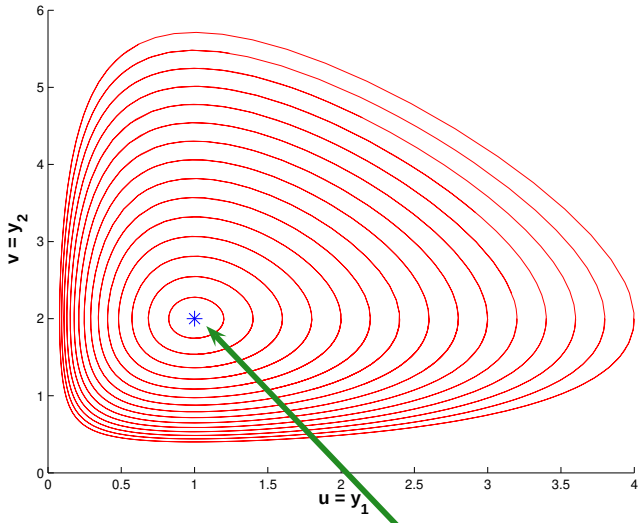


Fig. 402

Solution curves for (11.1.2.6)

stationary point

(Parameter values for Fig. 402, 401: $\alpha = 2, \beta = 1, \delta = 1, \gamma = 1$)

EXAMPLE 11.1.2.7 (Heartbeat model → [Dea80, p. 655]) This example deals with a *phenomenological model* from physiology. A model is called phenomenological, if it is entirely motivated by observations without appealing to underlying mechanisms or first principles.

State of heart described by quantities:

$$\begin{aligned} l = l(t) &\;\hat{=}\; \text{length of muscle fiber} \\ p = p(t) &\;\hat{=}\; \text{electro-chemical potential} \end{aligned}$$

Phenomenological model:

$$\begin{aligned} \dot{l} &= -(l^3 - \alpha l + p)\,, \\ \dot{p} &= \beta l\,, \end{aligned} \tag{11.1.2.8}$$

with parameters:

$$\begin{aligned} \alpha &\;\hat{=}\; \text{pre-tension of muscle fiber} \\ \beta &\;\hat{=}\; \text{(phenomenological) feedback parameter} \end{aligned}$$

This is the so-called Zeeman model: it is a phenomenological model entirely based on macroscopic observations without relying on knowledge about the underlying molecular mechanisms.

Plots of vector fields for (11.1.2.8) and solutions for different choices of parameters are given next:

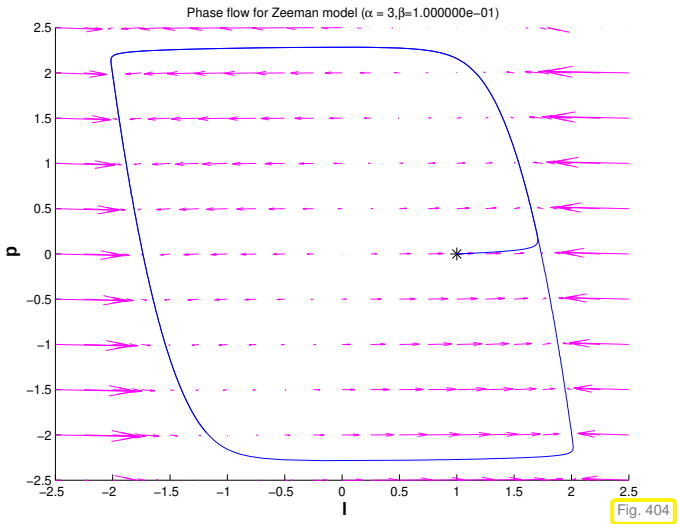


Fig. 403

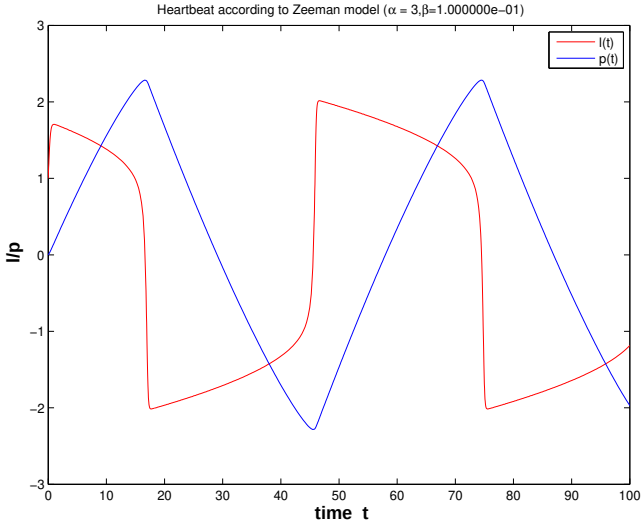


Fig. 404

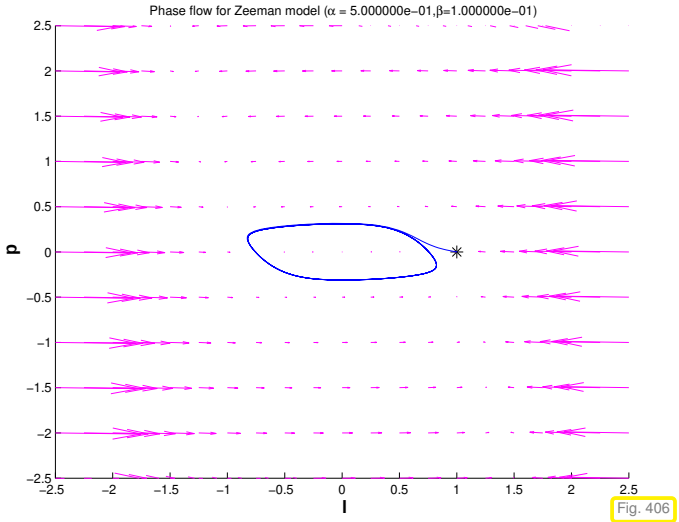


Fig. 405

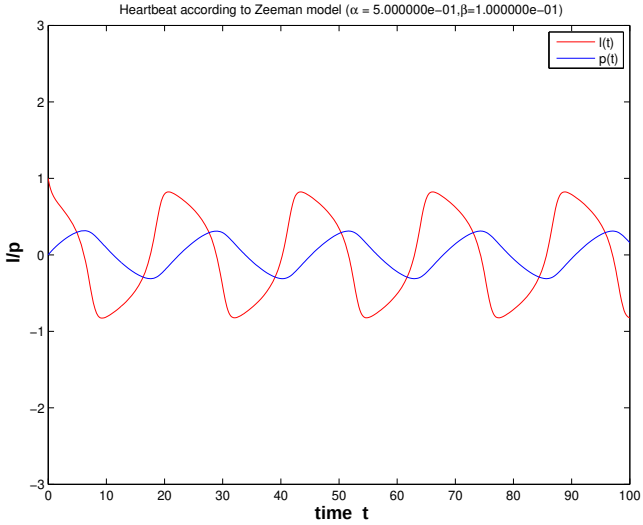


Fig. 406

Observation: $\alpha \ll 1$ (bottom plots) ➤ ventricular fibrillation, a life-threatening condition. ⌟

EXAMPLE 11.1.2.9 (SIR model for spread of local epidemic [Het00]) The field of epidemiology tries to understand the spread of contagious diseases in populations. It heavily relies on ODEs in its mathematical modeling. This example presents a particularly simple model for an epidemic in a large, stable, isolated, and vigorously mixing homogeneous population.

With respect to the disease we partition the population into three different groups and introduce time-dependent variables for their fractions $\in [0,1]$:

(I) $S = S(t) \hat{=}$ fraction of susceptible persons, who can still contract the disease,

(II) $I = I(t) \hat{=}$ fraction of infected/infectious persons, who can pass on the disease,

(III) $R = R(t) \hat{=}$ fraction of recovered/removed persons, who are immune or have died.

These three quantities enter the SIR model named after the groups it considers. Besides, the model involves two crucial model parameters, which have to be determined from data:

1. A parameter $\beta > 0$, whose value expresses the probability of transmission, and
2. a parameter $r > 0$, taking into account how quickly sick people recover or die.

With these notation the ODE underlying the SIR model can be stated as

$$\dot{S}(t) = -\beta S(t)I(t) \quad , \quad \dot{I}(t) = \beta S(t)I(t) - rI(t) \quad , \quad \dot{R}(t) = rI(t) \, . \tag{11.1.2.10}$$

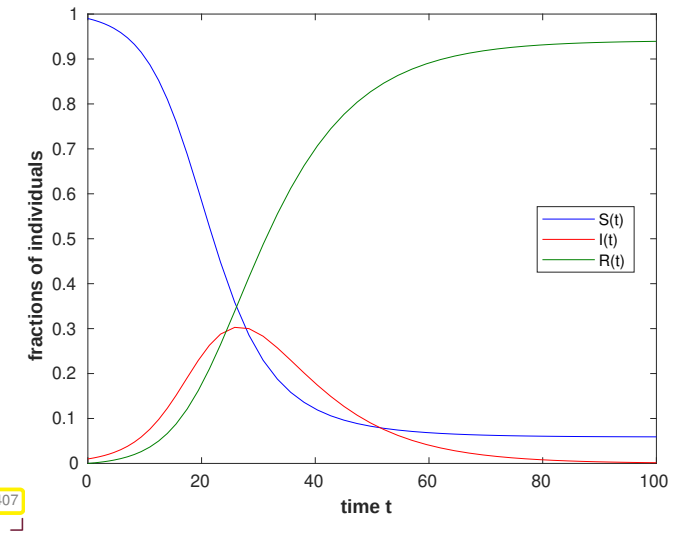


Fig. 407

◁ Evolution of an epidemic according to the SIR model (11.1.2.10) for
- $\beta = 0.3$, $r = 0.1$,
- $S(0) = 0.99$, $I(0) = 0.01$, $R(0) = 0.0$

(non-dimensionalized time)

Note that in this case not all people end up infected, $\lim_{t\to\infty} S(t) > 0$!

EXAMPLE 11.1.2.11 (Transient circuit simulation [Han02, Ch. 64]) Chapter 1 and Chapter 8 discuss circuit analysis as a source of linear and non-linear systems of equations, see Ex. 2.1.0.3 and Ex. 8.1.0.1. The former example admitted time-dependent currents and potentials, but dependence on time was confined to be "sinusoidal". This enabled us to switch to frequency domain, see (2.1.0.6), which gave us a complex linear system of equations for the complex nodal potentials. Yet, this trick is only possible for *linear* circuits. In the general case, circuits have to be modelled by ODEs connecting time-dependent potentials and currents. This will be briefly explained now.

The approach is transient nodal analysis, *cf.* Ex. 2.1.0.3, based on the Kirchhoff current law (2.1.0.4), which reads for the node • of the simple circuit drawn in Fig. 408

$$i_R(t) - i_L(t) - i_C(t) = 0\,. \tag{11.1.2.12}$$

In addition we rely on known transient constitutive relations for basic linear circuit elements:

resistor: $$i_R(t) = R^{-1}u_R(t)\,, \tag{11.1.2.13}$$

capacitor: $$i_C(t) = C\frac{du_C}{dt}(t)\,, \tag{11.1.2.14}$$

coil: $$u_L(t) = L\frac{di_L}{dt}(t)\,. \tag{11.1.2.15}$$

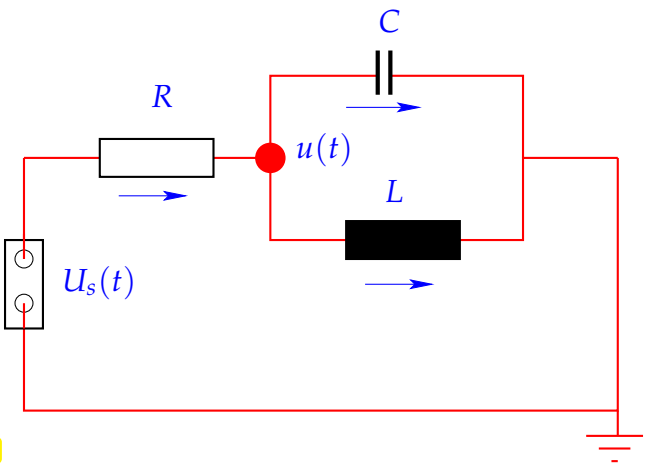


Fig. 408

We assume that the source voltage $U_s(t)$ is given. To apply nodal analysis to the circuit of Fig. 408 we differentiate (11.1.2.12) w.r.t. t

$$\frac{di_R}{dt}(t) - \frac{di_L}{dt}(t) - \frac{di_C}{dt}(t) = 0\,,$$

and plug in the above constitutive relations for circuit elements:

$$R^{-1}\frac{du_R}{dt}(t) - L^{-1}u_L(t) - C\frac{d^2u_C}{dt^2}(t) = 0\,.$$

We continue following the policy of nodal analysis and express all voltages by potential differences between nodes of the circuit.

$$u_R(t) = U_s(t) - u(t)\,,\quad u_C(t) = u(t) - 0\,,\quad u_L(t) = u(t) - 0\,.$$

For this simple circuit there is only one node with unknown potential, see Fig. 408. Its time-dependent potential will be denoted by $u(t)$ and this is the unknown of the model, a function of time satisfying the ordinary differential equation

$$R^{-1}(\dot{U}_s(t) - \dot{u}(t)) - L^{-1}u(t) - C\frac{d^2u}{dt^2}(t) = 0\,.$$

This is an autonomous 2nd-order ordinary differential equation:

$$C\ddot{u} + R^{-1}\dot{u} + L^{-1}u = R^{-1}\dot{U}_s\,. \tag{11.1.2.16}$$

The attribute "2nd-order" refers to the occurrence of a second derivative with respect to time. ⌟

11.1.3 Theory of Initial-Value-Problems (IVPs)

§11.1.3.1 (Initial value problems) We start with an abstract mathematical description that also introduces key terminology:

A generic **Initial value problem** (IVP) for a first-order ordinary differential equation (ODE) (→ [Str09, Sect. 5.6], [DR08, Sect. 11.1]) can be stated as:

Find a function $\mathbf{y} : I \to D$ that satisfies, *cf.* Def. 11.1.1.2,

$$\dot{\mathbf{y}} = \mathbf{f}(t,\mathbf{y}) \quad , \quad \mathbf{y}(t_0) = \mathbf{y}_0\,. \tag{11.1.3.2}$$

- $\mathbf{f} : I \times D \mapsto \mathbb{R}^N \triangleq$ right hand side (r.h.s.) ($N \in \mathbb{N}$),
- $I \subset \mathbb{R} \triangleq$ (time)interval $\leftrightarrow$ "time variable" t,
- $D \subset \mathbb{R}^N \triangleq$ **state space**/phase space $\leftrightarrow$ "state variable" $\mathbf{y}$,
- $\Omega := I \times D \triangleq$ extended state space (of tupels $(t,\mathbf{y})$),
- $t_0 \in I \triangleq$ initial time, $\mathbf{y}_0 \in D \triangleq$ initial state ➢ **initial conditions**.

The time interval I may be finite or infinite. Frequently, the extended state space is not specified, but assumed to coincide with the maximal domain of definition of $\mathbf{f}$. Sometimes, the model suggests constraints on D, for instance, positivity of certain components that represent a density. ⌟

§11.1.3.3 (IVPs for autonomous ODEs) Recall Def. 11.1.2.4: For an autonomous ODE $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$, that is the right hand side $\mathbf{f}$ does not depend on time t.

Hence, for autonomous ODEs we have $I = \mathbb{R}$ and the right hand side function $\mathbf{y} \mapsto \mathbf{f}(\mathbf{y})$ can be regarded as a stationary vector field (velocity field), see Fig. 400 or Fig. 403.

An important observation: If $t \mapsto \mathbf{y}(t)$ is a solution of an autonomous ODE, then, for any $\tau \in \mathbb{R}$, also the shifted function $t \mapsto \mathbf{y}(t-\tau)$ is a solution.

Initial time for autonomous ODEs

For initial value problems for autonomous ODEs the initial time is irrelevant and therefore we can always make the "canonical" choice $t_0 = 0$.

Autonomous ODEs naturally arise when modeling time-invariant systems or phenomena. All examples for Section 11.1.2 belong to this class. ⌟

§11.1.3.5 (Autonomization: Conversion into autonomous ODE) In fact, autonomous ODEs already represent the general case, because any ODE can be converted into an autonomous one:

The idea is to include time as an extra $N+1$-st component of an extended state vector $\mathbf{z}(t)$. This solution component has to grow linearly $\Leftrightarrow$ its temporal derivative must be $= 1$

$$\mathbf{z}(t) := \begin{bmatrix} \mathbf{y}(t) \\ t \end{bmatrix} = \begin{bmatrix} \mathbf{z}' \\ z_{N+1} \end{bmatrix}: \quad \dot{\mathbf{y}} = \mathbf{f}(t,\mathbf{y}) \quad \leftrightarrow \quad \dot{\mathbf{z}} = \mathbf{g}(\mathbf{z}) \,, \quad \mathbf{g}(\mathbf{z}) := \begin{bmatrix} \mathbf{f}(z_{N+1},\mathbf{z}') \\ 1 \end{bmatrix} .$$

This means $\dot{z}_{N+1} = 1$ and implies $z_{N+1}(t) = t + t_0$, if t_0 stands for the initial time in the original non-autonomous IVP.

➢ We restrict ourselves to autonomous ODEs in the remainder of this chapter. ⌟

Remark 11.1.3.6 (From higher order ODEs to first order systems [DR08, Sect. 11.2])

An ordinary differential equation of **order** $n \in \mathbb{N}$ has the form

$$\boxed{\mathbf{y}^{(n)} = \mathbf{f}(t,\mathbf{y},\dot{\mathbf{y}},\ldots,\mathbf{y}^{(n-1)})} \,. \tag{11.1.3.7}$$

where, with notations from § 11.1.3.1, $\mathbf{f} : I \times D \times \cdots \times D \to \mathbb{R}^N$ is a function of time t and n state arguments.

✎ Notation: superscript $^{(n)} \hat{=}$ n-th temporal derivative t: $\frac{d^n}{dt^n}$

No special treatment of higher order ODEs is necessary, because (11.1.3.7) can be turned into a 1st-order ODE (a system of size nN) by adding all derivatives up to order $n-1$ as additional components to the state vector. This extended state vector $\mathbf{z}(t) \in \mathbb{R}^{nd}$ is defined as

$$\mathbf{z}(t) := \begin{bmatrix} \mathbf{y}(t) \\ \mathbf{y}^{(1)}(t) \\ \vdots \\ \mathbf{y}^{(n-1)}(t) \end{bmatrix} = \begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \\ \vdots \\ \mathbf{z}_n \end{bmatrix} \in \mathbb{R}^{Nn}: \quad (11.1.3.7) \quad \leftrightarrow \quad \dot{\mathbf{z}} = \mathbf{g}(\mathbf{z}) \,, \quad \mathbf{g}(\mathbf{z}) := \begin{bmatrix} \mathbf{z}_2 \\ \mathbf{z}_3 \\ \vdots \\ \mathbf{z}_n \\ \mathbf{f}(t,\mathbf{z}_1,\ldots,\mathbf{z}_n) \end{bmatrix} . \tag{11.1.3.8}$$

Note that the extended system requires initial values $\mathbf{y}(t_0), \dot{\mathbf{y}}(t_0), \ldots, \mathbf{y}^{(n-1)}(t_0)$:

For ODEs of order $n \in \mathbb{N}$ well-posed initial value problems need to specify initial values for the first $n-1$ derivatives.

⌟

§11.1.3.10 (Smoothness classes for right-hand side functions) Now we review results about existence and uniqueness of solutions of initial value problems for first-order ODEs. These are surprisingly general and do not impose severe constraints on right hand side functions. Some kind of smoothness of the right-hand side function $\mathbf{f}$ is required, nevertheless and the following definitions describe it in detail.

Definition 11.1.3.11. Lipschitz continuous function ($\rightarrow$ [Str09, Def. 4.1.4])

Let $\Theta := I \times D$, $I \subset \mathbb{R}$ an interval, $D \subset \mathbb{R}^N$, $N \in \mathbb{N}$, an open domain. A function $\mathbf{f} : \Theta \mapsto \mathbb{R}^N$ is Lipschitz continuous (in the second argument) on Θ, if

$$\exists L > 0: \quad \|\mathbf{f}(t,\mathbf{w}) - \mathbf{f}(t,\mathbf{z})\| \le L\|\mathbf{w} - \mathbf{z}\| \quad \forall (t,\mathbf{w}),(t,\mathbf{z}) \in \Theta \,. \tag{11.1.3.12}$$

Definition 11.1.3.13. Local Lipschitz continuity (→ [Str09, Def. 4.1.5])

Let $\Omega := I \times D$, $I \subset \mathbb{R}$ an interval, $D \subset \mathbb{R}^N$, $N \in \mathbb{N}$, an open domain. A functions $\mathbf{f} : \Omega \mapsto \mathbb{R}^N$ is locally Lipschitz continuous, if for every $(t, \mathbf{y}) \in \Omega$ there is a closed box B with $(t, \mathbf{y}) \in B$ such that $\mathbf{f}$ is Lipschitz continuous on B:

$$\begin{aligned} \forall (t,\mathbf{y}) \in \Omega\colon \quad & \exists \delta > 0,\, L > 0\colon \\ & \|\mathbf{f}(\tau,\mathbf{z}) - \mathbf{f}(\tau,\mathbf{w})\| \le L\|\mathbf{z}-\mathbf{w}\| \\ & \forall \mathbf{z},\mathbf{w} \in D\colon \|\mathbf{z}-\mathbf{y}\| \le \delta,\ \|\mathbf{w}-\mathbf{y}\| \le \delta, \forall \tau \in I\colon |t-\tau| \le \delta\ . \end{aligned} \tag{11.1.3.14}$$

The property of local Lipschitz continuity means that the function $(t, \mathbf{y}) \mapsto \mathbf{f}(t, \mathbf{y})$ has "locally finite slope" in $\mathbf{y}$. ⌟

EXAMPLE 11.1.3.15 (A function that is not locally Lipschitz continuous [Str09, Bsp. 6.5.3]) The meaning of local Lipschitz continuity is best explained by giving an example of a function that fails to possess this property.

Consider the square root function $t \mapsto \sqrt{t}$ on the *closed* interval $[0, 1]$. Its slope in $t = 0$ is infinite and so it is not locally Lipschitz continuous on $[0, 1]$.

However, if we consider the square root on the *open* interval $]0, 1[$, then it is locally Lipschitz continuous there. ⌟

The next lemma gives a simple criterion for local Lipschitz continuity, which can be proved by the mean value theorem, *cf.* the proof of Lemma 8.3.2.9.

Lemma 11.1.3.16. Criterion for local Liptschitz continuity

If $\mathbf{f}$ and $D_{\mathbf{y}}\mathbf{f}$ are continuous on the extended state space Ω, then $\mathbf{f}$ is locally Lipschitz continuous (→ Def. 11.1.3.13).

✎ Notation: $\mathrm{D}_{\mathbf{y}}\,\mathbf{f} \hat{=}$ the derivative of $\mathbf{f}$ w.r.t. the state variable $\mathbf{y}$, a Jacobian matrix $\in \mathbb{R}^{N,N}$ as defined in (8.3.2.8).

The following is the the most important mathematical result in the theory of initial-value problems for ODEs:

Theorem 11.1.3.17. Theorem of Peano & Picard-Lindelöf [Ama83, Satz II(7.6)], [Str09, Satz 6.5.1], [DR08, Thm. 11.10], [Han02, Thm. 73.1]

If the right hand side function $\mathbf{f} : \Omega \mapsto \mathbb{R}^N$ is locally Lipschitz continuous (→ Def. 11.1.3.13) then for all initial conditions $(t_0, \mathbf{y}_0) \in \Omega$ the IVP

$$\dot{\mathbf{y}} = \mathbf{f}(t,\mathbf{y}) \quad , \quad \mathbf{y}(t_0) = \mathbf{y}_0\ . \tag{11.1.3.2}$$

has a solution $\mathbf{y} \in C^1(J(t_0,\mathbf{y}_0), \mathbb{R}^N)$ with maximal (temporal) domain of definition $J(t_0, \mathbf{y}_0) \subset \mathbb{R}$.

In light of § 11.1.3.5 and Thm. 11.1.3.17 henceforth we mainly consider

autonomous IVPs: $\boxed{\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}) \quad , \quad \mathbf{y}(0) = \mathbf{y}_0}$, (11.1.3.18)

with locally Lipschitz continuous (→ Def. 11.1.3.13) right hand side $\mathbf{f}$.

§11.1.3.19 (Domain of definition of solutions of IVPs) We emphasize a subtle message of Thm. 11.1.3.17.

Solutions of an IVP have an *intrinsic* maximal domain of definition

! Also not that the domain of definition/domain of existence $J(t_0, \mathbf{y}_0)$ of the solution usually depends on the initial values $(t_0, \mathbf{y}_0)$!

Terminology: if $J(t_0, \mathbf{y}_0) = I$, I the maximal temporal domain of definition of $\mathbf{f}$, we say that the solution $\mathbf{y} : I \mapsto \mathbb{R}^N$ is global.

Notation: For autonomous ODE we always have $t_0 = 0$, and therefore we write $J(\mathbf{y}_0) := J(0, \mathbf{y}_0)$. ⌟

EXAMPLE 11.1.3.20 ("Explosion equation": finite-time blow-up) Let us explain the still mysterious "maximal domain of definition" in statement of Thm. 11.1.3.17. It is related to the fact that every solution of an initial value problem (11.1.3.18) has its own largest possible time interval $J(\mathbf{y}_0) \subset \mathbb{R}$ on which it is defined naturally.

As an example we consider the autonomous scalar ($d = 1$) initial value problem, modeling "explosive growth" with a growth rate increasing linearly with the density:

$$\dot{y} = y^2 \quad , \quad y(0) = y_0 \in \mathbb{R} \, . \tag{11.1.3.21}$$

We choose $I = D = \mathbb{R}$. Clearly, $y \mapsto y^2$ is locally Lipschitz-continuous, but only locally! Why not globally?

We find the solutions

$$y(t) = \begin{cases} \frac{1}{y_0^{-1} - t} & \text{, if } y_0 \neq 0 \, , \\ 0 & \text{, if } y_0 = 0 \, , \end{cases} \tag{11.1.3.22}$$

with domains of definition

$$J(y_0) = \begin{cases}]-\infty, y_0^{-1}[& \text{, if } y_0 > 0 \, , \\ \mathbb{R} & \text{, if } y_0 = 0 \, , \\ ]y_0^{-1}, \infty[& \text{, if } y_0 < 0 \, . \end{cases}$$

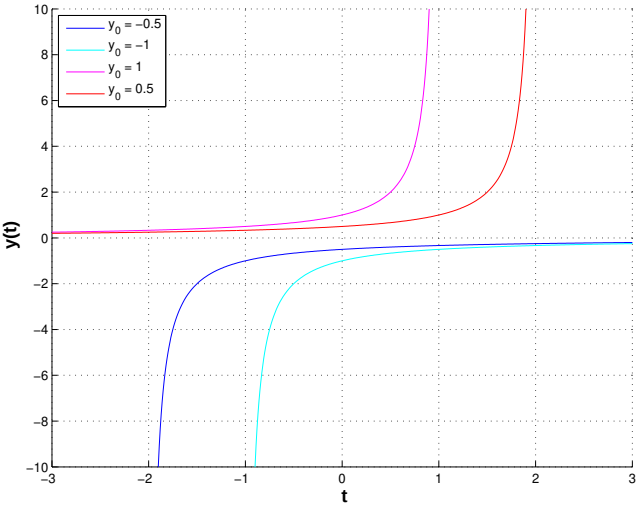


Fig. 409

In this example, for $y_0 > 0$ the solution experiences a blow-up in finite time and ceases to exists afterwards. ⌟

Supplementary literature. For other concise summaries of the theory of IVPs for ODEs refer to [QSS00, Sect. 11.1], [DR08, Sect. 11.3].

11.1.4 Evolution Operators

Now we examine a difficult but fundamental concept for time-dependent models stated by means of ODEs. For the sake of simplicity we restrict the discussion to autonomous initial-value problems (IVPs)

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}) \quad , \quad \mathbf{y}(0) = \mathbf{y}_0 \, , \tag{11.1.3.18}$$

with locally Lipschitz continuous ($\rightarrow$ Def. 11.1.3.13) right hand side $\mathbf{f} : D \subset \mathbb{R}^N \to \mathbb{R}^N$, $N \in \mathbb{N}$, and make the following assumption. A more general treatment is given in [DB02].

Assumption 11.1.4.1. Global solutions

Solutions of (11.1.3.18) are global: $J(\mathbf{y}_0) = \mathbb{R}$ for *all* $\mathbf{y}_0 \in D$.

Now we return to the study of a generic ODE (ODE) instead of an IVP (11.1.3.2). We do this by temporarily changing the perspective: we fix a "time of interest" $t \in \mathbb{R} \setminus \{0\}$ and follow all trajectories for the duration t. This induces a mapping of points in state space:

$$\succ \text{ mapping } \quad \boldsymbol{\Phi}^t : \begin{cases} D & \mapsto & D \\ \mathbf{y}_0 & \mapsto & \mathbf{y}(t) \end{cases} \quad , \quad t \mapsto \mathbf{y}(t) \text{ solution of IVP (11.1.3.18)} \, , \tag{11.1.4.2}$$

This is a well-defined mapping of the state space into itself, by Thm. 11.1.3.17 and Ass. 11.1.4.1.

Now, we may also let t vary, which spawns a *family* of mappings $\left\{\boldsymbol{\Phi}^t\right\}_{t\in\mathbb{R}}$ of the state space D into itself. However, it can also be viewed as a mapping with two arguments, a duration t and an initial state value $\mathbf{y}_0$!

Definition 11.1.4.3. Evolution operator/mapping

Under Ass. 11.1.4.1 the mapping

$$\boldsymbol{\Phi} : \begin{cases} \mathbb{R} \times D & \mapsto & D \\ (t, \mathbf{y}_0) & \mapsto & \boldsymbol{\Phi}^t \mathbf{y}_0 := \mathbf{y}(t) \end{cases} ,$$

where $t \mapsto \mathbf{y}(t) \in C^1(\mathbb{R}, \mathbb{R}^N)$ is the unique (global) solution of the IVP $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$, $\mathbf{y}(0) = \mathbf{y}_0$, is the **evolution operator**/mapping for the autonomous ODE $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$.

Note that $t \mapsto \boldsymbol{\Phi}^t \mathbf{y}_0$ describes the solution of $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$ for $\mathbf{y}(0) = \mathbf{y}_0$ (a trajectory). Therefore, by virtue of definition, we have

$$\frac{\partial \boldsymbol{\Phi}}{\partial t}(t, \mathbf{y}) = \mathbf{f}(\boldsymbol{\Phi}^t \mathbf{y}) \, . \tag{11.1.4.4}$$

Let us repeat the different kinds of information contained in an evolution operator when viewed from different angles:

$$\begin{aligned} t \mapsto \boldsymbol{\Phi}^t \mathbf{y}_0 \, , \quad \mathbf{y}_0 \in D \text{ fixed} \quad &\widehat{=} \quad \text{a trajectory = solution of an IVP} \, , \\ \mathbf{y} \mapsto \boldsymbol{\Phi}^t \mathbf{y} \, , \quad t \in \mathbb{R} \text{ fixed} \quad &\widehat{=} \quad \text{a mapping of the state space onto itself} \, . \end{aligned}$$

EXAMPLE 11.1.4.5 (Evolution operator for Lotka-Volterra ODE (11.1.2.6)**)** For $N = 2$ the action of an evolution operator can be visualized by tracking the movement of point sets in state space. Here this is done for the Lotka-Volterra ODE

$$\begin{aligned} \dot{u} &= (\alpha - \beta v)u \\ \dot{v} &= (\delta u - \gamma)v \end{aligned} \quad \leftrightarrow \quad \dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}) \quad \text{with} \quad \mathbf{y} = \begin{bmatrix} u \\ v \end{bmatrix}, \quad \mathbf{f}(\mathbf{y}) = \begin{bmatrix} (\alpha - \beta v)u \\ (\delta u - \gamma)v \end{bmatrix}, \tag{11.1.2.6}$$

with positive model parameters $\alpha, \beta, \gamma, \delta > 0$.

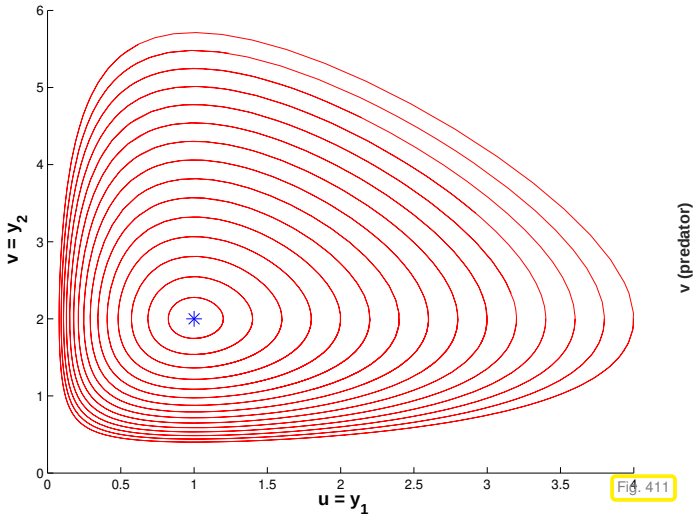


Fig. 410

trajectories $t \mapsto \mathbf{\Phi}^t \mathbf{y}_0$

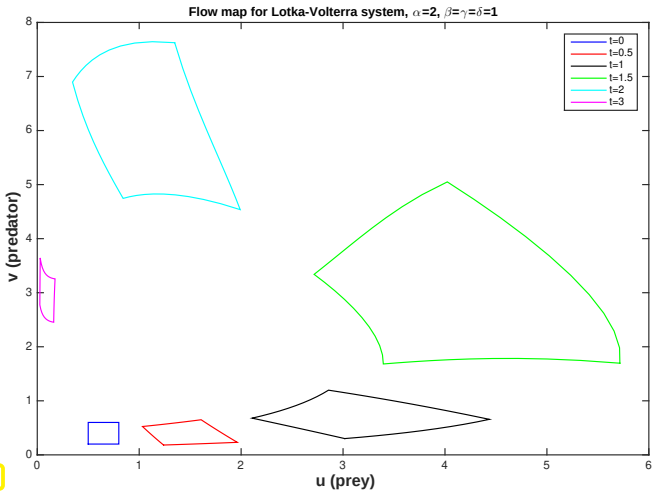


Fig. 411

state mapping $\mathbf{y} \mapsto \Phi^t \mathbf{y}$

Think of $\mathbf{y} \in \mathbb{R}^2 \mapsto \mathbf{f}(\mathbf{y}) \in \mathbb{R}^2$ as the velocity of the surface of a fluid. Specks of floating dust will be carried along by the fluid, patches of dust covering parts of the surface will move and deform over time. This can serve as a "mental image" of $\mathbf{\Phi}$. ⌟

Given an evolution operator, we can recover the right-hand side function $\mathbf{f}$ of the underlying autonomous ODE as $\mathbf{f}(\mathbf{y}) = \frac{\partial \mathbf{\Phi}}{\partial t}(0, \mathbf{y})$: There is a one-to-one relationship between ODEs and their evolution operators, and those are the key objects behind an ODE.

An ODE "encodes" an evolution operator.

Understanding the concept of evolution operators is indispensable for numerical integration, that the is the construction of numerical methods for the solution of IVPs for ODEs:

Numerical integration is concerned with the approximation of evolution operators.

Remark 11.1.4.6 (Group property of autonomous evolutions) Under Ass. 11.1.4.1 the evolution operator gives rise to a group of mappings $D \mapsto D$:

$$\mathbf{\Phi}^s \circ \mathbf{\Phi}^t = \mathbf{\Phi}^{s+t} \quad , \quad \mathbf{\Phi}^{-t} \circ \mathbf{\Phi}^t = Id \quad \forall t \in \mathbb{R} \, . \tag{11.1.4.7}$$

This is a consequence of the uniqueness theorem Thm. 11.1.3.17. It is also intuitive: following an evolution up to time t and then for some more time s leads us to the same final state as observing it for the whole time $s + t$. ⌟

Review question(s) 11.1.4.8 (IVPs for ODEs)

(Q11.1.4.8.A) A simple model for the spread of a viral epidemic like SARS-Cov2 is the **SIR model**:

$$\dot{S}(t) = -\beta I(t) S(t) \, , \quad \dot{I}(t) = \beta I(t) S(t) - \gamma I(t) \, , \quad \dot{R}(t) = \gamma I(t) \, , \tag{11.1.4.9}$$

with parameters $\beta, \gamma > 0$. Here $t \mapsto S(t)$ is the fraction of susceptible individuals, $t \mapsto I(t)$ that of infected (and infectious) individuals, and $t \mapsto R(t)$ stands for the removed (immune or dead) individuals.

- Write (11.1.4.9) in the form $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$.
- What is a meaningful state space for (11.1.4.9).
- Show that $S(t) + I(t) + R(t) \equiv \text{const}$.
- Show that $t \mapsto R(t)$ is non-decreasing.

(Q11.1.4.8.B) Determine the one-parameter family of solutions of the scalar autonomous ODE $\dot{y} = 1 + y^2$. Can you expect global solutions defined for all times $t \in \mathbb{R}$?

Hint. $\frac{d}{dy}\{y \mapsto \arctan(y)\} = \frac{1}{1+y^2}$

(Q11.1.4.8.C) Consider the autonomous scalar ODE $\dot{y} = \cos^2 y$.

- What are the stationary states, that is the states y^* that are zeros of the right-hand-side function?
- Compute the (analytic))solution of a related initial-value problem with $y(0) = y_0 \in \mathbb{R}$.
- The evolution operator Φ belonging to $\dot{y} = \cos^2 y$ will satisfy $\Phi^t \circ \Phi^s = \Phi^{t+s}$. Verify this formula based on what you found as analytic solution.

Hint. Remember that $\tan' = \cos^{-2}$.

(Q11.1.4.8.D) Show that the scalar autonomous initial-value problem

$$\dot{y} = \sqrt{y} \quad , \quad y(0) = 0 \, ,$$

has at least two solutions in the state space $\mathbb{R}_0^+$ according to the following definition.

Definition Def. 11.1.1.2. Solution of an ordinary differential equation

A solution of the ODE $\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y})$ with continuous right hand side function $\mathbf{f}$ is a continuously differentiable *function* "of time t" $\mathbf{y} : J \subset I \to D$, defined on an open interval J, for which $\dot{\mathbf{y}}(t) = \mathbf{f}(t, \mathbf{y}(t))$ holds for all $t \in J$ ($\hat{=}$ "*pointwise*").

How can this be reconciled with the assertion of the main theorem?

Theorem Thm. 11.1.3.17. Theorem of Peano & Picard-Lindelöf

If the right hand side function $\mathbf{f} : \hat{\Omega} \mapsto \mathbb{R}^N$ is locally Lipschitz continuous ($\to$ Def. 11.1.3.13) then for all initial conditions $(t_0, \mathbf{y}_0) \in \hat{\Omega}$ the IVP

$$\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y}) \quad , \quad \mathbf{y}(t_0) = \mathbf{y}_0 \, . \tag{11.1.3.2}$$

has a solution $\mathbf{y} \in C^1(J(t_0, \mathbf{y}_0), \mathbb{R}^N)$ with maximal (temporal) domain of definition $J(t_0, \mathbf{y}_0) \subset \mathbb{R}$.

Hint. Consider the function $y(t) = \left(\frac{1}{2}t\right)^2$.

(Q11.1.4.8.E) For the autonomous scalar ODE $\dot{y} = \sin\frac{1}{y} - 2$ answer the following questions

- What is the maximal state space?
- Which initial values for $t_0 = 0$ will allow a solution on $[0, \infty[$,
- and for which will the solution be defined for a finite time interval only?

Hint. Make use of the geometrically intuitive statement: If a differentiable function $f : [t_0, T] \to \mathbb{R}$ satisfies $\dot{f}(t) \leq C$ for all $t_0 \leq t \leq T$, then $f(t) \leq f(t_0) + Ct$.

(Q11.1.4.8.F) Rewrite the matrix differential equation $\dot{\mathbf{Y}}(t) = \mathbf{A}\mathbf{Y}(t)$ for $\mathbf{Y} : \mathbb{R} \to \mathbb{R}^{n,n}$, $n \in \mathbb{N}$, in the standard form $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$ with right-hand-side function $\mathbf{f} : \mathbb{R}^N \to \mathbb{R}^N$ and suitable $N \in \mathbb{N}$.

(Q11.1.4.8.G) What "ingredients" does it take to define an initial value problem for an ODE?

△

11.2 Introduction: Polygonal Approximation Methods

Video tutorial for Section 11.2: Introduction: Polygonal Approximation Methods: (17 minutes)
Download link, tablet notes

In this section we will see the first simple methods for the numerical integration (= solution) of initial-value problems (IVPs). We target an initial value problem (11.1.3.2) for a first-order ordinary differential equation

$$\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y}) \quad , \quad \mathbf{y}(t_0) = \mathbf{y}_0 \, . \tag{11.1.3.2}$$

As usual, the right hand side function $\mathbf{f} : D \subset \mathbb{R}^N \to \mathbb{R}^N$, $N \in \mathbb{N}$, may be given only in *procedural form*, for instance, in a C++ code as an functor object providing an evaluation operator

```
Eigen::VectorXd operator () (double t, const Eigen::VectorXd &y)
  const;
```

cf. Rem. 5.1.0.9. Occasionally the evaluation of $\mathbf{f}$ may involve costly computations.

§11.2.0.1 (Objectives of numerical integration) Two basic tasks can be identified in the field of numerical integration, which aims for the approximate solution of initial value problems for ODEs (Please distinguish from "numerical quadrature", see Chapter 7.):

(I) Given initial time t_0, final time T, and initial state $\mathbf{y}_0$ compute an approximation of $\mathbf{y}(T)$, where $t \mapsto \mathbf{y}(t)$ is the solution of (11.1.3.2). A corresponding function in C++ could look like

```
State solveivp(double t0, double T, State y0);
```

Here **State** is a type providing a fixed size or variable size vector $\in \mathbb{R}^N$, e.g.,

```
using State = Eigen::Matrix<double, N, 1>;
```

(II) Output an *approximate* solution $t \to \mathbf{y}_h(t)$ of (11.1.3.2) on $[t_0, T]$ up to final time $T \neq t_0$ for "all times" $t \in [t_0, T]$ (in practice, of course, only for finitely many times $t_0 < t_1 < t_2 < \cdots < t_{M-1} < t_M = T$, $M \in \mathbb{N}$, consecutively)

```
std::vector<State>
solveivp(State y0,const std::vector<double> &tvec);
```

This is the "plot solution" task, because we need to know $\mathbf{y}(t)$ for many times, if we want to create a faithful plot of $t \mapsto \mathbf{y}(t)$.

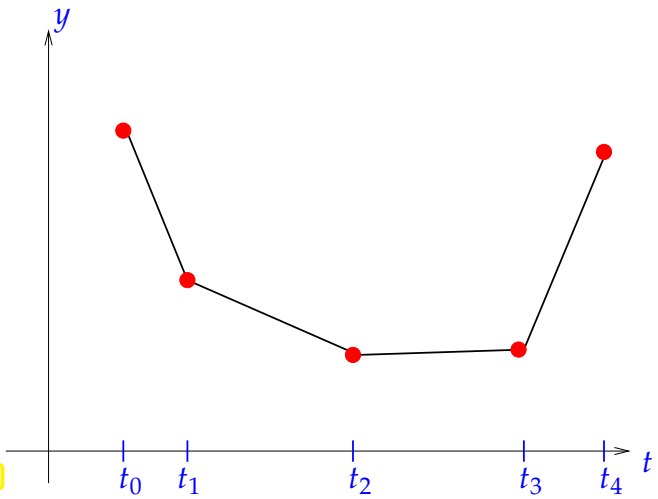


Fig. 412

This section presents three methods that provide a piecewise linear, that is, "polygonal" approximation of solution trajectories $t \mapsto \mathbf{y}(t)$.

◁ A piecewise linear function, aka a polygonal curve, approximating a function $t \mapsto y(t) \in \mathbb{R}$ in grid points $t_0 < t_1 < \cdots < t_4$.

§11.2.0.2 (Temporal mesh) As in Section 6.6.1 the polygonal approximation in this section will be based on a (temporal) mesh with $M+1$ mesh points (→ § 6.6.0.1)

$$\mathcal{M} := \{t_0 < t_1 < t_2 < \cdots < t_{M-1} < t_M := T\} \subset [t_0, T]\,, \tag{11.2.0.3}$$

covering the time interval of interest between initial time t_0 and final time $T > t_0$. We assume that the interval of interest is contained in the domain of definition of the solution of the IVP: $[t_0, T] \subset J(t_0, \mathbf{y}_0)$. ⌟

The next three sections will derive three simple mesh-based numerical integration methods, each in two ways:

(i) Based on geometric reasoning we interpret $\dot{\mathbf{y}}$ as the slope/direction of a tangent line.

(ii) In the spirit of numerical differentiation § 5.2.3.16, we replace the derivative $\dot{\mathbf{y}}$ with a mesh-based difference quotient.

11.2.1 Explicit Euler method

EXAMPLE 11.2.1.1 (Tangent field and solution curves) For $N = 1$ polygonal methods can be constructed by geometric considerations in the $t-y$ plane, a model for the extended state space. We explain this for the Riccati differential equation, a scalar ODE:

$$\dot{y} = f(t,y) := y^2 + t^2 \quad \blacktriangleright \quad N = 1, \quad I, D = \mathbb{R}^+\,. \tag{11.2.1.2}$$

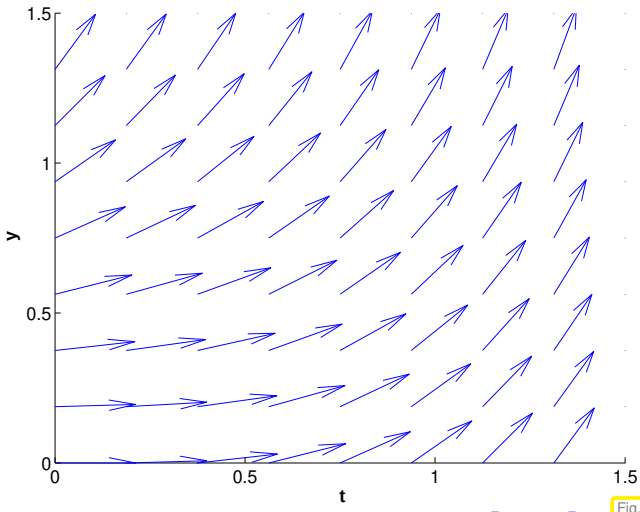


Fig. 413

tangent field $(t,y) \mapsto \frac{1}{\sqrt{f^2(t,y)+1}} \begin{bmatrix} 1 \\ f(t,y) \end{bmatrix}$

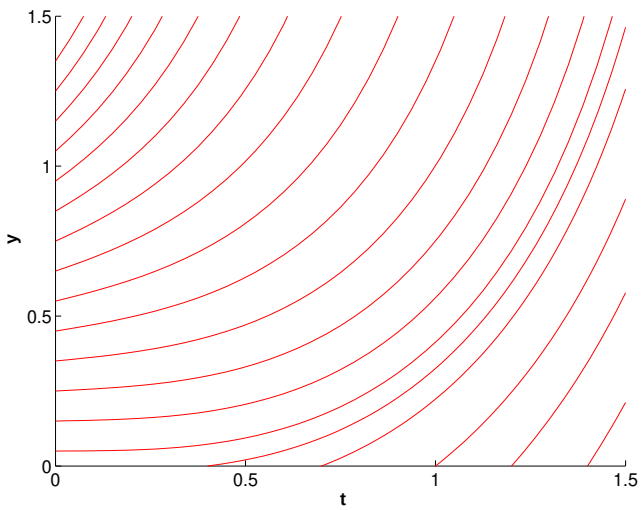
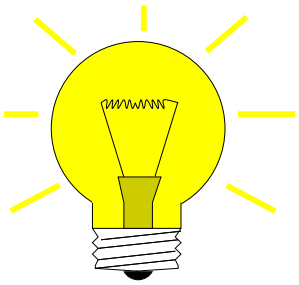


Fig. 414

solution curves

The solution curves run tangentially to the tangent field in each point of the extended state space. ⌟



Idea: "follow the tangents over short periods of time"

❶ timestepping: successive approximation of evolution on *mesh intervals* $[t_{k-1}, t_k]$, $k = 1, \ldots, M$, $t_M := T$,

❷ approximation of solution on $[t_{k-1}, t_k]$ by tangent line to solution trajectory through $(t_{k-1}, \mathbf{y}_{k-1})$.

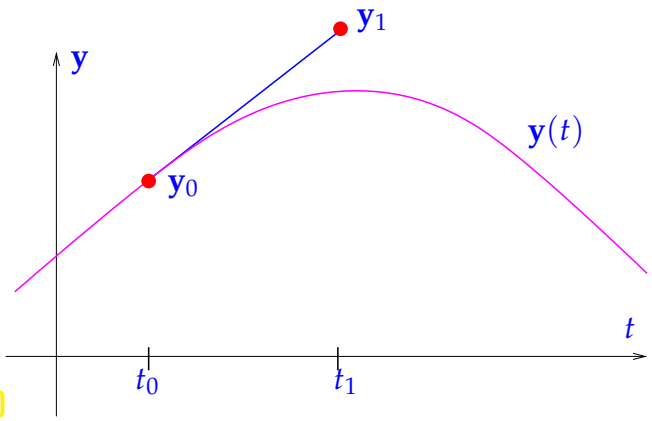


Fig. 415

explicit Euler method (Euler 1768)

◁ First step of explicit Euler method ($N = 1$):

Slope of tangent $= f(t_0, \mathbf{y}_0)$

$\mathbf{y}_1$ serves as initial value for next step!

See also [Han02, Ch. 74], [DR08, Alg. 11.4]

EXAMPLE 11.2.1.3 (Visualization of explicit Euler method)

We use the temporal mesh

$$\mathcal{M} := \{t_j := j/5 : j = 0, \ldots, 5\} ,$$

and solve an IVP for the Riccati differential equation, see Ex. 11.2.1.1

$$\dot{y} = y^2 + t^2 . \qquad (11.2.1.2)$$

Here: $y_0 = \frac{1}{2}$, $t_0 = 0$, $T = 1$, ▷

— ≙ "Euler polygon" for uniform timestep $h = 0.2$

↦ ≙ tangent field of Riccati ODE

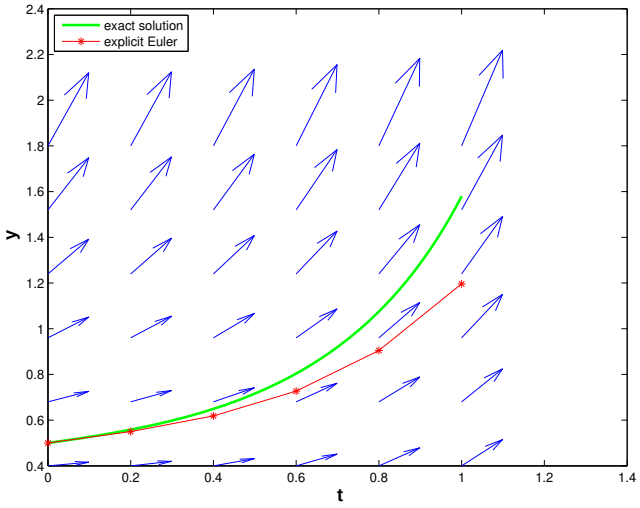


Fig. 416

⌟

§11.2.1.4 (Recursion for explicit Euler method) We translate the graphical construction of Fig. 415 into a formula. Given a temporal mesh $\mathcal{M} := \{t_0 < t_1 < t_2 < \cdots < t_{M-1} < t_M\}$ and applied to a general IVP

$$\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y}) \quad , \quad \mathbf{y}(t_0) = \mathbf{y}_0 . \qquad (11.1.3.2)$$

the explicit Euler method generates a *sequence* $(\mathbf{y}_k)_{k=0}^N$ of states by the recursion

$$\boxed{\mathbf{y}_{k+1} = \mathbf{y}_k + h_k \mathbf{f}(t_k, \mathbf{y}_k) , \quad k = 0, \ldots, M-1 \quad ,} \qquad (11.2.1.5)$$

with local (size of) timestep (stepsize) $h_k := t_{k+1} - t_k$.

The state $\mathbf{y}_k$ is supposed to approximate $\mathbf{y}(t_k)$, where $t \mapsto \mathbf{y}(t)$ is the exact solution of the IVP (11.1.3.2). ⌟

Remark 11.2.1.6 (Explicit Euler method as a difference scheme) One can obtain (11.2.1.5) by approximating the derivative $\frac{d}{dt}$ by a forward difference quotient on the (temporal) mesh $\mathcal{M} := \{t_0, t_1, \ldots, t_M\}$:

$$\dot{\mathbf{y}}(t_k) \approx \frac{\mathbf{y}(t_k + h_k) - \mathbf{y}(t_k)}{h_k}$$

$$\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y}) \quad \longleftrightarrow \quad \frac{\mathbf{y}_{k+1} - \mathbf{y}_k}{h_k} = \mathbf{f}(t_k, \mathbf{y}_k) \,, \quad k = 0, \ldots, M-1 \,. \tag{11.2.1.7}$$

Why a "forward difference quotient"? Because the difference quotient in (11.2.1.7) relies on the "future state" $y_{k+1} \approx \mathbf{y}(t_{k+1})$ to approximate $\dot{\mathbf{y}}(t_k)$.

In general, Difference schemes follow a simple policy for the *discretization* of differential equations: replace all derivatives by difference quotients connecting solution values on a set of discrete points (the mesh). ⌟

Remark 11.2.1.8 (Output of explicit Euler method) To begin with, the explicit Euler recursion (11.2.1.5) produces a sequence $\mathbf{y}_0, \ldots, \mathbf{y}_M$ of states. How does it deliver on the task (I) and (II) stated in § 11.2.0.1? By "geometric insight" we expect

$$\boxed{\mathbf{y}_k \approx \mathbf{y}(t_k)} \,.$$

(As usual, we use the notation $t \mapsto \mathbf{y}(t)$ for the exact solution of an IVP.)

No let us discuss to what extent the explicit Euler method delivers on the tasks formulated in § 11.2.0.1. Task (I): Easy, because $\mathbf{y}_M$ already provides an approximation of $\mathbf{y}(T)$.

Task (II): The trajectory $t \mapsto \mathbf{y}(t)$ is approximated by the piecewise linear function ('Euler polygon")

$$\mathbf{y}_h : [t_0, t_N] \to \mathbb{R}^N \quad , \quad \mathbf{y}_h(t) := \mathbf{y}_k \frac{t_{k+1} - t}{t_{k+1} - t_k} + \mathbf{y}_{k+1} \frac{t - t_k}{t_{k+1} - t_k} \quad \text{for} \quad t \in [t_k, t_{k+1}] \,, \tag{11.2.1.9}$$

see Fig. 416. This function can easily be sampled on any grid of $[t_0, t_M]$. In fact, it is the $\mathcal{M}$-piecewise linear interpolant of the data points $(t_k, \mathbf{y}_k)$, $k = 0, \ldots, N$, see Section 5.3.2).

The same considerations apply to the methods discussed in the next two sections and will not be repeated there. ⌟

11.2.2 Implicit Euler method

Recall the discussions of Rem. 11.2.1.6. Why should be use a forward difference quotient and not a backward difference quotient, which relies on "states in the past". Let's try!

On (temporal) mesh $\mathcal{M} := \{t_0 < t_1 < \cdots < t_M\}$ we obtain

$$\dot{\mathbf{y}} = f(t, \mathbf{y}) \quad \longleftrightarrow \quad \frac{\mathbf{y}_{k+1} - \mathbf{y}_k}{h_k} = f(t_{k+1}, \mathbf{y}_{k+1}) \,, \quad k = 0, \ldots, M-1 \,. \tag{11.2.2.1}$$

backward difference quotient

This leads to another simple timestepping scheme analoguous to (11.2.1.5):

$$\boxed{\mathbf{y}_{k+1} = \mathbf{y}_k + h_k \mathbf{f}(t_{k+1}, \mathbf{y}_{k+1}) \ , \quad k = 0, \ldots, M-1} \ , \tag{11.2.2.2}$$

with local timestep (stepsize) $h_k := t_{k+1} - t_k$.

(11.2.2.2) = implicit Euler method

Note: (11.2.2.2) requires solving a (possibly non-linear) system of equations to obtain $\mathbf{y}_{k+1}$!
(➤ Terminology "implicit")

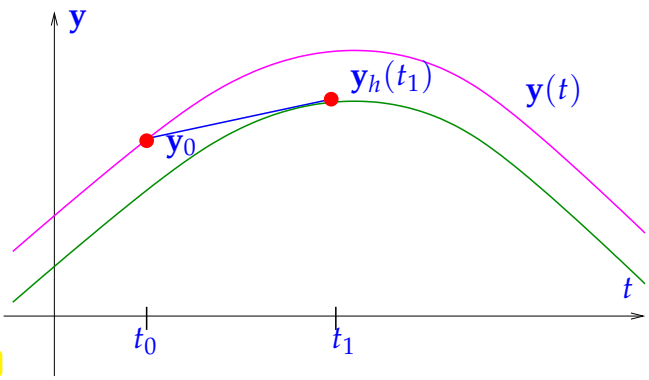


Fig. 417

Geometry of implicit Euler method:

Approximate solution through (t_0, y_0) on $[t_0, t_1]$ by
- straight line through (t_0, y_0)
- with slope $f(t_1, y_1)$

◁ — ≙ trajectory through (t_0, y_0),
— ≙ trajectory through (t_1, y_1),
— ≙ tangent at — in (t_1, y_1).

Remark 11.2.2.3 (Feasibility of implicit Euler timestepping) The issue is whether (11.2.2.2) well defined, that is, whether we can solve it for $\mathbf{y}_{k+1}$ and whether this solution unique. The intuition is that for *small timestep size* $h > 0$ the right hand side of (11.2.2.2) is a "small perturbation of the identity".

Let us give a formal argument. Consider an autonomous ODE $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$, assume a continuously differentiable right hand side function $\mathbf{f}$, $\mathbf{f} \in C^1(D, \mathbb{R}^N)$, and regard (11.2.2.2) as an h-dependent non-linear system of equations:

$$\mathbf{y}_{k+1} = \mathbf{y}_k + h_k \mathbf{f}(t_{k+1}, \mathbf{y}_{k+1}) \quad \Leftrightarrow \quad G(h, \mathbf{y}_{k+1}) = 0 \quad \text{with} \quad G(h, \mathbf{z}) := \mathbf{z} - h\mathbf{f}(t_{k+1}, \mathbf{z}) - \mathbf{y}_k \ .$$

To investigate the solvability of this non-linear equation we start with an observation about a partial derivative of G:

$$\frac{dG}{d\mathbf{z}}(h, \mathbf{z}) = \mathbf{I} - h \, \mathsf{D}_{\mathbf{y}} \mathbf{f}(t_{k+1}, \mathbf{z}) \quad \Rightarrow \quad \frac{dG}{d\mathbf{z}}(0, \mathbf{z}) = \mathbf{I} \ .$$

In addition, $G(0, y_k) = \mathbf{0}$. Next, recall the implicit function theorem [Str09, Thm. 7.8.1]:

Theorem 11.2.2.4. Implicit function theorem

Let $G = G(\mathbf{x}, \mathbf{y})$ a continuously differentiable function of $\mathbf{x} \in \mathbb{R}^k$ and $\mathbf{y} \in \mathbb{R}^\ell$, defined on the open set $\Omega \subset \mathbb{R}^k \times \mathbb{R}^\ell$ with values in $\mathbb{R}^\ell$: $G : \Omega \subset \mathbb{R}^k \times \mathbb{R}^\ell \to \mathbb{R}^\ell$.

Assume that G has a zero in $\mathbf{z}_0 := \begin{bmatrix} \mathbf{x}_0 \\ \mathbf{y}_0 \end{bmatrix} \in \Omega$, $\mathbf{x}_0 \in \mathbb{R}^k$, $\mathbf{y}_0 \in \mathbb{R}^\ell$: $G(\mathbf{z}_0) = 0$.

If the Jacobian $\frac{\partial G}{\partial \mathbf{y}}(\mathbf{p}_0) \in \mathbb{R}^{\ell,\ell}$ is invertible, then there is an open neighborhood U of $\mathbf{x}_0 \in \mathbb{R}^k$ and a continuously differentiable function $\mathbf{g} : U \to \mathbb{R}^l$ such that

$$\mathbf{g}(\mathbf{x}_0) = \mathbf{y}_0 \quad \textit{and} \quad G(\mathbf{x}, \mathbf{g}(\mathbf{x})) = 0 \quad \forall \mathbf{x} \in U \ .$$

For *sufficiently small* $|h|$ it permits us to conclude that the equation $G(h,\mathbf{z}) = 0$ defines a continuous function $\mathbf{g} = \mathbf{g}(h)$ with $\mathbf{g}(0) = \mathbf{y}_k$.

Corollary 11.2.2.5. Solvability of implicit Euler recursion

For sufficiently small $h > 0$ *the equation* (11.2.2.2) *has a unique solution* $\mathbf{y}_{k+1}$.

⌟

11.2.3 Implicit midpoint method

Beside using forward or backward difference quotients, the derivative $\dot{\mathbf{y}}$ can also be approximated by the symmetric difference quotient, see also (5.2.3.17),

$$\dot{\mathbf{y}}(t) \approx \frac{\mathbf{y}(t+h) - \mathbf{y}(t-h)}{2h} , \quad h > 0 . \tag{11.2.3.1}$$

The idea is to apply this formula in $t = \frac{1}{2}(t_k + t_{k+1})$ with $h = h_k/2$, which transforms the ODE into

$$\dot{\mathbf{y}} = f(t,\mathbf{y}) \quad \longleftrightarrow \quad \frac{\mathbf{y}_{k+1} - \mathbf{y}_k}{h_k} = f\big(\tfrac{1}{2}(t_k + t_{k+1}), \mathbf{y}_h(\tfrac{1}{2}(t_k + t_{k+1}))\big) , \quad k = 0,\ldots,M-1 . \tag{11.2.3.2}$$

The trouble is that the value $\mathbf{y}_h(\frac{1}{2}(t_{k+1} + t_{k+1}))$ does not seem to be available, unless we recall that the approximate trajectory $t \mapsto \mathbf{y}_h(t)$ is supposed to be piecewise linear, which implies $\mathbf{y}_h(\frac{1}{2}(t_{k+1} + t_{k+1})) = \frac{1}{2}(\mathbf{y}_h(t_k) + \mathbf{y}_h(t_{k+1}))$. This gives the recursion formula for the implicit midpoint method in analogy to (11.2.1.5) and (11.2.2.2):

$$\boxed{\mathbf{y}_{k+1} = \mathbf{y}_k + h_k \mathbf{f}\big(\tfrac{1}{2}(t_k + t_{k+1}), \tfrac{1}{2}(\mathbf{y}_k + \mathbf{y}_{k+1})\big) , \quad k = 0,\ldots,N-1} , \tag{11.2.3.3}$$

with local timestep (stepsize) $h_k := t_{k+1} - t_k$.

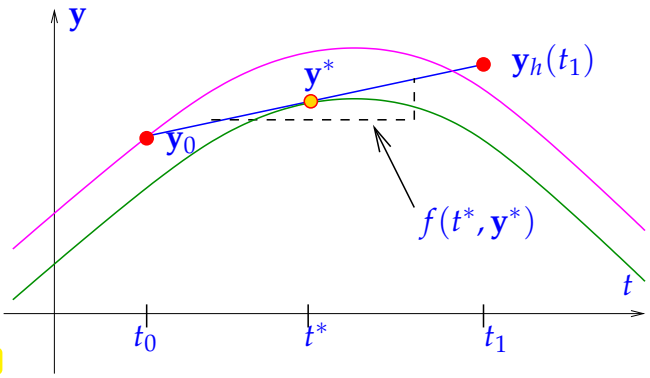


Fig. 418

Implicit midpoint method, a geometric view:

Approximaate trajectory through $(t_0,\mathbf{y}_0)$ on $[t_0,t_1]$ by

- straight line through $(t_0,\mathbf{y}_0)$
- with slope $f(t^*,\mathbf{y}^*)$, where $t^* := \frac{1}{2}(t_0+t_1)$, $\mathbf{y}^* = \frac{1}{2}(\mathbf{y}_0+\mathbf{y}_1)$

◁ — $\hat{=}$ trajectory through (t_0,y_0),
— $\hat{=}$ trajectory through (t^*,y^*),
— $\hat{=}$ tangent at — in (t^*,y^*).

As in the case of (11.2.2.2), also (11.2.3.3) entails solving a (non-linear) system of equations in order to obtain $\mathbf{y}_{k+1}$. Rem. 11.2.2.3 also holds true in this case: for sufficiently small h (11.2.3.3) will have a unique solution $\mathbf{y}_{k+1}$, which renders the recursion well defined.

Review question(s) 11.2.3.4 (Polygonal approximation methods)

(Q11.2.3.4.A) We consider the scalar linear IVP

$$\dot{y} = \lambda y \quad , \quad y(0) = 1$$

on the interval $[0,1]$. We use $M \in \mathbb{N}$ equidistant steps of the explicit Euler method to compute an approximation y_M for $y(1)$.

- Derive a formula for y_M.
- Which known result from calculus is equivalent to the convergence $y_M \to y(1)$ for $M \to \infty$?

(Q11.2.3.4.B) For an ODE

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}) \quad , \quad \mathbf{f} := \begin{bmatrix} f_1(\mathbf{y}) \\ \vdots \\ f_N(\mathbf{y}) \end{bmatrix} : D \subset \mathbb{R}^N \mapsto \mathbb{R}^N \, , \tag{*}$$

we know that

$$\sum_{\ell=1}^{N} f_\ell(\mathbf{y}) = 0 \quad \forall \mathbf{y} \in D \, .$$

- Show that the sum of the components of every solution $t \mapsto \mathbf{y}(t)$ is constant in time.
- Show that the sums of the components of the vectors $\mathbf{y}_0, \mathbf{y}_1, \mathbf{y}_2, \ldots$ generated by either the explicit Euler method, the implicit Euler method, or the implicit midpoint method, all applied to solve some IVP for (*), are the same for all vectors $\mathbf{y}_k$.

(Q11.2.3.4.C) We consider the implicit Euler method for the scalar autonomous "explosion ODE" $\dot{y} = y^2$. Given an *explicit* formula for y_{k+1} in terms of y_k and the timestep size $h_k > 0$. Specify potentially necessary constraints on the size of h_k.

The defining equation for recursion of the implicit Euler method (on some temporal mesh) applied to the ODE $\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y})$ is

$$\mathbf{y}_{k+1}: \quad \mathbf{y}_{k+1} = \mathbf{y}_k + h_k \mathbf{f}(t_{k+1}, \mathbf{y}_{k+1}) \, . \tag{11.2.2.2}$$

(Q11.2.3.4.D) The recursion of the implicit midpoint rule for the ODE $\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y})$ is

$$\mathbf{y}_{k+1}: \quad \mathbf{y}_{k+1} = \mathbf{y}_k + h_k \mathbf{f}(\frac{1}{2}(t_k + t_{k+1}), \tfrac{1}{2}(\mathbf{y}_k + \mathbf{y}_{k+1})) \, .$$

Give an *explicit* form of this recursion for the linear ODE $\dot{\mathbf{y}} = \mathbf{A}(t)\mathbf{y}$, where $\mathbf{A} : \mathbb{R} \to \mathbb{R}^{N,N}$ is a matrix-valued function. When will this recursion break down?

(Q11.2.3.4.E) For a twice continuously differentiable function $f : I \subset \mathbb{R} \to \mathbb{R}^N$ we can use the second symmetric difference quotient as an approximation of the second derivative $f''(x)$, $x \in I$:

$$\frac{f(x+h) - 2f(x) + f(x-h)}{h^2} \approx f''(x) \quad \text{for} \quad |h| \ll 1 \, .$$

Based on this approximation propose an explicit finite-difference timstepping scheme on a uniform temporal mesh for the second-order ODE $\ddot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$.

(Q11.2.3.4.F) Formulate the equation that defines that single-step method for the IVP $\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y})$, $\mathbf{y}(t_0) = \mathbf{y}_0$, that arises from the difference quotient approximation

$$\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y}) \quad \to \quad \frac{\mathbf{y}_{k+1} - \mathbf{y}_k}{h_k} \approx \tfrac{1}{2}(f(t_k, \mathbf{y}(t_k)) + f(t_{k+1}, \mathbf{y}(t_{k+1}))) \, , \quad h_k := t_{k+1} - t_k \, .$$

A temporal mesh $\mathcal{M} := \{t_0 < t_1 < t_2 < \cdots < t_{M-1} < t_M := T\}$ can be taken for granted.

△

11.3 General Single-Step Methods

Video tutorial for Section 11.3: General Single-Step Methods: (14 minutes) Download link, tablet notes

Now we fit the numerical schemes introduced in the previous section into a more general class of methods for the solution of (autonomous) initial value problems (11.1.3.18) for ODEs. Throughout we assume that all times considered belong to the domain of definition of the unique solution $t \to \mathbf{y}(t)$ of (11.1.3.18), that is, for $T > 0$ we take for granted $[0,T] \subset J(\mathbf{y}_0)$ (temporal domain of definition of the solution of an IVP is explained in § 11.1.3.19).

11.3.1 Definition

§11.3.1.1 (Discrete evolution operators) From Section 11.2.1 and Section 11.2.2 recall the two Euler methods for an autonomous ODE $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$:

$$\begin{aligned} &\text{explicit Euler:} \quad \mathbf{y}_{k+1} = \mathbf{y}_k + h_k \mathbf{f}(\mathbf{y}_k) \,, \\ &\text{implicit Euler:} \quad \mathbf{y}_{k+1}\text{:} \quad \mathbf{y}_{k+1} = \mathbf{y}_k + h_k \mathbf{f}(\mathbf{y}_{k+1}) \,, \end{aligned} \qquad h_k := t_{k+1} - t_k \,.$$

Both formulas, for sufficiently small h_k ($\to$ Rem. 11.2.2.3), provide a mapping

$$(\mathbf{y}_k, h_k) \mapsto \boldsymbol{\Psi}(h, \mathbf{y}_k) := \mathbf{y}_{k+1} \,. \tag{11.3.1.2}$$

If $\mathbf{y}_0$ is the initial value, then $\mathbf{y}_1 := \boldsymbol{\Psi}(h, \mathbf{y}_0)$ can be regarded as an approximation of $\mathbf{y}(h)$, the value returned by the evolution operator $\boldsymbol{\Phi}$ ($\to$ Def. 11.1.4.3) for $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$ applied to $\mathbf{y}_0$ over the period h. $\mathbf{y}(t_k)$:

$$\mathbf{y}_1 = \boldsymbol{\Psi}(h, \mathbf{y}_0) \quad \longleftrightarrow \quad \mathbf{y}(h) = \boldsymbol{\Phi}^h \mathbf{y}_0 \quad \succ \quad \boxed{\boldsymbol{\Psi}(h, \mathbf{y}) \approx \boldsymbol{\Phi}^h \mathbf{y}} \,, \tag{11.3.1.3}$$

In a sense the polygonal approximation methods are based on approximations for the evolution operator associated with the ODE.

This is what every single step method does: it tries to approximate the evolution operator $\boldsymbol{\Phi}$ for an ODE by a mapping $\boldsymbol{\Psi}$ of the kind as described in (11.3.1.2).

➙ A mapping $\boldsymbol{\Psi}$ as in (11.3.1.2) is called (a) **discrete evolution** operator.

✎ Notation: In analogy to $\boldsymbol{\Phi}^h$ for discrete evolutions we often write $\boldsymbol{\Psi}^h \mathbf{y} := \boldsymbol{\Psi}(h, \mathbf{y})$ ⌟

Remark 11.3.1.4 (Discretization) The adjective "discrete" used above designates (components of) methods that attempt to approximate the solution of an IVP by a sequence of finitely many states. "Discretization" is the process of converting an ODE into a discrete model. This parlance is adopted for all procedures that reduce a "continuous model" involving ordinary or partial differential equations to a form with a finite number of unknowns. ⌟

Above we identified the discrete evolutions underlying the polygonal approximation methods. Vice versa,

a mapping Ψ as given in (11.3.1.2) defines a single step method.

Definition 11.3.1.5. Single step method (for autonomous ODE) → [QSS00, Def. 11.2]

Given a discrete evolution $\Psi : \Omega \subset \mathbb{R} \times D \mapsto \mathbb{R}^N$, an initial state $\mathbf{y}_0$, and a temporal mesh $\mathcal{M} := \{0 =: t_0 < t_1 < \cdots < t_M := T\}$, $M \in \mathbb{N}$, the recursion

$$\mathbf{y}_{k+1} := \Psi(t_{k+1} - t_k, \mathbf{y}_k) , \quad k = 0, \ldots, M-1 , \tag{11.3.1.6}$$

defines a **single-step method** (SSM) for the autonomous IVP $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$, $\mathbf{y}(0) = \mathbf{y}_0$ on the interval $[0, T]$.

☞ In a sense, a single step method defined through its associated discrete evolution does not approximate a concrete initial value problem, but tries to approximate an ODE in the form of its evolution operator.

In C++ a discrete evolution operator can be incarnated by a functor type offering an evaluation operator

```
State operator ()(double h, const State &y) const;
```

see § 11.2.0.1 for the **State** data type.

Remark 11.3.1.7 (Discrete evolutions for non-autonomous ODEs) The concept of single step method according to Def. 11.3.1.5 can be generalized to non-autonomous ODEs, which leads to recursions of the form:

$$\mathbf{y}_{k+1} := \Psi(t_k, t_{k+1}, \mathbf{y}_k) , \quad k = 0, \ldots, M-1 ,$$

for a discrete evolution operator Ψ defined on $I \times I \times D$. ⌟

§11.3.1.8 (Consistent single step methods) Now we state a first quantification of the goal that the "discrete evolution should be an approximation of the evolution operator": $\Psi \approx \Phi$, *cf.* (11.3.1.3). We want the discrete evolution Ψ to inherit key properties of the evolution operator Φ. One such property is

$$\frac{d}{dt}\Phi^t\mathbf{y}\Big|_{t=0} = \mathbf{f}(\mathbf{y}) \quad \forall \mathbf{y} \in D . \tag{11.3.1.9}$$

Compliance of Ψ with (11.3.1.9) is expressed through the property of consistency, which, roughly speaking, demands that a viable discrete evolution operator methods is structurally similar to that for the explicit Euler method (11.2.1.5).

Consistent discrete evolution

The discrete evolution Ψ defining a single step method according to Def. 11.3.1.5 and (11.3.1.6) for the autonomous ODE $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$ must be of the form

$$\Psi^h\mathbf{y} = \mathbf{y} + h\psi(h, \mathbf{y}) \quad \text{with} \quad \begin{array}{l} \psi : I \times D \to \mathbb{R}^N \text{ continuous,} \\ \psi(0, \mathbf{y}) = \mathbf{f}(\mathbf{y}) . \end{array} \tag{11.3.1.11}$$

Differentiating $\left(h \mapsto \Psi^h\mathbf{y}\right)$ relying on the product rule confirms that (11.3.1.9) remains true for Ψ instead of Φ.

Definition 11.3.1.12. Consistent single step methods

A single step method according to Def. 11.3.1.5 based on a discrete evolution of the form (11.3.1.11) is called **consistent** with the ODE $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$.

⌟

EXAMPLE 11.3.1.13 (Consistency of implicit midpoint method) The discrete evolution Ψ and, hence, the function $\psi = \psi(h, \mathbf{y})$ for the implicit midpoint method are defined only implicitly, of course. Thus, consistency cannot immediately be seen from a formula for ψ.

We examine consistency of the implicit midpoint method for the autonomous ODE $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$. The corresponding discrete evolution Ψ is definied by:

$$\Psi^h \mathbf{y} = \mathbf{y} + h\mathbf{f}\left(\tfrac{1}{2}(\mathbf{y} + \Psi^h \mathbf{y})\right) , \quad h \in \mathbb{R},\ |h| \text{ "sufficiently small"}, \quad \mathbf{y} \in D \,. \tag{11.3.1.14}$$

Assume that

- the right hand side function $\mathbf{f} : D \subset \mathbb{R}^N \to \mathbb{R}^N$ is locally Lipschitz continuous, $\mathbf{f} \in C^0(D)$,
- and that $|h|$ is "sufficiently small" to guarantee the existence of a solution $\Phi^h \mathbf{y}$ of (11.3.1.14) as explained in Rem. 11.2.2.3.

Then we infer from the implicit function theorem Thm. 11.2.2.4 that the solution $\Psi^h \mathbf{y}$ of (11.3.1.14) will continuously depend on h: $h \mapsto \Psi^h \mathbf{y} \in C^0(]-\delta, \delta[, \mathbb{R}^N)$ for small $\delta > 0$. Knowing this, we plug (11.3.1.14) into itself and obtain

$$\Psi^h \mathbf{y} = \mathbf{y} + h\mathbf{f}(\tfrac{1}{2}(\mathbf{y} + \Psi^h \mathbf{y})) \overset{(11.3.1.14)}{=} \mathbf{y} + h \underbrace{\mathbf{f}(\mathbf{y} + \tfrac{1}{2}h\mathbf{f}(\tfrac{1}{2}(\mathbf{y} + \Psi^h \mathbf{y})))}_{=:\psi(h,\mathbf{y})} .$$

We repeat that by the implicit function theorem Thm. 11.2.2.4 $\Psi^{hj}\mathbf{y}$ depends continuously on h and $\mathbf{y}$. This means that $\psi(h, \mathbf{y})$ has the desired properties, in particular $\psi(0, \mathbf{y}) = \mathbf{f}(\mathbf{y})$ is clear. ⌟

Remark 11.3.1.15 (Notation for single step methods) Many authors specify a single step method by writing down the first step for a general stepsize h

$$\mathbf{y}_1 = \text{(implicit) expression in } \mathbf{y}_0, h \text{ and } \mathbf{f} \,,$$

for instance, for the implicit midpoint rule

$$\mathbf{y}_1 = \mathbf{y}_0 + h\mathbf{f}(\tfrac{1}{2}(\mathbf{y}_0 + \mathbf{y}_1)) \,.$$

Actually, this fixes the underlying discrete evolution. Also this course will sometimes adopt this practice. ⌟

§11.3.1.16 (Output of single step methods) Here we resume and continue the discussion of Rem. 11.2.1.8 for general single step methods according to Def. 11.3.1.5. Assuming unique solvability of the systems of equations faced in each step of an implicit method, every single step method based on a mesh $\mathcal{M} = \{0 = t_0 < t_1 < \cdots < t_M := T\}$ produces a finite sequence $(\mathbf{y}_0, \mathbf{y}_1, \ldots, \mathbf{y}_M)$ of states, where the first agrees with the initial state $\mathbf{y}_0$.

We expect that the states provide a pointwise approximation of the solution trajectory $t \to \mathbf{y}(t)$:

$$\mathbf{y}_k \approx \mathbf{y}(t_k) \,, \quad k = 1, \ldots, M \,.$$

Thus task (I) from § 11.2.0.1, computing an approximation for $\mathbf{y}(T)$, is again easy: output $\mathbf{y}_M$ as an approximation of $\mathbf{y}(T)$.

Task (II) from § 11.2.0.1, computing the solution trajectory, requires interpolation of the data points $(t_k, \mathbf{y}_k)$ using some of the techniques presented in Chapter 5. The natural option is $\mathcal{M}$-piecewise polynomial interpolation, generalizing the polygonal approximation (11.2.1.9) used in Section 11.2.

Note that from the ODE $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$ the derivatives $\dot{\mathbf{y}}_h(t_k) = \mathbf{f}(\mathbf{y}_k)$ are available without any further approximation. This facilitates cubic Hermite interpolation (→ Def. 5.3.3.1), which yields

$$\mathbf{y}_h \in C^1([0,T]): \quad \mathbf{y}_{h\,|[x_{k-1},x_k]} \in \mathcal{P}_3 \,, \quad \mathbf{y}_h(t_k) = \mathbf{y}_k \,, \quad \frac{d\mathbf{y}_h}{dt}(t_k) = \mathbf{f}(\mathbf{y}_k) \,.$$

Summing up, an approximate trajectory $t \mapsto \mathbf{y}_h(t)$ is built in two stages:

(i) Compute sequence $(\mathbf{y}_k)_k$ by running the single step method.

(ii) Post-process the obtained sequence, usually by applying interpolation, to get $\mathbf{y}_h$.

⌟

Review question(s) 11.3.1.17 (General single-step methods)

(Q11.3.1.17.A) Explain the concepts

- **evolution operator** and
- **discrete evolution operator**

in connection with the numerical integration of initial-value problems for the ODE $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$, $\mathbf{f} : D \subset \mathbb{R}^N \mapsto \mathbb{R}^N$.

(Q11.3.1.17.B) [Single-step methods and numerical quadrature] There is a connection between numerical integration (the design and analysis of numerical methods for the solution of initial-value problems for ODEs) and numerical quadrature (study of numerical methods for the evaluation of integrals).

- Explain, how a class of single-step methods for the solution of scalar initial-value problems

$$\dot{y} = f(t,y) \quad , \quad y(t_0) = y_0 \in \mathbb{R} \,,$$

  can be used for the approximate evaluation of integrals $\int_a^b \varphi(\tau)\,\mathrm{d}\tau$, $\varphi : [a,b] \to \mathbb{R}$.
- If the considered single-step methods are of order p, what does this mean for the induced quadrature method.
- Which quadrature formula does the implicit midpoint method yield?

(Q11.3.1.17.C) [Adjoint single-step method] Let a single-step method for the autonomous ODE $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$, $\mathbf{f} : D \subset \mathbb{R}^N \to \mathbb{R}^N$ be defined by its discrete evolution operator $\boldsymbol{\Psi} : I \times D \mapsto D$. Then the **adjoint single-step method** is spawned by the discrete evolution operator $\widetilde{\boldsymbol{\Psi}} : I \times D \mapsto D$ defined according to

$$\widetilde{\boldsymbol{\Psi}}^h \mathbf{y} := \left(\boldsymbol{\Psi}^{-h}\right)^{-1} \,, \quad \mathbf{y} \in D,\ h \in \mathbb{R} \quad \text{sufficiently small} \,.$$

What is the adjoint of the explicit Euler method?

(Q11.3.1.17.D) We have seen three simple single-step methods for the autonomous ODE $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$, $\mathbf{f} : D \subset \mathbb{R}^N \to \mathbb{R}^N$, here defined by describing the first step $\mathbf{y}_0 \to \mathbf{y}_1$ with stepsize $h \in \mathbb{R}$ ("sufficiently small"):

- The explicit Euler method:

$$\mathbf{y}_1 = \mathbf{y}_0 + h\mathbf{f}(\mathbf{y}_0) \ .$$

- The implicit Euler method:

$$\mathbf{y}_1 : \quad \mathbf{y}_1 = \mathbf{y}_0 + h\mathbf{f}(\mathbf{y}_1) \ .$$

- The implicit midpoint method:

$$\mathbf{y}_1 : \quad \mathbf{y}_1 = \mathbf{y}_0 + h\mathbf{f}(\tfrac{1}{2}(\mathbf{y}_0 + \mathbf{y}_1)) \ .$$

For which methods does the associated discrete evolution operator $\Psi : [-\delta, \delta] \times D \to D$, $\delta > 0$ sufficiently small, satisfy

$$\Psi^h \circ \Psi^{-h} = \mathrm{Id} \quad \forall h \in [0, \delta] \ ? \tag{11.3.1.18}$$

Try to find a simple (scalar) counterexample, if you think that a method does not have property (11.3.1.18).

△

11.3.2 (Asymptotic) Convergence of Single-Step Methods

Video tutorial for Section 11.3.2:(Asymptotic) Convergence of Single-Step Methods: (20 minutes) Download link, tablet notes

Of course, the accuracy of the solution sequence $(\mathbf{y}_k)_k$ obtained by a particular single-step method ($\to$ Def. 11.3.1.5) is a central concern. This motivates studying the dependence of suitable norms of the so-called discretization error on the choice of the temporal mesh $\mathcal{M} := \{0 = t_0 < t_1 < \cdots < t_M = T\}$.

§11.3.2.1 (Discretization error of single step methods) Approximation errors in numerical integration are also called discretization errors, *cf.* Rem. 11.3.1.4.

Depending on the objective of numerical integration as stated in § 11.2.0.1 different (norms of) discretization errors are of interest:

(I) If only the solution at final time T is sought, the relevant norm of the discretization error is

$$\epsilon_M := \|\mathbf{y}(T) - \mathbf{y}_M\| \ ,$$

where $\|\cdot\|$ is some vector norm on $\mathbb{R}^N$.

(II) If we want to approximate the solution trajectory for (11.1.3.18) the discretization error is the function

$$t \mapsto \mathbf{e}(t) \quad , \quad \mathbf{e}(t) := \mathbf{y}(t) - \mathbf{y}_h(t) \ ,$$

where $t \mapsto \mathbf{y}_h(t)$ is the approximate trajectory obtained by post-processing, see § 11.3.1.16. In this case accuracy of the method is gauged by looking at norms of the function $\mathbf{e}$, see § 5.2.4.4 for examples.

(III) Between (I) and (II) is the pointwise discretization error, which is the sequence (a so-called grid function)

$$\mathbf{e} : \mathcal{M} \to D \quad , \quad \mathbf{e}_k := \mathbf{y}(t_k) - \mathbf{y}_k \, , \quad k = 0, \ldots, M \, . \tag{11.3.2.2}$$

In this case one usually examines the maximum error in the mesh points

$$\|(\mathbf{e})\|_\infty := \max_{k \in \{1,\ldots,N\}} \|\mathbf{e}_k\| \, ,$$

where $\|\cdot\|$ is a suitable vector norm on $\mathbb{R}^N$, customarily the Euclidean vector norm.

⌟

§11.3.2.3 (Asymptotic convergence of single step methods) Once the discrete evolution Ψ associated with the ODE $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$ is specified, the single step method according to Def. 11.3.1.5 is fixed:

$$\mathbf{y}_{k+1} := \Psi(t_{k+1} - t_k, \mathbf{y}_k) \, , \quad k = 0, \ldots, M-1 \, , \tag{11.3.1.6}$$

The only way to control the accuracy of the solution $\mathbf{y}_N$ or $t \mapsto \mathbf{y}_h(t)$ is through the selection of the mesh $\mathcal{M} = \{0 = t_0 < t_1 < \cdots < t_N = T\}$.

Hence we study convergence of single step methods for families of meshes $\{\mathcal{M}_\ell\}$ and track the decay of (a norm) of the discretization error ($\to$ § 11.3.2.1) as a function of the number $M := \sharp\mathcal{M}$ of mesh points. In other words, we examine **h-convergence**. Convergence through mesh refinement is discussed for piecewise polynomial interpolation in Section 6.6.1 and for composite numerical quadrature in Section 7.5.

When investigating asymptotic convergence of single step methods we often resort to families of equidistant meshes of $[0, T]$:

$$\mathcal{M}_M := \{t_k := \frac{k}{M} T : k = 0 \ldots, M\} \, . \tag{11.3.2.4}$$

We also call this the use of uniform timesteps of size $h := \frac{T}{M}$. ⌟

EXPERIMENT 11.3.2.5 (Speed of convergence of polygonal methods)

The setting for this experiment is a follows:

- We consider the following IVP for the logistic ODE, see Ex. 11.1.2.1

$$\dot{y} = \lambda y(1-y) \quad , \quad y(0) = 0.01 \, .$$

- We apply explicit and implicit Euler methods (11.2.1.5)/(11.2.2.2) with uniform timestep $h = 1/M$, $M \in \{5, 10, 20, 40, 80, 160, 320, 640\}$.
- Monitored: Error at final time $E(h) := |y(1) - y_M|$

We are mainly interested in the qualitative nature of the asymptotic convergence as $h \to 0$ in the sense of the types of convergence introduced in Def. 6.2.2.7 with N there replaced with h^{-1}. Abbreviating some error norm with $EN = EN(h)$, recall the classification of asymptotic convergence from Def. 6.2.2.7:

$$\begin{array}{lll} \exists\, p > 0: & EN(h) \le h^p & : \textbf{algebraic convergence}, \text{ with order/rate } p > 0 \, , \\ \exists\, 0 < q < 1: & EN(h) \le q^{1/h} & : \textbf{exponential convergence} \, , \end{array} \quad \forall h > 0; \, .$$

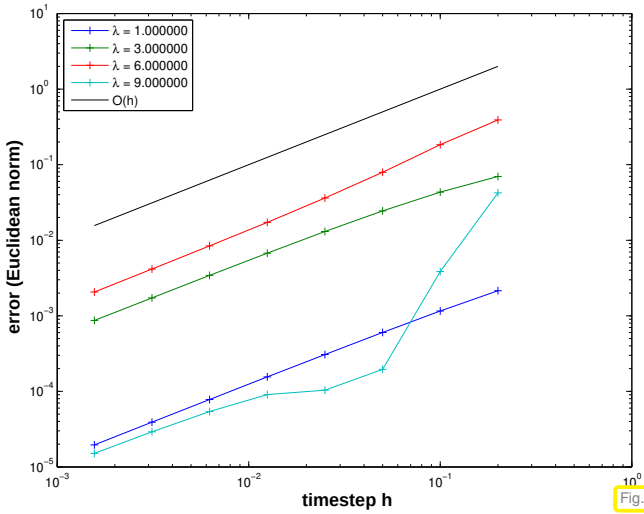


Fig. 419

explicit Euler method

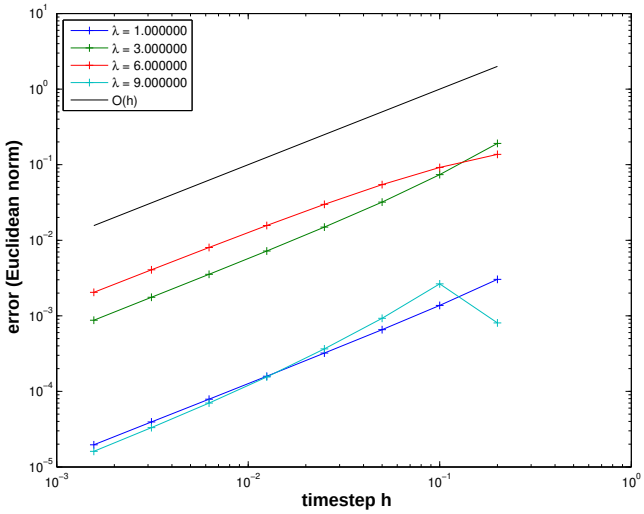
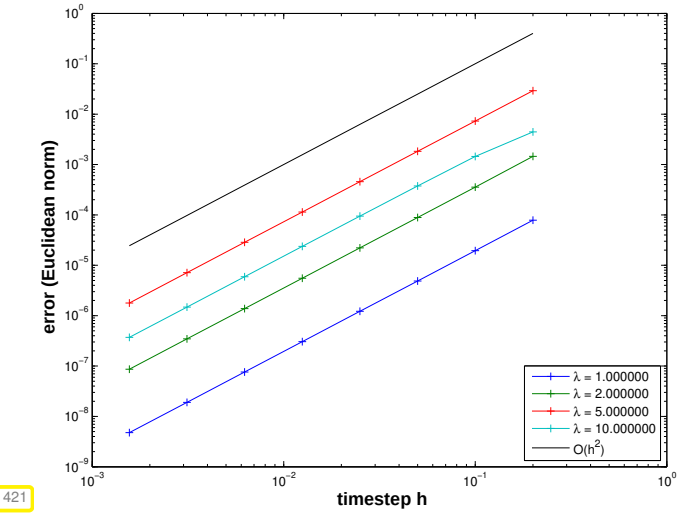


Fig. 420

implicit Euler method



$O(M^{-1}) = O(h)$ algebraic convergence with order/rate 1 in both cases for $h \to 0$

This matches our expectations, because, as we see from (11.2.1.7) and (11.2.2.1), both Euler methods can be introduced via an approximation of $\dot{y}$ by a *one-sided difference quotient*, which offers an $O(h)$ approximation of the derivative as $h \to 0$.

Fig. 421

implicit midpoint method

However, polygonal approximation methods can do better:

◁ We study the convergence of the implicit midpoint method (11.2.3.3) in the above setting.

We observe algebraic convergence $O(h^2)$, that is with order/rate 2 for $h \to 0$.

Also this is expected, because symmetric difference quotients of width h offer $O(h^2)$-approximation of the derivative for $h \to 0$.

Parlance: Based on the observed rate of algebraic convergence, the two Euler methods are said to "converge with first order", whereas the implicit midpoint method is called "second-order convergent".

The observations made for polygonal timestepping methods reflect a general pattern:

Algebraic convergence of single step methods

Consider the numerical integration of an initial value problem

$$\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y}) \quad , \quad \mathbf{y}(t_0) = \mathbf{y}_0 \, , \tag{11.1.3.2}$$

with sufficiently smooth right hand side function $\mathbf{f} : I \times D \to \mathbb{R}^N$.

Then conventional single step methods ($\to$ Def. 11.3.1.5) will enjoy *asymptotic algebraic conver-*

gence in the meshwidth, more precisely, see [DR08, Thm. 11.25],

there is a $p \in \mathbb{N}$ such that the sequence $(\mathbf{y}_k)_k$ generated by the single step method for $\dot{\mathbf{y}} = \mathbf{f}(t,\mathbf{y})$ on a mesh $\mathcal{M} := \{t_0 < t_1 < \cdots < t_M = T\}$ satisfies

$$\max_k \|\mathbf{y}_k - \mathbf{y}(t_k)\| \leq Ch^p \quad \text{for} \quad h := \max_{k=1,\ldots,M} |t_k - t_{k-1}| \to 0\,, \tag{11.3.2.7}$$

with $C > 0$ independent of $\mathcal{M}$

Definition 11.3.2.8. Order of a single step method

The maximal integer $p \in \mathbb{N}$ for which (11.3.2.7) holds for a single step method when applied to an ODE with (sufficiently) smooth right hand side, is called the **order** of the method.

As in the case of quadrature rules (→ Def. 7.4.1.1) their order is the principal intrinsic indicator for the "quality" of a single step method.

§11.3.2.9 (Convergence analysis for the explicit Euler method [Han02, Ch. 74]) We consider the simplest single-step method, namely the explicit Euler method (11.2.1.5) on a mesh $\mathcal{M} := \{0 = t_0 < t_1 < \cdots < t_M = T\}$ for a generic autonomous IVP

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}) \quad , \quad \mathbf{y}(0) = \mathbf{y}_0 \in D\,,$$

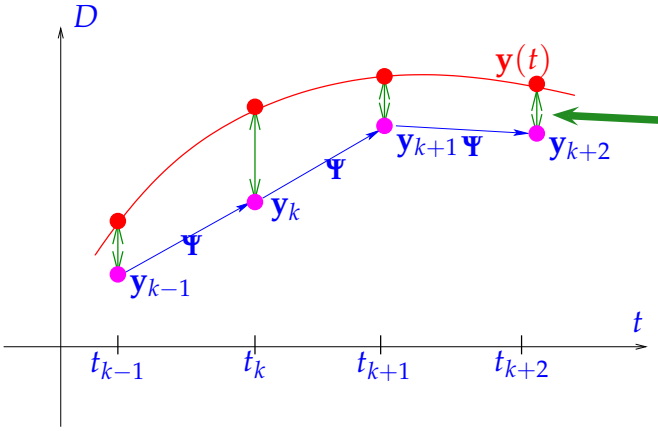
with sufficiently smooth and (*globally*) Lipschitz continuous $\mathbf{f} : D \subset \mathbb{R}^N \to \mathbb{R}^N$, that is,

$$\exists L > 0\colon \quad \|\mathbf{f}(\mathbf{y}) - \mathbf{f}(\mathbf{z})\| \leq L\|\mathbf{y} - \mathbf{z}\| \quad \forall \mathbf{y},\mathbf{z} \in D\,, \tag{11.3.2.10}$$

cf. Def. 11.1.3.13, and C^1 exact solution $t \mapsto \mathbf{y}(t)$. Throughout we assume that solutions of $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$ are defined on $[0,T]$ for all initial states $\mathbf{y}_0 \in D$.

Recall the recursion defining the explicit Euler method

$$\mathbf{y}_{k+1} = \mathbf{y}_k + h_k\mathbf{f}(\mathbf{y}_k)\,, \quad h_k := t_{k+1} - t_k\,, \quad k = 1,\ldots,M-1\,. \tag{11.2.1.5}$$



In numerical analysis one studies the

error sequence: $\mathbf{e}_k := \mathbf{y}_k - \mathbf{y}(t_k)$.

◁ — ≙ trajectory $t \mapsto \mathbf{y}(t)$
— ≙ Euler polygon,
• ≙ $\mathbf{y}(t_k)$,
• ≙ $\mathbf{y}_k$,
⟶ ≙ discrete evolution $\Psi^{t_{k+1}-t_k}$

The approach to estimate $\|\mathbf{e}_k\|$ follows a fundamental policy that comprises three key steps. To explain them we rely on the abstract concepts of the

- evolution operator Φ associated with the ODE $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$ (→ Def. 11.1.4.3) and
- discrete evolution operator Ψ defining the explicit Euler single step method, see Def. 11.3.1.5:

$$\text{(11.2.1.5)} \quad \Rightarrow \quad \Psi^h \mathbf{y} = \mathbf{y} + h\mathbf{f}(\mathbf{y}) \,. \tag{11.3.2.11}$$

We argue that in this context abstraction pays off, because it helps elucidate a general technique for the convergence analysis of single step methods.

① Abstract **splitting of error**:

Fundamental error splitting:

$$\begin{aligned} \mathbf{e}_{k+1} &= \Psi^{h_k}\mathbf{y}_k - \Phi^{h_k}\mathbf{y}(t_k) \\ &= \underbrace{\Psi^{h_k}\mathbf{y}_k - \Psi^{h_k}\mathbf{y}(t_k)}_{\text{propagated error}} \\ &\quad + \underbrace{\Psi^{h_k}\mathbf{y}(t_k) - \Phi^{h_k}\mathbf{y}(t_k)}_{\text{one-step error}} \,. \end{aligned} \tag{11.3.2.12}$$

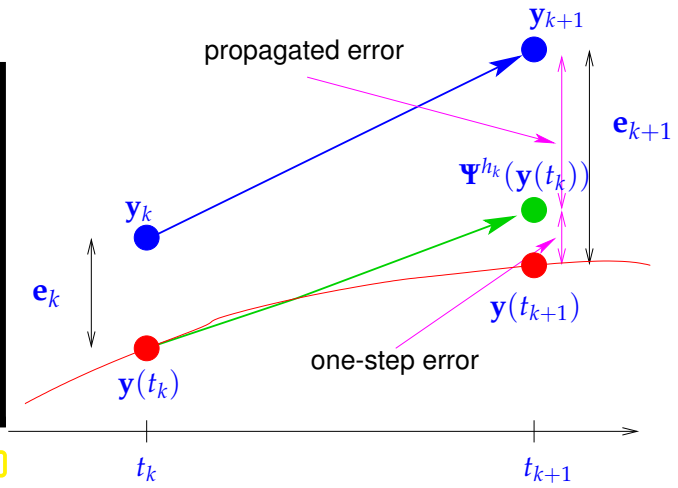


Fig. 422

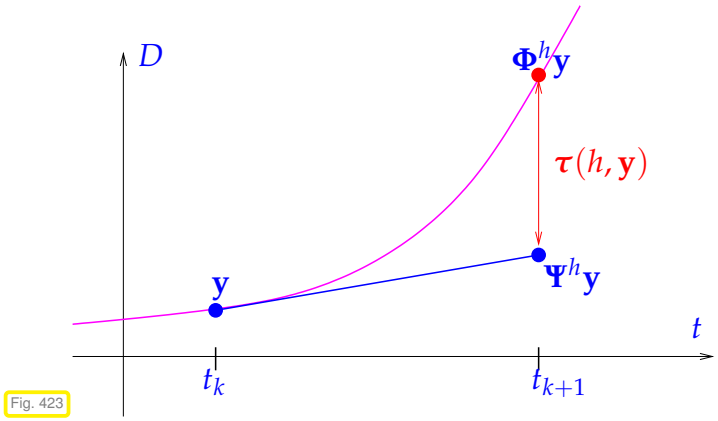


Fig. 423

A generic one-step error expressed through continuous and discrete evolutions reads:

$$\tau(h,\mathbf{y}) := \Psi^h\mathbf{y} - \Phi^h\mathbf{y} \,. \tag{11.3.2.13}$$

◁ geometric visualisation of one-step error for explicit Euler method (11.2.1.5), *cf.* Fig. 415, $h := t_{k+1} - t_k$

—: solution trajectory through $(t_k, \mathbf{y})$

② **Estimate for one-step error $\tau(h_k, \mathbf{y}(t_k))$**:

Geometric considerations: distance of a smooth curve and its tangent shrinks as the square of the distance to the intersection point (curve locally looks like a parabola in the $\xi - \eta$ coordinate system, see Fig. 425).

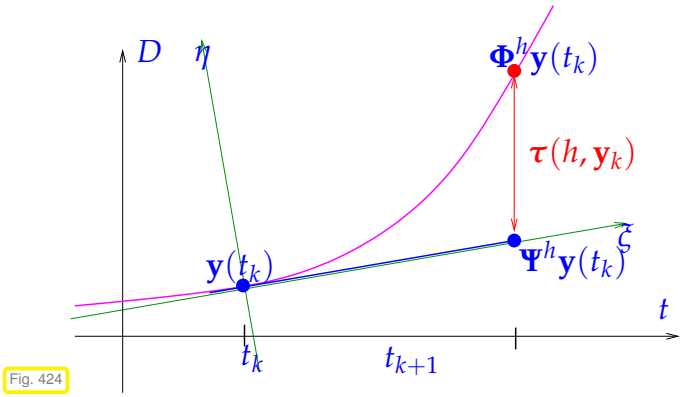


Fig. 424

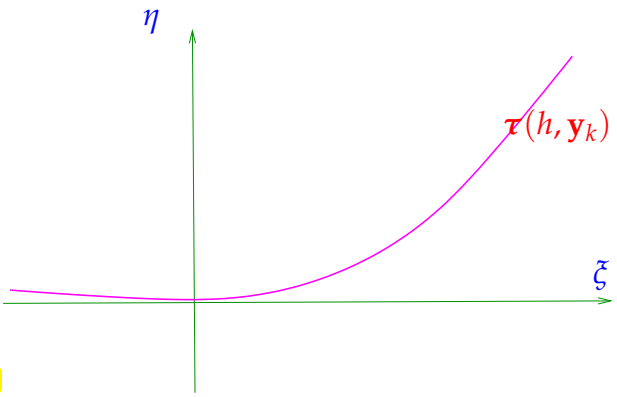


Fig. 425

The geometric considerations can be made rigorous by analysis: recall Taylor's formula for the function

$\mathbf{y} \in C^{K+1}$ [Str09, Satz 5.5.1]:

$$\mathbf{y}(t+h) - \mathbf{y}(t) = \sum_{j=1}^{K} \mathbf{y}^{(j)}(t)\frac{h^j}{j!} + \underbrace{\int_t^{t+h} \mathbf{y}^{(K+1)}(\tau)\frac{(t+h-\tau)^K}{K!}\,\mathrm{d}\tau}_{= \dfrac{\mathbf{y}^{(K+1)}(\xi)}{K!}h^{K+1}} , \tag{11.3.2.14}$$

for some $\xi \in [t, t+h]$. We conclude that, if $\mathbf{y} \in C^2([0,T])$, which is ensured for smooth $\mathbf{f}$, see Lemma 11.1.1.3, then

$$\mathbf{y}(t_{k+1}) - \mathbf{y}(t_k) = \dot{\mathbf{y}}(t_k)h_k + \tfrac{1}{2}\ddot{\mathbf{y}}(\xi_k)h_k^2 = \mathbf{f}(\mathbf{y}(t_k))h_k + \tfrac{1}{2}\ddot{\mathbf{y}}(\xi_k)h_k^2 ,$$

for some $t_k \le \xi_k \le t_{k+1}$.This leads to an expression for the one-step error from (11.3.2.13)

$$\begin{aligned} \boldsymbol{\tau}(h_k, \mathbf{y}(t_k)) &= \boldsymbol{\Psi}^{h_k}\mathbf{y}(t_k) - \mathbf{y}(t_{k+1}) \\ &\overset{(11.3.2.11)}{=} \mathbf{y}(t_k) + h_k\mathbf{f}(\mathbf{y}(t_k)) - \mathbf{y}(t_k) - \mathbf{f}(\mathbf{y}(t_k))h_k + \tfrac{1}{2}\ddot{\mathbf{y}}(\xi_k)h_k^2 \\ &= \tfrac{1}{2}\ddot{\mathbf{y}}(\xi_k)h_k^2 . \end{aligned} \tag{11.3.2.15}$$

Sloppily speaking, we observe $\boxed{\boldsymbol{\tau}(h_k, \mathbf{y}(t_k)) = O(h_k^2)}$ uniformly for $h_k \to 0$.

③ **Estimate for the propagated error** from (11.3.2.12)

$$\begin{aligned} \left\| \boldsymbol{\Psi}^{h_k}\mathbf{y}_k - \boldsymbol{\Psi}^{h_k}\mathbf{y}(t_k) \right\| &= \|\mathbf{y}_k + h_k\mathbf{f}(\mathbf{y}_k) - \mathbf{y}(t_k) - h_k\mathbf{f}(\mathbf{y}(t_k))\| \\ &\overset{(11.3.2.10)}{\le} (1 + Lh_k)\|\mathbf{y}_k - \mathbf{y}(t_k)\| . \end{aligned} \tag{11.3.2.16}$$

Thus we obtain *recursion* for error norms $\epsilon_k := \|\mathbf{e}_k\|$ by simply applying the $\triangle$-inequality:

$$\boxed{\epsilon_{k+1} \le (1 + h_k L)\epsilon_k + \rho_k , \quad \rho_k := \tfrac{1}{2}h_k^2 \max_{t_k \le \tau \le t_{k+1}} \|\ddot{\mathbf{y}}(\tau)\|} . \tag{11.3.2.17}$$

Taking into account $\epsilon_0 = 0$, this leads to

$$\epsilon_k \le \sum_{l=1}^{k}\prod_{j=1}^{l-1}(1 + Lh_j)\,\rho_l , \quad k = 1, \ldots, N . \tag{11.3.2.18}$$

Use the elementary estimate $(1 + Lh_j) \le \exp(Lh_j)$ (by convexity of the exponential function):

$$\text{(11.3.2.18)} \quad \Rightarrow \quad \epsilon_k \le \sum_{l=1}^{k}\prod_{j=1}^{l-1}\exp(Lh_j)\cdot\rho_l = \sum_{l=1}^{k}\exp(L\textstyle\sum_{j=1}^{l-1}h_j)\rho_l .$$

Note: $\sum_{j=1}^{l-1} h_j \le T$ for final time T and conclude

$$\epsilon_k \le \exp(LT)\sum_{l=1}^{k}\rho_l \le \exp(LT)\max_k\frac{\rho_k}{h_k}\sum_{l=1}^{k}h_l \le T\exp(LT)\max_{l=1,\ldots,k}h_l \cdot \max_{t_0\le\tau\le t_k}\|\ddot{\mathbf{y}}(\tau)\| .$$

$$\Rightarrow \quad \|\mathbf{y}_k - \mathbf{y}(t_k)\| \le T\exp(LT)\max_{l=1,\ldots,k}h_l \cdot \max_{t_0\le\tau\le t_k}\|\ddot{\mathbf{y}}(\tau)\| . \tag{11.3.2.19}$$

We can summarize the insight gleaned through this theoretical analysis as follows:

Total error arises from accumulation of propagated one-step errors!

From (11.3.2.19) we can conclude

- an error bound $= O(h)$, $h := \max_l h_l$ (➤ 1st-order algebraic convergence)
- and that the error bound grows *exponentially* with the length T of the integration interval.

⌟

§11.3.2.20 (One-step error and order of a single step method) In the analysis of the global discretization error of the explicit Euler method in § 11.3.2.9 a one-step error of size $O(h_k^2)$ led to a total error of $O(h)$ through the effect of error accumulation over $M \approx h^{-1}$ steps. This relationship remains valid for almost all single step methods [DB02, Theorem 4.10]:

Order of algebraic convergence of single-step methods

Consider an IVP (11.1.3.2) with solution $t \mapsto \mathbf{y}(t)$ and a single step method defined by the discrete evolution $\boldsymbol{\Psi}$ ($\rightarrow$ Def. 11.3.1.5). If the *one-step error along the solution trajectory* satisfies ($\boldsymbol{\Phi}$ is the evolution map associated with the ODE, see Def. 11.1.4.3)

$$\left\| \boldsymbol{\Psi}^h \mathbf{y}(t) - \boldsymbol{\Phi}^h \mathbf{y}(t) \right\| \leq Ch^{p+1} \quad \forall h \text{ sufficiently small}, \ t \in [0, T] \ , \tag{11.3.2.22}$$

for some $p \in \mathbb{N}$ and $C > 0$, then, usually,

$$\max_k \|\mathbf{y}_k - \mathbf{y}(t_k)\| \leq \overline{C} h_{\mathcal{M}}^p \ ,$$

with $\overline{C} > 0$ independent of the temporal mesh $\mathcal{M}$: The (pointwise) discretization error *converges algebraically* with order/rate p.

A rigorous statement as a theorem would involve some particular assumptions on $\boldsymbol{\Psi}$, which we do not want to give here. These assumptions are satisfied, for instance, for all the methods presented in the sequel. You may refer to [DB02, Sect. 4.1] for further information.

In fact, it is remarkable that a local condition like (11.3.2.22) permits us to make a quantitative prediction of global convergence. This close relationship has made researchers introduce "order" also as a property of discrete evolutions.

Definition 11.3.2.23. Order of a discrete evolution operator

Let $\boldsymbol{\Psi} : I \times D \mapsto \mathbb{R}^N$ be a discrete evolution for the autonomous ODE $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$ (with associated evolution operator $\boldsymbol{\Phi} : I \times D \mapsto \mathbb{R}^N \rightarrow$ Def. 11.1.4.3). The largest integer $q \in \mathbb{N}_0$ such that

$$\forall \mathbf{y} \in D \quad \exists \tau_0 > 0: \quad \|\boldsymbol{\Psi}^\tau \mathbf{y} - \boldsymbol{\Phi}^\tau \mathbf{y}\| \leq C(\mathbf{y}) \tau^{q+1} \quad \forall |\tau| \leq \tau_0 \tag{11.3.2.24}$$

is called the **order** of the discrete evolution. .

This notion of "order of a discrete evolution" allows a concise summary:

A single-step method (SSM, Def. 11.3.1.5) based on the discrete evolution $\boldsymbol{\Psi}$ satisfies
$\boldsymbol{\Psi}$ of order $q \in \mathbb{N}$ ⟹ SSM converges algebraically with order q.

⌟

EXAMPLE 11.3.2.25 (Orders of finite-difference single-step methods) Let us determine orders of the discrete evolutions for the three simple single-step methods introduced in Section 11.2, here listed with their corresponding discrete evolution operators $\boldsymbol{\Psi}$ (→ § 11.3.1.1) when applied to an autonomous ODE $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$: for $\mathbf{y}_0 \in D \subset \mathbb{R}^N$,

explicit (forward) Euler method (11.2.1.5):	$\boldsymbol{\Psi}^\tau \mathbf{y}_0 := \mathbf{y}_0 + \tau \mathbf{f}(\mathbf{y}_0)$,	(11.3.2.26)
implicit (backward) Euler method (11.2.2.2):	$\boldsymbol{\Psi}^\tau \mathbf{y}_0 := \mathbf{w}$: $\mathbf{w} = \mathbf{y}_0 + \tau \mathbf{f}(\mathbf{w})$,	(11.3.2.27)
implicit midpoint method (11.2.3.3):	$\boldsymbol{\Psi}^\tau \mathbf{y}_0 := \mathbf{w}$: $\mathbf{w} = \mathbf{y}_0 + \tau \mathbf{f}(\frac{1}{2}(\mathbf{y}_0 + \mathbf{w}))$.	(11.3.2.28)

The computation of their orders will rely on a fundamental technique for establishing (11.3.2.24) based on Taylor expansion, which asserts that for a function $\mathbf{g} \in C^{m+1}(]t_0 - \delta, t_0 + \delta[, \mathbb{R}^N)$, $\delta > 0$,

$$\mathbf{g}(t_0 + \tau) = \sum_{k=0}^{m} \frac{1}{m!} \mathbf{g}^{(k)}(t_0) h \tau^k + O(\tau^{m+1}) \quad \text{for} \quad \tau \to 0 \,. \tag{11.3.2.29}$$

Of course the arguments hinge on the smoothness of the vectorfield $\mathbf{f} = \mathbf{f}(\mathbf{y})$, which will ensure smoothness of solutions of the associated ODE $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$. Thus, we make the following simplifying assumption:

Assumption 11.3.2.30. Smoothness of right-hand side vectorfield

The vectorfield $\mathbf{y} \mapsto \mathbf{f}(\mathbf{y})$ is C^∞ on $\mathbb{R}^N$

Let $\boldsymbol{\Phi} = \boldsymbol{\Phi}(t, \mathbf{y})$ denote the evolution operator (→ Def. 11.1.4.3) induced by $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$, which, by definition, satisfies

$$\frac{\partial \boldsymbol{\Phi}}{\partial t}(t, \mathbf{y}_0) = \mathbf{f}(\boldsymbol{\Phi}^t \mathbf{y}_0) \quad \forall \mathbf{y}_0 \in D,\ t \in J(\mathbf{y}_0) \,. \tag{11.1.4.4}$$

Setting $\mathbf{v}(\tau) := \boldsymbol{\Phi}^\tau \mathbf{y}_0$, which is a solution of the initial-value problem $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$, $\mathbf{y}(0) = \mathbf{y}_0$, we find for small τ, appealing to the one-dimensional chain rule and (11.1.4.4),

$$\frac{d\mathbf{v}}{d\tau}(\tau) = \mathbf{f}(\mathbf{v}(\tau)) \quad , \quad \frac{d^2\mathbf{v}}{d\tau^2}(\tau) = \frac{\partial \mathbf{f}}{\partial \mathbf{y}}(\mathbf{v}(\tau)) \frac{d\mathbf{v}}{d\tau}(\tau) \,. \tag{11.3.2.31}$$

This yields the following truncated Taylor expansion

$$\begin{aligned} \boldsymbol{\Phi}^\tau \mathbf{y}_0 = \mathbf{v}(\tau) &= \mathbf{v}(0) + \tau \frac{d\mathbf{v}}{d\tau}(0) + \tfrac{1}{2}\tau^2 \frac{d^2\mathbf{v}}{d\tau^2}(0) + O(\tau^3) \\ &= \mathbf{y}_0 + \tau\, \mathbf{f}(\mathbf{y}_0) + \tfrac{1}{2}\tau^2\, \mathrm{D}\mathbf{f}(\mathbf{y}_0)\mathbf{f}(t, \mathbf{y}_0) + O(\tau^3) \end{aligned} \tag{11.3.2.32}$$

for $\tau \to 0$. Note that the derivative $\mathrm{D}\mathbf{f}(\mathbf{y}_0)$ is an $N \times N$ Jacobi matrix. Explicit expressions for the remainder term involve second derivatives of $\mathbf{f}$.

❶ For the explicit Euler method (11.3.2.26) we immediately have from (11.3.2.32)

$$\boldsymbol{\Psi}^\tau \mathbf{y}_0 - \boldsymbol{\Phi}^\tau \mathbf{y}_0 = \mathbf{y}_0 + \tau \mathbf{f}(\mathbf{y}_0) - \mathbf{y}_0 - \tau \mathbf{f}(\mathbf{y}_0) + O(\tau^2) = O(\tau^2) \quad \text{for} \quad \tau \to 0 \,.$$

The explicit Euler method is of order 1.

❷ It is not as straightforward for the implicit Euler method

$$\boldsymbol{\Psi}^\tau \mathbf{y}_0 := \mathbf{w}(\tau): \quad \mathbf{w}(\tau) = \mathbf{y}_0 + \tau \mathbf{f}(\mathbf{w}(\tau)) \,. \tag{11.3.2.27}$$

First, we plug (11.3.2.27) into itself

$$\mathbf{w}(\tau) = \mathbf{y}_0 + \tau\mathbf{f}(\mathbf{w}(\tau)) = \mathbf{y}_0 + \tau\mathbf{f}(\mathbf{y}_0 + \tau\mathbf{f}(\mathbf{w}(\tau))) \,,$$

and then use the truncated Taylor expansion of $\mathbf{f}$ around $\mathbf{y}_0$

$$\mathbf{f}(\mathbf{y}_0 + \mathbf{v}) = \mathbf{f}(\mathbf{y}_0) + \mathrm{D}\,\mathbf{f}(\mathbf{y}_0)\mathbf{v} + O(\|\mathbf{v}\|^2) \quad \text{for} \quad \mathbf{v} \to \mathbf{0} \,. \tag{11.3.2.33}$$

This gives

$$\mathbf{w}(\tau) = \mathbf{y}_0 + \tau(\mathbf{f}(\mathbf{y}_0) + \tau\,\mathrm{D}\,\mathbf{f}(\mathbf{y}_0)\mathbf{f}(\mathbf{w}(\tau))) + O(\tau^3) \quad \text{for} \quad \tau \to 0 \,.$$

Since $\boldsymbol{\Psi}^\tau\mathbf{y}_0 = \mathbf{w}(\tau)$, matching terms with (11.3.2.32) we obtain

$$\boldsymbol{\Psi}^\tau\mathbf{y}_0 - \boldsymbol{\Phi}^\tau\mathbf{y}_0 = \tau^2\,\mathrm{D}\,\mathbf{f}(\mathbf{y}_0)\mathbf{f}(\mathbf{w}(\tau)) + O(\tau^3) = O(\tau^2) \quad \text{for} \quad \tau \to 0 \,.$$

Thanks to the smoothness of $\mathbf{f}$ the remainder terms will depend continuously on $\mathbf{y}_0$.

▶ The implicit Euler method has order 1.

❸ For the implicit midpoint rule

$$\boldsymbol{\Psi}^\tau\mathbf{y}_0 := \mathbf{w}(\tau): \quad \mathbf{w}(\tau) = \mathbf{y}_0 + \tau\mathbf{f}(\tfrac{1}{2}(\mathbf{y}_0 + \mathbf{w}(\tau))) \,, \tag{11.3.2.28}$$

we follow the same idea and consider

$$\begin{aligned}\mathbf{w}(\tau) &= \mathbf{y}_0 + \tau\mathbf{f}(\tfrac{1}{2}(\mathbf{y}_0 + \mathbf{w}(\tau))) = \mathbf{y}_0 + \tau\mathbf{f}(\mathbf{y}_0 + \tfrac{1}{2}\tau\mathbf{f}(\tfrac{1}{2}(\mathbf{y}_0 + \mathbf{w}(\tau))) \\ &= \mathbf{y}_0 + \tau\mathbf{f}(\mathbf{y}_0 + \tfrac{1}{2}\tau\mathbf{f}(\mathbf{y}_0 + O(\tau))) \quad \text{for} \quad \tau \to 0 \,.\end{aligned}$$

Then we resort to the truncated Taylor expansion (11.3.2.33) and get for $\tau \to 0$

$$\begin{aligned}\mathbf{w}(\tau) &= \mathbf{y}_0 + \tau\Big(\mathbf{f}(\mathbf{y}_0) + \mathrm{D}\,\mathbf{f}(\mathbf{y}_0)\,\tfrac{1}{2}\tau\mathbf{f}(\mathbf{y}_0 + O(\tau))\Big) + O(\tau^3) \\ &= \mathbf{y}_0 + \tau\Big(\mathbf{f}(\mathbf{y}_0) + \mathrm{D}\,\mathbf{f}(\mathbf{y}_0)\,\tfrac{1}{2}\tau(\mathbf{f}(\mathbf{y}_0) + O(\tau))\Big) + O(\tau^3) \,.\end{aligned}$$

Matching with (11.3.2.32) shows $\mathbf{w}(\tau) - \boldsymbol{\Phi}^\tau\mathbf{y}_0 = O(\tau^3)$ where the "O" just comprises *continuous* higher order derivatives of $\mathbf{f}$.

▶ The implicit midpoint method is an order-2 method.

Hardly surprising, these analytic results match the orders of algebraic convergence observed in Exp. 11.3.2.5. ⌟

Review question(s) 11.3.2.34 (Asymptotic convergence of single-step methods)

(Q11.3.2.34.A) We consider an autonomous ODE $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$ with smooth $\mathbf{f} : D \subset \mathbb{R}^N \to \mathbb{R}^n$. Explain, why the one-step error

$$\boldsymbol{\tau}(h, \mathbf{y}) = \boldsymbol{\Psi}^h\mathbf{y} - \boldsymbol{\Phi}^h\mathbf{y} \,, \quad \mathbf{y} \in \ , \quad h \text{ "sufficiently small"} \,,$$

for a *consistent* single-step method defined by the discrete evolution operator $\boldsymbol{\Psi}$ satisfies

$$\forall \mathbf{y} \in D: \quad \boldsymbol{\tau}(h, \mathbf{y}) = O(h) \quad \text{for} \quad h \to 0 \,.$$

Definition 11.3.1.12. Consistent single step methods

A single step method according to Def. 11.3.1.5 based on a discrete evolution of the form

$$\Psi^h \mathbf{y} = \mathbf{y} + h\boldsymbol{\psi}(h,\mathbf{y}) \quad \text{with} \quad \begin{array}{l} \boldsymbol{\psi} : I \times D \to \mathbb{R}^N \text{ continuous,} \\ \boldsymbol{\psi}(0,\mathbf{y}) = \mathbf{f}(\mathbf{y}) \,. \end{array} \tag{11.3.1.11}$$

is called **consistent** with the ODE $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$.

(Q11.3.2.34.B) Let $t \in I \mapsto \mathbf{y}(t)$, $I \subset \mathbb{R}$ an interval containing 0, denote the solution of the autonomous IVP

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}) \quad , \quad \mathbf{y}(0) = \mathbf{y}_0 \,.$$

Assume that $\mathbf{f}$ is continuously differentiable.

Use the chain rule to express $\dot{\mathbf{y}}(t^*)$ and $\ddot{\mathbf{y}}(t^*)$ by means of $\mathbf{f}$ and its Jacobian.

(Q11.3.2.34.C) Based on the answer to Question (**Q**11.3.2.34.B), determine the order of a single-step method for the autonomous ODE $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$, $\mathbf{f} : D \subset \mathbb{R}^N \to \mathbb{R}^N$ smooth, whose discrete evolution operator is given by

$$\Psi^h \mathbf{y} := \mathbf{y} + h\mathbf{f}(\mathbf{y}) + \tfrac{1}{2}h^2 \,\mathrm{D}\,\mathbf{f}(\mathbf{y})\mathbf{f}(\mathbf{y}) \,,$$

where $\mathrm{D}\,\mathbf{f}(\mathbf{y}) \in \mathbb{R}^{N,N}$ is the Jacobian of $\mathbf{f}$ in $\mathbf{y} \in D$.

△

11.4 Explicit Runge-Kutta Single-Step Methods (RKSSMs)

Video tutorial for Section 11.4: Explicit Runge-Kutta Single-Step Methods (RKSSMs): (27 minutes) Download link, tablet notes

So far we only know first and second order methods from 11.2: the explicit and implicit Euler method (11.2.1.5) and (11.2.2.2), respectively, are of first order, the implicit midpoint rule of second order. We observed this in Exp. 11.3.2.5 and it can be proved rigorously for all three methods adapting the arguments of § 11.3.2.9.

Thus, barring the impact of roundoff, the low-order polygonal approximation methods are guaranteed to achieve any prescribed accuracy provided that the mesh is fine enough. Why should we need any other timestepping schemes?

Remark 11.4.0.1 (Rationale for high-order single step methods ***cf.*** **[DR08, Sect. 11.5.3])** We argue that the use of higher-order timestepping methods is highly *advisable for the sake of efficiency*. The reasoning is very similar to that of Rem. 7.4.3.12, when we considered numerical quadrature. The reader is advised to study that remark again.

As we saw in § 11.3.2.3 error bounds for single step methods for the solution of IVPs will inevitably feature unknown constants "$C > 0$". Thus they do **not** give useful information about the discretization error for a concrete IVP and mesh. Hence, it is too ambitious to ask how many timesteps are needed so that $\|\mathbf{y}(T) - \mathbf{y}_N\|$ stays below a prescribed bound.

However an easier question can be answered by *asymptotic estimates* like (11.3.2.7) and this questions reads:

What extra computational effort buys a prescribed *reduction of the error*?

The usual concept of "computational effort" for single step methods (→ Def. 11.3.1.5) is as follows

Computational effort
- ∼ total number of $\mathbf{f}$-evaluations for approximately solving the IVP,
- ∼ number of timesteps, if evaluation of discete evolution Ψ^h (→ Def. 11.3.1.5) requires fixed number of $\mathbf{f}$-evaluations,
- ∼ h^{-1}, in the case of uniform timestep size $h > 0$ (equidistant mesh (11.3.2.4)).

Now, let us consider a single step method of order $p \in \mathbb{N}$, employed with a uniform timestep h_{old}. We focus on the maximal discretization error in the mesh points, see § 11.3.2.1. We make the crucial assumption that the asymptotic error bounds are *sharp*:

$$\operatorname{err}(h) \approx Ch^p \quad \text{for small meshwidth } h > 0\,, \tag{11.4.0.2}$$

with a "generic constant" $C > 0$ independent of the mesh.

Goal: $\dfrac{\operatorname{err}(h_{\text{new}})}{\operatorname{err}(h_{\text{old}})} \overset{!}{=} \dfrac{1}{\rho}$ for reduction factor $\rho > 1$.

(11.3.2.7) $\Rightarrow$ $\dfrac{h_{\text{new}}^p}{h_{\text{old}}^p} \overset{!}{=} \dfrac{1}{\rho}$ $\Leftrightarrow$ $\boxed{h_{\text{new}} = \rho^{-1/p} h_{\text{old}}}$. (11.4.0.3)

For single step method of order $p \in \mathbb{N}$ we conclude the behavior:

increase effort by factor $\rho^{1/p}$ ⟹ reduce error by factor $\rho > 1$

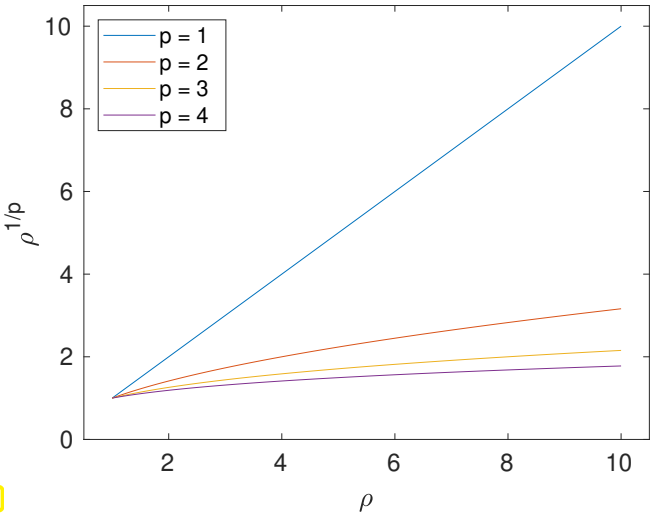


Fig. 426

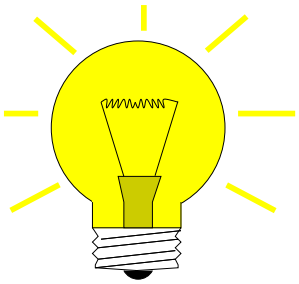
◁ Plots of $\rho^{1/p}$ vs. ρ

☞ the larger the order p, the less effort for a prescribed reduction of the error!

We remark that another (minor) rationale for using higher-order methods is to curb impact of roundoff errors (→ Section 1.5.3) accumulating during timestepping [DR08, Sect. 11.5.3]. ⌟

§11.4.0.4 (Bootstrap construction of explicit single step methods) Now we will build a class of methods that are explicit and achieve orders $p > 2$. The starting point is a simple *integral equation* satisfied by any solution $t \mapsto \mathbf{y}(t)$ of an initial value problems for the general ODE $\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y})$:

IVP: $\dot{\mathbf{y}}(t) = \mathbf{f}(t, \mathbf{y}(t))\,,$ $\mathbf{y}(t_0) = \mathbf{y}_0$ $\Rightarrow$ $\mathbf{y}(t_1) = \mathbf{y}_0 + \int_{t_0}^{t_1} \mathbf{f}(\tau, \mathbf{y}(\tau))\,\mathrm{d}\tau$



Idea: approximate the integral by means of s-point quadrature formula → Section 7.2, defined on the reference interval $[0,1]$) with nodes $c_1,\ldots,c_s$, weights $b_1,\ldots,b_s$.

$$\mathbf{y}(t_1) \approx \mathbf{y}_1 = \mathbf{y}_0 + h\sum_{i=1}^{s} b_i \mathbf{f}(t_0 + c_i h, \boxed{\mathbf{y}(t_0 + c_i h)}) , \quad h := t_1 - t_0 . \tag{11.4.0.5}$$

Obtain these values by bootstrapping

"Bootstrapping" = use the same idea in a simpler version to get $\mathbf{y}(t_0 + c_i h)$, noting that these values can be replaced by other approximations obtained by methods already constructed (this approach will be elucidated in the next example).

What error can we afford in the approximation of $\mathbf{y}(t_0 + c_i h)$ (under the assumption that $\mathbf{f}$ is Lipschitz continuous)? We take the cue from the considerations in § 11.3.2.9.

Goal: aim for one-step error bound $\mathbf{y}(t_1) - \mathbf{y}_1 = O(h^{p+1})$

Note that there is a factor h in front of the quadrature sum in (11.4.0.5). Thus, our goal can already be achieved, if only

$$\mathbf{y}(t_0 + c_i h) \text{ is approximated up to an error } O(h^p),$$

again, because in (11.4.0.5) a factor of size h multiplies $\mathbf{f}(t_0 + c_i, \mathbf{y}(t_0 + c_i h))$.

This is accomplished by a less accurate discrete evolution than the one we are about to build. Thus, we can construct discrete evolutions of higher and higher order, in turns, starting with the explicit Euler method. All these methods will be explicit, that is, $\mathbf{y}_1$ can be computed directly from point values of $\mathbf{f}$. ⌟

EXAMPLE 11.4.0.6 (Simple Runge-Kutta methods by quadrature & boostrapping) Now we apply the boostrapping idea outlined above. We write $\mathbf{k}_\ell \in \mathbb{R}^N$ for the approximations of $\mathbf{y}(t_0 + c_i h)$.

- Quadrature formula = trapezoidal rule (7.3.0.5):

  $$Q(f) = \tfrac{1}{2}(f(0) + f(1)) \quad \leftrightarrow \quad s = 2\text{:} \quad c_1 = 0, c_2 = 1 , \quad b_1 = b_2 = \frac{1}{2} , \tag{11.4.0.7}$$

  and $\mathbf{y}(t_1)$ approximated by explicit Euler step (11.2.1.5)

  $$\mathbf{k}_1 = \mathbf{f}(t_0, \mathbf{y}_0) , \quad \mathbf{k}_2 = \mathbf{f}(t_0 + h, \mathbf{y}_0 + h\mathbf{k}_1) , \quad \mathbf{y}_1 = \mathbf{y}_0 + \tfrac{h}{2}(\mathbf{k}_1 + \mathbf{k}_2) . \tag{11.4.0.8}$$

  (11.4.0.8) = explicit trapezoidal method (for numerical integration of ODEs).

- Quadrature formula → simplest Gauss quadrature formula = midpoint rule → Ex. 7.3.0.3 & $\mathbf{y}(\frac{1}{2}(t_1 + t_0))$ approximated by explicit Euler step (11.2.1.5)

  $$\mathbf{k}_1 = \mathbf{f}(t_0, \mathbf{y}_0) , \quad \mathbf{k}_2 = \mathbf{f}(t_0 + \tfrac{h}{2}, \mathbf{y}_0 + \tfrac{h}{2}\mathbf{k}_1) , \quad \mathbf{y}_1 = \mathbf{y}_0 + h\mathbf{k}_2 . \tag{11.4.0.9}$$

  (11.4.0.9) = explicit midpoint method (for numerical integration of ODEs) [DR08, Alg. 11.18].

⌟

EXAMPLE 11.4.0.10 (Convergence of simple Runge-Kutta methods) We perform an empiric study of the order of the explicit single step methods constructed in Ex. 11.4.0.6.

- IVP: $\dot{y} = 10y(1-y)$ (scalar logistic ODE (11.1.2.2)), initial value $y(0) = 0.01$, final time $T = 1$,
- Explicit single step methods, uniform timestep h.

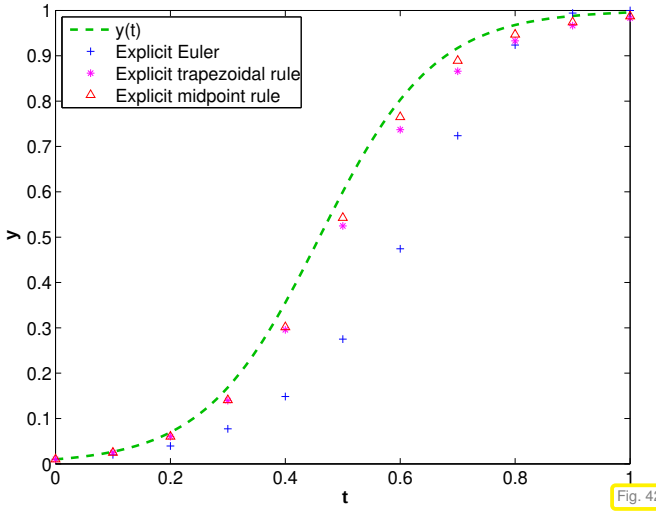


Fig. 427

$y_h(j/10)$, $j = 1, \ldots, 10$ for explicit RK-methods

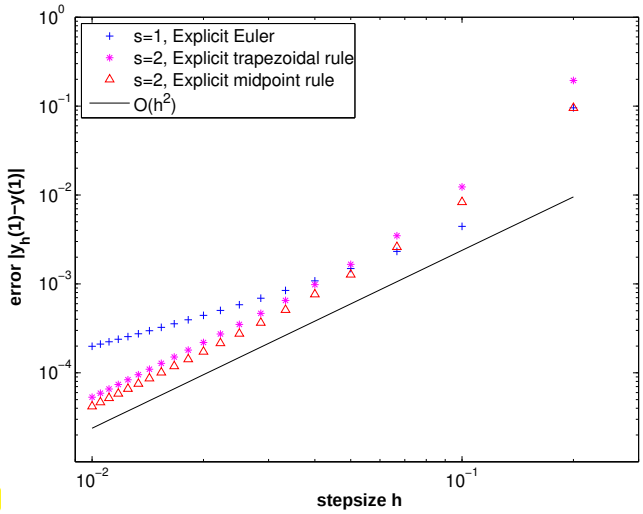


Fig. 428

Errors at final time $y_h(1) - y(1)$

Observation: obvious *algebraic convergence* in meshwidth h with integer rates/orders:

explicit trapezoidal method (11.4.0.8) $\rightarrow$ order 2
explicit midpoint method (11.4.0.9) $\rightarrow$ order 2

This is what one expects from the considerations in Ex. 11.4.0.6. ⌟

The formulas that we have obtained follow a general pattern:

Definition 11.4.0.11. Explicit Runge-Kutta single-step method

For $b_i, a_{ij} \in \mathbb{R}$, $c_i := \sum_{j=1}^{i-1} a_{ij}$, $i, j = 1, \ldots, s$, $s \in \mathbb{N}$, an **s-stage explicit Runge-Kutta single step method** (RK-SSM) for the ODE $\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y})$, $\mathbf{f} : \Omega \to \mathbb{R}^N$, is defined by ($\mathbf{y}_0 \in D$)

$$\mathbf{k}_i := \mathbf{f}(t_0 + c_i h, \mathbf{y}_0 + h \sum_{j=1}^{i-1} a_{ij} \mathbf{k}_j) \ , \quad i = 1, \ldots, s \quad , \quad \mathbf{y}_1 := \mathbf{y}_0 + h \sum_{i=1}^{s} b_i \mathbf{k}_i \ .$$

The vectors $\mathbf{k}_i \in \mathbb{R}^N$, $i = 1, \ldots, s$, are called **increments**, $h > 0$ is the size of the timestep.

Recall Rem. 11.3.1.15 to understand how the discrete evolution for an explicit Runge-Kutta method is specified in this definition by giving the formulas for the first step. This is a convention widely adopted in the literature about numerical methods for ODEs. Of course, the increments $\mathbf{k}_i$ have to be computed anew in each timestep.

The implementation of an s-stage explicit Runge-Kutta single step method according to Def. 11.4.0.11 is straightforward: The increments $\mathbf{k}_i \in \mathbb{R}^N$ are computed successively, starting from $\mathbf{k}_1 = \mathbf{f}(t_0 + c_1 h, \mathbf{y}_0)$.

Only s $\mathbf{f}$-evaluations and AXPY operations ($\rightarrow$ Section 1.3.2) are required to compute the next state vector from the current.

In books and research articles a particular way to write down the coefficients characterizing RK-SSMs is widely used:

Butcher scheme notation for explicit RK-SSM

Shorthand notation for (explicit) Runge-Kutta methods [DR08, (11.75)]

Butcher scheme ▷

(Note that $\mathfrak{A}$ is a strictly lower triangular $s \times s$-matrix)

$$\begin{array}{c|c} \mathbf{c} & \mathfrak{A} \\ \hline & \mathbf{b}^T \end{array} := \begin{array}{c|ccccc} c_1 & 0 & & \cdots & & 0 \\ c_2 & a_{21} & \ddots & & & \vdots \\ \vdots & \vdots & \ddots & & & \vdots \\ \vdots & \vdots & & \ddots & & \vdots \\ \vdots & \vdots & & & \ddots & \vdots \\ c_s & a_{s1} & \cdots & & a_{s,s-1} & 0 \\ \hline & b_1 & \cdots & & b_{s-1} & b_s \end{array} . \tag{11.4.0.13}$$

Now we restrict ourselves to the case of an autonomous ODE $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$. Matching Def. 11.4.0.11 and Def. 11.3.1.5, we see that the discrete evolution induced by an explicit Runge-Kutta single-step method is

$$\Psi^h \mathbf{y} = \mathbf{y} + h \sum_{i=1}^{s} b_i \mathbf{k}_i , \quad h \in \mathbb{R} , \quad \mathbf{y} \in D , \tag{11.4.0.14}$$

where the increments $\mathbf{k}_i$ are defined by the increment equations

$$\mathbf{k}_i := \mathbf{f}\left(\mathbf{y} + h \sum_{j=1}^{i-1} a_{ij} \mathbf{k}_j\right) .$$

In line with (11.3.1.11), this discrete evolution can be written as

$$\Psi^h \mathbf{y} = \mathbf{y} + h \boldsymbol{\psi}(h, \mathbf{y}) \quad , \quad \boldsymbol{\psi}(h, \mathbf{y}) = \sum_{i=1}^{s} b_i \mathbf{k}_i .$$

Is this discrete evolution consistent in the sense of § 11.3.1.8, that is, does $\boldsymbol{\psi}(0, \mathbf{y}) = \mathbf{f}(\mathbf{y})$ hold? If $h = 0$, the increment equations yield

$$h = 0 \quad \Rightarrow \quad \mathbf{k}_1 = \cdots = \mathbf{k}_s = \mathbf{f}(\mathbf{y}) . \quad \blacktriangleright \quad \boldsymbol{\psi}(0, \mathbf{y}) = \left(\sum_{i=1}^{s} b_i\right) \mathbf{f}(\mathbf{y}) .$$

Corollary 11.4.0.15. Consistent Runge-Kutta single step methods

A Runge-Kutta single step method according to Def. 11.4.0.11 is consistent (→ Def. 11.3.1.12) with the ODE $\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y})$, if and only if

$$\sum_{i=1}^{s} b_i = 1 .$$

Remark 11.4.0.16 (RK-SSM and quadrature rules) Note that in Def. 11.4.0.11 the coefficients C_i and b_i, $1 \in \{1, \ldots, s\}$, can be regarded as nodes and weights of a quadrature formula (→ Def. 7.2.0.1) on $[0, 1]$: apply the explicit Runge-Kutta single step method to the "ODE" $\dot{y} = f(t)$, $f \in C^0([0, 1])$, on $[]0, 1]$ with timestep $h = 1$ and initial value $y(0)$, with exact solution

$$\dot{y}(t) = f(t) , \quad y(0) = 0 \quad \Rightarrow \quad y(t) = \int_0^t f(\tau) \, \mathrm{d}\tau .$$

Then the formulas of Def. 11.4.0.11 reduce to

$$y_1 = 0 + \sum_{i=1}^{s} b_i f(c_i) \approx \int_0^1 f(\tau)\,\mathrm{d}\tau\,.$$

Recall that the quadrature rule with these weights and nodes c_j will have order ≥ 1 ($\rightarrow$ Def. 7.4.1.1), if the weights add up to 1! ⌟

EXAMPLE 11.4.0.17 (Butcher schemes for some explicit RK-SSM [DR08, Sect. 11.6.1]) The following explicit Runge-Kutta single step methods are often mentioned in literature.

- Explicit Euler method (11.2.1.5):
$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array} \quad \succ \quad \text{order = 1}$$

- explicit trapezoidal method (11.4.0.8):
$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \quad \succ \quad \text{order = 2}$$

- explicit midpoint method (11.4.0.9):
$$\begin{array}{c|cc} 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \hline & 0 & 1 \end{array} \quad \succ \quad \text{order = 2}$$

- Classical 4th-order RK-SSM:
$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \hline & \frac{1}{6} & \frac{2}{6} & \frac{2}{6} & \frac{1}{6} \end{array} \quad \succ \quad \text{order = 4}$$

- Kutta's 3/8-method:
$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 \\ \frac{2}{3} & -\frac{1}{3} & 1 & 0 & 0 \\ 1 & 1 & -1 & 1 & 0 \\ \hline & \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8} \end{array} \quad \succ \quad \text{order = 4}$$

Hosts of (explicit) Runge-Kutta methods can be found in the literature, see for example the Wikipedia page. They are stated in the form of Butcher schemes (11.4.0.13) most of the time. ⌟

Remark 11.4.0.18 (Construction of higher order Runge-Kutta single step methods) Runge-Kutta single step methods of order $p > 2$ are not found by bootstrapping as in Ex. 11.4.0.6, because the resulting methods would have quite a lot of stages compared to their order.

Rather one derives order conditions yielding large non-linear systems of equations for the coefficients a_{ij} and b_i in Def. 11.4.0.11, see [DB02, Sect .4.2.3] and [HLW06, Ch. III]. This approach is similar to the construction of a Gauss quadrature rule in Ex. 7.4.2.2. Unfortunately, the systems of equations are

very difficult to solve and no universal recipe is available. Nevertheless, through massive use of symbolic computation, explicit Runge-Kutta methods of order up to 19 have been constructed in this way. ⌟

Remark 11.4.0.19 ("Butcher barriers" for explicit RK-SSM) The following table gives lower bounds for the number of stages needed to achieve order p for an explicit Runge-Kutta method.

order p	1	2	3	4	5	6	7	8	≥ 9
minimal no. s of stages	1	2	3	4	6	7	9	11	$\geq p+3$

No general formula is has been discovered. What is known is that for explicit Runge-Kutta single step methods according to Def. 11.4.0.11

$$\text{order } p \quad \leq \quad \text{number } s \text{ of stages of RK-SSM}$$

⌟

Supplementary literature. Runge-Kutta methods are presented in every textbook covering numerical integration: [DR08, Sect. 11.6], [Han02, Ch. 76], [QSS00, Sect. 11.8].

Review question(s) 11.4.0.20 (Explicit Runge-Kutta single-step methods)

(Q11.4.0.20.A) How many parameters describe a *consistent* 2-stage explicit Runge-Kutta method for the autonomous ODE $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$?

Definition 11.4.0.11. Explicit Runge-Kutta method

For $b_i, a_{ij} \in \mathbb{R}$, $c_i := \sum_{j=1}^{i-1} a_{ij}$, $i,j = 1,\ldots,s$, $s \in \mathbb{N}$, an **s-stage explicit Runge-Kutta single step method** (RK-SSM) for the ODE $\dot{\mathbf{y}} = \mathbf{f}(t,\mathbf{y})$, $\mathbf{f}: \Omega \to \mathbb{R}^N$, is defined by ($\mathbf{y}_0 \in D$)

$$\mathbf{k}_i := \mathbf{f}(t_0 + c_i h, \mathbf{y}_0 + h\sum_{j=1}^{i-1} a_{ij}\mathbf{k}_j)\,, \quad i = 1,\ldots,s \quad , \quad \mathbf{y}_1 := \mathbf{y}_0 + h\sum_{i=1}^{s} b_i \mathbf{k}_i\,.$$

The vectors $\mathbf{k}_i \in \mathbb{R}^N$, $i = 1,\ldots,s$, are called **increments**, $h > 0$ is the size of the timestep.

(Q11.4.0.20.B) Recall that by "autonomization" the initial value problem

$$\dot{\mathbf{y}} = \mathbf{f}(t,\mathbf{y}) \quad , \quad \mathbf{y}(t_0) = \mathbf{y}_0\,, \tag{11.4.0.21}$$

with $\mathbf{f}: I \times D \to \mathbb{R}^N$ can be converted into the equivalent IVP for the extended state $\mathbf{z} = [z_1,\ldots,z_N,z_{N+1}]^\top := [\mathbf{y}\ t]^\top \in \mathbb{R}^{N+1}$:

$$\dot{\mathbf{z}} = \mathbf{g}(\mathbf{z})\,, \quad \mathbf{g}(\mathbf{z}) := \begin{bmatrix} \mathbf{f}(z_{N+1}, [z_1,\ldots,z_N]^\top) \\ 1 \end{bmatrix} \quad , \quad \mathbf{z}(0) = \begin{bmatrix} \mathbf{y}_0 \\ t_0 \end{bmatrix}. \tag{11.4.0.22}$$

Let us apply the same 2-stage explicit Runge-Kutta method to (11.4.0.21) and (11.4.0.22). When will both approaches produce the same sequence of states $\mathbf{y}_k \in D$?

(Q11.4.0.20.C) Formulate a generic 2-stage explicit Runge-Kutta method for the autonomous *second-order* ODE $\ddot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$, $\mathbf{f}: D \subset \mathbb{R}^N \to \mathbb{R}^N$.

Hint. Apply a standard 2-stage explicit Runge-Kutta method after transformation to an equivalent first-order ODE.

△

11.5 Adaptive Stepsize Control

Video tutorial for Section 11.5: Adaptive Stepsize Control: (32 minutes) Download link, tablet notes

Section 7.6, in the context of numerical quadrature, teaches an a-posteriori way to adjust the mesh underlying a composite quadrature rule to the integrand: During the computation we estimate the local quadrature error by comparing the approximations obtained by using quadrature formulas of different order. The same policy for adapting the integration mesh is very popular in the context of numerical integration, too. Since the size $h_k := t_{k+1} - t_k$ of the cells of the temporal mesh is also called the timestep size, this kind of a-posteriori mesh adaptation is also known as stepsize control.

11.5.1 The Need for Timestep Adaptation

EXAMPLE 11.5.1.1 (Oregonator reaction) Chemical reaction kinetics is a field where ODE based models are very common. This example presents a famous reaction with extremely abrupt dynamics. Refer to [Han02, Ch. 62] for more information about the ODE-based modelling of kinetics of chemical reactions.

This is a apecial case of an "oscillating" Zhabotinski-Belousov reaction [Gra02]:

$$\begin{aligned} \mathrm{BrO_3^-} + \mathrm{Br^-} &\mapsto \mathrm{HBrO_2} \\ \mathrm{HBrO_2} + \mathrm{Br^-} &\mapsto \mathrm{Org} \\ \mathrm{BrO_3^-} + \mathrm{HBrO_2} &\mapsto 2\,\mathrm{HBrO_2} + \mathrm{Ce(IV)} \\ 2\,\mathrm{HBrO_2} &\mapsto \mathrm{Org} \\ \mathrm{Ce(IV)} &\mapsto \mathrm{Br^-} \end{aligned} \tag{11.5.1.2}$$

By the laws of reaction kinetics of physical chemistry from (11.5.1.2) we can extract the following (system of) ordinary differential equation(s) for the concentrations of the different compounds:

$$\begin{aligned} y_1 := c(\mathrm{BrO_3^-}): \quad \dot{y}_1 &= -k_1 y_1 y_2 - k_3 y_1 y_3 \,, \\ y_2 := c(\mathrm{Br^-}): \quad \dot{y}_2 &= -k_1 y_1 y_2 - k_2 y_2 y_3 + k_5 y_5 \,, \\ y_3 := c(\mathrm{HBrO_2}): \quad \dot{y}_3 &= k_1 y_1 y_2 - k_2 y_2 y_3 + k_3 y_1 y_3 - 2k_4 y_3^2 \,, \\ y_4 := c(\mathrm{Org}): \quad \dot{y}_4 &= k_2 y_2 y_3 + k_4 y_3^2 \,, \\ y_5 := c(\mathrm{Ce(IV)}): \quad \dot{y}_5 &= k_3 y_1 y_3 - k_5 y_5 \,, \end{aligned} \tag{11.5.1.3}$$

with (non-dimensionalized) reaction constants:

$$k_1 = 1.34\,, \quad k_2 = 1.6 \cdot 10^9\,, \quad k_3 = 8.0 \cdot 10^3\,, \quad k_4 = 4.0 \cdot 10^7\,, \quad k_5 = 1.0\,.$$

periodic chemical reaction ➽ Video 1, Video 2

These are results from highly accurate simulations with initial state $y_1(0) = 0.06$, $y_2(0) = 0.33 \cdot 10^{-6}$, $y_3(0) = 0.501 \cdot 10^{-10}$, $y_4(0) = 0.03$, $y_5(0) = 0.24 \cdot 10^{-7}$:

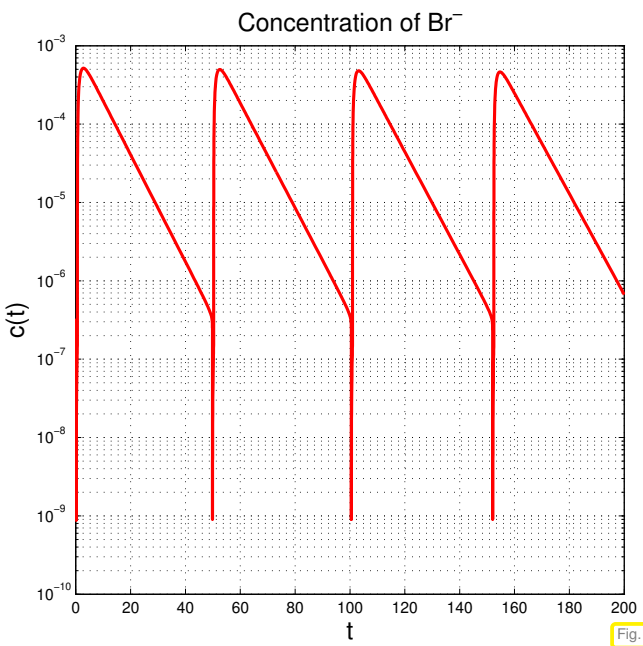


Fig. 429

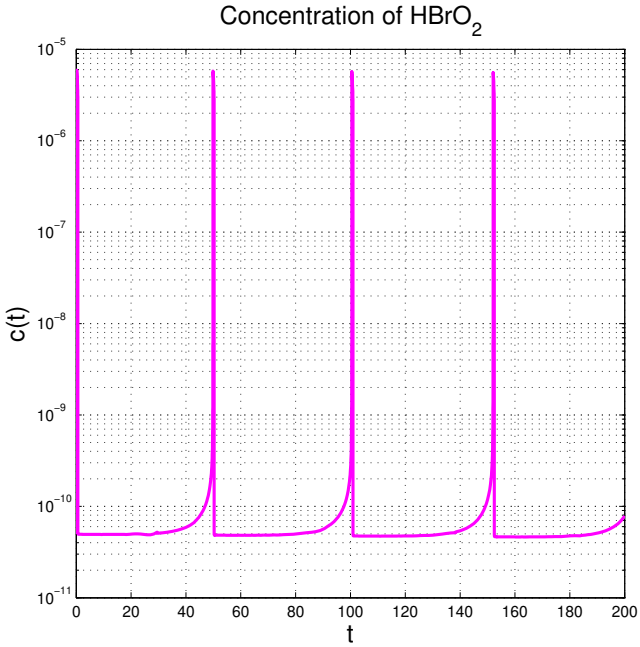


Fig. 430

We observe a strongly non-uniform behavior of the solution in time.

This is very common with evolutions arising from practical models (circuit models, chemical reaction models, mechanical systems)

⌟

EXAMPLE 11.5.1.4 (Blow-up)

We return to the "explosion ODE" of Ex. 11.1.3.20 and consider the scalar autonomous IVP:

$$\dot{y} = y^2 \,, \quad y(0) = y_0 > 0 \,.$$

$$\blacktriangleright \quad y(t) = \frac{y_0}{1 - y_0 t} \,, \quad t < 1/y_0 \,.$$

As we have seen a solution exists only for finite time and then suffers a **blow-up**, that is, $\lim_{t \to 1/y_0} y(t) = \infty$: $J(y_0) =]-\infty, 1/y_0]$!

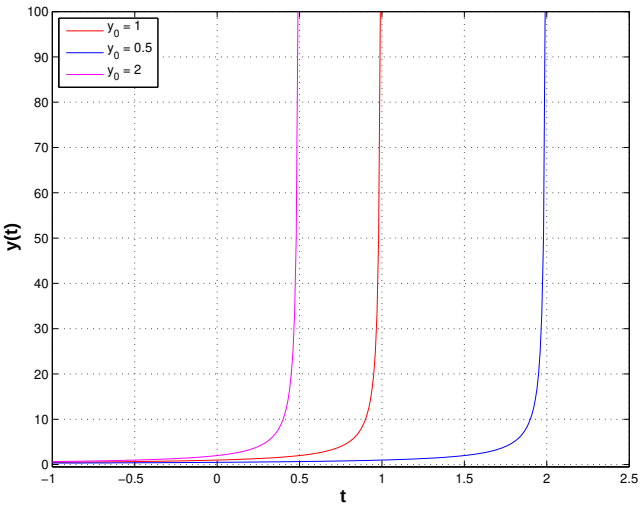


Fig. 431

How to choose temporal mesh $\{t_0 < t_1 < \cdots < t_{N-1} < t_N\}$ for single step method in case $J(y_0)$ is not known, even worse, if it is not clear a priori that a blow up will happen?

Just imagine: what will result from equidistant explicit Euler integration (11.2.1.5) applied to the above IVP?

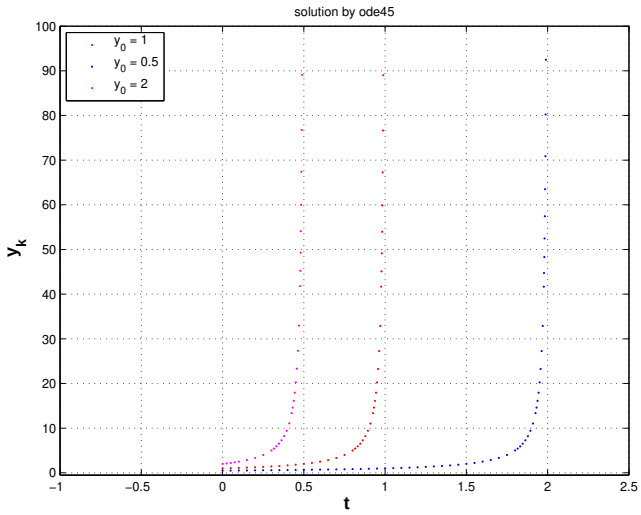


Fig. 432

A preview: There are single-step methods that can detect and resolve blow-ups!

◁ Simulations conducted with the numerical integrator **Ode45** introduced in § 11.5.3.3.

We observe that **Ode45** manages to reduce stepsize more and more as it approaches the singularity of the solution! How can it accomplish this feat!

11.5.2 Local-in-Time Stepsize Control

We identify as key challenge (discussed for autonomous ODEs below):

How to choose a *good temporal mesh* $\{0 = t_0 < t_1 < \cdots < t_{M-1} < t_M\}$
for a given single step method applied to a concrete IVP?

What does "good" mean ?

Be efficient! Be reliable!

Stepsize adaptation for single step methods

Objective: M as small as possible **&** $\max_{k=1,\ldots,N} \|\mathbf{y}(t_k) - \mathbf{y}_k\| < \mathtt{TOL}$ or $\|\mathbf{y}(T) - \mathbf{y}_M\| < \mathtt{TOL}$, $\mathtt{TOL}$ = tolerance

Policy: Try to curb/balance one-step error by

- ✦ adjusting *current* stepsize h_k,
- ✦ predicting suitable *next* timestep h_{k+1}

} local-in-time stepsize control

Tool: Local-in-time one-step error estimator (*a posteriori*, based on $\mathbf{y}_k, h_{k-1}$)

Why do we embrace local-in-time timestep control (based on estimating only the one-step error)**?** One could raise a serious objection: If a small time-local error in a single timestep leads to large error $\|\mathbf{y}_k - \mathbf{y}(t_k)\|$ at later times, then local-in-time timestep control is powerless about it and will not even notice!

Nevertheless, local-in-time timestep control is used almost exclusively,

- ☞ because we do not want to discard past timesteps, which could amount to tremendous waste of computational resources,
- ☞ because it is inexpensive and it works for many practical problems,
- ☞ because there is no reliable method that can deliver guaranteed accuracy for general IVP.

§11.5.2.2 (Local-in-time error estimation) We "recycle" heuristics already employed for adaptive quadrature, see Section 7.6, § 7.6.0.10. There we tried to get an idea of the local quadrature error by comparing two approximations of different order. Now we pursue a similar idea over a single timestep.

Idea: *Estimation of one-step error*

Compare results for two discrete evolutions $\boldsymbol{\Psi}^h$, $\widetilde{\boldsymbol{\Psi}}^h$ of different order over *current timestep* h:

If $\mathrm{Order}(\widetilde{\boldsymbol{\Psi}}) > \mathrm{Order}(\boldsymbol{\Psi})$, then we expect (for small h!)

$$\underbrace{\boldsymbol{\Phi}^h\mathbf{y}_k - \boldsymbol{\Psi}^h\mathbf{y}_k}_{\text{one-step error}} \approx \mathtt{EST}_k := \widetilde{\boldsymbol{\Psi}}^h\mathbf{y}_k - \boldsymbol{\Psi}^h\mathbf{y}_k\,. \tag{11.5.2.3}$$

Heuristics for concrete $h > 0$ A computable quantity!

§11.5.2.4 ((Crude) local timestep control) We take for granted the availability of a local error estimate $\mathtt{EST}_k$ that we have computed for a current stepsize h. We specify target values $\mathtt{ATOL} > 0$, $\mathtt{RTOL} > 0$ of absolute and relative tolerances to be met by the local error and implement the following policy:

absolute tolerance

$$\text{Compare}\quad \begin{array}{l}\mathtt{EST}_k \leftrightarrow \mathtt{ATOL}\\ \mathtt{EST}_k \leftrightarrow \mathtt{RTOL}\|\mathbf{y}_k\|\end{array} \quad\succ\quad \text{Reject/accept current step} \tag{11.5.2.5}$$

relative tolerance

Both tolerances $\mathtt{RTOL} > 0$ and $\mathtt{ATOL} > 0$ have to be supplied by the user of the adaptive algorithm. The absolute tolerance is usually chosen significantly smaller than the relative tolerance and merely serves as a safeguard against non-termination in case $\mathbf{y}_k \approx \mathbf{0}$. For a similar use of absolute and relative tolerances see Section 8.2.3, which deals with termination criteria for iterations, in particular (8.2.3.3).

☞ Simple accept/reject algorithm:

$\mathtt{EST}_k < \max\{\mathtt{ATOL}, \|\mathbf{y}_k\|\mathtt{RTOL}\}$:	Accept current step: • *Advance* by one timestep (stepsize h), • use larger stepsize (αh with some $\alpha > 1$) for next step $(*)$
$\mathtt{EST}_k > \max\{\mathtt{ATOL}, \|\mathbf{y}_k\|\mathtt{RTOL}\}$:	Reject current step: • *Repeat* current timestep • with smaller stepsize $< h$, e.g., $\frac{1}{2}h$

The rationale behind the adjustment of the timestep size in $(*)$ is the following: if the current stepsize guarantees sufficiently small one-step error, then it might be possible to obtain a still acceptable one-step error with a larger timestep, which would enhance efficiency (fewer timesteps for total numerical integration). This should be tried, since timestep control will usually provide a safeguard against undue loss of accuracy.

The following C++ code implements a wrapper function `odeintadapt()` for a general adaptive single-step method according to the policy outlined above. The arguments are meant to pass the following information:

- `Psilow`, `Psihigh`: functors passing discrete evolution operators for autonomous ODE of different order, type `@(y,h)`, expecting a state (usually a column vector) as first argument, and a stepsize as second,

- `T`: final time $T > 0$,
- `y0`: initial state $\mathbf{y}_0$,
- `h0`: stepsize h_0 for the first timestep
- `reltol`, `abstol`: relative and absolute tolerances, see (11.5.2.5),
- `hmin`: minimal stepsize, timestepping terminates when stepsize control $h_k < h_{\min}$, which is relevant for detecting blow-ups or collapse of the solution.

C++ code 11.5.2.6: Simple local stepsize control for single step methods ↪ GITLAB

```
2  // Auxiliary function:  default norm for an EIGEN vector type
3  template <class State>
4  double _norm(const State &y) { return y.norm(); }
5
6  // Adaptive numerical integrator based on local-in-time stepsize control
7  template <class DiscEvolOp, class State,
8            class NormFunc = decltype(_norm<State>)>
9  std::vector<std::pair<double, State>> odeintadapt(
10     DiscEvolOp &&Psilow, DiscEvolOp &&Psihigh, const State &y0, double T,
11     double h0, double reltol, double abstol, double hmin,
12     NormFunc &norm = _norm<State>) {
13   double t = 0;   // initial time t_0 = 0
14   State y = y0;   // current state
15   double h = h0;  // timestep to start with
16   // vector of times/computed states:  (t_k, y_k)_k
17   std::vector<std::pair<double, State>> states;
18   states.push_back({t, y});  // initial time and state
19   // Main timestepping loop
20   while ((states.back().first < T) && (h >= hmin)) {  //
21     State yh = Psihigh(h, y); // high-order discrete evolution Ψ̃^h
22     State yH = Psilow(h, y); // low-order discrete evolution Ψ^h
23     double est = norm(yH - yh);  // local error estimate EST_k
24     if (est < std::max(reltol * norm(y), abstol)) {  // step accepted
25       y = yh;           // use high order approximation
26       t = t + std::min(T - t, h);  // next time t_k
27       states.push_back({t, y});    // store approximate state
28       h = 1.1 * h;  // try with increased stepsize
29     } else {        // step rejected
30       h = h / 2;    // try with half the stepsize
31     }
32     // Numerical integration has ground to a halt !
33     if (h < hmin) {
34       std::cerr << "Warning: Failure at t=" << states.back().first
35         << ". Unable to meet integration tolerances without reducing "
36      << "the step size below the smallest value allowed (" << hmin
37         << ") at time t = " << t << "." << std::endl;
38     }
39   } // end main loop
40   return states; // ok thanks to return value optimization (RVO)
41 }
```

Comments on Code 11.5.2.6:

- line 20: check whether final time is reached or timestepping has ground to a halt ($h_k < h_{\min}$).
- line 21, 22: advance state by low and high order integrator.
- line 23: compute norm of estimated error, see (11.5.2.3).

- line 24: make comparison (11.5.2.5) to decide whether to accept or reject local step.
- line 27, 28: step accepted, update state and current time and suggest 1.1 times the current stepsize for next step.
- line 30 step rejected, try again with half the stepsize.
- Return value is a vector of pairs consisting of
  - times `t` $\leftrightarrow$ temporal mesh $t_0 < t_1 < t_2 < \ldots < t_N < T$, where $t_N < T$ indicated premature termination (collapse, blow-up),
  - states `y` $\leftrightarrow$ sequence $(\mathbf{y}_k)_{k=0}^N$.

⌟

Remark 11.5.2.7 (Choice of norm) In Code 11.5.2.6 the norm underlying timestep control is passed through a functor. This is related to the important fact that that norm has to be chosen in a problem-dependent way. For instance, in the case of systems of ODEs different components of the state vector $\mathbf{y}$ may correspond to different physical quantities. Hence, if we used the Euclidean norm $\|\cdot\| = \|\cdot\|_2$, the choice of physical units might have a strong impact on the selection of timesteps, which is clearly not desirable and destroys the scale-invariance of the algorithm, *cf.* Rem. 2.3.3.11. ⌟

Remark 11.5.2.8 (Estimation of "wrong" error?) We face the same conundrum as in the case of adaptive numerical quadrature, see Rem. 7.6.0.17:

! By the heuristic considerations, see (11.5.2.3) it seems that EST_k measures the one-step error for the low-order method Ψ and that we should use $\mathbf{y}_{k+1} = \Psi^{h_k}\mathbf{y}_k$, if the timestep is accepted.

However, it would be foolish not to use the better value $\mathbf{y}_{k+1} = \widetilde{\Psi}^{h_k}\mathbf{y}_k$, since it is available for free. This is what is done in every implementation of adaptive methods, also in Code 11.5.2.6, and this choice can be justified by control theoretic arguments [DB02, Sect. 5.2]. ⌟

EXPERIMENT 11.5.2.9 (Simple adaptive stepsize control) We test adaptive timestepping routine from Code 11.5.2.6 for a scalar IVP and compare the estimated local error and true local error.

- IVP for ODE $\dot{y} = \cos(\alpha y)^2$, $\alpha > 0$, solution $y(t) = \arctan(\alpha(t-c))/\alpha$ for $y(0) \in]-\pi/2, \pi/2[$
- Simple adaptive timestepping based on explicit Euler (11.2.1.5) and explicit trapezoidal rule (11.4.0.8)

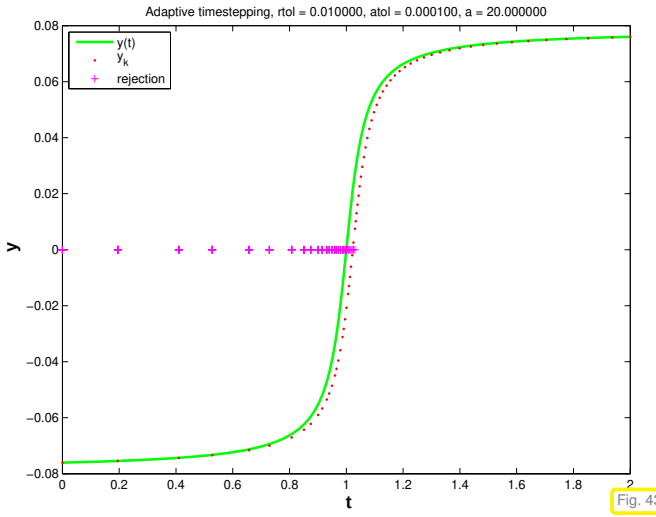


Fig. 433

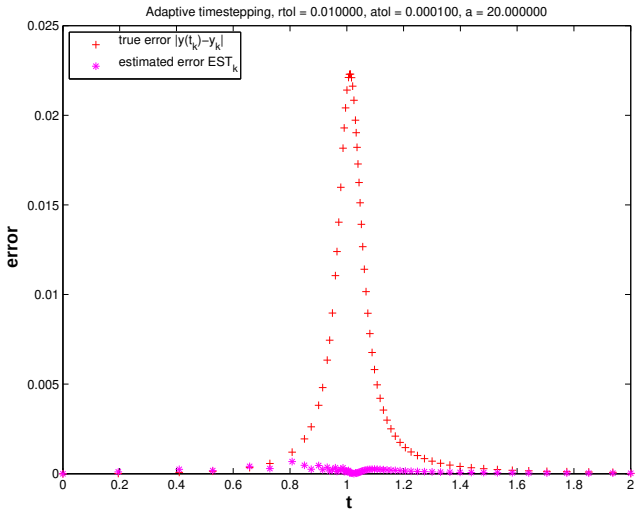


Fig. 434

Statistics: `66 timesteps, 131 rejected timesteps`

Observations:

- ☞ Adaptive timestepping well resolves local features of solution $y(t)$ at $t = 1$
- ☞ Estimated error (an estimate for the one-step error) and true error are **not** related! To understand this recall Rem. 11.5.2.8.

⌟

EXPERIMENT 11.5.2.10 (Gain through adaptivity → Exp. 11.5.2.9) In this experiment we want to explore whether adaptive timestepping is worth while, as regards reduction of computational effort without sacrificing accuracy.

We retain the simple adaptive timestepping from previous experiment Exp. 11.5.2.9 and also study the same IVP.

New: initial state $y(0) = 0$!

Now we examine the dependence of the maximal discretization error in mesh points on the computational effort. The latter is proportional to the number of timesteps.

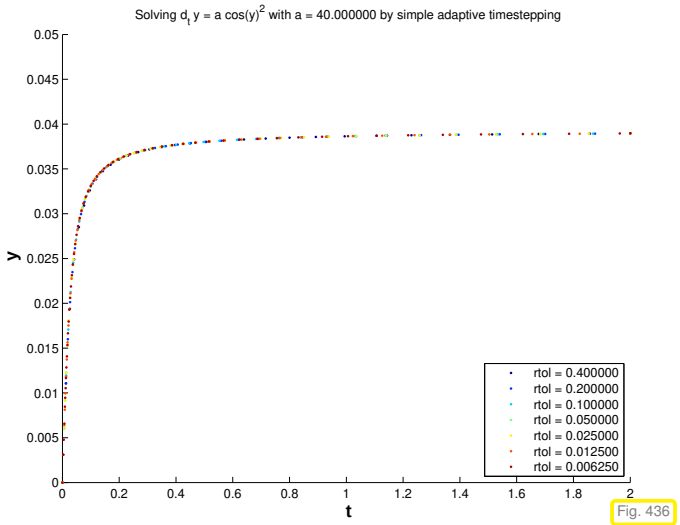


Fig. 435

Solutions $(\mathbf{y}_k)_k$ for different values of `rtol`

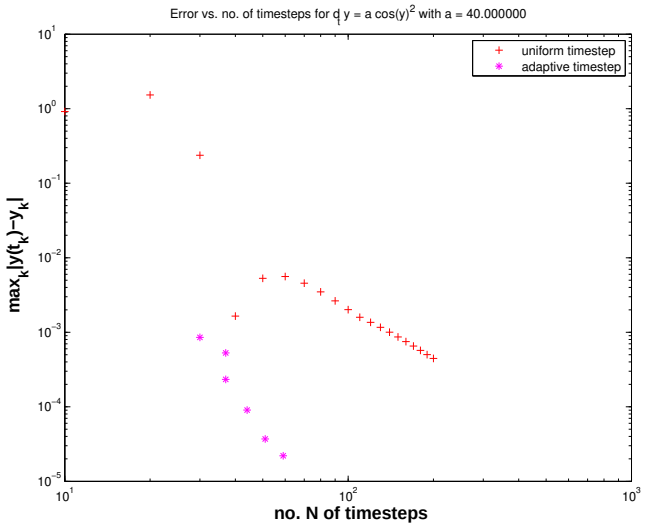


Fig. 436

Error vs. computational effort

Observations:

- ☞ Adaptive timestepping achieves much better accuracy for a fixed computational effort.

⌟

EXPERIMENT 11.5.2.11 ("Failure" of adaptive timestepping → Exp. 11.5.2.10)

Same ODE and simple adaptive timestepping as in previous experiment Exp. 11.5.2.10.

$$\dot{y} = \cos^2(\alpha y) \quad \Rightarrow \quad y(t) = \arctan(\alpha(t-c))/\alpha \,, y(0) \in]-\frac{\pi}{2\alpha}, -\frac{\pi}{2\alpha}[\,,$$

for $\alpha = 40$.

Now: initial state $y(0) = -0.0386 \approx \frac{\pi}{2\alpha}$ as in Exp. 11.5.2.9

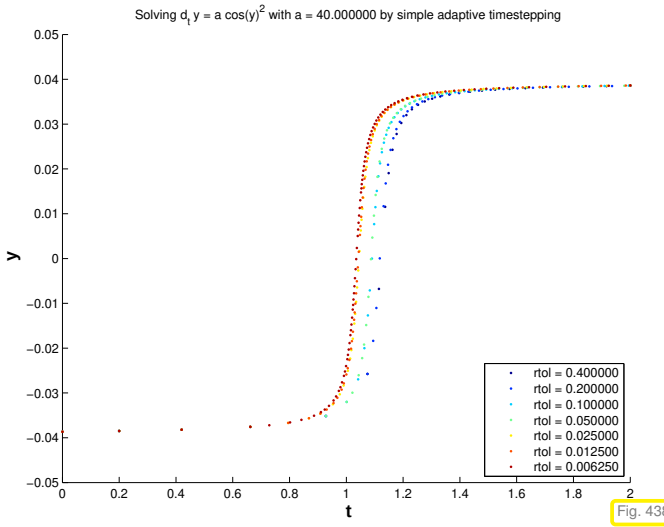


Fig. 437

Solutions $(\mathbf{y}_k)_k$ for different values of `rtol`

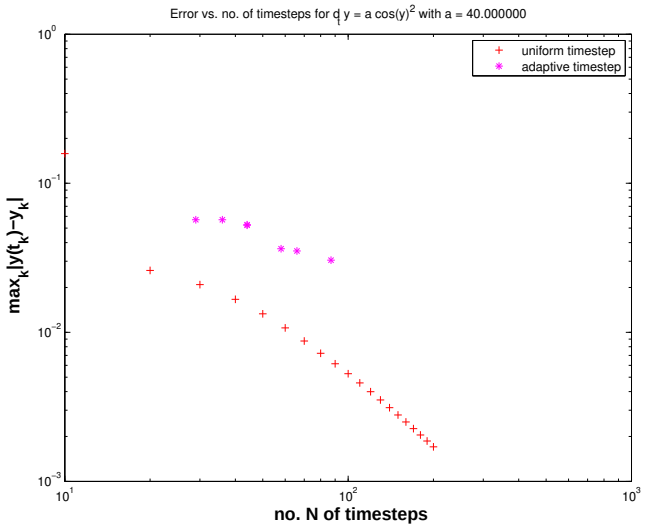


Fig. 438

Error vs. computational effort

Observations:

☞ Adaptive timestepping leads to larger errors at the same computational cost as uniform timestepping**!**

Explanation: the position of the steep step of the solution has a sensitive dependence on an initial value, if $y(0) \approx \frac{\pi}{2\alpha}$:

$$y(t) = \frac{1}{\alpha}\arctan(\alpha(t + \tan(y_0/\alpha))) \;, \quad \text{step at} \quad \approx -\tan(y_0/\alpha) \;.$$

Hence, small local errors in the initial timesteps will lead to large errors at around time $t \approx 1$. The stepsize control is mistaken in condoning these small one-step errors in the first few steps and, therefore, incurs huge errors later.

However, the perspective of backward error analysis (→ § 1.5.5.18) rehabilitates adaptive stepsize control in this case: it gives us a numerical solution that is very close to the exact solution of the ODE with slightly perturbed initial state y_0. ⌟

§11.5.2.12 (Refined local stepsize control → [DR08, Sect. 11.7]) The above algorithm (Code 11.5.2.6) is simple, but the rule for increasing/shrinking of timestep "squanders" the information contained in EST_k : `TOL`:

More ambitious goal **!**	When $\mathrm{EST}_k > \mathtt{TOL}$:	stepsize adjustment better $h_k =$ **?**
	When $\mathrm{EST}_k < \mathtt{TOL}$:	stepsize prediction good $h_{k+1} =$ **?**

Assumption: At our disposal are two discrete evolutions:

- Ψ with order$(\Psi) = p$ (➙ "low order" single step method)
- $\widetilde{\Psi}$ with order$(\widetilde{\Psi}) > p$ (➙ "higher order" single step method)

These are the same building blocks as for the simple adaptive strategy employed in Code 11.5.2.6 (passed as arguments `Psilow`, `Psihigh` there).

According to (11.3.2.22) we can expect the following asymptotic decay of the one-step errors for $h \to 0$:

$$\begin{aligned} \Psi^{h_k}\mathbf{y}(t_k) - \Phi^{h_k}\mathbf{y}(t_k) &= ch^{p+1} + O(h_k^{p+2}) \;, \\ \widetilde{\Psi}^{h_k}\mathbf{y}(t_k) - \Phi^{h_k}\mathbf{y}(t_k) &= O(h_k^{p+2}) \;, \end{aligned} \tag{11.5.2.13}$$

with some (unknown) constant $c > 0$. Why h^{p+1}? Remember the estimate (11.3.2.15) from the error analysis of the explicit Euler method: we also found $O(h_k^2)$ there for the one-step error of a single step method of order 1.

Heuristic reasoning: The timestep h_k is small ➢ "higher order terms" $O(h_k^{p+2})$ can be ignored.

$$\begin{array}{l} \Psi^{h_k}\mathbf{y}(t_k) - \Phi^{h_k}\mathbf{y}(t_k) \doteq ch_k^{p+1} + O(h_k^{p+2}) , \\ \widetilde{\Psi}^{h_k}\mathbf{y}(t_k) - \Phi^{h_k}\mathbf{y}(t_k) \doteq O(h_k^{p+2}) . \end{array} \quad \Rightarrow \quad \boxed{\mathrm{EST}_k \doteq ch_k^{p+1}} . \tag{11.5.2.14}$$

✎ notation: $\doteq$ equality up to higher order terms in h_k

$$\mathrm{EST}_k \doteq ch_k^{p+1} \quad \Rightarrow \quad c \doteq \frac{\mathrm{EST}_k}{h_k^{p+1}} . \tag{11.5.2.15}$$

Available in algorithm, see (11.5.2.3)

For the sake of *accuracy* (demands "$\mathrm{EST}_k < \mathrm{TOL}$") & *efficiency* (favors "$>$") we aim for

$$\mathrm{EST}_k \overset{!}{=} \mathrm{TOL} := \max\{\mathrm{ATOL}, \|\mathbf{y}_k\|\mathrm{RTOL}\} . \tag{11.5.2.16}$$

What timestep h_* can actually achieve (11.5.2.16), if we "believe" (heuristics!) in (11.5.2.14) (and, therefore, in (11.5.2.15))?

$$\text{(11.5.2.15) \& (11.5.2.16)} \quad \Rightarrow \quad \mathrm{TOL} = \frac{\mathrm{EST}_k}{h_k^{p+1}} h_*^{p+1} . \tag{11.5.2.17}$$

"'Optimal timestep": (stepsize prediction)

$$\boxed{h_* = h \sqrt[p+1]{\frac{\mathrm{TOL}}{\mathrm{EST}_k}}} . \tag{11.5.2.18}$$

The proposed timestep size h_* will be used in both cases:

(Rejection of current timestep): In case $\mathrm{EST}_k > \mathrm{TOL}$ ➢ *repeat step* with stepsize h_*.

(Acceptance of current timstep): If $\mathrm{EST}_k \leq \mathrm{TOL}$ ➢ use h_* as stepsize for *next step*.

C++ code 11.5.2.19: Refined local stepsize control for single step methods ↬ GITLAB

```
// Auxiliary function:  default norm for an EIGEN vector type
template <class State>
double defaultnorm(const State &y) { return y.norm(); }
// Auxilary struct to hold user options
struct Odeintssctrl_options {
  double T;       // terminal time
  double h0;      // initial time step
  double reltol;  // norm-relative error tolerance
  double abstol;  // absolute error tolerance
  double hmin;    // smallest allowed time step
} __attribute__((aligned(64)));
// Adaptive single-step integrator
template <class DiscEvolOp, class State,
          class NormFunc = decltype(defaultnorm<State>)>
std::vector<std::pair<double, State>> odeintssctrl(
    DiscEvolOp &&Psilow, unsigned int p, DiscEvolOp &&Psihigh, const State &y0,
    const Odeintssctrl_options &opt,
```

```
19      NormFunc &norm = defaultnorm<State>) {
20    double t = 0;           // initial time t_0 = 0
21    State y = y0;           // current state, initialized here
22    double h = opt.h0;  // timestep to start with
23    // Array for returning pairs of times/states (t_k, y_k)_k
24    std::vector<std::pair<double, State>> states;
25    states.push_back({t, y}); // initial time and state
26    // Main timestepping loop
27    while ((states.back().first < opt.T) && (h >= opt.hmin)) {  //
28      State yh = Psihigh(h, y);  // high-order discrete evolution Ψ̃^h
29      State yH = Psilow(h, y);  // low-order discrete evolution Ψ^h
30      const double est = norm(yH - yh);  // ↔ EST_k
31      const double tol = std::max(opt.reltol * norm(y), opt.abstol);  // effective
            tolerance
32      // Optimal stepsize according to (11.5.2.18)
33      if (est < tol) {  // step accepted
34        // store next approximate state
35        states.push_back({t = t + std::min(opt.T - t, h), y = yh});
36      }
37      // New timestep size according to (11.5.2.18)
38      h *= std::max(0.5, std::min(2., 0.9 * std::pow(tol / est, 1. / (p + 1))));  //
39      if (h < opt.hmin) {
40        std::cerr << "Warning: Failure at t=" << states.back().first
41            << ". Unable to meet integration tolerances without reducing the step"
42            << " size below the smallest value allowed (" << opt.hmin
43            << ") at time t = " << t << "." << std::endl;
44      }
45    } // end main loop
46    return states;
47  }
```

Comments on Code 11.5.2.19 (see comments on Code 11.5.2.6 for more explanations):

- Input arguments as for Code 11.5.2.6, except for $\mathrm{p} \hat{=}$ order of lower order discrete evolution.
- line 38: compute presumably better local stepsize according to (11.5.2.18),
- line 33: decide whether to repeat the step or advance,
- line 33: extend output arrays if current step has not been rejected.

11.5.3 Embedded Runge-Kutta Methods

For higher order RK-SSM with a considerable number of stages computing different sets of increments ($\rightarrow$ Def. 11.4.0.11) for two methods of different order just for the sake of local-in-time stepsize control would mean undue effort. This makes the following embedding idea attractive:

Embedding idea for RK-SSM

Use two RK-SSMs based on the same increments, that is, built with the same coefficients a_{ij}, but different weights b_i, see Def. 11.4.0.11 for the formulas, and *different orders* p and $p+1$.

Augmented Butcher scheme for embedded explicit Runge-Kutta methods ▷

(The lower-order scheme has weights $\widehat{b}_i$.)

$$\begin{array}{c|c} \mathbf{c} & \mathfrak{A} \\ \hline & \mathbf{b}^T \\ \hline & \widehat{\mathbf{b}}^T \end{array} := \begin{array}{c|ccc} c_1 & a_{11} & \cdots & a_{1s} \\ \vdots & \vdots & & \vdots \\ c_s & a_{s1} & \cdots & a_{ss} \\ \hline & b_1 & \cdots & b_s \\ \hline & \widehat{b}_1 & \cdots & \widehat{b}_s \end{array} .$$

EXAMPLE 11.5.3.2 (Commonly used embedded explicit Runge-Kutta methods) The following two embedded RK-SSM, presented in the form of their extended Butcher schemes, provided single step methods of orders 4 & 5 [HNW93, Tables 4.1 & 4.2].

$$\begin{array}{c|ccccc} 0 & & & & & \\ \frac{1}{3} & \frac{1}{3} & & & & \\ \frac{1}{3} & \frac{1}{6} & \frac{1}{6} & & & \\ \frac{1}{2} & \frac{1}{8} & 0 & \frac{3}{8} & & \\ 1 & \frac{1}{2} & 0 & -\frac{3}{2} & 2 & \\ \hline y_1 & \frac{1}{6} & 0 & 0 & \frac{2}{3} & \frac{1}{6} \\ \hline \widehat{y}_1 & \frac{1}{10} & 0 & \frac{3}{10} & \frac{2}{5} & \frac{1}{5} \end{array}$$

Merson's embedded RK-SSM

$$\begin{array}{c|ccccc} 0 & & & & & \\ \frac{1}{2} & \frac{1}{2} & & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & & \\ 1 & 0 & 0 & 1 & & \\ \frac{3}{4} & \frac{5}{32} & \frac{7}{32} & \frac{13}{32} & -\frac{1}{32} & \\ \hline y_1 & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} & \\ \hline \widehat{y}_1 & -\frac{1}{2} & \frac{7}{3} & \frac{7}{3} & \frac{13}{6} & -\frac{16}{3} \end{array}$$

Fehlberg's embedded RK-SSM

§11.5.3.3 (EIGEN-compatible adaptive explicit embedded Runge-Kutta integrator) An implementation of an explicit embedded Runge-Kutta single-step method with adaptive stepsize control for solving an autonomous IVP is provided by the utility class **Ode45** ⇝ GITLAB: The class name already indicates the orders of the pair of single step methods used for adaptive stepsize control:

Refer to the class implementation ⇝ GITLAB and the data members `Ode45::_mA`, `Ode45::_vb4`, `Ode45::_vb5` for the underlying coefficients in $\mathfrak{A}$, $\mathbf{b}$, and $\widehat{\mathbf{b}}$, as introduced in Section 11.5.3.

The class is templated with two type parameters:

```
template <class StateType,
class RhsType = std::function<StateType(const StateType &)>>
class Ode45 { ... };
```

(i) **StateType**: type for vectors in state space V, e.g. a fixed size vector type of EIGEN: `Eigen::Matrix<double,N,1>`, where `N` is an integer constant § 11.2.0.1.

(ii) **RhsType**: a functor type, see Section 0.3.3, for the right hand side function $\mathbf{f}$; must match **StateType**, default type provided.

The functor for the right hand side $\mathbf{f} : D \subset V \to V$ of the ODE $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$ is specified as an argument of the constructor. The single-step numerical integrator is invoked by the templated method

```
template<class NormFunc = decltype(_norm<StateType>)>
std::vector< std::pair<StateType, double>>
solve(const StateType & y0,double T,const NormFunc & norm =
   _norm<StateType>);
```

The following arguments have to be supplied:

1. `y0`: the initial value $\mathbf{y}_0$
2. `T`: the final time T, initial time $t_0 = 0$ is assumed, because the class can deal with autonomous ODEs only, recall § 11.1.3.3.
3. `norm`: a functor returning a suitable norm for a state vector. Defaults to EIGEN's maximum vector norm.

The method returns a vector of 2-tuples $(\mathbf{y}_k, t_k)$ (note the order!), $k = 0, \ldots, N$, of temporal mesh points t_k, $t_0 = 0$, $t_N = T$, see § 11.2.0.2, and approximate states $\mathbf{y}_k \approx \mathbf{y}(t_k)$, where $t \mapsto \mathbf{y}(t)$ stands for the exact solution of the initial value problem.

The arguments of `solve()` are not sufficient to control the behavior of the adaptive integrator. In addition, one can set data members of the data structure **Ode45.options** to configure an instance `ode45obj` of **Ode45**:

```
ode45obj.options.<option_you_want_to_set> = <value>;
```

In particular, the key data fields for adaptive timestepping are

- `rtol`: relative tolerance for error control (default is 10^{-6})
- `atol`: absolute tolerance for error control (default is 10^{-8})

For complete information about all control parameters and their default values see the **Ode45** class definition ➙ GITLAB.

The following self-explanatory code snippet uses the numerical integrator class **Ode45** for solving a scalar autonomous ODE.

C++ code 11.5.3.4: Invocation of adaptive embedded Runge-Kutta-Fehlberg integrator ➙ GITLAB

```
2  int main(int /*argc*/, char** /*argv*/) {
3    // Types to be used for a scalar ODE with state space ℝ
4    using StateType = double;
5    using RhsType = std::function<StateType(StateType)>;
6    // Logistic differential equation (11.1.2.2)
7    const RhsType f = [](StateType y) { return 5 * y * (1 - y); };
8    const StateType y0 = 0.2;  // Initial value
9    // Exact solution of IVP, see (11.1.2.3)
10   auto y = [y0](double t) { return y0 / (y0 + (1 - y0) * std::exp(-5 * t)); };
11   // State space ℝ, simple modulus supplies norm
12   auto normFunc = [](StateType x) { return std::fabs(x); };
13
14   // Invoke explicit Runge-Kutta method with stepsize control provided
15   // by a dedicated class implemented in ode45.h
16   Ode45<StateType, RhsType> integrator(f);
17   // Do the timestepping with adaptive strategy of Section 11.5
18   const std::vector<std::pair<StateType, double>> states =
19       integrator.solve(y0, 1.0, normFunc);
20   // Output information accumulation during numerical integration
```

```
21    integrator.options().do_statistics = true;
22    integrator.print();
23    // Output norm of discretization error in nodes of adaptively
          generated
24    // temporal mesh
25    std::cout << "Errors at points of temporal mesh:" << std::endl;
26    for (auto state : states) {
27      std::cout << "t = " << state.second << ", y = " << state.first
28                << ", |err| = " << fabs(state.first - y(state.second))
29                << std::endl;
30    }
31    return 0;
32  }
```

Remark 11.5.3.5 (Tolerances and accuracy) As we have learned in § 11.5.3.3 for objects of the class **Ode45** tolerances for the refined local stepsize control of § 11.5.2.12 can be specified by setting the member variables `options.rtol` and `options.atol`.

The possibility to pass tolerances to numerical integrators based on adaptive timestepping may tempt the user into believing that they allow to control the accuracy of the solutions. However, as is clear from § 11.5.2.12, these tolerances are solely applied to local error estimates and, inherently, have nothing to do with global discretization errors, see Exp. 11.5.2.9.

No global error control through local-in-time adaptive timestepping

The absolute/relative tolerances imposed for local-in-time adaptive timestepping do *not* allow to predict accuracy of solution!

EXAMPLE 11.5.3.7 (Adaptive timestepping for mechanical problem) We test the effect of adaptive stepsize control in MATLAB for the equations of motion describing the planar movement of a point mass in a conservative force field $\mathbf{x} \in \mathbb{R}^2 \mapsto F(\mathbf{x}) \in \mathbb{R}^2$: Let $t \mapsto \mathbf{y}(t) \in \mathbb{R}^2$ be the trajectory of point mass (in the plane).

$$\text{From Newton's law:} \quad \ddot{\mathbf{y}} = F(\mathbf{y}) := -\frac{2\mathbf{y}}{\|\mathbf{y}\|_2^2} \,. \tag{11.5.3.8}$$

acceleration force

As in Rem. 11.1.3.6 we can convert the second-order ODE (11.5.3.8) into an equivalent 1st-order ODE by introducing the velocity $\mathbf{v} := \dot{\mathbf{y}}$ as an extra solution component:

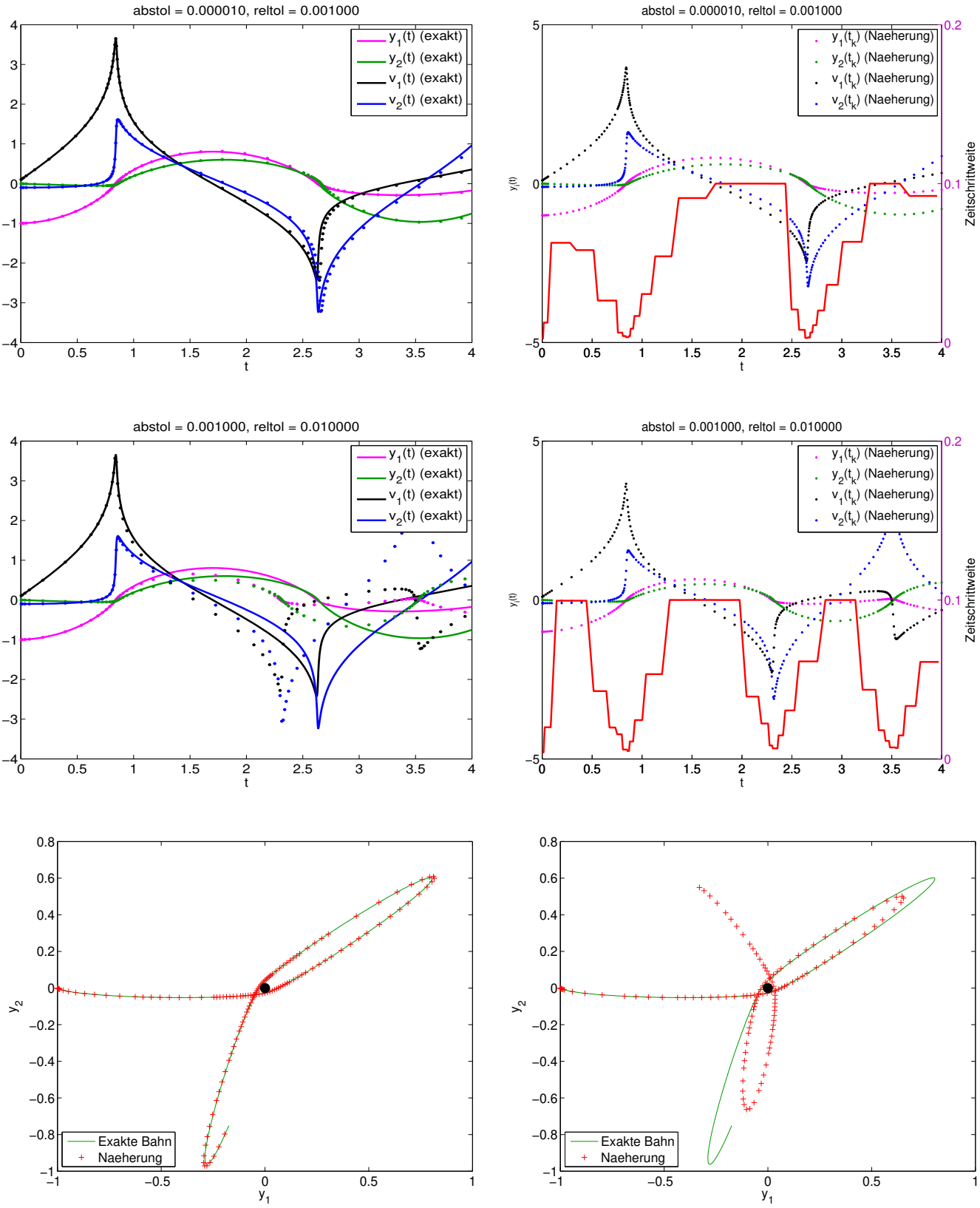
$$(11.5.3.8) \quad \Rightarrow \quad \begin{bmatrix} \dot{\mathbf{y}} \\ \dot{\mathbf{v}} \end{bmatrix} = \begin{bmatrix} \mathbf{v} \\ -\frac{2\mathbf{y}}{\|\mathbf{y}\|_2^2} \end{bmatrix} . \tag{11.5.3.9}$$

The following initial values used in the experiment:

$$\mathbf{y}(0) := \begin{bmatrix} -1 \\ 0 \end{bmatrix} \quad , \quad \mathbf{v}(0) := \begin{bmatrix} 0.1 \\ -0.1 \end{bmatrix}$$

Adaptive numerical integration with adaptive numerical integrator **Ode45** (*to* § 11.5.3.3) with

❶ `options.rtol = 0.001, options.atol = 1.0E-5,`
❷ `options.rtol = 0.01, options.atol = 1.0E-3,`



`reltol=0.001, abstol=1e-5` `reltol=0.01, abstol=1e-3`

Observations:

- ☞ Fast changes in solution components captured by adaptive approach through very small timesteps.
- ☞ Completely wrong solution, if tolerance reduced slightly.

In this example we face a rather sensitive dependence of the trajectories on initial states or intermediate states. Small perturbations at one instance in time can be have a massive impact on the solution at later

times. Local stepsize control is powerless about preventing this. ⌟

Review question(s) 11.5.3.10 (Adaptive timestep control)

(Q11.5.3.10.A) Explain how the blow-up of solutions of an initial-value problem can be captured by a single-step numerical integrator with adaptive stepsize control.

(Q11.5.3.10.B) Code 11.5.2.19 contains the line

```
h = h * std::max(0.5, std::min(2.0, std::pow(tol/est, 1.0/(p+1))));
```

Explain all parts and variables occurring in this expression.

△

Bibliography

[Ama83] H. Amann. *Gewöhnliche Differentialgleichungen*. 1st. Berlin: Walter de Gruyter, 1983 (cit. on pp. 764, 765, 771).

[DR08] W. Dahmen and A. Reusken. *Numerik für Ingenieure und Naturwissenschaftler*. Heidelberg: Springer, 2008 (cit. on pp. 765, 769–772, 778, 790, 796–798, 800–802, 810).

[Dea80] M.A.B. Deakin. "Applied catastrophe theory in the social and biological sciences". In: *Bulletin of Mathematical Biology* 42.5 (1980), pp. 647–679 (cit. on p. 766).

[DB02] P. Deuflhard and F. Bornemann. *Scientific Computing with Ordinary Differential Equations*. 2nd ed. Vol. 42. Texts in Applied Mathematics. New York: Springer, 2002 (cit. on pp. 761, 773, 793, 801, 808).

[Gra02] C.R. Gray. "An analysis of the Belousov-Zhabotinski reaction". In: *Rose-Hulman Undergraduate Math Journal* 3.1 (2002) (cit. on p. 803).

[HLW06] E. Hairer, C. Lubich, and G. Wanner. *Geometric numerical integration*. 2nd ed. Vol. 31. Springer Series in Computational Mathematics. Heidelberg: Springer, 2006 (cit. on pp. 761, 765, 801).

[HNW93] E. Hairer, S.P. Norsett, and G. Wanner. *Solving Ordinary Differential Equations I. Nonstiff Problems*. 2nd ed. Berlin, Heidelberg, New York: Springer-Verlag, 1993 (cit. on pp. 761, 813).

[HW11] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II. Stiff and Differential-Algebraic Problems*. Vol. 14. Springer Series in Computational Mathematics. Berlin: Springer-Verlag, 2011 (cit. on p. 761).

[Han02] M. Hanke-Bourgeois. *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens*. Mathematische Leitfäden. Stuttgart: B.G. Teubner, 2002 (cit. on pp. 764, 765, 768, 771, 778, 790, 802, 803).

[Het00] Herbert W. Hethcote. "The mathematics of infectious diseases". In: *SIAM Rev.* 42.4 (2000), pp. 599–653. DOI: 10.1137/S0036144500371907 (cit. on p. 767).

[NS02] K. Nipp and D. Stoffer. *Lineare Algebra*. 5th ed. Zürich: vdf Hochschulverlag, 2002 (cit. on p. 764).

[QSS00] A. Quarteroni, R. Sacco, and F. Saleri. *Numerical mathematics*. Vol. 37. Texts in Applied Mathematics. New York: Springer, 2000 (cit. on pp. 772, 784, 802).

[Str09] M. Struwe. *Analysis für Informatiker*. Lecture notes, ETH Zürich. 2009 (cit. on pp. 761, 764, 769–771, 780, 792).

Chapter 12

Single-Step Methods for Stiff Initial-Value Problems

Explicit Runge-Kutta methods with stepsize control (→ Section 11.5) seem to be able to provide approximate solutions for any IVP with good accuracy provided that tolerances are set appropriately. Does this mean that everything is settled about numerical integration**?**

EXAMPLE 12.0.0.1 (Explicit adaptive RK-SSM for stiff IVP) In this example we will witness the near failure of a high-order adaptive explicit Runge-Kutta method for a simple scalar autonomous ODE.

$$\text{IVP considered:} \quad \dot{y} = \lambda y^2(1-y)\,, \quad \lambda := 500 \quad , \quad y(0) = \tfrac{1}{100}\,. \tag{12.0.0.2}$$

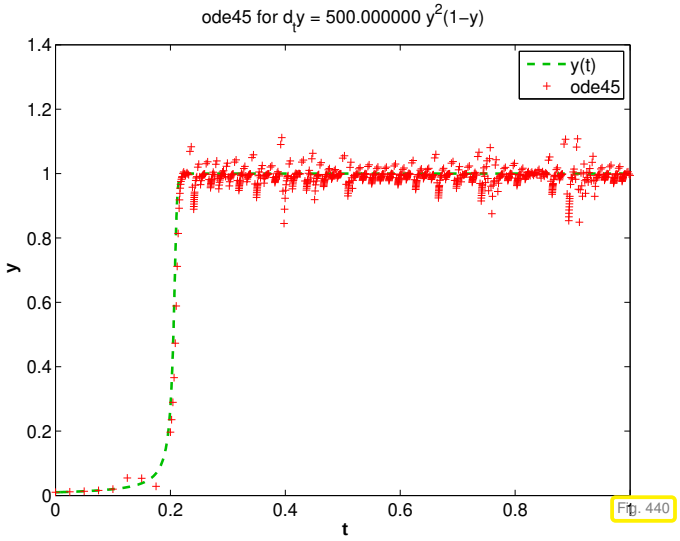
This is a logistic ODE as introduced in Ex. 11.1.2.1. We try to solve it by means of an explicit adaptive embedded Runge-Kutta-Fehlberg method (→ Section 11.5.3) using the embedded Runge-Kutta single-step method offered by **Ode45** as explained in § 11.5.3.3 (Preprocessor switch `MATLABCOEFF` activated).

C++ code 12.0.0.3: Solving (12.0.0.2) **with Ode45 numerical integrator ⇝ GITLAB**

```
2       // Types to be used for a scalar ODE with state space ℝ
3       using StateType = double;
4       using RhsType = std::function<StateType(StateType)>;
5       // Logistic differential equation (11.1.2.2)
6       const double lambda = 500.0;
7       const RhsType f = [lambda](StateType y) { return lambda * y * y * (1 - y); };
8       const StateType y0 = 0.01;  // Initial value, will create a STIFF IVP
9       // State space ℝ, simple modulus supplies norm
10      const auto normFunc = [](StateType x) { return fabs(x); };
11
12      // Invoke explicit Runge-Kutta method with stepsize control
13      Ode45<StateType, RhsType> integrator(f);
14      // Set rather loose tolerances
15      integrator.options().rtol = 0.1;
16      integrator.options().atol = 0.001;
17      integrator.options().min_dt = 1E-18;
18      const std::vector<std::pair<StateType, double>> states =
19          integrator.solve(y0, 1.0, normFunc);
20      // Output information accumulation during numerical integration
21      integrator.options().do_statistics = true;
22      integrator.print();
```

Statistics of the integrator run ▷

```
1 - number of steps:                183
2 - number of rejected steps:       185
3 - function calls:                 1302
```



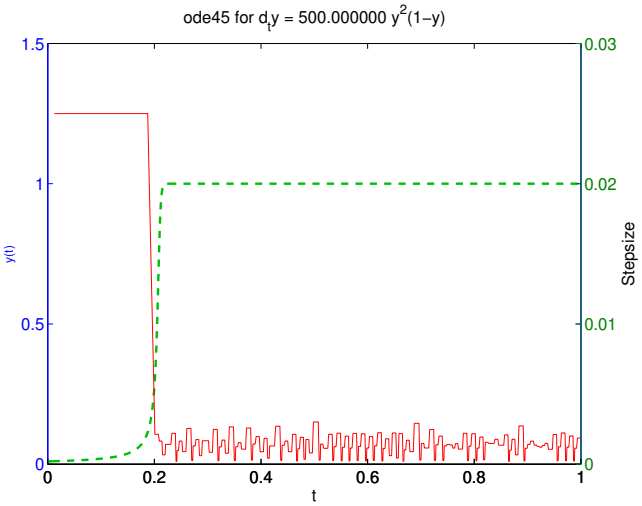


Fig. 439 Fig. 440

Stepsize control of **Ode45** running amok!

? The solution is virtually constant from $t > 0.2$ and, nevertheless, the integrator uses tiny timesteps until the end of the integration interval. Why this crazy behavior? ⌟

Contents

12.1 Model Problem Analysis 820
12.2 Stiff Initial-Value Problems 834
12.3 Implicit Runge-Kutta Single-Step Methods 840
12.3.1 The Implicit Euler Method for Stiff IVPs 840
12.3.2 Collocation Single-Step Methods 841
12.3.3 General Implicit Runge-Kutta Single-Step Methods (RK-SSMs) 845
12.3.4 Model Problem Analysis for Implicit Runge-Kutta Single-Step Methods (IRK-SSMs) 847
12.4 Semi-Implicit Runge-Kutta Methods 855
12.5 Splitting Methods 858

Supplementary literature. [DR08, Sect. 11.9]

12.1 Model Problem Analysis

Video tutorial for Section 12.1:Model Problem Analysis: (40 minutes) Download link, tablet notes

Fortunately, full insight into the observations made in Ex. 12.0.0.1 can already be gleaned from a scalar linear model problem that is extremely easy to analyze.

EXPERIMENT 12.1.0.1 (Adaptive explicit RK-SSM for scalar linear decay ODE) To rule out that what we observed in Ex. 12.0.0.1 might have been a quirk of the IVP (12.0.0.2) we conduct the same investigations for the simple linear, scalar ($N = 1$), autonomous IVP

$$\dot{y} = \lambda y \quad , \quad \lambda := -80 \quad , \quad y(0) = 1 \, . \tag{12.1.0.2}$$

We use the adaptive integrator of **Ode45** (→ § 11.5.3.3) to solve (12.1.0.2) with the same parameters as in Code 11.5.3.4. ▷

```
1 - number of steps:                33
2 - number of rejected steps:       32
3 - function calls:                 231
```

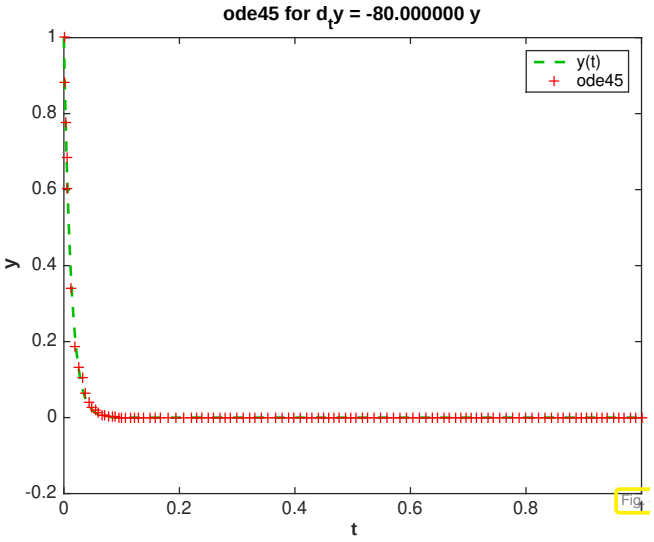


Fig.

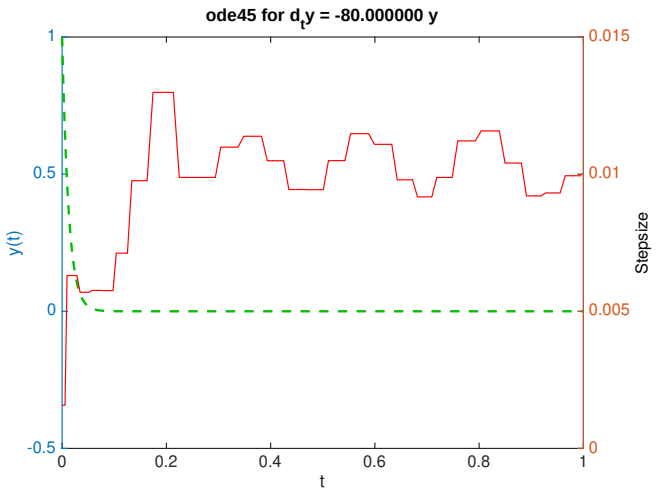


Fig.

Observation: Though $y(t) \approx 0$ for $t > 0.1$, the integrator keeps on using "unreasonably small" timesteps even then. ⌟

In this section we will discover a simple explanation for the startling behavior of the adaptive timestepping **Ode45** in Ex. 12.0.0.1.

EXAMPLE 12.1.0.3 (Blow-up of explicit Euler method) The simplest explicit RK-SSM is the explicit Euler method, see Section 11.2.1. We know that it should converge like $O(h)$ for meshwidth $h \to 0$. In this example we will see that this may be true only for sufficiently small h, which may be extremely small.

✦ We consider the IVP for the scalar linear decay ODE:

$$\dot{y} = f(y) := \lambda y \,, \quad \lambda \ll 0 \quad , \quad y(0) = 1 \,.$$

✦ We apply the explicit Euler method (11.2.1.5) with uniform timestep $h = 1/M$, $M \in \{5, 10, 20, 40, 80, 160, 320, 640\}$.

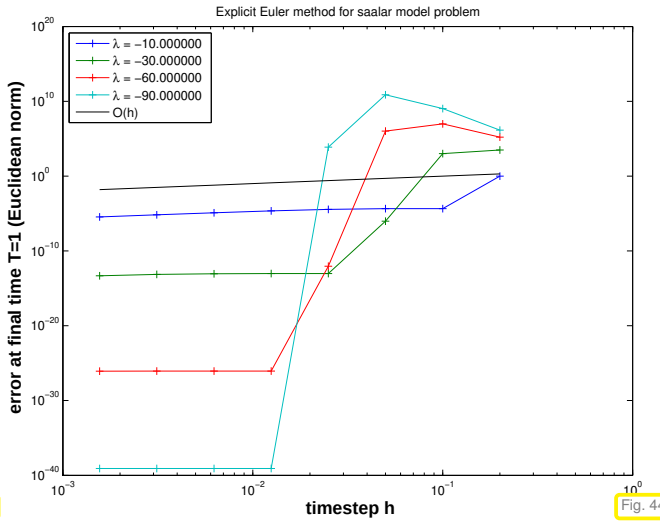


Fig. 443

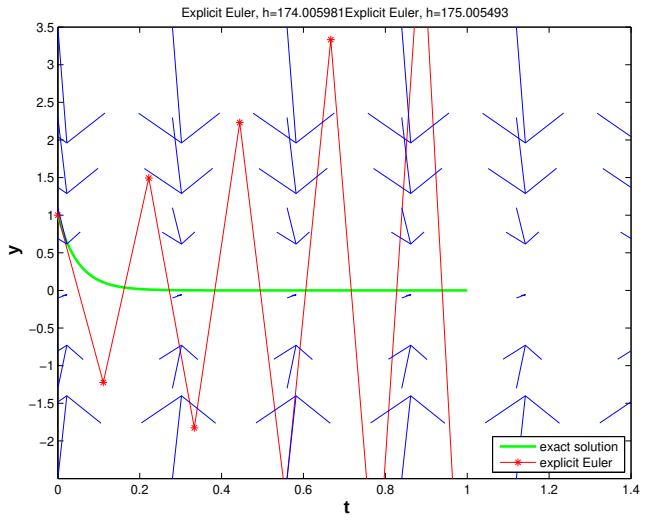


Fig. 444

$\lambda \ll 0$: blow-up of y_k for large timestep h

$\lambda = 20$: — $\hat{=}$ $y(t)$, — $\hat{=}$ Euler polygon

Explanation: From Fig. 444 we draw the geometric conclusion that, if h is "large in comparison with λ^{-1}", then the approximations y_k way miss the stationary point $y = 0$ due to overshooting.

This leads to a sequence $(y_k)_k$ with exponentially increasing oscillations.

✦ Now we look at an IVP for the logistic ODE, see Ex. 11.1.2.1:

$$\dot{y} = f(y) := \lambda y(1-y) \quad , \quad y(0) = 0.01 \, .$$

✦ As before, we apply the explicit Euler method (11.2.1.5) with uniform timestep $h = 1/M$, $M \in \{5, 10, 20, 40, 80, 160, 320, 640\}$.

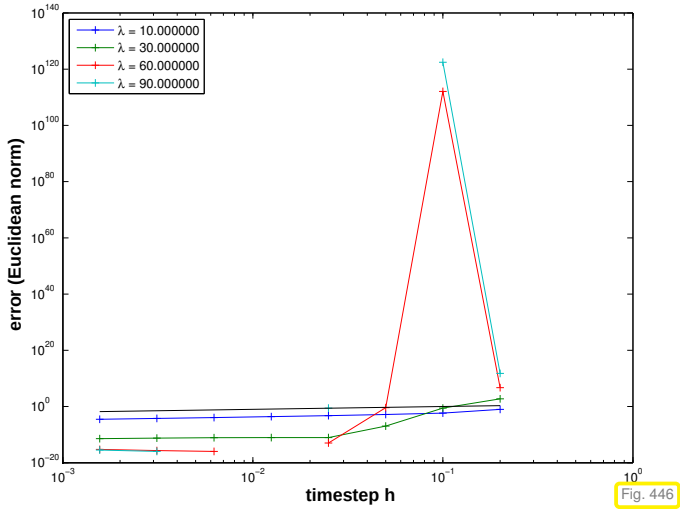


Fig. 445

λ large: blow-up of y_k for large timestep h

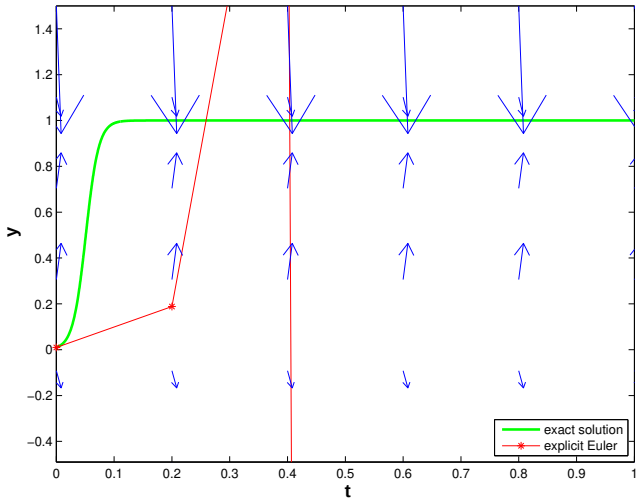


Fig. 446

$\lambda = 90$: — $\hat{=}$ $y(t)$, — $\hat{=}$ Euler polygon

For large timesteps h we also observe oscillatory blow-up of the sequence $(y_k)_k$.

Deeper analysis:

For $y \approx 1$: $f(y) \approx \lambda(1-y)$ ➢ If $y(t_0) \approx 1$, then the solution of the IVP will behave like the solution of $\dot{y} = \lambda(1-y)$, which is a linear ODE. Similary, $z(t) := 1 - y(t)$ will behave like the solution of the "decay equation" $\dot{z} = -\lambda z$. Thus, around the stationary point $y = 1$ the explicit Euler method behaves like it did for $\dot{y} = \lambda y$ in the vicinity of the stationary point $y = 0$; it grossly overshoots. ⌟

§12.1.0.4 (Linear model problem analysis: explicit Euler method) The phenomenon observed in the two previous examples is accessible to a remarkably simple rigorous analysis: Motivated by the considerations in Ex. 12.1.0.3 we study the explicit Euler method (11.2.1.5) for the

$$\text{linear model problem:} \quad \dot{y} = \lambda y \, , \quad y(0) = y_0 \, , \quad \text{with} \quad \lambda \ll 0 \, , \tag{12.1.0.5}$$

which has an *exponentially decaying* exact solution

$$y(t) = y_0 \exp(\lambda t) \to 0 \quad \text{for } t \to \infty \, .$$

Recall the recursion for the explicit Euler with uniform timestep $h > 0$ method for (12.1.0.5):

$$\text{(11.2.1.5) for } f(y) = \lambda y\text{:} \qquad y_{k+1} = y_k(1 + \lambda h) \, . \tag{12.1.0.6}$$

We easily get a closed form expression for the approximations y_k:

$$\blacktriangleright \quad y_k = y_0(1+\lambda h)^k \quad \Rightarrow \quad |y_k| \to \begin{cases} 0 & \text{, if } \lambda h > -2 \quad \text{(qualitatively correct)} \, , \\ \infty & \text{, if } \lambda h < -2 \quad \text{(qualitatively wrong)} \, . \end{cases}$$

Observed: stability-induced timestep constraint

Only if $|\lambda|h < 2$ we obtain a decaying solution by the explicit Euler method!

Could it be that the timestep control is desperately trying to enforce the qualitatively correct behavior of the numerical solution in Ex. 12.1.0.3? Let us examine how the simple stepsize control of Code 11.5.2.6 fares for model problem (12.1.0.5):

EXPERIMENT 12.1.0.8 (Simple adaptive timestepping for fast decay) In this example we let a transparent adaptive timestep struggle with "overshooting":

- "Linear model problem IVP": $\dot{y} = \lambda y$, $y(0) = 1$, $\lambda = -100$
- Simple adaptive timestepping method as in Exp. 11.5.2.9, see Code 11.5.2.6. Timestep control based on the pair of 1st-order explicit Euler method and 2nd-order explicit trapezoidal method.

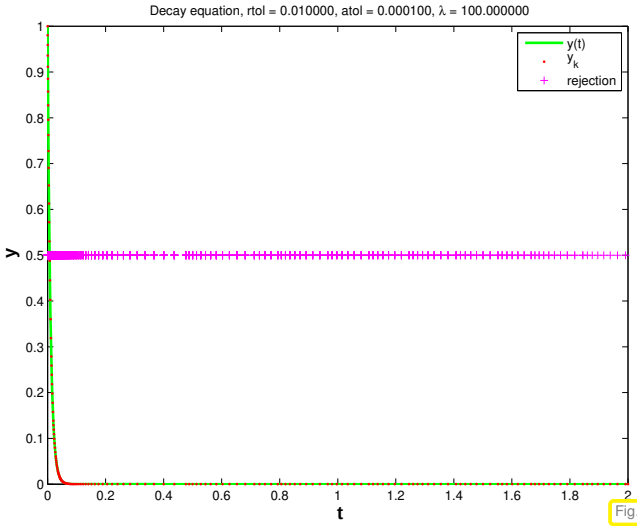


Fig. 447

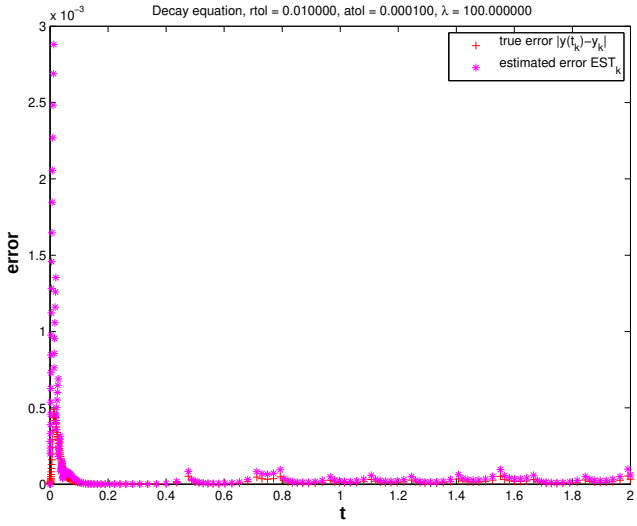


Fig. 448

Observation: in fact, stepsize control enforces small timesteps even if $\mathbf{y}(t) \approx 0$ and persistently triggers rejections of timesteps. This is necessary to prevent overshooting in the Euler method, which contributes to the estimate of the one-step error.

We see the purpose of *stepsize control thwarted*, because after only a very short time the solution is almost zero and then, in fact, large timesteps should be chosen.

Are these observations a particular "flaw" of the explicit Euler method? Let us study the behavior of another simple explicit Runge-Kutta method applied to the linear model problem.

EXAMPLE 12.1.0.9 (Explicit trapzoidal method for decay equation → [DR08, Ex. 11.29])

The explicit trapezoidal method is a 2-stage explicit Ruge-Kutta method, whose Butcher scheme is given in Ex. 11.4.0.17 and which was derived in Ex. 11.4.0.6. We state its recursion for the ODE $\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y})$ in terms of the first step $\mathbf{y}_0 \to \mathbf{y}_1$:

$$\mathbf{k}_1 = \mathbf{f}(t_0, \mathbf{y}_0)\,, \quad \mathbf{k}_2 = \mathbf{f}(t_0 + h, \mathbf{y}_0 + h\mathbf{k}_1)\,, \quad \mathbf{y}_1 = \mathbf{y}_0 + \tfrac{h}{2}(\mathbf{k}_1 + \mathbf{k}_2)\,. \tag{11.4.0.8}$$

Apply it to the model problem (12.1.0.5), that is, the scalar autonomous ODE with right hand side function $\mathbf{f}(y) = f(y) = \lambda y$, $\lambda < 0$:

$$k_1 = \lambda y_0\,, \quad k_2 = \lambda(y_0 + hk_1) \quad \Rightarrow \quad y_1 = \underbrace{(1 + \lambda h + \tfrac{1}{2}(\lambda h)^2)}_{=:S(h\lambda)} y_0\,. \tag{12.1.0.10}$$

The sequence of approximations generated by the explicit trapezoidal rule can be expressed in closed form as

$$y_k = S(h\lambda)^k y_0 \,, \quad k = 0, \ldots, N \,. \tag{12.1.0.11}$$

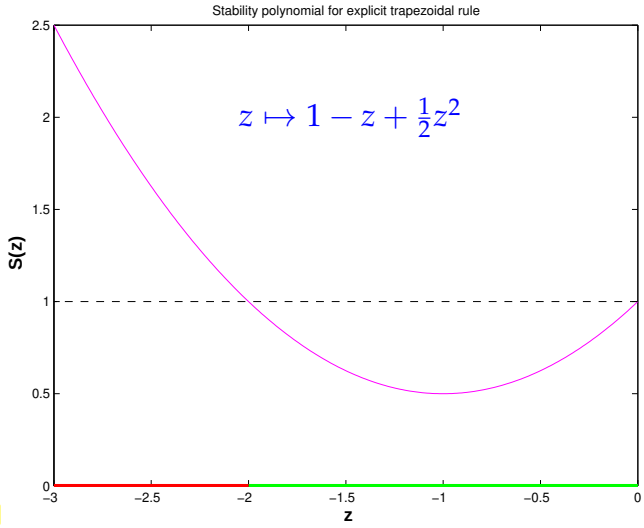


Fig. 449

Clearly, blow-up can be avoided only if $|S(h\lambda)| \leq 1$:

$$|S(h\lambda)| < 1 \quad \Leftrightarrow \quad -2 < h\lambda < 0 \,.$$

Qualitatively correct decay behavior of $(y_k)_k$ only under timestep constraint

$$h \leq |2/\lambda| \,. \tag{12.1.0.12}$$

◁ the **stability function** for the explicit trapezoidal method

§12.1.0.13 (Model problem analysis for general explicit Runge-Kutta single step methods) We generalize the approach taken in Ex. 12.1.0.9 and apply an explicit s-stage Runge-Kutta method (→ Def. 11.4.0.11) encoded by the Butcher scheme $\begin{array}{c|c} \mathbf{c} & \mathfrak{A} \\ \hline & \mathbf{b}^T \end{array}$, $\mathfrak{A} \in \mathbb{R}^{s,s}$ strictly lower-triangular, to the autonomous scalar linear ODE (12.1.0.5) ($\dot{y} = \lambda y$). We write down the equations for the increments and y_1 from Def. 11.4.0.11 for $f(y) := \lambda y$ and then convert the resulting system of equations into matrix form:

$$\begin{aligned} k_i &= \lambda(y_0 + h\sum_{j=1}^{i-1} a_{ij}k_j) \,, \\ y_1 &= y_0 + h\sum_{i=1}^{s} b_i k_i \end{aligned} \quad \Rightarrow \quad \begin{bmatrix} \mathbf{I} - z\mathfrak{A} & 0 \\ -z\mathbf{b}^\top & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{k} \\ y_1 \end{bmatrix} = y_0 \begin{bmatrix} \mathbf{1} \\ 1 \end{bmatrix} \,, \tag{12.1.0.14}$$

where $\boldsymbol{k} \in \mathbb{R}^s \,\hat{=}\,$ denotes the vector $[k_1, \ldots, k_s]^\top / \lambda$ of increments, and $z := \lambda h$. Next we apply block Gaussian elimination (→ Rem. 2.3.1.11) to solve for y_1 and obtain

$$y_1 = S(z)y_0 \quad \text{with} \quad S(z) := 1 + z\mathbf{b}^T(\mathbf{I} - z\mathfrak{A})^{-1}\mathbf{1} \,, \quad \mathbf{1} = [1, \ldots, 1]^\top \,. \tag{12.1.0.15}$$

Alternatively we can express y_1 through determinants appealing to Cramer's rule,

$$y_1 = y_0 \frac{\det \begin{bmatrix} \mathbf{I} - z\mathfrak{A} & \mathbf{1} \\ -z\mathbf{b}^\top & 1 \end{bmatrix}}{\det \begin{bmatrix} \mathbf{I} - z\mathfrak{A} & 0 \\ -z\mathbf{b}^\top & 1 \end{bmatrix}} \quad \Rightarrow \quad S(z) = \det(\mathbf{I} - z\mathfrak{A} + z\mathbf{1}\mathbf{b}^T) \,, \tag{12.1.0.16}$$

and note that $\mathfrak{A}$ is a strictly lower triangular matrix, which means that $\det(\mathbf{I} - z\mathfrak{A}) = 1$. Thus we have proved the following theorem.

Theorem 12.1.0.17. Stability function of some explicit Runge-Kutta methods → [Han02, Thm. 77.2], [QSS00, Sect. 11.8.4]

The discrete evolution Ψ_λ^h of an explicit s-stage Runge-Kutta single step method (→ Def. 11.4.0.11) with Butcher scheme $\begin{array}{c|c} \mathbf{c} & \mathfrak{A} \\ \hline & \mathbf{b}^T \end{array}$ *(see* (11.4.0.13)*) for the ODE $\dot{y} = \lambda y$ amounts to a multiplication with the number*

$$\Psi_\lambda^h = S(\lambda h) \quad \Leftrightarrow \quad y_1 = S(\lambda h) y_0 \,,$$

where S is the ***stability function*** *(SF)*

$$S(z) := 1 + z\mathbf{b}^T(\mathbf{I} - z\mathfrak{A})^{-1}\mathbf{1} = \det(\mathbf{I} - z\mathfrak{A} + z\mathbf{1}\mathbf{b}^T) \,, \quad \mathbf{1} := [1,\ldots,1]^\top \in \mathbb{R}^s \,. \tag{12.1.0.18}$$

⌟

EXAMPLE 12.1.0.19 (Stability functions of explicit Runge-Kutta single step methods) From Thm. 12.1.0.17 and their Butcher schemes we can instantly compute the stability functions of explicit RK-SSM. We do this for a few methods whose Butcher schemes were listed in Ex. 11.4.0.17

- Explicit Euler method (11.2.1.5): $\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}$ ➢ $S(z) = 1 + z$.
- Expl. trapezoidal method (11.4.0.8): $\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$ ➢ $S(z) = 1 + z + \frac{1}{2}z^2$.
- Classical RK4 method: $\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \hline & \frac{1}{6} & \frac{2}{6} & \frac{2}{6} & \frac{1}{6} \end{array}$ ➢ $S(z) = 1 + z + \frac{1}{2}z^2 + \frac{1}{6}z^3 + \frac{1}{24}z^4$.

These examples confirm an immediate consequence of the determinant formula for the stability function $S(z)$.

Corollary 12.1.0.20. Polynomial stability function of explicit RK-SSM

For a consistent (→ Def. 11.3.1.12) s-stage explicit Runge-Kutta single step method according to Def. 11.4.0.11 the stability function S defined by (12.1.0.56) *is a non-constant polynomial of degree $\leq s$, that is, $S \in \mathcal{P}_s$.*

⌟

Remark 12.1.0.21 (Stability function and exponential function) Let us compare the two evolution operators:

- $\Phi \triangleq$ evolution operator (→ Def. 11.1.4.3) for $\dot{y} = \lambda y$,

- $\Psi \triangleq$ discrete evolution operator ($\rightarrow$ § 11.3.1.1) for an s-stage Runge-Kutta single step method.

$$\Phi^h y = e^{\lambda h} y \quad \longleftrightarrow \quad \Psi^h y = S(\lambda h) y \, .$$

In light of that Ψ is supposed to be an approximation for Φ, $\Psi \approx \Phi$, see (11.3.1.3), we expect that

$$S(z) \approx \exp(z) \quad \text{for small} \quad |z| \, . \tag{12.1.0.22}$$

A more precise statement is made by the following lemma:

Lemma 12.1.0.23. Stability function as approximation of $\exp$ for small arguments

Let S denote the stability function of an s-stage explicit Runge-Kutta single step method of order $q \in \mathbb{N}$. Then

$$|S(z) - \exp(z)| = O(|z|^{q+1}) \quad \textit{for} \quad |z| \to 0 \, . \tag{12.1.0.24}$$

This means that the lowest $q+1$ coefficients of $S(z)$ must be equal to the first coefficients of the exponential series:

$$S(z) = \sum_{j=0}^{q} \frac{1}{j!} z^j + z^{q+1} p(z) \quad \text{with some} \quad p \in \mathcal{P}_{s-q-1} \, .$$

In order to match the first q terms of the exponential series, we need at least $S \in \mathcal{P}_q$, which entails a minimum of q stages.

Corollary 12.1.0.25. Stages limit order of explicit RK-SSM

An explicit s-stage RK-SSM has maximal order $q \leq s$.

⌟

§12.1.0.26 (Stability induced timestep constraint) In § 12.1.0.13 we established that for the sequence $(y_k)_{k=0}^{\infty}$ produced by an explicit Runge-Kutta single step method applied to the linear scalar model ODE $\dot{y} = \lambda y$, $\lambda \in \mathbb{R}$, with uniform timestep $h > 0$ holds

$$y_{k+1} = S(\lambda h) y_k \quad \Rightarrow \quad y_k = S(\lambda h^k y_0 \, .$$

$$\blacktriangleright \quad \begin{array}{lll} (y_k)_{k=0}^{\infty} \text{ non-increasing} & \Leftrightarrow & |S(\lambda h)| \leq 1 \, , \\ (y_k)_{k=0}^{\infty} \text{ exponentially increasing} & \Leftrightarrow & |S(\lambda h)| > 1 \, . \end{array} \tag{12.1.0.27}$$

where $S = S(z)$ is the stability function of the RK-SSM as defined in (12.1.0.56).

Invariably polynomials tend to $\pm\infty$ for large (in modulus) arguments:

$$\forall S \in \mathcal{P}_s, S \neq \text{const} : \quad \lim_{|z| \to \infty} S(z) = \infty \quad \text{uniformly} \, . \tag{12.1.0.28}$$

So, for any $\lambda \neq 0$ there will be a threshold $h_{\max} > 0$ so that $|y_k| \to \infty$ as $|h| > h_{\max}$.

Reversing the argument we arrive at a timestep constraint, as already observed for the explicit Euler methods in § 12.1.0.4.

Only if one ensures that $|\lambda h|$ is sufficiently small, one can avoid exponentially increasing approximations y_k (qualitatively wrong for $\lambda < 0$) when applying an explicit RK-SSM to the model problem (12.1.0.5) with uniform timestep $h > 0$,

For $\lambda \ll 0$ this stability induced timestep constraint may force h to be much *smaller than required by demands on accuracy*: in this case timestepping becomes inefficient. ⌟

Remark 12.1.0.29 (Stepsize control detects instability) Ex. 12.0.0.1, Exp. 12.1.0.8 send the message that local-in-time stepsize control as discussed in Section 11.5 selects timesteps that avoid blow-up, with a hefty price tag however in terms of computational cost and poor accuracy. ⌟

Objection: simple linear scalar IVP (12.1.0.5) may be an oddity rather than a model problem: the weakness of explicit Runge-Kutta methods discussed above may be just a peculiar response to an unusual situation. Let us extend our investigations to systems of linear ODEs of dimension $N > 1$ of the state space.

§12.1.0.30 (Systems of linear ordinary differential equations, § 11.1.1.8 revisited) A generic linear ordinary differential equation with constant coefficients on the state space $\mathbb{R}^N$ has the form

$$\dot{\mathbf{y}} = \mathbf{M}\mathbf{y} \quad \text{with a matrix} \quad \mathbf{M} \in \mathbb{R}^{N,N} . \tag{12.1.0.31}$$

As explained in [NS02, Sect. 8.1], (12.1.0.31) can be solved by diagonalization: If we can find a *regular* matrix $\mathbf{V} \in \mathbb{C}^{N,N}$ such that

$$\mathbf{M}\mathbf{V} = \mathbf{V}\mathbf{D} \quad \text{with diagonal matrix} \quad \mathbf{D} = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_N \end{bmatrix} \in \mathbb{C}^{N,N} , \tag{12.1.0.32}$$

then the 1-parameter family of global solutions of (12.1.0.31) is given by

$$\mathbf{y}(t) = \mathbf{V} \begin{bmatrix} \exp(\lambda_1 t) & & 0 \\ & \ddots & \\ 0 & & \exp(\lambda_N t) \end{bmatrix} \mathbf{V}^{-1}\mathbf{y}_0 , \quad \mathbf{y}_0 \in \mathbb{R}^N . \tag{12.1.0.33}$$

The columns of $\mathbf{V}$ are a basis of eigenvectors of $\mathbf{M}$, the $\lambda_j \in \mathbb{C}$, $j = 1, \ldots, N$ are the associated eigenvalues of $\mathbf{M}$, see Def. 9.1.0.1.

The idea behind diagonalization is the transformation of (12.1.0.31) into N *decoupled* scalar linear ODEs:

$$\dot{\mathbf{y}} = \mathbf{M}\mathbf{y} \xrightarrow{\mathbf{z}(t) := \mathbf{V}^{-1}\mathbf{y}(t)} \dot{\mathbf{z}} = \mathbf{D}\mathbf{z} \quad \leftrightarrow \quad \begin{aligned} \dot{z}_1 &= \lambda_1 z_1 \\ &\vdots \\ \dot{z}_N &= \lambda_N z_N \end{aligned} \quad \text{, since} \quad \mathbf{M} = \mathbf{V}\mathbf{D}\mathbf{V}^{-1} .$$

The formula (12.1.0.33) can be generalized to

$$\mathbf{y}(t) = \exp(\mathbf{M}t)\mathbf{y}_0 \quad \text{with the matrix exponential} \quad \exp(\mathbf{B}) := \sum_{k=0}^{\infty} \frac{1}{k!}\mathbf{B}^k , \quad \mathbf{B} \in \mathbb{C}^{N,N} . \tag{12.1.0.34}$$

⌟

EXAMPLE 12.1.0.35 (Transient simulation of RLC-circuit) This example demonstrates the diagonalization of a linear system of ODEs.

Consider the circuit from Ex. 11.1.2.11 ▷

Transient nodal analysis (→ Ex. 2.1.0.3) leads to the second-order linear ODE

$$\ddot{u} + \alpha\dot{u} + \beta u = g(t) \,,$$

with coefficients $\alpha := (RC)^{-1}$, $\beta = (LC)^{-1}$, $g(t) = \alpha\dot{U}_s(t)$.

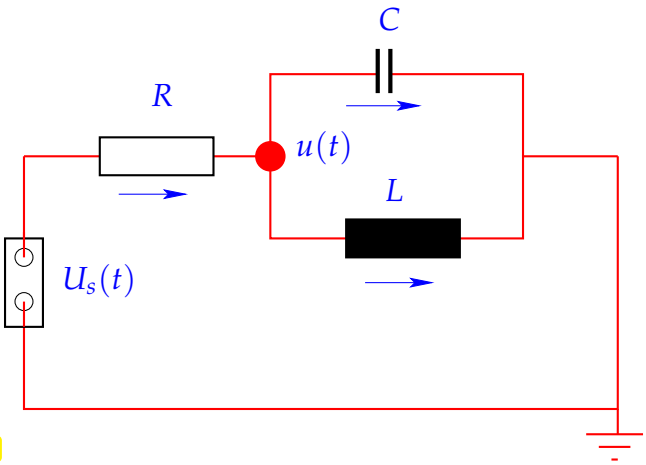


Fig. 450

We transform it to a linear 1st-order ODE as in Rem. 11.1.3.6 by introducing $v := \dot{u}$ as additional solution component:

$$\underbrace{\begin{bmatrix}\dot{u}\\ \dot{v}\end{bmatrix}}_{=:\dot{\mathbf{y}}} = \underbrace{\begin{bmatrix}0 & 1\\ -\beta & -\alpha\end{bmatrix}\begin{bmatrix}u\\ v\end{bmatrix} - \begin{bmatrix}0\\ g(t)\end{bmatrix}}_{=:\mathbf{f}(t,\mathbf{y})} \,, \qquad \text{with } \beta \gg \alpha \gg 1 \text{ in usual settings.}$$

We integrate IVPs for this ODE by means of the adaptive integrator **Ode45** from § 11.5.3.3.

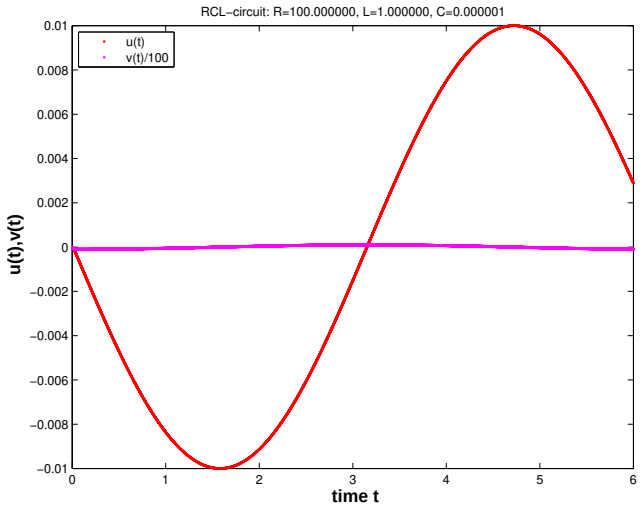


Fig. 451

$R = 100\Omega$, $L = 1\mathrm{H}$, $C = 1\mu\mathrm{F}$, $U_s(t) = 1\mathrm{V}\sin(t)$, $u(0) = v(0) = 0$ ("switch on")

Ode45 statistics:

```
17897 successful steps
1090 failed attempts
113923 function evaluations
```

Inefficient: way more timesteps than required for resolving smooth solution, *cf.* remark in the end of § 12.1.0.26.

Maybe the time-dependent right hand side due to the time-harmonic excitation severly affects **ode45**? Let us try a constant exciting voltage:

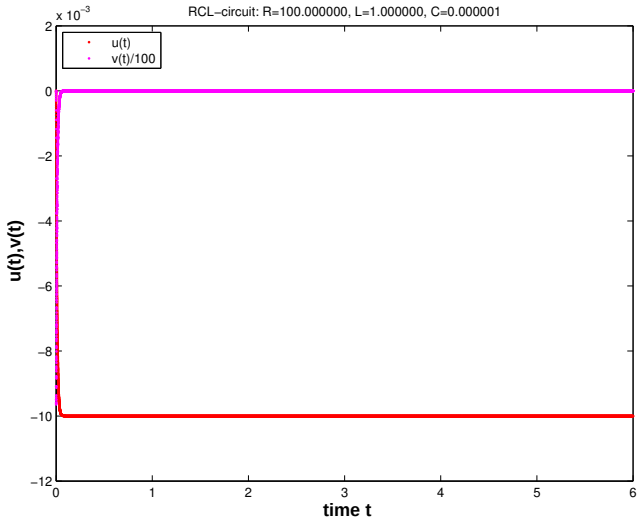


Fig. 452

$R = 100\Omega$, $L = 1\mathrm{H}$, $C = 1\mu\mathrm{F}$, $U_s(t) = 1\mathrm{V}$, $u(0) = v(0) = 0$ ("switch on")

Ode45 statistics:

```
17901 successful steps
1210 failed attempts
114667 function evaluations
```

Tiny timesteps despite virtually constant solution!

We make the same observation as in Ex. 12.0.0.1, Exp. 12.1.0.8: the local-in-time stepsize control of

ode45 (→ Section 11.5) enforces extremely small timesteps though the solution almost constant except at $t = 0$.

To understand the structure of the solutions for this transient circuit example, let us apply the diagonalization technique from § 12.1.0.30 to the linear ODE

$$\dot{\mathbf{y}} = \underbrace{\begin{bmatrix} 0 & 1 \\ -\beta & -\alpha \end{bmatrix}}_{=:\mathbf{M}} \mathbf{y} \quad , \quad \mathbf{y}(0) = \mathbf{y}_0 \in \mathbb{R}^2 \, . \tag{12.1.0.36}$$

Above we face the situation $\beta \gg \frac{1}{4}\alpha^2 \gg 1$.

We can obtain the general solution of $\dot{\mathbf{y}} = \mathbf{M}\mathbf{y}$, $\mathbf{M} \in \mathbb{R}^{2,2}$, by diagonalization of $\mathbf{M}$ (if possible):

$$\mathbf{M}\mathbf{V} = \mathbf{M}(\mathbf{v}_1, \mathbf{v}_2) = (\mathbf{v}_1, \mathbf{v}_2) \begin{bmatrix} \lambda_1 & \\ & \lambda_2 \end{bmatrix} . \tag{12.1.0.37}$$

where $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^2 \setminus \{0\}$ are the the eigenvectors of $\mathbf{M}$, λ_1, λ_2 are the eigenvalues of $\mathbf{M}$, see Def. 9.1.0.1. The latter are the roots of the characteristic polynomial $t \mapsto \chi(t) := t^2 + \alpha t + \beta$ in $\mathbb{C}$, and we find

$$\lambda_{1/2} = \tfrac{1}{2}(-\alpha \pm D) \, , \quad D := \begin{cases} \sqrt{\alpha^2 - 4\beta} & \text{, if } \alpha^2 \geq 4\beta \, , \\ \imath\sqrt{4\beta - \alpha^2} & \text{, if } \alpha^2 < 4\beta \, . \end{cases}$$

Note that the eigenvalues have a large (in modulus) negative real part and a non-vanishing imaginary part in the setting of the experiment.

Then we transform $\dot{\mathbf{y}} = \mathbf{M}\mathbf{y}$ into *decoupled* scalar linear ODEs:

$$\dot{\mathbf{y}} = \mathbf{M}\mathbf{y} \quad \Leftrightarrow \quad \mathbf{V}^{-1}\dot{\mathbf{y}} = \mathbf{V}^{-1}\mathbf{M}\mathbf{V}(\mathbf{V}^{-1}\mathbf{y}) \overset{\mathbf{z}(t):=\mathbf{V}^{-1}\mathbf{y}(t)}{\Leftrightarrow} \quad \dot{\mathbf{z}} = \begin{bmatrix} \lambda_1 & \\ & \lambda_2 \end{bmatrix} \mathbf{z} \, . \tag{12.1.0.38}$$

This yields the general solution of the ODE $\dot{\mathbf{y}} = \mathbf{M}\mathbf{y}$, see also [Str09, Sect. 5.6]:

$$\mathbf{y}(t) = A\mathbf{v}_1 \exp(\lambda_1 t) + B\mathbf{v}_2 \exp(\lambda_2 t) \, , \quad A, B \in \mathbb{R} \, . \tag{12.1.0.39}$$

Note: $t \mapsto \exp(\lambda_i t)$ is general solution of the ODE $\dot{z}_i = \lambda_i z_i$. ⌟

§12.1.0.40 ("Diagonalization" of explicit Euler method) Recall the discrete evolution of the explicit Euler method (11.2.1.5) for the linear ODE $\dot{\mathbf{y}} = \mathbf{M}\mathbf{y}$, $\mathbf{M} \in \mathbb{R}^{N,N}$:

$$\Psi^h\mathbf{y} = \mathbf{y} + h\mathbf{M}\mathbf{y} \quad \leftrightarrow \quad \mathbf{y}_{k+1} = \mathbf{y}_k + h\mathbf{M}\mathbf{y}_k \, .$$

As in § 12.1.0.30 we assume that $\mathbf{M}$ can be diagonalized, that is (12.1.0.32) holds: $\mathbf{V}^{-1}\mathbf{M}\mathbf{V} = \mathbf{D}$ with a diagonal matrix $\mathbf{D} \in \mathbb{C}^{N,N}$ containing the eigenvalues of $\mathbf{M}$ on its diagonal. Next, apply the decoupling by diagonalization idea to the recursion of the explicit Euler method.

$$\mathbf{V}^{-1}\mathbf{y}_{k+1} = \mathbf{V}^{-1}\mathbf{y}_k + h\mathbf{V}^{-1}\mathbf{M}\mathbf{V}(\mathbf{V}^{-1}\mathbf{y}_k) \overset{\mathbf{z}_k:=\mathbf{V}^{-1}\mathbf{y}_k}{\Leftrightarrow} \underbrace{(\mathbf{z}_{k+1})_i = (\mathbf{z}_k)_i + h\lambda_i(\mathbf{z}_k)_i}_{\hat{=} \text{ explicit Euler step for } \dot{z}_i = \lambda_i z_i} \, , \tag{12.1.0.41}$$

with $i \in \{1, \ldots, N\}$. This gives us a crucial insight:

The explicit Euler method generates uniformly bounded solution sequences $(\mathbf{y}_k)_{k=0}^{\infty}$ for $\dot{\mathbf{y}} = \mathbf{M}\mathbf{y}$ with diagonalizable matrix $\mathbf{M} \in \mathbb{R}^{N,N}$ with eigenvalues $\lambda_1, \ldots, \lambda_N$, if and only if it generates uniformly bounded sequences for **all** the scalar ODEs $\dot{z} = \lambda_i z$, $i = 1, \ldots, N$.

⌟

So far we conducted the model problem analysis under the premises $\lambda < 0$.

However, in Ex. 12.1.0.35 we face $\lambda_{1/2} = -\frac{1}{2}\left(\alpha \pm i\sqrt{4\beta - \alpha^2}\right)$ (complex eigenvalues!). Let us now examine how the explicit Euler method and even general explicit RK-methods respond to them.

Remark 12.1.0.42 (Explicit Euler method for damped oscillations) Consider linear model IVP (12.1.0.5) for $\lambda \in \mathbb{C}$:

$$\operatorname{Re}\lambda < 0 \quad \Rightarrow \quad \text{exponentially decaying solution} \quad y(t) = y_0 \exp(\lambda t)\,,$$

because $|\exp(\lambda t)| = \exp(\operatorname{Re}\lambda \cdot t)$.

The model problem analysis from Ex. 12.1.0.3, Ex. 12.1.0.9 can be extended verbatim to the case of $\lambda \in \mathbb{C}$. It yields the following insight for the for the explicit Euler method and $\lambda \in \mathbb{C}$:

The sequence generated by the explicit Euler method (11.2.1.5) for the model problem (12.1.0.5) satisfies

$$y_{k+1} = y_k(1 + h\lambda) \quad \blacktriangleright \quad \lim_{k\to\infty} y_k = 0 \quad \Leftrightarrow \quad |1 + h\lambda| < 1\,. \tag{12.1.0.6}$$

timestep constraint to get decaying (discrete) solution !

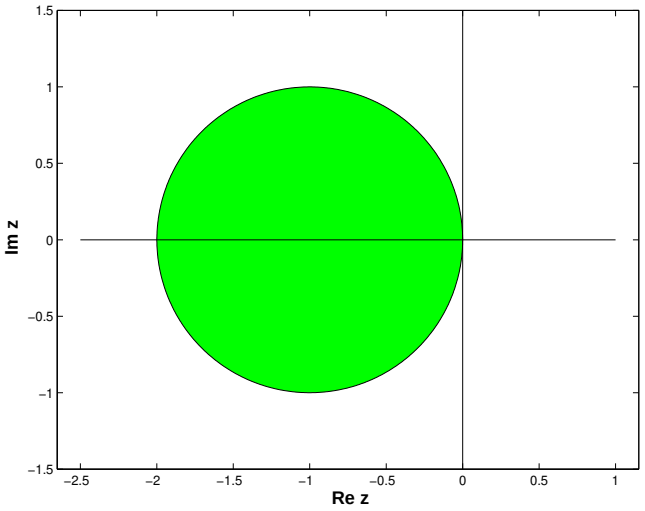


Fig. 453

$\triangleleft \quad \{z \in \mathbb{C}:\ |1 + z| < 1\}$

The green region of the complex plane marks values for λh, for which the explicit Euler method will produce exponentially decaying solutions.

Now we can conjecture what happens in Ex. 12.1.0.35: the eigenvalues $\lambda_{1/2} = -\frac{1}{2}\alpha \pm i\sqrt{\beta - \frac{1}{4}\alpha^2}$ of $\mathbf{M}$ have a very large (in modulus) negative real part. Since the integrator of **Ode45** can be expected to behave as if it integrates $\dot{z} = \lambda_2 z$, it faces a severe timestep constraint, if exponential blow-up is to be avoided, see Ex. 12.1.0.3. Thus stepsize control must resort to tiny timesteps. ⌟

§12.1.0.43 (Extended model problem analysis for explicit Runge-Kutta single step methods) Recall the definition of a generic explicit RK-SSM for the ODE $\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y})$:

Definition 11.4.0.11. Explicit Runge-Kutta method

For $b_i, a_{ij} \in \mathbb{R}$, $c_i := \sum_{j=1}^{i-1} a_{ij}$, $i, j = 1, \ldots, s$, $s \in \mathbb{N}$, an **s-stage explicit Runge-Kutta single step method** (RK-SSM) for the ODE $\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y})$, $\mathbf{f} : \Omega \to \mathbb{R}^N$, is defined by ($\mathbf{y}_0 \in D$)

$$\mathbf{k}_i := \mathbf{f}(t_0 + c_i h, \mathbf{y}_0 + h\sum_{j=1}^{i-1} a_{ij}\mathbf{k}_j)\,, \quad i = 1, \ldots, s \quad , \quad \mathbf{y}_1 := \mathbf{y}_0 + h\sum_{i=1}^{s} b_i \mathbf{k}_i\,.$$

The vectors $\mathbf{k}_i \in \mathbb{R}^N$, $i = 1, \ldots, s$, are called **increments**, $h > 0$ is the size of the timestep.

We apply such an explicit s-stage RK-SSM described by the Butcher scheme $\begin{array}{c|c} \mathbf{c} & \mathfrak{A} \\ \hline & \mathbf{b}^T \end{array}$ to the autonomous linear ODE $\dot{\mathbf{y}} = \mathbf{M}\mathbf{y}$, $\mathbf{M} \in \mathbb{C}^{N,N}$, and obtain (for the first step with timestep size $h > 0$)

$$\mathbf{k}_\ell = \mathbf{M}(\mathbf{y}_0 + h\sum_{j=1}^{\ell-1} a_{\ell j}\mathbf{k}_j)\,, \quad \ell = 1,\ldots,s \quad , \quad \mathbf{y}_1 = \mathbf{y}_0 + h\sum_{\ell=1}^{s} b_i\mathbf{k}_\ell\,. \tag{12.1.0.44}$$

Now assume that $\mathbf{M}$ can be diagonalized, that is (12.1.0.32) holds: $\mathbf{V}^{-1}\mathbf{M}\mathbf{V} = \mathbf{D}$ with a diagonal matrix $\mathbf{D} \in \mathbb{C}^{N,N}$ containing the eigenvalues $\lambda_1,\ldots,\lambda_N \in \mathbb{C}$ of $\mathbf{M}$ on its diagonal. Then apply the substitutions

$$\widehat{\mathbf{k}}_\ell := \mathbf{V}^{-1}\mathbf{k}_\ell\,, \quad \ell = 1,\ldots,s \quad , \quad \widehat{\mathbf{y}}_k := \mathbf{V}^{-1}\mathbf{y}_k\,, \quad k = 0,1\,,$$

to (12.1.0.44), which yield

$$\widehat{\mathbf{k}}_\ell = \mathbf{D}(\widehat{\mathbf{y}}_0 + h\sum_{j=1}^{s-1} a_{\ell j}\widehat{\mathbf{k}}_j)\,, \quad \ell = 1,\ldots,s \quad , \quad \widehat{\mathbf{y}}_1 = \widehat{\mathbf{y}}_0 + h\sum_{\ell=1}^{s} b_i\widehat{\mathbf{k}}_\ell\,. \tag{12.1.0.45}$$

$$\Updownarrow$$

$$\left(\widehat{\mathbf{k}}_\ell\right)_i = \lambda_i((\mathbf{y}_0)_i + h\sum_{j=1}^{s-1} a_{\ell j}\left(\widehat{\mathbf{k}}_j\right)_i)\,, \quad (\widehat{\mathbf{y}}_1)_i = (\widehat{\mathbf{y}}_0)_i + h\sum_{\ell=1}^{s} b_i\left(\widehat{\mathbf{k}}_\ell\right)_i\,, \quad i = 1,\ldots,N\,. \tag{12.1.0.46}$$

We infer that, if $(\mathbf{y}_k)_k$ is the sequence produced by an explicit RK-SSM applied to $\dot{\mathbf{y}} = \mathbf{M}\mathbf{y}$, then

$$\mathbf{y}_k = \mathbf{V}\begin{bmatrix} y_k^{[1]} & & 0 \\ & \ddots & \\ 0 & & y_k^{[d]} \end{bmatrix}\mathbf{V}^{-1}\,,$$

where $\left(y_k^{[i]}\right)_k$ is the sequence generated by the same RK-SSM with the same sequence of timesteps for the IVP $\dot{y} = \lambda_i y$, $y(0) = \left(\mathbf{V}^{-1}\mathbf{y}_0\right)_i$.

The RK-SSM generates uniformly bounded solution sequences $(\mathbf{y}_k)_{k=0}^{\infty}$ for the ODE $\dot{\mathbf{y}} = \mathbf{M}\mathbf{y}$ with diagonalizable matrix $\mathbf{M} \in \mathbb{R}^{N,N}$ with eigenvalues $\lambda_1,\ldots,\lambda_N$, if and only if it generates uniformly bounded sequences for **all** the scalar ODEs $\dot{z} = \lambda_i z$, $i = 1,\ldots,N$.

Stability analysis: reduction to scalar case

Understanding the behavior of RK-SSM for autonomous scalar linear ODEs $\dot{y} = \lambda y$ with $\lambda \in \mathbb{C}$ is enough to predict their behavior for general autonomous linear systems of ODEs.

From the considerations of § 12.1.0.26 we deduce the following fundamental result.

Theorem 12.1.0.48. (Absolute) stability of explicit RK-SSM for linear systems of ODEs

The sequence $(\mathbf{y}_k)_k$ of approximations generated by an explicit RK-SSM (→ Def. 11.4.0.11) with stability function S (defined in (12.1.0.56)*) applied to the linear autonomous ODE $\dot{\mathbf{y}} = \mathbf{M}\mathbf{y}$, $\mathbf{M} \in \mathbb{C}^{N,N}$, with uniform timestep $h > 0$ decays exponentially for* every *initial state $\mathbf{y}_0 \in \mathbb{C}^N$, if and only if $|S(\lambda_i h)| < 1$ for all eigenvalues λ_i of $\mathbf{M}$.*

Please note that

$$\operatorname{Re}\lambda_i < 0 \quad \forall i \in \{1,\ldots,N\} \quad \Longrightarrow \quad \|\mathbf{y}(t)\| \to 0 \quad \text{for} \quad t \to \infty\,,$$

for any solution of $\dot{\mathbf{y}} = \mathbf{M}\mathbf{y}$. This is obvious from the representation formula (12.1.0.33). ⌟

§12.1.0.49 (Region of (absolute) stability of explicit RK-SSM) We consider an explicit Runge-Kutta single step method with stability function S for the model linear scalar IVP $\dot{y} = \lambda y$, $y(0) = y_0$, $\lambda \in \mathbb{C}$. From Thm. 12.1.0.17 we learn that for uniform stepsize $h > 0$ we have $y_k = S(\lambda h)^k y_0$ and conclude that

$$\boxed{y_k \to 0 \quad \text{for} \quad k \to \infty \quad \Leftrightarrow \quad |S(\lambda h)| < 1} \,. \tag{12.1.0.50}$$

Hence, the modulus $|S(\lambda h)|$ tells us for which combinations of λ and stepsize h we achieve exponential decay $y_k \to \infty$ for $k \to \infty$, which is the desirable behavior of the approximations for $\operatorname{Re}\lambda < 0$.

Definition 12.1.0.51. Region of (absolute) stability

Let the discrete evolution Ψ for a single step method applied to the scalar linear ODE $\dot{y} = \lambda y$, $\lambda \in \mathbb{C}$, be of the form

$$\Psi^h y = S(z)y\,, \quad y \in \mathbb{C}, h > 0 \quad \text{with} \quad z := h\lambda \tag{12.1.0.52}$$

and a function $S : \mathbb{C} \to \mathbb{C}$. Then the **region of (absolute) stability** of the single step method is given by

$$\mathcal{S}_\Psi := \{z \in \mathbb{C}\colon |S(z)| < 1\} \subset \mathbb{C}\,.$$

Of course, by Thm. 12.1.0.17, in the case of explicit RK-SSM the function S will coincide with their stability function from (12.1.0.56).

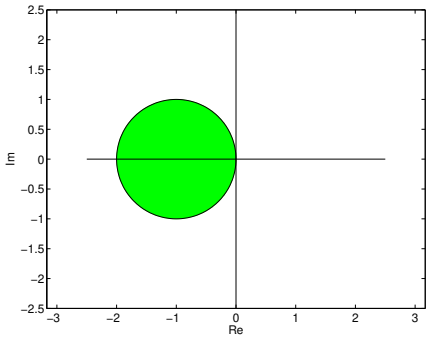
We can easily combine the statement of Thm. 12.1.0.48 with the concept of a region of stability and conclude that an explicit RK-SSM will generate expoentially decaying solutions for the linear ODE $\dot{\mathbf{y}} = \mathbf{M}\mathbf{y}$, $\mathbf{M} \in \mathbb{C}^{N,N}$, for every initial state $\mathbf{y}_0 \in \mathbb{C}^N$, if and only if $\lambda_i h \in \mathcal{S}_\Psi$ for all eigenvalues λ_i of $\mathbf{M}$.

Adopting the arguments of § 12.1.0.26 we conclude from Cor. 12.1.0.20 that

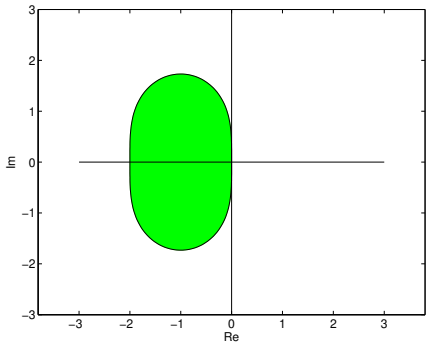
- the regions of (absolute) stability of explicit RK-SSM are bounded,
- a timestep constraint depending on the eigenvalues of $\mathbf{M}$ is necessary to have a guaranteed exponential decay RK-solutions for $\dot{\mathbf{y}} = \mathbf{M}\mathbf{y}$.

⌟

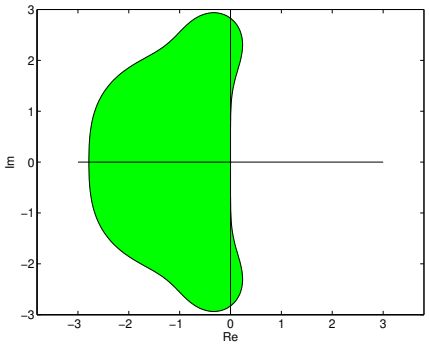
EXAMPLE 12.1.0.53 (Regions of stability of some explicit RK-SSM) The green domains $\subset \mathbb{C}$ depict the bounded regions of stability for some RK-SSM from Ex. 11.4.0.17.



$\mathcal{S}_\Psi$: explicit Euler (11.2.1.5)



$\mathcal{S}_\Psi$: explicit trapezoidal method



$\mathcal{S}_\Psi$: classical RK4 method

In general we have for a consistent RK-SSM ($\to$ Def. 11.3.1.12) that their stability functions staidfy $S(z) = 1 + z + O(z^2)$ for $z \to 0$. Therefore, $\mathcal{S}_\Psi \neq \emptyset$ and the imaginary axis will be tangent to $\mathcal{S}_\Psi$ in $z = 0$. ⌟

Review question(s) 12.1.0.54 (Model problem analysis)

(Q12.1.0.54.A) We consider the autonomous linear ODE

$$\dot{\mathbf{y}}(t) = \mathbf{M}\mathbf{y}(t) \quad \text{with} \quad \mathbf{M} := \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} . \tag{12.1.0.55}$$

on the state space $\mathbb{R}^2$.

(i) Sketch the right-hand-side vectorfield $\mathbf{y} \mapsto \mathbf{f}(\mathbf{y})$ for this ODE.

(ii) Compute the two-parameter family of its solutions.

Hint. The functions $\mathbf{y}(t) := \begin{bmatrix} \cos t \\ \sin t \end{bmatrix}$ and $\mathbf{y}(t) := \begin{bmatrix} -\sin t \\ \cos t \end{bmatrix}$ are solutions of (12.1.0.55).

(iii) Based on geometric reasoning, predict how the explicit Euler method will perform when applied to (12.1.0.55).

(Q12.1.0.54.B) Explain why explicit Runge-Kutta single step methods (RK-SSMs) are well suited for the autonomous scalar linear ODE $\dot{y} = \lambda y$ ("growth ODE") provided that $\lambda \geq 0$.

(Q12.1.0.54.C) What can you say about the stability function of a 3-stage explicit Runge-Kutta single step method of order 3?

(Q12.1.0.54.D) Compute the stability function for the "generic" second-order two-stage Runge-Kutta single step method, whose Butcher scheme is

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \alpha & \alpha & 0 \\ \hline & 1-\frac{1}{2\alpha} & \frac{1}{2\alpha} \end{array} , \quad \alpha \in \mathbb{R} \setminus \{0\} .$$

Theorem 12.1.0.17. Stability function of some explicit Runge-Kutta methods

The discrete evolution Ψ_λ^h of an explicit s-stage Runge-Kutta single step method ($\to$ Def. 11.4.0.11) with Butcher scheme $\begin{array}{c|c} \mathbf{c} & \mathfrak{A} \\ \hline & \mathbf{b}^T \end{array}$ (see (11.4.0.13)) for the ODE $\dot{y} = \lambda y$ amounts to a multiplication with the number

$$\Psi_\lambda^h = S(\lambda h) \quad \Leftrightarrow \quad y_1 = S(\lambda h) y_0 ,$$

*where S is the **stability function** (SF)*

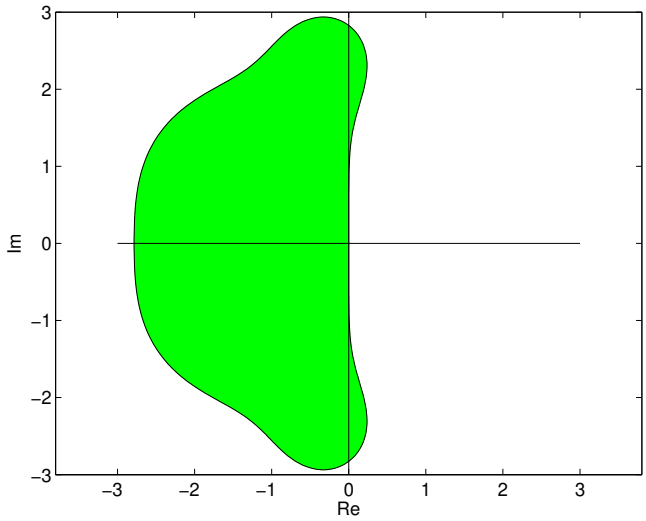
$$S(z) := 1 + z\mathbf{b}^T(\mathbf{I} - z\mathfrak{A})^{-1}\mathbf{1} = \det(\mathbf{I} - z\mathfrak{A} + z\mathbf{1}\mathbf{b}^T) , \quad \mathbf{1} := [1, \ldots, 1]^\top \in \mathbb{R}^s . \tag{12.1.0.56}$$

(Q12.1.0.54.E) Compute the stability function for the three-stage Runge-Kutta single step method, defined through the Butcher scheme

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 \\ 2/3 & 0 & 2/3 & 0 \\ \hline & 1/4 & 0 & 3/4 \end{array} .$$

(Q12.1.0.54.F) What is the stability-induced timestep constraint for the classical 4-stage explicit Runge-Kutta single step method of order 4, when applied to the ODE

$$\dot{\mathbf{y}}(t) = \mathbf{M}\mathbf{y}(t) \quad \text{with} \quad \mathbf{M} := \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} .$$



◁ Stability domain $\mathcal{S}_\Psi$ of the classical 4-stage Runge-Kutta single step method of order 4.

Hint. You may use that

$$\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ \imath & -\imath \end{bmatrix}^{\mathrm{H}} \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ \imath & -\imath \end{bmatrix} = \begin{bmatrix} -\imath & 0 \\ 0 & \imath \end{bmatrix} .$$

△

Supplementary literature. Related to this section are [Han02, Ch. 77] and [QSS00, Sect. 11.3.3].

12.2 Stiff Initial-Value Problems

Video tutorial for Section 12.2: Stiff Initial-Value Problems: (24 minutes) Download link, tablet notes

This section will reveal that the behavior observed in Ex. 12.0.0.1 and Ex. 12.1.0.3 is typical for a large class of problems and that the model problem (12.1.0.5) really represents a "generic case". This justifies the attention paid to linear model problem analysis in Section 12.1.

EXAMPLE 12.2.0.1 (Kinetics of chemical reactions → [Han02, Ch. 62]) In Ex. 11.5.1.1 we already saw an ODE model for the dynamics of a chemical reaction. Now we study an abstract reaction.

$$\text{reaction:} \quad \underbrace{\mathrm{A} + \mathrm{B} \underset{k_1}{\overset{k_2}{\rightleftharpoons}} \mathrm{C}}_{\text{fast reaction}} \quad , \quad \underbrace{\mathrm{A} + \mathrm{C} \underset{k_3}{\overset{k_4}{\rightleftharpoons}} \mathrm{D}}_{\text{slow reaction}} , \tag{12.2.0.2}$$

assuming vastly different reaction constants: $k_1, k_2 \gg k_3, k_4$

If $c_A(0) > c_B(0)$ ➢ 2nd reaction determines overall long-term reaction dynamics

Mathematical model: non-linear ODE involving concentrations $\mathbf{y}(t) = [c_A(t), c_B(t), c_C(t), c_D(t)]^\top$

$$\dot{\mathbf{y}} := \frac{d}{dt}\begin{bmatrix} c_A \\ c_B \\ c_C \\ c_D \end{bmatrix} = \mathbf{f}(\mathbf{y}) := \begin{bmatrix} -k_1 c_A c_B + k_2 c_C - k_3 c_A c_C + k_4 c_D \\ -k_1 c_A c_B + k_2 c_C \\ k_1 c_A c_B - k_2 c_C - k_3 c_A c_C + k_4 c_D \\ k_3 c_A c_C - k_4 c_D \end{bmatrix} . \tag{12.2.0.3}$$

Concrete choice of parameters: $t_0 = 0$, $T = 1$, $k_1 = 10^4$, $k_2 = 10^3$, $k_3 = 10$, $k_4 = 1$, initial value $\mathbf{y}_0 = [1, 1, 10, 0]^\top$.

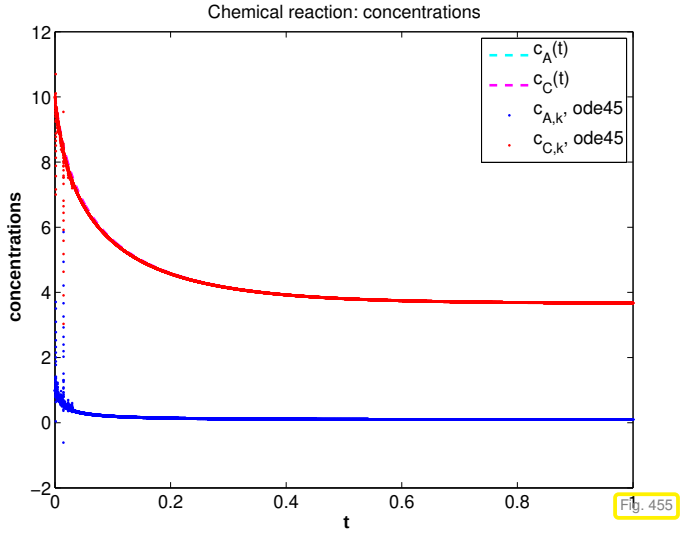


Fig. 454

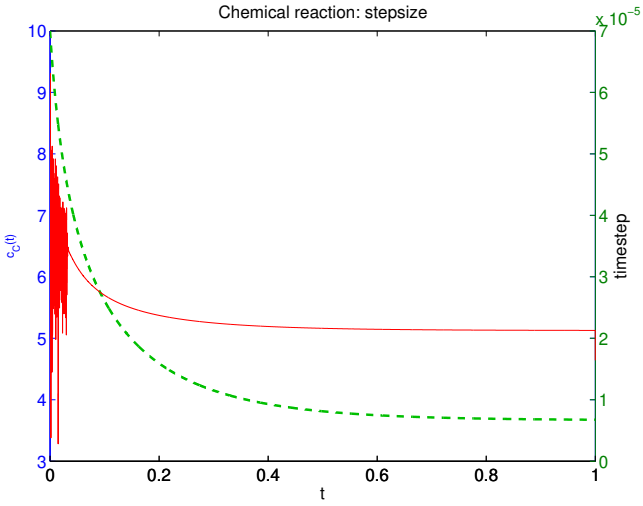


Fig. 455

Observations: After a fast initial transient phase, the solution shows only slow dynamics. Nevertheless, the explicit adaptive integrator used for this simulation insists on using a tiny timestep. It behaves very much like **Ode45** in Ex. 12.0.0.1. ⌟

EXAMPLE 12.2.0.4 (Strongly attractive limit cycle) We consider the non-linear Autonomous ODE $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$ with

$$\mathbf{f}(\mathbf{y}) := \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \mathbf{y} + \lambda(1 - \|\mathbf{y}\|^2)\,\mathbf{y}\,, \tag{12.2.0.5}$$

on the state space $D = \mathbb{R}^2 \setminus \{0\}$.

For $\lambda = 0$, the initial value problem $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$, $\mathbf{y}(0) = \begin{bmatrix} \cos\varphi \\ \sin\varphi \end{bmatrix}$, $\varphi \in \mathbb{R}$ has the solution

$$\mathbf{y}(t) = \begin{bmatrix} \cos(t - \varphi) \\ \sin(t - \varphi) \end{bmatrix}\,, \quad t \in \mathbb{R}\,. \tag{12.2.0.6}$$

For this solution we have $\|\mathbf{y}(t)\|_2 = 1$ for all times.

(12.2.0.6) provides a solution even for $\lambda \neq 0$, if $\|\mathbf{y}(0)\|_2 = 1$, because in this case the term $\lambda(1 - \|\mathbf{y}\|^2)\,\mathbf{y}$ will never become non-zero on the solution trajectory.

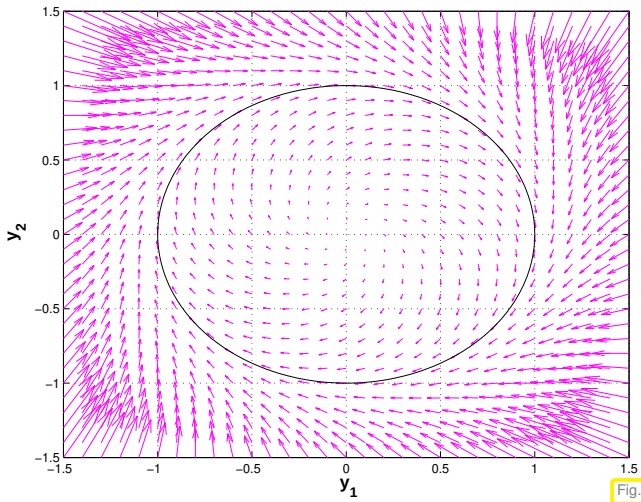


Fig. 456

vectorfield $\mathbf{f}$ ($\lambda = 1$)

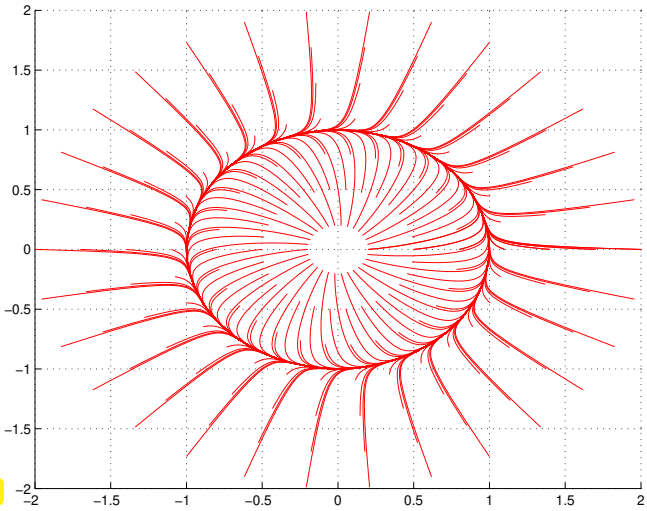


Fig. 457

solution trajectories ($\lambda = 10$)

We study the response of **Ode45** introduced in § 11.5.3.3 to different choice of λ with initial state $\mathbf{y}_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$.

According to the above considerations this initial state should completely "hide the impact of λ from our view".

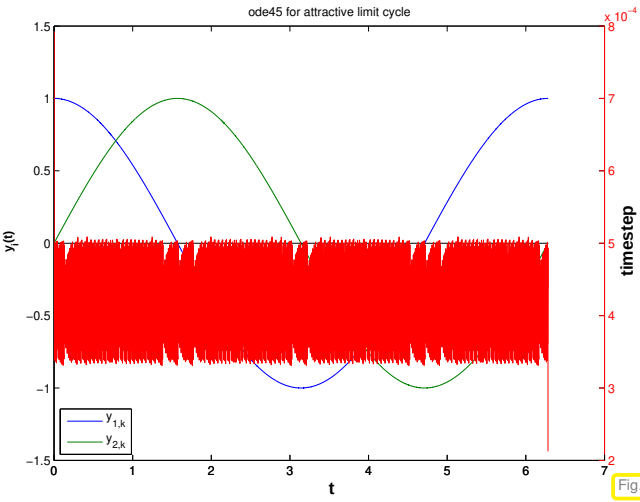


Fig. 458

many (3794) steps ($\lambda = 1000$)

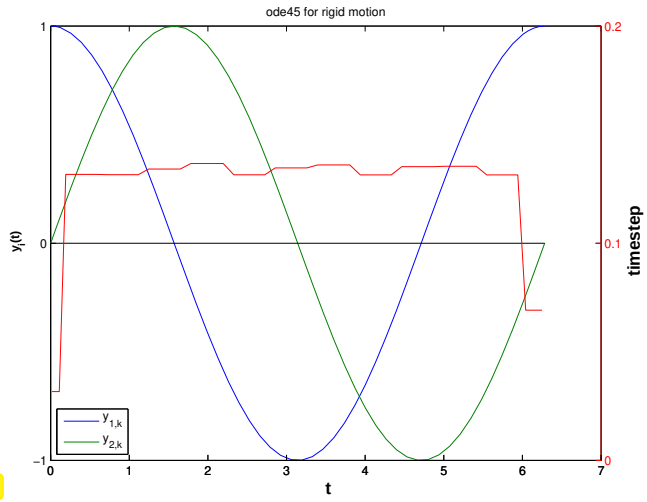


Fig. 459

accurate solution with few steps ($\lambda = 0$)

Confusing observation: we have $\|\mathbf{y}_0\| = 1$, which implies $\|\mathbf{y}(t)\| = 1 \quad \forall t$!

Thus, the term of the right hand side, which is multiplied by λ will always vanish on the exact solution trajectory, which stays on the unit circle.

Nevertheless, **Ode45** is forced to use tiny timesteps by the *mere presence* of this term! ⌟

We want to find criteria that allow to predict the massive problems haunting explicit single step methods in the case of the *non-linear* IVP of Ex. 12.0.0.1, Ex. 12.2.0.1, and Ex. 12.2.0.4. Recall that for *linear* IVPs of the form $\dot{\mathbf{y}} = \mathbf{M}\mathbf{y}$, $\mathbf{y}(0) = \mathbf{y}_0$, the model problem analysis of Section 12.1 tells us that, given knowledge of the region of stability of the timestepping scheme, the eigenvalues of the matrix $\mathbf{M} \in \mathbb{C}^{N,N}$ provide full information about timestep constraint we are going to face. Refer to Thm. 12.1.0.48 and § 12.1.0.49.

The ODEs we saw in Ex. 12.2.0.1 and Ex. 12.2.0.4 are *non-linear*. Yet, the entire stability analysis of Section 12.1 was based on linear ODEs. Thus, we need to extend the stability analysis to non-linear ODEs.

We start with a "phenomenological notion", just a keyword to refer to the kind of difficulties presented by the IVPs of Ex. 12.0.0.1, Ex. 12.2.0.1, Exp. 12.1.0.8, and Ex. 12.2.0.4.

Notion 12.2.0.7. Stiff IVP

An initial value problem is called stiff, if stability imposes much tighter timestep constraints on *explicit single step methods* than the accuracy requirements.

§12.2.0.8 (Linearization of ODEs) Linear ODEs, though very special, are highly relevant as "local model" for general ODEs: We consider a general autonomous ODE

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}) \quad , \quad \mathbf{f} : D \subset \mathbb{R}^N \to \mathbb{R}^N \, .$$

As usual, we assume $\mathbf{f}$ to be C^2-smooth and that it enjoys local Lipschitz continuity ($\rightarrow$ Def. 11.1.3.13) on D so that unique solvability of IVPs is guaranteed by Thm. 11.1.3.17.

We *fix* a state $\mathbf{y}^* \in D$, D the state space, write $t \mapsto \mathbf{y}(t)$ for the solution with $\mathbf{y}(0) = \mathbf{y}^*$. We set $\mathbf{z}(t) = \mathbf{y}(t) - \mathbf{y}^*$, which satisfies

$$\mathbf{z}(0) = 0 \quad , \quad \dot{\mathbf{z}} = \mathbf{f}(\mathbf{y}^* + \mathbf{z}) = \mathbf{f}(\mathbf{y}^*) + \mathrm{D}\,\mathbf{f}(\mathbf{y}^*)\mathbf{z} + R(\mathbf{y}^*, \mathbf{z}) \, , \quad \text{with} \quad \|R(\mathbf{y}^*, \mathbf{z})\| = O(\|\mathbf{z}\|^2) \, .$$

This is obtained by Taylor expansion of $\mathbf{f}$ at $\mathbf{y}^*$, see [Str09, Satz 7.5.2]. Hence, in a neighborhood of a state $\mathbf{y}^*$ on a solution trajectory $t \mapsto \mathbf{y}(t)$, the deviation $\mathbf{z}(t) = \mathbf{y}(t) - \mathbf{y}^*$ satisfies

$$\dot{\mathbf{z}} \approx \mathbf{f}(\mathbf{y}^*) + \mathrm{D}\,\mathbf{f}(\mathbf{y}^*)\mathbf{z} \, . \tag{12.2.0.9}$$

The short-time evolution of $\mathbf{y}$ with $\mathbf{y}(0) = \mathbf{y}^*$ is approximately governed by the affine-linear ODE

$$\boxed{\dot{\mathbf{y}} = \mathbf{M}(\mathbf{y} - \mathbf{y}^*) + \mathbf{b} \, , \quad \mathbf{M} := \mathrm{D}\,\mathbf{f}(\mathbf{y}^*) \in \mathbb{R}^{N,N} \, , \quad \mathbf{b} := \mathbf{f}(\mathbf{y}^*) \in \mathbb{R}^N} \, . \tag{12.2.0.10}$$

In the scalar case we have come across this linearization already in Ex. 12.1.0.3. ⌟

§12.2.0.11 (Linearization of explicit Runge-Kutta single step methods) We consider one step of a general s-stage RK-SSM according to Def. 11.4.0.11 for the autonomous ODE $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$, with smooth right hand side function $\mathbf{f} : D \subset \mathbb{R}^N \to \mathbb{R}^N$:

$$\mathbf{k}_i = \mathbf{f}(\mathbf{y}_0 + h\sum_{j=1}^{i-1} a_{ij}\mathbf{k}_j) \, , \quad i = 1, \ldots, s \quad , \quad \mathbf{y}_1 = \mathbf{y}_0 + h\sum_{i=1}^{s} b_i \mathbf{k}_i \, .$$

We perform linearization at $\mathbf{y}^* := \mathbf{y}_0$ and ignore all terms at least quadratic in the timestep size h (this is indicated by the $\approx$ symbol):

$$\mathbf{k}_i \approx \mathbf{f}(\mathbf{y}^*) + \mathrm{D}\,\mathbf{f}(\mathbf{y}^*)h\sum_{j=1}^{i-1} a_{ij}\mathbf{k}_j \, , \quad i = 1, \ldots, s \quad , \quad \mathbf{y}_1 = \mathbf{y}_0 + h\sum_{i=1}^{s} b_i \mathbf{k}_i \, .$$

The defining equations for the same RK-SSM applied to

$$\dot{\mathbf{z}} = \mathbf{M}\mathbf{z} + \mathbf{b} \, , \quad \mathbf{M} := \mathrm{D}\,\mathbf{f}(\mathbf{y}^*) \in \mathbb{R}^{N,N} \, , \quad \mathbf{b} := \mathbf{f}(\mathbf{y}^*) \, ,$$

which agrees with (12.2.0.10) after substitution $\mathbf{z}(t) - \mathbf{y}(t) - \mathbf{y}^*$, are

$$\mathbf{k}_i \approx \mathbf{b} + \mathbf{M}h\sum_{j=1}^{i-1} a_{ij}\mathbf{k}_j \, , \quad i = 1, \ldots, s \quad , \quad \mathbf{y}_1 = \mathbf{y}_0 + h\sum_{i=1}^{s} b_i \mathbf{k}_i \, .$$

We find that for small timesteps

> the discrete evolution of the RK-SSM for $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$ in the state $\mathbf{y}^*$ is close to the discrete evolution of the same RK-SSM applied to the linearization (12.2.0.10) of the ODE in $\mathbf{y}^*$.

By straightforward manipulations of the defining equations of an explicit RK-SSM we find that, if

- $(\mathbf{y}_k)_k$ is the sequence of states generated by the RK-SSM applied to the affine-linear ODE $\dot{\mathbf{y}} = \mathbf{M}(\mathbf{y} - \mathbf{y}_0) + \mathbf{b}$, $\mathbf{M} \in \mathbb{C}^{N,N}$ regular,
- $(\mathbf{w}_k)_k$ is the sequence of states generated by the same RK-SSM applied to the linear ODE $\dot{\mathbf{w}} = \mathbf{M}\mathbf{w}$ and $\mathbf{w}_0 := \mathbf{M}^{-1}\mathbf{b}$, then

$$\mathbf{w}_k = \mathbf{y}_k - \mathbf{y}_0 + \mathbf{M}^{-1}\mathbf{b} \,.$$

➢ The analysis of the behavior of an RK-SSM for an autonomous affine-linear ODE can be reduced to understanding its behavior for an autonomous linear ODE with the same matrix.

Combined with the insights from § 12.1.0.43 this means that

> the behavior of an explicit Runge-Kutta single-step method applied to $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$ close to the state $\mathbf{y}^*$ is determined by the eigenvalues of the Jacobian $\mathrm{D}\,\mathbf{f}(\mathbf{y}^*)$.

In particular, if $\mathrm{D}\,\mathbf{f}(\mathbf{y}^*)$ has at least one eigenvalue whose modulus is large, then an exponential drift-off of the approximate states $\mathbf{y}_k$ away from $\mathbf{y}^*$ can only be avoided for sufficiently small timestep, again a timestep constraint.

How to distinguish stiff initial value problems

An initial value problem for an autonomous ODE $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$ will probably be stiff, if, for substantial periods of time,

$$\min\{\operatorname{Re}\lambda : \lambda \in \sigma(\mathrm{D}\,\mathbf{f}(\mathbf{y}(t)))\} \ll 0 \,, \tag{12.2.0.13}$$

and

$$\max\{\operatorname{Re}\lambda : \lambda \in \sigma(\mathrm{D}\,\mathbf{f}(\mathbf{y}(t)))\} \lesssim 0 \,, \tag{12.2.0.14}$$

where $t \mapsto \mathbf{y}(t)$ is the solution trajectory and $\sigma(\mathbf{M})$ is the spectrum of the matrix $\mathbf{M}$, see Def. 9.1.0.1.

The condition (12.2.0.14) has to be read as "the real parts of all eigenvalues are below a bound with small modulus". If this is not the case, then the exact solution will experience blow-up. It will change drastically over very short periods of time and small timesteps will be required anyway in order to resolve this. ⌟

EXAMPLE 12.2.0.15 (Predicting stiffness of non-linear IVPs)

❶ We consider the IVP from Ex. 12.0.0.1:

$$\text{IVP considered:} \quad \dot{y} = f(y) := \lambda y^2(1-y) \,, \quad \lambda := 500 \quad , \quad y(0) = \tfrac{1}{100} \,.$$

We find

$$f'(y) = \lambda(2y - 3y^2) \quad \Rightarrow \quad f'(1) = -\lambda \,.$$

Hence, in case $\lambda \gg 1$ as in Fig. 440, we face a stiff problem close to the stationary state $y = 1$. The observations made in Fig. 440 exactly match this prediction.

❷ The solution of the IVP from Ex. 12.2.0.4

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}) := \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \mathbf{y} + \lambda(1 - \|\mathbf{y}\|^2)\,\mathbf{y} \,, \quad \|\mathbf{y}_0\|_2 = 1 \,. \tag{12.2.0.5}$$

satisfies $\|\mathbf{y}(t)\|_2 = 1$ for all times. Using the product rule (8.5.1.17) of multi-dimensional differential calculus, we find

$$\mathrm{D}\,\mathbf{f}(\mathbf{y}) = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} + \lambda\left(-2\mathbf{y}\mathbf{y}^\top + (1 - \|\mathbf{y}\|_2^2\mathbf{I})\right) .$$

$$\sigma(\mathrm{D}\,\mathbf{f}(\mathbf{y})) = \left\{-\lambda - \sqrt{\lambda^2 - 1}, -\lambda + \sqrt{\lambda^2 - 1}\right\} \quad \text{, if} \quad \|\mathbf{y}\|_2 = 1 .$$

Thus, for $\lambda \gg 1$, $\mathrm{D}\,\mathbf{f}(\mathbf{y}(t))$ will always have an eigenvalue with large negative real part, whereas the other eigenvalue is close to zero: the IVP is stiff.

Remark 12.2.0.16 (Characteristics of stiff IVPs) Often one can already tell from the expected behavior of the solution of an IVP, which is often clear from the modeling context, that one has to brace for stiffness.

Typical features of stiff IVPs:
- Presence of fast transients in the solution, see Ex. 12.1.0.3, Ex. 12.1.0.35,
- Occurrence of strongly attractive fixed points/limit cycles, see Ex. 12.2.0.4

Review question(s) 12.2.0.17 (Stiff Initial-Value Problems)

(Q12.2.0.17.A) Then following rules of thumb were given:

How to distinguish stiff initial value problems

An initial value problem for an autonomous ODE $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$ will probably be stiff, if, for substantial periods of time,

$$\min\{\mathrm{Re}\,\lambda : \lambda \in \sigma(\mathrm{D}\,\mathbf{f}(\mathbf{y}(t)))\} \ll 0 , \tag{12.2.0.13}$$

and $$\max\{\mathrm{Re}\,\lambda : \lambda \in \sigma(\mathrm{D}\,\mathbf{f}(\mathbf{y}(t)))\} \lesssim 0 , \tag{12.2.0.14}$$

where $t \mapsto \mathbf{y}(t)$ is the solution trajectory and $\sigma(\mathbf{M})$ is the spectrum of the matrix $\mathbf{M}$, see Def. 9.1.0.1.

Explain, why the condition (12.2.0.14) is has been included.

(Q12.2.0.17.B) [Damped pendulum] The non-dimensional equation for a damped pendulum is

$$\ddot{w} = -\sin w - \lambda\dot{w} .$$

When is this second-order ODE stiff in a neighborhood of the rest state $w = 0$, $\dot{w} = 0$?

(Q12.2.0.17.C) [Transient circuit ODE] In Ex. 12.1.0.35 we found the non-dimensional non-autonomous ODE describing the transient behavior of an RLC-circuit:

$$\underbrace{\begin{bmatrix} \dot{u} \\ \dot{v} \end{bmatrix}}_{=:\dot{\mathbf{y}}} = \underbrace{\begin{bmatrix} 0 & 1 \\ -\beta & -\alpha \end{bmatrix}\begin{bmatrix} u \\ v \end{bmatrix} - \begin{bmatrix} 0 \\ g(t) \end{bmatrix}}_{=:\mathbf{f}(t,\mathbf{y})} , \quad \beta \gg \tfrac{1}{4}\alpha \gg 1 , \tag{12.2.0.19}$$

where $t \mapsto g(t)$ is a given smooth excitation. Find out, whether initial-value problems for (12.2.0.19) whose solution satisfies

$$y(t) \approx \begin{bmatrix} -\beta^{-1}g(t) \\ 0 \end{bmatrix}$$

are stiff.

Hint. The eigenvalues of the matrix $\begin{bmatrix} 0 & 1 \\ -\beta & -\alpha \end{bmatrix}$ are $\lambda_1 = \frac{1}{2}(-\alpha + \imath\sqrt{4\beta - \alpha^2})$, $\lambda_1 = \frac{1}{2}(-\alpha - \imath\sqrt{4\beta - \alpha^2})$.

△

Supplementary literature. [QSS00, Sect. 11.10]

12.3 Implicit Runge-Kutta Single-Step Methods

Video tutorial for Section 12.3: Implicit Runge-Kutta Single-Step Methods: (50 minutes) Download link, tablet notes

Explicit Runge-Kutta single step method cannot escape tight timestep constraints for stiff IVPs that may render them inefficient, see § 12.1.0.49. In this section we are going to augment the class of Runge-Kutta methods by timestepping schemes that can cope well with stiff IVPs.

12.3.1 The Implicit Euler Method for Stiff IVPs

EXPERIMENT 12.3.1.1 (Euler methods for stiff decay IVP) We revisit the setting of Ex. 12.1.0.3 and again consider Euler methods for the decay IVP

$$\dot{y} = \lambda y \quad , \quad y(0) = 1 \, , \quad \lambda < 0 \, .$$

We apply both the explicit Euler method (11.2.1.5) and the implicit Euler method (11.2.2.2) with uniform timesteps $h = 1/N$, $N \in \{5, 10, 20, 40, 80, 160, 320, 640\}$ and monitor the error at final time $T = 1$ for different values of λ.

Explicit Euler method (11.2.1.5)

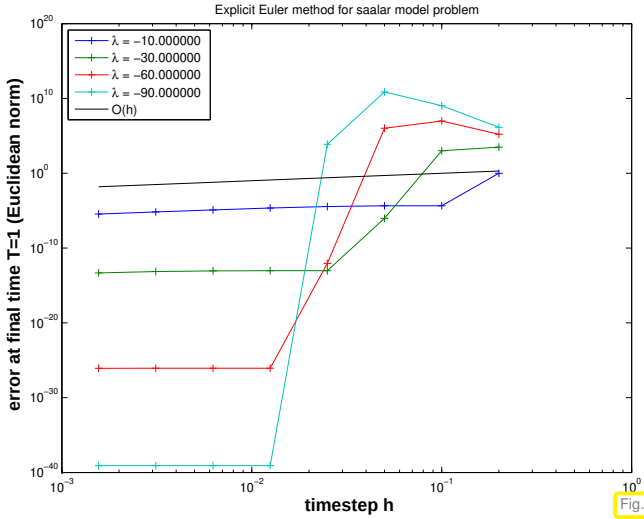


Fig. 460

λ large: blow-up of y_k for large timestep h

Implicit Euler method (11.2.2.2)

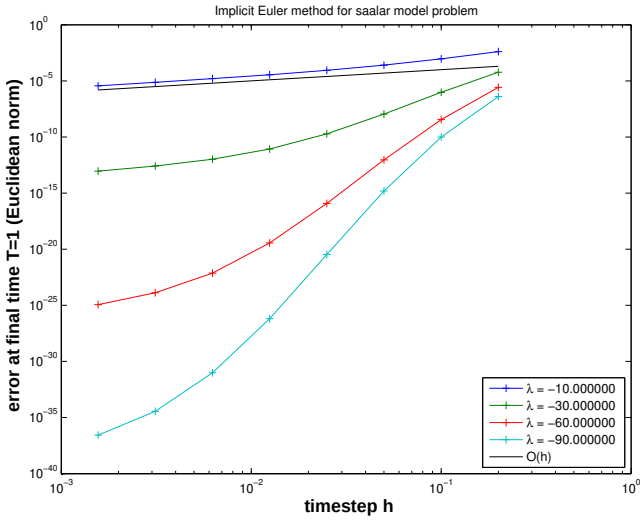


Fig. 461

λ large: stable for *all* timesteps $h > 0$!

We observe onset of convergence of the implicit Euler method already for large timesteps h. ⌟

§12.3.1.2 (Linear model problem analysis: implicit Euler method) We follow the considerations of § 12.1.0.4 and consider the *implicit* Euler method (11.2.2.2) for the

$$\text{linear model problem: } \dot{y} = \lambda y \, , \quad y(0) = y_0 \, , \quad \text{with } \operatorname{Re}\lambda \ll 0 \, , \tag{12.1.0.5}$$

with *exponentially decaying* (maybe osscillatory for $\operatorname{Im}\lambda \neq 0$) exact solution

$$y(t) = y_0 \exp(\lambda t) \to 0 \quad \text{for } t \to \infty \,.$$

The recursion of the implicit Euler method for (12.1.0.5) is defined by

$$\text{(11.2.2.2) for } \quad f(y) = \lambda y \quad \Rightarrow \quad y_{k+1} = y_k + \lambda h y_{k+1} \,, \quad k \in \mathbb{N}_0 \,. \tag{12.3.1.3}$$

$$\blacktriangleright \quad \text{generated sequence} \qquad y_k := \left(\frac{1}{1-\lambda h}\right)^k y_0 \,. \tag{12.3.1.4}$$

$$\blacktriangleright \quad \boxed{\operatorname{Re}\lambda < 0 \quad \Rightarrow \quad \lim_{k\to\infty} y_k = 0 \quad \forall h > 0\,!} \tag{12.3.1.5}$$

Without any timestep constraint we obtain the qualitatively correct behavior of $(y_k)_k$ for $\operatorname{Re}\lambda < 0$ and **any** $h > 0$!

As in § 12.1.0.40 this analysis can be extended to linear systems of ODEs $\dot{\mathbf{y}} = \mathbf{M}\mathbf{y}$, $\mathbf{M} \in \mathbb{C}^{N,N}$, by means of diagonalization.

As in § 12.1.0.30 and § 12.1.0.40 we assume that $\mathbf{M}$ can be diagonalized, that is (12.1.0.32) holds: $\mathbf{V}^{-1}\mathbf{M}\mathbf{V} = \mathbf{D}$ with a regular matrix $\mathbf{V} \in \mathbb{C}^{N,N}$ and a diagonal matrix $\mathbf{D} \in \mathbb{C}^{N,N}$ containing the eigenvalues $\lambda_1, \ldots, \lambda_N$ of $\mathbf{M}$ on its diagonal. Next, apply the decoupling by diagonalization idea to the recursion of the implicit Euler method.

$$\mathbf{V}^{-1}\mathbf{y}_{k+1} = \mathbf{V}^{-1}\mathbf{y}_k + h\underbrace{\mathbf{V}^{-1}\mathbf{M}\mathbf{V}}_{=\mathbf{D}}(\mathbf{V}^{-1}\mathbf{y}_{k+1}) \overset{\mathbf{z}_k := \mathbf{V}^{-1}\mathbf{y}_k}{\Leftrightarrow} \underbrace{(\mathbf{z}_{k+1})_i = \frac{1}{1-\lambda_i h}(\mathbf{z}_k)_i}_{\hat{=}\text{ implicit Euler step for } \dot{z}_i = \lambda_i z_i} \,. \tag{12.3.1.6}$$

Crucial insight:

For any timestep, the implicit Euler method generates exponentially decaying solution sequences $(\mathbf{y}_k)_{k=0}^{\infty}$ for $\dot{\mathbf{y}} = \mathbf{M}\mathbf{y}$ with diagonalizable matrix $\mathbf{M} \in \mathbb{R}^{N,N}$ with eigenvalues $\lambda_1, \ldots, \lambda_N$, if $\operatorname{Re}\lambda_i < 0$ for all $i = 1, \ldots, N$.

Thus we expect that the implicit Euler method will not face stability induced timestep constraints for stiff problems (→ Notion 12.2.0.7). ⌟

12.3.2 Collocation Single-Step Methods

Unfortunately the implicit Euler method is of first order only, see Exp. 11.3.2.5. This section presents an algorithm for designing higher order single step methods generalizing the implicit Euler method.

Setting: We consider the general ordinary differential equation $\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y})$, $\mathbf{f} : I \times D \to \mathbb{R}^N$ locally Lipschitz continuous, which guarantees the local existence of unique solutions of initial value problems, see Thm. 11.1.3.17.

We define the single step method through specifying the first step $\mathbf{y}_0 = \mathbf{y}(t_0) \to \mathbf{y}_1 \approx \mathbf{y}(t_1)$, where $\mathbf{y}_0 \in D$ is the initial step at initial time $t_0 \in I$, *cf.* Rem. 11.3.1.15. We assume that the exact solution trajectory $t \mapsto \mathbf{y}(t)$ exists on $[t_0, t_1]$. Use as a timestepping scheme on a temporal mesh (→ § 11.2.0.2) in the sense of Def. 11.3.1.5 is straightforward.

§12.3.2.1 (Collocation approach)

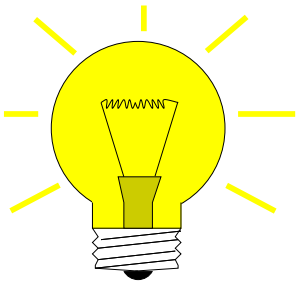
Abstract collocation idea

Collocation is a paradigm for the **discretization** (→ Rem. 11.3.1.4) of *differential equations*:

(I) Write the discrete solution u_h, a function, as linear combination of $N \in \mathbb{N}$ sufficiently smooth (basis) functions ➢ N unknown coefficients.

(II) Demand that u_h satisfies the differential equation at N points/times ➢ N equations.

We apply this policy to the differential equation $\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y})$ on $[t_0, t_1]$:



Idea:

❶ Approximate $t \mapsto \mathbf{y}(t)$, $t \in [t_0, t_1]$, by a function $t \mapsto \mathbf{y}_h(t) \in V$, V an $N \cdot (s+1)$-dimensional trial space V comprising functions $[t_0, t_1] \mapsto \mathbb{R}^N$, *cf.* Item (I).

❷ Fix $\mathbf{y}_h \in V$ by imposing collocation conditions

$$\begin{aligned} \mathbf{y}_h(t_0) &= \mathbf{y}_0 \,, \\ \dot{\mathbf{y}}_h(\tau_j) &= \mathbf{f}(\tau_j, \mathbf{y}_h(\tau_j)) \,, \quad j = 1, \ldots, s \,, \end{aligned} \tag{12.3.2.3}$$

for collocation points $t_0 \le \tau_1 < \ldots < \tau_s \le t_1$ → Item (II).

❸ Choose $\mathbf{y}_1 := \mathbf{y}_h(t_1)$.

⌟

§12.3.2.4 (Polynomial collocation) Existence of the function $\mathbf{y}_h : [t_0, t_1] \to \mathbb{R}^N$ satisfying (12.3.2.3) and the possibility to compute it efficiently will crucially depend on the choice of the trial space V.

Our choice (the "standard option"): (Componentwise) polynomial trial space $V = (\mathcal{P}_s)^N$

Recalling $\dim \mathcal{P}_s = s+1$ from Thm. 5.2.1.2, we see that our choice makes the number $= N(s+1)$ of collocation conditions matches the dimension of the trial space V.

Now we want to derive a concrete representation for the polynomial $\mathbf{y}_h$. We draw on concepts introduced in Section 5.2.2. We define the collocation points as

$$\tau_j := t_0 + c_j h \,, \quad j = 1, \ldots, s \,, \quad \text{for} \quad 0 \le c_1 < c_2 < \ldots < c_s \le 1 \,, \quad h := t_1 - t_0 \,.$$

Let $\{L_j\}_{j=1}^s \subset \mathcal{P}_{s-1}$ denote the set of Lagrange polynomials of degree $s-1$ associated with the node set $\{c_j\}_{j=1}^s$, see (5.2.2.4). They satisfy $L_j(c_i) = \delta_{ij}$, $i, j = 1, \ldots, s$ and form a basis of $\mathcal{P}_{s-1}$.

In each of its N components, the derivative $\dot{\mathbf{y}}_h$ is a polynomial of degree $s-1$: $\dot{\mathbf{y}} \in (\mathcal{P}_{s-1})^N$. Hence, it has the following representation, compare (5.2.2.6).

$$\dot{\mathbf{y}}_h(t_0 + \xi h) = \sum_{j=1}^{s} \dot{\mathbf{y}}_h(t_0 + c_j h) L_j(\xi) \,, \quad 0 \le \xi \le 1. \tag{12.3.2.5}$$

As $\tau_j = t_0 + c_j h$, the collocation conditions (12.3.2.3) make it possible to replace $\dot{\mathbf{y}}_h(c_j h)$ with an expres-

sion in the right hand side function $\mathbf{f}$:

(12.3.2.3)
$$\dot{\mathbf{y}}_h(t_0+\xi h) = \sum_{j=1}^{s} \mathbf{k}_j L_j(\xi) \quad \text{with "coefficients"} \quad \mathbf{k}_j := f(t_0+c_j h, \mathbf{y}_h(t_0+c_j h)) \,.$$

Next we integrate and use $\mathbf{y}_h(t_0) = \mathbf{y}_0$

$$\mathbf{y}_h(t_0+\xi h) = \mathbf{y}_0 + h\sum_{j=1}^{s} \mathbf{k}_j \int_0^{\xi} L_j(\zeta)\,\mathrm{d}\zeta \,.$$

This yields the following formulas for the computation of $\mathbf{y}_1$, which characterize the s-stage collocation single step method induced by the (normalized) collocation points $c_j \in [0,1]$, $j = 1,\ldots,s$.

$$\begin{aligned} \mathbf{k}_i &= f(t_0+c_i h, \mathbf{y}_0 + h\sum_{j=1}^{s} a_{ij}\mathbf{k}_j) \,, \\ \mathbf{y}_1 &:= \mathbf{y}_h(t_1) = \mathbf{y}_0 + h\sum_{i=1}^{s} b_i \mathbf{k}_i \,. \end{aligned} \quad \text{where} \quad \begin{aligned} a_{ij} &:= \int_0^{c_i} L_j(\tau)\,\mathrm{d}\tau \,, \\ b_i &:= \int_0^{1} L_i(\tau)\,\mathrm{d}\tau \,. \end{aligned} \tag{12.3.2.6}$$

Note that, since arbitrary $\mathbf{y}_0 \in D$, $t_0, t_1 \in I$ were admitted, this defines a discrete evolution $\boldsymbol{\Psi} : I \times I \times D \to \mathbb{R}^N$ by $\boldsymbol{\Psi}^{t_0,t_1}\mathbf{y}_0 := \mathbf{y}_h(t_1)$. ⌟

Remark 12.3.2.7 (Implicit nature of collocation single step methods) Note that (12.3.2.6) represents a generically non-linear system of $s \cdot N$ equations for the $s \cdot N$ components of the vectors $\mathbf{k}_i$, $i = 1,\ldots,s$. Usually, it will not be possible to obtain the increments $\mathbf{k}_i \in \mathbb{R}^N$ by a fixed number of evaluations of $\mathbf{f}$. For this reason the single step methods defined by (12.3.2.6) are called implicit.

With similar arguments as in Rem. 11.2.2.3 one can prove that for sufficiently small $|t_1 - t_0|$ a unique set of solution vectors $\mathbf{k}_1,\ldots,\mathbf{k}_s$ can be found. ⌟

§12.3.2.8 (Collocation single step methods and quadrature) Clearly, in the case $N = 1$, $f(t,\mathbf{y}) = f(t)$, $y_0 = 0$ the computation of y_1 boils down to the evaluation of a quadrature formula on $[t_0,t_1]$, because from (12.3.2.6) we get

$$y_1 = h\sum_{i=1}^{s} b_i f(t_0+c_i h) \,, \quad b_i := \int_0^1 L_i(\tau)\,\mathrm{d}\tau \,, \tag{12.3.2.9}$$

which is a polynomial quadrature formula (7.3.0.2) on $[0,1]$ with nodes c_j transformed to $[t_0,t_1]$ according to (7.2.0.5). ⌟

EXPERIMENT 12.3.2.10 (Empiric Convergence of collocation single step methods) We consider the initial value problem for the scalar logistic ODE

$$\dot{y} = \lambda y(1-y) \,, \quad y(0) = 0.01 \quad , \quad \lambda = 100 \,,$$

which is mildly stiff, over the time interval $[0,1]$

We perform numerical integration by timestepping with uniform timestep h based on a collocation single

step method (12.3.2.6).

❶ Equidistant collocation points, $c_j = \frac{j}{s+1}$, $j = 1, \cdots, s$.

We observe algebraic convergence with the empiric rates

$$\begin{aligned} s = 1 &: p = 1.96 \\ s = 2 &: p = 2.03 \\ s = 3 &: p = 4.00 \\ s = 4 &: p = 4.04 \end{aligned}$$

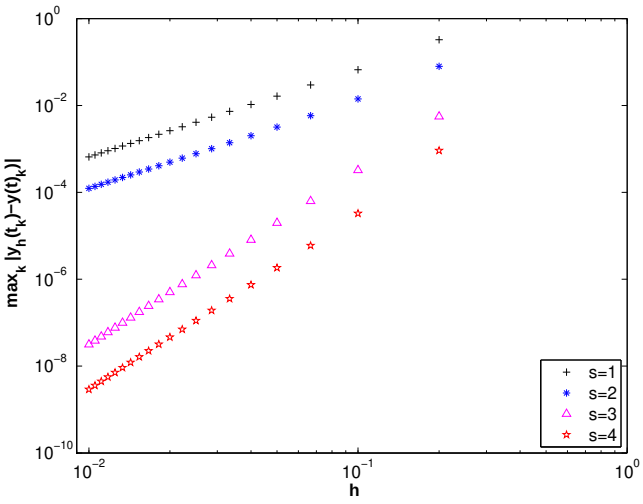


Fig. 462

In this case we conclude the following (empiric) order ($\to$ Def. 11.3.2.8) of the collocation single step method:

$$\text{(empiric) order} = \begin{cases} s & \text{for even } s\,, \\ s+1 & \text{for odd } s\,. \end{cases}$$

Next, we recall from § 7.4.2.15 an exceptional set of quadrature points, the Gauss points, provided by the zeros of the $L^2([-1,1])$-orthogonal Legendre polynomials, see Fig. 273.

❶ Gauss points in $[0,1]$ as normalized collocation points c_j, $j = 1, \ldots, s$.

We observe algebraic convergence with the empiric rates

$$\begin{aligned} s = 1 &: p = 1.96 \\ s = 2 &: p = 4.01 \\ s = 3 &: p = 6.00 \\ s = 4 &: p = 8.02 \end{aligned}$$

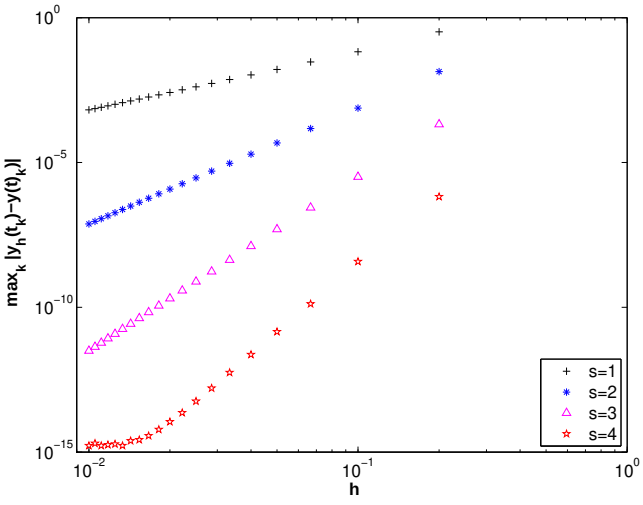


Fig. 463

Obviously, for the (empiric) order ($\to$ Def. 11.3.2.8) of the Gauss collocation single step method holds

$$\text{(empiric) order} = 2s\,.$$

Note that the 1-stage Gauss collocation single step method is the implicit midpoint method from Section 11.2.3. ⌟

§12.3.2.11 (Order of collocation single step method) What we have observed in Exp. 12.3.2.10 reflects

a fundamental result on collocation single step methods as defined in (12.3.2.6).

Theorem 12.3.2.12. Order of collocation single step method [DB02, Satz .6.40]

Provided that $\mathbf{f} \in C^p(I \times D)$, *the order (→ Def. 11.3.2.8) of an* s*-stage collocation single step method according to* (12.3.2.6) *agrees with the order (→ Def. 7.4.1.1) of the quadrature formula on* $[0,1]$ *with nodes* c_j *and weights* b_j, $j = 1, \ldots, s$.

This also explains the surprisingly high order of the Gauss collocation single-step method, because for s-point Gauss-Legendre numerical quadrature, the family of quadrature rules based on Gauss points as nodes, Section 7.4.2 derived the order $2s$.

➢ By Thm. 7.4.2.11 the s-stage Gauss collocation single step method whose nodes c_j are chosen as the s Gauss points on $[0,1]$ is of order $2s$.

⌟

12.3.3 General Implicit Runge-Kutta Single-Step Methods (RK-SSMs)

The notations in (12.3.2.6) have deliberately been chosen to allude to Def. 11.4.0.11. In that definition it takes only letting the sum in the formula for the increments run up to s to capture (12.3.2.6).

Definition 12.3.3.1. General Runge-Kutta single step method (*cf.* Def. 11.4.0.11)

For $b_i, a_{ij} \in \mathbb{R}$, $c_i := \sum_{j=1}^{s} a_{ij}$, $i, j = 1, \ldots, s$, $s \in \mathbb{N}$, a single step of size $h > 0$ of an **s-stage Runge-Kutta single step method** (RK-SSM) for the IVP (11.1.3.2) is defined by

$$\mathbf{k}_i := \mathbf{f}(t_0 + c_i h, \mathbf{y}_0 + h \sum_{j=1}^{s} a_{ij} \mathbf{k}_j) \;, \quad i = 1, \ldots, s \quad , \quad \mathbf{y}_1 := \mathbf{y}_0 + h \sum_{i=1}^{s} b_i \mathbf{k}_i \;.$$

As before, the vectors $\mathbf{k}_i \in \mathbb{R}^N$ are called **increments**.

Note that the computation of the increments $\mathbf{k}_i$ may now require the solution of *(non-linear) systems of equations* of size $s \cdot N$. In this case we speak about an "implicit" method, *cf.* Rem. 12.3.2.7.

The Butcher schenme notation introduced in (11.4.0.13) can easily be adapted to the the case of general RK-SSMs by dropping the requirement that the Butcher matrix be strictly lower triangular.

General Butcher scheme notation for RK-SSM

Shorthand notation for Runge-Kutta methods

Butcher scheme ▷

$$\begin{array}{c|c} \mathbf{c} & \mathfrak{A} \\ \hline & \mathbf{b}^T \end{array} := \begin{array}{c|ccc} c_1 & a_{11} & \cdots & a_{1s} \\ \vdots & \vdots & & \vdots \\ c_s & a_{s1} & \cdots & a_{ss} \\ \hline & b_1 & \cdots & b_s \end{array} \;. \qquad (12.3.3.3)$$

Note that now, in contrast to (11.4.0.13), $\mathfrak{A}$ can be a general $s \times s$-matrix.

Summary: terminology for Runge-Kutta single step methods:

$\mathfrak{A}$ strict lower triangular matrix	➤	explicit Runge-Kutta method, Def. 11.4.0.11
$\mathfrak{A}$ lower triangular matrix	➤	diagonally-implicit Runge-Kutta method (DIRK)

Many of the techniques and much of the theory discussed for explicit RK-SSMs carry over to general (implicit) Runge-Kutta single step methods:

- Sufficient condition for consistence from Cor. 11.4.0.15
- Algebraic convergence for meshwidth $h \to 0$ and the related concept of order ($\to$ Def. 11.3.2.8)
- Embedded methods and algorithms for adaptive stepsize control from Section 11.5

§12.3.3.4 (Butcher schemes for Gauss collocation RK-SSMs) As in (11.4.0.13) we can arrange the coefficients of Gauss collocation single-step methods in the form of a Butcher scheme and get

$$\text{for } s=1: \quad \begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array}\,, \tag{12.3.3.5a}$$

$$\text{for } s=2: \quad \begin{array}{c|cc} \frac{1}{2}-\frac{1}{6}\sqrt{3} & \frac{1}{4} & \frac{1}{4}-\frac{1}{6}\sqrt{3} \\ \frac{1}{2}+\frac{1}{6}\sqrt{3} & \frac{1}{4}+\frac{1}{6}\sqrt{3} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}\,, \tag{12.3.3.5b}$$

$$\text{for } s=3: \quad \begin{array}{c|ccc} \frac{1}{2}-\frac{1}{10}\sqrt{15} & \frac{5}{36} & \frac{2}{9}-\frac{1}{15}\sqrt{15} & \frac{5}{36}-\frac{1}{30}\sqrt{15} \\ \frac{1}{2} & \frac{5}{36}+\frac{1}{24}\sqrt{15} & \frac{2}{9} & \frac{5}{36}-\frac{1}{24}\sqrt{15} \\ \frac{1}{2}+\frac{1}{10}\sqrt{15} & \frac{5}{36}+\frac{1}{30}\sqrt{15} & \frac{2}{9}+\frac{1}{15}\sqrt{15} & \frac{5}{36} \\ \hline & \frac{5}{18} & \frac{4}{9} & \frac{5}{18} \end{array}\,. \tag{12.3.3.5c}$$

⌟

Remark 12.3.3.6 (Stage form equations for increments) In Def. 12.3.3.1 instead of the increments we can consider as unknowns the so-called **stages**

$$\mathbf{g}_i := h\sum_{j=1}^{s} a_{ij}\mathbf{k}_j \in \mathbb{R}^N\,, \quad i=1,\ldots,s\,, \quad \Leftrightarrow \quad \mathbf{k}_i = \mathbf{f}(t_0+c_ih, \mathbf{y}_0+\mathbf{g}_i)\,. \tag{12.3.3.7}$$

This leads to the equivalent defining equations in "stage form" for an implicit RK-SSM

$$\mathbf{k}_i := \mathbf{f}(t_0+c_ih, \mathbf{y}_0+h\sum_{j=1}^{s} a_{ij}\mathbf{k}_j)\,, \quad i=1,\ldots,s\,,$$

$$\Downarrow$$

$$\mathbf{g}_i = h\sum_{j=1}^{s} a_{ij}\mathbf{f}(t_0+c_ih, \mathbf{y}_0+\mathbf{g}_j) \quad , \quad \mathbf{y}_1 = \mathbf{y}_0 + h\sum_{i=1}^{s} b_i\mathbf{f}(t_0+c_jh, \mathbf{y}_0+\mathbf{g}_i)\,. \tag{12.3.3.8}$$

In terms of implementation there is no difference: Also the **stage equations** (12.3.3.8) are usually solved by means of Newton's method, see next remark. ⌟

Remark 12.3.3.9 (Solving the stage equations for implicit RK-SSMs) We reformulate the increment equations in stage form (12.3.3.8) as a non-linear system of equations in standard form $F(\mathbf{x}) = 0$.

Unknowns are the total $s \cdot N$ components of the stage vectors $\mathbf{g}_i \in \mathbb{R}^N$, $i = 1,\ldots,s$ as defined in (12.3.3.7).

$$\mathfrak{g} = [\mathbf{g}_1,\ldots,\mathbf{g}_s]^\top \in \mathbb{R}^{s\cdot N}\,, \quad \mathbf{g}_i := h\sum_{j=1}^{s} a_{ij}\mathbf{f}(t_0 + c_j h, \mathbf{y}_0 + \mathbf{g}_j) \quad \blacktriangleright \quad F(\mathfrak{g}) = \mathfrak{g} - h(\mathfrak{A}\otimes \mathbf{I}_N)\begin{bmatrix}\mathbf{f}(t_0+c_1h,\mathbf{y}_0+\mathbf{g}_1)\\ \vdots \\ \mathbf{f}(t_0+c_sh,\mathbf{y}_0+\mathbf{g}_s)\end{bmatrix} \stackrel{!}{=} \mathbf{0}\,,$$

where $\mathbf{I}_N$ is the $N \times N$ identity matrix and $\otimes$ designates the Kronecker product introduced in Def. 1.4.3.7.

We compute an approximate solution of $F(\mathfrak{g}) = \mathbf{0}$ iteratively by means of the simplified Newton method presented in Rem. 8.5.1.43. This is a Newton method with "frozen Jacobian". As $\mathfrak{g} \to \mathbf{0}$ for $h \to 0$, we choose zero as initial guess:

$$\mathfrak{g}^{(k+1)} = \mathfrak{g}^{(k)} - \mathrm{D}F(\mathbf{0})^{-1}F(\mathfrak{g}^{(k)}) \quad k = 0,1,2,\ldots \quad , \quad \mathfrak{g}^{(0)} = \mathbf{0}\,. \tag{12.3.3.10}$$

with the Jacobian

$$\mathrm{D}F(\mathbf{0}) = \begin{bmatrix}\mathbf{I} - ha_{11}\frac{\partial \mathbf{f}}{\partial \mathbf{y}}(t_0,\mathbf{y}_0) & \cdots & -ha_{1s}\frac{\partial \mathbf{f}}{\partial \mathbf{y}}(t_0,\mathbf{y}_0)\\ \vdots & \ddots & \vdots \\ -ha_{s1}\frac{\partial \mathbf{f}}{\partial \mathbf{y}}(t_0,\mathbf{y}_0) & \cdots & \mathbf{I} - ha_{ss}\frac{\partial \mathbf{f}}{\partial \mathbf{y}}(t_0,\mathbf{y}_0)\end{bmatrix} \in \mathbb{R}^{sN,sN}\,. \tag{12.3.3.11}$$

Obviously, $\mathrm{D}F(\mathbf{0}) \to \mathbf{I}$ for $h \to 0$. Thus, $\mathrm{D}F(\mathbf{0})$ will be regular for sufficiently small h.

In each step of the simplified Newton method we have to solve a linear system of equations with coefficient matrix $\mathrm{D}F(\mathbf{0})$. If $s \cdot N$ is large, an efficient implementation has to reuse the LU-decomposition of $\mathrm{D}F(\mathbf{0})$, see Code 8.5.1.44 and Rem. 2.5.0.10. ⌟

12.3.4 Model Problem Analysis for Implicit Runge-Kutta Single-Step Methods (IRK-SSMs)

Model problem analysis for general Runge-Kutta single step methods ($\to$ Def. 12.3.3.1) runs parallel to that for explicit RK-methods as elaborated in Section 12.1, § 12.1.0.13. Familiarity with the techniques and results of this section is assumed. The reader is asked to recall the concept of stability function from Thm. 12.1.0.17, the diagonalization technique from § 12.1.0.43, and the definition of region of (absolute) stability from Def. 12.1.0.51.

We apply the implicit RK-SSM according to Def. 12.3.3.1 to the autonomous linear scalar ODE $\dot{y} = \lambda y$, $\lambda \in \mathbb{C}$, and utterly parallel to the considerations in § 12.1.0.13, (12.1.0.14) we obtain

$$\blacktriangleright \quad k_i = \lambda(y_0 + h\sum_{j=1}^{s} a_{ij}k_j)\,, \quad y_1 = y_0 + h\sum_{i=1}^{s} b_i k_i \quad \Rightarrow \quad \begin{bmatrix}\mathbf{I} - z\mathfrak{A} & 0\\ -z\mathbf{b}^\top & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{k}\\ y_1\end{bmatrix} = y_0\begin{bmatrix}\mathbf{1}\\ 1\end{bmatrix}\,, \tag{12.3.4.1}$$

where $\boldsymbol{k} \in \mathbb{R}^s \mathrel{\hat{=}}$ denotes the vector $[k_1,\ldots,k_s]^\top/\lambda$ of increments, and $z := \lambda h$. As in § 12.1.0.13 we can eliminate the increments and obtain an expression for y_1:

$$y_1 = S(z)y_0 \quad \text{with} \quad S(z) := 1 + z\mathbf{b}^T(\mathbf{I} - z\mathfrak{A})^{-1}\mathbf{1}\,. \tag{12.3.4.2}$$

Alternatively, Cramer's rule supplies a formula for y_1 in terms of determinants:

$$y_1 = y_0 \frac{\det \begin{bmatrix} \mathbf{I} - z\mathfrak{A} & \mathbf{1} \\ -z\mathbf{b}^\top & 1 \end{bmatrix}}{\det \begin{bmatrix} \mathbf{I} - z\mathfrak{A} & 0 \\ -z\mathbf{b}^\top & 1 \end{bmatrix}} \quad \Rightarrow \quad S(z) = \frac{\det(\mathbf{I} - z\mathfrak{A} + z\mathbf{1}\mathbf{b}^T)}{\det(\mathbf{I} - z\mathfrak{A})} \,. \tag{12.3.4.3}$$

The next theorem summarizes these findings:

Theorem 12.3.4.4. Stability function of Runge-Kutta methods, *cf.* Thm. 12.1.0.17

The discrete evolution Ψ_λ^h of an s-stage Runge-Kutta single step method ($\to$ Def. 12.3.3.1) with Butcher scheme $\begin{array}{c|c} \mathbf{c} & \mathfrak{A} \\ \hline & \mathbf{b}^T \end{array}$ *(see* (12.3.3.3)*) for the ODE $\dot{y} = \lambda y$ is given by a multiplication with*

$$S(z) := \underbrace{1 + z\mathbf{b}^T(\mathbf{I} - z\mathfrak{A})^{-1}\mathbf{1}}_{\text{stability function}} = \frac{\det(\mathbf{I} - z\mathfrak{A} + z\mathbf{1}\mathbf{b}^T)}{\det(\mathbf{I} - z\mathfrak{A})} \,, \quad z := \lambda h \,, \quad \mathbf{1} = [1, \ldots, 1]^T \in \mathbb{R}^s \,.$$

EXAMPLE 12.3.4.5 (Regions of stability for simple implicit RK-SSM) We determine the Butcher schemes (12.3.3.3) for simple implicit RK-SSM and apply the formula from Thm. 12.3.4.4 to compute their stability functions.

- Implicit Euler method: $\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}$ ➢ $S(z) = \dfrac{1}{1-z} \,.$

- Implicit midpoint method: $\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array}$ ➢ $S(z) = \dfrac{1 + \frac{1}{2}z}{1 - \frac{1}{2}z} \,.$

Their regions of stability $\mathcal{S}_\Psi$ as defined in Def. 12.1.0.51,

$$\mathcal{S}_\Psi := \{z \in \mathbb{C} \colon |S(z)| < 1\} \subset \mathbb{C} \,,$$

can easily found from the respective stability functions:

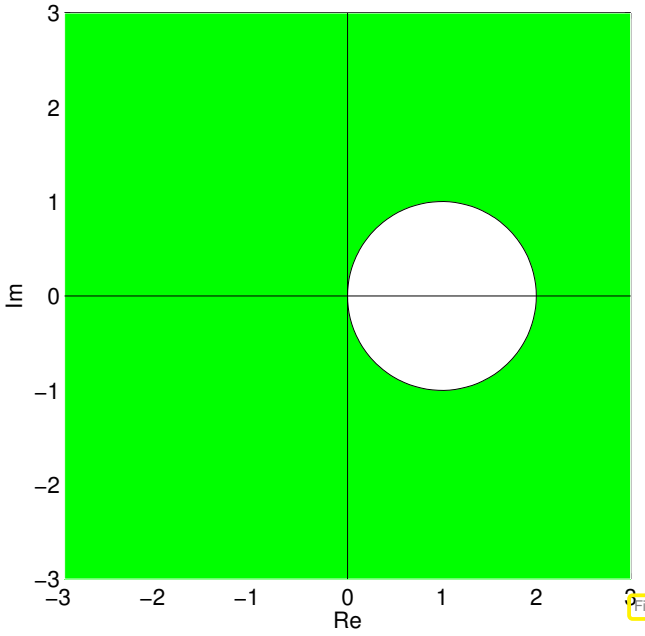


Fig. 464

$\mathcal{S}_\Psi$: implicit Euler method (11.2.2.2)

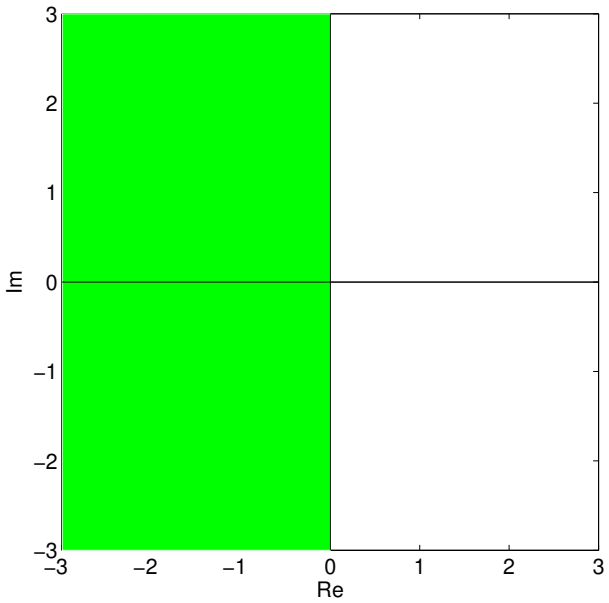


Fig. 465

$\mathcal{S}_\Psi$: implicit midpoint method (11.2.3.3)

We see that in both cases $|S(z)| < 1$, if $\operatorname{Re} z < 0$. ⌟

From the determinant formula for the stability function $S(z)$ we can conclude a generalization of Cor. 12.1.0.20.

Corollary 12.3.4.6. Rational stability function of explicit RK-SSM

For a consistent ($\rightarrow$ Def. 11.3.1.12) s-stage general Runge-Kutta single step method according to Def. 12.3.3.1 the stability function S is a non-constant rational function of the form $S(z) = \dfrac{P(z)}{Q(z)}$ with polynomials $P \in \mathcal{P}_s$, $Q \in \mathcal{P}_s$.

Of course, a rational function $z \mapsto S(z)$ can satisfy $\lim_{|z|\to\infty} |S(z)| < 1$ as we habe seen in Ex. 12.3.4.5. As a consequence, the region of stability for implicit RK-SSM need not be bounded.

§12.3.4.7 (A-stability) A general RK-SSM with stability function S applied to the scalar linear IVP $\dot{y} = \lambda y$, $y(0) = y_0 \in \mathbb{C}$, $\lambda \in \mathbb{C}$, with uniform timestep $h > 0$ will yield the sequence $(y_k)_{k=0}^{\infty}$ defined by

$$y_k = S(z)^k y_0 \quad , \quad z = \lambda h \, . \tag{12.3.4.8}$$

Hence, the next property of a RK-SSM guarantees that the sequence of approximations decays exponentially whenever the exact solution of the model problem IVP (12.1.0.5) does so.

Definition 12.3.4.9. A-stability of a Runge-Kutta single step method

A Runge-Kutta single step method with stability function S is **A-stable**, if

$$\mathbb{C}^- := \{z \in \mathbb{C} \colon \operatorname{Re} z < 0\} \subset \mathcal{S}_\Psi \, . \quad (\mathcal{S}_\Psi \hat{=} \text{ region of stability Def. 12.1.0.51})$$

From Ex. 12.3.4.5 we conclude that both the implicit Euler method and the implicit midpoint method are

A-stable.

A-stable Runge-Kutta single step methods will not be affected by stability induced timestep constraints when applied to stiff IVP ($\rightarrow$ Notion 12.2.0.7).

⌟

§12.3.4.10 ("Ideal" region of stability) In order to reproduce the qualitative behavior of the exact solution, a single step method when applied to the scalar linear IVP $\dot{y} = \lambda y$, $y(0) = y_0 \in \mathbb{C}$, $\lambda \in \mathbb{C}$, with uniform timestep $h > 0$,

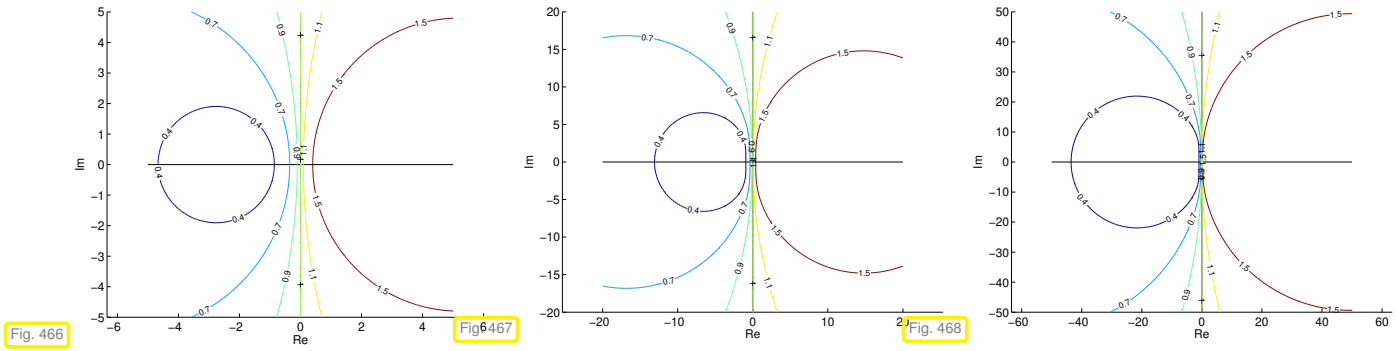
- should yield an *exponentially decaying* sequence $(y_k)_{k=0}^{\infty}$, whenever $\operatorname{Re}\lambda < 0$,
- should produce an *exponentially increasing* sequence sequence $(y_k)_{k=0}^{\infty}$, whenever $\operatorname{Re}\lambda > 0$.

Thus, in light of (12.3.4.8), we agree that the stability if

$$\text{"ideal" region of stability is} \quad \boxed{\mathcal{S}_\Psi = \mathbb{C}^-} \,. \tag{12.3.4.11}$$

Are there RK-SSMs that can boast of an ideal region of stability?

Regions of stability of Gauss collocation single step methods, see Exp. 12.3.2.10:



Implicit midpoint method $s = 2$ (order 4) $s = 4$ (order 8)

Level lines for $|S(z)|$ for Gauss collocation methods

Theorem 12.3.4.12. Region of stability of Gauss collocation single step methods [DB02, Satz 6.44]

s-stage Gauss collocation single step methods defined by (12.3.2.6) *with the nodes c_s given by the s Gauss points on $[0,1]$, feature the "ideal" stability domain:*

$$\mathcal{S}_\Psi = \mathbb{C}^- \,. \tag{12.3.4.11}$$

In particular, all Gauss collocation single step methods are A-stable.

⌟

EXPERIMENT 12.3.4.13 (Implicit RK-SSMs for stiff IVP) We consider the stiff IVP

$$\dot{y} = -\lambda y + \beta \sin(2\pi t) \,, \quad \lambda = 10^6, \beta = 10^6 \,, \quad y(0) = 1 \,,$$

whose solution essentially is the smooth function $t \mapsto \sin(2\pi t)$. Applying the criteria (12.2.0.13) and (12.2.0.14) we immediately see that this IVP is extremely stiff.

We solve it with different implicit RK-SSM on $[0,1]$ with large uniform timestep $h = \frac{1}{20}$.

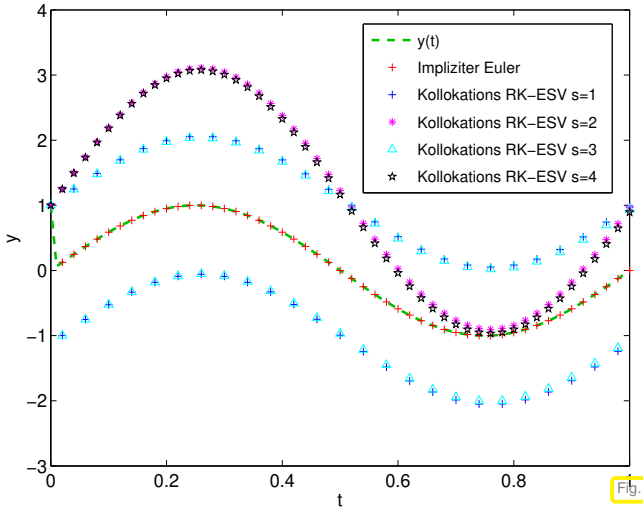


Fig. 469

Solutions by RK-SSMs

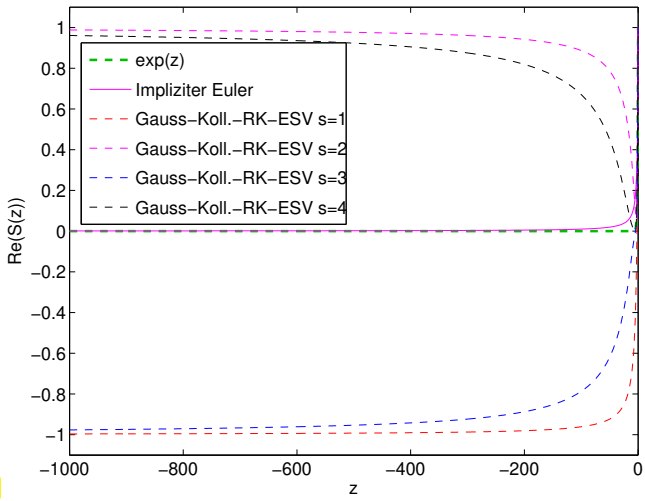


Fig. 470

Stability functions on $\mathbb{R}^-$

We observe that Gauss collocation RK-SSMs incur a huge discretization error, whereas the simple implicit Euler method provides a perfect approximation!

Explanation: The stability functions for Gauss collocation RK-SSMs satisfy

$$\lim_{|z|\to\infty} |S(z)| = 1 \,.$$

Hence, when they are applied to $\dot{y} = \lambda y$ with extremely large (in modulus) $\lambda < 0$, they will produce sequences that decay only very slowly or even oscillate, which misses the very rapid decay of the exact solution. The stability function for the implicity Euler method is $S(z) = (1-z)^{-1}$ and satisfies $\lim_{|z|\to\infty} S(z) = 0$, which will mean a fast exponential decay of the y_k. ⌟

§12.3.4.14 (L-stability) In light of what we learned in the previous experiment we can now state what we expect from the stability function of a Runge-Kutta method that is suitable for stiff IVP (→ Notion 12.2.0.7):

Definition 12.3.4.15. L-stable Runge-Kutta method → [Han02, Ch. 77]

A Runge-Kutta method (→ Def. 12.3.3.1) is L-stable/asymptotically stable, if its stability function (→ Thm. 12.3.4.4) satisfies

$$(i) \qquad \operatorname{Re} z < 0 \;\Rightarrow\; |S(z)| < 1 \,, \tag{12.3.4.16}$$

$$(ii) \qquad \lim_{\operatorname{Re} z \to -\infty} S(z) = 0 \,. \tag{12.3.4.17}$$

Remember: L-stable :⇔ A-stable & "$S(-\infty) = 0$" ⌟

Remark 12.3.4.18 (Necessary condition for L-stability of Runge-Kutta methods)

Consider a Runge-Kutta single step method (→ Def. 12.3.3.1) described by the Butcher scheme $\begin{array}{c|c} \mathbf{c} & \mathfrak{A} \\ \hline & \mathbf{b}^T \end{array}$.

Assume that $\mathfrak{A} \in \mathbb{R}^{s,s}$ is regular, which can be fulfilled only for an implicit RK-SSM.

For a rational function $S(z) = \frac{P(z)}{Q(z)}$ the limit for $|z| \to \infty$ exists and can easily be expressed by the leading coefficients of the polynomials P and Q:

$$\text{Thm. 12.3.4.4} \quad \Rightarrow \quad S(-\infty) = 1 - \mathbf{b}^T \mathfrak{A}^{-1} \mathbf{1} \,. \tag{12.3.4.19}$$

$$\blacktriangleright \quad \boxed{\text{If } \mathbf{b}^T = (\mathfrak{A})^T_{:,j} \text{ (row of } \mathfrak{A}) \quad \Rightarrow \quad S(-\infty) = 0} \,. \tag{12.3.4.20}$$

Butcher scheme (12.3.3.3) for L-stable RK-methods, see Def. 12.3.4.15

$$\triangleright \quad \begin{array}{c|c} \mathbf{c} & \mathfrak{A} \\ \hline & \mathbf{b}^T \end{array} := \begin{array}{c|ccc} c_1 & a_{11} & \cdots & a_{1s} \\ \vdots & \vdots & & \vdots \\ c_{s-1} & a_{s-1,1} & \cdots & a_{s-1,s} \\ 1 & b_1 & \cdots & b_s \\ \hline & b_1 & \cdots & b_s \end{array} .$$

A closer look at the coefficient formulas of (12.3.2.6) reveals that the algebraic condition (12.3.4.20) will automatically satisfied for a collocation single step method with $c_s = 1$! ⌟

EXAMPLE 12.3.4.21 (L-stable implicit Runge-Kutta methods) There is a family of s-point quadrature formulas on $[0,1]$ with a node located in 1 and (maximal) order $2s-1$: Gauss-Radau formulas. They induce the L-stable Gauss-Radau collocation single step methods of order $2s-1$ according to Thm. 12.3.2.12.

$$\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}$$

Implicit Euler method

$$\begin{array}{c|cc} \frac{1}{3} & \frac{5}{12} & -\frac{1}{12} \\ 1 & \frac{3}{4} & \frac{1}{4} \\ \hline & \frac{3}{4} & \frac{1}{4} \end{array}$$

Radau RK-SSM, order 3

$$\begin{array}{c|ccc} \frac{4-\sqrt{6}}{10} & \frac{88-7\sqrt{6}}{360} & \frac{296-169\sqrt{6}}{1800} & \frac{-2+3\sqrt{6}}{225} \\ \frac{4+\sqrt{6}}{10} & \frac{296+169\sqrt{6}}{1800} & \frac{88+7\sqrt{6}}{360} & \frac{-2-3\sqrt{6}}{225} \\ 1 & \frac{16-\sqrt{6}}{36} & \frac{16+\sqrt{6}}{36} & \frac{1}{9} \\ \hline & \frac{16-\sqrt{6}}{36} & \frac{16+\sqrt{6}}{36} & \frac{1}{9} \end{array}$$

Radau RK-SSM, order 5

The stability functions of s-stage Gauss-Radau collocation SSMs are rational functions of the form

$$S(z) = \frac{P(z)}{Q(z)} \,, \quad P \in \mathcal{P}_{s-1}, Q \in \mathcal{P}_s \,.$$

Beware that also "$S(\infty) = 0$", which means that Gauss-Radau methods when applied to problems with fast exponential blow-up may produce a spurious decaying solution.

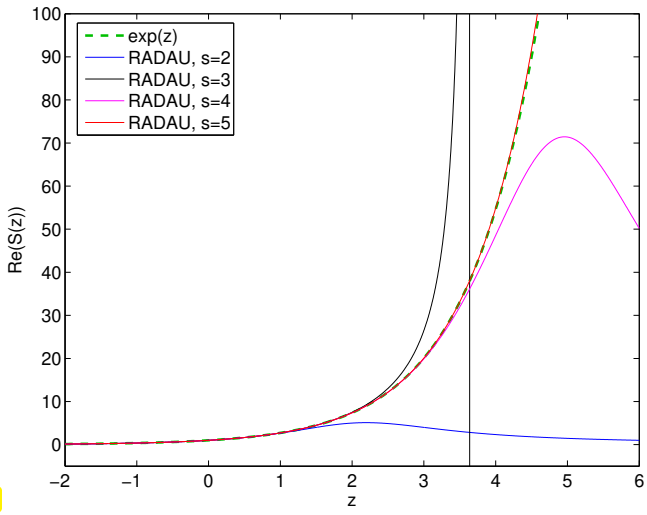
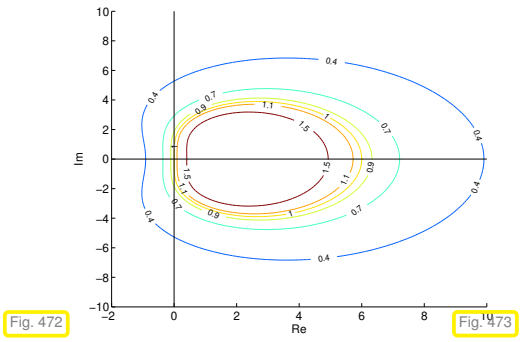
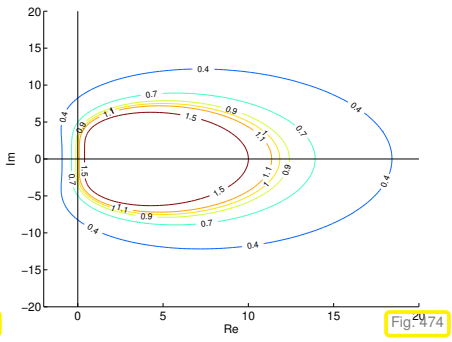


Fig. 471

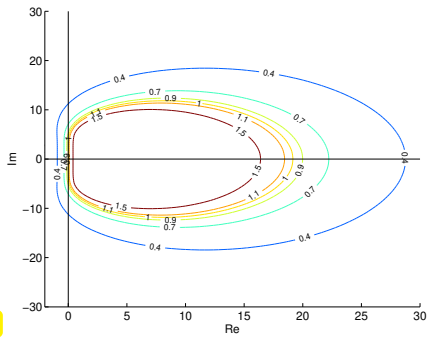
Level lines of stability functions of s-stage Gauss-Radau collocation SSMs:



$s = 2$



$s = 3$



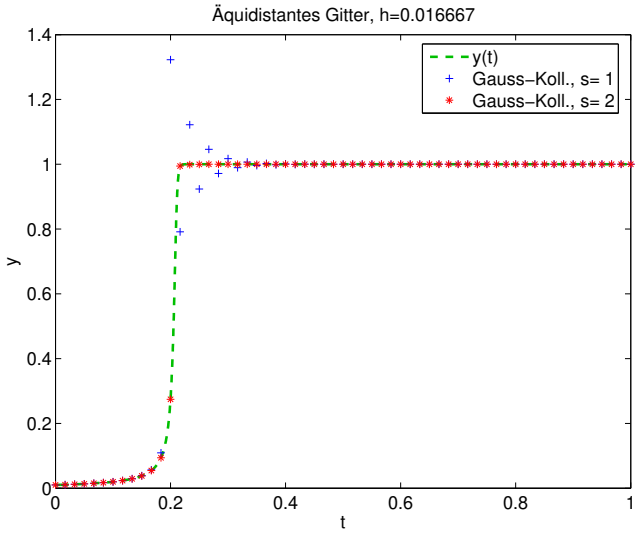
$s = 4$

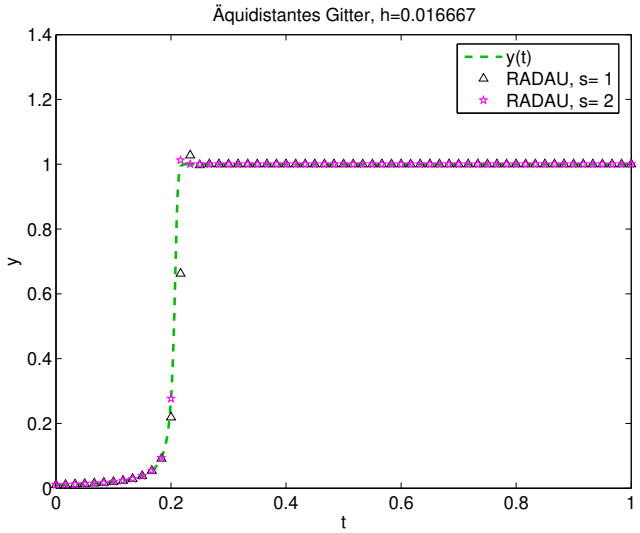
Further information about Radau-Runge-Kutta single step methods can be found in [Han02, Ch. 79]. ⌟

EXPERIMENT 12.3.4.22 (Gauss-Radau collocation SSM for stiff IVP) We revisit the stiff IVP from Ex. 12.0.0.1

$$\dot{y}(t) = \lambda y^2(1-y)\,, \quad \lambda = 500 \quad , \quad y(0) = \frac{1}{100}\,.$$

We compare the sequences generated by 1-stage and 2-stage Gauss collocation and Gauss-Radau collocation SSMs, respectively (uniform timestep).





The 2nd-order Gauss collocation SSM (implicit midpoint method) suffers from spurious oscillations when homing in on the stable stationary state $y = 1$. The explanation from Exp. 12.3.4.13 also applies to this example.

The fourth-order Gauss method is already so accurate that potential overshoots when approaching $y = 1$ are damped fast enough. ⌟

Review question(s) 12.3.4.23 (Implicit Runge-Kutta single-step methods)

(Q12.3.4.23.A) The stability function for the implicit Euler method for the ODE $\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y})$,

$$\mathbf{y}_{k+1} = \mathbf{y}_k + h_k \mathbf{f}(t_{k+1}, \mathbf{y}_{k+1})\,, \quad h_k := t_{k+1} - t_k\,,$$

is $S(z) = \dfrac{1}{1-z}$.

When will one observe a totally wrong qualitative behavior of the sequence (y_k) of states generated by the implicit Euler method applied to the scalar growth ODE $\dot{y} = \lambda y$, $\lambda > 0$?

(Q12.3.4.23.B) Gauss collocation RK-SSMs possess the *ideal stability domain* $\mathcal{S}_\Psi = \mathbb{C}^- := \{z \in \mathbb{C} : \operatorname{Re} z < 0\}$. Argue, why necessarily $\lim_{z\to\infty} S(z) = \pm 1$.

Hint. What kind of function is the stability function for an implicit RK-SSM?

(Q12.3.4.23.C) We apply a general Runge-Kutta single-step method to the autonomous affine-linear ODE $\dot{\mathbf{y}} = \mathbf{M}\mathbf{y} + \mathbf{b}$, $\mathbf{M} \in \mathbb{R}^{N,N}$, $\mathbf{b} \in \mathbb{R}^N$, $N \in \mathbb{N}$. Describe the linear system of equations that has to be solved in every timestep.

The definition of a general Runge-Kutta single-step method applied to the ODE is $\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y})$ is as follows:

Definition 12.3.3.1. Runge-Kutta single-step method

For $b_i, a_{ij} \in \mathbb{R}$, $c_i := \sum_{j=1}^s a_{ij}$, $i, j = 1, \ldots, s$, $s \in \mathbb{N}$, a single step of size $h > 0$ of an **s-stage Runge-Kutta single step method** (RK-SSM) for the IVP (11.1.3.2) is defined by

$$\mathbf{k}_i := \mathbf{f}(t_0 + c_i h, \mathbf{y}_0 + h \sum_{j=1}^s a_{ij} \mathbf{k}_j) \;, \quad i = 1, \ldots, s \quad , \quad \mathbf{y}_1 := \mathbf{y}_0 + h \sum_{i=1}^s b_i \mathbf{k}_i \;.$$

As before, the vectors $\mathbf{k}_i \in \mathbb{R}^N$ are called **increments**.

(Q12.3.4.23.D) Show that the single-step methods arising from the polynomial collocation approach with $s \in \mathbb{N}$ collocation points will always be consistent.

Hint. The general formulas for a single-step method constructed via the polynomial collocation approach with normalized collocation points $c_1, c_2, \ldots, c_s$ are

$$\begin{aligned} \mathbf{k}_i &= f(t_0 + c_i h, \mathbf{y}_0 + h \sum_{j=1}^s a_{ij} \mathbf{k}_j) \;, \\ \mathbf{y}_1 := \mathbf{y}_h(t_1) &= \mathbf{y}_0 + h \sum_{i=1}^s b_i \mathbf{k}_i \;. \end{aligned} \qquad \text{where} \quad \begin{aligned} a_{ij} &:= \int_0^{c_i} L_j(\tau)\,\mathrm{d}\tau \;, \\ b_i &:= \int_0^1 L_i(\tau)\,\mathrm{d}\tau \;. \end{aligned} \tag{12.3.2.6}$$

where $\{L_1, \ldots, L_s\} \subset \mathcal{P}_{s-1}$ are the Lagrange polynomials associated with the node set $\{c_1, c_2, \ldots, c_s\}$ on $[0, 1]$.

Also remember that a single-step method for the ODE $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$ is consistent, if and only if, its associated discrete evolution is of the form

$$\Psi^h \mathbf{y} = \mathbf{y} + h\boldsymbol{\psi}(h, \mathbf{y}) \qquad \text{with} \quad \begin{aligned} &\boldsymbol{\psi} : I \times D \to \mathbb{R}^N \text{ continuous,} \\ &\boldsymbol{\psi}(0, \mathbf{y}) = \mathbf{f}(\mathbf{y}) \;. \end{aligned} \tag{11.3.1.11}$$

(Q12.3.4.23.E) Let $\begin{array}{c|c} \mathbf{c} & \mathfrak{A} \\ \hline & \mathbf{b}^\top \end{array}$ be the Butcher scheme for an s-stage collocation single-step method. Show that

$$(\mathfrak{A})_{s,:} = \mathbf{b}^\top \;,$$

which, for an A-stable method is a sufficient condition for L-stability.

△

Supplementary literature. [DR08, Sect. 11.6.2], [QSS00, Sect. 11.8.3]

12.4 Semi-Implicit Runge-Kutta Methods

Video tutorial for Section 12.4: Semi-Implicit Runge-Kutta Methods: (13 minutes) Download link, tablet notes

From Section 12.3.3 recall the formulas for general/implicit Runge-Kutta single-step methods for the ODE $\dot{\mathbf{y}} = \mathbf{f}(t,\mathbf{y})$:

Definition 12.3.3.1. General Runge-Kutta single-step method

For $b_i, a_{ij} \in \mathbb{R}$, $c_i := \sum_{j=1}^{s} a_{ij}$, $i,j = 1,\ldots,s$, $s \in \mathbb{N}$, an s-stage Runge-Kutta single step method (RK-SSM) for the IVP (11.1.3.2) is defined by

$$\mathbf{k}_i := \mathbf{f}(t_0 + c_i h, \mathbf{y}_0 + h\sum_{j=1}^{s} a_{ij}\mathbf{k}_j)\ , \quad i = 1,\ldots,s \quad , \quad \mathbf{y}_1 := \mathbf{y}_0 + h\sum_{i=1}^{s} b_i \mathbf{k}_i\ .$$

As before, the vectors $\mathbf{k}_i \in \mathbb{R}^N$ are called **increments**.

The equations fixing the increments $\mathbf{k}_i \in \mathbb{R}^N$, $i = 1,\ldots,s$, for an s-stage implicit RK-method constitute a (Non-)linear system of equations with $s \cdot N$ unknowns.

Several expensive (Newton) iterations needed to find $\mathbf{k}_i$?

Remember that we compute approximate solutions anyway, and the increments *are weighted with the stepsize* $h \ll 1$, see Def. 12.3.3.1. So there is no point in determining them with high accuracy!

Idea: Use only a fixed *small* number of Newton steps to solve for the $\mathbf{k}_i$, $i = 1,\ldots,s$.

Extreme case: use only a single Newton step! Let's try.

EXAMPLE 12.4.0.1 (Semi-implicit Euler single-step method) We apply the above idea to the implicit Euler method introduced in Section 11.2.2. For the sake of simplicity we consider the autonomous ODE $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$, $\mathbf{f} : D \subset \mathbb{R}^N \to \mathbb{R}^N$.

The recursion for the implicit Euler method with (local) stepsize $h > 0$ is

$$\mathbf{y}_{k+1}\colon \quad \mathbf{y}_{k+1} = \mathbf{y}_k + h\mathbf{f}(\mathbf{y}_{k+1})\ . \tag{11.2.2.2}$$

We recast it as a non-linear system of N equations in "standard form $F(\mathbf{x}) = \mathbf{0}$":

$$\mathbf{y}_{k+1} = \mathbf{y}_k + h\mathbf{f}(\mathbf{y}_{k+1}) \quad \Leftrightarrow \quad F(\mathbf{y}_{k+1}) := \mathbf{y}_{k+1} - h\mathbf{f}(\mathbf{y}_{k+1}) - \mathbf{y}_k = 0\ .$$

A single Newton step (8.5.1.6) applied to $F(\mathbf{y}) = 0$ with the natural initial guess $\mathbf{y}_k$ yields

$$\mathbf{y}_{k+1} = \mathbf{y}_k - \mathrm{D}\,\mathbf{f}(\mathbf{y}_k)^{-1}F(\mathbf{y}_k) = \mathbf{y}_k + (\mathbf{I} - hD\mathbf{f}(\mathbf{y}_k))^{-1}h\mathbf{f}(\mathbf{y}_k)\ . \tag{12.4.0.2}$$

This defines the recursion for the **semi-implicit Euler method**.

Note that for a linear ODE with $\mathbf{f}(\mathbf{y}) = \mathbf{M}\mathbf{y}$, $\mathbf{M} \in \mathbb{R}^{N,N}$, we recover the original implicit Euler method! ⌟

EXPERIMENT 12.4.0.3 (Empiric convergence of semi-implicit Euler single-step method)

- We consider an Initial value problem for logistic ODE, see Ex. 11.1.2.1

$$\dot{y} = \lambda y(1-y) \quad , \quad y(0) = 0.1 \quad , \quad \lambda = 5 \, .$$

- We run the implicit Euler method (11.2.2.2) and the semi-implicit Euler method (12.4.0.2) with uniform timestep $h = 1/n$, $n \in \{5, 8, 11, 17, 25, 38, 57, 85, 128, 192, 288, 432, 649, 973, 1460, 2189, 3284, 4926, 7389\}$.
- Measured error $\mathrm{err} = \max\limits_{j=1,\ldots,n} |y_j - y(t_j)|$

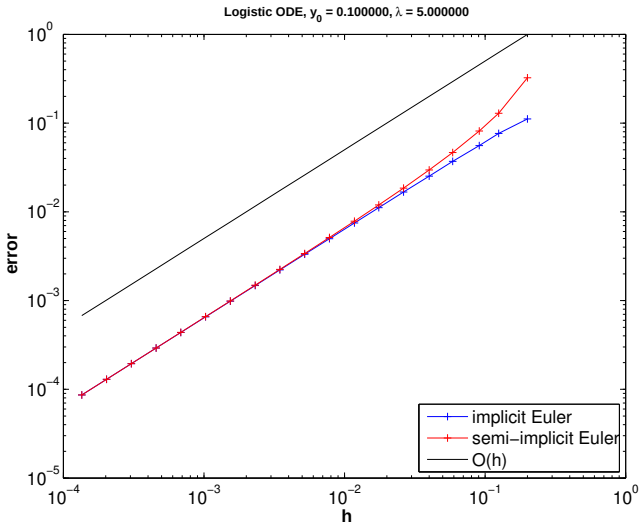


Fig. 475

We observe that the approximate solution of the defining equation for y_{k+1} by a single Newton step preserves the 1st-order convergence of the implicit Newton method. Also the semi-implicit Euler methods seems to be of first order. ⌟

EXPERIMENT 12.4.0.4 (Convergence of semi-implicit midpoint method) Again, we tackle the IVP from Exp. 12.4.0.3.

- Now, implicit midpoint method (11.2.3.3) uniform timesteps $h = 1/n$ as above

  & approximate computation of y_{k+1} by **1 Newton step**, initial guess y_k
- Measured error $\mathrm{err} = \max\limits_{j=1,\ldots,n} |y_j - y(t_j)|$

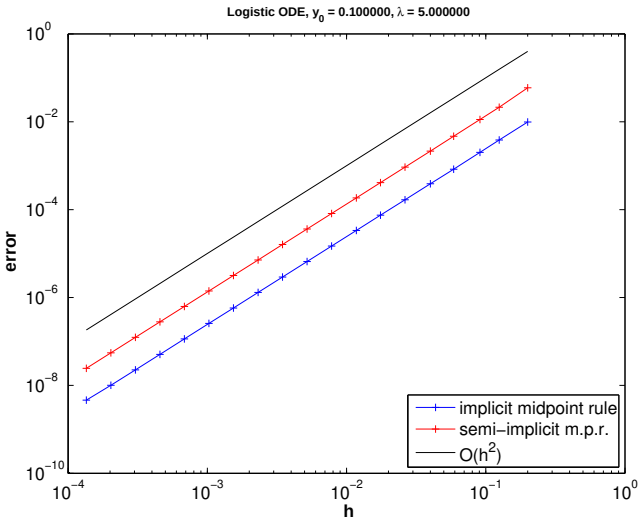
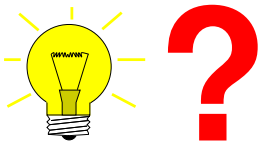


Fig. 476

We still observe second-order convergence! ⌟

Try: Use linearized increment equations for implicit RK-SSM



$$\mathbf{k}_i := \mathbf{f}(\mathbf{y}_0 + h\sum_{j=1}^{s} a_{ij}\mathbf{k}_j) \, , \quad i = 1, \ldots, s$$

$$\Downarrow$$

$$\mathbf{k}_i = \mathbf{f}(\mathbf{y}_0) + h\,\mathrm{D}\,\mathbf{f}(\mathbf{y}_0)\left(\sum_{j=1}^{s} a_{ij}\mathbf{k}_j\right) \, , \quad i = 1, \ldots, s \, . \tag{12.4.0.5}$$

The good news is that all results about stability derived from model problem analysis (→ Section 12.1) remain valid despite linearization of the increment equations:

Linearization does nothing for linear ODEs ➢ stability function (→ Thm. 12.3.4.4) not affected!

The bad news is that the preservation of the order observed in Exp. 12.4.0.3 will no longer hold in the general case.

EXPERIMENT 12.4.0.6 (Convergence of naive semi-implicit Radau method)

- We consider an IVP for the logistic ODE from Ex. 11.1.2.1:

$$\dot{y} = \lambda y(1-y) \quad , \quad y(0) = 0.1 \quad , \quad \lambda = 5 \, .$$

- 2-stage Radau RK-SSM, Butcher scheme

$$\begin{array}{c|cc} \frac{1}{3} & \frac{5}{12} & -\frac{1}{12} \\ 1 & \frac{3}{4} & \frac{1}{4} \\ \hline & \frac{3}{4} & \frac{1}{4} \end{array} \, , \qquad (12.4.0.7)$$

  order = 3, see Ex. 12.3.4.21.

- Increments from linearized equations (12.4.0.5)
- We monitor the error through $\mathrm{err} = \max_{j=1,\ldots,n} |y_j - y(t_j)|$

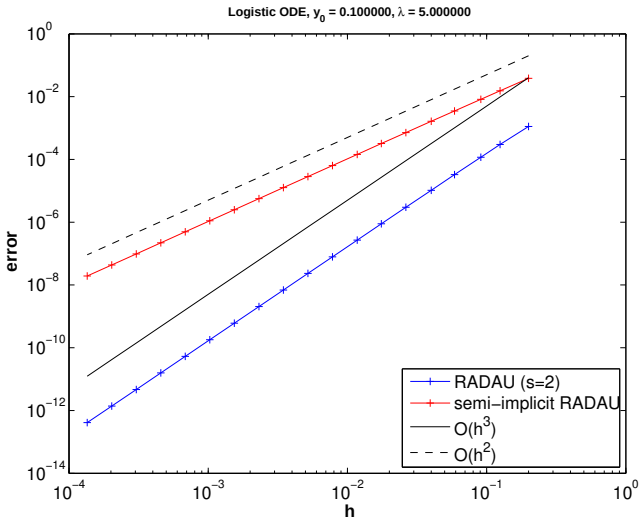


Fig. 477

Loss of order due to linearization !

§12.4.0.8 (Rosenbrock-Wanner methods) We have just seen that the simple linearization according to (12.4.0.5) will degrade the order of implicit RK-SSMs and leads to a substantial loss of accuracy. This is not an option.

Yet, the idea behind (12.4.0.5) has been refined. One does not start from a known RK-SSM, but introduces general coefficients for structurally linear increment equations.

Class of s-stage semi-implicit (linearly implicit) Runge-Kutta methods (Rosenbrock-Wanner (ROW) methods):

$$\begin{aligned} (\mathbf{I} - ha_{ii}\mathbf{J})\mathbf{k}_i &= \mathbf{f}(\mathbf{y}_0 + h\sum_{j=1}^{i-1}(a_{ij}+d_{ij})\mathbf{k}_j) - h\mathbf{J}\sum_{j=1}^{i-1} d_{ij}\mathbf{k}_j \, , \quad \mathbf{J} = \mathrm{D}\,\mathbf{f}(\mathbf{y}_0) \, , \\ \mathbf{y}_1 &:= \mathbf{y}_0 + h\sum_{j=1}^{s} b_j \mathbf{k}_j \, . \end{aligned} \qquad (12.4.0.9)$$

Then the coefficients a_{ij}, d_{ij}, and b_i are determined from order conditions by solving large non-linear systems of equations.

In each step s linear systems with coefficient matrices $\mathbf{I} - ha_{ii}\mathbf{J}$ have to be solved. For methods used in practice one often demands that $a_{ii} = \gamma$ for all $i = 1, \ldots, s$. As a consequence, we have to solve s linear

systems with the *same* coefficient matrix $\mathbf{I} - h\gamma\mathbf{J} \in \mathbb{R}^{N,N}$, which permits us to reuse LU-factorizations, see Rem. 2.5.0.10. ⌟

Supplementary literature. A related discussion can be found in [Han02, Ch. 80].

Review question(s) 12.4.0.10 (Semi-implicit Runge-Kutta single-step methods)

(Q12.4.0.10.A) [Semi-implicit midpoint method] The implicit midpoint single-step method applied to the autonomous ODE $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$ and with timestep h leads to the recursion

$$\mathbf{y}_{k+1}: \quad \mathbf{y}_{k+1} = \mathbf{y}_k + h\mathbf{f}(\tfrac{1}{2}(\mathbf{y}_k + \mathbf{y}_{k+1})) \ .$$

Derive the defining equation of the semi-implicit variant, which arises from solving the defining equation for $\mathbf{y}_{k+1}$ by a single Newton step with initial guess $\mathbf{y}_k$.

(Q12.4.0.10.B) [Stability function of ROW-SSM] A Rosenbrock-Wanner (ROW) single-step method for the autonomous ODE $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$ can be defined by

$$\begin{aligned} (\mathbf{I} - ha_{ii}\mathbf{J})\mathbf{k}_i &= \mathbf{f}(\mathbf{y}_0 + h\sum_{j=1}^{i-1}(a_{ij}+d_{ij})\mathbf{k}_j) - h\mathbf{J}\sum_{j=1}^{i-1} d_{ij}\mathbf{k}_j \ , \quad \mathbf{J} = \mathrm{D}\,\mathbf{f}(\mathbf{y}_0) \ , \\ \mathbf{y}_1 &:= \mathbf{y}_0 + h\sum_{j=1}^{s} b_j\mathbf{k}_j \ . \end{aligned} \tag{12.4.0.9}$$

Derive its stability functions for $s = 2$.

△

12.5 Splitting Methods

Video tutorial for Section 12.5: Splitting Methods: (21 minutes) Download link, tablet notes

§12.5.0.1 (Splitting idea: composition of partial evolutions) Many relevant ordinary differential equations feature a right hand side function that is the sum to two (or more) terms. Consider an autonomous IVP with a right hand side function that can be split in an additive fashion:

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}) + \mathbf{g}(\mathbf{y}) \quad , \quad \mathbf{y}(0) = \mathbf{y}_0 \ , \tag{12.5.0.2}$$

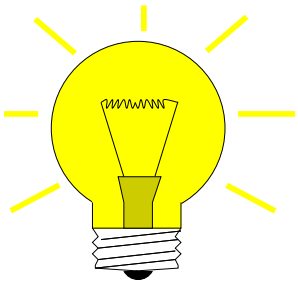
with $\mathbf{f} : D \subset \mathbb{R}^N \mapsto \mathbb{R}^N$, $\mathbf{g} : D \subset \mathbb{R}^N \mapsto \mathbb{R}^N$ "sufficiently smooth", locally Lipschitz continuous (→ Def. 11.1.3.13).

Let us introduce the evolution operators (→ Def. 11.1.4.3) for both summands:

$$\text{(Continuous) evolution maps:} \quad \begin{aligned} \boldsymbol{\Phi}_f^t &\leftrightarrow \text{ODE} \quad \dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}) \ , \\ \boldsymbol{\Phi}_g^t &\leftrightarrow \text{ODE} \quad \dot{\mathbf{y}} = \mathbf{g}(\mathbf{y}) \ . \end{aligned}$$

Temporarily we assume that both $\boldsymbol{\Phi}_f^t$, $\boldsymbol{\Phi}_g^t$ are available in the form of analytic formulas or highly accurate

approximations.



Idea: Build single step methods ($\rightarrow$ Def. 11.3.1.5) based on the following discrete evolutions

$$\text{Lie-Trotter splitting:} \quad \Psi^h = \Phi_g^h \circ \Phi_f^h \,, \tag{12.5.0.3}$$

$$\text{Strang splitting:} \quad \Psi^h = \Phi_f^{h/2} \circ \Phi_g^h \circ \Phi_f^{h/2} \,. \tag{12.5.0.4}$$

These splittings are easily remembered in graphical form:

(12.5.0.3) $\leftrightarrow$

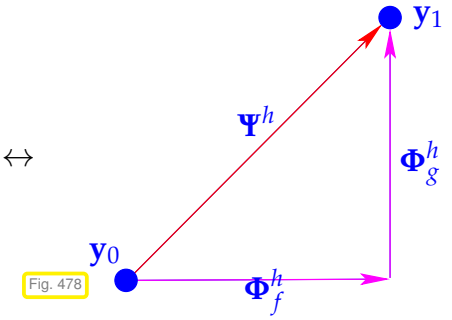


Fig. 478

(12.5.0.4) $\leftrightarrow$

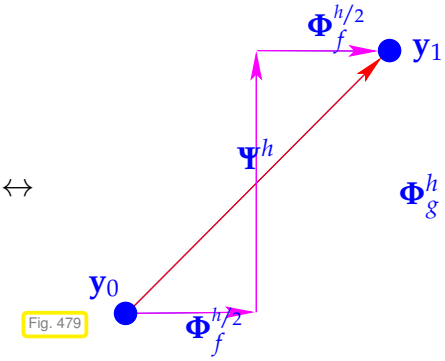


Fig. 479

Note that over many timesteps the Strang splitting approach is not more expensive than Lie-Trotter splitting, because the actual implementation of (12.5.0.4) should be done as follows:

$$\begin{aligned} \mathbf{y}_{1/2} &:= \Phi_f^{h/2} \,, & \mathbf{y}_1 &:= \Phi_g^h \mathbf{y}_{1/2} \,, \\ \mathbf{y}_{3/2} &:= \Phi_f^h \mathbf{y}_1 \,, & \mathbf{y}_2 &:= \Phi_g^h \mathbf{y}_{3/2} \,, \\ \mathbf{y}_{5/2} &:= \Phi_f^h \mathbf{y}_2 \,, & \mathbf{y}_3 &:= \Phi_g^h \mathbf{y}_{5/2} \,, \\ \vdots & & \vdots & \,, \end{aligned}$$

because $\Phi_f^{h/2} \circ \Phi_f^{h/2} = \Phi_f^h$. This means that a Strang splitting SSM differs from a Lie-Trotter splitting SSM in the first and the last step only. ⌟

EXPERIMENT 12.5.0.5 (Convergence of simple splitting methods) We consider the following IVP whose right hand side function is the sum of two functions for which the ODEs can be solved analytically:

$$\dot{y} = \underbrace{\lambda y(1-y)}_{=:f(y)} + \underbrace{\sqrt{1-y^2}}_{=:g(y)} \quad , \quad y(0) = 0 \,.$$

- $\Phi_f^t y = \dfrac{1}{1+(y^{-1}-1)e^{-\lambda t}} \,, t > 0, y \in]0,1]$ (logistic ODE (11.1.2.2))
- $\Phi_g^t y = \begin{cases} \sin(t + \arcsin(y)) & \text{, if } t + \arcsin(y) < \frac{\pi}{2} \,, \\ 1 & \text{, else,} \end{cases} \quad t > 0, y \in [0,1] \,.$

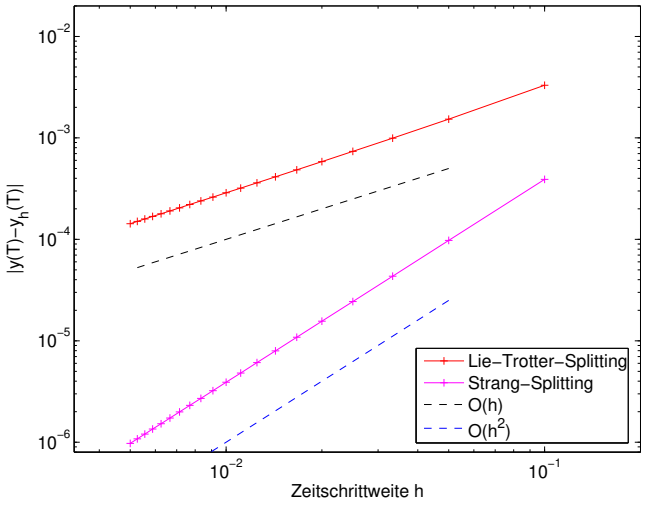


Fig. 480

Numerical experiment:

For $T = 1$, $\lambda = 1$, we compare the two splitting methods for uniform timesteps with a very accurate reference solution obtained in MATLAB by

```
f=@(t,x) λ*x*(1-x)+sqrt(1-x^2);
options=odeset('reltol',1.0e-10,...
               'abstol',1.0e-12);
[t,yex]=ode45(f,[0,1],y0,options);
```

◁ Error at final time $T = 1$

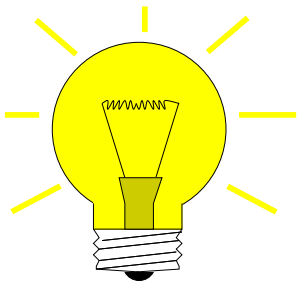
We observe algebraic convergence of the two splitting methods, order 1 for (12.5.0.3), oder 2 for (12.5.0.4). ⌟

The observation made in Exp. 12.5.0.5 reflects a general truth:

Theorem 12.5.0.6. Order of simple splitting methods

Die single step methods defined by (12.5.0.3) *or* (12.5.0.4) *are of order (→ Def. 11.3.2.8)* 1 *and* 2*, respetively.*

§12.5.0.7 (Inexact splitting methods) Of course, the assumption that $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$ and $\dot{\mathbf{y}} = \mathbf{g}(\mathbf{y})$ can be solved exactly will hardly ever be met. However, it should be clear that a "sufficiently accurate" approximation of the evolution maps Φ_g^h and $\boldsymbol{\Phi}_f^h$ is all we need



Idea: In (12.5.0.3)/(12.5.0.4) replace

exact evolutions $\longrightarrow$ discrete evolutions

$\boldsymbol{\Phi}_g^h$, $\boldsymbol{\Phi}_f^h$ $\longrightarrow$ $\boldsymbol{\Psi}_g^h$, $\boldsymbol{\Psi}_f^h$.

EXPERIMENT 12.5.0.8 (Convergence of inexact simple splitting methods) Again we consider the IVP of Exp. 12.5.0.5 and inexact splitting methods based on different single step methods for the two ODE corresponding to the summands.

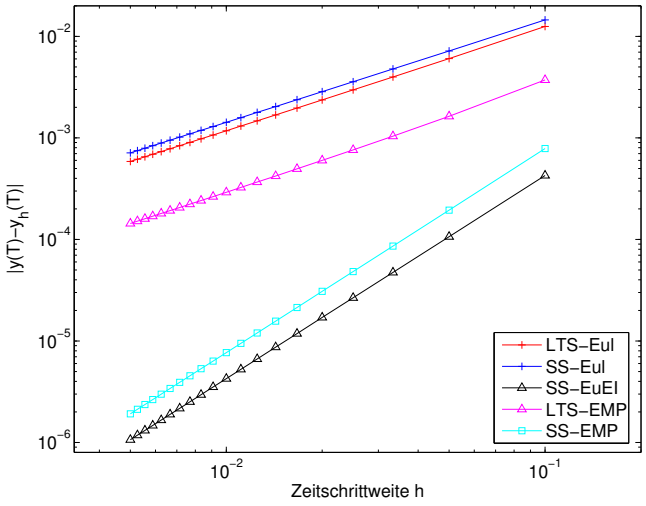


Fig. 481

LTS-Eul	explicit Euler method (11.2.1.5) $\to \boldsymbol{\Psi}^h_{h,g}$, $\boldsymbol{\Psi}^h_{h,f}$ + Lie-Trotter splitting (12.5.0.3)
SS-Eul	explicit Euler method (11.2.1.5) $\to \boldsymbol{\Psi}^h_{h,g}$, $\boldsymbol{\Psi}^h_{h,f}$ + Strang splitting (12.5.0.4)
SS-EuEI	Strang splitting (12.5.0.4): explicit Euler method (11.2.1.5) $\circ$ exact evolution $\boldsymbol{\Phi}^h_g \circ$ implicit Euler method (11.2.2.2)
LTS-EMP	explicit midpoint method (11.2.3.3) $\to \boldsymbol{\Psi}^h_{h,g}$, $\boldsymbol{\Psi}^h_{h,f}$ + Lie-Trotter splitting (12.5.0.3)
SS-EMP	explicit midpoint method (11.4.0.9) $\to \boldsymbol{\Psi}^h_{h,g}$, $\boldsymbol{\Psi}^h_{h,f}$ + Strang splitting (12.5.0.4)

☞ The order of splitting methods may be (but need not be) limited by the order of the SSMs used for $\boldsymbol{\Phi}^h_f$, $\boldsymbol{\Phi}^h_g$.

§12.5.0.9 (Application of splitting methods) In the following situation the use splitting methods seems advisable:

"Splittable" ODEs

$\dot{\mathbf{y}} = f(\mathbf{y}) + g(\mathbf{y})$ "'difficult" (e.g., stiff $\to$ Section 12.2)	$\dot{\mathbf{y}} = f(\mathbf{y}) \to$ stiff, but with an analytic solution $\dot{\mathbf{y}} = g(\mathbf{y})$ "'easy", amenable to explicit integration.

EXPERIMENT 12.5.0.11 (Splitting off stiff components) Recall Ex. 12.0.0.1 and the IVP studied there:

$$\text{IVP} \quad \dot{y} = \lambda y(1-y) + \alpha \sin(y)\,, \quad \lambda = 100\,, \quad \alpha = 1\,, \quad y(0) = 10^{-4}\,.$$

small perturbation

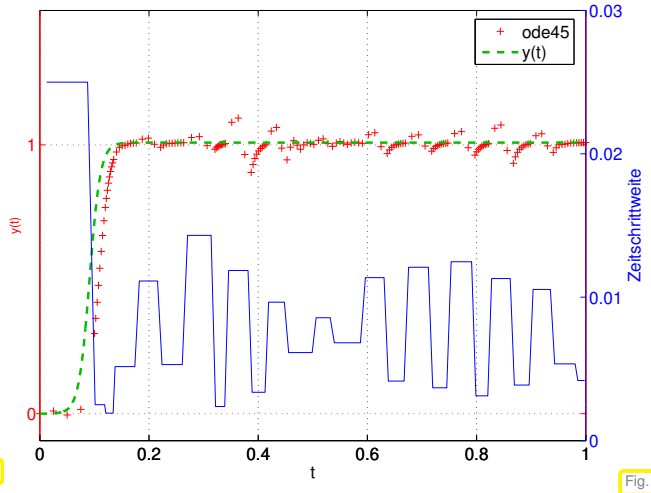


Fig. 482

Solution by **ode45**, see Ex. 12.0.0.1

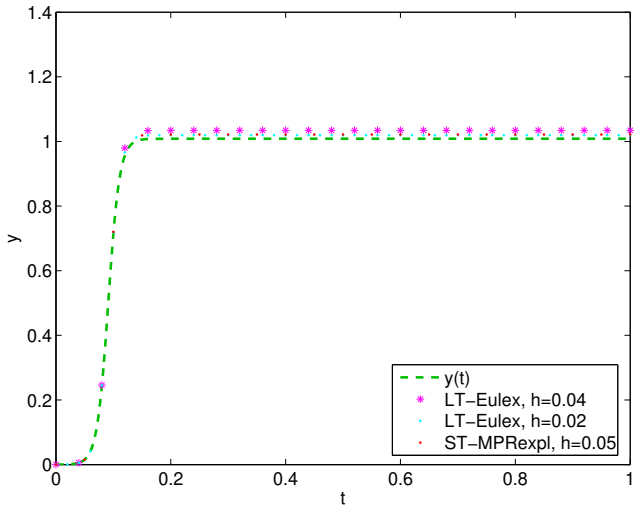


Fig. 483

inexact splitting method: solution (y_k)

Total number of timesteps

ode45:	152
LT-Eulex, $h = 0.04$:	25
LT-Eulex, $h = 0.02$:	50
ST-MPRexpl, $h = 0.05$:	20

Details of the methods:

LT-Eulex: $\dot{y} = \lambda y(1-y) \to$ exact evolution, $\dot{y} = \alpha \sin y \to$ expl. Euler (11.2.1.5) **&** Lie-Trotter splitting (12.5.0.3)
ST-MPRexpl: $\dot{y} = \lambda y(1-y) \to$ exact evolution, $\dot{y} = \alpha \sin y \to$ expl. midpoint rule (11.4.0.9) **&** Strang splitting (12.5.0.4)

We observe that this splitting scheme can cope well with the stiffness of the problem, because the stiff term on the right hand side is integrated exactly. ⌟

EXAMPLE 12.5.0.12 (Splitting linear and decoupled terms) In the numerical treatment of partial differential equation one commonly encounters ODEs of the form

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}) := -\mathbf{A}\mathbf{y} + \begin{bmatrix} g(y_1) \\ \vdots \\ g(y_N) \end{bmatrix}, \quad \mathbf{A} = \mathbf{A}^\top \in \mathbb{R}^{N,N} \text{ positive definite } (\to \text{Def. 1.1.2.6}), \tag{12.5.0.13}$$

with state space $D = \mathbb{R}^N$, where $\lambda_{\min}(\mathbf{A}) \approx 1$, $\lambda_{\max}(\mathbf{A}) \approx N^2$, and the derivative of $g : \mathbb{R} \to \mathbb{R}$ is bounded. Then IVPs for (12.5.0.13) will be stiff, since the Jacobian

$$\mathrm{D}\,\mathbf{f}(\mathbf{y}) = -\mathbf{A} + \begin{bmatrix} g'(y_1) & & \\ & \ddots & \\ & & g'(y_N) \end{bmatrix} \in \mathbb{R}^{N,N}$$

will have eigenvalues "close to zero" and others that are large (in modulus) and negative. Hence, $\mathrm{D}\,\mathbf{f}(\mathbf{y})$ will satisfy the criteria (12.2.0.13) and (12.2.0.14) for any state $\mathbf{y} \in \mathbb{R}^N$.

The natural splitting is

$$\mathbf{f}(\mathbf{y}) = \mathbf{g}(\mathbf{y}) + \mathbf{q}(\mathbf{y}) \quad \text{with} \quad \mathbf{g}(\mathbf{y}) := -\mathbf{A}\mathbf{y} \quad , \quad \mathbf{q}(\mathbf{y}) := \begin{bmatrix} g(y_1) \\ \vdots \\ g(y_N) \end{bmatrix}.$$

- For the linear ODE $\dot{\mathbf{y}} = \mathbf{g}(\mathbf{y})$ we have to use an L-stable ($\to$ Def. 12.3.4.15) single step method, for instance a second-order *implicit* Runge-Kutta method. Its increments can be obtained by solving a *linear system of equations*, whose coefficient matrix will be the same for every step, if uniform timesteps are used.
- The ODE $\dot{\mathbf{y}} = \mathbf{q}(\mathbf{y})$ boils down to decoupled scalar ODEs $\dot{y}_j = g(y_j)$, $j = 1, \ldots, N$. For them we can use an inexpensive *explicit RK-SSM* like the explicit trapezoidal method (11.4.0.8). According to our assumptions on g these ODEs are not haunted by stiffness.

⌟

Review question(s) 12.5.0.14 (Splitting single-step methods)

(Q12.5.0.14.A) The ODE describing the motion of a mathematical pendulum is

$$\begin{bmatrix} \dot{w} \\ \dot{v} \end{bmatrix} = \begin{bmatrix} v \\ -\sin w \end{bmatrix} = \begin{bmatrix} v \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ -\sin w \end{bmatrix}, \tag{12.5.0.15}$$

with state space $\mathbb{R}^2$. Derive formulas for the Strang-splitting single step method applied to the mathematical pendulum equation and using the additive decomposition of the right-hand-side vectorfield suggested in (**??**). Distinguish the initial step, a regular step, and the final step.

Hint. What is the analytic solution of the ODE $\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} \varphi(y) \\ 0 \end{bmatrix}$ for an arbitrary function $\varphi : \mathbb{R} \to \mathbb{R}$?

(Q12.5.0.14.B) Elaborate the extension of the Strang splitting single step method to the ODE

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}) + \mathbf{g}(\mathbf{y}) + \mathbf{r}(\mathbf{y}) \,, \quad \mathbf{f}, \mathbf{g}, \mathbf{r} : D \subset \mathbb{R}^N \to \mathbb{R}^N \,.$$

Develop formulas relying on the exact evolutions $\boldsymbol{\Phi}_f^h$, $\boldsymbol{\Phi}_g^h$, and $\boldsymbol{\Phi}_r^h$ for the ODEs $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$, $\dot{\mathbf{y}} = \mathbf{g}(\mathbf{y})$, and $\dot{\mathbf{y}} = \mathbf{r}(\mathbf{y})$.

(Q12.5.0.14.C) For a symmetric positive definite matrix $\mathbf{A} \in \mathbb{R}^{N,N}$ consider the autonomous ODE on state space $\mathbb{R}^N$:

$$\dot{\mathbf{y}} = -\mathbf{A}\mathbf{y} + \left[\sin(\pi y_j)\right]_{j=1}^N \,, \quad \mathbf{y} = \left[y_j\right]_{j=1}^N \,. \tag{12.5.0.16}$$

We know that the smallest eigenvalue $\lambda_{\min}(\mathbf{A})$ of $\mathbf{A}$ is 1 and the largest $\lambda_{\max}(\mathbf{A})$ can be as big as 10^9.

(i) Based on the Strang splitting Propose an efficient second-order single-step method for (12.5.0.16).

(ii) What is the computational effort for every regular timestep of your method, asymptotically for $N \to \infty$?

△

Supplementary literature. [MQ02] offers a comprehensive review of splitting methods.

Bibliography

[DR08] W. Dahmen and A. Reusken. *Numerik für Ingenieure und Naturwissenschaftler*. Heidelberg: Springer, 2008 (cit. on pp. 820, 823, 854).

[DB02] P. Deuflhard and F. Bornemann. *Scientific Computing with Ordinary Differential Equations*. 2nd ed. Vol. 42. Texts in Applied Mathematics. New York: Springer, 2002 (cit. on pp. 845, 850).

[Han02] M. Hanke-Bourgeois. *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens*. Mathematische Leitfäden. Stuttgart: B.G. Teubner, 2002 (cit. on pp. 825, 834, 851, 853, 858).

[Jax+21] Tim Jax, Andreas Bartel, Matthias Ehrhardt, Michael Günther, and Gerd Steinebach, eds. *Rosenbrock–Wanner-type methods—theory and applications*. Mathematics Online First Collections. Springer, Cham, 2021, pp. vii+120. DOI: 10.1007/978-3-030-76810-2.

[MQ02] R.I. McLachlan and G.R.W. Quispel. "Splitting methods". In: *Acta Nmerica* 11 (2002) (cit. on p. 863).

[NS02] K. Nipp and D. Stoffer. *Lineare Algebra*. 5th ed. Zürich: vdf Hochschulverlag, 2002 (cit. on p. 827).

[QSS00] A. Quarteroni, R. Sacco, and F. Saleri. *Numerical mathematics*. Vol. 37. Texts in Applied Mathematics. New York: Springer, 2000 (cit. on pp. 825, 834, 840, 854).

[Ran15] Joachim Rang. "Improved traditional Rosenbrock-Wanner methods for stiff ODEs and DAEs". In: *J. Comput. Appl. Math.* 286 (2015), pp. 128–144. DOI: 10.1016/j.cam.2015.03.010.

[Ste23] Gerd Steinebach. "Construction of Rosenbrock-Wanner method Rodas5P and numerical benchmarks within the Julia differential equations package". In: *BIT* 63.2 (2023), Paper No. 27, 26. DOI: 10.1007/s10543-023-00967-x.

[Str09] M. Struwe. *Analysis für Informatiker*. Lecture notes, ETH Zürich. 2009 (cit. on pp. 829, 837).

Main Index: Terms and Keywords

LU-decomposition
existence, 156
L^2-inner product, 568
h-convergence, 547
p-convergence, 549
BLAS
axpy, 79
C++
move semantics, 41
EIGEN: triangularView, 87
MATLAB: cumsum, 88
PYTHON: reshape, 68
3-term recursion
for Chebychev polynomials, 501
for Legendre polynomials, 572
3-term recusion
orthogonal polynomials, 521
5-point stencil, 371

a posteriori
adaptive quadrature, 588
a posteriori adaptive, 500
a posteriori error bound, 613
a posteriori termination, 611
a priori
adaptive quadrature, 588
a priori termination, 611
A-inner product, 734
A-orthogonal, 743
A-stable single step method, 849
absolute error, 98
absolute tolerance, 611, 643, 806
adaptive
a posteriori, 500
adaptive multigrid quadrature, 590
adaptive quadrature, 588
a posteriori, 588
a priori, 588
Adding EPS to 1, 100
affine invariance, 645
AGM, 608
Aitken-Neville scheme, 402
algebraic convergence, 484, 487, 788
algebraic dependence, 84
algebraically equivalent, 107
aliasing, 535
alternation theorem, 526
approximation
uniform, 473
approximation error, 473
arrow matrix, 196
Ass: "Axiom" of roundoff analysis, 99
Ass: Analyticity of interpoland, 496
Ass: Global solutions, 773
Ass: Sampling in a period, 456
Ass: Self-adjointness of multiplication operator, 520
Ass: Sharpness of error bounds, 485
Ass: Smoothness of right-hand side vectorfield, 794
asymptotic complexity, 83
sharp bounds, 83
asymptotic rate of linear convergence, 620
audio signal, 306
augmented normal equations, 301
autonomization, 770
autonomous ODE, 762
AXPY operation, 745
axpy operation, 79

back substitution, 137
backward error analysis, 122
backward substitution, 149
Bandbreite
Zeilen-, 200
banded matrix, 199
bandwidth, 199
lower, 199
minimizing, 204
upper, 199
barycentric interpolation formula, 399
basis
cosine, 374
orthonormal, 686

sine, 368
trigonometric, 325
Belousov-Zhabotinsky reaction, 803
bending energy, 437
Bernstein approximant, 477
Bernstein polynomials, 451, 477
Besetzungsmuster, 205
best approximation
uniform, 526
best approximation error, 479, 518
Bezier curve, 448
bicg, 757
BiCGStab, 757
bisection, 623
BLAS, 76
block LU-decomposition, 150
block matrix multiplication, 76
blow-up, 804
blurring operator, 345
Broyden
quasi-Newton method, 664
Broyden-Verfahren
ceonvergence monitor, 665
Butcher scheme, 800, 845

cache miss, 69
cache thrashing, 69
CAD, 445
cancellation, 102, 104
capacitance, 128
capacitor, 128
cardinal
spline, 439
cardinal basis, 390, 394
cardinal basis function, 439
cardinal interpolant, 439, 559
Cauchy product
of power series, 315
Cauchy-Schwarz ineqaulity, 490
causal filter, 306
CCS format, 181
cell
of a mesh, 545
CG
convergence, 750
preconditioned, 752
termination criterion, 745
CG = conjugate gradient method, 741
CG algorithm, 745
chain rule, 646
1D, 50
channel, 306
Characteristic parameters of IEEE floating point numbers, 97
characteristic polynomial, 685
Chebychev expansion, 510
Chebychev nodes, 503
Chebychev polynomials, 501, 748
3-term recursion, 501
Chebychev-interpolation, 499
chemical reaction kinetics, 834
Cholesky decomposition
costs, 211
circuit simulation
transient, 768
circulant matrix, 319
Classical Runge-Kutta method
Butcher scheme, 801
Clenshaw algorithm, 511
cluster analysis, 294
coefficient matrix, 127
coil, 128
collocation, 841
collocation conditions, 842
collocation points, 842
collocation single step methods, 841
column
of a matrix, 55
column major matrix format, 65
column sum norm, 120
column transformation, 75
combinatorial graph Laplacian, 189
complexity
asymptotic, 83
linear, 86
of SVD, 272
composite quadrature formulas, 580
Compressed Column Storage (CCS), 183
compressed row storage, 181
Compressed Row Storage (CRS), 183
computational cost
Gaussian elimination, 139
computational costs
LU-decomposition, 148
QR-decomposition, 256
computational effort, 82, 638
eigenvalue computation, 688
concave
data, 419
function, 420
condition number
of a matrix, 133
spectral, 740

conjugate gradient method, 741
consistency
of iterative methods, 603
fixed point iteration, 615
constant
Lebesgue, 505
constitutive relations, 128, 386
constrained least squares, 299
control points, 445
convergence
algebraic, 484, 487, 788
asymptotic, 635
exponential, 484, 487, 490, 506, 788
global, 604
iterative method, 602
linear, 605
linear in Gauss-Newton method, 678
local, 604
numerical quadrature, 559
quadratic, 608
rate, 605
convergence monitor, 665
of Broyden method, 665
convex
data, 419
function, 420
convex combination, 448
convolution
discrete, 312, 314
discrete periodic, 317
of sequences, 311, 315
Corollary: "Optimality" of CG iterates, 748
Corollary: Best approximant by orthogonal projection, 517
Corollary: Composition of orthogonal transformations, 242
Corollary: Consistent Runge-Kutta single step methods, 800
Corollary: Continuous local Lagrange interpolants, 546
Corollary: Dimension of $\mathcal{P}_{2n}^T$, 457
Corollary: Euclidean matrix norm and eigenvalues, 121
Corollary: Invariance of order under affine transformation, 565
Corollary: Lagrange interpolation as linear mapping, 395
Corollary: ONB representation of best approximant, 518
Corollary: Periodicity of Fourier transforms, 352
Corollary: Piecewise polynomials Lagrange interpolation operator, 546
Corollary: Polynomial stability function of explicit RK-SSM, 825
Corollary: Principal axis transformation, 686
Corollary: Rational stability function of explicit RK-SSM, 849
Corollary: Smoothness of cubic Hermite polynomial interpolant, 423
Corollary: Solvability of implicit Euler recursion, 781
Corollary: Stages limit order of explicit RK-SSM, 826
Corollary: Uniqueness of least squares solutions, 225
Corollary: Uniqueness of QR-factorization, 242
cosine
basis, 374
transform, 374
cosine matrix, 375
cosine transform, 374, 375
costs
Cholesky decomposition, 211
Crout's algorithm, 147
CRS, 181
CRS format
diagonal, 181
cubic complexity, 84
cubic Hermite interpolation, 397, 423, 424
cubic spline interpolation
error estimates, 551
curve, 446
cyclic permutation, 198

damped Newton method, 659
damping factor, 660
data fitting, 464, 670
linear, 465
polynomial, 467
data interpolation, 384
deblurring, 345
definite, 119
dense matrix, 179
derivative
in vector spaces, 646
descent methods, 734
destructor, 42
DFT, 321, 326
two-dimensional, 340
Diagonal dominance, 208
diagonal matrix, 56
diagonalization
for solving linear ODEs, 827

of a matrix, 686
diagonalization of local translation invariant linear operators, 371
diagonally implicit Runge-Kutta method, 846
difference quotient, 104
backward, 779
forward, 779
symmetric, 781
difference scheme, 779
differential, 646
dilation, 456
direct power method, 695
DIRK-SSM, 846
discrete L^2-inner product, 520, 523
discrete convolution, 312, 314
discrete evolution, 783
discrete Fourier transform, 321, 326
two-dimensional, 341
discrete periodic convolution, 317
two-dimensional, 343
discretization
of a differential equation, 783
discretization error, 787
discriminant formula, 103, 108
divided differences, 410
domain specific language (DSL), 58
domain of definition, 599
dot product, 70
double nodes, 397
double precision, 96
DSL: domain specific language, 58

economical singular value decomposition, 268
efficiency, 638
Eigen, 58
arrays, 61
data types, 58
initialisation, 60
sparse matrices, 183
eigen
accessing matrix entries, 60
Eigen: `LDLT()`, 211
Eigen: `LLT()`, 211
eigenspace, 685
eigenvalue, 685
generalized, 687
eigenvalue problem
generalized, 687
eigenvector, 685
generalized, 687
electric circuit, 127, 598
resonant frequencies, 682
element
of a matrix, 55
elementary arithmetic operations, 94, 99
elimination matrix, 145
embedded Runge-Kutta methods, 812
entry
of a matrix, 55
envelope
matrix, 200
Equation
non-linear, 599
equidistant mesh, 545
equidistribution principle
for quadrature error, 589
equivalence
of norms, 119
ergodicity, 694
error
absolute, 98
relative, 98
error estimator
a posteriori, 613
error indicator
for extrapolation, 407
Euler method
explicit, 778
implicit, 780
implicit, stability function, 848
semi implicit, 855
Euler polygon, 778
Euler's formula, 456
Euler's iteration, 633
evolution operator, 773
expansion
asymptotic, 405
explicit Euler method, 778
Butcher scheme, 801
explicit midpoint rule
Butcher scheme, 801
for ODEs, 798
explicit Runge-Kutta method, 799
explicit trapzoidal rule
Butcher scheme, 801
exponential convergence, 506
extended normal equations, 235
extended state space
of an ODE, 769
extrapolation, 404

fast Fourier transform, 360
FFT, 360
fill-in, 195

filter
high pass, 336
low pass, 336
Finding out EPS in C++, 99, 100
finite filter, 306
fixed point, 615
fixed point form, 615
fixed point interation, 614
fixed point iteration
consistency, 615
Newton's method, 654
floating point number, 95
floating point numbers, 94, 95
forward elimination, 136
forward substitution, 149
Fourier
matrix, 325
Fourier coefficient, 356
discrete, 335
Fourier coefficients, 534
Fourier modes, 535
Fourier series, 352, 534
Fourier transform, 352
discrete, 321, 326
fractional order of convergence, 635
frequency domain, 128, 332
frequency filtering, 331
full-rank condition, 225
function
concave, 420
convex, 420
function object, 43
function representation, 387

Gauss collocation single step method, 844
Gauss Quadrature, 564
Gauss-Legendre quadrature formulas, 572
Gauss-Newton method, 675
Gauss-Radau quadrature formulas, 852
Gauss-Seidel preconditioner, 753
Gaussian elimination, 135
block version, 142
by rank-1 modifications, 142
for non-square matrices, 140
general least squares problem, 230
generalization error, 468
Givens rotation, 245, 263
Givens-Rotation, 250
global solution
of an IVP, 772
GMRES, 757
Golub-Welsch algorithm, 573
gradient, 647, 736
Gram-Schmidt
Orthonormalisierung, 726
Gram-Schmidt orthogonalisation, 91, 239
Gram-Schmidt orthogonalization, 518, 726, 744
Gram-Schmidt orthonormalization, 716
graph partitioning, 705
grid, 545
grid cell, 545
grid function, 371
grid interval, 545

Halley's iteration, 633
harmonic mean, 427
hat function, 389
heartbeat model, 766
Hermite integral formula, 497
Hermite interpolation
cubic, 397
Hermitian matrix, 57
Hessenberg matrix, 260
Hessian, 57
Hessian matrix, 647
high pass filter, 336
Hilbert matrix, 107
homogeneous, 119
Hooke's law, 711
Horner scheme, 393
for Bezier polynomials, 451
Householder matrix, 243
Householder reflection, 243

I/O-complexity, 83
identity matrix, 56
IEEE standard 754, 96
ill conditioned, 133
ill-conditioned problem, 123
image segmentation, 697
image space, 219
implicit differentiation, 651
implicit Euler method, 780
implicit function theorem, 780
implicit midpoint method, 781
impulse response, 308
of a filter, 306
in place, 147, 148
in situ, 142, 148
increment equations
linearized, 856
increments
Runge-Kutta, 799, 845
inductance, 128

inductor, 128
inexact splitting methods, 860
inf, 96
infinity, 96
initial guess, 602, 614
initial value problem
    stiff, 837
initial value problem (IVP), 769
inner product
    A-, 734
intermediate value theorem, 623
interpolant
    piecewise linear, 421
interpolation
    barycentric formula, 399
    Chebychev, 499
    complete cubic spline, 434
    cubic Hermite, 424
    Hermite, 397
    Lagrange, 394
    natural cubic spline, 434
    periodic cubic spline, 434
    piecewise linear, 389
    spline cubic, 432
    spline cubic, locality, 439
    spline shape preserving, 440
    trigonometric, 456
interpolation operator, 391
interpolation problem, 386
interpolation scheme, 386
inverse interpolation, 636
inverse iteration, 706
    preconditioned, 707
inverse matrix, 130
invertible matrix, 130
iteration, 601
    Halley's, 633
    Euler's, 633
    quadratical inverse interpolation, 633
iteration function, 602, 614
iterative method, 601
    convergence, 602
IVP, 769

Jacobi preconditioner, 753
Jacobian, 619, 643, 646

kernel, 219
kinetics
    of chemical reaction, 834
Kirchhoff (current) law, 128
knots
    spline, 430
Konvergenz
    Algebraische, Quadratur, 577
Kronecker product, 89
Kronecker symbol, 55
Krylov space, 742
    for Ritz projection, 722

L-stable, 851
Lagrange function, 300
Lagrange interpolation approximation scheme, 483
Lagrange multiplier, 300
Lagrangian, 300
Lagrangian multiplier, 300
lambda function, 33, 43
Landau symbol, 484
Landau-O, 83
Lapack, 140
leading coefficient
    of polynomial, 392
Least squares
    with linear constraint, 299
least squares
    total, 298
least squares problem, 227
least-squares problem
    non-linear, 671
least-squares solution
    non-linear, 671
Lebesgue
    constant, 505
Lebesgue constant, 416, 491
Lemma: $\mathbf{r}_k \perp U_k$, 741
Lemma: Absolute conditioning of polynomial interpolation, 416
Lemma: Affine pullbacks preserve polynomials, 480
Lemma: Bases for Krylov spaces in CG, 744
Lemma: Basis property of Bernstein polynomials, 451
Lemma: Cholesky decomposition, 210
Lemma: Criterion for local Liptschitz continuity, 771
Lemma: Cubic convergence of modified Newton methods, 632
Lemma: Decay of Fourier coefficients, 537
Lemma: Diagonal dominance and definiteness, 209
Lemma: Diagonalization of circulant matrices, 326

Lemma: Equivalence of Gaussian elimination and LU-factorization, 159
Lemma: Error representation for polynomial Lagrange interpolation, 489
Lemma: Exact quadrature by equidistant trapezoidal rule, 586
Lemma: Existence of LU-decomposition, 146
Lemma: Existence of LU-factorization with pivoting, 156
Lemma: Formula for Euclidean norm of a Hermitian matrix, 121
Lemma: Fourier coefficients of derivatives, 536
Lemma: Gerschgorin circle theorem, 686
Lemma: Group of regular diagonal/triangular matrices, 74
Lemma: Higher order local convergence of fixed point iterations, 621
Lemma: Interpolation error estimates for exponentially decaying Fourier coefficients, 542
Lemma: Kernel and range of (Hermitian) transposed matrices, 224
Lemma: LU-factorization of diagonally dominant matrices, 208
Lemma: Ncut and Rayleigh quotient (→ [SM00, Sect. 2]), 700
Lemma: Necessary conditions for s.p.d., 57
Lemma: Perturbation lemma, 132
Lemma: Positivity of Gauss-Legendre quadrature weights, 575
Lemma: Principal branch of the square root, 543
Lemma: Properties of cosine matrix, 375
Lemma: Properties of Fourier matrices, 325
Lemma: Properties of the sine matrix, 368
Lemma: Quadrature error estimates for C^r-integrands, 576
Lemma: Quadrature formulas from linear interpolation schemes, 559
Lemma: Residue formula for quotients, 496
Lemma: S.p.d. LSE and quadratic minimization problem, 734
Lemma: Sherman-Morrison-Woodbury formula, 175
Lemma: Similarity and spectrum → [Gut09, Thm. 9.7], [DR08, Lemma 7.6], [NS02, Thm. 7.2], 686
Lemma: Smoothness of solutions of ODEs, 763
Lemma: Space of solutions of linear ODEs, 763
Lemma: Stability function as approximation of $\exp$ for small arguments, 826
Lemma: Sufficient condition for linear convergence of fixed point iteration, 620
Lemma: Sufficient condition for local linear convergence of fixed point iteration, 619
Lemma: SVD and Euclidean matrix norm, 278
Lemma: SVD and rank of a matrix → [NS02, Cor. 9.7], 269
Lemma: Taylor expansion of inverse distance function, 712
Lemma: Theory of Arnoldi process, 727
Lemma: Transformation of norms under affine pullbacks, 481
Lemma: Tridiagonal Ritz projection from CG residuals, 724
Lemma: Unique solvability of linear least squares fitting problem, 467
Lemma: Uniqueness of orthonormal polynomials, 521
Lemma: Zeros of Legendre polynomials, 572
Levinson algorithm, 380
Lie-Trotter splitting, 859
limit cycle, 835
limiter, 426
line search, 735
linear complexity, 84, 86
linear correlation, 290
linear data fitting, 465
linear electric circuit, 127
linear filter, 306
linear ODE, 763
linear operator, 391
  diagonalization, 371
linear ordinary differential equation, 684
linear regression, 84, 215
linear system of equations
  multiple right hand sides, 141
linear system of equations, 127
Lloyd-Max algorithm, 294
local Lagrange interpolation, 546
local linearization, 643
locality
  of interpolation, 438
logistic differential equation, 764
Lotka-Volterra ODE, 765
low pass filter, 336
lower triangular matrix, 56
LU-decomposition
  blocked, 150
  computational costs, 148
  envelope aware, 202
  existence, 146
  in place, 148

LU-factorization
envelope aware, 202
of sparse matrices, 194
with pivoting, 154

machine number, 95
exponent, 95
machine numbers, 94, 96
distribution, 96
extremal, 95
machine precision, 99
mantissa, 95
Markov chain, 377, 690
stationary distribution, 690
mass matrix, 712
Matrix
adjoint, 56
Hermitian, 686
Hermitian transposed, 56
normal, 686
skew-Hermitian, 686
transposed, 56
unitary, 686
matrix
banded, 199, 200
condition number, 133
dense, 179
diagonal, 56
envelope, 200
Fourier, 325
Hermitian, 57
Hessian, 647
lower triangular, 56
normalized, 56
orthogonal, 238
positive definite, 57
positive semi-definite, 57
rank, 130
sine, 368
sparse, 179
storage formats, 65
structurally symmetric, 204
symmetric, 57
tridiagonal, 199
unitary, 238
upper triangular, 56
matrix block, 55
matrix compression, 280
matrix exponential, 827
matrix factorization, 143, 144
matrix norm, 120
column sums, 120
row sums, 120
matrix storage
envelope oriented, 204
maximum likelihood, 220
member function, 31
mesh, 545
equidistant, 545
in time, 784
temporal, 777
mesh width, 545
mesh adaptation, 590
mesh refinement, 590
Method
Quasi-Newton, 663
method, 31
midpoint method
implicit, stability function, 848
midpoint rule, 562, 798
Milne rule, 563
min-max theorem, 701
minimal residual methods, 756
model function, 625
model reduction, 474
Modellfunktionsverfahren, 625
modification techniques, 259
modified Newton method, 632
monomial, 392
monomial representation
of a polynomial, 392
Moore-Penrose pseudoinverse, 230
move semantics, 41
multi-point methods, 625, 633
multiplicity
geometric, 685
of a spline knot, 453
of an interpolation node, 397

NaN, 96
Ncut, 698
nested
subspaces, 741
nested spaces, 477
Newton
basis, 409
damping, 660
damping factor, 660
monotonicity test, 660, 663
simplified method, 653
Newton correction, 643
simplified, 658
Newton iteration, 642
numerical Differentiation, 653

termination criterion, 657
Newton method
1D, 625
damped, 659
local quadratic convergence, 654
modified, 632
Newton's law of motion, 712
nodal analysis, 127, 598
transient, 768
nodal polynomial, 500
nodal potentials, 128
node
double, 397
for interpolation, 394
in electric circuit, 127
multiple, 396
multiplicity, 397
of a mesh, 545
quadrature, 557
nodes, 394
Chebychev, 503
Chebychev nodes, 505
for interpolation, 394
non-linear least-squares problem, 671
non-linear least-squares solution, 671
non-normalized numbers, 96
norm, 119
L^1, 415
L^2, 415
∞-, 119
1-, 119
energy-, 734
Euclidean, 119
Frobenius norm, 281
of matrix, 120
Sobolev semi-, 491
supremum, 415
normal equations, 222, 517
augmented, 301
extended, 235
with constraint, 301
normalization, 695
normalized lower triangular matrix, 145
normalized triangular matrix, 56
not a number, 96
nullspace, 219
Nullstellenbestimmung
Modellfunktionsverfahren, 625
numerical Differentiation
Newton iteration, 653
numerical rank, 276
Numerical differentiation
roundoff, 106
numerical differentiation, 104
numerical quadrature, 555
numerical rank, 272

ODE, 769
autonomous, 762
linear, 763
scalar, 763, 777
ODE, right-hand-side function, 762
Ohmic resistor, 128
one-point methods, 625
one-step error, 791
order
of a discrete evolution, 793
of an ODE, 770
of quadrature formula, 565
order of convergence
fractional, 635
ordinary differential equation
linear, 684
ordinary differential equation (ODE), 769
oregonator, 803
orthogonal complement, 270
orthogonal matrix, 238
orthogonal polynomials, 571
orthogonal projection, 517
orthonormal basis, 518, 686
overfitting, 468
overflow, 96, 101
overloading
of functions, 30
of operators, 31

page rank, 690
stochastic simulation, 690
parameter estimation, 215
parameterization, 446
PARDISO, 193
partial pivoting, 153, 155
pattern
of a matrix, 72
PCA, 286
PCG, 752
Peano
Theorem of, 771
penalization, 702
penalty parameter, 703
periodic
function, 456
periodic sequence, 316

periodic signal, 316
permutation, 156
permutation matrix, 156, 264
perturbation lemma, 132
Petrov-Galerkin condition, 757
phase space
of an ODE, 769
phenomenological model, 766
Picard-Lindelöf
Theorem of, 771
PINVIT, 707
pivot, 136–138
pivot row, 136, 138
pivoting, 152
partial, 153
plane wave, 340
point spread function, 345
polar decomposition, 270
polynomial
characteristic, 685
generalized Lagrange, 397
Lagrange, 394
polynomial curve, 446
polynomial fitting, 467
polynomial interpolation
existence and uniqueness, 395
generalized, 396
polynomial space, 392
positive definite
criteria, 57
matrix, 57
potentials
nodal, 128
power series, 493
power spectrum
of a signal, 336
preconditioned CG method, 752
preconditioned inverse iteration, 707
Preconditioner, 751
preconditioner, 751
preconditioning, 750
predator-prey model, 765
principal, 763
principal axis, 294
principal axis transformation, 738
principal component, 291
principal component analysis (PCA), 286
principal minor, 149
principal orthogonal decompostion (POD), 292
problem
ill conditioned, 133
ill-conditioned, 123
sensitivity, 131
well conditioned, 133
procedural form, 555
product rule, 647
1D, 50
propagated error, 791
pullback, 480, 558
Punkt
stationär, 766
pwer method
direct, 695
Python, 64

QR algorithm, 687
QR-algorithm with shift, 688
QR-decomposition, 93, 240
computational costs, 256
QR-factorization, QR-decomposition, 244
quadratic complexity, 84
quadratic convergence, 621
quadratic eigenvalue problem, 682
quadratic functional, 734
quadratic inverse interpolation, 637
quadratical inverse interpolation, 633
quadrature
adaptive, 588
polynomial formulas, 561
quadrature formula
order, 565
quadrature node, 557
quadrature numerical, 555
quadrature point, 557
quadrature weight, 557
quasi-linear system, 649
Quasi-Newton method, 663, 664

Radau RK-method
order 3, 852
order 5, 852
Rader's algorithm, 364
radiative heat transfer, 317
radius of convergence
of a power series, 494
range, 219
rank
column rank, 130
computation, 272
numerical, 272, 276
of a matrix, 130
row rank, 130
rank-1 modification, 142, 259

rank-1-matrix, 86
rank-1-modification, 174, 664
rate
of algebraic convergence, 484, 487, 788
of convergence, 605
Rayleigh quotient, 696, 701
Rayleigh quotient iteration, 707
region of absolute stability, 832
regular matrix, 130
relative error, 98
relative tolerance, 611, 643, 806
rem:Fspec, 326
residual quantity, 708
Ricati differential equation, 777
Riccati differential equation, 778
Riemann sum, 355
right hand side
of an ODE, 769
right hand side vector, 127
rigid body mode, 713
Ritz projection, 718, 722
Ritz value, 719
Ritz vector, 719
root of unity, 324
roots of unity, 585
rounding, 98
rounding up, 98
roundoff
for numerical differentiation, 106
row
of a matrix, 55
row major matrix format, 65
ROW methods, 857
row sum norm, 120
row transformation, 75, 136, 144
Runge's example, 415
Runge-Kutta
increments, 799, 845
Runge-Kutta method, 799, 845
L-stable, 851
Runge-Kutta methods
embedded, 812
semi-implicit, 855
stability function, 825, 848

saddle point problem, 300
matrix form, 301
scalar ODE, 777
scaling
of a matrix, 74
scaling invariance, 101
scheme
Horner, 393
Schur
Komplement, 150
Schur complement, 150, 171, 174
scientific notation, 95
secant condition, 664
secant method, 634, 637, 663
segmentation
of an image, 697
semi-implicit Euler method, 855
seminorm, 491
sensitive dependence, 123
sensitivity
of a problem, 131
of polynomial interpolation, 414
shape
preservation, 422
preserving spline interpolation, 440
Sherman-Morrison-Woodbury formula, 175
shifted inverse iteration, 706
signal
periodic, 316
time-discrete, 305
similarity
of matrices, 686
similarity function
for image segmentation, 698
similarity transformations, 686
similary transformation
unitary, 687
Simpson rule, 563
sine
basis, 368
matrix, 368
transform, 369
Sine transform, 368
single precicion, 96
single step method
A-stability, 849
single-step method, 784
singular value decomposition, 266, 267
singular vectors, 267
SIR model, 767
slopes
for cubic Hermite interpolation, 423
Sparse matrix, 179
sparse matrix, 179
COO format, 180
LU-factorization, 194
triplet format, 180
sparse matrix storage formats, 180

spectral condition number, 740
spectral partitioning, 705
spectral radius, 685
spectrum, 685
    of a matrix, 737
spline, 430
    cardinal, 439
    complete cubic, 434
    cubic, 432
    cubic, locality, 439
    knots, 430
    natural cubic, 434
    periodic cubic, 434
    physical, 437
    shape preserving interpolation, 440
splitting
    Lie-Trotter, 859
    Strang, 859
splitting methods, 858
    inexact, 860
spy, 73
stability
    region of, 832
stability function
    of explicit Runge-Kutta methods, 825
    of Runge-Kutta methods, 848
stable
    algorithm, 122
    numerically, 122
stages, 846
state space
    of an ODE, 769
stationary distribution, 690
steepest descent, 735
Stiff IVP, 837
stiffness matrix, 712
stochastic matrix, 691
stochastic simulation of page rank, 690
stopping rule, 610
Strang splitting, 859
Strassen's algorithm, 85
structurally symmetric matrix, 204
sub-matrix, 55
sub-multiplicative, 120
subspace correction, 741
subspace iteration
    for direct power method, 720
subspaces
    nested, 741
SuperLU, 193
surrogate function, 474
SVD, 266, 267
symmetric matrix, 57
symmetry
    structural, 204
system matrix, 127
system of equations
    linear, 127

tangent field, 777
Taylor expansion, 51, 621
Taylor polynomial, 476
Taylor series, 495
Taylor's formula, 476
template, 32
tensor product, 70
tent function, 389
Teopltiz matrices, 376
termination criterion, 610
    ideal, 611
    Newton iteration, 657
    residual based, 611
Theorem: $\rightarrow$ [Han02, Thm. 25.4], 707
Theorem: L^∞ polynomial best approximation estimate, 479
Theorem: (Absolute) stability of explicit RK-SSM for linear systems of ODEs, 831
Theorem: 2D convolution theorem, 344
Theorem: 3-term recursion for Chebychev polynomials, 501
Theorem: 3-term recursion for orthogonal polynomials, 522
Theorem: Courant-Fischer min-max theorem $\rightarrow$ [GV89, Thm. 8.1.2], 701
Theorem: **Variation -diminishing property** of Bezier curves 450
Theorem: Banach's fixed point theorem, 618
Theorem: Basis property of B-splines, 454
Theorem: Bernstein basis representation of Bezier curves, 451
Theorem: Best low rank approximation, 281
Theorem: Bezier curves stay in convex hull, 450
Theorem: Bound for spectral radius, 685
Theorem: Cauchy integral theorem, 540
Theorem: Chebychev alternation theorem, 526
Theorem: Commuting matrices have the same eigenvectors, 323
Theorem: Composition and products of analytic functions, 498
Theorem: Conditioning of LSEs, 132
Theorem: Consistency and convergence, 603

Theorem: Convergence of approximation by cubic Hermite interpolation, 550
Theorem: Convergence of CG method, 748
Theorem: Convergence of direct power method → [DR08, Thm. 25.1], 697
Theorem: Convergence of gradient method/steepest descent, 740
Theorem: Convolution of sequences commutes, 312
Theorem: Convolution theorem, 330, 357
Theorem: Cost for solving triangular systems, 166
Theorem: Cost of Gaussian elimination, 166
Theorem: Criteria for invertibility of matrix, 130
Theorem: Dimension of space of polynomials, 392
Theorem: Dimension of spline space, 431
Theorem: Divergent polynomial interpolants, 486
Theorem: Elementary properties of B-splines, 454
Theorem: Envelope and fill-in, 201
Theorem: Equivalence of all norms on finite dimensional vector spaces, 606
Theorem: Existence & uniqueness of generalized Lagrange interpolation polynomials, 397
Theorem: Existence & uniqueness of Lagrange interpolation polynomial, 395
Theorem: Existence of n-point quadrature formulas of order $2n$, 570
Theorem: Existence of least squares solutions, 220
Theorem: Exponential convergence of trigonometric interpolation for analytic interpolands, 542
Theorem: Exponential decay of Fourier coefficients of analytic functions, 540
Theorem: Finite-smoothness L^2-error estimate for trigonometric interpolation, 537
Theorem: Formula for generalized least-squares solution, 229
Theorem: Gaussian elimination for s.p.d. matrices, 209
Theorem: Gram-Schmidt orthonormalization, 519
Theorem: Implicit function theorem, 780
Theorem: Isometry property of the Fourier transform, 358
Theorem: Kernel and range of $\mathbf{A}^\top\mathbf{A}$, 224
Theorem: Least squares solution of data fitting problem, 466
Theorem: Local quadratic convergence of Newton's method, 656
Theorem: Local shape preservation by piecewise linear interpolation, 422
Theorem: Maximal order of n-point quadrature rule, 567
Theorem: Mean square norm best approximation through normal equations, 517
Theorem: Mean square (semi-)norm/Inner product (semi-)norm, 516
Theorem: Minimax property of the Chebychev polynomials, 502
Theorem: Monotonicity preservation of limited cubic Hermite interpolation, 428
Theorem: Obtaining least squares solutions by solving normal equations, 222
Theorem: Optimality of natural cubic spline interpolant, 437
Theorem: Order of collocation single step method, 845
Theorem: Order of simple splitting methods, 860
Theorem: Positivity of Clenshaw-Curtis weights, 564
Theorem: Preservation of Euclidean norm, 238
Theorem: Property of linear, monotonicity preserving interpolation into C^1, 428
Theorem: Pseudoinverse and SVD, 276
Theorem: QR-decomposition, 241
Theorem: QR-decomposition "preserves bandwidth", 250
Theorem: Quadrature error estimate for quadrature rules with positive weights, 576
Theorem: Rayleigh quotient, 701
Theorem: Region of stability of Gauss collocation single step methods, 850
Theorem: Representation of interpolation error, 488
Theorem: Residue theorem, 495
Theorem: Schur's lemma, 686
Theorem: Sensitivity of full-rank linear least squares problem, 231
Theorem: Series of analytic functions, 544
Theorem: Singular value decomposition (SVD), 266
Theorem: Solution of POD problem, 293
Theorem: Span property of G.S. vectors, 239
Theorem: Stability function of general Runge-Kutta methods, 848
Theorem: Stability function of some explicit

Runge-Kutta methods, 825
Theorem: Stability of Gaussian elimination with partial pivoting, 160
Theorem: Stability of Householder QR [Hig02, Thm. 19.4], 252
Theorem: Sufficient order conditions for quadrature rules, 565
Theorem: Taylor's formula, 621
Theorem: Theorem of Peano & Picard-Lindelöf [Ama83, Satz II(7.6)], [Str09, Satz 6.5.1], [DR08, Thm. 11.10], [Han02, Thm. 73.1], 771
Theorem: Uniform approximation by polynomials, 477
time domain, 332
time-invariant filter, 306
timestep (size), 779
timestep constraint, 826
timestepping, 778
Toeplitz solvers
fast algorithms, 382
tolerance, 611
absolute, 806
absoute, 611, 643
for adaptive timestepping for ODEs, 805
for termination, 611
realtive, 806
relative, 611, 643
total least squares, 298
trajectory, 766
transform
cosine, 374
fast Fourier, 360
sine, 369
transformation matrix, 76
trapezoidal rule, 562, 585, 798
for ODEs, 798
trend, 286
trial space
for collocation, 842
triangle inequality, 119
triangular linear systems, 171
triangulation, 185
tridiagonal matrix, 199
trigonometric basis, 325
trigonometric interpolation, 456, 531
trigonometric polynomial, 357
trigonometric polynomials, 456, 531
trigonometric transformations, 368
tripled format, 180
truss structure
vibrations, 710
trust region method, 679

UMFPACK, 193
underflow, 96, 101
uniform approximation, 473
uniform best approximation, 526
Uniform convergence
of Fourier series, 353
unit vector, 55
unitary matrix, 238
unitary similary transformation, 687
upper Hessenberg matrix, 727
upper triangular matrix, 56, 136, 145

Vandermonde matrix, 396
variation-diminishing property, 450
variational calculus, 437
vector field, 769
vectorization
of a matrix, 67
Vieta's formula, 108

WAV file format, 306
Weddle rule, 563
weight
quadrature, 557
weight function, 520
weighted L^2-inner product, 520
well conditioned, 133

Young's modulus, 711

Zerlegung
LU, 149
zero padding, 315, 320, 379, 462

List of Symbols

$(\mathbf{A})_{i,j} \hat{=}$ reference to entry a_{ij} of matrix $\mathbf{A}$, 55
$(\mathbf{A})_{k:l,r:s} \hat{=}$ reference to submatrix of $\mathbf{A}$ spanning rows $k,\ldots,l$ and columns $r,\ldots,s$, 55
$(\mathbf{x})_i \hat{=}$ i-th component of vector $\mathbf{x}$, 54
$(x_k) *_n (y_k) \hat{=}$ discrete periodic convolution, 317
$C^0(I) \hat{=}$ space of continuous functions $I \to \mathbb{R}$, 415
$C^1([a,b]) \hat{=}$ space of continuously differentiable functions $[a,b] \mapsto \mathbb{R}$, 423
$J(t_0,\mathbf{y}_0) \hat{=}$ maximal domain of definition of a solution of an IVP, 771
$O \hat{=}$ zero matrix, 56
$O(\cdot) \hat{=}$ Landau symbol, 83
$V^\perp \hat{=}$ orthogonal complement of a subspace, 225
$\mathcal{E} \hat{=}$ expected value of a random variable, 378
$\mathcal{P}_n^T \hat{=}$ space of trigonometric polynomials of degree n, 456
$\mathcal{R}_k(m,n) \hat{=}$ set of rank-k matrices, 281
$\mathrm{DFT}_{m,n} \hat{=}$ 2D discrete Fourier transform, 341
$\mathrm{DFT}_n \hat{=}$ discrete Fourier transform of length n, 326
$\mathrm{D}\,\Phi \hat{=}$ Jacobian of $\Phi : D \mapsto \mathbb{R}^n$ at $\mathbf{x} \in D$, 619
$\mathrm{D}_{\mathbf{y}}\,\mathbf{f} \hat{=}$ Derivative of $\mathbf{f}$ w.r.t. $\mathbf{y}$ (Jacobian), 771
$\mathtt{EPS} \hat{=}$ machine precision, 99
$\mathbf{EigA}\lambda \hat{=}$ eigenspace of $\mathbf{A}$ for eigenvalue λ, 685
$\mathsf{I}_{\mathcal{T}} \hat{=}$ Lagrange polynomial interpolation operator based on node set $\mathcal{T}$, 398
$\mathcal{R}\mathbf{A} \hat{=}$ range/column space of matrix $\mathbf{A}$, 269
$\mathcal{N}(\mathbf{A}) \hat{=}$ kernel/nullspace of a matrix, 130, 219
$\mathcal{N}\mathbf{A} \hat{=}$ nullspace of matrix $\mathbf{A}$, 269
$\mathcal{K}_l(\mathbf{A},\mathbf{z}) \hat{=}$ Krylov subspace, 742
$\mathsf{L}_{\mathcal{T}} \hat{=}$ Lagrangian (interpolation polynomial) approximation scheme on node set $\mathcal{T}$, 483
$\|\mathbf{Ax}-\mathbf{b}\|_2 \to \min \hat{=}$ minimize $\|\mathbf{Ax}-\mathbf{b}\|_2$, 227
$\|\mathbf{A}\|_F^2$, 281
$\|\mathbf{x}\|_A \hat{=}$ energy norm induced by s.p.d. matrix $\mathbf{A}$, 734
$\|\cdot\| \hat{=}$ Euclidean norm of a vector $\in \mathbb{K}^n$, 91
$\|\cdot\| \hat{=}$ norm on vector space, 119
$\mathcal{P}_k$, 392
$\boldsymbol{\Psi}^h\mathbf{y} \hat{=}$ discrete evolution for autonomous ODE, 783
$\mathcal{R}(\mathbf{A}) \hat{=}$ image/range space of a matrix, 130, 219
$(\cdot,\cdot)_V \hat{=}$ inner product on vector space V, 515
$\mathcal{S}_{d,\mathcal{M}}$, 430
$\mathbf{A}^\dagger \hat{=}$ Moore-Penrose pseudoinverse of $\mathbf{A}$, 230
$\mathbf{A}^\top \hat{=}$ transposed matrix, 56
$\mathbf{I} \hat{=}$ identity matrix, 56
$\mathbf{h} * \mathbf{x} \hat{=}$ discrete convolution of two vectors, 314
$\mathbf{x} *_n \mathbf{y} \hat{=}$ discrete periodic convolution of vectors, 317
$\bar{z} \hat{=}$ complex conjugation, 56
$\mathbb{C}^- := \{z \in \mathbb{C}\colon \mathrm{Re}\, z < 0\}$, 849
$\mathbb{K} \hat{=}$ generic field of numbers, either $\mathbb{R}$ or $\mathbb{C}$, 54
$\mathbb{K}_*^{n,n} \hat{=}$ set of invertible $n \times n$ matrices, 131
$\mathbb{M} \hat{=}$ set of machine numbers, 94
$\delta_{ij} \hat{=}$ Kronecker symbol, 394
$\delta_{ij} \hat{=}$ Kronecker symbol, called "Kronecker delta" by Wikipedia, 55
$\delta_{ij} \hat{=}$ Kronecker symbol, called "Kronecker delta" by Wikipedia, 308
$\ell^\infty(\mathbb{Z}) \hat{=}$ space of bounded bi-infinite sequences, 307
$\imath \hat{=}$ imaginary unit, "$\imath := \sqrt{-1}$", 128
$\kappa(\mathbf{A}) \hat{=}$ spectral condition number, 740
$\lambda_{\mathcal{T}} \hat{=}$ Lebesgue constant for Lagrange interpolation on node set $\mathcal{T}$, 416
$\lambda_{\max} \hat{=}$ largest eigenvalue (in modulus), 740
$\lambda_{\min} \hat{=}$ smallest eigenvalue (in modulus), 740
$(x_k) * (h_k) \hat{=}$ convolution of two sequences, 311
$\mathbf{1} = [1,\ldots,]^\top$, 825, 833
$\mathrm{Ncut}(\mathcal{X}) \hat{=}$ normalized cut of subset of weighted graph, 698
$\mathrm{argmin} \hat{=}$ (global) minimizer of a functional, 735
$\mathrm{cond}(\mathbf{A})$, 133
$\mathrm{cut}(\mathcal{X}) \hat{=}$ cut of subset of weighted graph, 698
$\mathrm{diag}(d_1,\ldots,d_n) \hat{=}$ square diagonal matrix, 56

$\mathrm{dist}_{\|\cdot\|}(x, V)$ ≙ distance of an element of a normed vector spcace from set V, 479
$\mathrm{env}(\mathbf{A})$, 200
$\mathrm{lsq}(\mathbf{A}, \mathbf{b})$ ≙ set of least squares solutions of $\mathbf{Ax} = \mathbf{b}$, 219
nnz, 179
$\mathrm{rank}(\mathbf{A})$ ≙ rank of matrix $\mathbf{A}$, 130
sgn ≙ sign function, $\mathrm{sgn}(0) := 0$, 246
sgn ≙ sign function, 426
$\mathrm{vec}(\mathbf{A})$ ≙ vectorization of a matrix, 67
$\mathrm{weight}(\mathcal{X})$ ≙ connectivity of subset of weighted graph, 698
$\overline{}$ ≙ complex conjugation, 515
$\overline{m}(\mathbf{A})$, 199
$\rho(\mathbf{A})$ ≙ spectral radius of $\mathbf{A} \in \mathbb{K}^{n,n}$, 685
$\rho_{\mathbf{A}}(\mathbf{u})$ ≙ Rayleigh quotient, 696
$\sigma(\mathbf{A})$ ≙ spectrum of matrix $\mathbf{A}$, 685
$\sigma(\mathbf{M})$ $hat{=}$ spectrum of matrix $\mathbf{M}$, 737
$\widetilde{\star}$, 99
$\mathbb{S}^1$ ≙ unit circle in the complex plane, 457
$\underline{m}(\mathbf{A})$, 199
$\widehat{f}_j$ ≙ j-th Fourier coefficient of periodic function f, 356
$f^{(k)}$ ≙ k-th derivative of function $f : I \subset \mathbb{R} \to \mathbb{K}$, 476
$f^{(k)}$ ≙ k derivative of f, 115
$m(\mathbf{A})$, 199
$y[t_i, \ldots, t_{i+k}]$ ≙ divided difference, 410
$\mathbf{f}$ ≙ right hand side of an ODE, 769
$\| \mathbf{x} \|_1$, 119
$\| \mathbf{x} \|_2$, 119
$\| \mathbf{x} \|_\infty$, 119
$\dot{}$ ≙ Derivative w.r.t. time t, 762

TOL tolerance, 805

Examples and Remarks

LU-decomposition of sparse matrices, 194
L^2-error estimates for polynomial interpolation, 490
h-adaptive numerical quadrature, 593
p-convergence of piecewise polynomial interpolation, 549
$L^2()0,1[]$: Natural setting for trigonometric interpolation, 534
'**auto**' in EIGEN codes, 62
(Nearly) singular LSE in shifted inverse iteration, 706
(Relative) point locations from distances, 217
Ex. 2.3.3.6 cnt'd, 157
Ex. 5.4.4.5 cnt'd, 444
General non-linear systems of equations, 599
Complex step differentiation [LM67], 116
$L^2([-1,1])$-orthogonal polynomials $\rightarrow$ [Han02, Bsp. 33.2], 522
BLAS calling conventions, 80
EIGEN-based code: debug mode and release mode, 63
EIGEN in use, 63
'Partial LU-decompositions" of principal minors, 149
"Annihilating" orthogonal transformations in 2D, 243
"Behind the scenes" of **MyVector** arithmetic, 46
"Butcher barriers" for explicit RK-SSM, 802
"Explosion equation": finite-time blow-up, 772
"Failure" of adaptive timestepping, 809
"Fast" matrix multiplication, 85
"Full-rank condition", 672
"Squeezed" DFT of a periodically truncated signal, 350
"**auto**" considered harmful, 37
"**Overfitting**", 468
$\mathbf{B} = \mathbf{B}^H$ s.p.d. mit Cholesky-Zerlegung, 687
L-stable implicit Runge-Kutta methods, 852
2-norm from eigenvalues, 740
3-Term recursion for Legendre polynomials, 572

Analytic solution of homogeneous linear ordinary differential equations , 684
Convergence of PINVIT , 709
Convergence of subspace variant of direct power method , 720
Direct power method , 696
Eigenvalue computation with Arnoldi process , 729
Impact of roundoff on Lanczos process , 725
Lagrange polynomials for uniformly spaced nodes, 394
Lanczos process for eigenvalue computation , 725
Page rank algorithm , 690
Power iteration with Ritz projection , 720
qr based orthogonalization , 717
Rayleigh quotient iteration , 707
Resonances of linear electrical circuits , 682
Ritz projections onto Krylov space , 722
Runtimes of `eig` , 689
Stabilty of Arnoldi process , 728
Subspace power iteration with orthogonal projection , 715
Vibrations of a truss structure , 710

A broader view of cancellation, 117
A data type designed for interpolation problems, 388
A function that is not locally Lipschitz continuous, 771
A parameterized curve, 446
A posteriori error bound for linearly convergent iteration, 613
A posteriori termination criterion for plain CG, 745
A priori and a posteriori choice of optimal interpolation nodes, 500
A special quasi-linear system of equations, 650
Access to LU-factors in EIGEN, 170
Accessing matrix data as a vector, 65
Adaptive explicit RK-SSM for scalar linear decay ODE, 820
Adaptive quadrature in PYTHON, 594

Adaptive timestepping for mechanical problem, 815
Adding EPS to 1, 100
Affine invariance of Newton method, 645
Algorithm for cluster analysis, 294
Analytic functions everywhere, 494
Angles in a triangulation, 216
Application of modified Newton methods, 633
Approximate computaton of Fourier coefficients, 586
Approximation by discrete polynomial fitting, 524
Arnoldi process Ritz projection, 727
Asymptotic behavior of Lagrange interpolation error, 483
Asymptotic complexity of Householder QR-factorization, 253
Asymptotic methods for the computation of Gauss-Legendre quadrature rules, 574
Auxiliary construction for shape preserving quadratic spline interpolation, 442
Average-based pecewise cubic Hermite interpolation, 424

Bad behavior of global polynomial interpolants, 420
Banach's fixed point theorem, 618
Bernstein approximants, 478
Block Gaussian elimination, 142
Block LU-factorization, 150
Blow-up, 804
Blow-up of explicit Euler method, 821
Blow-up solutions of vibration equations, 710
Bound for *asymptotic* rate of linear convergence, 620
Breakdown of associativity, 99
Broyden method for a large non-linear system, 669
Broyden's quasi-Newton method: convergence, 665
Butcher scheme for some explicit RK-SSM, 801

Calling BLAS routines from C/C++, 80
Cancellation during the computation of relative errors, 107
Cancellation in decimal system, 104
Cancellation in Gram-Schmidt orthogonalisation, 107
Cancellation when evaluating difference quotients, 104
Cancellation: roundoff error analysis, 107
Cardinal shape preserving quadratic spline, 444
CG convergence and spectrum, 750
Characteristics of stiff IVPs, 839
Chebychev interpolation errors, 506
Chebychev interpolation of analytic function, 509
Chebychev interpolation of analytic functions, 507
Chebychev nodes on arbitrary interval, 503
Chebychev representation of built-in functions, 514
Chebychev vs equidistant nodes, 504
Choice of norm, 808
Class PolyEval, 412
Classification from measured data, 286
Clenshaw-Curtis quadrature rules, 564
Commonly used embedded explicit Runge-Kutta methods, 813
Communicating special properties of system matrices in EIGEN, 166
Commutativity of discrete convolution, 314
Composite quadrature and piecewise polynomial interpolation, 582
Composite quadrature rules vs. global quadrature rules, 584
Computation of nullspace and image space of matrices, 273
Computational effort for eigenvalue computations, 688
Computing Gauss nodes and weights, 573
Computing the zeros of a quadratic polynomial, 103
Conditioning and relative error, 162
Conditioning of conventional row transformations, 251, 252
Conditioning of normal equations, 233
Conditioning of the extended normal equations, 235
Connetion with linear least squares problems Chapter 3, 518
Consistency of implicit midpoint method, 785
Consistent right hand side vectors are highly improbable, 219
Constitutive relations from measurements, 386
Construction of higher order Runge-Kutta single step methods, 801
Contiguous arrays in C++, 39
Control of a robotic arm, 435
Convergence monitors, 665
Convergence of CG as iterative solver, 747
Convergence of Fourier sums, 353
Convergence of global quadrature rules, 577
Convergence of gradient method, 738
Convergence of Hermite interpolation, 550

Convergence of Hermite interpolation with exact slopes, 549
Convergence of inexact simple splitting methods, 860
Convergence of Krylov subspace methods for non-symmetric system matrix, 758
Convergence of naive semi-implicit Radau method, 857
Convergence of Newton's method for matrix inversion, 655
Convergence of Newton's method in 2D, 654
Convergence of quadratic inverse interpolation, 637
Convergence of Remez algorithm, 529
Convergence of secant method, 635
Convergence of semi-implicit midpoint method, 856
Convergence of simple Runge-Kutta methods, 798
Convergence of simple splitting methods, 859
Convergence rates for CG method, 749
Convergence theory for PCG, 753
Convex least squares functional, 226
Convolution of causal sequences, 315
Cosine transforms for compression, 376
Curve design based on B-splines, 455
Curves from interpolation, 447

Damped Broyden method, 666
Damped Newton method, 662
Data points confined to a subspace, 291
Deblurring by DFT, 345
Decay conditions for bi-infinite signals, 353
Decay of cardinal basis functions for natural cubic spline interpolation, 439
Decimal floating point numbers, 95
Denoising by frequency filtering, 336
Derivative of a bilinear form, 647
Derivative of matrix inversion, 651
Derivative of Euclidean norm, 648
Derivative-based local linear approximation of functions, 642
Detecting linear convergence, 606
Detecting order $p > 1$ of convergence, 608
Detecting periodicity in data, 333
Determining the domain of analyticity, 498
Diagonalization of local translation invariant linear grid operators, 371
diagonally dominant matrices from nodal analysis, 206
Different choices for consistent fixed point iterations (II), 616
Different choices for consistent iteration functions (III), 621
Different meanings of "convergence", 484
Differentiating and integrating splines, 431
Discrete evolutions for non-autonomous ODEs, 784
Discretization, 783
Distribution of machine numbers, 96
Divided differences and derivatives, 413

Economical vs. full QR-decomposition, 259
Efficiency of FFT for different backend implementations, 366
Efficiency of FFT-based solver, 374
Efficiency of iterative methods, 641
Efficient associative matrix multiplication, 86
Eigenvectors of circulant matrices, 322
Eigenvectors of commuting matrices, 323
Empiric Convergence of collocation single step methods, 843
Empiric convergence of equidistant trapezoidal rule, 584
Empiric convergence of semi-implicit Euler single-step method, 855
Envelope of a matrix, 200
Envelope oriented matrix storage, 204
Error of polynomial interpolation, 489
Error representation for generalized Lagrangian interpolation, 489
Estimation of "wrong quadrature error"?, 593
Estimation of "wrong" error?, 808
Euler methods for stiff decay IVP, 840
Evolution operator for Lotka-Volterra ODE, 774
Explicit adaptive RK-SSM for stiff IVP, 819
Explicit Euler method as a difference scheme, 779
Explicit Euler method for damped oscillations, 830
Explicit representation of error of polynomial interpolation, 488
Explicit trapzoidal rule for decay equation, 823
Exploiting trigonometric identities to avoid cancellation, 109
Extracting triplets from **Eigen::SparseMatrix**, 184
Extremal numbers in $\mathbb{M}$, 95

Failure of damped Newton method, 662
Failure of Krylov iterative solvers, 758
Fast Toeplitz solvers, 382
Feasibility of implicit Euler timestepping, 780
FFT algorithm by matrix factorization, 362

FFT based on general factorization, 364
FFT for prime vector length, 364
Filtering in Fourier domain, 357
Fit of hyperplanes, 278
Fixed points in 1D, 617
Fractional order of convergence of secant method, 635
Frequency identification with DFT, 332
Frobenius' derivation of the Hermite integral formula, 497
From higher order ODEs to first order systems, 770
Full-rank condition, 225

Gain through adaptivity, 809
Gauss-Newton versus Newton, 677
Gauss-Radau collocation SSM for stiff IVP, 853
Gaussian elimination, 136
Gaussian elimination and LU-factorization, 144
Gaussian elimination for non-square matrices, 140
Gaussian elimination via rank-1 modifications, 142
Gaussian elimination with pivoting for 3×3-matrix, 153
Generalization of data, 385
Generalized bisection methods, 625
Generalized eigenvalue problems and Cholesky factorization, 687
Generalized Lagrange polynomials for Hermite Interpolation, 397
Generalized polynomial interpolation, 396
Gibbs phenomenon, 533
Gradient method in 2D, 737
Gram-Schmidt orthogonalization of polynomials, 569
Gram-Schmidt orthonormalization based on **MyVector** implementation, 47
Group property of autonomous evolutions, 774
Growth with limited resources, 764

Halley's iteration, 630
Heartbeat model, 766
Heating production in electrical circuits, 556
Hesse matrix of least squares functional, 226
Hidden summation, 87
Hump function, 493

Image compression, 284
Image segmentation, 697
Impact of choice of norm, 605
Impact of matrix data access patterns on runtime, 68
Impact of roundoff errors on CG, 746
Implicit nature of collocation single step methods, 843
Implicit RK-SSMs for stiff IVP, 850
Importance of numerical quadrature, 555
In-situ LU-decomposition, 148
Inequalities between vector norms, 119
Initial guess for power iteration, 697
Initialization of sparse matrices in Eigen, 184
Inner products on spaces $\mathcal{P}_m$ of polynomials, 520
Input errors and roundoff errors, 97
Instability of multiplication with inverse, 164
interpolation
  piecewise cubic monotonicity preserving, 427
  shape preserving quadratic spline, 444
Interpolation and approximation: enabling technologies, 475
Interpolation error estimates and the Lebesgue constant, 491
Interpolation error: trigonometric interpolation, 533
Interpolation of vector-valued data, 385
Intersection of lines in 2D, 134

Justification of Ritz projection by min-max theorem, 719

Kinetics of chemical reactions, 834
Krylov methods for complex s.p.d. system matrices, 734
Krylov subspace methods for generalized EVP, 730

Least squares data fitting, 670
Lebesgue Constant for Chebychev nodes, 505
Lebesgue constant for equidistant nodes, 416
Linear Filtering, 337
Linear parameter estimation = linear data fitting, 466
Linear parameter estimation in 1D, 215
linear regression, 219
Linear regression for stationary Markov chains, 377
Linear regression: Parameter estimation for a linear model, 215
Linear system for quadrature weights, 565
Linear systems with arrow matrices, 171
Linearly convergent iteration, 606
Local approximation by piecewise polynomials, 545

Local convergence of Newton's method, 659
local convergence of the secant method, 636
Loss of sparsity when forming normal equations, 234
LSE: key components of mathematical models in many fields, 127
LU-decomposition of flipped "arrow matrix", 196

Machine precision for IEEE standard, 99
Magnetization curves, 419
Many choices for consistent fixed point iterations, 615
Many sequential solutions of LSE, 169
Mathematical functions in a numerical code, 387
Matrix inversion by means of Newton's method, 652
Matrix norm associated with ∞-norm and 1-norm, 120
Matrix representation of interpolation operator, 396
Meaning of full-rank condition for linear models, 225
Meaningful "O-bounds" for complexity, 83
Measuring the angles of a triangle, 216
Midpoint rule, 562
Min-max theorem, 701
Minimality property of Broyden's rank-1-modification, 664
Model reduction in circuit simulation, 474
Modified **Horner scheme** for evaluation of Bezier polynomials, 451
Monitoring convergence for Broyden's quasi-Newton method, 666
Monomial representation, 392
Multi-dimensional data interpolation, 386
Multidimensional fixed point iteration, 620
Multiplication of Kronecker product with vector, 89
Multiplying matrices in EIGEN, 77
Multiplying triangular matrices, 74

Necessary condition for L-stability, 851
Necessity of iterative approximation, 601
Newton iteration by implicit differentiation of F, 627
Newton method and minimization of quadratic functional, 675
Newton method in 1D, 626
Newton's iteration; computational effort and termination, 657
Newton-Cotes formulas, 562
Nodal analysis of linear electric circuit, 127
Non-linear cubic Hermite interpolation, 427
Non-linear data fitting (II), 677
Non-linear electric circuit, 598
Normal equations for some examples from Section 3.0.1, 223
Normal equations from gradient, 224
Notation for single step methods, 785
Numerical Differentiation for computation of Jacobian, 653
Numerical stability and sensitive dependence on data, 123
Numerical summation of Fourier series, 354
NumPy command reshape, 68

Orders of finite-difference single-step methods, 794
Orders of simple polynomial quadrature formulas, 566
Oregonator reaction, 803
Origin of the term "Spline", 437
Oscillating polynomial interpolant, 415
Output of explicit Euler method, 779
Overflow and underflow, 101

Parameter identification for linear time-invariant filters, 376
PCA for data classification , 290
PCA of stock prices, 288
Perceived smoothness of cubic splines, 431
Piecewise cubic interpolation schemes, 435
Piecewise linear interpolation, 389
Piecewise polynomial interpolation, 547
Piecewise quadratic interpolation, 422
Pivoting and numerical stability, 152
Pivoting destroys sparsity, 198
Polynomial fitting, 467
Polynomial interpolation vs. polynomial fitting, 468
Polynomial planar curves, 446
Poor shape-fidelity of high-degree Bezier curves, 450
Power iteration, 694
Preconditioned Newton method, 631
Predator-prey model, 765
Predicting stiffness of non-linear IVPs, 838
Principal axis of a point cloud, 293
Pseudoinverse and SVD, 276

QR-Algorithm, 687
QR-based solution of banded LSE, 257
QR-based solution of linear systems of equations, 256

QR-decomposition of "fat" matrices, 244
QR-decomposition of banded matrices, 250
Quadratic convergence, 608
Quadratic functional in 2D, 734
Quadratur
Gauss-Legendre Ordnung 4, 568
Quadrature errors for composite quadrature rules, 583

Radiative heat transfer, 317
Rank defect in linear least squares problems, 226
Rationale for adaptive quadrature, 589
Rationale for high-order single step methods, 796
Rationale for partial pivoting policy, 155
Rationale for using LU-decomposition in algorithms, 149
Recursive LU-factorization, 148
Reducing bandwidth by row/column permutations, 204
Reducing fill-in by reordering, 205
Reduction of discrete convolution to periodic convolution, 320
Regions of stability for simple implicit RK-SSM, 848
Regions of stability of some explicit RK-SSM, 832
Relative error and number of correct digits, 98
Relevance of asymptotic complexity, 84
Removing a singularity by transformation, 577
Reshaping matrices in EIGEN, 67
Residual based termination of Newton's method, 658
Resistance to currents map, 176
Restarted GMRES, 757
Reuse of LU-decomposition in inverse power iteration, 169
RK-SSM and quadrature rules, 800
Roundoff effects in normal equations, 234
Row and column transformations, 75
Row swapping commutes with forward elimination, 157
Row-wise & column-wise view of matrix product, 71
Runge's example, 485, 490
Runtime comparison for computation of coefficient of trigonometric interpolation polynomials, 461
Runtime of Gaussian elimination, 139
Runtimes of DFT implementations, 359
Runtimes of elementary linear algebra operations in EIGEN, 85

S.p.d. Hessians, 57
Sacrificing numerical stability for efficiency, 173
Sampled audio signals, 305
Scaling a matrix, 74
Semi-implicit Euler single-step method, 855
Sensitivity of linear mappings, 131
Shape preservation of cubic spline interpolation, 438
Shape preserving quadratic spline interpolation, 444
Shifted inverse iteration, 706
Significance of smoothness of interpoland, 489
Simple adaptive stepsize control, 808
Simple adaptive timestepping for fast decay, 823
Simple composite polynomial quadrature rules, 580
Simple preconditioners, 753
Simple Runge-Kutta methods by quadrature & boostrapping, 798
Simplified Newton method, 652
Sine transform via DFT of half length, 370
SIR model, 767
Small residuals by Gaussian elimination, 163
Smoothing of a triangulation, 185
Solving the stage equations for implicit RK-SSMs, 846
Sound filtering by DFT, 336
Sparse LU-factors, 195
Sparse elimination for arrow matrix, 191
Sparse LSE in circuit modelling, 179
Sparse matrices from the discretization of linear partial differential equations, 179
Special cases in IEEE standard, 96
Spectrum of Fourier matrix, 326
Speed of convergence of polygonal methods, 788
spline
interpolants, approx. complete cubic, 551
Splitting linear and decoupled terms, 862
Splitting off stiff components, 861
Square root function, 493
Square root iteration as a Newton iteration, 626
Square root of a s.p.d. matrix, 750
Stability by small random perturbations, 161
Stability function and exponential function, 825
Stability functions of explicit Runge-Kutta single step methods, 825
Stable discriminant formula, 108

Stable implementation of Householder reflections, 245
Stable orthonormalization by QR-decomposition, 93
Stable solution of LSE by means of QR-decomposition, 258
Stage form equations for increments, 846
Standard EIGEN lu() operator versus triangularView() , 167
Stepsize control detects instability, 827
STOP, when stationary in $\mathbb{M}$, 612
Storing orthogonal transformations, 248
Strongly attractive limit cycle, 835
Subspace power methods, 721
Summation of exponential series, 113
SVD and additive rank-1 decomposition, 268
SVD-based computation of the rank of a matrix, 272
Switching to equivalent formulas to avoid cancellation, 110

Tables of quadrature rules, 559
Tangent field and solution curves, 777
Taylor approximation, 476
Termination criterion for contrative fixed point iteration, 622
Termination criterion for direct power iteration, 697
Termination criterion in `pcg`, 755
Termination of PCG, 754
Testing
`= 0.0` in Code 3.3.3.17, 248
Testing equality with zero, 100
Testing stability of matrix×vector multiplication, 122
The "matrix×vector-multiplication problem", 119
The case of finite signals and filters, 312
The inverse matrix and solution of a LSE, 131
The message of asymptotic estimates, 578
The zoo of sparse matrix formats, 180
Timing polynomial evaluations, 403
Timing sparse elimination for the combinatorial graph Laplacian, 192
Tolerances and accuracy, 815
Trade cancellation for approximation, 115
Transformation of quadrature rules, 558
Transient circuit simulation, 768
Transient simulation of RLC-circuit, 827
Trend analysis, 286
Tridiagonal preconditioning, 753
Trigonometric interpolation of 1-periodic analytic functions, 539
Two-dimensional DFT in EIGEN and PYTHON, 342

Understanding the structure of product matrices, 71
Uniqueness of SVD, 269
Unitary similarity transformation to tridiagonal form, 688
Unstable Gram-Schmidt orthonormalization, 92
Using Intel Math Kernel Library (Intel MKL) from EIGEN, 81

Vandermonde matrix, 396
Vectorisation of a matrix, 67
Visualization of explicit Euler method, 778
Visualization: local affine linear approximation for $n = 2$, 643
Visualization: superposition of impulse responses, 310

Why using $\mathbb{K} = \mathbb{C}$?, 324
Wilkinson's counterexample, 160

Abbreviations and Acronyms

CAD ≙ computer aided design, 445
CHI ≙ cubic Hermite interpolation, 423
CM ≙ circulant matrix, 319
CSI ≙ cubic spline interpolation, 431

DCONV ≙ discrete convolution, 314
DCONV ≙ discrete convolution, 314

EV ≙ eigenvector, 325

FRC ≙ full-ranjk condition, 225

GE ≙ Gaussian elimination, 135

HHM ≙ Householder matrix, 243

IC ≙ interpolation conditions, 385
IR ≙ impulse response, 308
IRK-SSM ≙ implicit Runge-Kutta single-step method, 847
IVP ≙ initial-value problem, 762

KCL ≙ Krichoff Current Law, 128

LIP ≙ Lagrange polynomial interpolation, 394
LPI ≙ Lagrangian (interpolation polynomial) approximation scheme, 483
LSE ≙ Linear System of Equations, 127
LT-FIR ≙ finite, linear, time-invariant, causal filter/channel, 308

N.I. ≙ Newton iteration, 626
NEQ ≙ normal eqations, 222
NMT ≙ natural monotonicity test, 660
NMT ≙ natural monotonicity test, 663

OD ≙ overdetermined (for LSE), 214
ODE ≙ ordinary differential equation, 762
ODE ≙ ordinary differential equation, 762

PCA ≙ principal component analysis, 286
PCONV ≙ discrete periodic convolution (of vectors), 317
POD ≙ principal orthogonal decompostion, 292
PP ≙ piecewise polynomial, 546
PPLIP ≙ piecewise polynomial Lagrange interpolation, 546
PSF ≙ point-spread function, 345

QR ≙ quadrature rule, 557

r.h.s ≙ right-hand side, 762

SF ≙ stability function, 825, 833
SSM ≙ single-step method, 784
SVD ≙ singular value decomposition, 266

TI ≙ time-invariant, 307

German terms

L^2-inner product = L^2-Skalarprodukt, 568

back substitution = Rücksubstitution, 137
bandwidth = Bandbreite, 199

cancellation = Auslöschung, 102
capacitance = Kapazität, 128
capacitor = Kondensator, 128
circulant = zirkulant, 319
circulant matrix = zirkulante Matrix, 319
coil = Spule, 128
column (of a matrix) = (Matrix)spalte, 55
column transformation = Spaltenumformung, 75
column vector = Spaltenvektor, 54
composite quadrature formulas = zusammengesetzte Quadraturformeln, 580
computational effort = Rechenaufwand, 82
consistency = Konsistenz, 603
consistent = konsistent, 615
constitutive relation = Kennlinie, 386
constitutive relations = Bauelementgleichungen, 128
constrained least squares = Ausgleichsproblem mit Nebenbedingungen, 299
convergence = Konvergenz, 602
convolution = Faltung, 312

damped Newton method = gedämpftes Newton-Verfahren, 659
deblurring = Entrauschen, 345
dense matrix = vollbesetzte Matrix, 179
descent methods = Abstiegsverfahren, 734
difference scheme = Differenzenverfahren, 779
discrete convolution = diskrete Faltung, 314
divided differences = dividierte Differenzen, 410
dot product = (Euklidisches) Skalarprodukt, 70

eigenspace = Eigenraum, 685
eigenvalue = Eigenwert, 685
electric circuit = elektrischer Schaltkreis/Netzwerk, 127
energy norm = Energienorm, 734
envelope = Hülle, 200
extended normal equations = erweiterte Normalengleichungen, 235

family of functions = Funktionenschar, 763
field of numbers = Zahlenkörper, 54
fill-in = “fill-in”, 195
fixed point = Fixpunkt, 615
fixed point interation = Fixpunktiteration, 614
floating point number = Gleitpunktzahl, 95
floating point numbers = Gleitkommazahlen, 95
forward elimination = Vorwärtselimination, 136
Fourier series = Fourierreihe, 352
frequency domain = Frequenzbereich, 332

Gaussian elimination = Gausselimination, 135

Hessian = Hesse-Matrix, 57
high pass filter = Hochpass, 336

identity matrix = Einheitsmatrix, 56
image segmentation = Bildsegmentierung, 697
impulse response = Impulsantwort, 308
in situ = am Ort, 148
in situ = an Ort und Stelle, 142
initial guess = Anfangsnäherung, 602
initial-value problem (IVP) = Anfangswertproblem, 762
intermediate value theorem = Zwischenwertsatz, 623
inverse iteration = inverse Iteration, 697

Kirchhoff (current) law = Kirchhoffsche Knotenregel, 128
knot = Knoten, 430
Krylov space = Krylovraum, 742

least squares = Methode der kleinsten Quadrate, 214
line search = Minimierung in eine Richtung, 735
linear system of equations = lineares Gleichungssystem, 127
low pass filter= Tiefpassfilter, 336

lower triangular matrix = untere Dreiecksmatrix, 56
LU-factorization = LR-Zerlegung, 143

machine number = Maschinenzahl, 95
machine numbers = Maschinenzahlen, 94
mass matrix = Massenmatrix, 712
matrix factorization = Matrixzerlegung, 143
mesh width = Gitterweite, 545
mesh/grid = Gitter, 545
multiplicity= Vielfachheit, 685

nested = geschachtelt, 477
nodal analysis = Knotenanalyse, 127
node = Knoten, 127
normal equations = Normalengleichungen, 222

order of convergence = Konvergenzordnung, 607
ordinary differential equation (ODE) = gewöhnliche Differentialgleichung, 762
overdetermined = überbestimmt, 214
overflow = Überlauf, 101

partial pivoting = Spaltenpivotsuche, 155
pattern (of a matrix) = Besetzungsmuster, 74
power method = Potenzmethode, 690
preconditioning = Vorkonditionierung, 750
principal = Stammfunktion, 763
principal = Stammgunktion, 50
principal axis transformation = Hauptachsentransformation, 686
principal component analysis = Hauptkomponentenanalyse, 286

quadratic functional = quadratisches Funktional, 734
quadrature node = Quadraturknoten, 557
quadrature weight = Quadraturgewicht, 557

resistor = Widerstand, 128
right hand side vector = rechte-Seite-Vektor, 127
rounding = Rundung, 98
roundoff error = Rundungsfehler, 94
row (of a matrix) = (Matrix)zeile, 55
row transformation = Zeilenumformung, 75
row vector = Zeilenvektor, 54

saddle point problem = Sattelpunktproblem, 300
Sampling = Abtasten, 305
scaling = Skalierung, 74
singular value decomposition = Singlärwertzerlegung, 266
sparse matrix = dünnbesezte Matrix, 179
speed of convergence = Konvergenzgeschwindigkeit, 605
steepest descent = steilster Abstieg, 735
stiffness matrix = Steifigkeitsmatrix, 712
storage format = Speicherformat, 65
subspace correction = Unterraumkorrektur, 741
system matrix = Koeffizientenmatrix, 127

tensor product = Tensorprodukt, 70
tent/hat function = Hutfunktion, 389
termination criterion = Abbruchbedingung, 610
time domain = Zeitbereich, 332
total least squares = totales Ausgleichsproblem, 298
transpose = transponieren/Transponierte, 54
truss = Stabwerk, 710

underflow = Unterlauf, 101
unit vector = Einheitsvektor, 55
upper triangular matrix = obere Dreiecksmatrix, 56

variational calculus = Variationsrechnung, 437

zero padding = Ergänzen durch Null, 462