Numerical Solution of Partial Differential Equations

Prof. Ralf Hiptmair, Prof. Christoph Schwab und Dr. H. Harbrecht
V1.0: summer term 2004, V2.0: winter term 2005/2006

Draft version February 9, 2006

(C) Seminar für Angewandte Mathematik, ETH Zürich
# Contents

1 Second-order scalar elliptic boundary value problems 11

1.1 Classification of boundary value problems 11

1.2 Stationary heat conduction 13

1.3 Boundary conditions 17

1.4 Characteristics of elliptic boundary value problems 19

1.5 Weak derivatives 23

1.6 Variational formulation of boundary value problem 26

1.7 Functional framework 30

1.8 Essential and natural boundary conditions 42

1.9 The Dirichlet principle 47
2 The Finite Element Method (FEM)

2.1 Fundamentals

2.1.1 Galerkin discretization
2.1.2 The (linear) algebraic setting
2.1.3 Principles of FEM
2.1.4 Linear $H^1$-conforming finite elements
2.1.5 Simplicial Lagrangian finite elements
2.1.6 Parametric finite elements
2.1.7 Lagrangian finite elements on quadrilaterals/hexahedra
2.1.8 Degrees of freedom

2.2 Finite Element Implementation

2.2.1 Mesh file format
2.2.2 Assembly
2.2.3 Mesh data structures
2.2.4 Algorithms
2.2.5 Local computations
2.2.6 Numerical quadrature
2.2.7 Treatment of essential boundary conditions
2.2.8 Boundary approximation
2.2.9 Treatment of hanging nodes
2.2.10 Static condensation
3 Special elliptic boundary value problems

3.1 Convection-diffusion problems

3.1.1 Heat conduction in fluid

3.1.2 Singular perturbation

3.1.3 Upwinding

3.1.4 Streamline diffusion

3.2 The Helmholtz equation

3.3 The Stokes Problem

3.3.1 Mixed variational formulation

3.3.2 Saddle point problems

3.3.3 Stable finite element schemes for the Stokes problem

3.4 Non-linear elliptic boundary value problems

3.4.1 Examples

3.4.2 Iterative solution methods

4 Iterative solution of discrete boundary value problems

4.1 Linear stationary iterative methods

4.2 Multigrid methods

4.3 Preconditioned conjugate gradients (PCG)
## Adaptive Finite Element Schemes

7.1 A priori adapted meshes .................................................. 404
7.2 A posteriori error estimation .................................................. 422
   7.2.1 Residual error estimators ................................................. 425
   7.2.2 Hierarchical error estimator ............................................. 428
   7.2.3 Recovery based error estimators ........................................ 432
   7.2.4 Goal oriented error estimators ......................................... 434
7.3 A posteriori mesh adaptation ................................................. 439
   7.3.1 Local mesh refinement .................................................. 439
   7.3.2 Control of local mesh refinement ...................................... 441

## Index

Keywords .................................................................................. 451
Examples .................................................................................. 458
Definitions ................................................................................ 460
MATLAB-CODE ....................................................................... 461
Symbols ................................................................................... 462

## Essential skills

A.1 Chapter 11 Second-order scalar elliptic boundary value problems .......... 466
A.2 Chapter 21 The Finite Element Method (FEM) ..................................... 467
A.3 Chapter 31 Special elliptic boundary value problems ......................... 470
A.4 Chapter 41 Solving discrete boundary value problems ....................... 471
A.5 Chapter 51 Parabolic Boundary Value Problems ................................ 472
A.6 Chapter 61 Numerical Methods for Conservation Laws ..................... 472
A.7 Chapter 71 Adaptive Finite Element Schemes ................................... 473
Reporting errors

Please report any error or dubious manipulation/assertion/reasoning by e-mail!

Examples:

From: "MrX" <mrx@student.ethz.ch>

To: hiptmair@sam.math.ethz.ch

Subject: NPDE05: Error

Error on page XX, Section XX, Formula (XX): index i has to be changed to j
Page XX, Section XX, Theorem XX:
the sign in front of the $\Psi$ seems to be wrong

Preamble

This course is part of the Computational Science and Engineering (CSE) curriculum.

Main skills to be acquired in this course:

- Ability to implement advanced numerical methods for the solution of partial differential equations in MATLAB efficiently
- Ability to modify and adapt numerical algorithms guided by awareness of their mathematical foundations
Ability to select and assess numerical methods in light of the predictions of theory

Ability to identify features of a model that are relevant for the selection and performance of a numerical algorithm

Ability to understand research publications on theoretical and practical aspects of numerical methods for partial differential equations.

This course ≠ Numerical analysis of PDE (→ mathematics curriculum)
Instruction on how to apply software packages
1 Second-order scalar elliptic boundary value problems

1.1 Classification of boundary value problems

*Boundary value problem*

Given a partial differential operator $\mathcal{L}$, a domain $\Omega \subset \mathbb{R}^d$, a boundary differential operator $\mathcal{B}$, boundary values $g$, and a source term $f$, seek a function $u : \Omega \mapsto \mathbb{R}^q$ such that

$$\mathcal{L}(u) = f \quad \text{in } \Omega,$$

$$\mathcal{B}(u) = g \quad \text{on part of boundary } \partial \Omega.$$  

Three main categories of boundary value problems (BVPs) for partial differential equations (PDE):
Elliptic BVPs

- Parabolic initial boundary value problems (IBVPs)
- Hyperbolic IBVPs, among them wave propagation problems and conservation laws.

What are “second-order scalar elliptic boundary value problems”?

Name: second-order: PDE features second spatial derivatives
scalar: Unknown is a function $u: \Omega \mapsto \mathbb{R}$
elliptic: “equilibrium character” (see following sections)
1.2 Stationary heat conduction

\( \Omega \subset \mathbb{R}^3 \): bounded open region occupied by solid object: \( \Omega \rightarrow \text{computational domain} \)

**Fourier's law**

\[ j(x) = -\sigma(x) \nabla u(x), \quad x \in \Omega. \quad (1.2.1) \]

Meaning of quantities:

- \( j \) = heat flux \( ([j] = \frac{W}{m^2}) \)
- \( u \) = temperature \( ([u] = 1K) \)
- \( \sigma \) = heat conductivity \( ([\sigma] = \frac{W}{Km}) \)

\((1.2.1) \Rightarrow \) Heat flow from hot to cold regions linearly proportional to gradient of temperature

\( \sigma \):
- \( \sigma = \sigma(x) \) for non-homogeneous materials (spatially varying heat conductivity)
- \( \sigma \) can even be discontinuous for composite materials
From thermodynamic principles:

\[
\exists \sigma^-, \sigma^+ > 0: \quad 0 < \sigma^- \leq \sigma(x) \leq \sigma^+ < \infty \quad \text{for almost all } x \in \Omega . \quad (1.2.2)
\]

**Terminology:** \((1.2.2) \iff \sigma = \text{uniformly positive}\)

**Remark 1.2.1.** In the case of *Anisotropic materials*:

\[
\sigma(x) \in \mathbb{R}^{3,3} , \quad \sigma(x) = \sigma(x)^T , \quad x \in \Omega \rightarrow \text{heat conductivity tensor}
\]

**Example 1.2.2** (Anisotropic material). Metal wires in polymer matrix modelled by *effective heat conductivity*:

\((1.2.2)\) becomes:

\[
\exists \sigma^-, \sigma^+ > 0: \quad \sigma^- |\xi|^2 \leq \xi^T \sigma(x) \xi \leq \sigma^+ |\xi|^2 \quad \forall \xi \in \mathbb{R}^3 , \text{ for almost all } x \in \Omega . \quad (1.2.3)
\]

**Terminology:** “\(\sigma\) is uniformly symmetric positive definite” \rightarrow **Def. 2.1.9**
Conservation of energy

\[ \int_{\partial V} \mathbf{j} \cdot \mathbf{n} \, dS = \int_V f \, d\mathbf{x} \quad \text{for all “control volumes” } V \]  

(1.2.4)

\( f = \) heat source/sink \( ([f] = \frac{W}{m^3}) \), \( f = f(x) \) and \( f \) can be discontinuous

By Gauss’ theorem we get local form of energy conservation:

\[ \text{div} \mathbf{j} = f . \]  

(1.2.5)

Combine equations

\[ \mathbf{j} = -\sigma(x) \, \text{grad} \, u \quad \text{(1.2.1)} \quad \text{div} \mathbf{j} = f \quad \text{(1.2.5)} \]

\[ -\text{div}(\sigma \, \text{grad} \, u) = f \quad \text{in } \Omega \]  

(1.2.6)

Linear scalar second order elliptic PDE
Remark 1.2.3 (Scaling).

Before numerical treatment: conversion into non-dimensional form by scaling

- length: \( \hat{x} = x/L \), \( L = \) “characteristic length”

- temperature: \( \hat{u} = u/T \), \( T = \) “characteristic temperature difference”

- heat source: \( \hat{f} = f/F \), \( F = \) “typical heat source”

\[
- \text{div} \left( \frac{T \sigma}{L^2 F} \text{grad} \hat{u} \right) = \hat{f} \quad \text{in} \ \Omega .
\]

(one non-dimensional “characteristic parameter” remains)

Note: Notations retained for scaled quantities

Remark 1.2.4. If \( \sigma \equiv \text{const.} \), by rescaling of (1.2.6):

\[
- \Delta u = f \quad \text{in} \ \Omega . \tag{1.2.7}
\]

\[
\Delta = \text{div} \circ \text{grad} = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} + \frac{\partial^2}{\partial x_3^2} = \text{Laplace operator}
\]

(1.2.7) is called Poisson equation
1.3 Boundary conditions

Boundary conditions on surface/boundary $\partial \Omega$ of $\Omega$:

(i) Temperature $u$ is fixed: with $g : \partial \Omega \mapsto \mathbb{R}$ prescribed

\[ u = g \quad \text{on } \partial \Omega . \]  

\[ (1.3.1) \]

Dirichlet boundary conditions

(ii) Heat flux $j$ through $\partial \Omega$ is fixed: with $h : \partial \Omega \mapsto \mathbb{R}$ prescribed, $n : \partial \Omega \mapsto \mathbb{R}^3$ exterior unit normal vectorfield on $\partial \Omega$

\[ j \cdot n = -h \quad \text{on } \partial \Omega . \]  

\[ (1.3.2) \]

Neumann boundary conditions

(iii) Heat flux through $\partial \Omega$ depends on (local) temperature: with increasing function $\Psi : \mathbb{R} \mapsto \mathbb{R}$

\[ j \cdot n = \Psi(u) \quad \text{on } \partial \Omega \]  

\[ (1.3.3) \]

radiation boundary conditions

Example 1.3.1 (Convective cooling (simple model)).

\[ j \cdot n = q(u - u_0) \quad \text{on } \partial \Omega , \quad \text{where } 0 < q^- \leq q(x) \leq q^+ < \infty \quad \text{for almost all } x \in \partial \Omega . \]
Example 1.3.2 (Radiative cooling (simple model)).

\[
\mathbf{j} \cdot \mathbf{n} = \alpha |u - u_0|(u - u_0)^3 \quad \text{on } \partial \Omega,
\]
with \(\alpha > 0\)

\[\rightarrow\quad \text{Non-linear boundary condition}\]

Terminology: If \(g = 0\) or \(h = 0\) → homogeneous Dirichlet or Neumann boundary conditions

Remark 1.3.3 (mixed boundary conditions).

Different boundary conditions can be prescribed on different parts of \(\partial \Omega\)
\(\rightarrow\) mixed boundary conditions

\[
\text{Example 1.3.4 (“Wrapped rock on a stove”).}
\]

- Non-homogeneous Dirichlet boundary conditions on \(\Gamma_D \subset \partial \Omega\)
- Homogeneous Neumann boundary conditions on \(\Gamma_N \subset \partial \Omega\)
- Convective cooling boundary conditions on \(\Gamma_R \subset \partial \Omega\)

Partition: \(\partial \Omega = \overline{\Gamma_D} \cup \overline{\Gamma_N} \cup \overline{\Gamma_R}, \quad \Gamma_D, \Gamma_N, \Gamma_R \text{ mutually disjoint}\)

\[- \text{div}(\sigma \, \text{grad} \, u) = f + \text{boundary conditions} \Rightarrow \text{elliptic boundary value problem (BVP)}\]
For second order elliptic boundary value problems exactly one boundary condition is needed on any part of $\partial \Omega$.

Remark 1.3.5. Solution operator $\begin{pmatrix} f \\ g \end{pmatrix} \mapsto u$ for (1.2.6), (1.3.1) is linear.

1.4 Characteristics of elliptic boundary value problems

Qualitative insights gained from heat conduction model:

- **continuity**: the temperature $u$ must be continuous (jump in $u \rightarrow j = \infty$).

- **normal component of $j$ across surfaces inside $\Omega$**: must be continuous (jump in $j \cdot n \rightarrow$ heat source $f$ of infinite intensity).

- **interior smoothness of $u$**: $u$ smooth where $f$ and $\sigma$ smooth.
non-locality: local alterations in \( f, g, h \) affect \( u \) everywhere in \( \Omega \).

quasi-locality: If local changes in \( f, g, h \) confined to \( \Omega' \subset \Omega \), their effects decay away from \( \Omega \).

maximum principle: (in the absence of heat sources extremal temperatures on the boundary)

\[
\text{if } f \equiv 0, \text{ then } \inf_{y \in \partial \Omega} u(y) \leq u(x) \leq \sup_{y \in \partial \Omega} u(y) \text{ for all } x \in \Omega
\]

Typical features of solutions of elliptic boundary value problems

Example 1.4.1 (Scalar elliptic boundary value problem in one space dimension).

Poisson equation \( \rightarrow (1.2.7) \) in 1D: \[ -u'' = f \]

\( f \) discontinuous, piecewise \( C^0 \) \( \Rightarrow \) \( u \in C^1 \), piecewise \( C^2 \)
Example 1.4.2 (Smoothness of solution of scalar elliptic boundary value problem).

\[-\Delta u = f(x) \quad \text{in } \Omega := ]0, 1[^2, \quad u = 0 \quad \text{on } \partial \Omega, \quad \text{(1.4.1)}\]

\[f(x) := \text{sign}(\sin(2\pi k_1 x_1) \sin(2\pi k_2 x_2)), \quad x \in \Omega, \quad k_1, k_2 \in \mathbb{N}.\]

Approximate solution computed by means of linear Lagrangian finite elements + lumping

(→ Sect. 2, details in Sect. 2.1.4, 2.2.6)

Source term \(f(x), k_1 = k_2 = 2\)

Solution of \((1.4.1)\)

“Smooth” \(u\) despite “rough” \(f\)!
Example 1.4.3 (Quasi-locality of solution of scalar elliptic boundary value problem).

\[ -\Delta u = f_\delta(x) \quad \text{in } \Omega := ]0, 1[^2, \quad u = 0 \quad \text{on } \partial\Omega, \]

\[ f_\delta(x) = \begin{cases} 
\delta^{-2}, & \text{if } \|x - \left(\frac{1}{2}\right)\|_2 \leq \delta, \\
0, & \text{elsewhere.}
\end{cases} \quad (1.4.3) \]
1.5 Weak derivatives

Assume translation symmetry in two coordinate directions ($\frac{\partial}{\partial x_2} = \frac{\partial}{\partial x_3} = 0$)

1D stationary heat conduction model (→ Ex. 1.4.1):

$$-\frac{d}{dx} (\sigma(x) \frac{d}{dx} u) = 0 \quad u(0) = 1 \quad u(1) = u_1 \in \mathbb{R},$$

where $\sigma(x) = \begin{cases} \sigma_1 & \text{for } 0 < x < \frac{1}{2}, \\ \sigma_2 & \text{for } \frac{1}{2} < x < 1. \end{cases}$

(Heat conduction in a big flat wall)

$$u(x) = \begin{cases} \frac{2u_1 \sigma_2 x}{\sigma_1 + \sigma_2} & \text{for } 0 < x < \frac{1}{2}, \\ \frac{\sigma_1 + \sigma_2}{2u_1 \sigma_1 (x - 1)} + u_1 & \text{for } \frac{1}{2} < x < 1. \end{cases}$$

Note: For discontinuous $\sigma$ (normal) continuity of $j$ rules out continuous differentiability of $u$

How to make sense of $\frac{d}{dx} u$?
Definition 1.5.1 (Weak derivative). The weak partial derivative $\frac{\partial u}{\partial x_j}$, $j = 1, \ldots, d$ (derivative in the sense of distributions), of a locally integrable function $u : \Omega \subset \mathbb{R}^d \mapsto \mathbb{R}$ is, if it exists, a locally integrable function $\partial_j u : \Omega \mapsto \mathbb{R}$ that satisfies

$$\int_{\Omega} \partial_j u \, v \, dx = - \int_{\Omega} u \, \frac{\partial v}{\partial x_j} \, dx \quad \forall v \in C^\infty_0(\Omega),$$

where $C^\infty_0(\Omega)$ is the space of compactly supported smooth functions $\Omega \mapsto \mathbb{R}$.

[Multidimensional integration by parts formula]

Weak derivative is genuine generalization of classical derivative

Example 1.5.1 (Weak derivative in 1D).

$$u(x) = \begin{cases} u_1(x) & \text{for } 0 < x < \frac{1}{2}, \\ u_2(x) & \text{for } \frac{1}{2} \leq x < 1, \end{cases}$$

with $u_1, u_2$ smooth, $u_1(1/2) = u_2(1/2)$. 

Answer: weak derivative (piecewise derivative almost everywhere)
We show that in this case weak derivative = piecewise derivative

that is

\[
\frac{du}{dx} = \begin{cases} 
  u'_1(x) & \text{for } 0 < x < \frac{1}{2}, \\
  u'_2(x) & \text{for } \frac{1}{2} \leq x < 1.
\end{cases}
\]

\[
\int_0^1 \frac{du}{dx} v \, dx = \int_0^{1/2} u'_1 v \, dx + \int_{1/2}^1 u'_2 v \, dx
\]

\[
= [u_1 v]_{0}^{1/2} - \int_0^{1/2} u_1 v' \, dx + [u_2 v]_{1/2}^{1} - \int_{1/2}^1 u_2 v' \, dx
\]

\[
= (u_1(1/2) - u_2(1/2)) v(1/2) - \int_{0}^{1} u v' \, dx,
\]

for \( v \in C^\infty([0,1]) \Rightarrow v(0) = v(1) = 0. \) fits definition of weak derivative

Remark 1.5.2. Weak \( \frac{\partial}{\partial x_j} \) → weak differential operators

\[
\text{grad } u := \left( \frac{\partial u}{\partial x_1}, \ldots, \frac{\partial u}{\partial x_d} \right)^T,
\]

\[
\text{div } j := \frac{\partial j_1}{\partial x_1} + \cdots + \frac{\partial j_d}{\partial x_d},
\]

\[
\Delta u, \text{ curl } j, \ldots
\]

Below: All differential operators will be understood in weak sense
New interpretation of a partial differential equation

PDE in weak sense $\iff$ PDE in classical sense

1.6 Variational formulation of boundary value problem

Formal approach:

STEP 1: test PDE with smooth functions (cf. weak derivative $\rightarrow$ Def. 1.5.1)

STEP 2: integrate over domain (cf. weak derivative $\rightarrow$ Def. 1.5.1)

STEP 3: perform integration by parts
Example 1.6.1 (Variational formulation of pure Dirichlet problem for heat equation).

\[ - \text{div}(\sigma \text{ grad } u) = f \quad \text{in } \Omega , \quad u = g \quad \text{on } \partial \Omega . \]  \hspace{1cm} (1.6.1)

STEP 1 & 2: test with \( v \in C_0^\infty (\Omega) \)

\[ - \int_\Omega \text{div}(\sigma \text{ grad } u) \ v \ dx = \int_\Omega f \ v \ dx . \]  \hspace{1cm} (1.6.2)

Note: \( v|_{\partial \Omega} = 0 \) for test function, because \( u \) already fixed on \( \partial \Omega \).

STEP 3: use Green’s formulas on \( \Omega \subset \mathbb{R}^d \) (multidimensional integration by parts)

\textbf{Theorem 1.6.1} (Green’s first formula). If \( \partial \Omega \) is piecewise smooth, then for all vector fields \( j \in (C^1(\overline{\Omega}))^3 \) and functions \( v \in C^1(\overline{\Omega}) \)

\[ \int_\Omega \text{div} \ j \ v + j \cdot \text{grad} \ v \ dx = \int_{\partial \Omega} j \cdot n \ v \ dS \]  \hspace{1cm} (1.6.3)

\textbf{Proof.} [by Gauss’ theorem \( \triangle \)]
Apply (1.6.3) to (1.6.2) with $j := \sigma \text{grad} u$:

\[
\int_{\Omega} \sigma \text{grad} u \cdot \text{grad} v \, dx - \int_{\partial \Omega} \sigma \text{grad} u \cdot n \, v \, dS = \int_{\Omega} f v \, dx \quad \forall v \in C_0^\infty(\Omega).
\]

Variational form of (1.6.1): seek $u : \Omega \mapsto \mathbb{R}$, $u = g$ on $\partial \Omega$ such that

\[
\int_{\Omega} \sigma \text{grad} u \cdot \text{grad} v \, dx = \int_{\Omega} f v \, dx \quad \forall v \in C_0^\infty(\Omega) \quad (1.6.4)
\]

**Example 1.6.2** (Variational formulation: heat conduction with general radiation boundary conditions).

\[
\text{BVP:} \quad -\text{div}(\sigma \text{grad} u) = f \quad \text{in} \ \Omega, \quad -\sigma \text{grad} u \cdot n = \Psi(u) \quad \text{on} \ \partial \Omega \quad (1.6.5)
\]

**STEP 1:** $u|_{\partial \Omega}$ not fixed $\Rightarrow$ test with $v \in C^\infty(\Omega)$

\[
-\int_{\Omega} \text{div}(\sigma \text{grad} u) \, v \, dx = \int_{\Omega} f v \, dx \quad \forall v \in C^\infty(\Omega).
\]

**STEP 2:** apply Green's first formula (1.6.3)

\[
\int_{\Omega} \sigma \text{grad} u \cdot \text{grad} v \, dx + \int_{\partial \Omega} \left[ -\sigma \text{grad} u \cdot n \right] v \, dS = \int_{\Omega} f v \, dx \quad \forall v \in C^\infty(\Omega). \]
Variational formulation of (1.6.5): seek $u : \Omega \rightarrow \mathbb{R}$ such that

$$
\int_{\Omega} \sigma \ \text{grad} \ u \cdot \text{grad} \ v \ dx + \int_{\partial \Omega} \Psi(u) \ v \ dS = \int_{\Omega} f \ v \ dx \quad \forall v \in C^\infty(\overline{\Omega}).
$$

(1.6.6)

Variational form of Neumann problem recovered for $\Psi(u) = -h$.

Observation: when we test (1.6.6) with $v \equiv 1$

$$
-\int_{\partial \Omega} h \ dS = \int_{\Omega} f \ dx
$$

(1.6.7)

This is a compatibility condition for the existence of (variational) solutions of the Neumann problem!

[In heat conduction: consequence of conservation of energy $\rightarrow (1.2.4)$]

(1.6.4) & (1.6.6) represent the variational interpretation of the boundary value problems.

**Theorem 1.6.2.** If $\sigma$ is smooth, classical solutions $u \in C^0(\overline{\Omega}) \cap C^2(\Omega)$ of the boundary value problems (1.6.1) and (1.6.5) are also variational solutions.
Proof. Apply Theorem 1.6.1 as in the derivation of the weak formulations.

Theorem 1.6.3. Solutions $u$ of (1.6.4) and (1.6.6), respectively, provide weak solutions of the heat equation $-\text{div}(\sigma \ \text{grad} \ u) = f$ in $\Omega$.

Proof. Straightforward from Definition 1.5.1.

1.7 Functional framework

(1.6.4) and (1.6.6) for $\Psi(u) = qu$ (convective cooling, Ex. 1.3.1) or $\Psi(u) = -h$ (Neumann boundary conditions) are linear variational problems of the form

$$ u \in V : \ a(u, v) = f(v) \quad \forall v \in V $$

(1.7.1)

where $V =$ real vector space of functions $\Omega \mapsto \mathbb{R}$ (trial and test space)

\begin{align*}
    a & = (\text{symmetric}) \ \text{bilinear form} \ a : V \times V \mapsto \mathbb{R} \\
    f & = \text{linear form} \ f : V \mapsto \mathbb{R}
\end{align*}
Definition 1.7.1. Given an \( \mathbb{R} \)-vector space \( V \), a linear form \( f \) is a mapping \( f : V \mapsto \mathbb{R} \) that satisfies

\[ f(\alpha u + \beta v) = \alpha f(u) + \beta f(v) \quad \forall u, v \in V, \forall \alpha, \beta \in \mathbb{R}. \]

A bilinear form \( a \) on \( V \) is a mapping \( a : V \times V \mapsto \mathbb{R} \), for which

\[ a(\alpha_1 v_1 + \beta_1 u_1, \alpha_2 v_2 + \beta_2 u_2) = \alpha_1 \alpha_2 a(v_1, v_2) + \alpha_1 \beta_2 a(v_1, u_2) + \beta_1 \alpha_2 a(u_1, v_2) + \beta_1 \beta_2 a(u_1, u_2) \]

for all \( u_i, v_i \in V, \alpha_i, \beta_i \in \mathbb{R}, i = 1, 2 \).

Reminder:

Definition 1.7.2 (Norm). A norm \( \| \cdot \|_V \) on an \( \mathbb{R} \)-vector space \( V \) is a mapping \( \| \cdot \|_V : V \mapsto \mathbb{R}^+_0 \), such that

\[ \| v \|_V = 0 \iff v = 0 \quad \forall v \in V \quad \text{(N1)} \]

\[ \| \lambda v \|_V = |\lambda| \| v \|_V \quad \forall \lambda \in \mathbb{R}, \forall v \in V, \quad \text{(N2)} \]

\[ \| w + v \|_V \leq \| w \|_V + \| v \|_V \quad \forall w, v \in V. \quad \text{(N3)} \]
**Definition 1.7.3** (Inner product). A bilinear form $a$ on a $\mathbb{R}$-vector space $V$ is called an **inner product**, if it is symmetric positive definite (s.p.d.), that is,

\[
\begin{align*}
    a(u, v) &= a(v, u) \quad \forall u, v \in V, \quad \text{(IP1)} \\
    v \neq 0 &\iff a(v, v) > 0 \quad \forall v \in V. \quad \text{(IP2)}
\end{align*}
\]

**Definition 1.7.4.** An $\mathbb{R}$-**Hilbert space** $V$ is a complete, normed vector space, whose norm $\| \cdot \|_V$ is derived from an inner product $(\cdot, \cdot)_V : V \times V \mapsto \mathbb{R}$ according to

\[
\|v\|_V^2 := (v, v) \quad \forall v \in V.
\]

**Important properties of (bi-)linear forms:**

**Definition 1.7.5** (Continuity of linear forms). A linear form $f$ on a Hilbert space $V$ is **continuous**, if

\[
\exists C_f > 0: \quad |f(v)| \leq C_f \|v\|_V \quad \forall v \in V.
\]
Definition 1.7.6 (Continuity of bilinear forms). A bilinear form $a$ on a Hilbert space $V$ is continuous, if

$$
\exists C_A > 0: \quad |a(u, v)| \leq C_A \|u\|_V \|v\|_V \quad \forall u, v \in V.
$$

Definition 1.7.7 (Ellipticity of bilinear forms). A bilinear form $a$ on a Hilbert space $V$ is $V$-elliptic if

$$
\exists \gamma > 0: \quad a(u, u) \geq \gamma \|u\|^2_V \quad \forall u \in V.
$$

$\gamma$ is called the ellipticity constant.
**Theorem 1.7.8** (Lax-Milgram Lemma). Assume that \( a \) and \( f \) are a continuous bilinear/linear form on the Hilbert space \( V \). If, moreover, \( a \) is \( V \)-elliptic, with ellipticity constant \( \gamma > 0 \), then the linear variational problem

\[
\begin{align*}
  u \in V: \quad a(u, v) &= f(v) \quad \forall v \in V \\
\end{align*}
\]

has a unique solution \( u \in V \) that satisfies the stability estimate

\[
\|u\|_V \leq \frac{1}{\gamma} \sup_{v \in V \setminus \{0\}} \frac{|f(v)|}{\|v\|_V}.
\]

**Proof.**

1. Uniqueness of solutions: if both \( u_1, u_2 \in V \) solve (1.7.1), then

\[
0 = a(u_1 - u_2, v) \quad \forall v \in V \quad \Rightarrow \quad \|u_1 - u_2\|_V^2 \leq \frac{1}{\gamma} a(u_1 - u_2, u_1 - u_2) = 0 \quad \Rightarrow \quad u_1 = u_2.
\]

2. Existence of solutions: \( \rightarrow \) profound functional analysis (requires completeness of \( V \))

3. Stability of solutions:

\[
\gamma \|u\|_V^2 \leq a(u, u) = f(u) \leq \sup_{v \in V \setminus \{0\}} \frac{|f(v)|}{\|v\|_V} \|u\|_V.
\]
Q: What is \( V \) for varational formulations (1.6.4) and (1.6.6)?

For (1.6.4):

\[
a(u, v) := \int_\Omega \sigma \text{grad} u \cdot \text{grad} v \, dx
\]

\( a = \text{symmetric bilinear form on } C_0^\infty(\overline{\Omega}) \)

Idea: use \( a(\cdot, \cdot) \) as inner product on \( C_0^\infty(\overline{\Omega}) \) \( \Rightarrow \) “energy norm” \( \| \cdot \|_A \)

Continuity/V-ellipticity of \( a(\cdot, \cdot) \) trivially satisfied

\( a: \) bilinear (ok) symmetric (ok) positive definite (?)

Estimate in 1D for \( \Omega = ]a, b[ \), \( a, b \in \mathbb{R} \):

\[
\int_a^b |u(x)|^2 \, dx \leq \frac{1}{\sigma_-(b-a)^2} \int_a^b \sigma(x)|u'(x)|^2 \, dx \quad \forall u \in C_0^\infty(]a, b[) .
\]

\[
a(u, v) = \int_a^b \sigma(x)u'(x) \, v'(x) \, dx \quad \text{symmetric positive definite on } C_0^\infty(]a, b[) .
\]

Generalization to \( \mathbb{R}^d \):

**Theorem 1.7.9** (First Poincaré-Friedrichs inequality). If \( \Omega \subset \mathbb{R}^d, d \in \mathbb{N}, \) is bounded, then

\[
\| u \|_0 \leq \text{diam}(\Omega) \| \text{grad} u \|_0 \quad \forall u \in C_0^\infty(\overline{\Omega}) .
\]
Notations: \[ \|v\|_0^2 := \int_\Omega |v|^2 \, dx \quad \text{("L}^2\text{-norm")}, \quad \text{diam}(\Omega) := \sup\{|x - y|, \ x, y \in \Omega\}. \]

Special energy norm: \[ \|u\|_1^2 := \|u\|_0^2 + \|\nabla u\|_0^2 \]

Assumption (1.2.2): \[ a(u, v) := \int_\Omega \sigma(x) \nabla u \cdot \nabla v \, dx \quad \text{is inner product on} \quad C_0^\infty(\Omega). \]

Example 1.7.1. \[ \Omega = ]0, 1[, \ \sigma \equiv 1 \]

"Hat function" \[ u(x) = \begin{cases} 2x & \text{for } 0 < x < \frac{1}{2}, \\ 2(1 - x) & \text{for } \frac{1}{2} < x < 1. \end{cases} \]

Compute \[ a(u, u) = \int_0^1 |u'(x)|^2 \, dx = 4 < \infty. \]

\[ \exists u : \Omega \mapsto \mathbb{R}, \ u(0) = u(1) = 0, \ a(u, u) < \infty, \ \text{but} \ u \not\in C_0^\infty(\Omega)! \]

[Meaningful solutions for temperature distribution]

\[ C_0^\infty(\Omega) \text{ not complete w.r.t. norm } \|u\|_A := a(u, u)^{1/2} \quad \Rightarrow \quad (C_0^\infty(\Omega), \|\cdot\|_A) \text{ no Hilbert space!} \]
Use “trick” from functional analysis: completion

normed vector space \rightarrow \text{complete vector space}

Use

\[ V = H_0^1(\Omega) := \{ v : \Omega \mapsto \mathbb{R} : \|v\|_A < \infty, v|_{\partial \Omega} = 0 \} \]

Sobolev space = Hilbert space with norm $\| \cdot \|_A$

Notation: $H_0^1(\Omega)$ ← superscript “1”, because first derivatives occur in norm

← subscript “0”, because zero on $\partial \Omega$

Another important space:

$L^2(\Omega) := \{ v : \Omega \mapsto \mathbb{R} : \|v\|_0^2 := \int_\Omega |v|^2 \, dx < \infty \}$

Notation: $L^2(\Omega)$ ← superscript “2”, because square in the definition of norm $\| \cdot \|_0$

Note: Discontinuities allowed for functions in $L^2(\Omega)$
Do not be afraid of Sobolev spaces!
It is only the norms that matter, the “spaces” are irrelevant!

Summary (relationships of concepts):

Variational problem ➞ norm (ensuring continuity/ellipticity) ➞ Sobolev space

Now: Neumann problem for \(-\text{div}(\sigma \, \text{grad } u) = f\): variational formulations (1.6.6)

\[
a(u, v) = \int_{\Omega} \sigma \, \text{grad } u \cdot \text{grad } v \, dx \quad u, v \in C^\infty(\overline{\Omega}).
\]

Same bilinear form as for Dirichlet problem, but different space!

Obvious:

\[
a(u, u) = 0 \quad \text{for } u \equiv \text{const.} \quad \text{on } \Omega
\]

\[
a(\cdot, \cdot) \text{ is no inner product on } C^\infty(\overline{\Omega}).
\]

Estimate in 1D for \(\Omega = ]a, b[, a, b \in \mathbb{R}\): for \(u \in C^\infty(\overline{\Omega})\)

\[
u(x) = u(y) + \int_{y}^{x} u'(t) \, dt \quad , x, y \in ]a, b[ \quad u(x) = \frac{1}{b - a} \int_{a}^{b} \left\{ u(y) + \int_{y}^{x} u'(t) \, dt \right\} \, dy.
\]
\[
\int_a^b |u(x)|^2 \, dx = \frac{1}{(b-a)^2} \int_a^b \left| \int_a^b u(y) \, dy + \int_a^b \int_y^x u'(t) \, dt \, dy \right|^2 \, dx
\]

\[
\leq^* \frac{2}{(b-a)^2} \left( \int_a^b \left| \int_a^b u(y) \, dy \right|^2 \, dx + \int_a^b \left| \int_y^x u'(t) \, dt \, dy \right|^2 \, dx \right)
\]

\[
\leq^* \frac{2}{b-a} \left| \int_a^b u(y) \, dy \right|^2 + \frac{2}{(b-a)^2} \int_a^b (b-a) \int_y^x \left| u'(t) \right|^2 \, dt \, dy \, dx
\]

\[
\leq^* \frac{2}{b-a} \left( \left| \int_a^b u(y) \, dy \right|^2 + \int_a^b \int_a^y |y-x| \, \left| u'(t) \right|^2 \, dt \, dy \, dx \right)
\]

\[
\leq \frac{2}{b-a} \left| \int_a^b u(y) \, dy \right|^2 + 2(b-a)^2 \int_a^b \left| u'(t) \right|^2 \, dt ,
\]

where \( \ast \) we used \((x + y)^2 \leq 2(x^2 + y^2), x, y \in \mathbb{R}, \)

at \( \ast \) we used the Cauchy-Schwarz inequality

\[
\left( \int_a^b f(x)g(x) \, dx \right)^2 \leq \int_a^b |f(x)|^2 \, dx \int_a^b |g(x)|^2 \, dx \quad \forall f, g \in L^2([a, b])
\]
\[
\int_a^b u(y) \, dy = 0 \quad \text{and} \quad \int_a^b |u'(t)|^2 \, dt = 0 \quad \Rightarrow \quad \int_a^b |u(x)|^2 \, dx = 0 \quad \Rightarrow \quad u = 0 .
\]

Restrict variational problem (1.6.6) to functions with vanishing mean.

Then \(a(u, u) = \text{inner product}\)

Generalization to \(\mathbb{R}^d\):

**Theorem 1.7.10** (Second Poincaré-Friedrichs inequality). If \(\Omega \subset \mathbb{R}^d\) is bounded, then there is \(C_{\text{PF}} = C_{\text{PF}}(\Omega) > 0\) such that

\[
\|u\|_0 \leq C_{\text{PF}} \|\nabla u\|_0 \quad \forall u \in C^\infty_*(\overline{\Omega}) := \{v \in C^\infty(\Omega), \int_\Omega v(x) \, dx = 0\} .
\]

\(a(\cdot, \cdot)\) is an inner product on \(C^\infty_*(\overline{\Omega})\).

Appropriate Sobolev space for Neumann problem:

\[
H^1_*(\Omega) := \{v : \Omega \mapsto \mathbb{R}: \int_\Omega v \, dx = 0, \|v\|_A < \infty\} .
\]

= **Hilbert space** with norm \(\|\cdot\|_A\).

---

1.7

p. 40
Variational formulation of homogeneous Neumann problem for $-\text{div}(\sigma \, \text{grad} \, u) = f$:
seek $u \in H^1_*(\Omega)$ such that

$$a(u, v) := \int_{\Omega} \sigma(x) \, \text{grad} \, u \cdot \text{grad} \, v \, dx = f(v) := \int_{\Omega} f \, v \, dx \quad \forall v \in H^1_*(\Omega). \quad (1.7.4)$$

**Corollary 1.7.11.** If $f \in L^2(\Omega)$, then (1.7.4) has a unique solution $u \in H^1_*(\Omega)$ fulfilling $\|u\|_A \leq C_{PF}^2(\sigma^-)^{-1} \|f\|_0$ with $C_{PF}$ from Thm. 1.7.10 and $\sigma^-$ from (1.2.2).

**Proof.** Use the general Cauchy-Schwarz inequality

$$\left( \int_{\Omega} f(x) \, g(x) \, dx \right)^2 \leq \int_{\Omega} f(x)^2 \, dx \int_{\Omega} g(x)^2 \, dx \quad \forall f, g \in L^2(\Omega), \quad (1.7.5)$$

which implies the continuity of source functional $f$ on $H^1_*(\Omega)$, the estimates

$$\|v\|_0 \leq C_{PF} \|\text{grad} \, v\|_0 \quad , \quad \|v\|_A \geq \sqrt{\sigma^-} \|\text{grad} \, v\|_0 ,$$

and

$$\sup_{v \in H^1_*(\Omega) \setminus \{0\}} \frac{|f(v)|}{\|v\|_A} \leq \sup_{v \in H^1_*(\Omega) \setminus \{0\}} \frac{\|v\|_0}{\|v\|_A} \frac{f(0)}{\|f\|_0} \leq \frac{C_{PF}}{\sqrt{\sigma^-}} \frac{f(0)}{\|f\|_0} .$$

Then, the assertion follows from Lax-Milgram Lemma Thm. 1.7.8. \qed
Remark 1.7.2. “Point source”: \( f = \delta_p, \ p \in \Omega \) \( \Rightarrow \) \( f(v) = v(p) \)

\[
v \in H^1_0(\Omega) \mapsto v(p) \in \mathbb{R} \text{ not continuous!}
\]

Example: \( u(x) = \log |\log(|x|/e)| \) on \( \Omega := \{x \in \mathbb{R}^2: |x| < 1\} \)

\[
\|u\|_A < \infty \quad u|_{\partial\Omega} = 0 \quad u \text{ unbounded}
\]

\[
\Rightarrow \quad \text{Point source meaningless in context of variational interpretation}
\]

1.8 Essential and natural boundary conditions

Issue: Linear variational problem (1.7.1) for Dirichlet problem?

(1.6.4) in Sobolev framework: seek \( u \in H^1_0(\Omega) + \tilde{g} \) with

\[
\int_{\Omega} \sigma(x) \nabla u \cdot \nabla v \, dx = \int_{\Omega} f v \, dx \quad \forall v \in H^1_0(\Omega),
\]

\( \tilde{g} = \text{extension of Dirichlet data } g: \quad \tilde{g} : \Omega \mapsto \mathbb{R}, \quad \tilde{g}|_{\partial\Omega} = g \)
Dirichlet boundary conditions for (1.2.6) directly imposed on test/trial space in (1.8.1):

\[ \quad \text{essential boundary conditions for (1.8.1)} \]

Yet, (1.8.1) does not match (1.7.1), because \( H_0^1(\Omega) + \tilde{g} \) is affine space.

Converting (1.8.1) into a linear variational problem: seek \( u_0 \in H_0^1(\Omega) \) such that

\[
\begin{align*}
\int_\Omega \sigma(x) \nabla(u_0 + \tilde{g}) \cdot \nabla v \, dx &= \int_\Omega f v \, dx \quad \forall v \in H_0^1(\Omega) .
\end{align*}
\]  
(1.8.2)

From \( u_0 \) we recover: \( u = u_0 + \tilde{g} \).

\[ (1.8.2) = \text{linear variational problem} \ (1.7.1) \text{ with} \]

\[
\begin{align*}
a(u, v) &= \int_\Omega \sigma(x) \nabla u \cdot \nabla v \, dx \quad \text{(see (1.6.4))} ,
\end{align*}
\]

\[
f(v) = \int_\Omega f v \, dx - \int_\Omega \sigma(x) \nabla \tilde{g} \cdot \nabla v \, dx .
\]

Has to be continuous on \( H_0^1(\Omega) \)

Continuity of source term \( \Rightarrow \) requirement on \( \tilde{g} \)

Demand

\[
\|\tilde{g}\|_1^2 := \|\tilde{g}\|_0^2 + \|\nabla \tilde{g}\|_0^2 < \infty
\]
\[ \tilde{g} \in H^1(\Omega) := \{ v : \Omega \leftrightarrow \mathbb{R} : \| v \|_1 < \infty \} \]

Continuity of \( f : H^1_0(\Omega) \leftrightarrow \mathbb{R} \) by Cauchy-Schwarz inequality (1.7.5)

**Remark.** If \( g : \partial \Omega \leftrightarrow \mathbb{R} \) piecewise smooth, but discontinuous \( \Rightarrow \) \( \mathbb{A} \tilde{g} \)

[Jump in temperature on \( \partial \Omega \) would entail infinitely thin perfect insulator]

What are valid Dirichlet data \( g \)?

1. Piecewise smooth, continuous functions on \( \partial \Omega \)
2. Functions belonging to the trace space \( (\Omega \subset \mathbb{R}^d) \)

\[ g \in H^{1/2}(\partial \Omega) := \{ v : \partial \Omega \leftrightarrow \mathbb{R} : \| v \|_{1/2, \partial \Omega} := \int_{\partial \Omega} |v|^2 \, dS + \int_{\partial \Omega} \int_{\partial \Omega} \frac{|v(x) - v(y)|^2}{|x - y|^{d+1}} \, dS(x, y) < \infty \}. \]
Theorem 1.8.1 (Trace theorem for $H^1(\Omega)$). The pointwise restriction mapping $\gamma : C^\infty(\overline{\Omega}) \mapsto C^0(\partial \Omega)$, $(\gamma u)(x) = u(x)$ for all $x \in \partial \Omega$, satisfies

$$\exists C = C(\Omega) > 0: \quad \|\gamma u\|_{1/2, \partial \Omega} \leq C \|u\|_1 \quad \forall u \in C^\infty(\overline{\Omega}) ,$$

and, hence, can be extended to a continuous trace operator

$$\gamma : H^1(\Omega) \mapsto H^{1/2}(\partial \Omega) ,$$

that has a continuous right inverse (an extension operator $E : H^{1/2}(\partial \Omega) \mapsto H^1(\Omega)$).

Extend (1.7.4): variational formulation of inhomogeneous Neumann problem:

seek $u \in H^1_*(\Omega)$

$$\int_{\Omega} \sigma(x) \text{grad} u \cdot \text{grad} v \, dx = \int_{\Omega} f v \, dx + \int_{\partial \Omega} h v \, dS \quad \forall v \in H^1_*(\Omega) \ . \quad (1.8.3)$$

$\rightarrow$

Linear variational problem according to (1.7.1)

Neumann boundary conditions for (1.2.6) do not show up in trial/test space:

$\rightarrow$

natural boundary conditions for (1.8.3)

Q: Requirements on $h$ to get continuous $v \in H^1_*(\Omega) \mapsto \int_{\partial \Omega} h v \, dS$ ?
Cauchy-Schwarz inequality (1.7.5) on $\partial \Omega$ + Trace theorem 1.8.1 $\Rightarrow h \in L^2(\partial \Omega)$ is enough!

**Remark 1.8.1.** Are Neumann boundary conditions “contained” in (1.8.3) ?

Test with $v \in C_0^\infty(\Omega)$ and Thm. 1.6.1: $-\text{div}(\sigma \, \text{grad} \, u) = f$ in $\Omega$

Then test with $v \in C^\infty(\overline{\Omega})$, use PDE, and Thm. 1.6.1: $\sigma \, \text{grad} \, u \cdot \mathbf{n} = -h$ on $\partial \Omega$

What about mixed boundary conditions: $\partial \Omega = \Gamma_D \cup \Gamma_N$, $\Gamma_D \neq \emptyset$, $\Gamma_N \neq \emptyset$

Use $V = H^1_{\Gamma_D}(\Omega) := \{ v : \Omega \to \mathbb{R} : \| v \|_A < \infty, v|_{\Gamma_D} = 0 \}$

Generalization of 1st Poincaré-Friedrichs inequality, Thm. 1.7.9:

**Theorem 1.8.2** (General Poincaré-Friedrichs inequality). There is $C > 0$ depending on $\Omega$, $\Gamma_D$ such that

$$\| u \|_0 \leq C \| \text{grad} \, u \|_0 \quad \forall u \in H^1_{\Gamma_D}(\Omega).$$

What about pure convective cooling conditions ($\rightarrow$ Ex. 1.3.1) $\mathbf{j} \cdot \mathbf{n} = q(u - u_0)$ in (1.6.6) ?
Variational formulation: seek \( u \in H^1(\Omega) \) such that
\[
\int_\Omega \sigma(x) \nabla u \cdot \nabla v \, dx + \int_{\partial \Omega} q u v \, dS = \int_\Omega f v \, dx + \int_{\partial \Omega} q u_0 v \, dS \quad \forall v \in H^1(\Omega) .
\]

Trace theorem 1.8.1
Cauchy-Schwarz inequality (1.7.5)
continuity of bilinear form/source functional

\( H^1(\Omega) \)-ellipticity of bilinear form \( \rightarrow \) exercises

### 1.9 The Dirichlet principle

Linear variational problem

\[ u \in V: \quad a(u, v) = f(v) \quad \forall v \in V , \tag{1.7.1} \]

\( V = \) Hilbert space with norm \( \| \cdot \|_V \), \( a(\cdot, \cdot) \) continuous bilinear form, \( f \) continuous linear form.

Assumption: \( V \)-ellipticity of \( a(\cdot, \cdot) \), see Def. 1.7.7:

\[ \exists \gamma > 0: \quad a(u, u) \geq \gamma \| u \|_V^2 \quad \forall u \in V . \tag{1.9.1} \]
Assumption: $a(\cdot, \cdot)$ symmetric: $a(u, v) = a(v, u) \quad \forall u, v \in V$

Theorem 1.9.1 (Dirichlet principle). Assuming (1.9.1), $u \in V$ solves (1.7.1), if and only if it is the unique solution of

$$u = \arg \min_{v \in V} J(v) \quad , \quad J(v) := \frac{1}{2} a(v, v) - f(v).$$

Proof. Lax-Milgram lemma Thm. 1.7.8 $\Rightarrow$ existence & uniqueness of solution $u$ of (1.7.1). Then for all $v \in V$, using $a(u, v) = f(v)$,

$$J(v) - J(u) = \frac{1}{2} (a(v, v) - a(u, u)) - f(v - u)$$
$$= \frac{1}{2} a(v, v) - \frac{1}{2} a(u, u) - a(u, v - u) \quad (1.9.2)$$
$$= \frac{1}{2} a(v - u, v - u) \geq \frac{1}{2} \gamma \|u - v\|_V^2.$$

$\Rightarrow \quad J(v) > J(u) \quad \text{and} \quad (J(v) = J(u) \iff u = v).$

Conversely, (1.9.2) $\Rightarrow \quad u$ is unique solution of minimization problem for $J$. $\square$
The Dirichlet principle

The variational formulations (1.8.2), (1.8.3) of linear scalar second-order elliptic boundary value problems are equivalent to minimization problems for quadratic functionals (also known as Dirichlet forms or energy functionals).

Example 1.9.1 (Quadratic functionals).

Analogy → parabola:

\[
J(v) = \frac{1}{2} a(v, v) - f(v)
\]

\[
f(x) = ax^2 + bx
\]

quadratic functional \( \mathbb{R}^2 \leftrightarrow \mathbb{R} \) \( \triangleright \)
The Finite Element Method (FEM)

Problem: scalar second-order elliptic boundary value problem
Perspective: variational interpretation in Sobolev spaces
Objective: algorithm for the computation of an approximate numerical solution

2.1 Fundamentals

Moot point: any computer can only handle a finite amount of information (reals)

Variational boundary value problem \[\xrightarrow{\text{DISCRETIZATION}}\] System of a finite number of equations for real unknowns
Abstract discussion: start from linear variational problem (see Sect. 1.7, (1.7.1))

\[
 u \in V: \quad a(u, v) = f(v) \quad \forall v \in V ,
\]

(1.7.1)

\[ V = \text{Hilbert space with norm } \| \cdot \|_V, \; a(\cdot, \cdot) \text{ continuous bilinear form, } f \text{ continuous linear form.} \]

Norm of \(a(\cdot, \cdot):\)

\[
 C_A := \sup_{v \in V \setminus \{0\}} \sup_{u \in V \setminus \{0\}} \frac{|a(u, v)|}{\|u\|_V \|v\|_V} < \infty
\]

Assumption: \(V\)-ellipticity of \(a(\cdot, \cdot),\) see Def. 1.7.7:

\[
 \exists \gamma > 0: \quad |a(u, u)| \geq \gamma \|u\|^2_V \quad \forall u \in V .
\]

(1.9.1)

Remark. If \(a(\cdot, \cdot)\) symmetric (\(\rightarrow\) inner product, see Def. 1.7.3) and \(\|\cdot\|_V = \text{energy norm } \|\cdot\|_A\)

\[
 \gamma, \; C_A = 1
\]

Idea of Galerkin discretization

Replace \(V\) in (1.7.1) with a finite dimensional subspace \(V_N\) (discrete trial/test space).
Notation: $N$ = formal index, tagging “discrete entities” (→ “finite amount of information”)

Discrete variational problem

$$u_N \in V_N: \quad a(u_N, v_N) = f(v_N) \quad \forall v_N \in V_N.$$  (2.1.1)

Lax-Milgram Lemma Thm. [1.7.8] ⇒ Existence & Uniqueness of solution $u_N \in V_N$, stability

$$\|u_N\|_V \leq \frac{1}{\gamma} \sup_{v_N \in V_N \setminus \{0\}} \frac{|f(v_N)|}{\|v_N\|_V}.$$  

Issues:

1. How “accurate” is the Galerkin solution $u_N$?
   (a) What measure for accuracy?
   (b) How to assess accuracy?

2. How to convert (2.1.1) into (linear) system of equations?

Ad 1(a): Focus on norm $\|\cdot\|_V$ (and $\|\cdot\|_A$, if $a(\cdot, \cdot)$ inner product)
Galerkin orthogonality

\[ a(u - u_N, v_N) = 0 \quad \forall v_N \in V_N . \quad (2.1.2) \]

[Geometric meaning for inner product \( a(\cdot, \cdot) \rightarrow \)]

Discretization error \( e_N := u - u_N \) "\( a(\cdot, \cdot)\)-orthogonal" to discrete [trial | test space] \( V_N \)

Remark 2.1.1. If \( a(\cdot, \cdot) \) is inner product on \( V \):

\[ \| u - u_N \|^2_A = \| u \|^2_A - \| u_N \|^2_A . \quad (2.1.3) \]

Theorem 2.1.1 (Cea’s lemma). If \( a(\cdot, \cdot) = \text{continuous}, \ V \text{-elliptic, bilinear form}, \ V_N \subset V \text{ finite dimensional subspace}, \ u \in V / u_N \in V_N \) solve \( (1.7.1)/ (2.1.1) \), then

\[ \| u - u_N \|_V \leq \frac{C_A}{\gamma} \inf_{v_N \in V_N \setminus \{0\}} \| u - v_N \|_V \]
Proof. By \textbf{Galerkin-orthogonality (2.1.2)}, for all $v_N \in V_N$

\[
\gamma \|u - u_N\|_V^2 \leq |a(u - u_N, u - u_N) + a(u - u_N, u_N - v_N)| \\
= |a(u - u_N, u - v_N)| \leq C_A \|u - u_N\|_V \|u - v_N\|_V .
\]

\[
\Rightarrow \|u - u_N\|_V \leq \frac{C_A}{\gamma} \inf_{v_N \in V_N \setminus \{0\}} \|u - v_N\|_V ,
\]

because $v_N$ arbitrary. \hfill \qed

---

**Quasi-optimality of Galerkin solutions:** with $C > 0$ \textit{independent} of $u$, $V_N$

\[
\|u - u_N\|_V \leq C \inf_{v_N \in V_N} \|u - v_N\|_V ,
\]

\[\uparrow\quad \uparrow\]

(norm of) discretization error \hspace{3cm} best approximation error

To assess accuracy of Galerkin solution: study capability of $V_N$ to approximate $u$!

---

"Monotonicity" of best approximation

Trial test spaces $V_N, V'_N \subset V$: $V_N \subset V'_N \Rightarrow \inf_{u_N \in V_N} \|u - u'_N\|_V \leq \inf_{u_N \in V_N} \|u - u_N\|_V .

Enhance accuracy by enlarging ("refining") trial space.
**Reminder:**

**Definition 2.1.2 (Linear operator).** Let $V$, $W$ be real vector spaces. A mapping $T : V \mapsto W$ is a linear operator, if

$$T(\alpha u + \beta v) = \alpha T(u) + \beta T(v) \quad \forall u, v \in V, \forall \alpha, \beta \in \mathbb{R}.$$ 

**Reminder:**

**Definition 2.1.3 (projection).** A linear operator $P : V \mapsto V$ on a vector space $V$ is a projection, if $P^2 = P$.

**Definition 2.1.4 (Galerkin projection).** Under the assumptions of Cea's lemma Thm. 2.1.1 the Galerkin projection $P_N : V \mapsto V_N \subseteq V$ is defined by

$$a(P_N u, v_N) = a(u, v_N) \quad \forall v_N \in V_N.$$  

[Lax-Milgram Lemma Thm 1.7.8 $\Rightarrow$ $P_N$ well defined and continuous]
2.1.2 The (linear) algebraic setting

[Now we tackle issue 2. (conversion of (2.1.1) into system of equations)]

I. Introduce (ordered) basis $\mathcal{B}_N$ of $V_N$:

$$\mathcal{B}_N := \{b_N^1, \ldots, b_N^N\} \subset V_N, \quad V_N = \text{Span}\{\mathcal{B}_N\}, \quad N := \dim(V_N).$$

II. Basis representations

$$u_N = \mu_1 b_N^1 + \cdots + \mu_N b_N^N, \quad \mu_i \in \mathbb{R}$$

$$v_N = \nu_1 b_N^1 + \cdots + \nu_N b_N^N, \quad \nu_i \in \mathbb{R}$$

in (2.1.1).

(2.1.1): $a(u_N, v_N) = f(v_N) \quad \forall v_N \in V_N$
\[ a(\mu_1 b_N^1 + \cdots + \mu_N b_N^N, \nu_1 b_N^1 + \cdots + \nu_N b_N^N) = f(\nu_1 b_N^1 + \cdots + \nu_N b_N^N) \quad \forall \nu_1, \ldots, \nu_N \in \mathbb{R} , \]

\[
\sum_{k=1}^{N} \sum_{j=1}^{N} \mu_k \nu_j a(b_N^k, b_N^j) = \sum_{j=1}^{N} \nu_j f(b_N^j) \quad \forall \nu_1, \ldots, \nu_N \in \mathbb{R} ,
\]

\[
\sum_{k=1}^{N} \mu_k a(b_N^k, b_N^j) = f(b_N^j) \quad \text{for} \ j = 1, \ldots, N .
\]

\[ \mathbf{A} \tilde{\mu} = \tilde{\varphi} , \quad \mathbf{A} = \left( a(b_N^k, b_N^j) \right)_{j,k=1}^{N} \in \mathbb{R}^{N,N}, \quad \tilde{\varphi} = \left( f(b_N^j) \right)_{j=1}^{N} , \]

\[ \tilde{\mu} = (\mu_1, \ldots, \mu_N)^T \in \mathbb{R}^N \]
Discrete variational problem

\[ u_N \in V_N : \quad a(u_N, v_N) = f(v_N) \quad \forall v_N \in V_N \]

Choosing basis \( \mathcal{B}_N \)

Linear system of equations

\[ A \tilde{\mu} = \tilde{\varphi} \]

Stiffness matrix:

\[ A = \left( a(b_N^k, b_N^j) \right)_{j,k=1}^N \in \mathbb{R}^{N \times N}, \]

Load vector:

\[ \tilde{\varphi} = \left( f(b_N^j) \right)_{j=1}^N \in \mathbb{R}^N, \]

Coefficient vector:

\[ \tilde{\mu} = (\mu_1, \ldots, \mu_N)^T \in \mathbb{R}^N, \]

Recovery of solution:

\[ u_N = \sum_{k=1}^N \mu_k b_N^k. \]

Corollary 2.1.5.

(2.1.1) has unique solution \( \iff \) \( A \) regular

Impact of choice of basis?

Choice of \( \mathcal{B}_N \) does not affect \( u_N \) \( \Rightarrow \) No impact on discretization error!
Lemma 2.1.6. Consider (2.1.1) and two bases of \( V_N \),

\[
\mathcal{B}_N := \{b^1_N, \ldots, b^N_N\} \quad \text{and} \quad \mathcal{B}_N' := \{b'^1_N, \ldots, b'^N_N\},
\]

related by

\[
b'^j_N = \sum_{k=1}^N s_{jk}b^k_N \quad \text{with} \quad S = (s_{jk})_{j,k=1}^N \in \mathbb{R}^{N,N} \text{ regular.}
\]

Stiffness matrices \( \mathbf{A}, \mathbf{A}' \in \mathbb{R}^{N,N} \), load vectors \( \bar{\mathbf{\varphi}}, \bar{\mathbf{\varphi}}' \in \mathbb{R}^N \), and coefficient vectors \( \bar{\mathbf{\mu}}, \bar{\mathbf{\mu}}' \in \mathbb{R}^N \), respectively, satisfy

\[
\mathbf{A} = SAS^T \quad , \quad \bar{\mathbf{\varphi}} = S\bar{\mathbf{\varphi}}' \quad , \quad \bar{\mathbf{\mu}} = S^{-T}\bar{\mathbf{\mu}}'.
\] (2.1.4)

Proof.

\[
\mathbf{A}_{lm} = a(b^m_N, b^l_N) = \sum_{k=1}^N \sum_{j=1}^N s_{mk}a(b^k_N, b^j_N)s_{lj} = \sum_{k=1}^N \left( \sum_{j=1}^N s_{lj}\mathbf{A}_{jk} \right) s_{mk} = (S\mathbf{A}S^T)_{lm},
\]
Reminder of linear algebra:

**Definition 2.1.7 (Congruent matrices).** Two matrices $A \in \mathbb{R}^{N,N}$, $B \in \mathbb{R}^{N,N}$, $N \in \mathbb{N}$, are called congruent, if there is a regular matrix $S \in \mathbb{R}^{N,N}$ such that $B = SAS^T$.

Equivalence relation on square matrices

**Lemma 2.1.8.** Matrix property invariant under congruence $\Leftrightarrow$ Property of stiffness matrix invariant under change of basis $\mathcal{B}_N$

Matrix properties invariant under congruence

=  

- regularity
- symmetry
- positive definiteness

Reminder:

**Definition 2.1.9 (Positive definite matrix).** Matrix $B \in \mathbb{R}^{N,N}$, $N \in \mathbb{N}$, is positive definite $\Leftrightarrow \tilde{\xi}^T B \tilde{\xi} > 0$ for all $\tilde{\xi} \in \mathbb{R}^N \setminus \{0\}$. 
2.1.3 Principles of FEM

\[ \Omega \subset \mathbb{R}^d, \; d = 2, 3, \text{ bounded computational domain: assumed polygonal } d = 2, \text{ polyhedral } d = 3 \]

First main ingredient: triangulation/mesh of \( \Omega \)

**Definition 2.1.10.** A *mesh (or triangulation)* of \( \Omega \subset \mathbb{R}^d \) is a finite collection \( \{K_i\}_{i=1}^M \), \( M \in \mathbb{N} \), of open non-degenerate polygons \( d = 2 \)/polyhedra \( d = 3 \) such that

1. \( \overline{\Omega} = \bigcup \{K_i, \; i = 1, \ldots, M\} \),
2. \( K_i \cap K_j = \emptyset \iff i \neq j \),
3. for all \( i, j \in \{1, \ldots, M\}, \; i \neq j \), the intersection \( \overline{K}_i \cap \overline{K}_j \) is a vertex, edge, or face of both \( K_i \) and \( K_j \).

“vertex”, ”edge”, ”face” of polygon/polyhedron: \( \rightarrow \) geometric intuition

**Terminology:** Given mesh \( \mathcal{M} := \{K_i\}_{i=1}^M \): \( K_i \) called cell or element.

Vertices of a mesh \( \rightarrow \) nodes \( (\text{set } \mathcal{N}(\mathcal{M})) \)
Types of meshes:

- Triangular mesh in 2D
- Quadrilateral mesh in 2D

If (C) does not hold

- Triangular non-conforming mesh (with hanging nodes)

$\overline{K}_i \cap \overline{K}_j$ is only part of an edge/face for at most one of the adjacent cells.

(However, conforming if degenerate quadrilaterals admitted)

Simplicial mesh = triangular mesh in 2D
tetrahedral mesh in 3D
Second main ingredient: space of piecewise polynomial functions

\[ V_N := \{v \in V: v|_K \in \mathcal{P}_p(K) \ \forall K \in \mathcal{M}\} , \]

\[ \mathcal{P}_p(K) = \text{polynomials of degree } \leq p \text{ on cell } K . \]

Note:

\[ v \in V \rightarrow \text{conformity conditions at interelement boundaries} \]

Lemma 2.1.11 (Conformity condition for \( H^1 \)). Let \( \mathcal{M} := \{K_i\}_{i=1}^M \) be a triangulation (→ Def. 2.1.10) of \( \Omega \subset \mathbb{R}^d \) and assume that \( v : \Omega \mapsto \mathbb{R} \) satisfies that \( v|_K \) can be extended to a function in \( C^\infty(\overline{K}) \) for any \( K \in \mathcal{M} \). Then

\[ v \in H^1(\Omega) \iff v \in C^0(\overline{\Omega}) . \]

Conformity condition for \( H^1 = \) global continuity \( (C^0, \text{not } C^1! \rightarrow \text{Ex. 1.7.1}) \)

(recall physical constraints on temperature distributions!)

Thanks to notion of weak derivative, Sect. 1.5 !
Definition 2.1.12 (Conformity). A $\mathcal{M}$-piecewise polynomial space $V_N$ is called $V$-conforming, if $V_N \subset V$.

Third main ingredient: Locally supported basis functions

Basis functions $b^1_N, \ldots, b^N_N$ for a finite element trial/test space $V_N$ built on a mesh $\mathcal{M}$ satisfy:

- each $b^i_N$ associated with a single cell/edge/face/vertex of $\mathcal{M}$,
- $\text{supp}(b^i_N) = \bigcup\{K : K \in \mathcal{M}, p \subset \overline{K}\}$, if $b^i_N$ associated with cell/edge/face/vertex $p$.

Finite element terminology: $b^i_N = \text{global shape/basis functions}$

Example 2.1.2 (Supports of global shape functions in 1D).
\[ \Omega = ]a, b[ \setminus \hat{=} \text{interval} \]

Equidistant mesh

\[ M := \{ x_{j-1}, x_j, j = 1, \ldots, N \}, \]
\[ x_j := a + hj, \quad h := (b - a)/N, \quad N \in \mathbb{N}. \]

Example 2.1.3 (Supports of global shape functions on triangular mesh).

Support of node-associated basis function

Support of edge-associated basis function

Support of cell-associated basis function

Rationale for small supports?
Recall bilinear form $\leftrightarrow -\Delta$: 

$$ a(u, v) := \int_{\Omega} \text{grad } u \cdot \text{grad } v \, dx $$

Use triangular mesh $\mathcal{M}$, test/trial space $V_N \subset H^1(\Omega)$ with basis $\mathcal{B}_N := \{b_1^N, \ldots, b_N^N\}$ (→ Sect. 2.1.2)

Stiffness matrix $A \in \mathbb{R}^{N \times N}$ with $a_{ij} := a(b_i^N, b_j^N)$, $i, j = 1, \ldots, N$

$b_i^N, b_j^N$ associated with nodes not linked by an edge

$$ a_{ij} = 0 \quad \text{(because } \text{vol}(\text{supp}(b_i^N) \cap \text{supp}(b_j^N)) = 0) $$

Finite element stiffness matrices are sparse.

**Definition 2.1.13** (Sparse matrix). A matrix $A \in \mathbb{R}^{N \times N}$ is called sparse, if $\text{nnz}(A) := \# \{ (i, j) : a_{ij} \neq 0 \} \ll N^2$. 
Example 2.1.4 (Sparse \textbf{stiffness matrices}).

$V_N$: one basis function associated with each vertex

\begin{align*}
0 & \quad 50 & \quad 100 & \quad 150 & \quad 200 & \quad 250 & \quad 300 & \quad 350 & \quad 400 \\
0 & \quad 50 & \quad 100 & \quad 150 & \quad 200 & \quad 250 & \quad 300 & \quad 350 & \quad 400 \\
\end{align*}

\text{Finite element mesh } \mathcal{M}

\text{Resulting sparsity pattern of stiffness matrix}

\text{Visualization of sparsity pattern: MATLAB-}\texttt{spy()}\texttt{()-Funktion}
Remark 2.1.5 (Storing sparse matrices).

Special (efficient) storage formats for sparse matrices, e.g., CRS-format

Special MATLAB commands: \texttt{sparse}, \texttt{spones}, \texttt{speye}, \texttt{spdiags}

\texttt{(!}} \texttt{use \textit{mandatory} !)}

\rightarrow \text{Sect. 2.1.2 “Matrix-Speicherformate”, course “Numerische Mathematik für CSE”}

2.1.4 Linear $H^1$-conforming finite elements

$\mathcal{M} =$ \texttt{simplicial mesh} of polygonal/polyhedral computational domain $\Omega \subset \mathbb{R}^d$, $d = 2, 3$

Linear $H^1$-conforming finite elements

= Simplest $H^1(\Omega)$-conforming finite element space

= Simplest finite element scheme for scalar second order elliptic BVP on $\Omega$
\[ \mathcal{S}_1^0(\mathcal{M}) := \{ v \in C^0(\overline{\Omega}) : v|_K \in \mathcal{P}_1(K) \quad \forall K \in \mathcal{M} \} \subset H^1(\Omega) \]

Representation: \[ \mathcal{P}_1(K) := \{ x \mapsto \alpha + \beta \cdot x, \quad x \in K, \quad \alpha \in \mathbb{R}, \quad \beta \in \mathbb{R}^d \} \]

(space of \( d \)-variate polynomials of total degree \( \leq 1 \))

\[ \dim \mathcal{P}_1(K) = d + 1 \]

Notation:

- \( \mathcal{S}_1^0(\mathcal{M}) \): continuous functions, cf. \( C^0(\Omega) \)
- \( \mathcal{P}_1 \): locally 1st degree polynomials, cf. \( \mathcal{P}_1 \)
Example 2.1.6. \((H^1(\Omega))-\text{conforming linear finite element space in 1D}\)

\(d = 1, \Omega = ]0, 1[,\)

mesh \(\mathcal{M} = \text{partition of } ]0, 1[ \text{ into intervals}

- **red**: function \(\in \mathcal{S}^0_1(\mathcal{M})\)
- **blue**: hat function basis of \(\mathcal{S}^0_1(\mathcal{M})\)

Locally supported basis functions in 2D?

On a triangle \(T\) with vertices \(a^1, a^2, a^3\): \(q \in P_1(T)\) uniquely determined by values \(q(a^i)\).

\[
v_N \in \mathcal{S}^0_1(\mathcal{M}) \text{ uniquely determined by } \{v_N(x), x \text{ node of } \mathcal{M}\}
\]

\[
\dim \mathcal{S}^0_1(\mathcal{M}) = \# \mathcal{V}(\mathcal{M}) \quad (\mathcal{V}(\mathcal{M}) = \text{set of vertices of } \mathcal{M})
\]

If \(\mathcal{V}(\mathcal{M}) = \{x^1, \ldots, x^N\}\), nodal basis \(\mathcal{B}_N := \{b^1_N, \ldots, b^N_N\}\) of \(\mathcal{S}^0_1(\mathcal{M})\) defined by \(b^i_N(x^j) = \delta_{ij}\).
Piecewise linear nodal basis function
("hat function")
(= global shape function for $\delta_1^0(\mathcal{M})$)

Coefficient $\mu_j = \text{"nodal value" of $u_N$ at } j\text{-th}\ \text{node of } \mathcal{M}$

Global shape functions $\xrightarrow{\text{Restriction to element}}$ local shape functions \hspace{1cm} (2.1.5)

Example 2.1.7 (Local shape functions for $\delta_1^0(\mathcal{M})$).

Triangle $K$ with vertices $a^1 = (0,0), a^2 = (1,0), a^3 = (0,1)$:

Local shape functions: $b_1^K(x) = 1 - x_1 - x_2$, $b_2^K(x) = x_1$, $b_3^K(x) = x_2$. 
Local shape functions for $\delta_1^0(\mathcal{M})$ on triangle/tetrahedron = barycentric coordinate functions

**Definition 2.1.14 (Barycentric coordinates).** Given $d + 1$ points $a^1, \ldots, a^{d+1} \in \mathbb{R}^d$ that do not lie in a hyperplane the barycentric coordinates $\lambda_1 = \lambda_1(x), \ldots, \lambda_{d+1} = \lambda_{d+1}(x) \in \mathbb{R}$ of $x \in \mathbb{R}^d$ are uniquely defined by

$$
\lambda_1(x) + \cdots + \lambda_{d+1}(x) = 1, \quad \lambda_1(x)a^1 + \cdots + \lambda_{d+1}(x)a^{d+1} = x \quad \forall x \in \mathbb{R}^d.
$$
Barycentric coordinates obtained by solving

\[
\begin{pmatrix}
    a_1^1 & \cdots & a_1^{d+1} \\
    \vdots & \ddots & \vdots \\
    a_d^1 & \cdots & a_d^{d+1} \\
    1 & \cdots & 1
\end{pmatrix}
\begin{pmatrix}
    \lambda_1 \\
    \vdots \\
    \lambda_d \\
    \lambda_{d+1}
\end{pmatrix}
= 
\begin{pmatrix}
    x_1 \\
    \vdots \\
    x_d \\
    1
\end{pmatrix}.
\]

(2.1.6)

Corollary 2.1.15. Given \( d + 1 \) points \( \mathbf{a}^1, \ldots, \mathbf{a}^{d+1} \in \mathbb{R}^d \) as in Def. 2.1.14, the barycentric coordinates are affine linear functions on \( \mathbb{R}^d \), which satisfy

\[
\lambda_j(\mathbf{a}^i) = \delta_{ij} := \begin{cases} 
1, & \text{if } i = j, \\
0, & \text{else ,}
\end{cases} \quad 1 \leq i, j \leq d + 1.
\]
Barycentric coordinate functions (= Local shape functions for $S^0(\mathcal{M})$) on a triangle

How to get $H^1_0(\Omega)$-conforming finite element space $S^0_{1,0}(\mathcal{M}) := S^0(\mathcal{M}) \cap H^1_0(\Omega)$?

- Discard nodal basis functions associated with vertices on $\partial \Omega$!

**Remark 2.1.8.** Piecewise linear finite element subspace of $H^1_0(\Omega)$?

- There exist no locally supported piecewise linear basis functions.
2.1.5 Simplicial Lagrangian finite elements

\[ \mathcal{M} = \text{simplicial mesh} \text{ of polygonal/polyhedral computational domain } \Omega \subset \mathbb{R}^d, \quad d = 2, 3 \]

Idea: Use higher degree polynomials \( \rightarrow \) “better accuracy” (cf. interpolation)

Higher degree polynomials

\[ \mathcal{P}_p(\mathbb{R}^d) := \{ x \in \mathbb{R}^d \mapsto \sum_{\alpha \in \mathbb{N}_0^d, \, |\alpha| \leq p} \kappa_{\alpha} x^\alpha, \quad \kappa_{\alpha} \in \mathbb{R} \} \]

Notation: \( \alpha = \text{“multiindex”} (\alpha_1, \ldots, \alpha_d), \quad |\alpha| = \alpha_1 + \cdots + \alpha_d, \quad x^\alpha := x_1^{\alpha_1} \cdots x_d^{\alpha_d} \).

Example:

\[ \mathcal{P}_2(\mathbb{R}^2) = \text{Span} \{ 1, x_1, x_2, x_1^2, x_2^2, x_1 x_2 \} \]

Lemma 2.1.16.

\[ \dim \mathcal{P}_p(\mathbb{R}^d) = \binom{d + p}{p} \quad \text{for all } p \in \mathbb{N}, \, d \in \mathbb{N} \]

Definition 2.1.17 (Higher order Lagrangian finite element spaces). *Space of \( p \)-th degree Lagrangian finite element functions on mesh \( \mathcal{M} \)

\[ \delta^0_p(\mathcal{M}) := \{ v \in C^0(\Omega) : v|_K \in \mathcal{P}_p(K) \quad \forall K \in \mathcal{M} \} . \]
Notation:

$S^0_p(M)$

continuous functions, cf. $C^0(\Omega)$

locally polynomials of degree $p$, cf. $P_p(\mathbb{R}^d)$

Construction:

Local shape functions

“Glueing”

Global FE space

(Glueing must ensure global continuity $\leftrightarrow H^1(\Omega)$-conformity)

Example 2.1.9 (Quadratic Lagrangian finite elements).

Design of local shape functions must make glueing possible

Local shape functions for $P_2(K)$, $K$ triangle:
\[ b^K_i = -\lambda_i (1 - 2\lambda_i), \quad b^K_{12} = 4\lambda_1 \lambda_2, \]
\[ b^K_2 = -\lambda_2 (1 - 2\lambda_2), \quad b^K_{13} = 4\lambda_1 \lambda_3, \]
\[ b^K_3 = -\lambda_3 (1 - 2\lambda_3), \quad b^K_{23} = 4\lambda_2 \lambda_3. \]

\[ \lambda_i = \text{barycentric coordinate function} \] (\text{Def. 2.1.14}) for vertex \( a_i \)

\[ b^K_j(a_i) = \delta_{ij}, \quad i, j \in \{1, 2, 3, (12), (23), (13)\}. \]

Local shape functions = Lagrangian (interpolatory) polynomials for local nodes in \( K \)

Specifying local interpolation nodes \( \Leftrightarrow \) specifying local shape functions
When are local nodes $q_i, i = 1, \ldots, Q$, (for $P_p(K)$) suitable for “glueing”? 

- **Unisolvence:**
  \[ v \in P_p(K) : \quad v(q_i) = 0 \quad \forall i \quad \Leftrightarrow \quad v \equiv 0 \]

- **Fixing traces:**
  locally unisolvent interpolation on each vertex/edge/face

Invalid choice of local nodes for $S_0^0(M)$

- **Interelement matching:**
  corresponding nodes on joint edges/faces
Matching nodes for quadratic Lagrangian finite elements

“Glueing”: edge-associated local and resulting global shape function for $\mathcal{B}_2^0(\mathcal{M})$, $\mathcal{M}$ triangular
Location of local (interpolation) nodes for triangular Lagrangian finite elements of degree 2 (left), degree 3 (middle), and degree 4 (right)

Local nodes for tetrahedral Lagrangian finite elements (left: $p = 2$, middle: $p = 3$, right: $p = 4$)

Can we find other locally supported bases for $\mathcal{B}_2^0(\mathcal{M})$? **YES!**
Alternative $p$-hierarchical local shape functions:

\[ b^K_1 = \lambda_1, \quad b^K_{12} = 4\lambda_1\lambda_2, \]
\[ b^K_2 = \lambda_2, \quad b^K_{13} = 4\lambda_1\lambda_3, \]
\[ b^K_3 = \lambda_3, \quad b^K_{23} = 4\lambda_2\lambda_3. \]

Set comprises local shape functions for $p = 1$.

Glueing can easily be accomplished

No “canonical” local shape functions/global basis functions for higher order Lagrangian finite elements.

Selection of $\mathcal{B}_N$ to get “good” matrix properties.

### 2.1.6 Parametric finite elements

**Definition 2.1.18** (Affine transformation). A mapping $\Phi : \mathbb{R}^d \mapsto \mathbb{R}^d$ is affine if $\Phi(x) = Fx + \tau$ with $F \in \mathbb{R}^{d \times d}$, $\tau \in \mathbb{R}^d$. 
Usually:

All elements of a mesh = affine images of reference element(s) \( \hat{K} \)

\[ \exists \hat{K} \quad \forall K \in \mathcal{M} : \exists F_K \in \mathbb{R}^{d,d} \text{ regular, } \tau_K \in \mathbb{R}^d : \quad K = \Phi_K(\hat{K}) \quad \text{with} \quad \Phi_K(\hat{x}) := F_K \hat{x} + \tau_K . \]

“Unit triangle”: \( \hat{K} = \{(0,0), (1,0), (0,1)\} \)

For \( K = \text{convex } \{a^1, a^2, a^3\} : \)

\[ F_K = \begin{pmatrix} a^2 - a^1 & a^3 - a^1 \\ a^2 - a^2 & a^3 - a^2 \end{pmatrix}, \quad \tau_K = a^1 . \]

Transformations of elements \( \Rightarrow \) transformation of functions:

**Definition 2.1.19 (Pullback).** Given domains \( \Omega, \hat{\Omega} \) and a bijective mapping \( \Phi : \hat{\Omega} \mapsto \Omega \), the pullback \( \Phi^*u : \hat{\Omega} \mapsto \mathbb{R} \) of a function \( u : \Omega \mapsto \mathbb{R} \) is defined by

\[ (\Phi^*u)(\hat{x}) := u(\Phi(\hat{x})) , \quad \hat{x} \in \hat{\Omega} . \]
Notation: If unambiguous: \( \widehat{u} := \Phi^* u \)

Note: Consider \( \delta_1^0(\mathcal{M}) \), triangle \( K \in \mathcal{M} \), unit triangle \( \widehat{K} \), affine mapping \( \Phi_K : \widehat{K} \mapsto K \)

- \( b_1^K, b_2^K, b_3^K \) (standard) local shape functions on \( K \)
- \( \widehat{b}_1, \widehat{b}_2, \widehat{b}_3 \) (standard) local shape functions on \( \widehat{K} \)

Observation: \( \widehat{b}_i = \Phi^*_K b_i^K \)

Terminology: affine equivalent finite elements

All families of Lagrangian finite elements (Sect. 2.1.5) (equipped with “natural” local shape functions) are affine equivalent.

STEP 1: define local shape functions on reference element \( \widehat{K} \)

STEP 2: local shape functions on \( K \in \mathcal{M} \) via pullback \( (\Phi^{-1})^* \) (\( \rightarrow \) Def. 2.1.19)

Parametric finite elements

Generalization: curvilinear meshes with curved edges/faces (\( \rightarrow \) Sect. 2.2.8)
Now: \( \Phi_K \) diffeomorphism!

\[ b^K_i := (\Phi_K^{-1}) \hat{b}_i. \]

Application: approximation of curved interfaces/boundaries (→ Sect 2.2.8)

2.1.7 Lagrangian finite elements on quadrilaterals/hexahedra

Parametric construction: start from reference element \( \hat{K} = ]0, 1[^d \) (unit cube)

Lowest polynomial degree \( p = 1, 2D \): piecewise bilinear finite elements
Local shape functions on $\mathcal{K} = ]0, 1[^2$

\[
\begin{align*}
\hat{b}_1 &= (1 - \hat{x}_1)(1 - \hat{x}_2), \\
\hat{b}_2 &= \hat{x}_1(1 - \hat{x}_2), \\
\hat{b}_3 &= \hat{x}_1\hat{x}_2, \\
\hat{b}_4 &= (1 - \hat{x}_1)\hat{x}_2.
\end{align*}
\]

Bilinear Lagrangian interpolation polynomials w.r.t. corner points $a_i$ of $\mathcal{K}$

Note: $\hat{b}_i$ linear on edges
Bilinear mapping to general quadrilateral:

\[ \Phi_K(\vec{x}) = \left( \begin{array}{c} \alpha_1 + \beta_1 \hat{x}_1 + \gamma_1 \hat{x}_2 + \delta_1 \hat{x}_1 \hat{x}_2 \\ \alpha_2 + \beta_2 \hat{x}_2 + \gamma_2 \hat{x}_2 + \delta_2 \hat{x}_1 \hat{x}_2 \end{array} \right), \quad \alpha_i, \beta_i, \gamma_i, \delta_i \in \mathbb{R}. \]

Example 2.1.10 (Triangle as degenerate quadrilateral).
Higher order quadrilateral Lagrangian finite elements:

**Definition 2.1.20 (Tensor product polynomials).** Space of tensor product polynomials of degree \( p \in \mathbb{N} \) in each coordinate direction

\[
\mathcal{Q}_p(\mathbb{R}^d) := \{ \mathbf{x} \mapsto p_1(x_1) \cdots p_d(x_d), \ p_i \in \mathcal{P}_p(\mathbb{R}), \ i = 1, \ldots, d \}.
\]

Local trial space \( \mathcal{Q}_p(\hat{K}) \) on \( \hat{K} = ]0, 1[^2 \) + parametric construction
2.1.8 Degrees of freedom

Recall: Lagrangian local shape functions $b^K_i$ fixed by $b^K_i(q_j) = \delta_{ij}$ for nodes $q_i, i, j = 1, \ldots, Q$, $Q \in \mathbb{N}$ (→ Sects. 2.1.5, 2.1.7).

**Definition 2.1.21** (Dual basis). *Given a vector space $V$ with basis $\mathcal{B} := \{b_1, \ldots, b_Q\}, Q \in \mathbb{N}$, the corresponding dual basis is a set $l_1, \ldots, l_Q$ of linear forms on $V$ such that*

\[
l_j(b_i) = \delta_{ij}, \quad i, j \in \{1, \ldots, Q\}.
\]
For Lagrangian finite elements $\mathcal{S}_p^0(M)$, on element $K$:

Nodal evaluation functionals $v \mapsto v(q_j), \ j = 1, \ldots, Q$, form dual basis w.r.t. basis $\{b^K_1, \ldots, b^K_Q\}$ of local trial space $\mathcal{S}_p(K)$.

**Definition 2.1.22** (Local degrees of freedom). A dual basis of the local trial space corresponding to the local shape functions provides local degrees of freedom (d.o.f.).

Role reversal: degrees of freedom $\Rightarrow$ local shape functions

**Example 2.1.11** (Cubic Hermitian Finite Elements on triangular mesh).

- **Local trial space** $V_K := \mathcal{P}_3(K)$ for each $K \in \mathcal{M}$,

- **Local degrees of freedom**, $K$ triangle with vertices $a^1, a^2, a^3 \in \mathbb{R}^2$

  \[
  l_1(v) = v(a^1), \quad l_2(v) = v(a^2), \quad l_3(v) = v(a^3),
  \]

  \[
  l_4(v) = \text{grad} v(a^1) \cdot (a^2 - a^1), \quad l_5(v) = \text{grad} v(a^2) \cdot (a^3 - a^2), \quad l_6(v) = \text{grad} v(a^3) \cdot (a^1 - a^3),
  \]

  \[
  l_7(v) = \text{grad} v(a^1) \cdot (a^3 - a^1), \quad l_8(v) = \text{grad} v(a^2) \cdot (a^1 - a^2), \quad l_9(v) = \text{grad} v(a^3) \cdot (a^2 - a^3),
  \]

  \[
  l_{10}(v) = v(\frac{1}{3}(a^1 + a^2 + a^3)).
  \]
Dual basis (→ Def. 2.1.21)?

- # functionals = \(\dim V_K = \dim \mathcal{P}_3(\mathbb{R}^2) = \binom{3+2}{2} = 10\),

- If \(l_j(v) = 0\) for all \(j = 1, \ldots, 10\), then \(\Rightarrow v(a_i) = 0\) and \(\nabla v(a_i) = 0\), \(i = 1, 2, 3\),
  \(\Rightarrow v \in \mathcal{P}_3(K) \equiv 0\) on any edge,
  \(\Rightarrow (= 0 \text{ at center of gravity}) v \equiv 0\) for any \(v \in V_K\).

**Unisolvence of local d.o.f.**

Suitable for glueing?

YES, because \(v_{\mid \text{edge}}\) uniquely determined by d.o.f. associated with the edge.

A **local degree of freedom** \(l\) is regarded as associated with an edge \(E\), if \(l(v)\) only depends on \(v_{\mid E}, \nabla v_{\mid E}, \ldots\).

Symbolic notation for local d.o.f. for cubic Hermitian elements:

(filled circle = nodal values, circle = first derivatives, arrows = directional derivatives)

Fig. 29 p. 90
HOWEVER, alternative choice of local degrees of freedom possible
(on triangle $K$ with vertices $a^1, a^2, a^3 \in \mathbb{R}^2$)

\begin{align*}
l_1(v) &= v(a^1), \\
l_4(v) &= \frac{\partial v}{\partial x_1}(a^1), \\
l_7(v) &= \frac{\partial v}{\partial x_2}(a^1), \\
l_{10}(v) &= v(a^{123}).
\end{align*}

\begin{align*}
l_2(v) &= v(a^2), \\
l_5(v) &= \frac{\partial v}{\partial x_1}(a^2), \\
l_8(v) &= \frac{\partial v}{\partial x_2}(a^2),
\end{align*}

\begin{align*}
l_3(v) &= v(a^3), \\
l_6(v) &= \frac{\partial v}{\partial x_1}(a^3), \\
l_9(v) &= \frac{\partial v}{\partial x_2}(a^3),
\end{align*}

Three d.o.f. associated with each vertex

Fewer global shape functions compared to previous choice!

Fig. 30
2.2 Finite Element Implementation

2.2.1 Mesh file format

Data flow in (most) finite element software packages:

Parameters

Mesh generator → Finite element solver (computational kernel) → Post-processor (e.g. visualization)

Example 2.2.1 (Mesh file format (triangular mesh of polygonal domain)).
Two-dimensional simplicial mesh

1 \( \xi_1 \) \( \eta_1 \)  
# Coordinates of first node

2 \( \xi_2 \) \( \eta_2 \)  
# Coordinates of second node

\[
N \quad \xi_N \quad \eta_N \]  
# Coordinates of \( N \)-th node  

1 \( n^1_1 \) \( n^1_2 \) \( n^1_3 \) \( X_1 \)  
# Indices of nodes of first triangle

2 \( n^2_1 \) \( n^2_2 \) \( n^2_3 \) \( X_2 \)  
# Indices of nodes of second triangle

\[
M \quad n^M_1 \quad n^M_2 \quad n^M_3 \quad X_M \]  
# Indices of nodes of \( M \)-th triangle

\( X_i, i = 1, \ldots, M \rightarrow \) extra information (e.g. material properties in triangle \#i).

Optional: additional information about edges (on \( \partial\Omega \)):

\[
K \in \mathbb{N} \]  
# Number of edges on \( \partial\Omega \)

1 \( n^1_1 \) \( n^1_2 \) \( Y_1 \)  
# Indices of endpoints of first edge

2 \( n^2_1 \) \( n^2_2 \) \( Y_2 \)  
# Indices of endpoints of second edge

\[
K \quad n^K_1 \quad n^K_2 \quad Y_K \]  
# Indices of endpoints of \( K \)-th edge
Example 2.2.2 (Mesh file format for MATLAB code).

Vertex coordinate file:

% List of vertices
1 +0.000000e+00 -1.000000e+00
2 +1.000000e+00 +0.000000e+00
3 +0.000000e+00 +1.000000e+00
4 -1.000000e+00 +0.000000e+00
5 +0.000000e+00 +0.000000e+00

Cell information file:

% List of elements
1 1 2 5
2 2 3 5
3 3 4 5
4 4 1 5
Loading a mesh

```matlab
m = load_Mesh('Coord_Circ.dat', ...
    'Elem_Circ.dat');
plot_Mesh(m, 'apts');
```

Option flags:

- `'a'`: with axes
- `'p'`: vertex labels on
- `'t'`: cell labels on
- `'s'`: caption/title on

How to create a mesh?
Mesh generation (beyond scope of this course)

http://www.andrew.cmu.edu/user/sowen/mesh.html

Free software:
- NETGEN (http://www.hpfem.jku.at/netgen/)
- Triangle (http://www.cs.cmu.edu/~quake/triangle.html)
- TETGEN (http://tetgen.berlios.de/index.html)

Example 2.2.3 (Mesh generation in MATLAB code).

Algorithm & details → [9]

MATLAB-CODE: mesh generation for circular domain

```matlab
BBOX = [-1 -1; 1 1];
H0 = 0.1;
DHD = @(x) sqrt(x(:,1).^2+x(:,2).^2)-1;
HHANDLE = @(x) ones(size(x,1),1);
Mesh = init_Mesh(BBOX,H0,DHD,...
                  HHANDLE,[],1);
save_Mesh(Mesh,'Coordinates.dat',...
          'Elements.dat');
```

Bounding box

Largest reasonable edge length

Signed distance function $\varphi(x)$:
(distance from $\partial \Omega$, $\varphi(x) < 0 \iff x \in \Omega$)

Element size function
(determines local edge length)
2.2.2 Assembly

→ term used for computing entries of stiffness matrix/load vector.

Discrete variational problem \((V_N = \text{FE space}, \dim V_N = N \in \mathbb{N}, \text{see Sect. 2.1.1})\)

\[ u_N \in V_N: \quad a(u_N, v_N) = f(v_N) \quad \forall v_N \in V_N. \]  

(2.1.1)

To be computed:

- **Stiffness matrix:**

\[ A = \left( a(b_N^j, b_N^i) \right)_{i,j=1}^N \in \mathbb{R}^{N,N} \]

- **Load vector:**

\[ \bar{\phi} := \left( f(b_N^i) \right)_{i=1}^N \in \mathbb{R}^N \]

both can be written in terms of cell contributions:

\[ a(u, v) = \sum_{K \in \mathcal{M}} a_K(u|_K, v|_K), \quad f(v) = \sum_{K \in \mathcal{M}} f_K(v|_K). \]  

(2.2.3)
Example: bilinear form/linear form arising from 2nd-order elliptic BVPs (→ Sect. 1.6)

\[
a(u, v) := \int_{\Omega} \sigma \, \text{grad} \, u \cdot \text{grad} \, v \, dx = \sum_{K \in \mathcal{M}} \int_{K} \sigma \, \text{grad} \, u \cdot \text{grad} \, v \, dx, \\

f(v) := \int_{\Omega} f \, v \, dx = \sum_{K \in \mathcal{M}} \int_{K} f \, v \, dx.
\]

Recall (2.1.5): Restrictions of \underline{global} shape functions to cells = \underline{local} shape functions

**Definition 2.2.1.** Given \underline{local shape functions} \(\{b^K_1, \ldots, b^K_Q\}\), we call

- element stiffness matrix \(\Lambda_K := \left(a_K(b^K_j, b^K_i)\right)_{i,j=1}^{Q} \in \mathbb{R}^{Q \times Q}\),
- element load vector \(\vec{\phi}_K := \left(f_K(b^K_i)\right)_{i=1}^{Q} \in \mathbb{R}^{Q}\).
Theorem 2.2.2. The **stiffness matrix** and **load vector** can be obtained from their cell counterparts by

\[
A = \sum_K T_K^T A_K T_K, \quad \bar{\varphi} = \sum_K T_K^T \bar{\varphi}_K,
\]

with the **index mapping matrices** ("T-matrices") \( T_K \in \mathbb{R}^{Q,N} \), defined by

\[
(T_K)_{ij} := \begin{cases} 
1, & \text{if } (b_N^j)|_K = b_i^K, \\
0, & \text{otherwise.}
\end{cases}
\]

Proof.

\[
(A)_{ij} = a(b_N^j, b_N^i) = \sum_{K \in \mathcal{M}} a_K (b_N^j|_K, b_N^i|_K) = \sum_{K \in \mathcal{M}, \text{supp}(b_N^j) \cap K \neq \emptyset} a_K (b_N^j|_K, b_N^i|_K) = \sum_{K \in \mathcal{M}, \text{supp}(b_N^j) \cap K \neq \emptyset} (A_K)_{l(i),l(j)}.
\]

\( l(i) \in \{1, \ldots, k_K\}, \ 1 \leq i \leq N \) index of the local shape function corresponding to the global shape function \( b_N^i \) on \( K \).

\[
\Rightarrow \quad (A)_{ij} = \sum_{K \in \mathcal{M}, \text{supp}(b_N^j) \cap K \neq \emptyset} \sum_{l=1}^Q \sum_{n=1}^Q (T_K)_{li} (A_K)_{ln} (T_K)_{nj}.
\]
Example 2.2.4 (Assembly for linear Lagrangian finite elements on triangular mesh).

Using the local/global numbering indicated beside

\[ \mathbf{T}_{K^*} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} \]

2.2.3 Mesh data structures
Bare minimum of information a mesh data structure has to provide is

1. a unique identification and the geometric location of global basis functions,
2. a possibility to traverse the local shape functions/degrees of freedom of every cell,
3. a way to run through the edges/faces of a cell in predefined order,
4. and a method for iterating through all cells of the mesh (→ global numbering)

Focus: array oriented data layout (→ MATLAB, FORTRAN)

Case: $d$-dimensional simplicial triangulation $\mathcal{M}$, minimal data structure (cf. Sect. 2.2.1)

→ Coordinates of vertices $\mathcal{N}(\mathcal{M})$: $\#\mathcal{N}(\mathcal{M}) \times d$-array Coordinates of double

→ Vertex indices for cells: $\#\mathcal{M} \times (d + 1)$-array Elements of int.
Example 2.2.5 (Arrays storing 2D triangular mesh).

<table>
<thead>
<tr>
<th>i</th>
<th>Coordinates</th>
<th>( K_j )</th>
<th>Vertex indices</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 -1 1</td>
<td>1</td>
<td>1 2 9</td>
</tr>
<tr>
<td>2</td>
<td>1 -1 0</td>
<td>2</td>
<td>2 5 9</td>
</tr>
<tr>
<td>3</td>
<td>1 -1 -1</td>
<td>3</td>
<td>5 8 9</td>
</tr>
<tr>
<td>4</td>
<td>0 0 -1</td>
<td>4</td>
<td>5 7 8</td>
</tr>
<tr>
<td>5</td>
<td>0 0 0</td>
<td>5</td>
<td>3 4 2</td>
</tr>
<tr>
<td>6</td>
<td>1 -1 -1</td>
<td>6</td>
<td>4 5 2</td>
</tr>
<tr>
<td>7</td>
<td>1 0 -1</td>
<td>7</td>
<td>4 7 5</td>
</tr>
<tr>
<td>8</td>
<td>1 1 -1</td>
<td>8</td>
<td>4 6 7</td>
</tr>
<tr>
<td>9</td>
<td>0 1 -1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Array Coordinates

Array Elements

Global shape functions associated with edges/faces \( \Rightarrow \) extra information required!

Example 2.2.6 (Extended MATLAB mesh data structure).

```
mesh = add_Edge2Elem(add_Edges(init_Mesh(BBOX,H0,DHD,HHANDLE,[],1)))
```

(init_Mesh \(\rightarrow\) Ex. 2.2.3)
mesh =

Coordinates: [5x2 double]

Elements: [4x3 double] \( \sharp \mathcal{N}(\mathcal{M}) \times \sharp \mathcal{N}(\mathcal{M}) \) sparse integer matrix:

Edges: [8x2 double] entry \((i, j) = \text{edge index, if } \neq 0\)

Vert2Edge: [5x5 double] \( \sharp \mathcal{E}(\mathcal{M}) \times 2 \) integer array:

Edge2Elem: [8x2 double] indices of adjacent cells in Elements array

EdgeLoc: [8x2 double] \( \sharp \mathcal{E}(\mathcal{M}) \times 2 \) integer array: local indices of edges w.r.t. adjacent cells

Notation: \( \mathcal{E}(\mathcal{M}) \) \( \hat{} \) edges of 2D mesh

---

How to number ↔ order

local shape functions

global shape functions

Elements, Edges arrays \( \Rightarrow \) ordering of vertices of cells/endpoints of edges

Arrays (of vertices, cells, edges) \( \Rightarrow \) array indices \( \Rightarrow \) numbering of global shape functions
2.2.4 Algorithms

Guideline:

Cell oriented assembly \(\iff\) (2.2.4)

Loop: \texttt{foreach } \(K \in \mathcal{M}\) \texttt{ do } \{ local operations on \(K\) \(\rightarrow A_K\) \ + \ \(A = A + T_K^T A_K T_K\) \}

Notion: local operations \(\triangleq\) required only data from fixed “neighbourhood” of \(K\)

\(\triangleright\) computational effort “\(O(1)\)”: independent of \(#\mathcal{M}\)

Cell oriented assembly in MATLAB

```matlab
function A = assemble(Mesh)

for k = Mesh.Elements'
    idx = \(\mathbf{1}\)
    Aloc = \(\mathbf{2}\)
    A(idx,idx) = A(idx,idx)+Aloc;
end
```

1. row vector of index numbers of global shape functions \(b_{i_1}, \ldots, b_{i_Q} \in V_N\)
corresponding to local shape functions \(b_{K_1}^1, \ldots, b_{K_Q}^K\):

\[\text{id}x = (i_1, \ldots, i_Q)\]
(encodes index mapping matrix \(T_K\))

2. \(Q \times Q\) element stiffness matrix

For Lagrangian FEM (\(\rightarrow\) Sect. 2.1.5):

the total computational effort is of the order \(O(#\mathcal{M}) = O(N)\), \(N := \text{dim } V_N\).
Example 2.2.7 (Assembly for quadratic Lagrangian FE in MATLAB code).

Setting: FE space $\mathcal{S}_2^0(\mathcal{M})$ on triangular mesh $\mathcal{M}$ of polygon $\Omega \subset \mathbb{R}^2$

Recall: 6 local shape functions: 3 vertex-associated, 3 edge-associated $\rightarrow$ Ex. 2.1.9, Sect. 2.1.5

Convention: vertex-associated global shape functions $\rightarrow b_1, \ldots, b_{\#\mathcal{M}}$

edge-associated global shape functions $\rightarrow b_{\#\mathcal{M}+1}, \ldots, b_{\#\mathcal{M}+\#\mathcal{E}(\mathcal{M})}$

Local numbering

Local numbering $\rightarrow$
function A = assemMat_QFE(Mesh, EHandle, varargin)

nV = size(Mesh.Coordinates,1);
nE = size(Mesh.Elements,1)

I = zeros(36*nE,1); J = I; a = I; offset = 0;
for k =1:nE
    vidx = Mesh.Elements(k,:)
    idx = [vidx, ...
        Mesh.Vert2Edge(vidx(1),vidx(2))+nV, ... 
        Mesh.Vert2Edge(vidx(2),vidx(3))+nV, ... 
        Mesh.Vert2Edge(vidx(3),vidx(1))+nV];
    Aloc = transpose(EHandle(Mesh.Coordinates(vidx,:), ...
        Mesh.ElemFlag(k),varargin{:}));

    Qsq = prod(size(Aloc)); range = offset + 1:Qsq;
    t = idx(ones(length(idx),1),:,:)'; I(range) = t(:);
    t = idx(ones(1,length(idx)),:,:); J(range) = t(:);
    a(range) = Aloc(:);
    offset = offset + Qsq;
end
A = sparse(I,J,a);
1: EHandle (function handle) → provides element stiffness matrix $A_K \in \mathbb{R}^{6,6}$

2: $I, J, a \rightarrow$ linear arrays storing $(i, j, a_{ij})$ for stiffness matrix $A$.
   Initialized with 0 for the sake of efficiency → Ex. 2.2.8

3: $\text{idx} \rightarrow$ index mapping vector, see above

4: $A_{loc} = A_K \in \mathbb{R}^{6,6}$ (element stiffness matrix)

5: $\text{Mesh.ElemFlag}(k)$ marks groups of elements (e.g. to select local heat conductivity $\sigma$ in (1.6.4))

6: Build sparse MATLAB-matrix (→ Def. 2.1.13) from index-entry arrays

Example 2.2.8 (Efficient implementation of assembly).

tic-toe-timing (min of 4v runs), MATLAB V7, Intel Pentium 4 Mobile CPU 1.80GHz, Linux
Computation of element stiffness matrices skipped!
Sparse assembly:
\[ A(idx, idx) = A(idx, idx) + Aloc; \]

Array assembly I: “growing arrays”
\[
I = []; \quad J = []; \quad a = []; \\
\ldots \\
t = idx(:, \text{ones}(\text{length}(idx), 1))'; \\
I = [I; t(:)]; \\
t = idx(:, \text{ones}(1, \text{length}(idx))); \\
J = [J; t(:)]; \\
a = [a; Aloc(:)]; \\
\]

Array assembly III

→ see code fragment above

2.2.5 Local computations

First option: analytic evaluations
We discuss bilinear form related to $-\Delta$, triangular Lagrangian finite elements of degree $p$:

$$K \text{ triangle: } a_K(u, v) := \int_K \nabla u \cdot \nabla v \, dx$$  \[\Rightarrow\] element stiffness matrix.

Use barycentric coordinate representations of local shape functions

$$b^K_i = \sum_{\alpha \in \mathbb{N}_0^3, |\alpha| \leq p} \kappa_\alpha \lambda_1^{\alpha_1} \lambda_2^{\alpha_2} \lambda_3^{\alpha_3} , \quad \kappa_\alpha \in \mathbb{R} ,$$  \text{(2.2.5)}

$$\Rightarrow \nabla b^K_i = \sum_{\alpha \in \mathbb{N}_0^3, |\alpha| \leq p} \kappa_\alpha \left( \begin{array}{c} \alpha_1 \lambda_1^{\alpha_1-1} \lambda_2^{\alpha_2} \lambda_3^{\alpha_3} \nabla \lambda_1 + \alpha_2 \lambda_1^{\alpha_1} \lambda_2^{\alpha_2-1} \lambda_3^{\alpha_3} \nabla \lambda_2 + \\ \alpha_3 \lambda_1^{\alpha_1} \lambda_2^{\alpha_2} \lambda_3^{\alpha_3-1} \nabla \lambda_3 \end{array} \right) .$$  \text{(2.2.6)}

To evaluate

$$\int_K \lambda_1^{\beta_1} \lambda_2^{\beta_2} \lambda_3^{\beta_3} \nabla \lambda_i \cdot \nabla \lambda_j \, dx , \quad i, j \in \{1, 2, 3\}, \beta_k \in \mathbb{N} .$$  \text{(2.2.7)}
If \( a^1, a^2, a^3 \) vertices of \( K \) (counterclockwise ordering):

\[
\lambda_1(x) = \frac{1}{2|K|} \left( x - \left( \begin{array}{c} a^2_1 \\ a^2_2 \end{array} \right) \right) \cdot \left( \begin{array}{c} a^2_1 - a^3_1 \\ a^3_1 - a^2_1 \end{array} \right),
\]

\[
\lambda_2(x) = \frac{1}{2|K|} \left( x - \left( \begin{array}{c} a^3_1 \\ a^3_2 \end{array} \right) \right) \cdot \left( \begin{array}{c} a^3_1 - a^1_1 \\ a^1_1 - a^3_1 \end{array} \right),
\]

\[
\lambda_3(x) = \frac{1}{2|K|} \left( x - \left( \begin{array}{c} a^1_1 \\ a^1_2 \end{array} \right) \right) \cdot \left( \begin{array}{c} a^1_2 - a^2_2 \\ a^2_2 - a^1_2 \end{array} \right).
\]

\[
\text{grad} \lambda_1 = \frac{1}{2|K|} \left( \begin{array}{c} a^2_2 - a^3_2 \\ a^3_1 - a^2_1 \end{array} \right), \quad \text{grad} \lambda_2 = \frac{1}{2|K|} \left( \begin{array}{c} a^3_2 - a^1_2 \\ a^1_1 - a^3_1 \end{array} \right), \quad \text{grad} \lambda_3 = \frac{1}{2|K|} \left( \begin{array}{c} a^1_2 - a^2_2 \\ a^2_1 - a^1_1 \end{array} \right).
\]

\[
\left( \int_K \text{grad} \lambda_i \cdot \text{grad} \lambda_j \, dx \right)^3_{i,j=1} = \frac{1}{2} \left( \begin{array}{ccc} \cot \omega_3 + \cot \omega_2 & - \cot \omega_3 & - \cot \omega_2 \\ - \cot \omega_3 & \cot \omega_3 + \cot \omega_1 & - \cot \omega_1 \\ - \cot \omega_2 & - \cot \omega_1 & \cot \omega_2 + \cot \omega_1 \end{array} \right). \tag{2.2.8}
\]
Exercise.

**Lemma 2.2.3** (Integration of powers of barycentric coordinate functions). For any non-degenerate $d$-simplex $K$ and $\alpha_j \in \mathbb{N}$, $j = 1, \ldots, d + 1$,

$$
\int_K \lambda_1^{\alpha_1} \cdots \lambda_{d+1}^{\alpha_{d+1}} \, d\mathbf{x} = d!|K| \frac{\alpha_1! \alpha_2! \cdots \alpha_{d+1}!}{(\alpha_1 + \alpha_2 + \cdots + \alpha_{d+1} + d)!} \quad \forall \alpha \in \mathbb{N}_0^{d+1}.
$$

(2.2.9)

**Proof for** $d = 2$ → Appendix B.1

**Remark.** Alternative: symbolic computing (MAPLE, Mathematica) for local computations

### 2.2.6 Numerical quadrature

Second option (for local evaluations): Numerical quadrature

$$
\int_{\Omega} f(\mathbf{x}) \, d\mathbf{x} \approx \sum_{K \in \mathcal{M}} \sum_{l=1}^{P_K} \omega^K_l f(\pi^K_l), \quad \pi^K_l \in K, \omega^K_l \in \mathbb{R}.
$$

(2.2.10)
Terminology:

\[ \omega^K_l \rightarrow \text{weights}, \quad \pi^K_l \rightarrow \text{quadrature nodes} \]

(2.2.10) = local quadrature rule

Mandatory ● for computation of **load vector** (\( f \) complicated/only available in procedural form)

● for computation of **stiffness matrix** if \( \sigma = \sigma(x) \) does not permit analytic integration.

Guideline: only quadrature rules with positive weights are numerically stable.

For **affine equivalent** finite elements (\( \rightarrow \) Sect. 2.1.6):

Parameter definition of local quadrature rules on **reference cell** \( \hat{K} \):

\[
\int_{\hat{K}} f(\hat{x}) \, d\hat{x} \approx |\hat{K}| \sum_{l=1}^{P} \hat{\omega}_l \, f(\hat{\pi}_l) \quad \Rightarrow \quad \int_{\Omega} f(x) \, dx \approx \sum_{K \in \mathcal{M}} |K| \sum_{l=1}^{P} \omega^K_l \, f(\pi^K_l) \\
\text{with} \quad \omega^K_l = \hat{\omega}_l, \quad \pi^K_l = \Phi_K(\hat{\pi}_l).
\]

How to gauge the quality of parametric local quadrature rules?

Quality of a parametric local quadrature rule on \( K \) \( \sim \) largest space of polynomials on \( \hat{K} \) integrated exactly by the corresponding quadrature rule on \( \hat{K} \).
Parlance: Quadrature rule exact for $\mathcal{P}_p(\hat{K}) \Rightarrow$ quadrature rule of order $p + 1$

degree of exactness $p$

Example 2.2.9 (Local quadrature rules on triangles).

If $K$ triangle $\Rightarrow \hat{K} := \text{convex}\left\{ (\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}), (\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}), (\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}) \right\}$.

Quadrature rules described by pairs $(\hat{\omega}_1, \hat{\pi}_1), \ldots, (\hat{\omega}_p, \hat{\pi}_p)$, $P \in \mathbb{N}$.

- Quadrature rule of order 2 (exact for $\mathcal{P}_1(\hat{K})$)

$$\left\{ \left( \frac{1}{3}, \left( \begin{smallmatrix} 0 \\ 0 \end{smallmatrix} \right) \right), \left( \frac{1}{3}, \left( \begin{smallmatrix} 0 \\ 1 \end{smallmatrix} \right) \right), \left( \frac{1}{3}, \left( \begin{smallmatrix} 1 \\ 0 \end{smallmatrix} \right) \right) \right\}.$$ (2.2.11)

- Quadrature rule of order 3 (exact for $\mathcal{P}_2(\hat{K})$)

$$\left\{ \left( \frac{1}{3}, \left( \begin{smallmatrix} \frac{1}{2} \\ 0 \end{smallmatrix} \right) \right), \left( \frac{1}{3}, \left( \begin{smallmatrix} 0 \\ \frac{1}{2} \end{smallmatrix} \right) \right), \left( \frac{1}{3}, \left( \begin{smallmatrix} \frac{1}{2} \\ 1/2 \end{smallmatrix} \right) \right) \right\}.$$ (2.2.12)

- One-point quadrature rule of order 2 (exact for $\mathcal{P}_1(\hat{K})$)

$$\left\{ \left( 1, \left( \begin{smallmatrix} 1/3 \\ 1/3 \end{smallmatrix} \right) \right) \right\}.$$ (2.2.13)
Quadrature rule of order 6 (exact for $P_5(\hat{K})$)

\[
\left\{ \left( \frac{9}{40}, \frac{1}{3} \right), \left( \frac{155 + \sqrt{15}}{1200}, \frac{6 + \sqrt{15}/21}{6 + \sqrt{15}/21} \right), \left( \frac{155 + \sqrt{15}}{1200}, \frac{9 - 2\sqrt{15}/21}{6 + \sqrt{15}/21} \right), \left( \frac{155 + \sqrt{15}}{1200}, \frac{6 - \sqrt{15}/21}{9 + 2\sqrt{15}/21} \right), \left( \frac{155 - \sqrt{15}}{1200}, \frac{9 + 2\sqrt{15}/21}{6 - \sqrt{15}/21} \right), \left( \frac{155 - \sqrt{15}}{1200}, \frac{6 - \sqrt{15}/21}{6 - \sqrt{15}/21} \right) \right\}
\]

(2.2.14)

Example 2.2.10 (Local quadrature rules on quadrilaterals).
If $K$ quadrilateral $\Rightarrow \hat{K} := \text{convex}\{\binom{0}{0}, \binom{1}{0}, \binom{0}{1}, \binom{1}{1}\}$ (unit square).

On $\hat{K}$: tensor product construction:

If $\{(\omega_1, \pi_1), \ldots, (\omega_P, \pi_P)\}$, $P \in \mathbb{N}$, quadrature rule on the interval $]0, 1[$, exact for $\mathcal{P}_0]0, 1[$, then

\[
\{ \binom{\omega_1^2}{\pi_1}, \binom{\omega_2}{\pi_2}, \ldots, \binom{\omega_P}{\pi_P} \} \quad \{ \binom{\omega_1 \omega_1}{\pi_1}, \binom{\omega_1 \omega_P}{\pi_P}, \ldots, \binom{\omega_P^2}{\pi_P} \}
\]

quadrature rule on $\hat{K}$, exact for $\mathcal{Q}_p(\hat{K})$.  

2.2 p. 115
Quadrature rules on \([0, 1]\) (\(\rightarrow\) basic numerics):

- Classical Newton-Cotes formulas (equidistant quadrature nodes).
- Gauss-Legendre quadrature rules, exact for \(\mathcal{P}_2P\]0, 1\]) using only \(P\) nodes.
- Gauss-Lobatto quadrature rules: \(P\) nodes including \(\{0, 1\}\), exact for \(\mathcal{P}_{2P-1}[0, 1]\).

**2.2.7 Treatment of essential boundary conditions**

Remember Sect. 1.8: extension \(g \rightarrow \tilde{g}\) of Dirichlet data into \(\Omega\) yielded linear variational problem.

Adaptation to finite element setting:
\( V_N = \) finite element space without constraints on \( \partial \Omega \).

**FIRST STEP:** Interpolation/projection of boundary data

FE-space \( V_N \) \( \Rightarrow \) \( W_N := V_N|_{\partial \Omega} \) (FE trace space)

Example: if \( V_N = \mathcal{S}_1^0(\mathcal{M}) \), then \( W_N \) = set of piecewise linear, continuous functions on boundary mesh \( \mathcal{M}|_{\partial \Omega} \).

**BUT,** not necessarily \( g \in W_N \) !

Replace \( g \) by (interpolant, least squares fit, etc.) \( g_N \in W_N \)

Example: if \( V_N = \mathcal{S}_1^0(\mathcal{M}) \) and \( g \in C^0(\partial \Omega) \), then choose \( g_N \) as p.w. linear interpolant.

**SECOND STEP:** Trivial extension of \( g_N \rightarrow \tilde{g}_N \in V_N \)

Only nodal basis functions associated with node/edge/face \( \subset \partial \Omega \) contribute to \( \tilde{g}_N \)!
Example: if \( V_N = \mathcal{S}_1^0(\mathcal{M}) \), \( g_N \) p.w. linear continuous on \( \mathcal{M}_{|\partial \Omega} \)

\[
\tilde{g}_N = \sum_{p \in \mathcal{N}(\mathcal{M}_{|\partial \Omega})} g_N(p) b^p_N , \quad \text{where } b^p_N = \text{"hat function"\} for node } p .
\]

\[
u_N \in V_{N,0}: \quad a(u_N + \tilde{g}_N, v_N) = f(v_N) \quad \forall v_N \in V_{N,0} . \quad (2.2.15)
\]

\( V_{N,0} := \{ v_N \in V_N : v_N = 0 \text{ on } \partial \Omega \} = \text{span of "interior" basis functions.} \)

Equivalent algebraic perspective: partitioning of (big, w.r.t. \( V_N \)) linear system \( A\tilde{\mu} = \tilde{\phi} \)

\[
\begin{pmatrix}
A_{\Omega\Omega} & A_{\Gamma\Omega} \\
A_{\Omega\Gamma} & A_{\Gamma\Gamma}
\end{pmatrix}
\begin{pmatrix}
\tilde{\mu}_\Omega \\
\tilde{\mu}_\Gamma
\end{pmatrix}
=
\begin{pmatrix}
\tilde{\phi}_\Omega \\
\tilde{\phi}_\Gamma
\end{pmatrix} . \quad (2.2.16)
\]

Note: coefficients \( \tilde{\mu}_\Gamma \) known = relevant coefficients of \( \tilde{g}_N \)

Elimination

\[
A_{\Omega\Omega}\tilde{\mu}_\Omega = \tilde{\phi}_\Omega - A_{\Gamma\Omega}\tilde{\mu}_\Gamma . \quad (2.2.17)
\]

Remark 2.2.11. Alternative: elimination on element level \( \Rightarrow \) modified \( \tilde{\phi}_K \)

\[
A_K = \begin{pmatrix}
A_{ii} & A_{bi} \\
A_{ib} & A_{bb}
\end{pmatrix} , \quad \tilde{\phi}_K = \begin{pmatrix}
\tilde{\phi}_i \\
\tilde{\phi}_b
\end{pmatrix} \quad \Rightarrow \quad \tilde{A}_K = A_{ii} , \quad \tilde{\phi}_K = \tilde{\phi}_i - A_{bi}\tilde{\mu}_{\Gamma,K} .
\]

Then do assembly based on \( \tilde{A}_K \) and \( \tilde{\phi}_K \).
2.2.8 Boundary approximation

Sect 2.1.6 → approximate treatment of curved $\partial \Omega$ by parametric FE:

Idea:

Piecewise polynomial approximation of boundary (boundary fitting)  
($\partial \Omega$ locally considered as function over straight edge of an element)

Example: Piecewise quadratic boundary approximation  
(Part of $\partial \Omega$ between $a^1$ and $a^2$ approximated by parabola)

Mapping $\tilde{K}$ → “curved element” $K$:

$$\Phi(\tilde{x}) := \tilde{x} + 4\delta \lambda_1(\tilde{x})\lambda_2(\tilde{x}) n.$$  

($\lambda_i$ barycentric coordinate functions on $\tilde{K}$, $n$ normal to $E_{\Gamma}$)

Note: Essential: $\Phi$ diffeomorphism $\leftrightarrow$ $\delta$ sufficiently small
Transformation formula for gradients: for \( u : K \mapsto \mathbb{R} \), diffeomorphism \( \Phi : \tilde{K} \mapsto K \)

\[
(\text{grad}_\tilde{x}(\Phi^* u))(\tilde{x}) = (D\Phi(\tilde{x}))^T (\text{grad}_x u)(\Phi(\tilde{x})) \quad \forall \tilde{x} \in \tilde{K}.
\] (2.2.18)

**Proof:** chain rule:

\[
\frac{\partial u}{\partial x_i}(x) = \frac{\partial}{\partial x_i} \Phi^* u(\Phi^{-1}(x)) = \sum_{j=1}^{d} \frac{\partial \Phi^* u}{\partial \tilde{x}_j}(\Phi^{-1}(x)) \frac{\partial \Phi^{-1}_j}{\partial x_i}(x).
\]

\[
\text{grad} u(x) = \left(D \Phi^{-1}(x)\right)^T \text{grad}_{\tilde{x}}(\Phi^* u)(\Phi^{-1}(x)) = D\Phi(\Phi^{-1}(x))^{-T} (\text{grad} \Phi^* u)(\Phi^{-1}(x)).
\]

Parametric construction:

\[
b_i^{\tilde{K}} = \Phi^* b_i^K, \quad i = 1, \ldots, Q
\]

Local shape functions on \( \tilde{K} \) Local shape functions on \( K \)

Local computations use (2.2.18) & transformation formula (for multidimensional integrals):

\[
\int_K f(\Phi(x)) \, dx = \int_{\tilde{K}} f(\tilde{x}) |\det D\Phi(\tilde{x})| \, d\tilde{x} \quad \text{for } f : K \mapsto \mathbb{R},
\]
\[ \int_K \text{grad} u \cdot \text{grad} v \, dx = \int_{\tilde{K}} (\text{grad} u)(\Phi(\tilde{x})) \cdot (\text{grad} v)(\Phi(\tilde{x})) \mid \text{det} D\Phi(\tilde{x}) \mid \, d\tilde{x} \]
\[ = \int_{\tilde{K}} D\Phi^{-T}(\tilde{x}) \text{grad}_{\tilde{x}}(\Phi^* u) \cdot D\Phi^{-T}(\tilde{x}) \text{grad}_{\tilde{x}}(\Phi^* v) \mid \text{det} D\Phi(\tilde{x}) \mid \, d\tilde{x} . \]

\[ \int_K \text{grad} b^K_i \cdot \text{grad} b^K_j \, dx = \int_{\tilde{K}} \left\{ D\Phi(\tilde{x})^T D\Phi(\tilde{x}) \right\}^{-1} \text{grad} b^K_i \cdot \text{grad} b^K_j \mid \text{det} D\Phi(\tilde{x}) \mid \, d\tilde{x} . \]

Note: local shape functions \( b^K_i \) simple polynomials!

For parabolic boundary fitting:
\[ D\Phi = Id + 4\delta n \cdot \text{grad}(\lambda_1 \lambda_2)^T \in \mathbb{R}^{2,2} , \quad \text{det}(D\Phi) = 1 + 4\delta n \cdot \text{grad}(\lambda_1 \lambda_2) . \]

Next: numerical quadrature (→ Sect. 2.2.6) on \( \tilde{K} \)
2.2.9 Treatment of hanging nodes

Reminder:

2D: if edge of a cell is union of edges of other cells ⇒ hanging node(s).

Discussion for $H^1(\Omega)$-conforming FE model mesh

Crucial: Global continuity of FE functions

Special case: Linear Lagrangian finite elements on model mesh $\mathcal{M}$, Fig. 37

$$\forall v_N \in \mathcal{S}^0_1(\mathcal{M}): \quad v_N(\zeta) = \frac{|\zeta - \mathbf{a}^1|}{|\mathbf{a}^2 - \mathbf{a}^1|} v_N(\mathbf{a}^2) + \frac{|\zeta - \mathbf{a}^2|}{|\mathbf{a}^2 - \mathbf{a}^1|} v_N(\mathbf{a}^1).$$

Then, if $v_N|_K \in \mathcal{P}_1(K) \forall K \in \mathcal{M}$ ⇒ $v_N$ continuous across $F$. (= linear interpolation from $\mathbf{a}^1, \mathbf{a}^2 \rightarrow \zeta$ ⇒ $\zeta$ “slave node”)

Assembly (→ Sect. 2.2.4) (for $V_N = \mathcal{S}^0_1(\mathcal{M})$ on mesh with hanging nodes):

Fig. 37
Assumption: mesh data structure (→ Sect. 2.2.3) of Ex. 2.2.6 (Coordinates, Elements)

- slave nodes = midpoints of (master) edges
- extra information on “masters” (a¹, a² in Fig. 37) of slave nodes (ζ in Fig. 37):
  - slaves(:,1) = indices of slave nodes in coordinates
  - slaves(:,2:3) = indices of master nodes

1: Assembly treating slave nodes and regular nodes alike → stiffness matrix \( \tilde{A} \)

2: Add \( \frac{1}{2} \) “slave row” of \( A \) to related “master rows”:

Post-processing of stiffness matrix for slave nodes

\[
\begin{align*}
N &= \text{size}(A,1); \\
S &= \text{speye}(N) + 0.5*\text{sparse}([\text{slave}(:,2); \text{slave}(:,3)],... \\
&\quad [\text{slave}(:,1); \text{slave}(:,1)],... \\
&\quad \text{ones}(2*\text{size}(<\text{slave},1)),N,N); \\
A &= S*A*S'; \\
A(\text{slave}(:,1), :) &= []; \quad A(:,\text{slave}(:,1)) = []; \\
\end{align*}
\]

A change of basis ! → Lemma 2.1.6

Deleting rows/columns ↔ slave nodes

Note: More efficient: process slave nodes during local assembly
2.2.10 Static condensation

interior basis functions = \textcolor{red}{\text{global shape functions}} supported inside a cell

(occur for $\delta^0_3(M)$ on triangular mesh $M$ in 2D)

Sorting of global basis functions: coefficients for interior basis functions last

Block structure of resulting linear system $A\bar{\mu} = \bar{\phi}$

$$A\bar{\mu} = \begin{pmatrix} A_{oo} & A_{oi} \\ A_{io} & A_{ii} \end{pmatrix} \begin{pmatrix} \mu_o \\ \mu_i \end{pmatrix} = \begin{pmatrix} \phi_o \\ \phi_i \end{pmatrix} = \bar{\phi} .$$ (2.2.19)

$A_{ii} \leftarrow$ coupling among interior basis functions

$A_{oi} \leftarrow$ coupling between interior b.f. & basis functions on nodes/edges

Note:

$A_{ii}$ is block-diagonal with small blocks $\Rightarrow$ “easy to invert”

[Elimination of $\mu_i$ (Static condensation)]

Schur complement system:

$$\left( A_{oo} - A_{oi} A_{ii}^{-1} A_{io} \right) \bar{\mu}_o = \bar{\phi}_o - A_{oi} A_{ii}^{-1} \phi_i .$$
2.3 Finite Difference Methods (FDM)

Finite difference methods:

- Perspective: classical interpretation of boundary value problems
- Idea: replace derivatives by difference quotients using values at nodes of a grid = “regular” mesh

2.3.1 From FEM to FD

Model problem (→ Sect. 1.2): 1D heat conduction, $\sigma = \sigma(x)$, homogeneous Dirichlet b.c.

$$\frac{d}{dx} \left( \sigma(x) \frac{d}{dx} u \right) = f \quad \text{in } \Omega := ]0, 1[, \quad u(0) = u(1) = 0.$$ 

Assumption:

$$\sigma \in C^0(]0, 1[), f \in C^0(]0, 1[).$$

Finite element Galerkin discretization → Sect 2.1.3:
\[
\mathcal{M} = \{ i h, (i + 1) h \mid i = 0, \ldots, N, N \in \mathbb{N} \}: \text{equidistant mesh with meshwidth } h := (N + 1)^{-1},
\]

- finite element space \( V_N = \mathcal{S}_0^0(\mathcal{M}) \) (p.w. linear Lagrangian FE), \( \mathcal{S}_0^0(\mathcal{M}) \subset H_0^1(\Omega) \).

**Local shape functions** on cell \( K = [i h, (i + 1) h[ \), \( i = 0, \ldots, N \): \( \rightarrow \) Ex. 2.1.6

\[
\begin{align*}
  b_1^K(x) &= 1 - \frac{x - i h}{h}, & \frac{d}{dx} b_1^K(x) &= -\frac{1}{h}, \\
  b_2^K(x) &= \frac{x - i h}{h}, & \frac{d}{dx} b_2^K(x) &= \frac{1}{h}.
\end{align*}
\]

Local computations based on numerical quadrature (\( \rightarrow \) Sect. 2.2.6):

- **element stiffness matrix**: midpoint quadrature \( \int_K f(x) \, dx \approx h f(\xi_{i+1/2}), \xi_{i+1/2} := (i + 1/2)h \),
- **element load vector**: trapezoidal rule \( \int_K f(x) \, dx \approx \frac{1}{2} h (f(\xi_i) + f(\xi_{i+1})), \xi_i := i h \).

**Element stiffness matrix**

\[
A_K = \left( h \sigma(\xi_{i+1/2}) \left( \frac{d}{dx} b^K_k(\xi_{i+1/2}) \left( \frac{d}{dx} b^K_l(\xi_{i+1/2}) \right) \right) \right)_{l,k=1}^2 = \frac{1}{h} \sigma_{i+1/2} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \in \mathbb{R}^{2,2}.
\]
Element load vector:

\[
\tilde{\phi}_K = \frac{1}{2} h \left( f(\xi_i) b_j^K(\xi_i) + f(\xi_{i+1}) b_j^K(\xi_{i+1}) \right)_{j=1}^{2} = \frac{1}{2} h \begin{pmatrix} f(\xi_i) \\ f(\xi_{i+1}) \end{pmatrix} \in \mathbb{R}^2.
\]

Tridiagonal linear system of equations \( \mathbf{A} \tilde{\mu} = \tilde{\phi} \), where \( \mu_i \approx u(\xi_i) \):

\[
\mathbf{A} = \frac{1}{h} \begin{pmatrix}
\sigma_{\frac{1}{2}} + \sigma_{\frac{3}{2}} & -\sigma_{\frac{3}{2}} & 0 & \cdots & 0 \\
-\sigma_{\frac{3}{2}} & \sigma_{\frac{3}{2}} + \sigma_{\frac{5}{2}} & -\sigma_{\frac{5}{2}} & 0 \\
0 & -\sigma_{\frac{5}{2}} & \sigma_{\frac{5}{2}} + \sigma_{\frac{7}{2}} & -\sigma_{\frac{7}{2}} & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
-\sigma_{\frac{N-3}{2}} & \sigma_{\frac{N-3}{2}} + \sigma_{\frac{N-1}{2}} & -\sigma_{\frac{N-1}{2}} & 0 & -\sigma_{\frac{N-1}{2}} \\
0 & -\sigma_{\frac{N-1}{2}} & \sigma_{\frac{N-1}{2}} + \sigma_{\frac{N+1}{2}} & \cdots & \cdots & \cdots
\end{pmatrix},
\]

\[
\tilde{\phi} = h \begin{pmatrix} f_1 \\ f_2 \\ \cdots \\ f_N \end{pmatrix}^T \in \mathbb{R}^N,
\]

\((\sigma_{i+1/2} := \sigma(\xi_{i+1/2}), f_i := f(\xi_i))\).
FDM approach: Approximation of derivative by symmetric difference quotient:

\[
\frac{d}{dx} u \bigg|_{x=\xi} \approx \frac{u(\xi + \frac{1}{2}h) - u(\xi - \frac{1}{2}h)}{h}.
\]

Applied twice at positions \( x = \xi_i, x = \xi_{i+1/2}, x = \xi_{i-1/2} \):

\[
\frac{d}{dx} \left( \sigma(x) \frac{d}{dx} u \right) \bigg|_{x=\xi_i} \approx \frac{1}{h} \left( \frac{\sigma_{i+1/2} \frac{d}{dx} u}{x=\xi_{i+1/2}} - \frac{\sigma_{i-1/2} \frac{d}{dx} u}{x=\xi_{i-1/2}} \right)
\]

\[
\approx \frac{1}{h^2} \left( \sigma_{i+1/2} (u(\xi_{i+1}) - u(\xi_i)) - \sigma_{i-1/2} (u(\xi_i) - u(\xi_{i-1})) \right).
\]

Use this approximation at grid points \( \xi_i, i = 1, \ldots, N \):

\[
\frac{1}{h^2} \left( \sigma_{i+1/2} (u(\xi_{i+1}) - u(\xi_i)) - \sigma_{i-1/2} (u(\xi_i) - u(\xi_{i-1})) \right) = f_i, \quad i = 1, \ldots, N.
\]  

(2.3.3)

Setting \( \mu_i := u(\xi_i) \) \( \Rightarrow \) (2.3.3) = \( h^{-1} A \bar{\mu} = h^{-1} \bar{\phi} \) from (2.3.1), (2.3.2).
Stencil notation for (row of) linear system (2.3.1), (2.3.2):

\[
\begin{bmatrix}
-\sigma_{i-1/2} & \sigma_{i-1/2} + \sigma_{i+1/2} & -\sigma_{i+1/2} \\
\uparrow & \uparrow & \uparrow \\
\xi_{i-1} & \xi_i & \xi_{i+1}
\end{bmatrix}_{h,x=\xi_i} \otimes \mathbf{\bar{u}} = h \begin{bmatrix} f_i \end{bmatrix}_h .
\]

\[
\frac{1}{h} \left( -\sigma_{i-1/2} \mu_{i-1} + (\sigma_{i-1/2} + \sigma_{i+1/2}) \mu_i - \sigma_{i+1/2} \mu_{i+1} \right) = h f_i , \quad i = 1, \ldots, N .
\]

Convention:

\[
\mu_0 = \mu_{N+1} = 0
\]

Homogeneous Dirichlet BVP for Laplacian:

\[
-\Delta u = f \quad \text{in} \quad \Omega := ]0, 1[^2 ,
\]
\[
u = 0 \quad \text{on} \quad \partial \Omega .
\]

Finite element Galerkin discretization:

- \( \mathcal{M} = \text{triangular tensor-product grid} \)
  (meshwidth \( h = (1 + N)^{-1} \), \( N \in \mathbb{N} \))
- \( V_N = \delta^{0}_{1,0}(\mathcal{M}) \), piecewise linear
  Lagrangian finite elements \( \rightarrow \) Sect. 2.1.4

Global shape functions: "hat functions"
Element stiffness matrix from (2.2.8):

\[ A_K = \frac{1}{2} \begin{pmatrix} 2 & -1 & -1 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix} . \]

(← numbering of local shape functions)

Element load vector: use three-point quadrature formula (2.2.11)

\[ \bar{\phi}_K = \frac{1}{6} h^2 \begin{pmatrix} f(a^1) \\ f(a^2) \\ f(a^3) \end{pmatrix} . \]
Contributions to load vector component associated with node $p$:

- From $K_1$: $(\phi_{K_1})_2$
- From $K_2$: $(\phi_{K_2})_3$
- From $K_3$: $(\phi_{K_3})_3$
- From $K_4$: $(\phi_{K_4})_1$
- From $K_5$: $(\phi_{K_5})_1$
- From $K_6$: $(\phi_{K_6})_2$

$$\bar{\phi}_p = h^2 f(p).$$
“center node” \( p \rightarrow \) grid position \((ih, jh), 1 \leq i, j \leq N\)

Indexing scheme for coefficient vector:

\[
\mu_{i,j} \leftrightarrow \text{hat function at } (ih, jh) \\
(1 \leq i, j \leq N)
\]

Row of \( A\tilde{\mu} = \tilde{\varphi} \leftrightarrow \) node \( p \):

\[
4\mu_{i,j} - \mu_{i-1,j} - \mu_{i+1,j} - \mu_{i,j-1} - \mu_{i,j+1} = h^2 f_{i,j}.
\]

\((\mu_{i,j} = 0 \leftrightarrow i = 0, N + 1 \lor j = 0, N + 1)\)

5-point stencil for \(-\Delta\):

\[
\begin{bmatrix}
-1 & 4 & -1 \\
-1 & h_{p} & -1 \\
-1 & 4 & -1 \\
\end{bmatrix} \circ \tilde{\mu} = h^2 f(p). \quad (2.3.4)
\]
Finite difference approach to $-\Delta$: approximation by symmetric difference quotients

\[
\begin{align*}
\frac{d^2}{dx^2}u \big|_{x=(\xi, \eta)} & \approx \frac{u(\xi - h, \eta) - 2u(\xi, \eta) + u(\xi + h, \eta)}{h^2}, \\
\frac{d^2}{dy^2}u \big|_{x=(\xi, \eta)} & \approx \frac{u(\xi, \eta - h) - 2u(\xi, \eta) + u(\xi, \eta + h)}{h^2}.
\end{align*}
\]

\[-\Delta u \big|_{x=(\xi, \eta)} \approx \frac{1}{h^2} \left( 4u(\xi, \eta) - u(\xi - h, \eta) - u(\xi + h, \eta) - u(\xi, \eta - h) - u(\xi, \eta + h) \right).\]

Using this approximation at grid point $p = (ih, jh)$:

\[
\text{(modulo scaling by $h^{-2}$)} \quad \text{5-point stencil (2.3.4)}
\]

(Most) finite difference schemes $\leftrightarrow$ finite element Galerkin schemes with numerical quadrature on structured meshes

More on the stencil notation:
9-point stencil
(on tensor product grid with meshwidth $h$)

\[
\begin{bmatrix}
\alpha_{nw} & \alpha_n & \alpha_{ne} \\
\alpha_w & \alpha_c & \alpha_e \\
\alpha_{sw} & \alpha_s & \alpha_{se}
\end{bmatrix}_h
\]

Terminology:
- $\alpha_x =$ weight
- $\alpha_c =$ center $\leftrightarrow$ off-center

Stencils offer a description of a finite dimensional linear operator \textit{without imposing an order on the vector components}

(Matrix notation requires ordering!)

(Difference) \text{ stencils} \leftrightarrow \text{ local linear operators on grid function space}

Bilinear form $a(\cdot, \cdot)$ on finite element space $V_N$, see (2.1.1)

[Choosing nodal FE basis functions]
Local linear operators on grid function space (\(\rightarrow\) stencils)

[Ordering the basis]

Stiffness matrix \(\mathbf{A}\)

Example: skew stencil

\[
\begin{pmatrix}
\alpha_w & \alpha_{ne} \\
\alpha_{sw} & \alpha_c & \alpha_e
\end{pmatrix}
\]

\(1/2h, 1/4\sqrt{5h}\)

Stencils also meaningful on unstructured meshes
2.3.2 The discrete maximum principle

Linear 2nd-order elliptic BVP, inhomogeneous Dirichlet problem, variational form

\[ u \in \widetilde{g} + H^1_0(\Omega): \quad \int_{\Omega} \sigma \ \text{grad} \ u \cdot \text{grad} \ v \, dx = \int_{\Omega} f \ v \, dx \quad \forall \ v \in H^1_0(\Omega). \] (2.3.5)

(\widetilde{g} = \text{extension of Dirichlet data } g, \rightarrow \text{ Sect. 1.8})

- FE Galerkin discretization \iff finite difference scheme (based on mesh \( \mathcal{M} \))

- Choice of nodal global shape functions: unknowns \( \mu_i, \ i = 1, \ldots, N \), approximate \( u(p) \) at certain geometric mesh locations (e.g. nodes, midpoints of edges, etc.)

If \( g \in C^0(\partial \Omega) \) & \( f \equiv 0 \), we know (maximum principle \rightarrow \text{ Sect 1.4}):

\[ \min_{x \in \partial \Omega} g(x) \leq \min_{x \in \Omega} u(x) \leq \max_{x \in \Omega} u(x) \leq \max_{x \in \partial \Omega} g(x). \]

True also for discrete solution \( u_N \)?
Definition 2.3.1 (Discrete maximum principle). A local linear operator on grid function space satisfies the discrete maximum principle, if, for all its stencils,

(i) all their off-center weights are non-positive,
(ii) there is a strictly negative off-center weight,
(iii) the sum of all weights is non-negative.

\[ a_{ij} \leq 0 \quad \Leftrightarrow \quad i \neq j, \ 1 \leq i, j \leq N , \]
\[ \sum_{j=1}^{N} a_{ij} \geq 0 \quad \forall i = 1, \ldots, N , \]
\[ \forall i \in \{1, \ldots, N\}: \exists j \in \{1, \ldots, N\}: \ a_{ij} < 0 \]

\[ \Rightarrow \quad a_{ii} > 0 . \quad (2.3.7) \]

Example: homogeneous Dirichlet BVP for \(-\Delta\) on \(\Omega \subset \mathbb{R}^2\), piecewise linear Lagrangian finite elements (\(\rightarrow\) Sect 2.1.4), triangular mesh \(\mathcal{M}\) of \(\Omega \subset \mathbb{R}^2\)
element stiffness matrix (for triangle with interior angles $\omega_i, i = 1, 2, 3$):

$$A_K = \frac{1}{2} \begin{pmatrix} \cot \omega_3 + \cot \omega_2 & - \cot \omega_3 & - \cot \omega_2 \\ - \cot \omega_3 & \cot \omega_3 + \cot \omega_1 & - \cot \omega_1 \\ - \cot \omega_2 & - \cot \omega_1 & \cot \omega_2 + \cot \omega_1 \end{pmatrix}.$$  \hfill (2.2.8)

Entry $\mathbf{(A)}_{ij}$ of stiffness matrix (w.r.t. $\mathcal{H}_1^0(\mathcal{M})$) corresponding to nodes $\mathbf{p}_i, \mathbf{p}_j$, $1 \leq i, j \leq N := \# \mathcal{N}(\mathcal{M})$:

$$\mathbf{(A)}_{ij} = -\frac{1}{2} \frac{\sin \omega_{ij} + \sin \omega'_{ij}}{\sin \omega_{ij} \sin \omega'_{ij}}.$$  \hfill \text{Note:}

$$\sum_{j=1}^{N} \mathbf{(A)}_{ij} = 0 \quad \forall i = 1, \ldots, N.$$  \hfill (2.3.7) for $\mathbf{A}$

angle conditions:

- sum of angles facing interior edge $\leq \pi$,
- angles facing boundary edges $\leq \pi/2$.

(for non-Dirichlet boundary conditions)
Example of triangular meshes satisfying angle condition: Delaunay triangulations

(→ see help for MATLAB `delaunay` command for more explanations)

---

**Definition 2.3.2** (Structurally symmetric matrix). A matrix $A = (a_{ij}) \in \mathbb{R}^{N,N}$ is structurally symmetric, if

$$
a_{ij} \neq 0 \iff a_{ji} \neq 0 \quad \forall i, j \in \{1, \ldots, N\}.
$$

**Theorem 2.3.3.** If $A \in \mathbb{R}^{N,N}$ is structurally symmetric, regular and satisfies (2.3.7), then $A^{-1} \geq 0$ componentwise.

**Proof.** By permutation: $A$ in block-diagonal form with connected blocks

$$
A = \begin{pmatrix}
A_{11} & 0 & \cdots & 0 \\
0 & A_{22} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & A_{JJ}
\end{pmatrix}.
$$
Then, focus on a single connected diagonal block (call it $A = (a_{ij}) \in \mathbb{R}^{M,M}$): ($\rightarrow$ $A$ invertible!)

For $j \in \{1, \ldots, M\}$:

$$\tilde{\mu} \in \mathbb{R}^M: \quad A\tilde{\mu} = \tilde{\epsilon}_j := (\delta_{ij})_{i=1}^M$$

$k \in \{1, \ldots, M\}$: $\mu_k = \min_{i=1,\ldots,M} \mu_i$.

Assumption:

$$\mu_k < 0$$

$$0 \leq a_{kk}\mu_k + \sum_{j \neq k, a_{kj} \neq 0} a_{kj}\mu_j = a_{kk}\mu_k - \sum_{j \neq k, a_{kj} \neq 0} |a_{kj}|\mu_j \overset{(*)}{\leq} \mu_k \sum_{j=1}^M a_{kj} \leq 0.$$  

$$(\ast): \quad \text{"<", if } \mu_j > \mu_k \Rightarrow \mu_j = \mu_k, \text{ if } a_{kj} \neq 0.$$  

All other coefficients of $\tilde{\mu}$ connected with index $k$ have to be equal to $\mu_k$.

$(A \text{ connected})$  

$$\tilde{\mu} = \mu_k \mathbf{1} = (\mu_k, \ldots, \mu_k)^T$$

$$\mu_k \begin{bmatrix} A \\ \mathbf{1} \end{bmatrix} \leq 0 \quad \text{componentwise}.$$  

Contradicts $A\tilde{\mu} = \tilde{\epsilon}_j \Rightarrow \tilde{\mu} \geq 0$ componentwise.

Revisit: Inhomogeneous Dirichlet problem (2.3.5) with $f \equiv 0$, $\sigma \equiv 1$,  
FE Galerkin discretization by Lagrangian FEM, nodal global shape functions
Related linear system of equations (see Sect. 2.2.7 (2.2.17)):

\[
A = \begin{pmatrix}
A_{\Omega\Omega} & A_{\Gamma\Omega} \\
A_{\Omega\Gamma} & A_{\Gamma\Gamma}
\end{pmatrix} \Rightarrow A_{\Omega\Omega} \vec{\mu}_\Omega = \vec{\phi}_\Omega - A_{\Gamma\Omega} \vec{\mu}_\Gamma.
\]

\(\vec{\mu}_\Gamma\) = values \(g_j\) of Dirichlet data \(g\) in nodes \(\in \partial \Omega\).

\[
\sum_{j=1}^{N} (A)_{ij} = 0 \iff A \mathbf{1} = 0.
\]

Note:

\[
\vec{\mu}_\Gamma \leftarrow \vec{\mu}_\Gamma - \gamma \mathbf{1} \Rightarrow \vec{\mu}_\Omega \leftarrow \vec{\mu}_\Omega - \gamma \mathbf{1}.
\]

Note: Matrix \(A_{\Omega\Omega}\) regular (by Thm. 1.7.8 & Thm. 1.7.9)

Assumption: stiffness matrix \(A\) (no boundary conditions!) satisfies (2.3.7)

Apply Thm. 2.3.3 (<, >, ≤, ≥ to be understood in componentwise sense):

\[
\vec{\mu}_\Gamma \geq 0 \quad A_{\Gamma\Omega} \leq 0 \quad A_{\Gamma\Omega} \vec{\mu}_\Gamma \leq 0 \quad \text{Thm. 2.3.3} \quad \vec{\mu}_\Omega = -(A_{\Omega\Omega})^{-1} A_{\Gamma\Omega} \vec{\mu}_\Gamma \geq 0.
\]

\[
\vec{\mu}_\Gamma := \vec{\mu}_\Gamma - \min\{g_j\} \mathbf{1} \geq 0 \quad \Rightarrow \quad \vec{\mu}_\Omega := \vec{\mu}_\Omega - \mathbf{1} \min\{g_j\} \geq 0 \quad \Rightarrow \quad \vec{\mu}_\Omega \geq \min\{g_j\} \mathbf{1}.
\]

\[
\vec{\mu}_\Gamma := \max\{g_j\} \mathbf{1} - \vec{\mu}_\Gamma \geq 0 \quad \Rightarrow \quad \vec{\mu}_\Omega := \max\{g_j\} \mathbf{1} - \vec{\mu}_\Omega \geq 0 \quad \Rightarrow \quad \vec{\mu}_\Omega \leq \max\{g_j\} \mathbf{1}.
\]
\[ \min\{g_j\} \leq (\mu_{\Omega})_i \leq \max\{g_j\} \quad \forall i = 1, \ldots, N. \]

Nodal values of \( u_N \) satisfy (discrete) maximum principle!

### 2.3.3 Finite difference convergence theory

- **Perspective**: classical interpretation of elliptic BVP, \( u \in C^2(\Omega), \ f \in C^0(\Omega) \)
- **Functional framework**: space \( C^0(\mathcal{N}) \cong \mathbb{R}^N \) of grid functions \( f : \mathcal{N} \mapsto \mathbb{R}, \mathcal{N} = \) set of “nodes”
- **Targetted norm**: maximum norm \( \|\cdot\|_{\infty} \) on \( C^0(\mathcal{N}) \)
- **Tool**: pointwise restriction operator \( R : C^0(\Omega) \mapsto C^0(\mathcal{N}), \ (Ru)(p) := u(p) \).

**Reminder:**

The maximum (operator) norm of a matrix \( A \in \mathbb{R}^{N,N} \) is defined by

\[
\|A\|_{\infty} := \sup_{\xi \in \mathbb{R}^N \setminus \{0\}} \frac{\|A\xi\|_{\infty}}{\|\xi\|_{\infty}}.
\]
For $A = (a_{ij}) \in \mathbb{R}^{N,N}$: 
$$\|A\|_\infty = \max\{\sum_{j=1}^{N} |a_{ij}|, i = 1, \ldots, N\}.$$ 

Linear system arising from a nodal FD/FE discretization of 2nd-order linear elliptic BVP $\mathcal{L}(u) = f$ (+boundary conditions, quadrature):

$$A\vec{\mu} = \vec{\phi} := R f. \quad (2.3.8)$$

(Linear systems arising from Lagrangian FE (with nodal bases) require scaling to get (2.3.8))

**Theorem 2.3.4** (Lax’ theorem: consistency + stability $\Rightarrow$ convergence). If $u \in C^2(\Omega)$ is a classical solution of the BVP and $A$ regular, then

$$\|Ru - \vec{\mu}\|_\infty \leq \|A^{-1}\|_\infty \|ARu - R\mathcal{L}(u)\|_\infty .$$

**Proof.** By submultiplicativity of matrix norm:

$$\|Ru - \vec{\mu}\|_\infty \leq \|A^{-1}\|_\infty \|ARu - A\vec{\mu} + \vec{\phi} - R\mathcal{L}(u)\|_\infty = \|A^{-1}\|_\infty \|ARu - R\mathcal{L}(u)\|_\infty . \quad \Box$$
- term $A u - R \mathcal{L}(u)$ is called consistency error
- Condition $\|A^{-1}\|_\infty$ uniformly bounded in some discretization parameter: stability

**Theorem 2.3.5.** If $A \in \mathbb{R}^{N \times N}$ regular, $A^{-1} \geq 0$ componentwise, then

$$\forall \tilde{\xi} \in \mathbb{R}^N: \quad A \tilde{\xi} \geq 1 \implies \|A^{-1}\|_\infty \leq \|\tilde{\xi}\|_\infty.$$ 

**Proof.**

$$\|A^{-1}\|_\infty = \sup_{\|\tilde{\eta}\|_\infty = 1} \|A^{-1} \tilde{\eta}\|_\infty \leq \|A^{-1}\|_\infty \leq \|A^{-1} A \tilde{\xi}\|_\infty.$$ 

**Example 2.3.1 (Convergence of simple FD).**

$$-\Delta u = f \quad \text{in } \Omega := [0, 1]^2, \quad u = 0 \quad \text{on } \partial \Omega.$$ 

Discretization: 5-point stencil on tensor product grid

$\mathcal{N} = \{(ih, jh) \in \Omega, \ 1 \leq i, j \leq N, \ h := (N + 1)^{-1}\}$

Resulting $A$ satisfies discrete maximum principle

$\rightarrow$ Sect. 2.3.2, Def. 2.3.1
“Comparison function”  

\[ w(x) := \frac{1}{2} x_1 (1 - x_1) \Rightarrow A(Rw) \geq 1 \text{ pointwise.} \]

(by Thm. 2.3.5)  

\[ \| A^{-1} \|_\infty \leq \frac{1}{8} \]  

for any dimension \( N \)/meshwidth \( h \) (ie. “uniformly”)

Uniform stability of 5-point stencil for \(-\Delta\) independently of meshwidth of tensor product grid

Tackling consistency error:

\[
  \frac{1}{h^2} (4u(\xi, \eta) - u(\xi - h, \eta) - u(\xi + h, \eta) - u(\xi, \eta - h) - u(\xi, \eta + h)) = \]

\[
= -\frac{\partial^2 u}{\partial x_1^2} - \frac{\partial^2 u}{\partial x_2^2} + R(\xi, \eta, h), \]
with remainder

$$|R(\xi, \eta, h)| \leq \frac{1}{6} h^2 \max \left\{ \left\| \frac{\partial^4 u}{\partial x_1^4} \right\|_\infty, \left\| \frac{\partial^4 u}{\partial x_2^4} \right\|_\infty \right\} \quad \forall (\xi, \eta) \in \Omega, \; h > 0.$$ 

Consistency term for 5-point stencil for $-\Delta$, tensor product grid, $u \in C^4(\Omega)$:

$$\|A(Ru) + R\Delta u\|_\infty \leq \frac{1}{6} h^2 \max \left\{ \left\| \frac{\partial^4 u}{\partial x_1^4} \right\|_\infty, \left\| \frac{\partial^4 u}{\partial x_2^4} \right\|_\infty \right\} \quad \forall N \in \mathbb{N}.$$ 

Asymptotic convergence estimate:

$$\|Ru - \bar{\mu}\|_\infty \leq C h^2 \text{ with } C = C(\Omega, u) > 0 \text{ independent of meshwidth } h > 0.$$
2.4 Finite Volume Methods (FVM)

2.4.1 Principles of FVM

Targeted problem: linear 2nd-order elliptic boundary value problem in 2D \( \rightarrow \) Sect. [1.1]

\[- \text{div}(\sigma \, \text{grad} \, u) = f \quad \text{in} \, \Omega \, , \quad u = 0 \quad \text{on} \, \partial \Omega .\]

First ingredient: control volumes
Control volumes =
(polygonal) cells of a mesh $\tilde{\mathcal{M}} = \{C_i\}_i$
covering computational domain $\Omega$.

Associate cell $C_i \leftrightarrow$ nodal value $\mu_i$

Meaning: $\mu_i \approx u(p_i), p_i = \text{“center” of } C_i$

Second ingredient: numerical fluxes

Two adjacent cells $C_k, C_i$ with common edge $\Gamma_{ik} := \overline{C_i} \cap \overline{C_k}$.

Numerical flux

$$J_{ik} = \Psi(\mu_i, \mu_k) \approx \int_{\Gamma_{ik}} \mathbf{j} \cdot \mathbf{n}_{ik} \, dS$$

($\Psi = \text{numerical flux function, } \mathbf{j} = \text{flux, see (1.2.1)}$)
By Gauss’ theorem

\[
\int_{\partial C_i} \mathbf{j} \cdot \mathbf{n} \, dS = \int_{C_i} f \, dx \quad \Rightarrow \quad \sum_{k \in \mathcal{U}_i} J_{ik} = |C_i| f(p_i).
\]

\[\mathcal{U}_i := \{ C_j \in \tilde{\mathcal{M}}: C_i \text{ and } C_j \text{ share edge } \}, \quad p_i = \text{node associated with control volume } C_i.\]

System of equations \((\tilde{\mathcal{M}} := \tilde{\mathcal{M}} \text{ equations/unknowns } \mu_i):\)

\[
\sum_{k \in \mathcal{U}_i} \Psi(\mu_i, \mu_k) = |C_i| f(p_i) \, dx \quad \forall i = 1, \ldots, \tilde{\mathcal{M}}.
\]  \hspace{1cm} (2.4.1)

Note: homogeneous Dirichlet problem \(\Rightarrow\) only “interior” control volumes in (2.4.1)

**2.4.2 Dual meshes**

\(\rightarrow\) widely used approach to construction of control volumes

\(\rightarrow\) based on conventional FE triangulation \(\mathcal{M}\) of \(\Omega\).

(Here: triangular mesh of 2D polygon \(\Omega\))
\[ \mathcal{N}(\mathcal{M}) = \{ \mathbf{p}_1, \ldots, \mathbf{p}_M \} = \text{nodes of } \mathcal{M} \]

\[ C_i := \{ \mathbf{x} \in \Omega : |\mathbf{x} - \mathbf{p}_i| < |\mathbf{x} - \mathbf{p}_j| \ \forall j \neq i \} . \]

Voronoi dual mesh \( \tilde{\mathcal{M}} := \{ C_i \}_{i=1}^M \)

(\( \rightarrow \) MATLAB command \texttt{voronoi})

Angle condition (Delaunay triangulation) \( \Rightarrow \) \( \exists \) Voronoi dual mesh
Construction of Voronoi dual cells:
edges $\rightarrow$ perpendicular bisectors
nodes $\rightarrow$ circumcenters of triangles

straightforward generalization to 3D

Obtuse angle $\omega$: $\Rightarrow$ circumcenter $\not\in$ triangle

Fig. 40
Fig. 41

p. 153
2.4.3 From FVM to FEM

We consider: homogeneous Dirichlet problem for $\Delta$

$$-\Delta u = f \quad \text{in } \Omega \ , \quad u = 0 \quad \text{on } \partial \Omega$$
Finite volume method on Voronoi dual cells → Fig. 39:

\[ \mathcal{M} = \text{triangular mesh of } \Omega, \text{ only non-obtuse triangles} \quad \Rightarrow \quad \text{angle condition} \]

Number of control volumes = number of interior nodes of \( \mathcal{M} \)

Numerical flux

\[ J_{ik} := - \int_{\Gamma_{ik}} \nabla l_1 u \cdot n_{ik} \, dS, \]

\( l_1 u \in S^0_{1,0}(\mathcal{M}) = \mathcal{M}\text{-piecewise linear interpolant of nodal values } \mu_i \text{ (zero on } \partial \Omega) \)

\[ \nabla (l_1 u) \text{ piecewise constant on triangles of } \mathcal{M}! \]

\[ \sum_{k \in \mathcal{U}_i} |\Gamma_{ik}| \frac{\mu_i - \mu_k}{|p_i - p_k|} = \left( \sum_{k \in \mathcal{U}_i} \frac{|\Gamma_{ik}|}{|p_i - p_k|} \right) \mu_i - \sum_{k \in \mathcal{U}_i} \frac{|\Gamma_{ik}|}{|p_i - p_k|} \mu_k = |C_i| f(p_i) \quad \forall i = 1, \ldots, \tilde{M}. \]

Entry of system matrix \( A \) (\( k \in \mathcal{U}_i, i = 1, \ldots, N \)): \( a_{ik} = -\frac{|\Gamma_{ik}|}{|p_i - p_k|}, \quad a_{ii} = \sum_{k \in \mathcal{U}_i} \frac{|\Gamma_{ik}|}{|p_i - p_k|}. \)
Local perspective:

**Element stiffness matrix** for finite volume scheme

For triangle $K$

$$
\Gamma_{ik}^K := \Gamma_{ik} \cap K ,
$$

$$
J_{ik}^K (u) := - \int_{\Gamma_{ik}^K} \nabla l_1 u \cdot n_{ik} \, dS .
$$

$$
A_K = \begin{pmatrix}
J_{ij}^K (\lambda_i) + J_{ik}^K (\lambda_i) & J_{ij}^K (\lambda_j) & J_{ik}^K (\lambda_k) \\
-J_{ij}^K (\lambda_i) & -J_{ij}^K (\lambda_j) + J_{jk}^K (\lambda_j) & J_{jk}^K (\lambda_k) \\
-J_{ik}^K (\lambda_i) & -J_{jk}^K (\lambda_j) & -J_{ik}^K (\lambda_k) - J_{jk}^K (\lambda_k)
\end{pmatrix}
$$

Note: there holds $J_{ij}^K (\lambda_i) = -J_{ij}^K (\lambda_j) = J_{ji}^K (\lambda_j) \sim$ zero row/column sum
With
\[ b_i = \text{hat function at node } p_i, \]
\[ E_{ik} = \text{edge connecting } p_i, p_k, \]
\[ E_{ik}^i = [p_i, 1/2(p_i + p_k)]. \]

Then:
integration by parts (1.6.3) on \( K \),
Gauss' theorem on control volumes \( \cap K \).

\[
\int_K \text{grad} |_1 u \cdot \text{grad} b_i \, dx = \int_{\partial K} \text{grad} |_1 u \cdot n_{\partial K} b_i \, dS = \int_{E_{ik} \cup E_{ij}} \text{grad} |_1 u \cdot n_{\partial K} b_i \, dS
\]

\[
= \int_{E_{ik}^i \cup E_{ij}^i} \text{grad} |_1 u \cdot n_{\partial K} \, dS
\]

\[
= -\int_{\Gamma_{ij}^K} \text{grad} |_1 u \cdot n_{ij} \, dS - \int_{\Gamma_{ik}^K} \text{grad} |_1 u \cdot n_{ik} \, dS
\]

Same matrix as for linear Lagrangian FE on \( M \) \( \rightarrow \) Sect. \textbf{2.2.5}

- Barycentric dual mesh \( \rightarrow \) exercise
2.5 Finite Element Convergence Theory

Focus: Finite element Galerkin discretization of 2nd-order elliptic BVPs.

- Boundary value problems (heat conduction): Sect. 1.1
- Variational formulation: Sect. 1.6
- Some Sobolev spaces: Sect. 1.7
- Abstract Galerkin discretization: Sect. 2.1.1
- Lagrangian finite elements: Sects. 2.1.4, 2.1.5, 2.1.7

2.5.1 A priori error estimates

Linear variational problem (1.7.1) & its Galerkin discretization (→ Sect. 2.1.1):

\[ V_N \subset V = \text{discrete trial/test space, } \dim V_N < \infty \]
\( u \in V: \quad a(u, v) = f(v) \quad \forall v \in V \quad \Rightarrow \quad u_N \in V_N: \quad a(u_N, v_N) = f(v_N) \quad \forall v_N \in V_N. \)

A priori error estimates provide bounds for some norm of the discretization error \( u - u_N \) depending on

1. problem parameters
2. (the class of) problem data
3. discretization parameters

For elliptic boundary value problems (1.2.6) + \{(1.3.1), (1.3.2), (1.3.3)\}, finite element scheme:

<table>
<thead>
<tr>
<th>Problem parameters</th>
<th>= computational domain ( \Omega ), boundary parts ( \Gamma_D, \Gamma_N, \Gamma_R ), heat conductivity ( \sigma ), cooling coefficient ( q )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Problem data</td>
<td>= Source term ( f ), Dirichlet data ( g ), Neumann data ( h )</td>
</tr>
<tr>
<td>Discretization parameters</td>
<td>= mesh ( \mathcal{M} ) (( \rightarrow ) meshwidth, mesh parameters), type of ( (H^1(\Omega))-conforming) finite element (( \rightarrow ) polynomial degree)</td>
</tr>
</tbody>
</table>
\[
\begin{aligned}
\{ & a(\cdot, \cdot) \text{ continuous (Def. 1.7.6)} \\
& a(\cdot, \cdot) \text{ V-elliptic (Def. 1.7.7)} \\
& f(\cdot) \text{ continuous (Def. 1.7.5)} \}
\quad \Rightarrow \quad \left\{ \begin{array}{l}
\text{Existence \& uniqueness of } u, u_N \\
\text{Quasi-optimality of } u_N \\
\quad \|u - u_N\|_V \leq \frac{C_\Delta}{\gamma} \inf_{v_N \in V_N} \|u - v_N\|_V.
\end{array} \right.
\end{aligned}
\]

Rule of thumb: \( V_N \subset V \rightarrow V_N' \subset V \), \( V_N \subset V_N' \) ("refinement") \( \Rightarrow \) improved accuracy

How to achieve refinement of FE space ?

- h-refinement: replace \( M \) (for \( V_N \)) \( \rightarrow \) \( M' \) (for \( V_N' \))

*Example 2.5.1* (regular refinement of triangular mesh in 2D).
Regular refinement of triangle $K$ into four congruent triangles $T_1, T_2, T_3, T_4$. 

PSfrag replacements

$K$

$T_1$

$T_2$

$T_3$

$T_4$
For **h-refinement**: asymptotic a priori error estimates in terms of meshwidth
\[ h_M := \max\{\text{diam } K : K \in \mathcal{M}\} , \quad \text{diam } K := \max\{|p - q| : p, q \in K\} : \]
\[ \|u - u_N\|_V \leq C(\text{problem data, problem parameters, mesh parameters, FE type}) \epsilon(h_M) , \]
with function \( \epsilon : \mathbb{R}^+ \mapsto \mathbb{R}^+ , \epsilon(h) \to 0 \) for \( h \to 0 \).

**Definition 2.5.1** (Order of convergence for h-refinement). *In the case of **h-refinement**: convergence of order \( s \), \( s \in \mathbb{R}^+ : \Leftrightarrow \epsilon(h) = h^s \).*

- **p-refinement**: replace \( V_N := \delta_p^0(\mathcal{M}), p \in \mathbb{N} \) with \( V'_N := \delta_{p+1}^0(\mathcal{M}) \Rightarrow V_N \subset V'_N \)

For **p-refinement**: asymptotic a priori error estimates in terms of polynomial degree \( p \)
\[ \|u - u_N\|_V \leq C(\text{problem data, problem parameters, mesh } \mathcal{M}) \epsilon(p) , \]
with function \( \epsilon : \mathbb{N} \mapsto \mathbb{R}^+ , \epsilon(p) \to 0 \) for \( p \to \infty \).

**Combination of h-refinement and p-refinement?** OF COURSE (hp-refinement, [12])
General asymptotic a priori error estimate in terms of $N := \dim V_N$:

$$
\| u - u_N \|_V \leq C(\text{problem data}, \text{problem parameters}, \text{mesh parameters}, \text{FE family}) \Phi(N),
$$

with $\Phi : \mathbb{N} \mapsto \mathbb{R}^+$, $\Phi(N) \to 0$ for $N \to \infty$.

**Definition 2.5.2** (Convergence rate).

- $\Phi(N) = O(N^{-\alpha}), \alpha > 0$ :\(\iff\) algebraic convergence with rate $\alpha$
- $\Phi(N) = O(\exp(-\gamma N^\delta)), \gamma, \delta > 0$ :\(\iff\) exponential convergence
Linear plot of qualitative convergence behavior: algebraic/exponential convergence rates

Exponential convergence will always win (asymptotically)
Log-linear plot of decrease of discretization error for algebraic/exponential convergence rates
Log-log plot of decrease of discretization error for algebraic/exponential convergence rates
A priori error estimates, how?

**Quasi-optimality** of Galerkin solution ➤ a priori error estimates

How to estimate best approximation error: \[ \inf_{v_N \in V_N} \| u - v_N \|_V \]?

> Well, given solution \( u \) seek candidate function \( w_N \in V_N \) with

\[ \| u - w_N \|_V \approx \inf_{v_N \in V_N} \| u - v_N \|_V . \]

Natural choice: \( w_N \) by interpolation/averaging of \( u \)

BUT, exact solution \( u \) unknown?

> Known: \( u \) solves BVP + A priori information about problem parameters & problem data

\( u \) will belong to a certain class of functions (e.g. subspace \( S \subset V \))
2.5.2 The Sobolev scales

Recall: Interpolation error bound for equidistant polynomial interpolation

\[(f - q)(a + k \frac{b - a}{N}) = 0, \quad k = 0, \ldots, N := p \quad \land \quad q \in \mathcal{P}_p([a, b])\]

\[
\max_{a \leq t \leq b} |f(t) - q(t)| \leq \frac{1}{(p + 1)!} (b - a)^{p+1} \max_{a \leq t \leq b} |f^{(p+1)}(t)| .
\]

\[\to \text{derivatives of function } f : [a, b] \mapsto \mathbb{R} \text{ required: } f \in C^{p+1}([a, b]).\]

Interpolation estimates hinge on smoothness of interpolands!

How to measure smoothness of solution \( u \) of elliptic BVP ?

**Definition 2.5.3** (Higher order Sobolev norms). The \( m \)-th order Sobolev norm, \( m \in \mathbb{N}_0 \), for \( u : \Omega \subset \mathbb{R}^d \mapsto \mathbb{R} \) (sufficiently smooth) is defined by

\[
\|u\|_m^2 := \sum_{k=0}^{m} \sum_{\alpha \in \mathbb{N}^d, |\alpha| = k} \int_{\Omega} |D^\alpha u|^2 \, dx , \quad \text{where} \quad D^\alpha u := \frac{\partial |\alpha| u}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}.
\]
Sobolev space (by completion of $C^\infty(\overline{\Omega})$) \[ H^m(\Omega) := \{ v : \Omega \rightarrow \mathbb{R} : \| v \|_m < \infty \} \].

Sobolev spaces provide framework for variational formulation of elliptic BVP \((\rightarrow \text{Sect. 1.7})\)

Sobolev spaces provide norms $\| \cdot \|_m$ that measure smoothness of functions

Sobolev scale:
\[ \ldots \subset H^3(\Omega) \subset H^2(\Omega) \subset H^1(\Omega) \subset L^2(\Omega) \]

**Theorem 2.5.4** (Sobolev embedding theorem).

\[ m > \frac{d}{2} \Rightarrow H^m(\Omega) \subset C^0(\overline{\Omega}) \land \exists C = C(\Omega) > 0: \| u \|_\infty \leq C \| u \|_m \quad \forall u \in H^m(\Omega) . \]

Another Sobolev scale:
Definition 2.5.5. $W^{m,\infty}$-norm, $m \in \mathbb{N}_0$, for $u : \Omega \mapsto \mathbb{R}$ sufficiently smooth is defined by

$$\|u\|_{m,\infty} := \max_{0 \leq k \leq m} \max_{\alpha \in \mathbb{N}_0^d, |\alpha| = m} \sup_{x \in \overline{\Omega}} |D^\alpha u(x)|.$$ 

Sobolev scale

$$\ldots \subset W^{3,\infty}(\Omega) \subset W^{2,\infty}(\Omega) \subset W^{1,\infty}(\Omega) \subset W^{0,\infty}(\Omega) = L^\infty(\Omega)$$

Note:

$W^{m,\infty}(\Omega)$ is not a Hilbert space, but a Banach space (→ functional analysis).

2.5.3 The Bramble-Hilbert lemma

Example: 1D, $\Omega = ]0, 1[, l_1 : C^0([0, 1]) \mapsto P_1([0, 1])$ linear interpolation in $x = 0, 1$

Taylor expansion for $u \in C^2([0, 1])$

$$u(0) = u(x) - u'(x)x - \int_x^0 tu''(t) \, dt,$$

$$u(1) = u(x) + u'(x)(1 - x) - \int_x^1 (1 - t)u''(t) \, dt.$$
\[
\int_0^1 |u - l_1 u|^2 \, dx = \int_0^1 |u(x) - u(0) (1-x) - u(1)x|^2 \, dx \\
= \int_0^1 \left| (1-x) \int_0^x t u''(t) \, dt - x \int_x^1 (1-t) u''(t) \, dt \right|^2 \, dx \\
\leq \int_0^1 \left( (1-x)^2 \int_0^x t^2 \, dt \int_0^x |u''(t)|^2 \, dt + x^2 \int_x^1 (1-t)^2 \, dt \int_x^1 |u''(t)|^2 \, dt \right) \, dx \\
\leq \int_0^1 \left( \frac{2}{3} ( (1-x)^2 + x^2 ) \int_0^1 |u''(t)|^2 \, dt \right) \, dx \\
= \frac{4}{9} \int_0^1 |u''(t)|^2 \, dt .
\]

\[\|u - l_1 u\|_0 \leq \frac{2}{3} \left( \int_0^1 |u''(x)|^2 \, dx \right)^{1/2} \quad \forall u \in C^\infty([0, 1]) .\]
**Definition 2.5.6** (Sobolev semi-norm). The $m$-th order Sobolev semi-norm, $m \in \mathbb{N}$, for sufficiently smooth $u : \Omega \rightarrow \mathbb{R}$ is defined by

$$|u|^2_m := \sum_{\alpha \in \mathbb{N}^d, |\alpha| = m} \int_{\Omega} |D^\alpha u|^2 \, dx.$$

---

**Theorem 2.5.7** (Bramble-Hilbert lemma). If $\Omega \subset \mathbb{R}^d$ bounded, $p \in \mathbb{N}$, then

$$\exists C = C(\Omega, p) > 0: \inf_{q \in \mathcal{P}_p(\Omega)} \|u - q\|_{p+1} \leq C |u|_{p+1} \quad \forall u \in H^{p+1}(\Omega).$$

**Proof.** by compactness arguments (→ functional analysis).

"Polynomials of degree $p$ can absorb all terms of order $\leq p$ in Sobolev norm"

---

**Note:** Thm. 2.5.7 = generalization of 2nd Poincaré-Friedrichs inequality → Thm. [1.7.10]
2.5.4 Transformation techniques

\( \rightarrow \) cf. use of reference element for parametric construction of local shape function \( \rightarrow \) Sect. 2.1.6.

**Definition 2.5.8.** Given the local trial space \( \delta^0_p(K) \) (see Sects. 2.1.5, 2.1.7) \( (K = \text{cell of a mesh}) \) and nodes \( q_1, \ldots, q_Q \), \( Q := \dim \delta^0_p(K) \), whose associated point evaluation functionals provide local degrees of freedom (\( \rightarrow \) Def. 2.1.22) for \( \delta^0_p(K) \), the **nodal interpolation operator** \( I_p : C^0(\bar{K}) \mapsto \delta^0_p(K) \) on \( K \) is defined by

\[
(I_pu)(q_i) = u(q_i) \quad \forall i = 1, \ldots, Q, \quad \forall u \in C^0(\bar{K}).
\]

**Example 2.5.2.** \( K = \text{convex}\{a^1, a^2, a^3\} \) triangle, \( p = 1 \) \( \Rightarrow \) linear interpolation:

\[
1_1u|_K = u(a^1)\lambda_1 + u(a^2)\lambda_2 + u(a^3)\lambda_3.
\]

\( (\lambda, i = 1, 2, 3 = \text{barycentric coordinate functions, Def. 2.1.14}) \)

**Example 2.5.3.** \( K = \text{triangle convex}\{a^1, a^2, a^3\}, \ p = 2 \) \( \Rightarrow \) quadratic interpolation:

\[
l_2u|_K = - \sum_{i=1}^{3} \lambda_i(1 - 2\lambda_i)u(a^i) + \sum_{1 \leq i < j \leq 3} 4\lambda_i\lambda_j u\left(\frac{1}{2}(a^i + a^j)\right).
\]
Goal: Asymptotic interpolation error estimates for $l_p$, $p \in \mathbb{N}$:

$$\exists C > 0: \| u - l_p u \|_X \leq C \varepsilon(h) \| u \|_Y,$$

for suitable Sobolev (semi-)norms $\| \cdot \|_X$, $\| \cdot \|_Y$.

$h$ : meshwidth of underlying triangulation

$\varepsilon(h)$ : “convergence rate function”, $\varepsilon(h) \to 0$ for $h \to 0$

$C$ : constant depending on “controllable” mesh parameters.

Main steps of transformation techniques:

1. Localization: express $\| u - l_p u \|_X$ through local contributions of cells $K \in \mathcal{M}$.

2. Pullback $\Phi_K^*(u - l_p u)|_K$, $K \in \mathcal{M}$, to the reference cell $\hat{K}$, $K = \Phi_K(\hat{K})$: relationship

   $$\| \Phi_K^*(u - l_p u) \|_{X, \hat{K}} \sim \| u - l_p u \|_{X, K}.$$

3. Bramble-Hilbert argument on $\hat{K}$:

   $$\exists \gamma > 0 : \| \hat{u} - \hat{l}_p \hat{u} \|_{X, \hat{K}} \leq \gamma \| \hat{u} \|_{Y, \hat{K}} \quad \forall u,$$

   where $\hat{l}_p := \Phi_K^* \circ l_p|_K$.

4. Determine the impact of the transformation on the $Y$-(semi-)norm.
Chain of estimates: with constants $C_K, C'_K > 0$, “universal constant” $\hat{C} > 0$:

\[
\|u - l_p u\|_X^2 \leq \sum_{K \in \mathcal{M}} \| (u - l_p u) |_K \|_{X,K}^2 \\
\leq \sum_{K \in \mathcal{M}} C_K \| \Phi_K^*(u - l_p u) \|_{X,\hat{K}}^2 \\
\leq \hat{C} \sum_{K \in \mathcal{M}} C_K \| \Phi_K^*(u) \|_{Y,\hat{K}}^2 \\
\leq \hat{C} \sum_{K \in \mathcal{M}} C_K C'_K \|u\|_{Y,K}^2 .
\]

Need to know behavior of (local) norms under pullback!

Concrete case: affine transformations $\Phi_K : \hat{K} \mapsto K$, see Def. 2.1.18.

\[
\Phi_K(\hat{x}) := F\hat{x} + \tau, F \in \mathbb{R}^{d,d} \text{ regular, } \tau \in \mathbb{R}^d
\]
Lemma 2.5.9. If \( \Phi_K : \hat{K} \rightarrow K \) is an affine mapping \( \hat{x} \mapsto F\hat{x} + \tau \), then, for all \( m \in \mathbb{N}_0 \),

\[
|u|_{m,\hat{K}} \leq \left( \frac{m + d}{d} \right) d^m |F|^m |\det(F)|^{-1/2} |u|_{m,K} \quad \forall u \in H^m(K),
\]

\[
|u|_{m,K} \leq \left( \frac{m + d}{d} \right) d^m |F^{-1}|^m |\det(F)|^{1/2} |\hat{u}|_{m,\hat{K}} \quad \forall u \in H^m(\hat{K}),
\]

with \(|F|\) denoting the matrix norm of \( F \) associated with the Euclidean vector norm.

Proof (for \( m = 1 \)).

chain rule \( \Rightarrow \)

\[
\text{grad}_{\hat{x}}(u \circ \Phi_K)(\hat{x}) = (D \Phi_K(\hat{x}))^T \text{grad}_{\hat{x}} u(\Phi_K(\hat{x})) \quad \forall \hat{x} \in \hat{K}.
\]

Note: \( D \Phi_K = F, \det(D \Phi_K) = \det F \neq 0 \).

\[
\int_{\hat{K}} |\text{grad}_{\hat{x}} \hat{u}|^2 d\hat{x} = \int_{\hat{K}} |(D \Phi_K(\hat{x}))^T \text{grad}_{\hat{x}} u(\Phi_K(\hat{x}))|^2 d\hat{x}
\]

\[
= \int_{\hat{K}} |F^T \text{grad}_{\hat{x}} u(\Phi_K(\hat{x}))|^2 d\hat{x}
\]

\[
(\ast) \int_{K} |F^T \text{grad}_{\hat{x}} u(\hat{x})|^2 |\det F|^{-1} d\hat{x}
\]

\[
\leq \int_{K} |F|^2 |\text{grad}_{\hat{x}} u|^2 |\det F|^{-1} d\hat{x}.
\]

\( \ast \) = application of transformation formula for integrals.
Role reversal of $K$, $\hat{K}$ ⇒ other estimate.

$|\mathbf{F}|$, $|\det \mathbf{F}|$: elusive quantities? NO ⇒ determined by geometry of cell $K$

**Definition 2.5.10.** For cell $K \in \mathcal{M}$ define its **diameter**

$$h_K = \text{diam } K := \sup\{|\mathbf{x} - \mathbf{y}|, \ \mathbf{x}, \mathbf{y} \in K\},$$

and the **maximum radius of an inscribed ball**

$$r_K := \sup\{r > 0 : \exists \mathbf{x} \in K : |\mathbf{x} - \mathbf{y}| < r \Rightarrow \mathbf{y} \in K\}.$$

Ratio $h_K / r_K = \text{shape regularity measure } \rho_K$ of $K$.

For triangle $K$: $\rho_K$ large $\Leftrightarrow$ $K$ “distorted”

Now: Focus on simplicial cells (triangles, tetrahedra)
Lemma 2.5.11. If the smallest angle of a triangle $K$ is bounded from below by $\alpha > 0$, then

$$\sin(\alpha/2)^{-1} \leq \rho_K \leq 2\sin(\alpha/2)^{-1}. $$

Proof. 

see figure $\rightarrow$

$$\frac{1}{2}h_K \sin(\alpha/2) \leq l \sin(\alpha/2) = r_K \leq h_K \sin(\alpha/2).$$

$\square$
Lemma 2.5.12. If $\hat{K}$, $K \subset \mathbb{R}^d$, $d = 2, 3$, are generic non-degenerate simplices and $\Phi_K : \hat{K} \mapsto K$, $\Phi_K(\hat{x}) := F\hat{\xi} + \tau$, the associated bijective affine mapping, then

$$
\left(\frac{h_k}{h_{\hat{k}}}\right)^d \rho_k^{1-d} = \frac{h_K r_k^{d-1}}{h_{\hat{k}}^d} \leq |\det(F)| = \frac{|K|}{|\hat{K}|} \leq \frac{h_K^d}{h_{\hat{k}}^d} \left(\frac{h_{\hat{k}}}{h_k}\right)^d \rho_{\hat{k}}^{d-1}, \tag{2.5.1}
$$

$$
|F| \leq \frac{h_{\hat{k}}}{2r_{\hat{k}}} = \frac{1}{2} \rho_{\hat{k}} \frac{h_k}{h_{\hat{k}}}, \quad |F^{-1}| \leq \frac{h_{\hat{k}}}{2r_K} = \frac{1}{2} \rho_k \frac{h_k}{h_{\hat{k}}}. \tag{2.5.2}
$$

Proof. \quad |\det F| = \frac{|K|}{|\hat{K}|} (\text{ratio of volumes}) \rightarrow (2.5.1) \text{ trivial}

For all directions $\hat{z} (|\hat{z}| = 1)$ we can find $\hat{x}, \hat{y}$ on the boundary of the largest inscribed ball of $\hat{K}$ such that $\hat{x} - \hat{y} = 2r_{\hat{k}} \hat{z}$. Thus, (2.5.2) from

$$
|F| = \sup\{|F\hat{z}|, \ |\hat{z}| = 1\} = \frac{1}{2} r_{\hat{k}}^{-1} \sup\{|\Phi(\hat{x}) - \Phi(\hat{y})|, \ |\hat{x} - \hat{y}| = 2r_{\hat{k}}\} \leq h_K / 2r_{\hat{k}},
$$

because $\Phi_K(\hat{x}), \Phi_K(\hat{y}) \in K$. \hfill \Box

---

uniform shape regularity $\Rightarrow$ control of change of norms under $\Phi_K^*$. 
Definition 2.5.13 (Mesh parameters). With notations from Def. 2.5.10:

- **meshwidth** \( h_{\mathcal{M}} := \max \{ h_K, \ K \in \mathcal{M} \} \),
- **shape regularity measure** \( \rho_{\mathcal{M}} := \max \{ \rho_K, \ K \in \mathcal{M} \} \),
- **quasi-uniformity measure** \( \mu_{\mathcal{M}} := \max \{ h_K / h_{K'}, \ K, K' \in \mathcal{M} \} \).

Standard choice of reference simplices:

For \( d = 2 \):

\[
\hat{K} := \text{convex } \left\{ \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\} \quad \text{for } d = 2 ,
\]

(2.5.3)

For \( d = 3 \):

\[
\hat{K} := \text{convex } \left\{ \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\} \quad \text{for } d = 3 .
\]

(2.5.4)
Corollary 2.5.14. \( \mathcal{M} = \) simplicial triangulation, \( \hat{\mathcal{K}}\) according to (2.5.3) and (2.5.4), respectively. Then the affine mappings \( \Phi_K : \hat{\mathcal{K}} \mapsto K, \Phi_K(\hat{\xi}) := \mathbf{F}_K \hat{\xi} + \tau_K, K \in \mathcal{M}, \) satisfy

\[
\frac{1-d}{\mu_{\mathcal{M}}} h_{\mathcal{M}}^d \leq |\det(\mathbf{F}_K)| \leq c h_{\mathcal{M}}^d, \quad |\mathbf{F}_K| \leq c h_{\mathcal{M}}, \quad |\mathbf{F}_K^{-1}| \leq c \rho_{\mathcal{M}} \mu_{\mathcal{M}} h_{\mathcal{M}}^{-1},
\]

with universal constant \( c > 0. \)

2.5.5 Interpolation error estimates

General formula for nodal Lagrangian interpolation operator \( I_p : C^0(\overline{\Omega}) \mapsto \mathcal{S}^0_p(\mathcal{M}) : \)

\[
\forall K \in \mathcal{M}: \quad I_p u|_K = \sum_{i=1}^{Q} u(\mathbf{q}_i^K) b_i^K, \quad (2.5.5)
\]

\( \mathbf{q}_i^K, i = 1, \ldots, Q = \) local interpolation nodes in cell \( K \in \mathcal{M}, \) see Sect. 2.1.5.

\( b_i^K, i = 1, \ldots, Q = \) local shape functions: \( b_i^K(\mathbf{q}_j^K) = \delta_{ij}. \)
I_p local: \[ u \in C^0(\overline{\Omega}), \text{supp } u \cap K = \emptyset \implies I_p u|_K = 0. \]

Notation: Local nodal interpolation operators: \[ I^K_p : C^0(\overline{K}) \mapsto \delta^0_p(K). \]

Consider: parametric Lagrangian FE \( \delta^0_p(\mathcal{M}) \), reference cell \( \hat{K} \),
diffeomorphisms \( \Phi_K : \hat{K} \mapsto K, K \in \mathcal{M} \):

Notation: Nodal interpolation operator on reference cell: \( \hat{I}_p : C^0(\overline{K}) \mapsto \delta^0_p(\hat{K}) \).

\textbf{Lemma 2.5.15} (Pullback and nodal interpolation commute).

\[ C^0(\overline{K}) \xrightarrow{I^K_p} \delta^0_p(K) \]
\[ \Phi^* \downarrow \quad \downarrow \Phi^* \]
\[ C^0(\overline{K}) \xrightarrow{\hat{I}_p} \delta^0_p(\hat{K}) \]

Lemma \textbf{2.5.15} \iff \[ \Phi^*_K(I^K_p u) = \hat{I}_p(\Phi^*_K u) \quad \forall u \in C^0(\overline{K}). \]
Proof. By parametric definition of local shape functions on $K$, see Sect. 2.1.6,

$$\Phi_K^*(l_p^K u) = \sum_{i=1}^{Q} u(q_i^K) \Phi_K^* b_i^K = \sum_{i=1}^{Q} (\Phi_K^* u)(\hat{q}_i^K) \hat{b}_i.$$ 

**Theorem 2.5.16.** On simplicial mesh $M$ nodal interpolation operators $l_p : C^0(\Omega) \mapsto \delta_p^0(M)$ satisfy for $2 \leq k \leq p + 1, m = 0, 1$

$$\exists C = C(k, m, \rho_M): \quad |u - l_p u|_m \leq C h_{\rho_M}^{k-m} |u|_k \quad \forall u \in H^k(\Omega).$$

Proof (For $k = 2, m = 1$): use transformation techniques and Bramble-Hilbert arguments

Pick $K \in \mathcal{M}, u \in H^2(K) \Rightarrow u \in C^0(\overline{K}),$ see Thm. 2.5.4

$\mathcal{M} = \text{triangular/tetrahedral mesh} \Rightarrow \text{mappings } \Phi_K \text{ affine, } \Phi_K(\tilde{x}) = F\tilde{x} + \tau, \text{ see Def. 2.1.18} \ $
Lemma 2.5.9 & Lemma 2.5.12 applicable:

\[
\left| u - l_p^K u \right|_{1,K} \leq \left( \frac{1 + d}{d} \right) d |F^{-1}| |\text{det} F|^{1/2} \left| \Phi^*_K (u - l_1^K) \right|_{1,\hat{K}}
\]

\[
= \left( \frac{1 + d}{d} \right) d |F^{-1}| |\text{det} F|^{1/2} |\hat{u} - \hat{I}_1 \hat{u}|_{1,\hat{K}}
\]

\[
\leq C \left( \frac{1 + d}{d} \right) d |F^{-1}| |\text{det} F|^{1/2} \inf_{q \in P_1(\hat{K})} |(I_d - \hat{I}_1)(\hat{u} - q)|_{1,\hat{K}}
\]

\[
\leq C C_{BH} \left( \frac{1 + d}{d} \right) d |F^{-1}| |\text{det} F|^{1/2} |\hat{u}|_{2,\hat{K}}
\]

\[
\leq C C_{BH} \left( \frac{1 + d}{d} \right) \left( \frac{2 + d}{d} \right) d^3 |F^{-1}| |F|^{2} |u|_{2,K}
\]

\[
\leq C (\rho_K h_K^{-1}) h_K^2 |u|_{2,K} = C \rho_K h_K |u|_{2,K}.
\]

\(1\) Lemma 2.5.9 for \(m = 1\).

\(2\) Lemma 2.5.15.

\(3\) \(\hat{I}_1 = \text{projection onto } P_1(\hat{K}) \) (→ Def. 2.1.3).
Sobolev embedding Thm. 2.5.4: \( \| \hat{u} \|_\infty \leq C_1 \| \hat{u} \|_{2, \hat{K}} \),
and continuity \( \| I_1 \hat{u} \|_{2, \hat{K}} \leq C_2 \| \hat{u} \|_\infty \)

\[ C_1 := C_1 C_2. \]

Lemma 2.5.17. If \( \dim V < \infty \) and \( \| \cdot \|_X, \| \cdot \|_Y \) norms on \( V \) (→ Def. 1.7.2), then

\[ \exists C, \overline{C} > 0: \ C \| v \|_X \leq \| v \|_Y \leq \overline{C} \| v \|_X \quad \forall v \in V. \]

Bramble-Hilbert lemma Thm. 2.5.7.

Lemma 2.5.9 for \( m = 2 \).

Lemma 2.5.12.

Finally, take squares and sum over all cells.

Moreover, Estimates of Thm. 2.5.16 not sharp

anisotropic interpolation estimates [2]

Triangular cells with “bad shape regularity” (\( \rho_K \) “large”): very small/large angles

However: Estimates of Thm. 2.5.16 not sharp

anisotropic interpolation estimates [2]
Inspect approximate interpolation constants

\[ C_K := \sup_{u \in H^2(K)} \frac{\|u - l_1 u\|_{1,K}}{\|u\|_{2,K}}, \]
\[ C_{K,2} := \sup_{u \in H^2(K)} \frac{\|u - l_1 u\|_{0,K}}{\|u\|_{2,K}}, \]

with triangle \( K := \text{convex} \left\{ (0,0), (1,0), (p_x, p_y) \right\}. \)

Sampling the space of triangles

(modulo similarity)

\[ 0 \leq p_x, p_y \leq 1. \]

+ Numerical computation of \( C_K, C_{K,2} \)

(implementation by A. Inci)
Example 2.5.4 (Linear interpolation on acute and obtuse triangles).
triangle $K := \text{convex } \left\{ \left( \begin{array}{c} 0 \\ 0 \\ \frac{1}{h} \end{array} \right), \left( \begin{array}{c} 1 \\ 0 \\ 0 \end{array} \right) \right\}$, $h > 0$, $u(x, y) = x(1 - x)$, $0 < x < 1$.

\[ \|u\|_{2, K}^2 = \frac{3031}{1440} h, \quad \|u - l_1 u\|_{1, K}^2 = \frac{29}{2880} h + \frac{1}{12} h + \frac{1}{32} h^{-1}, \quad \|u - l_1 u\|_{0, K}^2 = \frac{29}{2889} h \]

\[ C_K^2 \geq \frac{\|u - l_1 u\|_{1, K}^2}{\|u\|_{2, K}^2} \geq \frac{269}{6062} + \frac{45}{3031} h^{-2}, \quad \frac{\|u - l_1 u\|_{0, K}^2}{\|u\|_{2, K}^2} = \frac{29}{6062}. \]

\[ l_1 \rightarrow \text{good approximation in } H^1(K)-\text{norm} \quad l_1 \rightarrow \text{bad approximation in } H^1(K)-\text{norm} \]

Watch out for obtuse angles!
However:
\[ \inf_{v_N \in S^0_1(M)} \| u - v_N \|_1 \ll \| u - l_p u \|_1 \] is possible!

Elementary cell of “bad mesh” \( M_{bad} \)

Elementary cell of “good mesh” \( M_{good} \)

On “bad” mesh:
\[ \sup_{u \in H^2(\Omega)} \frac{\| u - 11u \|_1}{\| u \|_2} \rightarrow \infty \quad \text{as} \quad h/\delta \rightarrow \infty, \]

On “good” mesh:
\[ \sup_{u \in H^2(\Omega)} \frac{\| u - 11u \|_1}{\| u \|_2} \] uniformly bounded in \( h/\delta \).

Yet,
\[ \inf_{v_N \in S^0_1(M_{bad})} \| u - v_N \|_1 \leq \inf_{v_N \in S^0_1(M_{good})} \| u - v_N \|_1 \quad \forall u \in H^2(\Omega). \]

### 2.5.6 Elliptic regularity theory

We consider: scalar second order elliptic BVP, see Sect. 1.1, (1.2.6) + boundary conditions
Question: What can we say about smoothness of solutions of 2nd-order elliptic BVPs?

Example: 1D, $\Omega = ]0, 1[$, coefficient $\sigma \equiv 1$, homogeneous Dirichlet boundary conditions

$$u'' = f, \quad u(0) = u(1) = 0.$$  

Obvious:

$$f \in H^k(\Omega) \Rightarrow u \in H^{k+2}(\Omega) \quad \text{(a lifting theorem)}$$

Generalization to higher dimensions ($\Omega \subset \mathbb{R}^d$)?

**Theorem 2.5.18** (Smooth elliptic lifting theorem). If $\partial \Omega$ is $C^\infty$-smooth, ie. possesses a local parameterization by $C^\infty$-functions, and $\sigma \in C^\infty(\overline{\Omega})$, then, for any $k \in \mathbb{N}$,

$$u \in H^1_0(\Omega) \quad \text{and} \quad -\text{div}(\sigma \, \text{grad} \, u) \in H^k(\Omega)$$

$$u \in H^1(\Omega), \quad -\text{div}(\sigma \, \text{grad} \, u) \in H^k(\Omega), \quad \text{and} \quad \text{grad} \, u \cdot n = 0 \text{ on } \partial \Omega$$

$$\Rightarrow u \in H^{k+2}(\Omega).$$

Remember: statements about Sobolev spaces $\equiv$ statements about their norms
In Thm. 2.5.18: \( u \in H^{k+2}(\Omega) \) means \( \exists C = C(\Omega, \sigma) > 0: \| u \|_{k+2} \leq C \| -\text{div}(\sigma \text{grad } u) \|_k \)

What about non-smooth \( \partial \Omega \), discontinuous \( \sigma \)?

2D, \( \Omega \) polygon:
- corners \( c_i \)
- angles \( \omega_i \)

\[ u \in H^1_0(\Omega): \Delta u = f \in C^\infty(\overline{\Omega}) \]

How will corners affect smoothness of \( u \)?
Local considerations at corner $c^i$:

- Introduce polar coordinates $(r, \phi)$ at $c^i$
- Separation of variables ansatz

\[ u(r, \phi) = u_r(r)u_\phi(\phi) . \]

\[ u = 0 \text{ on } \partial \Omega \Rightarrow u_\phi(0) = u_\phi(\omega_i) = 0 . \]

In polar coordinates:

\[ \Delta u = \frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial u}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2 u}{\partial \phi^2} \]

\[ \Delta(u_r(r)u_\phi(\phi)) = 0 \Leftrightarrow u(r, \phi) = r^\lambda(\kappa \sin(\lambda \phi) + \nu \cos(\lambda \phi)) , \ \lambda > 0, \kappa, \nu \in \mathbb{R} . \]

[+ boundary condition $u_{\mid \partial \Omega} = 0$]

\[ u(r, \phi) = r^{\lambda_{ik}} \sin(\lambda_{ik}\phi) , \ \lambda_{ik} = \frac{k\pi}{\omega_i} , k \in \mathbb{N} . \]  

(2.5.7)

Terminology:

$u$ from (2.5.7) is called Dirichlet corner singular function for $\Delta$ at $c^i$. 
Corner singular functions locally (in a neighborhood of $c^i$) satisfy the homogeneous PDE and boundary conditions.

Singular functions for $k = 1, 2, 3$ on L-shaped domain ($\omega = 3\pi/2$).
Theorem 2.5.19 (Corner singular function decomposition). Let $\Omega \subset \mathbb{R}^2$ be a polygon with $J$ corners $c^i$. Denote the polar coordinates in the corner $c^i$ by $(r_i, \phi_i)$ and the inner angle at the corner $c^i$ by $\omega_i$. Additionally, let $f \in H^1(\Omega)$ with $l \in \mathbb{N}_0$ and $l \neq \lambda_{ik} - 1$, where the $\lambda_{ik}$ are given by the singular exponents

$$\lambda_{ik} = \frac{k\pi}{\omega_i} \quad \text{for } k \in \mathbb{N}.$$  \hfill (2.5.8)

Then $u \in H^1_0(\Omega)$ with $-\Delta u = f$ in $\Omega$ can be decomposed

$$u = u^0 + \sum_{i=1}^{J} \psi(r_i) \sum_{\lambda_{ik} < l+1} \kappa_{ik} s_{ik}(r_i, \phi_i), \quad \kappa_{ik} \in \mathbb{R},$$ \hfill (2.5.9)

with regular part $u^0 \in H^{l+2}(\Omega)$, cut-off functions $\psi \in C^\infty(\mathbb{R}^+) \ (\psi \equiv 1 \text{ in a neighborhood of } 0)$, and corner singular functions

$$\begin{align*}
\lambda_{ik} \notin \mathbb{N}: & \quad s_{ik}(r, \phi) = r^{\lambda_{ik}} \sin(\lambda_{ik} \phi), \\
\lambda_{ik} \in \mathbb{N}: & \quad s_{ik}(r, \phi) = r^{\lambda_{ik}} (\ln r) \sin(\lambda_{ik} \phi).
\end{align*}$$ \hfill (2.5.10)

Why should corner singular functions bother us?

Radial behavior:

$$|s_{ik}| = O(r^{\lambda_{ik}}) \quad \Rightarrow \quad |D^\alpha s_{ik}| = O(r^{\lambda_{ik} - |\alpha|}), \quad \alpha \in \mathbb{N}^2.$$
First two corner singular functions \( s_{i1}, s_{i2} \) for \( \omega_i = \frac{3\pi}{2} \): 

- Re-entrant corner
- Qualitative radial behavior

\[ \frac{\partial^2}{\partial r^2} s_{i1} \notin L^2(\Omega)! \]

**Theorem 2.5.20** (Sobolev regularity of corner singular functions). 

\[ l > \lambda_{ik} := \frac{k\pi}{\omega_i} \notin \mathbb{N} \quad \Rightarrow \quad s_{ik} \notin H^{l+1}(\Omega), \]

for corner singular functions \( s_{ik} \) from (2.5.10).
If $\omega_i > \pi \Rightarrow s_{i1} \notin H^2(\Omega)$.

$\Omega \subset \mathbb{R}^2$ has re-entrant corners $\Rightarrow$ Solution of homogeneous Dirichlet problem for $-\Delta \notin H^2(\Omega)$ in general.

Theorem 2.5.21. If $\Omega \subset \mathbb{R}^d$ convex, $u \in H^1_0(\Omega)$, $\Delta u \in L^2(\Omega) \Rightarrow u \in H^2(\Omega)$.

Terminology: if conclusion of Thm. 2.5.21 true $\rightarrow$ Dirichlet problem 2-regular.

Other causes for poor regularity of solution $u$: 
• Discontinuities of $\sigma$
  $\rightarrow$ singular functions at “material corners”,
• Mixed boundary conditions ($\rightarrow$ exercises),
• Non-smooth source function $f$.

2.5.7 Convergence of finite element solutions

Model problem: heat conduction in homogeneous body, (homogeneous) Dirichlet boundary conditions (Poisson equation)

$$-\Delta u = f \in L^2(\Omega) \quad \text{in } \Omega, \quad u = g \in C^0(\partial \Omega) \quad \text{on } \partial \Omega.$$
Example 2.5.5 (FE convergence for Poisson equation).

Setting:
\[ \Omega = ]0, 1[^2, \ f(x, y) = 2\pi^2 \sin(\pi x) \sin(\pi y), \ (x, y)^T \in \Omega \]
\[ u(x, y) = \sin(\pi x) \sin(\pi y), \ g = 0. \]

- Galerkin finite element discretization:
  - linear Lagrangian finite elements, \( V_N = \mathcal{S}^0_{1,0}(\mathcal{M}) \subset H^1_0(\Omega) \) (\( \rightarrow \) Sect. 2.1.4),
  - quadratic Lagrangian finite elements, \( V_N = \mathcal{S}^0_{2,0}(\mathcal{M}) \subset H^1_0(\Omega) \) (\( \rightarrow \) Sect. 2.1.5).

- Quadrature rule (2.2.14) for assembly of local load vectors (\( \rightarrow \) Sect. 2.2.6).

Monitoring the discretization error:

Approximate (*) computation of \( \|u - u_N\|_1 \) and \( \|u - u_N\|_0 \) on a sequence of meshes (created by successive regular refinement (\( \rightarrow \) Sect. 2.5.1) of coarse initial mesh).

(*): use of quadrature rule (on FE mesh)

**Numerical quadrature must** match expected order of convergence!

(If \( V_N = \mathcal{S}^0_{p,0}(\mathcal{M}) \), use quadrature rule of order \( 2p - 1 \))

\( \rightarrow \) Quadrature rule (2.2.14) “safe best” (for this example).
Unstructured triangular meshes of $\Omega = [0, 1]^2$ (two coarsest specimens)
\( H^1(\Omega) \)-semi-norm of discretization error on unit square \((- \leftrightarrow p = 1, - \leftrightarrow p = 2)\)

Observations:
- Algebraic rates of convergence in terms of \( N \) and \( h \)
- Quadratic Lagrangian FE converge with double the rate of linear Lagrangian FE
Discretization errors with respect to $L^2$ norm

$L^2(\Omega)$-norm of discretization error on unit square ($\leftrightarrow p = 1$, $\leftrightarrow p = 2$)

Observations:
- Linear Lagrangian FE ($p = 1$) \[ \| u - u_N \|_0 = O(N^{-1}) \]
- Quadratic Lagrangian FE ($p = 2$) \[ \| u - u_N \|_0 = O(N^{-1.5}) \]

Example 2.5.6 (FE-convergence on L-shaped domain).

Setting:
\[ \Omega = ] -1, 1[^2 \setminus ]0, 1[ \times ]-1, 0[ \), \( u(r, \phi) = r^{2/3} \sin(2/3\phi) \) (polar coordinates)
\[ f = 0, g = u|_{\partial \Omega} \]
\( (u = \text{corner singular function} \notin H^2(\Omega)) \)

Same evaluations as above:
Unstructured triangular meshes of $\Omega = [-1, 1[ \times [0, 1[ \times [-1, 0[)$ (two coarsest specimens)
$H^1(\Omega)$-semi-norm of discretization error on “L-shaped” domain ($\leftrightarrow p = 1, \leftrightarrow p = 2$)

Observations:

- For both $p = 1, 2$: $\|u - u_N\|_1 = O(N^{-1/3})$
- No gain from higher polynomial degree

Consider: scalar linear elliptic BVP, exact solution $u \in H^1(\Omega)$, $\Omega \subset \mathbb{R}^2$

Galerkin FE discretization based on $V_N := S_p^0(\mathcal{M})$ on triangular meshes $\mathcal{M}$.

Goal; Asymptotic a priori error estimates in terms of $N := \dim S_p^0(\mathcal{M})$: 
For triangular meshes $\mathcal{M}$: by Lemma 2.1.16
\[
\dim \delta_p^0(\mathcal{M}) = \#\{\text{vertices}(\mathcal{M})\} + \#\{\text{edges}(\mathcal{M})\} (p - 1) + \#\mathcal{M} \frac{1}{2}(p - 1)(p - 2) .
\]

Lemma 2.5.11 \[\exists C = C(\rho_M): \#\{K_j \in \mathcal{M}: \overline{K_i} \cap \overline{K_j} \neq \emptyset\} \leq C \quad (i = 1, 2, \ldots, \#\mathcal{M}) .\]

\[\#\{\text{vertices}(\mathcal{M})\}, \#\{\text{edges}(\mathcal{M})\} \approx \#\mathcal{M} .\]

\[
\dim \delta_p^0(\mathcal{M}) \approx (\#\mathcal{M})p^2 ,
\]
with constants depending on $\rho_M$.

Consider meshes with “all cells of about the same size”
(cf. quasi-uniformity measure $\mu_M \rightarrow$ Def. 2.5.13)

For $d = 2$: \[\exists C = C(\Omega, \rho_M, \mu_M): \#\mathcal{M} \leq C h^{-2}_M .\] (2.5.12)

Example: sequence of triangular meshes created by regular refinement

Interpolation error estimate Thm. 2.5.16 & quasi-optimality of FE solution $u_N$: If $u \in H^{p+1}(\Omega)$

\[\exists C = C(\Omega, \rho_M, \mu_M): \|u - u_N\|_1 \leq C h^p_M \|u\|_{p+1} \leq C (\#\mathcal{M})^{-p/2} \leq C N^{-p/2} .\]
\[ \|u - u_N\|_1 = O(N^{-p/2}) \, , \quad N := \dim \mathcal{S}_p(M) \, . \]

In the case of \textit{h-refinement}: if solution \( u \) “very smooth”, then raising polynomial degree \( p \) improves asymptotic convergence in terms of \( N \).

### 2.5.8 Variational crimes

= replacing (exact) \textbf{discrete variational problem}

\[ u_N \in V_N: \ a(u_N, v_N) = f(v_N) \quad \forall v_N \in V_N \, , \quad (2.1.1) \]

by perturbed variational problem

\[ \tilde{u}_N \in V_N: \ a_N(\tilde{u}_N, v_N) = f_N(v_N) \quad \forall v_N \in V_N \, . \quad (2.5.13) \]

Approximations \( a_N(\cdot, \cdot) \approx a(\cdot, \cdot), \ f_N(\cdot) \approx f(\cdot) \) due to
Approximation $a \to a_N$, $f \to f_N$ → additional consistency error.

Guideline:

<table>
<thead>
<tr>
<th>Consistency error</th>
<th>≈ discretization error of (2.1.1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asymptotic behavior</td>
<td>Asymptotic behavior of discretization error of unperturbed FE scheme</td>
</tr>
</tbody>
</table>

2.5.8.1 Abstract theory

Linear variational problem on Hilbert space $V$:

$$u \in V: \quad a(u, v) = f(v) \quad \forall v \in V.$$  \hspace{2cm} (1.7.1)
Galerkin discretization using \( V_N \subset V \) ➞ discrete variational problem \( \rightarrow \) Sect. 2.1.1

\[
\forall u_N \in V_N : \quad a(u_N, v_N) = f(v_N) \quad \forall v_N \in V_N .
\] (2.1.1)

Approximation \( a \rightarrow a_N, f \rightarrow f_N \) ➞ perturbed discrete variational problem

\[
\forall u_N \in V_N : \quad a_N(u_N, v_N) = f_N(v_N) \quad \forall v_N \in V_N .
\] (2.5.13)

\( a_N : V_N \times V_N \rightarrow \mathbb{R} \) : continuous (\( \rightarrow \) Def. 1.7.6), bilinear form.

\( f_N : V_N \rightarrow \mathbb{R} \) : continuous (\( \rightarrow \) Def. 1.7.5) linear form.

**Theorem 2.5.22 (Strang’s first lemma).** Beside the assumptions on \( a \) and \( a_N \) stated above we demand that

\[
\exists \tilde{\gamma} > 0 : \quad a_N(v_N, v_N) \geq \tilde{\gamma} \left\| v_N \right\|_V^2 \quad \forall v_N \in V_N .
\] (2.5.14)

Then (2.5.13) will have a unique solution \( u_N \in V_N \), which satisfies the a-priori error estimate

\[
\left\| u - \tilde{u}_N \right\|_V \leq C \left( \inf_{v_N \in V_N} \left( \left\| u - v_N \right\|_V + \sup_{w_N \in V_N} \frac{|a(v_N, w_N) - a_N(v_N, w_N)|}{\left\| w_N \right\|_V} \right) \right.

\left. + \sup_{w_N \in V_N} \frac{|f(w_N) - f_N(w_N)|}{\left\| w_N \right\|_V} \right),
\]

with \( C = C(C_A, \tilde{\gamma}) > 0 \).
Proof. (2.5.14) ⇒ for any $v_N \in V_N$

\[
\| \widetilde{u}_N - v_N \|_V^2 \leq a_N(\widetilde{u}_N - v_N, \widetilde{u}_N - v_N)
= a(u - v_N, \widetilde{u}_N - v_N) + (a(v_N, \widetilde{u}_N - v_N) - a_N(v_N, \widetilde{u}_N - v_N))
+ (a_N(\widetilde{u}_N, \widetilde{u}_N - v_N) - a(u, \widetilde{u}_N - v_N))
= a(u - v_N, \widetilde{u}_N - v_N) + (a(v_N, \widetilde{u}_N - v_N) - a_N(v_N, \widetilde{u}_N - v_N))
- f(\widetilde{u}_N - v_N) + f_N(\widetilde{u}_N - v_N).
\]

Continuity of $a$ & dividing by $\| \widetilde{u}_N - v_N \|_V$:

\[
\| \widetilde{u}_N - v_N \|_V \leq C_A \| u - v_N \|_V + \frac{|a(v_N, \widetilde{u}_N - v_N) - a_N(v_N, \widetilde{u}_N - v_N)|}{\| \widetilde{u}_N - v_N \|_V}
\]

\[
+ \frac{|f(\widetilde{u}_N - v_N) - f_N(\widetilde{u}_N - v_N)|}{\| \widetilde{u}_N - v_N \|_V}
\]

\[
\leq C_A \| u - v_N \|_V + \sup_{w_N \in V_N} \frac{|a(v_N, w_N) - a_N(v_N, w_N)|}{\| w_N \|_V}
\]

\[
+ \sup_{w_N \in V_N} \frac{|f(w_N) - f_N(w_N)|}{\| w_N \|_V}.
\]

\forall v_N & triangle-inequality: $\| u - \widetilde{u}_N \|_V \leq \| u - v_N \|_V + \| \widetilde{u}_N - v_N \|_V \Rightarrow \square$
Terminology: assumption (2.5.14) = h-ellipticity of $a_N$.

If $\tilde{\gamma}$ independent of the meshwidth $\Rightarrow$ uniform h-ellipticity.

The two terms

$$\sup_{w_N \in V_N} \frac{|a(v_N, w_N) - a_N(v_N, w_N)|}{\|w_N\|_V}, \quad \sup_{w_N \in V_N} \frac{|f(w_N) - f_N(w_N)|}{\|w_N\|_V},$$

are called consistency (error) terms.

### 2.5.8.2 Impact of numerical quadrature

Model problem: on polygonal/polyhedral $\Omega \subset \mathbb{R}^d$:

$$u \in H^1_0(\Omega): \quad a(u, v) := \int_\Omega \sigma(x) \nabla u \cdot \nabla v \, dx = f(v) := \int_\Omega f v \, dx. \quad (2.5.15)$$

Assumptions:

- $\sigma$ satisfies (1.2.2), $\sigma \in C^0(\overline{\Omega}), f \in C^0(\overline{\Omega})$

- Galerkin discretization, $V_N := \delta_p^0(M)$ on simplicial mesh $M$

- Approximate evaluation of $a(u_N, v_N), f(v_N)$ by stable local numerical quadrature rule (→ Sect. 2.2.6)

- perturbed $a_N, f_N$ (see (2.5.13))

Assumed: uniform quadrature rule (defined parametrically)
Consistency error estimates from finite element theory [5, Ch. 4,§4.1]:

\[ \text{Theorem 2.5.23 (Consistency error estimate for right hand side). If } f_N \text{ for (2.5.15) is based on a parametrically defined local quadrature rule that is exact for polynomials up to degree } 2p - 2 \text{ and } f \in W^{p, \infty}(\Omega), \text{ then} \]

\[ |f(w_N) - f_N(w_N)| \leq C h_M^p \| f \|_{p, \infty} \| w_N \|_{1, \Omega} \quad \forall w_N \in s^0_p(\mathcal{M}), \]

\[ \text{where } C = C(p, \rho, \mathcal{M}) > 0. \]

\[ \| u - u_N \|_1 = O(h_M^p) \text{ at best} \]

Quadrature rule of order \(2p - 1\) sufficient for \(f_N\).
Theorem 2.5.24 (Consistency error estimate for approximate bilinear form). Assume

- the exact solution $u$ of (2.5.15) belongs to $H^{p+1}(\Omega)$, $p, m \in \mathbb{N}$, and $\sigma \in W^{p, \infty}(\Omega)$,
- $a_N$ arises from parametrically defined local numerical quadrature exact for polynomials up to degree $2p - 2$,

Then

$$a(l_p u, w_N) - a_N(l_p u, w_N) \leq C h^p_{\mathcal{M}} \|\sigma\|_{p, \infty} \|u\|_{p+1, \Omega} \|w_N\|_{1, \Omega} \quad \forall w_N \in S^0_p(\mathcal{M}),$$

where $C = C(p, \rho_{\mathcal{M}}, \mu_{\mathcal{M}}) > 0$.

\[\|u - u_N\|_1 = O(h^p_{\mathcal{M}})\] at best \[\Rightarrow\] Quadrature rule of order $2p - 1$ sufficient for $a_N$.

Example 2.5.7 (Impact of numerical quadrature on finite element discretization error).

$\Omega = ]0, 1[^2$, $\sigma \equiv 1$, $f(x, y) = 2\pi^2 \sin(\pi x) \sin(\pi y)$, $(x, y)^T \in \Omega$

$\Rightarrow$ $u(x, y) = \sin(\pi x) \sin(\pi y)$, $g = 0$.

- Quadratic Lagrangian FE ($V_N = S^0_2(\mathcal{M})$) on triangular meshes $\mathcal{M}$, obtained by regular refinement.
• Exact evaluation of bilinear form

• $f_N$ from one point quadrature rule (2.2.13) of order 2
  (Thm. 2.5.23 → quadrature rule insufficient)

\[ H^1(\Omega) \text{-norm of discretization error on unit square} (- \leftrightarrow \text{rule (2.2.13)}, - \leftrightarrow \text{rule (2.2.14)}) \]

Observation: Use of quadrature rule of order 2 \( \Rightarrow \) Algebraic rate of convergence drops from \( \alpha = 1 \) to \( \alpha = 1/2 \)!
2.5.8.3 Approximation of boundary

Model problem \( (f \in L^2(\Omega)) \):

\[
u \in H^1_0(\Omega): \quad a(u, v) := \int_{\Omega} \nabla u \cdot \nabla v \, dx = f(v) := \int_{\Omega} f \, v \, dx \quad \forall v \in H^1_0(\Omega).
\]

- \( \Omega \subset \mathbb{R}^2 \) with piecewise \( C^2 \)-boundary.
- Linear Lagrangian finite elements \( S_{1,0}(\mathcal{M}) \) on triangular mesh \( \mathcal{M} \).
- Mesh resolves corners of \( \partial \Omega \).
- Linear boundary fitting (\( \to \) Sect. 2.2.8).

\[
\Omega_h = \text{(interior of) union of (closed) straight-edged triangles of } \mathcal{M}
\]

Note:

\[
S_{1,0}(\mathcal{M}) \subset H^1_0(\Omega_h), \quad \text{but } S_{1,0}^0(\mathcal{M}) \nsubseteq H^1_0(\Omega)!
\]
Approximate bilinear form on $\mathcal{S}_{1,0}^0(M)$:

$$a_N(u_N, v_N) := \int_{\Omega_h} \nabla u_N \cdot \nabla v_N \, dx.$$ 

Approximate linear form on $\mathcal{S}_{1,0}^0(M)$:

$$f_N(v_N) := \int_{\Omega \cap \Omega_h} f \, v_N \, dx.$$

How to apply Strang's first lemma (→ Thm. 2.5.22) ?

What is $V$ ?  What is $\| \cdot \|_V$ ?

Not necessarily $V = H^1_0(\Omega)$: use $V = H^1(\Omega)$ instead: $\| \cdot \|_V = \| \cdot \|_{1,\Omega}$

$K$ : curvilinear triangle

$\tilde{K}$ : straight-edged triangle

$C := (\tilde{K} \setminus K) \cup (K \setminus \tilde{K})$

Reinterpretation:

$$v_N \in \mathcal{S}_{1,0}^0(M) \rightarrow \tilde{v}_N \in H^1(\Omega)$$

$$\tilde{v}_N \in H^1(\Omega)$$ by local affine extension of $v_N|_{\tilde{K}}$ into $C \setminus \tilde{K}$
One step of proof fails:

\[ a(u, \tilde{u}_N - \tilde{v}_N) \neq f(\tilde{u}_N - \tilde{v}_N) \]

- **Modified consistency error terms:**

\[
\sup_{w_N \in V_N} \frac{|a(\tilde{v}_N, \tilde{w}_N) - a_N(v_N, w_N)|}{\|w_N\|_V} , \quad \sup_{w_N \in V_N} \frac{|a(u, \tilde{w}_N) - f_N(w_N)|}{\|w_N\|_V} . \tag{2.5.16}
\]

---

**Step I:** Estimation of consistency error \( a \leftrightarrow a_N \):
Local parameterization over edge $E_\Gamma$:

$$\partial \Omega = \{(\xi, \eta) : \eta = g(\xi)\}$$

with $g \in C^2$ > interpolation estimate

$$|C| \leq \int_0^h |g(t) - g(0)(1 - t) - g(h)t| \, dt$$

$$\leq h^3 K \max_{0 \leq t \leq h} |g''(t)|,$$

$$|K| \geq \rho_K^{-1} h_K^2.$$

(2.5.17)

\[\text{grad} \tilde{v}_N|_K \text{ const} \]

existence: $C = C(\rho_M, \partial \Omega) : (1 - Ch_M) \|v_N\|_{1, \Omega_h}^2 \leq \|\tilde{v}_N\|_{1, \Omega}^2 \leq (1 + Ch_M) \|v_N\|_{1, \Omega_h}^2.$

(2.5.18)

Thm. 1.7.9 ➤ $a_N(v_N, v_N) \geq (1 + \text{diam} \Omega_h)^{-1} \|v_N\|_{1, \Omega_h}^2 \quad \forall v_N \in \mathcal{S}_{1,0}^0(M).$

+ (2.5.18) ➤ uniform $h$-ellipticity of $a_N$ (since $\text{diam} \Omega_h \leq \text{diam} \Omega$ independent of $M$):

$$a_N(v_N, v_N) \geq (1 - Ch_M)(1 + \text{diam} \Omega)^{-1} \|\tilde{v}_N\|_{1, \Omega}^2 \quad \forall v_N \in \mathcal{S}_{1,0}^0(M).$$
Local estimation of consistency error:

$$\left| a(\tilde{v}_N, \tilde{w}_N) - a_N(v_N, w_N) \right| = \int_{\Omega_h \setminus \Omega} \nabla v_N \cdot \nabla w_N \, dx + \int_{\Omega \setminus \Omega_h} \nabla \tilde{v}_N \cdot \nabla \tilde{w}_N \, dx$$

\begin{align*}
&\leq C h_M \sum_{K \in \mathcal{M}, \tilde{K} \cap \partial \Omega \neq \emptyset} | \nabla v_N \cdot \nabla w_N | \, dx \\
&\leq C h_M \| \tilde{v}_N \|_{1, \Omega} \| \tilde{w}_N \|_{1, \Omega} .
\end{align*}

\((*) \leftarrow \tilde{v}_N, \tilde{w}_N \text{ constant on } K \in \mathcal{M} \quad \& \quad \text{estimate (2.5.17)}

---

Step II: Estimation of consistency error \(a(u, \tilde{w}_N) - f_N(w_N)\):

\[
a(u, \tilde{w}_N) - f_N(w_N) = \int_{\Omega} \nabla u \cdot \nabla \tilde{w}_N \, dx - \int_{\Omega \cap \Omega_h} f \, w_N \, dx \\
= - \int_{\Omega} \Delta u \, \tilde{w}_N \, dx + \int_{\partial \Omega} \frac{\partial u}{\partial n} \cdot \tilde{w}_N \, dS - \int_{\Omega \cap \Omega_h} f \, w_N \, dx \\
= - \int_{\Omega \setminus \Omega_h} f \, \tilde{w}_N \, dx + \int_{\partial \Omega} \frac{\partial u}{\partial n} \cdot \tilde{w}_N \, dS .
\]
Lifting theorem for Neumann data: \[ \exists C = C(\Omega): \left\| \frac{\partial u}{\partial n} \right\|_{0, \partial \Omega} \leq \| f \|_0 \]

\[ |a(u, \tilde{w}_N) - f_N(w_N)| \leq C h_M \| f \|_0 \| w_N \|_0 + \left\| \frac{\partial u}{\partial n} \right\|_{0, \partial \Omega} \| \tilde{w}_N \|_{0, \partial \Omega} \]

\[ \leq C h_M \| f \|_0 \| w_N \|_0 + C \| f \|_0 h^{3/2} \mathcal{M} \| w_N \|_{1, \Omega_h} \cdot \]

\( (*) \leftarrow \) Cauchy-Schwarz inequality (1.7.5) \& (2.5.17)

A technical estimate \([4, \text{Lemma 1.6}]:\)

(\( \tilde{w}_N \) affine linear !)

\( (\diamond) \leftarrow \)

\[ \| \tilde{w}_N \|_{0, \Gamma_K} \leq C h^{3/2}_K \| w_N \|_{1, K} \]

\[ C = C(\rho_M, \partial \Omega) \]

For h-refinement (uniform shape-regularity of meshes assumed):

Consistency error = \( O(h_M) \)

(matches optimal asymptotic convergence of linear FE in \( H^1(\Omega) \)-norm)
Example 2.5.8 (Impact of linear boundary fitting on FE convergence).

\[ \Omega := B_1(0) := \{ \mathbf{x} \in \mathbb{R}^2 : |\mathbf{x}| < 1 \}, \ u(r, \phi) = 1 - r^2 \text{ (polar coordinates)} \Rightarrow f \equiv 4. \]

- Sequences of unstructured triangular meshes \( \mathcal{M} \) obtained by regular refinement (of coarse mesh with 4 triangles) + linear boundary fitting.
- Galerkin FE discretization based on \( V_N := S_{1,0}(\mathcal{M}) \) or \( V_N := S_{2,0}(\mathcal{M}) \).
- Recorded: approximate norm \( |u - u_N|_{1,\Omega_h} \), evaluated using numerical quadrature rule (2.2.14).
Linearly boundary fitted unstructured triangular meshes of $\Omega = B_1(0)$. 
Discretization errors with respect to $H^1$ semi-norm

$H^1(\Omega)$-norm of discretization error on unit ball ($- \leftrightarrow p = 1, - \leftrightarrow p = 2$)

Observations:
- Linear Lagrangian FE retain algebraic convergence  $\|u - u_N\|_1 = O(N^{-1/2})$
- Bad example: exact solution by quadratic Lagrangian FE!

Modified setting:

$\Omega := B_1(0) := \{x \in \mathbb{R}^2: |x| < 1\}, u(r, \phi) = \cos(\pi/2r)$ (polar coordinates)

$\Rightarrow f = \frac{\pi}{2r} \sin(\pi/2r) + \frac{\pi}{2} \cos(\pi/2r)$
$H^1(\Omega)$-norm of discretization error on unit ball ($- \leftrightarrow p = 1, - \leftrightarrow p = 2$)

Observations: Despite linear boundary fitting:

Lagrangian FE achieve optimal algebraic convergence

$$\|u - u_N\|_1 = O(N^{-1/2}) \text{ for } p = 1, \|u - u_N\|_1 = O(N^{-1}) \text{ for } p = 2$$

Theory not sharp?

Theoreticians are pessimists!

(The worst case $\in \{\text{class of problems under scrutiny}\}$ governs the result)
Theoretical guideline:

If \( V_N = \delta^0_p(M) \), use boundary fitting with polynomials of degree \( p \).

### 2.5.9 Duality estimates

**Abstract setting:** linear variational problem on Hilbert space \( V \) with norm \( \| \cdot \|_V \)

\[
    u \in V: \quad a(u, v) = f(v) \quad \forall v \in V. \tag{1.7.1}
\]

- \( a : V \times V \mapsto \mathbb{R} \): continuous (\( \rightarrow \) Def. 1.7.6), \( V \)-elliptic (\( \rightarrow \) Def. 1.7.7) bilinear form.
- \( f : V \mapsto \mathbb{R} \): continuous (\( \rightarrow \) Def. 1.7.5) linear form.

**Galerkin discretization using** \( V_N \subset V \) \( \Rightarrow \) **discrete variational problem**

\[
    u_N \in V_N: \quad a(u_N, v_N) = f(v_N) \quad \forall v_N \in V_N. \tag{2.1.1}
\]

**Continuous output functional**

\[
    F : V \mapsto \mathbb{R}
\]

**Expected:**

\[
    |F(u) - F(u_N)| \leq C_f \|u - u_N\|_V.
\]

**A priori estimates for** \( \|u - u_N\|_V \) \( \Rightarrow \) **estimates for** \( |F(u) - F(u_N)| \)
Example 2.5.9 (Approximation of mean temperature).

(heat conduction): \[
\text{mean temperature } F(u) = \frac{1}{|\Omega|} \int_{\Omega} u \, dx
\]

\[
u \in H^1_0(\Omega): \quad a(u, v) := \int_{\Omega} \nabla u \cdot \nabla v \, dx = f(v) := \int_{\Omega} f \, v \, dx.
\] \hspace{1cm} (2.5.15)

Domain \(\Omega = ]0, 1[^2\), source function \(f(x, y) = 2\pi^2 \sin(\pi x) \sin(\pi y)\), \((x, y)^T \in \Omega\)

\(\Rightarrow\) solution \(u(x, y) = \sin(\pi x) \sin(\pi y), \ g = 0\).

- Sequence of triangular meshes \(\mathcal{M}\) created by regular refinement.
- Galerkin discretization: \(V_N := \delta^0_{1,0}(\mathcal{M})\) (linear Lagrangian finite elements).
- Quadrature rule (2.2.14) of order 6 for right hand side
  (more than sufficiently accurate \(\rightarrow\) Thm. 2.5.23).

Expected: algebraic convergence of order 1 of approximate mean temperature
Error in mean value on unit square ($-\leftrightarrow p = 1$, $-\leftrightarrow p = 2$)

Observation: Mean value converges twice as fast as expected!
Theorem 2.5.25 (Duality estimate for linear functional output). Define the dual solution $g^F \in V$ to $F$ as solution of

$$g^F \in V: \quad a(v, g^F) = F(v) \quad \forall v \in V.$$ 

Then

$$|F(u) - F(u_N)| \leq C_A \|u - u_N\|_V \inf_{v_N \in V_N} \|g^F - v_N\|_V.$$ 

Proof. For any $v_N \in V_N$:

$$F(u) - F(u_N) = a(u - u_N, g^F) = a(u - u_N, g^F - v_N) \leq C_A \|u - u_N\|_V \|g^F - v_N\|_V. \quad (*)$$

$(*) \leftarrow$ by Galerkin orthogonality (2.1.2).

If $g^F$ can be approximated well in $V_N$, then error in linear output functional $F$ can converge to $0$ (much) faster than $\|u - u_N\|_V$.

Application: a priori error estimate for approximation of mean temperature $\int \frac{1}{|\Omega|} \int \Omega u - u_N \, dx \leq |u - u_N|_1 \inf_{v_N \in \mathcal{S}_{1,0}^0(M)} |g - v_N|_1$.
where \( g \in H^1_0(\Omega) \) solves

\[
\int_{\Omega} \text{grad } g \cdot \text{grad } v \, dx = \frac{1}{|\Omega|} \int_{\Omega} v \, dx \quad \forall v \in H^1_0(\Omega).
\]

If \( g, u \in H^2(\Omega) \) by using interpolation estimates, Thm.\[2.5.16\]:

\[
\exists C = C(\rho, \mathcal{M}) : \frac{1}{|\Omega|} \int_{\Omega} u - u_N \, dx \leq C h^2_{\mathcal{M}}.
\]

**Example 2.5.10 (Boundary flux computation).**

Long pipe carrying turbulent flow of coolant (water)

\( \Omega \subset \mathbb{R}^2 \) : cross-section of pipe

\( \sigma \) : heat conductivity of pipe material (assumed homogeneous isotropic)

Assumption: Constant temperatures \( u_o, u_i \) at outer/inner wall \( \Gamma_o, \Gamma_i \) of pipe

Task: Compute heat flow pipe \( \rightarrow \) water
Mathematical model: elliptic boundary value for stationary heat conduction (→ Sect 1.2)

Numerical tool: finite element computation of heat conduction in pipe
  (e.g. linear Lagrangian finite elements)

\[- \text{div}(\sigma \, \text{grad} \, u) = 0 \quad \text{in} \, \Omega \quad , \quad u = u_x \quad \text{on} \, \Gamma_x, \ x \in \{i, o\}. \] (2.5.19)

Heat flux through \( \Gamma_i \):
\[ J = \int_{\Gamma_i} \sigma \, \text{grad} \, u \cdot n \, dS. \] (2.5.20)

In abstract framework:
\[(2.5.19) \cong (1.7.1), \quad V = H^1_0(\Omega) \text{ (see Sect. 1.8)}\]

BUT functional \( J \) from \( (2.5.20) \) is not continuous on \( H^1(\Omega) \) ! (→ exercise)

Thm. 2.5.25 cannot be applied

(Potentially) poor convergence of flux obtained from straightforward evaluation of \( J(u_N) \) for FE solution \( u_N \in \mathcal{H}^1(\mathcal{M}) \!)

Trick: use fixed cut-off function \( \psi \in C^0(\overline{\Omega}) \cap H^1(\Omega), \psi \equiv 1 \text{ on} \, \Gamma_i, \text{ supp} \psi \cap \Gamma_o = \emptyset \text{ [6, Sect. 2.1]}:

\[ \int_{\Gamma_i} \sigma \, \text{grad} \, u \cdot n \, dS = \int_{\Gamma_i} (\sigma \, \text{grad} \, u \cdot n) \, \psi \, dS = \int_{\Omega} \underbrace{\text{div}(\sigma \, \text{grad} \, u)}_{=0} \psi + \sigma \, \text{grad} \, u \cdot \text{grad} \, \psi \, dx \]

use
\[ J^*(u) := \int_{\Omega} \sigma \, \text{grad} \, u \cdot \text{grad} \, \psi \, dx. \] (2.5.21)
Obviously: \( J^* : H^1(\Omega) \mapsto \mathbb{R} \) continuous \& \( J^*(u) = J(u) \) for solution of (2.5.19)

Domain \( \Omega = B_{R_o}(0) \setminus B_{R_i}(0) := \{ x \in \mathbb{R}^2 : R_i < |x| < R_o \} \) with \( R_o = 1 \) and \( R_i = 1/2 \), boundary data \( u_i = 60^\circ C \) on \( \Gamma_i \), \( u_o = 10^\circ C \) on \( \Gamma_o \), heat source \( f \equiv 0 \) and heat conductivity \( \sigma = 1 \).

- Solution: \( u(r, \phi) = C_1 \ln(r) + C_2 \),
- Heat flux: \( J = 2\pi \sigma C_1 \),

with constants \( C_1 := (u_o - u_i)/(\ln R_i - \ln R_o) \) and \( C_2 := (\ln R_o u_i - \ln R_i u_o)/(\ln R_i - \ln R_o) \).

- Sequences of unstructured triangular meshes \( \mathcal{M} \) obtained by regular refinement of coarse mesh obtained from unstructured grid generator.
- Galerkin FE discretization based on \( V_N := \mathcal{S}^0_{1,0}(\mathcal{M}) \) or \( V_N := \mathcal{S}^0_{2,0}(\mathcal{M}) \).
- Approximate evaluation of \( a(u_N, v_N), f(v_N) \) and \( J^*(u_N) \) by six point quadrature rule (2.2.14).
- Approximate evaluation of \( J(u_N) \) by 4 point Gauss-Legendre quadrature rule.
- Recorded: \( |J - J(u_N)| \) and \( |J - J^*(u_N)| \), evaluated using numerical quadrature rule (2.2.14).
Unstructured triangular meshes for $\Omega = B_1(0) \setminus B_{1/2}(0)$ (two coarsest specimens).
Convergence rates for $|J - J(u_N)|$ and $|J - J^*(u_N)|$. ($\rightarrow p = 1$, $\leftrightarrow p = 2$)

**Observations:**
- For both $p = 1, 2$: $|J - J(u_N)| = O(h)$ and $|J - J^*(u_N)| = O(h^2)$.
- Algebraic convergence of $J^*$ twice as fast as that of $J$

---

Special linear output functionals:

$$F(v) := b(u - u_N, v) \quad \text{with continuous bilinear form} \quad b : V \times V \rightarrow \mathbb{R}.$$
\[ b(u - u_N, u - u_N) \leq C_A \|u - u_N\|_V \inf_{v_N \in V_N} \|g^b - v_N\|_V, \quad (2.5.22) \]

where

\[ g^b \in V: \quad a(v, g^b) = b(u - u_N, v) \quad \forall v \in V. \]

**Application:** homogeneous Dirichlet problem for scalar linear 2nd-order elliptic BVP

\( (V = H^1_0(\Omega), a, f \text{ from (1.7.4)}) \)

Lagrangian FE discretization, \( V_N := S_0^p(\mathcal{M}) \text{ on mesh } \mathcal{M} \text{ of } \Omega \)

\[ b(v, w) := \int_\Omega v w \, \text{d}x \quad \triangleright \quad \text{inner product in } L^2(\Omega). \]

[Apply (2.5.22)]

\[ \|u - u_N\|_0^2 \leq \|u - u_N\|_1 \inf_{v_N \in S_0^p(\mathcal{M})} \|g - v_N\|_1, \quad (2.5.23) \]

where

\[ g \in H^1_0(\Omega): \quad \int_\Omega \sigma(x) \text{grad } g \cdot \text{grad } v \, \text{d}x = \int_\Omega (u - u_N) v \, \text{d}x \quad \forall v \in H^1_0(\Omega) \]

\[ \triangleright \]

\[ -\text{div}(\sigma \text{grad } g) = u - u_N \quad \text{in } \Omega, \quad g = 0 \quad \text{on } \partial \Omega. \]
**Assumption:** Homogeneous Dirichlet boundary value problem for \(- \text{div}(\sigma \, \text{grad} \, u) = f\) is 2-regular on \(\Omega\):

\[
\exists C_2 = C_2(\Omega) > 0: \quad \|u\|_2 \leq C_2 \|f\|_0, \quad \|g\|_2 \leq C_2 \|\text{div}(\sigma \, \text{grad} \, g)\|_0 = C_2 \|u - u_N\|_0 \tag{2.5.24}
\]

(Thm. 2.5.21 ➤ assumption satisfied for \(\sigma \equiv 1, \Omega \text{ convex.}\))

\[\text{(2.5.23) & interpolation estimates, Thm. 2.5.16: } \|g - l_p g\|_1 \leq C_1 h_M \|g\|_2\]

\[
\|u - u_N\|_0^2 \leq C_2^2 C_1^2 h_M^2 \|f\|_0 \|u - u_N\|_0.
\]

**Theorem 2.5.26** (A priori \(L^2(\Omega)\)-error estimate). Let the variational formulation of a second-order elliptic BVP be given by (2.5.15) and assume 2-regularity. Then the Galerkin solution \(u_N \in \mathcal{S}_p^0(\mathcal{M})\), \(p \in \mathbb{N}\), satisfies the a priori \(L^2(\Omega)\)-error estimate

\[
\exists C = C(\Omega, \rho_\mathcal{M}): \quad \|u - u_N\|_0 \leq C h_M^2 \|f\|_0 \quad \forall \, f \in L^2(\Omega). 
\]

If \(u \in H^{p+1}(\Omega)\), then

\[
\exists C = C(\Omega, \rho_\mathcal{M}): \quad \|u - u_N\|_0 \leq C h_M^{p+1} |u|_{p+1}. 
\]
Example 2.5.11 ($L^2$-convergence of FE solutions).

Setting \(^1\) :
\[ \Omega = [0, 1]^2, \sigma = 1, f(x, y) = 2\pi^2 \sin(\pi x) \sin(\pi y), (x, y)^T \in \Omega \]
\[ u(x, y) = \sin(\pi x) \sin(\pi y). \]

Setting \(^2\) :
\[ \Omega = (-1, 1]^2 \setminus (0, 1] \times [0], \sigma = 1, u(r, \phi) = r^{2/3} \sin(2/3\phi) \text{ (polar coordinates)} \]
\[ f = 0, \text{ Dirichlet data } g = u|_{\partial \Omega}. \]

\begin{itemize}
  \item Sequence of triangular meshes \( \mathcal{M} \), created by regular refinement.
  \item FE Galerkin discretization based on \( S^0_0(\mathcal{M}) \) or \( S^0_2(\mathcal{M}) \).
  \item Quadrature rule (2.2.14) for assembly of local load vectors (→ Sect. 2.2.6).
  \item Approximate \( L^2(\Omega) \)-norm by means of quadrature rule (2.2.14).
\end{itemize}
$L^2(\Omega)$-norm of discretization error on unit square ($\iff p = 1$, $\iff p = 2$)

Observations:  
- Linear Lagrangian FE ($p = 1$)  $\gg \| u - u_N \|_0 = O(N^{-1})$
- Quadratic Lagrangian FE ($p = 2$)  $\gg \| u - u_N \|_0 = O(N^{-1.5})$
$L^2(\Omega)$-norm of discretization error on “L-shaped” domain ($-\leftrightarrow p = 1$, $-\leftrightarrow p = 2$)

Observation: For both ($p = 1, 2$) algebraic convergence $\|u - u_N\|_0 = O(N^{-2/3})$

(But we still gain a power of $h$ compared to $H^1(\Omega)$-estimate.)

2.5.10 Pointwise estimates

Goal: a priori finite element error estimate for $\|u - u_N\|_\infty$
Considered: 2nd-order elliptic BVP, Galerkin FE discretization by means \( V_N := S^0_{1,0}(\mathcal{M}) \) on simplicial mesh \( \mathcal{M} \) of \( \Omega \subset \mathbb{R}^d \)

Assumption: Source function \( f \in L^2(\Omega) \), BVP 2-regular.

**Lemma 2.5.27** \((L^\infty \text{ interpolation error estimate})\). If \( u \in H^2(\Omega) \), and \( I_1 \) denotes the nodal interpolation operator \( I_1 : C^0(\overline{\Omega}) \mapsto S^0_1(\mathcal{M}) \), then

\[
\exists C = C(\rho, \mathcal{M}): \| u - I_1 u \|_\infty \leq C h_{\mathcal{M}} |u|_2.
\]

**Proof.** By transformation techniques (\( \rightarrow \) Sect. 2.5.4), \( \hat{K} \) reference simplex (2.5.3), (2.5.4).
Pick $K \in \mathcal{M}$ with affine transformation $\Phi_K : \hat{K} \mapsto K$

$$\left\| u - 1^K_1 u \right\|_{\infty, K} = \left\| \Phi_K^* (u - 1^K_1 u) \right\|_{\infty, \hat{K}}$$

$$= \left\| \hat{u} - \hat{1}\hat{u} \right\|_{\infty, \hat{K}}$$

$$= \inf_{q \in P_1(\hat{K})} \left\| (\hat{u} - \hat{1})(\hat{u} - q) \right\|_{\infty, \hat{K}}$$

$$\leq C_1 \inf_{q \in P_1(\hat{K})} \left\| \hat{u} - q \right\|_{2, \hat{K}}$$

$$\leq C \rho_K h_K |u|_{2, K}.$$  

1. $\rightsquigarrow$ by definition of $\Phi_K^*$, see Def. 2.1.19
2. $\rightsquigarrow$ by Lemma 2.5.15
3. $\hat{1}_1 = \text{projection onto } P_1(\hat{K}) (\rightarrow \text{Def. 2.1.3})$
4. $\rightsquigarrow$ Sobolev embedding Thm. 2.5.4: $\|\hat{u}\|_{\infty} \leq C_1 \|\hat{u}\|_{2, \hat{K}}$
5. $\rightsquigarrow$ see final steps in (2.5.6).
Lemma 2.5.28 ($L^\infty$-$L^2$ inverse estimate).

\[ \exists C = C(d): \quad \| v_N \|_\infty \leq C \left( \min_{K \in \mathcal{M}} h_K \right)^{-1} \| v_N \|_0 \quad \forall v_N \in \delta_1^0(\mathcal{M}). \quad (2.5.25) \]

Proof. By transformation techniques (\(\rightarrow\) Sect. 2.5.4), \(\hat{K}\) reference simplex (2.5.3), (2.5.4).

Pick \(K \in \mathcal{M}\) with affine transformation \(\Phi_K : \hat{K} \mapsto K\)

\[ \| v_N \|_{\infty, K} \overset{1}{=} \| \Phi_K^* v_N \|_{\infty, \hat{K}} \overset{2}{=} C \| \hat{v}_N \|_{0, \hat{K}} \overset{3}{=} C \| v_N \|_{0, K}. \]

1 \(\leftarrow\) by definition of \(\Phi_K^*\), see Def. 2.1.19

2 \(\leftarrow\) by Lemma 2.5.17

3 \(\leftarrow\) by Lemma 2.5.9 for \(m = 0\)

Terminology: (2.5.25) is an example of an inverse estimate
Definition 2.5.29. Let \( \| \cdot \|_X \) and \( \| \cdot \|_Y \) be two norms on a function space \( V \) such that

\[
\exists C_1 > 0: \| v \|_X \leq C_1 \| v \|_Y \quad \forall v \in V, \quad \forall C > 0: \| v \|_Y \leq C \| v \|_X \quad \forall v \in V,
\]

(i.e. \( \| \cdot \|_Y \) is strictly stronger than \( \| \cdot \|_X \)). However, by Lemma 2.5.17 for \( V_N \subset V \), \( \dim V_N < \infty \)

\[
\exists \tilde{C} = \tilde{C}(V_N): \| v_N \|_Y \leq \tilde{C} \| v_N \|_X \quad \forall v_N \in V_N.
\]

Assertions about the dependence of \( \tilde{C} \) from (discretization) parameters defining \( V_N \) are inverse estimates.

Finish a priori \( L^\infty(\Omega) \)-error estimate:

Lemma 2.5.28
\[ \|u - u_N\|_\infty \leq \|u - l_1u\|_\infty + \|l_1u - u_N\|_\infty \]

\[ \leq \|u - l_1u\|_\infty + C \left( \min_{K \in \mathcal{M}} h_K \right)^{-1} \|l_1u - u_N\|_0 \]

\[ \leq C_1 h_{\mathcal{M}} |u|_2 + C' h_{\mathcal{M}}^{-1} (\|l_1u - u\|_0 + \|u - u_N\|_0) \]

\[ \leq C_1 h_{\mathcal{M}} |u|_2 + C'' h_{\mathcal{M}} |u|_2 \]

\[ \leq C_2 h_{\mathcal{M}} \|f\|_0. \]

1. \( \triangle \)-inequality
2. by Lemma 2.5.28
3. by Lemma 2.5.27 & appealing to quasi-uniformity (\( \rightarrow \) Def. 2.5.13) of the mesh \( \mathcal{M} \), and \( \triangle \)-inequality
4. by Thm. 2.5.26 and interpolation estimate, see Thm. 2.5.16: \( C'' = C''(\Omega, \rho_{\mathcal{M}}, \mu_{\mathcal{M}}) \)
5. by assumption of 2-regularity, \( f = \) source function

\[ \exists C = C(\Omega, \rho_{\mathcal{M}}, \mu_{\mathcal{M}}): \quad \|u - u_N\|_\infty \leq C h_{\mathcal{M}} \|f\|_0. \]
Sharper estimate for \( d = 2 \) (requires sophisticated FE theory [5, Ch. 4]):

\[
\exists C = C(\Omega, \rho_M, \mu_M): \|u - u_N\|_\infty \leq C h_M^2 |\log h|^{3/2} \|f\|_0.
\]

**Example 2.5.12** \( L^\infty \)-estimate for FE solutions.

Setting: second-order elliptic BVP, homogeneous Dirichlet boundary conditions,
\[
\Omega = ]0, 1[^2], \sigma \equiv 1, f(x, y) = 2\pi^2 \sin(\pi x) \sin(\pi y), (x, y)^T \in \Omega
\]
\[
\Rightarrow u(x, y) = \sin(\pi x) \sin(\pi y).
\]

- Sequence of unstructured triangular meshes \( M \) obtained by regular refinement + smoothing
- FE Galerkin discretization by means of linear/quadratic Lagrangian finite elements:
  \[
  V_N = \delta_{1,0}^0(M)/V_N = \delta_{2,0}^0(M)
  \]
- Approximate load vector by local numerical quadrature (2.2.14) of order 6
- Monitored
  \[
  \|u - u_N\|_{\infty, M} := \max\{(u - u_N)(p)\}: p \text{ vertex of } M.
  \]
$L^\infty(\Omega)$-norm of discretization error on unit square ($-\leftrightarrow p = 1$, $-\leftrightarrow p = 2$)

Example 2.5.13 ($L^\infty$-estimate on regular grid).
Example: same problem as above

- Galerkin FE discretization by means of linear and quadratic Lagrangian finite elements on triangular tensor-product grids
- Approximate load vector by vertex based quadrature (2.2.11)

Monitor $\|u - u_N\|_{\infty,M}$ as above
Discretization error with respect to $L^\infty$ norm

$L^\infty(\Omega)$-norm of discretization error on unit square, ($\leftrightarrow p = 1$, $\leftrightarrow p = 2$)

Observation: Significantly improved algebraic convergence $\|u - u_N\|_{\infty,M} = O(N^{-1})$!

Superconvergence of point values on regular grids
3 Special elliptic boundary value problems

3.1 Convection-diffusion problems

3.1.1 Heat conduction in fluid

Computational domain $\Omega \subset \mathbb{R}^d = \text{container filled with (homogeneous) fluid } (d = 3)$

Vectorfield $\mathbf{v} : \Omega \mapsto \mathbb{R}^d = \text{constant velocity field } (\leftrightarrow \text{movement of fluid})$

Elliptic BVP governing stationary temperature distribution $u = u(\mathbf{x}), \mathbf{x} \in \Omega$, in fluid

$$- \text{div}(\sigma \text{ grad } u) + \rho \mathbf{v} \cdot \text{grad } u = f,$$

+ boundary conditions (→ Sect [1.3]) on $\partial \Omega$. 

σ: heat conductivity
ρ: heat capacity \([\rho] = \frac{J}{\text{km}^3}\) (→ stationary heat conduction, Sect. 1.2)
f: heat source/sink

\[
- \text{div}(\sigma \ \text{grad} \ u) + \rho \mathbf{v} \cdot \text{grad} \ u = f.
\]

Terminology:

\[
\downarrow \quad \downarrow
\]

diffusive term  convective term

\(\sigma, \rho = \text{const.} \quad \Rightarrow \quad \text{non-dimensional (scaled) form: with } |\mathbf{v}| \leq 1, \epsilon > 0\)

\[
- \epsilon \Delta u + \mathbf{v} \cdot \text{grad} \ u = f \quad \text{in } \Omega.
\] (3.1.1)

Standard approach (Thm 1.6.1, → Sect. 1.6) \(\Rightarrow\) variational formulation of BVPs for (3.1.1)

For pure Dirichlet problem \((u = g, g \in H^{1/2}(\partial \Omega), \text{on } \partial \Omega)\):

\[
u \in \tilde{g} + H^1_0(\Omega): \quad \int_{\Omega} \epsilon \text{grad} \ u \cdot \text{grad} \ v + (\mathbf{v} \cdot \text{grad} \ u) \ v \, \text{d}x = \int_{\Omega} f \ v \, \text{d}x \quad \forall v \in H^1_0(\Omega).\] (3.1.2)
Linear variational problem (1.7.1) with non-symmetric bilinear form

$$a(u, v) := \int_{\Omega} \epsilon \mathbf{grad} u \cdot \mathbf{grad} v + (\mathbf{v} \cdot \mathbf{grad} u) v \, dx \quad u, v \in H^1(\Omega).$$  \hspace{1cm} (3.1.3)

$$\int_{\Omega} (\mathbf{v} \cdot \mathbf{grad} u) u \, dx = - \int_{\Omega} u \, \mathbf{div}(\mathbf{vu}) \, dx = - \int_{\Omega} u (\mathbf{v} \cdot \mathbf{grad} u + \mathbf{div}(\mathbf{v}) u) \, dx.$$

Lemma 3.1.1 (Ellipticity of bilinear form for convection-diffusion problems). If $\mathbf{div} \mathbf{v} = 0$, then $a(\cdot, \cdot)$ from (3.1.3) is $H^1_0(\Omega)$-elliptic.

Remark. $\mathbf{div} \mathbf{v} = 0$ means that the fluid is incompressible (e.g., water, oil).

Remark. Maximum principle

If $u \in C^0(\overline{\Omega})$ solves (3.1.1) with $f \equiv 0$ \Rightarrow $\inf_{y \in \partial \Omega} u(y) \leq u(x) \leq \sup_{y \in \partial \Omega} u(y)$.

3.1.2 Singular perturbation

Focus: Dirichlet problem for (3.1.1) with $|\mathbf{v}| \approx 1, \epsilon \ll 1$ (convection dominated case)
Model problem for $d = 1$: $\Omega := ]0, 1[, \ f = 0, \ v = 1$

\[-\varepsilon u'' + u' = 0 \quad \text{in } \Omega, \]
\[u(0) = 0 , \quad u(1) = 1 .\]

For $\varepsilon \ll 1$:

boundary layer at $x = 1$

For $\varepsilon \to 0$:

$|u|_m \to \infty \quad \forall m \in \mathbb{N}$.

Idea: Study limit problem, \textit{ie.} $\varepsilon = 0$:

- Imposing boundary conditions at $x = 0, 1$ impossible for limit problem!
Definition 3.1.2 (Singularly perturbed problem). A boundary value problem depending on parameter \( \epsilon \approx \epsilon_0 \) is called singularly perturbed, if the limit problem for \( \epsilon \to \epsilon_0 \) is ill-posed.

Especially in the case of elliptic boundary value problems:

Singular perturbation = low order terms become dominant for \( \epsilon \to \epsilon_0 \)

Guideline:

Numerical methods for singularly perturbed problems must “work” for the limit problem

The multidimensional case: \( \nabla \cdot \nabla u = f \) in \( \Omega \). (3.1.4)

(3.1.4) = linear hyperbolic equation (transport equation)
\( \gamma : [0, s_+] \to \Omega, \gamma(0) \in \partial \Omega, \gamma' = v \circ \gamma = \text{characteristic (curve) of (3.1.4)} \)

\[ u(\gamma(s)) = u(\gamma(0)) + \int_0^s f(\gamma(\tau)) \, d\tau, \quad s \in ]0, s_+[, \]

If \( f = 0 \Rightarrow \) boundary values carried along characteristics!

- For (3.1.4): Dirichlet boundary conditions only on

  inflow boundary \( \Gamma_{\text{in}} := \{ x \in \partial \Omega : n(x) \cdot v(x) < 0 \} \)

  or

  outflow boundary \( \Gamma_{\text{out}} := \{ x \in \partial \Omega : n(x) \cdot v(x) > 0 \} \).

If \( 0 < \epsilon \ll 1, |v| \approx 1 \Rightarrow \) expect boundary layers at outflow boundaries!

- If \( 0 < \epsilon \ll 1, |v| \approx 1 \Rightarrow \) “rough” boundary data will create internal layers!

Example:

\( \Omega = ]0, 1[^2, (3.1.1) \text{ with } v = \left( \frac{2}{1} \right), \epsilon = 10^{-4}, \)

Dirchlet b.c.: \( u = 1 \text{ on } \{ x_1 = 0 \} \cup \{ x_2 = 1 \}, u = 0 \text{ on } \{ x_1 = 1 \} \cup \{ x_2 = 0 \} \) (\( \not\in H^{1/2}(\partial \Omega) \))
Example:

Two steep layers:

- boundary layer at outflow boundary \( \{x_1 = 1\} \cap \{1/2 < x_2 < 1\} \) with mismatch of boundary values
- internal layer due to convected discontinuity of boundary data
Resolution of layers:

- use adapted meshes
- use slender cells aligned with layer

Mesh from [10]

3.1.3 Upwinding

Continued: model problem for $d = 1$: $\Omega := ]0, 1[ $, $f = 0$, $v = 1$

$$-\varepsilon u'' + u' = f \quad \text{in } \Omega \quad , \quad u(0) = 0 \quad , \quad u(1) = 1. $$

Galerkin discretization of (3.1.2) with linear Lagrangian FE on equidistant mesh

$$\mathcal{M} := \{ \xi_i, \xi_{i+1} : \xi_i = ih, i = 0, \ldots N \}, \; h = 1/N, \; N \in \mathbb{N}, \; \text{ of } ]0, 1[.$$
Difference stencil, approximate r.h.s. based on trapezoidal rule (→ Sect. 2.3.1):

\[
\left[ -\frac{\epsilon}{h} - \frac{1}{2} \quad \frac{2\epsilon}{h} - \frac{\epsilon}{h} + \frac{1}{2} \right]_{h,x=\xi_i} \odot \tilde{\mu} = hf(\xi_i) \quad i = 1, \ldots, N = 1,
\]

\[
\left\{ \frac{\epsilon}{h} \begin{bmatrix} -1 & 2 & -1 \end{bmatrix}_h + \frac{1}{2} \begin{bmatrix} -1 & 0 & 1 \end{bmatrix}_h, x = \xi_i \right\} \odot \tilde{\mu} = hf(\xi_i) \quad i = 1, \ldots, N = 1, \tag{3.1.5}
\]

where \( \mu_i \approx u(\xi_i) \). Boundary conditions: \( \mu_0 = 0, \mu_N = 1 \).

For \( f \equiv 0 \) ➤ linear difference equation

\[
\left( -\frac{2\epsilon}{h} - 1 \right) \mu_{i-1} + \frac{4\epsilon}{h} \mu_i + \left( -\frac{2\epsilon}{h} + 1 \right) \mu_{i+1} = 0 \quad i = 1, \ldots, N - 1.
\]

Try \( \mu_i = \zeta^i \) and use \( \mu_0 = 0, \mu_N = 1 \):

\[
\mu_i = \frac{1 - \left( \frac{2\epsilon + h}{2\epsilon - h} \right)^i}{1 - \left( \frac{2\epsilon + h}{2\epsilon - h} \right)^N}, \quad i = 0, \ldots, N.
\]
Exact solution, $\varepsilon=0.01$  
Galerkin solution, $N=30$

$\delta^0_1(M)$-Galerkin solution for convection-diffusion problem on $]0, 1[$.

Observations:
If $2\varepsilon < h$: oscillations in $u_N$ (no discrete maximum principle → Def. 2.3.1)

Even for $h = \varepsilon$: in the limit $h \to 0$

$$u(\xi_{N-1}) = \frac{e^{-(N-1)} - 1}{e^{-(N-1)} - e^1} \to e^{-1} \quad \iff \quad \mu_{N-1} = \frac{1 - 3^{N-1}}{1 - 3^N} \to \frac{1}{3}.$$  

What went wrong? [Consider limit case $\varepsilon = 0$]

(3.1.3): $u' = f \iff \frac{1}{2}(\mu_{i+1} - \mu_{i-1}) = hf(\xi_i) \quad i = 1, \ldots, N$,

Obvious (cf. $f \equiv 0$): flawed discretization for ODE $u' = f$!

Meaningful scheme for $u' = f$ on equidistant mesh:

Explicit/implicit Euler method: $\mu_{i+1} - \mu_i = hf(\xi_i) \quad i = 0, \ldots, N$.

Use one-sided difference quotients for discretization of convective term!

Which type? (Explicit or implicit Euler?)

Selection criterium: Ensure discrete maximum principle for resulting stencil

= upwinding
\(-\epsilon u'' + u' = f\)

[backward difference]
\[
\begin{bmatrix}
-\frac{\epsilon}{h} - 1 & \frac{2\epsilon}{h} + 1 & -\frac{\epsilon}{h}
\end{bmatrix}_{h,x=\xi_i} \circ \tilde{\mu} = hf(\xi_i)
\]

[forward difference]
\[
\begin{bmatrix}
-\frac{\epsilon}{h} & \frac{2\epsilon}{h} - 1 & -\frac{\epsilon}{h} + 1
\end{bmatrix}_{h,x=\xi_i} \circ \tilde{\mu} = hf(\xi_i)
\]

fits Def. 2.3.1

Rewriting upwind stencil:
\[
\begin{bmatrix}
-\frac{\epsilon}{h} - 1 & \frac{2\epsilon}{h} + 1 & -\frac{\epsilon}{h}
\end{bmatrix}_{h} = (\epsilon + h/2) \frac{1}{h} \begin{bmatrix}
-1 & 2 & -1
\end{bmatrix}_h + \frac{1}{2} \begin{bmatrix}
-1 & 0 & 1
\end{bmatrix}_h .
\]

Upwinding = \(h\)-dependent enhancement of diffusive term

artificial viscosity/diffusion
Upwind solution by (3.1.6) for convection-diffusion problem on ]0, 1[.
Convection-diffusion problem in higher dimensions, \( \mathbf{v} = \mathbf{v}(\mathbf{x}), |\mathbf{v}| \approx 1, \epsilon \ll 1: \)

\[
- \epsilon \Delta u + \mathbf{v} \cdot \nabla u = f \quad \text{in } \Omega \subset \mathbb{R}^d , \quad u = g \quad \text{on } \partial \Omega .
\]  

(3.1.1)

Isotropic “upwinding” in higher dimensions

\( \Rightarrow \) smearing of internal layers

Heat conduction:

Increasing \( \epsilon \) \( \Rightarrow \) transported layers in temperature distribution disappear away from \( \partial \Omega \)

excessive crosswind diffusion
Idea:
Anisotropic artificial diffusion

On cell $K$ replace:

$$
\epsilon \leftarrow \epsilon \mathbf{I} + \delta_K \mathbf{v}_K \mathbf{v}_K^T \in \mathbb{R}^{3,3}.
$$

$\mathbf{v}_K = $ local velocity (e.g., obtained by averaging)

“Streamline diffusion bilinear form” (= starting point for Galerkin-FE discretization):
\( a_{SD}(u, v) = \sum_{K \in \mathcal{M}} \int_{K} (\epsilon I + \delta_K \nabla_K \nabla_K^T) \nabla u \cdot \nabla v + v \cdot \nabla u \ n \ dx, \quad u, v \in H^1(\Omega) \). (3.1.7)

In the case of Galerkin discretization by means of linear Lagrangian FE (2D, \( V_N = \mathcal{S}_1^0(M) \)):

\[
\delta_K := \begin{cases} 
\epsilon^{-1} h_K^2 & \text{if } \frac{\|v\|_{K, \infty} h_K}{2\epsilon} \leq 1 \\
h_K & \text{if } \frac{\|v\|_{K, \infty} h_K}{2\epsilon} > 1.
\end{cases}
\]

Terminology: \( \frac{\|v\|_{K, \infty} h_K}{2\epsilon} = \text{Cell Péclet number} \)

### 3.2 The Helmholtz equation

Mathematical model for propagation of time-harmonic acoustic waves in isotropic homogeneous medium bounded by sound-soft walls (scaled equations)

\[
\Delta u + \kappa^2 u = f \quad \text{in } \Omega \subset \mathbb{R}^d, \quad u = 0 \quad \text{on } \partial \Omega.
\] (3.2.1)
\( u \) : complex amplitude of pressure [physical pressure \( p(t, \mathbf{x}) := \text{Re}\{u(\mathbf{x}) \exp(-i\omega t)\} \)]

\( f \) : sound source (loudspeaker)

\( \kappa \) : (normalized) wave number, \( \kappa = \frac{L\omega}{c} \)

\( \omega \) : angular frequency

\( c \) : speed of sound in medium

\( L \) : “size” of computational domain \( \Omega \), \( L \approx \text{diam}(\Omega) \)

\[
\begin{align*}
    u \in H^1_0(\Omega): \\
    a(u, v) := -\int_{\Omega} \nabla u \cdot \nabla v + \kappa^2 uv \, dx = \int_{\Omega} f v \, dx \quad \forall v \in H^1_0(\Omega). 
\end{align*}
\] (3.2.2)

\( a(\cdot, \cdot) \) not \( H^1_0(\Omega) \)-elliptic, but indefinite!

Note: plane wave solutions of homogeneous equation \( \Delta u + \kappa^2 u = 0 \):

\[
    u(\mathbf{x}) = \exp(i\mathbf{k} \cdot \mathbf{x}) \quad \mathbf{x} \in \mathbb{R}^d, \quad |\mathbf{k}| = \kappa \quad \text{(dispersion relation)}.
\]

If \( \kappa \gg 1 \Rightarrow \) singularly perturbed problem: “limit problem: \( u = 0 \)”
Model problem for $d = 1$: $\Omega := ]0, 1[$, $f = 0$, $u(0) = 1$

Absorbing boundary conditions at $x = 1$: $i \kappa u(1) = u'(1)$.

$\Rightarrow$ $u(x) = \exp(i \kappa x)$ (Wavelength $2\pi/\kappa$).

Galerkin discretization: linear Lagrangian FE on equidistant mesh

$\mathcal{M} := \{\xi_j, \xi_{j+1} | \xi_j = jh, j = 0, \ldots, N\}$, $h = 1/N$, $N \in \mathbb{N}$, of $]0, 1[$.

Difference stencils (\implies Sect. 2.3.1):

\[
S_h := \begin{bmatrix}
\frac{h \kappa^2}{6} + \frac{1}{h} \quad \frac{2h \kappa^2}{3} - \frac{2}{h} \quad \frac{h \kappa^2}{6} + \frac{1}{h}
\end{bmatrix}
\]

for $j = 1, \ldots, N - 1$, \hspace{1cm} (3.2.3)

\[
\begin{bmatrix}
\frac{h \kappa^2}{6} + \frac{1}{h} \quad \frac{h \kappa^2}{3} + i \kappa - \frac{1}{h}
\end{bmatrix}
\]

Norm related to (3.2.2):

\[\|v\|^2_\kappa := \kappa^{-2} \|v\|^2_1 + \|v\|^2_0\]

Estimate: error of p.w. linear interpolation on $\mathcal{M}$: with $C > 0$ independent of $\kappa, h$

\[
\begin{align*}
|u - l_1 u|_1 & \leq C h |u|_2 \leq C h \kappa^2, \\
\|u - l_1 u\|_0 & \leq C h |u|_1 \leq C h \kappa.
\end{align*}
\]

\[\Rightarrow\|u - l_1 u\|_\kappa \leq C h \kappa.\]

If Galerkin-FE solution quasi-optimal (indepedent of $\kappa$!), then $\|u - u_N\|_\kappa = O(h \kappa)$. 

Experiment 3.2.1. $L^2(\Omega)$-norm of the finite element discretization error for various $N$, $h\kappa = \text{const}$
\[ \|u - u_N\|_0^2: \text{blue: } h\kappa = 1/5, \text{magenta: } h\kappa = 1/10, \text{red: } h\kappa = 1/20 \]
Helmholtz BVP: real parts of exact solution (blue) and p.w. linear Galerkin solution (magenta) for 
\( \kappa = 20, N = 50 \)
Helmholtz BVP: real parts of exact solution (blue) and p.w. linear Galerkin solution (magenta) for $\kappa = 40, N = 50$
Helmholtz BVP: real parts of exact solution (blue) and p.w. linear Galerkin solution (magenta) for 
\( \kappa = 60, \ N = 50 \)

Observation: 
- \( \|u - u_N\|_0 \sim \kappa \), even if \( h\kappa = \text{const.} \).
- Phase error of Galerkin solution (grows as \( \kappa \) increases)

Numerical dispersion

Gauging numerical dispersion:

Consider difference stencil (3.2.3) on infinite translation invariant grid:

\[ \mathcal{M} := \{j\xi, j\xi + 1[: \xi_j := jh, \ j \in \mathbb{Z}, \ h > 0 \} . \]

“Plane wave grid function”

\( \eta_j = \exp(i\omega jh), \ j \in \mathbb{Z}, \ |\omega| < \pi/h \)

Idea: Determine discrete wave number \( \omega = \omega(h, \kappa) \) such that \( S_H\tilde{\eta} = 0 \).

\[
\left[ \frac{h\kappa^2}{6} + \frac{1}{h} \frac{2h\kappa^2}{3} - \frac{2}{h} \frac{h\kappa^2}{6} + \frac{1}{h} \right]_h \odot \tilde{\eta} = \frac{1}{h} \left( 2 \cos(\omega h) - 2 + \frac{\kappa^2 h^2}{6} (4 + 2 \cos(\omega h)) \right) \tilde{\eta} .
\]

\[ S_H \tilde{\eta} = 0 \ \Rightarrow \ \omega = \frac{1}{h} \cos^{-1} \left( \frac{6 - 2\kappa^2 h^2}{6 + \kappa^2 h^2} \right) = \kappa - \frac{1}{24} \kappa (\kappa h)^2 + O(\kappa (\kappa h)^4) . \]
Numerical dispersion: red: $h\kappa = 1/3$, blue: $h\kappa = 1/5$, green: $h\kappa = 1/10$
Relationship between numerical dispersion and $L^2(\Omega)$-error:

\[
\int_0^1 |\exp(i\kappa x) - \exp(i\omega x)|^2 \, dx = 2 \int_0^1 \sin^2(1/2(\kappa - \omega)x) \, dx = (\kappa - \omega)(1 - 1/2 \sin(\kappa - \omega)) .
\]

\[
\|u - u_N\|_0 \approx \sqrt{|\kappa - \omega|} \sim \sqrt{\text{phase error}}
\]

3.3 The Stokes Problem

Mathematical model for slow stationary flow of viscous incompressible fluid:

\[
-\nu \Delta u - \text{grad } p = f \quad \text{in } \Omega , \tag{3.3.1}
\]

\[
\text{div } u = 0 \quad \text{in } \Omega , \tag{3.3.2}
\]

\[
u u = 0 \quad \text{on } \partial \Omega . \tag{3.3.3}
\]
\( \Omega \): container = computational domain \( \subset \mathbb{R}^d, d = 2, 3, \) connected

\( \mathbf{u} \): velocity field \( \Omega \mapsto \mathbb{R}^d, [\mathbf{u}] = \text{ms}^{-1} \)

\( p \): pressure, \( [p] = \text{Nm}^{-2} \)

\( \mathbf{f} \): force density, \( [\mathbf{f}] = \text{Nm}^{-3} \)

\( \nu > 0 \): kinematic viscosity, \( [\nu] = \text{Nsm}^{-2} \)

Notation: \( \Delta = \text{vectorial Laplacian (applied to Cartesian components of vectorfield)} \)

(3.3.1) \( \rightarrow \) conservation of linear momentum

(3.3.2) \( \rightarrow \) incompressibility condition

(3.3.3) \( \rightarrow \) no-slip boundary conditions

Note: pressure \( p \) unique only up to constants!

### 3.3.1 Mixed variational formulation

Minimization problem underlying (3.3.1)-(3.3.3) (\( \rightarrow \) Sect. 1.9, “Dirichlet principle”):
Hilbert space $V := \{v \in (H^1_0(\Omega))^d : \text{div } v = 0 \}$.

Stokes problem (3.3.1)-(3.3.3) equivalent to

$$u \in V : u = \arg \min_{v \in V} J(v), \quad J(v) := \int_\Omega \frac{1}{2} \| \nabla v \|_F^2 - f \cdot v \, dx,$$  \hspace{1cm} (3.3.4)

where $\nabla v = \text{Jacobi matrix } (\frac{\partial v_i}{\partial x_j})_{i,j=1}^d : \Omega \mapsto \mathbb{R}^{d,d}$, $\| \cdot \|_F = \text{Frobenius norm of matrix}$, $|M|_F^2 = M : M$, with $A : B := \sum_{i,j} (A)_{ij} (B)_{ij}$.

$$(3.3.4) = \text{ (linearly) constrained minimization problem}$$

Idea: introduce Lagrangian multiplier $\in L^2_*(\Omega)$

[general strategy for treating (linear) constraints]

Notation:

$$L^2_*(\Omega) := \{v \in L^2(\Omega) : \int_\Omega v \, dx = 0\}$$

$$(3.3.4) \Leftrightarrow u = \arg \inf_{w \in (H^1_0(\Omega))^d} \sup_{z \in L^2_*(\Omega)} \tilde{J}(w, z), \quad \tilde{J}(w, z) := J(w) + \int_\Omega \text{div } w \, z \, dx.$$  \hspace{1cm} (3.3.5)

Saddle point problem: necessary conditions for stationary point $(u, p)$

$$(D_w \tilde{J})(u, p) = 0 \quad , \quad (D_z \tilde{J})(u, p) = 0.$$
Mixed variational formulation of Stokes problem: seek \( u \in (H^1_0(\Omega))^d, \, p \in L^2_*(\Omega) \)

\[
\begin{align*}
\int_{\Omega} \nu \nabla u : \nabla v \, dx &+ \int_{\Omega} \div v \cdot p \, dx = \int_{\Omega} f \cdot v \, dx \quad \forall v \in (H^1_0(\Omega))^d, \\
\int_{\Omega} \div u \, q \, dx & = 0 \quad \forall q \in L^2_*(\Omega).
\end{align*}
\]

= variational saddle point problem

Remark. Componentwise integration by parts (\( \rightarrow \) Thm. 1.6.1) of first equation of (3.3.6) recovers (3.3.1).

Parlance: (3.3.6) = “mixed” \( \iff \) two different spaces \((H^1_0(\Omega))^d, \, L^2_*(\Omega)\) involved
3.3.2 Saddle point problems

General structure: Hilbert spaces $V, Q$, bilinear forms $a : V \times V \mapsto \mathbb{R}$, $b : V \times Q \mapsto \mathbb{R}$, linear forms $f : V \mapsto \mathbb{R}$, $g : Q \mapsto \mathbb{R}$

\[
\begin{align*}
\mathbf{u} \in V & : \quad a(\mathbf{u}, \mathbf{v}) + b(\mathbf{v}, p) = f(\mathbf{v}) \quad \forall \mathbf{v} \in V, \\
p \in Q & : \quad b(\mathbf{u}, q) = g(q) \quad \forall q \in Q. 
\end{align*}
\]  

(3.3.7)

Analogue of Lax-Milgram Lemma Thm. \[1.7.8\] for variational saddle point problems:
Theorem 3.3.1. If the bilinear forms $a, b$ are continuous, and satisfy

- the inf-sup condition

$$\exists \beta > 0: \sup_{v \in V} \frac{b(v, q)}{\|v\|_V} \geq \beta \|q\|_Q \quad \forall q \in Q,$$  \hspace{1cm} (3.3.8)

- and the ellipticity on the kernel

$$\exists \alpha > 0: a(v, v) \geq \alpha \|v\|^2_V \quad \forall v \in \mathcal{N}(B) := \{w \in V : b(w, q) = 0 \ \forall q \in Q\},$$  \hspace{1cm} (3.3.9)

then for any r.h.s. $f, g$ (3.3.7) has a unique solution $(u, p) \in V \times Q$, which satisfies

$$\|u\|_V \leq \frac{1}{\alpha} \sup_{v \in V} \frac{f(v)}{\|v\|_V} + \frac{1}{\beta} \left(1 + \frac{C_A}{\alpha}\right) \sup_{q \in Q} \frac{g(q)}{\|q\|_Q},$$

$$\|p\|_Q \leq \frac{1}{\beta} \left(\sup_{v \in V} \frac{f(v)}{\|v\|_V} + C_A \|u\|_V\right).$$

Assumptions can be verified for variational Stokes problem (3.3.6), see [4, III, § 5].

Galerkin discretization:

$V \to V_N \subset V$, $N := \dim V_N < \infty$ + bases

$Q \to Q_N \subset Q$, $M := \dim Q_N < \infty$
Discrete variational saddle point problem:

\[ u_N \in V_N : \quad a(u_N, v_N) + b(v_N, p_N) = f(v_N) \quad \forall v_N \in V_N , \]
\[ p_N \in Q_N : \quad b(u_N, q_N) = g(q_N) \quad \forall q_N \in Q_N . \]

Linear system (saddle point form), \( N + M \) equations: with \( A \in \mathbb{R}^{N,N} \), \( B \in \mathbb{R}^{N,M} \)

\[
\begin{pmatrix}
A & B \\
B^T & 0
\end{pmatrix}
\begin{pmatrix}
\vec{\mu} \\
\vec{\pi}
\end{pmatrix}
=
\begin{pmatrix}
\vec{\varphi} \\
\vec{\gamma}
\end{pmatrix}.
\]

If \( A = A^T \) ➞ symmetric, indefinite linear system of equations

Analogue of Cea's lemma \( \rightarrow \) Thm. [2.1.1]:

\[ p. 279 \]
Theorem 3.3.2 (Babuška-Brezzi theory). Let the assumption of Thm. 3.3.1 be satisfied and assume the Babuška-Brezzi conditions

\[ \exists \beta_N > 0 : \sup_{v_N \in V_N} \frac{b(v_N, q_N)}{\Vert v_N \Vert_V} \geq \beta_N \Vert q_N \Vert_Q \quad \forall q_N \in Q_N , \]  

(LBB1)

\[ \exists \alpha_N > 0 : \quad a(v_N, v_N) \geq \alpha_N \Vert v_N \Vert_V^2 \quad \forall v_N \in N(B_n) , \]  

(LBB2)

are satisfied, where

\[ N(B_n) := \{ v_N \in V_N : b(v_N, q_N) = 0 \quad \forall q_N \in Q_N \} . \]  

(3.3.11)

Let \((u, p) \in V \times Q\) and \((u_N, p_N) \in V_N \times Q_N\) stand for the unique solutions of (3.3.7) and (3.3.10), respectively. Then

\[ \| u - u_N \|_V + \| p - q_N \|_Q \leq C(C_A, C_B, \alpha_N, \beta_N) \left\{ \inf_{v_N \in V_N} \| u - v_N \|_V + \inf_{q_N \in Q_N} \| p - q_N \|_Q \right\} . \]

Moreover, if \( N(B_n) \subset N(B) \), then

\[ \| u - u_N \|_V \leq C(C_A, C_B, \alpha_N, \beta_N) \inf_{v_N \in V_N} \| u - v_N \|_V . \]  

(3.3.12)

For Stokes problem (3.3.6): Poincaré-Friedrichs inequality (Thm. 1.7.9) ➤ (LBB2)
Necessary condition for (LBB1): \( \dim Q_N \leq \dim V_N \)

Babuška-Brezzi conditions (LBB1) and (LBB2) entail matching of \( V_N \) and \( Q_N \)

What does this mean for the Stokes problem (3.3.6)?
Example (violation of \((LBB1)\)):

Regular triangular mesh of \(\mathcal{M} = ]0, 1[^2\).

Finite element spaces:

\[
V_N := (\delta^{0}_{1,0}(\mathcal{M}))^2,
Q_N := \{\mathcal{M}\text{-piecewise constants}\}.
\]

If \(K \in \mathbb{N}\) mesh cells in one coordinate direction,

\[
\dim V_N = 2(K - 1)^2, \quad \dim Q_N = 2K^2 - 1.
\]

\[
\dim Q_N > \dim V_N
\]

### 3.3.3 Stable finite element schemes for the Stokes problem

Setting:

\(d = 2\), FE-Galerkin discretization of \((3.3.6)\) in triangular meshes of \(\Omega\)
Definition 3.3.3 (Stability of Stokes FE). A finite element scheme for (3.3.6) is stable, if positive $\beta_N$ and $\alpha_N$ from (LBB1) and (LBB2), respectively, can be chosen independently of discretization parameters.

Thm. 3.3.2 ➤ stability prerequisite for quasi-optimality

Goal: Stable $V_N$-$Q_N$ pair with $V_N$ as small as possible ("balanced pair")

[Thm. 3.3.2 ➤ also $Q_N$ limits accuracy !]

Example: MINI-element

$$V_N := \left\{ v \in (H^1_0(\Omega))^2 : v|_K \in (P_1(K))^2 + \text{Span} \left\{ \lambda_1^K \lambda_2^K \lambda_3^K \right\}^2 \forall K \in \mathcal{M} \right\},$$

$$Q_N := \mathcal{H}_1^0(\mathcal{M}) \cap L^2(\Omega).$$
Terminology: for triangle $K$: $\lambda_1^K \lambda_2^K \lambda_3^K = \text{bubble function}$

$V_N = \text{bubble augmented linear Lagrangian finite element space}$

Example: Nonconforming Crouzeix-Raviart element
Definition 3.3.4 (Non-conforming FE). (→ Def. 2.1.12) A finite element space $V_N$ for a variational problem posed on the function space $V$ is called non-conforming, if $V_N \nsubseteq V$.

\[
V_N := \left\{ \begin{array}{l}
v \in (L^2(\Omega))^2: v|_K \in (P_1(K))^2 \ \forall K \in \mathcal{M}, \\
v \text{ continuous at midpoints of interior edges, } \\
v = 0 \text{ at midpoints of boundary edges} \end{array} \right\}
\]

\[
Q_N := \{ q \in L^2(\Omega): q|_K \in P_0(K) \ \forall K \in \mathcal{M}, \int_{\Omega} q \, d\mathbf{x} = 0 \}.
\]

Lemma 2.1.11 $V_N \nsubseteq (C^0(\Omega))^2 \Rightarrow V_N \nsubseteq (H^1(\Omega))^2 \Rightarrow (3.3.13) = \text{non-conforming}$

Observation: continuity of averages across edges for $v \in V_N$  

= relaxed compatibility condition for $H^1(\Omega)$

Remark. global d.o.f. for $V_N$ = point evaluations at midpoints of interior edges
Global shape function (for one velocity component) in $V_N$

- associated with edge
- discontinuous
- continuous at midpoint of edge
How to enforce the constraint \( p_N \in L^2_\ast(\Omega) \)?

cf. Rem. 2.1.8: no local basis functions for locally polynomial subspace

Idea: enforce linear constraint \( \int_{\Omega} p \, dx = 0 \) by means of Lagrangian multiplier \( \lambda \in \mathbb{R} \).

Augmented saddle point problem (see (3.3.5)):

\[
\mathbf{u} = \arg\inf_{\mathbf{w} \in (H^1_0(\Omega))^d} \sup_{z \in L^2(\Omega)} \inf_{\eta \in \mathbb{R}} \tilde{J}(\mathbf{w}, z) + \eta \int_{\Omega} z \, dx.
\]
Augmented variational saddle point problem:

\[
\int_{\Omega} v \nabla u : \nabla v \, dx + \int_{\Omega} \text{div} v \ p \, dx = \int_{\Omega} f \cdot v \, dx \quad \forall v \in (H^1_0(\Omega))^d ,
\]

\[
\int_{\Omega} \text{div} u \ q \, dx + \lambda \int_{\Omega} q \, dx = 0 \quad \forall q \in L^2(\Omega) ,
\]

\[
\int_{\Omega} p \, dx = 0
\].

3.4 Non-linear elliptic boundary value problems

No universal theory of “non-linear elliptic boundary value problems”: too huge a class

Discussion of a few specimens only

Not addressed:  
- Multiple solutions  
- Stable, instable solutions and bifurcation  
- ...  

“Non-linear phenomena”
3.4.1 Examples

Example 1: Stationary heat conduction & radiation b.c. ⇒ Sect. 1.3

\[-\Delta u = f \quad \text{in } \Omega \subset \mathbb{R}^2 , \quad \text{grad} u \cdot n = -\alpha |u - u_0|(u - u_0)^3 \quad \text{on } \partial \Omega .\]

[Multiply with test function & integrate by parts, see Sect. 1.6]

Non-linear variational problem (→ (1.6.6)):

\[ u \in H^1(\Omega) : \quad \int_{\Omega} \text{grad} u \cdot \text{grad} v \, dx + \int_{\partial \Omega} \alpha |u - u_0|(u - u_0)^3 v \, dS = \int_{\Omega} f v \, dx \quad \forall v \in H^1(\Omega) . \]

Notation: \[a(u; v)\]

Related minimization problem ("Dirichlet principle", see Sect. 1.9)

\[ u = \arg \min_{v \in H^1(\Omega)} J(v) , \quad J(v) := \frac{1}{2} \int_{\Omega} |\text{grad} v|^2 \, dx + \frac{1}{5} \int_{\partial \Omega} \alpha |v - u_0|^5 \, dS - \int_{\Omega} f v \, dx . \quad (3.4.2)\]
Definition 3.4.1 (Convex/coercive functional). \( V \) Hilbert/Banach space, \( J : V \mapsto \mathbb{R} \) = strictly convex, if

\[
J(tv + (1 - t)u) < tJ(v) + (1 - t)J(u) \quad \forall u, v \in V, \quad \forall t \in [0, 1].
\]

\( J : V \mapsto \mathbb{R} \) = coercive, if

\[
\frac{J(v)}{\|v\|_V} \to \infty \quad \text{for} \quad \|v\|_V \to \infty.
\]
Theorem 3.4.2 (Convex & coercive minimization problem). \( V \) Hilbert space, \( J : V \rightarrow \mathbb{R} \) coercive & strictly convex:

\[ \Rightarrow \exists_1 u \in V : J(u) \leq J(v) \quad \forall v \in V . \]

Lemma 3.4.3. \( J : H^1(\Omega) \rightarrow \mathbb{R} \) from (3.4.2) is Frechet-differentiable, coercive & strictly convex.

Proof. \( \rightarrow \) exercise \( \square \)

Abstract framework:

\( a : V \times V \rightarrow \mathbb{R} \) continuous, linear in second argument, \( f : V \rightarrow \mathbb{R} \) continuous linear form:

non-linear variational problem: \( u \in V : a(u; v) = f(v) \quad \forall v \in V . \) \hfill (3.4.3)

Galerkin discretization (see Sect. 2.1.1): subspace \( V_N \subset V, \dim V_N < \infty \)

\( u_N \in V_N : a(u_N; v_N) = f(v_N) \quad \forall v_N \in V_N . \) \hfill (3.4.4)
[Choosing basis for $V_N$]

Non-linear system of equations

For example: use $V_N = \delta_1^0(\mathcal{M})$ for Galerkin discretization of (3.4.1).

Example 2: Minimal surface problem for graphs over $\Omega \subset \mathbb{R}^2$

Surfaces $\Sigma := \{(x, v(x)) \in \mathbb{R}^3: v : \Omega \mapsto \mathbb{R} \text{ continuous, } v = g \text{ on } \partial \Omega \}$.

Area functional

$$J(v) = \int_{\Omega} \sqrt{1 + |\nabla v|^2} \, dx.$$  \hfill (3.4.5)

minimal surface problem:

$$u = \arg \min_{v \in H^1(\Omega), v|_{\partial \Omega} = g} J(v).$$  \hfill (3.4.6)

$J : H^1(\Omega) \mapsto \mathbb{R}$ continuous, strictly convex, but not coercive!

[> intricate functional framework, see [5, Sect. 5.2] ]

Frechet derivative:

$$J'(u; v) = \int_{\Omega} \frac{\nabla u \cdot \nabla v}{\sqrt{1 + |\nabla u|^2}} \, dx, \quad v \in H^1_0(\Omega)$$
Variational minimal surface problem: for \( g \in C^0(\partial \Omega) \cap H^{1/2}(\partial \Omega) \)

\[
   u \in H^1(\Omega) \, , u|_{\partial \Omega} = g : \quad \int_{\Omega} \frac{\nabla u \cdot \nabla v}{\sqrt{1 + |\nabla u|^2}} \, dx = 0 \quad \forall v \in H^1_0(\Omega) .
\] (3.4.7)

\[\blacksquare\]

Galerkin discretization: \( H^1(\Omega) \rightarrow \mathcal{V}^1(M) \), \( u_N = g_N \) on \( \partial \Omega \), see Sect. 2.2.7

### 3.4.2 Iterative solution methods

Nonlinear system of equations:

\[
   A(\vec{\xi}) = \vec{\varphi} \text{ where } A : \mathbb{R}^N \mapsto \mathbb{R}^N , \ \vec{\varphi} \in \mathbb{R}^N .
\] (2.3.8)

Assume: \( \exists \vec{\mu} \in \mathbb{R}^N : \ A(\vec{\mu}) = \vec{\varphi} , \ A \text{ continuously differentiable} \)

Remark: If (2.3.8) \( \leftrightarrow \) minization problem

\[\blacksquare\] amenable to algorithms of non-linear optimizations
Definition 3.4.4. A mapping \( \Phi : \mathbb{R}^N \mapsto \mathbb{R}^N \) defines an iteration

\[
\tilde{\mu}^{(k+1)} := \Phi(\tilde{\mu}^{(k)}) , \quad \tilde{\mu}^{(0)} \in \mathbb{R}^N ,
\]

that is consistent with \( A(\tilde{\xi}) = \tilde{\varphi} \), if

\[
\mu \in \mathbb{R}^N : \quad \Phi(\tilde{\mu}) = \tilde{\mu} \iff A(\tilde{\mu}) = \tilde{\varphi}.
\]

Definition 3.4.5. An iteration \( \tilde{\mu}^{(k+1)} := \Phi(\tilde{\mu}^{(k)}) \) converges to \( \tilde{\mu}^* \in \mathbb{R}^N \)

- **linearly**, if \( \exists 0 < L < 1 : |\tilde{\mu}^{(k+1)} - \tilde{\mu}^*| \leq L|\tilde{\mu}^{(k)} - \tilde{\mu}^*| \quad \forall k \in \mathbb{N} , \)

- **quadratically**, if \( \exists C > 0 : |\tilde{\mu}^{(k+1)} - \tilde{\mu}^*| \leq C|\tilde{\mu}^{(k)} - \tilde{\mu}^*|^2 \quad \forall k \in \mathbb{N} . \)

Important concepts:

- local convergence \( \iff \) global convergence

3.4.2.1 Fixed point iteration
Special structure of (2.3.8): \[ A(\vec{\xi}) = \tilde{A}(\vec{\xi}, \vec{\xi}) , \quad \tilde{A} \text{ linear in 2nd argument}. \]

Fixed point iteration for \( A(\vec{\xi}) = \phi \): \[ \Phi(\vec{\xi}) = \tilde{A}(\vec{\xi}, \cdot)^{-1} \phi. \]

Example: minimal graph surface problem (3.4.7):

\[ \tilde{a}(z, u; v) := \int_{\Omega} \frac{\nabla u \cdot \nabla v}{\sqrt{1 + |\nabla z|^2}} \, dx , \quad u, z \in H^1(\Omega), \{u, z\}_{\partial \Omega} = g, \, v \in H^1_0(\Omega). \]

In each step of fixed point iteration: solve discrete linear variational problem

\[ u_{N}^{(k+1)} \in \delta^0_{1,0}(M) + \tilde{g}_N : \quad \int_{\Omega} \frac{\nabla u_{N}^{(k+1)} \cdot \nabla v_N}{\sqrt{1 + |\nabla u_{N}^{(k)}|^2}} \, dx = 0 \quad \forall v_N \in \delta^0_{1,0}(M). \]

(On algebraic level)

solve linear system of equations in each step

Remark. If convergent, fixed point iteration “usually” displays linear convergence

[General convergence theory elusive]
3.4.2.2 Newton’s method

Assumption:

\[ A : \mathbb{R}^N \mapsto \mathbb{R}^N \text{ from (2.3.8) continuously differentiable} \]

Idea: Linearization around “current iterate \( \overline{\xi}^{(k)} \)” ➤ correction equation

\[ \overline{\xi}^{(k+1)} \in \mathbb{R}^N : \quad A(\overline{\xi}^{(k)}) + A'(\overline{\xi}^{(k)})(\overline{\delta}) = \overline{\varphi} \quad , \quad \overline{\xi}^{(k+1)} = \overline{\xi}^{(k)} + \overline{\delta} . \quad (3.4.8) \]

correction equation = linear system of equations for Jacobi matrix \( A'(\overline{\xi}) \)

**Theorem 3.4.6.** If \( A : \mathbb{R}^N \mapsto \mathbb{R}^N \) is continuously differentiable, \( A(\overline{\mu}) = \overline{\varphi} \), and \( A'(\overline{\mu}) \in \mathbb{R}^{N \times N} \) regular, then the sequence of iterates \( \overline{\xi}^{(k)} \), \( k \in \mathbb{N}_0 \), converges quadratically to \( \overline{\mu} \) provided that \( \overline{\xi}^{(0)} \) is sufficiently close to \( \overline{\mu} \).

Problem: quadratic convergence only in very small neighborhood of \( \overline{\mu} \)

Remedy: damped Newton method: for damping factors \( s_k \in ]0, 1] \)

\[ \overline{\xi}^{(k+1)} \in \mathbb{R}^N : \quad A(\overline{\xi}^{(k)}) + A'(\overline{\xi}^{(k)})(\overline{\delta}) = \overline{\varphi} \quad , \quad \overline{\xi}^{(k+1)} = \overline{\xi}^{(k)} + s_k \overline{\delta} . \quad (3.4.9) \]
How to determine $s_k$ ("damping strategy") in (3.4.9)?

Example: Armijo’s rule based on some (appropriate) norm $\| \cdot \|$ on $\mathbb{R}^N$ ($\tau = \text{tolerance}$):

$$
\tilde{x}^{(0)} \in \mathbb{R}^N; k := 0; \quad \text{initial guess}
$$

do {
    \[ \tilde{d} := A'\left(\tilde{x}^{(k)}\right)^{-1}(\tilde{\varphi} - A(\tilde{x}^{(k)})); \]
    \[ s := 1; \]
    repeat {
        \[ \text{if } (s \leq s_{\text{min}}) \text{ stop} \]
        \[ \tilde{x}^{(k+1)} := \tilde{x}^{(k)} + s \tilde{d}; \]
        \[ \tilde{d}^{\star} := A'\left(\tilde{x}^{(k)}\right)^{-1}(\tilde{\varphi} - A(\tilde{x}^{(k+1)})); \]
        \[ s := \frac{1}{2}s; \]
    }
    until \( \| \tilde{d}^{\star} \| \leq (1 - s)\| \tilde{d} \| \);
    \[ k := k + 1; \]
} while \( \| \tilde{d} \| > \tau \); \quad \text{termination criterion}
Computation of Newton correction for \((3.4.4) \ (u \mapsto a(u; \cdot))\) assumed to be continuously differentiable:

\[ u_N \in V_N: \quad a(u_N; v_N) = f(v_N) \quad \forall v_N \in V_N. \tag{3.4.4} \]

\[ \delta := A'(\vec{\mu})^{-1}\vec{\zeta} = \text{coefficient vector (w.r.t. to some basis of } V_N) \text{ of solution of} \]

\[ w_N \in V_N: \quad b(u_N; w_N, v_N) = z(v_N) \quad \forall v \in V_N, \tag{3.4.10} \]

where \(u_N \in V_N \leftrightarrow \vec{\mu}, \vec{\zeta} \leftrightarrow \text{linear form } z : V_N \leftrightarrow \mathbb{R}^N, \text{ and} \]

\[ b(u_N; w_N, v_N) = \lim_{t \to 0} \frac{a(u_N + tw_N; v_N) - a(u_N; v_N)}{t}. \tag{3.4.11} \]

Example: minimal graph surface variational problem \((3.4.7), V_N = \mathcal{S}^0_{1,0}(\mathcal{M}): \)

\[ a(u; v) = \int_{\Omega} \frac{\text{grad } u \cdot \text{grad } v}{\sqrt{1 + |\text{grad } u|^2}} \, dx, \quad u \in H^1(\Omega), u|_{\partial \Omega} = g, \ v \in H_0^1(\Omega). \]
\[ b(u; w, v) = \lim_{t \to 0} \frac{1}{t} \int_{\Omega} \frac{\text{grad}(u + tw) \cdot \text{grad} v}{\sqrt{1 + |\text{grad}(u + tw)|^2}} - \frac{\text{grad} u \cdot \text{grad} v}{\sqrt{1 + |\text{grad} u|^2}} \, dx \]

\[ = \int_{\Omega} \frac{\partial}{\partial u} \left\{ \frac{\text{grad} u \cdot \text{grad} v}{\sqrt{1 + |\text{grad} u|^2}} \right\} (w) \, dx \]

\[ = \int_{\Omega} \frac{\text{grad} w \cdot \text{grad} v}{\sqrt{1 + |\text{grad} u|^2}} - \frac{(\text{grad} u \cdot \text{grad} v) (\text{grad} u \cdot \text{grad} w)}{(1 + |\text{grad} u|^2)^{3/2}} \, dx. \]

Computing Newton update from \( A'(\tilde{\mu}) \delta = \tilde{\varphi} \iff \text{seek } w_N \in V_N \iff \delta: \)

\[ \int_{\Omega} \frac{\text{grad} w_N \cdot \text{grad} v_N}{\sqrt{1 + |\text{grad} u_N|^2}} - \frac{(\text{grad} u_N \cdot \text{grad} v_N) (\text{grad} u_N \cdot \text{grad} w_N)}{(1 + |\text{grad} u_N|^2)^{3/2}} \, dx = f(v_N) \quad \forall v_N \in V_N. \]

= discrete linear variational problem (1.7.1) with symmetric positive definite (\( \to \text{Def. 1.7.3} \)) bilinear form (\( \to \text{exercise} \)).
4.1 Linear stationary iterative methods

Discrete variational problem (→ Sect. 2.1.1)

\[ V_N \subset V, \quad \dim V_N := N < \infty, \quad u_N \in V_N: \quad a(u_N, v_N) = f(v_N) \quad \forall v_N \in V_N \quad (2.1.1) \]
• \( a : V_N \times V_N \rightarrow \mathbb{R} \) symmetric, \( V \)-elliptic bilinear form (→ Def. 1.7.7)

• \( f : V_N \rightarrow \mathbb{R} \) continuous linear form

• Ordered basis \( \mathcal{B}_N := \{ b_1^N, \ldots, b_N^N \} \subset V_N \).
  \[ (2.1.1) \quad \Rightarrow \quad \text{linear system of equations } A \tilde{\mu} = \bar{\varphi} \text{ (see Sect. 2.1.1)} \]

• Given: “Initial guess” = crude approximation of \( u_N \): \( u_N^{(0)} \in V_N \leftrightarrow \tilde{\mu}^{(0)} \in \mathbb{R}^N \)

  Idea: iterated successive subspace correction

  Given: decomposition of trial space

  \[ V_N = U_1 + \cdots + U_J , \quad U_i \subset V_N \text{ subspace, } J \in \mathbb{N} . \quad (4.1.1) \]

  Improve \( u_N^{(0)} \) by successive corrections “in the direction of the subspaces \( U_i \”).

Algorithm: Iterative method for (2.1.1): successive subspace correction in the direction of \( U_i \)

(\( \tau = \) tolerance, termination threshold, \( \| \cdot \| \) = some norm on \( V_N \))
Initial guess $u_{N}^{(0)} \in V_{N}$

$k=0$

**do** {

$k=k+1$

$u_{N}^{(k)} := u_{N}^{(k-1)}$

**for** ($i = 1; i \leq J; i ++$) {

$c_{i} \in U_{i}: \quad a(u_{N}^{(k)} + c_{i}, v_{i}) = f(v_{i}) \quad \forall v_{i} \in U_{i}$  \hfill (*)

$u_{N}^{(k)} := u_{N}^{(k)} + c_{i}$

}}

**while** ($\|u_{N}^{(k)} - u_{N}^{(k-1)}\| > \tau$)  \hfill termination criterion

(4.1.2)

\((*) \iff\) Solving error equation (7.2.3) in subspace $U_{i}$:

$c_{i} \in U_{i}: \quad a(c_{i}, v_{i}) = r(v_{i}): \quad r(v) := f(v) - a(u_{N}^{(k)}, v) =\text{weak residual, Sect. 7.2.1.}$

**Remark.** Sect. 1.9  \(\geq\) (2.1.1) related to minimization problem

\[ u_{N} = \arg \min_{v_{N} \in V_{N}} J(v_{N}), \quad J(v_{N}) := 1/2a(v_{N}, v_{N}) - f(v_{N}). \]  \hfill (4.1.3)
\((*) \iff \text{minimization of } J \text{ in the direction of } U_i:\)
\[
c_i = \arg \min_{v_i \in U_i} J(u_N^{(k)} + v_i).
\]

**Convergence of iteration (4.1.2):** \( u_N^{(k)} \to u_N \) as \( k \to \infty \), \( u \) solves (2.1.1) (Why? \( \rightarrow \) exercise)

**Algebraic perspective:**

- \( \mathcal{B}_i := \{ q_i^1, \ldots, q_i^{N_i} \} \), \( N_i := \dim U_i \), basis of \( U_i \subset V_N \)
- Basis transformation: \( \exists \mathbf{R}_i = (r_{jl}^i) \in \mathbb{R}^{N_i \times N} \) regular: \( q_i^j = \sum_{l=1}^{N_i} r_{jl}^i b_N^l \), \( j = 1, \ldots, N_i \).

**Linear system of equations for \( \vec{\gamma}_i \in \mathbb{R}^N \) = coefficient vector for \( c_i \) w.r.t. \( \mathcal{B}_N \):**
\[
\vec{\gamma}_i = \mathbf{R}_i^T \mathbf{A}_i^{-1} \mathbf{R}_i \vec{\rho}, \quad \vec{\rho} := \vec{\varphi} - \mathbf{A} \vec{\mu}_N^{(k)}, \quad \mathbf{A}_i = \left( a(q_i^j, q_i^l) \right)_{j, l=1}^{N_i} = \mathbf{R}_i \mathbf{A} \mathbf{R}_i^T \in \mathbb{R}^{N_i \times N_i}.
\] (4.1.4)

If \( N_i \ll N \) \( \Rightarrow (4.1.4) \) easier to solve than \( \mathbf{A} \vec{\mu} = \vec{\varphi}! \)

**Algorithm:** (Algebraic version of (4.1.2))

Iterative method for (2.1.1): successive subspace correction in the direction of \( U_i \)
(\( \tau = \text{tolerance, termination threshold, } \| \cdot \| \triangleq \text{some norm on } \mathbb{R}^N \))
Initial guess $\tilde{\mu}^{(0)} \in \mathbb{R}^N$; \(k=0\)

\begin{verbatim}
    do 
    k=k+1
    $\tilde{\mu}^{(k)} := \tilde{\mu}^{(k-1)}$
    for (i = 1; i ≤ J; i++) {
        $\bar{\rho} := \bar{\varphi} - A_i \tilde{\mu}^{(k)}_N$
        Solve (for $\delta \in \mathbb{R}^{Ni}$): $A_i \delta_i = R_i \bar{\rho}$
        $\bar{\mu}^{(k)} := \bar{\mu}^{(k)} + R_i^T \delta$
    }
    while (\|$\bar{\mu}^{(k)} - \bar{\mu}^{(k-1)}$\| > $\tau$) ← termination criterion
\end{verbatim}

(4.1.5)

Special (successive) subspace correction method: $U_i = \text{Span} \left\{ b_i^N \right\}, i = 1, \ldots, N$
Algorithm (without termination criterion), $A = (a_{ij})_{i,j=1}^N$, $\vec{\varphi} = (\varphi_i)_{i=1}^N$

Initial guess $\vec{\mu} \in \mathbb{R}^N$

\begin{align*}
&\text{do} \\
&\quad \text{for } (i = 1; i \leq N; i++) \\
&\quad \quad \mu_i := \mu_i + \frac{1}{a_{ii}} (\varphi_i - \sum_{k=1}^N a_{ik} \mu_k) \\
&\text{while } (\ldots) \\
\end{align*}

(4.1.6) $\leftrightarrow$ Gauss-Seidel iteration for $A\vec{\mu} = \vec{\varphi}$ (Makes sense for any matrix with regular diagonal)

$\text{Inner loop = Gauss-Seidel sweep: computational effort } \approx \text{nnz}(A)$

$\text{Matrix formulation of inner loop of (4.1.6)}$

\[ \vec{\mu} := \vec{\mu} + (L + D)^{-1}(\vec{\varphi} - A\vec{\mu}) \quad \text{where} \quad A = L + D + U. \quad \text{(4.1.7)} \]

$L = (l_{ij})_{i,j=1}^N \in \mathbb{R}^{N,N}$ $\triangleq$ strictly lower triangular part of $A$: $l_{ij} := \begin{cases} a_{ij} , & \text{for } 1 \leq j < i \leq N , \\ 0 , & \text{else.} \end{cases}$

$U \triangleq$ strictly upper triangular part of $A$ ($U = L^T$, if $A = A^T$),

$D = (d_{ij})_{i,j=1}^N \in \mathbb{R}^{N,N}$ $\triangleq$ diagonal of $A$: $d_{ij} := \begin{cases} a_{ij} , & \text{for } i = j , \\ 0 , & \text{for } i \neq j. \end{cases}$
Definition 4.1.1 (Linear stationary iterative method). Given $A, B \in \mathbb{R}^{N,N}$ regular.

$$\mu^{(k+1)} = \mu^{(k)} + B(\varphi - A\mu^{(k)})$$  \hspace{1cm} (4.1.8)

defines a linear stationary iterative method for $A\mu = \varphi$. $B$ is called preconditioner.

Note: $A \in \mathbb{R}^{N,N}$ regular $\Rightarrow \mu = A^{-1}\varphi \Leftrightarrow \mu$ fixed point of (4.1.8) (consistent iteration $\Rightarrow$ Def. 3.4.4)

For Gauss-Seidel method (4.1.6): $B = (L + D)^{-1}$

Other choices:

$$B = D^{-1} \leftarrow \text{Jacobi method}$$

$$B = (U + D)^{-1} \leftarrow \text{Backward Gauss-Seidel}$$

\{ Computational effort for single sweep $\approx \text{nnz}(A)$ \}
Remark. Generalisation: iteration (4.1.5) with $\mathcal{B}_i \subset \mathcal{B}_N$, $N_i \geq 1$ \hfill \blacktriangleright \text{ block-Gauss-Seidel}

\[ A = \begin{pmatrix} L & 0 \\ 0 & D \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & R \end{pmatrix} \]

Definition 4.1.2 (Iteration matrix). $M := I - BA = \text{iteration matrix (error propagation operator)}$ of the linear stationary iterative method (4.1.8) for the solution of $A\tilde{\mu} = \varphi$, $A \in \mathbb{R}^{N,N}$.

Iteration error for (4.1.8):
\[ \tilde{\delta}^{(k)} = \tilde{\mu}^* - \tilde{\mu}^{(k)} , \quad \tilde{\mu}^* := A^{-1}\varphi \]
\[ \Rightarrow \quad \tilde{\delta}^{(k+1)} = \tilde{\mu}^* - \tilde{\mu}^{(k+1)} = \tilde{\mu}^* - \tilde{\mu}^{(k)} - B(A\tilde{\mu}^* - A\tilde{\mu}^{(k)}) \]
\[ = (I - BA)(\tilde{\mu}^* - \tilde{\mu}^{(k)}) = M\tilde{\delta}^{(k)} , \]
\[ \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quito
Note: If \( \vec{\varphi} = 0 \), then algorithm for (4.1.8) provides definition of iteration matrix.

**Example.** Successive subspace correction, operator perspective

In (4.1.2): if \( f \equiv 0 \), then \( c_i = -P_i u_N^{(k)} \), \( P_i = \text{Galerkin projection} \) (\( \rightarrow \) Def. 2.1.4) onto \( U_i \)

\[ \text{for (4.1.2)}: \quad \mathbf{M} \approx (\text{Id} - P_J)(\text{Id} - P_{J-1}) \cdots (\text{Id} - P_1). \quad (4.1.9) \]

---

**Theorem 4.1.3** (Convergence of stationary linear iterative methods). Let \( \mathbf{A} \in \mathbb{R}^{N \times N} \) be regular and \( \| \cdot \| \) some norm on \( \mathbb{R}^N \). If \( \rho := \| \mathbf{M} \|_x < 1 \), \( \mathbf{M} \in \mathbb{R}^{N \times N} \) = iteration matrix of (4.1.8) for \( \mathbf{A} \vec{\mu} = \vec{\varphi} \) (\( \| \cdot \|_x = \text{matrix norm associated with } \| \cdot \| \)), then

\[ \| \vec{\delta}(K) \| \leq \rho^K \| \vec{\delta}(0) \| \quad \forall \vec{\varphi} \in \mathbb{R}^N \quad (4.1.10) \]

Terminology: \( \rho = \| \mathbf{M} \|_x \) = contraction number of iteration (rate of linear convergence)

**Definition 4.1.4.** A stationary linear iteration (4.1.8) with iteration matrix \( \mathbf{M} \in \mathbb{R}^{N \times N} \) enjoys the rate of convergence

\[ \rho := \lambda_{\max}(\mathbf{M}) := \max\{|\lambda|: \lambda \text{ is eigenvalue of } \mathbf{M}\}. \]
rate of convergence $\rho \leftrightarrow$ measures *asymptotic convergence* of iteration (4.1.8):

**Lemma 4.1.5 (Asymptotic convergence of linear stationary iterative method).** *For any norm $\| \cdot \|$ on $\mathbb{R}^N$, and any $\tilde{\mu}^{(0)}$ the iterates $\tilde{\mu}^{(k)}$ of (4.1.8) satisfy (with convention “$0/0 = 0$”)*

$$
\lim_{k \to \infty} \frac{\| \tilde{\mu}^{(k+1)} - A^{-1}\tilde{\phi} \|}{\| \tilde{\mu}^{(k)} - A^{-1}\tilde{\phi} \|} \leq \lambda_{\text{max}}(M).
$$

*Remark.* A “good” preconditioner $B$

- is “close to” $A^{-1}$,
- $B \times$ vector is easy to evaluate.

**Example:** $A \in \mathbb{R}^{N,N}$ symmetric, positive definite

$B = \rho(A)^{-1}I$, $\rho(A) = \text{spectral radius}$: $\rho(A) := \max\{ |\lambda| : \lambda \text{ eigenvalue of } A \}$

Richardson iteration for $A\tilde{\mu} = \tilde{\phi}$:

$$
\tilde{\mu}^{(k+1)} = \tilde{\mu}^{(k)} - \rho(A)^{-1}(\tilde{\phi} - A\tilde{\mu}^{(k)})
$$

(4.1.11)

*Iteration matrix:* $M := I - \rho(A)^{-1}A$

$$
\rho = \lambda_{\text{max}}(I - \rho(A)^{-1}A) = 1 - \lambda_{\text{max}}(A)^{-1}\lambda_{\text{min}}(A)
$$

($\lambda_{\text{max}}(A), \lambda_{\text{min}}(A) =$ largest/smallest eigenvalue (in modulus) of $A$)
For Richardson iteration (4.1.11), $A = A^T$ positive definite: 
$$
\rho = 1 - \frac{\lambda_{\text{min}}(A)}{\lambda_{\text{max}}(A)} = 1 - \frac{1}{\kappa(A)} < 1
$$

### Definition 4.1.6
For $A \in \mathbb{R}^{N,N}$

$$
\kappa(A) := \frac{\max\{|\lambda| : \lambda \text{ eigenvalue of } A\}}{\min\{|\lambda| : \lambda \text{ eigenvalue of } A\}} = \frac{\lambda_{\text{max}}(A)}{\lambda_{\text{min}}(A)}
$$

is the spectral condition number of the matrix $A$ ($= \infty$ for singular $A$).

### Theorem 4.1.7
(A posteriori iteration error estimate). Given a stationary linear iterative method for $A\tilde{\mu} = \tilde{\varphi}$, $A \in \mathbb{R}^{n,n}$ regular, with iteration matrix $M \in \mathbb{R}^{n,n}$ that satisfies (4.1.8)

$$
\rho := \|M\|_x < 1 \text{ there holds}
$$

$$
\|\tilde{\mu}^{(k)} - A^{-1}\tilde{\varphi}\| \leq \frac{1}{1 - \rho} \|\tilde{\mu}^{(k)} - \tilde{\mu}^{(k+1)}\|.
$$

(4.1.12)

**Proof.** Set $\tilde{\mu}^* := A^{-1}\tilde{\varphi}$. By Theorem 4.1.3

$$
\|\tilde{\mu}^{(k)} - \tilde{\mu}^*\| \leq \|\tilde{\mu}^{(k)} - \tilde{\mu}^{(k+1)}\| + \|\tilde{\mu}^{(k+1)} - \tilde{\mu}^*\| \leq \|\tilde{\mu}^{(k)} - \tilde{\mu}^{(k+1)}\| + \rho \|\tilde{\mu}^{(k)} - \tilde{\mu}^*\|
$$

4.1

p. 310
If $\rho \ll 1$, $\|\tilde{\mu}^{(k)} - \tilde{\mu}^{(k+1)}\| < \tau$ provides viable termination criterion for a linear iteration

**Experiment 4.1.1.**

- $-\Delta u = f$ on $\Omega := B_1(0) := \{x \in \mathbb{R}^2: |x| < 1\}$, $u = 0$ on $\partial \Omega$,
- sequence of triangular unstructured meshes $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4$ (→ Example 2.5.8),
- Galerkin FE discretization based on $V_N := \mathcal{S}_0^{1,0}(\mathcal{M})$ or $V_N := \mathcal{S}_0^{2,0}(\mathcal{M})$.
- Gauss-Seidel iteration (4.1.11), r.h.s $\bar{\varphi} = 0$, $\tilde{\mu}^{(0)} = 1$

Recorded: approximate Gauss-Seidel convergence rate (→ Def. 4.1.4) $\rho := \frac{|\tilde{\mu}^{(19)}|}{|\tilde{\mu}^{(20)}|}$

<table>
<thead>
<tr>
<th>FE space</th>
<th>$V_N := \mathcal{S}_0^{1,0}(\mathcal{M})$</th>
<th>$V_N := \mathcal{S}_0^{2,0}(\mathcal{M})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>mesh</td>
<td>$\mathcal{M}_1$</td>
<td>$\mathcal{M}_2$</td>
</tr>
<tr>
<td>$\rho$</td>
<td>0.9276</td>
<td>0.9805</td>
</tr>
</tbody>
</table>
approximate Gauss-Seidel rate of convergence deteriorates on fine meshes: $\rho \approx 1 - O(N^{-1})$
4.2 Multigrid methods

Experiment 4.2.1.

\[-\Delta u = f \quad \text{in } \Omega := ]0, 1[^2, \quad u = 0 \quad \text{on } \partial \Omega.\]

FE-Galerkin discretization:

- triangular tensor product mesh
- \( V_N = \mathcal{S}_1^{0,0}(\mathcal{M}) \)

Gauss-Seidel iterations (4.1.7),
initial guess \( \tilde{\mu}^{(0)} = \text{random} \), r.h.s. \( \tilde{\varphi} = 0 \)

\( > \) iterates = error
Recorded: “shape” of iterates

Observation: slow convergence of smooth error components

Bilinear interpolants with nodal values $\tilde{\mu}^{(k)}$, $\tilde{\mu}^{(k)} = \text{Gauss-Seidel iterates}$

Culprit: only local (“non-smooth”) corrections, slow flow of global information
[for symmetric positive definite matrix $A$]

also use non-local directions in successive subspace correction!

Question: How can this be achieved economically?

Idea:

Non-local subspace provided by FE space on coarse mesh.

← mesh $\mathcal{M}$ (black)
← coarse mesh $\mathcal{M}_0$ (magenta)

$$U_{N+1} := \mathcal{S}_{1,0}(\mathcal{M}_0) \subset V_N = \mathcal{S}_{1,0}(\mathcal{M})$$

Iteration (4.1.2) based on (→ (4.1.1))

$$\mathcal{S}_{1,0}(\mathcal{M}) = \sum_{i=1}^{N} \text{Span} \{ b_i^i \} + \mathcal{S}_{1,0}(\mathcal{M}_0) \quad (4.2.1)$$

$b_i^i = \text{hat function}$

Experiment 4.2.2. (Two-grid iteration)
● setting of Exp. 4.2.1, $\bar{\mu}^{(0)} = 1$

● successive subspace correction (4.1.5) based on (4.2.1)

● monitor “shape” of iterates (= iteration errors), plotted: bilinear interpolants

Observation: Rapid decay of all error components
Goal: Retain speed of two-grid method at reduced cost ↔ using smaller subspaces

Example 4.2.3 (Hierarchical basis for $S_{1,0}^0(M)$ in 1D).

(⇒ Lecture “Numerische Mathematik für CSE”, Sect. 4.5 “Multiskalenbasen”)

- 1D, $\Omega = [0, 1]$, $-u'' = f$, $u(0) = u(1) = 0$
- $M = \{\xi_{j-1}, \xi_j, j = 1, \ldots, N\}, \xi_j = hj, h = N^{-1}, N = 2^L, L \in \mathbb{N}$
- FE-Galerkin discretization, $V_N = S_{1,0}^0(M)$
Fig. 59
Nodal basis
(Level $L = 3$)

Fig. 60
Hierarchical basis
($- \triangleleft$ level 0, $- \triangleleft$ level 1, $- \triangleleft$ level 2)

Resulting stiffness matrices w.r.t hierarchical basis
\[ \mathbf{A}_{\text{nod}} = h^{-1} \begin{pmatrix} 2 & -1 \\ -1 & 2 & -1 \\ & \ddots & \ddots & \ddots \\ & -1 & 2 & -1 \\ & & 2 & -1 \end{pmatrix} \]

\[ \mathbf{A}_h = \text{diag}(2/h, \ldots, 2/h, 1/h, \ldots, 1/h, \ldots, 1/4) \]

\[ \begin{array}{c}
2^{L-1} \\
2^{L-2} 
\end{array} \text{ times } \]

Stiffness matrix \( \mathbf{A}_h \) diagonal for hierarchical basis!

Gauss-Seidel (4.1.7) applied to \( \mathbf{A}_h \vec{\mu}_h = \vec{\varphi}_h \) will "converge" in one step.

\((\leftrightarrow \text{ (successive) subspace corrections in directions of hierarchical basis functions provide direct solver.)}\)

Implementation:

1. Transformation: nodal basis \( \rightarrow \) hierarchical basis
2. Scaling (by diagonal entries of \( \mathbf{A}_h \))
3. Transformation: hierarchical basis \( \rightarrow \) nodal basis
Input: $\tilde{v} = (v_0, \ldots, v_N) \in \mathbb{R}^N$: Nodal basis coefficients, $v_0 = v_N = 0$
Output: $\tilde{\eta} = (\eta_0, \ldots, \eta_N) \in \mathbb{R}^N$: Hierarchical basis coefficients

\[ \tilde{\eta} = \tilde{v}; \]
\[ \text{for } l = L \text{ to } 1 \text{ step } -1 \text{ do } \{ \]
\[ \quad \text{for } j = 2^{L-l} \text{ to } 2^L - 2^{L-l} \text{ step } 2^{L-l+1} \text{ do } \{ \]
\[ \quad \quad \eta_j = \eta_j - \frac{1}{2}(\eta_{j-2^{L-l}} + \eta_{j+2^{L-l}}); \]
\[ \quad \} \]
\[ \} \]

Computational effort for basis transformation $\approx N!$

Generalization to linear Lagrangian finite elements in higher dimensions:
Linear Lagrangian FE: \( V_N = \mathcal{S}_{1,0}^0(\mathcal{M}) \)

- \( \hat{\phi} \) basis function level 0: \( b_0^1 \)
- \( \hat{\phi} \) level 1 basis functions: \( b_1^j \)
- \( \hat{\phi} \) level 2 basis functions: \( b_2^j \)

However: \( A_h \) no longer diagonal

Nevertheless, Gauss-Seidel (4.1.7) applied to \( A_h \) may converge fast!

**Experiment 4.2.4.**

- \( -\Delta u = f \) on \( \Omega = ]0, 1[^2 \), \( u = 0 \) on \( \partial \Omega \),
- 32 \( \times \) 32 triangular tensor product mesh, Galerkin FE discretization based on \( V_N = \mathcal{S}_{1,0}^0(\mathcal{M}) \),
- Hierarchical basis Gauss-Seidel (ordering: fine \( \rightarrow \) coarse), \( \tilde{\mu}^{(0)} = 1 \), \( \tilde{\varphi} = 0 \).
Recorded: “shape” & “size” of functions corresponding to iterates \( \tilde{\mu}^{(k)} = \) iteration errors

Observation: fast convergence of smooth error components

Bilinear interpolants with nodal values \( \tilde{\mu}^{(k)}, \tilde{\nu}^{(k)} = \) iterates of hierarchical multigrid

Do the subspaces \( U_i \) in (4.1.1) have to form a direct sum?
Multigrid idea (Galerkin FE, s.p.d. case):
Successive subspace correction using one-dimensional subspaces spanned by global basis functions on a hierarchy of nested triangulations

Experiment 4.2.5. Repetition of Exp. 4.2.4 with true multigrid Gauss-Seidel iteration:

Bilinear interpolants with nodal values $\bar{\mu}(k), \bar{\mu}(k) = \text{iterates of multigrid V}(1,0)$-cycle
\[ -\Delta u = f \text{ on } \Omega = ]0, 1[, u = 0 \text{ on } \partial \Omega, \]

- \( N \times N \) triangular tensor product mesh \( \mathcal{M} \), Galerkin FE discretization based on \( V_N = \mathcal{S}_{1,0}(\mathcal{M}) \),
- Multigrid Gauss-Seidel (ordering: fine \( \rightarrow \) coarse), \( \mu^{(0)} = 1, \bar{\varphi} = 0 \).

**Recorded:** approximate rate of linear convergence

\[
\rho = \frac{|\mu^{(15)}|}{|\mu^{(14)}|}
\]

<table>
<thead>
<tr>
<th>( N )</th>
<th>8</th>
<th>16</th>
<th>32</th>
<th>64</th>
<th>128</th>
<th>256</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \rho )</td>
<td>0.4559</td>
<td>0.4174</td>
<td>0.4401</td>
<td>0.4463</td>
<td>0.4483</td>
<td>0.4490</td>
</tr>
</tbody>
</table>

**Observation:**

Contraction number of multigrid iteration \( \approx \) independent of \( N \)

**Terminology:** Iterative solution method for FEM equations “optimal”

\[ \uparrow \]

Speed of convergence independent of discretization parameters

**Efficient implementation of multigrid**
Setting: discrete variational problem (→ Sect. 2.1.1) with $V$-elliptic $a(\cdot, \cdot)$:

\[ V_N \subset V, \quad \dim V_N =: N < \infty, \quad v_N \in V_N: \quad a(v_N, v_N) = f(v_N) \quad \forall v_N \in V_N \tag{2.1.1} \]

Linear system of equations: \[ A\tilde{\mu} = \tilde{\varphi} \]

\[ M_0 : \text{coarse mesh} \quad \Rightarrow \quad \text{FE space } V_H := \delta_1^0(M_0) : \text{basis } \mathcal{B}_H = \{b_H^1, \ldots, b_H^M\}, M \in \mathbb{N}, \]

\[ M : \text{fine mesh} \quad \Rightarrow \quad \text{FE space } V_N := \delta_1^0(M) : \text{basis } \mathcal{B}_h = \{b_h^1, \ldots, b_h^N\}, N \in \mathbb{N}. \]

\[ \text{Definition 4.2.1. Two meshes } M_0, M \text{ of a computational domain } \Omega \text{ are nested, write } M_0 \prec M, \text{ if} \]

\[ \forall K \in M_0: \quad \exists n \in \mathbb{N},, K_1, \ldots, K_n \in M: \quad \overline{K} = \overline{K}_1 \cup \ldots \cup \overline{K}_n. \]

Remark. Local mesh refinement (→ Sect. 7.3.1) creates sequences of nested meshes!

Note:

\[ \text{nesting } M_0 \prec M \quad \Rightarrow \quad \text{nesting } \quad V_H \subset V_h \]

\[ \Rightarrow \quad b_H^j = \sum_{k=1}^N r_{jk} b_h^k, \quad R := (r_{jk}) \in \mathbb{R}^{M,N} \text{ regular}. \]

\[ b_h^k, b_H^j \text{ locally supported } \Rightarrow \quad R \text{ sparse}. \]
Example: 1D, $\mathcal{B}_{1,0}(\mathcal{M})$ on equidistant mesh
magenta: basis functions on fine mesh
green: basis functions on coarse mesh

\[
R = \begin{pmatrix}
\frac{1}{2} & 1 & \frac{1}{2} & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & \frac{1}{2} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \frac{1}{2} & 1 & \frac{1}{2}
\end{pmatrix}
\]

Correction (obtained by 1D minimization of Dirichlet functional) in direction of $b^j_H$:

\[
u_N \leftarrow u_N + \frac{f(b^j_H) - a(u_N, b^j_H)}{a(b^j_H, b^j_H)} b^j_H = u_N + \frac{\sum_{k=1}^{N} r_{jk}(f(b^k_H) - a(u_N, b^k_H))}{a(b^j_H, b^j_H)} b^j_H.
\]

Perspective: iteration in function space (i.e. in finite element space $V_N$)
Correction expressed in terms of basis of $V_N$ (on fine mesh):

$$
\tilde{\mu} \leftarrow \tilde{\mu} + R^T \tilde{e}_j \frac{\tilde{e}_j^T R \tilde{\rho}}{(A_H)_{jj}} \text{ residual } \tilde{\rho} := \tilde{\varphi} - A \tilde{\mu} \in \mathbb{R}^N ,
$$

$\tilde{e}_j \in \mathbb{R}^M = j$-th unit vector, $A_H =$ stiffness matrix w.r.t. basis $\mathcal{B}_H$ of $V_H$.

Compiled corrections in the direction of basis functions of $V_H$ (w.r.t. $\mathcal{B}_H$)

$$
\text{Gauss-Seidel sweep on } A_H \tilde{\mu}_H = R \tilde{\rho}, \text{ initial guess } \tilde{\mu}_H^{(0)} = 0
$$

Algorithm: two-grid multigrid sweep (correction scheme), $V(1,0)$-cycle  (→ Exp. 4.2.2)
Input : $\tilde{\mu} \in \mathbb{R}^N$

Output : updated coefficient vector $\tilde{\mu} \in \mathbb{R}^N$

for $j = 1$ to $N$ do { 
    $\mu_j := \mu_j + (A)^{-1}_{jj}(\varphi_j - (A\tilde{\mu})_j)$
}

$\tilde{\rho}_H := R(\varphi - A\tilde{\mu})$

$\tilde{\gamma}_H = 0$;

for $j = 1$ to $M$ do { 
    $\gamma_j := \gamma_j + (A_H)^{-1}_{jj}(\tilde{\rho}_H - A_H\tilde{\gamma}_H)_j$
}

$\tilde{\mu} := \tilde{\mu} + R^T\tilde{\gamma}_H$;

Gauss Seidel sweep on fine mesh $\mathcal{M}$

restriction of residual

Gauss Seidel sweep on coarse mesh $\mathcal{M}_H$

prolongation of correction of residual

Extensions:

- $m_1$ Gauss-Seidel sweeps on $\mathcal{M}$,
- $m_2$ Gauss-Seidel sweeps after returning from coarse mesh.
Algorithm:

full multigrid sweep (correction scheme) for solving $A\tilde{\mu} = \tilde{\varphi}$ (arising from discrete variational problem (2.1.1)) on $L \in \mathbb{N}$ levels

(FE spaces $V_0 \subset V_1 \subset \cdots \subset V_L = V_N$, $N_l := \dim V_l$, $0 \leq l \leq L$

(sparse) basis transformation matrices $R_l \in \mathbb{R}^{N_{l-1}, N_l}$)

Notation: $A_l =$ stiffness matrix “on level $l$”, $0 \leq l \leq L$ w.r.t. FE space $V_l$

Input : $\tilde{\mu} \in \mathbb{R}^{N_L}$

Output : updated coefficient vector $\tilde{\mu} \in \mathbb{R}^{N_L}$
\[ \tilde{\varphi}_L = \tilde{\varphi}; \]

for \( l = L \) to 1 step \(-1\) do {
\[ m_1 \text{ Gauss-Seidel sweeps on } A_l \tilde{\mu}_l = \tilde{\varphi}_l; \] 
pre-smoothing
\[ \tilde{\varphi}_{l-1} := R_{l-1}(\tilde{\varphi}_l - A_l \tilde{\mu}_l); \] 
restriction
\[ \tilde{\mu}_{l-1} = 0; \]
}

Solve (small) linear system \( A_0 \tilde{\mu}_0 = \tilde{\varphi}_0 \).

for \( l = 1 \) to \( L \) do {
\[ \tilde{\mu}_l := \tilde{\mu}_l + R_{l-1}^T \tilde{\mu}_{l-1}; \] 
prolongation
\[ m_2 \text{ Gauss-Seidel sweeps on } A_l \tilde{\mu}_l = \tilde{\varphi}_l; \] 
post-smoothing
}


Computational effort:

- on each level $l$: $\text{work} \approx (m_1 + m_2 + 2)N_l + 2N_{l-1}$, $1 \leq l \leq L$,

- if geometric increase of space dimensions

  \[ \exists 0 < q < 1: \quad N_{l-1} \approx qN_l, \quad 1 \leq l \leq L, \]

  then total work

  \[ \text{work}_{\text{tot}} \leq \frac{m_1 + m_2 + 4}{1 - q} N_L. \]

  “Optimal effort” for a single sweep, proportional to dim $V_L$

Remark. drawback of standard multigrid: needs sequence of nested FE spaces

automated construction of “coarse spaces” in algebraic multigrid (AMG)

4.3 Preconditioned conjugate gradients (PCG)

Lecture “Numerische Mathematik für CSE”, Sect. 2.5 “Krylov-Verfahren für lineare Gleichungssysteme”
Meta Galerkin discretization in Krylov subspace of FE space \( V_N \)

1 matrix×vector product \( A\hat{\xi} \) per iteration

For \( A \) sparse: computational effort \( O(N) \), \( N := \text{dim } V_N \), per iteration

**Theorem 4.3.1** (Convergence of CG). \( A \in \mathbb{R}^{N,N} \) symmetric & positive definite, \( \tilde{\varphi} \in \mathbb{R}^{N,N} \), \( \tilde{\mu}^* := A^{-1}\tilde{\varphi} \), \( \tilde{\mu}^{(L)} \) iterate after \( L \) CG steps:

\[
\| \tilde{\mu}^{(L)} - \tilde{\mu}^* \|_A \leq \frac{2q^L}{1 + q^2L} \| \tilde{\mu}^{(0)} - \tilde{\mu}^* \|_A ,
\]

\[
q := \frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1} < 1 .
\]

Notation: \( \text{energy norm on } \mathbb{R}^N : \)

\[
\| \tilde{\xi} \|_A^2 := \tilde{\xi}^T A\tilde{\xi} \quad (\rightarrow \text{Sect. 1.7})
\]

**Example 4.3.1** (Convergence of CG for discrete 2nd order elliptic BVP).
\[-\Delta u = f \text{ on } \Omega := B_1(0) := \{ \mathbf{x} \in \mathbb{R}^2: |\mathbf{x}| < 1 \}, \ u = 0 \text{ on } \partial\Omega,\]

- sequence of triangular unstructured meshes \( \mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4 \) (\( \rightarrow \) Ex. 2.5.8),

- Galerkin FE discretization based on \( V_N := \delta^0_{1,0}(\mathcal{M}) \text{ or } V_N := \delta^0_{2,0}(\mathcal{M}). \)

- 40 conjugate gradient iterations (4.1.11), r.h.s \( \varphi = 1, \bar{\mu}^{(0)} = 0, \) using MATLAB `pcg` command.

Recorded: approximate conjugate gradient “convergence rate” \( \rho := \left( \frac{|\bar{\rho}^{(40)}|}{|\bar{\rho}^{(0)}|} \right)^{1/40} \).

<table>
<thead>
<tr>
<th>FE space</th>
<th>( V_N := \delta^0_{1,0}(\mathcal{M}) )</th>
<th>( V_N := \delta^0_{2,0}(\mathcal{M}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>mesh ( \mathcal{M}_1 ) ( \mathcal{M}_2 ) ( \mathcal{M}_3 ) ( \mathcal{M}_4 )</td>
<td>( \mathcal{M}_1 ) ( \mathcal{M}_2 ) ( \mathcal{M}_3 ) ( \mathcal{M}_4 )</td>
<td></td>
</tr>
<tr>
<td>( \rho )</td>
<td>0.3416</td>
<td>0.6049</td>
</tr>
</tbody>
</table>
Decay of relative Euclidean residual norm $|	ilde{\rho}^{(l)}|/|\tilde{\varphi}$ during CG iterations ($\mathcal{S}_1(M)$)
Thm. 4.3.1  ➤  CG slows down as spectral condition number (→ Def. 4.1.6) $\kappa(A) \uparrow$

Is this a reason to worry, when $A$ = finite element stiffness matrix?

*Example 4.3.2 (Spectral condition numbers of stiffness matrices).*

Spectral conditions numbers of stiffness matrices from Ex. 2.5.8:
(as computed approximately with MATLAB `eigs` functions)

<table>
<thead>
<tr>
<th>FE space</th>
<th>$V_N := \delta^{0}_{1,0}(\mathcal{M})$</th>
<th>$V_N := \delta^{0}_{2,0}(\mathcal{M})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>mesh</td>
<td>$\mathcal{M}_1$</td>
<td>$\mathcal{M}_2$</td>
</tr>
<tr>
<td>$\lambda_{\text{min}}$</td>
<td>0.1329</td>
<td>0.0335</td>
</tr>
<tr>
<td>$\lambda_{\text{max}}$</td>
<td>6.0620</td>
<td>5.5640</td>
</tr>
<tr>
<td>$\kappa(A)$</td>
<td>45.6</td>
<td>166</td>
</tr>
</tbody>
</table>
Spectral conditions numbers of stiffness matrices as function of $N := \dim V_N$
The conditioning of finite element stiffness matrices

Setting:

\[ V = H^1_0(\Omega), \ V_N = \delta^0_{1,0}(\mathcal{M}) \] on triangular mesh \( \mathcal{M} \) of computational domain \( \Omega \subset \mathbb{R}^2 \), bilinear form \( a(u, v) = \int_{\Omega} \nabla u \cdot \nabla v \ dx \)

\[ \Rightarrow \text{stiffness matrix } A \in \mathbb{R}^{N \times N}, \text{ see Sect. 2.1.2.} \]

Lemma 4.3.2. Let \( \mathcal{B} = \{b^1_N, \ldots, b^N_N\} \) be the nodal basis of \( \delta^0_{1,0}(\mathcal{M}) \) (\( \rightarrow b^j_N \equiv \text{"hat functions"} \)). Then

\[
\sum_{j=1}^{N} \eta_j^2 \|b^j_N\|_0^2 \leq \|v_N\|_0^2 \leq 4 \sum_{j=1}^{N} \eta_j^2 \|b^j_N\|_0^2 \quad \forall v_N = \sum_{j=1}^{N} \eta_j b^j_N \in V_N .
\]

Proof. Pick triangle \( K := \text{convex}\{a^1, a^2, a^3\} \in \mathcal{M} \), \( v_N \in \delta^0_1(\mathcal{M}) \).

\[
\xi_j := v_N(a_j) \Rightarrow \|v_N\|_{0,K}^2 = 1/12|K| \left( (\xi_1 + \xi_2)^2 + (\xi_1 + \xi_3)^2 + (\xi_2 + \xi_3)^2 \right) . \quad (4.3.1)
\]

\[
b^K_j \equiv \text{local shape function belonging to vertex } a^j, j = 1, 2, 3: \quad \|b^K_j\|_{0,K}^2 = \frac{1}{12}|K| . \quad (4.3.2)
\]

\[(4.3.1) \ & (4.3.2) \Rightarrow \sum_{j=1}^{3} \xi_j^2 \|b^K_j\|_{0,K}^2 \leq \|v_N\|_{0,K}^2 \leq 4 \sum_{j=1}^{3} \xi_j^2 \|b^K_j\|_{0,K}^2 .\]
Lemma 4.3.3 (Numerical range of s.p.d. matrix).

\[
\forall \mathbf{A} = \mathbf{A}^T \in \mathbb{R}^{N,N}: \quad \lambda_{\text{max}}(\mathbf{A}) = \max_{\xi \neq 0} \frac{\xi \cdot \mathbf{A} \xi}{\xi \cdot \xi}, \quad \lambda_{\text{min}}(\mathbf{A}) = \min_{\xi \neq 0} \frac{\xi \cdot \mathbf{A} \xi}{\xi \cdot \xi}.
\]

Application to stiffness matrix ($\vec{\eta} = $ coefficient vector for $v_N \in V_N$)

\[
\lambda_{\text{max}}(\mathbf{A}) = \max_{\vec{\eta} \neq 0} \frac{\vec{\eta} \cdot \mathbf{A} \vec{\eta}}{\vec{\eta} \cdot \vec{\eta}} = \max_{\vec{\eta} \neq 0} \frac{|v_N|^2}{|\vec{\eta}|^2} \leq C(\rho,\mathcal{M}) (\min\{h_K : K \in \mathcal{M}\})^{-2} \frac{\|v_N\|_0^2}{\vec{\eta} \cdot \vec{\eta}}
\]

\[
\leq C(\rho,\mathcal{M},\mu,\mathcal{M}) h^{-2} \|b\|_0^2 \sum_{j=1}^N \eta_j^2 \|b_j^K\|_0^2 \leq C(\rho,\mathcal{M},\mu,\mathcal{M}) .
\]

1: Use inverse estimates ($\rightarrow$ Def. 2.5.29, cf. Lemma 2.5.28)

\[
\exists C = C(\rho,\mathcal{M}): \quad |v_N|_1 \leq C (\min\{h_K : K \in \mathcal{M}\})^{-1} \|v_N\|_0 \quad \forall v_N \in S_1^0(\mathcal{M}), \tag{4.3.3}
\]

2: Quasi-uniformity ($\rightarrow$ Def. 2.5.13) & Lemma 4.3.2.

3: Norm of hat functions: $\|b_j^N\|_0^2 \leq C(\rho,\mathcal{M}) h_{\mathcal{M}}^2$ for all $j = 1, \ldots, N$. 
\[ \lambda_{\text{min}}(A) = \min_{\vec{\eta} \neq 0} \frac{\vec{\eta} \cdot A \vec{\eta}}{\vec{\eta} \cdot \vec{\eta}} = \min_{\vec{\eta} \neq 0} \frac{|v_N|^2_1}{|\vec{\eta}|^2} \geq \text{diam}(\Omega)^2 \frac{\|v_N\|_0^2}{|\vec{\eta}|^2} \]

\[ \geq \text{diam}(\Omega)^2 \frac{1}{|\vec{\eta}|^2} \sum_{j=1}^{N} \eta_j^2 \left\| b_{N}^j \right\|_0^2 \geq C(\Omega, \rho_M) \min\{h_K : K \in \mathcal{M}\}^2. \]

1: \( v_N \in H^1_0(\Omega) \) ➤ Poincaré-Friedrichs inequality Thm. 1.7.9.

2: Lower estimate of Lemma 4.3.2

3: Norm of hat functions: \( \left\| b_{N}^j \right\|_0^2 \geq C(\rho_M) \min\{h_K : K \in \mathcal{M}\}^2 \) for all \( j = 1, \ldots, N \).

Estimate for spectral condition number of stiffness matrix \( A \):

\[ \kappa(A) \leq C(\Omega, \rho_M, \mu_M) h_M^{-2} \]

Numerical experiment ➤ estimate (seems to be) sharp

\( A = \text{FE-stiffness matrix on fine mesh/high polynomial degree} \) ➞ conjugate gradient iteration very slow!
Idea: Preconditioning

Apply conjugate gradient iterations to transformed system

\[ \tilde{A} \tilde{\mu} = \tilde{\varphi} \quad , \quad \tilde{A} := B^{1/2} A B^{1/2} \quad , \quad \tilde{\mu} := B^{-1/2} \bar{\mu} \quad , \quad \tilde{\varphi} := B^{1/2} \bar{\varphi} \, , \]

where \( \kappa(\tilde{A}) = \) “small”, \( B = B^T \in \mathbb{R}^{N,N} \) positive definite.

Suitable \( B \) supplied by preconditioners of “fast” symmetric (ie. \( B = B^T \)) stationary linear iterations (\( \rightarrow \) Def. 4.1.1):

\[ \rho := \lambda_{\text{max}}(I - BA) = \lambda_{\text{max}}(I - B^{1/2}AB^{1/2}) \quad \Rightarrow \quad \begin{cases} 
\lambda_{\text{min}}(B^{1/2}AB^{1/2}) \geq 1 - \rho \\
\lambda_{\text{max}}(B^{1/2}AB^{1/2}) \leq 1 + \rho 
\end{cases} , \]

\[ \Rightarrow \quad \kappa(B^{1/2}AB^{1/2}) \leq \frac{1 + \rho}{1 - \rho} . \]

Implementation: CG iteration formulated in original variables \( \bar{\mu} \) (intrinsic transformation)

Algorithm: Preconditioned conjugate gradient iteration (PCG)
Input: initial guess $\tilde{\mu} 
cong \tilde{\mu}^{(0)} \in \mathbb{R}^N$, termination threshold $\tau > 0$

Output: improved solution vector $\tilde{\mu} 
cong \tilde{\mu}^{(L)}$

$$\tilde{\pi} := \tilde{\rho} := \tilde{\varphi} - A \tilde{\mu}; \quad \tilde{\pi} := B \tilde{\rho}; \quad \tilde{v} := \tilde{\pi};$$

for $l = 1$ to $L_{\text{max}}$ do {

$\beta := \tilde{\rho} \cdot \tilde{v}; \quad \tilde{\eta} := A \tilde{\pi}; \quad \alpha := \frac{\beta}{\tilde{\pi} \cdot \tilde{\eta}};$

$\tilde{\mu} := \tilde{\mu} + \alpha \tilde{\pi};$

$\tilde{\rho} := \tilde{\rho} - \alpha \tilde{\eta};$

$\tilde{v} := B \tilde{\rho}; \quad \beta := \frac{\tilde{\rho} \cdot \tilde{v}}{\beta};$

if $|\tilde{v} \cdot \tilde{\rho}| \leq \tau$ then stop;

$\tilde{\pi} := \tilde{v} + \beta \tilde{\pi};$

} 

Remark. Many variants or Krylov subspace based iterative solution exist [3]:
(also for non-positive definite, non-symmetric linear systems)
MATLAB functions: bicg, bicgstab, cgs, gmres, minres, qmr, symmlq
5.1 Transient heat conduction

\[ \Omega \subset \mathbb{R}^d \]: space occupied by solid body (spatial computational domain),

\( \sigma \): heat conductivity (in general \( \sigma = \sigma(x), \ x \in \Omega \)), see Sect. 1.2,

\( u_0 \): initial temperature distribution in \( \Omega \), \( u_0 \in L^2(\Omega) \)

+ boundary conditions from Sect. 1.3 (and combinations).

Goal: Approximate calculation of evolution of temperature \( u = u(t, x) \)

with initial condition \( u(0, \cdot) = u_0 \).

\( t = \) temporal independent variable, \( 0 < t < T \), \( T = \) end time,

\( x = \) spatial independent variable, \( x \in \Omega \).
Computational domain:

\[ \tilde{\Omega} := ]0, T[ \times \Omega \subset \mathbb{R}^{d+1} \] .

space-time cylinder

On \{0\} \times \Omega \rightarrow \text{initial conditions},
on \]0, T[ \times \partial \Omega \rightarrow \text{boundary conditions.}
Conservation of energy:

\[
\frac{d}{dt} \int_V \rho u \, dx + \int_{\partial V} \mathbf{j} \cdot \mathbf{n} \, dS = \int_V f \, dx \quad \text{for all "control volumes" } V
\] (5.1.1)

\[f = f(t, x) : \text{time-dependent heat source/sink, } t \mapsto f(t, \cdot) \in L^2(\Omega) \text{ continuous,}\]

\[\rho = \rho(x) : \text{heat capacity } (\lbrack \rho \rbrack = JK^{-1}), \text{uniformly s.p.d., cf. (1.2.2).}\]
Local form of energy balance law

\[ \frac{d}{dt}(\rho u) + \text{div}\, \mathbf{j} = f \quad \text{in} \, \widetilde{\Omega} . \]  
\[ \text{(5.1.2)} \]

[using Fourier's law (1.2.1)]

\[ \frac{d}{dt}(\rho u) - \text{div}(\sigma \, \text{grad} \, u) = f \quad \text{in} \, \widetilde{\Omega} . \]  
\[ \text{(5.1.3)} \]

\[ + \text{ boundary conditions on } ]0, T[ \times \partial \Omega, \text{ e.g.,} \]

Dirichlet boundary conditions:  \[ u(t, \cdot) = g(t) \quad \text{on} \, ]0, T[ \times \partial \Omega , \]  
\[ \text{(5.1.4)} \]

Neumann boundary conditions:  \[ \sigma \, \text{grad} \, u \cdot n = h(t) \quad \text{on} \, ]0, T[ \times \partial \Omega , \]  
\[ \text{(5.1.5)} \]

\[ + \text{ initial conditions for } t = 0 \]

\[ u(0, \cdot) = u_0 \quad \text{in} \, \Omega . \]  
\[ \text{(5.1.6)} \]

\[ = \text{parabolic initial boundary value problem (IBVP)} \]

Assumption:

\[ g \, \text{continuous in time} \quad \& \quad \text{compatibility condition} \, u_0|_{\partial \Omega} = g(0) \]
Theorem 5.1.1 (Maximum principle for transient heat conduction). If $u_0 \in C^0(\overline{\Omega})$, $g \in C^0([0, T] \times \partial \Omega)$, and $f \equiv 0$, then a continuous (strong) solution of (5.1.3), (5.1.4), and (5.1.6) satisfies

$$
\min\{u_0, g\} \leq u(t, x) \leq \max\{u_0, g\} \quad \forall (t, x) \in \tilde{\Omega}.
$$

As in Sect. 1.6 & Sect. 1.8: variational formulation in space (for pure Dirichlet b.c. (5.1.4)):

seek $t \in ]0, T[ \mapsto u(t) \in g(t) + H^1_0(\Omega)$, $u(0) = u_0$, such that

$$
\int_{\Omega} \frac{d}{dt}(\rho u(t)) \ v \ dx + \int_{\Omega} \sigma \ \text{grad} u(t) \cdot \ \text{grad} v \ dx = \int_{\Omega} f(t) \ v \ dx \quad \forall v \in H^1_0(\Omega).
$$

General form of spatial variational formulation of parabolic evolution problem:

$$
t \in ]0, T[ \mapsto u(t) \in V : \begin{cases}
\frac{d}{dt}m(u(t), v) + a(u(t), v) = f(t)(v) & \forall v \in V, \\
u(0) = u_0 \in V.
\end{cases}
$$
with \( V = \text{Hilbert space}, \text{see Sect. 1.7} \).

\( V \)-elliptic (\( \rightarrow \) Def. 1.7.7) symmetric bilinear form \( a : V \times V \mapsto \mathbb{R} \) (independent of time),

symmetric positive definite bilinear form \( m : V \times V \mapsto \mathbb{R} \) (independent of time),

time-dependent linear form \( f(t) : V \mapsto \mathbb{R} \), \( 0 < t < T \).

Notation: “energy norm” \( \| u \|^2_A := a(u, u) \), “\( L^2 \)-type norm” \( \| u \|^2_M := m(u, u) \)

Assumption: \( \| \cdot \|_A \) is stronger norm:

\[
\exists c^* > 0: \quad c^* \| v \|_M \leq \| v \|_A \quad \forall v \in V . \quad (5.1.8)
\]

**Remark.** for (5.1.3) with (5.1.4) \( V \in H^1_0(\Omega) \): Thm. 1.7.9 \( \rightarrow \) assumption (5.1.8)

“dissipative” nature of (5.1.7):

\[
f \equiv 0 \quad \Rightarrow \quad \frac{1}{2} \frac{d}{dt} \| u \|^2_M = - \| u \|^2_A \leq - (c^*)^2 \| u \|^2_M \quad \Rightarrow \quad \| u(t) \|^2_M \leq \exp(-2(c^*)^2t) \| u(0) \|^2_M .
\]
Method of lines

Method of lines = technique for spatial semi-discretization of (5.1.7):

Idea: Galerkin discretization of (5.1.7): \( V \rightarrow V_N \subset V, \dim V_N < \infty, \) see Sect. 2.1.1.

\[
\begin{align*}
t \in [0, T[ \mapsto u(t) \in V_N & : \\
& \begin{cases} m\left(\frac{d}{dt}u_N(t), v_N\right) + a(u_N(t), v_N) = f(t)(v_N) \quad \forall v_N \in V_N, \\
u_{N,0} = \text{projection/interpolant of } u_0 \text{ in } V_N. \end{cases}
\end{align*}
\]

(5.2.1)

\( \Rightarrow \)

(5.2.1) \Rightarrow \begin{cases} \mathbf{M} \left\{ \frac{d}{dt} \tilde{\mu}(t) \right\} + \mathbf{A} \tilde{\mu}(t) = \tilde{\varphi}(t) \quad \text{for } 0 < t < T, \\
\tilde{\mu}(0) = \tilde{\mu}_0. \end{cases}

(5.2.2)

\( \Rightarrow \) s.p.d. stiffness matrix \( \mathbf{A} \in \mathbb{R}^{N, N}, (\mathbf{A})_{ij} := a(b^j_N, b^i_N) \) (independent of time),

\( \Rightarrow \) s.p.d. mass matrix \( \mathbf{M} \in \mathbb{R}^{N, N}, (\mathbf{M})_{ij} := m(b^j_N, b^i_N) \) (independent of time),

\( \Rightarrow \) source (load) vector \( \tilde{\varphi}(t) \in \mathbb{R}^{N}, (\tilde{\varphi}(t))_i := f(t)(b^i_N) \) (time-dependent),

\( \Rightarrow \) \( \tilde{\mu}_0 \) \( \overset{\Delta}{=} \) coefficient vector of a projection of \( u_0 \) onto \( V_N \).
Diagonalization:

\[ \begin{align*}
A, M & \quad \text{symmetric positive definite} \quad \Rightarrow \quad M^{-1/2}AM^{-1/2} \quad \text{symmetric positive definite} \\
\Rightarrow & \quad \exists \text{ orthogonal } T \in \mathbb{R}^{N,N} : \quad T^TM^{-1/2}AM^{-1/2}T = D := \text{diag}(\lambda_1, \ldots, \lambda_N),
\end{align*} \]

where the \( \lambda_i > 0 \) are generalized eigenvalues for \( A\xi = \lambda M\xi \quad \Rightarrow \quad \lambda_i \geq (c^*)^2 \) for all \( i \).

Transformation ("diagonalization") of (5.2.2):

\[ \begin{align*}
\ddot{\eta} & = T^T M^{1/2} \ddot{\mu} \\
\frac{d}{dt}\ddot{\eta}(t) + D\ddot{\eta} & = T^T M^{-1/2} \dot{\varphi}(t).
\end{align*} \]  \hfill (5.2.3)

\( \Rightarrow \) decoupled scalar ODEs (for eigencomponents \( \eta_i \) of \( \ddot{\mu} \)).

Note: for \( \ddot{\varphi} \equiv 0, \lambda > 0 \) : \( \eta_i(t) = \exp(-\lambda_i t)\eta_i(0) \to 0 \) for \( t \to \infty \)
How do the $\lambda_i$ behave?

Model problem:

$\mathcal{M} = \text{simplicial mesh of } \Omega \subset \mathbb{R}^2, \ g \equiv 0, \ \sigma = \rho \equiv 1, \ \mathcal{V}_N = \mathcal{S}^0_{1,0}(\mathcal{M})$:

\[ \|u\|_A = |u|_1, \quad \|u\|_M = \|u\|_0 \]

\[
\begin{align*}
\lambda_{\max}(M^{-1/2}A M^{-1/2}) &= \max_{\tilde{\xi} \neq 0} \frac{\tilde{\xi} \cdot \dot{A} \tilde{\xi}}{\tilde{\xi} \cdot \tilde{M} \tilde{\xi}} = \max_{v_N \in \mathcal{V}_N \setminus \{0\}} \frac{|v_N|^2_1}{\|v_N\|^2_0}, \\
\lambda_{\min}(M^{-1/2}A M^{-1/2}) &= \min_{\tilde{\xi} \neq 0} \frac{\tilde{\xi} \cdot \dot{A} \tilde{\xi}}{\tilde{\xi} \cdot \tilde{M} \tilde{\xi}} = \min_{v_N \in \mathcal{V}_N \setminus \{0\}} \frac{|v_N|^2_1}{\|v_N\|^2_0}.
\end{align*}
\]

\[ \text{diam}(\Omega)^{-2} \leq \lambda_{\min}(M^{-1/2}A M^{-1/2}), \quad \lambda_{\max}(M^{-1/2}A M^{-1/2}) \leq C(\rho,\mathcal{M}) \min\{h_K : K \in \mathcal{M}\}^{-2}. \]

1: Poincaré-Friedrichs inequality Thm. 1.7.9,

2: Inverse estimate (4.3.3) (cf. Def. 2.5.29, cf. Lemma 2.5.28).

Set $v_N := b_N^j, \ b_N^j$ “hat function” at vertex of smallest triangle \quad \Rightarrow \quad \text{bound for } \lambda_{\max} \text{ sharp}

Set $v_N = \text{interpolant of lowest Dirichlet eigenmode of } -\Delta \text{ on } \Omega \quad \Rightarrow \quad \text{bound for } \lambda_{\min} \text{ sharp}

\[ \lambda_{\min}(M^{-1/2}A M^{-1/2}) \approx \text{const.}, \]

\[ \lambda_{\max}(M^{-1/2}A M^{-1/2}) \approx C(\rho,\mathcal{M}, \mu,\mathcal{M}) h^{-2}, \quad \text{as } h,\mathcal{M} \to 0. \]
5.3 Timestepping

Focus: single step methods for (5.2.2), equidistant temporal mesh \( \{ t_j := j \Delta t : j = 0, \ldots, M \} \),
timestep \( \Delta t := T/M, M \in \mathbb{N} \).

\[ t_j VD_j = \Delta t \quad \ldots \quad M \]

\[ VD_j = \Delta t \quad \ldots \quad M \]

\[ \mu^{(j+1)} = \Phi_{M,A,\bar{\varphi}(t)}(j, \Delta t, \bar{\mu}^{(j)}) \quad , \quad \bar{\mu}^{(0)} = \bar{\mu}_0 , \] (5.3.1)

where \( \bar{\mu}^{(j)} \) is approximation for \( \bar{\mu}(t_j), t_j := j \Delta t \).

Assumptions:

- similarity transformation invariance: for regular \( S \in \mathbb{R}^{N,N} \)
  \[ \Phi_{M,A,\bar{\varphi}(t)}(j, \Delta t, \bar{\xi}) = S^{-1} \Phi_{S^{-1}MS,S^{-1}AS,S^{-1}\bar{\varphi}(t)}(j, \Delta t, S\bar{\xi}) . \]

- If \( \bar{\varphi} \) constant in time \( \Rightarrow \Phi_{M,A,\bar{\varphi}}(j, \Delta t, \bar{\xi}) = \Phi_{M,A,\bar{\varphi}}(\Delta t, \bar{\xi}) \).
If $M, A$ diagonal, separation of components

$$
\Phi_{M,A,\varphi(t)}(j, \Delta t, \xi) = \begin{pmatrix}
\Phi_{m11,a11,\varphi_1(t)}(j, \Delta t, \xi_1) \\
\vdots \\
\Phi_{mNN,aNN,\varphi_N(t)}(j, \Delta t, \xi_N)
\end{pmatrix}
$$

- To study (5.3.1) apply single step method to diagonalized system (5.2.3):

$$
\tilde{\eta}^{(j+1)} = \Phi_{I, D, T^T M^{1/2} \varphi(t)}(j, \Delta t, \tilde{\eta}^{(j)}) \quad , \quad \tilde{\eta}^{(0)} = T^T M^{1/2} \tilde{\mu}_0. \quad (5.3.2)
$$

- Examine single step method for scalar ODEs

$$
\frac{d}{dt} \eta(t) = -\lambda \eta(t) + \rho(t) , \quad \lambda > 0 , \quad \rho : [0, T] \mapsto \mathbb{R} \text{ Lipschitz-continuous} . \quad (5.3.3)
$$

[ solution \( \eta(t) = \eta(0) \exp(-\lambda t) + \int_0^t \rho(s) \exp(-\lambda (t - s)) \, ds \) ]

5.3.1 Stability

Crucial: (5.2.4) \(\rightarrow\) uniform stability/accuracy of the method for all \(\lambda \in ]0, \infty[\) !
Example: Explicit Euler scheme for (5.3.3), \( \rho \equiv 0 \)

\[
\eta^{(j+1)} = (1 - \Delta t \lambda) \eta^{(j)} \Rightarrow \eta^{(j)} = (1 - \Delta t \lambda)^j \eta^{(0)} .
\]

Example: Implicit Euler scheme for (5.3.3), \( \rho \equiv 0 \)

\[
\eta^{(j+1)} = (1 + \Delta t \lambda)^{-1} \eta^{(j)} \Rightarrow \eta^{(j)} = \left( \frac{1}{1 + \Delta t \lambda} \right)^j \eta^{(0)} .
\]

\\
\textbf{⇒ Implicit Euler timestepping for (5.2.2):}

\[
(M + \Delta t A) \bar{\mu}^{(j+1)} = M \bar{\mu}^{(j)} + \Delta t \bar{\varphi}((j + 1) \Delta t) , \quad j = 0, \ldots, M - 1 .
\] (5.3.4)

Single step method applied to (5.3.3) with \( \rho = 0 \) \Rightarrow difference equation of the form

\[
\eta^{(j+1)} = R(-\Delta t \lambda) \eta^{(j)} \text{ with rational function } R .
\] (5.3.5)

Explicit Euler: \( R(z) = 1 + z \),

Implicit Euler: \( R(z) = (1 - z)^{-1} \).
Definition 5.3.1. A single step method $\eta^{(j+1)} = \Phi(\Delta t, \eta^{(j)})$ for (5.3.3) with $\rho \equiv 0$ is called stable, if

$$\forall \lambda > 0, \Delta t > 0, \eta^{(0)} \in \mathbb{R}: \eta^{(j)} \text{ bounded } \forall j \in \mathbb{N}.$$ 

Corollary 5.3.2. A single step method with associated rational function $R$ according to (5.3.5) is stable, if $|R(z)| \leq 1$ $\forall z < 0$.

- Explicit Euler timestepping unstable: oscillations & blow-up for $\Delta t \lambda > 2$
- Implicit Euler $L(\pi)$-stable: iterates $\eta^{(j)}$ decay for all $\lambda, \Delta t > 0$.

Definition 5.3.3. A stable single step method with associated rational function $R$ according to (5.3.5) is $L(\pi)$-stable, if $\lim_{z \to -\infty} R(z) = 0$.

$L(\pi)$-stability $\Rightarrow$ rapid decay of $|\eta^{(j)}|$ for $\lambda \gg 1$ (mimics behavior of $\eta(t)$) $\Rightarrow$ timestepping schemes for parabolic problems should be $L(\pi)$-stable.
Widely used: Runge-Kutta (RK) methods (→ higher order timestepping)

Description of RK-methods by Butcher’s tableau: for RK method with \( s, s \in \mathbb{N} \), stages:

\[
\begin{array}{c|cccc}
\vec{c} & \mathbf{A} & \vec{b} \\
\hline
c_1 & a_{11} & a_{12} & \cdots & a_{1s} \\
c_2 & a_{21} & \ddots & & a_{2s} \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
c_s & a_{s1} & \ddots & & a_{ss} \\
\end{array}
\]

\( \vec{c}, \vec{b} \in \mathbb{R}^s, \quad \mathbf{A} \in \mathbb{R}^{s \times s} \).

Timestepping scheme for (5.2.2): compute \( \tilde{\mu}^{(j+1)} \) from \( \tilde{\mu}^{(j)} \) through

\[
(\kappa_i)_{i=1,\ldots,s} \in (\mathbb{R}^N)^s: \quad M \vec{\kappa}_i + \sum_{j=1}^{s} \Delta t a_{ij} A \vec{\kappa}_j = \varphi(t_j + c_i \Delta t) - A \tilde{\mu}^{(j)}, \quad i = 1, \ldots, s,
\]

\[
\tilde{\mu}^{(j+1)} = \tilde{\mu}^{(j)} + \Delta t \sum_{j=1}^{s} \vec{\kappa}_j b_j.
\]
Variational notation:

\[ k_i \in V_N: \quad m(k_i, v_N) + \sum_{j=1}^{s} \Delta t a_i j a(k_j, v_N) = f(t + c_i \Delta t)(v_N) - a(u^{(j)}_N, v_N) \quad \forall v_N \in V_N , \]

\[ u^{(j+1)}_N = u^{(j)}_N + \sum_{j=1}^{s} b_j k_j . \]

Note: \( \vec{b}, \vec{c}, \) and \( \mathcal{A} \) need to be carefully chosen to get consistent timestepping → literature [8].

Example: RADAU-3 scheme (for (5.2.2)):

\[
\begin{array}{c|cc}
1/3 & 5/12 & -1/12 \\
1 & 3/4 & 1/4 \\
3/4 & 1/4 & 1/4
\end{array}
\]

\[ (M + \frac{5}{12} \Delta t A)k_1 - \frac{1}{12} \Delta t A k_2 = \phi(t_j + \frac{1}{3} \Delta t) - A \mu^{(j)} , \]

\[ \frac{3}{4} \Delta t A k_1 + (M + \frac{1}{4} \Delta t A) k_2 = \phi(t_j + \Delta t) - A \mu^{(j)} , \]

\[ \mu^{(j+1)} = \mu^{(j)} + \frac{3}{4} \vec{k}_1 + \frac{1}{4} \vec{k}_2 . \]

Example: SDIRK-2 scheme:

\[
\begin{array}{c|ccc}
\lambda & \lambda & 0 \\
1 & 1 - \lambda & \lambda \\
1 - \lambda & \lambda & 1 - \lambda \\
\end{array}
\]

with \( \lambda := 1 - \frac{1}{2} \sqrt{2} . \)
5.3.2 Convergence

- Implicit Euler timestepping for (5.2.1) (fully discrete scheme)

Continuous, semi-discrete, and fully discrete variational formulations:

\[ m\left(\frac{d}{dt}u(t), v\right) + a(u(t), v) = f(t)(v) \quad \forall v \in V , \quad (5.3.6) \]
\[ m\left(\frac{d}{dt}u_N(t), v_N\right) + a(u_N(t), v_N) = f(t)(v_N) \quad \forall v_N \in V_N , \quad (5.3.7) \]
\[ m\left(\frac{1}{\Delta t}(u_N^{(j+1)} - u_N^{(j)}), v_N\right) + a(u_N^{(j+1)}, v_N) = f(t_{j+1})(v_N) \quad \forall v_N \in V_N . \quad (5.3.8) \]

Tool: Galerkin projection \( P_N : V \mapsto V_N \), see Def. [2.1.4]:

\[ P_N u \in V_N : \quad a(P_N u, v_N) = a(u, v_N) \quad \forall v_N \in V_N . \quad (5.3.9) \]
Note:

\[ P_N \leftrightarrow \text{“Galerkin solution operator”:} \]

\[
\begin{align*}
  u \in V & : \quad a(u, v) = f(v) \quad \forall v \in V, \\
  u_N \in V_N & : \quad a(u_N, v) = f(v_N) \quad \forall v \in V_N,
\end{align*}
\]

\[ \Rightarrow u_N = P_N u. \]

\[ V = H^1_0(\Omega), \quad V_N = \delta^0_{1,0}(\mathcal{M}), \quad a(u, v) = \int_{\Omega} \nabla u \cdot \nabla v \, dx, \quad \text{Dirichlet problem for } -\Delta \]

2-regular:

\[ \text{Thm. 2.5.26 } \Rightarrow \exists C = C(\Omega, \rho_M): \quad ||u - P_N u||_0 \leq C h_M^2 ||u||_2. \quad (5.3.10) \]

Idea: first investigate \( w_N^{(j)} := P_N u(t_j) - u_N^{(j)} \in V_N \) (instead of \( u(t_j) - u_N^{(j)} \in V \) directly)

\[
\begin{align*}
m\left( \frac{w^{(j+1)} - w^{(j)}}{\Delta t}, v_N \right) + a(w^{(j+1)}, v_N) & = m\left( P_N \left\{ \frac{u(t_{j+1}) - u(t_j)}{\Delta t} \right\}, v_N \right) + a(P_N u(t_{j+1}), v_N) - f(v_N) \\
& = m\left( \frac{u(t_{j+1}) - u(t_j)}{\Delta t}, v_N \right) - \frac{du}{dt}(t_{j+1}, v_N) + m((P_N - Id)\left\{ \frac{u(t_{j+1}) - u(t_j)}{\Delta t} \right\}, v_N) \\
& = m(e^j_1 + e^j_2, v_N),
\end{align*}
\]
where

\[ e_1^j = -\frac{1}{\Delta t} \int_{t_j}^{t_{j+1}} \int_\sigma \frac{d^2u}{dt^2}(\tau) \, d\tau \, d\sigma \quad \Rightarrow \quad \|e_1^j\|_M \leq \frac{1}{\Delta t} \int_{t_j}^{t_{j+1}} \left\| \frac{d^2u}{dt^2}(\tau) \right\|_M \, d\tau , \]

\[ e_2^j = \frac{1}{\Delta t} \int_{t_j}^{t_{j+1}} \left( P_N - Id \right) \frac{du}{dt}(\tau) \, d\tau \quad \Rightarrow \quad \|e_2^j\|_M \leq \frac{1}{\Delta t} \int_{t_j}^{t_{j+1}} \left\| P_N - Id \right\| \frac{du}{dt}(\tau) \right\|_M \, d\tau . \]

1: use (5.3.8)
2: use (5.3.9) & (5.3.6) for \( t = t_{j+1} \)
3: use fundamental theorem of calculus twice

Set \( v_N := w_N^{(j+1)} \):

\[ \frac{1}{\Delta t} \left( \|w_N^{(j+1)}\|_M^2 - m(w_N^{(j+1)}, w_N^{(j)}) \right) + \|w_N^{(j+1)}\|_A^2 \leq (\|e_1^j\|_M + \|e_2^j\|_M) \|w_N^{(j+1)}\|_M , \]

\[ \Rightarrow \quad \|w_N^{(j+1)}\|_M - \|w_N^{(j)}\|_M \leq \Delta t \left( \|e_1^j\|_M + \|e_2^j\|_M \right) . \]
(*) use Cauchy-Schwarz inequality

\[ |m(w^{(j+1)}_N, w^{(j)}_N)| \leq \|w^{(j+1)}_N\|_M \|w^{(j)}_N\|_M. \]

\[ \|w^{(j)}_N\|_M \leq \|w^{(0)}_N\|_M + \Delta t \sum_{k=0}^{j-1} \left( \|e^k_1\|_M + \|e^k_2\|_M \right) \]

\[ \leq \|w^{(0)}_N\|_M + \Delta t \int_0^{t_j} \|\frac{d^2u}{dt^2} (\tau)\|_M \, d\tau + \int_0^{t_j} \|(P_N - I) \{\frac{du}{dt} (\tau)\}\|_M \, d\tau. \]

+ \triangle - inequality

\[ \|u(t_j) - u^{(j)}_N\|_M \leq \|(I - P_N)u(t_j)\|_M + \|P_Nu_0 - u_{N,0}\|_M + \]

\[ \Delta t \int_0^{t_j} \|\frac{d^2u}{dt^2} (\tau)\|_M \, d\tau + \int_0^{t_j} \|(P_N - I) \{\frac{du}{dt} (\tau)\}\|_M \, d\tau. \]

Example:

\[ V = H^1_0(\Omega), \quad V_N = \mathcal{S}^0_{1,0}(\mathcal{M}), \quad a(u, v) = \int_\Omega \nabla u \cdot \nabla v \, dx, \]

Dirichlet problem for \(-\triangle\) regular.
Assumptions: (on regularity of $u(t)$)

\[
\int_0^T \left\| \frac{d^2 u}{dt^2}(\tau) \right\|_0^2 \, d\tau < \infty, \quad \int_0^T \left\| \frac{du}{dt}(\tau) \right\|_2^2 \, d\tau < \infty, \quad \sup_{0 \leq t \leq T} \|u(t)\|_2 < \infty.
\]

(5.3.10) \Rightarrow \left\| u(t_j) - u_N^{(j)} \right\|_0 \leq C h_M^2 \left\{ \sup_{0 \leq t \leq T} \|u(t)\|_2 + \int_0^T \left\| \frac{du}{dt}(\tau) \right\|_2^2 \, d\tau \right\} + \Delta t \int_0^T \left\| \frac{d^2 u}{dt^2}(\tau) \right\|_0^2 \, d\tau,

with $C = C(\Omega, \rho_M)$.

Short notation: \[ \left\| u(t_j) - u_N^{(j)} \right\|_0 = O(h_M^2 + \Delta t). \]

"Second order in space & first order in time"

Required: balancing of $\Delta t$ and $h_M$.

**Remark.** Assuming sufficient regularity of $t \mapsto u(t)$

- for SDIRK-2: \[ \left\| u(t_j) - u_N^{(j)} \right\|_0 = O(h_M^2 + \Delta t^2) \] (second order in time)
- for RADAU-3: \[ \left\| u(t_j) - u_N^{(j)} \right\|_0 = O(h_M^2 + \Delta t^3) \] (third order in time)
6.1 Traffic flow

Simple mathematical model for non-stationary traffic flow on a single highway lane:

- spatial domain $\Omega = ] - L, L [ \triangleq \text{“long” highway lane}$,
- computational domain $\tilde{\Omega} = ]0, T [ \times \Omega$, $T > 0 = \text{end time}$
- $u : \tilde{\Omega} \mapsto \mathbb{R} \triangleq \text{density of cars \,(\text{cars per unit length})\,\,\text{scaling} \rightarrow 0 \leq u \leq 1}$,
- constitutive law for speed of cars

$$v = v_{\text{max}}(1 - u) > 0 \quad \Rightarrow \quad \text{cars moving in \,+\,-direction,}$$
• boundary condition \( u(t, -L) = g(t) \) at inflow boundary,
• initial condition \( u(0, x) = u_0(x), \ x \in \Omega \).

\[
\text{Flux of cars:} \quad F(u) = v_{\text{max}}u(1 - u). \quad (6.1.1)
\]

\[\begin{align*}
\text{Balance law (integral form) for } u: & \quad \text{for all } [t_0, t_1] \times [x_0, x_1] \subset \tilde{\Omega} \\
& \quad \int_{x_0}^{x_1} u(t_1, x) \, dx - \int_{x_0}^{x_1} u(t_0, x) \, dx = \int_{t_0}^{t_1} F(u(t, x_0)) \, dt - \int_{t_0}^{t_1} F(u(t, x_1)) \, dt \\
\end{align*} \quad (6.1.2)
\]

\[\begin{align*}
\text{Balance law (differential form)} \rightarrow \text{ evolution equation} \\
\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} (F \circ u) &= 0 \quad \text{in } \tilde{\Omega}, \quad u(0, \cdot) = u_0 \quad \text{in } \Omega, \quad u(\cdot, -L) = g \quad \text{for } 0 < t < T. \quad (6.1.3)
\end{align*} \]

**Transient heat conduction (5.1.3), Sect. 5.1:**
\( \Rightarrow \) parabolic evolution problem: flux = function of spatial derivatives of \( u \) (\( j = -\sigma \ \text{grad} \ u \))

**Traffic flow equations (6.1.3):**
\( \Rightarrow \) hyperbolic evolution problem: flux = function of \( u \) (\( F = v_{\text{max}}u(1 - u) \))
6.2 General conservation laws

- Domain $\tilde{\Omega} := ]0, T[ \times \Omega$, $\Omega \subset \mathbb{R}^d$, unbounded $\Omega$ and $T = \infty$ possible
- Unknown state vector $u : \tilde{\Omega} \mapsto \mathbb{R}^q$, $q \in \mathbb{N}$ $\mapsto \mathbb{R}^q \equiv$ state space (densities)
- Flux function $F : \mathbb{R}^q \mapsto \mathbb{R}^{q,d}$
- Source term $g : \tilde{\Omega} \mapsto \mathbb{R}^q$
- Initial conditions $u_0 : \Omega \mapsto \mathbb{R}^q$

\[ \frac{\partial u}{\partial t} + \text{div}(F \circ u) = g \quad \text{in } \tilde{\Omega}, \quad (6.2.1) \]
\[ u(0, \cdot) = u_0 \quad \text{in } \Omega, \quad (6.2.2) \]

+ boundary conditions at inflow boundaries.

Terminology: $q = 1 \iff (6.2.1) \Rightarrow$ scalar conservation law
Focus: Scalar conservation laws in 1D, i.e., $d = 1, q = 1$: $\tilde{\Omega} = ]0, T[ \times I, I \subset \mathbb{R}$ connected

$$\frac{\partial u}{\partial t}(t, x) + \frac{\partial}{\partial x} F(u(t, x)) = 0 \quad \text{in } \tilde{\Omega}, \quad (6.2.3)$$

$$u(0, x) = u_0(x) \quad \forall x \in I.$$  

with differentiable flux function $F : \mathbb{R} \mapsto \mathbb{R}$.

Example: Pure linear advection on $\mathbb{R}$, $F(u) = au$, cf. limit case of Sect. 3.1: for $a \in \mathbb{R} \setminus \{0\}$

$$\frac{\partial u}{\partial t} + a \frac{\partial u}{\partial x} = 0 \quad \text{in } \mathbb{R}, \quad u(0, x) = u_0(x) \quad \forall x \in \mathbb{R}. \quad (6.2.4)$$

Exact solution $u(t, x) = u_0(x - ta)$.

6.2.1 Characteristics

$u \in C^1(\tilde{\Omega})$ \& $u$ solves $(6.2.3)$ $\Rightarrow$ $u =$ classical solution of $(6.2.3)$. 
Lemma 6.2.1. Every classical solution of (6.2.3) is constant on the characteristic curves $s : ]0, T[ \rightarrow \mathbb{R}$ that solve

$$\frac{ds}{dt} = F'(u(t, s(t))) \quad \text{for } 0 < t < T \quad , \quad s(0) = x_0 \in I.$$ 

Proof. Chain rule & (6.2.3)

$$\frac{d}{dt} u(t, s(t)) = \frac{\partial u}{\partial t}(t, s(t)) + \frac{\partial u}{\partial x}(t, s(t)) F'(u(t, s(t))) = \frac{\partial u}{\partial t}(t, s(t)) + \frac{\partial}{\partial x} (F \circ u)(t, s(t)) = 0.$$ 

Characteristic through $(0, x_0), x_0 \in I$, is straight line with slope $F'(u_0(x_0))$ in the $(t, x)$-plane.
Example: pure linear advection

 Characteristics $s(t) = vt + c$, $c \in \mathbb{R}$.

For conservation laws: information propagates along characteristics (with finite speed)
solution value $u(t^*, x^*)$ of (6.2.1) not affected by small perturbations of $u_0$, $g$, outside a finite region

$$D(t^*, x^*) \subset \{ (t, x) \in \tilde{\Omega}: t < t^* \}.$$ 

**Maximal** region of dependence for scalar conservation law (6.2.3):

$$D(t^*, x^*) \subset \{(x, t) \in \tilde{\Omega}: |x^* - x| \leq \sup_{u \in \mathcal{R}} |F'(u)| (t^* - t) \},$$

(6.2.5)

$$\mathcal{R} := [\inf_{x \in I} u_0(x), \sup_{x \in I} u_0(x)], \text{ cf. Thm. 6.2.5}.$$
Example: traffic flow (6.1.1) & (6.1.3)

\[ F'(u) = v_{\text{max}}(1 - 2u) \quad \text{(decreasing)} \]

\( u_0 \) smooth & increasing

Characteristics intersect!

Evolution of classical solution for traffic flow and smooth increasing \( u_0 \):
As $t \to 0.7$: $u(t, \cdot)$ steeper and steeper

The “wave breaks” $\rightarrow$ shock develops

Classical solutions “break down” in finite time $\leftrightarrow$ inadequate concept for (6.2.3)

### 6.2.2 Weak solutions

We consider (6.2.3) on $\tilde{\Omega} := \mathbb{R}_{\geq 0} \times \mathbb{R}$:

$$\frac{\partial u}{\partial t}(t, x) + \frac{\partial}{\partial x} F(u(t, x)) = 0 \quad \text{in } \mathbb{R}_{\geq 0} \times \mathbb{R}, \quad u(0, x) = u_0(x) \quad \forall x \in \mathbb{R}. \tag{6.2.6}$$
**Definition 6.2.2** (Weak solutions of scalar conservation law). \( u : \mathbb{R}_{\geq 0} \times \mathbb{R} \mapsto \mathbb{R} \) is a weak solution of (6.2.6), if

\[
\int_{\mathbb{R}} \int_{0}^{\infty} u(t, x) \frac{\partial \phi}{\partial t} + F(u(t, x)) \frac{\partial \phi}{\partial x} \, dt \, dx = -\int_{\mathbb{R}} u_0(x) \phi(0, x) \, dx \quad \forall \phi \in C^1_0(\mathbb{R}_{\geq 0} \times \mathbb{R}) .
\]

Notation: \( C^1_0(\mathbb{R}_{\geq 0} \times \mathbb{R}) := \{ u \in C^1(\mathbb{R}_{\geq 0} \times \mathbb{R}) : \text{supp}(u) \text{ compact} \} \)

**Definition 6.2.3** (Riemann problem).

\[
u_0(x) = \begin{cases} 
    u_l & \text{if } x < x_0 , \\
    u_r & \text{if } x \geq x_0 ,
\end{cases} \quad \overset{\approx}{=} \text{Riemann problem for (6.2.6)} .
\]

Riemann problems for traffic flow equation (6.1.3): \( F(u) = v_{\text{max}} u (1 - u) , \ X_0 = 0 \)

Note: \( F'(u) = v_{\text{max}} (1 - 2u) \) decreasing, flux \( u \mapsto F(u) \) strictly concave
If $u_l < u_r$:

- moving discontinuity = shock

$$u(t, x) = \begin{cases} u_l & \text{if } x < st, \\ u_r & \text{if } x > st, \end{cases} \quad (6.2.10)$$

with shock speed $s \in \mathbb{R}$ according to Rankine-Hugoniot condition

$$s = \frac{F(u_l) - F(u_r)}{u_l - u_r}. \quad (6.2.11)$$

If $u_l > u_r$:

- transition = rarefaction wave

$$u(t, x) = \begin{cases} u_l & \text{if } x < F'(u_l)t, \\ v(x/t) & \text{if } F'(u_l) \leq x/t \leq F'(u_r), \\ u_r & \text{if } x > F'(u_r)t, \end{cases} \quad (6.2.12)$$

where $v(\xi)$ solves $F'(v(\xi)) = \xi$. 

6.2 p. 373
Both, (6.2.10) (shock) and (6.2.12) (rarefaction wave) are weak solutions of the Riemann problem for (6.2.6).

Riemann problem for traffic flow: shock configurations (blue = shock in \((t,x)\)-plane):
Riemann problem for traffic flow: rarefaction waves ($\text{red} = u_0$):

**Remark.** If flux $u \mapsto F(u)$ in (6.2.6) strictly convex:

- Reversal: shock, if $u_l > u_r$ & rarefaction wave, if $u_l < u_r$

**Remark.** Non-physical shocks

Riemann problem for traffic flow, $u_l > u_r$ ("rarefaction situation"): 
Another weak solution:

\[ u(t, x) = \begin{cases} 
  u_l & \text{if } x < st , \\
  u_r & \text{if } x > st ,
\end{cases} \]

with \( s \) given by Rankine-Hugoniot condition (6.2.11).

**Non-physical shock**

with characteristics emenating from shock

(\(\rightarrow\) creation of information in shock)

Additional entropy conditions for selection meaningful weak solutions.

Example of an entropy condition:
Definition 6.2.4 (Oleinik’s entropy condition). A weak solution \( u : \tilde{\Omega} \rightarrow \mathbb{R} \) of (6.2.6) satisfies the entropy condition, if at any curve of discontinuity \( t \mapsto s(t), 0 \leq t_1 < t < t_2, \)

\[
\frac{F(u_-(t)) - F(v(t))}{u_-(t) - v(t)} \geq \frac{ds}{dt}(t) \geq \frac{F(u_+(t)) - F(v(t))}{u_+(t) - v(t)} \quad \forall t \in ]t_1, t_2[, \\
\text{and all } u_-(t) < v(t) < u_+(t), \text{ where}
\]

\[
u_\pm(t) = \lim_{\epsilon \to 0} u(t, s(t) \pm \epsilon), \quad t_1 < t < t_2.
\]

“Characteristics must impinge on shock”

Theorem 6.2.5 (Monotonicity for scalar 1D conservation laws). If Oleinik’s entropy condition (→ Def. 6.2.4) is imposed and \( u_0 \in L^\infty(\mathbb{R}) \), then (6.2.6) has a unique weak solution \( u \in L^\infty(\tilde{\Omega}) \) that satisfies

\[
\|u(t, \cdot)\|_{\infty, \mathbb{R}} \leq \|u_0\|_{\infty, \mathbb{R}} \quad \forall t > 0.
\]

If \( v \) denotes the solution for initial condition \( v_0 \in L^\infty(\mathbb{R}) \), then

\[
u_0 \geq v_0 \quad \text{a.e. in } \mathbb{R} \quad \Rightarrow \quad u(t, \cdot) \geq v(t, \cdot) \quad \text{a.e. in } \mathbb{R} \quad \text{for all } t > 0.
\]
6.3 Finite volume methods for scalar conservation laws

Focus: 1D scalar conservation law, periodic boundary conditions,
\( \tilde{\Omega} := ]0, T[ \times ]-1, 1[ \), \( T > 0 \):

\[
\frac{\partial u}{\partial t}(t, x) + \frac{\partial}{\partial x} F(u(t, x)) = 0 \quad \text{in } \tilde{\Omega} ,
\]

\[
u(t, -1) = u(t, 1) \quad \forall t \in ]0, T[ ,
\]

\[
u(0, x) = u_0(x) \quad \forall x \in ]-1, 1[ .
\]

Equidistant space-time mesh: with \( N, M \in \mathbb{N}, \tau = T/M, h = 2/N \)

\[
\mathcal{M} := \{ \eta^n, \eta^n \otimes x_j, x_{j+1} : \eta^n := n \tau, x_j := jh, j = -N, \ldots, N - 1, n = 1, \ldots, M \} .
\]

\[ \text{Dual spatial cells: } ]x_{j-1/2}, x_{j+1/2}[ , \quad x_{j+1/2} := (j + 1/2)h , \quad j = -N, \ldots, N - 1 . \]

Terminology:
\[ \tau \triangleq \text{timestep}, \quad h \triangleq \text{meshwidth} \]

Remark. "space wrapping" for torus \( ]-1, 1[ = \mathbb{R}/] -1, 1[ : \quad \text{regard } \mathcal{M} \text{ as mesh of } ]0, T[ \times \mathbb{R}. \)

Principle: finite volume methods approximate (dual) cell averages on \( \mathcal{M} : \)

\[
\mu_j^n \approx \frac{1}{h} \int_{x_{j-1/2}}^{x_{j+1/2}} u(t^n, x) \, dx , \quad j = -N + 1, \ldots, N ,
\]

\[
n = 1, \ldots, M .
\]
Notation:  
- approximation in \( n \)-th timestep: \( \vec{\mu}^n := (\mu_j^n)_{j=-N+1}^N \in \mathbb{R}^{2N} \)
- piecewise constant function associated with \( \vec{\mu}^n \)

\[
u^n(x) := \sum_{j=-N}^{N-1} \mu_j^n \chi_{[x_{j-1/2}, x_{j+1/2}]}(x), \quad \chi_I \triangleq \text{characteristic function of } I \subset \mathbb{R}.
\]

\[\nabla \text{ Timestepping: } \vec{\mu}^{n+1} = \mathcal{H}(\vec{\mu}^n), \quad n = 0, \ldots, M - 1, \quad \text{with } \mathcal{H} : \mathbb{R}^{2N} \mapsto \mathbb{R}^{2N}. \quad (6.3.2)\]

### 6.3.1 Difference equations in conservation form

From integral form of 1D conservation law \((6.2.1)\) (with \( g = 0 \)):

\[
\int_{x_{j-1/2}}^{x_{j+1/2}} u(t^{n+1}, x) \, dx - \int_{x_{j-1/2}}^{x_{j+1/2}} u(t^n, x) \, dx = \int_{t^n}^{t^{n+1}} F(u(t, x_{j-1/2})) \, dt - \int_{t^n}^{t^{n+1}} F(u(t, x_{j+1/2})) \, dt
\]
Explicit 3-point difference equation for cell averages: with Lipschitz continuous $\Phi : \mathbb{R}^{2q} \mapsto \mathbb{R}^q$:

$$\mu_{j}^{n+1} - \mu_{j}^{n} = -\frac{\tau}{h} \left\{ \Phi(\mu_{j}, \mu_{j+1}^{n}) - \Phi(\mu_{j-1}^{n}, \mu_{j}^{n}) \right\} . \tag{6.3.3}$$

Difference stencil for explicit 3-point scheme:
$$(\mu_{j}^{n+1} \text{ depends on } \mu_{j-1}^{n}, \mu_{j}^{n}, \mu_{j+1}^{n})$$

**Definition 6.3.1.** (6.3.3) represents the conservation form of an explicit 3-point difference scheme for (6.2.1) with numerical flux function $\Phi : \mathbb{R}^{2q} \mapsto \mathbb{R}^q$.

Why “conservation form”?
**Telescopic sum argument**:

\[
\int_{-1}^{1} u^{n+1}(x) \, dx = h \sum_{j=-N}^{N-1} \mu_j^{n+1} = h \sum_{j=-N}^{N-1} \mu_j^n - \frac{\tau}{h} \left( \Phi(\mu_j^n, \mu_{j+1}^n) - \Phi(\mu_{j-1}^n, \mu_j^n) \right)
\]

\[
= h \sum_{j=-N}^{N-1} \mu_j^n = \int_{-1}^{1} u^n(x) \, dx .
\]

cf. for (6.2.1) with \( g = 0 \):

\[
\int_{-1}^{1} u(x, t) \, dx = \text{const}. .
\]

**scalar case**: correct speed of “numerical shock” for Riemann problem

---

**Numerical region of dependence**:

\[
D_N(t^n, x_j) := \{(t, x) \in ]0, T[ \times \mathbb{R} : \frac{|x_j - x|}{t_j - t} \leq \frac{h}{\tau} \}.
\]

---

6.3

p. 381
**Definition 6.3.2.** An explicit 3-point difference scheme satisfies the **Courant-Friedrichs-Lewy (CFL) condition**, if for all $n, j$

\[
\text{Region of dependence } D(t^n, x_j) \subset \text{Numerical region of dependence } D_N(t^n, x_j).
\]

CFL-condition necessary for convergence of explicit 3-point scheme as $h, \tau \to 0$.

CFL-condition $\Rightarrow$ timestep constraint: $\exists C = C(F, u): \tau \leq Ch$.

For 1D scalar conservation law (6.2.3)

\[
(6.2.5) \Rightarrow \frac{\tau}{h} \sup_{\inf u_0 \leq u \leq \sup u_0} |F'(u)| \leq 1 \text{ sufficient for CFL condition.} \tag{6.3.4}
\]

Simplest (and usual) choice in (6.3.3): $\tau : h = \text{const.}$. \tag{6.3.5}

Notation:

\[
\gamma := \frac{\tau}{h}
\]
6.3.2 Consistency

\[
\begin{align*}
\int_{x_{j-1/2}}^{x_{j+1/2}} u(t^{n+1}, x) \, dx - \int_{x_{j-1/2}}^{x_{j+1/2}} u(t^n, x) \, dx &= \int_{t^n}^{t^{n+1}} F(u(t, x_{j-1/2})) \, dt - \int_{t^n}^{t^{n+1}} F(u(t, x_{j+1/2})) \, dt \\
&\Downarrow \\
\mu_{j}^{n+1} - \mu_{j}^{n} &= -\gamma \left\{ \Phi(\mu_{j}^{n}, \mu_{j+1}^{n}) - \Phi(\mu_{j-1}^{n}, \mu_{j}^{n}) \right\} \quad &\& \mu_{j}^{n} &\approx \frac{1}{h} \int_{x_{j-1/2}}^{x_{j+1/2}} u(t^n, x) \, dx
\end{align*}
\]

Numerical flux
\[
\Phi(\mu_{j}^{n}, \mu_{j+1}^{n}) \approx \frac{1}{\tau} \int_{t^n}^{t^{n+1}} F(u(t, x_{j+1/2})) \, dt
\]

**Definition 6.3.3.** Numerical flux \( \Phi : \mathbb{R}^2 \mapsto \mathbb{R} \) is consistent, if \( \Phi \) Lipschitz continuous and \( \Phi(u, u) = F(u) \) for all \( u \in \mathbb{R} \).

Recall: Taylor expansion technique for consistency error estimate for difference schemes \( \rightarrow \) Sect. 2.3.
Plug (restriction of) exact solution \( u = u(t^n, x_j) \) into 3-point difference formula \((6.3.3)\)

**Definition 6.3.4.** An explicit 3-point difference scheme \((6.3.3)\) for \((6.2.3)\) is consistent with \((6.2.3)\) of order \( p \) in time and order \( q \) in space, \( p, q \in \mathbb{N} \), if, for \( h, \tau \to 0 \) (with bounded \( \tau/h \)) and all \( (t, x) \in \tilde{\Omega} \)

\[
\frac{u(t + \tau, x) - u(t, x)}{\tau} + \frac{\Phi(u(t, x), u(t, x + h)) - \Phi(u(t, x - h), u(t, x))}{h} = O(\tau^p, h^q),
\]

for any smooth (classical) solution \( u(t, x) \) of \((6.2.3)\).

Example: natural choice centered flux \( \Phi(u, v) = \frac{1}{2}(F(u) + F(v)) \)

\[\text{(6.3.3)}: \quad \mu_{j+1}^n - \mu_j^n = -\frac{1}{2} \gamma \left( F(\mu_{j+1}^n) - F(\mu_{j-1}^n) \right), \quad \gamma := \tau/h. \quad (6.3.6)\]

If \( u \in C^3(\tilde{\Omega}), \ F \in C^2(\mathbb{R}) \) first order in time, second order in space
6.3.3 Stability

Consider scalar advection \( F(u) = au, \ a \in \mathbb{R} \setminus \{0\} \):

\[
\frac{\partial u}{\partial t}(t, x) + a \frac{\partial u}{\partial x}(t, x) = 0 \quad \text{in} \ [0, T] \times ]-1, 1[ , \quad u(0, x) = u_0(x) . \tag{6.2.4}
\]

CFL-condition (6.3.4) for 3-point upwind scheme for scalar advection: \( |a| \gamma \leq 1 \)

(6.3.6) \( \Rightarrow \) “Natural” explicit 3-point centered finite difference scheme

\[
\mu_j^{n+1} - \mu_j^n = -\frac{1}{2} a \gamma (\mu_{j+1}^n - \mu_{j-1}^n) . \tag{6.3.7}
\]

CAUTION!

(6.3.7) \( \triangleq \) centered finite difference stencil for \( \frac{\partial}{\partial x}u \):

deja vu \( \rightarrow \) Sect. 3.1.3

(Linear) Von Neumann stability analysis:

If \( \tilde{\mu}^n = (\exp(i h j))^N_{j=-N}, \ \omega \in \frac{\pi}{N} \mathbb{Z} \ \Rightarrow \ \tilde{\mu}^{n+1} = (1 - \frac{1}{2} i \gamma a \sin(\omega)) \tilde{\mu}^n . \)
inevitable blow-up of solutions as $\tau \to 0$ (instability).

Recall remedy (→ Sect. 3.1.3): upwinding ($\equiv$ one-sided difference quotients for $\frac{\partial u}{\partial x}$)

$$
a > 0: \quad \mu_j^{n+1} - \mu_j^n = -a\gamma (\mu_j^n - \mu_{j-1}^n), \quad (6.3.8)
$$

$$
a < 0: \quad \mu_j^{n+1} - \mu_j^n = -a\gamma (\mu_{j+1}^n - \mu_j^n). \n$$

\ Philosophical Society

Upwind numerical flux for $6.2.4$: \hspace{1cm} $\Phi(u, v) = \begin{cases} au \ (\equiv F(u)) \quad \text{if} \ a > 0, \\ av \ (\equiv F(v)) \quad \text{if} \ a < 0. \end{cases}$ \hspace{1cm} (6.3.9)

Von Neumann stability analysis:

If $\mu^n = (\exp(ihj))_{j=-N}^N, \omega\in\frac{\pi}{N}\mathbb{Z} \Rightarrow \mu^{n+1} = (1 - \gamma a + \gamma a \exp(-i\omega)) \mu^n$.

$L^\infty$-stability (for $a > 0$):

$$
|\mu_j^{n+1}| = |(1 - a\gamma)\mu_j^n + a\gamma \mu_{j-1}^n| \leq \max\{|\mu_j^n|, |\mu_{j-1}^n|\}, \quad \text{if} \ a\gamma \leq 1. \hspace{1cm} (6.3.10)
$$

*: convex combination, if $a\gamma \leq 1$
**Definition 6.3.5.** A difference scheme for a 1D scalar conservation law (6.2.3) is **monotone**, if for all timesteps $n$

$$
\mu_j^n \leq v_j^n \quad \forall j \quad \Rightarrow \quad \mu_j^{n+1} \leq v_j^{n+1} \quad \forall j .
$$

(6.3.10) $\Rightarrow$ upwind scheme (6.3.8) for (6.2.4) monotone, if CFL-condition satisfied

**Thm. 6.2.5** $\Rightarrow$ monotone schemes mimic qualitative behavior of solutions.

**Remark:** Monotonicity is a tool to establish non-linear stability

**Lemma 6.3.6** (Criterion for monotone schemes). A timestepping scheme $\tilde{\mu}^{n+1} = \mathcal{H}(\tilde{\mu}^n)$, see (6.3.2), is monotone (→ Def. 6.3.5), if

$$
\frac{\partial \mathcal{H}_j}{\partial \mu_i} \geq 0 \quad \forall i, j .
$$

**Example:** Lax-Friedrichs method for 1D scalar conservation law (6.2.3)

$$
\Phi(u, v) = \frac{F(u) + F(v)}{2} + \frac{1}{2\gamma}(u - v) .
$$
\[ \mu_{j+1} = \mu_j - \gamma \frac{F(\mu_{j+1}^n) - F(\mu_{j-1}^n)}{2} + \frac{\mu_{j+1}^n - 2\mu_j^n + \mu_{j-1}^n}{2}. \]  
\hfill (6.3.11)

\[ \frac{\partial \mathcal{H}_j}{\partial \mu_i} = \begin{cases} \frac{1}{2} \left(1 \mp F'(\mu_i)\gamma\right), & \text{if } i = j \pm 1, \\ 0, & \text{else}. \end{cases} \]

If \(|F'(u)|\gamma \leq 1, \gamma := \tau/h, \leftrightarrow (6.3.4)\), then the Lax-Friedrichs method (6.3.11) is monotone.

**Remark.**

\[ \frac{\mu_{j+1}^n - 2\mu_j^n + \mu_{j-1}^n}{2} \approx \frac{1}{2} h^2 u''(t^n, x_j) \rightarrow \text{artificial viscosity}, \text{ see Sect. 3.1.3, (3.1.6)}. \]

**Theorem 6.3.7** ("Order barrier theorem" for monotone methods). A monotone difference scheme for (6.2.3) is at most of first order in space.

**Upwind scheme (6.2.4):** one-sided difference quotient \( \leftrightarrow \) first order approximation (\( \rightarrow \) Def. 6.3.4)

### 6.3.4  Godunov scheme

**Setting:**Scalar conservation in 1D, (6.2.3)
Idea behind upwind flux (6.3.9): follow flow of information

“Non-linear upwinding”

**Principles of Godunov’s method:**

To compute $\mu^{n+1}$ from $\mu^n$

1. solve (6.2.3) on $]t^n, t^{n+1}[ \times I$ with piecewise constant initial data $u^n$ ("local Riemann problems", see Def. 6.2.3),

$$\frac{\partial w}{\partial t}(t, x) + \frac{\partial (F \circ w)}{\partial t}(t, x) = 0 \quad \text{in } ]t^n, t^{n+1}[ \times I, \quad w(t^n, x) = u^n(x) \quad \text{in } I,$$

(with suitable boundary conditions in $]0, T[ \times \partial I$)

2. obtain $\mu^{n+1}$ by averaging, cf. (6.3.1),

$$\mu_j^{n+1} \approx \frac{1}{h} \int_{x_{j-1/2}}^{x_{j+1/2}} w(t^{n+1}, x) \, dx$$
Reformulation as difference scheme:

\[
\int_{x_{j-1/2}}^{x_{j+1/2}} w(t^{n+1}, x) \, dx - \int_{x_{j-1/2}}^{x_{j+1/2}} w(t^n, x) \, dx = \int_{t^n}^{t^{n+1}} F(w(t, x_{j-1/2})) \, dt - \int_{t^n}^{t^{n+1}} F(w(t, x_{j+1/2})) \, dt
\]

\[\mu_{j}^{n+1} - \mu_{j}^{n} = -\gamma \left\{ \frac{1}{\tau} \int_{t^n}^{t^{n+1}} F(w(t, x_{j-1/2})) \, dt - \frac{1}{\tau} \int_{t^n}^{t^{n+1}} F(w(t, x_{j+1/2})) \, dt \right\} \]  

(6.3.12)

How to evaluate time-averaged fluxes \(F(w(t, x_{j-1/2})), F(w(t, x_{j+1/2}))\)?

Recall from Sect. 6.2.2: Rankine-Hugoniot condition (6.2.11) & (6.2.10), (6.2.12):

If \(F : \mathbb{R} \mapsto \mathbb{R}\) is strictly concave (\(\rightarrow\) traffic flow, Sect. 6.1), then the Riemann problem (\(\rightarrow\) Def. 6.2.3) for (6.2.3) has the solution:

1. If \(u_l < u_r \quad \rightarrow \quad \text{discontinuous solution, shock} \rightarrow (6.2.10)\)

\[u(t, x) = \begin{cases} 
    u_l & \text{if } x < st \\
    u_r & \text{if } x > st
\end{cases}, \quad s = \frac{F(u_l) - F(u_r)}{u_l - u_r}.\]
If $u_l \geq u_r$ ➤ continuous solution, rarefaction wave \( \rightarrow (6.2.12) \)

\[
\begin{align*}
\begin{cases}
  u_l & \text{if } x < F'(u_l)t, \\
  v(x/t) & \text{if } F'(u_l) \leq x/t \leq F'(u_r), \\
  u_r & \text{if } x > F'(u_r)t,
\end{cases}
\quad v := (F')^{-1}.
\end{align*}
\]

- \( \hat{u} \) = piecewise constant function \( u^n(x) \)
- \( \hat{u} \) = shock in \((t, x)\)-plane
- \( \hat{u} \) = rarefaction wave in \((t, x)\)-plane

\[
2\gamma |F'(u)| \leq 1, \min u^n \leq u \leq \max u^n \quad (6.3.13) \quad \Rightarrow \text{no interaction of adjacent shocks/rarefaction waves!}
\]

CFL(-type)-condition \((6.3.13)\) \(\Rightarrow\) Godunov scheme = explicit 3-point difference scheme in conservation form with numerical flux, \textit{cf.} \((6.3.12)\),
\[ \Phi(u, v) = \frac{1}{\tau} \int_0^\tau F(w(t, 0)) \, dt, \quad u, v \in \mathbb{R}, \text{ where } w(t, x) \text{ is entropy solution of} \]

\[ \frac{\partial w}{\partial t}(t, x) + \frac{\partial (F \circ w)}{\partial x}(t, x) = 0 \quad \text{in } ]0, \tau[ \times \mathbb{R}, \quad w(0, x) = \begin{cases} u, & \text{if } x < 0, \\ v, & \text{if } x > 0. \end{cases} \]

\[ (6.3.15) \]

\[ (6.2.10) \text{ & } (6.2.12) \quad \blacktriangleright \quad \text{explicit formulas: for strictly concave flux } F \in C^1(\mathbb{R}): \]

\[ w(t, x) = u \quad \blacktriangleright \quad \text{shock } (6.2.10) \]

\[ u < v \rightarrow \text{ shock speed } (6.2.11): \quad s = \frac{F(u) - F(v)}{u - v}. \]

\[ \blacktriangleright \quad w(t, 0) = \begin{cases} u, & \text{if } s > 0, \\ v, & \text{if } s < 0. \end{cases} \]

\[ \blacktriangleright \quad \Phi(u, v) = \begin{cases} F(v), & \text{if } F(u) > F(v) \\ F(u), & \text{if } F(v) > F(u) \end{cases} = \min\{F(u), F(v)\} = \min\{F(\xi), u \leq \xi \leq v\}. \]
\[ u > v \quad \text{→ rarefaction wave } (6.2.12) \]

\[
\begin{align*}
\text{speed of left edge:} & \quad s = F'(u), \\
\text{speed of right edge:} & \quad s = F'(v).
\end{align*}
\]

\[ w(t, 0) = \left\{ \begin{array}{ll}
F(u), & \text{if } F'(u) \geq 0, \\
F(z), & \text{if } F'(u) < 0 < F'(v), \\
F(v), & \text{if } F'(v) \leq 0,
\end{array} \right. \]

where \( F'(z) = 0 \) (\( z = \text{position of maximum} \)).

Since \( F \) strictly concave,

\[
\Phi(u, v) = \left\{ \begin{array}{ll}
F(u), & \text{if } F'(u) > 0, \\
F(z), & \text{if } F'(u) < 0 < F'(v), \\
F(v), & \text{if } F'(v) \leq 0
\end{array} \right.
\]

\[ = \max\{F(\xi), u \leq \xi \leq v\}. \]
Numerical flux for Godunov scheme ($F$ concave):

$$
\Phi(u, v) = \begin{cases} 
\min\{F(\xi), u \leq \xi \leq v\}, & \text{if } u < v, \\
\max\{F(\xi), u \leq \xi \leq v\}, & \text{if } u \geq v.
\end{cases}
$$

(6.3.16)

(If $F$ convex → swap min ↔ max.)

**Remark.** Upwind numerical flux (6.3.9) = Godunov numerical flux (6.3.16) for $F(u) = au$

### 6.3.5 Shock resolution

Example: monotone Lax-Friedrichs scheme (6.3.11), expressed by means of centered difference quotients

$$
\mu_{j+1}^{n+1} = \mu_j^{n} - \gamma \frac{F(\mu_{j+1}^{n}) - F(\mu_{j-1}^{n})}{2} + \frac{\mu_{j+1}^{n} - 2\mu_j^{n} + \mu_{j-1}^{n}}{2} ,
$$

$$
\Phi
$$

$$
\frac{\mu_{j+1}^{n+1} - \mu_j^{n}}{\tau} + \frac{F(\mu_{j+1}^{n}) - F(\mu_{j-1}^{n})}{2h} = h \frac{1}{2\gamma} \frac{\mu_{j+1}^{n} - 2\mu_j^{n} + \mu_{j-1}^{n}}{h^2} .
$$

p. 394
Lax-Friedrichs scheme (6.3.11) provides second-order spatial approximation of

\[ \frac{\partial u}{\partial t}(t, x) + \frac{\partial (F \circ u)}{\partial x} - \frac{1}{2}h \frac{\partial^2 u}{\partial x^2} = 0. \]

**Definition 6.3.8 (Modified equation).** Let a difference scheme be consistent with (6.2.3) of order \( p, \ p \in \mathbb{N}, \) in space (\( \rightarrow \) Def. 6.3.4). A partial differential equation is called a modified equation for the scheme, if the latter is consistent with it of order \( p + 1 \) is space.

Extra diffusion term in modified equation for (6.3.11) (\( \rightarrow \) artificial viscosity, Sect. 3.1.3)

**Experiment 6.3.1.** Linear scalar advection (6.2.4) with \( a = 1, \ \tilde{\Omega} = [0, 1[, \) periodic boundary conditions \( u(t, -1) = u(t, 1) \) for all \( 0 \leq t \leq 1. \)

Discontinuous initial data:

\[ u_0(x) = \begin{cases} 1 & \text{if } -1/2 < x < 1/2, \\ 0 & \text{elsewhere}. \end{cases} \]

Recorded: \( \bar{\mu}^n \) for Lax-Friedrichs scheme, \( h = 1/100, \ \tau \in \{1/100, 1/200, 1/400\}. \)
Lax-Friedrichs solution $\mu_j^n$, $j = -100, \ldots, 100$, $n = 0, \ldots, 400$, $\tau = 1/400$
Lax-Friedrichs solution $\mu^n_j, j = -100, \ldots, 100, n = 0, \ldots, 400, \tau = 1/200$

Shock smearing caused by (excessive) artificial viscosity
Lax-Friedrichs solution $\mu^n_j, j = -100, \ldots, 100, n = 0, \ldots, 400, \tau = 1/100$

CFL-limit for $\tau : h$: “magic time step”
Example:

\[ \Phi(u, v) = F\left(\frac{\gamma}{2}(F(u) - F(v)) + \frac{u + v}{2}\right). \]

Residual term in consistency error analysis (\(u\) smooth solution of (6.2.3)):

\[ \text{residual} = O(\tau^3) + \tau \frac{h^3 \partial^3 u}{\tau \partial x^3}(t_0, x_0) + O(h^4). \]

Lax-Wendroff scheme second-order in space, first order in time

with modified equation (\(\rightarrow\) Def. 6.3.8)

\[ \frac{\partial u}{\partial t}(t, x) + \frac{\partial (F \circ u)}{\partial x}(t, x) - \frac{1}{6} \frac{h^3 \partial^3 u}{\tau \partial x^3}(t, x) = 0. \]

\[ \frac{\partial^3 u}{\partial x^3} = \text{dispersive term}. \]

Experiment 6.3.2. linear scalar advection (6.2.4) with \(a = 1\), \(\Omega = ]0, 1[ \times ]-1, 1[\), periodic boundary conditions \(u(t, -1) = u(t, 1)\) for all \(0 \leq t \leq 1\).

Discontinuout initial data: \(u_0(x) = \begin{cases} 1 & \text{if } -1/2 < x < 1/2, \\ 0 & \text{elsewhere}. \end{cases} \)

Recorded: \(\tilde{\mu}^n\) for Lax-Wendroff scheme, \(h = 1/100\), \(\tau \in \{1/100, 1/200, 1/400\}\).
Lax-Wendroff solution $\mu_j^n, j = -100, \ldots, 100, n = 0, \ldots, 400, \tau = 1/400$
Lax-Wendroff solution $\mu_j^n, j = -100, \ldots, 100, n = 0, \ldots, 400, \tau = 1/200$

Artificial dispersion  ➤  spurious oscillations at shocks
(non-monotone scheme, cf. Thm. 6.3.7)
6.4 Limiting
Adaptive Finite Element Schemes

Focus: Finite element Galerkin discretization of 2nd-order linear elliptic BVPs.

- Scalar elliptic boundary value problems (heat conduction): Sect. 1.1
- Variational formulation: Sect. 1.6
- Some Sobolev spaces: Sect. 1.7
- Abstract Galerkin discretization: Sect 2.1.1
- Lagrangian finite elements: Sects. 2.1.4, 2.1.5, 2.1.7

We assume: **quasi-optimality** of Galerkin solution $u_N \in V_N$

$$\exists C = C(\text{problem parameters}) > 0: \quad \|u - u_N\|_V \leq C \inf_{v_N \in V_N} \|u - v_N\|_V.$$  

$(u \in V = \text{exact solution of variational problem})$
\( V_N \) adjusted to \( u \) \( \Rightarrow \) adaptive Galerkin scheme

- \( V_N \) adapted to \( u \), but fixed
- \( V_N \) built based on information gleaned from preliminary solutions during computation

\( \Rightarrow \) a priori adaptive approach

\( \Rightarrow \) a posteriori adaptive approach

Example: For linear variational problem (1.7.1) & (2.1.1): if \( V_N = \text{Span} \{u\} \Rightarrow u_N = u \)

Adaptivity can dramatically improve accuracy and efficiency (dim \( V_N \) small)

Adaptive FE schemes marked by:
- non-uniform size of elements
- non-uniform polynomial degrees of local trial spaces
- h-adaptivity & p-adaptivity & hp-adaptivity

### 7.1 A priori adapted meshes

Sect. 2.5.6: singular function decomposition (2.5.9) of solution \( u \) of scalar elliptic BVP near corners of \( \Omega \), Thm. 2.5.19.
2D, Ω polygon, corners $c^i$ with angles $\omega_i, i = 1, \ldots, J$:

$$u \in H^1_0(\Omega): \quad \Delta u = f \in H^1(\Omega).$$

$$u = u^0 + \sum_{i=1}^{J} \psi(r_i) \sum_{\lambda_{ik} < l+1} \kappa_{ik} s_{ik}(r_i, \phi_i), \kappa_{ik} \in \mathbb{R},$$

(2.5.9)

regular part $u^0 \in H^{l+2}(\Omega)$, cut-off functions $\psi \in C^\infty(\mathbb{R}^+)$ ($\psi \equiv 1$ in a neighborhood of 0), corner singular functions: with $\lambda_{ik} = \frac{k\pi}{\omega_i}$

$$\lambda_{ik} \notin \mathbb{N}: \quad s_{ik}(r, \phi) = r^{\lambda_{ik}} \sin(\lambda_{ik}\phi),$$

$$\lambda_{ik} \in \mathbb{N}: \quad s_{ik}(r, \phi) = r^{\lambda_{ik}}(\ln r)(\sin(\lambda_{ik}\phi) + \phi \cos(\lambda_{ik}\phi)).$$

(2.5.10)

(Singular) radial behavior of $u$ close to corners is known (qualitatively)

Example: 1D, $\Omega = ]0, 1[$, $u(x) = x^\lambda$, $\lambda > 1/2 \Rightarrow u \in H^1(\Omega)$,

equidistant mesh $\mathcal{M} = \{ih: i = 1, \ldots, N\}, N \in \mathbb{N}, \quad h := h_{\mathcal{M}} = N^{-1},$

$l_1 : C^0([0, 1]) \mapsto \delta^0_1(\mathcal{M})$ piecewise linear interpolation.
$H^1([0, 1])$-seminorm of interpolation error $u - l_1u$, $u(x) = x^\lambda$ on equidistant grid.
Algebraic convergence rate of error of linear interpolation for $u(x) = x^\lambda$ on equidistant mesh, $N = 1, \ldots, 1000$, linear fitting of slope.
Deterioration of rate of algebraic convergence for $\lambda \leq \frac{3}{2}$.
(Recall: poor convergence of FE solutions $\in S^0_1(\mathcal{M})$ on “L-shaped” domain, Sect. 2.5.7)

Rigorous estimates:

Use local interpolation error estimates from Sect. 2.5.5, Thm. 2.5.16:

$[a, b] \subset [0, 1]$, $u \in C^2([0, 1])$, $l(x) = \frac{b-x}{b-a} u(a) + \frac{x-a}{b-a} u(b)$

$|u - l|_{1, [a, b]} \leq C(b - a) |u|_{2, [a, b]}$,
with universal constant $C > 0$.

\[
|u - l_1 u|^2_{1,0,1} = |u - l_1 u|^2_{1,0,h} + \sum_{k=1}^{N-1} |u - l_1 u|^2_{1,kh,(k+1)h} \\
\leq \int_0^h |x^\lambda - h^\lambda - 1|^2 \, dx + C^2 \sum_{k=1}^{N-1} h^2 |u|^2_{2,kh,(k+1)h} \\
= \frac{(\lambda - 1)^2}{2\lambda - 1} h^{2\lambda - 1} + C^2 h^2 \lambda^2 (\lambda - 1)^2 \int_h^1 x^{2\lambda - 4} \, dx \\
= \frac{(\lambda - 1)^2}{2\lambda - 1} h^{2\lambda - 1} + C^2 h^2 \lambda^2 (\lambda - 1)^2 \begin{cases} 
\frac{1}{2\lambda - 3} (1 - h^{2\lambda - 3}) & \text{for } \lambda > \frac{3}{2} , \\
\frac{1}{3 - 2\lambda} (h^{2\lambda - 3} - 1) & \text{for } \lambda < \frac{3}{2} , \\
\log(h) & \text{for } \lambda = \frac{3}{2} , 
\end{cases}
\]

\[
\leq C(\lambda) \begin{cases} 
h^2 & \text{for } \lambda > \frac{3}{2} , \\
h^2 \log(h) & \text{for } \lambda = \frac{3}{2} , \\
h^{2\lambda - 1} & \text{for } \lambda < \frac{3}{2} .
\end{cases}
\]

Idea: ● use non-uniform meshes with $N$ nodes
● aim for “Equidistribution of interpolation error”

Bad luck: untractable optimization problem!
Examine classes of non-uniform meshes, generated by parameterized grading functions.

Grading function: 

\[ g \in C^1([0, 1]), \quad g(0) = 1, \quad g(1) = 1, \quad g \text{ monotone} \]

\[ \Rightarrow \text{ graded mesh } \quad \mathcal{M} := \{g(i/N) : i = 0, \ldots, N\} \quad \text{of } [0, 1]. \]
Heuristics: "$H^1([0, 1])$-optimal" grading function $g \in C^1([0, 1])$, for piecewise linear interpolation of $u \in C^2([0, 1])$

- Aim for equidistribution of local interpolation errors:

$$|u - 1_1 u|^2_{1, [\tau^N_{k-1}, \tau^N_k]} \approx \text{const.} \quad \forall k = 1, \ldots, N, \quad N \in \mathbb{N}, \quad \tau^N_k := g(k/N).$$

Thm. 2.5.16 \Rightarrow |u - 1_1 u|^2_{1, [\tau^N_{k-1}, \tau^N_k]} \approx h^2_k |u|^2_{2, [\tau^N_{k-1}, \tau^N_k]} = h^3_k |u''(\xi)|^2 \quad \text{for some } \xi_k \in ]\tau^N_{k-1}, \tau^N_k[ .

For $N \gg 1$:

$$\begin{cases}
|u - 1_1 u|^2_{1, [\tau^N_{k-1}, \tau^N_k]} \approx h^3_k |u''(\tau^N_k)|^2 , \\
h_k := \tau^N_k - \tau^N_{k-1} = g(k/N) - g(k-1/N) \approx N^{-1} g'(k/N) .
\end{cases}$$

Goal: $N^{-3} (g'(k/N))^3 |u''(g(k/N))|^2 = \text{const.} \quad \forall k = 1, \ldots, N .

- Ordinary differential equation for grading function

$$g' |u'' \circ g|^{2/3} = \text{const.} , \quad g(0) = 0 , \quad g(1) = 1 . \quad (7.1.1)$$

(Meaningful only, if $u$ strictly convex/concave !)
Example:

\[ u(x) = x^\lambda, \lambda > 1/2, \lambda \neq 1 \quad \Rightarrow \quad u''(x) = \lambda(\lambda - 1)x^{\lambda-2} \]

\[ (7.1.1) \quad \Rightarrow \quad g'(\xi) g(\xi)^{2/3\lambda - 4/3} = \text{const.} \quad \Rightarrow \quad g(\xi) = \xi^{2\lambda - \frac{3}{\lambda - 1}}, \ 0 \leq \xi \leq 1. \]

**Definition 7.1.1** (Algebraically graded meshes in 1D). A graded mesh of \([0, 1]\) generated by a grading function \(g : [0, 1] \mapsto [0, 1]\) is called algebraically graded with grading factor \(\beta > 0\), if \(g(\xi) = \xi^\beta\).

Linear interpolation of \(u(x) = x^\lambda\) on algebraically graded meshes with \(\beta > 1\):

Write \(\tau_k^N = (k/N)^\beta\), ie. \(\tau_k^N = g(k/N), k = 0, \ldots, N, N \in \mathbb{N}\).

Local meshwidth from mean value theorem:

\[ g'(\xi) = \beta \xi^{\beta - 1} \quad \Rightarrow \quad \frac{\beta}{N} \left( \frac{k}{N} \right)^{\beta - 1} \leq \tau_{k+1}^N - \tau_k^N \leq \frac{\beta}{N} \left( \frac{k + 1}{N} \right)^{\beta - 1}, \ k = 1, \ldots, N - 1. \quad (7.1.2) \]
\[ I_1 = \text{linear interpolation operator on algebraically graded mesh } \mathcal{M} := \{(i/N)\beta\}, \beta > 1 \]

\[
|u - I_1 u|_{1, [0, 1]}^2 \leq \int_0^{N-\beta} |\lambda x^\lambda - 1 - N^{-\beta}|^2 \, dx + C\lambda^2(1 - \lambda)^2 \sum_{k=1}^{N-1} (\tau_{k+1}^N - \tau_k^N)^2 \int \frac{(k+1)^\beta}{(k/N)^\beta} x^{2\lambda-4} \, dx
\]

\[
\leq \frac{(\lambda - 1)^2}{2\lambda - 1} N^{-\beta(2\lambda - 1)} + C\lambda^2(1 - \lambda)^2 \sum_{k=1}^{N-1} \frac{\beta^2}{N^2} \left( \frac{k + 1}{k} \right)^{2\beta - 2} \int \frac{(k+1)^\beta}{(k/N)^\beta} x^{2(1-1/\beta)\lambda - 4} \, dx
\]

\[
\leq \frac{(\lambda - 1)^2}{2\lambda - 1} N^{-\beta(2\lambda - 1)} + C\lambda^2(1 - \lambda)^2 \beta^2 \sum_{k=1}^{N-1} \frac{2^{2\beta - 2}}{N^2} \int \frac{(k+1)^\beta}{(k/N)^\beta} x^{2(\lambda - 1/\beta) - 2} \, dx
\]

\[
\leq \frac{(\lambda - 1)^2}{2\lambda - 1} N^{-\beta(2\lambda - 1)} + C\lambda^2(1 - \lambda)^2 \beta^2 \sum_{k=1}^{N-1} \frac{2^{2\beta - 2}}{N^2} \int \frac{1}{(k/N)^\beta} x^{2(\lambda - 1/\beta) - 2} \, dx
\]

\[
\leq C_1 N^{-\beta(2\lambda - 1)} + C_2 N^{-2} \leq C' N^{-2} , \text{ if } \beta > \frac{2}{2\lambda - 1} , C' = C'(\beta, \lambda) .
\]
\[(7.1.2) \quad \Rightarrow \quad \tau_{k+1}^N - \tau_k^N \leq \frac{\beta}{N} \left(\frac{k+1}{k}\right)^{\beta-1} \left[(k/N)^\beta\right]^{(1-1/\beta)}.
\]

Optimal algebraic convergence $|u - I_1u|_{1,\,\,0,\,\,1} = O(N^{-1})$ on algebraically graded mesh with grading factor $\beta \geq \frac{2}{2\lambda - 1}$!
$H^1([0, 1])-\text{seminorm of interpolation error } u - I_1 u, u(x) = x^\lambda$, on algebraically graded mesh with grading factor $\beta = \frac{2}{2\lambda - 1} + \frac{1}{2}$. 
Algebraic convergence rate of linear interpolation error for $u(x) = x^\lambda$ on algebraically graded mesh with $\beta = \frac{2}{2\lambda - 1} + \frac{1}{2}$, linear fitting of slope.
Remark: meshwidth of algebraically graded mesh $\mathcal{M}$ for $\beta > 1$ (→ Def. 7.1.1)

\[(7.1.2) \quad \Rightarrow \quad h_{\mathcal{M}} \leq \frac{\beta}{N} .\]

Triangular algebraically graded meshes in 2D:

Construction on unit triangle, $N = 5$, $\beta = 3/2 \Rightarrow \tau_j^5 = (j/5)^\beta$

1. Algebraically graded mesh on axes define layers $L_1, \ldots, L_4$.

2. Triangulate layer $L_i$ with “nice” triangles of size $\tau_{i+1}^5 - \tau_i^5$. 
Application in 2D: resolution of corner singular functions (→ Thm. 2.5.19)
Theorem 7.1.2. Let $\Omega = \text{convex}\{(0,0), (1,0), (0,1)\}$ and $s(r, \theta) = r^\lambda \Theta(\theta)$ with $\lambda > 0$ and $\Theta \in C^\infty([0, \pi/2])$ for $(r, \theta) \in \Omega$. Let $\mathcal{M}$ be an algebraically graded mesh of $\Omega$ with grading factor $\beta > \max\{\lambda^{-1}, 1\}$ (see above).

Then, with $N := \dim \delta_1^0(\mathcal{M})$,

$$\min_{v_N \in \delta_1^0(\mathcal{M})} |s - v_N|_1 \leq C N^{-1/2},$$

where $C = C(\lambda, \beta) > 0$, i.e. for $\beta > \lambda^{-1}$ the optimal asymptotic convergence rate is recovered.

Model problem:

$$-\Delta u = f \text{ in polygon } \Omega \subset \mathbb{R}^2, \; f \in C^\infty(\overline{\Omega}), \text{ homogeneous Dirichlet b.c.}$$

Step I: Fix target meshwidth $h > 0$.

Step II: Choose grading factor $\beta > \max\{\frac{\pi}{\omega_i}, 1\}$ for each corner $c^i$ (with angle $\omega_i$) of $\Omega$. 
Step III:

Split zone around $\mathbf{c}^j$ into triangles & triangles that match target meshwidth in outer layer.

*Example 7.1.1 (Linear Lagrangian finite elements on graded meshes).*

- L-shaped domain $\Omega = \mathbb{R} \setminus (\mathbb{R}^2 \setminus ([0, 1] \times [-1, 0]))$
- Dirichlet problem for $-\Delta u = 0$, Dirichlet data matching exact solution $u(r, \varphi) = r^{2/3} \sin(2/3\varphi)$ (in polar coordinates), load vector computed by means of numerical quadrature of order 8 (→ Sect. 2.2.6)
Linear Lagrangian finite elements (→ Sect. 2.1.4) on algebraically graded triangular mesh with grading factor $\beta \geq 1$

**Fig. 62**
Apriori adapted mesh ($\beta = 1.5, N = 10$)

- # Vertices: 341
- # Elements: 600
- # Edges: 940

**Fig. 63**
Apriori adapted mesh ($\beta = 3, N = 10$)

- # Vertices: 341
- # Elements: 600
- # Edges: 940

Grading factor $\beta = \frac{3}{2}$

Grading factor $\beta = 3$
Monitored: • Discretization error $|u - u_N|_1$ as function of $N := \dim \mathcal{S}_1^0(\mathcal{M})$
• $N$-asymptotics of discretization error: rate of algebraic convergence.

(Error computations based on quadrature rule of order 8)

Convergence rate for energy error

Relationship between beta and convergence rate

7.2 A posteriori error estimation

Question: How can you tell that a finite element Galerkin solution $u_N$ is accurate enough?
No clue from asymptotic a priori error estimates of Sect. 2.5! (Constants remain elusive!)

Use $u_N$ itself to guess some norm of $u - u_N$ (a posteriori).

Setting: Linear variational problem (1.7.1) & its Galerkin discretization ($\rightarrow$ Sect. 2.1.1):

$(V_N \subset V = \text{discrete trial/test space}, \ dim \ V_N < \infty)$

$$u \in V: \quad a(u, v) = f(v) \quad \forall v \in V$$

$$u \in V_N: \quad a(u_N, v_N) = f(v_N) \quad \forall v_N \in V_N.$$

Assumptions: $a(\cdot, \cdot)$ continuous ($\rightarrow$ Def. 1.7.6) and V-elliptic ($\rightarrow$ Def. 1.7.7)

Model problem: Galerkin FE-discretization of elliptic boundary value problem, Sect. 2.1.1, 2.1.3.
Definition 7.2.1 (A posteriori error estimator).

A posteriori error estimator = mapping $\eta : V_N \mapsto \mathbb{R}$

- An a posteriori error estimator $\eta$ is called **reliable**, if

  \[ \exists C = C(\text{problem parameters} | \text{mesh parameters} | \text{FE type}): \quad \| u - u_N \|_V \leq C \eta + \text{hot}. \]  

  \[ \eta \text{ is asymptotically exact for a sequence } \{V_N\}_{N \in \mathbb{N}} \text{ of trial/test spaces, if the constant in (7.2.1) } \to 1 \text{ for } N \to \infty. \]  

- An a posteriori error estimator $\eta$ is called **efficient**, if

  \[ \exists C = C(\text{problem parameters} | \text{mesh parameters} | \text{FE type}): \quad \| u - u_N \|_V \geq C \eta + \text{hot}. \]  

  \[ \text{[hot = "higher order terms", tending to 0 faster (in terms of meshwidth) than } \| u - u_N \|_V] \]

Important: Quantitative information about constant $C$ in (7.2.1) must be available.

A posteriori error estimator is the answer to the question!

Computation of $\eta$ must not cost more than computation of $u_N$
7.2.1 Residual error estimators

weak residual = linear form $r : V \mapsto \mathbb{R}$: $r(v) := f(v) - a(u_N, v)$ \quad $v \in V$.

error equation: $a(\varepsilon, v) = r(v)$ \quad $\forall v \in V$, \quad $\varepsilon := u - u_N \in V$. \quad (7.2.3)

$[a(\cdot, \cdot) \text{ } V\text{-elliptic } & \text{ continuous}]$

\[ \gamma \|\varepsilon\|_V \leq \sup_{v \in V \setminus \{0\}} \frac{|a(\varepsilon, v)|}{\|v\|_V} = \sup_{v \in V \setminus \{0\}} \frac{|r(v)|}{\|v\|_V} =: \|r\|_{V^*} \leq C_A \|\varepsilon\|_V. \quad \text{(7.2.4)} \]

The dual norm $\|r\|_{V^*}$ of the weak residual functional provides an upper and lower bound for the $V$-norm of the Galerkin discretization error.

1. By Galerkin orthogonality (2.1.2)

\[ r(v) = r(v - v_N) \quad \forall v_N \in V_N \quad \Rightarrow \quad \|r\|_{V^*} = \sup_{v \in V \setminus \{0\}} \frac{|r(v - P v)|}{\|v\|_V}, \]

for a suitable local & continuous interpolation/projection operator $P : V \mapsto V_N$. 
Lower bounds from appropriate “candidate functions”

\[
\|r\|_{V^*} = \sup_{v \in V \setminus \{0\}} \frac{|r(v)|}{\|v\|_V} \geq \frac{|r(\tilde{v})|}{\|\tilde{v}\|_V} \forall \tilde{v} \in V.
\]

Model problem:

\[
u \in H^1_0(\Omega), -\Delta u = f \text{ in polygon } \Omega \subset \mathbb{R}^2, f \in L^2(\Omega), \quad V_N = \mathcal{S}_0^1(\mathcal{M}) \text{ on triangular mesh } \mathcal{M}, \text{ Galerkin solution } u_N \in V_N.
\]

\[
\Rightarrow \quad u \in H^1_0(\Omega): \quad a(u, v) := \int_{\Omega} \text{grad } u \cdot \text{grad } v \, dx = f(v) := \int_{\Omega} f v \, dx \quad \forall v \in H^1_0(\Omega).
\]

Local integration by parts (→ Thm. 1.6.1), for \( v \in H^1_0(\Omega) \)

\[
r(v - Pv) = \sum_{K \in \mathcal{M}} \int_K f(v - Pv) - \text{grad } u_N \cdot \text{grad } (v - Pv) \, dx
\]

\[
= \sum_{K \in \mathcal{M}} \left\{ \int_K f(v - Pv) \, dx - \int_{\partial K} (\text{grad } u_N \cdot n_{\partial K}) (v - Pv) \, dS \right\}
\]

\[
= \sum_{K \in \mathcal{M}} \int_K f(v - Pv) \, dx + \sum_{F \in \partial(\mathcal{M})} \int_F [\text{grad } u_N \cdot n_F]_F (v - Pv) \, dS,
\]

7.2
where \( \mathcal{E}_\Omega(\mathcal{M}) \) : set of interior edges of \( \mathcal{M} \),

\( \mathbf{n}_F \) : suitably fixed normal vector for edge \( F \),

\( [\cdot]_F \) : jump across \( F \).

Use local estimates for \( \| u - P_0 u \|_{K} \), \( \| u - P_0 u \|_{0,F} \):

\[
\eta = \left( \sum_{K \in \mathcal{M}} h_K^2 \| f \|_{0,K}^2 \right)^{1/2} + \left( \sum_{F \in \mathcal{E}_\Omega(\mathcal{M})} h_F \left\| \left[ \text{grad} \, u_N \cdot \mathbf{n}_F \right]_F \right\|_{0,F}^2 \right)^{1/2}.
\]  \hspace{1cm} (7.2.5)

**Theorem 7.2.2** (Residual based a posteriori error estimator for \( -\Delta u = f \)). With \( \underline{C}, \overline{C} > 0 \) only depending on \( \rho_{\mathcal{M}} \) & local quasi-uniformity of \( \mathcal{M} \):

\[
\underline{C} |e|_1 \leq \eta \leq \overline{C} \left( |e|_1 + \sum_{K \in \mathcal{M}} h_K \| f - \overline{f}_K \|_{0,K} \right),
\]

with local means \( \overline{f}_K := |K|^{-1} \int_K f \, dx \). \( \Rightarrow \) \( h_K \| f - \overline{f}_K \|_{0,K} = \text{hot} \rightarrow \text{exercise} \).

**Proof.** See [13], [1, Sect. 2.2].

\[
\eta = \text{element residuals} + \text{jumps of} \left( \text{grad} \, u_N \cdot \mathbf{n} \right) \text{ across interior edges}
\]
normal flux $\mathbf{j} \cdot \mathbf{n} := \text{grad} u \cdot \mathbf{n}$ continuous across lines,
BUT discrete flux $\text{grad} u_N$ has discontinuous normal components

\[ [\text{grad} u_N \cdot \mathbf{n}_F]_F \leftrightarrow \text{discretization error} \]

7.2.2 Hierarchical error estimator

Idea: Galerkin solution of error equation (7.2.3)

\[ a(e, v) = r(v) \quad \forall v \in V \]

based on refined trial/test space $\tilde{V}_N \subset V: \ V_N \subset \tilde{V}_N$.

Saturation assumption: if $\tilde{u}_N$ solves $a(\tilde{u}_N, \tilde{v}_N) = f(\tilde{v}_N) \ \forall \tilde{v}_N \in \tilde{V}_N$, then

\[ \exists \beta < 1: \ \| u - \tilde{u}_N \|_V \leq \beta \| u - u_N \|_V. \quad (7.2.6) \]
Lemma 7.2.3 (Reliability & efficiency of hierarchical a posteriori error estimator). If the energy norm $\| \cdot \|_V$ is induced by $a(\cdot, \cdot)$ (→ Sect. 1.7), and the saturation assumption (7.2.6) holds, then

$$\| \tilde{u}_N - u_N \|_V \leq \| u - u_N \|_V \leq \frac{1}{\sqrt{1 - \beta^2}} \| \tilde{u}_N - u_N \|_V .$$

Proof. By Galerkin orthogonality (2.1.2), because $V \subset \tilde{V}$

$$\| u - u_N \|_V^2 = \| u - \tilde{u}_N + \tilde{u}_N - u_N \|_V^2 = \| u - \tilde{u}_N \|_V^2 + \| \tilde{u}_N - u_N \|_V^2 \leq \beta^2 \| u - u_N \|_V^2 + \| \tilde{u}_N - u_N \|_V^2 .$$

$$\| u - u_N \|_V^2 \leq \frac{1}{1 - \beta^2} \| \tilde{u}_N - u_N \|_V^2 .$$

Get $\tilde{u}_N - u_N$ from: $a(\tilde{u}_N - u_N, \tilde{v}_N) = r(\tilde{v}_N)$. (7.2.7)

Main example: 2nd order elliptic BVP, $V_N = \delta_{1,0}^0(\mathcal{M})$, $\tilde{V}_N = \delta_{2,0}^0(\mathcal{M})$

However: solving (7.2.7) too expensive ($\dim \tilde{V}_N \gg \dim V_N$)
Lemma 7.2.4. If \( \hat{\alpha} : V \times V \rightarrow \mathbb{R} \) is another symmetric positive definite bilinear form that satisfies

\[
\exists C, \overline{C} > 0: \ C\hat{\alpha}(v, v) \leq a(v, v) \leq \overline{C}\hat{\alpha}(v, v) \quad \forall v \in V,
\]

i.e. \( \hat{\alpha} \) and \( a \) are spectrally equivalent, and

\[
e \in V: \ a(e, v) = r(v) \quad \forall v \in V \quad \text{and} \quad \hat{\epsilon} \in V: \ \hat{\alpha}(\hat{\epsilon}, v) = r(v) \quad \forall v \in V,
\]

then

\[
C^2a(e, e) \leq a(\hat{\epsilon}, \hat{\epsilon}) \leq \overline{C}^2a(e, e).
\]

Proof. By spectral equivalence & Cauchy-Schwarz inequality

\[
a(e, e) = \hat{\alpha}(\hat{\epsilon}, e) \leq \hat{\alpha}(\hat{\epsilon}, \hat{\epsilon})^{1/2}\hat{\alpha}(e, e)^{1/2} \leq \frac{1}{C}\hat{\alpha}(\hat{\epsilon}, \hat{\epsilon})^{1/2}a(e, e)^{1/2},
\]

\[
a(\hat{\epsilon}, \hat{\epsilon}) \leq \overline{C}\hat{\alpha}(\hat{\epsilon}, \hat{\epsilon}) = \overline{C}a(\hat{\epsilon}, \hat{\epsilon}) \leq \overline{C}\hat{\alpha}(\hat{\epsilon}, \hat{\epsilon})^{1/2}a(e, e)^{1/2}.
\]

Setting:

\[
a(u, v) = \int_{\Omega} \sigma \, \text{grad} \, u \cdot \text{grad} \, v \, dx, \ \sigma \ \text{satisfying (1.2.2)}.
\]
Construction of spectrally equivalent bilinear form for $V_N = \delta_{1,0}^0(\mathcal{M})$, $\tilde{V}_N = \delta_{2,0}^0(\mathcal{M})$:

direct splitting: $\tilde{V}_N \ni \tilde{v}_N = v_N^{\text{lin}} + \sum_{F \in \mathcal{E}(\mathcal{M})} \kappa_F(\tilde{v}_N) b_F$, $v_N^{\text{lin}} \in \delta_{1,0}^0(\mathcal{M})$, $\kappa_F \in \mathbb{R}$,

$b_F = \text{nodal basis function associated with interior edge } F$.

\[
\hat{a} (\tilde{w}_N, \tilde{v}_N) := a (w_N^{\text{lin}}, v_N^{\text{lin}}) + \sum_{F \in \mathcal{E}(\mathcal{M})} \kappa_F(\tilde{w}_N) \kappa_F(\tilde{v}_N) a (b_F, b_F). \tag{7.2.8}
\]

**Lemma 7.2.5.** $\hat{a}$ from (7.2.8) is spectrally equivalent to $a$ with constants only depending on $\mu, \mathcal{M}$ and $\sigma$.

**Proof.** → exercise

Replace $a$ by $\hat{a}$ in error equation (7.2.7)
\[ \tilde{a}(\tilde{w}_N, \tilde{v}_N) = a(w_{N}^{\text{lin}}, v_{N}^{\text{lin}}) + \sum_{F \in \mathcal{E}_\Omega(\mathcal{M})} \kappa_F(\tilde{w}_N) \kappa_F(\tilde{v}_N) a(b_F, b_F) = r(\tilde{v}_N) \quad \forall \tilde{v}_N \in \tilde{V}_N. \]

\[ r(v_N) = 0 \quad \forall v_N \in V_N \subset \tilde{V}_N \quad \Rightarrow \quad w_{N}^{\text{lin}} = 0. \]

\[ \eta = a(\sum_{F \in \mathcal{E}_\Omega(\mathcal{M})} \kappa_F(\tilde{w}_N) b_F, \sum_{F \in \mathcal{E}_\Omega(\mathcal{M})} \kappa_F(\tilde{w}_N) b_F)^{1/2}, \quad \kappa_F(\tilde{w}_N) = \frac{r(b_F)}{a(b_F, b_F)} \rightarrow \text{inexpensive!} \]

(7.2.9)

**Corollary 7.2.6.** Under the saturation assumption (7.2.6) \( \eta \) from (7.2.9) provides a reliable and efficient a posteriori error estimator (w.r.t. to energy norm).

### 7.2.3 Recovery based error estimators

**Setting:** 2nd order elliptic BVP in 2D, Galerkin discretization \( V_N = \mathcal{S}_1^0(\mathcal{M}) \), triangular mesh \( \mathcal{M} \)

**Notations:**
- \( \mathcal{N}(\mathcal{M}) \): nodes of \( \mathcal{M} \),
- \( b_p \): hat function associated with \( p \in \mathcal{N}(\mathcal{M}) \),
- \( \mathcal{U}_p \): elements of \( \mathcal{M} \) sharing node \( p \in \mathcal{N}(\mathcal{M}) \).
Definition 7.2.7 (Gradient recovery operator).

\[
G : \mathcal{S}_1^0(\mathcal{M}) \mapsto (\mathcal{S}_1^0(\mathcal{M}))^2 \quad , \quad (G(v_N))(p) := \frac{1}{|\mathcal{U}_p|} \int_{\mathcal{U}_p} \text{grad} \, v_N \, dx \quad , \quad p \in \mathcal{N}(\mathcal{M}) .
\]

Note: A posteriori error estimator \( \eta \) from (7.2.10) independent of problem data!

Example 7.2.1 (Comparative performance of local error estimators).
L-shaped computational domain:
\[ \Omega := \left[ -1, 1 \right]^2 \setminus \left( [0, 1] \times [-1, 0] \right) \]

Triangular meshes created by \textit{regular refinement} (\(\rightarrow\) Ex. 2.5.1) of initial mesh consisting of four triangles

Dirichlet BVP for \(-\Delta u = f\) with (singular \(\rightarrow\) Sect. 2.5.6) exact solution (polar coordinates)
\[
 u(r, \varphi) = r^{2/3} \sin(2/3\varphi), \quad r = \sqrt{x_1^2 + x_2^2}.
\]

Monitored: Ratios of estimated and “true” (\(\rightarrow\) order 8 numerical quadrature, Sect. 2.2.6) error for estimators of Sect. 7.2.1, 7.2.2, 7.2.3.

\section*{7.2.4 Goal oriented error estimators}

Setting: Linear variational problem (1.7.1) & its Galerkin discretization (\(\rightarrow\) Sect. 2.1.1):
\(V_N \subset V = \text{discrete trial/test space, dim } V_N < \infty\)
\( u \in V: \ a(u, v) = f(v) \ \forall v \in V \quad \Rightarrow \quad u \in V_N: \ a(u_N, v_N) = f(v_N) \ \forall v_N \in V_N. \)

Assumptions: \( a(\cdot, \cdot) \) continuous (\( \rightarrow \) Def. 1.7.6) and \( V \)-elliptic (\( \rightarrow \) Def. 1.7.7)

Model problem: Galerkin FE-discretization of elliptic boundary value problem, Sect. 2.1.1, 2.1.3.

(Non-linear) continuous output functional \( F : V \mapsto \mathbb{R} \quad \Rightarrow \quad \) quantity of interest \( F(u) \)

Goal: A posteriori error estimate: bound for \( F(u) - F(u_N) \)

Assumption: \( F : V \mapsto \mathbb{R} \) twice continuously (Frechet-)differentiable:
\[
F' : V \mapsto \{ \text{linear forms on } V \}, \quad F'' : V \mapsto \{ \text{symmetric bilinear forms on } V \times V \}.
\]

Recall: duality based a priori error estimates \( \rightarrow \) Sect. 2.5.9
**Theorem 7.2.8.** In first order approximation holds

\[ |F(u) - F(u_N)| \leq \eta := \inf_{v_N \in V_N} |r(u_N, g^F - v_N)| \quad \text{with} \quad r(v) := f(v) - a(u_N, v) , \]

where \( g^F \in V \) is the solution of the linearized dual problem

\[ g^F \in V: \quad a(v, g^F) = F'(u_N; v) \quad \forall v \in V . \]

**Proof.** “Taylor expansion”, error \( e := u - u_N \in V \):

\[
F(u) - F(u_N) = F'(u_N; e) - \int_0^1 D^2 F(u_N + \tau e; e, e)(\tau - 1) \, d\tau \\
= a(e, g^F) - \int_0^1 D^2 F(u_N + \tau e; e, e)(\tau - 1) \, d\tau \\
= r(g^F) + R(e) \quad \text{with} \quad |R(e)| = O(\|e\|_V^2) .
\]

Next, use Galerkin orthogonality (2.1.2) \( \square \)
Model problem:

\[ u \in H^1_0(\Omega), \ - \text{div}(\sigma \ \text{grad} \ u) = f \text{ in polygon } \Omega \subset \mathbb{R}^2, \ f \in L^2(\Omega), \]
coefficient \( \sigma \) piecewise constant,

\[ V_N = \mathcal{S}_1^0(\mathcal{M}) \text{ on triangular mesh } \mathcal{M} \text{ (that resolves jumps of } \sigma), \]

\[ \Rightarrow \text{ Galerkin solution } u_N \in V_N. \]

\[ u \in H^1_0(\Omega): \ a(u, v) := \int_\Omega \sigma \ \text{grad} \ u \cdot \text{grad} \ v \ dx = f(v) := \int_\Omega f v \ dx \ \forall v \in H^1_0(\Omega). \]

Integration by parts according to 1.6.1 \( \rightarrow \) residual error estimator, Sect. 7.2.1:

\[ \eta = \inf_{v_N \in V_N} \sum_{K \in \mathcal{M}} \int_K f(g^F - v_N) - \sigma \ \text{grad} u_N \cdot \text{grad}(g^F - v_N) \ dx \]

\[ \leq \inf_{v_N \in V_N} \sum_{K \in \mathcal{M}} \left\{ \int_K (f + \text{div}(\sigma \ \text{grad} u_N))(g^F - v_N) \ dx + \frac{1}{2} \int_{\partial K} [\text{grad} u_N \cdot n]_{\partial K} (g^F - v_N) \ dS \right\} \]

\[ \leq \inf_{v_N \in V_N} \sum_{K \in \mathcal{M}} \left( \|f + \text{div}(\sigma \ \text{grad} u_N)\|_{0,K}^2 + \frac{1}{4} h_K^{-1} \| [\text{grad} u_N \cdot n]_{\partial K} \|_{0,\partial K}^2 \right)^{1/2} \]

\[ \left( \|g^F - v_N\|_{0,K}^2 + h_K \|g^F - v_N\|_{0,\partial K}^2 \right)^{1/2} \]

\[ \leq C \sum_{K \in \mathcal{M}} \rho_K(u_N) h_K^2 \left| \frac{g^F}{g_{\partial K}} \right|_{2,K}, \]

“practical” estimator \( \eta \).
with \( C = C(\rho_M) > 0 \) and local residual

\[
\rho_K(u_N) := \| f + \text{div}(\sigma \nabla u_N) \|_{0,K} + \frac{1}{2} h_K^{-1/2} \| [\nabla u_N \cdot \mathbf{n}]_{\partial K} \|_{0,\partial K} .
\]

\( h_K^2 |g^F|_{2,K} \) = sensitivity factor (measuring significance of \( K \) for output \( F(u_N) \))

**Issue**: How to get \( g^F \) (= exact solution of an elliptic BVP !)

**Option**: [Recovery of Hessian]

- Given \( u_N \), compute FE Galerkin approximation \( g_N^F \in S_1^0(M) \) of \( g^F \).
- Pick \( K \in M \), \( \mathcal{U}_K := \bigcup \{ K' : K' \cap K \neq \emptyset \} \), and compute

\[
w_K \in \mathcal{P}_2(\mathbb{R}^2): \quad w_K = \arg \min_{q \in \mathcal{P}_2(\mathbb{R}^3)} \| g_N^F - q \|_{0,\mathcal{U}_K} .
\]

- Approximate \( |g^F|_{2,K} \approx |w_K|_{2,K} \).

Final error estimator: \( \eta = \sum_{K \in M} \rho_K(u_N) h_K^2 |w_K|_{2,K} \). \hspace{1cm} (7.2.11)
7.3 A posteriori mesh adaptation

7.3.1 Local mesh refinement

= refinement of a subset $\Sigma \subset \mathcal{M}$ of the elements of a mesh $\mathcal{M}$

Example: longest edge bisection
$\rightarrow$ local refinement of a triangular mesh in 2D

Algorithm: (Longest edge bisection)

Input: • triangular mesh of $\Omega \subset \mathbb{R}^2$,
• subset $\Sigma \subset \mathcal{M}$ of marked cells,
• marked vertex ("newest vertex") for each triangle
Bisection of triangle $K$ through vertex opposite to longest edge

while $\Sigma \neq \emptyset$ \{ Bisect any $K \in \Sigma$ through its longest edge
$\Sigma = \{\text{cells with hanging node on at least one edge}\}$
\}

Repeated longest edge bisection of a triangle will create only a small number of congruence classes of triangles!

Reference: [11]
Lemma 7.3.1. Given initial mesh with minimal angle $\alpha$, steps of newest vertex bisection will always create meshes with minimal angle $> \alpha/2$.

7.3.2 Control of local mesh refinement

Basic assumption:

Justified for elliptic boundary value problems (→ Sect. 1.4)
Assumption:  
- (efficient & reliable) a posteriori error estimator $\eta$ available,  
- $\eta$ arising from local contributions

\[ \eta^2 = \sum_{K \in \mathcal{M}} \eta_K^2 \quad (\eta_K \leftarrow \text{local error indicators}) \]  

- $\eta_K \leftarrow \text{“local error contributions”}$

Note: all a posteriori error estimators introduced above fit (7.3.1)

**Basic adaptive algorithm:**

Input:  
- coefficients and right hand side for boundary value problem,  
- initial triangulation $\mathcal{M}_1$ with good shape regularity, measure  
- tolerance $\tau > 0$

1. Set $k := 1$.
2. Compute finite element solution $u_N$ on $\mathcal{M}_k$
3. Compute the local error indicator $\eta_K$ for each $K \in \mathcal{M}_k$.
4. **IF** $\eta < \tau$ for $\eta$ according to (7.3.1), **THEN** STOP.
5. Mark all cells $\in M_k$ for which

$$\eta_K > \theta \max\{\eta_K, \ K \in M_k\} \quad \text{for some} \quad \theta \in ]0, 1[ \quad \text{e.g.} \ \theta \approx 0.9 \ . \quad (7.3.2)$$

6. One refinement step (*eg.* newest vertex bisection) with $\Sigma = \{\text{marked cells}\} \rightarrow \mathcal{M}_{k+1}$

7. $k := k + 1$ and GOTO STEP 2.

Rationale for marking strategy *(7.3.2)*: equidistribution of error

**Example 7.3.1 (Efficacy of adaptive mesh refinement).**

- L-shaped computational domain: $\Omega := ] - 1, 1[^2 \setminus ([0, 1] \times [-1, 0])$

- Triangular meshes created by either
  - regular refinement (*→* Ex. [2.5.1]), or *adaptive refinement* (*→* see above)
    of initial mesh consisting of four triangles.

- Dirichlet BVP for $-\Delta u = f$ with (singular *→* Sect. [2.5.6]) exact solution (polar coordinates)

$$u(r, \varphi) = r^{2/3} \sin(2/3\varphi) \ , \quad r = \sqrt{x_1^2 + x_2^2} \ .$$
Distribution of energy norm of error for regular refinement:

Distribution of energy norm of error for adaptively refined meshes
(Mesh adaptation controlled by a posteriori error estimators from Sect. 7.2.1, 7.2.2, 7.2.3) Both, “true error” (order 8 numerical quadrature, left) and estimated error (right):

Fig. 67

Energy Error distribution for iteration 10 (RES)

# Coordinates : 71
# Elements : 120
# Edges : 190

Err = 1.15e−01

Fig. 68

ErrEst distribution for iteration 10 (RES)

# Coordinates : 71
# Elements : 120
# Edges : 190

Eta = 3.65e−01
Residual based error estimator, Sect. [7.2.1]
Energy Error distribution for iteration 10 (HIER)

Err = 1.15e−01

# Coordinates : 71
# Elements : 120
# Edges : 190

Fig. 71

ErrEst distribution for iteration 10 (HIER)

Eta = 1.05e−01

# Coordinates : 71
# Elements : 120
# Edges : 190

Fig. 72
Hierarchical error estimator, Sect. [7.2.2]
**Energy Error distribution for iteration 10 (REC)**

Err = 1.61e−01

- Coordinates : 42
- Elements : 63
- Edges : 104

**ErrEst distribution for iteration 10 (REC)**

Eta = 2.53e−01

- Coordinates : 42
- Elements : 63
- Edges : 104

Fig. 75

Fig. 76
Example 7.3.2 (Performance of error estimators in adaptive setting).

Setting of Ex. 7.3.1

- Monitored: decrease of “true” (order 8 quadrature) and estimated error (in energy norm):
**Convergence rates for energy error**

![Graph](image)

- RES based ad. ref.
- REC based ad. ref.
- HIER based ad. ref.
- Reg. ref.

**Convergence rates for energy error and estimators**

![Graph](image)

- Err. est. RES ad. ref.
- Err. est. REC ad. ref.
- Err. est. HIER ad. ref.
- Disc. err. reg. ref.

**Fig. 79**

**Fig. 80**

Convergence rates for energy and estimators

Dofs [log]

Errors [log]

$p = -0.5$

$p = -0.3$

Errors [log]

Dofs [log]
2. Adaptive refinement: ratio of “true” and estimated errors, see Ex. 7.2.1
Index

p-hierarchical local shape functions, 81
5-point stencil, 135

a-orthogonal, 53
affine equivalent finite elements, 83
affine transformation, 81
angle conditions, 140
arithmetic finite elements, 81
assembly, 97, 122
  cell oriented, 104

barycentric coordinates, 72
barycentric dual mesh, 153
basis, 56
  change of, 59
basis functions, 64
basis representation, 56
best approximation error, 54
beta function, 476
bilinear form, 31
  continuous, 33
  elliptic, 33
  symmetric positive definite, 32
boundary conditions
  Dirichlet, 17, 43
  essential, 43, 116
  homogeneous, 18
  mixed, 46
  natural, 45
  Neumann, 17
  radiation, 17
boundary fitting, 119
  parabolic, 119
boundary value, 11
Boundary value problem, 11
boundary value problem
  elliptic, 19
Cauchy-Schwarz inequality, 39, 41
Cea’s lemma, 53
cell, 61
cell contributions, 97
CG method, 331
chain rule, 120
circumcenter, 153
classical solution, 29
coefficient vector, 58
compatibility condition, 29
complete, 32, 37
completion, 37
computational domain, 13
conforming space, 64
conformity condition, 63
conjugate gradients, 331
consistency, 145
consistency error, 146, 208, 211
consistency error estimate, 212, 213
continuity, 43
continuous, 32, 33
control volume, 149
CRS-format, 68
cubic Hermitian Finite Elements, 89
curl, 25
curvilinear mesh, 83
degenerate quadrilateral, 86
degrees of freedom
  local, 89
Delaunay triangulation, 140, 152
Differential operator, 11
Dirichlet boundary conditions, 17, 43
Dirichlet data, 42, 44
Dirichlet principle, 48
discrete maximum principle, 138
  angle conditions, 140
discrete variational problem, 52, 58
discretization, 50
discretization error, 54
distributional derivative, 24
div, 25
domain, 11
  computational, 13
dual basis, 88
dual mesh, 151
  angle condition, 152
degree, 61
element, 61
element load vector, 98
element stiffness matrix, 98
elliptic, 12
  linear scalar second order PDE, 15
elliptic boundary value problem, 19
Ellipticity, 33
Energy functional, 48
energy norm, 35, 332
equidistant mesh, 65
essential boundary conditions, 43, 116
ex:errestcmpadap, 449
extension operator, 45
face, 61
finite difference stencil, 128
finite elements
  basis functions, 64
Examples and Remarks

$L^2$-convergence of FE solutions, 236
$L^\infty$-estimate for FE solutions, 244
$L^\infty$-estimate on regular grid, 245

Adaptive mesh refinement, 443
Anisotropic material, 14
Approximation of mean temperature, 226
Arrays storing 2D triangular mesh, 102
Assembly for linear Lagrangian finite elements on triangular mes, 100
Assembly for quadratic Lagrangian FEM, 105

Boundary flux computation, 229
Comparative performance of local error estimators, 433
Convective cooling, 17
Convergence of CG for discrete 2nd order elliptic BVP, 332
Cubic Hermitian Finite Elements, 89

Extended MATLAB mesh data structure, 102
heat conduction, 18
Hierarchical basis for $s^0_{1,0}(M)$ in 1D, 317

Impact of linear boundary fitting on FE convergence, 221
Impact of numerical quadrature on finite element discretization error, 213

Linear interpolation, 173
Linear interpolation on acute and obtuse triangles, 188
Linear Lagrangian finite elements on graded meshes, 420
Local quadrature rules on quadrilaterals, 114
Local quadrature rules on triangles, 113

mesh file format, 92
Mesh file format for MATLAB code, 94

Performance of error estimators in adaptive setting, 449

Quadratic interpolation, 173
Quadratic Lagrangian finite elements, 76
Quasi-locality of solution of scalar elliptic boundary value problem, 22

Radiative cooling, 18
Scalar elliptic boundary value problem in one space dimension, 20
Smoothness of solution of scalar elliptic boundary value problem, \textsuperscript{21}
Sparse stiffness matrices, \textsuperscript{67}
Spectral condition numbers of stiffness matrices, \textsuperscript{385}
Supports of global shape functions in 1D, \textsuperscript{64}
Supports of global shape functions on triangular mesh, \textsuperscript{65}
Triangle as degenerate quadrilateral, \textsuperscript{86}
Variational formulation of pure Dirichlet problem for heat equation, \textsuperscript{27}
Variational formulation: heat conduction with general radiation boundary conditions, \textsuperscript{28}
Weak derivative in 1D, \textsuperscript{24}
weak differential operators, \textsuperscript{25}
Definitions

Affine transformation, 81
barycentric coordinates, 72
conformity, 64
congruent matrices, 60
Continuity of bilinear forms, 33
Continuity of linear forms, 32
discrete maximum principle, 139
distributional derivative, 24
dual basis, 88
element load vector, 98
element stiffness matrix, 98
Ellipticity of bilinear forms, 33
Galerkin projection, 55
Higher order Lagrangian finite element spaces, 75
Hilbert space, 32
inner product, 32
linear operator, 55
local degrees of freedom, 89
mesh, 61
norm, 31
positive definite matrix, 60
projection, 55
pullback, 82
Second Poincaré-Friedrichs inequality, 40
sparse matrix, 66
structurally symmetric matrix, 141
tensor product polynomials, 87
weak derivative, 24
MATLAB-CODE-Codes

assemMat_QFE, 106
sparse (MATLAB-function), 106
add_Edge2Elem, 102
add_Edges, 102
init_Mesh, 102
Loading a mesh from file, 95
List of Symbols

$C^\infty(\Omega)$, 40
$\mathbb{C}$, 40
$V_{N,0}$, 118
$P_p(K)$, 63
$P_p(\mathbb{R}^d)$, 75
$Q_p(\mathbb{R}^d)$, 87
$\Delta$, 116
$\text{div}$, 25
$\mathcal{E}(\mathcal{M})$, 103
$H^{1/2}(\partial \Omega)$, 44
$\delta_1^0(\mathcal{M})$, 69
$I_1$, 155
$\delta_p^0(\mathcal{M})$, 75
$L^2(\Omega)$, 37
$\| \cdot \|_{1/2, \partial \Omega}$, 44
$\| \cdot \|_0$, 37
$\| \cdot \|_0^2$, 36
$\| \cdot \|_\infty$, 144
$\mathcal{N}(\mathcal{M})$, 61
$\Omega$, 111

$\Phi^*$, 82
$A, B, C, \ldots$ (matrices), 58
$n$, 17
$x^\alpha$, $\alpha = \text{multiindex}$, 75
$H^1_{1,p}(\Omega)$, 46
$\delta_{ij}$, 73
$\frac{\partial u}{\partial x_j}$, 24
$\gamma$, 45
$\text{grad}$, 25
$\lambda_i$, 77
$H^1(\Omega)$, 40
$\delta^\theta$,$_{1,0}(\mathcal{M})$, 74
$diam(\Omega)$, 36
$\vec{\mu}, \vec{\phi}, \vec{\xi}, \ldots$ (coefficient vectors), 58
$\hat{K}$, 82
$H^1_0(\Omega)$, 37
Bibliography


Appendix
Essential skills

This chapter lists essential skills that you possess after having studied the individual chapters of the course.

A.1 Chapter 1: Second-order scalar elliptic boundary value problems

You should know

- the concept of a second-order scalar elliptic boundary value problem together with appropriate boundary conditions (Dirichlet, Neumann, radiation, mixed).
- the concept of uniform positivity of the conductivity/coefficient of a second-order scalar differential operator and its consequences for the associated bilinear form.
● the definition of weak derivatives and the rationale for introducing them
● to derive the complete variational form (bilinear form, right hand side functional and Hilbert space framework) for any linear second-order scalar elliptic boundary value problem.
● fundamental notions like ellipticity and continuity (of linear/bilinear forms)
● that functions in $H^1(\Omega)$ can be unbounded for $d > 1$.
● the Lax-Milgram lemma and how to prove the ellipticity/continuity of bilinear forms arising from linear second-order scalar elliptic boundary value problems (Poincaré-Friedrichs inequalities)
● the compatibility condition for the pure Neumann problem.
● the trace theorem from $H^1(\Omega)$ and its significance for admissible Dirichlet and Neumann boundary data.
● how to tell invalid and valid source terms and boundary data from each other.
● how to express the variational form of a linear second-order scalar elliptic boundary value problem as an equivalent minimization problem.

A.2 Chapter 2: The Finite Element Method (FEM)

You should know
- the idea of the Galerkin approximation, [Galerkin-orthogonality (2.1.2)], and [Cea's Lemma Thm. 2.1.1] (including the proof)
- how to derive a linear system of equations from linear variational problems
- the terms stiffness matrix and load vector
- the impact of a change of bases on the stiffness matrix and the Galerkin solution
- the concept of a mesh
- that FE functions for $H^1(\Omega)$ have to be continuous
- that FE spaces possess bases of locally supported functions associated with vertices/edges/cells
- the rationale behind the use of locally supported basis functions in FEM
- simplicial and quadrilateral Lagrangian FE (their local polynomial spaces and interpolation points)
- the concept of parametric (particularly affine equivalent) Lagrangian FE
- the concept of local assembly for the efficient computation of the stiffness matrix and load vector
- the use of parametric FE for the approximation of curved boundaries
- how to use numerical quadrature to approximate the coefficients of the stiffness matrix and load vector
- how to deal with non-homogeneous Dirichlet boundary conditions
- the notation of difference stencils
the discrete maximum principle and its consequences for the discrete solution

the idea behind finite volume methods and the construction of dual meshes

different types of refinement and convergence and how to tell them from raw error data

what the [Bramble-Hilbert lemma Thm. 2.5.7](#) does tell

the concept of shape regularity of simplicial meshes and its role in the transformation estimates (Lemma 2.5.9) for norms

that acute angles do not affect accuracy of FE Galerkin solutions but obtuse angles are harmful

the notion 2-regularity of a 2nd order elliptic boundary value problem

what happens in case of reentrant corners to solutions of 2nd order elliptic boundary value problem

the impact of numerical quadrature on the convergence rate of Lagrangian FE

the convergence rates of Lagrangian FE in the energy and the $L^2$-norm in case of $h$-refinement

that you can gain up to twice the convergence rate in the energy norm for the evaluation of $H^1(\Omega)$-continuous linear functionals

Practical: Implementing Lagrangian FE for 2nd order boundary value problems in 2D using the MATLAB environment of the exercises
You should know

- what a singular perturbed problem is
- what special phenomena are encountered in the case of convection-diffusion problems
- the idea behind upwinding and streamline diffusion
- the result of quasi optimality of Galerkin solutions and the notion dispersion in the case of the Helmholtz equation
- that saddle point problems lead to mixed formulations
- the notions continuous and discrete inf-sup condition and their importance for the discretization of saddle point problems
- the Stokes equation and some stable pairs for its FE discretization
- what consistent iteration does mean
- the principle underlying a fix point iteration
- how to derive Newton’s method for non-linear elliptic boundary value problems
A.4 Chapter 4: Solving discrete boundary value problems

You should know

- what the idea behind successive subspace correction is
- the terms iteration matrix, contraction number, and rate of convergence of linear stationary iterative methods
- how the hierarchical basis is defined
- the idea behind multigrid methods
- what the idea behind the cg- and the pcg-method is
- that the condition number of the stiffness matrix to 2nd order elliptic boundary value problems grows like $h^{-2}M$
- that increasing condition numbers of the iteration matrix generally slow down the convergence of linear stationary iterative methods
A.5  Chapter 5: Parabolic Boundary Value Problems

You should know

- how a 2nd order parabolic initial BVP looks like
- what the method of lines is
- what a stable single step method is, particularly the notion $L(\pi)$-stability
- why implicit timestepping schemes have to be used
- that spatial and temporal errors enter the a priori estimates in an additive fashion

A.6  Chapter 6: Numerical Methods for Conservation Laws

You should know

- how (nonlinear) conservation laws look like
• the concept of characteristics and their importance
• that classical solutions make no sense in case of shocks
• how to derive physically meaningful weak solutions
• the solution of a Riemann problem
• what a method in conservation form is
• what the idea behind the Godunov scheme is and its properties
• what the Lax-Friedrichs and Lax-Wendroff schemes are
• monotone schemes and the order barrier theorem
• the continuous and numerical region of dependence and their consequences (→ CFL-condition)

A.7 Chapter 7: Adaptive Finite Element Schemes

You should know

• the idea behind a priori adaptivity
• a few important a posteriori estimators for 2nd order elliptic BVPs
• how to derive goal oriented error estimators
• the properties of reliable and efficient error estimators
• the algorithm for adaptive local mesh refinement controlled by an a posteriori error estimator
B

Some technical details

B.1 Proof of Lemma 2.2.3

Proof. Step #1: transformation $K \rightarrow \text{“unit triangle”} \quad \hat{K} := \text{convex} \left\{ \left( \begin{array}{c} 0 \\ 0 \\ 0 \end{array} \right), \left( \begin{array}{c} 1 \\ 0 \\ 0 \end{array} \right), \left( \begin{array}{c} 0 \\ 1 \\ 0 \end{array} \right) \right\}$,

\[ \Rightarrow \int_{K} \lambda_1^{\beta_1} \lambda_2^{\beta_2} \lambda_3^{\beta_3} \, dx = 2|K| \int_{0}^{1} \int_{0}^{1-x_1} x_1^{\beta_1} x_2^{\beta_2} (1 - x_1 - x_2)^{\beta_3} \, dx_2 dx_1 \]

\[ = 2|K| \int_{0}^{1} x_1^{\beta_1} \int_{0}^{1} (1 - x_1)^{\beta_2 + \beta_3 + 1} s^{\beta_3} (1 - s)^{\beta_3} \, ds \, dx_1 \]

\[ = 2|K| \int_{0}^{1} x_1^{\beta_1} (1 - x_1)^{\beta_2 + \beta_3 + 1} \, dx_1 \cdot B(\beta_2 + 1, \beta_3 + 1) \]

\[ = 2|K| \cdot B(\beta_1 + 1, \beta_2 + \beta_3 + 2) \cdot B(\beta_2 + 1, \beta_3 + 1) , \]
(\ast) \triangleq \text{substitution } s(1 - x_1) = x_2, \quad B(\cdot, \cdot) \triangleq \text{Euler's beta function}

\[
B(\alpha, \beta) := \int_0^1 t^{\alpha-1} (1 - t)^{\beta-1} \, dt, \quad 0 < \alpha, \beta < \infty.
\]

Using \(\Gamma(\alpha + \beta) B(\alpha, \beta) = \Gamma(\alpha) \Gamma(\beta)\), \(\Gamma \triangleq \text{Gamma function}, \Gamma(n) = (n - 1)!\),

\[
\Rightarrow \quad \int_K \lambda_1^{\beta_1} \lambda_2^{\beta_2} \lambda_3^{\beta_3} \, dx = 2|K| \cdot \frac{\Gamma(\beta_1 + 1) \Gamma(\beta_2 + 1) \Gamma(\beta_3 + 1)}{\Gamma(\beta_1 + \beta_2 + \beta_3 + 3)} \quad \square.
\]