Numerical Methods for Ordinary Differential Equations

Habib Ammari

Wei Wu

Sanghyeon Yu

Contents

CHAPTER 1

Some basics

1.1. What is a differential equation?

An ordinary differential equation (ODE) is an equation that contains one or more derivatives of an unknown function $x(t)$. The equation may also contain x itself and constants. We say that an ODE is of order n if the n-th derivative of the unknown function is the highest order derivative in the equation. The following equations are examples of ODEs:

Membrane equation as a neuron model:

$$C\frac{dx(t)}{dt} + gx(t) = f(t), \tag{1.1}$$

where $x(t)$ is the membrane potential, i.e., the voltage difference between the inside and the outside of the neuron, $f(t)$ is the current flow due to excitation, C is the capacitance and g is the conductance (the inverse of the resistance) of the membrane.

Equation (1.1) is linear ODE of order 1.

The theta model: The theta model is a simple one-dimensional model for the spiking of a neuron. It takes the form

$$\frac{d\theta(t)}{dt} = 1 - \cos\theta(t) + (1 + \cos\theta(t))f(t), \tag{1.2}$$

where $f(t)$ are the inputs to the model. The variable θ lies on the unit circle and ranges between 0 and 2π. When $\theta = \pi$ the neuron spikes, that is, it produces an action potential. By the change of variables, $x(t) = \tan(\theta(t)/2)$, (1.2) leads to the quadratic model

$$\frac{dx(t)}{dt} = x^2(t) + f(t). \tag{1.3}$$

Population growth under competition for resources:

$$\frac{dx(t)}{dt} = rx(t) - \frac{r}{k}x^2(t), \tag{1.4}$$

where r and k are positive parameters. In (1.4), $x(t)$ is the number of cells at time instant t, $rx(t)$ is the growth rate and $-(r/k)x^2(t)$ is the death rate. Equations (1.2), (1.3), and (1.4) are nonlinear ODEs of order 1.

FitzHugh-Nagumo model:

$$\begin{cases} \frac{dV}{dt} = f(V) - W + I \\ \frac{dW}{dt} = a(V - bW), \end{cases} \tag{1.5}$$

where $f(V)$ is a polynomial of third degree, and a and b are constant parameters. The FitzHugh-Nagumo model is a two-dimensional simplification of the Hodgkin-Huxley model of spike generation in squid giant axons. It aims at isolating the mathematical properties of excitation and propagation from the electrochemical properties of sodium and potassium ion flow. In (1.5), V is the membrane potential, W is a recovery variable, and I is the magnitude of stimulus current. Equation (1.5) is a system of nonlinear ODEs of order 1.

Langevin equation of motion for a single particle:

$$\frac{dx(t)}{dt} = ax(t) + \eta, \tag{1.6}$$

where $x(t)$ is the position of the particle at time instant t and η is a random variable that represents some uncertainties or stochastic effects perturbing the particle. Equation (1.6) represents diffusion-like motion from the probabilistic perspective of a single microscopic particle moving in a fluid medium. Equation (1.6) is a linear stochastic ODE of order 1.

Vander der Pol equation:

$$\frac{d^2x(t)}{dt^2} - a(1 - x^2(t))\frac{dx(t)}{dt} + x(t) = 0, \tag{1.7}$$

where a is a positive parameter, which controls the nonlinearity and the strength of the damping. Equation (1.7) is used to generate waveforms corresponding to electrocardiogram patterns. Equation (1.7) is a nonlinear ODE of order 2.

1.1.1. Higher order ODEs. Here we introduce higher order ODEs. Let $\Omega \subset \mathbb{R}^{n+2}$ and $n \in \mathbb{N}$. Then an ODE of order n is an equation of the form:

$$F(t, x(t), \frac{dx}{dt}(t), ..., \frac{d^n x}{dt^n}(t)) = 0,$$

where x is a real-valued unknown function and $dx(t)/dt, ..., d^n x(t)/dt^n$ are its derivatives. We say that $\varphi \in \mathcal{C}^n(I)$ is a solution of the differential equation if I is an open interval,

$$(t, \varphi(t), \frac{\partial \varphi}{\partial t}(t), ..., \frac{\partial^n \varphi}{\partial t^n}(t)) \in \Omega$$

for all $t \in I$, and

$$F(t, \varphi(t), \frac{\partial \varphi}{\partial t}(t), ..., \frac{\partial^n \varphi}{\partial t^n}(t)) = 0$$

for all $t \in I$. When x is a vector valued function, *i.e.*, $x(t) \in \mathbb{R}^d$, then $\Omega \subset \mathbb{R} \times \mathbb{R}^{(n+1)d}$.

Next we consider the following form of n-th order ODE:

$$x^{(n)}(t) = f(t, x, \frac{dx}{dt}, ..., \frac{d^{n-1}x}{dt^{n-1}}), \quad t \in I. \tag{1.8}$$

where $x(t) \in \mathbb{R}^d$ and $f : I \times \mathbb{R}^{nd} \to \mathbb{R}^d$. To ensure uniqueness of the solution, (1.8) has to be augmented with the initial condition:

$$(x(t_0), x'(t_0), x''(t_0), ..., x^{(n-1)}(t_0))^\top.$$

We can reduce the high order ODE (1.8) into a first order ODE. Let us define

$$y(t) := (x(t), dx(t)/dt, ..., d^{n-1}x(t)/dt^{n-1})^\top \in \mathbb{R}^{nd}$$

and

$$F(t, y) := (y_2, ..., y_n, f(t, y_1, ..., y_n))^\top$$

for $y = (y_1, ..., y_n)^\top \in \mathbb{R}^{nd}$ and $y_i \in \mathbb{R}^d$ for $i = 1, 2, ..., n$. Then the n-th order ODE (1.8) is equivalent to the following first order ODE:

$$\frac{dy}{dt} = F(t, y(t)),$$

EXAMPLE 1.1. *Consider the second order ODE given by*

$$\frac{d^2x}{dt^2} + p(t)\frac{dx}{dt} + q(t)x(t) = g(t).$$

Then we have

$$\frac{d}{dt}\begin{bmatrix} x \\ \frac{dx}{dt} \end{bmatrix} = \begin{bmatrix} \frac{dx}{dt} \\ -p(t)\frac{dx}{dt} - q(t)x(t) + g(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -q(t) & -p(t) \end{bmatrix}\begin{bmatrix} x \\ \frac{dx}{dt} \end{bmatrix} + \begin{bmatrix} 0 \\ g(t) \end{bmatrix}.$$

The main problems concerning ordinary differential equations are:

(i) Existence of solutions;
(ii) Uniqueness of solutions with suitable initial conditions;
(iii) Regularity and stability of solutions (e.g. dependence on the initial conditions, large time stability, higher regularity);
(iv) Computation of solutions.

The existence of solutions can be proved by fixed point theorems, by the implicit function theorem in Banach spaces, and by functional analysis techniques. The problem of uniqueness is typically more difficult. Only in a very few special cases is it possible to compute solutions in some explicit form.

1.2. Some methods of resolution

In the following subsections, we present several examples of exactly solvable ODEs and then explain how to solve them.

1.2.1. Separation of variables. Let I and J be two open intervals and let $f \in \mathcal{C}^0(I)$ and $g \in \mathcal{C}^0(J)$ be two continuous functions. We look for solutions to the first order equation

$$\frac{dx}{dt} = f(t)g(x). \tag{1.9}$$

Let $t_0 \in I$ and $x_0 \in J$. If $g(x_0) = 0$ for some $x_0 \in J$, then the constant function $x(t) = x_0$ for $t \in I$ is a solution to (1.9). Suppose that $g(x_0) \neq 0$. Then $g \neq 0$ in a neighborhood of x_0 and we can divide (1.9) by $g(x)$ and hence, **separate the variables**. We find

$$\frac{dx}{g(x)} = f(t)dt. \tag{1.10}$$

Integrating (1.10) gives

$$\int \frac{dx}{g(x)} = \int f(t)dt + c,$$

where the constant c is uniquely determined by the initial condition.

Let F and G be the primitives of f and $1/g$, respectively. The function G is strictly monotonic, because $G'(x) \neq 0$, and thus invertible. The solution of the differential equation (1.9) is then

$$x(t) = G^{-1}(F(t) + c).$$

This method of solving ODEs is called the **method of separation of variables** and (1.9) is called a **separable equation**.

Example 1.2. *Consider the following ODE:*

$$\begin{cases} \dfrac{dx}{dt} = \dfrac{1+2t}{\cos x(t)}, \\ x(0) = \pi. \end{cases}$$

In this case, we have $g(x) = 1/\cos x$ *and* $f(t) = 1+2t$. *Note that* g *is defined for* $x \neq \pi/2 + k\pi, k \in \mathbb{Z}$. *By separating variables, we get*

$$\cos x dx = 1 + 2t dt.$$

By integration, we have

$$\sin x(t) = t^2 + t + C,$$

for some constant $C \in \mathbb{R}$. *Then, from the initial condition* $x(0) = \pi$, *we see that* $C = 0$.

One might think that we can obtain the solution by taking the arcsin. *But the function* $x(t) = \arcsin(t^2+t)$ *is not the solution because* $x(0) = \arcsin(0) = 0$. *In order to get the correct solution, we note that* arcsin *is the inverse of* sin *on* $[-\pi/2, \pi/2]$, *whereas* $x(t)$ *takes the values in a neighborhood*

of π. Letting $w(t) = x(t) - \pi$, we have $w(0) = x(0) - \pi = 0$. So, we have $w(t) = -\arcsin(t^2 + t)$. Therefore, we get the following correct solution:

$$x(t) = \pi - \arcsin(t^2 + t).$$

1.2.2. Change of variables. Consider the following ODE:

$$\frac{dx}{dt} = f\Big(\frac{x(t)}{t}\Big), \tag{1.11}$$

where $f : I \subset \mathbb{R} \to \mathbb{R}$ is a continuous function on some open interval $I \subset \mathbb{R}$. By the **change of variable** $x(t) = ty(t)$ where $y(t)$ is the new unknown function, the above ODE can be changed to a separable equation. Since

$$\frac{dx}{dt} = y(t) + t\frac{dy}{dt} = f(y(t)),$$

we have a separable equation for y, which reads:

$$\frac{dy}{f(y) - y} = \frac{dt}{t}.$$

Therefore, (1.11) can be solved by the method of separation of variables.

Example 1.3. *Consider*

$$\frac{dx}{dt} = \frac{t^2 + x^2}{xt}.$$

In this case, $f(s) = s + 1/s$ with $s = x/t$. By letting $y(t) = x(t)/t$, we get $ydy = dt/t$. So, we have $(1/2)y^2 = \ln t + C$. Therefore, we obtain

$$x(t) = \pm t\sqrt{2(\ln t + C)}.$$

1.2.3. Method of integrating factors. Consider

$$\frac{dx(t)}{dt} = f(t). \tag{1.12}$$

By integrating (1.12), it follows that the solution $x(t)$ is given by

$$x(t) = x(0) + \int_0^t f(s)\, ds.$$

Consider

$$\frac{dx}{dt} + p(t)x(t) = g(t), \tag{1.13}$$

where p and g are functions of t.

If (1.13) were of the form (1.12), then we could immediately write down a solution in terms of integrals. By (1.13) being of the form (1.12), we mean that the left-hand side is expressed as the derivative of our unknown quantity. To make this happen, we can multiply (1.13) by a function, $\mu(t)$, and ask whether the resulting equation can be put in the form (1.12).

Let us look for $\mu(t)$ such that

$$\mu(t)\frac{dx}{dt} + \mu(t)p(t)x(t) = \frac{d}{dt}(\mu(t)x(t)).$$

Taking derivatives, we have $(1/\mu)d\mu/dt = p(t)$ or

$$\frac{d}{dt}\ln\mu(t) = p(t). \tag{1.14}$$

Integrating (1.14) gives

$$\mu(t) = \exp(\int_0^t p(s)ds),$$

up to a multiplicative constant. The equation (1.13) is transformed to

$$\frac{d}{dt}(\mu(t)x(t)) = \mu(t)g(t).$$

This equation is precisely of the form (1.12), so we can immediately conclude

$$x(t) = \frac{1}{\mu(t)}\Big(\int_0^t \mu(s)g(s)ds\Big) + \frac{C}{\mu(t)},$$

where the constant C can be determined from the initial condition $x(0) = x_0$. The function $\mu(t)$ is called the **integrating factor**.

EXAMPLE 1.4. *Consider*

$$\begin{cases} \dfrac{dx}{dt} + \dfrac{1}{t+1}x(t) = (1+t)^2, \qquad t \geq 0, \\ x(0) = 1. \end{cases}$$

In this case, $p(t) = 1/(t+1)$ and $g(t) = (1+t)^2$. Then the integrating factor μ is

$$\mu(t) = \exp(\int_0^t p(s)ds) = e^{\ln(t+1)} = t+1.$$

Therefore, we get

$$x(t) = \frac{1}{t+1}\int_0^t (s+1)^3 ds + \frac{C}{t+1} = \frac{(t+1)^3}{4} + \frac{C-\frac{1}{4}}{t+1}.$$

Then, from the initial condition $x(0) = 1$, we obtain $C = 1$.

EXAMPLE 1.5. *(Bernoulli's equation) Consider*

$$\frac{dx}{dt} + p(t)x(t) = g(t)x^\alpha(t). \tag{1.15}$$

Here α is a real parameter satisfying $\alpha \notin \{0, 1\}$. Letting $x = z^{\frac{1}{1-\alpha}}$, we get

$$\frac{dx}{dt} = \frac{1}{1-\alpha} z^{\frac{\alpha}{1-\alpha}} \frac{dz}{dt}.$$

Then (1.15) can be reduced to the following linear equation:

$$\frac{dz}{dt} + (1-\alpha)p(t)z(t) = (1-\alpha)g(t),$$

which can be solved by the method of integrating factors.

1.3. Important examples of ODEs

1.3.1. Autonomous ODEs.

DEFINITION 1.6. *The equation*

$$\frac{dx(t)}{dt} = f(t, x(t)) \tag{1.16}$$

*is called **autonomous** if f is independent of t.*

Any ODE can be rewritten as an autonomous ODE on a higher-dimensional space. Writing $y = (t, x(t))$, (1.16) is equivalent to the autonomous ODE

$$\frac{dy(t)}{dt} = F(y(t)),$$

where $F(y) = \begin{pmatrix} 1 \\ f(t, x(t)) \end{pmatrix}$.

1.3.2. Exact equations. Let $\Omega = I \times \mathbb{R} \subset \mathbb{R}^2$ with $I \subset \mathbb{R}$ being an open interval. Let $f, g \in \mathcal{C}^0(\Omega)$. We look for a solution $x \in \mathcal{C}^1(I)$ of the differential equation

$$f(t, x(t)) + g(t, x(t))\frac{\mathrm{d}x}{\mathrm{d}t} = 0 \tag{1.17}$$

satisfying the initial condition $x(t_0) = x_0$ for some $(t_0, x_0) \in \Omega$.

Consider the differential form

$$\omega = f(t, x)\mathrm{d}t + g(t, x)\mathrm{d}x.$$

Definition 1.7. *The differential form is called **exact** if there exists $F \in \mathcal{C}^1(\Omega)$ such that*

$$\omega = \mathrm{d}F = \frac{\partial F}{\partial t}\mathrm{d}t + \frac{\partial F}{\partial x}\mathrm{d}x.$$

*The function F is called a **potential** of ω. In this case the differential equation* (1.17) *is called an **exact equation**.*

Theorem 1.8 (Implicit function theorem). *Suppose that $F(t, x)$ is continuously differentiable in a neighborhood of $(t_0, x_0) \in \mathbb{R} \times \mathbb{R}^d$ and $F(t_0, x_0) = 0$. Suppose that $\partial F/\partial x(t_0, x_0) \neq 0$. Then there exists a $\delta > 0$ and $\epsilon > 0$ such that for each t satisfying $|t - t_0| < \delta$, there exists a unique x such that $|x - x_0| < \epsilon$ for which $F(t, x) = 0$. This correspondence defines a function $x(t)$ continuously differentiable on $\{|t - t_0| < \delta\}$ such that*

$$F(t, x) = 0 \Leftrightarrow x = x(t).$$

Theorem 1.9. *Suppose that ω is an exact form with potential F such that*

$$\frac{\partial F}{\partial x}(t_0, x_0) \neq 0,$$

then the equation $F(t, x) = F(t_0, x_0)$ implicitly defines a function $x \in \mathcal{C}^1(I)$ for some open interval I containing t_0, which solves (1.17) *with the initial condition $x(t_0) = x_0$. This solution is unique on I.*

Proof. Suppose without loss of generality that $F(t_0, x_0) = 0$. By the **implicit function theorem**, there exists $\delta, \eta > 0$ and $x \in \mathcal{C}^1(t_0 - \delta, t_0 + \delta)$ such that

$$\{(t, x) \in \Omega : |t - t_0| < \delta, |x - x_0| < \eta, F(t, x) = 0\} = \{(t, x(t)) \in \Omega : |t - t_0| < \delta\}.$$

By differentiating the identity $F(t, x(t)) = 0$, we get

$$0 = \frac{\mathrm{d}}{\mathrm{d}t}F(t, x(t)) = \frac{\partial F}{\partial t}(t, x(t)) + \frac{\partial F}{\partial x}(t, x(t))\frac{\mathrm{d}x}{\mathrm{d}t} = f(t, x(t)) + g(t, x(t))\frac{\mathrm{d}x}{\mathrm{d}t},$$

and hence $x(t)$ is a solution of the differential equation. Moreover, $x(t_0) = x_0$.

On the other hand, if $z \in \mathcal{C}^1(I)$ is a solution to (1.17) such that $z(t_0) = x_0$, then

$$\frac{\mathrm{d}}{\mathrm{d}t}F(t, z(t)) = 0 \Longrightarrow F(t, z(t)) = F(t_0, z(t_0)) = 0 \Longrightarrow z(t) = x(t).$$

□

Proposition 1.10. *Let $f, g \in \mathcal{C}^1(\Omega)$. The differential form $\omega = f\mathrm{d}t + g\mathrm{d}x$ is **closed** in Ω if*

$$\frac{\partial f}{\partial x} = \frac{\partial g}{\partial t}$$

for all $(t, x) \in \Omega$.

Proposition 1.11. *An exact differential form $\omega = f\mathrm{d}t + g\mathrm{d}x$ with a potential $F \in \mathcal{C}^2$ is closed if*

$$\frac{\partial^2 F}{\partial t \partial x} = \frac{\partial^2 F}{\partial x \partial t}$$

for all $(t, x) \in \Omega$. The converse is also true if Ω is simply connected.

Closed forms always have a potential (at least locally).

EXAMPLE 1.12. *Consider the equation*

$$tx^2 + x - t\frac{\mathrm{d}x}{\mathrm{d}t} = 0. \tag{1.18}$$

Here, $f(t,x) = tx^2 + x$ *and* $g(t,x) = -t$. *Since*

$$\frac{\partial f}{\partial x} = 2xt + 1 \neq \frac{\partial g}{\partial t} = -1,$$

equation (1.18) is not exact.

EXAMPLE 1.13. *The equation*

$$t + \frac{1}{x} - \frac{t}{x^2}\frac{\mathrm{d}x}{\mathrm{d}t} = 0$$

is exact with the potential function F *given by*

$$F(t,x) = \frac{t^2}{2} + \frac{t}{x} + C, \quad C \in \mathbb{R}.$$

The equation $F(t,x) = 0$ *implicitly defines the solutions (locally for* $t \neq 0$ *and* $x \neq 0$ *such that* $\partial F/\partial x(t,x) \neq 0$*).*

1.3.3. Hamiltonian systems.

DEFINITION 1.14. *Let* M *be a subset of* $\mathbb{R}^d$ *and let* $H : \mathbb{R}^d \times M \to M$ *be a* $\mathcal{C}^1$ *function. The Hamiltonian system with Hamiltonian* H *is given by the first-order system of ODEs*

$$\begin{cases} \dfrac{\mathrm{d}p}{\mathrm{d}t} = -\dfrac{\partial H}{\partial q}(p,q), \\ \dfrac{\mathrm{d}q}{\mathrm{d}t} = \dfrac{\partial H}{\partial p}(p,q). \end{cases} \tag{1.19}$$

EXAMPLE 1.15. *An important basic example of a Hamiltonian system is the simple harmonic oscillator with Hamiltonian*

$$H(p,q) = \frac{1}{2}\frac{p^2}{m} + \frac{1}{2}kq^2,$$

where m *and* k *are positive constants. Given a potential* V*, Hamiltonian systems of the form*

$$H(p,q) = \frac{1}{2}p^\top M^{-1}p + V(q),$$

where M *is symmetric positive definite matrix and* $\top$ *denotes the transpose, are widely used in* ***molecular*** *and* ***biological dynamics****.*

We now introduce the notion of an invariant (also called **first integral**) for a system of ODEs.

DEFINITION 1.16. *Let* $\Omega = I \times D$*, where* $I \subset \mathbb{R}$ *and* $D \subset \mathbb{R}^d$*. Consider*

$$\frac{\mathrm{d}x}{\mathrm{d}t} = f(t, x(t)), \tag{1.20}$$

where $f : \Omega \to \mathbb{R}^d$*. We call* $F : D \to \mathbb{R}$ *an* ***invariant*** *of* (1.20) *if* $F(x(t)) = Constant$*. A point* $(t,x) \in I \times D$ *is called a* ***stationary point*** *if* $f(t,x) = 0$*.*

EXAMPLE 1.17. *Consider the system of* ***Lotka-Volterra****'s ODEs given by*

$$\begin{cases} \dfrac{\mathrm{d}u}{\mathrm{d}t} = u(v-2), \\ \dfrac{\mathrm{d}v}{\mathrm{d}t} = v(1-u). \end{cases} \tag{1.21}$$

The system of ODEs (1.21) is used to describe the dynamics of biological systems in which two species interact, one as a predator and the other as prey.

Define

$$F(u,v) := \ln u - u + 2\ln v - v.$$

$F(u,v)$ *is an invariant of* (1.21). *In fact, by differentiating with respect to time, we have*

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}F(u,v) &= \frac{1}{u}\frac{\mathrm{d}u}{\mathrm{d}t} - \frac{\mathrm{d}u}{\mathrm{d}t} + \frac{2}{v}\frac{\mathrm{d}v}{\mathrm{d}t} - \frac{\mathrm{d}v}{\mathrm{d}t}\\ &= v - 2 - \frac{\mathrm{d}u}{\mathrm{d}t} + 2(1-u) - \frac{\mathrm{d}v}{\mathrm{d}t}\\ &= (v-2) - u(v-2) + 2(1-u) + v(1-u)\\ &= (v-2)(1-u) + (2-v)(1-u)\\ &= 0.\end{aligned}$$

For the system (1.21), $(u,v) = (1,2)$ *and* $(u,v) = (0,0)$ *are two stationary points.*

Lemma 1.18. *The Hamiltonian H is an invariant of the associated Hamiltonian system (1.19).*

Proof. We have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}H(p(t),q(t)) &= \frac{\partial H}{\partial p}(p(t),q(t))\frac{\mathrm{d}p}{\mathrm{d}t} + \frac{\partial H}{\partial q}(p(t),q(t))\frac{\mathrm{d}q}{\mathrm{d}t}\\ &= -\frac{\partial H}{\partial p}(p(t),q(t))\frac{\partial H}{\partial q}(p(t),q(t)) + \frac{\partial H}{\partial q}(p(t),q(t))\frac{\partial H}{\partial p}(p(t),q(t)) = 0.\end{aligned}$$

Hence, $H(p,q)$ is an invariant of the system of equations (1.19). □

Example 1.19. *Consider the system of equations*

$$\begin{cases}\dfrac{\mathrm{d}p}{\mathrm{d}t} = -\sin q,\\ \dfrac{\mathrm{d}q}{\mathrm{d}t} = p.\end{cases}$$

Here, $H(p,q) = \frac{1}{2}p^2 - \cos q$ *is the Hamiltonian of the above system, because*

$$\begin{cases}\dfrac{\partial H}{\partial q} = \sin q = -\dfrac{\mathrm{d}p}{\mathrm{d}t}\\ \dfrac{\partial H}{\partial p} = p = \dfrac{\mathrm{d}q}{\mathrm{d}t}\end{cases}$$

There is another equivalent expression for Hamiltonian systems. Let $x = (p,q)^\top$ (note that $p, q \in \mathbb{R}^d$), and let

$$J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}, \tag{1.22}$$

where I denotes the $d \times d$ identity matrix. We can rewrite the Hamiltonian system (1.19) in the form

$$\frac{\mathrm{d}x}{\mathrm{d}t} = J^{-1}\nabla H(x). \tag{1.23}$$

Note that $J^{-1} = J^\top$.

Definition 1.20 (Symplectic linear mapping). *A matrix $A \in \mathbb{R}^{2d} \times \mathbb{R}^{2d}$ (which is also a linear mapping from $\mathbb{R}^{2d}$ to $\mathbb{R}^{2d}$) is called* ***symplectic*** *if $A^\top JA = J$.*

Definition 1.21 (Symplectic mapping). *A differentiable map $g : U \to \mathbb{R}^{2n}$ is called* ***symplectic*** *if the* ***Jacobian matrix*** *$g'(p,q)$ is everywhere symplectic, i.e., if*

$$g'(p,q)^\top J g'(p,q) = J.$$

Taking the transpose of both sides of the above equation, we also have

$$g'(p,q)^\top J^\top g'(p,q) = J^\top,$$

or equivalently,

$$g'(p,q)^\top J^{-1} g'(p,q) = J^{-1}.$$

THEOREM 1.22. *If g is a symplectic mapping, then it preserves the Hamiltonian form of the equation.*

PROOF. Let $x = (p,q)^\top$, $y = g(p,q)^\top$ and let $G(y) := H(x)$. By using the chain rule, we have

$$\begin{aligned}\frac{\partial}{\partial x} H(x) &= \frac{\partial}{\partial x} G(y) \\ &= \frac{\partial}{\partial y} G(y) \nabla_x y(x) \\ &= \nabla_y G(y) g'^\top(p,q).\end{aligned}$$

Then,

$$\begin{aligned}\frac{\mathrm{d}y}{\mathrm{d}t} &= g'^\top(p,q) \frac{\mathrm{d}x}{\mathrm{d}t} \\ &= g'^\top(p,q) J^{-1} \left(\frac{\partial H(x)}{\partial x}\right)^\top \\ &= g'^\top J^{-1} g' \nabla_y G(y) \\ &= J^{-1} \nabla_y G(y),\end{aligned}$$

and therefore,

$$\frac{\mathrm{d}y}{\mathrm{d}t} = J^{-1} \nabla_y G(y).$$

□

DEFINITION 1.23 (Flow). *We define the **flow** by $\phi_t(p_0,q_0) = (p(t,p_0,q_0), q(t,p_0,q_0))$, $\phi_t : U \to \mathbb{R}^{2d}$, $U \subset \mathbb{R}^{2d}$, and p_0 and q_0 are the initial data at $t = 0$.*

THEOREM 1.24 (Poincaré's theorem). *Suppose that H is twice differentiable. Then the flow ϕ_t is a symplectic transformation whenever it is defined.*

PROOF. Let $y_0 = (p_0, q_0)$. We have

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}\left(\left(\frac{\partial \phi_t}{\partial y_0}\right)^\top J \left(\frac{\partial \phi_t}{\partial y_0}\right)\right) \\ =&\left(\frac{\partial \phi_t}{\partial y_0}\right)'^\top J \left(\frac{\partial \phi_t}{\partial y_0}\right) + \left(\frac{\partial \phi_t}{\partial y_0}\right)^\top J \left(\frac{\partial \phi_t}{\partial y_0}\right)' \\ =&\left(\frac{\partial \phi_t}{\partial y_0}\right)^\top \nabla^2 H J^{-\top} J \left(\frac{\partial \phi_t}{\partial y_0}\right) + \left(\frac{\partial \phi_t}{\partial y_0}\right)^\top J J^{-1} \nabla^2 H \left(\frac{\partial \phi_t}{\partial y_0}\right) \\ =&0,\end{aligned}$$

where $\nabla^2 H$ is the Hessian matrix of $H(p,q)$ (and is symmetric). Moreover, since $\partial \phi_t / \partial y_0$ at $t = 0$ is the identity map, the identity

$$\left(\frac{\partial \phi_t}{\partial y_0}\right)^\top J \left(\frac{\partial \phi_t}{\partial y_0}\right) = J$$

is satisfied for all t and all (p_0, q_0) as long as the solution remains in the domain of definition of H. □

The following result shows that the symplecticity of the flow is a characteristic property of the Hamiltonian system.

THEOREM 1.25. *Let $f : U \to \mathbb{R}^{2n}$ be continuously differentiable. Then $\frac{\mathrm{d}x}{\mathrm{d}t} = f(x)$ is locally Hamiltonian if and only if $\phi_t(x)$ is symplectic for all $x \in U$ and for all sufficiently small t.*

PROOF. The necessity follows from Theorem 1.24. We therefore suppose that ϕ_t is symplectic, and we have to prove the local existence of a Hamiltonian H such that $f(x) = J^{-1}\nabla H(s)$. Using the fact that $\frac{\partial \phi_t}{\partial y_0}$ is a solution of

$$\frac{\mathrm{d}y}{\mathrm{d}t} = f'(\phi_t(y_0))y,$$

we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\left(\frac{\partial \phi_t}{\partial y_0}\right)^\top J \left(\frac{\partial \phi_t}{\partial y_0}\right)\right) = \left(\frac{\partial \phi_t}{\partial y_0}\right)^\top [f'(\phi_t(y_0))^\top J + Jf'] \left(\frac{\partial \phi_t}{\partial y_0}\right) = 0.$$

Putting $t = 0$, it follows from $J = -J^\top$ that $Jf'(y_0)$ is a symmetric matrix for all y_0. The **integrability lemma** below shows that $Jf(y)$ can be written as the gradient of a function H. □

LEMMA 1.26 (Integrability lemma). *Let $D \subset \mathbb{R}^d$ be an open set and let $g : D \to \mathbb{R}^d$ be of class $\mathcal{C}^1$. Suppose that the Jacobian $g'(y)$ is symmetric for all $y \in D$. Then, for every $y_0 \in D$, there exists a neighborhood of y_0 and a function $H(y)$ such that*

$$g(y) = \nabla H(y)$$

on this neighborhood.

PROOF. Suppose that $y_0 = 0$, and consider a ball around y_0 which is contained in D. On this ball we define

$$H(y) = \int_0^1 y^\top g(ty)\mathrm{d}t.$$

Differentiating with respect to y_k, and using the symmetry assumption

$$\frac{\partial g_i}{\partial y_k} = \frac{\partial g_k}{\partial y_i}$$

yields

$$\begin{aligned}\frac{\partial H}{\partial y_k} &= \int_0^1 (g_k(ty) + y^\top \frac{\partial g}{\partial y_k}(ty)t)\mathrm{d}t \\ &= \int_0^1 \frac{\mathrm{d}}{\mathrm{d}t}(tg_k(ty))\mathrm{d}t = g_k(y)\end{aligned}$$

which proves that

$$\nabla H = g.$$

□

Finally, consider the gradient system:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -\nabla F(x), \tag{1.24}$$

where F is the potential function. Equation (1.24) is a particular example of exact equations.

LEMMA 1.27. *The Hamiltonian system (1.19) is a **gradient system** if and only if the function H is harmonic.*

PROOF. Suppose that H is harmonic, i.e.,

$$\frac{\partial^2 H}{\partial p^2} + \frac{\partial^2 H}{\partial q^2} = 0.$$

Then the Jacobian of $J^{-1}\nabla H$ given by

$$(J^{-1}\nabla H)' = \begin{pmatrix} -\frac{\partial^2 H}{\partial p \partial q} & -\frac{\partial^2 H}{\partial q^2} \\ \frac{\partial^2 H}{\partial p^2} & \frac{\partial^2 H}{\partial p \partial q} \end{pmatrix}$$

is symmetric. The integrability lemma shows that there exists V such that $J^{-1}\nabla H = \nabla V$ and therefore, the Hamiltonian system is a gradient system.

Suppose that the Hamiltonian system is a gradient system. Then, there exists V such that

$$\frac{\partial V}{\partial p} = \frac{\partial H}{\partial q} \quad \text{and} \quad \frac{\partial V}{\partial q} = -\frac{\partial H}{\partial p}.$$

Therefore,

$$\Delta H := \frac{\partial^2 H}{\partial p^2} + \frac{\partial^2 H}{\partial q^2} = 0.$$

□

EXAMPLE 1.28. *The Hamiltonian system with $H(p,q) = p^2 - q^2$ is a gradient system.*

CHAPTER 2

Existence, uniqueness, and regularity in the Lipschitz case

2.1. Banach fixed point theorem

DEFINITION 2.1 (Contraction). *Let (X, d) be a metric space. A mapping $F : X \to X$ is a* ***contraction*** *if there exists $0 < \lambda < 1$ such that*

$$d(F(x), F(y)) \leq \lambda d(x, y)$$

for all $x, y \in X$.

THEOREM 2.2 (Banach fixed point theorem). *Let (X, d) be a* ***complete*** *metric space (i.e., every Cauchy sequence of elements of X is convergent) and let $F : X \to X$ be a contraction. Then there exists a unique $x \in X$ such that*

$$F(x) = x.$$

2.2. Gronwall's lemma

LEMMA 2.3 (**Gronwall's lemma**). *Let $I = [0, T]$ and let $\phi \in \mathcal{C}^0(I)$. If there exist two constants $\alpha, \beta \in \mathbb{R}$, $\beta \geq 0$, such that*

$$\phi(t) \leq \alpha + \beta \int_0^t \phi(s) \mathrm{d}s \quad \text{for all } t \in I, \tag{2.1}$$

then

$$\phi(t) \leq \alpha e^{\beta t} \quad \text{for all } t \in I.$$

PROOF. Let $\varphi : I \to \mathbb{R}$ be the function

$$\varphi(t) := \alpha + \beta \int_0^t \phi(s) \mathrm{d}s.$$

Since $\phi \in \mathcal{C}^0$, we conclude that $\varphi \in \mathcal{C}^1$, and

$$\frac{\mathrm{d}\varphi}{\mathrm{d}t} = \beta \phi(t) \qquad \text{for all } t \in I.$$

By using (2.1), it follows that

$$\frac{\mathrm{d}\varphi}{\mathrm{d}t} \leq \beta \varphi.$$

Let $\psi(t) := \exp(-\beta t)\varphi(t)$ for $t \in I$. Then

$$\begin{aligned} \frac{\mathrm{d}\psi}{\mathrm{d}t} &= -\beta e^{-\beta t} \varphi(t) + e^{-\beta t} \frac{\mathrm{d}\varphi}{\mathrm{d}t} \\ &= e^{-\beta t} \left(-\beta \varphi(t) + \frac{\mathrm{d}\varphi}{\mathrm{d}t} \right) \leq 0. \end{aligned}$$

Since $\psi(0) = \varphi(0) = \alpha$, we have $\psi(t) \leq \alpha$ for $t \in I$, and hence

$$\varphi(t) \leq \alpha e^{\beta t},$$

which implies that $\phi(t) \leq \varphi(t) \leq \alpha e^{\beta t}$ for all $t \in I$. □

2.3. Cauchy-Lipschitz theorem

Let $I=[0,T]$, let d be a positive integer, and let $f: I\times\mathbb{R}^d\to\mathbb{R}^d$. Suppose that $f\in\mathcal{C}^0(I\times\mathbb{R}^d)$.

DEFINITION 2.4 (Lipschitz condition). *If there exists a constant $C_f\geq 0$ such that, for any $x_1,x_2\in\mathbb{R}^d$ and any $t\in I$, the following inequality holds:*

$$|f(t,x_1)-f(t,x_2)|\leq C_f|x_1-x_2|, \tag{2.2}$$

*then we say that f satisfies a **Lipschitz condition** on I. The constant C_f is called the **Lipschitz constant** for f.*

THEOREM 2.5 (Cauchy-Lipschitz theorem). *Consider the initial value problem*

$$\begin{cases} \dfrac{\mathrm{d}x}{\mathrm{d}t}=f(t,x), \quad t\in[0,T],\\ x(0)=x_0, \quad x_0\in\mathbb{R}^d. \end{cases} \tag{2.3}$$

If $f\in\mathcal{C}^0(I\times\mathbb{R}^d)$ satisfies the Lipschitz condition (2.2) on $[0,T]$, then there exists a unique solution $x\in\mathcal{C}^1(I)$ to (2.3) on $[0,T]$.

PROOF. By (2.3), we have

$$x(t)=x_0+\int_0^t f(s,x(s))\mathrm{d}s, \quad \forall t\in[0,T].$$

Define the functional $F:\mathcal{C}^0([0,T];\mathbb{R}^d)\to\mathcal{C}^0([0,T];\mathbb{R}^d)$ by

$$F(y):=x_0+\int_0^t f(s,y(s))\mathrm{d}s.$$

For $y\in\mathcal{C}^0([0,T];\mathbb{R}^d)$, defined the norm of y by

$$\|y\|:=\sup_{t\in[0,T]}\{|y(t)|e^{-C_f t}\}, \tag{2.4}$$

where C_f is the Lipschitz constant for f. It is easy to prove that (2.4) is equivalent to the usual norm $\sup_{t\in[0,T]}|y(t)|$ and hence, $\mathcal{C}^0([0,T];\mathbb{R}^d)$ equipped with (2.4) is complete.

With (2.4), we compute

$$\begin{aligned} \|F[y_1]-F[y_2]\| &= \sup_{t\in[0,T]}|F[y_1](t)-F[y_2](t)|e^{-C_f t}\\ &\leq \sup_{t\in[0,T]} e^{-C_f t}\int_0^t |f(s,y_1(s))-f(s,y_2(s))|\mathrm{d}s\\ &\leq \sup_{t\in[0,T]} e^{-C_f t}C_f\int_0^t |y_1(s)-y_2(s)|\mathrm{d}s\\ &\leq \sup_{t\in[0,T]} e^{-C_f t}C_f\int_0^t e^{C_f s}e^{-C_f s}|y_1(s)-y_2(s)|\mathrm{d}s\\ &\leq \sup_{t\in[0,T]}\{e^{-C_f t}C_f\int_0^t e^{C_f s}\mathrm{d}s\}\|y_1-y_2\|\\ &\leq (1-e^{-C_f T})\|y_1-y_2\|. \end{aligned}$$

By Banach fixed point theorem in a complete metric space (Theorem 2.2), there exists a unique $y\in\mathcal{C}^0([0,T];\mathbb{R}^d)$ such that $F(y)=y$. The **Picard iteration**

$$y^{(n+1)}=F[y^{(n)}]$$

is a Cauchy sequence and converges to the unique fixed point y. Therefore, there exists a unique solution to (2.3). □

REMARK 2.6. *Theorem 2.5 holds true if $\mathbb{R}^d$ is replaced with a* ***Banach space*** *(a complete normed vector space). The proof is the same.*

If f is continuous, there is no guarantee that the initial value problem (2.3) possesses a unique solution.

EXAMPLE 2.7. *Consider*

$$\frac{\mathrm{d}x}{\mathrm{d}t} = x^{\frac{2}{3}}, \quad x(0) = 0. \tag{2.5}$$

Then there are two solutions to (2.5) given by $x_1(t) = \frac{t^3}{27}$ and $x_2(t) = 0$.

THEOREM 2.8 (Cauchy-Peano existence theorem). *If f is continuous, then* (2.3) *admits a solution $x(t)$ that is, at least, defined for small t.*

This theorem can be proved by using the **Arzela-Ascoli theorem**.

DEFINITION 2.9 (Equicontinuity). *A family of functions $\mathcal{F}$ is said to be* ***equicontinuous*** *on $[a, b]$ if for any given $\epsilon > 0$, there exists $\delta > 0$ such that*

$$|f(t) - f(s)| < \epsilon$$

whenever $|t - s| < \delta$ for every function $f \in \mathcal{F}$ and $t, s \in [a, b]$.

DEFINITION 2.10 (Uniform boundedness). *A family of continuous functions $\mathcal{F}$ on $[a, b]$ is uniformly bounded if there exists a positive number M such that $|f(t)| \leq M$ for every function $f \in \mathcal{F}$ and $t \in [a, b]$.*

THEOREM 2.11 (Arzela-Ascoli). *Suppose that the sequence of functions $\{f_n(t)\}_{n\in\mathbb{N}}$ on $[a, b]$ is uniformly bounded and equicontinuous, then there exists a subsequence $\{f_{n_k}(t)\}_{k\in\mathbb{N}}$ that is uniformly convergent on $[a, b]$.*

EXAMPLE 2.12. *Consider*

$$\frac{\mathrm{d}x}{\mathrm{d}t} = x^2, \quad x(0) = x_0 \neq 0.$$

By ***separation of variables****, we obtain*

$$\frac{\mathrm{d}x}{x^2} = \mathrm{d}t.$$

Thus,

$$-\frac{1}{x} = \int \frac{\mathrm{d}x}{x^2} = t + C,$$

and hence,

$$x = -\frac{1}{t + C}.$$

Since $x(0) = x_0$,

$$x(t) = \frac{x_0}{1 - x_0 t}.$$

If $x_0 > 0$, $x(t)$ blows up when $t \to \frac{1}{x_0}$ from below. If $x_0 < 0$, the singularity is in the past($t < 0$). The only solution defined for all positive and negative t is the constant solution $x(t) = 0$, corresponding to $x_0 = 0$.

Now we turn to the continuity of the solution of (2.3).

THEOREM 2.13 (**Continuity with respect to the initial data**). *Suppose that f satisfies the Lipschitz condition (2.2). Let $x_1(t)$ and $x_2(t)$ be the solutions of* (2.3) *corresponding to the initial data $x_1(0)$ and $x_2(0)$, respectively. Then we have*

$$|x_1(t) - x_2(t)| \leq e^{C_f t}|x_1(0) - x_2(0)| \quad \textit{for all } t \in [0, T]. \tag{2.6}$$

PROOF. Since

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}|x_1(t)-x_2(t)|^2 &= 2(f(t,x_1(t))-f(t,x_2(t)))(x_1(t)-x_2(t))\\ &\le 2C_f|x_1(t)-x_2(t)|^2, \quad t\in[0,T],\end{aligned}$$

we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(|x_1(t)-x_2(t)|^2 e^{-2C_f t}\right)\le 0. \tag{2.7}$$

Integrating (2.7) from 0 to t gives

$$|x_1(t)-x_2(t)|^2 e^{-2C_f t}\le |x_1(0)-x_2(0)|^2,$$

or equivalently,

$$|x_1(t)-x_2(t)|\le |x_1(0)-x_2(0)|e^{C_f t},$$

which yields the desired inequality. □

Next we discuss the differentiability of the solution of (2.3) with respect to the initial data.

Formally, taking the derivative of the solution x of (2.3) with respect to the initial data, we obtain that $\partial x(t)/\partial x_0$ is the solution of the linear equation

$$\begin{cases}\dfrac{\mathrm{d}}{\mathrm{d}t}\dfrac{\partial x(t)}{\partial x_0}=\dfrac{\partial f}{\partial x}(t,x(t))\dfrac{\partial x(t)}{\partial x_0},\\ \dfrac{\partial x(t)}{\partial x_0}=1.\end{cases} \tag{2.8}$$

THEOREM 2.14. *Suppose that f is of class $\mathcal{C}^1$. Then $x_0\mapsto x(t)$ is differentiable and $\partial x(t)/\partial x_0$ is the unique solution of the linear equation (2.8).*

PROOF. Let $\Delta x(t,x_0,h):=x(t,x_0+h)-x(t,x_0)$ be the difference quotient. By using the **mean-value theorem**, we have

$$\begin{aligned}\Delta x(t,x_0,h) &= h+\int_0^t (f(s,x(s,x_0+h))-f(s,x(s,x_0)))\mathrm{d}s\\ &= h+\int_0^t (f(s,x(t,x_0)+\Delta x(s,x_0,h))-f(s,x(s,x_0)))\mathrm{d}s\\ &= h+\int_0^t \frac{\partial f}{\partial x}(s,x(s,x_0)+\tau\Delta x)\Delta x\mathrm{d}s,\end{aligned}$$

where $\tau=\tau(s,x_0,h)\in[0,1]$. Since there exists a positive constant M such that $|\frac{\partial f}{\partial x}|<M$, it holds that

$$|\Delta x|\le |h|+M\int_0^t |\Delta x(s,x_0,h)|\mathrm{d}s,$$

By Gronwall's lemma (Lemma 2.3),

$$|\Delta x(t,x_0,h)|\le |h|e^{MT}.$$

Let $v(t)$ be the unique solution of (2.8). We compute

$$\begin{aligned}\frac{\Delta x(t,x_0,h)}{h}-v(t) &= \int_0^t\left(\frac{f(s,x(s,x_0+h))-f(s,x(s,x_0))}{h}-\frac{\partial f}{\partial x}(s,x(s,x_0))v(s)\right)\mathrm{d}s\\ &= \int_0^t \frac{\Delta x(s,x_0,h)}{h}\left[\frac{\partial f}{\partial x}(s,x(s,x_0)+\tau\Delta x(s,x_0,h))-\frac{\partial f}{\partial x}(s,x(s,x_0))\right]\mathrm{d}s\\ &\quad+\int_0^t \frac{\partial f}{\partial x}(s,x(s,x_0))\left(\frac{\Delta x(s,x_0,h)}{h}-v(s)\right)\mathrm{d}s.\end{aligned}$$

By using the uniform continuity of $\frac{\partial f}{\partial x}$, we have that for any $\epsilon > 0$ there exists $h_0 > 0$ such that, for any $|h| \le h_0$, the first term on the right-hand side is of order $O(\epsilon)$. Then, again by Gronwall's lemma,

$$|\frac{\Delta x(t, x_0, h)}{h} - v| \le \epsilon M T e^{MT},$$

for $|h|$ small enough, which proves that $x_0 \mapsto x(t)$ is differentiable and its derivative is given by

$$\frac{\partial x}{\partial x_0} = v,$$

where v is the solution of (2.8). □

2.4. Stability

THEOREM 2.15 (Strong continuity theorem). *Let*

$$\frac{\mathrm{d}x}{\mathrm{d}t} = f(t, x) \quad \textit{and} \quad \frac{\mathrm{d}y}{\mathrm{d}t} = g(t, y)$$

be two ODEs on $[0, T]$. If f satisfies the Lipschitz condition (2.2) on $[0, T]$ and there exists $\epsilon > 0$ such that, for any $x \in \mathbb{R}^d$, $t \in [0, T]$,

$$|f(t, x) - g(t, x)| \le \epsilon,$$

then the following inequality holds:

$$|x(t) - y(t)| \le |x(0) - y(0)| e^{C_f t} + \frac{\epsilon}{C_f}(e^{C_f t} - 1), \quad t \in [0, T].$$

REMARK 2.16. *The function g may not satisfy a Lipschitz condition.*

PROOF. Since

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}|x(t) - y(t)|^2 &= 2(f(t, x(t)) - g(t, y(t)))(x(t) - y(t)) \\ &= 2(f(t, x(t)) - f(t, y(t)))(x(t) - y(t)) + 2(f(t, y(t)) - g(t, y(t)))(x(t) - y(t)),\end{aligned}$$

we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}|x(t) - y(t)|^2 &\le \left|\frac{\mathrm{d}}{\mathrm{d}t}|x(t) - y(t)|^2\right| \\ &\le 2|f(t, x(t)) - f(t, y(t))|\,|x(t) - y(t)| + 2|f(t, y(t)) - g(t, y(t))|\,|x(t) - y(t)| \\ &\le 2C_f\,|x(t) - y(t)|^2 + 2\epsilon|x(t) - y(t)| \\ &\le 2C_f\,|x(t) - y(t)|^2 + 2\epsilon\sqrt{|x(t) - y(t)|^2}.\end{aligned}$$

If we denote by $h(t) := |x(t) - y(t)|^2$, then

$$\frac{\mathrm{d}h}{\mathrm{d}t} \le 2C_f h + 2\epsilon\sqrt{h}.$$

Consider the following initial value problem:

$$\begin{cases} \dfrac{\mathrm{d}u}{\mathrm{d}t} = 2C_f u + 2\epsilon\sqrt{u}, \\ u(0) = |x(0) - y(0)|^2. \end{cases} \tag{2.9}$$

Since $C_f > 0$, $u(0) > 0$, it follows that $\frac{\mathrm{d}u}{\mathrm{d}t}$ is always non-negative when $t \ge 0$, and hence u is increasing.

Let $z(t) := \sqrt{u(t)}$ and suppose that $h(0) > 0$. Then (2.9) is equivalent to

$$\begin{cases} \dfrac{\mathrm{d}z}{\mathrm{d}t} - C_f z = \epsilon, \quad t \in [0, T], \\ z(0) = \sqrt{u(0)}. \end{cases}$$

This gives the solution of (2.4):

$$\sqrt{u(t)} = z(t) = \sqrt{u(0)}e^{C_f t} + \frac{\epsilon}{C_f}(e^{C_f t} - 1).$$

Moreover,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}(h(t) - u(t)) &\le 2C_f(h(t) - u(t)) + 2\epsilon(\sqrt{h(t)} - \sqrt{u(t)}) \\ &= 2C_f(h(t) - u(t)) + 2\epsilon\frac{h(t) - u(t)}{\sqrt{h(t)} + \sqrt{u(t)}} \\ &\le 2C_f(h(t) - u(t)) + 2\epsilon\frac{h(t) - u(t)}{\sqrt{u(0)}} \\ &= (2C_f + \frac{2\epsilon}{\sqrt{u(0)}})(h(t) - u(t)),\end{aligned}$$

which implies

$$\frac{\mathrm{d}}{\mathrm{d}t}\left((h(t) - u(t))\exp(-2C_f + \frac{2\epsilon}{\sqrt{u(0)}})t\right) \le 0.$$

By integrating the last inequality from 0 to t, we obtain

$$(h(t) - u(t))\exp(-2C_f + \frac{2\epsilon}{\sqrt{u(0)}}t) \le h(0) - u(0).$$

Since $u(0) = h(0)$, we have $h(t) \le u(t)$ for $t \in [0, T]$, and hence

$$\begin{aligned}|x(t) - y(t)| &\le \sqrt{u(t)} \\ &= \sqrt{u(0)}e^{C_f t} + \frac{\epsilon}{C_f}(e^{C_f t} - 1) \\ &= \sqrt{h(0)}e^{C_f t} + \frac{\epsilon}{C_f}(e^{C_f t} - 1).\end{aligned}$$

Therefore, the desired estimate

$$|x(t) - y(t)| \le |x(0) - y(0)|e^{C_f t} + \frac{\epsilon}{C_f}(e^{C_f t} - 1)$$

holds.

If $h(0) = 0$, then, instead of (2.9), we consider the following equation:

$$\begin{cases} \dfrac{\mathrm{d}u_n}{\mathrm{d}t} = 2C_f u_n + 2\epsilon\sqrt{u_n}, \quad t \in [0, T], \\ u_n(0) = \dfrac{1}{n}, \end{cases} \tag{2.10}$$

which, analogously to (2.9), has the explicit solution

$$u_n(t) = \left[\frac{1}{\sqrt{n}}e^{C_f t} + \frac{\epsilon}{C_f}(e^{C_f t} - 1)\right]^2.$$

We only need to prove that for each $n \in \mathbb{N}$,

$$h(t) \le u_n(t) \tag{2.11}$$

holds for $t \in [0, T]$. Then by letting $n \to +\infty$, $u_n \to u$, where u is the solution to (2.9), and hence $h(t) \le u(t)$.

Inequality (2.11) can be proved by contradiction. Suppose that there exists $t_1 > 0$ such that $h(t_1) > u_n(t_1)$. Let t_0 be the largest t in the interval $0 < t \le t_1$ such that $h(t_0) \le u_n(t_0)$. By the continuity of $h(t)$ and $u_n(t)$, we assert that

$$h(t_0) = u_n(t_0) > 0,$$

and $h(t) > u_n(t)$ on $(t_0, t_0 + \epsilon)$, a small right-neighborhood of t_0. But this is impossible according to the discussion in the case where $h(0) > 0$. The proof of the theorem is now complete. □

2.5. Regularity

THEOREM 2.17. *If $f \in \mathcal{C}^n$ for $n \geq 0$, then the solution x of (2.3) is of class $\mathcal{C}^{n+1}$.*

PROOF. The proof is by induction, the case $n = 0$ being clear. If $f \in \mathcal{C}^n$ then x is at least of class $\mathcal{C}^n$, by the inductive assumption. Then the function $t \mapsto f(t, x(t)) = dx(t)/dt$ is also of class $\mathcal{C}^n$. The function $x(t)$ is then of class $\mathcal{C}^{n+1}$. □

REMARK 2.18. *If f is a real analytic function, then it can be proved that x is also real analytic.*

CHAPTER 3

Linear systems

3.1. Exponential of a matrix

Let $\mathbb{M}_d(\mathbb{C})$ be the vector space of $d \times d$ matrices with entries in $\mathbb{C}$. Let $GL_d(\mathbb{C}) \subset \mathbb{M}_d(\mathbb{C})$ be the group of invertible matrices.

DEFINITION 3.1 (Matrix norm). *The matrix norm of $A \in \mathbb{M}_d(\mathbb{C})$ is*

$$\|A\| = \max_{|y|=1} |Ay|.$$

LEMMA 3.2. *The matrix norm has the following properties:*

(i) $|Ay| \leq \|A\||y|$ *for all* $y \in \mathbb{C}^d$;
(ii) $\|A + B\| \leq \|A\| + \|B\|$ *for all* $A, B \in \mathbb{M}_d(\mathbb{C})$;
(iii) $\|AB\| \leq \|A\|\,\|B\|$ *for all* $A, B \in \mathbb{M}_d(\mathbb{C})$.

LEMMA 3.3 (Jordan-Chevalley decomposition). *Let $A \in \mathbb{M}_d(\mathbb{C})$. Then there exists $C \in GL_d(\mathbb{C})$ such that A has a unique decomposition*

$$C^{-1}AC = D + N,$$

where D is diagonal, N is nilpotent (i.e., $N^d = 0$), and $ND = DN$.

We now define the **exponential of a matrix**.

DEFINITION 3.4. *For a matrix $A \in \mathbb{M}_d(\mathbb{C})$, we define*

$$e^A = \sum_{n \geq 0} \frac{A^n}{n!}.$$

We list some properties of the exponential of a matrix.

LEMMA 3.5. *The exponential of a matrix has the following properties:*

(i) *(exponential of the sum) Let $A, B \in \mathbb{M}_d(\mathbb{C})$. If $AB = BA$, then $e^{A+B} = e^A e^B$;*
(ii) *(conjugation and exponentiation) Let $A, B \in \mathbb{M}_d(\mathbb{C})$ and $C \in GL_d(\mathbb{C})$ be such that $A = C^{-1}BC$. Then we have*

$$e^A = C^{-1} e^B C.$$

In fact,

$$e^A = \sum_{n \geq 0} \frac{A^n}{n!} = \sum_{n \geq 0} \frac{(C^{-1}BC)^n}{n!} = \sum_{n \geq 0} \frac{C^{-1}B^nC}{n!} = C^{-1}e^B C;$$

(iii) *(exponential of a diagonalizable matrix) If A is a diagonalizable matrix of the form*

$$A = C^{-1} \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_d \end{pmatrix} C,$$

where $\lambda_1, \ldots, \lambda_d \in \mathbb{C}$ and $C \in GL_d(\mathbb{C})$, then

$$e^A = C^{-1} \begin{pmatrix} e^{\lambda_1} & & 0 \\ & \ddots & \\ 0 & & e^{\lambda_d} \end{pmatrix} C;$$

(iv) *(exponential of a block matrix) Let $A_j \in \mathbb{M}_{h_j}(\mathbb{C})$ for $j = 1, ..., p$. Let A be a block matrix of the form*

$$A = \begin{pmatrix} A_1 & & 0 \\ & \ddots & \\ 0 & & A_p \end{pmatrix}.$$

Then

$$e^A = \begin{pmatrix} e^{A_1} & & 0 \\ & \ddots & \\ 0 & & e^{A_p} \end{pmatrix};$$

(v) *(derivative) Let $A \in \mathbb{M}_d(\mathbb{C})$. We have*

$$\frac{d}{dt} e^{tA} = Ae^{tA} = e^{tA} A.$$

3.2. Linear systems with constant coefficients

Let $A \in \mathbb{M}_d(\mathbb{C}))$ be independent of t. Let $f \in \mathcal{C}^0([0, T])$. Consider the following linear ODE with constant coefficients:

$$\begin{cases} \dfrac{dx}{dt} = Ax(t) + f(t), \quad t \in [0, T], \\ x(0) = x_0 \in \mathbb{R}^d. \end{cases} \tag{3.1}$$

Since

$$|A(x - y)| \leq \|A\||x - y| \quad \text{for all } x, y \in \mathbb{C}^d,$$

by the Cauchy-Lipschitz theorem there exists a unique solution x to (3.1). The system of equations (3.1) is an **autonomous** system.

If $d = 1$ (i.e., $A = a \in \mathbb{C}$), then by the method of integrating factors,

$$x(t) = e^{at} x_0 + \int_0^t e^{a(t-s)} f(s) ds. \tag{3.2}$$

In the general case ($d \geq 1$), if $f = 0$, then, from Lemma 3.5 (v), it follows that the solution x of (3.1) is $x(t) = e^{tA} x_0$.

For an arbitrary f, we have

$$\frac{d}{dt}(e^{-tA} x) = e^{-tA} f(t),$$

and hence the solution $x(t)$ of (3.1) is given by

$$x(t) = e^{tA} x_0 + \int_0^t e^{(t-s)A} f(s) ds. \tag{3.3}$$

Observe that the solution of (3.1) has been reduced in (3.3) to matrix calculations and integration.

3.3. Linear system with non-constant real coefficients

3.3.1. The homogeneous case. Let $\mathbb{M}_d(\mathbb{R})$ be the vector space of $d \times d$ matrices with entries in $\mathbb{R}$.

PROPOSITION 3.6. *Let $A : [0, T] \to \mathbb{M}_d(\mathbb{R})$ be continuous. The set S of solutions of $dx/dt = A(t)x$ defined by*

$$S = \left\{ x \in \mathcal{C}^1([0, T]; \mathbb{R}^d) : x \text{ satisfies } \frac{dx}{dt} = A(t)x \right\} \tag{3.4}$$

is a linear subspace of $\mathcal{C}^1([0, T]; \mathbb{R}^d)$ of dimension d.

PROOF. If $x, y \in S$, then, for any $\alpha, \beta \in \mathbb{R}$, $\alpha x + \beta y \in \mathcal{C}^1([0,T];\mathbb{R}^d)$ is also a solution. Then S is a linear subspace of $\mathcal{C}^1([0,T];\mathbb{R}^d)$. We show that the dimension of S is d. Let the mapping $F : S \to \mathbb{R}^d$ be defined by

$$F[x] = x(t_0) \tag{3.5}$$

for some $t_0 \in [0,T]$. Then F is linear: $F[\alpha x + \beta y] = \alpha x(t_0) + \beta y(t_0) = \alpha F[x] + \beta F[y]$. F is injective, i.e., $F[x] = 0$ implies that $x = 0$. In fact, x solves $\frac{dx}{dt} = A(t)x(t)$ with the initial condition $x(t_0) = 0$. The solution to this problem is unique (by the Cauchy-Lipschitz theorem) and 0 is a solution. Then $x = 0$. Finally, F is surjective because for any $x_0 \in \mathbb{R}^d$ the equation

$$\begin{cases} \dfrac{dx}{dt} = A(t)x(t), \quad t \in [0,T], \\ x(t_0) = x_0, \end{cases} \tag{3.6}$$

has a solution $x \in \mathcal{C}^1([0,T];\mathbb{R}^d)$.

□

PROPOSITION 3.7. *Let S be defined by (3.4) and let $x_1, \dots, x_d \in S$. The following statements are equivalent:*

(i) *$\{x_1, ..., x_d\}$ is a basis of S;*
(ii) *$\det[x_1(t), ..., x_d(t)] \neq 0$ for all $t \in [0,T]$.*
(iii) *$\det[x_1(t_0), ..., x_d(t_0)] \neq 0$ for some $t_0 \in [0,T]$.*

Here, det *denotes the determinant of a matrix and $[x_1, \dots, x_d]$ is the $d \times d$ matrix with columns $x_1, \dots, x_d \in \mathbb{R}^d$.*

PROOF. It is clear that (i) is equivalent to (ii). To see that (i) implies (iii), let $\{x_1, ..., x_d\}$ be a basis of S. Then $\{F[x_1], ..., F[x_d]\}$ forms a basis of $\mathbb{R}^d$, where the isomorphism F relative to t_0 is defined by (3.5). Next let us check that (iii) implies (i). Let t_0 be such that (iii) holds and let $F : S \to \mathbb{R}^d$ be the isomorphism relative to t_0 defined by (3.5). Then the inverse $F^{-1} : \mathbb{R}^d \to S$ is also an isomorphism. It follows that $x_1 = F^{-1}[x_1(t_0)], \dots, x_d = F^{-1}[x_d(t_0)]$ is a basis of S. □

DEFINITION 3.8 (**Fundamental matrix**). *If one of the three equivalent conditions of Proposition 3.7 holds, then the functions $x_1, \dots, x_d$ are called a **fundamental system** of solutions of the differential equation $\frac{dx}{dt} = A(t)x$. The matrix $X = [x_1, \dots, x_d]$ is then called a **fundamental matrix** of the equation.*

We now introduce the **Wronskian determinant**.

DEFINITION 3.9 (**Wronskian determinant**). *Let $x_1, ..., x_d \in S$. The Wronskian determinant $w \in \mathcal{C}^1([0,T];\mathbb{R})$ of $x_1, \dots, x_d$ is defined by*

$$w(t) = \det[x_1(t), ..., x_d(t)].$$

THEOREM 3.10. *Let $x_1, ..., x_d \in S$ and let $w \in \mathcal{C}^1([0,T];\mathbb{R}^d)$ be the Wronskian determinant of $x_1, \dots, x_d$. Then w solves the differential equation*

$$\frac{dw}{dt} = (\mathrm{tr}A(t))w \quad \text{for } t \in [0,T]. \tag{3.7}$$

Here, tr *denotes the trace of a matrix.*

PROOF. If $x_1, ..., x_d$ are linearly dependent, then $w = 0$ and (3.7) trivially holds. Suppose that $x_1, ..., x_d$ are linearly independent, i.e., $w(t) \neq 0$ for all $t \in [0,T]$.

Let $X : [0,T] \to \mathbb{M}_d(\mathbb{R})$ be the fundamental matrix having as columns the solutions $x_1, ..., x_d$, i.e.,

$$X(t) = (x_{ij}(t))_{i,j=1,\dots,d}, \quad t \in [0,T],$$

where $x_j = (x_{1j}, ..., x_{dj})^\top$ for $j = 1, ..., d$.

Let z_j be the solution of

$$\begin{cases} \dfrac{dz_j}{dt} = A(t)z_j(t), \\ z_j(t_0) = e_j, \end{cases}$$

where $\{e_j\}_{j=1,\dots,d}$ is the standard unit orthonormal basis in $\mathbb{R}^d$.

Then $\{z_1, \dots, z_d\}$ is a basis of the space of solutions to $dz/dt = Az$. Moreover, there exists $C \in GL_d(\mathbb{R}^d)$ such that

$$X(t) = CZ(t), \quad t \in [0, T],$$

where $Z = [z_1, \dots, z_d]$.

Let $v(t) := \det Z(t)$. Then v solves

$$\frac{dv}{dt}(t_0) = \mathrm{tr}A(t_0).$$

In fact, by the definition of the determinant of a matrix, we have

$$\frac{dv}{dt}(t) = \frac{d}{dt} \sum_{\sigma \in S_d} (-1)^{\mathrm{sgn}\,\sigma} \prod_{i=1}^{d} z_{i\sigma(i)}(t) = \sum_{\sigma \in S_d} (-1)^{\mathrm{sgn}\,\sigma} \sum_{j=1}^{d} \frac{d}{dt} z_{j\sigma(j)}(t) \prod_{i \neq j} z_{i\sigma(i)}(t),$$

where S_d is the set of all permutations of the d elements $\{1, 2, \dots, d\}$ and $\mathrm{sgn}\,\sigma$ is the signature of the permutation σ. Note that

$$\prod_{i \neq j} z_{i\sigma(i)}(t_0) = 0 \quad \text{unless } \sigma = \text{identity},$$

and

$$\begin{aligned} \frac{dz_{jj}}{dt}(t_0) &= (A(t_0)z_j(t_0))_j \\ &= \sum_{h=1}^{d} a_{jh}(t_0) z_{hj}(t_0) = \sum_{h=1}^{d} a_{jh}(t_0)\delta_{hj}(t_0) \\ &= a_{jj}(t_0). \end{aligned}$$

Therefore,

$$\frac{dv}{dt}(t_0) = \sum_{j=1}^{d} a_{jj}(t_0) = \mathrm{tr}A(t_0).$$

Now the general result follows from the differentiation of the following identity:

$$w = \det X = \det(CZ) = (\det C) \det Z = (\det C)v.$$

In fact, we have

$$\frac{dw}{dt}(t_0) = (\det C)\frac{dv}{dt}(t_0) = (\det C)\mathrm{tr}A(t_0).$$

Therefore,

$$\frac{dw}{dt}(t_0) = \mathrm{tr}A(t_0)w(t_0),$$

since $v(t_0) = 1$. □

REMARK 3.11. *Let $t_0 \in [0, T]$. From (3.7), it follows that*

$$w(t) = w(t_0)e^{\int_{t_0}^{t} \mathrm{tr}A(s)\,ds} \quad \textit{for } t \in [0, T]. \tag{3.8}$$

*This is known as **Abel's identity** or **Liouville's formula**. Identity (3.8) shows that it suffices to check that the determinant of the fundamental matrix is nonzero for one $t_0 \in [0, T]$.*

3.3.2. The inhomogeneous case. Consider the inhomogeneous linear differential equation of the form

$$\begin{cases} \dfrac{dx}{dt} = A(t)x + f(t), \end{cases} \tag{3.9}$$

where $A(t) \in \mathcal{C}^0([0,T]; \mathbb{M}_d(\mathbb{R}))$ and $f \in \mathcal{C}^0([0,T]; \mathbb{R}^d)$.

Let X be a fundamental matrix for the homogeneous equation $dx(t)/dt = A(t)x(t)$, i.e.,

$$\frac{dX}{dt} = AX \quad \text{and} \quad \det X \neq 0 \quad \text{for all } t \in [0,T].$$

Then, any solution x to the homogeneous equation is of the form

$$x(t) = X(t)c, \quad t \in [0,T], \tag{3.10}$$

for some (column) vector $c \in \mathbb{R}^d$.

By using the method of integrating factors, we look for a solution to (3.9) of the form (3.10) with $c \in \mathcal{C}^1([0,T]; \mathbb{R}^d)$. In this case, we have

$$\frac{dx}{dt} = \frac{dX}{dt}c + X\frac{dc}{dt} = AXc + X\frac{dc}{dt} = Ax + X\frac{dc}{dt},$$

which implies $X\frac{dc}{dt} = f(t)$. Since X is invertible, we obtain

$$\frac{dc}{dt} = X^{-1}f(t).$$

Therefore, we find

$$c(t) = c_0 + \int_0^t X(s)^{-1}f(s)ds,$$

for some $c_0 \in \mathbb{R}^d$.

THEOREM 3.12. *Let X be a fundamental matrix for the homogeneous equation $dx/dt = Ax$. Then, for all $c_0 \in \mathbb{R}^d$, the function*

$$x(t) = X(t)\big(c_0 + \int_0^t X(s)^{-1}f(s)ds\big) \tag{3.11}$$

is a solution to (3.9). Moreover, any solution to (3.9) is of the form (3.11) for some $c_0 \in \mathbb{R}^d$.

PROOF. The first statement is already proved. To prove the second statement, let x_2 be a solution to (3.9). Since

$$\frac{d}{dt}(x_2 - x(t)) = A(x_2 - x),$$

where x is given by (3.11), we get $x_2 - x = Xc_1$ for some $c_1 \in \mathbb{R}^d$ and the claim follows. □

Formula (3.11) is called **Duhamel'sformula**.

3.4. Second order linear equations

Let $d = 1$ and consider the following second order ODE:

$$\frac{d^2x}{dt^2} = f(t, x, \frac{dx}{dt}),$$

for a given scalar function f. The above ODE is linear if f is linear in x and dx/dt, namely,

$$f(t, x, \frac{dx}{dt}) = g(t) - p(t)\frac{dx}{dt} - q(t)x,$$

where g, p, q are (scalar) functions of t but do not depend on x. Then the ODE becomes

$$\frac{d^2x}{dt^2} + p(t)\frac{dx}{dt} + q(t)x = g(t). \tag{3.12}$$

The initial value problem consists of (3.12) together with a pair of initial conditions

$$x(t_0) = x_0, \quad \frac{dx}{dt}(t_0) = x_0', \quad x_0, x_0' \in \mathbb{R}^d. \tag{3.13}$$

The second order ODE (3.12) is called **homogeneous** if $g = 0$ and **inhomogeneous** otherwise. If $p(t)$ and $q(t)$ are constant, then (3.12) is called linear ODE with constant coefficients.

Suppose that

$$p, q \in \mathcal{C}^0([0,T]). \tag{3.14}$$

If the condition (3.14) fails, then the points at which either p or q fail to be continuous are called **singular points**. The following are important examples:

Bessel's equation: $p(t) = \dfrac{1}{t}, q(t) = 1 - \dfrac{\nu}{t^2}, \quad (\text{at } t = 0);$

Legendre's equation: $p(t) = \dfrac{2t}{1-t^2}, q(t) = \dfrac{n(n+1)}{1-t^2}, n \in \mathbb{N} \quad (\text{at } t = \pm 1).$

THEOREM 3.13. *Suppose that $p, q, g \in \mathcal{C}^0([0,T], \mathbb{R}^d)$. Then there exists a unique solution $x(t)$ on $[0,T]$ to (3.12) with the initial conditions (3.13).*

3.4.1. Structure of the general solution. Here we discuss the structure of the general solution to the second order ODE (3.12).

First we consider the homogeneous case. We need the following results regarding the Wronskian determinant.

DEFINITION 3.14. *Two functions x_1 and x_2 on $[0,T]$ are called **linearly independent** if neither of them is a multiple of the other. Otherwise, they are called **linearly dependent**.*

PROPOSITION 3.15. *Let w be the Wronskian determinant given by*

$$w(t) := x_1(t)\frac{dx_2}{dt}(t) - x_2(t)\frac{dx_1}{dt}(t) = \det \begin{pmatrix} x_1 & x_2 \\ \dfrac{dx_1}{dt} & \dfrac{dx_2}{dt} \end{pmatrix}.$$

If $w(t)$ is not zero at some $t_0 \in [0,T]$, then x_1 and x_2 are linearly independent.

PROOF. Let us prove that if x_1 and x_2 are linearly dependent, then $w(t) = 0$ for all $t \in [0,T]$. Suppose that x_1 and x_2 are linearly dependent. Then, with respect to (α_1, α_2), the following system:

$$\begin{cases} \alpha_1 x_1 + \alpha_2 x_2 = 0, \\ \alpha_1 \dfrac{dx_1}{dt} + \alpha_2 \dfrac{dx_2}{dt} = 0, \end{cases} \qquad \text{for all } t \in [0,T],$$

has a non-trivial solution. Therefore,

$$w(t) = \det \begin{pmatrix} x_1 & x_2 \\ \dfrac{dx_1}{dt} & \dfrac{dx_2}{dt} \end{pmatrix} = 0, \qquad \text{for all } t \in [0,T].$$

This completes the proof. □

PROPOSITION 3.16. *If x_1 and x_2 solve (3.12) on $[0,T]$ then $w(t)$ is either identically zero or not equal to zero at any point of $[0,T]$.*

PROOF. We have

$$w'(t) = x_1 \frac{d^2x_2}{dt^2} - x_2 \frac{d^2x_1}{dt^2}.$$

We also have, by the assumption that x_1, x_2 solve (3.12), that

$$\frac{d^2x_i}{dt^2} = -p(t)\frac{dx_i}{dt} - q(t)x_i, \quad i = 1, 2.$$

So we get

$$\frac{dw}{dt} = -p(t)(x_1\frac{dx_2}{dt} - \frac{dx_1}{dt}x_2) = -p(t)w(t).$$

Therefore $w(t) = w(t_0)e^{-\int_{t_0}^{t} p(s)ds}$, which is either identically zero or never vanishes depending on $w(t_0)$. □

Now we discuss the structure of the general solution to the homogeneous system.

THEOREM 3.17. *Suppose that x_1 and x_2 solve the equation (3.12) with $g = 0$. Suppose also that x_1 and x_2 are linearly independent. Then the general solution is of the form $c_1x_1 + c_2x_2$, where c_1 and c_2 are constant coefficients.*

PROOF. Let $\tilde{x}$ be an arbitrary solution with the initial condition $\tilde{x}(t_0) = \tilde{x}_0, d\tilde{x}/dt(t_0) = \tilde{x}_0'$. Consider the system of equations for (c_1, c_2)

$$\begin{cases} c_1x_1(t_0) + c_2x_2(t_0) = \tilde{x}_0, \\ c_1\frac{dx_1}{dt}(t_0) + c_2\frac{dx_2}{dt}(t_0) = \tilde{x}_0'. \end{cases}$$

Since $x_1\frac{dx_2}{dt} - x_2\frac{dx_1}{dt} \neq 0$ at $t = t_0$, there exists a unique nontrivial solution $(c_1, c_2) = (\tilde{c}_1, \tilde{c}_2)$ to the above system. Then, by the existence and uniqueness theorem for the initial value problem of the second order ODE, we conclude that $\tilde{c}_1x_1 + \tilde{c}_2x_2 = \tilde{x}$. □

3.4.2. Linear n-th order ODE with constant coefficients. Here we discuss the approach to solving a linear n-th order ODE with constant coefficients. Consider

$$\frac{d^nx}{dt^n} + a_{n-1}\frac{d^{n-1}x}{dt^{n-1}} + ... + a_1\frac{dx}{dt} + a_0x = 0, \tag{3.15}$$

where $a_i \in \mathbb{R}$ for $i = 0, ..., n-1$.

The general solution has the form

$$x(t) = c_1x_1 + ... + c_nx_n,$$

where $\{x_i\}_{i=1}^n$ is the set of linearly independent solutions (a fundamental set of solutions) and c_i are constant coefficients.

Let $W(t)$ be the Wronskian determinant of the set $\{x_1, ..., x_n\}$, i.e.,

$$W(t) = \det\begin{bmatrix} x_1 & x_2 & ... & x_n \\ \frac{dx_1}{dt} & & & \\ \vdots & & & \vdots \\ \frac{d^{n-1}}{dt^{n-1}}x_1 & \frac{d^{n-1}}{dt^{n-1}}x_2 & ... & \frac{d^{n-1}}{dt^{n-1}}x_n \end{bmatrix}.$$

If $w(t_0) \neq 0$ for some t_0, then $(x_1, ..., x_n)$ forms a fundamental set of solution.

We solve the equation through the characteristic equation

$$\lambda^n + a_{n-1}\lambda^{n-1} + ... + a_1\lambda + a_0 = 0. \tag{3.16}$$

This equation is derived by guessing a solution $x(t)$ has the form $e^{\lambda t}$ with $\lambda \in \mathbb{C}$.

The characteristic equation (3.16) has n complex roots $\hat{\lambda}_j$ counted with their multiplicities l_j. In other words, equation (3.16) can be rewritten in the form

$$\prod_{j=1}^{m}(\lambda - \hat{\lambda}_j)^{l_j} = 0$$

with $\sum_{j=1}^m l_j = n$. In fact, the general solution $x(t)$ is a linear combination of $t^ke^{\hat{\lambda}_j t}$ for $0 \leq k < l_j$ and $j = 1, \ldots, m$. In particular, if $m = n$, then $x(t)$ is a linear combination of $e^{\hat{\lambda}_j t}$.

THEOREM 3.18. *Let $\hat{\lambda}_j, 1 \le j \le m$, be the zeros of the characteristic polynomial (3.16) associated with (3.15) and let l_j be the corresponding multiplicities. Then the functions*

$$x_{j,k}(t) = t^k e^{\hat{\lambda}_j t}, \quad 0 \le k < l_j, \quad 1 \le j \le m, \tag{3.17}$$

are n linearly independent solutions of (3.15). In particular, any other solution can be written as a linear combination of these solutions.

3.4.3. Reduction of order. Here we discuss a method for finding a second solution to the homogeneous second order ODE when a first solution is known by reducing the order.

Suppose that x_1 a solution of (3.12). Let

$$x(t) = v(t)x_1(t).$$

Then

$$\frac{dx}{dt}(t) = \frac{dv}{dt}x_1 + v\frac{dx_1}{dt}$$

and

$$\frac{d^2x}{dt^2}(t) = \frac{d^2v}{dt^2}x_1 + 2\frac{dv}{dt}\frac{dx_1}{dt} + v\frac{d^2x_1}{dt^2}.$$

So, we get

$$\frac{d^2v}{dt^2} + (p + 2\frac{(dx_1/dt)}{x_1})\frac{dv}{dt} = 0.$$

By letting $u = dv/dt$, the equation above can be rewritten as a first order ODE

$$\frac{du}{dt} + (p + 2\frac{(dx_1/dt)}{x_1})u = 0.$$

Therefore,

$$u(t) = ce^{-\int^t (p+2\frac{(dx_1/dt)}{x_1})ds} = \frac{c}{(x_1(t))^2}e^{-\int^t p(s)ds}. \tag{3.18}$$

Since $v = \int^t u(s)ds$, we get

$$x(t) = x_1(t)\int^t u(s)ds. \tag{3.19}$$

In conclusion, if one solution to (3.12) is known, then a second solution can be found and it is expressed by (3.19), where u is given by (3.18).

CHAPTER 4

Numerical solution of ordinary differential equations

4.1. Introduction

This chapter is concerned with the numerical solution of initial value problems for systems of ordinary differential equations. Since there is no hope of solving the vast majority of differential equations in explicit and analytic form, the design of suitable numerical schemes for accurately approximating solutions is essential. Explicit solutions, when they are known, can also be used as test cases for tracking the reliability and accuracy of a chosen numerical scheme. In this chapter, we survey the most basic numerical methods for solving initial value problems. It goes without saying that some equations are more difficult to accurately approximate than others, and a variety of more specialized techniques are employed when confronted with a recalcitrant system. However, all of the more advanced developments build on the basic schemes and ideas laid out in this chapter.

4.2. The general explicit one-step method

4.2.1. Consistency, stability and convergence. Consider the initial value problem

$$\begin{cases} \dfrac{\mathrm{d}x}{\mathrm{d}t} = f(t,x), \quad t \in [0,T], \\ x(0) = x_0, \quad x_0 \in \mathbb{R}, \end{cases} \tag{4.1}$$

where $f \in \mathcal{C}^0([0,t] \times \mathbb{R})$ satisfies the Lipschitz condition (2.2).

Starting at the initial time $t = 0$, we introduce successive discretization points

$$t_0 = 0 < t_1 < t_2 < \dots,$$

continuing on until we reach the final time T. To keep the analysis as simple as possible, we use a uniform **step size**, and so

$$\Delta t := t_{k+1} - t_k > 0, \tag{4.2}$$

does not dependent on k and is assumed to be relatively small, with $t_k = k\Delta t$. We also suppose that $K = T/(\Delta t)$ is an integer.

A general **explicit one-step method** may be written in the form:

$$x^{k+1} = x^k + \Delta t\, \Phi(t_k, x^k, \Delta t), \tag{4.3}$$

for some continuous function $\Phi(t,x,h)$. In (4.3), taking in succession $k = 0, 1, \dots, K-1$, **one-step** at a time, the approximate values x^k of x at t_k can be easily obtained. Scheme (4.3) is called **explicit** since x^{k+1} is obtained from x^k. x^{k+1} appears only on the left-hand side of (4.3).

We define the **truncation error** of the numerical scheme (4.3) by

$$T_k(\Delta t) = \frac{x(t_{k+1}) - x(t_k)}{\Delta t} - \Phi(t_k, x(t_k), \Delta t). \tag{4.4}$$

As $\Delta t \to 0, k \to +\infty$, $k\Delta t = t$,

$$T_k(\Delta t) \to \frac{dx}{dt} - \Phi(t,x,0).$$

Definition 4.1 (**Consistency**). *The numerical scheme (4.3) is **consistent** with (4.1) if*

$$\Phi(t,x,0) = f(t,x) \quad \textit{for all } t \in [0,T] \textit{ and } x \in \mathbb{R}.$$

DEFINITION 4.2 (**Stability**). *The numerical scheme (4.3) for solving (4.1) is* ***stable*** *if Φ is Lipschitz continuous in x, i.e., there exist positive constants C_Φ and h_0 such that*

$$|\Phi(t,x,h)-\Phi(t,y,h)|\le C_\Phi|x-y|,\quad t\in[0,T], h\in[0,h_0], x,y\in\mathbb{R}. \tag{4.5}$$

Define **global error** the of the numerical scheme (4.3) by

$$e_k = x^k - x(t_k). \tag{4.6}$$

DEFINITION 4.3 (**Convergence**). *The numerical scheme (4.3) for solving (4.1) is* ***convergent*** *if*

$$|e_k|\to 0 \quad \textit{as } \Delta t\to 0, k\to+\infty, k\Delta t = t\in[0,T].$$

THEOREM 4.4 (Dahlquist-Lax equivalence theorem). *The numerical scheme (4.3) is convergent if and only if it is consistent and stable.*

PROOF. From (4.1), it follows that

$$x(t_{k+1})-x(t_k)=\int_{t_k}^{t_{k+1}} f(s,x(s))\,ds,$$

which gives

$$x(t_{k+1})-x(t_k)=(\Delta t)f(t_k,x(t_k))+\int_{t_k}^{t_{k+1}}\big[f(s,x(s))-f(t_k,x(t_k))\big]\,ds.$$

Therefore,

$$\left|x(t_{k+1})-x(t_k)-(\Delta t)f(t_k,x(t_k))\right| = \left|\int_{t_k}^{t_{k+1}}\big[f(s,x(s))-f(t_k,x(t_k))\big]\,ds\right| \le (\Delta t)\,\omega_1(\Delta t), \tag{4.7}$$

where

$$\omega_1(\Delta t):=\sup\big\{|f(t,x(t))-f(s,x(s))|, 0\le s,t\le T, |t-s|\le\Delta t\big\}. \tag{4.8}$$

Note that $\omega_1(\Delta t)\to 0$ as $\Delta t\to 0$. Moreover, if f is Lipschitz in t, then $\omega_1(\Delta t)=O(\Delta t)$.

From (4.3) and

$$e_{k+1}-e_k = x^{k+1}-x^k-(x(t_{k+1})-x(t_k)),$$

we obtain

$$e_{k+1}-e_k=\Delta t\,\Phi(t_k,x^k,\Delta t)-(x(t_{k+1})-x(t_k)),$$

or equivalently,

$$e_{k+1}-e_k=\Delta t\,\big[\Phi(t_k,x^k,\Delta t)-f(t_k,x(t_k))\big]-\big[x(t_{k+1})-x(t_k)-\Delta t\,f(t_k,x(t_k))\big].$$

Write

$$\begin{aligned} e_{k+1}-e_k &= \Delta t\,\big[\Phi(t_k,x^k,\Delta t)-\Phi(t_k,x(k),\Delta t)+\Phi(t_k,x(k),\Delta t)\\ &\quad -f(t_k,x(t_k))\big]-\big[x(t_{k+1})-x(t_k)-\Delta t\,f(t_k,x(t_k))\big]. \end{aligned} \tag{4.9}$$

Let

$$\omega_2(\Delta t):=\sup\big\{|\Phi(t,x,h)-f(t,x)|, t\in[0,T], x\in\mathbb{R}, 0<h\le(\Delta t)\big\}. \tag{4.10}$$

Since the numerical scheme is consistent,

$$\left|\Phi(t_k,x^k,\Delta t)-f(t_k,x^k)\right|\le\omega_2(\Delta t)\to 0 \text{ as } \Delta t\to 0. \tag{4.11}$$

On the other hand, from the stability condition (4.5), it follows that

$$\left|\Phi(t_k,x^k,\Delta t)-\Phi(t_k,x(k),\Delta t)\right|\le C_\Phi|e_k|. \tag{4.12}$$

Combining (4.7), (4.9), (4.11), and (4.12) yields

$$|e_{k+1}|\le\big(1+C_\Phi\Delta t\big)|e_k|+\Delta t\omega_3(\Delta t),\quad 0\le k\le K-1, \tag{4.13}$$

where $K = T/(\Delta t)$ and $\omega_3(\Delta t) := \omega_1(\Delta t) + \omega_2(\Delta t) \to 0$ as $\Delta t \to 0$. By induction, we deduce from (4.13) that

$$|e_{k+1}| \leq (1 + C_\Phi \Delta t)^k |e_0| + (\Delta t)\, \omega_3(\Delta t) \sum_{l=0}^{k-1} (1 + C_\Phi \Delta t)^l, \quad 0 \leq k \leq K. \tag{4.14}$$

Estimate (4.14) together with

$$\sum_{l=0}^{k-1} (1 + C_\Phi \Delta t)^l = \frac{(1 + C_\Phi \Delta t)^k - 1}{C_\Phi \Delta t},$$

and

$$(1 + C_\Phi \Delta t)^K \leq (1 + C_\Phi \frac{T}{K})^K \leq e^{C_\Phi T},$$

yields

$$|e_k| \leq e^{C_\Phi T} |e_0| + \frac{e^{C_\Phi T} - 1}{C_\Phi} \omega_3(\Delta t). \tag{4.15}$$

Therefore, if $e_0 = 0$, then as $\Delta t \to 0, k \to +\infty$ such that $k\Delta t = t \in [0, T]$

$$\lim_{k \to +\infty} |e_k| = 0,$$

which shows that the scheme is in fact convergent. □

Definition 4.5. *An explicit one-step method is said to be of* ***order*** *p if there exist positive constants h_0 and C such that*

$$|T_k(\Delta t)| \leq C(\Delta t)^p, \quad 0 < \Delta t \leq h_0, k = 0, \ldots, K-1,$$

where the truncation error $T_k(\Delta t)$ is defined by (4.4).

If the explicit one-step method is stable, then the global error is bounded by the truncation error.

Proposition 4.6. *Consider the explicit one-step scheme (4.3), where Φ satisfies the stability condition (4.5). Suppose that $e_0 = 0$. Then*

$$|e_{k+1}| \leq \frac{(e^{C_\Phi T} - 1)}{C_\Phi} \max_{0 \leq l \leq k} |T_l(\Delta t)| \quad \textit{for } k = 0, \ldots, K-1, \tag{4.16}$$

where the truncation error T_l and the global error e_k are defined by (4.4) and (4.6), respectively.

Proof. From (4.9), we have

$$e_{k+1} - e_k = -(\Delta t) T_k(\Delta t) + (\Delta t) \Big[\Phi(t_k, x^k, \Delta t) - \Phi(t_k, x(t_k), \Delta t) \Big],$$

so we get

$$\begin{aligned} |e_{k+1}| &\leq (1 + C_\Phi(\Delta t))|e_k| + (\Delta t)|T_k(\Delta t)| \\ &\leq (1 + C_\Phi(\Delta t))|e_k| + (\Delta t) \max_{0 \leq l \leq k} |T_l(\Delta t)|. \end{aligned}$$

In exactly the same manner as in the proof of Theorem 4.4, we obtain estimate (4.16). □

4.2.2. Explicit Euler's method. Let $\Phi(t,x,h) = f(t,x)$. The numerical method (4.3) reduces to

$$x^{k+1} = x^k + (\Delta t) f(t, x^k). \tag{4.17}$$

The numerical method (4.17) is called the **explicit Euler scheme**.

THEOREM 4.7. *Consider the initial value problem (4.1). Suppose that f satisfies the Lipschitz condition (2.2) and f is Lipschitz with respect to t. Then the explicit Euler scheme (4.17) is convergent and the global error e_k is of order Δt. If $f \in \mathcal{C}^1$, then (4.17) is of order one.*

PROOF. Since f satisfies the Lipschitz condition (2.2) then the numerical scheme with $\Phi(t,x,h) = f(t,x)$ is stable. Moreover, it is consistent since $\Phi(t,x,0) = f(t,x)$ for all $t \in [0,T]$ and $x \in \mathbb{R}$. Therefore, by Theorem 4.4, (4.17) is convergent. Furthermore, since f is Lipschitz in t, $\omega_1(\Delta t) = O(\Delta t)$, where ω_1 is defined by (4.8). On the other hand, $\omega_2(\Delta t) = 0$, and hence $\omega_3(\Delta t) = O(\Delta t)$, where ω_2 is defined by (4.10) and $\omega_3 = \omega_1 + \omega_2$. Then, from (4.15), we have $|e_k| = O(\Delta t)$ for $1 \le k \le K$. Now if $f \in \mathcal{C}^1$, then from Theorem 2.17 $x \in \mathcal{C}^2$. By using the mean-value theorem, we have

$$\begin{aligned} T_k(\Delta t) &= \frac{1}{\Delta t}\Big(x(t_{k+1}) - x(t_k)\Big) - f(t_k, x(t_k)) \\ &= \frac{1}{\Delta t}\Big(x(t_k) + (\Delta t)\frac{dx}{dt}(t_k) + \frac{(\Delta t)^2}{2}\frac{d^2x}{dt^2}(\tau) - x(t_k)\Big) - f(t_k, x(t_k)) \\ &= \frac{\Delta t}{2}\frac{d^2x}{dt^2}(\tau), \end{aligned} \tag{4.18}$$

for some $\tau \in [t_k, t_{k+1}]$, which shows that (4.17) is of first order. □

REMARK 4.8 (**Round off error effects**). *Theorem 4.7 is true provided the arithmetic in calculating the numerical approximation is perfect, that is, when performing the operations required by (4.17) no errors occur. However computers always round off real numbers. In numerical methods rounding errors become important when the step size Δt is comparable with the precision of the computations. Thus, when running Euler's method (4.17), the best we can do is to compute the solution of the perturbed scheme:*

$$\widetilde{x}^{k+1} = \widetilde{x}^k + \Delta t f(t_k, \widetilde{x}^k) + (\Delta t)\mu^k + \rho^k,$$

where μ^k and ρ^k represent the errors in f and in the assembling, respectively. Assume that $|\mu^k| \le \mu$ and $|\rho^k| \le \rho$ for all k and $f \in \mathcal{C}^1$. Defining $\widetilde{e}^k = x(t_k) - \widetilde{x}^k$, we have

$$|\widetilde{e}^{k+1}| \le (1 + C_f \Delta t)\widetilde{e}^k + (\Delta t)\mu + \rho,$$

and hence

$$|\widetilde{e}^k| \le e^{C_f T}|\widetilde{e}^0| + (\Delta t)e^{C_f T}\int_0^T |\frac{d^2x}{dt^2}|(s)\, ds + \mu(\Delta t)\frac{e^{C_f T}}{C_f} + \rho\frac{T}{\Delta t}e^{C_f T},$$

where C_f is the Lipschitz constant for f.

Introduce

$$\varphi(\Delta t) = \frac{\mu e^{C_f T}}{C_f}\Delta t + \frac{T\rho e^{C_f T}}{\Delta t}.$$

One can see that φ attains its minimum at $\sqrt{\rho C_f T/\mu}$ and diverges for $\Delta t \to 0$. From a practical point of view, it is better to take time steps that are larger than $\sqrt{\rho C_f T/\mu}$.

REMARK 4.9 (**Control of the time step**). *In (4.17) the time step is uniform and is chosen such that the global error $|e_k|$ is smaller than a given tolerance. In view of (4.18) this supposes a good knowledge of the exact solution. An alternative method consists in computing the numerical solution for an arbitrary Δt and then for $2\Delta t$. If the discrepancy between the two numerical solutions is smaller than the tolerance, we keep Δt. If not, we restart the calculations with a smaller step size, say $\Delta t/2$, until we reach the target.*

4.2.3. High-order methods. In general, the order of a numerical solution method governs both the accuracy of its approximations and the speed at which they converge to the true solution as the step size $\Delta t \to 0$. Although the explicit Euler method is simple and easy to implement, it is only a first order scheme as shown in Theorem 4.7, and therefore of limited use. So, the goal is to devise simple numerical methods that enjoy a higher order of accuracy. The higher its order, the more accurate the numerical scheme, and hence the larger the step size that can be used to produce the solution to a desired accuracy. However, this should be balanced with the fact that higher order methods inevitably require more computational effort at each step.

4.2.3.1. *Taylor methods.* The explicit Euler scheme is based on a first order Taylor approximation to the solution. The Taylor expansion of the solution $x(t)$ at the discretization points t_{k+1} has the form

$$x(t_{k+1}) = x(t_k + \Delta t) = x(t_k) + (\Delta t)\frac{dx}{dt}(t_k) + \frac{(\Delta t)^2}{2}\frac{d^2x}{dt^2}(t_k) + \frac{(\Delta t)^3}{6}\frac{d^3x}{dt^3}(t_k) + \dots. \tag{4.19}$$

We can evaluate the first derivative term by using the differential equation

$$\frac{dx}{dt} = f(t, x). \tag{4.20}$$

The second derivative can be found by differentiating the equation with respect to t. Invoking the chain rule,

$$\frac{d^2x}{dt^2} = \frac{d}{dt}f(t, x) = \frac{\partial f}{\partial t}(t, x) + \frac{\partial f}{\partial x}(t, x)\frac{dx}{dt}. \tag{4.21}$$

Substituting (4.20) and (4.21) into (4.19) and truncating at order $(\Delta t)^2$ leads to the **second order Taylor method**

$$x^{k+1} = x^k + (\Delta t)f(t_k, x^k) + \frac{(\Delta t)^2}{2}\left(\frac{\partial f}{\partial t}(t_k, x^k) + \frac{\partial f}{\partial x}(t_k, x^k)f(t_k, x^k)\right), \tag{4.22}$$

in which we have replaced the solution value $x(t_k)$ by its computed approximation x^k. The resulting method is of second order.

PROPOSITION 4.10. *Suppose that $f \in \mathcal{C}^2$. Then (4.22) is of second order.*

PROOF. If f is of class C^2, then by Theorem 2.17 $x \in \mathcal{C}^3$. Therefore, by using the Taylor expansion (4.19), we obtain that the truncation error T_k is given by

$$T_k(\Delta t) = \frac{(\Delta t)^2}{6}\frac{d^3x}{dt^3}(\tau),$$

for some $\tau \in [t_k, t_{k+1}]$ and so, (4.22) is of second order. □

Higher order Taylor methods are obtained by including further terms in the expansion (4.19). Whereas higher order Taylor methods are easy to motivate, they are rarely used in practice. There are two principal difficulties:

(i) Owing to their dependence upon the partial derivatives of f, f needs to be smooth;
(ii) Efficient evaluation of the terms in the Taylor approximation and avoidance of round off errors are significant concerns.

4.2.3.2. *Integral equation method.* In order to design high-order numerical schemes that avoid the complications inherent in a direct Taylor expansion, we replace the differential equation by an equivalent **integral equation**. The solution $x(t)$ of (4.1) coincides with the solution to the **integral equation**

$$x(t) = x_0 + \int_0^t f(s, x(s))\, ds, \quad t \in [0, T]. \tag{4.23}$$

Starting at the discretization point t_k instead of 0, and integrating until time $t = t_{k+1}$ gives an expression

$$x(t_{k+1}) = x(t_k) + \int_{t_k}^{t_{k+1}} f(s, x(s))\, ds, \tag{4.24}$$

that implicitly computes the value of the solution at the subsequent discretization point. Comparing formula (4.24) with the explicit Euler method

$$x^{k+1} = x^k + (\Delta t) f(t_k, x^k),$$

where Δt is defined by (4.2) and assuming for the moment that $x^k = x(t_k)$ is exact, we see that we are merely approximating the integral by

$$\int_{t_k}^{t_{k+1}} f(s, x(s))\, ds \approx (\Delta t) f(t_k, x(t_k)), \tag{4.25}$$

which is the **left endpoint rule** for numerical integration–that approximates the integral of $f(t, x(t))$ between $t_k \le t \le t_{k+1}$ by the area of the rectangle whose height $f(t_k, x(t_k))$ is prescribed by the left endpoint of the curve $t \mapsto f(t, x(t))$. Approximation (4.25) is not an especially accurate method of numerical integration. Better methods include the **Trapezoid rule**, which approximates the integral of the function $f(t, x(t))$ between $t_k \le t \le t_{k+1}$ by the area of the trapezoid obtained by connecting the points $f(t_k, x(t_k))$ and $f(t_{k+1}, x(t_{k+1}))$ of the curve $t \mapsto f(t, x(t))$ by a straight line.

We recall the following basic numerical integration formulas for continuous functions.

(i) **Trapezoidal rule**:

$$\int_{t_k}^{t_{k+1}} g(s)\, ds \approx \frac{\Delta t}{2}\Big(g(t_{k+1}) + g(t_k)\Big); \tag{4.26}$$

(ii) **Simpson's rule**:

$$\int_{t_k}^{t_{k+1}} g(s)\, ds \approx \frac{\Delta t}{6}\Big(g(t_{k+1}) + 4g(\frac{t_k + t_{k+1}}{2}) + g(t_k)\Big); \tag{4.27}$$

(iii) The Trapezoidal rule is **exact** for polynomials of order one, while the Simpson's rule is exact for polynomials of second order.

Replacing (4.25) by the more accurate Trapezoidal approximation

$$\int_{t_k}^{t_{k+1}} f(s, x(s))\, ds \approx \frac{(\Delta t)}{2}\Big[f(t_k, x(t_k)) + f(t_{k+1}, x(t_{k+1}))\Big], \tag{4.28}$$

and substituting (4.28) into the integral equation (4.24) leads to the **Trapezoidal scheme**

$$x^{k+1} = x^k + \frac{(\Delta t)}{2}\Big[f(t_k, x^k) + f(t_{k+1}, x^{k+1})\Big]. \tag{4.29}$$

The Trapezoidal scheme is an **implicit numerical method**, since the updated value x^{k+1} appears on both sides of the equation, and hence is only defined implicitly. Only for very simple functions $f(t, x)$ can one expect to solve (4.29) explicitly for x^{k+1} given t_k, x^k, and t_{k+1}.

PROPOSITION 4.11. *Suppose that $f \in \mathcal{C}^2$ and*

$$\frac{(\Delta t) C_f}{2} < 1, \tag{4.30}$$

where C_f is the Lipschitz constant for f in x defined by (2.2). Then the Trapezoidal scheme (4.29) is convergent and is of second order.

PROOF. Let

$$\Phi(t, x, \Delta t) := \frac{1}{2}\Big[f(t, x) + f(t + \Delta t, x + (\Delta t)\Phi(t, x, \Delta t))\Big].$$

The scheme (4.29) is clearly consistent. In order to show that it converges, according to Theorem 4.4, we must establish the stability condition (4.5). We have

$$\big|\Phi(t, x, \Delta t) - \Phi(t, y, \Delta t)\big| \le C_f |x - y| + \frac{\Delta t}{2} C_f \big|\Phi(t, x, \Delta t) - \Phi(t, y, \Delta t)\big|.$$

Hence

$$\big(1 - \frac{(\Delta t)C_f}{2}\big)\big|\Phi(t,x,\Delta t) - \Phi(t,y,\Delta t)\big| \le C_f|x-y|,$$

and therefore, (4.5) holds with

$$C_\Phi = \frac{C_f}{1 - \frac{(\Delta t)C_f}{2}},$$

provided that Δt satisfies (4.30). Now we prove that (4.29) is of second order.

By the mean-value theorem,

$$\begin{aligned} T_k(\Delta t) &= \frac{x(t_{k+1}) - x(t_k)}{\Delta t} - \frac{1}{2}\Big[f(t_k, x(t_k)) + f(t_{k+1}, x(t_{k+1}))\Big] \\ &= -\frac{1}{12}(\Delta t)^2 \frac{d^3x}{dt^3}(\tau), \end{aligned}$$

for some $\tau \in [t_k, t_{k+1}]$, and therefore (4.29) is of second order, provided that $f \in \mathcal{C}^2$ (and consequently $x \in \mathcal{C}^3$). □

An alternative is to replace in (4.29) x^{k+1} by $x^k + (\Delta t)f(t_k, x^k)$. This yields the **improved Euler scheme**

$$x^{k+1} = x^k + \frac{(\Delta t)}{2}\Big[f(t_k, x^k) + f(t_{k+1}, \mathbf{x^k} + (\mathbf{\Delta t})\mathbf{f}(\mathbf{t_k}, \mathbf{x^k}))\Big]. \tag{4.31}$$

Proposition 4.12. *The numerical scheme (4.31) is convergent and is of second order.*

The improved Euler scheme (4.31) performs comparably to the Trapezoidal scheme (4.29), and significantly better than the Euler scheme (4.17). The improved Euler scheme (4.31) is the simplest of a large family of so-called **predictor-corrector algorithms**. In general, one begins by using a relatively crude method–in this case the explicit Euler method–to predict a first approximation $\widetilde{x}^{k+1}$ to the desired solution value $x(t_{k+1})$. One then employs a more sophisticated, typically implicit, method to correct the original prediction, by replacing the required update x^{k+1} on the right-hand side of the implicit scheme by a less accurate prediction $\widetilde{x}^{k+1}$. The resulting explicit, corrected value x^{k+1} will be an improved approximation of the true solution, provided the method has been designed with due care.

We can design a range of numerical solution schemes by implementing alternative numerical approximations to the integral equation (4.24). For example, the **midpoint rule** approximates the integral of $f(t, x(t))$ between $t_k \le t \le t_{k+1}$ by the area of the rectangle whose height is the value of f at the midpoint $t = t_k + (\Delta t)/2$

$$\int_{t_k}^{t_{k+1}} f(s, x(s))\, ds \approx (\Delta t) f(t_k + \frac{\Delta t}{2}, x(t_k + \frac{\Delta t}{2})). \tag{4.32}$$

The midpoint rule has the same order of accuracy as the trapezoid rule. Substituting (4.32) into (4.24) leads to the **midpoint scheme**

$$x^{k+1} = x^k + (\Delta t) f\big(t_k + \frac{\Delta t}{2}, x^k + \frac{\Delta t}{2} f(t_k, x^k)\big), \tag{4.33}$$

where we have approximated $x(t_k + \frac{\Delta t}{2})$ by $x^k + \frac{\Delta t}{2} f(t_k, x^k)$.

A comparison between the terms in the Taylor expansion (4.19) of $x(t_{k+1})$ and (4.33) reveals that the midpoint scheme is also of second order.

4.3. Example of linear systems

Let $A \in \mathbb{M}_d(\mathbb{C})$ be independent of t. Consider the following linear system of ODEs:

$$\begin{cases} \dfrac{dx}{dt} = Ax(t), \quad t \in [0, +\infty[, \\ x(0) = x_0 \in \mathbb{R}^d. \end{cases} \tag{4.34}$$

By Lemma 3.3, there exists $C \in GL_d(\mathbb{C})$ such that

$$C^{-1}AC = D + N,$$

where D is diagonal, N is nilpotent, and $ND = DN$. Let $\lambda_j, j = 1, \ldots, J$ be the (distinct) eigenvalues of A. Let m_j be the multiplicity of λ_j and denote by $E_j = \ker(A - \lambda_j I)^{m_j}$ the characteristic subspace associated with λ_j. We have $\oplus E_j = \mathbb{C}^d$.

The system (4.34) is said to be **stable** if there exists a positive constant C_0 such that

$$|x(t)| \leq C_0|x_0| \quad \text{for all } t \in [0, +\infty[. \tag{4.35}$$

LEMMA 4.13. *The system (4.34) is stable if and only if $\Re\lambda_j < 0$ or $\Re\lambda_j = 0$ and $N|_{E_j} = 0$ for $j = 1, \ldots, J$.*

PROOF. Let $\widetilde{x}(t) = Cx(t)$ and $\widetilde{x}_0 = Cx_0$. By Lemma 3.5,

$$\widetilde{x}(t) = e^{tD+tN}\widetilde{x}_0, \quad t \in [0, +\infty[. \tag{4.36}$$

Since $DN = ND$, (4.36) yields

$$\widetilde{x}(t) = \Big(\sum_{i=0}^{d-1} \frac{(tN)^i}{i!}\Big)e^{tD}\widetilde{x}_0, \quad t \in [0, +\infty[. \tag{4.37}$$

If $\widetilde{x}_0$ belongs to the vector eigenspace associated with the eigenvalue λ_j, then

$$\widetilde{x}(t) = e^{t\lambda_j}\Big(\sum_{i=0}^{d-1} \frac{(tN)^i}{i!}\Big)\widetilde{x}_0, \quad t \in [0, +\infty[. \tag{4.38}$$

Therefore, $x(t)$ satisfies (4.35) for some positive constant C_0 if and only if $\Re\lambda_j < 0$ or $\Re\lambda_j = 0$ and $N|_{E_j} = 0$. □

A one-step numerical scheme for solving (4.34) is said to be **stable** if there exists a positive constant C_0 such that

$$|x^{k+1}| \leq C_0|x^0| \quad \text{for all } k \in \mathbb{N}. \tag{4.39}$$

Consider the following schemes for solving (3.1):

(i) Explicit Euler's scheme

$$x^{k+1} = x^k + (\Delta t)Ax^k; \tag{4.40}$$

(ii) Implicit Euler's scheme

$$x^{k+1} = x^k + (\Delta t)Ax^{k+1}; \tag{4.41}$$

(iii) Trapezoidal scheme:

$$x^{k+1} = x^k + \frac{(\Delta t)}{2}\Big[Ax^k + Ax^{k+1}\Big], \tag{4.42}$$

where $k \in \mathbb{N}$, and $x^0 = x_0$.

PROPOSITION 4.14. *Suppose that $\Re\lambda_j < 0$ for all j. The following results hold:*

(i) *The explicit Euler scheme (4.40) is stable for Δt small enough;*
(ii) *The implicit Euler scheme is unconditionally stable;*
(iii) *The Trapezoidal scheme (4.42) is unconditionally stable.*

PROOF. Consider the explicit Euler scheme (4.40). By a change of basis, we have

$$\widetilde{x}^{k+1} = (I + \Delta t(D+N))^k \widetilde{x}^0,$$

where $\widetilde{x}^k = Cx^k$. If $\widetilde{x}^0 \in E_j$, then

$$\widetilde{x}^k = \sum_{l=0}^{\min\{k,d\}} C_k^l (1+\Delta t\lambda_j)^{k-l} (\Delta t)^l N^l \widetilde{x}^0,$$

where C_k^l is the binomial coefficient.

If $|1+(\Delta t)\lambda_j| < 1$, then $\widetilde{x}^k$ is bounded. If $|1+(\Delta t)\lambda_j| > 1$, then one can find $\widetilde{x}^0$ such that $|\widetilde{x}^k| \to +\infty$ (exponentially) as $k \to +\infty$. If $|1+(\Delta t)\lambda_j| = 1$ and $N \neq 0$, then for all $\widetilde{x}^0$ such that $N\widetilde{x}^0 \neq 0, N^2\widetilde{x}^0 = 0$, it can be seen that

$$\widetilde{x}^k = (1+(\Delta t)\lambda_j)^k \widetilde{x}^0 + (1+(\Delta t)\lambda_j)^{k-1} k\Delta t N\widetilde{x}^0$$

goes to infinity as $k \to +\infty$.

The stability condition $|1+(\Delta t)\lambda_j| < 1$ is equivalent to

$$\Delta t < -2\frac{\Re\lambda_j}{|\lambda_j|^2},$$

and therefore holds for Δt small enough.

For the implicit Euler scheme (4.40), we have

$$\widetilde{x}^{k+1} = (I - \Delta t(D+N))^{-k} \widetilde{x}^0.$$

Note that all the eigenvalues of the matrix $(I-\Delta t(D+N))^{-1}$ are of modulus strictly smaller than 1. Therefore, the implicit Euler scheme (4.40) is unconditionally stable.

For the Trapezoidal scheme, we have

$$\widetilde{x}^{k+1} = (I - \frac{(\Delta t)}{2}(D+N))^{-k}(I + \frac{(\Delta t)}{2}(D+N))^k \widetilde{x}^0.$$

Therefore, the stability condition is

$$|1 + \frac{(\Delta t)}{2}\lambda_j| < |1 - \frac{(\Delta t)}{2}\lambda_j|,$$

which holds for all $\Delta t > 0$ since $\Re\lambda_j < 0$.

□

Note that while the explicit and implicit Euler schemes are of order one, the Trapezoidal scheme is of order two.

REMARK 4.15. *If $\Re\lambda_j = 0$ for some j, then the explicit Euler scheme may be unstable for any $\Delta t > 0$. Consider the second order linear equation*

$$\begin{cases} \dfrac{d^2x}{dt^2} + x = 0, \quad t \in [0, +\infty[, \\ x(0) = x_0, \dfrac{dx}{dt}(0) = x_0', \quad x_0, x_0' \in \mathbb{R}^d. \end{cases} \tag{4.43}$$

We first reduce (4.43) to the first order linear equation

$$\begin{cases} \dfrac{dX}{dt} = AX, \quad t \in [0, +\infty[, \\ X(0) = (x_0, x_0')^\top \in \mathbb{R}^{2d}, \end{cases} \tag{4.44}$$

where $X = (x, dx/dt)^\top$ and $A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$. The eigenvalues of A are $\pm i$. Consequently, the explicit Euler scheme is unstable since $|1 \pm \Delta t i| > 1$ for any $\Delta t > 0$. However, the implicit Euler scheme is stable since $|1 \pm \Delta t i|^{-1} < 1$.

4.4. Runge-Kutta methods

The **Runge-Kutta methods** are by far the most popular and powerful general-purpose numerical methods for integrating ordinary differential equations.

The idea behind the Runge-Kutta methods is to evaluate f at carefully chosen values of its arguments, t and x, in order to create an approximation that is as accurate as a higher-order Taylor expansion of $x(t+\Delta t)$ without evaluating derivatives of f. Runge-Kutta schemes are time-stepping schemes that can be derived by matching **multivariable Taylor series expansions** of $f(t,x)$ with the Taylor series expansion of $x(t+\Delta t)$. To find the right values of t and x at which to evaluate f, we need to take a Taylor expansion of f evaluated at these (unknown) values, and then match the resulting numerical scheme to a Taylor series expansion of $x(t+\Delta t)$ around t. Towards this, we state a generalization of Taylor's theorem to functions of two variables.

THEOREM 4.16. *Let $f(t,x) \in \mathcal{C}^{n+1}([0,T]\times\mathbb{R})$. Let $(t_0,x_0) \in [0,T]\times\mathbb{R}$. There exist $t_0 \le \tau \le t$, $x_0 \le \xi \le x$, such that*

$$f(t,x) = P_n(t,x) + R_n(t,x),$$

where $P_n(t,x)$ is the nth Taylor polynomial of f around (t_0,x_0),

$$\begin{aligned} P_n(t,x) &= f(t_0,x_0) + \left[(t-t_0)\frac{\partial f}{\partial t}(t_0,x_0) + (x-x_0)\frac{\partial f}{\partial x}(t_0,x_0)\right] \\ &\quad + \left[\frac{(t-t_0)^2}{2}\frac{\partial^2 f}{\partial t^2}(t_0,x_0) + (t-t_0)(x-x_0)\frac{\partial^2 f}{\partial t\partial x}(t_0,x_0) + \frac{(x-x_0)^2}{2}\frac{\partial^2 f}{\partial x^2}(t_0,x_0)\right] \\ &\quad \ldots + \left[\frac{1}{n!}\sum_{j=0}^{n} C_j^n (t-t_0)^{n-j}(x-x_0)^j \frac{\partial^n f}{\partial t^{n-j}\partial x^j}(t_0,x_0)\right], \end{aligned}$$

and $R_n(t,x)$ is the remainder term associated with $P_n(t,x)$,

$$R_n(t,x) = \frac{1}{(n+1)!}\sum_{j=0}^{n+1} C_j^{n+1}(t-t_0)^{n+1-j}(x-x_0)^j \frac{\partial^{n+1} f}{\partial t^{n+1-j}\partial x^j}(\tau,\xi).$$

We now illustrate the proposed approach in order to obtain a second-order accurate method, that is, its local truncation error is $O((\Delta t)^2)$. This involves matching

$$x + \Delta t f(t,x) + \frac{(\Delta t)^2}{2}\big[\frac{\partial f}{\partial t}(t,x) + \frac{\partial f}{\partial x}(t,x)f(t,x)\big] + \frac{(\Delta t)^3}{6}\frac{d^2}{dt^2}[f(\tau,x)]$$

to

$$x + (\Delta t) f(t+\alpha_1, x+\beta_1),$$

where $\tau \in [t, t+\Delta t]$ and α_1 and β_1 are to be found. After simplifying by removing terms that already match, we see that we only need to match

$$f(t,x) + \frac{(\Delta t)}{2}\big[\frac{\partial f}{\partial t}(t,x) + \frac{\partial f}{\partial x}(t,x)f(t,x)\big] + \frac{(\Delta t)^2}{6}\frac{d^2}{dt^2}[f(t,x)]$$

with $f(t+\alpha_1, x+\beta_1)$ at least up to terms of the order of $O(\Delta t)$, so that the local truncation error will be $O((\Delta t)^2)$. Applying the multivariable version of Taylor's theorem to f, we obtain

$$f(t+\alpha_1,x+\beta_1) = f(t,x) + \alpha_1\frac{\partial f}{\partial t}(t,x) + \beta_1\frac{\partial f}{\partial x}(t,x) + \frac{\alpha_1^2}{2}\frac{\partial^2 f}{\partial t^2}(\tau,\xi) + \alpha_1\beta_1\frac{\partial^2 f}{\partial t\partial x}(\tau,\xi) + \frac{\beta_1^2}{2}\frac{\partial^2 f}{\partial x^2}(\tau,\xi),$$

where $t \le \tau \le t+\alpha_1$ and $x \le \xi \le x+\beta_1$. Hence comparing terms yields

$$\alpha_1 = \frac{\Delta t}{2} \quad \text{and} \quad \beta_1 = \frac{\Delta t}{2} f(t,x).$$

The resulting numerical scheme is therefore the **explicit midpoint method** (4.33), which is the simplest example of a Runge-Kutta method of second order. The **improved Euler method** (4.31) is also another often-used Runge-Kutta method.

The most general Runge-Kutta method takes the form

$$x^{k+1} = x^k + \Delta t \sum_{i=1}^{m} c_i f(t_{i,k}, x_{i,k}), \tag{4.45}$$

where m stands for the number of terms in the method. Each $t_{i,k}$ denotes a point in $[t_k, t_{k+1}]$. The second argument $x_{i,k} \approx x(t_{i,k})$ can be viewed as an approximation to the solution at the point $t_{i,k}$, and so is computed by a similar but simpler formula of the same type. To construct an nth order Runge-Kutta method, we need to take at least $m \geq n$ terms in (4.45).

The best-known Runge-Kutta method is the **fourth-order Runge-Kutta method**, which uses four evaluations of f during each step. The method proceeds as follows:

$$\begin{cases} \kappa_1 := f(t_k, x^k), \\ \kappa_2 := f(t_k + \frac{\Delta t}{2}, x^k + \frac{\Delta t}{2}\kappa_1), \\ \kappa_3 := f(t_k + \frac{\Delta t}{2}, x^k + \frac{\Delta t}{2}\kappa_2), \\ \kappa_4 := f(t_{k+1}, x^k + \kappa_3), \\ x^{k+1} = x^k + \dfrac{(\Delta t)}{6}(\kappa_1 + 2\kappa_2 + 2\kappa_3 + \kappa_4). \end{cases} \tag{4.46}$$

In (4.46), the values of f at the midpoint in time are given four times as much weight as values at the endpoints t_k and t_{k+1}, which is similar to Simpson's rule (4.27) from numerical integration.

4.4.1. Construction of Runge-Kutta methods. In this subsection we first construct Runge-Kutta methods by generalizing **collocation methods**. Then we discuss their consistency, stability, and order.

4.4.1.1. *Collocation methods.* Let $\mathcal{P}_m$ denote the space of real polynomials of degree $\leq m$. Given a set of m **distinct** quadrature points $c_1 < c_2 < \ldots < c_m$ in $\mathbb{R}$, and corresponding data $g_1, \ldots, g_m$, there exists a unique polynomial, called the **interpolating polynomial**, $P(t) \in \mathcal{P}_{m-1}$ satisfying $P(c_i) = g_i, i = 1, \ldots, m$.

Define the ith **Lagrange interpolating polynomial** $l_i(t)$, $i = 1, \ldots, m$, for the set of quadrature points $\{c_j\}$ by

$$l_i(t) := \prod_{j \neq i, j=1}^{m} \frac{t - c_j}{c_i - c_j}.$$

The set of Lagrange interpolating polynomials form a basis of $\mathcal{P}_{m-1}$ and the interpolating polynomial P corresponding to the data $\{g_j\}$ is given by

$$P(t) := \sum_{i=1}^{m} g_i l_i(t). \tag{4.47}$$

Consider first a smooth function g on $[0, 1]$. We can approximate the integral of g on $[0, 1]$ by exactly integrating the Lagrange interpolating polynomial of order $m-1$ based on m **quadrature points** $0 \leq c_1 < c_2 < \ldots < c_m \leq 1$. The data are the values of g at the quadrature points $g_i = g(c_i)$, $i = 1, \ldots, m$.

Define the weights

$$b_i = \int_0^1 l_i(s)\, ds. \tag{4.48}$$

The **quadrature formula** is

$$\int_0^1 g(s)\, ds \approx \int_0^1 P(s)\, ds = \sum_{i=1}^{m} b_i g(c_i),$$

where P is defined by (4.47).

Now let f be a smooth function on $[0,T]$ and let $t_k = k\Delta t$ for $k = 0, \ldots, K = T/(\Delta t)$, be the discretization points in $[0,T]$. The integral $\int_{t_k}^{t_{k+1}} f(s)\,ds$ can be approximated by

$$\int_{t_k}^{t_{k+1}} f(s)\,ds = (\Delta t)\int_0^1 f(t_k + \Delta t\tau)\,d\tau \approx (\Delta t)\sum_{i=1}^m b_i f(t_k + (\Delta t)c_i). \tag{4.49}$$

Next let x be a polynomial of degree m satisfying

$$\begin{cases} x(0) = x_0, \\ \dfrac{dx}{dt}(c_i\Delta t) = F_i, \end{cases} \tag{4.50}$$

where $F_i \in \mathbb{R}, i = 1, \ldots, m$.

From the Lagrange interpolation formula (4.47), it follows that for t in the first time-step interval $[0, \Delta t]$,

$$\frac{dx}{dt}(t) = \sum_{i=1}^m F_i l_i(\frac{t}{\Delta t}). \tag{4.51}$$

Integrating (4.51) over the intervals $[0, c_i\Delta t]$ gives

$$x(c_i\Delta t) = x_0 + (\Delta t)\sum_{j=1}^m F_j \int_0^{c_i} l_j(s)\,ds = x_0 + (\Delta t)\sum_{j=1}^m a_{ij}F_j, \quad i = 1, \ldots, m, \tag{4.52}$$

where

$$a_{ij} := \int_0^{c_i} l_j(s)\,ds. \tag{4.53}$$

Integrating (4.51) over $[0, \Delta t]$ yields

$$x(\Delta t) = x_0 + (\Delta t)\sum_{i=1}^m F_i \int_0^1 l_i(s)\,ds = x_0 + (\Delta t)\sum_{i=1}^m b_i F_i, \tag{4.54}$$

where b_i is defined by (4.48).

Writing $dx/dt = f(x(t))$, we obtain from (4.52) and (4.54) on the first time step interval $[0, \Delta t]$

$$\begin{cases} F_i = f(x_0 + (\Delta t)\sum_{j=1}^m a_{ij}F_j), \quad i = 1, \ldots, m, \\ \\ x(\Delta t) = x_0 + (\Delta t)\sum_{i=1}^m b_i F_i. \end{cases} \tag{4.55}$$

Similarly, we have on $[t_k, t_{k+1}]$

$$\begin{cases} F_{i,k} = f(x(t_k) + (\Delta t)\sum_{j=1}^m a_{ij}F_{j,k}), \quad i = 1, \ldots, m, \\ \\ x(t_{k+1}) = x(t_k) + (\Delta t)\sum_{i=1}^m b_i F_{i,k}. \end{cases} \tag{4.56}$$

In the **collocation method** (4.56), one first solves the coupled nonlinear system to obtain $F_{i,k}$, $i = 1, \ldots, m$, and then computes $x(t_{k+1})$ from $x(t_k)$.

REMARK 4.17. *Since*

$$t^{l-1} = \sum_{i=1}^m c_i^{l-1} l_i(t), \quad t \in [0,1], l = 1, \ldots, m,$$

we have

$$\sum_{i=1}^m b_i c_i^{l-1} = \frac{1}{l}, \quad l = 1, \ldots, m,$$

and

$$\sum_{j=1}^{m} a_{ij} c_j^{l-1} = \frac{c_i^l}{l}, \quad i, l = 1, \ldots, m.$$

4.4.2. Runge-Kutta methods as generalized collocation methods. In (4.56), the coefficients b_i and a_{ij} are defined by certain integrals of the Lagrange interpolating polynomials associated with a chosen set of quadrature nodes c_i, $i = 1, \ldots, m$.

A natural generalization of collocation methods is obtained by allowing the coefficients c_i, b_i, and a_{ij} to take arbitrary values, not necessary related to quadrature formulas. In fact, we no longer assume the c_i to be distinct. However, we should assume that

$$c_i = \sum_{j=1}^{m} a_{ij}, \quad i = 1, \ldots, m. \tag{4.57}$$

The result is the class of Runge-Kutta methods for solving (4.1), which can be written as

$$\begin{cases} F_{i,k} = f(t_{i,k}, x^k + (\Delta t) \sum_{j=1}^{m} a_{ij} F_{j,k}), \\ x^{k+1} = x^k + (\Delta t) \sum_{i=1}^{m} b_i F_{i,k}, \end{cases} \tag{4.58}$$

where $t_{i,k} = t_k + c_i \Delta t$, or equivalently,

$$\begin{cases} x_{i,k} = x^k + (\Delta t) \sum_{j=1}^{m} a_{ij} f(t_{j,k}, x_{j,k}), \\ x^{k+1} = x^k + (\Delta t) \sum_{i=1}^{m} b_i f(t_{i,k}, x_{i,k}). \end{cases} \tag{4.59}$$

Let

$$\kappa_j := f(t + c_j \Delta t, x_j), \tag{4.60}$$

and define the function Φ by

$$\begin{cases} x_i = x + (\Delta t) \sum_{j=1}^{m} a_{ij} \kappa_j, \\ \Phi(t, x, \Delta t) = \sum_{i=1}^{m} b_i f(t + c_i \Delta t, x_i). \end{cases} \tag{4.61}$$

One can see that the scheme (4.59) is a one step method. Moreover, if $a_{ij} = 0$ for $j \geq i$, then (4.59) is explicit.

It is also easy to see that with definition (4.59), explicit Euler's method and Trapezoidal scheme are Runge-Kutta methods. For example, explicit Euler's method (4.17) can be put into the form (4.59) with $m = 1, b_1 = 1, a_{11} = 0$. The Trapezoidal scheme (4.29) has $m = 2, b_1 = b_2 = 1/2, a_{11} = a_{12} = 0, a_{21} = a_{22} = 1/2$. Finally, for the fourth-order Runge-Kutta method (4.46), we have $m = 4, c_1 = 0, c_2 = c_3 = 1/2, c_4 = 1, b_1 = 1/6, b_2 = b_3 = 1/3, b_4 = 1/3, a_{21} = a_{32} = 1/2, a_{43} = 1$, and all the other a_{ij} entries are zero.

4.4.3. Consistency, stability, convergence, and order of Runge-Kutta methods. From (4.61), the Runge-Kutta scheme is consistent if and only if

$$\sum_{j=1}^{m} b_j = 1. \tag{4.62}$$

Let $|A|$ be the matrix defined by $(|a_{ij}|)_{i,j=1}^m$. Let the **spectral radius** $\rho(|A|)$ of the matrix $|A|$ be defined by

$$\rho(|A|) := \max\{|\lambda_j|, \lambda_j \text{ is an eigenvalue of } |A|\}. \tag{4.63}$$

The following stability result holds.

THEOREM 4.18. *Let C_f be the Lipschitz constant for f. Suppose that*

$$(\Delta t) C_f \rho(|A|) < 1. \tag{4.64}$$

Then the Runge-Kutta method (4.59) for solving (4.1) is stable.

PROOF. Let Φ be defined by (4.61). We have

$$\Phi(t, x, \Delta t) - \Phi(t, y, \Delta t) = \sum_{i=1}^m b_i \Big[f(t + c_i \Delta t, x_i) - f(t + c_i \Delta t, y_i) \Big], \tag{4.65}$$

where

$$x_i = x + (\Delta t) \sum_{j=1}^m a_{ij} f(t + c_j \Delta t, x_j), \tag{4.66}$$

and

$$y_i = y + (\Delta t) \sum_{j=1}^m a_{ij} f(t + c_j \Delta t, y_j). \tag{4.67}$$

Subtracting (4.67) from (4.66) yields

$$x_i - y_i = x - y + (\Delta t) \sum_{j=1}^m a_{ij} \Big[f(t + c_j \Delta t, x_j) - f(t + c_j \Delta t, y_j) \Big]. \tag{4.68}$$

Therefore, for $i = 1, \ldots, m$,

$$|x_i - y_i| \le |x - y| + (\Delta t) C_f \sum_{j=1}^m |a_{ij}| |x_j - y_j|, \tag{4.69}$$

where C_f is the Lipschitz constant for f. Let the vectors X and Y be defined by

$$X = \begin{bmatrix} |x_1 - y_1| \\ \vdots \\ |x_m - y_m| \end{bmatrix} \quad \text{and} \quad Y = \begin{bmatrix} |x - y| \\ \vdots \\ |x - y| \end{bmatrix}.$$

From (4.69), it follows that

$$X \le Y + (\Delta t) C_f |A| X, \tag{4.70}$$

and therefore,

$$X \le (I - (\Delta t) C_f |A|)^{-1} Y, \tag{4.71}$$

provided that condition (4.64) holds. Finally, combining (4.65) and (4.71) yields the stability of the Runge-Kutta scheme (4.59). □

By the Dahlquist-Lax equivalence theorem (Theorem 4.4), it follows that the Runge-Kutta scheme (4.59) is convergent provided that (4.62) and (4.64) hold.

In order to establish the order of the Runge-Kutta scheme (4.59), we compute the order as $\Delta t \to 0$ of the truncation error

$$T_k(\Delta t) = \frac{x(t_{k+1}) - x(t_k)}{\Delta t} - \Phi(t_k, x(t_k), \Delta t),$$

where Φ is defined by (4.61). We write

$$T_k(\Delta t) = \frac{x(t_{k+1}) - x(t_k)}{\Delta t} - \sum_{i=1}^m b_i f(t_k + c_i \Delta t, x(t_k) + \Delta t \sum_{j=1}^m a_{ij} \kappa_j).$$

Suppose that f is smooth enough. We have

$$f(t_k+c_i\Delta t, x(t_k)+\Delta t\sum_{j=1}^{m} a_{ij}\kappa_j) = f(t_k,x(t_k))+\Delta t\left[c_i\frac{\partial f}{\partial t}(t_k,x(t_k))+(\sum_{j=1} a_{ij}\kappa_j)\frac{\partial f}{\partial x}(t_k,x(t_k))\right]+O((\Delta t)^2).$$

Suppose that (4.57) holds. Then, from

$$\sum_{j=1} a_{ij}\kappa_j = (\sum_{j=1} a_{ij})f(t_k,x(t_k)) = c_i f(t_k,x(t_k)) + O(\Delta t),$$

it follows that

$$f(t_k+c_i\Delta t, x(t_k)+\Delta t\sum_{j=1}^{m} a_{ij}\kappa_j) = f(t_k,x(t_k))+\Delta t c_i\left[\frac{\partial f}{\partial t}(t_k,x(t_k))+\frac{\partial f}{\partial x}(t_k,x(t_k))f(t_k,x(t_k))\right]+O((\Delta t)^2).$$

Therefore, we obtain the following theorem.

THEOREM 4.19. *Assume that f is smooth enough. Then the Runge-Kutta scheme (4.59) for solving (4.1) is of order 2 provided that the conditions (4.62) and*

$$\sum_{i=1}^{m} b_i c_i = \frac{1}{2} \tag{4.72}$$

hold.

One can prove by higher-order Taylor expansions that the following results hold.

THEOREM 4.20. *Assume that f is smooth enough. Then the Runge-Kutta scheme (4.59) for solving (4.1) is of order 3 provided that the conditions (4.62), (4.72), and*

$$\sum_{i=1}^{m} b_i c_i^2 = \frac{1}{3}, \quad \sum_{i=1}^{m}\sum_{j=1}^{m} b_i a_{ij} c_j = \frac{1}{6} \tag{4.73}$$

hold. It is of order 4 provided that (4.62), (4.72), (4.73), and

$$\sum_{i=1}^{m} b_i c_i^3 = \frac{1}{4}, \quad \sum_{i=1}^{m}\sum_{j=1}^{m} b_i c_i a_{ij} c_j = \frac{1}{8}, \quad \sum_{i=1}^{m}\sum_{j=1}^{m} b_i c_i a_{ij} c_j^2 = \frac{1}{12}, \quad \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{l=1}^{m} b_i a_{ij} a_{jl} c_l = \frac{1}{24} \tag{4.74}$$

hold.

The Runge-Kutta scheme (4.46) satisfies the four conditions (4.62), (4.72), (4.73), and (4.74). Hence, (4.46) is of order 4.

4.5. Multi-step methods

While Runge-Kutta methods present an improvement over Euler's methods in terms of accuracy, this is achieved by investing additional computational effort. For example, the fourth-order method (4.46) involves four function evaluations per step. For comparison, by considering three consecutive points t_{k-1}, t_k, t_{k+1}, integrating the differential equation between t_{k-1} and t_{k+1}, and applying **Simpson's rule** to approximate the resulting integral yields

$$\begin{aligned} x(t_{k+1}) &= x(t_{k-1}) + \int_{t_{k-1}}^{t_{k+1}} f(s,x(s))\,ds \\ &\approx x(t_{k-1}) + \frac{(\Delta t)}{3}\left[f(t_{k-1},x(t_{k-1})) + 4f(t_k,x(t_k)) + f(t_{k+1},x(t_{k+1}))\right], \end{aligned}$$

which leads to the method

$$x^{k+1} = x^{k-1} + \frac{(\Delta t)}{3}\left[f(t_{k-1},x^{k-1}) + 4f(t_k,x^k) + f(t_{k+1},x^{k+1})\right]. \tag{4.75}$$

In contrast with the one-step methods considered in the previous sections where only a single value of x^k was required to compute the next approximation x^{k+1}, in (4.75) we need two preceding values, x^k and x^{k-1} in order to calculate x^{k+1}, and therefore (4.75) is a **two-step method**.

The general n-step method is of the form

$$\sum_{j=0}^{n} \alpha_j x^{k+j} = (\Delta t) \sum_{j=0}^{n} \beta_j f(t_{k+j}, x^{k+j}), \tag{4.76}$$

where the coefficients α_j and β_j are real constants and $\alpha_n \neq 0$.

If $\beta_n = 0$, then x^{k+n} is obtained explicitly from previous values of x^j and $f(t_j, x^j)$, and the n-step method is **explicit**. Otherwise, the n-step method is **implicit**.

EXAMPLE 4.21. (i) *The two-step* ***Adams-Bashforth method***

$$x^{k+2} = x^{k+1} + \frac{(\Delta t)}{2}\left[3f(t_{k+1}, x^{k+1}) - f(t_k, x^k)\right] \tag{4.77}$$

is an example of an ***explicit*** *two-step method;*

(ii) *The three-step Adams-Bashforth method*

$$x^{k+3} = x^{k+2} + \frac{(\Delta t)}{12}\left[23f(t_{k+2}, x^{k+2}) - 16f(t_{k+1}, x^{k+1}) + f(t_k, x^k)\right] \tag{4.78}$$

is an example of an explicit three-step method;

(iii) *The four-step Adams-Bashforth method*

$$x^{k+4} = x^{k+3} + \frac{(\Delta t)}{24}\left[55f(t_{k+3}, x^{k+3}) - 59f(t_{k+2}, x^{k+2}) + 37f(t_{k+1}, x^{k+1}) - 9f(t_k, x^k)\right] \tag{4.79}$$

is an example of an explicit four-step method;

(iv) *The two-step* ***Adams-Moulton method***

$$x^{k+2} = x^{k+1} + \frac{(\Delta t)}{12}\left[5f(t_{k+2}, x^{k+2}) + 8f(t_{k+1}, x^{k+1}) + f(t_k, x^k)\right] \tag{4.80}$$

is an example of an ***implicit*** *two-step method;*

(v) *The three-step Adams-Moulton method*

$$x^{k+3} = x^{k+2} + \frac{(\Delta t)}{24}\left[9f(t_{k+3}, x^{k+3}) + 19f(t_{k+2}, x^{k+2}) - 5f(t_{k+1}, x^{k+1}) - 9f(t_k, x^k)\right] \tag{4.81}$$

is an example of an implicit three-step method.

The construction of general classes of linear multi-step methods is discussed in the next subsection.

4.5.1. Construction of linear multi-step methods. Suppose that $x^k, k \in \mathbb{N}$, is a sequence of real numbers. We introduce the **shift operator** E, the **forward difference operator** Δ_+ and the **backward difference operator** Δ_- by

$$E : x^k \mapsto x^{k+1}, \quad \Delta_+ : x^k \mapsto x^{k+1} - x^k, \quad \Delta_- : x^k \mapsto x^k - x^{k-1}.$$

Since $\Delta_+ = E - I$ and $\Delta_- = I - E^{-1}$, it follows that, for any $n \in \mathbb{N}$,

$$(E - I)^n = \sum_{j=0}^{n} (-1)^j C_j^n E^{n-j},$$

and

$$(I - E^{-1})^n = \sum_{j=0}^{n} (-1)^j C_j^n E^{-j}.$$

Therefore,

$$\Delta_+^n x^k = \sum_{j=0}^{n} (-1)^j C_j^n x^{k+n-j}$$

and

$$\Delta_-^n x^k = \sum_{j=0}^{n} (-1)^j C_j^n x^{k-j}.$$

Now let $y(t) \in \mathcal{C}^\infty(\mathbb{R})$ and let $t_k = k\Delta t, \Delta t > 0$. By applying the Taylor series we find that, for any $s \in \mathbb{N}$,

$$E^s y(t_k) = y(t_k + s\Delta t) = \Big(\sum_{l=0}^{+\infty} \frac{1}{l!} (s\Delta t \frac{\partial}{\partial t})^l y \Big)(t_k) = \big(e^{s(\Delta t)\frac{\partial}{\partial t}} y\big)(t_k),$$

and hence

$$E^s = e^{s(\Delta t)\frac{\partial}{\partial t}}.$$

Thus, formally,

$$(\Delta t)\frac{\partial}{\partial t} = \ln E = -\ln(I - \Delta_-) = \Delta_- + \frac{1}{2}\Delta_-^2 + \frac{1}{3}\Delta_-^3 + \dots \tag{4.82}$$

Therefore, if $x(t)$ is the solution of (4.1), then by using (4.82) we find that

$$(\Delta t) f(t_k, x(t_k)) = \Big(\Delta_- + \frac{1}{2}\Delta_-^2 + \frac{1}{3}\Delta_-^3 + \dots \Big) x(t_k). \tag{4.83}$$

The successive truncation of the infinite series on the right-hand side of (4.83) yields

$$\begin{aligned} &x^k - x^{k-1} = (\Delta t) f(t_k, x^k), \\ &\frac{3}{2}x^k - 2x^{k-1} + \frac{1}{2}x^{k-2} = (\Delta t) f(t_k, x^k), \\ &\frac{11}{6}x^k - 3x^{k-1} + \frac{3}{2}x^{k-2} - \frac{1}{3}x^{k-3} = (\Delta t) f(t_k, x^k), \end{aligned} \tag{4.84}$$

and so on. This gives rise to a class of implicit multi-step methods called **backward differentiation formulas**.

Similarly,

$$E^{-1}((\Delta t)\frac{\partial}{\partial t}) = (\Delta t)\frac{\partial}{\partial t} E^{-1} = -(I - \Delta_-)\ln(I - \Delta_-),$$

and hence,

$$((\Delta t)\frac{\partial}{\partial t}) = -E(I - \Delta_-)\ln(I - \Delta_-) = -(I - \Delta_-)\ln(I - \Delta_-)E. \tag{4.85}$$

Therefore, if $x(t)$ is the solution of (4.1), then we find that

$$(\Delta t) f(t_k, x(t_k)) = \Big(\Delta_- - \frac{1}{2}\Delta_-^2 - \frac{1}{6}\Delta_-^3 + \dots \Big) x(t_{k+1}). \tag{4.86}$$

The successive truncation of the infinite series on the right-hand side of (4.86) yields the following explicit numerical schemes:

$$\begin{aligned} &x^{k+1} - x^k = (\Delta t) f(t_k, x^k), \\ &\frac{1}{2}x^{k+1} - \frac{1}{2}x^{k-1} = (\Delta t) f(t_k, x^k), \\ &\frac{1}{3}x^{k+1} + \frac{1}{2}x^k - x^{k-1} + \frac{1}{6}x^{k-2} = (\Delta t) f(t_k, x^k), \\ &\vdots \end{aligned} \tag{4.87}$$

The first of these numerical scheme is the explicit Euler method, while the second is the explicit mid-point method.

In order to construct further classes of multi-step methods, we define, for $y \in \mathcal{C}^\infty$,

$$D^{-1}y(t_k) = y(t_0) + \int_{t_0}^{t_k} y(s)\,ds,$$

and observe that

$$(E-I)D^{-1}y(t_k) = \int_{t_k}^{t_{k+1}} y(s)\,ds.$$

Now, from

$$(E-I)D^{-1} = \Delta_+ D^{-1} = E\Delta_- D^{-1} = (\Delta t)E\Delta_-((\Delta t)D)^{-1},$$

it follows that

$$(E-I)D^{-1} = -(\Delta t)E\Delta_-\big(\ln(I-\Delta_-)\big)^{-1}. \tag{4.88}$$

Furthermore,

$$(E-I)D^{-1} = E\Delta_- D^{-1} = \Delta_- E D^{-1} = \Delta_-(DE^{-1})^{-1} = (\Delta t)\Delta_-\big((\Delta t)DE^{-1}\big)^{-1}.$$

Thus,

$$(E-I)D^{-1} = -(\Delta t)\Delta_-\bigg((I-\Delta_-)\ln(I-\Delta_-)\bigg)^{-1}. \tag{4.89}$$

By using (4.88) and (4.89), we deduce from

$$x(t_{k+1}) - x(t_k) = \int_{t_k}^{t_{k+1}} f(s,x(s))\,ds = (E-I)D^{-1}f(t_k,x(t_k)),$$

that

$$x(t_{k+1}) - x(t_k) = \begin{cases} -(\Delta t)\Delta_-\big((I-\Delta_-)\ln(I-\Delta_-)\big)^{-1}f(t_k,x(t_k)) \\ -(\Delta t)E\Delta_-\big(\ln(I-\Delta_-)\big)^{-1}f(t_k,x(t_k)), \end{cases} \tag{4.90}$$

where $x(t)$ is the solution of (4.1).

On expanding $\ln(I-\Delta_-)$ into a Taylor series on the right-hand side of (4.90) we find that

$$x(t_{k+1}) - x(t_k) = (\Delta t)\bigg[I + \frac{1}{2}\Delta_- + \frac{5}{12}\Delta_-^2 + \frac{3}{8}\Delta_-^3 + \dots\bigg]f(t_k,x(t_k)), \tag{4.91}$$

and

$$x(t_{k+1}) - x(t_k) = (\Delta t)\bigg[I - \frac{1}{2}\Delta_- - \frac{1}{12}\Delta_-^2 - \frac{1}{24}\Delta_-^3 + \dots\bigg]f(t_{k+1},x(t_{k+1})). \tag{4.92}$$

The successive truncation of (4.91) yields the family (4.79) of (explicit) Adams-Bashforth methods, while similar successive truncation of (4.92) gives rise to the family (4.81) of (implicit) Adams-Moulton methods.

4.5.2. Consistency, stability, and convergence. In this subsection, we introduce the concepts of consistency, stability, and convergence for analyzing linear multi-step methods.

DEFINITION 4.22 (Consistency). *The n-step method (4.76) is* ***consistent*** *with (4.1) if the* ***truncation error*** *defined by*

$$T_k = \frac{\sum_{j=0}^n \big[\alpha_j x(t_{k+j}) - (\Delta t)\beta_j \frac{dx}{dt}(t_{k+j})\big]}{(\Delta t)\sum_{j=0}^n \beta_j}$$

is such that for any $\epsilon > 0$ there exists h_0 for which

$$|T_k| \le \epsilon \quad \text{for } 0 < \Delta t \le h_0$$

and any $(n+1)$ points $\big((t_j,x(t_j)),\dots,(t_{j+n},x(t_{j+n}))\big)$ on any solution $x(t)$.

DEFINITION 4.23 (Stability). *The n-step method (4.76) is* ***stable*** *if there exists a constant C such that, for any two sequences (x^k) and $(\widetilde{x}^k)$ which have been generated by the same formulas but different initial data $x^0, x^1, \dots, x^{k-1}$ and $\widetilde{x}^0, \widetilde{x}^1, \dots, \widetilde{x}^{k-1}$, respectively, we have*

$$|x^k - \widetilde{x}^k| \le C \max\{|x^0 - \widetilde{x}^0|, |x^1 - \widetilde{x}^1|, \dots, |x^{k-1} - \widetilde{x}^{k-1}|\} \tag{4.93}$$

as $\Delta t \to 0$.

THEOREM 4.24 (Convergence). *Suppose that the n-step method (4.76) is consistent with (4.1). The stability condition (4.93) is necessary and sufficient for the convergence. Moreover, if $x \in \mathcal{C}^{p+1}$ and the truncation error $O((\Delta t)^p)$, the the global error $e_k = x(t_k) - x^k$ is also $O((\Delta t)^p)$.*

4.6. Stiff equations and systems

Let $\epsilon > 0$ be a small parameter. Consider the initial value problem

$$\begin{cases} \dfrac{\mathrm{d}x(t)}{\mathrm{d}t} = -\dfrac{1}{\epsilon} x(t), \quad t \in [0, T], \\ x(0) = 1, \end{cases} \tag{4.94}$$

which has an exponential solution $x(t) = e^{-t/\epsilon}$. The explicit Euler method with step size Δt relies on the iterative scheme

$$x^{k+1} = (1 - \frac{\Delta t}{\epsilon}) x^k, \quad x^0 = 1, \tag{4.95}$$

with solution

$$x^k = (1 - \frac{\Delta t}{\epsilon})^k.$$

Since $\epsilon > 0$ the exact solution is exponentially decaying and positive. But now, if $1 - \frac{\Delta t}{\epsilon} < -1$, then the iterates (4.95) grow exponentially fast in magnitude, with alternating signs. In this case, the numerical solution is nowhere close to the true solution. If $-1 < 1 - \frac{\Delta t}{\epsilon} < 0$, then the numerical solution decays in magnitude, but continue to alternate between positive and negative values. Thus, to correctly model the qualitative features of the solution and obtain a numerically accurate solution, we need to choose the step size Δt so as to ensure that $1 - \frac{\Delta t}{\epsilon} > 0$, and hence $\Delta t < \epsilon$.

Equation (4.94) is the simplest example of what is known as a **stiff differential equation**. In general, an equation or system is stiff if it has one or more very rapidly decaying solutions. In the case of the autonomous constant coefficient linear system (4.34), stiffness occurs whenever the coefficient matrix A has an eigenvalues λ_{j_0} with large negative real part: $\Re \lambda_{j_0} \ll 0$, resulting in a very rapidly decaying eigensolution. It only takes one such eigensolution to render the equation stiff, and ruin the numerical computation of even well behaved solutions. Even though the component of the actual solution corresponding to λ_{j_0} is almost irrelevant, as it becomes almost instantaneously tiny, its presence continues to render the numerical solution to the system very difficult. Stiff equations require more sophisticated numerical schemes to integrate.

Most of the numerical methods derived above also suffer from instability due to stiffness of (4.94) for sufficiently small positive ϵ. Interestingly, stability of (4.94) suffices to characterize acceptable step sizes Δt, depending on the size of $-1/\epsilon$, which, in the case of linear systems, is the eigenvalue. Applying the Trapezoidal scheme (4.29) to (4.94) leads to

$$x^{k+1} = x^k - \frac{\Delta t}{2\epsilon}\big(x^k + x^{k+1}\big), \quad x^0 = 1, \tag{4.96}$$

which we solve for

$$x^{k+1} = \frac{1 - \frac{\Delta t}{2\epsilon}}{1 + \frac{\Delta t}{2\epsilon}} x^k, \quad x^0 = 1. \tag{4.97}$$

Thus, the behavior of the numerical solution is entirely determined by the size of the coefficient

$$\mu := \frac{1 - \frac{\Delta t}{2\epsilon}}{1 + \frac{\Delta t}{2\epsilon}}.$$

Since $|\mu| < 1$ for all $\epsilon > 0$, the Trapezoidal scheme (4.96) is not affected by stiffness.

In the system of equations (1.5), the parameter satisfies $0 < a \ll 1$. This makes (1.5) a stiff system of ODEs.

4.7. Perturbation theories for differential equations

4.7.1. Regular perturbation theory. Let $\epsilon > 0$ be a small parameter and consider the differential equation

$$\begin{cases} \dfrac{\mathrm{d}x}{\mathrm{d}t} = f(t, x, \epsilon), \quad t \in [0, T], \\ x(0) = x_0, \quad x_0 \in \mathbb{R}. \end{cases} \tag{4.98}$$

If we suppose that $f \in \mathcal{C}^1$, then (4.98) is a **regular perturbation problem**. The solution $x(t, \epsilon)$ is in $\mathcal{C}^1$ and has the following Taylor expansion:

$$x(t, \epsilon) = x^{(0)}(t) + \epsilon x^{(1)}(t) + o(\epsilon) \tag{4.99}$$

with respect to ϵ in a neighborhood of 0.

Clearly, the unperturbed term $x^{(0)}$ is given as the solution of the unperturbed equation

$$\begin{cases} \dfrac{\mathrm{d}x^{(0)}}{\mathrm{d}t} = f_0(t, x^{(0)}), \quad t \in [0, T], \\ x^{(0)}(0) = x_0, \quad x_0 \in \mathbb{R}, \end{cases} \tag{4.100}$$

where $f_0(t, x) := f(t, x, 0)$. Moreover, the first-order correction term $x^{(1)}$, which is the derivative of $x(t, \epsilon)$ with respect to ϵ at 0,

$$x^{(1)}(t) = \frac{\partial x}{\partial \epsilon}(t, 0),$$

solves the equation

$$\begin{cases} \dfrac{\mathrm{d}x^{(1)}}{\mathrm{d}t} = \dfrac{\partial f}{\partial x}(t, x^{(0)}, 0)x^{(1)} + \dfrac{\partial f}{\partial \epsilon}(t, x^{(0)}, 0), \quad t \in [0, T], \\ x^{(1)}(0) = 0. \end{cases} \tag{4.101}$$

The initial condition $x^{(1)}(0) = 0$ follows from the fact that the initial condition x_0 does not depend on ϵ.

The numerical methods described in Section 4.4 can be used to efficiently compute the unperturbed solution $x^{(0)}$ and the first-order correction $x^{(1)}$.

REMARK 4.25. *Consider the equation*

$$\begin{cases} \dfrac{\mathrm{d}x}{\mathrm{d}t} = -\epsilon x + 1, \quad t \in [0, +\infty[, \\ x(0) = 0. \end{cases} \tag{4.102}$$

The solution can be easily found

$$x(t, \epsilon) = \frac{e^{-\epsilon t} - 1}{\epsilon}. \tag{4.103}$$

If we apply the perturbation theory to (4.102), then by solving (4.100) and (4.101) with

$$f(t, x, \epsilon) = -\epsilon x + 1,$$

we find

$$x^{(0)}(t) = -t \quad \textit{and} \quad x^{(1)}(t) = \frac{t^2}{2},$$

which gives

$$x(t, \epsilon) = -t + \epsilon \frac{t^2}{2} + o(\epsilon). \tag{4.104}$$

The approximation (4.104) of course coincides with the Taylor expansion of the exact solution given by (4.103). However, note that the approximation is valid only for fixed $t = O(1)$ and diverges to

$+\infty$ as t increases while the exact solution converges to $-1/\epsilon$. The limits $\epsilon \to 0$ and $t \to +\infty$ do not commute. Expansion (4.104) is not uniformly valid in time.

4.7.2. Singular perturbation theory. In this subsection we consider a system of ordinary differential equations (together with appropriate boundary conditions) in which the highest derivative is multiplied by a small, positive parameter ϵ. In what follows we give the general (nonlinear) form of the system:

$$\begin{cases} \epsilon \dfrac{d^2x}{dt^2} = f(t, x, \dfrac{dx}{dt}), \quad t \in [0, T], \\ x(0) = x_0, \quad x(T) = x_1. \end{cases} \tag{4.105}$$

The problem above is called a **singular perturbation problem**, and is characterized by the fact that its order reduces when the problem parameter ϵ equals zero. In such a situation, the problem becomes singular since, in general, not all of the original boundary conditions can be satisfied by the reduced problem. Singular perturbed problems form a particular class of **stiff problems**.

Consider the following linear, scalar and of second-order ODE which is subject to Dirichlet boundary conditions:

$$\begin{cases} \epsilon \dfrac{d^2x}{dt^2} + 2\dfrac{dx}{dt} + x = 0, \quad t \in [0, 1], \\ x(0) = 0, \quad x(1) = 1. \end{cases} \tag{4.106}$$

Let

$$\alpha(\epsilon) := \frac{1 - \sqrt{1-\epsilon}}{\epsilon} \quad \text{and} \quad \beta(\epsilon) := 1 + \sqrt{1-\epsilon}.$$

The solution of equation (4.106) is given by

$$x(t, \epsilon) = \frac{e^{-\alpha t} - e^{-\beta t/\epsilon}}{e^{-\alpha} - e^{-\beta/\epsilon}}, \quad t \in [0, 1]. \tag{4.107}$$

The solution $x(t, \epsilon)$ involves two terms which vary on widely different length-scales. Let us consider the behavior of $x(t, \epsilon)$ as $\epsilon \to 0^+$. The asymptotic behavior is nonuniform, and there are two cases, which lead to matching **outer** and **inner** solutions.

(i) **Outer limit**: $t > 0$ fixed and $\epsilon \to 0^+$. Then $x(t, \epsilon) \to x^{(0)}(t)$, where

$$x^{(0)}(t) := e^{(1-t)/2}. \tag{4.108}$$

This leading-order **outer solution** satisfies the boundary condition at $t = 1$ but not the boundary condition at $t = 0$. Indeed, $x^{(0)}(0) = e^{1/2}$.

(ii) **Inner limit**: $t/\epsilon = \tau$ fixed and and $\epsilon \to 0^+$. Then $x(\epsilon\tau, \epsilon) \to X^{(0)}(\tau) := e^{1/2}(1 - e^{-2\tau})$. This leading-order **inner solution** satisfies the boundary condition at $t = 0$ but not the one at $t = 1$, which corresponds to $\tau = 1/\epsilon$. Indeed, $\lim_{\tau \to +\infty} X^{(0)}(\tau) = e^{1/2}$.

(iii) **Matching**: Both the inner and outer expansions are valid in the region $\epsilon \ll t \ll 1$, corresponding to $t \to 0$ and $\tau \to +\infty$ as $\epsilon \to 0^+$. They satisfy the **matching condition**

$$\lim_{t \to 0^+} x^{(0)}(t) = \lim_{\tau \to +\infty} X^{(0)}(\tau). \tag{4.109}$$

Let us now construct an asymptotic solution of (4.106) without relying on the fact that we can solve it exactly.

We begin with the outer solution. We look for a straightforward expansion

$$x(t, \epsilon) = x^{(0)}(t) + \epsilon x^{(1)}(t) + O(\epsilon^2). \tag{4.110}$$

We use this expansion in (4.106) and equate the coefficients of the leading-order terms to zero. Guided by our analysis of the exact solution, we only impose the boundary condition at $t = 1$. We will see later that matching is impossible if, instead, we attempt to impose the boundary condition at $t = 0$. We obtain that

$$\begin{cases} 2\dfrac{dx^{(0)}}{dt} + x^{(0)} = 0, \quad t \in [0, 1], \\ x^{(0)}(1) = 1. \end{cases} \tag{4.111}$$

The solution of (4.111) is given by (4.108), in agreement with the expansion of the exact solution $x(t,\epsilon)$.

Next we consider the inner solution. We suppose that there is a **boundary layer** at $t = 0$ of width $\delta(\epsilon)$, and introduce a **stretched variable** $\tau = t/\delta$. We look for an inner solution $X(\tau,\epsilon) = x(t,\epsilon)$. Since

$$\frac{d}{dt} = \frac{1}{\delta}\frac{d}{d\tau},$$

we find from (4.106) that X satisfies

$$\frac{\epsilon}{\delta^2}\frac{d^2X}{d\tau^2} + \frac{2}{\delta}\frac{dX}{d\tau} + X = 0.$$

There are two possible dominant balances in this equation:

(i) $\delta = 1$, leading to the outer solution;
(ii) $\delta = \epsilon$, leading to the inner solution.

Thus we conclude that the boundary layer thickness is of the order of ϵ, and the appropriate inner variable is $\tau = t/\epsilon$. The equation for X is then

$$\begin{cases} \dfrac{d^2X}{d\tau^2} + 2\dfrac{dX}{d\tau} + \epsilon X = 0, \\ X(0,\epsilon) = 0. \end{cases}$$

We impose only the boundary condition at $\tau = 0$, since we do not expect the inner expansion to be valid outside the boundary layer where $t = O(\epsilon)$.

We seek an inner expansion

$$X(\tau,\epsilon) = X^{(0)}(\tau) + \epsilon X^{(1)}(\tau) + O(\epsilon^2)$$

and find that

$$\begin{cases} \dfrac{d^2X^{(0)}}{d\tau^2} + 2\dfrac{dX^{(0)}}{d\tau} = 0, \\ X^{(0)}(0) = 0. \end{cases} \tag{4.112}$$

The general solution of (4.112) is

$$X^{(0)}(\tau) = c(1 - e^{-2\tau}), \tag{4.113}$$

where c is an arbitrary constant of integration.

We can determine the unknown constant c in (4.113) by requiring that the inner solution (4.113) matches with the outer solution (4.108). Here the matching condition is simply

$$\lim_{t\to 0^+} x^{(0)}(t) = \lim_{\tau\to+\infty} X^{(0)}(\tau),$$

which implies that $c = e^{1/2}$.

In summary, the asymptotic solution as $\epsilon \to 0^+$ is given by

$$x(t,\epsilon) = \begin{cases} e^{1/2}(1 - e^{-2\tau}) & \text{as } \epsilon \to 0^+ \text{ with } t/\epsilon \text{ fixed}, \\ e^{(1-t)/2} & \text{as } \epsilon \to 0^+ \text{ with } t \text{ fixed}. \end{cases}$$

CHAPTER 5

Geometrical numerical integration methods for differential equation

5.1. Introduction

Geometric integration is the numerical integration of a differential equation, while preserving one or more of its geometric properties exactly, i.e., to within round-off error. Many of these geometric properties are of crucial importance in physical applications: preservation of energy, momentum, volume, symmetries, time-reversal symmetry, dissipation, and symplectic structure being examples. The aim of this chapter is to present geometric numerical integration methods for ordinary differential equations. We concentrate mainly on Hamiltonian systems and on methods that preserve their symplectic structure, invariants, symmetries, or volume.

5.2. Structure preserving methods for Hamiltonian systems

The numerical methods discussed in Chapter 4 are designed for general differential equations, and a distinction was drawn only between stiff and nonstiff problems. As shown in Chapter 1, Hamiltonian systems are an important class of differential equations with a geometric structure (their flow has the geometric property of being symplectic), whose preservation in the numerical discretization leads to substantially better methods, especially when integrating over long times. In general, most geometric properties are not preserved by the standard numerical methods presented in Chapter 4.

Some of the reasons we are motivated to preserve structure are

(i) it may yield methods that are faster, simpler, more stable, and/or more accurate for some types of ODEs;
(ii) it may yield more robust and quantitatively better results than standard methods for the long-time integration of Hamiltonian systems.

The standard problem in numerical ODEs discussed in the previous chapter is to compute the solution to an initial value problem at a fixed time, to within a given global error, as efficiently as possible. The class of method, its order and local error, and choice of time steps are all tailored to this end. In contrast, a typical application of a geometric numerical method is to fix a (sometimes moderately large) time step and compute solutions with perhaps many different initial conditions over very long time intervals.

5.2.1. Symplectic methods. Consider the Hamiltonian system

$$\begin{cases} \dfrac{\mathrm{d}p}{\mathrm{d}t} = -\dfrac{\partial H}{\partial q}(p,q), \\ \dfrac{\mathrm{d}q}{\mathrm{d}t} = \dfrac{\partial H}{\partial p}(p,q), \\ p(0) = p_0, q(0) = q_0, \end{cases} \tag{5.1}$$

where $p_0, q_0 \in \mathbb{R}^d$, and the Hamiltonian function $H : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ is a smooth function.

Definition 5.1. *Let J be defined by (1.22). A numerical one-step method $(p^{k+1}, q^{k+1}) = \Phi_{\Delta t}(p^k, q^k)$ for solving (5.1) is called* ***symplectic*** *if the* ***numerical flow*** *$\Phi_{\Delta t}$ is a symplectic*

map:

$$\Phi'_{\Delta t}(p,q)^\top J \Phi'_{\Delta t}(p,q) = J, \tag{5.2}$$

for all (p,q) *and all step sizes* Δt.

THEOREM 5.2. *The implicit Euler method for solving (5.1)*

$$\begin{cases} p^{k+1} &= p^k - \Delta t \dfrac{\partial H}{\partial q}(p^{k+1}, q^k), \\ q^{k+1} &= q^k + \Delta t \dfrac{\partial H}{\partial p}(p^{k+1}, q^k), \end{cases} \tag{5.3}$$

is symplectic. Moreover, if the Hamiltonian function $H(p,q) = T(p) + V(q)$ *is* ***separable****, then (5.3) is explicit.*

PROOF. Let $\Phi_{\Delta t}$ be the numerical flow associated with (5.3). We have

$$\Phi'_{\Delta t}(p^k, q^k) = \frac{(p^{k+1}, q^{k+1})}{\partial (p^k, q^k)}.$$

From

$$\begin{pmatrix} I + \Delta t \dfrac{\partial^2 H}{\partial p \partial q} & 0 \\ -\Delta t \dfrac{\partial^2 H}{\partial p^2} & I \end{pmatrix} \Phi'_{\Delta t}(p^k, q^k) = \begin{pmatrix} I & -\Delta t \frac{\partial^2 H}{\partial q^2} \\ 0 & I + \Delta t \dfrac{\partial^2 H}{\partial p \partial q} \end{pmatrix}, \tag{5.4}$$

where the matrices $\frac{\partial^2 H}{\partial p^2}$, $\frac{\partial^2 H}{\partial q^2}$, and $\frac{\partial^2 H}{\partial p^2}$ are evaluated at (p^{k+1}, q^{k+1}), one can easily verify by computing $\Phi'_{\Delta t}(p^k, q^k)$ from (5.4) that the symplecticity condition (5.2) holds. □

A variant of (5.3) is

$$\begin{cases} p^{k+1} &= p^k - \Delta t \dfrac{\partial H}{\partial q}(p^k, q^{k+1}), \\ q^{k+1} &= q^k + \Delta t \dfrac{\partial H}{\partial p}(p^k, q^{k+1}). \end{cases} \tag{5.5}$$

Analogously to (5.3), the Euler method (5.5) is symplectic and turns out to be explicit for separable Hamiltonian functions.

THEOREM 5.3. *The composition of two symplectic one-step methods for solving (5.1) is also symplectic.*

PROOF. Let $\Phi^{(1)}_{\Delta t}$ and $\Phi^{(2)}_{\Delta t}$ be the numerical flows associated with two symplectic one-step methods for solving (5.1). Let $\Phi_{\Delta t} := \Phi^{(2)}_{\Delta t} \circ \Phi^{(1)}_{\Delta t}$. We have

$$\begin{aligned} (\Phi'_{\Delta t}(x))^\top J \Phi'_{\Delta t}(x) &= ((\Phi^{(2)}_{\Delta t})'(x^*)(\Phi^{(1)}_{\Delta t})'(x))^\top J (\Phi^{(2)}_{\Delta t})'(x^*)(\Phi^{(1)}_{\Delta t})'(x) \\ &= ((\Phi^{(1)}_{\Delta t})'(x))^\top ((\Phi^{(2)}_{\Delta t})'(x^*))^\top J (\Phi^{(2)}_{\Delta t})'(x^*)(\Phi^{(1)}_{\Delta t})'(x) \\ &= ((\Phi^{(1)}_{\Delta t})'(x))^\top J (\Phi^{(1)}_{\Delta t})'(x) = J, \end{aligned}$$

where $x^* = \Phi^{(1)}_{\Delta t}(x)$. That is, the composition of symplectic one-step methods is again a symplectic one-step method. □

Define the **leapfrog method** (Verlet method and Strömer-Verlet method are also often-used names) for solving the Hamiltonian system (5.1) by

$$\begin{cases} p^{k+\frac{1}{2}} = p^k - \dfrac{\Delta t}{2}\dfrac{\partial H}{\partial q}(p^{k+\frac{1}{2}}, q^k), \\ q^{k+1} = q^k + \dfrac{\Delta t}{2}\left(\dfrac{\partial H}{\partial p}(p^{k+\frac{1}{2}}, q^k) + \dfrac{\partial H}{\partial p}(p^{k+\frac{1}{2}}, q^{k+1})\right), \\ p^{k+1} = p^{k+\frac{1}{2}} - \dfrac{\Delta t}{2}\dfrac{\partial H}{\partial q}(p^{k+\frac{1}{2}}, q^{k+1}). \end{cases} \tag{5.6}$$

THEOREM 5.4. *The leapfrog method (5.6) for solving the Hamiltonian system (5.1) is symplectic.*

PROOF. The leapfrog method (5.6) can be interpreted as the composition of the symplectric Euler method

$$\begin{cases} p^{k+\frac{1}{2}} &= p^k - \dfrac{\Delta t}{2}\dfrac{\partial H}{\partial q}(p^{k+\frac{1}{2}}, q^k), \\ q^{k+\frac{1}{2}} &= q^k + \dfrac{\Delta t}{2}\dfrac{\partial H}{\partial p}(p^{k+\frac{1}{2}}, q^k), \end{cases} \tag{5.7}$$

and its adjoint

$$\begin{cases} q^{k+1} &= q^{k+\frac{1}{2}} + \dfrac{\Delta t}{2}\dfrac{\partial H}{\partial p}(p^{k+\frac{1}{2}}, q^{k+1}), \\ p^{k+1} &= p^{k+\frac{1}{2}} - \dfrac{\Delta t}{2}\dfrac{\partial H}{\partial q}(p^{k+\frac{1}{2}}, q^{k+1}). \end{cases} \tag{5.8}$$

The methods (5.7) and (5.8) are symplectic. Hence their composition (5.6) is also symplectic. □

Let $x = (p, q)^\top$. The Hamiltonian system of equations (5.1) can be rewritten as a first-order differential equation

$$\begin{cases} \dfrac{dx}{dt} = f(x), \\ x(0) = x_0 \in \mathbb{R}^{2d}, \end{cases} \tag{5.9}$$

where $x_0 = (p_0, q_0)^\top$ and

$$\begin{aligned} f \quad : \quad & \mathbb{R}^{2d} \to \mathbb{R}^{2d} \\ & x \mapsto J^{-1}\nabla H(x). \end{aligned}$$

5.2.2. Preserving time-reversal symmetry and invariants.

5.2.2.1. *Preserving time-reversal symmetry.* The leapfrog method (5.6) is symmetric with respect to changing the direction of time: replacing Δt by $-\Delta t$ and exchanging the superscripts k and $k+1$ results in the same method. In terms of the numerical one-step map $\Phi_{\Delta t} : (p^k, q^k) \mapsto (p^{k+1}, q^{k+1})$, the symmetry property is stated as follows.

DEFINITION 5.5. *The numerical one-step map $\Phi_{\Delta t}$ is said to be **symmetric** if*

$$\Phi_{\Delta t} = \Phi_{-\Delta t}^{-1}. \tag{5.10}$$

Relation (5.10) does not hold for the symplectic Euler methods (5.7) and (5.8), where the time reflection transforms (5.7) to (5.8) and vice versa.

The time-symmetry of the leapfrog method (5.6) implies an important geometric property of the numerical map, namely **reversibility**.

Assume that

$$H(-p, q) = H(p, q). \tag{5.11}$$

Then the system (5.1) has the property that inverting the direction of the initial p_0 does not change the solution trajectory. The flow ϕ_t associated with (5.1) satisfies

$$\phi_t(p_0, q_0) = (p, q) \Rightarrow \phi_t(-p, q) = (-p_0, q_0). \tag{5.12}$$

Relation (5.12) shows that ϕ_t is **reversible** with respect to the reflection $(p, q) \mapsto (-p, q)$.

DEFINITION 5.6. *The numerical one-step map $\Phi_{\Delta t}$ is said to be **reversible** if*

$$\Phi_{\Delta t}(p, q) = (\hat{p}, \hat{q}) \Rightarrow \Phi_{\Delta t}(-\hat{p}, \hat{q}) = (-p, q), \tag{5.13}$$

for all p, q and all Δt.

Since

$$\Phi_{\Delta t}(p, q) = (\hat{p}, \hat{q}) \Rightarrow \Phi_{-\Delta t}(-p, q) = (-\hat{p}, \hat{q}), \tag{5.14}$$

the symmetry (5.10) of the leapfrog method (5.6) is therefore equivalent to the reversibility (5.12).

THEOREM 5.7. *The leapfrog method (5.6) applied to (5.1) with H satisfying (5.11) is both symmetric and reversible, i.e., its one-step map satisfies (5.10) and (5.12).*

5.2.2.2. *Preserving invariants.*

DEFINITION 5.8. *A numerical one-step method $\Phi_{\Delta t}$ for solving (5.9) is said to **preserve the invariant** F if $F(\Phi_{\Delta t}(p, q)) = $ Constant for all p, q and all Δt. If $F = H$, then we say that the scheme preserves **energy**.*

THEOREM 5.9. *The leapfrog method (5.6) applied to (5.1) preserves linear invariants and quadratic invariants of the form*

$$F(p, q) = p^\top (Bq + b). \tag{5.15}$$

PROOF. Let the linear invariant be $F(p, q) = b^\top q + c^\top p$, so that

$$b^\top \frac{\partial H}{\partial p}(p, q) - c^\top \frac{\partial H}{\partial q}(p, q) = 0,$$

for all p, q. Multiplying the formulas for $\Phi_{\Delta t}(p, q)$ in (5.6) by $(c, b)^\top$ thus yields the desired result on linear invariants.

Next we turn to the conservation by the leapfrog method of quadratic invariants of the form (5.15). In order to prove that (5.6) applied to (5.1) preserves quadratic invariants of the form $F(p, q) = p^\top (Bq + b)$, we write (5.6) as the composition of the two symplectic Euler methods (5.7) and (5.8). For the first half-step, we obtain

$$(p^{k+\frac{1}{2}})^\top (Bq^{k+\frac{1}{2}} + b) = (p^k)^\top (Bq^k + b).$$

For the second half-step, we obtain in the same way

$$(p^{k+1})^\top (Bq^{k+1} + b) = (p^{k+\frac{1}{2}})^\top (Bq^{k+\frac{1}{2}} + b),$$

and the result follows. □

The energy is generally not preserved by the leapfrog method (5.6). Consider $H(p, q) = \frac{1}{2}(p^2 + q^2)$. Applying (5.6) gives

$$\begin{pmatrix} p^{k+1} \\ q^{k+1} \end{pmatrix} = \begin{bmatrix} 1 - \frac{(\Delta t)^2}{2} & -\frac{\Delta t}{2}(1 - \frac{(\Delta t)^2}{4}) \\ \frac{\Delta t}{2} & 1 - \frac{(\Delta t)^2}{2} \end{bmatrix} \begin{pmatrix} p^k \\ q^k \end{pmatrix}. \tag{5.16}$$

Since the **propagation matrix** in (5.16) is not orthogonal, $H(p, q)$ is not preserved along numerical solutions.

Consider the Hamiltonian

$$H(p, q) := \frac{1}{2} p^\top M^{-1} p + V(q), \tag{5.17}$$

where M is a symmetric positive definite matrix and the potential V is a smooth function.

In the particular case of the Hamiltonian (5.17), the leapfrog method (5.6) reduces to the explicit method

$$\begin{cases} p^{k+\frac{1}{2}} = p^k - \dfrac{\Delta t}{2}\nabla V(q^k), \\ q^{k+1} = q^k + \Delta t M^{-1} p^{k+\frac{1}{2}}, \\ p^{k+1} = p^{k+\frac{1}{2}} - \dfrac{\Delta t}{2}\nabla V(q^{k+1}). \end{cases} \tag{5.18}$$

Note that the Hamiltonian (5.17) is invariant under $p \mapsto -p$ and the corresponding Hamiltonian system (5.1) is invariant under the transformation

$$\begin{bmatrix} p \\ t \end{bmatrix} \mapsto \begin{bmatrix} -p \\ -t \end{bmatrix}. \tag{5.19}$$

The **time-reversal symmetry** of (5.18) is preserved by the leapfrog method (5.18).

5.2.2.3. *Preserving volume.* Recall that, due to equality of mixed partial derivatives, (5.9) is divergence-free, i.e.,

$$\nabla \cdot f := \sum_{i=1}^{2d} \frac{\partial f_i}{\partial x_i} = 0.$$

A remarkable feature of divergence-free vector fields is that the associated flows are volume preserving.

Given a map $\phi : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ and a domain Ω, by change of variables

$$\mathrm{vol}(\phi(\Omega)) = \int_\Omega |\det \phi'(y)|\, dy,$$

where ϕ' is the Jacobian of ϕ. It follows that ϕ preserves volume provided that

$$|\det \phi'(y)| = 1 \quad \text{for } y \in \Omega. \tag{5.20}$$

Let ϕ_t be the flow associated with (5.9). Then ϕ_t satisfies

$$\frac{d\phi_t(y)}{dt} = f(\phi_t(y)),$$

and therefore, its Jacobian ϕ' satisfies

$$\frac{d\phi_t'(y)}{dt} = f'(\phi_t(y))\phi_t'(y).$$

Assuming ϕ_t' is invertible yields

$$\mathrm{tr}\left[\frac{d\phi_t'(y)}{dt}\phi_t'(y)^{-1}\right] = \mathrm{tr} f'(\phi_t(y)).$$

Combining $\mathrm{tr} f' = \nabla \cdot f = 0$ and Jacobi's formula for the derivative of a determinant gives

$$\mathrm{tr}\left[\frac{d\phi_t'(y)}{dt}\phi_t'(y)^{-1}\right] = \frac{1}{\det \phi_t'(y)}\frac{d}{dt}\det \phi_t'(y) = 0.$$

Hence,

$$\det \phi_t'(y) = \det \phi_{t=0}'(y) = 1.$$

The following result holds.

THEOREM 5.10 (**Liouville's theorem**). *The flow ϕ_t associated with the system*

$$\begin{cases} \dfrac{dx}{dt} = f(x), \\ x(0) = x_0 \in \mathbb{R}^{2d}, \end{cases} \tag{5.21}$$

*where the vector field f is **divergence-free**, is a **volume preserving map** (for all t).*

Note that if the system (5.21) is Hamiltonian, then Theorem 5.10 can be immediately obtained from the **symplecticity** of the associated flow.

DEFINITION 5.11. *A numerical one-step method for solving (5.21) is said to be* ***volume preserving*** *if* $|\det \Phi'_{\Delta t}(p,q)| = 1$ *for all* p, q.

Note that if (5.21) is a Hamiltonian system, then any symplectic numerical method preserves the volume. However, no standard methods can be volume-preseving for all divergence-free vector fields.

EXAMPLE 5.12. *Consider the divergence-free problem*

$$\begin{cases} \dfrac{dx}{dt} = Ax, \\ x(0) = x_0 \in \mathbb{R}^{2d}, \end{cases} \tag{5.22}$$

where $A \in \mathbb{M}_{2d}(\mathbb{R})$ *and* $\mathrm{tr}A = 0$. *The Explicit and implicit Euler's schemes for solving (5.22)*

$$x^{k+1} = x^k + \Delta t A x^k,$$
$$x^{k+1} = x^k + \Delta t A x^{k+1},$$

are volume-preserving if and only if

$$|\det(I + \Delta t A)| = 1,$$

and

$$|\det(I - \Delta t A)| = 1,$$

respectively.

5.2.3. Composition methods. Now using the fact that (5.9) is divergence-free, we have

$$\begin{aligned} f_{2d}(x) &= f_{2d}(\overline{x}) + \int_{\overline{x}}^{x_{2d}} \frac{\partial f_{2d}}{\partial x_{2d}}\, dx_{2d} \\ &= f_{2d}(\overline{x}) - \int_{\overline{x}}^{x_{2d}} \left(\sum_{i=1}^{2d-1} \frac{\partial f_i(x)}{\partial x_i} \right) dx_{2d}, \end{aligned} \tag{5.23}$$

where $\overline{x}$ is an arbitrary point which can be chosen conveniently (e.g., if possible such that $f_{2d}(\overline{x}) = 0$).

Substituting (5.23) into (5.9) yields

$$\begin{aligned} \frac{dx_1}{dt} &= f_1(x), \\ &\vdots \\ \frac{dx_{2d-1}}{dt} &= f_{2d-1}(x), \\ \frac{dx_{2d}}{dt} &= f_{2d}(\overline{x}) - \sum_{i=1}^{2d-1} \int_{\overline{x}}^{x_{2d}} \frac{\partial f_i(x)}{\partial x_i}\, dx_{2d}. \end{aligned} \tag{5.24}$$

We now split this as the sum of $2d-1$ vector fields

$$\begin{aligned} \frac{dx_i}{dt} &= 0, \quad i \neq j, 2d-1, \\ \frac{dx_j}{dt} &= f_j(x), \\ \frac{dx_{2d}}{dt} &= f_{2d}(\overline{x})\delta_{j,2d-1} - \int_{\overline{x}}^{x_{2d}} \frac{\partial f_j(x)}{\partial x_j}\, dx_{2d}, \end{aligned} \tag{5.25}$$

for $j = 1, \ldots, 2d-1$. Here δ is the Kronecker delta function.

Note that each of the $2d-1$ vector fields is divergence-free. Moreover, we have split (5.24) into the $2d-1$ problems (5.25). Each of these problems has a simpler structure than (5.9). In fact, each of them corresponds to a two-dimensional Hamiltonian system

$$\begin{aligned} \frac{dx_j}{dt} &= \frac{\partial H_j}{\partial x_{2d}}, \\ \frac{dx_{2d}}{dt} &= -\frac{\partial H_j}{\partial x_j}, \end{aligned} \tag{5.26}$$

with Hamiltonian

$$H_j(x) := f_{2d}(\overline{x})\delta_{j,2d-1}x_j - \int_{\overline{x}}^{x_{2d}} f_j(x)\,dx_{2d}, \tag{5.27}$$

treating x_i for $i \neq j, 2d$ as fixed parameters.

Each of the two-dimensional problems (5.26) can either be solved exactly (if possible), or approximated with a symplectic integrator $\Phi_{\Delta t}^{(j)}$. A volume-preserving integrator for f is then given by

$$\Phi_{\Delta t} = \Phi_{\Delta t}^{(1)}\Phi_{\Delta t}^{(2)}\ldots\Phi_{\Delta t}^{(2d-1)}. \tag{5.28}$$

5.2.4. Splitting methods. Consider a Hamiltonian system

$$\frac{dx}{dt} = J^{-1}\nabla H(x), \quad H(x) = H_1(x) + H_2(x),$$

and suppose the flows

$$\frac{dx}{dt} = J^{-1}\nabla H_1(x) \quad \text{and} \quad \frac{dx}{dt} = J^{-1}\nabla H_2(x),$$

can be exactly integrated. Define the corresponding flow maps $\phi_t^{(1)}$ and $\phi_t^{(2)}$. Since the exact solution of a Hamiltonian system defines a symplectic map, we have

$$((\phi_t^{(1)})')^\top J(\phi_t^{(1)})' = J \quad \text{and} \quad ((\phi_t^{(2)})')^\top J(\phi_t^{(2)})' = J.$$

Next consider the numerical method defined by composing these two exact flows:

$$\Phi_{\Delta t}(x) := \phi_{\Delta t}^{(2)} \circ \phi_{\Delta t}^{(1)}(x).$$

This map is also symplectic, since

$$\begin{aligned} (\Phi'_{\Delta t}(x))^\top J\Phi'_{\Delta t}(x) &= ((\phi_{\Delta t}^{(2)})'(x^*)(\phi_{\Delta t}^{(1)})'(x))^\top J(\phi_{\Delta t}^{(2)})'(x^*)(\phi_{\Delta t}^{(1)})'(x) \\ &= ((\phi_{\Delta t}^{(1)})'(x))^\top((\phi_{\Delta t}^{(2)})'(x^*))^\top J(\phi_{\Delta t}^{(2)})'(x^*)(\phi_{\Delta t}^{(1)})'(x) \\ &= ((\phi_{\Delta t}^{(1)})'(x))^\top J(\phi_{\Delta t}^{(1)})'(x) = J, \end{aligned}$$

where $x^* = \phi_{\Delta t}^{(1)}(x)$. That is, as shown in Theorem 5.3, the composition of symplectic maps is again a symplectic map.

EXAMPLE 5.13. *Consider the separable Hamiltonian $H(p,q) = T(p)+V(q)$. Based on splitting the Hamiltonian H into T and V, we derive the following methods for solving (5.9): The symplectic Euler method*

$$\begin{cases} p^{k+1} = p^k - \Delta t\nabla V(q^k), \\ q^{k+1} = q^k + \Delta t\nabla T(p^{k+1}), \end{cases}$$

and the leapfrog method

$$\begin{cases} p^{k+\frac{1}{2}} = p^k - \dfrac{\Delta t}{2}\nabla V(q^k), \\ q^{k+1} = q^k + \Delta t\nabla T(p^{k+\frac{1}{2}}), \\ p^{k+1} = p^{k+\frac{1}{2}} - \dfrac{\Delta t}{2}\nabla V(q^{k+1}). \end{cases}$$

5.3. Runge-Kutta methods

Now we turn to Runge-Kutta methods

$$\begin{cases} x_{i,k} = x^k + (\Delta t)\sum_{j=1}^{m} a_{ij} f(x_{j,k}), \\ x^{k+1} = x^k + (\Delta t)\sum_{i=1}^{m} b_i f(x_{i,k}), \end{cases} \tag{5.29}$$

for solving (5.9).

THEOREM 5.14. (i) *All the Runge-Kutta methods (5.29) preserve linear invariants;*
(ii) *The Runge-Kutta method (5.29) whose coefficients satisfy the condition*

$$b_i a_{ij} + b_j a_{ji} - b_i b_j = 0, \quad i, j = 1, \dots m, \tag{5.30}$$

preserves all quadratic invariants.

PROOF. Define $\Phi_{\Delta t}$ by $x^{k+1} = \Phi_{\Delta t}(x^k)$. Let $F(x) = d^\top x$, where $d \in \mathbb{R}^{2d}$. We compute

$$F(\Phi_{\Delta t}(x^k)) = d^\top (x^k + \Delta t \sum_{i=1}^{m} b_i f(x_{i,k})) = d^\top x^k,$$

since $d^\top x$ is assumed to be an invariant of (5.9) and hence $d^\top f(x_{i,k}) = 0$.

Next, let $F(x) = x^\top C x$, where C is a symmetric $2d \times 2d$ matrix. Assume that F is an invariant of (5.9). We have

$$x^\top C f(x) = 0 \quad \text{for all } x. \tag{5.31}$$

On the other hand, we have

$$\begin{aligned} F(\Phi_{\Delta t}(x^k)) &= (x^k + \Delta t \sum_{j=1}^{m} b_j f(x_{j,k}))^\top C (x^k + \Delta t \sum_{i=1}^{m} b_i f(x_{i,k})) \\ &= (x^k)^\top C x^k + (\Delta t)\sum_{i=1}^{m} (x^k)^\top C b_i f(x_{i,k}) + (\Delta t)\sum_{j=1}^{m} b_j f(x_{j,k})^\top C x^k \\ &\quad + (\Delta t)^2 \sum_{i,j=1}^{m} b_i b_j f(x_{j,k})^\top C f(x_{i,k}). \end{aligned}$$

From (5.31), we obtain

$$(x_{i,k})^\top C f(x_{i,k}) = 0,$$

and hence, by writing

$$x^k = x^k + \Delta t \sum_{j=1}^{m} a_{ij} f(x_{j,k}) - \Delta t \sum_{j=1}^{m} a_{ij} f(x_{j,k}) = x_{i,k} - \Delta t \sum_{j=1}^{m} a_{ij} f(x_{j,k}),$$

we get

$$\begin{aligned} F(\Phi_{\Delta t}(x^k)) &= (x^k)^\top C x^k - (\Delta t)^2 \sum_{i,j=1}^{m} b_i a_{ij} f(x_{j,k})^\top C f(x_{i,k}) - (\Delta t)^2 \sum_{i,j=1}^{m} b_j a_{ji} f(x_{j,k})^\top C f(x_{i,k}) \\ &\quad + (\Delta t)^2 \sum_{i,j=1}^{m} b_i b_j f(x_{j,k})^\top C f(x_{i,k}) \\ &= (x^k)^\top C x^k - (\Delta t)^2 \Big(\sum_{i,j=1}^{m} (b_i a_{ij} + b_j a_{ji} - b_i b_j) f(x_{j,k})^\top C f(x_{i,k}) \Big). \end{aligned}$$

Therefore, the Runge-Kutta method (5.29) preserves the quadratic invariant F provided that (5.30) holds. □

Lemma 1.18 shows that H is an invariant of (5.9). If H is quadratic, then Theorem 5.14 says that the energy is preserved by the Runge-Kutta method (5.29) provided that condition (5.30) holds.

The following characterization of **symplectic Runge-Kutta methods** for solving (5.9) holds.

THEOREM 5.15. *The Runge-Kutta method (5.29) for solving (5.9) whose coefficients satisfy condition (5.30) is symplectic.*

PROOF. Theorem 1.24 shows that the flow ϕ_t is a symplectic transformation (if H is smooth enough). Let $\Psi(t) := \frac{\partial \phi_t(x_0)}{\partial x_0} = \phi_t'$, where x_0 is the initial condition. We have

$$\begin{cases} \dfrac{d\Psi}{dt} = f'(x)\Psi, \\ \Psi(0) = I. \end{cases} \tag{5.32}$$

Apply a Runge-Kutta method satisfying (5.30) to (5.9) and (5.32) to obtain the approximations x^{k+1} and Ψ^{k+1} from x^k and Ψ^k. Since $\Psi^\top J\Psi$ is a quadratic invariant of (5.32), we obtain

$$(\Psi^k)^\top J\Psi^k = J \quad \text{for all } k.$$

Suppose for a moment that

$$\Psi^{k+1} = \frac{\partial x^{k+1}}{\partial x^k}. \tag{5.33}$$

We obtain

$$(\frac{\partial x^{k+1}}{\partial x^k})^\top J \frac{\partial x^{k+1}}{\partial x^k} = J,$$

which means that the Runge-Kutta method for solving (5.9) whose coefficients satisfy condition (5.30) is symplectic.

In order to complete the proof, we prove (5.33). We want to show that the result of first applying $\Phi_{\Delta t}$ and then differentiating with respect to x^k is the same as applying the same Runge-Kutta method to (5.32).

In fact, on the one hand, by differentiating (5.29) with respect to x^k we obtain

$$\begin{cases} \dfrac{\partial x_{i,k}}{\partial x^k} = I + (\Delta t)\displaystyle\sum_{j=1}^m a_{ij} f'(x_{j,k}) \frac{\partial x_{j,k}}{\partial x^k}, \\ \dfrac{\partial x^{k+1}}{\partial x^k} = I + (\Delta t)\displaystyle\sum_{i=1}^m b_i f'(x_{i,k}) \frac{\partial x_{i,k}}{\partial x^k}. \end{cases} \tag{5.34}$$

Multiplying the first equation in (5.34) by $f'(x_{i,k})$ yields the linear system in the unknowns $f'(x_{i,k})\frac{\partial x_{i,k}}{\partial x^k}$

$$f'(x_{i,k})\frac{\partial x_{i,k}}{\partial x^k} = f'(x_{i,k})\bigg(I + (\Delta t)\sum_{j=1}^m a_{ij} f'(x_{j,k})\frac{\partial x_{j,k}}{\partial x^k}\bigg), \tag{5.35}$$

$$\frac{\partial x^{k+1}}{\partial x^k} = I + (\Delta t)\sum_{i=1}^m b_i f'(x_{i,k})\frac{\partial x_{i,k}}{\partial x^k}. \tag{5.36}$$

On the other hand, applying the same Runge-Kutta method to (5.32) yields

$$\Psi_{i,k} = f'(x^k + \Delta t\sum_{j=1}^m a_{ij} x_{j,k})\bigg(I + (\Delta t)\sum_{j=1}^m a_{ij}\Psi_{j,k}\bigg), \tag{5.37}$$

$$\Psi^{k+1} = I + (\Delta t)\sum_{i=1}^m b_i \Psi_{i,k}. \tag{5.38}$$

We conclude the proof by observing that (5.37) is the same system as (5.35) but in the unknowns $\Psi_{i,k}, i = 1, \ldots, m$. It is easily seen that this system has a unique solution for sufficiently small Δt, so it must be

$$\Psi_{i,k} = \frac{\partial x_{i,k}}{\partial x^k} \quad \text{for } i = 1, \ldots, m,$$

which, in view of (5.36) and (5.38), yields (5.33).

□

For arbitrary Hamiltonians, the only known symplectic one-step numerical methods are the symplectic Runge-Kutta methods of the form (4.59) that satisfy the symplectic condition (5.30).

EXAMPLE 5.16. *The midpoint scheme for solving (5.9)*

$$x^{k+1} = x^k + \Delta t f(\frac{x^k + x^{k+1}}{2}), \tag{5.39}$$

is symplectic and preserves linear and quadratic invariants. Moreover, it is time-reversible.

5.4. Long-time behaviour of numerical solutions

In (5.16) we have seen that the energy is not exactly preserved by the leapfrog method (5.6). In that example, it is however, approximately preserved. As shown in the following theorem, the symplecticity of a one-step numerical method yields an approximate conservation of energy over very long times for general Hamiltonian systems.

THEOREM 5.17. *For an analytic Hamiltonian H and a symplectic one-step numerical method $\Phi_{\Delta t}$ of order n, if the numerical trajectory remains in a compact subset, then there exist $h > 0$ and $\Delta t^* > 0$ such that, for $\Delta t \leq \Delta t^*$,*

$$H(p^k, q^k) = H(p^0, q^0) + O((\Delta t)^n), \tag{5.40}$$

for exponentially long times $k\Delta t \leq e^{\frac{h}{\Delta t}}$. Here, $(p^{k+1}, q^{k+1}) = \Phi_{\Delta t}(p^k, q^k)$.

Theorem (5.17) is based on simplicticity. It can be proved via backward error analysis. The idea is to deduce the long-time behavior estimate (5.40) from properties of the solution of the equation corresponding to an approximation $H_{\Delta t}$ of the Hamiltonian H.

CHAPTER 6

Finite difference methods

6.1. Introduction

Finite difference methods are basic numerical solution methods for partial differential equations. They are obtained by replacing the derivatives in the equation by the appropriate numerical differentiation formulas. However, there is no guarantee that the resulting numerical scheme will accurately approximate the true solution. Further analysis is required. In this chapter, we establish some of the most basic finite difference schemes for the heat and the wave equations.

6.2. Numerical algorithms for the heat equation

6.2.1. Finite difference approximations. Consider the **heat equation**

$$\begin{cases} \dfrac{\partial u}{\partial t} - \gamma \dfrac{\partial^2 u}{\partial x^2} = 0, \quad x \in [0,1], t \geq 0, \\ u(t,0) = u(t,1) = 0, \quad t \geq 0, \\ u(0,x) = u_0(x), \quad x \in [0,1], \end{cases} \tag{6.1}$$

where $\gamma > 0$ is the thermal conductivity.

In order to design a numerical approximation to the solution u of (6.1), we begin by introducing a rectangular mesh consisting of points (t_k, x_j) with

$$0 = t_0 < t_1 < t_2 < \dots \quad \text{and} \quad 0 = x_0 < x_1 < \dots < x_{N+1} = 1.$$

For simplicity, we maintain a uniform mesh spacing in both directions, with

$$\Delta t = t_{k+1} - t_k, \quad \Delta x = x_{j+1} - x_j = \frac{1}{N},$$

representing, respectively, the time step size and the spatial mesh size. We shall use the notation

$$u_j^k \approx u(t_k, x_j) \quad \text{where } t_k = k\Delta t, \quad x_j = j\Delta x,$$

to denote the numerical approximation of u at the mesh point (t_k, x_j).

The Dirichlet boundary conditions $u(t,0) = u(t,1) = 0$, $t \geq 0$, yield

$$u_0^k = u_{N+1}^k = 0 \quad \text{for all } k > 0.$$

As a first attempt at designing a numerical method, we shall employ the simplest finite difference approximations to the derivatives. The second order space derivative is approximated by

$$\begin{aligned} \frac{\partial^2 u}{\partial x^2}(t_k, x_j) &\approx \frac{u(t_k, x_{j-1}) + 2u(t_k, x_j) - u(t_k, x_{j+1})}{(\Delta x)^2} + O((\Delta x)^2) \\ &\approx \frac{u_{j-1}^k - 2u_j^k + u_{j+1}^k}{(\Delta x)^2} + O((\Delta x)^2). \end{aligned} \tag{6.2}$$

Similarly, the time derivative can be approximated by

$$\frac{\partial u}{\partial t}(t_k, x_j) \approx \frac{u(t_{k+1}, x_j) - u(t_k, x_j)}{\Delta t} + O(\Delta t) \approx \frac{u_j^{k+1} - u_j^k}{\Delta t} + O(\Delta t). \tag{6.3}$$

Replacing the derivatives in the heat equation (6.1) by their finite difference approximations (6.2) and (6.3), we end up with the **explicit scheme**

$$\frac{u_j^{k+1}-u_j^k}{\Delta t}+\gamma\frac{-u_{j-1}^k+2u_j^k-u_{j+1}^k}{(\Delta x)^2}=0 \tag{6.4}$$

for $k\geq 0$ and $j\in\{1,\dots,N\}$.

Let

$$\mu:=\frac{\gamma\Delta t}{(\Delta x)^2}, \tag{6.5}$$

and let

$$u^{(k)}:=(u_1^k,u_2^k,\dots,u_N^k)^\top\approx(u(t_k,x_1),u(t_k,x_2),\dots,u(t_k,x_N))^\top, \tag{6.6}$$

be the vector whose entries are the numerical approximations to the solution values at time t_k at the interior nodes.

The scheme (6.4) can be written in the matrix form

$$u^{(k+1)}=Au^{(k)}, \tag{6.7}$$

where

$$A:=\begin{pmatrix}1-2\mu & \mu & & & \\ \mu & 1-2\mu & \mu & & \\ & \mu & 1-2\mu & \mu & \\ & \ddots & \ddots & \ddots & \\ & & \mu & 1-2\mu & \mu \\ & & & \mu & 1-2\mu\end{pmatrix}. \tag{6.8}$$

The matrix A is symmetric and tridiagonal.

6.2.2. Consistency, stability, and convergence. A general **finite difference method** is defined by

$$F_{\Delta t,\Delta x}(\{u_{j+n}^{k+m}\}_{m^-\leq m\leq m^+,n^-\leq n\leq n^+})=0, \tag{6.9}$$

where the integers $m^\pm,n^\pm$ define the width of the stencil of the scheme.

Definition 6.1 (**Consistency and order**). *The finite difference scheme (6.9) is consistent with the equation $F(u)=0$ if, for any smooth solution $u(x,t)$, the* ***truncation error*** *defined by*

$$F_{\Delta t,\Delta x}(\{u(t_{k+m},x_{j+n})\}_{m^-\leq m\leq m^+,n^-\leq n\leq n^+}) \tag{6.10}$$

goes to zero as Δt and Δx go to zero independently. Moreover, the scheme is said to be of order p in time and order q in space if the truncation error is of the order of $O((\Delta t)^p+(\Delta x)^q)$ as Δt and Δx go to zero.

Theorem 6.2. *The explicit scheme (6.4) is consistent with the heat equation (6.1), of order one in time and two in space. Moreover, if*

$$\frac{\gamma\Delta t}{(\Delta x)^2}=\frac{1}{6}, \tag{6.11}$$

then it is of order two in time and four in space.

Proof. Let $v(t,x)\in\mathcal{C}^6$. By the Taylor expansion of v evaluated at (t,x),

$$\begin{aligned}\frac{v(t+\Delta t,x)-v(t,x)}{\Delta t}+\gamma\frac{-v(t,x-\Delta x)+2v(t,x)-v(t,x+\Delta x)}{(\Delta x)^2}&=(\frac{\partial v}{\partial t}-\gamma\frac{\partial^2 v}{\partial x^2})(t,x)\\ +\frac{\Delta t}{2}\frac{\partial^2 v}{\partial t^2}(t,x)-\frac{\gamma(\Delta x)^2}{12}\frac{\partial^4 v}{\partial x^4}(t,x)&+O((\Delta t)^2+(\Delta x)^4).\end{aligned} \tag{6.12}$$

If v is a solution to (6.1), then it follows from (6.12) that the truncation error goes to zero as $\Delta t, \Delta x \to 0$ and hence, the explicit scheme is consistent. Moreover, it is of order 1 in time and 2 in space. If we suppose that (6.11) holds, then the terms in Δt and $(\Delta x)^2$ cancel out since

$$\frac{\partial^2 v}{\partial t^2} = \gamma \frac{\partial^3 v}{\partial t \partial x^2} = \gamma^2 \frac{\partial^4 v}{\partial x^4}.$$

Thus, the explicit scheme is of order 2 in time and 4 in space. □

DEFINITION 6.3 (**Stability**). *A finite difference scheme is stable with respect to the norm $\| \ \|_r$ defined by*

$$\|u^{(k)}\|_r := \left(\sum_{j=1}^{N} \Delta x |u_j^k|^r \right)^{\frac{1}{r}}, \quad 1 \leq r \leq +\infty, \tag{6.13}$$

where $u^{(k)}$ is given by (6.6), if there exists a positive constant C independent of Δt and Δx such that

$$\|u^{(k)}\|_r \leq C \|u^{(0)}\|_r \quad \textit{for all } k \geq 0. \tag{6.14}$$

DEFINITION 6.4 (**linear scheme**). *A finite difference scheme defined by (6.9) is said to be linear if (6.9) is linear with respect to its arguments u_{j+n}^{k+m}.*

If a finite difference scheme is linear, then it can be written in the form

$$u^{(k+1)} = A u^{(k)}, \tag{6.15}$$

where A is the **iteration matrix**. From (6.15), it follows that

$$u^{(k+1)} = A^{k+1} u^{(0)},$$

and therefore, the stability of (6.15) is equivalent to

$$\|A^k u^{(0)}\|_r \leq C \|u^{(0)}\|_r, \quad \text{for all } k \geq 0 \text{ and } u^{(0)} \in \mathbb{R}^N. \tag{6.16}$$

Introduce the matrix norm

$$\|M\|_r = \sup_{u \in \mathbb{R}^N, u \neq 0} \frac{\|Mu\|_r}{\|u\|_r}.$$

The stability of (6.15) with respect to $\| \ \|_r$ is equivalent to

$$\|A^k\|_r \leq C, \quad \text{for all } k \geq 0.$$

6.2.2.1. *Stability in the L^∞ norm.* Introduce the **implicit scheme**

$$\frac{u_j^{k+1} - u_j^k}{\Delta t} + \gamma \frac{-u_{j-1}^{k+1} + 2u_j^{k+1} - u_{j+1}^{k+1}}{(\Delta x)^2} = 0 \tag{6.17}$$

for $k \geq 0$ and $j \in \{1, \ldots, N\}$. The scheme (6.17) is well defined since $u^{(k+1)}$ can be obtained from $u^{(k)}$ by inverting the definite positive matrix

$$\begin{pmatrix} 1+2\mu & -\mu & & & \\ -\mu & 1+2\mu & -\mu & & \\ & -\mu & 1+2\mu & -\mu & \\ & \ddots & \ddots & \ddots & \\ & & -\mu & 1+2\mu & -\mu \\ & & & -\mu & 1+2\mu \end{pmatrix}. \tag{6.18}$$

THEOREM 6.5. (i) *The explicit scheme (6.4) is stable with respect to the L^∞ norm if and only if the following Courant-Friedrichs-Lewy (CFL) condition holds:*

$$2\gamma \Delta t \leq (\Delta x)^2. \tag{6.19}$$

(ii) *The implicit scheme (6.17) is unconditionally stable with respect to the L^∞ norm.*

6.2.2.2. *Stability in the L^2 norm.* In order to investigate the stability of a finite difference scheme for solving the heat equation with the respect to the L^2 norm we consider (6.1) with the **periodic boundary conditions**

$$u(t, x+1) = u(t, x) \quad \text{for all } x \in [0,1], \quad t \geq 0.$$

For any $u^{(k)} = (u_j^k)_{j=0,\dots,N}$, we associate a piece-wise constant function $u^{(k)}(x)$, periodic with period 1, defined on $[0,1]$ by

$$u^{(k)}(x) := u_j^k \quad \text{for } x_{j-\frac{1}{2}} < x < x_{j+\frac{1}{2}},$$

where

$$x_{j+\frac{1}{2}} = (j + \frac{1}{2})\Delta x, \quad j = 0, \dots, N, \quad x_{-\frac{1}{2}} = 0, x_{N+1+\frac{1}{2}} = 1.$$

The Fourier series of $u^{(k)}$ reads

$$u^{(k)}(x) = \sum_{n \in \mathbb{Z}} \hat{u}_n^{(k)} e^{2\pi i n x},$$

where

$$\hat{u}_n^{(k)} := \int_0^1 u^{(k)}(x) e^{-2\pi i n x}\, dx.$$

Moreover, by **Plancherel's formula**, we have

$$\int_0^1 |u^{(k)}(x)|^2\, dx = \sum_{n \in \mathbb{Z}} |\hat{u}_n^{(k)}|^2.$$

Furthermore, an important property of Fourier series of periodic functions is that

$$v^{(k)}(x) = u^{(k)}(x + \Delta x) \Rightarrow \hat{v}_n^{(k)} = \hat{u}_n^{(k)} e^{2\pi i n \Delta x}.$$

With this notation, one can rewrite the explicit scheme (6.4) in the form

$$\frac{u^{k+1}(x) - u^k(x)}{\Delta t} + \gamma \frac{-u^k(x - \Delta x) + 2u^k(x) - u^k(x + \Delta x)}{(\Delta x)^2} = 0. \tag{6.20}$$

Applying the Fourier transform yields

$$\hat{u}_n^{(k+1)} = \left(1 - \frac{\gamma \Delta t}{(\Delta x)^2}(e^{-2\pi i n \Delta x} + 2 - e^{2\pi i n \Delta x})\right) \hat{u}_n^{(k)},$$

or equivalently,

$$\hat{u}_n^{(k+1)} = \alpha(n) \hat{u}_n^{(k)} = \alpha(n)^{k+1} \hat{u}^{(0)}(n) \quad \text{with } \alpha(n) := 1 - \frac{4\gamma \Delta t}{(\Delta x)^2} (\sin(\pi n \Delta x))^2.$$

Therefore, $\hat{u}_n^{(k)}$ is bounded as $k \to +\infty$ if and only if the amplification factor $\alpha(n)$ satisfies

$$|\alpha(n)| \leq 1 \quad \text{for all } n \in \mathbb{Z}.$$

Similarly, the implicit scheme (6.17) can be rewritten in the form

$$\frac{u^{k+1}(x) - u^k(x)}{\Delta t} + \gamma \frac{-u^{k+1}(x - \Delta x) + 2u^{k+1}(x) - u^{k+1}(x + \Delta x)}{(\Delta x)^2} = 0. \tag{6.21}$$

Again, by applying the Fourier transform, it follows that

$$\hat{u}_n^{(k+1)} = \beta(n) \hat{u}_n^{(k)} = \beta(n)^{k+1} \hat{u}^{(0)}(n),$$

where

$$\beta(n) := \left(1 + \frac{4\gamma \Delta t}{(\Delta x)^2} (\sin(\pi n \Delta x))^2\right)^{-1}.$$

THEOREM 6.6. (i) *The explicit scheme (6.4) is stable with respect to the L^2 norm if and only if the CFL condition (6.19) holds.*

(ii) *The implicit scheme (6.17) is unconditionally stable with respect to the L^2 norm.*

6.2.3. Convergence.

THEOREM 6.7 (**Lax theorem**). *Let u be a smooth solution of the heat equation (6.1). Suppose that the finite difference scheme for computing the numerical solution u_j^k is linear, consistent, and stable with respect to the norm $\| \ \|_r$. Let $e_j^k := u_j^k - u(t_k, x_j)$ and $e^{(k)} = (e_1^k, e_2^k, \dots, e_N^k)^\top$. Assume that $u_j^0 = u_0(x_j)$. Then,*

$$\lim_{\Delta t, \Delta x \to 0} \big(\sup_{t_k \le T} \|e^{(k)}\|_r \big) = 0 \quad \textit{for all } T > 0.$$

Moreover, if the scheme is of order p in time and q in space, then there exists a constant $C_T > 0$ such that

$$\sup_{t_k \le T} \|e^{(k)}\|_r \le C_T \big((\Delta t)^p + (\Delta x)^q \big).$$

PROOF. Let $u^{(k+1)} = Au^{(k)}$, where A is the iteration matrix, and let $\widetilde{u}_j^k = u(t_k, x_j)$. Since the scheme is consistent, there exists $\epsilon^{(k)}$ such that

$$\widetilde{u}^{(k+1)} = A\widetilde{u}^{(k)} + (\Delta t)\epsilon^{(k)} \quad \text{and} \quad \lim_{\Delta t, \Delta x \to 0} \|\epsilon^{(k)}\|_r = 0, \tag{6.22}$$

uniformly in k. If the scheme is of order p in time and q in space, then

$$\|\epsilon^{(k)}\|_r \le C((\Delta t)^p + (\Delta x)^q \big).$$

By subtracting (6.22) from (6.15), we obtain

$$e^{(k+1)} = Ae^{(k)} - \Delta t \epsilon^k, \tag{6.23}$$

and therefore, by induction,

$$e^{(k)} = A^k e^{(0)} - \Delta t \sum_{l=1}^{k} A^{k-l} \epsilon^{l-1}. \tag{6.24}$$

The stability of the scheme yields

$$\|A^k\|_r \le C'$$

for some positive constant C'. Therefore, since $e^{(0)} = 0$, (6.24) yields

$$\|e^{(k)}\|_r \le (\Delta t) k C C' ((\Delta t)^p + (\Delta x)^q \big) \le T C C' ((\Delta t)^p + (\Delta x)^q \big). \tag{6.25}$$

The proof is then complete. □

6.3. Numerical algorithms for the wave equation

Consider the wave equation

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2} = c^2 \dfrac{\partial^2 u}{\partial x^2}, \quad 0 < x < 1, \quad t \ge 0, \\ u(t, x+1) = u(t, x), \quad 0 < x < 1, \quad t \ge 0, \\ u(0, x) = u_0(x), \quad 0 < x < 1, \\ \dfrac{\partial u}{\partial t}(0, x) = u_1(x), \quad 0 < x < 1, \end{cases} \tag{6.26}$$

where $c > 0$ is the wave speed.

Suppose that

$$\int_0^1 u_1(x)\, dx = 0. \tag{6.27}$$

A standard finite difference scheme for solving (6.26) is the **θ-centered scheme**

$$\begin{cases} \dfrac{u_j^{k+1}-2u_j^k+u_j^{k-1}}{(\Delta t)^2}+\theta c^2\dfrac{-u_{j-1}^{k+1}+2u_j^{k+1}-u_{j+1}^{k+1}}{(\Delta x)^2} \\ \quad +(1-2\theta)c^2\dfrac{-u_{j-1}^{k}+2u_j^{k}-u_{j+1}^{k}}{(\Delta x)^2}+\theta c^2\dfrac{-u_{j-1}^{k-1}+2u_j^{k-1}-u_{j+1}^{k-1}}{(\Delta x)^2}=0, \end{cases} \tag{6.28}$$

where $0 \le \theta \le 1/2$.

If $\theta = 0$, then the scheme is explicit, while it is implicit if $\theta \neq 0$.

The initial conditions can be expressed by

$$u_j^0 = u_0(x_j) \quad \text{and} \quad \frac{u_j^1 - u_j^0}{\Delta t} = \int_{x_{j-1/2}}^{x_{j+1/2}} u_1(x)\,dx,$$

which shows that (6.27) is satisfied by the numerical solution.

THEOREM 6.8. *If $1/4 \le \theta \le 1/2$, then the θ-centered scheme (6.28) is unconditionally stable with respect to the L^2 norm. If $0 \le \theta < 1/4$, (6.28) is stable provided that the CFL condition*

$$\frac{c\Delta t}{\Delta x} < \sqrt{\frac{1}{1-4\theta}} \tag{6.29}$$

holds and is unstable if $c\Delta t/\Delta x > 1/\sqrt{1-4\theta}$.

Index