Chapter 7

Numerical ordinary differential equations

Introduction

This chapter concerns the numerical solution of ordinary differential equations (ODEs) of the following form:

$$\begin{cases} \boldsymbol{x}'(t) = \boldsymbol{f}\big(t, \boldsymbol{x}(t)\big), \\ \boldsymbol{x}(t_0) = \boldsymbol{x}_0. \end{cases} \tag{7.1}$$

Here $\boldsymbol{f}\colon \mathbf{R} \times \mathbf{R}^n \to \mathbf{R}$ and $\boldsymbol{x}_0$ is the initial condition. Equations of this type are the building blocks of a plethora mathematical models in science and engineering. They have applications in celestial dynamics, molecular simulation and fluid mechanics, to mention just a few. Ordinary differential equations also arise after discretization of time-dependent partial differential equations, which are also ubiquitous in science. More often than not, it is not possible to find an explicit solution of (7.1), and so one has to resort to numerical simulation. The rest of chapter is organized as follows:

- In Section 7.1, we define the concepts of local and global solutions for the continuous-time problem (7.1), and we recall fundamental results concerning the existence and uniqueness of a solution.
- In Section 7.2, we analyze the so-called *one-step* numerical methods to solve (7.1). We emphasize in particular the concepts of *consistency*, *stability* and *convergence*.
- In Section 7.3, we present *multistep* methods to solve (7.1), and discuss their drawbacks and advantages compared to one-step methods.
- Finally, in Section 7.4, we introduce the concept of *absolute stability* and discuss its relevance in the context of *stiff* differential equations.

7.1 Analysis of the continuous problem

A differentiable function $\boldsymbol{x}\colon I \to \mathbf{R}^n$, where I denotes an interval of $\mathbf{R}$ containing t_0, is a solution of (7.1) if $\boldsymbol{x}(t_0) = \boldsymbol{x}_0$ and the equation (7.1) is satisfied for all $t \in I$. The solution is called global if $I = \mathbf{R}$, and local otherwise.

Integral formulation. If $\boldsymbol{x}$ is a solution to (7.1), then it holds that

$$\forall t \in I, \qquad \boldsymbol{x}(t) = \boldsymbol{x}_0 + \int_{t_0}^{t} \boldsymbol{f}\big(s, \boldsymbol{x}(s)\big)\,\mathrm{d}s. \tag{7.2}$$

The converse statement is not true in general, because a solution to (7.2) need not necessarily be differentiable everywhere. However, if the integral formulation (7.2) holds, then necessarily $\boldsymbol{x}$ is absolutely continuous and (7.1) is satisfied for almost every t. Additionally, if (7.2) is satisfied and the function $\boldsymbol{f}$ is continuous, then the function $s \mapsto \boldsymbol{f}\big(s, \boldsymbol{x}(s)\big)$ is continuous, and so (7.1) is satisfied for all $t \in I$ by the fundamental theorem of analysis. We now focus on the integral formulation (7.2), and begin by establishing existence of a local solution.

Theorem 7.1 (Existence of a solution)**.** *Let $\boldsymbol{x}_0 \in \mathbf{R}^n$ and let $\Omega_{\mathcal{T},\mathcal{R}}$ denote the set*

$$\big\{(t, \boldsymbol{x}) \in \mathbf{R} \times \mathbf{R}^n : |t - t_0| \leqslant \mathcal{T} \text{ and } \|\boldsymbol{x} - \boldsymbol{x}_0\| \leqslant \mathcal{R}\big\},$$

Assume that the following conditions are satisfied for some $\mathcal{T} > 0$ and $\mathcal{R} > 0$:

- *The function $\boldsymbol{f}$ is uniformly bounded on $\Omega_{\mathcal{T},\mathcal{R}}$:*

$$\forall (t, \boldsymbol{x}) \in \Omega_{\mathcal{T},\mathcal{R}}, \qquad \|\boldsymbol{f}(t, \boldsymbol{x})\| \leqslant M. \tag{7.3}$$

- *The function $\boldsymbol{f}$ satisfies the following Lipschitz condition: there is $L > 0$ such that*

$$\forall \big((t, \boldsymbol{x}_1), (t, \boldsymbol{x}_2)\big) \in \Omega_{\mathcal{T},\mathcal{R}} \times \Omega_{\mathcal{T},\mathcal{R}}, \qquad \|\boldsymbol{f}(t, \boldsymbol{x}_1) - \boldsymbol{f}(t, \boldsymbol{x}_2)\| \leqslant L\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|. \tag{7.4}$$

Then there exists $T \in (0, \mathcal{T}]$ depending on $\mathcal{R}$, M and L such that the differential equation (7.2) has a local solution $\boldsymbol{x}\colon [t_0 - T, t_0 + T] \to \mathbf{R}^n$.

Proof. Fix $T \in (0, \mathcal{T}]$ and let $I = [t_0 - T, t_0 + T]$. Let also $\mathcal{X}$ denote the following subset of continuous functions defined from I to $\mathbf{R}^n$:

$$\mathcal{X} := \left\{\boldsymbol{x} \in C\left(I, \mathbf{R}^n\right) : \sup_{t \in I} \big\|\boldsymbol{x}(t) - \boldsymbol{x}_0\big\| \leqslant \mathcal{R}\right\}$$

The set $\mathcal{X}$ endowed with supremum metric is a closed subset of $C(I, \mathbf{R}^n)$. Since $\mathcal{X}$ is a closed subset of a complete metric space, it is itself complete. Let $\Phi\colon \mathcal{X} \to C(I, \mathbf{R}^n)$ denote the mapping

$$\Phi(\boldsymbol{x})\colon t \mapsto \boldsymbol{x}_0 + \int_0^t \boldsymbol{f}\big(s, \boldsymbol{x}(s)\big)\,\mathrm{d}s.$$

The right-hand side, being the integral of a bounded function, is indeed a continuous function. We will show that, for T sufficiently small,

- the mapping Φ maps $\mathcal{X}$ into $\mathcal{X}$;
- the mapping Φ is a contraction.

From (7.3), it follows that

$$\forall \boldsymbol{x} \in \mathcal{X}, \qquad \forall t \in I, \qquad \|\Phi(\boldsymbol{x})(t) - \boldsymbol{x}_0\| = \left\| \int_{t_0}^{t} \boldsymbol{f}\big(s, \boldsymbol{x}(s)\big)\, \mathrm{d}s \right\| \leqslant MT.$$

On the other hand, from the Lipschitz condition (7.4), it holds that

$$\forall (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathcal{X}, \qquad \|\Phi(\boldsymbol{x}) - \Phi(\boldsymbol{y})\| \leqslant LT \|\Phi(\boldsymbol{x}) - \Phi(\boldsymbol{y})\|.$$

Therefore, it suffices to take $T < \min\left\{\mathcal{T}, \frac{\mathcal{R}}{M}, \frac{1}{L}\right\}$ to ensure that the above conditions are satisfied. For a value of T in this range, the Banach fixed point theorem, Theorem A.3, gives the existence of a unique fixed point $\boldsymbol{x}_* \in \mathcal{X}$ of Φ. Since a fixed point of Φ is a solution to (7.2) in view of the definition of Φ, the statement is proved. □

It may seem at first glance that uniqueness of the solution to (7.2) follows from the uniqueness of the fixed point guaranteed by Theorem A.3. However, this theorem implies uniqueness *only in the set* $\mathcal{X}$, a property known as *conditional uniqueness*. In order to prove that the solution is unique over the full space $C([t_0 - T, t_0 + T], \mathbf{R}^n)$, additional assumptions and arguments are required. A simple approach is to rely on Grönwall's lemma.

Lemma 7.2 (Grönwall's lemma, simplified integral form)**.** *Suppose that* $u\colon [t_0 - T, t_0 + T] \to \mathbf{R}_{\geqslant 0}$ *is continuous, nonnegative, and satisfies*

$$\forall t \in [t_0, t_0 + T], \qquad u(t) \leqslant \alpha + \int_{t_0}^{t} \beta(s)\, u(s)\, \mathrm{d}s, \tag{7.5}$$

where $\alpha \geqslant 0$ *and* $\beta\colon [t_0, t_0 + T] \to \mathbf{R}_{\geqslant 0}$ *is continuous and nonnegative. Then*

$$\forall t \in [t_0, t_0 + T], \qquad u(t) \leqslant \alpha \exp\left(\int_{t_0}^{t} \beta(s)\, \mathrm{d}s \right). \tag{7.6}$$

Proof. Assume first that $\alpha > 0$, so that the logarithm in (7.7) is well-defined. By the fundamental theorem of calculus and (7.5), it holds that

$$\frac{\mathrm{d}}{\mathrm{d}t} \left(\alpha + \int_{t_0}^{t} \beta(s)\, u(s)\, \mathrm{d}s \right) \leqslant \beta(t) \left(\alpha + \int_{t_0}^{t} \beta(s)\, u(s)\, \mathrm{d}s \right)$$

Therefore we have

$$\frac{\mathrm{d}}{\mathrm{d}t} \log \left(\alpha + \int_{t_0}^{t} \beta(s)\, u(s)\, \mathrm{d}s \right) \leqslant \beta(t), \tag{7.7}$$

and after integrating and exponentiating, we obtain

$$\alpha + \int_{t_0}^{t} \beta(s)\, u(s)\, \mathrm{d}s \leqslant \alpha \exp\left(\int_{t_0}^{t} \beta(s)\, \mathrm{d}s \right)$$

The statement then follows by using (7.5) again. Assume next that $\alpha = 0$. If (7.5) is satisfied for $\alpha = 0$, then this condition is also satisfied for all $\alpha > 0$. Therefore the conclusion (7.6) holds for all $\alpha > 0$, and taking the limit $\alpha \to 0$ in this equation, we obtain the statement. □

Note that the estimate (7.6) is sharp, since the function $v\colon [t_0, t_0+T]$ given by

$$v(t) = \alpha \exp\left(\int_{t_0}^{t} \beta(s)\,\mathrm{d}s\right)$$

satisfies (7.5) with equality. We are now ready to prove uniqueness under an appropriate condition.

Theorem 7.3 (Uniqueness of the solution)**.** *Let $\boldsymbol{x}_0 \in \mathbf{R}^n$ and let*

$$\Omega_{\mathcal{T},\mathcal{R}}\big\{(t,\boldsymbol{x}) \in \mathbf{R}\times\mathbf{R}^n : |t-t_0| \leqslant \mathcal{T} \text{ and } \|\boldsymbol{x}-\boldsymbol{x}_0\| \leqslant \mathcal{R}\big\},$$

Assume that for all $\mathcal{T} \in \mathbf{R}_{>0}$ and $\mathcal{R} \in \mathbf{R}_{>0}$, there is $L_{\mathcal{T},\mathcal{R}}$ such that

$$\forall\big((t,\boldsymbol{x}_1),(t,\boldsymbol{x}_2)\big) \in \Omega_{\mathcal{T},\mathcal{R}}\times\Omega_{\mathcal{T},\mathcal{R}}, \qquad \|\boldsymbol{f}(t,\boldsymbol{x}_1)-\boldsymbol{f}(t,\boldsymbol{x}_2)\| \leqslant L_{\mathcal{T},\mathcal{R}}\|\boldsymbol{x}_1-\boldsymbol{x}_2\|. \tag{7.8}$$

Then if $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ in $C\big([t_0-T,t_0+T],\mathbf{R}^n\big)$ are local solutions to (7.2), it holds that $\boldsymbol{x}_1 = \boldsymbol{x}_2$.

Proof. Suppose that $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are solutions to (7.2). Let $I = [t_0-T, t_0+T]$ and

$$R := \max\left\{\sup_{t\in I}\|\boldsymbol{x}_1(t)-\boldsymbol{x}_0\|, \sup_{t\in I}\|\boldsymbol{x}_2(t)-\boldsymbol{x}_0\|\right\} < \infty,$$

Since $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are solutions, it holds that

$$\forall t \in [t_0-T,t_0+T], \qquad \boldsymbol{x}_1(t)-\boldsymbol{x}_2(t) = \int_{t_0}^{t}\Big(f\big(s,\boldsymbol{x}_1(s)\big)-f\big(s,\boldsymbol{x}_2(s)\big)\Big)\,\mathrm{d}s.$$

Taking the norm and using (7.8), we obtain

$$\forall t \in [t_0,t_0+T], \qquad \|\boldsymbol{x}_1(t)-\boldsymbol{x}_2(t)\| \leqslant L_{T,R}\int_{t_0}^{t}\|\boldsymbol{x}_1(s)-\boldsymbol{x}_2(s)\|\,\mathrm{d}s$$

Using Grönwall's lemma, we deduce that $\boldsymbol{x}_1(t) = \boldsymbol{x}_2(t)$ for all $t \in [t_0, t_0+T]$. A similar argument can be employed to show that $\boldsymbol{x}_1 = \boldsymbol{x}_2$ on $[t_0-T, t_0]$. □

Corollary 7.4 (Maximal solutions)**.** *Assume that $\boldsymbol{f}$ is continuous in t and satisfies the local Lipschitz condition (7.8). Then there exists $-\infty \leqslant T_- < T_+ \leqslant \infty$ such that $t_0 \in (T_-,T_+)$ and the following properties are satisfied.*

- *there exists a solution $\boldsymbol{x}_*\colon (T_-,T_+) \to \mathbf{R}^n$ to (7.2);*
- *if $\boldsymbol{x}\colon I \to \mathbf{R}^n$ is a local solution of (7.2), then $I \subset (T_-,T_+)$ and $\boldsymbol{x}(t) = \boldsymbol{x}_*(t)$ for all $t \in I$.*
- *If T_+ is finite, then $\lim_{t\to T_+}\big\|\boldsymbol{x}(t)\big\| = \infty$, and if T_- is finite, then $\lim_{t\to T_-}\big\|\boldsymbol{x}(t)\big\| = \infty$.*

The solution $\boldsymbol{x}_*$ is called the maximal solution of (7.2).

Proof. Let $\mathcal{I}$ denote the union of all the open intervals I such that there exists a solution

in $C(I, \mathbf{R}^n)$ to (7.2). The open set $\mathcal{I}$ is connected and, by Theorem 7.1, it contains a neighborhood of t_0. Therefore $\mathcal{I}$ is of the form (T_-, T_+), where $-\infty \leqslant T_- < t_0 < T_+ \leqslant \infty$. In view of Theorem 7.3, all the local solutions coincide where they are defined, and so they can be patched together in order to construct a solution $\boldsymbol{x}_*\colon (T_-, T_+) \to \mathbf{R}$. It remains to prove the third item. To this end, suppose for contradiction that T_+ was finite and that there was $(t_n)_{n\in\mathbf{N}}$ such that $t_n \to T_+$ in the limit $n \to \infty$ and

$$K := \sup_{n\in\mathbf{N}} \|\boldsymbol{x}_*(t_n)\| < \infty.$$

Since $\boldsymbol{f}$ is continuous, there is M such that $|f(t, \boldsymbol{x})|$ is uniformly bounded from above by M for all $(t, \boldsymbol{x}) \in [T_- - 1, T_+ + 1] \times B_{K+1}(\mathbf{0})$. Furthermore, by the assumption (7.8), there is L such that for all $t \in [T_- - 1, T_+ + 1]$, the following Lipschitz condition holds:

$$\forall (\boldsymbol{x}_1, \boldsymbol{x}_2) \in B_{K+1}(\mathbf{0}) \times B_{K+1}(\mathbf{0}), \qquad \|\boldsymbol{f}(t, \boldsymbol{x}_1) - \boldsymbol{f}(t, \boldsymbol{x}_2)\| \leqslant L\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|.$$

Consequently, Theorem 7.1 with $\mathcal{T} = \mathcal{R} = 1$ implies for all n the existence of a solution to

$$\begin{cases} \boldsymbol{x}'(t) = \boldsymbol{f}\big(t, \boldsymbol{x}(t)\big), \\ \boldsymbol{x}(t_n) = \boldsymbol{x}_*(t_n). \end{cases}$$

over the time interval $[t_n - T, t_n + T]$, where $T > 0$ depends only on M and L, and not on n. But then, for n sufficiently large, this solution extends beyond T_+, which contradicts the maximality of $\mathcal{I}$. An analogous reasoning can be employed for T_-. □

Example 7.1. Consider the ODE

$$\begin{cases} x'(t) = x(t)^2, \\ x(0) = 1. \end{cases}$$

The maximal solution is $x_*\colon (-\infty, 1) \to \mathbf{R}$ given by

$$\boldsymbol{x}_*(t) = \frac{1}{1 - t}.$$

Existence of a unique global solution. In certain settings, it is possible to prove the maximal solution to (7.2) is globally defined for any initial condition. We discuss a few important examples.

- The first case is when $\boldsymbol{f}\colon \mathbf{R} \times \mathbf{R}^n$ is globally Lipschitz in its second argument, with a Lipschitz constant that depends continuously on the first argument.
- The second case, generalizing the first, is when the growth of $\boldsymbol{f}(t, \bullet)$ is at most affine:

  $$\forall (t, \boldsymbol{x}) \in \mathbf{R} \times \mathbf{R}^n, \qquad \|\boldsymbol{f}(t, \boldsymbol{x})\| \leqslant C(t) + L(t)\|\boldsymbol{x}\|,$$

  with continuous constants $C(t)$ and $L(t)$.

- The third case is when $\boldsymbol{f}$ is independent of t and there is a function $W \in C^1(\mathbf{R}^n)$ such that $W(\boldsymbol{x}) \to \infty$ in the limit as $\|\boldsymbol{x}\| \to \infty$ and

$$\forall \boldsymbol{x} \in \mathbf{R}^n, \qquad \nabla W(\boldsymbol{x}) \cdot \boldsymbol{f}(\boldsymbol{x}) \leqslant c < \infty$$

  Such a function is called a *Lyapunov function*.

The strategy of proof for global existence usually relies on an argument by contradiction. Consider for example the third setting. Since the assumptions of Corollary 7.4 are satisfied, there exists a maximal solution $\boldsymbol{x}_* \colon (T_-, T_+) \to \infty$. Assume for contradiction that T_+ is finite. Then the third item in Corollary 7.4 implies that $\lim_{t \to T_+} \|\boldsymbol{x}_*(t)\| \to \infty$, and so $W\big(\boldsymbol{x}_*(t)\big)$ blows up as t approaches T^+. On the other hand, we have

$$\frac{\mathrm{d}}{\mathrm{d}t} W\big(\boldsymbol{x}_*(t)\big) = \nabla W\big(\boldsymbol{x}_*(t)\big) \cdot \boldsymbol{f}\big(\boldsymbol{x}_*(t)\big) \leqslant c.$$

Therefore $\lim_{t \to T_+} W\big(\boldsymbol{x}_*(t)\big) \leqslant W\big(\boldsymbol{x}_*(t_0)\big) + |c|(T_+ - t_0)$, which is a contradiction.

7.2 One-step methods

From now on, we assume for simplicity that $t_0 = 0$ and that the initial value problem (7.1) admits a unique solution over the interval $[0, T]$. Most numerical methods for ODEs construct an approximation of the solution at discrete points:

$$\boldsymbol{x}_n \approx \boldsymbol{x}(t_n), \qquad n = 0, 1, 2, \dots .$$

The discretization points $(t_n)_{n \in \mathbf{N}}$ are commonly equidistant, i.e. $t_n = n\Delta$ where Δ is the *discretization step*. Sometimes, it is useful to employ a variable time step, but we assume throughout this section that the time step is fixed, for simplicity. We begin in Section 7.2.1 and Section 7.2.2 by studying the simplest one-step methods, namely the forward and backward Euler methods. Then, in Section 7.2.3, we present a general approach to the analysis of one-step methods. Finally, in Section 7.2.4, we present other widely used one-step methods in applications.

7.2.1 Forward Euler method

Assume that (7.1) has a unique solution $\boldsymbol{x}(t)$ over the interval $[0, T]$. If $\boldsymbol{x}(t)$ is twice continuously differentiable, then by Taylor's formula, we have

$$\boldsymbol{x}(t + \Delta) = \boldsymbol{x}(t) + \Delta \boldsymbol{f}(t, \boldsymbol{x}) + \frac{\Delta^2}{2} \boldsymbol{x}''(\tau), \qquad \tau \in (t, t + \Delta). \tag{7.9}$$

This motivates a method known as the *forward* or *explicit* Euler method:

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n + \Delta \boldsymbol{f}(t_n, \boldsymbol{x}_n),$$

with the same initial condition as for the continuous equation (7.1). The convergence of this method can be proved under a global Lipschitz assumption on the function $\boldsymbol{f}$.

Theorem 7.5 (Convergence of the forward Euler method)**.** *Assume that there is $L \in \mathbf{R}_{>0}$ such that*

$$\forall (t, \boldsymbol{x}, \boldsymbol{y}) \in \mathbf{R} \times \mathbf{R}^n \times \mathbf{R}^n, \qquad \|\boldsymbol{f}(t, \boldsymbol{x}) - \boldsymbol{f}(t, \boldsymbol{y})\| \leqslant L\|\boldsymbol{x} - \boldsymbol{y}\|. \tag{7.10}$$

Suppose in addition that there exists a unique, twice continuously differentiable of (7.1) over the interval $[0, T]$, and let

$$M = \sup_{t \in [0,T]} \|\boldsymbol{x}''(t)\|$$

Then the following error estimate holds:

$$\forall n \in \left\{0, 1, \ldots, \left\lfloor \frac{T}{\Delta} \right\rfloor\right\}, \qquad \|\boldsymbol{x}(t_n) - \boldsymbol{x}_n\| \leqslant \frac{\Delta M}{2}\left(\frac{\mathrm{e}^{Lt_n} - 1}{L}\right). \tag{7.11}$$

Proof. By Taylor's theorem, it holds that

$$\boldsymbol{x}(t_n) = \boldsymbol{x}(t_{n-1}) + \Delta \boldsymbol{f}\big(t_{n-1}, \boldsymbol{x}(t_{n-1})\big) + \frac{\Delta^2}{2}\boldsymbol{\alpha}_n, \qquad \boldsymbol{\alpha}_n := 2\int_0^1 (1-s)\boldsymbol{x}''(t_n + \Delta s)\,\mathrm{d}s.$$

Notice that that $\|\boldsymbol{\alpha}_n\| \leqslant M$. Therefore, it holds that

$$\begin{aligned}
\boldsymbol{x}(t_n) - \boldsymbol{x}_n &= \left(\boldsymbol{x}(t_{n-1}) + \Delta\, \boldsymbol{f}\big(t_{n-1}, \boldsymbol{x}(t_{n-1})\big) + \frac{\Delta^2}{2}\boldsymbol{\alpha}_n\right) - \Big(\boldsymbol{x}_{n-1} + \Delta\, \boldsymbol{f}\big(t_{n-1}, \boldsymbol{x}_{n-1}\big)\Big) \\
&= \big(\boldsymbol{x}(t_{n-1}) - \boldsymbol{x}_{n-1}\big) + \Delta\Big(\boldsymbol{f}\big(t_{n-1}, \boldsymbol{x}(t_{n-1})\big) - \boldsymbol{f}\big(t_{n-1}, \boldsymbol{x}_{n-1}\big)\Big) + \frac{\Delta^2}{2}\boldsymbol{\alpha}_n,
\end{aligned}$$

Let $\boldsymbol{e}_n = \boldsymbol{x}(t_n) - \boldsymbol{x}_n$ and $\boldsymbol{\varepsilon}_n = \frac{\Delta^2}{2}\boldsymbol{\alpha}_n$. The first term is the error at iteration $n-1$, and the second may be bounded from (7.10), which gives

$$\|\boldsymbol{e}_n\| \leqslant (1 + \Delta L)\|\boldsymbol{e}_{n-1}\| + \|\boldsymbol{\varepsilon}_n\|.$$

The structure of this equation is important, as it appears in the analysis of all one-step methods for ODEs. The first term is an amplification of the error at the previous iteration, and the second term is an upper bound on the additional error introduced at step n. Applying this inequality to the previous time steps, we obtain

$$\begin{aligned}
\|\boldsymbol{e}_n\| &\leqslant (1 + \Delta L)\Big((1 + \Delta L)\|\boldsymbol{e}_{n-2}\| + \|\boldsymbol{\varepsilon}_{n-1}\|\Big) + \|\boldsymbol{\varepsilon}_n\| \\
&\leqslant \ldots \leqslant (1 + \Delta L)^n\|\boldsymbol{e}_0\| + \sum_{i=1}^{n}(1 + \Delta L)^{n-i}\|\boldsymbol{\varepsilon}_i\|.
\end{aligned} \tag{7.12}$$

Since $\|\boldsymbol{\varepsilon}_i\| \leqslant \Delta^2 M/2$, we have by using the formula for geometric series that

$$\|\boldsymbol{e}_n\| \leqslant (1 + \Delta L)^n\|\boldsymbol{e}_0\| + \frac{(1 + \Delta L)^n - 1}{\Delta L}\left(\frac{\Delta^2 M}{2}\right).$$

The first term is zero because $\|\boldsymbol{e}_0\| = 0$. Using the bound $(1 + \Delta L)^n \leqslant \big(\exp(\Delta L)\big)^n = \mathrm{e}^{Lt_n}$ in

the second term and rearranging, we finally obtain the statement (7.11). □

7.2.2 Backward Euler method

If we apply the Taylor expansion (7.9) backward around $t + \Delta$, instead of forward around t, then we obtain

$$\boldsymbol{x}(t) = \boldsymbol{x}(t + \Delta) - \Delta \boldsymbol{f}(t + \Delta, \boldsymbol{x}) + \mathcal{O}(\Delta^2).$$

This motivates the so-called *backward* or *implicit* Euler method:

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n + \Delta \boldsymbol{f}(t_{n+1}, \boldsymbol{x}_{n+1}). \tag{7.13}$$

Observe that the right-hand side depends on $\boldsymbol{x}_{n+1}$. Therefore, given t_n and $\boldsymbol{x}_n$, this is a nonlinear equation for the unknown $\boldsymbol{x}_{n+1}$, which can be solved by using any of the methods studied in Chapter 5. Finding a solution to (7.13) amounts to finding a fixed point of the function

$$\boldsymbol{y} \mapsto \boldsymbol{F}(\boldsymbol{y}) := \boldsymbol{x}_n + \Delta \boldsymbol{f}(t_n, \boldsymbol{y}).$$

A priori, the existence and uniqueness of such a fixed point is not guaranteed. We proved in Theorem 5.2 that a sufficient condition for these two properties to hold is that $\boldsymbol{F}$ is globally Lipschitz with a constant strictly less than 1, which holds if and only if the function $\boldsymbol{y} \mapsto \boldsymbol{f}(t_n, \boldsymbol{y})$ is globally Lipschitz with a constant strictly less than $1/\Delta$. If the condition (7.10) holds, for example, then the backward Euler method (7.13) is guaranteed to be well defined for $\Delta < \frac{1}{L}$. Theorem 5.2 also ensures that, if $\boldsymbol{F}$ is globally Lipschitz with a constant less than 1, then the fixed point can be approximated by using the iteration

$$\boldsymbol{y}_{k+1} = \boldsymbol{F}(\boldsymbol{y}_k). \tag{7.14}$$

and there is exponential convergence $\boldsymbol{y}_k \to \boldsymbol{x}_{n+1}$ in the limit as $k \to \infty$. A natural starting point for (7.14) is $\boldsymbol{y}_0 = \boldsymbol{x}_n$. An alternative approach to the fixed point iteration (7.14) is to use the Newton–Raphson method for (7.13), which is faster in principle but must be initialized sufficiently close to the fixed point.

Using a reasoning similar to that employed for proving Theorem 7.5, we can prove the following result.

Theorem 7.6 (Convergence of the backward Euler method)**.** *If the assumptions of Theorem 7.5 hold and $\Delta < \frac{1}{L}$, then the following error estimate holds:*

$$\forall n \in \left\{0, 1, \dots, \left\lfloor \frac{T}{\Delta} \right\rfloor\right\}, \qquad \|\boldsymbol{x}(t_n) - \boldsymbol{x}_n\| \leqslant \frac{\Delta M}{2} \left(\frac{\left(\frac{1}{1-\Delta L}\right)^n - 1}{L} \right). \tag{7.15}$$

Proof. The proof is left as an exercise. □

Remark 7.1. Note that, if $\Delta < \frac{1}{2L}$, then

$$\begin{aligned}\frac{1}{1-\Delta L} &= 1 + \Delta L + (\Delta L)^2 + (\Delta L)^3 + (\Delta L)^4 \dots \\ &\leqslant 1 + \Delta L + (\Delta L)^2 + \frac{1}{2}(\Delta L)^2 + \frac{1}{4}(\Delta L)^2 + \dots \\ &\leqslant 1 + \Delta L + 2(\Delta L)^2 \leqslant \exp\bigl(\Delta L + (\Delta L)^2\bigr),\end{aligned}$$

and so the error estimate (7.15) gives

$$\|\boldsymbol{x}(t_n) - \boldsymbol{x}_n\| \leqslant \frac{\Delta M}{2}\left(\frac{\exp(Lt_n + \Delta L^2 t_n) - 1}{L}\right),$$

which makes it clear that the right-hand side of (7.15) is close, in absolute and relative terms, to that of (7.11) when $\Delta \ll 1$.

At this point, the reader may be wondering why one would use the backward Euler method instead of the forward Euler method, given that both methods have same order of convergence but iterations of the former are more computationally costly. The reason is that the backward Euler method, like many implicit methods, is more *stable* than its forward counterpart. Implicit methods are especially attractive in the context of *stiff* differential equations. We shall elaborate on this subject in Section 7.4.

7.2.3 Analysis of general one-step methods

In general, one-step methods to solve differential equations are of the abstract form

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n + \Delta \boldsymbol{\Phi}_\Delta(t_n, \boldsymbol{x}_n). \tag{7.16}$$

where $\boldsymbol{\Phi}_\Delta \colon \mathbf{R} \times \mathbf{R}^n \to \mathbf{R}^n$ is a function such that

$$\boldsymbol{\Phi}_\Delta(t, \boldsymbol{x}) \approx \frac{1}{\Delta}\int_t^{t+\Delta} \boldsymbol{f}\bigl(s, \boldsymbol{x}^{t,\boldsymbol{x}}(s)\bigr)\,\mathrm{d}s = \frac{\boldsymbol{x}^{t,\boldsymbol{x}}(t+\Delta) - \boldsymbol{x}}{\Delta}. \tag{7.17}$$

Here $\boldsymbol{x}^{t,\boldsymbol{x}}$ denotes the solution to the differential equation (7.1) with initial condition $\boldsymbol{x}(s) = \boldsymbol{x}$. The main goal of this section is to establish general conditions, known as *consistency* and *stability*, under which the numerical scheme (7.16) is convergent. As we observed in the proof of Theorem 7.5 – specifically in equation (7.12) – the error at the final iteration for the forward Euler method is a sum of local errors, each amplified by a factor depending to the number of iterations left to reach the final time. *Consistency* of a numerical method enables to control the size of local errors when they arise, while *stability* enables to control their growth.

We emphasize that both the forward and the backward Euler methods can be recast in the form (7.16). For the forward Euler method $\boldsymbol{\Phi}_\Delta(t, \boldsymbol{x}) = \boldsymbol{f}(t, \boldsymbol{x})$, while for the backward Euler method, the function $\boldsymbol{\Phi}_\Delta$ is defined implicitly as the function which to $(t, \boldsymbol{x})$ associates the solution $\boldsymbol{\phi} \in \mathbf{R}^n$ to the equation

$$\boldsymbol{\phi} = \boldsymbol{f}(t + \Delta, \boldsymbol{x} + \Delta\boldsymbol{\phi}).$$

Local truncation error and consistency

The local truncation error is the residual error obtained when substituting the exact solution of the differential equation in (7.16):

$$\boldsymbol{\eta}_{n+1} := \frac{\boldsymbol{x}(t_{n+1}) - \boldsymbol{x}(t_n)}{\Delta} - \boldsymbol{\Phi}_\Delta\big(t_n, \boldsymbol{x}(t_n)\big).$$

Since there is a division by Δ, the local truncation error has the same physical dimension as that of $\boldsymbol{x}'$, and so it should be viewed as an error *per time unit*.

Definition 7.1 (Consistency)**.** A numerical method is consistent if

$$\lim_{\Delta \to 0} \left(\max_{1 \leqslant n \leqslant N} \|\boldsymbol{\eta}_n\| \right) = 0, \qquad N = \left\lfloor \frac{T}{\Delta} \right\rfloor.$$

It is consistent with order p if there exists C such that

$$\forall \Delta > 0, \qquad \max_{1 \leqslant n \leqslant N} \|\boldsymbol{\eta}_n\| \leqslant C\Delta^p.$$

Proving the consistency of a numerical method is usually achieved on a case-by-case basis by application of Taylor's formula.

Stability

The stability of a numerical method qualifies its sensitivity to perturbations. Roughly speaking, it expresses that small perturbation of the right-hand side of (7.16) lead to small perturbations of the numerical solution.

Definition 7.2 (Stability)**.** A numerical method of the form (7.16) is stable if there exists a constant $S(T) > 0$ independent of Δ such that for all sequence $(\boldsymbol{y}_n)_{1 \leqslant n \leqslant N}$ satisfying

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + \Delta \boldsymbol{\Phi}_\Delta(t_n, \boldsymbol{y}_n) + \Delta \boldsymbol{\delta}_{n+1}, \qquad \boldsymbol{y}_0 = \boldsymbol{x}_0, \tag{7.18}$$

it holds that

$$\max_{1 \leqslant n \leqslant N} \|\boldsymbol{x}_n - \boldsymbol{y}_n\| \leqslant S(T)\Delta \sum_{n=1}^{N} \|\boldsymbol{\delta}_n\|. \tag{7.19}$$

It is convenient to introduce the following norms for sequences of vectors $(\boldsymbol{u}_n)_{1 \leqslant n \leqslant N}$:

$$\|\boldsymbol{u}_\bullet\|_{\ell^1_T} = \sum_{n=1}^{N} \|\boldsymbol{u}_n\|, \qquad \|\boldsymbol{u}_\bullet\|_{\ell^\infty_T} = \max_{1 \leqslant n \leqslant N} \|\boldsymbol{u}_n\|,$$

With these notations, equation (7.19) may be rewritten compactly as follows:

$$\|\boldsymbol{x}_\bullet - \boldsymbol{y}_\bullet\|_{\ell^\infty_T} \leqslant S(T)\Delta \|\boldsymbol{\delta}_\bullet\|_{\ell^1_T}$$

One could argue that this equation is neater than (7.19); it bounds a norm of just one math-

ematical object, namely the sequence $(\boldsymbol{x}_n - \boldsymbol{y}_n)_{1\leqslant n\leqslant N}$, by a norm of another object, namely the sequence $(\boldsymbol{\delta}_n)_{1\leqslant n\leqslant N}$. Arguments for proving that a numerical scheme is stable often rely on some form of Lipschitz continuity. If the function $\boldsymbol{\Phi}_\Delta(t, \boldsymbol{y})$ is globally Lipschitz continuous with respect to $\boldsymbol{y}$, then stability is particularly simple to prove, as we now demonstrate.

Proposition 7.7. *Assume that there is $L_\Phi > 0$ such that for all $t \in [0, T]$ and $\Delta > 0$, the function $\boldsymbol{\Phi}_\Delta(t, \bullet)$ is globally Lipschitz continuous with constant L_Φ. Then the one-step method (7.16) is stable.*

Proof. By (7.16) and (7.18), it holds that

$$\boldsymbol{x}_n - \boldsymbol{y}_n = \boldsymbol{x}_{n-1} - \boldsymbol{y}_{n-1} + \Delta\Big(\boldsymbol{\Phi}_\Delta(t_{n-1}, \boldsymbol{x}_{n-1}) - \boldsymbol{\Phi}_\Delta(t_{n-1}, \boldsymbol{y}_{n-1})\Big) - \Delta\boldsymbol{\delta}_n.$$

Taking the Euclidean norm and employing the Lipschitz continuity assumption, we obtain

$$\|\boldsymbol{x}_n - \boldsymbol{y}_n\| \leqslant (1 + \Delta L_\Phi)\|\boldsymbol{x}_{n-1} - \boldsymbol{y}_{n-1}\| + \Delta\|\boldsymbol{\delta}_n\|.$$

By a reasoning similar to that in the proof of Theorem 7.5, we then obtain

$$\|\boldsymbol{x}_n - \boldsymbol{y}_n\| \leqslant (1 + \Delta L_\Phi)^n\|\boldsymbol{x}_0 - \boldsymbol{y}_0\| + \sum_{i=1}^{n}(1 + \Delta L_\Phi)^{n-i}\Delta\|\boldsymbol{\delta}_i\| \leqslant 0 + \mathrm{e}^{L_\Phi t_n}\Delta\sum_{i=1}^{n}\|\boldsymbol{\delta}_i\|.$$

We conclude that (7.19) is satisfied with $S(T) = \mathrm{e}^{L_\Phi T}$. □

Convergence

We are now ready to prove that consistency and stability of the numerical (7.16) together imply convergence, in the sense that

$$\lim_{\Delta\to 0}\left(\max_{1\leqslant n\leqslant N}\|\boldsymbol{x}(t_n) - \boldsymbol{x}_n\|\right) = 0, \qquad N = \left\lfloor\frac{T}{\Delta}\right\rfloor.$$

This result is an instance of the *Lax equivalence theorem*, a pillar of numerical analysis with far-reaching applications.

Theorem 7.8 (Consistence and stability imply convergence)**.** *Assume that the one-step numerical method (7.16) is consistent and stable. Then the method is also convergent.*

Proof. By definition of the local truncation error, it holds that

$$\boldsymbol{x}(t_{n+1}) = \boldsymbol{x}(t_n) + \Delta\boldsymbol{\Phi}_\Delta\big(t_n, \boldsymbol{x}(t_n)\big) + \Delta\boldsymbol{\eta}_{n+1}.$$

Therefore, the sequence $\big(\boldsymbol{x}(t_n)\big)_{1\leqslant n\leqslant N}$ satisfies (7.18) with $\boldsymbol{\delta}_n = \boldsymbol{\eta}_n$, and so the stability estimate (7.19) implies that

$$\max_{1\leqslant n\leqslant N}\|\boldsymbol{x}(t_n) - \boldsymbol{x}_n\| \leqslant S(T)\Delta\sum_{n=1}^{N}\|\boldsymbol{\eta}_n\|.$$

By consistency, the right-hand side converges to zero in the limit as $\Delta \to 0$, which concludes the proof. □

Remark 7.2. If we assume in Theorem 7.8 that the method is consistent with order p, then by adapting the proof, we find that the error satisfies

$$\max_{1 \leqslant n \leqslant N} \|\boldsymbol{x}(t_n) - \boldsymbol{x}_n\| \leqslant CS(T)\Delta^p.$$

In this setting, the numerical scheme is said to be *convergent with order p*.

7.2.4 Widely used one-step methods

In this section, we motivate and describe some of the other widely-used one-step methods, namely methods of Taylor and Runge–Kutta type. We assume in this section that the equation (7.1) admits a unique smooth solution over the interval $[0, T]$.

Taylor methods

In order to construct a method with a smaller local truncation error than that of the forward Euler method, a Taylor expansion of higher order than (7.9) can be employed:

$$\boldsymbol{x}(t + \Delta) = \boldsymbol{x}(t) + \Delta \boldsymbol{x}'(t) + \cdots + \frac{\Delta^p}{p!} \boldsymbol{x}^{(p)}(t) + \mathcal{O}(\Delta^{p+1}). \tag{7.20}$$

Since $\boldsymbol{x} \colon [0, T] \to \mathbf{R}$ is a smooth solution to (7.1) by assumption, the time derivatives of $\boldsymbol{x}$ can be obtained by differentiation of (7.1):

$$\boldsymbol{x}'(t) = \boldsymbol{f}\big(t, \boldsymbol{x}(t)\big), \qquad \boldsymbol{x}''(t) = \partial_t \boldsymbol{f}\big(t, \boldsymbol{x}(t)\big) + \Big(\boldsymbol{f}\,(t, \boldsymbol{x}(t)) \cdot \nabla_{\boldsymbol{x}}\Big) \boldsymbol{f}\big(t, \boldsymbol{x}(t)\big), \qquad \dots$$

In general, it is immediate to show inductively that $\boldsymbol{x}^{(p)}(t) = \boldsymbol{f}^{(p-1)}\big(t, \boldsymbol{x}(t)\big)$, where the functions $\boldsymbol{f}^{(p)} \colon \mathbf{R} \times \mathbf{R}^n \to \mathbf{R}$ are defined recursively from the following equation:

$$\boldsymbol{f}^{(p+1)} = \partial_t \boldsymbol{f}^{(p)}\big(t, \boldsymbol{x}(t)\big) + \Big(\boldsymbol{f}\,(t, \boldsymbol{x}(t)) \cdot \nabla_{\boldsymbol{x}}\Big) \boldsymbol{f}^{(p)}\big(t, \boldsymbol{x}(t)\big).$$

The Taylor expansion (7.20) motivates the so-called Taylor methods for integrating (7.1) numerically, which are based on the following iteration:

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n + \Delta \boldsymbol{T}^p_\Delta\big(t_n, \boldsymbol{x}_n\big), \tag{7.21}$$

where

$$\boldsymbol{T}^p_\Delta(t, \boldsymbol{x}) := \boldsymbol{f}(t, \boldsymbol{x}) + \frac{\Delta}{2!} \boldsymbol{f}^{(1)}(t, \boldsymbol{x}) + \cdots + \frac{\Delta^{p-1}}{p!} \boldsymbol{f}^{(p-1)}(t, \boldsymbol{x}).$$

Note that, for any p, the Taylor scheme (7.21) may be rewritten as

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n + \Delta \frac{\mathrm{d}\boldsymbol{x}^{t_n, \boldsymbol{x}_n}}{\mathrm{d}t}(t_n) + \cdots + \frac{\Delta^p}{p!} \frac{\mathrm{d}^p \boldsymbol{x}^{t_n, \boldsymbol{x}_n}}{\mathrm{d}t^p}(t_n).$$

For $p = 1$, this scheme coincides with the forward Euler scheme.

Runge–Kutta methods

Runge–Kutta methods resemble Taylor methods, but they do not require to calculate the derivatives of the function $\boldsymbol{f}$. This is achieved by approximating the derivatives in Taylor methods by difference quotients. Consider for example the Taylor method of order $p = 2$:

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n + \Delta \frac{\mathrm{d}\boldsymbol{x}^{t_n,\boldsymbol{x}_n}}{\mathrm{d}t}(t_n) + \frac{\Delta^2}{2!}\frac{\mathrm{d}^2\boldsymbol{x}^{t_n,\boldsymbol{x}_n}}{\mathrm{d}t^2}(t_n). \tag{7.22}$$

Substituting the approximation

$$\begin{aligned}\frac{\mathrm{d}^2\boldsymbol{x}^{t_n,\boldsymbol{x}_n}}{\mathrm{d}t^2}(t_n) &\approx \frac{1}{\Delta}\left(\frac{\mathrm{d}\boldsymbol{x}^{t_n,\boldsymbol{x}_n}}{\mathrm{d}t}(t_n+\Delta) - \frac{\mathrm{d}\boldsymbol{x}^{t_n,\boldsymbol{x}_n}}{\mathrm{d}t}(t_n)\right)\\ &= \frac{1}{\Delta}\Big(\boldsymbol{f}\big(t_n+\Delta, \boldsymbol{x}^{t_n,\boldsymbol{x}_n}(t_n+\Delta)\big) - \boldsymbol{f}\big(t_n,\boldsymbol{x}_n\big)\Big)\\ &\approx \frac{1}{\Delta}\Big(\boldsymbol{f}\big(t_n+\Delta, \boldsymbol{x}_n+\Delta\boldsymbol{f}(t_n,\boldsymbol{x}_n)\big) - \boldsymbol{f}\big(t_n,\boldsymbol{x}_n\big)\Big)\end{aligned} \tag{7.23}$$

in (7.22), we obtain an explicit method known as *Heun's method*:

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n + \frac{\Delta}{2}\boldsymbol{f}(t_n,\boldsymbol{x}_n) + \frac{\Delta}{2}\boldsymbol{f}\big(t_n+\Delta, \boldsymbol{x}_n+\Delta\boldsymbol{f}(t_n,\boldsymbol{x}_n)\big).$$

It is possible to show that the local truncation error for this method also scales as Δ^2. Heun's method is a particular instance of a Runge–Kuta method. In general, an explicit Runge–Kutta method with s stages is of the form

$$\begin{aligned}\boldsymbol{x}_{n+1} &= \boldsymbol{x}_n + \Delta\sum_{i=1}^{s} b_i\boldsymbol{k}_i\\ \boldsymbol{k}_1 &= \boldsymbol{f}(t_n,\boldsymbol{x}_n),\\ \boldsymbol{k}_2 &= \boldsymbol{f}\big(t_n + c_2\Delta, \boldsymbol{x}_n + \Delta(a_{21}\boldsymbol{k}_1)\big),\\ \boldsymbol{k}_3 &= \boldsymbol{f}\big(t_n + c_3\Delta, \boldsymbol{x}_n + \Delta(a_{31}\boldsymbol{k}_1 + a_{32}\boldsymbol{k}_2)\big),\\ &\vdots\\ \boldsymbol{k}_s &= \boldsymbol{f}\left(t_n + c_s\Delta, \boldsymbol{x}_n + \Delta\sum_{j=1}^{s-1} a_{sj}\boldsymbol{k}_j\right),\end{aligned}$$

with appropriate coefficients c_i and a_{ij}. Heun's iteration can be recast in this form as follows:

$$\begin{aligned}\boldsymbol{x}_{n+1} &= \boldsymbol{x}_n + \frac{\Delta}{2}(\boldsymbol{k}_1+\boldsymbol{k}_2)\\ \boldsymbol{k}_1 &= \boldsymbol{f}(t_n,\boldsymbol{x}_n)\\ \boldsymbol{k}_2 &= \boldsymbol{f}\big(t_n+\Delta, \boldsymbol{x}_n+\Delta\boldsymbol{k}_1\big).\end{aligned}$$

The approach we employed to construct Heun's method may be generalized to higher orders. For example, the most widely known Runge–Kutta method approximates the Taylor method of

order $p = 4$ with the following iteration:

$$\begin{aligned}
\boldsymbol{x}_{n+1} &= \boldsymbol{x}_n + \frac{\Delta}{6}(\boldsymbol{k}_1 + 2\boldsymbol{k}_2 + 2\boldsymbol{k}_3 + \boldsymbol{k}_4), \\
\boldsymbol{k}_1 &= \boldsymbol{f}(t_n, \boldsymbol{x}_n), & \boldsymbol{k}_2 &= \boldsymbol{f}\left(t_n + \frac{\Delta}{2}, \boldsymbol{x}_n + \Delta\frac{\boldsymbol{k}_1}{2}\right), \\
\boldsymbol{k}_3 &= \boldsymbol{f}\left(t_n + \frac{\Delta}{2}, \boldsymbol{x}_n + \Delta\frac{\boldsymbol{k}_2}{2}\right), & \boldsymbol{k}_4 &= \boldsymbol{f}\left(t_n + \Delta, \boldsymbol{x}_n + \Delta\boldsymbol{k}_3\right).
\end{aligned}$$

The local truncation error for this method scales as Δ^4 and, when $\boldsymbol{f}(t, \boldsymbol{x}) = \boldsymbol{f}(t)$, this method coincides with Simpson's formula (3.6) for the approximation of the integral in (7.17). The systematic derivation of Runge–Kutta methods is cumbersome, and so we do not address this issue in this course.

Remark 7.3. Explicit Runge–Kutta methods of a given order are not uniquely defined. For example, if we employ instead of (7.23) the approximation

$$\frac{\mathrm{d}^2\boldsymbol{x}^{t_n,\boldsymbol{x}_n}}{\mathrm{d}t^2}(t_n) \approx \frac{2}{\Delta}\left(\boldsymbol{f}\left(t_n + \frac{\Delta}{2}, \boldsymbol{x}_n + \frac{\Delta}{2}\boldsymbol{f}(t_n, \boldsymbol{x}_n)\right) - \boldsymbol{f}\big(t_n, \boldsymbol{x}_n\big)\right),$$

then we obtain by substitution in (7.22) the so-called *explicit midpoint method*, which is also a Runge–Kutta method of order 2:

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n + \Delta\boldsymbol{f}\left(t_n + \frac{1}{2}\Delta, \boldsymbol{x}_n + \frac{\Delta}{2}\boldsymbol{f}(t_n, \boldsymbol{x}_n)\right).$$

Implicit methods

To conclude this section, we mention two common implicit methods with a better order of convergence than that of the backward Euler method.

- The Crank–Nicolson method:

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n + \frac{\Delta}{2}\big(f(t_n, \boldsymbol{x}_n) + f(t_n + \Delta, \boldsymbol{x}_{n+1})\big). \tag{7.24}$$

  When $\boldsymbol{f}$ is independent of $\boldsymbol{x}$ and depends only on t, this method coincides with the trapezoidal rule for numerical integration.

- The implicit midpoint method:

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n + \Delta f\left(t_n + \frac{\Delta}{2}, \frac{\boldsymbol{x}_n + \boldsymbol{x}_{n+1}}{2}\right).$$

Similarly to the backward Euler method, each iteration of these methods requires the resolution of a nonlinear equation. Implicit methods often enjoy better stability than their explicit counterparts. This subject is further discussed in Section 7.4.

7.3 Multistep methods

The idea of multistep methods is to use, in the construction of a new iterate, information from not only the current but also previous iterations. This degree of freedom enables to construct more economical numerical methods than one-step methods for the same order of convergence, at the cost of a more difficult initialization. In this section we focus on *linear* multistep methods of the form

$$\begin{aligned}\boldsymbol{x}_{n+1} = a_0\boldsymbol{x}_n + a_1\boldsymbol{x}_{n-1} + \cdots + a_k\boldsymbol{x}_{n-k} \\ + \Delta\Big(b_{-1}\boldsymbol{f}(t_{n+1},\boldsymbol{x}_{n+1}) + b_0\boldsymbol{f}(t_n,\boldsymbol{x}_n) + \cdots + b_k\boldsymbol{f}(t_{n-k},\boldsymbol{x}_{n-k})\Big).\end{aligned} \tag{7.25}$$

This equation defines an explicit method if $b_{-1} = 0$, and an implicit method if $b_{-1} \neq 0$. Note that explicit methods of the form (7.25) require only one additional evaluation $f(t_n, \boldsymbol{x}_n)$ per iteration, in contrast with Runge–Kutta methods. When $b_{-1} \neq 0$, the iteration (7.25) is a nonlinear equation for the unknown $\boldsymbol{x}_{n+1}$, which must itself be solved by resorting to a numerical method.

Initialization. In order to initiate the numerical method (7.25), the values $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_k$ are required. These can be calculated by using a one-step method with an order of convergence matching that of the multistep method.

Local truncation error. Consistently with the setting of one-step methods, the local truncation error for (7.25) is defined as the residual error left when the exact solution is substituted in the numerical scheme:

$$\begin{aligned}\Delta\boldsymbol{\eta}_{n+1} := \boldsymbol{x}(t_n+\Delta) - a_0\boldsymbol{x}(t_n) - a_1\boldsymbol{x}(t_n-\Delta) - \cdots - a_k\boldsymbol{x}(t_n-k\Delta) \\ - \Delta\Big(b_{-1}\boldsymbol{x}'(t_n+\Delta) + b_0\boldsymbol{x}'(t_n) + \cdots + b_k\boldsymbol{x}'(t_n-k\Delta)\Big).\end{aligned} \tag{7.26}$$

The multistep method (7.25) is said to be of order p if the maximum local truncation error over all the discretization points, in norm, scales as $\mathcal{O}(\Delta^p)$. The following result is useful for estimating the order of consistency of a linear multistep method.

Proposition 7.9. *The linear multistep method (7.25) is consistent with order p for any smooth $\boldsymbol{x}\colon [0,T] \to \mathbf{R}^n$ if and only if the local truncation error (7.26) is everywhere zero when $\boldsymbol{x}(t)$ is of the scalar form*

$$x(t) = t^q, \qquad q \in \{0, \ldots, p\}. \tag{7.27}$$

Proof. Assume that the method is consistent with order p, fix $q \in \{1, \ldots, p\}$, and let $x(t) = t^q$. Fix also $t \in [0,T]$ and consider the function $\xi\colon \{\Delta : t/\Delta \in \mathbf{N}_{>0}\} \to \mathbf{R}$ given by

$$\begin{aligned}\Delta\xi(\Delta) = \Delta\eta_{(t/\Delta)+1} = x(t+\Delta) - a_1x(t) - a_2x(t-\Delta) - \cdots - a_kx\big(t-(k-1)\Delta\big) \\ - \Delta\Big(b_0x'(t+\Delta) + b_1x'(t) + \cdots + b_kx'\big(t-(k-1)\Delta\big)\Big).\end{aligned}$$

The quantity $\xi(\Delta)$ should be understood as the local truncation error at t for time step Δ. It is a polynomial in Δ of degree at most p and scaling as $\mathcal{O}(\Delta^{p+1})$. Therefore, it holds necessarily that $\xi(\Delta) = 0$.

Conversely, assume that the right-hand side of (7.26) is equal to zero for any function of the form (7.27). If $\boldsymbol{x}(t)$ denotes a smooth solution of (7.1), then by Taylor's theorem there is $C > 0$ independent of t_n such that

$$\forall t \in [0, T], \qquad \begin{cases} \|\boldsymbol{x}(t) - \boldsymbol{y}(t)\| \leqslant C|t - t_n|^{p+1} \\ \|\boldsymbol{x}'(t) - \boldsymbol{y}'(t)\| \leqslant C|t - t_n|^{p} \end{cases}, \qquad \boldsymbol{y}(t) := \boldsymbol{x}(t_n) + \sum_{i=1}^{p} \boldsymbol{e}_i (t - t_n)^i,$$

for appropriate vectors $\boldsymbol{e}_i \in \mathbf{R}^n$ depending on t_n. Substituting $\boldsymbol{x}(t) = \boldsymbol{y}(t) + \big(\boldsymbol{x}(t) - \boldsymbol{y}(t)\big)$ in the right-hand side of (7.26), we obtain

$$\Delta \|\boldsymbol{\eta}_{n+1}\| = \mathcal{O}(\Delta^{p+1}) + \Delta \mathcal{O}(\Delta^p) = \mathcal{O}(\Delta^{p+1}),$$

with the constant implicit in the big $\mathcal{O}$ notation independent of n. This concludes the proof. □

Example 7.2. In the one-dimensional setting, we wish to find parameters a_0, a_1 and b_1 such that the order of consistency of the following multistep scheme is as high as possible:

$$x_{n+1} = a_0 x_n + a_2 x_{n-1} + b_0 \Delta f(t_n, x_n).$$

Substituting $x(t) = 1$ in the formula (7.26) for the local truncation error, we obtain

$$\eta_{n+1} = x(t_n + \Delta) - a_0 x(t_n) - a_1 x(t_n - \Delta) - b_0 \Delta x'(t_n) = 1 - a_0 - a_1.$$

Therefore $a_1 = (1 - a_0)$. Next, substituting $x(t) = t - t_n$ in (7.26), we obtain

$$\eta_{n+1} = \Delta(2 - a_0 - b_0),$$

which gives $b_0 = 2 - a_0$. Finally, substituting $x(t) = (t - t_n)^2$, we obtain

$$\eta_{n+1} = \Delta^2 a_0,$$

and so $a_0 = 1$. We conclude that the best parameters, leading to a local truncation error scaling as $\mathcal{O}(\Delta^2)$, are given by $a_0 = 0$, $a_1 = 1$ and $b_0 = 2$. The resulting method reads

$$x_{n+1} = x_{n-1} + 2\Delta f(t_n, x_n),$$

and is known as the *multistep midpoint method*.

We now present two widely used systematic approaches for constructing multistep methods, known as the Adams–Bashforth and Adams–Moulton approaches.

7.3.1 Adams–Bashforth methods

Let $\boldsymbol{x}\colon [0,T] \to \mathbf{R}^n$ denote the exact solution to the differential equation (7.1). Integrating this equation between t_n and t_{n+1}, we obtain

$$\boldsymbol{x}(t_{n+1}) = \boldsymbol{x}(t_n) + \int_{t_n}^{t_{n+1}} \boldsymbol{f}\big(t, \boldsymbol{x}(t)\big)\,\mathrm{d}t. \tag{7.28}$$

The key idea of the Adams–Bashforth method is to approximate the function $t \mapsto \boldsymbol{f}\big(t, \boldsymbol{x}(t)\big)$ by the interpolating polynomial $\widehat{\boldsymbol{f}}$ of degree k at the nodes $t_{n-k}, \dots, t_n$:

$$\widehat{\boldsymbol{f}}(t) = \sum_{i=0}^{k} \boldsymbol{f}\big(t_{n-i}, \boldsymbol{x}(t_{n-i})\big) L_i(t), \qquad L_i(t) := \prod_{\substack{j=0 \\ j\neq i}}^{k} \frac{t - t_{n-j}}{t_{n-i} - t_{n-j}}. \tag{7.29}$$

Substituting this approximation in (7.28), we obtain

$$\boldsymbol{x}(t_{n+1}) \approx \boldsymbol{x}(t_n) + \sum_{i=0}^{k} \boldsymbol{f}\big(t_{n-i}, \boldsymbol{x}(t_{n-i})\big) \int_{t_n}^{t_{n+1}} L_i(t)\,\mathrm{d}t.$$

This motivates the following *Adams–Bashforth* numerical scheme:

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n + \Delta \sum_{i=0}^{k} b_i \boldsymbol{f}\big(t_{n-i}, \boldsymbol{x}_{n-i}\big), \qquad b_i := \int_0^1 \prod_{\substack{j=0 \\ j\neq i}}^{k} \frac{s+j}{-i+j}\,\mathrm{d}s. \tag{7.30}$$

Since the Lagrange polynomials $(L_i)_{0\leqslant i\leqslant k}$ depend on k, so do the coefficients b_i. However, these are independent of Δ, and so they can be tabulated. The value of these coefficients for the first few Adams–Bashforth methods are collated in Table 7.1.

i	0	1	2	3
$k=0$	1			
$k=1$	$\frac{3}{2}$	$-\frac{1}{2}$		
$k=2$	$\frac{23}{12}$	$-\frac{16}{12}$	$\frac{5}{12}$	
$k=3$	$\frac{55}{24}$	$-\frac{59}{24}$	$\frac{37}{24}$	$-\frac{9}{24}$

Table 7.1: Coefficients $(b_i)_{0\leqslant i\leqslant k}$ of the Adams–Bashforth methods.

Local truncation error. Assuming $\boldsymbol{x} \in C^{k+2}\big([0,T], \mathbf{R}^n\big)$ and applying Theorem 2.3 for the interpolation error component-wise, we obtain

$$\forall t \in [0,T], \qquad \Big\|\boldsymbol{x}'(t) - \widehat{\boldsymbol{f}}(t)\Big\|_\infty \leqslant \frac{|t-t_{n-k}|\cdots|t-t_n|}{(k+1)!} \sup_{t\in[0,T]} \Big\|\boldsymbol{x}^{(k+2)}(t)\Big\|_\infty,$$

where $\widehat{\boldsymbol{f}}$ is the function defined in (7.29). Since

$$\Delta\boldsymbol{\eta}_{n+1} = \boldsymbol{x}(t_{n+1}) - \boldsymbol{x}(t_n) - \Delta\sum_{i=0}^{k} b_i \boldsymbol{f}\big(t_{n-i}, \boldsymbol{x}(t_{n-i})\big) = \int_{t_n}^{t_{n+1}} \Big(\boldsymbol{x}'(t) - \widehat{\boldsymbol{f}}(t)\Big)\,\mathrm{d}t,$$

we deduce that

$$\|\boldsymbol{\eta}_{n+1}\| \leqslant C_k M_{k+2} \Delta^{k+1}, \qquad M_{k+2} := \sup_{t\in[0,T]} \Big\|\boldsymbol{x}^{(k+2)}(t)\Big\|_\infty, \tag{7.31}$$

for an appropriate numerical constant C_k independent of n and of the problem data. Therefore the Adams–Bashforth method (7.30) is consistent with order $k+1$.

Convergence. By using a reasoning similar to that in the proof of Theorem 7.5, we can prove a convergence result of the Adams–Bashforth method.

Theorem 7.10. *Assume that the solution $\boldsymbol{x}\colon [0,T] \to \mathbf{R}^n$ to (7.1) is $k+2$ times continuously differentiable and that the global Lipschitz condition (7.10) is satisfied. Suppose also that*

$$\forall i \in \{0,\dots,k\}, \qquad \|\boldsymbol{x}(t_i) - \boldsymbol{x}_i\| \leqslant \delta.$$

Then the following error estimate holds for the Adams–Bashforth method (7.30):

$$\forall n \in \left\{0,1,\dots,\left\lfloor \frac{T}{\Delta}\right\rfloor\right\}, \qquad \|\boldsymbol{x}(t_n) - \boldsymbol{x}_n\| \leqslant \delta \mathrm{e}^{LB} + C_k M_{k+2}\Delta^{k+1}\left(\frac{\mathrm{e}^{LBt_n}-1}{LB}\right),$$

where C_k and M_{k+2} are the constants from (7.31), and with $B := |b_0| + \cdots + |b_k|$.

Sketch of proof. Let $\boldsymbol{e}_n := \boldsymbol{x}(t_{n+1}) - \boldsymbol{x}_{n+1}$. From the equation

$$\boldsymbol{x}(t_{n+1}) - \boldsymbol{x}_{n+1} = \boldsymbol{x}(t_n) - \boldsymbol{x}_n + \Delta\sum_{i=0}^{k} b_i\Big(\boldsymbol{f}\big(t_{n-i}, \boldsymbol{x}(t_{n-i})\big) - \boldsymbol{f}\big(t_{n-i}, \boldsymbol{x}_{n-i}\big)\Big) + \Delta\boldsymbol{\eta}_{n+1},$$

which is valid for $n \geqslant k$, we deduce that

$$\max\Big\{\|\boldsymbol{e}_0\|,\dots,\|\boldsymbol{e}_{n+1}\|\Big\} \leqslant (1+\Delta LB)\max\Big\{\|\boldsymbol{e}_0\|,\dots,\|\boldsymbol{e}_n\|\Big\} + C_k M_{k+2}\Delta^{k+2}.$$

Since $\max\Big\{\|\boldsymbol{e}_0\|,\dots,\|\boldsymbol{e}_k\|\Big\} \leqslant \delta$ by assumption, the statement easily follows. □

7.3.2 Adams–Moulton methods

The Adams–Moulton methods are very similar to their Adams–Bashforth cousins. The only difference is that the former are obtained by interpolating the function $t \mapsto \boldsymbol{f}\big(t, \boldsymbol{x}(t)\big)$ in (7.28) at nodes shifted forward by Δ, i.e. at the nodes $t_{n-k+1},\dots,t_{n+1}$. This leads to the method

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n + \Delta\sum_{i=-1}^{k-1} b_i \boldsymbol{f}\big(t_{n-i}, \boldsymbol{x}_{n-i}\big), \qquad b_i := \int_0^1 \prod_{\substack{j=-1\\ j\neq i}}^{k-1} \frac{s+j}{-i+j}\,\mathrm{d}s. \tag{7.32}$$

Unlike the Adams–Bashforth methods, which are *explicit*, the Adams–Moulton methods are *implicit*. The value of the coefficients for the first few Adams–Moulton methods are collated in Table 7.2. Notice that, for $k = 0$, the Adams–Moulton method coincides with the backward Euler method, and for $k = 1$ it coincides with the Crank–Nicolson method.

i	-1	0	1	2
$k = 0$	1			
$k = 1$	$\frac{1}{2}$	$\frac{1}{2}$		
$k = 2$	$\frac{5}{12}$	$\frac{8}{12}$	$-\frac{1}{12}$	
$k = 3$	$\frac{9}{24}$	$\frac{19}{24}$	$-\frac{5}{24}$	$\frac{1}{24}$

Table 7.2: Coefficients $(b_i)_{0 \leqslant i \leqslant k}$ of the Adams–Moulton methods.

7.4 Absolute stability

To conclude this chapter, we introduce the notion of *absolute stability* and explain its relevance. Absolute stability is a property of a numerical method in relation to the model equation

$$\begin{cases} x'(t) = \lambda x(t), \\ x(0) = 1. \end{cases} \tag{7.33}$$

A numerical scheme for approximating (7.33) is called *absolutely stable* if

$$|x_n| \to 0 \qquad \text{in the limit as } n \to \infty. \tag{7.34}$$

where $(x_n)_{n=0,1,\ldots}$ denotes the numerical solution to (7.33). Whether a numerical method is absolutely stable or not depends on the parameters λ and Δ.

Example 7.3. The forward Euler method for (7.33) reads

$$x_{n+1} = x_n + \Delta\lambda x_n = (1 + \Delta\lambda)x_n.$$

Therefore $x_n \to 0$ if and only if $|1 + \Delta\lambda| \leqslant 1$.

As Example 7.3 illustrates, whether absolute stability holds for the forward Euler methods depends only the value of the product $\Delta\lambda \in \mathbf{C}$. This dependence on λ and Δ only through the product $\Delta\lambda$ holds in fact generally. Indeed, all the numerical schemes we considered in this chapter are invariant under linear time rescaling of the ordinary differential equation: the numerical solution of the rescaled equation, when the time step is rescaled by the same factor, coincides with the discrete function obtained by linear rescaling of the numerical solution to the original equation. This motivates the definition of *absolute stability region* as

$$\mathcal{A} := \{z \in \mathbf{C} : (7.34) \text{ holds when } \Delta\lambda = z\} \subset \mathbf{C}.$$

The exact solution to the model equation (7.33) diverges to ∞ as $t \to \infty$ if $\Re(\lambda) > 0$, and it converges to 0 if $\Re(\lambda) < 0$. Numerical schemes which exhibit a similar property at the discrete level are called *A-stable*. More precisely, a numeric method is A-stable if the absolute stability region $\mathcal{A}$ contains the left half-plane, i.e. if

$$\{z \in \mathbf{C} : \Re(z) < 0\} \subset \mathcal{A}.$$

Before investigating whether the numerical schemes introduced previously in this chapter are absolutely stable, we address the following natural question: why focus on the simple model equation (7.33)? We provide a couple of motivations:

- First, note that equations of the form (7.33) are more relevant in science than might appear at first glance. Indeed, when discretizing in space a linear parabolic partial differential equation, one often obtains a linear differential equation of the form

  $$\boldsymbol{x}'(t) = \mathsf{A}\boldsymbol{x},$$

  where $\mathsf{A} \in \mathbf{C}^{n \times n}$. If the matrix $\mathsf{A} = \mathsf{Q}\mathsf{D}\mathsf{Q}^*$ is diagonalizable, then the vector $\boldsymbol{z}(t) := \mathsf{Q}^*\boldsymbol{x}(t)$ satisfies the differential equation

  $$\boldsymbol{z}'(t) = \mathsf{D}\boldsymbol{z}.$$

  In other words, each component of $\boldsymbol{z}$ satisfies an ordinary differential equation of the same form as the model equation (7.33). In applications, the components of $\boldsymbol{z}$ often encode the amplitudes of Fourier modes of the solution to the partial differential equation, and for dissipative equations all the eigenvalues of A have a negative real part. However, the spectral radius of A usually diverges as the number of discretization points increases. In this context, A-stability is particularly attractive, as it ensures that the numerical approximation remains well-behaved as the number of discretization points increases.

- Second, the model equation (7.33) may be viewed as a linearized approximation of a more interesting equation. Consider, for example, the following one-dimensional autonomous differential equation:

  $$\begin{cases} x'(t) = f\big(x(t)\big), \\ x(0) = x_0. \end{cases} \tag{7.35}$$

  Assume that $f(x_*) = 0$ for some $x_* \in \mathbf{R}$. Such a point is called a *critical point* of the differential equation. If $f'(x_*) < 0$, then x_* is an attractor of the equation, in the sense that $x(t) \to x_*$ provided that $x(0)$ is sufficiently close to x_*. This result, which is the counterpart of Proposition 5.5 for differential equations, is a particular case of a theorem due to Poincaré and Lyapunov; see [16, Theorem 7.1]. If $|x_0 - x_*|$ is sufficiently small, then the solution to (7.35) is expected to be close to that of the linearized equation

  $$\begin{cases} y'(t) = f'\big(x_*\big)\big(y(t) - x_*\big), \\ y(0) = x_0, \end{cases} \tag{7.36}$$

  which is of the form (7.33). Often, studying the linearized equation (7.36) enables to gain

insight into the behavior of the original equation (7.35), and analyzing the performance of a numerical method for the linearized equation (7.36) is useful to inform the choice of a numerical scheme for (7.35).

- More generally, if $x(t)$ and $x_\varepsilon(t)$ are respectively the solutions to

$$\begin{cases} x'(t) = f\big(t, x(t)\big), \\ x(0) = x_0 + \varepsilon, \end{cases} \quad \text{and} \quad \begin{cases} x'_\varepsilon(t) = f\big(t, x_\varepsilon(t)\big), \\ x_\varepsilon(0) = x_0 + \varepsilon, \end{cases} \tag{7.37}$$

then the difference $e(t) := x_\varepsilon(t) - x(t)$ satisfies the equation

$$\begin{aligned} e'(t) &= f\big(t, x_\varepsilon(t)\big) - f\big(t, x(t)\big) \approx \partial_x f\big(t, x(t)\big) e(t), \\ e(0) &= \varepsilon, \end{aligned} \tag{7.38}$$

which looks similar to (7.33) with $\partial_x f\big(t, x(t)\big)$ in place of λ. At a given time t, the solutions tend to converge to each other as time increases if $\partial_x f\big(t, x(t)\big) < 0$, and diverge from each other if $\partial_x f\big(t, x(t)\big) > 0$. Testing absolute stability with $\lambda = \partial_x f\big(t, x(t)\big)$ enables to determine whether this property holds true also at the discrete level. Although the latter statement is difficult to state precisely and prove generally, we illustrate its validity for the forward Euler method in Example 7.4.

Example 7.4. Let (x_n) and (x_n^ε) denote the numerical solutions obtained by applying the forward Euler method to the differential equations in (7.37). If $\varepsilon \ll 1$, then

$$\begin{aligned} x_{n+1}^\varepsilon - x_{n+1} &= x_n^\varepsilon - x_n + \Delta f(t_n, x_n^\varepsilon) - \Delta f(t_n, x_n) \\ &\approx x_n^\varepsilon - x_n + \Delta \partial_x f(t_n, x_n)(x_n^\varepsilon - x_n) = \big(1 + \Delta \partial_x f(t_n, x_n)\big)(x_n^\varepsilon - x_n). \end{aligned}$$

Therefore, the numerical solutions (x_n^ε) and (x_n) tend to become closer as n increases if

$$\Delta \partial_x f(t_n, x_n) \in \mathcal{A}. \tag{7.39}$$

The absolute stability regions of the forward and backward Euler methods are illustrated in green in Figure 7.1. For the forward Euler method, absolute stability holds if and only if $|1 + \Delta\lambda| < 1$, as we proved in Example 7.3. A similar reasoning gives that the absolute stability region for backward Euler method is given by $\{z \in \mathbf{C} : |1 - z|^{-1} < 1\}$. The backward Euler method is A-stable but the forward Euler method is not. Notice that, if the time step is sufficiently large, then the backward Euler method is absolutely stable even for values of λ with a positive real part, for which exact solutions to the model equation are divergent.

Example 7.5 (Absolute stability region of the Taylor methods). When applied to (7.33), the Taylor method of order p given in (7.21) reads

$$x_{n+1} = \left(1 + \Delta\lambda + \frac{\Delta^2\lambda^2}{2} + \cdots + \frac{\Delta^p\lambda^p}{p!}\right) x_n.$$

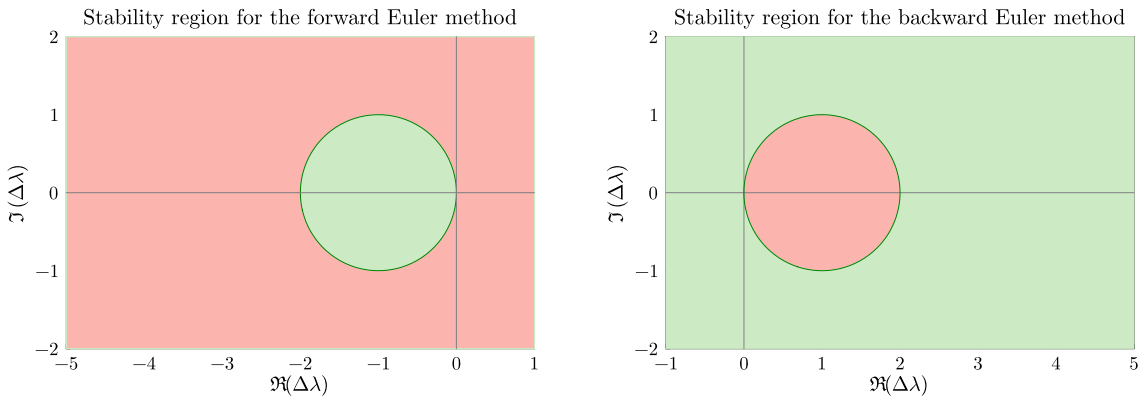

Figure 7.1: Absolute stability regions for the forward (left) and backward (right) Euler methods.

Thus, the absolute stability region is given by

$$\left\{ z \in \mathbf{C} : \left| 1 + z + \frac{z^2}{2} + \cdots + \frac{z^p}{p!} \right| < 1 \right\}.$$

This region is illustrated for various values of p in Figure 7.2. We observe that the absolute stability region grows as p increases.

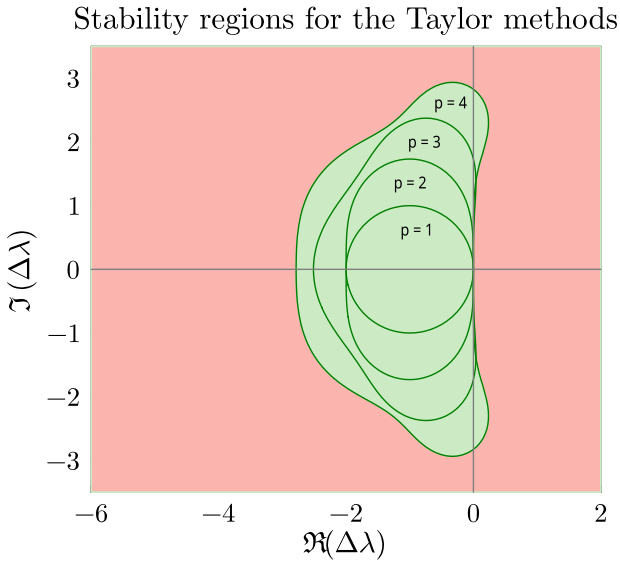

Figure 7.2: Stability regions for the first few Taylor methods.

Stiff differential equations

In the context of ordinary differential equations, stiffness is not a precisely defined concept, but rather rather a generic term employed to describe equations with widely separated time scales. Roughly speaking, a differential equation of the form (7.1) is called stiff if the Jacobian matrix of f, with respect to the variable $\boldsymbol{x}$, has at least one eigenvalue with a large negative real part. In the one-dimensional setting, the solutions to stiff differential equations which are close at the initial time tend to converge quickly to each other, in view of (7.38). This is illustrated in Example 7.6.

Example 7.6 (Stiff differential equation). Consider the following equation [7, Chapter 4]:

$$\begin{cases} x'(t) = -\alpha\big(x(t) - \sin(t)\big) + \cos(t) \\ x(0) = x_0 \end{cases} \tag{7.40}$$

The exact solution to this equation is given by

$$x(t) = \sin(t) + x_0\, \mathrm{e}^{-\alpha t}.$$

When $\alpha \in \mathbf{R}$ is large, the distance between the solution and the function $t \mapsto \sin(t)$ converges to zero very quickly, regardless of the initial condition. This behavior is illustrated in Figure 7.3.

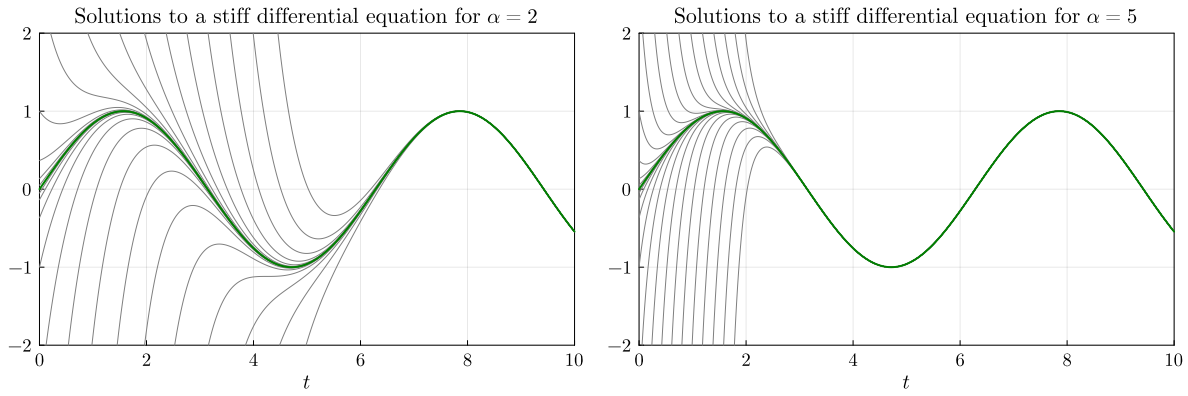

Figure 7.3: Solutions to (7.40) for various initial conditions when $\alpha = 2$ (left) and $\alpha = 5$ (right).

In the rest of this section, we use the differential equation (7.40) as a guiding example. For this problem, we have $\partial_x f(t, x) = \alpha$. Therefore, in view of (7.39), we expect that the forward Euler scheme is non-divergent if $|1 - \alpha\Delta| < 1$, i.e. if

$$\Delta < \Delta_* = \frac{2}{\alpha}.$$

It turns out that this prediction is precise, as depicted in Figure 7.4. Note that if the equation is very stiff, that is to say if $\alpha \gg 1$, then a very small time step is required to ensure stability.

In contrast with the forward Euler scheme, the backward Euler scheme is stable regardless of the time step. Since the right-hand side of (7.40) is linear in x, the value of the iterate x_{n+1} can be calculated explicitly from x_n for the backward scheme:

$$x_{n+1} = \frac{x_n + \Delta\alpha \sin(t_{n+1}) + \Delta \cos(t_{n+1})}{1 + \Delta\alpha}.$$

Numerical approximations obtained using this scheme are illustrated in Figure 7.5. We observe that the method is stable even for the large time step $\Delta = 2\Delta_*$.

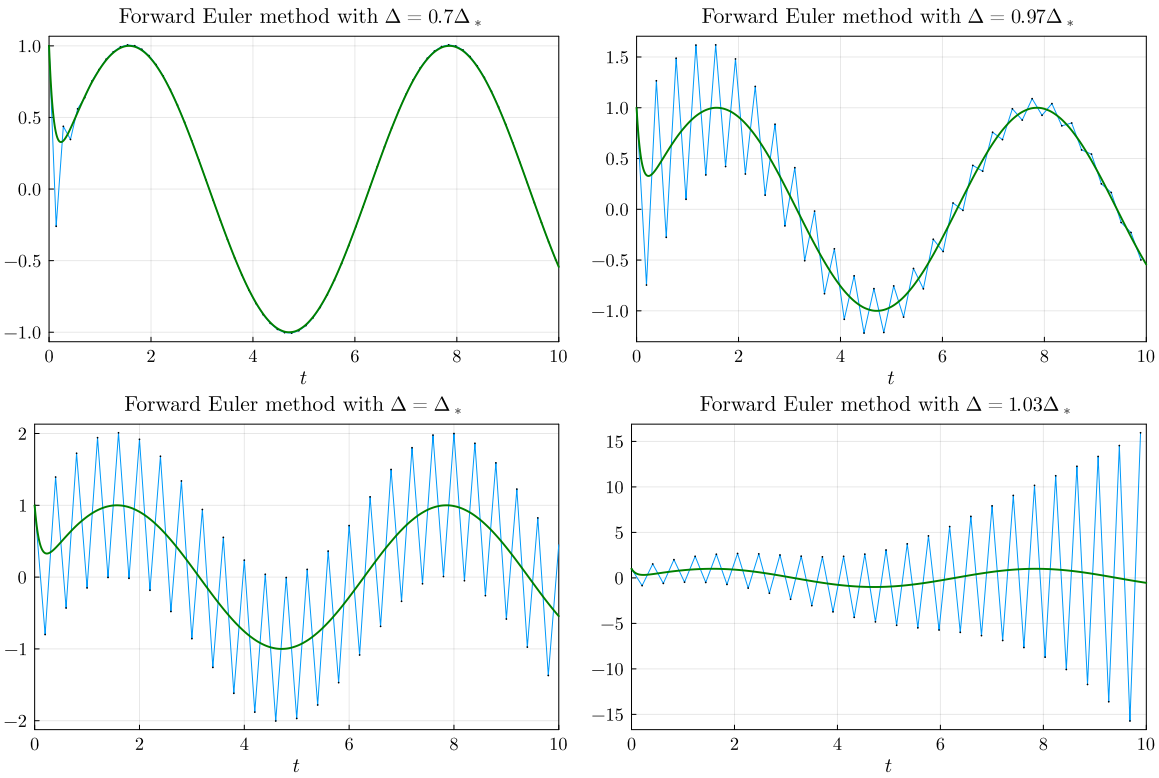

Figure 7.4: Numerical approximations of the solution to (7.40) with $\alpha = 10$ obtained with the forward Euler method, for four different values of Δ.

7.5 Exercises

Exercise 7.1. *Show that the absolute stability region of the Crank–Nicolson method* (7.24) *is given by the left half-plane; see Figure 7.6.*

Exercise 7.2. *Calculate the absolute stability region for Gear's method.*

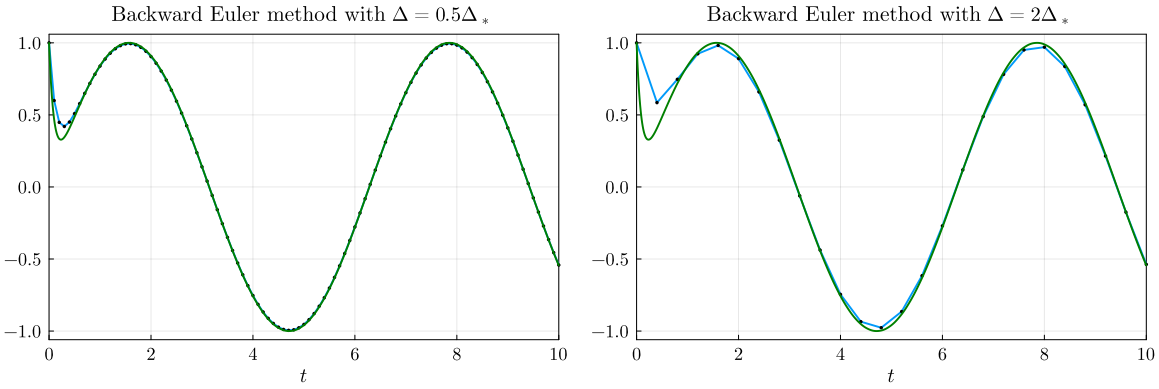

Figure 7.5: Numerical approximations of the solution to (7.40) with $\alpha = 10$ obtained with the backward Euler method, for two different values of Δ.

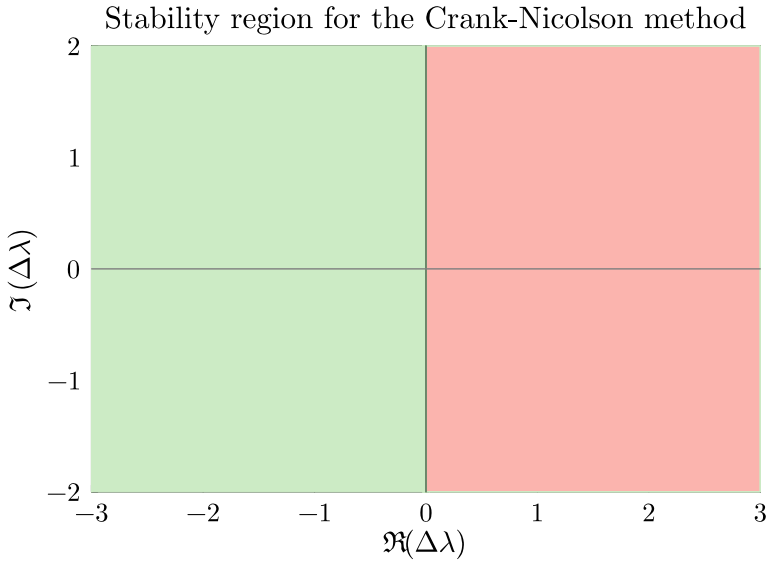

Figure 7.6: Absolute stability regions for the Crank Nicolson method.