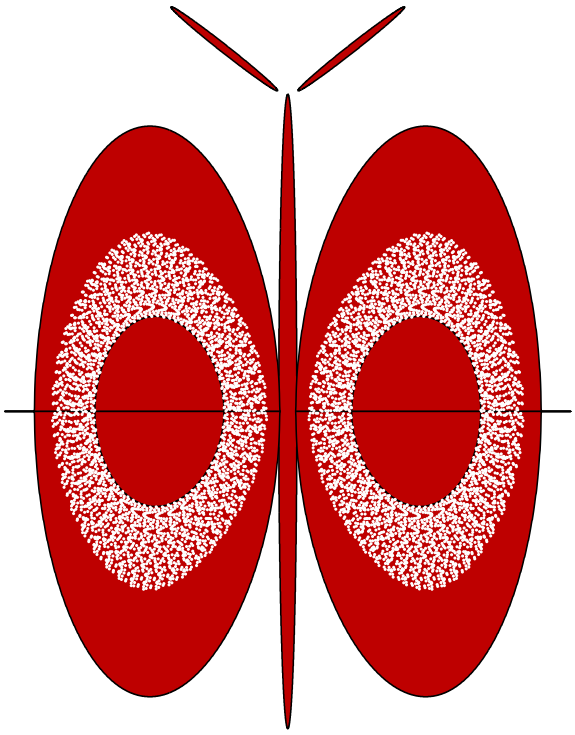
Numerical methods for models in atmospheric sciences

Mechthild Thalhammer



Leopold–Franzens Universität Innsbruck

Winter term 2016/17

Course.

Numerical methods for models in atmospheric sciences
Numerische Methoden für Modelle der Atmosphärenwissenschaften
707706 & 707707, VO 2 & PS 1, 3.5 & 1.5 ECTS
Pflichtmodul Numerische Methoden

Lecture: Description of numerical methods for partial differential equations, with regard to applications in the atmospheric sciences. Assessment based upon final written examination.

Exercises: Discussion of practical aspects and implementation of low-order space and time discretisation methods for fundamental model equations. Assessment based upon written and oral contributions.

Main reference.

DALE R. DURRAN
Numerical Methods for Fluid Dynamics: With Applications to Geophysics
Texts in Applied Mathematics 32, Springer, New York, 2010.

For helpful suggestions, thanks to

Alexander Gohm, Georg Mayr, Mathias Rotach

Contents

Chapter 1

Basic means

Preliminary remarks. In this chapter, we recall the definitions of partial and total derivatives for functions in several variables. Moreover, we introduce basic means that are needed in different contexts.

- *Taylor series expansions.* Taylor series expansions yield local approximations to smooth functions by polynomials. In numerical mathematics, Taylor series expansions are frequently used for the construction and error analysis of numerical methods.
- *Fourier series expansions.* For specific types of (linear) partial differential equations, it is beneficial to employ solution representations based on Fourier series expansions; in conjunction with the Fast Fourier transform, this approach then leads to approximations that are superior in accuracy and efficiency compared to other space discretisation methods.
- *Finite difference approximations.* Various space and time discretisation methods for partial differential equations rely on the idea to replace the arising differential quotients by finite difference approximations.

1.1 Derivatives

Functions in a single variable. For a real-valued function in a single variable

$$f : \mathbb{R} \longrightarrow \mathbb{R} : x \longmapsto f(x) ,$$

the first derivative is defined as (provided that limit exists, differential quotient)

$$f' : \mathbb{R} \longrightarrow \mathbb{R} : x \longmapsto f'(x) = \lim_{\xi \to 0} \frac{f(x+\xi) - f(x)}{\xi} .$$

Evidently, each value of the first derivative yields a linear function (indicate dependence on $x \in \mathbb{R}$, trivial case where linear function corresponds to (1×1)-matrix)

$$A_x : \mathbb{R} \longrightarrow \mathbb{R} : \zeta \longmapsto A_x \zeta = f'(x)\, \zeta .$$

The tangent line in $x \in \mathbb{R}$ is given by the affine-linear function (indicate dependence on $x \in \mathbb{R}$, tangent line is determined by conditions $T_x(x) = f(x)$ and $T_x'(x) = f'(x)$, i.e., same value and slope at $x \in \mathbb{R}$)

$$T_x : \mathbb{R} \longrightarrow \mathbb{R} : \zeta \longmapsto T_x(\zeta) = f(x) + f'(x)\,(\zeta - x) .$$

In the present situation, differentiability is equivalent to the property that function values can be represented as (use relation for tangent line with $\zeta = x + \xi$ and $\xi = \zeta - x$)

$$\begin{aligned} f'(x) = \lim_{\xi \to 0} \frac{f(x+\xi) - f(x)}{\xi} \quad &\Longleftrightarrow \quad \lim_{\xi \to 0} \Big(\frac{f(x+\xi) - f(x) - f'(x)\,\xi}{\xi} \Big) = 0 \\ &\Longleftrightarrow \quad f(x+\xi) - f(x) - f'(x)\,\xi = r(\xi) \text{ with } \lim_{\xi \to 0} \frac{r(\xi)}{\xi} = 0 \\ &\Longleftrightarrow \quad f(\zeta) = T_x(\zeta) + r(\zeta - x) \text{ with } \lim_{\zeta \to x} \frac{r(\zeta - x)}{\zeta - x} = 0 , \end{aligned}$$

that is, for values $\zeta \in \mathbb{R}$ close to $x \in \mathbb{R}$, the associated affine-linear function provides an adequate approximation to the function.

Extension. It is straightforward to extend the above approach to functions of the form

$$f : \mathbb{R} \longrightarrow \mathbb{R}^d : x \longmapsto f(x) = \big(f_1(x), \ldots, f_d(x) \big)^T$$

by considering each component function.

Functions in several variables. For a real-valued function in several variables

$$f : \mathbb{R}^d \longrightarrow \mathbb{R} : x = (x_1, \ldots, x_d)^T \longmapsto f(x) ,$$

differentiability along the cartesian coordinate axes is no longer equivalent to the property that the function can be adequately approximated by an affine-linear function; the second

concept leads to the definition of the total derivative and the first to the weaker notion of partial derivatives. By fixing all components but for instance the first one, the considered function reduces to a real-valued function in a single variable

$$f(\cdot, x_2, \ldots, x_d) : \mathbb{R} \longrightarrow \mathbb{R} : x_1 \longmapsto f(x);$$

accordingly, the partial derivative with respect to the first coordinate is defined as (derivative in direction of first standard unit vector $e_1 = (1, 0, \ldots, 0)^T \in \mathbb{R}^d$, components $x_2, \ldots, x_d$ fixed)

$$\partial_{x_1} f : \mathbb{R}^d \longrightarrow \mathbb{R} : x \longmapsto \partial_{x_1} f(x) = \lim_{\xi \to 0} \frac{f(x + \xi e_1) - f(x)}{\xi} = \lim_{\xi \to 0} \frac{f(x_1 + \xi, x_2, \ldots, x_d) - f(x)}{\xi}.$$

Analogous definitions hold for the partial derivatives $\partial_{x_2} f, \ldots, \partial_{x_d} f : \mathbb{R}^d \to \mathbb{R}$. In general, the existence of all partial derivatives *does not* imply the existence of the tangent plane; as the function might be discontinuous, approximations by affine-linear functions might be rather poor. Under the additional requirement that any partial derivative is continuous, however, the tangent plane is given by the affine-linear function

$$T_x : \mathbb{R}^d \longrightarrow \mathbb{R} : \zeta \longmapsto f(x) + f'(x)\,(\zeta - x),$$

which involves the (total) derivative

$$f : \mathbb{R} \longrightarrow \mathbb{R}^d : x \longmapsto f'(x) = \big(\partial_{x_1} f(x), \ldots, \partial_{x_d} f(x)\big).$$

In particular, the following relation holds

$$f(\zeta) = T_x(\zeta) + r(\zeta - x) \text{ with } \lim_{\zeta \to x} \frac{r(\zeta - x)}{\|\zeta - x\|} = 0,$$

that is, for values $\zeta \in \mathbb{R}^d$ close to $x \in \mathbb{R}^d$, the associated affine-linear function provides an adequate approximation to the function.

Extension. As before, it is straightforward to extend the above approach to functions of the form

$$f : \mathbb{R}^d \longrightarrow \mathbb{R}^d : x = (x_1, \ldots, x_d)^T \longmapsto f(x) = \big(f_1(x), \ldots, f_d(x)\big)^T$$

by considering each component function.

Evaluation of first derivative. For a real-valued function in several variables

$$f : \mathbb{R}^d \longrightarrow \mathbb{R} : x = (x_1, \ldots, x_d)^T \longmapsto f(x),$$

the first derivative is given by

$$f'(x) : \mathbb{R}^d \longrightarrow \mathbb{R} : \zeta \longmapsto f'(x)\,\zeta,$$

$$f'(x)\,\zeta = \big(\partial_{x_1} f(x), \ldots, \partial_{x_d} f(x)\big) \begin{pmatrix} \zeta_1 \\ \vdots \\ \zeta_d \end{pmatrix} = \partial_{x_1} f(x)\,\zeta_1 + \cdots + \partial_{x_d} f(x)\,\zeta_d.$$

More generally, for a vector-valued function such as

$$f:\mathbb{R}^d \longrightarrow \mathbb{R}^d : x=(x_1,\dots,x_d)^T \longmapsto \big(f_1(x),\dots,f_d(x)\big)^T,$$

we obtain the representation

$$f'(x):\mathbb{R}^d \longrightarrow \mathbb{R}^d : \zeta \longmapsto f'(x)\,\zeta,$$
$$f'(x)\,\zeta = \begin{pmatrix} \partial_{x_1} f_1(x) & \dots & \partial_{x_d} f_1(x) \\ \vdots & & \vdots \\ \partial_{x_1} f_d(x) & \dots & \partial_{x_d} f_d(x) \end{pmatrix} \begin{pmatrix} \zeta_1 \\ \vdots \\ \zeta_d \end{pmatrix}.$$

Evaluation of higher derivatives. For a real-valued function in several variables

$$f:\mathbb{R}^d \longrightarrow \mathbb{R} : x=(x_1,\dots,x_d)^T \longmapsto f(x),$$

the values of the second derivative can be computed by matrix-vector multiplications (second derivative acts as a bilinear form, use that $\partial_{x_k x_\ell} f = \partial_{x_\ell x_k} f$ and thus $f''(x) = (f''(x))^T$ if function is twice continuously differentiable)

$$f''(x):\mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R} : \big(\zeta,\widetilde{\zeta}\big) \longmapsto f''(x)\,\big(\zeta,\widetilde{\zeta}\big),$$
$$f''(x) = \begin{pmatrix} \partial_{x_1 x_1} f(x) & \dots & \partial_{x_1 x_d} f(x) \\ \vdots & & \vdots \\ \partial_{x_1 x_d} f(x) & \dots & \partial_{x_d x_d} f(x) \end{pmatrix}$$
$$f''(x)\,(\zeta,\widetilde{\zeta}) = \big(\zeta_1 \dots,\zeta_d\big) \begin{pmatrix} \partial_{x_1 x_1} f(x) & \dots & \partial_{x_1 x_d} f(x) \\ \vdots & & \vdots \\ \partial_{x_1 x_d} f(x) & \dots & \partial_{x_d x_d} f(x) \end{pmatrix} \begin{pmatrix} \widetilde{\zeta}_1 \\ \vdots \\ \widetilde{\zeta}_d \end{pmatrix} = \sum_{k,\ell=1}^{d} \partial_{x_k x_\ell} f(x)\,\zeta_k \widetilde{\zeta}_\ell.$$

For higher derivatives, however, a representation by matrices is no longer possible; the values of the m-th derivative are computed as

$$f^{(m)}(x):\mathbb{R}^d \times \dots \times \mathbb{R}^d \longrightarrow \mathbb{R} : \big(\zeta^{(1)},\dots,\zeta^{(m)}\big) \longmapsto f^{(m)}(x)\,\big(\zeta^{(1)},\dots,\zeta^{(m)}\big),$$
$$f^{(m)}(x)\,\big(\zeta^{(1)},\dots,\zeta^{(m)}\big) = \sum_{k_1,\dots,k_m=1}^{d} \partial_{x_{k_1}\cdots x_{k_m}} f(x)\,\zeta^{(1)}_{k_1} \cdots \zeta^{(m)}_{k_m}.$$

For a vector-valued function in several variables, the above relation holds for each component function.

Gradient, Divergence. For a real-valued function in several variables, we employ the common notation (gradient)

$$f:\mathbb{R}^d \longrightarrow \mathbb{R} : x=(x_1,\dots,x_d)^T \longmapsto f(x),$$
$$\nabla f(x) = \begin{pmatrix} \partial_{x_1} f(x) \\ \vdots \\ \partial_{x_d} f(x) \end{pmatrix};$$

we note that $\nabla f(x) = (f'(x))^T$. Furthermore, for a vector-valued function in several variables, we set (divergence, omit parantheses $(\nabla \cdot F)(x) = \nabla \cdot F(x)$)

$$F : \mathbb{R}^d \longrightarrow \mathbb{R}^d : x = (x_1, \ldots, x_d)^T \longmapsto F(x) = \begin{pmatrix} F_1(x) \\ \vdots \\ F_d(x) \end{pmatrix},$$

$$\nabla \cdot F(x) = \mathrm{div} F(x) = \partial_{x_1} F_1(x) + \cdots + \partial_{x_d} F_d(x)\,.$$

1.2 Taylor series expansions

Situation. We consider a sufficiently often differentiable function

$$f : \mathbb{R} \longrightarrow \mathbb{R} : x \longmapsto f(x) .$$

A basic mean for the derivation of the Taylor series expansion with remainder in integral form is the integration-by-parts formula; an alternative representation of the remainder is obtained with the help of the mean value theorem.

Integration-by-parts. For (sufficiently regular) functions $u, v : \mathbb{R} \to \mathbb{R}$, the integration-by-parts formula

$$\int_a^b u'(\xi)\, v(\xi)\, \mathrm{d}\xi = u(\xi)\, v(\xi)\Big|_a^b - \int_a^b u(\xi)\, v'(\xi)\, \mathrm{d}\xi$$

follows at once from the product rule

$$\begin{aligned} \tfrac{\mathrm{d}}{\mathrm{d}\xi}\big(u(\xi)\, v(\xi)\big) &= u'(\xi)\, v(\xi) + u(\xi)\, v'(\xi) , \\ u'(\xi)\, v(\xi) &= \tfrac{\mathrm{d}}{\mathrm{d}\xi}\big(u(\xi)\, v(\xi)\big) - u(\xi)\, v'(\xi) . \end{aligned}$$

Mean value theorem. Assume that $v : \mathbb{R} \to \mathbb{R}$ is non-negative or non-positive, respectively, over the considered interval of integration. The mean value theorem states that there exists an element $\zeta \in [a, b]$ such that the relation

$$\int_a^b u(\xi)\, v(\xi)\, \mathrm{d}\xi = u(\zeta) \int_a^b v(\xi)\, \mathrm{d}\xi$$

holds; in particular, by setting $v = 1$, the identity

$$\int_a^b u(\xi)\, \mathrm{d}\xi = u(\zeta)\, (b - a)$$

is obtained.

Taylor series expansion. A repeated application of the integration-by-parts formula implies the following Taylor series expansion with remainder in integral form (for any $x \in \mathbb{R}$ and $m \in \mathbb{N}$, with center $a \in \mathbb{R}$)

$$\begin{aligned} f(x) = f(a) + f'(a)\,(x-a) + \cdots + \tfrac{1}{m!}\, f^{(m)}(a)\,(x-a)^m \\ + \int_0^1 \tfrac{1}{m!}\,(1-\xi)^m\, f^{(m+1)}\big(\xi x + (1-\xi)a\big)\,(x-a)^{m+1}\, \mathrm{d}\xi ; \end{aligned}$$

with the help of the mean value theorem, the remainder takes the form (with a certain node $\zeta \in [\min\{a, x\}, \max\{a, x\}]$)

$$\begin{aligned} f(x) = f(a) + f'(a)\,(x-a) + \cdots + \tfrac{1}{m!}\, f^{(m)}(a)\,(x-a)^m \\ + \tfrac{1}{(m+1)!}\, f^{(m+1)}(\zeta)\,(x-a)^{m+1} . \end{aligned}$$

In particular, for $m = 1$ and $m = 2$, respectively, we get the expansions (with certain nodes $\zeta_1, \zeta_2 \in [\min\{a,x\}, \max\{a,x\}]$)

$$\begin{aligned}
f(x) &= f(a) + f'(a)\,(x-a) + \int_0^1 (1-\xi)\, f''\big(\xi x + (1-\xi)a\big)\,(x-a)^2\,\mathrm{d}\xi \\
&= f(a) + f'(a)\,(x-a) + \tfrac{1}{2}\, f''(\zeta_1)\,(x-a)^2\,, \\
f(x) &= f(a) + f'(a)\,(x-a) + \tfrac{1}{2}\, f''(a)\,(x-a)^2 + \int_0^1 \tfrac{1}{2}\,(1-\xi)^2\, f'''\big(\xi x + (1-\xi)a\big)\,(x-a)^3\,\mathrm{d}\xi \\
&= f(a) + f'(a)\,(x-a) + \tfrac{1}{2}\, f''(a)\,(x-a)^2 + \tfrac{1}{6}\, f'''(\zeta_2)\,(x-a)^3\,.
\end{aligned}$$

Explanation. The derivation of the Taylor series expansion with remainder in integral form relies on the main theorem of differential and integral calculus

$$\begin{aligned}
f(x) &= f(a) + f\big(\xi x + (1-\xi)a\big)\Big|_{\xi=0}^{1} \\
&= f(a) + \int_0^1 \tfrac{\mathrm{d}}{\mathrm{d}\xi}\, f\big(\xi x + (1-\xi)a\big)\,\mathrm{d}\xi \\
&= f(a) + \int_0^1 f'\big(\xi x + (1-\xi)a\big)\,(x-a)\,\mathrm{d}\xi\,.
\end{aligned}$$

A first application of integration-by-parts yields (set $u'(\xi) = 1$ and employ special choice $u(\xi) = -(1-\xi)$, for $v(\xi) = f'(\xi x + (1-\xi)a)\,(x-a)$ obtain $v'(\xi) = f''(\xi x + (1-\xi)a)\,(x-a)^2$)

$$\begin{aligned}
f(x) &= f(a) - (1-\xi)\, f'\big(\xi x + (1-\xi)a\big)\,(x-a)\Big|_{\xi=0}^{1} \\
&\qquad + \int_0^1 (1-\xi)\, f''\big(\xi x + (1-\xi)a\big)\,(x-a)^2\,\mathrm{d}\xi \\
&= f(a) + f'(a)\,(x-a) \\
&\qquad + \int_0^1 (1-\xi)\, f''\big(\xi x + (1-\xi)a\big)\,(x-a)^2\,\mathrm{d}\xi\,;
\end{aligned}$$

analogously, employing integration-by-parts, we obtain

$$\begin{aligned}
f(x) &= f(a) + f'(a)\,(x-a) - \tfrac{1}{2}\,(1-\xi)^2\, f''\big(\xi x + (1-\xi)a\big)\,(x-a)^2\Big|_{\xi=0}^{1} \\
&\qquad + \int_0^1 \tfrac{1}{2}\,(1-\xi)^2\, f'''\big(\xi x + (1-\xi)a\big)\,(x-a)^3\,\mathrm{d}\xi \\
&= f(a) + f'(a)\,(x-a) + \tfrac{1}{2}\, f''(a)\,(x-a)^2 \\
&\qquad + \int_0^1 \tfrac{1}{2}\,(1-\xi)^2\, f'''\big(\xi x + (1-\xi)a\big)\,(x-a)^3\,\mathrm{d}\xi
\end{aligned}$$

as well as

$$\begin{aligned} f(x) &= f(a) + f'(a)\,(x-a) + \tfrac{1}{2}\,f''(a)\,(x-a)^2 - \tfrac{1}{6}\,(1-\xi)^3\,f'''\big(\xi x + (1-\xi)a\big)\,(x-a)^3\Big|_{\xi=0}^{1} \\ &\quad + \int_0^1 \tfrac{1}{6}\,(1-\xi)^3\,f^{(4)}\big(\xi x + (1-\xi)a\big)\,(x-a)^4\,\mathrm{d}\xi \\ &= f(a) + f'(a)\,(x-a) + \tfrac{1}{2}\,f''(a)\,(x-a)^2 + \tfrac{1}{6}\,f'''(a)\,(x-a)^3 \\ &\quad + \int_0^1 \tfrac{1}{6}\,(1-\xi)^3\,f^{(4)}\big(\xi x + (1-\xi)a\big)\,(x-a)^4\,\mathrm{d}\xi\,. \end{aligned}$$

Moreover, by induction, the expansion

$$\begin{aligned} f(x) &= f(a) + f'(a)\,(x-a) + \cdots + \tfrac{1}{m!}\,f^{(m)}(a)\,(x-a)^m \\ &\quad + \int_0^1 \tfrac{1}{m!}\,(1-\xi)^m\,f^{(m+1)}\big(\xi x + (1-\xi)a\big)\,(x-a)^{m+1}\,\mathrm{d}\xi \end{aligned}$$

follows. Applying the mean value theorem with $v(\xi) = \frac{1}{m!}(1-\xi)^m$ finally yields

$$\begin{aligned} &\int_0^1 \tfrac{1}{m!}\,(1-\xi)^m\,f^{(m+1)}\big(\xi x + (1-\xi)a\big)\,(x-a)^{m+1}\,\mathrm{d}\xi \\ &\qquad = f^{(m+1)}(\zeta)\,(x-a)^{m+1}\int_0^1 \tfrac{1}{m!}\,(1-\xi)^m\,\mathrm{d}\xi \\ &\qquad = \tfrac{1}{(m+1)!}\,f^{(m+1)}(\zeta)\,(x-a)^{m+1}\,, \end{aligned}$$

which is the stated result. ◇

Extension. The Taylor series expansion with remainder in integral form immediately extends to sufficiently often differentiable vector-valued functions in several variables such as $f : \mathbb{R}^d \to \mathbb{R}^d$. With regard to the estimate (higher derivative defines multi-linear form)

$$\begin{aligned} &\Big\| f(x) - \big(f(a) + f'(a)\,(x-a) + \cdots + \tfrac{1}{m!}\,f^{(m)}(a)(x-a,\ldots,x-a)\big)\Big\| \\ &\qquad \leq \tfrac{1}{(m+1)!}\sup_{\xi\in[0,1]}\Big\| f^{(m+1)}\big(\xi x + (1-\xi)a\big)\Big\|\,\|x-a\|^{m+1}\,, \end{aligned}$$

it is often convenient to employ the symbolic notation

$$\begin{aligned} f(x) &= f(a) + f'(a)\,(x-a) + \cdots + \tfrac{1}{m!}\,f^{(m)}(a)(x-a,\ldots,x-a) \\ &\quad + \mathscr{O}\big(f^{(m+1)}, \|x-a\|^{m+1}\big)\,. \end{aligned}$$

We note that the attempt to generalise the mean value theorem to vector-valued functions in several variables fails, in general.

1.3 Fourier series expansions

Situation. We consider a complex-valued function that is defined on a bounded interval (with $a, b \in \mathbb{R}$ such that $a < b$)

$$f : [a,b] \subset \mathbb{R} \longrightarrow \mathbb{C} : x \longmapsto f(x);$$

at first, no additional regularity requirement is needed.

Space of square-integrable functions. In the context of Fourier series expansions, it is natural to consider the linear space of square-integrable functions

$$L^2\big([a,b],\mathbb{C}\big) = \left\{ f : [a,b] \longrightarrow \mathbb{C} \text{ such that } \int_a^b \big|f(\xi)\big|^2 \,\mathrm{d}\xi < \infty \right\},$$

which forms a Hilbert space when endowed with inner product and associated norm (complex conjugation in second argument, norm provides mean to quantify distance of functions)

$$\big(f\big|g\big)_{L^2} = \int_a^b f(\xi)\,\overline{g(\xi)}\,\mathrm{d}\xi\,, \qquad f,g \in L^2\big([a,b],\mathbb{C}\big),$$

$$\big\|f\big\|_{L^2} = \sqrt{\big(f\big|f\big)_{L^2}} = \sqrt{\int_a^b \big|f(\xi)\big|^2 \,\mathrm{d}\xi}\,, \qquad f \in L^2\big([a,b],\mathbb{C}\big).$$

Orthogonality. A family of functions $f_1,\ldots,f_M \in L^2([a,b],\mathbb{C})$ is called orthonormal iff the condition (Kronecker-delta, orthogonality, normalisation $\|f_m\|_{L^2} = 1$)

$$\big(f_\ell\big|f_m\big)_{L^2} = \delta_{\ell m} = \begin{cases} 1 & \text{if } \ell = m, \\ 0 & \text{otherwise}, \end{cases}$$

is satisfied for all $\ell, m \in \{1,\ldots,M\}$. The representation of a function $g \in L^2([a,b],\mathbb{C})$ with respect to a family of orthonormal functions is particularly simple (assume that function is given as linear combination involving complex coefficients $\alpha_1,\ldots,\alpha_M \in \mathbb{C}$)

$$g = \sum_{m=1}^{M} \alpha_m f_m \quad \Longrightarrow \quad g = \sum_{m=1}^{M} \big(g\big|f_m\big)_{L^2} f_m\,,$$

since (determine inner product with f_m, use linearity in first component)

$$g = \sum_{\ell=1}^{M} \alpha_\ell f_\ell\,,$$

$$\big(g\big|f_m\big)_{L^2} = \left(\sum_{\ell=1}^{M} \alpha_\ell f_\ell \middle| f_m\right)_{L^2} = \sum_{\ell=1}^{M} \alpha_\ell \big(f_\ell\big|f_m\big)_{L^2} = \sum_{\ell=1}^{M} \alpha_\ell \delta_{\ell m} = \alpha_m\,.$$

This immediately implies that orthonormal functions are in particular linearly independent.

Fourier basis functions. The Fourier basis functions are defined as

$$\mathscr{F}_m : \mathbb{R} \longrightarrow \mathbb{C} : x \longmapsto \tfrac{1}{\sqrt{b-a}}\, \mathrm{e}^{\frac{2\pi}{b-a}\,\mathrm{i}\,m\,(x-a)}\,, \qquad m \in \mathbb{Z}\,;$$

evidently, they are periodic on the considered interval

$$\mathscr{F}_m(a) = \tfrac{1}{\sqrt{b-a}} = \mathscr{F}_m(b)\,, \qquad m \in \mathbb{Z}\,.$$

A straightforward calculation (for all $\ell, m \in \mathbb{Z}$, use periodicity)

$$\begin{aligned}
\ell \neq m\,: \qquad \big(\mathscr{F}_\ell \big| \mathscr{F}_m\big)_{L^2} &= \int_a^b \mathscr{F}_\ell(\xi)\, \overline{\mathscr{F}_m(\xi)}\, \mathrm{d}\xi \\
&= \tfrac{1}{b-a} \int_a^b \mathrm{e}^{\frac{2\pi}{b-a}\,\mathrm{i}\,\ell\,(x-a)}\, \mathrm{e}^{-\frac{2\pi}{b-a}\,\mathrm{i}\,m\,(x-a)}\, \mathrm{d}\xi \\
&= \tfrac{1}{b-a} \int_a^b \mathrm{e}^{\frac{2\pi}{b-a}\,\mathrm{i}\,(\ell-m)\,(x-a)}\, \mathrm{d}\xi \\
&= \tfrac{1}{2\pi\,\mathrm{i}\,(\ell-m)}\, \mathrm{e}^{\frac{2\pi}{b-a}\,\mathrm{i}\,(\ell-m)\,(x-a)} \Big|_a^b \\
&= 0\,, \\
\ell = m\,: \qquad \big\|\mathscr{F}_m\big\|_{L^2}^2 &= \big(\mathscr{F}_m \big| \mathscr{F}_m\big)_{L^2} \\
&= \tfrac{1}{b-a} \int_a^b 1\, \mathrm{d}\xi \\
&= 1\,,
\end{aligned}$$

confirms that the Fourier basis functions are orthonormal

$$\big(\mathscr{F}_\ell \big| \mathscr{F}_m\big)_{L^2} = \delta_{\ell m}\,, \qquad \ell, m \in \mathbb{Z}\,;$$

furthermore, the family of Fourier basis functions forms a complete orthonormal system, that is, for any function $f \in L^2([a,b],\mathbb{C})$ the following representation as infinite series, commonly referred to as Fourier series expansion, is valid

$$f = \sum_{m\in\mathbb{Z}} f_m\, \mathscr{F}_m\,, \qquad f_m = \big(f \big| \mathscr{F}_m\big)_{L^2} = \int_a^b f(\xi)\, \overline{\mathscr{F}_m(\xi)}\, \mathrm{d}\xi \in \mathbb{C}\,, \qquad m \in \mathbb{Z}\,.$$

In addition, by Parseval's identity, we get (recall definition of norm)

$$\big\|f\big\|_{L^2} = \sqrt{\int_a^b \big|f(\xi)\big|^2\, \mathrm{d}\xi} = \sqrt{\sum_{m\in\mathbb{Z}} \big|f_m\big|^2}\,.$$

Provided that the periodic continuation of the considered function is differentiable, the representation as Fourier series holds pointwise (for all $x \in [a,b]$)

$$f(x) = \sum_{m\in\mathbb{Z}} \big(f \big| \mathscr{F}_m\big)_{L^2}\, \mathscr{F}_m(x)\,.$$

Numerical realisation. The numerical realisation of the representation

$$f(x) = \sum_{m\in\mathbb{Z}} f_m \mathscr{F}_m(x)\,, \qquad f_m = \big(f\big|\mathscr{F}_m\big)_{L^2} = \int_a^b f(\xi)\,\overline{\mathscr{F}_m(\xi)}\,\mathrm{d}\xi\,, \qquad m\in\mathbb{Z}\,,$$

relies on a truncation of the infinite series and a quadrature approximation of the spectral coefficients.

(i) *Approximation by finite series.* For a sufficiently large even integer number $M\in\mathbb{N}$, we replace the infinite Fourier series expansion by a sum involving M Fourier basis functions

$$\sum_{m=-\frac{M}{2}}^{\frac{M}{2}-1} \big(f\big|\mathscr{F}_m\big)_{L^2}\,\mathscr{F}_m(x) \;\approx\; f(x) = \sum_{m\in\mathbb{Z}} \big(f\big|\mathscr{F}_m\big)_{L^2}\,\mathscr{F}_m(x)\,.$$

(ii) *Trapezoidal rule.* In connection with Fourier basis functions, it is most natural to consider the trapezoidal rule on a uniform mesh for the approximation of integrals (sufficiently large integer number $K\in\mathbb{N}$, grid width $h>0$ and associated equidistant grid points, evidently $x_0=a$ as well as $x_K=b$, area of trapezoid)

$$h = \tfrac{b-a}{K}\,, \qquad x_k = a+kh\,, \qquad k\in\{0,\dots,K\}\,,$$
$$h\,g(x_k) + \tfrac{h}{2}\big(g(x_{k+1})-g(x_k)\big) = \tfrac{h}{2}\big(g(x_k)+g(x_{k+1})\big) \;\approx\; \int_{x_k}^{x_{k+1}} g(\xi)\,\mathrm{d}\xi\,,$$
$$\tfrac{h}{2}\sum_{k=0}^{K-1}\big(g(x_k)+g(x_{k+1})\big) = \tfrac{h}{2}\,g(x_0) + h\sum_{k=1}^{K-1} g(x_k) + \tfrac{h}{2}\,g(x_K) \;\approx\; \int_a^b g(\xi)\,\mathrm{d}\xi\,.$$

If the considered function is periodic, the quadrature approximation reduces as follows (use that $g(x_K)=g(x_0)$)

$$g(a)=g(b):\qquad h\sum_{k=0}^{K-1} g(x_k) \;\approx\; \int_a^b g(\xi)\,\mathrm{d}\xi\,.$$

(iii) *Approximations to spectral coefficients.* Under the reasonable presumption that the function $f:\mathbb{R}\to\mathbb{R}$ is periodic on the interval $[a,b]$, the application of the trapezoidal rule yields the following approximations to the spectral coefficients

$$h\sum_{k=0}^{K-1} f(x_k)\,\overline{\mathscr{F}_m(x_k)} \;\approx\; f_m = \int_a^b f(\xi)\,\overline{\mathscr{F}_m(\xi)}\,\mathrm{d}\xi\,, \qquad m\in\big\{-\tfrac{M}{2},\dots,\tfrac{M}{2}-1\big\}\,.$$

Implementation by fast Fourier transform (FFT). In practice, we choose $K=M$. For the efficient numerical approximation of the spectral coefficients collected in a column, we employ the fast Fourier transform (meanwhile set $f_k=f(x_k)$ for $k\in\{0,\dots,M-1\}$)

$$f_m^{(s)} = h\sum_{k=0}^{M-1} f_k\,\overline{\mathscr{F}_m(x_k)}\,, \qquad m\in\big\{-\tfrac{M}{2},\dots,\tfrac{M}{2}-1\big\}\,.$$

The inverse transform yields approximations to the function values at the grid points

$$f_k = \sum_{m=-\frac{M}{2}}^{\frac{M}{2}-1} f_m^{(s)}\, \mathscr{F}_m(x_k)\,, \qquad k \in \{0,\dots,M-1\}\,.$$

With suitably chosen constants, the implementation in MATLAB reads as follows.

```
function fs = Fourier_Real2Spectral(f)
    fs = Const*fftshift(fftn(f));
end

function f = Fourier_Spectral2Real(fs)
    f = Const*ifftn(ifftshift(fs));
end
```

Extension. Provided that the underlying domain is of the special form (cartesian product of bounded intervals)

$$\Omega = [a_1, b_1] \times \cdots \times [a_d, b_d] \subset \mathbb{R}^d\,,$$

it is straightfoward to extend the considerations to square-integrable functions in several variables

$$f : \Omega \longrightarrow \mathbb{C}\,, \qquad \|f\|_{L^2} = \sqrt{\int_\Omega |f(\xi)|^2 \,\mathrm{d}\xi} < \infty,$$

by using the tensor product of Fourier basis functions

$$\mathscr{F}_m(x) = \mathscr{F}_{(m_1,\dots,m_d)}(x_1,\dots,x_d) = \mathscr{F}_{m_1}(x_1)\cdots\mathscr{F}_{m_d}(x_d)\,, \qquad x \in \Omega\,, \qquad m \in \mathbb{Z}^d\,,$$
$$f = \sum_{m\in\mathbb{Z}^d} f_m\, \mathscr{F}_m\,, \qquad f_m = \big(f \big| \mathscr{F}_m\big)_{L^2} = \int_\Omega f(\xi)\, \overline{\mathscr{F}_m(\xi)}\,\mathrm{d}\xi \in \mathbb{C}\,, \qquad m \in \mathbb{Z}^d\,.$$

Simplifying assumption. By means of a bijective linear function, any interval $[a,b]$ transforms to the symmetric interval $[-\pi,\pi]$.

$$[a,b] \longrightarrow [0,1] \longrightarrow [-\pi,\pi] : x \longmapsto t = \tfrac{x-a}{b-a} \longmapsto -\pi + 2\pi\, t = -\pi + 2\pi\, \tfrac{x-a}{b-a}\,;$$

thus, it suffices to study the Fourier series expansion of a function $f \in L^2([-\pi,\pi],\mathbb{C})$

$$\mathscr{F}_m : [-\pi,\pi] \longrightarrow \mathbb{C} : x \longmapsto \tfrac{1}{\sqrt{2\pi}}\, \mathrm{e}^{\mathrm{i}\, m\,(x+\pi)}\,, \qquad m \in \mathbb{Z}\,,$$
$$f = \sum_{m\in\mathbb{Z}} f_m\, \mathscr{F}_m\,, \qquad f_m = \big(f \big| \mathscr{F}_m\big)_{L^2} = \int_{-\pi}^{\pi} f(\xi)\, \overline{\mathscr{F}_m(\xi)}\,\mathrm{d}\xi \in \mathbb{C}\,, \qquad m \in \mathbb{Z}\,.$$

Even functions. An even functions is characterised by the condition (for all $x \in \mathbb{R}$)

$$f(-x) = f(x)\,.$$

Euler's identity for the complex exponential function (for any $x \in \mathbb{R}$)

$$\mathrm{e}^{\mathrm{i}x} = \cos(x) + \mathrm{i}\sin(x)$$

indicates that cosine functions are well-suited to represent even functions (for $\xi, x \in \mathbb{R}$)

$$\tfrac{1}{2}\left(\mathrm{e}^{\mathrm{i}\xi} + \mathrm{e}^{-\mathrm{i}\xi}\right) = \cos(\xi)\,,$$
$$\tfrac{1}{2}\left(\mathscr{F}_m(x) + \mathscr{F}_m(-x)\right) = \tfrac{1}{2\sqrt{2\pi}}\left(\mathrm{e}^{\mathrm{i}m(x+\pi)} + \mathrm{e}^{-\mathrm{i}m(x+\pi)}\right) = \tfrac{1}{\sqrt{2\pi}}\cos\left(m(x+\pi)\right),$$

since (insert Fourier series expansion, reduction to non-negative integers)

$$\begin{aligned} f(x) &= \tfrac{1}{2}\left(f(x) + f(-x)\right) \\ &= \tfrac{1}{2}\left(\sum_{m\in\mathbb{Z}} f_m \mathscr{F}_m(x) + \sum_{m\in\mathbb{Z}} f_m \mathscr{F}_m(-x)\right) \\ &= \sum_{m\in\mathbb{Z}} f_m \tfrac{1}{2}\left(\mathscr{F}_m(x) + \mathscr{F}_m(-x)\right) \\ &= \sum_{m\in\mathbb{Z}} \tfrac{1}{\sqrt{2\pi}} f_m \cos\left(m(x+\pi)\right) \\ &= \tfrac{1}{\sqrt{2\pi}} f_0 + \tfrac{1}{\sqrt{2\pi}} \sum_{m\in\mathbb{N}} \left(f_m + f_{-m}\right)\cos\left(m(x+\pi)\right); \end{aligned}$$

that is, based on the family of cosine functions (derivative vanishes at boundary, orthogonality ensured, suitable choice of normalisation constants $c_m \in \mathbb{R}$)

$$\mathscr{C}_m : \mathbb{R} \longrightarrow \mathbb{R} : x \longmapsto c_m \cos\left(m(x+\pi)\right), \qquad m \in \mathbb{N}_0\,,$$
$$\mathscr{C}_m'(-\pi) = 0 = \mathscr{C}_m'(\pi)\,, \qquad m \in \mathbb{N}_0\,,$$
$$\left(\mathscr{C}_\ell \middle| \mathscr{C}_m\right)_{L^2} = \int_{-\pi}^{\pi} \mathscr{C}_\ell(\xi)\,\mathscr{C}_m(\xi)\,\mathrm{d}\xi = \delta_{\ell m}\,, \qquad \ell, m \in \mathbb{N}_0\,,$$

for any even function $f \in L^2([-\pi,\pi],\mathbb{C})$, the following series expansion holds (use again notation f_m for coefficients)

$$f = \sum_{m\in\mathbb{N}_0} f_m \mathscr{C}_m\,, \qquad f_m = \left(f \middle| \mathscr{C}_m\right)_{L^2} = \int_{-\pi}^{\pi} f(\xi)\,\mathscr{C}_m(\xi)\,\mathrm{d}\xi \in \mathbb{C}\,, \qquad m \in \mathbb{N}_0\,.$$

Odd functions. For odd functions, characterised by the condition (for all $x \in \mathbb{R}$)

$$f(-x) = -f(x)\,,$$

analogous arguments are applicable (for $\xi, x \in \mathbb{R}$)

$$\tfrac{1}{2}\left(\mathrm{e}^{\mathrm{i}\xi} - \mathrm{e}^{-\mathrm{i}\xi}\right) = \mathrm{i}\sin(\xi)\,,$$
$$\tfrac{1}{2}\left(\mathscr{F}_m(x) - \mathscr{F}_m(-x)\right) = \tfrac{1}{2\sqrt{2\pi}}\left(\mathrm{e}^{\mathrm{i}m(x+\pi)} - \mathrm{e}^{-\mathrm{i}m(x+\pi)}\right) = \tfrac{1}{\sqrt{2\pi}}\,\mathrm{i}\sin\left(m(x+\pi)\right),$$
$$f(x) = \tfrac{1}{2}\left(f(x) - f(-x)\right) = \tfrac{1}{\sqrt{2\pi}}\,\mathrm{i}\sum_{m\in\mathbb{N}} \left(f_m - f_{-m}\right)\sin\left(m(x+\pi)\right);$$

thus, considering the family of sine functions (vanish at boundary, orthogonality ensured, suitable choice of normalisation constants $c_m \in \mathbb{R}$)

$$\begin{gathered}\mathscr{S}_m : \mathbb{R} \longrightarrow \mathbb{R} : x \longmapsto c_m \sin\big(m\,(x+\pi)\big)\,, \qquad m \in \mathbb{N}\,,\\ \mathscr{S}_m(-\pi) = 0 = \mathscr{S}_m(\pi)\,, \qquad m \in \mathbb{N}\,,\\ \big(\mathscr{S}_\ell \big| \mathscr{S}_m\big)_{L^2} = \int_{-\pi}^{\pi} \mathscr{S}_\ell(\xi)\,\mathscr{S}_m(\xi)\,\mathrm{d}\xi = \delta_{\ell m}\,, \qquad \ell, m \in \mathbb{N}\,,\end{gathered}$$

yields the following series expansion for any odd function $f \in L^2([-\pi,\pi],\mathbb{C})$

$$f = \sum_{m\in\mathbb{N}} f_m \mathscr{S}_m\,, \qquad f_m = \big(f\big|\mathscr{S}_m\big)_{L^2} = \int_{-\pi}^{\pi} f(\xi)\,\mathscr{S}_m(\xi)\,\mathrm{d}\xi \in \mathbb{C}\,, \qquad m \in \mathbb{N}\,.$$

1.4 Finite difference approximations

Aim. We introduce numerical approximations to certain values of the first and second derivatives of a real-valued function in a single variable

$$f : \mathbb{R} \longrightarrow \mathbb{R} : x \longmapsto f(x)$$

and study the accuracy of the approximations.

Derivative. For convenience, we recall the definition of the first derivative (for any $x \in \mathbb{R}$, differential quotient)

$$f'(x) = \lim_{\xi \to 0} \frac{f(x+\xi) - f(x)}{\xi} .$$

Approximations to first derivative. The above relation for the first derivative motivates the following approximations by forward finite differences or backward finite differences, respectively (for any $x \in \mathbb{R}$, difference quotient, with suitably chosen small increment $\xi > 0$, for notational simplicity omit parentheses $\Delta_+ f(x) = (\Delta_+ f)(x)$)

$$\Delta_+ f(x) = \frac{f(x+\xi) - f(x)}{\xi} \approx f'(x) ,$$
$$\Delta_- f(x) = \frac{f(x) - f(x-\xi)}{\xi} \approx f'(x) .$$

Obviously, the finite difference approximation $\Delta_- f$ is equivalent to $\Delta_+ f$ applied with negative increment $-\xi < 0$; with regard to the discretisation of advection equations, we restrict ourselves to positive increments. As shown below, central finite differences yield improved approximations for sufficiently often differentiable functions (construction by Taylor series expansions)

$$\Delta f(x) = \frac{f(x+\xi) - f(x-\xi)}{2\xi} \approx f'(x) .$$

Approximations to second derivative. Approximations to the second derivative are for instance obtained by repeated applications of finite difference approximations to the first derivative (for any $x \in \mathbb{R}$, with suitably chosen small increments $\xi > 0$, set $\Delta_+^2 f = \Delta_+(\Delta_+ f)$)

$$\Delta_+^2 f(x) = \frac{\Delta_+ f(x+\xi) - \Delta_+ f(x)}{\xi} = \frac{f(x+2\xi) - 2f(x+\xi) + f(x)}{\xi^2} \approx f''(x) ,$$
$$\Delta_+ \Delta_- f(x) = \frac{\Delta_+ f(x) - \Delta_+ f(x-\xi)}{\xi} = \frac{f(x+\xi) - 2f(x) + f(x-\xi)}{\xi^2} \approx f''(x) ,$$
$$\Delta^2 f(x) = \frac{\Delta f(x+\xi) - \Delta f(x-\xi)}{2\xi} = \frac{f(x+2\xi) - 2f(x) + f(x-2\xi)}{4\xi^2} \approx f''(x) .$$

The approximation $\Delta_+ \Delta_- f$ is commonly referred to as central finite difference approximation; evidently, the approximation $\Delta^2 f$ is equivalent to $\Delta_+ \Delta_- f$ applied with increment 2ξ.

Approximation errors. Provided that the considered function is sufficiently often differentiable with bounded derivatives, the following estimates are valid. The forward and backward finite difference approximations to the first derivative satisfy the relations (for any $x \in \mathbb{R}$, with increment $\xi > 0$)

$$\left| \frac{f(x+\xi) - f(x)}{\xi} - f'(x) \right| \leq \tfrac{1}{2} \xi \sup_{\zeta \in [x, x+\xi]} \left| f''(\zeta) \right|,$$
$$\left| \frac{f(x) - f(x-\xi)}{\xi} - f'(x) \right| \leq \tfrac{1}{2} \xi \sup_{\zeta \in [x-\xi, x]} \left| f''(\zeta) \right|.$$

The central finite difference approximations to the first and second derivatives fulfill (for any $x \in \mathbb{R}$, with increment $\xi > 0$)

$$\left| \frac{f(x+\xi) - f(x-\xi)}{2\xi} - f'(x) \right| \leq \tfrac{1}{6} \xi^2 \sup_{\zeta \in [x-\xi, x+\xi]} \left| f'''(\zeta) \right|,$$
$$\left| \frac{f(x+\xi) - 2f(x) + f(x-\xi)}{\xi^2} - f''(x) \right| \leq \tfrac{1}{12} \xi^2 \sup_{\zeta \in [x-\xi, x+\xi]} \left| f^{(4)}(\zeta) \right|.$$

Henceforth, we also employ the convenient symbolic notation

$$\frac{f(x+\xi) - f(x)}{\xi} = f'(x) + \mathscr{O}\big(\xi, f''\big),$$
$$\frac{f(x) - f(x-\xi)}{\xi} = f'(x) + \mathscr{O}\big(\xi, f''\big),$$
$$\frac{f(x+\xi) - f(x-\xi)}{2\xi} = f'(x) + \mathscr{O}\big(\xi^2, f'''\big),$$
$$\frac{f(x+\xi) - 2f(x) + f(x-\xi)}{\xi^2} = f''(x) + \mathscr{O}\big(\xi^2, f^{(4)}\big).$$

Explanation. We employ Taylor series expansions with remainder (with certain $\zeta_+ \in [x, x+\xi]$, $\zeta_- \in [x-\xi, x]$)

$$f(x+\xi) = f(x) + \xi f'(x) + \tfrac{1}{2} \xi^2 f''(\zeta_+),$$
$$f(x-\xi) = f(x) - \xi f'(x) + \tfrac{1}{2} \xi^2 f''(\zeta_-),$$

to obtain the following relations for forward and backward finite differences

$$\frac{f(x+\xi) - f(x)}{\xi} = f'(x) + \tfrac{1}{2} \xi f''(\zeta_+),$$
$$\frac{f(x) - f(x-\xi)}{\xi} = f'(x) - \tfrac{1}{2} \xi f''(\zeta_-).$$

In a similar manner, by means of the Taylor series expansions (with certain $\zeta_+ \in [x, x+\xi]$, $\zeta_- \in [x-\xi, x]$, possibly different values at different occurences)

$$\begin{aligned} f(x+\xi) &= f(x) + \xi f'(x) + \tfrac{1}{2} \xi^2 f''(x) + \tfrac{1}{6} \xi^3 f'''(\zeta_+) \\ &= f(x) + \xi f'(x) + \tfrac{1}{2} \xi^2 f''(x) + \tfrac{1}{6} \xi^3 f'''(x) + \tfrac{1}{24} \xi^4 f^{(4)}(\zeta_+), \\ f(x-\xi) &= f(x) - \xi f'(x) + \tfrac{1}{2} \xi^2 f''(x) - \tfrac{1}{6} \xi^3 f'''(\zeta_-) \\ &= f(x) - \xi f'(x) + \tfrac{1}{2} \xi^2 f''(x) - \tfrac{1}{6} \xi^3 f'''(x) + \tfrac{1}{24} \xi^4 f^{(4)}(\zeta_-), \end{aligned}$$

we get the following relations for central finite differences

$$\frac{f(x+\xi)-f(x-\xi)}{2\,\xi} = f'(x) + \tfrac{1}{12}\,\xi^2\left(f'''(\zeta_+) + f'''(\zeta_-)\right),$$
$$\frac{f(x+\xi)-2\,f(x)+f(x-\xi)}{\xi^2} = f''(x) + \tfrac{1}{24}\,\xi^2\left(f^{(4)}(\zeta_+) + f^{(4)}(\zeta_-)\right).$$

A straightforward estimation of these relations implies the stated bounds. ◇

Realisation. For the realisation of finite difference approximations with the help of computers, one has to keep in mind that the number of digits is limited. In order to balance the approximations errors and round-off errors, it is essential to choose the increments in dependence of the machine accuracy.

Extension. Using instead a representation of the remainders in integral form, it is straightforward to extend the above considerations to vector-valued functions $f : \mathbb{R} \to \mathbb{R}^d$.

Chapter 2

Basic differential equations

Preliminary remarks. In this chapter, we introduce basic ordinary and partial differential equations. The description of favourable numerical methods will be the objective of subsequent chapters. For a start, with regard to the construction and convergence analysis of higher-order space and time discretisation methods, we suppose that the problem data and the solution satisfy suitable regularity and consistency requirements. We primarily focus on elementary linear differential equations with explicit solution representations, since these are useful initial test equations for discretisation methods. Moreover, we do not take into account the occurrence of (small) parameters, which might effect the performance of numerical methods. We point out that these presumptions are employed to provide a first insight into the topic and that the treatment of relevant practical applications will require additional, more sophisticated considerations.

- Dahlquist test equation
- Ordinary differential equations
- Linear advection equations
- Nonlinear conservation laws (Burgers' equation)
- Linear diffusion equations

2.1 Dahlquist test equation

Dahlquist test equation. In numerical mathematics, the homogeneous linear scalar ordinary differential equation

$$\begin{cases} y'(t) = \lambda\, y(t)\,, & t \in (0,\infty)\,, \qquad \lambda \in \mathbb{C}\,, \\ y(0) = y_0\,, \end{cases}$$

with solution given by the exponential function

$$y : [0,\infty) \longrightarrow \mathbb{C} : t \longmapsto y(t) = \mathrm{e}^{\,t\,\lambda}\, y_0$$

is commonly mentioned as Dahlquist[1] test equation; the real part $\lambda_1 = \Re\lambda$ determines growth or decay, respectively, and the imaginary part $\lambda_2 = \Im\lambda$ the frequency of oscillations

$$\mathrm{e}^{\,t\,\lambda} = \mathrm{e}^{\,t\,\lambda_1}\big(\cos(t\,\lambda_2) + \mathrm{i}\,\sin(t\,\lambda_2)\big)\,, \qquad t \in [0,\infty)\,.$$

Due to the fact that the considered differential equation is linear, it suffices to study the function (evolution)

$$E : [0,\infty) \longrightarrow \mathbb{C} : t \longmapsto E(t) = \mathrm{e}^{\,t\,\lambda}\,.$$

Evidently, the solution value at time $t \in [0,\infty)$ is given by

$$y(t) = E(t)\, y_0\,.$$

Practical relevance. Despite its simplicity, the Dahlquist test equation is a valuable tool to detect stability issues and thus serves as an initial test equation for time integration methods.

[1] Germund Dahlquist (January 16, 1925 to February 8, 2005)

2.2 Ordinary differential equations

Nonlinear ordinary differential equations. We consider an initial value problem of the form (defining function $F : \mathbb{R}^d \to \mathbb{R}^d$, initial value $y_0 \in \mathbb{R}^d$, solution $y : [0, T] \to \mathbb{R}^d$)

$$\begin{cases} y'(t) = F\big(y(t)\big), & t \in (0, T), \\ y(0) = y_0. \end{cases}$$

Mainly for theoretical purposes, it is convenient to introduce the function (evolution, indicate dependence on defining function, nonlinear function with respect to initial value)

$$E : [0, T] \times \mathbb{R}^d \longrightarrow \mathbb{R}^d : (t, y_0) \longmapsto E(t, y_0) = y(t).$$

Transformation to autonomous form. By adding the trivial scalar differential equation

$$\tfrac{\mathrm{d}}{\mathrm{d}t} t = 1,$$

any ordinary differential equation involving an explicit time dependency can be rewritten in autonomous form (where $G : [0, T] \times \mathbb{R}^d \to \mathbb{R}^d$, employ differential equation to obtain new defining function $F : \mathbb{R}^{d+1} \to \mathbb{R}^{d+1}$)

$$\begin{aligned} z'(t) &= G\big(t, z(t)\big), & t &\in (t_0, T), \\ y(t) &= \begin{pmatrix} t \\ z(t) \end{pmatrix}, & t &\in [t_0, T], \\ y'(t) = \begin{pmatrix} 1 \\ z'(t) \end{pmatrix} &= \begin{pmatrix} 1 \\ G\big(t, z(t)\big) \end{pmatrix} = F\big(y(t)\big), & t &\in (t_0, T), \\ y'(t) &= F\big(y(t)\big), & t &\in (t_0, T). \end{aligned}$$

For theoretical considerations, it thus suffices to study autonomous systems. Moreover, the linear transformation $t \leftrightarrow t - t_0$ permits to choose $t_0 = 0$ as initial time.

Linear ordinary differential equations. The solution to a system of linear ordinary differential equations (defining matrix $A \in \mathbb{R}^{d \times d}$, initial value $y_0 \in \mathbb{R}^d$, solution $y : [0, \infty) \to \mathbb{R}^d$)

$$\begin{cases} y'(t) = A\, y(t), & t \in (0, \infty), \\ y(0) = y_0, \end{cases}$$

is given by the matrix exponential (use notation $E(t)\, y_0 = E(t, y_0)$ to reveal linear dependence on initial value)

$$y(t) = E(t)\, y_0, \qquad E(t) = \mathrm{e}^{tA} = \sum_{k=0}^{\infty} \tfrac{1}{k!}\, t^k A^k, \qquad t \in [0, \infty).$$

Existence and uniqueness results. We note that for a linear differential equation the existence of a global solution is ensured, in general; employing in addition the prescribed initial condition, implies its uniqueness. For a nonlinear differential equation, however, the existence and uniqueness of a solution can only be ensured locally, in general.

Applications and numerical methods. Systems of nonlinear ordinary differential equations arise in the mathematical modelling of chemical reactions and hence as subproblems in connection with diffusion-advection-reaction equations. In most cases, due to the complexity of the systems, it is not possible to find explicit solution representations by elementary functions. Thus, it is essential to construct numerical methods that capture correctly the quantitative and qualitative solution behaviour.

Connection to Dahlquist test equation. For certain systems of nonlinear ordinary differential equations, it is possible to draw conclusions on the (qualitative) solution behaviour from the study of a linearised system (consider for instance equilibrium $y^* \in \mathbb{R}^d$ and use that $F(y^*) = 0$, denote $A = F'(y^*)$)

$$\begin{aligned} y'(t) &= F\big(y(t)\big), \qquad t \in (0,T], \\ F\big(y(t)\big) &\approx F(y^*) + F'(y^*)\,(y-y^*)(t) = A\,(y-y^*)(t), \\ (y-y^*)'(t) &\approx A\,(y-y^*)(t), \qquad t \in (0,\infty). \end{aligned}$$

Provided that the defining matrix is diagonalisable, by the eigenvalue decomposition

$$A = V\,\Lambda\,V^{-1}, \qquad \Lambda = \operatorname{diag}(\lambda_1,\dots,\lambda_d),$$

the linearised system reduces to decoupled scalar differential equations (for simplicity denote $y - y^* \leftrightarrow y$, set $z = V^{-1}y$)

$$\begin{aligned} y'(t) &= A\,y(t) = V\,\Lambda\,V^{-1}y(t), \qquad t \in (0,\infty), \\ \big(V^{-1}y\big)'(t) &= \Lambda\big(V^{-1}y\big)(t), \qquad t \in (0,\infty), \\ z'(t) &= \Lambda\,z(t), \qquad t \in (0,\infty), \\ z_k'(t) &= \lambda_k\,z_k(t), \qquad t \in (0,\infty), \qquad k \in \{1,\dots,d\}. \end{aligned}$$

2.3 Linear advection equations

Linear advection equations. We consider the homogeneous linear advection equation in a single space dimension (unbounded space domain, speed of propagation $0 \neq c \in \mathbb{R}$, regular initial state $u_0 \in \mathscr{C}^1(\mathbb{R})$)

$$\begin{cases} \partial_t u(x,t) = c\,\partial_x u(x,t)\,, & (x,t) \in \mathbb{R} \times (0,\infty)\,, \\ u(x,0) = u_0(x)\,, & x \in \mathbb{R}\,. \end{cases}$$

The solution $u : \mathbb{R} \times [0,\infty) \to \mathbb{R} : (x,t) \mapsto u(x,t)$ is given by the relation

$$u(x,t) = u_0(x + c\,t)\,, \qquad (x,t) \in \mathbb{R} \times [0,\infty)\,.$$

If $c > 0$, the prescribed initial profile is shifted to the left with constant velocity; if $c < 0$, it is shifted to the right.
Verification. Straightforward differentiation (chain rule)

$$\partial_x u(x,t) = u_0'(x + c\,t)\,, \quad \partial_t u(x,t) = c\,u_0'(x + c\,t)\,, \qquad (x,t) \in \mathbb{R} \times (0,\infty)\,,$$

verifies the above solution representation. ◇

Method of characteristics. The method of characteristics is a valuable mean for the analytical and numerical treatment of conservation laws. With regard to the above solution representation, the basic idea is to connect all points which yield the same value

$$\big\{(x,t) \in \mathbb{R} \times [0,\infty) : u(x,t) = \text{constant}\big\}\,.$$

In order to illustrate the general approach, we consider the simplest case of the one-dimensional homogeneous linear advection equation.

(i) *Approach.* We replace the space variable by a time-dependent scalar function and fix a starting value $x_0 \in \mathbb{R}$

$$\xi : [0,\infty) \longrightarrow \mathbb{R} : t \longmapsto \xi(t)\,, \qquad \xi(0) = x_0\,.$$

Moreover, we require that the time-dependent function

$$v : [0,\infty) \longrightarrow \mathbb{R} : t \longmapsto v(t) = u\big(\xi(t), t\big)$$

is constant in time; that is, there holds

$$\begin{aligned} v'(t) &= 0\,, \qquad t \in (0,\infty)\,, \\ v(t) &= v(0)\,, \qquad t \in [0,\infty)\,. \end{aligned}$$

(ii) *Initial value problem.* Differentiating and employing the partial differential equation yields the condition (partial derivatives with respect to first and second arguments are denoted by $\partial_x u = \partial_1 u$ as well as $\partial_t u = \partial_2 u$, exclude trivial case $\partial_x u(\xi(t), t) = 0$)

$$0 = v'(t) = \partial_x u\big(\xi(t), t\big)\, \xi'(t) + \partial_t u\big(\xi(t), t\big) = \partial_x u\big(\xi(t), t\big)\, \big(\xi'(t) + c\big)\,,$$
$$\xi'(t) = -\,c\,.$$

This leads to the initial value problem

$$\begin{cases} \xi'(t) = -\,c\,, & t \in (0,\infty)\,, \\ \xi(0) = x_0\,, \end{cases}$$

with solution given by

$$\xi(t) = x_0 - c\,t\,, \qquad t \in [0,\infty)\,.$$

(iii) *Solution representation.* Altogether, this implies (set $x = x_0 - c\,t$ and $x_0 = x + c\,t$)

$$u(x_0 - c\,t, t) = u\big(\xi(t), t\big) = v(t) = v(0) = u(x_0, 0) = u_0(x_0)\,,$$
$$u(x, t) = u_0(x + c\,t)\,, \qquad (x, t) \in \mathbb{R} \times [0,\infty)\,.$$

Weak solutions. In view of relevant applications such as the propagation of shock waves, it is desirable to consider functions that are less regular, i.e., discontinuous in single points. Strictly speaking, if the initial state $u_0 : \mathbb{R} \to \mathbb{R}$ is *not* differentiable, the associated function

$$u(x, t) = u_0(x + c\,t)\,, \qquad (x, t) \in \mathbb{R} \times [0,\infty)\,,$$

does not satisfy the advection equation

$$\begin{cases} \partial_t u(x, t) = c\,\partial_x u(x, t)\,, & (x, t) \in \mathbb{R} \times (0,\infty)\,, \\ u(x, 0) = u_0(x)\,, & x \in \mathbb{R}\,. \end{cases}$$

However, in order to retain the compact formulation by a partial differential equation, it is common practice to call such a function a *solution in a weak sense.*

(i) *Integral solution.* Using as a starting point the linear advection equation, multiplication by a so-called test function leads to a related integral equation (assume that the function $v : \mathbb{R} \times [0,\infty) \to \mathbb{R}$ is sufficiently often differentiable and has compact support, thus the relation $v(x, t) \to 0$ holds for $x \to \pm\infty$ as well as $t \to \infty$, integration with respect to space variable $x \in \mathbb{R}$ and time variable $t \in [0,\infty)$, integration-by-parts, assumptions on test

functions ensure that certain boundary terms vanish, employ initial condition)

$$\begin{aligned}
&\partial_t u(x,t) - c\,\partial_x u(x,t) = 0\,,\\
&\int_0^\infty \int_{\mathbb{R}} \big(\partial_t u(x,t) - c\,\partial_x u(x,t)\big)\, v(x,t)\,\mathrm{d}x\,\mathrm{d}t = 0\,,\\
&-\int_0^\infty \int_{\mathbb{R}} u(x,t)\,\big(\partial_t v(x,t) - c\,\partial_x v(x,t)\big)\,\mathrm{d}x\,\mathrm{d}t\\
&\qquad + \int_{\mathbb{R}} \big(u(x,t)\,v(x,t)\big)\Big|_{t=0}^{\infty}\mathrm{d}x - \int_0^\infty \big(c\,u(x,t)\,v(x,t)\big)\Big|_{x=-\infty}^{\infty}\mathrm{d}t = 0\,,\\
&-\int_0^\infty \int_{\mathbb{R}} u(x,t)\,\big(\partial_t v(x,t) - c\,\partial_x v(x,t)\big)\,\mathrm{d}x\,\mathrm{d}t - \int_{\mathbb{R}} u_0(x)\,v(x,0)\,\mathrm{d}x = 0\,.
\end{aligned}$$

A function $u \in L^\infty(\mathbb{R}\times(0,\infty))$ that satisfies the following integral equation for all test functions $v : \mathbb{R}\times[0,\infty) \to \mathbb{R}$ is called an integral solution of the advection equation

$$\int_0^\infty \int_{\mathbb{R}} u(x,t)\,\big(\partial_t v(x,t) - c\,\partial_x v(x,t)\big)\,\mathrm{d}x\,\mathrm{d}t + \int_{\mathbb{R}} u_0(x)\,v(x,0)\,\mathrm{d}x = 0\,.$$

(ii) *Heaviside function.* As initial state, we consider the Heaviside function with a discontinuity at the origin

$$H : \mathbb{R} \longrightarrow \mathbb{R} : x \longmapsto H(x) = \begin{cases} 1\,, & x \geq 0\,,\\ 0\,, & x < 0\,, \end{cases}$$

The related function (initial profile is shifted to the left or right with constant velocity)

$$u(x,t) = u_0(x + c\,t) = H(x + c\,t) = \begin{cases} 1\,, & x + c\,t \geq 0\,,\\ 0\,, & x + c\,t < 0\,, \end{cases} \qquad (x,t) \in \mathbb{R}\times[0,\infty)\,,$$

is an integral solution of the linear advection equation.
Explanation. Assume for instance $c > 0$. By the definition of the Heaviside function, we obtain (use that $x + c\,t \geq 0$ is equivalent to $t \geq -\frac{x}{c}$ or $x \geq -c\,t$, respectively, note that $t \geq 0$)

$$\begin{aligned}
&\int_0^\infty \int_{\mathbb{R}} u(x,t)\,\big(\partial_t v(x,t) - c\,\partial_x v(x,t)\big)\,\mathrm{d}x\,\mathrm{d}t + \int_{\mathbb{R}} u_0(x)\,v(x,0)\,\mathrm{d}x\\
&\quad = \int_{\mathbb{R}}\int_0^\infty H(x + c\,t)\,\partial_t v(x,t)\,\mathrm{d}t\,\mathrm{d}x - c\int_0^\infty\int_{\mathbb{R}} H(x + c\,t)\,\partial_x v(x,t)\,\mathrm{d}x\,\mathrm{d}t\\
&\qquad + \int_{\mathbb{R}} H(x)\,v(x,0)\,\mathrm{d}x\\
&\quad = \int_{\mathbb{R}}\int_{\max\{0,-\frac{x}{c}\}}^\infty \partial_t v(x,t)\,\mathrm{d}t\,\mathrm{d}x - c\int_0^\infty\int_{-c\,t}^\infty \partial_x v(x,t)\,\mathrm{d}x\,\mathrm{d}t + \int_0^\infty v(x,0)\,\mathrm{d}x\\
&\quad = \int_{\mathbb{R}} v(x,t)\Big|_{t=\max\{0,-\frac{x}{c}\}}^\infty \mathrm{d}x - c\int_0^\infty v(x,t)\Big|_{x=-c\,t}^\infty \mathrm{d}t + \int_0^\infty v(x,0)\,\mathrm{d}x\\
&\quad = -\int_{\mathbb{R}} v\Big(x,\max\{0,-\tfrac{x}{c}\}\Big)\,\mathrm{d}x + c\int_0^\infty v(-c\,t,t)\,\mathrm{d}t + \int_0^\infty v(x,0)\,\mathrm{d}x\\
&\quad = -\int_{-\infty}^0 v\big(x,-\tfrac{x}{c}\big)\,\mathrm{d}x - \int_0^\infty v(x,0)\,\mathrm{d}x + c\int_0^\infty v(-c\,t,t)\,\mathrm{d}t + \int_0^\infty v(x,0)\,\mathrm{d}x\,;
\end{aligned}$$

moreover, the substitution $\xi = -\frac{x}{c}$ or $x = -c\,\xi$, respectively, implies

$$\begin{aligned}
&\int_0^\infty \int_{\mathbb{R}} u(x,t)\,\big(\partial_t v(x,t) - c\,\partial_x v(x,t)\big)\,\mathrm{d}x\,\mathrm{d}t + \int_{\mathbb{R}} u_0(x)\,v(x,0)\,\mathrm{d}x \\
&\quad = -c\int_0^\infty v(-c\,\xi,\xi)\,\mathrm{d}\xi - \int_0^\infty v(x,0)\,\mathrm{d}x + c\int_0^\infty v(-c\,t,t)\,\mathrm{d}t + \int_0^\infty v(x,0)\,\mathrm{d}x \\
&\quad = 0\,.
\end{aligned}$$

Analogous arguments hold for the case $c < 0$. ◇

Connection to Dahlquist test equation. In order to justify the following considerations, we need to assume that the initial state is localised; thus, for a certain period of time, we may also suppose that the associated solution to the linear advection equation is localised. In this situation, it is reasonable to restrict the unbounded spatial domain to a sufficiently large bounded interval and to employ a Fourier series expansion of the solution (recall that $\mathscr{F}_m(x) = \frac{1}{\sqrt{b-a}}\,\mathrm{e}^{\mathrm{i}\,\mu_m\,(x-a)}$ with $\mu_m = \frac{2\pi}{b-a}\,m$ for any $m \in \mathbb{Z}$)

$$u(x,t) = \sum_{m\in\mathbb{Z}} u_m(t)\,\mathscr{F}_m(x)\,, \qquad (x,t) \in (a,b)\times[0,\infty)\,.$$

Inserting this representation into the linear advection equation implies (for all (x,t), use that $\partial_x \mathscr{F}_m(x) = \mathrm{i}\,\mu_m\,\mathscr{F}_m(x)$, employ orthogonality of Fourier basis functions)

$$\begin{aligned}
\partial_t u(x,t) &= \sum_{m\in\mathbb{Z}} u_m'(t)\,\mathscr{F}_m(x)\,, \\
\partial_x u(x,t) &= \sum_{m\in\mathbb{Z}} \mathrm{i}\,\mu_m\,u_m(t)\,\mathscr{F}_m(x)\,, \\
0 = \partial_t u(x,t) - c\,\partial_x u(x,t) &= \sum_{m\in\mathbb{Z}} \big(u_m'(t) - \mathrm{i}\,c\,\mu_m\,u_m(t)\big)\,\mathscr{F}_m(x)\,, \\
u_m'(t) &= \mathrm{i}\,c\,\mu_m\,u_m(t)\,, \qquad m\in\mathbb{Z}\,.
\end{aligned}$$

This explains the significance of the scalar test equation (here $\mu = c\,\mu_m \in \mathbb{R}$)

$$y'(t) = \mathrm{i}\,\mu\,y(t)\,, \qquad t\in(0,\infty)\,, \qquad \mu\in\mathbb{R}\,,$$

in the context of advection equations.

Higher space dimensions. It is straightforward to extend the above considerations for the one-dimensional linear advection equation to higher space dimensions. Indeed, the solution $u : \mathbb{R}^d \times [0,\infty) \to \mathbb{R} : (x,t) \mapsto u(x,t)$ to the homogeneous linear advection equation with constant coefficients (where $0 \neq c \in \mathbb{R}^d$)

$$\begin{cases} \partial_t u(x,t) = \operatorname{div}\big(c\,u(x,t)\big)\,, & (x,t)\in\mathbb{R}^d\times(0,\infty)\,, \\ u(x,0) = u_0(x)\,, & x\in\mathbb{R}^d\,, \end{cases}$$

is given by the relation

$$u(x,t) = u_0(x + c\,t)\,, \qquad (x,t) \in \mathbb{R}^d \times [0,\infty)\,,$$

as verified by differentiation ($u_0 : \mathbb{R}^d \to \mathbb{R}$, chain rule)

$$\begin{aligned}
u_0(x + c\,t) &= u_0(x_1 + c_1\,t, \ldots, x_d + c_d\,t)\,,\\
\partial_t u_0(x + c\,t) &= c_1\,\partial_1 u_0(x + c\,t) + \cdots + c_d\,\partial_d u_0(x + c\,t)\,,\\
c\,u(x,t) = c\,u_0(x + c\,t) &= \begin{pmatrix} c_1\,u_0(x + c\,t) \\ \vdots \\ c_d\,u_0(x + c\,t) \end{pmatrix},\\
\operatorname{div}\bigl(c\,u(x,t)\bigr) &= c_1\,\partial_1 u_0(x + c\,t) + \cdots + c_d\,\partial_d u_0(x + c\,t)\,.
\end{aligned}$$

For theoretical purposes, in order to reveal similarities to ordinary differential equations, it is convenient to formulate the partial differential equation as evolution equation and to introduce the associated linear evolution operator (where $U(t) = u(\cdot,t) : \mathbb{R}^d \to \mathbb{R}$ and $v : \mathbb{R}^d \to \mathbb{R}$)

$$\begin{aligned}
A &= \operatorname{div}\bigl(c\,(\cdot)\bigr)\,,\\
&\begin{cases} U'(t) = A\,U(t)\,, & t \in (0,\infty)\,,\\ U(0) = U_0\,, \end{cases}\\
\bigl(\mathscr{E}(t)\,v\bigr)(x) &= v(x + c\,t)\,, \quad (x,t) \in \mathbb{R}^d \times [0,\infty)\,.
\end{aligned}$$

2.4 Nonlinear conservation laws

Nonlinear conservation laws. For simplicity, we restrict ourselves to the one-dimensional case. The conservation law associated with a nonlinear function $f : \mathbb{R} \to \mathbb{R}$ is given by (solution $u : \mathbb{R} \times [0, T] \to \mathbb{R}$, prescribed initial state $u_0 : \mathbb{R} \to \mathbb{R}$)

$$\begin{cases} \partial_t u(x,t) = \partial_x f\big(u(x,t)\big) = f'\big(u(x,t)\big)\,\partial_x u(x,t)\,, & (x,t) \in \mathbb{R} \times (0,T)\,, \\ u(x,0) = u_0(x)\,, & x \in \mathbb{R}\,; \end{cases}$$

the designation is explained by the fact that the integral is a conserved quantity (corresponds to total mass for positive function, interchange derivative and integral, employ partial differential equation, assumption that solution is integrable over whole line implies $u(x,t) = 0$ for $x \to \pm\infty$)

$$\begin{gathered} M(t) = \int_{\mathbb{R}} u(x,t)\,\mathrm{d}x\,, \qquad t \in [0,T]\,, \\ M'(t) = \int_{\mathbb{R}} \partial_t u(x,t)\,\mathrm{d}x = \int_{\mathbb{R}} \partial_x f\big(u(x,t)\big)\,\mathrm{d}x = f\big(u(x,t)\big)\Big|_{x=-\infty}^{\infty} = 0\,, \qquad t \in (0,T)\,, \\ M(t) = M(0) = \int_{\mathbb{R}} u_0(x)\,\mathrm{d}x\,, \qquad t \in [0,T]\,. \end{gathered}$$

Method of characteristics. In special situations, the method of characteristics permits to determine a useful representation of the solution to a nonlinear conservation law.

(i) *Approach.* As before, we replace the space variable by a time-dependent scalar function and fix a starting value $x_0 \in \mathbb{R}$

$$\xi : [0,T] \longrightarrow \mathbb{R} : t \longmapsto \xi(t)\,, \qquad \xi(0) = x_0\,.$$

Moreover, we require that the time-dependent function

$$v : [0,T] \longrightarrow \mathbb{R} : t \longmapsto v(t) = u\big(\xi(t), t\big)$$

is constant in time; that is, there holds

$$\begin{gathered} v'(t) = 0\,, \qquad t \in (0,T)\,, \\ u\big(\xi(t),t\big) = v(t) = v(0) = u(x_0,0) = u_0(x_0)\,, \qquad t \in [0,T]\,. \end{gathered}$$

(ii) *Initial value problem.* Differentiating and employing the partial differential equation yields the condition (again with $\partial_x u = \partial_1 u$ and $\partial_t u = \partial_2 u$, exclude trivial case $\partial_x u(\xi(t),t) = 0$, for $t \in (0,T)$)

$$\begin{gathered} 0 = v'(t) = \partial_x u\big(\xi(t),t\big)\,\xi'(t) + \partial_t u\big(\xi(t),t\big) = \partial_x u\big(\xi(t),t\big)\Big(\xi'(t) + f'\Big(u\big(\xi(t),t\big)\Big)\Big)\,, \\ \xi'(t) = -f'\Big(u\big(\xi(t),t\big)\Big) = -f'\big(u_0(x_0)\big)\,. \end{gathered}$$

This leads to the following initial value problem

$$\begin{cases} \xi'(t) = -f'\big(u_0(x_0)\big), & t \in (0,T), \\ \xi(0) = x_0, \end{cases}$$

with solution given by

$$\xi(t) = x_0 - f'\big(u_0(x_0)\big)\, t, \qquad t \in [0,T].$$

(iii) *Solution representation.* Altogether, this implies the solution representation

$$u\Big(x_0 - f'\big(u_0(x_0)\big)\, t, t\Big) = u\big(\xi(t), t\big) = v(t) = v(0) = u(x_0, 0) = u_0(x_0), \qquad t \in [0,T].$$

Advection equation. For the simplest case of the linear advection equation, where $f(v) = c\,v$ with $c \in \mathbb{R}$ and thus $f'(v) = c$, the formerly stated solution representation follows at once from the above relation (set $x = x_0 - c\,t$ as well as $x_0 = x - c\,t$)

$$\begin{aligned} f'(v) = c: \qquad & u(x_0 - c\,t, t) = u_0(x_0), \qquad t \in [0,\infty), \\ & u(x,t) = u_0(x + c\,t), \qquad t \in [0,\infty). \end{aligned}$$

Burgers' equation. The special choice $f(v) = \frac{1}{2}\,v^2$ and thus $f'(v) = v$ leads to Burgers' equation for $u : \mathbb{R} \times [0,T] \to \mathbb{R} : (x,t) \mapsto u(x,t)$

$$\begin{cases} \partial_t u(x,t) = u(x,t)\,\partial_x u(x,t), & (x,t) \in \mathbb{R} \times (0,T), \\ u(x,0) = u_0(x), \end{cases}$$

with the characteristic formation of shock waves, even for regular initial states; in particular, the notion of a weak solution is needed in this context. The method of characteristics yields the relation

$$f'(v) = v: \qquad u\big(x_0 - u_0(x_0)\, t, t\big) = u_0(x_0), \qquad t \in [0,T];$$

as a consequence, the following implicit solution representation is obtained (set $x = x_0 - u_0(x_0)\,t$ as well as $x_0 = x + u_0(x_0)\,t$, note that $u(x,t) = u_0(x_0)$ implies $u(x,t) = u_0(x_0) = u_0(x + u_0(x_0)\,t) = u_0(x + u(x,t)\,t)$)

$$f'(v) = v: \qquad u(x,t) = u_0\big(x + u(x,t)\, t\big), \qquad t \in [0,T].$$

2.5 Linear diffusion equations

Linear diffusion equations. We consider the homogeneous linear diffusion (or heat) equation (consider unbounded domain $\Omega = \mathbb{R}^d$ or bounded open domain $\Omega \subset \mathbb{R}^d$, positive coefficients $c_1, \dots c_d > 0$, initial state $u_0 : \overline{\Omega} \to \mathbb{R}$)

$$\begin{cases} \partial_t u(x,t) = c_1 \, \partial_{x_1 x_1} u(x,t) + \cdots + c_d \, \partial_{x_d x_d} u(x,t) \,, & (x,t) \in \Omega \times (0,\infty) \,, \\ u(x,0) = u_0(x) \,, & x \in \overline{\Omega} \,. \end{cases}$$

By a scaling of the spatial variables, we obtain (Laplace operator $\Delta = \partial_{x_1 x_1} + \cdots + \partial_{x_d x_d}$, new domain due to transformation)

$$\begin{cases} \partial_t u(x,t) = \Delta u(x,t) \,, & (x,t) \in \Omega \times (0,\infty) \,, \\ u(x,0) = u_0(x) \,, & x \in \overline{\Omega} \,. \end{cases}$$

Again, for theoretical purposes, it is useful to rewrite the partial differential equation as an abstract ordinary equation and to introduce the associated evolution operator (commonly, we do not distinguish between $u(t)$ and $u(\cdot, t)$ in notation)

$$A = \Delta \,,$$

$$\begin{cases} u'(t) = A \, u(t) \,, & t \in (0,\infty) \,, \\ u(0) = u_0 \,, \end{cases}$$

$$u(t) = \mathscr{E}(t) \, u_0 \,, \quad t \in [0,\infty) \,.$$

Boundary conditions. In order to attain uniqueness of solutions, additional boundary conditions have to be imposed. In view of practial applications, it is of relevance to consider boundary conditions of Dirichlet or Neumann type, respectively. Homogeneous Dirichlet conditions require the solution to vanish at the boundary (bounded domain $\Omega \subset \mathbb{R}$, cooling to zero temperature at boundary, exponential decay of L^2-norm)

$$u(x,t)\Big|_{x \in \partial\Omega} = 0 \,, \qquad t \in [0,\infty) \,;$$

homogeneous Neumann conditions require the normal derivative of the solution to vanish at the boundary (no flux over boundary, isolated system)

$$\partial_\nu u(x,t)\Big|_{x \in \partial\Omega} = 0 \,, \qquad t \in [0,\infty) \,.$$

Continuation to odd and even functions. In situations, where the underlying domain is given as the cartesian product of bounded intervals, it is often beneficial to use Fourier series expansions (see below); the consideration of periodic boundary conditions, however, is often unnatural. Employing continuations of the solutions to odd functions permits to realise homogeneous Dirichlet boundary conditions, at the cost of a significantly enlarged space domain; as explicated before, the series expansions involve sine basis functions. A similar approach based on continuations to even functions is used for homogeneous Neumann boundary conditions.

Connection to Dahlquist test equation. For simplicity, we again restrict ourselves to the consideration of the homogeneous linear diffusion equation in a single space dimension (unbounded space domain, with $c > 0$, initial state $u_0 : \mathbb{R} \to \mathbb{R}$)

$$\begin{cases} \partial_t u(x,t) = c\,\partial_{xx} u(x,t)\,, & (x,t) \in \mathbb{R} \times (0,\infty)\,, \\ u(x,0) = u_0(x)\,, & x \in \mathbb{R}\,. \end{cases}$$

As for the homogeneous linear advection equation, we can justify a Fourier series representation for a localised solution (restriction of unbounded spatial domain to sufficiently large bounded interval, recall that by definition $\mathscr{F}_m(x) = \frac{1}{\sqrt{b-a}}\,\mathrm{e}^{\mathrm{i}\mu_m(x-a)}$ with $\mu_m = \frac{2\pi}{b-a}\,m$ for any $m \in \mathbb{Z}$)

$$u(x,t) = \sum_{m\in\mathbb{Z}} u_m(t)\,\mathscr{F}_m(x)\,, \qquad (x,t) \in [a,b] \times [0,\infty)\,.$$

Inserting this series expansion into the linear advection equation implies (relation valid for all $(x,t) \in (a,b) \times (0,\infty)$, use that $\partial_{xx}\mathscr{F}_m(x) = -\mu_m^2\,\mathscr{F}_m(x)$, employ orthogonality of Fourier basis functions)

$$\begin{aligned} \partial_t u(x,t) &= \sum_{m\in\mathbb{Z}} u_m'(t)\,\mathscr{F}_m(x)\,, \\ \partial_{xx} u(x,t) &= -\sum_{m\in\mathbb{Z}} \mu_m^2\,u_m(t)\,\mathscr{F}_m(x)\,, \\ 0 = \partial_t u(x,t) - c\,\partial_{xx} u(x,t) &= \sum_{m\in\mathbb{Z}} \big(u_m'(t) + c\,\mu_m^2\,u_m(t)\big)\,\mathscr{F}_m(x)\,, \\ u_m'(t) &= -\,c\,\mu_m^2\,u_m(t)\,, \qquad m \in \mathbb{Z}\,. \end{aligned}$$

This explains the significance of the scalar test equation (here $\lambda = c\,\mu_m^2 > 0$)

$$y'(t) = -\,\lambda\,y(t)\,, \qquad t \in (0,\infty)\,, \qquad \lambda > 0\,,$$

in the context of diffusion equations.

Solution representation by Fourier series expansion. The above considerations imply the solution representation (use that $u_m(t) = \mathrm{e}^{-c\,\mu_m^2\,t}\,u_m(0)$ and $u_m(0) = (u_0|\mathscr{F}_m)_{L^2}$ for $m \in \mathbb{Z}$)

$$u(x,t) = \big(\mathscr{E}(t)\,u_0\big)(x) = \sum_{m\in\mathbb{Z}} \mathrm{e}^{-c\,\mu_m^2\,t}\,\big(u_0\big|\mathscr{F}_m\big)_{L^2}\,\mathscr{F}_m(x)\,, \qquad (x,t) \in [a,b] \times [0,\infty)\,.$$

An application of Parseval's identity shows that the L^2-norm of the solution is non-growing (recall $\mu_m = \frac{2\pi}{b-a}\,m$ for $m \in \mathbb{Z}$, note that $\mu_0 = 0$, use bound $\mathrm{e}^{-2\,c\,\mu_m^2\,t} \le 1$ for all $t \in [0,\infty)$)

$$\big\|u(\cdot,t)\big\|_{L^2}^2 = \sum_{m\in\mathbb{Z}} \mathrm{e}^{-2\,c\,\mu_m^2\,t}\,\big|u_m(0)\big|^2 \le \sum_{m\in\mathbb{Z}} \big|u_m(0)\big|^2 = \big\|u_0\big\|_{L^2}^2\,, \qquad t \in [0,\infty)\,.$$

Numerical realisation. The numerical realisation of the above stated solution representation relies on the following steps, see Section 1.3.

(i) Computation of approximations to spectral coefficients (implementation based on FFT)

$$u_m^{(s)} = \big(u_0 \big| \mathscr{F}_m\big)_{L^2}\,, \qquad m \in \big\{-\tfrac{M}{2},\dots,\tfrac{M}{2}-1\big\}\,.$$

(ii) Computation of coefficients (implementation based on pointwise multiplication)

$$c_m = \mathrm{e}^{-c\,\mu_m^2\,t}\,u_m^{(s)}\,, \qquad m \in \big\{-\tfrac{M}{2},\dots,\tfrac{M}{2}-1\big\}\,.$$

(iii) Computation of approximations to solution values at grid points (implementation based on IFFT)

$$\sum_{m=-\frac{M}{2}}^{\frac{M}{2}-1} c_m\,\mathscr{F}_m(x_k) \;\approx\; u(x_k,t)\,, \qquad k \in \{0,\dots,M-1\}\,.$$

Solution representation by Sine series expansion. In connection with homogeneous Dirichlet boundary conditions, it is natural to employ a solution representation by sine basis functions to obtain

$$u_m(0) = \big(u_0 \big| \mathscr{S}_m\big)_{L^2}\,, \qquad m \in \mathbb{N},$$
$$u(x,t) = \sum_{m\in\mathbb{Z}} u_m(t)\,\mathscr{S}_m(x) = \sum_{m\in\mathbb{N}} \mathrm{e}^{-c\,\mu_m^2\,t}\,u_m(0)\,\mathscr{S}_m(x)\,, \qquad (x,t) \in [a,b]\times[0,\infty)\,.$$

The application of Parseval's identity implies exponential decay of the L^2-norm (regularising effect, recall $\mu_m = \frac{2\pi}{b-a}\,m$ for $m \in \mathbb{N}$, note that minimal value is obtained for $\mu_1 \neq 0$, use estimate $\mathrm{e}^{-2c\mu_m^2 t} \leq \mathrm{e}^{-2c\mu_1^2 t}$ for all $t \in [0,\infty)$)

$$\big\|u(\cdot,t)\big\|_{L^2} = \mathrm{e}^{-c\,\mu_1^2\,t}\,\big\|u_0\big\|_{L^2}^2\,, \qquad t \in [0,\infty)\,.$$

Chapter 3

Basic discretisation methods

Contents. In this chapter, in order to suggest certain issues that are treated in more detail later on, we introduce basic space and time discretisation methods for diffusion-advection-reaction equations. We consider a simple one-dimensional model equation comprising a linear diffusion term, a linear advection term, and a nonlinear reaction term. For the space discretisation, we consider an elementary approach based on central finite difference approximations for uniform meshes. The resulting system of nonlinear ordinary differential equations is solved numerically by the explicit and implicit Euler methods, applied with constant time stepsizes. As alternative time integration methods, we mention the semi-implicit Euler method and the Lie–Trotter splitting method.

3.1 Model equation

Initial-boundary value problem. As model equation, we consider a one-dimensional diffusion-advection-reaction equation subject to homogeneous Dirichlet boundary conditions and an initial condition (bounded space interval $[a,b] \subset \mathbb{R}$, final time $T > 0$, parameters $\alpha, \beta, \gamma \in \mathbb{R}$ with $\alpha > 0$, initial state $u_0 : [a,b] \to \mathbb{R}$, nonlinearity $g : \mathbb{R} \to \mathbb{R}$, solution $u : [a,b] \times [0,T] \to \mathbb{R}$)

$$\begin{cases} \partial_t u(x,t) = \alpha\, \partial_{xx} u(x,t) + \beta\, \partial_x u(x,t) + \gamma\, g\big(u(x,t)\big)\,, & (x,t) \in (a,b) \times (0,T)\,, \\ u(a,t) = 0 = u(b,t)\,, & t \in [0,T]\,, \\ u(x,0) = u_0(x)\,, & x \in [a,b]\,. \end{cases}$$

For consistency, we require the initial state to satisfy homogenous Dirichlet conditions.

Evolution operator. Formally, the solution is given by the relation

$$u(x,t) = \big(\mathscr{E}(t,u_0)\big)(x)\,, \qquad (x,t) \in [a,b] \times [0,T]\,.$$

Objective. Our objective is the introduction of a discrete evolution operator yielding approximations to the exact solution at certain space and time grid points

$$\mathscr{S} \approx \mathscr{E}\,.$$

Special cases. Evidently, when setting $\beta = 0 = \gamma$, we obtain a linear diffusion equation; for $\alpha = 0 = \gamma$, a linear advection equation results. The special case $\alpha = 0 = \beta$ corresponds to a nonlinear ordinary differential equation.

3.2 Space discretisation methods

Uniform mesh. We consider a uniform spatial mesh (integer number $M \in \mathbb{N}$ defines grid width $h > 0$ and associated equidistant grid points, evidently $x_0 = a$ as well as $x_{M+1} = b$)

$$h = \tfrac{b-a}{M+1}\,, \qquad x_m = a + mh\,, \qquad m \in \{0, \ldots, M+1\}\,.$$

Approach. For convenience, we restate the initial-boundary value problem, evaluated at the space grid points; with regard to the subsequent considerations, we employ the compact formulation

$$\begin{cases} \begin{pmatrix} \partial_t u(x_1,t) \\ \vdots \\ \partial_t u(x_M,t) \end{pmatrix} = \alpha \begin{pmatrix} \partial_{xx} u(x_1,t) \\ \vdots \\ \partial_{xx} u(x_M,t) \end{pmatrix} + \beta \begin{pmatrix} \partial_x u(x_1,t) \\ \vdots \\ \partial_x u(x_M,t) \end{pmatrix} + \gamma \begin{pmatrix} g\big(u(x_1,t)\big) \\ \vdots \\ g\big(u(x_m,t)\big) \end{pmatrix}, \qquad t \in (0,T)\,, \\ u(x_0,t) = 0 = u(x_{M+1},t)\,, \qquad t \in [0,T]\,, \\ \begin{pmatrix} u(x_1,0) \\ \vdots \\ u(x_M,0) \end{pmatrix} = \begin{pmatrix} u_0(x_1) \\ \vdots \\ u_0(x_M) \end{pmatrix}; \end{cases}$$

we recall that by assumption the initial state satisfies the conditions $u_0(x_0) = 0 = u_0(x_{M+1})$. A common approach for the space discretisation of such a problem is to replace the spatial derivatives by finite difference approximations; this yields a system of nonlinear ordinary differential equations for a vector-valued function comprising approximations to the solution values at the grid points

$$\begin{pmatrix} U_1 \\ \vdots \\ U_M \end{pmatrix} : [0,T] \longrightarrow \mathbb{R}^M : t \longmapsto \begin{pmatrix} U_1(t) \\ \vdots \\ U_M(t) \end{pmatrix} \approx \begin{pmatrix} u(x_1,t) \\ \vdots \\ u(x_M,t) \end{pmatrix}.$$

Central finite differences. At the interior grid points $x_2, \ldots, x_{M-1}$, we employ the central finite difference approximations

$$\begin{aligned} \partial_x u(x_m,t) &\approx \tfrac{1}{2h}\big(u(x_m+h,t) - u(x_m-h,t)\big) \\ &= \tfrac{1}{2h}\big(u(x_{m+1},t) - u(x_{m-1},t)\big)\,, \qquad m \in \{2,\ldots,M-1\}\,, \\ \partial_{xx} u(x_m,t) &\approx \tfrac{1}{h^2}\big(u(x_m+h,t) - 2\,u(x_m,t) + u(x_m-h,t)\big) \\ &= \tfrac{1}{h^2}\big(u(x_{m+1},t) - 2\,u(x_m,t) + u(x_{m-1},t)\big)\,, \qquad m \in \{2,\ldots,M-1\}\,, \end{aligned}$$

see Section 1.4; in accordance with the conditions $u(x_0,t) = 0$ and $u(x_{M+1},t) = 0$, we set

$$\begin{aligned} \partial_x u(x_1,t) &\approx \tfrac{1}{2h}\big(u(x_2,t) - u(x_0,t)\big) = \tfrac{1}{2h}\,u(x_2,t)\,, \\ \partial_x u(x_M,t) &\approx \tfrac{1}{2h}\big(u(x_{M+1},t) - u(x_{M-1},t)\big) = -\tfrac{1}{2h}\,u(x_{M-1},t)\,, \\ \partial_{xx} u(x_1,t) &\approx \tfrac{1}{h^2}\big(u(x_2,t) - 2\,u(x_1,t) + u(x_0,t)\big) = \tfrac{1}{h^2}\big(u(x_2,t) - 2\,u(x_1,t)\big)\,, \\ \partial_{xx} u(x_M,t) &\approx \tfrac{1}{h^2}\big(u(x_{M+1},t) - 2\,u(x_M,t) + u(x_{M-1},t)\big) = \tfrac{1}{h^2}\big(-2\,u(x_M,t) + u(x_{M-1},t)\big)\,. \end{aligned}$$

In compact matrix-vector notation, we thus obtain

$$\begin{pmatrix} \partial_x u(x_1,t) \\ \vdots \\ \partial_x u(x_m,t) \\ \vdots \\ \partial_x u(x_M,t) \end{pmatrix} \approx \tfrac{1}{2h} \begin{pmatrix} 0 & 1 & & & \\ -1 & 0 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 0 & 1 \\ & & & -1 & 0 \end{pmatrix} \begin{pmatrix} u(x_1,t) \\ \vdots \\ u(x_m,t) \\ \vdots \\ u(x_M,t) \end{pmatrix},$$

$$\begin{pmatrix} \partial_{xx} u(x_1,t) \\ \vdots \\ \partial_{xx} u(x_m,t) \\ \vdots \\ \partial_{xx} u(x_M,t) \end{pmatrix} \approx \tfrac{1}{h^2} \begin{pmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{pmatrix} \begin{pmatrix} u(x_1,t) \\ \vdots \\ u(x_m,t) \\ \vdots \\ u(x_M,t) \end{pmatrix}.$$

Initial value problem. Altogether, employing the abbreviations (initial value $U_0 \in \mathbb{R}^M$, defining matrices $A, B \in \mathbb{R}^{M\times M}$ and nonlinearity $g : \mathbb{R} \to \mathbb{R}$, solution $U : [0,T] \to \mathbb{R}^M$)

$$U(t) = \begin{pmatrix} U_1(t) \\ \vdots \\ U_M(t) \end{pmatrix}, \qquad U_0 = \begin{pmatrix} u_0(x_1) \\ \vdots \\ u_0(x_M) \end{pmatrix}, \qquad G\big(U(t)\big) = \begin{pmatrix} g\big(U_1(t)\big) \\ \vdots \\ g\big(U_M(t)\big) \end{pmatrix}, \qquad t \in [0,T],$$

$$A = \tfrac{1}{h^2} \begin{pmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{pmatrix}, \qquad B = \tfrac{1}{2h} \begin{pmatrix} 0 & 1 & & & \\ -1 & 0 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 0 & 1 \\ & & & -1 & 0 \end{pmatrix},$$

we obtain the following initial value problem

$$\begin{cases} U'(t) = \alpha\, A U(t) + \beta\, B U(t) + \gamma\, G\big(U(t)\big), & t \in (0,T), \\ U(0) = U_0. \end{cases}$$

We point out that the imposed homogeneous Dirichlet boundary conditions are included in the matrices A, B.

Connection to polynomial approximations. We note that central finite differences correspond to approximations by polynomials.

(i) For instance, replacing the values of a function $f : [a,b] \to \mathbb{R}$ on the subinterval $[x_m, x_{m+1}]$ by a poynomial of degree two

$$p : \mathbb{R} \longrightarrow \mathbb{R} : x \longmapsto p(x) = f(x_m) + f'(x_m)\,(x - x_m) + \tfrac{1}{2} f''(x_m)\,(x - x_m)^2 \approx f(x),$$

at once implies (since $p'(x) = f'(x_m) + f''(x_m)\,(x - x_m)$, evident from Taylor series expansion of function)

$$p(x_m) = f(x_m)\,, \qquad p'(x_m) = f'(x_m)\,, \qquad p''(x) = f''(x_m)\,.$$

Evaluation at $x_m + h$ and $x_m - h$, respectively, yields

$$\begin{aligned} p(x_m + h) &= f(x_m) + f'(x_m)\,h + \tfrac{1}{2}\,f''(x_m)\,h^2\,, \\ p(x_m - h) &= f(x_m) - f'(x_m)\,h + \tfrac{1}{2}\,f''(x_m)\,h^2\,; \end{aligned}$$

as a consequence, we obtain the relation

$$\tfrac{1}{2h}\big(p(x_m + h) - p(x_m - h)\big) = f'(x_m)\,.$$

(ii) Considering instead the approximation by a poynomial of degree three

$$\begin{aligned} p : \mathbb{R} \longrightarrow \mathbb{R} : x \longmapsto p(x) &= f(x_m) + f'(x_m)\,(x - x_m) + \tfrac{1}{2}\,f''(x_m)\,(x - x_m)^2 \\ &\quad + \tfrac{1}{6}\,f'''(x_m)\,(x - x_m)^3 \\ &\approx f(x)\,, \end{aligned}$$

implies the identities

$$\begin{aligned} p(x_m + h) &= f(x_m) + f'(x_m)\,h + \tfrac{1}{2}\,f''(x_m)\,h^2 + \tfrac{1}{6}\,f'''(x_m)\,h^3\,, \\ p(x_m - h) &= f(x_m) - f'(x_m)\,h + \tfrac{1}{2}\,f''(x_m)\,h^2 - \tfrac{1}{6}\,f'''(x_m)\,h^3\,, \end{aligned}$$

which further show

$$\tfrac{1}{h^2}\big(p(x_m + h) - 2\,p(x_m) + p(x_m - h)\big) = f''(x_m)\,.$$

Alternative approaches. Alternative space discretisation methods such as spectral methods and the finite element method rely on a representation of the solution by a linear combination of certain *basis functions* (infinite index set $\mathscr{M} \subseteq \mathbb{Z}$, real or complex coefficient functions $u_m : [0, T] \to \mathbb{C}$, essential that family $(\mathscr{B}_m)_{m \in \mathscr{M}}$ generates considered function space, linear independence ensures uniqueness of coefficient functions)

$$u(x, t) = \sum_{m \in \mathscr{M}} u_m(t)\,\mathscr{B}_m(x)\,, \qquad (x, t) \in [a, b] \times [0, T]\,.$$

Under the presumption that the spatial derivatives $\partial_x \mathscr{B}_m, \partial_{xx} \mathscr{B}_m$ can be computed in an efficient manner, a suitable truncation of the infinite series leads to a system of nonlinear ordinary differential equations (choose finite index set $\mathscr{M}_M \subset \mathscr{M}$ comprising M elements, replace coefficient functions by approximations $v_m(t) \approx u_m(t)$ for $m \in \mathscr{M}_M$)

$$\sum_{m \in \mathscr{M}_M} v_m'(t)\,\mathscr{B}_m(x) = \sum_{m \in \mathscr{M}_M} v_m(t)\,\big(\alpha\,\partial_{xx}\mathscr{B}_m(x) + \beta\,\partial_x \mathscr{B}_m(x)\big) + \gamma\,g\Big(\sum_{m \in \mathscr{M}_M} v_m(t)\,\mathscr{B}_m(x)\Big).$$

In view of a practical implementation, this relation is evaluated at certain space grid points.

(i) *Spectral methods.* Spectral methods rely on the consideration of a differential operator and the associated eigenfunctions

$$\mathscr{A}\,\mathscr{B}_m = \lambda_m\,\mathscr{B}_m\,, \qquad m \in \mathscr{M}\,;$$

fundamental presumptions are that the principal linear part of the partial differential equation is defined by this operator and that the family of eigenfunctions forms a complete orthonormal system of the underlying function space. Relevant examples include the Fourier spectral method (Laplacian), the Hermite spectral method (Laplacian with additional quadratic potential), and the generalised Laguerre–Fourier–Hermite spectral method (Laplacian with additional quadratic potential and rotation term). With regard to applications in atmospheric sciences, spectral methods based on scalar or vector spherical harmonics, i.e. special functions on the surface of the sphere, are of importance.

(ii) *Fourier spectral methods.* In particular, in connection with the Fourier spectral method, we use the eigenvalue relations (with $\mu_m = \frac{2\pi}{b-a}\,m$ and $\mathscr{M} = \mathbb{Z}$)

$$\partial_x \mathscr{F}_m = \mathrm{i}\,\mu_m\,\mathscr{F}_m\,, \qquad \partial_{xx}\mathscr{F}_m = -\,\mu_m^2\,\mathscr{F}_m\,, \qquad m \in \mathscr{M}\,,$$
$$\mathscr{A} = \alpha\,\partial_{xx} + \beta\,\partial_x\,, \qquad \mathscr{A}\,\mathscr{F}_m = \lambda_m\,\mathscr{F}_m\,, \qquad \lambda_m = \alpha\,\mathrm{i}\,\mu_m - \beta\,\mu_m^2\,,$$

which lead to the following system of nonlinear ordinary differential equations (with $\mathscr{M}_M = \{-\frac{M}{2}, \ldots, \frac{M}{2} - 1\}$)

$$\sum_{m \in \mathscr{M}_M} v_m'(t)\,\mathscr{F}_m(x) = \sum_{m \in \mathscr{M}_M} \lambda_m\,v_m(t)\,\mathscr{F}_m(x) + \gamma\,g\Bigg(\sum_{m \in \mathscr{M}_M} v_m(t)\,\mathscr{F}_m(x)\Bigg).$$

If $\gamma = 0$, the system decouples and the exact solution is given by the exponential function (employ orthogonality of Fourier basis functions)

$$\gamma = 0: \qquad v_m'(t) = \lambda_m\,v_m(t)\,, \qquad m \in \mathscr{M}_M$$
$$v_m(t) = \mathrm{e}^{\lambda_m\,t}\,v_m(0)\,, \qquad m \in \mathscr{M}_M\,;$$

otherwise, a suitable time integration method has to be used for the resolution of the system.

(iii) *Finite element method.* In the context of the finite element method, the fundamental idea is to define basis functions that are non-zero on a small region. In most cases, the basis functions are given by piecewise polynomials (non-zero on subinterval); thus, a space discretisation by central finite differences is related to the application of the finite element method with a special choice of polynomial basis functions. In addition, a weak formulation of the partial differential equation, obtained by integration over the space domain, is employed. As the oversimplified example (consider $\mathscr{M} = \{0,1,2,3\}$, computation of coefficients by resolution of linear systems)

$$\mathscr{B}_0(x) = 1\,, \qquad \mathscr{B}_1(x) = 1 + x\,, \qquad \mathscr{B}_2(x) = 1 + x + x^2\,, \qquad \mathscr{B}_3(x) = 1 + x + x^2 + x^3\,,$$
$$\partial_x \mathscr{B}_3(x) = 1 + 2\,x + 3\,x^2 = c_0\,\mathscr{B}_0(x) + c_1\,\mathscr{B}_1(x) + c_2\,\mathscr{B}_2(x)\,,$$
$$\partial_{xx}\mathscr{B}_3(x) = 2 + 6\,x = d_0\,\mathscr{B}_0(x) + d_1\,\mathscr{B}_1(x)\,,$$

indicates, the derivatives of the basis functions involve basis functions of lower degrees; even if $\gamma = 0$, the resulting system will not decouple. However, due to the fact that the basis functions are only defined locally, the arising matrices will be sparse, i.e. they involve only few non-zero elements.

Contrary to spectral methods, which need the underlying spatial domain and the defining differential operators to be of a special form, the finite element method is well-suited for the treatment of general partial differential equations. Furthermore, its extension to higher space dimensions and domains of arbitrary shape is straightforward.

3.3 Time discretisation methods

Situation. By a finite difference space discretisation of the model equation, we have obtained a high-dimensional system of nonlinear ordinary differential equations (with solution $U : [0,T] \to \mathbb{R}^M$, favourable approximations to exact solution values expected for *large* integer numbers $M \in \mathbb{N}$)

$$\begin{cases} U'(t) = F\big(U(t)\big) = \alpha\, A\, U(t) + \beta\, B\, U(t) + \gamma\, G\big(U(t)\big), & t \in (0,T), \\ U(0) = U_0\,. \end{cases}$$

Time discretisation methods. For a (sufficiently small) time increment $\tau = \frac{T}{N} > 0$, the numerical solution values at equidistant time grid points $t_n = n\tau$ for $n \in \{1,\dots,N\}$ are determined by a recurrence relation of the form (slight misuse of notation $U_1 \leftrightarrow U_1(t)$)

$$U_n = S(\tau, U_{n-1}) \;\approx\; U(t_n) = E\big(\tau, U(t_{n-1})\big);$$

it suffices to specify the initial step.

(i) *Explicit Euler method.* The explicit Euler method relies on an evaluation of the differential equation at time $t = 0$ and the application of the forward finite difference approximation ($\frac{1}{\tau}\,(U(\tau) - U(0)) \approx U'(0)$, see Section 1.4)

$$\begin{aligned} \tfrac{1}{\tau}\big(U_1 - U_0\big) &= F(U_0)\,, \\ U_1 &= U_0 + \tau F(U_0)\,. \end{aligned}$$

Due to stability issues, the application of the explicit Euler method to diffusion-advection-reaction equations requires tiny time increments.

(ii) *Implicit Euler method.* The implicit Euler method relies on an evaluation of the differential equation at time $t = \tau$ and the application of the backward finite difference approximation ($\frac{1}{\tau}\,(U(\tau) - U(0)) \approx U'(\tau)$, see Section 1.4)

$$\begin{aligned} \tfrac{1}{\tau}\big(U_1 - U_0\big) &= F(U_1)\,, \\ U_1 &= U_0 + \tau F(U_1)\,. \end{aligned}$$

Compared to the explicit Euler method, the implicit Euler method has a favourable stability behviour for the solution of (semi-linear) diffusion-reaction systems; the treatment of equations involving an addition advection part, is more delicate. The computation of the numerical solution values requires the resolution of nonlinear systems; for moderate values of $M \gg 1$, Newton's method is usually the method of choice (solve system defined by nonlinear function H, here $H : \mathbb{R}^M \to \mathbb{R}^M : z \mapsto H(z) = z - U_0 - \tau F(z)$, Taylor series expansion yields $H(z_{k+1}) \approx H(z_k) + H'(z_k)(z_{k+1} - z_k)$, aim is to achieve $H(z_{k+1}) \approx 0$, yields

relation $z_{k+1} = z_k - (H'(z_k))^{-1} H(z_k)$, realisation by resolution of linear systems)

$$\begin{aligned} &H(z) = 0\,, \\ &\begin{cases} H'(z_k)\,\zeta_k = H(z_k)\,, \\ z_{k+1} = z_k - \zeta_k\,, \end{cases} \qquad k = 0, 1, \dots \end{aligned}$$

(iii) *Semi-implicit Euler method.* As the resolution of nonlinear systems is connected with high computational costs, explicit-implicit schemes are often used in practice; again, stability issues have to be treated with care. The fact that special solvers are available for the resolution of systems with symmetric matrices suggests the choice

$$\begin{aligned} \tfrac{1}{\tau}\big(U_1 - U_0\big) &= \alpha\, A\, U_1 + \beta\, B(U_0) + \gamma\, G(U_0)\,, \\ U_1 &= \big(I - \tau\,\alpha\, A\big)^{-1}\big(U_0 + \tau\,\beta\, B(U_0) + \tau\,\gamma\, G(U_0)\big)\,. \end{aligned}$$

(iv) *Symplectic Euler method.* The symplectic Euler method or Lie–Trotter splitting method, respectively, uses a natural decomposition of the right-hand side into several parts and the fact that efficient solvers are available for the associated subproblems. For instane, solving the initial value problem involving A, B over a subinterval

$$\begin{cases} V'(t) = \alpha\, A\, V(t) + \beta\, B\, V(t)\,, \qquad t \in (0, \tau)\,, \\ V(0) = U_0\,, \end{cases}$$

and subsequently the initial value problem

$$\begin{cases} W'(t) = \gamma\, G\big(W(t)\big)\,, \qquad t \in (0, \tau)\,, \\ W(0) = V(\tau)\,, \end{cases}$$

yields an approximation at time $t = \tau$

$$W(\tau) \approx U(\tau)\,.$$

Alternatively, a splitting into three parts can be used.

Splitting approach. The Lie–Trotter splitting method can also be applied to diffusion-advection-reaction equations, before discretisation in space. This provides the possibility to employ space and time discretisation methods that are well adapted to each subproblem. For the considered model equation, the splitting approach permits to make use of the fact that the exact solution to the linear advection equation is known.

Chapter 4

Basic time integration methods

Contents. In this chapter, we present fundamental concepts that are employed for a stability and error analysis of time integration methods. As the simplest representatives for widely used classes of time integration methods, explicit as well as implicit Runge–Kutta and linear multi-step methods, we study the explicit and implicit Euler methods. Presuming that a spatial semi-discretisation has been realised for the considered class of diffusion-advection-reaction equations, for instance by finite differences, finite elements, finite volums, or spectral methods, we may restrict ourselve to the study of a system of ordinary differential equations. In order to lower technical difficulties, we provide details for the scalar Dahlquist equation only.

Attention! In general, it is *not* justified to draw conclusions on partial differential equations from the treatment of ordinary differential equations. As well, it is *not* justified to draw conclusions on nonlinear differential equations from the treatment of related linear differential equations. In particular, the study of the Dahlquist test equation is only a first step; it provides some insight in view of the fact that it may lead to the exclusion of a class of numerical methods for a particular class of differential equations.

4.1 Time stepping approach

Differential equation. We consider the following initial value problem comprising a nonlinear ordinary differential equation and an initial condition (initial and final time $t_0, T \in \mathbb{R}$, assume $T > t_0$, prescribed defining function $F : [t_0, T] \times \mathbb{R}^d \to \mathbb{R}^d$, prescribed initial value $y_0 \in \mathbb{R}^d$, unknown solution $y : [t_0, T] \to \mathbb{R}^d$, in view of practical implementation include explicit time dependency, transformation to autonomous form only used for theoretical investigations)

$$\begin{cases} y'(t) = F\big(t, y(t)\big), & t \in (t_0, T), \\ y(t_0) = y_0. \end{cases}$$

General requirement. We focus on situations, where the data and the solution to the considered differential equation is sufficiently often differentiable with derivatives bounded by moderate constants.

Integral equation. For different purposes, it is useful to consider a reformulation of the differential equation as integral equation

$$y(t) = y_0 + \int_{t_0}^{t} F\big(s, y(s)\big)\,\mathrm{d}s, \qquad t \in [t_0, T].$$

In view of the introduction of numerical approximations, we choose time grid points certain $t_n, t_{n+1} \in [t_0, T]$ such that $t_n < t_{n+1}$ and consider the relation

$$y(t_{n+1}) = y(t_n) + \int_{t_n}^{t_{n+1}} F\big(s, y(s)\big)\,\mathrm{d}s.$$

Explanation. The integral equation follows at once by intergration and an application of the fundamental theorem of calculus

$$\begin{aligned} y'(s) &= F\big(y(s), s\big), \qquad s \in (t_0, T), \\ \int_{t_n}^{t_{n+1}} y'(t)\,\mathrm{d}t &= \int_{t_n}^{t_{n+1}} F\big(s, y(s)\big)\,\mathrm{d}s, \\ y(t_{n+1}) - y(t_n) &= \int_{t_n}^{t_{n+1}} F\big(s, y(s)\big)\,\mathrm{d}s. \end{aligned}$$

Objective. Our objective is to introduce approximations to the exact solution values at certain time grid points by recurrence (for some positive integer $N \in \mathbb{N}$, additional dependence of evolution operators on starting time due to fact that differential equation is non-autonomous)

$$\begin{gathered} t_0 < t_1 < \cdots < t_N = T, \\ y_n = S(t_n - t_{n-1}, t_{n-1}, y_{n-1}) \approx y(t_n) = E\big(t_n - t_{n-1}, t_{n-1}, y(t_{n-1})\big), \qquad n \in \{1, \ldots, N\}, \end{gathered}$$

and to characterise the quality of the time-discrete solution (stability, rate of convergence, efficiency). The consideration of constant time increments slightly reduces the technicalities

$$\tau = \tfrac{T-t_0}{N}\,, \qquad t_n = t_0 + n\,\tau\,, \qquad n \in \{0,\dots,N\}\,,$$
$$y_n = S(\tau, t_{n-1}, y_{n-1}) \;\approx\; y(t_n) = E\big(\tau, t_{n-1}, y(t_{n-1})\big)\,, \qquad n \in \{1,\dots,N\}\,,$$

Remark. We point out that the exact evolution operator satisfies the relation (for simplicity assume global existence of solution)

$$y(t_0+s+t) = E\big(s+t, t_0, y(t_0)\big) = E\Big(t, t_0+s, E\big(s, t_0, y(t_0)\big)\Big)\,, \qquad s,t \in \mathbb{R}\,,$$

contrary to the numerical evolution operator.

Adaptivity. In view of efficiency, it is desirable to use variable time increments in order to optimally adapt the numerical approximations to the exact solution profile. Whenever the solution varies slowly, larger time stepsizes can be used; in other regions, smaller stepsizes are required.

4.2 Explicit Euler method

Explicit Euler method . The explicit Euler method is given by the recurrence relation

$$y_{n+1} = y_n + (t_{n+1} - t_n)\, F(t_n, y_n)\,, \qquad n \in \{0, \dots, N-1\}\,;$$

that is, we have

$$S(t, v) = v + t\, F(t_n, y_n)\,, \qquad n \in \{0, \dots, N-1\}\,;$$

In particular, for equidistant time grid points, we have

$$y_{n+1} = y_n + \tau\, F(t_n, y_n)\,, \qquad n \in \{0, \dots, N-1\}\,.$$

Explanation. As stated before, the explicit Euler method relies on the approach to consider the differential equation and to replace the differential quotient by a forward finite difference approximtion

$$\begin{aligned} y'(t_n) &= F\big(t_n, y(t_n)\big)\,, \\ \frac{y(t_{n+1}) - y(t_n)}{t_{n+1} - t_n} &\approx F\big(t_n, y(t_n)\big)\,, \\ y(t_n + \tau) &\approx y(t_n) + (t_{n+1} - t_n)\, F\big(t_n, y(t_n)\big)\,, \\ y_{n+1} &= y_n + (t_{n+1} - t_n)\, F(t_n, y_n)\,. \end{aligned}$$

An alternative approach is to start from the integral equation and to employ a quadrature approximation by the left-hand rule

$$\begin{aligned} y(t_{n+1}) &= y(t_n) + \int_{t_n}^{t_{n+1}} F\big(s, y(s)\big)\, \mathrm{d}s\,, \\ \int_{t_n}^{t_{n+1}} F\big(s, y(s)\big)\, \mathrm{d}s &\approx (t_{n+1} - t_n)\, F\big(t_n, y(t_n)\big)\,, \\ y(t_{n+1}) &\approx y(t_n) + (t_{n+1} - t_n)\, F\big(t_n, y(t_n)\big)\,, \\ y_{n+1} &= y_n + (t_{n+1} - t_n)\, F(t_n, y_n)\,. \end{aligned}$$

This second approach is valuable with regard to the introduction of higher-order Runge–Kutta and linear multi-step methods. ◇

Question. How does the accuracy of the approximations improve when the time increments are refined?

Aim. Our objective is to study the approximation errors

$$\big\| y_n - y(t_n) \big\|\,, \qquad n \in \{1, \dots, N\}\,.$$

In particular, we aim at deducing an estimate for the global error

$$\big\| y_N - y(T) \big\|$$

in terms of the (maximal) time increment.

Approach. As a direct estimation of the difference $y_N - y(T)$ is difficult, a basic approach is to relate the global error to local errors. Estimates for the local errors together with stability bounds yield a global error bound.

Global error estimate for test equation . For simplicity, we consider the Dahlquist test equation discretised in time by the explicit Euler method with constant time increments. We recall that the special case

$$\lambda = \mathrm{i}\,\mu\,, \qquad \mu \in \mathbb{R}\,,$$

is of relevance in the context of advection equations; in the context of diffusion equations, it is of interest to study the case

$$\lambda = -\,\mu^2 \leq 0\,, \qquad \mu \in \mathbb{R}\,.$$

(i) *Exact solution.* For the Dahlquist test equation (defining function $F : [0,\infty) \times \mathbb{R} \to \mathbb{R} : (t,z) \mapsto \lambda\, z$ with $\lambda \in \mathbb{C}$, for autonomous equation sufficient to consider $t_0 = 0$)

$$\begin{cases} y'(t) = \lambda\, y(t)\,, & t \in (0,\infty)\,, \\ y(0) = y_0\,, \end{cases}$$

the exact solution value at a final time $T > 0$ is given by the exponential function

$$E(t) = \mathrm{e}^{\,t\,\lambda}\,, \qquad t \in [0,\infty)\,,$$
$$y(T) = E(T)\, y_0\,.$$

(ii) *Explicit Euler solution.* Resolving the recurrence for the numerical solution, obtained by the explicit Euler method, yields (constant time increment $\tau = \frac{T}{N}$ for some $N \in \mathbb{N}$)

$$\begin{aligned} y_N &= y_{N-1} + \tau\,\lambda\, y_{N-1} \\ &= (1 + \tau\,\lambda)\, y_{N-1} \\ &= (1 + \tau\,\lambda)^2\, y_{N-2} \\ &= \ldots \\ &= (1 + \tau\,\lambda)^N\, y_0\,. \end{aligned}$$

In analogy to the exact solution, it is convenient to introduce the function

$$S(t) = 1 + t\,\lambda\,, \qquad t \in [0,\infty)\,,$$
$$y_n = S(\tau)\, y_{n-1}\,, \qquad n \in \{1,\ldots,N\}\,,$$
$$y_N = \big(S(\tau)\big)^N\, y_0\,.$$

(iii) *Group property.* We note that the exponential function satisfies the relation

$$E(s)\,E(t) = E(s+t)\,, \qquad s,t > 0\,,$$

in contrast to the numerical approximation

$$S(s)\,S(t) \neq S(s+t)\,, \qquad s,t > 0\,.$$

(iv) *Auxiliary relation.* For later use, we recall the elementary relations (where $a, b, q \in \mathbb{R}$, telescopic identity corresponds to geometric series, generalisation)

$$\sum_{n=0}^{N-1} q^n = \frac{q^N - 1}{q-1}, \qquad q^N - 1 = (q-1) \sum_{n=0}^{N-1} q^n,$$

$$a \neq b: \qquad a^N - b^N = b^N \Big(\big(\tfrac{a}{b}\big)^N - 1 \Big) = b^N \big(\tfrac{a}{b} - 1\big) \sum_{n=0}^{N-1} \big(\tfrac{a}{b}\big)^n = (a-b) \sum_{n=0}^{N-1} a^n \, b^{N-1-n}.$$

(v) *Global error.* The global error is given by

$$\begin{aligned} \big(S(\tau)\big)^N - E(T) &= (1+\tau\lambda)^N - \mathrm{e}^{\lambda T}, \\ y_N - y(T) &= \Big(\big(S(\tau)\big)^N - E(T) \Big) y_0, \\ \big| y_N - y(T) \big| &\leq \Big| \big(S(\tau)\big)^N - E(T) \Big| \, \big| y_0 \big|. \end{aligned}$$

(vi) *Approach.* In order to expand the global error in terms of the time increment, we employ the basic relation for the exponential function and the telescopic identity (recall that $t_n = n\tau$ and $T = N\tau$, corresponding relations for E, S)

$$\begin{aligned} \big(S(\tau)\big)^N - E(T) &= \big(S(\tau)\big)^N - \big(E(\tau)\big)^N \\ &= \big(S(\tau) - E(\tau)\big) \sum_{n=0}^{N-1} \big(S(\tau)\big)^n E(t_{N-1-n}), \\ (1+\tau\lambda)^N - \mathrm{e}^{T\lambda} &= (1+\tau\lambda)^N - \big(\mathrm{e}^{\tau\lambda}\big)^N \\ &= \big(1 + \tau\lambda - \mathrm{e}^{\tau\lambda}\big) \sum_{n=0}^{N-1} (1+\tau\lambda)^n \, \mathrm{e}^{t_{N-1-n}\lambda}. \end{aligned}$$

(vii) *Stability estimates.* Stability bounds for the exponential function and its numerical counterpart on bounded time intervals (use that $\mathrm{e}^{t\lambda} \leq \mathrm{e}^{T|\lambda|}$, the elementary estimate $1 + x \leq \mathrm{e}^x$ for $x \geq 0$ and thus $(1+\tau\lambda)^N \leq \mathrm{e}^{N\tau|\lambda|} = \mathrm{e}^{T|\lambda|}$)

$$\begin{gathered} E(t) = \mathrm{e}^{t\lambda}, \qquad S(t) = 1 + t\lambda, \qquad t \in \mathbb{R}, \qquad \lambda \in \mathbb{C}, \\ \big| E(t) \big| \leq M_E, \qquad t \in [0,T], \qquad \Big| \big(S(\tau)\big)^N \Big| \leq M_S, \qquad N\tau = T, \end{gathered}$$

are essential means for the estimation of the global error. We point out that boundedness of $S(\tau)$ by a constant is not sufficient to ensure boundedness of $\big(S(\tau)\big)^N$ for $\tau \to 0$ and $N \to \infty$ such that $N\tau = T$. Often, coarse bounds that remain valid for arbitrary complex numbers $\lambda \in \mathbb{C}$ are employed; the consideration of special cases permits to obtain refined estimates.

(viii) *Special cases.* In connection with partial differential equations, stability issues have to be treated with care, as the following considerations show.

(a) *Explicit Euler method for diffusion equations.* With regard to the time integration of diffusion equations, we consider the special case $\lambda = -\mu^2 \leq 0$ with $\mu \in \mathbb{R}$, see Chapter 3. We note that the Fourier spectral method shows that the limiting case $\mu \to \infty$ occurs; as well, for the spatially discretised system large values $\mu \gg 1$ arise. Unlike the exponential function, which remains bounded for all positive times

$$\big|E(t)\big| \leq 1\,, \qquad t \in [0,\infty)\,,$$

the modulus of the numerical counterpart goes beyond one, in general, leading to a highly oscillatory behaviour of the explicit Euler solution

$$\begin{aligned} \tau\mu^2 = 1 + c\,, \quad c \gg 1 \quad \Longrightarrow \quad & S(\tau) = 1 - \tau\mu^2 = -c\,, \\ & \big|S(\tau)\big| = \tau\mu^2 - 1 = c\,, \\ & \big(S(\tau)\big)^2 = c^2\,, \quad \big(S(\tau)\big)^3 = -c^3\,, \quad \text{etc}\,. \end{aligned}$$

Only for sufficiently small time increments, associated with high computational costs, stability of the explicit Euler method can be ensured.

(b) *Explicit Euler method for advection equations.* In view of the time integration of advection equations, the special case $\lambda = \mathrm{i}\,\mu$ with $\mu \in \mathbb{R}$ is of interest, see Chapter 3. Again, the exponential function remains bounded for all positive times; more precisely, we even have

$$\big|E(t)\big| = 1\,, \qquad t \in \mathbb{R}\,.$$

The numerical counterpart, obtained by the explicit Euler method, however, is unstable for all time increments (but the trivial case $\mu = 0$)

$$S(\tau) = 1 + \mathrm{i}\,\tau\mu\,, \qquad \big|S(\tau)\big| = \sqrt{1 + \tau^2\mu^2} > 1\,.$$

(ix) *Local error estimate.* An estimate for the difference between the exact and numerical solutions after a single step is obtained by means of a Taylor series expansion (local error, note that same starting value is used, integration-by-parts with $f'(\sigma) = 1$, $g(\sigma) = \mathrm{e}^{\sigma x}$, $f(\sigma) = -(1-\sigma)$, $g'(\sigma) = x\,\mathrm{e}^{\sigma x}$)

$$\begin{aligned} \mathrm{e}^x - 1 &= \mathrm{e}^{\sigma x}\Big|_{\sigma=0}^1 \\ &= \int_0^1 \tfrac{\mathrm{d}}{\mathrm{d}\sigma}\,\mathrm{e}^{\sigma x}\,\mathrm{d}\sigma \\ &= x\int_0^1 \mathrm{e}^{\sigma x}\,\mathrm{d}\sigma \\ &= -x\,(1-\sigma)\,\mathrm{e}^{\sigma x}\Big|_{\sigma=0}^1 + x^2\int_0^1 (1-\sigma)\,\mathrm{e}^{\sigma x}\,\mathrm{d}\sigma \\ &= x + x^2\int_0^1 (1-\sigma)\,\mathrm{e}^{\sigma x}\,\mathrm{d}\sigma\,. \end{aligned}$$

This further implies the local error bound

$$S(\tau) - E(\tau) = 1 + \tau\,\lambda - \mathrm{e}^{\tau\lambda} = -\,(\lambda\tau)^2 \int_0^1 (1-\sigma)\,\mathrm{e}^{\sigma\lambda\tau}\,\mathrm{d}\sigma\,,$$
$$\big|S(\tau) - E(\tau)\big| \le |\lambda|^2\,\mathrm{e}^{\tau|\lambda|}\,\tau^2 \le M_L\,\tau^2\,.$$

(x) *Global error estimate.* Combining the stated stability and local error estimates, we obtain the bound

$$\begin{aligned}\big|\big(S(\tau)\big)^N - E(T)\big| &\le \big|S(\tau) - E(\tau)\big| \sum_{n=0}^{N-1} \big|\big(S(\tau)\big)^n\big|\big|E(t_{N-1-n})\big| \\ &\le M_L\,\tau^2 \sum_{n=0}^{N-1} M_S\,M_E \\ &\le M_S\,M_E\,M_L\,T\,\tau\,.\end{aligned}$$

Altogether, this implies the global error estimate (convergence order $p = 1$)

$$p = 1: \qquad \big|y_N - y(T)\big| \le C\,\tau^p\,, \qquad C = C(M_S, M_E, M_L, T)\,.$$

This is particular ensures convergence of the numerical solution values towards the exact solution values (on bounded time intervals)

$$\lim_{\tau\to 0} \big|y_N - y(N\tau)\big| = 0\,.$$

(xi) *Variable time stepsizes.* Similar considerations apply to the more general case of non-equidistant time grid points.

4.3 Implicit Euler method

Implicit Euler method. The implicit Euler method is given by the recurrence relation

$$y_{n+1} = y_n + (t_{n+1} - t_n)\, F(t_{n+1}, y_{n+1}), \qquad n \in \{0, \dots, N-1\}.$$

It is obtained by the application of the backward finite difference approximation or the right-hand rule, respectively.

Stability. The poor stability behaviour of the explicit Euler method for *stiff equations*, for instance, for systems of ordinary differential equations obtained from a space discretisation of diffusion equations, motivates the introduction of the implicit Euler method (numerical solution to Dahlquist test equation given by $y_1 = y_0 + \tau\lambda\, y_1$)

$$S(\tau) = \left(1 - \tau\lambda\right)^{-1} = \left(1 + \tau\mu^2\right)^{-1},$$
$$\left|S(\tau)\right| \leq 1, \qquad \tau > 0,$$

which is stable for arbitrary positive time increments.

Global error. The above approach for the derivation of a global error estimate extends to the implicit Euler method and shows a first-order convergence bound. We note that the size of the constant M_L is effected by the size of $\lambda \in \mathbb{C}$; thus, in the context of partial differential equations, a more careful treatment is required.