NUMERICS OF DIFFERENTIAL EQUATIONS

MICHAEL INNERBERGER AND DIRK PRAETORIUS

2. Explicit one-step methods

Throughout this section, we consider the following model problem: Let $[t_0, T]$ be a given time-interval. For given $n \in \mathbb{N}$, let $f \in C([t_0, T] \times \mathbb{R}^n; \mathbb{R}^n)$ and suppose that f is Lipschitz continuous in y, i.e.,

$$\forall\, t \in [t_0, T]\, \forall\, y, \widetilde{y} \in \mathbb{R}^n : \quad \|f(t, y) - f(t, \widetilde{y})\| \le L\, \|y - \widetilde{y}\|, \tag{2.1}$$

where $\|\cdot\|$ is an arbitrary norm on $\mathbb{R}^n$ and $L > 0$ is a fixed constant. Then, for any initial value $y_0 \in \mathbb{R}^n$, the Picard–Lindelöf theorem guarantees existence and uniqueness of $y \in C^1([t_0, T]; \mathbb{R}^n)$ such that

$$y(t_0) = y_0 \quad \text{and} \quad y'(t) = f(t, y(t)) \quad \text{for all } t \in [t_0, T]. \tag{2.2}$$

Remark 2.1. *If $f(t, y) = g(t)$, then the exact solution of (2.2) reads*

$$y(t) = y_0 + \int_{t_0}^{t} g \,\mathrm{d}s,$$

*i.e., y is the antiderivative of g. Therefore, the solution of an initial value problem (2.2) ist also called **integration of the ODE** and the numerical solvers are also called **(numerical) integrators**.*

2.1. Notation.

In practice, the solution of the initial boundary value problem (2.2) cannot be computed in closed form. If quantitative results are required, one usually considers numerical methods, which provide approximations $y_\ell \approx y(t_\ell)$ at certain time-steps $t_\ell \in [t_0, T]$. From now on, we shall implicitly use the following notation:

Definition 2.2. *Suppose that we have given time-steps $t_0 < t_1 < \cdots < t_N = T$. The set $\Delta := \{t_0, \dots, t_N\}$ is called **mesh** (or **grid**) of the time interval $[t_0, T]$. The quantities $h_\ell := t_{\ell+1} - t_\ell$ are called **step-sizes**. Moreover, the **maximum step-size** of Δ is defined as $h_\Delta := \max_{\ell=0,\dots,N-1} h_\ell$.*

Definition 2.3. *Given an **incremental function** $\Phi : [t_0, T] \times \mathbb{R}^n \times \mathbb{R}_{>0} \to \mathbb{R}^n$, the inductive procedure*

$$y_{\ell+1} := y_\ell + h_\ell\, \Phi(t_\ell, y_\ell, h_\ell) \quad \textit{for all } \ell = 0, \dots, N-1 \tag{2.3}$$

Note: These a *posteriori* lecture notes provide the way in which I would have liked to present the course material. If you face any typos, please let me know: `dirk.praetorius@asc.tuwien.ac.at`.

is called ***explicit one-step method****. Given $y_\ell \approx y(t_\ell)$, we can compute $y_{\ell+1} \approx y(t_{\ell+1})$. Note that this computation involves only the last time-step, but* not *the full history $y_0, \dots, y_{\ell-1}$. This will be different for* multi-step methods *considered below.*

Remark 2.4. *Later, we will also consider* implicit one-step methods*, where Φ depends additionally on the sought approximation $y_{\ell+1}$, i.e.,*

$$y_{\ell+1} = y_\ell + h_\ell \, \Phi(t_\ell, y_\ell, y_{\ell+1}, h_\ell).$$

Then, each time-step of the implicit one-step method will require the solution of one (possibly nonlinear) equation.

Example 2.5 (Explicit Euler method / forward Euler method). *If $f \in C^1([t_0, T] \times \mathbb{R}^n; \mathbb{R}^n)$ and hence $y \in C^2([t_0, T]; \mathbb{R}^n)$, the Taylor theorem proves that*

$$y(t+h) = y(t) + h\, y'(t) + \mathcal{O}(h^2) = y(t) + h\, f\big(t, y(t)\big) + \mathcal{O}(h^2). \tag{2.4}$$

With $y(t_{\ell+1}) \approx y_{\ell+1}$ and $y(t_\ell) \approx y_\ell$, the explicit Euler method reads

$$y_{\ell+1} = y_\ell + h_\ell \, \Phi(t_\ell, y_\ell, h_\ell) \quad \text{with} \quad \Phi(t_\ell, y_\ell, h_\ell) := f\big(t_\ell, y_\ell\big). \tag{2.5}$$

Example 2.6 (Implicit Euler method / backward Euler method). *If $f \in C^1([t_0, T] \times \mathbb{R}^n; \mathbb{R}^n)$ and hence $y \in C^2([t_0, T]; \mathbb{R}^n)$, the Taylor theorem proves that*

$$y(t) = y(t+h) - h\, y'(t+h) + \mathcal{O}(h^2) = y(t+h) - h\, f\big(t+h, y(t+h)\big) + \mathcal{O}(h^2). \tag{2.6}$$

With $y(t_{\ell+1}) \approx y_{\ell+1}$ and $y(t_\ell) \approx y_\ell$, the implicit Euler method reads

$$y_{\ell+1} = y_\ell + h_\ell \, \Phi(t_\ell, y_\ell, y_{\ell+1}, h_\ell) \quad \text{with} \quad \Phi(t_\ell, y_\ell, y_{\ell+1}, h_\ell) := f\big(t_{\ell+1}, y_{\ell+1}\big), \tag{2.7}$$

where we note that $t_{\ell+1} = t_\ell + h_\ell$.

2.2. Consistency.

Definition 2.7. *Let $\Phi(t, y, h)$ be the incremental function of an explicit one-step method. We say that the one-step method is* ***consistent****, if*

$$\forall\, t \in [t_0, T): \quad \lim_{h \to 0^+} \frac{\big\| y(t+h) - \big[y(t) + h\, \Phi(t, y(t), h)\big] \big\|}{h} = 0, \tag{2.8}$$

i.e., if we apply the one-step method to the exact solution, the discretization error of one step *vanishes as $h \to 0$. For $p \geq 1$, we say that the one-step method has* ***consistency***

Leonhard Euler (1707–1783) was a Swiss mathematician. From 1720–1723, he was studying at the University of Basel (attending also classes by Johann I Bernoulli). In 1727, he moved to the Imperial Russian Academy of Sciences in Saint Petersburg, where he became professor of mathematics in 1733. In 1741, he moved to the Prussian Academy of Sciences in Berlin, but returned to the Imperial Russian Academy of Sciences in Saint Petersburg in 1766.

order *p, if for all $f \in C^p([t_0,T]\times\mathbb{R}^n;\mathbb{R}^n)$ and hence $y \in C^{p+1}([t_0,T];\mathbb{R}^n)$ for the solution of (2.2), it holds that*

$$\exists\, C>0 :\ \forall\, t\in[t_0,T)\,\forall\, h\in(0,T-t]: \quad \big\|y(t+h)-\big[y(t)+h\,\Phi(t,y(t),h)\big]\big\| \le C\,h^{p+1}. \tag{2.9}$$

Remark 2.8. *Suppose that the incremental function $\Phi(t,y(t),h)$ is continuous at $h=0$. Then,*

$$\begin{aligned} \textit{consistency (2.8)} \quad &\iff \quad \forall\, t\in[t_0,T): \quad \lim_{h\to 0^+}\left\|\frac{y(t+h)-y(t)}{h}-\Phi(t,y(t),h)\right\|=0 \\ &\iff \quad \forall\, t\in[t_0,T): \quad f(t,y(t))=\Phi(t,y(t),0), \end{aligned}$$

since $f(t,y(t)) = y'(t) = \lim_{h\to 0^+}\dfrac{y(t+h)-y(t)}{h}$.

Example 2.9 (Explicit Euler has consistency order $p=1$). *The explicit Euler method (2.5) is always consistent, since $\Phi(t,y,h)=f(t,y)$. According to (2.4) it has consistency order $p=1$. Moreover, the constant C in the consistency estimate (2.9) takes the form $C=\|y''\|_{\infty,[t_0,T]}/2$, since the Taylor expansion yields that*

$$y(t+h)=y(t)+h\,y'(t)+\frac{h^2}{2}\,y''(\xi) \overset{(2.5)}{=} y(t)+h\,\Phi(t,y(t),h)+\frac{h^2}{2}\,y''(\xi)$$

for some appropriate $\xi\in[t,t+h]$.

Remark 2.10. *If the mesh Δ has step-size h, then it takes $N\sim\frac{T-t_0}{h}=\mathcal{O}(h^{-1})$ steps of the numerical scheme to obtain the approximation $y_N\approx y(T)$. If each step has a (cumulative consistency) error of order $\mathcal{O}(h^{p+1})$, one can hope for a total error estimate*

$$\|y(T)-y_N\| \sim \sum_{j=0}^{N-1}\mathcal{O}(h^{p+1}) = \mathcal{O}(Nh^{p+1}) = \mathcal{O}(h^p).$$

In the following section, we aim to rigorously prove this expectation.

2.3. Convergence.

Lemma 2.11 (Discrete Gronwall lemma). *Let $A>0$, $B\ge 0$, $h_\ell, a_\ell \ge 0$ such that*

$$0\le a_{\ell+1}\le (1+h_\ell A)a_\ell + h_\ell B \quad \textit{for all } \ell\in\mathbb{N}_0. \tag{2.10}$$

Then, it holds that

$$0\le a_\ell \le \frac{B}{A}\left(\exp\Big(A\sum_{j=0}^{\ell-1}h_j\Big)-1\right)+a_0\exp\Big(A\sum_{j=0}^{\ell-1}h_j\Big) \quad \textit{for all } \ell\in\mathbb{N}_0. \tag{2.11}$$

Proof. It holds that

$$1 + x \leq \sum_{j=0}^{\infty} \frac{x^j}{j!} = \exp(x) \quad \text{for all } x \geq 0. \tag{2.12}$$

Note that (2.11) holds for $\ell = 0$ (even with equality $a_\ell = a_0$). Arguing by induction on ℓ, we may suppose that (2.11) holds up to some $\ell \in \mathbb{N}_0$. For $\ell + 1$, note that

$$\begin{aligned}
a_{\ell+1} &\overset{(2.10)}{\leq} (1 + h_\ell A) a_\ell + h_\ell B \\
&\overset{(2.11)}{\leq} (1 + h_\ell A) \left[\frac{B}{A} \left(\exp\Big(A \sum_{j=0}^{\ell-1} h_j\Big) - 1 \right) + a_0 \exp\Big(A \sum_{j=0}^{\ell-1} h_j\Big) \right] + h_\ell B \\
&\leq \frac{B}{A} \Bigg(\underbrace{(1 + h_\ell A)}_{\overset{(2.12)}{\leq} \exp(h_\ell A)} \exp\Big(A \sum_{j=0}^{\ell-1} h_j\Big) - (1 + h_\ell A) \Bigg) + a_0 \underbrace{(1 + h_\ell A)}_{\overset{(2.12)}{\leq} \exp(h_\ell A)} \exp\Big(A \sum_{j=0}^{\ell-1} h_j\Big) + h_\ell B \\
&\leq \frac{B}{A} \left(\exp\Big(A \sum_{j=0}^{\ell} h_j\Big) - (1 + h_\ell A) \right) + a_0 \exp\Big(A \sum_{j=0}^{\ell} h_j\Big) + h_\ell B \\
&= \frac{B}{A} \left(\exp\Big(A \sum_{j=0}^{\ell} h_j\Big) - 1 \right) + a_0 \exp\Big(A \sum_{j=0}^{\ell} h_j\Big).
\end{aligned}$$

This concludes the proof. □

Theorem 2.12 (Consistency plus stability implies convergence). *Let $\Phi(t, y, h)$ be the incremental function of an explicit one-step method. Suppose stability*

$$\forall\, t \in [t_0, T)\, \forall\, h \in (0, T - t]\, \forall\, y, \widetilde{y} \in \mathbb{R}^n : \quad \|\Phi(t, y, h) - \Phi(t, \widetilde{y}, h)\| \leq C_{\mathrm{stab}} \|y - \widetilde{y}\| \tag{2.13}$$

for some constant $C_{\mathrm{stab}} > 0$. Let $f \in C^p([t_0, T] \times \mathbb{R}^n; \mathbb{R}^n)$ with corresponding solution $y \in C^{p+1}([t_0, T]; \mathbb{R}^n)$. Then, consistency order $p \geq 1$ implies that

$$\max_{\ell=1,\dots,N} \|y_\ell - y(t_\ell)\| \leq \frac{C_{\mathrm{cons}}}{C_{\mathrm{stab}}} \left[\exp\big(C_{\mathrm{stab}}(T - t_0)\big) - 1 \right] h_\Delta^p, \tag{2.14}$$

where $C_{\mathrm{cons}} > 0$ is the consistency constant; see (2.9) in Definition 2.7.

Proof. We define the consistency error

$$\eta(t_{\ell+1}) := y(t_{\ell+1}) - \big[y(t_\ell) + h_\ell\, \Phi(t_\ell, y(t_\ell), h_\ell)\big] \overset{(2.9)}{=} \mathcal{O}(h_\ell^{p+1}).$$

With this notation, the error satisfies

$$\begin{aligned}
E_{\ell+1} &:= y_{\ell+1} - y(t_{\ell+1}) \\
&= \big[y_\ell + h_\ell\, \Phi(t_\ell, y_\ell, h_\ell)\big] - \big[y(t_\ell) + h_\ell \Phi(t_\ell, y(t_\ell), h_\ell)\big] - \eta(t_{\ell+1}) \\
&= \big[y_\ell - y(t_\ell)\big] + h_\ell \big[\Phi(t_\ell, y_\ell, h_\ell) - \Phi(t_\ell, y(t_\ell), h_\ell)\big] - \eta(t_{\ell+1}).
\end{aligned}$$

With $a_\ell := \|y_\ell - y(t_\ell)\|$, stability (2.13) and consistency (2.9) lead to

$$0 \le a_{\ell+1} \le a_\ell\,(1 + h_\ell C_{\mathrm{stab}}) + C_{\mathrm{cons}}\,h_\ell^{p+1} \le a_\ell\,(1 + h_\ell C_{\mathrm{stab}}) + C_{\mathrm{cons}}\,h_\Delta^p\,h_\ell.$$

With $a_0 = 0$ and Lemma 2.11, we infer that

$$\|y_\ell - y(t_\ell)\| = a_\ell \overset{(2.11)}{\le} \frac{C_{\mathrm{cons}}}{C_{\mathrm{stab}}} \left(\exp\Big(C_{\mathrm{stab}} \sum_{j=0}^{\ell-1} h_j\Big) - 1 \right) h_\Delta^p \quad \text{for all } \ell \in \mathbb{N}_0.$$

The claim follows from $\sum_{j=0}^{\ell-1} h_j \le T - t_0$. □

Remark 2.13 (Stability of explicit Euler method). *For the explicit Euler method, it holds that $\Phi(t, y, h) = f(t, y)$. Hence, stability* (2.13) *with $C_{\mathrm{stab}} = L$ follows from Lipschitz continuity of f in y.*

Corollary 2.14 (One step methods are stable). *Under the assumptions of Theorem 2.12, we suppose that the* computed approximations *face rounding errors such that*

$$\widetilde{y}_0 = y_0 + \varepsilon_0 \quad \text{and} \quad \widetilde{y}_{\ell+1} = \widetilde{y}_\ell + h_\ell\,\Phi(t_\ell, \widetilde{y}_\ell, h_\ell) + \delta_\ell \quad \text{for all } \ell = 0, \dots, N-1, \tag{2.15}$$

where $\|\varepsilon_0\| \le \varepsilon$ and $\|\delta_\ell\| \le \delta\,h_\ell$. Then,

$$\max_{\ell=1,\dots,N} \|y_\ell - y(t_\ell)\| \le C\,(h_\Delta^p + \delta) + \varepsilon\,\exp\big(C_{\mathrm{stab}}(T - t_0)\big), \tag{2.16}$$

where $C = \dfrac{\max\{1, C_{\mathrm{cons}}\}}{C_{\mathrm{stab}}} \big[\exp\big(C_{\mathrm{stab}}(T - t_0)\big) - 1\big]$.

Proof. We argue as for the proof of Theorem 2.12. Define the consistency error

$$\eta(t_{\ell+1}) := y(t_{\ell+1}) - \big[y(t_\ell) + h_\ell\,\Phi(t_\ell, y(t_\ell), h_\ell)\big] = \mathcal{O}(h_\ell^{p+1}).$$

With this notation, the perturbed error satisfies

$$\begin{aligned}
\widetilde{E}_{\ell+1} &:= \widetilde{y}_{\ell+1} - y(t_{\ell+1}) \\
&= \big[\widetilde{y}_\ell + h_\ell\,\Phi(t_\ell, \widetilde{y}_\ell, h_\ell) + \delta_\ell\big] - \big[y(t_\ell) + h_\ell \Phi(t_\ell, y(t_\ell), h_\ell)\big] - \eta(t_{\ell+1}) \\
&= \big[\widetilde{y}_\ell - y(t_\ell)\big] + h_\ell\,\big[\Phi(t_\ell, \widetilde{y}_\ell, h_\ell) - \Phi(t_\ell, y(t_\ell), h_\ell)\big] - \eta(t_{\ell+1}) + \delta_\ell
\end{aligned}$$

With $a_\ell := \|\widetilde{y}_\ell - y(t_\ell)\|$, stability (2.13) and consistency (2.9) lead to

$$\begin{aligned}
0 \le a_{\ell+1} &\le a_\ell\,(1 + h_\ell C_{\mathrm{stab}}) + C_{\mathrm{cons}}\,h_\ell^{p+1} + \delta_\ell \\
&\le a_\ell\,(1 + h_\ell C_{\mathrm{stab}}) + \max\{1, C_{\mathrm{cons}}\}\,(h_\Delta^p + \delta)\,h_\ell
\end{aligned}$$

With $a_0 = \|\widetilde{y}_0 - y_0\| \le \varepsilon$ and Lemma 2.11, we infer that

$$\begin{aligned}
\|\widetilde{y}_\ell - y(t_\ell)\| = a_\ell \overset{(2.11)}{\le} (h_\Delta^p + \delta)\,\frac{\max\{1, C_{\mathrm{cons}}\}}{C_{\mathrm{stab}}} \left(\exp\Big(C_{\mathrm{stab}} \sum_{j=0}^{\ell-1} h_j\Big) - 1 \right) \\
+ \varepsilon\,\exp\left(C_{\mathrm{stab}} \sum_{j=0}^{\ell-1} h_j\right) \quad \text{for all } \ell \in \mathbb{N}_0.
\end{aligned}$$

With $\sum_{j=0}^{\ell-1} h_j \le T - t_0$, this concludes the proof. □

2.4. Examples. We aim to construct explicit one-step methods with higher convergence order. One natural approach is to start from the Taylor expansion of the exact solution

$$y(t+h) = y(t) + \sum_{k=1}^{p} \frac{y^{(k)}(t)}{k!}\, h^k + \mathcal{O}(h^{p+1}), \tag{2.17}$$

where we implicitly assume smoothness of $f \in C^p([t_0,T]\times\mathbb{R}^n;\mathbb{R}^n)$ and hence of $y \in C^{p+1}([t_0,T];\mathbb{R}^n)$. Since y is unknown (and so is $y^{(k)}$), we need to express the derivatives in terms of the given right-hand side f. First, recall that

$$y'(t) = f\big(t, y(t)\big). \tag{2.18}$$

To write the second derivative of f, define $g(t) := \big(t, y(t)\big)$. Then, the chain rule gives

$$\begin{aligned} y''(t) = \mathrm{d}_t f\big(t,y(t)\big) = \mathrm{d}_t\big[f\big(g(t)\big)\big] &= Df\big(t,y(t)\big)Dg(t) \\ &= (\partial_t f)\big(t,y(t)\big) + (\partial_y f)\big(t,y(t)\big)\, y'(t) \\ &= (\partial_t f)\big(t,y(t)\big) + (\partial_y f)\big(t,y(t)\big)\, f\big(t,y(t)\big). \end{aligned} \tag{2.19}$$

Proceeding with the chain rule, we can express all derivatives of y in terms of partial derivatives of f. We leave this to the interested reader, but state the following example for $p=2$.

Example 2.15 (Second-order one-step method based on Taylor expansion). *Given $f \in C^2([t_0,T]\times\mathbb{R}^n;\mathbb{R}^n)$, define the incremental function*

$$\Phi(t,y,h) := f(t,y) + \frac{h}{2}\big[(\partial_t f)(t,y) + (\partial_y f)(t,y)\, f(t,y)\big]. \tag{2.20}$$

Then, the corresponding explicit one-step method has consistency order $p=2$.

One drawback of the latter construction is that it requires high regularity of f to write down the incremental function. Moreover, one has to provide explicit formulas for the derivatives of f and, more importantly, stability (2.13) of the one-step method requires additional assumptions.

Alternatively, one can consider nested evaluations of f to avoid the evaluation of the derivatives of f.

Proposition 2.16. *Given $a, b_1, b_2, c \in \mathbb{R}$, define the incremental function*

$$\Phi(t,y,h) := b_1 f(t,y) + b_2 f\big(t+ch, y+ah\, f(t,y)\big). \tag{2.21}$$

Then, the following statements are equivalent:

(i) *The corresponding one-step method has consistency order $p=2$.*

(ii) $a = c, \quad b_1 + b_2 = 1, \quad$ *and* $\quad b_2 a = 1/2.$

Remark 2.17. *The conditions in Proposition 2.16*(ii) *show that one does not have four degrees of freedom (i.e., $a, b_1, b_2, c \in \mathbb{R}$), but only one to ensure second-order consistency. Fixing any of these constants, all other constants are determined as well. Moreover, second-order consistency requires $a \neq 0 \neq c$. Finally, we note that a reasonable method*

will additionally impose the restriction $c \in (0,1]$ to ensure that the evaluation point $t+ch$ will belong to the time interval $[t_0, T]$.

Proof of Proposition 2.16. Let $f \in C^2([t_0,T]\times\mathbb{R}^n;\mathbb{R}^n)$ and hence $y \in C^3([t_0,T];\mathbb{R}^n)$ be the exact solution. The Taylor expansion of Φ around $h=0$ proves that

$$\Phi(t,y,h) = \Phi(t,y,0) + h\,\partial_h\Phi(t,y,0) + \mathcal{O}(h^2).$$

Note that

$$\begin{aligned}\Phi(t,y,0) &= b_1 f(t,y) + b_2 f(t,y) = (b_1+b_2)\,f(t,y),\\ \partial_h\Phi(t,y,0) &= b_2 c\,\partial_t f(t,y) + b_2 a\,\partial_y f(t,y)\,f(t,y).\end{aligned}$$

Hence,

$$\Phi(t,y,h) = (b_1+b_2)\,f(t,y) + h\,\big[b_2 c\,\partial_t f(t,y) + b_2 a\,\partial_y f(t,y)\,f(t,y)\big] + \mathcal{O}(h^2). \tag{2.22}$$

Taylor expansion of y around t proves that

$$\begin{aligned} y(t+h) &= y(t) + h\,y'(t) + \frac{h^2}{2}\,y''(t) + \mathcal{O}(h^3)\\ &\overset{(2.19)}{=} y(t) + h\,f\big(t,y(t)\big) + \frac{h^2}{2}\,\big[(\partial_t f)\big(t,y(t)\big) + (\partial_y f)\big(t,y(t)\big)\,f\big(t,y(t)\big)\big]\end{aligned} \tag{2.23}$$

Combining (2.22)–(2.23), we can identify the consistency error. To ease the presentation, we omit the arguments $(t,y(t))$ for the f-terms. Then,

$$\begin{aligned} &y(t+h) - \big[y(t) + h\,\Phi\big(t,y(t),h\big)\big]\\ &\quad= h\,f + \frac{h^2}{2}\,\big[\partial_t f + (\partial_y f)\,f\big] - h\,(b_1+b_2)\,f - h^2\,\big[b_2 c\,\partial_t f + b_2 a\,(\partial_y f)\,f\big] + \mathcal{O}(h^3)\\ &\quad= h\,\big[1-(b_1+b_2)\big]\,f + h^2\,\big[1/2 - b_2 c\big]\,\partial_t f + h^2\,\big[1/2 - b_2 a\big]\,(\partial_y f)\,f + \mathcal{O}(h^3).\end{aligned}$$

Consequently,

$$y(t+h) - \big[y(t) + h\,\Phi\big(t,y(t),h\big)\big] = \mathcal{O}(h^3)$$

is equivalent to the fact that all lower-order powers of h vanish, i.e.,

$$b_1 + b_2 = 1, \quad b_2 c = 1/2, \quad b_2 a = 1/2.$$

In particular, it follows that $a \neq 0$, $c \neq 0$, $b_2 \neq 0$, and $a = c$. This concludes the proof. □

Example 2.18 (Heun method (1900)). *Choose $a = 1 = c$. Then, Proposition 2.16*(ii) *suggests the choice $b_1 = 1/2 = b_2$. Define*

$$\Phi(t,y,h) := \frac{1}{2}\,f(t,y) + \frac{1}{2}\,f\big(t+h,\, y + h\,f(t,y)\big). \tag{2.24}$$

Karl Heun (1859–1929) was a German mathematician. He took his PhD from the University of Göttingen in 1881 and completed his habilitation at LMU München. From 1890–1902, he worked as a teacher in Berlin, before he obtained the chair of theoretical mechanics at Technische Hochschule Darmstadt, where he retired in 1922.

According to Proposition 2.16, the resulting one-step method has consistency order $p = 2$. Clearly, Lipschitz continuity of f in y yields stability (2.13)*. For the implementation, the Heun method computes*

$$\begin{aligned} k_1 &:= f(t, y), \\ k_2 &:= f(t + h,\, y + hk_1), \\ \Phi(t, y, h) &:= \frac{k_1 + k_2}{2}. \end{aligned}$$

Example 2.19 (Modified Euler method (Runge 1895)). *Choose $a = 1/2 = c$. Then, Proposition 2.16*(ii) *suggests the choice $b_1 = 0$ and $b_2 = 1$. Define*

$$\Phi(t, y, h) := f\Big(t + \frac{h}{2}\,,\, y + \frac{h}{2}\, f(t, y)\Big). \tag{2.25}$$

According to Proposition 2.16, the resulting one-step method has consistency order $p = 2$. Clearly, Lipschitz continuity of f in y yields stability (2.13)*. For the implementation, Runge's modified Euler method computes*

$$\begin{aligned} k_1 &:= f(t, y), \\ k_2 &:= f\Big(t + \frac{h}{2}\,,\, y + \frac{h}{2}\, k_1\Big), \\ \Phi(t, y, h) &:= k_2. \end{aligned}$$

Remark 2.20. *As far as the effectivity of an integrator is concerned, one should plot the error vs. the number of f-evaluations. Then, higher-order methods pay, if the solution is smooth: For a uniform time-step size h, the number of time-steps satisfies $N = (T - t_0)/h$.*

- *The explicit Euler method has N evaluations of f, and the error decays like $\mathcal{O}(h) = \mathcal{O}(N^{-1})$, i.e., we get an algebraic decay with rate 1.*
- *The modified Euler method has $2N$ evluations of f, and the error decays like $\mathcal{O}(h^2) = \mathcal{O}(N^{-2})$, i.e., we get an algebraic decay with rate 2.*

Asymptotically, the modified Euler method will thus beat the explicit Euler method with respect to accuracy and computational time.

Remark 2.21. *Second-order consistency / convergence is only visible if f (and hence y) are sufficiently smoooth. If f is not smooth, then numerical experiments with uniform meshes lead to order reductions.*

Carl Runge (1856–1927) was a German mathematician. He studied at LMU München and University of Bremen, before he took his PhD in 1880 at HU Berlin supervised by Weierstrass. He did his habilitation in 1883 at HU Berlin and became Professor at University of Hannover in 1886. In 1904, he accepted the Chair of Applied Mathematics at University of Göttingen (the first chair for applied mathematics in Germany). Runge is the father-in-law of Richard Courant.

2.5. Explicit Runge–Kutta methods. In this section, we generalize / formalize the idea of Proposition 2.16.

Definition 2.22. *Let $A \in \mathbb{R}^{m\times m}$ be strictly lower triangular (i.e., $A_{ij} = 0$ for $i \le j$), $b, c \in \mathbb{R}^m$ with $0 \le c_1 \le c_2 \le \cdots \le c_m \le 1$. Then, a one-step method with incremental function*

$$\Phi(t, y, h) := \sum_{j=1}^{m} b_j k_j, \tag{2.26}$$

where the so-called ***increments*** *satisfy that*

$$k_i = f\Big(t + c_i h,\, y + h \sum_{j=1}^{i-1} A_{ij} k_j\Big) \quad \text{for all } i = 1, \dots, m, \tag{2.27}$$

is called ***explicit m-stage Runge–Kutta method****. The y-arguments $y + h\sum_{j=1}^{i-1} A_{ij}k_j$ of the increments k_i are called* ***stages****. Usually, Runge–Kutta methods are denoted by their* ***Butcher tableau*** $\begin{array}{c|c} c & A \\ \hline & b^\top \end{array}$*. If the data are explicitly given, usually the zero entries of A are omitted (see examples below).*

Example 2.23.

- *The* ***explicit Euler method*** *has the Butcher tableau*

$$\begin{array}{c|c} 0 & \\ \hline & 1 \end{array} = \begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}.$$

- *The* ***modified Euler*** *has the Butcher tableau*

$$\begin{array}{c|cc} 0 & & \\ 1/2 & 1/2 & \\ \hline & 0 & 1 \end{array} = \begin{array}{c|cc} 0 & 0 & 0 \\ 1/2 & 1/2 & 0 \\ \hline & 0 & 1 \end{array}.$$

- *The* ***Heun method*** *has the Butcher tableau*

$$\begin{array}{c|cc} 0 & & \\ 1 & 1 & \\ \hline & 1/2 & 1/2 \end{array} = \begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & 1/2 & 1/2 \end{array}.$$

Martin Wilhelm Kutta (1867–1944) was a German mathematician. He studied mathematics at TU München. In 1901, we took his PhD with the thesis *"Beitrag zur näherungsweisen Integration totaler Differentialgleichungen"*, where he was generalizing the ideas of Runge. Later, he was associate professor at University of Jena (1909–1910), RWTH Aachen (1910–1912), before he became full professor at University of Stuttgart in 1912. He retired in 1935.

John Charles Butcher (born 1933) is a mathematician from New Zealand. He took his PhD at University of Sydney in 1961. In 1966, he became professor at the University of Auckland. He retired in 1999. He has made fundamental contributions to the mathematical understanding of Runge–Kutta methods. In 2010, Butcher was awarded the Jones Medal from the Royal Society of New Zealand for his "exceptional lifetime work on numerical methods for the solution of differential equations".

Proposition 2.24 (Consistent Runge–Kutta methods provide quadrature rules). *Let* $\begin{array}{c|c} c & A \\ \hline & b^\top \end{array}$ *be an (explicit) m-stage Runge–Kutta method with consistency order $p \geq 1$. Then, the vectors $c, b \in \mathbb{R}^m$ provide a quadrature, which is exact for polynomials $q \in \mathbb{P}_{p-1}$, i.e.,*

$$\sum_{j=1}^{m} b_j q(c_j) = \int_0^1 q(s)\,\mathrm{d}s \quad \text{for all } q \in \mathbb{P}_{p-1}. \tag{2.28}$$

Proof. Consider the integration problem $f(t,y) := t^p$ and $y(0) = 0$. The exact solution is the antiderivative $y(t) = \int_0^t t^p\,\mathrm{d}s = \frac{t^{p+1}}{p+1}$. On the one hand, it holds that

$$k_j \overset{(2.27)}{=} (t + c_j h)^p = \sum_{i=0}^{p} \binom{p}{i} t^{p-i}\, c_j^i\, h^i$$

and hence

$$\Phi(t,y,h) = \sum_{j=1}^{m} b_j k_j = \sum_{i=0}^{p} \binom{p}{i} t^{p-i}\, h^i \sum_{j=1}^{m} b_j c_j^i. \tag{2.29}$$

On the other hand, note that

$$y^{(i)}(t) = \frac{1}{p+1}\,\frac{(p+1)!}{(p+1-i)!}\,t^{p+1-i} = \frac{p!}{(p+1-i)!}\,\frac{(i-1)!}{(i-1)!}\,t^{p+1-i} = \binom{p}{i-1}\,(i-1)!\,t^{p+1-i}$$

and hence

$$y(t+h) - y(t) = \sum_{i=1}^{p+1} \frac{y^{(i)}(t)}{i!}\,h^i = \sum_{i=1}^{p+1} \binom{p}{i-1}\,\frac{1}{i}\,t^{p+1-i}\,h^i = \sum_{i=0}^{p} \binom{p}{i}\,\frac{1}{i+1}\,t^{p-i}\,h^{i+1}. \tag{2.30}$$

Combining (2.29)–(2.30), we obtain for the consistency error that

$$\mathcal{O}(h^{p+1}) = y(t+h) - \big[y(t) + h\,\Phi\big(t, y(t), h\big)\big] = \sum_{i=0}^{p} \binom{p}{i}\,t^{p-i}\,h^{i+1}\left[\frac{1}{i+1} - \sum_{j=1}^{m} b_j c_j^i\right].$$

Consequently, the lower-order powers of h must vanish, i.e.,

$$\int_0^1 s^i\,\mathrm{d}s = \frac{1}{i+1} = \sum_{j=1}^{m} b_j c_j^i \quad \text{for all } i = 0, \ldots, p-1. \tag{2.31}$$

The identity (2.31) proves that the quadrature rule is exact for all monomials $1, s, s^2, \ldots, s^{p-1}$. Due to linearity, this proves (2.28). □

Example 2.25 (Classical Runge–Kutta method RK4 (Runge 1901)). *Recall the Simpson rule*

$$\int_0^1 g\,\mathrm{d}s \approx \frac{1}{6}\big[g(0) + 4\,g(1/2) + g(1)\big], \tag{2.32}$$

which is exact for polynomials of degree 3*, i.e.,*

$$\int_0^1 q\,\mathrm{d}s = \frac{1}{6}\big[q(0) + 4\,q(1/2) + q(1)\big]. \tag{2.33}$$

Due to Proposition 2.24, we aim to extend the Simpson rule to a Runge–Kutta method of order $p = 4$*. Runge introduced the method*

$$\begin{array}{c|cccc} 0 & & & & \\ 1/2 & 1/2 & & & \\ 1/2 & 0 & 1/2 & & \\ 1 & 0 & 0 & 1 & \\ \hline & 1/6 & 1/3 & 1/3 & 1/6 \end{array}\,. \tag{2.34}$$

In explicit terms, the method reads

$$\begin{aligned} k_1 &= f(t, y), \\ k_2 &= f\Big(t + \frac{h}{2}\,,\, y + \frac{h}{2}\,k_1\Big), \\ k_3 &= f\Big(t + \frac{h}{2}\,,\, y + \frac{h}{2}\,k_2\Big), \\ k_4 &= f(t + h\,,\, y + h\,k_3), \\ \Phi(t, y, h) &= \frac{1}{6}\big[k_1 + 2\,k_2 + 2\,k_3 + k_4\big]. \end{aligned}$$

Note that (2.34) *in fact induces the Simpson rule. By Taylor expansion, one can show that this method has consistency order* $p = 4$*. Moreover, we will see below that it needs (at least)* $m = 4$ *stages to get consistency order* $p = 4$*. In this sense, RK4 is optimal.*

Example 2.26 (Another extension of the Simpson rule). *Consider the Butcher tableau*

$$\begin{array}{c|ccc} 0 & & & \\ 1/2 & 1/2 & & \\ 1 & -1 & 2 & \\ \hline & 1/6 & 2/3 & 1/6 \end{array}\,. \tag{2.35}$$

Note that (2.35) *also induces the Simpson rule. Together with Runge's method from Example 2.25, we thus see that the extension of a quadrature rule to a Runge–Kutta method is not unique. By Taylor expansion, one can show that the method* (2.35) *has consistency order* $p = 3$*. Moreover, we will see below that this consistency order is maximal (due to* $m = 3$ *stages).*

Theorem 2.27 (Stability of explicit Runge–Kutta methods). *Let* $\begin{array}{c|c} c & A \\ \hline & b^\top \end{array}$ *be an (explicit) m-stage Runge–Kutta method. Then, there exist $C_{\mathrm{stab}} > 0$ such that*

$$\forall\, t \in [t_0, T)\, \forall\, h \in (0, T-t]\, \forall\, y, \widetilde{y} \in \mathbb{R}^n : \quad \|\Phi(t, y, h) - \Phi(t, \widetilde{y}, h)\| \leq C_{\mathrm{stab}} \, \|y - \widetilde{y}\|. \tag{2.36}$$

Let $L > 0$ be the Lipschitz constant of f in y. Then, it holds that

$$C_{\mathrm{stab}} \leq L\, p(hL), \quad \text{where} \quad p(s) = \sum_{j=0}^{m-1} \mu_j s^j \in \mathbb{P}_{m-1} \tag{2.37}$$

with $\mu_j \geq 0$ and $\mu_0 = \sum_{j=1}^m |b_j|$

Proof. Recall the increments $k_i = f\big(x + c_i h,\, y + h \sum_{j=1}^{i-1} A_{ij} k_j\big)$. Let $\widetilde{k}_i$ denote the increments with respect to $\widetilde{y}$. The proof is split into two steps.

Step 1. We prove by induction on $i = 1, \ldots, m$ that

$$\|k_i - \widetilde{k}_i\| \leq L\, q_i(hL)\, \|y - \widetilde{y}\| \quad \text{with} \quad q_i(s) = \sum_{j=0}^{i-1} \lambda_j s^j \in \mathbb{P}_{i-1}, \quad \lambda_j \geq 0, \quad \lambda_0 = 1. \tag{2.38}$$

For $i = 1$, Lipschitz continuity of f in y proves that

$$\|k_1 - \widetilde{k}_1\| = \|f(t, y) - f(t, \widetilde{y})\| \leq L\, \|y - \widetilde{y}\|.$$

This proves (2.38) for $i = 1$. In the induction step $(i-1) \rightsquigarrow i$, note that

$$\begin{aligned} \|k_i - \widetilde{k}_i\| &\leq L \Big\| y - \widetilde{y} + h \sum_{j=1}^{i-1} A_{ij} (k_j - \widetilde{k}_j) \Big\| \\ &\leq L \left[\|y - \widetilde{y}\| + h \sum_{j=1}^{i-1} |A_{ij}| \, \|k_j - \widetilde{k}_j\| \right] \\ &\leq L\, \|y - \widetilde{y}\| \Bigg[\underbrace{1 + hL \sum_{j=1}^{i-1} |A_{ij}|\, q_j(hL)}_{=: q_i(hL)} \Bigg]. \end{aligned}$$

Hence, (2.38) holds for all $i = 1, \ldots, m$.

Step 2. Finally, it follows that

$$\|\Phi(t, y, h) - \Phi(t, \widetilde{y}, h)\| = \Big\| \sum_{j=1}^m b_j (k_j - \widetilde{k}_j) \Big\| \overset{(2.38)}{\leq} L \underbrace{\sum_{j=1}^m |b_j|\, q_j(hL)}_{=: p(hL)} \, \|y - \widetilde{y}\|.$$

This concludes the proof. □

Theorem 2.28 (Necessary consistency conditions for Runge–Kutta methods). *Let* $\begin{array}{c|c} c & A \\ \hline & b^\top \end{array}$ *be an (explicit) m-stage Runge–Kutta method with consistency order $p \geq 1$. Then, there hold the following statements:*

(i) $\displaystyle\sum_{j=1}^{m} b_j = 1.$

(ii) $\displaystyle\sum_{j=1}^{m} b_j c_j^\ell = \frac{1}{\ell+1}$ *for all* $\ell = 0, \dots, p-1.$

(iii) $\displaystyle\sum_{j=1}^{m} b_j (A^\ell \mathbb{1})_j = \frac{1}{(\ell+1)!}$ *for all* $\ell = 0, \dots, p-1$, *where* $\mathbb{1} = (1, \dots, 1)^\top \in \mathbb{R}^m$.

Proof. (i) follows already from (ii) and (iii) for $\ell = 0$. (ii) has already been proved in Proposition 2.24; see the identity (2.31). Hence, it only remains to prove (iii).

To prove (iii), we consider the scalar problem $y' = f(t,y) := y$ on $[0,1]$ with $y(0) = 1$. The unique solution is $y(t) = e^t$. Let $k := (k_1, \dots, k_m)^\top$ and note that

$$k_i = f\Big(t + c_i h,\, y + h \sum_{j=1}^{i-1} A_{ij} k_j\Big) = y + h \sum_{j=1}^{i-1} A_{ij} k_j = y + h\,(Ak)_i \quad \text{for all } i = 1, \dots, m.$$

In vector form, this identity reads

$$(I - hA)\, k = k - h\, Ak = y\, \mathbb{1}.$$

For small h, it holds that $h\,\|A\| < 1$. Hence, the so-called Neumann series proves that

$$I - hA \quad \text{is invertible with} \quad (I - hA)^{-1} = \sum_{\ell=0}^{\infty} (hA)^\ell.$$

Therefore,

$$\Phi(t, y, h) = \sum_{j=1}^{m} b_j k_j = b \cdot (I - hA)^{-1}(y\mathbb{1}) = b \cdot \sum_{\ell=0}^{\infty} (hA)^\ell (y\mathbb{1}) = y \sum_{\ell=0}^{\infty} h^\ell b \cdot (A^\ell \mathbb{1}).$$

On the other hand, Taylor expansion for $y(t) = e^t$ gives

$$y(t+h) - y(t) = \sum_{\ell=1}^{\infty} \frac{y^{(\ell)}(t)}{\ell!}\, h^\ell = y(t) \sum_{\ell=1}^{\infty} \frac{h^\ell}{\ell!} = y(t) \sum_{\ell=0}^{\infty} \frac{h^{\ell+1}}{(\ell+1)!}.$$

Combining these two identities, we obtain for the consistency error that

$$\mathcal{O}(h^{p+1}) = y(t+h) - \big[y(t) + h\,\Phi\big(t, y(t), h\big)\big] = y(t) \sum_{\ell=0}^{\infty} h^{\ell+1} \Big[\frac{1}{(\ell+1)!} - b \cdot (A^\ell \mathbb{1})\Big].$$

Hence, lower-order powers of h must vanish, i.e.,

$$\frac{1}{(\ell+1)!} = b \cdot (A^\ell \mathbb{1}) \quad \text{for all } \ell = 0, \dots, p-1.$$

This concludes the proof. □

Corollary 2.29 (Naive Butcher barriers for explicit Runge–Kutta methods). *For any explicit m-step Runge–Kutta method, the consistency order $p \geq 1$ satisfies that $p \leq m$.*

Proof. Let $\begin{array}{c|c} c & A \\ \hline & b^\top \end{array}$ be the Butcher tableau of an explicit m-stage Runge–Kutta method. Recall that $A = (A_{ij}) \in \mathbb{R}^{m \times m}$ is strictly lower triangular, i.e., $A_{ij} = 0$ for all $i \leq j$. Let $A^\ell = (A_{ij}^{(\ell)})$. We show by induction on $\ell = 1, \dots, m$ that

$$A_{ij}^{(\ell)} = 0 \quad \text{for all } i \leq j + \ell - 1. \tag{2.39}$$

Obviously, the claim is OK for $\ell = 1$. For the induction step $(\ell - 1) \rightsquigarrow \ell$, note that

$$A_{ij}^{(\ell)} = \sum_{k=1}^{m} A_{ik} A_{kj}^{(\ell-1)} = \sum_{k=1}^{i-1} A_{ik} A_{kj}^{(\ell-1)} \overset{(2.39)}{=} \sum_{k=j+\ell-1}^{i-1} A_{ik} A_{kj}^{(\ell-1)}.$$

The latter sum is empty (and hence 0) if $i - 1 < j + \ell - 1$, i.e., $i < j + \ell$. This concludes the proof of (2.39).

For $\ell = m$, it follows from (2.39) that $A^m = 0$. Hence, Theorem 2.28(iii) cannot be satisfied for $\ell = m$. Therefore, we conclude that $p \leq m$. □

Exercise 2.30. *Let $\begin{array}{c|c} c & A \\ \hline & b^\top \end{array}$ be an (explicit) m-stage Runge–Kutta method. Show that the method has at least consistency order $p = 1$, if and only if $\sum_{j=1}^{m} b_j = 1$.*

Exercise 2.31. *Let $\begin{array}{c|c} c & A \\ \hline & b^\top \end{array}$ be an (explicit) m-stage Runge–Kutta method. Suppose that*

$$\sum_{j=1}^{m} b_j = 1, \quad \sum_{j=1}^{m} b_j c_j = \frac{1}{2}, \quad \text{and} \quad \sum_{j=1}^{i-1} A_{ij} = c_i \quad \text{for all } i = 1, \dots, m. \tag{2.40}$$

Argue as for Proposition 2.16 to show that the considered Runge–Kutta method has at least consistency order $p = 2$.

Remark 2.32. *In Exercise 2.31, we have seen that the consistency conditions of Theorem 2.28 are (essentially) sharp for $m = 2 = p$. However, it is known that the Butcher barriers $p \leq m$ are* not sharp *in general. For $m \geq 5$, there are no explicit Runge–Kutta methods with consistency order $p = m$, i.e., it holds that $m - p \geq 1$; see* [But08, Theorem 324B]. *Moreover $m - p \geq 2$ for $p \geq 7$ and $m - p \geq 3$ for $p \geq 8$; see* [But08, page 188].

The precise growth of the maximal order $p_{\max}(m)$ with respect to m as well as the the minimal stage number $m_{\min}(p)$ with respect to p are still unknown, however $p_{\max}(m) \to \infty$ as $m \to \infty$; see [But08, Theorem 324C].

The monograph [SWP12, Satz 2.4.6] *refers to the original literature for the table*

p	1	2	3	4	5	6	7	8	$p \geq 9$
$m_{\min}(p)$	1	2	3	4	6	7	9	11	$m_{\min}(p) \geq p+3$

Conversely,

m	1	2	3	4	5	6	7	8	9	$m \geq 9$
$p_{\max}(m)$	1	2	3	4	4	5	6	6	7	$p_{\max}(m) \leq m-2$

For $p = 10$, Hairer (1978) found an explicit Runge–Kutta method with $m = 17$. If $m_{\min}(10) < 17$ appears to be still open.

A nice motivation for RK methods is as follows: We employ quadrature on $[0,1]$ to approximate

$$\begin{aligned} y(t+h) = y(t) + \int_t^{t+h} y'(s)\,\mathrm{d}s &= y(t) + h\int_0^1 y'(t+sh)\,\mathrm{d}s \\ &= y(t) + h\int_0^1 f\big(t+sh, y(t+sh)\big)\,\mathrm{d}s \\ &\approx y(t) + h\sum_{j=1}^{m} b_j \underbrace{f\big(t+c_jh, y(t+c_jh)\big)}_{\approx k_j} \end{aligned}$$

This idea is also reflected by the following exercise.

Exercise 2.33. *Let* $\begin{array}{c|c} c & A \\ \hline & b^\top \end{array}$ *be an (explicit) m-stage Runge–Kutta method. Suppose that*

$$\sum_{j=1}^{m} b_j = 1 \quad \text{and} \quad \sum_{j=1}^{i-1} A_{ij} = c_i \quad \text{for all } i = 1,\dots,m. \tag{2.41}$$

Let $f \in C^1([t_0,T]\times\mathbb{R}^n;\mathbb{R}^n)$ and suppose that f is Lipschitz continuous in y. Prove that

$$y'(t+c_jh) - k_j = f\big(t+c_jh, y(t+c_jh)\big) - k_j = \mathcal{O}(h^2). \tag{2.42}$$

2.6. Adaptive time-step control. In this section, we aim to (heuristically) design an algorithm of the following type:

Algorithm 2.34 (Theoretical adaptive algorithm).

Input: *time interval $[t_0,T]$, initial value y_0, right-hand side $f(\cdot,\cdot)$, one-step method with incremental function $\Phi(t,y,h)$, tolerance $\tau > 0$, initial time-step size h_0, counter $\ell = 0$.*

`REPEAT`

Ernst Hairer (born 1949) is an Austrian mathematician. He took is PhD in 1972 at University of Innsbruck (supervized by Gerhard Wanner). In 1985, he became Associate Professor at University of Geneva, Switzerland. Since 1999, he is full professor at University of Geneva. He has coauthored the two-volume monograph "Solving Ordinary Differential Equations", which is the main reference for research in the field of numerical integrators. Ernst Hairer is the father of the mathematician Martin Hairer, who won the fields medal in 2014.

- *Determine $h > 0$ such that $t + h \le T$ and $\|z(t_\ell + h) - \big[y_\ell + h\,\Phi(t_\ell, y_\ell, h)\big]\| \approx \tau\, h$, where z solves $z(t_\ell) = y_\ell$ and $z' = f(t, z)$ in $[t_\ell, T]$.*
- *Define $h_\ell := h, \quad t_{\ell+1} := t_\ell + h_\ell, \quad y_{\ell+1} := y_\ell + h_\ell\,\Phi(t_\ell, y_\ell, h_\ell)\big]$.*
- *Update counter $\ell \mapsto \ell + 1$.*

`UNTIL` $t_\ell = T$

Output: *mesh $\Delta = \{t_0, \dots, t_N = T\}$, approximations $y_\ell \approx y(t_\ell)$ for all $\ell = 0, \dots, N$.*

Remark 2.35 (Why adaptive time-step control?).

- ***Efficiency:*** *We want to compute / approximate $y(T)$ as cheaply as possible, i.e., with as few evaluations of f as possible such that the error satisfies that $\|y(T) - y_N\| \approx \tau$.*
- ***Reliability:*** *Even if the solution is smooth, the error behavior $\|y(T) - y_N\| = \mathcal{O}(h_\Delta^p)$ hides a multiplicative constant. Possibly, a chosen uniform step-size $h = h_\Delta > 0$ is to coarse to ensure that $\|y(T) - y_N\| \approx \tau$.*
- *If the solution y is non-smooth, then this will spoil the experimental convergence order. For uniform meshes, one usually obtains a convergence $\mathcal{O}(h^q) = \mathcal{O}(N^{-q})$, where $q > 0$ is (much) smaller than the consistency order $p \ge 1$. However, if the mesh is locally adapted to the points of reduced regularity (i.e., the* singularities*), in many cases one can recover $\|y(T) - y_N\| = \mathcal{O}(N^{-p})$ for the error.*

Remark 2.36 (What is the challenge?). *Since the exact solutions to ODEs are in general unknown, the algorithm cannot evaluate the consistency error to ensure that*

$$\|z(t_\ell + h) - \big[y_\ell + h\,\Phi(t_\ell, y_\ell, h)\big]\| \approx \tau\, h,$$

where z solves $z(t_\ell) = y_\ell$ and $z' = f(t, z)$ in $[t_\ell, T]$. Therefore, we require appropriate heuristics to circumvent this lack of knowledge.

Heuristics 2.37 (Step-size control based on an $p - (p+1)$ strategy). *Let $\Phi(t, y, h)$ be the incremental function of an explicit one-step method with consistency order $p \ge 1$. Let $\widetilde{\Phi}(t, y, h)$ be the incremental function of an explicit one-step method with consistency order $p + 1$. Let $h > 0$ be given and $\Phi(t, y, h)$ as well as $\widetilde{\Phi}(t, y, h)$ be computed. Then, (up to higher-order terms)*

$$H := \left[\frac{\tau}{\|\Phi\big(t, z(t), h\big) - \widetilde{\Phi}\big(t, z(t), h\big)\|}\right]^{1/p} h \tag{2.43a}$$

would be OK to ensure that

$$z(t + H) - \big[z(t) + H\,\Phi\big(t, z(t), H\big)\big] \approx H\,\tau. \tag{2.43b}$$

In explicit terms: If you compute with step-size h, you get a feedback on the appropriate step-size H.

To see that Heuristics 2.37 makes sense, recall that consistency proofs are obtained by means of the Taylor expansion. For instance, let us consider the explicit Euler method.

If the solution is smooth, then

$$z(t+h) = z(t) + h\,z'(t) + \frac{h^2}{2}\,z''(t) + \mathcal{O}(h^3)$$
$$= z(t) + h\,\Phi\big(t, z(t), h\big) + \Big[\frac{h^2}{2}\,z''(t) + \mathcal{O}(h^3)\Big].$$

For a general one-step method with consistency order $p \geq 1$, one can similarly show that

$$z(t+h) = z(t) + h\,\Phi\big(t, z(t), h\big) + h^{p+1}\,c(t) + \mathcal{O}(h^{p+2}), \tag{2.44}$$

where $c(t) \sim z^{(p+1)}(t)$. Define

$$z_1 := z(t) + h\,\Phi\big(t, z(t), h\big),$$
$$\widetilde{z}_1 := z(t) + h\,\widetilde{\Phi}\big(t, z(t), h\big).$$

Then, it holds that

$$z(t+h) - z_1 = h^{p+1}\,c(t) + \mathcal{O}(h^{p+2}),$$
$$z(t+h) - \widetilde{z}_1 = \mathcal{O}(h^{p+2}),$$

where we recall that $\widetilde{\Phi}(t, y, h)$ belongs to a method of order $p+1$. Hence,

$$z_1 - \widetilde{z}_1 = \big[z(t+h) - \widetilde{z}_1\big] - \big[z(t+h) - z_1\big] = -c(t)\,h^{p+1} + \mathcal{O}(h^{p+2}).$$

We obtain that

$$\big\|z_1 - \widetilde{z}_1\big\| = \big\|c(t)\big\|\,h^{p+1} + \mathcal{O}(h^{p+2})$$

and consequently

$$\big\|c(t)\big\| = \frac{\big\|z_1 - \widetilde{z}_1\big\|}{h^{p+1}} + \mathcal{O}(h). \tag{2.45}$$

With the ansatz

$$\tau\,H \stackrel{!}{=} \Big\|z(t+H) - \big[z(t) + H\,\Phi\big(t, z(t), H\big)\big]\Big\| \stackrel{(2.44)}{=} \big\|c(t)\big\|\,H^{p+1} + \mathcal{O}(H^{p+2})$$
$$\stackrel{(2.45)}{=} \frac{\big\|z_1 - \widetilde{z}_1\big\|}{h^{p+1}} H^{p+1} + \mathcal{O}(hH^{p+1}) + \mathcal{O}(H^{p+2}).$$

If we neglect the higher-order terms $\mathcal{O}(hH^{p+1})$ and $\mathcal{O}(H^{p+2})$, we obtain that

$$\tau\,H \approx \frac{\big\|z_1 - \widetilde{z}_1\big\|}{h^{p+1}} H^{p+1}.$$

Rearranging this estimate, we are led to

$$H^p \approx \frac{\tau\,h^{p+1}}{\big\|z_1 - \widetilde{z}_1\big\|} = \frac{\tau\,h^p}{\big\|\Phi\big(t, z(t), h\big) - \widetilde{\Phi}\big(t, z(t), h\big)\big\|},$$

where we recall that $z_1 - \widetilde{z}_1 = h\,\big[\Phi\big(t, z(t), h\big) - \widetilde{\Phi}\big(t, z(t), h\big)\big]$. This proves (2.43).

Algorithm 2.38 (Practical adaptive algorithm).

Input: *time interval* $[t_0, T]$, *initial value* y_0, *right-hand side* $f(\cdot,\cdot)$, *one-step method* $\Phi(t, y, h)$ *of order* $p \geq 1$, *auxiliary one-step method* $\widetilde{\Phi}(t, y, h)$ *of order* $p+1$, *tolerance* $\tau > 0$, *initial time-step size* $h > 0$, *minimal time-step size* $h_{\min} > 0$, *conformity factor* $\lambda \geq 1$, *safety factor* $0 < \varrho \leq 1$, *counter* $\ell = 0$.

REPEAT

[1] $h := \min\big\{T - t_\ell,\, \max\{h_{\min}, h\}\big\}$

[2] $F := \widetilde{\Phi}\big(t_\ell, y_\ell, h\big)$

[3] $H := \varrho \left[\dfrac{\tau}{\|\Phi\big(t_\ell, y_\ell, h\big) - F\|}\right]^{1/p} h$

[4] IF $h \le H$ OR $h \le h_{\min}$

[5] $t_{\ell+1} := t_\ell + h$

[6] $y_{\ell+1} := y_\ell + hF$

[7] IF $t_{\ell+1} < T$

[8] $h := \min\{H,\, \lambda h\}$

[9] *Update counter* $\ell \mapsto \ell + 1$

[10] END IF

[11] ELSE

[12] $h := \min\{H,\, h/\lambda\}$

[13] END IF

UNTIL $t_{\ell+1} = T$

Output: *mesh* $\Delta = \{t_0, \dots, t_N = T\}$ *and corresponding approximations* $y_\ell \approx y(t_\ell)$ *for all* $\ell = 0, \dots, N$*).*

Remark 2.39 (Comments on Algorithm 2.38).

Line [1] and [4]: *To ensure that the algorithm is finite, we have to guarantee that* $h \not\to 0$*. In fact, the algorithm ensures that* $h \ge h_{\min}$ *up to the final time-step (where possibly* $T - t < h_{\min}$*).*

Line [3] (and [2]): *Up to the safety factor* $0 < \varrho \le 1$*, we use the formula for* H *derived in* (2.43)*. The safety factor tries to cover the fact that* (2.43) *was derived by neglecting higher-order terms (which could have an impact in our computation).*

Line [4]: *In order to avoid* f*-evaluations, we accept the time-step, if* $h \le H$*, i.e., if the current step-size is smaller than the step-size allowed by our heuristics* (2.43) *(i.e., a longer time-step would have been OK).*

Line [6] (and [2]): *Usually* $\widetilde{\Phi}(t_\ell, y_\ell, h)$ *is more accurate. Therefore, we use* $\widetilde{\Phi}(t_\ell, y_\ell, h)$ *instead of* $\Phi(t_\ell, y_\ell, h)$ *for the next time step.*

Line [7] and [8]: *If the time-step was accepted, but the final time* T *has not been reached, the algorithm requires a guess for the step-size in the next time-step. To this end, we aim to choose* $h = H$*, but we enforce that the growth of the step-size is not too big (i.e., the ratio is bounded by* λ *if the step-size is increased). Again, we intend to avoid* f*-evaluations. Therefore, the step-size guess should be accepted in the next time-step.*

Line [11] and [12]: *If* $H < h$*, then we cannot accept the time-step* $t_\ell + h$ *and we have to recompute with a smaller* h*. To this end, we aim to choose* $h = H$*, but enforce at least a uniform reduction of the current step-size (by the factor* λ^{-1}*).*

Attention: *Since we have made certain simplifications to get the error estimate* (2.43) *for the consistency error, we cannot rigorously guarantee that the adaptive algorithm is mathematically reliable in the sense that* $\|y(T) - y_N\| \le \tau$ *is guaranteed.*

Remark 2.40. *Practical choices of the parameters are the following:*

- $h \sim \tau^{1/p}$, *e.g.,* $h := \tau^{1/p}/10$, *since* $\tau \approx \|y(T) - y_N\| = \mathcal{O}(h^p)$ *for smooth* y.
- $h_{\min} = \tau$.
- $\lambda = 2$.
- $\varrho = 0.8$.

Remark 2.41. *For the implementation of Algorithm 2.38, one aims to choose* $\Phi(t, y, h)$ *and* $\widetilde{\Phi}(t, y, h)$ *in a way that minimizes the number of* f*-evaluations. Practically relevant are so-called* ***embedded Runge–Kutta methods****, which use the same increments* k_j *(and* c_j*), but differ only for the vector* $b \in \mathbb{R}^m$*. They are usually denoted by*

$$\begin{array}{c|c} c & A \\ \hline & b^\top \\ & \beta^\top \end{array}$$

where $b \in \mathbb{R}^m$ *belongs to the higher-order method and* $\beta \in \mathbb{R}^m$ *belongs to the lower-order method, i.e.,*

$$\Phi(t, y, h) = \sum_{j=1}^{m} \beta_j k_j \quad \text{and} \quad \widetilde{\Phi}(t, y, h) = \sum_{j=1}^{m} b_j k_j.$$

Example 2.42 (Bogacki–Shampine pair RK3(2)). *In 1989, Bogacki and Shampine proposed the embedded 4-stage Runge–Kutta method*

$$\begin{array}{c|cccc} 0 & & & & \\ 1/2 & 1/2 & & & \\ 3/4 & 0 & 3/4 & & \\ 1 & 2/9 & 1/3 & 4/9 & \\ \hline & 2/9 & 1/3 & 4/9 & 0 \\ & 7/24 & 1/4 & 1/3 & 1/8 \end{array}$$

The vector $b = (2/9, 1/3, 4/9, 0)^\top$ *belongs to a third-order method, while the vector* $\beta = (7/24, 1/4, 1/3, 1/8)^\top$ *belongs to a second-order method.*

For one time-step, the method needs 4 f*-evaluations instead of* $3 + 2 = 5$ *for non-embedded methods. Moreover, after the first time-step, RK3(2) has only* 3 f*-evaluations because of the* ***FSAL property*** *(i.e., first same as last): Since the last row of* A *coincides with* b *(together with* $c_1 = 0$ *and* $c_4 = 1$*), the last increment of the current time-step coincides with the first increment of the next time-step (i.e.,* $k_1(t_{\ell+1}) = k_4(t_\ell)$*).*

An adaptive Bogacki–Shampine method is provided by MATLAB function `ode23`.

Przemyslaw Bogacki, Professor at Old Dominion University, Virginia, USA; see [webpage]

Lawrence F. Shampine, Professor Emeritus at Southern Methodist University, Texas, USA; see [webpage]

Example 2.43 (Dormand–Prince pair RK5(4)). *In 1980, Dormand and Prince proposed the embedded 7-stage Runge–Kutta method*

$$
\begin{array}{c|ccccccc}
0 & & & & & & & \\
1/5 & 1/5 & & & & & & \\
3/10 & 3/40 & 9/40 & & & & & \\
4/5 & 44/45 & -56/15 & 32/9 & & & & \\
8/9 & 19372/6561 & -25360/2187 & 64448/6561 & -212/729 & & & \\
1 & 9017/3168 & -355/33 & 46732/5247 & 49/176 & -5103/18656 & & \\
1 & 35/384 & 0 & 500/1113 & 125/192 & -2187/6784 & 11/84 & \\
\hline
 & 35/384 & 0 & 500/1113 & 125/192 & -2187/6784 & 11/84 & 0 \\
 & 5179/57600 & 0 & 7571/16695 & 393/640 & -92097/339200 & 187/210 & 1/40
\end{array}
$$

The vector $b = (35/384, 0, 500/1113, 125/192, -2187/6784, 11/84, 0)^\top \in \mathbb{R}^7$ *belongs to the fifth-order method, while* $\beta = (5179/57600, 0, 7571/16695, 393/640, -92097/339200, 187/210, 1/40)^\top$ *belongs to the fourth-order method.*

For one time-step, the method needs 7 f*-evaluations instead of* $5 + 4 = 9$ *for non-embedded methods. As the Bogacki–Shampine pair RK3(2) from Example 2.42, RK5(4) has the FASL property and hence requires only* 6 f*-evaluatiosn after the first time-step.*

An adaptive Dormand–Prince method is provided by MATLAB function `ode45`.

2.7. Extrapolation.

If no auxiliary one-step method of order $p + 1$ is at hand, one can use the Richardson extrapolation to obtain the higher-order method.

Heuristics 2.44 (Two-step Richardson extrapolation of one-step method). *Let* $\Phi(t, y, h)$ *be the incremental function of an explicit one-step method with consistency order* $p \geq 1$*. Let*

$$z_1 := z(t) + h\,\Phi\big(t, z(h), h\big) \tag{2.46}$$

be one step of this method with step-size h*. Moreover, let*

$$\widehat{z}_1 := \widehat{z}_{1/2} + \frac{h}{2}\,\Phi\big(t + h/2,\, \widehat{z}_{1/2},\, h/2\big) \quad \textit{with} \quad \widehat{z}_{1/2} := z(t) + \frac{h}{2}\,\Phi\big(t, z(t),\, h/2\big) \tag{2.47}$$

be the result of two successive steps with step-size $h/2$*. Then, it holds that*

$$z(t + h) - \frac{2^p\,\widehat{z}_1 - z_1}{2^p - 1} = \mathcal{O}(h^{p+2}). \tag{2.48}$$

This procedure thus gives rise to an auxiliary one-step method of order $p + 1$*.*

John R. Dormand: According to his book on "Numerical methods for differential equations", Dormand took his PhD in Physics at the Unviersity of York and then used to be a senior lecturer at the Department of Mathematics and Statistics, Teesside Polytechnique, Middlesbrough, Cleveland, UK. According to Scopus, his 1980 paper on RK5(4) has been cited more than 1400 times. His last paper appeared in 2003.

Peter J. Prince used to work at Department of Mathematics and Statistics, Teesside Polytechnique, Middlesbrough, Cleveland, UK. From 1978–1999, he was publishing papers on Runge–Kutta methods (mainly together with John R. Dormand).

To see that Heuristics 2.45 makes sense, we argue as before: Recall that consistency proofs are obtained by means of the Taylor expansion. For instance, let us consider the explicit Euler method. If the solution is smooth, then

$$z(t+h) = z(t) + h\,\Phi\big(t, z(t), h\big) + \Big[\frac{h^2}{2}\,z''(t) + \mathcal{O}(h^3)\Big].$$

For a general one-step method with consistency order $p \geq 1$, one can similarly show that

$$z(t+h) = z(t) + h\,\Phi\big(t, z(t), h\big) + h^{p+1}\,c(t) + \mathcal{O}(h^{p+2}), \tag{2.49}$$

where $c(t) \sim z^{(p+1)}(t)$ is at least C^1. Define

$$z_1^\star := z(t+h/2) + \frac{h}{2}\,\Phi\big(t+h/2,\, z(t+h/2),\, h/2\big).$$

Then,

$$\begin{aligned} z(t+h) - \widehat{z}_1 &= \big[z(t+h) - z_1^\star\big] + \big[z_1^\star - \widehat{z}_1\big] \\ &\overset{(2.49)}{=} \big[(h/2)^{p+1}\,c(t+h/2) + \mathcal{O}(h^{p+2})\big] \\ &\quad + \Big[z(t+h/2) - \widehat{z}_{1/2} + \frac{h}{2}\Big(\Phi\big(t+h/2,\, z(t+h/2),\, h/2\big) - \Phi\big(t+h/2,\, \widehat{z}_{1/2},\, h/2\big)\Big)\Big]. \end{aligned}$$

If Φ is stable in the sense of (2.13), consistency order p guarantees that

$$\begin{aligned} \big\|\Phi\big(t+h/2,\, z(t+h/2),\, h/2\big) - \Phi\big(t+h/2,\, \widehat{z}_{1/2},\, h/2\big)\big\| &\leq L\,\|z(t+h/2) - \widehat{z}_{1/2}\| \\ &= \mathcal{O}(h^{p+1}). \end{aligned}$$

Hence, we get that

$$\begin{aligned} z(t+h) - \widehat{z}_1 &= (h/2)^{p+1}\,c(t+h/2) + \big[z(t+h/2) - \widehat{z}_{1/2}\big] + \mathcal{O}(h^{p+2}) \\ &\overset{(2.49)}{=} (h/2)^{p+1}\,\big[c(t+h/2) + c(t)\big] + \mathcal{O}(h^{p+2}) \\ &= 2\,(h/2)^{p+1}\,c(t) + \mathcal{O}(h^{p+2}), \end{aligned} \tag{2.50}$$

where we have finally used the Taylor expansion $c(t+h/2) = c(t) + \mathcal{O}(h)$. Together with

$$z(t+h) - z_1 \overset{(2.49)}{=} c(t)\,h^{p+1} + \mathcal{O}(h^{p+2}),$$

we are led to

$$\widehat{z}_1 - z_1 = \big[z(t+h) - z_1\big] - \big[z(t+h) - \widehat{z}_1\big] = c(t)\,h^{p+1}\,\big[1 - 2^{-p}\big] + \mathcal{O}(h^{p+2}).$$

This yields that

$$c(t) = \frac{\widehat{z}_1 - z_1}{1 - 2^{-p}}\,h^{-(p+1)} + \mathcal{O}(h) \tag{2.51}$$

and hence

$$\begin{aligned} z(t+h) - \widehat{z}_1 &\overset{(2.50)}{=} 2\,(h/2)^{p+1}\,c(t) + \mathcal{O}(h^{p+2}) \\ &\overset{(2.51)}{=} 2\,(h/2)^{p+1}\,\frac{\widehat{z}_1 - z_1}{1-2^{-p}}\,h^{-(p+1)} + \mathcal{O}(h^{p+2}) \\ &= 2^{-p}\,\frac{\widehat{z}_1 - z_1}{1-2^{-p}} + \mathcal{O}(h^{p+2}) \\ &= \frac{\widehat{z}_1 - z_1}{2^p - 1} + \mathcal{O}(h^{p+2}). \end{aligned} \tag{2.52}$$

Since

$$\widehat{z}_1 + \frac{\widehat{z}_1 - z_1}{2^p-1} = \frac{(2^p-1)\,\widehat{z}_1 + (\widehat{z}_1 - z_1)}{2^p-1} = \frac{2^p\,\widehat{z}_1 - z_1}{2^p-1},$$

we conclude that

$$z(t+h) - \frac{2^p\,\widehat{z}_1 - z_1}{2^p-1} = z(t+h) - \Big[\widehat{z}_1 + \frac{\widehat{z}_1 - z_1}{2^p-1}\Big] \overset{(2.52)}{=} \mathcal{O}(h^{p+2}),$$

i.e., the extrapolated method has consistency order $p+1$.

Heuristics 2.45 (Step-size control based on an $h - h/2$ strategy). *We use the notation of Heuristics 2.44. For given $h > 0$, let z_1 and $\widehat{z}_1$ be computed by* (2.46)–(2.47)*. Then, (up to higher-order terms)*

$$H := \left[\frac{2^p-1}{2^p}\,\frac{\tau\,h^{p+1}}{\|z_1 - \widehat{z}_1\|}\right]^{1/p} \tag{2.53a}$$

would be OK to ensure that

$$z(t+H) - \big[z(t) + H\,\Phi\big(t, z(t), H\big)\big] \approx H\,\tau. \tag{2.53b}$$

In explicit terms: If you compute with step-size h, you get a feedback on the appropriate step-size H. In particular, we can also employ (2.53) *to steer Algorithm 2.38.*

We build on Heuristics 2.37 and use the auxiliary method provided by Heuristics 2.44: Note that

$$\begin{aligned} \Phi\big(t, z(t), h\big) - \widetilde{\Phi}\big(t, z(t), h\big) = z_1 - \frac{2^p\,\widehat{z}_1 - z_1}{2^p-1} &= \frac{(2^p-1)\,z_1 - (2^p\,\widehat{z}_1 - z_1)}{2^p-1} \\ &= \frac{2^p}{2^p-1}\,(z_1 - \widehat{z}_1). \end{aligned}$$

Therefore, (2.43) implies (2.53).

3. Implicit one-step methods

Throughout this section, we consider the following model problem: Let $[t_0, T]$ be a given time-interval. For given $n \in \mathbb{N}$, let $f \in C([t_0,T]\times\mathbb{R}^n;\mathbb{R}^n)$ and suppose that f is Lipschitz continuous in y, i.e.,

$$\forall\, t \in [t_0,T]\,\forall\, y, \widetilde{y} \in \mathbb{R}^n: \quad \|f(t,y) - f(t,\widetilde{y})\| \le L\,\|y - \widetilde{y}\|, \tag{3.1}$$

where $\|\cdot\|$ is an arbitrary norm on $\mathbb{R}^n$ and $L > 0$ is a fixed constant. Then, for any initial value $y_0 \in \mathbb{R}^n$, the Picard–Lindelöf theorem guarantees existence and uniqueness of $y \in C^1([t_0, T]; \mathbb{R}^n)$ such that

$$y(t_0) = y_0 \quad \text{and} \quad y'(t) = f(t, y(t)) \quad \text{for all } t \in [t_0, T]. \tag{3.2}$$

Let $\Delta = \{t_0 < t_1 < \cdots < t_N = T\}$ be a given mesh with local mesh-sizes $h_\ell := t_{\ell+1} - t_\ell > 0$ for all $\ell = 0, \dots, N-1$ and maximum mesh-size $h_\Delta := \max_{\ell=0,\dots,N-1} h_\ell$. As before, our task is to compute approximations

$$y_\ell \approx y(t_\ell) \quad \text{for all } \ell = 1, \dots, N. \tag{3.3}$$

3.1. Fundamentals. This section aims to briefly transfer concepts and results from *explicit* one-step methods to *implicit* one-step methods.

Definition 3.1. *For a given* ***incremental function*** $\Phi : [t_0, T] \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}_{>0} \to \mathbb{R}^n$, *the inductive procedure*

$$y_{\ell+1} := y_\ell + h_\ell \, \Phi(t_\ell, y_\ell, y_{\ell+1}, h_\ell) \quad \textit{for all } \ell = 0, \dots, N-1 \tag{3.4}$$

is called ***implicit one-step method****.*

Exercise 3.2. *Suppose that the incremental function $\Phi(t, y, z, h)$ is Lipschitz continuous in z, i.e.,*

$$\begin{aligned} &\exists\, L > 0 \;\forall\, t \in [t_0, T] \;\forall\, y \in \mathbb{R}^n \;\forall\, z, \widetilde{z} \in \mathbb{R}^n \;\forall\, h \in (0, T-t) : \\ &\qquad \|\Phi(t, y, z, h) - \Phi(t, y, \widetilde{z}, h)\| \le L \, \|z - \widetilde{z}\|. \end{aligned} \tag{3.5}$$

Then, it holds that

$$\begin{aligned} &\exists\, H > 0 \;\forall\, t \in [t_0, T) \;\forall\, y \in \mathbb{R}^n \;\forall\, 0 < h \le \min\{T - t, H\} \;\exists!\, z \in \mathbb{R}^n : \\ &\qquad z = y + h \, \Phi(t, y, z, h). \end{aligned} \tag{3.6}$$

In particular, for sufficiently small $h_\Delta > 0$, the implicit one-step method (3.4) *is well-defined.*

Remark 3.3. (i) *While the proof of the last exercise is based on the Banach fixpoint theorem, one rather uses the Newton method than the Banach fixpoint iteration to compute $y_{\ell+1}$ from* (3.4) *in practice. Usually, the fixpoint iteration requires the restrictive condition $h_\Delta L < 1$, while the Newton method converges (in practice!) under much weaker conditions.*

(ii) *If $\Phi(t, y, z, h) = M(t, y, h) z + b(t, y, h)$ is affine in z, then one has to solve only one linear system*

$$\big(I - h_\ell \, M(t_\ell, y_\ell, h_\ell)\big) y_{\ell+1} = y_\ell + h_\ell \, b(t_\ell, y_\ell, h_\ell)$$

to compute the solution $y_{\ell+1}$ of (3.4).

Definition 3.4. *The implicit one-step method corresponding to the incremental function $\Phi(t,y,z,h)$ has consistency order $p \geq 1$, if the following is satisfied: Given $f \in C^p([t_0,T]\times\mathbb{R}^n;\mathbb{R}^n)$, the exact solution $y \in C^{p+1}([t_0,T];\mathbb{R}^n)$ of (3.2) satisfies that*

$$\begin{aligned} &\exists\, C>0\, \forall\, t\in[t_0,T)\, \forall\, h\in(0,T-t]: \\ &\qquad \big\| y(t+h) - \big[y(x) + h\,\Phi(t,y(t),y(t+h),h)\big]\big\| \leq C\,h^{p+1}. \end{aligned} \tag{3.7}$$

Example 3.5 (Implicit Euler method). *Recall that*

$$y_{\ell+1} := y_\ell + h_\ell\, f(t_{\ell+1}, y_{\ell+1}), \quad \textit{i.e.,} \quad \Phi(t,y,z,h) := f(t+h,z).$$

According to the Taylor theorem for $t = (t+h) - h$, it holds that

$$y(t) = y(t+h) - h\,y'(t+h) + \mathcal{O}(h^2).$$

Hence, we see that

$$\begin{aligned} y(t+h) - \big[y(t) + h\,\Phi(t,y(t),y(t+h),h)\big] &= y(t+h) - \big[y(t) + h\,f(t+h,y(t+h))\big] \\ &= y(t+h) - \big[y(t) + h\,y'(t+h)\big] = \mathcal{O}(h^2). \end{aligned}$$

Therefore, the implicit Euler method has consistency order $p=1$.

Example 3.6 (Implicit midpoint rule). *The implicit midpoint rule is defined as*

$$y_{\ell+1} := y_\ell + h_\ell\, f\Big(t_\ell + \frac{h_\ell}{2}, \frac{y_\ell + y_{\ell+1}}{2}\Big), \quad \textit{i.e.,} \quad \Phi(t,y,z,h) := f\Big(t+\frac{h}{2}, \frac{y+z}{2}\Big).$$

According to the Taylor theorem for $(t+h/2) \pm h/2$, it holds that

$$\begin{aligned} y(t) &= y(t+h/2) - \frac{h}{2}\,y'(t+h/2) + \frac{h^2}{8}\,y''(t+h/2) + \mathcal{O}(h^3), \\ y(t+h) &= y(t+h/2) + \frac{h}{2}\,y'(t+h/2) + \frac{h^2}{8}\,y''(t+h/2) + \mathcal{O}(h^3). \end{aligned}$$

On the one hand, this shows that

$$y(t+h) - y(t) = h\,y'(t+h/2) + \mathcal{O}(h^3) = h\,f(t+h/2, y(t+h/2)) + \mathcal{O}(h^3).$$

On the other hand, this shows that

$$\frac{y(t+h)+y(t)}{2} = y(t+h/2) + \mathcal{O}(h^2)$$

and hence

$$f(t+h/2, y(t+h/2)) = f\Big(t+h/2, \frac{y(t+h)+y(t)}{2}\Big) + \mathcal{O}(h^2).$$

Overall, this results in

$$y(t+h) - \Big[y(t) + h\,f\Big(t+\frac{h}{2}, \frac{y(t+h)+y(t)}{2}\Big)\Big] = \mathcal{O}(h^2).$$

Therefore, the implicit midpoint rule has consistency order $p=2$.

Exercise 3.7 (Trapezoidal rule). *The trapezoidal rule is defined as*

$$y_{\ell+1} := y_\ell + h_\ell \, \frac{f(t_\ell, y_\ell) + f(t_{\ell+1}, y_{\ell+1})}{2}, \quad \textit{i.e.,} \quad \Phi(t,y,z,h) := \frac{f(t,y) + f(t+h,z)}{2}.$$

Show that the trapezoidal rule has consistency order $p = 2$.

Exercise 3.8 (Stability plus consistency implies convergence). *Let $\Phi(t,y,z,h)$ be the incremental function of an implicit one-step method with consistency order $p \geq 1$. Suppose that $\Phi(t,y,z,h)$ is stable, i.e.,*

$$\begin{aligned} &\exists\, L > 0 \;\forall\, t \in [t_0, T) \;\forall\, y, \widetilde{y} \in \mathbb{R}^n \;\forall\, z, \widetilde{z} \in \mathbb{R}^n \;\forall\, h \in (0, T-t] : \\ &\qquad \big\| \Phi(t,y,z,h) - \Phi(t,\widetilde{y},\widetilde{z},h) \big\| \leq L \, \big(\|y - \widetilde{y}\| + \|z - \widetilde{z}\| \big). \end{aligned} \tag{3.8}$$

Then, the solution $y \in C^{p+1}([t_0,T];\mathbb{R}^n)$ satisfies that

$$\max_{\ell=1,\dots,N} \|y(t_\ell) - y_\ell\| = \mathcal{O}(h_\Delta^p), \tag{3.9}$$

whenever the discrete solutions exist (e.g., $h_\Delta \leq H$).

The following proposition shows that all implicit one-step methods are locally explicit. As far as, e.g., the adaptive step-size control is concerned, we can thus simply employ the ideas developed for explicit one-step methods (since our ideas above have built only on local Taylor expansions).

Proposition 3.9 (Implicit methods are locally explicit). *Let $\Phi(t,y,z,h)$ be the incremental function of an implicit one-step method with consistency order $p \geq 1$. Suppose that $\Phi(t,y,z,h)$ is Lipschitz continuous in z (see (3.5)) and continuously differentiable in (t,y,z). Let $h > 0$ and (t,y,z) with*

$$z = y + h\,\Phi(t,y,z,h) \quad \textit{and} \quad h\,\|D_z\Phi(t,y,z,h)\| < 1. \tag{3.10}$$

Then, there exist open sets $U \subset [0,T] \times \mathbb{R}^n$ and $V \subset \mathbb{R}^n$ with $(t,y) \in U$ and $z \in V$ as well as a function $g \in C^1(U;V)$ such that

$$\forall\, (\widetilde{t},\widetilde{y}) \in U \;\forall\, \widetilde{z} \in V : \quad \big(\widetilde{z} = \widetilde{y} + h\,\Phi(\widetilde{t},\widetilde{y},\widetilde{z},h) \quad \Longleftrightarrow \quad \widetilde{z} = \widetilde{g}(\widetilde{t},\widetilde{y}) \big). \tag{3.11}$$

In particular, one step of the one-step method (3.4) with step-size h is well-defined and even explicit, since $\Phi\big(t,y,z,h\big) = \Phi\big(t,y,g(t,y),h\big)$.

Proof. Consider $F\big((\widetilde{t},\widetilde{y}),\widetilde{z}\big) := \widetilde{z} - \big[\widetilde{q} + h\,\Phi(\widetilde{t},\widetilde{y},\widetilde{z},h)\big]$. Then, $F \in C^1([t_0,T]\times\mathbb{R}^n\times\mathbb{R}^n;\mathbb{R}^n)$ and $F\big((t,y),z\big) = 0$ by assumption. Moreover, $\kappa := h\,\|D_z\Phi(t,y,z,h)\| < 1$ allows to employ the Neumann series to see that

$$D_z F\big((t,y),z\big) = I - h\,D_z\Phi(t,y,z,h)$$

is regular with

$$\big[D_z F\big((t,y),z\big)\big]^{-1} = \sum_{k=0}^{\infty} \big(h\,D_z\Phi(t,y,z,h)\big)^k.$$

Therefore, the claim follows from the implicit function theorem. □

3.2. Implicit Runge–Kutta methods.

Definition 3.10. *Let $A \in \mathbb{R}^{m\times m}$, $b, c \in \mathbb{R}^m$ with $0 \le c_1 \le c_2 \le \cdots \le c_m \le 1$. Then, a one-step method with incremental function*

$$\Phi(t, y, h) := \sum_{j=1}^{m} b_j k_j, \tag{3.12}$$

where the so-called ***stages*** *satisfy the implicit conditions*

$$k_j = f\Big(t + c_j h,\, y + h \sum_{\ell=1}^{m} A_{j\ell} k_\ell\Big) \quad \text{for all } j = 1, \dots, m, \tag{3.13}$$

is called ***m-stage Runge–Kutta method****. A method is called* ***implicit m-stage Runge–Kutta method****, if the matrix A is* not *strictly lower triangular. Usually, Runge–Kutta methods are denoted by their* ***Butcher tableau*** $\begin{array}{c|c} c & A \\ \hline & b^\top \end{array}$.

We stress that well-posedness of an implicit Runge–Kutta method is not obvious, since the equations for the stages (3.13) are (nonlinearly) coupled and implicit. However, well-posedness is shown in the following proposition, if the step-size h is sufficiently small.

Proposition 3.11. *Let* $\begin{array}{c|c} c & A \\ \hline & b^\top \end{array}$ *be an m-stage Runge–Kutta method. Then, there exists $H > 0$ such that the stages* (3.13) *are well-defined, i.e.,*

$$\forall\, t \in [t_0, T]\ \forall\, y \in \mathbb{R}^n\ \forall\, 0 < h < \min\{T - t, H\}\ \exists!\, k_1, \dots, k_m \in \mathbb{R}^n\ \forall\, j = 1, \dots, m:$$
$$k_j = f\Big(t + c_j h,\, y + h \sum_{\ell=1}^{m} A_{j\ell} k_\ell\Big) \tag{3.14}$$

Proof. We write $K := (k_1, \dots, k_m) \in \mathbb{R}^{n\times m}$. Consider

$$\Psi : \mathbb{R}^{n\times m} \to \mathbb{R}^{n\times m}, \quad \big(\Psi(K)\big)_j := f\Big(t + c_j h,\, y + h \sum_{\ell=1}^{m} A_{j\ell} k_\ell\Big) \in \mathbb{R}^n \quad \text{for all } j = 1, \dots, m.$$

Note that the stage conditions (3.13) are equivalent to $\Psi(K) = K$. Hence, we only need to show that Ψ has a unique fixpoint. To this end, we aim to apply the Banach fixpoint theorem. Note that

$$\begin{aligned}
\|\Psi(K) - \Psi(\widetilde{K})\|_\infty &:= \max_{j=1,\dots,m} \Big\| f\Big(t + c_j h,\, y + h \sum_{\ell=1}^{m} A_{j\ell} k_\ell\Big) - f\Big(t + c_j h,\, y + h \sum_{\ell=1}^{m} A_{j\ell} \widetilde{k}_\ell\Big) \Big\| \\
&\le hL \max_{j=1,\dots,m} \Big\| \sum_{\ell=1}^{m} A_{j\ell}(k_\ell - \widetilde{k}_\ell) \Big\| \le hL \max_{j=1,\dots,m} \sum_{\ell=1}^{m} |A_{j\ell}| \max_{i=1,\dots,m} \|k_i - \widetilde{k}_i\| \\
&= \Big(hL \max_{j=1,\dots,m} \sum_{\ell=1}^{m} |A_{j\ell}| \Big) \|K\|_\infty.
\end{aligned}$$

For sufficiently small $h > 0$, the mapping Ψ is hence a contraction. Therefore, the Banach fixpoint theorem concludes the proof. □

Example 3.12. *We consider the Runge–Kutta method* $\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}$. *By definition, one step of this method leads to* $y_{\ell+1} = y_\ell + h\,k_1$, *where*

$$k_1 = f(t_\ell + h_\ell,\, y_\ell + h_\ell k_1) = f(t_{\ell+1},\, y_\ell + h_\ell k_1) = f(t_{\ell+1}, y_{\ell+1}).$$

We hence obtain that

$$y_{\ell+1} = y_\ell + h\, f(t_{\ell+1}, y_{\ell+1}),$$

which is the implicit Euler method.

Example 3.13. *We consider the Runge–Kutta method* $\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1/2 & 1/2 \\ \hline & 1/2 & 1/2 \end{array}$. *Then,*

$$\begin{aligned} k_1 &= f(t_\ell, y_\ell), \\ k_2 &= f\Big(t_\ell + h_\ell,\, y_\ell + h_\ell\, \frac{k_1 + k_2}{2}\Big), \\ y_{\ell+1} &= y_\ell + h_\ell\, \frac{k_1 + k_2}{2}. \end{aligned}$$

From the last equation, we obtain that $k_2 = f(t_{\ell+1}, y_{\ell+1})$. *Overall, we see that*

$$y_{\ell+1} = y_\ell + h_\ell \frac{f(t_\ell, y_\ell) + f(t_{\ell+1}, y_{\ell+1})}{2},$$

which is the implicit trapezoidal rule.

Example 3.14. *We consider the Runge–Kutta method* $\begin{array}{c|c} 1/2 & 1/2 \\ \hline & 1 \end{array}$. *Then,*

$$\begin{aligned} k_1 &= f\Big(t_\ell + \frac{h_\ell}{2},\, y_\ell + \frac{h_\ell}{2}\, k_1\Big), \\ y_{\ell+1} &= y_\ell + h_\ell\, k_1. \end{aligned}$$

From the last equation, we obtain that $h_\ell\, k_1 = y_{\ell+1} - y_\ell$ *and hence* $k_1 = f(t_\ell + \frac{h_\ell}{2},\, \frac{y_\ell + y_{\ell+1}}{2})$. *Overall, we see that*

$$y_{\ell+1} = y_\ell + h_\ell\, f\Big(t_\ell + \frac{h_\ell}{2},\, \frac{y_\ell + y_{\ell+1}}{2}\Big),$$

which is the implicit midpoint rule.

Exercise 3.15. *Consider a general m-stage Runge–Kutta method* $\begin{array}{c|c} c & A \\ \hline & b^\top \end{array}$. *Note that the corresponding stages* $k_j = k_j(t, y, h)$ *depend on* (t, y, h) *and that* $\Phi(t, y, h) = \sum_{j=1}^m b_j k_j$.

By exploiting the Lipschitz continuity of f in y, show that the Runge–Kutta method is stable in the sense of

$$\exists\, H > 0\, \exists\, C > 0\, \forall t \in [t_0, T]\, \forall y, \widetilde{y} \in \mathbb{R}^n\, \forall\, 0 < h < \min\{T - t, H\} : \\ \|\Phi(t, y, h) - \Phi(t, \widetilde{y}, h)\| \leq C \|y - \widetilde{y}\|. \tag{3.15}$$

Using stability and consistency, formulate and prove a convergence theorem for general Runge–Kutta methods.

Remark 3.16. *We have already proved that an m-stage Runge–Kutta method with consistency order $p \geq 1$ satisfies the following statements* (i)–(iii)*:*

(i) $\displaystyle\sum_{j=1}^{m} b_j = 1.$

(ii) $\displaystyle\sum_{j=1}^{m} b_j c_j^\ell = \frac{1}{\ell + 1}$ *for all* $\ell = 0, \ldots, p - 1.$

(iii) $\displaystyle\sum_{j=1}^{m} b_j (A^\ell \mathbb{1})_j = \frac{1}{(\ell + 1)!}$ *for all* $\ell = 0, \ldots, p - 1$, *where* $\mathbb{1} = (1, \ldots, 1) \in \mathbb{R}^m$.

Proposition 3.17 (Consistency $\leq 2m$). *Consider a general m-stage Runge–Kutta method* $\begin{array}{c|c} c & A \\ \hline & b^\top \end{array}$ *and suppose consistency order $p \geq 1$. Then, it holds that $p \leq 2m$. Moreover, if $p = 2m$, then the vectors $b, c \in \mathbb{R}^m$ are unique: The coefficients of c are the nodes of the Gaussian quadrature rule on $[0, 1]$ and that of b are the corresponding weights.*

Proof. Consistency order p yields that

$$\sum_{j=1}^{m} b_j c_j^\ell \overset{!}{=} \frac{1}{\ell + 1} = \int_0^1 t^\ell \,\mathrm{d}t \quad \text{for all } \ell = 0, \ldots, p - 1.$$

Hence, this quadrature is exact for polynomials of degree $p - 1$, i.e.,

$$\sum_{j=1}^{m} b_j q(c_j) = \int_0^1 q(t) \,\mathrm{d}t \quad \text{for all } q \in \mathbb{P}_{p-1}.$$

The claim follows from the basic lecture on numerical analysis: First, the maximum exactness for a quadrature rule with m nodes is $2m - 1$ and hence $p \leq 2m$. Second, the Gaussian quadrature is the unique quadrature rule with exactness $2m - 1$. □

Remark 3.18. *Runge–Kutta methods can be interpreted as the application of appropriate quadrature schemes employed to the integral representation of the exact solution. A simple*

substitution shows that

$$y(t+h) - y(t) = \int_t^{t+h} y'(s)\,\mathrm{d}s = h\int_0^1 y'(t+sh)\,\mathrm{d}s = h\int_0^1 f(t+sh, y(t+sh))\,\mathrm{d}s$$
$$\approx h\sum_{j=1}^m b_j f\big(t+c_jh, y(t+c_jh)\big) \approx h\sum_{j=1}^m b_jk_j.$$

To derive a formula for $k_j \approx f(t+c_jh, y(t+c_jh))$, we proceed analogously:

$$y(t+c_jh) - y(t) = \int_t^{t+c_jh} y'(s)\,\mathrm{d}s = h\int_0^{c_j} y'(t+sh)\,\mathrm{d}s = h\int_0^{c_j} f(t+sh, y(t+sh))\,\mathrm{d}s$$
$$\approx h\sum_{\ell=1}^m A_{j\ell} f\big(t+c_\ell h, y(t+c_\ell h)\big) \approx h\sum_{\ell=1}^m A_{j\ell}k_\ell.$$

With the foregoing interpretation, the following proposition gives a sufficient criterion for consistency of Runge–Kutta methods in terms of quadrature.

Proposition 3.19 (Consistency in terms of quadrature rules). *Let $p \geq 1$. Consider a general Runge–Kutta method* $\begin{array}{c|c} c & A \\ \hline & b^\top \end{array}$ *and suppose that*

$$\sum_{j=1}^m b_j q(c_j) = \int_0^1 q\,\mathrm{d}t \quad \text{for all } q \in \mathbb{P}_{p-1} \tag{3.16a}$$

as well as

$$\sum_{j=1}^m A_{\ell j} q(c_j) = \int_0^{c_\ell} q\,\mathrm{d}t \quad \text{for all } q \in \mathbb{P}_{p-2} \text{ and all } \ell = 1,\dots,m, \tag{3.16b}$$

where $\mathbb{P}_s$ denotes the space of all polynomials of degree $\leq s$. Then, the Runge–Kutta method has at least consistency order p.

Lemma 3.20 (Lagrange interpolation). *Let $g \in C^s[a,b]$ and $a \leq t_1 < \cdots < t_s \leq b$. Then, there exists a unique polynomial*

$$q \in \mathbb{P}_{s-1} \quad \text{such that} \quad q(t_j) = g(t_j) \quad \text{for all } j = 1,\dots,s, \tag{3.17}$$

and it holds that

$$q(t) = \sum_{j=1}^s g(t_j)L_j(t) \quad \text{with the } \textbf{\textit{Lagrange polynomials}} \quad L_j(t) := \prod_{\substack{k=1\\k\neq j}}^s \frac{t-t_k}{t_j-t_k}. \tag{3.18}$$

Moreover, for all $k = 0,\dots,s-1$ and all $t \in [a,b]$, there holds the error identity

$$g^{(k)}(t) - q^{(k)}(t) = \frac{g^{(s)}(\xi)}{(s-k)!}\prod_{\ell=1}^{s-k}(t-\zeta_\ell) \tag{3.19}$$

with appropriate scalars $\xi = \xi(k,t), \zeta_\ell = \zeta_\ell(k) \in [a,b]$. In particular, it follows that

$$\|g^{(k)} - q^{(k)}\|_{\infty,[a,b]} \le \frac{\|g^{(s)}\|_{\infty,[a,b]}}{(s-k)!}\,|b-a|^{s-k}. \tag{3.20}$$

Proof. The proof is split into four steps.

Step 1. Consider the operator

$$\mathrm{T} : \mathbb{P}_{s-1} \to \mathbb{R}^s, \quad \mathrm{T}p := \big(p(t_1), \dots, p(t_s)\big).$$

Clearly, T is linear and $\dim \mathbb{P}_{s-1} = s$. Hence, T is bijective if and only if it is surjective (or injective). Given $z \in \mathbb{R}^s$, define $p := \sum_{j=1}^s z_j L_j$. Note that $L_j(t_j) = 1$, while $L_j(t_k) = 0$ for all $j,k = 1,\dots,n$ with $j \neq k$. Therefore, we obtain that $\mathrm{T}p = z$, i.e., T is surjective and hence bijective. In explicit terms, we have thus shown that $p = \sum_{j=1}^s z_j L_j$ is the unique polynomial in $\mathbb{P}_{s-1}$ such that

$$\forall j = 1,\dots,s: \quad p(t_j) = z_j.$$

This concludes the proof of (3.17).

Step 2. The error $e := g - q \in C^s[a,b]$ has at least s zeros in $[a,b]$ (at the t_j). According to the mean value theorem, between two zeros of e, one has one zero of e'. Hence, $e' \in C^{s-1}[a,b]$ has at least $s-1$ zeros. Inductively, we obtain that $e^{(k)} \in C^{s-k}[a,b]$ has at least $s-k$ zeros $\zeta_1,\dots,\zeta_{s-k} \in [a,b]$. For $t = \zeta_\ell$, the error identity (3.19) is trivial. Without loss of generality, we can thus assume that $t \notin \{\zeta_1,\dots,\zeta_{s-k}\}$.

Step 3. We consider the function

$$G(x) := e^{(k)}(t)\,\omega(x) - \omega(t)\,e^{(k)}(x), \quad \text{where} \quad \omega(x) := \prod_{\ell=1}^{s-k}(x-\zeta_\ell).$$

Clearly, $G \in C^{s-k}[a,b]$ has at least $s-k+1$ zeros in $[a,b]$. Inductively, the mean value theorem shows that $G^{(s-k)}$ has at least one zero $\xi \in [a,b]$. Hence,

$$0 = G^{(s-k)}(\xi) = e^{(k)}(t)\,\omega^{(s-k)}(\xi) - \omega(t)\,e^{(s)}(\xi) = e^{(k)}(t)\,(s-k)! - \omega(t)\,g^{(s)}(\xi).$$

Rearranging this estimate, we prove the error identity (3.19).

Step 4. The final estimate (3.20) follows from $|t-\zeta_\ell| \le |b-a|$ and hence

$$|g^{(k)}(t) - q^{(k)}(t)| \overset{(3.19)}{\le} \frac{\|g^{(s)}\|_{\infty,[a,b]}}{(s-k)!}\,|b-a|^{s-k}.$$

Taking the supremum over all $t \in [a,b]$, we conclude the proof. □

Lemma 3.21 (Interpolatory quadrature). *Let $g \in C^s[a,b]$ and $a \le t_1 < \cdots < t_s \le b$. Then,*

$$\Big| \int_a^b g\,\mathrm{d}t - \sum_{j=1}^s g(t_j) \int_a^b L_j\,\mathrm{d}t \Big| \le \frac{\|f^{(s)}\|_{\infty,[a,b]}}{s!}\,|b-a|^{s+1}. \tag{3.21}$$

Proof. Let $q := \sum_{j=1}^{s} g(t_j) L_j \in \mathbb{P}_{s-1}$ and recall the Lagrange interpolation (3.17)–(3.18). Then,

$$\sum_{j=1}^{s} g(t_j) \int_a^b L_j \,\mathrm{d}t = \int_a^b q \,\mathrm{d}t$$

and hence

$$\Big| \int_a^b g \,\mathrm{d}t - \sum_{j=1}^{s} g(t_j) \int_a^b L_j \,\mathrm{d}t \Big| = \Big| \int_a^b (g-q) \,\mathrm{d}t \Big| \overset{(3.20)}{\leq} \frac{\|g^{(s)}\|_{\infty,[a,b]}}{s!} \, |b-a|^{s+1}.$$

This concludes the proof. □

Proof of Proposition 3.19. The proof is split into two steps.

Step 1. First, note that

$$\begin{aligned} y(t + c_\ell h) - y(t) &= \int_t^{t+c_\ell h} y'(s) \,\mathrm{d}s = h \int_0^{c_\ell} y'(t+sh) \,\mathrm{d}s \\ &= h \int_0^{c_\ell} f\big(t+sh, y(t+sh)\big) \,\mathrm{d}s \overset{(3.16b)}{=} h\Big(\sum_{j=1}^{m} A_{\ell j} f\big(t + c_j h, y(t+c_j h)\big) + \mathcal{O}(h^{p-1}) \Big). \end{aligned}$$

Recall that $k_j = f\big(t + c_j h, y(t) + h \sum_{i=1}^{m} A_{ji} k_i\big)$. Using the last identity, we get that

$$\begin{aligned} r_\ell &:= \Big\| y(t+c_\ell h) - \Big[y(t) + h \sum_{j=1}^{m} A_{\ell j} k_j \Big] \Big\| \\ &= h \Big\| \sum_{j=1}^{m} A_{\ell j} \Big[f\big(t+c_j h, y(t+c_j h)\big) - f\big(t + c_j h, y(t) + h \sum_{i=1}^{m} A_{ji} k_i\big) \Big] \Big\| + \mathcal{O}(h^p) \\ &\leq hL \sum_{j=1}^{m} |A_{\ell j}| \, \Big\| y(t+c_j h) - \Big[y(t) + h \sum_{i=1}^{m} A_{ji} k_i \Big] \Big\| + \mathcal{O}(h^p) \\ &\leq hL \max_{i=1,\dots,m} \sum_{j=1}^{m} |A_{ij}| \, \|r_j\| + \mathcal{O}(h^p). \end{aligned}$$

Note that the right-hand side is independent of ℓ. Therefore, we get that

$$R := \max_{\ell=1,\dots,m} r_\ell \leq \Big(hL \max_{i=1,\dots,m} \sum_{j=1}^{m} |A_{ij}| \Big) R + \mathcal{O}(h^p). \tag{3.22}$$

For sufficiently small $h > 0$, we thus obtain that $R = \mathcal{O}(h^p)$.

Step 2. Similarly, we see that

$$\begin{aligned} y(t+h) - y(t) &= \int_t^{t+h} y'(s) \,\mathrm{d}s = h \int_0^1 y'(t+sh) \,\mathrm{d}s \\ &= h \int_0^1 f\big(t+sh, y(t+sh)\big) \,\mathrm{d}s \overset{(3.16a)}{=} h\Big(\sum_{j=1}^{m} b_j f\big(t + c_j h, y(t+c_j h)\big) + \mathcal{O}(h^p) \Big). \end{aligned}$$

Finally, we can estimate the consistency error by

$$\begin{aligned}
\big\|y(t+h) - \big[y(t) + h\,\Phi(t,y(t),h)\big]\big\| &= \Big\|y(t+h) - \Big[y(t) + h\sum_{j=1}^m b_j k_j\Big]\Big\| \\
&= h\,\Big\|\sum_{j=1}^m b_j\Big[f\big(t+c_jh, y(t+c_jh)\big) - f\big(t+c_jh, y(t) + h\sum_{i=1}^m A_{ji}k_i\big)\Big]\Big\| + \mathcal{O}(h^{p+1}) \\
&\le hL\sum_{j=1}^m |b_j|\,\Big\|y(t+c_jh) - \Big[y(t) + h\sum_{i=1}^m A_{ji}k_i\Big]\Big\| + \mathcal{O}(h^{p+1}) \\
&\le hL\sum_{j=1}^m |b_j|\,R + \mathcal{O}(h^{p+1}) \overset{(3.22)}{=} \mathcal{O}(h^{p+1}).
\end{aligned}$$

This concludes the proof. □

3.3. Collocation methods.

Definition 3.22. *Let $0 \le c_1 < \cdots < c_m \le 1$. Then, the following inductive procedure is called **collocation method**: Given a time-step t_ℓ and the corresponding approximation $y_\ell \in \mathbb{R}^n$, let $q_\ell \in \mathbb{P}_m$ satisfy*

$$q_\ell(t_\ell) = y_\ell \quad \text{and} \quad q_\ell'(t_\ell + c_jh_\ell) = f\big(t_\ell + c_jh_\ell,\, q_\ell(t_\ell + c_jh_\ell)\big) \quad \text{for all } j = 1,\dots,m. \tag{3.23}$$

Then define $y_{\ell+1} := q_\ell(t_{\ell+1})$.

Note that a collocation method is a very natural strategy to solve (3.2). A collocation method provides a continuous piecewise polynomial, which satisfies the given ODE pointwise at finitely many collocation nodes $t_\ell + c_jh_\ell$. We stress that well-posedness of a collocation method is not obvious, since the interpolation conditions for q_ℓ' in (3.23) are implicit and (thus possibly) nonlinear. However, well-posedness is shown in the following proposition, if the step-size h is sufficiently small.

Proposition 3.23. *Let $0 \le c_1 < \cdots < c_m \le 1$. Then, it holds that*

$$\begin{aligned}
&\exists\, H > 0\, \forall t \in [t_0, T)\, \forall y \in \mathbb{R}^n\, \forall 0 < h \le \min\{T-t, H\}\, \exists!\, q \in \mathbb{P}_m : \\
&\qquad \Big(q(t) = y \quad \text{and} \quad \forall j = 1,\dots,m: \quad q'(t+c_jh) = f\big(t+c_jh,\, q(t+c_jh)\big)\Big).
\end{aligned} \tag{3.24}$$

In particular, there exists a unique polynomial $q \in \mathbb{P}_m$ which satisfies (3.23), if the step-size h_ℓ is sufficiently small.

Proof. The proof is split into two steps.

Step 1. We consider the operator

$$\mathrm{T} : \mathbb{P}_m \to \mathbb{R}^{m+1}, \quad \mathrm{T}q := \big(q(t), q'(t+c_1h), \dots, q'(t+c_mh)\big).$$

Clearly, T is linear and $\dim \mathbb{P}_m = m+1$. Hence, T is bijective if and only if T is injective (or surjective). Let $q \in \mathbb{P}_m$ with $\mathrm{T}q = 0$. Then, $q' \in \mathbb{P}_{m-1}$ has at least m zeros (at $t + c_jh$). Hence, the fundamental theorem of algebra shows that $q' = 0$ and hence q is

constant. Since $q(t) = 0$, it follows that $q = 0$. Overall, we have shown that T is injective and hence bijective. In explicit terms, it thus holds that

$$\forall z_0, z_1, \dots, z_m \in \mathbb{R}\, \exists!\, q \in \mathbb{P}_m : \quad \Big(q(t) = z_0 \quad \text{and} \quad \forall j = 1, \dots, m : \quad q'(t + c_j h) = z_j \Big).$$

Step 2. By abuse of notation, we abbreviate $(\mathbb{P}_m)^n$ by simply $\mathbb{P}_m$ in the following. Let $q \in \mathbb{P}_m$. According to Step 1, there exists a unique $\Psi q \in \mathbb{P}_m$ such that

$$(\Psi q)(t) = y \quad \text{and} \quad \forall j = 1, \dots, m : \quad (\Psi q)'(t + c_j h) = f\big(t + c_j h,\, q(t + c_j h)\big)$$

This gives rise to an operator $\Psi : \mathbb{P}_m \to \mathbb{P}_m$. In view of (3.24), we have to show that the mapping Ψ has as unique fixpoint, provided that h is sufficiently small.

To this end, note that $(\Psi q)' \in \mathbb{P}_{m-1}$ and Lemma 3.20 yields that

$$(\Psi q)'(s) \overset{(3.18)}{=} \sum_{j=1}^{m} f\big(t + c_j h,\, q(t + c_j h)\big)\, L_j(s).$$

For $0 \le \lambda \le 1$, it follows that

$$\begin{aligned}(\Psi q)(t + \lambda h) &= (\Psi q)(t) + \int_t^{t+\lambda h} (\Psi q)'(s)\, \mathrm{d}s = y + h \int_0^\lambda (\Psi q)'(s)\, \mathrm{d}s \\ &= y + h \sum_{j=1}^{m} f\big(t + c_j h,\, q(t + c_j h)\big) \int_0^\lambda L_j(s)\, \mathrm{d}s.\end{aligned}$$

For $q, \widetilde{q} \in \mathbb{P}_m$, it follows that

$$\begin{aligned}&\big\|(\Psi q)(t + \lambda h) - (\Psi \widetilde{q})(t + \lambda h)\big\| \\ &\quad = h \Big\| \sum_{j=1}^{m} \big[f\big(t + c_j h,\, q(t + c_j h)\big) - f\big(t + c_j h,\, \widetilde{q}(t + c_j h)\big) \big] \int_0^\lambda L_j(s)\, \mathrm{d}s \Big\| \\ &\quad \le h \sum_{j=1}^{m} \Big| \int_0^\lambda L_j(s)\, \mathrm{d}s \Big| \, \big\| f\big(t + c_j h,\, q(t + c_j h)\big) - f\big(t + c_j h,\, \widetilde{q}(t + c_j h)\big) \big\| \\ &\quad \le hL \sum_{j=1}^{m} \Big(\int_0^\lambda |L_j|\, \mathrm{d}s \Big) \big\| q(t + c_j h) - \widetilde{q}(t + c_j h) \big\| \\ &\quad \le hL \sum_{j=1}^{m} \Big(\int_0^1 |L_j|\, \mathrm{d}s \Big) \, \|q - \widetilde{q}\|_{\infty, [t, t+h]}.\end{aligned}$$

Note that the right-hand side is independent of λ. Thus, we are led to

$$\|\Psi q - \Psi \widetilde{q}\|_{\infty, [t, t+h]} \le hL \sum_{j=1}^{m} \Big(\int_0^1 |L_j|\, \mathrm{d}s \Big) \, \|q - \widetilde{q}\|_{\infty, [t, t+h]}.$$

This shows that Ψ is a contraction, if h is sufficiently small. Therefore, the Banach fixpoint theorem concludes the proof. □

Example 3.24 (Collocation methods with $m = 1$). *Let $m = 1$. Then, the collocation polynomial q_ℓ from* (3.23) *is linear and hence its derivative is constant $q_\ell'(t) = \frac{y_{\ell+1} - y_\ell}{h_\ell}$.*

- *If $c_1 = 0$, then $\frac{y_{\ell+1}-y_\ell}{h_\ell} = q_\ell'(t_\ell) = f(t_\ell, q_\ell(t_\ell)) = f(t_\ell, y_\ell)$. Hence, we obtain the explicit Euler method.*
- *If $c_1 = 1$, then $\frac{y_{\ell+1}-y_\ell}{h_\ell} = q_\ell'(t_{\ell+1}) = f(t_{\ell+1}, q_\ell(t_{\ell+1})) = f(t_{\ell+1}, y_{\ell+1})$. Hence, we obtain the implicit Euler method.*
- *If $c = 1/2$, then $t_\ell + \frac{h_\ell}{2} = \frac{t_\ell + t_{\ell+1}}{2}$ and $q_\ell(t_\ell + \frac{h_\ell}{2}) = \frac{q_\ell(t_\ell)+q_\ell(t_{\ell+1})}{2} = \frac{y_\ell + y_{\ell+1}}{2}$. Hence,*

$$\frac{y_{\ell+1} - y_\ell}{h_\ell} = q_\ell'\Big(\frac{t_\ell + t_{\ell+1}}{2}\Big) = f\Big(t_\ell + \frac{h_\ell}{2}\,,\, q_\ell\Big(t_\ell + \frac{h_\ell}{2}\Big)\Big) = f\Big(\frac{t_\ell + t_{\ell+1}}{2}\,,\, \frac{y_\ell + y_{\ell+1}}{2}\Big),$$

and we obtain the implicit midpoint scheme.

Theorem 3.25 (Collocation methods are Runge–Kutta methods). *For any nodes $0 \le c_1 < \cdots < c_m \le 1$, the corresponding collocation scheme is an (possibly implicit) m-stage Runge–Kutta method with consistency order $p \ge m$, where*

$$A_{ij} := \int_0^{c_i} L_j \,\mathrm{d}t \quad \text{and} \quad b_j := \int_0^1 L_j \,\mathrm{d}t \quad \text{for all } i,j = 1,\dots,m \tag{3.25}$$

with the Lagrange polynomials $L_j(t) := \prod_{\substack{k=1\\k\neq j}}^{m} \frac{t - c_k}{c_j - c_k}$.

Proof. The proof is split into two steps.

Step 1. Let $q_\ell \in \mathbb{P}_m$ be the collocation polynomial (3.23). Then, $q_\ell' \in \mathbb{P}_{m-1}$ and hence

$$q_\ell'(s) \overset{(3.18)}{=} \sum_{j=1}^{m} f\big(t + c_j h,\, q_\ell(t + c_j h)\big)\, L_j(s).$$

Therefore, we see that

$$\begin{aligned} q_\ell(t_\ell + c_i h_\ell) &= q_\ell(t_\ell) + \int_{t_\ell}^{t_\ell + c_i h_\ell} q_\ell'(t)\,\mathrm{d}t = y_\ell + h_\ell \int_0^{c_i} q_\ell'(t_\ell + s h_\ell)\,\mathrm{d}s \\ &= y_\ell + h_\ell \sum_{j=1}^{m} f\big(t + c_j h,\, q_\ell(t + c_j h)\big) \int_0^{c_i} L_j(s)\,\mathrm{d}s \\ &\overset{(3.25)}{=} y_\ell + h_\ell \sum_{j=1}^{m} A_{ij} f\big(t + c_j h,\, q_\ell(t + c_j h)\big). \end{aligned}$$

With $k_i := f\big(t + c_i h,\, q_\ell(t + c_i h)\big)$, the last identity proves that

$$k_i = f\big(t_\ell + c_i h_\ell,\, q_\ell(t_\ell + c_i h_\ell)\big) = f\Big(t_\ell + c_i h_\ell,\, y_\ell + h_\ell \sum_{j=1}^{m} A_{ij} k_j\Big).$$

The same argument proves that

$$y_{\ell+1} = q_\ell(t_{\ell+1}) = q_\ell(t_\ell) + \int_{t_\ell}^{t_\ell+h_\ell} q_\ell'(t)\,\mathrm{d}t = y_\ell + h_\ell \int_0^1 q_\ell'(t_\ell + sh_\ell)\,\mathrm{d}s$$

$$= y_\ell + h_\ell \sum_{j=1}^m f\big(t + c_j h,\, q_\ell(t + c_j h)\big) \int_0^1 L_j(s)\,\mathrm{d}s \overset{(3.25)}{=} y_\ell + h_\ell \sum_{j=1}^m b_j k_j.$$

This shows that the collocation method is indeed a Runge–Kutta method.

Step 2. For each polynomial $q \in \mathbb{P}_{m-1}$, it holds that

$$q(s) \overset{(3.18)}{=} \sum_{j=1}^m q(c_j) L_j(s).$$

Hence, it follows that

$$\sum_{j=1}^m b_j q(c_j) \overset{(3.25)}{=} \sum_{j=1}^m q(c_j) \int_0^1 L_j\,\mathrm{d}t = \int_0^1 q\,\mathrm{d}t$$

as well as

$$\sum_{j=1}^m A_{ij} q(c_j) \overset{(3.25)}{=} \sum_{j=1}^m q(c_j) \int_0^{c_i} L_j\,\mathrm{d}t = \int_0^{c_i} q\,\mathrm{d}t.$$

Therefore, Proposition 3.19 proves that the consistency order is at least m. □

One major advantage of collocation methods is that they provide an approximation of the sought solution y in the full time interval $[t_0, T]$ (and not only at the time-steps t_j). Moreover, we even get natural approximations of the derivatives of y.

Theorem 3.26 (Global convergence of collocation methods). *Let $0 \le c_1 < \cdots < c_m \le 1$ be given nodes of a collocation method with consistency order p. Define the spline $q : [t_0, T] \to \mathbb{R}^n$ by $q|_{[t_\ell, t_{\ell+1}]} := q_\ell$, where $q_\ell \in \mathbb{P}_m$ are the collocation polynomials (3.23) for all $\ell = 0, \dots, N-1$. Then, $q \in C([t_0, T]; \mathbb{R}^n)$. Let $r := \min\{p, m+1\}$ and note that $m \le r \le 2m$. Suppose that $f \in C^r([t_0, T] \times \mathbb{R}^n; \mathbb{R}^n)$. Then, the solution $y \in C^{r+1}([t_0, T]; \mathbb{R}^n)$ of (3.2) satisfies that*

$$\|y^{(k)} - q^{(k)}\|_{\infty,[t_0,T]} = \mathcal{O}(h_\Delta^{r-k}) \quad \text{for all } 0 \le k \le r, \tag{3.26}$$

where $q^{(k)}$ is understood elementwise on $[t_\ell, t_{\ell+1}]$ for all $\ell = 0, \dots, N-1$.

Lemma 3.27 (Polynomial approximation). *Let $z \in C^{r+1}[t, t+h]$, $q \in \mathbb{P}_r$, and $k \in \{1, \dots, r\}$. Then, it holds that*

$$\|z^{(k)} - q^{(k)}\|_{\infty,[t,t+h]} \le C\,\big(h^{r+1-k}\,\|z^{(r+1)}\|_{\infty,[t,t+h]} + h^{-k}\,\|z - q\|_{\infty,[t,t+h]}\big), \tag{3.27}$$

where $C > 0$ depends only on r, but is independent of z, q, and h.

Proof. The proof follows from a so-called scaling argument and is split in three steps.

Step 1. Define $\widehat{z}(s) := z(t+sh)$ and $\widehat{q}(s) := q(t+sh)$. Note that, e.g., $\widehat{z}^{(k)}(s) = h^k\, z^{(k)}(t+sh)$. In particular, one sees that

$$\|\widehat{z}^{(k)} - \widehat{q}^{(k)}\|_{\infty,[0,1]} = h^k\, \|z^{(k)} - q^{(k)}\|_{\infty,[t,t+h]},$$
$$\|\widehat{z}^{(r+1)}\|_{\infty,[0,1]} = h^{r+1}\, \|z^{(r+1)}\|_{\infty,[t,t+h]}.$$

Step 2. Let $0 \le s_0 < \cdots < s_r \le 1$ be interpolation nodes. Define

$$(I\widehat{z})(s) := \sum_{j=0}^{r} \widehat{z}(s_j) L_j(s), \quad \text{where} \quad L_j(s) := \prod_{\substack{k=0 \\ k\neq j}}^{r} \frac{s - s_k}{s_j - s_k}.$$

The error estimate (3.20) for the Lagrange interpolation yields that

$$\|\widehat{z} - I\widehat{z}\|_{\infty,[0,1]} \le \frac{1}{(r+1)!}\, \|\widehat{z}^{(r+1)}\|_{\infty,[0,1]},$$
$$\|\widehat{z}^{(k)} - (I\widehat{z})^{(k)}\|_{\infty,[0,1]} \le \frac{1}{(r+1-k)!}\, \|\widehat{z}^{(r+1)}\|_{\infty,[0,1]}.$$

Moreover, norm equivalence on the finite-dimensional space $\mathbb{P}_r$ provides a constant $C_k > 0$ such that

$$\begin{aligned}\|(I\widehat{z})^{(k)} - \widehat{q}^{(k)}\|_{\infty,[0,1]} &\le \|I\widehat{z} - \widehat{q}\|_{\infty,[0,1]} + \|(I\widehat{z})^{(k)} - \widehat{q}^{(k)}\|_{\infty,[0,1]} \\ &\le C_k \max_{j=0,\dots,r} \big|(I\widehat{z})(s_j) - \widehat{q}(s_j)\big|,\end{aligned}$$

where we note that $\|p\| := \max_{j=0,\dots,r} \big|p(z_j)\big|$ is a norm on $\mathbb{P}_r$ due to Lemma 3.20. Since $(I\widehat{z})(s_j) = z(s_j)$ for all $j = 0, \dots, r$, we obtain that

$$\|(I\widehat{z})^{(k)} - \widehat{q}^{(k)}\|_{\infty,[0,1]} \le C_k\, \|\widehat{z} - \widehat{q}\|_{\infty,[0,1]}.$$

With the triangle inequality and $C := \max_{k=1,\dots,r} \max\{C_k,\, 1/(r+1-k)!\}$, we conclude that

$$\begin{aligned}\|\widehat{z}^{(k)} - \widehat{q}^{(k)}\|_{\infty,[0,1]} &\le \|z^{(k)} - (I\widehat{z})^{(k)}\|_{\infty,[0,1]} + \|(I\widehat{z})^{(k)} - \widehat{q}^{(k)}\|_{\infty,[0,1]} \\ &\le \frac{1}{(r+1-k)!}\, \|\widehat{z}^{(r+1)}\|_{\infty,[0,1]} + C_k\, \|\widehat{z} - \widehat{q}\|_{\infty,[0,1]} \\ &\le C\, \big(\|\widehat{z}^{(r+1)}\|_{\infty,[0,1]} + \|\widehat{z} - \widehat{q}\|_{\infty,[0,1]}\big).\end{aligned}$$

We stress that this already proves the claim (3.27) on the reference interval $[0,1]$.

Step 3. It holds that

$$\begin{aligned}\|z^{(k)} - q^{(k)}\|_{\infty,[t,t+h]} &= h^{-k}\, \|\widehat{z}^{(k)} - \widehat{q}^{(k)}\|_{\infty,[0,1]} \\ &\le C\, h^{-k}\, \big(\|\widehat{z}^{(r+1)}\|_{\infty,[0,1]} + \|\widehat{z} - \widehat{q}\|_{\infty,[0,1]}\big) \\ &\le C\, h^{-k}\, \big(h^{r+1}\, \|z^{(r+1)}\|_{\infty,[t,t+h]} + \|z - q\|_{\infty,[t,t+h]}\big).\end{aligned}$$

This concludes the proof. We stress that the norm equivalence argument has to be used on $[0,1]$ instead of $[t, t+h]$ to ensure that C is independent of h. □

Proof of Theorem 3.26. Note that

$$\|g\|_{\infty,[t_0,T]} = \max_{\ell=0,\dots,N-1} \|g\|_{\infty,[t_\ell,t_{\ell+1}]}.$$

Therefore, it suffices to prove (3.26) for one time interval $[t_\ell, t_{\ell+1}]$. Moreover, recall that $q|_{[t_\ell,t_{\ell+1}]} = q_\ell \in \mathbb{P}_m$ and $r \geq m$. Therefore, Lemma 3.27 proves that

$$\|y^{(k)} - q^{(k)}\|_{\infty,[t_\ell,t_{\ell+1}]} \leq C \big(h_\ell^{r+1-k} \, \|y^{(r+1)}\|_{\infty,[t_\ell,t_{\ell+1}]} + h_\ell^{-k} \, \|y - q\|_{\infty,[t_\ell,t_{\ell+1}]}\big).$$

Hence, it suffices to consider the case $k = 0$.

Step 1. Let $t \in [t_0, T)$ and $h > 0$. Let $Iy \in \mathbb{P}_m$ be the (unique) polynomial such that

$$Iy(t) = y(t) \quad \text{and} \quad \forall j = 1, \dots, m: \quad (Iy)'(t + c_j h) = y'(t + c_j h).$$

Let $0 \leq \lambda \leq 1$. With $L_j(s) = \displaystyle\prod_{\substack{k=1\\k\neq j}}^{m} \frac{s - c_k}{c_j - c_k}$, Lemma 3.21 yields that

$$\begin{aligned} y(t + \lambda h) - Iy(t + \lambda h) &= \int_t^{t+\lambda h} (y - Iy)'(s) \, \mathrm{d}s = h \int_0^\lambda (y - Iy)'(t + sh) \, \mathrm{d}s \\ &\overset{(3.21)}{=} h \Big[\sum_{j=1}^{m} (y - Iy)'(t + c_j h) \int_0^\lambda L_j(s) \, \mathrm{d}s + \mathcal{O}(h^m) \Big] = \mathcal{O}(h^{m+1}). \end{aligned}$$

Since the right-hand side is independent of λ, this proves that

$$\|y - Iy\|_{\infty,[t,t+h]} = \mathcal{O}(h^{m+1}).$$

Step 2. Employ the notation $Iy \in \mathbb{P}_m$ from Step 1 for the interval $[t_\ell, t_\ell + h_\ell] = [t_\ell, t_{\ell+1}]$. The triangle inequality and Step 1 prove that

$$\begin{aligned} \|y - q\|_{\infty,[t_\ell,t_{\ell+1}]} &\leq \|y - Iy\|_{\infty,[t_\ell,t_{\ell+1}]} + \|Iy - q\|_{\infty,[t_\ell,t_{\ell+1}]} \\ &= \mathcal{O}(h_\ell^{m+1}) + \|Iy - q\|_{\infty,[t_\ell,t_{\ell+1}]}. \end{aligned} \tag{3.28}$$

Let $0 \leq \lambda \leq 1$. We argue as before:

$$\begin{aligned} Iy(t_\ell + \lambda h_\ell) - q(t_\ell + \lambda h_\ell) &= y(t_\ell) - q(t_\ell) + \int_{t_\ell}^{t_\ell + \lambda h_\ell} (Iy - q)'(s) \, \mathrm{d}s \\ &= y(t_\ell) - q(t_\ell) + h_\ell \int_0^\lambda (Iy - q)'(t + sh) \, \mathrm{d}s \\ &= y(t_\ell) - q(t_\ell) + h_\ell \sum_{j=1}^{m} (Iy - q)'(t_\ell + c_j h_\ell) \int_0^\lambda L_j(s) \, \mathrm{d}s. \end{aligned}$$

Recall that collocation methods are implicit Runge–Kutta schemes and hence stable. From consistency, we thus get convergence

$$y(t_\ell) - q(t_\ell) = y(t_\ell) - y_\ell = \mathcal{O}(h_\Delta^p).$$

Note that

$$(Iy - q)'(t_\ell + c_j h_\ell) = f\big(t_\ell + c_j h_\ell, \, y(t_\ell + c_j h_\ell)\big) - f\big(t_\ell + c_j h_\ell, \, q(t_\ell + c_j h_\ell)\big).$$

Combining the last three identities with the Lipschitz continuity of f, we obtain that

$$\|(Iy-q)(t_\ell+\lambda h_\ell)\| \le h_\ell L \sum_{j=1}^m \Big| \int_0^\lambda L_j(s)\,\mathrm{d}s\Big| \,\|(y-q)(t_\ell+c_j h_\ell)\| + \mathcal{O}(h_\Delta^p)$$

$$\le \left(h_\ell L \sum_{j=1}^m \int_0^1 |L_j(s)|\,\mathrm{d}s \right) \|y-q\|_{\infty,[t_\ell,t_{\ell+1}]} + \mathcal{O}(h_\Delta^p).$$

Note that the right-hand side is independent of $0 \le \lambda \le 1$. Hence, we are led to

$$\|Iy-q\|_{\infty,[t_\ell,t_{\ell+1}]} \le \left(h_\ell L \sum_{j=1}^m \int_0^1 |L_j(s)|\,\mathrm{d}s \right) \|y-q\|_{\infty,[t_\ell,t_{\ell+1}]} + \mathcal{O}(h_\Delta^p).$$

In combination with (3.28), we see that

$$\|y-q\|_{\infty,[t_\ell,t_{\ell+1}]} \le \left(h_\ell L \sum_{j=1}^m \int_0^1 |L_j(s)|\,\mathrm{d}s \right) \|y-q\|_{\infty,[t_\ell,t_{\ell+1}]} + \mathcal{O}(h_\Delta^p) + \mathcal{O}(h_\ell^{m+1}).$$

For sufficiently small $h_\ell \le h_\Delta$, we thus infer that

$$\|y-q\|_{\infty,[t_\ell,t_{\ell+1}]} = \mathcal{O}(h_\Delta^p) + \mathcal{O}(h_\Delta^{m+1}) = \mathcal{O}(h_\Delta^r).$$

□

The next goal is the theorem that a collocation method has consistency order p if and only if the induced quadrature on $[0,1]$ (see Proposition 3.17) is exact for all $q \in \mathbb{P}_{p-1}$. In particular, the m-stage Gaussian collocation method is the only collocation scheme with consistency order $p = 2m$.

Remark 3.28. *Let $f \in C^p([t_0,T]\times\mathbb{R}^n;\mathbb{R}^n)$ and suppose that f is Lipschitz in y. Let $(t_0,y_0)\in\mathbb{R}\times\mathbb{R}^n$. Then,*

$$Y(t,t_0,y_0) := y_0 + \int_{t_0}^t f\big(s, Y(s,t_0,y_0)\big)\,\mathrm{d}s \tag{3.29}$$

is the unique solution of

$$y'(t) = f\big(t,y(t)\big) \text{ in } [t_0,T] \quad \text{subject to} \quad y(t_0)=y_0.$$

Clearly, $y(t)=Y(t,t_0,y_0)$ satisfies $y\in C^{p+1}([t_0,T],\mathbb{R}^n)$. Moreover, one can show that $Y(t,t_0,y_0)$ is also C^p with respect to t_0 and y_0, and all derivatives depend only on the derivatives of f. We refer to [Wal00, Part III, Section §13, Subsection XI, Corollar].

Lemma 3.29 (Gröbner & Alekseev). *Let $f,\varepsilon\in C^1([t,t+h]\times\mathbb{R}^n;\mathbb{R}^n)$ be Lipschitz in y. Let $y,z\in C^2([t,t+h],\mathbb{R}^n)$ such that*

$$\begin{aligned} y(t) &= y_0, \quad y'(t) = f\big(t,y(t)\big), \\ z(t) &= y_0, \quad z'(t) = f\big(t,z(t)\big) + \varepsilon\big(t,z(t)\big). \end{aligned} \tag{3.30}$$

Then, it follows that

$$z(t+h) - y(t+h) = \int_t^{t+h} \partial_3 Y\big(t+h,s,z(s)\big)\varepsilon\big(s,z(s)\big) \tag{3.31}$$

where $Y\big(t,s,z(s)\big)$ is given by (3.29).

Proof. Let $N \in \mathbb{N}$. For $\ell = 0, \dots, N$, let $t_\ell := t + \ell H$ with $H := h/N$. Note that $t_0 = t$ and $t_N = t + h$. The Taylor expansion shows that

$$\begin{aligned} d_i &:= z(t_{i+1}) - Y\big(t_{i+1}, t_i, z(t_i)\big) \\ &= \big[\, z(t_i) + H\,\big\{ f\big(t_i, z(t_i)\big) + \varepsilon\big(t_i, z(t_i)\big)\big\} + \mathcal{O}(H^2)\,\big] - \big[\, z(t_i) + H\, f\big(t_i, z(t_i)\big) + \mathcal{O}(H^2)\,\big] \\ &= H\,\varepsilon\big(t_i, z(t_i)\big) + \mathcal{O}(H^2) \\ &= H\,\varepsilon\big(t_{i+1}, z(t_{i+1})\big) + \mathcal{O}(H^2). \end{aligned}$$

By definition of d_i, the uniqueness of ODE solutions proves that

$$Y\big(t+h, t_i, z(t_i)\big) = Y\big(t+h, t_{i+1}, z(t_{i+1}) - d_i\big).$$

Therefore, the Taylor expansion shows that

$$\begin{aligned} D_i &:= Y\big(t+h, t_{i+1}, z(t_{i+1})\big) - Y\big(t+h, t_i, z(t_i)\big) \\ &= Y\big(t+h, t_{i+1}, z(t_{i+1})\big) - Y\big(t+h, t_{i+1}, z(t_{i+1}) - d_i\big) \\ &= \partial_3 Y\big(t+h, t_{i+1}, z(t_{i+1})\big) d_i + \mathcal{O}(|d_i|^2) \\ &= H\, \underbrace{\partial_3 Y\big(t+h, t_{i+1}, z(t_{i+1})\big)\, \varepsilon\big(t_{i+1}, z(t_{i+1})\big)}_{=:g(t_{i+1})} + \mathcal{O}(H^2). \end{aligned}$$

Note that $\mathcal{O}(H^2) = \mathcal{O}(N^{-2})$. Together with Riemann sum theory, it follows that

$$\sum_{i=0}^{N-1} D_i = \mathcal{O}(N^{-1}) + \sum_{i=0}^{N-1} \Big((t_{i+1} - t_i)\, g(t_{i+1}) \Big) \xrightarrow{N\to\infty} \int_t^{t+h} g(s)\,\mathrm{d}s,$$

since g is (at least) continuous. Moreover, the telescopic series proves that

$$\begin{aligned} \sum_{i=0}^{N-1} D_i &= Y\big(t+h, t_N, z(t_N)\big) - Y\big(t+h, t_0, z(t_0)\big) \\ &= Y\big(t+h, t+h, z(t+h)\big) - Y\big(t+h, t, y(t)\big) = z(t+h) - y(t+h). \end{aligned}$$

Combining the latter two identities, we prove (3.31). □

Theorem 3.30 (Quadrature vs. collocation). *Let $0 \le c_1 < \cdots < c_m \le 1$ and $b_j := \int_0^1 L_j\,\mathrm{d}t$, where $L_j \in \mathbb{P}_{m-1}$ are the Lagrange polynomials. Let $p \in \mathbb{N}$. Then, the following statements* (i)–(ii) *are equivalent:*

(i) $\displaystyle\sum_{j=1}^{m} b_j q(c_j) = \int_0^1 q\,\mathrm{d}t$ *for all $q \in \mathbb{P}_{p-1}$, i.e., the quadrature has exactness $p-1$.*

(ii) *The corresponding collocation method has consistency order p.*

In any case, it holds that $m \le p \le 2m$.

Proof. The implication (ii) $\Longrightarrow$ (i) as well as the bound $m \le p \le 2m$ are already known; see Proposition 3.17. It thus only remains to prove that (i) $\Longrightarrow$ (ii).

Step 1. Let $g \in C^p[t, t+h]$. Choose $\widetilde{c}_{m+1}, \dots, \widetilde{c}_p \in [0,1]$ such that $c_1, \dots, c_m, \widetilde{c}_{m+1}, \dots, \widetilde{c}_p$ are pairwise distinct. According to Lemma 3.20, there exists a unique polynomial $q \in \mathbb{P}_{p-1}$ such that

$$q(t + c_j h) = g(t + c_j h) \quad \text{for all } j = 1, \dots, m,$$
$$q(t + \widetilde{c}_j h) = g(t + \widetilde{c}_j h) \quad \text{for all } j = m+1, \dots, p.$$

Then, it holds that

$$\|g - q\|_{\infty,[t,t+h]} \overset{(3.20)}{\leq} \frac{\|g^{(p)}\|_{\infty,[t,t+h]}}{p!}\, h^p. \tag{3.32}$$

By assumption (i), it holds that

$$\int_t^{t+h} q(s)\,\mathrm{d}s = h \int_0^1 q(t+sh)\,\mathrm{d}s \overset{\text{(i)}}{=} h \sum_{j=1}^m b_j q(t + c_j h) = h \sum_{j=1}^m b_j g(t + c_j h).$$

Therefore, it follows that

$$\begin{aligned} \Big| \int_t^{t+h} g\,\mathrm{d}s - h \sum_{j=1}^m b_j g(t + c_j h) \Big| &= \Big| \int_t^{t+h} (g - q)\,\mathrm{d}s \Big| \leq h\, \|g - q\|_{\infty,[t,t+h]} \\ &\overset{(3.32)}{\leq} \frac{\|g^{(p)}\|_{\infty,[t,t+h]}}{p!}\, h^{p+1}. \end{aligned} \tag{3.33}$$

Step 2. Let $q \in \mathbb{P}_m$ be the collocation polynomial on $[t, t+h]$, i.e.,

$$q(t) = y(t) \quad \text{and} \quad q(t + c_j h) = f\big(t + c_j h,\, q(t + c_j h)\big) \quad \text{for all } j = 1, \dots, m.$$

We apply Lemma 3.29 for the collocation polynomial and

$$q'(t) = f\big(t, q(t)\big) + \big[q'(t) - f\big(t, q(t)\big)\big], \quad \text{i.e.,} \quad \varepsilon\big(t, q(t)\big) = q'(t) - f\big(t, q(t)\big).$$

This leads to

$$q(t+h) - y(t+h) = \int_t^{t+h} \underbrace{\partial_3 Y\big(t+h, s, q(s)\big)\big[q'(s) - f\big(s, q(s)\big)\big]}_{=:g(s)}\,\mathrm{d}s,$$

where we also define g for which we aim to apply Step 1. Note that $g(t + c_j h) = 0$ for all $j = 1, \dots, m$ by the collocation conditions. From Step 1, we thus infer that the consistency error satisfies

$$|y(t+h) - q(t+h)| = \Big| \int_t^{t+h} g\,\mathrm{d}s - \underbrace{h \sum_{j=1}^m b_j g(t + c_j h)}_{=0} \Big| \overset{(3.33)}{\leq} \frac{\|g^{(p)}\|_{\infty,[t,t+h]}}{p!}\, h^{p+1}.$$

It only remains to argue that $\|g^{(p)}\|_{\infty,[t,t+h]}$ is uniformly bounded in terms of y and f. Note that

$$g(s) = \partial_3 Y\big(t+h, s, q(s)\big)\big[q'(s) - f\big(s, q(s)\big)\big],$$

where $Y\big(t+h, s, q(s)\big)$ solves

$$\partial_1 Y\big(t, s, q(s)\big) = f\big(t, Y(t, s, q(s))\big) \text{ in } [t, t+h] \quad \text{subject to} \quad Y\big(s, s, q(s)\big) = q(s).$$

According to the chain rule (and Remark 3.28), $g^{(p)}$ depends only on partial derivatives of f and derivatives of q. For $0 \le k \le m$, it holds that

$$\|q^{(k)}\|_{\infty,[t,t+h]} \le \|y^{(k)} - q^{(k)}\|_{\infty,[t,t+h]} + \|y^{(k)}\|_{\infty,[t,t+h]} \overset{(3.26)}{=} \mathcal{O}(h^{m-k}) + \|y^{(k)}\|_{\infty,[t,t+h]} = \mathcal{O}(1).$$

For $k > m$, it holds that $q^{(k)} = 0$. Overall, we thus obtain that $\|g^{(p)}\|_{\infty,[t,t+h]} = \mathcal{O}(1)$. This concludes the proof. □

4. Stiff ODEs

4.1. Introduction.

In 1952, it was observed by Curtiss and Hirschfelder that explicit methods fail for the numerical integration of certain ODEs which model chemical reactions. They said that the ODE is stiff, if certain components of the solution arrive in a very short time in their equilibrium (i.e., the fast reacting components), while other slowly changing components are more or less fixed (i.e., stiff).

Mathematically, this is, e.g., the case for

$$y'(t) = My(t) + f(t) \tag{4.1}$$

if the eigenvalues $\lambda_1, \dots, \lambda_n \in \mathbb{C}$ of $M \in \mathbb{R}^{n\times n}$, sorted by $\operatorname{Re}\lambda_1 \ge \cdots \ge \operatorname{Re}\lambda_n$, satisfy that

$$|\operatorname{Re}\lambda_1| \sim 1, \quad \text{but} \quad \operatorname{Re}\lambda_n \ll 0. \tag{4.2}$$

Example 4.1. *Consider the symmetric matrix* $M = \begin{pmatrix} a & b \\ b & a \end{pmatrix} \in \mathbb{R}^{2\times 2}$ *with* $a, b \in \mathbb{R}$. *The characteristic polynomial is* $p(\lambda) = \det(A - \lambda I) = (a-\lambda)^2 - b^2 = \lambda^2 - 2a\lambda + (a^2 - b^2)$. *The eigenvalues of* M *are the zeros of* p *and hence* M *has the eigenvalues* $\lambda = a \pm b$.

Joseph-Louis de Lagrange (1736–1813) was an Italian mathematician (born in Turino as *Giuseppe Lodovico Lagrangia*). In 1755, he got the chair for mathematics at the military academy in Turino. In 1766, he became the successor of Euler at the Prussian Academy if Sciences in Berlin. In 1786, he moved to Paris to the French Academy of Sciences. From 1797, he was professor at the Ecole Polytechnique in Paris, where we was promoting the young Augustin-Louis Cauchy.

Charles Francis Curtiss (1921–2007) was a American chemist. He lived his academic live at the University of Wisconsin, where he did his bachelor in 1942 and his PhD (supervized by Hirschfelder) in 1948. He became assistant professor in 1949, associate professor in 1954, and full professor in 1960. He retired in 1989.

Joseph Oakland Hirschfelder (1911–1990) was an American chemist in physicist. He studied natural sciences at University of Minnesota (1927–1929) and Yale (1929–1931) and finished his PhD in chemistry and physics at Princeton in 1936. From 1936–1937, he was postdoc with John von Neumann at the Princeton Institute for Advanced Studies. In 1937, he went to the University of Wisconsin, where he became assistant professor in chemistry in 1941. 1944/45, he was group leader at the Manhattan project at Los Alamos and later leading researcher in the American nuclear program. In 1946, he became full professor at the University of Wisconsin. He retired in 1981.

Charles Francis Curtiss, Joseph Oakland Hirschfelder: *Integration of stiff equations*, Proceedings of the National Academy of Sciences of the USA; 38 (1952), 235–243. [This work introduces the BDF multistep methods for the solution of stiff ODEs]

Even for simple $a = -50$, $b = -51$, we get eigenvalues $\lambda_1 = 1$ and $\lambda_2 = -101$, so that the corresponding ODE would be stiff.

Proposition 4.2. *Given $y_0 \in \mathbb{R}^n$ and $M \in \mathbb{R}^{n\times n}$, consider the initial value problem*

$$y'(t) = My(t) \text{ in } [t_0, T] \quad \text{subject to} \quad y(t_0) = y_0. \tag{4.3}$$

Then, the unique solution $y \in C^1([t_0, T]; \mathbb{R}^n)$ reads

$$y(t) = e^{M(t-t_0)} y_0 \tag{4.4}$$

with the matrix exponential function

$$e^X := \sum_{k=0}^{\infty} \frac{1}{k!} X^k \quad \text{for } X \in \mathbb{R}^{n\times n}. \tag{4.5}$$

Suppose that M is diagonalizable, i.e., there exist $\{v_1, \dots v_n\} \subset \mathbb{C}^n$ linearly independent eigenvectors for eigenvalues $\lambda_1, \dots, \lambda_n \in \mathbb{C}$ such that $Mv_j = \lambda_j v_j$ for all $j = 1, \dots, n$. Then,

$$y_0 = \sum_{j=1}^{n} \alpha_j v_j \quad \Longrightarrow \quad y(t) = \sum_{j=1}^{n} \left(\alpha_j e^{\lambda_j (t-t_0)}\right) v_j, \tag{4.6}$$

i.e., $y = \sum_{j=1}^n y_j v_j$, where the scalar functions $y_j \in C^1[t_0, T]$ solve the ODEs

$$y_j'(t) = \lambda_j y_j(t) \text{ in } [t_0, T] \quad \text{subject to} \quad y_j(t_0) = \alpha_j. \tag{4.7}$$

Proof. Existence and uniqueness of the solution of (4.3) (resp. (4.7)) follows from the Picard–Lindelöf theorem. Consider y given by (4.4). Then, $y(t_0) = y_0$ and

$$\begin{aligned} y'(t) &= d_t \Big(\sum_{k=0}^{\infty} \frac{1}{k!} M^k (t-t_0)^k y_0 \Big) \\ &= \sum_{k=1}^{\infty} \frac{1}{k!} k\, M^k (t-t_0)^{k-1} y_0 \\ &= M \Big(\sum_{k=1}^{\infty} \frac{1}{(k-1)!} M^{k-1} (t-t_0)^{k-1} y_0 \Big) \\ &= M \Big(\sum_{k=0}^{\infty} \frac{1}{k!} M^k (t-t_0)^k y_0 \Big) = My(t). \end{aligned}$$

If M is diagonalizable and $y_0 = \sum_{j=1}^n \alpha_j v_j$, then

$$\begin{aligned} y(t) &= \sum_{k=0}^{\infty} \frac{1}{k!} M^k (t-t_0)^k y_0 \\ &= \sum_{j=1}^{n} \left(\sum_{k=0}^{\infty} \frac{1}{k!} M^k (t-t_0)^k \alpha_j v_j \right) \\ &= \sum_{j=1}^{n} \left(\alpha_j \sum_{k=0}^{\infty} \frac{1}{k!} \lambda_j^k (t-t_0)^k v_j \right) \\ &= \sum_{j=1}^{n} \left(\alpha_j e^{\lambda_j (t-t_0)} \right) v_j. \end{aligned}$$

Since $y_j = e^{\lambda_j (t-t_0)} \alpha_j$, this concludes the proof. □

More generally, one can even prove the following result. We note that the solution $z \in C^1([t_0, T]; \mathbb{R}^n)$ of (4.9) can be obtained by solving n scalar ODEs (if $\widetilde{g}(t) := V^{-1} g(t)$ is explicitly known).

Proposition 4.3. *Given $g \in C([t_0, T]; \mathbb{R}^n)$, $M \in \mathbb{R}^{n \times n}$, and $y_0 \in \mathbb{R}^n$, consider the inhomogeneous initial value problem*

$$y'(t) = My(t) + g(t) \text{ in } [t_0, T], \quad y(t_0) = y_0. \tag{4.8}$$

Suppose that the matrix $M \in \mathbb{R}^n$ is diagonalizable, i.e., there exist $\{v_1, \dots v_n\} \subset \mathbb{C}^n$ linearly independent eigenvectors for eigenvalues $\lambda_1, \dots, \lambda_n \in \mathbb{C}$ such that $Mv_j = \lambda_j v_j$ for all $j = 1, \dots, n$. Define $\Lambda := \operatorname{diag}(\lambda_1, \dots, \lambda_n) \in \mathbb{R}^{n \times n}$ and $V := (v_1, \dots, v_n) \in \mathbb{R}^{n \times n}$ and consider the problem

$$z'(t) = \Lambda z(t) + V^{-1} g(t) \text{ in } [t_0, T], \quad z(t_0) = V^{-1} y_0. \tag{4.9}$$

Then, (4.8)–(4.9) have unique solutions and it holds that $y(t) = Vz(t)$.

Proof. Existence and uniqueness of the solutions of (4.8)–(4.9) follows from the Picard–Lindelöf theorem. Clearly, V is invertible and independent of t. Therefore, (4.9) implies that

$$(Vz)'(t) = V\Lambda z(t) + g(t).$$

Note that $V\Lambda = MV$, since the columns of V are the eigenvectors of M. Hence, $\widetilde{y} := Vz$ solves that

$$\widetilde{y}'(t) = M\widetilde{y}(t) + g(t) \quad \text{together with} \quad \widetilde{y}(t_0) = Vz(t_0) = y_0.$$

From the uniqueness of solutions, we thus conclude that $\widetilde{y} = y$. □

Exercise 4.4. *Consider the setting of Proposition 4.3. Let* $\begin{array}{c|c} c & A \\ \hline & b^\top \end{array}$ *be an explicit m-stage Runge–Kutta method. Let $y_\ell, z_\ell \in \mathbb{R}^n$ be the resulting Runge–Kutta iterates of (4.8)–(4.9). Prove that $y_\ell = Vz_\ell$ for all $\ell = 0, \dots, N$, i.e., explicit Runge–Kutta methods commute with diagonalization of the ODE system.*

If we consider the simple homogeneous model problem from Proposition 4.2, it follows that, for a stiff ODE, we should employ Runge–Kutta methods, which appropriately integrate the scalar ODEs

$$y_j'(t) = \lambda_j y(t) \quad \text{in } [t_0, T]$$

simultaneously for all eigenvalues $\lambda_j \in \mathbb{C}$ of $M \in \mathbb{R}^{n\times n}$.

Remark 4.5. *For given $\lambda \in \mathbb{C}\backslash\{0\}$, let us consider the scalar ODE*

$$y'(t) = \lambda y(t) \text{ in } \mathbb{R}_{\geq 0}, \quad y(0) = y_0. \tag{4.10}$$

The unique solution is $y(t) = e^{\lambda t} y_0$. If we apply the explicit Euler method with constant time-step size $h > 0$, we get that

$$y_{\ell+1} = y_\ell + h\, f(t_\ell, y_\ell) = (1 + h\lambda)\, y_\ell \quad \text{for all } \ell \geq 0$$

and hence

$$y_\ell = (1 + h\lambda)^\ell\, y_0 \quad \text{for all } \ell \geq 0. \tag{4.11}$$

If we apply the implicit Euler method, we get that

$$y_{\ell+1} = y_\ell + h\, f(t_{\ell+1}, y_{\ell+1}) = y_\ell + h\lambda\, y_{\ell+1} \quad \text{for all } \ell \geq 0$$

and hence (if $h\lambda \neq 1$)

$$y_\ell = (1 - h\lambda)^{-\ell}\, y_0 \quad \text{for all } \ell \geq 0. \tag{4.12}$$

We make the following observation: If $\operatorname{Re}\lambda \gg 0$, then both methods will need small h, since the exact solution grows exponentially. If $\operatorname{Re}\lambda < 0$, then the explicit Euler method needs small h so that (4.11) leads to $y_\ell \to 0$ as $\ell \to \infty$. If $|1 + h\lambda| > 1$, then (4.11) leads to blow-up and oscillations. On the other hand, the implicit Euler method (4.12) guarantees always $y_\ell \to 0$ as $\ell \to \infty$, if $\operatorname{Re}\lambda < 0$.

Example 4.6. *We consider the ODE*

$$y' = My \text{ in } \mathbb{R}_{\geq 0} \quad \text{with} \quad M = \begin{pmatrix} a & b \\ b & a \end{pmatrix}.$$

Obviously, the eigenvalues of M are $\lambda_{1/2} = a \pm b$ with eigenvectors $v_{1/2} = (1, \pm 1)^\top$. According to Proposition 4.2, the unique solution satisfies

$$y(t) = \alpha_1 e^{(a+b)t} v_1 + \alpha_2 e^{(a-b)t} v_2 \quad \text{provided that} \quad y(0) = \alpha_1 v_1 + \alpha_2 v_2.$$

From the initial condition, we get that

$$\left.\begin{array}{rcl} y_1(0) &=& \alpha_1 + \alpha_2 \\ y_2(0) &=& \alpha_1 - \alpha_2 \end{array}\right\} \quad \text{and hence} \quad \left\{\begin{array}{rcl} \alpha_1 &=& \frac{y_1(0)+y_2(0)}{2} \\ \alpha_2 &=& \frac{y_1(0)-y_2(0)}{2}. \end{array}\right.$$

Altogether, the solution reads

$$y_1(t) = \frac{y_1(0) + y_2(0)}{2}\, e^{(a+b)t} + \frac{y_1(0) - y_2(0)}{2}\, e^{(a-b)t},$$

$$y_2(t) = \frac{y_1(0) + y_2(0)}{2}\, e^{(a+b)t} - \frac{y_1(0) - y_2(0)}{2}\, e^{(a-b)t}.$$

Let $a = -51$, $b = -50$, so that $a + b = -101$ and $a - b = -1$. For large $t > 0$, a good approximation of $y(t)$ is obtained, if we neglect the first summands with $e^{(a+b)t}$. But the ODE is stiff. The explicit Euler method requires $|1 + h(a \pm b)| < 1$ to avoid oscillations (in both components) of y. However, note that the "component" $y(t) \cdot (1, -1)^\top$ (which corresponds to the "good" eigenvalue $\lambda_2 = a - b = -1$ will nevertheless be approximated well for any $0 < h < 1$.

4.2. Stability domains. The idea of all notions of stability is that the discrete solution should reflect certain properties of the continuous solution, at least for model problems, where the solution (or its qualitative behavior) is known. For stability domains (as well as A-stability and L-stability), one considers the scalar model problem

$$y'(t) = \lambda y(t) \text{ in } \mathbb{R}_{\geq 0}, \quad y(0) = y_0, \tag{4.13}$$

where $\lambda \in \mathbb{C}$ with $\operatorname{Re} \lambda < 0$.

Definition 4.7. *Let $\Phi(t, y, z, h)$ be the incremental function of a one-step method. A function $R : \mathbb{C} \to \mathbb{C}$ is called **stability function** of the method, if*

$$y_{\ell+1} := y_\ell + h_\ell \Phi(t_\ell, y_\ell, y_{\ell+1}, h_\ell) \stackrel{!}{=} R(\lambda h_\ell)\, y_\ell. \tag{4.14}$$

Example 4.8. *We rephrase Remark 4.5: The stability function of the explicit Euler method is $R(z) = 1 + z$, and it is well-defined for all $z \in \mathbb{C}$. The stability function of the implicit Euler method is $R(z) = 1/(1 - z)$, and it is well-defined for all $z \in \mathbb{C}\backslash\{1\}$.*

Theorem 4.9 (Stability function of Runge–Kutta methods). *Let* $\begin{array}{c|c} c & A \\ \hline & b^\top \end{array}$ *be an m-step Runge–Kutta method. Then, the stability function reads*

$$R(z) = 1 + z b^\top (I - zA)^{-1} \mathbb{1}, \quad \text{where} \quad \mathbb{1} = (1, \dots, 1)^\top \in \mathbb{R}^m. \tag{4.15}$$

It is well-defined for all $z \in \mathbb{C}$ such that $1/z \notin \sigma(A) := \big\{\lambda \in \mathbb{C} : \lambda \text{ is eigenvalue of } A\big\}$. Moreover, there hold the following statements:

(i) *If the method is explicit, then $R \in \mathbb{P}_m$.*

(ii) *If the method is implicit, then $R = P/Q$ with $P, Q \in \mathbb{P}_m$.*

Proof. Recall the Runge–Kutta increments

$$k_i = f\Big(t + c_i h\,,\, y_\ell + h \sum_{j=1}^m A_{ij} k_j\Big) \stackrel{!}{=} \lambda \Big(y_\ell + h \sum_{j=1}^m A_{ij} k_j\Big) \quad \text{for all } i = 1, \dots, m.$$

With the vector $k = (k_1, \dots, k_m)^\top$, it follows that $k = \lambda y_\ell \mathbb{1} + \lambda h A\, k$ and hence

$$k = \lambda y_\ell (I - \lambda h A)^{-1} \mathbb{1}, \quad \text{provided that} \quad 1/(\lambda h) \notin \sigma(A).$$

By definition of a Runge–Kutta method, it follows that

$$y_{\ell+1} = y_\ell + h \sum_{i=1}^m b_i k_i = y_\ell + h b^\top k = y_\ell \left[1 + \lambda h\, b^\top (I - \lambda h A)^{-1} \mathbb{1}\right] = y_\ell\, R(\lambda h).$$

This proves (4.15).

For the remainder of the proof, recall **Cramer's rule**: Let $V = (v_1, \dots, v_n) \in \mathbb{R}^n$ be invertible with columns $v_j \in \mathbb{R}^n$. Let $w \in \mathbb{R}^n$. Define $V_i := (v_1, \dots, v_{i-1}, w, v_{i+1}, \dots, v_n) \in \mathbb{R}^{n\times n}$. Then, the vector $x := V^{-1}w \in \mathbb{R}^n$ satisfies that

$$x_i = \frac{\det(V_i)}{\det(V)} \quad \text{for all } i = 1, \dots, m. \tag{4.16}$$

To apply Cramer's rule, we write

$$R(z) = 1 + zb^\top (I - zA)^{-1}\mathbb{1} = 1 + zb^\top \gamma \quad \text{where} \quad (I - zA)\gamma = \mathbb{1}.$$

This is rewritten as

$$\underbrace{\begin{pmatrix} I - zA & 0 \\ -zb^\top & 1 \end{pmatrix}}_{=:V} \begin{pmatrix} \gamma \\ R(z) \end{pmatrix} = \begin{pmatrix} \mathbb{1} \\ 1 \end{pmatrix}.$$

According to the Laplace cofactor expansion (for the $(m+1)$-th column), it holds that

$$\det(V) = (-1)^{(m+1)+(m+1)} \det(I - zA) = \det(I - zA) \neq 0 \quad \text{provided that} \quad 1/z \notin \sigma(A).$$

Hence, V is invertible and Cramer's rule yields (for $R(m)$ being the $(m+1)$-th coefficient of the solution) that

$$R(z) = \frac{\det \begin{pmatrix} I - zA & \mathbb{1} \\ -zb^\top & 1 \end{pmatrix}}{\det \begin{pmatrix} I - zA & 0 \\ -zb^\top & 1 \end{pmatrix}} = \frac{\det \begin{pmatrix} I - zA & \mathbb{1} \\ -zb^\top & 1 \end{pmatrix}}{\det(I - zA)} =: \frac{P(z)}{Q(z)}.$$

Clearly, the characteristic polynomial $Q(z) = \det(I - zA)$ satisfies $Q \in \mathbb{P}_m$. Moreover, if the method is explicit, then A is strictly lower triangular and hence $\det(I - zA) = 1$. Hence, it only remains to show that the numerator satisfies that $P \in \mathbb{P}_m$. With the Laplace cofactor expansion (for the $(m+1)$-th column), we write

$$P(z) = \det \begin{pmatrix} I - zA & \mathbb{1} \\ -zb^\top & 1 \end{pmatrix} =: \det \begin{pmatrix} S - zT & \mathbb{1} \end{pmatrix} = \sum_{i=1}^{m+1} (-1)^{i+(m+1)} \det \big(S_i - zT_i\big),$$

where $S_i, T_i \in \mathbb{R}^{m\times m}$ are obtained by canceling the i-th row of $S, T \in \mathbb{R}^{(m+1)\times m}$. A simple induction on the dimension m (together with the Laplace cofactor expansion) proves that $\det\big(S_i - zT_i\big) \in \mathbb{P}_m$. Therefore, we conclude that $P \in \mathbb{P}_m$. □

Corollary 4.10 (Stability function is an approximation of exp). *Let $R(z)$ be the stability function of a Runge–Kutta method with consistency order $p \geq 1$. Then,*

$$R(z) = \exp(z) + \mathcal{O}(z^{p+1}). \tag{4.17}$$

Gabriel Cramer (1704–1752) was a Swiss Mathematician. He studied Mathematics at the University of Geneva and took his PhD in 1722. In 1724, he became professor at the University of Geneva. He had personal contact with his colleagues Johann Bernoulli, Leonhard Euler, and James Stirling.

Pierre-Simon (Marquis de) Laplace (1749–1827) was a French Mathematician and, from 1768–1771, a student of d'Alembert in Paris. In 1771, he became teacher for geometry, trigonometry, analysis, and statistics at the military academy in Paris. In 1773, be became associate member of the French Academy of Sciences.

Moreover, if the method is explicit, then it is a polynomial with

$$R(z) = \sum_{j=1}^{p} \frac{z^j}{j!} + \mathcal{O}(z^{p+1}), \tag{4.18}$$

and $\mathcal{O}(z^{p+1})$ vanishes for an explicit p-stage method of order p.

Proof. Consider the first step of the Runge–Kutta method for $\lambda = -1$ and $y_0 = 1$. Then,

$$\exp(-h) = y(h) \approx y_1 = R(-h)y_0 = R(-h).$$

Moreover, consistency order p thus implies that

$$\exp(-h) - R(-h) = \mathcal{O}(h^{p+1}). \tag{4.19}$$

Note that $R(z) = 1 + zb^\top(I - zA)^{-1}\mathbb{1}$ is smooth locally around $z = 0$ (i.e., analytic). Hence, it holds that

$$f(z) := R(z) - \exp(z) = \sum_{n=0}^{p} \frac{f^{(n)}(0)}{n!}\, z^n + \mathcal{O}(z^{p+1}).$$

Due to (4.19), we see that $f^{(n)}(0) = 0$ for all $n = 0, \dots, p$. Therefore, it follows that

$$R(z) = \exp(z) + \big(R(z) - \exp(z)\big) = \exp(z) + \mathcal{O}(z^{p+1}).$$

If the method is explicit, then $R \in \mathbb{P}_m$, i.e., $R(z) = \sum_{j=0}^{m} a_j z^j$. Then,

$$\sum_{j=0}^{p} \Big(a_j - \frac{1}{j!}\Big)\, z^j = R(z) - \exp(z) + \mathcal{O}(z^{p+1}) \overset{(4.17)}{=} \mathcal{O}(z^{p+1}).$$

Consequently, the lower-order powers of z vanish, i.e., $a_j = 1/j!$ for $j = 0, \dots, p$. This concludes the proof. □

Example 4.11. *The stability function of the Heun method (Example 2.18) and the modified Euler method (Example 2.19) is $R(z) = 1 + z + \frac{z^2}{2}$, since both methods are explicit 2-stage methods of order $p = 2$. The stability function of the classical RK4 method (Example 2.25) is $R(z) = 1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \frac{z^4}{24}$, since RK4 is a 4-step method of order $p = 4$.*

Definition 4.12. *Let $R(z)$ be the stability function of a one-step method. Then,*

$$S := \big\{z \in \mathbb{C} \,:\, |R(z)| \le 1\big\} \tag{4.20}$$

denotes the corresponding ***stability domain****.*

Remark 4.13. *If $\lambda h = z \in S$, then the approximations $y_\ell \approx y(t_\ell)$ of the model problem (4.13) remain bounded. Note that $y(t) = e^{\lambda t} y_0 \to 0$ as $t \to \infty$, since $\operatorname{Re}\lambda < 0$.*

Example 4.14. *The stability domain of the explicit Euler method ist*

$$S = \big\{z \in \mathbb{C} \,:\, |1 + z| \le 1\big\} = \overline{U_1(-1)}.$$

The stability domain of the implicit Euler method is

$$S = \left\{ z \in \mathbb{C} \,:\, \left| \frac{1}{1-z} \right| \leq 1 \right\} = \left\{ z \in \mathbb{C} \,:\, 1 \leq |1-z| \right\} = \mathbb{C} \backslash U_1(1).$$

Corollary 4.15. *For all Runge–Kutta methods with consistency order $p \geq 1$, it holds that $0 \in \partial S$.*

Proof. According to Corollary 4.10, it holds that

$$R(z) = \exp(z) + \mathcal{O}(z^{p+1}) = 1 + z + \mathcal{O}(z^2).$$

For all sufficiently small $z = \pm h$, it hence follows that

- $R(+h) = 1 + h + \mathcal{O}(h^2) \geq 1 + h/2 > 1$, i.e., $+h \notin S$;
- $R(-h) = 1 - h + \mathcal{O}(h^2) \leq 1 - h/2 < 1$, i.e., $-h \in S$.

Hence, each neighborhood of $z = 0$ contains one point $-h \in S$ and $+h \notin S$. This proves that $0 \in \partial S$. □

4.3. A-stability and L-stability.

Definition 4.16. *Let $R(z)$ be the stability function of a one-step method. The one-step method is*

- ***A-stable**, if $\sup_{\operatorname{Re} z \leq 0} |R(z)| \leq 1$;*
- ***L-stable**, if $\lim_{\operatorname{Re} z \to -\infty} |R(z)| = 0$.*

Remark 4.17. *Due to the definition of the stability domain S in* (4.20)*, A-stability is equivalent to $\mathbb{C}^- := \left\{ z \in \mathbb{C} \,:\, \operatorname{Re} z \leq 0 \right\} \subseteq S$. Hence, the explicit Euler method is* not *A-stable (see Example 4.14).*

Remark 4.18. *If a method is A-stable, then*

$$|y_{\ell+1}| = |R(\lambda h)||y_\ell| \leq |y_\ell| \quad \text{for all } h > 0 \text{ and } \operatorname{Re} \lambda \leq 0,$$

i.e., the discrete solutions of the model problem (4.13) *are non-expansive and, in particular, remain at least bounded.*

If a method is L-stable, then

$$|y_{\ell+1}| = |R(\lambda h)||y_\ell| \to 0 \quad \text{as } h \to \infty \text{ for all } \operatorname{Re} \lambda \leq 0,$$

i.e., the discrete solutions of the model problem (4.13) *decay with larger time-steps.*

Example 4.19. *The implicit Euler method satisfies that $R(z) = \frac{1}{1-z}$ and $S = \mathbb{C} \backslash U_1(1) \supset \mathbb{C}^-$. Hence, the implicit Euler method is L-stable and A-stable.*

Exercise 4.20. *Show that the stability function of the implicit midpoint rule is* $R(z) = \frac{1+z/2}{1-z/2}$. *Determine the stability domain of the implicit midpoint rule! Is the implicit midpoint rule A-stable and/or L-stable?*

Theorem 4.21 (Explicit Runge–Kutta methods fail). *No explicit Runge–Kutta method is L-stable. No consistent explicit Runge–Kutta method is A-stable. In particular, any A-stable and/or L-stable Runge–Kutta method is implicit.*

Proof (L-stability fails). Recall that the stability function of an m-stage Runge–Kutta method satisfies that $R = P/Q$ with $P, Q \in \mathbb{P}_m$. Obviously, L-stability thus is equivalent to the fact that $P \in \mathbb{P}_{\nu-1}$ and $Q \in \mathbb{P}_\nu \backslash \mathbb{P}_{\nu-1}$ for some $1 \le \nu \le m$ to ensure that $P/Q = \mathcal{O}(1/z)$ as $|z| \to \infty$. □

Proof (A-stability fails). The stability function of an explicit m-stage Runge–Kutta method $\begin{array}{c|c} c & A \\ \hline & b^\top \end{array}$ satisfies that $R \in \mathbb{P}_m$. Consistency implies that $\sum_{j=1}^m b_j = 1$ and hence the method has consistency order $p \ge 1$. Together with Corollary 4.10, this implies that

$$R(z) = \sum_{j=0}^{\nu} a_j z^j \quad \text{for some } 1 \le \nu \le m \text{ with } a_\nu \neq 0.$$

The triangle inequality thus proves that

$$|R(z)| \ge |a_\nu||z|^\nu - \sum_{j=0}^{\nu-1} |a_j||z|^j = |z|^\nu \Big(|a_\nu| - \sum_{j=0}^{\nu-1} |a_j| \frac{1}{|z|^{\nu-j}} \Big) \xrightarrow{\operatorname{Re} z \to -\infty} \infty.$$

Hence, $\sup_{\operatorname{Re} z \le 0} |R(z)| = \infty$ and the method is *not* A-stable. □

Theorem 4.22. *Let* $\begin{array}{c|c} c & A \\ \hline & b^\top \end{array}$ *be an implicit Runge–Kutta method such that* A *is invertible. Then, the following statements* (i)–(ii) *are equivalent:*

(i) $b^\top A^{-1} \mathbb{1} = 1 \quad$ *with* $\quad \mathbb{1} = (1, \dots, 1)^\top \in \mathbb{R}^m$.

(ii) *The method is L-stable.*

Proof. If $\| \frac{1}{z} A^{-1} \| < 1$ (e.g., $|z| > 2\|A^{-1}\|$), then the Neumann series proves that

$$\Big(I - \frac{1}{z} A^{-1} \Big)^{-1} = \sum_{j=0}^{\infty} \big(\frac{1}{z} A^{-1} \big)^j = I + \sum_{j=1}^{\infty} \big(\frac{1}{z} A^{-1} \big)^j = I + \frac{1}{z} A^{-1} \Big(I + \sum_{j=1}^{\infty} \big(\frac{1}{z} A^{-1} \big)^j \Big).$$

Moreover,

$$I - zA = (A^{-1} - zI)A = -z \Big(I - \frac{1}{z} A^{-1} \Big) A.$$

This leads to

$$(I - zA)^{-1} = -\frac{1}{z} A^{-1} \Big(I + \frac{1}{z} A^{-1} \Big)^{-1} = -\frac{1}{z} A^{-1} \Big[I + \frac{1}{z} A^{-1} \Big(I + \sum_{j=1}^{\infty} \big(\frac{1}{z} A^{-1} \big)^j \Big) \Big].$$

With Theorem 4.9, it follows that

$$R(z) \overset{(4.15)}{=} 1 + zb^\top (I - zA)^{-1}\mathbb{1} = 1 - b^\top A^{-1}\left[I + \frac{1}{z}A^{-1}\left(I + \sum_{j=1}^{\infty}\left(\frac{1}{z}A^{-1}\right)^j\right)\right]\mathbb{1}$$

$$= 1 - b^\top A^{-1}\mathbb{1} - \frac{1}{z}\left[b^\top A^{-1}A^{-1}\left(I + \sum_{j=1}^{\infty}\left(\frac{1}{z}A^{-1}\right)^j\right)\right]\mathbb{1}.$$

Hence, L-stability is equivalent to $1 - b^\top A^{-1}\mathbb{1} = 0$. This concludes the proof. □

Remark 4.23. *Suppose the assumptions of Theorem 4.22. If the j-th column of A is constant $b_j \neq 0$, then the j-th unit vector $e_j \in \mathbb{R}^m$ leads to $Ae_j = b_j\mathbb{1}$. This implies that $A^{-1}\mathbb{1} = \frac{1}{b_j}e_j$ and results in*

$$b^\top A^{-1}\mathbb{1} = \frac{1}{b_j}b^\top e_j = 1.$$

According to Theorem 4.22, the method is L-stable.

Remark 4.24 (Radau-IIA methods). *The Radau-IIA methods are collocation methods with $c_m = 1$. The remaining nodes c_j of $c \in \mathbb{R}^m$ are chosen such that the induced quadrature rule has maximal exactness. Arguing as for the Gaussian quadrature rule, one obtains that the nodes $0 < c_1 < c_2 < \cdots < c_m = 1$ are unique. The quadrature rule has exactness $2m - 2$ (i.e., one order less than the Gaussian quadrature). Hence, the Radau-IIA methods have consistency order $p = 2m - 1$. Recall from Theorem 3.25 that the Runge–Kutta data of a collocation method read*

$$A_{ij} = \int_0^{c_i} L_j \,\mathrm{d}t \quad \text{and} \quad b_j = \int_0^1 L_j \,\mathrm{d}t \quad \text{for all } i, j = 1, \ldots, m.$$

For Radau-IIA methods, the last row of $A \in \mathbb{R}^{m\times m}$ thus coincides with $b \in \mathbb{R}^m$ (since $c_m = 1$). Moreover, A is invertible, since $c_j > 0$. With the m-th unit vector $e_m \in \mathbb{R}^m$, this means that $b^\top = e_m^\top A$ and hence $b^\top A^{-1} = e_m^\top$. Consequently,

$$b^\top A^{-1}\mathbb{1} = e_m^\top\mathbb{1} = 1.$$

According to Theorem 4.22, the Radau-IIA methods are L-stable.

Exercise 4.25. *Show that the implicit Euler method is the unique Radau-IIA method for $m = 1$.*

Remark 4.26. *The A-stability of collocation methods is fully understood mathematically; see* [But08, Theorem 358A]*. One can show that all Radau-IIA methods are A-stable and L-stable. Moreover, one can show that all Gauss methods are A-stable, but fail to be L-stable.*

Jean Charles Rodolphe Radau (1835–1911) was a German-French mathematician. Born in Germany, he studied mathematics and astronomy at the University of Königsberg. In 1858, he moved to Paris and later became French citizen in 1873. In 1897, he became meber of the French Academy of Sciences.

Theorem A.1 (Banach fixpoint theorem). *Let M be a complete metric space and $\mathrm{A} : M \to M$ be a contraction, i.e., there exists a contraction constant $0 < \kappa < 1$ such that*

$$\forall x, y \in M : \quad d(\mathrm{A}x, \mathrm{A}y) \leq \kappa\, d(x, y). \tag{A.1}$$

Then, A has a unique fixpoint $w \in M$, i.e., $\mathrm{A}w = w$. Moreover, for any initial value $w_0 \in M$, the sequence of Banach iterates $w_{n+1} := \mathrm{A}w_n$ converges to w and it holds that

$$d(w, w_n) \leq \frac{\kappa}{1-\kappa} d(w_{n+1}, w_n) \leq \frac{\kappa^n}{1-\kappa} d(w_1, w_0) \quad \text{for all } n \in \mathbb{N}_0. \tag{A.2}$$

Remark A.2. *Often, the Banach fixpoint theorem is applied in the following setting: Let X be a Banach space and $M \subseteq X$ be a closed subspace. Suppose that $\mathrm{A} : M \to M$ is a contraction with respect to the natural metric, i.e., there exists a contraction constant $0 < \kappa < 1$ such that*

$$\forall x, y \in M : \quad \|\mathrm{A}x - \mathrm{A}y\|_X \leq \kappa\, \|x - y\|_X. \tag{A.3}$$

Then, the Banach fixpoint theorem applies.

Theorem A.3 (Implicit function theorem). *Let $f \in C^1(\mathbb{R}^m \times \mathbb{R}^n; \mathbb{R}^n)$. Let $(x, y) \in \mathbb{R}^m \times \mathbb{R}^n$ with $f(x, y) = 0$. Suppose that the Jacobi matrix $D_y f(x, y) \in \mathbb{R}^{n \times n}$ is invertible. Then, there exist open sets $U \subset \mathbb{R}^m$ and $V \subset \mathbb{R}^n$ with $(x, y) \in U \times V$ as well as a function $g \in C^1(U; V)$ such that*

$$\forall\, (\widetilde{x}, \widetilde{y}) \in U \times V : \quad \big[\, f(\widetilde{x}, \widetilde{y}) = 0 \quad \Longleftrightarrow \quad \widetilde{y} = g(\widetilde{x}) \,\big] \tag{A.4}$$

References

[But08] John C. Butcher: Numerical methods for ordinary differential equations, Wiley, Chichester, second edition, 2008.

[SWP12] Karl Strehmel, Rüdiger Weiner, Helmut Podhaisky: Numerik gewöhnlicher Differentialgleichungen, Springer, Berlin, second edition, 2012 [in German].

[Wal00] Wolfgang Walter: Gewöhnliche Differentialgleichungen, Springer, Berlin, second edition, 2000 [in German].

Stefan Banach (1892–1945) was a Polish mathematician. Being the founder of modern functional analysis, he is amongst the most important mathematicians of the 20th century. His major work was the 1932 book *"Théorie des opérations linéaires"*, the first monograph on the general theory of functional analysis. Banach finished his PhD in mathematics at Lviv University in 1922 (the thesis also contained the Banach fixpoint theorem). In the same year, he completed the habilitation and became associate professor at Lviv University. In 1926, he was promoted to full professor. In August 1945, he died of lung cancer.