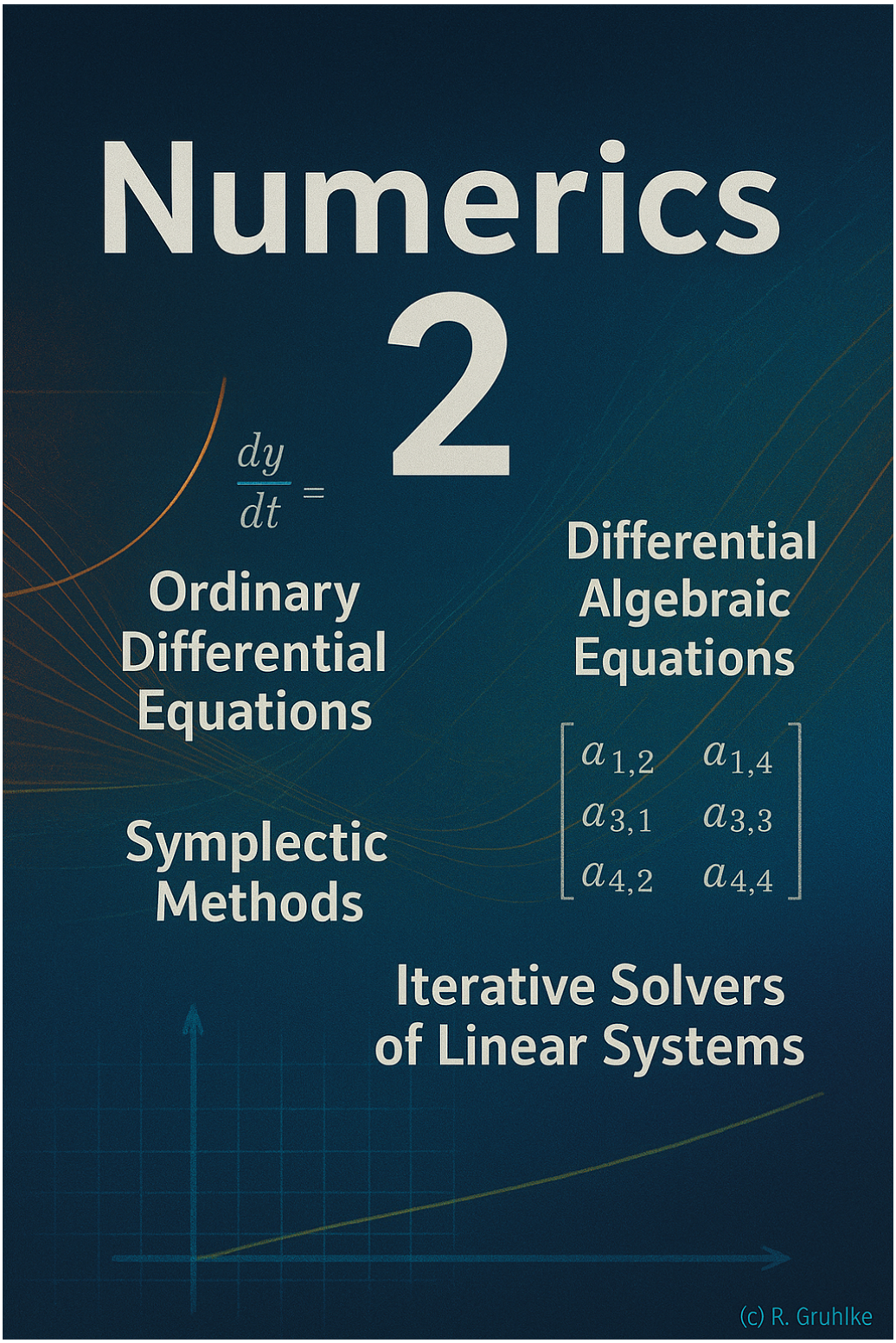
Freie Universität Berlin

FB Mathematik und Informatik

Numerics II

Winter term 2025/2026

Dr. Robert Gruhlke



Stand: November 7, 2025

Preface

This script accompanies the lecture Numerics II, taught by the author at Freie Universität Berlin during the winter term 2025/2026. It draws on several sources and follows presentation styles inspired by Numerics II lectures from Prof. Harry Yserentant (TU Berlin) and Prof. Marcus Bachmayr (RWTH Aachen).

This lecture script is intended as an aid for reviewing and following up on the material covered in the course. It does not, and is not meant to, reproduce the lecture verbatim. In the lectures, we will motivate the definitions, theorems, and algorithms, discuss their results, and illustrate them with examples.

Suggestions for improvements of any kind are very welcome: **r.gruhlke@fu-berlin.de**

Contents

1. Numerical Methods for Ordinary Differential Equations

Examples and Motivation

For reasons of limited time, various (derivation) examples for ordinary differential equations are omitted in this lecture and maybe added at a later stage for the interested reader. For some applications, please refer to the introductory information slides or to previous courses such as Computerorientierte Mathematik II (FU Berlin) or Numerics I (FU Berlin).

1.1 Regularity, Existence and Uniqueness

Consider the ordinary differential equation (ODE) of **first order** for $G \subset \mathbb{R}^n$ open for $n \in \mathbb{N}$ and $t \in [a,b]$ and $f\colon [a,b] \times G \to \mathbb{R}^n$ given as

$$y'(t) = f(t, y(t)), \quad a \leq t \leq b. \tag{1.1}$$

A solution (if existing) $y\colon [a,b] \to \mathbb{R}^n$ of (1.1) is called *classical solution* if it continuous differentiable, i.e. $y \in \mathcal{C}([a,b]; \mathbb{R}^n)$.

Theorem 1.1 (Regularity) If $f\colon [a,b] \times G \to \mathbb{R}^n$ is continuous, then y is continuously differentiable. If f is r-times continuously differentiable, then y is $(r+1)$-times continuously differentiable.

Proof. By induction. □

In general there are infinite solutions of the ODE 1.1. In order to fix a possible unique solution, additional contraints are required.

Definition 1.2 (Initial value problem (IVP)) An initial value problem is given for $y_0 \in G$ as

$$\begin{cases} y'(t) = f(t, y(t)), & a \leq t \leq b, \\ y(a) = y_0 \end{cases} \tag{IVP}$$

Remark 1.3 (General form and high order ODEs) A general m-th order ordinary differential equation is given of the form

$$F(t, y(t), y'(t), \dots, y^{(m)}(t)) = 0.$$

The equation is called explicit, if it can be written as

$$y^{(m)} = f\left(t, y(t), y'(t), \dots, y^{(m-1)}(t)\right). \tag{1.2}$$

Every explicit differential equation of m-th order can be transformed to a system of m differential equations of first order using

$$\begin{aligned} y_i'(t) &= y_{i+1}(t), \qquad i = 1, \dots, m-1, \\ y_m'(t) &= f(t, y_1(t), \dots, y_m(t)) \end{aligned}$$

or compactly written for $\boldsymbol{y}(t) = (y_1(t), \dots, y_m(t))^\top$ as

$$\boldsymbol{y}'(t) = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ \vdots & & & \ddots & 1 \\ 0 & \cdots & \cdots & \cdots & 0 \end{pmatrix} \boldsymbol{y}(t) + \begin{pmatrix} 0 \\ \vdots \\ \vdots \\ 0 \\ f(t, y_1, \dots, y_m) \end{pmatrix}$$

Then, y_1 solves (1.2). Moreover, together with the conditions

$$y(a) = y_0, \quad y'(a) = y_0', \quad \dots, y^{(m-1)}(a) = y_{m-1}'$$

the general form becomes an initial value problem. △

Theorem 1.4 (Peano - Existence of solutions) Assume $f\colon [a,b] \times \mathbb{R}^n \to \mathbb{R}^n$ is continuous and bounded. Then, for any $y_0 \in \mathbb{R}^n$ the initial value problem (IVP) has a solution.

Proof. See, e.g., Theorem 1.2. in [3]. □

We call a continuous function $f\colon [a,b] \times \mathbb{R}^n \to \mathbb{R}^n$, Lipschitz w.r.t. to its second input if there exists $L > 0$ such that

$$\|f(t, y_1) - f(t, y_2)\| \le L\|y_1 - y_2\|, \qquad \forall t \in [a,b], y_1, y_2 \in \mathbb{R}^n. \tag{1.3}$$

Theorem 1.5 (Picard-Lindelöf - Uniqueness of solution) Assume $f\colon [a,b] \times \mathbb{R}^n \to \mathbb{R}^n$ is continuous w.r.t. t and Lipschitz-continuous w.r.t. y. Then, for any $y_0 \in \mathbb{R}^n$ the initial value problem (IVP) has a **unique** solution.

Proof. See, e.g., Theorem 1.1. in [3]. □

Remark 1.6 Theorem 1.4 has a local version for $f\colon [a,b] \times G$ continuous only. Theorem 1.5 has a local version for a continuous differentiable right hand side f, yielding local existence and uniqueness. Moreover, the boundedness in Theorem 1.4 is needed for the existence of the solution

y on the whole interval $[a,b]$. For this consider the ODE $y(t)' = y(t)^2$ on the interval $[0,2]$ with general form of solutions $y(t) = \frac{1}{c-t}$ for any $c \in \mathbb{R}$. So let $y(0) = 1$, then $c = 1$ and the solution is not defined at $t = 1$. In particular the right hand side does not satisfy the assumptions of Picard-Lindelöf. △

We introduce the space $\mathcal{C}_{\mathrm{PL}}$ of right hand sides for our IVP satisfying the conditions for the Picard-Lindelöf theorem, i.e.

$$\begin{aligned}\mathcal{C}_{\mathrm{PL}} &\coloneqq \mathcal{C}_{\mathrm{PL}}([a,b]\times\mathbb{R}^n;\mathbb{R}^n)\\ &\coloneqq \{f\colon [a,b]\times\mathbb{R}^n \to \mathbb{R}^n \text{ continuous and satisfies (1.3)}\}.\end{aligned}$$

Example 1.7 A linear ODE is of the form

$$y'(t) = A(t)y(t) + g(t), \qquad a \le t \le b,$$

with continuous matrix-valued function A and continuous function g. The right hand side is Lipschitz w.r.t. the y coordinate with constant

$$L = \max_{a\le t\le b} \|A(t)\|_2.$$

Consequently, IVPs based on linear ODEs are uniquely solvable by Theorem 1.5. △

As a next result we want to understand the effect of perturbation of the initial value for general IVPs and for the case of dissipative systems. For this we need an auxillary result

Lemma 1.8 (Gronwall Lemma) Let $\Phi\colon [a,b] \to \mathbb{R}$ continuous, $\alpha \in \mathbb{R}$ and $\beta > 0$ such that

$$\Phi(t) \le \alpha + \beta \int_a^t \Phi(s)\,\mathrm{d}s, \qquad a \le t \le b.$$

Then, it holds $\Phi(t) \le \alpha e^{\beta(t-a)}$ for all $t \in [a,b]$.

Proof. Let $\epsilon > 0$ and $\Psi(t) := (\alpha+\epsilon)e^{\beta(t-a)}$. Then either $\Phi(t) < \Psi(t)$ or the set $\{t \in [a,b] \mid \Phi(t) \ge \Psi(t)\}$ is not empty. Since the latter set is compact subset of $\mathbb{R}$, there exists a minimal t_0. Since, for this t_0 it holds $\Phi(t_0) \ge \Psi(t_0)$ and $\Phi(a) \le \alpha < \alpha + \epsilon = \Psi(a)$ it must holds $t_0 > a$. Consequently, $\Phi(t) < \Psi(t)$ for $a \le t \le t_0$ and since $\beta > 0$ it holds

$$\Phi(t_0) \le \alpha + \beta\int_a^{t_0} \Phi(t)\,\mathrm{d}t < \alpha + \beta\int_a^{t_0} \Psi(t)\,\mathrm{d}t = \alpha + \int_a^{t_0} \Psi'(t)\,\mathrm{d}t = \Psi(t_0) - \epsilon < \Psi(t_0).$$

This is a contradiction to the definition of t_0. It follows that

$$\Phi(t) < \Psi(t) = (\alpha+\epsilon)e^{\beta(t-a)}.$$

Since $\epsilon > 0$ can be chosen arbitrary small, it yields the claim. □

Theorem 1.9 Let $f \in \mathcal{C}_{\mathrm{PL}}$ with Lipschitz constant $L > 0$ and let $y, z\colon [a,b] \to \mathbb{R}^n$ be the solution of the IVPs

$$y'(t) = f(t, y(t)), \quad y(a) = y_0 \qquad \text{and} \qquad z'(t) = f(t, z(t)), z(a) = z_0.$$

Then, it holds

$$\|y(t) - z(t)\| \leq e^{L(t-a)} \|y_0 - z_0\|, \qquad a \leq t \leq b. \tag{1.4}$$

Proof. For all $t \in [a,b]$ it holds

$$y(t) = y_0 + \int_a^t f(s, y(s))\,\mathrm{d}s, \qquad z(t) = z_0 + \int_a^t f(s, z(s))\,\mathrm{d}s.$$

Hence,

$$\begin{aligned}
\|y(t) - z(t)\| &= \left\| (y_0 - z_0) + \int_a^t f(s, y(s)) - f(s, z(s))\,\mathrm{d}s \right\| \\
&\leq \|y_0 - z_0\| + \int_a^t \|f(s, y(s)) - f(s, z(s))\|\,\mathrm{d}s \\
&\leq \|y_0 - z_0\| + L \int_a^t \|y(s) - z(s)\|\,\mathrm{d}s
\end{aligned}$$

Then the claim follows by application of Gronwall Lemma 1.8 with $\Phi(t) = \|y(t) - z(t)\|$, $\alpha = \|y_0 - z_0\|$ and $\beta = L$. □

Theorem 1.10 Let $f \in \mathcal{C}_{\mathrm{PL}}$ and it satisfies

$$\langle y_1 - y_2, f(t, y_1) - f(t, y_2)\rangle \leq \theta \|y_1 - y_2\|^2 \qquad \forall t \in [a,b], y_1, y_2 \in \mathbb{R}^n$$

for an inner product $\langle \cdot, \cdot \rangle$ on $\mathbb{R}^n$ with induced norm $\|\cdot\|$ and for $\theta \in \mathbb{R}$. Then, the solutions $t \mapsto y_1(t), \mapsto y_2(t)$ of the IVPs

$$y'(t) = f(t, y(t)), \quad t \in [a,b], \quad y(a) \in \{y_1(a), y_2(a)\},$$

satisfy

$$\|y_1(t) - y_2(t)\| \leq e^{\theta(t-a)} \|y_1(a) - y_2(a)\|.$$

Proof. Define the function $\Psi(t) = \|y_1(t) - y_2(t)\|^2 > 0$. It satisfies

$$\begin{aligned}
\Psi'(t) &= 2\langle y_1(t) - y_2(t), y_1'(t) - y_2'(t)\rangle \\
&= 2\langle y_1(t) - y_2(t), f(t, y_1(t)) - f(t, y_2(t))\rangle \\
&\leq 2\theta \|y_1(t) - y_2(t)\|^2 \\
&= 2\theta \Psi(t).
\end{aligned}$$

Hence,

$$\frac{\Psi'(s)}{\Psi(s)} \leq 2\theta, \qquad \forall s \in [a,b].$$

Integration over $[a,t]$, yields

$$\ln(\Psi(t)) - \ln(\Psi(a)) \leq 2\theta(t-a) \qquad \Leftrightarrow \qquad \Psi(t) \leq e^{2\theta(t-a)}\Psi(a).$$

□

Remark 1.11 If $f \in \mathcal{C}_{\mathrm{PL}}$ with Lipschitz constant L then automatically

$$\langle y_1 - y_2, f(t,y_1) - f(t,y_2)\rangle \leq \|y_1 - y_2\|\|f(t,y_1) - f(t,y_2)\| \leq L\|y_1 - y_2\|^2.$$

However, Theorem 1.10 includes the case of arbitrary $\theta \in \mathbb{R}$. △

The relevance of Theorem 1.10 becomes visible in the case of dissipative systems, see Example 1.12. We call a system $y'(t) = f(t,y(t))$ *monotone or dissipative in the sense of Dahlquist* with respect to a inner product $\langle\cdot,\cdot\rangle$ on $\mathbb{R}^n$ with induced norm $\|\cdot\|$ if there exists $\eta > 0$ such that for all $y_1, y_2 \in \mathbb{R}^n$:

$$\langle f(t,y_1) - f(t,y_2), y_1 - y_2\rangle \leq -\eta\|y_1 - y_2\|^2. \tag{1.5}$$

Example 1.12 Consider the IVPs for $\lambda \in \mathbb{R}$ given as

$$y'(t) = \lambda y(t), \quad y(a) = y_0 \qquad \text{and} \qquad z'(t) = \lambda z(t), \quad z(a) = z_0$$

with Lipschitz constant $L = |\lambda|$ of the right hand side. In this case

$$y(t) - z(t) = e^{\lambda(t-a)}(y(a) - z(a)).$$

Consequently, following Remark 1.11 the choice $\theta = L = |\lambda|$ is only sharp if $\lambda > 0$ and else represents a huge overestimation for the case of $\lambda < 0$. △

Example 1.13 Consider

$$y'(t) = A(t)y(t), \quad y(0) = y_0$$

with

$$A(t) \equiv A := \begin{pmatrix} \alpha & \beta \\ -\beta & \alpha \end{pmatrix}$$

for $\alpha, \beta \in \mathbb{R}$. Then, $L = \sqrt{\alpha^2 + \beta^2}$ and $\theta = \alpha$. Exercise. △

1.2 Explicit One-step methods

We seek for approximate values Y_k for the solution $y(t_k)$ for points t_k of a grid

$$a = t_0 < t_1 < \ldots < t_N = b, \tag{1.6}$$

of the interval $[a,b]$. In view of a one-step scheme to generate the sequence $(Y_k)_k$, we denote $h_k = t_{k+1} - t_k$ the step size. For equidistant grids, i.e. $h_0 = h_1 = \ldots = h_{N-1}$ the notation of h is used.

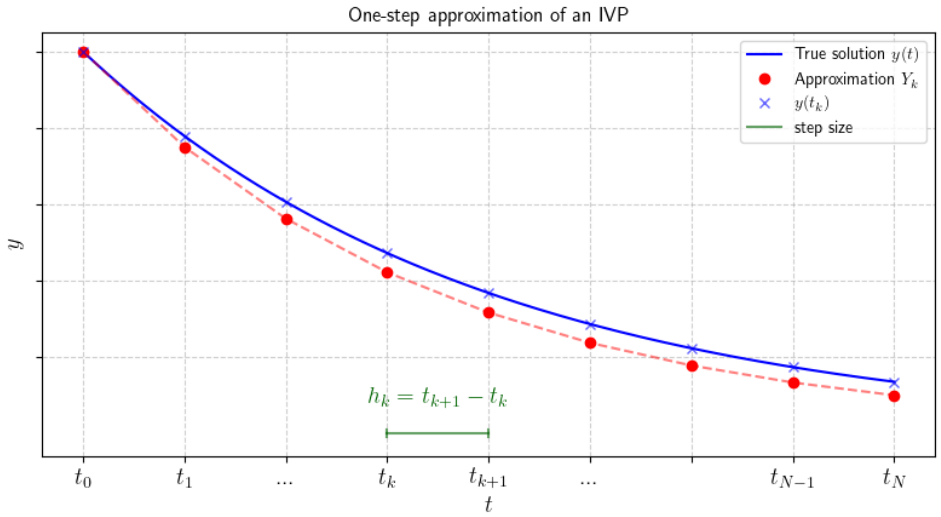


Figure 1.1: Illustration of approximation of $t \mapsto y(t)$ on grid points t_k.

Definition 1.14 (Incremental function) An explicit one-step method is given as

$$Y_{k+1} = Y_k + h_k \Phi(t_k, Y_k, h_k), \quad k = 0, \dots, N-1, \qquad Y_0 = y(t_0), \tag{1.7}$$

and determined by an *incremental function* $\Phi \colon t \times \mathbb{R}^n \times \mathbb{R}_+ \to \mathbb{R}^n$.

Example 1.15 (Explicit Euler method) Motivated by the first order approximation using rectangle rule given as

$$y(t_{k+1}) = y(t_k) + \int_{t_k}^{t_{k+1}} f(t, y(t)\,\mathrm{d}t \approx y(t_k) + h_k f(t_k, y(t_k))$$

the *explicit Euler method* reads

$$Y_{k+1} = Y_k + h_k f(t_k, Y_k). \tag{1.8}$$

Consequently, $\Phi(t, Y, h) = f(t, Y)$. △

Example 1.16 (Improved Euler method) Motivated by the midpoint rule for integration taking the form

$$y(t_{k+1}) = y(t_k) + \int_{t_k}^{t_{k+1}} f(t, y(t))\,\mathrm{d}t \approx y(t_k) + h_k f(t_k + \tfrac{h_k}{2}, y(t_k + \tfrac{h_k}{2}))$$

one obtains the scheme

$$Y_{k+1} = Y_k + h_k f\left(t_k + \tfrac{h_k}{2}, Y_k + \tfrac{h_k}{2} f(t_k, Y_k)\right). \tag{1.9}$$

Consequently, $\Phi(t, Y, h) = f\left(t + \frac{h}{2}, Y + \frac{h}{2} f(t, Y)\right)$. △

Example 1.17 Consider the initial value problem

$$\begin{aligned} y'(t) &= \cos(t)y \qquad -8 \le t \le 0 \\ y(-8) &= \exp(\sin(-8)) \end{aligned} \tag{1.10}$$

with solution $y(t) = \exp(\sin(t))$.

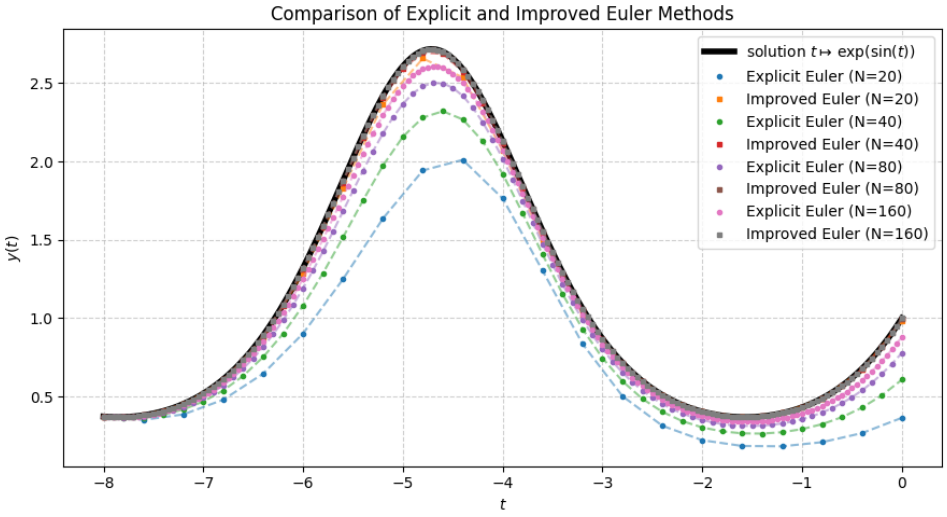


Figure 1.2: Numerical results of explicit and improved Euler methods for various equidistant grids for Example 1.17.

Table 1.1: Error $|Y_N - y(t)|$ at endpoint $t = 0$ for the explicit and improved Euler for Example 1.17

N	**Explicit Euler error**	**Improved Euler error**
20	6.362×10^{-1}	1.928×10^{-2}
40	3.929×10^{-1}	6.105×10^{-3}
80	2.218×10^{-1}	1.719×10^{-3}
160	1.184×10^{-1}	4.549×10^{-4}

Observation: Explicit Euler converges as $\mathcal{O}(h)$ and improved Euler method as $\mathcal{O}(h^2)$. △

There are two main sources of errors for the discretization of IVP:

1. **Local error**: Each iteration introduces a new error.
2. **Error propagation**: Errors introduced once are propagated and possible amplified by the recursion.

1.2.1 Consistency, Stability and Convergence

This section is devoted to the general theory leading to convergence of one-step methods.
As a first step we formalize the introduced error by a single step of the method.

Definition 1.18 (Local truncation error) The quantity

$$\tau(t,h) := \frac{1}{h}(y(t+h) - y(t)) - \Phi(t, y(t), h) \tag{1.11}$$

is called the local truncation error (*dt. Abbruchfehler*) of the one-step method at the solution point $y(t)$.

In other words, it holds

$$y(t_{k+1}) = y(t_k) + h_k\Phi(t_k, y(t_k), h_k) + h_k\tau(t_k, h_k).$$

Then, the one-step method at k-th step yields a value

$$\widehat{y}_{k+1} = y(t_k) + h_k\Phi(t_k, y(t_k), h_k)$$

leading to the local deviation $\widehat{d}_{k+1} = y(t_{k+1}) - \widehat{y}_{k+1}$ with error

$$\|\widehat{d}_{k+1}\| = \|y(t_{k+1}) - \hat{y}_{k+1}\| = \|y(t_{k+1}) - (y(t_k) + h_k\Phi(t_k, y(t_k), h_k)\|.$$

Hence, the design of Φ should be necessarily meaningful in terms of controlling the local truncation error, ideally $\tau(t,h) \in \mathcal{O}(h^q)$ for some $q > 0$.

Definition 1.19 (Consistency of one-step methods) Let $h = \max_k h_k$. Then, the one-step method (1.7) is called consistent if for any $f \in \mathcal{C}_{\mathrm{PL}}$ and corresponding unique solution y of (IVP), it holds

$$\lim_{h\to 0}\left(\max_{k=0,\dots,N-1}\|\tau(t_k,h_k)\|\right) = 0. \tag{1.12}$$

The condition (1.12) is equivalent to

$$\lim_{h\to 0}\left(\max_{k=0,\dots,N-1}\|f(t_k, y(t_k)) - \Phi(t_k, y(t_k), h_k)\|\right) = 0.$$

In particular, consistency requires that the incremental function approximates the derivative y' of the function y sufficiently well. This allows a comparison of different one-step methods.
Let us for simplicity assume that $h_k \equiv h$ in the following.

Definition 1.20 (Order of consistency) An explicit one-step method has consistency order $q \in \mathbb{N}$ if, q is the largest natural number such that for any $f \in \mathcal{C}_{\mathrm{PL}}$ and related incremental function $\Phi = \Phi(f)$ it holds

$$\|\tau(t_k,h)\| \le Ch^q, \qquad \forall k = 0,\dots,N-1$$

for a constant $C = C(y,f)$ and all $h \in (0,\overline{h})$ for some $\overline{h} > 0$.

Remark 1.21 The constant $C = C(y,f)$ may depend on (partial) derivatives of y and f. △

Example 1.22 The explicit Euler method has consistency order $q = 1$. Assume $y \in \mathcal{C}^2([a,b];\mathbb{R}^n)$. By Taylor expansion with integral remainder it holds

$$y(t_k + h) = y(t_k) + h_k y'(t_k) + \int_{t_k}^{t_k+h_k} (t_k + h_k - s)y''(s)\mathrm{d}s.$$

Consequently, since for the explicit Euler method $\Phi(t_k, y(t_k), h_k) = f(t, y(t_k)) = y'(t_k)$ it holds

$$\begin{aligned}
h_k\|\tau(t_k,h_k)\| &= \|y(t_k+h) - \widehat{y}_{k+1}\| \\
&= \left\| y(t_k) + h_k y'(t_k) + \int_{t_k}^{t_k+h} (t_k + h_k - s)y''(s)\mathrm{d}s - y(t_k) - h_k f(t_k, y(t_k)) \right\| \\
&= \left\| \int_{t_k}^{t_k+h_k} (t_k + h_k - s)y''(s)\mathrm{d}s \right\| \\
&\leq \left(\int_{t_k}^{t_k+h_k} (t_k + h_k - s)y''(s)\mathrm{d}s \right) \max_{s\in[t_k,t_k+h_k]} \|y''(s)\| \\
&\leq \frac{h_k^2}{2}\|y''\|_\infty
\end{aligned}$$

. Hence $\|\tau(t_k,h_k)\| \in \mathcal{O}(h_k)$. △

Definition 1.23 (Global error) The global error of the one-step method is defined as

$$\max_{k=0,\dots,N-1} \|Y_k - y(x_k)\|.$$

Now the goal is to estimate the global error with the local truncation error.

Theorem 1.24 (Stability) Let y be solution of (IVP) defined on the whole interval $[a,b]$ and for some $\overline{h} > 0$ assume that the incremental function $\Phi\colon [a,b] \times \mathbb{R}^n \times [0,\overline{h}]$ is uniformly L-Lipschitz w.r.t the second component, i.e.

$$\|\Phi(t,y_1,h) - \Phi(t,y_2,h)\| \leq L\|y_1 - y_2\|, \qquad y_1, y_2 \in \mathbb{R}^n,$$

for $t \in [a,b], h \in [0,\overline{h}]$. Then, it holds the stability estimate

$$\|y(t_k) - Y_k\| \leq e^{L(t_k-a)} \sum_{j=0}^{k-1} h_j \|\tau(t_j,h_j)\|, \qquad k = 1,\dots,N.$$

Proof. For $k = 0, \dots, N-1$ it holds that

$$\begin{aligned}\|y(t_{k+1}) - Y_{k+1}\| &= \|y(t_k) + h_k\Phi(t_k, y(t_k), h_k) + h_k\tau(t_k, h_k) - [Y_k + h_k\Phi(t_k, Y_k, h_k)]\| \\ &\le \|y(t_k) - Y_k\| + h_k\|\Phi(t_k, y(t_k), h_k) - \Phi(t_k, Y_k, h_k)\| + h_k\|\tau(t_k, h_k)\| \\ &\le \|y(t_k) - Y_k\| + h_k L\|y(t_k) - Y_k\| + h_k\|\tau(x_k, h_k)\|.\end{aligned}$$

Since $1 + t \le e^t$ for $t \ge 0$, it follows

$$\|y(t_{k+1}) - Y_{k+1}\| \le e^{Lh_k}\|y(t_k) - Y_k\| + h_k\|\tau(t_k, h_k)\|.$$

Hence, by induction

$$\begin{aligned}\|y(t_{k+1}) - Y_{k+1}\| &\le e^{Lh_k}\left(e^{L(t_k-a)}\sum_{j=0}^{k-1} h_j\|\tau(t_j, h_j)\|\right) + h_k\|\tau(t_k, h_k)\| \\ &\le e^{L(t_{k+1}-a)}\sum_{j=0}^{k} h_j\|\tau(t_j, h_j)\|\end{aligned}$$

□

Example 1.25 Let $f \in \mathcal{C}_{\mathrm{PL}}$ with Lipschitz constant L, then the improved Euler-Method with increment function $\Phi(t, y, h) = f(t + \frac{h}{2}, y + \frac{h}{2}f(t, y))$ satisfies

$$\|\Phi(t, y_1, h) - \Phi(t, y_2, h)\| \le (1 + \tfrac{h}{2}L)L\|y_1 - y_2\|$$

for $0 \le h \le b - a$, $a \le t \le b - h$ and all $y_1, y_2 \in \mathbb{R}^n$. △

Remark 1.26 The constant $e^{L(t_k-a)}$ in Theorem 1.24 can become very large due to the dependence of L from the incremental function Φ, that in turn is dependent on the Lipschitz constant of f. An improvement is obtained with the next Theorem. △

Theorem 1.27 (improved Stability) Let y be solution of (IVP) defined on the whole $[a, b]$. Let $\|\cdot\| = \langle\cdot,\cdot\rangle^{\frac{1}{2}}$ and for some $\overline{h} > 0$ assume that the incremental function $\Phi\colon [a, b] \times \mathbb{R}^n \times [0, \overline{h}]$ is uniformly L-Lipschitz w.r.t the second component, i.e.

$$\|\Phi(t, y_1, h) - \Phi(t, y_2, h)\| \le L\|y_1 - y_2\|, \qquad y_1, y_2 \in \mathbb{R}^n,$$

for $t \in [a, b], h \in [0, \overline{h}]$ and satisfies

$$\langle y_1 - y_2, \Phi(t, y_1, h) - \Phi(t, y_2, h)\rangle \le \theta\|y_1 - y_2\|^2$$

for $\theta \in \mathbb{R}$. Then for $h = \max_k h_k$ it holds the estimate

$$\|y(t_k) - Y_k\| \le e^{\max\left\{\theta + \frac{h}{2}L^2, 0\right\}(t_k-a)}\sum_{j=0}^{k-1} h_j\|\tau(t_j, h_j)\|, \qquad k = 1, \dots, N.$$

Proof. Exercise :). □

Now we define the concept of convergence for the one-step methods in mind.

Definition 1.28 (Convergence) A one-step method converges for our (IVP) on the interval $[a, b]$ if for each sequence of step sizes (h_k) such that $h = \max_k h_k \to 0$ it holds that the global error goes to zero. In particular

$$\max_{k=0,\dots,N-1} \|Y_k - y(t_k)\| \to 0 \text{ as } h \to 0.$$

The method has order of convergence q if q is the largest natural number such that there exists $C > 0$ and for some $\overline{h} > 0$ and all $h \in (0, \overline{h})$ it holds

$$\max_{k=0,\dots,N-1} \|Y_k - y(t_k)\| \leq Ch^q.$$

It holds

$$\sum_{j=0}^{k-1} h_j \|\tau(t_j, h_j)\| \leq (x_k - a) \max_{j=0,\dots,k-1} \|\tau(t_j, h_j)\| \leq (b-a) \max_{j=0,\dots,N-1} \|\tau(t_j, h_j)\|.$$

So assume that we have a consistency estimate of the form

$$\|\tau(t, h)\| \leq Kh^q,$$

for some $K > 0$ and $q \in \mathbb{N}$ and $h \in (0, \overline{h}]$. Then for $h = \max_j h_j$, it holds

$$\sum_{j=0}^{k-1} h_j \|\tau(t_j, h_j)\| \leq (b-a)Kh^q.$$

So under the setup of Theorem 1.24 it follows that

$$\|y(t_k) - Y_k\| \leq e^{L(t_k - a)} \sum_{j=0}^{k-1} h_j \|\tau(t_j, h_j)\| \leq e^{L(t_k-a)}(b-a)Kh^q.$$

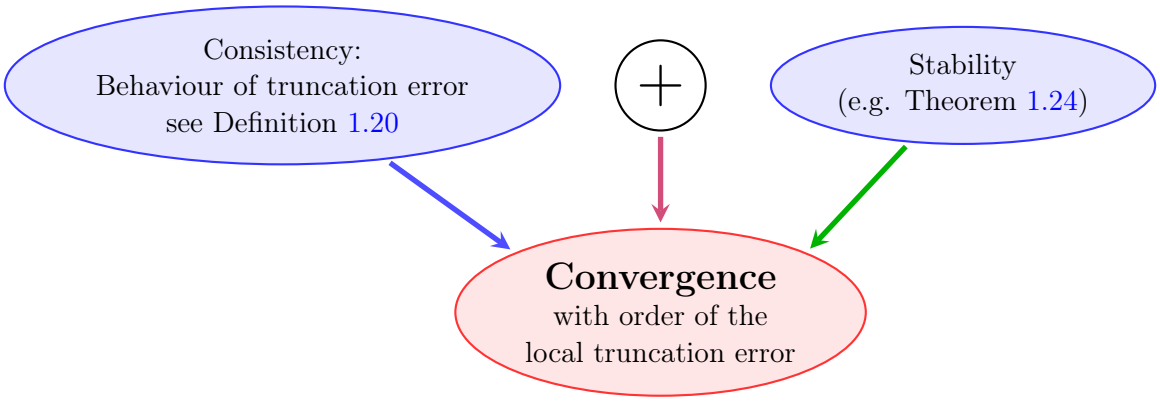
Taking the maximum on both sides it follows for $C = e^{L(b-a)}(b-a)K$

$$\max_k \|y(t_k) - Y_k\| \leq Ch^q.$$

For many ODEs, the right hand side f **is not defined** on the whole strip $[a, b] \times \mathbb{R}^n$ and in particular **not Lipschitz**.

Theorem 1.29 Consider our (IVP) with right hand side $f\colon [a, b] \times G \to \mathbb{R}^n$ and solution $y\colon [a, b] \to G$ for $G \subset \mathbb{R}^n$ open. Assume the increment function Φ is defined on

$$S_\delta := \{(t, Y, h) \mid 0 \leq h \leq \overline{h}, t \in [a, b-h], \|Y - y(t)\| \leq \delta\}$$



for some $\delta > 0$ and for all $(t, Y, u) \in S_\delta$ it holds

$$\|\Phi(t, y(t), h) - \Phi(t, Y, h)\| \leq L\|y(t) - Y\|.$$

Now additional assume that

$$e^{L(b-a)} \sum_{j=0}^{N-1} h_j \|\tau(t_j, h_j)\| \leq \delta.$$

Then, the iterates $Y_{k+1} = Y_k + h_k \Phi(t_k, Y_k, h_k)$ for $k = 0, \ldots, N-1$ with $Y_0 = y_0$ are welldefined and satisfy

$$\|y(t_k) - Y_k\| \leq e^{L(t_k - a)} \sum_{j=0}^{k-1} h_j \|\tau(t_j, h_j)\|.$$

Proof. Similar to Theorem 1.24 using induction. Exercise. □

At the end of this section we discuss the effect of perturbation errors.

Theorem 1.30 (Effect of perturbation errors) Let the assumptions of Theorem 1.29 hold. Moreover, let σ_k, for $k = 0, \ldots, N-1$ be a sequence of perturbations satisfying

$$e^{L(b-a)} \sum_{j=0}^{N-1} \|\sigma_j\| \leq \tfrac{\delta}{2}$$

and assume

$$e^{L(b-a)} \sum_{j=0}^{N-1} h_j \|\tau(t_j, h_j)\| \leq \tfrac{\delta}{2}.$$

Then, the iteration

$$Y_0 = y_0, \qquad Y_{k+1} = Y_k + h_k \Phi(t_k, Y_k, h_k) + \sigma_k$$

is well-defined and it holds

$$\|y(t_k) - Y_k\| \leq e^{L(t_k - a)} \left[\sum_{j=0}^{k-1} h_j \|\tau(t_j, h_j)\| + \sum_{j=0}^{k-1} \|\sigma_j\| \right]$$

Proof. As in Theorem 1.29. □

So assume

$$\|\sigma_j\| \sim \epsilon > 0 \qquad \Rightarrow \qquad \sum_{j=0}^{k-1} \|\sigma_j\| \sim k\epsilon.$$

The influence of the perturbations growth proportional to the number of grid points and at some points dominates the influence of the local truncation errors.

Example 1.31 Consider the IVP

$$y'(t) = -200ty(t)^2, \qquad -1 \le t \le 0, \qquad y(-1) = \frac{1}{101}.$$

with exact solution $y(t) = \frac{1}{100t^2+1}$. Compute approximations of $y(0) = 1$ with the Euler scheme and the improved Euler scheme with constant step size using float and double floating point accuracy. Exercise. △

Remark 1.32 One-step methods yield approximation Y_k to $y(t_k)$ at each grid point t_k for $k = 1, \dots, N$. In order to make the approximation a function on $[a, b]$ a typical approach is piecewise linear interpolation, i.e. connecting (t_k, Y_k) linearly with (t_{k+1}, Y_{k+1}). △

1.2.2 Explicit Runge-Kutta schemes

Idea: Construct an incremental function Φ, that is a linear combination of values of $f(t, Y)$ in different points to obtain methods of high order at the cost of evaluating f more often per iteration step.

For simplicity assume $y'(t) = f(t)$, hence

$$y(t_{k+1}) = y(t_k) + \int_{t_k}^{t_{k+1}} f(t)\,\mathrm{d}t.$$

Then, we can apply an arbitrary quadrature rule of level $s \in \mathbb{N}$ of the form

$$\int_{t_k}^{t_{k+1}} f(t)\,\mathrm{d}t \approx h_k \sum_{j=1}^{s} b_j f(t_k + c_j h_k)$$

for suitable $b_j, c_j \in \mathbb{R}$. The local error at starting value $Y_k = y(t_k)$ is given as

$$y(t_{k+1}) - Y_{k+1} = y(t_{k+1}) - y(t_k) - h_k \sum_{j=1}^{s} b_j f(t_k + c_j h_k) = \int_{t_k}^{t_{k+1}} f(t)\,\mathrm{d}t - h_k \sum_{j=1}^{s} b_j f(t_k + c_j h_k),$$

that is exactly the quadrature error. Basic concept for a Runge-Kutta scheme for our problem (IVP) with right hand side $f(t, y)$ is a similar ansatz

$$\int_{t_i}^{t_{k+1}} f(t, y(t))\,\mathrm{d}t \approx h_k \sum_{j=1}^{s} b_j f(t_k + c_j h_k, \boldsymbol{\eta}_j), \tag{1.13}$$

where now additional approximations $\boldsymbol{\eta}_j$ for intermediate values $y(t_k + c_j h_k)$ have to be considered for $j = 1, \dots, s$.

Example 1.33 (Improved Euler method) Making the ansatz with midpoint rule we first obtain

$$Y_{k+1} = Y_k + h_k f(t_k + \tfrac{h_k}{2}, \boldsymbol{\eta}_1)$$

for an approximation $\boldsymbol{\eta}_1 \approx y(t_k + \frac{h_k}{2})$ to be choosen. Using the explicit Euler scheme with step size $\frac{h_k}{2}$ yields

$$\boldsymbol{\eta}_1 = Y_k + \tfrac{h_k}{2} f(t_k, Y_k),$$

and in total

$$Y_{k+1} = Y_k + h_k f(t_k + \tfrac{h}{2}, Y_k + \tfrac{h_k}{2} f(t_k, Y_k))$$

which is the improved Euler scheme from (1.16) known as *Runge scheme* (1895, aka *explicit midpoint rule*). Let f be sufficiently smooth, then for $h_k \equiv h$ by Taylor expansion:

$$\begin{aligned} Y_{k+1} &= Y_k + hf(t_k + \tfrac{h}{2}, Y_k + \tfrac{h}{2} f(t_k, Y_k)) \\ &= Y_k + h\left[f(t_k, Y_k) + \tfrac{h}{2}(\partial_t f)(t_k, Y_k) + \tfrac{h}{2}(\partial_y f)(t_k, Y_k) + \mathcal{O}(h^2).\right] \end{aligned}$$

Differentiating the differential equation leads

$$y''(t) = \frac{\partial}{\partial t} f(t, y(t)) = (\partial_t f)(t, y(t)) + (\partial_y f)(t, y(t)) f(t, y(t)).$$

Hence, for the local error starting at $Y_k = y(t_k)$ we conclude by Taylor expansion that

$$\begin{aligned} y(t_{k+1}) &= y(t_k) + hy'(t_k) + \tfrac{h}{2}^2 y''(t_k) + \mathcal{O}(h^3) \\ &= Y_k + hf(t_k, Y_k) + \tfrac{h}{2}^2 \left[(\partial_t f)(t_k, Y_k) + (\partial_y f)(t_k, Y_k)\right] + \mathcal{O}(h^3). \end{aligned}$$

Consequently,

$$\|Y_{k+1} - y(t_{k+1})\| \in \mathcal{O}(h^3).$$

The method has **consistency order** $q = 2$**.** △

Definition 1.34 (explicit Runge-Kutta methods) Let $s \in \mathbb{N}$. A s-stage explicit Runge-Kutta method has the form

$$Y_0 = y(t_0), \qquad Y_{k+1} = Y_k + h\Phi(t, Y_k, h), \qquad k = 0, 1, \ldots$$

with

$$\Phi(t, Y, h) = \sum_{i=1}^{s} b_i K_i(t, Y, h),$$

based on increments

$$\begin{aligned} K_1(t, Y, h) &= f(t, Y) \\ K_i(t, Y, h) &= f\left(t + c_i h, Y + h \sum_{j=1}^{i-1} a_{ij} K_j(t, Y, h)\right), \quad i = 2, \ldots, s \end{aligned}$$

and constants $c_2, \ldots, c_s; b_1, \ldots b_s; a_{ij} \in \mathbb{R}$ for $i = 1, \ldots, s$ and $j = 1, \ldots, i-1$.

In the spirit of (1.13) let $\boldsymbol{\eta}_{k+1}^{(i)} = Y_k + h\sum_{j=1}^{i-1} a_{ij} f(t_k + c_j h, \boldsymbol{\eta}_{k+1}^{(j)})$, then we can equivalently represent the increment function as

$$\Phi(t, Y, h) = \sum_{i=1}^{s} b_i f(t_k + c_i h, \boldsymbol{\eta}_{k+1}^{(i)}).$$

This representation is useful, when extending explicit Runge-Kutta methods to the implicit case, see also Remark 1.35.

We can summarize an explicit Runge Kutta scheme in a so-called *Butcher tableau*:

$$\begin{array}{c|cccc} 0 & 0 & \cdots & \cdots & 0 \\ c_2 & a_{21} & \ddots & & \vdots \\ c_3 & a_{31} & a_{32} & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & 0 \\ c_s & a_{s1} & a_{s2} & \cdots & a_{s,s-1} \\ \hline & b_1 & b_2 & \cdots & b_s \end{array} = \begin{array}{c|c} \boldsymbol{c} & \boldsymbol{A} \\ \hline & \boldsymbol{b}^\top \end{array}$$

with nodes $\boldsymbol{c}$, matrix $\boldsymbol{A}$ of the method and weights $\boldsymbol{b}$.

Remark 1.35 The matrix $\boldsymbol{A}$ is strictly lower triangular for explicit Runge-Kutta methods. For so called *implicit* Runge Kutta methods, that we introduce later, $\boldsymbol{A}$ might be lower triangular (non-zero diagonal) or even dense. Then, $\boldsymbol{\eta}_{k+1}^{(i)} = Y_k + h\sum_{j=1}^{s} a_{ij} f(t_k + c_j h, \boldsymbol{\eta}_{k+1}^{(j)})$. and in general $\boldsymbol{\eta}_j$ is given implicitly. △

Example 1.36 The following Butcher tableaus

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}, \quad \begin{array}{c|cc} 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \hline & 0 & 1 \end{array}, \quad \begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}, \quad \begin{array}{c|cccc} 0 & 0 & \cdots & \cdots & 0 \\ \frac{1}{2} & \frac{1}{2} & \ddots & & \\ \frac{1}{2} & 0 & \frac{1}{2} & \ddots & \vdots \\ 1 & 0 & 0 & 1 & 0 \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array} \tag{1.14}$$

represent the *explicit Euler method*, the *Runge method* (aka *improved Euler method*), the *Heun method* (based on trapezoid + explicit Euler for $\boldsymbol{\eta}_1$) and the *classical Runge-Kutta method* (based on Simpson rule), respectively. △

Lemma 1.37 Assume a one-step method of the form

$$Y_{k+1} = Y_k + h\sum_{i=1}^{s} b_i f(t_k + c_i h, \boldsymbol{\eta}_j)$$

has consistency order $q \in \mathbb{N}$. Then, the quadrature rule

$$\mathcal{Q}[g] := \sum_{i=1}^{s} b_i g(c_i) \approx \int_0^1 g(t)\,\mathrm{d}t$$

is exact for polynomials up to degree $q-1$.

Proof. W.l.o.g. let $n = 1$. Let $p = 0, \dots, q-1$ and $y'(t) = (t-a)^p$, $y(a) = 0$, with unique solution $y(t) = \frac{(t-a)^{p+1}}{p+1}$ on $[a, b]$. By consistency assumption

$$|y(t_1) - Y_1| = |y(a+h) - Y_1| = \left| \tfrac{1}{p+1} h^{p+1} - h \sum_{j=1}^{s} b_j (c_j h)^p \right| \in \mathcal{O}(h^{p+1}), \quad \text{for } h \to 0.$$

Hence,

$$\left| \frac{1}{p+1} - \sum_{j=1}^{s} b_j c_j^p \right| \in \mathcal{O}(h^{q-p}), \quad \text{for } h \to 0.$$

Consequently, in the limit $h \to 0$

$$\sum_{j=1}^{s} b_j c_j^p = \frac{1}{p+1} = \int_0^1 t^p \, \mathrm{d}t,$$

which means exactness for all monomials t^p for $p = 0, \dots, q-1$. □

Remark 1.38 As a consequence, a Runge-Kutta method of stage s can have **maximum consistency order** $2s$ due to the limit of exactness order $2s-1$ for quadrature formulas with s abscissas. However for Runge-Kutta methods also $\boldsymbol{\eta}_j$ is designed by quadrature and this will limit the consistency order further for fixed s (in the explicit case). △

Our next goal is the discussion of the degrees of freedoms of a Runge-Kutta scheme, namely $\boldsymbol{c}, \boldsymbol{b}$ and $\boldsymbol{A}$.

For this we will consider test IVPs to discuss preliminary constraints on the degrees of freedom. Let $y'(t) = 1$, i.e. $f(t, y) \equiv 1$. Then, $y(t) = y_0 + (t-a)$ and $y(t_k + h) = y(t_k) + h$. Then, stage computation leads (using $c_1 = 0$):

$$K_i(t_k, Y_k, h) = f(t_k + c_i h, Y_k + h \sum_{j=1}^{i-1} a_{ij} K_j(t_k, Y_k, h)) \equiv 1.$$

Hence, for the iteration

$$Y_{k+1} = Y_k + h \sum_{i=1}^{s} b_i K_i(t_k, Y_k, h) = Y_k + h \sum_{i=1}^{s} b_i = y_0 + (t_k - a) \sum_{i=1}^{s} b_i.$$

Comparing with $y(t) = y_0 + (t-a)$ consistency is only possible for

$$\boxed{\sum_{i=1}^{s} b_i = 1.} \tag{1.15}$$

Now recall for each $i = 1, \dots, d$ the ansatz of the Runge-Kutta scheme (choice of $\boldsymbol{\eta}_i$) generalized to the implicit regime as in Remark 1.35, s.t.

$$\begin{aligned} y(t_k + c_i h) &= y(t_k) + \int_{t_k}^{t_k + c_i h} f(t, y(t)) \, \mathrm{d}t \\ &\approx \boldsymbol{\eta}_i = Y_k + h \sum_{j=1}^{s} a_{ij} f(t_k + c_j h, \boldsymbol{\eta}_j). \end{aligned}$$

With the choice of $f \equiv 1$ yields the necessary condition

$$\boxed{\sum_{j=1}^{s} a_{ij} = \frac{1}{h} \int_{t_k}^{t_k + c_i h} 1 \, \mathrm{d}t = c_i.} \tag{1.16}$$

Both necessary conditions have in fact a bigger impact.

Theorem 1.39 Let f be continuous on $[a, b] \times \mathbb{R}^n$. Then, an explicit Runge-Kutta scheme is consistent if and only if

$$\sum_{i=1}^{s} b_i = 1. \tag{1.17}$$

Proof. By continuity of f and the definition of K_i it holds

$$\lim_{h \to 0} K_i(t, Y, h) = f(t, Y), \qquad \forall (t, Y) \in [a, b] \times \mathbb{R}^n.$$

Then, for $t = t_k, Y = y(t_k)$ it follows

$$\begin{aligned} \lim_{h \to 0} |f(t_k, y(t_k) - \Phi(t_k, y(t_k), h)| &= \lim_{h \to 0} \left| f(t_k, y(t_k)) - \sum_{i=1}^{s} b_i K_i(t_k, y(t_k), h) \right| \\ &= \left| f(t_k, y(t_k)) \left(1 - \sum_{i=1}^{s} b_i \right) \right| = 0 \end{aligned}$$

if and only if $\sum_{i=1}^{s} b_i = 1$. □

Theorem 1.40 Let (IVP) have a solution $y \in \mathcal{C}^2([a, b])$ and let $f \in \mathcal{C}_{\mathrm{PL}}$. Let $Y_k = y(t_k)$ and

$$c_i = \sum_{j=1}^{i-1} a_{ij}, \qquad i \geq 2.$$

Then, $K_i(t_k, Y_k, h)$ is an first order approximation of $y'(t_k + c_i h)$, i.e.

$$\| y'(t_k + c_i h) - K_i(t_k, Y_k, h) \| \in \mathcal{O}(h^2).$$

Proof. By induction: Let $i = 2$, by Lipschitz continuity and Taylor it follows

$$\begin{aligned} \| y'(t_k + c_2 h) - K_2(t_k, Y_k, h) \| &= \| f(t_k + c_2 h, y(t_k + c_2 h)) - f(t_k + c_2 h, y(t_k) + h a_{21} f(t_k, y(t_k))) \| \\ &\leq L \| y(t_k + c_2 h) - y(t_k) - h a_{21} f(t_k, y(t_k)) \| \\ &= L \| y(t_k) + c_2 h y'(t_k) + \mathcal{O}(h^2) - y(t_k) - h a_{21} y'(t_k) \| \\ &= L \| (c_2 - a_{21}) h y'(t_k) + \mathcal{O}(h^2) \| \end{aligned}$$

Hence for $c_2 = a_{21}$, the approximation error is of order $\mathcal{O}(h^2)$. For $i > 2$, let the statement holds for all indices $2, \dots, i-1$. Then, similarily

$$\begin{aligned}
\|y'(t_k + c_i h) - K_i(t_k, Y_k, h)\| &= \left\| f(t_k + c_i h, y(t_k + c_i h)) - f(t_k + c_i h, y(t_k) + h \sum_{j=1}^{i-1} a_{ij} K_j(t_k, Y_k, h)) \right\| \\
&\le L \left\| y(t_k + c_i h) - y(t_k) - h \sum_{j=1}^{i-1} a_{ij} K_j(t_k, Y_k, h)) \right\| \\
&= L \left\| c_i h y'(t_k) + \mathcal{O}(h^2) - h \sum_{j=1}^{i-1} a_{ij} (y'(t_k + c_j h) + \mathcal{O}(h^2)) \right\| \\
&= L \left\| c_i h y'(t_k) + \mathcal{O}(h^2) - h \sum_{j=1}^{i-1} a_{ij} (y'(t_k + c_j h) + \mathcal{O}(h^2)) \right\| \\
&= L \left\| c_i h y'(t_k) + \mathcal{O}(h^2) - h \sum_{j=1}^{i-1} a_{ij} (y'(t_k) + \mathcal{O}(h)) \right\| \\
&= L \left\| \left(c_i - \sum_{j=1}^{i-1} a_{ij} \right) h y'(t_k) + \mathcal{O}(h^2) \right\|,
\end{aligned}$$

where we used Taylor expansion (using the integral remainder) of $y'(t_k + c_j h) = y'(t_k) + \mathcal{O}(h)$. The approximation error is of order $\mathcal{O}(h^2)$ if $c_i = \sum_{j=1}^{i-1} a_{ij}$. □

Remark 1.41 For $f \in C^1([a,b] \times \mathbb{R}^n)$ it can directly be shown that the explicit Runge-Kutta schema is consistent of order 1 if the condition $\sum_{i=1}^{s} b_i = 1$ holds. Exercise △

A generalized result to Remark 1.41 is giving without proof.

Theorem 1.42 (See [2], Theorem 4.18) An explicit Runge-Kutta method is consistent for any right hand side $f \in \mathcal{C}^q([a,b] \times \mathbb{R}^n; \mathbb{R}^n)$ with order

1. $q = 1$: if
$$\sum_{i=1}^{s} b_i = 1$$
2. $q = 2$: if additional
$$\sum_{i=1}^{s} b_i c_i = 1/2$$
3. $q = 3$: if additional
$$\sum_{i=1}^{s} b_i c_i^2 = 1/3, \qquad \sum_{i,j=1}^{s} b_i a_{ij} c_j = 1/6$$

4. $q = 4$: if additional

$$\sum_{i=1}^{s} b_i c_i^3 = 1/4, \qquad \sum_{i,j=1}^{s} b_i c_i a_{ij} c_j = 1/8, \qquad \sum_{i,j=1}^{s} b_i a_{ij} c_j^2 = 1/12, \qquad \sum_{i,j,k=1}^{s} b_i a_{ij} a_{jk} c_k = 1/24$$

Note that the sufficient conditions for $p > 1$ become non-linear.

Remark 1.43 For a requested consistency order p a number N_p of conditions has to be fulfilled. Up to $p = 8$ the original work of J.C.Butcher, see [1], contains those conditions. It turns out that the number N_p can be calculated by combinatorial arguments, but their number grows exponentially in p as illustrated in the following table:

Table 1.2: Number of order conditions for Runge-Kutta-methods

p	1	2	3	4	5	10	20
N_p	1	2	4	8	17	1205	20247374.

△

Corollary 1.44 The explicit Euler method has consistency order $q = 1$, the Runge and Heun method have consistency order $q = 2$ and the classical Runge Kutta scheme has consistency order $q = 4$ for any right hand side $f \in \mathcal{C}^q([a, b] \times \mathbb{R}^n; \mathbb{R}^n)$.

As a next result we will refine Remark 1.38 and show that explicit stage s Runge-Kutta methods are of consistency order at most s

For this we need to derive the **stability function** (including the implicit case).

Consider the scalar (in particular $n = 1$) linear test equation

$$y'(t) = \lambda y(t), \qquad \lambda \in \mathbb{C},$$

with initial value $y(0) = y_0$. Let an s-stage Runge–Kutta method with Butcher tableau

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

be applied to this problem.

The internal stages of the Runge–Kutta scheme are defined as

$$K_i = f\left(t_n + c_i h,\, Y_k + h \sum_{j=1}^{s} a_{ij} K_j\right), \qquad i = 1, \ldots, s.$$

For the test equation $f(t, y) = \lambda y$, this becomes

$$K_i = \lambda\left(Y_k + h \sum_{j=1}^{s} a_{ij} K_j\right), \qquad i = 1, \ldots, s.$$

In vector form, introducing $\boldsymbol{K} = (K_1, \dots, K_s)^\top \in \mathbb{C}^s$ and $\mathbf{1} = (1, \dots, 1)^\top \in \mathbb{R}^s$, we can write

$$\boldsymbol{K} = \lambda \left(Y_k \mathbf{1} + h \boldsymbol{A} \boldsymbol{K} \right).$$

Rearranging terms yields (note that $Y_k \in \mathbb{C}$)

$$(\boldsymbol{I} - h\lambda \boldsymbol{A}) \boldsymbol{K} = \lambda Y_k \mathbf{1}.$$

Assuming $\boldsymbol{I} - h\lambda \boldsymbol{A}$ is invertible, we obtain

$$\boldsymbol{K} = \lambda (\boldsymbol{I} - h\lambda \boldsymbol{A})^{-1} Y_k \mathbf{1}.$$

Hence, the Runge–Kutta update reads

$$Y_{k+1} = Y_k + h\, \boldsymbol{b}^\top \boldsymbol{K}.$$

Substituting the expression for $\boldsymbol{K}$ gives

$$Y_{k+1} = Y_k + h\, \boldsymbol{b}^\top \big(\lambda (\boldsymbol{I} - h\lambda \boldsymbol{A})^{-1} Y_k \mathbf{1} \big) = \Big(1 + h\lambda\, \boldsymbol{b}^\top (\boldsymbol{I} - h\lambda \boldsymbol{A})^{-1} \mathbf{1} \Big) Y_k.$$

Defining $z = h\lambda \in \mathbb{C}$, we can write

$$Y_{k+1} = R(z) Y_k, \qquad R(z) = 1 + z \boldsymbol{b}^\top (\boldsymbol{I} - z\boldsymbol{A})^{-1} \mathbf{1}.$$

Definition 1.45 (Stability function) Let $\overline{\mathbb{C}} \coloneqq \mathbb{C} \cup \{\infty\}$. Then, the associated *stability function* $R \colon \overline{\mathbb{C}} \to \overline{\mathbb{C}}$ for a (possible implicite) Runge Kutta scheme with coefficients $\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}$ is given as

$$R(z) = 1 + z \boldsymbol{b}^\top (\boldsymbol{I} - z\boldsymbol{A})^{-1} \mathbf{1}.$$

Note that R is welldefined for all $z = \lambda h$, which are not eigenvalue of $\boldsymbol{A}$.

Theorem 1.46 (Consistency of maximum order s) A explicit Runge-Kutta scheme of stage $s \in \mathbb{N}$ has consistency order of maximum s.

Proof. In the case of explicit Runge-Kutta the matrix $\boldsymbol{A} \in \mathbb{R}^{s,s}$ is strictly lower triangular and hence nilpotent with $\boldsymbol{A}^s = \mathbf{0}$. It holds that

$$(\boldsymbol{I} - z\boldsymbol{A})(\boldsymbol{I} + z\boldsymbol{A} + \cdots + z^{s-1}\boldsymbol{A}^{s-1}) = \boldsymbol{I} - z^s \boldsymbol{A}^s = \boldsymbol{I}$$

and thus $(\boldsymbol{I} - z\boldsymbol{A})^{-1}$ is invertible for each $z \in \mathbb{C}$ given as a matrix polynomial of degree $s-1$. Consequently, R is a polynomial of degree s. Assume for some $q \in \mathbb{N}$ that the explicit Runge-Kutta schema is of consistency order q. Then, for the test equation

$$y'(t) = y(t), \qquad y(a) = 1,$$

and step size $h > 0$ the local error satisfies

$$|Y_1 - y(h)| = |Y_1 - e^h| = |R(h) - e^h| \in \mathcal{O}(h^{q+1}), \qquad \text{as } h \to 0.$$

Since $R(z)$ is a polynomial in z and e^z can be written as a local convergent Taylor expansion at $z = 0$, their Taylor polynomials must match up to order q. It follows $q \le s$ and if $q = s$ then $R(z) = \sum_{k=0}^{q} \frac{z^k}{k!}$. □

So we have seen with Corollary 1.44, that the explicit Euler-, Runge-, Heun- and the classical Runge-Kutta method, all archive maximum possible consistency order. The question arises, if the same holds true for higher number of stages $s \geq 5$.

Remark 1.47 (Order limitations, Butcher (1963,1965,1985)) For $s \geq 5$, the order $q = s$ can no longer be achieved by explicit Runge–Kutta methods. A more detailed analysis yields the following values $s_{\min}(q)$ for the minimal number of stages required by an explicit method to attain order q:

q	1	2	3	4	5	6	7	8
$s_{\min}(q)$	1	2	3	4	6	7	9	11

As we shall see later, for implicit Runge–Kutta methods one always has $q = 2s$. △

Example 1.48 (stage 2 Runge-Kutta schemes of consistency order 2) For simplicity let $n = 1$. Runge-Kutta schemes of stage $s = 2$ have the incremental function

$$\Phi(t, Y, h) = b_1 f(x, Y) + b_2 f(t + c_2 h, Y + h a_{21} f(t, Y)) \in \mathbb{R}.$$

Taylor expansion around $h = 0$ and requiring consistency order $q = s = 2$, yields the condition for $q = 2$ from Theorem 1.42. Cp. Tutorium. △

Theorem 1.49 (Stability of explicit Runge-Kutta methods) Let $f \in \mathcal{C}_{\mathrm{PL}}$ with Lipschitz constant L. Then, the incremental function Φ satisfies

$$\|\Phi(t, y_1, h) - \Phi(t, y_2, h)\| \leq L(h)\|y_1 - y_2\|$$

with $L(h) = \sum_{i=1}^{s} |b_i| L_i(h)$ and

$$L_1(h) = L, \qquad L_i(h) = \left(1 + h \sum_{j=1}^{i-1} |a_{ij}| L_j(h)\right) L, \quad i = 2, \ldots, s.$$

For $i = 1, \ldots, s$ it holds

$$\|K_i(t, y_1, h) - K_i(t, y_2, h)\| \leq L_i(h)\|y_1 - y_2\|. \tag{1.18}$$

Proof. Show (1.18) via induction and conclude the claim. □

It follows

$$\lim_{h \to 0^+} L_i(h) = L \qquad \text{and} \qquad \lim_{h \to 0^+} L(h) = \left(\sum_{i=1}^{s} |b_i|\right) L.$$

So if $b_i \geq 0$, by consistency $\sum_{i=1}^{s} |b_i| = \sum_{i=1}^{s} b_i = 1$.

Remark 1.50 (Stability in light of dissipative systems) If additional to the assumption of Theorem 1.49 it holds that

$$\langle y_1 - y_2, f(t, y_1) - f(t, y_2) \rangle \leq \theta \|y_1 - y_2\|^2$$

and $b_i \geq 0$ for all $i = 1, \ldots, s$. Then,

$$\langle y_1 - y_2, \Phi(t, y_1, h) - \Phi(t, y_2, h) \rangle \leq \theta(h) \|y_1 - y_2\|^2$$

with $\lim_{h \to 0^+} \theta(h) = \theta \sum_{i=1}^{s} b_i = \theta$. △

Theorem 1.51 (Convergence of explicit Runge-Kutta methods) Explicit Runge-Kutta methods of stage s with consistency order $q \leq s$ are convergent with order q. The constant in the convergence estimate can be improved for dissipative systems.

Proof. Using Theorem 1.49 Runge-Kutta methods are stable by application of Theorem 1.24 and with improved constants due to Remark 1.50 and application of Theorem 1.27. Then, **stability** + **Consistency** yields **convergence**. □

1.2.3 Step length control

When solving an ordinary differential equation numerically, the choice of the step size plays a decisive role in the quality and efficiency of the computed solution. If the step size is too large, the numerical method may lose accuracy or even become unstable, producing results that no longer resemble the true behavior of the system. If the step size is too small, the computation becomes unnecessarily expensive, and rounding errors may begin to accumulate.

In practice, the solution of an ODE often exhibits regions of very different behavior: it may vary slowly for a long time and then change rapidly within a short interval. Using a fixed step size throughout such a computation is inefficient—either one wastes computational effort in the smooth parts or loses accuracy in the rapidly changing ones.

Step size control addresses this problem by adapting the step length dynamically to the local behavior of the solution. The idea is simple yet powerful: estimate the local error made in each step, and adjust the next step size so that the error remains within a prescribed tolerance. In this way, the numerical method automatically takes small steps where the solution demands high resolution and larger steps where the dynamics are smooth.

This adaptive approach not only improves computational efficiency but also enhances robustness, allowing numerical solvers to handle stiff or highly variable problems with minimal user intervention. Step size control is thus a central component of modern ODE solvers and an essential tool in scientific computing.

Theorem 1.52 Consider different incremental functions $\widetilde{\Phi}$ and Φ an associated one-step scheme

$$Y_{k+1} = Y_k + h_k \Phi(t_k, Y_k, h_k), \qquad \widetilde{Y}_{k+1} = Y_k + h_k \widetilde{Y}_{k+1}(t_k, Y_k, h_k)$$

for the solution of (IVP) with step sizes $h_k > 0$ and nodes $t_{k+1} = t_k + h_k$. Assume $\widetilde{\Phi}$ is L-Lipschitz w.r.t. to the second input uniformly in t, h, i.e.

$$\|\widetilde{\Phi}(t, y_1, h) - \widetilde{\Phi}(t, y_2, h) \leq L \|y_1 - y_2\|, \qquad \forall t \in [a, b], y_1, y_2 \in \mathbb{R}^n, h > 0.$$

Then

$$\|y(t_k) - Y_k\| \leq e^{L(t_k - t_0)} \sum_{j=0}^{k-1} \left(\|Y_{j+1} - \widetilde{Y}_{j+1}\| + h_j \|\widetilde{\tau}(t_j, h_j)\| \right), \qquad k = 1, 2, \ldots,$$

where $\widetilde{\tau}$ denotes the local truncation error with respect to $\widetilde{\Phi}$ as in Definition 1.18.

Proof. Analogue to the proof of Theorem 1.24 based on $1 + t \leq e^t$ and using

$$\begin{aligned}\|y(t_{k+1}) - Y_{k+1}\| &= \|y(t_k) + h_k\widetilde{\Phi}(t_k, y(t_k), h_k) + h_k\widetilde{\tau}(t_k, h_k) \underbrace{-Y_k - h_k\widetilde{\Phi}(t_k, Y_k, h_k) + \widetilde{Y}_{k+1}}_{=0} - Y_{k+1}\| \\ &\leq \|y(t_k) - Y_k + h_k(\widetilde{\Phi}(t_k, y(t_k), h_k) - \widetilde{\Phi}(t_k, Y_k, h_k))\| + \|\widetilde{Y}_{k+1} - Y_{k+1}\| + h_k\|\widetilde{\tau}(t_k, h_k)\| \\ &\leq e^{h_k L}\|y(t_k) - Y_k\| + \|\widetilde{Y}_{k+1} - Y_{k+1}\| + h_k\|\widetilde{\tau}(t_k, h_k)\|\end{aligned}$$

Unrolling the recurrence yields

$$\begin{aligned}\|y(t_{k+1}) - Y_{k+1}\| &\leq \sum_{j=0}^{k} \left((\|Y_{j+1} - \widetilde{Y}_{j+1}\| + h_j\|\widetilde{\tau}(t_j, h_j)\|) \prod_{i=j+1}^{k} e^{h_i L} \right) \\ &= \sum_{j=0}^{k} \left((\|Y_{j+1} - \widetilde{Y}_{j+1}\| + h_j\|\widetilde{\tau}(t_j, h_j)\|) e^{L(t_{k+1} - t_{j+1})} \right).\end{aligned}$$

Then, the claim follows by noting that $t_{k+1} - t_{j+1} \leq t_{k+1} - t_0$. □

As a consequence of Theorem 1.52 if Φ is of consistency order q and $\widetilde{\Phi}$ is of consistency order $\geq q+1$, then the term $\sum_{j=0}^{k-1} h_j\|\widetilde{\tau}(t_k, h_k)\|$ is of higher order and dominated by $\sum_{j=0}^{k-1} \|Y_{j+1} - \widetilde{Y}_{j+1}\|$. The goal is now to choose step sizes h_j making the latter controlled, such that for some prescribed $\epsilon > 0$

$$\|Y_{j+1} - \widetilde{Y}_{j+1}\| \approx h_j\epsilon \qquad \Rightarrow \qquad \sum_{j=0}^{k-1} \|Y_{j+1} - \widetilde{Y}_{j+1}\| \approx (t_k - t_0)\epsilon.$$

One approach to choose h_k is to analyze the behavior of

$$\|\{Y_k + h_k\Phi(t_k, Y_k, h)\} - \{Y_k + h\widetilde{\Phi}(t_k, Y_k, h)\}\|, \qquad \text{as } h \to 0.$$

So let z be solution of

$$z'(t) = f(t, z(t)), \qquad z(t_k) = Y_k.$$

Then, by construction for some function $e(t)$ we have

$$\frac{z(t+h) - z(t)}{h} - \Phi(t, z(t), h) = e(t)h^q + \mathcal{O}(h^{q+1}) \qquad \text{and} \qquad \frac{z(t+h) - z(t)}{h} - \widetilde{\Phi}(t, z(t), h) = \mathcal{O}(h^{q+1}).$$

Hence, for a single step of approximation of $z(t+h)$ given as $Y = Y_k + h\Phi(t_k, Y_k, h)$ and $\widetilde{Y} = Y_k + h\widetilde{\Phi}(t_k, Y_k, h)$ it holds

$$\begin{aligned}Y - \widetilde{Y} &= \{z(t_k) + h\Phi(t_k, Y_k, h)\} - \{z(t_k) + h\widetilde{\Phi}(t_k, Y_k, h)\} \\ &= h\left\{ \frac{z(t_k + h) - z(t_k)}{h} - \widetilde{\Phi}(t_k, Y_k, h) \right\} - h\left\{ \frac{z(t_k + h) - z(t_k)}{h} - \Phi(t_k, Y_k, h) \right\} \\ &= e(t_k)h^{q+1} + \mathcal{O}(h^{q+2}).\end{aligned}$$

As a consequence up to higher order terms for

$$Y_{k+1} = Y_k + h_k\Phi(t_k, Y_k, h_k) \qquad \text{and} \qquad \widetilde{Y}_{k+1} = Y_k + h_k\widetilde{\Phi}(t_k, Y_k, h_k)$$

with the representation $h_k = sh$ for some unknown $s > 0$, it holds

$$\|Y_{k+1} - \widetilde{Y}_{k+1}\| \approx Ch_k^{q+1} = C(sh)^{q+1} = s^{q+1}(Ch^{q+1}) = s^{q+1}\|Y - \widetilde{Y}\|.$$

Hence, the requirement $\|Y_{k+1} - \widetilde{Y}_{k+1}\| \approx h_k\epsilon = sh\epsilon$ leads for given h to the scaling factor

$$s = \left(\frac{h\epsilon}{\|Y - \widetilde{Y}\|}\right)^{1/q}. \tag{1.19}$$

and hence for the recommended step size $h_k = sh \approx h\left(\frac{h\epsilon}{\|Y-\widetilde{Y}\|}\right)^{1/q}$.

Algorithmus 1.2.1 Adaptive Step Size Control

Require: Current step size estimate h, current value Y_k at time t_k, increment functions Φ of order q and $\widetilde{\Phi}$ of order $q+1$, tolerance $\epsilon > 0$.

Ensure: Next value Y_{k+1} and next step size estimate h_{next}.

1: Compute trial steps:

$$Y = Y_k + h\,\Phi(t_k, Y_k, h), \qquad \widetilde{Y} = Y_k + h\,\widetilde{\Phi}(t_k, Y_k, h).$$

2: Estimate the scaling factor:

$$s = \left(\frac{h\epsilon}{\|Y - \widetilde{Y}\|}\right)^{1/q}.$$

3: **if** $s \geq 1$ **then**

4: Accept the step: $Y_{k+1} \leftarrow \widetilde{Y}$ (typically).

5: Advance time: $t_{k+1} \leftarrow t_k + h$.

6: Update step size estimate: {Possibly double the step size}

$$h_{\text{next}} \leftarrow \min\{2, s\}\, h.$$

7: **else**

8: Reject the step.

9: Reduce step size: {possibly half the step size}

$$h \leftarrow \max\{1/2, s\}\, h.$$

10: Repeat from line 1 with new h.

11: **end if**

Richardson Extrapolation The approach of halfing the step size can be used to construct schemes of higher order based on extrapolation.

Let us assume we have a reference scheme of order q with incremental function $\widehat{\Phi}$. Let us compute Y_{k+1} based on two half steps using this incremental function, s.t.

$$\begin{aligned} Y_{k+\frac{1}{2}} &= Y_k + \tfrac{1}{2}\widehat{\Phi}(t_k, Y_k, \tfrac{h_k}{2}) \\ Y_{k+1} &= Y_{k+\frac{1}{2}} + \tfrac{h_k}{2}\widehat{\Phi}(t_k + \tfrac{h_k}{2}, Y_{k+\frac{1}{2}}, \tfrac{h_k}{2}). \end{aligned}$$

This corresponds to a one step method

$$Y_{k+1} = Y_k + h_k \Phi(t_k, Y_k, h_k)$$

with incremental function

$$\Phi(t, y, h) = \tfrac{1}{2}\widehat{\Phi}(t, y, \tfrac{h}{2}) + \tfrac{1}{2}\widehat{\Phi}(t + \tfrac{h}{2}, y + \tfrac{h}{2}\widehat{\Phi}(t, y, \tfrac{h}{2}), \tfrac{h}{2}).$$

Now, let

$$\widehat{Y}_{k+1} = Y_k + h_k \widehat{\Phi}(t_k, Y_k, h_k)$$

and by extrapolation define

$$\widetilde{Y}_{k+1} = \frac{1}{2^q - 1}\left(2^q Y_{k+1} - \widehat{Y}_{k+1}\right).$$

This step can be interpreted as a one-step method with incremental function

$$\widetilde{\Phi}(t, y, h) = \frac{1}{2^q - 1}\left(2^q \Phi(t, y, h) - \widehat{\Phi}(t, y, h)\right).$$

Theorem 1.53 Let y be solution of the ODE $y'(t) = f(t, y(t))$. Assume for some $\delta > 0$ that the incremental function $\widehat{\Phi}$ is well-defined on the tube

$$\{(x, u, h) \mid 0 \le h \le \overline{h}, a \le t \le b, \|u - y(t)\| < \delta\}$$

where it satisfies the uniform Lipschitz condition

$$\|\widehat{\Phi}(t, y(t), h) - \widehat{\Phi}(t, u, h)\| \le L\|y(t) - u\|.$$

. Assume that the associated local truncation error $\widehat{\tau}$ is of the form

$$\widehat{\tau}(t, h) = e(t)h^q + r_0(t, h)h^{q+1},$$

for some continuously differentiable function $e\colon [a, b] \to \mathbb{R}$ and a bounded function r_0. Moreover let Φ and $\widetilde{\Phi}$ be defined on the same tube as $\widehat{\Phi}$. Then, the associated local truncation errors τ and $\widetilde{\tau}$ fulfill

$$\begin{aligned}\tau(t, h) &= e(t)\left(\tfrac{h}{2}\right)^q + r(t, h)h^{q+1},\\ \widetilde{\tau}(t, h) &= R(t, h)h^{q+1}.\end{aligned}$$

with bounded functions r and R.

Proof. The proof is optional. See below in gray. □

Proof. It holds

$$\begin{aligned}\tau(t, h) &= \frac{y(t + h) - y(t)}{h} - \Phi(t, y(t), h)\\ &= \frac{1}{2}\left\{\frac{y(t + \frac{h}{2}) - y(t)}{\frac{h}{2}} - \widehat{\Phi}(t, y(t), \tfrac{h}{2})\right\}\\ &\quad + \frac{1}{2}\left\{\frac{y(t + h) - y(t + \frac{h}{2})}{\frac{h}{2}} - \widehat{\Phi}(t + \tfrac{h}{2}, y(t + \tfrac{h}{2}), \tfrac{h}{2})\right\}\\ &\quad - \frac{1}{2}\left\{\widehat{\Phi}\left(t + \tfrac{h}{2}, y(t) + \tfrac{h}{2}\widehat{\Phi}(t, y(t), \tfrac{h}{2}), \tfrac{h}{2}\right) - \widehat{\Phi}\left(t + \tfrac{h}{2}, y(t + \tfrac{h}{2}), \tfrac{h}{2}\right)\right\}\end{aligned} \tag{1.20}$$

By assumption it holds

$$\frac{y(t+\frac{h}{2})-y(t)}{\frac{h}{2}} - \widehat{\Phi}(t,y(t),\tfrac{h}{2}) = \widehat{\tau}(t,\tfrac{h}{2}) = e(t)\left(\tfrac{h}{2}\right)^q + r_0(t,\tfrac{h}{2})\left(\tfrac{h}{2}\right)^{q+1} \tag{1.21}$$

and

$$\begin{aligned}\frac{y(t+h)-y(t+\frac{h}{2})}{\frac{h}{2}} - \widehat{\Phi}(t+\tfrac{h}{2},y(t+\tfrac{h}{2}),\tfrac{h}{2}) &= e(t+\tfrac{h}{2})\left(\tfrac{h}{2}\right)^q + r_0(t+\tfrac{h}{2},\tfrac{h}{2})\left(\tfrac{h}{2}\right)^{q+1} \\ &= e(t)\left(\tfrac{h}{2}\right)^q + \int_0^1 e'(t+\tfrac{h}{2}\eta)\,\mathrm{d}\eta\left(\tfrac{h}{2}\right)^{q+1} + r_0(t+\tfrac{h}{2},\tfrac{h}{2})\left(\tfrac{h}{2}\right)^{q+1} \\ &= e(t)\left(\tfrac{h}{2}\right)^q + r_1(t,\tfrac{h}{2})\left(\tfrac{h}{2}\right)^{q+1}\end{aligned} \tag{1.22}$$

and for the third term

$$\begin{aligned}&\left\|\widehat{\Phi}\left(t+\tfrac{h}{2},y(t)+\tfrac{h}{2}\widehat{\Phi}(t,y(t),\tfrac{h}{2}),\tfrac{h}{2}\right) - \widehat{\Phi}\left(t+\tfrac{h}{2},y(t+\tfrac{h}{2}),\tfrac{h}{2}\right)\right\| \\ &\le L\|y(t)+\tfrac{h}{2}\widehat{\Phi}(t,y(t),\tfrac{h}{2}) - y(t+\tfrac{h}{2})\| \\ &= L\|y(t+\tfrac{h}{2}) - \tfrac{h}{2}\widehat{\tau}(t,\tfrac{h}{2}) - y(t+\tfrac{h}{2})\| \\ &= L\left\|\tfrac{h}{2}\left(e(t)\left(\tfrac{h}{2}\right)^q + r_0(t,\tfrac{h}{2})\left(\tfrac{h}{2}\right)^{q+1}\right)\right\|\end{aligned} \tag{1.23}$$

Hence,

$$\widehat{\Phi}\left(t+\tfrac{h}{2},y(t)+\tfrac{h}{2}\widehat{\Phi}(t,y(t),\tfrac{h}{2}),\tfrac{h}{2}\right) - \widehat{\Phi}\left(t+\tfrac{h}{2},y(t+\tfrac{h}{2}),\tfrac{h}{2}\right) = r_2(t,\tfrac{h}{2})\left(\tfrac{h}{2}\right)^{q+1} \tag{1.24}$$

with a bounded function r_2. So combining (1.20),(1.21),(1.22) and (1.24) yields

$$\frac{y(t+h)-y(t)}{h} - \Phi(t,y(t),h) = e(t)\left(\tfrac{h}{2}\right)^q + r(t,h)\left(\tfrac{h}{2}\right)^{q+1}$$

for a bounded function r depending on r_0, r_1, r_2 yielding the first claim.
Using this result we obtain

$$\begin{aligned}\frac{y(t+h)-y(t)}{h} - \widetilde{\Phi}(t,y(t),h) &= \frac{2^q}{2^q-1}\left\{\frac{y(t+h)-y(t)}{h} - \Phi(t,y(t),h)\right\} \\ &\quad - \frac{1}{2^q-1}\left\{\frac{y(t+h)-y(t)}{h} - \widehat{\Phi}(t,y(t),h)\right\} \\ &= \frac{2^q}{2^q-1}\left\{e(t)\left(\tfrac{h}{2}\right)^q + r(t,h)\left(\tfrac{h}{2}\right)^{q+1}\right\} \\ &\quad - \frac{1}{2^q-1}\left\{e(t)h^q + r_0(t,h)h^{q+1}\right\} \\ &= \frac{1}{2^q-1}\left\{\frac{1}{2}r(t,h) - r_0(t,h)\right\}h^{q+1} \\ &= R(t,h)h^{q+1}.\end{aligned}$$

□

Note that explicit Runge-Kutta methods for smooth enough right hand sides f satisfy the assumptions of Theorem 1.53. Consequently, originating from a Runge-Kutta scheme of order q, we can design a one-step method based on extrapolation as described before with order $q+1$. The new scheme can be seen as an explicit Runge-Kutta method with **triple** amount of stages. This is **far from optimal**, in particular requiring many evaluations of the right hand side.

Embedded Runge-Kutta methods The idea of embedded Runge-Kutta methods is at step k to construct two iterates of the form

$$Y_{k+1} = Y_k + h\sum_{i=1}^{s} b_i K_i(t, Y_k, h), \qquad \text{and} \qquad \widetilde{Y}_{k+1} = Y_k + h\sum_{i=1}^{s} \widetilde{b}_i K_i(t, Y_k, h)$$

where only the parameters b_i and $\widetilde{b}_i$ possibly differ and the Runge-Kutta scheme using parameters $(\widetilde{b}_i)$ has a different order $\widetilde{q}$ than the one using (b_i) with order q, typically $\widetilde{q} = q+1$ or $\widetilde{q} = q-1$. Consequently, the computation of $\widetilde{Y}_{k+1}$ does **not require additional evaluations** of the right hand side f of the IVP. Such an embedded Runge-Kutta method is denoted as RK$\widetilde{q}(q)$.

The embedded Runge-Kutta scheme can be summarized as

$$\begin{array}{c|ccccc} 0 & 0 & \cdots & \cdots & \cdots & 0 \\ c_2 & a_{21} & \ddots & & & \vdots \\ c_3 & a_{31} & a_{32} & \ddots & & \vdots \\ \vdots & \vdots & \vdots & \ddots & & \vdots \\ c_s & a_{s1} & a_{s2} & \cdots & a_{s,s-1} & 0 \\ \hline & b_1 & b_2 & \cdots & b_{s-1} & b_s \\ & \widetilde{b}_1 & \widetilde{b}_2 & \cdots & \widetilde{b}_{s-1} & \widetilde{b}_s \end{array} = \begin{array}{c|c} \boldsymbol{c} & \boldsymbol{A} \\ \hline & \boldsymbol{b}^\top \\ & \widetilde{\boldsymbol{b}}^\top \end{array}.$$

Then, Theorem 1.42 can be applied to construct such a parameter sets for various schemes, assuming they exist!

Example 1.54 (Construction is not trivial) This example shows, that naive embedding may not work. Let us start with the classical Runge-Kutta scheme of order $\widetilde{q} = 4$ from Example 1.36 given as

$$\begin{array}{c|cccc} 0 & 0 & \cdots & \cdots & 0 \\ \frac{1}{2} & \frac{1}{2} & \ddots & & \vdots \\ \frac{1}{2} & 0 & \frac{1}{2} & \ddots & \vdots \\ 1 & 0 & 0 & 1 & 0 \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array}$$

We now seek for an embedded scheme by determining b_1, b_2, b_3, b_4 such that its corresponding incremental function Φ is of order $q = 3$. Application of Theorem 1.42 leads the only solution $\boldsymbol{b}^\intercal = (\frac{1}{6}, \frac{1}{3}, \frac{1}{3}, \frac{1}{6})$, which is nothing but the classical Runge-Kutta scheme of order $q = 4$. The construction failed. △

Surprisingly, adding a stage s may be required to actually allow for the construction of embedded pairs. To compensate for the increased number of evaluations of the right hand side f, the so called *Fehlberg-Trick* (FSAL, first same as last) requires, that the s-th stage K_s of the current iterate and the first stage K_1^* of the next iteration coincide in case of the same choosen (adaptive) stepsize h. In particular, adding the constrains

$$K_s = f(t+c_s h, t+h\sum_{j=1}^{s-1} a_{sj} K_j(t,y,h)) \overset{!}{=} f(t+h, y+h\sum_{j=1}^{s} b_j K_j(t,y,h)) = f(t+h, y+h\Phi(t,y,h)) = K_1^*. \tag{1.25}$$

Note that, in general there must not exist a solution for coefficients. But when they exist, they yield an efficient solution scheme.

Example 1.55 (Erwin Fehlberg's RK4(3), 1969) Based on classical Runge-Kutta method of order $\widetilde{q} = 4$ we consider the embedded tableau

$$\begin{array}{c|ccccc}
0 & 0 & \cdots & \cdots & \cdots & 0 \\
\frac{1}{2} & \frac{1}{2} & \ddots & & & \vdots \\
\frac{1}{2} & 0 & \frac{1}{2} & \ddots & & \vdots \\
1 & 0 & 0 & 1 & \ddots & \vdots \\
c_5 & a_{51} & a_{52} & a_{53} & a_{54} & 0 \\
\hline
 & b_1 & b_2 & b_3 & b_4 & b_5 \\
 & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} & 0
\end{array}$$

Applying Fehlberg-Trick and Theorem 1.42 leads parameters for order $q = 3$, Exercise. △

The RK4(3) is only a order 3 method, although its cost behave like a order 4 method. As an opposite idea one can use the control scheme of lower order, still obtain error estimation but having higher order scheme as the following example.

Example 1.56 (DOPRI5, $s = 6$, Dormand& Prince 1980) On of the most popular schemes is up to day is the RK4(5) scheme with an effective stage number $s = 6$ (due to Fehlberg Trick) by J.R. Dormand and P.J.Prince (1980) given as

$$\begin{array}{c|ccccccc}
0 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\
\frac{1}{5} & \frac{1}{5} & \ddots & & & & & \vdots \\
\frac{3}{10} & \frac{3}{40} & \frac{9}{40} & \ddots & & & & \vdots \\
\frac{4}{5} & \frac{44}{45} & -\frac{56}{15} & \frac{32}{9} & \ddots & & & \vdots \\
\frac{8}{9} & \frac{19372}{6561} & -\frac{25360}{2187} & \frac{64448}{6561} & -\frac{212}{729} & \ddots & & \vdots \\
1 & \frac{9017}{3168} & -\frac{355}{33} & \frac{46732}{5247} & \frac{49}{176} & -\frac{5103}{18656} & \ddots & \vdots \\
1 & \frac{35}{384} & 0 & \frac{500}{1113} & \frac{125}{192} & -\frac{2187}{6784} & \frac{11}{84} & 0 \\
\hline
 & \frac{35}{384} & 0 & \frac{500}{1113} & \frac{125}{192} & -\frac{2187}{6784} & \frac{11}{84} & 0 \quad q = 5 \\
 & \frac{5179}{57600} & 0 & \frac{7571}{16695} & \frac{393}{640} & -\frac{92097}{339200} & \frac{187}{2100} & \frac{1}{40} \quad \widetilde{q} = 4
\end{array}$$

It is implemented in DOPRI5 and Matlabs ode45, relies on a lower order control scheme and is of order 5, which is optimal according to Remark 1.47. △

Remark 1.57 Other explicit one step methods in praxis are RKF4(5) with $s = 6$ (Fehlberg (1964)), RKF7(8) with $s = 13$ (Fehlberg, 1969) or DOPRI7(8) with $s = 13$ (Prince&Dormand (1981)). △

1.2.4 Approximation of solution between grid points

Recalling Remark 1.32, a common approach to obtain approximate solution between grid points is obtained by piecewise affine linear interpolation between data pairs (t_k, Y_k) and (t_{k+1}, Y_{k+1}). However, not high accuracy has to be expected then. However in some applications this is required.

Continuous Runge-Kutta-Methods

For this reason *continuous Runge-Kutta schemes* have been developed. They aim for high accuracy between grid points, with neglectable additional cost, i.e. without additional evaluation of the right hand side f.
The ansatz is to parametrize the nodes $\boldsymbol{b} \to \boldsymbol{b}(\theta)$ for some $\theta \in [0,1]$ and approximate

$$y(t_k + \theta h_k) \approx Y_{k+\theta} := Y_k + h_k \sum_{i=1}^{s} b_i(\theta) K_i(t_k, Y_k, h_k),$$

while keeping the increments K_i untouched. Their Butcher-tableau reads

$$\begin{array}{c|ccc} 0 & 0 & \cdots & 0 \\ c_2 & a_{21} & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ c_s & a_{s1} & \cdots & a_{s,s-1} \quad 0 \\ \hline & b_1(\theta) & \cdots & b_s(\theta) \end{array} = \begin{array}{c|c} \boldsymbol{c} & \boldsymbol{A} \\ \hline & \boldsymbol{b}(\theta)^\top \end{array}$$

The goal then, is to choose coefficients c_i, α_{ij} and b_i such that the consistency order is uniformly as a high as possible for all $\theta \in [0,1]$.

Example 1.58 The order-3 RK scheme from Heun denoted as *Heun3* and its continuous extension with consistency order 2 uniform in $\theta \in (0,1)$ are given as

$$\begin{array}{c|ccc} 0 & 0 & \cdots & 0 \\ \frac{1}{3} & \frac{1}{3} & \ddots & \vdots \\ \frac{2}{3} & 0 & \frac{2}{3} & 0 \\ \hline & \frac{1}{4} & 0 & \frac{3}{4} \end{array} \qquad \begin{array}{c|ccc} 0 & 0 & \cdots & 0 \\ \frac{1}{3} & \frac{1}{3} & \ddots & \vdots \\ \frac{2}{3} & 0 & \frac{2}{3} & 0 \\ \hline & b_1(\theta) & b_2(\theta) & b_3(\theta) \end{array}$$

$$b_i(\theta) = \begin{cases} \frac{3}{2}\theta^3 - \frac{9}{4}\theta^2 + \theta & i = 1, \\ 3\theta^2(1-\theta) & i = 2, \\ \frac{3}{4}\theta^2(2\theta - 1) & i = 3. \end{cases}$$

Note that for $\theta = 1$ the continuous scheme recovers the order-3 scheme from Heun.

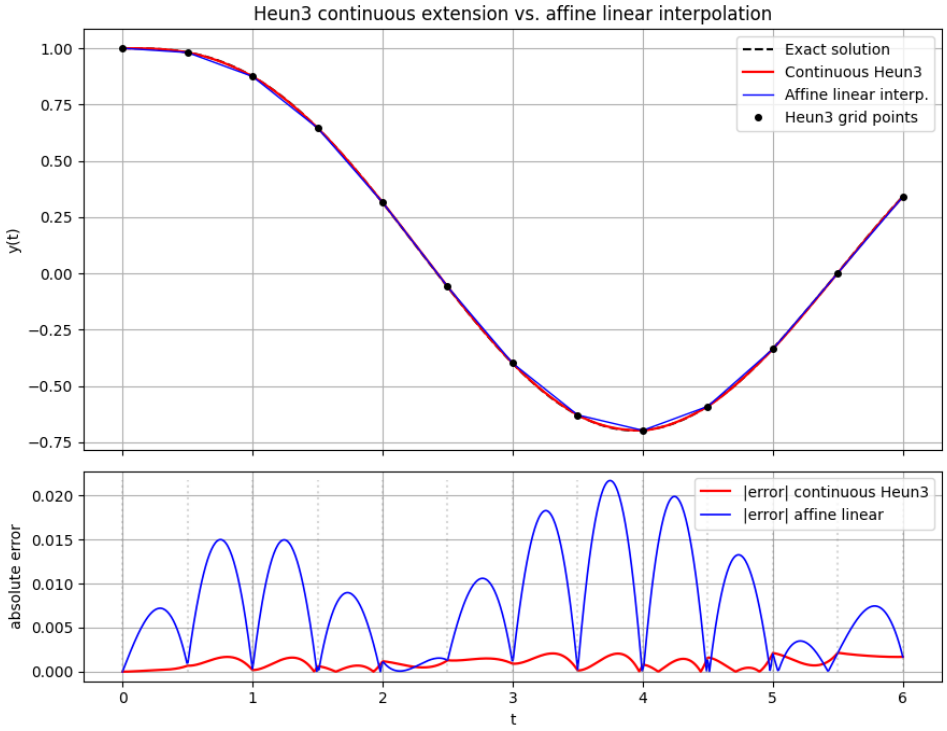


Figure 1.3: Comparison of Heun3 and continuous Heun3 for the IVP $y'(t) = cos(t) - y(t)$ with $y(0) = 1$ with exact solution $y(t) = 0.5(\sin(t) + \cos(t)) + 0.5\exp(-t)$.

△

Example 1.59 A convenient, commonly used dense-output (continuous extension) for DOPRI5 is a quartic polynomial in $\theta \in [0, 1]$ written in the nested form (Numerical Recipes / Shampine style):

$$y(t_k + \theta h) = r_1 + \theta\big(r_2 + (1-\theta)\big(r_3 + \theta(r_4 + (1-\theta)r_5)\big)\big),$$

with

$$\begin{aligned} r_1 &= Y_k, \\ r_2 &= Y_{k+1} - Y_k, \\ r_3 &= Y_k + hf_k - Y_{k+1}, \\ r_4 &= 2(Y_{k+1} - Y_k) - h(f_k + f_{k+1}), \\ r_5 &= d_1 h f_n + d_3 K_3 + d_4 K_4 + d_5 K_5 + d_6 K_6 + d_7 h f_{n+1}, \end{aligned}$$

where k_i are the stage increments $K_i = hf(t_k + c_i h,\ \cdot)$ for stages $i = 3, 4, 5, 6$ and $f_n = f(t_n, y_n)$, $f_{n+1} = f(t_{n+1}, y_{n+1})$. The numerical constants (as used in Numerical Recipes / ode45 implementations) are

$$\begin{aligned} d_1 &= -\frac{12715105075}{11282082432}, \\ d_3 &= \frac{87487479700}{32700410799}, \quad d_4 = -\frac{10690763975}{1880347072}, \\ d_5 &= \frac{701980252875}{199316789632}, \quad d_6 = -\frac{1453857185}{822651844}, \quad d_7 = \frac{69997945}{29380423}. \end{aligned}$$

Expanding the nested polynomial yields an explicit quartic polynomial in θ. This dense output

uses only the already computed stages and is the formula used in many implementations (e.g. ‘ode45‘), see Shampine (1986) and discussion in Numerical Recipes. Note that this construction is of the form

$$y(t_k + \theta h) = y(t_k) + h \sum_{i=1}^{s=7} b_i(\theta) K_i$$

with $b_i(0) = 0$ and $b_i(1) = b_i$ with b_i from non-continuous DOPRI5. The choice of polynomials satisfying this is not unique, and in the construction above $b_2(\theta) \equiv 0$ leading no K_2 depends at all (cp. to definition of r_5.)

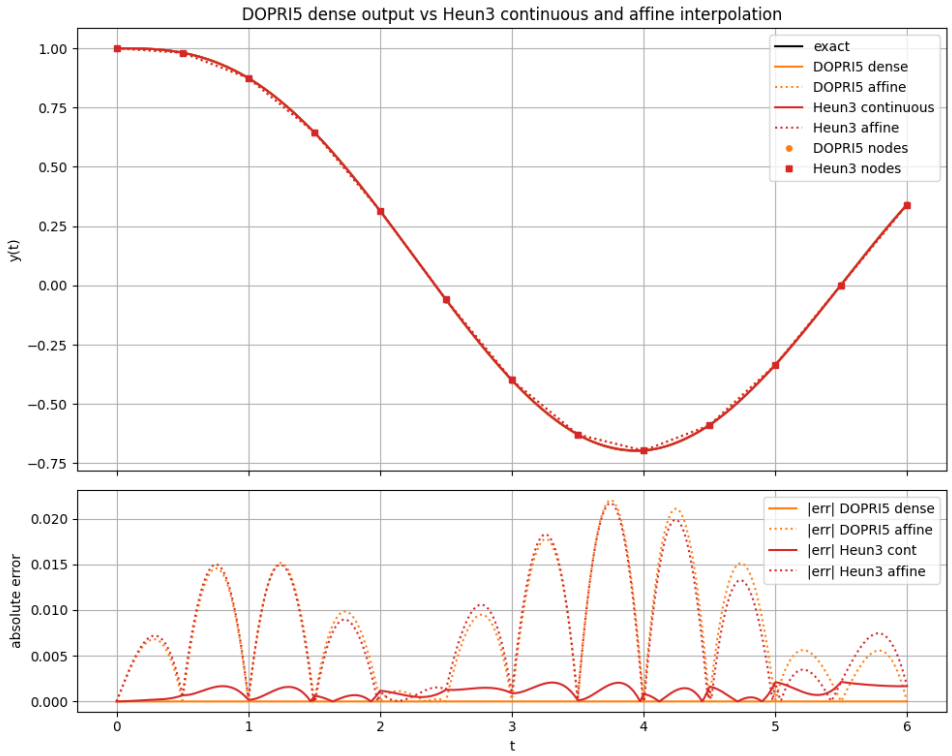


Figure 1.4: Comparison of continuous Heun3 and continuous DOPRI5 for the IVP $y'(t) = cos(t) - y(t)$ with $y(0) = 1$ with exact solution $y(t) = 0.5(\sin(t) + \cos(t)) + 0.5\exp(-t)$.

△

Bibliography

[1] John C Butcher. Coefficients for the study of runge-kutta integration processes. *Journal of the Australian Mathematical Society*, 3(2):185–201, 1963.

[2] Peter Deuflhard and Folkmar A Bornemann. Numerische mathematik. ii. 1994.

[3] Philip Hartman. Ordinary differential equations. 2002.