

Chapter 1

Explicit One-Step Methods

Remark 1.1. Contents. This course presents methods for the numerical solution of explicit systems of initial value problems for ordinary differential equations of first order

$$\mathbf{y}'(x) = \mathbf{f}(x, \mathbf{y}(x)), \quad \mathbf{y}(x_0) = \mathbf{y}_0.$$

For the most part, only initial value problems for scalar ordinary differential equations of first order

$$y'(x) = f(x, y(x)), \quad y(x_0) = y_0, \tag{1.1}$$

are considered, for simplicity of presentation. The extension of the results and the methods to systems is generally straightforward.

It will be always assumed that there is a unique solution of the initial value problem in a neighborhood of the initial value. In applications, the independent variable is often the time. \square

1.1 Consistency and Convergence

Definition 1.2. Grid, step size. A grid is a decomposition I_h of the interval $I = [x_0, x_e]$

$$I_h = \{x_0, x_1, \dots, x_N = x_e\}$$

with $x_0 < x_1 < \dots < x_N$. The differences between neighboring grid points $h_k = x_{k+1} - x_k$ are called step sizes. For an equidistant grid, the notation $h = h_k$ will be used for the step size, see Figure 1.1. \square

Remark 1.3. Explicit and implicit methods. Let $y(x_k)$ denote the solution of (1.1) in the node x_k and y_k a numerical approximation of $y(x_k)$. A numerical method for the solution of (1.1) on a grid I_h is called explicit, if an approximation y_{k+1} in x_{k+1} can be calculated directly by inserting already computed values y_i , $i \leq k$, in some formula(s). Otherwise, the method is called implicit method. Implicit methods require in each step the solution of a generally nonlinear equation for computing y_{k+1} . \square



Fig. 1.1 Equidistant grid.

Definition 1.4. One-step method, incremental function. A one-step method for the computation of an approximation y_{k+1} of the solution of (1.1) on a grid I_h has the form

$$y_{k+1} = y_k + h_k \Phi(x, y, h_k), \quad k = 0, 1, \dots, \quad x \in [x_k, x_{k+1}], \quad y_0 = y(x_0). \quad (1.2)$$

Here, $\Phi(\cdot, \cdot, \cdot)$ is called incremental function of the one-step method. \square

Example 1.5. One-step methods, incremental functions. The explicit or forward Euler method

$$y_{k+1} = y_k + h_k f(x_k, y_k), \quad k = 0, 1, 2, \dots, \quad y_0 = y(x_0),$$

is an explicit one-step method with the incremental function

$$\Phi(x, y, h_k) = f(x_k, y_k).$$

The computation of y_{k+1} requires only the substitution of already computed values in the function $f(x, y)$ from the initial value problem (1.1).

The implicit or backward Euler method

$$y_{k+1} = y_k + h_k f(x_{k+1}, y_{k+1}), \quad k = 0, 1, 2, \dots, \quad y_0 = y(x_0),$$

is an implicit one-step method with the incremental function

$$\Phi(x, y, h_k) = f(x_{k+1}, y_{k+1}).$$

One has to solve an equation for computing y_{k+1} . The complexity of this step depends on $f(x, y)$. \square

Remark 1.6. Representation of implicit one-step methods. Explicit one-step methods require only that known values are inserted in the incremental function. Hence, their incremental function can be written finally in the form $\Phi(x, y, h_k) = \Phi(x_k, y_k, h_k)$. For the considerations in this section, one can adopt the point of view that also implicit one-step methods can be written as explicit one-step methods, because the data for the nonlinear equation are x_k, y_k , and h_k . However, generally one does not know the concrete form of the incremental function. \square

Example 1.7. Incremental function of the implicit Euler method. The incremental function of the implicit Euler method on an equidistant grid can be written in the form

$$\Phi(x, y, h) = f(x + h, y + h\Phi(x, y, h)),$$

which allows formally the representation of this method as explicit one-step scheme. \square

Definition 1.8. Local error. Let \hat{y}_{k+1} be the result of one step of an explicit one-step method (1.2) with the initial value $y(x_k)$, i.e.,

$$\hat{y}_{k+1} = y(x_k) + h_k \Phi(x_k, y(x_k), h_k).$$

Then,

$$\begin{aligned} \text{le}(x_{k+1}) &= \text{le}_{k+1} = y(x_{k+1}) - \hat{y}_{k+1} \\ &= y(x_{k+1}) - (y(x_k) + h_k \Phi(x_k, y(x_k), h_k)) \end{aligned} \quad (1.3)$$

is called local error, see Figure 1.2 \square

Remark 1.9. The local error. In the literature, sometimes

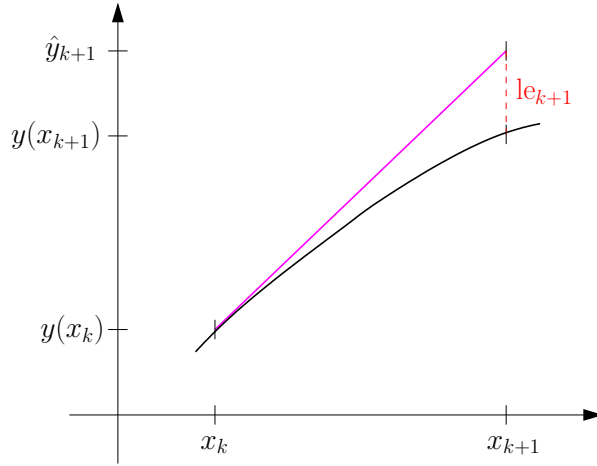


Fig. 1.2 The local error.

$$\frac{y(x_{k+1}) - y(x_k)}{h_k} - \Phi(x_k, y(x_k), h_k)$$

is defined to be the local error.

For the local error, one starts from the solution of the initial value problem and considers the error after one step of the numerical method.

One should require for a reasonable method that the local error is small in an appropriate sense. \square

Definition 1.10. Consistent method. Let $y(x)$ be the solution of the initial value problem (1.1), $h_{\max} = \max_k h_k$, and

$$S := \{(x, y) : x \in [x_0, x_e], y \in \mathbb{R}\}.$$

The one-step method (1.2) is said to be consistent, if for all $f \in C(S)$, which satisfy in S a Lipschitz condition with respect to y , it holds

$$\lim_{h_{\max} \rightarrow 0} \left(\max_{x_k \in I_h} \frac{|\text{le}(x_{k+1})|}{h_k} \right) = 0$$

or

$$\lim_{h_{\max} \rightarrow 0} \left(\max_{x_k \in I_h} |f(x_k, y(x_k)) - \Phi(x_k, y(x_k), h_k)| \right) = 0. \quad (1.4)$$

Both conditions are equivalent, compare Remark 1.11. \square

Remark 1.11. Approximation of the derivative with the incremental function. For bounded incremental functions, it is obvious that the local error converges to zero if $h_{\max} \rightarrow 0$, because in this case it holds $h_k \rightarrow 0$ and $y(x_{k+1}) \rightarrow y(x_k)$, such that this statement follows from (1.3). Consistency requires more, namely that the incremental function approximates the derivative of the solution sufficiently well. Applying (1.3) and (1.1) yields

$$\begin{aligned} \frac{\text{le}(x_{k+1})}{h_k} &= \frac{y(x_{k+1}) - y(x_k)}{h_k} - \Phi(x_k, y(x_k), h_k) \\ &\approx y'(x_k) - \Phi(x_k, y(x_k), h_k) \\ &= f(x_k, y_k) - \Phi(x_k, y(x_k), h_k), \end{aligned}$$

compare (1.4). \square

Example 1.12. Consistency of the explicit Euler method. For the explicit Euler method, it is $\Phi(x_k, y(x_k), h_k) = f(x_k, y(x_k))$. Hence, condition (1.4) from Definition 1.10 is satisfied and the method is consistent. \square

Remark 1.13. Quality of the approximation of the incremental function. For practical purposes, not only the consistency itself but the quality of the approximation of the derivative by the incremental function is essential. The quality allows a comparison of different one-step methods. For simplicity of presentation, let $h_k = h$ for all k . \square

Definition 1.14. Order of consistency. A one-step method (1.2) has the consistency order $p \in \mathbb{N}$, if p is the largest natural number such that for all functions $f \in C(S)$, which satisfy a Lipschitz condition with respect to y , it holds

$$|\text{le}(x_k + h)| \leq Ch^{p+1}$$

for all $x_k \in I_h$, for all I_h with $h \in (0, H]$, and with the constant $C > 0$ being independent of h . The constant C might depend on derivatives of $y(x)$, on $f(x, y)$, and on partial derivatives of $f(x, y)$. \square

Example 1.15. Order of consistency of the explicit Euler method. Consider the explicit Euler method and assume that the function $y(x)$ is two times continuously differentiable. Then, it follows with Taylor series expansion and using the differential equation that

$$\begin{aligned} |\text{le}(x_k + h)| &= |y(x_k + h) - \hat{y}_{k+1}| \\ &= |y(x_k) + hy'(x_k) + \frac{h^2}{2}y''(x_k + \theta h) - y(x_k) - \underbrace{hf(x_k, y(x_k))}_{=y'(x_k)}| \\ &= \frac{h^2}{2} |y''(x_k + \theta h)| \leq \frac{h^2}{2} \|y\|_{C^2([x_0, x_e])}, \end{aligned}$$

with $\theta \in (0, 1)$. Since there is no way to replace the term on the right-hand side by a term with a larger power of h , the method has consistency order 1. \square

Remark 1.16. Consistency and convergence. The consistency is a local property of a one-step method. For practical purposes, it is important that the computed solution converges to the analytic solution if the grid becomes finer and finer. Of course, the order of convergence is of importance, too.

It will be shown that, under certain conditions, the convergence of a one-step method follows from its consistency and that the order of convergence equals the consistency order. \square

Definition 1.17. Convergent method, order of convergence. A one-step method (1.2) converges for the initial value problem (1.1) on the interval $I = [x_0, x_e]$, if for each sequence of grids $\{I_h\}$ with $h_{\max} = \max_k h_k \rightarrow 0$ for the global error

$$e(x_k, h) = y(x_k) - y_k, \quad x_k \in I_h,$$

it follows that

$$\max_{x_k \in I_h} |e(x_k, h)| \rightarrow 0 \quad \text{for} \quad h_{\max} \rightarrow 0.$$

The one-step method has the order of convergence p^* , if p^* is the largest natural number such that for all step lengths $h_{\max} \in (0, H]$, for some $H > 0$, it holds

$$|e(x_k, h)| \leq Ch_{\max}^{p^*} \quad \forall x_k \in I_h,$$

where $C > 0$ is independent of h_{\max} . \square

Lemma 1.18. Estimate for a sequence of real numbers. Assume that for real numbers x_n , $n = 0, 1, \dots$, the inequality

$$|x_{n+1}| \leq (1 + \delta) |x_n| + \beta$$

holds with constants $\delta > 0$, $\beta \geq 0$. Then, it holds that

$$|x_n| \leq e^{n\delta} |x_0| + \frac{e^{n\delta} - 1}{\delta} \beta, \quad n = 0, 1, \dots$$

Proof. With induction, problem for exercises. ■

Theorem 1.19. Connection of consistency and convergence. Let $y(x)$ be the solution of the initial value problem (1.1) with $f \in C(S)$. Let a Lipschitz condition hold for the second argument of the incremental function

$$\begin{aligned} |\Phi(x, y_1, h) - \Phi(x, y_2, h)| &\leq M |y_1 - y_2| \\ \forall x \in [x_0, x_e], y_1, y_2 \in \mathbb{R}, h \in (0, H), \end{aligned} \quad (1.5)$$

with $M \in \mathbb{R}$, $M > 0$, fixed. Assume that for the local error the estimate

$$|\text{le}(x_k + h)| \leq Ch^{p+1} \quad \forall x_k \in I_h, h \in (0, H) \quad (1.6)$$

is valid and assume that $y_0 = y(x_0)$.

Then, it follows for the global error that

$$|e(x_{k+1}, h)| \leq C \frac{e^{M(x_{k+1} - x_0)} - 1}{M} h^p, \quad (1.7)$$

where C is independent of h .

Proof. Using Remark 1.6 and the representation of the local error (1.3), one finds that

$$\begin{aligned} y_{k+1} &= y_k + h\Phi(x_k, y_k, h), \\ y(x_{k+1}) &= y(x_k) + h\Phi(x_k, y(x_k), h) + \text{le}(x_{k+1}), \quad k = 0, 1, \dots \end{aligned}$$

Then, it follows with the triangle inequality, the assumption on the local error (1.6), and the Lipschitz condition of the incremental function (1.5) that

$$\begin{aligned} |e(x_{k+1}, h)| &= |y(x_{k+1}) - y_{k+1}| \\ &= |y(x_k) - y_k + \text{le}(x_{k+1}) + h(\Phi(x_k, y(x_k), h) - \Phi(x_k, y_k, h))| \\ &= |e(x_k, h) + \text{le}(x_{k+1}) + h(\Phi(x_k, y(x_k), h) - \Phi(x_k, y_k, h))| \\ &\leq |e(x_k, h)| + |\text{le}(x_{k+1})| + h|\Phi(x_k, y(x_k), h) - \Phi(x_k, y_k, h)| \\ &\leq |e(x_k, h)| + Ch^{p+1} + hM |y(x_k) - y_k| \\ &= (1 + hM) |e(x_k, h)| + Ch^{p+1}. \end{aligned}$$

This sequence of inequalities has the form that was considered in Lemma 1.18. One obtains with $e(x_0) = 0$

$$|e(x_{k+1}, h)| \leq e^{(k+1)hM} |e(x_0)| + C \frac{e^{(k+1)hM} - 1}{hM} h^{p+1} = C \frac{e^{M(x_{k+1} - x_0)} - 1}{M} h^p. \quad \blacksquare$$

Remark 1.20. To Theorem 1.19.

- The constant in the error bound might be very large because of the exponential term, in particular if M is large or the interval is long.
- The consideration of a constant step length is only for simplicity of presentation. The result of the theorem holds also for non-constant step lengths with $h = \max_k h_k$.
- One-step methods compute an approximation y_k of the solution in the grid points x_k , $k = 0, 1, \dots, N$. To enable a better comparison with the analytic solution, one connects these points linearly from (x_k, y_k)

to (x_{k+1}, y_{k+1}) . In this way, one obtains a continuous piecewise linear approximation (polygonal approximation) of the solution that is defined on $[x_0, x_e]$. This function is called $y^h(x)$. The considerations from above can be extended to $y^h(x)$. □

1.2 Explicit Runge–Kutta Schemes

Remark 1.21. Idea. The Euler methods are only of first order. The idea of Runge¹–Kutta² methods consists in using an incremental function $\Phi(x, y, h)$ that is a linear combination of values of $f(x, y)$ in different points. With this approach, one obtains methods of higher order for the cost of evaluating more values of $f(x, y)$.

This approach can be illustrated well at the integral equation that is equivalent to the initial value problem (1.1). For simplicity, let the right-hand side of (1.1) depend only on x . Then, the integral equation has the form

$$y(x) = y_0 + \int_{x_0}^x f(t) dt. \quad (1.8)$$

The idea of the Runge–Kutta methods consists in approximating the right-hand side by a quadrature rule, e.g., in the interval $[x_k, x_{k+1}]$ by

$$\int_{x_k}^{x_{k+1}} f(t) dt \approx h_k \sum_{j=1}^s b_j f(x_k + c_j h_k)$$

with the weights b_j and the nodes $x_k + c_j h$.

In the following, only $h_k = h$ for all k will be considered for the sake of simplicity. □

Definition 1.22. Runge–Kutta methods, increments, and stages. A Runge–Kutta method has the form

$$y_{k+1} = y_k + h\Phi(x, y, h), \quad k = 0, 1, \dots, \quad y_0 = y(x_0),$$

where the incremental function is defined with the help of

$$K_i(x, y, h) = f\left(x_k + c_i h, y_k + h \sum_{j=1}^s a_{ij} K_j(x, y, h)\right)$$

by

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

with $c_1, \dots, c_s, b_1, \dots, b_s, a_{ij} \in \mathbb{R}$, $i, j = 1, \dots, s$. The quantities $K_i(x, y, h)$, $i = 1, \dots, s$, are called increments. The natural number $s \in \mathbb{N}$ is the number of stages of the method.

An equivalent definition is as follows

$$y_{k+1}^{(i)} = y_k + h \sum_{j=1}^s a_{ij} f\left(x_k + c_j h, y_{k+1}^{(j)}\right), \quad (1.9)$$

$$\Phi(x, y, h) = \sum_{i=1}^s b_i f\left(x_k + c_i h, y_{k+1}^{(i)}\right). \quad (1.10)$$

¹ Carle David Tolmé Runge (1856 – 1927)

² Martin Kutta (1867 – 1944)

The intermediate values $y_{k+1}^{(i)}$ are called stages. □

Remark 1.23. Butcher³ tableau. For the reason of clarity, one writes a Runge–Kutta scheme in general in form of a tableau, the so-called Butcher tableau

$$\begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \cdots & a_{1s} \\ c_2 & a_{21} & a_{22} & \cdots & a_{2s} \\ c_3 & a_{31} & a_{32} & \cdots & a_{3s} \\ \vdots & \vdots & \vdots & & \vdots \\ c_s & a_{s1} & a_{s2} & \cdots & a_{ss} \\ \hline & b_1 & b_2 & \cdots & b_s \end{array} \implies \frac{\mathbf{c}}{\mathbf{b}^T} A.$$

Here, \mathbf{c} are the nodes, A is the matrix of the method, and \mathbf{b} are the weights. The notions for \mathbf{c} and \mathbf{b} come from their roles in the numerical quadrature. □

Remark 1.24. Increments and Butcher tableau. For explicit Runge–Kutta schemes, the increments can be computed one after the other by inserting values in given formulas

$$\begin{aligned} K_1(x, y, h) &= f(x_k, y_k), \\ K_2(x, y, h) &= f(x_k + c_2h, y_k + ha_{21}K_1(x, y, h)), \\ &\vdots \\ K_s(x, y, h) &= f\left(x_k + c_sh, y_k + h \sum_{j=1}^{s-1} a_{sj}K_j(x, y, h)\right). \end{aligned} \tag{1.11}$$

The Butcher tableau has the form

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

A Runge–Kutta method is explicit if and only if the matrix of the method is a strict lower triangular matrix. □

Example 1.25. Explicit Euler scheme. The explicit Euler scheme is an explicit Runge–Kutta scheme with the Butcher tableau

$$\frac{0}{1}.$$

In the integral equation, the approximation

$$\int_{x_k}^{x_{k+1}} f(t, y(t)) dt \approx hf(x_k, y(x_k))$$

is used, see the proof of the Theorem of Peano, lectures notes of Numerical Mathematics I or literature. □

Theorem 1.26. Consistency of explicit Runge–Kutta schemes. *Let $f \in C(S)$, see Definition 1.10. An explicit Runge–Kutta scheme is consistent if and only if*

³ John C. Butcher, born 1933

$$\sum_{i=1}^s b_i = 1. \quad (1.12)$$

Proof. From the continuity of $f(x, y)$ and the definition (1.11) of the increments of an explicit Runge–Kutta scheme, it follows that

$$\lim_{h \rightarrow 0} K_i(x, y, h) = f(x_k, y(x_k)), \quad \forall (x, y) \in S, \quad i = 1, \dots, s,$$

for the case that the initial value of this step is $y_k = y(x_k)$. The continuity of the absolute value function gives

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

if and only if $\sum_{i=1}^s b_i = 1$. If and only if this equality is satisfied the condition (1.4) in Definition 1.10 is satisfied. \blacksquare

Theorem 1.27. Interpretation of the increments. *Let for the solution of (1.1) hold $y \in C^2([x_0, x_e])$, let $f \in C(S)$, and let f be Lipschitz continuous in the second argument. If $y_k = y(x_k)$ and*

$$c_i = \sum_{j=1}^{i-1} a_{ij}, \quad i \geq 2, \quad (1.13)$$

holds, then $K_i(x, y, h)$ is an approximation of at least first order (of consistency) to $y'(x_k + c_i h)$, i.e.,

$$y'(x_k + c_i h) - K_i(x, y, h) = \mathcal{O}(h^2).$$

Proof. For interested students only.

The proof follows by induction.

$i = 2$. For $i = 2$, it follows with (1.1), the Lipschitz continuity, and Taylor series expansion that

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

Hence, in the case $c_2 = a_{21}$, the difference is of order $\mathcal{O}(h^2)$.

$i > 2$. Let the asymptotic order of the errors be proved for all indices $2, \dots, i-1$. Then, one gets in the same way as for $i = 2$

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

$$\begin{aligned}
&= L \left| c_i h y'(x_k) + \mathcal{O}(h^2) - h \sum_{j=1}^{i-1} (a_{ij} (y'(x_k) + \mathcal{O}(h))) \right| \\
&= L \left| h \left(c_i - \sum_{j=1}^{i-1} a_{ij} \right) y'(x_k) + \mathcal{O}(h^2) \right|.
\end{aligned}$$

The order of the difference $\mathcal{O}(h^2)$ is given, if $c_i = \sum_{j=1}^{i-1} a_{ij}$. ■

Remark 1.28. Conditions on the coefficients for certain orders of convergence. The condition from Theorem 1.26 is satisfied for all explicit Runge–Kutta schemes and the condition from Theorem 1.27 is satisfied for many explicit Runge–Kutta schemes.

The goal consists in determining the coefficients $c_1, \dots, c_2, b_1, \dots, b_s$, and a_{ij} in such a way that one obtains an order of consistency as high as possible. The consistency order of a Runge–Kutta scheme with s stages can be derived from the Taylor series expansion of the local error. Let (1.12) be valid, then one obtains, e.g.,

- A Runge–Kutta scheme with the parameters $(A, \mathbf{b}, \mathbf{c})$ has at least consistency order $p = 2$ if, in addition to (1.12),

$$\sum_{j=1}^s b_j c_j = \frac{1}{2}. \quad (1.14)$$

This condition will be shown in Example 1.29 for $s = 2$.

- If in addition

$$\sum_{j=1}^s b_j c_j^2 = \frac{1}{3} \quad \text{and} \quad \sum_{j=1}^s b_j \sum_{k=1}^s a_{jk} c_k = \frac{1}{6}$$

hold, then the order of consistency is at least $p = 3$. A special case will be studied in the exercise problems. Proofs for the general conditions and conditions for even higher order consistency can be found in the literature, e.g. in (Strehmel & Weiner, 1995; Strehmel *et al.*, 2012, Section 2.4.2). □

Example 1.29. Runge–Kutta methods with 2 stages. For the investigation of 2-stage Runge–Kutta schemes, one considers for simplicity the so-called autonomous initial value problem

$$y'(x) = f(y(x)), \quad y(x_0) = y_0.$$

One has for the increments

$$\begin{aligned}
K_1(y, h) &= f(y_k), \\
K_2(y, h) &= f(y_k + h a_{21} K_1(y_k, h)) = f(y_k + h a_{21} f(y_k)) \\
&= f(y_k) + h a_{21} f(y_k) \partial_y f(y_k) + \mathcal{O}(h^2).
\end{aligned}$$

If the initial value is exact, it follows for the incremental function that

$$\begin{aligned}
\Phi(y(x_k)) &= b_1 K_1(y, h) + b_2 K_2(y, h) \\
&= (b_1 + b_2) f(y(x_k)) + h b_2 a_{21} f(y(x_k)) \partial_y f(y(x_k)) + \mathcal{O}(h^2).
\end{aligned} \quad (1.15)$$

The Taylor series expansion of the solution has the form

$$y(x_k + h) = y(x_k) + h \underbrace{y'(x_k)}_{=f(y(x_k))} + \frac{h^2}{2} y''(x_k) + \mathcal{O}(h^3).$$

One obtains with the chain rule

$$y''(x) = \frac{d}{dx}y'(x) = \frac{d}{dx}f(y(x)) = \partial_y f(y)y'(x) = \partial_y f(y)f(y(x)).$$

Now, it follows for the local error, using the Taylor series expansion and (1.15), that

$$\begin{aligned} \text{le}(x_k + h) &= y(x_k + h) - y(x_k) - h\Phi(y(x_k)) \\ &= y(x_k) + hf(y(x_k)) + \frac{h^2}{2}(\partial_y f(y(x_k))f(y(x_k))) + \mathcal{O}(h^3) - y(x_k) \\ &\quad - h\left((b_1 + b_2)f(y(x_k)) + hb_2a_{21}f(y(x_k))\partial_y f(y(x_k)) + \mathcal{O}(h^2)\right) \\ &= h(1 - (b_1 + b_2))f(y(x_k)) + h^2\left(\frac{1}{2} - b_2a_{21}\right)f(y(x_k))\partial_y f(y(x_k)) + \mathcal{O}(h^3). \end{aligned}$$

To achieve an order of consistency as large as possible, the first two terms have to vanish. One obtains with the condition $c_2 = a_{21}$ that

$$b_1 + b_2 = 1, \quad b_2a_{21} = \frac{1}{2} \iff b_2c_2 = \frac{1}{2}.$$

The first equation is the general condition for consistency (1.12) and the second condition is exactly (1.14) for $s = 2$. These two conditions characterize all 2-stage explicit Runge–Kutta methods that possess consistency and convergence order 2

$$\frac{c_2}{1 - \frac{c_2}{2c_2} \frac{1}{2c_2}}, \quad \text{with } c_2 \neq 0.$$

In the case $c_2 = 1/2$, one obtains the method of Runge (1895)

$$\frac{1/2 \mid 1/2}{\mid 0 \quad 1}.$$

This method corresponds with respect to the approximation of the integral in (1.8) to the application of the mid point rule.

For $c_2 = 1$, one gets the method of Heun⁴ (1900)

$$\frac{1 \mid 1}{\mid 1/2 \quad 1/2},$$

which corresponds to the use of the trapezoidal rule for the numerical quadrature in (1.8). □

Remark 1.30. Autonomous ordinary differential equations. Every explicit first order ordinary differential equation

$$\mathbf{y}'(x) = \mathbf{f}(x, \mathbf{y}(x))$$

can be transformed into an autonomous form

$$\tilde{\mathbf{y}}'(x) = \tilde{\mathbf{f}}(\tilde{\mathbf{y}}(x)) = \begin{pmatrix} \mathbf{f}(x, \mathbf{y}(x)) \\ 1 \end{pmatrix}$$

by introducing the function

$$\bar{y}(x) := x \quad \text{and} \quad \tilde{\mathbf{y}}(x) := \begin{pmatrix} \mathbf{y}(x) \\ \bar{y}(x) \end{pmatrix}$$

and noting that $(\mathbf{y}(x), x)$ are just the components of $\tilde{\mathbf{y}}(x)$. □

⁴ Karl Heun (1859 – 1929)

Theorem 1.31. Consistency and convergence of explicit Runge–Kutta methods. Let $y(x)$ be the solution of the initial value problem (1.1) with $f \in C(S)$ and let $f(x, y)$ satisfy a Lipschitz condition in the second argument. Then, an explicit Runge–Kutta scheme that is consistent of order p converges also with order p .

Proof. The incremental function of an explicit Runge–Kutta scheme is a linear combination of values of the right-hand side $f(x, y)$. Thus, the assumptions of Theorem 1.19 are satisfied, since the Lipschitz condition in this theorem follows from the assumed Lipschitz condition on the right-hand side of the differential equation. The statement of the theorem follows now directly from Theorem 1.19. ■

Remark 1.32. Explicit Runge–Kutta methods of higher order. Analogously to 2-stage methods, it is possible to derive conditions on the coefficients of an explicit Runge–Kutta scheme in order to construct methods of higher order. An important question is the minimal number of stages that is necessary to be able to reach a certain order. Some answers to this question are from Butcher (1963, 1965, 1985):

$$\frac{p}{\min s} \left| \begin{array}{cccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ \hline 1 & 2 & 3 & 4 & 6 & 7 & 9 & 11 \end{array} \right.$$

□

Example 1.33. Classical Runge–Kutta scheme (1901). The so-called classical Runge–Kutta scheme has four stages and the Butcher tableau

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

It is based on the Simpson⁵ rule. The center node of the Simpson rule is used twice, $c_2 = c_3$, but with a different second argument for the computation of the increments. This method is of fourth order. □

1.3 Step Length Control

Remark 1.34. Motivation. The considerations so far did not provide a way for estimating a good step length for solving a given initial value problem with prescribed accuracy and with as little work as possible.

- If the steps are too large, then the numerical solution might be too inaccurate.
- If the steps are too small, then the numerical simulation might take much longer than necessary.

A good step length depends certainly on the concrete problem and generally it will change within the considered interval. For these reasons, the step length should be controlled during the numerical simulation of the initial value problem.

A typical approach consists in computing two approximations of the solution in a node with different methods and to draw conclusions on the size of the local error, based on the difference of these approximations. Of course, the consideration of the global error would be better. However, Theorem 1.19 shows that on the one hand, the global error is influenced by problem-dependent terms, like the length of the interval $[x_0, x_e]$ or the Lipschitz constant. On the other hand, the global error is expected to be small only if the local errors are small. □

⁵ Thomas Simpson (1710 – 1761)

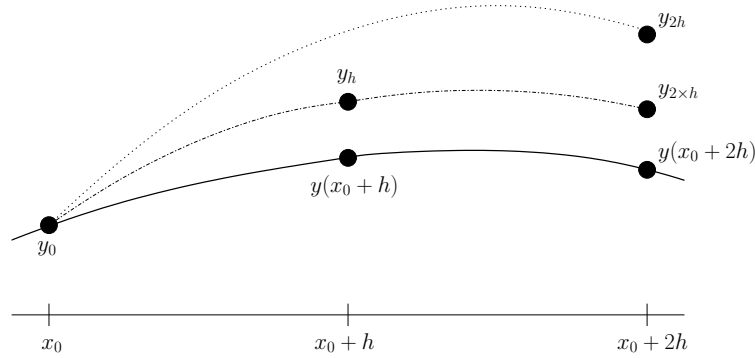


Fig. 1.3 Sketch of the Richardson method.

1.3.1 The Richardson Method

Remark 1.35. Idea. Given a numerical method for solving an initial value problem and given a step length h . The Richardson⁶ method consists of the following steps, see also Figure 1.3:

1. Starting from a node (x_0, y_0) and using a step length of $2h$, an approximation y_{2h} at the node $x_0 + 2h$ will be computed.
2. Two approximations y_h and $y_{2 \times h}$ in $x_0 + h$ and $x_0 + 2h$ are computed with two steps of length h .
3. The step length will be controlled by comparing y_{2h} and $y_{2 \times h}$.

In general, the more accurate approximation will be $y_{2 \times h}$. In addition, it will be demonstrated that it is possible to improve the accuracy of $y_{2 \times h}$ with the information obtained by this method. \square

Example 1.36. Richardson method for an explicit 2-stage Runge–Kutta method. Consider an explicit 2-stage Runge–Kutta scheme. One obtains in the first step of the Richardson method, using (1.9), (1.10),

$$\begin{aligned} y_{2h}^{(1)} &= y_0, \\ y_{2h}^{(2)} &= y_0 + 2ha_{21}f(x_0, y_0), \\ y_{2h} &= y_0 + 2h [b_1K_1(x, y) + b_2K_2(x, y)] \\ &= y_0 + 2h \left[b_1f \left(x_0, y_{2h}^{(1)} \right) + b_2f \left(x_0 + 2c_2h, y_{2h}^{(2)} \right) \right], \end{aligned}$$

or written as Butcher tableau

$$\begin{array}{c|cc} 0 & & \\ \hline 2c_2 & 2a_{21} & \\ \hline & 2b_1 & 2b_2 \end{array}.$$

Notice that because of the step length $2h$, the weights sum up to 2.

The second step of the Richardson method yields

$$\begin{aligned} y_{2 \times h}^{(1)} &= y_0, \\ y_{2 \times h}^{(2)} &= y_0 + ha_{21}f(x_0, y_0), \\ y_{2 \times h}^{(3)} &= y_h = y_0 + h \left[b_1f \left(x_0, y_{2 \times h}^{(1)} \right) + b_2f \left(x_0 + c_2h, y_{2 \times h}^{(2)} \right) \right], \\ y_{2 \times h}^{(4)} &= y_h + ha_{21}f(x_0 + h, y_h), \\ y_{2 \times h} &= y_h + h \left[b_1f(x_0 + h, y_h) + b_2f \left(x_0 + h + c_2h, y_{2 \times h}^{(4)} \right) \right]. \end{aligned}$$

⁶ Lewis Fry Richardson (1881 – 1953)

Inserting the formula for y_h in the last two lines, one sees that the Butcher tableau of this method is

$$\begin{array}{c|ccc} 0 & & & \\ c_2 & a_{21} & & \\ 1 & b_1 & b_2 & \\ \hline 1 + c_2 & b_1 & b_2 & a_{21} \\ \hline & b_1 & b_2 & b_1 & b_2 \end{array}.$$

That means, the computation of $y_{2 \times h}$ is equivalent to the computation of an approximation with the help of an explicit 4-stage Runge–Kutta scheme.

Altogether, five function evaluations are needed:

$$f(x_0, y_0), f(x_0 + 2c_2h, y_{2h}^{(2)}), f(x_0 + c_2h, y_{2 \times h}^{(2)}), f(x_0 + h, y_h), f(x_0 + h + c_2h, y_{2 \times h}^{(4)}).$$

In the case of a s -stage Runge–Kutta method, $(3s - 1)$ function evaluations are required. This number is rather large and the high costs per time step are a disadvantage of the Richardson method. \square

Remark 1.37. Comparison of both approximations. Consider a one-step method

$$y_{k+1} = y_k + h\Phi(x, y, h)$$

of order p . Let the initial value $y(x_0)$ be exact, then it follows for the local error in $x_0 + 2h$ that

$$y(x_0 + 2h) - y_{2h} = C(x_0)(2h)^{p+1} + \mathcal{O}(h^{p+2}). \quad (1.16)$$

For estimating the local error of $y_{2 \times h}$ it will be assumed that the incremental function $\Phi(x, y, h)$ is Lipschitz continuous in the second argument. This assumption is always satisfied for explicit Runge–Kutta schemes if $f(x, y)$ possesses this property, see the proof of Theorem 1.31. It is

$$y_{2 \times h} = y_h + h\Phi(x + h, y_h, h). \quad (1.17)$$

Let

$$\hat{y}_{2 \times h} = y(x_0 + h) + h\Phi(x + h, y(x_0 + h), h) \quad (1.18)$$

be the iterate that is computed with the exact starting value in $x_0 + h$. Using the definition of the consistency order, one obtains with (1.17) and (1.18)

$$\begin{aligned} & y(x_0 + 2h) - y_{2 \times h} \\ &= (y(x_0 + 2h) - \hat{y}_{2 \times h}) + (\hat{y}_{2 \times h} - y_{2 \times h}) \\ &= \left[C(x_0 + h)h^{p+1} + \mathcal{O}(h^{p+2}) \right] + \left[y(x_0 + h) + h\Phi(x + h, y(x_0 + h), h) - y_h - h\Phi(x + h, y_h, h) \right]. \end{aligned}$$

For the terms with the incremental function, one gets from the Lipschitz continuity and the consistency order for the first step

$$|h\Phi(x + h, y(x_0 + h), h) - h\Phi(x + h, y_h, h)| \leq hL \underbrace{|y(x_0 + h) - y_h|}_{\mathcal{O}(h^{p+1})} = \mathcal{O}(h^{p+2}).$$

It follows, applying again the consistency error for the first step, that

$$\begin{aligned} y(x_0 + 2h) - y_{2 \times h} &= C(x_0 + h)h^{p+1} + y(x_0 + h) - y_h + \mathcal{O}(h^{p+2}) \\ &= C(x_0 + h)h^{p+1} + C(x_0)h^{p+1} + \mathcal{O}(h^{p+2}) + \mathcal{O}(h^{p+2}) \end{aligned}$$

$$= 2C(x_0)h^{p+1} + \mathcal{O}(h^{p+2}), \quad (1.19)$$

where one assumes that $C(x_0 + h) = C(x_0) + \mathcal{O}(h)$, i.e., that the constants do not change too rapidly.

Neglecting in (1.16) and (1.19) the higher order terms allows to eliminate $y(x_0 + 2h)$ and solve for the constant, yielding

$$C(x_0) = \frac{1}{2} \left(\frac{y_{2 \times h} - y_{2h}}{2^p - 1} \right) \frac{1}{h^{p+1}}. \quad (1.20)$$

From (1.19), it follows for the local error of the more accurate method that

$$y(x_0 + 2h) - y_{2 \times h} = \frac{y_{2 \times h} - y_{2h}}{2^p - 1} + \mathcal{O}(h^{p+2}). \quad (1.21)$$

The first term on the right-hand side is a computable approximation of this local error. \square

Remark 1.38. Increasing the accuracy, local Richardson extrapolation. Rearranging terms in (1.21) gives

$$y(x_0 + 2h) - \left(y_{2 \times h} + \frac{y_{2 \times h} - y_{2h}}{2^p - 1} \right) = \mathcal{O}(h^{p+2}).$$

Then,

$$\bar{y}_{2 \times h} = y_{2 \times h} + \frac{y_{2 \times h} - y_{2h}}{2^p - 1}$$

is an approximation of the solution of order $p + 1$. This approach is called local Richardson extrapolation. \square

Remark 1.39. Automatic step length control. From (1.21) and (1.20), it follows that

$$\text{err} = \frac{|y_{2 \times h} - y_{2h}|}{2^p - 1} \approx 2C(x_0)h^{p+1} \quad (1.22)$$

is a computable approximation of the local error. This approximation will be compared with a prescribed tolerance. Often, a so-called scaled tolerance sc is used, (Hairer *et al.*, 1993, p. 167) or (Strehmel *et al.*, 2012, p. 61). The scaled tolerance is a combination of an absolute tolerance atol and a relative tolerance rtol

$$\text{sc} = \text{atol} + \max\{|y_0|, |y_{2 \times h}|\} \text{rtol}.$$

Then, the scaled error

$$\text{err}_{\text{sc}} = \frac{|y_{2 \times h} - y_{2h}|}{(2^p - 1)\text{sc}}$$

is defined.

- If $\text{err}_{\text{sc}} \leq 1 \iff \text{err} \leq \text{sc}$, then the performed step will be accepted. Starting from $y_{2 \times h}$ or $\bar{y}_{2 \times h}$, the next step will be performed.

An important aspect is the choice of the step length h_{new} for the next step. The guideline is that the scaled error for the next step should be on the one hand still smaller than or equal to 1 but on the other hand as close to 1 as possible. Following (1.22), it should hold

$$\begin{aligned} 1 &= \frac{\text{err}_{\text{new}}}{\text{sc}} = \frac{2C(x_0 + 2h)h_{\text{new}}^{p+1}}{\text{sc}} \approx \frac{2C(x_0)h_{\text{new}}^{p+1}}{\text{sc}} \\ &= \frac{2C(x_0)h^{p+1}}{\text{sc}} \left(\frac{h_{\text{new}}}{h} \right)^{p+1} \approx \text{err}_{\text{sc}} \left(\frac{h_{\text{new}}}{h} \right)^{p+1}, \end{aligned}$$

i.e., h_{new} has to be chosen such that