J. Stoer
R. Bulirsch

Introduction to Numerical Analysis

Translated by R. Bartels, W. Gautschi, and C. Witzgall

Springer-Verlag
New York Berlin Heidelberg London
Paris Tokyo Hong Kong Barcelona

Lemma (7.2.2.2), since $\tilde{e}_0 = \tilde{\eta}_0 - y_0 = 0$, gives

$$|\tilde{e}_k| \leqslant N|h|^p \frac{e^{k|h|M} - 1}{M}. \tag{7.2.2.9}$$

Now let $x \in [a, b]$, $x \neq x_0$, be fixed and $h := h_n = (x - x_0)/n$, $n > 0$ an integer. Then $x_n = x_0 + nh = x$, and from (7.2.2.9) with $k = n$, since $\tilde{e}(x; h_n) = \tilde{e}_n$, it follows at once that

$$|\tilde{e}(x; h_n)| \leqslant N|h_n|^p \frac{e^{M|x-x_0|} - 1}{M} \tag{7.2.2.10}$$

for all $x \in [a, b]$ and h_n with $|h_n| \leqslant h_0$. Since $|x - x_0| \leqslant |b - a|$ and $\gamma > 0$, there exists an $\bar{h}$, $0 < \bar{h} \leqslant h_0$, such that $|\tilde{e}(x; h_n)| \leqslant \gamma$ for all $x \in [a, b]$, $|h_n| \leqslant \bar{h}$, i.e., for the one-step method generated by Φ,

$$\eta_0 = y_0,$$
$$\eta_{i+1} = \eta_i + \Phi(x_i, \eta_i; h),$$

we have for $|h| \leqslant \bar{h}$, according to the definition of $\tilde{\Phi}$,

$$\tilde{\eta}_i = \eta_i, \qquad \tilde{e}_i = e_i, \quad \text{and} \quad \tilde{\Phi}(x_i, \tilde{\eta}_i; h) = \Phi(x_i, \eta_i; h).$$

The assertion of the theorem,

$$|e(x; h_n)| \leqslant |h_n|^p N \frac{e^{M|x-x_0|} - 1}{M},$$

thus follows for all $x \in [a, b]$ and all $h_n = (x - x_0)/n$, $n = 1, 2, \ldots$, with $|h_n| \leqslant \bar{h}$. □

From the preceding theorem it follows in particular that methods of order $p > 0$ which in the neighborhood of the exact solution satisfy a Lipschitz condition of the form (7.2.2.4) are convergent in the sense (7.2.2.1). Observe that the condition (7.2.2.4) is fulfilled, e.g., if $(\partial/\partial y)\Phi(x, y; h)$ exists and is continuous in a domain G of the form stated in the theorem.

Theorem (7.2.2.3) also provides an upper bound for the discretization error, which in principle can be evaluated if one knows M and N. One could use it, e.g., to determine the steplength h which is required to compute $y(x)$ within an error ε, given x and $\varepsilon > 0$. Unfortunately, in practice this is doomed by the fact that the constants M and N are not easily accessible, since an estimation of M and N is only possible via estimates of higher derivatives of f. Already in the simple Euler's method, $\Phi(x, y; h) := f(x, y)$, e.g., one has [see (7.2.1.8) f.]

$$N \approx \tfrac{1}{2}|f_x(x, y(x)) + f_y(x, y(x))f(x, y(x))|,$$
$$M \approx \left|\frac{\partial \Phi}{\partial y}\right| = |f_y(x, y)|.$$

For the Runge–Kutta method, one would already have to estimate derivatives of f of the fourth order.

7.2.3 Asymptotic Expansions for the Global Discretization Error of One-Step Methods

It may be conjectured from Theorem (7.2.2.3) that the approximate solution $\eta(x; h)$, furnished by a method of order p, possesses an asymptotic expansion in powers of h of the form

$$\eta(x; h) = y(x) + e_p(x)h^p + e_{p+1}(x)h^{p+1} + \cdots \tag{7.2.3.1}$$

for all $h = h_n = (x - x_0)/n$, $n = 1, 2, \ldots$, with certain coefficient functions $e_i(x)$, $i = p, p + 1, \ldots$, that are independent of h. This is indeed true for general one-step methods of order p, provided only that $\Phi(x, y; h)$ and f satisfy certain additional regularity conditions. One has [see Gragg (1963)]

(7.2.3.2) Theorem. *Let $f(x, y) \in F_{N+2}(a, b)$ [cf. (7.1.3)] and let $\eta(x; h)$ be the approximate solution obtained by a one-step method of order p, $p \leqslant N$, to the solution $y(x)$ of the initial value problem*

$$\text{(I)} \qquad y' = f(x, y), \qquad y(x_0) = y_0, \quad x_0 \in [a, b].$$

Then $\eta(x; h)$ has an asymptotic expansion of the form

$$\eta(x; h) = y(x) + h^p e_p(x) + h^{p+1} e_{p+1}(x) + \cdots + h^N e_N(x) + h^{N+1} E_{N+1}(x; h) \quad \text{with } e_k(x_0) = 0, \quad k = p, p + 1, \ldots \tag{7.2.3.3}$$

which is valid for all $x \in [a, b]$ and all $h = h_n = (x - x_0)/n$, $n = 1, 2, \ldots$. The functions $e_i(x)$ therein are independent of h, and the remainder term $E_{N+1}(x; h)$ is bounded for fixed x and all $h = h_n = (x - x_0)/n$, $n = 1, 2, \ldots$.

Asymptotic laws of the type (7.2.3.1) or (7.2.3.3) are significant in practice for two reasons. In the first place, one can use them to estimate the global discretization error $e(x; h)$. Suppose the method of order p has an asymptotic expansion of the form (7.2.3.1), so that

$$e(x; h) = \eta(x; h) - y(x) = h^p e_p(x) + O(h^{p+1}).$$

Having found the approximate value $\eta(x; h)$ with stepsize h, one computes for the same x, but with another stepsize (say $h/2$), the approximation $\eta(x; h/2)$. For sufficiently small h [and $e_p(x) \neq 0$] one then has in first approximation

$$\eta(x; h) - y(x) \doteq e_p(x) \cdot h^p, \tag{7.2.3.4}$$

$$\eta\left(x; \frac{h}{2}\right) - y(x) \doteq e_p(x) \cdot \left(\frac{h}{2}\right)^p. \tag{7.2.3.5}$$

Subtracting the second equation from the first gives

$$\eta(x; h) - \eta\left(x; \frac{h}{2}\right) \doteq e_p(x) \cdot \left(\frac{h}{2}\right)^p (2^p - 1),$$

$$e_p(x)\left(\frac{h}{2}\right)^p \doteq \frac{\eta(x; h) - \eta(x; h/2)}{2^p - 1},$$

and one obtains, by substitution in (7.2.3.5),

$$\eta\left(x; \frac{h}{2}\right) - y(x) \doteq \frac{\eta(x; h) - \eta(x; h/2)}{2^p - 1}. \tag{7.2.3.6}$$

For the Runge–Kutta method one has $p = 4$ and obtains the frequently used formula

$$\eta\left(x; \frac{h}{2}\right) - y(x) \doteq \frac{\eta(x; h) - \eta(x; h/2)}{15}.$$

The other, more important, significance of asymptotic expansions lies in the fact that they justify the application of extrapolation methods (see Section 3.4). Since a little later [see (7.2.12.7) f.] we will get to know a discretization method for which the asymptotic expansion of $\eta(x; h)$ contains only even powers of h and which, therefore, is more suitable for extrapolation algorithms (see Section 3.5) than Euler's method, we defer the description of extrapolation algorithms to Section 7.2.14.

7.2.4 The Influence of Rounding Errors in One-Step Methods

If a one-step method

$$\begin{aligned} &\eta_0 := y_0; \\ &\text{for } i = 0, 1, 2, \ldots: \\ &\eta_{i+1} := \eta_i + h\Phi(x_i, \eta_i; h), \\ &x_{i+1} := x_i + h \end{aligned} \tag{7.2.4.1}$$

is executed in floating-point arithmetic (t decimal digits) with relative precision $\text{eps} = 5 \times 10^{-t}$, then instead of the η_i one obtains other numbers $\tilde{\eta}_i$, which satisfy a recurrence formula of the form

$$\begin{aligned} &\tilde{\eta}_0 := y_0; \\ &\text{for } i = 0, 1, 2, \ldots: \\ &c_i := \text{fl}(\Phi(x_i, \tilde{\eta}_i; h)), \\ &d_i := \text{fl}(hc_i), \\ &\tilde{\eta}_{i+1} := \text{fl}(\tilde{\eta}_i + d_i) = \tilde{\eta}_i + h\Phi(x_i, \tilde{\eta}_i; h) + \varepsilon_{i+1}, \end{aligned} \tag{7.2.4.2}$$

where the total rounding error ε_{i+1}, in first approximation, is made up of three components:

$$\varepsilon_{i+1} \doteq h\Phi(x_i, \tilde{\eta}_i; h)(\alpha_{i+1} + \mu_{i+1}) + \tilde{\eta}_{i+1}\sigma_{i+1}.$$

Here

$$\alpha_{i+1} = \frac{\text{fl}(\Phi(x_i, \tilde{\eta}_i; h)) - \Phi(x_i, \tilde{\eta}_i; h)}{\Phi(x_i, \tilde{\eta}_i; h)}$$

is the relative rounding error committed in the floating-point computation of Φ, μ_{i+1} the relative rounding error committed in the computation of the product hc_i, and σ_{i+1} the relative rounding error which occurs in the addition $\tilde{\eta}_i + d_i$. Normally, in practice, the stepsize h is so small that $|h\Phi(x_i, \tilde{\eta}_i; h)| \ll |\tilde{\eta}_i|$, and if $|\alpha_{i+1}| \leqslant \text{eps}$ and $|\mu_{i+1}| \leqslant \text{eps}$, one thus has $\varepsilon_{i+1} \doteq \tilde{\eta}_{i+1}\sigma_{i+1}$, i.e., the influence of rounding errors is determined primarily by the addition error σ_{i+1}.

Remark. It is natural, therefore, to reduce the influence of rounding errors by carrying out the addition in double precision ($2t$ decimal places). Denoting by $\text{fl}_2(a + b)$ a double-precision addition, by $\bar{\eta}_i$ a double-precision number ($2t$ decimal places), and by $\bar{\eta}_i := \text{rd}_1(\bar{\eta}_i)$ the number $\bar{\eta}_i$ rounded to single precision, then the algorithm, instead of (7.2.4.2), now runs as follows,

$$\begin{aligned} &\tilde{\eta}_0 := y_0; \\ &\text{for } i = 0, 1, 2, \ldots: \\ &\bar{\eta}_i := \text{rd}_1(\tilde{\eta}_i), \\ &c_i := \text{fl}(\Phi(x_i, \bar{\eta}_i; h)), \\ &d_i := \text{fl}(hc_i), \\ &\tilde{\eta}_{i+1} := \text{fl}_2(\tilde{\eta}_i + d_i). \end{aligned} \tag{7.2.4.3}$$

Let us now briefly estimate the total influence of all rounding errors ε_i. For this, let $y_i = y(x_i)$ be the values of the exact solution of the initial-value problem, $\eta_i = \eta(x_i; h)$ the discrete solutions produced by the one-step method (7.2.4.1) in exact arithmetic, and finally $\tilde{\eta}_i$ the approximate values of η_i actually obtained in t-digit floating-point arithmetic. The latter satisfy relations of the form

$$\begin{aligned} &\tilde{\eta}_0 = y_0; \\ &\text{for } i = 0, 1, 2, \ldots: \\ &\tilde{\eta}_{i+1} = \tilde{\eta}_i + h\Phi(x_i, \tilde{\eta}_i; h) + \varepsilon_{i+1}. \end{aligned} \tag{7.2.4.4}$$

For simplicity, we also assume

$$|\varepsilon_{i+1}| \leqslant \varepsilon \quad \text{for all } i \geqslant 0.$$