II.3 Error Estimation and Convergence for RK Methods

> Es fehlt indessen noch der Beweis dass diese Näherungs-Verfahren convergent sind oder, was practisch wichtiger ist, es fehlt ein Kriterium, um zu ermitteln, wie klein die Schritte gemacht werden müssen, um eine vorgeschriebene Genauigkeit zu erreichen. (Runge 1905)

Since the work of Lagrange (1797) and, above all, of Cauchy, a numerically established result should be accompanied by a reliable error estimation ("... l'erreur commise sera inférieure à ..."). Lagrange gave the well-known error bounds for the Taylor polynomials and Cauchy derived bounds for the error of the Euler polygons (see Section I.7). A couple of years after the first success of the Runge-Kutta methods, Runge (1905) also required error estimates for these methods.

Rigorous Error Bounds

Runge's device for obtaining bounds for the error in one step ("local error") can be described in a few lines (free translation):

"For a method of order p consider the local error

$$e(h) = y(x_0 + h) - y_1 \tag{3.1}$$

and use its Taylor expansion

$$e(h) = e(0) + he'(0) + \ldots + \frac{h^p}{p!} e^{(p)}(\theta h) \tag{3.2}$$

with $0 < \theta < 1$ and $e(0) = e'(0) = \ldots = e^{(p)}(0) = 0$. Now compute explicitly $e^{(p)}(h)$, which will be of the form

$$e^{(p)}(h) = E_1(h) + hE_2(h), \tag{3.3}$$

where $E_1(h)$ and $E_2(h)$ contain partial derivatives of f up to order $p-1$ and p respectively. Further, because of $e^{(p)}(0) = 0$, we have $E_1(0) = 0$. Thus, if all partial derivatives of f up to order p are bounded, we have $E_1(h) = \mathcal{O}(h)$ and $E_2(h) = \mathcal{O}(1)$. So there is a constant C such that $|e^{(p)}(h)| \le Ch$ and

$$|e(h)| \le C\frac{h^{p+1}}{p!}. \text{ "} \tag{3.4}$$

A slightly different approach is adopted by Bieberbach (1923, 1. Abschn., Kap. II, §7), explained in more detail in Bieberbach (1951): we write

$$e(h) = y(x_0+h) - y_1 = y(x_0+h) - y_0 - h\sum_{i=1}^{s} b_i k_i \tag{3.5}$$

and use the Taylor expansions

$$y(x_0+h) = y_0 + y'(x_0)h + y''(x_0)\frac{h^2}{2!} + \ldots + y^{(p+1)}(x_0+\theta h)\frac{h^{p+1}}{(p+1)!}$$

$$k_i(h) = k_i(0) + k_i'(0)h + \ldots + k_i^{(p)}(\theta_i h)\frac{h^p}{p!}, \tag{3.6}$$

where, for vector valued functions, the formula is valid componentwise with possibly different θ's. The first terms in the h expansion of (3.5) vanish because of the order conditions. Thus we obtain

Theorem 3.1. *If the Runge-Kutta method (1.8) is of order p and if all partial derivatives of $f(x,y)$ up to order p exist (and are continuous), then the local error of (1.8) admits the rigorous bound*

$$\begin{aligned}\|y(x_0+h) - y_1\| \le h^{p+1}\Big(&\frac{1}{(p+1)!}\max_{t\in[0,1]}\|y^{(p+1)}(x_0+th)\| \\ &+ \frac{1}{p!}\sum_{i=1}^{s}|b_i|\max_{t\in[0,1]}\|k_i^{(p)}(th)\|\Big)\end{aligned} \tag{3.7}$$

and hence also

$$\|y(x_0+h) - y_1\| \le Ch^{p+1}. \tag{3.8}$$

□

Let us demonstrate this result on Runge's first method (1.4), which is of order $p=2$, applied to a scalar differential equation. Differentiating (1.1) we obtain

$$y^{(3)}(x) = \Big(f_{xx} + 2f_{xy}f + f_{yy}f^2 + f_y(f_x + f_y f)\Big)\big(x, y(x)\big) \tag{3.9}$$

while the second derivative of $k_2(h) = f(x_0+\frac{h}{2}, y_0+\frac{h}{2}f_0)$ is given by

$$k_2^{(2)}(h) = \frac{1}{4}\Big(f_{xx}\big(x_0+\frac{h}{2}, y_0+\frac{h}{2}f_0\big) + 2f_{xy}(\ldots)f_0 + f_{yy}(\ldots)f_0^2\Big) \tag{3.10}$$

(f_0 stands for $f(x_0,y_0)$). Under the assumptions of Theorem 3.1 we see that the expressions (3.9) and (3.10) are bounded by a constant independent of h, which gives (3.8).

The Principal Error Term

For higher order methods rigorous error bounds, like (3.7), become very unpractical. It is therefore much more realistic to consider the first non-zero term in the Taylor expansion of the error. For autonomous systems of equations (2.2), the error term is best obtained by subtracting the Taylor series and using (2.14) and (2.7;q).

Theorem 3.2. *If the Runge-Kutta method is of order p and if f is $(p+1)$-times continuously differentiable, we have*

$$y^J(x_0+h) - y_1^J = \frac{h^{p+1}}{(p+1)!} \sum_{t\in T_{p+1}} \alpha(t)e(t)F^J(t)(y_0) + \mathcal{O}(h^{p+2}) \tag{3.11}$$

where

$$e(t) = 1 - \gamma(t) \sum_{j=1}^{s} b_j \Phi_j(t). \tag{3.12}$$

□

$\gamma(t)$ and $\Phi_j(t)$ are given in Definitions 2.9 and 2.10; see also formulas (2.17) and (2.19). The expressions $e(t)$ are called the *error coefficients.*

Example 3.3. For the two-parameter family of 4th order RK methods (1.17) the error coefficients for the 9 trees of Table 2.2 are ($c_2 = u$, $c_3 = v$):

$$\begin{aligned}
&e(t_{51}) = -\frac{1}{4} + \frac{5}{12}(u+v) - \frac{5}{6}uv, && e(t_{52}) = \frac{5}{12}v - \frac{1}{4},\\
&e(t_{53}) = \frac{5}{8}u - \frac{1}{4}, && e(t_{54}) = -\frac{1}{4},\\
&e(t_{55}) = 1 - \frac{5(b_4 + b_3(3-4v)^2)}{144 b_3 b_4 (1-v)^2}, &&\\
&e(t_{56}) = -4e(t_{51}), && e(t_{57}) = -4e(t_{52}),\\
&e(t_{58}) = -4e(t_{53}), && e(t_{59}) = -4e(t_{54}).
\end{aligned} \tag{3.13}$$

Proof. The last four formulas follow from (1.12). $e(t_{59})$ is trivial, $e(t_{58})$ and $e(t_{57})$ follow from (1.11h). Further

$$e(t_{51}) = 5\int_0^1 t(t-1)(t-u)(t-v)\,dt$$

expresses the quadrature error. For $e(t_{55})$ one best introduces $c_i' = \sum_j a_{ij}c_j$ such that $e(t_{55}) = 1 - 20\sum_i b_i c_i' c_i'$. Then from (1.11d,f) one obtains

$$c_1' = c_2' = 0, \qquad b_3 c_3' = \frac{1}{24(1-v)}, \qquad b_4 c_4' = \frac{3-4v}{24(1-v)}.$$

□

For the classical 4th order method (Table 1.2a) these error coefficients are given by Kutta (1901), p. 448 (see also Lotkin 1951) as follows

$$\Big(-\frac{1}{24}, -\frac{1}{24}, \frac{1}{16}, -\frac{1}{4}, -\frac{2}{3}, \frac{1}{6}, \frac{1}{6}, -\frac{1}{4}, 1\Big)$$

Kutta remarked that for the second method (Table 1.2b) ("Als besser noch erweist sich ...") the error coefficients become

$$\Big(-\frac{1}{54}, \frac{1}{36}, -\frac{1}{24}, -\frac{1}{4}, -\frac{1}{9}, \frac{2}{27}, -\frac{1}{9}, \frac{1}{6}, 1\Big)$$

which, with the exception of the 4th and 9th term, are all smaller than for the above method. A tedious calculation was undertaken by Ralston (1962) (and by many others) to determine optimal coefficients of (1.17). For solutions which minimize the constants (3.13), see Exercise 3 below.

Estimation of the Global Error

Das war auch eine aufregende Zeit ... (P. Henrici 1983)

The global error is the error of the computed solution after *several* steps. Suppose that we have a one-step method which, given an initial value (x_0, y_0) and a step size h, computes a numerical solution y_1 approximating $y(x_0 + h)$. We shall denote this process by Henrici's notation

$$y_1 = y_0 + h\Phi(x_0, y_0, h) \tag{3.14}$$

and call Φ the *increment function* of the method.

The numerical solution for a point $X > x_0$ is then obtained by a step-by-step procedure

$$y_{i+1} = y_i + h_i\Phi(x_i, y_i, h_i), \qquad h_i = x_{i+1} - x_i, \qquad x_N = X \tag{3.15}$$

and our task is to estimate the *global error*

$$E = y(X) - y_N. \tag{3.16}$$

This estimate is found in a simple way, very similar to Cauchy's convergence proof for Theorem 7.3 of Chapter I: *the local errors are transported to the final point x_N and then added up.* This "error transport" can be done in two different ways:

a) either along the exact solution curves (see Fig. 3.1); this method can yield sharp results when sharp estimates of error propagation for the exact solutions are known, e.g., from Theorem 10.6 of Chapter I based on the logarithmic norm $\mu(\partial f/\partial y)$.

b) or along $N - i$ steps of the numerical method (see Fig. 3.2); this is the method used in the proofs of Cauchy (1824) and Runge (1905), it generalizes easily to multistep methods (see Chapter III) and will be an important tool for the existence of asymptotic expansions (see II.8).

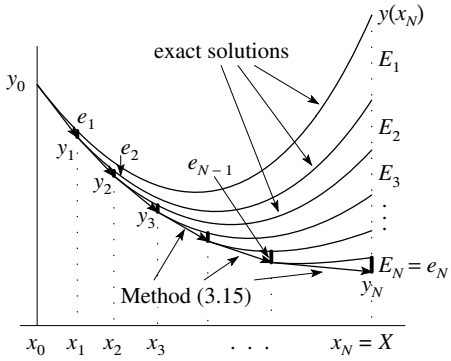

Fig. 3.1. Global error estimation, method (a)

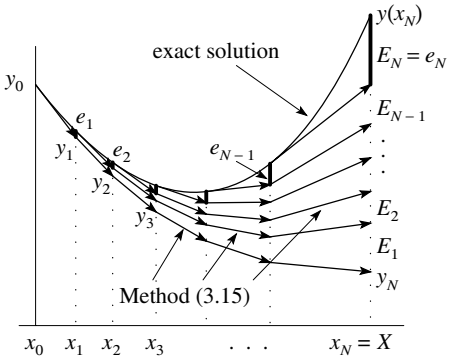

Fig. 3.2. Global error estimation, method (b)

In both cases we first estimate the local errors e_i with the help of Theorem 3.1 to obtain

$$\|e_i\| \le C \cdot h_{i-1}^{p+1}. \tag{3.17}$$

Warning. The e_i of Fig. 3.1 and Fig. 3.2, for $i \neq 1$, are *not* the same, but they allow similar estimates.

We then estimate the transported errors E_i: for method (a) we use the known results from Chapter I, especially Theorem I.10.6, Theorem I.10.2, or formula (I.7.17). The result is

Theorem 3.4. *Let U be a neighbourhood of $\{(x, y(x)) | x_0 \le x \le X\}$ where $y(x)$ is the exact solution of (1.1). Suppose that in U*

$$\left\| \frac{\partial f}{\partial y} \right\| \le L \qquad \text{or} \qquad \mu\Big(\frac{\partial f}{\partial y}\Big) \le L, \tag{3.18}$$

and that the local error estimates (3.17) are valid in U. Then the global error (3.16) can be estimated by

$$\|E\| \le h^p \, \frac{C'}{L} \Big(\exp\big(L(X - x_0)\big) - 1 \Big) \tag{3.19}$$

where $h = \max h_i$,

$$C' = \begin{cases} C & L \ge 0 \\ C \exp(-Lh) & L < 0\,, \end{cases}$$

and h is small enough for the numerical solution to remain in U.

Remark. For $L \to 0$ the estimate (3.19) tends to $h^p \, C \, (x_N - x_0)$.

Proof. From Theorem I.10.2 (with $\varepsilon = 0$) or Theorem I.10.6 (with $\delta = 0$) we obtain

$$\|E_i\| \le \exp\big(L(x_N - x_i)\big)\|e_i\|. \tag{3.20}$$

We then insert this together with (3.17) into

$$\|E\| \le \sum_{i=1}^{N} \|E_i\|.$$

Using $h_{i-1}^{p+1} \le h^p \cdot h_{i-1}$ this leads to

$$\|E\| \le h^p C\Big(h_0 \exp\big(L(x_N - x_1)\big) + h_1 \exp\big(L(x_N - x_2)\big) + \ldots\Big).$$

The expression in large brackets can be bounded by

$$\int_{x_0}^{x_N} \exp(L(x_N - x))dx \qquad \text{for} \qquad L \ge 0 \tag{3.21}$$

$$\int_{x_0}^{x_N} \exp(L(x_N - h - x))dx \qquad \text{for} \qquad L < 0 \tag{3.22}$$

(see Fig. 3.3). This gives (3.19). □

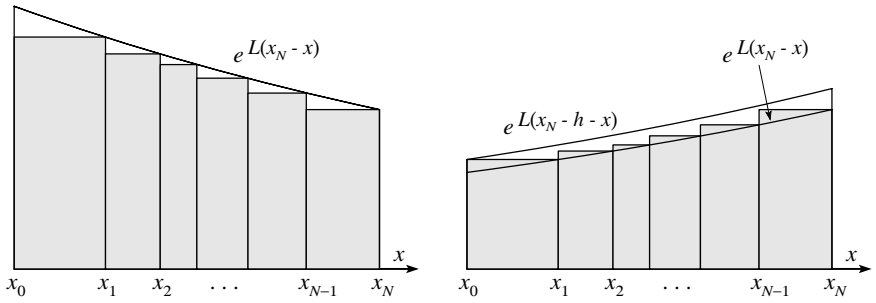

Fig. 3.3. Estimation of Riemann sums

For the second method (b) we need an estimate for $\|z_{i+1} - y_{i+1}\|$ in terms of $\|z_i - y_i\|$, where, besides (3.15),

$$z_{i+1} = z_i + h_i \Phi(x_i, z_i, h_i)$$

is a second pair of numerical solutions. For RK-methods z_{i+1} is defined by

$$\ell_1 = f(x_i, z_i),$$
$$\ell_2 = f(x_i + c_2 h_i, z_i + h_i a_{21} \ell_1), \quad \text{etc.}$$

We now subtract formulas (1.8) from this and obtain

$$\|\ell_1 - k_1\| \le L\|z_i - y_i\|,$$
$$\|\ell_2 - k_2\| \le L(1 + |a_{21}|h_i L)\|z_i - y_i\|, \quad \text{etc.}$$

This leads to the following

Lemma 3.5. *Let L be a Lipschitz constant for f and let $h_i \le h$. Then the increment function Φ of method (1.8) satisfies*

$$\|\Phi(x_i, z_i, h_i) - \Phi(x_i, y_i, h_i)\| \le \Lambda\|z_i - y_i\| \tag{3.23}$$

where

$$\Lambda = L\Big(\sum_i |b_i| + hL\sum_{i,j}|b_i a_{ij}| + h^2L^2\sum_{i,j,k}|b_i a_{ij}a_{jk}| + \dots\Big). \tag{3.24}$$

□

From (3.23) we obtain

$$\|z_{i+1} - y_{i+1}\| \le (1 + h_i\Lambda)\|z_i - y_i\| \le \exp(h_i\Lambda)\|z_i - y_i\| \tag{3.25}$$

and for the errors in Fig. 3.2,

$$\|E_i\| \le \exp\big(\Lambda(x_N - x_i)\big)\|e_i\| \tag{3.26}$$

instead of (3.20). The same proof as for Theorem 3.4 now gives us

Theorem 3.6. *Suppose that the local error satisfies, for initial values on the exact solution,*

$$\|y(x+h) - y(x) - h\Phi(x, y(x), h)\| \le Ch^{p+1}, \tag{3.27}$$

and suppose that in a neighbourhood of the solution the increment function Φ satisfies

$$\|\Phi(x, z, h) - \Phi(x, y, h)\| \le \Lambda\|z - y\|. \tag{3.28}$$

Then the global error (3.16) can be estimated by

$$\|E\| \le h^p\,\frac{C}{\Lambda}\Big(\exp\big(\Lambda(x_N - x_0)\big) - 1\Big) \tag{3.29}$$

where $h = \max h_i$. □

Exercises

1. (Runge 1905). Show that for explicit Runge Kutta methods with $b_i \geq 0$, $a_{ij} \geq 0$ (all i, j) of order s the Lipschitz constant Λ for Φ satisfies

$$1 + h\Lambda < \exp(hL)$$

and that (3.29) is valid with Λ replaced by L.

2. Show that $e(t_{55})$ of (3.13) becomes

$$e(t_{55}) = 1 - 5\,\frac{(4v^2 - 15v + 9) - u(6v^2 - 42v + 27) - u^2(26v - 18)}{12(1-2u)(6uv - 4(u+v) + 3)}$$

after inserting (1.17).

3. Determine u and v in (1.17) such that in (3.13)

a) $\max_{i=5,6,7,8} |e(t_{5i})| = \min$ b) $\sum_{i=1}^{9} |e(t_{5i})| = \min$

c) $\max_{i=5,6,7,8} \alpha(t_{5i})\,|e(t_{5i})| = \min$ d) $\sum_{i=1}^{9} \alpha(t_{5i})\,|e(t_{5i})| = \min$

Results.

a) $u = 0.3587$, $v = 0.6346$, $\min = 0.1033$;
b) $u = 0.3995$, $v = 0.6$, $\min = 1.55$;
c) $u = 0.3501$, $v = 0.5839$, $\min = 0.1248$;
d) $u = 0.3716$, $v = 0.6$, $\min = 2.53$.

Such optimal formulas were first studied by Ralston (1962), Hull & Johnston (1964), and Hull (1967).

4. Apply an explicit Runge-Kutta method to the problem $y' = f(x, y)$, $y(0) = 0$, where

$$f(x,y) = \begin{cases} \dfrac{\lambda}{x}\, y + g(x) & \text{if } x > 0 \\ (1-\lambda)^{-1} g(0) & \text{if } x = 0, \end{cases}$$

$\lambda \leq 0$ and $g(x)$ is sufficiently differentiable (see Exercise 10 of Section I.5).

a) Show that the error after the first step is given by

$$y(h) - y_1 = C_2 h^2 g'(0) + \mathcal{O}(h^3)$$

where C_2 is a constant depending on λ and on the coefficients of the method. Also for high order methods we have in general $C_2 \neq 0$.

b) Compute C_2 for the classical 4th order method (Table 1.2).