4. Einschrittverfahren, insbesondere Runge–Kutta–Verfahren

A. Allgemeines.

Es geht in diesem Abschnitt um die numerische Lösung einer AWA

$$y'(t) \;=\; f(t,y(t)), \qquad y(t_0) \;=\; y_0\,. \tag{4.1}$$

Aufgabe ist es, zu vorgegebenem $t_b \neq t_0$ eine numerische Approximation für die Lösung $y(t_b)$ zu berechnen. O.B.d.A. sei hierbei $t_b > t_0$ und wir setzen voraus, dass f auf einem Gebiet $I \times \mathbb{R}^n$ mit $[t_0, t_b] \subset I$ hinreichend oft stetig differenzierbar ist und die AWA (4.1) auch eine (eindeutig bestimmte) Lösung y besitzt, die im gesamten Intervall $[t_0, t_b]$ erklärt ist. Mitunter wird eine Lipschitz-Konstante der rechten Seite f benötigt. Damit ist stets eine (lokale) Lipschitz-Konstante gemeint, die zu einem kompakten Quader $Q = [t_0, t_b] \times \widetilde{Q}$ gehört, der die Lösung umfasst, d.h. $y(t) \in \widetilde{Q}^0$, für alle $t \in [t_0, t_b]$.

An dieser Stelle ist eine besondere *Warnung* angebracht. In vielen Anwendungen tauchen DGLn auf, deren rechte Seite sich in gewissen Zeitpunkten nichtdifferenzierbar oder sogar unstetig ändert. Sie können beispielsweise von folgender Form sein

$$y'(t) \;=\; \begin{cases} f_1(t,y), & \text{falls } S(y(t)) \leq 0 \\ f_2(t,y), & \text{falls } S(y(t)) > 0. \end{cases} \tag{4.2}$$

Hierbei ist S eine so genannte *Schaltfunktion*. Ein klassisches Beispiel in der Mechanik ist das Phänomen der *trockenen Reibung*, bei der die Richtung der Reibungskraft von der Geschwindigkeitsrichtung abhängt. Die rechte Seite der zugehörigen DGL hängt damit vom Vorzeichen einer Zustandsgröße (abhängige Variable der DGL) ab. Eine solche DGL erfüllt jedenfalls die genannten Voraussetzungen *nicht*, und man hat besondere Vorkehrungen zu treffen, um diese numerisch zu lösen.

Numerische Integratoren arbeiten mit einer *Diskretisierung*, d.h., anstelle einer kontinuierlichen Lösung $y(t)$, $t_0 \leq t \leq t_b$ betrachtet man eine Zerlegung des Integrationsintervalls

$$t_0 \;<\; t_1 \;<\; \ldots \;<\; t_m \;=\; t_b \tag{4.3}$$

und Näherungen $Y_j \approx y(t_j),\ \ j = 0, 1, \ldots, m$.

Die t_j heißen *Integrationsknoten*, $I_h = \{t_0, \ldots, t_m\}$ heißt das *Integrationsgitter*, die $h_j := t_{j+1} - t_j$, $j = 0, \ldots, m-1$ heißen *Schrittweiten*, $\delta_h := \max h_j$ heißt die *Feinheit* des Gitters I_h.

Schließlich lassen sich die Näherungen auch als Funktionswerte einer so genannten *Gitterfunktion* $Y_h : I_h \to \mathbb{R}^n$ interpretieren. Unter einem *Diskretisierungsverfahren* versteht man dann eine Vorschrift, die jeden Gitter I_h eine Gitterfunktion Y_h zuordnet.

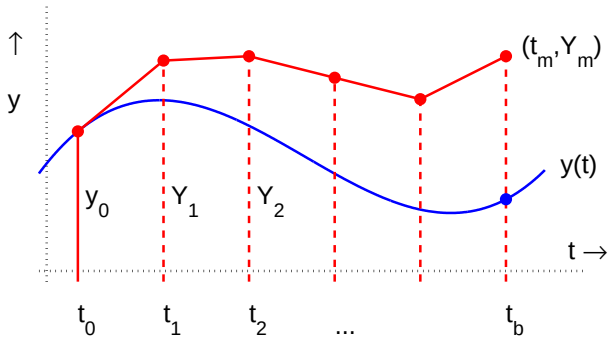

Abb. 4.1. Diskrete Lösung einer DGL

Die numerischen Verfahren zur Integration gewöhnlicher DGLn werden üblicherweise in die Klassen *Einschritt-, Mehrschritt- und Extrapolationsverfahren* unterteilt (wobei nicht immer eine scharfe Abgrenzung dieser Klassen möglich ist).

Einschrittverfahren (ESV) verwenden jeweils nur die zuletzt berechnete Näherung (t_j, Y_j), um hieraus eine nächste Näherung (t_{j+1}, Y_{j+1}) zu bestimmen. Sie haben die allgemeine Form

$$Y_{j+1} \;=\; Y_j + h_j\,\Phi\,(t_j, Y_j, Y_{j+1}; h_j)\,. \tag{4.4}$$

Die Funktion Φ heißt *Verfahrensfunktion* oder *Inkrementfunktion* des konkreten Verfahrens. Sie gibt die „Fortschreiterichtung“ eines Integrationsschrittes wieder. Sie hängt natürlich von der rechten Seite f der DGL ab und wird im Allgemeinen mit Hilfe mehrerer f–Auswertungen berechnet. Ist Φ unabhängig von Y_{j+1}, so definiert (4.4) ein *explizites ESV* – Y_{j+1} kann dann direkt mittels (4.4) ausgewertet werden, andernfalls ist (4.4) ein *implizites Verfahren*.

Mehrschrittverfahren (MSV) verwenden dagegen *mehrere* zuvor berechnete Näherungen $(t_i, Y_i),\ j < i < j+s$, um hieraus eine neue Näherung Y_{j+s} zu berechnen. Zumeist schränkt man sich dabei auf lineare Ansätze in den f_i–Daten (Quadraturformeln) und den Y_i–Daten (Differentiationsformeln) ein. Im Fall äquidistanter Schrittweiten erhält man somit den folgenden allgemeinen Ansatz für ein *lineares Mehrschrittverfahren*:

$$\sum_{i=0}^{s} \alpha_i\, Y_{j+i} \;=\; h \sum_{i=0}^{s} \beta_i\, f_{j+i}\,. \tag{4.5}$$

Dabei sind $\alpha_i,\ \beta_i \in \mathbb{R},\ i = 0,\ldots,s,\ \alpha_s \neq 0,\ t_i := a + ih,\ h := (b-a)/m$ und $f_i := f(t_i, Y_i),\ i = 0, 1, \ldots$.

Ist $\beta_s = 0$, so lässt sich (4.5) nach Y_{j+s} auflösen; man hat dann ein *explizites Verfahren*. Ist dagegen $\beta_s \neq 0$, so ist (4.5) eine implizite Gleichung zur (numerischen) Berechnung von Y_{j+s}, man spricht dann von einem *impliziten Verfahren*.

Extrapolationsverfahren beruhen auf einem, im Allgemeinen einfachen Ein– oder Mehrschrittverfahren. Es werden Näherungen für $y(t_b)$ zu *verschiedenen Schrittweiten* berechnet. Diese Näherungen werden durch Extrapolation bzgl. der Schrittweite verbessert.

In diesem Kapitel wollen wir uns zunächst mit expliziten ESV vom Typ (4.4) beschäftigen. Das einfachste ESV ist das bereits erwähnte *Eulersche Polygonzugverfahren.*

$$Y_{j+1} \;=\; Y_j + h_j\, f(t_j, Y_j)\,. \tag{4.6}$$

Man erhält das Verfahren, wenn man in der DGL $y'(t_j) = f(t_j, Y_j)$ die Ableitung durch den Vorwärts–Differenzenquotienten $(Y_{j+1} - Y_j)/h_j$ ersetzt.

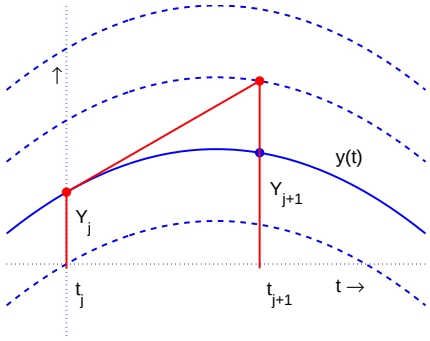

Abb. 4.2. Explizites Euler-Verfahren

Nimmt statt dessen den Rückwärts–Differenzenquotienten $(Y_j - Y_{j-1})/h_{j-1}$, so ergibt sich nach Indexverschiebung das so genannte *implizite Euler-Verfahren*

$$Y_{j+1} \;=\; Y_j + h_j\, f(t_{j+1}, Y_{j+1})\,. \tag{4.7}$$

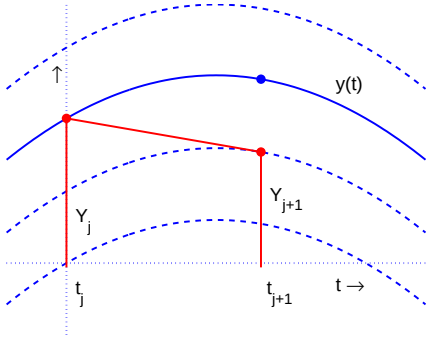

Abb. 4.3. Implizites Euler-Verfahren

Man beachte, dass man auch dieses Verfahren formal als ESV in der Form (4.4) schreiben kann, allerdings ist Y_{j+1} und damit auch $\Phi(t_j, Y_j; h_j)$ implizit durch die Beziehung (4.7) festgelegt. Diese ist im Übrigen eine Fixpunktgleichung, die für hinreichend kleine Integrationsschrittweiten kontrahiert.

Aus der geometrischen Bedeutung der beiden Euler-Verfahrens (die Fortschreiterichtung ist gleich der Tangentenrichtung im linken bzw. rechten Punkt) lassen sich sofort „Verbesserungen“ des Euler-Verfahrens finden:

Wählt man als Fortschreiterichtung etwa den Mittelwert zweier Steigungen, so erhält man das *Verfahren von Heun*:

$$\begin{aligned} k_1 &:= f(t_j, Y_j)\,, \\ k_2 &:= f(t_j + h_j, Y_j + h_j\, k_1)\,, \\ Y_{j+1} &:= Y_j + h_j \left(\frac{1}{2}\, k_1 + \frac{1}{2}\, k_2\right)\,. \end{aligned} \tag{4.8}$$

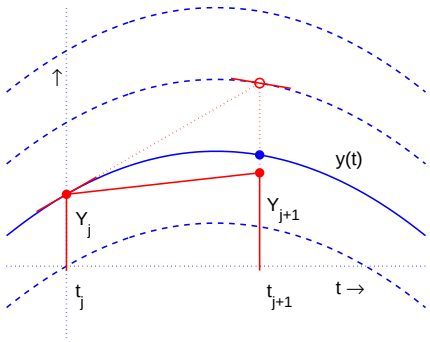

Abb. 4.4. Verfahren von Heun

Alternativ könnte man als Fortschreiterichtung auch eine mittlere Steigung wählen. Ein zugehöriges ESV ist beispielsweise das *modifizierte Euler-Verfahren* (Runge, 1895):

$$\begin{aligned} k_1 &:= f(t_j, Y_j)\,, \\ k_2 &:= f\left(t_j + \tfrac{1}{2}\, h_j, Y_j + h_j \left(\tfrac{1}{2}\, k_1\right)\right)\,, \\ Y_{j+1} &:= Y_j + h_j\, k_2\,. \end{aligned} \tag{4.9}$$

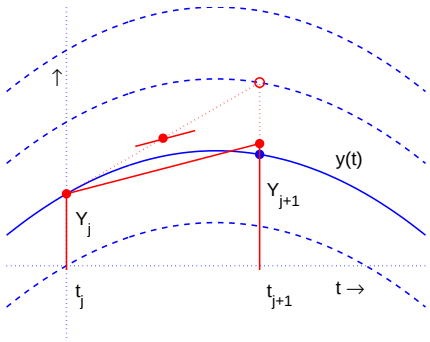

Abb. 4.5. Modifiziertes Euler-Verfahren

Beispiel (4.10) Wir greifen nochmal das Beispiel (3.43) auf und bestimmen numerisch die Lösung der AWA

$$y' \;=\; t^2 + y^2, \quad y(0) \;=\; 1.$$

im Punkt $t_b = 0.95$ mit Hilfe des expliziten Euler-Verfahrens, des Heunschen Verfahrens und des modifizierten Euler-Verfahrens. Wir verwenden zur Integration jeweils eine konstante Schrittweite und bestimmen den *relativen Fehler* der Näherungslösung in t_b. Der Referenzwert für die Lösung ist

$$y(t_b) \approx 0.50471867247946 \times 10^2.$$

Tabelle 4.1: Relative Fehler für Beispiel (4.10).

m	Schrittweite	Euler-Verf.	Heun-Verf.	Modif. Euler-V.
19	0.500D–01	0.82984D+00	0.46801D+00	0.51635D+00
95	0.100D–01	0.59076D+00	0.82046D–01	0.10688D+00
190	0.500D–02	0.44575D+00	0.25811D–01	0.35798D–01
950	0.100D–02	0.15551D+00	0.12034D–02	0.17809D–02
1900	0.500D–03	0.86164D–01	0.30536D–03	0.45585D–03
9500	0.100D–03	0.18896D–01	0.12350D–04	0.18564D–04
19000	0.500D–04	0.95643D–02	0.30915D–05	0.46510D–05
95000	0.100D–04	0.19319D–02	0.12379D–06	0.18636D–06
190000	0.500D–05	0.96718D–03	0.30951D–07	0.46600D–07

B. Konsistenz, Ordnung und Konvergenz.

Die Güte eines ESVs wird durch den so genannten *lokalen Diskretisierungsfehler* gemessen. Dieser gibt an, wie sich für einen einzelnen Integrationsschritt die Fortschreiterichtung $\Phi(t_j, Y_j; h_j)$ von der theoretisch exakten Fortschreiterichtung unterscheidet.

Definition (4.11)

Zu einer aktuellen Näherung $(t_j, Y_j) \in Q$ bezeichne $z(t)$, genauer $z(t; t_j, Y_j)$ die Lösung der *lokalen AWA*

$$z' \;=\; f(t, z(t)), \qquad z(t_j) \;=\; Y_j\,.$$

a) Zu hinreichend kleinem $h > 0$ heißt dann

$$\Delta(t_j, Y_j; h) \quad := \quad \frac{z(t_j + h) - Y_j}{h} \tag{4.12}$$

das *exakte Inkrement* oder *die exakte Fortschreiterichtung*.

b) Die Differenz zwischen exakter und numerischer Fortschreiterichtung

$$\tau(t_j, Y_j; h) \quad := \quad \Delta(t_j, Y_j; h) \;-\; \Phi(t_j, Y_j; h) \tag{4.13}$$

heißt der *lokale Diskretisierungsfehler* des ESVs.

c) Ein ESV heißt *konsistent*, falls für alle hinreichend oft stetig differenzierbaren rechten Seiten f und Näherungen $(t_j, Y_j) \in Q$ eine Abschätzung der folgenden Form gilt:

$$\| \tau(t_j, Y_j; h) \| \quad \leq \quad \sigma(h), \quad \text{mit } \sigma(h) \to 0 \ (h \to 0). \tag{4.14}$$

Hier und im Folgenden sei mit $\| \cdot \|$ die Maximumsnorm im $\mathbb{R}^n$ bezeichnet. Die Konsistenzbedingung fordert also die *gleichmäßige* Konvergenz des lokalen Diskretisierungsfehlers für alle Schrittweitenfolgen $h \to 0$.

d) Wir sagen, ein ESV besitzt die *Konsistenzordnung* $p \in \mathbb{N}$, falls gilt:

$$\| \tau(t_j, Y_j, h) \| \quad \leq \quad \sigma(h), \quad \text{mit } \sigma(h) \ = \ O(h^p), \tag{4.15}$$

d.h., es gibt nur von f und Q abhängige Konstante C, $h_0 > 0$, so dass gilt:
$\forall\, h \in]0, h_0] : \ \| \tau(t_j, Y_j; h) \| \ \leq \ C\, h^p$.

Bemerkungen (4.16)

a) Der lokale Diskretisierungsfehler wird mitunter in der Literatur etwas anders definiert, nämlich durch $\varepsilon := z(t_{j+1}) - Y_{j+1}$; dies entspricht gerade dem lokalen Fehler in den Funktionswerten. Dieser hängt aber auch direkt mit dem Fehler in den Steigungen (unsere Definition) zusammen. Man erhält nämlich durch einfache Umformung mittels (4.4) und (4.12)

$$z(t_{j+1}) - Y_{j+1} \quad = \quad h_j \, \tau(t_j, Y_j; h_j) \, ,$$

d.h., τ ist der (absolute) *Integrationsfehler pro Schrittweite (local error per unit step).*

b) Eine andere Interpretation des lokalen Diskretisierungsfehlers ist die folgende: τ ist das Residuum, welches man erhält, wenn man in der Relation (4.4)

$$\frac{Y_{j+1} - Y_j}{h} \quad = \quad \Phi(t_j, Y_j, h)$$

Y_{j+1} durch die exakte (lokale) Lösung $z(t_{j+1})$ ersetzt.

Zur Bestimmung der Ordnung eines vorgegebenen Einschrittverfahrens vergleicht man die Taylor–Entwicklungen von $\Phi(t, Y; h)$ bezüglich h (Entwicklungspunkt: $h = 0$) mit der entsprechenden Entwicklung von $\Delta(t, Y; h)$.

Für $\Delta(t, Y; h)$ finden wir mittels Taylor–Entwicklung von $z(t + h)$ um $h = 0$

$$\begin{aligned} \Delta(t, Y; h) \ &= \ \frac{z(t+h) - Y}{h} \ = \ \frac{z(t + h; t, Y) - Y}{h} \\ &= \ z'(t) \ + \ \frac{h}{2\,!}\, z''(t) \ + \ \frac{h^2}{3\,!}\, z'''(t) \ + \ \ldots \end{aligned}$$

Hierin verwenden wir nun die vorgegebene DGL $z' = f(t,z)$, die wir, so oft wie benötigt, mittels Kettenregel weiter differenzieren, also

$$\begin{aligned} z''(t) &= f_t(t,z) + f_y(t,z)\cdot f(t,z)\,, \quad z(t)=Y\,, \\ z'''(t) &= f_{t\,t} + 2\,f_{t\,y}\,f + f_{y\,y}\,(f,f) + f_y\,f_t + f_y\,f_y\,f \end{aligned}$$

und so fort. Damit finden wir:

$$\begin{aligned} \Delta &= f + \frac{h}{2}\,(f_t + f_y\,f) \\ &+ \frac{h^2}{6}\,(f_{tt} + 2f_{ty}\,f + f_{yy}\,(f,f) + f_y\,f_t + f_y\,f_y\,f) + O(h^3)\,. \end{aligned} \tag{4.17}$$

Hierbei sind f und sämtliche partiellen Ableitungen von f (außer denen im hier nicht angegebenem Restglied) jeweils im aktuellen Bezugspunkt (t,Y) auszuwerten.

Ferner ist zu beachten, dass es sich sowohl bei f wie bei y um vektorwertige Funktionen handeln kann. Der Term $f_{yy}\,(f,f)$ beispielsweise ist in Koordinaten folgendermaßen zu lesen: $\sum\limits_{k,\ell} \frac{\partial^2 f}{\partial y_k \partial y_\ell} f_k f_\ell$. Analog ist $f_y\,f_y\,f = \sum\limits_{k,\ell} \frac{\partial f}{\partial y_k}\frac{\partial f_k}{\partial y_\ell}\,f_\ell\,.$

Beispiele (4.18)

a) Für das Euler–Verfahren ist $\Phi = f(t,Y) = f$, also:

$$\tau \;=\; \Delta - \Phi \;=\; \frac{h}{2}\,(f_t + f_y\,f) + O(h^2)\,.$$

Das Euler–Verfahren ist also konsistent und hat die Ordnung $p = 1$.

b) Für das Heun–Verfahren erhält man durch Taylor–Entwicklung

$$\begin{aligned} \Phi &= \frac{1}{2}\,f(t,Y) + \frac{1}{2}\,f\,(t+h, Y + h\,f(t,Y)) \\ &= f + h\left\{\frac{1}{2}\,(f_t + f_y f)\right\} + \frac{h^2}{2}\left\{\frac{1}{2}\,(f_{tt} + 2f_{ty}f + f_{yy}(f,f))\right\} + O(h^3)\,. \end{aligned}$$

Zusammen mit (14) ergibt sich

$$\tau \;=\; \Delta - \Phi \;=\; h^2\left\{\frac{1}{6}\,(f_y f_t + f_y f_y f) - \frac{1}{12}\,(f_{tt} + 2f_{ty}f + f_{yy}(f,f))\right\} + O(h^3)\,.$$

Das Heun–Verfahren ist also konsistent und besitzt die Ordnung $p = 2$.

Aufgabe: Berechnen Sie genauso den führenden Term des lokalen Diskretisierungsfehlers für das modifizierte Euler–Verfahren.

Satz (4.19) (Konvergenzsatz)

Die Lösung y der AWA (4.1) existiere im Intervall $t_0 \le t \le t_b$. Ein ESV sei konsistent und besitze die Ordnung p, es gelte also $\| \tau(t, Y; h)\| \le C\, h^p$.
Ferner sei die Verfahrensfunktion Φ des ESVs auf dem Quader Q Lipschitz–stetig bezüglich der Variablen Y:

$$\|\Phi(t, \widetilde{Y}; h) - \Phi(t, Y; h)\| \quad \le \quad L_\Phi \, \|\widetilde{Y} - Y\| \, .$$

für alle (t, Y), $(t, \widetilde{Y}) \in Q$ und hinreichend kleinen Schrittweiten $h > 0$.

Dann liegen alle auf einem hinreichend feinen Gitter I_h berechneten Näherungen (t_j, Y_j) im Quader Q und für die Näherungen $Y_m = Y(t_b; I_h)$ im Endpunkt t_b gelten:

$$\|Y(t_b; I_h) - y(t_b)\| \;\le\; \frac{C}{L_\Phi} \left(\mathrm{e}^{L_\Phi (t_b - t_0)} - 1\right) \cdot h_{\max}^p \, . \tag{4.20}$$

Beweis:

Wir schätzen für einen Integrationsschritt $t_j \to t_{j+1}$ mit Schrittweite $h_j > 0$ ab

$$\begin{aligned}
\|Y_{j+1} - y(t_{j+1})\| &= \| \left(Y_j + h_j \, \Phi(t_j, Y_j; h_j)\right) \, - \, \left(y(t_j) + h_j \, \Delta(t_j, Y_j; h_j)\right) \| \\
&= \|(Y_j - y(t_j)) \, + \, h_j \, \left(\Phi(t_j, Y_j; h_j) - \Phi(t_j, y(t_j); h_j)\right) \, + \\
&\qquad h_j \left(\Phi(t_j, y(t_j); h_j) - \Delta(t_j, y(t_j); h_j)\right) \| \\
&\le (1 + h_j \, L_\Phi) \, \|Y_j - y(t_j)\| \, + \, h_j \, \| \tau(t_j, y(t_j); h_j)\| \\
&\le \mathrm{e}^{L_\Phi (t_{j+1} - t_j)} \, \|Y_j - y(t_j)\| \, + \, h_j \, \| \tau(t_j, y(t_j); h_j)\| \, .
\end{aligned}$$

Setzt man diese Abschätzung nun iterativ ineinander ein und beachtet $Y_0 = y(t_0) = y_0$, so erhält man

$$\|Y_{j+1} - y(t_{j+1})\| \;\le\; \sum_{k=0}^{j} \mathrm{e}^{L_\Phi (t_{j+1} - t_{k+1})} \, h_k \, \| \tau(t_k, y(t_k); h_k)\| \, .$$

Mittels vollständiger Induktion zeigt man nun die Abschätzung (Übungsaufgabe!)

$$\sum_{k=0}^{j} \mathrm{e}^{L_\Phi (t_{j+1} - t_{k+1})} \, h_k \;\le\; \frac{1}{L_\Phi} \left(\mathrm{e}^{L_\Phi (t_{j+1} - t_0)} - 1\right) ,$$

woraus sich zusammen mit der vorausgesetzten Ordungseigenschaft schließlich ergibt:

$$\|Y_{j+1} - y(t_{j+1})\| \;\le\; \frac{C}{L_\Phi} \left(\mathrm{e}^{L_\Phi (t_{j+1} - t_0)} - 1\right) \cdot h_{\max}^p \, .$$

Insbesondere lässt sich also - in Abhängigkeit der Konstanten C, L_Φ und der Integrationslänge $(t_b - t_0)$ - eine Schrittweite $h > 0$ angeben, so dass die diskrete Lösung zu jedem Gitter I_h mit Feinheit $h_{\max} \le h$ ganz in einem ε-Streifen verläuft, der selbst im vorgebenen Quader Q liegt.

Für $j = m - 1$ ergibt sich ferner die gewünschte Abschätzung (4.20). □

Bemerkungen (4.21)

a) Der Satz (4.19) zeigt, dass die (lokale) Konsistenzordung mit der globalen Konvergenzordnung übereinstimmt (*Konsistenzordnung = Konvergenzordnung*). Im Übrigen ist jedes konsistente ESV auch konvergent. Diese Aussage lässt sich, wie wir sehen werden, auf Mehrschrittverfahren nicht übertragen!

b) Aus (4.20) folgt unmittelbar die mitunter nützliche, jedoch etwas schwächere Abschätzung

$$\|Y(t_b; I_h) - y(t_b)\| \;\leq\; C\,(t_b - t_0)\,\mathrm{e}^{L_\Phi (t_b - t_0)} \cdot h_{\max}^p\,. \tag{4.22}$$

c) Für viele ESV impliziert die lokale Lipschitz-Eigenschaft der rechten Seite f auch unmittelbar die (lokale) Lipschitz-Stetigkeit der Verfahrensfunktion.

So folgt beispielsweise für das Verfahren von Heun (4.8) mit $\Phi(t,Y;h) = 0.5\,(f(t,Y) + f(t+h, Y+hf(t,Y)))$ die Abschätzung

$$\|\Phi(t,\widetilde{Y};h) - \Phi(t,Y;h)\| \;\leq\; 0.5\,L\,\|\widetilde{Y} - Y\| \;+\; 0.5\,L\,\|\widetilde{Y}^* - Y^*\|,$$

wobei $\widetilde{Y}^* := \widetilde{Y} + hf(t,\widetilde{Y})$, $Y^* := Y + hf(t,Y)$ und L eine lokale Lipschitzkonstante der rechten Seite f bezeichnen. Damit ergibt sich weiter

$$\|\widetilde{Y}^* - Y^*\| \;\leq\; (1 + h\,L)\,\|\widetilde{Y} - Y\|$$

und somit insgesamt

$$\|\Phi(t,\widetilde{Y};h) - \Phi(t,Y;h)\| \;\leq\; (L \;+\; 0.5\,h\,L^2)\;\|\widetilde{Y} - Y\|\,.$$

$L_\Phi := L + 0.5\,(t_b - t_0)\,L^2$ ist also eine (von h unabhängige) lokale Lipschitz-Konstante der Verfahrensfunktion Φ.

Auf die gleiche Art lässt sich auch für die Verfahrensfunktionen der Runge-Kutta Methoden (vgl. Abschnitt D) die lokale Lipschitz-Stetigkeit zeigen.

d) Man könnte versuchen, aus der Abschätzung (4.20) des Konvergenzsatzes eine optimale, äquidistante Schrittweite zu berechnen. Hierzu würde man fordern

$$\frac{C}{L_\Phi}\,\left(\mathrm{e}^{L_\Phi (t_b - t_0)} - 1\right) \cdot h^p \;=\; \|Y_h\|_\infty \cdot \mathrm{tol},$$

wobei tol (von Toleranz) eine (vom Benutzer) vorzugebende Schranke für den relativen Fehler bezeichnet und $\|Y_h\|_\infty$ die Maximumnorm der Gitterfunktion ist, also $\|Y_h\|_\infty := \max\{\|Y_j\| :\; j = 0, \ldots, m\}$.

Aus dem obigen Ansatz würde man also die folgende Formel für die Schrittweite erhalten:

$$h \;=\; \left[\,\frac{L_\Phi\,\|Y_h\|_\infty\,\mathrm{tol}}{C\,(\mathrm{e}^{L_\Phi (t_b - t_0)} - 1)}\,\right]^{1/p}\,. \tag{4.23}$$

Natürlich ist diese Beziehung für die praktische Wahl der Schrittweite wenig hilfreich, da die Größen C und L_Φ im Allgemeinen kaum abgeschätzt werden können. Sie zeigt jedoch ein Phänomen, dem wir bereits im Beispiel (4.10) begegnet sind: Je größer die Ordnung des Verfahrens ist, desto größere Schrittweiten werden wir im Allgemeinen verwenden können, um eine vorgegebene Genauigkeit zu erreichen. Insbesondere scheint es also numerisch sinnvoll zu sein, Verfahren höherer Ordung zu konstruieren.

C. Rundungsfehler.

Nach der Abschätzung (4.20) des Konvergenzsatzes konvergiert der absolute (aber auch der relative) Fehler der Näherungslösungen $Y(t_b, I_h)$ eines ESVs der Ordnung p für eine Gitterfolge mit $h_{\max} \to 0$ wie $h_{\max}^p$ gegen Null. Dabei sind wir von exakter Rechnung ausgegangen, haben also die bei der Durchführung auf einem Computer auftretenden Rundungsfehler vernachlässigt. Diese können natürlich insbesondere bei kleiner Schrittweite die erreichte Genauigkeit erheblich beeinflussen.

Wir wollen in diesem Abschnitt den Einfluss der Rundungsfehler überschlagsmäßig erfassen, wobei wir von exakter Gleitpunkt-Arithmetik ausgehen, d.h. alle elementaren Operationen $+, -, *, /$ werden mit einer relativen Genauigkeit durchgeführt, die durch eine universelle Konstante, der Maschinengenauigkeit eps, beschränkt ist. Diese liegt bei einfacher Genauigkeit bei eps $\approx 10^{-7}$, bei doppelter Genauigkeit etwa bei eps $\approx 10^{-14}$. Sind a, b Maschinenzahlen und ist $\circ \in \{+, -, *, /\}$, so gilt also für die auf einem Computer berechnete Verknüpfung (fl bezeichnet die in Gleitpunktrechnung ausgeführte Operation)

$$fl(a \circ b) \;=\; (a \circ b)\,(1+\varepsilon), \qquad |\varepsilon| \le \text{eps}.$$

Desweiteren nehmen wir an, dass die Verfahrensfunktion numerisch stabil ausgewertet werden kann, so dass für alle (t, Y, h) gilt

$$fl(\Phi(t, Y; h)) \;=\; \Phi(t, Y; h)\,(1+\alpha), \qquad \|\alpha\| \le K\,\text{eps}$$

mit einer nicht zu großen und von (t, Y, h) unabhängigen Konstanten $K > 0$.

Berücksicht man nun bei der Auswertung eines ESVs (4.4) die Rundungsfehler (einschließlich der Eingangsfehler!), so ergibt sich für die numerisch berechneten Näherungen $\widetilde{Y}_j,\ j = 0, \ldots, m,$ die folgende Rekursion:

$$\begin{aligned}
\widetilde{Y}_0 \quad &= \quad y_0 \,+\, \Delta y_0, \\[2ex]
\widetilde{Y}_{j+1} \quad &= \quad \Big[\widetilde{Y}_j + h_j\,\Phi(t_j, \widetilde{Y}_j; h_j)\,(1+\alpha_j)\,(1+\mu_j)\Big]\,(1+\sigma_j), \\[2ex]
&\text{mit} \qquad \|\,\alpha_j\| \;\le\; K\,\text{eps},\ \ \|\,\mu_j\|,\ \ \|\,\sigma_j\| \;\le\; \text{eps}.
\end{aligned}$$

Die Linearisierung des Fehlers, d.h. die Vernachlässigung aller Terme der Größenordnung eps^2 ergibt

$$\begin{aligned}
\widetilde{Y}_{j+1} \quad &= \quad \widetilde{Y}_j + h_j\,\Phi(t_j, \widetilde{Y}_j; h_j) \,+\, \varepsilon_{j+1} \\[2ex]
\varepsilon_{j+1} \quad &:= \quad \widetilde{Y}_j\,\sigma_j \,+\, h_j\,\Phi(t_j, \widetilde{Y}_j; h_j)\,[(1+\alpha_j)\,(1+\mu_j)\,(1+\sigma_j) \,-\, 1] \qquad (4.24) \\[2ex]
&\approx \quad \widetilde{Y}_j\,\sigma_j \,+\, h_j\,\Phi(t_j, \widetilde{Y}_j; h_j)\,(\alpha_j + \mu_j + \sigma_j).
\end{aligned}$$

Ist, wie wir bereits angenommen hatten, K nicht zu groß und sind die Schrittweiten h_j so klein, dass die Terme $h_j\Phi$ gegenüber $\widetilde{Y}_j$ vernachlässigt werden können, so kann man in erster Näherung abschätzen:

$$\|\,\varepsilon_{j+1}\| \;\le\; \|\widetilde{Y}_h\|_\infty\ \text{eps}. \qquad (4.25)$$

Von der ersten Gleichung in (4.24) wird nun die exakte Rekursion (4.4) subtrahiert. Wir erhalten:

$$\widetilde{Y}_{j+1} - Y_{j+1} \;=\; (\widetilde{Y}_j - Y_j) \;+\; h_j\,(\Phi(t_j, \widetilde{Y}_j; h_j) \;-\; \Phi(t_j, Y_j; h_j)) \;+\; \varepsilon_{j+1}$$

und damit – mit Hilfe der Lipschitz-Bedingung für Φ:

$$\begin{aligned} \|\widetilde{Y}_{j+1} - Y_{j+1}\| \;&\leq\; (1 + h_j\,L_\Phi)\,\|\widetilde{Y}_j - Y_j\| \;+\; \|\varepsilon_{j+1}\| \\ &\leq\; \mathrm{e}^{L_\Phi(t_{j+1}-t_j)}\,\|\widetilde{Y}_j - Y_j\| \;+\; \|\varepsilon_{j+1}\|. \end{aligned}$$

Diese Abschätzung ist genau von der Art, wie wir sie im Beweis des Konvergenzsatzes (4.19) kennengelernt haben. Mit gleicher Technik (ineinander einsetzen!) folgt daher

$$\begin{aligned} \|\widetilde{Y}_{j+1} - Y_{j+1}\| \;&\leq\; \mathrm{e}^{L_\Phi(t_{j+1}-t_0)}\,\|\Delta y_0\| \;+\; \sum_{k=0}^{j} \mathrm{e}^{L_\Phi(t_{j+1}-t_{k+1})}\,\|\,\varepsilon_{k+1}\| \\ &\leq\; \mathrm{e}^{L_\Phi(t_{j+1}-t_0)}\,\|\Delta y_0\| \;+\; \frac{1}{L_\Phi}\,\left(\mathrm{e}^{L_\Phi(t_{j+1}-t_0)} - 1\right) \cdot \max_k \frac{\|\varepsilon_{k+1}\|}{h_k}\,. \end{aligned}$$

Zur Abschätzung des tatsächlichen Fehlers $\|\widetilde{Y}_{j+1} - y(t_{j+1})\|$ verwenden wir nun noch die Beziehung (4.20) sowie die Dreieckungleichung. Wir fassen das Ergebnis im folgenden Satz zusammen:

Satz (4.26) (Rundungsfehler bei ESV)

Die Lösung y der AWA (4.1) existiere im Intervall $t_0 \leq t \leq t_b$. Ein ESV (4.4) sei konsistent und besitze die Ordnung p, es gelte also $\|\,\tau(t, Y; h)\| \leq C\,h^p$. Ferner sei die Verfahrensfunktion Φ auf dem Quader Q Lipschitz–stetig bezüglich der Variablen Y mit der Lipschitz-Konstanten L_Φ .

Liegen dann die auf einem hinreichend feinen Gitter I_h numerisch berechneten Näherungen $(t_j, \widetilde{Y}_j)$ im Quader Q, so gilt für die Näherung $\widetilde{Y}_m = \widetilde{Y}(t_b; I_h)$ im Endpunkt t_b die Abschätzung:

$$\begin{aligned} \|\widetilde{Y}(t_b; I_h) - y(t_b)\| \;&\leq\; \mathrm{e}^{L_\Phi(t_b-t_0)}\,\|\Delta y_0\| \;+ \\ &+\; \frac{1}{L_\Phi}\,\left(\mathrm{e}^{L_\Phi(t_b-t_0)} - 1\right) \cdot \left(C\,h_{\max}^p \;+\; \max_j \frac{\|\varepsilon_{j+1}\|}{h_j}\right)\,. \end{aligned} \tag{4.27}$$

In der Beziehung (4.27) beschreibt Δy_0 den absoluten Fehler im Anfangswert (*Einlesefehler*), der Term mit $C\,h_{\max}^p$ beschreibt den *Diskretisierungsfehler* und schließlich der Term mit $\|\varepsilon_{j+1}\|/h_j$ den *Rundungsfehlereinfluß*.

Unter den bei (4.25) genannten Voraussetzungen lässt sich der Ausdruck in (4.27) nochmals vereinfachen zu

$$\begin{aligned} \|\widetilde{Y}(t_b; I_h) - y(t_b)\| \;&\leq\; \mathrm{e}^{L_\Phi(t_b-t_0)}\,\|\Delta y_0\| \;+ \\ &+\; \frac{1}{L_\Phi}\,\left(\mathrm{e}^{L_\Phi(t_b-t_0)} - 1\right) \cdot \left(C\,h_{\max}^p \;+\; \frac{\|\widetilde{Y}_h\|_\infty\,\mathrm{eps}}{h_{\min}}\right)\,. \end{aligned} \tag{4.28}$$

Für den Fall äquidistanter Schrittweite (also $h_{\min} = h_{\max}$) sind die Ausdrücke in der rechten Klammer von (4.28) in der Abbildung 4.6 qualitativ wiedergegeben.

Man erkennt, dass es bei vorgegebener Maschinengenauigkeit eine Grenzgenauigkeit und eine zugehörige optimale Schrittweite gibt, bis zu der der Gesamtfehler des numerischen Ergebnisses fällt. Bei weiterer Verkleinerung der Schrittweite wächst jedoch der Fehler dann aufgrund der Rundungsfehler wieder an.

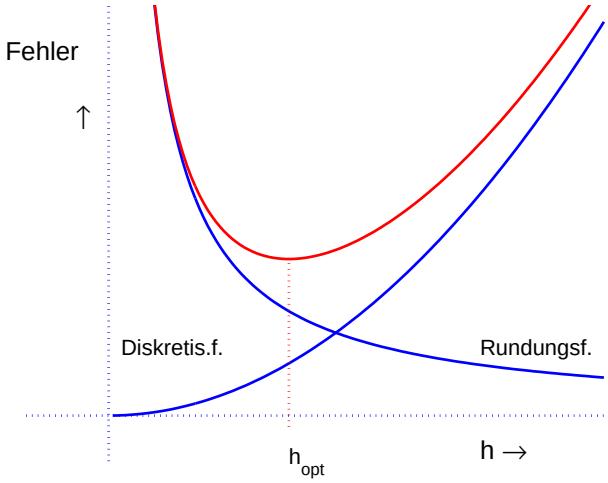

Abb. 4.6. Gesamtfehler bei ESV

D. Runge–Kutta–Verfahren.

Die meistgebräuchlichen Einschrittverfahren sind die so genannten Runge–Kutta–Verfahren (kurz: RK–Verfahren) benannt nach Carl Runge (1856–1927) und Martin Wilhelm Kutta (1867–1927). Es handelt sich dabei um ESV, die die Verfahrensfunktion als Linearkombination von Auswertungen der rechten Seite f ansetzen. Insoweit sind dies direkte Verallgemeinerungen des Heunschen Verfahrens (4.8) bzw. des modifizierten Euler-Verfahrens (4.9). Erste Verfahren dieser Art wurden von Runge (1895), Heun (1900) und Kutta (1901) angegeben. Letzterer gab auch *das klassische Runge-Kutta Verfahren* vierter Ordnung an. Erst fünfzig Jahre später bemühte man sich um die Konstruktion von RK–Verfahren höherer Ordnung.

Die allgemeine Form eines (expliziten) RK–Verfahrens lautet:

$$\begin{aligned} Y_{j+1} &= Y_j + h_j \sum_{i=1}^{s} b_i \, k_i(t_j, Y_j; h_j) \\ k_1(t, Y; h) &= f(t, Y) \\ k_i(t, Y; h) &= f(t + c_i h, Y + h \sum_{\ell=1}^{i-1} a_{i\ell} \, k_\ell \, (t, Y; h)) \, , \end{aligned} \tag{4.29}$$

Dabei heißt s die *Stufenzahl* des RK–Verfahrens, die c_i heißen *Knoten*, die b_i *Gewichte* und die (a_{ij}) werden zu einer *Verfahrensmatrix* zusammengefasst. Alle Koefizienten b_i, c_i und $a_{i\ell}$, welche ja das konkrete Verfahren festlegen, werden üblicherweise in einem Tableau, dem so genannten *Butcher–Schema* angeordnet:

Tabelle 4.2: Allgemeines Butcher–Schema.

0					
c_2	a_{21}				
c_3	a_{31}	a_{32}			
$\vdots$	$\vdots$	$\vdots$	$\ddots$		
c_s	a_{s1}	a_{s2}	$\dots$	$a_{s,s-1}$	
	b_1	b_2	$\dots$	b_{s-1}	b_s

Beispiele hatten wir schon kennengelernt. So sind die Schemata für das Heun-Verfahren (4.8) bzw. für das modifizierte Euler-Verfahren (4.9) (für beide ist $s = 2$) in der Tabelle 4.3 angegeben.

Tabelle 4.3: Heun–Verfahren (p=2) und Modifiziertes Euler–Verfahren (p=2)

0		
1	1	
	$1/2$	$1/2$

0		
$1/2$	$1/2$	
	0	1

Für die so genannte *Kutta–Regel* und ein weiteres auf Heun zurückgehendes dreistufiges Verfahren (beide mit $s = p = 3$) sind die Schemata der Tabelle 4.4 zu entnehmen.

Tabelle 4.4: Kutta–Regel (p=3) und Heun–Verfahren (p=3)

0			
$1/2$	$1/2$		
1	-1	2	
	$1/6$	$2/3$	$1/6$

0			
$1/3$	$1/3$		
$2/3$	0	$2/3$	
	$1/4$	0	$3/4$

Zwei Beispiele vierter Ordnung gehen auf Kutta zurück. Dies ist zum Einen *das klassische Runge-Kutta Verfahren* RK4 und zum Andern die so genannte 3/8–Regel. Die Schemata sind in Tabelle 4.5 angegeben.

In allen bisher angegebenen Beispielen stimmt jeweils die Stufenzahl s mit der Ordnung p des Verfahrens überein. Dies ist allerding für Verfahren höherer Ordnung nicht mehr der Fall. So werden für ein RK–Verfahren der Ordung sieben bereits neun, für ein Verfahren der Ordnung acht sogar elf Stufen benötigt (Butcher, 1987).

Tabelle 4.5: Klassisches RK4–Verfahren und 3/8–Regel (p=4)

0				
1/2	1/2			
1/2	0	1/2		
1	0	0	1	
	1/6	1/3	1/3	1/6

0				
1/3	1/3			
2/3	−1/3	1		
1	1	−1	1	
	1/8	3/8	3/8	1/8

Zur Konstruktion von RK–Verfahren hat man für die allgemeine Verfahrensfunktion nach (4.29):

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

einen Taylor–Abgleich mit dem exakten Inkrement $\Delta(t, Y; h)$ gemäß (4.12) und (4.17) durchzuführen.

Durch Abgleich des Absolutterms ($h = 0$) findet man beispielsweise sofort:

Satz (4.30)

Ein RK–Verfahren ist genau dann konsistent, falls $\sum\limits_{i=1}^{s} b_i \;=\; 1$.

Zumeist schränkt man sich bei der Aufstellung der Ordnungsbedingungen auf den Fall *autonomer* DGLn ein. Dies ist, wie der folgende Satz zeigt, gerechtfertigt, wenn die so genannte *Knotenbedingung* erfüllt ist:

$$c_i \;=\; \sum_{j=1}^{i-1} a_{ij} \;, \quad i = 2, \ldots s \;. \tag{4.31}$$

Satz (4.32)

Hat ein RK–Verfahren die Ordnung p für alle autonomen DGLn und ist die Knotenbedingung (4.31) erfüllt, so ist das Verfahren auch von gleicher Ordnung für nichtautonome DGLn.

Beweis: Eine nichtautonome AWA

$$z'(t) \;=\; g(t, z(t)), \quad z(t_0) \;=\; z_0$$

lässt sich wie folgt in ein autonomes Problem $y' = f(y)$, $y(t_0) = y_0$ transformieren:

$$y(t) := \begin{pmatrix} t \\ z(t) \end{pmatrix}, \quad f(y) := \begin{pmatrix} 1 \\ g(y_1, y_2) \end{pmatrix}, \quad y_0 := \begin{pmatrix} t_0 \\ z_0 \end{pmatrix}.$$

Das RK–Verfahren für dieses Problem lautet

$$Y_{j+1} \;=\; Y_j \;+\; h_j \sum_{i=1}^{s} b_i\, k_i\,, \qquad k_i \;=\; f(Y_j + h_j \sum_{\ell=1}^{i-1} a_{i\ell}\, k_\ell) \tag{4.33}$$

Wir schreiben dies wieder in Koordinaten mit $Y = (t, Z)^{\mathrm{T}}$ und $k_i = (1, \widetilde{k}_i)^{\mathrm{T}}$ und erhalten

$$\begin{aligned} t_{j+1} &= t_j \;+\; h_j \sum_{i=1}^{s} b_i\,, \\ Z_{j+1} &= Z_j \;+\; h_j \sum_{i=1}^{s} b_i\, \widetilde{k}_i\,, \\ \widetilde{k}_i &= g(t_j + h_j \sum_{\ell=1}^{i-1} a_{i\ell}\,,\; Z_j + h_j \sum_{\ell=1}^{i-1} a_{i\ell}\, \widetilde{k}_\ell)\,. \end{aligned}$$

Wegen der Konsistenz (4.30) des Verfahrens und der vorausgesetzten Knotenbedingung (4.31) ist dies aber genau das RK–Verfahren für das nichtautonome Anfangswertproblem und dieses hat demnach die gleiche Konsistenzordnung wie (4.33). □

Die in den Tabellen 4.3–4.5 angegebenen RK-Verfahren erfüllen alle die Knotenbedingung, es genügt dort also zur Ordnungsbestimmung, sich auf autonome Differentialgleichungen einzuschränken.

Die Einschränkung auf autonome Differentialgleichungen bedeutet eine erhebliche Vereinfachung für das Aufstellen der Ordnungsbedingungen.

Beispiel (4.34) Will man beispielsweise die Ordnungsbedingungen für ein dreistufiges RK-Verfahren der Ordnung $p = 3$ aufstellen, so hat man die Funktionen

$$\begin{aligned} \Phi &= b_1\, k_1 + b_2\, k_3 + b_3\, k_3, \\ k_1 &= f(Y), \quad k_2 \;=\; f(Y + h\, a_{21}\, k_1), \\ k_3 &= f(Y + h\,(a_{31}\, k_1 + a_{32}\, k_2)) \end{aligned}$$

bzgl. der Schrittweite h in eine Taylor-Reihe zu entwickeln und diese bis zur Potenz h^2 mit der der exakten Inkrementfunktion

$$\Delta \;=\; f + \frac{h}{2}\,(f'f) + \frac{h^2}{6}\,(f''(f,f) + f'f'f) + O(h^3)\,,$$

abzugleichen. Wir schreiben hier f' anstelle von f_y, vgl. auch (4.17).

Die Terme f, $f'f$, $f''(f,f)$ und $f'f'f$ heißen *elementare Differentiale*. Sie treten ganz analog bei der Taylor–Entwicklung der Verfahrensfunktion auf. Man erhält

$$\begin{aligned} \Phi \;=\; & (\textstyle\sum b_i)\; f \;+\; h\,[b_2\, a_{21} + b_3(a_{31} + a_{32})]\; f'\, f \\ & +\; \frac{h^2}{2}\,[(b_2\, a_{21}^2 \;+\; b_3\,(a_{31} + a_{32})^2)\; f''(f,f) \;+\; 2\, b_3\, a_{21}\, a_{32}\; f'\, f'\, f] \;+\; O(h^3). \end{aligned}$$

Nun sind die elementaren Differentiale (bei hinreichend großer Dimension n) linear unabhängig (vgl. Lemma 4.25 in Deuflhard, Bornemann (2002)), so dass die Entwicklungen von Δ und Φ nicht nur bzgl. der h-Potenzen, sondern auch bzgl. der elementaren Differentiale übereinstimmen müssen.

Mittels Koeffizientenvergleichs erhalten wir damit die folgenden vier Ordnungsgleichungen für die sechs Unbekannten $b_i,\ a_{ik}$:

$$\begin{aligned} b_1 \ + \ b_2 \ + \ b_3 \ &= \ 1 \\ b_2\, a_{21} \ + \ b_3\, (a_{31} + a_{32}) \ &= \ 1/2 \\ b_2\, a_{21}^2 \ + \ b_3\, (a_{31} + a_{32})^2 \ &= \ 1/3 \\ b_3\, a_{21}\, a_{32} \ &= \ 1/6. \end{aligned} \tag{4.35}$$

Ein dreistufiges RK-Verfahren besitzt also genau dann die Konsistenzordnung $p = 3$, wenn das obige Gleichungssystem erfüllt ist. Man überzeugt sich unmittelbar, dass die in Tabelle 4.4 angegebenen dreistufigen RK-Verfahren dieses Gleichungsystem lösen, und damit die Konsistenzordnung $p = 3$ besitzen. Die Parameter c_i sind jeweils durch die Knotenbedingung (4.31) festgelegt.

E. Ordnungsbedingungen nach Butcher.

J.C. Butcher (1963) hat zur Aufstellung der Ordnungsgleichungen ein relativ einfaches graphentheoretisches Verfahren angegeben. Will man die Gleichungen dafür aufstellen, dass ein s-stufiges RK–Verfahren die Ordnung $\geq$ p besitzt, so hat man sämtliche, paarweise nicht isomorphen Wurzelbäume mit höchstens p Knoten aufzustellen. Diese Wurzelbäume entsprechen genau den elementaren Differentialen in den Taylor-Entwicklungen von Δ und Φ.

Wir benötigen einige Grundbegriffe aus der Graphentheorie, die wir zunächst hier zusammenstellen wollen. Ein *Graph* $\mathbf{g} = (P, K, v)$ besteht aus einer endlichen Menge $P = \{x_1, \ldots, x_q\}$ von Knoten (Punkte), einer endlichen Menge K von Kanten und einer Abbildung v mit

(i) für *ungerichtete Graphen*: $v : K \to \wp(P),\ \ v(k) = \{a, b\}$ ($a = b$ ist zugelassen).

(ii) für *gerichtete Graphen*: $v : K \to P \times P,\ \ v(k) = (a, b)$. In diesem Fall heißt $v_A(k) := a$ der *Anfangspunkt* und $v_E(k) := b$ der *Endpunkt* der Kante k.

Aus jedem gerichteten Graphen lässt sich natürlich durch Vergessen der Richtungen ein ungerichteter Graph machen.

Mit $\sharp\,\mathbf{g}$ wird die Knotenzahl des Graphen $\mathbf{g}$ bezeichnet.

Zwei Graphen $\mathbf{g}_1 = (P_1, K_1, v_1)$ und $\mathbf{g}_2 = (P_2, K_2, v_2)$ heißen *isomorph*, falls es Bijektionen $\phi : P_1 \to P_2$ und $\psi : K_1 \to K_2$ mit der folgenden Eigenschaft gibt: Für jede Kante $k \in K_1$ mit $v_1(k) = \{a, b\}$ bzw. $v_1(k) = (a, b)$ ist $v_2(\psi(k)) = \{\phi(a), \phi(b)\}$ bzw. $v_2(\psi(k)) = (\phi(a), \phi(b))$.

Für einen gerichteten Graphen $\mathbf{g}$ und $x \in P$ heißt $g^-(x) := \sharp\{k : v_E(k) = x\}$ der *negative*

Grad (Zahl der einlaufenden Kanten) und $g^+(x) := \sharp\{k : v_A(k) = x\}$ der *positive Grad* des Knotens x (Zahl der auslaufenden Kanten). $g(x) := g^-(x) + g^+(x)$ heißt der *Grad* von x.

Eine endliche Folge $\omega = \langle x_1, k_1, x_2, \ldots, x_{m-1}, k_{m-1}, x_m \rangle$ von Knoten und Kanten heißt ein *ungerichteter* bzw. *gerichteter Kantenzug*, falls $v(k_i) = \{x_i, x_{i+1}\}$ bzw. $v(k_i) = (x_i, x_{i+1})$ für alle $i = 1, \ldots, m-1$. Wir sagen auch, der Kantenzug ω verbindet x_1 und x_m. Im Fall $x_1 = x_m$ heißt der Kantenzug ein *Kreis*.

Ein Graph $\mathbf{g}$ heißt *zusammenhängend*, falls je zwei (verschiedene) Knoten durch einen Kantenzug verbunden werden können.

Ein ungerichteter, zusammenhängender Graph ohne Kreise heißt ein *Baum*. Schließlich heißt ein gerichteter Graph ein *Wurzelbaum*, falls er als ungerichteter Graph ein Baum ist, und es einen ausgezeichneten Knoten $x_1 \in P$ gibt, die *Wurzel*, mit der Eigenschaft:

$$g^-(x_1) \;=\; 0 \quad \text{und} \quad \forall\, x \neq x_1 : \;\; g^-(x) \;=\; 1\,.$$

Mit dieser Eigenschaft ist klar, dass ein Wurzelbaum mit q Knoten genau $q-1$ Kanten besitzt.

Ist $\mathbf{g}$ ein Wurzelbaum und $x \in P$ ein Knoten von $\mathbf{g}$, so erzeugt x einen Teil-Wurzelbaum von $\mathbf{g}$ mit der Wurzel x, der aus allen Knoten von $\mathbf{g}$ besteht, die durch einen gerichteten Kantenzug – von x ausgehend – erreicht werden können, zusammen mit den zugehörigen Kanten. Dieser von x erzeugte Teil-Wurzelbaum werde mit $[x]$ bezeichnet.

In der Abbildung 4.7 sind sämtliche paarweise nicht isomorphe Wurzelbäume mit maximal vier Knoten aufgezeichnet. Die Richtung der Kanten weist dabei stets von unten nach obe

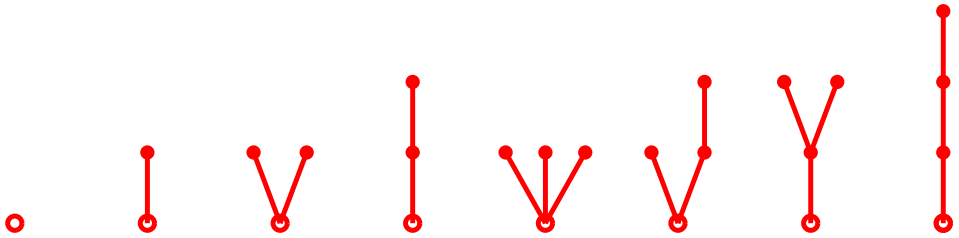

Abb. 4.7. Alle Wurzelbäume mit bis zu vier Knoten

Den Wurzelbäumen entsprechen in eineindeutiger Weise die elementaren Differentiale.

So gehören die in Abbildung 4.7 dargestellten Wurzelbäume (in dieser Reihenfolge) zu den folgenden elementaren Differentialen f, $f'\,f$, $f''(f,f)$, $f'\,f'\,f$, $f'''\,(f,f,f)$, $f''\,(f,f'\,f)$, $f'\,f''\,(f,f)$ und $f'\,f'\,f'\,f$.

Es sei nun $\mathbf{g}$ ein Wurzelbaum mit der Knotenmenge $P = \{x_1, \ldots, x_q\}$. x_1 sei die Wurzel von $\mathbf{g}$. Zu einem Knoten $x_j \in P$ bezeichne J_j die Indizes der Nachfolgerknoten, also

$$J_j \;:=\; \{\ell : \;\; \exists\, k \in K : v(k) = (x_j, x_\ell)\,\}$$

Wir ordnen dem Wurzelbaum $\mathbf{g}$ nun den folgenden polynomialen Ausdruck in den RK–Koeffizienten $b_i\, a_{jk}$ zu:

$$p(\mathbf{g}) \;:=\; \sum_{i_1,\ldots,i_q=1}^{s} b_{i_1} \prod_{j=1}^{q} \left(\prod_{\ell \in J_j} a_{i_j\, i_\ell} \right) . \tag{4.36}$$

Hierbei ist zu beachten, dass, wie üblich, leere Produkte ($J_j = \emptyset$) Eins gesetzt werden. Ferner sind alle Koeffizienten a_{jk} mit $k \geq j$, die also im expliziten Butcher-Schema nicht auftreten, Null zu setzen.

Sodann wird dem Wurzelbaum $\mathbf{g}$ noch eine natürliche Zahl $\gamma(\mathbf{g})$ zugeordnet, nämlich:

$$\gamma(\mathbf{g}) \;:=\; \prod_{i=1}^{q} \sharp\,[x_i] \,. \tag{4.37}$$

Mit diesen Größen lässt sich nun der folgende Satz von Butcher (1963) formulieren:

Satz (4.38) (Ordungsbedingungen für RK–Verfahren)

Ein s-stufiges RK-Verfahren mit Koeffizienten (b_i, a_{jk}) und der Knotenbedingung (4.31) besitzt genau dann die Konsistenzordnung p, wenn für alle Wurzelbäume $\mathbf{g}$ mit maximal p Knoten gilt: $p(\mathbf{g}) = 1/\gamma(\mathbf{g})$.

Beweise dieses Satzes findet man u.a. in den Lehrbüchern von Hairer, Norsett und Wanner, von Strehmel und Weiner, sowie von Deuflhard und Bornemann.

Wir wollen uns die Aussage des Butcherschen Satzes für den Fall eines RK–Verfahren der Ordnung vier ansehen. Hierzu sind genau die Wurzelbäume aus Abbildung 4.7 zu betrachten. Diese werden im Folgenden mit $\mathbf{g}_1, \ldots \mathbf{g}_8$ bezeichnet.

Es ergeben sich damit die folgenden acht Ordnungsbedingungen:

$$\begin{aligned}
p(\mathbf{g}_1) &= \sum_{i_1=1}^{s} b_{i_1} = \sum_i b_i = && \gamma(\mathbf{g}_1)^{-1} = 1,\\
p(\mathbf{g}_2) &= \sum_{i_1,i_2=1}^{s} b_{i_1} a_{i_1 i_2} = \sum_i b_i\, c_i = && \gamma(\mathbf{g}_2)^{-1} = 1/2,\\
p(\mathbf{g}_3) &= \sum_{i_1,i_2,i_3=1}^{s} b_{i_1} a_{i_1 i_2} a_{i_1 i_3} = \sum_i b_i\, c_i^2 = && \gamma(\mathbf{g}_3)^{-1} = 1/3,\\
p(\mathbf{g}_4) &= \sum_{i_1,i_2,i_3=1}^{s} b_{i_1} a_{i_1 i_2} a_{i_2 i_3} = \sum_{i,j} b_i\, a_{ij}\, c_j = && \gamma(\mathbf{g}_4)^{-1} = 1/6,\\
p(\mathbf{g}_5) &= \sum_{i_1,i_2,i_3,i_4=1}^{s} b_{i_1} a_{i_1 i_2} a_{i_1 i_3} a_{i_1 i_4} = \sum_i b_i\, c_i^3 = && \gamma(\mathbf{g}_5)^{-1} = 1/4,\\
p(\mathbf{g}_6) &= \sum_{i_1,i_2,i_3,i_4=1}^{s} b_{i_1} a_{i_1 i_2} a_{i_1 i_3} a_{i_3 i_4} = \sum_{i,j} b_i\, c_i\, a_{ij}\, c_j = && \gamma(\mathbf{g}_6)^{-1} = 1/8,
\end{aligned}$$

$$p(\mathbf{g}_7) \;=\; \sum_{i_1,i_2,i_3,i_4=1}^{s} b_{i_1} a_{i_1 i_2} a_{i_2 i_3} a_{i_2 i_4} \;=\; \sum_{i,j} b_i\, a_{ij}\, c_j^2 \;=\; \gamma(\mathbf{g}_7)^{-1} \;=\; 1/12,$$

$$p(\mathbf{g}_8) \;=\; \sum_{i_1,i_2,i_3,i_4=1}^{s} b_{i_1} a_{i_1 i_2} a_{i_2 i_3} a_{i_3 i_4} \;=\; \sum_{i,j,k} b_i\, a_{ij}\, a_{jk}\, c_k \;=\; \gamma(\mathbf{g}_8)^{-1} \;=\; 1/24.$$

Man kann sich nun wiederum davon überzeugen, dass die in Tabelle 4.5 angegebenen vierstufigen RK–Verfahren tatsächlich diese Ordnungsgleichungen erfüllen.

Die Anzahl der Ordnungsgleichungen nimmt mit wachsender Ordnung p stark zu. So hat man für ein Verfahren der Ordnung sieben schon 85 nichtlineare Gleichungen zu lösen, wozu übrigens ein wenigstens neunstufiges RK–Verfahren benötigt wird. Für ein Verfahren der Ordnung zehn sind es sogar 1205 nichtlineare Gleichungen und man benötigt wenigstens 13 Stufen.

F. Schrittweitensteuerung.

Hierbei geht es um die automatische Generierung eines Integrationsgittes I_h, das einerseits fein genug sein soll, um eine vorgegebene Genauigkeit der numerischen Lösung zu garantieren, andererseits aber auch nicht feiner, um den numerischen Aufwand (dieser schließt auch die Gittererzeugung selbst ein) und den Einfluss von Rundungsfehlern möglichst gering zu halten.

Die Schrittweitensteuerung wird dabei ein *lokaler* Prozess sein, also im Allgemeinen eine *nicht äquidistante* Schrittweite generieren. Dass die Wahl einer äquidistanten Schrittweite über größere Integrationsdistanzen häufig zu einem unvertretbaren numerischen Aufwand führt, zeigt sehr eindrucksvoll das folgende Beispiel des restringierten Dreikörperproblems, dass als numerisches Testproblem für Anfangswertproblemlöser vielfach in der Literatur verwendet worden ist.

Beispiel (4.39) (Das restringierte Dreikörperproblem)

Wir betrachten die bereits in (1.18) vorgestellte AWA zur Beschreibung einer ebenen periodischen Satellitenbahn im Gravitationsfeld von Erde und Mond. Mit den dort genannten Bezeichnungen haben wir die AWA

$$\begin{aligned}
\ddot{x} &= x + 2\,\dot{y} - \hat{\mu}\,\frac{x+\mu}{[(x+\mu)^2+y^2]^{3/2}} - \mu\,\frac{x-\hat{\mu}}{[(x-\hat{\mu})^2+y^2]^{3/2}} \\
\ddot{y} &= y - 2\,\dot{x} - \hat{\mu}\,\frac{y}{[(x+\mu)^2+y^2]^{3/2}} - \mu\,\frac{y}{[(x-\hat{\mu})^2+y^2]^{3/2}}, \qquad (4.40)\\
x(0) &= 1.2, \quad y(0) = 0, \quad \dot{x}(0) = 0, \quad \dot{y}(0) = -1.049357510
\end{aligned}$$

Eine Lösung dieser AWA soll im Periodenintervall $[0, t_b]$ mit $t_b \doteq 6.1921\,69331$ berechnet werden.

Wir lösen die AWA (transformiert in ein System erster Ordnung) auf zwei Arten, wobei wir jeweils ein fünfstufiges RK–Verfahren der Ordnung vier anwenden, dass auf Fehlberg (1969) zurückgeht.

Zum Einen arbeiten wir mit der konstanten Schrittweite $h := t_b/1000$, führen also 1000 Integrationsschritte mit je fünf Auswertungen der rechten Seite aus. Die sich ergebenden Integrationspunkte sind in Abbildung 4.8 blau eingezeichnet. Man erkennt, dass jeweils bei den erdnahen Bereichen die Schrittweite noch zu groß ist und daher Integrationsfehler auftreten, die schließlich die Bahn völlig verfälschen.

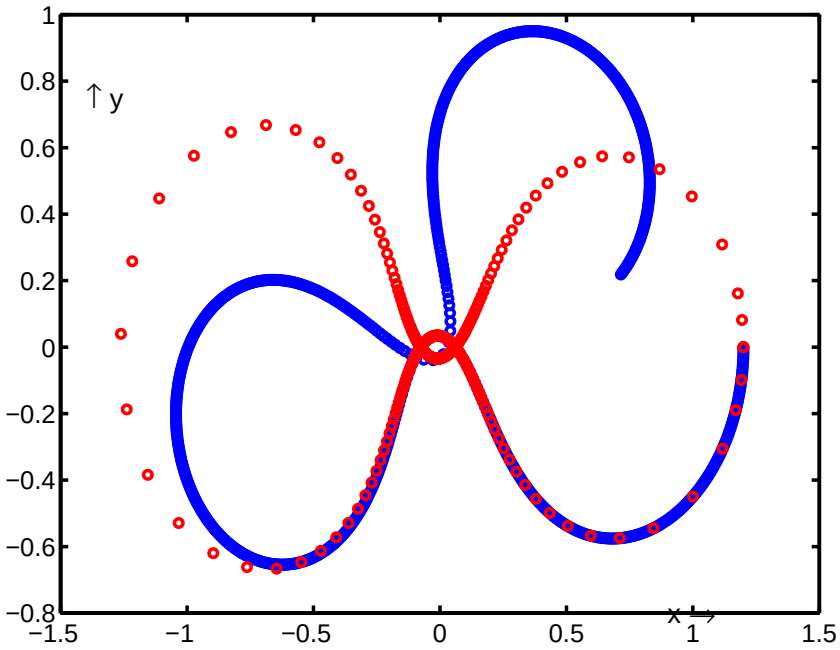

Abb. 4.8. Dreikörperproblem.

Zum Anderen arbeiten wir mit dem gleichen Integrationsverfahren, benutzen jedoch eine automatische Schrittweitensteuerung nach Fehlberg. Die sich ergebenden Integrationsknoten bei vorgegebener Toleranzanforderung tol $= 10^{-5}$ sind ebenfalls in Abbildung 3.8 eingezeichnet (rot). Man erkennt, dass sich die Bahn hier tatsächlich (im Rahmen der Zeichengenauigkeit) schließt. Im Endpunkt ergibt sich sogar ein relativer/absoluter Fehler $\leq 1.4 \times 10^{-4}$.

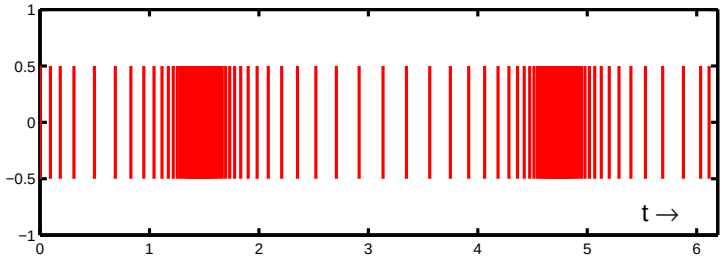

Abb. 4.9. Integrationsgitter.

In der Abbildung 4.9 ist das von der Schrittweitensteuerung erzeugte Gitter dargestellt. Die von der Schrittweitensteuerung erzeugten Schrittweiten sind fern der Erde relativ groß (Größenordnung ≈ 0.3) und werden erdnahe auf etwa 2×10^{-4} reduziert. Das schrittweitengesteuerte Verfahren kommt insgesamt mit nur 2196 Auswertungen der rechten Seite aus, ist also weniger als halb so teuer wie unsere Rechnung mit konstanter Schrittweite!

Die Algorithmen zur Schrittweitensteuerung arbeiten mit einer *Schätzung des lokalen Diskretisierungsfehlers*, d.h. zu jedem Integrationsschritt

$$(t_j, Y_j) \;\rightarrow\; (t_j + h, Y_j + h\Phi(t_j, Y_j; h))$$

mit einer aktuellen Schrittweite $h > 0$ wird zugleich ein (numerisch berechenbarer!) Schätzwert τ_{est} (est von Estimation) ermittelt:

$$\tau_{\text{est}}(t_j, Y_j; h) \;\approx\; \tau(t_j, Y_j; h) \;=\; (1/h)\;[y(t_{j+1}; t_j, Y_j) - Y_j - h\Phi(t_j, Y_j; h)]\;.$$

Von einer optimalen lokalen Schrittweite h_j^* wird nun mit einer vom Benutzer vorzugebenden Genauigkeitsanforderung tol (von Toleranz) gefordert:

$$\| \tau(t_j, Y_j; h_j^*) \| \;=\; \text{tol}.$$

Zusammen mit der Ordnungseigenschaft: $\tau(t_j, Y_j; h) = C(t_j)\, h^p + O(h^{p+1})$ folgt hiermit die folgende Heuristik:

$$\begin{aligned} \text{tol} \;&=\; \| \tau(t_j, Y_j; h_j^*) \| \;\approx\; \| C(t_j) \| \, (h_j^*)^p \;=\; \| C(t_j)\, h^p \| \, (h_j^*/h)^p \\ &\approx\; \| \tau(t_j, Y_j; h) \| \, (h_j^*/h)^p \;\approx\; \| \tau_{\text{est}} \| \, (h_j^*/h)^p \end{aligned}$$

und somit

$$h_j^* \;\approx\; \left(\frac{\text{tol}}{\| \tau_{\text{est}} \|} \right)^{(1/p)} \; h\;. \tag{4.41}$$

Diese Beziehung muss für die praktische Anwendung noch modifiziert werden. Die Schrittweite wird dazu etwas kleiner gewählt (Sicherheitsfaktor $q \in]0, 1[$) als optimal, um die Zahl der nicht erfolgreichen Integrationsschritte ($\|\tau_{\text{est}}\| > \text{tol}$) klein zu halten. Ferner werden Schranken $0 < \nu < 1 < \mu$ eingeführt, mit denen ein starkes Oszillieren der Schrittweite vermieden werden soll. Insgesamt erhält man folgenden Grundalgorithmus zur adaptiven Schrittweitenwahl.

Algorithmus (4.42) (Schrittweitensteuerung)

Start: Toleranzschranke: $\text{tol} > 0$, Parameter: $q,\ \nu \in]0, 1[,\ \mu > 1$,
$j := 0$, $Y_0 := y_0$, Startschrittweite: h_0 mit $0 < h_0 \le (t_b - t_0)$,
Minimale Schrittweite: $h_{\min} > 0$;

Iteration: $Y_{j+1} := Y_j + h_j\, \Phi(t_j, Y_j; h_j)$, $\tau_{\text{est}}(t_j, Y_j; h_j)$,
$h := q\, (\text{tol}/\|\tau_{\text{est}}\|)^{(1/p)}\, h_j$,
$h := \max[\, \min(h,\ \mu\, h_j),\ \nu\, h_j\,]$, Falls: $h < h_{\min}$, Stop!

Falls: $\|\tau_{\text{est}}\| > \text{tol}$ (Integrationsschritt wird verworfen)
$h_j := h$, gehe zu Iteration;

Sonst: (Integrationsschritt wird akzeptiert)
$t_{j+1} := t_j + h_j, \quad h_{j+1} := \min(h, t_b - t_{j+1}), \quad j := j+1;$
Falls: $t_j = t_b$, Stop! Sonst: Gehe zu Iteration.

Natürlich gibt es in den professionellen Realisierungen der Schrittweitensteuerung verschiedene Varianten und Verfeinungen des obigen Grundalgorithmus. So werden häufig Skalierungen verwendet und es wird an Stelle einer universellen Toleranzanforderung tol (hier absoluter Fehler pro Schrittweite) mit relativen und absoluten Genauigkeitsforderungen gearbeitet, die auch komponentenweise unterschiedlich vorgegeben werden können.

G. Eingebettete Runge-Kutta Verfahren.

Eine effiziente Methode, den lokalen Diskretisierungsfehler zu schätzen, besteht in der Verwendung so genannter eingebetteter RK-Verfahren. Hierunter versteht man ein Paar von RK-Verfahren mit gemeinsamen Knoten c_i, gemeinsamer Verfahrensmatrix (a_{ij}), aber unterschiedlichen Gewichten.

Tabelle 4.6: Eingebettete RK-Verfahren.

0					
c_2	a_{21}				
c_3	a_{31}	a_{32}			
$\vdots$	$\vdots$	$\vdots$	$\ddots$		
c_s	a_{s1}	a_{s2}	$\dots$	$a_{s,s-1}$	
	b_1	b_2	$\dots$	b_{s-1}	b_s
	$\widehat{b}_1$	$\widehat{b}_2$	$\dots$	$\widehat{b}_{s-1}$	$\widehat{b}_s$

Das Verfahren $\Phi(t, Y; h) = \sum b_i\, k_i$ habe hierbei die Konsistenzordnung p, das zweite Verfahren $\widehat{\Phi}(t, Y; h) = \sum \widehat{b}_i k_i$ habe die Konsistenzordung $\widehat{p}$, wobei üblicherweise $\widehat{p} = p-1$ oder $\widehat{p} = p+1$ ist. Man bezeichnet solche Verfahren auch kurz mit RK $p\,(\widehat{p})$. Das Verfahren Φ ist das eigentliche Integrationsverfahren, das Verfahren $\widehat{\Phi}$ dient zur Schätzung des lokalen Diskretisierungsverfahrens:

$$\tau_{\text{est}}(t_j, Y_j; h_j) \;:=\; \sum_{i=1}^{s} (\widehat{b}_i - b_i)\, k_i(t_j, Y_j; h_j)\,. \tag{4.42}$$

Die ersten eingebetteten RK–Verfahren sind von Merson (1957), Ceschino (1962) und Zonneveld (1963) konstruiert worden. Viele Verfahren dieser Klasse , die in den Anwendungen

besonders erfolgreich warnen und sind, gehen auf Fehlberg zurück. In Tabelle 4.7 sind die Koeffizienten des Fehlbergschen RKF4(5) Verfahrens angegeben. Dieses Verfahren haben wir zur Lösung des Beispiels (4.39) verwendet.

Tabelle 4.7: RK4(5)–Verfahren von Fehlberg (1969)

0						
$\frac{1}{4}$	$\frac{1}{4}$					
$\frac{3}{8}$	$\frac{3}{32}$	$\frac{9}{32}$				
$\frac{12}{13}$	$\frac{1932}{2197}$	$-\frac{7200}{2197}$	$\frac{7296}{2197}$			
1	$\frac{439}{216}$	-8	$\frac{3680}{513}$	$-\frac{845}{4104}$		
$\frac{1}{2}$	$-\frac{8}{27}$	2	$-\frac{3544}{2565}$	$\frac{1859}{4104}$	$-\frac{11}{40}$	
	$\frac{25}{216}$	0	$\frac{1408}{2565}$	$\frac{2197}{4104}$	$-\frac{1}{5}$	0
	$\frac{16}{135}$	0	$\frac{6656}{12825}$	$\frac{28561}{56430}$	$-\frac{9}{50}$	$\frac{2}{55}$

Ein weiteres außerordentlich erfolgreiches Verfahren ist ein RK7(8) Verfahren von Fehlberg, dessen Koeffizienten in Tabelle 4.8 angegeben sind.

Alle von Fehlberg angegebenen Verfahren sind vom Typ RK$p(q)$ mit $q > p$. Dabei ist das Verfahren der Konsistenzordnung p das eigentliche Integrationsverfahren, das Verfahren höherer Ordnung, in der Regel $q = p + 1$ dient lediglich zur Schrittweitensteuerung. Die Verfahren sind daher auch so konzipiert worden, dass die Abbrechfehler der Verfahren niedrigerer Ordnung möglichst kleine Koeffizienten (Faktoren bei den elementaren Differentialen) besitzen.

Dormand und Price (1980) bemühten sich statt dessen, eingebettete RK-Verfahren zu konstruieren, bei denen die Fehlerkoeffizienten des Verfahrens *höherer* Ordnung minimiert werden und verwenden natürlich dann auch dieses Verfahren als eigentliches Integrationsverfahren. Sie konstruierten so die vielfach verwendeten Verfahren vom Typ RK5(4) und RK8(7), genannt DOPRI5 und DOPRI8. Die Koeffizienten beider Verfahren (respektive eine genaue rationale Approximation dieser) findet man bei Deuflhard, Bornemann. Anwendungen dieser Verfahren auf das restringierte Dreikörperproblem sind in Hairer, Norsett und Wanner beschrieben.

Tabelle 4.8: RK7(8)–Verfahren von Fehlberg (1968)

0													
$\frac{2}{27}$	$\frac{2}{27}$												
$\frac{1}{9}$	$\frac{1}{36}$	$\frac{1}{12}$											
$\frac{1}{6}$	$\frac{1}{24}$	0	$\frac{1}{8}$										
$\frac{5}{12}$	$\frac{5}{12}$	0	$-\frac{25}{16}$	$\frac{25}{16}$									
$\frac{1}{2}$	$\frac{1}{20}$	0	0	$\frac{1}{4}$	$\frac{1}{5}$								
$\frac{5}{6}$	$-\frac{25}{108}$	0	0	$\frac{125}{108}$	$-\frac{65}{27}$	$\frac{125}{54}$							
$\frac{1}{6}$	$\frac{31}{300}$	0	0	0	$\frac{61}{225}$	$-\frac{2}{9}$	$\frac{13}{900}$						
$\frac{2}{3}$	2	0	0	$-\frac{53}{6}$	$\frac{704}{45}$	$-\frac{107}{9}$	$\frac{67}{90}$	3					
$\frac{1}{3}$	$-\frac{91}{108}$	0	0	$\frac{23}{108}$	$-\frac{976}{135}$	$\frac{311}{54}$	$-\frac{19}{60}$	$\frac{17}{6}$	$-\frac{1}{12}$				
1	$\frac{2383}{4100}$	0	0	$-\frac{341}{164}$	$\frac{4496}{1025}$	$-\frac{301}{82}$	$\frac{2133}{4100}$	$\frac{45}{82}$	$\frac{45}{164}$	$\frac{18}{41}$			
0	$\frac{3}{205}$	0	0	0	0	$-\frac{6}{41}$	$-\frac{3}{205}$	$-\frac{3}{41}$	$\frac{3}{41}$	$\frac{6}{41}$	0		
1	$-\frac{1777}{4100}$	0	0	$-\frac{341}{164}$	$\frac{4496}{1025}$	$-\frac{289}{82}$	$\frac{2193}{4100}$	$\frac{51}{82}$	$\frac{33}{164}$	$\frac{12}{41}$	0	1	
	$\frac{41}{840}$	0	0	0	0	$\frac{34}{105}$	$\frac{9}{35}$	$\frac{9}{35}$	$\frac{9}{280}$	$\frac{9}{280}$	$\frac{41}{840}$	0	0
	0	0	0	0	0	$\frac{34}{105}$	$\frac{9}{35}$	$\frac{9}{35}$	$\frac{9}{280}$	$\frac{9}{280}$	0	$\frac{41}{840}$	$\frac{41}{840}$