Somit ist

$$y(t_0 + h) = y_0 + hf(t_0, y_0) + \frac{1}{2}h^2(f_t(t_0, y_0) + f_y(t_0, y_0)f(t_0, y_0)) + O(h^3).$$

Setzen wir

$$\Phi(t, y, h) := f(t, y) + \frac{1}{2}h \left( f_t(t, y) + f_y(t, y) f(t, y) \right),$$

so erhalten wir ein Verfahren mit der Konsistenzordnung p = 2. Wir sehen ferner, daß prinzipiell Verfahren beliebig hoher Ordnung auf diese Weise konstruiert werden können. Allerdings werden dann immer höhere Ableitungen von f benötigt. Eine Möglichkeit, die benötigten Ableitungen zu erhalten, bietet die automatische Differentiation, [5, 4].

# 2.3.1 Explizite Runge-Kutta-Verfahren

## Einige wichtige explizite Runge-Kutta-Verfahren

Bei Taylorverfahren wie sie in Beispiel 2.12 vorgeführt wurden, werden die Ableitungen von f benötigt. Verfahren, die ohne explizite Kenntnis der Ableitungen von f auskommen, werden in der Praxis bevorzugt, weil geschlossene Ausdrücke für die Ableitungen von f oft nicht vorliegen und geeignete Approximationen (z.B. durch Differenzenquotienten) aufwendig zu bestimmen sind. Die meistverwendeten Verfahren sind die sog.  $Runge-Kutta-Verfahren^2$ .  $Explizite\ Runge-Kutta-Verfahren$  haben die folgende Form:

Definition 2.13 (Explizite Runge-Kutta-Verfahren) Eine Inkrementfunktion  $\Phi = \Phi(t, y, h)$  gehört zu einen s-stufiges explizites Runge-Kutta-Verfahren, falls sie für Zahlen  $a_{ij}, b_i \in \mathbb{R}$ ,  $c_i \in [0, 1]$  von der folgenden Form ist:

$$\Phi(t, y, h) := \sum_{i=1}^{s} b_{i}k_{i}, 
k_{1} := f(t, y), 
k_{2} := f(t + c_{2}h, y + ha_{21}k_{1}), 
k_{3} := f(t + c_{3}h, y + h(a_{31}k_{1} + a_{32}k_{2})), 
\vdots \vdots \vdots \vdots 
k_{s} := f(t + c_{s}h, y + h(a_{s1}k_{1} + \dots + a_{s s-1}k_{s-1})).$$

Zudem müssen die Koeffizienten  $b_i$  die Konsistenzbedingung

$$\sum_{i=1}^{s} b_i = 1 \tag{2.13}$$

erfüllen.

<sup>&</sup>lt;sup>2</sup>Runge, Carle David Tolmé, 1856–1927, Kutta, Martin Wilhelm, 1867–1944

Die Zahlen  $a_{ij}$ ,  $c_i$ ,  $b_i$ , die ein Runge-Kutta-Verfahren festlegen, werden üblicherweise kompakt in einem Tableau wie folgt notiert:

Die Bedingung (2.13) ergibt sich aus dem folgenden Satz:

Satz 2.14 (Konsistenz von expliziten RK-Verfahren) Jedes explizite RK-Verfahren im Sinne von Definition 2.13 ist konsistent.

**Beweis:** Für h = 0 gilt  $k_i = f(t, y), i = 1, ..., s$ . Damit folgt für die Inkrementfunktion

$$\Phi(t, y, 0) = \sum_{i=1}^{s} b_i k_i = f(t, y) \sum_{i=1}^{s} b_i = f(t, y),$$

weil (2.13) gefordert wird. Aus der Charakterisierung von Konsistenz in Lemma 2.7 folgt damit die Behauptung.

Das einfachste explizite Runge-Kutta-Verfahren ist das explizite Eulerverfahren: Es ist ein 1-stufiges Verfahren mit

$$s = 1,$$
  $b_1 = 1.$ 

Nach Satz 2.14 sind also alle RK-Verfahren konsistent. Sinnvollerweise wird man die die Parameter so wählen, daß die Konsistenzordnung möglichst hoch ist. Wir illustrieren das prinzipielle Vorgehen am Fall expliziter zweistufiger RK-Verfahren:

Beispiel 2.15 (zweistufige RK-Verfahren) Zum Anfangswert  $y_0$  liefert das allgemeine zweistufige RK-Verfahren die Approximation

$$y_1 = y_0 + h [b_1 k_1 + b_2 k_2] = y_0 + h [b_1 f(t_0, y_0) + b_2 f(t_0 + c_2 h, y_0 + a_{21} h f(t_0, y_0))].$$

Wir entwickeln nun die rechte Seite in eine Taylorreihe in h und schreiben kurz  $f = f(t_0, y_0)$ ,  $f_t = f_t(t_0, y_0)$ ,  $f_y = f_y(t_0, y_0)$  und erhalten

$$y_1 = y_0 + h(b_1 + b_2)f + h^2[b_2c_2f_t + b_2a_{21}ff_y] + O(h^3).$$

Um den Konsistenzfehler zu bestimmen, entwickeln wir die exakte Lösung y(t) in eine Taylorreihe um  $t_0$ . Durch Differenzieren der Differentialgleichung y'(t) = f(t, y(t)) erhalten wir

$$y'(t_0) = f$$
,  $y''(t_0) = f_t + f f_y$ ,

und somit

$$y(t_0 + h) = y_0 + hy'(t_0) + \frac{h^2}{2}y''(t_0) + O(h^3).$$

Will man nun die Parameter  $c_2$ ,  $b_1$ ,  $b_2$ ,  $a_{21}$  so bestimmen, daß ein Verfahren möglichst hoher Konsistenzordnung entsteht, so ergeben sich durch Koeffizientenvergleich die drei Bestimmungsgleichungen

$$b_1 + b_2 = 1, b_2 c_2 = \frac{1}{2}, b_2 a_{21} = \frac{1}{2}.$$
 (2.15)

Dieses nichtlineare System besitzt mehrere Lösungen. Zwei bekannte sind das Verfahren von Heun<sup>3</sup> und das modifizierte Eulerverfahren mit den folgenden Tableaus:

$$\begin{array}{c|cccc}
0 & & & & & & & & & \\
1 & 1 & & & & & & \frac{1}{2} & \frac{1}{2} \\
\hline
& \frac{1}{2} & \frac{1}{2} & & & & & & 0 & 1
\end{array}$$

Eine naheliegende Frage ist, ob man die Koeffizienten  $b_i$ ,  $c_i$  und  $a_{ij}$  auch so bestimmen kann, daß sogar ein Verfahren dritter Ordnung entsteht. Dies ist nicht möglich, wie Proposition 2.17 unten zeigt.

Beispiel 2.16 (RK4) Das klassische Runge-Kutta-Verfahren 4. Ordnung wird sehr oft eingesetzt. Es ist ein 4-stufiges Verfahren gegeben durch

$$\Phi(t, y, h) := \frac{1}{6} [k_1 + 2k_2 + 2k_3 + k_4],$$

$$k_1 := f(t, y),$$

$$k_2 := f\left(t + \frac{h}{2}, y + \frac{1}{2}hk_1\right),$$

$$k_3 := f\left(t + \frac{h}{2}, y + \frac{1}{2}hk_2\right),$$

$$k_4 := f(t + h, y + hk_3).$$

Das entsprechende Tableau ist

Man kann durch Taylorentwicklungen nachrechnen, daß es ein Verfahren mit Konsistenzordnung p=4 ist.

Für Funktionen f, die nur von t abhängen, ist das klassische RK4-Verfahren gerade die Simpsonregel.

#### Butcherschranken

Die Konstruktion von expliziten Runge-Kutta-Verfahren ist prinzipiell einfach, wie wir in Beispiel 2.15 gesehen haben: Will man für eine gegebene Stufenanzahl s ein Verfahren möglichst hoher Ordnung erzeugen, so ergeben sich durch Taylorentwicklung Bedinungen an die gesuchten

<sup>&</sup>lt;sup>3</sup>Heun, Karl, 1859–1929

Koeffizienten  $c_i$ ,  $a_{ij}$  und  $b_i$ . Diese nichtlinearen Gleichungen sind für höhere Stufenanzahl jedoch nicht einfach lösbar. Schon die Frage, was die maximale erreichbare Ordnung p bei s Stufen ist, ist schwer zu beantworten. Antworten auf diese Frage und die Entwicklung von Systematiken zur Behandlung der Bedingungsgleichungen gehen auf Butcher<sup>4</sup> zurück. Es gilt z.B.

**Proposition 2.17 (Butcherschranken)** Für die maximale erreichbare Ordnung p eines expliziten Runge-Kutta-Verfahrens mit s Stufen gilt  $p \leq s$ . Zudem gelten folgende verschärfte Abschätzungen:

**Beweis:** Die Aussage, daß  $p \leq s$  gelten muß, folgt durch Betrachtung des konkreten Anfangswertproblems

$$y' = y,$$
  $y(0) = 1.$ 

Es ergibt sich dann für den ersten Schritt mit Schrittweite h:

$$k_1 = 1$$
,  $k_2 = 1 + ha_{21}k_1 = 1 + ha_{21}$ ,  $k_3 = 1 + h(a_{31}k_1 + a_{32}k_2) = 1 + (a_{31} + a_{21})h + a_{32}a_{21}h^2$ 

und allgemein durch Induktion, daß  $k_i$  ein Polynom vom Grad i-1 in der Variablen h ist. Für ein explizites RK-Verfahren mit s Stufen gilt somit, daß nach dem ersten Schritt die Approximation  $\tilde{y}(h)$ 

$$\tilde{y}(h) = 1 + h\Phi(0, 1, h)$$

ein Polynom vom Grad s in der Variablen h ist. Die Taylorentwicklung von  $e^h$  um h=0 ist  $e^h=\sum_{i=0}^\infty \frac{1}{i!}h^i$ ; somit muß  $e^h-\tilde{y}(h)=O(h^p)$  für ein  $p\leq s+1$  sein, d.h. ein explizites RK-Verfahren kann höchstens die Ordnung s haben.

Für die Beweise der angegebenen Butcherschranken sei bemerkt, daß auch [6] nur auf die Originalliteratur verweist.

Die Aussage, daß  $p \leq s-2$  für  $s \geq 9$  ist so zu verstehen, daß eine scharfe Butcherschranke noch nicht bekannt ist. Es war deshalb lange ein "Sport", Verfahren möglichst hoher Ordnung bei möglichst geringer Stufenzahl zu konstruieren. Den Rekord hält immer noch E. Hairer mit einem 17-stufigen Verfahren der Ordnung 10.

### Beispiele für das Konvergenzverhalten

Das explizite Eulerverfahren ist ein Verfahren erster Ordnung. Da pro Schritt eine Auswertung der Funktion f benötigt wird, ist die erwartete Konvergenz in der Perspektive "Fehler gegen Anzahl Funktionsauswertungen":

$$\max_{i=0,\dots,N} |y_{ex}(t_i) - y_i| \le CF^{-1}.$$
(2.16)

Das Runge-Kutta-Verfahren aus Beispiel 2.16 benötigt nur 4 Auswertungen von f pro Schritt; andererseits erwarten wir nach Satz 2.10, daß der Fehler sich wie  $O(h^4)$  verhält. Umgerechnet in "Genauigkeit gegen Fehler" erwarten wir also

$$\max_{i=0,\dots,N} |y_{ex}(t_i) - y_i| \le CF^{-4},\tag{2.17}$$

<sup>&</sup>lt;sup>4</sup>John C. Butcher, 1933–, neuseeländischer Mathematiker

h	$2^{-1}$	$2^{-2}$	$2^{-3}$	$2^{-4}$	$2^{-5}$	$2^{-6}$	$2^{-7}$
$F_{Euler}$	$2^1$	$2^2$	$2^3$	$2^{4}$	$2^{5}$	$2^{6}$	$2^7$
exp. Euler	0.468	0.277	0.152	$0.80_{-1}$	$0.412_{-1}$	$0.209_{-1}$	$0.105_{-1}$
$F_{RK4}$	$2^3$	$2^4$	$2^5$	$2^{6}$	$2^7$	$2^{8}$	$2^{9}$
RK 4	$0.936_{-3}$	$0.719_{-4}$	$0.498_{-5}$	$0.328_{-6}$	$0.2105_{-7}$	$0.133_{-8}$	$0.838_{-10}$

**Tabelle 2.2:** Vergleich explizites Eulerverfahren und RK4. Fehler gegen Schrittweite h sowie Anzahl Funktionsauswertungen F bei glatter Lösung.

wobei F die Anzahl Auswertungen von f mißt. Vergleicht man die Aufwandsabschätzung (2.16) mit (2.17), so sieht man, daß das Runge-Kutta-Verfahren mit seiner höheren Ordnung eine wesentlich höhere Genauigkeit bei vergleichbarem Aufwand erzielt als das explizite Eulerverfahren.

Beispiel 2.18 (optimales Konvergenzverhalten von Euler und RK4) Wir wenden das explizte Eulerverfahren aus Beispiel 2.1 und das Runge-Kutta-Verfahren 4. Ordnung (RK4) aus Beispiel 2.16 auf das Anfangswertproblem

$$y'(t) = y(t), \qquad y(0) = 1$$

an und vergleichen den Fehler an der Stelle T=1. Die Ergebnisse sind in Tabelle 2.2 zusammengestellt. Man sieht, daß das Eulerverfahren ein Verfahren 1. Ordnung ist (der Fehler reduziert sich um einen Faktor 2 bei Halbierung der Schrittweite) und RK4 ein Verfahren 4. Ordnung (der Fehler reduziert sich um einen Faktor 16 bei Halbierung der Schrittweite). Wenn wir Effizienz als Verhältnis "Genauigkeit pro Anzahl Funktionsauswertungen" messen, ist das Verfahren 4. Ordnung erheblich besser als das explizite Eulerverfahren, wie die Abschätzungen (2.16) und (2.17) zeigen.

Wir haben in Beispiel 2.18 gesehen, daß Verfahren hoher Ordnung in gewissen Situationen effizienter arbeiten als Verfahren niedriger Ordnung. Im Wesentlichen ist dies gegeben, wenn die gesuchte Lösung hinreichend oft differenzierbar ist (beim RK4-Verfahren aus Beispiel 2.16 z.B. sollte die gesuchte Lösung  $y \in C^5$  sein). An den folgenden zwei Beispielen zeigen wir nun, daß man von Verfahren hoher Ordnung nicht erwarten kann, daß die maximale Konvergenzordnung errreicht wird, wenn die gesuchte Lösung nicht hinreichend oft differenzierbar ist.

Beispiel 2.19 (Reduzierte Konvergenzordnung des RK4 bei nicht-glatter Lösung) Wir betrachten das Anfangswertproblem

$$y'(t) = f(t)$$
 auf  $[0, 1]$ ,  $y(0) = 0$ ,  $f(t) = \begin{cases} 0 & \text{für } t \le 1/2 \\ t - 1/2 & \text{für } t > 1/2 \end{cases}$ 

mit Lösung

$$y(t) = \int_0^t f(\tau) d\tau = \begin{cases} 0 & \text{für } t < 1/2\\ \frac{1}{2}(t - 1/2)^2 & \text{für } t \ge 1/2. \end{cases}$$

Da die rechte Seite der Differentialgleichung nicht explizit von der Variablen y abhängt, ist für beliebige Wahl von Knoten  $t_i$  das explizite Eulerverfahren gerade die summierte Trapezregel und das RK4-Verfahren die summierte Simpsonregel, d.h. z.B. für das RK4-Verfahren gilt

$$y_i = \sum_{j=0}^{i-1} \frac{h_j}{6} \left( f(t_j) + 4f(t_j + h_j/2) + f(t_{j+1}) \right).$$

(Übung: Man überzeuge sich von dieser Beobachtung.) Definiert man den Index M so, daß

$$t_M \le 1/2 < t_{M+1}$$

haben wir, da die summierte Simpsonregel exakt ist für Polynome vom Grad 3 und f auf dem Intervall  $[0, t_M]$  ein quadratisches Polynom ist,

$$y_M = \int_0^{t_M} f(t) dt.$$

Weil f auf dem Intervall  $[t_{M+1}, t_N]$  wieder ein quadratisches Polynom ist, gilt weiter:

$$y_N = y_{M+1} + \int_{t_{M+1}}^{t_N} f(t) dt.$$

Damit ergibt sich, daß der Gesamtfehler vollständig von dem Fehler bestimmt wird, der im m-ten Schritt gemacht wird:

$$y_N - \int_0^{t_N} f(t) dt = y_{M+1} - y_M - \int_{t_M}^{t_{M+1}} f(t) dt.$$

Da die rechte Seite f bei t = 1/2 nicht glatt ist, ist auch die Lösung y nicht glatt, so daß nicht mit Konvergenz vierter Ordnung gerechnet werden kann. Um dies einzusehen, betrachten wir nun das folgende Gitter: Die Knoten  $t_i$ , i = 0, ..., N, sind gegeben durch

$$t_0 = 0$$
,  $t_i = \frac{h}{2} + (i-1)h$ ,  $i = 1, ..., N-1$ ,  $t_N = 1$ ,

wobei h=1/N mit  $N=2M,\ M\in\mathbb{N}$ . Wesentlich ist diese Wahl motiviert von der Beobachtung, daß der Punkte 1/2 kein Knoten ist, und daß gilt  $1/2=t_M+\frac{h}{2}=t_{M+1}-\frac{h}{2}$ . Wir erhalten damit wegen  $t_M=1/2-h/2$ ,  $t_M=1/2,\ t_{M+1}=1/2+h/2$ 

$$\int_{t_M}^{t_{M+1}} f(t) \, dt - (y_{M+1} - y_M) = \int_{t_M}^{t_{M+1}} f(t) \, dt - \frac{h_M}{6} \left( f(t_M) + 4 f(t_M + h_M/2) + f(t_{M+1}) \right) = \frac{1}{8} h^2 - \frac{1}{12} h^2 = \frac{1}{24} h^2.$$

Dieses  $O(N^{-2})$ -Verhalten wird numerisch in Fig. 2.2 illustriert.

Daß man nicht Konvergenz vom Typ  $O(h^4)$  erwarten kann, ist ist auch aus dem Konvergenzresultat Satz 2.10 ersichtlich, denn die Voraussetzung (iii) kann nur mit p=1 erfüllt werden: Sei zu h>0 hierzu t=1/2-h/2 gewählt; dann berechnen wir  $\Phi(t,y(t),h)$  aus:

$$k_1 = f(t) = 0,$$
  $k_2 = f(t + h/2) = f(1/2) = 0,$   $k_3 = f(t + h/2) = f(1/2) = 0,$   $k_4 = f(t + h) = f(1/2 + h/2) = h/2$ 

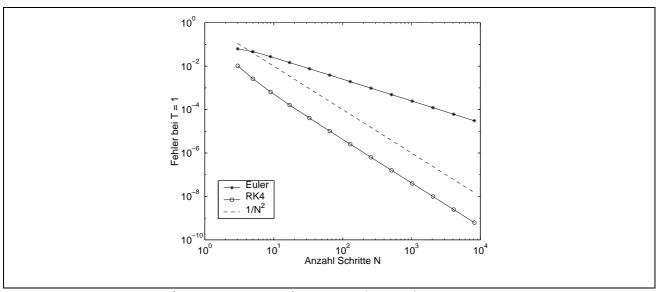
die Inkrementfunktion  $\Phi_{RK4}(t, y(t), h) = \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4) = \frac{1}{12}h$  und damit

$$y(t+h) - (y(t) + h\Phi_{RK4}(t, y(t), h)) = y(1/2 + h/2) - \left(0 + \frac{1}{12}h^2\right) = \left(\frac{1}{8} - \frac{1}{12}\right)h^2.$$

Mithin können wir nur p = 1 in (iii) von Satz 2.10 wählen.

In Beispiel 2.19 liefert das RK4-Verfahren (und sogar das Eulerverfahren) den exakten Werte, wenn der Punkt 1/2 ein Knoten ist. Dies ist ein besonders einfaches Beispiel, bei dem man aus Kenntnis des Verhaltens der Lösung schließen kann, wie man die Knoten zu wählen hat, um ein möglichst effizientes Verfahren zu erhalten. Das folgende Beispiel gibt weitere Fälle an, in denen durch geschickte Wahl der Knoten die optimale Konvergenzrate erreicht werden kann.

#### Beispiel 2.20 (Optimale Konvergenzordnung von RK4 mit angepaßten Gittern)



Figur 2.2: (vgl. Beispiel 2.19) RK4-Verfahren für nicht-glatte Lösung.

- 1. Wir betrachten die Differentialgleichung  $y'(t)=1.1\cdot t^{0.1}$  auf dem Intervall [0,1] mit Anfangsbedingung y(0)=0. Für konstante Schrittweite h=1/N wird in Fig. 2.3 das Verhalten des expliziten Eulerverfahrens und des RK4-Verfahrens illustriert. Im linken Bild beobachten wir, daß in diesem Beispiel das RK4-Verfahren nur unwesentlich besser ist als das Eulerverfahren mit den Konvergenzverhalten O(h). Der Grund hierfür sind die mangelnden Differenzierbarkeitseigenschaften der Lösung  $y(t)=t^{1.1}$  am Anfangspunkt t=0. Wählt man die Knoten  $t_i$  nicht uniform, so kann durch geeignete Wahl der Knotenpositionen die optimale Konvergenzrate (gemessen in Fehler gegen Anzahl Schritte N) erreicht werden. In rechten Bild von Fig. 2.3 ist das Verhalten des RK4-Verfahren für die Wahl  $t_i=\left(\frac{i}{N}\right)^{5/1.1}$ ,  $i=0,\ldots,N$ , angegeben. Wir beobachten, daß das Verfahren wie  $O(N^{-4})$  konvergiert.
- 2. Wir betrachten das Anfangswertproblem

$$y'(t) = \tilde{f}(t) + y(t)$$
 für  $t \in [0, 1],$   $y(0) = e^{1/3},$ 

wobei die Funktion  $\tilde{f}$  gegeben ist durch

$$\tilde{f}(t) = \begin{cases} -2(1 - e^{|t-1/3|}) & t \le 1/3\\ 0 & t > 1/3. \end{cases}$$

Die Lösung  $y \in C^1$  ist

$$y(t) = \begin{cases} 2 - e^{|t-1/3|} & t \le 1/3\\ e^{|t-1/3|} & t > 1/3. \end{cases}$$

Verglichen werden das explizite Eulerverfahren und das RK4-Verfahren mit konstanter Schrittweite h=1/N. In Fig. 2.4 ist der Fehler am Endpunkt T=1 gegen die Anzahl Schritte N für beide Verfahren angegeben. Wir beobachten, daß das RK4-Verfahren nur das Konvergenzverhalten  $O(N^{-3})$  hat. Der Grund für das nicht optimale Verhalten des RK4-Verfahrens ist ähnlich wie im Beispiel 2.19, daß beim Überspringen der Stelle t=1/3 ein (relativ) großer Fehler gemacht wird, der dann vom Verfahren nicht mehr korrigiert werden kann sondern bis zu t=T "weitertransportiert" wird. Eine Möglichkeit, die optimale Konvergenzrate wiederherzustellen, ist auf den Intervallen [0/1/3] und [1/3,1] mit je konstanter Schrittweite zu arbeiten und sicherzustellen, daß  $t^*=1/3$  ein Knoten ist. Dies wurde im rechten Bild von 2.4 gemacht; in der Tat beobachtet man nun eine Konvergenz 4. Ordnung.