2 Einschrittverfahren

2.1 Einführung

Im folgenden werden wir uns bei der Beschreibung und Analyse von numerischen Verfahren für Anfangswertprobleme auf den Fall $n = 1$ beschränken. Dies wird nur gemacht, um die Notation einfacher zu gestalten. Die vorgestellten Verfahren lassen sich in natürlicher Weise auf Probleme für $n > 1$ übertragen. Auch die Konvergenztheorie kann einfach auf diese Fälle erweitert werden.

I.a. kann ein Anfangswertproblem der Form (1.2) nicht exakt gelöst werden. Man muß deshalb numerische Verfahren entwickeln, um eine approximative Lösung zu erhalten. Ziel unseres numerischen Verfahrens ist, die gesuchte Funktion y auf einem gegebenen Intervall $[t_0, T]$ zu bestimmen. Hierzu unterteilen wir das Intervall $[t_0, T]$ in N Teilintervalle mit Knoten

$$t_0 < t_1 < \cdots < t_N = T$$

und sucht Approximationen y_i an die exakte Lösung $y(t_i)$, $i = 0, \ldots, N$. Wir fassen die Punkte t_i, $i = 0, \ldots, N$, zu einem *Gitter* $\Delta := \{t_i \,|\, i = 0, \ldots, N\}$ zusammen und definieren die i-te *Schrittweite*

$$h_i := t_{i+1} - t_i, \qquad i = 0, \ldots, N-1.$$

Wir schreiben weiterhin $h_\Delta := \max_{i=0,\ldots,N-1} h_i$.
Da $y(t_0) = y_0$ vorgeschrieben ist, ist y_0 bekannt. Bei Einschrittverfahren wird nun mithilfe von y_0 eine Approximation $y_1 \approx y(t_1)$ bestimmt, anschließend eine Approximation $y_2 \approx y(t_2)$, etc. bis man bei $t_N = T$ angelangt ist. Das allereinfachste Einschrittverfahren ist das sog. *explizite Eulerverfahren*[1], welches auch als "Euler vorwärts" oder als Eulersches Polygonzugverfahren bekannt ist:

Beispiel 2.1 (explizites Eulerverfahren)**.** Die Lösung y ist für $t = t_0$ durch $y(t_0) = y_0$ bereits vorgegeben. Die Differentialgleichung $y'(t) = f(t, y(t))$ legt damit auch die Steigung der Kurve $t \mapsto y(t)$ an der Stelle t fest. Aus dem Taylorschen Satz erhalten wir damit eine Approximation für $y(t_1)$ durch

$$y(t_1) \approx y_1 := y_0 + (t_1 - t_0)y'(t_0) = y_0 + (t_1 - t_0)f(t_0, y_0) = y_0 + h_0 f(t_0, y_0).$$

Offensichtlich kann man für die Approximation $y_2 \approx y(t_2)$ ähnlich vorgehen: Zwar ist nun die exakte Steigung der Funktion $t \mapsto y(t)$ im Punkte $t = t_1$ nicht bekannt, aber die Differentialgleichung $y'(t_1) = f(t_1, y(t_1)) \approx f(t_1, y_1)$ liefert eine gute Approximation an $y'(t_1)$, falls y_1 eine gute Approximation an $y(t_1)$ ist. Wir setzen also $y_2 := y_1 + h_1 f(t_1, y_1)$. Ganz allgemein erhalten wir damit die Vorschrift

$$y_{i+1} = y_i + h_i f(t_i, y_i), \qquad i = 0, \ldots, N-1. \tag{2.1}$$

■

Die "Herleitung" des expliziten Eulerverfahrens in Beispiel 2.1 legt ein weiteres Verfahren nahe, das sog. *implizite Eulerverfahren*:

[1] Euler, Leonhard, 1707–1783

Beispiel 2.2 (implizites Eulerverfahren)**.** Beim expliziten Eulerverfahren wird die Approximation y_{i+1} durch Taylorentwicklung um die Stelle t_i motiviert. Man kann die gesuchte Approximation y_{i+1} auch dadurch motivieren, daß man um die Stelle t_{i+1} eine Taylorentwicklung macht und mithilfe des bereits bekannten Wertes y_i eine Gleichung herleitet: Nach dem Taylorschen Satz erwartet man

$$y_i \approx y_{i+1} + (t_i - t_{i+1})y'(t_{i+1}) = y_{i+1} + (t_i - t_{i+1})f(t_{i+1}, y_{i+1}).$$

Ersetzt man nun $\approx$ durch $=$, so erhält man eine Gleichung für das zu bestimmende y_{i+1}:

$$\text{Finde } y_{i+1} \text{ so, daß } y_{i+1} = y_i + h_i f(t_{i+1}, y_{i+1}), \qquad i = 0, \ldots, N. \tag{2.2}$$

■

Bemerkung 2.3. Im Unterschied zum expliziten Eulerverfahren aus Bsp. 2.1 ist beim impliziten Eulerverfahren in Beispiel 2.2 die Approximation y_{i+1} nicht mehr explizit gegeben, sondern es muß eine Gleichung gelöst werden (y_{i+1} ist implizit bestimmt). Da i.a. die Funktion f nicht linear im zweiten Argument ist, ist ihre Lösung aufwendiger als bei expliziten Verfahren. Man wird deshalb die Verwendung von impliziten Verfahren vermeiden, wenn man kann. Wie wir im Kapitel 3 sehen werden, sind implizite Verfahren bei *steifen* Differentialgleichungen dennoch von Vorteil. ■

Die Form (2.1) des expliziten Eulerverfahrens und die Form (2.2) des impliziten Eulerverfahrens legen es nahe, das allgemeine Einschrittverfahren in der folgenden Form zu schreiben:

Definition 2.4 (Einschrittverfahren)**.** *Ein numerisches Verfahren, bei dem zu gegebenem y_0 die Werte y_i, $i = 1, \ldots, N$, durch eine Rekursion*

$$y_{i+1} = y_i + h_i \Phi(t_i, y_i, y_{i+1}, h_i), \qquad i = 0, \ldots, N-1, \tag{2.3}$$

bestimmt werden, heißt Einschrittverfahren*. Die Funktion Φ heißt* Inkrementfunktion*. Hängt die Funktion Φ nicht explizit von y_{i+1}, so spricht man von einem* expliziten Einschrittverfahren*; andernfalls spricht man von einem* impliziten Einschrittverfahren*.*

Explizites und implizites Eulerverfahren sind gegeben durch die Wahl

$$\Phi(t_i, y_i, y_{i+1}, h_i) = f(t_i, y_i), \qquad \Phi(t_i, y_i, y_{i+1}, h_i) = f(t_i + h_i, y_{i+1}).$$

Bemerkung 2.5. Bei impliziten Verfahren muß die eindeutige Lösbarkeit der Gleichung

$$y_{i+1} = y_i + h\Phi(t_i, y_i, y_{i+1}, h)$$

für hinreichend kleine h sichergestellt werden. Falls Φ eine glatte Funktion ist, folgt aus dem Banachschen Fixpunktsatz und dem Satz über implizite Funktionen, daß es eine Funktion $\tilde{\Phi}$ gibt, so daß

$$y_{i+1} = y_i + h\tilde{\Phi}(t_i, y_i, h)$$

gilt (vgl. den Beweis von Satz 3.7). Für theoretische Zwecke wie z.B. die Konvergenzanalyse im folgenden Abschnitt kann damit ein implizites Verfahren auch als explizites Vefahren aufgefaßt werden. ■

2.2 Konvergenzanalyse von expliziten Verfahren

Ein explizites Verfahren wie (2.3) liefert Approximationen y_i an die gesuchte Lösung $y(t)$ in den Knoten t_i. Für ein Einschrittverfahren der Form (2.3) sprechen wir deshalb von Konvergenz, falls für die Approximationen y_i und die gesuchte Lösung $y(t)$ gilt:

$$\max_{i=0,\ldots,N} |y(t_i) - y_i| \to 0 \qquad \text{falls} \qquad h_\Delta = \max_{i=0\ldots,N} h_i \to 0. \tag{2.4}$$

2.2.1 Konsistenz

Offensichtlich muß die Funktion Φ in (2.3), die das Einschrittverfahren definiert, gewisse Eigenschaften haben, damit man Konvergenz erwarten kann. Da die Inkrementfunktion Φ die einzige Verbindung zur Differentialgleichung darstellt, die von der gesuchten Lösung erfüllt wird, muß sie eng mit Lösungen der gesuchten Differentialgleichung zusammenhängen. Dieser Zusammenhang wird im Begriff der *Konsistenz* in Definition 2.6 genauer erfaßt.

Wir betrachten explizite Einschrittverfahren der Form

$$y_{i+1} = y_i + h_i \Phi(t_i, y_i, h_i). \tag{2.5}$$

Offensichtlich ist eine Mindestvoraussetzung für Konvergenz, daß der Fehler, der in jedem Schritt gemacht wird, "klein" ist. Ein Maß für diesen *lokalen* Fehler ist der sog. *Konsistenzfehler*, den wir wie folgt definieren:

Definition 2.6 (Konsistenzfehler)**.** *Sei $G \subset \mathbb{R}^2$ ein Gebiet, $f \in C(G)$ lokal lipschitzstetig im zweiten Argument. Sei die Inkrementfunktion Φ für ein $\underline{h} > 0$ auf der Menge $\mathcal{G} := G \times [0, \underline{h}] \subset \mathbb{R}^3$ definiert. Für $(\tilde{t}, \tilde{y}, h) \in \mathcal{G}$ ist der* Konsistenzfehler *$\tau(\tilde{t}, \tilde{y}, h)$ definiert als*

$$\tau(\tilde{t}, \tilde{y}, h) = y_{\tilde{t},\tilde{y}}(\tilde{t} + h) - (\tilde{y} + h\Phi(\tilde{t}, \tilde{y}, h))$$

wobei die Funktion $t \mapsto y_{\tilde{t},\tilde{y}}(t)$ die Lösung von

$$y'_{\tilde{t},\tilde{y}}(t) = f(t, y_{\tilde{t},\tilde{y}}(t)), \qquad y(\tilde{t}) = \tilde{y}.$$

ist. Gilt für jedes $(\tilde{t}, \tilde{y}) \in G$

$$\lim_{h \to 0+} \frac{\tau(\tilde{t}, \tilde{y}, h)}{h} = 0,$$

so heißt das Einschrittverfahren (2.5) konsistent *auf G. Das Verfahren heißt* konsistent von der Ordnung $p > 0$*, falls es für jede kompakte Teilmenge $K \subset G$ eine Konstante $C > 0$ und ein $h' > 0$ gibt, so daß*

$$|\tau(\tilde{t}, \tilde{y}, h)| \le C h^{p+1} \qquad \forall (\tilde{t}, \tilde{y}) \in K \quad \text{und alle } h \in [0, h'].$$

Der Konsistenzfehler τ mißt den Fehler, den das numerische Verfahren in *einem* Schritt der Länge h macht, d.h. die exakte Lösung $y_{\tilde{t},\tilde{y}}$ zum Zeitpunkt $\tilde{t} + h$ wird verglichen mit der numerischen Approximation $\tilde{y} + h\Phi(\tilde{t}, \tilde{y}, h)$. Der Konsistenzfehler τ, wie in Definition 2.6 eingeführt, ist damit für $h > 0$

$$\begin{aligned} \tau(\tilde{t}, \tilde{y}, h) &= y_{\tilde{t},\tilde{y}}(\tilde{t} + h) - (\tilde{y} + h\Phi(\tilde{t}, \tilde{y}, h)) \\ &= h \left(\frac{y_{\tilde{t},\tilde{y}}(\tilde{t} + h) - y_{\tilde{t},\tilde{y}}(\tilde{t})}{h} - \Phi(\tilde{t}, y_{\tilde{t},\tilde{y}}(\tilde{t}), h) \right), \end{aligned} \tag{2.6}$$

was eine Darstellung des Konsistenzfehlers ist, die oft in der Literatur als Definition des Konsistenzfehlers verwendet wird.
Das folgende Lemma erlaubt eine schnelle Überprüfung der Konsistenz eines Einschrittverfahrens:

Lemma 2.7. *Sei $G \subset \mathbb{R}^2$ ein Gebiet, $f \in C(G)$ sei lokal lipschitzstetig im zweiten Argument. Sei die Inkrementfunktion Φ für ein $\underline{h} > 0$ auf $\mathcal{G} = G \times [0, \underline{h}]$ definiert und sei $\Phi \in C(\mathcal{G})$. Dann sind die folgenden Aussagen äquivalent:*

(i) Das Einschrittverfahren (2.5) ist konsistent im Sinne von Definition 2.6.

(ii) $\Phi(t, y, 0) = f(t, y) \qquad \forall (t, y) \in G.$

Beweis: Aus der Konsistenz folgt $\lim_{h\to 0+} \frac{\tau(\tilde{t},\tilde{y},h)}{h} = 0$. Die Gleichung (2.6) impliziert damit

$$\begin{aligned} 0 &= \lim_{h\to 0+} \frac{\tau(\tilde{t},\tilde{y},h)}{h} = \lim_{h\to 0+} \frac{y_{\tilde{t},\tilde{y}}(\tilde{t}+h) - y_{\tilde{t},\tilde{y}}(\tilde{t})}{h} - \lim_{h\to 0+} \Phi(\tilde{t},\tilde{y},h) && \text{(2.7a)} \\ &= y'_{\tilde{t},\tilde{y}}(\tilde{t}) - \Phi(\tilde{t},\tilde{y},h) = f(\tilde{t},\tilde{y}) - \Phi(\tilde{t},\tilde{y},0). && \text{(2.7b)} \end{aligned}$$

Damit ergibt sich die Behauptung (i) $\Longrightarrow$ (ii). Umgekehrt folgern wir (ii) $\Longrightarrow$ (i) aus der Voraussetzung $\Phi(\tilde{t},\tilde{y},0) = f(\tilde{t},\tilde{y})$, indem wir die Schritte in (2.7) rückwärts durchführen. □

finis 2.DS

Bemerkung 2.8 (Konsistenz von Verfahren)**.** Man kann die Funktion f als zusätzlichen Parameter in die Inkrementfunktion Φ aufnehmen. Man spricht dann von Konsistenz eines Verfahrens, wenn für jede Funktion f, die die Voraussetzungen der Definition 2.6 erfüllt, das zugehörige Verfahren konsistent ist. Analog spricht man von einem Verfahren der Ordnung p, wenn für jedes $f \in C^p(G)$ das Verfahren Konsistenzordnung p hat. Im folgenden werden wir diese Sprechweise übernehmen. ∎

Der Grund für die Forderung $f \in C^p(G)$ für Verfahren der Konsistenzordnung p liegt darin begründet, daß die Konsistenzordnung typischerweise mit Hilfe der Taylorentwicklung der Lösung $y_{\tilde{t},\tilde{y}}$ ausgerechnet wird; die Forderung $f \in C^p(G)$ garantiert dann nach Satz 1.12, daß $y_{\tilde{t},\tilde{y}} \in C^{p+1}$ auf einer Umgebung von $\tilde{t}$. Wir führen dies exemplarisch für das Eulerverfahren vor:

Beispiel 2.9 (Konsistenzordnung beim expliziten Eulerverfahren)**.** Das explizite Eulerverfahren hat die Konsistenzordnung 1. Um dies einzusehen betrachten wir $f \in C^1(G)$.
1. Schritt: Wir betrachten zuerst einen festen Punkt $(\tilde{t},\tilde{y}) \in G$. Nach Satz 1.12 ist dann die Lösung $y_{\tilde{t},\tilde{y}} \in C^2(U)$ für eine Umgebung $U = (\tilde{t}-\alpha, \tilde{t}+\alpha)$ des Punktes $\tilde{t}$. Nach dem Taylorschen Satz erhalten wir damit für $h \in (-\alpha/2, \alpha/2)$

$$y_{\tilde{t},\tilde{y}}(\tilde{t}+h) = \tilde{y} + h y'_{\tilde{t},\tilde{y}}(\tilde{t}) + r(\tilde{t},h)$$

wobei das Restglied r gegeben ist durch

$$|r(\tilde{t},h)| = \left| \int_{x=\tilde{t}}^{\tilde{t}+h} (\tilde{t}+h-x) y''_{\tilde{t},\tilde{y}}(x)\, dx \right| \le \frac{1}{2} h^2 \|y''_{\tilde{t},\tilde{y}}\|_{C([\tilde{t}-\alpha/2,\tilde{t}+\alpha/2])}.$$

Somit ergibt sich für den Konsistenzfehler

$$\begin{aligned}\tau(\tilde{t},\tilde{y},h) &= y_{\tilde{t},\tilde{y}}(\tilde{t}+h) - \big(\tilde{y}+h\Phi(\tilde{t},\tilde{y},h)\big) = \tilde{y} + hy'_{\tilde{t},\tilde{y}}(\tilde{t}) + r(\tilde{t},h) - \big(\tilde{y}+hf(\tilde{t},\tilde{y})\big)\\ &= r(\tilde{t},h),\end{aligned}$$

weil wegen der Differentialgleichung $y'_{\tilde{t},\tilde{y}}(\tilde{t}) = f(\tilde{t},\tilde{y})$ gilt. Damit erhalten wir damit

$$|\tau(\tilde{t},\tilde{y},h)| \leq \frac{1}{2}\|y''_{\tilde{t},\tilde{y}}\|_{C([\tilde{t}-\alpha/2,\tilde{t}+\alpha/2])}h^{1+1},$$

d.h. wir erwarten, daß das explizite Eulerverfahren ein Verfahren der Ordnung 1 ist.
2. Schritt: Um den Beweis, daß das explizite Eulerverfahren ein Verfahren erster Ordnung ist, formal abzuschließen, benötigen wir noch ein Kompaktheitsargument. Das generelle Vorgehen, das wir hier vorstellen, ist das typische Vorgehen.
Sei $K \subset G$ kompakt. Wir führen für Punkte $(t,y) \in K$ die Rechtecksumgebungen $R_{2\delta}(t,y) := (t-2\delta, t+2\delta) \times (y-2\delta, y+2\delta)$ ein. Wegen der Kompaktheit von K (und G offen) existiert dann ein $\delta > 0$, so daß

$$K \subset \cup_{(t,y)\in K} R_{2\delta}(t,y) \subset G \tag{2.8}$$

gilt. Weiter gilt

$$K \subset \tilde{K} := \overline{\cup_{(t,y)\in K} R_{\delta}(t,y)} \subset \cup_{(t,y)\in K} R_{2\delta}(t,y) \subset G.$$

Da K kompakt ist, ist es beschränkt; somit ist auch $\tilde{K}$ beschränkt. Als abgeschlossene Menge ist somit $\tilde{K}$ kompakt. Weil $f \in C^1(G)$, ist somit

$$M := \|f\|_{C(\tilde{K})} + \|f_t\|_{C(\tilde{K})} + \|f_y\|_{C(\tilde{K})} \tag{2.9}$$

endlich. Insbesondere sind damit für jedes $(\tilde{t},\tilde{y}) \in K$ die Funktionen f, f_t, f_y auf dem Rechteck $R := [\tilde{t}-\delta, \tilde{t}+\delta] \times [\tilde{y}-\delta, \tilde{y}+\delta] \subset \tilde{K}$ definiert und durch M beschränkt. Nach dem Satz von Picard-Lindelöf (Satz 1.3) existiert damit ein $\alpha > 0$, so daß die Lösung $y_{\tilde{t},\tilde{y}} \in C^1(\tilde{t}-\alpha, \tilde{t}+\alpha)$ für ein α, welches nur von M und δ abhängt, und der Graph von $y_{\tilde{t},\tilde{y}}$ ist in $R \subset \tilde{K}$ enthalten. Mit der Kettenregel und $y' = f(t,y)$ ergibt sich damit als Abschätzung für $y''_{\tilde{t},\tilde{y}}$

$$\|y''_{\tilde{t},\tilde{y}}\|_{C([\tilde{t}-\alpha/2,\tilde{t}+\alpha/2])} \leq \|f_t\|_{C(R)} + \|f_y f\|_{C(R)} \leq M + M^2.$$

Somit schließen wir aus dem ersten Schritt, daß für $h \leq \alpha/2$ gilt:

$$|\tau(\tilde{t},\tilde{y},h)| \leq \frac{1}{2}(M+M^2)h^2,$$

wobei α und M lediglich von der kompakten Menge K und f abhängen. ∎

2.2.2 Konvergenzanalyse

Der Begriff der Konsistenzordnung in Definition 2.6 quanitifiziert den *lokalen* Fehler, der durch das Verfahren in jedem Schritt eingeführt wird. Wir wenden uns nun dem Problem zu, den *globalen* Fehler

$$\max_{i=0,\ldots,N} |y(t_i) - y_i|$$

abzuschätzen. Wir werden sehen, daß die Konsistenzordnung so eingeführt wurde, daß unter vernünftigen Annahmen das Einschrittverfahren mit der Ordnung p konvergiert, d.h.

$$\max_{i=0,\dots,N} |y(t_i) - y_i| \le Ch^p, \qquad h = \max_{i=0,\dots,N-1} h_i = \max_{i=0,\dots,N-1} t_{i+1} - t_i.$$

Da wir nur an der numerischen Approximation der exakten Lösung y_{ex} interessiert sind, verlangen wir Bedingungen an die Inkrementfunktion Φ nur in einer Umgebung der gesuchten Lösung y_{ex}. Insbesondere reicht es, die Konsistenz des Verfahrens für die gesuchte Lösung zu überprüfen (vgl. (iii) in Satz 2.10). Wir erhalten das folgende Konvergenzresultat:

Satz 2.10 (Konvergenz von Einschrittverfahren)**.**
Voraussetzungen: Sei $J \subset \mathbb{R}$ ein offenes Intervall und $y_{ex} \in C^1(J)$. Sei $[t_0, T] \subset J$ ein Intervall. Erfülle die Inkrementfunktion Φ für ein $\delta > 0$ und ein $\underline{h} > 0$ folgende Bedingungen:

(i) Φ ist definiert und stetig auf $\mathcal{G} := S_\delta \times [0, \underline{h}]$, wobei (vgl. Fig. 2.1)

$$S_\delta = \bigcup_{t\in[t_0,T]} \{t\} \times [y_{ex}(t) - \delta, y_{ex}(t) + \delta]$$

(ii) Φ ist lipschitzstetig bzgl. des zweiten Arguments, d.h. es existiert $L_\Phi > 0$ derart, daß

$$|\Phi(t, y, h) - \Phi(t, \hat{y}, h)| \le L_\Phi |y - \hat{y}| \qquad \forall (t, y, h), (t, \hat{y}, h) \in \mathcal{G}.$$

(iii) ("Konsistenz auf der Lösung") Es existiert $C_\tau > 0$, $p \in \mathbb{N}$, so daß für alle $t \in [t_0, T]$ und $h > 0$ mit $t + h \le T$ gilt:

$$|y_{ex}(t + h) - (y_{ex}(t) + h\Phi(t, y_{ex}(t), h))| \le C_\tau h^{p+1}.$$

Behauptung: Es gibt ein $\overline{h} \in (0, \underline{h})$, so daß für jedes Gitter $\Delta = \{t_i \,|\, i = 0, \dots, N\}$ mit $h_\Delta \le \overline{h}$ das folgende gilt:

1. *die durch (2.5) gegebenen Approximationen y_i, $i = 0, \dots, N$, existieren, und*
2. *sie erfüllen die Abschätzung*

$$|y_{ex}(t_i) - y_i| \le C_\tau (t_i - t_0) e^{L_\Phi (t_i - t_0)} h_\Delta^p,$$

Insbesondere gilt damit

$$\max_{i=0,\dots,N} |y_{ex}(t_i) - y_i| \le C_\tau (T - t_0) e^{L_\Phi (T - t_0)} h_\Delta^p.$$

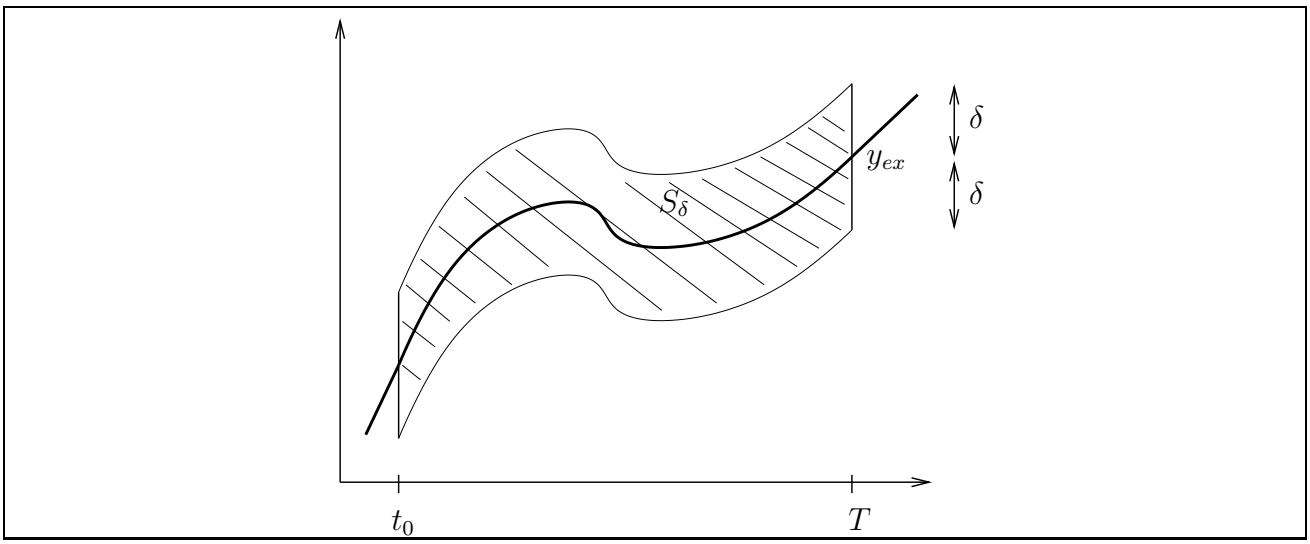
Bevor wir diesen Satz beweisen, formulieren wir ein Lemma, das wir im weiteren benötigen. Es stellt eine Variante des Gronwall-Lemmas dar, welches wir bereits in Lemma 1.4 kennengelernt haben.

Lemma 2.11 (Gronwall)**.** *Seien $(\delta_i)_{i=0}^N$, $(e_i)_{i=0}^N$, $(\eta_i)_{i=0}^N$ gegeben mit $\delta_i \ge 0$, $e_i \ge 0$, $\eta_i \ge 0$ für $i = 0, \dots, N$. Es gelte*

$$e_{i+1} \le (1 + \delta_i) e_i + \eta_i, \qquad i = 0, \dots, N - 1.$$

Dann ist

$$|e_i| \le \left(e_0 + \sum_{j=0}^{i-1} \eta_j \right) e^{\sum_{j=0}^{i-1} \delta_j} \qquad i = 0, 1, \dots, N. \qquad \text{(Konvention: leere Summe} = 0\text{)}$$



Figur 2.1: Skizze zum Beweis des Konvergenzsatzes für Einschrittverfahren

Beweis: Siehe Übung. □

Beweis von Satz 2.10: Die Inkrementfunktion Φ ist nur in der Nähe der exakten Lösung $(t, y_{ex}(t))$ definiert. Es ist also nicht von vorneherein klar, daß wir die Werte y_i in (2.5) tatsächlich bilden können. Um dieses Problem zu meistern, gehen wir in dem Beweis in zwei Schritten vor: Im ersten Schritt konstruieren wir eine Inkrementfunktion $\widetilde{\Phi}$, die auf $[t_0, T] \times \mathbb{R} \times [0, \underline{h}]$ definiert ist. Damit ist das Verfahren

$$\widetilde{y}_{i+1} := \widetilde{y}_i + h_i \widetilde{\Phi}(t_i, \widetilde{y}_i, h), \qquad \widetilde{y}_0 := y_0 \tag{2.10}$$

wohldefiniert. Wir zeigen, daß dieses Hilfsverfahren konvergiert. Im zweiten Schritt werden wir dann zeigen, daß für hinreichend kleine h die Approximationen $\widetilde{y}_i$ mit den y_i übereinstimmen. Daraus folgt dann die Konvergenz des Verfahrens.
1. Schritt: Wir definieren $\widetilde{\Phi}$ auf $[t_0, T] \times \mathbb{R} \times [0, \underline{h}]$ durch

$$\widetilde{\Phi}(t, y, h) := \begin{cases} \Phi(t, y, h) & \text{falls } y \in [y_{ex}(t) - \delta, y_{ex}(t) + \delta] \\ \Phi(t, y_{ex}(t) + \delta, h) & \text{falls } y > y_{ex}(t) + \delta \\ \Phi(t, y_{ex}(t) - \delta, h) & \text{falls } y < y_{ex}(t) - \delta \end{cases}$$

Die Voraussetzung (ii) impliziert zudem die Lipschitzstetigkeit von $\widetilde{\Phi}$ bzgl. des zweiten Argumentes:

$$\left| \widetilde{\Phi}(t, y, h) - \widetilde{\Phi}(t, \hat{y}, h) \right| \leq L_\Phi |y - \hat{y}| \qquad \forall t \in [t_0, T], \quad h \in [0, \underline{h}], \qquad y, \hat{y} \in \mathbb{R}.$$

Ferner stimmen auf $S_\delta \times [0, \underline{h}]$ die Funktionen $\widetilde{\Phi}$ mit Φ auf $S_\delta \times [0, \underline{h}]$ überein. Also folgt für den Konsistenzfehler des Verfahrens mit Inkrementfunktion $\widetilde{\Phi}$ für $t \in [t_0, T]$ und $h > 0$ mit $t + h \leq T$:

$$\begin{aligned} \tilde{\tau}(t, y_{ex}(t), h) &= y_{ex}(t+h) - \Big(y_{ex}(t) + h\widetilde{\Phi}(t, y_{ex}(t), h) \Big) \\ &= y_{ex}(t+h) - (y_{ex}(t) + h\Phi(t, y_{ex}(t), h)) \end{aligned}$$

und damit aus der Voraussetzung (iii):

$$|\tilde{\tau}(t, y_{ex}(t), h)| \leq C_\tau h^{p+1} \qquad \forall t \in [t_0, T], h > 0 \quad \text{mit } t + h \leq T. \tag{2.11}$$

Wir leiten nun eine Gleichung für den Fehler $\widetilde{e}_i := y_{ex}(t_i) - \widetilde{y}_i$ her. Durch Subtraktion der Gleichungen

$$\begin{aligned} \widetilde{y}_{i+1} &= \widetilde{y}_i + h_i \widetilde{\Phi}(t_i, \widetilde{y}_i, h_i) \\ y_{ex}(t_{i+1}) &= y_{ex}(t_i) + h_i \widetilde{\Phi}(t_i, y_{ex}(t_i), h_i) + \tilde{\tau}(t_i, h_i, y_{ex}(t_i)) \end{aligned}$$

erhalten wir

$$\widetilde{e}_{i+1} = \widetilde{e}_i + h_i \left[\widetilde{\Phi}(t_i, y_{ex}(t_i), h_i) - \widetilde{\Phi}(t_i, \widetilde{y}_i, h_i) \right] + \tilde{\tau}(t_i, y_{ex}(t_i), h_i).$$

Unter Ausnutzung der Lipschitzstetigkeit von $\widetilde{\Phi}$ und der Abschätzung (2.11) für $\tilde{\tau}$ erhalten wir

$$|\widetilde{e}_{i+1}| \leq |\widetilde{e}_i| + h_i L_\Phi |\widetilde{e}_i| + C_\tau h_i^{p+1}, \qquad i = 0, \ldots, N-1.$$

Aus $h_i \leq h_\Delta$ und dem Gronwall-Lemma 2.11 erhalten wir damit für $i \in \{0, \ldots, N\}$ wegen $\widetilde{e}_0 = 0$ und $\sum_{j=0}^{i-1} h_i = t_i - t_0$

$$|\widetilde{e}_i| \leq \left(\sum_{j=0}^{i-1} C_\tau h_j^{p+1} \right) e^{\sum_{j=0}^{i-1} L_\Phi h_j} \leq C_\tau h_\Delta^p \sum_{j=0}^{i-1} h_j e^{L_\Phi (t_i - t_0)} \leq C_\tau (t_i - t_0) e^{L_\Phi (t_i - t_0)} h_\Delta^p.$$

Insbesondere gilt damit

$$\max_{i=0,\ldots,N} |y_{ex}(t_i) - \widetilde{y}_i| \leq C_\tau (T - t_0) e^{L_\Phi (T - t_0)} h_\Delta^p. \tag{2.12}$$

2. Schritt: Die Konvergenzaussage (2.12) impliziert die Existenz eines $\overline{h} > 0$, so daß für $h_\Delta \in (0, \overline{h}]$ gilt

$$\widetilde{y}_i \in [y_{ex}(t_i) - \delta, y_{ex}(t_i) + \delta], \qquad \forall i \in \{0, \ldots, N\},$$

d.h. für $h_\delta \leq \overline{h}$ gilt $(t_i, \widetilde{y}_i) \in S_\delta$ für alle $i \in \{0, \ldots, N\}$. Da auf $S_\delta \times [0, \underline{h}]$ die Funktionen $\widetilde{\Phi}$ und Φ übereinstimmen, gilt also $y_i = \widetilde{y}_i$. Dies schließt den Beweis ab. □

Satz 2.10 besagt, daß die Konsistenz des Verfahrens bereits die Konvergenz impliziert. Dies ist bemerkenswert, wenn man bedenkt, daß Konsistenz nur den lokalen, in jedem Schritt gemachten Fehler mißt und nicht die Fortpflanzung der lokalen Fehler berücksichtigt. Wir werden später bei Mehrschrittverfahren sehen, daß dort über die Konsistenz hinaus auch noch eine sog. Stabilität des Verfahrens verlangt werden muß. finis 3.DS

2.3 Explizite Einschrittverfahren höherer Ordnung

Einschrittverfahren höherer Ordnung werden typischerweise auf eine von zwei Arten erzeugt: Es gibt die Runge-Kutta-Verfahren, die wir in Abschnitt 2.3.1 behandeln und die Extrapolationsverfahren, die wir in ansprechen. Diesen Konstruktionen von Verfahren höherer Ordnung schicken wir zwei wichtige Gründe für ihren Einsatz voraus:

N	x	Fehler
10	2.5937426	1.2_{-01}
10^2	2.7048137	1.3_{-02}
10^3	2.7169204	1.4_{-03}
10^4	2.7181435	1.4_{-04}
10^5	2.7182324	4.9_{-05}
10^6	2.7180855	2.0_{-04}
10^7	2.5960464	1.2_{-01}
10^8	1	1.7_{+00}

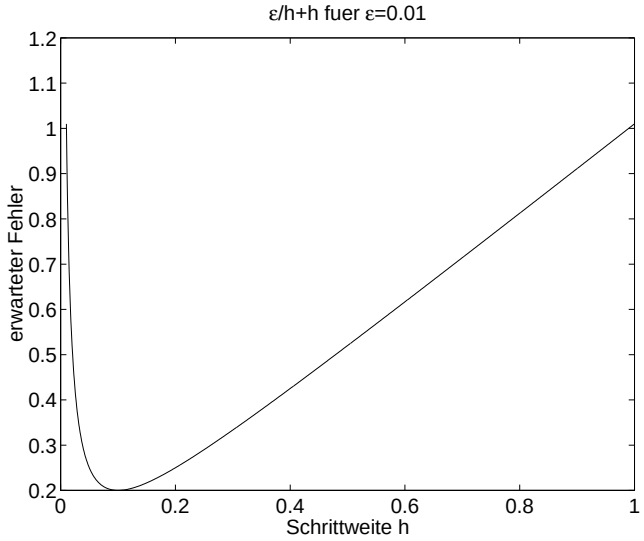


Tabelle 2.1: Links: Explizites Eulerverfahren für $y' = y$, $y(0) = 1$ bei Rechnung mit einfacher Genauigkeit (ca. 8 Ziffern). Rechts: erwartetes qualitatives Fehlerverhalten unter Berücksichtigung von Rundungsfehlern.

1. Es stellt sich heraus (siehe z.B. Tabelle 2.2), daß bei Verfahren höherer Ordnung die Relation "Fehler gegen Anzahl Funktionsauswertungen" günstiger ist. Dies ist vor allem in Anwendungen von Interesse, bei denen jede Funktionsauswertung teuer bis sehr teuer ist (weil z.B. die Auswertung der Funktion selbst die Lösung eines aufwendigen Problems beinhaltet).

2. Ein weiteres Argument für den Einsatz von Verfahren höherer Ordnung ist deren geringere Anfälligkeit für Rundungsfehler. Rundungsfehler sind bei der endlichen Rechnengenauigkeit von Computern unvermeidbar und limitieren deshalb die erreichbare Genauigkeit. Diesen Punkt beleuchten wir in den Übungen und dem numerischen Beispiel in Tabelle 2.1.

Wir illustrieren nun im folgenden Beispiel, daß es prinzipiell möglich ist, Verfahren beliebig hoher Ordnung zu konstruieren:

Beispiel 2.12 (Taylorverfahren)**.** Satz 2.10 zeigt, daß wir Konvergenzordnung p erhalten, wenn das Verfahren konsistent von der Ordnung p ist (vgl. Definition 2.6). Es bietet sich deshalb an, die Funktion Φ durch Taylorentwicklung der Lösung $t \mapsto y(t)$ des Anfangswertproblems (1.2) zu definieren. Unter Ausnutzung der Differentialgleichung können dabei Ausdrücke hergeleitet werden, in denen zwar Ableitungen von f, nicht aber die unbekannte Lösung y auftreten. Wir führen dies für die Konstruktion eines Verfahrens zweiter Ordnung vor: Sei $t \mapsto y(t)$ Lösung von $y'(t) = f(t, y(t))$ und $y(t_0) = t_0$. Durch Taylorentwicklung gilt dann

$$y(t_0 + h) \;=\; y(t_0) + hy'(t_0) + \frac{1}{2}h^2 y''(t_0) + O(h^3).$$

Durch Differentiation der Differentialgleichung erhalten wir

$$y''(t) = f_t(t, y(t)) + f_y(t, y(t))y'(t) = f_t(t, y(t)) + f_y(t, y(t))f(t, y(t)).$$

Somit ist

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

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, [4, 3]. ∎

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:*

$$\begin{aligned}
\Phi(t,y,h) &:= \sum_{i=1}^{s} b_i k_i, \\
k_1 &:= f(t,y), \\
k_2 &:= f(t+c_2h, y+ha_{21}k_1), \\
k_3 &:= f\left(t+c_3h, y+h(a_{31}k_1+a_{32}k_2)\right), \\
&\vdots \quad \vdots \quad \vdots \\
k_s &:= f\left(t+c_sh, y+h(a_{s1}k_1+\cdots+a_{s\,s-1}k_{s-1})\right).
\end{aligned}$$

Zudem müssen die Koeffizienten b_i *die Konsistenzbedingung*

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

erfüllen.

[2] 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:

$$\begin{array}{c|ccccc} 0 & & & & & \\ c_2 & a_{21} & & & & \\ c_3 & a_{31} & a_{32} & & & \\ \vdots & \vdots & \vdots & \ddots & & \\ c_s & a_{s1} & a_{s2} & \cdots & a_{s\,s-1} & \\ \hline & b_1 & b_2 & \cdots & b_{s-1} & b_s \end{array} \tag{2.14}$$

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 $k_i = f(t, y)$, $i = 1, \ldots, s$. 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

$$s = 1, \qquad b_1 = 1.$$

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

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

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

$$y_1 = y_0 + h(b_1 + b_2) f + h^2 \left[b_2 c_2 f_t + b_2 a_{21} f f_y\right] + O(h^3).$$

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

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

und somit

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

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

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

Dieses nichtlineare System besitzt mehrere Lösungen. Zwei bekannte sind das Verfahren von Heun[3] und das modifizierte Eulerverfahren mit den folgenden Tableaus:

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

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

$$\begin{aligned} \Phi(t,y,h) &:= \frac{1}{6}\left[k_1 + 2k_2 + 2k_3 + k_4\right], \\ k_1 &:= f(t,y), \\ k_2 &:= f\left(t + \frac{h}{2}, y + \frac{1}{2}hk_1\right), \\ k_3 &:= f\left(t + \frac{h}{2}, y + \frac{1}{2}hk_2\right), \\ k_4 &:= f\left(t + h, y + hk_3\right). \end{aligned}$$

Das entsprechende Tableau ist

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

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 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

[3] Heun, Karl, 1859–1929

Stufen ist, ist schwer zu beantworten. Antworten auf diese Frage und die Entwicklung von Systematiken zur Behandlung der Bedingungsgleichungen gehen auf Butcher[4] 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:*

s	1	2	3	4	5	6	7	8	9	$s \geq 9$
p	1	2	3	4	4	5	6	6	7	$p \leq s-2$

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

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

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

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

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 [5] 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":

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

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

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

[4] 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

$$y\prime(t) = f(t) \quad \text{auf } [0,1], \quad y(0) = 0, \qquad f(t) = \begin{cases} 0 & \text{für } t \le 1/2 \\ t - 1/2 & \text{für } t > 1/2 \end{cases}$$

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 eine summierte Rechtecksregel und das RK4-Verfahren die summierte Simpsonregel, d.h. z.B. für das RK4-Verfahren gilt

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

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

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

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,

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

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

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

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, $i = 0, \ldots, N$, sind gegeben durch

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

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 $t_M = 1/2 - h/2$, $t_M = 1/2$, $t_{M+1} = 1/2 + h/2$

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

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:

$$\begin{aligned} k_1 = f(t) = 0, \qquad & k_2 = f(t + h/2) = f(1/2) = 0, \\ k_3 = f(t + h/2) = f(1/2) = 0, \qquad & k_4 = f(t + h) = f(1/2 + h/2) = h/2 \end{aligned}$$

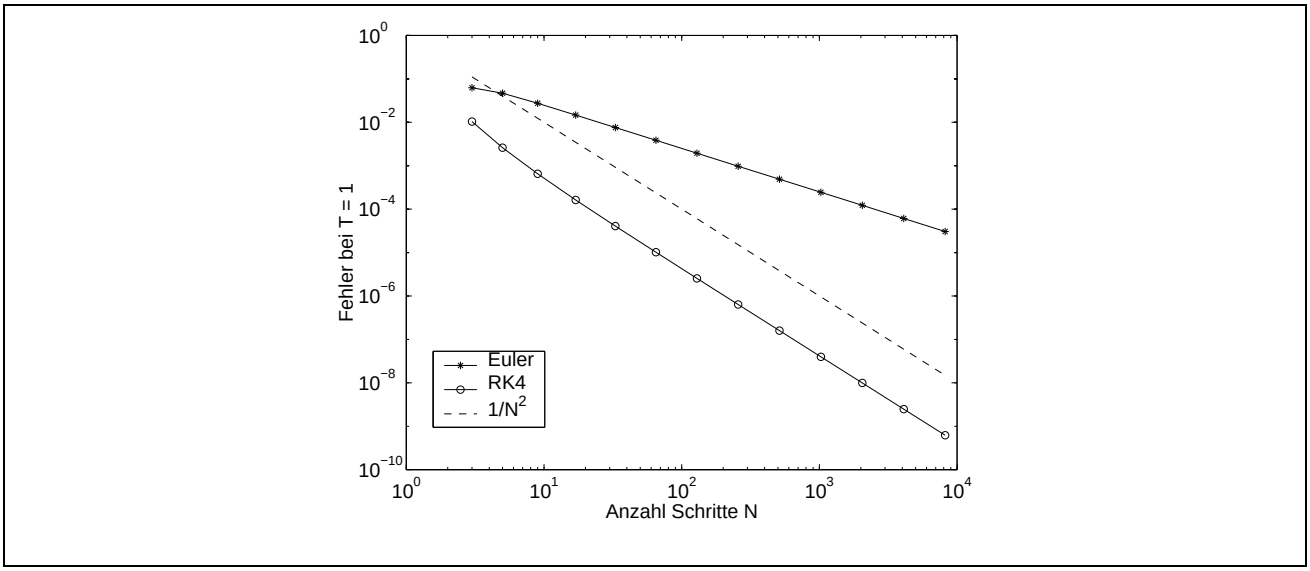
die Inkrementfunktion $\Phi_{RK4}(t, y(t), h) = \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4) = \frac{1}{12}h$ und damit

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

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

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

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.

■