Numerik von gewöhnlichen Differentialgleichungen[1]

Martin Kiehl

7. März 2012

[1]Skriptum der gleichnamigen Vorlesung in WS 2011/12 an der TUD.

Inhaltsverzeichnis

Kapitel 1

Anfangswertprobleme

1.1 Theoretische Grundlagen:

Anfangswertprobleme treten in fast allen technischen Anwendungsgebieten auf (z.B., Flugbahnoptimierung, Schaltkreisentwurf, Robotersteuerung, Fahrzeugdynamik, Reaktionskinetik).
Im einfachsten Fall ist eine Funktion $y : x \in \mathbb{R} \mapsto y(x) \in \mathbb{R}$ gesucht, welche die Differentialgleichung

$$y' = f(x, y)$$

und die Anfangsbedingung

$$y(x_0) = y_0$$

erfüllt. Die Existenz und Eindeutigkeit kann noch für relativ viele Probleme bewiesen werden. Im Fall der Eindeutigkeit bezeichnen wir die Lösung mit $y(x)$ bzw. $y(x, x_0, y_0)$, wenn wir die Abhängigkeit von den Anfangsdaten betonen wollen.
Im eindimensionalen Fall existieren für viele wichtige Fälle auch analytische Lösungsformeln. Im mehrdimensionalen Fall ($y \in \mathbb{R}^n$) beherrscht man immerhin noch den sehr wichtigen linearen Fall $y' = Ay$ mit konstanter Koeffizientenmatrix A. In der Praxis sind jedoch fast alle Differentialgleichungen hochdimensional und nichtlinear, so daß wir uns mit numerischen Lösungsverfahren behelfen müssen. Die Kenntnis einiger analytischer Lösungsmethoden ist dennoch sehr hilfreich bei der Konstruktion geeigneter Verfahren.

Wir betrachten im Folgenden Systeme von n Differentialgleichungen:

$$y' = \frac{\mathrm{d}}{\mathrm{d}x}\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = f(x, y(x)) = \begin{pmatrix} f_1(x, y_1(x), \ldots, y_n(x)) \\ f_2(x, y_1(x), \ldots, y_n(x)) \\ \vdots \\ f_n(x, y_1(x), \ldots, y_n(x)) \end{pmatrix} , \qquad (1.1.1)$$

mit $f : D \subseteq \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$, D offen zusammenhängend.
Desweiteren seien n **Anfangswerte** gegeben:

$$y_0 := y(x_0) = \begin{pmatrix} y_{0,1} \\ y_{0,2} \\ \vdots \\ y_{0,n} \end{pmatrix} \quad ; \quad (x_0, y_0) \in D \ . \qquad (1.1.2)$$

Differentialgleichungen höherer Ordnung:
Sie können auf Systeme erster Ordnung zurückgeführt werden. Im Falle

$$y^{(m)} = f(x, y(x), y'(x), \ldots, y^{(m-1)}(x)) \in \mathbb{R}^n \ . \qquad (1.1.3)$$

definiert man dazu die Hilfsfunktionen

$$\begin{aligned} z_1(x) &:= y(x) \ , \\ z_2(x) &:= y'(x) \ , \\ &\vdots \\ z_m(x) &:= y^{(m-1)}(x) \ . \end{aligned}$$

Dann gilt:

$$z' = \begin{bmatrix} z_1' \\ \vdots \\ z_{m-1}' \\ z_m' \end{bmatrix} = \begin{bmatrix} z_2 \\ \vdots \\ z_m \\ f(x, z_1, z_2, \ldots, z_m) \end{bmatrix} =: F(x, z) \in \mathbb{R}^{nm}, \qquad (1.1.4)$$

mit Anfangswerten

$$z(x_0) = \begin{pmatrix} z_1(x_0) \\ \vdots \\ z_m(x_0) \end{pmatrix} = \begin{pmatrix} y(x_0) \\ y'(x_0) \\ \vdots \\ y^{(m-1)}(x_0) \end{pmatrix} .$$

Autonome Differentialgleichungen:
Manche Programme erlauben nur die Behandlung autonomer Differentialgleichungen $y' = f(y)$. Dies ist jedoch keine wesentliche Einschränkung. Durch Einführung einer zusätzlichen Variablen $y_{n+1} := x$ erhält man eine autonome Differentialgleichung

$$z' := \frac{\mathrm{d}}{\mathrm{d}x} \begin{pmatrix} y \\ y_{n+1} \end{pmatrix} = \begin{pmatrix} y' \\ y'_{n+1} \end{pmatrix} = \begin{pmatrix} f(y_{n+1}, y) \\ 1 \end{pmatrix} \; ; \; z_0 = \begin{pmatrix} y_0 \\ x_0 \end{pmatrix} \; . \tag{1.1.5}$$

Differentialgleichungen mit Parametern:
Viele Differentialgleichungen enthalten Parameter, (Reibungskoeffizienten (Maschinenbau), Widerstände (Elektrotechnik), Geschwindigkeitskonstanten (Reaktionschemie)).

$$y' = f(x, y, p) \; ; \; p \in \mathbb{R}^{n_p} \; ; \; y(x_0) = y_0 \in \mathbb{R}^n \; .$$

Durch Einführung sogenannter trivialer Differentialgleichungen $v' = 0$ für die Parameter, mit Anfangswerten $v(x_0) = p$, erhält man dann ein System, bei dem die Parameter die Rolle von Anfangswerten spielen.

$$z' := \frac{\mathrm{d}}{\mathrm{d}x} \begin{pmatrix} y \\ v \end{pmatrix} = \begin{pmatrix} y' \\ v' \end{pmatrix} = \begin{pmatrix} f(x, y, v) \\ 0 \end{pmatrix} \; ; \; z_0 = \begin{pmatrix} y_0 \\ p \end{pmatrix} \; . \tag{1.1.6}$$

Umgekehrt kann man jeden Anfangswert als Parameter auffassen:

$$z := y - y_0 \Rightarrow z' = f(x, z + y_0) =: F(x, z, y_0) \; ; \; z(x_0) = 0 \; . \tag{1.1.7}$$

Aus der Theorie ist bekannt:

Satz 1.1.1 (Existenz– und Eindeutigkeitssatz)
Sei $S \subseteq \mathbb{R} \times \mathbb{R}^n$ *ein* **Schlauch** *um* $y(x)$*:*

$$S := \{(x, y) \in \mathbb{R}^{n+1} |\; a \le x \le b, \; l_i(x) \le y_i(x) \le u_i(x)\} \; , \tag{1.1.8}$$

mit $l_i(x)$ *untere,* $u_i(x)$ *obere Schranke von* $y_i(x)$ [1]*.*
Es existiere eine **Lipschitz-Konstante** $L < \infty$ *mit*

$$\|f(x, y) - f(x, \bar{y})\| \le L \, \|y - \bar{y}\| \quad \forall (x, y), (x, \bar{y}) \in S \; . \tag{1.1.9}$$

Dann gibt es für alle $(x_0, y_0) \in S/\partial S$ *genau eine Funktion* $y(x)$ *mit:*

[1] S ist also ein Normalbereich. Dies ist technisch einfacher und für die Anwendungen hinreichend allgemein. Satz 1.1.1 gilt aber auch für einfach susammenhängende abgeschlossene beschränkte Umgebungen von (x_0, y_0).

(i) *y ist in einer Umgebung von x_0 differenzierbar, d.h., es gilt:* $y \in C^1[a', b']$ *mit* $x_0 \in [a', b'] \subseteq [a, b]$.

(ii) *y ist Lösung des Anfangswertproblems* $y' = f(x, y(x))$ *mit* $y(x_0) = y_0$.

(iii) *Die Lösung endet nicht im Inneren von S , d.h.:* $(x, y(x)) \in S \setminus \partial S \forall x \in]a'; b'[$ *und* $y(a'), y(b') \in \partial S$.

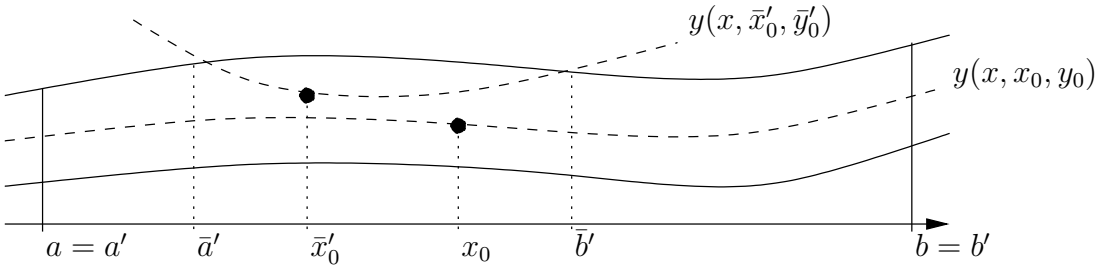


Abbildung 1.1: Existenz und Eindeutigkeitsgebiet

Beweis: Annahme $y(x)$ endet in $(\bar{x}, \bar{y}) \in S \setminus \partial S$.
Picard-Iteration:

$$y_{i+1}(t) := \Phi[y_i](t) := \bar{y} + \int_{\bar{x}}^{\bar{x}+t} f(\tau, y_i(\tau))\mathrm{d}\tau \quad ; \quad y_0(t) := \bar{y}$$

Im Raum $C^0([\bar{x}; \bar{x} + \varepsilon])$ ist die Iteration Φ bezüglich der Maximumnorm kontrahierend falls $\varepsilon < 1/L$ denn

$$\max |\Phi[u](\bar{x}+t) - \Phi[v](\bar{x}+t)| \leq \max_t \int_{\bar{x}}^{\bar{x}+t} |f(\tau, u) - f(\tau, v)|\mathrm{d}\tau \leq \varepsilon L \max |v - u|$$

Es existiert daher ein Fixpunkt y^* und es gilt:

$$y^*(t) = \bar{y} + \int_{\bar{x}}^{\bar{x}+t} f(\tau, y^*(\tau))\mathrm{d}\tau \ .$$

y^* ist also Lösung der Differentialgleichung in einer kleinen Umgebung rechts von $(\bar{x}, \bar{y})$, endet also nicht dort. Analog Fortsetzung nach links. ■

Beispiel 1.1.2

$$y' = y^2 \ , \quad [a,b] = [-2,2] \ , \quad l = 0 \ , \quad u = 10$$

$f(x,y) = y^2$, $|f_y| = |2y| \leq 20$ falls $l < y < u$.

$$y(0) = 1 \Longrightarrow y(x) = \frac{1}{1-x}$$

existiert nicht für $x \geq 1$. Die Lösung verläßt den Schlauch S aber schon bei $x = b' = 0.9$. Bis dahin existiert die Lösung und ist dort eindeutig.

Satz 1.1.3 *Ist f auf dem Schlauch S stetig differenzierbar, so ist (1.1.9) erfüllt, da S abgeschlossen und beschränkt.*

Ganz allgemein beeinflußt die Glattheit von f auch die Glattheit der Lösung.

Definition 1.1.4 *Sei $F_N(a,b)$ die Menge aller Funktionen mit im Intervall $[a;b]$ stetigen und beschränkten partiellen Ableitungen bis zur Ordnung N .*[2]

Eine Funktion $f \in F_1(a,b)$ erfüllt dann die Bedingung von Satz 1.1.3

Bemerkung 1.1.5
In manchen Büchern betrachtet man unendliche **Streifen** $S = [a,b] \times \mathbb{R}^n$.
Gilt dann (1.1.9) in S , so existiert stets sogar eine Lösung $y \in C^1[a,b]$.
In der Praxis ist aber (1.1.9) für solche S meist nicht nachweisbar, selbst wenn die Lösung eindeutig ist.
Hat man dagegen erst einmal eine Lösung $y(x, x_0, y_0)$ numerisch berechnet oder sonstwie geschätzt, so läßt sich $\partial f / \partial y$ in der Umgebung dieser Lösung oft abschätzen und man erhält die Existenz und Eindeutigkeit a posteriori.
Daher geht man folgendermaßen vor:

1. Annahme der Existenz und Eindeutigkeit.
2. Berechnung einer numerischen Approximation $\eta(x)$ der Lösung.
3. $S_\varepsilon : \quad l_i(x) := \eta_i(x) - \varepsilon \leq \eta_i(x) \leq \eta_i(x) + \varepsilon \leq u_i(x)$.

[2]Funktionen deren partielle Ableitungen nur auf einem Schlauch S beschränkt sind, lassen sich außerhalb S mit beschränkten partielle Ableitungen fortsetzen. Verwendet man f also nur innerhalb eines Schlauchs S , auf dem f stetige, beschränkte partielle Ableitungen bis einschließlich Ordnung N besitzt, so kann man $f \in F_N(a,b)$ annehmen. Sonst wären die meisten folgenden Definitionen und Sätze bedeutungslos.

4. Falls $\|f_y(x,\eta(x)\|$ beschränkt in S_ε, $\Longrightarrow$ Existenz und Eindeutigkeit in S_ε.

Beispiel 1.1.6 $y' = f(y) = \sqrt[3]{y^2} = y^{\frac{2}{3}}$
Allgemeine Lösung: $y(x) = \frac{(x-a)^3}{27}$ oder $y(x) \equiv 0$ Jacobimatrix: $f_y(x,y) = \frac{2}{3}\frac{1}{\sqrt[3]{y}}$
$\Longrightarrow$ Satz 1.1.3 in Umgebung von $y = 0$ nicht anwendbar.

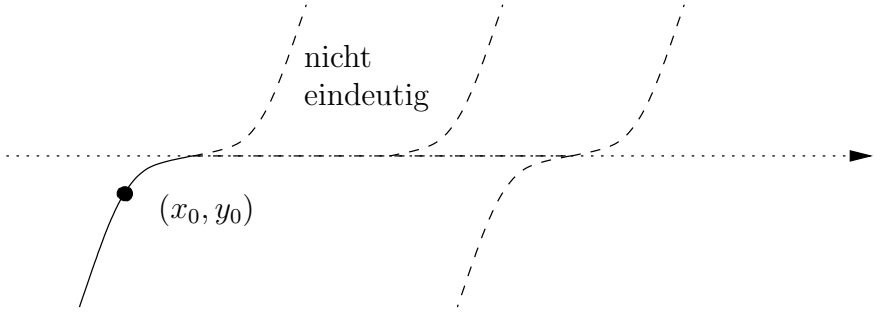


Abbildung 1.2: Lösung nur lokal eindeutig

Da die Anfangsdaten oft nur verfälscht vorliegen und die rechte Seite f oft mit kleinen Modellfehlern behaftet ist und auch nur ungenau ausgewertet werden kann, ist es wichtig zu wissen, wie dies die Lösung beeinflußt.

Lemma 1.1.7 (Fundamentallemma)
Sei $\eta(x)$ eine Approximation die ganz in einem Schlauchumgebung S der Lösung des AWPs $y' = f(x,y)$, $y(x_0) = y_0$ liegt mit

$$\|\eta(x_0) - y(x_0)\| \leq \rho \tag{1.1.10}$$
$$\|\eta' - f(x,\eta(x))\| \leq \varepsilon \tag{1.1.11}$$
$$\|f(x,\eta) - f(x,y)\| \leq L\,\|\eta - y\| \quad \forall (x,\eta),(x,y) \in S\,, \tag{1.1.12}$$

dann ist $e(x) := y(x) - \eta(x)$ beschränkt $\forall x \geq x_0$ durch:

$$\|e(x)\| \leq \rho e^{L(x-x_0)} + \frac{\varepsilon}{L}\left(e^{L(x-x_0)} - 1\right) \tag{1.1.13}$$

Beweis: Beweis siehe Fußnote[3] ∎

[3] Sei $\eta' =: g(x,\eta(x))$. Dann ist e Lösung des AWPs

$$e'(x) = y'(x) - \eta'(x), \quad e(x_0) = \rho_0 \quad \text{mit } |\rho_0| \leq \rho\,.$$

- Falls $f \in C^1(M)$ mit einer kompakten Menge M, so existiert eine Lipschitzkonstante $L := \max_{(x,y)\in M} \|f_y\|$.
- ε beschränkt den Modellfehler.
- Achtung falls auch in einer kleinen Umgebung keine Lipschitz-Bedingung erfüllt ist!

Beispiel 1.1.8

$$y' = f(y) = 2\mathrm{sgn}(y)\sqrt{|y|} \left(\stackrel{y\neq 0}{=} 2\frac{\sqrt{|y^3|}}{y} \right) \quad , \quad y(0) = 0$$

In einer Umgebung der Lösung $y = 0$ existiert keine *Lipschitz*konstante da $f_y = \mathrm{sgn}(y)/\sqrt{y}$ unbeschränkt ist. Die Lösung ist auch tatsächlich nicht eindeutig.

$$y(x) = \begin{cases} 0 & \text{für } x \leq t \\ \pm(x-t)^2 & \text{für } x > t \end{cases}$$

ist für jedes beliebige $t \geq 0$ auch eine Lösung.
Kleine Änderungen der Anfangswerte haben daher großen Einfluß auf die Lösung.
$y(0) = \varepsilon \Rightarrow y(x) = (x - \sqrt{|\varepsilon|})^2$ aber $y(0) = -\varepsilon \Rightarrow y(x) = -(x - \sqrt{|\varepsilon|})^2$.

Wegen (1.1.11) and (1.1.12) ist $e'(x)$ beschränkt durch

$$\begin{aligned} \|e'(x)\| &= \|f(x,y) - g(x,\eta)\| = \|f(x,y) - f(x,\eta) + f(x,\eta) - g(x,\eta)\| \\ &\leq L\|\eta - y\| + \varepsilon = L\|e\| + \varepsilon \ . \end{aligned}$$

Wir vergleichen e mit der Lösung u_δ der skalaren linearen Differentialgleichung

$$u'_\delta(x) = Lu_\delta(x) + \varepsilon \in \mathbb{R} \quad , \quad u_\delta(x_0) = \rho + \delta$$

mit $\delta > 0$. Die Lösung ist gegeben durch:

$$u_\delta(x) = (\rho + \delta)e^{L(x-x_0)} + \frac{\varepsilon}{L}\left(e^{L(x-x_0)} - 1\right) \ .$$

Daher erhalten wir

$$\left.\begin{array}{l} u_\delta(x_0) > \|e(x_0)\| \\ u_\delta(x) > \|e(x)\| \Rightarrow u'_\delta(x) > \|e'(x)\| \end{array}\right\} \Rightarrow u_\delta(x) > \|e(x)\| \text{ für alle } x \geq x_0 \ .$$

Die Behauptung erhält man für $\delta \to 0$.

$$\|e(x)\| \leq \lim_{\delta\to 0} u_\delta(x) = \rho e^{L(x-x_0)} + \frac{\varepsilon}{L}\left(e^{L(x-x_0)} - 1\right)$$

Meistens ist die Lösung $y(x, x_0, y_0, p)$ sogar differenzierbar abhängig von den Anfangswerten y_0 und dem Startpunkt x_0 und Parametern p.

Lemma **??** macht deutlich, daß lineare Differentialgleichungen bei der Stabilitätsanalyse nichtlinearer Differentialgleichungen eine wesentliche Rolle spielen. Jedes nichtlineare autonome System $y' = f(y)$ kann in einer Umgebung der Lösung $\varphi(x)$ für kurze Zeit durch ein lineares System approximiert werden.

$$y'(x) = f(\varphi) + f_y(\varphi_0)(y - \varphi) + \mathcal{O}(y - \varphi)^2 \ .$$

$\bar{y} := y - \varphi$ genügt daher der Differentialgleichung

$$\bar{y}'(x) \approx f_y(\varphi_0)\bar{y} = A\bar{y} \ .$$

Daher ist die Lösungstheorie für lineare Differentialgleichungen mit konstanten Koeffizienten auch für die Analyse allgemeiner nichtlinearer Differentialgleichungssysteme von großer Bedeutung, insbesonderer bei Stabilitätsuntersuchungen.

Lineare Differentialgleichungen mit konstanten Koeffizienten:
Gegeben sei

$$y' = Ay \tag{1.1.14}$$

mit $y \in \mathbb{R}^n$, $A \in \mathbb{R}^{(n,n)}$. Es existiert dann stets eine nicht-singuläre Matrix T mit $T^{-1}AT = \mathrm{diag}(J_1, \ldots, J_m)$ **Jordan'sche Normalform** mit **Jordan-Blocks**

$$J_i = \begin{bmatrix} \lambda_i & 1 & & \cdots & 0 \\ 0 & \lambda_i & \ddots & & \vdots \\ & \ddots & \ddots & \ddots & \\ \vdots & & \ddots & \ddots & 1 \\ 0 & \cdots & & 0 & \lambda_i \end{bmatrix}$$

der Dimension d_i mit $d_1 + d_2 + \cdots + d_m = n$ und den Eigenwerten λ_i von A.

An Stelle von (1.1.14) betrachten wir nun $w = T^{-1}y$ und

$$w' = T^{-1}y' = T^{-1}Ay = T^{-1}ATw = \mathrm{diag}(J_1, \ldots, J_m)w \ .$$

Sie zerfällt in m unabhängige Gleichungen

$$w_i' = J_i w_i \quad , \quad i = 1, \ldots, m$$

mit $w_i \in \mathbb{R}^{d_i}$. Die Lösung ist gegeben durch:

$$w_i = F_i w_i(x_0) := \begin{bmatrix} 1 & x & x^2/2 & \cdots & \frac{x^{d_i-1}}{(d_i-1)!} \\ 0 & 1 & x & & \vdots \\ & \ddots & \ddots & \ddots & x^2/2 \\ \vdots & & \ddots & \ddots & x \\ 0 & \cdots & & 0 & 1 \end{bmatrix} e^{\lambda_i(x-x_0)} w_i(x_0) \ .$$

Im Falle eines positiven Realteils von λ_i , also für $\text{Re}\,(\lambda_i) > 0$ wächst $\|w_i\|$ exponentiell über alle Grenzen. Für $\text{Re}\,(\lambda_i) < 0$ dagegen konvergiert w_i assymptotisch gegen Null. Der Exponentialterm bestimmt also das Wachstumsverhalten von $\|w_i\|$. Nur im Falle $\text{Re}\,(\lambda_i) = 0$ werden die Polynomeinträge der Matrix bedeutsam. In diesem Fall spricht man von polynomialer Instabilität.
Die Lösung von (1.1.14) ist dann gegeben durch

$$y(x) = T\text{diag}(F_1, \ldots, F_m)T^{-1}y_0 \ .$$

Die Eigenwerte von A beschreiben also das Wachstum der Lösung des linearen Systems. Das wird entscheidend sein bei der Diskussion der Stabilität und der Behandlung sogenannter steifer Differentialgleichungen.

1.2 Einschrittverfahren – Einführung

Wir wenden unser Augenmerk nun auf numerische Methoden zur näherungsweisen Lösung gewöhnlicher Differentialgleichung erster Ordnung. Dabei wird die Lösung nicht für alle x, sondern nur an diskreten Stellen x_i approximiert. Man bestimmt also Näherungen $\eta_i := \eta(x_i)$ der exakten Lösungen $y_i := y(x_i)$ an diskreten Stellen x_i, $i = 0, 1, \ldots$. Im einfachsten Fall sind die x_i äquidistant, $x_i = x_0 + i\,h$. In diesem Fall schreiben wir $\eta_i = \eta(x_i; h)$, da η_i und x_i von der *Schrittweite* h abhängen. Eine der wichtigsten Fragen ist dann, ob und wie schnell die Näherungen gegen $y(x)$ konvergieren, falls $h \to 0$.
Wir erhalten dabei nur Näherungen $\eta(x; h)$ an diskreten Punkten

$$x \in R_h := \{x_0 + i\,h| \quad i = 0, 1, 2, \ldots\} ,$$

bzw. an beliebig vorgegebenen Punkten für diskrete Schrittweiten

$$h \in H_x := \left\{ \frac{x - x_0}{n} | \quad n = 1, 2, \ldots \right\} .$$

Wir nehmen im Folgenden an, das Anfangswertproblem

$$y' = f(x, y) , \quad y(x_0) = y_0 \tag{1.2.1}$$

sei immer eindeutig lösbar.

Beispiel 1.2.1 (Euler-Verfahren)
Idee: $y(x)$ gegeben, $y(x + h)$ gesucht.
Taylorentwicklung liefert falls $y \in C^2$:

$$y(x + h) \approx y(x) + hy'(x) + \frac{h^2}{2} y''(x + \xi) = y(x) + hf(x, y(x)) + \mathcal{O}(h^2)$$

Zu Gitter $x_0, x_1, \ldots \quad x_i = x_0 + ih$
berechne Näherungen $\eta(x_i, h) \approx y(x_i)$ gemäß:

$$\begin{aligned} \eta(x_0, h) &:= y_0 \\ \eta(x_{i+1}, h) &:= \eta(x_i, h) + hf(x_i, \eta(x_i, h)) . \end{aligned} \tag{1.2.2}$$

Nach Wahl einer Schrittweite $h \neq 0$ sind ausgehend von den Anfangswerten x_0, und $y_0 = y(x_0)$ Approximationen $\eta_i \approx y_i = y(x_i)$ an äquidistanten Punkten $x_i = x_0 + i\,h$, $i = 1, 2, \ldots$ bestimmt.

Algorithmus des Euler Polygonzugverfahrens.

`Gegeben:` $x_0,\ y_0,\ h$

`Start:` $\eta_0 := y_0$

`Rekursion für` $i = 0, 1, 2, \ldots$:

$$\eta_{i+1} := \eta_i + h\, f(x_i, \eta_i)$$
$$x_{i+1} := x_i + h$$

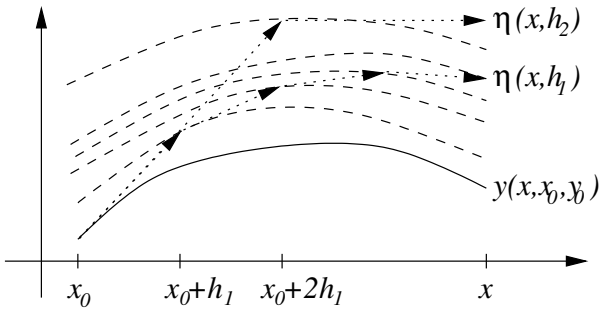


Abbildung 1.3: Euler Polygonzugverfahrens

Eine andere Herleitung verwendet die Identität

$$y(x+h) = y(x) + \int_x^{x+h} f(t, y(t))\mathrm{d}t \tag{1.2.3}$$

und als Integralnäherung die *Rechteckregel linker Rand* als Quadraturformel.

$$y(x+h) = y(x) + hf(x, y(x)) + \mathcal{O}(h^2) \tag{1.2.4}$$

Beispiel 1.2.2 (implizites Euler-Verfahren) Verwendet man zur Approximation des Integrals in (1.2.3) die Rechteckregel am rechten Rand, so erhält man:

$$y(x+h) = y(x) + \int_x^{x+h} f(t, y(t))\mathrm{d}t = y(x) + hf(x+h, y(x+h)) + \mathcal{O}(h^2) \tag{1.2.5}$$

Bei Vernachlässigung des Terms $\mathcal{O}(h^2)$ erhält man die Rekursion des impliziten Euler-Verfahrens.

Beispiel 1.2.3 (semi-implizites Euler-Verfahren)
Das implizite Euler-Verfahren führt auf eine i.a. nichtlineare Bestimmungsgleichung zur Bestimmung von $\eta_i \approx y(x_i)$.

$$\eta_{i+1} := \eta_i + hf(x_i + h, \eta_{i+1})$$

bzw.

$$\eta_{i+1} - \eta_i - hf(x_i + h, \eta_{i+1}) =: F(\eta_{i+1}) \stackrel{!}{=} 0 \ .$$

Verwendet man zur Lösung dieser Gleichung das Newtonverfahren und $\eta_{i+1}^{(0)} := \eta_i$ als Startwert, so erhält man im ersten Newtonschritt

$$\begin{aligned} [I - hf_y]\Delta\eta_{i+1}^{(0)} &= -F(\eta_{i+1}^{(0)}) = hf(x_i + h, \eta_{i+1}^{(0)}) = hf(x_i + h, \eta_i) \\ \eta_{i+1}^{(1)} &= \eta_i + \Delta\eta_{i+1}^{(0)} \ . \end{aligned}$$

Bricht man an dieser Stelle ab, so erhält man das **semi-implizite Euler-Verfahren**.

$$[I - hf_y](\eta_{i+1} - \eta_i) = hf(x_i + h, \eta_i)$$

Beispiel 1.2.4 (implizite Trapezregel) Die Trapezregel führt in (1.2.3) auf

$$y(x+h) = y(x) + \int_x^{x+h} f(t, y(t))\mathrm{d}t = y(x) + \frac{h}{2}(f(x, y(x)) + f(x+h, y(x+h))) + \mathcal{O}(h^3) \tag{1.2.6}$$

Beispiel 1.2.5 (Heun-Verfahren) In der impliziten Gleichung der Trapezregel wird das Argument $y(x+h)$ von f durch eine explizite Euler-Approximation ersetzt

$$\begin{aligned} y(x+h) &= y(x) + \frac{h}{2}(f(x, y(x)) + f(x+h, y(x+h))) + \mathcal{O}(h^3) \\ &= y(x) + \frac{h}{2}(f(x, y(x)) + f(x+h, y(x) + hf(x, y(x)) + \mathcal{O}(h^2))) + \mathcal{O}(h^3) \\ &= y(x) + \frac{h}{2}(f(x, y(x)) + f(x+h, y(x) + hf(x, y(x)))) + \mathcal{O}(h^3) \end{aligned}$$

Die angegebenen Verfahren sind typische Einschrittverfahren.

Definition 1.2.6 (Einschrittverfahren)
Ein Verfahren bei dem die numerischen Näherungen $\eta(x_i, h)$ durch eine Rekursionsformel der Art

$$\eta(x_{i+1}, h) := \eta(x_i, h) + h\Phi(x_i, \eta(x_i), h, f) \ . \tag{1.2.7}$$

berechnet werden, heißt **Einschrittverfahren**. Φ *heißt* **Inkrementfunktion** . *Oft wird* Φ *mit dem Verfahren identifiziert.*

Beispiel 1.2.7 (Inkrementfunktion des expliziten Euler-Verfahrens)

$\Phi(x, y, h, f) = f(x, y(x))$.

Bemerkung: Φ muß nicht explizit (als analytischer Ausdruck von $(x_i, \eta(x_i), h, f)$) gegeben sein.

Beispiel 1.2.8 (Inkrementfunktion des impliziten Euler-Verfahrens)

$$\begin{aligned} \eta(x_{i+1}, h) &:= \eta(x_i, h) + hf(x_{i+1}, \eta(x_{i+1}), h, f) \ . \qquad (1.2.8) \\ &=: \eta(x_i, h) + h\Phi(x_i, \eta(x_i), h, f) \\ \Rightarrow \Phi(x_i, \eta(x_i), h, f) &= f(x_{i+1}, \eta(x_i, h) + h\Phi(x_i, \eta(x_i), h, f)) \end{aligned}$$

Nach dem Satz über implizite Funktionen ist damit Φ eindeutig bestimmt, falls

$$\frac{\partial}{\partial \Phi}[\Phi(x_i, \eta(x_i), h, f) - f(x_{i+1}, \eta(x_i, h) + h\Phi(x_i, \eta(x_i), h, f))] = I - hf_y(x_{i+1}, \eta(x_i, h)$$

regulär. Dies ist für $\|f_y\|$ beschränkt und h klein genug sicher erfüllt.

Welche Eigenschaften muß nun aber Φ haben, damit (1.2.7) eine möglichst gute Lösung von (1.1.1), (1.1.2) ist?

Definition 1.2.9 *Es sei* x , y *beliebig aber fest.* $z(t)$ *sei die exakte Lösung des AWPs*

$$z'(t) = f(t, z(t)), \quad z(x) = y \ . \qquad (1.2.9)$$

Die Funktion

$$\Delta(x, y, h, f) := \begin{cases} \frac{z(x+h)-y}{h} & \text{falls } h \neq 0 \\ f(x, y) & \text{falls } h = 0 \end{cases} \qquad (1.2.10)$$

heißt **exaktes relatives Inkrement**.

Sie gibt die Steigung der Sekante an diejenige Kurve der Lösungsschar der Differentialgleichung an, die durch $(x, z(x))$ und $(x + h, z(x + h))$ geht.

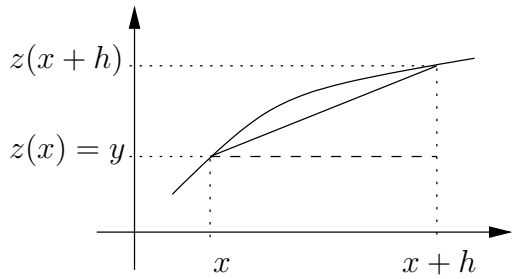


Abbildung 1.4: Exaktes relatives Inkrement

Wegen

$$z(x+h) = y + h\Delta(x, y, h, f) \tag{1.2.11}$$

sollte Φ möglichst gut mit Δ übereinstimmen.

Definition 1.2.10

$$\tau(x, y, h, f) := \Delta(x, y, h, f) - \Phi(x, y, h, f) \tag{1.2.12}$$

heißt **lokaler Diskretisierungsfehler**.

$$le(x, y, h, f) := h\tau(x, y, h, f) \tag{1.2.13}$$

heißt **lokaler Fehler**.

$h\tau$ gibt an, wie gut die exakte Lösung der Differentialgleichung durch den Punkt (x, y) die Differenzengleichung des Einschrittverfahrens (1.2.7) erfüllt, bzw. wie weit die numerische Lösung $\eta(x+h)$ von der exakten Lösung durch die letzte numerische Approximation $\eta(x)$ abweicht.

$$\begin{aligned}\tau(x, y, h, f) &= \frac{z(x+h) - y}{h} - \Phi(x, y, h, f) = \frac{z(x+h) - y - h\Phi(x, y, h, f)}{h} \\ &= \frac{z(x+h) - \eta(x+h)}{h} = \frac{\text{lokaler Fehler in einem Schritt}}{h}\end{aligned}$$

"local error per unit step." Der lokale Fehler ist also um eine h-Ordnung kleiner als der lokale Diskretisierungsfehler.
Man wird hoffen und erwarten, daß τ für kleine Schrittweiten h schnell klein wird.

Definition 1.2.11
Ein 1-Schritt-Verfahren heißt **konsistent** *falls gilt:*

$$\lim_{h\to 0} \tau(x,y,h,f) = 0 \quad \forall x \in [a,b],\ y \in \mathbb{R}^n,\ f \in F_1(a,b)\ . \tag{1.2.14}$$

Wegen (1.2.10) und (1.2.12) erhält man:

Folgerung 1.2.12 *Ein Einschrittverfahren ist genau dann konsistent, falls gilt:*

$$\lim_{h\to 0} \Phi(x,y,h,f) = f(x,y) \quad \forall f \in F_1(a,b)\ . \tag{1.2.15}$$

Definition 1.2.13
Ein 1-Schritt-Verfahren heißt **konsistent von der Ordnung** p *(hat* **Konsistenzordnung** p*) falls gilt:*

$$\tau(x,y,h,f) \le \delta(h,f) = \mathcal{O}(h^p) \quad \forall f \in F_p(a,b)\ . \tag{1.2.16}$$

Die Konsistenz eines Verfahrens kann man nachweisen durch Taylor-Entwicklung von Φ und Δ und Koeffizientenvergleich. Zunächst entwickelt man die exakte Lösung

$$\begin{aligned}
z(x+h) &= z(x) + hz'(x) + \frac{h^2}{2!}z''(x) + \cdots + \frac{h^{p+1}}{(p+1)!}z^{(p+1)}(x+\xi h) \quad \text{mit } 0<\xi<1\\
z(x) &= y\ ,\\
z'(x) &= f(x,y(x))\ ,\\
z''(x) &= \frac{\mathrm{d}}{\mathrm{d}x}f(x,y(x)) = f_x(x,y) + f_y(x,y)\frac{\mathrm{d}}{\mathrm{d}x}y(x) = f_x(x,y) + f_y(x,y)f(x,y)\\
z'''(x) &= f_{xx}(x,y) + f_{xy}(x,y)f(x,y) + f_{yx}(x,y)f(x,y) + f_{yy}(x,y)f(x,y)f(x,y) +\\
&\quad + f_y(x,y)f_x(x,y) + f_y(x,y)f_y(x,y)f(x,y)
\end{aligned}$$

$\Longrightarrow$

$$\begin{aligned}
\Delta &= (z(x+h)-y)/h = z'(x) + \frac{h}{2!}z''(x) + \frac{h^2}{3!}z'''(x) + \mathcal{O}(h^3) = \quad (1.2.17)\\
&= f + \frac{h}{2}[f_x + f_y f] + \frac{h^2}{6}[f_{xx} + 2f_{xy}f + f_{yy}f^2 + f_y f_x + f_y^2 f] + \mathcal{O}(h^3)
\end{aligned}$$

Der Rechenaufwand für diese Entwicklung wächst schnell. Untersucht man speziell nur autonome Systeme $y' = f(y)$ [4] so vereinfacht sich die Entwicklung ganz wesentlich und wir erhalten:

$$\begin{aligned}
z(x) &= y\ , \\
z'(x) &= f(y)\ , \\
z''(x) &= f_y(y)f(y) \\
z'''(x) &= f_{yy}ff + f_y^2 f \\
z''''(x) &= f_{yyy}fff + f_{yy}(f_y f)f + f_{yy}f(f_y f) + f_{yy}f(f_y f) + f_y f_{yy}ff + f_y f_y f_y f = \\
&= f_{yyy}fff + 3f_{yy}(f_y f)f + f_y f_{yy}ff + f_y f_y f_y f
\end{aligned}$$

Bemerkung 1.2.14 Man beachte, daß für $y \in \mathbb{R}^n$ der Term f_{yy} ein Tensor 3-ter Stufe ist und z.B. $f_{yy}ff$ ein Vektor mit i-ter Komponente

$$(f_{yy}ff)_i = \sum_{j=1}^{n}\sum_{k=1}^{n} \frac{\partial^2 f_i}{\partial y_j \partial y_k} f_j f_k$$

$f_{yy}ff$ ist also eigentlich $f_{yy}(f, f)$ zu lesen.
In $z^{(4)}$ treten dabei die Terme $f_{yy}f_yff = f_{yy}(f_y f, f)$ und $f_y f_{yy}(f, f)$ auf, die im allgemeinen verschieden sind. Bei Verfahren höherer Ordnung genügt es daher nicht, nur skalare Probleme zu betrachten.

Beispiel 1.2.15 (Konsistenz des expliziten Euler-Verfahrens)

$$\begin{aligned}
\Phi(x, y, h, f) &= f(x, y) \\
\Longrightarrow \tau(x, y, h, f) &= \Delta - \Phi = \frac{h}{2}[f_x + f_y f]_{(x,y)} + \mathcal{O}(h^2) = \mathcal{O}(h^1)
\end{aligned}$$

Das Euler-Verfahren ist also konsistent von der Ordnung 1.

Beispiel 1.2.16 (Konsistenzordnung des Heun-Verfahrens)

$$\eta_{i+1} := \eta_i + \frac{h}{2} f(x_i, \eta_i) + \frac{h}{2} f(x_{i+1}, \eta_i + hf(x_i, \eta_i)) \tag{1.2.18}$$

[4] Jedes System läßt sich autonom machen

$$\begin{aligned}\Phi(x,y,h,f) &= \frac{1}{2}f(x,y)+\frac{1}{2}f(x+h,y+hf(x,y)) = \\ &= \frac{1}{2}f+\frac{1}{2}[f+f_x h+f_y hf+f_{xx}\frac{h^2}{2}+f_{yy}\frac{h^2}{2}f^2+f_{xy}h^2 f]+\mathcal{O}(h^3) \\ &= f+\frac{h}{2}[f_x+f_y f]+\frac{h^2}{4}[f_{xx}+f_{yy}f^2+2f_{xy}f]+\mathcal{O}(h^3)\end{aligned}$$

$$\overset{(1.2.17)}{\Longrightarrow} \tau = \Delta - \Phi = \mathcal{O}(h^2)$$

Das Heun-Verfahren ist also konsistent von der Ordnung 2.

Bei der Konstruktion von Verfahren hoher Ordnung muß dann ein großes System nichtlinearer Gleichungen möglichst exakt gelöst werden.

Konvergenz von Einschrittverfahren:

Zur Bestimmung einer Näherung am Punkt $x \in [a,b]$ werde ein Einschrittverfahren mit Schrittweite $h_n = \frac{x-x_0}{n}$ verwendet. Die numerische Näherung $\eta(x,h_n)$ hängt also von h_n ab. Man hofft, daß diese für $n \to \infty$ gegen die exakte Lösung konvergiert.

Definition 1.2.17 *Die Funktion*

$$e(x,h) := \eta(x,h) - y(x) \tag{1.2.19}$$

heißt **globaler Diskretisierungsfehler**.

($e(x,h)$ ist nur für diskrete $h = h_n = \frac{x-x_0}{n}$ definiert.)

Definition 1.2.18

Sei $e_j(h) = \eta_j - y(x_j)$ *und* $E(h) := \max_{j=0,\ldots,n(h)} \|e_j(h)\|$ *dann heißt ein Einschrittverfahren* **konvergent**, *falls*

$$\lim_{h\to 0} E(h) = 0 \Rightarrow \lim_{n\to\infty} e(x,h_n) = 0 \quad \textit{bzw.} \quad \lim_{h\to 0} e(x,h) = 0\ . \tag{1.2.20}$$

In jedem Integrationsschritt wird nun ein lokaler Fehler $l_i = h_i\tau_i(x_i,\eta_i,h_i)$ begangen. Er beschreibt wie weit das numerische Rekursionsschema von der lokalen Lösung der Differentialgleichung abweicht. Der Fehler verstärkt sich in den folgenden Schritten zu einem Fehler E_i am Ende. Am Ende erhält man den Fehler $e(x,h) = \sum_i E_i$.

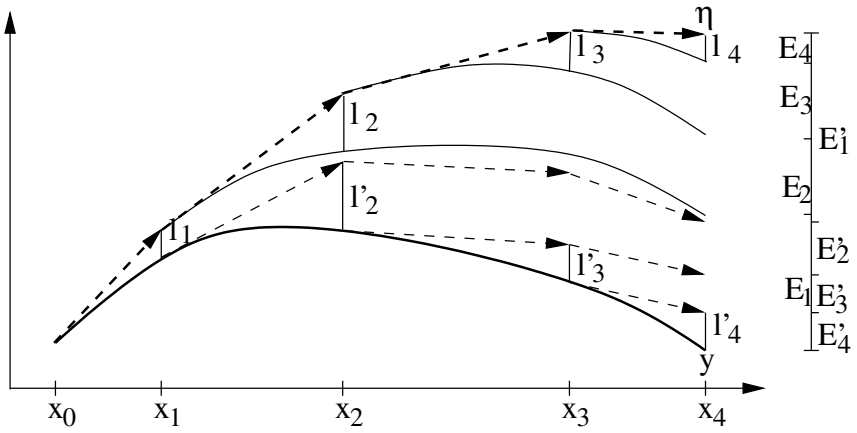


Abbildung 1.5: Globaler Diskretisierungsfehler

Abbildung 1.5 zeigt die Situation für das Euler-Vefahren. Lösungen der Differentialgleichungen sind durchgezogen dargestellt, die gestrichelten Pfeile bezeichnen Schritte des Eulerverfahrens ausgehend von verschiedenen Startpunkten. Die zu (x_0, y_0) gehörende Lösung und numerische Näherung sind jeweils fett gezeichnet.
Statt der lokalen Fehler $l_i := h_i\tau_i(x_i, \eta_i, h_i)$ an den numerischen Approximationsstellen kann man auch die lokalen Fehler $l'_i := h_i\tau_i(x_i, y(x_i), h_i)$ entlang der Lösung betrachten. Sie beschreiben wie gut die Lösung der Differentialgleichung das numerische Rekursionsschema erfüllt, und verstärken sich bis zum Ende zu E'_i.
E_i beschreibt also die Fehlerverstärkung der Differenzenapproximation bei anschließender exakter Lösung der Differentialgleichung.
E'_i beschreibt die Fehlerverstärkung bei Approximation von η_i durch $y(x_i)$ und anschließender exakter Lösung der Differenzengleichung.
Dabei gilt $e(x, h) = \sum_i E_i = \sum_i E'_i$. Insbesondere bei den später zu beschreibenden Mehrschrittverfahren ist dieser Ansatz geeigneter.
Man kann nun zeigen, daß bei Einschrittverfahren aus Konsistenz auch die Konvergenz folgt:
Dazu benötigen wir ein Lemma, welches wir auch noch später verwenden können.

Lemma 1.2.19 (Gronwall) *Sei z_i eine Zahlenfolge mit*

$$|z_{j+1}| \leq (1+A)|z_j| + B$$

mit $A, B > 0$. Dann gilt:

$$|z_j| \leq |z_0|e^{jA} + \frac{B}{A}\left(e^{jA} - 1\right)$$

Beweis: durch Induktion: Behauptung klar für $j = 0$.
$j \to j+1$:

$$\begin{aligned}
|z_{j+1}| &\leq (1+A)|z_j| + B \\
&\leq (1+A)\left[|z_0|e^{jA} + \frac{B}{A}\left(e^{jA} - 1\right)\right] + B \\
&\overset{1+A<e^A}{\leq} |z_0|e^{jA+A} + \frac{B}{A}e^{jA+A} - (1+A)\frac{B}{A} + B \\
&= |z_0|e^{(j+1)A} + \frac{B}{A}e^{(j+1)A} - \frac{B}{A}
\end{aligned}$$

■

Satz 1.2.20 (Konvergenzsatz für Einschrittverfahren)
Sei y die Lösung von $y' = f(x,y)$; $y(x_0 = y_0)$; $x_0 \in [a,b]$; $f \in F_p$ und es gelte

(i) Φ *stetig auf* $G := \{(x,y,h) \,|\, a \leq x \leq b\,,\ \|y - y(x)\| \leq \gamma\,,\ 0 \leq |h| \leq h_0\}$.

(ii) $\exists M > 0$ *mit* $\|\Phi(x,y,h) - \Phi(x,\bar{y},h)\| \leq M\|y - \bar{y}\|$ $\forall (x,y,h),(x,\bar{y},h) \in G$.

(iii) $\exists N > 0$ *mit* $\|\tau(x,y(x),h)\| \leq N|h|^p \,\forall h \leq h_0$ *und* $p > 0$.

Dann gilt für ein $\bar{h} < h_0$ und alle $|h| < \bar{h}$

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

D.h., Konsistenzordnung und Konvergenzordnung sind gleich, insbesondere folgt aus der Konsistenz die Konvergenz.

Beweis: Wir wollen den Fehler $e_i := \eta_i - y(x_i)$ abschätzen.

$$\begin{aligned}
\eta_n &= \eta_{n-1} + h\Phi(x_{n-1}, \eta_{n-1}; h) \\
y_n &= y_{n-1} + h\Phi(x_{n-1}, y_{n-1}; h) + h\tau(x_{n-1}, y_{n-1}; h) \\
\Rightarrow e_n &= e_{n-1} + h[\Phi(x_{n-1}, \eta_{n-1}; h) - \Phi(x_{n-1}, y_{n-1}; h)] - h\tau(x_{n-1}, y_{n-1}; h) \\
\Rightarrow \|e_n\| &\leq \|e_{n-1}\| + hM\|e_{n-1}\| + hNh^p
\end{aligned}$$

Dies sind genau die Voraussetzungen von Lemma 1.2.19. Mit $A = hM$ und $B = hNh^p$. Daher gilt

$$\begin{aligned}
\|e_n\| &\leq \underbrace{\|e_0\|}_{=0} e^{nhM} + \frac{hNh^p}{hM}\left(e^{nhM} - 1\right) \\
&\leq \frac{Nh^p}{M}\left(e^{M|x_n - x_0|} - 1\right)
\end{aligned}$$

■

Bemerkung: Bei einem Verfahren der Ordnung p gilt also:
Lokaler Diskretisierungsfehler $\tau = \mathcal{O}(h^p)$.
Globaler Diskretisierungsfehler $e = \mathcal{O}(h^p)$ = **globaler Fehler**.
aber lokaler Fehler $\eta_{i+1} - z(x_{i+1}) = h\tau = \mathcal{O}(h^{p+1})$!

Bemerkung 1.2.21
Satz 1.2.20 läßt sich verallgemeinern für Iterationsverfahren $\eta_{n+1} = \Psi(x_n, \eta_n, h)$.
Im Beweis wird nur benötigt:

$$\|\Psi(x, y, h) - \Psi(x, \bar{y}, h)\| \leq (1 + hM)\|y - \bar{y}\| \ \forall (x, y, h), (x, \bar{y}, h) \in G \ .$$

Dies folgt für Einschrittverfahren $\Psi(x, y, h) := y + h\Phi(x, y, h)$ aus 1.2.20(ii).

Fehlerfortpflanzung: Wegen Lemma 1.1.7 werden lokale Fehler bei x_i höchstens um den Faktor $e^{L(x-x_i)}$ verstärkt. Insgesamt gilt für den globalen Fehler:

$$\|e(x, h)\| \leq \sum_{i=0}^{n-1} h_i\tau(x_i, \eta_i, h_i)e^{L(x-x_i)} \leq e^{L|x-x_0|}\sum_{i=0}^{n-1} h_i\tau(x_i, \eta_i, h_i)$$

Gelingt es in jedem Schritt für den lokalen Fehler

$$h_i\tau(x_i, \eta_i, h_i) \leq \frac{h_i\varepsilon}{|x - x_0|}e^{-L|x-x_0|}$$

zu erreichen, so gilt

$$\|e(x,h)\| \leq \sum_{i=0}^{n-1} \frac{h_i}{|x-x_0|}\varepsilon = \varepsilon \ .$$

Falls L nicht abgeschätzt werden kann, was in der Praxis meistens der Fall ist, ist diese Bedingung an $h_i\tau_i$ nicht überprüfbar. In den meisten Verfahren werden daher nur entsprechende heuristische Bedingungen zur Fehlerkontrolle gestellt. Etwa

$$h_i\tau_i \leq \frac{h_i\varepsilon}{|x-x_0|} \Rightarrow e(x,h) \leq \varepsilon e^{L|x-x_0|}$$

Gelingt es zudem nur

$$h_i\tau(x_i, y(x_i), h) \leq \frac{h_i\varepsilon}{|x-x_0|}$$

zu erreichen, also eine Abschätzung von τ entlang der exakten Lösung, so kann man nicht mit Lemma 1.1.7 argumentieren. Man benötigt dann ein analoges Lemma für das numerische Rekursionsschemas bezüglich der Anfangsdaten. Eine Voraussetzung ist eine Lipschitzbedingung für die Inkrementfunktion

$$\|\Phi(x,u,h) - \Phi(x,v,h)\| \leq M\|u-v\| \ .$$

Bei allen hier diskutierten Verfahren folgt diese Bedingung stets aus einer entsprechenden Lipschitzbedingung für f.

Bei vorgegebener Schrittweite besteht jedes Einschrittverfahren aus einer Folge differenzierbarer Rechenschritte, und ist daher bei entsprechender Differenzierbarkeit von f selbst bezüglich der Anfangsdaten entsprechend differenzierbar. Für den lokalen Fehler existiert daher bei festem h eine Taylorentwicklung

$$\begin{aligned} h\tau(x,y(x),h) &= y(x+h) - y(x) - h\Phi(x,y(x),h) \\ &= \sum_{i=p+1}^{N+1} g_i(x)h^i + \mathcal{O}(h^{N+2}) \ . \end{aligned} \tag{1.2.21}$$

Bemerkung 1.2.22 In vielen Anwendungen ist f nur abschnittsweise in $F_p[a,b]$. Dann gilt Satz 1.2.20 für jeden Abschnitt. Die numerische Integration muß dann aber an den Grenzen angehalten werden, d.h., die Abschnittsgrenzen müssen Knoten x_i der Integration werden. Dazu muß ihre Lage natürlich bekannt sein, was problematisch sein kann.

Beispiel 1: Die Dynamik beim Durchbruch der Schallmauer ändert sich schlagartig. Der genaue Zeitpunkt des Durchbruchs ist a priori nicht bekannt sondern ergibt sich erst während der Lösung der Differentialgleichungen. Die Integration muß an dieser Stelle angehalten werden. Dies erfordert eine aufwendige Lokalisation.
Beispiel 2: Kennlinien elektronischer Bauelemente sind oft nur stückweise definiert und global lediglich C^0 oder C^1. Bei der Simulation großer Schaltungen überschreitet fast ständig eines der Bauteile eine Unstetigkeitsstelle. Ein ständiges Anhalten ist hier nicht sinnvoll, so daß nur Verfahren sehr geringer Ordnung (1-3) eingesetzt werden können.
Beispiel 3: Kubische Splinapproximation an Daten liefert C^2 Funktionen. An jedem Knoten muß die Integration gestoppt werden, oder Verfahren niedriger Ordnung.

1.2.1 Explizite Runge-Kutta-Verfahren

Nach dem Prinzip des Verfahrens von Heun:

$$\begin{aligned} k_1 &= f(x,y) \\ k_2 &= f(x+1\cdot h, y+1\cdot hk_1) \\ \Phi &= \frac{1}{2}k_1+\frac{1}{2}k_2 \end{aligned}$$

lassen sich weitere Verfahren konstruieren.

Beispiel 1.2.23 Verallgemeinerung von (1.2.18)

$$\Phi(x,y,h,f) := b_1 f(x,y) + b_2 f(x+ch, y+ahf(x,y)) \ . \qquad (*)$$

Man könnte versuchen jetzt die Parameter b_1, b_2, c und a so zu wählen, daß die Konsistenzordnung maximal wird.
Taylorentwicklung von Φ ergibt mit (1.2.17)

$$\begin{aligned} \Phi &= b_1 f(x,y) + b_2 f(x,y) + b_2 ch f_x(x,y) + f_y(x,y) b_2 ahf(x,y) + \mathcal{O}(h^2) \\ &= (b_1+b_2) f(x,y) + b_2 ch f_x + b_2 ahf_y f + \mathcal{O}(h^2) \\ \Delta &= f + \frac{h}{2}[f_x + f_y f] + \frac{h^2}{6}[f_{xx} + 2f_{yx}f + f_{yy}f^2 + f_y f_x + f_y^2 f] + \mathcal{O}(h^3) \end{aligned}$$

Koeffizientenvergleich:

$$\begin{aligned} b_1 + b_2 &= 1 \\ b_2 c &= \frac{1}{2} = b_2 a \end{aligned}$$

$b_1 = 1 - b_2$; $c = a = \frac{1}{2b_2}$; b_2 frei wählbar.
$b_2 = \frac{1}{2} \longrightarrow$ Methode von Heun.
$b_2 = 1 \longrightarrow$ "Modifiziertes Euler-Verfahren" oder Mittelpunktsregel.
Man erhält für jedes b_2 ein Verfahren der Ordnung 2. Ordnung 3 ist nicht möglich, da z.B., der Term $\frac{h^2}{6} f_y^2 f$ in der Entwicklung von Φ nicht auftaucht.

Abarbeitung von (*)

$$\begin{aligned} k_1 &= f(x,y) \\ k_2 &= f(x + ch, y + ahf(x,y)) \\ \Phi &= b_1 k_1 + b_2 k_2 \end{aligned}$$

D.h., es werden pro Schritt 2 Auswertungen von f benötigt.

Verallgemeinerung: Mit 4 Auswertungen pro Schritt erhält man bereits das bekannte (klassische) Runge-Kutta Verfahren 4-ter Ordnung:

$$\begin{aligned} k_1 &= f(x,y) \\ k_2 &= f(x + \frac{1}{2}h, y + \frac{1}{2}hk_1) \\ k_3 &= f(x + \frac{1}{2}h, y + \frac{1}{2}hk_2) \\ k_4 &= f(x + h, y + hk_3) \\ \Phi &= \frac{1}{6}k_1 + \frac{1}{3}k_2 + \frac{1}{3}k_3 + \frac{1}{6}k_4 \end{aligned}$$

Man prüft leicht Konsistenz nach:

$$\Phi = \frac{1}{6}f(x,y) + \frac{1}{3}f(x,y) + \mathcal{O}(h) + \frac{1}{3}f(x,y) + \frac{1}{6}f(x,y) = f(x,y) + \mathcal{O}(h) \ .$$

Bemerkung 1.2.24 Man beachte, daß im Falle $y \in \mathbb{R}^n$ f_y eine Matrix ist und z.B. f_y und f nicht vertauschbar sind. Der Nachweis hoher Konsistenzordnung erfordert daher den Umgang mit Tensoren der Stufe p.

Noch allgemeiner:

Definition 1.2.25 (Explizite Runge-Kutta-Verfahren)
ein Einschrittverfahren mit der Inkrementfunktion

$$\Phi(x,y;h) := \sum_{j=1}^{s} b_j \, k_j(x,y;h)$$

und s *Auswertungen der rechten Seite* f *pro Schritt*

$$\begin{aligned} k_1 &= f(x,y) \\ k_j &= f(x + c_j h, y + h \sum_{l=1}^{j-1} a_{jl}\, k_l) \quad j = 2, \ldots, s\ . \end{aligned}$$

heißt s**-stufiges explizites Runge-Kutta-Verfahren (ERK):** *Das Verfahren ist bestimmt durch die Koeffizienten des* **Runge-Kutta-Tableaus**

$$\begin{array}{c|cccc} c_1 & 0 & & & \\ c_2 & a_{2,1} & 0 & & \\ \vdots & \vdots & \ddots & 0 & \\ c_s & a_{s,1} & \cdots & a_{s,s-1} & 0 \\ \hline & b_1 & b_2 & \cdots & b_s \end{array}$$

In jeder Stufe ist also ein k_i durch Auswertung von f zu bestimmen.

Bemerkung 1.2.26 (Allgemeine Runge Kutta Verfahren) Ersetzt man bei der Berechnung von k_i $\sum_{l=1}^{j-1}$ durch $\sum_{l=1}^{j}$ oder $\sum_{l=1}^{s}$, so erhält man s implizite Gleichungen zur Berechnung der k_i oder ein implizites System zur Berechnung aller k_i. Das Tableau lautet dann

$$\begin{array}{c|cccc} c_1 & a_{1,1} & & & \\ c_2 & a_{2,1} & a_{2,2} & & \\ \vdots & \vdots & \ddots & \ddots & \\ c_s & a_{s,1} & \cdots & a_{s,s-1} & a_{s,s} \\ \hline & b_1 & b_2 & \cdots & b_s \end{array}$$

oder

$$\begin{array}{c|cccc} c_1 & a_{1,1} & a_{1,2} & \cdots & a_{1,s} \\ c_2 & a_{2,1} & a_{2,2} & \cdots & a_{2,s} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ c_s & a_{s,1} & \cdots & a_{s,s-1} & a_{s,s} \\ \hline & b_1 & b_2 & \cdots & b_s \end{array}$$

Reduktion der Konsistenzgleichugen: Die Koeffizienten werden in der Regel so gewählt, daß die Ordnung maximal wird. Um dabei die Anzahl der

Konsistenzgleichungen zu reduzieren, macht man apriori gewisse Zusatzeinschränkungen.

Die wesentliche Vereinfachung erhält man, wenn man fordert, daß das numerische Verfahren invariant ist gegen eine autonome Umformulierung eines Anfangswertproblems.

$$y' = f(x,y) \longrightarrow z' = \begin{pmatrix} y \\ x \end{pmatrix} = \begin{pmatrix} f(x,y) \\ 1 \end{pmatrix} = F(z)$$

$$\boxed{\Rightarrow \sum a_{i,j} = c_i} \; .$$

Es hat sich gezeigt, das dies keine wesentliche Beschränkung an die erreichbare Ordnung darstellt. Man kann sich dann aber bei den Konsistenzbedingungen auf autonome Probleme beschränken. Damit verschwinden alle partiellen Ableitungen von f nach x.

Die Konsistenz erfordert zudem sofort daß die Differentialgleichung $y' = 1$ exakt integriert wird, da in diesem Fall $\Phi = \sum b_i$ und $\Delta = 1$ von h unabhängig sind.

$$\boxed{\Rightarrow \sum b_i = 1} \; ,$$

Allgemein erhält man unter der vereinfachenden Annahme $\sum a_{i,j} = c_i$ als Konsistenzbedingung für Ordnung 1 bis p:

Ordnung	Bedingungen
1	$\sum b_i = 1$
2	$\sum b_i c_i = \frac{1}{2}$
3	$\sum b_i c_i^2 = \frac{1}{3}$ $\sum b_i a_{i,j} c_j = \frac{1}{6}$
4	$\sum b_i c_i^3 = \frac{1}{4}$ $\sum b_i c_i a_{i,j} c_j = \frac{1}{8}$ $\sum b_i a_{i,j} c_j^2 = \frac{1}{12}$ $\sum b_i a_{i,j} a_{j,k} c_k = \frac{1}{24}$
5	$\sum b_i c_i^4 = \frac{1}{5}$ $\sum b_i c_i a_{i,j} c_j^2 = \frac{1}{15}$ $\sum b_i c_i a_{i,j} a_{j,k} c_k = \frac{1}{30}$ $\sum b_i a_{i,j} c_j a_{i,k} c_k = \frac{1}{20}$ $\sum b_i a_{i,j} c_j^3 = \frac{1}{20}$ $\sum b_i a_{i,j} c_j a_{j,k} c_k = \frac{1}{40}$ $\sum b_i a_{i,j} a_{j,k} c_k^2 = \frac{1}{60}$ $\sum b_i a_{i,j} a_{j,k} a_{k,l} c_l = \frac{1}{120}$ $\sum b_i c_i^2 a_{i,j} c_j = \frac{1}{10}$

Die Anzahl q der zu lösenden Gleichungen und die Anzahl s nötiger Stufen steigt dabei auch wenn man sich auf autonome Probleme $y' = f(y)$ beschränkt, rasch mit der gewünschten Ordnung p.

p	1	2	3	4	5	6	7	8	9	10
q	1	2	4	8	17	37	85	200	486	1205
s	1	2	3	4	6	7	9	11	≥ 12	17

Die Herleitung von Bestimmungsgleichungen für Verfahren hoher Ordnung geschieht am besten mit Hilfe sogenannter Butcher-Bäume [**?**].
Oft wählt man zusätzlich $c_1 = 0$. Dies hat insbesondere Vorteile bei den später beschriebenen Fehlberg-Verfahren, und bei einer genauen Ausgabe auch an Zwischenpunkten durch Hermite-Interpolation z.B., zum Plotten.
Das klassische Runge-Kutta-Verfahren (RK4):

0				
$\frac{1}{2}$	$\frac{1}{2}$			
$\frac{1}{2}$	0	$\frac{1}{2}$		
1	0	0	1	
	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{6}$

hat Konsistenzordnung 4 (erhalten aus 11 Gleichungen für 13 Parameter).

1.3 Schrittweitensteuerung

Außer den Diskretisierungsfehlern macht man in jedem Schritt auch Rundungsfehler: Statt η_{i+1} berechnet man $\bar{\eta}_{i+1}$ mit $\|\bar{\eta}_{i+1} - \eta_{i+1}\| < C\varepsilon$. Dies entspricht einer Modelländerung von $f \to \bar{f}$ mit $\|\bar{f} - f\| \leq \frac{C\varepsilon}{h}$ und erzeugt nach Satz 1.1.7 einen zusätzlichen globalen Fehler von

$$\|\bar{\eta}(x,h) - \eta(x,h)\| \leq \frac{C\varepsilon}{h}\frac{1}{L}\left(e^{L(x-x_0)} - 1\right) = \mathcal{O}\left(\frac{\varepsilon}{h}\right) .$$

Einfluß der Schrittweite auf Rundungs- und Diskretisierungsfehler:

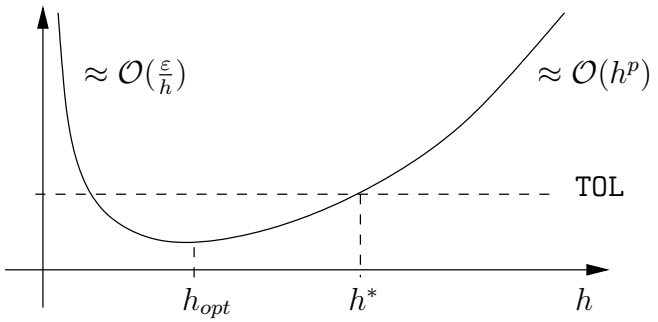


Abbildung 1.6: Optimale und effizienteste Schrittweite

$$\bar{e}(x,h) = \mathcal{O}(h^p) + \mathcal{O}(\varepsilon/h)$$

h zu groß $\Longrightarrow e(x,h) \approx \mathcal{O}(h^p)$ zu groß.
h zu klein $\Longrightarrow$ Rundungsfehlereinfluß $\approx \mathcal{O}(\frac{\varepsilon}{h})$ zu groß.
h_{opt} liefert genaueste Resultate.
Strategie: Wähle $h = h^*$ möglichst groß, so daß τ gerade noch die Genauigkeit einhält. Dabei ist das Problem, wie die Genauigkeit geschätzt werden kann.

1.3.1 Zwei Verfahren eine Schrittweite

Man kann zwei verschiedene Verfahren verschiedener Ordnung p und $p+1$ miteinander koppeln.

Sei

$$\begin{aligned} & \hat{\eta}(x+h,h) = \eta(x) + h\hat{\Phi}(x,\eta,h,f)\ ; \quad \hat{\tau} \le \hat{M}h^p \\ & \eta(x+h,h) = \eta(x) + h\Phi(x,\eta,h,f)\ ; \quad \tau \le Mh^{p+1} \\ \Longrightarrow \quad & \hat{\eta}(x+h,h) - z(x+h) = h\hat{\tau}(x) \\ & \eta(x+h,h) - z(x+h) = h\tau(x) \end{aligned}$$

Genauigkeitsschätzer für den lokalen Fehler von $\hat{\eta}(x+h,h)$:

$$\begin{aligned} Err_2 \ &:= \ \hat{\eta}(x+h,h) - \eta(x+h,h) = h\hat{\tau}(x) - h\tau(x) \doteq \\ &\doteq \ \hat{\eta}(x+h,h) - z(x+h) = h\hat{\tau}(x) \le Mh^{p+1}\ . \end{aligned} \tag{1.3.1}$$

Nachteil: Im Falle $\hat{M} >> M$ ist die Annahme $h\hat{\tau} - h\tau \approx h\hat{\tau}$ falsch.
Zur Effizienzsteigerung sucht man 2 Verfahren mit möglichst vielen gleichen Zeilen im RK-Tableau (eingebettete Verfahren).

Beispiel 1.3.1

$$\begin{array}{c|cc} 0 & 0 & \\ 1 & 0 & 0 \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \qquad \begin{array}{c|ccc} 0 & 0 & & \\ 1 & 0 & 0 & \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & \\ \hline & \frac{1}{6} & \frac{1}{6} & \frac{4}{6} \end{array}$$

ergibt

$$\begin{array}{c|ccc} 0 & 0 & & \\ 1 & 0 & 0 & \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & \\ \hline & \frac{1}{6} & \frac{1}{6} & \frac{4}{6} \\ & \frac{1}{2} & \frac{1}{2} & \end{array}$$

1.3.2 Ein Verfahren zwei Schrittweiten

Sei $z(x)$ die Lösung von $z' = f(x,z)$ mit $z(x_0) = \eta(x_0)$, Φ ein Verfahren der Ordnung p und

$$\eta(x+h,h) = \eta(x,h) + h\Phi(x,h)$$

die Iterationsvorschrift des ESV. Die Näherung habe einen globalen Fehler der Bauart:

$$\eta(x,h) - z(x) = e_p(x)h^p + e_{p+1}(x)h^{p+1} + \mathcal{O}(h^{p+2})\ ,$$

d.h., der Fehler läßt sich entwickeln (vgl. Kapitel Extrapolationsverfahren). Dann berechne man wie bei der Extrapolation von Quadraturformeln Näherungen zu zwei verschiedenen Schrittweiten. Daraus läßt sich eine weitere Näherung der Ordnung $p+1$ extrapolieren. Der Vergleich dieser drei Näherungen dient dann zur Abschätzung des Fehlers.

Beispiel 1.3.2

$$h_1 = h \; ; \quad h_2 = \frac{h}{2} \; ;$$

$$\eta(x+h,h_1) - z(x+h) = e_p(x+h)h^p + e_{p+1}(x+h)h^{p+1} \tag{1.3.2}$$
$$+\mathcal{O}(h^{p+2}) \tag{1.3.3}$$
$$\eta(x+h,h_2) - z(x+h) = e_p(x+h)\left(\frac{h}{2}\right)^p + e_{p+1}(x+h)\left(\frac{h}{2}\right)^{p+1} \tag{1.3.4}$$
$$+\mathcal{O}(h^{p+2})$$

$\Longrightarrow$

$$\eta(x+h,h) - \eta(x+h,\tfrac{h}{2}) = e_p(x+h)\,(\tfrac{h}{2})^p\,(2^p-1) + \mathcal{O}(h^{p+2}) \qquad (*)$$

da $e_{p+1}(x+h) \doteq \underbrace{e_{p+1}(x)}_{=0} + he'_{p+1}(x) = \mathcal{O}(h)$ ($\longrightarrow$ Satz 1.2.20).

$\Longrightarrow$ Genauigkeitsschätzer für den Fehler von $\eta(x+h,\frac{h}{2})$:

$$Err_1 := \frac{\eta(x+h,h) - \eta(x+h,\frac{h}{2})}{2^p-1}$$
$$\overset{(*)}{\doteq} e_p(x+h)\,(\tfrac{h}{2})^p + \mathcal{O}(h^{p+2}) \overset{(1.3.4)}{\doteq} \eta(x+h,\tfrac{h}{2}) - z(x+h) \tag{1.3.5}$$

Extrapolation: Ziehe Fehler Err_1 von $\eta(x+h,\frac{h}{2})$ ab (*) $\overset{(1.3.5)}{\Longrightarrow}$

$$\eta_{\text{ext}}(x+h,h) := \eta(x+h,\frac{h}{2}) - \frac{\eta(x+h,h) - \eta(x+h,\frac{h}{2})}{2^p-1} =$$
$$\overset{(1.3.5)}{=} z(x+h) + \mathcal{O}(h^{p+2}) \tag{1.3.6}$$

Man verwendet Err_1 als Genauigkeitsschätzer für $\eta(x+h,h/2)$ und rechnet mit $\eta_{\text{ext}}(x+h,h)$ weiter. Dadurch ist das Ergebnis oft genauer als verlangt.

Ein Nachteil dieser Methode ist der große Aufwand. Hat das Basisverfahren die Ordnung p und berechnet man eine weitere Näherung mit halber Schrittweite, so gewinnt man nur eine Ordnung, bei etwa dreimal soviel Funktionsaufrufen.

1.3.3 Schrittweitenwahl

Der Benutzer gibt eine gewünschte Genauigkeit TOL von y bei $x = x_f$ vor. Bei Schrittweite h werden $\frac{x_f - x_0}{h}$ Schritte benötigt.
Geht man davon aus, daß sich die lokalen Fehler zwar nicht verstärken,[5]aber doch addieren, so muß man mit einem globalen Fehler

$$\frac{x_f - x_0}{h} Err \overset{!}{<} \texttt{TOL}$$

rechnen $\Longrightarrow$ In jedem Schritt verlangt man:

$$Err \doteq h\tau \approx h\texttt{TOL} \quad \left(\text{bzw.} \frac{h\texttt{TOL}}{x_f - x_0}\right) \tag{1.3.7}$$

Für den lokalen Fehler eines Verfahrens der Ordnung p gilt:

$$e_p(x_0) = 0$$

$$Err(x+h) \doteq e_p(x+h)h^p \doteq e_p'(x)h^{p+1} \tag{1.3.8}$$

$$\overset{(1.3.7)}{\Longrightarrow} \frac{Err(x+h)}{h\texttt{TOL}} \doteq \frac{e_p'(x)h^p}{\texttt{TOL}} \overset{!}{\approx} 1 \tag{1.3.9}$$

Falls (1.3.9) nicht erfüllt ist wähle im nächsten Schritt h^* so daß:

$$\begin{aligned} Err(x+h^*) &\overset{(1.3.8)}{\doteq} e_p'(x)h^{*p+1} = e_p'(x)h^p \frac{h^{*p}}{h^p} h^* \\ &\overset{(1.3.9)}{=} \frac{Err(x+h)}{h} \frac{h^{*p}}{h^p} h^* \overset{!}{=} h^*\texttt{TOL} \end{aligned}$$

$$\Longrightarrow h^* := h \sqrt[p]{\frac{h\texttt{TOL}}{Err(x+h)}} \tag{1.3.10}$$

War $Err(x+h)$ zu groß, so muß der Schritt mit h^* wiederholt werden. Ist dagegen $Err(x+h) < h\texttt{TOL}$, so kann der nächste Schritt vermutlich mit Schrittweite h^* erfolgreich durchgeführt werden
Bemerkung: Fordert man statt (1.3.7)

[5]Lokale Fehler verstärken sich maximal um den Faktor $F := e^{L(x_f - x_0)}$ mit unbekanntem L. Dies ist der Unsicherheitsfaktor bei jeder numerischen Integration. Man wählt daher zur Sicherheit TOL meist etwas kleiner als benötigt. Dennoch liefert das keine Garantie für hinreichende Genauigkeit des Ergebnisses.

$$Err \leq \texttt{TOL} \tag{1.3.7'}$$

so erhält man statt (1.3.10)

$$h^* := h \sqrt[p+1]{\frac{\texttt{TOL}}{Err}} \tag{1.3.10'}$$

Zur Sicherheit wählt man h_{neu} noch etwas kleiner

$$h_{\text{neu}} = \alpha h^* \; ; \quad \text{mit } \alpha < 1 \; ,$$

und vermeidet auch eine zu extreme Schrittweitenreduktion oder Schrittweitenvergrößerung[6].Daher verlangt man z.B:

$$h_{\text{neu}} = \alpha h^* \in [\beta h_{\text{alt}}, \gamma h_{\text{alt}}] \; ; \quad \text{mit } \beta < 1 < \gamma \; .$$

Z.B. mit $\alpha = 0.9\,,\ \beta = 0.5\,,\ \gamma = 1.5\,.$

$$h_{\text{neu}} = \min\{\max\{0.5 h_{\text{alt}}, 0.9 h^*\}, 1.5 h_{\text{alt}}\} \tag{1.3.11}$$

[6]Der Sicherheitsfaktor $\alpha = 0.9$ führt z.B. dazu, daß etwa 10% mehr Aufwand getrieben wird, falls das Krümmungsverhalten von f nicht schlechter wird. In 50% der Fälle wird das Krümmungsverhalten von f jedoch schlechter. Ohne Sicherheitsfaktor würden daher etwa die Hälfte aller Schritte verworfen. Mit Sicherheitsfaktor müssen nur dann Schritte verworfen werden, wenn sich das Krümmungsverhalten von f massiv verschlechtert.

Ergibt die Schrittweitenschätzung dagegen $h^* = \gamma h_{\text{alt}}$ mit $\gamma > 1$ so muß der Fehler kleiner als TOL gewesen sein, und zwar wegen (1.3.10) bei einem Verfahren der Ordnung p:

$$Err < h\texttt{TOL}\gamma^{-p} \; .$$

Ist Err in der Größenordnung der Maschinengenauigkeit ε, so liegt dies jedoch oft nicht mehr nur an der Approximationsgüte, sondern auch an Rundungsfehlern, d.h., beide Approximationen werden auf den selben Wert gerundet. Dadurch wird eine zu große Genauigkeit vorgegaukelt. In diesem Fall verläßt man sich nicht mehr auf die Heuristik (1.3.10) und erhöht h vorsichtig. Eine Schranke für γ erhält man aus

$$\varepsilon = h\texttt{TOL}\gamma^{-p} \Longrightarrow \gamma = \sqrt[p]{\frac{h\texttt{TOL}}{\varepsilon}} \; .$$

Wegen (1.3.7) ist dabei $\frac{h\texttt{TOL}}{\varepsilon} > 1$ (sonst Programmabbruch TOL zu klein). Üblich ist etwa $\gamma = \sqrt[p]{10}$ oder $\gamma = 1.5 \approx \sqrt[6]{10}$.

Ergibt die Schrittweitenschätzung dagegen $h^* = \beta h_{\text{alt}}$ mit $\beta << 1$, so muß sich das Verhalten der Lösung plötzlich stark geändert haben. Starke Änderungen in der Schrittweite deuten darauf hin, daß die Information aus der Vergangenheit in Bezug auf die Schrittweitenwahl nicht besonders zuverlässig war. In diesem Fall ist es nicht unbedingt sinnvoll, sie in Zukunft als strikte Grundlage zu verwenden. In der Praxis beobachtet man sonst sehr oft stark schwankende Schrittweiten, mit sehr kleinen (erfolgreichen) und sehr großen (erfolglosen) Schritten. Häufig ist eine isolierte Problemstelle die Ursache.

Bemerkung 1.3.3 Zur Vermeidung von Überlauf in (1.3.10) rechnet man:
`if` $\gamma^p Err(x+h) \geq h$`TOL then` $h^* = \alpha\gamma h$
`elseif` $\beta^p Err(x+h) \geq h$`TOL then` $h^* = \alpha\beta h$
`else` $h^* = \alpha h \sqrt[p]{\frac{h\texttt{TOL}}{Err(x+h)}}$ `fi`

Bemerkung 1.3.4 Eine weitere Verbesserung erhält man, wenn man sich nach einem erfolgreichen Schritt die Schrittweite bis zum Abschluß des nächsten Schrittes speichert. Dann läßt sich verhindern, daß Schrittweiten die gerade erst verkleinert bzw. vergrößert wurden im nächsten Schritt gleich wieder vergrößert bzw. verkleinert werden.

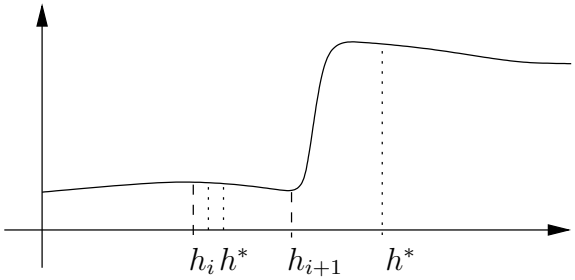


Abbildung 1.7: Isolierte Problemstelle

Die Lokalisierung einer solchen Stelle geschieht am besten durch Bisektion.

1.4 Mehrschrittverfahren

Die wesentliche Idee der Mehrschrittverfahren ist, mühsam gesammelte Informationen so weit wie möglich auszunutzen. Insbesondere sollen Informationen über Funktionswerte und Ableitungen der Lösung $y(x)$ einer Differentialgleichung ausgenutzt werden, auch wenn sie zu viel kleineren x-Werten gehören als momentan interessieren. Dies ist gleichzeitig einer der Nachteile, wenn sich die Dynamik schnell ändert.

1.4.1 Interpolatorische Verfahren

Einen ersten Ansatz erhält man aus der Integraldarstellung der Lösung an der Stelle $x_{r+1} = x_r + h$

$$y(x_{r+1}) = y(x_r) + \int_{x_r}^{x_r+h} f(t, y(t))\mathrm{d}t \ . \tag{1.4.1}$$

Sind Näherungen

$$\eta_i := \eta(x_i) \quad f(x_i, \eta_i) \quad i = 1, \ldots, r$$

gegeben, so liegt es nahe f in (1.4.1) durch das interpolierende Polynom p mit $p(x_i) = f(x_i, \eta_i)$ zu ersetzen. Dann erhält man die Rekursionsformel

$$\eta_{r+1} =: \eta_r + \int_{x_r}^{x_r+h} p(t)\mathrm{d}t = \eta_r + \sum_{i=1}^{r} f(x_i, \eta_i) \int_{x_r}^{x_r+h} L_i(t)\mathrm{d}t \ , \tag{1.4.2}$$

bei der r Näherungen bereits bekannt sein müssen.
Allgemeiner betrachtet man

$$y(x_{p+k}) - y(x_{p-j}) = \int_{x_{p-j}}^{x_{p+k}} f(t, y(t))\, dt \ . \tag{1.4.3}$$

Ersetzt man den Integrand in (1.4.3) durch das interpolierende Polynom P_q mit

$$P_q(x_i) = f(x_i, y(x_i)) \ , \quad i = p, p-1, \ldots, p-q \ ,$$

und $\operatorname{grad} P_q(x) \leq q$ so erhält man eine explizite (Lagrange) Darstellung von P_q

$$P_q(x) = \sum_{i=0}^{q} f(x_{p-i}, y_{p-i})\, L_i(x)$$

mit $y_i := y(x_i)$ und

$$L_i(x) := \prod_{\substack{l=0 \\ l \neq i}}^{q} \frac{x - x_{p-l}}{x_{p-i} - x_{p-l}} \ ,$$

sowie die Rekursionsformel

$$\begin{aligned} y_{p+k} - y_{p-j} &\approx \sum_{i=0}^{q} f(x_{p-i}, y_{p-i}) \int_{x_{p-j}}^{x_{p+k}} L_i(x)\, dx \\ &= h \sum_{i=0}^{q} \beta_{qi}\, f(x_{p-i}, y_{p-i}) \end{aligned}$$

wobei sich die auftretenden Faktoren β_{qi} bei konstanter Schrittweite für jedes Verfahren einmal vorab berechnen lassen.

$$\beta_{qi} := \frac{1}{h} \int_{x_{p-j}}^{x_{p+k}} L_i(x)\, dx = \int_{-j}^{k} \prod_{\substack{l=0 \\ l \neq i}}^{q} \frac{s+l}{-i+l}\, ds \ , \quad i = 0, 1, \ldots, q \ . \tag{1.4.4}$$

Ersetzt man in (1.4.4) die y_i durch die Näherungen η_i, so erhält man das Rekursionsschema

$$\eta_{p+k} = \eta_{p-j} + h \sum_{i=0}^{q} \beta_{qi}\, f_{p-i} \ , \quad f_i := f(x_i, \eta_i) \ . \tag{1.4.5}$$

Je nach Wahl von k, j, und q, erhält man dabei verschiedene Mehrschrittverfahren.
Für $k = 1$, $j = 0$, und $q = 0, 1, 2, \ldots,$ (vergleiche (1.4.1)) erhält man die sogenannten **Adams-Bashforth Verfahren**.

$$\eta_{p+1} = \eta_p + h\, [\beta_{q0} f_p + \beta_{q1} f_{p-1} + \cdots + \beta_{qq} f_{p-q}]$$

mit

$$\beta_{qi} := \frac{1}{h} \int_0^1 \prod_{\substack{l=0 \\ l \neq i}}^{q} \frac{s+l}{-i+l}\, ds \ , \quad i = 0, 1, \ldots, q \ .$$

Beispiel 1.4.1 Für $q = 0$, erhält man

$$\eta_{p+1} = \eta_p + h \int_{x_p}^{x_{p+1}} L_i(x)\, dx = \eta_p + h \int_{x_p}^{x_{p+1}} f(x_p, \eta_p)\, dx = \eta_p + h f(x_p, \eta_p) \ ,$$

also das explizite Euler-Verfahren als Spezialfall der Adams-Bashforth-Verfahren.

Für $k = 1$, $j = 1$, und $q = 0, 1, 2, \ldots$ erhält man die sogenannten **Nyström-Verfahren**.

$$\eta_{p+1} = \eta_{p-1} + h\,[\beta_{q0} f_p + \beta_{q1} f_{p-1} + \cdots + \beta_{qq} f_{p-q}]$$

mit

$$\beta_{qi} := \tfrac{1}{h} \textstyle\int_{-1}^{1} \prod_{\substack{l=0 \\ l \neq i}}^{q} \frac{s+l}{-i+l}\, ds\ , \quad i = 0, 1, \ldots, q\ ,$$

Beispiel 1.4.2 (Mittelpunktsregel)
Für $q = 0$ erhält man die sogenannte **Mittelpunktsregel**

$$\eta_{i+2} = \eta_i + 2h f(x_{i+1}, \eta_{i+1}) \tag{1.4.6}$$

Beide Verfahrensklassen sind explizit.
Für $k = 0$, $j = 1$, $q = 0, 1, 2, \ldots$ erhält man die *impliziten* **Adams-Moulton-Verfahren**

$$\eta_p = \eta_{p-1} + h\,[\beta_{q0} f(x_p, \eta_p) + \beta_{q1} f_{p-1} + \cdots + \beta_{qq} f_{p-q}] \tag{1.4.7}$$

bzw.—

$$\eta_{p+1} = \eta_p + h\,[\beta_{q0} f(x_{p+1}, \eta_{p+1}) + \beta_{q1} f_p + \cdots + \beta_{qq} f_{p+1-q}] \tag{1.4.8}$$

mit

$$\beta_{qi} := \frac{1}{h} \int_{-1}^{0} \prod_{\substack{l=0 \\ l \neq i}}^{q} \frac{s+l}{-i+l}\, ds\ , \quad i = 0, 1, \ldots, q\ .$$

Beispiel 1.4.3 Für $q = 0$ erhält man das implizite Euler-Verfahren

$$\eta_{i+1} = \eta_i + h f(x_{i+1}, \eta_{i+1})\ ,$$

und für $q = 1$ die implizite Trapezregel

$$\eta_{i+1} = \eta_i + h \left[\frac{1}{2} f(x_i, \eta_i) + \frac{1}{2} f(x_{i+1}, \eta_{i+1})\right]\ .$$

Die implizite Rekursionsgleichung (1.4.8) kann gleichzeitig als Fixpunktiterationsgleichung dienen:

$$\eta_{p+1}^{(i+1)} = \Phi(\eta_{p+1}^{(i)}) := \eta_p + h \left[\beta_{q0} f(x_{p+1}, \eta_{p+1}^{(i)}) + \beta_{q1} f_p + \cdots + \beta_{qq} f_{p+1-q}\right]$$

Für h klein genug und $f \in F_1(a,b)$ gilt

$$\|\Phi'(z)\| = \|h\beta_{q0}\frac{\partial f}{\partial y}(x_{p+1},z)\| \leq |h| \cdot \text{const} < 1 \ ,$$

und die Iteration ist kontraktiv für alle $f \in F_1(a,b)$ und hinreichend kleine h.
Sind alle $\eta_p, \eta_{p-1}, \ldots, \eta_{p+1-q}$ gegeben, so erhält man eine gute Startschätzung für $\eta_{p+1}^{(0)}$ durch irgendein anderes explizites Verfahren, z.B. ein Adams-Bashforth-Verfahren[7].Daher heißen die Adams-Bashforth-Verfahren auch **Prädiktor-Verfahren** und die Adams-Moulton-Verfahren **Korrektor-Verfahren**

1.4.2 BDF Verfahren

Eine andere Klasse von Funktionen erhält man, wenn man ein Polynom q_k konstruiert, welches $\eta_{p-k+1}, \ldots, \eta_{p+1}$ interpoliert und verlangt, daß es wenigstens an einem Punkt x_{p+j}, in der Regel dem neuen Punkt x_{p+1}, die Differentialgleichung erfüllt, d.h.,

$$q_k'(x_{p+j}) = \frac{\mathrm{d}}{\mathrm{d}x}\sum_{i=p-k+1}^{p+1}\eta_i L_i(x) \ . \tag{1.4.9}$$

Für $j = 1$ erhält man dann die impliziten sogenannten **BDF Verfahren** (*backward differentiation formulas*).

$$\begin{aligned} q_k(x) &:= \sum_{i=p-k+1}^{p+1}\eta_i L_i(x) \\ L_i(x) &:= \Pi_{\substack{i=p-k+1 \\ l\neq i}}^{p+1}\frac{x-x_l}{x_i-x_l} \\ q_k'(x_{p+1}) &:= f(x_{p+1}, q_k(x_{p+1}) \ . \end{aligned}$$

Dies ist eine Gleichung der Bauart

$$\sum_{i=p-k+1}^{p+1} a_i\eta_i = f(x_{p+1},\eta_{p+1}) \tag{1.4.10}$$

[7]Bei steifen Problemen ist oft $\eta_{p+1}^{(0)} := \eta_p$ besser als ein konsistenter expliziter Prädiktor.

1.4.3 Grundlagen allgemeiner Mehrschrittverfahren

Eine formale Gemeinsamkeit aller interpolatorischen Verfahren war, daß in der Rekursionsformel viele Auswertungen $f(x_j, \eta_j)$ aber nur ein η_j auftritt. Dagegen tritt bei den BDF-Verfahren in der Rekursionsformel (1.4.10) nur eine Funktionsauswertung, aber viele Stützwerte η_j auf.
Noch allgemeiner bezeichnet man als r**-Schritt-Verfahren** Verfahren, die durch eine Rekursion der Bauart

$$\eta_{j+r} + a_{r-1}\eta_{j+r-1} + \cdots + a_0\eta_j = h\,F(x_j; \eta_{j+r}, \eta_{j+r-1}, \ldots, \eta_j; h; f)\ . \quad (1.4.11)$$

beschrieben werden, und insbesondere **lineare** r**-Schritt-Verfahren**

$$\eta_{j+r} + a_{r-1}\eta_{j+r-1} + \cdots + a_0\eta_j = hF = h[b_r f(x_{j+r}, \eta_{j+r}) + \cdots + b_0 f(x_j, \eta_j)]\ , \quad (1.4.12)$$

die die interpolatorischen Verfahren und die BDF-Verfahren als Spezialfall enthalten und im weiteren behandelt werden.
Die Begriffe des lokalen und globalen Diskretisierungsfehlers lassen sich dann in einfacher Weise übertragen.

Definition 1.4.4 (Lokaler Diskretisierungsfehler)
Sei $z(t)$ *die exakte Lösung von* $z'(t) = f(t, z(t))\,,\ z(x) = y$ *. Dann heißt*

$$\tau(x, y; h) = \frac{1}{h}\left[z(x + rh) + \sum_{i=0}^{r-1} a_i z(x + ih)\right] - \sum_{i=0}^{r} b_i f(x + ih, z(x + ih))\ . \quad (1.4.13)$$

der **lokale Diskretisierungsfehler** *des Mehrschrittverfahrens (1.4.11).*

Interessant ist der Fall $x = x_j$ und $y = \eta_j$.
Man beachte, daß hier die Version der Definition gewählt wird, bei der die exakte Lösung in das Rekursionsschema eingesetzt wird, und nicht die Lösung des Rekursionsschema in die Differentialgleichung. Bei Mehrschrittverfahren ist auch nur dieser Weg möglich, da nach einigen Schritten Näherungen vorliegen, die inkonsistent sind[8]. Durch diese Näherungen existiert also gar keine exakte Lösung. Der lokale Diskretisierungsfehler der neuen Näherung als Abweichung von der exakten Lösung wäre dann gar nicht definiert.
Analog zu Einschrittverfahren definieren wir auch die Konsistenz:

[8] Bei Einschrittverfahren geht dagegen auch durch leicht verfälschte Näherungen stets eine benachbarte Lösung der Differentialgleichung.

Definition 1.4.5 (Konsistenz von Mehrschrittverfahren)
Ein Mehrschrittverfahren (1.4.11), (1.4.12) heißt **konsistent**, *falls für jedes* $f \in F_1(a,b)$ *eine Funktion* $\sigma(h)$ *existiert mit* $\lim_{h\to 0}\sigma(h) = 0$, *so daß*

$$|\tau(x,y;h)| \le \sigma(h) \quad \textit{für alle } x \in [a,b]\,,\ y \in \mathbb{R}^n \ . \tag{1.4.14}$$

Es heißt **konsistent von der Ordnung** p *falls für jedes* $f \in F_p(a,b)$,

$$\sigma(h) = \mathcal{O}(h^p) \ . \tag{1.4.15}$$

Für die Konsistenz eines linearen Mehrschrittverfahrens (1.4.12) sind die assoziierten Polynome von Bedeutung.

Definition 1.4.6 (Assoziierte Polynome)

$$\Psi(\mu) := \mu^r + a_{r-1}\mu^{r-1} + a_{r-2}\mu^{r-2} + \cdots + a_1\mu + a_0$$

heißt das zum linearen Mehrschrittverfahren (1.4.12) **erste assoziierte Polynom**

$$\chi(\mu) := b_r\mu^r + b_{r-1}\mu^{r-1} + b_{r-2}\mu^{r-2} + \cdots + b_1\mu + b_0$$

heißt **zweites assoziiertes Polynom**.

Durch Einsetzen des einfachen Vertreters $y' = f(x,y) = 0 \in F_1$ folgt notwendig für die Konsistenz

$$1 + a_{r-1} + \cdots + a_0 = 0 = \Psi(1) \ . \tag{1.4.16}$$

Diese Bedingung gilt auch für nichtlineare Mehrschrittverfahren.
Der Beweis einer höheren Konsistenzordnung ist für Mehrschrittverfahren (1.4.12) wesentlich einfacher, als für Einschrittverfahren, weil alle Auswertungen von f an Approximationen der gleichen Ordnung erfolgen, bzw, man in (1.4.12) alle Argumente von f durch die exakte Lösung ersetzt um τ zu berechnen. Ist $f \in F_{N-1}$, so gilt $z \in C^N$ und mit $z(x) = y$ erhalten wir

den lokalen Fehler

$$\begin{aligned} h\tau_j(x,y;h) &= \left[z(x_j+rh)+\sum_{i=0}^{r-1} a_i(z(x_j+ih)\right] \\ &\quad -h\sum_{i=0}^{r} b_i z'(x_j+ih) = \\ &= \sum_{k=0}^{N} z^{(k)}(x_j)\frac{(rh)^k}{k!}+\sum_{i=0}^{r-1} a_i \sum_{k=0}^{N} z^{(k)}(x_j)\frac{(ih)^k}{k!} \\ &\quad -h\sum_{i=0}^{r} b_i \sum_{k=0}^{N-1} z^{(k+1)}(x_j)\frac{(ih)^k}{k!}+\mathcal{O}(h^{N+1}) \qquad (1.4.17) \\ &= z(x_j)\left(1+\sum_{i=0}^{r-1} a_i\right)+\sum_{k=1}^{N} z^{(k)}(x_j)\frac{h^k}{k!}\left[r^k+\sum_{i=0}^{r-1} i^k a_i-\sum_{i=0}^{r} k i^{k-1} b_i\right] \\ &\quad +\mathcal{O}(h^{N+1}) \\ &=: z(x_j)c_0+\sum_{k=1}^{N} z^{(k)}(x_j)h^k c_k+\mathcal{O}(h^{N+1}) \end{aligned}$$

mit

$$\begin{aligned} c_0 &= 1+a_0+\cdots+a_{r-1}=\Psi(1) \\ c_1 &= r\cdot 1+a_1+2a_2+3a_3+\cdots+(r-1)a_{r-1}-(b_0+b_1+\cdots+b_r)=\Psi'(1)-\chi(1) \\ &\vdots \\ c_m &= \frac{1}{m!}[r^m+\sum_{i=1}^{r-1} a_i i^m-\sum_{i=0}^{r} b_i m i^{m-1}] \qquad (1.4.18) \end{aligned}$$

Für glattes f sollten die Terme niedriger h-Ordnung verschwinden. Daraus erhält man:

Satz 1.4.7 (Konsistenz linearer Mehrschrittverfahren)
Ein lineares Mehrschrittverfahren (1.4.12) ist konsistent genau dann, wenn gilt:

$$\Psi(1)=0 \quad ; \quad \Psi'(1)-\chi(1)=0$$

Beweis: Zu zeigen ist nur noch, daß die Bedingung notwendig ist. (Die c_i könnten ungleich Null sein, die Terme sich aber dennoch wegheben.)

Dies ersieht man, wenn man die Klasse linearer Differentialgleichungen $y' = f(x,y) = \alpha$ einsetzt, bei der die $\mathcal{O}$-Terme in (1.4.17) verschwinden. In diesem Fall erhält man exakt den lokalen Diskretisierungsfehler

$$\tau(x,y;h) = \frac{1}{h} z(x)\Psi(1) + \alpha z(x)[\Psi'(1) - \chi(1)]$$

der für alle α für $h \to 0$ verschwinden muß. ∎

Satz 1.4.8 (Ordnung eines linearen Mehrschrittverfahrens)
Das lineare Mehrschrittverfahren (1.4.12) ist **konsistent von der Ordnung** p*, genau dann, wenn*

$$\varphi(\mu) := \Psi(\mu) - \chi(\mu)\ln\mu$$

die $p+1$*-fache Nullstelle* $\mu = 1$ *besitzt.*

Beweis: Man ersetze $\mu := e^h$ in Ψ und χ:

$$\begin{aligned}
\Psi(\mu) &= \Psi(e^h) = e^{rh} + \sum_{k=0}^{r-1} a_k e^{kh} = \\
&= (1 + rh + \frac{1}{2}(rh)^2 + \cdots) + \sum_{k=0}^{r-1} a_k (1 + kh + \frac{1}{2}(kh)^2 + \cdots) \\
\chi(\mu) &= \chi(e^h) = \sum_{k=0}^{r} b_k (1 + kh + \frac{1}{2}(kh)^2 + \cdots) \\
&\overset{(1.4.18)}{\Rightarrow} \\
\Psi(e^h) - h\chi(e^h) &= c_0 + c_1 h + c_2 h^2 + \cdots + c_p h^p + c_{p+1} h^{p+1} + \cdots
\end{aligned}$$

Ordnung $p \Leftrightarrow [c_0 = c_1 = \cdots = c_p = 0] \Leftrightarrow$ [$h = 0$ ist $p+1$-fache Nullstelle von $\bar{\varphi}(h) := \Psi(e^h) - h\chi(e^h)$]
$\Leftrightarrow$ [$\mu = 1$ ist $p+1$-fache Nullstelle von $\varphi(h)$].[9] ∎

Beispiel 1.4.9 Die Mittelpunktsregel ist konsistent und von der Ordnung 2.

$$\eta_{i+1} - \eta_{i-1} = 2hf(x_i, \eta_i)$$

[9] Die Abbildung $h \mapsto e^h$ ist C^∞ und streng monoton mit Ableitung ungleich Null. Sie erhält daher die Vielfachheit der Nullstellen.

Mittels Taylorentwicklung erhält man:

$$\begin{aligned}\tau(x,y;h) &= \frac{1}{h}\left[z(x+2h)-z(x)-2hf(x+h,z(x+h))\right]\\ &= \frac{1}{h}\left[z(x+2h)-z(x)-2hz'(x+h)\right]\\ &= \frac{h^2}{3}z'''(x)+\mathcal{O}(h^3)\end{aligned}$$

oder mit Hilfe von φ:

$$\begin{aligned}\Psi(\mu) &= \mu^2-1 & \chi(\mu) &= 2\mu\\ \varphi(\mu) &= \mu^2-1-2\mu\ln\mu & \varphi(1) &= 0\\ \varphi'(\mu) &= 2\mu-2\ln\mu-2 & \varphi'(1) &= 0\\ \varphi''(\mu) &= 2-\frac{2}{\mu} & \varphi''(1) &= 0\\ \varphi^{(3)}(\mu) &= \frac{2}{\mu^2} & \varphi^{(3)}(1) &= 2\neq 0\end{aligned}$$

oder mit Hilfe von (1.4.18):

$$\begin{aligned}c_0 &= 1+a_0+a_1 = 1-1+0 = 0\\ c_1 &= r+a_1-b_0 = 2+0-2 = 0\\ 2c_2 &= r^2+a_1-b_1\cdot 2-b_2\cdot 2\cdot 2 = 4+0-2\cdot 2-0 = 0\end{aligned}$$

Wie bereits erwähnt, müssen zum Start eines r-Schrittverfahrens Näherungen an r Stellen bereits gegeben sein. In der Regel werden sie durch k-Schrittverfahren mit $k<r$ erzeugt. Bereits die Startwerte sind daher verfälscht mit Fehlern

$$\epsilon_i := \eta_i - y(x_i) \quad \text{für } i=0,1,\ldots,r-1\,.$$

Zusätzlich wird bei der Ausführung der Rekursionsformel ein neuer Fehler ϵ_{r+j}, $j=0,1,\ldots$ begangen.

$$\eta_{j+r}+a_{r-1}\eta_{j+r-1}+\cdots+a_0\eta_j =: h\,F(x_j;\eta_{j+r},\eta_{j+r-1},\ldots,\eta_j;h;f)+h\epsilon_{j+r}\,.$$

Die Lösung η_i des Mehrschrittverfahrens hängt daher von h und den ϵ_j ab und ist daher eine Funktion $\eta(x;\epsilon;h)$ wobei $\epsilon=\epsilon(x;h)$ eine Fehlerfunktion

mit $\epsilon(x_i;h) := \epsilon_i$ ist. $\eta(x;\epsilon;h)$ und $\epsilon(x;h)$ sind dabei beide wieder nur an bestimmten Stellen $x \in R_h$ bzw. für bestimmte Schrittweiten $h \in H_x$ definiert.
Wegen der Abhängigkeit des **globalen Diskretisierungsfehlers**

$$e(x;\epsilon;h) := \eta(x;\epsilon;h) - y(x) \tag{1.4.19}$$

von den Startwerten kann man nur Konvergenz erwarten, wenn mit der Schrittweite h auch die Fehler ϵ_i klein werden. Nur in diesem Fall können wir Konvergenz erhoffen.

Definition 1.4.10 (Konvergenz eines Mehrschrittverfahrens)
Ein Mehrschrittverfahren (1.4.11) heißt **konvergent** *falls*

$$\lim_{n\to\infty} \eta(x;\epsilon;h_n) = y(x) \ , \qquad h_n := (x - x_0)/n \ , \quad n = 1, 2, \dots \ , \tag{1.4.20}$$

für alle $x \in [a,b]$ *, alle* $f \in F_1(a,b)$ *, und alle Funktionen* $\epsilon(z;h)$ *mit*

$$|\epsilon(z;h)| \le \rho(h) \quad \textit{für alle } z \in R_h \textit{ mit} \quad \lim_{h\to 0} \rho(h) = 0 \ .$$

Ist also f glatt genug und verkleinern wir sowohl Schrittweite als auch Startfehler und Rundungsfehler bei der Erfüllung der Rekursion, so erwarten wir stets Konvergenz. Ansonsten hat das Verfahren einen gravierenden Mangel und wird nicht als konvergent bezeichnet.
Bei Einschrittverfahren war die Ordnung von globalem Fehler und lokalem Diskretisierungsfehler gleich. Daher könnte man auch bei Mehrschrittverfahren hoffen, daß Verfahren hoher Ordnung auch schnelle globale Konvergenz bedeuten. Dies ist jedoch ohne weitere Zusatzbedingungen nicht immer richtig.
Bei Einschrittverfahren folgte Konvergenz aus Konsistenz, weil bei lipschitzstetigem f lokale Fehler auch für $h \to 0$ nur begrenzt verstärkt werden. Bei MSV ist dafür die Stabilität der Rekursionsvorschrift nötig.

Beispiel 1.4.11 (Verfahren maximaler Ordnung)
Man betrachte das explizite lineare 2-Schritt-Verfahren

$$\eta_{j+2} + a_1\eta_{j+1} + a_0\eta_j = h\,[b_1 f_{j+1} + b_0 f_j] \ .$$

Dabei seien die Parameter a_0, a_1, b_0, b_1 so gewählt, daß ein Verfahren mit möglichst hoher Ordnung entsteht.

$$\begin{aligned}
\Psi(\mu) &= \mu^2 + a_1\mu + a_0 & \chi(\mu) &= b_1\mu + b_0 \\
\varphi(\mu) &= \mu^2 + a_1\mu + a_0 - (b_1\mu + b_0)\ln\mu & \varphi(1) &= 1 + a_1 + a_2 \\
\varphi'(\mu) &= 2\mu + a_1 - b_1\ln\mu - b_1 - b_0\frac{1}{\mu} & \varphi'(1) &= 2 + a_1 - b_1 - b_0 \\
\varphi''(\mu) &= 2 - b_1\frac{1}{\mu} + b_0\frac{1}{\mu^2} & \varphi''(1) &= 2 - b_1 + b_0 \\
\varphi^{(3)}(\mu) &= b_1\frac{1}{\mu^2} - b_0\frac{2}{\mu^3} & \varphi^{(3)}(1) &= b_1 - 2b_0
\end{aligned}$$

Ordnung 3 erfordert 4 Gleichungen für die 4 Koeffizienten

$$\begin{array}{rcl@{\qquad}rcl}
1 + a_1 + a_0 &=& 0 & a_1 &=& 4 \\
2 + a_1 - b_1 - b_0 &=& 0 & a_0 &=& -5 \\
2 - b_1 + b_0 &=& 0 & b_1 &=& 4 \\
b_1 - 2b_0 &=& 0 & b_0 &=& 2
\end{array}$$

Das Verfahren lautet dann

$$\eta_{j+2} + 4\eta_{j+1} - 5\eta_j = h[4f_{j+1} + 2f_j] \tag{1.4.21}$$

und ist ein Verfahren der Ordnung 3. Wir werden ab sehen, daß das Verfahren jedoch wegen Stabilitätsproblemen unbrauchbar ist.
Wir wenden das Verfahren auf das Anfangswertproblem

$$y' = -y\ , \quad y(0) = 1$$

an. Die exakte Lösung lautet $y(x) = e^{-x}$. Startet man mit den exakten Näherungen $\eta_0 = 1$ und $\eta_1 = e^{-h}$, und verwendet man eine kleine Schrittweite $h = 10^{-2}$, so erhält man bei einfach genauer Rechnung eine Folge η_i mit zunächst fallenden, dann aber betragsmäsig wachsenden Werten wechselnden Vorzeichens. Z.B. $\eta_{99} = .13 \cdot 10^{60}$ und $\eta_{100} = -.65 \cdot 10^{60}$.
Dies erklärt sich folgendermaßen: Die Folge η_j muß der Differenzengleichung

$$\eta_{j+2} + 4(1+h)\,\eta_{j+1} + (-5+2h)\,\eta_j = 0 \quad \text{für } j = 0, 1, \ldots$$

mit Startwerten $\eta_0 = 1$ und $\eta_1 = e^{-h}$ genügen. Jede lineare Differenzengleichung (ohne Anfangswerte) hat aber Lösungen der Form

$$\eta_j = \lambda^j\ .$$

Wir verifizieren den Ansatz durch einsetzen und erhalten als Bedingung für λ:

$$\begin{aligned} & \lambda^2 + 4(1+h)\lambda + (2h-5) = 0 \\ \Rightarrow \qquad & \lambda_{1,2} = -2(1+h) \pm 3\sqrt{1 + \frac{2}{3}h + \frac{4}{9}h^2}\ . \end{aligned}$$

Taylorentwicklung liefert

$$\begin{aligned} \lambda_1 &= 1 - h + \frac{1}{2}h^2 - \frac{1}{6}h^3 + \frac{1}{72}h^4 + \mathcal{O}(h^5)\ , \\ \lambda_2 &= -5 - 3h + \mathcal{O}(h^2)\ . \end{aligned}$$

Die allgemeine Lösung der Differenzengleichung lautet also

$$\eta_j = \alpha\,\lambda_1^j + \beta\,\lambda_2^j$$

wobei die Konstanten α und β durch die Startwerte festgelegt sind. In diesem Fall also:

$$\alpha = \frac{\lambda_2 - e^{-h}}{\lambda_2 - \lambda_1}\ , \quad \beta = \frac{e^{-h} - \lambda_1}{\lambda_2 - \lambda_1}\ .$$

Daraus ergibt sich, wieder durch Taylorentwicklung, $\alpha = 1 + \mathcal{O}(h^2)$ und $\beta = -\frac{1}{216}h^4 + \mathcal{O}(h^5)$. Dies ergibt für die Näherung

$$\begin{aligned} \eta_n \;=\; & \left[1 + \mathcal{O}\left((xn)^2\right)\right] \cdot \left[1 - \frac{x}{n} + \mathcal{O}\left(\left(\frac{x}{n}\right)^2\right)\right]^n \\ & -\frac{1}{216}\left(\frac{x}{n}\right)^4 \left[1 + \mathcal{O}\left(\frac{x}{n}\right)\right] \cdot \left[-5 - 3\,\frac{x}{n} + \mathcal{O}\left(\left(\frac{x}{n}\right)^2\right)\right]^n\ . \end{aligned} \tag{1.4.22}$$

Beachten wir $\lim_{n\to\infty}\left(1 \pm \frac{x}{n}\right)^n = \exp(\pm x)$, so sehen wir, daß der erste Term in (1.4.22) die Lösung e^{-x} approximiert, während der zweite Term für $n \to \infty$ ein oszillierendes asymptotisches Verhalten zeigt.

$$-\frac{x^4}{216}\,\frac{(-5)^n}{n^4}\,\exp\left(\frac{3}{5}\,x\right) \to \pm\infty\ . \tag{1.4.23}$$

Selbst für sehr kleines β wird dieser Term dominant.

Wir untersuchen den Fall für allgemeine lineare r-Schrittverfahren.
Jedes r-Schrittverfahren kann als Abbildung

$$(\eta_{j+r-1}, \ldots, \eta_j) \to (\eta_{j+r}, \ldots, \eta_{j+1})$$

interpretiert werden. Ist das Verfahren und die Funktion f linear, so wird die Abbildung durch eine Matrix repräsentiert. Das Verfahren

$$\begin{aligned} &\eta_{j+r} + a_{r-1}\eta_{j+r-1} + \cdots + a_1\eta_{j+1} + a_0\eta_j = \\ &\quad = h[b_r f(x_{j+r}, \eta_{j+r}) + b_{r-1} f(x_{j+r-1}, \eta_{j+r-1}) + \cdots + b_1 f(x_{j+1}, \eta_{j+1}) + b_0 f(x_j, \eta_j)] \end{aligned} \tag{1.4.24}$$

führt etwa in dem Fall $f = 0$ auf die Differenzengleichung

$$\eta_{j+r} + a_{r-1}\eta_{j+r-1} + \cdots + a_1\eta_{j+1} + a_0\eta_j = 0 \ .$$

Ein Vektor von Näherungen $(\eta_j, \ldots, \eta_{j+r-1})$ wird dadurch abgebildet auf einen Vektor

$$\bar{\eta}_j := (\eta_j, \ldots, \eta_{j+r-1})^T \mapsto \bar{\eta}_{j+1} = (\eta_{j+1}, \ldots, \eta_{j+r})^T =: \Phi(\bar{\eta}_j) \ .$$

Die zugehörige lineare Abbildung lautet

$$\bar{\eta}_j := \begin{bmatrix} \eta_j \\ \eta_{j+1} \\ \vdots \\ \eta_{j+r-1} \end{bmatrix} \to \begin{bmatrix} 0 & 1 & & \\ 0 & \ddots & \ddots & \\ & & 0 & 1 \\ -a_0 & \cdots & -a_{r-2} & -a_{r-1} \end{bmatrix} \begin{bmatrix} \eta_j \\ \eta_{j+1} \\ \vdots \\ \eta_{j+r-1} \end{bmatrix}$$

$$= \begin{bmatrix} \eta_{j+1} \\ \eta_{j+2} \\ \vdots \\ \eta_{j+r} \end{bmatrix} = \bar{\eta}_{j+1}$$

Die Abbildungsmatrix A hat als charakteristisches Polynom

$$\begin{aligned} |A - \mu I| &= \begin{vmatrix} -\mu & 1 & & \\ & \ddots & \ddots & \\ & & -\mu & 1 \\ -a_0 & \cdots & -a_{r-2} & -(a_{r-1} + \mu) \end{vmatrix} \\ &= \begin{vmatrix} -\mu & 1 & & & \\ 0 & \ddots & & \ddots & \\ & & 0 & & 1 \\ -a_0 & \cdots & -(a_{r-2} + \mu(a_{r-1} + \mu) & & -(a_{r-1} + \mu) \end{vmatrix} \end{aligned}$$

Dabei wurde die letzte Spalte μ-fach zur vorletzten addiert. Eliminiert man auf diese Weise sukzessive die Einträge unterhalb der Diagonalen, indem man das μ-fache der j-ten Spalte zur $j-1$-ten Spalte addiert, so erhält man eine Matrix in der links unten bis auf das Vorzeichen gerade das erste assozierte Polynom

$$\Psi(\mu) := \mu^r + a_{r-1}\mu^{r-1} + a_{r-2}\mu^{r-2} + \cdots + a_1\mu + a_0$$

in Hornerform entsteht. Entwicklung nach der letzten Spalte liefert dann

$$|A - \mu I| = (-1)^r \Psi(\mu)|I_{r-1}| = (-1)^r \Psi(\mu)$$

Hat das Polynom $\Psi(\mu)$ Nullstellen μ_i mit $|\mu_i| > 1$, so wächst der entsprechende Eigenvektoranteil exponentiell wie μ_i^j an, obwohl die Differentialgleichung nur konstante Lösungen besitzt ($f = 0$). Dies darf offensichtlich nicht passieren. Selbst für $|\mu_i| = 1$ aber k-fach existieren polynomial wie $j^{k-1}\mu_i^j$ wachsende Anteile, denn $A-\mu I$ hat stets Rang $r-1$. Die geometrische Vielfachheit der Eigenwerte von A ist also stets 1 und es existiert bei k-fachem Eigenwert eine Hauptvektorkette der Länge k.

Definition 1.4.12 (Stabilität)
Das lineares Mehrschrittverfahren (1.4.24) heißt **stabil**, *falls für alle Nullstellen μ des* **ersten assozierten Polynoms**

$$\Psi(\mu) := \mu^r + a_{r-1}\mu^{r-1} + a_{r-2}\mu^{r-2} + \cdots + a_1\mu + a_0$$

gilt:

$$|\mu \leq 1| \quad ; \quad |\mu| = 1 \Rightarrow \mu \text{ einfache Nullstelle .}$$

Da diese Bedingung aus der Analyse der Differentialgleichung $y' = 0$ gewonnen wurde, nennt man die Verfahren auch **nullstabil**, *bzw. zu Ehren von Dahlquist* **D-stabil**.

Aufgabe 1.4.13 (Stabilität eines Einschrittverfahrens)
Man zeige: Einschrittverfahren sind stabil

Aufgabe 1.4.14 (Stabilität interpolatorischer Mehrschrittverfahren)

Man zeige: Interpolatorische Mehrschrittverfahren vom Typ (1.4.5) sind stabil!

Aufgabe 1.4.15 (Stabilität eines Mehrschrittverfahrens)
Das Mehrschrittverfahren

$$\eta_{j+2} + 4\eta_{j+1} - 5\eta_j = h[4f_{j+1} + 2f_j]$$

werde auf das Anfangswertproblem

$$y' = 0 \ , \quad y(0) = 1$$

angewendet. Dabei sei $\eta_0 = y(0) = 1$ exakt und $\eta_1 = 1 + \varepsilon$ leicht verfälscht mit Fehler ε. Man gebe η_j in analytischer Form an!

Globaler Diskretisierungsfehler:
Um den globalen Diskretisierungsfehler abzuschätzen, formulieren wir Mehrschrittverfahren als verallgemeinerte Einschrittverfahren $y_{i+1} = \Psi(x_i, y_i, h, f)$ (vergleiche die Bemerkung nach Satz 1.2.20). Ohne Beschränkung der Allgemeinheit können wir den Fall einer Differentialgleichung ($n = 1$) untersuchen. Ein lineares Mehrschrittverfahren (1.4.24) läßt sich auch formulieren als

$$\eta_{j+r} = -\sum_{i=0}^{r-1} a_i \eta_{j+i} + h\Theta$$

mit (falls $b_r \neq 0$) implizit definiertem Θ

$$\begin{aligned}\Theta &= [b_r f(x_{j+r}, \eta_{j+r}) + b_{r-1} f(x_{j+r-1}, \eta_{j+r-1}) + \cdots + b_1 f(x_{j+1}, \eta_{j+1}) + b_0 f(x_j, \eta_j)] \\ &= [b_r f(x_{j+r}, -\sum_{i=0}^{r-1} a_i \eta_{j+i} + h\Theta) + b_{r-1} f(x_{j+r-1}, \eta_{j+r-1}) + \cdots \\ &\quad \cdots + b_1 f(x_{j+1}, \eta_{j+1}) + b_0 f(x_j, \eta_j)] \ .\end{aligned}$$

Für h klein genug und $f \in F_1$ ist nach dem Satz über implizite Funktionen ist Θ eindeutig bestimmt und hängt differenzierbar ab von

$$\bar{\eta}_j := (\eta_j, \eta_{j+1}, \ldots, \eta_{j+r-1})^T$$

ist also insbesondere Lipschitz-stetig

$$\|\Theta(U) - \Theta(V)\| \leq M \|U - V\| \ .$$

Die Iterationsfunktion $\bar{\eta}_{j+1} = F(\bar{\eta}_j)$ ist dann implizit gegeben durch das nichtlineare Gleichungssystem

$$I\bar{\eta}_{j+1} = \underbrace{\begin{bmatrix} 0 & 1 & & \\ 0 & \ddots & \ddots & \\ & & 0 & 1 \\ -a_0 & \cdots & -a_{r-2} & -a_{r-1} \end{bmatrix}}_{=:A} \bar{\eta}_j + h \underbrace{\begin{bmatrix} 0 \\ \vdots \\ 0 \\ \sum_{i=0}^{r} b_i f(\eta_{j+i}) \end{bmatrix}}_{=:g(\bar{\eta}_{j+1})} . \quad (1.4.25)$$

Für h klein genug ist der nichtlineare Anteil hg vernachlässigbar und nach dem Satz über implizite Funktionen Ψ seinerseits wohldefiniert, differenzierbar und daher lipschitzstetig.
Analog bildet man aus der exakten Lösung $y(x)$ der Differentialgleichung den erweiterten Vektor

$$\bar{y}_j := (y(x_j), \ldots, y(x_{j+r-1}))^T$$

Wegen Bemerkung 1.2.21 ist dann nur zu zeigen, ob sich $\|\bar{y}_j - \bar{\eta}_j\|$ in einer geeigneten Norm abschätzen lassen.

Satz 1.4.16 (Konvergenz von linearen Mehrschrittverfahren)
Ist ein stabiles lineares Mehrschrittverfahren konsistent, so konvergiert die numerische Lösung für $f \in F_1[a,b]$ gegen die exakte Lösung. Hat das Verfahren die Konsistensordnung p so konvergiert die numerische Lösung für $f \in F_p[a,b]$ mit Ordnung $\mathcal{O}(h^p)$.

Beweis: Wir interpretieren das Mehrschrittverfahren als Einschrittiteration auf $\bar{\eta}$, allerdings nicht als Einschrittverfahren, denn $\bar{\eta}_{i+1} = F(\bar{\eta}_i) \neq \bar{\eta}_i + h\Phi(h, \bar{\eta}_i, f)$, (vgl. Lemma 1.2.19), und müssen nur zeigen:

$$\|F(U) - F(V)\| \leq (1 + hM)\|U - V\|$$

Zunächst erhalten wir aus der Stabilitätsbedingung: Spektralradius $\rho(A) \leq 1$ sowie Eigenwerte auf dem Einheitskreis sind einfach. Die Jordansche Normal-

form von A mit betragsmäßig geordneten Eigenwerten hat dann die Bauart

$$T^{-1}AT = J = \begin{bmatrix} \lambda_1 & & & & & & & \\ & \ddots & & & & & & \\ & & \lambda_m & & & & & \\ & & & \lambda_{m+1} & \delta_{m+1} & & & \\ & & & & \ddots & \ddots & & \\ & & & & & \lambda_{r-1} & \delta_{r-1} & \\ & & & & & & \lambda_r & \end{bmatrix}$$

mit $|\lambda_i| < 1$ für $i = m+1, \ldots, r$ und $\delta_i \in \{0; 1\}$. Durch weitere Transformation läßt sich erreichen, daß an Stelle der Einsen in der Nebendiagonale nur Einträge kleiner $\varepsilon := 1 - |\lambda_{m+1}|$ stehen. Dazu bilde man $D^{-1}T^{-1}ATD$ mit $D := \operatorname{diag}(\varepsilon^1, \ldots, \varepsilon^r)$ Für die spezielle Norm $\|x\| := \|D^{-1}T^{-1}x\|_\infty$ gilt dann

$$\|A\|_{lub} := \max \frac{\|Ax\|}{\|x\|} \leq 1$$

Damit erhält man für h klein genug die notwendige Abschätzung

$$\|F(U) - F(V)\| \leq \|A(U-V)\| + \|h\Theta(U) - h\Theta(V)\| \leq (1 + hM)\|U - V\| \ .$$

Nach Lemma 1.2.19 folgt damit, daß sich ein lokaler Fehler e_i' an der Stelle x_i nur mit $exp(\frac{x-x_i}{h}hM)e_i'$ auf das Endergebnis auswirkt, so daß wir für den globalen Fehler die Abschätzung haben

$$\begin{aligned} \|\eta(x) - y(x)\| &\leq \sum_{i=1}^{m} e_i' e^{\frac{x-x_i}{h}hM} \leq \frac{x - x_0}{h} \max_i h\|\tau_i\| e^{\frac{x-x_0}{h}hM} \\ &= (x - x_0) \max_i \|\tau_i\| e^{(x-x_0)M} \ . \end{aligned}$$

Für $\max_i \tau_i \to 0$ strebt der globale Fehler also mit gleicher Ordnung gegen 0. ∎

Das Ergebnis läßt sich kurz zusammenfassen zu:

$$\text{Konsistenz} + \text{Stabilität} = \text{Konvergenz}$$

1.5 Extrapolationsverfahren

Statt durch immer größere Runge-Kutta Tableaus und immer mehr Bedingungsgleichungen, kann man auch versuchen durch Extrapolation Verfahren hoher Ordnung zu konstruieren.
Für die Schrittweitensteuerung sehr wichtig war die Tatsache, daß nicht an der Stelle $x+h$ eine Entwicklung in h-Potenzen existiert, sondern daß die numerische Approximation an einer festen Stelle $x > x_0$ eine Entwicklung in h-Potenzen besitzt. Dies ist bei genauerem Hinsehen erst einmal erstaunlich. Wir betrachten jetzt nicht den Fehler eines Schrittes in Abhängigkeit von h, sondern den Fehler an einer festen Stelle x in Abhängigkeit von h. Bei einer Änderung der Schrittweite ändert sich jedoch die Anzahl der benötigten Schritte bis zu einem vorgegebenen x. Außerdem sind nur diskrete Schrittweiten möglich, so daß von Differenzierbarkeit nach h keine Rede sein kann, und also auch keine Taylorentwicklung definiert ist.

Satz 1.5.1 (Asymptotischen Entwicklung des globalen Fehlers)
(Gragg 1964) Sei $f \in F_{N+2}[a,b]$, $\Phi \in C^{N+2}$ genügend oft differenzierbar[10] *ein Einschrittverfahren der Ordnung p. Dann existieren differenzierbare Funktionen $e_k(x)$ unabhängig von h mit $e_k(x_0) = 0$, so daß für die numerische Lösung $\eta(x,h)$ gilt:*

$$\eta(x,h) = y(x) + \sum_{k=p}^{N} e_k(x)h^k + E_{N+1}(x,h)h^{N+1} \tag{1.5.1}$$

für alle $h = \frac{x-x_0}{n}$, mit $E_{N+1}(x,h)$ bei festem x für alle h beschränkt.

Die Konsistenzordnung überträgt sich also auf die Ordnung des globalen Fehlers und durch Extrapolation lassen sich Verfahren höherer Ordnung gewinnen.
Beweis: Wir verwenden die Entwicklung des lokalen Fehlers (1.2.21), insbesondere

$$h\tau(x,y(x),h) = \sum_{i=p+1}^{N+1} g_i(x)h^i + \mathcal{O}(h^{N+2}) \ .$$

[10] Bei praktisch allen wichtigen Verfahren ist $\Phi \in C^\infty$, beziehungsweise genauso glatt wie f, so daß diese Bedingung oft gar nicht angegeben wird.

Bei konsistenten Verfahren mit differenzierbarer Verfahrensvorschrift und hinreichend glatter exakter Lösung gilt dann

$$\begin{aligned} \Phi(x,y,h) &= \frac{1}{h}\int_x^{x+h} f(t,y(t))\mathrm{d}t - \tau \\ \Rightarrow \Phi_y(x,y,h) &= f_y(x,y) + \mathcal{O}(h)\ . \end{aligned}$$

Wir zeigen nun die Existenz einer differenzierbaren Funktion $e_p(x)$ mit

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

Existiert $e_p(x)$, so wäre die korrigierte Näherung

$$\begin{aligned} \bar\eta(x+h,h) &:= \eta(x+h,h) - e_p(x+h)h^p \\ &= \eta(x,h) + h\Phi(x,\eta(x,h),h) - e_p(x+h)h^p \\ &= \bar\eta(x,h) + \underbrace{e_p(x)h^p + h\Phi(x,\bar\eta(x,h)+e_p(x)h^p,h) - e_p(x+h)h^p}_{h\bar\Phi} \\ &= \bar\eta(x,h) + h\bar\Phi(x,\bar\eta(x,h),h) \end{aligned}$$

konsistent von der Ordnung $p+1$, mit

$$\begin{aligned} \bar\Phi(x,y,h) &= \Phi(x,y+e_p(x)h^p,h) - [e_p(x+h) - e_p(x)]h^{p-1} \\ &= \Phi(x,y,h) + \frac{\partial\Phi}{\partial y}(x,y,h)e_p(x)h^p + \mathcal{O}(h^{p+1}) - [e_p'(x)h + \mathcal{O}(h^2)]h^{p-1} \\ &\overset{(1.2.14)}{=} \Phi(x,y,h) + [f_y(x,y) + \mathcal{O}(h)]e_p(x)h^p + \mathcal{O}(h^{p+1}) - e_p'(x)h^p \\ &= \Phi(x,y,h) + [f_y(x,y)e_p(x) - e_p'(x)]h^p + \mathcal{O}(h^{p+1}) \end{aligned}$$

Das heißt:

$$\begin{aligned} &y(x+h) - y(x) - h\bar\Phi(x,y,h) = \\ &\quad = y(x+h) - y(x) - h\Phi(x,y,h) + [f_y(x,y)e_p(x) - e_p'(x)]h^{p+1} + \mathcal{O}(h^{p+2}) \\ &\quad = g_{p+1}(x)h^{p+1} + [f_y(x,y)e_p(x) - e_p'(x)]h^{p+1} + \mathcal{O}(h^{p+2}) \\ &\quad = [g_{p+1}(x) + f_y(x,y)e_p(x) - e_p'(x)]h^{p+1} + \mathcal{O}(h^{p+2}) \end{aligned}$$

Das zu $\bar\Phi$ gehörige Einschrittverfahren besitzt also die Ordnung $p+1$, wenn man $e_p(x)$ als Lösung des folgenden AWPs wählt:

$$e_p'(x) = f_y(x,y(x))e_p(x) + g_{p+1}(x) \quad ; \quad e_p(x_0) = 0$$

Dann gilt nach Satz 1.2.20

$$\bar{\eta}(x,h) - y(x) = \mathcal{O}(h^{p+1})$$

und somit

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

Man wiederholt nun die Argumentation für das Verfahren $\bar{\Phi}$ statt Φ und erhält

$$\bar{\eta}(x,h) - y(x) - e_{p+1}(x)h^{p+1} = \mathcal{O}(h^{p+2}) \ .$$

falls e_{p+1} das AWP

$$e'_{p+1}(x) = f_y(x,y(x))e_{p+1}(x) + g_{p+2}(x) \quad ; \quad e_{p+1}(x_0) = 0$$

löst. Nach $N-p+1$ Schritten erhält man den Satz. ∎

Die besondere Bedeutung von Satz 1.5.1 besteht darin, daß sich damit in einfacher Weise Verfahren beliebig hoher Ordnung konstruieren lassen. Ausgangspunkt ist die assymptotische Entwicklung (1.5.1) des globalen Diskretisierungsfehlers

$$\eta(x,h) = y(x) + \sum_{k=p}^{N} e_k(x)h^k + E_{N+1}(x,h)h^{N+1} \ .$$

Hat man mit zwei verschiedenen Schrittweiten $h_1 = \frac{h}{n_1}$ und $h_2 = \frac{h}{n_2}$ Näherungen an der Stelle $x = x_0 + h$ berechnet, so gilt:

$$\begin{aligned}
\eta(x,h_1) &= y(x) + \sum_{k=p}^{N} e_k(x)h_1^k + E_{N+1}(x,h_1)h_1^{N+1} \\
\eta(x,h_2) &= y(x) + \sum_{k=p}^{N} e_k(x)h_2^k + E_{N+1}(x,h_2)h_2^{N+1} \\
\frac{h_1^p\eta(x,h_2) - h_2^p\eta(x,h_1)}{h_1^p - h_2^p} &= y(x) + \frac{h_1^p}{h_1^p - h_2^p}e_{p+1}(x)h_2^{p+1} - \frac{h_2^p}{h_1^p - h_2^p}e_{p+1}(x)h_1^{p+1} \\
&\quad + \mathcal{O}(h_1^{p+2} + h_2^{p+2})
\end{aligned}$$

Man gewinnt also ein neues Verfahren der Ordnung $p+1$ durch

$$\bar{\eta}(x,h) := \frac{h_1^p\eta(x,h_2) - h_2^p\eta(x,h_1)}{h_1^p - h_2^p} \ .$$

Ist p klein, so steht dem höheren Aufwand eine relativ große Ordnungserhöhung gegenüber. Besonders geeignet sind daher Verfahren niedriger Ordnung.

1.5.1 Das extrapolierte Eulerverfahren

Das extrapolierte Eulerverfahren ist ein Einschrittverfahren, bei dem in jedem Makroschritt mit Schrittweite H ausgehend von $(x, \eta(x))$ Näherungen $\eta(x+H; h_i)$ an der Stelle $x+H$ durch n_i Euler-Schritte mit verschiedenen Schrittweiten $h_i = \frac{H}{n_i}$ [11]berechnet werden.
Der zugehörige Algorithmus lautet dann:
Extrapoliertes Eulerverfahren, Teil 1.

`Gegeben:`	$x_0, \quad y_0$
	H: Makroschrittweite
	k: Anzahl der Näherungen, bzw. Ordnung
	n_i: Schritteweitenfolge $= h_i = H/n_i$
`Vorbereitung:`	$f_0 := f(x_0, y_0)$: Auswertung am Startpunkt
`Berechne:`	$\hat{x} := x_0 + H$
	$h_i := H/n_i$: Mikroschrittweite
	$\eta_0 := y_0$
	$\eta_1 := y_0 + h_i\, f_0$
	Für $j = 1, 2, \ldots, n-1$
	$\quad \eta_{j+1} := \eta_j + h_i\, f(x_j, \eta_j), \quad x_{j+1} := x_j + h_i$
	$\eta_{i,0} := \eta_{n_i}$

Dann gilt

$$\begin{aligned} \eta(x,h) &= y(x) + \sum_{k=1}^{N} e_k(x) h^k + E_{N+1}(x,h) h^{N+1} \\ &= p_k(h) + \sum_{j=k+1}^{N} e_j(x) h^j + E_{N+1}(x,h) h^{N+1} = p_k(h) + \mathcal{O}(h^{k+1}) \end{aligned}$$

mit $e_k(x_0) = 0$ und $e_k(x)$ differenzierbar. $\eta(x,h)$ kann also näherungsweise geschrieben werden als ein Polynom p_k vom Grad k.

$$p_k(t) = y(x) + \sum_{j=1}^{k} e_j(x) h^j$$

[11]Die Schrittweitenfolge n_i beeinflußt nicht nur über die Anzahl benötigter Funktionsaufrufe den Aufwand, sondern über die Extrapolationsgewichte auch den Einfluß von Rundungsfehlern.

Hat man $\eta_{i,0} := \eta(x, h_i), \ldots, \eta_{i+k,0} := \eta(x, h_{i+k})$ berechnet, so erhält man durch Interpolation ein Polynom

$$\begin{aligned}
\eta_{i,k}(t) &:= \sum_{j=i}^{i+k} \eta_{j,0} L_j(t) \\
&= \sum_{j=i}^{i+k} [p_k(h_j) + \sum_{l=k+1}^{N} e_l(x) h_j^l + E_{N+1}(x, h_j) h_j^{N+1}] L_j(t) \\
&= \sum_{j=i}^{i+k} p_k(h_j) L_j(t) + \underbrace{\sum_{j=i}^{i+k} \sum_{l=k+1}^{N} [e_l(x) h_j^l + E_{N+1}(x, h_j) h_j^{N+1}] L_j(t)}_{\mathcal{O}(h_i \cdots h_{i+k})} \\
&= \underbrace{\sum_{j=i}^{i+k} p_k(h_j) L_j(t)}_{p_k(t)} + \mathcal{O}(h_i \cdots h_{i+k})
\end{aligned}$$

Auswertung an der Stelle $t = 0$ ergibt

$$\eta_{i,k} := \eta_{i,k}(0) = y(x) + \mathcal{O}(h_i \cdots h_{i+k})$$

Die $\eta_{i,k}$ lassen sich nach dem Algorithmus von Aitken-Neville effizient berechnen.

$$\begin{array}{ccccccc}
\eta_{1,0} & & & & & & \\
 & \searrow & & & & & \\
\eta_{2,0} & \rightarrow & \eta_{2,1} & & & & \\
 & \searrow & & \searrow & & & \\
\eta_{3,0} & \rightarrow & \eta_{3,1} & \rightarrow & \eta_{3,2} & & \\
 & \searrow & & \searrow & & \ddots & \\
\vdots & & \vdots & & \vdots & &
\end{array}$$

In der ersten Spalte stehen die Näherungen $\eta_{i,0} := \eta(x, h_i)$ des Eulerverfahrens. in Spalte k jeweils eine extrapolierte Näherung der Ordnung k. Die tieferen Spalten wurden dabei mit kleinerern Schrittweiten erzeugt, d.h., der Fehler in Zeile i und Spalte k ist von der Ordnung $\mathcal{O}(h_i \cdots h_{i+k-1})$.

Extrapoliertes Eulerverfahren, Teil 2:.

`Berechne:` Für $i = 1, 2, 3, \ldots$
$\eta_{i,0} := \eta(x, h_i)$
For $1 \leq k \leq i-1$
$$\eta_{i,k} := \eta_{i,k-1} + \frac{\eta_{i,k-1} - \eta_{i-1,k-1}}{\left(\frac{h_{i-k}}{h_i}\right) - 1}$$

Im Fall $n_i = i\,,\ i = 1, 2, \ldots, k$ erhält man aus k Näherungen mit Schrittweiten $h_i = H/i$ des expliziten Euler-Verfahrens eine extrapolierte Näherung der Ordnung k.

Dabei werden $\left(\sum_{i=1}^{k} i\right) - (k-1) = (k-2)(k-1)/2$ Funktionsauswertungen benötigt. (Die Auswertung am linken Rand f_0 kann mehrfach verwendet werden. Mit 36 Funktionsauswertungen erreicht man also z.B. Ordnung $k = 10$.

Bemerkung: Teil 1 und 2 werden ineinander verwoben. Insbesondere wird nach einer neuen Näherung $\eta_{i,0}$ das Extrapolationstableau so weit wie möglich vervollständigt. Insbesondere sind dann $\eta_{k,k-1}$ und $\eta_{k-1,k-1}$ Näherungen der Ordnung $k-1$ mit unterschiedlichen Schrittweiten. Aus ihnen läßt sich die Näherung $\eta_{k,k}$ der Ordnung $\mathcal{O}(H^k)$ berechnen sowie eine Schrittweite schätzen.

Reicht die Genauigkeit von $\eta_{k-1,k-1}$ aus, so werden keine weiteren Näherungen zur Extrapolation berechnet. Ist die geforderte Genauigkeit jedoch auch mit dem vollständigen Extrapolationstableau noch nicht erreicht so wird gegebenfalls eine weitere Zeile des Extrapolationstableaus berechnet.

Das Verfahren von Aitken-Neville ist zwar etwas aufwendiger als die Newton'sche Interpolationsformel mit dividierten Differenzen, dafür lassen sich alle Zwischenergebnisse des Tableaus als Näherung verschiedener Ordnung und verschiedener Genauigkeit interpretieren. Das Verfahren erlaubt daher ständig Näherungen verschiedener Ordnung auf ihre Genauigkeit und Effizienz hin zu überprüfen und eignet sich daher hervorragend für einer Ordnungssteuerung. Dabei wird auch der mit der erhöhten Ordnung verbundene höhere Aufwand (Auswertungen von f) mit in Betracht gezogen.

Als weiteres Basisverfahren der Ordnung 1 für die Extrapolation ist vor allem

auch noch das **implizite Eulerverfahren**[12]

$$\eta_{i+1} = \eta_i + hf(x_{i+1}, \eta_{i+1}) \ ,$$

und das **semiimplizite Eulerverfahren**

$$\begin{aligned} & \eta_{i+1} - \eta_i \approx h[f(x_{i+1}, \eta_i) + f_y(x_i, \eta_i)(\eta_{i+1} - \eta_i)] \\ \Rightarrow \quad & [I - hf_y(x_i, \eta_i)](\eta_{i+1} - \eta_i) = hf(x_{i+1}, \eta_i) \end{aligned}$$

besonders interessant.

Bemerkung 1.5.2 Extrapolationsverfahren fester Ordnung lassen sich auch als Einschrittverfahren interpretieren.

Aufgabe 1.5.3 (Das extrapolierte Eulerverfahrenals Einschrittverfahren) Die Näherung $\eta(x + H)$ werde mit dem extrapolierten Eulerverfahren der Ordnung 3 mit Schrittweiten $h_i = H/n_i\,,\ n_i = 1, 2, 3$ berechnet. Dies kann als RK-Verfahren interpretiert werden. Man gebe das Buttcherarray an.

1.5.2 Die extrapolierte Mittelpunktsregel

Verfahren höherer Ordnung sind für die Extrapolation weniger geeignet, da der Aufwand schnell steigt, die Ordnung sich jedoch viel langsamer verdoppelt als beim Eulerverfahren. Eine Ausnahme stellen Verfahren der Ordnung 2 dar, die aufgrund ihrer Symmetrie eine besondere Form der Fehlerentwicklung besitzen.
Wir untersuchen dies exemplarisch an der expliziten Mittelpunktsregel

$$\eta_{i+1} = \eta_{i-1} + 2hf(x_i, \eta_i) \ .$$

[12] Für h klein genug gilt für die Funktion $g(\eta_{i+1}) := \eta_{i+1} - \eta_i - hf(x_{i+1}, \eta_{i+1}) :\ g_{\eta_{i+1}} \approx I$. Nach Satz über implizite Funktionen ist das nichtlineare Gleichungssystem dann nach η_{i+1} auflösbar, und für $f \in C^k$ ist η_{i+1} sogar k-fach differenzierbar abhängig von x_i, h und η_i. Damit ist dann auch $\Phi := f(x_{i+1}, \eta_{i+1})$ k-fach differenzierbar, wenn auch nicht explizit bekannt.

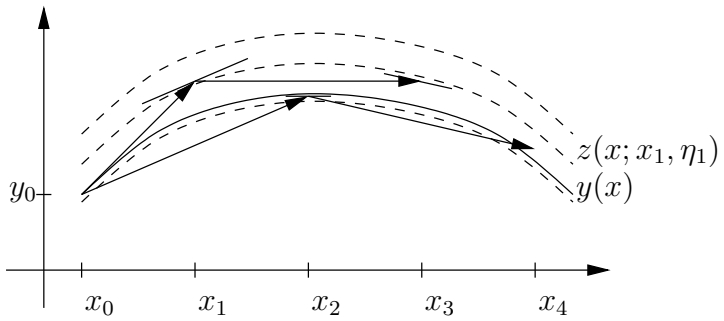


Abbildung 1.8: Mittelpunktsregel

Dies ist kein Einschrittverfahren, da η_{i-1} und η_i benötigt werden um η_{i+1} zu berechnen. Zum Start wird außer $\eta_0 = y_0$ noch η_1 benötigt. η_1 wird z.B. durch einen Eulerschritt berechnet

$$\begin{aligned} h &= \frac{x - x_0}{n} =: h_n \\ \eta_0 &= y_0 \\ \eta_1 &= \eta_0 + hf(x_0, \eta_0) \\ i &= 2, \ldots, n \\ & \quad x_{i-1} = x_0 + (i-1)h \\ & \quad \eta_i = \eta_{i-2} + 2hf(x_{i-1}, \eta_{i-1}) \end{aligned}$$

Dann gilt:

Satz 1.5.4 (Globaler Fehler der Mittelpunktsregel)
(Gragg 1965) Sei $f \in F_{2N+2}[a,b]$, $y(x)$ die exakte Lösung des Anfangswertproblems

$$y' = f(x,y) \quad ; \quad y(x_0) = y_0$$

Für $x \in R_h = \{x_0 + ih \,|\, i = 0, 1, \ldots, \}$ sei

$$\begin{aligned} \eta(x_0; h) &:= y_0 \\ \eta(x_0 + h; h) &:= y_0 + hf(x_0, y_0) \\ \eta(x + h; h) &:= \eta(x - h; h) + 2hf(x, \eta(x; h)) \end{aligned} \tag{1.5.2}$$

dann existieren differenzierbare Funktionen u_k, v_k unabhängig von h und eine beschränkte Funktion $E_{2n+2}(x; h)$, so daß für alle $x \in [a,b]$ und al-

le $h = (x - x_0)/n\,,\ n = 1, 2, \ldots$ *für den Fehler von* $\eta(x; h)$ *gilt:*

$$\eta(x; h) = y(x) + \sum_{k=1}^{N} h^{2k} \left[u_k(x) + (-1)^{(x-x_0)/h}\, v_k(x)\right] + h^{2N+2}\, E_{2n+2}(x; h) \tag{1.5.3}$$

Die Entwicklung (1.5.3) ist nicht von der Bauart (1.5.1). (Die Mittelpunktsregel ist auch kein Einschrittverfahren.)
Unter den Annahmen von Satz 1.5.4 gilt für den Fehler in erster Näherung

$$e(x; h) := \eta(x; h) - y(x) \doteq h^2[u_1(x) + (-1)^n\, v_1(x)]$$

Der zweite Term oszilliert. Die Mittelpunktsregel heißt daher auch "schwach instabil". Der führende oszillierende Fehlerterm $(-1)^n\, v_1(x)$ kann jedoch durch einen Trick eliminiert werden. Dazu definiert man die Graggfunktion

$$S(x; h) = \frac{1}{2}\left[\eta_{n+1} + \eta_n - h\, f(x, \eta_n)\right]\ .$$

Gragg konnte zeigen, daß diese Funktion ebenfalls eine quadratische Entwicklung besitzt, bei der der führende Fehlerterm jedoch keine oszillierend Komponente mehr besitzt. Dies kostet jedoch eine zusätzliche Funktionsauswertung für jede Schrittweite.
Dennoch muß man bei den Näherungen unterscheiden mit wie vielen Schritten sie berechnet wurden. Nur wenn man sich darauf beschränkt nur gerade oder ungerade Schrittzahlen zu verwenden, sind die Näherungen von der Bauart (1.5.1). Da sich alle Näherungen die mit ungerader Schrittzahl berechnet wurden auf $\eta(x_0+h; h)$ "abstützen", also auf einer Näherung erster Ordnung sind sie alle etwas schlechter als die Näherungen die mit gerader Schrittzahl berechnet wurden. Man verwendet daher in der Praxis stets nur Schrittweiten $h_i = \frac{x-x_0}{2k}$. Dennoch ist der "schlechte" Startschritt verantwortlich für den oszillierenden Fehlerterm. Ersetzt man den Startschritt durch ein geeignetes Verfahren der Ordnung 2, so kann der führende oszillierenden Fehlerterm ebenfalls eliminiert werden. Ein Vorteil ist, daß die nachfolgenden Auswertungen von f bereits mit genaueren Approximationen als Argumenten gemacht werden, und nicht erst im nachinein, durch Rechentricks Fehlerterme eliminiert werden müssen. Ein anderer Vorteil ist, daß dabei Funktionsauswertungen gespart werden können. Der Startschritt

$$\eta(x_0 + h; h) := y_0 + hf(x_0, y_0)$$

wird dabei ergänzt um

$$\begin{aligned}
\bar{\eta}(x_0 + \bar{h}; \bar{h}) &:= y_0 + \bar{h} f(x_0, y_0) \\
c(x_0, y_0) &:= \frac{f(x_0 + \bar{h}, \bar{\eta}(x_0 + \bar{h}; \bar{h})) - f(x_0, y_0)}{\bar{h}} \\
\eta(x_0 + h; h) &:= y_0 + h f(x_0, y_0) + \frac{h^2}{2} c(x_0, y_0)
\end{aligned}$$

Die Berechnung von $c(x_0, y_0)$ kostet dabei nur eine zusätzliche Auswertung, kann aber für alle $\eta(x_0 + h_i; h_i)$ verwendet werden. $\bar{h}$ wird dabei in der Größenordnung der kleineren h_i gewählt. (Computing 41, 131-136 (1989))

Die extrapolierte Mittelpunktsregel ist nun ein Einschrittverfahren, bei dem in jedem Makroschritt mit Schrittweite H ausgehend von $(x, \eta(x))$ Näherungen $\eta(x + H; h_i)$ an der Stelle $x + H$ mit der Mittelpunktsregel als 2-Schrittverfahren und verschiedenen Schrittweiten $h_i - \frac{H}{n_i}$ berechnet werden. Dabei werden nur gerade n_i verwendet.

Üblich ist etwa die **doppelt harmonische Folge** $n_i = 2i + 2\,,\ i = 0, 1, \ldots$. Dies führt zu der höchsten Ordnung bei geringster Zahl von Funktionsaufrufen. Verwendet wird aber auch die sogenannte **Bulirsch Folge**[13] $n_i = 2, 4, 6, 8, 12, 16, \ldots$, d.h., $n_i = 2n_{i-2}$ für $i \geq 3$. Dies verkleinert die Extrapolationsgewichte und reduziert damit den Einfluß von Rundungsfehlern und ist vorteilhaft wenn die geforderte Genauigkeit in der Nähe der Maschinengenauigkeit liegt ($TOL/\varepsilon \leq 10^3$).

Der zugehörige Algorithmus lautet dann:

[13] Bulirsch konnte zeigen, daß bei dieser Folge die Summe der Extrapolationsgewichtsbeträge beschränkt bleibt, auch wenn man die Ordnung immer weiter erhöht. Dies erlaubt es in den meißten Fällen Konvergenz durch fortgesetzte Ordnungserhöhung zu erzielen (extrapolation to the limit). Da in der Praxis aber stets eine maximale Ordnung festgeschrieben wird und Konvergenz durch Verkleinerung der Schrittweite erreicht wird, ist dieses Argument nicht so wichtig. Die Rundungsfehler sind aber bei Verwendung der Bulirsch Folge bei hoher Ordnung (16-20) etwa um den Faktor 10 kleiner als bei der doppelt harmonischen Folge.

Extrapolierte Mittelpunktsregel, Teil 1.

`Gegeben:`	$x_0, \quad y_0$
	H: Makroschrittweite
	k: Anzahl der Näherungen
	n_i: Schritteweitenfolge
`Vorbereitung:`	$\bar{h} = H/n_k$: kleinste verwendete Schritteweite
	$f_0 := f(x_0, y_0)$: Auswertung am Startpunkt
	$c := \frac{f(x_0+\bar{h}, y_0+\bar{h}f_0)-f(x_0,y_0)}{\bar{h}}$: Näherung für $y''(x_0)$
`Berechne:`	$\hat{x} := x_0 + H$
	$h_i := H/n_i$: Mikroschritteweite
	$\eta_0 := y_0$
	$\eta_1 := \eta_0 + h_i\, f(x_0, \eta_0) + \frac{h_i^2}{2} c, \quad x_1 := x_0 + h_i$
	Für $j = 1, 2, \ldots, n-1$
	$\quad \eta_{j+1} := \eta_{j-1} + 2h_i\, f(x_j, \eta_j), \quad x_{j+1} := x_j + h_i$

Dann gilt

$$\eta(x,h) = y(x) + \sum_{j=1}^{N} e_j(x) h^{2j} + E_{N+1}(x,h) h^{2N+1} = p(h^2; x) + E_{N+1}(x,h) h^{2N+1}$$

mit $e_j(x_0) = 0$ und $e_j(x)$ differenzierbar. $\eta(x,h)$ kann also wieder näherungsweise geschrieben werden als ein Polynom p_k vom Grad k, jedoch diesmal in $t := h^2$.

$$p_k(t) = y(x) + \sum_{j=1}^{k} e_j(x) h^{2j} = y(x) + \sum_{j=1}^{k} e_j(x) t^j$$

Die $\eta_{i,k}$ lassen sich wieder nach dem Algorithmus von Aitken-Neville berechnen. In der ersten Spalte stehen dann die Näherungen $\eta_{i,0} := \eta(x, h_i)$ der Mittelpunktsregel. In Spalte k jeweils eine extrapolierte Näherung der Ordnung $2k$. Die tieferen Spalten wurden dabei mit kleinerern Schrittweiten erzeugt, d.h., der Fehler in Zeile $i+1$ und Spalte k ist von der Ordnung $\mathcal{O}(h_i^2 \cdots h_{i+k}^2)$.

Extrapolierte Mittelpunktsregel, Teil 2:.

`Berechne:` Für $i = 0, 1, 2, \ldots$
$\eta_{i,0} := \eta(x, h_i)$
For $1 \leq k \leq i$
$$\eta_{i,k} := \eta_{i,k-1} + \frac{\eta_{i,k-1} - \eta_{i-1,k-1}}{\left(\frac{h_{i-k}}{h_i}\right)^2 - 1}$$

Bemerkung: Die Auswertung von p_k an der Stelle $t = 0$, also außerhalb des Bereiches $t > 0$ in dem die Stützstellen liegen sollte nach den Resultaten der Interpolation eigentlich zu sehr schlechter Kondition führen. In diesem Fall entspricht dies jedoch wegen der Symmetrie einer Interpolation von $q(h) = p_k(t)$ zu Stützstellen $h := \pm h_i$. In Wahrheit haben wir es also mit einer Interpolation in der Intervallmitte und zur Mitte gehäuften Stützstellen zu tun. Dies ist besonders stabil. Dennoch ist die Summe der Interpolationsgewichte bei der harmonischen Folge $n_i = 2i + 2$ und hoher Ordnung größer als 100. Man verliert also 2 Stellen der Maschinengenauigkeit. Bei hohen Genauigkeitsanforderungen, der Stärke der Extrapolationsverfahren, kann dies entscheidend sein und zu anderen Schrittweitenfolgen zwingen.
Implementationen:
DIFSY1: Bulirsch, Stoer (letzte Version 1973)
DIFEX1: Deuflhard (letzte Version 1983) [**?**]
DIFEXM: Kiehl, Zenger, (letzte Version 1987)
ODEX: [**?**]
Außer der expliziten Mittelpunktsregel wird noch die **implizite Mittelpunktsregel**

$$\begin{aligned} k_1 &= f(x_i + h/2, \eta_i + h/2k_1) \\ \eta_{i+1} &= \eta_{i-1} + hk_1 \ , \end{aligned}$$

und die **implizite Trapezregel**

$$\eta_{i+1} = \eta_{i-} + h/2[f(x_i, \eta_i) + f(x_{i+1}, \eta_{i+1})]$$

mit jeweils quadratischer Fehlerentwicklung zur mehrfachen Extrapolation verwendet

Aufgabe 1.5.5 Eulerverfahren und Mittelpunktsregel Mit dem extrapolierten Eulerverfahren und Schrittweiten $h_i = H/1, H/2, \ldots, H/20$ und der extrapolierten Mittelpunktsregel und Schrittweiten $h_i = H/2, H/4, \ldots, H/20$ werde die Näherung $\eta(x + H)$ mit Konsistenzordnung 20 approximiert.
Wieviele Auswertungen der rechten Seite f sind jeweils für einen Makroschritt notwendig?
Mit welchem Faktor gehen die Auswertungen von f maximal in die nächste Näherung ein? Man gebe jeweils das betragsgrößte Extrapolationsgewicht an.

1.5.3 Ordnungssteuerung

Für die Effizienz eines bestimmten Verfahrens ist nicht nur die Ordnung von Bedeutung, sondern auch die Fehlerkonstanten vor den entsprechenden Fehlertermen. Hat ein Verfahren eine groß Fehlerkonstante vor dem Term $f_{yy}f_yfff$, so ist dies bei Problemen mit $f_{yy}f_yfff \approx 0$ nicht besonders tragisch. Die optimale Verfahrenswahl hängt damit auch entscheidend vom Problem ab. Insbesondere können bei verschiedenen Problemen Verfahren verschiedener Ordnung sinnvoll sein. Die optimale Ordnung kann sich dabei sogar bei der Behandlung eines einzigen Problems während der Integration ändern.
Verfahren die es erlauben ohne großen Aufwand Approximationen verschiedener Ordnungen zu berechnen haben dann einen wesentlichen Vorteil. Zu diesen Verfahren gehören insbesondere die Extrapolationsverfahren.
Berechnet man mit einem Extrapolationstableau Näherungen $\eta_{0,0}, \ldots, \eta_{k,k}$, so ist $\eta_{i,i}$ eine Approximation der Ordnung $\mathcal{O}(h^{p_i})$, mit $p_i = i$ oder $p_i = 2i$, je nachdem das Ausgangsverfahren die Ordnung 1 (Euler) oder 2 (Trapezregel oder Mittelpunktsregel) hat.
$\|\eta_{i,i-1} - \eta_{i,i}\|$ ist dann ein Maß für den Fehler von $\eta_{i,i-1}$ bzw. $\eta_{i,i}$ und gemäß (1.3.10)

$$h_i^* := h \sqrt[p_i]{\frac{h\mathtt{TOL}}{Err(x+h)}} = h \sqrt[p_i]{\frac{h\mathtt{TOL}}{\|\eta_{i,i-1} - \eta_{i,i}\|}} ,$$

so erhält man zu jeder Ordnung p_i, $i = 0, ldots, k$ eine zugehörige Schrittweite. Bei höherer Ordnung ist dabei in der Regel eine größerer Schrittweite möglich. Allerdings bedeutet hohe Ordnung auch einen größeren Aufwand A_i. Den Aufwand eines Verfahrens mißt man grob durch die Anzahl der notwendigen Stufen. Bei expliziten Verfahren entspricht dies der Anzahl der Auswertungen der rechten Seite.

Beispiel 1.5.6 Bei der Mittelpunktsregel mit Schrittweitenfolge n_i, gilt $A_0 = n_0 + 1$ und $A_{k+1} = A_k + n_{k+1}$ bzw. $A_{k+1} = A_k + n_{k+1} - 1$, je nachdem, ob eine Startschrittkorrektur oder der Gragg-Schlußschritt verwendet wird. Der Ausdruck

$$W_i := \frac{A_{i+1}}{h_i^*}$$

gibt dann den normierten Aufwand pro Schrittweite (work per unit step) an. ($\eta_{k+1,0}$ benötigt man, um den Fehler von $\eta_{k,k}$ bzw $\eta_{k+1,k}$ zu schätzen.)
Man kann dann die Ordnung des Tableaus folgendermaßen wählen: Man versucht

$$W_i = \min_{j=0,\ldots,k-1} W_j$$

zu erreichen. Ist dies für $i < k-1$ der Fall, so veringert man die Ordnung um eins und berechnet im nächsten Schritt nur $\eta_{0,0}, \ldots, \eta_{k-1,0}$ und wählt die Schrittweite entsprechend. Ist dies für $i = k-1$ der Fall, und gilt sogar

$$W_{k-1} < 0.9 W_{k-2} \ ,$$

so erhöht man die Ordnung um eins und berechnet im nächsten Schritt nur $\eta_{0,0}, \ldots, \eta_{k+1,0}$. In allen anderen Fällen behält man die Ordnung bei.

1.6 Steife Differentialgleichungen

Problem: In vielen Anwendungen ändert sich die Lösung des AWPs

$$y' = f(x, y) \ , \quad y(x_0) = y_0 \tag{1.6.1}$$

(ab einer gewissen Zeit) sehr wenig, aber kleine Störungen werden schnell gedämpft.

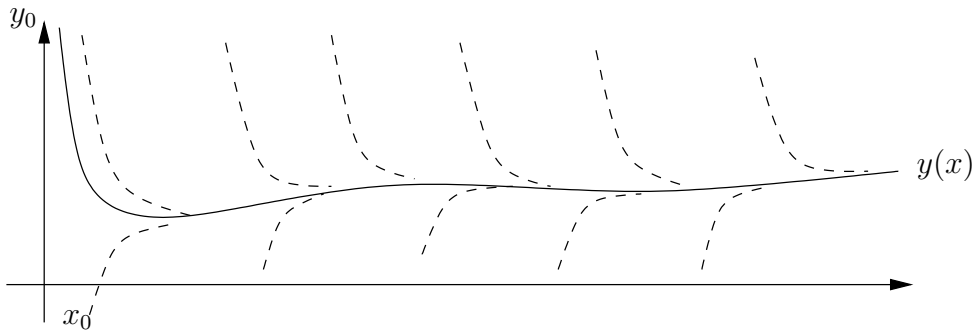


Abbildung 1.9: Stabile Lösung

Die Stabilität einer Lösung kann folgendermaßen charakterisiert werden. Ist

$$v' = f(x, v) \ , \quad v(x_0) = v_0 := y_0 + \delta_0$$

ein benachbartes Problem, so genügt die Abweichung $e(x, \delta_0)$ dem Anfangswertproblem

$$e'(x, \delta_0) = g(x, e) = f(x, y(x) + e) - f(x, y(x)) \quad ; \quad e(x_0, \delta_0) = \delta_0 \tag{1.6.2}$$

bzw näherungsweise für kleine Intervalle

$$e'(x, \delta_0) = f_y(x, y(x))e(x, \delta_0) + \mathcal{O}(e(x, \delta_0)^2) \approx f_y(x_0, y(x_0))e(x, \delta_0)$$

Die Störung genügt also in erster Näherung einer linearen Differentialgleichung mit Koeffizientenmatrix $f_y(x_0, y(x_0))$, und die Störungsanteile in Richtung der Eigenvektoren von $f_y(x_0, y(x_0))$ werden unabhängig voneinander exponentiell gedämpft oder verstärkt (vgl. Lösungstheorie zu gewöhnlichen Differentialgleichungen).

Definition 1.6.1 *Die Lösung $y(x)$ von (1.6.1) heißt* **stabil** *oder* **Ljapunov-stabil**, *wenn für alle $\varepsilon > 0$ ein $\delta > 0$ gibt, so daß*

$$\|e(x, \delta_0)\| \leq \varepsilon \ \forall x \geq x_0 \quad \textit{falls} \quad \|\delta_0\| \leq \delta \ .$$

Da bei numerischen Verfahren kleine Fehler zu Beginn nicht vermieden werden können ist Ljapunovstabilität eine wesentliche Voraussetzung, wenn man Anfangswertprobleme über sehr lange Zeiträume hinweg numerisch lösen will. Ist ein Problem Ljapunov-stabil, so erwartet man natürlich auch von vernünftigen numerischen Verfahren eine analoge Eigenschaft. Dies ist jedoch nicht selbstverständlich.

Definition 1.6.2 *Die Lösung $y(x)$ von (1.6.1) bzw. das Anfangswertproblem (1.6.1) heißt* **asymptotisch stabil** *wenn sie Ljapunov-stabil ist und zusätzlich gilt*

$$\lim_{x\to\infty} \|e(x,\delta_0)\| = 0 \quad \textit{falls} \quad \|\delta_0\| \le \delta \ .$$

Kleine Fehler bleiben in ihrere Wirkung also nicht nur beschränkt, sondern sie klingen irgendwann vollständig ab.
Bei Systemen mit Erhaltungsgrößen H klingen Fehler in H nicht ab, werden aber auch nicht verstärkt. Eine sehr große und wichtige Klasse von Problemen ist daher Ljapunov-stabil, aber nicht asymptotisch stabil.

Beispiel 1.6.3 Wir betrachten ein Beispiel aus der Reaktionskinetik.

$$\begin{array}{lcl} Y_1 + Y_2 & \underset{1}{\overset{100}{\rightleftarrows}} & Y_3 \\ Y_2 + Y_2 & \underset{1}{\overset{10^4}{\rightleftarrows}} & Y_4 \end{array}$$

mit der zugehörigen Differentialgleichung und Anfangswerten

$$\begin{array}{rclrcl} y_1' &=& y_3 - 100 y_1 y_2 & y_1(0) &=& 1 \\ y_2' &=& y_3 - 100 y_1 y_2 - 2\cdot 10^4 y_2^2 + 2 y_4 & y_2(0) &=& 1 \\ y_3' &=& -y_3 + 100 y_1 y_2 & y_3(0) &=& 0 \\ y_4' &=& 10^4 y_2^2 - y_4 & y_4(0) &=& 0 \end{array}$$

Das Problem ist Ljapunov-stabil. Die Konzentrationen y_i streben sehr schnell einem stationären Punkt zu und ändern sich dann nicht mehr. An diesem Grenzzustand y_l gilt dann $y' = 0$. y_l ist gegeben durch

$$y_l \approx (0.63976, 0.00563, 0.36024, 0.317065).$$

Bemerkung: y_2 startet bei $y_2(0) = 1$, erreicht seinen Endzustand $y_2 \approx 0$ aber so schnell, daß auf dem Plot von dem Übergang fast nichts zu sehen ist.

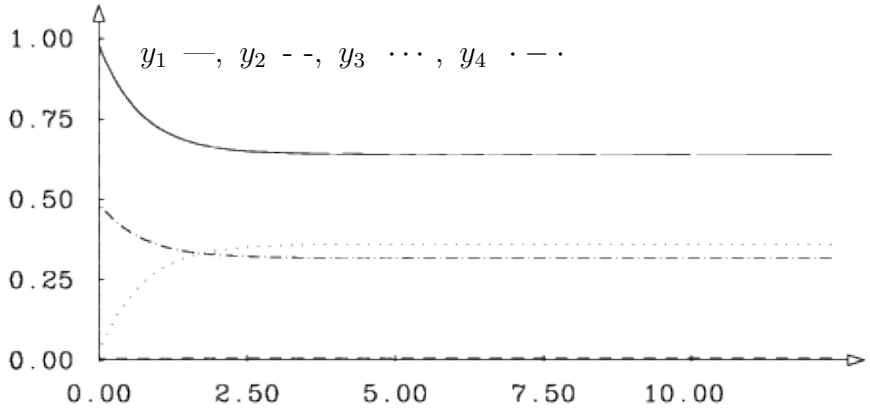


Abbildung 1.10: Problem D3 aus STIFF-DETEST

Beispiel 1.6.4 Beispiel 1.6.3 ist Ljapunov-stabil aber nicht asymptotisch stabil, denn es besitzt die Erhaltungsgrößen $y_1 + y_2 + 2y_3 + 2y_4 = 2$ und $y_1 + y_3 = 1$.

Manchmal kann die Abklinggeschwindigkeit von Störungen sogar abgeschätzt werden.

Definition 1.6.5 *Die Lösung $y(x)$ von (1.6.1) bzw. das Anfangswertproblem (1.6.1) heißt* **exponentiell stabil** *wenn gilt: Es gibt $a, b, \delta > 0$ mit*

$$\|e(x, \delta_0)\| \leq a e^{(-b(x-x_0))} \|\delta_0\| \ \forall\ x > x_0 \quad \textit{falls} \quad \|\delta_0\| \leq \delta\ .$$

Ist etwa $f_y(x, y)$ unabhängig von x, und gilt für alle Eigenwerte von $f_y(y)$: $\mathrm{Re}\,(\lambda) < b < 0$, so ist das Problem exponentiell stabil.

Beispiel 1.6.6 Beispiel 1.6.3 ist nach Transformation auf den Raum $\mathrm{span}\{(1, 1, -1, 0)^T, (0, 2, 0, -1)^T)\}$ exponentiell stabil.
Es gilt stets:

$$y(x) = y(0) + u_1(x) r_1 + u_2(x) r_2$$

mit

$$r_1 = (-1, -1, 1, 0)^T \quad ; \quad r_2 = (0, -2, 0, 1)^T\ .$$

Aufgabe 1.6.7 Man gebe für Beispiel 1.6.3 bzw. Beispiel 1.6.3 die Differentialgleichung $u' = g(x, u)$ an und untersuche sie auf exponentielle Stabilität.

Bei sehr stabilen Problemen haben manche numerische Verfahren jedoch unerwartete Schwierigkeiten.

Beispiel 1.6.8 Die lineare Differentialgleichung

$$y' = f(y) := \begin{bmatrix} \mu & 0 \\ 0 & \lambda \end{bmatrix} y \quad ; \quad y(0) = y_0$$

hat die Lösung

$$y(t) = \begin{bmatrix} y_1(0)e^{\mu t} \\ y_2(0)e^{\lambda t} \end{bmatrix} y \ .$$

Das explizite Eulerverfahren liefert

$$y(x+h) = y(x) + h \begin{bmatrix} \mu & 0 \\ 0 & \lambda \end{bmatrix} y(x) = \begin{bmatrix} (1+h\mu)y_1(x) \\ (1+h\lambda)y_2(x) \end{bmatrix}$$

und wir erhalten

$$y_1(x+ih) = (1+h\mu)^i y_1(x) \quad ; \quad y_2(x+ih) = (1+h\lambda)^i y_2(x)$$

Im Falle $\mu = -1000$ und $\lambda = 1$ explodiert $y_1(x+ih)$ falls nicht $h < 1/5000$ während die tatsächliche Lösung durch die zweite Komponente bestimmt wird. Das implizite Eulerverfahren liefert dagegen

$$y(x+h) = y(x) + h \begin{bmatrix} \mu & 0 \\ 0 & \lambda \end{bmatrix} y(x+h)$$

$$\Rightarrow y(x+h) = \begin{bmatrix} \frac{1}{1-h\mu} y_1(x) \\ \frac{1}{1-h\lambda} y_2(x) \end{bmatrix}$$

und wir erhalten

$$y_1(x+ih) = \frac{1}{(1-h\mu)^i} y_1(x) \quad ; \quad y_2(x+ih) = \frac{1}{(1-h\lambda)^i} y_2(x)$$

so daß der y_2-Anteil, welcher schnell gedämpft wird, auch numerisch klein bleibt.

Dieses Verhalten ist typisch für steife Differentialgleichungen.
Bei exponentiell stabilen Problemen, aber auch bei instabilen Problemen ist die Fehlerabschätzungfortpflanzung mit einer Lipschitzkonstanten $L = \|f_y\|$ gemäß Fundamentallemma 1.1.7 sehr grob. In Beispiel 1.6.8 ist im Falle $\mu < 0 < \lambda$ und $|\lambda| \ll |\mu|$, zwar $\|f_y\| = |\mu|$ und damit $|\mu|$ die kleinste Lipschitzkonstante gemäß Fundamentallemma 1.1.7. Die Instabilität der Lösung wird aber bestimmt durch $|\lambda|$.
Als geeigneter erweist sich dafür die einseitige Lipschitzbedingung.

Definition 1.6.9 *Sei* $< .,. >$ *ein Skalarprodukt und* $\|x\|^2 =< x, x >$. $f(t,u)$ *genügt einer* **einseitigen Lipschitzbedingung** *in* S, *falls*

$$< f(t,u) - f(t,v); u - v > \leq l(t)\|u - v\|^2 \quad \forall t \geq t_0,\ u, v \in S\ .$$

$l(t)$ *heißt* **einseitige Lipschitzkonstante**. *(Sie kann auch kleiner 0 sein.)*

In Beispiel 1.6.8 ist $|\lambda|$ eine einseitige Lipschitzkonstante.
Existiert eine Lipschitzkonstante, so auch eine einseitige Lipschitzkonstante, aber nicht notwendig umgekehrt.

Definition 1.6.10 *Das Anfangswertproblem*

$$y' = f(y) \quad ; \quad y(t_0) = y_0$$

heißt **dissipativ** *oder kontraktiv, wenn für zwei benachbarte Lösungen* $y_1(t), y_2(t)$ *und eine geeignete Norm gilt:*

$$\|y_1(t_2) - y_2(t_2)\| \leq \|y_1(t_1) - y_2(t_1)\| \ \forall t_0 < t_1 < t_2 < \infty$$

Bei solchen Problemen nähern sich je zwei Lösungen monoton an. Die Probleme sind daher Ljapunov-stabil, jedoch nicht notwendig asymptotisch stabil.

Satz 1.6.11 $f(t,y)$ *besitze in einem Schlauch* S *eine einseitige Lipschitzkonstante* $l(t)$. *Dann gilt für 2 beliebige Lösungen* $u, v \in S$ *von* $y' = f(t,y)$ *in der entsprechenden Norm:*

$$\|u(t_2) - v(t_2)\| \leq \exp\left(\int_{t_1}^{t_2} l(\tau)\mathrm{d}\tau\right) \|u(t_1) - v(t_1)\|\ .$$

Beweis: Definiere

$$\varphi(t) := \|u(t) - v(t)\|^2 =< u(t) - v(t); u(t) - v(t) >$$

φ ist stetig differenzierbar mit

$$\begin{aligned} \varphi'(t) &= 2 < u'(t) - v'(t); u(t) - v(t) > \\ &= 2 < f(t,u(t)) - f(t,v(t)); u(t) - v(t) > \\ &\leq l(t) < u(t) - v(t); u(t) - v(t) >= 2l(t)\varphi(t) \end{aligned}$$

Sei weiter

$$\eta(t) := \exp\left(-2\int_{t_0}^{t} l(\tau)\mathrm{d}\tau\right)\ .$$

Dann gilt:

$$\begin{aligned}(\varphi\eta)' &= \varphi'\eta + \varphi\eta' = \varphi'\eta - 2l(t)\varphi\eta \\ &= \eta(\varphi' - 2l(t)\varphi) \leq 0 \quad \forall\, t \geq t_0\ .\end{aligned}$$

also ist $\varphi(t)\eta(t)$ monoton fallend und es gilt

$$\varphi(t_2) \leq \varphi(t_1)\frac{\eta(t_1)}{\eta(t_2)} = \varphi(t_1)\exp\left(2\int_{t_1}^{t_2} l(\tau)\mathrm{d}\tau\right)\ .$$

■

Es gilt also:

Lemma 1.6.12 *Sei $l(t) \leq 0$ eine einseitige Lipschitzkonstante von $f(t,y)$, dann ist die Differentialgleichung $y' = f(t,y)$ dissipativ. Gilt $l(t) \leq l_0 < 0$, dann ist die Lösung $y(t)$ von $y' = f(t,y)$, exponentiell stabil.*

Beweis: Folgt aus Satz 1.6.11 ■

Bei diagonalisierbaren f_y bestimmte also $l = \max$ Re λ_i die Empfindlichkeit der Lösung gegen kleine Störungen.
Die Beschränkung des lokalen Fehlers bei Verwendung eines Verfahrens der Ordnung p und einer gewünschten Genauigkeit TOL erzwingt

$$h\tau = C(x)h^{p+1} < TOL \Rightarrow h \leq \sqrt[p+1]{\frac{TOL}{C(x)}} \tag{1.6.3}$$

$C(x)$ hängt dabei von den höheren partiellen Ableitungen von f ab.
Die Fehlerfortpflanzung des lokalen Fehlers erzwingt zusätzlich

$$h\tau e^{l(x-x_0)} \approx C(x)h^{p+1}e^{l(x-x_0)} \leq \frac{h\,TOL}{x-x_0} \Rightarrow h \leq \sqrt[p]{\frac{TOL}{C(x)\,(x-x_0)}e^{-l(x-x_0)}}\ . \tag{1.6.4}$$

Dies ist bei instabilen Systemen ($l > 0$) die stärkere Beschränkung
Im Falle $l < 0$ genügt dagegen meist $h\tau < TOL$ Leider kann man aber nicht immer erwarten, daß solche exponentiell stabile Differentialgleichungen automatisch auch numerisch leicht bzw. billig gelöst werden können.
Im Beispiel 1.6.3 (siehe Abb. 1.6.3) wird y_l in der sogenannten **transienten Phase** sehr schnell erreicht.
Während des schnellen Übergangs erwartet man kleine Schrittweiten da $C(x)$ hier sehr großt ist. Verwendet man aber z.B. das explizite Euler-Verfahren,

so sind die Schrittweiten auch in der sogenannten **stationären Phase**, in der sich fast nichts mehr ändert ($C(x)$ klein), sehr klein.
Erstaunlicherweise tritt dieses Problem nicht auf, wenn man das implizite Euler-Verfahren verwendet. Dies wollen wir genauer untersuchen.

1.6.1 Stabilitätsfunktion, A-Stabilität

Zur einfacheren Analyse dieses Phänomens betrachten wir erst einmal eine lineare Differentialgleichung.

Beispiel 1.6.13 Gegeben sei das Anfangswertproblem

$$\begin{aligned} y_1' &= \frac{\lambda_1+\lambda_2}{2} y_1 + \frac{\lambda_1-\lambda_2}{2} y_2 \\ y_2' &= \frac{\lambda_1-\lambda_2}{2} y_1 + \frac{\lambda_1+\lambda_2}{2} y_2 \\ \lambda_1, \lambda_2 &< 0 \ . \end{aligned}$$

Das System ist linear von der Form $y' = Ay$. Die Matrix

$$A = \begin{pmatrix} \frac{\lambda_1+\lambda_2}{2} & \frac{\lambda_1-\lambda_2}{2} \\ \frac{\lambda_1-\lambda_2}{2} & \frac{\lambda_1+\lambda_2}{2} \end{pmatrix}$$

besitzt die Eigenwerte λ_1, λ_2 und die Eigenvektoren $v_1 = (1,1)^T$, $v_2 = (1,-1)^T$.
Die allgemeine Lösung des Systems lautet also

$$\begin{aligned} y_1(x) &= c_1 e^{\lambda_1 x} + c_2 e^{\lambda_2 x} \\ y_2(x) &= c_1 e^{\lambda_1 x} - c_2 e^{\lambda_2 x} \end{aligned}$$

mit Konstanten c_1, c_2. Für $\lambda_1, \lambda_2 < 0$ gilt $\lim_{x\to\infty} y(x) = 0$. Das Problem ist also exponentiell stabil.
Wendet man jedoch das expliziten Euler-Verfahren auf das Problem an, so erhält man

$$\eta_i := \begin{pmatrix} \eta_{1,i} \\ \eta_{2,i} \end{pmatrix} \to \eta_{i+1} = \eta_i + hA\eta_i = (I + hA)\eta_i \ .$$

Wegen

$$Av_i = \lambda_i v_i \Rightarrow (I + hA)v_i = (1 + h\lambda_i)v_i$$

ist jeder Eigenvektor v_i von A auch Eigenvektor von $I + hA$, jedoch zum Eigenwert $\mu_i := 1 + h\lambda_i$. Daher gilt für die Eulerapproximationen

$$\eta_{1,i} = c_1(1 + h\lambda_1)^i + c_2(1 + h\lambda_2)^i$$
$$\eta_{2,i} = c_1(1 + h\lambda_1)^i - c_2(1 + h\lambda_2)^i$$

Also

$$[\lim_{x\to\infty} \eta(x;h) = 0] \Rightarrow [|(1 + h\lambda_1)| < 1 \text{ und } |(1 + h\lambda_2)| < 1] .$$

Im Falle $\lambda_1 = -1$, $\lambda_2 = -1000$, $y(0) = (2, 0)^T$ ($c_1 = c_2 = 1$) erfordert dies: $h < |2/\lambda_2| = 0.002$, wobei die Komponente $e^{\lambda_2 x} = e^{-1000x}$, welche die Schrittweite beschränkt zur Lösung praktisch nichts beiträgt.

$$e^{\lambda_2 x}/e^{\lambda_1 x} = e^{-1000x}/e^{-x} = e^{-999x} \approx 4.10^{-5} \text{ für } x > 0.01$$

Dies gilt allgemeiner.

Lemma 1.6.14 (Fundamentallösung numerischer Approximationen)

Sei $y' = Ay$ mit $T^{-1}AT = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. Die numerische Approximation erfülle

$$Q(hA)\eta_{i+1} = P(hA)\eta_i \tag{1.6.5}$$

mit Polynomen P und Q. Dann gilt

$$\eta_i = \sum_{j=1}^{n} c_j (R(h\lambda_j))^i v_j$$

mit Eigenwerten λ_j, Eigenvektoren v_j von A. und $R(z) = P(z)/Q(z)$.

D.h., jede Eigenkomponente $c_j v_j$ entwickelt sich unabhängig und wird behandelt wie das skalare Problem $y' = \lambda_j y$. Damit ist die Testgleichung (1.6.6) repräsentativ auch für mehrdimensionale Probleme (1.6.7) falls A diagonaliserbar. Ist A nicht diagonaliserbar, so existiert aber eine beschränkte Matrix B mit $A+\varepsilon B$ diagonaliserbar für alle $0 < \varepsilon < \varepsilon_0$. Ist das numerische Verfahren stetig bezüglich der rechten Seite f, dann ist (1.6.6) auch repräsentativ für nicht diagonaliserbares A.

Beweis: v Eigenvektor von A zum Eigenwert $\lambda \Rightarrow v$ Eigenvektor von $P(A)$ und $Q(A)$ mit Eigenwerten $P(\lambda)$ bzw. $Q(\lambda)$. $\Rightarrow v$ Eigenvektor des verallgemeinerten Eigenwertproblems $Q(hA)v = P(hA)v$ mit Eigenwert

$R(h\lambda)$. Daher ist v Eigenvektor der Rekursionsgleichung und die zugehörigen Eigenwerte sind $R(h\lambda_j)$. ∎
Wir sehen also, daß das Verhalten eines numerischen Verfahrens bei linearen Problemen bereits durch sein Verhalten bei der **skalaren Testgleichung**

$$y' = \lambda y \ , \quad y \in \mathbb{R} \ , \quad \lambda \in \mathbb{C} \tag{1.6.6}$$

wesentlich bestimmt wird. (Im höherdimensionalen Fall spielt λ dann die Rolle der Eigenwerte von f_y .) Für diese Testgleichung liefern lineare Verfahren (fast alle verwendeten Verfahren insbesondere alle hier besprochenen, sind linear) Rekursionen des Typs (1.6.5).

Definition 1.6.15 (Stabilitätsfunktion)
Erzeugt ein numerisches Verfahren angewendet auf (1.6.6) Näherungen

$$\eta(x+h,h) = \eta(x,h)R(\lambda h) \ ,$$

so heißt $R(z)$ *mit* $z = \lambda h$ **Stabilitätsfunktion** *des Verfahrens.*

Lemma 1.6.16 *Runge-Kutta-Verfahre und ROW-Verfahren erfüllen (1.6.5) aus Lemma 1.6.14 und (1.6.15) aus Definition 1.6.15. Dabei gilt:*
(i) Im Falle eines s *-stufigen expliziten Runge-Kutta-Verfahrens ist* $R(z)$ *ein Polynom vom Grad* s *.*
(ii) Bei s *-stufigen impliziten Runge-Kutta-Verfahren und ROW-Verfahren ist* $R(z)$ *eine rationale Funktion mit Zählergrad und Nennergrad* $\leq s$ *.*

Beweis: Zur Übung (Man zeige entsprechende Behauptungen für jedes k_i .) ∎

Beispiel 1.6.17 (Stabilitätsfunktion des explizitenEuler-Verfahrens)

Das explizite Euler-Verfahren liefert ganz allgemein bei dem linearen Problem

$$y' = Ay \tag{1.6.7}$$

eine Rekursion der Art

$$\eta_{i+1} = \eta_i + hf(x_i,\eta_i) = \eta_i + hA\eta_i = (1+hA)\eta_i =: R(hA)\eta_i \tag{1.6.8}$$

Beispiel 1.6.18 (Stabilitätsfunktion des Heun-Verfahrens)

$$\begin{aligned}
\eta_{i+1} &= \eta_i + \frac{h}{2} f(x_i, \eta_i) + \frac{h}{2} f(x_{i+1}, \eta_i + h f(x_i, \eta_i)) = \\
&\overset{(1.6.6)}{=} \eta_i + \frac{h}{2}\lambda\eta_i + \frac{h}{2} f(x_{i+1}, \eta_i + h\lambda\eta_i) = \\
&\overset{(1.6.6)}{=} \eta_i + \frac{h}{2}\lambda\eta_i + \frac{h}{2}\lambda(\eta_i + h\lambda\eta_i) = \\
&= \eta_i \left(1 + \lambda h + \frac{(\lambda h)^2}{2}\right) = \eta_i R(\lambda h) \\
R(z) &:= 1 + z + \frac{z^2}{2}
\end{aligned}$$

η_{i+1} berechnet sich also aus η_i durch Multiplikation mit R, wobei R einem Polynom zweiten Grades in hA ist

Beispiel 1.6.19 (Stabilitätsfunktion desimpliziten Euler-Verfahren)

$$\eta_{i+1} = \eta_i + h f(x_i + h, \eta_{i+1}) \Longleftrightarrow \eta_{i+1} - \eta_i + h f(x_i + h, \eta_{i+1}) \overset{!}{=} 0$$

$$\eta_{i+1} = \eta_i + h\lambda\eta_{i+1} \Longrightarrow \eta_{i+1} = \eta_i \tfrac{1}{1-h\lambda} \Longrightarrow R(z) = \tfrac{1}{1-z}$$

Kleine Störungen der Lösung $y(x, x_0, y_0)$ von (1.6.6) werden von dem numerischen Lösungsverfahren nicht verstärkt, falls gilt:

$$|R(z)| \leq 1$$

Sie sind also eigentlich für die weitere Simulation unerheblich. Sie werden aber bei der numerischen Rechnung bedeutsam, wenn sie verstärkt werden. Jedes numerische Verfahren verstärkt Störungen unterschiedlich, und abhängig von den entsprechenden Eigenwerte von f_y.

Definition 1.6.20 (Stabilitätsgebiet)
Gilt für ein numerisches Verfahren angewendet auf (1.6.6) $\eta_{i+1} = R(\lambda h)\eta_i$, *so heißt* $\mathcal{S} := \{z \in \mathbb{C} \mid |R(z)| \leq 1\}$ **Stabilitätsgebiet** *des Verfahrens.*

Beispiel 1.6.21

$$y' = -1000y\ , \quad y(0) = 1$$

hat exakte Lösung

$$y(x) = y_0 e^{-1000x} \approx 0 \quad \text{für } x \gg 0 \ .$$

Sei $\texttt{TOL} = 10^{-8}$ die verlangte absolute Genauigkeit. Für $x > 0.02$ gilt dann $0 < y(x) < 10^{-8}$, d.h., jeder Wert in $[0, 10^{-8}]$ ist eine hinreichende Approximation. Das explizite Eulerverfahren mit Schrittweite h liefert jedoch bei einer Approximation $0 \neq \eta_i(x) = \delta < 10^{-8}$ eine Folge

$$\eta_j(x) = \delta R^{j-i}(-1000h) = \delta(1 - 1000h)^{j-i}$$

Z.B., für $h = \frac{1}{100}$ wächst $\eta_j(x)$ dann in jedem Schritt um den Faktor 9 und ist bald nicht mehr zulässig. Man ist also gezwungen die Schrittweite $h < \frac{1}{500}$ zu wählen, obwohl die Lösung fast stationär ist. Verwendet man statt dessen das implizite Eulerverfahren, mit

$$|R(h\lambda)| = |R(-1000h)| = \frac{1}{1 + 1000h} < 1 \ ,$$

so können beliebig große Schrittweiten verwendet werden, sobald $|\eta_i(x)| < 10^{-8}$.

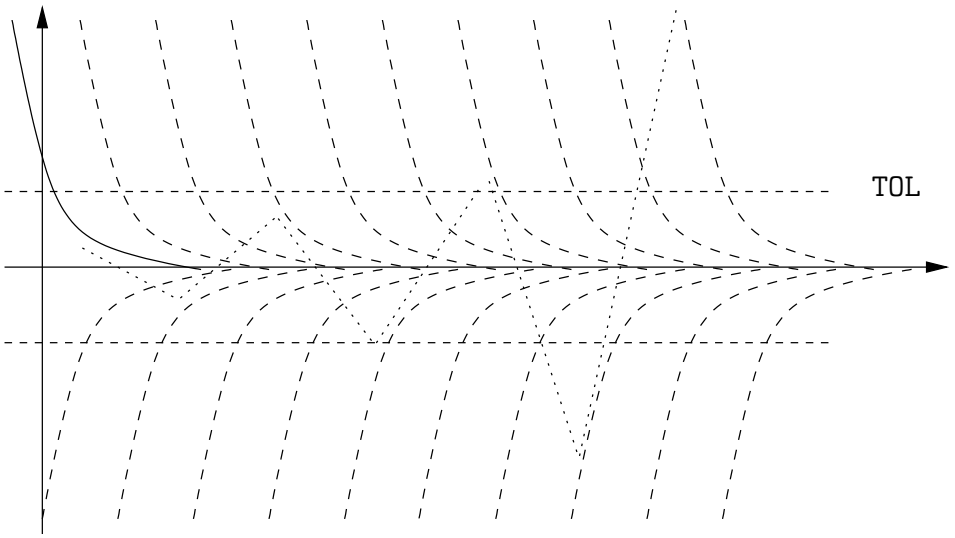


Abbildung 1.11: Instabilität des expliziten Euler-Verfahrens

Die Bedingung $\lambda h \in \mathcal{S}$ beschränkt also die Schrittweite, und zwar umso stärker, je stabiler die stationäre Lösung ist.
Dieses Phänomen tritt nun nicht nur bei dieser speziellen Gleichung, sondern auch ganz allgemein auf.

Beispiel 1.6.22 (von Prothero-Robinson) Sei

$$\dot{y}(t) = f(t,y) = \lambda(y - F(t)) + \dot{F}(t)$$

wobei $F(t)$ irgendeine bestimmte glatte Lösung einer beliebigen Differentialgleichung sei. z.B., $F(t) = e^{-t}$.
Die Lösung ist stabil, d.h., alle anderen Lösungen der Umgebung streben gegen $F(t)$, falls $\operatorname{Re}\lambda < 0$.

$$\lim_{t\to\infty} y(t) - F(t) = 0$$

Wir untersuchen speziell den Fall $\lambda << 0$, $F(0) = F_0 = 1$

$$y(0) = y_0 \Rightarrow y(t) = F(t) + e^{\lambda t}(y_0 - F_0)$$

Das Problem ist dann extrem gut konditioniert. Fehler werden exponentiell gedämpft.
Wir untersuchen nun die numerischen Näherungen die mit dem expliziten Euler-Verfahren im Interval $[0;1]$ und konstanter Schrittweite $h = 1/n$ berechnet werden.

$$y_{m+1} = y_m + hf(t_m, y_m)$$

Wir untersuchen den lokalen Fehler $l_m := l_h(t_m)$ und den globalen Fehler $e_m := e_h(t_m)$

$$\begin{aligned} e_{m+1} &:= y(t_{m+1}) - y_{m+1} \\ l_{m+1} &:= y(t_{m+1}) - [y(t_m) + hf(t_m, y(t_m))] \end{aligned}$$

Dann gilt

$$\begin{aligned} e_{m+1} &:= y(t_{m+1}) - y_{m+1} = y(t_{m+1}) - y_m - hf(t_m, y_m) \\ &= y(t_{m+1}) - y(t_m) - hf(t_m, y(t_m)) + \\ &\qquad + y(t_m) + hf(t_m, y(t_m)) - y_m - hf(t_m, y_m) \\ &= l_{m+1} + e_m + h[f(t_m, y(t_m)) - f(t_m, y_m)] \\ &= l_{m+1} + e_m + h\lambda[y(t_m) - y_m] = l_{m+1} + e_m + h\lambda e_m \\ &= l_{m+1} + (1 + h\lambda)e_m = l_{m+1} + R(h\lambda)e_m \\ &= \sum_{j=1}^{m+1} (1 + h\lambda)^{m+1-j} l_j = \sum_{j=1}^{m+1} R(h\lambda)^{m+1-j} l_j \end{aligned}$$

Falls $|1+h\lambda| >> 1$ so explodiert der globale Fehler auch im Fall kleiner lokaler Fehler. Die Schrittweite h muß daher aus Stabilitätsgründen sehr klein gewählt werden.
Die gleiche Überlegung für das implizite Euler-Verfahren

$$y_{m+1} = y_m + hf(t_{m+1}, y_{m+1}) \ .$$

In diesem Fall ist der globale Fehler definiert durch

$$l_{m+1} := y(t_{m+1}) - [y(t_m) + hf(t_{m+1}, \tilde{u})] = y(t_{m+1}) - \tilde{u}$$

wobei $\tilde{u} = y(t_m) + hf(t_{m+1}, \tilde{u})$ die Näherung des impliziten Euler-Verfahrens mit Startpunkt $y(t_m)$ ist. Dann gilt

$$\begin{aligned}
e_{m+1} &:= y(t_{m+1}) - y_{m+1} = y(t_{m+1}) - y_m - hf(t_{m+1}, y_{m+1}) \\
&= y(t_{m+1}) - y(t_m) - hf(t_{m+1}, \tilde{u}) + \\
&\qquad + y(t_m) + hf(t_{m+1}, \tilde{u}) - y_m - hf(t_{m+1}, y_{m+1}) \\
&= l_{m+1} + e_m + h[f(t_{m+1}, \tilde{u}) - f(t_{m+1}, y_{m+1})] \\
&= l_{m+1} + e_m + h\lambda[\tilde{u} - y_{m+1}] \\
&= l_{m+1} + e_m + h\lambda[y(t_{m+1}) - l_{m+1} - y_{m+1}] \\
&= (1 - h\lambda)l_{m+1} + e_m + h\lambda[e_{m+1}] \\
\Rightarrow e_{m+1} &= l_{m+1} + \frac{1}{(1-h\lambda)}e_m = l_{m+1} + R(h\lambda)e_m \ .
\end{aligned}$$

Die Fehler werden also im Fall $\lambda < 0$ gedämpft.

Das Beispiel von Prothero-Robinson zeigt, daß für Stabilitätsuntersuchungen linearer Systeme mit konstanter Jacobimatrix f_y die Analyse des Systems $y' = Ay$ genügt. λ wird dann ersetzt durch $f_y = A$. Da für kurze Zeiträume f_y stets näherungsweise konstant ist, sind die Ergebnisse auch für allgemeine Probleme von Bedeutung.
Wie in Abschnitt 1.1 dargestellt, läßt sich die Lösung linearer Systeme darstellen als Linearkombination von Eigenlösungen der Bauart $v_i e^{\lambda_i(x-x_0)}$, mit v_i Eigenvektor von A zum Eigenwert λ_i.
Nach Lemma 1.6.14 lassen sich die numerischen Näherungen bezüglich der Eigenvektoren zerlegen, nur mit anderen Wachstumsfunktionen. Daher genügt es, statt allgemeiner linearer Systeme, die skalare Testgleichung zu betrachten.

Unerwünscht ist nun der Fall, wo ein Eigenwerte λ_i von f_y negativen Realteil besitzt, und die Lösung $y = F(t)$ also in der entsprechenden Richtung stabil ist, eine kleine Störung in Richtung dieser stabilen Komponente v_i mit $\text{Re}\ \lambda_i < 0$ bei der gewählten positiven Schrittweite aber numerisch verstärkt wird, d.h., $\lambda_i h \notin \mathcal{S}$, so daß h verkleinert werden muß, und zwar mehr als nötig ist, um die Approximation der empfindlichsten Komponente mit $\text{Re}\ \lambda_i$ maximal sicher zu stellen.

Wunsch: Falls alle Störungen der exakten Lösung abklingen $\text{Re}\ (\lambda) < 0$, so sollte die numerische Lösung auch bei großer Schrittweite $h > 0$ nicht anwachsen.

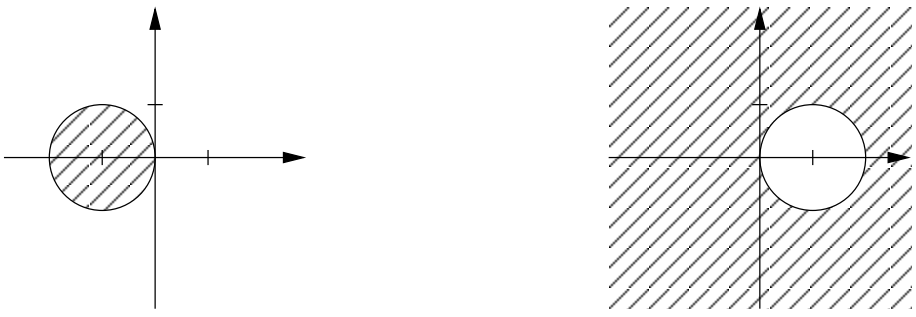
Definition 1.6.23 (A-stabil)
Enthält das Stabilitätsgebiet die linke komplexe Halbebene

$$\{z \in \mathbb{C} \mid \text{Re}\ (z) \leq 0\} \subseteq \mathcal{S}$$

so heißt das Verfahren **absolut stabil** *oder* **A-stabil**.

Beispiel 1.6.24 Das implizite Euler-Verfahren ist A-stabil.

$$R(z) = \frac{1}{1-z} \Longrightarrow |R(z)| \leq 1 \Leftrightarrow |1-z| = |z-1| \geq 1$$



$|z-(-1)| \leq 1$ $|z-1| \geq 1$

Abbildung 1.12: Stabilitätsgebiete des expliziten und impliziten Eulerverfahrens

Beispiel 1.6.25 (Die implizite Trapez-Regel ist A-stabil)

$$\begin{aligned}\eta(x+h;h) &= \eta(x;h) + \frac{h}{2}f(x,\eta(x;h)) + \frac{h}{2}f(x,\eta(x+h;h)) \\ &= \eta(x;h) + \frac{h}{2}\lambda\eta(x;h) + \frac{h}{2}\lambda\eta(x+h;h) \\ &\Rightarrow \\ \eta(x+h;h) &= \frac{1+h\lambda/2}{1-h\lambda/2}\eta(x;h)\end{aligned}$$

Wir erhalten

$$R(z) = \frac{1+z/2}{1-z/2} = \frac{2+z}{2-z} = -\frac{z-(-2)}{z-2} \ .$$

$|R(z)| = 1 \Leftrightarrow |z-(-2)| = |z-2|$ und wir erhalten exakt die linke komplexe Halbebene als Stabilitätsgebiet. Das ist besonders vorteilhaft bei Problemen mit oszillierenden Lösungsanteilen. Oszillationen mit $\mathrm{Re}\,(\lambda) = 0$ werden weder gedämpft, noch verstärkt.

Aufgabe 1.6.26 Man zeige: Ein explizites Runge-Kutta-Verfahren ist nicht A-stabil. (Vergleiche Lemma 1.6.16

Insbesondere bei stabilen Systemen ($l \leq 0$) wird die Bedingung daß Fehleranteile stabiler Komponenten, die hinreichend abgeklungen sind ($< TOL$), nicht mehr anwachsen ($|R(h\lambda_i)| \leq 1$) neben den beiden natürlichen Bedingungen (1.6.3) und (1.6.4) immer wichtiger.
Durch die Bedingung $|R(z)| < 1$ begrenzt nun oft der betragsgrößte Eigenwert von f_y mit negativem Realteil die Schrittweite. Man muß h so klein wählen daß $h\lambda_i \in \mathcal{S}$. Dies kann bei vielen Verfahren dazu führen, daß λ_i mit $\mathrm{Re}\ \lambda_i$ sehr klein die Schrittweite beschränkt. Probleme bei denen dies relevant werden kann heißen steif. Dies hängt auch von TOL und p und $x - x_0$ ab, aber auch vom Startwert, bzw. davon ob die stabilen Lösungsanteile schon abgeklungen sind oder von $C(x)$ ab. Löst man eine Differentialgleichung numerisch in k Schritten, mit Schrittweite $h_i = t_i - t_{i-1}$ und macht man dabei in jedem Schritt einen lokalen Fehler e_i, so erhält man mit dem Fundamentallemma am Ende einen Gesamtfehler

$$E_k \leq \sum_{i=1}^{k} e_i \exp\left(\int_{t_i}^{t_k} l(\tau)\mathrm{d}\tau\right) \leq \max_i e_i e^{\max_t l(t)(t_k - t_0)}$$

Soll dieser Fehler klein sein, erzwingt dies bei großem $l(t)$ kleine lokale Fehler bzw. kleine Schrittweiten.

Ist die Lösung y von $y' = f(t,y)\,,\ y(t_0) = y_0$ lokal durch einen Streckenzug nicht gut darstellbar, erzwingt bereits die Approximation von y kleine Schrittweiten (transiente Phase).
Ist $l(t) < l_0$ und y glatt, so spielen lokale Fehler eine beschränkte Rolle und es genügt

$$\sum_{i=1}^{k} e_i e^{l_0(t_k - t_0)} < E_k$$

zu erfüllen.
Es kann aber sein, daß sein, daß $L := \|f_y(t,y)\| >> |l_0|$, und daß dies zwar durch exponentiell stabile Anteile der Lösung verursacht wird, dies beim numerischen Lösen jedoch besondere Schwierigkeiten erzeugt, und gegebenfalls kleinere Schrittweiten erzwingt, als aufgrund einer leichten Instabilität erforderlich. Differentialgleichungen mit
$\|f_y(t,y)\| \gg \mu[f_y(t,y)]$ bezeichnet man daher als steif. Sie erfordern in der Regel implizite Verfahren.

Definition 1.6.27 *Ein Anfangswertproblem*

$$y'(x) = f(x, y(x)) \quad ; \quad y(a) = y_a \quad ; \quad x \in [a,b]$$

heißt steif, falls eine einseitige Lipschitzkonstante $l(x)$ von $f(x,y)$ existiert mit

$$\max\{l(x), 0\}|l(x)| << \|f_y(x,y)\|.$$

Ein Anfangswertproblem kann also auch in Teilbereichen des relevanten Intervalles steif sein.

Bemerkung 1.6.28 Die Definition 1.6.27 ist nicht einheitlich in der Literatur. Andere gebräuchliche Definitionen sind:

- Ein AWP ist steif $:\Longleftrightarrow\ \exists\lambda_i :\ \text{Re}\ (\lambda_i)[b-a] << 0$
- Ein AWP ist steif $:\Longleftrightarrow\ |\lambda_{\max}|/|\lambda_{\min}| << 1$
- Ein AWP ist steif $:\Longleftrightarrow$ bezüglich mindestens einer Störungsrichtung extrem gut konditioniert.
- Ein AWP ist steif $:\Longleftrightarrow$ Die Schrittweite ist nicht durch die Genauigkeit beschränkt.

- Ein AWP ist steif $:\Longleftrightarrow$ implizite Methoden sind effizienter als explizite.

Bei steifen Problemen lohnen sich implizite Verfahren, obwohl man in jedem Schritt eine implizite Gleichung durch ein Iterationsverfahren lösen muß.

Beispiel 1.6.29 (Das implizite Eulerverfahren)

$$\eta(x+h;h) = \eta(x;h) + hf(x, \eta(x+h;h))$$

Startschätzung

$$\eta(x+h;h)^{(0)} = \eta(x;h)$$

Fixpunktiteration

$$\eta(x+h;h)^{(k+1)} = \eta(x;h) + hf(x, \eta(x+h;h)^{(k)})$$

Die Iterationsfunktion ist kontraktiv für h klein genug.
Alternativ: Newtonverfahren zur Lösung der Gleichung

$$\eta(x+h;h) - \eta(x;h) - hf(x, \eta(x+h;h)) = F(\eta(x+h;h)) \stackrel{!}{=} 0$$

erfordert die Jacobimatrix

$$\mathcal{J}_F = I - hf_y \ ,$$

die für h klein genug nicht singulär ist. Das Verfahren konvergiert also lokal quadratisch und für h klein genug ist auch die Startschätzung im lokalen Konvergenzbereich.

1.6.2 Berechnung der Stabilitätsfunktion

Bei einem gegebenen Runge-Kutta-Verfahren läßt sich nun die Stabilitätsfunktion sehr leicht angeben.

Satz 1.6.30 Stabilitätsfunktion des Runge–Kutta–Verfahrens.
Die Stabilitätsfunktion eines impliziten Runge–Kutta*–Verfahrens mit Tableau*

$$\begin{array}{c|c} * & A \\ \hline & b^T \end{array}$$

ist gegeben durch

$$R(z) = \frac{det(I - zA + z\vec{1}b^T)}{det(I - zA)}.$$

Beweis: Zur Übung.
Hinweis: Man schreiben dazu die Verfahrensvorschrift des RUNGE–KUTTA–Verfahrens als lineares Gleichungssystem und verwende die CRAMERsche Regel.

■

Aufgabe 1.6.31 Verifizieren Sie, daß die Stabilitätsfunktion des klassischen Runge-Kutta-Verfahrens 4. Ordnung durch das Polynom

$$R(z) = 1 + z + \frac{1}{2}z^2 + \frac{1}{6}z^3 + \frac{1}{24}z^4$$

gegeben ist.

Satz 1.6.32 *Die Gauß-Verfahren sind A-stabil*

Satz 1.6.33 *Es existiert kein A-stabiles ERK-Verfahren.*

Beweis: Zur Übung.
Hinweis: Man zeige: Für ein s-stufiges ERK-Verfahren ist die Stabilitätsfunktion ein Polynom vom Grade s. $\Rightarrow \lim_{|z|\to\infty} |R(z)| = \infty$. ■
Bei A-stabilen Verfahren muß also stets ein implizites Gleichungssystem iterativ gelöst werden.
Als Startnäherung verwendet man ein explizites Verfahren, oder einfach die Approximation an der letzten Stelle

$$\eta^{(0)}(x+h) := \eta(x)$$

beziehungsweise

$$k_i(x_l) := f(x_l + c_i h_l, y_l + h_l \sum_{j=1}^{s} \alpha_{i,j} k_j(x_l)) \approx k_i(x_{l-1}) =: k_i^{(0)}(x_l)$$

Aus der impliziten Gleichung kann man eine Fixpunktiteration ableiten.

$$k_i^{(m+1)}(x_l) := f(x_l + c_i h_l, y_l + h_l \sum_{j=1}^{s} \alpha_{i,j} k_j^{(m)}(x_l))$$

mit

$$\left| \frac{\partial}{\partial k_j} \right| = |h_l \alpha_{i,j} f_y| < 1$$

falls h klein genug. Allerdingg erreicht man nur lineare Konvergenz.
Verwendet man ein Newtonähnliches Verfahren, so benötigt man die Jacobimatrix f_y und muß in jeder Iteration lineare Gleichungssysteme lösen. Man hat lokal quadratische Konvergenz und bei sehr guten Startwerten auch beim vereinfachten Newtonverfahren noch nahezu quadratische Konvergenz. Dies erreicht man wieder durch h hinreichend klein. Man benötigt dann nur einmal die Jacobimatrix f_y und wenige Iterationsschritte. Dennoch ist dies extrem aufwendig, weshalb man wann immer möglich explizite Verfahren vorzieht.

Padé Approximationen:
Hat man ein numerisches Verfahren konstruiert, so kann man die Stabilitätsfunktion aufstellen und auf ihre Eigenschaften hin untersuchen. Man kann aber auch umgekehrt vorgehen, und eine rationale Funktion $R_{jk}(z) = P_k(z)/Q_j(z)$ suchen, die die gewünschten Stabilitätseigenschaften hat. Interessante Kandidaten sind etwa die rationalen sogenannten **Padé Approximationen** der e-Funktion, denn für die skalare Testgleichung gilt $\eta(x+h) = \eta(x)e^{\lambda h}$. Allgemein gilt für relativ beliebige Funktionen $g(z)$:

$$\begin{aligned} \forall j,k \quad R_{jk}(z) = \frac{P_k(z)}{Q_j(z)} : \; R_{jk}(z) - g(z) &= \mathcal{O}|z|^{j+k+1} \\ \Rightarrow P_k(z) - Q_j(z)g(z) &= \mathcal{O}|z|^{j+k+1} \; . \end{aligned}$$

Also für $g(z) = e^z$

$$P_k(z) - Q_j(z)e^z = \mathcal{O}|z|^{j+k+1} \tag{1.6.9}$$

R_{jk} ist dann eindeutig bestimmt, und damit nach Normierung von Q_j auch P_k und Q_j. Man erhält sie durch den Ansatz (1.6.9) und Koeffizientenvergleich, wenn man für e^z bzw. allgemein für $g(z)$ die Taylorreihe einsetzt.

j/k	0	1	2	3
0	$\frac{1}{1}$	$\frac{1+z}{1}$	$\frac{1+z+z^2/2}{1}$	$\dots$
1	$\frac{1}{1-z}$	$\frac{1+z/2}{1-z/2}$	$\frac{1+z/3}{1-z/3}$	$\dots$
2	$\frac{1}{1-z+z^2/2}$	$\frac{1+z/3+z^2/12}{1-2z/3+z^2/6}$	$\frac{1+z/2+z^2/12}{1-z/2+z^2/12}$	$\dots$

$(j,k) = (0,1)$ entspricht dem expliziten Euler-Verfahren
$(j,k) = (1,0)$ entspricht dem impliziten Euler-Verfahren
$(j,k) = (1,1)$ entspricht der impliziten Trapezregel
Für $k < j$ gilt $\lim_{z\to-\infty} R_{jk} = 0$ wie e^{-z}.

Für $k = j$ gilt $\lim_{z\to-\infty} R_{jk}$ beschränkt. Damit sind die Verfahren mit Einschränkungen für steife Probleme einsetzbar.
Für $k > j$ gilt $|\lim_{z\to-\infty} R_{jk}| = \infty$, die Verfahren sind also für steife Probleme unbrauchbar
Durch Analyse der rationalen Funktionen kann man auch allgemeine Aussagen über mögliche Stabilitätseigenschaften von Einschrittverfahren treffen. Dazu plotte man die Grenzlinie des Stabilitätsgebietes

$$\partial S = \{z \in \mathbb{C} \,|\, |R(z)| = 1\} \ .$$

Bei Mehrschrittverfahren betrachtet man an Stelle der Stabilitätsfunktion die assoziierten Polynome. Ein lineares Mehrschrittverfahren mit Schrittweite h angewendet auf die Testgleichung ist eine lineare Abbildung. Ist μ ein Eigenwert und $(\eta_j, \ldots, \eta_{j+k})$ Eigenvektor des Iterationsverfahrens, d.h., $\eta_{j+i} = \mu^i \eta_j$, dann gilt

$$\Psi(\mu)\bar{\eta}_j = h\lambda\chi(\mu)\bar{\eta}_j$$

das heißt, es gilt die Beziehung

$$z := h\lambda = \frac{\Psi(\mu)}{\chi(\mu)} \ .$$

P_k und Q_j entsprechen dann den Funktionen Ψ und χ, d.h., zu vorgegebener Padé Approximation kann man sofort ein zugehöriges Mehrschrittverfahren angeben (eindeutig, falls Ψ und χ teilerfremd).
Kritisch wird die Sache, falls $h\lambda$ so gewählt sind, daß $|\mu| > 1$. Man plotte daher die Kurve

$$\frac{\Psi(e^{i\alpha})}{\chi(e^{i\alpha})} \quad ; \quad \alpha \in [0, 2\pi]$$

und man erhält ∂S. Schließlich teste man $|R(z)|$ in jeder Zusammenhangskomponente von $\mathbb{C}/\partial S$. Auf diese Weise zeigt man etwa:

Satz 1.6.34 (Dahlquist Barriere, (1963))
Es gibt kein A-stabiles Mehrschrittverfahren der Ordnung $p > 2$.

Trotzdem gehören Mehrschrittverfahren, insbesondere BDF-Verfahren, zu den am häufigsten verwendeten Verfahren für steife Differentialgleichungen. Dabei ist eine Ordnungssteuerung implementiert, die sehr geringe Ordnung (1 oder 2) wählt, wenn A-Stabilität gebraucht wird und ansonsten bis zur Ordnung 6 hoch schaltet.
BDF1 (impliziter Euler) und BDF2 sind A-stabil. BDF3-6 sind nur $A(\alpha)$-stabil mit abnehmendem α.

Kapitel 2

Randwertprobleme bei gewöhlichen Differentialgleichungen

2.1 Einleitung

In vielen Anwendungen ist der Zustand eines Systems zu Beginn gar nicht vollständig bestimmt. Vielmehr ist der Zustand, oder Teile des Systems zu Beginn so einzustellen, daß ein gewünschtes Ergebnis erzielt wird.
Parameter, die ja als Komponenten des Zustandvektors y mit trivialen Differentialgleichungen formuliert werden können, sollen z.B. so eingestellt werden, daß das System am Ende des Intervall einen bestimmten Zustand einnimmt. Allgemein betrachtet man also ein **nichtlineares 2-Punkt-Randwertprobleme:** der Bauart

$$y'(x) = f(x, y(x)) \quad ; \quad a \le x \le b \tag{2.1.1}$$

$$r(y(a), y(b)) = 0 \quad ; \quad r : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n \tag{2.1.2}$$

Beispiel 2.1.1 Z.B., könnte verlangt sein, das die erste schwingende Komponente y_1 von y mit einer bestimmten Periode $b-a$ schwingt. Diese Bedingung könnte man in der Form

$$y_1(b) - y_1(a) = 0$$

formulieren.

Häufig trifft man auch auf einfachere Spezialfälle:

Randwertprobleme höherer Ordnung:
Eine Differentialgleichung höherer Ordnung

$$y^{(m)}(x) = f(x, y(x), \ldots, y^{(m-1)}(x)) \tag{2.1.3}$$

kann in ein System erster Ordnung transformiert werden. Dabei verliert man allerdings manchmal wesentliche Information, so daß auch oft (2.1.3) direkt behandelt wird.

Lineare Randwertprobleme:
Es besteht aus einer linearen Differentialgleichung und linearen Randbedingungen

$$\begin{aligned} y'(x) &= A(x)y(x) + q(x) \quad ; \quad a \le x \le b \\ r(y(a), y(b)) &= B_a y(a) + B_b y(b) - c = 0 \quad ; \quad B_a, B_b : \mathbb{R}^{n\times n},\ c \in \mathbb{R}^n \end{aligned}$$

Separierte Randbedingungen:
Oft sind die Randbedingungen nicht wie in Beispiel 2.1.1 verknüpft, sondern einzelne Bedingungen betreffen nur Zustände am Anfang, andere nur am Endpunkt.

$$\begin{aligned} r_1(y(a)) &= 0 \quad ; \quad r_1 : \mathbb{R}^n \to \mathbb{R}^p \\ r_2(y(b)) &= 0 \quad ; \quad r_2 : \mathbb{R}^n \to \mathbb{R}^q \quad ; \quad p + q = n \end{aligned}$$

Durch Einführung von neuen Variablen w für y_b erhält man ein Problem mit separierten Randwerten:

$$\begin{aligned} y'(x) &= f(x, y(x)) \\ w'(x) &= 0 \end{aligned} \quad \text{mit} \quad \begin{aligned} r(y(a), w(a)) &= 0 \\ w(b) - y(b) &= 0 \end{aligned} .$$

Dieses System hat die doppelte Dimension, weshalb man die Transformation in der Praxis selten durchführt. Theoretische Aussagen bei separierten Randbedingungen lassen sich aber entsprechend übertragen.

Randwertprobleme mit freiem Rand:
Manchmal ist die rechte (oder linke) Grenze des Intervalls nicht festgelegt.

$$\begin{aligned} y'(x) &= f(x, y(x)) \quad ; \quad a \le x \le b \quad ; \quad b \text{ frei} \\ r(y(a), y(b)) &= 0 \quad ; \quad r \in \mathbb{R}^{n+1} \end{aligned}$$

b soll dabei geeignet bestimmt werden.

Beispiel 2.1.2 Will man z.B. eine periodische Lösung eines Systems ermitteln

$$\begin{aligned} y'(x) &= f(x, y(x)) \\ r(y(a), y(b)) &= y(a) - y(b) = 0 \ . \end{aligned}$$

mit unbekannter Periode,
so macht man durch $x := a + t(b-a)$ eine Transformation auf das Intervall $[0;1]$, und führt eine Variable $y_{n+1} = b - a$ ein. Wegen $\frac{\mathrm{d}x}{\mathrm{d}t} = y_{n+1}$ müssen die Differentialgleichungen für die Zustände $z(t) := y(x(t))$ dabei modifiziert werden.

$$y'(x) = \frac{\mathrm{d}y}{\mathrm{d}x} = \frac{\mathrm{d}y}{\mathrm{d}t}\frac{\mathrm{d}t}{\mathrm{d}x} = \frac{\mathrm{d}\dot{z}(t)}{\mathrm{d}z_{n+1}}$$

Das neue normalisierte Randwertproblem lautet also

$$\begin{aligned} \dot{z}(t) &= z_{n+1}(t) f(a + t z_{n+1}(t), z(t)) \\ \dot{z}_{n+1}(t) &= 0 \\ r(z(0), z(1)) &= 0 \in \mathbb{R}^{n+1} \end{aligned}$$

Die z_{n+1} Komponente der Lösung enthält die Periode. Die Randbedingung $r = 0$ ist dabei nicht eindeutig, da jeder Zustand eines periodischen Orbitsals $z(0), z(1)$ die Bedingungen erfüllt. Zur Eindeutigkeit kann man eine Bedingungvon r weglassen und durch eine einseitige Randbedingunug ersetzen. Z.B. $z_i(0) = \alpha$ statt $z_i(0) = z_i(1)$.

Die Methode der Transformation auf ein Einheitsinterval benutzt man oft standardmäßig. Das Problem wird dabei autonom, die Intervalllänge fest.[1] Lineare Differentialgleichungen mit nichtlinearer Inhomogenität werden dabei jedoch nichtlinear. In diesem Fall keine Transformation.

Mehrpunkt-Randwertprobleme:
Manchmal sind nicht nur am Anfang und Ende Bedingungen vorgegeben, sondern auch im Inneren.

$$\begin{aligned} y'(x) &= f(x, y(x)) \quad ; \quad y \in \mathbb{R}^n \quad ; \quad a \le x \le b \\ 0 &= r(y(x_0), y(x_1), \ldots, y(x_{s-1}), y(x_s)) \end{aligned}$$

mit

$$a =: x_0 < x_1 < \cdots < x_{s-1} < x_s := b$$

[1] Dieses System ist meist nicht eindeutig lösbar. Im Falle periodischer Lösungen sind die Bedingungen $y(a) = y(b)$ teilweise redundant. Die Lösung dann daher nicht eindeutig.

Z.B., muß bei einer Reflexion das reflektierte Teilchen am sich zum Zeitpunkt der Reflexion an der reflektierenden Wand befinden.
An solchen inneren Punkte ändert sich auch oft die Dynamik des Systems. Beim Eintauchen in ein anderes Medium ändert sich z.B. der Widerstand. Wir haben es daher allgemein mit einer abschnittsweise definierten rechten Seite zu tun

$$y' = f(x,y) = \begin{cases} f_1(x,y) & \text{für} \quad x_0 \le x \le x_1 \\ f_2(x,y) & \text{für} \quad x_1 \le x \le x_2 \\ \vdots & \\ f_s(x,y) & \text{für} \quad x_{s-1} \le x \le x_s \end{cases}$$

Dabei sind die sogenannten Schaltpunkte s_i entweder direkt vorgegeben, oder es werden weitere Randbedingungen gestellt. Diese Randbedingungen können sich auch auf Zustände an inneren Punkten beziehen. Die Zustandsgrößen selbst sind oft an den Schaltpunkten sogar unstetig. Z.B. ändert sich die Masse einer Rakete an dem Zeitpunkt der Stufentrennung sprunghaft. Die Unstetigkeit hängt dabei aber meist vom Zustand vor dem Sprung statt ab. Zur Differentialgleichung kommen dann noch die Bedingungen

$$\begin{aligned} y(x_i^+) &= \sigma_i(x_i, y(x_i^-)) \quad ; \quad i = 1, \ldots, s-1 \\ 0 &= r(y(x_0), y(x_1^-), \ldots, y(x_{s-1}^-), y(x_s)) \in \underbrace{\mathbb{R}^n \times \mathbb{R}^n \times \cdots \times \mathbb{R}^n}_{s+1} \to \mathbb{R}^{n+s-1} \; . \end{aligned}$$

Auch dieses Problem kann auf Standardform gebracht werden. Dazu definiert man in jedem Teilintervall $[x_{i-1}; x_i]$ eine Variable

$$z_i(\tau) := y(x_{i-1} + \tau(x_i - x_{i-1}))$$

und Variablen $\Delta_1 \ldots, \Delta_s$ mit $\Delta_i := x_i - x_{i-1}$. Dann erhält man das normalisierte Randwertproblem:

$$\begin{aligned} \dot{z}_i(\tau) &= \Delta_i f(a + \sum_{j=1}^{i-1} \Delta_j + \Delta_i \tau, z_i) \quad ; \quad i = 1, \ldots, s \\ \dot{\Delta}_i &= 0 \quad ; \quad i = 1, \ldots, s \\ 0 &= r(z_1(0), z_1(0), \ldots, z_s(0), z_s(1)) \\ z_i(0) &= \sigma_i(a + \sum_{j=1}^{i-1} \Delta_j, z_{i-1}(1)) \quad ; \quad i = 1, \ldots, s-1 \; . \end{aligned}$$

Dieser Weg wird allerdings in der Praxis nicht bestritten, weil dabei unterschiedliche Intervalle numerisch mit gleicher Schrittweite gelöst werden müßten, was sehr ineffizient wäre. Theoretische Überlegungen lassen sich durch diese Äquivalenz aber vom Standardfall übertragen.

2.2 Existenz und Eindeutigkeit

Leider ist bei Randwertproblemen auch bei gutartiger rechter Seite weder Existenz noch Eindeutigkeit garantiert.

Beispiel 2.2.1 Die Schwingung einer idealen Hook'schen Feder mit Federkonstante 1 genügt der Differentialgleichung

$$u''(x) = -u(x)$$

Die allgemeine Lösung lautet:

$$u(x) = c_1 \sin(x) + c_2 \cos(x)$$

Wir unterscheiden drei Fälle:
$u(0) = 0\,,\ u(\frac{\pi}{2}) = 1: \Rightarrow u(x) = \sin(x)$ (eindeutig).
$u(0) = 0\,,\ u(\pi) = 0: \Rightarrow u(x) = c_1 \sin(x)$ (unendlich viele Lösungen).
$u(0) = 0\,,\ u(\pi) = 1: \Rightarrow$ keine Lösung.
Mit $y_1 := u\,,\ y_2 := u'$ erhält man alternativ die Differentialgleichung erster Ordnung

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}' = \begin{bmatrix} y_2 \\ -cy_1 \end{bmatrix}$$

Als Randbedingungen sind dann nur Werte für y_1 am linken und rechten Rand vorgegeben und wir erhalten die gleiche Fallunterscheidung.

Ganz allgemein kann man zu jedem Randwertproblem

$$y' = f(x, y) \quad ; \quad r(y(a), y(b)) = 0 \tag{2.2.1}$$

ein Anfangswertproblem

$$y' = f(x, y) \quad ; \quad y(a) = s$$

mit noch unbekanntem Parameter s angeben. Liefern alle diese Anfangswerte eindeutige Lösungen $w(x, s) := y(x; a, s)$, so existiert eine eindeutige Funktion

$$F(s) := r(s, w(b, s))\ . \tag{2.2.2}$$

Das Randwertproblem ist eindeutig lösbar, falls $F(s) = 0$ genau eine Lösung besitzt.

Diese Bedingung ist allerdings schwer nachprüfbar,[2]
Die Existenz und die lokale Eindeutigkeit einer Lösung läßt sich dagegen im nachhinein oft bestimmen, also nachdem man z.B. die Lösung numerisch berechnet hat.

Definition 2.2.2 (Lokale Eindeutigkeit)
Eine Lösung $y(x)$ des RWPs(2.2.1) heißt **lokal eindeutig**, *wenn es einen Streifen um $y(x)$ gibt, in dem $y(x)$ die einzige Lösung des RWPs ist, d.h.,*

$$\exists \delta > 0 \, : \forall u(x) \not\equiv y(x) \, : [\|u - y\|_\infty > \delta \vee u \text{ ist keine Lösung von (2.2.1)}]$$

Diese Eigenschaft läßt sich zumindest numerisch überprüfen.
Sei $u_i(x)$ eine Funktionenfolge mit $\|u_i(x) - y\|_\infty =: \delta_i \to 0$,
$0 \neq u_i(a) - s^* = \alpha_i \to 0$
und $u_i(b) - w(b, s^*) = \beta_i \to 0$. Dann gilt

$$\frac{F(s^* + \alpha_i)}{\|\alpha_i\|} = \left[\underbrace{F(s^*)}_{=0} + F_s(s^*)\alpha_i + \mathcal{O}(\alpha_i^2) \right] / \|\alpha_i\|$$

$$\frac{F(s^* + \alpha_i)}{\|\alpha_i\|} = 0 \Longrightarrow \qquad F(s^*)\frac{\alpha_i}{\|\alpha_i\|} \to 0$$

$\frac{\alpha_i}{\|\alpha_i\|}$ liegt auf der n-dimensionalen Einheitskugel S_n. S_n ist kompakt, daher hat $\frac{\alpha_i}{\|\alpha_i\|}$ einen Häufungspunkt α mit $\|\alpha\| = 1$. Ist F stetig, so gilt dann $F_s(s^*)\alpha = 0$. Es gilt aber

$$F_s(s^*)\alpha = \frac{\partial r(s^*, w(b, s^*))}{\partial y(a)}\alpha + \frac{\partial r(s^*, w(b, s^*))}{\partial y(b)}\frac{\partial w(b, s^*)}{\partial s}\alpha$$

Ist also die Lösung nicht lokal eindeutig, so ist F_s singulär und es gibt eine Richtung α den Anfangswert s^* zu variieren mit

$$\frac{\partial r(s^*, w(b, s^*))}{\partial y(a)}\alpha + \frac{\partial r(s^*, w(b, s^*))}{\partial y(b)} \underbrace{\frac{\partial w(b, s^*)}{\partial s}\alpha}_{=:\beta} = 0 \, .$$

[2]Sie liefert aber eine Motivation für numerische Verfahren, den sogenannten Anfangswertmethoden oder Schießverfahren.

Dabei nutzen wir aus, daß $w(x,s)$ differenzierbar von s abhängt, wobei $z(x) := \frac{\partial w(x;s)}{\partial s}$ der Variationsdifferentialgleichung

$$z' = \underbrace{\frac{\partial f(x,y(x))}{\partial y}}_{\text{Jacobimatrix}} z \tag{2.2.3}$$

genügt. Für $z(a) = \alpha$ ergibt sich dann $z(b) = \beta$ und es muß gelten:

$$\underbrace{\frac{\partial r(y(a),y(b))}{\partial y(a)}}_{=:B_a} z(a) + \underbrace{\frac{\partial r(y(a),y(b))}{\partial y(b)}}_{=:B_b} z(b) = 0 \ . \tag{2.2.4}$$

Definition 2.2.3 (Isolierte Lösung) *Eine Lösung $y(x)$ des RWPs(2.2.1) heißt* **isoliert**, *wenn das Variationsproblem (2.2.3), (2.2.4) eine eindeutige Lösung $z(x) \equiv 0$ besitzt.*

Satz 2.2.4 *Sei $f \in C^2$, so daß (2.2.3) wohl definiert ist. Ist dann $y(x)$ eine isolierte Lösung von (2.2.1), dann ist es auch lokal eindeutig. In diesem Fall ist s^* eine einfache Nullstelle von F, d.h F_s ist nicht singulär*

Die Umkehrung gilt nicht, d.h., auch wenn F_s singulär ist, kann die Lösung lokal eindeutig sein.
Wir müssen also die Eindeutigkeit des Variationsproblems zeigen und betrachten jetzt noch geringfügig allgemeiner das inhomogenelineare Randwertproblem:

$$y' = A(x)y + q(x) \tag{2.2.5}$$

$$B_a y(a) + B_b y(b) = c \ . \tag{2.2.6}$$

Für $q(x) \equiv 0$ erhält man eine **Fundamentallösung** $Y(x;t) \in \mathbb{R}^{n,n}$ von (2.2.5), also von

$$y' = Ay$$

durch Lösung der Matrixdifferentialgleichung[3]

$$\frac{\mathrm{d}}{\mathrm{d}x} Y(x;t) =: Y' = A(x)Y \tag{2.2.7}$$

$$Y(t,t) = I \tag{2.2.8}$$

[3]Diese Differentialgleichung läßt sich spaltenweise lesen, d.h., jede Spalte von Y genügt der Differentialgleichung $y' = Ay$. Als Startwerte an der Stelle $x = t$ sind alle Einheitsrichtungen gewählt. Wegen des Superpositionsprinzips für lineare homogene Differentialgleichungen erhält man daraus die Lösung des Anfangswertproblems zu beliebigen Startwerten $y(t) = y_t$ aus $y(x) = Y(x,t)y_t$.

Für das inhomogene Anfangswertproblem

$$y' = A(x)y + q(x) \tag{2.2.9}$$
$$y(a) = s \ , \tag{2.2.10}$$

erhält man mit $Y(x) := Y(x,a)$ die Lösung

$$y(x) = Y(x)\left[s + \int_a^x Y^{-1}(t)q(t)\mathrm{d}t\right] \ . \tag{2.2.11}$$

Beweis: durch Nachrechnen (vgl Variation der Konstanten).

Lemma 2.2.5 (Existenz und Eindeutigkeit) *Sei A und q stetig auf $[a;b]$. Das lineare Randwertproblem (2.2.5), (2.2.6) hat genau dann eine eindeutige Lösung, falls gilt:*

$$Q := B_a\Phi(a) + B_b\Phi(b) \quad \textit{nicht singulär.} \tag{2.2.12}$$

In diesem Fall gilt

$$y(x) = \Phi(x)Q^{-1}\left[c - B_b\Phi(b)\int_a^b \Phi^{-1}(t)q(t)\mathrm{d}t\right] + \Phi(x)\int_a^x \Phi^{-1}(t)q(t)\mathrm{d}t \ . \tag{2.2.13}$$

Dabei ist $\Phi \in \mathbb{R}^{n,n}$ irgendeine Fundamentallösung[4] von (2.2.7), die jedoch nicht (2.2.8) erfüllen muß. Allerdings muß $\Phi(x)$ zu irgend einem Zeitpunkt $x = t$ nicht singulär[5] sein.

Beweis: Jede Lösung von (2.2.9), (2.2.10 läßt sich schreiben als

$$y(x) = \underbrace{\Phi(x)\Phi^{-1}(a)s}_{=:Y(x)} + y_p(x) \ , (**)$$

[4] Die Verallgemeinerung ist sinnvoll, weil manchmal bestimmte Fundamentalsysteme $\Phi \neq Y$ gegeben sind, oder sich leichter berechnen lassen. Später werden wir auch formal ein Fundamentalsystem $\bar{\Phi}$ benutzen.

[5] Dann ist $\Phi(x)$ für alle x nicht singulär, da $\Phi(x) = Y(x,t)\Phi(t)$.

wobei y_p eine partikuläre Lösung zu homogenem Anfangswert $y_p(a) = 0$ ist. Damit die Randbedingungen (2.2.6) erfüllt sind muß für s gelten:

$$\begin{aligned} B_a y(a) + B_b y(b) &\overset{(**)}{=} B_a[Y(a)s + \underbrace{y_p(a)}_{=0}] + B_b[Y(b)s + y_p(b)] = \\ &\overset{(2.2.12)}{=} Q\Phi^{-1}(a)s + B_b y_p(b) \\ &\overset{(2.2.11)}{=} Q\Phi^{-1}(a)s + B_b \underbrace{\Phi(b)\Phi^{-1}(a)}_{Y(b)} \int_a^b \underbrace{\Phi(a)\Phi^{-1}(t)}_{Y^{-1}(t)} q(t)\mathrm{d}t \\ &= c \end{aligned}$$

Dies ist ein lineares Gleichungssystem für s. Es hat genau dann eine eindeutige Lösung, falls Q nicht singulär. ∎

Kondition eines Randwertproblems: Damit hat man ein Kriterium in der Hand, um wenigstens im nachhinein auch die lokale Eindeutigkeit nicht linearer Randwertprobleme zu zeigen. Man bestimmt dabei numerisch die Lösung der Variationsdifferentialgleichung (2.2.7), (2.2.8), mit $A = f_y(x, \eta(x))$ und $\eta(x)$ die numerisch bestimmte Lösung des Randwertproblems. Sodann bestimmt man den Rang von $Q := B_a Y(a) + B_b Y(b)$.
A priori Aussagen über Existenz und Eindeutigkeit, d.h. ohne vorherige Berechnung der Lösung, sind im nicht linearen Fall kaum möglich, beziehungsweise die Bedingungen fast nie nachprüfbar.

Abhängigkeit der Lösung von den Daten:
Aus der Darstellung (2.2.11) für die partikuläre Lösung und der Beziehung

$$\Phi(x)\Phi^{-1}(a) = Y(x)$$

erhält man für die Lösung des Anfangswertproblems (2.2.9), (2.2.10) die Darstellung

$$y(x) = \Phi(x)\left[\Phi^{-1}(a)s + \int_a^x \Phi^{-1}(t)q(t)\mathrm{d}t\right], \tag{2.2.14}$$

wobei Φ wieder irgendeine Fundamentallösung ist. Mit der sogenannten **Green'schen Funktion**

$$G(x,t) := \begin{cases} Y(x,t) = \Phi(x)\Phi^{-1}(t) & \text{falls} \quad t \le x \\ 0 & \text{falls} \quad t > x \end{cases} \tag{2.2.15}$$

ergibt das

$$y(x) = \Phi(x)\Phi^{-1}(a)s + \int_a^b G(x,t)q(t)\mathrm{d}t \ . \tag{2.2.16}$$

$G(x,a)$ beschreibt also die Kondition bezüglich des Anfangswertes und $\int_a^b \|G(x,a)\|$ die Kondition bezüglich der Inhomogenität.
Ist $Q = B_a\Phi(a) + B_b\Phi(b)$ regulär, so ist

$$\bar{\Phi}(x) := \Phi(x)Q^{-1}$$

eine andere Fundamentallösung mit

$$\bar{Q} = B_a\bar{\Phi}(a) + B_b\bar{\Phi}(b) = I$$

Für das Randwertproblem muß dann das lineare Gleichungssystem

$$\underbrace{\bar{Q}}_{=I}\bar{\Phi}^{-1}(a)s = c - B_b\bar{\Phi}(b)\int_a^b \bar{\Phi}^{-1}(t)q(t)\mathrm{d}t$$

gelöst werden,

$$s = \bar{\Phi}(a)\left[c - B_b\bar{\Phi}(b)\int_a^b \bar{\Phi}^{-1}(t)q(t)\mathrm{d}t\right]$$

und man erhält die Lösung

$$\begin{aligned}
y(x) &= \bar{\Phi}(x)\bar{\Phi}^{-1}(a)\bar{\Phi}(a)\left[c - B_b\bar{\Phi}(b)\int_a^b \bar{\Phi}^{-1}(t)q(t)\mathrm{d}t\right] \\
&\quad + \int_a^x \bar{\Phi}(x)\bar{\Phi}^{-1}(t)q(t)\mathrm{d}t \\
&= \bar{\Phi}(x)c + \int_a^x \bar{\Phi}(x)\underbrace{[I - B_b\bar{\Phi}(b)]}_{=B_a\bar{\Phi}(a)}\bar{\Phi}^{-1}(t)q(t)\mathrm{d}t - \int_x^b \bar{\Phi}(x)B_b\bar{\Phi}(b)\bar{\Phi}^{-1}(t)q(t)\mathrm{d}t \\
&= \bar{\Phi}(x)c + \int_a^x \bar{\Phi}(x)B_a\bar{\Phi}(a)\bar{\Phi}^{-1}(t)q(t)\mathrm{d}t - \int_x^b \bar{\Phi}(x)B_b\bar{\Phi}(b)\bar{\Phi}^{-1}(t)q(t)\mathrm{d}t \\
&= \bar{\Phi}(x)c + \int_a^b \bar{G}(x,t)q(t)\mathrm{d}t
\end{aligned}$$

mit der **Green'schen Funktion**

$$\bar{G}(x,t) := \begin{cases} \bar{\Phi}(x)B_a\bar{\Phi}(a)\bar{\Phi}^{-1}(t) & t \le x \\ -\bar{\Phi}(x)B_b\bar{\Phi}(b)\bar{\Phi}^{-1}(t) & t > x \end{cases} \tag{2.2.17}$$

Daraus folgt

$$\|y\|_\infty \leq \kappa(\|c\|_\infty + \int_a^b \|q(t)\|_\infty \mathrm{d}t) \tag{2.2.18}$$

mit $\kappa := \max\{\|\bar{G}\|_\infty, \|\bar{\Phi}\|_\infty\}$. κ heißt **Konditionskonstante** oder **Stabilitätskonstante** des linearen Randwertproblems.
Zur Berechnung bestimme man zuerst für eine Fundamentallösung Φ die Werte $\Phi(a)$ und $\Phi(b)$, z.B. $Y(a) = I$ und $Y(b)$, damit Q und $\bar{\Phi}$. Danach berechne man erneut $\Phi(t)$ für $a \leq t \leq b$ und schätze damit κ ab.

2.3 Anfangswertmethoden

Gegeben sei das Randwertproblem

$$y' = f(t,y) \in \mathbb{R}^n \qquad ; \quad a \le t \le b \tag{2.3.1}$$
$$r(y(a), y(b)) = 0 \in \mathbb{R}^n \tag{2.3.2}$$

Man schätze $s := y(a)$ und löse das Anfangswertproblem

$$y' = f(t,y) \quad ; \quad y(a) = s \ . \tag{2.3.3}$$

Die erhaltene (numerische) Lösung $y(t;a,s)$ werte man am rechten Rand aus und berechne

$$F(s) := r(y(a;a,s), y(b;a,s)) = r(s, y(b;a,s)) \tag{2.3.4}$$

Gesucht ist dann eine Nullstelle der nichtlinearen Funktion F.
Die Funktion $F(s)$ in (2.2.2) liefert den Zugang zu einem numerischen Verfahren. Dabei wird im wesentlichen nur ein Verfahren zur Lösung eines nichtlinearen Gleichungssytems eingesetzt, welches aber auf die vorliegende Struktur hin etwas optimiert wird.
Notation:
In den Anwendungen ist die unabhängige Variable (bislang x) meist die Zeit und wird daher mit t (time) bezeichnet. Die unbekannte Funktion (bislang y) wird dagegen oftmals mit x bezeichnet. Um Verwechslungen zu vermeiden, wird hier daher $y(t)$ verwendet.

2.3.1 Einfachschießen

Beim klassischen **Einfachschießverfahren** wird ein modifiziertes inexaktes Newtonverfahren eingesetzt. Die Iterationsgleichung lautet dann

$$\begin{aligned} DF(s^{(i)})\Delta s^{(i)} &= -F(s^{(i)}) \\ s^{(i+1)} &= s^{(i)} + \lambda^{(i)}\Delta s^{(i)} \ , \end{aligned}$$

wobei die Jacobimatrix DF von F nicht exakt berechnet, sondern numerisch approximiert wird.
Eine Iterationon besteht dabei aus den Schritten:
0. Schätze Startwert $s^{(0)}$, $i := 0$, $\lambda^{(-1)} := 1$, berechne $F(s^{(0)})$

1. Berechne $DF(s^{(i)})$
2. Berechne $\Delta s^{(i)}$, $\lambda^{(i)} = \min\{1, 2\lambda^{(i-1)}\}$.
3. $s^{(i+1)} = s^{(i)} + \lambda^{(i)}\Delta s^{(i)}$
4. Berechne $F(s^{(i+1)})$
5. Falls $\|F(s^{(i+1)})\| < TOL$ Ende sonst weiter mit 6
6. Falls $\|F(s^{(i+1)})\| > \|F(s^{(i)})\|$, $\lambda^{(i)} = \lambda^{(i)}/2$ und gehe nach 3. sonst weiter mit 7.
7. $i := i + 1$ gehe nach 1.

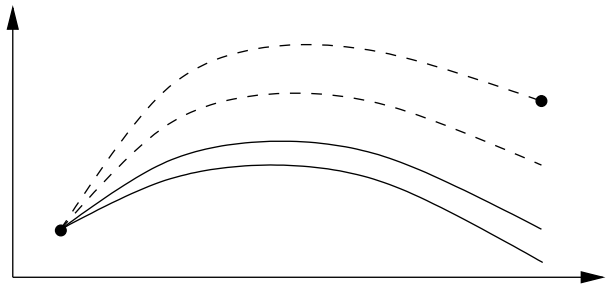


Abbildung 2.1: Einfachschießen

Abbildung 2.1 beschreibt die Situation einer Differentialgleichung zweiter Ordnung $w'' = f(t, w, w')$ mit $w(a) = \alpha$ und $w(b) = \beta$ vorgegeben. Gesucht ist noch die Anfangssteigung $w'(a) = s$. Durch zwei benachbarte Lösungen mit geschätztem s und leichter Variation (durchgezogene Kurven) läßt sich dann der Einfluß von s auf die Lösung des Anfangswertproblems schätzen (gestrichelte Kurven) und s für die nächste Iteration geeignet wählen.

Eine Auswertung von F erfordert dabei die Lösung eines Anfangswertproblems und die Auswertung der Randbedingungen. DF berechnet sich nach

$$DF = \frac{\partial r(y(a;a,s), y(b;a,s))}{\partial y_a} + \frac{\partial r(y(a;a,s), y(b;a,s))}{\partial y_b} \frac{\partial y(b;a,s)}{\partial s}. \quad (2.3.5)$$

$G(b;a,s) := \frac{\partial y(b;a,s)}{\partial s}$ ist dabei die Sensitivitätsmatrix des Anfangswertproblems (2.3.3) und kann durch numerische Differenzenapproximation berechnet werden. Die i-te Spalte von $G(t;a,s)$ enthält gerade

$$\frac{\partial y(t;a,s)}{\partial s_i} \doteq \frac{y(t;a,s+\delta e_i) - y(t;a,s)}{\delta} \quad (2.3.6)$$

Dabei ist δ ein kleiner Diskretisierungsparameter und e_i der i-te Einheitsvektor. Zur Auswertung von DF sind daher noch weitere n Anfangswertprobleme $y(t; a, s+\delta e_i)$, $i = 1.\ldots, n$, mit leicht veränderten Anfangswerten zu lösen. Daher wird oft versucht, die teure Auswertung durch Broydenkorrekturen zu ersetzen.
Ein großes Problem ist dabei die Wahl des Diskretisierungsparameters δ. Bei Differenzenapproximationen der Art (2.3.6) kommt es zu zwei Arten von Fehlern. Dem Approximationsfehler von der Ordnung $\mathcal{O}(\delta)$ und dem Rundungsfehler aufgrund der Auslöschung von der Ordnung $\mathcal{O}(\varepsilon/\delta)$. Je kleiner nämlich δ, umso mehr Stellen von $y(t; a, s + \delta e_i)$ und $y(t; a, s)$ stimmen überein. ε ist dabei nicht die Maschinengenauigkeit,[6]sondern die Genauigkeit der numerischen Lösungen der Anfangswertprobleme.
Als guter Kompromiß hat sich daher die Wahl $\delta = \sqrt{\varepsilon}$ herausgestellt, der dann zu einer Approximation von G mit Genauigkeit Ordnung $\mathcal{O}(\delta)$ führt. Stellt man hohe Genauigkeitsanforderungen δ an S, so ist $\varepsilon \approx \delta^2$ extrem klein zu wählen, und die numerische Lösung der Anfangswertprobleme wird sehr aufwendig. Es gibt aber Methoden diese aufwendige Sensitivitätsanalyse wesentlich effizienter durchzuführen (siehe Abschnitt **??**).

Bemerkung 2.3.1 Im Falle eines linearen Randwertproblems ist (2.3.4) ein lineares Gleichungssystem und das Newtonverfahren konvergiert in einem Schritt.

Bemerkung 2.3.2 Bei nichtlinearen Randwertproblemen existiert zu einem falschen Schätzwertvektor s unter Umständen keine Lösung des Anfangswertproblems (2.3.3), obwohl das Randwertproblem eine lokal eindeutige und gut konditionierte Lösung besitzt. (Bewegliche, von s abhängige Singularität)

Beispiel 2.3.3 (Bewegliche Singularität) Wir betrachten die Differentialgleichung

$$\dot{y} = y^2$$

[6]Bei der Differenzenapproximation einfacher Funktionen ohne Auslöschung im Funktionsrumpf, geht man in der Regel davon aus, daß ein Unterprogrammaufruf das Resultat bis auf etwa Maschinengenauigkeit liefert. Bei komplizierteren Funktionen, und Funktionen die implizite Gleichungen enthalten die iterativ bis zu einer vorgebbaren Genauigkeit gelöst werden, ist dies nicht der Fall.

mit der allgemeinen Lösung

$$y = \frac{1}{c - t}$$

mit c beliebig. Fordert man die Randbedingung $y(10) - y(0) = \frac{1}{20}$, so ergibt dies $a = 20$, $y(t) = \frac{1}{20-t} \Rightarrow s = y(0) = \frac{1}{20} = 0.05$.
Startet man mit der Approximation $s^{(0)} = y^{(0)}(0) = 0.2 \Rightarrow c = 5 \Rightarrow y = \frac{1}{5-t}$ so hat das Anfangswertproblem eine Singularität bei $t = 5 \in [0; 10]$. und die Funktion F kann gar nicht erst ausgewertet werden (Overflow). Erst für $s^{(0)} < 0.1$ konvergiert das Verfahren.

Wegen des in Beispiel 2.3.3 beschriebenen Effektes ist eine gute Startschätzung von s sehr wichtig. Selbst wenn die Anfangswertprobleme lösbar sind, ist auch die Konvergenz des Newtonverfahrens ja nur lokal garantiert und quadratisch. In der Praxis reicht Einfachschießen daher bei hochnichtlinearen Problemen nicht aus, oder ist sehr ineffizient. Eine Abhilfe bietet dann das Mehrfachschießverfahren.

2.3.2 Mehrfachschießen

Wir betrachten wieder das Randwertproblem (2.3.1), (2.3.2).
Nach Lemma 1.1.7 gilt

$$\|y(t, s_1) - y(t, s_2)\| \leq \|s_1 - s_2\| e^{L|t-a|}$$

dabei ist L in der Regel nur innerhalb eines Streifens S um die Lösung beschränkt, und die Abschätzung gilt nur solange $(t, y(t, s_1)), (t, y(t, s_2)) \in S$. Für $|t-a|$ groß, ist dies aber oft nicht mehr erfüllt. Für $|t-a|$ klein genug, bleibt aber $(t, s_1), (t, s_2) \in S$, wenn dies für (a, s_1) und (a, s_2) galt. Die Lösungen existieren dann und weichen kaum voneinander ab. daher versucht man die Intervalllänge der Anfangswertprobleme zu reduzieren.
Wir zerlegen daher das betrachtete Intervall in m Teilintervalle

$$a = t_1 \leq t_2 \leq \cdots \leq t_{m+1} = b \ .$$

Hat man Approximationen η_k der Lösung $y(.)$ an den Stellen $t = t_k$, so kann man sie überprüfen, indem man die m Anfangswertprobleme

$$y'(t) = f(t, y) \ , \quad y(t_k) = \eta_k \ , \quad t_k \leq t \leq t_{k+1} \ , \quad k = 1, \ldots, m$$

löst. Sind die Approximationen exakt, so ergibt sich

$$d_k := y(t_{k+1}, t_k, \eta_k) - \eta_{k+1} = 0 \quad \text{für} \quad k = 1, \cdots, m \ . \tag{2.3.7}$$

und

$$r(\eta_1, \eta_{m+1}) = 0 \tag{2.3.8}$$

Das Randwertproblem (2.3.1), (2.3.2) hat genau dann eine Lösung, falls

$$\mathcal{F}(\eta_1, \ldots, \eta_{m+1}) := \begin{pmatrix} d_1 \\ \vdots \\ d_m \\ r \end{pmatrix} = 0 \in \mathbb{R}^{(n+1)m}. \tag{2.3.9}$$

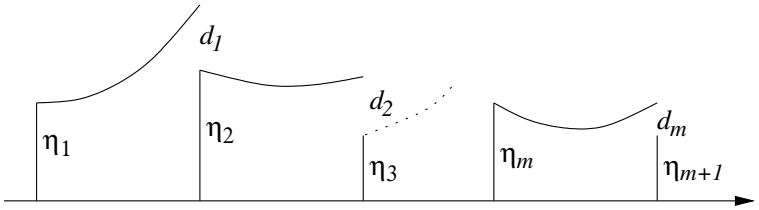


Abbildung 2.2: Mehrfachschießen

Die Jacobimatrix $\mathcal{J}$ von $\mathcal{F}$ hat dabei eine sehr spezielle Struktur, da alle d_i und auch r jeweils nur von wenigen η_j abhängen. Das lineare Gleichungssystem zur Berechnung der Korrekturen hat dann die Form:

$$\mathcal{J}\ \Delta\eta := \begin{pmatrix} G_1 & -I & & & \\ & G_2 & -I & & \\ & & \ddots & \ddots & \\ & & & G_m & -I \\ A & & & & B \end{pmatrix} \begin{pmatrix} \Delta\eta_1 \\ \Delta\eta_2 \\ \vdots \\ \Delta\eta_m \\ \Delta\eta_{m+1} \end{pmatrix} = -\lambda \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_m \\ r \end{pmatrix}, \tag{2.3.10}$$

mit

$$A := \frac{\partial r(\eta_1, \eta_{m+1})}{\partial y_a} \quad ; \quad B := \frac{\partial r(\eta_1, \eta_{m+1})}{\partial y_b} \ .$$

Die

$$G_k := \partial d_k/\partial \eta_k = \partial y(t_{k+1}, t_k, \eta_k)/\partial \eta_k. \tag{2.3.11}$$

sind dabei wieder Sensitivitätsmatrizen der Anfangswertprobleme $y(t_{k+1}, t_k, \eta_k)$ auf den Teilintervallen und genügen selbst den Variationsdifferentialgleichungen

$$G_k'(t) = f_y(t, y(t, t_k, \eta_k))G_k(t) \ , \quad G_k(t_k) = I \ , \quad k = 1, \ldots, m \tag{2.3.12}$$

Die Spalten jeder dieser Matrizen $G_k = \partial d_k/\partial \eta_k$ können wieder approximiert werden durch

$$\frac{\partial d_k}{\partial \eta_{k,j}} \doteq \frac{y(t_{k+1}, t_k, \eta_k + \delta \cdot e_j) - y(t_{k+1}, t_k, \eta_k)}{\delta} \quad j = 1, \ldots, n \ , \quad k = 1, \ldots, m \ , \tag{2.3.13}$$

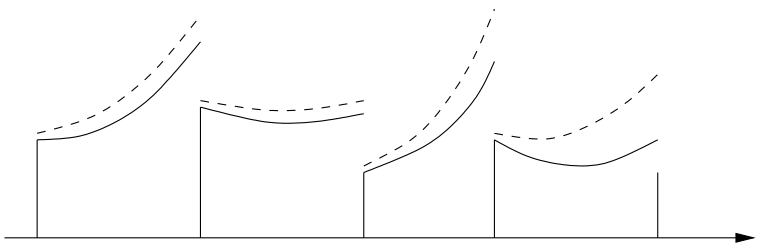


Abbildung 2.3: Variation der Anfangwerte

Lösung des linearen Gleichungssystems:
Wegen der speziellen Struktur, läßt sich das Gleichungssytem sehr einfach auf ein Gleichungssystem kleinerer Dimension reduzieren. Dazu eliminiere man mit Hilfe der ersten Blockzeile die Matrix G_2 aus der zweiten Blockzeile. Die zweite Blockzeile erhält dann die neue Form

$$(G_2G_1, 0, -I, 0, \ldots, 0) \begin{pmatrix} \Delta\eta_1 \\ \Delta\eta_2 \\ \vdots \\ \Delta\eta_m \\ \Delta\eta_{m+1} \end{pmatrix} = -\lambda[G_2d_1 + d_2] \ .$$

Damit läßt sich analog die Matrix G_3 aus der dritten Blockzeile eliminieren. Am Ende erhält man das kondensierte Gleichungssystem

$$[B\underbrace{G_m G_{m-1}\cdots G_2 G_1}_{\doteq Y(t_{m+1},t_1,\eta_1)}+A]\Delta\eta_1 = \quad r + (Bd_m + G_m(d_{m-1} + G_{m-1}(d_{m-2} + \cdots \\ \cdots + G_3(d_2 + G_2 d_1)\cdots)))$$

der Dimension n für $\Delta\eta_1$. (Dies entspricht dem Gleichungssystem beim Einfachschießen (vgl. (**??**)). Ist das gelöst, so erhält man alle anderen $\Delta\eta_i$ aus den ersten m Gleichungen von (2.3.10).
Man beachte jedoch, daß dies für $\|G_k\| > 1$ einer Gaußelimination ohne Pivotsuche entspricht. Daher wird die Kondition des Gleichungssystems oft zu sehr verschlechtert. Bei einer orthogonalen Dreieckszerlegung, etwa mit Houshouldertransformation, bleibt die Struktur jedoch im wesentlichen erhalten. Lediglich in der letzten Blockzeile kommt es zu "fill in".[7]
Es bleibt dabei das Problem geeigneter Startwerte. Dies ist sogar noch etwas erschwert, da jetzt die Lösung auch an Zwischenpunkten geschätzt werden muß.

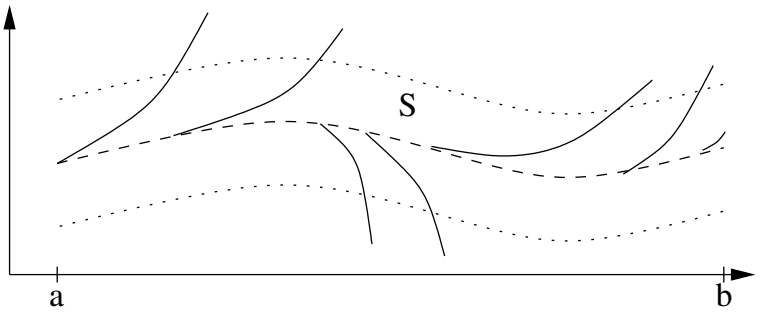


Abbildung 2.4: Anfangsdaten

Ist ein beschränkter Streifen S bekannt, in dem die Lösung vermutet wird, so kann man wie folgt vorgehen. Man startet bei a mit einem Startwert in S, und verfolgt die Lösung des Anfangswertproblems, bis sie den S verläßt. Dort wählt man den ersten Zwischenknoten und startet neu.
In Abbildung 2.4 ist S ein Streifen um eine irgendwie geschätzte Lösung. An jedem Knoten startet man daher in der Mitte des Streifens. In Bereichen in denen das Anfangswertproblem sehr empfindlich von den Startdaten

[7]Bei Mehrpunktrandwertproblemen hängt r von $\eta_1 \ldots, \eta_{m+1}$ ab und die letzte Blockzeile ist ohnehin "voll" besetzt.

abhängt werden dabei automatisch mehr Gitterpunkte plaziert. Glücklicherweise ist diese etwas umständliche Vorbereitung oft unnötig. Es genügt oft, bei Versagen des Verfahrens das Gitter äquidistant zu verfeinern.

Homotopieverfahren:
Oft ist ein aktuelles gegebenes Problem eine Erweiterung oder Veränderung eines früher schon gelösten Problems. Gegeben ist dann eine Problemschar

$$F(t,y,\alpha) := \begin{pmatrix} y' - f(t,y,\alpha) \\ r(y(a),y(b),\alpha) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

mit Lösungen $y(t,\alpha)$. Der Parameter α beschreibt dabei die verschiedenen Modelle. Für α_0 sei das Problem leicht lösbar, bzw. schon gelöst. Gesucht ist die Lösung eines neuen Problems $y(t,\bar{\alpha})$ man wählt dann $y(t,\alpha_0)$ als Startnäherung zur Bestimmung von $y(t,\alpha_1)$ mit $\alpha_1 \approx \alpha_0$. So nähert man sich mit der Folge der α_i langsam an $\bar{\alpha}$ an. Die Schrittweite der Iteration, also $|\alpha_{i+1} - \alpha_i|$ wählt man dabei kleiner, wenn bei der Lösung des Randwertproblems Konvergenzprobleme auftauchen. Am besten so klein, daß man sich im quadratischen Konvergenzbereich des Newtonverfahrens bewegt.

Beispiel 2.3.4 Man versuche etwa eine Raumsonde knapp an einem kleinen Asteroiden vorbei zu steuern, daß er von diesem so abgelenkt wird, daß er eine gewünschte Endposition P erreicht. Bekannt ist dabei der Abschußort, die Position des Asteroiden. sowie die gewünschte neue Richtung. Gesucht ist der Abschußwinkel. Das Problem reagiert extrem empfindlich, weil die Sonde sehr nah am Asteroiden vorbei fliegen muß. Kleinste Fehler im Abschußwinkels führen zu gänzlich falschen Bahnen.

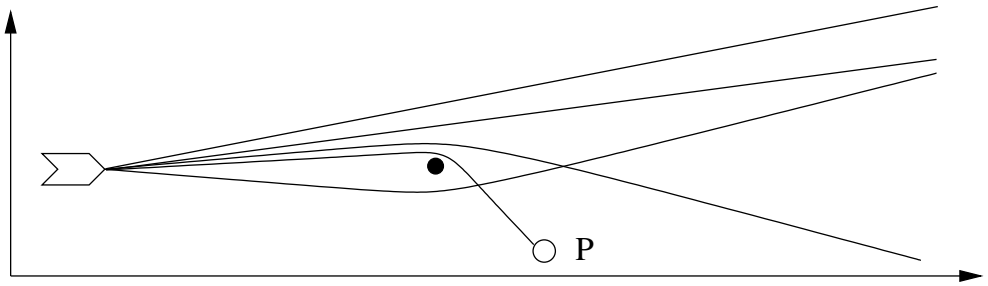


Abbildung 2.5: Swingby-Manöver und Homotopie

Vergrößert man dagegen die Masse des Asteroiden, so kann man in großer Entfernung auf einer Ellipsenbahn vorbei fliegen. Die Bahn hängt dabei in der Nähe der Lösung nur wenig vom Abschußwinkel ab (Strichpunktierte Linien).
Durch sukzessive Verkleinerung des Asteroiden (gepunktete, gestrichelte, durchgezogene Linien) erhält man stets gute Näherungen für das jeweils nächste Problem.

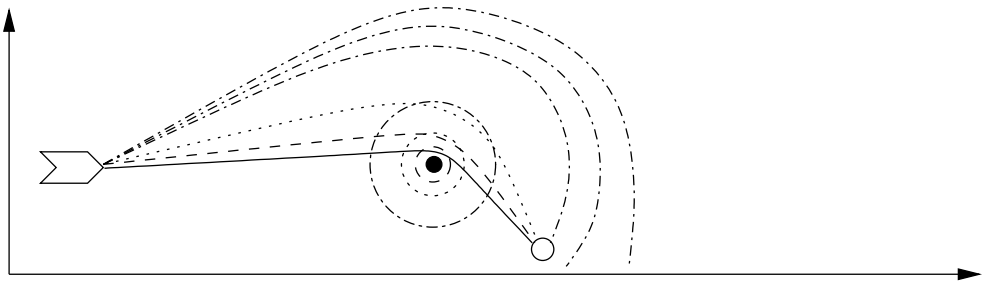


Abbildung 2.6: Homotopie über Masse

Bemerkung 2.3.5 (Natürliche Parameter) Ist überhaupt keine Lösung eines ähnlichen Problems bekannt, so wäre folgender Weg möglich. Bestimme die Lösung irgend eines Randwertproblems

$$y' = g(t,y) \quad ; \quad s(y(a),y(b)) = 0$$

und definiere

$$\begin{aligned} y' &= f(t,y,\alpha) := \alpha f(t,y) + (1-\alpha) g(t,y) \\ 0 &= \alpha r(y(a),y(b)) + (1-\alpha) s(y(a),y(b)) \end{aligned}$$

mit $\alpha \in [0;1]$. Leider funktioniert dieser Trick meist nicht. Erfahrungen zeigen, daß man dagegen gute Chancen hat, wenn der Parameter α eine natürliche Interpretation besitzt, und alle Probleme der Schar in gewisser Weise verwandt sind. D.h., die Lösungen der Probleme zu verschiedenen Parametern alle die gleichen charakteristischen Eigenschaften aufweisen.

Beispiel 2.3.6 Das Problem $f(x) = \ln(x+10) = 0$ ist bei schlechtem Startwert mit dem Newtonverfahren schwer zu lösen. Für $x < 10$ ist f nicht definiert und für $x \gg -10$ führt der erste Newtonschritt in dieses Gebiet.

Löst man statt dessen das völlig andere Problem $g(x) = x^2 - 4$, so findet man mit dem Newtonverfahren leicht eine der beiden Lösungen $x = \pm 2$. Die Nullstellenmenge N_α der eingebetteten Probleme

$$F(x, \alpha) := \alpha f(x) + (1 - \alpha g(x)$$

ändert sich wärend der Homotopie sprunghaft. Sie enthält für $\alpha = 0$ die zwei Lösungen $x = \pm 2$, für kleines positives α ändern sich diese Lösungen stetig (gepunktete Linie), es kommt aber noch eine Lösung in der Nähe von -10 dazu. Diese neue Lösung wird während der Homotopie jedoch nie gefunden. Die Homotopie bricht ab, sobald die beiden größeren Lösungen in der Nähe von 0 verschmelzen und im nächsten Homotopieschritt nur noch eine Lösung existiert (strichpunktierte Linie). Die Chance durch Zufall von diesem Startwert aus in das kleine Konvergenzgebiet des neuen Problems zu treffen ist gering.

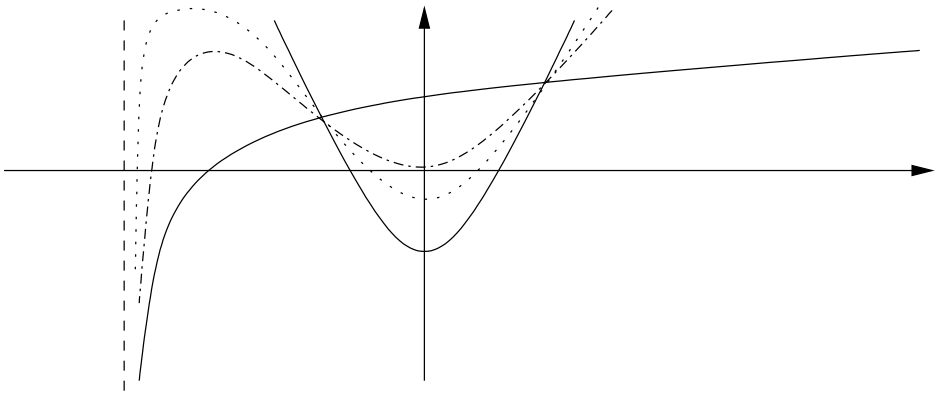


Abbildung 2.7: Homotopie mit künstlichem Parameter

2.3.3 Parameteroptimierung

Eine der wichtigsten Aufgaben der technischen Simulation ist die Bestimmung unbekannter Parameter eines Systems. So kann ein Robotergelenk nur dann richtig gesteuert werden, wenn die Gelenkreibung bekannt ist. Diese ändert sich jedoch während des Betriebs fortlaufend. Aus der Beobachtung der tatsächlichen Roboterbewegung und dem Vergleich mit Simulationen müssen dann die Parameter ständig neu bestimmt und die Steuerung so

angepaßt werden. Andere Parameter können frei gewählt werden. Dies soll so geschehen, daß weitere Bedingungen erfüllt werden.
Wir betrachten also die parameterabhängige Differentialgleichung

$$y' = f(t, y, p) \quad , \quad y \in \mathbb{R}^n \quad , \quad p \in \mathbb{R}^{n_p} \tag{2.3.14}$$

mit dem Parametervektor p. Entsprechend bezeichnen wir die Lösung eines Anfangswertproblems nun mit $y(t, t_0, y_0, p)$.
Formal kann man p auch als weitere konstante Zustandsgrößen ansehen, d.h., man betrachtet das theoretisch equivalente Problem

$$z := \begin{pmatrix} y \\ p \end{pmatrix}, \Rightarrow \dot{z}(t) = F(t, z) := \begin{pmatrix} f(t, y, p) \\ 0 \end{pmatrix}, \quad z(t_0) = \begin{pmatrix} y_0 \\ p \end{pmatrix} \in \mathbb{R}^{n+n_p} \ , \tag{2.3.15}$$

mit Randbedingungen

$$\bar{r}(z_a, z_b) := \bar{r}(y_a, p, y_b, p) := r(y_a, y(b, a, y_a, p)) = 0 \in \mathbb{R}^{n+n_p} \ . \tag{2.3.16}$$

In der Praxis ist dies leicht implementierbar aber nicht effizient, weil die Bibliotheksverfahren zur numerischen Integration die Struktur der Differentialgleichung (2.3.15) nicht ausnützen. Im Falle n_p nicht wesentlich größer als n wird das aber dennoch manchmal gemacht.
Auf jeden Fall können nun also mehr Bedingungen gestellt werden.
Häufig ist z.B. der Zustand y am linken und rechten Rand komplett vorgegeben, d.h., das dynamische System soll von einem gegebenen Ausgangszustand in einen gewünschten Endzustand überführt werden.
Im Fall $n_p = n$ kann man dann eine eindeutige Lösung erhoffen, steht aber vor dem Existenz- und Eindeutigkeitsproblem des modifizierten Randwertproblems (2.3.15), (2.3.16).

Unterbestimmte Probleme: Bei ihnen sind wesentlich mehr Parameter zu wählen als Randbedingungen vorgeschrieben sind. Dies ist etwa der Fall, wenn die Gelenksteuerfunktion eines Roboters durch eine in n_p Teilintervallen stückweise konstante Steuerfunktion approximiert wird und nur die n_p Konstanten beliebig gewählt werden kann.
In solchen Fällen kann man zusätzliche Optimalitätskriterien fordern. Z.B. minimaler Energieverbrauch oder zeitminimale Bewegung.

Überbestimmte Probleme: Ebenso häufig ist der umgekehrte Fall. Entweder sind weniger Parameter vorhanden als Bedingungen, oder Änderungen

der Parameter führen zu linear abhängigen Zustandsänderungen und die Bedingungen sind nicht zu erfüllen. In diesem Fall löst man statt dessen ein nichtlineares Ausgleichsproblem

$$\|r(z_a, z_b)\| = \min$$

Will man ein dynamisches System möglichst gut durch ein Modell beschreiben, dessen prinzipielle Struktur man kennt (oder annimmt) und nur n_p Parameter unbekannt sind, so versucht man diese so zu wählen, daß das mathematische Modell in der Simulation alle Beobachtungen möglichst gut wiedergibt.
Dabei sind oft sehr viele Beobachtungen zu verschiedenen Zeitpunkten vorhanden, die allerdings oft nur einen Teil der Zustände messen.
In der Regel erhebt man zu $k+1$ Zeitpunkten t_j, $j = 0, \ldots, k$ jeweils n_m Messungen $m_{j,i}$, $i = 1, \ldots, n_m$, für die ein funktionaler Zusammenhang $m_{j,i} = M_i(t_j, y(t_j), p)$ bekannt ist. Aufgrund der Messung ist die i-te Komponente von m_j dabei mit einem zufälligen Fehler mit Streuung $\sigma_{j,i}$ behaftet[8]. In der Praxis mißt $m_{j,i}$ oft eine Zustandsgröße $y_i(t_j)$, und zu einem Messzeitpunkt werden meist mehrere (aber nicht alle) Zustände gemessen.
Es kann aber auch sein, daß nur eine Messgröße, z,B. die Extinktion (Lichtabsorbtion) einer Flüssigkeit erhoben wird, die von mehreren Zustandsgrößen abhängt.
Wären Anfangswerte und die Parameter bekannt, und das mathematische Modell exakt, so müsste gelten

$$M_i(t_j, y(t_j, t_0, y_0, p), p) = m_{j,i} \quad ; \quad j = 1, \ldots, k$$

dies ist schon aufgrund der Meßfehler nicht erreichbar. Man verlangt daher

$$\min_{y_0, p} \sum_{\substack{i=1,n_m \\ j=0,k}} \left[\frac{|M_i(y(t_j, t_0, y_0, p), p) - m_{j,i}|}{\sigma_{j,i}} \right]^2 . \tag{2.3.17}$$

Dies ist ein nichtlineares Minimierungsproblem. In jeder Iteration sind dabei Anfangswertprobleme zu lösen bei denen die Zwischenwerte $y(t_j, t_0, y_0, p)$ ausgegeben werden müssen, d.h., die Integration muß bei t_j angehalten werden.

[8] Meßgeräte messen oft mit relativer Genauigkeit $\Rightarrow \sigma_{j,i}$ ist proportional $m_{j,i}$. Manchmal aber auch absolute Meßgenauigkeit (Beispiel Ablesefehler beim Maßband).

In der Regel ist t_0 auch der erste Meßzeitpunkt und die Abweichungen von tatsächlichem Zustand und Simulation bei falsch gewählten Parametern wird mit fortschreitender Zeit immer größer. Die Konvergenz ist daher oft sehr langsam.
Analog der Idee die zum Mehrfachschießen führte, kann man versuchen die Konvergenz dadurch zu beschleunigen, daß man die Simulation in jedem Teilintervall stets mit Anfangsdaten ausführt, die im Bereich der Messungen liegen.
Man verwendet also eine Intervallunterteilung entsprechend der Messpunkte, und in jedem Teilintervall möglichst konsistente Anfangswerte η_j , d,h., $M_i(\eta_j, p) \approx m_{j,i}$ Man lößt dann das Minimierungsproblem

$$\min_{\eta_0 \ldots, \eta_k, p} \sum_{j=1,k} \|\eta_j - y(t_j, t_{j-1}, \eta_{j-1}, p)\|^2 + \sum_{\substack{i=1,n_m \\ j=0,k}} \frac{\|M_i(\eta_j, p) - m_{j,i}\|_2^2}{\sigma_{j,i}^2} \ .$$

Geht man jedoch davon aus, daß das mathematische Modell exakt ist, so erwartet man stets $\eta_j - y(t_j, t_{j-1}, \eta_{j-1}, p) = 0$. Die Auswertung (Messung) von $y(t_j, t_{j-1}, \eta_{j-1}, p)$ hat dabei eine extrem geringe Streuung (bei hoher Integrationsgenauigkeit). Berücksichtigt man dies, und behandelt alle Bedingungen gleichberechtigt, so erhält man

$$\min_{\eta_0 \ldots, \eta_k, p} \sum_{\substack{i=1,n_m \\ j=0,k}} \frac{\|M_i(\eta_{j,i}, p) - m_{j,i}\|_2^2}{\sigma_{j,i}^2} \ . \tag{2.3.18}$$

mit den Nebenbedingungen

$$\begin{aligned} \eta_j - y(t_j, t_{j-1}, \eta_{j-1}, p) &= 0 \quad , \quad j = 1, \ldots, k \\ r(\eta_0, \eta_k, p) &= 0 \ , \end{aligned} \tag{2.3.19}$$

Dazu kommen oft noch weitere Ungleichungebeschränkungen wie etwa $y_i \geq 0$ oder $p_j \geq 0$, bzw allgemein

$$h(t, y, p) \geq 0$$

Dies ist ein nichtlineares Minimierungsproblem der Form

$$\begin{aligned} L(t, y) &\rightarrow \min \\ g(y) &= o \\ h(y) &\geq 0 \end{aligned}$$

das etwa durch SQP-Verfahren gelöst werden kann.

Sequentielle quadratische Programmierung (SQP):
Bei SQP-Verfahren approximiert man das Problem in der Umgebung einer Näherung $y^{(i)}$, indem man die Nebenbedingungen linearisiert und das Zielfunktional quadratisch approximiert, und in jeder Iteration ein restringiertes quadratisches Optimierungsproblem löst.

$$\begin{aligned} g(y^{(i)} + \Delta y) &\approx g(y^{(i)}) + \frac{\partial g}{\partial y}\Delta y \\ h(y^{(i)} + \Delta y) &\approx h(y^{(i)}) + \frac{\partial h}{\partial y}\Delta y \\ L(t, y^{(i)} + \Delta y) &\approx L(t, y^{(i)}) + L_y \Delta y + \Delta y^T A^{(i)} \Delta y \to \min \end{aligned}$$

Dabei ist $A^{(i)}$ im Idealfall die Hessematrix $L_{yy}(t, y^{(i)})$ oder eine positiv definite Approximation davon.

Im speziellen Fall

$$L(t, y) = \|F(t, y)\|_2^2 = F(t, y)^T F(t, y)$$

erhält man eine solche Approximation z.B., indem man $F(t, y)$ linearisiert, d.h.,

$$L(t, y^{(i)} + \Delta y) \approx \left[F(t, y^{(i)}) + \mathcal{J}_F(t, y^{(i)})\Delta y\right]^T \left[F(t, y^{(i)}) + \mathcal{J}_F(t, y^{(i)})\Delta y\right] \ .$$

Im Laufe der Iteration kann man dann die Approximation der Hessematrix L_{yy} nach und nach verbessern.
Man benötigt dann nur die **Sensitivitätsmatrizen**

$$G(t, t_0, y_0, p) := \frac{\partial y(t, t_0, y_0, p)}{\partial y_0}$$

und

$$\bar{G}(t, t_0, y_0, p) := \frac{\partial y(t, t_0, y_0, p)}{\partial p} \ ,$$

aber keine höheren Ableitungen von L , g oder h .

2.4 Finite Differenzenverfahren

Bei den Finite Differenzenverfahren wird versucht, die Lösung y auf einem feinen Gitter zu approximieren. Die Differentialgleichung soll dann an diesen Gitterpunkten näherungsweis erfüllt werden. Näherungsweise bedeutet hier, daß man die Ableitungen durch Differenzenquotienten ersetzt, die nur auf y-Werte des Gitters aufbauen. Man erhält dann ein großes i.a. nichtlineares Gleichungssystem spezieller Struktur.
Vorgehen:

- Wähle Gitter
$$a = x_1 < x_2 < \dots x_N < x_{N+1} = b \tag{2.4.1}$$
Gesucht ist $y(x_i)\,,\ i = 1\dots, N+1\,.$
- Stelle die Differentialgleichung an den Gitterpunkten auf und ersetze Ableitungen von y durch Differenzenquotienten.
- Ordne die Gleichungen in der Reihenfolge der Gitterpunkte.
- Löse das Gleichungssystem.

Bezeichnung: Zu $v \in C([a,b])$ sei
$$R_h v := (v(x_1), \dots, v(x_N))^T$$
der **Restriktionsoperator**.
Differenzenapproximationen der ersten und zweiten Ableitungen sind etwa:
$$y'(x_j) = \underbrace{\frac{y(x_{j+1}) - y(x_{j-1})}{2h}}_{=:D^0 y(x_j)} + \mathcal{O}(h^2) \text{ falls } y \in C^3 \tag{2.4.2}$$
$$y'(x_j) = \underbrace{\frac{y(x_{j+1}) - y(x_j)}{h}}_{=:D^+ y(x_j)} + \mathcal{O}(h) \text{ falls } y \in C^2 \tag{2.4.3}$$
$$y'(x_j) = \underbrace{\frac{y(x_j) - y(x_{j-1})}{h}}_{=:D^- y(x_j)} + \mathcal{O}(h) \text{ falls } y \in C^2 \tag{2.4.4}$$
$$y''(x_j) = \underbrace{\frac{y(x_{j+1}) - 2y(x_j) + y(x_{j-1})}{h^2}}_{=D^+D^- y(x_j)} + \mathcal{O}(h^2) \text{ falls } y \in C^4 \tag{2.4.5}$$

2.4.1 Ein Beispiel

Beispiel 2.4.1

$$\begin{aligned} &y'' - p(x)y' - r(x)y = q(x) \quad ; \quad p, r, q \in C[0;1] \\ &y(0) = c_1 \quad ; \quad y(1) = c_2 \quad ; \quad c_1, c_2 \in \mathbb{R} \end{aligned} \tag{2.4.6}$$

Gitter (hier äquidistant): $\pi := \{x_j = (j-1)h,\, h := 1/N,\, j = 1, \ldots, N+1\}$.
Gesucht ist eine **Gitterfunktion**

$$\eta_j := \eta(x_j, \pi) =: \eta_\pi(x_j) \approx y(x_j),\, j = 1, \ldots, N+1$$

Damit lautet die diskretisierte Differentialgleichung z.B.:

$$\frac{\eta_{j+1} - 2\eta_j + \eta_{j-1}}{h^2} - p_j \frac{\eta_{j+1} - \eta_{j-1}}{2h} - r_j \eta_j = q_j \; ; 2 \le j \le N \tag{2.4.7}$$

mit $p_j = p(x_j)$, $q_j = q(x_j)$, $r_j = r(x_j)$ und den zusätzlichen Randbedingungen

$$\eta_1 = c_1 \quad ; \quad \eta_{N+1} = c_2 \ .$$

Eine lineare Differentialgleichung mit linearen Randbedingungenführt dann immer auf ein lineares Gleichungssystem. Die vorliegende auf ein Gleichungssystem der Bauart

$$\underbrace{\begin{bmatrix} \alpha_1 & \gamma_1 & & & & & \\ \beta_2 & \alpha_2 & \gamma_2 & & & & \\ & \ddots & \ddots & \ddots & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \ddots & \ddots & \gamma_N & \\ & & & & \beta_{N+1} & \alpha_{N+1} \end{bmatrix}}_{A} \eta = \underbrace{\begin{bmatrix} c_1 \\ -q_2 \\ -q_3 \\ \vdots \\ -q_N \\ c_2 \end{bmatrix}}_{b} \tag{2.4.8}$$

mit

$$\left.\begin{array}{rcl} \alpha_j &:=& (\frac{2}{h^2} + r_j) \\ \beta_j &:=& -\frac{1}{h^2} + \frac{p_j}{2h} \\ \gamma_j &:=& -\frac{1}{h^2} - \frac{p_j}{2h} \end{array}\right\} \quad 2 \le j \le N$$

$$\alpha_1 = \alpha_{N+1} = 1 \quad ; \quad \gamma_1 = \beta_{N+1} = 0$$

Bemerkung: Bei nicht äquidistantem Gitter ist eine Equilibrierung des linearen Gleichungssystems vorteilhaft. Dazu multipliziert man jede Zeile mit dem entsprechenden h^2 und erhält als Einträge der j-ten Zeile von A

$$-1 - \frac{h}{2} p(x_i) \quad 2 + h^2 r(x_j) \quad -1 + \frac{h}{2} p(x_i) \ ,$$

also für kleines h Einträge von etwa ähnlicher Größenordnung.

Die so entstehenden Gleichungssysteme sind groß und dünn besetzt mit Bandstruktur (hier tridiagonal) oftmals symmetrisch (hier im Fall $p(x) = 0$) und diagonaldominant (hier wenn $h \leq 2/\|p\|_\infty$ und $r \geq 0$. Daher ist Gaußelimination ohne Pivotsuche möglich (oder Cholesky-Zerlegung). Die Bandstruktur bleibt dabei also erhalten und die Zerlegung ist billig.
Im Fall der Differentialgleichung (2.4.6) ist A zwar nicht notwendig positiv definit (da nicht symmetrisch) aber diagonaldominant falls $p_j/(2h) \leq 1/h^2$ bzw.

$$h \leq \frac{2}{\|p\|_\infty} \quad \text{und} \quad r_j \geq 0 \; . \tag{2.4.9}$$

Unter der Voraussetzung (2.4.9) ist also (2.4.8) eindeutig lösbar und es ist keine Pivotwahl notwendig.[9]
Wir untersuchen nun die Approximationsgüte der Lösung von (2.4.8). Dazu betrachten wir den sogenannten **Differentialoperator**[10]

$$Lu(x) := L(u,x) := u''(x) - p(x)u'(x) - r(x)u(x) \tag{2.4.10}$$

und einen passenden zugehörigen **Differenzenoperator**[11]

$$\begin{aligned} L_\pi u(x) \;\; := \;\; & \underbrace{\frac{u(x+h) - 2u(x) + u(x-h)}{h^2}}_{=D^+D^-u(x)} \\ & -p(x) \underbrace{\frac{u(x+h) - u(x-h)}{2h}}_{=D^0u(x)} - r(x)u(x) \; . \end{aligned} \tag{2.4.11}$$

[9] Für $r_j = \varepsilon > 0$, $p_j = 0$ ist A strikt diagonaldominant und positiv definit. Ohne Pivotwahl sind alle Eliminantionsfaktoren betragsmäsig < 1. Für $\varepsilon \to 0$ bleiben die Eliminationsfaktoren ≤ 1. Unter der Bedingung (2.4.9) wird der erste Eliminationsfaktor $l_{2,1}$ sogar betragsmäßig noch kleiner und negativ, da $-1 - p(x_i)h/2 > 0$ und daher im ersten Eliminationsschritt das 2-te Diagonalelement noch größer, als im Fall $r_j = \varepsilon > 0$, $p_j = 0$. Durch Induktion folgt damit für alle Eliminationsfaktoren $|l_{i+1,i}| \leq 1$.

[10] Ein Operator bildet Funktionen auf Funktionen ab

$$L : C^k[a;b] \to C^k[a;b] \quad ; \quad L : u \mapsto L[u] : C^k[a;b] \to \mathbb{R}^n$$

hier mit $k = 2$ und $n = 1$.
$Lu := L[u]$, $Lu(x) := L[u](x) = (L[u])(x)$. D.h. L und u binden stärker als $u(x)$.

[11] L_π ist auch für punktweise definierte Funktionen definiert. Z.B. auf $\eta : \{x_i\} \mapsto \{\eta_i\}$.

Damit läßt sich (2.4.7) schreiben als

$$L_\pi \eta_j = q(x_j) \quad , \quad 2 \leq j \leq N \ . \tag{2.4.12}$$

Bemerkung: D^+D^- und D^0 passen zusammen, weil sie die gleiche Approximationsordnung besitzen. $\Rightarrow \tau = \mathcal{O}(h^2)$.

Aufgabe 2.4.2 Man zeige, daß $D^+D^-D^+D^-u(x) = u^{(4)}(x)\mathcal{O}(h^2)$

Definition 2.4.3 (Lokaler Abbrechfehler)
Man betrachte die Differentialgleichung

$$L[u] = f(x) \qquad (*)$$

Sei u eine beliebige glatte Funktion und

$$\tau_j[u] := L_\pi[u](x_j)) - L[u](x_j) \quad 2 \leq j \leq N \ ,$$

dann heißt

$$\tau[u] := \max_{2 \leq j \leq N} |\tau_j[u]|$$

der **lokale Abbrechfehler** *der Differenzengleichung (bezüglich u).*[12] *Für u Lösung von (*) und η Lösung von (2.4.8) heißt $R_h u - \eta$* **Diskretisierungsfehler**

Bei linearen Differentialgleichungen $L[u] = f$ gilt $L_\pi u = AR_h u$ und $A\eta = R_h f$. Damit ist der Diskretisierungsfehler

$$(R_h u - \eta) = A^{-1}A(R_h u - \eta) = A^{-1}(AR_h u - R_h f)$$

Der lokale Abbrechfehler der exakten Lösung $AR_h u - R_h f$ heißt auch **lokaler Defekt** von (2.4.8).[13]

Beispiel 2.4.4 Für $u \in C^4[0;1]$ gilt gemäß der Konstruktion der Differenzenquotienten in Beispiel 2.4.1 $\tau \leq ch^2$, mit c abhängig von u.

[12] τ beschreibt, wie gut Differenzenoperator und Differentialoperator für beliebiges u übereinstimmen. Nicht nur entlang der Lösung.

[13] Er beschreibt, wie gut die Lösung die Differenzengleichung erfüllt und überträgt sich mit A^{-1} auf den Diskretisierungsfehler. Daher ist die Bedingung $\|A^{-1}\| < K$ entscheidend.

Wegen

$$\lim_{h\to 0} \tau[u] = 0 \Rightarrow \lim_{h\to 0} L_\pi = L$$

ist man daher an Abbrechfehlern hoher Ordnung interessiert.

Definition 2.4.5 (Konsistenz) *Ein Differenzenverfahren heißt* **konsistent**, *falls*

$$\lim_{h\to 0} \tau[u] = 0 \ .$$

Es ist von der **Konsistenzordnung** p, *falls für jedes genügend glatte vorgegebene* u *gilt:*

$$\tau[u] = \mathcal{O}(h^p) \ .$$

Definition 2.4.6 (Konvergenzordnung)
Ein finites Differenzenverfahren heißt **konvergent**, *wenn*

$$\max_{1\le j\le N+1} \|y(x_j) - \eta_j\| \to 0 \quad \textit{für} \quad h \to 0$$

Es hat die **Konvergenzordnung** p, *wenn*

$$\max_{1\le j\le N+1} |e_j| = \mathcal{O}(h^p)$$

Natürlich können wir Konvergenz nur an diskreten Punkten erwarten.
$\eta_k = \eta_k(h)$ hängt von der Schrittweite ab und liefert eine Näherung an der Stelle $x_0 + kh$. Für festes x kommen nur Schrittweiten $h = (x - x_0)/n$ in Frage. In diesem Fall ist $\eta_{1+n} = \eta_{1+(x-x_0)/h} \approx y(x)$. Unter **diskreter Konvergenz** verstehen wir dann

$$\lim_{\substack{n\to\infty \\ h=(x-x_0)/n}} \eta_{1+(x-x_0)/h}(h) = y(x) \ .$$

(Entsprechend für nicht äquidistante Stützstellen.) Dabei gehen wir davon aus, daß das spezielle Randwertproblem (2.4.6) gut gestellt ist, d.h., es soll gelten:

$$\|y\|_\infty \le \kappa \max\{|c_1|, |c_2|, \| \underbrace{Ly}_{q} \|_\infty\} \tag{2.4.13}$$

Definition 2.4.7 (Stabilität) *Ein Differenzenschema heißt* **stabil**,
wenn für alle Gitter π *mit* h *genügend klein und alle Gitterfunktionen* $u_\pi = R_h u$ *gilt:*

$$|u_j| \le K \max\{|c_1|, |c_2|, \max_{2\le j\le N} |L_\pi u_j|\} \ \textit{für}\ j = 1, \ldots, N+1 \tag{2.4.14}$$

mit K *unabhängig von* π *bzw.* h *und nicht zu groß, wenn* κ *nicht zu groß.*

Das Differenzenscheme soll also nicht wesentlich schlechter konditioniert sein wie das Randwertproblem.
Für das Differenzenverfahren (2.4.8) folgt mit $\eta_\pi \equiv u_\pi$, daß (2.4.14) äquivalent ist mit der Bedingung

$$\|A^{-1}\| \leq K \tag{2.4.15}$$

Beweis:($\Leftarrow$) $\eta_\pi = A^{-1}b \Rightarrow \|\eta_\pi\|_\infty \leq \|A^{-1}\|_\infty \|b\|_\infty \leq K\|b\|_\infty \Rightarrow$ (2.4.14), da die rechte Seite b von (2.4.8) nur abhängt von c_1, c_2 und $L_\pi u_j$.
($\Rightarrow$) A ist regulär, denn sei u_π Eigenvektor zum Eigenwert 0, dann gilt $Au_\pi = 0$. u_π erfüllt dann nicht (2.4.14), da $\|u_\pi\| > 0 = K \max\{0, \ldots, 0\}$. Zu b beliebig existiert also $\eta_\pi = A^{-1}b$.

$$\begin{aligned} \|\eta_\pi\|_\infty \leq K\|b\| \quad &\Rightarrow \quad \|A^{-1}b\|_\infty \leq K\|b\|_\infty \\ &\Rightarrow \quad \frac{\|A^{-1}b\|_\infty}{\|b\|_\infty} \leq K \ \forall b \neq 0 \Rightarrow \|A^{-1}\| \leq K \end{aligned}$$

■

Wir betrachten nun den **globalen Fehler**

$$e_j := y(x_j) - \eta_j \tag{2.4.16}$$

Da L_π linear $\Rightarrow$

$$L_\pi e_j = L_\pi y(x_j) - L_\pi \eta_j = L_\pi y(x_j) - q(x_j) = L_\pi y(x_j) - Ly(x_j) = \tau_j[y] \tag{2.4.17}$$

mit $e_1 = e_{N+1} = 0$. Der globale Fehler genügt also den Differenzengleichungen mit dem lokalen Abbrechfehler als Inhomogenität
Aus (2.4.14) mit $u_j \equiv e_j$ folgt sofort

$$\|e_j\| \leq K\tau[y] \leq Kch^2 \quad ; \quad 2 \leq j \leq N \tag{2.4.18}$$

falls $u \in C^4[0;1]$
Bemerkungen:

- In Definition 2.4.6 bedeutet $h \to 0$: Man betrachte einen festen Punkt $\hat{x}$ und wähle eine beliebige Familie von Gittern mit $h \to 0$ und $\hat{x}$ als gemeinsamen Gitterpunkt $\hat{x} = j_h h$. Die Folge von Approximationen η_{j_h} konvergiert dann gegen $y(\hat{x})$.
- Wegen (2.4.18) gilt:
  Konsistenzordnung = Konvergenzordnung =: Ordnung.

- Für das betrachtete Beispiel gilt:
  Konsistenz + Stabilität $\rightarrow$ Konvergenz
- In (2.4.18) stammt die Ordnung von der Approximationsformel. K hängt aber auch vom Problem ab. Man kann daher abhängig vom Problem ein geeignetes Differenzenverfahren zu wählen.

2.4.2 Lineare Randwertprobleme 1. Art

Wir betrachten wieder das RWP

$$\begin{aligned} Lu(x) &= -u''(x) + b(x)u' + c(x)u = f(x) \quad a \le x \le b \\ u(a) = \alpha \quad ; \quad u(b) &= \beta \end{aligned} \tag{2.4.19}$$

Zunächst transformieren wir das Problem auf das Einheitsuintervall nit homogenmen Randbedingungen. Dazu führen wir die Transformationen

$$u(x) = v(x) + \alpha\frac{x-b}{a-b} + \beta\frac{x-a}{a-b}$$

und

$$x = (b-a)\xi + a \quad \Rightarrow \xi \in [0;1]$$

aus.
Wir betrachten also ohne Einschränkung

$$\begin{aligned} Lu(x) &= -u''(x) + b(x)u' + c(x)u = f(x) \quad 0 \le x \le 1 \\ u(0) = 0 &= u(1) \end{aligned} \tag{2.4.20}$$

Im symmetrischen Fall gilt dann

Satz 2.4.8 *In (2.4.20) gelte $c, f \in C[0;1]$, $b(x) = 0$ und $c(x) \ge 0$, dann existiert genau eine Lösung $u \in C^2[0;1]$ von (2.4.20).*

Beweis: Eindeutigkeit: Seien u_1 und u_2 zwei Lösungen, dann löst $u := u_1 - u_2$ das zugehörige homogene RWP

$$-u'' + cu = 0 \quad , \quad u(0) = u(1) = 0$$

$$\begin{aligned} \Rightarrow 0 &= \int_0^1 (-u'' + cu)u\mathrm{d}x \\ &= -u\ddot{u}|_0^1 + \int_0^1 (u')^2 + cu^2\mathrm{d}x \\ &= \int_0^1 (u')^2 + cu^2\mathrm{d}x \end{aligned}$$

$$\stackrel{c\geq 0}{\Rightarrow} u \equiv 0$$

Existenz: Die allgemeine Lösung von (2.4.20) hat die Bauart

$$u(x) = \alpha_1 u_1(x) + \alpha_2 u_2(x) + u_p(x)$$

Mit homogenen Lösungen u_1 und u_2. Zu gegebener partikulärer Lösung u_p löst man

$$\begin{aligned} u_1(0)\alpha_1 + u_2(0)\alpha_2 &= -u_p(0) \\ u_1(1)\alpha_1 + u_2(1)\alpha_2 &= -u_p(1) \end{aligned}$$

immer lösbar, da homogene Lösung eindeutig lösbar, also .

$$\begin{pmatrix} u_1(0) & u_2(0) \\ u_1(1) & u_2(1) \end{pmatrix}$$

regulär. ■

Das allgemeine Problem (2.4.20) läßt sich symmetrisch machen. Dazu setzt man

$$u(x) := v(x) \exp\left(\frac{1}{2}\int_0^x b(t)\mathrm{d}t\right)$$

und erhält

$$\tilde{L}v = -v"(x) + \tilde{c}(x)v(x) = \tilde{f}(x) \quad ; \quad 0 \leq x \leq 1 \quad ; \quad v(0) = v(1) = 0$$

mit

$$\begin{aligned} \tilde{c}(x) &= c(x) + \tfrac{1}{4}b^2(x) - \tfrac{1}{2}b'(x) \\ \tilde{f}(x) &= f(x)\exp\left(-\tfrac{1}{2}\int_0^x b(t)\mathrm{d}t\right) \end{aligned}$$

Damit erhält man auch im unsymmetrischen Fall:

Lemma 2.4.9 *Gilt $\tilde{c}(x) > 0$ und $b'(x) \in C[0,1]$, so existiert ein eindeutige Lösung von (2.4.19)*

Bei der diskretisierten symmetrischen Gleichung

$$\begin{aligned} Lu(x) &= -u"(x) + c(x)u = f(x) \quad 0 \leq x \leq 1 \\ &\quad u(0) = 0 = u(1) \end{aligned} \tag{2.4.21}$$

entsteht ein Gleichungssystem

$$A\eta = R_h f$$

mit symmetrischer zeilendiagonaldominanter Matrix A. die Außerdiagonalelemente sind alle negativ und die Diagonale ist positiv.

Definition 2.4.10 *Eine Matrix A heißt L_0-**Matrix**, wenn $a_{i,j} \leq 0 \; \forall i \neq j$. Sie heißt* **inversmonoton** *wenn gilt:*

$$Ax \leq Ay \Rightarrow x \leq y \; .$$

(Die Ungleichung ist dabei komponentenweise zu verstehen.)
*Eine inversmonotone L_0-Matrix heißt M-**Matrix**.*

Bemerkung 2.4.11 A ist inversmonoton genau dann, wenn A^{-1} existiert und $A^{-1} \geq 0$ (komponentenweise).

Beweis: Übung.

Lemma 2.4.12 ((M-Kriterium)) *Eine L_0-Matrix A ist inversmonoton genau dann, wenn $v > 0$ existiert mit $Av > 0$ (komponentenweise). Es gilt dann*

$$\|A\|_\infty^{-1} \leq \frac{\|v\|_\infty}{\min_k (Av)_k} \; .$$

Beweis: ($\Leftarrow$): Ist A inversmonoton, so setze $v = A^{-1}(1, \ldots, 1)^T$. Dann gilt $v > 0$, da $A^{-1} > 0$ und nicht singulär und natürlich $Av > 0$.
($\Rightarrow$): Sei $v > 0$ mit $Av > 0$. A ist L_0-Matrix, also gilt $a_{i,j} v_j \leq 0$ falls $i \neq j \; \Rightarrow a_{ii} > 0$. Also ist die Diaginaolmatrix $A_D := diag(a_{ii})$ invertierbar. Sei $P := A_D^{-1}(A_D - A)$, Dann gilt $P \geq 0$ und $A = A_D(I - P)$ sowie $(I - P)v = A_D^{-1} A v > 0$, also $Pv < v$. Mit der Norm

$$\|x\|_v := \max_i \frac{x_i}{v_i}$$

erhält man für ein x mit $\|x\|_v = 1$ die (komponentenweise) Abschätzung $|x| \leq |v|$ und $Px \leq P|x| \leq Pv \Rightarrow \|Px\|_v \leq \|Pv\|_v$. Für die zugeordnete Matrixnorm gilt dann

$$\|P\|_v = sup_{\|x\|_v = 1} \|Px\|_v = \|Pv\|_v = \|v\|_v = 1$$

Damit existiert $(I-P)^{-1}$ und es gilt

$$(I-P)^{-1} = \sum_{j=0}^{\infty} P^j > 0 \quad \text{(Neumannsche Reihe)}$$

Da $A = A_D(I-P)$, existiert auch $A^{-1} \geq 0$.
Mit $\vec{\mathbf{1}} := (1,\ldots,1)^T$ folgt $Av \geq \min_k (Av)_k \cdot \vec{\mathbf{1}}$ und da A inversmonoton

$$\|A^{-1}\vec{\mathbf{1}} \leq \frac{v}{\min_k (Av)_k} \Rightarrow \|A\|_\infty^{-1} \leq \frac{\|v\|_\infty}{\min_k (Av)_k} \; .$$

■

2.4.3 Lineare Randwertprobleme

Das Einführungsbeispiel war sehr speziell (linear, eindimensional)
Wir betrachten jetzt den allgemeineren linearen Fall

$$\begin{aligned} y' &= A(x)y + q(x) \quad ; \quad a \le x \le b && (2.4.1)\\ c &= B_a y(a) + B_b y(b) \quad ; \quad A, B_a, B_b \in \mathbb{R}^{n\times n} && (2.4.2) \end{aligned}$$

Wendet man auf die Differentialgleichung ein Ein- oder Mehrschrittverfahren mit festgelegter Schrittweitenfolge an, so erhält man automatisch ein Differenzenverfahren, sobald man sich entschieden hat, durch welche Differenzenapproximation y' ersetzt werden soll.

Beispiel 2.4.1 (Trapezregel)

$$\frac{\eta_{j+1}-\eta_j}{h_j} = \frac{1}{2}[A(x_{j+1})\eta_{j+1} + A(x_j)\eta_j] + \frac{1}{2}[q(x_{j+1}) + q(x_j)] \quad ; \quad 1 \le j \le N$$

$$r(\eta_1, \eta_{N+1}) = 0$$

Beispiel 2.4.2 (Mittelpunktsregel)

$$\frac{\eta_{j+1}-\eta_j}{h_j} = \frac{1}{2}A(x_{j+1/2})(\eta_{j+1}+\eta_j) + q(x_{j+1/2})$$

Beide Schemata führen auf ein großes lineares Gleichungssystem für η_π der Form

$$\underbrace{\begin{bmatrix} S_1 & R_1 & & & \\ & S_2 & R_2 & & \\ & & \ddots & \ddots & \\ & & & S_N & R_N \\ B_a & & & & B_b \end{bmatrix}}_{A} \eta_\pi = \underbrace{\begin{bmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_N \\ \eta_{N+1} \end{bmatrix}}_{b} \qquad (2.4.3)$$

Dabei gilt im Falle der Trapezregel:

$$\begin{aligned} S_j &= -\frac{1}{h_j}I - \frac{1}{2}A(x_j)\\ R_j &= \frac{1}{h_j}I - \frac{1}{2}A(x_{j+1}) \quad ; \quad 1 \le j \le N\\ q_j &= \frac{1}{2}[q(x_{j+1}) - q(x_j)] \end{aligned}$$

und im Falle der Mittelpunktsregel:

$$\begin{aligned} S_j &= -\frac{1}{h_j} I - \frac{1}{2} A(x_{j+1/2}) \\ R_j &= \frac{1}{h_j} I - \frac{1}{2} A(x_{j+1/2}) \quad ; \quad 1 \le j \le N \\ q_j &= q(x_{j+1/2}) \end{aligned}$$

Die Matrix hat die gleiche Blockstruktur wie die Mehrzielmatrix. Im Unterschied dazu stehen dort in der Diagonalen jedoch Einheitsmatrizen. Dies liegt an der Asymmetrie der Schießverfahren, die eine Schießrichtung auszeichnen. Bei Differenzenverfahren wäre dies auch möglich, wenn man etwa ein explizites Eulerverfahren zur Generierung des Differenzenschemas verwendet. Da am Ende jedoch ohnehin ein großes Gleichungssystem gelöst werden muß, und bei Randwertproblemen keine Richtung ausgezeichnet ist, werden bevorzugt implizite symmetrische Verfahren eingesetzt.

2.4.4 Nichtlineare Randwertprobleme

Wir betrachten jetzt den nichtlinearen Fall

$$\begin{aligned} y' &= f(x,y) \quad ; \quad a \le x \le b \\ 0 &= r(y(a), y(b)) \end{aligned}$$

Beispiel 2.4.3 (Trapezregel)

$$\begin{aligned} \frac{\eta_{j+1} - \eta_j}{h_j} &= \frac{1}{2}[f(x_{j+1}, \eta_{j+1}) + f(x_j, \eta_j)] \quad ; \quad 1 \le j \le N \\ r(\eta_1, \eta_{N+1}) &= \end{aligned}$$

Beispiel 2.4.4 (Mittelpunktsregel)

$$\begin{aligned} \frac{\eta_{j+1} - \eta_j}{h_j} &= f(x_{j+1/2}, \frac{1}{2}(\eta_{j+1} + \eta_j)) \quad ; \quad 1 \le j \le N \\ r(\eta_1, \eta_{N+1}) &= 0 \end{aligned}$$

Beide Schemata führen auf ein großes nichtlineares Gleichungssystem für η_π vom Typ

$$F(s) = 0 \quad ; \quad s \equiv \eta = (\eta_1 \ldots, \eta_{N+1})^T \in \mathbb{R}^{n(N+1)}$$

Die Lösung mit dem Newtonverfahren erhält man gemäß:

$$\begin{aligned} DF(s^{(m)})\Delta s^{(m)} &= -F(s^{(m)}) \\ s^{(m+1)} &= s^{(m)} + \Delta s^{(m)} \quad ; \quad m = 0, 1, \ldots \end{aligned}$$

Die Trapezregel führt auf den Differenzenoperator $L_{\pi,T}$ (T für Trapez) mit

$$L_{\pi,T}\eta_j := \frac{\eta_{j+1} - \eta_j}{h_j} - \frac{1}{2}[f(x_{j+1}, \eta_{j+1}) + f(x_j, \eta_j)]$$

und wir erhalten

$$F(s) = (L_{\pi,T}\eta_1, \ldots, L_{\pi,T}\eta_N, r(\eta_1, \eta_{N+1}))^T$$

Die Newtonkorrekturgleichungen lauten dann

$$\frac{\Delta\eta_{j+1} - \Delta\eta_j}{h_j} - \frac{1}{2}[A(x_{j+1})\Delta\eta_{j+1} + A(x_j)\Delta\eta_j] = -L_{\pi,T}\eta_j^{(i)} \; ; \quad 1 \le j \le n$$

und

$$B_a\Delta\eta_1 + B_b\Delta\eta_{N+1} = -r(\eta_1^{(i)}, \eta_{N+1}^{(i)})$$

mit

$$\begin{aligned} \Delta s^{(i)} &\equiv \Delta\eta_\pi = (\Delta\eta_1, \ldots, \Delta\eta_{N+1})^T \\ A(x_j) &:= \frac{\partial f}{\partial y}(x_j, \eta_j^{(i)}) \\ B_a &:= \frac{\partial r}{\partial y_a}(\eta_1^{(i)}, \eta_{N+1}^{(i)}) \quad ; \quad B_b := \frac{\partial r}{\partial y_b}(\eta_1^{(i)}, \eta_{N+1}^{(i)}) \end{aligned}$$

Daraus ergeben sich die nächsten Iterierten

$$\eta_j^{(i+1)} = \eta_j^{(i)} + \Delta\eta_j^{(i)} \quad ; \quad j = 1, \ldots, N+1$$

Ein Vergleich zeigt, daß die Newtonkorrekturen einer linearen Differenzengleichung genügen, die man erhält, wenn man die Trapezregel auf das lineare Randwertproblem

$$\begin{aligned} y' &= \frac{\mathrm{d}}{\mathrm{d}t}y^{(i)} + A(x)(y - y^{(i)}) \quad ; \quad a \le x \le b \\ c &= r(\eta_1^{(i)}, \eta_{N+1}^{(i)}) + B_a(y(a) - \eta_1^{(i)}) + B_b(y(b - \eta_{N+1}^{(i)}) \end{aligned}$$

vom Typ (2.4.1), (2.4.2) anwendet. $y^{(i)}$ ist dabei eine beliebige glatte Funktion die die $\eta_j^{(i)}$ interpoliert, d.h., $y^{(i)}(x_j) = \eta_j^{(i)}$. Die Konstruktion von $y^{(i)}$ ist dabei unnötig, da $y^{(i)}$ nur an den Interpolationsstellen benötigt wird.
Man kann also entweder das exakte nichtlineare Randwertproblem diskretisieren, und das entstehende nichtlineare Gleichungssystem linearisieren und iterativ lösen, oder in jeder Iteration das Randwertproblem um die letzte Näherung linearisieren und das bei der Diskretisierung entstehende lineare Gleichungssystem exakt lösen. Letzteres nennt man auch **Quasilinearisierung** Beide Verfahren sind in diesem Fall äquivalent. Dies gilt für viele Differenzenverfahren.

2.4.5 Differenzenverfahren höherer Ordnung

Alle bisherigen Differenzenverfahren hatten nur geringe Ordnung (≤ 2). Höhere Ordnung erhält man nur, wenn man Differenzenapproximationen höherer Ordnung für y' verwendet. Solche Approximationen gewinnt man aus den Ein- und Mehrschrittverfahren für Anfangswertprobleme. Mehrschrittverfahren lassen sich direkt in entsprechende Differenzenverfahren umwandeln.
Bei Einschrittverfahren ist eine kleine Modifikation der Notation notwendig, da bei einem s-stufigen Einschrittverfahren die rechte Seite f nicht nur an Näherungen hoher Ordnung ausgewertet werden, sondern auch an $s-1$ Zwischenpunkten.
Das Interval $[a;b]$ sei also zerlegt in

$$a = x_1 < x_2 < \cdots < x_{N+1} = b \ .$$

Jedes Teilintervall $[x_i; x_{i+1}]$ sei noch einmal unterteilt gemäß

$$x_i = x_{i,1} \leq \cdots \leq x_{i,s} \leq x_{i+1,1} = x_{i+1} \ .$$

Man beachte, daß die sogenannten **Kollokationspunkte** $x_{i,j}$ nicht notwendig verschieden sein müssen.
Das allgemeinste (vollimplizite) lineare Einschrittverfahren hat die Bauart

$$\eta_{i+1,1} := \eta_{i+1} = \eta_i + h_i \sum_{j=1}^{s} b_j k_{i,j} \ .$$

In der j-ten Stufe eines Einschrittverfahrens löst man iterativ

$$k_{i,j} := f\left(\underbrace{x_i + c_j h_i}_{x_{i,j}}, \underbrace{\eta_i + h_i \sum_{l=1}^{s} a_{j,l} k_{i,l}}_{=:\eta_{i,j}}\right) =: f(x_{i,j}, \eta_{i,j}) \ ,$$

wobei in jeder Iteration genau einmal die rechte Seite f aufgerufen wird. Die x-Argumente zweier solcher Aufrufe können gleich sein, die y-Argumente sind es in der Regel nicht. Packt man alle $\eta_{i,j}$, $i = 1, \ldots, N$, $j = 1, \ldots, s$ und η_{N+1} in einen Vektor η, so erhält man ein Differenzenverfahren, bei dem jedoch nur die $\eta_{i,1}$ mit entsprechend hoher Ordnung konvergieren. Das nichtlineare Gleichungssystem hat die Dimension $n \times N \times s$ und wird wieder iterativ gelöst. Verwendet man explizite Runge-Kutta Ansätze, so ließen sich die Variablen auf $n \times N$ reduzieren. Dennoch müßte ein nichtlineares Gleichungssystem gelöst werden. Der große Vorteil expliziter Verfahren bei Anfangswertproblemen entfällt also. Man verwendet daher vorwiegend vollimplizite Runge-Kutta-Verfahren und nutzt die damit verbundenen Freiheitsgrade zur Erhöhung der Ordnung voll aus.
Dies führt auf die **Gauß-Verfahren**, hier für $s = 2$ und $s = 3$.

$\frac{1}{2} - \frac{\sqrt{3}}{6}$	$\frac{1}{4}$	$\frac{1}{4} - \frac{\sqrt{3}}{6}$
$\frac{1}{2} + \frac{\sqrt{3}}{6}$	$\frac{1}{4} + \frac{\sqrt{3}}{6}$	$\frac{1}{4}$
	1/2	1/2

$\frac{1}{2} - \frac{\sqrt{5}}{10}$	$\frac{5}{36}$	$\frac{2}{9} - \frac{\sqrt{15}}{15}$	$\frac{5}{36} - \frac{\sqrt{15}}{30}$
$\frac{1}{2}$	$\frac{5}{36} + \frac{\sqrt{15}}{24}$	$\frac{2}{9}$	$\frac{5}{36} - \frac{\sqrt{15}}{24}$
$\frac{1}{2} + \frac{\sqrt{5}}{10}$	$\frac{5}{36} + \frac{\sqrt{15}}{30}$	$\frac{2}{9} + \frac{\sqrt{15}}{15}$	$\frac{5}{36}$
	$\frac{5}{18}$	$\frac{4}{9}$	$\frac{5}{18}$

bei denen die Ordnung $p = 2s$ erreicht wird.
Bei **Lobatto-Verfahren** für $s = 2$ und $s = 3$

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

0	0	0	0
1/2	5/24	1/3	-1/24
1	1/6	2/3	1/6
	1/6	2/3	1/6

erreicht man nur $p = 2s - 2$, dafür kann eine Stufe doppelt (in zwei benachbarten Intervallen) verwendet werden, und das Verfahren entspricht im Aufwand einem Gauß-Verfahren mit $s - 1$ Stufen.

Bei impliziten Verfahren werden in jeder Stufe nichtlineare Gleichungssysteme iterativ nahezu exakt gelöst. η_{i+1} wird also mit hohem Aufwand konsistent zu η_i bestimmt. Nachdem bei Randwertproblemen $\eta_{i,1}$ selbst nur eine Approximation ist, die im Rahmen der äußeren Iteration ständig verbessert wird, lohnt es sich nicht großen Aufwand für die Konsistenz zu treiben. Man wird daher bei der Lösung der impliziten Stufengleichungen die Iteration abbrechen sobald die Genauigkeit deutlich kleiner ist als der Fehler der Differenzengleichungen. Oft verwendet man nur eine Iteration.

2.4.6 Extrapolationsverfahren

Man kann die Ordnung auch durch Extrapolation erhöhen. Kombiniert man 2 Differenzenverfahren mit Schrittweite h und $h/2$, so entsprechen die zusätzlichen Punkte des feineren Gitters den Kollokationspunkten. Auf den gemeinsamen Punkten lassen sich die Approximationen dann extrapolieren.
Jedes übliche Extrapolationsverfahren ist bei fester Extrapolationsordnung und fester Schrittweitenfolge äquivalent zu einem Runge-Kutta-Verfahren. Damit ist es äquivalent zu einem Differenzenverfahren. Man erhält es genau durch extrapolation der verschiedenen Lösungen der Differenzenlösungen.
D.h., Einschrittverfahren $\rightarrow$ Extrapolation $\rightarrow$ äquivalentes Differenzenverfahren
und Einschrittverfahren $\rightarrow$ äquivalentes Differenzenverfahren $\rightarrow$ Extrapolation
liefern die gleichen Resultate.

2.4.7 Kollokationsverfahren

Verlangt man von den Runge-Kutta-Verfahren,
daß alle Kollokationspunkte $x_{i,j}\,,\ j = 1, \ldots, s$ verschieden sein müssen,
so existiert bei Vorgabe von $\eta_{i,j}\,,\ 1 \leq j \leq s$ im Intervall $[x_i; x_{i+1}]$ eindeutig ein Polynom $\eta'_\pi(x) \in \Pi_{s-1}$ mit

$$\eta'_\pi(x_{i,j}) = f(x_{i,j}, \eta_{i,j}) \quad ; \quad 1 \leq j \leq s\ ,$$

und damit ein eindeutiges Polynom

$$\eta_\pi(x) := \eta_i + \int_{x_i}^{x} \eta'_\pi(\xi)\mathrm{d}\xi \in \Pi_s$$

Man erhält also eine stückweise aus Polynomen $\in \Pi_s$ stetig zusammengesetzte Approximation an die Lösung. Die Approximation ist an den Knoten x_i besonders gut (maximal $p = 2s$), allerdings können dort Knicke auftreten, d.h., die allgemeine Konvergenz ist oft deutlich geringer. In diesem Fall spricht man von **Superkonvergenz** an den Knoten.
Man kann nun auf die Berechnung der $\eta_{i,j}$ nach dem Vorbild von Einschrittverfahren verzichten, und einfach nach stückweise Polynomen suchen, die an gewissen Punkten die Differentialgleichung erfüllen. Es bietet sich an zu verlangen

$$\eta_\pi(x_{i,j}) = \eta_i + \int_{x_i}^{x_{i,j}} \eta'_\pi(\xi)\mathrm{d}\xi \in \Pi_s = \eta_{i,j}$$

An den Kollokationspunkten $x_{i,j}$ ist dann zwar die Approximation oft von geringerer Ordnung, dafür erfüllt die numerische Näherungsfunktion $\eta_\pi(x)$ an den Kollokationspunkten die Differentialgleichung exakt.

Definition 2.4.5 (Kollokationslösung)
Gegeben sei ein Gitter π *und* **kanonische Punkte**

$$0 \leq \alpha_1 < \alpha_2 < \cdots < \alpha_{s-1} < \alpha_s \leq 1 \ .$$

Eine **Kollokationslösung** *für das nicht lineare Randwertproblem*

$$\begin{aligned} y' &= f(x,y) \quad ; \quad a \leq x \leq b \\ 0 &= r(y(a), y(b)) \end{aligned}$$

ist eine stetige, stückweise polynomiale Funktion $\eta_\pi(x)$ *mit*

$$\eta_\pi(x)|_{[x_i;x_{i+1}]} \in \Pi_s \quad ; \quad 1 \leq i \leq N$$

die die Differentialgleichung an den Kollokationspunkten erfüllt, d.h.,

$$\eta'_\pi(x_i + \alpha_j h_i) = f(x_i + \alpha_j h_i, \eta_\pi(x_i + \alpha_j h_i))$$

und den Randbedingungen

$$r(\eta_\pi(a), \eta_\pi(b)) = 0$$

genügt.

Satz 2.4.6 (Äquivalenz von Kollokation und Runge-Kutta-Verfahren)

i) Jedes Kollokationsverfahren ist äquivalent zu einem voll-impliziten Runge-Kutta-Verfahren.
ii) Ein voll-implizites s-stufiges Runge-Kutta-Verfahren der Ordnung $p \geq s$, mit paarweise verschieden $c_1 \ldots, c_s$, bei denen alle $k_{i,j}$ gemäß

$$k_{i,j} = \eta_i + \sum_{l=1}^{s} a_{i,l} \underbrace{f(x_{i,l}, k_{i,l})}_{f(x_{i,l})} \approx \eta_i \int_{x_i}^{x_{i,j}} f(\xi)\mathrm{d}\xi$$

mit einer Quadraturformel berechnet werden, welche die Approximationsordnung s besitzt ist äquivalent zu einem Kollokationsverfahren.

Beweis: (i):

$$\eta'_\pi(x) = \sum_{l=1}^{s} L_{i,l}(x) f(x_{i,l}, \eta_{i,j})$$

$$\eta_\pi(x) := \eta_i + \int_{x_i}^{x} \sum_{l=1}^{s} L'_{i,l}(x) f(x_{i,l}, \eta_{i,l})\mathrm{d}\xi$$

$$\eta_\pi(x_{i,j}) = \eta_i + \sum_{l=1}^{s} f(x_{i,l}, \underbrace{\eta_{i,l}}_{k_{i,l}}) \underbrace{\int_{x_i}^{x_{i,j}} L_{i,l}(\xi)\mathrm{d}\xi}_{=:a_{i,l}} = \underbrace{\eta_{i,j}}_{k_{i,j}}$$

Ein Kollokationsverfahren hat also die Bauart eines impliziten RK-Verfahrens. Die Koeffizienten des RK-Tableaus sind Integrale über die Ableitungen der Lagrangepolynome und nach Wahl der Knoten $x_{i,j}$ leicht berechenbar.
(ii): Für den zweiten Teil des Satzes zeigt man, daß die Ordnungsbedingung das Runge-Kutta-Verfahren bereits eindeutig bestimmt. Da ein Kollokationsverfahren zu einem Runge-Kutta-Verfahren äquivalent ist, das diese Bedingung erfüllt, folgt die Umkehrung. (siehe Ascher, Mattheij, Russel: Numerical Solution of Boundary Value Problems for ODEs) ∎

Verallgemeinerungen:
Die bisher betrachteten Kollokationsverfahren waren alle eine Untermenge der Runge-Kutta-Verfahren.
Man kann an Stelle der vielen Kollokationsbedingungen im Inneren der Teilintervalle aber auch andere Bedingungen stellen und die Parameter der Kollokationspolynome dazu benutzen, um höhere Differenzierbarkeit an den Knoten x_i zu erreichen. Verlangt man z.B., $\eta_\pi \in C^1[a;b]$, so ist dies noch

mit Einschrittverfahren zu erreichen, z.B. mit Lobatto-Verfahren. Ab $\eta_\pi \in C^2[a;b]$ erhält man jedoch Kollokations-Verfahren, die sich nicht aus Einschrittverfahren ableiten lassen.

2.4.8 Mehrpunktformeln - kompakte Schemata

Zur Erhöhung der Konsistenzordnung eines Differenzenverfahrens könnte man für die Differenzenapproximation der Ableitungen mehr Gitterpunkte verwenden.

$$u(k)(x_i) \approx \frac{1}{h} \sum_{j=-l}^{l} c_j u(x_i + jh)$$

(Ableitung des entsprechenden Interpolationspolynoms.) Das Problem hierbei ist, daß an den Rändern nur Gitterpunkte auf einer Seite vorliegen und andere Formeln verwendet werden müssen und daß die entstehende Koeffizientenmatrix nicht mehr diagonaldominant ist. Der Nachweis der Regularität unabhängig von h ist dann nicht mehr so einfach.
statt dessen nutzt man die Linearität der Differentialgleichung aus, und konstruiert speziell optimierte Verfahren für spezielle lineare Operatoren.

Definition 2.4.7 *Ein Differenzenverfahren für eine Differentialgleichung m-ter Ordnung heißt* **kompakt**, *wenn es nur $m+1$-Approximationen verwendet.*

Sie sind von der Bauart

$$\frac{1}{h^m}(\alpha_{k,0}u_k + \alpha_{k,1}u_{k+1} + \cdots + \alpha_{k,m}u_{k+m}) = \sum_{j=1}^{J} \beta_{k,j} f(\tau_{k,j})$$

Dabei nutzt man insbesondere aus, daß f an mehr und ganz anderen Stellen ausgewertet werden kann.
Es gilt dann

Satz 2.4.8 *Zu lösen sei eine lineare Differentialgleichungen m-ter Ordnung*

$$L[u](x) = f(x)$$

Es gelte
(i) die Normierungsbedingung

$$\sum_{j=1}^{J} \beta_{k,j} = 1$$

(ii) Die Koeffizienten $\alpha_{k,j}$ und $\beta_{k,j}$ seien beschränkt für alle $h < h_0$ und für die Matrix des entstehenden Differenzenschemas $M(h)$ gelte $\|M^{-1}\| < K$ für alle $h < h_0$.
(iii) Das Differenzenverfahren sei exakt für alle Polynome $p \in \prod_l$, d.h., es gelte

$$\frac{1}{h^m}(\alpha_{k,0}p_k + \alpha_{k,1}p_{k+1} + \cdots + \alpha_{k,m}p_{k+m}) = \sum_{j=1}^{J} \beta_{k,j} L[p](\tau_{k,j}) \;\forall\, p \in \prod_l$$

Dann ist das Verfahren konsistent von der Ordnung $l - m + 1$.

Beweis: Sei $p^* \in \prod_l$ das Bestapproximierende Polynom an die Lösung u, dann gilt $\|p^* - u\|_\infty < u^{(l+1)}(\xi)\mathcal{O}(h^{l+1})$. Das Differenzenschema verwendet Differenzen für u und für f, die durch Matrizen D_1 und D_2 dargestellt werden können. Das Schema zur Berechnung der Approximation η lautet dann

$$\frac{1}{h^m} D_1 \eta = D_2 f = D_2 L[u] \;.$$

(iii) bedeutet

$$\frac{1}{h^m} D_1 p = D_2 L[p] \forall\, p \in \prod_l$$

Dann gilt:

$$\begin{aligned}
(\frac{1}{h^m}D_1 - D_2 L)[p^*] &= 0 \\
(\frac{1}{h^m}D_1 - D_2 L)[p^* - u] &\overset{(ii)}{=} \mathcal{O}(h^{l+1-m}) \\
(\frac{1}{h^m}D_1 - D_2 L)[u] = (\frac{1}{h^m}D_1 - D_2 L)[u - p^* + p^*] &= \mathcal{O}(h^{l+1-m}) \\
\frac{1}{h^m} D_1 u &= D_2 L u + \mathcal{O}(h^{l+1-m}) \\
&= D_2 f + \mathcal{O}(h^{l+1-m}) \\
\frac{1}{h^m} D_1 \eta &= D_2 f \\
\|\eta - u\|_\infty \leq \left[\frac{1}{h^m} D_1\right]^{-1} \mathcal{O}(h^{l+1-m}) &\overset{(ii)}{\leq} K\mathcal{O}(h^{l+1-m})
\end{aligned}$$

■

Die Umkehrung gilt auch, denn sonst existiert für die homogene Gleichung

$$Lu = 0$$

mit der trivialen Lösung $u = 0$ eine nichttriviale Approximation $p \in \prod_l$.

Beispiel 2.4.9 Für das RWP

$$-u"(x) = f(x) \quad , \quad 0 < x < 1 \quad , \quad u(0) = u(1) = 0$$

erhält man das kompakte Schema

$$-\frac{u_{i-1} - 2u_i + u_{i+1}}{h^2} = \frac{1}{24}(5f((x_i - \tau)) + 14f(x_i)5f((x_i + \tau)))$$

mit $\tau = \sqrt{2/5}$.

Für jeden linearen Operator L erhält man so andere Differenzenverfahren.

Aufgabe 2.4.10 Für das RWP

$$-u"(x) + cu(x) = f(x) \quad , \quad 0 < x < 1 \quad , \quad u(0) = u(1) = 0$$

versuche man in Abhängigkeit von c ein kompaktes symmetrisches Verfahren maximaler Ordnung der Art

$$-\frac{a_{i-1}u_{i-1} + a_iu_i + a_{i+1}u_{i+1}}{h^2} = b_{i-1}f((x_i - \tau) + b_if(x_i) + b_{i+1}f((x_i + \tau))$$

herzuleiten.

2.4.9 Boxverfahren

In vielen Anwendungen beschreibt die Differentialgleichung Austauschvorgänge zwischen Teilgebieten (Boxen) bzw. Teilintervallen. Die Knotenwerte u_i der zu approximierenden Funktion u sind dann repräsentativ für den Wert von u im Teilintervall

$$[x_{i-1/2}; x_{i+1/2}] := [x_i - h/2; x_i + h/2] \ ,$$

und wir erhalten Differenzenschamata, die teilweise Integrale über eine Box und teilweise den Gradient benachbarter Boxen approximiert.

Beispiel 2.4.11 Für das RWP

$$-(K(x)u')' + c(x)u(x) = f(x) \quad , \quad 0 < x < 1 \quad , \quad u(0) = u(1) = 0$$

mit $K(x), c(x) \geq 0$ gilt

$$\int_{x_{i-1/2}}^{x_{i+1/2}} -(K(x)u')' + c(x)u(x)\mathrm{d}x = -(K(x)u'(x))|_{x_{i-1/2}}^{x_{i+1/2}} + \int_{x_{i-1/2}}^{x_{i+1/2}} c(x)u(x)\mathrm{d}x$$
$$\stackrel{!}{=} \int_{x_{i-1/2}}^{x_{i+1/2}} f(x)u(x)\mathrm{d}x$$

und man erhält das einfache Differenzenschema

$$\frac{1}{h}\left(K_{i+1/2}\frac{u_{i+1}-u_i}{h} - K_{i-1/2}\frac{u_i - u_{i-1}}{h}\right) + c_i u_i = f_i \ .$$

Es besitzt die Konsistenzordnung 2. Der Stabilitätsnachweis ist bei diesen Verfahren üblicherweise viel einfacher.

2.5 Finite-Element-Verfahren

Viele Differentialgleichungen (insbesondere auch partielle Differentialgleichungen) werden unter der Voraussetzung hergeleitet, daß die Lösung genügend glatt ist und ein gewisse Minimaleigenschaft besitzt.
Wir betrachten Differentialgleichungen der Art

$$\begin{aligned} -u'' + b(x)u'(x) + c(x)u(x) &= f(x) \\ 0 < x < 1 \quad , \quad u(0) &= u(1) = 0 \ , \end{aligned} \tag{2.5.1}$$

und erwarten natürlich eine zweimal stetig differenzierbare Lösung

$$u \in C^2(0,1) \cap C^0[0,1] \ .$$

Ein solches u nennen wir auch **klassische Lösung** von (2.5.1).
Analog zum Minimierungsproblem einer Funktion $f : x \in \mathbb{R} \mapsto \mathbb{R}$, bei der man unter der Annahme der Differenzierbarkeit Kandidaten für ein Minimum aus der Lösung von $f'(x) = 0$ gewinnt, obwohl die Differenzierbarkeit gar nicht garantiert ist (Beispiel: $f(x) = |x|$), so kann es auch hier sein, daß (2.5.1) gar keine Lösung besitzt, aber ein passendes Minimierungsproblem die Situation besser beschreibt und eine Lösung besitzt, die nur dann, wenn sie hinreichend glatt ist auch (2.5.1) löst.
Wir entwickeln daher einen abgeschwächten Lösungsbegriff, der uns die Chance läßt auch tatsächlich eine Lösung angeben zu können. Dazu sei

$$X := \{v \in C^1(0,1) \cap C^0[0,1] : \ v(0) = v(1) = 0\} \ ,$$

Dann gilt (Multiplikation von (2.5.1) mit v):

$$\int_0^1 (-u'' + b(x)u'(x) + c(x)u(x))v(x)\mathrm{d}x = \int_0^1 f(x)v(x)\mathrm{d}x \quad \forall v \in X \ .$$

Partielle Integration ergibt:

$$\begin{aligned} &\int_0^1 (-u''(x) + b(x)u'(x) + c(x)u(x))v(x)\mathrm{d}x = \\ &= \underbrace{-u'(x)v(x)|_0^1}_{=0} + \int_0^1 u'(x)v'(x) + b(x)u'(x)v(x) + c(x)u(x)v(x)\mathrm{d}x \\ &= \int_0^1 f(x)v(x)\mathrm{d}x \quad \forall v \in X \ . \end{aligned} \tag{2.5.2}$$

Existiert ein klassische Lösung u^*, und liegt ein $u \in C^2(0,1) \cap C^0[0,1]$ vor, welches in einem kleinen Intervall von u^* abweicht, so kann man ein $v \in X$ angeben, welches diesen Unterschied aufdeckt indem (2.5.2) nicht erfüllt ist. (2.5.2) erzwingt jetzt aber nicht mehr $u \in C^2(0,1) \cap C^0[0,1]$, sondern ist auch für $u \in X$ definiert.
Das allgemeinere Problem lautet also: Find $u \in X$ mit

$$\underbrace{\int_0^1 u'(x)v'(x) + b(x)u'(x)v(x) + c(x)u(x)v(x)\mathrm{d}x}_{=:a(u,v)} =$$

$$= \underbrace{\int_0^1 f(x)v(x)\mathrm{d}x}_{=:f(v)} \quad \forall v \in X \tag{2.5.3}$$

Zur Approximation der Lösung wollen wir aber in einem noch etwas anderen Raum arbeiten, der einfachere Funktionen enthält, mit denen die Lösung aber beliebig gut approximiert werden kann.

Definition 2.5.1 *Den Abschluß (closure) der Teilmenge A von V in der Topologie des Raumes V bezeichnen wir mit $cl_v(A)$ und*

$$supp\ v := cl_{\mathbb{R}}\{x \in (0,1)\ :\ v(x) \neq 0\}$$

bezeichnet man als **Träger** *von v.*

Wir werden uns insbesondere für Funktionen interessieren, bei denen der Träger klein ist.

Definition 2.5.2 *Wir definieren*

$$C_0^\infty := \{v \in C^\infty\ :\ supp\ v \subset (0,1)\}$$

$$L^2(0,1) := \{v : [0,1] \to \mathbb{R}\ mit\ \int_0^1 v^2(x)\mathrm{d}x < \infty\}$$

sowie die Norm:

$$\|u\|^2_{L^2(0,1)} := \int_0^1 u^2(x)\mathrm{d}x$$

und das Skalarprodukt

$$(u,v)_{L^2(0,1)} := \int_0^1 u(x)v(x)\mathrm{d}x$$

Wegen

$$\int_0^1 u'(x)v(x)\mathrm{d} = \underbrace{u(x)v(x)|_0^1}_{=0 \forall v \in C_0^\infty} - \int_0^1 u(x)v'(x)\mathrm{d}$$

definieren wir weiter:

Definition 2.5.3 *$w \in L^2(0,1)$ heißt* **verallgemeinerte Ableitung** *von $u \in L^2(0,1)$, falls*

$$\int_0^1 w(x)v(x)\mathrm{d} = -\int_0^1 u(x)v'(x)\mathrm{d} \quad \forall v \in C_0^\infty(0,1)$$

und schreibt dann $w = u'$.

Definition 2.5.4 *$H^1(0,1) := \{v \in L^2(0,1) : \exists v' \in L^2(0,1)\}$ heißt* **Sobolev-Raum** *der Funktionen mit auf $(0,1)$ verallgemeinerter quadratisch integrierbarer Ableitung. In $H^1(0,1)$ verwenden wir als Norm:*

$$\|u\|^2_{H^1(0,1)} := \int_0^1 u'^2(x) + u^2(x)\mathrm{d}x$$

Dann ist $H^1(0,1)$ ein Hilbertraum mit Skalarprodukt

$$(u,v)_{H^1(0,1)} := \int_0^1 u'(x)v'(x) + u(x)(x)\mathrm{d}x \ .$$

Definition 2.5.5

$$H_0^1(0,1) := cl_{H^1(0,1)} C_0^\infty(0,1)$$

sei der Abschluß aller glatten Funktionen mit Nullrandbedingungen bezüglich der Norm in $H^1(0,1)$.
Die Lösung der Aufgabe

$$\textit{Finde } u \in H^1(0,1) : \ a(u,v) = f(v) \ \forall\, v \in H_0^1(0,1) \tag{2.5.4}$$

heißt **schwache Lösung** *von (2.5.1).*

2.5.1 Ritz-Galerkin-Verfahren

Wir betrachten nun allgemeine Aufgaben der Art

$$\text{Finde } u \in V \ : \ a(u,v) = f(v) \ \forall \, v \in V \tag{2.5.5}$$

mit Bilinearform a und linearem Funktional f in einem Hilbertraum V. Die Idee des Ritz-Galerkin-Verfahrens besteht nun darin, das Variationsproblem (2.5.5) in einem endlichdimensionalen Vektorraum $V_n \subset V$ zu lösen, d. h., man betrachtet das Problem

$$\text{Finde } u_n \in V_n \ : \ a(u_n,v) = f(v) \ \forall \, v \in V_n \tag{2.5.6}$$

Ist $\{\Phi_i\}_{i=1,\ldots,n}$ eine Basis von V_n, so ist wegen a, f linear (2.5.6) äquivalent zu

$$\text{Finde } u_n \in V_n \ : \ a(u_n,\Phi_i) = f(\Phi_i) \ \forall \, i = 1,\ldots,n \tag{2.5.7}$$

Mit

$$u_n(x) = \sum_{i=1}^{n} u_{n,i}\Phi_i(x)$$

führt dies auf ein lineares Gleichungssysten für die Koeffizienten $u_{n,i}$

$$a(u_n,\Phi_i) = a\left(\sum_{j=1}^{n} u_{n,j}\Phi_j(x),\Phi_i(x)\right) = \sum_{j=1}^{n} u_{n,j}a\left(\Phi_j(x),\Phi_i(x)\right) = f(\Phi_i) \ ,$$

oder

$$A_n U_n F_n$$

mit

$$\begin{aligned} U_n &:= (u_{n,1},\ldots,u_{n,n})^T \\ A_n &:= (a_{ij})_{i,j=1,2,\ldots,n} = a\left(\Phi_i(x),\Phi_j(x)\right)_{i,j=1,2,\ldots,n} \in \mathbb{R}^{n\times n} \\ F_n &:= (f(\Phi_1),\ldots,f(\Phi_n))^T \in \mathbb{R}^n \ . \end{aligned}$$

Die Frage ist dann, ist das Gleichungssystem immer lösbar, und approximiert die Lösung u_n die Lösung des eigentlichen Problems?

Satz 2.5.6 *Sei V ein normierter Vektorraum, $V_n \subset V$, $dim V_n = n < \infty$ und*

$$a(.,.) : V \times V \to \mathbb{R}$$

eine V-elliptische Bilinearform, d.h.,

$$\exists \gamma > 0 : \ a(v,v) \geq \gamma \|v\|_V^2 \quad \forall \ v \in V$$

und $f : V \to \mathbb{R}$ linear und stetig, d.h.,

$$\exists K > 0 : \ |f(v)| \leq K \|v\|_V \quad \forall \ v \in V \ .$$

Dann gilt:
(i) $A_n U_n = F_n$ ist eindeutig lösbar, also A_n regulär.
(ii) Es gilt die a-priori Abschätzung

$$\|u_n\|_V \leq \frac{K}{\gamma}$$

Bemerkung 2.5.7 (i) V-elliptische ist eine Verallgemeinerung von *positiv definit* Ist $a(.,.) : V \times V \to \mathbb{R}$ V-elliptisch, und $\{\Phi_1, \ldots, \Phi_n\}$ eine Basis von $V_n \subset V$, so ist die Matrix A_{V_n} mit $a_{i,j} := a(\Phi_i, \Phi_j)$ positiv definit mit kleinstem Eigenwert $\lambda_{min} \geq \gamma$.
(ii) ist V unendlichdimensional, so reicht dabei nicht, daß A_{V_n} positiv definit ist für jede endlichdimensionale Teilmenge $V_n \subset V$, sondern es muß eine einheitliche untere Schranke geben

$$\min \lambda(A_{V_n}) \leq \gamma \forall V_n \subset V \ .$$

Beweis: (i) Sei $W_n := (w_{n,1}, \ldots, w_{n,n})^T \neq 0 \in \mathbb{R}^n$, und $w_n := \sum_{i=1}^n w_{n,i} \Phi_i$, dann gilt

$$A_n W_n \cdot W_n = a(w_n, w_n) \geq \gamma \|w_n\|_V^2 > 0$$

also $A_n W_n \neq 0 \ \forall W_n \neq 0$. Damit gilt

$$\gamma \|u_n\|_V^2 \leq a(u_n, u_n) = f(u_n) \leq K \|u_n\|_V$$

■

Die so erhaltene Lösung in V_n ist dabei nicht immer die bestmögliche Approximation von u in V_n, aber es gilt immerhin

Satz 2.5.8 *(Cea Lemma):*
Sei V ein Hilbertraum, $V_n \subset V$, $dim V = n < \infty$

$$a(.,.) : V \times V \to \mathbb{R}$$

eine V-elliptische Bilinearform und stetig d.h., es gilt

$$\exists M > 0 \ : \ |a(u,v)| \leq M\|u\|_V\|v\|_V \ \forall u,v \in V \ ,$$

dann gilt:

$$\|u - u_n\|_V \leq \frac{M}{\gamma} \inf_{v \in V_n} \|u - v\|_V \ .$$

Der Fehler u_n ist also von der Größenordnung des unvermeidbaren Fehlers.
Beweis:

$$a(u - u_n, v) = a(u,v) - a(u_n,v) = f(v) - f(v) = 0 \quad \forall\, v \in V_n \subset V$$

$$\begin{aligned}
\gamma\|u - u_n\|_V^2 &\leq a(u - u_n, u - u_n) = a(u - u_n, u - v + v - u_n) \\
&= a(u - u_n, u - v) + \underbrace{a(u - u_n, v - u_n)}_{=0} = a(u - u_n, u - v) \\
&\leq M\|u - u_n\|_V\|u - v\|_V \quad \forall\, v \in V_n \\
&\Rightarrow \gamma\|u - u_n\|_V \leq M\|u - v\|_V, \quad \forall\, v \in V_n
\end{aligned}$$

Also auch

$$\gamma\|u - u_n\|_V \leq M \inf_{v} \|u - v\|_V \ .$$

■

Folgerung 2.5.9 *Enthält V_n Polynome vom Grad k, so kann die Approximationsgüte der Interpolation von u in $\prod_k$ auf die Approximationsgüte $u_n - u$ in V_n übertragen werden.*

Die Idee ist nun mit zunehmendem n auch Polynome immer höheren Grades hinreichend gut approximieren zu können. Der Aufwand besteht dabei vorwiegend in der Aufstellung und Lösung der linearen Gleichungssysteme $A_nU_n = F_n$.

2.5.2 Finite Elemente

Gegeben sei nun ein Gitter

$$\pi : a = x_1 < x_2 \cdots < x_N < x_{N+1} = b$$

und

$$S^l_{\pi,k} := \{s \in C^{l-1}[a;b], s|_{[x_j;x_{j+1}]} \in \Pi_k, s(a) = s(b) = 0\}$$

der Raum aller Splinefunktionen $\in C^{l-1}$ vom Höchstgrad k
Die Berechnung der Skalarprodukte zur Bestimmung der Koeffizienten von A ist bei allgemeiner Wahl der Basisfunktionen Φ_j sehr aufwendig. Der Aufwand für das Ritzsche Verfahren besteht hauptsächlich in der Aufstellung des linearen Gleichungssystems und in der Lösung. Man versucht deshalb eine Basis zu wählen, bei der jeweils zwei verschiedene Basisfunktionen fast nie gleichzeitig ungleich Null sind. Man wählt daher Basisfunktionen mit möglichst kleinem kompakten Träger, also etwa Polygondachfunktionen

$$\Phi(x) := \begin{cases} \frac{x-xj-1}{x_j-x_{j-1}} & \text{falls} \quad x_{j-1} \le x \le x_j \\ 1 - \frac{x-xj}{x_{j+1}-x_j} & \text{falls} \quad x_j \le x \le x_{j+1} \\ 0 & \text{sonst} \end{cases}$$

In diesem Fall ist A tridiagonal und daher billig zu zerlegen, und jedes $a_{i,j}$ und $\hat{q}_j$ erfordert die Berechnung eines Integral über maximal 2 Teilintervalle.

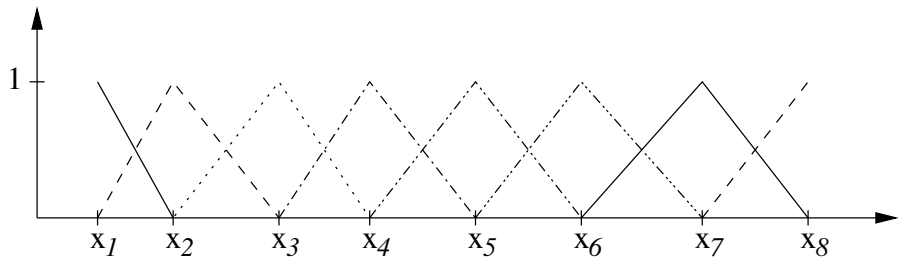


Abbildung 2.8: Basisfunktionen mit kompaktem Träger

Man kann auch glattere Ansatzfunktionen wählen, also z.B $S^2_{\pi,2}$, dann sind die Basisfunktionen alle mindestens C^1 und der kleinste Träger umfaßt mindestens 3 Teilintervalle. A hat dann Bandbreite 5. Dafür genügt meist eine gröbere Intervallunterteilung, d.h., die Dimension von A ist kleiner. Insbesondere falls eine echte Lösung von $L(y) = q$ existiert, sind glättere Ansatzfunktionen vorzuziehen.
Existiert nur eine schwache Lösung, bringt der Aufwand mit glatten Ansatzfunktionen weniger.

2.5.3 Ein Beispiel

Wir betrachten

$$-u''(x) = f(x) \quad 0 < x < 1 \quad u(0) = 0 = u(1)$$

mit dem zugehörigen Variationsproblem

$$\text{Finde } u \in V = H_0^1(0,1) \ : \ a(u,v) = f(v) \ \forall v \in V$$

mit

$$a(u,v) := \int_0^1 u'(x)v'(x)\mathrm{d}x \ ' \quad f(v) := \int_0^1 f(x)v(x)\mathrm{d}x$$

Lemma 2.5.10 *Für $u \in H_0^1(0,h)$ mit $h > 0$ gilt:*

$$\|u\|_{L^2(0,h)} \leq \frac{1}{2\sqrt{2}} h \|u'\|_{L^2(0,h)} \tag{2.5.8}$$

Beweis: Mit der Cauchy-Schwarzschen Ungleichung ($< u, v >^2 \leq \|u\|^2\|v\|^2$) zeigt man zuerst

$$\begin{aligned}
\int_0^{h/2} |u(s)|^2 \mathrm{d}s &= \int_0^{h/2} \left| \int_0^s 1 \cdot u'(x)\mathrm{d}x \right|^2 \mathrm{d}s \\
&\overset{CSUgl.}{\leq} \int_0^{h/2} \left| \int_0^s 1^2 \mathrm{d}x \cdot \int_0^s [u'(x)]^2 \mathrm{d}x \right| \mathrm{d}s \\
&\leq \int_0^{h/2} s\mathrm{d}s \cdot \int_0^{h/2} [u'(x)]^2 \mathrm{d}x = \frac{h^2}{8} \|u'\|^2_{L^2(0,h/2)}
\end{aligned}$$

und analog

$$\int_{h/2}^h |u(s)|^2 \mathrm{d}s \leq \frac{h^2}{8} \|u'\|^2_{L^2(h/2,h)}$$

■

Die Bilinearform ist stetig

$$|a(u,v)| = \left| \int_0^1 u'v'\mathrm{d}x \right| \leq \|u'\|_{L^2(0,1)} \|v'\|_{L^2(0,1)} \leq \|u'\|_{H_0^1(0,1)} \|v'\|_{H_0^1(0,1)}$$

Also ist a stetig mit $M = 1$. Die Bilinearform ist V-elliptisch denn

$$\begin{aligned} a(u,u) &= \left| \int_0^1 u'u' \mathrm{d}x \right| = \|u'\|^2_{L^2(0,1)} \\ &= \frac{8}{9}\|u'\|^2_{L^2(0,1)} + \frac{1}{9}\|u'\|^2_{L^2(0,1)} \\ &\overset{(2.5.8)}{\geq} \frac{8}{9}(\|u'\|^2_{L^2(0,1)} + \|u\|^2_{L^2(0,1)}) = \underbrace{\frac{8}{9}}_{=:\gamma} \|u\|^2_{H^1(0,1)} \end{aligned}$$

(2.5.10)

Ist dann noch $f \in L^2(0,1)$ so gilt

$$|f(v)| = \left| \int_0^1 f(x)v(x)\mathrm{d}x \right| \leq \|f\|_{L^2(0,1)}\|v\|_{L^2(0,1)} \leq \|f\|_{L^2(0,1)}\|\|v\|_{H^1_0(0,1)} \ ,$$

Damit ist f stetig mit $K = \|f\|_{L^2(0,1)}\|$ und nach Satz 2.5.6 existiert eine Lösung.
Wir zerlegen das Intervall in Teilintervalle der Breite h und verwenden als V_n die stückweise linearen Funktionen

$$V_n := \{v \in C[0,1] \ : \ v(0) = v(1) = 0, \ v|_{x_i,x_{i+1}} \in \Pi_1\}$$

Als Basis verwenden wir die Dachfunktionen $\Phi_i(x) \in V_n$ mit

$$\Phi_i(x_j) = 1 \text{ falls } i = j \text{ und } 0 \text{ sonst}$$

Für das lineare Gleichungssystem $A_nU_n = F_n$ generieren wir

$$a_{i,j} = \int_0^1 \Phi_i'(x)\Phi_j'(x)\mathrm{d}x = 0 \forall |i-j| \geq 2$$

und

$$A = tridiag\{-\frac{1}{h_i}; \frac{1}{h_i} + \frac{1}{h_{i+1}}; -\frac{1}{h_{i+1}}\},$$

und

$$F_i = \int_0^1 f(x)\Phi_i(x)\mathrm{d}x = \int_{x_{i-1}}^{x_{i+1}} \Phi_i(x)\mathrm{d}x$$

Lemma 2.5.11 *Unter der Annahme* $u \in H_0^1(0,1)$ *und* $u'' \in L^2(0,1)$ *gilt*

$$\|u - u_n\|_{H_0^1(0,1)} \leq Ch$$

und damit

$$\|u - u_n\|_{L^2(0,1)} \leq Ch \text{ und } \|u' - u_n'\|_{L^2(0,1)} \leq Ch$$

mit $h = \max h_i$.

Beweis: der Beweis läuft über die verallgemeinerte Restgliedformel der Interpolation für die Funktion und deren Ableitungen. Ist $\Pi_n u$ die stückweise lineare Interpolierende von u und $e := \Pi_n u - u$ der Interpolationsfehler, so gilt nach Lemma 2.5.10 in jedem Teilintevall I_i

$$\|e\|_{L^2(I_i)}^2 \leq \frac{1}{8} h^2 \|e'\|_{L^2(I_i)}^2$$

und summiert über alle Teilintervalle

$$\|e\|_{L^2(0,1)}^2 \leq \frac{1}{8} \sum_{i=1}^{n+1} h_i^2 \|e'\|_{L^2(I_i)}^2 \leq \frac{1}{8} h^2 \|e'\|_{L^2(0,1)}^2 \tag{2.5.11}$$

Damit erhalten wir:

$$\begin{aligned}
&\|(u - \Pi_n u)'\|_{L^2(0,1)}^2 - \|u'\|_{L^2(0,1)}^2 + \|\Pi_n u'\|_{L^2(0,1)}^2 \\
&= \int_0^1 ((u - \Pi_n u)')^2 - (u')^2 + (\Pi_n u)')^2 \mathrm{d}x \\
&= \int_0^1 (u')^2 - 2u' \Pi_n u' + ((\Pi_n u)')^2 - (u')^2 + (\Pi_n u)')^2 \mathrm{d}x \\
&= \int_0^1 2(\Pi_n u)'((\Pi_n u)' - u') \mathrm{d}x \\
&= 2 \sum_{j=1}^{n+1} \int_{x_{j-1}}^{x_j} 2(\Pi_n u)'((\Pi_n u)' - u') \mathrm{d}x \\
&= 2 \sum_{j=1}^{n+1} \left\{ [(\Pi_n u)' \underbrace{((\Pi_n u) - u)}_{=0}]_{x_{j-1}}^{x_j} - \int_{x_{j-1}}^{x_j} \underbrace{(\Pi_n u)''}_{=0} ((\Pi_n u) - u) \mathrm{d}x \right\} = 0
\end{aligned}$$

Also

$$\|e'\|_{L^2(0,1)}^2 = \|u'\|_{L^2(0,1)}^2 - \|\Pi_n u'\|_{L^2(0,1)}^2 \leq \|u'\|_{L^2(0,1)}^2 \overset{(2.5.11)}{\leq} \frac{1}{8} h^2 \|u'\|_{L^2(0,1)}^2 \quad (*)$$

und

$$\begin{aligned}\|e'\|^2_{L^2(0,1)} &= \int_0^1 (u-\Pi_n u)'(u-\Pi_n u)'\mathrm{d}x \\ &= \int_0^1 u'(u-\Pi_n u)'\mathrm{d}x - \underbrace{\int_0^1 \Pi_n u'(u-\Pi_n u)'\mathrm{d}x}_{=0} \\ &\overset{p.Int}{=} \int_0^1 u''(-e)\mathrm{d}x \leq \|u''\|_{L^2(0,1)}\|e\|_{L^2(0,1)} \\ &\overset{2.5.11}{=} \frac{1}{2\sqrt{2}}h\|e'\|_{L^2(0,1)}\|u''\|_{L^2(0,1)} \\ \Rightarrow \|e'\|_{L^2(0,1)} &\leq \frac{1}{2\sqrt{2}}h\|u''\|_{L^2(0,1)} \end{aligned} \tag{2.5.12}$$

Mit dem Cea-Lemma kann man dann folgern:

$$\|u-u_n\|^2_{H^1_0(0,1)} \leq c_1\|e\|^2_{H^1_0(0,1)} \leq c_2 h^2$$

■

Bemerkung: unter den gleichen Voraussetzungen kann man auch zeigen:

$$\|u-u_n\|_{L^2(0,1)} \leq Ch^2 \ .$$

Wir benötigen also viel geringere Regularitätsanforderungen an u, und können immer noch quadratische Konvergenz zeigen.
In diesem einfachen Beispiel ist der Nachweis der V-Elliptizität sehr einfach, weil die Bilinearform symmetrisch ist und eine einfache Abschätzung gegen die $\|.\|_{H^1_0(0,1)}$-Norm erlaubt. Bei anderen Differentialgleichungen sind analoge Überlegungen anzustellen, die dabei in der regel weit schwieriger sind. Dies lohnt nur bei Differentialgleichungen mit denen man es sehr oft zu tun hat. Imk Kontext gewöhnlicher Differentialgleichungen lohnt dieser Aufwand oft nicht. Finite Element Methoden werden gleichwohl angewendet