

Numerik von Differentialgleichungen

(Numerik 2B)

Bernhard Schmitt

Wintersemester 2012/13

Inhaltsverzeichnis

1	Numerische Integration und Differentiation	3
1.1	Quadratur	3
1.2	Newton-Cotes-Formeln	4
1.3	Gauß-Quadratur	8
1.4	Adaptive Integration	11
1.5	Richardson-Extrapolation, Romberg-Integration	13
1.6	Numerische Differentiation	17
2	Gewöhnliche Differentialgleichungen	19
2.1	Theoretische Grundlagen	19
2.2	Einschrittverfahren für Anfangswertprobleme	23
2.2.1	Herleitung	23
2.2.2	Konsistenz	26
2.2.3	Stabilität	28
2.2.4	Konvergenz	29
2.2.5	Schrittweitensteuerung	30
2.3	Mehrschrittverfahren	37
2.3.1	Adams-Verfahren	37
2.3.2	Lineare Mehrschrittverfahren und Stabilität	40

2.4	Extrapolationsverfahren	43
2.5	Schießverfahren für Randwertprobleme	46
2.6	Differenzenverfahren für Randwertprobleme	52
3	Partielle Differentialgleichungen	59
3.1	Allgemeine Eigenschaften	59
3.2	Differenzenverfahren für elliptische Randwertprobleme	60
3.2.1	Die Poissongleichung auf einfachen Gebieten	60
3.2.2	Stabilität und Konvergenz	66
3.2.3	Allgemeinere Gebiete und Gleichungen	70
3.3	Finite-Elemente-Verfahren für elliptische Probleme	72
3.3.1	Variationsformulierung	72
3.3.2	Rayleigh-Ritz-Galerkin-Verfahren	74
3.3.3	Spline-Räume, <i>Finite Elemente</i>	76

Einleitung

In den (Natur-) Wissenschaften kann man viele Systeme durch Differentialgleichungen unterschiedlicher Art beschreiben. Verschiedene Fragestellungen bei der Modellierung führen zu unterschiedlichen Beschreibungen. Bei Himmelskörpern etwa kann man sich zunächst für die Bewegung im Weltraum interessieren. Dazu reicht es aus, den Körper als punktförmige Masse im Schwerfeld von Sonnen, Planeten usw. zu betrachten. Die Bahn wird in Abhängigkeit von der Zeit dann durch ein System von gewöhnlichen Differentialgleichungen beschrieben, wie sie im zweiten Kapitel dieser Vorlesung behandelt werden.

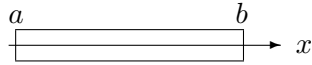
Treten dagegen (Beinahe-) Kollisionen oder Strahlungseffekte auf, spielt die Beschaffenheit der Körper (Gas, Gestein, Eis) eine Rolle. Bei einem Kometen etwa kommt die sichtbare Gestalt am Himmel durch die Einwirkung der Sonnenstrahlung und des Sonnenwinds zustande. Der Schweif bildet sich aus verdampfter Kometenmaterie, seine Form wird durch die Strömung im Sonnenwind bestimmt. Die Modellierung der Materiedichte im dreidimensionalen Raum erfordert partielle Differentialgleichungen. Einfache Fälle solcher Dgln werden in Kapitel 3 besprochen.



Komet Hale-Bopp
(Quelle Wikipedia)

Auch in anderen Disziplinen kann die gewünschte Genauigkeit der Modellierung auf verschiedene Arten von Dgln führen. Betrachtet man in der Biologie/Medizin in einer Population das Gesamtwachstum oder die Ausbreitung einer Infektion, so kommt man mit einem Modell aus gewöhnlichen Dgln aus, wenn sich die Population stark mischt (Bakterien im Reagenzglas, Menschen in einer Stadt). Ist dies nicht der Fall, müssen Wanderungsprozesse mit berücksichtigt werden und führen wieder auf partielle Dgln. Dieses Beispiel wird in der folgenden Aufstellung aufgegriffen.

Einige durch Differentialgleichungen beschriebene Problemstellungen

<p>Einfache Modelle physikalischer, biologischer, chemischer,.. Prozesse; z.B.,</p> <p>a) <i>Massenbewegung unter Gravitation.</i></p> <p>Vereinfachende Annahmen: Vakuum, Massen starr, punktförmig, konzentriert,...</p> <p>a1) Wurf auf Erde: Flugbahn $u(t) = \begin{pmatrix} x(t) \\ y(t) \end{pmatrix}$, Newtonsches Gesetz:</p> $mu''(t) = m \begin{pmatrix} x''(t) \\ y''(t) \end{pmatrix} = -m \begin{pmatrix} 0 \\ g \end{pmatrix}. \text{ Explizite Lösung Wurfparabel}$ <p>Sinnvolle Problemstellungen:</p> <p>"Anfangswertproblem": Start-Ort &-Geschwindigkeit gegeben $\Rightarrow \exists_1$ Lösung!</p> <p>"Randwertproblem": Start-& Zielpunkt gegeben, Lösung?</p> <p>a2) Astronomie (Raumflug, n-Körper-Problem): nicht explizit lösbar, numerische Verfahren erforderlich (hoher Genauigkeit)</p> <p>b) <i>Wachsende Population in Nährmedium</i> (Bakterien, Protozoen,..)</p> <p>b1) Kleiner Behälter, gute Durchmischung:</p> <p>Population $p(t)$ zur Zeit t wächst nach Gesetz</p> $\frac{dp}{dt} = \alpha p - \beta p^2 = \alpha \left(1 - \frac{\beta}{\alpha} p\right) p.$ <p>Dabei: α = Vermehrungsrate, $-\beta p^2$ = Wettbewerb untereinander.</p> <p>Lösung: logistische Kurve</p>	<p>Gewöhnliche Dgln</p>
<p>b2) Rohrförmiger Behälter, p abhängig von x und t: </p> <p>Wanderung (Diffusion) im Rohr prop. zu p_{xx}, für $p(t, x)$ gilt</p> $\frac{\partial p}{\partial t} = \alpha p - \beta p^2 + \gamma \frac{\partial^2 p}{\partial x^2}$ <p>dabei: γ = Diffusionskonstante</p> <p>Zusatzangaben: Startpopulation $p(0, x)$, Rohrende: $p_x(t, a) = p_x(t, b) = 0$, d.h., Anfangswerte in t-Richtung, Randwerte in x-Richtung.</p> <p><<----->></p> <p>Partielle Dgln 2.Ordnung, Einteilung nach Typen (hier in Orts-Gebiet $D \subset \mathbb{R}^2$)</p> <p>Aufgabenstellung nur mit <i>geeigneten Randwerten</i> (typ-abhängig) sinnvoll!</p> <p>a) Parabolische Dgln: Ausgleichsvorgänge $u(t, x, y)$, z.B., Wärmeleitungs-Gl.</p> $u_t = \gamma(u_{xx} + u_{yy}).$ <p>Sinnvolle Aufgabe: Anfangswerte in $t = 0$, Randwerte auf ∂D</p> <p>b) Elliptische Dgln: Gleichgewichtszustände $u(x, y)$, z.B., Potentialgleichung</p> $u_{xx} + u_{yy} = 0.$ <p>Sinnvolle Aufgabe: Randwerte auf ganzem Rand ∂D</p> <p>c) Hyperbolische Dgln: Schwingungen/Wellen $u(t, x, y)$, z.B., Wellen-Gl.</p> $u_{tt} = \alpha(u_{xx} + u_{yy}).$ <p>Sinnvolle Aufgabe: Anfangswerte u, u_t in $t = 0$, Randwerte auf ∂D</p>	<p>Partielle Dgln</p>

1 Numerische Integration und Differentiation

Nur sehr wenige Differentialgleichungen sind explizit lösbar, hier ist der Bedarf nach numerischen Verfahren offensichtlich. Die einfachste Form einer Differentialgleichung ist $u'(x) = f(x)$, bei der nur die Ableitung u' auftritt. Dieses Problem ist natürlich durch einfache Integration zu lösen, $u(x) = u(a) + \int_a^x f(t) dt$. Daher wird dieser einfache Spezialfall wegen seiner besonderen Bedeutung zuerst behandelt, es kommen aber auch schon einige Techniken für allgemeinere Dgl'n zum Einsatz.

Prinzipiell kann man eine aufwändige, oder nicht explizit durchführbare Operation mit einer Funktion f dadurch approximieren, dass man f durch ein einfacheres Modell annähert und die Operation dann mit diesem Modell ausführt. Bei *Integration und Differentiation* bietet sich hier die Polynom-Interpolierende in der Lagrange-Form aus der Numerik 1 an:

$$\begin{aligned}
 f(x) &\cong p_n(x) = \sum_{j=0}^n L_j(x) f(x_j) && \stackrel{?}{\implies} \\
 \int_a^b f(x) dx &\cong \int_a^b p_n(x) dx = \sum_{j=0}^n \left(\int_a^b L_j(x) dx \right) f(x_j) =: \sum_{j=0}^n \alpha_j f(x_j), && (1.0.1) \\
 f'(0) &\cong p'_n(0) = \sum_{j=0}^n L'_j(0) f(x_j) =: \sum_{j=0}^n \delta_j f(x_j).
 \end{aligned}$$

Näherungen für Integral- und Ableitungswerte erhält man also einfach als Linearkombinationen von Funktionswerten. Daher werden jetzt solche Näherungsformeln behandelt, zunächst für die Integration.

1.1 Quadratur

Bekanntlich ist Integrierbarkeit eine sehr schwache Einschränkung an Funktionen, auch sehr irreguläre Funktionen sind integrierbar. Da Fehleraussagen bei Polynominterpolation aber von Ableitungen der Funktion abhängen, ist es bei weniger glatten Integranden vorteilhaft, diese als Produkt einer glatten Funktion f und einer speziellen, schwierigen *Gewichtsfunktion* $g(x) > 0$ (z.B. $\sqrt{x}, 1/\sqrt{x}, \dots$) aufzufassen. Im folgenden wird daher die Integration mit einer festen Gewichtsfunktion g betrachtet, Π_k bezeichnet die Menge der Polynome vom Maximalgrad k .

Definition 1.1.1 Falls das Integral $\int_a^b f(x)g(x)dx$ existiert, heißt

$$\sum_{j=0}^n \alpha_j f(x_j) = \int_a^b f(x)g(x)dx - R_n(f) \tag{1.1.1}$$

eine Quadraturformel und $R_n(f)$ der Quadraturfehler bzw. das Restglied. Die Koeffizienten α_j nennt man Gewichte zu den Knoten

$$a \leq x_0 < x_1 < \dots < x_n \leq b.$$

Die Formel (1.1.1) besitzt die Ordnung $m \in \mathbb{N}$, wenn gilt

$$R_n(p) = 0 \quad \forall p \in \Pi_{m-1}.$$

Für die speziellen Quadraturformeln (1.0.1), die man durch Integration des Interpolationspolynoms aus Π_n erhält, bekommt man explizit die Gewichte

$$\alpha_j = \int_a^b L_j(x)g(x)dx, \quad j = 0, \dots, n. \quad (1.1.2)$$

Aus der Darstellung (Num1/2.1.15) des Interpolationsfehlers, $r_n = f - p_n = \omega_{n+1}(x)f[x_0, \dots, x_n, x]$ mit dem Knotenpolynom $\omega_{n+1}(x) = (x - x_0) \cdots (x - x_n)$ folgt für den Quadraturfehler die Formel

$$\begin{aligned} R_n(f) &= \int_a^b f(x)g(x)dx - \sum_{j=0}^n \alpha_j f(x_j) = \int_a^b r_n(x)g(x)dx \\ &= \int_a^b \omega_{n+1}(x)f[x_0, \dots, x_n, x]g(x)dx. \end{aligned} \quad (1.1.3)$$

Für $f \in C^{n+1}[a, b]$ führt eine Betragsabschätzung auf die Schranke

$$|R_n(f)| \leq \int_a^b |\omega_{n+1}(x)|g(x)dx \frac{1}{(n+1)!} \|f^{(n+1)}\|_\infty. \quad (1.1.4)$$

Dies zeigt insbesondere den

Satz 1.1.2 Für die Ordnung m einer interpolatorischen Quadraturformel (1.1.1), (1.1.2) gilt $m \geq n + 1$.

Die Betragsschranke kann noch verschärft werden, wenn das Knotenpolynom ω_{n+1} im Intervall keine Vorzeichenwechsel besitzt, wenn also gilt

$$\omega_{n+1}(x) = (x - x_0) \cdots (x - x_n) \geq 0 \quad \forall x \in [a, b] \quad (\text{bzw. } \leq 0 \quad \forall x). \quad (1.1.5)$$

Dann existieren nach dem Mittelwertsatz der Integralrechnung sogar Zwischenstellen $\xi_1, \xi_2 \in (a, b)$ mit

$$R_n(f) = \int_a^b \omega_{n+1}(x)g(x)dx \cdot f[x_0, \dots, x_n, \xi_1] = \varrho_n \cdot f^{(n+1)}(\xi_2), \quad (1.1.6)$$

und die Fehlerkonstante ist

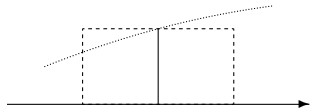
$$\varrho_n := \frac{1}{(n+1)!} \int_a^b \omega_{n+1}(x)g(x)dx.$$

1.2 Newton-Cotes-Formeln

Für eine spezifische Quadraturformel sind Gewicht g und Knoten zu wählen. Mit der einfachsten Wahl $g \equiv 1$ und äquidistanten Knoten $x_j = x_0 + jh, j = 0, \dots, n$, die die Randpunkte enthalten

(d.h. $x_0 = a$, $h = (b - a)/n$, "abgeschlossene Formeln") oder ausschließen ($a < x_0$, $x_n < b$, "offene Formeln") bekommt man die *Newton-Cotes-Formeln*. Für die folgenden, einfachsten Spezialfälle kann man dabei mit Zusatzüberlegungen die Fehlerdarstellung (1.1.6) einsetzen:

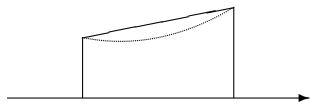
Rechteckregel: $n = 0$, offene Formel, Ordnung 2:



$$\int_a^b f(x)dx = (b - a) f\left(\frac{a+b}{2}\right) + \frac{(b-a)^3}{24} f''(\xi). \quad (1.2.1)$$

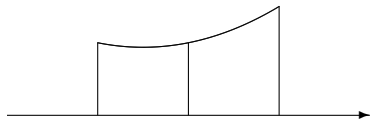
Elementargeometrisch: die Rechteckfläche ist Breite $(b - a)$ mal Höhe $f(\frac{a+b}{2})$, aber auch jedes Trapez mit x -Achse und einer Geraden durch $(\frac{a+b}{2}, f(\frac{a+b}{2}))$ als Seitenlinien hat die gleiche Fläche!

Trapezregel: $n = 1$, abgeschlossene Formel, Ordnung 2:



$$\int_a^b f(x)dx = \frac{b-a}{2} [f(a) + f(b)] - \frac{(b-a)^3}{12} f''(\xi). \quad (1.2.2)$$

Simpsonregel: $n = 2$, abgeschlossene Formel, Ordnung 4 (Keplersche Faßregel):



$$\int_a^b f(x)dx = \frac{b-a}{6} \left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right] - \frac{(b-a)^5}{2880} f^{(4)}(\xi). \quad (1.2.3)$$

In diesen Formeln steht ξ jeweils für eine unbekannte Zwischenstelle in (a, b) . Eine Tabelle der Gewichte α_j und der Fehlerkoeffizienten ϱ_n für weitere Ordnungen folgt weiter unten. Zunächst wird aber nachgeprüft, dass bei der Rechteck- und Simpsonregel die Ordnung tatsächlich $m = n + 2$ ist statt $n + 1$, wie nach Satz 1.1.2 zu erwarten.

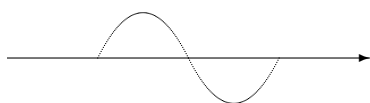
Beweis von (1.2.3): Für $a = 0$ (oBdA) sind die Lagrangepolynome zu $x_0 = 0$, $x_1 = h$, $x_2 = 2h$:

$$L_0(x) = \frac{(x-h)(x-2h)}{2h^2}, \quad L_1(x) = \frac{x(x-2h)}{-h^2}, \quad L_2(x) = \frac{x(x-h)}{2h^2} \in \Pi_2.$$

Daraus berechnen sich die Gewichte

$$\alpha_0 = \int_0^{2h} L_0(x)dx = \frac{1}{2h^2} \left[\frac{1}{3}x^3 - \frac{3h}{2}x^2 + 2h^2x \right]_0^{2h} = \frac{h}{3} = \alpha_2, \quad \alpha_1 = \int_0^{2h} L_1(x)dx = \frac{4}{3}h.$$

Die Fehlerdarstellungen (1.1.4) bzw. (1.1.6) stimmen allerdings noch nicht mit der in (1.2.3) überein. In einer Zusatzüberlegung wird eine *zusätzliche* Interpolationsbedingung $p'(x_1) = f'(x_1)$ herangezogen. Diese Zusatzinformation geht aber überhaupt nicht ein, da ihr Gewicht



$$\int_0^{2h} \omega_3(x)dx = \int_0^{2h} x(x-h)(x-2h)dx = 0 \quad (1.2.4)$$

null ist. Denn in der Newton-Darstellung gilt mit $p_2 = f(x_0)L_0 + f(x_1)L_1 + f(x_2)L_2$ die Dar-

stellung $p_3(x) = p_2(x) + \omega_3(x)f[x_0, x_1, x_1, x_2]$ und somit

$$\int_0^{2h} p_3(x)dx = \int_0^{2h} p_2(x)dx + \underbrace{\int_0^{2h} \omega_3(x)dx}_{=0} f[x_0, x_1, x_1, x_2] = \alpha_0 f(x_0) + \alpha_1 f(x_1) + \alpha_2 f(x_2).$$

Also sind die Integrale über p_2 und p_3 gleich, sogar kubische Polynome werden exakt integriert, und es gilt (1.1.3) sogar mit dem Polynomgrad $n = 3$ und auch (1.1.6) mit $\varrho_3 = -\frac{1}{90}h^5$, da

$$\omega_4(x) = x(x-h)^2(x-2h) \leq 0 \quad \forall x \in [0, 2h]. \quad \blacksquare$$

Dieser Beweis überträgt sich auf alle Newton-Cotes-Formeln mit geradem n , die Ordnung ist dort $n+2$ statt $n+1$. Die folgende Tabelle enthält die Gewichte, Fehlerkonstanten und Ordnungen der Newton-Cotes-Formeln. Da die Gewichte i.w. rational sind, erfolgt die Angabe in der Form $\alpha_j = (b-a)\beta_j/\gamma$,

$$\int_a^b f(x)dx = \frac{b-a}{\gamma} \sum_{j=0}^n \beta_j f(x_j) + c \left(\frac{b-a}{n} \right)^{m+1} f^{(m)}(\xi), \quad \xi \in (a, b). \quad (1.2.5)$$

Abgeschlossene Newton-Cotes-Formeln, $x_j = a + jh$, $j = 0, \dots, n$, $h = \frac{b-a}{n}$.

n	γ	β_0	β_1	β_2	β_3	β_4	c	m
1	2	1	1				-1/12	2
2	6	1	4	1			-1/90	4
3	8	1	3	3	1		-3/80	4
4	90	7	32	12	32	7	-8/945	6
5	288	19	75	50	50	75	-275/12096	6
6	840	41	216	27	272	27	-9/1400	8
7	17280	751	3577	1323	2989	2989	-8183/518400	8
8	28350	989	5888	-928	10496	-4540	-2368/467775	10

Die fehlenden Koeffizienten β_5, \dots, β_8 sind symmetrisch zu ergänzen. Die Gewichte für $n \geq 8$ sind nicht mehr positiv. Dies erhöht etwas die Anfälligkeit der Formeln für Rundungsfehler.

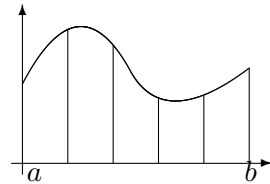
Für die Konvergenz ($n \rightarrow \infty$) der Quadraturformel (1.2.5) gilt das gleiche wie bei der Polynom-Interpolation. Das Anwachsen der höheren Ableitungen $f^{(m)}$, $m = 2(\lfloor \frac{n}{2} \rfloor + 1)$, kann die Konvergenz zerstören. Andererseits wird der Fehler

$$R_n(f) = c \left(\frac{b-a}{n} \right)^{m+1} f^{(m)}(\xi)$$

sehr schnell klein für $(b-a) \rightarrow 0$. Dies kann durch *Unterteilung* des Gesamtintervalls ausgenutzt werden. Dabei wird eine Quadraturformel fester Ordnung *iteriert* angewendet. Bei der Trapezregel (1.2.2), z.B., ergibt sich mit Teilintervallen $[x_{i-1}, x_i]$, $x_j = a + jh$, $j = 0, \dots, n$,

$h := (b - a)/n$, die einfach aufgebaute Näherung

$$\begin{aligned} \int_a^b f(x) dx &= \sum_{i=1}^n \int_{x_{i-1}}^{x_i} f(x) dx \cong \sum_{i=1}^n \frac{h}{2} [f(x_i) + f(x_{i-1})] \\ &= \frac{h}{2} [f(x_0) + 2f(x_1) + \dots + 2f(x_{n-1}) + f(x_n)] =: T_h(f). \end{aligned}$$



mit folgender Fehleraussage.

Satz 1.2.1 Für $n \in \mathbb{N}$ sei $h := (b - a)/n$, $x_i = a + ih$, $f_i := f(x_i)$, $i = 0, \dots, n$. Dann gilt mit einer Zwischenstelle $\xi \in (a, b)$

a) für die iterierte Trapezregel bei $f \in C^2[a, b]$:

$$\int_a^b f(x) dx = \frac{h}{2} (f_0 + 2f_1 + \dots + 2f_{n-1} + f_n) - \frac{b-a}{12} h^2 f''(\xi), \quad (1.2.6)$$

b) für die iterierte Simpsonregel bei $f \in C^4[a, b]$ und n gerade:

$$\int_a^b f(x) dx = \frac{h}{3} (f_0 + 4f_1 + 2f_2 + 4f_3 + \dots + 4f_{n-1} + f_n) - \frac{b-a}{180} h^4 f^{(4)}(\xi). \quad (1.2.7)$$

Beweis Die Beweise von a) und b) verlaufen analog. Bei b) wird die Formel (1.2.3) über die $\frac{n}{2} =: k$ Intervalle $[x_{2i-2}, x_{2i}]$ summiert. Beim Restglied erhält man dabei mit $\xi_i \in (x_{2i-2}, x_{2i})$

$$-\frac{1}{2880} \sum_{i=1}^k (x_{2i} - x_{2i-2})^5 f^{(4)}(\xi_i) = -\frac{32}{2880} h^5 \sum_{i=1}^k f^{(4)}(\xi_i) = -\frac{b-a}{180} h^4 f^{(4)}(\xi). \quad \blacksquare$$

Bemerkung: Man beachte, dass auch die iterierten Formeln (1.2.6), (1.2.7) die Gestalt (1.1.1) aus Definition 1.1.1 besitzen, sogar mit positiven Gewichten, aber nicht in der Form (1.1.2).

Durch die Summation über die Teilintervalle geht eine der h -Potenzen im Fehler verloren und man sieht, dass es in (1.2.5) sinnvoll war als Ordnung der Formel die Zahl m und nicht $m + 1$ anzugeben. Ein großer praktischer Vorteil der iterierten Trapez- und Simpsonregel (z.B. gegenüber der Rechteckregel) ist, dass bei einer Verdopplung der Intervallzahl nur die jeweils neuen Funktionswerte auszuwerten sind, z.B. gilt für die Trapezregel

$$T_{h/2}(f) = \frac{1}{2} T_h(f) + \frac{b-a}{2n} \sum_{j=1}^n f\left(a + \frac{b-a}{2n}(2j-1)\right). \quad (1.2.8)$$

Der wesentliche Rechenaufwand bei beiden iterierten Formeln fällt bei den $n + 1$ Funktionsauswertungen f_0, \dots, f_n an. Bei einem Vergleich der Fehler sieht man, dass gilt

$$R_n(f) \sim \frac{1}{n^2} \text{ (Trapezregel)} \quad \text{bzw.} \quad R_n(f) \sim \frac{1}{n^4} \text{ (Simpsonregel)}$$

bei einem genügend glatten Integranden f . Eine Verdopplung der Zahl n der Funktionsauswertungen verkleinert den Fehler $R_n(f)$ bei der Trapezregel um den Faktor $\frac{1}{4}$, bei der Simpsonregel

dagegen um den Faktor $\frac{1}{16}$ bei vergleichbarem Aufwand. Daher sind bei *glatten* Integranden Formeln hoher Ordnung günstiger.

Beispiel 1.2.2 $\int_1^2 \frac{dx}{x} = \ln 2 = 0.6931472\dots$

h	Trapez	Fehler	Simpson	Fehler
1	0.75	0.056		
1/2	0.7083..	0.015..	0.694..	0.00129..
1/4	0.6970..	0.0038..	0.69325..	0.00010..
1/8	0.6941..	0.00097..	0.69315..	0.000007..
1/16	0.69339..	0.0002..	0.6931476..	0.0000004..

Wegen dieser Beobachtung ist es interessant, mit einer festen Knotenzahl $n + 1$ möglichst hohe Ordnungen zu erreichen. Dies erreicht man bei der folgenden Klasse von Quadraturformeln.

1.3 Gauß-Quadratur

Bei geradem Polynomgrad n hatten die Newton-Cotes-Formeln eine um eins erhöhte Konvergenzordnung, da der erste Fehlerterm im Newton-Polynom wegen der Eigenschaft

$$\int_a^b \omega_{n+1}(x) dx = 0, \quad \omega_{n+1} = (x - x_0) \cdots (x - x_n),$$

des Knotenpolynoms verschwand. Der Grund war die Symmetrie der Knoten x_i . Das Prinzip läßt sich verallgemeinern, durch geeignete Wahl der Knoten x_i können weitere Fehlerterme eliminiert und die Ordnung sogar verdoppelt werden. Dann gilt für den Quadraturfehler (1.1.1)

$$R_n(f) = 0 \quad \forall f \in \Pi_{2n+1}. \quad (1.3.1)$$

Zur Herleitung sei nun $f \in \Pi_{2n+1}$ und $p_n(x) = \sum_{i=0}^n f(x_i)L_i(x)$ das Interpolationspolynom dazu in Π_n . Also gilt jetzt

$$f = p_n + \omega_{n+1} \cdot q_n \quad \text{mit } q_n \in \Pi_n.$$

Daher entspricht die Forderung (1.3.1) der Bedingung

$$R_n(f) = \int_a^b f(x)g(x) dx - \int_a^b p_n(x)g(x) dx = \int_a^b \omega_{n+1}(x)q_n(x)g(x) dx \stackrel{!}{=} 0.$$

Daher ist (1.3.1) offensichtlich genau dann erfüllt, wenn

$$\int_a^b \omega_{n+1}(x)q(x)g(x)dx = 0 \quad \forall q \in \Pi_n. \quad (1.3.2)$$

Prinzipiell gilt dazu: Für eine auf (a, b) positive Gewichtsfunktion g stellt die *Bilinearform*

$$(u, v)_g := \int_a^b u(x)v(x)g(x)dx \quad (1.3.3)$$

ein *Innenprodukt* in $C[a, b]$ dar. Die Forderung (1.3.2) entspricht daher der Orthogonalität

$$\omega_{n+1} \perp_g \Pi_n \iff (\omega_{n+1}, q)_g = 0 \quad \forall q \in \Pi_n. \quad (1.3.4)$$

Für Ordnung $2n + 2$ in (1.3.1) ist ω_{n+1} also g -orthogonal zu allen Polynomen vom Grad n zu wählen!

Die Konstruktion einer Familie von Orthogonalpolynomen ω_k , $k \in \mathbb{N}$, zum Innenprodukt (1.3.3) läßt sich einfach mit dem Gram-Schmidt-Orthogonalisierungsverfahren durchführen ausgehend von der Monom-Basis $\{1, x, x^2, \dots\}$. Diese Orthogonalpolynome besitzen glücklicherweise nur einfache, reelle Nullstellen im Intervall (a, b) , die als Knoten x_0, \dots, x_n einer Quadraturformel verwendbar sind. Der nächste Satz faßt diese Aussagen zusammen, später folgt ein Überblick über einige wichtige Klassen von Orthogonalpolynomen.

Satz 1.3.1 *Die Stützstellen x_i , $i = 0, \dots, n$, der Quadraturformel (1.1.1), (1.1.2) seien die Nullstellen des Orthogonalpolynoms vom Grad $n + 1$ zur Gewichtsfunktion $g \in C[a, b]$ mit $g > 0$ in (a, b) , es gelte also $\omega_{n+1} \in \Pi_{n+1}$, $\omega_{n+1} \perp_g \Pi_n$. Dann besitzt diese Gaußsche Quadraturformel die Ordnung $2n + 2$, d.h. es ist $R_n(f) = 0 \quad \forall f \in \Pi_{2n+1}$. Für Integranden $f \in C^{2n+2}[a, b]$ gilt mit $\xi \in (a, b)$ die Fehleraussage*

$$R_n(f) = \int_a^b f(x)g(x)dx - \sum_{i=0}^n \alpha_i f(x_i) = \frac{1}{(2n+2)!} \int_a^b \omega_{n+1}^2(x)g(x)dx f^{(2n+2)}(\xi). \quad (1.3.5)$$

Bemerkung: Gegenüber den Newton-Cotes-Formeln erhält man also ungefähr die doppelte Konvergenzordnung ohne Mehraufwand. Bei iterierter Anwendung gleicht sich dieser Vorteil aber teilweise aus, da mit Gaußformeln bei Verdopplung der Intervallzahl alte Funktionswerte nicht weiter verwendet werden können. Als Bestandteile anderer numerischer Verfahren, z.B. der Finite-Element-Methode aus §3.3, können Gaußformeln aber sehr effizient sein.

Beweis Hier ist nur noch zu zeigen, dass das Restglied die von (1.1.6) abweichende Form hat. Analog zum Beweis bei der Simpsonregel (1.2.3) werden formal die zusätzlichen Interpolationsdaten $f'(x_0), \dots, f'(x_n)$ benutzt. Das Interpolationspolynom $p_{2n+1} \in \Pi_{2n+1}$ zu den Hermite-Bedingungen $p_{2n+1}(x_i) = f(x_i)$, $p'_{2n+1}(x_i) = f'(x_i)$, $i = 0, \dots, n$, hat im Vergleich zum einfachen Polynom $p_n \in \Pi_n$ mit $p_n(x_i) = f(x_i)$, $i = 0, \dots, n$, in der Newton-Darstellung die Gestalt

$$\begin{aligned} p_{2n+1}(x) &= p_n(x) + \underbrace{\omega_{n+1}(x)}_{f \dots = 0} f[x_0, \dots, x_n, x_0] + \underbrace{\omega_{n+1}(x)(x-x_0)}_{f \dots = 0} f[x_0, \dots, x_n, x_0, x_1] + \\ &\quad \dots + \underbrace{\omega_{n+1}(x)(x-x_0) \cdots (x-x_{n-1})}_{f \dots = 0} f[x_0, \dots, x_n, x_0, \dots, x_n]. \end{aligned}$$

Außerdem gilt nach (2.1.15/Numerik-1) für dessen Fehler

$$f(x) = p_{2n+1}(x) + \omega_{n+1}^2(x) f[x_0, \dots, x_n, x_0, \dots, x_n, x]. \quad (1.3.6)$$

Aufgrund der Orthogonalität (1.3.4) fallen im Integral aber die markierten Zusatzterme weg:

$$\int_a^b \omega_{n+1}(x)(x-x_0)\cdots(x-x_k)g(x)dx = 0 \quad \forall k < n \quad \Rightarrow$$

$$\int_a^b p_{2n+1}(x)g(x)dx = \int_a^b p_n(x)g(x)dx = \sum_{i=0}^n \alpha_i f(x_i). \quad (1.3.7)$$

Aus (1.3.6) folgt die Aussage (1.3.5), da die Fehlerformel (1.1.6) wegen $\omega_{n+1}^2 \geq 0$ jetzt auf p_{2n+1} angewendet werden kann. ■

Bemerkung: In (1.3.7) zeigt sich, dass die Gewichte β_i in der zu p_{2n+1} gehörigen erweiterten Quadraturformel $\int_a^b p_{2n+1}(x)g(x)dx = \sum_{i=0}^n [\alpha_i f(x_i) + \beta_i f'(x_i)]$ alle verschwinden: $\beta_i \equiv 0$.

Je nach Art der im Integranden abgespaltenen Singularität (\rightarrow Gewichtsfunktion) erhält man verschiedene Polynomfamilien (vgl. folgende Tabelle). Analog zu den Tschebyscheff-Polynomen aus der Numerik I gelten auch für diese jeweils Drei-Term-Rekursionen. Die wichtigste Familie zur Gewichtsfunktion $g \equiv 1$ ist die der Legendrepolynome. Zur Vereinfachung wird meist das Standardintervall $[-1, 1]$, bzw. bei uneigentlichen Integralen $[0, \infty)$, zugrundegelegt. Andere Intervalle werden durch Variablensubstitution darauf zurückgeführt.

Polynom-Orthogonalfamilien nach Satz 1.3.1:

Intervall	Gewichtsfunkt.	Bezeichnung	Rekursionsformel
$[-1, 1]$	$g \equiv 1$	Legendre-Pol.	$P_{n+1} = \frac{2n+1}{n+1}xP_n - \frac{n}{n+1}P_{n-1}$, $P_0 = 1$, $P_1 = x$
$[-1, 1]$	$g(x) = \sqrt{1-x^2}$	Tscheby. 2.Art	$U_{n+1} = 2xU_n - U_{n-1}$, $U_0 = 1$, $U_1 = 2x$
$(-1, 1)$	$g(x) = \frac{1}{\sqrt{1-x^2}}$	Tscheby. 1.Art	$T_{n+1} = 2xT_n - T_{n-1}$, $T_0 = 1$, $T_1 = x$ Gewichte konstant, $\alpha_j = \frac{\pi}{n}$
$[0, \infty)$	$g(x) = e^{-x}$	Laguerre-Pol.	$L_{n+1} = \frac{2n+1-x}{n+1}L_n - \frac{n}{n+1}L_{n-1}$ $L_0 = 1$, $L_1 = 1-x$
$(-\infty, \infty)$	$g(x) = e^{-x^2}$	Hermite-Pol.	$H_{n+1} = 2xH_n - 2nH_{n-1}$, $H_0 = 1$, $H_1 = 2x$

Die angegebenen Polynomfamilien haben eine erhebliche, weitergehende Bedeutung v.a. im Zusammenhang mit Reihenentwicklungen für Lösungen von Differentialgleichungen. Daher gibt es eine umfangreiche Literatur (z.B. Courant-Hilbert), Quadratur-Knoten und -Gewichte sind tabelliert, die Berechnung ist bei Stoer, §3.5, besprochen.

Eine explizite Darstellung von Orthogonalpolynomen ist oft mit Hilfe von *Rodriguez-Formeln* möglich. Die Legendre-Polynome, z.B., haben, bis auf Konstanten, die Gestalt

$$P_n(x) = \frac{n!}{(2n)!} \frac{d^n}{dx^n} (x^2 - 1)^n, \quad n = 0, 1, \dots \quad (1.3.8)$$

Durch partielle Integration kann man mit dieser Formel sofort die Orthogonalitätsbedingungen (1.3.2) verifizieren. Aus der Darstellung (1.3.8) folgt übrigens direkt die Existenz von n einfachen reellen Nullstellen in $(-1, 1)$ nach dem Satz von Rolle.

Zum Vergleich der verschiedenen Quadraturverfahren dient folgendes

Beispiel 1.3.2 $\int_{-1}^1 e^x dx = e^1 - e^{-1} = 2 \sinh 1 = 2.3504024$. Näherungen durch

a) Trapezregel, Ordnung 2, 2 Schritte, 3 Punkte:

$$\frac{b-a}{4}[f(-1) + 2f(0) + f(1)] = 2 \cosh^2 \frac{1}{2} = 2.54308\dots, \quad \text{Fehler} = 0.193\dots$$

b) Simpsonregel, Ordnung 4, 3 Punkte:

$$\frac{b-a}{6}[f(-1) + 4f(0) + f(1)] = \frac{1}{3}[4 + 2 \cosh 1] = 2.36205\dots, \quad \text{Fehler} = 0.012\dots$$

c) Gaußquadratur mit 2 Punkten, Ordnung 4. Bestimmung der Orthogonalpolynome: $\omega_0 = P_0 \equiv 1$, $\omega_1(x) = P_1(x) = x \perp \omega_0$, Ansatz $\omega_2 = x^2 + ax + b \perp \Pi_1$, d.h.

$$\alpha) \quad 0 \stackrel{!}{=} \int_{-1}^1 [x^2 + ax + b] dx = \left[\frac{x^3}{3} + a \frac{x^2}{2} + bx \right]_{-1}^1 = \frac{2}{3} + 2b \Rightarrow b = -\frac{1}{3},$$

$$\beta) \quad 0 \stackrel{!}{=} \int_{-1}^1 x[x^2 + ax + b] dx = \left[\frac{x^4}{4} + a \frac{x^3}{3} + b \frac{x^2}{2} \right]_{-1}^1 = \frac{2}{3}a \Rightarrow a = 0.$$

Dies ergibt i.w. das Legendrepolynom $\omega_2(x) = x^2 - \frac{1}{3} = \frac{2}{3}P_2(x)$. Dieses besitzt die Nullstellen $x_0 = -\frac{1}{\sqrt{3}}$, $x_1 = \frac{1}{\sqrt{3}} = 0.57735\dots$. Aus Symmetriegründen ist $\alpha_0 = \alpha_1 = 1$. Die Gaußnäherung hat daher den Wert

$$f(x_0) + f(x_1) = 2 \cosh(x_1) = 2.342696\dots \quad \text{Fehler} = -0.0077\dots$$

1.4 Adaptive Integration

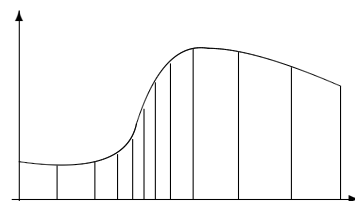
Die Verfahren sind in der besprochenen Form zu unhandlich für den praktischen Einsatz, da bei einem gegebenem Integranden das Quadraturverfahren und dessen Parameter (Schrittweite h , Ordnung) zu wählen sind, um den Integralwert mit einer bestimmten Genauigkeit bzw. Toleranz ε möglichst effizient zu berechnen. Falls man den Integranden analytisch kennt, kann im Prinzip aus der Restglieddarstellung (1.1.4) die Ordnung der Formel und die Zahl der Teilintervalle so bestimmt werden, dass die Fehlerschranke den Wert ε nicht überschreitet (vgl. Übungen).

Diese Methode ist für den Anwender aufwändig und man umgeht ihn durch Verbindung der Quadraturformel mit einer Fehler(ab)schätzung. Bei der iterierten Trapezregel etwa gilt bei einer Zerlegung in Teilintervalle $[x_{i-1}, x_i]$ in jedem Teil mit einer Zwischenstelle $\xi_i \in (x_{i-1}, x_i)$:

$$\int_{x_{i-1}}^{x_i} f(x) dx - \frac{1}{2}(x_i - x_{i-1})[f(x_{i-1}) + f(x_i)] = -\frac{1}{12}(x_i - x_{i-1})^3 f''(\xi_i) =: R[x_{i-1}, x_i]. \quad (1.4.1)$$

Der Fehler zu einem einzelnen Teilintervall hängt also nur von der aktuellen Schrittweite $x_i - x_{i-1}$ und dem *lokalen* Wert der zweiten Ableitung zusammen.

Eine effiziente Strategie bei der Quadratur hält die vorgegebene Toleranz mit möglichst wenigen Funktionsauswertungen ein. Im Bereich kleiner Ableitungen kann dazu mit großen Schrittweiten vorgegangen werden, während bei großen Ableitungswerten feinere Unterteilungen nötig sind. Diese erreicht man mit Teil-



Intervallen $[x_{i-1}, x_i]$, $i = 1, \dots, n$, durch *proportionale Gleichverteilung* des Fehlers

$$\left| R[x_{i-1}, x_i] \right| \stackrel{!}{\leq} \frac{x_i - x_{i-1}}{b - a} \varepsilon, \quad i = 1, \dots, n. \quad (1.4.2)$$

Dann gilt nämlich

$$\left| \int_a^b f(x) dx - \sum_{i=1}^n \frac{x_i - x_{i-1}}{2} [f(x_{i-1}) + f(x_i)] \right| \leq \sum_{i=1}^n |R[x_{i-1}, x_i]| \leq \sum_{i=1}^n \frac{x_i - x_{i-1}}{b - a} \varepsilon = \varepsilon. \quad (1.4.3)$$

Zur Anwendung der Strategie (1.4.2) gibt es zwei Anforderungen:

- (Ab-) Schätzung des lokalen Fehlers $R[x_{i-1}, x_i]$.
- Konstruktion einer Unterteilung, die (1.4.2) erfüllt.

Zu a): Außer bei einfachen Integranden, bei denen $\|f''\|_{[x_{i-1}, x_i]}$ explizit abgeschätzt werden kann (evtl. mit Intervallrechnung), begnügt man sich mit einer *Schätzung* des lokalen Fehlers $R[x_{i-1}, x_i]$. So gilt, z.B., mit $h_i := x_i - x_{i-1}$ für ein $\zeta_i \in (x_{i-1}, x_i)$ die Aussage

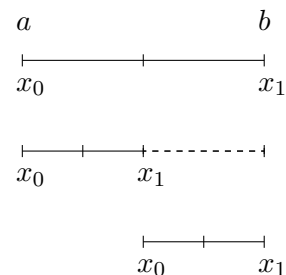
$$f(x_{i-1}) - 2f\left(\frac{x_{i-1} + x_i}{2}\right) + f(x_i) = \frac{1}{4} h_i^2 f''(\zeta_i) \cong -\frac{3}{h_i} R[x_{i-1}, x_i]. \quad (1.4.4)$$

Dies folgt aus den Eigenschaften dividierter Differenzen oder durch Taylorentwicklung um $\frac{1}{2}(x_{i-1} + x_i)$. Mit der Approximation (1.4.4) wird das Kriterium (1.4.2) *ersetzt* durch

$$\left| f(x_{i-1}) - 2f\left(\frac{x_{i-1} + x_i}{2}\right) + f(x_i) \right| \stackrel{!}{\leq} \frac{3\varepsilon}{b - a}. \quad (1.4.5)$$

Zu b): Ein Gitter mit Eigenschaft (1.4.5) konstruiert man adaptiv, etwa in folgender Weise:

- wähle $x_0 := a$, $x_1 := b$;
- Falls (1.4.5) verletzt ist, halbiere das Intervall, setze $x_1 := (x_0 + x_1)/2$;
- Falls (1.4.5) gilt, akzeptiere den Integralwert für $\int_{x_0}^{x_1}$ und fahre wie oben fort mit der Integration im Restintervall $\int_{x_1}^b$, d.h. mit $x_0 := x_1$, $x_1 := b$.



Algorithmus 1.4.1 Einfache adaptive Quadratur

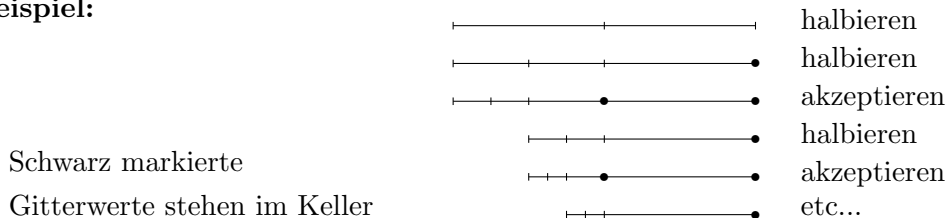
```

x0 := a; f0 := f(x0); x1 := b; f1 := f(x1); w := 0; e := 3ε/(b - a);
wiederhole
  xm := 0.5 * (x0 + x1); fm := f(xm);
  falls abs(f0 - 2 * fm + f1) <= e dann { //akzeptieren
    w := w + 0.5 * (x1 - x0) * (f0 + f1);
    x0 := x1; f0 := f1; x1 := b; f1 := f(x1)}
  sonst//halbieren
    { x1 := xm; f1 := fm }
bis x0 >= b;

```

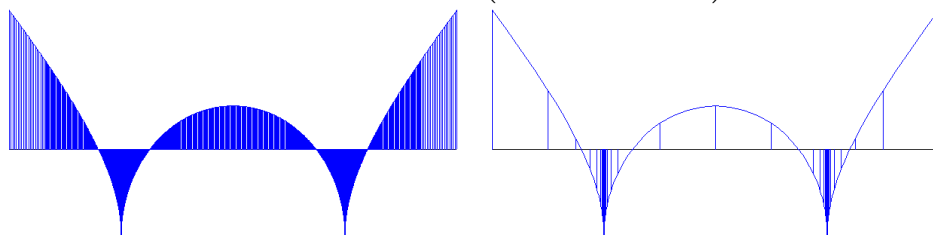
In der Praxis sind Funktionsauswertungen zu kostbar, um sie im Falle eines nicht akzeptierten Teilintegrals zu verwerfen. Wenn man die Teilintervallängen auf feste Bruchteile $(b-a)2^{-j}$, $j \in \mathbb{N}$, einschränkt und durch einfache Zusatzbedingungen kann man erreichen, dass alle einmal *berechneten* Funktionswerte später auch *verwendet* werden. Die Verwaltung der gespeicherten Funktionswerte läßt sich elegant in einem *Kellerspeicher* (Stapel, stack, FILO) realisieren.

Beispiel:



Professionelle Programme steuern lokal außer der Schrittweite auch die Ordnung der Formel.

Demo-Beispiel: Adaptive Integration von $\int_{-1}^1 \left(\sqrt{|x^2 - 0.25|} - \frac{1}{3} \right) dx$.



Bei Toleranz $\varepsilon = 10^{-6}$ benötigt die Trapezregel (linke Graphik) für das Beispiel 2837 Funktionswerte, während die Simpsonregel (rechts) mit nur 241 Werten auskommt. Beide unterschätzen hier aber den exakten Fehler, der bei der Trapezregel 1.2ε und bei Simpson 5ε ist.

1.5 Richardson-Extrapolation, Romberg-Integration

Schon im letzten Abschnitt wurde die Kenntnis des theoretischen Fehlerverhaltens von Quadraturformeln für $h \rightarrow 0$ praktisch zur Steuerung genutzt. Alternativ ist aber auch eine erhebliche Konvergenzbeschleunigung möglich. Dazu sei noch einmal an die Werte aus Beispiel 1.2.2 erinnert:

Beispiel 1.5.1 Iterierte Trapezregel zu $W := \int_1^2 \frac{1}{x} dx = \ln 2 = 0.6931472\dots$

h	1	1/2	1/4	1/8	1/16
T_h	0.75	0.70833	0.69702	0.694122	0.693391
Fehler	0.0568	0.01518	0.00387	0.000975	0.000244
Fehler-Quotient		3.74	3.92	3.99	3.996

Nach (1.2.6) hat der Fehler die Form $W - T_h = a(h)h^2$. Offensichtlich konvergiert sogar die Quotienten aufeinanderfolgender Fehler für $h \rightarrow 0$,

$$\frac{W - T_h}{W - T_{h/2}} = 4 \frac{a(h)}{a(h/2)} \rightarrow 4.$$

Daher gilt für den Vorfaktor tatsächlich $a(h) = a_1 + o(1)$ und für den Fehler also

$$W - T_h = a_1 h^2 + o(h^2) \quad (1.5.1)$$

mit einer festen Konstante a_1 . Der Term $o(h^k)$, ($h \rightarrow 0$), bezeichnet eine Funktion, die schneller gegen Null geht als h^k . Mit dieser Erkenntnis kann man durch eine Linearkombination aus zwei W -Näherungen den h^2 -Anteil in (1.5.1) eliminieren, sogar ohne Kenntnis der Konstante a_1 . Mit der Linearkombination

$$\widehat{T}_{h/2} := \frac{4}{3}T_{h/2} - \frac{1}{3}T_h = \left(\frac{4}{3} - \frac{1}{3}\right)W + \left(\frac{4}{3}\left(\frac{h}{2}\right)^2 - \frac{1}{3}h^2\right)a_1 + o(h^2) = W + o(h^2) \quad (1.5.2)$$

erhält man im Beispiel bessere Näherungen, die ungefähr die doppelte Anzahl von gültigen Ziffern wie bei der Trapezregel aufweisen.

$h =$	1	1/2	1/4	1/8	1/16
$\widehat{T}_h =$		0.69444	0.6932	0.69316	0.693147

Der Erfolg des Vorgehens beruht auf folgender Eigenschaft.

Definition 1.5.2 Die Größe T_h , $0 < h \in \mathbb{R}$, besitzt eine asymptotische Entwicklung (nach Potenzen von h^k), wenn $q \in \mathbb{N}$ und von h unabhängige Koeffizienten $a_j \in \mathbb{R}$ existieren mit

$$T_h = W + a_1 h^k + a_2 h^{2k} + \dots + a_q h^{qk} + \mathcal{O}(h^{qk+1}), \quad h \rightarrow 0. \quad (1.5.3)$$

Der folgende Satz zeigt diese Eigenschaft (1.5.3) für die Trapezregel mit $k = 2$.

Satz 1.5.3 (Euler-McLaurin-Summenformel) Für $f \in C^{2q+2}[a, b]$ besitzt die iterierte Trapezregel (1.2.6) folgende h^2 -Entwicklung mit $h = (b - a)/n$ und $\xi \in (a, b)$,

$$\begin{aligned} T_h &= \frac{h}{2}[f_0 + 2f_1 + \dots + 2f_{n-1} + f_n] = \int_a^b f(x) dx \\ &\quad + \sum_{j=1}^q b_j \left(f^{(2j-1)}(b) - f^{(2j-1)}(a) \right) h^{2j} + c_{2q+2} h^{2q+2} f^{(2q+2)}(\xi). \end{aligned} \quad (1.5.4)$$

Bemerkung: Ein wichtige Folge von (1.5.4) ist, dass bei einem $(b - a)$ -periodischen Integranden die Trapezregel beliebig hohe Ordnung (hier $2q + 2$) besitzt.

Beweis (-Idee, vgl. Stoer) Auf einem einzelnen Intervall, z.B. $[0, h]$, gilt

$$\frac{h}{2}[f(0) + f(h)] = \left[\left(x - \frac{h}{2}\right) f(x) \right]_0^h = \int_0^h \left[\left(x - \frac{h}{2}\right) f(x) \right]' dx = \int_0^h f(x) dx + \int_0^h \left(x - \frac{h}{2}\right) f'(x) dx.$$

Bei partieller Integration lassen sich die Integrationskonstanten hier so wählen, dass Terme mit ungeraden h -Potenzen nicht auftreten. Im ersten Schritt etwa ist $p_2(x) := \frac{1}{2}(x^2 - hx + h^2/6)$ eine Stammfunktion von $(x - \frac{h}{2})$ mit $\int_0^h p_2(x) dx = 0$. Daher hat p_2 eine Stammfunktion $p_3(x) = \int_0^x p_2(t) dt$ mit $p_3(0) = p_3(h) = 0$ und es gilt

$$\frac{h}{2}[f(0) + f(h)] = \int_0^h f(x) dx + \frac{h^2}{12}[f'(h) - f'(0)] + \int_0^h p_3(x) f'''(x) dx,$$

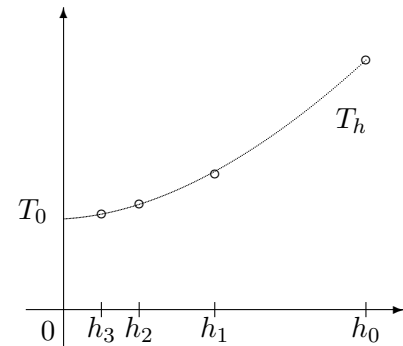
also $b_1 = 1/12$. Das Gesamtergebnis (1.5.4) folgt durch Summation über die Teilintervalle. ■

Den Effekt der Linearkombination (1.5.2) auf die Entwicklung (1.5.3) kann man bei der Trapezregel einfach nachvollziehen, die Approximation $\widehat{T}_{h/2}$ hat jetzt Ordnung 4:

$$\begin{aligned}\widehat{T}_{h/2} &= \frac{4}{3}T_{h/2} - \frac{1}{3}T_h = W + a_1h^2\left(\frac{4}{3}\frac{1}{4} - \frac{1}{3}\right) + a_2h^4\left(\frac{4}{3}\frac{1}{16} - \frac{1}{3}\right) + \dots \\ &= W - \frac{1}{4}a_2h^4 + \dots\end{aligned}$$

Der analytische Hintergrund ist recht einfach, $\widehat{T}_{h/2}$ ist nämlich das lineare Interpolationspolynom $p_1(s)$ zu den Stützstellen $s_0 = h^2$, $s_1 = h^2/4$, welches an der bei Quadratur nicht realisierbaren Stelle $s = 0$ ausgewertet wird. Diese Interpolation eliminiert den h^2 -Fehleranteil in der Entwicklung und ergibt Ordnung 4. Die Methode ist daher ein Extrapolationsverfahren.

Extrapolationsverfahren arbeiten nach folgendem Prinzip. Der Grenzwert $W = \lim_{h \searrow 0} T_h =: T_0$ soll bestimmt werden, wobei die Größe T_h aber nur für (einzelne) positive Werte $h > 0$ tatsächlich berechnet werden kann. Anstatt nun die Elemente einer Folge von Werten T_{h_0}, T_{h_1}, \dots jeweils einzeln zu betrachten, ist es günstiger, ein Interpolationspolynom zu allen Daten $(h_0, T_{h_0}), \dots, (h_m, T_{h_m})$ zu berechnen und dessen Wert in $h = 0$ als Näherung für $T_0 = W$ heranzuziehen. Zur Auswertung des Polynoms in der Stelle 0 eignet sich am besten der *Neville-Algorithmus* (Num1/2.1.9).



Dazu wird (1.5.3) als Entwicklung nach der Variablen $s := h^k$ aufgefaßt. Wegen der Auswertung in $s = 0$ vereinfacht sich der Neville-Algorithmus, er hängt nur von den Quotienten h_i/h_{i+j} ab. Dies führt auf die folgende *Richardson-Extrapolation* für die Polynomwerte $p_{ij} = p_{ij}(0)$:

$$\begin{aligned}p_{i0} &:= T_{h_i}, \quad i = 0, \dots, m \\ p_{ij} &:= p_{i+1, j-1} + \frac{p_{i+1, j-1} - p_{i, j-1}}{(h_i/h_{i+j})^k - 1}, \quad \begin{cases} i = 0, \dots, m-j \\ j = 1, \dots, m \end{cases} \end{aligned} \quad (1.5.5)$$

Die Berechnung erfolgt wieder in einem Dreieckschema spalten- oder zeilenweise. In der Praxis werden meist feste Schrittweitenfolgen, z.B. $h_i/h_{i+j} = 2^j$, verwendet. Dann benötigt (1.5.5) nur 3 Operationen pro Schritt und bei glatten Integranden f gilt für den Fehler $p_{ij} - W = \mathcal{O}(h_{i+j}^{k(j+1)})$. Die neuen Näherungen haben jetzt tatsächlich die höhere Ordnung $k(j+1)$.

Praktische Durchführung

1. Bei der Trapezregel werden nach einer Schrittweithalbung bei $T_{h/2}$ die in T_h enthaltenen Funktionswerte wiederverwendet, vgl (1.2.8). Bei Extrapolation arbeitet man daher mit folgenden Schrittweitenfolgen.
 - a) *Romberg-Folge*: $h_0 = b - a$, $h_i = 2^{-i}h_0$. Diese Quadratur-Methode heißt *Romberg-*

Verfahren, die Genauigkeit ist

$$p_{ij} = W + h_{i+j}^{2j+2} \bar{a}_{j+1} + \dots \quad (1.5.6)$$

Bei hohen Ordnungen werden dabei aber sehr viele Funktionsauswertungen verwendet, nämlich $2^j + 1$ für p_{0j} . Die Formel der Ordnung 16 etwa benötigt 129 Punkte, während die Gaußformel dafür mit 8 Punkten auskommt!

b) *Bulirsch-Folge*: $h_0 = b - a$ und für $i = 1, 2, \dots$

$$h_i/h_0 = \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{6}, \frac{1}{8}, \frac{1}{12}, \dots = \begin{cases} 2^{-(i+1)/2}, & i \text{ ungerade,} \\ \frac{2}{3} 2^{-i/2}, & i \text{ gerade.} \end{cases}$$

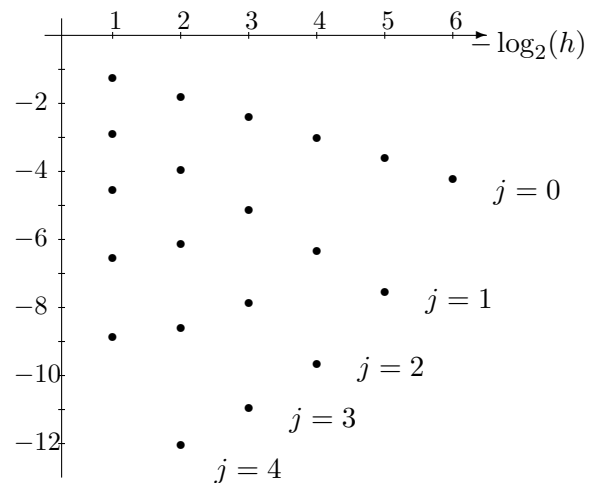
Die Anzahl der für eine bestimmte Ordnung benötigten Stützstellen wächst hier langsamer, Ordnung 16, z.B., erhält man jetzt mit 25 Punkten.

2. Bei keinem Quadraturverfahren kennt man i.d.R. die günstigste Ordnung (Diffbarkeitsordnung von f ?). Die Zulässigkeit der Extrapolation sollte man daher in der Praxis durch Überwachung der Konvergenzaussage (1.5.6) in den Spalten der Extrapolationstabelle prüfen.

Beispiel 1.5.4 $\int_1^2 \frac{1}{x} dx = \ln 2 = 0.69314718056$, Rombergfolge.

h	j=0	1	2	3	4
1/2	0.7083333333	0.69325396825	0.69314790148	0.69314718307	0.69314718056
1/4	0.69702380952	0.69315453065	0.69314719430	0.69314718057	
1/8	0.69412185037	0.69314765282	0.69314718079		
1/16	0.69339120221	0.69314721029			
1/32	0.69320820827				

Die Tabelle enthält die Werte p_{ij} , $i = 0, \dots, 4$, $0 \leq i + j \leq 4$. Der Wert p_{04} ist auf alle 11 Stellen genau. Die Graphik zeigt die Genauigkeit aller Näherungen in den verschiedenen Spalten der Tabelle. An der horizontalen Achse ist der negative Zweier-Logarithmus der Schrittweiten abgetragen, an der vertikalen der Zehner-Logarithmus der Fehler. Die Ordnung der Näherungen in einer Spalte der Tabelle läßt sich an der Steigung der Punktereihen ablesen.



Schlußbemerkung: Die Richardson-Extrapolation beruht allein auf der *Existenz* einer Entwicklung (1.5.3). Solche Entwicklungen existieren auch bei vielen anderen Verfahren, z.B. für Differentialgleichungen. Dort läßt sich die Extrapolation daher analog einsetzen (\rightarrow §2.4).

1.6 Numerische Differentiation

Approximationsformeln für Ableitungen werden auch bei Randwertproblemen für Differentialgleichungen benötigt (\rightarrow §3.2). Ableitungen einer Funktion kann man analog zu den Quadraturformeln über das Interpolationspolynom approximieren, für die k -te Ableitung ist

$$f^{(k)}(\hat{x}) \cong p_n^{(k)}(\hat{x}) = \sum_{j=0}^n f(x_j) L^{(k)}(\hat{x}).$$

Da hier weniger Gestaltungsmöglichkeiten bestehen als bei Quadratur, werden nur kurz einige Beispiele und Besonderheiten beim Fehlerverhalten besprochen. Im folgenden sei $[a, b]$ ein (beliebig kleines) Intervall positiver Länge, das die Stelle \hat{x} enthält.

Satz 1.6.1 *Es sei $p_n \in \Pi_n$ das Interpolationspolynom zu $f \in C^{n+1}[a, b]$ und einfachen Stützstellen $x_j \in [a, b]$, $j = 0, \dots, n$. Dann gilt für $k \leq n$ und $\hat{x} \in [a, b]$ die Aussage*

$$\begin{aligned} p_n^{(k)}(\hat{x}) - f^{(k)}(\hat{x}) &= r_{kn}(\hat{x}) \quad \text{mit} \\ r_{kn}(x) &= (x - x_0^{(k)}) \cdots (x - x_{n-k}^{(k)}) \frac{f^{(n+1)}(\xi_k)}{(n+1-k)!}, \end{aligned} \quad (1.6.1)$$

$\xi_k \in (a, b)$, und $a < x_0^{(k)} < \dots < x_{n-k}^{(k)} < b$. Mit $\varrho := \max_{j=0}^n |\hat{x} - x_j| \leq b - a$ gilt daher

$$|r_{kn}(\hat{x})| \leq \frac{\varrho^{n+1-k}}{(n+1-k)!} \|f^{(n+1)}\|_\infty.$$

Beweis Iterierter Satz von Rolle, vgl. [Stummel-Hainer]

Bemerkung: Nach (1.6.1) gibt es offensichtlich in $[a, b]$ Stellen $x_i^{(k)}$, in denen der Fehler verschwindet. Zur Auswertungsstelle \hat{x} kann man daher geeignete Stützstellen wählen, um eine bessere Approximation zu erhalten, analog zur Quadratur. Zur Untersuchung wird wieder eine zusätzliche Interpolationsstelle x_{n+1} benutzt. In der Newtondarstellung ist dann

$$f(x) = p_n(x) + \omega_{n+1}(x) f[x_0, \dots, x_{n+1}] + r_{n+1}(x),$$

mit $\omega_{n+1}(x) = (x - x_0) \cdots (x - x_n)$. Damit folgt aus dem letzten Satz, dass

$$f^{(k)}(\hat{x}) - p_n^{(k)}(\hat{x}) = \omega_{n+1}^{(k)}(\hat{x}) f[x_0, \dots, x_{n+1}] + r_{k,n+1}(\hat{x}), \quad (1.6.2)$$

mit $|r_{k,n+1}(\hat{x})| \leq \varrho^{n+2-k} \|f^{(n+2)}\|_\infty / (n+2-k)!$. Bei Auswertung in den Nullstellen der Ableitung $\omega_{n+1}^{(k)}$ erhält man also eine um eins erhöhte Konvergenzordnung, wenn $\varrho \rightarrow 0$. Bei Approximationen mit minimaler Knotenzahl, d.h. $n = k$, gilt dazu

$$\omega_{n+1}(x) = x^{n+1} - \left(\sum_{j=0}^n x_j \right) x^n + \dots \quad \Rightarrow \quad \omega_{n+1}^{(n)}(x) = (n+1)! x - n! \sum_{j=0}^n x_j.$$

Diese Ableitung verschwindet also im Schwerpunkt der Knoten

$$\hat{x} = \frac{1}{n+1} \sum_{j=0}^n x_j, \quad (1.6.3)$$

etwa wenn man die Knoten symmetrisch um \hat{x} verteilt. Die höchste Ableitung $p_n^{(n)} \equiv n!f[x_0, \dots, x_n]$ ist konstant und wegen (1.6.2) ist dieser Wert also eine besonders gute Approximation für die Ableitung $f^{(n)}(\hat{x})$ im Punkt (1.6.3). In diesem Fall gilt somit

$$\left| f^{(n)}(\hat{x}) - n!f[x_0, \dots, x_n] \right| \leq \frac{\varrho^2}{2} \|f^{(n+2)}\|_\infty.$$

Bei äquidistanten Knoten $x_j = x_0 + jh$ ist $\varrho \leq nh$, in den einfachsten Fällen erhält man folgende Approximationen (sog. *Differenzenformeln*) für die Stelle $\hat{x} = 0$:

$$\begin{aligned} f'(0) &= \frac{1}{h}(f(h) - f(0)) && + chf''(\xi) \\ f'(0) &= \frac{1}{2h}(f(h) - f(-h)) && + ch^2f^{(3)}(\xi), \quad \text{vgl. (1.6.3)} \\ f''(0) &= \frac{1}{h^2}(f(-h) - 2f(0) + f(h)) && + ch^2f^{(4)}(\xi), \quad \text{vgl. (1.6.3)} \end{aligned}$$

Dabei steht c für i. A. unterschiedliche Fehlerkonstanten. Die Formel für f'' wurde schon bei der adaptiven Quadratur benutzt, (1.4.4). Eine Formel höherer Ordnung ist z.B.

$$f'(0) = \frac{1}{12h}(f(-2h) - 8f(-h) + 8f(h) - f(2h)) + ch^4f^{(5)}(\xi). \quad (1.6.4)$$

Die Differenzenformeln liefern umso genauere Werte, je kleiner die Schrittweite h ist. Allerdings wird dann in der Formel durch die kleinen Zahlen h bzw. h^2 dividiert. Daher wird der bei der Berechnung der Werte $f(x_j)$ im Computer gemachte Rundungsfehler vergrößert und die Verwendung sehr kleiner Schrittweiten ist unsinnig. Da dieser Effekt hier besonders kritisch ist, soll er an der symmetrischen ersten Differenz erläutert werden. Daher seien jetzt $\tilde{f}(x_j) = f(x_j) + \epsilon_j$ die tatsächlich berechneten Funktionswerte, deren Fehler ϵ_j durch eine kleine Konstante \mathcal{E} (ca. Maschinengenauigkeit, z.B. 10^{-15}) beschränkt sind, $|\epsilon_j| \leq \mathcal{E}$. Dann gilt für den Gesamtfehler der tatsächlich *berechneten* Approximation

$$\left| \frac{\tilde{f}(h) - \tilde{f}(-h)}{2h} - f'(0) \right| = \left| ch^2f''(\xi) + \frac{\epsilon_1 - \epsilon_0}{2h} \right| \leq \underbrace{ch^2\|f''\|_\infty}_{\searrow 0} + \underbrace{\frac{\mathcal{E}}{h}}_{\nearrow \infty} \quad (h \rightarrow 0). \quad (1.6.5)$$

Zu dem mit $h \searrow 0$ fallenden Approximationsfehler kommt ein zwar kleiner, aber wachsender Rundungsfehler \mathcal{E}/h . Die Schranke wird minimal bei $\hat{h} \cong \mathcal{E}^{1/3}$ mit einem Minimalwert $\cong \mathcal{E}^{2/3}$. Man kann also bei diesem Beispiel nicht erwarten, eine Genauigkeit von mehr als $\frac{2}{3}$ der im Computer verfügbaren Stellen zu erreichen. Bei Formeln höherer Konvergenzordnung, wie (1.6.4), ist die maximal erreichbare Genauigkeit besser.

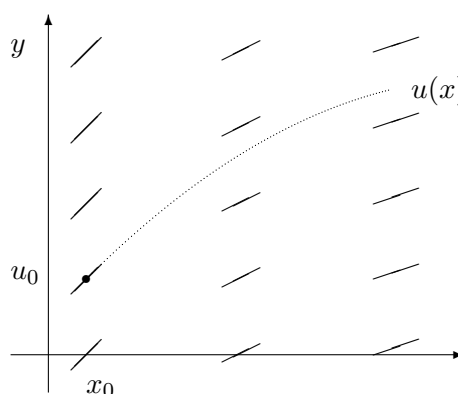
2 Gewöhnliche Differentialgleichungen

2.1 Theoretische Grundlagen

Im allgemeinsten Fall ist eine *gewöhnliche Differentialgleichung* ein System von (nichtlinearen) Gleichungen, in dem verschiedene Ableitungen einer Vektor-Funktion $u : \mathbb{R} \rightarrow \mathbb{R}^n$ einer Variablen miteinander verknüpft sind. Zur Standardisierung betrachtet man aber meist explizite Systeme von Gleichungen erster Ordnung in der Form

$$u'(x) = \frac{du}{dx} = f(x, u(x)), \quad \text{kurz} \quad u' = f(x, u), \quad (2.1.1)$$

wobei $f : \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ eine gegebene stetige Funktion ist. Die Funktion f definiert im Bereich $\mathbb{R} \times \mathbb{R}^n$ ein *Richtungsfeld*, das an jedem Ort (x, u) eine Steigung u' festlegt (vgl. Diagramm). Als *Lösung* einer solchen Differentialgleichung (Dgl) ist eine stetig differenzierbare Vektorfunktion $u : [a, b] \rightarrow \mathbb{R}^n$, also eine Raumkurve gesucht, die für jeden x -Wert aus $[a, b]$ im Punkt $y = u(x)$ den Ableitungswert $u'(x) = f(x, y) = f(x, u(x))$ besitzt, sich also in das gegebene Richtungsfeld einpaßt.



Differentialgleichungen höherer Ordnung können durch eine einfache Substitution auf ein System erster Ordnung umgeformt werden. Die Dgl (2.1.1) alleine definiert durch jeden Punkt $(x, y) \in \mathbb{R} \times \mathbb{R}^n$ eine eigene Lösungskurve. Zur Festlegung einer *eindeutigen* Lösung sind weitere Bedingungen erforderlich. In dieser Vorlesung werden dazu zwei Standard-Aufgabenstellungen behandelt. Beim *Anfangswertproblem* (AWP) wird der Funktionswert an einer Stelle $x_0 \in [a, b]$ vollständig vorgeschrieben (im folgenden oBdA $x_0 = a$)

$$u(a) = u_0 \in \mathbb{R}^n. \quad (2.1.2)$$

Die Lösung ist dann in einem (evtl. unbeschränkten) Intervall $[a, b]$ gesucht. Für dieses Problem gibt es recht allgemeine Existenz- und Eindeutigkeitsaussagen. Sowohl aus theoretischer als auch praktischer Sicht schwieriger ist das *Randwertproblem* (RWP), wo nach Lösungen gesucht wird, für die eine bestimmte Beziehung der Funktionswerte am Anfangs- und Endpunkt eines Intervalls $[a, b]$ gefordert wird, etwa in Form eines zusätzlichen (nichtlinearen) Gleichungssystems

$$r(u(a), u(b)) = 0, \quad r : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^n. \quad (2.1.3)$$

Es folgen einige grundlegende theoretische Aussagen zu den genannten Problemen.

Der Satz von Picard-Lindelöf gibt eine recht einfache und allgemeine Existenzaussage für das Anfangswertproblem (2.1.1),(2.1.2). Er beruht auf dem Übergang von der Dgl durch Integration

zur äquivalenten *Integralgleichung*

$$u(x) = u_0 + \int_a^x f(t, u(t)) dt, \quad (2.1.4)$$

den man auch für die Konstruktion numerischer Verfahren ausnutzt.

Satz 2.1.1 Die Funktion f sei auf dem Streifen $\Omega := \{(x, y) : x \in [a, b], y \in \mathbb{R}^n\}$ über dem endlichen Intervall $[a, b]$ definiert und stetig. Außerdem gelte dort die Lipschitzbedingung

$$\|f(x, y) - f(x, v)\| \leq L\|y - v\| \quad \forall x \in [a, b], y, v \in \mathbb{R}^n, \quad (2.1.5)$$

mit einer Lipschitzkonstanten $L \geq 0$. Dann existiert zu jedem $u_0 \in \mathbb{R}^n$ genau eine Lösung $u \in (C^1[a, b])^n$ für das Anfangswertproblem (2.1.1), (2.1.2). Diese Lösung hängt stetig vom Anfangswert ab. Für zwei Lösungen $u, v \in (C^1[a, b])^n$ mit

$$\begin{aligned} u' &= f(x, u), & u(x_0) &= u_0, \\ v' &= f(x, v), & v(x_0) &= v_0, \end{aligned}$$

gilt nämlich die Schranke

$$\|u(x) - v(x)\| \leq e^{L|x-a|} \|u_0 - v_0\| \quad \forall x \in [a, b]. \quad (2.1.6)$$

Beweis a) Bei (2.1.4) handelt es sich um ein Fixpunktproblem $u = Tu$ mit der Abbildung

$$T : \begin{cases} (C[a, b])^n \rightarrow (C[a, b])^n, \\ v \mapsto Tv, \end{cases} \quad \text{mit } (Tv)(x) := u_0 + \int_a^x f(t, v(t)) dt.$$

Für $y \in (C[a, b])^n$ wird die gewichtete Norm

$$\|y\|_L := \max\{\|y(x)\|e^{-L|x-a|} : x \in [a, b]\}$$

eingeführt, mit der $(C[a, b])^n$ ein Banachraum wird. Aus Voraussetzung (2.1.5) läßt sich direkt die Kontraktivität von T ableiten, denn für beliebige $y, v \in (C[a, b])^n$ gilt

$$\begin{aligned} \|Ty - Tv\|_L &= \max_{x \in [a, b]} e^{-L|x-a|} \left\| \int_a^x (f(t, y(t)) - f(t, v(t))) dt \right\| \\ &\leq \max_{x \in [a, b]} e^{-L|x-a|} \int_a^x \|f(t, y(t)) - f(t, v(t))\| dt \\ &\leq L \max_{x \in [a, b]} e^{-L|x-a|} \int_a^x e^{L|t-a|} \underbrace{e^{-L|t-a|} \|y(t) - v(t)\|}_{\|y - v\|_L} dt \\ &\leq L \max_{x \in [a, b]} e^{-L|x-a|} \int_a^x e^{L|t-a|} dt \|y - v\|_L \\ &\leq (1 - e^{-L|b-a|}) \|y - v\|_L. \end{aligned}$$

Da der Vorfaktor $q = 1 - e^{-L|b-a|}$ kleiner als eins ist wegen $b - a < \infty$, ist T eine Kontraktion, $\|Ty - Tv\|_L \leq q\|y - v\|_L$, und nach dem Banachschen Fixpunktsatz existiert daher ein eindeutiger

Fixpunkt $u = Tu \in (C[a, b])^n$. Mit u ist auch die Funktion $f(x, u(x)) = u'(x)$ stetig, also $u \in (C^1[a, b])^n$.

b) Aus der Differenz der beiden für u und v geltenden Gleichungen (2.1.4) folgt wie eben

$$\begin{aligned} \|u(x) - v(x)\| &= \|u_0 - v_0 + \int_a^x (f(t, u(t)) - f(t, v(t))) dt\| \\ &\leq \|u_0 - v_0\| + L \int_a^x \|u(t) - v(t)\| dt. \end{aligned}$$

Für die skalare Funktion $\delta(x) := \|u(x) - v(x)\|$ gilt also die Integral-Ungleichung $\delta(x) \leq \|u_0 - v_0\| + L \int_a^x \delta(t) dt$. Mit dem folgenden Gronwall-Lemma folgt daraus die Schranke (2.1.6). ■

Im Beweis wurde die folgende Aussage benutzt.

Lemma 2.1.2 (Gronwall-Lemma) Die Funktion $\delta \in C[a, b]$ erfülle die Integralungleichung

$$\delta(x) \leq \alpha + L \int_a^x \delta(t) dt \quad \forall x \in [a, b],$$

mit $\alpha, L \geq 0$. Dann folgt

$$\delta(x) \leq \alpha e^{L(x-a)} \quad \forall x \in [a, b].$$

Beweis Es sei $\varepsilon > 0$. Für die Funktion $\gamma(x) := (\alpha + \varepsilon)e^{L(x-a)}$ gilt $\gamma(x) = \alpha + \varepsilon + L \int_a^x \gamma(t) dt$. Offensichtlich ist $\delta(a) < \gamma(a)$. Sei nun $a < x_1 \leq b$ die erste Stelle mit $\delta(x_1) = \gamma(x_1)$, insbesondere gelte also $\delta(x) \leq \gamma(x) \forall a \leq x \leq x_1$. Dann folgt in x_1 mit

$$\delta(x_1) - \gamma(x_1) \leq L \int_a^{x_1} \delta(t) dt - \varepsilon + L \int_a^{x_1} \gamma(t) dt = -\varepsilon + L \int_a^{x_1} \underbrace{(\delta(t) - \gamma(t))}_{\leq 0} dt < 0$$

aber ein Widerspruch. Daher gilt $\delta(x) < (\alpha + \varepsilon)e^{L(x-a)}$ für jedes $\varepsilon > 0$. ■

Besonders einfach ist das Studium von linearen Differentialgleichungen,

$$u'(x) = A(x)u(x) + g(x), \quad \text{d.h., } f(x, y) = A(x)y + g(x), \quad (2.1.7)$$

mit einer stetigen Matrixfunktion $A(x) \in \mathbb{R}^{n \times n}$ und einer Vektorfunktion $g(x) \in \mathbb{R}^n$, da man hier explizite Lösungsdarstellungen besitzt. Zunächst sieht man, dass die Differenz $y = u - v$ von zwei Lösungen u, v der linearen Dgl wegen

$$u'(x) - v'(x) = A(x)u(x) + g(x) - A(x)v(x) - g(x) = A(x)(u(x) - v(x))$$

die *homogene* Dgl $y' = A(x)y$ erfüllt und dass Linearkombinationen von Lösungen der homogenen Dgl wieder Lösungen sind. Daher definiert die Dgl (2.1.7) alleine einen affinen Lösungsraum der Dimension n . Eine Basis des linearen Raums der homogenen Lösungen faßt man zusammen zu einem Fundamentalsystem der Dgl (2.1.7). Dies ist eine reguläre Matrixfunktion $W(x) \in \mathbb{R}^{n \times n}$, die die Matrix - Dgl

$$W'(x) = A(x)W(x)$$

erfüllt. Mit der Methode der *Variation der Konstanten* läßt sich die eindeutige Lösung des Anfangswertproblems (2.1.7), (2.1.2) damit explizit angeben:

$$u(x) = W(x)W(a)^{-1}u_0 + W(x) \int_a^x W(t)^{-1}g(t) dt.$$

Da mit $W(x)$ auch $W(x)M$ (M regulär) ein Fundamentalsystem ist, vereinfacht sich diese Darstellung durch Übergang zum ausgezeichneten Fundamentalsystem $Y(x) := W(x)W(a)^{-1}$ mit Anfangsbedingung $Y(a) = I$.

Satz 2.1.3 Die Lösung des linearen Anfangswertproblems (2.1.7), (2.1.2) mit einer beschränkten, stetigen Matrixfunktion $A(x)$ ist

$$u(x) = Y(x)u_0 + Y(x) \int_a^x Y(t)^{-1}g(t) dt, \quad (2.1.8)$$

wobei $Y(x)$ das spezielle Fundamentalsystem mit $Y' = AY$, $Y(a) = I$, ist.

Für eine konstante Matrix A ist das Fundamentalsystem sogar explizit bekannt. Der Ansatz $u(t) = e^{\lambda x}y$ führt bei der homogenen Dgl auf das Eigenwertproblem $(A - \lambda I)y = 0$. Bei einer diagonalisierbaren Matrix A mit Eigen-Wert/-Vektor-Paaren (λ_j, v_j) ist die Lösung $u(x) = \sum_{j=1}^n \alpha_j e^{\lambda_j x} v_j$. Dies ist aber gerade die Eigenvektorzerlegung für $u(x) = Y(x)u_0$ mit der Matrix-Exponentialfunktion als Fundamentalsystem. Denn wenn $V = (v_1, \dots, v_n)$ die Matrix der Eigenvektoren ist und $A = V\Lambda V^{-1}$ die Jordan-Normalform, dann gilt

$$Y(x) = e^{(x-a)A} := \sum_{k=0}^{\infty} \frac{(x-a)^k}{k!} A^k = V e^{(x-a)\Lambda} V^{-1}.$$

Bei diagonalisierbarer Matrix A ist $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$. Dies vereinfacht die Lösungsformel (2.1.8) weiter zu

$$u(x) = e^{(x-a)A}u_0 + \int_a^x e^{(x-t)A}g(t) dt.$$

Es gibt eine direkte Beziehung dieser expliziten Formeln zu Satz 2.1.1. Bei (2.1.7) gilt die Lipschitzbedingung (2.1.5) mit $L = \max_x \|A(x)\|$, aus der für die Differenz von Lösungen folgt

$$\|u(x) - v(x)\| = \|Y(x)(u_0 - v_0)\| \leq e^{L|x-a|}\|u_0 - v_0\|,$$

also $\|Y(x)\| \leq e^{L|x-a|}$. Im Fall konstanter Koeffizienten reduziert sich dies auf die Ungleichung

$$\|e^{tA}\| \leq e^{t\|A\|}, \quad t \geq 0, \quad (2.1.9)$$

die auch direkt aus der Reihendarstellung folgt. Im Hinblick auf §2.5 sei an dieser Stelle aber erwähnt, dass diese Ungleichung (2.1.9) sehr unrealistisch sein kann.

Mit Hilfe der expliziten Lösungsdarstellungen soll kurz die Situation bei linearen *Randwertproblemen* diskutiert werden. Gesucht ist eine Lösung von

$$u'(x) = A(x)u(x) + g(x), \quad x \in (a, b), \quad R_0u(a) + R_1u(b) = d, \quad (2.1.10)$$

mit festen $n \times n$ -Matrizen R_0, R_1 und $d \in \mathbb{R}^n$. Dabei ist $\text{rang}(R_0, R_1) = n$ vorauszusetzen. Im Gegensatz zum Anfangswertproblem können hier alle Lösbarkeitsfälle auftreten. Die Lösung u wird durch die explizite Formel (2.1.8) mit einem unbekanntem Startvektor $\eta \in \mathbb{R}^n$ dargestellt,

$$u(x) = Y(x)\eta + \gamma(x), \quad \gamma(x) := \int_a^x Y(x)Y(t)^{-1}g(t) dt.$$

Nach Einsetzen in die Randbedingung bekommt man für η das lineare Gleichungssystem

$$\left(R_0 Y(a) + R_1 Y(b) \right) \eta = d - R_1 \gamma(b) \quad (2.1.11)$$

mit $Y(a) = I$. Prinzipiell gilt also hier die Fallunterscheidung: wenn

$$R_0 + R_1 Y(b) \quad \left\{ \begin{array}{l} \text{regulär:} \quad \text{es existiert eine eindeutige Lösung,} \\ \text{singulär:} \quad \left\{ \begin{array}{l} \text{es gibt keine Lösung,} \\ \text{es gibt unendlich viele Lösungen.} \end{array} \right. \end{array} \right.$$

Die Situation hängt also vom Zusammenspiel zwischen den Randbedingungen (R_0, R_1) und dem Fundamentalsystem $Y(b)$ ab, beim Randwertproblemen gibt es also keine zum Satz 2.1.1 vergleichbare allgemeine Existenz- oder Eindeutigkeitsaussage.

2.2 Einschrittverfahren für Anfangswertprobleme

2.2.1 Herleitung

Die Lösung der Dgl $u' = f(x, u)$ ist diejenige Kurve $(x, u(x))$, die dem Richtungsfeld $(x, y, f(x, y))$ folgt. Als einfaches Approximationsverfahren kann man vom Startwert $u_0 = u(x_0)$ "ein Stück weit" in Richtung der Ableitung $u'(x_0) = f(x_0, u_0)$ gehen und dann dort dieses Verfahren wiederholen (Eulerscher Polygonzug, s.u.). Einen allgemeineren Zugang zu einem solchen schrittweisen Vorgehen bekommt man über die Integralgleichung (2.1.4), allerdings auf kürzeren Intervallen. Dazu wählt man im Punkt \bar{x} eine *Schrittweite* $h > 0$ und integriert die Dgl von \bar{x} bis $\bar{x} + h$. Dies führt auf

$$\frac{1}{h} [u(\bar{x} + h) - u(\bar{x})] = \frac{1}{h} \int_{\bar{x}}^{\bar{x}+h} f(t, u(t)) dt = \int_0^1 f(\bar{x} + hs, u(\bar{x} + hs)) ds. \quad (2.2.1)$$

Für eine numerische Lösung müssen auf der rechten Seite sowohl das Integral (durch eine Quadraturformel) als auch die unbekannte Lösung $u(\bar{x} + hs)$ im Integranden approximiert werden. In diesem Abschnitt werden dabei Näherungen der Form

$$\int_0^1 f(\bar{x} + hs, u(\bar{x} + hs)) ds = \underbrace{f_h(\bar{x}, u(\bar{x}))}_{\text{Näherung}} + \underbrace{T_h(\bar{x})}_{\text{Fehler}}. \quad (2.2.2)$$

betrachtet, in denen die *Verfahrensfunktion* f_h nur vom Wert $(\bar{x}, u(\bar{x}))$ abhängt. Der *lokale Fehler* T_h wird im Verfahren vernachlässigt, bei der Analyse aber genau untersucht, und die

Differentialgleichung (2.1.1) durch eine *Differenzgleichung* ersetzt. Mit dem Startpunkt (x_0, u_0) berechnet man eine Näherung zunächst an der Stelle $x_0 + h$ und dann an weiteren Punkten. Näherungswerte $u_\Delta(x)$ für $u(x)$ erhält man so aus der Formel

$$\frac{1}{h}[u_\Delta(\bar{x} + h) - u_\Delta(\bar{x})] := f_h(\bar{x}, u_\Delta(\bar{x}))$$

nur an endlich vielen Stellen x_i , also auf einem *Gitter* Δ .

Definition 2.2.1 Ein Einschrittverfahren zur Lösung des Anfangswertproblems (2.1.1), (2.1.2) besteht aus der Wahl einer Verfahrensfunktion $f_h(x, y)$ und eines Gitters

$$\Delta: \quad a := x_0 < x_1 < \dots < x_N = b, \quad h_i := x_{i+1} - x_i, \quad H := \max_{i=0}^{N-1} h_i, \quad |\Delta| := H.$$

Damit berechnet sich die Näherungslösung $u_\Delta(x_i) = y_i$, $i = 0, \dots, N$ auf dem Gitter schrittweise

$$\begin{aligned} y_0 &:= u_0, \\ y_{i+1} &:= y_i + h_i f_h(x_i, y_i), \quad i = 0, \dots, N-1. \end{aligned} \quad (2.2.3)$$

Die Bezeichnung Einschrittverfahren bezieht sich auf die Tatsache, dass in (2.2.3) nur Näherungswerte y_i, y_{i+1} aus einem Schritt $x_i \rightarrow x_{i+1}$ verknüpft werden. Es folgen vier einfache Beispiele.

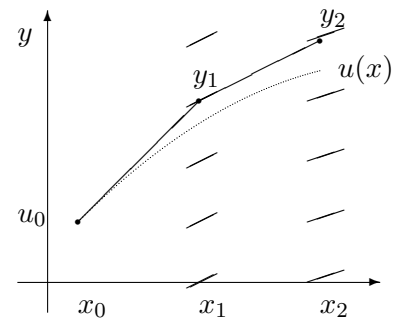
Verfahren 1, Euler-Cauchy-Polygonzug: Das Integral in (2.2.2) wird bei

$$\int_0^1 f(\bar{x} + hs, u(\bar{x} + hs)) ds \cong f(\bar{x}, u(\bar{x})) =: f_h(\bar{x}, u(\bar{x})),$$

durch den Funktionswert im bekannten linken Randwert y_i ersetzt, das Verfahren lautet also

$$y_{i+1} := y_i + h_i f(x_i, y_i), \quad i = 0, 1, \dots$$

Klartext: "In x_i geht man einen Schritt der Länge h_i in Richtung des Richtungsfeldes"



Verfahren 2 von Runge: Integralapproximation durch die Rechteckregel,

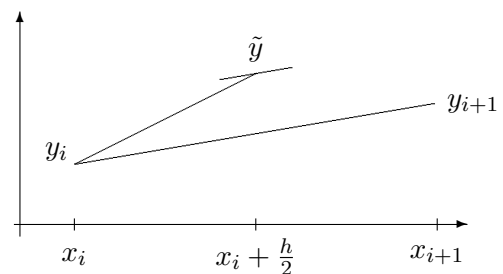
$$\int_0^1 f(\bar{x} + hs, u(\bar{x} + hs)) ds \cong f(\bar{x} + \frac{h}{2}, u(\bar{x} + \frac{h}{2})).$$

Der unbekannte Wert $u(\bar{x} + \frac{h}{2})$ darin wird durch einen Euler-Schritt (Verf. 1) angenähert,

$$u(\bar{x} + \frac{h}{2}) \cong u(\bar{x}) + \frac{h}{2} f(\bar{x}, u(\bar{x})) =: \tilde{y},$$

die Verfahrensfunktion ist hier also

$$f_h(x, y) := f(x + \frac{h}{2}, y + \frac{h}{2} f(x, y)).$$



Verfahren 3 von Heun: Integralapproximation durch die Trapezregel,

$$\int_0^1 f(\bar{x} + hs, u(\bar{x} + hs)) ds \cong \frac{1}{2}[f(\bar{x}, u(\bar{x})) + f(\bar{x} + h, u(\bar{x} + h))].$$

Auch hier wird der unbekannte Wert $u(\bar{x} + h)$ durch Verfahren 1 angenähert, die Verfahrensfunktion ist

$$f_h(x, y) := \frac{1}{2} [f(x, y) + f(x + h, y + hf(x, y))].$$

Verfahren 4, Klassisches Runge-Kutta-Verfahren: Integralapproximation durch eine modifizierte Simpsonregel, bei der der mittlere Wert zweimal approximiert wird:

$$\begin{aligned} f_h(x, y) &:= \frac{1}{6}[k_1 + 2k_2 + 2k_3 + k_4], & \text{mit} \\ k_1 &:= f(x, y) & \cong u'(x) \\ k_2 &:= f(x + \frac{h}{2}, y + \frac{h}{2}k_1) & \cong u'(x + \frac{h}{2}) \\ k_3 &:= f(x + \frac{h}{2}, y + \frac{h}{2}k_2) & \cong u'(x + \frac{h}{2}) \\ k_4 &:= f(x + h, y + hk_3) & \cong u'(x + h) \end{aligned}$$

Beispiel 2.2.2 Die Exponentialfunktion $u(x) = e^x$ löst das AWP $u' = u =: f(x, u)$, $u(0) = 1$. Mit Schrittweite $h > 0$ liefern die behandelten Verfahren folgende Näherungen:

Verfahren	Näherung	lokal.Fehler
1	$y_1 = y_0 + hy_0 = 1 + h$	$h^2/2$
2	$\tilde{y}_{1/2} = 1 + \frac{h}{2}, y_1 = 1 + h(1 + \frac{h}{2}) = 1 + h + \frac{1}{2}h^2$	$h^3/6$
3	$\tilde{y}_1 = 1 + h, y_1 = 1 + \frac{h}{2}(1 + 1 + h) = 1 + h + \frac{1}{2}h^2$	$h^3/6$
4	$k_1 = 1, k_2 = 1 + \frac{h}{2}, k_3 = 1 + \frac{h}{2} + \frac{h^2}{4}, k_4 = 1 + h + \frac{h^2}{2} + \frac{h^3}{4} \Rightarrow y_1 = 1 + h + \frac{1}{2}h^2 + \frac{1}{6}h^3 + \frac{1}{24}h^4$	$h^5/120$

Also approximieren die Verfahren die spezielle Lösung e^x dadurch, dass sie hier unterschiedliche lange Anfangsteile der Exponentialreihe für $u(h) = e^h$ erzeugen. Die Verfahren sind Spezialfälle einer allgemeinen Verfahrensklasse, der *Runge-Kutta-Verfahren*. Diese bestehen aus $m \in \mathbb{N}$ *Stufen* und können durch eine Tabelle von Koeffizienten (Butcher-Tableau) charakterisiert werden,

$$\begin{array}{c|c} c & A \\ \hline & \mathbf{b} \end{array} = \begin{array}{c|ccc} c_1 & a_{11} & \cdots & a_{1m} \\ \vdots & \vdots & & \vdots \\ c_m & a_{m1} & \cdots & a_{mm} \\ \hline & b_1 & \cdots & b_m \end{array}$$

wobei $c_i = \sum_{j=1}^m a_{ij}$, $i = 1, \dots, m$ gilt. Dazu gehört die Verfahrensfunktion

$$f_h(x, y) := \sum_{i=1}^m b_i k_i, \tag{2.2.4}$$

wobei man die Stufen aus folgenden Stufengleichungen berechnet:

$$k_i := f(x + hc_i, y + h \sum_{j=1}^m a_{ij}k_j), \quad i = 1, \dots, m.$$

Wenn die Koeffizientenmatrix A strikt untere Dreiecksgestalt besitzt, also für $a_{ij} = 0 \forall j \geq i$, berechnet man die Hilfsgrößen k_i der Reihe nach explizit aus diesen Gleichungen, man hat ein *explizites Runge-Kutta-Verfahren*. Die Verfahren 1–4 gehören zu dieser Klasse, zum klassischen Runge-Kutta-Verfahren 4 etwa gehört die Tabelle

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

Bei allgemeiner Koeffizientenmatrix ($\exists a_{ij} \neq 0, j \geq i$) ist (2.2.4) aber ein *implizites Runge-Kutta-Verfahren*, zur Berechnung der k_i sind (i.a. nichtlineare) Gleichungssysteme zu lösen. Dieses wird man normalerweise vermeiden. Solche Verfahren können aber sehr gute Stabilitäts-Eigenschaften besitzen und werden daher in schwierigen Situationen verwendet.

Bei den numerischen Verfahren (2.2.3) will man natürlich **Konvergenz** der Näherungswerte bei feiner werdendem Gitter haben. Bei der Konvergenz betrachtet man Näherungen zu immer kleineren Schrittweiten an einer festen Stelle \bar{x} . Dazu wurde die diskrete Näherungsfunktion u_Δ eingeführt, Konvergenz bedeutet dann einfach $u_\Delta(\bar{x}) \rightarrow u(\bar{x}), |\Delta| \rightarrow 0$. Man beachte aber, dass bei Gitter-Näherungen y_k zu verschiedenen Schrittweiten an der gleichen Stelle \bar{x} unterschiedliche Indizes k gehören, hier ist die Verwendung der Gitterfunktion u_Δ günstiger. Der Nachweis der Konvergenz erfolgt nach einem sehr allgemeinen Prinzip, das auch für viele andere Verfahren anwendbar ist, zweistufig in der Form

$$\boxed{\text{Konsistenz} + \text{Stabilität} \implies \text{Konvergenz}}$$

2.2.2 Konsistenz

Eine Verfahrensfunktion f_h sollte sinnvollerweise so gewählt sein, dass die exakten Lösungen u der Differentialgleichung auch die Differenzgleichung (2.2.3) näherungsweise erfüllen.

Definition 2.2.3 Das Verfahren (2.2.3) heißt konsistent, wenn mit jeder Lösung u der Dgl (2.1.1) bei stetig differenzierbarer rechter Seite f für den lokalen Fehler

$$T_h(x) := \frac{1}{h}(u(x+h) - u(x)) - f_h(x, u(x)), \quad x \in [a, b-h], \quad (2.2.5)$$

gilt $T_h \rightarrow 0$ für $h \rightarrow 0$. Das Verfahren heißt konsistent mit Ordnung $p > 0$, wenn

$$T_h(x) = \mathcal{O}(h^p), \quad h \rightarrow 0, \quad x \in [a, b], \quad (2.2.6)$$

gilt für genügend oft differenzierbare rechte Seiten f .

Beispiel 2.2.4 Verfahren 2 mit $f_h = f(x + \frac{h}{2}, y + \frac{h}{2}f(x, y))$ Die Taylorentwicklung einer genügend glatten Lösung u in x **nach h (!)** ist

$$\frac{1}{h}(u(x+h) - u(x)) = u'(x) + \frac{h}{2}u''(x) + \mathcal{O}(h^2). \quad (2.2.7)$$

Für $f \in C^2$ erhält man die Ableitungen $u^{(j)}(x)$ der Lösung aus der Dgl $u' = f(x, u) \Rightarrow u'' = f_x + f_y u' = f_x + f_y f$ durch die Kettenregel (alle Funktionen werden in $(x, u(x))$ ausgewertet). Bei allen expliziten Verfahren gilt für die erste Stufe mit Startwert $u(x)$ dass $k_1 = f(x, u(x)) = u'(x)$ ist. Taylorentwicklung von $k_2 = f_h$ nach h liefert

$$f_h(x, u(x)) = f\left(x + \frac{h}{2}, u + \frac{h}{2}f\right) = \underbrace{f(x, u(x))}_{=u'(x)} + \frac{h}{2} \underbrace{(f_x + f_y f)}_{=u''(x)} + \mathcal{O}(h^2).$$

Somit stimmt die Taylorentwicklung von f_h nach h mit (2.2.7) in den ersten beiden Gliedern überein, Verfahren 2 hat Konsistenzordnung $p = 2$.

Die Taylorentwicklung aus dem Beispiel ist die Standardmethode für Konsistenznachweise. Diese sind also einfach durchführbar, aber aufwändig. Umgekehrt können durch Vergleich der Taylorentwicklungen von u und f_h im allgemeinen Ansatz (2.2.4) die erforderlichen Bedingungen an die Koeffizienten hergeleitet und damit geeignete Verfahren konstruiert werden. Für ein explizites dreistufiges Verfahren ist $c_1 = 0$ und für Ordnung drei, z.B., sind folgende *Ordnungsbedingungen* zu erfüllen:

$$\begin{aligned} b_1 + b_2 + b_3 &= 1 \\ b_2 c_2 + b_3 c_3 &= \frac{1}{2} \\ b_2 c_2^2 + b_3 c_3^2 &= \frac{1}{3} \\ b_3 a_{32} c_2 &= \frac{1}{6} \end{aligned} \quad (2.2.8)$$

Offensichtlich handelt es sich hier um ein nichtlineares Gleichungssystem. Dabei entsprechen die ersten drei Bedingungen denen an die Quadraturformel mit Stützstellen c_i und Gewichten b_i (Parabeln werden exakt integriert). Da das Verfahren 6 freie Parameter besitzt, gibt es also eine zweiparametrische Schar von dreistufigen Verfahren mit Ordnung drei. Analoges gilt bei 4-stufigen Verfahren (8 Bedingungen für 10 Parameter). Keines dieser Verfahren hat aber eine höhere Ordnung als drei bzw. vier. Die Zahl der Ordnungsbedingungen wächst zu höheren Ordnungen stark an, sodass dann (im Unterschied zu den Quadraturformeln) $m > p$ Stufen für Ordnung p erforderlich werden. Die mit einer bestimmten Stufenzahl m erreichbare Ordnung p ist in der folgenden Tabelle aufgeführt:

$m =$	1	2	3	4	5	6	7	8	9	10	11	..	17
$p =$	1	2	3	4	4	5	6	6	7	7	8	..	10

Die speziellen Verfahren 1–4 besitzen jeweils die höchsten erreichbaren Ordnungen $p = m$. Die höchste bisher durch explizite Konstruktion erreichte Ordnung ist $p = 10$ mit $m = 17$ Stufen. Später werden zwar Einschrittverfahren beliebig hoher Ordnung auftreten, doch werden bei diesen noch sehr viel höhere Stufenzahlen benutzt.

Zum Nachweis der Konvergenz eines Verfahrens (d.h. der zugehörigen Näherungen) reicht die Konsistenz nicht aus. Das Verfahren muß eine zweite zentrale Eigenschaft erfüllen.

2.2.3 Stabilität

Prinzipiell bedeutet Stabilität, dass sich die von einem Verfahren berechneten Näherungen bei kleinen Störungen nur wenig (Lipschitz-stetig) ändern. Beim diskreten Anfangswertproblem (2.2.3) bietet sich wegen der einfachen Struktur eine etwas speziellere Formulierung an.

Definition 2.2.5 Ein Einschrittverfahren (2.2.3) heißt stabil, wenn Konstanten S, r, H_0 existieren so, dass für je zwei gestörte Lösungen $\{y_i\}, \{z_i\}$ mit

$$\left. \begin{aligned} \frac{1}{h_i}(y_{i+1} - y_i) - f_{h_i}(x_i, y_i) &= \eta_i \\ \frac{1}{h_i}(z_{i+1} - z_i) - f_{h_i}(x_i, z_i) &= \zeta_i \end{aligned} \right\} i = 0, 1, \dots, N-1, \quad (2.2.9)$$

gilt

$$\max_{i=0}^N \|y_i - z_i\| \leq S \left(\|y_0 - z_0\| + \max_{i=0}^{N-1} \|\eta_i - \zeta_i\| \right) \quad (2.2.10)$$

für alle Gitter Δ mit $|\Delta| \leq H_0$ und solange $\|y_0 - z_0\| \leq r$, $\|\eta_i - \zeta_i\| \leq r$, $i = 0, \dots, N-1$.

Im Unterschied zu den später behandelten Mehrschrittverfahren ist die Stabilität von Einschrittverfahren unter einfachen Voraussetzungen nachweisbar. Dabei und auch in Bezug auf die verwendeten Beweistechniken liegen starke Ähnlichkeiten zu den Differentialgleichungen aus §2.1 vor. Umgekehrt kann übrigens auch der Existenzsatz 2.1.1 mit Hilfe des Euler-Cauchy-Verfahrens bewiesen werden. Als Beweishilfsmittel im Stabilitätssatz benötigt man ein diskretes Analogon zum Gronwall-Lemma (2.1.2).

Lemma 2.2.6 Für Werte $\delta_i \in \mathbb{R}_+$ gelte mit $\alpha, \ell, h_i \geq 0$ rekursiv

$$\delta_{i+1} \leq (1 + h_i \ell) \delta_i + h_i \alpha, \quad i = 0, \dots, N-1.$$

Dann folgt (mit $x_i := x_0 + h_0 + \dots + h_{i-1}$) die explizite Schranke

$$\delta_i \leq e^{\ell(x_i - x_0)} \left(\delta_0 + (x_i - x_0) \alpha \right), \quad i = 0, \dots, N.$$

Beweis Da $1 \leq 1 + \ell h_i \leq e^{\ell h_i}$ ist und δ_0 die Behauptung erfüllt, folgt induktiv

$$\begin{aligned} \delta_{i+1} &\leq e^{\ell h_i} \delta_i + h_i \alpha \leq e^{\ell h_i} \left(e^{\ell(x_i - x_0)} (\delta_0 + (x_i - x_0) \alpha) \right) + h_i \alpha \\ &\leq e^{\ell(x_{i+1} - x_0)} (\delta_0 + (x_i - x_0) \alpha + h_i \alpha). \quad \blacksquare \end{aligned}$$

Beim Stabilitätsbeweis von Einschrittverfahren wird als wesentliche Voraussetzung wieder eine Lipschitzbedingung benötigt.

Satz 2.2.7 (Stabilitätssatz) Die Verfahrensfunktion f_h des Einschrittverfahrens (2.2.3) erfülle in dem in Satz 2.1.1 definierten Streifen Ω eine Lipschitzbedingung

$$\|f_h(x, y) - f_h(x, z)\| \leq \ell \|y - z\| \quad \forall x \in [a, b], y, z \in \mathbb{R}^n. \quad (2.2.11)$$

Dann ist das Verfahren stabil, es gilt (2.2.10) mit der (von Δ unabhängigen) Konstanten $S = e^{(b-a)\ell} \max\{1, b-a\}$ und $H_0 = b-a, r = \infty$.

Beweis Die Subtraktion der beiden Gleichungen in (2.2.9) ergibt

$$y_{i+1} - z_{i+1} = y_i - z_i + h_i (f_{h_i}(x_i, y_i) - f_{h_i}(x_i, z_i)) + h_i(\eta_i - \zeta_i).$$

Mit der Lipschitzbedingung (2.2.11) folgt daraus die Ungleichungskette

$$\|y_{i+1} - z_{i+1}\| \leq \|y_i - z_i\| + h_i \ell \|y_i - z_i\| + h_i \|\eta_i - \zeta_i\| = (1 + h_i \ell) \|y_i - z_i\| + h_i \|\eta_i - \zeta_i\|.$$

Für die Differenzen $\delta_i := \|y_i - z_i\|$ entspricht dies den Voraussetzungen von Lemma 2.2.6 mit $\alpha = \max_j \|\eta_j - \zeta_j\|$ und liefert mit $(x_i - x_0)e^{\ell(x_i - x_0)} \leq (b-a)e^{\ell(b-a)}$ die Behauptung. ■

Eine Lipschitzbedingung für die Verfahrensfunktion f_h folgt bei expliziten Runge-Kutta-Verfahren aus derjenigen von f selber, eine Lipschitzkonstante ℓ läßt sich sogar explizit angeben. Beim Euler-Cauchy-Verfahren ist dies wegen $f_h = f$ trivial, beim Verfahren 2 gilt Folgendes.

Beispiel 2.2.8 Verfahren 2, $f_h(x, y) = f(y + \frac{h}{2}f(y))$ die Abhängigkeit von x ist unwesentlich. Aus der Annahme $\|f(y) - f(z)\| \leq L\|y - z\|$ folgt (2.2.11) mit $\ell = L + \frac{h}{2}L^2$, denn

$$\begin{aligned} \|f_h(y) - f_h(z)\| &= \|f(y + \frac{h}{2}f(y)) - f(z + \frac{h}{2}f(z))\| \\ &\leq L\|y + \frac{h}{2}f(y) - z - \frac{h}{2}f(z)\| \leq L(\|y - z\| + \frac{h}{2}L\|y - z\|). \end{aligned}$$

2.2.4 Konvergenz

Die Doppelbezeichnung der Näherungen mit $y_i, u_\Delta(x_i)$ wurde eingeführt, um einerseits mit den y_i die Schreibweise im Verfahren zu vereinfachen, andererseits aber mit u_Δ auch die Konvergenzaussage klarer formulieren zu können. Konvergenz der Näherungslösungen kann natürlich nur bei feiner werdendem Gitter $\Delta, |\Delta| \rightarrow 0$, also auch $N \rightarrow \infty$ erwartet werden. Bezogen auf einen festen Punkt \bar{x} im Integrationsintervall, z.B. $\bar{x} = b$, betrifft die Aussage $u_\Delta(\bar{x}) \rightarrow u(\bar{x})$ ($|\Delta| \rightarrow 0$) Näherungswerte y_k mit wachsendem Index k auf verschiedenen Gittern.

Satz 2.2.9 Ein konsistentes und stabiles Verfahren ist konvergent:

Das Einschrittverfahren (2.2.3) sei konsistent mit der Ordnung $p > 0$, d.h. es gelte (2.2.6). Das Verfahren sei außerdem stabil, etwa aufgrund der Lipschitzbedingung (2.2.11) für f_h . Dann konvergieren die Näherungswerte $\{y_i = u_\Delta(x_i), x_i \in \Delta\}$ bei feiner werdendem Gitter Δ ($|\Delta| \rightarrow$

0) gegen die Lösung u des AWP (2.1.1), (2.1.2). Die Konvergenzordnung ist mindestens p , d.h. es gibt eine Konstante K so, dass für alle $|\Delta| \leq H_0$ für den globalen Fehler gilt

$$\max_{x \in \Delta} \|u_\Delta(x) - u(x)\| = \max_{i=0}^{|\Delta|} \|y_i - u(x_i)\| \leq K|\Delta|^p. \quad (2.2.12)$$

Beweis Nach Definition (2.2.3) erfüllen die Näherungen y_i die Beziehung

$$\frac{1}{h_i}(y_{i+1} - y_i) - f_{h_i}(x_i, y_i) = 0.$$

Für die Werte $z_i := u(x_i)$ der Lösung der Dgl gilt andererseits die Konsistenzaussage (2.2.6)

$$\frac{1}{h_i}(z_{i+1} - z_i) - f_{h_i}(x_i, z_i) = T_{h_i}(x_i) = \mathcal{O}(h_i^p).$$

Diese Gleichungen entsprechen denen aus der Stabilitätsdefinition, (2.2.9), mit $\eta_i = 0$ und $\zeta_i = T_{h_i}(x_i)$. Wegen $y_0 = u_0$ und $h_i^p \leq |\Delta|^p$ folgt daher die Schranke (2.2.12). ■

Der Beweis des Satzes ist natürlich trivial **nach** geeigneter Begriffsbildung, da zwei Definitionen einfach zusammengefügt werden. Die eigentliche Arbeit liegt im Einzelnachweis von Konsistenz und Stabilität.

Bemerkung: Für stabile Verfahren gilt Konvergenzordnung \geq Konsistenzordnung.

Der Satz zeigt, dass die Näherungslösung u_Δ genügend schnell gegen u konvergiert für $|\Delta| \rightarrow 0$. Wie bei den Quadraturformeln kann man bei Ordnung p und einer Halbierung der Schrittweiten eine Verkleinerung des Fehlers um den Faktor 2^{-p} erwarten. Der Verwendung von Verfahren möglichst hoher Ordnung sind allerdings dadurch Grenzen gesetzt, dass die Fehlerkonstante K in (2.2.12) u.a. von immer höheren Ableitungen $u^{(p)}$ der Lösung abhängt. Die Konstante K ist nicht explizit bekannt und in der Praxis kaum berechenbar, daher kann man aus dieser Fehlerschranke die tatsächliche Größe des Fehlers u_Δ bei gegebenem Gitter nicht ableiten. Genausowenig kann man umgekehrt ein Gitter konstruieren, auf dem ein geforderter Fehler erreicht wird. In der Praxis löst man das Problem wie bei der Quadratur durch eine adaptive Wahl des Gitters Δ .

2.2.5 Schrittweitensteuerung

Bei realen Problemen variiert die Gestalt der Lösungen in unterschiedlichen Bereichen des Integrationsintervalls $[a, b]$ oft sehr stark (Beispiel Satellitenbahn: stark/schwach gekrümmt nahe/entfernt von Himmelskörpern). Ein Verfahren kann hier nur dann effizient (und auch mit akzeptablen Rundungsfehlern, s.u.) arbeiten, wenn sich die Schrittweite lokal an diesen Verlauf anpaßt. Durch eine Inspektion von Lemma 2.2.6 sieht man, dass die Fehlerschranke (2.2.12) aus

$$\max_{x \in \Delta} \|u_\Delta(x) - u(x)\| \leq \tilde{K} \sum_{j=0}^{N-1} h_j \|T_{h_j}\|, \quad \tilde{K} = e^{\ell(b-a)} \quad (2.2.13)$$

hervorgeht. Ein Gesamtfehler der Größenordnung $(b-a)\tilde{K}\varepsilon$ kann man daher durch die Forderung

$$\|T_{h_i}(x_i)\| \stackrel{!}{\leq} \varepsilon \quad \forall i = 0, \dots, N-1 \tag{2.2.14}$$

erreichen, vgl. §1.4 (adaptive Integration). Da der exakte lokale Fehler T_h nicht bekannt ist, arbeitet man wieder mit einer Schätzung und nutzt das bekannte Verhalten $T_h \doteq \gamma h^p$ mit der Forderung (2.2.14) als Richtlinie zur lokalen, fortlaufenden Schrittweitensteuerung. Hat man nämlich nach einem mit Schrittweite h ausgeführten Integrationsschritt (eine Schätzung für) den aktuellen lokalen Fehler T_h , dann kann man die Fehlerkonstante durch $\gamma \doteq T_h/h^p$ und damit die eigentlich für (2.2.14) erforderliche Schrittweite \hat{h} schätzen über

$$T_{\hat{h}} \doteq \gamma \hat{h}^p \doteq \left(\frac{\hat{h}}{h}\right)^p T_h \stackrel{!}{\leq} \varepsilon \quad \Rightarrow \quad \hat{h} \doteq h \sqrt[p]{\frac{\varepsilon}{T_h}}. \tag{2.2.15}$$

Natürlich nutzt man die Schrittweite \hat{h} aus (2.2.15) nur dann, wenn der aktuelle Fehler zu groß ist, $T_h > \varepsilon$, dann wiederholt man den Schritt ab x_i mit Schrittweite $h_i = \hat{h}$. Bei akzeptablem Fehler $T_h \leq \varepsilon$ verwendet man aber erst im nächsten Intervall \hat{h} als Schätzung für h_{i+1} .

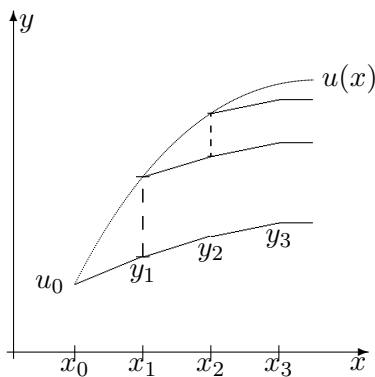
Zur *Schätzung* des lokalen Fehlers T_h können zwei Verfahren verschiedener Ordnung, etwa p und $p+1$, verwendet werden, ihre Verfahrensfunktionen seien f_h und \bar{f}_h . Dazu berechnet man im i -ten Schritt

$$\left. \begin{aligned} y_{i+1} &= y_i + h f_h(x_i, y_i) \\ \bar{y}_{i+1} &= y_i + h \bar{f}_h(x_i, y_i) \end{aligned} \right\} \text{mit lokalem Fehler } \begin{cases} \mathcal{O}(h^p) \\ \mathcal{O}(h^{p+1}) \end{cases}. \tag{2.2.16}$$

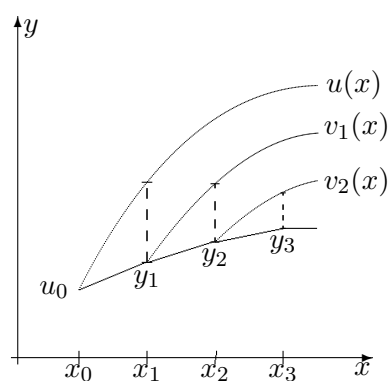
Es sei jetzt $v_i(x)$ diejenige Lösung der Dgl, die die Anfangsbedingung $v_i(x_i) = y_i$ erfüllt. Dann gilt bezüglich dieser Lösung für den lokalen Fehler des Grundverfahrens mit f_h die Beziehung

$$\begin{aligned} T_h(x_i) &= \frac{1}{h}(v_i(x_i+h) - y_i) - f_h(x_i, y_i) = \frac{1}{h}(\bar{y}_{i+1} - y_i) + \mathcal{O}(h^{p+1}) - f_h(x_i, y_i) \\ &= \bar{f}_h(x_i, y_i) - f_h(x_i, y_i) + \mathcal{O}(h^{p+1}). \end{aligned} \tag{2.2.17}$$

Da nur $T_h = \mathcal{O}(h^p)$ gilt, ist also (für genügend kleines h) die Differenz der Verfahrensfunktionen $\bar{f}_h(x_i, y_i) - f_h(x_i, y_i)$ bis auf den $\mathcal{O}(h^{p+1})$ -Term eine Schätzung des lokalen Fehlers T_h .



In Satz 2.2.9: Lösungsfächer Differenzengl. mit $\zeta_j = T_{h_j}(x_j)$, $j = 0, 1, \dots$



In (2.2.17): Lösungsfächer Dgl

Die Unterschiede zwischen der Beweismethode zur Fehlerabschätzung in Satz 2.2.9 und der hier verwendeten zur Fehlerschätzung lassen sich graphisch mit den *Lösungsfächern* von Differentialgleichung (2.1.1) und Differenzgleichung (2.2.3) veranschaulichen. Die gestrichelten Fehlerbeiträge der beiden Ansätze sind dabei vergleichbar, aber nicht identisch.

Aufbauend auf einer Schätzung des lokalen Fehlers wird in Algorithmus 2.2.5 eine einfache Schrittweitensteuerung formuliert, vorzugeben ist dabei nur die Toleranz ε und eine Startschrittweite h_0 .

Algorithmus 2.2.10 Schrittweitensteuerung beim AWP:

$x := a; y := u_0; h = h_0; \varepsilon = \text{Toleranz}$ wiederhole { falls $x + h > b$, dann $h := b - x$; $y_0 := y + hf_h(x, y)$; $y_1 := y + hf_{\bar{h}}(x, y)$; $q := \varepsilon h / \ y_1 - y_0\ $; falls $q > 1$ dann $\{x := x + h; y := y_0\}$ $h := h * \sqrt[p]{q}$; } bis $x \geq b$; 	rechten Rand ansteuern Grundverfahren besseres Verfahren Fehlerquotient Schritt akzeptieren neue Schrittweite
---	--

Sinnvolle Modifikationen dieses Algorithmus sind eine geringfügig kleinere Schrittweite $h := 0.9 * h * \sqrt[p]{q}$ als in (2.2.15), da eine Schrittwiederholung wesentlich teurer wird als ein etwas zu kleiner Schritt. Außerdem ist es ratsam, einen zu großen Wechsel bei den Schrittweiten durch Zusatzabfragen zu verhindern. Zusätzlich wird meist auch mit der genaueren Lösung $y_1 = \bar{y}_{i+1}$ weitergerechnet ($y := y_1$ statt $y := y_0$ nach $x := x + h$), da die lokale Fehlerschätzung nur zur Steuerung verwendet wird. Beim Einsatz ist es schwer, die Toleranz ε für den *lokalen* Fehler T_h zu schätzen, wenn ein bestimmter *globaler* Fehler $\|u_\Delta - u\|$ gewünscht wird, da die Konstante \tilde{K} in (2.2.13) kaum genau bekannt ist.

Zur Konstruktion von Verfahrenspaaren (2.2.16) gibt es zwei übliche Zugänge unterschiedlicher Effizienz und Allgemeinheit:

a) *Eingebettete Runge-Kutta-Verfahren*: Das Entwurfsziel bei Runge-Kutta-Verfahren ist meistens eine möglichst hohe Ordnung bei möglichst wenigen Funktionsauswertungen. Nach diesem Prinzip sucht man Paare von Verfahren (2.2.4) mit m bzw. $m+1$ Stufen, bei denen das zweite die Funktionswerte k_i des ersten *mitverwendet*. Beide verwenden also ein gemeinsames Stufenschema (Matrix A), nur die Linearkombination mit b_i, \bar{b}_i ist verschieden. Die Verfahrensformulierung und das Butcher-Tableau haben die folgende Gestalt.

$$\begin{aligned}
 y_{i+1} &= y_i + h_i \sum_{j=1}^m b_j k_j, \\
 \bar{y}_{i+1} &= y_i + h_i \sum_{j=1}^{m+1} \bar{b}_j k_j, \\
 k_j &= f(x + h_i c_j, y_i + h_i \sum_{l=1}^{j-1} a_{jl} k_l), \quad j = 1, \dots, m+1.
 \end{aligned}$$

c_1	0	...	0	0
\vdots		\ddots		\vdots
c_{m+1}	a_{j1}			0
	b_1	...	b_m	0
	\bar{b}_1	\bar{b}_{m+1}

Die lokale Fehlerschätzung zum i -ten Schritt ist für dieses " $p(p+1)$ -Paar" dann

$$T_{h_i}(x_i) \cong \frac{1}{h_i} (\bar{y}_{i+1} - y_{i+1}) = \sum_{j=1}^m (\bar{b}_j - b_j) k_j + \bar{b}_{m+1} k_{m+1}.$$

Beispiel 2.2.11 Konstruktion einer eingebetteten Methode zu Verfahren 3 mit Hilfe der Ordnungsbedingungen (2.2.8). Hier sind $b_1 = b_2 = \frac{1}{2}$ und $c_2 = a_{21} = 1$ vorgegeben. Für die restlichen Parameter des eingebetteten Verfahrens bleiben die Bedingungen

$$\begin{aligned}
 \bar{b}_1 + \bar{b}_2 + \bar{b}_3 &= 1, \\
 \bar{b}_2 + \bar{b}_3 c_3 &= \frac{1}{2}, \\
 \bar{b}_2 + \bar{b}_3 c_3^2 &= \frac{1}{3}, \\
 \bar{b}_3 a_{32} &= \frac{1}{6}.
 \end{aligned}$$

Die Größen \bar{b}_1, a_{32} treten nur in der ersten bzw. letzten Bedingung auf. Aus der 2. und 3. ergibt sich $\bar{b}_3 c_3 (1 - c_3) = \frac{1}{6}$. Für $c_3 = \frac{1}{2}$ erhält man daraus die Koeffizienten in der folgenden rechten Tabelle. Die linke enthält ein 2(3)-Paar zu Verfahren 2.

2(3)-Paar zu Verf. 2	2(3)-Paar zu Verf. 3
0	0
$\frac{1}{2}$	1
1	$\frac{1}{2}$

Bekannt sind Runge-Kutta-Paare mit

Ordnungen	1(2)	2(3)	3(4)	4(5)	4(5)	..	7(8)
Stufen	(1,2)	(2,3)	(3,5)	(4,6)	(5,6)	..	(11,13)

Bei den 3-stufigen Verfahren gibt es keine eingebetteten 4-stufigen Verfahren. Allerdings fand Fehlberg ein 5-stufiges Paar der Ordnung 3(4), bei dem der letzte Funktionswert $f(x + c_5 h, \dots)$ mit $c_5 = 1$ nur in die Fehlerschätzung eingeht ($b_5 = 0$) und daher als erster Funktionswert im nächsten Intervall, $f(x + h + c_1 h, \dots)$, $c_1 = 0$, wiederverwendet werden kann (*FSAL=first same as last*). Pro Schritt fallen also doch nur 4 Auswertungen an. Auch zum klassischen vierstufigen Runge-Kutta-Verfahren gibt es kein eingebettetes mit 5 Stufen, da überhaupt keine 5-stufigen Verfahren mit Ordnung 5 existieren (s.o.). Somit werden sowieso 6 Stufen benötigt und für

ein 4(5)-Verfahren ist es günstiger, das Stufenpaar (5,6) zu verwenden. Ein sehr effizientes Referenz-Verfahren ist das 5(4)-Paar DOPRI5 von Dormand und Prince mit 6+1 Stufen, bei dem der Fehler des Verfahrens fünfter Ordnung optimiert wurde. Es ist die Grundlage der MATLAB-Routine ode45. Auch dieses verwendet den Fehlberg-Trick, wegen $b^\Gamma = (a_{7j})_j$ stimmt der letzte Funktionswert $k_7 = f(x + h_i, y_{i+1})$ mit dem ersten des nächsten Schritts überein. Von den gleichen Autoren existiert sogar ein 8(5,3)-Tripel DOP853. Das Koeffizienten-Tableau für DOPRI5 ist

0	0								
$\frac{1}{5}$	$\frac{1}{5}$	0							
$\frac{3}{10}$	$\frac{3}{40}$	$\frac{9}{40}$	0						
$\frac{4}{5}$	$\frac{44}{45}$	$-\frac{56}{15}$	$\frac{32}{9}$	0					
$\frac{8}{9}$	$\frac{19372}{6561}$	$-\frac{25360}{2187}$	$\frac{64448}{6561}$	$-\frac{212}{729}$	0				
1	$\frac{9017}{3168}$	$-\frac{355}{33}$	$\frac{46732}{5247}$	$\frac{49}{176}$	$-\frac{5103}{18656}$	0			
1	$\frac{35}{384}$	0	$\frac{500}{1113}$	$\frac{125}{192}$	$-\frac{2187}{6784}$	$\frac{11}{84}$	0		
$b =$	$\frac{35}{384}$	0	$\frac{500}{1113}$	$\frac{125}{192}$	$-\frac{2187}{6784}$	$\frac{11}{84}$	0		
$\bar{b} - b =$	$\frac{71}{57600}$	0		$-\frac{71}{16695}$	$\frac{71}{1920}$	$-\frac{17253}{339200}$	$\frac{22}{525}$	$-\frac{1}{40}$	

In praktischen Anwendungen benötigt man oft die Näherung u_Δ tatsächlich als Funktion von x (Graphiken, Wegintegrale, etc.). Zu diesem Zweck betrachtet man *stetige Erweiterungen* bzw. Einbettungen. Dabei sind die Koeffizienten in (2.2.4) Polynome $\bar{b}_i(t)$ mit $\bar{b}_i(0) = 0, \bar{b}_i(1) = b_i$. Dann wird durch

$$u_\Delta(x_i + th_i) = y_i + h_i \sum_{j=1}^m b_j(t)k_j,$$

wird eine stetige Funktion $u_\Delta(x)$ definiert. Für das Verfahren 3 etwa erhält man mit

$$b_1(t) = t - \frac{1}{2}t^2, \quad b_2(t) = \frac{1}{2}t^2$$

eine stetige Erweiterung $u_\Delta(x)$, die für jedes t Konvergenzordnung 2 besitzt. Meist bekommt man aber für stetige Erweiterungen nicht die volle normale Ordnung der Endformel mit $\bar{b}_i(1)$. Für $m=3$ Stufen ist nur die glm. Ordnung 2 erreichbar, zu DOPRI5 existiert eine Erweiterung mit glm. Ordnung 4.

b) *Richardson-Extrapolation*: Bei vielen Verfahren, z.B. auch den Runge-Kutta-Verfahren, existiert für äquidistante Gitter $\Delta = \{x_i = a + ih\}$ eine asymptotische Entwicklung des Fehlers

$$u_\Delta(x) - u(x) = h^p g_p(x) + \mathcal{O}(h^{p+1}), \quad h \rightarrow 0, \tag{2.2.18}$$

vgl. (1.5.3). Durch Linearkombination von zwei Näherungen zu Schrittweiten h und $h/2$ kann der dominierende Fehleranteil $h^p g_p$ eliminiert (\rightarrow Richardson-Extrapolation) oder aber geschätzt werden, da für eine zweite Lösung $u_{\Delta'}$, zu $\Delta' = \{x_i = a + i\frac{h}{2} : i = 0, \dots, 2N\}$, gilt

$$\begin{aligned} u_{\Delta'}(x) - u(x) &= 2^{-p} h^p g_p(x) + \mathcal{O}(h^{p+1}) \Rightarrow \\ (1 - 2^{-p}) h^p g_p(x) &= u_\Delta(x) - u_{\Delta'}(x) + \mathcal{O}(h^{p+1}). \end{aligned}$$

Daraus kann man eine bessere Approximation $\bar{u}_\Delta(x) := u_{\Delta'}(x) + \frac{1}{2^p-1}[u_{\Delta'}(x) - u_\Delta(x)]$, $x \in \Delta$, oder eben die Fehlerschätzungen

$$u_\Delta - u \doteq \frac{2^p}{2^p-1}[u_\Delta - u_{\Delta'}], \quad u_{\Delta'} - u \doteq \frac{1}{2^p-1}[u_\Delta - u_{\Delta'}],$$

berechnen. Wie in (2.2.17) ergibt sich

$$\begin{aligned} T_h(x_i) &= \frac{1}{h}[v_i(x_i+h) - u_\Delta(x_i+h)] \doteq \frac{1}{h}[\bar{u}_\Delta(x_i+h) - u_\Delta(x_i+h)] \\ &= \frac{1}{1-2^{-p}} \frac{1}{h}[u_{\Delta'}(x_i+h) - u_\Delta(x_i+h)] \doteq -h^{p-1}g_p(x_i+h). \end{aligned} \quad (2.2.19)$$

Diese Größe ist im angegebenen Fall von der Ordnung $\mathcal{O}(h^p)$. Denn da man sich hier in (2.2.18) auf (Näherungs-)Lösungen bezieht mit $u_\Delta(x_i) = v_i(x_i) = y_i$, gilt $g_p(x_i) = 0$ und somit nach dem Mittelwertsatz $h^{p-1}g_p(x_i+h) = \mathcal{O}(h^p)$. Daher ist (2.2.19) eine Fehlerschätzung von der in Algorithmus 2.2.5 verwendeten Gestalt. Im Vergleich zu eingebetteten Verfahren werden hier aber wesentlich mehr zusätzliche Funktionsauswertungen verwendet (bezogen auf die genauere Lösung $u_{\Delta'}$ nach 2 Schritten mit $h/2$ also $m/2$ pro Schritt).

Beispiel 2.2.12 Periodische Satellitenbahn um Erde und Mond: In der Ebene des rotierenden Erde-Mond-System bewegt sich ein Satellit mit den Koordinaten $u(t) \in \mathbb{R}^2$ nach der Dgl

$$u'' = u + \begin{pmatrix} 0 & 2 \\ -2 & 0 \end{pmatrix} u' - \nu \frac{u-E}{\|u-E\|^3} - \mu \frac{u-M}{\|u-M\|^3},$$

wobei $\mu = 1/82.45$, $\nu = 1 - \mu$ die Massenverhältnisse und $E = (-\mu, 0)^\top$ bzw. $M = (\nu, 0)^\top$ die Positionen von Erde und Mond sind. Das rotierende Referenzsystem erzeugt die von u' abhängige *Corioliskraft*. Es gibt eine periodische Bahn, die zweimal nahe an der Erde vorbeiführt und dort sehr kleine Schrittweiten erfordert. Ergebnis bei Toleranz $\varepsilon =_{10} -4$:

	f -Auswert.	Endfehler		
RuKuFe-2(3)	1506	$5.1 \cdot 10^{-3}$		
RuKuFe-4(5)	762	$9.8 \cdot 10^{-3}$		
DoPri-5(4)	798	$2.3 \cdot 10^{-3}$		

Die Graphiken zeigen jeweils logarithmisch den Verlauf der Schrittweiten (oben, grün) und der Fehlerschätzung (darunter, rot) mit der Zeit. Im linken Diagramm sieht man das Ergebnis für RKFB-2(3), das den lokalen Fehler sehr gut unter der Toleranz hält, trotz der beiden starken Schrittweitenreduktionen auf $\cong 10^{-4}$ bei den Erd-Passagen. DoPri-5(4) im rechten Diagramm arbeitet mit größeren Schrittweiten, hat daher wesentlich mehr Schrittweiterholungen (offene Quadrate), benötigt dennoch nur ca die halbe Anzahl von f -Auswertungen.

Schrittweitensteuerung hatte vorrangig das Ziel, die geforderte Genauigkeit der Näherung mit möglichst wenig Aufwand einzuhalten. Zusätzlich reduziert sie aber auch den Einfluß von *Rundungsfehlern*. Bei Maschinenrechnung werden in (2.2.3) gestörte Näherungen \tilde{y}_i berechnet, die pro Schritt anfallenden Rundungsfehler ε_i führen also zu einer verfälschten Rekursion

$$\tilde{y}_{i+1} = \tilde{y}_i + h_i f_{h_i}(x_i, \tilde{y}_i) + \varepsilon_{i+1}, \quad i = 0, \dots, N-1,$$

$\tilde{y}_0 = u_0$. Daher kommen bei praktischer Rechnung zum Diskretisierungsfehler $y_i - u(x_i)$ noch Rundungsfehler hinzu. Unter Voraussetzung (2.2.11) gilt dafür

$$\begin{aligned} \|\tilde{y}_{i+1} - y_{i+1}\| &\leq \|\tilde{y}_i - y_i\| + h_i \|f_{h_i}(x_i, \tilde{y}_i) - f_{h_i}(x_i, y_i)\| + \|\varepsilon_{i+1}\| \\ &\leq (1 + h_i \ell) \|\tilde{y}_i - y_i\| + \hat{\varepsilon}_{i+1}, \quad \hat{\varepsilon}_{i+1} := \|\varepsilon_{i+1}\|. \end{aligned}$$

Analog zu Lemma 2.2.6 ergibt sich mit $\tilde{y}_0 = y_0$ die Abschätzung

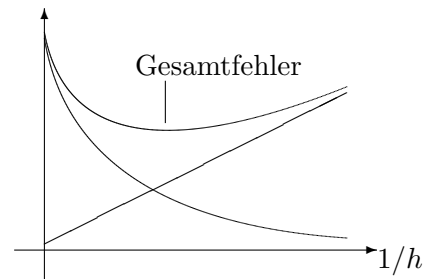
$$\|\tilde{y}_i - y_i\| \leq \sum_{j=1}^i e^{\ell(x_i - x_j)} \hat{\varepsilon}_j \leq S \sum_{j=1}^N \hat{\varepsilon}_j, \quad S := e^{\ell(b-a)}. \quad (2.2.20)$$

Für den Gesamtfehler bei einem äquidistanten Gitter bekommt man somit die Schranke

$$\|\tilde{y}_i - u(x_i)\| \leq S(b-a) \frac{\varepsilon}{h} + Kh^p, \quad \varepsilon := \max_i \hat{\varepsilon}_i.$$

Analog zu (1.6.5) rechnet man nach, dass ein minimaler Gesamtfehler für $h \cong \varepsilon^{1/(p+1)}$ erreicht wird in der Größenordnung $\|\tilde{y}_i - u(x_i)\| = \mathcal{O}(\varepsilon^{p/(p+1)})$ und mit wachsender Ordnung kleiner wird. Allerdings muß bei äquidistantem Gitter die feste Schrittweite h nach der ungünstigsten Stelle im Intervall gewählt werden und erzwingt daher evtl. eine sehr

große Zahl N von Schritten, welche nach (2.2.20) zu entsprechend großen Rundungsfehlern führt. Die Gesamtzahl der Schritte ist bei variablen Gittern i.d.R. viel kleiner, da in leichten Passagen große Schritte und daher insgesamt weniger verwendet werden.



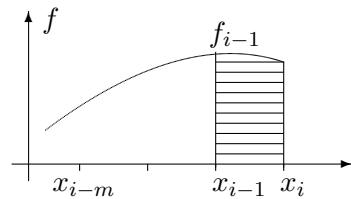
2.3 Mehrschrittverfahren

2.3.1 Adams-Verfahren

Bei Einschrittverfahren benötigt man für Ordnung p mindestens $m \geq p$ Stufen und daher sind zur Approximation des Integrals in (2.2.1),

$$\int_0^1 f(x_{i-1} + ht, u(x_{i-1} + ht)) dt = \frac{1}{h}[u(x_{i-1} + h) - u(x_{i-1})]$$

beim Schritt $x_{i-1} \rightarrow x_i$ mindestens p zusätzliche Funktionswerte $k_j \cong f(x_{i-1} + hc_j, u(x_{i-1} + hc_j))$ neu zu berechnen. Als Alternative kann man stattdessen schon bekannte, aber vor x_{i-1} gelegene Funktionswerte $f(x_{i-j}, y_{i-j}) =: f_{i-j}$, $j = 1, 2, \dots, m$, im Verfahren verwenden. Das Interpolationspolynom $p_{m-1} \in \Pi_{m-1}$ zu den Werten f_{i-1}, \dots, f_{i-m} wird mit den Lagrange polynomen $L_j(x)$ zu den Stützstellen x_{i-1}, \dots, x_{i-m} dargestellt (vgl. Numerik I, §3.1). Durch Integration von p_{m-1} erhält man für obiges Integral die Approximation



$$f_h(x_{i-1}, y_{i-m}, \dots, y_{i-1}) := \sum_{j=-m}^{-1} f_{i+j} \underbrace{\int_0^1 L_j(x_{i-1} + ht) dt}_{=: \beta_{-j}}. \quad (2.3.1)$$

Diese Verfahrensfunktion hängt also von m früheren und bekannten Lösungswerten y_{i-j} ab. Für gleich große Schrittweiten $h_i \equiv h$ erhält man damit die *Methode von Adams-Bashforth* (A-B):

$$\left. \begin{aligned} \frac{1}{h}(y_i - y_{i-1}) &= f_h(x_{i-1}, y_{i-m}, \dots, y_{i-1}) \\ f_h(x_{i-1}, y_{i-m}, \dots, y_{i-1}) &:= \sum_{j=1}^m \beta_j f(x_{i-j}, y_{i-j}) \end{aligned} \right\} i = m, m+1, \dots, N. \quad (2.3.2)$$

Dieses Verfahren ist ein explizites, lineares, m -Schrittverfahren. Pro Schritt benötigt man hier tatsächlich nur *eine* einzige neue Funktionsauswertung, f_{i-1} . Andererseits ist das Verfahren erst ab Schritt m einsetzbar, Startwerte y_1, \dots, y_{m-1} müssen anderweitig berechnet werden. Explizit heißt das Verfahren wieder, da die Verfahrensfunktion f_h nicht vom noch unbekanntem Wert y_i abhängt. Durch Einsetzen der Dgl-Lösung u in f_h erhält man wegen $f(x, u(x)) = u'(x)$ den Konsistenzfehler einfach als Quadraturfehler

$$\frac{1}{h}(u(t_i) - u(t_{i-1})) - \sum_{j=1}^m \beta_j f(x_{i-j}, u(t_{i-j})) = \frac{1}{h}(u(t_i) - u(t_{i-1})) - \sum_{j=1}^m \beta_j u'(t_{i-j}),$$

das Verfahren hat daher Ordnung $p = m$ (mehr s.u.). Die Koeffizienten (2.3.1) der ersten A-B-Verfahren enthält die folgende Tabelle.

m	$j =$	1	2	3	4	Fehler
1	$\beta_j =$	1				$\frac{1}{2}hu''$
2	$2\beta_j =$	3	-1			$\frac{5}{12}h^2u'''$
3	$12\beta_j =$	23	-16	5		$\frac{3}{8}h^3u^{(4)}$
4	$24\beta_j =$	55	-59	37	-9	$\frac{251}{720}h^4u^{(5)}$

Bei der Approximation (2.3.1) werden nur Daten links vom Integrationsintervall $(x_{i-1}, x_i]$ benutzt. Man kann versuchen, die Genauigkeit durch Berücksichtigung des Wertes im Endpunkt x_i zu verbessern, die Zahl m der Teilintervalle würde sich dadurch nicht erhöhen. Man interpoliert also in den Stellen x_{i-m}, \dots, x_i mit einem Polynom vom Grad m . Dies führt auf die *Methode von Adams-Moulton* (A-M)

$$\left. \begin{aligned} \frac{1}{h}(y_i - y_{i-1}) &= f_h(x_{i-1}, y_{i-m}, \dots, y_i) \\ f_h(x_{i-1}, y_{i-m}, \dots, y_i) &:= \sum_{j=0}^m \mu_j f(x_{i-j}, y_{i-j}) \end{aligned} \right\} i = m, m+1, \dots, N. \quad (2.3.3)$$

Dieses ist jetzt aber ein *implizites*, lineares, m -Schritt-Verfahren, denn die Vorschrift (2.3.3) ist für y_i ein (nichtlineares) Gleichungssystem. Auf diese Schwierigkeit wird nach dem nächsten Satz eingegangen. Die Koeffizienten der ersten A-M-Verfahren lauten

m	$j =$	0	1	2	3	Fehler
0	$\mu_j =$	1				$-\frac{1}{2}hu''$
1	$2\mu_j =$	1	1			$-\frac{1}{12}h^2u'''$
2	$12\mu_j =$	5	8	-1		$-\frac{1}{24}h^3u^{(4)}$
3	$24\mu_j =$	9	19	-5	1	$-\frac{19}{720}h^4u^{(5)}$

Die Konsistenz überprüft man bei Mehrschrittverfahren leicht, bei beiden Verfahren folgt sie aus Approximationseigenschaften der Interpolationspolynome beim Integranden $f(x, u(x)) = u'(x)$. Der folgende Satz behandelt beide Verfahren, wobei das A-B-Verfahren als A-M-Verfahren mit Gewicht $\beta_0 = 0$ interpretiert wird.

Satz 2.3.1 *Der lokale Fehler bei den Adams-Bashforth-Verfahren (2.3.2), $\gamma_0 := 0$, $\gamma_j := \beta_j$, $1 \leq j \leq m$, bzw. Adams-Moulton-Verfahren (2.3.3), $\gamma_j := \mu_j$, $0 \leq j \leq m$, ist gegeben durch*

$$T_h(x_{i-1}) := \frac{1}{h}(u(x_i) - u(x_{i-1})) - \sum_{j=0}^m \gamma_j u'(x_{i-j}).$$

Er hat die Form

$$T_h(x_{i-1}) = \begin{cases} c_m^B h^m u^{(m+1)}(x_{i-1}) + \mathcal{O}(h^{m+1}) & \text{bei A-B,} \\ c_m^M h^{m+1} u^{(m+2)}(x_{i-1}) + \mathcal{O}(h^{m+2}) & \text{bei A-M.} \end{cases} \quad (2.3.4)$$

Beweis Beim A-B-Verfahren sei $p \in \Pi_{m-1}$ das Interpolationspolynom zu den Wertepaaren $(x_{i-j}, u'(x_{i-j}))$, $j = 1, \dots, m$. Dann folgt aus der Fehlerformel der Interpolation (Numerik I, 2.1.15) und durch Taylorentwicklung die Aussage

$$\begin{aligned} T_h(x_{i-1}) &= \int_0^1 (u'(x_{i-1} + ht) - p(x_{i-1} + ht)) dt \\ &= \frac{1}{m!} \int_0^1 (x_{i-1} + ht - x_{i-1}) \dots (x_{i-1} + ht - x_{i-m}) u^{(m+1)}(\xi(t)) dt \\ &= h^m \underbrace{\frac{1}{m!} \int_0^1 t(t+1) \dots (t+m-1) dt}_{=: c_m^B} u^{(m+1)}(x_{i-1}) + \mathcal{O}(h^{m+1}). \end{aligned}$$

Bei A-M betrachtet man das Polynom vom Grad m zu $(x_{i-j}, u'(x_{i-j}))$, $j = 0, \dots, m$. ■

Die Fehlerkoeffizienten c_m wurden schon in den obigen Tabellen angegeben. Das A-M-Verfahren mit m Stufen besitzt offensichtlich die gleiche Konvergenzordnung h^{m+1} wie das A-B-Verfahren mit $m + 1$ Stufen, hat aber kleinere Fehlerkoeffizienten $|c_m^M| < |c_{m+1}^B|$, $m > 0$. Daher ist das Adams-Moulton-Verfahren vorzuziehen. Die Auflösung der dabei auftretenden nichtlinearen Gleichung für y_i in (2.3.3) kann durch Kombination mit dem A-B-Verfahren umgangen werden, indem man mit A-B zunächst eine Approximation \tilde{y}_i für den unbekanntten Wert berechnet und diesen in A-M einsetzt. So ergibt sich das *Prädiktor-Korrektor-Verfahren* (P-K)

$$\left. \begin{aligned} \tilde{y}_i &:= y_{i-1} + h \sum_{j=1}^m \beta_j f(x_{i-j}, y_{i-j}) \\ y_i &:= y_{i-1} + h\mu_0 f(x_i, \tilde{y}_i) + h \sum_{j=1}^m \mu_j f(x_{i-j}, y_{i-j}) \end{aligned} \right\} i = m, m+1, \dots \quad (2.3.5)$$

Hier werden pro Schritt zwei Funktionsaufrufe benutzt, $f(x_{i-1}, y_{i-1})$ und $f(x_i, \tilde{y}_i)$. Der lokale Fehler dieses Verfahren ist tatsächlich von der Ordnung h^{m+1} . Denn nach Satz 2.3.1 gilt für Lipschitz-stetiges f mit $\tilde{u}_i := u(x_{i-1}) + h \sum_{j=1}^m \beta_j f(x_{i-j}, u(x_{i-j}))$ hier

$$\begin{aligned} \|T_h^{PK}\| &= \|T_h^{AM} + \mu_0(u'(x_i) - f(x_i, \tilde{u}_i))\| \\ &\leq \|T_h^{AM}\| + |\mu_0|L\|u(x_i) - \tilde{u}_i\| = \|T_h^{AM}\| + |\mu_0|Lh\|T_h^{AB}\| \\ &= h^{m+1}(|c_m^M|\|u^{(m+2)}(x_{i-1})\| + |\mu_0|Lc_m^B\|u^{(m+1)}(x_{i-1})\|) + \mathcal{O}(h^{m+2}). \end{aligned}$$

Bemerkung: (Nur) für $m = 1$ sind die Adams-Verfahren auch Runge-Kutta-Verfahren. Das A-B-1-Verfahren ist gerade das explizite Euler-Verfahren, das einstufige P-K-Verfahren (A-B-M-1) entspricht dem R-K-Verfahren 3 von Heun.

Der Fehler bei Runge-Kutta-Verfahren hat nicht die einfache Form wie bei Mehrschrittverfahren, er hängt nicht nur von Ableitungen $u^{(m)}$ der Lösung ab. Meist ist er aber für ein Verfahren gleicher Ordnung kleiner als bei Mehrschrittverfahren, da die Quadraturformel (2.2.2) näher bei x_i liegende Daten als (2.3.1) heranzieht. Für das Problem $u' = u$, $u(0) = 1$, ergeben sich bei den Verfahren vierter Ordnung nach einem Schritt

Verfahren:	Runge-Kutta	Adams-Bashforth	Adams-Moulton
lokaler Fehler:	$\frac{1}{120}h^5$	$\frac{251}{6} \frac{1}{120}h^5$	$\frac{19}{6} \frac{1}{120}h^5$

Bei den Mehrschrittverfahren sind für gleiche Fehlergrößen also kleinere Schrittweiten erforderlich. Außerdem ist die Programmierung aufwändiger:

Fragen zum praktischen Einsatz:

- *Anlaufrechnung:* Die m -Schritt-Verfahren können, wie aus (2.3.2)-(2.3.5) ersichtlich, erstmals nach einer Anlaufphase von m Integrationsschritten eingesetzt werden. Daher sind Startwerte y_0, \dots, y_{m-1} ausreichender Genauigkeit auf andere Weise bereitzustellen, entweder mit Runge-Kutta-Verfahren oder Mehrschrittverfahren niedriger (wachsender) Ordnung und sehr kleinen Schrittweiten.

- *Schrittweitenwechsel:* Die Adams-Formeln (2.3.2)-(2.3.5) beziehen sich auf äquidistante Punkte x_{i-m}, \dots, x_i in konstantem Abstand h . Die Koeffizienten β_j, μ_j ergaben sich dabei durch Integration der Lagrange-Darstellung des Interpolationspolynoms, bei Adams-Bashforth also von $p(x) = \sum_{j=1}^m L_j(x) f_{i-j}$ in der Form

$$f_h^{AB} = \sum_{j=1}^m \beta_j f_{i-j} = \int_0^1 \sum_{j=1}^m L_j(x_{i-1} + ht) f_{i-j} dt.$$

Wird p dagegen als Newton-Polynom (Numerik I, §3.1) dargestellt, gilt

$$f_h^{AB} = \sum_{j=0}^{m-1} h^j \delta_j f[x_{i-j-1}, \dots, x_{i-1}].$$

Ein Wechsel der Schrittweite $h \rightarrow \hat{h}$ ist in dieser Form jetzt einfach möglich durch Ersetzung der Faktoren $h^j \rightarrow \hat{h}^j$. Die schon berechneten Differenzen $f[\dots, x_{i-1}]$ werden übernommen. In der Praxis sind dabei zusätzliche Vorsichtsmaßnahmen zu treffen.

- *Schrittweiten-, Ordnungs-Steuerung:* Da die A-B-Verfahren Ordnung m und die A-M-Verfahren Ordnung $m+1$ besitzen, sind die beiden Teile des Prädiktor-Korrektor-Verfahrens (2.3.5) als Verfahrenspaar zur Schätzung des lokalen Fehlers einsetzbar. Algorithmus 1.2.10 könnte damit also auch hier direkt eingesetzt werden (vgl. letzte Anmerkung). Darüberhinaus läßt sich aber auch noch die Ordnung des Verfahrens steuern, da es Adams-Methoden beliebiger Ordnung gibt. Eine einfache Strategie hierfür erhält man aus der Regel:

Wurde ein Schritt in $[x_{i-1}, x_i]$ mit einem Verfahren der Ordnung m akzeptiert, so wähle für $[x_i, x_{i+1}]$ dasjenige der drei Verfahren der Ordnungen $m-1, m, m+1$, das die größte Schrittweite \hat{h} gestattet.

Hierzu wird insbesondere eine Schätzung für die nächste Ableitung $u^{(m+2)}$ benötigt.

Die erwähnten Maßnahmen ermöglichen auch bei Mehrschrittverfahren eine selbststeuernden Einsatz. Allerdings ist der damit verbundene Verwaltungsaufwand erheblich höher als bei Einzschrittverfahren und muß bei einem Vergleich berücksichtigt werden.

2.3.2 Lineare Mehrschrittverfahren und Stabilität

Die Verfahren (2.3.2, 2.3.3) nutzen zwar die Funktionswerte $f_{i-j} \cong y'_{i-j}$ aus m zurückliegenden Punkten ("rechte Seite"), Lösungswerte dagegen nur von zwei Stellen, y_i, y_{i-1} . Man kann sich daher bei gleicher Schrittzahl m eine Verbesserung durch Einführung zusätzlicher Parameter vorstellen. Der Ansatz für allgemeinere Verfahren ist

$$\sum_{j=0}^m \alpha_j y_{i-j} = h \sum_{j=0}^m \beta_j f_{i-j}, \quad i = m, m+1, \dots, \quad (2.3.6)$$

mit $x_j = a + jh$, $f_j := f(x_j, y_j)$ und $\alpha_0 = 1$. Für $\beta_0 = 0$ ist dies ein explizites, andernfalls ein implizites, *lineares m -Schrittverfahren*. Den lokalen bzw. Konsistenzfehler erhält man wie üblich durch Einsetzen von u in (2.3.6) mit $f(x, u(x)) = u'(x)$:

$$T_h(x_{i-1}) = \frac{1}{h} \sum_{j=0}^m \alpha_j u(x_{i-j}) - \sum_{j=0}^m \beta_j u'(x_{i-j}). \quad (2.3.7)$$

Zur Überprüfung werden die Taylorentwicklungen in beiden Summen verglichen. Mit

$$\begin{aligned} u(x_i - jh) &= \sum_{k=0}^p \frac{(-jh)^k}{k!} u^{(k)}(x_i) + \mathcal{O}(h^{p+1}) \\ u'(x_i - jh) &= \sum_{k=0}^{p-1} \frac{(-jh)^k}{k!} u^{(k+1)}(x_i) + \mathcal{O}(h^p) \end{aligned}$$

in (2.3.7) ergibt sich

$$\begin{aligned} T_h(x_{i-1}) &= \sum_{j=0}^m \left[\sum_{\ell=0}^p \alpha_j \frac{(-j)^\ell}{\ell!} h^{\ell-1} u^{(\ell)} - \sum_{k=0}^{p-1} \beta_j \frac{(-j)^k}{k!} h^k u^{(k+1)} \right] + \mathcal{O}(h^p) \\ &= \frac{1}{h} u(x_i) \left(\sum_{j=0}^m \alpha_j \right) - \sum_{k=0}^{p-1} \frac{(-h)^k}{k!} u^{(k+1)}(x_i) \sum_{j=0}^m \left(\alpha_j \frac{j^{k+1}}{k+1} + \beta_j j^k \right) + \mathcal{O}(h^p). \end{aligned}$$

In dieser Entwicklung nach h -Potenzen liest man die Ordnungsbedingungen direkt ab:

Satz 2.3.2 *Das Verfahren (2.3.6) besitzt Konsistenzordnung p genau dann, wenn seine Koeffizienten das folgende lineare Gleichungssystem erfüllen, $\alpha_0 = 1$,*

$$\begin{aligned} \sum_{j=0}^m \alpha_j &= 0, \\ \sum_{j=0}^m \left(\frac{j^{k+1}}{k+1} \alpha_j + j^k \beta_j \right) &= 0, \quad k = 0, \dots, p-1. \end{aligned} \quad (2.3.8)$$

Ordnungsbedingungen können hier also wesentlich leichter aufgestellt und gelöst werden als bei den Einschrittverfahren. Man könnte also versuchen, durch geeignete Wahl der $2m(+1)$ freien Parameter α_j, β_j eine möglichst hohe Ordnung zu erreichen.

Beispiel 2.3.3 Konstruktion eines expliziten ($\beta_0 = 0$) 2-Schrittverfahrens maximaler Ordnung $p = 3$.

$$\left. \begin{aligned} 1 + \alpha_1 + \alpha_2 &= 0 \\ \alpha_1 + 2\alpha_2 + \beta_1 + \beta_2 &= 0 \\ \frac{1}{2}(\alpha_1 + 4\alpha_2) + \beta_1 + 2\beta_2 &= 0 \\ \frac{1}{3}(\alpha_1 + 8\alpha_2) + \beta_1 + 4\beta_2 &= 0 \end{aligned} \right\} \implies \begin{aligned} (\alpha_0, \alpha_1, \alpha_2) &= (1, 4, -5) \\ (\beta_0, \beta_1, \beta_2) &= (0, 4, 2) \end{aligned}$$

Das Verfahren lautet also

$$y_i + 4y_{i-1} - 5y_{i-2} = h(4f_{i-1} + 2f_{i-2}), \quad i \geq 2.$$

Zahlenbeispiel mit der Testgleichung $u' = -u$, $u(0) = 1$, und exakten Startwerten $y_0 := 1$, $y_1 = e^{-h}$, die Berechnung der Näherungen erfolgt durch $y_i := -4(1+h)y_{i-1} + (5-2h)y_{i-2}$ mit $h = 0.2$.

Numerisches Ergebnis:

i	1	2	3	4	5	6	7	8	9	⚡
y_i	0.8187	0.6701	0.5497	0.4438	0.3986	0.1283	1.2177	-5.2551	30.8262	

Die Fehler in den Approximationen oszillieren und wachsen ungefähr um einen Faktor 5 pro Schritt, das Verfahren ist offensichtlich **instabil**, Störungen wachsen extrem an. Im Gegensatz zu den Einschrittverfahren ist die Stabilitätsforderung eine gravierende Einschränkung für Mehrschrittverfahren! Denn es gibt eine sogenannte *Ordnungsbarriere* (von Dahlquist):

Ein stabiles m -Schritt-Verfahren besitzt eine (Konsistenz-)

$$\text{Ordnung } p \leq \begin{cases} m+1, & \text{wenn } m \text{ ungerade,} \\ m+2, & \text{wenn } m \text{ gerade.} \end{cases}$$

Gegenüber den Adams-Verfahren kann man die Ordnung also höchstens um 1 erhöhen, wenn m gerade ist. Die zugehörigen Verfahren besitzen eine den Adams-Verfahren ähnliche Struktur,

$$y_i - y_{i-k} = h \sum_{j=0}^m \beta_j f_{i-j}, \quad i \geq m, \quad 1 \leq k \leq m. \quad (2.3.9)$$

Der folgende Konvergenzsatz enthält für solche Verfahren eine zu den Einschrittverfahren ähnliche Stabilitätsaussage.

Satz 2.3.4 Die rechte Seite f der Dgl erfülle eine Lipschitzbedingung (2.1.5) mit der Konstanten L . Die Schrittweite im Verfahren sei $h \leq h_0 := 1/(2L|\beta_0|)$ ($:= \infty$ für $\beta_0 = 0$). Für die Startwerte gelte $\|y_i - u(x_i)\| \leq \varphi(h)$, $i = 0, \dots, m-1$, und für die Konsistenzfehler $\|T_h(x_{i-1})\| \leq \tau(h)$, $i = m, \dots, N$. Dann gilt die Fehlerschranke

$$\|y_i - u(x_i)\| \leq e^{\lambda(x_i - x_{m-1})}(\varphi(h) + 2(x_i - x_{m-1})\tau(h)), \quad i = m, \dots, N, \quad (2.3.10)$$

mit $\lambda = L \sum_{j=0}^m |\beta_j| / (1 - hL|\beta_0|) \leq 2L \sum_{j=0}^m |\beta_j|$. Für genügend oft differenzierbares f und $\tau(h) = \mathcal{O}(h^p)$, $\varphi(h) = \mathcal{O}(h^p)$ ist also insbesondere auch $\max_{\Delta} \|u_{\Delta} - u\| = \mathcal{O}(h^p)$, $h \rightarrow 0$.

Beweis Für $hL|\beta_0| < 1$ ist die Gleichung (2.3.9) nach y_i eindeutig auflösbar (Banachscher Fixpunktsatz). Wieder durch Subtraktion der Gleichungen (2.3.9) und (2.3.7),

$$\begin{aligned} y_i - y_{i-k} &= h \sum_{j=0}^m \beta_j f(x_{i-j}, y_{i-j}) \\ u(x_i) - u(x_{i-k}) &= h \sum_{j=0}^m \beta_j f(x_{i-j}, u(x_{i-j})) + hT_h(x_{i-1}) \end{aligned}$$

folgt für die Fehler $\varepsilon_j := y_j - u(x_j)$ die Beziehung

$$\|\varepsilon_i - \varepsilon_{i-k}\| = h \left\| \sum_{j=0}^m \beta_j (f(x_{i-j}, y_{i-j}) - f(x_{i-j}, u(x_{i-j}))) - T_h(x_{i-1}) \right\|$$

$$\leq hL \sum_{j=0}^m |\beta_j| \|\varepsilon_{i-j}\| + h\tau(h).$$

Daraus erhält man die Schranke

$$(1 - hL|\beta_0|) \|\varepsilon_i\| \leq \|\varepsilon_{i-k}\| + hL \sum_{j=1}^m |\beta_j| \max_{j=1}^m \|\varepsilon_{i-j}\| + h\tau(h).$$

Da auf der rechten Seite viele ältere Fehlerwerte auftreten, liegt es nahe deren Maximum

$$\delta_j := \max_{\nu \leq j} \|\varepsilon_\nu\|, \quad j = m-1, \dots, N,$$

zu betrachten. Nach Division durch $1 - hL|\beta_0|$ und Berücksichtigung der Identität $1/(1 - hL|\beta_0|) = 1 + hL|\beta_0|/(1 - hL|\beta_0|)$ ergibt sich so

$$\|\varepsilon_i\| \leq (1 + h\lambda)\delta_{i-1} + \frac{h}{1 - hL|\beta_0|} \tau(h), \quad i = m, \dots, N. \quad (2.3.11)$$

Da nun $\delta_i = \max\{\|\varepsilon_i\|, \delta_{i-1}\}$ und die rechte Seite größer als δ_{i-1} ist, kann die linke Seite in (2.3.11) durch δ_i ersetzt werden. Dann liegt aber die in Lemma 2.2.6 behandelte Situation vor, die auf die Schranke (2.3.10) führt. Als Startpunkt ist dabei allerdings x_{m-1} zu nehmen. ■

Die stabilen Verfahren höchster Ordnung $m+2$, m gerade, bekommt man mit der Wahl $k = m$ in (2.3.9) aus den abgeschlossenen Newton-Cotes-Quadraturformeln (1.2.5) zum Intervall $[x_{i-m}, x_i]$,

$$\frac{1}{h}(y_i - y_{i-m}) = \sum_{j=0}^m \beta_j f_{i-j} \cong \frac{1}{h} \int_{x_{i-m}}^{x_i} f(x, u(x)) dx. \quad (2.3.12)$$

Bei diesen Verfahren ist allerdings, im Gegensatz zu den Adams-Verfahren, die wie $e^{\lambda x}$ wachsende Stabilitätsschranke (2.3.10) scharf, selbst wenn im Problem nur exponentiell fallende Lösungen auftreten, vgl. Bsp 2.4.1. Daher verwendet man in der Praxis doch lieber Adams-Verfahren.

Aufgrund der Symmetrie um $\frac{1}{2}(x_{i-m} + x_i)$ besitzen Verfahren der Form (2.3.12) aber eine weitere wichtige Eigenschaft. Das einfachste, explizite Verfahren dieser Form mit $m = 2$ liefert die Rechteckregel zum Intervall $[-2h, 0]$, seine Symmetrie ist erkennbar in der Schreibweise

$$y_{i+1} - y_{i-1} = 2hf(x_i, y_i), \quad i = 1, 2, \dots$$

Diese *explizite Mittelpunkregel* hat nicht nur Ordnung 2, sondern besitzt sogar eine *asymptotische Entwicklung* (1.5.3) nach Potenzen von h^2 . Mit Richardson-Extrapolation (1.5.5) kann daher die Ordnung erhöht werden. Da das resultierende Gesamt-Verfahren vom Typus her aber wieder ein Einschrittverfahren ist, wird es jetzt getrennt behandelt.

2.4 Extrapolationsverfahren

Die explizite Mittelpunkregel benötigt zur Durchführung noch einen Näherungswert y_1 . Dieser kann ohne Verlust der Gesamtordnung 2 mit dem Eulerverfahren (Verfahren 1 in §1.2.1)

bestimmt werden. Das betrachtete Verfahren ist daher

$$\begin{aligned} y_1 - y_0 &= hf(x_0, y_0), \quad y_0 = u_0, \\ y_{i+1} - y_{i-1} &= 2hf(x_i, y_i), \quad i = 1, 2, \dots, N(-1). \end{aligned} \quad (2.4.1)$$

Der erste Schritt führt eine Asymmetrie ein, die von der Mittelpunkregel als oszillierende Störung weitergegeben wird. Daher besitzt die Näherungslösung u_Δ eine ungewöhnliche asymptotische Entwicklung der Gestalt

$$u_\Delta(x_j) - u(x_j) = \sum_{k=1}^q h^{2k} \left[g_{2k}(x_j) + (-1)^j \tilde{g}_{2k}(x_j) \right] + \mathcal{O}(h^{2q+2}), \quad h \rightarrow 0, \quad (2.4.2)$$

mit vom Gitter unabhängigen Koeffizientenfunktionen g_{2k}, \tilde{g}_{2k} . Insbesondere haben also Punkte x_j mit geradem und ungeradem Index verschiedene Entwicklungen und dürfen daher nicht gleichzeitig zur Extrapolation herangezogen werden.

Beispiel 2.4.1 Beim Problem $u' = \lambda u$, $u(0) = 1$ lautet das Verfahren $y_0 = 1$, $y_1 = 1 + h\lambda$, $y_{i+1} = y_{i-1} + 2h\lambda y_i$. Wie bei den linearen Dgln führt der Ansatz $y_j = c\xi^j$ in der Differenzgleichung $y_{i+1} - 2h\lambda y_i - y_{i-1} = 0$ auf $0 = c(\xi^{i+1} - 2h\lambda\xi^i - \xi^{i-1})\forall i$ und damit auf die quadratische Gleichung

$$\xi^2 - 2h\lambda\xi - 1 = 0 \Rightarrow \xi_{1/2} = h\lambda \pm \sqrt{1 + h^2\lambda^2} = \begin{cases} \sqrt{1 + h^2\lambda^2} + h\lambda & > 0 \\ -(\sqrt{1 + h^2\lambda^2} - h\lambda) & < 0 \end{cases}.$$

Die y_j besitzen daher die Darstellung $y_j = C_1\xi_1^j + C_2\xi_2^j$, wobei die Koeffizienten C_1, C_2 aus den Anfangsbedingungen $y_0 = 1$, $y_1 = 1 + h\lambda$ zu bestimmen sind. Um den Bezug auf das Gitter herzustellen, wird der Gitterindex bei $y_j = u_\Delta(jh)$ in der Form $j = x/h$ geschrieben. So ergibt sich die Näherungslösung explizit in der Form

$$\begin{aligned} 2u_\Delta(x) &= \left(1 + \frac{1}{\sqrt{1 + h^2\lambda^2}}\right) \left(\sqrt{1 + h^2\lambda^2} + h\lambda\right)^{x/h} \\ &\quad + (-1)^{x/h} \frac{h^2\lambda^2}{1 + h^2\lambda^2 + \sqrt{1 + h^2\lambda^2}} \left(\sqrt{1 + h^2\lambda^2} - h\lambda\right)^{x/h}. \end{aligned}$$

Man sieht leicht, dass mit Ausnahme von $(-1)^{x/h}$ alle Ausdrücke eine Taylorentwicklung nach h um $h = 0$ besitzen. Daraus folgt hier für den Anfang der Entwicklung (2.4.2)

$$u_\Delta(x) = e^{\lambda x} + h^2 \left[\underbrace{-\lambda^2 \left(\frac{1}{6}\lambda x + \frac{1}{4}\right) e^{\lambda x}}_{g_2} + (-1)^{x/h} \underbrace{\frac{1}{4}\lambda^2 e^{-\lambda x}}_{\tilde{g}_2} \right] + \mathcal{O}(h^4), \quad h \rightarrow 0.$$

Wenn die Lösung der Dgl fallend ist bei $\lambda < 0$, kann der oszillierende *und wachsende* Störterm $\tilde{g}_2(x)$ die numerische Lösung erheblich beeinträchtigen. Dieses Problem läßt sich durch folgenden Glättungsschritt etwas abschwächen.

$$\hat{y}_N := \frac{1}{4}(y_{N-1} + 2y_N + y_{N+1}).$$

Die Extrapolation bei der Mittelpunkregel ist als *Verfahren von Gragg-Bulirsch-Stoer* bekannt. Für das aktuelle Teilintervall $[x_0, x_0 + H]$ wählt man eine wachsende Folge gerader Gittergrößen $N_j \geq 2$, etwa $\{N_j\} \subseteq \{2, 4, 6, 8, 10, \dots\}$, und Teil-Schrittweiten $H_j := H/N_j$ sowie Gitter $x_i^{[j]} := x_0 + iH_j$ ($i = 0, \dots, N_j$). Mit den zugehörigen Näherungen $\hat{y}_{N_j}^{[j]} =: p_{j0}$ kann dann die

Richardson-Extrapolation (1.5.5) angewendet werden. Die Extrapolation eliminiert die führenden Terme der asymptotischen Entwicklung (2.4.2). Im j -ten Lauf wird dabei mit der Näherung $p_{j,0}$ vom j -ten Gitter eine zusätzliche Schrägzeile unten an das Extrapolations-Tableau angehängt.

p_{00}	p_{01}	p_{02}	\cdots	p_{0j}
p_{10}	p_{11}			
p_{20}	\vdots	\cdot		
\vdots	$p_{j-1,1}$			
p_{j0}				

Danach gilt die (lokale) Fehleraussage

$$\frac{1}{H}(p_{0j} - u(x_0 + H)) = \mathcal{O}(H^{2j+2}), \quad j \leq q. \tag{2.4.3}$$

Dieses Endergebnis p_{0j} gehört zu einem Einschrittverfahren der Ordnung $2j + 2$ für den Schritt $x_0 \rightarrow x_0 + H$, da es nicht auf Werte vor der Stelle x_0 zugreift, vgl. (2.4.1). Damit wurde insbesondere folgende Existenzaussage zur Theorie der Einschrittverfahren geliefert, die sich durch Abzählung der insgesamt benötigten Funktionsauswertungen ergibt.

Satz 2.4.2 *Für jedes $p \in 2\mathbb{N}$ gibt es ein Einschrittverfahren mit Ordnung p und $\frac{1}{4}p^2 + 1$ Stufen.*

Das Extrapolationsverfahren benutzt also viel mehr Funktionsauswertungen als die in §2.2 besprochenen Runge-Kutta-Verfahren. Aufgrund der einfachen und einheitlichen Konstruktion von Verfahren verschiedener Ordnung kann hier aber eine Schrittweiten- und Ordnungssteuerung wie bei den Mehrschrittverfahren durchgeführt werden. Insbesondere wird die aktuell verwendete Ordnung $2j + 2 (\leq 2q + 2)$ in (2.4.3) in der Praxis nicht vorgegeben, sondern durch Inspektion des Tableaus p_{ik} bestimmt.

Demo-Beispiel: Extrapolationsverfahren mit Ordnungs- und Schrittweitensteuerung (vgl. Ende von §2.3.1) beim Erde-Mond-Orbit und beim AWP $u' = f(x, u), u(-\frac{1}{2}) = u_0$ auf $[-\frac{1}{2}, 1]$ mit $(a := 800, b = \ln(1 + a) - 3)$

$$f(x, y) = \begin{cases} -2axy^2, & x \leq 0 \\ 2\left(-\frac{ax}{1+ax^2} + bx\right)y^2, & x > 0 \end{cases} \quad \text{mit } u(x) = \begin{cases} \frac{1}{1+ax^2}, & x \leq 0 \\ \frac{1}{1+\ln(1+ax^2)-bx^2}, & x > 0 \end{cases} .$$

Bei $x = 0$ hat f'' eine Unstetigkeit (Sprung).

Vergleich der Verfahren:

	Einschritt	Extrapolation	Mehrschritt
Konsistenz-Ordnung	aufwändig	einfach	einfach
Stabilität	immer	→ ungünstiger	← Einschränkung
Aufwand	merklich	hoch	gering
Genauigkeit	hoch	hoch	mittel
Fehlerschätzung	Einbettung	einfach	einfach
Schrittweitenänd.	immer	→ immer	aufwändig
Ordnungswechsel	nein	einfach	möglich

Die *praktische* Erfahrung mit diesen Verfahrensklassen ergibt ein differenziertes Bild, keines kann als Universalverfahren gelten. Zwar schneiden bei der Gesamtzahl der Funktionsaufrufe von f die Prädiktor-Korrektor-Verfahren am günstigsten ab. Da sie bei gleicher Schrittweite größere Fehler und im adaptiven Einsatz beim Schrittweitenwechsel den höchsten Verwaltungsaufwand haben, sind die Rechenzeiten nur selten kürzer als bei Einschrittverfahren, nämlich dort, wo die numerische Auswertung $f(x, y)$ sehr aufwändig ist. Im Vergleich der Einschrittverfahren, Runge-Kutta und Extrapolation, zeigt sich ein Vorteil der möglichen Ordnungssteuerung bei letzteren erst für (extrem) hohe Genauigkeitsanforderungen. Das Runge-Kutta-Tripel DOP853 wird erst für extreme Toleranzen ($< 10^{-12}$) von Extrapolationsverfahren übertroffen.

Zusatzanforderungen: Mit den behandelten Verfahren können viele Anfangswertprobleme effizient und verlässlich gelöst werden. Aus der Praxis oder aufgrund der aktuellen Computerentwicklung kommen aber immer wieder neue Anforderungen an die Numerik. Schwierigkeiten bekommen die Standardverfahren, wenn die Stabilitätsschranken in den Sätzen 2.2.7 bzw. 2.3.4 unbrauchbar werden wegen $L(b - a) \rightarrow \infty$. So weisen viele Probleme extrem große Lipschitzkonstanten L auf trotz glatter Lösungen u (\rightarrow steife AWP). Bei anderen Anwendungen (Molekularphysik) ist man an sehr langen Zeitintervallen interessiert und will dabei Erhaltungssätze (Energie) möglichst exakt einhalten (\rightarrow geometrische Verfahren). Ganz andere Anforderungen kommen aus der Entwicklung von Parallelrechnern (aktuell Multi-Core-Prozessoren), keines der behandelten Verfahren erlaubt eine Parallelisierung in der Methode. Parallelität wird in einer Spezialvorlesung im nächsten Semester behandelt.

2.5 Schießverfahren für Randwertprobleme

Hier wird das standardisierte, nichtlineare Randwertproblem (RWP, vgl. §2.1)

$$u'(x) = f(x, u(x)), \quad x \in (a, b), \quad r(u(a), u(b)) = 0, \quad (2.5.1)$$

betrachtet mit differenzierbaren Funktionen $f(x, y)$, $r(y_0, y_1)$. Für die numerische Lösung solcher gewöhnlicher RWPe kann man zunächst Verfahren für Anfangswertprobleme als Hilfsmittel verwenden, aber auch neue Verfahren untersuchen, die auf partielle Randwertprobleme verallgemeinert werden können. Aus der Reihe der letzteren wird eine auf Differenzenapproximationen

basierende Klasse schon hier besprochen, da deren wesentliche Eigenschaften im gewöhnlichen Fall einfacher zu diskutieren sind. Methoden, die auf AWP-Verfahren aufbauen, haben den praktischen Vorteil, dass man dafür adaptive Standard-Programme für AWPe einsetzen kann.

Schießverfahren

In §2.1 wurde das Schießprinzip schon beim linearen Randwertproblem

$$u'(x) = A(x)u(x) + g(x), \quad x \in (a, b), \quad R_0u(a) + R_1u(b) = d, \quad (2.5.2)$$

eingesetzt. Mit einer Fundamentallösung $Y(x)$ des Matrix-Anfangswertproblems $Y'(x) = A(x)Y(x)$, $Y(a) = I$, wurde über die Lösungsdarstellung $u(x) = Y(x)\eta + \dots$ der unbekannte Anfangswert $\eta = u(a) \in \mathbb{R}^n$ durch Einsetzen in die Randbedingung bestimmt: $R_0\eta + R_1(Y(b)\eta + \dots) = d$. Damit wurde also die Lösungskurve $u(x)$ durch Variation des Startwertes $\eta = u(a)$ so angepasst, dass ihr Wert $u(b)$ am rechten Rand (zusammen mit $u(a) = \eta$) die Randbedingung erfüllt (Bildlich *Schießen/Werfen*: Anpassung der Wurfrichtung so, dass Treffer erzielt wird). Zur Verallgemeinerung dieses Prinzips muss die Abhängigkeit der Lösung vom Anfangswert betont werden. Daher bezeichne jetzt $u(x; a, \eta)$ die Lösungsschar des AWP

$$\frac{d}{dx}u(x; a, \eta) = f(x, u(x; a, \eta)), \quad x > a, \quad u(a; a, \eta) = \eta. \quad (2.5.3)$$

Die spezielle Lösung des RWP (2.5.1) ist dann diejenige, $u(x) = u(x; a, \hat{\eta})$, aus der Schar, deren Anfangswert $\hat{\eta}$ das nichtlineare Gleichungssystem

$$F(\eta) = 0, \quad \text{mit } F : \begin{cases} \mathbb{R}^n & \rightarrow \mathbb{R}^n \\ \eta & \mapsto r(\eta, u(b; a, \eta)) \end{cases} \quad (2.5.4)$$

erfüllt. Zur Lösung solcher Gleichungssysteme kommt vor allem das Newton-Verfahren in Betracht, vgl. Numerik I, §5.2. Sein Einsatz setzt aber die genügende Differenzierbarkeit der Funktion F voraus. Die Lipschitz-stetige Abhängigkeit der Funktion $u(x; a, \eta)$ vom Anfangswert η wurde schon in Satz 2.1.1 gezeigt. Darüber hinaus gilt aber auch

Satz 2.5.1 Die rechte Seite f sei zweimal stetig differenzierbar, ihre Ableitung $f_y = \frac{\partial f}{\partial y}$ beschränkt auf $[a, b] \times \mathbb{R}^n$. Dann hängt die Lösungsschar $u(x; a, \eta)$ von (2.5.3) differenzierbar vom Anfangswert η ab. Ihre Ableitung an der Stelle $(x; a, \eta)$ ist

$$\frac{\partial}{\partial \eta}u(x; a, \eta) = Y(x; a, \eta), \quad (2.5.5)$$

wobei Y das Fundamentalsystem zur linearisierten Differentialgleichung (Variationsgleichung) ist,

$$\frac{d}{dx}Y(x; a, \eta) = f_y(x, u(x; a, \eta)) \cdot Y(x; a, \eta), \quad Y(a; a, \eta) = I. \quad (2.5.6)$$

Beweis Die Ableitung wird entsprechend der allgemeinen Definition von Differenzierbarkeit als in der Störung $s \in \mathbb{R}^n$ linearer Anteil der Differenz $v(x) := u(x; a, \eta + s) - u(x; a, \eta)$ bestimmt. Offensichtlich ist $v(a) = \eta + s - \eta = s$ und linear in s . Nach Satz 2.1.1 gilt daher die Schranke $\|v(x)\| \leq \exp(L(x-a))\|s\|$. Da beide Funktionen u Lösungen der Dgl sind, gilt auch für $x \in (a, b)$ (mit $' = \partial/\partial x$)

$$\begin{aligned} v'(x) &= u'(x; a, \eta + s) - u'(x; a, \eta) = f(x, u(x; a, \eta) + v(x)) - f(x, u(x; a, \eta)) \\ &= f_y(x, u(x; a, \eta))v(x) + \mathcal{O}(\|v(x)\|^2) = f_y(x, u(x; a, \eta))v(x) + \mathcal{O}(\|s\|^2). \end{aligned}$$

Die Funktion $\hat{v}(x) := Y(x; a, \eta)s$ mit dem Fundamentalsystem aus (2.5.6) ist Lösung des AWP

$$\hat{v}'(x) = f_y(x, u(x; a, \eta))\hat{v}(x), \quad \hat{v}(a) = s.$$

Daher gilt insbesondere $v(x) - \hat{v}(x) = \mathcal{O}(\|s\|^2)$ und somit

$$u(x; a, \eta + s) = u(x; a, \eta) + Y(x; a, \eta)s + \mathcal{O}(\|s\|^2), \quad s \rightarrow 0.$$

Damit ist $\hat{v}(x) = Y(x)s$ die Richtungsableitung von u in Richtung s , und damit ist $Y = \frac{\partial}{\partial \eta}u$ die totale Ableitung der Lösungsschar, wie in (2.5.5) behauptet. ■

Bemerkung: Die Ableitung (2.5.5) der Lösungsschar nach dem Anfangswert η ist also eine Matrixfunktion $Y(x)$, welche das lineare AWP $Y'(x) = A(x)Y(x)$, $Y(a) = I$ erfüllt, wobei die Ableitungsmatrix $A(x) = f_y(x, u(x))$ in der Lösung $u(x)$ der Ursprungs-Dgl mit $u(a) = \eta$ ausgewertet wird.

Die im Newtonverfahren zur Lösung des Systems (2.5.4) benötigte Ableitung erhält man bei differenzierbarer Funktion $r = r(y_0, y_1)$ somit nach der Kettenregel

$$\begin{aligned} F'(\eta) &= \frac{d}{d\eta}r(\eta, u(b; a, \eta)) = \frac{\partial r}{\partial y_0}(\eta, u(b; a, \eta)) + \frac{\partial r}{\partial y_1}(\eta, u(b; a, \eta))\frac{\partial u}{\partial \eta}(b; a, \eta) \\ &= \frac{\partial r}{\partial y_0}(\eta, u(b; a, \eta)) + \frac{\partial r}{\partial y_1}(\eta, u(b; a, \eta))Y(b; a, \eta). \end{aligned} \quad (2.5.7)$$

Die Bestimmung von F' ist somit durch Lösung des Matrix-AWP (2.5.6), also durch (numerische) Berechnung von n Lösungen $y_i(x) = Y(x)e_i$ der linearen AWPe

$$y_i' = f_y(x, u(x; a, \eta))y_i, \quad y_i(a) = e_i, \quad i = 1, \dots, n,$$

für die Spalten von Y möglich (e_i : i -ter Einheitsvektor). Zur praktischen Durchführung können alle im §2.2-2.4 behandelten Verfahren herangezogen werden, in der folgenden Diskussion wird aber jetzt zur Vereinfachung von einer exakten Lösung dieser AWPe ausgegangen. Damit sind alle benötigten Größen beim Newtonverfahren

$$\eta^{[k+1]} := \eta^{[k]} - \left(F'(\eta^{[k]}) \right)^{-1} F(\eta^{[k]}), \quad k = 0, 1, \dots$$

bekannt. Das Verfahren wird noch einmal ausführlich formuliert, wobei bei den Hilfsgrößen $u(x; a, \eta^{[k]})$, $Y(x; a, \eta^{[k]})$ der Index k und die Abhängigkeit vom Startwert unterdrückt wird.

Algorithmus 2.5.2 *Schießverfahren*: Gegeben sei $\eta^{[0]} \in \mathbb{R}^n$. Für $k = 0, 1, \dots$ löse man

$$\text{die AWPe } \begin{cases} u'(x) = f(x, u(x)), & u(a) = \eta^{[k]}, \\ Y'(x) = f_y(x, u(x))Y(x), & Y(a) = I, \end{cases} \quad (2.5.8)$$

$$\text{das LGS } \left(r_{y_0} + r_{y_1} Y(b) \right) \left(\eta^{[k+1]} - \eta^{[k]} \right) = -r(\eta^{[k]}, u(b)). \quad (2.5.9)$$

Die partiellen Ableitungen r_{y_0}, r_{y_1} sind nach (2.5.7) ebenfalls an der Stelle $(\eta^{[k]}, u(b))$ auszuwerten. Zunächst sind also in (2.5.8) jeweils $n + 1$ gekoppelte AWPes zu lösen und danach in (2.5.9) ein einzelnes lineares $n \times n$ -Gleichungssystem.

Im linearen Spezialfall (2.5.2) ist $r(y_0, y_1) = R_0 y_0 + R_1 y_1 - d$, d.h. $\frac{\partial r}{\partial y_0} = R_0$, $\frac{\partial r}{\partial y_1} = R_1$. Auch $f_y(x, y) = A(x)$ ist unabhängig von $u(x; a, \eta)$ also auch von η . Somit hängt auch Y nicht von η ab. Ein Schritt des Schießverfahrens mit Startwert $\eta^{[0]} = 0$ liefert daher

$$\eta^{[1]} = - \left(R_0 + R_1 Y(b) \right)^{-1} (0 + R_1 u(b) - d).$$

Dies ist genau der Lösungswert, der bei der expliziten Lösung in (2.1.11) berechnet wurde, (2.5.9) konvergiert also bei linearen Problemen erwartungsgemäß in einem Schritt.

Wie auch sonst beim Newton-Verfahren kann die Konvergenz dieser Iteration für "genügend genaue" Startwerte $\eta^{[0]}$ gezeigt werden, wenn die zweite Ableitung f_{yy} beschränkt ist, und in der exakten Lösung u des RWPes die Randmatrix $F'(u(a)) = r_{y_0} + r_{y_1} Y(b)$ regulär ist, mit $Y(b) = Y(b; a, u(a))$. In der Praxis ist diese Aussage aber wertlos, wenn die Lösungen $u(x; a, \eta)$ sich sehr schnell von der gesuchten Lösung $u(x; a, u(a))$ entfernen. Dann ist der *Einzugsbereich* der Iteration (2.5.9), und eventuell sogar der Definitionsbereich der Funktion F überhaupt, sehr klein. Das folgende Beispiel illustriert das Problem.

Beispiel 2.5.3 Die allgemeine Lösung der folgenden skalaren Dgl ist bekannt

$$u' = \frac{1 + u^2}{1 + x^2} =: f(x, u) \quad \Rightarrow \quad u(x; a, \eta) = \frac{\eta - a + (1 + a\eta)x}{1 + a\eta - (\eta - a)x}.$$

Für $a = 0$ soll das RWP mit der Randbedingung $u(0) + u(b) = b$, $b > 0$, (Lösung $u(x) \equiv x$) mit dem Schießverfahren gelöst werden. Damit aber die Lösung $u(x; 0, \eta) = (\eta + x)/(1 - \eta x)$ überhaupt den rechten Intervallrand erreicht, ist ein Startwert $\eta < 1/b$ zu wählen. Mit der angegebenen allgemeinen Lösung hat die Randgleichung (2.5.4) die Form

$$F(\eta) = \eta + u(b; 0, \eta) - b = \eta - b + \frac{b + \eta}{1 - b\eta} \stackrel{!}{=} 0.$$

Diese(!) Gleichung besitzt formal 2 Lösungen $\eta = 0$ und $\tilde{\eta} = b + 2/b$. Allerdings gehört zu $\tilde{\eta}$ keine Lösung des RWPes, da das zugehörige $u(x)$ zwischen 0 und b einen Pol besitzt. Die Ableitung $F'(\eta)$ läßt sich natürlich direkt berechnen. Andererseits gilt tatsächlich

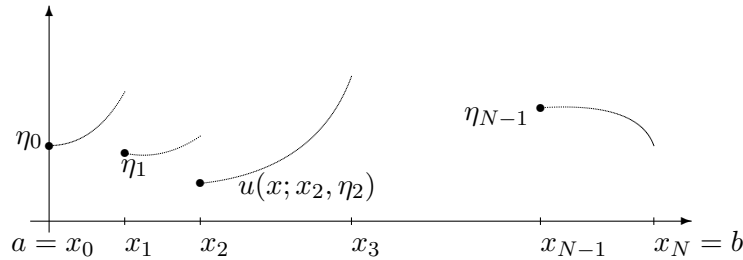
$$Y(x) = \frac{\partial u}{\partial \eta}(x; 0, \eta) = \frac{1 + x^2}{(1 - \eta x)^2} \quad \text{und} \quad Y'(x) = 2 \frac{\eta + x}{(1 - \eta x)(1 + x^2)} Y(x) = \underbrace{\frac{2u(x; 0, \eta)}{1 + x^2}}_{f_y(x, u)} Y(x).$$

Bei positiven Startwerten konvergiert das Newtonverfahren ebenfalls für $\eta < 1/b$ (Ü-Aufgabe).

Die oben erwähnten Schwierigkeiten bei rasch divergierenden Lösungen des Anfangswertproblems kann man dadurch entschärfen, dass man die Lösungen jeweils nur über kurze Intervalle integriert, das Gesamtintervall also in mehrere Teile zerlegt.

Mehrzielmethode: Das Gesamtintervall $[a, b]$ wird unterteilt durch ein Gitter

$$\Delta : a = x_0 < x_1 < \dots < x_N = b.$$



In jedem Teilintervall $[x_i, x_{i+1}]$ wird das Schießverfahren mit einem *eigenen Startwert* η_i angewendet. Dabei werden N Teil-Lösungen $u(x; x_i, \eta_i)$ der AWPes

$$u'(x; x_i, \eta_i) = f(x, u(x; x_i, \eta_i)), \quad \underline{u(x_i; x_i, \eta_i)} = \eta_i, \quad i = 0, \dots, N-1, \quad (2.5.10)$$

berechnet. Eine (stetige!) Lösung u des Randwertproblems (2.5.1) setzt sich aus diesen Stücken genau dann zusammen, wenn diese an den Gitterpunkten zusammenpassen und außerdem die Randbedingung erfüllen, wenn also gilt

$$\begin{aligned} u(x_i; x_{i-1}, \eta_{i-1}) \stackrel{!}{=} \eta_i &= u(x_i; x_i, \eta_i), \quad i = 1, \dots, N-1, \\ r(\eta_0, u(x_N; x_{N-1}, \eta_{N-1})) \stackrel{!}{=} 0. \end{aligned}$$

Wegen der Stetigkeit von f ist die so zusammengesetzte Lösung auch stetig differenzierbar. Im Unterschied zum Einfach-Schießverfahren sind nun N unbekannte Startwerte η_i zu bestimmen als Lösung des nichtlinearen Gleichungssystems für $\vec{\eta} \in \mathbb{R}^{nN}$,

$$\vec{\eta} := \begin{pmatrix} \eta_0 \\ \eta_1 \\ \vdots \\ \eta_{N-1} \end{pmatrix} \text{ mit } F(\vec{\eta}) := \begin{pmatrix} u(x_i; x_{i-1}, \eta_{i-1}) - \eta_i, & i = 1, \dots, N-1, \\ r(\eta_0, u(x_N; x_{N-1}, \eta_{N-1})), & i = N \end{pmatrix} \stackrel{!}{=} 0. \quad (2.5.11)$$

Hierin ist das einfache Schießverfahren mit $N = 1$ enthalten. Das Newtonverfahren zur Lösung dieses Systems (2.5.11) lautet formal

$$\frac{\partial F}{\partial \vec{\eta}}(\vec{\eta}^{[k]}) (\vec{\eta}^{[k+1]} - \vec{\eta}^{[k]}) = -F(\vec{\eta}^{[k]}), \quad k = 0, 1, \dots \quad (2.5.12)$$

und berechnet im k -ten Schritt den Vektor der Korrekturen

$$v_i := \eta_i^{[k+1]} - \eta_i^{[k]}, \quad i = 0, \dots, N-1.$$

Dazu werden die Ableitungen der einzelnen Zeilenblöcke von F benötigt. Nach Satz 2.5.1 sind die partiellen Ableitungen der Teillösungen nach ihrem Anfangswert wieder

$$\begin{aligned} \frac{\partial}{\partial \eta_i} u(x; x_i, \eta_i) &= Y(x; x_i, \eta_i) \\ Y'(x; x_i, \eta_i) &= f_y(x, u(x; x_i, \eta_i))Y(x; x_i, \eta_i), \quad Y(x_i; x_i, \eta_i) = I. \end{aligned} \quad (2.5.13)$$

In jedem Teilintervall $[x_i, x_{i+1}]$ sind also wieder n lineare Hilfs-Dgln mit der Variationsgleichung zu lösen, jeweils mit eigenem Anfangswert $Y(x_i; x_i, \cdot) = I$. Daher bestimmt der Newtonschritt (2.5.12) die Lösung (v_0, \dots, v_{n-1}) des folgenden linearen Gleichungssystems

$$\begin{aligned} Y(x_i; x_{i-1}, \eta_{i-1})v_{i-1} - v_i &= \eta_i - u(x_i; x_{i-1}, \eta_{i-1}), \quad i = 1, \dots, N-1, \\ R_0 v_0 + R_1 Y(b; x_{N-1}, \eta_{N-1})v_{N-1} &= -r(\eta_0, u(b; x_{N-1}, \eta_{N-1})), \end{aligned} \quad (2.5.14)$$

mit $R_0 = \frac{\partial r}{\partial y_0}$, $R_1 = \frac{\partial r}{\partial y_1}$, ausgewertet jeweils an der Stelle $(\eta_0, u(b; x_{N-1}, \eta_{N-1}))$. Die Aufstellung dieses Systems erfordert wieder die (numerische) Lösung der Teil-AWPe (2.5.10), (2.5.13). Die Struktur des Systems (2.5.14) entspricht der der Ableitungsmatrix $F' = \partial F / \partial \vec{\eta}$ von (2.5.11). Mit den Matrizen $Y_{i,i-1} := Y(x_i; x_{i-1}, \eta_{i-1})$ hat diese die Form

$$F'(\vec{\eta}^{[k]}) = \begin{pmatrix} Y_{1,0} & -I & & & \\ & Y_{2,1} & -I & & \\ & & \ddots & \ddots & \\ & & & Y_{N-1,N-2} & -I \\ R_0 & 0 & \cdots & 0 & R_1 Y_{N,N-1} \end{pmatrix}. \quad (2.5.15)$$

Dies ist eine $(nN) \times (nN)$ -Matrix mit zyklischer Block-Bidiagonalgestalt (Treppenform) aus $n \times n$ -Blöcken. Bei linearen Problemen ist $Y_{i,i-1} = Y(x_i)Y(x_{i-1})^{-1}$ unabhängig von η_{i-1} .

Die *Auflösung* der linearen Systeme in (2.5.14) kann mit dem Gauß-Algorithmus erfolgen. Bei Pivotisierung mit Spaltenvertauschung werden dabei nur Elemente in der untersten (Block-) Zeile neu eingeführt und es wird nur mit dem skizzierten Bereich gearbeitet:

$$\begin{pmatrix} \square & \square & & & & \\ & \square & \square & & & \\ & & \square & \square & & \\ & & & \ddots & \ddots & \\ & & & & \square & \square \\ \square & \square & \square & \cdots & \square & \square \end{pmatrix}$$

Dieses Verfahren ist auch in kritischen Fällen (z.B., wenn die Matrizen $Y_{i,i-1}$ große Normen besitzen) numerisch recht stabil. In vielen Fällen kann man dieses Gleichungssystem aber noch einfacher durch *Kondensation* lösen. Dazu behandelt man nur v_0 als Unbekannte und eliminiert die restlichen v_i mit Hilfe der ersten Gleichungen in (2.5.14):

$$\begin{aligned} v_1 &= Y_{10}v_0 + w_1, \quad v_2 = Y_{21}v_1 + w_2 = Y_{21}Y_{10}v_0 + Y_{21}w_1 + w_2, \quad \dots \\ v_{N-1} &= Y_{N-1,N-2} \cdots Y_{21}Y_{10}v_0 + \hat{w}. \end{aligned}$$

Mit der Randbedingung in (2.5.14) ergibt dies also das System

$$\left(R_0 + R_1 Y_{N,N-1} Y_{N-1,N-2} \dots Y_{21} Y_{10}\right) v_0 = -r(\eta_0, \dots) - R_1 Y_{N,N-1} \hat{w}, \quad (2.5.16)$$

das nur noch von der Größe $n \times n$ ist. Diese Kondensation kann übrigens zusammen mit der Berechnung der $Y_{i,i-1}$, etc., erfolgen. Dann sind für einen Newtonschritt jeweils zwei Läufe durch das Intervall nötig, nämlich:

AWPe mit Kondensation, Lösung v_0 aus (2.5.16), Berechnung von v_1, \dots, v_{N-1} .

Im linearen Fall sind die Mehrzielmethode mit Kondensation und das einfache Schießverfahren gleich. Denn, hier ist $Y_{i,i-1} = Y(x_i; x_{i-1}) = Y(x_i)Y(x_{i-1})^{-1}$ unabhängig von η_{i-1} und für die Produkte gilt daher

$$Y_{N,N-1} \dots Y_{21} Y_{10} = Y(b; x_{N-1}) \dots Y(x_2; x_1) Y(x_1; a) = Y(b; a) = Y_{N,0}.$$

Die Matrix $R_0 + R_1 Y_{N,N-1} \dots Y_{10} = R_0 + R_1 Y_{N0}$ ist hier identisch mit der in (2.5.9). Der wesentliche Vorteil der Mehrzielmethode ist die (starke) Vergrößerung des Konvergenzbereichs beim Newtonverfahren im nichtlinearen Fall. Die Wahl der Zwischenzielpunkte x_1, \dots, x_{N-1} kann dabei leicht während der Durchführung des Verfahrens modifiziert werden. Z. B. ist es sinnvoll, bei starkem Anwachsen einer Zwischenlösung $u(\cdot; x_i, \eta_i)$ neue Zielpunkte einzufügen. Die Auswirkungen der Mehrzielmethode wird anhand des einführenden Beispiels diskutiert.

Beispiel 2.5.4 Als Mindestvoraussetzung für die Durchführbarkeit des Verfahrens wird die Existenz der Teillösungen in $[x_i, x_{i+1}]$ überprüft. Beim Problem aus Beispiel 2.5.3 lautet der Wert der Lösungsschar in x_{i+1} ,

$$u(x_{i+1}; x_i, \eta_i) \equiv \frac{\eta_i(1 + x_i x_{i+1}) + (x_{i+1} - x_i)}{1 + x_i^2 - (x_{i+1} - x_i)(\eta_i - x_i)}$$

und existiert insbesondere nur dann, wenn der Nenner positiv ist, also für

$$\eta_i - x_i = \eta_i - u(x_i) \leq \frac{1 + x_i^2}{x_{i+1} - x_i}, \quad i = 0, \dots, N-1.$$

Diese Genauigkeitsforderungen an die lokalen Startwerte η_i sind schwächer als beim Einfach-Schießverfahren. Bei äquidistanter Unterteilung, $x_i = ib/N$, etwa ist für η_0 nur zu fordern $\eta_0 - u(x_0) \leq N/b$, während in Beispiel 2.5.3 $\eta - u(x_0) < 1/b$ erforderlich war.

Demo-Beispiel: Mehrzielmethode bei den Randwertproblemen 2. Ordnung $u'' = \frac{3}{2}u^2$, $u(0) = 4$, $u(1) = 1$ und $u'' = 10(u^2 - 1)$, $u(0) = u'(1) = 0$. Beide besitzen jeweils 2 Lösungen.

2.6 Differenzenverfahren für Randwertprobleme

Ansatzpunkt für die Konstruktion numerischer Verfahren war bisher die zur Dgl erster Ordnung äquivalente Integralgleichung. Bei allgemeineren Problemstellungen, in denen verschiedene Ableitungen von u auftreten, ist dies unbequem oder nicht mehr möglich. Hier approximiert

man Ableitungen, auch höherer Ordnung, direkt durch Differenzenausdrücke und vermeidet den Übergang zum System erster Ordnung. Differenzenformeln können explizit konstruiert oder systematisch durch Ableitung von Interpolationspolynomen bestimmt werden (vgl. §1.6). Eine besonders einfache Form besitzen die symmetrischen Differenzenquotienten der Ordnung 2.

Lemma 2.6.1 Für $u \in C^{m+2}[-\epsilon, \epsilon]$, $0 < \frac{m}{2}h \leq \epsilon$, ist

$$\frac{1}{h^m} \sum_{j=0}^m \binom{m}{j} (-1)^j u\left(j - \frac{m}{2}\right)h = u^{(m)}(0) + C_m h^2 u^{(m+2)}(\xi), \quad \xi \in (-\epsilon, \epsilon). \quad (2.6.1)$$

”Beweis”: Die Koeffizienten der Formel sind die aus dem binomischen Satz. Die Herleitung der Formel ist durch formale Rechnung mit Operatorreihen bei $C^\infty(\mathbb{R})$ -Funktionen u (z.B. Polynomen) einfach möglich. Mit $h > 0$ werden Verschiebungs- und Ableitungsoperatoren punktweise definiert durch

$$(S_h u)(x) := u(x+h), \quad (Du)(x) := u'(x).$$

Für $r \in \mathbb{R}$ ist dann $(S_h^r u)(x) = (S_{rh} u)(x) = u(x+rh)$. Der Satz von Taylor besagt

$$S_h u(x) = u(x+h) = \sum_{k=0}^{\infty} \frac{h^k}{k!} u^{(k)}(x) = \left(\sum_{k=0}^{\infty} \frac{h^k}{k!} D^k \right) u(x) = e^{hD} u(x) \Rightarrow \boxed{S_h = e^{hD}} = \exp(hD).$$

Eine Differenzbildung liefert Darstellungen für Differenzenoperatoren, etwa bei $S_h - I = e^{hD} - I$. Der symmetrische Differenzenausdruck in (2.6.1) ergibt sich aus der binomischen Formel bei der einfachen Differenz $u(\frac{h}{2}) - u(-\frac{h}{2}) = (S_{h/2} - S_{-h/2})u(0)$:

$$\begin{aligned} S_{h/2} - S_{-h/2} &= e^{\frac{h}{2}D} - e^{-\frac{h}{2}D} = 2 \sinh\left(\frac{h}{2}D\right) \Rightarrow \\ (S_h - I)^m S_{-h/2}^m &= (S_{h/2} - S_{-h/2})^m = \left(2 \sinh\frac{h}{2}D\right)^m \\ &= \left(hD + \frac{h^3}{24}D^3 + \dots\right)^m = h^m (D^m + \frac{m}{24}h^2 D^{m+2} + \dots). \end{aligned} \quad (2.6.2)$$

Dies entspricht (2.6.1), wobei durch einige Zusatzüberlegungen die Darstellung des (Taylor-) Restglieds vereinfacht wurde. ■

Differenzenverfahren lohnen sich insbesondere bei (Systemen von) Dgln höherer Ordnung, da man so den Übergang zum größeren System erster Ordnung vermeiden kann. Als Vorbereitung auf die partiellen Dgln wird das häufig auftretende lineare Randwertproblem zweiter Ordnung mit Koeffizienten p, q, g betrachtet

$$-u''(x) + p(x)u'(x) + q(x)u(x) = g(x), \quad u(a) = \alpha, \quad u(b) = \beta. \quad (2.6.3)$$

Als Randbedingungen sind in beiden Randpunkten jeweils nur die Funktionswerte vorgeschrieben. Unter der Voraussetzung $p, p', q, g \in C[a, b]$ und $q^* + \frac{1}{4}p^2 - \frac{1}{2}p' + (b-a)^2/\pi^2 > 0$, wobei

Definition 2.6.2 Eine Matrix $A = (a_{ij}) \in \mathbb{R}^{n \times n}$ heißt M-Matrix, wenn gilt

$$\left. \begin{array}{l} a_{ii} > 0 \quad \forall i = 1, \dots, n, \\ a_{ij} \leq 0 \quad \forall i \neq j, \end{array} \right\} A \text{ regulär und } A^{-1} \geq 0. \quad (2.6.7)$$

Nach einer Zerlegung der Matrix $A = D - N$ in Diagonale D und Nebendiagonalen $-N$ lautet die Vorzeichenbedingung (2.6.7) kürzer $N \geq 0, a_{ii} > 0 \forall i$. Im Zusammenhang mit der Maximumnorm wurden wichtige Schlußweisen für nichtnegative Vektoren $v \geq 0$ bzw. Matrizen $B \geq 0$ schon in der Numerik 1 behandelt (§4). So gilt in diesem Fall $\|v\|_\infty \leq a \iff v \leq a\mathbb{1}$ mit $\mathbb{1} = (1, \dots, 1)^\top$ und $\|B\|_\infty \leq b \iff B\mathbb{1} \leq b\mathbb{1}$. Außerdem ist Produktbildung monoton, $v \leq w, B \geq 0 \Rightarrow Bv \leq Bw$. Kriterien für M-Matrizen enthält

Lemma 2.6.3 a) Eine Matrix $A = D - N$ mit der Vorzeichenverteilung in (2.6.7), $D, N \geq 0$, ist genau dann M-Matrix, wenn $\rho(D^{-1}N) < 1$ gilt.

b) Eine Matrix A mit der Vorzeichenverteilung in (2.6.7) ist (genau dann) M-Matrix, wenn ein Vektor $z > 0$ existiert mit $w := Az > 0$. Mit diesem gilt

$$\|A^{-1}\|_\infty \leq \frac{\max_i z_i}{\min_i w_i}. \quad (2.6.8)$$

Beweis a) Für $\rho(D^{-1}N) < 1$ konvergiert die Neumannreihe

$$A^{-1} = (I - D^{-1}N)^{-1}D^{-1} = \sum_{j=0}^{\infty} \underbrace{(D^{-1}N)^j}_{\geq 0} D^{-1} \geq 0,$$

die Inverse existiert also und ist nichtnegativ. Sei umgekehrt A M-Matrix und $v \neq 0$ ein Eigenvektor von $D^{-1}N$ zum Eigenwert λ . Wegen $N \geq 0$ gilt $|Nv| \leq N|v|$. Aus $\lambda Dv = Nv$ folgt daher

$$-N|v| \leq -|\lambda|D|v| = -|\lambda Dv| \quad \Rightarrow \quad A|v| = (D - N)|v| \leq (1 - |\lambda|)D|v|.$$

Multiplikation mit $A^{-1} \geq 0$ führt dann auf die Ungleichung $|v| \leq (1 - |\lambda|)A^{-1}D|v|$, die für $|\lambda| \geq 1$ zum Widerspruch $0 \leq |v| \leq 0$ mit $v \neq 0$ führt.

b) Zu $z > 0$ wird die reguläre Matrix $Z := \text{diag}(z_i) \geq 0$ definiert, es gilt dann $z = Z\mathbb{1}$. Damit ist

$$w = Az = Dz - Nz > 0 \iff NZ\mathbb{1} < DZ\mathbb{1} \iff Z^{-1}D^{-1}NZ\mathbb{1} < \mathbb{1}.$$

Die letzte Ungleichung für diese nichtnegative Matrix bedeutet aber gerade $\|Z^{-1}D^{-1}NZ\|_\infty < 1$, woraus insbesondere $\rho(Z^{-1}D^{-1}NZ) = \rho(D^{-1}N) < 1$ folgt. Nach Teil a) ist A daher M-Matrix. Mit $\alpha := \min_i w_i$ gilt laut Voraussetzung auch $Az = w \geq \alpha\mathbb{1}$. Aus $A^{-1} \geq 0$ folgt damit $\alpha A^{-1}\mathbb{1} \leq z \leq \|z\|_\infty \mathbb{1}$. Dies ist äquivalent zur Behauptung (2.6.8). ■

Für Matrizen mit (2.6.7) ist die strenge Diagonaldominanz (vgl. Numerik 1)

$$a_{ii} > \sum_{j \neq i} |a_{ij}| \quad \forall i,$$

ein spezielles M-Matrix-Kriterium und mit $z = \mathbf{1}$ in Lemma 2.6.3 enthalten. Praktische Bedeutung hat dieses Ergebnis, weil bei diagonaldominanten Matrizen der Gauß-Algorithmus ohne Pivotisierung durchführbar ist mit Aufwand $\mathcal{O}(N)$. Für die Differenzenmatrix A aus (2.6.5) liefert das Lemma für positives $q(x) \geq q^* > 0$, dass

$$\left(A\mathbf{1} \right)_{i=2}^{N-1} = \frac{1}{h^2} \left(-1 - \frac{h}{2}p_i + 2 + h^2q_i - 1 + \frac{h}{2}p_i \right)_{i=2}^{N-1} \geq (q_i)_i \geq q^*\mathbf{1} \quad (2.6.9)$$

gilt unter einfachen Voraussetzungen. In den inneren Gleichungen $2 \leq i \leq N-1$ fallen tatsächlich alle Terme außer q_i weg. In den Randgleichungen ($i = 1, N$) ist zusätzlich $h|p_i| \leq 2$ zu fordern. Die analoge Bedingung sorgt in den Nebendiagonalen auch für die M-Matrix-Eigenschaft $a_{i,i\pm 1} \leq 0$ und liefert damit über Lemma 2.6.3 eine Schranke für $\|A^{-1}\|_\infty$. Diese Normschranke $\|A^{-1}\| \leq \frac{1}{q^*}$ ist eine Stabilitätsaussage für das diskrete RWP (2.6.6), aus der im nächsten Satz eine Fehlerschranke für das Verfahren hergeleitet wird. Aus $A^{-1} \geq 0$ folgt außerdem, dass die Lösung y monoton von der Inhomogenität g abhängt, dass daher, z.B., $y \geq 0$ gilt für $g \geq 0, \alpha, \beta \geq 0$.

Satz 2.6.4 *Für die Koeffizientenfunktionen in (2.6.3) gelte $p, q, g \in C^2[a, b]$ und $q(x) \geq q^* > 0$. Die Schrittweite im Verfahren sei so, dass $h|p(x)| \leq 2 \forall x \in \Delta$ ist. Dann ist die Differenzenmatrix (2.6.5) eine M-Matrix, es gilt*

$$A^{-1} \geq 0 \quad \text{sowie} \quad \|A^{-1}\|_\infty \leq \frac{1}{q^*}.$$

Der globale Fehler der Näherung u_Δ aus (2.6.6) hat Ordnung 2, mit einer von Δ unabhängigen Konstanten C ist

$$\max_{x \in \Delta} |u(x) - u_\Delta(x)| \leq Ch^2(\|u^{(4)}\|_\infty + \|u^{(3)}\|_\infty).$$

Beweis Die Aussagen zu A folgen aus (2.6.9) und Lemma 2.6.3. Unter den Voraussetzungen des Satzes ist $u \in C^4[a, b]$. Mit (2.6.4) und Lemma 2.6.1 gilt daher für die Konsistenzfehler

$$\begin{aligned} |T_h(x_i)| &= \left| \frac{1}{h^2}(-u_{i-1} + 2u_i - u_{i+1}) + \frac{p_i}{2h}(u_{i+1} - u_{i-1}) + q_i u_i - g_i \right| \\ &\leq \underbrace{\left| -u''(x_i) + p_i u'(x_i) - q_i u(x_i) - g_i \right|}_{=0} + h^2(C_2|u^{(4)}(\xi_i)| + 4C_1|p_i u^{(3)}(\zeta_i)|) \\ &\leq C'h^2(|p_i| \|u^{(3)}\|_\infty + \|u^{(4)}\|_\infty). \end{aligned}$$

Nach (2.6.4) erfüllt der Gittervektor $U = (u_1, \dots, u_{N-1})^\top$ der Lösung u das System $AU = g + \tau$, $\tau = (T_h(x_1), \dots, T_h(x_N))^T$. Durch Subtraktion von (2.6.6) folgt

$$A(U - y) = \tau \quad \Rightarrow \quad \|U - y\|_\infty \leq \|A^{-1}\|_\infty \|\tau\|_\infty \leq \frac{1}{q^*} \|\tau\|_\infty \leq Ch^2 \sum_{k=3}^4 \|u^{(k)}\|_\infty. \quad \blacksquare$$

Dieses einfache Differenzenverfahren erzeugt also Näherungen an u von der (niedrigen) Konvergenzordnung zwei. Bei gewöhnlichen Randwertproblemen ist dies im Vergleich zur Genauigkeit

des Schießverfahrens zusammen mit Anfangswertverfahren höherer Ordnung uninteressant. Mit Hilfe der Identität $D_h := S_{h/2} - S_{-h/2} = 2 \sinh \frac{h}{2} D$ aus dem Beweis von Lemma 2.6.1 kann man aber 2. Ableitung formal exakt durch Differenzen-Entwicklungen darstellen

$$D^2 = \left(\frac{2}{h} \operatorname{arsinh} \frac{1}{2} D_h \right)^2 = \frac{1}{h^2} \left(D_h^2 - \frac{1}{12} D_h^4 + \dots \right).$$

Durch Partialsummen erhält man Differenzenapproximationen höherer Ordnung. So führen die in der letzten Gleichung angegebenen Terme auf die (wieder symmetrische) Differenzenformel

$$\begin{aligned} \frac{1}{h^2} \left(D_h^2 - \frac{1}{12} D_h^4 \right) u(x) &= \frac{1}{h^2} \left(S_{-h} - 2I + S_h - \frac{1}{12} (S_{-2h} - 4S_{-h} + 6I - 4S_h + S_{2h}) \right) u(x) \\ &= \frac{1}{12h^2} \left(-u(x-2h) + 16u(x-h) - 30u(x) + 16u(x+h) - u(x+2h) \right) \\ &= u''(x) - \frac{h^4}{90} u^{(6)}(\xi). \end{aligned} \tag{2.6.10}$$

In den randnahen Punkten x_1, x_N werden allerdings andere, unsymmetrische Formeln benötigt, um das Intervall $[a, b]$ nicht zu verlassen, ebenso bei u' . Da die Formel (2.6.10) Funktionswerte aus 5 Gitterpunkten benutzt, führt die Approximation statt der Tridiagonalmatrix (2.6.5) auf eine mit fünf besetzten Diagonalen. Auch gehen weitere Eigenschaften, wie Diagonaldominanz, verloren, sodass auch die einfache Stabilitätsaussage von Satz 2.6.4 nicht mehr möglich ist.

Bei Dgln ohne erste Ableitung u' (d.h. $p \equiv 0$) läßt sich dennoch ein $\mathcal{O}(h^4)$ -Verfahren ohne diese Nachteile angeben. Dabei wird ausgenutzt, dass hier bei Kenntnis von $u(x)$ auch $u''(x) = q(x)u(x) - g(x) =: f(x, u(x))$ bekannt ist. Damit ersetzt man den h^2 -Korrekturterm $\frac{1}{h^2} D_h^4 u(x) \cong h^2 u^{(4)}(x_i)$ in (2.6.10) durch $D_h^2 u''(x_i)$. Dieses *Mehrstellenverfahren* führt nun auf das Gleichungssystem

$$\frac{1}{h^2} (-y_{i-1} + 2y_i - y_{i+1}) + \frac{1}{12} (f(x_{i-1}, y_{i-1}) + 10f(x_i, y_i) + f(x_{i+1}, y_{i+1})) = 0,$$

$i = 1, \dots, N$, dessen Koeffizientenmatrix wieder tridiagonal und diagonaldominant ist für $q^* > 0$.

Abstrakter Hintergrund, gemeinsame Beweisprinzipien

Trotz sehr unterschiedlicher Strukturen verliefen die Konvergenzbeweise bei Anfangs- und Randwertproblemen sehr ähnlich. In beiden Fällen war eine Funktion $u \in V$ eines Funktionenraums V (z.B. $V = C^k[a, b]$) gesucht, die ein (System von) Gleichung(en) erfüllt,

$$u \in V : F(u) = 0, \quad F : V \rightarrow W. \quad \text{(Operatorgleichung)}$$

Zum Beispiel verwendet man beim AWP bzw. RWP die Definitionen

$$F(u) = \begin{pmatrix} u' - f(\cdot, u) \\ u(a) - u_0 \end{pmatrix}, \quad F(u) = \begin{pmatrix} -u'' + pu' + qu - g \\ u(a) - \alpha \\ u(b) - \beta \end{pmatrix}.$$

Durch Diskretisierung mit Gittern Δ wird dieses System approximiert durch eine Folge von endlichdimensionalen Problemen

$$u_\Delta \in V_\Delta : F_\Delta(u_\Delta) = 0, \quad F_\Delta : V_\Delta \rightarrow W_\Delta, \quad \text{(Diskretisierung)}$$

wobei bei feiner werdenden Gittern gilt $\dim(V_\Delta) \rightarrow \infty$ ($|\Delta| \rightarrow 0$). Da V_Δ in der Regel nicht in V enthalten ist, betrachtet man eine Restriktion $Q_\Delta : V \rightarrow V_\Delta$. In unseren Beispielen war $V_\Delta \sim \mathbb{R}^{N+1}$ ein Raum von Gittervektoren, bei dem als Restriktion die Werte auf dem Gitter $Q_\Delta u = (u(x_i))_{i=0}^N$ verwendet werden können. Die zentralen Begriffe sind jetzt sehr einfach zu beschreiben. Der Konsistenzfehler entsteht durch Einsetzen der (Restriktion der) exakten Lösung u in das Verfahren F_Δ , für Konsistenzordnung p heißt dies

$$T_\Delta := F_\Delta(Q_\Delta u), \quad \|T_\Delta\| = \mathcal{O}(|\Delta|^p), \quad |\Delta| \rightarrow 0. \quad \text{(Konsistenz)}$$

Stabilität bedeutet, dass sich kleine Änderungen im diskreten System nur wenig auf die Lösung auswirken. Diese Lipschitzstetigkeit der Umkehrabbildung F_Δ^{-1} läßt sich so fordern:

$$\|y_\Delta - z_\Delta\| \leq S \|F_\Delta(y_\Delta) - F_\Delta(z_\Delta)\| \quad \forall y_\Delta, z_\Delta \in V_\Delta, \quad \text{(Stabilität)}$$

vgl. (2.2.10). Entscheidend ist dabei, dass diese Abschätzung mit einer festen Konstanten S gleichmäßig für alle Gitter Δ gilt. Durch Kombination der Ungleichungen zur Konsistenz und Stabilität folgt dann direkt die Konvergenz (wähle $y_\Delta := u_\Delta$, $z_\Delta := Q_\Delta u$):

$$\|u_\Delta - Q_\Delta u\| \leq S \|F_\Delta(u_\Delta) - F_\Delta(Q_\Delta u)\| = S \|T_\Delta\| = \mathcal{O}(|\Delta|^p), \quad |\Delta| \rightarrow 0. \quad \text{(Konvergenz)}$$

3 Partielle Differentialgleichungen

3.1 Allgemeine Eigenschaften

Partielle Dgln sind Beschreibungen für Funktionen $u(x_1, \dots, x_n)$ von mehr als einer Veränderlichen. Dieser Problembereich ist sehr umfangreich und wird hier nur ansatzweise diskutiert mit linearen Problemen in zwei (Raum-) Dimensionen für Funktionen $u(x, y)$. Da bei partiellen Differentialgleichungen Ableitungen nach mehr als einer Veränderlichen auftreten, können sie i.A. nicht auf eine Standardform wie (2.1.1) gebracht werden. Geht man nur von Systemen erster Ordnung für Funktionen $u(x, y) := (u_1(x, y), \dots, u_n(x, y))^T$ aus, dann ist ein recht allgemeine Gestalt

$$u_x = A(x, y)u_y + B(x, y)u + g(x, y). \quad (3.1.1)$$

Charakteristische Eigenschaften dieser Dgl führen zu einer Einteilung in verschiedene *Typen*, nach denen sich auch die Wahl der zusätzlichen Randbedingungen zu richten hat. Der Typ wird vor allem durch die $n \times n$ -Koeffizientenmatrix A bestimmt. Zwei der Sonderfälle sind:

Definition 3.1.1 a) Das System (3.1.1) heißt elliptisch im Punkt (x, y) , wenn $A(x, y)$ keine reellen Eigenwerte besitzt.

b) Das System (3.1.1) heißt hyperbolisch in (x, y) , wenn $A(x, y)$ reell diagonalisierbar ist.

Für $n = 1$ ist jede reelle Gleichung hyperbolisch. Für $n > 2$ sind diese beiden Fälle nicht erschöpfend. Die wichtigsten Beispiele bei $n = 2$ Gleichungen für $u = (v, w)^T$ sind folgende

Beispiel 3.1.2 Die Cauchy-Riemann-Dgln der Funktionentheorie

$$v_x + w_y = 0, \quad w_x - v_y = 0 \quad \iff \quad u_x = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} u_y$$

sind überall elliptisch (EWe $\pm i$). Wenn u sogar zweimal differenzierbar ist, führt einmalige Differentiation der Gleichungen auf $v_{xx} + w_{yy} = 0$, $w_{xy} - v_{yy} = 0$, woraus die *Potentialgleichung* $v_{xx} + v_{yy} = 0$ für v (analog auch für w) folgt, eine einzelne partielle Dgl zweiter Ordnung.

Beispiel 3.1.3 Aus dem hyperbolischen System (EWe ± 1)

$$v_x + w_y = 0, \quad w_x + v_y = 0 \quad \iff \quad u_x = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix} u_y$$

wird analog die *Wellengleichung* $v_{xx} - v_{yy} = 0$ für v und w .

Beispiel 3.1.4 Beim System

$$v_x + w_y = 0, \quad w_x + v = 0 \quad \iff \quad u_x = - \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} u_y + \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} u$$

tritt der doppelte Eigenwert 0 auf, es ist nicht diagonalisierbar und führt auf die *Wärmeleitungsgleichungen* $v_{xx} - v_y = 0$ bzw. $w_{xx} - w_y = 0$.

Das letzte Beispiel ist durch die Fälle aus Definition 3.1.1 nicht abgedeckt, sein Typ ist *parabolisch*. Da viele Probleme direkt als Gleichung zweiter Ordnung auftreten, wird eine vollständige Typeneinteilung nur dafür formuliert.

Definition 3.1.5 Die allgemeine lineare Dgl in zwei Variablen x, y

$$a(x, y)u_{xx} + 2b(x, y)u_{xy} + d(x, y)u_{yy} + e(x, y)u_x + f(x, y)u_y = g(x, y)$$

heißt im Punkt (x, y)

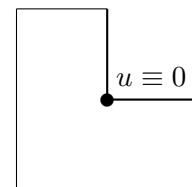
- a) elliptisch, wenn $a(x, y)d(x, y) - b^2(x, y) > 0$,
- b) hyperbolisch, wenn $a(x, y)d(x, y) - b^2(x, y) < 0$,
- c) parabolisch, wenn $a(x, y)d(x, y) - b^2(x, y) = 0$ und $\text{Rang} \begin{pmatrix} a & b & e \\ b & d & f \end{pmatrix} = 2$.

Der Typ entscheidet, welche zusätzlichen Randbedingungen mathematisch sinnvoll sind. Bei einer (überall) elliptischen Dgl, die auf einem Gebiet $\Omega \subseteq \mathbb{R}^2$ gilt, sind auf dem ganzen Rand von Ω Bedingungen an u und/oder u_x, u_y vorzuschreiben. Bei parabolischen und hyperbolischen dagegen nur auf einem Teil des Randes. Eine Eigenschaft der Lösung im nächsten Beispiel unterscheidet sich grundlegend vom gewöhnlichen Fall.

Beispiel 3.1.6 Auf dem L -Gebiet $\Omega := (-1, 1)^2 \setminus [0, 1]^2$ wird in Polarkoordinaten $x = r \cos \phi, y = r \sin \phi$ folgende Funktion betrachtet

$$u(r, \phi) = r^{2/3} \sin \frac{2\phi - \pi}{3}, \quad \frac{\pi}{2} \leq \phi \leq 2\pi.$$

Sie erfüllt die Potentialgleichung $u_{xx} + u_{yy} = 0$ auf Ω und hat glatte Randwerte, z.B. $u(x, y) = 0$ für $x = 0, y \geq 0$ sowie $x \geq 0, y = 0$. Die ersten Ableitungen von u sind jedoch nicht beschränkt in der einspringenden Ecke im markierten Nullpunkt $r \rightarrow 0$. In den übrigen Ecken treten schwächere Singularitäten auf.



Diese geringe Regularität von Lösungen führt auf neue Rahmenbedingungen für numerische Verfahren. Wenn die Koeffizienten einer gewöhnlichen Dgl genügend glatt sind, ist dies auch die Lösung und macht den Einsatz von Verfahren hoher Ordnung sinnvoll. Bei partiellen Dgln kann dagegen die Lösung schon bei ganz einfachen Problemen Singularitäten (von Ableitungen) besitzen. Da aus diesen und anderen, praktischen Gründen der Einsatz von Verfahren hoher Ordnung begrenzte Bedeutung hat, werden v.a. Verfahren der Ordnung zwei oder vier behandelt.

3.2 Differenzenverfahren für elliptische Randwertprobleme

3.2.1 Die Poissongleichung auf einfachen Gebieten

Es sei $\Omega \subseteq \mathbb{R}^2$ ein beschränktes Gebiet, d.h. eine offene und zusammenhängende Menge und $\Gamma = \partial\Omega$ sein Rand, der aus stückweise stetig differenzierbaren Kurven bestehen soll. Aus praktischen Gründen werden in diesem Paragraphen zunächst nur Polygone mit achsenparallelen oder

diagonalen Kanten behandelt. $\bar{\Omega} = \Omega \cup \Gamma$ ist der Abschluß von Ω . Mit einer stetigen Funktion $g(x, y)$ lautet die *Poissongleichung*

$$-u_{xx} - u_{yy} = g \quad \text{für } (x, y) \in \Omega. \quad (3.2.1)$$

Diese Gleichung ist die Normalform elliptischer Dgln. Sie beschreibt, z.B., die Wärmeverteilung in einem homogenen Körper mit Wärmequellen g . Die homogene Gleichung ($g \equiv 0$) heißt *Potentialgleichung*, ihre Lösungen harmonische Funktionen (\rightarrow Funktionentheorie). Für eine sinnvolle Problemstellung sind bei dieser Dgl auf ganz Γ Randwerte vorzugeben. Historisch unterscheidet man drei Arten von Randbedingungen, die den Funktionswert oder die Ableitung $\partial u / \partial \mathbf{n}$ in Richtung der äußeren Normalen \mathbf{n} enthalten. Dazu wird der Rand disjunkt zerlegt in $\Gamma = \Gamma_d \cup \Gamma_n \cup \Gamma_c$ und auf jedem Teil eine der folgenden Bedingungen gefordert mit stetigen Funktionen ϕ, ψ .

$$u(x, y) = \phi_d(x, y), \quad (x, y) \in \Gamma_d \quad (\text{Dirichlet - RB}) \quad (3.2.2)$$

$$\frac{\partial u}{\partial \mathbf{n}}(x, y) = \phi_n(x, y), \quad (x, y) \in \Gamma_n \quad (\text{Neumann - RB}) \quad (3.2.3)$$

$$\frac{\partial u}{\partial \mathbf{n}}(x, y) + \psi(x, y)u(x, y) = \phi_c(x, y), \quad (x, y) \in \Gamma_c \quad (\text{Cauchy - RB}). \quad (3.2.4)$$

In der genannten physikalischen Interpretation entspricht die Dirichlet-Randbedingung einer Temperaturvorgabe auf Γ_d , die Neumann-RB (mit $\phi_n \equiv 0$) einer isolierten Wand Γ_n , während die Cauchy-Bedingung sich als Wärmeabstrahlung durch Γ_c interpretieren läßt. Wenn jeweils nur eine der Randbedingungen auftritt, redet man vom Dirichletschen ($\Gamma = \Gamma_d$) bzw. Neumannschen ($\Gamma = \Gamma_n$) Randwertproblem.

Definition 3.2.1 Eine Funktion $u(x, y)$ heißt (klassische) Lösung des Randwertproblems (3.2.1), (3.2.2-3.2.4), wenn $u \in C^2(\Omega) \cap C(\bar{\Omega}) \cap C^1(\Omega \cup \Gamma_n \cup \Gamma_c)$ gilt und die Gleichungen (3.2.1), (3.2.2-3.2.4) erfüllt sind.

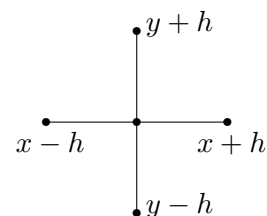
Beim Dirichletproblem existiert eine eindeutige Lösung, das Neumannproblem aber hat Lösungen nur unter der Voraussetzung $\int_{\Gamma} \phi_n ds = \int_{\Omega} g(x, y) dx dy$. Diese Lösungen sind dann auch nur bis auf eine Konstante eindeutig.

Wie im eindimensionalen Fall (§2.7) approximiert man die Ableitungen in der Dgl durch Differenzenquotienten. Mit Schrittweiten $h_x, h_y > 0$ gilt für $u \in C^4$ nach Lemma 2.6.1, dass

$$u_{xx}(x, y) = \frac{1}{h_x^2}(u(x - h_x, y) - 2u(x, y) + u(x + h_x, y)) + \mathcal{O}(h_x^2)$$

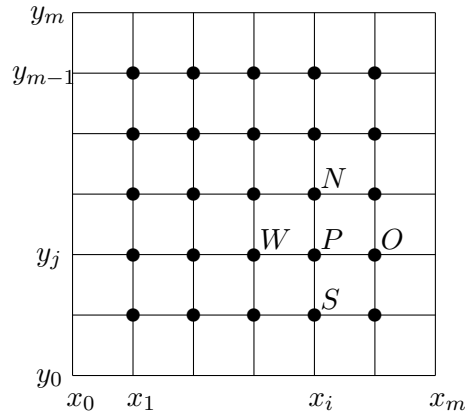
$$u_{yy}(x, y) = \frac{1}{h_y^2}(u(x, y - h_y) - 2u(x, y) + u(x, y + h_y)) + \mathcal{O}(h_y^2)$$

Bei Addition dieser beiden Näherungen werden zur Approximation der Poissongleichung im Punkt (x, y) fünf sternförmig verteilte Funktionswerte miteinander verknüpft. Für den Standardfall $h_x = h_y = h$ haben die vier Strahlen des Sterns alle die Länge h .



Als erstes Modellproblem wird das Einheitsquadrat $\Omega = (0, 1) \times (0, 1)$ betrachtet. Wie im eindimensionalen lassen sich mit einer Differenzenapproximation am einfachsten Funktionswerte eines äquidistanten Gitters verknüpfen. Dazu sei mit $h := 1/m$, $m \in \mathbb{N}$, das Gitter

$$\Delta_h := \{(x_i, y_j), i, j = 0, \dots, m\}, \text{ mit } x_i = y_i := ih, i = 0, \dots, m. \quad (3.2.5)$$



Da die Handhabung im zweidimensionalen schwieriger ist als früher, wird diese Indizierung der Gitterpunkte mit zwei Indizes auch für die Näherungswerte $v_{ij} = u_\Delta(x_i, y_j) \cong u(x_i, y_j)$ benutzt. Zur Beschreibung der Nachbarwerte eines Punktes $P = (x_i, y_j)$ ist es jedoch einprägsamer, Himmelsrichtungen zu verwenden. Zu einem inneren Gitterpunkt $P \in \Delta_h \cap \Omega$ sei also

$$v_P = v_{i,j} \quad \text{und} \quad v_N := v_{i,j+1}, \quad v_O := v_{i+1,j}, \quad v_S := v_{i,j-1}, \quad v_W := v_{i-1,j}, \quad (3.2.6)$$

wie in der obigen Skizze angedeutet. Im Punkt P wird die Poissongleichung $-u_{xx}(P) - u_{yy}(P) = g(P)$ approximiert durch die Differenzengleichung $-\frac{1}{h^2}(v_W - 2v_P + v_O) - \frac{1}{h^2}(v_S - 2v_P + v_N) = g_P$ mit $g_P := g(x_i, x_j)$, d.h.,

$$\begin{array}{c} -1 \\ | \\ -1 \text{---} 4 \text{---} -1 \\ | \\ -1 \end{array} \cdot \frac{1}{h^2} \qquad \frac{1}{h^2}(4v_P - v_N - v_O - v_S - v_W) = g_P. \quad (3.2.7)$$

Die Linearkombination in den Klammern wird gerne kompakt durch den links dargestellten *Fünf-Punkte-Stern* beschrieben. Bei der (homogenen) Potentialgleichung kann man diese Beziehung auch in der Form

$$v_P = \frac{1}{4}(v_N + v_O + v_S + v_W)$$

ausdrücken, die besagt, dass in jedem Punkt P aus dem Gitter v_P der Mittelwert der Funktionswerte seiner vier Nachbarpunkte ist (Interpretation: "aufgespanntes Gummnetz").

Dirichletproblem: Hier sind die Funktionswerte der Lösung auf dem ganzen Rand bekannt. Daher können die auf den Rand fallenden Variablen in (3.2.7) durch die vorgegebenen Werte von ϕ_d ersetzt werden. Aufgrund der einfachen Gestalt des Einheitsquadrats Ω ergeben sich so für die Werte auf den im Bild bei (3.2.5) durch einen Punkt markierten inneren $(m-1)^2$ Gitterpunkte gerade $(m-1)^2$ Gleichungen (3.2.7). In Anlehnung an die Formulierung des RWP's

(3.2.1),(3.2.2), kann man durch Einführung der Bezeichnungen

$$\begin{aligned}\Omega_h &:= \{(x_i, y_j) : i, j = 1, \dots, m-1\}, \\ \Gamma_h &:= \{(x_i, y_j) : i = 0 \vee i = m \vee j = 0 \vee j = m\},\end{aligned}\tag{3.2.8}$$

d.h., $\Omega_h \cup \Gamma_h = \Delta_h$, dieses Gleichungssystem ebenfalls in kompakter Form schreiben:

$$\begin{aligned}\frac{1}{h^2}(4v_{ij} - v_{i,j+1} - v_{i+1,j} - v_{i,j-1} - v_{i-1,j}) &= g_{ij}, & \text{für } (x_i, y_j) \in \Omega_h, \\ v_{ij} &= \phi_d(x_i, y_j) & \text{für } (x_i, y_j) \in \Gamma_h.\end{aligned}\tag{3.2.9}$$

Wie erwähnt, werden die trivialen Randgleichungen dazu benutzt, die Randwerte aus den Gleichungen auf Ω_h zu eliminieren (vgl. auch (2.6.6)). Das lineare Gleichungssystem besteht dann aus $(m-1)^2$ Gleichungen für ebenso viele Unbekannte. Bei den partiellen Dgln ist es angebracht, sich bei der Formulierung zunächst stärker auf die zugehörige lineare Abbildung L_Δ von Gitterfunktionen zu konzentrieren. Definiert man zur Gitterfunktionen $u_\Delta : \Omega_h \rightarrow \mathbb{R}$ die lineare Abbildung L_Δ punktweise durch

$$(L_\Delta u_\Delta)(x, y) := \frac{1}{h^2}(4u_\Delta(x, y) - u_\Delta(x, y+h) - u_\Delta(x, y-h) - \dots), \quad (x, y) \in \Omega_h,$$

mit Modifikationen in den randnahen Punkten, lautet das Gleichungssystem (3.2.9) einfach

$$L_\Delta u_\Delta = g_\Delta \quad \text{auf } \Omega_h,$$

mit der schon verwendeten Abkürzung $v_{ij} = u_\Delta(x_i, y_j)$. Die zugehörige Matrix hängt aber von der Nummerierung der bisher doppelt indizierten Unbekannten ab!

Bei *zeilenweiser Nummerierung* (lexikographische) des Gitters erhalten die drei Unbekannten v_W, v_P, v_O aufeinanderfolgende Indizes, während die Nummern von v_S, v_N genau um $m-1$ kleiner bzw. größer sind als die von v_P . Die Nummern der Gleichungen (3.2.7)/(3.2.9) seien die von v_P , sodass also das Hauptdiagonalelement immer $4/h^2$ ist. Dann ist die zu dieser Nummerierung gehörende Matrix A sehr regelmäßig aufgebaut, sie hat eine Block- und Bandstruktur mit Bandbreite $2m-1$ in der Form

$$A = \frac{1}{h^2} \begin{pmatrix} T & -I & & & \\ -I & T & -I & & \\ & -I & T & -I & \\ & & & \ddots & \\ & & & & -I & T \end{pmatrix}, \quad T := \begin{pmatrix} 4 & -1 & & & \\ -1 & 4 & -1 & & \\ & -1 & 4 & -1 & \\ & & & \ddots & \\ & & & & -1 & 4 \end{pmatrix}.\tag{3.2.10}$$

Dabei besitzen die Matrizen I und T jeweils die Größe $(m-1) \times (m-1)$, die Gesamtmatrix A die Größe $(m-1)^2 \times (m-1)^2$. Werden die Gitterpunkte dagegen wie ein *Schachbrett* schwarz-weiß eingefärbt, besitzen in (3.2.9) die Nachbarpunkte N, O, S, W immer die zum Punkt P entgegengesetzte Farbe. Nummeriert man dann zunächst weiße und dann schwarze Punkte, ergibt sich im Großen eine Blockstruktur der Form

$$\frac{1}{h^2} \begin{pmatrix} 4I & -M \\ -M^\top & 4I \end{pmatrix},$$

wobei M in jeder Zeile höchstens vier Einsen und sonst Nullen enthält. Für ungerades $m - 1$ haben die Blöcke $4I$ geringfügig unterschiedliche Größen. Weitere Eigenschaften dieser Matrizen werden später diskutiert. Lösungsverfahren für solche Gleichungssysteme, die die geringe Anzahl der nichttrivialen Matrixelemente nutzen, sind Thema der Vorlesung Numerik 2A.

Neumann-Randbedingung: Da jetzt der Wert der Normalenableitung vorgeschrieben ist, ist auch in der Randbedingung eine Differenzenapproximation erforderlich. Beim Einheitsquadrat sind die Normalen parallel zu den Achsen. Daher gilt entweder $\partial u / \partial \mathbf{n} = \pm u_x$ (auf der Ost- und Westseite), oder $\partial u / \partial \mathbf{n} = \pm u_y$ (auf der Nord- und Südseite). Als Beispiel werde die Approximation der Randbedingung $u_x(1, y) = \phi_n(1, y)$, $y \in (0, 1)$ betrachtet. Dabei bieten sich folgende Möglichkeiten an:

1. Verwendung des Gitters (3.2.5), Approximation der Randbedingung durch den einseitigen Differenzenquotienten mit Schrittweite h : Im Beispiel sei $P = (1 - h, y_j)$ ein dem Rand benachbarter Gitterpunkt. Die Randbedingung wird hier ersetzt durch $\frac{1}{h}(v_O - v_P) = \phi_n(O)$. Da die Variable v_O sonst nur noch im Differenzenstern von P auftaucht, kann die Bedingung dazu benutzt werden, diese Unbekannte sofort aus diesem Stern zu eliminieren. Dann hat das Gleichungssystem wieder die gleiche Struktur wie beim Dirichletproblem, z.B. (3.2.10), nur die Sterne in den Randnachbarpunkten sind abgeändert, im Beispiel ergibt sich statt (3.2.7) jetzt

$$\frac{1}{h^2}(3v_P - v_N - v_S - v_W) = g_p + \frac{1}{h}\phi_n(O).$$

Die Approximation der Randbedingung ist ungenauer und hat wegen der Verwendung unsymmetrischer Differenzenquotienten nur einen $\mathcal{O}(h)$ -Fehler. Es wird sich aber zeigen, dass eine geringere Genauigkeit in wenigen Punkten unschädlich ist, vgl. (3.2.23).

2. Der symmetrische Differenzenquotient $\frac{1}{2h}(u(x + h, y) - u(x - h, y))$ ist eine $\mathcal{O}(h^2)$ -Approximation für den Wert $u_x(x, y)$. Um ihn einsetzen zu können, muss der entsprechende Randpunkt mit ins Berechnungsgitter Ω_h aufgenommen werden, dies sei jetzt $P = (1, y_j) \in \Gamma_n$. Die approximierte Randbedingung

$$\frac{1}{2h}(v_O - v_W) = \phi_n(P) \tag{3.2.11}$$

verwendet dann den Punkt O außerhalb von Ω . Die Unbekannte v_O kann aber, wie vorher, aus dem Fünfpunktestern im Punkt P eliminiert werden. Dies führt für v_P auf den einseitigen Stern

$$\begin{array}{c} -1 \\ -2 \text{ --- } | \\ -1 \end{array} \left. \begin{array}{c} 4 \cdot \frac{1}{h^2} \\ \\ \end{array} \right\} \frac{1}{h^2}(4v_P - 2v_W - v_N - v_S) = g_p + \frac{2}{h}\phi_n(P).$$

Diese Methode ist besonders naheliegend, da (homogene) Neumann-Randbedingungen oft aufgrund von Symmetrieüberlegungen beim Dirichletproblem auftreten. Zum Beispiel gilt

für die Lösung des Problems

$$-u_{xx} - u_{yy} = 1 \text{ auf } (-1, 1) \times (-1, 1), \quad u = 0 \text{ auf dem Rand,}$$

die Symmetrieaussage $u(x, y) = u(\pm x, \pm y)$, also $u_x(0, y) \equiv 0, u_y(x, 0) \equiv 0$. Daher kann dieses Problem auf ein Viertel des ursprünglichen Quadrats reduziert werden, etwa das Einheitsquadrat, auf dessen linken und unteren Rand gilt dann die Neumann-Bedingung (3.2.3) mit $\phi_n \equiv 0$. Wird diese mit Hilfe von (3.2.11) diskretisiert, ergibt sich einfach eine Symmetriebedingung für $v = u_\Delta$, nämlich in $P = (0, y_j)$ die Gleichung $\frac{1}{h}(v_O - v_W) = 0 \iff v_O = v_W$ und in $P = (x_i, 0)$ die Bedingung $v_N = v_S$. Diese Approximation der Randbedingung führt daher auf das gleiche Ergebnis, das man durch Berücksichtigung der Symmetrie beim diskreten Problem (3.2.9) selbst erhalten würde. Die Symmetriebetrachtung reduziert dabei die Größe des Gleichungssystems auf ein Viertel.

Die Struktur der Matrix zu diesem reduzierten Problem

$$-u_{xx} - u_{yy} = g \quad \text{in} \quad \Omega = (0, 1) \times (0, 1), \quad (3.2.12)$$

$$u = 0 \quad \text{in} \quad \Gamma_d = \{(x, y) \in \bar{\Omega} : x = 1 \vee y = 1\}, \quad (3.2.13)$$

$$\frac{\partial}{\partial \mathbf{n}} u = 0 \quad \text{in} \quad \Gamma_n = \{(x, y) : x = 0, y \in [0, 1) \vee y = 0, x \in [0, 1)\} \quad (3.2.14)$$

ist ähnlich zu (3.2.10), wobei allerdings in den Nebendiagonalen auch der Wert $-2/h^2$ auftritt. Dies zerstört die Symmetrie der Matrix. Durch Skalierung der Gleichungen (Multiplikation mit $\frac{1}{2}, \frac{1}{4}$) kann die Symmetrie wiederhergestellt werden. Dies entspricht der Verwendung folgender Differenzensterne auf Γ_n (zusätzlicher Vorfaktor $1/h^2$):

$$\begin{array}{c} -\frac{1}{2} \left| \begin{array}{cc} & -1 \\ -1 & \end{array} \right. \\ 2 \left| \begin{array}{cc} -\frac{1}{2} & -1 \\ & -\frac{1}{2} \end{array} \right. \\ -\frac{1}{2} \left| \begin{array}{cc} & -\frac{1}{2} \\ & \end{array} \right. \end{array} \quad \begin{array}{c} -\frac{1}{2} \left| \begin{array}{cc} & -1 \\ -\frac{1}{2} & \end{array} \right. \\ 2 \left| \begin{array}{cc} -\frac{1}{2} & -1 \\ & -\frac{1}{2} \end{array} \right. \\ -\frac{1}{2} \left| \begin{array}{cc} & -\frac{1}{2} \\ & \end{array} \right. \end{array} \quad \begin{array}{c} -\frac{1}{2} \left| \begin{array}{cc} & -\frac{1}{2} \\ & \end{array} \right. \\ 1 \left| \begin{array}{cc} & -\frac{1}{2} \\ & \end{array} \right. \end{array}$$

Die Matrix ist auch etwas größer, unbekannt sind jetzt die Werte auf dem gesamten Gitter $\Omega_h := \{(x_i, y_j) : i, j = 0, \dots, m-1\}$. Zeilenweise Numerierung dieser Unbekannten $v_{ij} \cong u(x_i, y_j)$, in der Reihenfolge $v_{00}, \dots, v_{m-1,0}, v_{0,1}, \dots, v_{m-1,1}, \dots$, führt bei der Systemmatrix zu (3.2.12) auf die Gestalt

$$A = \frac{1}{h^2} \begin{pmatrix} \frac{1}{2}T & -E & & & \\ -E & T & -E & & \\ & -E & T & -E & \\ & & & \ddots & \\ & & & & -E & T \end{pmatrix}, \quad (3.2.15)$$

$$T := \begin{pmatrix} 2 & -1 & & & \\ -1 & 4 & -1 & & \\ & -1 & 4 & -1 & \\ & & & \ddots & \\ & & & & -1 & 4 \end{pmatrix}, \quad E := \begin{pmatrix} \frac{1}{2} & & & & \\ & 1 & & & \\ & & 1 & & \\ & & & \ddots & \\ & & & & & 1 \end{pmatrix}.$$

Die Größe von T und E ist $m \times m$, die der Gesamtmatrix A ist $m^2 \times m^2$.

Die in (3.2.10) und (3.2.15) beobachtete Symmetrie der Matrizen macht den Einsatz effizienterer Verfahren möglich und ist auch bei der Analyse (Stabilität) hilfreich.

3.2.2 Stabilität und Konvergenz

Für den Stabilitätsnachweis des Differenzenverfahrens sind die Inversen A^{-1} der Matrizen (3.2.10), (3.2.15) unabhängig von der Gittergröße gleichmäßig abzuschätzen. Dazu kann man zwei verschiedene Matrix-Eigenschaften einsetzen. Definitheit liefert Schranken in der Euklidnorm, die M-Matrix-Eigenschaft solche in der Maximumnorm. Symmetrie und Definitheit haben den praktischen Vorteil, dass dann effizientere Lösungsverfahren eingesetzt werden können (Cholesky-Zerlegung, spezielle Iterationsverfahren). Die M-Matrix-Eigenschaft gilt aber auch in allgemeineren Situationen und liefert aussagekräftigere Fehlerschranken in der Maximumnorm.

Die Differenzenmatrizen A bzw L_Δ aus §3.2.1 besitzen bei der Poissongleichung die Vorzeichenverteilung von M-Matrizen. Den *Testvektor* für die Schranke von Lemma 2.6.3 kann man bei L_Δ einfach angeben, er liefert folgende Stabilitätsaussage auf dem Einheitsquadrat.

Satz 3.2.2 *Für das diskrete Dirchletproblem (3.2.9) auf dem Einheitsquadrat mit homogenen Randbedingungen gilt $\|L_\Delta^{-1}\|_\infty \leq \frac{1}{8}$, d.h., die Lösung v erfüllt*

$$\max_{i,j=1}^{m-1} |v_{ij}| \leq \frac{1}{8} \max_{i,j=1}^{m-1} |g_{ij}|,$$

Beweis Differenzen bei konstanten Funktionen sind null, in x -Richtung hat die 2. Differenz bei der Funktion $(x - \frac{1}{2})^2$ den Wert $(x - h - \frac{1}{2})^2 - 2(x - \frac{1}{2})^2 + (x + h - \frac{1}{2})^2 = 2h^2$. Für die Testfunktion $z_\Delta(x, y) := r - (x - \frac{1}{2})^2 - (y - \frac{1}{2})^2$ gilt daher $L_\Delta z_\Delta(x, y) \geq 4$ mit Gleichheit in den randfernen Punkten. Außerdem ist $0 < z_\Delta \leq r$ in Ω_h für $r = \frac{1}{2}$. Lemma 2.6.3 zeigt so die Aussage. ■

Beim Problem (3.2.12) mit Neumann-Bedingung auf zwei Seiten des Quadrats ist die Konstante größer mit Wert 1, beim vollständigen Neumann-Problem ist die Matrix aber natürlich singulär.

Da die Konsistenz des Differenzenverfahrens (für C^4 -Lösungen) aus Lemma 2.6.1 folgt und die Stabilität gerade geklärt wurde, kann sofort auf die Konvergenz der Näherungen geschlossen werden, wenn das Randwertproblem hinreichend glatte Lösungen besitzt. Nach Beispiel 3.1 ist dies aber nicht immer der Fall. Daher wird zunächst eine abgeschwächte Konsistenzaussage behandelt, danach Möglichkeiten zur Beeinflussung des Fehlers durch Gitteranpassung oder Verbesserung des Verfahrens. Bei der folgenden Aussage wird der eigentlich nur auf Gitterfunktionen ($\Omega_h \subseteq \Omega$) definierte Differenzenoperator L_Δ auch auf normale Funktionen angewendet (formal wäre eine Gitter-Restriktion Q_Δ erforderlich, $L_\Delta Q_\Delta u$, vgl. §2.6).

Satz 3.2.3 *Es sei $u \in C^k(\Omega)$, $2 \leq k \leq 4$. Dann gilt für $(x, y) \in \Omega_h$ die Konsistenz-Aussage*

$$|(L_\Delta u)(x, y) + u_{xx}(x, y) + u_{yy}(x, y)| \leq \quad (3.2.16)$$

$$Ch^{k-2} \left(\max\left\{ \left| \frac{\partial^k}{\partial x^k} u(\xi, y) \right| : |\xi - x| < h \right\} + \max\left\{ \left| \frac{\partial^k}{\partial y^k} u(x, \eta) \right| : |\eta - y| < h \right\} \right).$$

Beweis Taylorentwicklung von u um (x, y) . ■

Wird im Satz die Lösung u der Poissongleichung eingesetzt, hat der lokale Fehler $T_\Delta := L_\Delta u - g_\Delta$ die Größenordnung $T_\Delta = \mathcal{O}(h^2)$, wenn $u \in C^4$ gilt. Wegen der Linearität von L_Δ führt die Subtraktion von (3.2.9), auf eine Gleichung für den Fehler $u_\Delta - u$

$$L_\Delta u_\Delta = g_\Delta \Rightarrow -T_\Delta = g_\Delta - L_\Delta u = L_\Delta u_\Delta - L_\Delta u = L_\Delta(u_\Delta - u). \quad (3.2.17)$$

Die tatsächliche Größe dieses globalen Fehlers liefert dann der Stabilitäts-Satz 3.2.2. Für die Formulierung wird für Gitterfunktionen g_Δ folgende Norm definiert,

$$\|g_\Delta\|_{\Delta, \infty} = \max_{(x, y) \in \Omega_h} |g_\Delta(x, y)|. \quad (3.2.18)$$

Für feiner werdendes Gitter approximiert diese die Funktionen-Normen,

$$\|g\|_\infty = \sup_{(x, y) \in \Omega} |g(x, y)|.$$

insbesondere gilt $\|Q_\Delta g\|_{\Delta, \infty} \leq \|g\|_\infty$ für die Gitterrestriktion. Durch Kombination der Ergebnisse aus Formel (3.2.17) und den Sätzen 3.2.2, 3.2.3, folgt die globale Fehleraussage

Satz 3.2.4 *Die Lösung u des Dirichletproblems auf dem Einheitsquadrat erfülle $u \in C^4(\bar{\Omega})$. Dann gilt für den Fehler der Näherungslösung u_Δ aus dem Differenzenverfahren (3.2.9) mit Schrittweite h in der Norm (3.2.18) mit der Konstanten $C_\infty = 1/96$ die Abschätzung*

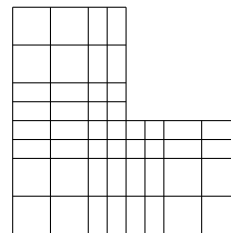
$$\|u_\Delta - u\|_{\Delta, \infty} \leq C_\infty h^2 (\|u_{xxxx}\|_\infty + \|u_{yyyy}\|_\infty).$$

Beweis Da die Werte von u und u_Δ auf dem Rand Γ_h übereinstimmen, treten in der Fehlergleichung (3.2.17) nur Beiträge aus dem Konsistenzfehler T_Δ auf, der mit Satz 3.2.3 (Konstante $C = 1/12$) abgeschätzt werden kann. Daher ist Satz 3.2.2 anwendbar (homogene Randbedingungen) und ergibt die Behauptung. ■

Für das gemischte Randwertproblem (3.2.12) erhält man nach dem selben Prinzip eine ähnliche Aussage mit etwas größeren Konstanten.

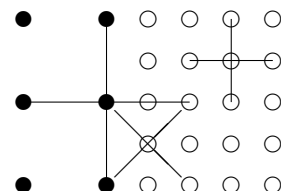
Nur auf den ersten Blick ist diese Fehleraussage zufriedenstellend. Denn in wichtigen Fällen erfüllt die Lösung u nicht die Voraussetzungen des Satzes, da dort von $g \in C^2(\bar{\Omega})$ nur auf $u \in C^4(\Omega)$, (Ω offen!) geschlossen werden kann (vgl. Beispiel 3.1.6). In diesem Beispiel wächst der lokale Konsistenzfehler (3.2.16) zum Rand hin stark an. Bei den gewöhnlichen Differentialgleichungen wurden solche Effekte durch Verkleinerung der lokalen Schrittweite kompensiert.

Überträgt man dieses Prinzip auf ein Produkt-Gitter (x_i, y_j) , indem man in bestimmten Teilen des x - und y -Intervalls die Punkte x_i, y_j dichter wählt, kann zwar an einer beliebigen Stelle im Gebiet ein feineres zweidimensionales Gitter erzeugt werden. Allerdings ergeben sich dann auch außerhalb dieses Bereichs Regionen mit kleinen Schrittweiten, die in x - und y -Richtung außerdem sehr unterschiedlich sind.

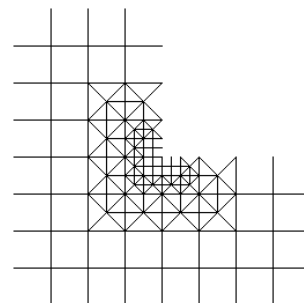


Auf einfache Weise läßt sich eine lokale Verfeinerung in einem kleinen Bereich dadurch erreichen, dass man ein Gitter mit halber Schrittweite $h/2$ in das grobe Gitter "einhängt". An einer Übergangsstelle wird das skizziert:

Im groben Gitter (ausgefüllte Punkte) wird der normale Fünf-punktestern mit Schrittweite h verwendet, der an den Übergangsstellen den ersten Punkt des feinen Gitters überspringt. In den Punkten des feinen Gitters (offene Kreise) wird der Stern mit Schrittweite $h/2$ eingesetzt. In den beiden Punkten des feinen Gitters, die keinen westlichen Nachbarn haben, kann ein diagonalaler Fünf-Punkte-Stern mit Schrittweite $h/\sqrt{2}$ benutzt werden (vgl. Lemma 3.2.6). Alle diese Sterne haben die Approximationsordnung 2. Das Verfahren führt nicht mehr auf eine symmetrische



Matrix, allerdings läßt sich die Inverse in der ∞ -Norm wieder mit Lemma 2.6.3 und dem Testvektor aus Satz 3.2.2 abschätzen. Die lokale Gitterverfeinerung kann in der Nähe einer Problemstelle natürlich mehrfach angewendet werden.



Beispiel 3.2.5 Dirichletproblem auf dem L-Gebiet $\Omega = (0, 1) \times (0, 1) \setminus [\frac{1}{2}, 1) \times [\frac{1}{2}, 1)$. Verwendung geschachtelter Gitter mit $h = 1/8, 1/16$ und $1/32$.

Wenn die (höheren) Ableitungen der Lösung zu groß werden, kann man dies also durch lokale Gitterverfeinerung kompensieren. Umgekehrt ist es bei genügend glatten Lösungen sinnvoll, den Aufwand durch Verfahren höherer Konvergenzordnung zu senken. Dazu kann man die genaue Kenntnis der Fehlerstruktur des oben erwähnten diagonalen Sterns verwenden.

Lemma 3.2.6 Für eine Funktion $u \in C^6(\bar{\Omega})$ gilt

$$\begin{aligned} & \frac{1}{2h^2} \left(4u(x, y) - u(x + h, y + h) - u(x + h, y - h) - u(x - h, y + h) - u(x - h, y - h) \right) \\ &= - \left(u_{xx} + u_{yy} + \frac{h^2}{12} (u_{xxxx} + 6u_{xxyy} + u_{yyyy}) \right) |_{(x,y)} + \mathcal{O}(h^4), \end{aligned}$$

Beweis Taylorentwicklung/Übungsaufgabe.

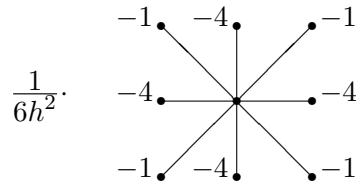
Durch Kombination mit dem normalen Fünf-Punkte-Stern, dessen Fehler (Satz 3.2.3)

$$\frac{h^2}{12} (u_{xxxx}(x, y) + u_{yyyy}(x, y)) + \mathcal{O}(h^4)$$

ist, kann bei der Poissongleichung ein h^2 -Fehlerterm der Gestalt

$$\frac{h^2}{12}(u_{xxxx} + 2u_{xxyy} + u_{yyyy}) = -\frac{h^2}{12}(g_{xx} + g_{yy}) \quad (3.2.19)$$

erzeugt werden, wobei der Laplace-Ausdruck $\frac{h^2}{12}(g_{xx} + g_{yy})$ wieder durch den Fünfpunktstern approximiert wird. Diese Linearkombination führt auf den folgenden *Neun-Punkte-Stern*, dessen Gewicht im Mittelpunkt $20/(6h^2)$ ist:

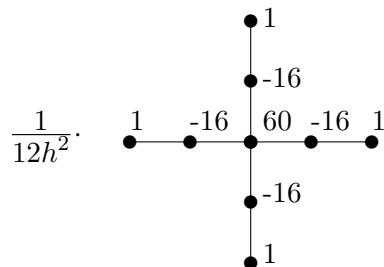


Eine einzelne Gleichung dieses *Mehrstellenverfahrens* (vgl. auch §2.6), im Punkt (x, y) wird der Übersicht halber durch Angabe der Gewichte in den Sternen abgekürzt:

$$\frac{1}{6h^2} \begin{bmatrix} -1 & -4 & -1 \\ -4 & 20 & -4 \\ -1 & -4 & -1 \end{bmatrix} u_{\Delta}(x, y) = \frac{1}{12} \begin{bmatrix} 1 & & \\ 1 & 8 & 1 \\ & 1 & \end{bmatrix} g(x, y), \quad (x, y) \in \Omega_h. \quad (3.2.20)$$

Der lokale Fehler hat die Form $\mathcal{O}(h^4)$ für $u \in C^6$, bei der Potentialgleichung mit $g \equiv 0$ besitzt das Verfahren sogar die Ordnung 6. Die Stabilität beim Dirichletproblem folgt wieder aus Lemma 2.6.3 wie in Satz 3.2.2 mit dem gleichen Ergebnis, $\|L_{\Delta}^{-1}\|_{\infty} \leq \frac{1}{8}$. Der wesentliche Vorteil des Verfahrens ist die Kompaktheit seines Differenzensterns, der bei Anwendung nur geringe Modifikationen gegenüber dem Fünfpunktstern erfordert (Randbedingungen, Matrixstruktur). Er beruht aber auf der speziellen Form der Poissongleichung und ist nur dort einsetzbar.

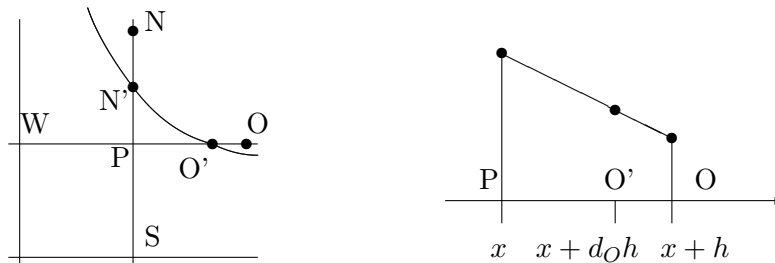
Eine allgemeinere Methode zur Verbesserung der Genauigkeit ist der Einsatz von höheren Differenzenapproximationen mit mehr als drei Punkten in jeder Ortsrichtung. Kombiniert man die symmetrische Approximation (2.6.10) für u_{xx} mit fünf Funktionswerten und eine entsprechende Formel für u_{yy} , führt dies bei der Poissongleichung auf folgenden Neun-Punkte-Stern



mit $\mathcal{O}(h^4)$ -Fehler. Die Verknüpfung von Werten mit einem maximalen Abstand von 4 Gitterpunkten führt auf erhebliche Probleme am Rand. Insbesondere bei nicht achsenparallelen Ecken können diese Schwierigkeiten auch nicht immer durch Verwendung unsymmetrischer Approximationen behoben werden. Dieses Verfahren wird daher nicht weiter betrachtet.

3.2.3 Allgemeinere Gebiete und Gleichungen

Der Fünfpunktstern ist auf allen Gebieten verwendbar, dessen Randpunkte auf ein geeignetes (rechteckiges, $h_x \neq h_y$) Gitter Δ_h fallen. Für nichtpolygonale ("krumme") Ränder kann man es aber meist nicht mehr so einrichten, dass alle Nachbarn eines inneren Gitterpunkts wieder Gitter- oder Randpunkte sind. Daher wird nun der Fall betrachtet, dass einer oder mehrere Nachbarn eines (randnahen) Gitterpunkts P außerhalb von Ω liegen wie in der gezeigten Skizze.



Der Rand Γ verlaufe also zwischen P und O durch den Punkt O' und (oder) zwischen P und N durch N' . Als allgemeinen Zugang kann man hier durch Taylor-Abgleich die Gewichte eines Fünfpunktsterns zu den Punkten P, N', O', S, W als Approximation an $-(u_{xx} + u_{yy})$ bestimmen. Auf das gleiche Ergebnis kommt man, wenn man Gitterwerte außerhalb von Ω , also v_N, v_O durch lineare Extrapolation des inneren Wertes v_P und der bekannten Randwerte $\phi_d(N'), \phi_d(O')$ (beim Dirichletproblem) approximiert. Sind dabei $d_N h, d_O h$, $0 \leq d_N, d_O \leq 1$ die Abstände der Randpunkte von P , dann ist

$$\phi_d(O') = d_O v_O + (1 - d_O) v_P \Rightarrow v_O = \frac{1}{d_O} (\phi_d(O') - (1 - d_O) v_P).$$

Durch Einsetzen in die zweite Differenz in x -Richtung, Anwendung auf u und Taylorentwicklung um $P = (x, y)$ führt dies auf

$$\begin{aligned} & -u(x - h, y) + (1 + \frac{1}{d_O})u(x, y) - \frac{1}{d_O}u(x + d_O h, y) = \\ & -u(W) + (1 + \frac{1}{d_O})u(P) - \frac{1}{d_O}u(O') = -\frac{h^2}{2}(1 + d_O)u_{xx}(P) + \frac{h^3}{6}(1 - d_O^2)u_{xxx} + \dots \quad (3.2.21) \end{aligned}$$

Division durch $h^2(1 + d_O)/2$ liefert eine Approximation der zweiten Ableitung. Man beachte aber den ungünstigeren Fehlerterm hu_{xxx} von nur erster Ordnung. Analoges gilt in y -Richtung. Im oben skizzierten Beispiel bekommt man so im Punkt P die Gleichung

$$\frac{1}{h^2} \left(\left(\frac{2}{d_O} + \frac{2}{d_N} \right) v_P - \frac{2}{1 + d_O} v_W - \frac{2}{1 + d_N} v_S - \underbrace{\frac{2}{d_O(1 + d_O)} v_{O'} - \frac{2}{d_N(1 + d_N)} v_{N'}}_{\text{bekannt}} \right) = g_P. \quad (3.2.22)$$

Da die Gewichte bei den Gitternachbarn v_S, v_W von den d -Werten abhängen, führt diese Approximation nicht mehr auf eine symmetrische, aber immer noch auf eine M-Matrix, da die Gleichung (3.2.22) weiterhin diagonaldominant ist. Die Normabschätzung in der ∞ -Norm aus §3.2.2 kann hier wiederholt werden, wobei die Konstanten aber vom Gebiet abhängen.

Die Taylorentwicklung (3.2.21) zeigt, dass die (Konsistenz-) Ordnung des jetzt asymmetrischen Sterns (3.2.22) nur noch eins ist. Durch eine präzisere Argumentation mit der M-Matrix-Eigenschaft bleibt aber die $\mathcal{O}(h^2)$ -Konvergenz des Gesamtverfahrens dennoch erhalten, da der größere lokale $\mathcal{O}(h)$ -Fehler nur am Rand auftritt. Dazu bezeichne L_Δ wieder die lineare Abbildung, die sich beim Dirichletproblem aus dem normalen Fünf-Punkte-Sterns in inneren Gitterpunkten und einer Gleichung der Form (3.2.22) in randnahen Punkten zusammensetzt. Das Bild $L_\Delta \mathbb{1}_\Delta$ (die Zeilensumme der Matrix) der konstanten Funktion $\mathbb{1}_\Delta(x, y) \equiv 1$ liefert nur in den randnahen Punkten einen nichtverschwindenden Beitrag, da die Dirichletrandwerte im Gleichungssystem $L_\Delta u_\Delta = g_\Delta$ auf der rechten Seite berücksichtigt werden (vgl.(3.2.10)). Speziell im Punkt P (3.2.22) bekommt man

$$(L_\Delta \mathbb{1}_\Delta)(P) = \frac{1}{h^2} \left(\frac{2}{d_O} + \frac{2}{d_N} - \frac{2}{1+d_O} - \frac{2}{1+d_N} \right) = \frac{2}{h^2} \left(\frac{1}{d_O(1+d_O)} + \frac{1}{d_N(1+d_N)} \right) > 0. \quad (3.2.23)$$

Bei einer anderen Anzahl von Randpunkten im Stern von P ist dieser Ausdruck zu modifizieren. Der Konsistenzfehler des Verfahrens hat hier die Form $T_\Delta = L_\Delta u - g_\Delta = hr_\Delta + h^2 t_\Delta$, wobei der h^2 -Anteil vom normalen Stern aus Satz 3.2.3 kommt, während hr_Δ nur Randfehler enthält. Im Randpunkt P liefert Gleichung (3.2.22) nach (3.2.21) hierzu den Beitrag

$$hr_\Delta(P) = \frac{h}{3} \left((1-d_O)u_{xxx}(x+\xi h, y) + (1-d_N)u_{yyy}(x, y+\eta h) \right), \quad \xi, \eta \in (0, 1).$$

Mit $K_j := \|\partial^j u / \partial x^j\|_\infty + \|\partial^j u / \partial y^j\|_\infty$, $j = 3, 4$, wird dieser Ausdruck mit dem Bild (3.2.23) des Testvektors $\mathbb{1}$ verglichen. Für den Fehlerbeitrag im Punkt P gilt

$$h|r_\Delta(P)| \leq K_3 \frac{h}{3} (1-d_O + 1-d_N) \leq K_3 \frac{h}{3} \left(\frac{1}{1+d_O} + \frac{1}{1+d_N} \right) \leq K_3 \frac{h^3}{6} (L_\Delta \mathbb{1}_\Delta)(P).$$

Mit analogen Abschätzungen gilt also auf dem gesamten Gitter punktweise $h|r_\Delta| \leq \hat{K}_3 h^3 L_\Delta \mathbb{1}_\Delta$. Aus der üblichen Fehlergleichung $L_\Delta(u_\Delta - u) = -T_\Delta = -(hr_\Delta + h^2 t_\Delta)$ folgen damit wegen $L_\Delta^{-1} \geq 0$ komponentenweise Abschätzungen, in jedem Gitterpunkt gilt

$$\begin{aligned} |u_\Delta - u| &= |L_\Delta^{-1}(hr_\Delta + h^2 t_\Delta)| \leq L_\Delta^{-1}(h|r_\Delta| + h^2|t_\Delta|) \leq L_\Delta^{-1}(\hat{K}_3 h^3 \cdot L_\Delta \mathbb{1}_\Delta + h^2|t_\Delta|) \\ &\leq (\hat{K}_3 h^3 + K_4 \frac{h^2}{12} \|L_\Delta^{-1}\|_\infty) \mathbb{1}_\Delta, \end{aligned}$$

also $\|u_\Delta - u\|_\infty \leq Ch^2$ für den globalen Fehler.

Für eine *allgemeine lineare, elliptische Dgl* 2.Ordnung

$$au_{xx} + 2bu_{xy} + du_{yy} + eu_x + fu_y + qu = g \quad (3.2.24)$$

($ad-b^2 > 0$) ist zu den Approximationen aus §2.6 und §3.2.1 noch die für die gemischte Ableitung u_{xy} nachzutragen. Diese läßt sich aus dem Produkt der ersten Differenzen in x - und y -Richtung zusammensetzen. Mit Schrittweite h hat der einfache Stern für u_{xy} die Form

$$\frac{1}{4h^2} \begin{array}{c} -1 & & 1 \\ * & \times & \\ 1 & & -1 \end{array} u(x, y) := \frac{u(x+h, y+h) + u(x-h, y-h) - u(x+h, y-h) - u(x-h, y+h)}{4h^2}$$

Dieser ist eine $O(h^2)$ -Approximation an $u_{xy}(x, y)$. Da die Nebendiagonalen dieses Sterns aber verschiedene Vorzeichen besitzen, müssen zum Erhalt der M-Matrix-Eigenschaft Linearkombinationen leicht verschobener Sterne eingesetzt werden [Hackbusch]. Zusätzlich sind bei allgemeinen Gebieten in jedem randnahen Punkt noch spezielle Randsterne analog zu (3.2.22) erforderlich. Das jetzt besprochene Verfahren ist wesentlich flexibler in Bezug auf die Geometrie des Problems, allerdings nicht in Bezug auf die Gestalt der Dgl (Ableitungen, Nichtlinearität).

3.3 Finite-Elemente-Verfahren für elliptische Probleme

3.3.1 Variationsformulierung

Nach einem fundamentalen physikalischen Prinzip nimmt jedes physikalische System den Zustand mit minimaler Gesamtenergie an. Dieser läßt sich dann *auch* durch eine Differentialgleichung beschreiben. Zunächst sei $u \in C^2[0, 1]$ die Lösung des folgenden Randwertproblems

$$-(p(x)u'(x))' + q(x)u(x) = g(x), \quad x \in (0, 1), \quad u(0) = u(1) = 0, \quad (3.3.1)$$

im eindimensionalen mit Koeffizienten-Funktionen $p \in C^1[0, 1]$, $q \in C[0, 1]$, $p(x) \geq p^* > 0$. Die Randbedingungen wurden zur Vereinfachung homogen gewählt, da dann die Räume

$$C_0^k[0, 1] := \{v \in C^k[0, 1] : v(0) = v(1) = 0\}, \quad k = 0, 1, 2,$$

linear sind. In (3.3.1) ist also $u \in C_0^2[0, 1]$ gesucht. Bildet man auf beiden Seiten der Dgl in (3.3.1) das Innenprodukt mit einem beliebigen Element $v \in C_0^2[0, 1]$, folgt durch partielle Integration

$$(g, v)_2 := \int_0^1 g(x)v(x)dx = \int_0^1 (-(pu')'v + quv)dx = \underbrace{-[pu'v]_0^1}_{=0} + \int_0^1 (pu'v' + quv)dx.$$

Der Ausdruck auf der rechten Seite ist aber schon für einmal differenzierbare Funktionen erklärt, er definiert folgende symmetrische *Bilinearform* $a : C_0^1[0, 1] \times C_0^1[0, 1] \rightarrow \mathbb{R}$:

$$a(u, v) := \int_0^1 (p(x)u'(x)v'(x) + q(x)u(x)v(x)) dx, \quad u, v \in C_0^1[0, 1]. \quad (3.3.2)$$

Das Innenprodukt $(g, v)_2$ ist ein beschränktes lineares Funktional, das mit f bezeichnet wird, $f : C[0, 1] \rightarrow \mathbb{R}$, $v \mapsto (g, v)_2$. Aus der obigen Umformung folgt also, dass die Lösung u des RWP's (3.3.1) auch folgende Gleichung erfüllt

$$a(u, v) = f(v) \quad \forall v \in C_0^1[0, 1]. \quad (3.3.3)$$

Für $p > 0, q \geq 0$ ist die Bilinearform $a(u, v)$ definit, stellt auf $C_0^1[0, 1]$ also ein Innenprodukt dar und $\sqrt{a(v, v)}$ eine Norm. Daher ist die Lösung u auch Minimalstelle eines *konvexen Funktionals*

$$J(u) = \min\{J(v) : v \in C_0^1[0, 1]\} \quad \text{mit} \quad J(v) := \frac{1}{2}a(v, v) - f(v). \quad (3.3.4)$$

Denn mit einem beliebigen $0 \neq v \in C_0^1[0, 1]$, $0 \neq t \in \mathbb{R}$, gilt für das Element $u + tv$:

$$\begin{aligned} J(u + tv) &= \frac{1}{2}a(u + tv, u + tv) - f(u + tv) & (3.3.5) \\ &= \frac{1}{2}a(u, u) - f(u) + t \underbrace{(a(u, v) - f(v))}_{=0} + \frac{t^2}{2} \underbrace{a(v, v)}_{>0} = J(u) + \frac{t^2}{2}a(v, v) \\ &> J(u). \end{aligned}$$

Die Definitheit von a ist essentiell: $a(v, v) > 0$, $v \neq 0$. Umgekehrt wäre für $a(u, v) - f(v) \neq 0$ auch u keine Minimalstelle. Daher stimmen die Lösungen des RWP's (3.3.1), von (3.3.3) und (3.3.4) überein, wenn sie in C_0^2 liegen. Es sei aber betont, dass bei weniger glatten Koeffizienten p, q, g das Minimum bei (3.3.4) in einer Funktion $u \notin C^2[0, 1]$ angenommen werden kann. Daher heißen Funktionen, die (3.3.3) oder (3.3.4) erfüllen, verallgemeinerte oder *schwache* Lösungen. Für eine korrekte Formulierung müssen dann andere Räume zugrundegelegt werden, vgl. (3.3.10).

Auch die Lösung eines partiellen, elliptischen Randwertproblems ist Minimalstelle eines *konvexen Funktionals*. Es sei wieder $\Omega \subseteq \mathbb{R}^2$ ein beschränktes Gebiet mit einem Rand Γ aus stückweise stetig differenzierbaren Kurven. Nach der Greenschen Formel gilt hier die Identität

$$\int \int_{\Omega} (u_x v_x + u_y v_y) dx dy = - \int \int_{\Omega} (u_{xx} + u_{yy})v dx dy + \oint_{\Gamma} \frac{\partial u}{\partial \mathbf{n}} v ds \quad (3.3.6)$$

für $u \in C^2(\bar{\Omega})$ und $v \in C^1(\bar{\Omega})$. Diese Relation tritt auf, wenn man Störungen $u + tv$ einer Lösung u betrachtet. Beim Dirichletproblem etwa müssen u und $u + tv$ die gleichen Randwerte besitzen, also v auf dem Rand verschwinden. Für dieses Problem entfällt das Randintegral in (3.3.6). Jetzt wird das Dirichletproblem bei der (etwas ergänzten) Poissongleichung mit $q \geq 0$ diskutiert,

$$-u_{xx} - u_{yy} + qu = g \quad \text{in } \Omega, \quad u = 0 \quad \text{auf } \Gamma. \quad (3.3.7)$$

Zu diesem Randwertproblem können schon für Funktionen aus dem Raum $C_0^1(\Omega)$, mit

$$C_0^k(\Omega) := \{u \in C^k(\Omega) : u|_{\Gamma} \equiv 0\}, \quad k = 0, 1,$$

die Bilinearform $a : C_0^1(\Omega) \times C_0^1(\Omega) \rightarrow \mathbb{R}$ und die Linearform $f : C_0^1(\Omega) \rightarrow \mathbb{R}$ gebildet werden,

$$\begin{aligned} a(u, v) &:= \int \int_{\Omega} (u_x v_x + u_y v_y + quv) dx dy, \\ f(v) &:= (g, v)_2 = \int \int_{\Omega} uv dx dy, \end{aligned} \quad u, v \in C_0^1(\Omega). \quad (3.3.8)$$

Wegen der Greenschen Formel (3.3.6) gilt also wieder:

$$u \in C_0^2(\Omega) \text{ löst (3.3.7)} \quad \Rightarrow \quad a(u, v) = f(v) \quad \forall v \in C_0^1(\Omega).$$

Neumann- und Cauchy-Randbedingungen lassen sich in einer ähnlichen Formulierung durch zusätzliche Randintegrale mit uv bzw. u in den Funktionalen a bzw. f berücksichtigen. Für das Dirichletproblem läßt sich zeigen, dass

$$\|v\|_{1,2} := \left(\int \int_{\Omega} (v_x^2 + v_y^2) dx dy \right)^{1/2} \quad (3.3.9)$$

auf $C_0^1(\Omega)$ eine Norm darstellt. Für $q \geq 0$ ist daher auch $a(v, v)$ wieder definit. Dies war die entscheidende Eigenschaft in der Entwicklung (3.3.5), die zeigt, dass auch die Lösung u von (3.3.7) als Minimalstelle (3.3.4) von $J(v) = \frac{1}{2}a(v, v) - f(v)$ charakterisiert werden kann mit den Bilinear- und Linearformen aus (3.3.8). Die Tatsache, dass dabei nur einmalige Differenzierbarkeit von u erforderlich ist, ist nach den Bemerkung zu Beginn über die Herkunft der Dgl zweiter Ordnung aus einem System 1. Ordnung, nicht so überraschend.

In diesem Minimalproblem dürfen im Prinzip alle Funktionen v betrachtet werden, für die das Funktional J sinnvoll definiert ist, also auch solche, bei der die ersten Ableitungen nur quadrat-integrierbar sind. Dazu gehören bei (3.3.8) die Elemente des Sobolev-Raums

$$W_0^1(\Omega) := \{u \in L_2(\Omega) : u_x, u_y \in L_2(\Omega), u|_\Gamma = 0\}. \quad (3.3.10)$$

Jede klassische Lösung des RWP (vgl. Def. 3.2.1) ist auch globale Minimalstelle von J in $W_0^1(\Omega)$. Diese Vergrößerung des Grundraums von C_0^1 auf W_0^1 hat eine entscheidende Bedeutung für die Konstruktion von Verfahren, da W_0^1 auch Funktionen enthält, die (nur) stückweise stetig differenzierbar sind.

3.3.2 Rayleigh-Ritz-Galerkin-Verfahren

Mit der Minimaleigenschaft der Lösung bekommt man auf elementare Weise Näherungsverfahren dadurch, dass man das Minimum des Energie-Funktional J nicht im Gesamtraum V ($= W_0^1(\Omega)$) sucht, sondern nur noch in einem *endlichdimensionalen* Unterraum $V_\Delta \subseteq V$. Für V_Δ kommen Polynome eines bestimmten Maximalgrads in Frage, besser sind aber die demnächst besprochenen Spline-Funktionen, die nur noch stückweise stetig differenzierbar sind. Es sei also

$$V_\Delta \subseteq V, \quad \dim V_\Delta = n \in \mathbb{N},$$

ein Unterraum und $\{B_i\}_{i=1}^n$ eine Basis von V_Δ . Die *Rayleigh-Ritz-Galerkin-Lösung* (RRG-Lösung) $u_\Delta \in V_\Delta$ des Minimalproblems (3.3.4) ist definiert durch

$$J(u_\Delta) = \min_{v \in V_\Delta} J(v) = \min_{v \in V_\Delta} \frac{1}{2}a(v, v) - f(v). \quad (3.3.11)$$

Ein allgemeines Prinzip erlaubt die Analyse dieses Verfahrens, wesentlich ist dabei die Vergleichbarkeit der von der Bilinearform a induzierten Norm und der Standard-(Sobolev-) Norm. Daher wird jetzt angenommen, dass Konstanten $\mu_0, \mu_1, M_1 > 0$ existieren mit

$$\mu_0 \|v\|_2^2 \leq \mu_1 \|v\|_{1,2}^2 \leq a(v, v) \leq M_1 \|v\|_{1,2}^2 \quad \forall v \in V. \quad (3.3.12)$$

Diese Voraussetzung ist in den betrachteten Beispielen (3.3.2) bzw. (3.3.8) erfüllt, wenn die Koeffizienten beschränkt sind und $p \geq p^* > 0$, $q \geq 0$ gilt.

Satz 3.3.1 *Unter der Voraussetzung (3.3.12) existiert eine eindeutige RRG-Lösung u_Δ von (3.3.11). Der Vektor $\eta = (\eta_j)_{j=1}^n$ der Koeffizienten von $u_\Delta = \sum_{j=1}^n \eta_j B_j$ löst das lineare System*

$$A\eta = \gamma, \quad A = \left(a(B_i, B_j) \right)_{i,j=1}^n, \quad \gamma = \left(f(B_i) \right)_{i=1}^n. \quad (3.3.13)$$

Die Matrix $A = (a_{ij})$ ist symmetrisch und positiv definit. Äquivalent zu (3.3.13) ist die Aussage

$$a(u_\Delta, v_\Delta) = f(v_\Delta) \quad \forall v_\Delta \in V_\Delta. \quad (3.3.14)$$

Bemerkung: Die Bedingung (3.3.14) heißt *Galerkin-Bedingung*. Sie definiert brauchbare Lösungen auch ohne Definitheit der Bilinearform a .

Beweis Es sei $v_\Delta = \sum_{j=1}^n \xi_j B_j \in V_\Delta$, $\xi := (\xi_i)$, beliebig. Dann gilt wegen der Bilinearität von $a(\cdot, \cdot)$ und wegen (3.3.12) zunächst für die quadratische Form

$$\xi^T A \xi = \sum_{i,j=1}^n \xi_i \xi_j a(B_i, B_j) = a\left(\sum_{i=1}^n \xi_i B_i, \sum_{j=1}^n \xi_j B_j\right) = a(v_\Delta, v_\Delta) \geq \mu_0 \|v_\Delta\|_2^2 > 0,$$

für $v_\Delta \neq 0$, d.h., $\xi \neq 0$. Also ist A definit und somit invertierbar. Daraus folgt wie eben durch quadratische Ergänzung

$$\begin{aligned} J(v_\Delta) &= \frac{1}{2} a\left(\sum_{i=1}^n \xi_i B_i, \sum_{j=1}^n \xi_j B_j\right) - f\left(\sum_{i=1}^n \xi_i B_i\right) = \frac{1}{2} \xi^T A \xi - \gamma^T \xi \\ &= \frac{1}{2} \left(\xi^T A \xi - 2(\gamma^T A^{-1}) A \xi + \gamma^T A^{-1} \gamma \right) - \frac{1}{2} \gamma^T A^{-1} \gamma \\ &= \frac{1}{2} \underbrace{(\xi - A^{-1} \gamma)^T A (\xi - A^{-1} \gamma)}_{\geq 0} - \frac{1}{2} \gamma^T A^{-1} \gamma. \end{aligned}$$

Das eindeutige Minimum liegt daher in $\xi = A^{-1} \gamma = \eta$, also $v_\Delta = u_\Delta$. Nach Definition entspricht (3.3.13) der Bedingung $a(u_\Delta, B_i) = f(B_i)$, $i = 1, \dots, n$. Da $\{B_i\}$ eine Basis von V_Δ darstellt (n.V.) ergibt sich (3.3.14). ■

Die Lösung u erfüllt (3.3.3), also $f(v) = a(u, v)$ auch mit $v = u$. Für beliebige v gilt daher

$$\begin{aligned} J(v) - J(u) &= \frac{1}{2} a(v, v) - f(v) - \frac{1}{2} a(u, u) + f(u) = \frac{1}{2} a(v, v) - a(u, v) + \frac{1}{2} a(u, u) \\ &= \frac{1}{2} a(v - u, v - u). \end{aligned}$$

Da die Minimalstellen der Funktionen $v \mapsto J(v) - J(u)$ und $v \mapsto J(v)$ übereinstimmen (auch in V_Δ), bekommt man daraus eine andere Charakterisierung der RRG-Lösung, die direkt auf Fehlerschranken führen. u_Δ ist nämlich die Bestapproximierende an u im Raum V_Δ , gemessen in der "a-Norm".

Satz 3.3.2 Für die RRG-Lösung u_Δ aus (3.3.11) gilt

$$\mu_1 \|u_\Delta - u\|_{1,2}^2 \leq a(u_\Delta - u, u_\Delta - u) = \inf_{v_\Delta \in V_\Delta} a(v_\Delta - u, v_\Delta - u) \leq M_1 \inf_{v_\Delta \in V_\Delta} \|v_\Delta - u\|_{1,2}^2.$$

Der Satz liefert durch Einsetzen einer beliebigen Funktion $v \in V_\Delta$, z.B., einer Interpolierenden der Lösung u , direkt eine globale Fehlerschranke in der 1,2-Norm (die Ableitung von u_Δ). Schranken für den Abstand $\text{dist}_{1,2}(u, V_\Delta) = \inf_{v_\Delta \in V_\Delta} \|v_\Delta - u\|_{1,2}$ werden später für konkrete Räume V_Δ hergeleitet.

3.3.3 Spline-Räume, *Finite Elemente*

Die Effizienz und Flexibilität des RRG-Verfahrens hängt wesentlich von der richtigen Wahl des endlichdimensionalen Raums V_Δ ab. Bei partiellen Problemen sind Polynome zu unhandlich. Für unregelmäßige Gebiete ist dagegen ein Zugang sehr anpassungsfähig, bei dem man das Gesamtgebiet in kleine, einfache *Elemente* zerlegt. Dabei ist es hilfreich, dass das Minimalproblem (3.3.4) schon für stückweise differenzierbare Funktionen erklärt ist.

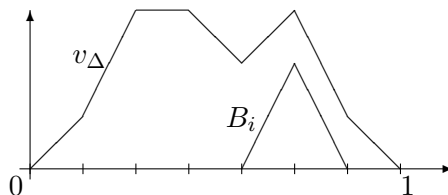
Beim gewöhnlichen Randwertproblem (3.3.1) auf $[0, 1]$ führt die Unterteilung des Intervalls auf ein Gitter $\Delta : 0 = x_0 < x_1 < \dots < x_m = 1$, mit Schrittweiten $h_i = x_{i+1} - x_i$. Die einfachste Funktionenklasse in $W_0^1[0, 1]$ sind die stückweise linearen Funktionen (lineare Splines, vgl. Numerik I, §2.2). Unter Berücksichtigung der Randbedingungen sei daher

$$V_\Delta := \{s \in C_0^0[0, 1] : s|_{[x_i, x_{i+1}]} \in \Pi_1, i = 0, \dots, m-1\} \quad (3.3.15)$$

der Raum aller stetigen Funktionen mit Randbedingung null, die in jedem Teilintervall Polynome vom Grad 1 ($\in \Pi_1$) sind. Eine einfache Basis dieses Raums ist durch die *Dach-Funktionen*

$$B_i(x) := \begin{cases} (x - x_{i-1})/h_{i-1}, & x \in [x_{i-1}, x_i) \\ (x_{i+1} - x)/h_i, & x \in [x_i, x_{i+1}) \\ 0, & \text{sonst} \end{cases}, \quad (3.3.16)$$

$i = 1, \dots, m-1$, gegeben, vgl. die Skizze.



Diese B_i bilden eine Kardinalbasis, da für $x = x_k$ gilt $u_\Delta(x_k) = \sum \eta_j B_j(x_k) = \eta_k$. Wichtiger ist aber die Lokalität dieser Basis. Daher sind nämlich nur wenige Matrixelemente $a(B_i, B_j)$ in (3.3.13) von Null verschieden. Dies reduziert den Rechenaufwand bei Aufstellung und Auflösung des Gleichungssystems. Als Beispiel werde die einfache Dgl $-u'' = g$ mit $p \equiv 1, q \equiv 0$, mit der Bilinearform $a(u, v) = \int_0^1 u'(x)v'(x)dx$ betrachtet. Für die Basisfunktionen gilt nach (3.3.16)

$$B_i'(x) := \begin{cases} 1/h_{i-1}, & x \in [x_{i-1}, x_i) \\ -1/h_i, & x \in [x_i, x_{i+1}) \\ 0, & \text{sonst} \end{cases}.$$

Daher ist $a(B_i, B_j) = 0$ für $|i - j| > 1$. Für $j \in \{i-1, i, i+1\}$ folgt

$$a(B_i, B_j) = \int_{x_{i-1}}^{x_i} \frac{1}{h_{i-1}} B_j' dx - \int_{x_i}^{x_{i+1}} \frac{1}{h_i} B_j' dx = \begin{cases} -\frac{1}{h_{i-1}}, & j = i-1 \\ \frac{1}{h_{i-1}} + \frac{1}{h_i}, & j = i \\ -\frac{1}{h_i}, & j = i+1 \end{cases}.$$

Das Verfahren arbeitet daher bei der Approximation der zweiten Ableitung wie das Differenzenverfahren, denn es führt auf das Gleichungssystem $(\eta_i = u_\Delta(x_i), \text{ s.o.})$

$$-\frac{1}{h_i}(\eta_{i+1} - \eta_i) + \frac{1}{h_{i-1}}(\eta_i - \eta_{i-1}) = \int_{x_{i-1}}^{x_{i+1}} g(x)B_i(x)dx, \quad i = 1, \dots, m-1.$$

Für das RRG-Verfahren im Raum (3.3.15) erwartet man daher auch einen Fehler der Größenordnung h^2 . Dieser läßt sich mit Hilfe von Satz 3.3.2 auf den Approximationsfehler $\inf\{\|u - v_\Delta\| : v_\Delta \in V_\Delta\}$ zurückführen. Der Fehler der linearen Interpolierenden ist auf einzelnen Teilintervallen gegeben durch die Standarddarstellung des Interpolationsfehlers für Polynome (Numerik I, 2.1.17). Die Interpolierende einer Funktion $v \in V$ ist eine Restriktion $Q_\Delta v$ im Sinne von §2.

Lemma 3.3.3 *Zu einer Funktion $v \in C_0^2[0, 1]$ sei $Q_\Delta v := \sum_{i=1}^{m-1} v(x_i)B_i$ die Interpolierende aus V_Δ . Dann gilt mit $H := \max_i h_i$ die Fehlerschranke*

$$\|(Q_\Delta v - v)^{(k)}\|_2 \leq \left(\frac{H}{\pi}\right)^{2-k} \|v''\|_2, \quad k = 0, 1.$$

Schätzt man das Infimum in Satz 3.3.2 durch Einsetzen der Interpolierenden $v = Q_\Delta u$ ab, bekommt man mit dem letzten Lemma die Fehlerschranke,

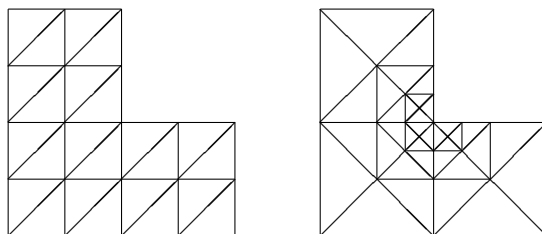
$$\|(u_\Delta - u)'\|_2 \leq \sqrt{\frac{M_1}{\mu_1} \frac{H}{\pi}} \|u''\|_2.$$

Mit einem zusätzlichen Argument ("Nitsche-Trick") kann man aus dieser $\mathcal{O}(H)$ -Schranke für die Ableitung bei vielen Randwertproblemen die globale Fehlerschranke

$$\|u_\Delta - u\|_2 \leq CH^2 \|u''\|_2 \leq \tilde{C}H^2 \|g\|_2$$

für die Funktion herleiten. Vorteilhaft ist hier, dass nur die L_2 -Integrierbarkeit von u'' benötigt wird im Vergleich zur Voraussetzung $u \in C^4$ bei Differenzenverfahren. RRG-Verfahren höherer Ordnung erhält man durch Verwendung von Splines aus Polynomen höheren Grads.

Die Möglichkeit, mit stückweise linearen Ansatzfunktionen zu arbeiten, ist bei partiellen RWPen wichtiger und von großer praktischer Bedeutung. Sehr einfach sind Zerlegungen des Gebiets Ω durch *Dreieckgitter*. Damit ist das RRG-Verfahren viel flexibler als Differenzenverfahren bei der Approximation komplizierterer Gebiete und der lokalen Gitterverfeinerung. Beispiele für regelmäßige bzw. lokal verfeinernde Triangulierungen beim L-Gebiet sind:



Dabei sind nur Triangulierungen aus offenen, paarweise disjunkten Dreiecken T_i zulässig,

$$\Delta = \{T_i\}_{i=1}^m, \quad \bar{\Omega} = \bigcup_{i=1}^m \bar{T}_i,$$

bei denen alle Ecken eines Dreiecks auf Ecken seiner Nachbardreiecke oder den Rand fallen. Für den Raum aller stetigen, stückweise linearen Funktionen auf Δ ,

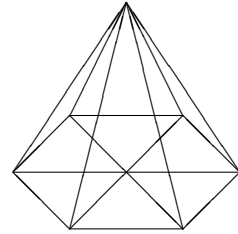
$$V_\Delta := \{s \in C_0(\Omega) : s|_{T_i} \text{ linear}\}, \quad (3.3.17)$$

existiert wieder eine Basis von Dachfunktionen. Es sei $\{P_i\}_{i=1}^n \subseteq \Omega$ die Menge aller Eckpunkte von Dreiecken (Knoten) im Innern des Gebiets. Dann bilden die Funktionen, $i = 1, \dots, n$,

$$B_i \in V_\Delta, \quad \text{mit } B_i(P_j) = \delta_{ij}, \quad j = 1, \dots, n,$$

eine Basis von V_Δ . Ein Beispiel einer solchen Basisfunktion bei sechs in einer Ecke zusammenstoßenden Dreiecken ist rechts gezeigt. Wesentliche Merkmale dieser Basis sind

wieder die Kardinalität, $u_\Delta = \sum_{i=1}^n u_\Delta(P_i)B_i$, und die Lokalität der Basis, da B_i nur auf den Dreiecken T_j mit $P_i \in \bar{T}_j$ nicht verschwindet. Die Lokalität bedeutet bei mehrdimensionalen Problemen einen großen Effizienzgewinn, da nur die wenigen nichttrivialen Integrale $a(B_i, B_j)$ in der Matrix A zu berechnen sind, und man bei diesen für $i \neq j$ nur über zwei Dreiecke integriert.



$$a(B_i, B_j) = \sum_{P_i \in T_k \wedge P_j \in T_k} \int \int_{T_k} [(B_i)_x(B_j)_x + (B_i)_y(B_j)_y + qB_iB_j] dx dy.$$

Für die praktische Berechnung der Integrale $a_{ij} = a(B_i, B_j)$, $\gamma_i = f(B_i)$, arbeitet man am einfachsten alle Dreiecke T_i ab und addiert die einzelnen Teilintegrale in a_{ij} , γ_i auf. Die Matrix A ist hier wieder dünn besetzt und symmetrisch, positiv definit. Daher können effiziente (insbesondere Iterations-) Verfahren zur Lösung des Gleichungssystems (3.3.13) eingesetzt werden.

Zur Fehlerabschätzung der zugehörigen RRG-Lösung nach Satz 3.3.2 läßt sich wieder eine Schranke für den Interpolationsfehler lokal herleiten. Dabei ergibt sich eine verwertbare Aussage allerdings nur, wenn die Dreiecke T_k nicht zu spitz werden.

Lemma 3.3.4 *Zu einer Funktion $v \in C_0^2(\Omega)$ sei $Q_\Delta v := \sum_{i=1}^n v(P_i)B_i$ die Interpolierende aus V_Δ , (3.3.17). In der Dreieckerlegung des Gebiets sei H die längste Dreiecksseite und α der kleinste Innenwinkel eines Dreiecks T_i , $i = 1, \dots, m$. Dann gilt die Fehlerschranke*

$$\|Q_\Delta v - v\|_{1,2} \leq C \frac{H}{\sin \alpha} \|v\|_{2,2},$$

$$\text{mit } \|v\|_{2,2}^2 := \sum_{j+k=2} \left\| \frac{\partial^2}{\partial x^j \partial y^k} v \right\|_2^2.$$

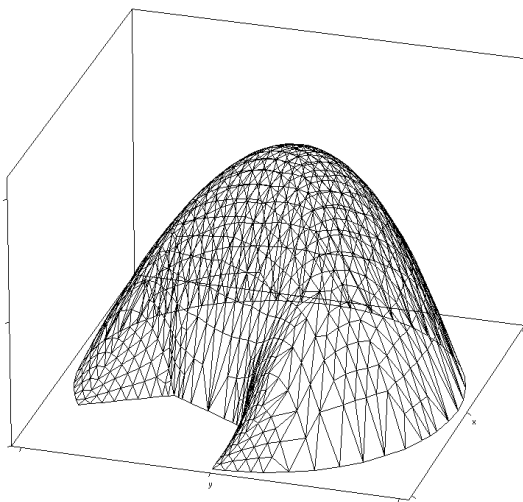
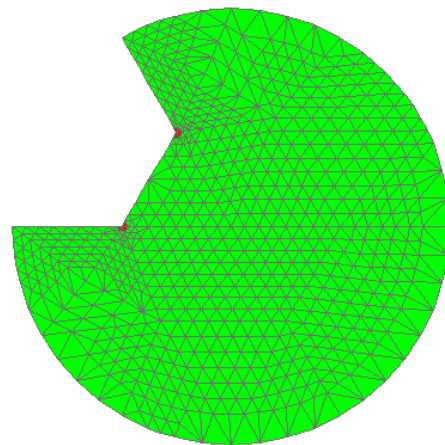
Damit kann man wie im gewöhnlichen Fall in zwei Schritten einen Fehler $\mathcal{O}(H/\sin \alpha)\|u\|_{2,2}$ in den ersten Ableitungen der Näherung u_Δ und dann eine Schranke $\mathcal{O}(H^2/\sin \alpha)\|u\|_{2,2}$ für den Fehler $\|u_\Delta - u\|_2$ der Funktionswerte herleiten. Zusätzlich kann man unter einfachen Annahmen an die Koeffizienten zumindest für konvexe Gebiete die Abschätzung $\|u\|_{2,2} \leq C\|g\|_2$ zeigen.

Daher hat man auch im partiellen Fall die H^2 -Konvergenz unter schwächeren Voraussetzungen als bei Differenzenverfahren.

Allerdings ist bei diesen starken Aussagen zu beachten, dass sie für das exakte RRG-Verfahren zutreffen, bei dem alle Integrale $a(B_i, B_j)$, $f(B_i) = (g, B_i)_2$ ohne Fehler bestimmt werden. In der Praxis berechnet man jedoch zumindestens die Integralanteile, in denen allgemeine Koeffizienten (hier q, g) auftreten, mit Hilfe von Quadraturformeln. Um auch für diese approximierten RRG-Verfahren einen $\mathcal{O}(H^2)$ -Fehler garantieren zu können, müssen doch wieder stärkere Annahmen an die Koeffizienten, also auch an die Regularität der Lösung gemacht werden. Verfahren höherer Konvergenzordnung ergeben sich bei Verwendung von stückweise stetigen Polynomen höheren Grades auf den Dreieckzerlegungen.

Zu Finite-Elemente-Verfahren wurden auch Schätzer entwickelt, die den lokalen Fehler einer Näherungslösung u_Δ schätzen. Damit ist dann eine schrittweise Anpassung des Gitters möglich, welche kritische Stellen im Gebiet weitgehend selbstständig identifiziert anhand einer aktuellen Näherung und dort für eine genügend feine Unterteilung sorgt.

Beispiel 3.3.5 Adaptive Gitterkonstruktion beim Dirichletproblem mit der Poissongleichung für $g \equiv 1$. Das "Pacman"-Gebiet hat 2 einspringende Ecken, wo das Verfahren (bei genügend scharfer Toleranzvorgabe) das Gitter tatsächlich über mehrere Schritte verfeinert. Die Bilder zeigen einen Gittergraphen der Lösung für $tol = 10^{-3}$ und den Fehlerindikator (grün/rot) nach 4 Verfeinerungen mit 639 inneren Gitterpunkten.

Lösung u_Δ 

Fehlerindikator auf Gitter

Index

- Adams-
 - Bashforth-Verfahren, 37
 - Moulton-Verfahren, 38
- adaptive Integration, 12
- Anfangswert, 47
- Anfangswertproblem, 19, 24
- Anlaufrechnung, 39
- asymptotische Entwicklung, 14

- Bilinearform, 72, 73
- Bulirsch-Folge, 16
- Butcher-Tableau, 25, 32

- Dahlquist, 42
- Diagonaldominanz, 55
- Differentiation, 3, 17
- Differenzen
 - Formel, 18
 - Gleichung, 24, 44
 - Quotienten, 53, 61
- Dirichlet-Problem, 61, 62, 67, 73
- Drei-Term-Rekursion, 10
- Dreieck
 - Schema, 15
- Dreieck-Gitter, 77

- Einschritt-Verfahren, 24, 28
- Einschrittverfahren, 28
- Entwicklung
 - asymptotische, 34, 43
- Euler-McLaurin-Summenformel, 14
- Euler-Verfahren, 24
- Extrapolations-Verfahren, 15

- Fehler
 - Gleichverteilung, 12
 - Schätzung, 11, 12, 31, 34
 - globaler, 30, 42, 56, 67, 71, 75, 77
 - lokaler, 26, 38, 39, 45, 54, 67, 69, 71

- Fixpunkt, 20
- FSAL, 33
- Fundamentalsystem, 21, 47
- Funktional
 - konvexes, 72, 73
 - lineares, 72

- Galerkin-Bedingung, 75
- Gauß
 - Quadratur, 11
- Gauß-
 - Algorithmus, 51, 56
- Gesamt-Fehler, 18
- Gewichtsfunktion, 3, 10
- Gragg-Bulirsch-Stoer-Verfahren, 45
- Green-Funktion, 54
- Greensche Formel, 73
- Gronwall-Lemma, 21
 - diskret, 28

- Innenprodukt, 9
- Integration, 3
- Interpolations-
 - Polynom, 8, 17
 - Lagrange-, 3
 - Newton-, 9, 17

- Keller-Speicher, 13
- Kondensation, 51, 52
- Konsistenz, 26, 37, 38, 41, 67
 - Ordnung, 26
- Konvergenz, 6

- L-Gebiet, 60
- Lösungsfächer, 32
- Lösungsschar, 47, 48
- Lipschitzbedingung, 20, 22, 28, 42

- M-Matrix, 55, 70
- Maschinengenauigkeit, 18

- Mehrschrittverfahren, 37
 - lineare, 41
- Mehrstellenverfahren, 57, 69
- Mehrzielmethode, 50
- Mittelpunktregel
 - explizite, 43
- Mittelwertsatz, 4
- Neumann-Randbedingung, 64
- Neumann-Reihe, 55
- Neville-Algorithmus, 15
- Newton
 - Cotes-Formeln, 5, 6, 43
- Newton-Verfahren, 48, 50
- Numerierung
 - lexikographisch, 63
 - Schachbrett-, 63
- Ordnung, 15, 17
- Ordnungs-
 - Barriere, 42
 - Bedingungen, 27, 41
 - Steuerung, 40, 45
- Orthogonalisierung, 9
- Part.Differentialgleichung
 - elliptische, 59–61, 73
 - hyperbolische, 59, 60
 - parabolische, 60
- Poissongleichung, 61, 73
- Polynom
 - Lagrange-, 40
 - Legendre-, 10, 11
 - Newton-, 40
 - Orthogonal-, 9, 11
- Prädiktor-Korrektor-Verfahren, 39
- Quadratur, 3
 - Fehler, 3, 13
 - Formel, 3
 - Gauß-, 9
 - Gewichte, 3, 10
 - iteriert, 6, 9
 - Ordnung, 4, 9
- Randbedingung, 53
- Randwertproblem, 19, 46, 53, 61
- Rechenaufwand, 7
- Rechteckregel, 5
- Reihenentwicklungen, 10
- Restglied, 3, 9, 11
- Restriktion, 66, 77
- Richardson-Extrapolation, 15, 34, 45
- Richtungsfeld, 19
- Rodriguez-Formel, 10
- Romberg-Folge, 15
- Rundungsfehler, 18, 35
- Runge-Kutta-Verfahren
 - allgemeine, 25
 - eingebettete, 32
 - explizite, 26
 - klassisches, 25
 - stetige, 34
- Satz von
 - Rolle, 10, 17
- Schrittweiten-Steuerung, 30, 32, 40, 45
- Simpsonregel, 5, 11
 - iteriert, 7
- Sobolev-Raum, 74
- Spline-Funktionen, 77
- Stabilität, 28, 29, 42, 66, 69
- Stabilität, 56
- Stern
 - Fünf-Punkte-, 62, 68
 - Neun-Punkte-, 69
- Stufen, 25
- Trapezregel, 5, 11
 - iteriert, 7, 11, 14
- Treppenmatrix, 51
- Triangulierung, 77
- Tridiagonalsystem, 54

Variationsgleichung, 47

Verfahren

explizite, 37

implizite, 38

Wärmeleitungsgleichung, 59

Wellengleichung, 59