Numerik IIB

— Numerische Verfahren für Differentialgleichungen —

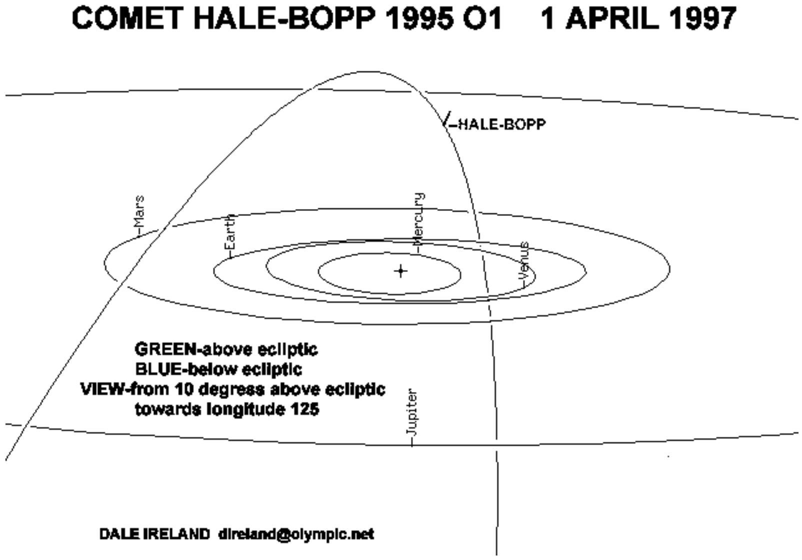
Bernhard Schmitt

Sommersemester 1997

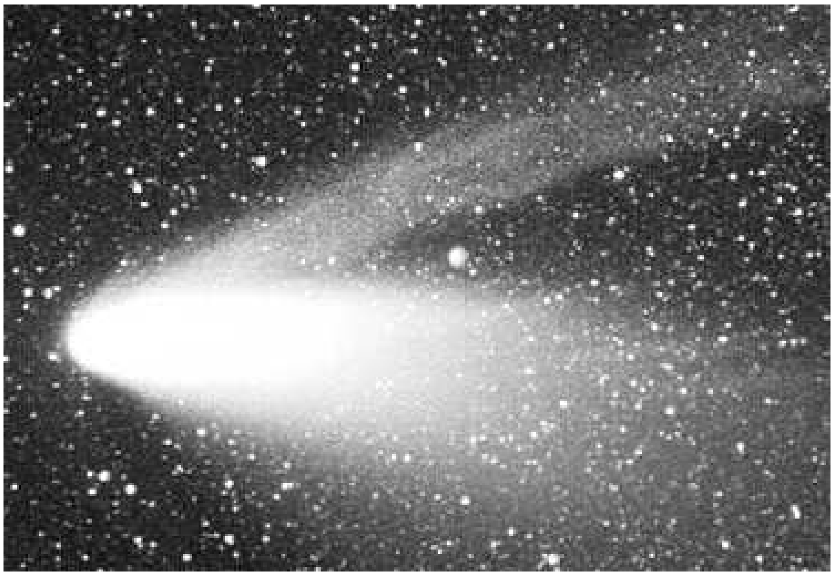
Inhaltsverzeichnis

Einleitung

In den (Natur-) Wissenschaften lassen sich viele der beobachteten Effekte durch Differentialgleichungen unterschiedlicher Art beschreiben. Ein gutes aktuelles Beispiel ist der Komet Hale-Bopp, da man sich hier tatsächlich für unterschiedliche Aspekte interessiert.



Zur Bestimmung seiner Bewegung durch das Sonnensystem reicht es aus, ihn als punktförmige Masse im Schwerefeld von Sonne, Planeten usw. zu betrachten. Die Bahn (in Abhängigkeit von der Zeit) wird dann durch ein System von gewöhnlichen Differentialgleichungen beschrieben, wie sie im ersten Teil der Vorlesung behandelt werden.



Die sichtbare Gestalt des Kometen am Himmel kommt dagegen 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 im zweiten Teil der Vorlesung besprochen.

Einige durch Differentialgleichungen beschriebene Problemstellungen

Einfache Modelle physik., biolog., chem.,.. Prozesse; z.B.,

a) *Massenbewegung unter Gravitation.*

Vereinfachende Annahmen: Vakuum, Massen starr, punktförmig, konzentriert,...

a1) Wurf auf Erde: Flugbahn $u(t) = \begin{pmatrix} x(t) \\ y(t) \end{pmatrix}$, Newtonsches Gesetz:

$mu''(t) = m \begin{pmatrix} x''(t) \\ y''(t) \end{pmatrix} = -m \begin{pmatrix} 0 \\ g \end{pmatrix}$. Explizite Lösung Wurfparabel

'Anfangswertproblem': Start-Ort &-Geschwindigkeit gegeben $\Rightarrow$ $\exists_1$ Lösung!

'Randwertproblem': Start-& Zielpunkt gegeben, Lösung?

a2) Astronomie (Raumflug, n-Körper-Problem): nicht explizit lösbar, numerische Verfahren erforderlich (hoher Genauigkeit)

b) *Wachsende Population in Nährmedium* (Bakterien, Protozoen,..)

b1) Kleiner Behälter, gute Durchmischung:

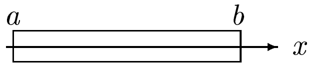
Population $p(t)$ zur Zeit t wächst nach Gesetz

$$\frac{dp}{dt} = \alpha p - \beta p^2 = \alpha(1 - \frac{\beta}{\alpha} p)p.$$

Dabei: α = Vermehrungsrate, βp^2 = Wettbewerb. Lösung: logistische Kurve

Gewöhn-
liche
Dgln

b2) Rohrförmiger Behälter, p abhängig von x und t:



Wanderung (Diffusion) im Rohr prop. zu p_{xx}, für $p(t,x)$ gilt

$$\frac{\partial p}{\partial t} = \alpha p - \beta p^2 + \gamma \frac{\partial^2 p}{\partial x^2}$$

Zusatzinformation: Startpopulation $p(0,x)$, Rohrende: $p_x(t,a) = p_x(t,b) = 0$, d.h., Anfangswerte bzgl. t, Randwerte bzgl. x.

<<<<————————————————————>>>>

Partielle Dgln 2.Ordnung, Einteilung nach Typen (hier in Orts-Gebiet $D \subset \mathbf{R}^2$)

a) Parabolische Dgln: Ausgleichsvorgänge $u(t,x,y)$, z.B., Wärmeleitungs-Gl.

$$u_t = \alpha(u_{xx} + u_{yy}).$$

Sinnvolle Aufgabe: Anfangswerte in $t = 0$, Randwerte auf ∂D

b) Elliptische Dgln: Gleichgewichtszustände $u(x,y)$, z.B., Potentialgleichung

$$u_{xx} + u_{yy} = 0.$$

Sinnvolle Aufgabe: Randwerte auf ∂D

c) Hyperbolische Dgln: Schwingungen/Wellen $u(t,x,y)$, z.B., Wellen-Gl.

$$u_{tt} = \alpha(u_{xx} + u_{yy}).$$

Sinnvolle Aufgabe: Anfangswerte u, u_t in $t = 0$, Randwerte auf ∂D

Parti-
elle
Dgln

1 Gewöhnliche Differentialgleichungen

1.1 Theoretische Grundlagen

Im allgemeinsten Fall sind gewöhnliche Differentialgleichungen Systeme von nichtlinearen Gleichungen, die verschiedene Ableitungen einer Vektor-Funktion $u : \mathbf{R} \to \mathbf{R}^n$ miteinander verknüpfen. Meist betrachtet man aber zur Standardisierung explizite Gleichungen der Form

$$u'(x) = \frac{du}{dx} = f(x, u(x)), \quad \text{kurz} \quad u' = f(x, u), \tag{1.1.1}$$

wobei $f : \mathbf{R} \times \mathbf{R}^n \to \mathbf{R}^n$ eine gegebene stetige Funktion ist. Als Lösung einer solchen Differentialgleichung (Dgl) ist eine stetig differenzierbare Vektorfunktion $u : [a, b] \to \mathbf{R}^n$, also eine Raumkurve gesucht, die für jeden x-Wert aus $[a, b]$ im Punkt $u(x)$ den Ableitungswert $f(x, u(x))$ annimmt, sich also in das durch $f(x, y)$, $x \in [a, b], y \in \mathbf{R}^n$ gegebene *Richtungsfeld* einpaßt.

Die ausführliche Bezeichnung für (1.1.1) ist *System von Dgln erster Ordnung*, da nur die erste Ableitung u' der gesuchten Funktion auftaucht. Jede Dgl m-ter Ordnung,

$$z^{(m)}(x) = g(x, z(x), z'(x), \ldots, z^{(m-1)}(x))$$

für eine Funktion $z : \mathbf{R} \to \mathbf{R}$ (zur Vereinfachung) kann durch Einführung von Hilfsfunktionen

$$u_1(x) := z(x),\ u_2(x) := z'(x),\ \ldots,\ u_m(x) := z^{(m-1)}(x),$$

in die Standardform (1.1.1) mit $n = m$ Komponenten gebracht werden, denn dann ist

$$u' = \begin{pmatrix} u_1' \\ \vdots \\ u_{m-1}' \\ u_m' \end{pmatrix} = \begin{pmatrix} u_2 \\ \vdots \\ u_m \\ g(x, u_1, \ldots, u_m) \end{pmatrix}.$$

Aus diesem Grund werden im folgenden meist Systeme erster Ordnung behandelt, und nur in wenigen Fällen der häufig auftretende Fall von Dgln zweiter Ordnung.

Die Dgl (1.1.1) definiert durch jeden Punkt $(x, y) \in \mathbf{R} \times \mathbf{R}^n$ eine eigene Lösungskurve. Zur Festlegung einer eindeutigen Lösung sind weitere Bedingungen erforderlich. Dazu werden zwei Standardfälle behandelt. Beim *Anfangswertproblem* (AWP) wird der Funktionswert an einer Stelle $x_0 \in [a, b]$ vollständig vorgeschrieben (im folgenden oBdA $x_0 = a$)

$$u(a) = u_0 \in \mathbf{R}^n. \tag{1.1.2}$$

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 (evtl. nichtlinearen) Gleichungssystems

$$r(u(a), u(b)) = 0, \quad r : \mathbf{R}^n \times \mathbf{R}^n \to \mathbf{R}^n. \tag{1.1.3}$$

Es folgen einige grundlegende theoretische Aussagen zu den genannten Problemen.

Für das Anfangswertproblem (1.1.1),(1.1.2) gibt es eine sehr einfache und allgemeine Existenzaussage im Satz von Picard-Lindelöf. Er beruht auf dem Übergang von der Dgl durch Integration zur äquivalenten *Integralgleichung*

$$u(x) = u_0 + \int_a^x f(t, u(t))\, dt, \tag{1.1.4}$$

der auch für die Konstruktion numerischer Verfahren grundlegend ist.

Satz 1.1.1 *Die Funktion f sei auf dem Streifen $\Omega := \{(x, y) : x \in [a, b], y \in \mathbf{R}^n\}$ über dem endlichen Intervall $[a, b]$ definiert und stetig. Außerdem gelte dort die Lipschitzbedingung*

$$\|f(x, u) - f(x, v)\| \le L\|u - v\| \quad \forall x \in [a, b], u, v \in \mathbf{R}^n, \tag{1.1.5}$$

mit einer Lipschitzkonstanten $L \ge 0$. Dann existiert zu jedem $u_0 \in \mathbf{R}^n$ genau eine Lösung $u \in (C^1[a, b])^n$ des Anfangswertproblems (1.1.1),(1.1.2). Diese Lösung hängt stetig vom Anfangswert ab. Für zwei Lösungen $u, v \in (C^1[a, b]))^n$ mit

$$\begin{array}{c} u' = f(x, u) \\ u(x_0) = u_0 \end{array}, \qquad \begin{array}{c} v' = f(x, v) \\ v(x_0) = v_0 \end{array},$$

gilt dabei die Schranke

$$\|u(x) - v(x)\| \le e^{L|x-a|}\, \|u_0 - v_0\| \quad \forall x \in [a, b]. \tag{1.1.6}$$

Beweis a) Bei (1.1.4) handelt es sich um ein Fixpunktproblem $u = Tu$ mit der Abbildung

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

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 (1.1.5) läßt sich direkt die Kontraktivität von T ableiten, denn für beliebige $u, v \in (C[a, b])^n$ gilt

$$\begin{aligned} \|Tu - Tv\|_L &= \max_{x\in[a,b]} e^{-L|x-a|} \| \int_a^x \Big(f(t, u(t)) - f(t, v(t))\Big) dt\| \\ &\le \max_{x\in[a,b]} e^{-L|x-a|} \int_a^x \|f(t, u(t)) - f(t, v(t))\|\, dt \\ &\le L \max_{x\in[a,b]} e^{-L|x-a|} \int_a^x e^{L|t-a|} e^{-L|t-a|} \|u(t) - v(t)\|\, dt \\ &\le L \max_{x\in[a,b]} e^{-L|x-a|} \int_a^x e^{L|t-a|}\, dt\ \|u - v\|_L \\ &\le \Big(1 - e^{-L|b-a|}\Big) \|u - v\|_L. \end{aligned}$$

Da der Vorfaktor kleiner als eins ist, ist T eine Kontraktion und nach dem Banachschen Fixpunktsatz existiert ein eindeutiger Fixpunkt $u = Tu \in (C[a,b])^n$. Mit u ist auch die Funktion $f(x,u(x)) = u'(x)$ stetig.

b) Aus der Differenz der beiden für u und v geltenden Gleichungen (1.1.4) folgt wie eben

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

Für die skalare Funktion $\delta(x) := \|u(x)-v(x)\|$ gilt also die Ungleichung $\delta(x) \le \|u_0 - v_0\| + L\int_a^x \delta(t)\,dt$. Mit dem folgenden Gronwall-Lemma ergibt sich daraus die Schranke (1.1.6). ∎

Im Beweis wurde die folgende Aussage benutzt.

Lemma 1.1.2 (Gronwall-Lemma) *Die Funktion $\delta \in C[a,b]$ erfülle die Ungleichung*

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

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

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

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 $a < x_1 \le b$ die erste Stelle mit $\delta(x_1) = \gamma(x_1)$, insbesondere gelte also $\delta(x) \le \gamma(x)\ \forall a \le x \le x_1$. Dann folgt in x_1 mit

$$\delta(x_1) \le \alpha + L\int_a^{x_1} \delta(t)\,dt < \alpha + \varepsilon + L\int_a^{x_1} \gamma(t)\,dt = \gamma(x_1)$$

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,

$$u'(x) = A(x)u(x) + g(x), \quad \text{d.h.,}\ f(x,y) = A(x)y + g(x), \tag{1.1.7}$$

mit einer Matrixfunktion $A(x) \in \mathbf{R}^{n\times n}$ und einer Vektorfunktion $g(x) \in \mathbf{R}^n$, da hier explizite Lösungsdarstellungen gelten. Mit Hilfe eines Fundamentalsystems der Dgl (1.1.7), d.h. einer regulären Matrixfunktion $W(x) \in \mathbf{R}^{n\times n}$, die die Dgl

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

erfüllt, ist die Lösung des Anfangswertproblems (1.1.7), (1.1.2) explizit gegeben durch

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

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

Satz 1.1.3 *Die Lösung des linearen Anfangswertproblems (1.1.7),(1.1.2) mit einer beschränkten, stetigen Matrixfunktion $A(x)$ ist*

$$u(x) = Y(x)u_0 + Y(x)\int_a^x Y(t)^{-1}g(t)\,dt, \tag{1.1.8}$$

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

Explizit angebbar ist das Fundamentalsystem bei konstanter Matrix A. Über den Ansatz $u(t) = e^{\lambda x}y$ läßt sich die homogene Dgl auf das Eigenwertproblem $(A - \lambda I)y = 0$ zurückführen. Bei einer diagonalisierbaren Matrix A mit Eigen-Wert-/-Vektor-Paaren (λ_j, y_j) ist $u(x) = \sum_{j=1}^n \alpha_j e^{\lambda_j x} y_j$. Äquivalent dazu ist analog zum skalaren Fall ($n = 1$) die Darstellung

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

Dies vereinfacht die Lösungsformel (1.1.8) weiter zu

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

Die Beziehung dieser expliziten Formeln zu Satz 1.1.1 ist folgende. Bei (1.1.7) gilt die Lipschitzbedingung (1.1.5) mit $L = \max_x \|A(x)\|$, aus der für die Differenz von Lösungen folgt

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

also $\|Y(x)\| \le e^{L|x-a|}$. Im Fall konstanter Koeffizienten reduziert sich dies auf die Ungleichung $\|e^{tA}\| \le e^{t\|A\|}$, $t \ge 0$, die auch direkt aus der Reihendarstellung folgt.

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

$$u'(x) = A(x)u(x) + g(x),\ x \in (a,b), \quad R_0u(a) + R_1u(b) = d, \tag{1.1.9}$$

mit festen $n \times n$-Matrizen R_0, R_1 und $d \in \mathbf{R}^n$. Dabei ist $\operatorname{rang}(R_0, R_1) = n$ vorauszusetzen. Im Gegensatz zum Anfangswertproblem können hier alle Lösbarkeitsfälle auftreten. Die Lösung u wird durch die explizite Formel (1.1.8) mit einem unbekannten Startvektor $\eta \in \mathbf{R}^n$ 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 ergibt sich das lineare Gleichungssystem

$$\Big(R_0Y(a) + R_1Y(b)\Big)\eta = d - R\gamma(b) \tag{1.1.10}$$

für η. Wie aus der Linearen Algebra bekannt, gilt die Fallunterscheidung,

$$\text{Wenn}\quad R_0Y(a) + R_1Y(b) \quad \begin{cases} \text{regulär:} & \text{es existiert eine eindeutige Lösung} \\ \text{singulär:} & \begin{cases} \text{es gibt keine Lösung} \\ \text{es gibt unendlich viele Lösungen} \end{cases} \end{cases}.$$

Es gibt bei Randwertproblemen also keine zum Satz 1.1.1 vergleichbare allgemeine Existenz- oder Eindeutigkeitsaussage.

1.2 Einschrittverfahren für Anfangswertprobleme

1.2.1 Herleitung

Die Lösung der Dgl $u' = f(x,u)$ ist diejenige Kurve $(x,u(x))$, die dem Richtungsfeld $(x,y,f(x,y))$ folgt. Ein elementares Approximationsverfahren beruht darauf, daß man vom Startwert $u_0 = u(x_0)$ 'ein Stück weit' in Richtung der Ableitung $u'(x_0) = f(x_0,u_0)$ geht und dann dort dieses Verfahren wiederholt (Eulerscher Polygonzug, s.u.). Einen allgemeinen Zugang zu solchen schrittweise vorgehenden numerischen Integrationsverfahren bekommt man über die Formel (1.1.4). Dazu wird im Punkt $\bar{x}$ eine *Schrittweite* $h > 0$ gewählt und die Dgl von $\bar{x}$ bis $\bar{x}+h$ integriert. Dies ergibt

$$\frac{1}{h}[u(\bar{x}+h) - u(\bar{x})] = \frac{1}{h}\int_{\bar{x}}^{\bar{x}+h} f(t,u(t))\,dt = \int_0^1 f(\bar{x}+hs, u(\bar{x}+hs))\,ds. \tag{1.2.1}$$

In numerischen Verfahren wird das Integral auf der rechten Seite approximiert, z.B., durch eine geeignete Quadraturformel. In diesem Abschnitt werden Näherungen der Form

$$\int_0^1 f(\bar{x}+hs, u(\bar{x}+hs))\,ds = \underbrace{f_h(\bar{x},u(\bar{x}))}_{\text{Näherung}} + \underbrace{T_h(\bar{x})}_{\text{Fehler}}\,. \tag{1.2.2}$$

betrachtet, in denen die Verfahrensfunktion f_h nur vom Wert $(\bar{x},u(\bar{x}))$ abhängt. Im Verfahren wird der lokale Fehler T_h vernachlässigt und die Differentialgleichung (1.1.1) durch eine *Differenzengleichung* ersetzt. Somit ergeben sich Näherungswerte $u_\Delta(x)$ an $u(x)$ 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 1.2.1 *Ein* Einschrittverfahren *zur Lösung des Anfangswertproblems (1.1.1), (1.1.2) besteht aus der Wahl einer Verfahrensfunktion $f_h(x,y)$ und eines Gitters*

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

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

$$\begin{aligned} y_0 &:= u_0,\\ y_{i+1} &:= y_i + h_i f_{h_i}(x_i,y_i), \quad i = 0,\ldots,N-1. \end{aligned} \tag{1.2.3}$$

Die Bezeichnung Einschrittverfahren bezieht sich auf die Tatsache, daß in (1.2.3) nur Näherungswerte y_i, y_{i+1} aus einem Schritt $x_i \to x_{i+1}$ verknüpft werden. Es folgen vier einfache Beispiele.

Verfahren 1, Euler-Cauchy-Polygonzug: Das Integral in (1.2.2) wird durch den Funktionswert am linken Rand ersetzt,

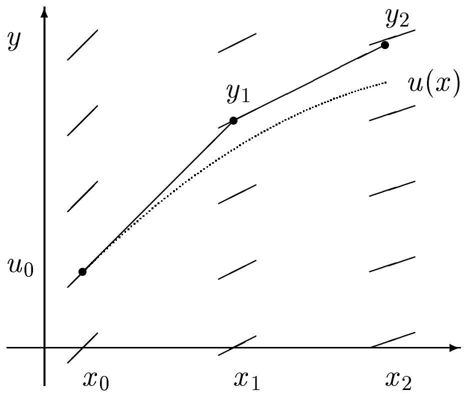
$$\int_0^1 f(\bar{x}+hs, u(\bar{x}+hs))\,ds \cong f(\bar{x},u(\bar{x})) =: f_h(\bar{x},u(\bar{x})),$$

das Verfahren ist also

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

$i = 0, 1, \ldots$.

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



Verfahren 2, von Runge: Integralapproximation durch die Rechteckregel,

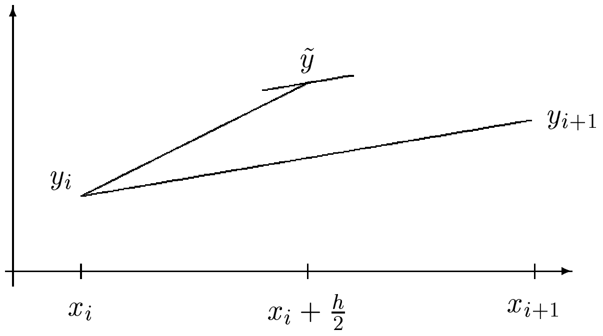
$$\int_0^1 f(\bar{x} + hs, u(\bar{x} + hs))\, ds \cong f(\bar{x} + \frac{h}{2}, u(\bar{x} + \frac{h}{2})).$$

Der unbekannte Wert $u(\bar{x} + \frac{h}{2})$ wird durch Verfahren 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,

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

Wieder wird $u(\bar{x} + h)$ durch Verfahren 1 angenähert. Jetzt ist

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

Verfahren 4, Klassisches Runge-Kutta-Verfahren: Integralapproximation durch eine modifizierte Simpsonregel,

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

Beispiel 1.2.2 Dgl $u' = u =: f(x,u),\ x_0 = 0, u_0 = 1 \Rightarrow u(x) = e^x$.

Verfahren	Näherung	lokal.Fehler
1	$y_1 = y_0 + hy_0 = 1 + h$	$h^2/2$
2	$y_{1/2} = 1 + \frac{h}{2},\ y_1 = 1 + h(1 + \frac{h}{2}) = 1 + h + \frac{1}{2}h^2$	$h^3/6$
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 erzeugen alle Verfahren den Anfang der Exponentialreihe unterschiedlich weit. Sie sind Spezialfälle einer allgemeinen Verfahrensklasse, der *Runge-Kutta-Verfahren.* Diese bestehen aus $m \in \mathbf{N}$ Stufen und werden durch eine Tabelle von Koeffizienten definiert,

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

wobei $c_i = \sum_{j=1}^m a_{ij},\ i = 1, \ldots, m$ gilt. Die Verfahrensfunktion beim allgemeinen Runge-Kutta-Verfahren ist damit

$$f_h(x,y) := \sum_{i=1}^m b_i\ k_i\,, \tag{1.2.4}$$

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

Wenn die Koeffizientenmatrix A strikt untere Dreieckgestalt besitzt, also für $a_{ij} = 0\,\forall j \geq i$, können die Hilfsgrößen k_i der Reihe nach explizit aus diesen Gleichungen berechnet werden. Dann liegt ein *explizites Runge-Kutta-Verfahren* vor. Die Verfahren 1–4 gehören zu dieser Klasse, das klassische Runge-Kutta-Verfahren 4 etwa hat die Tabelle

$$\begin{array}{c|cccc} 0 & 0 & \cdot & \cdot & \cdot \\ \frac{1}{2} & \frac{1}{2} & 0 & \cdot & \cdot \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & \cdot \\ 1 & 0 & 0 & 1 & 0 \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array}$$

Im allgemeinen Fall $(\exists a_{ij} \neq 0, j \geq i)$ ist (1.2.4) ein *implizites* Runge-Kutta-Verfahren, da zur Berechnung der k_i i.a. nichtlineare Gleichungssysteme zu lösen sind. Solche Verfahren können sehr gute Eigenschaften besitzen und werden daher später noch einmal betrachtet.

Die entscheidende Frage bei den numerischen Verfahren (1.2.3) ist die nach der Konvergenz der Näherungswerte $u_\Delta(x) \to u(x),\ |\Delta| \to 0,\ (x \text{ fest})$ bei feiner werdendem Gitter. Der Nachweis dieser Tatsache erfolgt nach einem sehr allgemeinen Prinzip zweistufig in der Form

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

1.2.2 Konsistenz

Eine Verfahrensfunktion f_h sollte sinnvollerweise so gewählt sein, daß die exakten Lösungen u der Differentialgleichung auch die Differenzengleichung (1.2.3) näherungsweise erfüllen.

Definition 1.2.3 *Das Verfahren (1.2.3) heißt* konsistent, *wenn mit jeder Lösung u der Dgl (1.1.1) bei stetig differenzierbarer rechter Seite f für den* lokalen Fehler

$$T_h(x) := \frac{1}{h}[u(x+h) - u(x)] - f_h(x, u(x)), \quad x \in [a, b-h], \tag{1.2.5}$$

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

$$T_h(x) = \mathcal{O}(h^p), \ h \to 0, \quad x \in [a, b), \tag{1.2.6}$$

für genügend oft differenzierbare rechte Seiten f.

Beispiel 1.2.4 Verfahren 2, $f_h = f(x + \frac{h}{2}, y + \frac{h}{2} f(x, y))$. Die Taylorentwicklung einer Lösung u in x ist

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

Für $f \in C^2$ erhält man die Ableitungen $u^{(j)}(x)$ der Lösung aus $u' = f(x, u), \ u'' = f_x + f_y u' = f_x + f_y f$. Daher stimmt die Taylorentwicklung von f_h (nach h),

$$f(x + \frac{h}{2}, u + \frac{h}{2} f) = f(x, u) + \frac{h}{2}(f_x + f_y f) + \mathcal{O}(h^2)$$

mit der obigen in den ersten beiden Gliedern überein, das Verfahren hat also Konsistenzordnung $p = 2$.

Konsistenznachweise sind zwar elementar durchführbar, aber sehr arbeitsaufwendig, Standardmethode ist dabei die im Beispiel angewendete Taylorentwicklung. Umgekehrt können durch Vergleich der Taylorentwicklungen von u und f_h in (1.2.4) die erforderlichen Bedingungen an die Koeffizienten hergeleitet und damit geeignete Verfahren konstruiert werden. Für ein dreistufiges Verfahren mit Ordnung drei sind, z.B., folgende *Ordnungsbedingungen* zu erfüllen:

$$\begin{array}{rcl} b_1 + b_2 + b_3 &=& 1 \\ b_2 c_2 + b_3 c_3 &=& \frac{1}{2} \\ b_2 c_2^2 + b_3 c_3^2 &=& \frac{1}{3} \\ b_3 a_{32} c_2 &=& \frac{1}{6} \end{array}. \tag{1.2.7}$$

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 zweiparametrige Schar von dreistufigen Verfahren mit Ordnung drei. Analoges gilt bei 4-stufigen (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 vorher behandelten Verfahren 1–4 besitzen jeweils die höchsten erreichbaren Ordnungen. 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.

Um die Konvergenz eines Verfahrens bzw. der durch es berechneten Näherungen nachzuweisen, reicht die Konsistenz i.a. nicht aus. Das Verfahren muß eine weitere zentrale Eigenschaft erfüllen.

1.2.3 Stabilität

Stabilität bedeutet allgemein, daß sich die von einem Verfahren berechneten Näherungen bei kleinen Störungen im Verfahren nur wenig (Lipschitz-stetig) änderen. Wegen der einfachen Struktur des diskreten Anfangswertproblems (1.2.3) bietet sich hier aber eine etwas speziellere Formulierung an.

Definition 1.2.5 *Ein Einschrittverfahren (1.2.3) heißt stabil, wenn Konstanten S, r, H_0 existieren so, daß für je zwei gestörte Lösungen $\{y_i\}, \{z_i\}$ mit*

$$\left.\begin{array}{rcl} \frac{1}{h_i}[y_{i+1} - y_i] - f_{h_i}(x_i, y_i) &=& \eta_i \\ \frac{1}{h_i}[z_{i+1} - z_i] - f_{h_i}(x_i, z_i) &=& \zeta_i \end{array}\right\} \; i = 0, 1, \ldots, N-1, \tag{1.2.8}$$

gilt

$$\max_{i=0}^{N} \|y_i - z_i\| \le S\Big(\|y_0 - z_0\| + \max_{i=0}^{N-1} \|\eta_i - \zeta_i\|\Big) \tag{1.2.9}$$

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

Im Gegensatz zu den später behandelten Mehrschrittverfahren gilt die Stabilität von Einschrittverfahren unter einfachen Voraussetzungen. In dieser Beziehung und auch in Bezug auf die verwendeten Beweistechniken liegen starke Ähnlichkeiten zu den Aussagen aus §1.1 vor. Umgekehrt kann übrigens auch der Existenzsatz 1.1.1 mit Hilfe des Euler-Cauchy-Verfahrens bewiesen werden. Als Beweishilfsmittel im folgenden Stabilitätssatz wird das diskrete Analogon zum Gronwall-Lemma benötigt.

Lemma 1.2.6 *Für Werte $\delta_i \in \mathbf{R}_+$ gelte mit $\alpha, \ell, h_i \ge 0$ rekursiv*

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

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

$$\delta_i \le e^{\ell(x_i - x_0)}\Big(\delta_0 + (x_i - x_0)\alpha\Big), \quad i = 0, \ldots, N.$$

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

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

Als wesentliche Voraussetzung wird beim Stabilitätsbeweis von Einschrittverfahren wieder eine Lipschitzbedingung benötigt.

Satz 1.2.7 (Stabilitätssatz) *Die das Einschrittverfahren (1.2.3) definierende Funktion f_h erfülle in dem in Satz 1.1.1 definierten Streifen Ω eine Lipschitzbedingung*

$$\|f_h(x,y) - f_h(x,z)\| \le \ell\|y - z\| \quad \forall x \in [a,b], y, z \in \mathbf{R}^n. \tag{1.2.10}$$

Dann ist das Verfahren stabil, es gilt (1.2.9) 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 (1.2.8) 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 folgt daraus die Ungleichungskette

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

F'ur die Differenzen $\delta_i := \|y_i - z_i\|$ entspricht dies den Voraussetzungen von Lemma 1.2.6 mit $\alpha = \max_j \|\eta_j - \zeta_j\|$ und liefert mit $(x_i - x_0)e^{\ell(x_i - x_0)} \le (b-a)e^{\ell(b-a)}$ die Behauptung. ■

Die im Satz benötigte Lipschitzbedingung der Verfahrensfunkton f_h folgt bei expliziten Runge-Kutta-Verfahren aus derjenigen von f. Beim Euler-Cauchy-Verfahren ist dies trivialerweise der Fall, elementar ist auch

Beispiel 1.2.8 Verfahren 2, $f_h(x,y) = f(y + \frac{h}{2}f(y))$ die Abhängigkeit von x ist unwesentlich. Aus der Bedingung $\|f(y) - f(z)\| \le L\|y - z\|$ folgt

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

d.h. $\ell = L + \frac{h}{2}L^2$.

1.2.4 Konvergenz

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

Satz 1.2.9 *Ein konsistentes und stabiles Verfahren ist konvergent:*
Das Einschrittverfahren (1.2.3) sei konsistent mit der Ordnung $p > 0$, d.h. es gelte (1.2.6). Das Verfahren sei außerdem stabil, etwa aufgrund der Lipschitzbedingung (1.2.10) für f_h. Dann konvergieren die Näherungswerte $\{y_i = u_\Delta(x_i)\}$ bei feiner werdendem Gitter Δ $(|\Delta| \to 0)$ gegen die Lösung u des AWPs (1.1.1),(1.1.2). Die Konvergenzordnung ist mindestens p, d.h. es gibt eine Konstante K so, daß für alle $|\Delta| \le H_0$ gilt

$$\max_{x \in \Delta} \|u_\Delta(x) - u(x)\| = \max_{i=0}^{|\Delta|} \|y_i - u(x_i)\| \le K|\Delta|^p. \tag{1.2.11}$$

Beweis Nach Definition 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 Gitterwerte $z_i := u(x_i)$ der Lösung der Dgl gilt andererseits die Konsistenzaussage (1.2.6)

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

Diese Gleichungen entsprechen denen aus der Stabilitätsdefinition, (1.2.8), mit $\eta_i = 0$ und $\zeta_i = T_{h_i}(x_i)$. Wegen $y_0 = u_0$ und $h_i^p \le |\Delta|^p$ folgt daher die Schranke (1.2.11). ■

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

Bemerkung: Für stabile Verfahren gilt Konvergenzordnung $\ge$ Konsistenzordnung.

Von theoretischer Bedeutung ist die Konvergenzaussage des Satzes, da sie zeigt, daß die Näherungslösung u_Δ genügend schnell konvergiert. Bei einer Halbierung der Schrittweiten im Gitter kann, z.B., eine Verkleinerung des Fehlers um den Faktor 2^{-p} erwartet werden. Dies legt die Verwendung von Verfahren möglichst hoher Ordnung nahe. Allerdings hängt die Fehlerkonstante K in (1.2.11) von immer höheren Ableitungen $u^{(p)}$ ab und begrenzt dadurch die verwendbare Ordnung. Diese Konstante K ist nicht explizit bekannt und in der Praxis kaum berechenbar, daher kann aus der Fehlerschranke keine Angabe über die tatsächliche Größe des Fehlers u_Δ bei gegebenem Gitter abgeleitet werden. Genausowenig kann man umgekehrt ein Gitter konstruieren, auf dem ein geforderter Fehler erreicht wird. Diese Frage wird jetzt behandelt.

1.2.5 Schrittweitensteuerung

In vielen realen Problemen kann die Gestalt der Lösungen in unterschiedlichen Bereichen des Integrationsintervalls $[a, b]$ sehr unterschiedlich ausfallen (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 Schrittweiten an diesen Verlauf anpassen. Durch eine Inspektion von Lemma 1.2.6 sieht man, daß die Fehlerschranke (1.2.11) aus

$$\max_{x\in\Delta} \|u_\Delta(x) - u(x)\| \le \tilde{K} \sum_{j=0}^{N-1} h_j \|T_{h_j}\|, \quad \tilde{K} = e^{\ell(b-a)} \tag{1.2.12}$$

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

$$\|T_{h_i}(x_i)\| \overset{!}{\le} \varepsilon \quad \forall i = 0, \ldots, N-1 \tag{1.2.13}$$

erreicht werden, vgl. Numerik I, §5.4 (adaptive Integration). Da der exakte lokale Fehler T_h nicht bekannt ist, arbeitet man mit einer Schätzung für diesen und nutzt das bekannte Verhalten $T_h \doteq \gamma h^p$ mit der Forderung (1.2.13) 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 läßt sich die eigentlich für (1.2.13) erforderliche Schrittweite $\hat{h}$ schätzen nach

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

Zur *Schätzung* des lokalen Fehlers 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

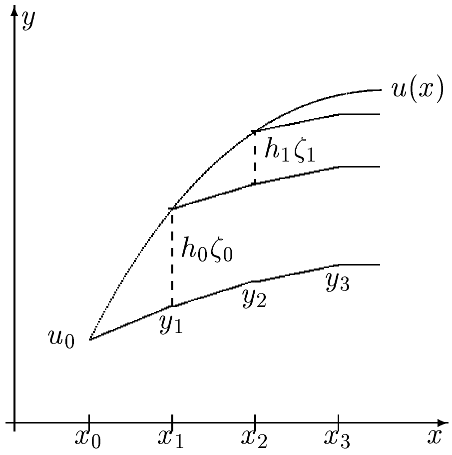
$$\left.\begin{array}{rcl} y_{i+1} &=& y_i + h f_h(x_i, y_i) \\ \bar{y}_{i+1} &=& y_i + h \bar{f}_h(x_i, y_i) \end{array}\right\} \text{ mit Fehler } \left\{\begin{array}{c} \mathcal{O}(h^p) \\ \mathcal{O}(h^{p+1}) \end{array}\right. . \tag{1.2.14}$$

Ist $v_i(x)$ diejenige Lösung der Dgl, die die Anfangsbedingung $v_i(x_i) = y_i$ erfüllt, dann gilt bezüglich dieser Lösung die Beziehung

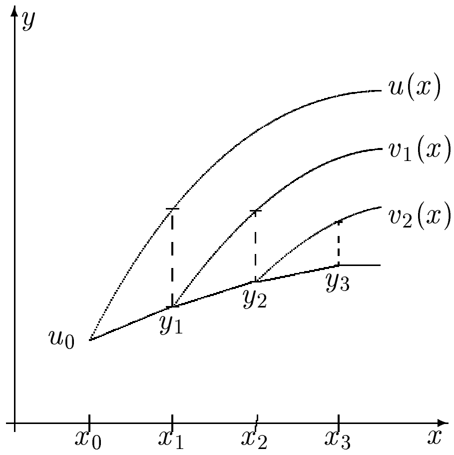
$$\begin{aligned} T_h(x_i) &= \frac{1}{h}[v_i(x_i+h) - y_i] - f_h(x_i, y_i) = \frac{1}{h}[v_i(x_i+h) - y_{i+1}] \\ &= \frac{1}{h}[\bar{y}_{i+1} - y_{i+1}] + \mathcal{O}(h^{p+1}) = \bar{f}_h(x_i, y_i) - f_h(x_i, y_i) + \mathcal{O}(h^{p+1}). \end{aligned} \tag{1.2.15}$$

Daher ist (für genügend kleine Schrittweite) die Differenz $\bar{f}_h(x_i, y_i) - f_h(x_i, y_i)$ eine Schätzung des lokalen Fehlers.

Die Unterschiede zwischen der Beweismethode zur Fehler**ab**schätzung in Satz 1.2.9 und der hier verwendeten zur Fehlerschätzung lassen sich graphisch mit den Lösungsfächern von Differentialgleichung (1.1.1) und Differenzengleichung (1.2.3) veranschaulichen. Die gestrichelten Fehlerbeiträge der beiden Ansätze sind dabei vergleichbar, aber nicht identisch.



In Satz 1.2.9: Lösungsfächer Differenzengl. mit $\zeta_j = T_{h_j}(x_j), j = 0, 1, ..$



In (1.2.15): Lösungsfächer Dgl

Mit einer solchen Schätzung des lokalen Fehlers kann auf sehr einfache Weise eine Schrittweitensteuerung formuliert werden:

Algorithmus 1.2.10 Schrittweitensteuerung beim AWP:

> $x := a;\ y := u_0;\ h =?,\ \varepsilon =?$
> wiederhole:
> $\quad y0 := y + hf_h(x, y);$
> $\quad y1 := y + h\bar{f}_h(x, y);$
> $\quad q := \varepsilon h / \|y1 - y0\|;$
> $\quad$ falls $q > 1$ dann {Schritt akzeptieren}
> $\qquad [\ x := x + h;\ y := y0;\ ]$
> $\quad h := h * \sqrt[p]{q};$
> $\quad$ falls $x + h > b$ dann $h := b - x;$
> bis $x \geq b;$

In der Praxis wird meist eine geringfügig kleinere Schrittweite $h := 0.9 * h * \sqrt[p]{q}$ genommen, 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 $y1$ weitergerechnet ($y := y1$ statt $y := y0$ nach $x := x + h$), da die lokale Fehlerschätzung nur zur Steuerung verwendet wird. Zur Wahl der Toleranz ε für den *lokalen* Fehler T_h in Relation zum gewünschten *globalen* Fehler $\|u_\Delta - u\|$ kann keine allgemeine Aussage gemacht werden, da die Konstante $\tilde{K}$ in (1.2.12) i.a. unbekannt ist.

Bei der Konstruktion von Verfahrenspaaren (1.2.14) gibt es zwei übliche Zugänge unterschiedlicher Effizienz und Allgemeinheit:

a) *Eingebettete Runge-Kutta-Verfahren:* Analog zum Konstruktionsprinzip der Runge-Kutta-Verfahren wird versucht, die gewünschte Näherung $\bar{y}_{i+1}$ mit möglichst wenig Funktionsauswertungen zu erreichen. Dazu sucht man nach zwei Verfahren (1.2.4) mit m bzw. $m+1$ Stufen, bei denen das zweite die Funktionswerte k_i des ersten mitverwendet,

$$\begin{aligned} y_{i+1} &= y_i + h_i \sum_{j=1}^{m} b_j k_j, \\ \bar{y}_{i+1} &= y_i + h_i \sum_{j=1}^{m+1} \bar{b}_j k_j, \end{aligned} \qquad \begin{array}{c|cccc} c_1 & 0 & \cdots & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ c_{m+1} & a_{jl} & & & 0 \\ \hline & b_1 & \cdots & b_m & 0 \\ & \bar{b}_1 & \cdots & \cdot & \bar{b}_{m+1} \end{array}$$

$$k_j = f(x + h_i c_j, y_i + h_i \sum_{l=1}^{j-1} a_{jl} k_l), \ j = 1, .., m+1,$$

Die lokale Fehlerschätzung zum i-ten Schritt ist damit (und $b_{m+1} := 0$)

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

Mit Hilfe der Ordnungsbedingungen (1.2.7) wird kurz die Konstruktion eines eingebetteten Verfahrens zu Verfahren 3 betrachtet. Hier sind $c_2 = a_{21} = 1$ vorgegeben. Für die restlichen Parameter bleiben die Bedingungen

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

Für $c_3 = \frac{1}{2}$ ergeben sich daraus die Koeffizienten in der folgenden rechten Tabelle. Die linke enthält ein (2,3)-Paar zu Verfahren 2.

(2,3)-Paar zu Verf. 2

$$\begin{array}{c|ccc} 0 & 0 & & \\ \frac{1}{2} & \frac{1}{2} & 0 & \\ 1 & -1 & 2 & 0 \\ \hline & 0 & 1 & 0 \\ & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \end{array}$$

(2,3)-Paar zu Verf. 3

$$\begin{array}{c|ccc} 0 & 0 & & \\ 1 & 1 & 0 & \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 \\ \hline & \frac{1}{2} & \frac{1}{2} & 0 \\ & \frac{1}{6} & \frac{1}{6} & \frac{2}{3} \end{array}$$

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 $f(x + c_5 h, ...)$ als erster Funktionswert im nächsten Intervall, $f(x + h + c_1 h, ..)$, verwendet wird, das also beinahe diese Anforderungen erfüllt. Auch zum klassischen 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 nötig und es ist günstiger, das Stufenpaar (5,6) zu verwenden. Sehr effiziente Verfahren (und Implementierungen) sind das (4,5)-Paar DOPRI5 bzw. (3,5,8)-Tripel DOP853 von Dormand und Prince.

In praktischen Anwendungen benötigt man oft (z.B. Graphik) die Näherung y bzw. u_Δ tatsächlich als Funktion von t. Zu diesem Zweck betrachtet man *stetige Erweiterungen*. Dabei sind die Koeffizienten in (1.2.4) Polynome $\bar b_i(t)$ mit $\bar b_i(0) = 0$, $\bar b_i(1) = b_i$, und 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_Δ definiert. Für das Verfahren 3 etwa ergibt sich mit

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

eine stetige Erweiterung $u_\Delta(x)$ der gleichmäßigen Konvergenzordnung 2. Im allgemeinen kann man aber für stetige Erweiterungen nicht die volle normale Ordnung der Endformel mit $\bar b_i(1)$ erwarten. Für 3-stufige Verfahren ist 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

$$u_\Delta(x) - u(x) = h^p g_p(x) + \mathcal{O}(h^{p+1}), \ h \to 0, \tag{1.2.16}$$

vgl. Numerik I, §5.5. Durch Linearkombination von zwei Näherungen zu Schrittweiten h und $h/2$ kann der dominierende Fehleranteil $h^p g_p$ eliminiert ($\to$ Richardson-Extrapolation) oder aber geschätzt werden, da für eine zweite Lösung $u_{\Delta'}$, zu $\Delta' = \{x_i = a + i\frac{h}{2} : i = 0, \ldots, 2N\}$, gilt

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

Daraus kann 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 die Fehlerschätzungen

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

berechnet werden. Wie in (1.2.15) ergibt sich

$$\begin{aligned} T_h(x_i) &= \frac{1}{h}[v_i(x_i + h) - u_\Delta(x_i + h)] \doteq \frac{1}{h}[\bar u_\Delta(x_i + h) - u_\Delta(x_i + h)] \\ &= \frac{1}{1 - 2^{-p}} \frac{1}{h}[u_{\Delta'}(x_i + h) - u_\Delta(x_i + h)] \doteq -h^{p-1} g_p(x_i + h). \end{aligned} \tag{1.2.17}$$

Diese Größe ist im speziellen Fall von der Ordnung $\mathcal{O}(h^p)$. Da man sich hier in (1.2.16) 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 (1.2.17) eine Fehlerschätzung von der in Algorithmus 1.2.5 verwendeten Gestalt. Im Vergleich zu eingebetteten Verfahren werden hier aber wesentlich mehr Zusatzfunktionsauswertungen verwendet (bezogen auf die genauere Lösung $u_{\Delta'}$ nach 2 Schritten mit $h/2$ also $m/2$ pro Schritt).

Beispiel 1.2.11 Periodische Satellitenbahn um Erde und Mond: In der Ebene des rotierenden Erde-Mond-System bewegt sich ein Satellit mit den Koordinaten $u(t) \in \mathbf{R}^2$ 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 $E = (-\mu, 0)^{\mathsf{T}}$ bzw. $M = (\nu, 0)^{\mathsf{T}}$ die Positionen von Erde und Mond sind. Es gibt eine periodische Bahn, die nahe an der Erde vorbeiführt und dort sehr kleine Schrittweiten erfordert. Ergebnis bei $\varepsilon =_{10} -4$:

	Anz.f-Werte	$h_{\min}$	$h_{\max}$	Endfehler
RuKuFe-2/3	35805	2_{10}-6	8_{10}-2	2_{10}-5
RuKuFe-4/5	972	4_{10}-4	6_{10}-1	9_{10}-3

Schrittweitensteuerung hatte zunächst nur das Ziel, die geforderte Genauigkeit der Näherung mit möglichst wenig Aufwand zu erreichen. Zusätzlich reduziert sie aber auch den Einfluß von *Rundungsfehlern*. Bei Maschinenrechnung werden in (1.2.3) gestörte Näherungen $\tilde{y}_i$ berechnet, die pro Schritt anfallenden Rundungsfehler ε_i führen also zu einer verfälschten Rekursion

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

$\tilde{y}_0 = u_0$. Daher kommen bei praktischer Rechnung zum Diskretisierungsfehler $y_i - u(x_i)$ noch Rundungsfehler hinzu. Unter Voraussetzung (1.2.10) gilt dafür

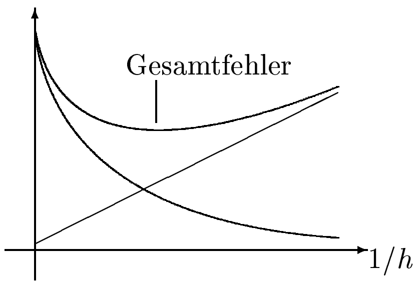
$$\begin{aligned} \|\tilde{y}_{i+1} - y_{i+1}\| &\le \|\tilde{y}_i - y_i\| + h_i \|f_{h_i}(x_i, \tilde{y}_i) - f_{h_i}(x_i, y_i)\| + \|\varepsilon_{i+1}\| \\ &\le (1 + h_i \ell)\|\tilde{y}_i - y_i\| + \hat{\varepsilon}_{i+1}, \quad \hat{\varepsilon}_{i+1} := \|\varepsilon_{i+1}\|. \end{aligned}$$

Analog zu Lemma 1.2.6 ergibt sich mit $\tilde{y}_0 = y_0$ die Abschätzung

$$\|\tilde{y}_i - y_i\| \le \sum_{j=1}^{i} e^{\ell(x_i - x_j)} \hat{\varepsilon}_j \le S \sum_{j=1}^{N} \hat{\varepsilon}_j, \quad S := e^{\ell(b-a)}. \tag{1.2.18}$$

Bei einem äquidistanten Gitter führt dies für den Gesamtfehler auf die Schranke

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



Bei äquidistantem Gitter muß die (zunächst unbekannte) Schrittweite h nach der ungünstigsten Stelle im Intervall gewählt werden und erzwingt daher evtl. eine sehr große Zahl N von Schritten. Da der Rundungsfehler nach (1.2.18) im wesentlichen von dieser Schrittezahl N abhängt, können diese bei variablem Gitter, das in leichten Passagen große und daher weniger Schritte verwendet, reduziert werden. Zusätzlich sei daran erinnert (vgl. Numerik I, §7.2), daß der kleinstmögliche Wert in der Schranke für $\|\tilde{y}_i - u(x_i)\|$ von der Größe $O(\varepsilon^{p/(p+1)})$ ist und mit wachsender Ordnung kleiner wird.

1.3 Mehrschrittverfahren

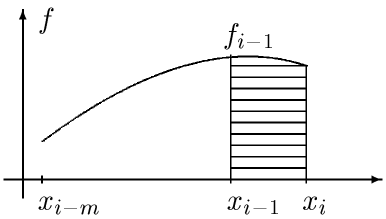
1.3.1 Adams-Verfahren

Bei Einschrittverfahren der Ordnung p sind $m \geq p$ Stufen erforderlich und daher zur Approximation des Integrals in (1.2.1),

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

beim Schritt $x_{i-1} \to x_i$ jeweils mindestens p Funktionswerte $k_j \cong f(x_{i-1}+hc_j, u(x_{i-1}+hc_j))$ neu zu berechnen. Eine Alternative ist, stattdessen schon bekannte, vor x_{i-1} gelegene Funktionswerte $f(x_{i-j}, y_{i-j}) =: f_{i-j},\ j = 1, 2, .., m,$ heranzuziehen.

Durch Integration des Interpolationspolynoms $p_{m-1}(x)$ vom Grad $m-1$ zu den Funktionswerten $f_{i-1}, ..., f_{i-m}$, das mit den Lagrangepolynomen $L_j(x)$ zu den Stützstellen $x_{i-1}, .., x_{i-m}$ dargestellt werden kann (vgl. Numerik I, §3.2), ergibt sich für dieses Integral die Approximation



$$f_h(x_{i-1}, y_{i-m}, \ldots, y_{i-1}) := \sum_{j=-m}^{-1} f_{i+j} \underbrace{\int_0^1 L_j(x_{i-1} + ht)\, dt}_{=:\beta_{-j}}. \tag{1.3.1}$$

Diese Verfahrensfunktion hängt also von m früheren Lösungswerten y ab. Speziell für äquidistante Punkte (d.h. $h_i \equiv h$) erhält man damit die *Methode von Adams-Bashforth* (A-B):

$$\left.\begin{array}{rcl} \frac{1}{h}(y_i - y_{i-1}) & = & f_h(x_{i-1}, y_{i-m}, .., y_{i-1}) \\ f_h(x_{i-1}, y_{i-m}, .., y_{i-1}) & := & \sum\limits_{j=1}^{m} \beta_j f(x_{i-j}, y_{i-j}) \end{array}\right\} \ i = m, m+1, .., N. \tag{1.3.2}$$

Dieses Verfahren ist ein explizites, lineares, m-Schrittverfahren. Pro Schritt fällt hier tatsächlich nur *eine* neue Funktionsauswertung, f_{i-1}, an. Explizit heißt das Verfahren wieder, da f_h nicht vom unbekannten Wert y_i abhängt.

Der Konsistenzfehler entspricht dem Quadraturfehler bei der Funktion $u'(x)$, das Verfahren hat daher Ordnung $p = m$. Die Koeffizienten der ersten A-B-Verfahren enthält die 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 (1.3.1) werden nur Daten auf einer Seite des Integrationsintervalls $[x_{i-1}, x_i]$ benutzt. Ohne Erhöhung der Zahl m der Teilintervalle kann das A-B-Verfahren durch Berücksichtigung des Wertes im Endpunkt x_i verbessert werden, man interpoliert also in den

Stellen $x_{i-m}, .., x_i$ mit einem Polynom vom Grad m. Dies führt auf die *Methode von Adams-Moulton* (A-M)

$$\left.\begin{array}{rcl} \frac{1}{h}(y_i - y_{i-1}) & = & f_h(x_{i-1}, y_{i-m}, .., y_i) \\ f_h(x_{i-1}, y_{i-m}, .., y_i) & := & \sum\limits_{j=0}^{m} \mu_j f(x_{i-j}, y_{i-j}) \end{array}\right\} \; i = m, m+1, .., N. \tag{1.3.3}$$

Hier liegt ein implizites, lineares, m-Schritt-Verfahren vor, da (1.3.3) für y_i ein (nichtlineares) Gleichungssystem darstellt. 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 von beiden Verfahrensklassen folgt aus den Approximationseigenschaften der Interpolationspolynome. Beide Verfahrenstypen werden in einem Satz behandelt. Dazu wird das A-B-Verfahren als A-M-Verfahren mit Gewicht $\beta_0 = 0$ interpretiert.

Satz 1.3.1 *Der lokale Fehler bei den Adams-Bashforth-Verfahren (1.3.2), $\gamma_0 := 0$, $\gamma_j := \beta_j, 1 \leq j \leq m$, bzw. Adams-Moulton-Verfahren (1.3.3), $\gamma_j := \mu_j, 0 \leq j \leq m$, 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

$$T_h(x_{i-1}) = \begin{cases} c_m^B \; h^m \; u^{(m+1)}(x_{i-1}) + \mathcal{O}(h^{m+1}) & \text{bei A-B} \\ c_m^M h^{m+1} u^{(m+2)}(x_{i-1}) + \mathcal{O}(h^{m+2}) & \text{bei A-M} \end{cases}. \tag{1.3.4}$$

Beweis Beim A-B-Verfahren sei p das Interpolationspolynom vom Grad $m-1$ zu den Werten $(x_{i-j}, u'(x_{i-j}))$, $j = 1, .., m$. Dann folgt aus der Fehlerformel der Interpolation (Numerik I, 3.3.11) und durch Taylorentwicklung die Aussage

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

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

Die Fehlerkoeffizienten c_m sind in den obigen Tabellen angegeben. Das A-M-Verfahren mit m Stufen besitzt 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 hier auftretenden nichtlinearen Gleichung für y_i in (1.3.3) kann durch Kombination mit dem A-B-Verfahren umgangen werden, indem mit A-B zunächst eine Approximation $\tilde{y}_i$ für den unbekannten Wert berechnet wird. So ergibt sich das *Prädiktor-Korrektor-Verfahren* (P-K)

$$\left.\begin{array}{lcl} \tilde{y}_i & := & y_{i-1} + h\sum\limits_{j=1}^{m} \beta_j f(x_{i-j}, y_{i-j}) \\ y_i & := & y_{i-1} + h\mu_0 f(x_i, \tilde{y}_i) + h\sum\limits_{j=1}^{m} \mu_j f(x_{i-j}, y_{i-j}) \end{array}\right\} \; i = m, m+1, \ldots \tag{1.3.5}$$

Hier werden pro Schritt zwei Funktionsaufrufe benutzt. Der lokale Fehler dieses Verfahren ist tatsächlich von der Ordnung h^{m+1}. Denn nach Satz 1.3.1 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

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

Der Fehler bei Runge-Kutta-Verfahren besitzt keine so einfache Gestalt. Im allgemeinen ist er aber für ein Verfahren gleicher Ordnung kleiner als bei Mehrschrittverfahren. Für das Problem $u' = u, u(0) = 1$, etwa ergeben sich bei den Verfahren vierter Ordnung nach einem Schritt

Verfahren:	Runge-Kutta	Adams-Bashf.	Adams-Moulton
lok. Fehler:	$\frac{1}{120}h^5$	$\frac{251}{720}h^5$	$\frac{19}{720}h^5$

Bei den Mehrschrittverfahren sind für gleiche Fehlerbeträge also kleinere Schrittweiten erforderlich. Außerdem ist die Programmierung aufwendiger.

Fragen zum praktischen Einsatz:

- *Anlaufrechnung:* Die m-Schritt-Verfahren können, wie aus (1.3.2)-(1.3.5) ersichtlich, erstmals nach einer Anlaufphase von m Integrationsschritten eingesetzt werden. Daher sind Startwerte $y_0, \ldots, y_{m-1}$ ausreichender Genauigkeit auf andere Weise bereitzustellen, entweder mit Runge-Kutta-Verfahren oder Mehrschrittverfahren niederer (wachsender) Ordnung und sehr kleinen Schrittweiten.

- *Schrittweitenwechsel:* Die Adams-Formeln (1.3.2)-(1.3.5) beziehen sich auf äquidistante Punkte $x_{i-m}, .., x_i$ im konstanten 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, §3.3) dargestellt, gilt

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

Ein Wechsel der Schrittweite $h \to \hat{h}$ ist in dieser Form jetzt einfach möglich durch Ersetzung der Faktoren $h^j \to \hat{h}^j$. Die schon berechneten Differenzen $f[..,x_{i-1}]$ 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 (1.3.5) 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:

  Ist ein Schritt in $[x_{i-1}, x_i]$ mit einem Verfahren der Ordnung m akzeptiert worden, 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.

Die erwähnten Maßnahmen ermöglichen auch bei Mehrschrittverfahren eine selbststeuernden Einsatz. Allerdings ist der damit verbundene Verwaltungsaufwand erheblich höher als bei Einschrittverfahren und muß bei einem Vergleich berücksichtigt werden (vgl. §1.4).

1.3.2 Lineare Mehrschrittverfahren und Stabilität

In den Verfahren (1.3.2,1.3.3) werden zwar die Funktionswerte f_{i-j} aus m zurückliegenden Punkten benutzt ('rechte Seite'), Lösungswerte dagegen nur von zwei Stellen, y_i, y_{i-1}. Ohne Erhöhung der Schrittzahl m kann man daher eine Verbesserung durch Einführung zusätzlicher Parameter erwarten. Dazu werden jetzt allgemeinere Verfahren der Form

$$\sum_{j=0}^{m} \alpha_j y_{i-j} = h \sum_{j=0}^{m} \beta_j f_{i-j}\,, \quad i = m, m+1, \ldots, \tag{1.3.6}$$

betrachtet 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.*

Eines der Ziele bei der Konstruktion von Verfahren ist eine möglichst hohe Ordnung des lokalen bzw. Konsistenzfehlers

$$T_h(x_{i-1}) = \frac{1}{h} \sum_{j=0}^{m} \alpha_j u(x_{i-j}) - \sum_{j=0}^{m} \beta_j u'(x_{i-j}). \tag{1.3.7}$$

Zur Überprüfung werden die Taylorentwicklungen in beiden Summen verglichen. Mit

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

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

in (1.3.7) ergibt sich

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

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

Satz 1.3.2 *Das Verfahren (1.3.6) besitzt Konsistenzordnung p genau dann, wenn seine Koeffizienten das folgende lineare Gleichungssystem erfüllen, $\alpha_0 = 1$,*

$$\begin{gathered} \sum_{j=0}^{m} \alpha_j = 0, \\ \sum_{j=0}^{m} \Big(\tfrac{j^{k+1}}{k+1} \alpha_j + j^k \beta_j \Big) = 0, \quad k = 0, \ldots, p-1. \end{gathered} \tag{1.3.8}$$

Ordnungsbedingungen sind hier also wesentlich leichter aufzustellen und Lösungen zu berechnen 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 1.3.3 Konstruktion eines expliziten ($\beta_0 = 0$) 2-Schrittverfahrens mit maximaler Ordnung $p = 3$.

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

Das Verfahren lautet also

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

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

Numerisches Ergebnis: Die Fehler in den Approximationen oszillieren und wachsen ungefähr um einen Faktor 5 pro Schritt.

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

Das Verfahren ist **instabil**, Störungen wachsen extrem an. Im Gegensatz zu den Einschrittverfahren ist die Stabilitätsforderung eine echte Einschränkung für Mehrschrittverfahren!

Es gibt sogar eine sogenannte *Ordnungsbarriere* (von Dahlquist):

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

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

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

$$y_i - y_{i-k} = h \sum_{j=0}^{m} \beta_j f_{i-j}, \quad i \geq m, \ 1 \leq k \leq m. \tag{1.3.9}$$

Der folgende Konvergenzsatz enthält für diese eine zu den Einschrittverfahren ähnliche Stabilitätsaussage. Auf die Stabilitätsforderungen an allgemeine lineare Mehrschrittverfahren wird kurz noch einmal in §1.4 eingegangen.

Satz 1.3.4 *Die rechte Seite f der Dgl erfülle eine Lipschitzbedingung (1.1.5) 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 $\|y_i - u(x_i)\| \leq \varphi(h), \ i = 0,..,m-1,$ und für die Konsistenzfehler $\|T_h(x_{i-1})\| \leq \tau(h), \ i = m, \dots, N$. Dann gilt die Fehlerschranke*

$$\|y_i - u(x_i)\| \leq e^{\lambda(x_i - x_{m-1})}[\varphi(h) + 2(x_i - x_{m-1})\tau(h)], \quad i = m, \dots, N, \tag{1.3.10}$$

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 $\max_\Delta \|u_\Delta - u\| = \mathcal{O}(h^p)$, $h \to 0$.

Beweis Für $hL|\beta_0| < 1$ ist die Gleichung (1.3.9) nach y_i eindeutig auflösbar (Banachscher Fixpunktsatz). Wieder durch Subtraktion der Gleichungen (1.3.9) und (1.3.7),

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

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

$$\begin{aligned} \|\varepsilon_i - \varepsilon_{i-k}\| &= h \| \sum_{j=0}^{m} \beta_j [f(x_{i-j}, y_{i-j}) - f(x_{i-j}, u(x_{i-j}))] - hT_h(x_{i-1}) \| \\ &\leq hL \sum_{j=0}^{m} |\beta_j| \|\varepsilon_{i-j}\| + h\tau(h). \end{aligned}$$

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

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

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

$$\|\varepsilon_i\| \le (1 + h\lambda)\delta_{i-1} + \frac{h}{1 - hL|\beta_0|}\tau(h), \quad i = m, .., N. \tag{1.3.11}$$

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 (1.3.11) durch δ_i ersetzt werden. Dann liegt aber die in Lemma 1.2.6 behandelte Situation vor, die auf die Schranke (1.3.10) führt. Als Startpunkt ist dabei allerdings x_{m-1} zu nehmen. ■

Die stabilen Verfahren höchster Ordnung $m+2$, m gerade, ergeben sich bei der Wahl $k = m$ in (1.3.9) aus den abgeschlossenen Newton-Cotes-Quadraturformeln zum Intervall $[x_{i-m}, x_i]$,

$$\frac{1}{h}[y_i - y_{i-m}] = \sum_{j=0}^{m} \beta_j f_{i-j} \quad \cong \frac{1}{h}\int_{x_{i-m}}^{x_i} f(x, u(x))\, dx. \tag{1.3.12}$$

Bei diesen Verfahren ist allerdings, im Gegensatz zu den Adams-Verfahren, die mit x wachsende Stabilitätsschranke (1.3.10) scharf, selbst dann, wenn im Problem nur exponentiell fallende Lösungen auftreten können (vgl. §1.4). Daher werden in der Praxis doch meist die Adams-Verfahren verwendet.

Aufgrund ihrer Symmetrie bzgl. $\frac{1}{2}(x_{i-m} + x_i)$ besitzen Verfahren der Form (1.3.12) aber noch eine weitere Eigenschaft, die von großer praktischer Bedeutung ist. Das einfachste, explizite Verfahren dieser Form ergibt sich aus der Rechteckregel zum Intervall $[-2h, 0]$, seine Symmetrie ist erkennbar in der Form

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

Diese *explizite Mittelpunktregel* hat nicht nur Ordnung 2, sondern besitzt sogar eine *asymptotische Entwicklung* nach Potenzen von h^2. Daher können mit Richardson-Extrapolation (vgl. Numerik I, §5.5) Näherungen beliebig hoher Ordnung berechnet werden. Das resultierende Gesamt-Verfahren ist vom Typus her aber wieder ein Einschrittverfahren und wird daher in einem eigenen Abschnitt behandelt.

1.4 Extrapolationsverfahren

Die explizite Mittelpunktregel 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

$$\begin{aligned} y_1 - y_0 &= hf(x_0, y_0), \quad y_0 = u_0, \\ y_{i+1} - y_{i-1} &= 2hf(x_i, y_i), \quad i = 1, 2, \ldots. \end{aligned} \tag{1.4.1}$$

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

$$u_\Delta(x_j) - u(x_j) = \sum_{k=1}^{q} h^{2k}\Big[g_{2k}(x_j) + (-1)^j \tilde{g}_{2k}(x_j)\Big] + \mathcal{O}(h^{2q+2}),\ h \to 0, \tag{1.4.2}$$

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 1.4.1 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 Differenzengleichung $y_{i+1} - 2h\lambda y_i - y_{i-1} = 0$ 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

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

Man sieht leicht, daß mit Ausnahme von $(-1)^{x/h}$ alle Ausdrücke eine Talyorentwicklung nach h um $h = 0$ besitzen. Daraus folgt hier für den Anfang der Entwicklung (1.4.2)

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

Wenn die Lösung der Dgl fällt, $\lambda < 0$, kann der oszillierende *und wachsende* Störterm $\tilde{g}_2$ die numerische Lösung erheblich beeinträchtigen. Dieses Problem läßt sich etwas durch einen zusätzlichen Glättungsschritt abschwächen.

Zur Extrapolation bei der Mittelpunktregel wählt man eine wachsende Folge gerader Zahlen $2 \le N_0 < \ldots < N_q$ und bildet Schrittweiten $H_j := H/N_j$ sowie Gitter $x_i^{[j]} := x_0 + iH_j$ $(i =$

$0,\dots,N_j$). Dies ergibt das *Verfahren von Gragg-Bulirsch-Stoer*:

	für $j = 0,\dots,q$:	
1	$y_0^{[j]} := u_0,\ y_1^{[j]} := y_0 + H_j f(x_0, y_0)$,	Startwerte
	für $i = 1,..,N_j$:	Mittelpunkt-
2	$y_{i+1}^{[j]} := y_{i-1}^{[j]} + 2H_j f(x_i^{[j]}, y_i^{[j]})$	-Regel
3	$P_{j,0} := \frac{1}{4}(y_{N_j-1}^{[j]} + 2y_{N_j}^{[j]} + y_{N_j+1}^{[j]})$,	Glättung
	für $k = 1,..,j$: $\quad i := j - k$;	Extrapo-
4	$P_{i,k} := P_{i+1,k-1} + \frac{1}{(N_j/N_i)^2-1}(P_{i+1,k-1} - P_{i,k-1})$	-lation

(1.4.3)

Im Verfahren (1.4.3) kann einfach die am langsamsten wachsende Folge $\{N_j\} = \{2,4,6,8,10,..\}$ verwendet werden. Schritt 3 ist der erwähnte Glättungsschritt.
In Schritt 4 wird die Richardson-Extrapolation ausgeführt, die dem Neville-Algorithmus (I.3.3.5) zur Polynominterpolation entspricht und der Reihe nach jeweils den führenden Term der asymptotischen Entwicklung (1.4.2) eliminiert (Numerik I, §5.5). Im j-ten Lauf durch Schritt 3-4 wird dabei mit der Näherung $P_{j,0}$ vom j-ten Gitter eine zusätzliche Schrägzeile unten an das Extrapolations-Tableau angehängt.

$$\begin{array}{lllll} P_{0,0} & P_{0,1} & P_{0,2} & \cdots & P_{0,j} \\ P_{1,0} & P_{1,1} & & & \\ P_{2,0} & \vdots & & & \\ \vdots & P_{j-1,1} & & & \\ P_{j,0} & & & & \end{array}$$

Danach gilt die (lokale) Fehleraussage

$$\frac{1}{H}[P_{0,j} - u(x_0 + H)] = \mathcal{O}(H^{2j+2}), \quad j \le q. \tag{1.4.4}$$

Dieses Endergebnis $P_{0,j}$ gehört zu einem Einschrittverfahren der Ordnung $2j+2$ für den Schritt $x \to x + H$, da es nicht auf Werte vor der Stelle x zugreift. Damit wurde insbesondere folgende Existenzaussage zur Theorie der Einschrittverfahren geliefert, die sich durch Abzählung der insgesamt in (1.4.3) benötigten Funktionsauswertungen ergibt.

Satz 1.4.2 *Für jedes $p \in 2\mathbf{N}$ gibt es ein Einschrittverfahren mit Ordnung p und $\frac{1}{4}p^2 + 1$ Stufen.*

Das Extrapolationsverfahren benutzt also mehr Funktionsauswertungen als die in §1.2 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 $2j + 2 (\le 2q + 2)$ in (1.4.4) in der Praxis nicht vorgegeben, sondern durch Inspektion des Tableaus $P_{i,k}$ bestimmt.

Demo-Beispiel: Extrapolationsverfahren mit Ordnungs- und Schrittweitensteuerung (vgl. Ende §1.3.1) beim AWP $u' = f(x,u), u(-\frac{1}{2}) = u_0$ auf $[-\frac{1}{2}, 1]$ mit $(a := 800, b = \ln(1+a) - 3)$

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

Vergleich der Verfahren:

	Einschritt		Extrapolation		Mehrschritt
Konsistenz-Ordnung	aufwendig		einfach	$\Leftarrow$	einfach
Stabilität	immer	$\Rightarrow$	immer		Einschränkung
Aufwand	merkbar		hoch		gering
Fehlerschätzung	möglich		einfach		einfach
Schrittweitenänd.	immer	$\Rightarrow$	immer		aufwendig
Ordnungswechsel	nein		einfach		möglich

Die *praktische* Erfahrung mit den drei behandelten Verfahrensklassen ergibt ein differenziertes Bild, keines kann als Universalverfahren bezeichnet werden. Zwar schneiden in Bezug auf die Gesamtanzahl der zur Lösung benötigten Funktionsaufrufe von f die Prädiktor-Korrektor-Verfahren am günstigsten ab. Da sie im adaptiven Einsatz beim Schrittweitenwechsel aber den höchsten Verwaltungsaufwand haben, sind sie tatsächlich nur selten schneller (in Bezug auf die Rechenzeiten) als Einschrittverfahren, nämlich in Fällen, wo die numerische Auswertung $f(x, y)$ sehr aufwendig 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.

1.5 Steife Anfangswertprobleme

In der Praxis trifft man oft auf Differentialgleichungen, deren Lösungen sehr unterschiedliche *Zeitskalen* besitzen, d.h., es können sowohl sehr schnell, als auch sehr langsam veränderliche Lösungen auftreten. Dies ist, z.B., bei chemischen Reaktionen der Fall. Von größerer Bedeutung im Rahmen dieser Vorlesung ist aber die Tatsache, daß bestimmte Approximationsverfahren für parabolische, partielle Differentialgleichungen (Linienmethode, vgl. §3.4) auf gewöhnliche Anfangswertprobleme dieser Art mit großen Dimensionen n führen. Die Behandlung solcher Probleme ist ein sehr aktuelles Forschungsgebiet, daher können nur die wichtigsten Themen angerissen werden. Eine ausführlichere Darstellung findet sich in den Büchern von Hairer/Wanner und Strehmel/Weiner. Zur Problematik:

Beispiel 1.5.1 $u(x) \in \mathbf{R}^2$,

$$u'(x) = Lu(x) := \begin{pmatrix} -500 & 499 \\ 499 & -500 \end{pmatrix} u(x), \quad u(0) = u_0 := \begin{pmatrix} 1 \\ 0 \end{pmatrix}. \tag{1.5.1}$$

Da die Eigenwerte der Koeffizientenmatrix L die Werte $\lambda_1 = -1$ und $\lambda_2 = -999$ sind, lautet die exakte Lösung dieses Problems

$$u(x) = \begin{pmatrix} C_1 e^{-x} + C_2 e^{\lambda_2 x} \\ C_1 e^{-x} - C_2 e^{\lambda_2 x} \end{pmatrix} \tag{1.5.2}$$

mit $C_1 = C_2 = \frac{1}{2}$. Bis auf eine sehr kurze Startphase $0 \le x \ll 1/1000$ verhält sich die Lösung wie die Funktion $\frac{1}{2}\binom{1}{1}e^{-x}$, trivialerweise gilt auch $u(x) \to 0,\ x \to \infty$. Wird das Problem, z.B., mit dem (expliziten) Euler-Verfahren und konstantem h behandelt, ergibt sich die numerische Lösung aus

$$y_{i+1} = (I + hL)y_i \quad \Rightarrow \quad y_i = \begin{pmatrix} C_1(1-h)^i + C_2(1+h\lambda_2)^i \\ C_1(1-h)^i - C_2(1+h\lambda_2)^i \end{pmatrix}.$$

Hier gilt offensichtlich $y_i \to 0$ $(i \to \infty)$ nur für $|1 + h\lambda_2| < 1$, d.h., wenn

$$h < \frac{2}{-\lambda_2} = \frac{1}{500}.$$

Im anderen Fall wächst y_i oszillierend stark an. Obwohl also (z.B.) $e^{\lambda_2 x} \le 10^{-8}$ für $x \ge 0.02$ ist und die Lösung für solche Argumente vollkommen glatt aussieht, auch durch einen Polygonzug mit größeren Schrittweiten gut approximiert werden kann, erzwingt das Vorhandensein des Eigenwerts λ_2 im Problem die Verwendung kleiner Schrittweiten *bei diesem Verfahren.*

Beim *impliziten* Euler-Verfahren wird das Integral $\int_0^1 f(x_i+th, u(x_i+th))\,dt$ in (1.2.1) durch den Funktionswert am *rechten* Rand ersetzt. Bei diesem Verfahren,

$$y_{i+1} = y_i + h_i f(x_{i+1}, y_{i+1}) \iff \begin{cases} k_1 &= f(x_i + h_i, y_i + h_i k_1) \\ y_{i+1} &= y_i + h_i k_1 \end{cases} \tag{1.5.3}$$

sind k_1 bzw. y_{i+1} nur implizit beschrieben, müssen also durch Auflösung eines (nicht-)linearen Gleichungssystems bestimmt werden. Im linearen Beispiel (1.5.1) führt dies auf die Beziehungen

$$y_{i+1} = y_i + h_i L y_{i+1} \iff (I - h_i L)y_{i+1} = y_i \iff y_{i+1} = (I - h_i L)^{-1} y_i.$$

Mit Hilfe der Eigenwertzerlegung von L kann auch hier wieder das Ergebnis für konstante Schrittweite h explizit bestimmt werden,

$$y_i = \begin{pmatrix} C_1(1+h)^{-i} + C_2(1-h\lambda_2)^{-i} \\ C_1(1+h)^{-i} - C_2(1-h\lambda_2)^{-i} \end{pmatrix}.$$

Da jetzt $1/(1+h) < 1$ und $1/(1-h\lambda_2) < 1$ gilt, *für alle* $h > 0$, ist zumindestens das asymptotische Verhalten der Lösung u reproduziert, $y_i \to 0,\ i \to \infty$ für alle Schrittweiten.

Die beiden Beispielverfahren unterscheiden sich also ganz wesentlich in ihren Stabilitätseigenschaften. Zur Überprüfung solcher Eigenschaften von numerischen Lösungen verschiedener Verfahren betrachtet man deren Verhalten bei bestimmten *Testproblemen* im Vergleich zur exakten Lösung. Die einfachste Testgleichung ist die skalare, linear-homogene

$$u'(x) = \lambda u(x), \quad u(0) = 1, \quad \lambda \in \mathbf{C}, \quad \operatorname{Re}\lambda \le 0, \tag{1.5.4}$$

die aber über die Eigenvektor-Zerlegung auch Aussagen über Systeme der Form (1.5.1) gestattet. Für deren Lösungen gilt $|u(x+s)| = |e^{\lambda s}u(x)| \le |u(x)|\ \forall s \ge 0$. Runge-Kutta-Verfahren (1.2.4) führen hier auf die skalaren Gleichungen

$$\begin{aligned} y_{i+1} &= y_i + h\sum_{j=1}^m b_j k_j, \\ k_j &= \lambda y_i + h\lambda \sum_{l=1}^m a_{jl} k_l \qquad \iff \qquad (I - zA)k = \lambda y_i \mathbf{1} \end{aligned}$$

wobei $z := h\lambda$, $k := (k_1, .., k_m)^T$, $A = (a_{jl})$, $\mathbf{1} := (1, .., 1)^T$ gesetzt wurde. Analog sei $b = (b_1, .., b_m)^T$. Aus dieser Beziehung ergibt sich die Rekursion $y_i \to y_{i+1}$ explizit in der Form

$$\begin{aligned} y_{i+1} &= y_i + hb^T k = \Big[1 + zb^T(I - zA)^{-1}\mathbf{1}\Big] y_i = \varphi(z) y_i, && (1.5.5) \\ & \varphi(z) := 1 + zb^T(I - zA)^{-1}\mathbf{1}. && (1.5.6) \end{aligned}$$

Einschrittverfahren führen bei (1.5.4) immer auf eine solche Beziehung $y_{i+1} = \varphi(h\lambda)y_i$. In (1.5.5) ist $\varphi(z)$ eine rationale Funktion der komplexen Variabeln $z = h\lambda$, die sogenannte *Stabilitätsfunktion*. Für explizite Verfahren wurden diese schon in Beispiel 1.2.1 bzw. bei (1.5.3) berechnet:

Verfahren:	expl. Euler	Runge u.Heun (Nr.2/3)	klass. Runge-Kutta (Nr.4)	impl. Euler
$\varphi(z) =$	$1 + z$	$1 + z + \frac{1}{2}z^2$	$1 + z + \frac{1}{2}z^2 + \frac{1}{6}z^3 + \frac{1}{24}z^4$	$\frac{1}{1-z}$

Offensichtlich sind die Stabilitätsfunktionen der expliziten Verfahren ganze rationale Funktionen, d.h., Polynome. Für diese ist ja A eine strikt untere Dreiecksmatrix, also nilpotent mit $A^m = 0$. Daher ist die Neumannreihe ein Polynom, $\varphi(z) = 1 + zb^T(I - zA)^{-1}\mathbf{1} = 1 + z\sum_{j=0}^{m-1} z^j b^T A^j \mathbf{1}$. Wenn ein Verfahren mit Ordnung p konvergiert, bildet die Stabilitätsfunktionen natürlich eine rationale Approximationen an die e-Funktion mit $\varphi(z) = e^z + \mathcal{O}(z^{p+1})$, $z \to 0$.

Bei der Testgleichung (1.5.4), die nur beschränkte Lösungen besitzt, interessieren diejenigen Argumentwerte $z = h\lambda$, für die die numerische Lösung y ebenfalls nicht wächst.

Definition 1.5.2 *Ein Einschrittverfahren besitze die Stabilitätsfunktion φ, d.h. bei Anwendung auf (1.5.4) führe es auf die Beziehung $y_{i+1} = \varphi(h\lambda)y_i$. Das Verfahren heißt* absolut stabil *bei $z \in \mathbf{C}$, falls $|\varphi(z)| \le 1$ ist. Die Menge*

$$S := \{z \in \mathbf{C} : \ |\varphi(z)| \le 1\} \qquad (1.5.7)$$

heißt Bereich der absoluten Stabilität *des Verfahrens.*

Zur stabilen Integration von (1.5.4) muß bei einem gegebenen Verfahren die Schrittweite h so gewählt werden, daß $h\lambda \in S$ gilt, da sonst unsinnige, wachsende Lösungen auftreten können. Für die in $\mathbf{R}$ liegenden *Stabilitätsintervalle* $S \cap \mathbf{R}$ der obigen Beispielverfahren gilt

Verfahren:	expl. Euler	Runge u.Heun (Nr.2/3)	klass.Runge-Kutta (Nr.4)	impl. Euler
Intervall:	$[-2, 0]$	$[-2, 0]$	$[-2.78, 0]$	$(-\infty, 0]$

Die in Beispiel 1.5 für das explizite Euler-Verfahren gefundene Schrittweiten-Einschränkung

$h \leq 2/(-\lambda)$ entspricht also der Bedingung $h\lambda \in [-2, 0] = S \cap \mathbf{R}$. Von allen erwähnten Verfahren kommt bei Anwendung auf Probleme der Form (1.5.1) nur das implizite Eulerverfahren ohne Schrittweitenbeschränkung aus, da dessen Stabilitätsbereich die negative reelle Achse umfaßt. Da bei expliziten Verfahren die Stabilitätsfunktion φ ein Polynom ist, mit $|\varphi(z)| \to \infty,\ z \to \infty$, besitzen diese nur beschränkte Stabilitätsbereiche.

Daher läßt sich allgemein feststellen, daß zur Lösung *steifer* Probleme, mit möglicherweise schnell ausklingenden Lösungen, nur implizite Verfahren sinnvoll sind, da nur diese sich ausschließlich nach Genauigkeitsforderungen richten müssen. Bei expliziten Verfahren dagegen wird die Schrittweite nicht durch die gewünschte Genauigkeit bestimmt, sondern (sehr restriktiv) durch Stabilitätsschwächen. Eine automatische Schrittweitensteuerung erzeugt hier eine um den Grenzfall schwankende Schrittweitenfolge (vgl. Demo-Beispiel). In Anlehnung an das Lösungsverhalten der Dgl (1.5.4) zeichnet man Verfahren mit analogen Dämpfungseigenschaften aus.

Definition 1.5.3 *Ein Einschrittverfahren heißt* A-stabil, *wenn gilt*

$$S \supseteq \mathbf{C}_- := \{z \in \mathbf{C} : \ Re\, z \leq 0\}, \quad d.h., \quad |\varphi(z)| \leq 1 \ \forall z \in \mathbf{C}_-.$$

Offensichtlich ist das implizite Eulerverfahren A-stabil, da hier die Bedingung $|\varphi(z)| = |\frac{1}{1-z}| \leq 1 \iff 1 \leq |1-z|^2 = 1 - 2\mathrm{Re}\, z + |z|^2$ für $\mathrm{Re}\, z \leq 0$ sicher erfüllt ist.

Implizite Runge-Kutta-Verfahren:

Bei den IRK-Verfahren gibt es eine Klasse, bei der gleichzeitig die höchstmöglichen Konvergenzordnungen und gute Stabilitätseigenschaften aufeinandertreffen. Diese Verfahren ergeben sich aus Anwendung der Gaußschen Quadraturformeln auf (1.2.1). Dazu werden die Stützstellen c_i in (1.2.4) als Gaußknoten, d.h. die Nullstellen des m-ten Legendre-Polynoms gewählt (vgl. Numerik I, §5.3). Die restlichen Koeffizienten bestimmen sich als Integrale der zugehörigen Lagrangepolynome

$$L_i(x) := \prod_{\substack{j=1 \\ j \neq i}}^{m} \frac{x - c_j}{c_i - c_j},\ a_{ij} := \int_0^{c_i} L_j(x)\, dx,\ b_i := \int_0^1 L_i(x)\, dx, \quad i, j = 1, .., m. \tag{1.5.8}$$

Satz 1.5.4 *Für beliebiges $m \in \mathbf{N}$ sind die Gauß-Runge-Kutta-Verfahren A-stabil. Die Konsistenzordnung des m-stufigen Verfahrens ist $2m$.*

Die Gauß-Runge-Kutta-Verfahren besitzen also exzellente theoretische Eigenschaften. Es können sogar Stabilitätsaussagen für nichtlineare Probleme gezeigt werden. Allerdings sind bei diesen Verfahren in jedem Verfahrensschritt nichtlineare Gleichungssysteme der Größe mn, z.B. mit dem Newtonverfahren, zu lösen. Daher sind sie für den praktischen Einsatz meist zu teuer. Es gibt verschiedene Ansätze zur effektiveren Implementierung von IRK-Verfahren. Diese schränken aber entweder die Struktur ein, und damit leider auch die erreichbare Ordnung auf $m + 1$, oder besitzen andere Nachteile.

Eine effiziente Alternative sind *Linear-implizite Runge-Kutta-Verfahren.* Zur Motivation werde die Lösung der Gleichung (1.5.3) beim impliziten Euler-Verfahren betrachtet mit sog. autonomer rechter Seite $f = f(u)$. Der erste Schritt der Newton-Iteration mit Startwert $y_{i+1}^{[0]} := y_i$ und $L_i := \frac{\partial f}{\partial y}(y_i)$ ergibt

$$y_{i+1}^{[1]} = y_i + h(I - hL_i)^{-1} f(y_i).$$

Man kann nun diese Vorschriftt als eigenständiges Verfahren studieren in Bezug auf Stabilität und Konsistenz. Bei mehrstufigen Verfahren sind dabei zusätzliche Parameter γ, γ_{ij} sinnvoll. Die Verfahrensvorschrift lautet dann mit $L := \frac{\partial f}{\partial y}(y_i)$,

$$\begin{aligned}(I - h\gamma L)k_l &= f\Big(y_i + h\sum_{j=1}^{l-1} \alpha_{lj} k_j\Big) + hL \sum_{j=1}^{l-1} \gamma_{lj} k_j, \quad l = 1, \ldots, m \qquad (1.5.9)\\ y_{i+1} &= y_i + h\sum_{j=1}^{m} b_j k_j.\end{aligned}$$

Die Stabilitätsfunktion dieser Verfahren entspricht der eines IRK-Verfahrens mit Koeffizienten $a_{lj} = \alpha_{lj} + \gamma_{lj}$, daher gibt es A-stabile Verfahren. Da in allen Stufen die gleiche Matrix $(I - h\gamma L)$ auftritt, ist pro Schritt $y_i \to y_{i+1}$ nur eine LR- oder QR-Zerlegung erforderlich. Die Verfahren gehen auf Rosenbrock und Wanner zurück. Man kann auch hier eingebettete Verfahren studieren zur Schrittweitensteuerung, z.B., existieren verschiedene (3,4)-Paare. Diese sind für mittlere Genauigkeitsforderungen konkurrenzfähig zur folgenden Verfahrensklasse.

Implizite Mehrschrittverfahren

Im praktischen Einsatz hat sich eine spezielle Verfahrensklasse bewährt, die auf den rückwärtigen Differenzenformeln beruht. Im Gegensatz zu den Adams-Verfahren wird in der Form des allgemeinen Mehrschrittverfahrens (1.3.6) nur ein nichtverschwindender Koeffizient $\beta_0 \neq 0$ gewählt, während die Koeffizienten α_j aus dem linksseitigen Differenzenquotienten abgeleitet werden mit

$$\frac{1}{h}\sum_{j=0}^{m} \alpha_j v(-jh) = \beta_0 v'(0) + \mathcal{O}(h^m), \; h \to 0, \qquad (1.5.10)$$

für $v \in C^m[-m, 0]$. Aufgrund von Satz 1.3.2 besitzt daher die m-stufige Version dieser BDF-Verfahren (backward difference formula) Ordnung m. Die Koeffizienten sind

m	$j =$	1	2	3	4	β_0	δ_m
1	$\alpha_j =$	-1				1	90^o
2	$3\alpha_j =$	-4	1			$\frac{2}{3}$	90^o
3	$11\alpha_j =$	-18	9	-2		$\frac{6}{11}$	86^o
4	$25\alpha_j =$	-48	36	-16	3	$\frac{12}{25}$	73^o

Ihre Stabilitätseigenschaften sind allerdings weniger gut als bei Einschrittverfahren. Bei der Testgleichung (1.5.4) haben die BDF-Verfahren die Gestalt ($z := h\lambda$)

$$(1 - \beta_0 z) y_i + \sum_{j=1}^{m} \alpha_j y_{i-j} = 0, \quad i \geq m. \qquad (1.5.11)$$

Über den Lösungsansatz $y_i = C\xi^i$ kommt man wieder auf die allgemeine Lösung $y_i = \sum_{j=1}^m C_j \xi_j^i$, wenn das *charakteristische Polynom*

$$p_m(\xi; z) = (1 - \beta_0 z)\xi^m + \alpha_1 \xi^{m-1} + \ldots + \alpha_m$$

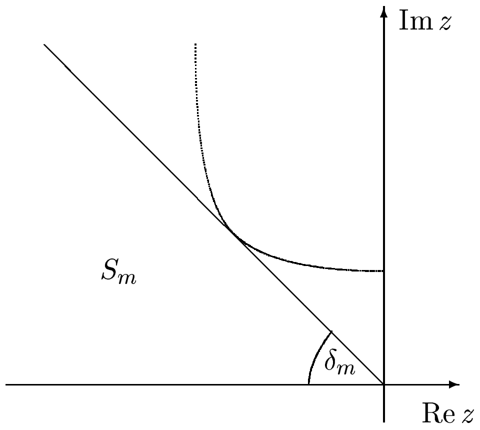
der Differenzengleichung (1.5.11) m verschiedene Nullstellen ξ_j besitzt. Bei mehrfachen Nullstellen ist die Darstellung durch Polynomanteile $i^l \xi^i$ zu modifizieren. In Analogie zu den Einschrittverfahren kann man die Menge dieser Nullstellen $\{\xi_j\} = \{\xi_j(z)\}_{j=1}^m$ als (mehrwertige) Stabilitätsfunktion des Mehrschrittverfahrens interpretieren. Insbesondere nennt man ein Mehrschrittverfahren absolut stabil an der Stelle $z \in \mathbf{C}$, wenn

$$p_m(\xi; z) = 0 \quad \Rightarrow \quad |\xi| < 1$$

gilt und A-stabil, wenn es für jedes $z \in \mathbf{C}_-$ absolut stabil ist. Allerdings gibt es eine zweite *Ordnungsbarriere* von Dahlquist:

Satz 1.5.5 *Die Ordnung eines A-stabilen Mehrschrittverfahrens ist höchstens* 2.

Daher können die Stabilitätsbereiche S_m der BDF-Verfahren höherer Ordnung $m \geq 2$ nicht mehr die ganze Halbebene $\mathbf{C}_-$ umfassen, in der obigen Tabelle ist der (halbe) Öffnungswinkel δ_m des größten in S_m gelegenen Winkelraums angegeben. Für $m > 6$ sind die Verfahren unbrauchbar, da dann nicht einmal mehr die negative reelle Achse vollständig in S enthalten ist. Auch für $m = 6$ ist der Winkel $\delta_6 = 17.8^o$ schon sehr klein (vgl. Diagramme).



Demo-Beispiel Das nichtlineare System von Dgln

$$\begin{aligned} u_j' &= -jdu_j, & j &\in \{1, n-1, n\} \\ u_j' &= -jdu_j + (u_{j-1}u_{j+2} - u_{j+1}u_j)^2, & j &= 2, \ldots, n-2 \end{aligned}$$

besitzt die Lösung $u_j(t) = e^{-jdt}$. Für große n ist die Lipschitzkonstante der rechten Seite groß ($\geq dn$). Die Stabilitätsprobleme expliziter Verfahren mit Schrittweitensteuerung sind daran erkennbar, daß Schrittweiten und Rechenzeit i.w. unabhängig von der geforderten Genauigkeit sind.

1.6 Schießverfahren für Randwertprobleme

Hier wird das standardisierte, nichtlineare Randwertproblem (RWP, vgl. §1.1)

$$u'(x) = f(x, u(x)),\ x \in (a, b), \quad r(u(a), u(b)) = 0, \tag{1.6.1}$$

betrachtet mit differenzierbaren Funktionen $f(x, y),\ r(y_0, y_1)$. Bei der numerischen Behandlung solcher gewöhnlicher RWPe kann man sowohl Lösungsverfahren von den Anfangswertproblemen übertragen, als auch neue Verfahren untersuchen, die auf partielle Randwertprobleme verallgemeinert werden können. Aus der Reihe der letzteren wird eine spezielle Klasse schon hier besprochen, da deren wesentliche Eigenschaften im gewöhnlichen Fall einfacher zu diskutieren sind. Zunächst wird mit Methoden begonnen, die auf AWP-Verfahren aufbauen. Diese haben den praktischen Vorteil, daß Standardprogramme aus diesem Bereich eingesetzt werden können.

Schießverfahren

Aufbauend auf dem letzten Paragraphen wird vorausgesetzt, daß Anfangswertprobleme (mit genügender Genauigkeit) numerisch gelöst werden können. In §1.1 war beim linearen Randwertproblem

$$u'(x) = A(x)u(x) + g(x),\ x \in (a, b), \quad R_0u(a) + R_1u(b) = d, \tag{1.6.2}$$

mit einer Fundamentallösung $Y(x)$ des Matrix-Anfangswertptoblems $Y'(x) = A(x)Y(x), Y(a) = I$, aus der Lösungsdarstellung $u(x) = Y(x)\eta + ..$ der unbekannte *Anfangswert* $\eta = u(a) \in \mathbf{R}^n$ durch Einsetzen in die Randbedingung bestimmt worden: $R_0\eta + R_1[Y(b)\eta + ..] = d$. 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 Lösung des AWPs

$$\frac{d}{dx}u(x; a, \eta) = f(x, u(x; a, \eta)),\ x > a, \quad u(a; a, \eta) = \eta. \tag{1.6.3}$$

Die spezielle Lösung des RWPs (1.6.1) ist dann diejenige, $u(x) = u(x; a, \hat{\eta})$, deren Anfangswert $\hat{\eta}$ das nichtlineare Gleichungssystem

$$F(\eta) = 0, \quad \text{mit } F : \begin{cases} \mathbf{R}^n & \to & \mathbf{R}^n \\ \eta & \mapsto & r\Big(\eta, u(b; a, \eta)\Big) \end{cases} \tag{1.6.4}$$

erfüllt. Zur Lösung solcher Gleichungssysteme kommt vor allem das Newton-Verfahren in Betracht, vgl. Numerik I, §6.2. 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 1.1.1 gezeigt. Darüber hinaus gilt aber auch

Satz 1.6.1 *Die rechte Seite f sei zweimal stetig differenzierbar, ihre Ableitung $f_y = \frac{\partial f}{\partial y}$ beschränkt auf $[a,b] \times \mathbf{R}^n$. Dann hängt die Lösung $u(x;a,\eta)$ von (1.6.3) differenzierbar vom Anfangswert η ab. Ihre Ableitung an der Stelle $(x, u(x;a,\eta))$ ist*

$$\frac{\partial}{\partial \eta} u(x;a,\eta) = Y(x;a,\eta), \tag{1.6.5}$$

wobei Y das Fundamentalsystem zur linearisierten Differentialgleichung ist,

$$\frac{\partial}{\partial x} Y(x;a,\eta) = f_y(x, u(x;a,\eta)) \cdot Y(x;a,\eta), \quad Y(a;a,\eta) = I. \tag{1.6.6}$$

Beweis Die Ableitung wird direkt als in s linearer Anteil der Differenz $v(x) := u(x;a,\eta+s) - u(x;a,\eta),\ s \in \mathbf{R}^n$, bestimmt. Dies entspricht der Definition der Differenzierbarkeit. Offensichtlich ist $v(a) = \eta + s - \eta = s$. Nach Satz 1.1.1 gilt daher die Schranke $\|v(x)\| \le \exp(L(x-a))\|s\|$. Da beide Funktionen u Lösungen der Dgl sind, gilt auch für $x \in (a,b)$ (mit $' = \partial/\partial x$)

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

Die Funktion $\hat{v}(x) := Y(x;a,\eta)s$ mit Y aus (1.6.6) ist Lösung des AWPs $\hat{v}'(x) = f_y(x, u(x;a,\eta))\hat{v}(x)$, $\hat{v}(a) = s$. Daher gilt insbesondere $v(x) - \hat{v}(x) = \mathcal{O}(\|s\|^2)$ und somit

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

Damit erfüllt der in s lineare Anteil Ys die Definition der Ableitung von u in Richtung s, also gilt $\frac{\partial}{\partial \eta} u = Y$, wie in (1.6.5) behauptet. ∎

Die im Newtonverfahren zur Lösung des Systems (1.6.4) benötigte Ableitung ergibt sich somit nach der Kettenregel

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

Die Bestimmung von F' ist somit durch Lösung des Matrix-AWPs (1.6.6), also durch (numerische) Berechnung von n Lösungen $y_i(x)$ der linearen AWPe $y_i' = f_y(x, u(x;a,\eta))y_i,\ y_i(a) = e_i, i = 1,..,n$, möglich ($e_i$: i-ter Einheitsvektor). Zur praktischen Durchführung können alle im §1.2-1.4 behandelten Verfahren herangezogen werden, es ist aber einfacher von einer exakten Lösung der AWPe auszugehen. Damit sind die beim Newtonverfahren

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

auftretenden Größen 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 1.6.2 *Schießverfahren:* Gegeben sei $\eta^{[0]} \in \mathbf{R}^n$. Für $k = 0, 1, \dots$ löse man

$$\text{die AWPe} \quad \begin{cases} u'(x) = f(x, u(x)), & u(a) = \eta^{[k]}, \\ Y'(x) = f_y(x, u(x))Y(x), & Y(a) = I, \end{cases} \tag{1.6.7}$$

$$\text{das LGS} \quad \Big(r_{y_0} + r_{y_1} Y(b)\Big)\Big(\eta^{[k+1]} - \eta^{[k]}\Big) = -r(\eta^{[k]}, u(b)). \tag{1.6.8}$$

Die partiellen Ableitungen von r sind ebenfalls an der Stelle $(\eta^{[k]}, u(b))$ auszuwerten. Zunächst sind also in (1.6.7) jeweils $n+1$ gekoppelte AWPe zu lösen und danach in (1.6.8) eine lineares $n \times n$-System.

Im linearen Spezialfall (1.6.2) 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

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

Dies ist genau der Lösungswert, der bei der expliziten Lösung in (1.1.10) berechnet wurde, (1.6.8) 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 RWPs die Randmatrix $F'(u(a)) = r_{y_0} + r_{y_1} Y(b)$ regulär ist, wobei $Y(b) = Y(b; a, u(a))$. In der Praxis ist diese Aussage aber oft wertlos, da die Lösungen $u(x; a, \eta)$ sich möglicherweise sehr schnell von der gesuchten Lösung $u(x; a, u(a))$ entfernen. Dann ist der *Einzugsbereich* der Iteration (1.6.8), und eventuell sogar der Definitionsbereich der Funktion F überhaupt, sehr klein.

Beispiel 1.6.3 Diese Probleme können anhand der allgemeinen Lösung der skalaren Dgl

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

erläutert werden. Soll das RWP mit der Randbedingung $u(0) + u(b) = b$, $b > 0$, (Lösung $u(x) \equiv x$) mit dem Schießverfahren gelöst werden, ist ein Startwert $\eta < 1/b$ zu wählen, damit die Lösung $u(x; 0, \eta) = (\eta + x)/(1 - \eta x)$ überhaupt den rechten Intervallrand erreicht. Mit der explizit bekannten allgemeinen Lösung erhält die Randgleichung (1.6.4) die Form

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

Die Ableitung $F'(\eta)$ kann hieraus natürlich direkt bestimmt werden. Andererseits gilt auch

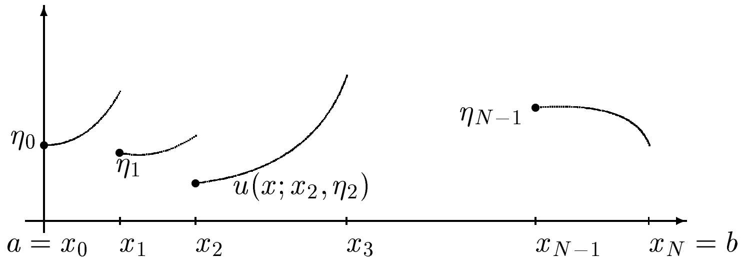
$$Y(x) = \frac{\partial u}{\partial \eta}(x; 0, \eta) = \frac{1 + x^2}{(1 - \eta x)^2} \quad \text{und} \quad Y'(x) = 2\frac{\eta + x}{(1 - \eta x)(1 + x^2)} Y(x) = \underbrace{\frac{2u(x; 0, \eta)}{1 + x^2}}_{f_y(x,u)} Y(x).$$

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

Die oben erwähnten Schwierigkeiten bei rasch divergierenden Lösungen des Anfangswertproblems können entschärft werden dadurch, daß man die Lösungen jeweils nur über kürzere Intervalle integriert, das Gesamtintervall also in mehrere Teile zerlegt. Dies führt auf die im folgenden beschriebene **Mehrzielmethode**:

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

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



In jedem Teilintervall $[x_i, x_{i+1}]$ wird das Schießverfahren getrennt angewendet. Dabei werden N Teil-Lösungen $u(x; x_i, \eta_i)$ zu

$$u'(x; x_i, \eta_i) = f(x, u(x; x_i, \eta_i)), \quad u(x_i; x_i, \eta_i) = \eta_i, \quad i = 0, \ldots, N-1, \tag{1.6.9}$$

berechnet. Eine Lösung u des Randwertproblems (1.6.1) kann nun aus diesen Stücken genau dann zusammengesetzt werden, wenn diese an den Gitterpunkten 'zusammenpassen' und außerdem die Randbedingung erfüllen, wenn also gilt

$$\begin{aligned} u(x_i; x_{i-1}, \eta_{i-1}) = \eta_i \qquad & \equiv u(x_i; x_i, \eta_i), \; i = 1, \ldots, N-1, \\ r(\eta_0, u(x_N; x_{N-1}, \eta_{N-1})) = 0. & \end{aligned}$$

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 eine Lösung des nichtlinearen Gleichungssystems für

$$\vec{\eta} := \begin{pmatrix} \eta_0 \\ \eta_1 \\ \vdots \\ \eta_{N-1} \end{pmatrix} \in \mathbf{R}^{nN} \text{ mit } F(\vec{\eta}) := \begin{pmatrix} u(x_i; x_{i-1}, \eta_{i-1}) - \eta_i, & i = 1, .., N-1, \\ r(\eta_0, u(x_N; x_{N-1}, \eta_{N-1})), & i = N \end{pmatrix} = 0 \tag{1.6.10}$$

Hierin ist das einfache Schießverfahren mit $N = 1$ enthalten. Beim linearen RWP (1.6.2) können die Stück-Lösungen (1.6.9) wieder explizit dargestellt werden (vgl. Satz 1.1.3) durch

$$u(x; x_i, \eta_i) = Y(x) Y(x_i)^{-1} \eta_i + \int_{x_i}^{x} Y(x) Y(t)^{-1} g(t) dt.$$

Dabei sind $Y(x)Y(x_i)^{-1} =: Y(x; x_i)$ diejenigen (hier von η_i unabhängigen) Fundamentallösungen der linearen Dgl mit $Y(x_i; x_i) = I$. Die Integrale entsprechen den speziellen Lösungen $u(x; x_i, 0)$. Das System (1.6.10) ist hier dann auch linear und lautet

$$\begin{aligned} Y(x_i; x_{i-1}) \eta_{i-1} - \eta_i &= -u(x_i; x_{i-1}, 0), \quad i = 1, .., N-1, \\ R_0 \eta_0 + R_1 Y(b; x_{N-1}) \eta_{N-1} &= d - Ru(b; x_{N-1}, 0). \end{aligned}$$

Die $(nN) \times (nN)$-Matrix dieses Systems hat eine spezielle Struktur. Sie besitzt eine zyklische Block-Bidiagonalgestalt (Treppenform) aus $n \times n$-Blöcken,

$$\begin{pmatrix} Y(x_1;x_0) & -I & & & \\ & Y(x_2;x_1) & -I & & \\ & & \ddots & \ddots & \\ & & & Y(x_{N-1};x_{N-2}) & -I \\ R_0 & & & & R_1 Y(b;x_{N-1}) \end{pmatrix} \tag{1.6.11}$$

Bei der Lösung des allgemeinen, nichtlinearen Problems (1.6.10) mit dem Newtonverfahren treten lineare Gleichungssysteme der gleichen Struktur auf. Denn die partiellen Ableitungen der Teillösungen nach ihrem Anfangswert sind nach Satz 1.6.1 wieder,

$$\begin{aligned} \tfrac{\partial}{\partial \eta_i} u(x;x_i,\eta_i) &= Y(x;x_i,\eta_i) \\ Y'(x;x_i,\eta_i) = f_y(x,u(x;x_i,\eta_i))Y(x;x_i,\eta_i), &\quad Y(x_i;x_i,\eta_i) = I. \end{aligned} \tag{1.6.12}$$

Daher besitzt die Ableitungsmatrix $F' = \partial F/\partial \vec{\eta}$ von (1.6.10) genau die Gestalt (1.6.11). Somit ist beim Newtonverfahren zur Lösung des Systems (1.6.10),

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

im k-ten Schritt für die Korrekturen

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

tatsächlich ein lineares Gleichungssystem der folgenden Form zu lösen,

$$\begin{aligned} Y(x_i;x_{i-1},\eta_{i-1})\delta_{i-1} - \delta_i &= \eta_i - u(x_i;x_{i-1},\eta_{i-1}), \quad i = 1,..,N-1, \\ R_0\delta_0 + R_1 Y(b;x_{N-1},\eta_{N-1})\delta_{N-1} &= -r(\eta_0, u(b;x_{N-1},\eta_{N-1})), \end{aligned} \tag{1.6.13}$$

mit $R_0 = \frac{\partial r}{\partial y_0}$, $R_1 = \frac{\partial r}{\partial y_1}$, 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 (1.6.9), (1.6.12).

Die *Auflösung* der linearen Systeme in (1.6.13) kann mit dem Gauß-Algorithmus durchgeführt werden. Bei Pivotisierung mit Spaltenvertauschung werden dabei Elemente nur in der untersten (Block-) Zeile neu eingeführt und es wird nur mit dem skizzierten Bereich gearbeitet:

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

Dieses Verfahren ist auch in kritischen Fällen (z.B., wenn die Matrizen Y große Normen besitzen) numerisch recht stabil. In vielen Fällen kann dieses Gleichungssystem aber noch einfacher durch

Kondensation gelöst werden. Dazu behandelt man nur δ_0 als Unbekannte und eliminiert die restlichen δ_i mit Hilfe der ersten Gleichungen in (1.6.13):

$$\begin{aligned} \delta_1 &= Y_1\delta_0 + r_1, \quad \delta_2 = Y_2\delta_1 + r_2 = Y_2Y_1\delta_0 + Y_2r_1 + r_2, \quad \ldots \\ \delta_{N-1} &= Y_{N-1}Y_{N-2}..Y_1\delta_0 + \hat{r}, \end{aligned}$$

$Y_i := Y(x_i; x_{i-1}, \eta_{i-1})$. Mit der letzten Gleichung aus (1.6.13) ergibt dies also das System

$$\Big(R_0 + R_1Y_NY_{N-1}..Y_2Y_1\Big)\delta_0 = -r(\eta_0, ..) - R_1Y_N\hat{r}, \tag{1.6.14}$$

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

AWPe mit Kondensation, Lösung δ_0 aus (1.6.14), Berechnung von $\delta_1, \ldots, \delta_{N-1}$.

Im linearen Fall und in der Nähe der Lösung u unterscheidet sich die Mehrzielmethode mit Kondensation nicht vom einfachen Schießverfahren. Denn, für die Matrix in (1.6.14) gilt mit exakten Startwerten $\eta_i \equiv u(x_i)$ die Aussage

$$R_0 + R_1Y_n \ldots Y_1 = R_0 + R_1Y(b; x_{N-1}, u(x_{N-1})) \ldots Y(x_1; a, u(a)) = R_0 + R_1Y(b; a, u(a)).$$

Letzteres ist aber genau die Matrix aus (1.6.8), da in der Lösung die (Halb-)Gruppenidentität $Y(x; s, u(s))Y(s; t, u(t)) = Y(x; t, u(t))\ \forall t < s < x$ gilt. Die lokale Konvergenz (bei u) des Newtonverfahrens hängt also insbesondere nur von der Regularität dieser Matrix ab. Der entscheidende Vorteil der Mehrzielmethode wirkt sich erst im nichtlinearen Fall aus durch (wesentliche) Vergrößerung des Einzugs-, d.h., Konvergenzbereichs beim Newtonverfahren. Die Wahl der Zwischenzielpunkte $x_1, .., x_{N-1}$ kann dabei leicht während der Durchführung des Verfahrens modifiziert werden. Z. B. ist es sinnvoll, bei starkem Anwachsen einer Zwischenlösung $u(.; x_i, \eta_i)$ einen neuen Zielpunkt einzufügen. Die Auswirkungen der Mehrzielmethode kann anhand des einführenden Beispiels diskutiert werden.

Beispiel 1.6.4 Eine Diskussion der Konvergenz ist hier naturgemäß viel schwieriger als im Fall $N = 1$. Als Mindestvoraussetzung kann aber die Existenz der Teillösungen in $[x_i, x_{i+1}]$ überprüft werden. Beim Problem aus Beispiel 1.6 lautet der Lösungswert in x_{i+1},

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

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

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

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 1.6 $\eta - u(x_0) < 1/b$ erforderlich war.

Demo-Beispiel: Mehrzielmethode beim Randwertproblem 2. Ordnung $u'' = \frac{3}{2}u^2$, $u(0) = 4, u(1) = 1$. Dieses besitzt 2 Lösungen.

1.7 Differenzenverfahren für Randwertprobleme

Die Konstruktion von numerischen Verfahren hatte bisher an der zur Dgl erster Ordnung äquivalenten Integralgleichung angesetzt. Bei allgemeineren Problemstellungen ist aber eine solche Umformung nicht einfach möglich bzw. unbequem, z.B. wenn ursprünglich Dgln höherer Ordnung vorliegen. Hier ist ein Zugang, der auf einer direkten Approximation der Ableitungen aufbaut, naheliegender. Die Standardmethode ist dabei die Approximation von Ableitungen durch Differenzenausdrücke. Diese können explizit konstruiert oder systematisch durch Ableitung von Interpolationspolynomen bestimmt werden (vgl. Numerik I, §5.6). Eine besonders einfache Form besitzen dabei die symmetrischen Differenzenquotienten der Ordnung 2. Hier gilt

Lemma 1.7.1 *Für* $u \in C^{m+2}[-\epsilon, \epsilon]$, $0 < \frac{m}{2}h \leq \epsilon$, *ist*

$$\frac{1}{h^m}\sum_{j=0}^{m}\binom{m}{j}(-1)^j u\Big((j-\frac{m}{2})h\Big) = u^{(m)}(0) + C_m h^2 u^{(m+2)}(\xi),\ \xi \in (-\epsilon, \epsilon). \tag{1.7.1}$$

'Beweis': Eine einfache Herleitung der Formel ist durch formale Rechnung mit Operatorreihen bei $C^\infty(\mathbf{R})$-Funktionen u (z.B. Polynomen) möglich. Mit $h > 0$ werden Verschiebungs- und Ableitungsoperatoren punktweise durch

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

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

$$S_h u(x) = u(x+h) = \sum_{j=0}^{\infty}\frac{h^j}{j!}u^{(j)}(x) = \sum_{j=0}^{\infty}\frac{h^j}{j!}D^j u(x) = e^{hD}u(x), \quad \text{d.h., } S_h = e^{hD}.$$

Durch Differenzbildung erhält man Darstellungen für Differenzenoperatoren, etwa $S_h - I = e^{hD} - I$. Der symmetrische Differenzenausdruck in (1.7.1) folgt aus der binomischen Formel bei

$$\begin{aligned} S_{h/2} - S_{-h/2} &= e^{\frac{h}{2}D} - e^{-\frac{h}{2}D} = 2\sinh\frac{h}{2}D \quad \Rightarrow \\ (S_h - I)^m S_{-h/2}^m &= (S_{h/2} - S_{-h/2})^m = \Big(2\sinh\frac{h}{2}D\Big)^m \\ &= \Big(hD + \frac{h^3}{24}D^3 + ..\Big)^m = h^m\Big(D^m + \frac{m}{24}h^2 D^{m+2} + ..\Big). \end{aligned} \tag{1.7.2}$$

Dies entspricht (1.7.1), wobei durch einige Zusatzüberlegungen die Darstellung des (Taylor-) Restglieds vereinfacht wurde. ■

Von besonderem Interesse sind Differenzenverfahren bei (Systemen von) Dgln höherer Ordnung, da sich mit ihnen der Übergang zum äquivalenten System erster Ordnung, mit einer vielfachen Anzahl von Variablen, umgehen läßt. Als Vorbereitung auf die partiellen Dgln wird das häufig auftretende lineare Randwertproblem zweiter Ordnung betrachtet

$$-u''(x) + p(x)u'(x) + q(x)u(x) = g(x), \quad u(a) = \alpha,\ u(b) = \beta. \tag{1.7.3}$$

In den beiden Randpunkten sind also jeweils 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 $q(x) \geq q^*$, besitzt (1.7.3) eine eindeutige Lösung. Zur Approximation auf einem äquidistanten Gitter

$$\Delta = \{x_i := a + ih,\ i = 0, \ldots, N+1\},\ h := \frac{b-a}{N+1},$$

werden die Ableitungen u'', u' durch symmetrische Differenzenquotienten ersetzt,

$$\begin{aligned} u''(x_i) &= \frac{u(x_{i+1}) - 2u(x_i) + u(x_{i-1})}{h^2} - C_2 h^2 u^{(4)}(\xi_i), \\ u'(x_i) &= \frac{u(x_{i+1}) - u(x_{i-1})}{2h} - 4C_1 h^2 u^{(3)}(\zeta_i). \end{aligned}$$

Die Gitterwerte $u_i := u(x_i),\ i = 0,..,N$, der Lösung (analog $p_i := p(x_i), q_i := .., g_i := ..$) sind am Rand gegeben $u_0 = \alpha,\ u_{N+1} = \beta$, und erfüllen für $i = 1, \ldots, N$ das Gleichungssystem

$$\begin{aligned} \frac{1}{h^2}(-u_{i-1} + 2u_i - u_{i+1}) + \frac{p_i}{2h}(u_{i+1} - u_{i-1}) + q_i u_i &= g_i + T_h(x_i), \quad \Longleftrightarrow \\ \frac{1}{h^2}\Big(-(1 + \frac{h}{2}p_i)u_{i-1} + (2 + h^2 q_i)u_i - (1 - \frac{h}{2}p_i)u_{i+1}\Big) &= g_i + T_h(x_i). \end{aligned} \tag{1.7.4}$$

Dabei enthält T_h die Fehler der Differenzenquotienten und besitzt die Größenordnung $\mathcal{O}(h^2)$, $h \to 0$ (s.u.). Die Matrix dieses Gleichungssystems ist tridiagonal und hat die Form

$$A := \frac{1}{h^2}\begin{pmatrix} 2 + h^2 q_1 & -1 + \frac{h}{2}p_1 & & & \\ -1 - \frac{h}{2}p_2 & 2 + h^2 q_2 & -1 + \frac{h}{2}p_2 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 - \frac{h}{2}p_{N-1} & 2 + h^2 q_{N-1} & -1 + \frac{h}{2}p_{N-1} \\ & & & -1 + \frac{h}{2}p_N & 2 + h^2 q_N \end{pmatrix}. \tag{1.7.5}$$

Durch Vernachlässigung der unbekannten lokalen Fehler T_h in (1.7.4) ergibt sich ein Gleichungssystem mit bekannten Koeffizienten für den Vektor $y := (y_1, \ldots, y_N)^\mathsf{T}$ einer Näherungslösung u_Δ, $y_i = u_\Delta(x_i)$. Das System besitzt die Gestalt

$$Ay = \tilde{g}, \qquad \tilde{g} = \begin{pmatrix} g_1 \\ g_2 \\ \vdots \\ g_{N-1} \\ g_N \end{pmatrix} + \frac{1}{h^2}\begin{pmatrix} \alpha(1 + \frac{h}{2}p_1) \\ 0 \\ \vdots \\ 0 \\ \beta(1 - \frac{h}{2}p_N) \end{pmatrix}, \tag{1.7.6}$$

mit der Matrix A aus (1.7.5). Es ist leicht zu zeigen, daß dieses Gleichungssystem (1.7.6) eindeutig lösbar ist, da die Matrix A viele Eigenschaften besitzt, die auch das ursprüngliche RWP (1.7.3) hat (bei geeigneter Übertragung der Begriffe). Zum Beispiel ist A für $p \equiv 0, q \geq 0$, symmetrisch und positiv definit, genauso wie die zum RWP gehörige lineare Abbildung (zwischen Funktionenräumen). Auf dieser Eigenschaft beruht ein Verfahren, das später behandelt wird. Zunächst wird aber nach einem anderen Beweisprinzip vorgegangen.

Satz 1.7.2 *Für die Koeffizientenwerte in (1.7.5) gelte* $q_i \geq q^* > 0, h|p_i| \leq 2, i = 1, \ldots, N$. *Dann ist die Matrix* A *streng diagonaldominant, also regulär, und es gilt (elementweise)*

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

Beweis Nach Voraussetzung sind die Nebendiagonalelemente von $A = (a_{ij})$ nicht-positiv. Daher gilt $A = D - B$ mit der Diagonalmatrix $D = \text{diag}(a_{ii}) = \text{diag}(\frac{2}{h^2} + q_1, .., \frac{2}{h^2} + q_N)$ und einer nichtnegativen Matrix B. Die Zeilensummen von $D^{-1}B$ für $i = 2, .., N-1$ sind

$$0 \leq \frac{1}{a_{ii}}(a_{i,i-1} + a_{i,i+1}) = \frac{1 + \frac{h}{2}p_i + 1 - \frac{h}{2}p_i}{2 + h^2 q_i} = \frac{2}{2 + h^2 q_i} \leq \frac{2}{2 + h^2 q^*} < 1.$$

Für $i = 1, N$ gilt eine analoge Aussage. Damit folgt $\|D^{-1}B\|_\infty < 1$ und die Neumann-Reihe

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

konvergiert. Durch Übergang zu Normen in dieser Reihe ergibt sich mit

$$\|A^{-1}\|_\infty \leq \sum_{j=0}^{\infty} \|D^{-1}B\|_\infty^j \|D^{-1}\|_\infty = \frac{\|D^{-1}\|_\infty}{1 - \|D^{-1}B\|_\infty} \leq \frac{h^2}{(2 + h^2 q^*)(1 - \frac{2}{2+h^2 q^*})} = \frac{1}{q^*}.$$

die zweite Aussage. ■

Der Satz erfordert ein positives q, hängt aber nicht von der Länge des Intervalls ab. Mit etwas mehr Aufwand läßt sich die Regularität von A unter Voraussetzungen zeigen, die nahe bei der Eindeutigkeitsbedingung an das RWP sind.

Der Satz ist auch von praktischem Interesse, da die Diagonaldominanz die Durchführbarkeit des Gaußalgorithmus ohne Zeilenvertauschung beim Tridiagonalsystem (1.7.6) garantiert mit $\mathcal{O}(N)$ Operationen. Aus $A^{-1} \geq 0$ folgt, daß die Lösung y monoton von der Inhomogenität g abhängt, und daß daher, z.B., $y \geq 0$ gilt, falls $g \geq 0, \alpha, \beta \geq 0$. Die Normschranke $\|A^{-1}\| \leq \frac{1}{q^*}$ schließlich ist eine Stabilitätsaussage, aus der direkt eine Fehlerschranke für das Verfahren folgt.

Satz 1.7.3 *Für die Koeffizientenfunktionen in (1.7.3) gelte* $p, q, g \in C^2[a, b]$ *und* $q(x) \geq q^* > 0$. *Die Schrittweite im Verfahren sei so, daß* $h|p(x)| \leq 2\ \forall x \in \Delta$. *Dann gilt für den globalen Fehler der Näherung* u_Δ *aus (1.7.6) mit einer von* Δ *unabhängigen Konstanten* C *die Aussage*

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

Beweis Nach den einleitenden Bemerkungen erfüllt der Gittervektor $\vec{u}$ der Lösung u das System $A\vec{u} = g + \tau,\ \tau = (T_h(x_1), .., T_h(x_N))^T$. Durch Subtraktion von (1.7.6) folgt

$$A(\vec{u} - y) = \tau \quad \Rightarrow \quad \|u - y\|_\infty \leq \|A^{-1}\|_\infty \|\tau\|_\infty \leq \frac{1}{q^*}\|\tau\|_\infty.$$

Unter den Voraussetzungen des Satzes ist $u \in C^4[a,b]$. Daher gilt für die Konsistenzfehler T_h nach (1.7.4) und Lemma 1.7.1

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

Dieses einfache Differenzenverfahren erzeugt also Näherungen an u von der (niedrigen) Konvergenzordnung zwei. Bei gewöhnlichen Randwertproblemen ist dies im Vergleich zu den Möglichkeiten des Schießverfahrens zusammen mit Anfangswertverfahren höherer Ordnung uninteressant. Mit Hilfe der im Beweis von Lemma 1.7.1 gefundenen formalen Identität, $\frac{1}{2}(S_{h/2} - S_{-h/2}) = \sinh \frac{h}{2} D$, aus der, z.B., folgt

$$
D^2 = \Big(\frac{2}{h}\operatorname{arsinh}\frac{1}{2}(S_{h/2}-S_{-h/2})\Big)^2 = \frac{1}{h^2}\Big((S_{h/2}-S_{-h/2})^2 - \frac{1}{12}(S_{h/2}-S_{-h/2})^4 + \dots\Big),
$$

können aber auch Differenzenapproximationen höherer Ordnung hergeleitet werden. So führen die in der letzten Gleichung angegebenen Terme auf die (wieder symmetrische) Differenzenformel

$$
\begin{aligned}
&\frac{1}{h^2}\Big(S_{-h} - 2I + S_h - \frac{1}{12}(S_{-2h} - 4S_{-h} + 6I - 4S_h + S_{2h})\Big)u(x) \\
&\quad= \frac{1}{12h^2}\Big(-u(x-2h) + 16u(x-h) - 30u(x) + 16u(x+h) - u(x+2h)\Big) \\
&\quad= u''(x) - \frac{h^4}{90}u^{(6)}(\xi). \qquad (1.7.7)
\end{aligned}
$$

In den randnahen Punkten x_1, x_N sind allerdings andere, unsymmetrische Formeln zu verwenden. Analoge Approximationen für u' sind ebenfalls leicht herzuleiten. Da die Formel (1.7.7) Funktionswerte aus 5 Gitterpunkten benutzt, ergibt sich nach Approximation der Dgl anstelle der Tridiagonalmatrix (1.7.5) eine mit fünf besetzten Diagonalen. Auch gehen weitere Eigenschaften, wie Diagonaldominanz, verloren, sodaß auch die einfache Stabilitätsaussage von Satz 1.7.2 nicht mehr möglich ist.

Bei speziellen Gleichungen ohne Beteiligung der ersten Ableitung u' (d.h. $p \equiv 0$) kann allerdings ein $\mathcal{O}(h^4)$-Verfahren ohne diese Nachteile konstruiert werden. 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. Daher ersetzt man den h^2-Korrekturterm $(S_{h/2} - S_{-h/2})^4 u(x) \cong h^4 u^{(4)}(x_i)$ in (1.7.7) durch $h^2(S_{h/2} - S_{-h/2})^2 u''(x_i)$. Dieses Mehrstellenverfahren führt nun auf das Gleichungssystem

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

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

2 Gemeinsame Prinzipien von Diskretisierungsverfahren

2.1 Konvergenz, Konsistenz, Stabilität

Den Fehlerbetrachtungen bei allen bisher behandelten Verfahren liegen gemeinsame Prinzipien zugrunde, die hier noch einmal herausgearbeitet werden sollen. Alle behandelten Probleme können als Gleichungssysteme geschrieben werden, wobei die jeweilige Lösung u implizit

$$F(u) = 0,\ u \in D, \qquad \text{mit } F : D \subset V \to W \tag{2.1.1}$$

definiert ist mit Hilfe einer Abbildung F zwischen normierten Räumen V, W. Der Definitionsbereich D von F ist dabei i. a. eine echte Untermenge von V. Bei der Lösung von Differentialgleichungen ist V ein Raum von stetigen, differenzierbaren Funktionen, z.B., $(C^1[a,b])^n$. Die Abbildung F besteht hauptsächlich aus einem Ausdruck, der Ableitungen der eingesetzten Funktion enthält, also einem Differentialoperator. Die zusätzlichen Anfangs- bzw. Randbedingungen können entweder in die Definition von F oder in die von V bzw. D eingebaut werden. Bei den Systemen von Dgln erster Ordnung läßt sich

$$\begin{aligned} V &= (C^1[a,b])^n \quad \text{mit Norm} \quad \|v\|_V := \max\{\|v(x)\| :\ x \in [a,b]\}, \\ W &= \mathbf{R}^n \times (C[a,b])^n \quad \text{mit Norm} \quad \left\| \begin{pmatrix} d \\ g \end{pmatrix} \right\|_W := \|d\| + \max\{\|g(x)\| :\ x \in [a,b]\} \end{aligned}$$

$D = V$, wählen. Dabei sei die nicht-indizierte Norm eine beliebige des $\mathbf{R}^n$. Man beachte, daß der so gewählte Raum V nicht vollständig ist, da die Norm keine Ableitungen einschließt. Als Abbildung ergibt sich beim

$$\underline{\text{AWP}}\quad F : v \mapsto \begin{pmatrix} v(a) - u_0 \\ v' - f(.,v) \end{pmatrix}, \qquad \underline{\text{RWP}}\quad F : v \mapsto \begin{pmatrix} r(v(a), v(b)) \\ v' - f(.,v) \end{pmatrix}. \tag{2.1.2}$$

Beim Randwertproblem (1.7.3) zweiter Ordnung können homogene Randbedingungen, $\alpha = \beta = 0$, besser in die Definition des Raums V eingebaut werden. Hier bieten sich an

$$V := \{v \in C^2[a,b] : v(a) = v(b) = 0\}, \qquad W := C[a,b], \tag{2.1.3}$$

$$F :\ v \mapsto -v'' + pv' + qv - g. \tag{2.1.4}$$

In V und W kann jeweils die Maximumnorm von v verwendet werden.

Zur numerischen Approximation solcher Probleme wird das Originalproblem durch eine *Folge von effektiv lösbaren endlichdimensionalen (diskreten) Problemen* ersetzt. Das Ziel bei der Konstruktion solcher Verfahren ist, bei wachsender Dimension im diskreten Problem den Fehler zwischen diskreter und exakter Lösung immer kleiner zu machen. In den bisher behandelten Beispielen entsprachen die endlichdimensionalen Räume jeweils einer Sammlung von Funktionswerten auf einem Gitter $\Delta \subset [a,b]$, wobei die Raumdimension bei feiner werdendem Gitter anwächst. Stattdessen kommen aber auch etwa Räume von Polynomen oder Splinefunktionen wachsenden Grads in Frage.

Ein Diskretisierungsverfahren besteht also aus einer Schar von endlichdimensionalen Räumen V_Δ, W_Δ, $\dim V_\Delta = \dim W_\Delta$, und Abbildungen $F_\Delta : V_\Delta \to W_\Delta$, mit denen diskrete Lösungen durch

$$F_\Delta(u_\Delta) = 0 \tag{2.1.5}$$

definiert werden. Um diese Näherungen mit der Originallösung u vergleichen zu können, wird mindestens ein Transfer von u nach V_Δ (Restriktion) oder von u_Δ nach V (Prolongation) benötigt. Bisher trat nur der erste Fall auf. Es sei also eine Restriktion bekannt, d.h., eine (lineare) Abbildung

$$R_\Delta : V \to V_\Delta, \text{ womit } \quad e_\Delta := u_\Delta - R_\Delta u \tag{2.1.6}$$

als Fehler der diskreten Lösung definiert wird. Normen in V_Δ und W_Δ werden beide mit $\|.\|_\Delta$ bezeichnet. Diese Normen sollten in Beziehung zu den Normen der Räume V, W stehen. Dazu wird angenommen, daß ein Maß $|\Delta|$ für die *Feinheit* der Gitter Δ existiert mit den Eigenschaften, daß für jedes $v \in V$ gilt

$$|\Delta| \to 0 \quad \Rightarrow \quad \dim V_\Delta \to \infty, \ \|R_\Delta v\|_\Delta \to \|v\|_V. \tag{2.1.7}$$

Für die einfachsten Gitter, die äquidistanten mit Schrittweite $h = (b-a)/N$, ist $|\Delta| := h$ ein solches Maß. Hier liegt sogar eine abzählbare Folge von Gittern vor, die durch die Zahl der Gitterpunkte indiziert werden kann. Die Beziehungen können durch folgendes Diagramm zusammengefaßt werden.

$$\begin{array}{rccc} & V & \xrightarrow{F} & W \\ R_\Delta & \downarrow & & \\ & V_\Delta & \xrightarrow{F_\Delta} & W_\Delta \end{array}$$

Beispiel 2.1.1 Einschrittverfahren bei Anfangs- und Randwertproblemen. Mit $\Delta = \{x_i := a + ih,\ i = 0,..,N\}$, $h = (b-a)/N$, $N \in \mathbf{N}$, $|\Delta| := h$, sei

$$V_\Delta = \left\{ y_\Delta = \begin{pmatrix} y_0 \\ \vdots \\ y_N \end{pmatrix} : y_i \in \mathbf{R}^n, i = 0,..,N \right\},\ \|y_\Delta\|_\Delta := \max_{i=0}^{N} \|y_i\|,\ R_\Delta v := \begin{pmatrix} v(x_0) \\ \vdots \\ v(x_N) \end{pmatrix},$$

$v \in V$, wobei $\dim V_\Delta = n(N+1)$. Die Restriktion $R_\Delta v$ besteht also aus den Gitterwerten von v. Der Bildraum W_Δ kann mit V_Δ identifiziert werden. Mit einem durch seine Verfahrensfunktion f_h definierten Einschrittverfahren erhält man die diskrete Abbildung

$$F_\Delta(y_\Delta) = \begin{pmatrix} y_0 - u_0 \\ \frac{1}{h}(y_1 - y_0) - f_h(x_0, y_0) \\ \vdots \\ \frac{1}{h}(y_N - y_{N-1}) - f_h(x_{N-1}, y_{N-1}) \end{pmatrix} \tag{2.1.8}$$

beim Anfangswertproblem. Beim Randwertproblem ist nur die erste Komponente, $y_0 - u_0$, von F_Δ durch $r(y_0, y_N)$ zu ersetzen. Beim AWP kann man das Gleichungssystem $F_\Delta(u_\Delta) = 0$

natürlich schrittweise nach $y_0, y_1, \ldots$ auflösen. Beim RWP entspricht dieses Gleichungssystem dem der Mehrzielmethode (1.6.10) und kann im Rahmen der Newtoniteration auf ein $n \times n$-System kondensiert werden.

Beispiel 2.1.2 Differenzenverfahren beim linearen RWP zweiter Ordnung mit homogenen Randbedingungen. Da die Randwerte durch die Randbedingungen bekannt sind, müssen sie nicht mehr als Unbekannte geführt werden. Mit dem Gitter aus Beispiel 2.1 wählt man hier

$$V_\Delta = \left\{ y_\Delta = \begin{pmatrix} y_1 \\ \vdots \\ y_{N-1} \end{pmatrix} : y_i \in \mathbf{R}, i = 1, .., N-1 \right\},\ \|y_\Delta\|_\Delta := \max_{i=1}^{N-1} |y_i|,\ R_\Delta v := \begin{pmatrix} v(x_1) \\ \vdots \\ v(x_{N-1}) \end{pmatrix},$$

$v \in V$, mit $\dim V_\Delta = N-1$. Das diskrete Problem hat damit die in §1.7 behandelte Form $F_\Delta(u_\Delta) = Au_\Delta - \tilde{g}$, mit der in (1.7.5) definierten Matrix A (hier mit $h = (b-a)/N$, Dimensionen $(N-1) \times (N-1)$).

Diese Bezeichnungen wurden vor allem eingeführt, um noch einmal die gemeinsame Struktur der alten Beweise hervorzuheben, insbesondere in Bezug auf das Zusammenspiel der Begriffe Konsistenz, Stabilität und Konvergenz, und deren Abhängigkeit von der geeigneten Wahl der Definitionen von F_Δ und $\|.\|_\Delta$. Im Beispiel 2.1 hätte man alle Komponenten von F_Δ mit h multiplizieren können, ohne das Verfahren zu ändern. In diesem Fall wäre das Verfahren zwar noch konsistent, aber nicht mehr stabil im Sinne der folgenden Definition.

Definition 2.1.3 *Ein Diskretisierungsverfahren heißt* konsistent *(von der Ordnung $p > 0$), wenn mit der Lösung $u \in V$ für den Konsistenzfehler*

$$T_\Delta := F_\Delta(R_\Delta u) \quad \textit{gilt} \quad \|T_\Delta\|_\Delta \to 0,\ \textit{bzw.}\ \|T_\Delta\|_\Delta = \mathcal{O}(|\Delta|^p), |\Delta| \to 0.$$

Das Verfahren heißt stabil, *wenn Konstanten r, S, H existieren so, daß für alle Gitter mit $|\Delta| \le H$ für je zwei Elemente $v_\Delta^{[j]} \in V_\Delta$ mit $F_\Delta(v_\Delta^{[j]}) = g_\Delta^{[j]}$ und $\|g_\Delta^{[j]}\|_\Delta \le r,\ j = 1, 2$, folgt*

$$\|v_\Delta^{[1]} - v_\Delta^{[2]}\|_\Delta \le S\|g_\Delta^{[1]} - g_\Delta^{[2]}\|_\Delta = S\|F_\Delta(v_\Delta^{[1]}) - F_\Delta(v_\Delta^{[2]})\|_\Delta. \tag{2.1.9}$$

Das wesentliche Merkmal der Stabilitätsforderung ist, daß die Schranke gleichmäßig gilt für die gesamte (unendliche) Folge von Gittern, die Stabilitätskonstante S also nicht von Δ abhängt. Daher bedeutet Stabilität insbesondere auch die gleichmäßig beschränkte Invertierbarkeit der Ableitungen F'_Δ für die ganze Folge der Gitter! Die Stabilitätsschwelle r beschränkt die Forderung auf eine nicht zu kleine Umgebung der diskreten Lösung, die der rechten Seite $g_\Delta = 0$ entspricht.

Diese Definition stimmt bis auf kleinere Modifikationen mit denen von früher überein, beim AWP war das Def. 1.2.3 und Def. 1.2.5, sowie Satz 1.2.7. Beim RWP zweiter Ordnung wurden die entsprechenden Aussagen direkt nachgewiesen, Konsistenz mit $p = 2$ in Satz 1.7.3), Stabilität mit $S = 1/q^*$ in Satz 1.7.2. Wie in der obigen Bemerkung gilt die Normschranke dieses Satzes gleichmäßig für die Inversen einer unendliche Folge von Matrizen wachsender Dimension $n \to \infty$.

Die Konvergenz eines Verfahrens folgt nun wieder durch Kombination der beiden Eigenschaften Konsistenz und Stabilität.

Satz 2.1.4 *Das diskrete Problem (2.1.5) sei stabil nach (2.1.9) und konsistent von der Ordnung p mit $\|T_\Delta\|_\Delta \le C_T|\Delta|^p$. Dann konvergieren die Näherungslösungen mit Ordnung p, d.h., für alle Gitter mit $|\Delta| \le H_0 := \min\{H, (r/C_T)^{1/p}\}$, gilt*

$$\|u_\Delta - R_\Delta u\|_\Delta \le SC_T|\Delta|^p.$$

Beweis In (2.1.9) wird gewählt $v_\Delta^{[1]} := u_\Delta$ mit $g_\Delta^{[1]} = 0$ und $v_\Delta^{[2]} := R_\Delta u$ mit $g_\Delta^{[2]} = T_\Delta$. Nach Voraussetzung ist $|\Delta| \le H$, $\|g_\Delta^{[j]}\|_\Delta \le r$, $j = 1, 2$, und die Behauptung folgt aus (2.1.9). ■

2.2 Allgemeine Verfahrens-Ansätze

In diesem allgemeinen Rahmen sollen noch kurz zwei Prinzipien skizziert werden, die jeweils einen allgemeinen Ansatz zur Konstruktion von Verfahren für unterschiedliche Problembereiche darstellen.

Projektions-Verfahren Näherungen für eine isolierte Lösung von $F(u) = 0, u \in V$, werden hier in einem endlichdimensionalen Unterraum $V_\Delta \subseteq V$ bestimmt, z.B., einem Raum von Polynomen oder Spline-Funktionen (vgl. Numerik I, §4). Für $v_\Delta \in V_\Delta$ kann dann $F(v_\Delta) \in W$ gebildet werden, allerdings ist das Gleichungssystem $F(u_\Delta) = 0$ in der Regel unlösbar. Hier wird nun ein Projektor $P_\Delta = P_\Delta^2$ eingeführt, und die Näherung $u_\Delta \in V_\Delta$ definiert durch

$$P_\Delta F(u_\Delta) = 0, \quad P_\Delta : W \to W_\Delta. \tag{2.2.1}$$

Bei geeigneter Wahl von P_Δ und $\dim W_\Delta = \dim V_\Delta$ ist dieses System (lokal) eindeutig lösbar, Stabilitätseigenschaften der Abbildung $P_\Delta F : V_\Delta \to W_\Delta$ müssen für das konkrete Verfahren untersucht werden. Aufgrund der speziellen Struktur läßt sich aber die Konvergenz nach einfachen Prinzipien analysieren. Im linearen Fall, $F(u) = Lu - r$, kann dazu das Hilfsproblem $Lv = LR_\Delta u$ betrachtet werden. Es hat die Eigenschaft, daß exakte und numerische Lösung übereinstimmen, denn $v = v_\Delta = R_\Delta u$. Durch Subtraktion der Beziehungen $P_\Delta L u_\Delta = P_\Delta r = P_\Delta L u$ und $P_\Delta(Lv_\Delta - LR_\Delta u) = 0$ folgt die Gleichung $P_\Delta L(u_\Delta - v_\Delta) = P_\Delta L(I - R_\Delta)u$. Damit gilt

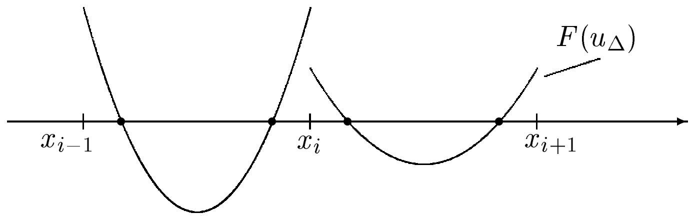
$$\|u - u_\Delta\| = \|u - v - (u_\Delta - v_\Delta)\| \le (1 + K)\|(I - R_\Delta)u\|.$$

Der globale Fehler kann also bis auf die Konstante $1 + K$ durch den Approximationsfehler $(I - R_\Delta)u$ bei der Lösung u abgeschätzt werden. Allerdings muß dazu wieder im konkreten Fall die Konstante K als Norm der Abbildung $(P_\Delta L|_{V_\Delta})^{-1} P_\Delta L : V \to V_\Delta$ ($\ne$ Id!) bestimmt werden.

Spezielle Projektionsverfahren sind die in §3.3 behandelten Finite-Element-Verfahren und die *Kollokationsverfahren*. Ihr Name kommt von der Interpretation, daß man (bei Funktionen $u = u(x), x \in \Omega$) für $u_\Delta(x)$ die Gleichung $F(u_\Delta(x)) = 0$ nicht mehr überall fordert, $\forall x \in \Omega$, sondern nur noch in endlich vielen Punkten $x \in \bar{\Delta} \subseteq \Omega$. Äquivalent dazu ist, daß man P_Δ als Interpolationsabbildung mit den $x \in \bar{\Delta}$ als Stützstellen definiert und (2.2.1) fordert. Als Beispiel werde eine Dgl k-ter Ordnung betrachtet, etwa (1.7.3) bei $k = 2$. Dazu wählt man ein Grundgitter $\Delta : a = x_0 < x_1 < \ldots < x_N = b$ und als Ansatzraum einen Raum von Spline-Funktionen

$$V_\Delta := S^{k-1}_{\Delta,p+k} = \{g \in C^{k-1}[a,b] : g|_{[x_i,x_{i+1})} \in \Pi_{p+k}\},$$

die stückweise Polynome vom Grad $p+k$ sind. Dabei ist nur $g \in C^{m-1}$ wirklich erforderlich, vgl. Satz von Picard-Lindelöf ($T : C^0 \to C^0$). Als Kollokationspunkte werden in jedem Teilintervall $[x_i, x_{i+1})$ jeweils p affin ähnliche Punkte verwendet $\bar{\Delta} := \{x_i + h_i\xi_j : j = 1, \ldots, p,\ i = 0, \ldots, N-1\}$. P_Δ entspricht der Polynom-Interpolation unabhängig in allen Teilintervallen, Funktionen $P_\Delta v$ sind also nur stückweise stetig. In der Skizze sind für $p = 2$ die Kollokationspunkte $\bar{\Delta}$ durch Kreise markiert, der Defekt $F(u_\Delta)$ verschwindet nur dort.



Das Verfahren soll nicht weiter diskutiert werden, es sei jedoch erwähnt, daß bei Wahl der Kollokationsstellen $\xi_1, \ldots, \xi_p \in (0,1)$ wie bei der Gauß-Quadratur (Numerik I, §5.3) sich eine *Superkonvergenz* auf den Gitterpunkten x_i einstellt, d.h., eine Konvergenzordnung, die über der Konsistenzordnung liegt. Dies war auch der Hintergrund für die hohe Ordnung in Satz 1.5.4, die Gauß-Runge-Kutta-Verfahren sind nämlich äquivalent zu Kollokationsverfahren in S^0_m.

Defektkorrektur-Verfahren Bei Verfahren höherer Ordnung reicht die Kopplung zwischen Unbekannten (Lösungs-Werten, -Koeffizienten) in der Regel weiter als bei solchen niederer Ordnung. Daher ist die Struktur der (bei RWPen) auftretenden Gleichungssysteme komplexer und die Auflösung aufwendiger. Bei den Defektkorrekturverfahren wird dies umgangen dadurch, daß ausgehend von einer Lösung niederer Ordnung schrittweise Näherungen wachsender Ordnung berechnet werden. Wegen der damit auch verfügbaren Fehlerschätzung bieten diese Verfahren als Bonus die Möglichkeit zu Gitter- und Ordnungssteuerung. Als Ansatzpunkt kann eine Newton-artige Iteration

$$L_\Delta(u_\Delta^{[k+1]} - u_\Delta^{[k]}) = -G_\Delta(u_\Delta^{[k]})$$

für eine nichtlineare Gleichung $G_\Delta(u_\Delta) = 0$ dienen. L_Δ ist dabei eine Approximation an G'_Δ, da für die (lokale) Konvergenz die Bedingung $\|I - L_\Delta^{-1}G'_\Delta(u_\Delta)\| = \|L_\Delta^{-1}(L_\Delta - G'_\Delta(u_\Delta))\| < 1$ erforderlich ist. Diese bietet aber auch den Spielraum merkliche Abweichungen $L_\Delta - G'_\Delta$ zuzulassen, wenn L_Δ dafür eine einfachere Struktur hat als G'_Δ. Z.B., kann L_Δ zu einem Verfahren niederer Ordnung gehören als G_Δ. Daher werde jetzt konkret folgende Iteration betrachtet,

$$F'_\Delta(u_\Delta^{[0]})(u_\Delta^{[k+1]} - u_\Delta^{[k]}) = -\hat{F}_\Delta(u_\Delta^{[k]}), \quad k = 0, 1, \ldots. \tag{2.2.2}$$

Dabei soll $\hat{F}_\Delta$ ein Verfahren höherer Ordnung realisieren, während F'_Δ die Ableitung zu einem Verfahren niedriger Ordnung darstellt. Als Startwert kann die Lösung $u^{[0]}$ dieses *Basisverfahrens* dienen, definiert durch $F_\Delta(u^{[0]}) = 0$. Dann kann in allen Schritten (2.2.2) die LR-Zerlegung von $F'_\Delta(u^{[0]}_\Delta)$ wiederverwendet werden.

Zur Illustration werden Differenzenverfahren beim linearen RWP (1.7.3) mit $p \equiv 0$ betrachtet. Zum Basisverfahren F_Δ der Ordnung 2 gehört die Tridiagonalmatrix (1.7.5). Die Abbildung $\hat{F}_\Delta$ zum Verfahren 4-ter Ordnung mit (1.7.7) ist für $i = 2, \ldots, N-2$

$$(\hat{F}_\Delta u_\Delta)(x_i) = q(x_i)y_i + \frac{1}{h^2}(-y_{i-1} + 2y_i - y_{i+1}) + \frac{1}{12h^2}(y_{i-2} - 4y_{i-1} + 6y_i - 4y_{i+1} + y_{i+2}) - g(x_i),$$

$(y_i := u_\Delta(x_i))$, mit Modifikationen in den randnahen Punkten x_1, x_{N-1}. Die Ableitungsmatrix $\hat{F}'$ hat ≥ 5 Diagonalen. Es sei $\hat{u}_\Delta$ die genauere Lösung mit $\hat{F}(\hat{u}_\Delta) = 0$. Betrachtet man nun den ersten Schritt der Iteration (2.2.2), ergibt sich für die Fehler dazu die Beziehung

$$F'_\Delta \cdot (u^{[1]}_\Delta - \hat{u}_\Delta) = (F'_\Delta - \hat{F}'_\Delta) \cdot (u^{[0]}_\Delta - \hat{u}_\Delta), \tag{2.2.3}$$

da $\hat{F}_\Delta(u^{[0]}_\Delta) = F_\Delta(u^{[0]}_\Delta) - \hat{F}_\Delta(\hat{u}_\Delta) = \hat{F}'_\Delta \cdot (u^{[0]}_\Delta - \hat{u}_\Delta)$. Wegen der Linearität aller Abbildungen ist bei diesen kein Argument angegeben. Die Differenz $(F'_\Delta - \hat{F}'_\Delta)v$ besteht jetzt nur noch aus dem Differenzenausdruck 4-ter Ordnung

$$\frac{1}{12h^2}(v_{i-2} - 4v_{i-1} + 6v_i - 4v_{i+1} + v_{i+2}) = 2h^2 v[x_{i-2}, \ldots, x_{i+2}].$$

In (2.2.3) ist $v = u^{[0]}_\Delta - \hat{u}_\Delta = O(h^2)$. Für genügend oft differenzierbare Koeffizienten der Dgl kann man (mit einigem Aufwand) nachweisen, daß dann auch für die dividierte Differenz $v[x_{i-2}, \ldots, x_{i+2}] = O(h^2)$ gilt, sodaß insgesamt die rechte Seite von (2.2.3) die Größenordnung $O(h^4)$ besitzt, d.h., schon ein Defektkorrektur-Schritt erzeugt eine h^4-Näherung $u^{[1]}_\Delta$.

Wesentlich an dem Verfahren ist, daß der *Defekt* $\hat{F}_\Delta(u^{[k]}_\Delta)$ auf der rechten Seite der Iteration (2.2.2) mit einem Verfahren höherer Ordnung gebildet wird. Wenn allgemein $\hat{F}_\Delta$ Konsistenz-Ordnung p hat und F_Δ die Ordnung 2, ergibt sich bei passender Konstruktion der Fehler $u^{[k]}_\Delta - R_\Delta u = O(h^{\min\{2k+2,p\}})$.

3 Partielle Differentialgleichungen

3.1 Allgemeine Eigenschaften

Im Unterschied zu den gewöhnlichen Differentialgleichungen werden jetzt Funktionen $u(x_1,..,x_n)$ von mehr als einer Veränderlichen betrachtet. Da partielle Dgln ein sehr umfangreiches Problemgebiet darstellen, das hier nur ansatzweise diskutiert werden kann, werden hier nur lineare Probleme in zwei (Raum-) Dimensionen diskutiert für Funktionen $u(x,y)$. Bei partiellen Differentialgleichungen treten Ableitungen der gesuchten Funktion nach mehr als einer Veränderlichen auf. Daher können sie auch i.a. nicht auf eine Standardform wie (1.1.1) gebracht werden. Geht man von Systemen erster Ordnung für Funktionen $u(x,y) := (u_1(x,y),..,u_n(x,y))^\mathsf{T}$ aus, dann ist ein recht allgemeine Form

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

Charakteristische Eigenschaften dieser Dgl, nach denen sich auch die Wahl der zusätzlichen Randbedingungen zu richten hat, werden vor allem durch die $n \times n$-Koeffizientenmatrix A bestimmt. Zwei Sonderfälle erhalten eine Bezeichnung.

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 *im Punkt* (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 in Dimension $n = 2$ für $u = (v,w)^\mathsf{T}$ sind folgende

Beispiel 3.1.2 Die Cauchy-Riemann-Dgln

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

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

Beispiel 3.1.3 Aus dem hyperbolischen System (EWe ± 1)

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

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

Beispiel 3.1.4 Das System

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

führt auf die *Wärmeleitungsgleichungen* $v_{xx} - v_y = 0$ und $w_{xx} - w_y = 0$.

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 Variabeln* x, y

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

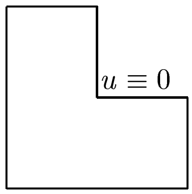
heißt im Punkt (x, y)

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 Rang* $\begin{pmatrix} a & b & e \\ b & d & f \end{pmatrix} = 2$.

Abhängig vom Typ sind nur bestimmte Randbedingungen sinnvoll. Bei einer (überall) elliptischen Dgln, die auf einem (evtl. unbeschränkten) Gebiet $\Omega \subseteq \mathbf{R}^2$ 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 davon.

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

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



definiert. Dann erfüllt u die Potentialgleichung $u_{xx} + u_{yy} = 0$ auf Ω und besitzt glatte Randwerte, z.B. $u(x,y) = 0$ für $x = 0, y \ge 0$ sowie $x \ge 0, y = 0$. Die ersten Ableitungen von u sind jedoch nicht beschränkt in der einspringenden Ecke im Nullpunkt $r \to 0$. In den übrigen Ecken des L treten etwas schwächere Singularitäten auf.

Diese Eigenschaft weicht von der Situation bei den gewöhnlichen Dgln ab und führt auf eine neue Rahmenbedingung bei der Diskussion numerischer Verfahren. Wenn eine gewöhnliche Dgl genügend oft differenzierbare Koeffizienten besitzt, kann durch Ableitung der Dgl auf eine entsprechende Differenzierbarkeit der Lösung geschlossen werden. Dies machte den Einsatz von Verfahren hoher Ordnung erst sinnvoll. Bei den partiellen Dgln dagegen kann die Lösung schon bei ganz einfachen Problemen Singularitäten (von Ableitungen) besitzen. Aus diesem und auch anderen, praktischen Gründen macht bei partiellen Dgln der Einsatz numerischer Verfahren hoher Ordnung Schwierigkeiten. Daher 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 $\Omega \subseteq \mathbf{R}^2$ 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*

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

Diese Gleichung ist elliptisch, sie stellt sogar die Normalform einer elliptischen Dgl dar. Physikalisch beschreibt sie, u.a., die Wärmeverteilung in einem homogenen Körper mit Wärmequellen g. Die homogene Gleichung ($g \equiv 0$) heißt *Potentialgleichung*, ihre Lösungen harmonische Funktionen ($\rightarrow$ Funktionentheorie). Zur Definition einer 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.

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

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

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

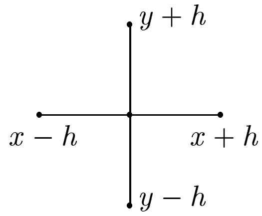
In der obigen physikalischen Interpretation entspricht die Dirichlet-Randbedingung einer Temperaturvorgabe auf Γ_d, die Neumann-RB (mit $\phi_n \equiv 0$) einer isolierten Wand Γ_n, während die Cauchy-Bedingung sich als Wärmeabstrahlung durch Γ_c interpretieren läßt. In der Praxis kann natürlich auch jeweils nur eine der Randbedingungen auftreten. In diesen Fällen redet man von dem 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.*

Während beim Dirichletproblem eine eindeutige Lösung existiert, besitzt das Neumannproblem Lösungen nur unter der Voraussetzung $\int_\Gamma \phi_n \, ds = \int \int_\Omega g(x, y) \, dxdy$. Die Lösung ist dann auch nur bis auf eine Konstante eindeutig.

Wie im eindimensionalen Fall (§1.7) lassen sich Näherungsverfahren gewinnen, indem man die Ableitungen in der Dgl durch Differenzenquotienten ersetzt. Mit Schrittweiten $h_x, h_y > 0$ gilt für $u \in C^4$ nach Lemma 1.7.1

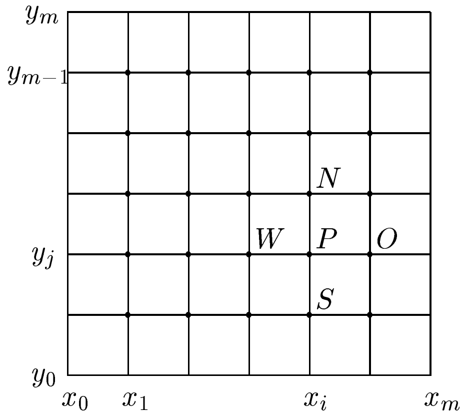
$$u_{xx}(x,y) = \frac{1}{h_x^2}[u(x-h_x,y) - 2u(x,y) + u(x+h_x,y)] + \mathcal{O}(h_x^2)$$
$$u_{yy}(x,y) = \frac{1}{h_y^2}[u(x,y-h_y) - 2u(x,y) + u(x,y+h_y)] + \mathcal{O}(h_y^2)$$



Bei Addition dieser beiden Näherungen werden zur Approximation der Poissongleichung in (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 $h := 1/m,\ m \in \mathbf{N}$, das Gitter

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



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 speziellen Punktes $P = (x_i,y_j)$ ist es jedoch einprägsamer, Himmelsrichtungen zu zu verwenden. Also gelte, zu

$$v_P = v_{i,j} \ \text{ sei } \ v_N := v_{i,j+1},\ v_O := v_{i+1,j},\ v_S := v_{i,j-1},\ v_W := v_{i-1,j}, \tag{3.2.6}$$

wie in der obigen Skizze angedeutet. Die Poissongleichung im Punkt P, $-u_{xx}(P) - u_{yy}(P) = g(P)$ wird nun 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.,

$$\begin{array}{c|c} & -1 \\ \hline -1 & 4 \quad -1 \\ & -1 \end{array} \cdot \frac{1}{h^2} \qquad\qquad \frac{1}{h^2}[4v_P - v_N - v_O - v_S - v_W] = g_P. \tag{3.2.7}$$

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

$$v_P = \frac{1}{4}[v_N + v_O + v_S + v_W]$$

ausgedrückt werden, die aussagt, daß v_P der Mittelwert der Funktionswerte in den vier Nachbarpunkten ist (Interpretation: 'aufgespanntes Gumminetz').

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

$$\Omega_h := \{(x_i, y_j) :\ i, j = 1, .., m-1\}, \quad \Gamma_h := \{(x_i, y_j) :\ i = 0 \vee i = m \vee j = 0 \vee j = m\}, \tag{3.2.8}$$

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

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

Wie erwähnt, werden die trivialen Randgleichungen dazu benutzt, die entsprechenden Unbekannten aus den Gleichungen auf Ω_h zu eliminieren (vgl. auch (1.7.6)). Dann liegt ein lineares Gleichungssystem aus $(m-1)^2$ Gleichungen für ebenso viele Unbekannte vor. Bei den partiellen Dgln ist es angebracht, in der Formulierung wieder stärker zwischen der mit dem Gleichungssystem zusammenhängenden linearen Abbildung L_Δ von Gitterfunktionen und den zugehörigen Matrizen zu unterscheiden. Wird zu Gitterfunktionen $u_\Delta : \Omega_h \to \mathbf{R}$ die lineare Abbildung L_Δ punktweise definiert durch

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

mit Modifikationen in den randnahen Punkten, entspricht das Gleichungssystem (3.2.9) einem Problem der Form

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

mit der schon verwendeten Abkürzung $v_{ij} = u_\Delta(x_i, y_j)$. 'Die' Matrix dieses linearen Systems hängt zumindest von einer Numerierung der bisher doppelt indizierten Unbekannten ab.

Bei *zeilenweiser Numerierung* (lexikographischer) des Gitters erhalten die drei Unbekannten v_W, v_P, v_O aufeinanderfolgende Nummern bzw. 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 Numerierung 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 sind die Matrizen I und T jeweils von der Größe $(m-1)\times(m-1)$, die Gesamtmatrix A von der Größe $(m-1)^2\times(m-1)^2$.

Stellt man sich andererseits vor, daß die Gitterpunkte wie ein *Schachbrett* schwarz-weiß eingefärbt sind, erkennt man, daß in (3.2.9) die Nachbarpunkte N, O, S, W immer die zum Punkt P entgegengesetzte Farbe besitzen. Wenn daher zunächst weiße und dann schwarze Punkte numeriert werden, ergibt sich im Großen eine Blockstruktur der Form

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

wobei M in jeder Zeile höchstens vier Einsen und sonst Nullen enthält. Wenn $m-1$ ungerade ist, 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 IIA.

Neumann-Randbedingung: Hier ist der Wert der Normalenableitung vorgeschrieben. Daher muß auch bei der Randbedingung eine Approximation durch Differenzenquotienten vorgenommen werden. 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 also 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 dem 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

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

   Nachteilig ist hierbei die schlechtere Approximation der Randbedingung mit einem $\mathcal{O}(h)$-Fehler wegen der Verwendung unsymmetrischer Differenzenquotienten.

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
$$\begin{array}{cc|c} & -1 & \\ -2 & & 4 \cdot \frac{1}{h^2} \\ & -1 & \end{array} \qquad\qquad \frac{1}{h^2}(4v_P - 2v_W - v_N - v_S) = g_p + \frac{2}{h}\phi_n(P).$$
Diese Methode ist besonders naheliegend, da (homogene) Neumann-Randbedingungen oft aufgrund von Symmetrieüberlegungen beim Dirchletproblem auftreten. Zum Beispiel gilt für die Lösung des Problems
$$-u_{xx} - u_{yy} = 1 \text{ auf } (-1,1)\times(-1,1),\ u=0 \text{ auf dem Rand,}$$
die Aussage $u(x,y)=u(\pm x,\pm y)$, also $u_x(0,y)\equiv 0, u_y(x,0)\equiv 0$. Daher kann dieses Problem auf ein Viertel des uspründlichen Quadrats eingeschränkt werden, etwa das Einheitsquadrat, wobei auf dessen linken und unteren Rand die Neumann-Bedingung (3.2.3) mit $\phi_n\equiv 0$ gilt. Wird diese mit Hilfe von (3.2.11) diskretisiert, ergibt sich ebenfalls 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) erhalten würde. Die Symmetriebetrachtung reduziert dabei die Größe des Gleichungssystems auf ein Viertel.

Die Struktur der Matrix zu diesem reduzierten Problem
$$\begin{aligned} -u_{xx} - u_{yy} &= g \quad \text{in} \quad \Omega = (0,1)\times(0,1), && (3.2.12)\\ u &= 0 \quad \text{in} \quad \Gamma_d = \{(x,y)\in\bar\Omega:\ x=1 \vee y=1\}, && (3.2.13)\\ \frac{\partial}{\partial \mathbf{n}} u &= 0 \quad \text{in} \quad \Gamma_n = \{(x,y):\ x=0, y\in[0,1) \vee y=0, x\in[0,1)\} && (3.2.14)\end{aligned}$$
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 (Vorfaktor $1/h^2$):
$$\begin{array}{c|c} -\frac{1}{2} & \\ 2 & -1 \\ -\frac{1}{2} & \end{array} \qquad \begin{array}{ccc} & -1 & \\ -\frac{1}{2} & 2 & -\frac{1}{2} \end{array} \qquad \begin{array}{c|c} -\frac{1}{2} & \\ 1 & -\frac{1}{2} \end{array}$$

Die Matrix ist auch etwas größer, da jetzt unbekannte Werte auf dem größeren Berechnungsgitter $\Omega_h := \{(x_i, y_j) : \ i, j = 0, .., m-1\}$ auftreten. Bei zeilenweiser Numerierung dieser Unbekannten $v_{ij} \cong u(x_i, y_j)$, in der Reihenfolge $v_{00}, .., v_{m-1,0}, v_{0,1}, .., v_{m-1,1}, \dots$, hat die Systemmatrix zu (3.2.12) die Gestalt

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

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

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 Eigenschaften der Differenzen-Matrizen

Zwei Matrix-Eigenschaften und damit verbundene Beweistechniken spielen bei der Behandlung elliptischer Probleme eine wichtige Rolle. Die erste Eigenschaft ist Symmetrie und Definitheit.

Definition 3.2.2 *Eine Matrix $A = (a_{ij}) \in \mathbf{R}^{n \times n}$ heißt symmetrisch, wenn $A = A^\mathsf{T}$ gilt, dh. $a_{ij} = a_{ji} \forall i, j = 1, .., n$. Sie heißt positiv definit bzw. semi-definit, wenn sie symmetrisch ist und ein $\alpha > 0$ bzw. $\alpha \geq 0$ existiert so, daß*

$$v^\mathsf{T} A v \geq \alpha\, v^\mathsf{T} v \quad \forall v \in \mathbf{R}^n. \tag{3.2.16}$$

Bei einer symmetrisch, positiv definiten ('SPD-') Matrix sind alle Eigenwerte reell positiv, insbesondere ist (3.2.16) mit dem kleinsten Eigenwert, $\alpha = \min \lambda(A)$, erfüllt. Hier ist die Verwendung der Euklidnorm und der dazu gehörigen Matrixnorm,

$$\|v\|_2 = \sqrt{v^\mathsf{T} v}, \ v \in \mathbf{R}^n, \quad \|A\|_2 := \sqrt{\varrho(A^\mathsf{T} A)},$$

der Spektralnorm angebracht. Dabei ist $\varrho(M)$ der Spektralradius der Matrix M, d. h. der Maximalbetrag ihrer Eigenwerte. Da jede symmetrische Matrix unitär ähnlich zu einer reellen Diagonalmatrix ist, gilt $\|A\|_2 = \varrho(A)$. Zur Abschätzung der Lösungen von linearen Gleichungssystemen gilt bei Definitheit die einfache Aussage,

$$A \text{ SPD } \Rightarrow A^{-1} \text{ SPD }, \quad \|A^{-1}\|_2 = \varrho(A^{-1}) \leq \frac{1}{\alpha}, \ \alpha \text{ aus (3.2.16)}.$$

Diese läßt sich auch auf allgemeinere Matrizen übertragen.

Satz 3.2.3 *Für die (i.a. nichtsymmetrischen) Matrizen $A_j \in \mathbf{R}^{n\times n}$, $j = 1,\dots,k$, gelte $v^\mathsf{T} A_j v \geq \alpha_j v^\mathsf{T} v \ \forall v \in \mathbf{R}^n$, und $\alpha_1 + \ldots + \alpha_k > 0$. Dann ist $A = A_1 + \ldots + A_k$ invertierbar und*

$$\|A^{-1}\|_2 = \|(A_1 + \ldots + A_k)^{-1}\|_2 \leq \frac{1}{\alpha_1 + \ldots + \alpha_k}.$$

Beweis Es sei $0 \neq v \in \mathbf{R}^n$ beliebig. Aus der Voraussetzung folgt durch Addition $v^\mathsf{T} A v \geq \alpha v^\mathsf{T} v$ wo $\alpha = \alpha_1 + \ldots + \alpha_k > 0$. Mit der Cauchy-Schwarz-Ungleichung führt dies auf die Aussage $\alpha\|v\|^2 \leq v^\mathsf{T} A v \leq \|v\|_2 \|Av\|_2 \Rightarrow \|Av\|_2 \geq \alpha\|v\|_2$. Für $v \neq 0$ ist insbesondere immer $Av \neq 0$. Daher ist A regulär und mit $v = A^{-1}w$ folgt $\|A^{-1}w\|_2 \leq \frac{1}{\alpha}\|w\|_2$. ■

Bei der Voraussetzung des Satzes spielt insbesondere nur der symmetrische Anteil

$$S(A) := \frac{1}{2}(A + A^\mathsf{T})$$

der Matrix A eine Rolle, da $v^\mathsf{T} A v = (v^\mathsf{T} A v)^\mathsf{T} = v^\mathsf{T} A^\mathsf{T} v = v^\mathsf{T} S(A) v$. Die entscheidende Voraussetzung des Satzes ist also die positive Definitheit des symmetrischen Anteils von A, $S(A) = S(A_1) + \ldots + S(A_k)$. Diese folgt für die Matrizen des letzten Paragraphen einfach aus der Tatsache, daß die Matrix A des Systems (3.2.9) nach Herleitung eine Summe $A = A_1 + A_2$ ist, wobei A_1, A_2 die Differenzenquotienten in x- bzw. y-Richtung enthalten. Beim diskreten Dirichlet-Problem (3.2.9) etwa wird die quadratische Form $v^\mathsf{T} A v$ betrachtet mit A aus (3.2.10),

$$\begin{aligned}
&\frac{1}{h^2}\sum_{i,j=1}^{m-1} v_{ij}\Big(4v_{ij} - v_{i-1,j} - v_{i+1,j} - v_{i,j-1} - v_{i,j+1}\Big) \\
&= \frac{1}{h^2}\sum_{j=1}^{m-1}\sum_{i=1}^{m-1} v_{ij}(-v_{i-1,j} + 2v_{ij} - v_{i+1,j}) + \frac{1}{h^2}\sum_{i=1}^{m-1}\sum_{j=1}^{m-1} v_{ij}(-v_{i,j-1} + 2v_{ij} - v_{i,j+1}).
\end{aligned}$$

Randwerte bei v sind hier null zu setzen. Zu den einzelnen Summen gilt

Lemma 3.2.4 *Gegeben seien $v_1,..,v_m \in \mathbf{R}$ mit $v_m = 0$. Dann gilt für $a \in \{1, 2\}$*

$$v_1(av_1 - v_2) + \sum_{j=2}^{m-1} v_j(-v_{j-1} + 2v_j - v_{j+1}) = (a-1)v_1^2 + \sum_{j=2}^{m-1}(v_j - v_{j-1})^2 + v_{m-1}^2 \geq C_a \sum_{j=1}^{m-1} v_j^2,$$

mit $C_2 = 4\sin^2\frac{\pi}{2m}$, $C_1 > 0$.

Beweis Links steht die quadratische Form $v^\mathsf{T} T v$ der symmetrischen Tridiagonalmatrix

$$T = \begin{pmatrix} a & -1 & & \\ -1 & 2 & -1 & \\ & \ddots & & \\ & & -1 & 2 \end{pmatrix}.$$

Die erste Identität ist elementar, die Ungleichung gilt, da Quadrate summiert werden. Für $a = 2$ ist die Konstante bekannt, da die Eigenwerte von T gleich $4\sin^2\frac{j\pi}{2m}$ sind. ■

Satz 3.2.5 *Die Matrix A (3.2.10) des Gleichungssystems (3.2.9) ist positiv definit mit $\|A^{-1}\|_2 \le C_h$, $C_h = \frac{h^2}{8}/\sin^2(h\pi/2) \le \frac{1}{16}$ für $h \le \frac{1}{2}$. Insbesondere gilt für die Lösung des diskreten Dirichletproblems (3.2.9) mit $\phi_d = 0$ die Aussage*

$$\sum_{i,j=1}^{m-1} v_{ij}^2 \le C_h^2 \sum_{i,j=1}^{m-1} g_{ij}^2.$$

Die Symmetrie und Definitheit hat auch praktische Bedeutung, da deswegen zur Lösung der linearen Gleichungssysteme effizientere Verfahren als im unsymmetrischen Fall eingesetzt werden können (Cholesky-Zerlegung, spezielle Iterationsverfahren). Ihre Bedeutung im Theoretischen ist nicht so entscheidend, da die Euklidnorm zwar hier die schärfsten Schranken liefert, diese Norm aber bei (diskreten) Funktionen weniger aussagekräftig ist als in der Maximumnorm. Für Abschätzungen in dieser Norm ist eine andere Eigenschaft entscheidend, die mit den Vorzeichen der Matrixelemente zu tun hat. Im folgenden sind Ungleichungszeichen $(\le, \ge, >, <)$ zwischen Vektoren und Matrizen sowie Beträge elementweise zu verstehen, $|v| = (|v_i|)$, $|A| = (|a_{ij}|)$.

Definition 3.2.6 *Eine Matrix $A = (a_{ij}) \in \mathbf{R}^{n\times n}$ heißt* M-Matrix, *wenn*

$$a_{ii} > 0 \ \forall i = 1,..,n, \quad a_{ij} \le 0 \ \forall i \ne j, \tag{3.2.17}$$

$$A \text{ regulär und } A^{-1} \ge 0. \tag{3.2.18}$$

Nach Zerlegung der Matrix $A = D - N$ in Diagonale D und Nebendiagonalen $-N$ lautet die Vorzeichenbedingung (3.2.17) kürzer $N \ge 0, a_{ii} > 0 \forall i$. Bei Normbetrachtungen in der Euklidnorm sind Eigenvektoren extremal. Diese Rolle wird hier von den positiven Vektoren übernommen. Wichtige Schlußweisen für nichtnegative Vektoren $v \ge 0$ bzw. Matrizen $B \ge 0$ sind $\|v\|_\infty \le a \iff v \le ae$ mit $e = (1,..,1)^\mathsf{T}$ und $\|B\|_\infty \le b \iff Be \le be$. Außerdem übertragen sich Ungleichungen wie $v \le w \Rightarrow Bv \le Bw$. Kriterien für M-Matrizen enthält

Lemma 3.2.7 *a) Eine Matrix $A = D - N$ mit der Vorzeichenverteilung (3.2.17) ist genau dann M-Matrix, wenn $\varrho(D^{-1}N) < 1$ gilt.*

b) Eine Matrix A mit der Vorzeichenverteilung (3.2.17) ist (genau dann) M-Matrix, wenn ein Vektor $z > 0$ existiert mit $w := Az > 0$. Damit gilt

$$\|A^{-1}\|_\infty \le \frac{\max_i z_i}{\min_i w_i}. \tag{3.2.19}$$

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

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

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

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

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 $Z := diag(z_i)$ definiert. Damit ist

$$Az = Dz - Nz > 0 \iff NZe < DZe \iff Z^{-1}D^{-1}NZe < e.$$

Die letzte Ungleichung für diese nichtnegative Matrix bedeutet aber gerade $\|Z^{-1}D^{-1}NZ\|_\infty < 1$, woraus insbesondere $\varrho(Z^{-1}D^{-1}NZ) = \varrho(D^{-1}N) < 1$ folgt. Nach Teil a) ist A daher M-Matrix. Mit $\alpha := \min_i w_i$ gilt laut Voraussetzung auch $Az = w \geq \alpha e$. Aus $A^{-1} \geq 0$ folgt damit $\alpha A^{-1}e \leq z \leq \|z\|_\infty e$. Dies ist äquivalent zur Behauptung (3.2.19). ∎

Für Matrizen mit (3.2.17) ist die strenge Diagonaldominanz

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

ein spezielles M-Matrix-Kriterium und mit $z = e$ in Lemma 3.2.7 in enthalten. Abschwächungen der Diagonaldominanz sind aufwendiger in der Formulierung und liefern nur die Regularitätsaussage, aber keine Schranke wie (3.2.19). Der in Lemma 3.2.7 benötigte *Testvektor* z kann bei L_Δ einfach angegeben werden. Dies führt direkt auf eine Stabilitätsaussage für die betrachteten Modellprobleme auf dem Einheitsquadrat.

Satz 3.2.8 *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 Für $z_\Delta(x,y) := r - (x-\frac{1}{2})^2 - (y-\frac{1}{2})^2$ ergibt sich sofort $L_\Delta z_\Delta(x,y) \geq 4$ mit Gleichheit in allen randfernen Punkten. Außerdem ist $0 < z_\Delta \leq r$ in Ω_h für $r = \frac{1}{2}$. Aus Lemma 3.2.7 folgt dann die Aussage. ∎

Für das Problem (3.2.12) mit Neumann-Bedingung auf zwei Seiten des Quadrats gilt eine analoge Aussage mit der größeren Konstanten 1, beim vollständigen Neumann-Problem ist die Matrix natürlich singulär.

3.2.3 Diskretisierungsfehler, Konvergenz

Da sich die Konsistenz des Differenzenverfahrens (für C^4-Lösungen) aus Lemma 1.7.1 ergibt und die Stabilität im letzten Abschnitt geklärt wurde, kann sofort auf die Konvergenz der Näherungen geschlossen werden, wenn das Randwertproblem hinreichend glatte Lösungen besitzt. Allerdings ist dies nicht immer der Fall, wie in Beispiel 3.1 gezeigt wurde. 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 R_Δ erforderlich, $L_\Delta R_\Delta u$, vgl. §2).

Satz 3.2.9 *Es sei $u \in C^k(\Omega)$, $2 \le k \le 4$. Dann gilt für $(x,y) \in \Omega_h$ die Konsistenz-Aussage*

$$|(L_\Delta u)(x,y) + u_{xx}(x,y) + u_{yy}(x,y)| \le \tag{3.2.20}$$
$$Ch^{k-2}\Big(\max\{|\frac{\partial^k}{\partial x^k}u(\xi,y)| : \ |\xi - x| < h\} + \max\{|\frac{\partial^k}{\partial y^k}u(x,\eta)| : \ |\eta - y| < h\}\Big).$$

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

Wird im Satz die Lösung u der Poissongleichung eingesetzt, ergibt sich für den lokalen 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 $u_\Delta - u$

$$L_\Delta u_\Delta = g_\Delta \Rightarrow -T_\Delta = g_\Delta - L_\Delta u = L_\Delta u_\Delta - L_\Delta u = L_\Delta(u_\Delta - u). \tag{3.2.21}$$

Auf die tatsächliche Größe dieses globalen Fehlers kann man daraus mit den Stabilitäts-Sätzen 3.2.5 (Euklidnorm) und 3.2.8 (Maximumnorm) schließen. Bei der *Quadratsummennorm* muß allerdings auf die richtige Skalierung geachtet werden. Denn bei Gittervektoren g_{ij} ist die reine Summe $\sum_{i,j} g_{ij}^2$ kein gutes Maß für die Größe von g, da sich wegen der $(m-1)^2$ Summanden bei einer konstanten Funktion der Wert $(m-1)^2 g^2$ ergibt. Daher werden nun für Gitterfunktionen g_Δ folgende skalierten Normen definiert,

$$\|g_\Delta\|_{\Delta,2} = \Big(h^2 \sum_{(x,y)\in\Omega_h} g_\Delta^2(x,y)\Big)^{1/2}, \quad \|g_\Delta\|_{\Delta,\infty} = \max_{(x,y)\in\Omega_h} |g_\Delta(x,y)|. \tag{3.2.22}$$

Beide Normen sind Approximationen für die jeweiligen Funktionen-Normen,

$$\|g\|_2 = \Big(\int_0^1\int_0^1 g(x,y)^2\,dxdy\Big)^{1/2}, \quad \|g\|_\infty = \sup_{(x,y)\in\Omega} |g(x,y)|,$$

insbesondere gilt $\|g_\Delta\|_{\Delta,2} \le \|g_\Delta\|_{\Delta,\infty}$. Durch Kombination der Ergebnisse aus Formel (3.2.21) und den Sätzen 3.2.5, 3.2.8, 3.2.9, folgt die globale Fehleraussage

Satz 3.2.10 *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 jeder der Normen aus (3.2.22) die Abschätzung*

$$\|u_\Delta - u\|_{\Delta,p} \le C_p h^2(\|u_{xxxx}\|_\infty + \|u_{yyyy}\|_\infty),$$

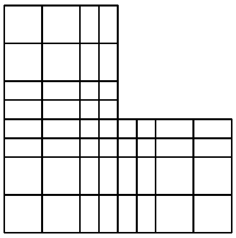
$p \in \{2,\infty\}$, mit $C_2 = 1/192$ $(h \le \frac{1}{2})$, $C_\infty = 1/96$.

Beweis Da die Werte von u und u_Δ auf dem Rand Γ_h übereinstimmen, treten in der Fehlergleichung (3.2.21) nur Beiträge aus dem Konsistenzfehler T_Δ auf, der mit Satz 3.2.9 (Konstante

$C = 1/12$) abgeschätzt werden kann. Daher sind die Sätze 3.2.5 und 3.2.8 anwendbar (homogene Randbedingungen) und ergeben die Behauptung. ∎

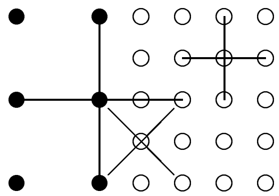
Für das gemischte Randwertproblem (3.2.12) erhält man nach dem selben Prinzip eine ähnliche Aussage mit etwas größeren Konstanten.

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). In diesem Beispiel wächst der lokale Konsistenzfehler (3.2.20) 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 kann das folgendermaßen skizziert werden:



Im groben Gitter (ausgefüllte Punkte) wird der normale Fünfpunktestern 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 diagonaler Fünf-Punkte-Stern mit Schrittweite $h/\sqrt{2}$ benutzt werden (vgl. Lemma 3.2.12). 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 3.2.7 und dem Testvektor aus Satz 3.2.8 abschätzen. Die lokale Gitterverfeinerung kann in der Nähe einer Problemstelle natürlich mehrfach angewendet werden.

Beispiel 3.2.11 Auf dem L-Gebiet $\Omega = (0,1) \times (0,1) \setminus [\frac{1}{2}, 1) \times [\frac{1}{2}, 1)$ gelte $-u_{xx} - u_{yy} = 25$, auf dem Rand $u = 0$. Einsatz ineinandergeschachtelte Gitter mit $h = 1/8,\ 1/16,\ 1/32$. Der steile Anstieg in der einspringenden Ecke wird erst durch die lokal feinere Unterteilung sichtbar.

Wenn die (höheren) Ableitungen der Lösung zu groß werden, kann dies also durch Gitterverfeinerung kompensiert werden. Umgekehrt ist es natürlich sinnvoll, den Aufwand durch Verfahren höherer Konvergenzordnung zu senken, wenn die höheren Ableitungen der Funktion gutartig sind. Die Kenntnis der genauen Fehlerstruktur des oben verwendeten diagonalen Sterns kann dazu verwendet werden.

Lemma 3.2.12 *Für eine Funktion $u \in C^6(\bar{\Omega})$ gilt*

$$\frac{1}{2h^2}\Big(4u(x,y) - u(x+h,y+h) - u(x+h,y-h) - u(x-h,y+h) - u(x-h,y-h)\Big)$$
$$= -\Big(u_{xx} + u_{yy} + \tfrac{h^2}{12}(u_{xxxx} + 6u_{xxyy} + u_{yyyy})\Big)\big|_{(x,y)} + \mathcal{O}(h^4),$$

Beweis Übungsaufgabe.

Zum Vergleich lautet der Fehler des normalen Fünf-Punkte-Sterns (Satz 3.2.9)

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

Bei der Poissongleichung kann nun durch eine geeignete Linearkombination dieser beiden Sterne ein h^2-Fehlerterm erzeugt werden, der sich mit Hilfe der Dgl wieder approximieren läßt. Wendet man nämlich den Laplace-Operator noch einmal auf die Poissongleichung an, folgt

$$-g_{xx} - g_{yy} = u_{xxxx} + 2u_{xxyy} + u_{yyyy}.$$

Diejenige Linearkombination der beiden Fünfpunktesterne, deren h^2-Fehleranteil den Ausdruck auf der rechten Seite enthält, entspricht dem folgenden *Neun-Punkte-Stern*

$$\frac{1}{6h^2}\cdot \quad \begin{array}{ccc} -1 & -4 & -1 \\ -4 & & -4 \\ -1 & -4 & -1 \end{array}$$

Das Gewicht im Mittelpunkt ist $20/(6h^2)$. Der Ausdruck $\frac{h^2}{12}(g_{xx}+g_{yy})$ im führenden Fehlerbeitrag wird natürlich wieder durch den Fünfpunktestern approximiert. Dies führt auf das folgende Verfahren. Eine einzelne Gleichung desselben im Punkt (x,y) wird der Übersicht halber durch Angabe der Gewichte in den Sternen abgekürzt:

$$\frac{1}{6h^2}\begin{bmatrix} -1 & -4 & -1 \\ -4 & 20 & -4 \\ -1 & -4 & -1 \end{bmatrix} u_\Delta(x,y) = \frac{1}{12}\begin{bmatrix} & 1 & \\ 1 & 8 & 1 \\ & 1 & \end{bmatrix} g(x,y), \quad (x,y) \in \Omega_h. \tag{3.2.23}$$

Dieses Verfahren wird *Mehrstellenverfahren* genannt, sein lokaler 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 kann wieder mit Lemma 3.2.7 wie in Satz 3.2.8 gezeigt werden mit dem gleichen Ergebnis, $\|L_\Delta^{-1}\|_\infty \le \frac{1}{8}$. Der wesentliche Vorteil des Verfahrens ist die Kompaktheit seines Differenzensterns, der bei Anwendung nur geringe Modifikationen gegenüber dem Fünfpunktestern erfordert (Randbedingungen, Matrixstruktur). Da er aber auf der speziellen Form der Poissongleichung beruht, kann er nur in diesem Spezialfall verwendet werden.

Eine auch bei allgemeineren Dgln verwendbare Methode zur Erhöhung der Konsistenzordnung ist der Einsatz von Differenzenausdrücken mit mehr als drei Punkten in jeder Ortsrichtung.

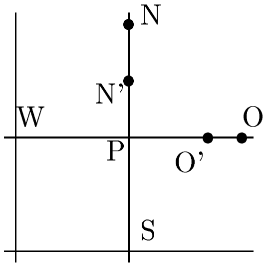
In (1.7.7) wurde eine symmetrische Approximation für u_{xx} mit fünf Funktionswerten angegeben. Die Kombination mit der entsprechenden Approximation für u_{yy} führt bei der Poissongleichung auf den Neun-Punkte-Stern

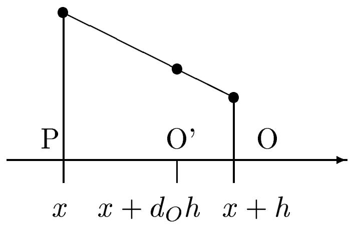
$$\frac{1}{12h^2}\cdot \begin{array}{ccccc} & & 1 & & \\ & & -16 & & \\ 1 & -16 & 60 & -16 & 1 \\ & & -16 & & \\ & & 1 & & \end{array}$$

mit einem $\mathcal{O}(h^4)$-Fehler. Da hier jetzt Werte verknüpft werden mit einem maximalen Abstand von 4 Gitterpunkten, ergeben sich erhebliche Probleme am Rand. Insbesondere bei nicht achsenparallelen Ecken können diese Schwierigkeiten auch nicht immer durch Verwendung unsymmetrischer Approximationen behoben werden. Dieses Verfahren wird nicht weiter betrachtet.

3.2.4 Allgemeinere Gebiete und Gleichungen

Der Fünfpunktestern kann in allen Gebieten verwendet werden, dessen Randpunkte auf ein geeignetes quadratisches oder rechteckiges ($h_x \neq h_y$) Gitter Δ_h fallen. Für nichtpolygonale Ränder ('krummlinige') läßt es sich aber meist nicht mehr erreichen, daß alle Nachbarn eines inneren Gitterpunkts entweder wieder Gitter- oder Randpunkte sind. Daher ist nun der Fall zu betrachten, daß ein 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 Fünfpunktesterns zu den Punkten P, N', O', S, W als Approximation an $-(u_{xx}+u_{yy})$ bestimmen. Auf das gleiche Ergebnis kommt man, wenn 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 werden. Wenn $d_N h, d_O h,\ 0 \le d_N, d_O \le 1$ die Abstände der Randpunkte von P sind, dann ist

$$\phi_d(O') = d_O v_O + (1-d_O)v_P \ \Rightarrow\ \ v_O = [\phi_d(O') - (1-d_O)v_P]/d_O.$$

Durch Einsetzen in die zweite Differenz in x-Richtung, Anwendung auf u und Taylorentwicklung führt dies auf

$$-u(x-h,y)+(1+\frac{1}{d_O})u(x,y)-\frac{1}{d_O}u(x+d_Oh,y)=$$

$$-u(W)+(1+\frac{1}{d_O})u(P)-\frac{1}{d_O}u(O')=-\frac{h^2}{2}(1+d_O)u_{xx}+\frac{h^3}{6}(1-d_O^2)u_{xxx}+\ldots \qquad (3.2.24)$$

Division durch $h^2(1+d_O)/2$ liefert eine Approximation der zweiten Ableitung. Analoges gilt in y-Richtung. Im oben skizzierten Beispiel ergibt sich so im Punkt P die Gleichung

$$\frac{1}{h^2}[(\frac{2}{d_O}+\frac{2}{d_N})v_P-\frac{2}{1+d_O}v_W-\frac{2}{d_O(1+d_O)}v_{O'}-\frac{2}{1+d_N}v_S-\frac{2}{d_N(1+d_N)}v_{N'}]=g_P. \quad (3.2.25)$$

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 (3.2.25) diagonaldominant ist. Die Normabschätzung in der ∞-Norm aus §3.2.2 kann hier wiederholt werden, wobei die Konstanten vom Gebiet abängen.

Bei der Diskussion des Gesamtfehlers ist allerdings eine ausführlichere Argumentation notwendig, da die (Konsistenz-) Ordnung des jetzt asymmetrischen Sterns (3.2.25) nur noch eins ist, wie man an der Taylorentwicklung (3.2.24) erkennt. Dennoch kann weiter $\mathcal{O}(h^2)$-Konvergenz des Gesamtverfahrens gezeigt werden, da der größere lokale Fehler nur am Rand auftritt. Hierbei spielt die M-Eigenschaft der Matrix wieder die entscheidende Rolle, da mit ihr Vergleiche zwischen Vektoren übertragen werden können. Es bezeichne L_Δ wieder die lineare Abbildung, die sich beim Dirichletproblem bei Anwendung des normalen Fünf-Punkte-Sterns in inneren Gitterpunkten und einer Gleichung der Form (3.2.25) in randnahen Punkten ergibt. Bei Anwendung dieser Abbildung auf die konstante Funktion $e_\Delta(x,y)\equiv 1$ liefert das Bild $L_\Delta e_\Delta$ (die Zeilensumme der Matrix) 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 in (3.2.25) ergibt dies

$$(L_\Delta e_\Delta)(P)=\frac{1}{h^2}[\frac{2}{d_O}+\frac{2}{d_N}-\frac{2}{1+d_O}-\frac{2}{1+d_N}]=\frac{2}{h^2}[\frac{1}{d_O(1+d_O)}+\frac{1}{d_N(1+d_N)}]>0. \quad (3.2.26)$$

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 dem des normalen Sterns aus Satz 3.2.9 entspricht, während hr_Δ nur Randfehler enthält. Im Randpunkt P liefert Gleichung (3.2.25) nach (3.2.24) hierzu den Beitrag

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

Mit $K_j := \|\partial^j u/\partial x^j\|_\infty + \|\partial^j u/\partial y^j\|_\infty$, $j=3,4$, kann dieser Beitrag mit (3.2.26) verglichen werden, dabei gilt

$$h|r_\Delta(P)|\le K_3\frac{h}{3}[1-d_O+1-d_N]\le K_3\frac{h}{3}[\frac{1}{1+d_O}+\frac{1}{1+d_N}]\le K_3\frac{h^3}{6}(L_\Delta e_\Delta)(P).$$

Aus der üblichen Fehlergleichung $L_\Delta(u_\Delta - u) = -T_\Delta = -(hr_\Delta + h^2 t_\Delta)$ kann man damit wegen $L_\Delta^{-1} \ge 0$ komponentenweise Abschätzungen herleiten. In jedem Gitterpunkt gilt

$$\begin{aligned} |u_\Delta - u| &= |L_\Delta^{-1}(hr_\Delta + h^2 t_\Delta)| \le L_\Delta^{-1}(h|r_\Delta| + h^2|t_\Delta|) \le L_\Delta^{-1}(K_3 \frac{h^3}{6} L_\Delta e_\Delta + h^2|t_\Delta|) \\ &\le (K_3 \frac{h^3}{6} + K_4 \frac{h^2}{12} \|L_\Delta^{-1}\|_\infty) e_\Delta, \end{aligned}$$

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

Beispiel 3.2.13 Auf dem Viertelkreis $\Omega = \{(x,y):\ x^2+y^2 < 1, x > 0, y > 0\}$ sei $-u_{xx} - u_{yy} = 4$ sowie $u = 0$ auf dem Kreis und $u_x(0,y) = 0, u_y(x,0) = 0$. Da die Lösung $u(x,y) = 1 - x^2 - y^2$ ist, arbeitet das Verfahren exakt.

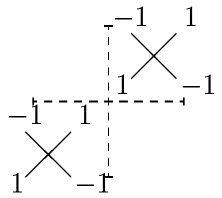
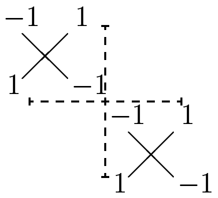
Bei einer *allgemeinen linearen, elliptischen Dgl* 2.Ordnung

$$au_{xx} + 2bu_{xy} + du_{yy} + eu_x + fu_y + qu = g \tag{3.2.27}$$

$(ad - b^2 > 0)$ ist gegenüber den schon aus §1.7 und §3.2.1 bekannten Approximationen nur noch die der gemischten Ableitung u_{xy} nachzutragen. Diese kann aus dem Produkt der ersten Differenzen in x- und y-Richtung zusammengesetzt werden. Mit Schrittweite t hat dieser einfache Stern für u_{xy} die Form

$$\frac{1}{t^2} * \begin{array}{ccc} -1 & & 1 \\ & \times & \\ 1 & & -1 \end{array} u(x,y) \quad := \quad \frac{1}{t^2}[u(x+t,y+t) + u(x-t,y-t) - u(x+t,y-t) - u(x-t,y+t)]$$

Er ist mit Zentrum (x,y) und Schrittweite $t = 2h$ eine $O(h^2)$-Approximation an $u_{xy}(x,y)$. Beim Einsatz dieses Sterns besitzen dann aber die Nebendiagonalelemente der entstehenden Matrix (die Gewichte der Sternstrahlen) nicht mehr alle das gleiche Vorzeichen, was für die M-Matrix-Eigenschaft entscheidend war. Dies kann jedoch durch die Verwendung von zwei anderen Sternen in Abhängigkeit vom Vorzeichen von b erreicht werden. Dazu setzt man jeweils zwei der vier Sterne mit Achsenlänge h, die um $(x \pm h/2, y \pm h/2)$ zentriert sind, durch Mittelung zu einem um (x,y) zentrierten zusammen, der einen Fehler $O(h^2)$ besitzt:

$\frac{1}{2h^2} *$  oder $\frac{1}{2h^2} *$ 

Von diesen Sternen wird der linke verwendet in Punkten, wo $b > 0$, der rechte, wenn $b < 0$ ist:

$$u_{xy} \cong \frac{1}{2h^2} \cdot \begin{array}{ccc} & -1 & 1 \\ -1 & 2 & -1 \\ 1 & -1 & \end{array} \cdot u, \qquad u_{xy} \cong \frac{1}{2h^2} \cdot \begin{array}{ccc} -1 & 1 & \\ 1 & -2 & 1 \\ & 1 & -1 \end{array} \cdot u.$$

Gemeinsam führen diese Approximationen für die allgemeine Dgl (3.2.27) in jedem Gitterpunkt (x, y) auf folgende Differenzengleichung, die wieder in Stern/Matrixform geschrieben wird. Alle Koeffizienten sind ebenfalls in (x, y) auszuwerten. Dabei bedeutet $b^+ := \max\{b, 0\} \geq 0$, $b^- := \min\{b, 0\} \leq 0$.

$$\frac{1}{h^2}\begin{bmatrix} -b^- & d-|b| & b^+ \\ a-|b| & -2(a+d-|b|) & a-|b| \\ b^+ & d-|b| & -b^- \end{bmatrix} u + \frac{1}{2h}\begin{bmatrix} & f & \\ -e & & e \\ & -f & \end{bmatrix} u + qu = g. \tag{3.2.28}$$

Insbesondere reicht also bei allgemeineren Gleichungen ein Fünfpunktestern nicht mehr aus. Unter einigen Einschränkungen an die Koeffizienten, die einerseits eine Verschärfung der Elliptizität der Dgl darstellen (vgl. Def. 3.1.5) bzw. der Schrittweitenbeschränkung aus Satz 1.7.2 entsprechen, läßt sich auch hier wieder zeigen, daß der Stern (3.2.28) eine M-Matrix erzeugt.

Satz 3.2.14 *Für die Koeffizienten der Dgl (3.2.27) gelte in jedem Gitterpunkt die Einschränkung*

$$|b| + \frac{h}{2}|e| < a, \quad |b| + \frac{h}{2}|f| < d, \ q \geq 0.$$

Für Gebiete, bei denen alle relevanten Randpunkte auf dem Gitter liegen, ist die negative Koeffizientenmatrix zu (3.2.28) beim Dirichletproblem eine M-Matrix.

Bei allgemeineren Gebieten oder Randbedingungen sind außerdem in jedem randnahen Punkt spezielle Randsterne analog zu (3.2.25) zu erstellen. Das im nächsten Paragraphen besprochene Verfahren ist im Vergleich dazu 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

Ein fundamentales physikalisches Prinzip besagt, daß jedes physikalische System den Zustand einnimmt, in dem seine Gesamtenergie minimal ist. Dieser Zustand kann dann *auch* durch eine Differentialgleichung beschrieben werden. Im eindimensionalen sei $u \in C^2[0, 1]$ die Lösung des folgenden Randwertproblems

$$-\left(p(x)u'(x)\right)' + q(x)u(x) = g(x), \ x \in (0, 1), \quad u(0) = u(1) = 0, \tag{3.3.1}$$

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. Wenn auf beiden Seiten der Dgl in (3.3.1) das Innenprodukt mit einem beliebigen anderen Element $v \in C_0^2[0, 1]$ gebildet wird, 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 schon für einmal differenzierbare Funktionen erklärt, damit läßt sich folgende symmetrische *Bilinearform* $a : C_0^1[0,1] \times C_0^1[0,1] \to \mathbf{R}$ definieren

$$a(u,v) := \int_0^1 [p(x)u'(x)v'(x) + q(x)u(x)v(x)]\,dx, \quad u,v \in C_0^1[0,1]. \tag{3.3.2}$$

Das Innenprodukt $(g,v)_2$ ist ein beschränktes lineares Funktional, das mit f bezeichnet wird, $f : C[0,1] \to \mathbf{R},\ v \mapsto (g,v)_2$. Aus der obigen Umformung folgt also, daß die Lösung u des RWPs (3.3.1) auch folgende Gleichung erfüllt

$$a(u,v) = f(v) \quad \forall v \in C_0^1[0,1]. \tag{3.3.3}$$

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. Aufgrund dieser Tatsache ist die Lösung u auch Minimalstelle des konvexen Funktionals

$$J(v) := \frac{1}{2}a(v,v) - f(v), \text{ d.h.,} \quad J(u) = \inf\{J(v) :\ v \in C_0^1[0,1]\}. \tag{3.3.4}$$

Mit beliebigem $0 \neq v \in C_0^1[0,1], 0 \neq t \in \mathbf{R}$, gilt für das Element $u + tv$ nämlich

$$\begin{aligned} J(u+tv) &= \frac{1}{2}a(u+tv, u+tv) - f(u+tv) \\ &= \frac{1}{2}a(u,u) - f(u) + t[a(u,v) - f(v)] + \frac{t^2}{2}a(v,v) = J(u) + \frac{t^2}{2}a(v,v) \\ &> J(u). \end{aligned} \tag{3.3.5}$$

Umgekehrt wäre für $a(u,v) - f(v) \neq 0$ auch u keine Minimalstelle. Daher stimmen die Lösungen des RWPs (3.3.1), von (3.3.3) und (3.3.4) überein, wenn sie in C_0^2 liegen. Es muß aber betont werden, 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 $\Omega \subseteq \mathbf{R}^2$ ein beschränktes Gebiet mit einem Rand Γ aus stückweise stetig differenzierbaren Kurven. Nach der Greenschen Formel gilt hier die Beziehung

$$\int\int_\Omega [u_x v_x + u_y v_y]\,dx\,dy = -\int\int_\Omega [u_{xx} + u_{yy}]v\,dx\,dy + \oint_\Gamma \frac{\partial u}{\partial \mathbf{n}} v\,ds \tag{3.3.6}$$

für $u \in C^2(\bar\Omega)$ und $v \in C^1(\bar\Omega)$. Diese Relation spielt wieder eine Rolle, wenn man Störungen $u + tv$ einer Lösung u betrachtet. Speziell beim Dirichletproblem müssen u und $u + tv$ die gleichen Randwerte besitzen, also v auf dem Rand verschwinden. In diesem Fall entfällt daher das Randintegral in (3.3.6). Zur Vereinfachung wird wieder das Dirichletproblem bei der (etwas erweitereten) Poissongleichung mit $q \geq 0$ diskutiert,

$$-u_{xx} - u_{yy} + qu = g \quad \text{in } \Omega, \quad u = 0 \quad \text{auf } \Gamma. \tag{3.3.7}$$

Zu diesem Randwertproblem kann schon für Funktionen aus dem Raum $C_0^1(\Omega)$, wobei

$$C_0^k(\Omega) := \{u \in C^k(\Omega) : \; u|_\Gamma \equiv 0\}, k = 0, 1,$$

die Bilinearform $a : C_0^1(\Omega) \times C_0^1(\Omega) \to \mathbf{R}$ und die Linearform $f : C_0^1(\Omega) \to \mathbf{R}$ gebildet werden,

$$\begin{aligned} a(u,v) &:= \int\int_\Omega [u_x v_x + u_y v_y + quv]\, dx\, dy, \\ f(v) &:= (g,v)_2 = \int\int_\Omega uv\, dx\, dy, \end{aligned} \qquad u, v \in C_0^1(\Omega). \tag{3.3.8}$$

Aufgrund der Greenschen Formel gilt also wieder:

$$u \in C_0^2(\Omega) \text{ löst } (3.3.7) \Rightarrow \quad a(u,v) = f(v) \quad \forall v \in C_0^1(\Omega).$$

Auch Neumann- oder Cauchy-Randbedingungen können in einer ähnlichen Formulierung berücksichtigt werden. Dazu werden die Restriktionen an u in der Definition des Raums C_0^1 nur auf Γ_d vorgenommen, für Γ_n, Γ_c werden dafür Randintegrale mit uv, u in die Funktionale a bzw. f aufgenommen. Im folgenden wird aber weiter nur das Dirichletproblem behandelt. Es läßt sich zeigen, daß

$$\|v\|_{1,2} := \Big(\int\int_\Omega [v_x^2 + v_y^2]\, dx\, dy\Big)^{1/2} \tag{3.3.9}$$

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 Mimimalstelle (3.3.4) von $J(v) = \frac{1}{2}a(v,v) - f(v)$ charakterisiert werden kann mit den Bi-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

$$W_0^1(\Omega) := \{u \in L_2(\Omega) : \; u_x, u_y \in L_2(\Omega), u|_\Gamma = 0\}. \tag{3.3.10}$$

Jede klassische Lösung des RWPs (vgl. Def. 3.2.1) ist auch globale Minimalstelle von J in $W_0^1(\Omega)$. Diese Erweiterung des Grundraums von C_0^1 auf W_0^1 hat große 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 im letzten Abschnitt hergeleiteten Minimaleigenschaft der Lösung können in elementarer Weise Näherungsverfahren konstruiert werden dadurch, daß das Minimum des Energie-Funktionals J nicht mehr im Gesamtraum V $(= W_0^1[0,1]$ bzw. $= W_0^1(\Omega))$ gesucht wird, sondern nur noch in einem endlichdimensionalen linearen Unterraum $V_\Delta \subseteq V$. Man kann bei V_Δ zunächst an die Menge aller Polynome eines bestimmten Maximalgrads denken, flexibler sind aber die im nächsten Abschnitt besprochenen Spline-Funktionen, die nur noch stückweise stetig differenzierbar sind. Es sei nun also

$$V_\Delta \subseteq V, \quad \dim V_\Delta = n \in \mathbf{N},$$

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

$$J(u_\Delta) = \min_{v\in V_\Delta} J(v) = \min_{v\in V_\Delta} \frac{1}{2}a(v,v) - f(v). \tag{3.3.11}$$

Für sie Analyse dieses Verfahrens gibt es ein allgemeines Prinzip. Als wesentliche Voraussetzung wird dabei die Vergleichbarkeit der von der Bilinearform a induzierten Norm und der Standard-(Sobolev-) Normen benötigt. Daher wird jetzt angenommen, daß Konstanten $\mu_0, \mu_1, M_1 > 0$ existieren so, daß

$$\mu_0\|v\|_2^2 \le \mu_1\|v\|_{1,2}^2 \le a(v,v) \le M_1\|v\|_{1,2}^2 \quad \forall v \in V. \tag{3.3.12}$$

Diese Voraussetzung ist in den betrachteten Beispielen (3.3.2) bzw. (3.3.8) erfüllt, wenn die Koeffizienten beschränkt sind und $p \ge p^* > 0,\ q \ge 0$ 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*

$$A\eta = \gamma, \quad A = \Big(a(B_i, B_j)\Big)_{i,j=1}^n,\ \gamma = \Big(f(B_i)\Big)_{i=1}^n. \tag{3.3.13}$$

Die Matrix $A = (a_{ij})$ ist symmetrisch, positiv definit. Äquivalent zu (3.3.13) ist die Aussage

$$a(u_\Delta, v) = f(v)\ \forall v \in V_\Delta. \tag{3.3.14}$$

Bemerkung: Die Bedingung (3.3.14) heißt *Galerkin-Bedingung*. Sie definiert auch dann brauchbare Lösungen, wenn die Bilinearform a nicht mehr definit ist, eine Minimallösung (3.3.11) also nicht mehr existiert.

Beweis Es sei $v = \sum_{j=1}^n \xi_j B_j \in V_\Delta,\ \xi := (\xi_i)$, beliebig. Dann gilt wegen der Bilinearität von $a(.,.)$ und wegen (3.3.12) zunächst

$$\xi^T A\xi = \sum_{i,j=1}^n \xi_i\xi_j a(B_i, B_j) = a\Big(\sum_{i=1}^n \xi_i B_i, \sum_{j=1}^n \xi_j B_j\Big) = a(v,v) \ge \mu_0\|v\|_2^2 > 0,$$

für $v \neq 0$, d.h., $\xi \neq 0$. Also ist A definit und somit invertierbar. Daraus folgt

$$\begin{aligned} J(v) &= \frac{1}{2}a\Big(\sum_{i=1}^{n}\xi_i B_i, \sum_{j=1}^{n}\xi_j B_j\Big) - f(\sum_{i=1}^{n}\xi_i B_i) = \frac{1}{2}\xi^T A\xi - \gamma^T\xi \\ &= \frac{1}{2}(\xi - A^{-1}\gamma)^T A(\xi - A^{-1}\gamma) - \frac{1}{2}\gamma^T A^{-1}\gamma. \end{aligned}$$

Dieser Ausdruck wird minimal genau dann, wenn $\xi = A^{-1}\gamma = \eta$ ist, also $v = u_\Delta$. Nach Definition entspricht (3.3.13) der Bedingung $a(u_\Delta, B_i) = f(B_i),\ i = 1,..,n$. 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 von $J(v) - J(u)$ und $J(v)$ 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*

$$\mu_1\|u_\Delta - u\|_{1,2}^2 \le a(u_\Delta - u, u_\Delta - u) = \inf_{v\in V_\Delta} a(v-u, v-u) \le M_1 \inf_{v\in V_\Delta} \|v-u\|_{1,2}^2.$$

Durch diesen Satz erhält man sofort eine globale Fehlerschranke in der $1,2$-Norm, also eine Schranke für den Fehler in der Ableitung von u_Δ, wenn man rechts eine beliebige, geeignete Funktion $v \in V_\Delta$ einsetzt, z.B., eine Interpolierende der Lösung u. Schranken für den Abstand $\mathrm{dist}_{1,2}(u, V_\Delta) = \inf_{v\in V_\Delta}\|v-u\|_{1,2}$ 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 Wahl des endlichdimensionalen Raums V_Δ ab. Man sieht schnell, daß ein Raum von Polynomen bei partiellen Problemen sehr unhandlich wird. Gerade bei der Behandlung unregelmäßiger Gebiete ist dagegen der folgende Zugang sehr anpassungsfähig, bei dem das Grundgebiet in kleine, einfache *Elemente* zerlegt wird. Hierbei ist es sehr hilfreich, daß das Minimalproblem (3.3.4) schon für stückweise differenzierbare Funktionen erklärt ist.

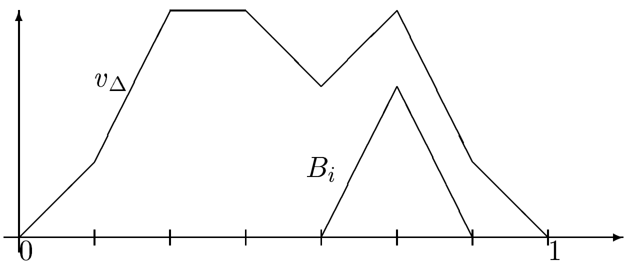
Zunächst wird das gewöhnliche Randwertproblem (3.3.1) auf $[0,1]$ betrachtet. Eine Unterteilung des Intervalls führt auf ein Gitter $\Delta:\ 0 = x_0 < x_1 < .. < x_m = 1$, mit Schrittweiten $h_i = x_{i+1} - x_i$. Die einfachste Funktionenklasse in $W_0^1[0,1]$ sind die stückweise linearen Funktionen (lineare Splines, vgl. Numerik I, §4). Unter Berücksichtigung der Randbedingungen sei

$$V_\Delta := \{s \in C_0^0[0,1]:\ s|_{[x_i,x_{i+1})} \in \Pi_1, i = 0,..,m-1\} \tag{3.3.15}$$

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

$$B_i(x) := \begin{cases} (x - x_{i-1})/h_{i-1}, & x \in [x_{i-1}, x_i) \\ (x_{i+1} - x)/h_i, & x \in [x_i, x_{i+1}) \\ 0, & \text{sonst} \end{cases}, \tag{3.3.16}$$

$i = 1, \ldots, m-1$, gegeben, vgl. die Skizze.



Diese B_i bilden eine Kardinalbasis, da für $x = x_k$ gilt $u_\Delta(x_k) = \sum \eta_j B_j(x_k) = \eta_k$. Der wichtigste praktische Vorteil der genannten Basis ist aber ihre Lokalität, die dazu führt, daß nur wenige der Matrixelemente $a(B_i, B_j)$ in (3.3.13) von Null verschieden sind. Dies reduziert den Rechenaufwand sowohl bei der Aufstellung des Gleichungssystems, als auch bei seiner Auflösung.

Als Beispiel werde die Dgl $-u'' = g$ mit $p \equiv 1, q \equiv 0$, also die 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 $j \in \{i-1, i, i-1\}$ gilt

$$a(B_i, B_j) = \int_{x_{i-1}}^{x_i} \frac{1}{h_{i-1}} B_j' dx - \int_{x_i}^{x_{i+1}} \frac{1}{h_i} B_j' dx = \begin{cases} -\frac{1}{h_{i-1}}, & j = i-1 \\ \frac{1}{h_{i-1}} + \frac{1}{h_i}, & j = i \\ -\frac{1}{h_i}, & j = i+1 \end{cases}.$$

Das Verfahren arbeitet daher, zumindestens bei der Approximation der zweiten Ableitung, ähnlich wie das Differenzenverfahren, denn es führt auf das Gleichungssystem ($\eta_i = u_\Delta(x_i)$, s.o.)

$$-\frac{1}{h_i}(\eta_{i+1} - \eta_i) + \frac{1}{h_{i-1}}(\eta_i - \eta_{i-1}) = \int_{x_{i-1}}^{x_{i+1}} g(x) B_i(x) dx, \; i = 1, .., m-1.$$

Für das RRG-Verfahren im Raum (3.3.15) sollte man daher auch einen Fehler der Größenordnung h^2 erwarten. Dieser läßt sich mit Hilfe von Satz 3.3.2 auf den Approximationsfehler $\inf\{\|u - v_\Delta\| : v_\Delta \in V_\Delta\}$ zurückführen. Bei der Interpolierenden in V_Δ kann der Fehler unabhängig auf den einzelnen Teilintervallen untersucht werden. Die Interpolierende einer Funktion $v \in V$ ist eine Restriktion $R_\Delta v$ im Sinne von §2.

Lemma 3.3.3 *Zu einer Funktion $v \in C_0^2[0,1]$ sei $R_\Delta v := \sum_{i=1}^{m-1} v(x_i)B_i$ die Interpolierende aus V_Δ. Dann gilt mit $H := \max_i h_i$ die Fehlerschranke*

$$\|(R_\Delta v - v)^{(k)}\|_2 \leq \left(\frac{H}{\pi}\right)^{2-k} \|v''\|_2, \ k = 0, 1.$$

Auf diese Aussage kommt man direkt über die Standarddarstellung des Interpolationsfehlers für Polynome (Numerik I, 3.3.11), zur Verifikation der angegebenen Konstanten ist allerdings eine diffizilere Argumentation nötig. Wird das Infimum in Satz 3.3.2 durch Einsetzen von $v = R_\Delta u$ abgeschätzt, ergibt sich mit dem letzten Lemma die Fehlerschranke,

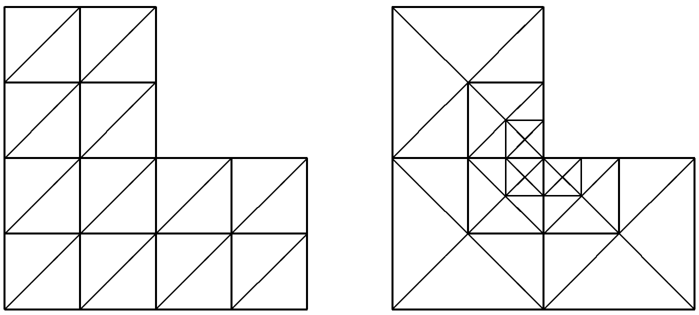
$$\|(u_\Delta - u)'\|_2 \leq \sqrt{\frac{M_1}{\mu_1}} \frac{H}{\pi} \|u''\|_2.$$

Mit einem zusätzlichen Argument ('Nitsche-Trick') kann 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 Funktionswerte hergeleitet werden. Im Vergleich zu den entsprechenden Aussagen bei Differenzenverfahren ist dabei hervorzuheben, daß hier nur die L_2-Integrierbarkeit von u'' ausgenutzt wurde im Vergleich zur Voraussetzung $u \in C^4$ bei den Differenzenverfahren. RRG-Verfahren höherer Ordnung erhält man durch Verwendung von Splines höherer Ordnung, wobei die Funktionen in V_Δ stückweise aus Polynomen höheren Grads zusammengesetzt sind.

Von größter praktischer Bedeutung ist die Möglichkeit, beim RRG-Verfahren mit stückweise linearen Ansatzfunktionen arbeiten zu können, bei den partiellen Problemen. Sehr einfach sind Zerlegungen bzw. Approximationen des Gebiets Ω durch *Dreieckgitter*. Daher ist das RRG-Verfahren wesentlich 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,

$$\Delta = \{T_i\}_{i=1}^m, \quad \bar{\Omega} = \bigcup_{i=1}^m \bar{T}_i,$$

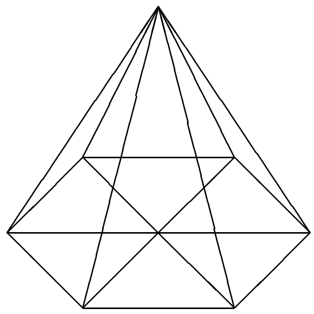
bei denen alle Ecken eines Dreiecks auf Ecken seiner Nachbardreiecke oder den Rand fallen. Der Raum aller stetigen, stückweise linearen Funktionen auf Δ,

$$V_\Delta := \{s \in C_0(\Omega) : \ s|_{T_i} \text{ linear}\}, \qquad (3.3.17)$$

kann wieder durch eine Basis von Dachfunktionen aufgespannt werden. Es sei $\{P_i\}_{i=1}^n \subseteq \Omega$ die Menge aller Eckpunkte von Dreiecken (Knoten) im Innern des Gebiets in einer geeigneten Numerierung. Dann wird durch

$$B_i \in V_\Delta,\ B_i(P_j) = \delta_{ij},\ j = 1, \ldots, n,$$

$i = 1, \ldots, n$, eine Basis von V_Δ definiert. Ein Beispiel einer solchen Basisfunktion bei sechs in einer Ecke zusammenstoßenden Dreiecken ist



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.

Diese Eigenschaft ist gerade bei mehrdimensionalen Problemen von großer Bedeutung, da dann nur wenige der Integrale $a(B_i, B_j)$ in der Matrix A von Null verschieden sind, und bei den restlichen für $i \neq j$ nur über zwei Dreiecke integriert werden muß.

$$a(B_i, B_j) = \sum_{P_i \in T_k \wedge P_j \in T_k} \int\int_{T_k} [(B_i)_x(B_j)_x + (B_i)_y(B_j)_y + qB_iB_j]\, dxdy$$

Bei der praktischen Berechnung der Integrale $a_{ij} = a(B_i, B_j)$, $\gamma_i = f(B_i)$, ist es am einfachsten, alle Dreiecke T_i abzuarbeiten und die einzelnen Teilintegrale in a_{ij}, γ_i aufzuaddieren. Die Matrix A ist hier wieder dünn besetzt und symmetrisch, pos. definit. Daher können sehr effiziente (Iterations-) Verfahren zur Lösung des Gleichungssystems (3.3.13) eingesetzt werden.

Zur Fehlerabschätzung der zugehörigen RRG-Lösung nach Satz 3.3.2 läßt sich eine Schranke für den Interpolationsfehler wieder 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 $R_\Delta v := \sum_{i=1}^n v(P_i)B_i$ die Interpolierende aus V_Δ, (3.3.17). In der Dreieckzerlegung des Gebiets sei h die längste Dreieckseite und α der kleinste Innenwinkel eines Dreiecks T_i, $i = 1, .., m$. Dann gilt die Fehlerschranke*

$$\|R_\Delta v - v\|_{1,2} \leq C \frac{h}{\sin\alpha} \|v\|_{2,2},$$

mit $\|v\|_{2,2}^2 := \sum_{j+k=2} \|\frac{\partial^2}{\partial x^j \partial y^k} v\|_2^2$.

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 RRG-Näherung u_Δ und eine Schranke $\mathcal{O}(h^2/\sin\alpha)\|u\|_{2,2}$ für den Fehler ihrer Funktionswerte herleiten. Zusätzlich kann unter einfachen Annahmen an die Koeffizienten zumindest für konvexe Gebiete die Abschätzung $\|u\|_{2,2} \le C\|g\|_2$ gezeigt werden. Daher hat man auch im partiellen Fall die h^2-Konvergenz unter schwächeren Voraussetzungen als bei den Differenzenverfahren.

Allerdings ist bei diesen starken Aussagen zu beachten, daß sie nur 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.

3.4 Partielle Anfangs-Randwert-Probleme

Die nicht-elliptischen partiellen Differentialgleichungen 2. Ordnung, die parabolischen und hyperbolischen, besitzen eine gemeinsame Eigenschaft. Schreibt man den jeweiligen einfachsten Vertreter (Wärmeleitungs-, Wellengleichung) in der Form

$$u_{xx} - u_y = 0, \quad \text{bzw.} \quad u_{xx} - u_{yy} = 0,$$

so lassen sich damit sinnvolle Problemstellungen formulieren, wenn Randbedingungen nur in x-Richtung vorgegeben werden, in y-Richtung dagegen nur einseitige, also Anfangsbedingungen. Im physikalischen Zusammenhang hat die erste Variable meist die Bedeutung einer Ortskoordinate, während die zweite einen Zeitablauf beschreibt. Daher wird sie im folgenden mit t bezeichnet. Da die Geometrie des Problems sich i.a. nicht mit der Zeit ändert, ist das Grundgebiet meist ein Zylinder, z. B., $[0,1] \times [0,\infty)$. Mit stetigen Koeffizienten $p, r, g,\ p > 0, p_x$ stetig werden also Anfangs-Randwertprobleme betrachtet mit einer der Dgln

$$u_t = [p(x,t)u_x]_x + q(x,t)u + g(x,t), \quad x \in (0,1),\ t > 0, \quad \text{(parabolisch)} \tag{3.4.1}$$

$$u_{tt} = [p(x,t)u_x]_x + q(x,t)u + g(x,t), \quad x \in (0,1),\ t > 0, \quad \text{(hyperbolisch)} \tag{3.4.2}$$

und den Rand- und Anfangsbedingungen

$$\begin{gathered} u(x,0) = \phi(x),\ x \in (0,1), \quad u(0,t) = u(1,t) = 0,\ t \ge 0, \\ u_t(x,0) = \psi(x),\ x \in (0,1) \text{ nur bei (3.4.2)}. \end{gathered} \tag{3.4.3}$$

Die Probleme (3.4.1,3.4.2),(3.4.3) verhalten sich also in Ortsrichtung wie (elliptische) Randwertprobleme, in Zeitrichtung wie Anfangswertprobleme erster bzw. zweiter Ordnung. Diese Tatsache kann bei der numerischen Approximation auch ausgenutzt werden durch Semidiskretisierung in der sogenannten *Linienmethode*. Dabei wird jeweils eines der von früher bekannten Verfahren zur Diskretisierung von Anfangs- oder Randwertproblemen getrennt auf die Zeit- und Ortsrichtung angewendet. Allerdings muß man sich dabei über die Eigenschaften der nach dem ersten Diskretisierungsschritt entstehenden gewöhnlichen Probleme klar werden. Bei der Zeitlinienmethode, z.B., wird das Problem nur in der Ortsvariablen diskretisiert, das AWP in Zeitrichtung (zunächst) beibehalten. Dabei werde zunächst das Differenzenverfahren betrachtet mit einem äquidistanten Gitter $\Delta = \{x_i = ih: \ i = 0,..,n+1\}, h = 1/(n+1)$, wie in §1.7. Der Ausdruck $[pu_x]_x$ wird dabei durch zweifache Differenzenquotienten approximiert,

$$[p(x,t)u_x(x,t)]_x = \frac{1}{h^2}[p(x_{i+1/2},t)(u(x_{i+1},t)-u(x_i,t))-p(x_{i-1/2},t)(u(x_i,t)-u(x_{i-1},t))]+\mathcal{O}(h^2),$$

$x_{i\pm 1/2} := x_i \pm h/2$. Unter Vernachlässigung des $\mathcal{O}(h^2)$-Fehlerterms ergibt sich für die Gitterwerte $u_\Delta(x_i,t)$ einer Näherung u_Δ, die immer noch Funktionen der Zeit sind, ein System von gewöhnlichen Anfangswertproblemen. Faßt man nämlich diese Funktionswerte als Komponenten einer Vektorfunktion auf, $w(t) = (w_i(t))_{i=1}^n$, $w_i(t) := u_\Delta(x_i,t)$ (und $w_0 = w_{n+1} \equiv 0$), dann führt die obige Differenzenapproximation im parabolischen Fall (3.4.1) auf das AWP erster Ordnung

$$\begin{aligned} w'(t) &= f(t,w(t)), \ t>0, \quad \text{mit} \\ f_i(t,y) &= \tfrac{1}{h^2}[p_{i+1/2}(y_{i+1}-y_i) - p_{i-1/2}(y_i - y_{i-1})] + q_i y_i + g_i, \\ w_i(0) &= \phi(x_i), \quad i = 1,\dots,n. \end{aligned} \tag{3.4.4}$$

Die Koeffizientenwerte sind, wie früher, $q_i = q(x_i,t), g_i = g(x_i,t)$ und $p_{i\pm 1/2} = p((i\pm\frac{1}{2})h,t)$. Beim hyperbolischen Problem (3.4.2) ergibt sich das entsprechende AWP mit der Dgl zweiter Ordnung $w'' = f(t,w(t))$ und der zusätzlichen Anfangsbedingung $w_i'(0) = \psi(x_i)$. Im betrachteten Beispiel ist die Funktion f affin linear, $f(t,y) = A(t)y + \vec{g}(t)$. Die Matrix A ist tridiagonal (vgl. (1.7.5)) und sogar symmetrisch mit den Elementen

$$a_{ii}(t) = -\frac{1}{h^2}[p_{i+1/2} + p_{i-1/2}] + q_i, \quad a_{i,i+1}(t) = a_{i+1,i}(t) = \frac{1}{h^2}p_{i+1/2}. \tag{3.4.5}$$

Wie in Lemma 3.2.4 kann für $p > 0$ bei dieser Matrix die Aussage $y^TA(t)y \le q^* y^T y$, $q^* = \max_i q_i$, gezeigt werden. Daher sind die Eigenwerte von A reell und durch q^* nach oben beschränkt, also insbesondere negativ für $q^* \le 0$. Allerdings wird die Norm $\|A(t)\| \cong \frac{4}{h^2}\|p\|_\infty$ für kleine Schrittweiten h sehr groß. Tatsächlich läßt sich zeigen, daß sehr kleine Eigenwerte von A (mit großem Betrag) $\lambda \ll 0$ auftreten können. Bei (3.4.4) handelt es sich somit um (eines der Standardbeispiele für) ein *steifes Anfangswertproblem*. Die Diskussion aus §1.5 ist also hier anwendbar. Die wichtigste Erkenntnis in diesem Zusammenhang war, daß zur Integration des AWPs (3.4.4) nur implizite Verfahren sinnvoll sind. Das läßt sich auch anhand der einfachsten Verfahren bei der Wärmeleitungsgleichung nachvollziehen. Bei dieser ($p \equiv 1$, $q = g \equiv 0$) wird das einfache Einschrittverfahren ($w(t_k) \cong w^{[k]} \in \mathbf{R}^n$)

$$\frac{1}{\tau}(w^{[j+1]} - w^{[j]}) = (1-\vartheta)Aw^{[j]} + \vartheta A w^{[j+1]}, \quad j = 0,1,\dots, \tag{3.4.6}$$

mit Zeitschrittweite τ und einem Parameter $\vartheta \in [0,1]$ betrachtet. Die Matrix A (3.4.5) ist konstant. Das Verfahren stimmt für $\vartheta = 0$ mit dem expliziten und für $\vartheta = 1$ mit dem impliziten

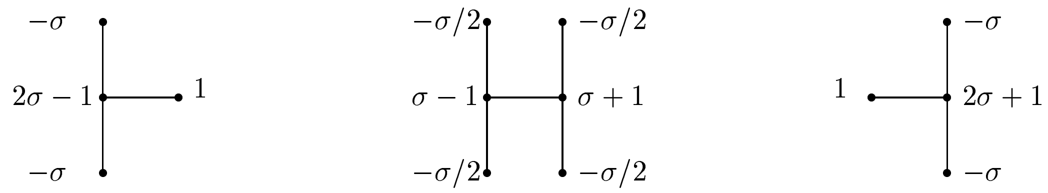
Eulerverfahren, überein, für $\vartheta = \frac{1}{2}$ ist es die implizite Mittelpunktregel. Es führt auf eine Serie von Gleichungssystemen ($\sigma := \tau/h^2$, $\vartheta' := 1 - \vartheta$),

$$-\vartheta\sigma w_{i-1}^{[j+1]} + (1+2\vartheta\sigma)w_i^{[j+1]} - \vartheta\sigma w_{i+1}^{[j+1]} = \vartheta'\sigma w_{i-1}^{[j]} + (1-2\vartheta'\sigma)w_i^{[j]} + \vartheta'\sigma w_{i+1}^{[j]}, \qquad (3.4.7)$$

$i = 1, \ldots, n$. Beim expliziten Euler-Verfahren, $\vartheta = 0$, reduziert sich dies auf

$$w_i^{[j+1]} = \sigma w_{i-1}^{[j]} + (1-2\sigma)w_i^{[j]} + \sigma w_{i+1}^{[j]},\ i = 1, \ldots, n. \qquad (3.4.8)$$

Wie im letzten Paragraphen können diese Beziehungen kompakt durch Differenzensterne beschrieben werden. Zu den verschiedenen Parameterwerten ϑ gehören die Sterne



$\vartheta = 0$ $\qquad$ $\vartheta = 1/2$ $\qquad$ $\vartheta = 1$

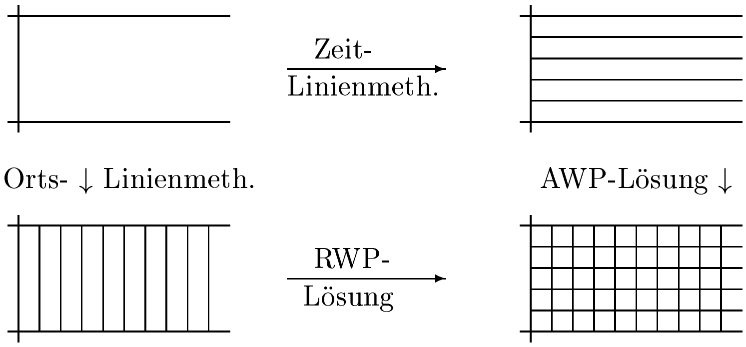
wobei die Zeitachse t nach rechts gezeichnet ist. Für $\sigma \le \frac{1}{2}$ sind alle Gewichte in (3.4.8) nichtnegativ, es findet also eine fortgesetzte Mittelwertbildung in den Zeitschritten statt. Dies ist ein gutes Modell für die durch die Wärmeleitungsgleichung beschriebenen Ausgleichsvorgänge. Allerdings muß dabei die Zeitschrittweite τ aus Stabilitätsgründen inakzeptabel klein gewählt werden, $\tau = h^2\sigma \le h^2/2$. Für $\tau > h^2/2$, d.h. $\sigma > 1/2$ wird das explizite Verfahren (3.4.8) nämlich instabil, denn dann tritt in der (scharfen) Abschätzung

$$\|w^{[j+1]}\|_\infty = \max_i |w_i^{[j+1]}| \le (2\sigma + |1-2\sigma|)\ \|w^{[j]}\|_\infty,$$

der Vorfaktor $4\sigma - 1 = 4\tau/h^2 - 1 > 1$ auf. Dies paßt zur Diskussion von §1.5. Denn, da für die Norm $\|A\| = 4/h^2$ gilt und die Stabilitätsabszisse des expliziten Eulerverfahren 2 ist, entspricht dieser Fall der instabilen Situation $\tau\|A\| > 2$.

Beim impliziten Eulerverfahren ($\vartheta = 1$) dagegen ist das Tridiagonalsystem (3.4.7) in jedem Zeitschritt diagonaldominant (mit einer M-Matrix) und hier kann ohne jede Schrittweitenbeschränkung die Schranke $\|w^{[j+1]}\|_\infty \le \|w^{[j]}\|_\infty$ gezeigt werden. Das allgemeine Verfahren (3.4.6) ist für alle $\vartheta \in [\frac{1}{2}, 1]$ A-stabil. Bei diesen Verfahren muß sich die Schrittweitenwahl nur nach den Genauigkeitsanforderungen richten. Dabei ist zu beachten, daß im AWP (3.4.4) auch schon ein Diskretisierungsfehler steckt, der Fehler beim Gesamtverfahren (3.4.6) setzt sich daher aus einem Orts- und einem Zeitfehler zusammen, die jeweils getrennt durch Wahl der Ortsschrittweite h und der Zeitschrittweite τ beeinflußt werden können. Für die Mittelpunktregel ($\vartheta = 1/2$) hat dieser globale Fehler die Form $C_xh^2 + C_t\tau^2$, für alle anderen $\vartheta \ne 0$ nur $C_xh^2 + C_t\tau$. Das Verfahren arbeitet dann am effizientestens, wenn beide Fehler ungefähr gleich groß sind. Bei der Zeitlinienmethode, mit festem Ortsgitter, kann nur der Zeitfehler beeinflußt werden. Die Zeitschrittweite τ ist daher so zu steuern, daß dieser ungefähr die Größe des Ortsfehlers besitzt.

Eine größere Anpassungsfähigkeit ist mit der Orts-Linienmethode möglich. Dabei werden die beiden Diskretisierungsschritte vertauscht. Die erste Diskretisierung erfolgt in Zeitrichtung mit einem impliziten Verfahren (z.B. dem impliziten Euler-Verfahren). Dies führt auf eine Serie von gewöhnlichen RWPen in jeder Zeitstufe. Das folgende Diagramm veranschaulicht die auftretenden Gitter für beide Möglichkeiten:



Wird bei den RWPen das Differenzenverfahren mit einer festen, konstanten Schrittweite h verwendet, ergibt sich wieder das Verfahren (3.4.6) und das im Bild gezeigte Endgitter. In dieser zweiten Interpretation ist es aber auch möglich, bei der Lösung der RWPe in jedem einzelnen Zeitschritt das Ortsgitter neu (adaptiv) zu wählen. Die im Verfahren verwendeten Werte der vorhergehenden Zeitstufe können dabei interpoliert werden. Hiermit läßt sich sowohl die Zeit- als auch die Ortsunterteilung leichter an den Lösungsverlauf anpassen, die Gitter sind dann irregulär.

Auch bei den **hyperbolisch**en Problemen (3.4.2) kann, im Prinzip, analog zum parabolischen Fall vorgegangen werden. Zur Diskretisierung der zweiten Zeitableitung u_{tt} bzw. w_{tt} (in der Zeitlinienmethode) müssen dabei allerdings drei Zeitstufen herangezogen werden. Ein explizites Verfahren ergibt sich direkt aus der Standardapproximation von u_{tt} mit Hilfe des zweiten Differenzenquotienten, für die Gitterwerte (wie oben) $w_i^{[j]} = u_\Delta(ih, j\tau), i = 1, .., n$, lautet das Gesamtverfahren

$$w_i^{[j+1]} = s^2 w_{i-1}^{[j]} + 2(1-s^2) w_i^{[j]} + s^2 w_{i+1}^{[j]} - w_i^{[j-1]},\ i = 1, \ldots, n, \tag{3.4.9}$$

mit $s = \tau/h$. Diese Rekursion kann im Zeitschritt $j = 0$ gestartet werden, wenn die zweite Anfangsbedingung aus (3.4.3) mit der symmetrischen ersten Differenz approximiert wird, $w_i^{[1]} - w_i^{[-1]} = 2\tau\psi(x_i)$, die eine künstliche Zeitebene $-\tau$ einbezieht. Durch eine etwas aufwendigere Analyse als bei den parabolischen Problemen läßt sich zeigen, daß das Verfahren ebenfalls nur stabil ist unter der (schwächeren) Schrittweiteneinschränkung $s = \tau/h \le 1$. Es können auch einfache implizite Verfahren angegeben werden, die ohne Schrittweitenbeschränkung stabil sind. Der Vergleich mit den expliziten Verfahren fällt dennoch anders aus als bei den parabolischen Problemen aufgrund der unterschiedlichen Eigenschaften der Lösung, die jetzt diskutiert werden.

Bei der parabolischen Gleichung (3.4.1) werden die Lösungen für wachsendes $t > 0$ immer glatter. Bei der Wärmeleitungsgleichung $u_t = u_{xx}$ mit $u(0,t) = u(1,t) = 0$ besteht die Lösung aus einer Linearkombination von Funktionen der Form

$$e^{-k^2\pi^2 t}\sin k\pi x, \quad k\in\mathbf{N},$$

die in t-Richtung umso schneller verschwinden, je stärker sie in x-Richtung variieren. Außerdem hängt der Lösungswert $u(x,t)$ von allen Daten im Bereich $[0,1]\times[0,t)$ ab. Beim hyperbolischen Problem ergibt sich dagegen ein anderes Bild. Die allgemeine Lösung der Wellengleichung kann explizit konstruiert werden. Jede Funktion der Form

$$u(x,t)=a(x+t)+b(x-t) \quad \text{löst} \quad u_{tt}=u_{xx} \tag{3.4.10}$$

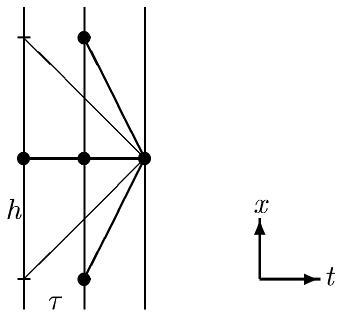
mit $a,b\in C^2(\mathbf{R})$. Wenn man die Randbedingungen in x-Richtung zur Vereinfachung außer Acht läßt, hat diese Lösung die Anfangsbedingungen

$$a(x)+b(x)=u(x,0)=\phi(x), \quad a'(x)-b'(x)=u_t(x,0)=\psi(x)$$

zu erfüllen. Durch Differentiation der ersten Gleichung können a und b bis auf Konstanten bestimmt werden. Daraus folgt die explizite Darstellung des AWPs bei der Wellengleichung,

$$u(x,t)=\frac{1}{2}\Big(\phi(x+t)+\phi(x-t)+\int_{x-t}^{x+t}\psi(\xi)\,d\xi\Big). \tag{3.4.11}$$

Zwei wichtige Eigenschaften können an dieser Lösungsdarstellung abgelesen werden. Erstens zeigt schon (3.4.10), daß die Eigenschaften der Lösung vom Startzeitpunkt praktisch unverändert fortgepflanzt werden, zweitens, daß die Lösung an einem Punkt (x,t) nur von den Startwerten im Intervall $[x-t,x+t]$, dem sogenannten Abhängigkeitsintervall, abhängt. Daher pflanzen sich Störungen (Sprünge, Schockwellen) längs der *Wellenfronten* $x\pm t=$ *const* ungedämpft fort. Diese Linien nennt man *Charakteristiken.* Die aufgezählten Eigenschaften sollten bei den numerischen Verfahren berücksichtigt werden. Dies bedeutet, daß implizite Verfahren (mit großen Schrittweiten) nur angebracht sind, wenn die Lösung von Anfang an sehr glatt ist und daß andernfalls bei expliziten Verfahren die diskreten Abhängigkeitsgebiete an die der Dgl angepaßt werden sollten. In folgender Skizze



ist zu erkennen, daß für $\tau<h$ die 'diskreten' Charakteristiken (dicke Linien) des Verfahrens (3.4.9) eine größere Steigung besitzen als die exakten (Steigung ±1), Störungen sich in der Näherung in x-Richtung also schneller fortpflanzen als in der Realität. Im Verfahren (3.4.9) ist daher das Schrittweitenverhältnis $s=1$ am günstigsten. Bei allgemeinen Dgln (3.4.2) hängen die Charakteristiken von den Koeffizienten ab, sind im allgemeinen also keine geraden Linien mehr. Hier wird die Diskussion sehr viel schwieriger.

Index