Numerische Simulation[1]

Heinrich Voß

Technische Universität Hamburg–Harburg
Arbeitsbereich Mathematik
2007

[1]27.03.2012: Ich bedanke mich bei Dominik Zobel für eine Reihe von Korrekturen

Inhaltsverzeichnis

1 Grundlagen **1**

1.1 Gewöhnliche Differentialgleichungen 1

1.2 Partielle Differentialgleichungen 10

2 Einschrittverfahren **23**

2.1 Das Eulersche Polygonzugverfahren 23

2.2 Allgemeine Einschrittverfahren 31

2.3 Extrapolationsverfahren 44

2.4 Software zu Einschrittverfahren 49

3 Mehrschrittverfahren **51**

4 Steife Probleme **63**

4.1 Motivation 63

4.2 Stabilitätsgebiete 65

4.3 Implizite Runge–Kutta Verfahren 77

4.4 Rosenbrock Verfahren 87

4.5 Extrapolation 91

4.6 Abschließende Bemerkungen zur Wahl der Verfahren 95

5 DAE vom Index 1 **97**

5.1 Einleitende Bemerkungen 97

5.2 Der Index eines DAE Systems 98

5.3 Eine Einbettungsmethode 101

5.4 Probleme mit Massenmatrizen 103

5.5 Mehrschrittverfahren 104

6 Randwertaufgaben **107**

6.1 Anfangswertmethoden 107

6.2 Differenzenverfahren 119

6.3 Variationsmethoden 127

7 Differenzenverfahren für Randwertaufgaben **143**

7.1 Das Modellproblem 143

7.2 Die Neumannsche Randwertaufgabe 148

7.3 Die Poisson Gleichung in allgemeinen Gebieten 152

7.4 Allgemeinere Differentialoperatoren 156

7.5 Idee der Methode der finiten Volumen 159

8 Finite Elemente **163**

8.1 Variationsmethoden 163

8.2 Methode der finiten Elemente 172

8.3 Fehlerabschätzung 176

8.4 Realisierung von Verfahren der finiten Elemente 180

8.5 Weitere ebene Elemente 183

8.6 Software 185

9 Parabolische Anfangsrandwertaufgaben **187**

9.1 Differenzenverfahren 187

9.2 Linienmethode 195

Literaturverzeichnis **200**

Kapitel 1

Grundlagen

Wir stellen in diesem Abschnitt Aussagen über gewöhnliche und partielle Differentialgleichungen zusammen, die wir in den folgenden Abschnitten über numerische Methoden benötigen. Verweise mit dem Zusatz MI (z.B. Satz 28.1 MI) beziehen sich dabei auf die Skripten "Mathematik für Studierende der Ingenieurwissenschaften I - IV", z.B. Satz 28.1 MI.

1.1 Gewöhnliche Differentialgleichungen

Wir betrachten die Anfangswertaufgabe

$$\boldsymbol{y}' = f(x, \boldsymbol{y}), \;\; \boldsymbol{y}(x_0) = \boldsymbol{y}^0. \tag{1.1}$$

Dabei ist $\boldsymbol{y}$ die gesuchte Funktion und $\boldsymbol{y}(x_0) = \boldsymbol{y}^0 \in \mathbb{R}^n$ der vorgegebene Anfangswert.

Für die Herleitung von Existenz- und Eindeutigkeitsresultaten wurde in Satz 28.1 MI die Anfangswertaufgabe (1.1) in eine Integralgleichung umgeformt. Wir werden diese Integralgleichung in Kapitel 3 verwenden, um numerische Verfahren zu begründen.

Satz 1.1. *Sei*

$$f \,:\, Q := \{(x, \boldsymbol{y}) \,:\, |x - x_0| \le a, \;\; \|\boldsymbol{y} - \boldsymbol{y}^0\| \le b\} \to \mathbb{R}^n$$

stetig und $\boldsymbol{y} \,:\, I := [x_0 - a, x_0 + a] \to \mathbb{R}^n$ *mit* $(x, \boldsymbol{y}(x)) \in Q$ *für alle* $x \in I$. *Dann sind äquivalent*

(i) $\boldsymbol{y}$ *ist in* I *stetig differenzierbar und löst die Anfangswertaufgabe*

$$\boldsymbol{y}'(x) = f(x, \boldsymbol{y}(x)) \quad \text{für alle } x \in I, \;\; \boldsymbol{y}(x_0) = \boldsymbol{y}^0$$

(ii) $\boldsymbol{y}$ *ist in* I *stetig und erfüllt die Integralgleichung*

$$\boldsymbol{y}(x) = \boldsymbol{y}^0 + \int_{x_0}^{x} f(t, \boldsymbol{y}(t))\, dt, \;\; x \in I. \tag{1.2}$$

Wendet man auf die Integralgleichung (1.2) den Fixpunktsatz für kontrahierende Abbildungen an, so erhält man

Satz 1.2. (Satz von Picard und Lindelöf)
Es sei $Q := \{(x, \boldsymbol{y}) \in \mathbb{R}^{n+1} : |x - x_0| \le a, \;\; \|\boldsymbol{y} - \boldsymbol{y}^0\| \le b\}$, *sei* $f : Q \to \mathbb{R}^n$ *stetig auf* Q *mit* $\|f(x, \boldsymbol{y})\| \le M$ *für alle* $(x, \boldsymbol{y}) \in Q$, *und es erfülle* f *eine Lipschitz Bedingung bzgl.* $\boldsymbol{y}$ *auf* Q, *d.h.*

$$\|f(x, \boldsymbol{y}) - f(x, \boldsymbol{z})\| \le L\|\boldsymbol{y} - \boldsymbol{z}\|$$

für alle $(x, \boldsymbol{y}), (x, \boldsymbol{z}) \in Q$.

Dann besitzt die Anfangswertaufgabe

$$\boldsymbol{y}' = f(x, \boldsymbol{y}), \;\; \boldsymbol{y}(x_0) = \boldsymbol{y}^0$$

eine eindeutige Lösung $\boldsymbol{y}(x)$, *die (wenigstens) auf dem Intervall* $[x_0 - \alpha, x_0 + \alpha]$ *mit* $\alpha := \min\,(a, \dfrac{b}{M})$ *definiert ist.*

In Satz 28.2 MI wurde dieses Ergebnis direkt (durch Konstruktion einer Folge von Funktionen, die gleichmäßig gegen eine Lösung von (1.2) konvergiert) gezeigt, da der Fixpunktsatz für kontrahierende Abbildungen im Funktionenraum nicht zur Verfügung stand.

Mit dem Fixpunktsatz von Schauder erhält man

Satz 1.3. (Existenzsatz von Peano)
Es sei $Q := \{(x, \boldsymbol{y}) \in \mathbb{R}^{n+1} : |x - x_0| \le a, \;\; \|\boldsymbol{y} - \boldsymbol{y}^0\| \le b\}$, *sei* $f : Q \to \mathbb{R}^n$ *stetig auf* Q *mit* $\|f(x, \boldsymbol{y})\| \le M$ *für alle* $(x, \boldsymbol{y}) \in Q$ *und sei* $\alpha = \min\Big(a, \dfrac{b}{M}\Big)$.

Dann besitzt die Anfangswertaufgabe

$$\boldsymbol{y}' = f(x, \boldsymbol{y}), \;\; \boldsymbol{y}(x_0) = \boldsymbol{y}^0$$

eine Lösung, die in $[x_0 - \alpha, x_0 + \alpha]$ *definiert ist.*

Bemerkung 1.4. Die Eindeutigkeit kann nicht mehr garantiert werden, denn die Funktion $f(x,y) = \sqrt{|y|}$ ist stetig, aber $y' = \sqrt{|y|}$, $y(0) = 0$, ist nicht eindeutig lösbar. □

Um die Abhängigkeit der Lösung der Anfangswertaufgabe (1.1) von den Anfangswerten und von Parametern zu diskutieren, wurde das Lemma von Gronwall benutzt.

Satz 1.5. (Lemma von Gronwall)
Es sei $\phi : I := [x_0 - a, x_0 + a] \to \mathbb{R}$ *stetig, und es gelte mit* $\alpha, \beta \geq 0$

$$0 \leq \phi(x) \leq \alpha + \beta \Big| \int_{x_0}^{x} \phi(t)\, dt \Big| \quad \text{für alle } x \in I.$$

Dann gilt

$$\phi(x) \leq \alpha \exp(\beta |x - x_0|) \quad \text{für alle } x \in I.$$

Für die Abhängigkeit der Lösungen von den Anfangswerten erhält man hiermit (vgl. Satz 28.7 MI):

Satz 1.6. *Es seien die Voraussetzungen des Satzes von Picard und Lindelöf erfüllt, und es sei* L *die Lipschitz Konstante von* f *in* Q*. Dann gilt*

$$\|\boldsymbol{y}(x; x_0, \boldsymbol{y}^0) - \boldsymbol{y}(x; x_0, \boldsymbol{z}^0)\| \leq e^{L|x - x_0|} \|\boldsymbol{y}^0 - \boldsymbol{z}^0\|$$

für alle $\boldsymbol{z}^0 \in \mathbb{R}^n$ *mit* $\|\boldsymbol{z}^0 - \boldsymbol{y}^0\| \leq b$ *und alle* $x \in [x_0 - \alpha, x_0 + \alpha]$*, für die*

$$\|\boldsymbol{y}(x; x_0, \boldsymbol{z}^0) - \boldsymbol{y}^0\| \leq b$$

gilt.

Wir betrachten nun Anfangswertaufgaben, bei denen die rechte Seite von einem Parameter $\boldsymbol{\lambda} \in \mathbb{R}^m$ abhängt:

$$\boldsymbol{y}' = f(x, \boldsymbol{y}, \boldsymbol{\lambda}), \;\; \boldsymbol{y}(x_0) = \boldsymbol{y}^0. \tag{1.3}$$

Hierfür gilt (vgl. Satz 28.9 MI)

Satz 1.7. *Die Funktion* f *besitze auf der Menge*

$$\tilde{Q} := \{(x, \boldsymbol{y}, \boldsymbol{\lambda}) : |x - x_0| \leq \alpha, \|\boldsymbol{y} - \boldsymbol{y}^0\| \leq b, \|\boldsymbol{\lambda} - \boldsymbol{\lambda}^0\| \leq c\}$$

stetige partielle Ableitungen erster Ordnung bzgl. der Komponenten von $\boldsymbol{y}$ *und* $\boldsymbol{\lambda}$.

Dann ist $\boldsymbol{y}(x;\boldsymbol{\lambda})$ *stetig differenzierbar auf* M. *Darüberhinaus existieren alle gemischten zweiten partiellen Ableitungen bzgl.* x *und der Komponenten von* $\boldsymbol{\lambda}$, *und diese sind stetig.*

Die Matrixfunktion $\boldsymbol{Z}(x;\boldsymbol{\lambda}) := \dfrac{\partial}{\partial\boldsymbol{\lambda}}\boldsymbol{y}(x;\boldsymbol{\lambda})$ *ist Lösung der Anfangswertaufgabe*

$$\begin{aligned} \boldsymbol{Z}'(x;\boldsymbol{\lambda}) &= \frac{\partial}{\partial\boldsymbol{y}} f(x,\boldsymbol{y}(x;\boldsymbol{\lambda}),\boldsymbol{\lambda})\boldsymbol{Z}(x;\boldsymbol{\lambda}) + \frac{\partial}{\partial\boldsymbol{\lambda}} f(x,\boldsymbol{y}(x;\boldsymbol{\lambda}),\boldsymbol{\lambda}), \\ \boldsymbol{Z}(x_0;\boldsymbol{\lambda}) &= \boldsymbol{0}. \end{aligned} \tag{1.4}$$

Auf Satz 1.7. kann man die Frage nach der Abhängigkeit der Lösung einer Anfangswertaufgabe von den Anfangswerten zurückführen. Man erhält hiermit

Korollar 1.8. *Ist die Funktion* $f : Q \to \mathbb{R}^n$ *stetig differenzierbar, so hängt die Lösung* $\boldsymbol{y}(x;x_0,\boldsymbol{y}^0)$ *der Anfangswertaufgabe*

$$\boldsymbol{y}' = f(x,\boldsymbol{y}),\ \ \boldsymbol{y}(x_0) = \boldsymbol{y}^0$$

stetig differenzierbar von x_0 *und* $\boldsymbol{y}^0$ *ab.*

Die Matrixfunktion $\boldsymbol{Z}(x) := \dfrac{\partial}{\partial\boldsymbol{y}^0}\,\boldsymbol{y}(x;x_0,\boldsymbol{y}^0)$ *ist Lösung der Anfangswertaufgabe*

$$\boldsymbol{Z}'(x) = \frac{\partial}{\partial\boldsymbol{y}} f(x,\boldsymbol{y}(x;x_0,\boldsymbol{y}^0))\,\boldsymbol{Z}(x),\ \ \boldsymbol{Z}(x_0) = \boldsymbol{E}, \tag{1.5}$$

wobei $\boldsymbol{E} \in \mathbb{R}^{(n,n)}$ *die Einheitsmatrix bezeichnet.*

Die Funktion $\boldsymbol{w}(x) := \dfrac{\partial}{\partial x_0}\,\boldsymbol{y}(x;x_0,\boldsymbol{y}^0)$ *ist Lösung der Anfangswertaufgabe*

$$\boldsymbol{w}'(x) = \frac{\partial}{\partial\boldsymbol{y}} f(x,\boldsymbol{y}(x;x_0,\boldsymbol{y}^0))\,\boldsymbol{w}(x),\ \ \boldsymbol{w}(x_0) = -f(x_0,\boldsymbol{y}^0). \tag{1.6}$$

Bemerkung 1.9. Satz 1.7. und Korollar 1.8. gelten entsprechend für höhere Ableitungen. Ist z.B. $f : Q \to \mathbb{R}^n$ eine C^m-Funktion, so ist auch $\boldsymbol{y}(x;x_0,\boldsymbol{y}^0)$ eine C^m-Funktion aller Variablen. □

Wir betrachten nun lineare Systeme

$$\boldsymbol{y}' = \boldsymbol{A}(x)\,\boldsymbol{y} + \boldsymbol{b}(x) \tag{1.7}$$

Hierfür existiert die Lösung einer Anfangswertaufgabe auf dem ganzen Intervall, auf dem $\boldsymbol{A}$ und $\boldsymbol{b}$ stetig sind. Es gilt

Satz 1.10. *Die lineare Anfangswertaufgabe*

$$\boldsymbol{y}' = \boldsymbol{A}(x)\,\boldsymbol{y} + \boldsymbol{b}(x), \;\; \boldsymbol{y}(x_0) = \boldsymbol{y}^0$$

mit stetigen Funktionen $\boldsymbol{A} : [a,b] \to \mathbb{R}^{(n,n)}$, $\boldsymbol{b} : [a,b] \to \mathbb{R}^n$ *bzw.* $\boldsymbol{A} : \mathbb{R} \to \mathbb{R}^{(n,n)}$, $\boldsymbol{b} : \mathbb{R} \to \mathbb{R}^n$ *besitzt eine eindeutige Lösung, die auf ganz* $[a,b]$ *bzw.* $\mathbb{R}$ *definiert ist.*

Wie für lineare Gleichungssysteme gilt

Satz 1.11. *Die allgemeine Lösung von (1.7) lautet*

$$\boldsymbol{y}(x) = \boldsymbol{y}_s(x) + \boldsymbol{y}_h(x).$$

Dabei ist $\boldsymbol{y}_s$ *eine spezielle Lösung von (1.7) und* $\boldsymbol{y}_h(x)$ *die allgemeine Lösung des zu (1.7) gehörenden homogenen Differentialgleichungssystems*

$$\boldsymbol{y}' = \boldsymbol{A}(x)\,\boldsymbol{y}. \tag{1.8}$$

Die Lösungen des homogenen Systems $\boldsymbol{y}' = \boldsymbol{A}(x)\,\boldsymbol{y}$ bilden offenbar einen Vektorraum. Eine Basis dieses Vektorraums kann man auf folgende Weise bestimmen:

> *Wir wählen ein* $x_0 \in \mathbb{R}$ *und eine Basis* $\boldsymbol{v}^1, \ldots, \boldsymbol{v}^n$ *des* $\mathbb{R}^n$. *Dann besitzt jede der Anfangswertaufgaben*
>
> $$\boldsymbol{y}' = \boldsymbol{A}(x)\,\boldsymbol{y}, \;\; \boldsymbol{y}(x_0) = \boldsymbol{v}^j, \;\; j = 1, \ldots, n,$$
>
> *eine eindeutige Lösung* $\boldsymbol{y}^j(x)$.

Definition 1.12. *Ist* $\boldsymbol{y}^1, \ldots, \boldsymbol{y}^n$ *eine beliebige Basis des Lösungsraums von (1.8), so heißt* $\boldsymbol{Y}(x) := (\boldsymbol{y}^1(x), \ldots, \boldsymbol{y}^n(x))$ *ein* **Fundamentalsystem** *oder eine* **Fundamentallösung** *von (1.2).*

Satz 1.13. *Es seien* $\boldsymbol{y}^1, \ldots, \boldsymbol{y}^n$ *n Lösungen des homogenen Systems* $\boldsymbol{y}' = \boldsymbol{A}(x)\,\boldsymbol{y}$ *und* $\boldsymbol{Y}(x) := (\boldsymbol{y}^1(x), \ldots, \boldsymbol{y}^n(x))$.

Dann gilt

(i) Ist $\boldsymbol{Y}$ *ein Fundamentalsystem von (1.8), so ist die allgemeine Lösung von (1.8) gegeben durch* $\boldsymbol{y}(x) = \boldsymbol{Y}(x)\,\boldsymbol{\alpha}$, $\boldsymbol{\alpha} \in \mathbb{R}^n$.

(ii) $\boldsymbol{Y}$ *ist genau dann ein Fundamentalsystem von (1.8), wenn für ein* $x_0 \in \mathbb{R}$ *die Matrix* $\boldsymbol{Y}(x_0)$ *regulär ist.*

(iii) Ist $\boldsymbol{Y}(x_0)$ *regulär für ein* $x_0 \in \mathbb{R}$, *so ist* $\boldsymbol{Y}(x)$ *regulär für alle* $x \in \mathbb{R}$.

Ist eine Fundamentallösung $\boldsymbol{Y}$ des homogenen Systems bekannt, so kann man die Lösung des inhomogenen Problems durch Variation der Konstanten ermitteln.

Satz 1.14. *Es sei $\boldsymbol{Y}(x)$ ein beliebiges Fundamentalsystem des homogenen Problems $\boldsymbol{y}' = \boldsymbol{A}(x)\,\boldsymbol{y}$.*

Dann ist

$$\boldsymbol{y}(x) = \boldsymbol{Y}(x)\Big(\boldsymbol{Y}^{-1}(x_0)\,\boldsymbol{y}^0 + \int_{x_0}^{x} \boldsymbol{Y}^{-1}(t)\,\boldsymbol{b}(t)\,dt\Big) \tag{1.9}$$

die eindeutige Lösung der Anfangswertaufgabe

$$\boldsymbol{y}' = \boldsymbol{A}(x)\,\boldsymbol{y} + \boldsymbol{b}(x),\ \ \boldsymbol{y}(x_0) = \boldsymbol{y}^0.$$

Ist (1.7) ein lineares System mit konstanten Koeffizienten:

$$\boldsymbol{y}' = \boldsymbol{A}\,\boldsymbol{y},\ \ \boldsymbol{A} \in \mathbb{R}^{(n,n)}, \tag{1.10}$$

so kann man ein Fundamentalsystem mit Methoden der linearen Algebra bestimmen.

Besitzt $\boldsymbol{A}$ die Jordansche Normalform

$$\boldsymbol{A} = \boldsymbol{V}\boldsymbol{J}\boldsymbol{V}^{-1}$$

mit

$$\boldsymbol{J} = \begin{pmatrix} \boldsymbol{J}_1 & & \boldsymbol{0} \\ & \ddots & \\ \boldsymbol{0} & & \boldsymbol{J}_m \end{pmatrix},\ \ \boldsymbol{J}_j = \begin{pmatrix} \lambda_j & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda_j \end{pmatrix},\ \ j = 1, \ldots, m,$$

und sind $\boldsymbol{v}^1, \ldots, \boldsymbol{v}^k$ die zu dem Jordan Kästchen $\boldsymbol{J}_j$ gehörenden Spalten von $\boldsymbol{V}$, so sind

$$\boldsymbol{v}^1\,\exp(\lambda x),\ (x\boldsymbol{v}^1 + \boldsymbol{v}^2)\,\exp(\lambda x),\ \Big(\frac{1}{2}\,x^2\boldsymbol{v}^1 + x\boldsymbol{v}^2 + \boldsymbol{v}^3\Big)\,\exp(\lambda x),$$

$$\ldots,\ \Big(\frac{1}{(k-1)!}\,x^{k-1}\boldsymbol{v}^1 + \cdots + x\boldsymbol{v}^{k-1} + \boldsymbol{v}^k\Big)\,\exp(\lambda x)$$

linear unabhängige Lösungen von (1.8). Fasst man diese Lösungen zu den verschiedenen Kästchen zusammen, so erhält man insgesamt ein Fundamentalsystem von (1.8).

Wir geben noch eine andere Gestalt einer Fundamentallösung an. Dazu definieren wir zunächst für $\boldsymbol{A} \in \mathbb{C}^{(n,n)}$

$$e^{\boldsymbol{A}} := \sum_{j=0}^{\infty} \frac{1}{j!}\,\boldsymbol{A}^j.$$

Dann ist

$$\boldsymbol{Y}(x) := e^{x\boldsymbol{A}}$$

die durch $\boldsymbol{Y}(0) = \boldsymbol{E}$ normierte Fundamentalmatrix und

$$\boldsymbol{y}(x) = e^{\boldsymbol{A}x}\Big\{\boldsymbol{y}^0 + \int_0^x e^{-\boldsymbol{A}t}\,\boldsymbol{b}(t)\,dt\Big\} \tag{1.11}$$

ist die Lösung der Anfangswertaufgabe

$$\boldsymbol{y}' = \boldsymbol{A}\boldsymbol{y} + \boldsymbol{b},\ \boldsymbol{y}(0) = \boldsymbol{y}^0.$$

Man beachte, dass diese Gestalt der Fundamentallösung und der Lösungsformel niemals zur praktischen Berechnung, sondern nur für qualitative Überlegungen verwendet werden.

Wir betrachten nun die lineare (2-Punkt) Randwertaufgabe

$$\left.\begin{aligned} L\boldsymbol{y}(x) :=&\ \ \boldsymbol{y}'(x) - \boldsymbol{C}(x)\,\boldsymbol{y}(x) = \boldsymbol{r}(x) \\ R\boldsymbol{y} :=&\ \ \boldsymbol{A}\boldsymbol{y}(a) + \boldsymbol{B}\boldsymbol{y}(b) = \boldsymbol{c} \end{aligned}\right\} \tag{1.12}$$

wobei $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{(n,n)}$, $\boldsymbol{c} \in \mathbb{R}^n$ und stetige Funktionen $\boldsymbol{C} : [a,b] \to \mathbb{R}^{(n,n)}$ und $\boldsymbol{r} : [a,b] \to \mathbb{R}^n$ gegeben sind. Hierfür gilt

Satz 1.15. *Gegeben sei die lineare Randwertaufgabe (1.12). Es sei $\boldsymbol{Y}(x)$ ein Fundamentalsystem von $\boldsymbol{y}' = \boldsymbol{C}(x)\,\boldsymbol{y}$. Dann sind äquivalent:*

(i) Die Randwertaufgabe hat für jede stetige rechte Seite $\boldsymbol{r}(x)$ und jeden Vektor $\boldsymbol{c} \in \mathbb{R}^n$ eine eindeutig bestimmte Lösung.

(ii) Die homogene Randwertaufgabe

$$\boldsymbol{y}' = \boldsymbol{C}(x)\,\boldsymbol{y},\ \ \boldsymbol{A}\boldsymbol{y}(a) + \boldsymbol{B}\boldsymbol{y}(b) = \boldsymbol{0}$$

hat nur die triviale Lösung $\boldsymbol{y}(x) \equiv \boldsymbol{0}$.

(iii) Die Matrix $\boldsymbol{D} := \boldsymbol{A}\boldsymbol{Y}(a) + \boldsymbol{B}\boldsymbol{Y}(b)$ ist regulär.

Für die Randwertaufgabe (1.12) kann man eine geschlossene Lösungsformel angeben, die aber (ähnlich wie die Lösungsformel (1.11) für Anfangswertaufgaben mit Hilfe der Fundamentalmatrix $e^{x\boldsymbol{A}}$) nur für theoretische Zwecke verwendet wird.

Die homogene Randwertaufgabe

$$L\boldsymbol{y}(x) = \boldsymbol{0}, \quad R\boldsymbol{y} = \boldsymbol{0},$$

besitze nur die triviale Lösung $\boldsymbol{y}(x) \equiv \boldsymbol{0}$, und es sei $\boldsymbol{Y}(x)$ die Fundamentallösung von

$$\boldsymbol{y}' = \boldsymbol{C}(x)\,\boldsymbol{y}$$

mit $\boldsymbol{Y}(a) = \boldsymbol{E}$.

Nach Satz 1.15. ist die Matrix $\boldsymbol{D} := \boldsymbol{A} + \boldsymbol{B}\,\boldsymbol{Y}(b)$ regulär, und daher ist die Matrix

$$\boldsymbol{G}(x,t) = -\boldsymbol{Y}(x) \left\{ \begin{array}{ll} (\boldsymbol{A} + \boldsymbol{B}\,\boldsymbol{Y}(b))^{-1}\,\boldsymbol{B}\,\boldsymbol{Y}(b) - \boldsymbol{E} & \\ (\boldsymbol{A} + \boldsymbol{B}\,\boldsymbol{Y}(b))^{-1}\,\boldsymbol{B}\,\boldsymbol{Y}(b) & \end{array} \right\} \boldsymbol{Y}(t)^{-1} \begin{array}{l} , \, t < x \\ , \, t > x \end{array} \tag{1.13}$$

für alle $x \in [a,b]$ und alle $t \in [a,b]$ definiert.

Definition 1.16. *Die Matrix $\boldsymbol{G}(x,t)$ aus (1.13) heißt die* **Greensche Matrix** *der linearen Randwertaufgabe (1.12).*

Satz 1.17. *Es sei die Randwertaufgabe (1.12) eindeutig lösbar und $\boldsymbol{Y}(x)$ Fundamentallösung von $\boldsymbol{y}' = \boldsymbol{C}(x)\,\boldsymbol{y}$ mit $\boldsymbol{Y}(a) = \boldsymbol{E}$. Dann ist mit der Greenschen Matrix $\boldsymbol{G}(x,t)$ aus (1.13) die Lösung von (1.12) darstellbar als*

$$\boldsymbol{y}(x) = \boldsymbol{Y}(x)\,(\boldsymbol{A} + \boldsymbol{B}\,\boldsymbol{Y}(b))^{-1}\,\boldsymbol{c} + \int_a^b \boldsymbol{G}(x,t)\,\boldsymbol{r}(t)\,dt.$$

Für eine große Klasse linearer Differentialgleichungen zweiter Ordnung erhält man die Existenz und Eindeutigkeit der Lösungen aus dem folgenden Satz.

Satz 1.18. *Es sei*

$$Ly := -y'' + p(x)y' + q(x)y, \;\; a < x < b, \tag{1.14}$$

mit $p, q \in C[a,b]$ und $q(x) \geq 0$ für alle $x \in [a,b]$.

Dann gilt

$$Ly(x) \geq 0 \text{ für alle } x \in [a,b], \;\; y(a) \geq 0, \;\; y(b) \geq 0 \;\Rightarrow\; y(x) \geq 0 \text{ für alle } x \in [a,b]. \tag{1.15}$$

Beweis: Angenommen y nimmt in $s \in [a,b]$ ein negatives Minimum an:

$$y(s) = \min_{a \le x \le b} y(x) < 0.$$

Dann gilt wegen $y(a) \ge 0$ und $y(b) \ge 0$ sogar $a < s < b$, und

$$y(s) < 0,\ y'(s) = 0,\ y''(s) \ge 0.$$

Wegen der Stetigkeit von y gibt es eine Umgebung $U \subset [a,b]$ von s mit $y(x) < 0$ für alle $x \in U$. Es gilt

$$-y''(x) + p(x)y'(x) \ge -q(x)y(x) \ge 0 \quad \text{für alle } x \in U.$$

Durch Multiplikation dieser Ungleichung mit

$$\rho(x) := \exp\Big(-\int_s^x p(t)\,dt\Big)$$

erhält man

$$-(\rho(x)y'(x))' \ge 0 \quad \text{für alle } x \in U,$$

und hieraus durch Integration unter Berücksichtigung von $y'(s) = 0$

$$y'(x) \le 0 \text{ für } x \in U,\ x \ge s \quad \text{und} \quad y'(x) \ge 0 \text{ für } x \in U,\ x \le s.$$

Da die Funktion y in s ihr Minimum annimmt, muss y in U – und mit demselben Schluss in ganz $[a,b]$ – konstant gleich $y(s) < 0$ sein im Widerspruch zu $y(a) \ge 0$.

∎

Ein Randwertproblem mit der Eigenschaft (1.15) heißt **invers monoton**.

Aus der Inversmonotonie folgt insbesondere, dass die Randwertaufgabe

$$Ly(x) = f(x),\ a < x < b, \quad y(a) = \gamma_1,\ y(b) = \gamma_2, \tag{1.16}$$

für jede stetige rechte Seite f und für alle $\gamma_1, \gamma_2 \in \mathbb{R}$ eine eindeutige Lösung besitzt, denn das homogene Problem ist nur trivial lösbar. Gilt nämlich $Ly(x) \equiv 0$, $y(a) = 0$, $y(b) = 0$, so folgt aus Satz 1.18. $y(x) \ge 0$ für $x \in [a,b]$, und da zugleich $L(-y) \ge 0$, $(-y)(a) \ge 0$, $(-y)(b) \ge 0$ gilt, ist auch $-y(x) \ge 0$ für $x \in [a,b]$. Zusammen folgt also $y(x) \equiv 0$, und wegen Satz 1.15. die eindeutige Lösbarkeit von (1.16).

Aus Satz 1.17. folgt, dass die Greensche Funktion von L mit Dirichletschen Randbedingungen nichtnegativ ist, denn existiert ein $(\tilde{x}, \tilde{t}) \in (a,b) \times (a,b)$ mit $g(x,t) < 0$, so gibt es wegen der Stetigkeit von g ein $\varepsilon > 0$ mit $g(x,t) < 0$ für alle $(x,t) \in$

$(\tilde{x}-\varepsilon,\tilde{x}+\varepsilon)\times(\tilde{t}-\varepsilon,\tilde{t}+\varepsilon)$. Es sei $f\in C[a,b]$ mit $f(x)=0$ für $|x-\tilde{t}|\geq\varepsilon$ und $f(x)>0$ für $|x-\tilde{t}|<\varepsilon$ und y die Lösung von

$$Ly(x)=f(x),\ y(a)=y(b)=0.$$

Dann gilt

$$y(\tilde{x})=\int_a^b g(\tilde{x},t)f(t)\,dt=\int_{\tilde{t}-\varepsilon}^{\tilde{t}+\varepsilon} g(\tilde{x},t)f(t)\,dt<0,$$

während aus der Inversmonotonie $y(x)\geq 0$ für alle $x\in[a,b]$ folgt.

1.2 Partielle Differentialgleichungen

Wir betrachten in diesem Abschnitt lineare partielle Differentialgleichungen zweiter Ordnung

$$Lu(\boldsymbol{x}):=\sum_{j,k=1}^{n} a_{jk}(\boldsymbol{x})\,u_{x_jx_k}(\boldsymbol{x})+\sum_{j=1}^{n} b_j(x)\,u_{x_j}(\boldsymbol{x})+c(\boldsymbol{x})\,u(\boldsymbol{x})=f(\boldsymbol{x}). \tag{1.17}$$

Dabei sind $a_{jk},b_j,c,f:\mathbb{R}^n\supset\Omega\to\mathbb{R}$, $j,k=1,\ldots,n$, gegebene stetige Funktionen und Ω ein Gebiet im $\mathbb{R}^n$. Wir beschränken uns meistens auf den Fall eines ebenen Systems ($n=2$). Die Matrix $\boldsymbol{A}:=(a_{jk})$ nehmen wir ohne Beschränkung der Allgemeinheit als symmetrisch an.

Eine Funktion $u\in C^2(\Omega)$ heißt eine **klassische Lösung** der Differentialgleichung (1.17), wenn die Gleichung (1.17) in jedem Punkt aus Ω erfüllt ist.

Wir unterscheiden drei Typen, elliptische, parabolische und hyperbolische Differentialgleichungen. Für verschiedene Typen sind verschiedene Aufgabentypen sachgemäß (d.h. eindeutig lösbar, wobei die Lösung stetig von den Eingangsdaten abhängt) und physikalisch sinnvoll. Die Theorie und die numerische Behandlung sind bei den verschiedenen Typen sehr unterschiedlich.

Definition 1.19. *Besitzt die Matrix $\boldsymbol{A}(\boldsymbol{x})$ Eigenwerte einheitlichen Vorzeichens, so heißt (1.17)* **elliptisch** *im Punkt $\boldsymbol{x}$, ist $\boldsymbol{A}(\boldsymbol{x})$ regulär und hat ein Eigenwert von $\boldsymbol{A}(\boldsymbol{x})$ ein anderes Vorzeichen als die übrigen $n-1$ Eigenwerte, so heißt die Gleichung (1.17)* **hyperbolisch** *in $\boldsymbol{x}$, ist schließlich die Matrix $\boldsymbol{A}(\boldsymbol{x})$ singulär, so heißt die Differentialgleichung (1.17)* **parabolisch** *in $\boldsymbol{x}$.*

Ist die Gleichung in allen Punkten von Ω elliptisch oder hyperbolisch oder parabolisch, so nennt man sie elliptisch oder hyperbolisch oder parabolisch in Ω.

Für $n \geq 4$ betrachtet man zusätzlich noch **ultrahyperbolische** *Probleme. Dies sind Aufaben, bei denen alle Eigenwerte von $\boldsymbol{A}$ von Null verschieden sind und es wenigstens zwei positive und zwei negative Eigenwerte gibt. Wir werden hierauf nicht weiter eingehen.*

Man kann eine lineare Differentialgleichung zweiter Ordnung stets auf eine der folgenden Normalformen transformieren.

Ist die Differentialgleichung elliptisch, so lautet die Normalform

$$\Delta_n u + \boldsymbol{b}^T \nabla u + cu = f \tag{1.18}$$

mit dem $n-$dimensionalen Laplace Operator

$$\Delta_n := \sum_{i=1}^{n} \frac{\partial^2}{\partial x_i^2}$$

und gewissen Funktionen b_i, c und f.

Speziell für $b_i \equiv 0$, $c \equiv 0$ heißt Gleichung (1.18) **Poisson Gleichung**. Gilt zusätzlich $f \equiv 0$, so erhält man die **Potentialgleichung**.

Die Normalform der hyperbolischen Aufgabe lautet

$$u_{tt} = \Delta_{n-1} u + \boldsymbol{b}^T \nabla_{n-1} u + b_n \frac{\partial u}{\partial t} + cu + f, \tag{1.19}$$

wobei

$$\Delta_{n-1} := \sum_{i=1}^{n-1} \frac{\partial^2}{\partial x_i^2}$$

und

$$\nabla_{n-1} := \Big(\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_{n-1}} \Big)^T$$

sich nur auf die (Orts-) Variable $\boldsymbol{x}$ beziehen.

Speziell für $b_i \equiv 0,\ i = 1, \ldots, n$, $c \equiv 0$ und $f \equiv 0$ erhält man die **Wellengleichung**

$$u_{tt} = \Delta_{n-1} u.$$

Die Normalform der parabolischen Aufgabe lautet

$$u_t = \Delta_{n-1} u + \boldsymbol{b}^T \nabla_{n-1} u + cu + f. \tag{1.20}$$

Ein typischer Vertreter ist die **Wärmeleitungsgleichung**

$$u_t = \Delta_{n-1} u,$$

wobei $n-1$ wieder die Raumdimension bezeichnet.

Etwas allgemeinere Klassen von Differentialgleichungen, bei denen die obige Klassifizierung ebenfalls verwendet wird, sind die halblinearen und die quasilinearen Differentialgleichungen.

Definition 1.20. *Die Differentialgleichung*

$$\sum_{i,j=1}^{n} a_{ij}(\boldsymbol{x}) \frac{\partial^2}{\partial x_i \partial x_j} u = f(\boldsymbol{x}, u, \nabla u),$$

in der die Funktion u und ihre ersten Ableitungen auch nichtlinear auftreten können, der Hauptteil aber nur von den unabhängigen Variablen $x_1, \ldots, x_n$ abhängt, heißt **halblinear** .

Die Differentialgleichung

$$\sum_{i,j=1}^{n} a_{ij}(\boldsymbol{x}, u, \nabla u) \frac{\partial^2}{\partial x_i \partial x_j} u = f(\boldsymbol{x}, u, \nabla u),$$

die linear in den zweiten Ableitungen ist, wobei der Hauptteil auch von u und/oder den ersten Ableitungen von u abhängt, heißt **quasilinear**.

Wie bei den gewöhnlichen Differentialgleichungen kann man nur dann eine eindeutige Lösung einer Differentialgleichung erwarten, wenn man zusätzlich Anfangsbedingungen oder Randbedingungen vorgibt.

Im linearen (und halblinearen) Fall hängt der Typ der Differentialgleichung nur von dem betrachteten Punkt $\boldsymbol{x} \in \Omega$ ab. Im quasilinearen Fall kann der Typ der Differentialgleichung nicht nur von $\boldsymbol{x}$, sondern auch von der Lösung u (also von den Randwerten oder Anfangswerten) abhängen.

Definition 1.21. *Eine Differentialgleichung*

$$Lu = f$$

mit zusätzlichen Anfangs- und/oder Randbedingungen heißt **sachgemäß**, *wenn sie eindeutig lösbar ist und die Lösung stetig von den Eingangsdaten abhängt.*

Die Forderung der stetigen Abhängigkeit (auch Stabilität) ist bei physikalischen Problemen sinnvoll, da Eingangsdaten häufig aus Messungen gewonnen werden, d.h. nur

bis auf eine gewisse Genauigkeit bekannt sind. Werden diese Eingabedaten in gewissen Grenzen variiert, so sollte sich die Lösung des Problems nicht zu dramatisch ändern, da sonst die Lösung wertlos ist.

Durch elliptische Differentialgleichungen werden in der Regel Gleichgewichtszustände beschrieben (z.B. stationäre Temperaturverteilungen in einem Körper), durch hyperbolische oder parabolische Gleichungen zeitabhängige Probleme (z.B. Ausbreitung von Wellen oder die zeitliche Entwicklung einer Temperaturverteilung). Es ist daher anschaulich klar, dass man für elliptische Probleme zusätzlich Randbedingungen vorzugeben hat und für die beiden anderen Typen Anfangs- und Randbedingungen.

1.2.1 Elliptische Probleme

Wir betrachten zunächst die elliptische Differentialgleichung

$$Lu(\boldsymbol{x}) := -\sum_{j,k=1}^{n} a_{jk}(\boldsymbol{x})\, u_{x_j x_k}(\boldsymbol{x}) + \sum_{j=1}^{n} b_j(\boldsymbol{x})\, u_{x_j}(\boldsymbol{x}) + c(\boldsymbol{x})\, u(\boldsymbol{x}) = f(\boldsymbol{x}), \quad \boldsymbol{x} \in \Omega \subset \mathbb{R}^n \tag{1.21}$$

und setzen voraus, dass sogar ein $\alpha_0 > 0$ existiert mit

$$\sum_{j,k=1}^{n} a_{jk}(\boldsymbol{x})\xi_j\xi_k \geq \alpha_0 \sum_{j=1}^{n} \xi_j^2 \quad \text{für alle } \boldsymbol{x} \in \Omega \text{ und alle } \xi_j \in \mathbb{R}. \tag{1.22}$$

In diesem Fall sind nach dem Rayleighschen Prinzip sogar alle Eigenwerte der Matrix $\boldsymbol{A}(\boldsymbol{x})$ für jedes $\boldsymbol{x} \in \Omega$ größer oder gleich α_0, also von einem Vorzeichen, und L ist elliptisch. Die Differentialgleichung (1.21) heißt dann **gleichmäßig elliptisch** .

Wir stellen uns vor, dass durch die Gleichung (1.21) die stationäre Temperaturverteilung in einem Körper Ω beschrieben wird. Dann sind die folgenden Randvorgaben physikalisch sinnvoll:

$$u(\boldsymbol{x}) = g(\boldsymbol{x}), \quad \boldsymbol{x} \in \partial\Omega. \tag{1.23}$$

mit einer gegebenen Funktion $g : \partial\Omega \to \mathbb{R}$. Dies bedeutet, dass der Wärmeaustausch zwischen dem Körper und seiner Umgebung so perfekt ist, dass die Oberflächentemperatur des Körpers gleich der gegebenen Umgebungstemperatur $g(\boldsymbol{x})$ ist. Die Randwertaufgabe (1.21), (1.23) heißt **erste Randwertaufgabe** oder **Dirichletsche Randwertaufgabe** .

Daneben betrachtet man die Bedingung

$$\frac{\partial}{\partial \boldsymbol{n}} u(\boldsymbol{x}) = g(\boldsymbol{x}), \quad \boldsymbol{x} \in \partial\Omega \tag{1.24}$$

mit gegebenem $g : \partial\Omega \to \mathbb{R}$, wobei $\boldsymbol{n}$ den äußeren Normalenvektor auf $\partial\Omega$ bezeichnet. Hier wird der Wärmefluss durch die Oberfläche des Körpers vorgegeben. Speziell für $g \equiv 0$ ist also die Oberfläche perfekt isoliert. Die Randwertaufgabe (1.21), (1.24) heißt **zweite Randwertaufgabe** oder **Neumannsche Randwertaufgabe** .

Realistisch ist für das stationäre Wärmeleitungsproblem die Randbedingung

$$a(\boldsymbol{x})u(\boldsymbol{x}) + b(\boldsymbol{x})\frac{\partial}{\partial \boldsymbol{n}}u(\boldsymbol{x}) = g(\boldsymbol{x}), \quad \boldsymbol{x} \in \partial\Omega \tag{1.25}$$

mit gegebenen Funktionen $a, b, g : \partial\Omega \to \mathbb{R}$. Hierdurch wird z.B. beschrieben, dass der Wärmefluss durch die Oberfläche proportional zur Differenz der Umgebungstemperatur u_0 und der Temperatur des Körpers am Rande von Ω ist:

$$\frac{\partial}{\partial \boldsymbol{n}}u(\boldsymbol{x}) = \alpha(u_0(\boldsymbol{x}) - u(\boldsymbol{x})).$$

(1.21), (1.25) heißt **dritte Randwertaufgabe** oder **Robinsche Randwertaufgabe**.

Eindeutigkeitsresultate und die stetige Abhängigkeit für elliptische Randwertaufgaben erhält man leicht aus der Inversmonotonie des Differentialoperators L. Es gilt

Satz 1.22. *Es sei $\Omega \subset \mathbb{R}^n$ ein beschränktes Gebiet, und es sei L gleichmäßig elliptisch mit $c \geq 0$. Für die Funktionen $v, w \in C^2(\Omega) \cap C(\bar{\Omega})$ seien die Ungleichungen*

$$Lv(\boldsymbol{x}) \leq Lw(\boldsymbol{x}) \quad \textit{für alle } \boldsymbol{x} \in \Omega, \tag{1.26}$$

$$v(\boldsymbol{x}) \leq w(\boldsymbol{x}) \quad \textit{für alle } \boldsymbol{x} \in \partial\Omega, \tag{1.27}$$

erfüllt. Dann gilt

$$v(\boldsymbol{x}) \leq w(\boldsymbol{x}) \quad \textit{für alle } \boldsymbol{x} \in \Omega. \tag{1.28}$$

Den Beweis findet man in *Protter, Weinberger* [39]. Als Spezialfall erhält man

Korollar 1.23. (schwaches Maximumprinzip)
Es seien die Voraussetzungen von Satz 1.22. erfüllt, und es gelte

$$c(\boldsymbol{x}) \geq 0 \quad \textit{für alle } \boldsymbol{x} \in \Omega.$$

Für $u \in C(\bar{\Omega}) \cap C^2(\Omega)$ gelte

$$Lu(\boldsymbol{x}) \geq 0 \ (\textit{bzw. } Lu(\boldsymbol{x}) \leq 0) \quad \textit{für } \boldsymbol{x} \in \Omega$$

Ist das Minimum von u auf $\bar{\Omega}$ negativ (bzw. das Maximum positiv), so wird es auf dem Rand $\partial\Omega$ angenommen.

Beweis: Das Minimum von u auf $\partial\Omega$ ist negativ, denn gilt $u(\boldsymbol{x}) \geq 0$ für alle $\boldsymbol{x} \in \partial\Omega$, so folgt mit $Lu(\boldsymbol{x}) \geq 0$ in Ω, dass $u(\boldsymbol{x}) \geq 0$ für alle $\boldsymbol{x} \in \overline{\Omega}$ gilt.

Es sei

$$u_0 := \min_{\boldsymbol{x}\in\partial\Omega} u(\boldsymbol{x}) < 0$$

und $v(\boldsymbol{x}) := u_0$ für alle $\boldsymbol{x} \in \bar{\Omega}$. Dann gilt

$$Lv(\boldsymbol{x}) = c(\boldsymbol{x})u_0 \leq 0 \leq Lu(\boldsymbol{x}) \quad \text{für alle } \boldsymbol{x} \in \Omega$$

und

$$v(\boldsymbol{x}) \leq u(\boldsymbol{x}) \quad \text{für alle } \boldsymbol{x} \in \partial\Omega.$$

Daher folgt aus Satz 1.22.

$$u_0 \leq u(\boldsymbol{x}) \quad \text{für alle } \boldsymbol{x} \in \bar{\Omega},$$

und die Funktion u nimmt ihr Minimum auf $\partial\Omega$ an.

Das positive Maximum behandelt man im Falle $Lu(\boldsymbol{x}) \leq 0$ genauso. ∎

Bemerkung 1.24. Der Beweis zeigt, dass man im Falle $c \equiv 0$ auf die Vorzeichenvoraussetzung für das Extremum verzichten kann. Insbesondere nimmt also eine subharmonische Funktion ihr Maximum und eine superharmonische ihr Minimum auf dem Rand $\partial\Omega$ an, und eine harmonische Funktion nimmt Minimum und Maximum auf $\partial\Omega$ an. □

Als erste Folgerung aus dem Maximumprinzip erhält man ein Eindeutigkeitsresultat.

Korollar 1.25. *Sei $\Omega \subset \mathbb{R}^n$ offen und beschränkt, und es sei $c(\boldsymbol{x}) \geq 0$ für alle $\boldsymbol{x} \in \Omega$. Dann besitzt die Dirichletsche Randwertaufgabe*

$$Lu(\boldsymbol{x}) = f(\boldsymbol{x}), \quad \boldsymbol{x} \in \Omega, \quad u(\boldsymbol{x}) = g(\boldsymbol{x}), \quad \boldsymbol{x} \in \partial\Omega$$

höchstens eine Lösung.

Beweis: Sind $u_1, u_2 \in C^2(\Omega) \cap C(\bar{\Omega})$ zwei Lösungen, so erfüllt $v(\boldsymbol{x}) := u_1(\boldsymbol{x}) - u_2(\boldsymbol{x})$ die homogene Randwertaufgabe

$$Lv(\boldsymbol{x}) = 0, \quad \boldsymbol{x} \in \Omega, \quad v(\boldsymbol{x}) = 0, \quad \boldsymbol{x} \in \partial\Omega.$$

Angenommen v besitzt ein negatives Minimum in Ω, so wird dieses wegen Korollar 1.23. auf dem Rand $\partial\Omega$ angenommen im Widerspruch zu $v \equiv 0$ auf $\partial\Omega$. Genauso führt man die Annahme, dass u ein positives Maximum besitzt zum Widerspruch.
∎

Bemerkung 1.26. Die Voraussetzung, dass Ω beschränkt ist, ist wesentlich, denn für $\Omega := \mathbb{R} \times (0, 2\pi)$ ist $u(x, y) := e^x \sin y$ eine Lösung der homogenen Randwertaufgabe

$$\Delta u(x, y) = 0, \quad (x, y)^T \in \Omega, \quad u(x, y) = 0, \quad (x, y)^T \in \partial\Omega.$$

Die Randwertaufgabe ist also nicht eindeutig lösbar. □

Wir untersuchen nun die stetige Abhängigkeit von den Randdaten.

Korollar 1.27. *Sei $\Omega \subset \mathbb{R}^n$ offen und beschränkt, L gleichmäßig elliptisch, und es sei $c(\boldsymbol{x}) \geq 0$ für alle $\boldsymbol{x} \in \Omega$. Seien $u_i \in C^2(\Omega) \cap C(\bar{\Omega})$, $i = 1, 2$, Lösungen der Randwertaufgaben*

$$Lu(\boldsymbol{x}) = f(\boldsymbol{x}), \quad \boldsymbol{x} \in \Omega, \quad u(\boldsymbol{x}) = g_i(\boldsymbol{x}), \quad \boldsymbol{x} \in \partial\Omega.$$

Dann gilt

$$\max_{\boldsymbol{x} \in \bar{\Omega}} |u_1(\boldsymbol{x}) - u_2(\boldsymbol{x})| \leq \max_{\boldsymbol{x} \in \partial\Omega} |g_1(\boldsymbol{x}) - g_2(\boldsymbol{x})|.$$

Beweis: Mit $v(\boldsymbol{x}) := u_1(\boldsymbol{x}) - u_2(\boldsymbol{x})$ folgt die Behauptung wegen $Lv(\boldsymbol{x}) = 0$ für alle $\boldsymbol{x} \in \Omega$ und $v(\boldsymbol{x}) = g_1(\boldsymbol{x}) - g_2(\boldsymbol{x})$ für $\boldsymbol{x} \in \partial\Omega$ aus Korollar 1.23. ∎

Korollar 1.28. *Sei $\Omega \subset \mathbb{R}^n$ offen und beschränkt, und seien $u_i \in C^2(\Omega) \cap C(\bar{\Omega})$, $i = 1, 2$, Lösungen der Randwertaufgaben*

$$Lu(\boldsymbol{x}) = f_i(\boldsymbol{x}), \quad \boldsymbol{x} \in \Omega, \quad u_i(\boldsymbol{x}) = g(\boldsymbol{x}), \quad \boldsymbol{x} \in \partial\Omega.$$

Dann gilt mit einer von den Funktionen f_i und g unabhängigen Konstante C

$$\max_{\boldsymbol{x} \in \bar{\Omega}} |u_1(\boldsymbol{x}) - u_2(\boldsymbol{x})| \leq C \max_{\boldsymbol{x} \in \bar{\Omega}} |f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})|. \tag{1.29}$$

Beweis: Es sei $R > 0$ so groß gewählt, dass

$$\Omega \subset \{\boldsymbol{x} \in \mathbb{R}^n \,:\, \|\boldsymbol{x}\|_2 \leq R\}.$$

Es sei

$$|a_{jk}(\boldsymbol{x})| \leq K, \ |b_j(\boldsymbol{x})| \leq K, \ c(\boldsymbol{x}) \leq K \quad \text{für alle } \boldsymbol{x} \in \Omega \text{ und alle } j, k \in \{1, \ldots, n\}.$$

Schließlich sei $M > 0$ so gewählt, dass

$$M^2\alpha_0 - K(M + 1) \geq 1$$

gilt mit der Elliptizitätskonstante α_0 aus (1.22).

Hiermit definieren wir die Funktion

$$w(\boldsymbol{x}) := (e^{2MR} - e^{M(x_1+R)}) \max_{\boldsymbol{x}\in\Omega} |f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})|$$

und vergleichen diese mit der Lösung $v(\boldsymbol{x}) := u_1(\boldsymbol{x}) - u_2(\boldsymbol{x})$ der Randwertaufgabe

$$Lv(\boldsymbol{x}) = f_1(\boldsymbol{x}) - f_2(\boldsymbol{x}) \text{ in } \Omega, \quad v(\boldsymbol{x}) = 0 \text{ auf } \partial\Omega.$$

Es gilt

$$\begin{aligned} Lw(\boldsymbol{x}) &= \Big\{e^{2MR}\underbrace{c(\boldsymbol{x})}_{\geq 0} + (M^2\underbrace{a_{11}(\boldsymbol{x})}_{\geq\alpha_0} - M\underbrace{b_1(\boldsymbol{x})}_{\leq K} - \underbrace{c(\boldsymbol{x})}_{\leq K}\Big)e^{M(x_1+R)}\Big\}\|f_1 - f_2\|_\infty \\ &\geq \underbrace{(M^2\alpha_0 - K(M+1))}_{\geq 1}\underbrace{e^{M(x_1+R)}}_{\geq 1}\|f_1 - f_2\|_\infty \\ &\geq \|f_1 - f_2\|_\infty \geq f_1(\boldsymbol{x}) - f_2(\boldsymbol{x}) = Lv(\boldsymbol{x}). \end{aligned}$$

Ferner gilt nach Wahl von R

$$w(\boldsymbol{x}) \geq 0 = v(\boldsymbol{x}) \quad \text{für alle } \boldsymbol{x} \in \partial\Omega.$$

Daher folgt aus Satz 1.22.

$$u_1(\boldsymbol{x}) - u_2(\boldsymbol{x}) = v(\boldsymbol{x}) \leq w(\boldsymbol{x}) \leq \max_{\boldsymbol{x}\in\bar{\Omega}} |w(\boldsymbol{x})| \leq e^{2MR} \max_{\boldsymbol{x}\in\partial\Omega} |f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})|.$$

Genauso folgt $-w \leq v$, und damit (1.29). ∎

Die drei Korollare besagen, dass die Randwertaufgabe

$$Lu(\boldsymbol{x}) = f(\boldsymbol{x}), \quad \boldsymbol{x} \in \Omega, \quad u(\boldsymbol{x}) = g(\boldsymbol{x}), \quad \boldsymbol{x} \in \partial\Omega,$$

für beschränkte Gebiete $\Omega \subset \mathbb{R}^n$ fast sachgemäß ist : Es existiert höchstens eine Lösung, und diese hängt (bzgl. der Chebyshev Norm) stetig von den Funktionen f und g ab. Zu zeigen bleibt "nur noch" die Existenz der Lösung. Die Existenzfrage ist jedoch für diese Vorlesung zu schwierig. Ohne zu präzisieren, was es bedeutet, dass der Rand $\partial\Omega$ genügend glatt ist und dass die Koeffizientenfunktionen und die rechten Seite f und g hinreichend glatt sind, erwähnen wir, dass unter diesen Voraussetzungen die Existenz einer klassischen Lösung der Dirichletschen Randwertaufgabe (1.21), (1.23) gesichert werden kann. Die Präzisierungen und einen Beweis findet man in *Hackbusch* [27].

Für die zweite und dritte Randwertaufgabe schließen wir einige Bemerkungen an. Wir beschränken uns dabei auf den Laplace Operator.

Die zweite Randwertaufgabe

$$\Delta u(\boldsymbol{x}) = f(\boldsymbol{x}), \quad \boldsymbol{x} \in \Omega, \quad \frac{\partial u}{\partial \boldsymbol{n}}(\boldsymbol{x}) = g(x), \quad \boldsymbol{x} \in \partial\Omega, \tag{1.30}$$

ist sicher nicht eindeutig lösbar, denn mit $u(\boldsymbol{x})$ ist auch $\tilde{u}(\boldsymbol{x}) := u(\boldsymbol{x}) + c$ für jede Konstante $c \in \mathbb{R}$ eine Lösung.

Bis auf eine solche additive Konstante ist die Lösung eindeutig, denn sind u_1, u_2 Lösungen der Randwertaufgabe (1.30), so löst $v(\boldsymbol{x}) := u_1(\boldsymbol{x}) - u_2(\boldsymbol{x})$ die homogene Aufgabe

$$\Delta v(\boldsymbol{x}) = 0, \quad \boldsymbol{x} \in \Omega, \quad \frac{\partial v}{\partial \boldsymbol{n}}(\boldsymbol{x}) = 0, \quad \boldsymbol{x} \in \partial\Omega,$$

und aus der ersten Greenschen Formel (vgl. IM III, Satz 25.73) folgt

$$\begin{aligned} 0 &= \int\limits_{\partial\Omega} v(\boldsymbol{x}) \frac{\partial v}{\partial \boldsymbol{n}}(\boldsymbol{x})\, do = \int\limits_{\Omega} \left\{ v(\boldsymbol{x})\Delta v(\boldsymbol{x}) + \|\nabla v(\boldsymbol{x})\|_2^2 \right\} d\boldsymbol{x} \\ &= \int\limits_{\Omega} \|\nabla v(\boldsymbol{x})\|_2^2 \, d\boldsymbol{x}, \end{aligned}$$

d.h. $\nabla v(\boldsymbol{x}) \equiv 0$ in Ω, und daher $v(\boldsymbol{x}) \equiv \text{const}$.

Ferner ist die zweite Randwertaufgabe nicht für alle rechten Seiten f und alle Randvorgaben g lösbar.

Wir betrachten zunächst

$$\Delta w(\boldsymbol{x}) = 0, \quad x \in \Omega, \quad \frac{\partial w}{\partial \boldsymbol{n}}(\boldsymbol{x}) = h(\boldsymbol{x}), \quad \boldsymbol{x} \in \partial\Omega.$$

Ist $w \in C^2(\Omega) \cap C^1(\bar{\Omega})$ eine Lösung, so gilt nach der ersten Greenschen Formel (mit der Funktion $\phi(\boldsymbol{x}) \equiv 1$)

$$\begin{aligned} 0 &= \int\limits_{\Omega} \phi(\boldsymbol{x})\Delta w(\boldsymbol{x})\, d\boldsymbol{x} = -\int\limits_{\Omega} \langle \nabla\phi(\boldsymbol{x}), \nabla w(\boldsymbol{x}) \rangle \, d\boldsymbol{x} + \int\limits_{\partial\Omega} \phi(\boldsymbol{x}) \frac{\partial w}{\partial \boldsymbol{n}}(\boldsymbol{x})\, do \\ &= \int\limits_{\partial\Omega} h(\boldsymbol{x})\, do. \end{aligned}$$

Ist $u \in C^2(\Omega) \cap C^1(\bar{\Omega})$ eine Lösung von

$$\Delta u(\boldsymbol{x}) = f(\boldsymbol{x}), \quad \boldsymbol{x} \in \Omega, \qquad \frac{\partial u}{\partial \boldsymbol{n}}(\boldsymbol{x}) = g(\boldsymbol{x}), \quad \boldsymbol{x} \in \partial\Omega, \tag{1.31}$$

so wählen wir eine Funktion $v(\boldsymbol{x}) \in C^2(\Omega) \cap C^1(\bar{\Omega})$ mit $\Delta v(\boldsymbol{x}) = f(\boldsymbol{x})$, $\boldsymbol{x} \in \Omega$, und hiermit $w(\boldsymbol{x}) := u(\boldsymbol{x}) - v(\boldsymbol{x})$.

Dann gilt $\Delta w(\boldsymbol{x}) = 0$, $\boldsymbol{x} \in \Omega$, und nach dem ersten Teil folgt mit dem Gaußschen Integralsatz

$$\begin{aligned} 0 &= \int\limits_{\partial\Omega} \frac{\partial w}{\partial \boldsymbol{n}}(\boldsymbol{x})\, do = \int\limits_{\partial\Omega} \frac{\partial u}{\partial \boldsymbol{n}}(\boldsymbol{x})\, do - \int\limits_{\partial\Omega} \frac{\partial v}{\partial \boldsymbol{n}}(\boldsymbol{x})\, do \\ &= \int\limits_{\partial\Omega} g(\boldsymbol{x})\, do - \int\limits_{\Omega} \Delta v(\boldsymbol{x})\, d\boldsymbol{x} = \int\limits_{\partial\Omega} g(\boldsymbol{x})\, do - \int\limits_{\Omega} f(\boldsymbol{x})\, d\boldsymbol{x}. \end{aligned}$$

Notwendig für die Lösbarkeit von (1.30) ist also die Bedingung

$$\int\limits_{\Omega} f(\boldsymbol{x})\, d\boldsymbol{x} = \int\limits_{\partial\Omega} g(\boldsymbol{x})\, do.$$

Bemerkung 1.29. Die Lösbarkeitsbedingung ist auch physikalisch einsichtig. Beschreibt nämlich die Lösung u von

$$\Delta u(\boldsymbol{x}) = f(\boldsymbol{x}), \quad \boldsymbol{x} \in \Omega, \quad \frac{\partial u}{\partial \boldsymbol{n}}(\boldsymbol{x}) = g(\boldsymbol{x}), \quad \boldsymbol{x} \in \partial\Omega,$$

die stationäre Temperaturverteilung in einem Körper Ω, so ist $\int\limits_{\Omega} f(\boldsymbol{x})\, d\boldsymbol{x}$ die Wärmeentwicklung in dem Körper und $\int\limits_{\partial\Omega} g(\boldsymbol{x})\, do$ der Wärmefluss durch den Rand (jeweils pro Zeiteinheit). Eine stationäre Temperaturverteilung kann sich sicher nur dann einstellen, wenn diese beiden Größen übereinstimmen. □

Ähnlich wie für die Neumannsche Randbedingung erhält man Eindeutigkeitsaussagen für die dritte Randwertaufgabe. Wir betrachten den Fall $b(\boldsymbol{x}) \equiv 1$, $a(\boldsymbol{x}) \geq 0$, $a(\boldsymbol{x}) \not\equiv 0$. Dann löst die Differenz $v(\boldsymbol{x})$ zweier Lösungen wieder die homogene Aufgabe

$$\Delta v(\boldsymbol{x}) = 0, \quad x \in \Omega, \quad a(\boldsymbol{x})v(\boldsymbol{x}) + \frac{\partial v}{\partial \boldsymbol{n}}(\boldsymbol{x}) = 0, \quad \boldsymbol{x} \in \partial\Omega.$$

Aus der Greenschen Formel folgt wie eben

$$\begin{aligned} 0 &\leq \int\limits_{\Omega} \|\nabla v(\boldsymbol{x})\|_2^2\, d\boldsymbol{x} = \int\limits_{\Omega} \left\{ v(\boldsymbol{x})\Delta v(\boldsymbol{x}) + \|\nabla v(\boldsymbol{x})\|_2^2 \right\} d\boldsymbol{x} \\ &= \int\limits_{\partial\Omega} v(\boldsymbol{x}) \frac{\partial v}{\partial \boldsymbol{n}}(\boldsymbol{x})\, do = -\int\limits_{\partial\Omega} a(\boldsymbol{x}) v^2(\boldsymbol{x})\, do \leq 0, \end{aligned}$$

d.h. $\nabla v(\boldsymbol{x}) \equiv 0$ in Ω, also $v(\boldsymbol{x}) \equiv \text{const}$, und die Randbedingung

$$\frac{\partial v}{\partial \boldsymbol{n}}(\boldsymbol{x}) + a(\boldsymbol{x})v(\boldsymbol{x}) = a(\boldsymbol{x}) \cdot \text{const} = 0$$

liefert wegen $a(\boldsymbol{x}) \not\equiv 0$ schließlich $v(\boldsymbol{x}) \equiv 0$.

1.2.2 Parabolische Pobleme

Für parabolische Differentialgleichungen erhält man Eindeutigkeitsaussagen und die stetige Abhängigkeit der Lösung von Anfangs- und Randwerten wieder aus der Inversmonotonie bzw. einem Maximumprinzip.

Anschaulich besagt das Maximumprinzip für die Wärmeleitungsgleichung: Bleibt die Temperatur am Rande eines Körpers zu jeder Zeit und im Inneren zum Anfangszeitpunkt unter einem Wert M und sind keine Wärmequellen sondern nur Senken vorhanden, so kann im Inneren des Körpers die Temperatur niemals den Wert M übersteigen.

Satz 1.30. *Es sei $\Omega \subset \mathbb{R}^n$ ein beschränktes Gebiet und L (wie in (1.21)) ein gleichmäßig elliptischer Differentialoperator mit $c(\boldsymbol{x}) \geq 0$.*

Die Funktionen $v, w : \bar{\Omega} \times [0,T] \to \mathbb{R}$ seien stetig und in $\Omega \times (0,T)$ zweimal stetig partiell differenzierbar nach den Komponenten von $\boldsymbol{x}$ und einmal stetig partiell differenzierbar nach t. Es gelte

$$\begin{aligned} \frac{\partial}{\partial t} v + Lv \leq \frac{\partial}{\partial t} w + Lw \quad &\text{in} \quad \Omega \times (0,T) \\ v \leq w \quad &\text{auf} \quad (\partial\Omega \times (0,T)) \cup (\Omega \times \{0\}). \end{aligned}$$

Dann folgt

$$v(\boldsymbol{x},t) \leq w(\boldsymbol{x},t) \quad \text{für alle } (\boldsymbol{x},t) \in \bar{\Omega} \times [0,T].$$

Den Beweis findet man wie für den elliptischen Fall in *Protter, Weinberger* [39]. Fast wörtlich wie dort erhält man als Folgerung ein Maximumprinzip, wobei man wiederum im Falle $c(\boldsymbol{x}) \equiv 0$ auf die Vorzeichenvoraussetzung für das Maximum verzichten kann:

Korollar 1.31. (Maximumprinzip)
Es seien die Voraussetzungen von Satz 1.30. erfüllt, und es gelte für die Funktion $u \in C^2(\Omega \times (0,T)) \cap C(\bar{\Omega} \times [0,T])$ die Ungleichung

$$\frac{\partial}{\partial t} u(\boldsymbol{x},t) + Lu(\boldsymbol{x},t) \leq 0, \quad \text{für alle } (\boldsymbol{x},t) \in \Omega \times (0,T).$$

Gilt dann

$$\max_{\boldsymbol{x} \in \bar{\Omega}} u(\boldsymbol{x}) > 0,$$

so wird dieses Maximum von u auf der Menge $(\bar{\Omega} \times \{0\}) \cup (\partial\Omega \times [0,T])$ angenommen.

Mit dem Maximumprinzip erhält man (ähnlich wie im elliptischen Fall) die Eindeutigkeit und stetige Abhängigkeit der Lösung (wenn sie existiert) von den Anfangs- und Randdaten und der Inhomogenität. Wir formulieren diese Ergebnisse in den folgenden drei Korollaren. Auf die Beweise verzichten wir wegen ihrer sehr großen Ähnlichkeit mit den entsprechenden Aussagen für elliptische Probleme.

Korollar 1.32. (Eindeutigkeit)
Die erste Randwertaufgabe

$$\begin{aligned} \tfrac{\partial}{\partial t}u(\boldsymbol{x},t) + Lu(\boldsymbol{x},t) &= f(\boldsymbol{x},t) \quad , \quad (\boldsymbol{x},t) \in \Omega \times (0,T], \\ u(\boldsymbol{x},0) &= \phi(\boldsymbol{x}) \quad , \quad \boldsymbol{x} \in \Omega, \\ u(\boldsymbol{x},t) &= \psi(\boldsymbol{x},t) \quad , \quad (\boldsymbol{x},t) \in \partial\Omega \times [0,T], \end{aligned}$$

besitzt höchstens eine Lösung.

Korollar 1.33. (Stetige Abhängigkeit von Anfangs- und Randdaten)
Lösen die Funktionen u_i, $i = 1, 2$, die Anfangsrandwertaufgaben

$$\begin{aligned} \tfrac{\partial}{\partial t}u_i(\boldsymbol{x},t) + Lu_i(\boldsymbol{x},t) &= f(\boldsymbol{x},t) \quad , \quad (\boldsymbol{x},t) \in \Omega \times (0,T], \\ u_i(\boldsymbol{x},0) &= \phi_i(\boldsymbol{x}) \quad , \quad \boldsymbol{x} \in \Omega, \\ u_i(\boldsymbol{x},t) &= \psi_i(\boldsymbol{x},t) \quad , \quad (\boldsymbol{x},t) \in \partial\Omega \times [0,T], \end{aligned}$$

so gilt

$$\begin{aligned} &\max\{|u_1(\boldsymbol{x},t) - u_2(\boldsymbol{x},t)| \,:\, \boldsymbol{x} \in \bar{\Omega},\, t \in [0,T]\} \\ &\quad \leq \max \left\{ \begin{array}{l} \max\{|\phi_1(\boldsymbol{x}) - \phi_2(\boldsymbol{x})| \,:\, \boldsymbol{x} \in \bar{\Omega}\}, \\ \max\{|\psi_1(\boldsymbol{x},t) - \psi_2(\boldsymbol{x},t)| \,:\, (\boldsymbol{x},t) \in \partial\Omega \times [0,T]\} \end{array} \right\}. \end{aligned}$$

Korollar 1.34. (Stetige Abhängigkeit vom Quellterm)
Es seien u_i, $i = 1, 2$, Lösungen der Anfangsrandwertaufgaben

$$\begin{aligned} \tfrac{\partial}{\partial t}u_i(\boldsymbol{x},t) + Lu_i(\boldsymbol{x},t) &= f_i(\boldsymbol{x},t) \quad , \quad (\boldsymbol{x},t) \in \Omega \times (0,T], \\ u_i(\boldsymbol{x},0) &= \phi(\boldsymbol{x}) \quad , \quad \boldsymbol{x} \in \Omega, \\ u_i(\boldsymbol{x},t) &= \psi(\boldsymbol{x},t) \quad , \quad (\boldsymbol{x},t) \in \partial\Omega \times [0,T]. \end{aligned}$$

Dann gilt

$$\begin{aligned} &\max\{|u_1(\boldsymbol{x},t) - u_2(\boldsymbol{x},t)| \,:\, \boldsymbol{x} \in \bar{\Omega},\, t \in [0,T]\} \\ &\quad \leq \; T \cdot \max\{|f_1(\boldsymbol{x},t) - f_2(\boldsymbol{x},t)| \,:\, \boldsymbol{x} \in \bar{\Omega},\, t \in [0,T]\}. \end{aligned}$$

Kapitel 2

Einschrittverfahren

2.1 Das Eulersche Polygonzugverfahren

Wir betrachten die Anfangswertaufgabe

$$y' = f(x, y), \;\; y(a) = y_0, \tag{2.1}$$

wobei die Lösung y im Intervall $[a, b]$ gesucht ist.

Dabei kann $\boldsymbol{y}$ auch vektorwertig, also (2.1) ein Differentialgleichungssystem erster Ordnung sein.

Es sei $a = x_0 < x_1 < x_2 < \ldots < x_N =: b$ eine (nicht notwendig äquidistante) Zerlegung von $[a, b]$. Wir nehmen diese Zerlegung zunächst als gegeben an. Tatsächlich wird die Folge der x_j im Verfahren mitbestimmt und an das Verhalten der Lösung der Anfangswertaufgabe angepasst.

Da $f(x_n, y(x_n))$ gerade die Steigung $y'(x_n)$ der gesuchten Lösung $y(x)$ von (2.1) ist, gilt näherungsweise bei nicht zu großer Schrittweite $h_n := x_{n+1} - x_n$

$$\frac{1}{h_n} \left(y(x_{n+1}) - y(x_n) \right) \approx f(x_n, y(x_n)),$$

d.h.

$$y(x_{n+1}) = y(x_n) + h_n \, f(x_n, y(x_n)) + \varepsilon_n. \tag{2.2}$$

Wir vernachlässigen nun in (2.2) den Fehler ε_n. Dann wird die entstehende Gleichung nicht mehr durch die Lösung $y(x_n)$ von (2.1) an den Knoten x_n erfüllt, sondern nur

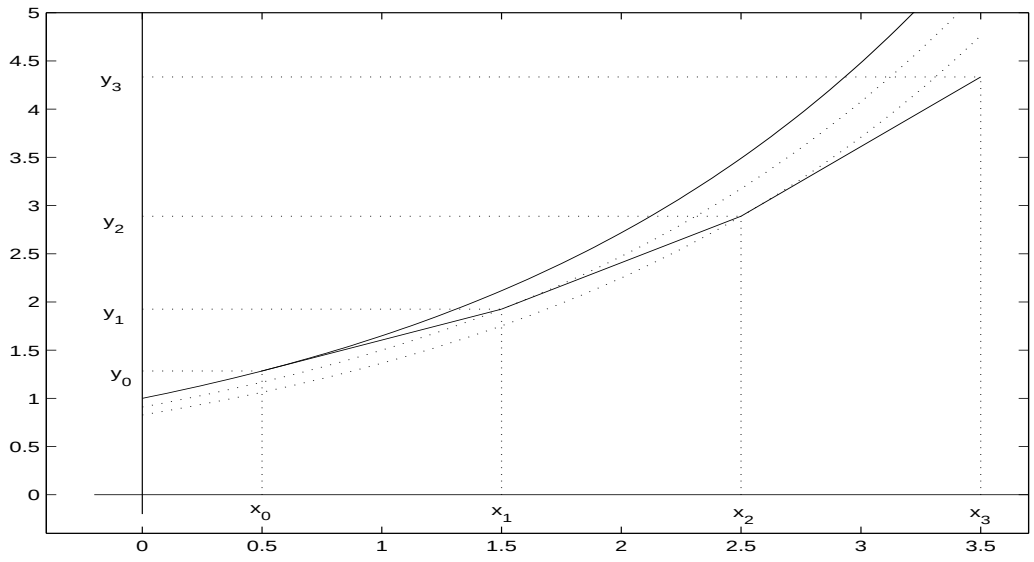

Abbildung 2.1: Knotenpunkte und Näherungswerte

noch durch Näherungswerte y_n für $y(x_n)$. Wir bestimmen also die y_n ausgehend von y_0 durch das Verfahren

$$y_{n+1} = y_n + h_n\, f(x_n, y_n), \quad n = 0, 1, \ldots, N-1, \tag{2.3}$$

wobei $h_n := x_{n+1} - x_n$ ist.

Definition 2.1. *Das durch (2.3) beschriebene Verfahren zur approximativen Lösung der Anfangswertaufgabe (2.1) heißt das* **Eulersche Polygonzugverfahren** *. Es wurde 1768 von L. Euler beschrieben.*

Beispiel 2.2.

$$y' = y^2,\ y(0.8) = \frac{5}{6}, \quad x \in [0.8, 1.8]$$

besitzt die Lösung $y(x) = \dfrac{1}{2-x}$.

Mit den äquidistanten Schrittweiten $h = \dfrac{1}{100}, \dfrac{1}{200}$ und $\dfrac{1}{400}$ liefert das Verfahren (2.3) Näherungen, deren Fehler in der Tabelle 2.1 enthalten sind. Man liest aus der Tabelle ab, dass die Fehler bei Halbierung der Schrittweite ebenfalls halbiert werden. □

Wir wollen nun im allgemeinen Fall den entstandenen Fehler abschätzen. Dieser setzt sich aus zwei Anteilen zusammen: Wir haben im n-ten Schritt die Lösung

Tabelle 2.1: Fehlertabelle von Beispiel 2.2.

x	$N = 100$	$N = 200$	$N = 400$
0.80	$0.00E+0$	$0.00E+0$	$0.00E+0$
0.90	$-7.09E-4$	$-3.57E-4$	$-1.79E-4$
1.00	$-1.79E-3$	$-9.04E-4$	$-4.54E-4$
1.10	$-3.49E-3$	$-1.76E-3$	$-8.84E-4$
1.20	$-6.20E-3$	$-3.13E-3$	$-1.57E-3$
1.30	$-1.07E-2$	$-5.43E-3$	$-2.73E-3$
1.40	$-1.86E-2$	$-9.47E-3$	$-4.77E-3$
1.50	$-3.35E-2$	$-1.71E-2$	$-8.66E-3$
1.60	$-6.48E-2$	$-3.33E-2$	$-1.69E-2$
1.70	$-1.41E-1$	$-7.38E-2$	$-3.77E-2$
1.80	$-3.90E-1$	$-2.08E-1$	$-1.08E-1$

$y(x_{n+1}; x_n, y(x_n))$ der Differentialgleichung $y' = f(x, y)$ mit dem Anfangswert $y(x_n)$ an der Stelle x_n zu bestimmen. Statt dessen betrachten wir die Anfangswertaufgabe

$$y' = f(x, y),\ y(x_n) = y_n$$

mit dem "falschen" Anfangswert y_n, und wir lösen diese auch nur näherungsweise, indem wir den Abbruchfehler vernachlässigen. Tatsächlich werden bei der Realisierung des Verfahrens auf einem Rechner noch bei der Auswertung von $f(x_n, y_n)$ und den Rechenoperationen Rundungsfehler gemacht. Diese wollen wir aber bei den folgenden Betrachtungen außer Acht lassen.

Wir schreiben das Polygonzugverfahren in der Form

$$y_{n+1} - y_n - h_n\, f(x_n, y_n) = 0. \tag{2.4}$$

Setzt man hier an Stelle der Werte y_n die Werte $y(x_n)$ der Lösung von (2.1) an den Knoten x_n ein, so erhält man (vgl. (2.2))

$$y(x_{n+1}) - y(x_n) - h_n\, f(x_n, y(x_n)) =: \varepsilon(x_n, h_n). \tag{2.5}$$

Definition 2.3. *$\varepsilon(x_n, h_n)$ heißt der* **lokale Fehler** *(auch* **Abbruchfehler***) des Polygonzugverfahrens an der Stelle x_n bei der Schrittweite h_n.*

Subtrahiert man die Gleichung (2.4) von (2.5), so folgt

$$y(x_{n+1}) - y_{n+1} = y(x_n) - y_n + h_n\,(f(x_n, y(x_n)) - f(x_n, y_n)) + \varepsilon(x_n, h_n). \tag{2.6}$$

Definition 2.4.

$$\delta_{n+1} := |y(x_{n+1}) - y_{n+1}|$$

heißt der **Fehler** *oder (zur besseren Unterscheidung)* **globale Fehler** *des Polygonzugverfahrens an der Stelle x_{n+1}.*

Um den globalen Fehler abschätzen zu können, setzen wir voraus, dass f auf $[a,b]\times\mathbb{R}$ einer globalen Lipschitz Bedingung

$$|f(x,y)-f(x,z)| \le L|y-z| \quad \text{für alle } y,z\in\mathbb{R} \text{ und alle } x\in[a,b]$$

bzgl. y genügt. Eine Lipschitz Bedingung in einem Rechteck (wie in dem Satz von Picard–Lindelöf) würde auch genügen. Man müsste dann nur aufpassen, dass die Näherungslösungen dieses Rechteck nicht verlassen.

Dann folgt aus (2.6) mit der Dreiecksungleichung und $\varepsilon_n := \varepsilon(x_n,h_n)$

$$\delta_{n+1} \le (1+Lh_n)\,\delta_n + |\varepsilon_n|, \tag{2.7}$$

und durch vollständige Induktion erhält man hieraus

$$\delta_n \le \left(\delta_0 + \sum_{j=0}^{n-1}|\varepsilon_j|\right)\cdot\exp\left(\sum_{j=0}^{n-1} h_j\,L\right), \tag{2.8}$$

denn für $n=0$ ist diese Aussage trivial, und ist (2.8) für ein $n<N$ erfüllt, so folgt

$$\begin{aligned}
\delta_{n+1} &\le (1+Lh_n)\,\delta_n + |\varepsilon_n| \le \exp(Lh_n)\,\delta_n + |\varepsilon_n| \\
&\le \exp(Lh_n)\cdot\left(\delta_0+\sum_{j=0}^{n-1}|\varepsilon_j|\right)\cdot\exp\left(\sum_{j=0}^{n-1}h_j\,L\right)+|\varepsilon_n| \\
&= \left(\delta_0+\sum_{j=0}^{n-1}|\varepsilon_j|\right)\cdot\exp\left(\sum_{j=0}^{n}h_j\,L\right)+|\varepsilon_n| \\
&\le \left(\delta_0+\sum_{j=0}^{n}|\varepsilon_j|\right)\cdot\exp\left(\sum_{j=0}^{n}h_j\,L\right)
\end{aligned}$$

Es ist

$$\sum_{j=0}^{n} h_j \le \sum_{j=0}^{N-1} h_j = b-a$$

und $\delta_0 = |y(a)-y_0| = 0$. Daher folgt

$$\delta_n \le e^{(b-a)L}\sum_{j=0}^{N-1}|\varepsilon_j|.$$

Bis auf eine multiplikative Konstante lässt sich der globale Fehler also durch die Summe der lokalen Fehler abschätzen.

Wir setzen nun weiter voraus, dass die Lösung y von (2.1) zweimal stetig differenzierbar ist. Dies ist z.B. erfüllt, wenn die rechte Seite f stetig differenzierbar in einer offenen Menge ist, die

$$\{(x,y(x)) \,:\, a\le x\le b\}$$

enthält. Dann gilt nach dem Taylorschen Satz für den lokalen Fehler

$$\varepsilon_j = y(x_j + h_j) - y(x_j) - h_j y'(x_j) = \frac{1}{2} h_j^2 y''(x_j + \theta_j h_j)$$

mit einem $\theta_j \in (0,1)$, und daher folgt für den globalen Fehler

$$\begin{aligned}
\delta_n &\leq \frac{1}{2} e^{(b-a)L} \sum_{j=0}^{N-1} h_j^2 |y''(x_j + \theta_j h_j)| \\
&\leq \frac{1}{2}\, e^{(b-a)L} \max_{a\leq x\leq b} |y''(x)| \sum_{j=0}^{N-1} h_j^2 \\
&\leq \frac{1}{2}\, e^{(b-a)L} \max_{a\leq x\leq b} |y''(x)| \max_{j=0,\ldots,N-1} h_j \sum_{j=0}^{N-1} h_j \\
&\leq \frac{1}{2}\, (b-a)\, e^{(b-a)L} \max_{a\leq x\leq b} |y''(x)| \max_{j=0,\ldots,N-1} h_j \\
&=: C \cdot \max_{j=0,\ldots,N-1} h_j
\end{aligned}$$

mit einer von den gewählten Schrittweiten h_n unabhängigen Konstante C.

Hieraus liest man ab, dass bei Halbierung der Schrittweiten der Fehler ebenfalls halbiert wird (vgl. Beispiel 2.2.). Man liest ferner ab, dass man $O(\delta^{-1})$ Schritte des Polygonzugverfahrens aufwenden muss, um den globalen Fehler δ zu erreichen (für den Fehler $\delta = 10^{-6}$ also $c \cdot 10^6$ Schritte). Dies zeigt, dass das Eulersche Polygonzugverfahren für die Praxis nicht geeignet ist und dass man schnellere Verfahren benötigt.

Wie im Falle der Quadratur wird man in der Praxis nicht mit vorgegebenen Schrittweiten rechnen, sondern die Schrittweite dem Lösungsverhalten anpassen. Dabei schätzt man wie bei den adaptiven Quadraturformeln den lokalen Fehler mit Hilfe einer zweiten Formel.

Wir verwenden hierzu zwei Schritte des Polygonzugverfahrens mit halber Schrittweite:

$$\begin{aligned}
\tilde{y}_{n+\frac{1}{2}} &= y_n + \frac{h_n}{2} f(x_n, y_n) \\
\tilde{y}_{n+1} &= \tilde{y}_{n+\frac{1}{2}} + \frac{h_n}{2} f(x_n + \frac{h_n}{2}, \tilde{y}_{n+\frac{1}{2}}) \\
&= y_n + \frac{h_n}{2} f(x_n, y_n) + \frac{h_n}{2} f(x_n + \frac{h_n}{2}, y_n + \frac{h_n}{2} f(x_n, y_n)).
\end{aligned}$$

Für den lokalen Fehler gilt mit der Lösung $z(x)$ der Anfangswertaufgabe $y' = f(x,y)$, $y(x_n) = y_n$ (im Falle $z \in C^3\,[a,b]$) nach dem Taylorschen Satz

$$\varepsilon(x_n, h_n) = z(x_n + h_n) - (y_n + h_n\, f(x_n, y_n))$$

$$
\begin{aligned}
&= z(x_n) + h_n\, z'(x_n) + \frac{1}{2}\, h_n^2 z''(x_n) + O(h_n^3) - z(x_n) - h_n\, z'(x_n) \\
&= \frac{1}{2}\, h_n^2 z''(x_n) + O(h_n^3) \qquad (2.9)
\end{aligned}
$$

und genauso für die zusammengesetzte Formel

$$
\begin{aligned}
&\tilde{\varepsilon}(x_n, h_n) = z(x_n + h_n) - \tilde{y}_{n+1} \\
&= y_n + h_n\, f(x_n, y_n) + \frac{1}{2}\, h_n^2 z''(x_n) + O(h_n^3) - y_n - \frac{h_n}{2}\, f(x_n, y_n) \\
&\quad - \frac{h_n}{2}(f(x_n, y_n) + \frac{h_n}{2}\, \frac{\partial}{\partial x}\, f(x_n, y_n) + \frac{h_n}{2}\, \frac{\partial}{\partial y}\, f(x_n, y_n) f(x_n, y_n) + O(h_n^2)) \\
&= \frac{1}{4}\, h_n^2 z''(x_n) + O(h_n^3)
\end{aligned}
$$

wegen

$$
z''(x) = \frac{d}{dx}\, f(x, z(x)) = \frac{\partial}{\partial x}\, f(x, z(x)) + \frac{\partial}{\partial y}\, f(x, z(x))\, z'(x).
$$

Durch Subtraktion dieser beiden Formeln erhält man

$$
\tilde{y}_{n+1} - y_{n+1} = \frac{1}{4}\, h_n^2 z''(x_n) + O(h_n^3).
$$

Setzt man dies in (2.9) unter Vernachlässigung des $O(h_n^3)$-Terms ein, so erhält man die Schätzung für den lokalen Fehler

$$
\varepsilon(x_n, h_n) \approx \phi(x_n, h_n) := 2(\tilde{y}_{n+1} - y_{n+1}). \qquad (2.10)
$$

Zugleich erhält man mit

$$
\begin{aligned}
\hat{y}_{n+1} &:= 2\tilde{y}_{n+1} - y_{n+1} \\
&= y_n + h_n f(x_n + \frac{1}{2} h_n, y_n + \frac{h_n}{2} f(x_n, y_n)) \qquad (2.11)
\end{aligned}
$$

eine Näherung für $y(x_n + h_n)$ mit dem lokalen Fehler

$$
\hat{\varepsilon}(x_n, h_n) = 2\,\tilde{\varepsilon}(x_n, h_n) - \varepsilon(x_n, h_n) = O(h_n^3).
$$

Verfahren mit dieser Eigenschaft werden wir Verfahren der Ordnung 2 nennen.

Die Formel (2.10) verwenden wir nun zur **Schrittweitensteuerung** :
Wir geben uns eine Toleranz $\tau > 0$ vor und bestimmen die Schrittweite in jedem Schritt so, dass

$$
\text{lokaler Fehler } \approx \tau \qquad (2.12)
$$

gilt.

Approximieren wir $\varepsilon(x_n, h)$ durch $\varepsilon(x_n, h) \approx \gamma h^2$, so kann man γ durch einen Probeschritt der Länge H schätzen:

$$\gamma \approx \frac{1}{H^2}\,\varepsilon(x_n, H).$$

Die optimale Wahl der Schrittweite wäre nach (2.12)

$$\tau = |\varepsilon(x_n, h)| \approx |\gamma| \cdot h^2 \approx \frac{h^2}{H^2}\,|\varepsilon(x_n, H)|,$$

d.h.

$$h = H\sqrt{\frac{\tau}{|\varepsilon(x_n, H)|}}.$$

Der folgende MATLAB-Programmteil verwendet eine ähnliche Schrittweitenkontrolle bei gegebenen Startwerten x und y und gegebener Probeschrittlänge h:

```
v=1.e-5*ones(1,n);
z = f(x,y);
while h>0
  y1 = y + h*z;
  y2 = y + h/2*z;
  y2 = y2 + h/2 * f(x+h/2,y2);
  d=max(v,max(abs(y),abs(y1)));
  phi = 2 * norm((y2 - y1)./d);
  hneu = h * min(max(0.9*sqrt(tol/phi),0.2),10);
  if phi > tol
    h = hneu;
  else
    x = x + h;
    y = 2*y2 - y1;                (*)
    z = f(x,y);
    h = min(b-x,hneu);
    end
  end
```

Bemerkung 2.5. Es wurde der absolute Fehler durch den "relativen" Fehler ersetzt. Ferner wurde dabei der Betrag des Funktionswerts nach unten komponentenweise durch 10^{-5} begrenzt. Zusätzlich wurde die "optimale" Schrittweite *hneu* mit

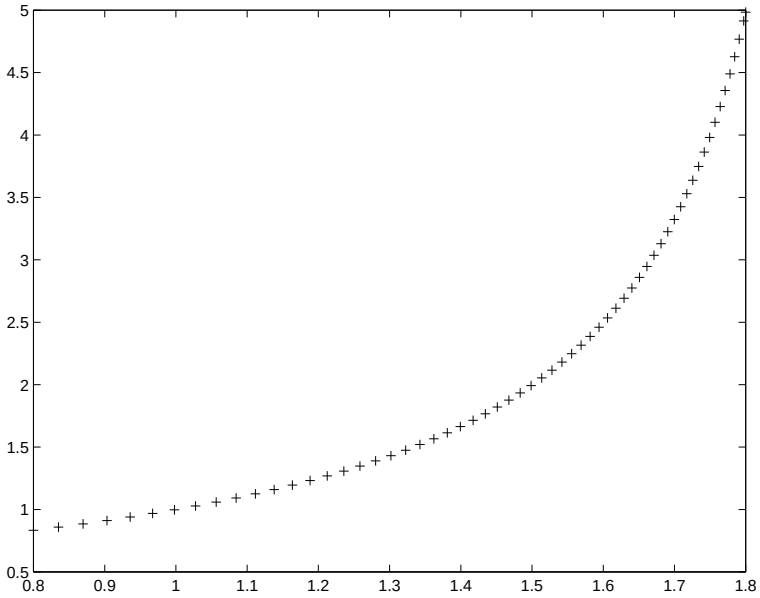

Abbildung 2.2: Schrittweitenkontrolle in Beispiel 2.7.

dem Faktor 0.9 verkleinert. Dies verringert die Wahrscheinlichkeit, dass der nächste Schritt verworfen wird. Schließlich wurde durch den minimalen Faktor 0.2 und den maximalen Faktor 10 dafür gesorgt, dass die Schrittweiten von Schritt zu Schritt sich nicht zu stark ändern. Die gewählten Konstanten 0.9, 0.2 und 10 lassen sich durch keine Theorie begründen, sondern haben sich in Codes bewährt. □

Bemerkung 2.6. Nach unserer Herleitung müsste in der Zeile $(*)$ $y = y1$ stehen. Da man aber ohne Mehrkosten die bessere Näherung $y = 2 * y2 - y1$ (Formel der Ordnung 2) zur Verfügung hat, verwendet man diese. Unsere Fehlerschätzung ist damit in der Regel pessimistisch. □

Beispiel 2.7.

$$y' = y^2, \;\; y(0.8) = \frac{5}{6}.$$

Mit $\tau = 1e-3$ benötigt man 61 Funktionsauswertungen für die numerische Lösung im Intervall $[0.8, 1.8]$. Der maximale absolute Fehler ist dabei $1.61 \cdot 10^{-2}$, der maximale relative Fehler $3.21e-3$, die maximale benutzte Schrittweite ist $3.52 \cdot 10^{-2}$ und die minimale Schrittweite ist $2.86 \cdot 10^{-3}$. Abbildung 2.2 zeigt die berechneten Punkte; die Schrittweite wird kleiner bei größerer Steigung.

Um dieselbe Genauigkeit mit äquidistanter Schrittweite zu erreichen, benötigt man 2776 Funktionsauswertungen. □

2.2 Allgemeine Einschrittverfahren

Das behandelte Polygonzugverfahren ist die einfachste Methode der großen Klasse der **Einschrittverfahren**, bei denen die Näherung y_{n+1} an dem neuen Punkt $x_{n+1} := x_n + h_n$ allein aus der Näherung y_n an der Stelle x_n und der Schrittweite h_n berechnet wird. Einschrittverfahren haben also die folgende Gestalt

$$y_{n+1} = y_n + h_n \, \Phi(x_n, y_n, h_n) \tag{2.13}$$

mit einer **Verfahrensfunktion** Φ.

Um die Güte von Einschrittverfahren zu beurteilen, führen wir die folgenden Begriffe ein:

Definition 2.8. *Es sei $z(x)$ die Lösung der Anfangswertaufgabe*

$$z' = f(x, z(x)), \;\; z(x_n) = y_n.$$

Dann heißt

$$\varepsilon(h) := z(x_n + h) - y_n - h \, \Phi(x_n, y_n, h)$$

der **lokale Fehler** *des durch (2.13) definierten Verfahrens.*

Das Verfahren (2.13) heißt **konsistent**, *falls $\varepsilon(h) = o(h)$ gilt, es heißt* **von der Ordnung** *p, wenn $\varepsilon(h) = O(h^{p+1})$ gilt.*

Wie im Falle des Polygonzugverfahrens gilt:

Satz 2.9. *Es seien die Näherungen y_n von $y(x_n)$ mit dem Einschrittverfahren*

$$y_{n+1} = y_n + h_n \Phi(x_n, y_n, h_n), \;\; n = 0, 1, \ldots, N,$$

berechnet.

Erfüllt die Verfahrensfunktion Φ eine Lipschitz Bedingung bzgl. y in $[a, b] \times \mathbb{R}$ (es genügt eine Umgebung der Lösung)

$$|\Phi(x, y, h) - \Phi(x, z, h)| \leq \Lambda |y - z| \tag{2.14}$$

und ist das Einschrittverfahren konsistent von der Ordnung p

$$|y(x + h) - y(x) - h\Phi(x, y(x), h)| \leq C \cdot h^{p+1}, \tag{2.15}$$

so gilt für den globalen Fehler

$$|\delta_n| = |y_n - y(x_n)| \leq C(b-a)e^{\Lambda(b-a)}\, h^p, \tag{2.16}$$

wobei

$$h := \max_{j=0,\ldots,N-1} h_j$$

gesetzt ist.

Beweis: Mit wörtlich demselben Beweis wie für das Polygonzugverfahren erhält man aus (2.13) und (2.14)

$$\delta_n \leq \exp\left(\sum_{j=0}^{n-1} h_j \Lambda\right) \cdot \sum_{j=0}^{n-1} |\varepsilon_j|,$$

und mit (2.15) folgt

$$\begin{aligned}
\delta_n &\leq \exp\left(\sum_{j=0}^{n-1} h_j \Lambda\right) \cdot \sum_{j=0}^{n-1} C h_j^{p+1} \\
&\leq \exp\left(\Lambda(b-a)\right) \cdot C \cdot \max_{j=0,\ldots,N-1} h_j^p \sum_{j=0}^{n-1} h_j \\
&\leq C(b-a)e^{\Lambda(b-a)} \cdot h^p.
\end{aligned}$$

∎

Bemerkung 2.10. Die Lipschitz Bedingung für die Verfahrensfunktion Φ erhält man in vielen Fällen aus der Lipschitz Bedingung für die rechte Seite f. □

Bemerkung 2.11. Wie beim Übergang von den Quadraturformeln zu summierten Quadraturformeln verliert man beim Übergang vom lokalen zum globalen Fehler eine h-Potenz. □

Beispiel 2.12. (Polygonzugverfahren) *(Euler 1768)*
Das einfachste Einschrittverfahren ist das in Abschnitt 2.1 betrachtete Eulersche Polygonzugverfahren

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

□

Beispiel 2.13. (Verbessertes Polygonzugverfahren)
(Coriolis 1837, Runge 1895)
Wir haben dieses Verfahren bereits durch Extrapolation aus dem Polygonzugverfahren mit den Schrittweiten h und $\frac{h}{2}$ hergeleitet (vgl. (2.11)):

$$y_{n+1} = y_n + h_n\, f\Big(x_n + \frac{h_n}{2}, y_n + \frac{h_n}{2}\, f(x_n, y_n)\Big).$$

Geometrisch kann man dieses Verfahren so interpretieren: Es wird zunächst eine Schätzung $y_{n+\frac{1}{2}} = y_n + \frac{h_n}{2}\, f(x_n, y_n)$ für $y\Big(x_n + \frac{h_n}{2}\Big)$ ermittelt, und die hiermit berechnete Näherung $f\Big(x_n + \frac{h_n}{2}, y_{n+\frac{1}{2}}\Big) \approx y'\Big(x_n + \frac{h_n}{2}\Big)$ für die Steigung von y im ganzen Intervall $[x_n, x_n + h_n]$ verwendet. Eine Veranschaulichung findet sich in der linken Skizze in Abbildung 2.3.

Wir haben bereits gesehen, dass das verbesserte Polygonzugverfahren die Ordnung 2 besitzt. Erfüllt f eine Lipschitzbedingung

$$|f(x,y) - f(x,z)| \le L|y - z|,$$

so erfüllt auch die Verfahrensfunktion

$$\Phi(x, y, h) = f(x + 0.5h, y + 0.5hf(x, y))$$

wegen

$$\begin{aligned}
&|\Phi(x,y,h) - \Phi(x,z,h)| \\
&\quad = \; |f(x + 0.5h, y + 0.5hf(x,y)) - f(x + 0.5h, z + 0.5hf(x,z))| \\
&\quad \le \; L|y + 0.5hf(x,y) - (z + 0.5hf(x,z))| \\
&\quad \le \; L(|y - z| + 0.5h|f(x,y) - f(x,z)|) \\
&\quad \le \; L(|y - z| + 0.5hL|y - z|) \\
&\quad \le \; L(1 + 0.5(b - a)L)|y - z| =: \Lambda|y - z|
\end{aligned}$$

eine Lipschitz Bedingung. Nach Satz 2.9. ist das verbesserte Polygonzugverfahren also ein Verfahren der Ordnung 2. □

Beispiel 2.14. Eine naheliegende Idee, Verfahren höherer Ordnung zu entwickeln, liefert der Taylorsche Satz. Ist f differenzierbar, so erhält man wegen

$$\begin{aligned}
y''(x) &= \frac{d}{dx} f(x, y(x)) = f_x(x, y(x)) + f_y(x, y(x)) y'(x) \\
&= f_x(x, y(x)) + f_y(x, y(x)) f(x, y(x))
\end{aligned}$$

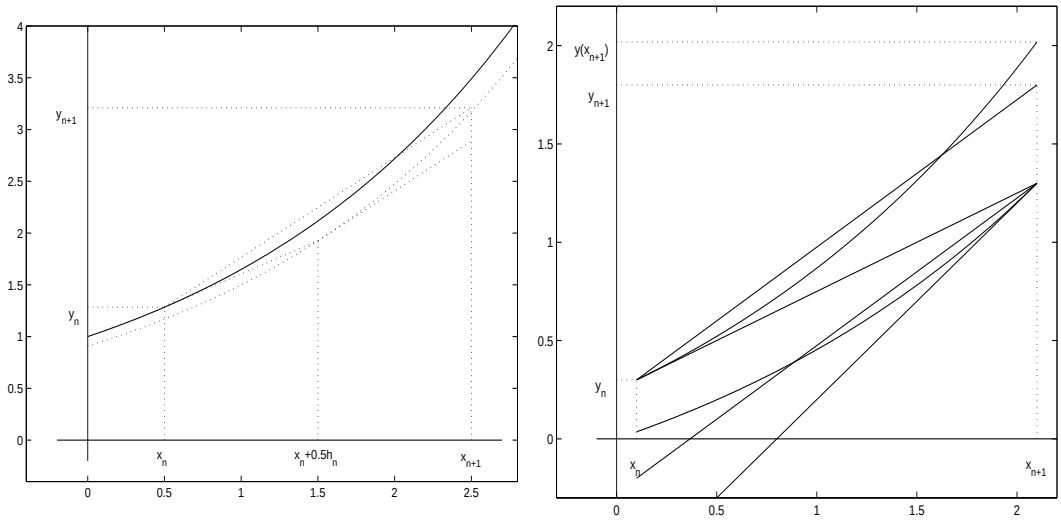

Abbildung 2.3: Verbessertes Polygonzugverfahren / Verfahren von Heun

und

$$\begin{aligned} y(x_n + h) &= y(x_n) + hy'(x_n) + \frac{1}{2}h^2 y''(x_n) + O(h^3) \\ &= y_n + hf(x_n, y_n) + \frac{1}{2}h^2(f_x(x_n, y_n) + f_y(x_n, y_n)f(x_n, y_n)) + O(h^3) \end{aligned}$$

mit

$$\Phi(x, y, h) = f(x, y) + 0.5h(f_x(x, y) + f_y(x, y)f(x, y))$$

ein Einschrittverfahren der Ordnung 2. Nachteil ist, dass man die Ableitung der rechten Seite benötigt, deren Auswertung häufig aufwendiger ist als die der Funktion.

Besitzt f höhere Ableitungen, so kann man auf dieselbe Weise Einschrittverfahren höherer Ordnung bestimmen. □

Beispiel 2.15. (Verfahren von Heun) *(Heun 1900)*
Dem verbesserten Polygonzugverfahren verwandt ist das Verfahren von Heun, das in der Literatur machmal auch Verbessertes Polygonzugverfahren genannt wird. Man verwendet den Mittelwert zweier Näherungen

$$k_1 := f(x_n, y_n), \quad k_2 := f(x_n + h_n, y_n + h_n k_1)$$

für die Steigung von y im Intervall $[x_n, x_{n+1}]$ und setzt hiermit

$$y_{n+1} = y_n + h_n \frac{k_1 + k_2}{2}.$$

Dieses Verfahren entspricht der Quadratur des Integrals in

$$y(x_{n+1}) - y(x_n) = \int_{x_n}^{x_{n+1}} f(t, y(t))\, dt$$

mit der Trapezregel, wenn man den unbekannten Punkt $(x_{n+1}, y(x_{n+1}))$ ersetzt durch $(x_{n+1}, y_n + h_n f(x_n, y_n))$. Eine Veranschaulichung findet man in der rechten Skizze von Abbildung 2.3.

Mit dem Taylorschen Satz kann man zeigen, dass für den lokalen Fehler

$$\varepsilon(h) = z(x_n + h) - y_n - \frac{h}{2}(f(x_n, y_n) + f(x_n + h, y_n + h\, f(x_n, y_n))) = O(h^3)$$

gilt, dass das Verfahren von Heun also wie das verbesserte Polygonzugverfahren die Ordnung 2 besitzt. □

Beispiel 2.16. (explizite Runge–Kutta Verfahren) *(Kutta 1901)*
Explizite Runge–Kutta Verfahren sind Verallgemeinerungen der Einschrittverfahren in Beispiel 2.12., Beispiel 2.13. und Beispiel 2.15.. Es wird ein gewichtetes Mittel von Approximationen der Steigung der Lösung y im Intervall $[x_n, x_{n+1}]$ bestimmt und hiermit ein Schritt der Länge h_n ausgeführt. Es sei

$$\begin{aligned} k_1 &:= f(x_n, y_n) \\ k_j &:= f(x_n + \alpha_j\, h_n, y_n + h_n \sum_{\ell=1}^{j-1} \beta_{j\ell}\, k_\ell), \quad j = 2, \ldots, s \\ y_{n+1} &:= y_n + h_n \sum_{j=1}^{s} \gamma_j\, k_j. \end{aligned} \tag{2.17}$$

Die Koeffizienten $\alpha_j, \beta_{j\ell}, \gamma_j$ werden dabei so gewählt, dass das Verfahren möglichst hohe Ordnung hat. s heißt die **Stufe** des Runge–Kutta Verfahrens. □

Als Beispiel betrachten wir die Herleitung von zweistufigen Runge–Kutta Verfahren. Es gilt

$$\begin{aligned} y_{n+1} &= y_n + h(\gamma_1 k_1 + \gamma_2 k_2) \\ &= y_n + h(\gamma_1 f(x_n, y_n) + \gamma_2 f(x_n + \alpha_2 h, y_n + h\beta_{21} f(x_n, y_n))). \end{aligned} \tag{2.18}$$

Wir bestimmen die Parameter γ_1, γ_2, α_2 und β_{21} so, dass das Verfahren möglichst große Ordnung hat.

Ist $f \in C^2$ (und damit $y \in C^3$), so gilt nach dem Taylorschen Satz

$$\begin{aligned} y(x+h) &= y(x) + hy'(x) + \frac{1}{2}h^2 y''(x) + \frac{1}{6}h^3 y'''(x) + o(h^3) \\ &= y(x) + hf(x, y(x)) + \frac{1}{2}h^2(f_x(x, y(x)) + f_y(x, y(x))f(x, y(x))) \\ &\quad + \frac{1}{6}h^3[f_{xx} + 2ff_{xy} + f^2 f_{yy} + f_x f_y + f f_y^2](x, y(x)) + o(h^3) \end{aligned}$$

und

$$\begin{aligned} & y(x) + h(\gamma_1 f(x, y(x)) + \gamma_2 f(x + \alpha_2 h, y(x) + h\beta_{21} f(x, y(x)))) \\ = \;& y(x) + h\gamma_1 f(x, y(x)) + h\gamma_2 [f + \alpha_2 h f_x + \beta_{21} h f_y f + \frac{1}{2}(\alpha_2 h)^2 f_{xx} \\ & + \alpha_2 \beta_{21} h^2 f_{xy} f + \frac{1}{2}(\beta_{21} h)^2 f^2 f_{yy}](x, y(x)) + o(h^3). \end{aligned}$$

Subtraktion und Ordnen nach Potenzen von h liefert

$$\begin{aligned} & y(x+h) - y(x) - h(\gamma_1 f(x, y(x)) + \gamma_2 f(x + \alpha_2 h, y(x) + h\beta_{21} f(x, y(x)))) \\ = \;& h(1 - \gamma_1 - \gamma_2) f(x, y(x)) \\ & + \frac{1}{2} h^2 [(1 - 2\gamma_2 \alpha_2) f_x + (1 - 2\gamma_2 \beta_{21}) f f_y](x, y(x)) \\ & + \frac{1}{6} h^3 [(1 - 3\gamma_2 \alpha_2^2) f_{xx} + 2(1 - 3\gamma_2 \alpha_2 \beta_{21}) f_{xy} f \\ & \quad + (1 - 3\gamma_2 \beta_{21}^2) f_{yy} f^2 + f_x f_y + f f_y^2](x, y(x)) + o(h^3). \end{aligned}$$

Da bei keiner Wahl der Parameter der Koeffizient bei h^3 für alle Funktionen f verschwindet, können wir keine höhere Konvergenzordnung als 2 erreichen.

Für Verfahren der Ordnung 2 muss gelten

$$\gamma_1 + \gamma_2 = 1, \; 2\gamma_2 \alpha_2 = 1, \; 2\gamma_2 \beta_{21} = 1. \tag{2.19}$$

Dieses nichtlineare System von 3 Gleichungen in 4 Unbekannten besitzt unendlich viele Lösungen. Wählt man γ_2 als freien Parameter, so erhält man die Lösungsschar

$$\gamma_1 = 1 - \gamma_2, \; \alpha_2 = \beta_{21} = \frac{1}{2\gamma_2}, \; \gamma_2 \neq 0. \tag{2.20}$$

Die bereits betrachteten Verfahren der Ordnung 2 sind hierin enthalten. Für $\gamma_2 = 1$ erhält man das verbesserte Polygonzugverfahren, für $\gamma_2 = 0.5$ das Verfahren von Heun.

Unter Verwendung von (2.19) geht der Term bei h^3 über in

$$\frac{1}{6}((1 - \frac{3}{4\gamma_2})(f_{xx} + 2 f f_{xy} + f^2 f_{yy}) + f_x f_y + f f_y^2)(x, y(x)). \tag{2.21}$$

Hier ist die Summe der Beträge der Koeffizienten vor den partiellen Ableitungen minimal für $\gamma_2 = \frac{3}{4}$. In diesem Sinne ist also die folgende Formel unter den Methoden (2.18) optimal:

$$y_{n+1} = y_n + \frac{1}{4} h_n (f(x_n, y_n) + 3 f(x_n + \frac{2}{3} h_n, y_n + \frac{2}{3} h_n f(x_n, y_n))). \tag{2.22}$$

Tabelle 2.2: Verfahren der Ordnung 2

h	verb. Poly.	Heun	optimal	Taylor
1/5	$1.01e+0$	$8.51e-1$	$9.58e-1$	$1.16e+0$
1/10	$4.34e-1$	$3.38e-1$	$4.02e-1$	$5.25e-1$
1/20	$1.47e-1$	$1.07e-1$	$1.34e-1$	$1.86e-1$
1/40	$4.27e-2$	$2.98e-2$	$3.84e-2$	$5.54e-2$
1/80	$1.14e-2$	$7.82e-3$	$1.02e-2$	$1.51e-2$
1/160	$2.96e-3$	$2.00e-3$	$2.64e-3$	$3.92e-3$
1/320	$7.51e-4$	$5.04e-4$	$6.69e-4$	$9.99e-4$
1/640	$1.89e-4$	$1.27e-4$	$1.48e-4$	$2.52e-4$
1/1280	$4.75e-5$	$3.17e-5$	$4.22e-5$	$6.33e-5$

Damit ist nicht gesagt, dass (2.22) unter allen möglichen Formeln (2.18) bei Anwendung auf eine spezielle Differentialgleichung das kleinste führende Fehlerglied hat, da sich in der Entwicklung von

$$y(x+h) - y(x) - h\Phi(x, y(x))$$

Terme mit entgegengesetztem Vorzeichen kompensieren können. Andere Optimalitätskriterien sind denkbar (vgl. Grigorieff [24]).

Beispiel 2.17. Wir betrachten erneut

$$y' = y^2,\ y(0.8) = \frac{5}{6},\ x \in [0.8, 1.8].$$

Dann erhält man mit dem verbesserten Polygonzugverfahren, dem Verfahren von Heun, dem "optimalen" Verfahren aus (2.22) und dem Verfahren aus Beispiel 2.14. die Fehler der Tabelle 2.2. □

Bevor wir Verfahren größerer Ordnung als 2 angeben, schicken wir einige Bemerkungen voraus über die mit einem s-stufigen Verfahren erreichbare Konsistenzordnung. Diese Frage ist keinesfalls leicht zu beantworten, da die beim Taylorabgleich entstehenden Bedingungsgleichungen nichtlinear in den Parametern sind. Eine sorgfältige Untersuchung mit Hilfe von Ordnungsbäumen findet man in *Hairer, Nørsett, Wanner* [29].

Die Zahl der zu erfüllenden Gleichungen steigt mit wachsender Ordnung p sehr stark an, wie die folgende Tabelle zeigt:

Ordnung p	1	2	3	4	5	6	7	8	9	10
Zahl der Gleichungen	1	2	4	8	17	37	85	200	486	1205.

Dabei wurden die Gleichungen

$$\sum_{j=1}^{k-1} \beta_{kj} = \alpha_k, \;\; k = 2, \ldots, m,$$

die wir stets als erfüllt annehmen, nicht mitgezählt.

Gibt man die Ordnung p vor und bestimmt dazu die Stufenzahl s des Runge–Kutta Verfahrens minimal, so gilt der folgende Zusammenhang

p	1	2	3	4	5	6	7	8	9	10
s	1	2	3	4	6	7	9	10	11	12.

Man sieht, dass sich mit wachsender Ordnung das Verhältnis von erreichbarer Ordnung p zur Zahl der dazu nötigen Stufen (also zur Zahl der Funktionsauswertungen in jedem Schritt) verschlechtert.

Entwickelt man (2.17) bis zu Termen mit h^2, so sieht man unmittelbar:

Satz 2.18. *Das Einschrittverfahren (2.17) ist genau dann konsistent, wenn gilt*

$$\sum_{j=1}^{s} \gamma_j = 1.$$

Wir betrachten nun den Fall $s = 4$: Eine etwas mühsame Taylorentwicklung (ähnlich wie im Fall $s = 2$; vgl. Gear [22], p. 32 ff) zeigt, dass man in der Darstellung des lokalen Fehlers durch keine Wahl der Parameter den Koeffizienten bei h^5 unabhängig von f zum Verschwinden bringen kann. Es ist also die Ordnung $p = 4$ erreichbar. Die acht Bedingungen der 10 Parameter (3 Gleichungen zur Bestimmung der Parameter α_j nicht mitgerechnet) lauten:

$$\begin{aligned}
1 &= \gamma_1 + \gamma_2 + \gamma_3 + \gamma_4 \\
\frac{1}{2} &= \alpha_2\gamma_2 + \alpha_3\gamma_3 + \alpha_4\gamma_4 \\
\frac{1}{3} &= \alpha_2^2\gamma_2 + \alpha_3^2\gamma_3 + \alpha_4^2\gamma_4 \\
\frac{1}{4} &= \alpha_2^3\gamma_2 + \alpha_3^3\gamma_3 + \alpha_4^3\gamma_4 \\
\frac{1}{6} &= \alpha_3\beta_{43}\gamma_4 + \alpha_2\beta_{42}\gamma_4 + \alpha_2\beta_{32}\gamma_3 \\
\frac{1}{8} &= \alpha_3\alpha_4\beta_{43}\gamma_4 + \alpha_2\alpha_4\beta_{42}\gamma_4 + \alpha_2\alpha_3\beta_{32}\gamma_3 \\
\frac{1}{12} &= \alpha_3^2\beta_{43}\gamma_4 + \alpha_2^2\beta_{42}\gamma_4 + \alpha_2^2\beta_{32}\gamma_3 \\
\frac{1}{24} &= \alpha_2\beta_{32}\beta_{43}\gamma_4.
\end{aligned}$$

Die Parameter sind durch die vorstehenden Gleichungen nicht eindeutig bestimmt. Wir geben 3 verschiedene Formeln an. Dabei verwenden wir das folgende Koeffiziententableau, mit dem man Runge–Kutta Verfahren auf übersichtliche Weise darstellen kann:

$$\begin{array}{c|cccc} 0 & & & & \\ \alpha_2 & \beta_{21} & & & \\ \alpha_3 & \beta_{31} & \beta_{32} & & \\ \vdots & & & & \\ \alpha_s & \beta_{s1} & \beta_{s2} & \cdots & \beta_{s,s-1} \\ \hline & \gamma_1 & \gamma_2 & \cdots & \gamma_{s-1} \quad \gamma_s. \end{array}$$

Die uns bekannten Verfahren der Ordnung 2 kann man damit so schreiben

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

Verfahren von Heun

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

verb. Polygonzugverfahren

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

optimales Verfahren

Am bekanntesten ist wohl das **klassische Runge–Kutta Verfahren** (1895) der Ordnung 4.

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

Ausführlich geschrieben lautet dieses

$$\begin{aligned} k_1 &= f(x_n, y_n) \\ k_2 &= f\Big(x_n + \frac{h_n}{2}, y_n + \frac{h_n}{2}\, k_1\Big) \\ k_3 &= f\Big(x_n + \frac{h_n}{2}, y_n + \frac{h_n}{2}\, k_2\Big) \\ k_4 &= f(x_n + h_n, y_n + h_n\, k_3) \\ y_{n+1} &= y_n + h_n\, \frac{k_1 + 2k_2 + 2k_3 + k_4}{6}. \end{aligned}$$

Tabelle 2.3: Verfahren der Ordnung 4

h	klassisch	3/8–Regel	Kuntzmann
1/5	$3.52e-2$	$3.42e-2$	$3.59e-2$
1/10	$3.39e-3$	$3.36e-3$	$3.51e-3$
1/20	$2.50e-4$	$2.38e-4$	$2.59e-4$
1/40	$1.65e-5$	$1.43e-5$	$1.67e-5$
1/80	$1.05e-6$	$8.25e-7$	$1.03e-6$
1/160	$6.58e-8$	$4.82e-8$	$6.37e-8$
1/320	$4.12e-9$	$2.89e-9$	$3.94e-9$

Weitere Verfahren der Ordnung 4 sind die **3/8–Regel** (Kutta 1901)

$$\begin{array}{c|cccc} 0 & & & & \\ \frac{1}{3} & \frac{1}{3} & & & \\ \frac{2}{3} & -\frac{1}{3} & 1 & & \\ 1 & 1 & -1 & 1 & \\ \hline & \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8} \end{array}$$

und die **optimale Formel von Kuntzmann** (Kuntzmann 1959)

$$\begin{array}{c|cccc} 0 & & & & \\ \frac{2}{5} & \frac{2}{5} & & & \\ \frac{3}{5} & -\frac{3}{20} & \frac{3}{4} & & \\ 1 & \frac{19}{44} & -\frac{15}{44} & \frac{40}{44} & \\ \hline & \frac{55}{360} & \frac{125}{360} & \frac{125}{360} & \frac{55}{360}. \end{array}$$

Die 3/8-Regel verallgemeinert die Newtonsche 3/8-tel Regel (oder Keplersche Fassregel oder pulcherima) der numerischen Integration. Die Formel von Kuntzmann erhält man ähnlich wie im Falle $s = 2$, indem man die Parameter so bestimmt, dass die Summe der Beträge der Koeffizienten im führenden Fehlerglied von $y(x+h)-y(x)-h\Phi(x,y(x))$ minimal wird.

Die klassische Runge–Kutta Regel ist die bekannteste, die 3/8-Regel häufig die genaueste der expliziten Runge–Kutta Verfahren der Ordnung 4.

Beispiel 2.19. Wir betrachten erneut

$$y' = y^2,\ y(0.8) = \frac{5}{6},\ x \in [0.8, 1.8].$$

Dann erhält man mit dem klassischen Runge–Kutta Verfahren, der 3/8–Regel und dem optimalen Verfahren von Kuntzmann die Fehler der Tabelle 2.3 □

Für Ordnungen $p \geq 5$ hat Butcher ab 1963 ein graphentheoretisches Hilfsmittel eingeführt, die sogenannten Butcher-Bäume , mit denen man die Bestimmungsgleichungen auf systematische Weise aufstellen kann.

Eine Schrittweitensteuerung kann man für die Runge–Kutta Verfahren prinzipiell wie für das Polygonzugverfahren durchführen. Um den Fehler zu schätzen, kann man zwei Schritte mit der halben Schrittweite ausführen. Im Falle der klassischen Runge–Kutta Verfahren hat man dabei die Funktion f an 7 zusätzlichen Punkten auszuwerten, so dass man in jedem Schritt insgesamt 11 Funktionsauswertungen benötigt.

Mit wesentlich weniger Aufwand kommt man bei den **eingebetteten Runge–Kutta Formeln** aus:

Die Idee ist — ähnlich wie bei den Kronrod-Formeln zur Quadratur — von einer Runge–Kutta Formel der Stufe s mit den Zuwächsen $k_1, \ldots, k_s$ und der Ordnung p auszugehen und hierzu bei erhöhter Stufenzahl σ weitere $k_{s+1}, \ldots, k_\sigma$ zu bestimmen, so dass die Formel

$$\tilde{y}_{n+1} = y_n + h_n\Big(\sum_{j=1}^{s} \tilde{\gamma}_j\, k_j + \sum_{j=s+1}^{\sigma} \tilde{\gamma}_j\, k_j\Big)$$

eine höhere Ordnung q als die Ausgangsformel hat.

Dann gilt für die lokalen Fehler $\varepsilon = C\,h^{p+1} + O(h^{p+2})$ und $\tilde{\varepsilon} = O(h^{q+1}) = O(h^{p+2})$, d.h. $\tilde{y}_{n+1} - y_{n+1} = C\,h^{p+1} + O(h^{p+2})$, und hiermit kann man bei vorgegebener Toleranz die optimale Schrittweite wie vorher schätzen.

Man führt also einen Probeschritt der Länge H aus, erhält hieraus

$$C \approx \frac{\tilde{y}_{n+1}(H) - y_{n+1}(H)}{H^{p+1}},$$

und die Forderung $|\varepsilon(x_n, h)| \approx |C| h^{p+1} = \tau$ liefert die neue Schrittweite

$$h = H\left(\frac{\tau}{|\tilde{y}_{n+1}(H) - y_{n+1}(H)|}\right)^{1/(p+1)}.$$

Wir geben einige **eingebettete Formelpaare** an:

Fehlberg (Ordnungen $p = 2$, $q = 3$)

0			
1	1		
$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	
$p = 2$	$\frac{1}{2}$	$\frac{1}{2}$	
$q = 3$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{2}{3}$.

England (Ordnungen $p = 4$, $q = 5$)

$$
\begin{array}{c|ccccccc}
0 & & & & & & & \\
\frac{1}{2} & \frac{1}{2} & & & & & & \\
\frac{1}{2} & \frac{1}{4} & \frac{1}{4} & & & & & \\
1 & 0 & -1 & 2 & & & & \\
\frac{2}{3} & \frac{7}{27} & \frac{10}{27} & 0 & \frac{1}{27} & & & \\
\frac{1}{5} & \frac{28}{625} & -\frac{125}{625} & \frac{546}{625} & \frac{54}{625} & -\frac{378}{625} & & \\
\hline
p = 4 & \frac{1}{6} & 0 & \frac{2}{3} & \frac{1}{6} & 0 & & \\
q = 5 & \frac{14}{336} & 0 & 0 & \frac{35}{336} & \frac{162}{336} & \frac{125}{336}. &
\end{array}
$$

Verner (Ordnungen $p = 5$, $q = 6$)

$$
\begin{array}{c|cccccccc}
0 & & & & & & & & \\
\frac{1}{6} & \frac{1}{6} & & & & & & & \\
\frac{4}{15} & \frac{4}{75} & \frac{16}{75} & & & & & & \\
\frac{2}{3} & \frac{5}{6} & -\frac{8}{3} & \frac{5}{2} & & & & & \\
\frac{5}{6} & -\frac{165}{64} & \frac{55}{6} & -\frac{425}{64} & \frac{85}{96} & & & & \\
1 & \frac{12}{5} & -8 & \frac{4015}{612} & -\frac{11}{36} & \frac{88}{255} & & & \\
\frac{1}{15} & -\frac{8263}{15000} & \frac{124}{75} & -\frac{643}{680} & -\frac{81}{250} & \frac{2484}{10625} & 0 & & \\
1 & \frac{3501}{1720} & -\frac{300}{43} & \frac{297275}{52632} & -\frac{319}{2322} & \frac{24068}{84065} & 0 & \frac{3850}{26703} & \\
\hline
p = 5 & \frac{13}{160} & 0 & \frac{2375}{5984} & \frac{5}{16} & \frac{12}{85} & \frac{3}{44} & 0 & \\
q = 6 & \frac{3}{40} & 0 & \frac{875}{2244} & \frac{23}{72} & \frac{264}{1955} & 0 & \frac{125}{11592} & \frac{43}{616}.
\end{array}
$$

Wie beim Polygonzugverfahren mit Schrittweitensteuerung verwendet man bei diesen sehr frühen Formelpaaren im weiteren Verlauf der Integration die Näherung für das Verfahren höherer Ordnung, obwohl der Fehler für das Verfahren niedrigerer Ordnung geschätzt wurde. Für diese Näherung liegt dann zwar keine Fehlerschätzung vor, aber ein Anwender wird sich wohl kaum beschweren, wenn der Code eine höhere Genauigkeit liefert als die Geforderte.

Neuere Formelpaare wurden so konstruiert, dass die neue Näherung y_{n+1} in x_{n+1} so gewählt ist, dass das im nächsten Schritt benötigte $k_1 = f(x_{n+1}, y_{n+1})$ bereits im letzten Schritt berechnet wurde. Verfahren dieses Typs werden als **FSAL-Verfahren** (*first* evaluation of f for the next step *same as last* of the current step). Bei ihnen bekommt man in jedem erfolgreichen Schritt eine Auswertung der rechten Seite geschenkt.

Das bekannteste und wohl am häufigsten verwendete Verfahren, das auch in der ODE-Suite von MATLAB als ODE45 implementiert ist, ist das Formelpaar von

Dormand und Prince [18], das Runge-Kutta Verfahren der Ordnungen 5 und 4 kombiniert.

Für etwas schwächere Genauigkeitsanforderungen wird eine Kombination von Verfahren der Ordnungen 3 und 2 von Bogacki und Shampine [6] in der ODE-Suite von MATLAB als ODE23 bereitgestellt.

Bogacki und Shampine (Ordnungen $p = 3$, $q = 2$)

0				
$\frac{1}{2}$	$\frac{1}{2}$			
$\frac{3}{4}$		$\frac{3}{4}$		
1	$\frac{2}{9}$	$\frac{1}{3}$	$\frac{4}{9}$	
$p = 3$	$\frac{2}{9}$	$\frac{1}{3}$	$\frac{4}{9}$	
$q = 2$	$\frac{7}{24}$	$\frac{1}{4}$	$\frac{1}{3}$	$\frac{1}{8}$.

Dormand und Prince (Ordnungen $p = 5$, $q = 4$)

0							
$\frac{1}{5}$	$\frac{1}{5}$						
$\frac{3}{10}$	$\frac{3}{40}$	$\frac{9}{40}$					
$\frac{4}{5}$	$\frac{44}{45}$	$-\frac{56}{15}$	$\frac{32}{9}$				
$\frac{8}{9}$	$\frac{19372}{6561}$	$-\frac{25360}{2187}$	$\frac{64448}{6561}$	$-\frac{212}{729}$			
1	$\frac{9017}{3168}$	$-\frac{355}{33}$	$\frac{46732}{5247}$	$\frac{49}{176}$	$-\frac{5103}{18656}$		
1	$\frac{35}{384}$	0	$\frac{500}{1113}$	$\frac{125}{192}$	$-\frac{2187}{6784}$	$\frac{11}{84}$	
$p = 5$	$\frac{35}{384}$	0	$\frac{500}{1113}$	$\frac{125}{192}$	$-\frac{2187}{6784}$	$\frac{11}{84}$	0
$q = 4$	$\frac{5179}{57600}$	0	$\frac{7571}{16695}$	$\frac{393}{640}$	$-\frac{92097}{339200}$	$\frac{187}{2100}$	$\frac{1}{40}$.

Beispiel 2.20.

$$y' = y^2, \;\; y(0.8) = \frac{5}{6}.$$

Um die Lösung $y(x) = \dfrac{1}{2-x}$ im Punkte $b = 1.8$ mit höchstens dem absoluten Fehler 0.0005 zu approximieren, benötigt man mit dem Formelpaar von Fehlberg 138 Funktionsauswertungen, mit dem von England 96, mit dem von Verner 84, mit dem Paar von Dormand und Prince 49 und mit dem Paar von Bogacki und Shampine 187 Funktionsauswertungen. □

2.3 Extrapolationsverfahren

Wir betrachten das Einschrittverfahren

$$y_{n+1} = y_n + h\Phi(x_n, y_n, h) \tag{2.23}$$

der Ordnung p. Ist die Schrittfunktion Φ genügend oft differenzierbar (dies ist bei Runge–Kutta Methoden eine Forderung an die Glattheit der rechten Seite f der Differentialgleichung (2.1)), so gilt für den lokalen Fehler

$$y(x+h) - y(x) - h\Phi(x, y(x), h) = \sum_{j=p+1}^{N+1} d_j(x)h^j + O(h^{N+2}). \tag{2.24}$$

Es sei nun $y_h(x) := y_n$ die Näherung von $y(a + nh)$, die man mit n äquidistanten Schritten mit dem Verfahren (2.23) erhält. Dann hat der globale Fehler die Größe $O(h^p)$. Wir zeigen, dass es eine Funktion e_p gibt, so dass der globale Fehler geschrieben werden kann als

$$y_h(x) - y(x) = e_p(x)h^p + O(h^{p+1}). \tag{2.25}$$

Dazu betrachten wir die Funktion

$$\hat{y}_h(x) := y_h(x) - e_p(x)h^p \tag{2.26}$$

als die Näherung, die man mit n Schritten eines Einschrittverfahrens

$$\hat{y}_{n+1} = \hat{y}_n + h\hat{\Phi}(x_n, \hat{y}_n, h) \tag{2.27}$$

erhalten hat. Es gilt

$$\begin{aligned}
\hat{y}_{n+1} &= y_{n+1} - e_p(x_n + h)h^p \\
&= y_n + h\Phi(x_n, y_n, h) - e_p(x_n + h)h^p \\
&= \hat{y}_n + h\Phi(x_n, \hat{y}_n + e_p(x_n)h^p, h) - e_p(x_n + h)h^p + e_p(x_n)h^p,
\end{aligned}$$

und daher ist die Schrittfunktion

$$\hat{\Phi}(x, \hat{y}, h) = \Phi(x, \hat{y} + e_p(x)h^p, h) - (e_p(x + h) - e_p(x))h^{p-1}. \tag{2.28}$$

Für den lokalen Fehler erhält man

$$\begin{aligned}
&y(x+h) - y(x) - h\hat{\Phi}(x, y(x), h) \\
&= y(x+h) - y(x) - h\Phi(x, y(x) + e_p(x)h^p, h) + (e_p(x + h) - e_p(x))h^p \\
&= y(x+h) - y(x) - h\Big(\Phi(x, y(x), h) + \frac{\partial}{\partial y}\Phi(x, y(x), h)e_p(x)h^p + O(h^{p+1})\Big) \\
&\quad + (e_p(x + h) - e_p(x))h^p \\
&= d_{p+1}(x)h^{p+1} - \frac{\partial}{\partial y}\Phi(x, y(x), 0)e_p(x)h^{p+1} + e_p'(x)h^{p+1} + O(h^{p+2}),
\end{aligned}$$

und wegen

$$\Phi(x, y(x), 0) = \lim_{h \to 0} \frac{y(x+h) - y(x)}{h} = y'(x) = f(x, y(x)),$$

d.h.

$$\frac{\partial}{\partial y} \Phi(x, y(x), 0) = \frac{\partial}{\partial y} f(x, y(x))$$

folgt schließlich

$$\begin{aligned} & y(x+h) - y(x) - h\hat{\Phi}(x, y(x), h) \\ &= \Big(d_{p+1}(x) - \frac{\partial}{\partial y} f(x, y(x)) e_p(x) + e_p'(x)\Big) h^{p+1} + O(h^{p+2}). \end{aligned} \tag{2.29}$$

Damit wird durch (2.27) ein Verfahren der Ordnung $p+1$ definiert, falls e_p Lösung der Anfangswertaufgabe

$$e_p'(x) = \frac{\partial}{\partial y} f(x, y(x)) e_p(x) - d_{p+1}(x), \;\; e_p(a) = 0, \tag{2.30}$$

ist. Nach Satz 2.9. gilt in diesem Fall

$$\hat{y}_h - y(x) = O(h^{p+1}),$$

d.h.

$$y_h(x) - y(x) = e_p(x) h^p + O(h^{p+1}). \tag{2.31}$$

Das obige Vorgehen kann man mit dem Einschrittverfahren (2.27) mit der Schrittfunktionen $\hat{\Phi}$ wiederholen. Es hat die Konsistenzordnung $p+1$ und erfüllt

$$\frac{\partial}{\partial y} \hat{\Phi}(x, y(x), 0) = \frac{\partial}{\partial y} f(x, y(x)).$$

Man erhält schließlich

Satz 2.21. *Die Schrittfunktion Φ des Einschrittverfahrens (2.23) sei hinreichend glatt, und es gelte für den lokalen Fehler die Entwicklung (2.24). Dann besitzt der globale Fehler eine asymptotische Entwicklung*

$$y_h(x) - y(x) = \sum_{j=p}^{N} e_j(x) h^j + E_h(x) h^{N+1}, \tag{2.32}$$

wobei die e_j Lösungen linearer Anfangswertaufgaben der Gestalt (2.30) sind und E_h beschränkt auf $[a, b]$ ist.

Satz 2.21. wurde nur für äquidistante Schritte der Länge h bewiesen. Bei unterschiedlichen Schrittlängen nehmen wir an, dass es eine Funktion $\tau : [a,b] \to \mathbb{R}^+$ gibt, so dass die Schrittlängen

$$x_{n+1} - x_n = \tau(x_n)h \tag{2.33}$$

von einem Parameter h abhängen. Dann geht die Entwicklung des lokalen Fehlers über in

$$\begin{aligned} & y(x+\tau(x)h) - y(x) - h\tau(x)\Phi(x, y(x), \tau(x)h) \\ = & \sum_{j=p+1}^{N+1} d_j(x)\tau(x)^j h^j + O(h^{N+2}), \end{aligned} \tag{2.34}$$

und mit ähnlichen Schlüssen wie oben erhält man, dass Satz 2.21. auch bei variabler Schrittweite richtig bleibt.

Satz 2.21. ist die Grundlage für Extrapolationsverfahren. Wir betrachten das Einschrittverfahren (2.23) für die Anfangswertaufgabe (2.1). Es sei $H > 0$ eine Basisschrittlänge. Wir wählen eine Folge positiver ganzer Zahlen

$$n_1 < n_2 < n_3 < \dots \tag{2.35}$$

und definieren hierzu die Schrittweiten

$$h_j := \frac{H}{n_j}.$$

Eine häufige Wahl ist die Romberg Folge $1, 2, 4, 8, 16, 32, \dots$, oder die Bulirsch Folge $1, 2, 3, 4, 6, 8, 12, 16, 24, 32, \dots$. Diese beiden Folgen haben sich bei der Extrapolation zur numerischen Integration bewährt, da man dann mit der zusammengesetzten Trapezregel auf jeweils vorher berechnete Trapezsummen zurückgreifen kann, und so ohne großen zusätzlichen Aufwand Näherungen für kleine Schrittweiten bestimmen kann. Bei Anfangswertaufgaben ist dies nicht der Fall, und hier ist häufig die harmonische Folge

$$1, 2, 3, 4, 5, 6, \dots$$

die beste Wahl, da hiermit die Zahl der Funktionsauswertungen am langsamsten steigt.

Ist eine Näherung y_n für $y(x_n)$ bereits bekannt, so berechnen wir mit n_j Schritten der äquidistanten Schrittweite h_j Näherungen

$$y_{h_j}(x_n + H) =: T_{j,1} \tag{2.36}$$

für $y(x_n + H)$ und benutzen diese, um durch Extrapolation eine bessere Näherung für $y(x_n + H)$ zu bestimmen. Dazu berechnen wir das Interpolationspolynom

$$p(h) = \eta_0 + \eta_p h^p + \eta_{p+1} h^{p+1} + \ldots + \eta_{p+k-2} h^{p+k-2}, \tag{2.37}$$

das die Bedingungen

$$p(h_i) = T_{i,1},\ i = j, j-1, j-2, \ldots, j-k+1, \tag{2.38}$$

erfüllt, und wählen

$$p(0) = \eta_0 =: T_{j,k} \tag{2.39}$$

als neue Näherung. (2.38) ist ein lineares System von k Gleichungen zur Bestimmung der k Unbekannten η_0, η_p,..., η_{p+k-2}.

Offensichtlich kann $T_{j,k}$ als Näherungswert aus einem Einschrittverfahren mit der Schrittweite H gedeutet werden.

Satz 2.22. *Besitzt das Basisverfahren (2.23) die Konsistenzordnung p, so hat das Einschrittverfahren, das $T_{j,k}$ durch Interpolation bestimmt, die Konsistenzordnung $p + k - 1$.*

Beweis: Berechnet man die $T_{i,1}$ ausgehend vom Punkt $(x, y(x))$, so gilt nach Satz 2.21. (mit $N = p + k - 1$)

$$T_{i,1} = y(x + H) + e_p(x + H)h_i^p + \ldots + e_{p+k-2}(x + H)h_i^{p+k-2} + \Delta_i \tag{2.40}$$

mit

$$\Delta_i = e_{p+k-1}(x + H)h_i^{p+k-1} + E_{h_i}(x + H)h_i^{p+k}.$$

Wegen $e_{p+k-1}(x) = 0$ gilt $e_{p+k-1}(x + H) = O(H)$, und wegen $h_i \leq H$ folgt

$$\Delta_i = O(H^{p+k}).$$

Ferner gilt nach (2.37) und (2.38)

$$T_{i,1} = \eta_0 + \eta_p h_i^p + \ldots + \eta_{p+k-2} h_i^{p+k-2}.$$

Daher gilt

$$(\eta_0 - y(x+H)) + (\eta_p - e_p(x+H))\Big(\frac{H}{n_i}\Big)^p + \ldots + (\eta_{p+k-2} - e_{p+k-2}(x+H))\Big(\frac{H}{n_i}\Big)^{p+k-2} = \Delta_i. \tag{2.41}$$

(2.41) ist ein lineares Gleichungssystem von k Gleichungen zur Bestimmung der k Unbekannten $\eta_0 - y(x+H)$, $(\eta_p - e_p(x+H))H^p, \ldots, (\eta_{p+k-2} - e_{p+k-2}(x+H))H^{p+k-2}$, und die Koeffizientenmatrix hat die Gestalt

$$\boldsymbol{A} = \begin{pmatrix} 1 & 1/n_j^p & 1/n_j^{p+1} & \ldots & 1/n_j^{p+k-2} \\ 1 & 1/n_{j-1}^p & 1/n_{j-1}^{p+1} & \ldots & 1/n_{j-1}^{p+k-2} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & 1/n_{j-k+1}^p & 1/n_{j-k+1}^{p+1} & \ldots & 1/n_{j-k+1}^{p+k-2} \end{pmatrix}.$$

Mit ähnlichen Argumenten wie in Mathematik I für die Vandermondesche Determinante sieht man ein, dass die Matrix $\boldsymbol{A}$ regulär ist. Daher folgt

$$|\eta_0 - y(x+H)| \leq \|\boldsymbol{A}^{-1}\|_\infty \max_i |\Delta_i| = O(H^{k+p}).$$

∎

Ist $T_{i,1}$ gegeben, so kann man die $T_{j,k}$ mit dem Neville–Aitken Schema

$$T_{j,k+1} = T_{j,k} + \frac{T_{j,k} - T_{j-1,k}}{(n_j/n_{j-k}) - 1} \tag{2.42}$$

(vgl. Mathematik II) auf effiziente Weise bestimmen. Man erhält die Werte

$$\begin{pmatrix} T_{11} & & & \\ T_{21} & T_{22} & & \\ T_{31} & T_{32} & T_{33} & \\ T_{41} & T_{42} & T_{43} & T_{44} \\ \ldots & \ldots & \ldots & \ldots \end{pmatrix}.$$

Ein großer Vorteil der Extrapolationsmethoden ist, dass in der ersten Spalte Näherungen der Ordnung p stehen, in der zweiten der Ordnung $p+1$, in der dritten der Ordnung $p+2$ usw. Man kann die Spalten als die Ergebnisse eingebetteter Verfahren auffassen. Damit erhält man eine billige Möglichkeit der Schrittweitenkontrolle.

Wählt man eine große Zahl von Spalten in dem Neville–Aitken Schema, also eine große Ordnung, so hat man als Preis zu zahlen, dass man sehr viele Funktionsauswertungen für einen Schritt vornehmen muss. Andererseits wird man bei großer Konsistenzordnung große Schritte ausführen können, um eine vorgegebene Genauigkeit zu erreichen. Ist A_k die Zahl der nötigen Funktionsauswertungen, um $T_{k,k}$ auszuwerten und H_k die optimale Schrittlänge im nächsten Schritt, so ist

$$W_k := \frac{A_k}{H_k}$$

die Arbeit pro Einheitsschritt. Man wird nun nicht nur die Schrittlänge optimal wählen, sondern zugleich die Ordnung des Verfahrens, so dass W_k minimal wird.

Methoden zur Ordnungs- und Schrittweitenkontrolle sind in Hairer, Nørsett, Wanner [29] beschrieben und in dem FORTRAN77 Code ODEX umgesetzt.

2.4 Software zu Einschrittverfahren

Das angegebene Verfahren von Dormand und Prince wird für geringe Genauigkeitsanforderungen (relativer Fehler: $10^{-3} - 10^{-7}$) als die effizienteste Methode angesehen. Für höhere Genauigkeitsanforderungen (relativer Fehler: $10^{-7} - 10^{-14}$) gilt ein Verfahren der Ordnung 8, das ebenfalls von Dormand und Prince publiziert wurde, mit einem Fehlerschätzer der Ordnung 5 und einer Korrektur der Ordnung 3, die von Hairer und Wanner vorgestellt wurden, als geeignet. Für sehr hohe Genauigkeitsanforderungen muss man Extrapolationsverfahren verwenden.

FORTRAN77 Codes der Methoden von Dormand und Prince und das Extrapolationsverfahren ODEX sind in dem Buch von *Hairer, Nørsett, Wanner* [29] abgedruckt. Man kann diese (oder auch C Versionen) erhalten unter

http://www.unige.ch/ hairer/software.html

In MATLAB sind als eingebettete Runge–Kutta Methoden für nicht-steife Anfangswertaufgaben das Verfahren von Dormand und Prince der Ordnung 5 als ODE45 und für den Fall, dass keine sehr hohe Genauigkeit benötigt wird, das Verfahren von Bogacki und Shampine (vgl. *Shampine* [47]) der Ordnung 3 als ODE23 implementiert.

ODEX wurde von D. Abendroth in MATLAB übertragen und kann von

http://www.tu-harburg.de/ins/download/software.html

geladen werden.

Weitere Sortware findet man in der Spiegelung der NETLIB in der elib des ZIB

http://elib.zib.de/netlib

in den Unterverzeichnissen ‘ode’ und ‘odepack’.

Das Konrad Zuse Zentrum für Informationsverarbeitung (ZIB), Berlin, stellt das ODELab zur Verfügung. Es enthält eine Reihe von Lösern für Anfangswertaufgaben, mit denen die Verfahren getestet werden können. Das ODELab ist unter

http://num-lab.zib.de/public/odelab/

zugänglich.

Kapitel 3

Mehrschrittverfahren

Ein weiterer, häufig benutzter Verfahrenstyp zur numerischen Lösung der Anfangswertaufgabe

$$y' = f(x, y),\ y(a) = y_0 \tag{3.1}$$

sind die linearen **Mehrschrittverfahren**, bei denen man zur Berechnung der Näherung y_{n+k} die bereits ermittelten Näherungen $y_{n+k-1}, y_{n+k-2}, \ldots, y_n$ verwendet. Dazu macht man den Ansatz

$$\sum_{\nu=0}^{k} a_\nu\, y_{n+\nu} = h \sum_{\nu=0}^{k} b_\nu\, f_{n+\nu} \tag{3.2}$$

mit $f_{n+\nu} := f(x_{n+\nu}, y_{n+\nu})$, wobei $a_k \neq 0$ vorausgesetzt wird.

Ist $b_k \neq 0$, so kommt y_{n+k} auf beiden Seiten von (3.2) vor. Es muss dann in jedem Schritt ein (i.a. nichtlineares) Gleichungssystem gelöst werden, um y_{n+k} zu bestimmen. In diesem Fall heißt (3.2) **implizites k-Schritt Verfahren**. Ist $b_k = 0$, so kann man (3.2) sofort nach y_{n+k} auflösen. In diesem Fall heißt (3.2) **explizites k-Schritt Verfahren**.

Offenbar ist der erste Wert, den man mit (3.2) berechnen kann, y_k. Neben dem gegebenen Wert y_0 müssen also zunächst Näherungen $y_1, \ldots, y_{k-1}$ für $y(x_1), \ldots, y(x_{k-1})$ zur Verfügung gestellt werden. Diese können z.B. mit einem Runge-Kutta Verfahren berechnet werden.

Wegen $a_k \neq 0$ können wir o.B.d.A. $a_k = 1$ annehmen. Wir bestimmen die übrigen a_ν, b_ν nun so, dass (3.2) zu einem brauchbaren Verfahren wird.

Den lokalen Fehler von (3.2) erhält man wieder, indem man die exakte Lösung $y(x)$ von (3.1) in (3.2) einsetzt:

Definition 3.1. *Der* **lokale Fehler** *des k-Schritt Verfahrens (3.2) ist definiert durch*

$$\varepsilon_n := \sum_{\nu=0}^{k} a_\nu\, y(x_n+\nu h) - h\sum_{\nu=0}^{k} b_\nu\, y'(x_n+\nu h). \tag{3.3}$$

wobei y die Lösung der Anfangswertaufgabe (3.1) bezeichnet.

Das Verfahren (3.2) heißt **konsistent**, *wenn $\varepsilon_n(h) = o(h)$ gilt, es heißt* **von der Ordnung** *p, wenn $\varepsilon_n(h) = O(h^{p+1})$ für alle genügend glatten Funktionen y gilt.*

Ist y $(m+1)$-mal differenzierbar, so liefert der Taylorsche Satz

$$\begin{aligned}\varepsilon_n \;=\; & \sum_{\nu=0}^{k} a_\nu\Big(\sum_{\mu=0}^{m}\frac{y^{(\mu)}(x_n)}{\mu!}\,\nu^\mu h^\mu + \frac{1}{(m+1)!}\,y^{(m+1)}(x_n+\theta_\nu\nu h)\,\nu^{m+1}h^{m+1}\Big)\\ & -\sum_{\nu=0}^{k} b_\nu\Big(\sum_{\mu=0}^{m-1}\frac{y^{(\mu+1)}(x_n)}{\mu!}\,\nu^\mu h^{\mu+1} + \frac{1}{m!}\,y^{(m+1)}(x_n+\hat\theta_\nu\nu h)\,\nu^m h^{m+1}\Big).\end{aligned}$$

Damit das Verfahren konsistent ist, müssen sich die Glieder mit dem Faktor h^0 und mit dem Faktor h^1 jeweils gegenseitig aufheben, d.h. es muss gelten

$$\sum_{\nu=0}^{k} a_\nu = 0, \quad \sum_{\nu=0}^{k}(\nu a_\nu - b_\nu) = 0. \tag{3.4}$$

Das Nächstliegende ist nun, die a_ν, b_ν so zu bestimmen, dass in ε_n möglichst hohe Potenzen von h abgeglichen werden.

Dies führt zu einem linearen (wegen $a_k = 1$ inhomogenen) Gleichungssystem.

Für $k = 2$ erhält man

$$\begin{array}{rcrcrcrcll} a_0 & + & a_1 & & & & & = & -1 & \text{(siehe (3.4))}\\ & & a_1 & - & b_0 & - & b_1 & - & b_2 = & -2 \quad \text{(siehe (3.4))}\\ & & a_1 & & & - & 2b_1 & - & 4b_2 = & -4\\ & & a_1 & & & - & 3b_1 & - & 12b_2 = & -8\\ & & a_1 & & & - & 4b_1 & - & 32b_2 = & -16 \end{array} \tag{3.5}$$

mit der Lösung $a_0 = -1$, $a_1 = 0$, $b_0 = \dfrac{1}{3}$, $b_1 = \dfrac{4}{3}$, $b_2 = \dfrac{1}{3}$.

Man erhält also das implizite Verfahren der Ordnung 4:

$$y_{n+2} = y_n + \frac{h}{3}\,(f_n + 4f_{n+1} + f_{n+2}).$$

Verlangt man, um die Auflösung einer nichtlinearen Gleichung (bzw. eines nichtlinearen Gleichungssystems) in y_{n+2} in jedem Schritt zu vermeiden, $b_2 = 0$, d.h. ein explizites Verfahren, so kann man nur die ersten vier Gleichungen von (3.5) erfüllen.

Tabelle 3.1: AWA mit Anfangsfehler 10^{-15} ($y \equiv 1$)

n	y_n	n	y_n
0	1.00000000000000000	16	0.99997176556842504
1	1.00000000000000111	17	1.00014117215787590
2	0.99999999999999556	18	0.99929413921062160
3	1.00000000000002331	19	1.00352930394689310
4	0.99999999999988454	20	0.98235348026553560
5	1.00000000000057843	21	1.08823259867232314
6	0.99999999999710898	22	0.55883700663838543
7	1.00000000001445621	$\vdots$	$\cdots\cdots\cdots\cdots\cdots\cdots$
8	0.99999999992772004	34	$-1.077058079E+0008$
9	1.00000000036140091	35	$5.385290456E+0008$
10	0.99999999819299656		

Lösung hiervon ist $a_0 = -5$, $a_1 = 4$, $b_0 = 2$, $b_1 = 4$, und man erhält das explizite Verfahren der Ordnung 3:

$$y_{n+2} = -4y_{n+1} + 5y_n + 2h(f_n + 2f_{n+1}). \tag{3.6}$$

Wendet man (3.6) auf die Anfangswertaufgabe $y' = 0$, $y(0) = 1$ mit dem (z.B. durch Rundungsfehler verfälschten) Anfangsfeld $y_0 = 1$, $y_1 = 1 + 10^{-15}$ an, so erhält man Tabelle 3.1. Kleinste Anfangsfehler schaukeln sich also auf und machen das Verfahren trotz der Ordnung 3 völlig unbrauchbar.

Die Fehlerordnung allein ist also kein geeignetes Mittel zur Bewertung eines Mehrschrittverfahrens.

Für den Fall $f(x, y) \equiv 0$ lautet (3.6)

$$y_{n+2} + 4y_{n+1} - 5y_n = 0. \tag{3.7}$$

Dies ist eine lineare homogene Differenzengleichung mit konstanten Koeffizienten. Der Ansatz $y_n = \lambda^n$ für eine Lösung von (3.7) führt auf die Bedingung

$$\lambda^{n+2} + 4\lambda^{n+1} - 5\lambda^n = 0$$

d.h. $\lambda^2 + 4\lambda - 5 = 0$ mit den Lösungen $\lambda_1 = 1$, $\lambda_2 = -5$.

Da (3.7) linear und homogen ist, ist auch

$$y_n = A\lambda_1^n + B\lambda_2^n = A + B(-5)^n, \quad A, B \in \mathbb{R},$$

eine Lösung (sogar die allgemeine Lösung). Die Konstanten A und B kann man aus dem Anfangsfeld y_0 und y_1 bestimmen. Man erhält

$$y_n = \frac{1}{6}(5y_0 + y_1) + (-5)^n \frac{1}{6}(y_0 - y_1).$$

Der zweite Term hiervon führt dazu, dass sich die Fehler (bei alternierendem Vorzeichen) aufschaukeln.

Im allgemeinen Fall (3.2) hätte man statt (3.7) für $f(x,y) \equiv 0$ die Differenzengleichung $\sum_{\nu=0}^{k} a_\nu \, y_{n+\nu} = 0$ mit der charakteristischen Gleichung

$$\rho(\lambda) := \sum_{\nu=0}^{k} a_\nu \, \lambda^\nu = 0.$$

Sind $\lambda_1, \ldots, \lambda_r$ die verschiedenen Nullstellen von ρ mit den Vielfachheiten $m_1, \ldots, m_r$, so sind alle Lösungen von

$$\sum_{\nu=0}^{k} a_\nu \, y_{n+\nu} = 0$$

Linearkombinationen von $\lambda_j^n, n\lambda_j^n, \ldots, n^{m_j-1}\lambda_j^n$, $j = 1, \ldots, r$ (vgl. die allgemeine Lösung der homogenen Differentialgleichung mit konstanten Koeffizienten).

Fehler im Anfangsfeld $y_0, \ldots, y_{k-1}$ werden daher nicht verstärkt, wenn $|\lambda_j| \leq 1$ für alle Nullstellen λ_j von ρ gilt und die Nullstellen mit $|\lambda_j| = 1$ einfach sind.

Definition 3.2. *Das Mehrschrittverfahren (3.2) heißt* **stabil***, wenn für alle Nullstellen λ_j des charakteristischen Polynoms $\rho(\lambda)$ gilt $|\lambda_j| \leq 1$ und wenn die Nullstellen mit $|\lambda_j| = 1$ einfach sind.*

Wegen der Konsistenzbedingung ist stets $\lambda = 1$ eine Nullstelle von ρ. Gilt $|\lambda_j| < 1$ für alle anderen Nullstellen λ_j von ρ, so heißt das Verfahren **stark stabil**.

Die obigen Überlegungen zeigen, dass die Stabilität neben der Konsistenz die Mindestanforderung an ein k-Schritt Verfahren ist. Umgekehrt kann man zeigen, dass konsistente und stabile Verfahren konvergieren (vgl. *Hairer, Nørsett, Wanner*[29], p. 395). Dabei müssen wir die Definition der Konvergenz gegenüber den Einschrittverfahren nun ein wenig modifizieren, da das Verfahren (3.2) nicht nur von dem in (3.1) gegebenen Anfangswert y_0 abhängt, sondern auch von dem gewählten Anfangsfeld $y_1, \ldots, y_{k-1}$.

Definition 3.3. *Das lineare k-Schritt Verfahren (3.2) heißt* **konvergent** *, wenn für jedes Anfangswertproblem (3.1)*

$$y(x) - y_n \to 0 \quad \textit{für } h \to 0 \textit{ und } n \to \infty \textit{ mit } x_0 + nh \to x$$

gilt für alle Anfangsfelder $y_1(h), \ldots, y_{k-1}(h)$ mit

$$y(x_0 + jh) - y_j(h) \to 0 \quad \textit{für } h \to 0, \; j = 1, \ldots, k-1.$$

Das Verfahren heißt konvergent von der Ordnung p, wenn für jede Anfangswertaufgabe (3.1) mit genügend glatter rechter Seite f es ein $h_0 > 0$ gibt mit

$$\|y(x_0 + jh) - y_j(h)\| \leq Ch^p \quad \textit{für } h \leq h_0$$

für alle Anfangsfelder mit

$$\|y(x_0 + jh) - y_j(h)\| \leq C_0 h^p \quad \textit{für } h \leq h_0 \textit{ und } j = 1, \ldots, k-1.$$

Fordert man in (3.5) neben $b_2 = 0$ (Explizitheit), dass $\rho(\lambda)$ die Nullstellen $\lambda_1 = 1$ (Konsistenz) und $\lambda_2 = 0$ (um die Stabilität zu erzwingen) besitzt, so kann man nur die ersten drei Gleichungen von (3.5) erfüllen und erhält das explizite Verfahren der Ordnung 2:

$$y_{n+2} = y_{n+1} + \frac{h}{2}(-f_n + 3f_{n+1}).$$

Wir geben nun einen Weg an, wie man stark stabile Mehrschrittverfahren konstruieren kann. In Übereinstimmung mit der Literatur numerieren wir dabei nun die an der Mehrschrittformel beteiligten Näherungswerte mit $y_{n-k+1}, \ldots, y_n, y_{n+1}$, wobei $y_{n-k+1}, \ldots, y_n$ als aus den vorhergehenden Schritten bekannt angenommen werden und y_{n+1} in dem aktuellen Schritt zu bestimmen ist.

Für die Lösung y der Anfangswertaufgabe (3.1) gilt

$$y(x_{n+1}) - y(x_n) = \int_{x_n}^{x_{n+1}} y'(t)\,dt = \int_{x_n}^{x_{n+1}} f(t, y(t))\,dt. \tag{3.8}$$

Wir ersetzen daher bei gegebenen Näherungen $y_j \approx y(x_j)$, $j = n, n-1, \ldots, n-k+1$, und damit bekannten Näherungen

$$f_j := f(x_j, y_j) \approx f(x_j, y(x_j)) = y'(x_j),\ j = n, n-1, \ldots, n-k+1,$$

die Funktion y' im Integranden durch ihr Interpolationspolynom

$$p \in \Pi_{k-1}\ :\ p(x_j) = f_j,\ j = n, n-1, \ldots, n-k+1,$$

und berechnen die neue Näherung gemäß

$$y_{n+1} = y_n + \int_{x_n}^{x_{n+1}} p(t)\,dt.$$

An der Lagrangeschen Integrationsformel

$$p(x) = \sum_{j=0}^{k-1} f_{n-j} \cdot \ell_j(x),\ \ell_j(x) := \prod_{\substack{i=0 \\ i \neq j}}^{k-1} (x - x_{n-i}) \Big/ \prod_{\substack{i=0 \\ i \neq j}}^{k-1} (x_{n-j} - x_{n-i}),$$

erkennt man, dass

$$y_{n+1} = y_n + \sum_{j=0}^{k-1} f_{n-j} \int_{x_n}^{x_{n+1}} \ell_j(t)\,dt$$

tatsächlich die Gestalt eines k–Schritt Verfahrens hat.

Mit der Variablentransformation $t := x_n + h \cdot s$ erhält man

$$\int_{x_n}^{x_{n+1}} \ell_j(t)\,dt = h \cdot \int_0^1 \prod_{\substack{i=0 \\ i\neq j}}^{k-1} (i+s) \Big/ \prod_{\substack{i=0 \\ i\neq j}}^{k-1} (i-j)\,ds =: h\alpha_j.$$

Die Integrale über das Interpolationspolynom lassen sich also schreiben als

$$\int_{x_n}^{x_{n+1}} p(t)\,dt = h \cdot \sum_{j=0}^{k-1} \alpha_j f_{n-j},$$

wobei die Koeffizienten α_j unabhängig von den y_j und von den speziellen Knoten x_j und der Schrittweite h sind, und daher in Dateien bereitgestellt werden können.

Die Mehrstellenformel erhält damit die Gestalt

$$y_{n+1} = y_n + h \cdot \sum_{j=0}^{k-1} \alpha_j f_{n-j}.$$

Das charakteristische Polynom ist (man beachte die geänderte Numerierung der Koeffizienten in der k-Schritt Formel)

$$\rho(\lambda) = \lambda^k - \lambda^{k-1}$$

mit der einfachen Nullstelle $\lambda = 1$ und der $(k-1)$-fachen Nullstelle 0. Die Mehrstellenformel ist also stark stabil.

So konstruierte Mehrstellenformeln heißen **Adams–Bashforth Verfahren**. Sie sind explizit und aus der Fehlerdarstellung des Interpolationspolynoms erhält man, dass ihre Ordnung k ist. Die ersten Adams–Bashforth Formeln sind:

$$\begin{aligned}
k = 1: &\quad y_{n+1} = y_n + hf_n \\
k = 2: &\quad y_{n+1} = y_n + 0.5h(3f_n - f_{n-1}) \\
k = 3: &\quad y_{n+1} = y_n + h(23f_n - 16f_{n-1} + 5f_{n-2})/12 \\
k = 4: &\quad y_{n+1} = y_n + h(55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3})/24 \\
k = 5: &\quad y_{n+1} = y_n + h(1901f_n - 2774f_{n-1} + 2616f_{n-2} - 1274f_{n-3} + 251f_{n-4})/720
\end{aligned}$$

Für $k = 1$ ergibt sich also das Eulersche Polygonzugverfahren.

Nachteil der Adams–Bashforth Formeln ist, dass bei ihrer Konstruktion das Interpolationspolynom p im Intervall $[x_n, x_{n+1}]$ verwendet wird, während die Interpolationsknoten außerhalb dieses Intervalls liegen. Wir wissen bereits, dass der Fehler

eines Interpolationspolynoms außerhalb des kleinsten Intervalls $[x_{n-k+1}, x_n]$, das alle Knoten enthält, sehr schnell anwächst. Es ist daher naheliegend, die Funktion y' in (3.8) durch das Interpolationspolynom

$$p \in \Pi_k \ : \ p(x_j) = f(x_j, y_j), j = n+1, n, n-1, \ldots, n-k+1$$

zu ersetzen.

Wie eben kann man das Verfahren schreiben als

$$y_{n+1} = y_n + h \sum_{j=0}^{k} \beta_j f_{n+1-j}$$

mit

$$\beta_j := \frac{1}{h} \int_{x_n}^{x_{n+1}} \prod_{\substack{i=0 \\ i \neq j}}^{k} (t - x_{n+1-i}) \Big/ \prod_{\substack{i=0 \\ i \neq j}}^{k} (x_{n+1-j} - x_{n+1-i})\, dt.$$

Diese Verfahren heißen **Adams–Moulton Verfahren**. Sie sind wie die Adams–Bashforth Verfahren stark stabil und haben die Ordnung $k+1$ (Beachten Sie, dass der Grad des Interpolationspolynoms hier k ist, beim Adams–Bashforth Verfahren aber nur $k-1$). Die ersten Adams-Moulton Formeln sind:

$$\begin{aligned}
k = 0: &\quad y_{n+1} = y_n + h f_{n+1} \\
k = 1: &\quad y_{n+1} = y_n + 0.5h(f_{n+1} + f_n) \\
k = 2: &\quad y_{n+1} = y_n + h(5f_{n+1} + 8f_n - f_{n-1})/12 \\
k = 3: &\quad y_{n+1} = y_n + h(9f_{n+1} + 19f_n - 5f_{n-1} + f_{n-2})/24 \\
k = 4: &\quad y_{n+1} = y_n + h(251f_{n+1} + 646f_n - 264f_{n-1} + 106f_{n-2} - 19f_{n-3})/720.
\end{aligned}$$

und für $k = 5$

$$y_{n+1} = y_n + h(475f_{n+1} + 1427f_n - 798f_{n-1} + 482f_{n-2} - 173f_{n-3} + 27f_{n-4})/1440.$$

Das Adams–Moulton Verfahren mit $k = 1$ heißt das **implizite Euler Verfahren**, dasjenige für $k = 2$ die **Trapezregel**.

Die Adams–Moulton Verfahren haben wesentlich bessere Konvergenzeigenschaften als die Adams–Bashforth Verfahren gleicher Ordnung. Nachteilig ist, dass sie implizit sind, man also in jedem Schritt ein nichtlineares Gleichungssystem zu lösen hat.

Man kombiniert daher beide Verfahren zu einem **Prädiktor-Korrektor Verfahren** :

Sind bereits Näherungen $y_j = y(x_j)$, $j = 0, \ldots, n$, bekannt ($n \geq k$), so bestimme man mit dem Adams–Bashforth Verfahren der Ordnung k eine vorläufige Näherung

$$\tilde{y}_0 := y_n + h \sum_{j=0}^{k-1} \alpha_j f_{n-j}$$

für $y(x_{n+1})$ und verbessere diese iterativ unter Benutzung der Adams–Moulton Formel der Ordnung $k+1$:

$$\tilde{y}_{i+1} = y_n + h\Big(\beta_0 f(x_{n+1}, \tilde{y}_i) + \sum_{j=1}^{k} \beta_j f_{n+1-j}\Big),\ i = 0, 1, \ldots.$$

Erfüllt f eine Lipschitz Bedingung und ist h genügend klein gewählt, so ist diese Iteration konvergent. In der Regel genügen ein oder zwei Verbesserungsschritte (sonst ist die Schrittweite h zu groß). Das so gefundene $\tilde{y}_1$ oder $\tilde{y}_2$ wird als y_{n+1} gewählt und es wird der nächste Prädiktor-Korrektor-Schritt ausgeführt.

Eine typische Implementierung eines Prädiktor-Korrektor Verfahrens hat also die folgende Gestalt: P (Auswertung der Prädiktorformel) E (Evaluation von f) C (Auswertung der Korrektorformel) E (Evaluation von f) oder PECECE. Mit den Adams-Bashforth und Adams-Moulton Formeln für $k = 1$ (dem expliziten Euler Verfahren und der Trapezregel) erhält man also das PECECE Verfahren

$$\begin{aligned}
P \quad & y_{n+1}^{[1]} = y_n + h f_n \\
E \quad & f_{n+1}^{[1]} = f(x_{n+1}, y_{n+1}^{[1]}) \\
C \quad & y_{n+1}^{[2]} = y_n + 0.5h(f_{n+1}^{[1]} + f_n) \\
E \quad & f_{n+1}^{[2]} = f(x_{n+1}, y_{n+1}^{[2]}) \\
C \quad & y_{n+1} = y_n + 0.5h(f_{n+1}^{[2]} + f_n) \\
E \quad & f_{n+1} = f(x_{n+1}, y_{n+1}^{[2]}).
\end{aligned}$$

Vorteil der Mehrschrittverfahren ist, dass auch bei größeren Ordnungen nur in jedem Schritt eine Funktionsauswertung von f im expliziten Fall bzw. 2 oder 3 Auswertungen beim Prädiktor-Korrektor Verfahren benötigt werden, während beim Einschrittverfahren die Zahl der Funktionsauswertungen bei Steigerung der Ordnung sehr rasch wächst. Mehrschrittverfahren werden daher vor allem verwendet, wenn die Auswertung von f sehr teuer ist.

Beispiel 3.4. Wir betrachten erneut die Anfangswertaufgabe

$$y' = y^2,\ y(0.8) = 5/6,\ 0.8 \leq x \leq 1.8.$$

Mit dem klassischen Runge–Kutta Verfahren mit äquidistanter Schrittweite $h = 1/64$ ist der Fehler im Endpunkt 1.8 des Intervalls $2.55 \cdot 10^{-6}$. Hierzu muss die rechte Seite an $4 \cdot 64 = 256$ Stellen ausgewertet werden.

Mit dem PECE Verfahren mit den Adams–Bashforth und Adams–Moulton Formeln für $k = 5$ erhält man einen vergleichbaren Fehler $2.53 \cdot 10^{-6}$ mit der Schrittweite $h = 1/80$. Hierzu benötigt man zur Bestimmung des Anfangsfeldes y_1, y_2, y_3, y_4 mit dem klassischen Runge–Kutta Verfahren 16 Funktionsauswertungen und für die weiteren 76 Schritte je zwei Auswertungen, insgesamt also 168. Das Prädiktor–Korrektor Verfahren erfordert also wesentlich geringeren Aufwand. □

Nachteil der Mehrschrittverfahren ist, dass die Schrittweitensteuerung komplizierter als beim Einschrittverfahren ist. Man muss

- entweder nicht äquidistante Knoten $x_n, x_{n-1}, \ldots, x_{n-k+1}$ verwenden und kann dann die α_j bzw. β_j nicht einer Tabelle entnehmen, sondern muss sie nach jeder Veränderung der Schrittweite während der nicht äquidistanten Phase neu berechnen
- oder bei geänderter Schrittweite $\tilde{h}$ Näherung für $y(x_n - j \cdot \tilde{h})$ aus einem Interpolationspolynom berechnen.

Der zweite Zugang wird in Hairer, Nørsett, Wanner [29] zusammen mit einer kombinierten Schrittweiten- und Ordnungskontrolle diskutiert.

In MATLAB ist als Funktion ODE113 ein PECE Verfahren von Adams-Bashforth-Moulton zur Lösung von nicht-steifen Anfangswertaufgaben implementiert.

Die Adams Formeln wurden durch numerische Berechnung des Integrals in (3.8) konstruiert. Es gibt weitere Mehrschrittverfahren, die auf Integralgleichungen beruhen, die der Anfangswertaufgabe (3.1) äquivalent sind. Betrachtet man z.B. die Integralgleichung

$$y(x_{n+1}) - y(x_{n-1}) = \int_{x_{n-1}}^{x_{n+1}} f(t, y(t))\, dt, \tag{3.9}$$

und ersetzt man die unbekannte Funktion $f(t, y(t))$ durch das Interpolationspolynom zu den Daten $(x_{n-k+1}, f_{n-k+1}), \ldots, (x_{n-1}, f_{n-1}), (x_n, f_n)$, so erhält man die expliziten **Nyström Formeln**. Für $k = 1$ (und $k = 2$) ist dies die **Mittelpunktregel**

$$y_{n+1} = y_{n-1} + 2hf_n$$

und für $k = 3$

$$y_{n+1} = y_{n-1} + h(7f_n - 2f_{n-1} + f_{n-2})/3.$$

Ersetzt man in (3.9) den Integranden $f(t, y(t))$ durch das Interpolationspolynom zu den Daten $(x_{n-k+1}, f_{n-k+1}), \ldots, (x_n, f_n), (x_{n+1}, f_{n+1})$, so erhält man die impliziten **Milne–Simpson Formeln** . Für $k = 0$ ist dies das **implizite Euler Verfahren**

$$y_{n+1} = y_{n-1} + 2hf_{n+1}$$

zur Schrittweite $2h$, für $k = 1$ erhält man erneut die Mittelpunktregel, für $k = 2$ die **Milne Formel**

$$y_{n+1} = y_{n-1} + h(f_{n+1} + 4f_n + f_{n-1})/3$$

und für $k = 4$

$$y_{n+1} = y_{n-1} + h(29f_{n+1} + 124f_n + 24f_{n-1} + 4f_{n-2} - f_{n-3})/90.$$

Es ist klar, dass die Nyström Formeln und die Milne–Simpson Regeln stabil, aber nicht stark stabil sind, denn das zugehörige Polynom $\rho(\lambda) = (\lambda^2 - 1)\lambda^{k-2}$ besitzt die einfachen Nullstellen 1 und -1 und die $(k-2)$-fache Nullstelle 0.

Die bisherigen Mehrschrittverfahren basierten auf der numerischen Lösung der Integralgleichung (3.8) bzw. (3.9). Die folgende Klasse von Verfahren wird mit Hilfe der numerischen Differentiation konstruiert.

Es seien bereits Näherungen $y_{n-k+1}, \ldots, y_n$ der Lösung der Anfangswertaufgabe (3.1) an den Knoten $x_{n-k+1}, \ldots, x_n$ bekannt. Um $y_{n+1} \approx y(x_{n+1})$ zu bestimmen betrachten wir das Interpolationspolynom q zu den Daten (x_{n-k+1}, y_{n-k+1}), $\ldots$, (x_n, y_n), (x_{n+1}, y_{n+1}). Dann kann man q nach der Newtonschen Darstellung des Interpolationspolynoms mit den **rückwärtsgenommenen Differenzen**

$$\nabla^0 y_n := y_n, \ \nabla^{j+1} y_n := \nabla^j y_n - \nabla^j y_{n-1}$$

schreiben als

$$q(s) = y(x_n + sh) = \sum_{j=0}^{k} (-1)^j \binom{-s+1}{j} \nabla^j y_{n+1}. \tag{3.10}$$

Der Unbekannte Wert y_{n+1} wird nun so bestimmt, dass das Polynom q die Differentialgleichung an einem Gitterpunkt erfüllt:

$$q'(x_{n+\ell}) = f(x_{n+\ell}, y_{n+\ell}), \ \ell \in \{0, 1, \ldots\}. \tag{3.11}$$

Für $\ell = 0$ erhält man explizite Formeln, und zwar für $k = 1$ das explizite Euler Verfahren und für $k = 2$ die Mittelpunktregel. Die Formeln für $k \geq 3$ sind instabil und daher wertlos.

Für $\ell = 1$ erhält man implizite Formeln, die **BDF Methoden** (backward differentiation formulas)

$$\sum_{j=0}^{k} \alpha_j \nabla^j y_{n+1} = h f_{n+1}$$

mit den Koeffizienten

$$\alpha_j = (-1)^j \frac{d}{ds} \binom{-s+1}{j} \Big|_{s=1}.$$

Mit

$$(-1)^j \binom{-s+1}{j} = \frac{1}{j!}(s-1)s(s+1)\dots(s+j-2)$$

folgt

$$\alpha_0 = 0,\ \alpha_j = \frac{1}{j} \text{ für } j \geq 1,$$

und daher

$$\sum_{j=1}^{k} \frac{1}{j} \nabla^j y_{n+1} = h f_{n+1}. \tag{3.12}$$

Für $k \leq 6$ rechnet man leicht aus, dass gilt

$$\begin{aligned}
k = 1 &: \ y_{n+1} - y_n = h f_{n+1} \\
k = 2 &: \ 3y_{n+1} - 4y_n + y_{n-1} = 2h f_{n+1} \\
k = 3 &: \ 11y_{n+1} - 18y_n + 9y_{n-1} - 2y_{n-2} = 6h f_{n+1} \\
k = 4 &: \ 25y_{n+1} - 48y_n + 36y_{n-1} - 16y_{n-2} + 3y_{n-3} = 12h f_{n+1} \\
k = 5 &: \ 137y_{n+1} - 300y_n + 300y_{n-1} - 200y_{n-2} + 75y_{n-3} - 12y_{n-4} = 60h f_{n+1} \\
k = 6 &: \ 147y_{n+1} - 360y_n + 450y_{n-1} - 400y_{n-2} + 225y_{n-3} - 72y_{n-4} + 10y_{n-5} = 60h f_{n+1}.
\end{aligned}$$

Durch Diskussion des Polynoms

$$\rho(\lambda) = \sum_{j=1}^{k} \frac{1}{j} \lambda^{k-j} (\lambda - 1)^j$$

sieht man für $k \leq 6$ leicht, dass diese BDF–Formeln stabil sind. Für $k \geq 7$ sind sie instabil (vgl. *Hairer, Nørsett, Wanner* [29], p. 380).

Kapitel 4

Steife Probleme

4.1 Motivation

Es gibt Differentialgleichungen mit Lösungen, zu deren Approximation bei Anwendung expliziter Verfahren viel kleinere Schrittweiten benötigt werden, als man erwartet. Diese Probleme werden **steif** genannt.

Beispiel 4.1. Die Anfangswertaufgabe

$$y' = -\lambda(y - e^{-x}) - e^{-x}, \quad y(0) = 1 \tag{4.1}$$

besitzt für alle $\lambda \in \mathbb{R}$ die eindeutige Lösung $y(x) = e^{-x}$.

Tabelle 4.1 und Tabelle 4.2 enthalten die Fehler der Näherungslösungen bei konstanter Schrittweite $h = 0.01$ für das Polygonzugverfahren, das verbesserte Polygonzugverfahren und das klassische Runge–Kutta Verfahren für die Parameter $\lambda = 1$ und $\lambda = 1000$.

Tabelle 4.1: Fehler für $\lambda = 1$

x	Polygonzug	verb. Polygonzug	Runge–Kutta
0.00	$0.00E+0$	$0.00E+0$	$0.00E+0$
0.10	$-4.55E-4$	$1.52E-6$	$7.60E-12$
0.20	$-8.24E-4$	$2.75E-6$	$1.38E-11$
0.30	$-1.12E-3$	$3.73E-6$	$1.87E-11$
0.40	$-1.35E-3$	$4.50E-6$	$2.25E-11$
0.50	$-1.52E-3$	$5.09E-6$	$2.55E-11$
0.60	$-1.66E-3$	$5.53E-6$	$2.77E-11$
0.70	$-1.75E-3$	$5.83E-6$	$2.92E-11$
0.80	$-1.81E-3$	$6.04E-6$	$3.02E-11$
0.90	$-1.84E-3$	$6.14E-6$	$3.07E-11$
1.00	$-1.85E-3$	$6.18E-6$	$3.09E-11$

Tabelle 4.2: Näherungen für $\lambda = 1000$

x	Polygonzug	verb. Polygonzug	Runge–Kutta
0.00	$0.00E+0$	$0.00E+0$	$0.00E+0$
0.01	$-4.98E-5$	$1.25E-4$	$1.04E-3$
0.02	$3.99E-4$	$5.24E-3$	$3.04E-1$
0.03	$-3.64E-3$	$2.15E-1$	$8.83E+1$
0.04	$3.27E-2$	$8.82E+0$	$2.57E+4$
0.05	$-2.95E-1$	$3.61E+2$	$7.48E+6$
0.06	$2.65E+0$	$1.48E+4$	$2.18E+9$
0.07	$-2.39E+1$	$6.08E+5$	$6.33E+11$
0.08	$2.15E+2$	$2.49E+7$	$1.84E+14$
0.09	$-1.93E+3$	$1.02E+9$	$5.36E+16$
0.10	$1.74E+4$	$4.19E+10$	$1.56E+19$

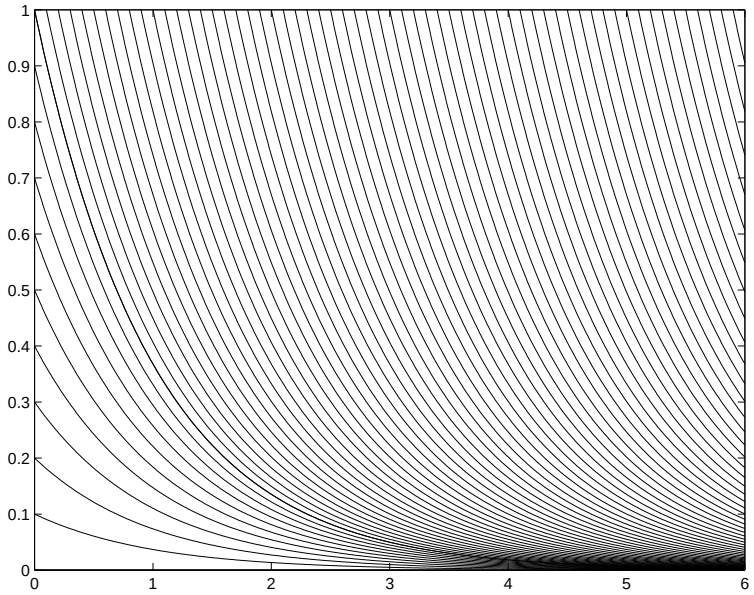

Abbildung 4.1: Lösungen für $\lambda = 1$

Abbildung 4.1 und Abbildung 4.2 zeigen die Lösungen der Anfangswertaufgaben

$$y' = -\lambda(y - e^{-x}) - e^{-x},\ y(x_0) = y_0$$

für verschiedene Werte von x_0 und y_0 für $\lambda = 1$ und $\lambda = 20$.

Wir möchten die Lösung $y(x) = e^{-x}$ von (4.1) für großes λ mit einem numerischen Verfahren verfolgen. Durch (Rundungs- oder Verfahrens-) Fehler werden wir ein wenig von der Lösung weggeführt. Ist $y_n \neq e^{-x_n}$, so konvergiert die Lösung

$$y(x; x_n, y_n) = (y_n - e^{-x_n})e^{-\lambda(x-x_n)} + e^{-x}$$

der Anfangswertaufgabe für großes λ sehr rasch gegen die quasi stationäre Lösung $\tilde{y}(x) = e^{-x}$. Die dem Betrage nach sehr große Steigung

$$y'(x_n; x_n, y_n) = -\lambda(y_n - e^{-x_n}) - e^{-x_n}$$

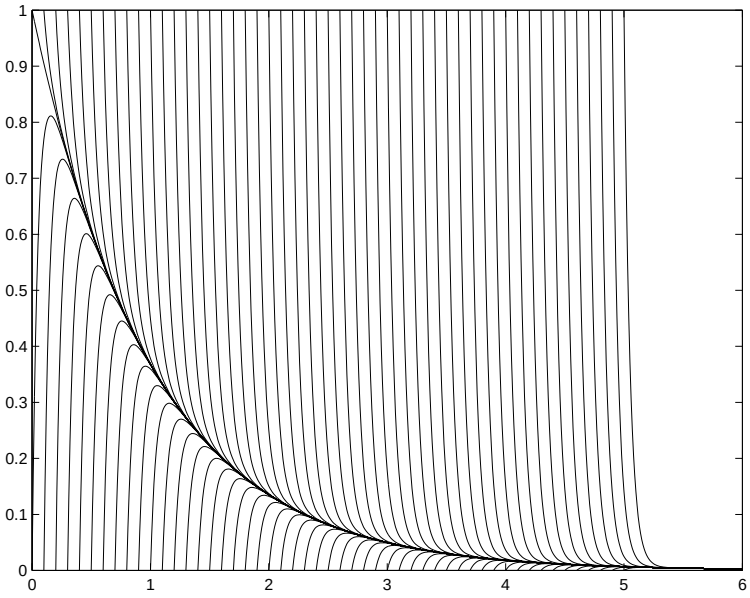

Abbildung 4.2: Lösungen für $\lambda = 20$

führt dazu, dass für das Euler Verfahren (und für die anderen Einschrittverfahren genauso) über das Ziel hinausgeschossen wird und die Fehler sich aufschaukeln. □

4.2 Stabilitätsgebiete

Wendet man das Eulersche Polygonzugverfahren auf die Testgleichung

$$y' = \lambda y, \ \mathrm{Re}(\lambda) < 0, \tag{4.2}$$

an, so erhält man bei konstanter Schrittweite $h > 0$

$$y_{n+1} = y_n + h\lambda y_n,$$

und daher

$$y_n = (1 + h\lambda)^n y_0.$$

Für

$$|1 + h\lambda| > 1$$

explodiert die numerische Lösung y_n, und zwar ist dieses Aufschaukeln um so rascher, je kleiner $\mathrm{Re}(\lambda)$ ist, je schneller die Lösung der Anfangswertaufgabe also abklingt. Das Mindeste, was man von einem Verfahren erwarten muss, ist aber, dass die numerische Lösung bei nicht zu kleinen Schrittweiten ebenfalls abklingt.

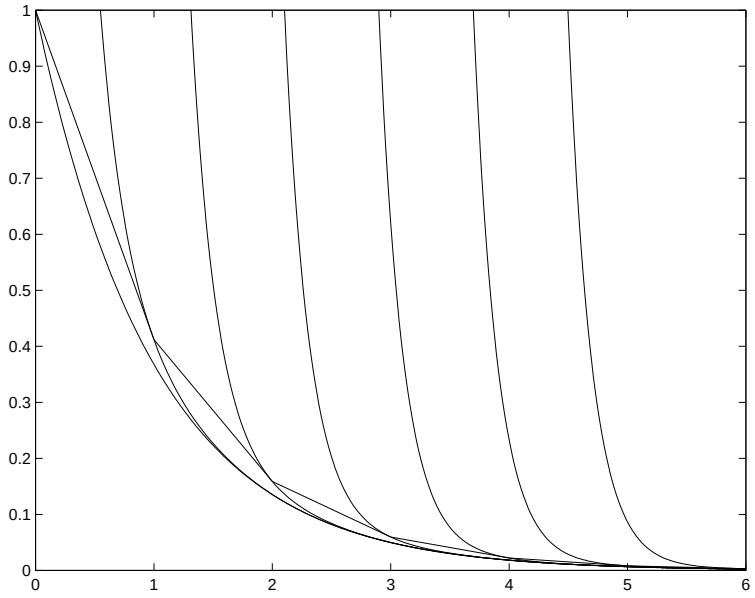

Abbildung 4.3: Implizites Euler Verfahren

Das einfachste Verfahren, dessen numerische Lösung der Testgleichung (4.2) bei annehmbaren Schrittweiten das Abklingverhalten der Lösung der Anfangswertaufgabe reproduziert, ist das **implizite Euler Verfahren**

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}^n + h\boldsymbol{f}(x_{n+1}, \boldsymbol{y}^{n+1}).$$

Mit ihm erhält man für (4.2)

$$y_{n+1} = y_n + h\lambda y_{n+1},$$

d.h.

$$y_n = \Big(\frac{1}{1-h\lambda}\Big)^n y_0 \to 0 \quad \text{für } n \to \infty$$

für jede Schrittweite $h > 0$. Man geht also einen linearen Schritt mit der Steigung weiter, die in dem Richtungsfeld der Differentialgleichung dort herrscht, wo man hinkommt (Abbildung 4.3 zeigt 6 Schritte des impliziten Euler Verfahrens für Beispiel 4.1. mit $\lambda = -5$ und der Schrittweite $h = 1$).

Der Preis, den man für dieses verbesserte Stabilitätsverhalten zu zahlen hat, ist, dass man im allgemeinen Fall in jedem Schritt ein nichtlineares Gleichungssystem

$$\boldsymbol{F}(\boldsymbol{y}^{n+1}) = \boldsymbol{y}^{n+1} - \boldsymbol{y}^n - h\boldsymbol{f}(x_{n+1}, \boldsymbol{y}^{n+1}) = \boldsymbol{0}$$

zu lösen hat. Dies kann man z.B. mit dem Newton Verfahren mit dem Startwert $\boldsymbol{y}^n$ tun.

Bemerkung 4.2. Die Testgleichung (4.2) ist aussagekräftig für allgemeinere Systeme, denn ist die Matrix $\boldsymbol{A} \in \mathbb{R}^{(n,n)}$ in dem linearen System

$$\boldsymbol{y}' = \boldsymbol{A}\boldsymbol{y} + \boldsymbol{g} \tag{4.3}$$

diagonalisierbar und gilt

$$\boldsymbol{X}^{-1}\boldsymbol{A}\boldsymbol{X} = \boldsymbol{\Lambda} = \operatorname{diag}\{\lambda_1, \ldots, \lambda_n\}$$

mit einer regulären Matrix $\boldsymbol{X}$, so erhält man mit der Variablentransformation $\boldsymbol{z} := \boldsymbol{X}^{-1}\boldsymbol{y}$

$$\boldsymbol{z}' = \boldsymbol{X}^{-1}\boldsymbol{y}' = \boldsymbol{X}^{-1}\boldsymbol{A}\boldsymbol{X}\boldsymbol{X}^{-1}\boldsymbol{y} + \boldsymbol{X}^{-1}\boldsymbol{g} = \boldsymbol{\Lambda}\boldsymbol{z} + \boldsymbol{X}^{-1}\boldsymbol{g} =: \boldsymbol{\Lambda}\boldsymbol{z} + \tilde{\boldsymbol{g}},$$

d.h. das entkoppelte System

$$z_j' = \lambda_j z_j + \tilde{g}_j, \quad j = 1, \ldots, n. \tag{4.4}$$

Wendet man auf das System (4.3) ein Mehrschrittverfahren

$$\sum_{j=0}^{m} \alpha_j \boldsymbol{y}^{k-j} - h \sum_{j=0}^{m} \beta_j (\boldsymbol{A}\boldsymbol{y}^{k-j} + \boldsymbol{g}^{k-j}) = \boldsymbol{0}$$

an, so ist dieses mit $\boldsymbol{z}^i := \boldsymbol{X}^{-1}\boldsymbol{y}^i$ äquivalent zu

$$\sum_{j=0}^{m} \alpha_j \boldsymbol{z}^{k-j} - h \sum_{j=0}^{m} \beta_j (\boldsymbol{\Lambda}\boldsymbol{z}^{k-j} + \tilde{\boldsymbol{g}}^{k-j}) = \boldsymbol{0},$$

d.h. zu dem Mehrschrittverfahren

$$\sum_{j=0}^{m} \alpha_j z_{i,k-j} - h\lambda_i \sum_{j=0}^{m} \beta_j (z_{i,k-j} + \tilde{g}_{i,k-j}) = 0, \quad i = 1, \ldots, n,$$

für die skalaren Gleichungen (4.4). □

Das Abklingverhalten des Mehrschrittverfahrens für das System (4.3) wird also bestimmt durch das Abklingverhalten für die skalaren Gleichungen (4.2) für die Eigenwerte $\lambda = \lambda_j$ der Matrix $\boldsymbol{A}$. Da diese nicht notwendig reell sind, auch wenn $\boldsymbol{A}$ nur reelle Einträge besitzt, müssen wir die Testgleichung (4.2) für $\lambda \in \mathbb{C}$ mit negativem Realteil untersuchen.

Wendet man das explizite Runge–Kutta Verfahren (2.17) auf das System (4.3) an, so ist dieses wieder äquivalent der Anwendung auf die skalaren Gleichungen. Das

Abklingverhalten der numerischen Lösung wird also wieder bestimmt durch das Abklingverhalten für die Testgleichung (4.2). Man erhält

$$\begin{aligned} k_1 &:= \lambda y_n \\ k_j &:= \lambda\Big(y_n + h\sum_{\ell=1}^{j-1}\beta_{j\ell}\,k_\ell\Big), \quad j = 2,\ldots,s, \\ y_{n+1} &:= y_n + h\sum_{j=1}^{s}\gamma_j\,k_j. \end{aligned} \tag{4.5}$$

Setzt man die k_j nacheinander ein, so folgt

$$y_{n+1} = R(h\lambda)y_n$$

mit

$$R(z) = 1 + z\sum_{j=1}^{s}\gamma_j + z^2\sum_{j=1}^{s}\sum_{k=1}^{j-1}\gamma_j\beta_{jk} + z^3\sum_{j=1}^{s}\sum_{k,\ell=1}^{j-1}\gamma_j\beta_{jk}\beta_{k\ell} + \ldots \in \Pi_s,$$

und die Folge $\{y_n\}$ ist offenbar beschränkt, wenn $|R(z)| \le 1$ gilt. Beachten Sie, dass diese Bedingung nur von $z := h\lambda$ abhängt.

Wir definieren

Definition 4.3. *Das* **Stabilitätsgebiet** $S \subset \mathbb{C}$ *eines Verfahrens ist die Menge aller* $z := h\lambda \in \mathbb{C}$*, so dass für alle Startwerte die erzeugte Folge* $\{y_n\}$ *mit der Schrittweite* h *für die Testgleichung (4.2) beschränkt ist.*

Das Stabilitätsgebiet eines Verfahrens hat die folgende Bedeutung. Will man ein lineares Differentialgleichungssystem $\boldsymbol{y'} = \boldsymbol{Ay}$ lösen und besitzt die Matrix $\boldsymbol{A}$ die Eigenwerte λ_j, $j = 1,\ldots,n$, mit $\operatorname{Re}\lambda_j < 0$, so kann man das System nur dann stabil mit diesem Verfahren lösen, wenn die Schrittweite h so klein gewählt ist, dass die Zahlen $h\lambda_j$ für alle $j = 1,\ldots,n$ in dem Stabilitätsgebiet S des Verfahrens liegen.

Wünschenswert ist für ein Verfahren, dass sein Stabilitätsgebiet die linke Halbebene umfasst.

Definition 4.4. *Ein Verfahren zur Lösung von Anfangswertaufgaben heißt* **A–stabil**, *wenn gilt*

$$S \supset \mathbb{C}_- := \{z \in \mathbb{C} : \operatorname{Re} z \le 0\}.$$

Für das explizite Runge–Kutta Verfahren (2.17) der Stufe s ist die Funktion R ein Polynom vom Höchstgrad s. Das Stabilitätsgebiet

$$S = \{z \in \mathbb{C} : |R(z)| \leq 1\}$$

ist daher mit Sicherheit eine beschränkte Menge in $\mathbb{C}$. Explizite Runge–Kutta Verfahren sind also niemals A–stabil.

Beispiel 4.5. Für das (explizite) Polygonzugverfahren ist das Stabilitätsgebiet

$$S_{\text{expliziter Euler}} = \{z \in \mathbb{C} : |1 + z| \leq 1\},$$

für das implizite Euler Verfahren

$$S_{\text{impliziter Euler}} = \{z \in \mathbb{C} : |1 - z| \geq 1\} \supset \mathbb{C}_-.$$

Das implizite Euler Verfahren ist also A–stabil. □

Beispiel 4.6. Wendet man das verbesserte Polygonzugverfahren auf die Testgleichung an, so erhält man

$$y_{n+1} = y_n + h\lambda\Big(y_n + \frac{1}{2}h\lambda y_n\Big) = \Big(1 + (h\lambda) + \frac{1}{2}(h\lambda)^2\Big)y_n.$$

Es ist also

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

Für das Verfahren von Heun erhält

$$y_{n+1} = y_n + \frac{h}{2}\Big(\lambda y_n + \lambda(y_n + h\lambda y_n)\Big) = \Big(1 + h\lambda + \frac{1}{2}(h\lambda)^2\Big)y_n,$$

d.h. dieselbe Funktion R wie für das verbesserte Polygonzugverfahren. Das Stabilitätsgebiet $S = \{z \in \mathbb{C} : |R(z)| \leq 1\}$ ist in Abbildung 4.4 dargestellt.

Dass man für das verbesserte Polygonzugverfahren und das Verfahren von Heun dieselbe Funktion R erhalten hat, ist kein Zufall. Besitzt das Runge–Kutta Verfahren die Ordnung p, so gilt

$$R(z) = \sum_{j=0}^{p} \frac{z^j}{j!} + O(z^{p+1}) \quad \text{für } z \to 0, \tag{4.6}$$

denn die Lösung der Anfangswertaufgabe $y' = y$, $y(0) = 1$, ist $y(x) = e^x$, und daher muss für die numerische Lösung $y_1 = R(h)y_0 = R(h)$ im ersten Schritt gelten

$$e^z - R(z) = e^h - R(h) = O(h^{p+1}) = O(z^{p+1}).$$

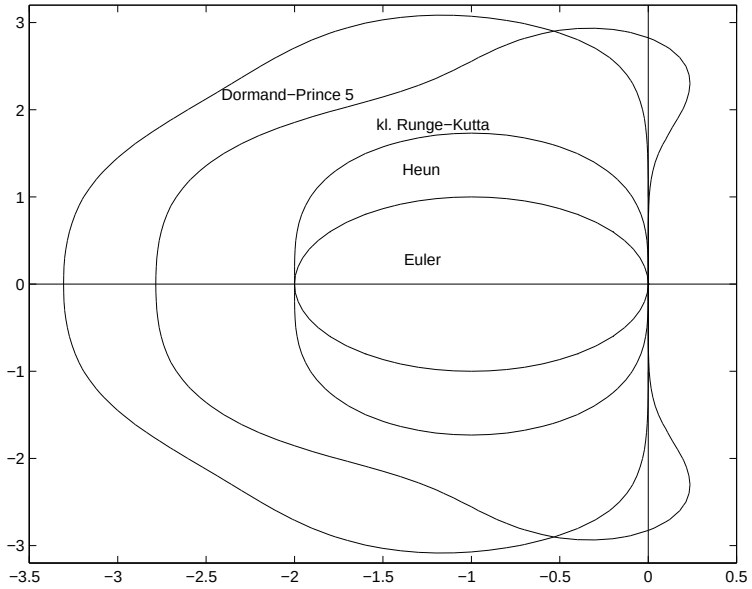

Abbildung 4.4: Stabilitätsgebiete von Runge–Kutta Verfahren

Da R ein Polynom ist, folgt hieraus (4.6)

Gilt zusätzlich $s = p$, so ist R ein Polynom vom Grade p, und es folgt

$$R(z) = \sum_{j=0}^{p} \frac{z^j}{j!}.$$

Für das klassische Runge–Kutta Verfahren, die 3/8–Regel und das Verfahren von Kuntzmann erhält man also

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

Auch hierfür findet man das Stabilitätsgebiet in Abbildung 4.4.

Das Verfahren von Dormand und Prince ist von der Ordnung $p = 5$ und benötigt $s = 6$ Stufen (die siebte Stufe wird nur für die Fehlerschätzung verwendet). Die Funktion R hat also die Gestalt

$$R(z) = 1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \frac{z^4}{24} + \frac{z^5}{120} + \alpha z^6,$$

und durch direkte Rechnung erhält man $\alpha = \frac{1}{600}$. Auch hierfür findet man das Stabilitätsgebiet in Abbildung 4.4 □

Beispiel 4.7. Wendet man das k-Schrittverfahren

$$\sum_{\nu=0}^{k} a_\nu y_{n+\nu} = h \sum_{\nu=0}^{k} b_\nu f_{n+\nu}$$

auf die Testgleichung $y' = \lambda y$ an, so erhält man

$$\sum_{\nu=0}^{k}(a_\nu - zb_\nu)y_{n+\nu} = 0,\ z = h\lambda. \tag{4.7}$$

Bezeichnet $\zeta_\mu(z)$, $\mu = 1, \ldots, \ell$, für $z \in \mathbb{C}$ die Nullstellen des charakteristischen Polynoms

$$\chi_z(\zeta) := \sum_{\nu=0}^{k}(a_\nu - zb_\nu)\zeta^\nu \tag{4.8}$$

und ist r_μ die Vielfachheit von ζ_μ, so sind die Lösung von (4.7) Linearkombinationen von

$$\zeta_\mu(z)^n, n\zeta_\mu(z)^n, \ldots, n^{r_\mu-1}\zeta_\mu(z)^n,\ \mu = 1, \ldots, \ell.$$

Damit besteht das Stabilitätsgebiet S genau aus den Punkten $z \in \mathbb{C}$, für die $|\zeta_\mu(z)| \leq 1$ für alle Nullstellen $\zeta_\mu(z)$ gilt und $|\zeta_\mu(z)| < 1$ für alle mehrfachen Nullstellen.

Ist $\zeta(z)$ für ein $z \in \mathbb{C}$ eine einfache Nullstelle von (4.8), so existiert eine lokale Umkehrabbildung $\zeta \mapsto z$, und diese erhält man durch Auflösung der Gleichung

$$\chi_z(\zeta) = \sum_{\nu=0}^{k}(a_\nu - zb_\nu)\zeta^\nu = 0$$

nach z:

$$z(\zeta) = \sum_{\nu=0}^{k} a_\nu\zeta^\nu \Big/ \sum_{\nu=0}^{k} b_\nu\zeta^\nu. \tag{4.9}$$

Das Bild des Äußeren des Einheitskreises in der ζ-Ebene unter dieser Abbildung beschreibt die Punkte, die nicht im Stabilitätsgebiet S liegen. Damit begrenzt die Bildkurve des Einheitskreises $\zeta = e^{i\theta}$, $0 \leq \theta \leq 2\pi$, das Stabilitätsgebiet. Genauer liegen diejenigen Punkte der z-Ebene in S, die links von der orientierten Kurve $\theta \mapsto z(e^{i\theta})$ liegen.

Für die Adams–Bashforth Formeln gilt

$$y_{n+1} - y_n = h\sum_{\nu=0}^{k-1}\alpha_\nu f_{n-\nu},$$

und daher ist das Stabilitätsgebiet bestimmt durch die Kurve

$$z = (\zeta^k - \zeta^{k-1}) \Big/ \sum_{\mu=0}^{k-1}\alpha_{k-1-\mu}\zeta^\mu = (\zeta - 1) \Big/ \sum_{\mu=0}^{k-1}\alpha_{k-1-\mu}\zeta^{\mu+1-k}, \quad \zeta = e^{i\theta}.$$

Für $k = 1$ gilt $z = \zeta - 1$. Die das Stabilitätsgebiet begrenzende Kurve ist also $\theta \mapsto e^{i\theta} - 1$, $0 \leq \theta \leq 2\pi$, und dies ist der im mathematisch positiven Sinn durchlaufene Kreis mit dem Mittelpunkt -1 und dem Radius 1. Damit ist das Stabilitätsgebiet

$$S = \{z \in \mathbb{C} : |z + 1| \leq 1\}.$$

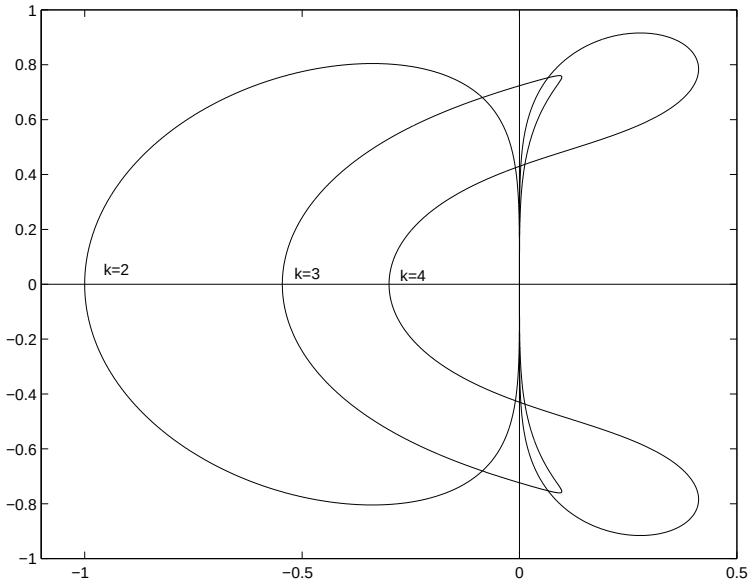

Abbildung 4.5: Stabilitätsgebiete von Adams–Bashforth Verfahren

Dies ist auch kein Wunder, denn das Adams–Bashforth Verfahren mit $k = 1$ ist das Polygonzugverfahren.

Für $k = 2, 3, 4$ sind diese Kurven in Abbildung 4.5 dargestellt. Die Stabilitätsgebiete werden also sehr rasch kleiner. Beachten Sie, dass für $k = 4$ nur das von der Kurve umschlossene beschränkte Gebiet der linken Halbebene links von der Kurve liegt. Nur dies ist also das Stabilitätsgebiet für $k = 4$.

Für die Adams–Moulton Verfahren

$$y_{n+1} - y_n = h \sum_{\nu=0}^{k} \beta_\nu f_{n+1-\nu}$$

erhält man genauso die das Stabilitätsgebiet umschließende Kurve

$$z = (\zeta^k - \zeta^{k-1}) \Big/ \sum_{\mu=0}^{k} \beta_{k-\mu} \zeta^\mu = (1 - \frac{1}{\zeta}) \Big/ \sum_{\mu=0}^{k} \beta_{k-\mu} \zeta^{\mu-k}.$$

Für $k = 0$ ist dies $\theta \mapsto 1 - e^{-i\theta}$, d.h. der im negativen Drehsinn durchlaufene Kreis mit dem Mittelpunkt 1 und dem Radius 1. Damit ist das Stabilitätsgebiet wie in Beispiel 4.5. (man beachte, dass das Adams–Moulton Verfahren mit $k = 0$ das implizite Euler Verfahren ist)

$$S = \{z \in \mathbb{C} : |z - 1| \geq 1\}.$$

Für $k = 1$, die **Trapezregel**,

$$y_{n+1} = y_n + 0.5h(f_{n+1} + f_n) \tag{4.10}$$

ist die das Stabilitätsgebiet begrenzende Kurve

$$\theta \mapsto (1 - e^{-i\theta}) \Big/ (0.5 + 0.5e^{-i\theta}), \quad 0 \le \theta \le 2\pi.$$

unter Benutzung von $e^{i\theta} = \cos\theta + i\sin\theta$ erhält man die Darstellung

$$\theta \mapsto \frac{2\sin\theta}{1+\cos\theta} i, \quad 0 \le \theta \le 2\pi,$$

und dies ist die (zweimal) von unten nach oden durchlaufene imaginäre Achse. Das Stabilitätsgebiet ist damit die Halbebene

$$S_{\text{Trapezregel}} = \{z \in \mathbb{C} \,:\, \operatorname{Re} z \le 0\} = \mathbb{C}_-.$$

Die Trapezregel ist also ebenfalls A–stabil.

Dieses Ergebnis kann man schneller so erhalten: Wendet man die Trapezregel auf die Testgleichung an, so erhält man

$$y_{n+1} = y_n + 0.5h(\lambda y_{n+1} + \lambda y_n),$$

d.h.

$$y_{n+1} = \frac{h\lambda - (-2)}{h\lambda - 2} y_n,$$

und damit ist

$$S_{\text{Trapezregel}} = \{z \in \mathbb{C} \,:\, \Big|\frac{z+2}{z-2}\Big| \le 1\} = C_-.$$

Für $k = 2, 3, 4$ enthält Abbildung 4.6 die Stabilitätsgebiete der Adams–Moulton Verfahren. Diese sind wie die der Adams–Bashforth Verfahren beschränkt, ihre Größe nimmt aber wesentlich langsamer ab. □

Für die BDF–Verfahren

$$\sum_{\nu=1}^{k} \frac{1}{\nu} \nabla^\nu y_{n+1} = h f_{n+1}$$

erhält man

$$z = \sum_{\mu=1}^{k} \frac{1}{\mu} \Big(1 - \frac{1}{\zeta}\Big)^\mu,$$

und hiermit die Stabilitätsgebiete in Abbildung 4.7.

Für $k \ge 3$ sind die BDF–Verfahren wieder nicht A–stabil. Sie haben aber die Eigenschaft, dass ein Sektor

$$S_\alpha := \{z \in \mathbb{C} \,:\, |\arg(-z)| \le \alpha\} \tag{4.11}$$

in der komplexen Ebene im Stabilitätsgebiet enthalten ist.

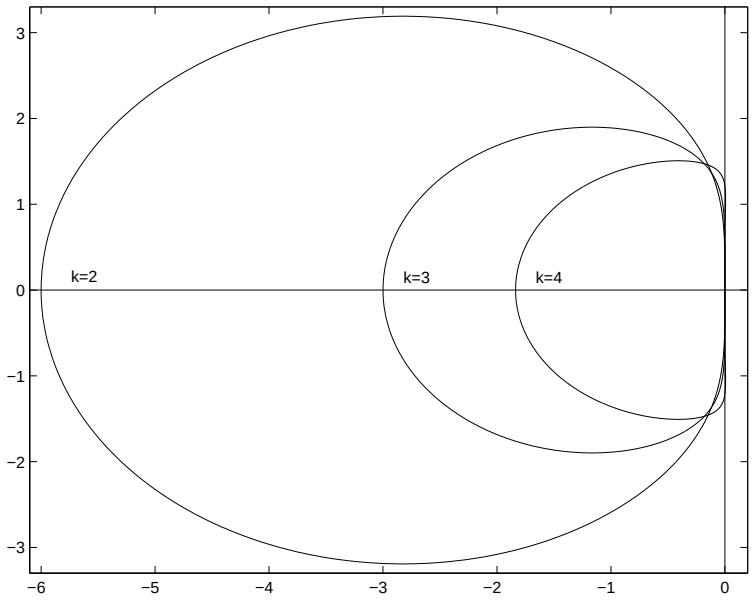

Abbildung 4.6: Stabilitätsgebiete von Adams–Moulton Verfahren

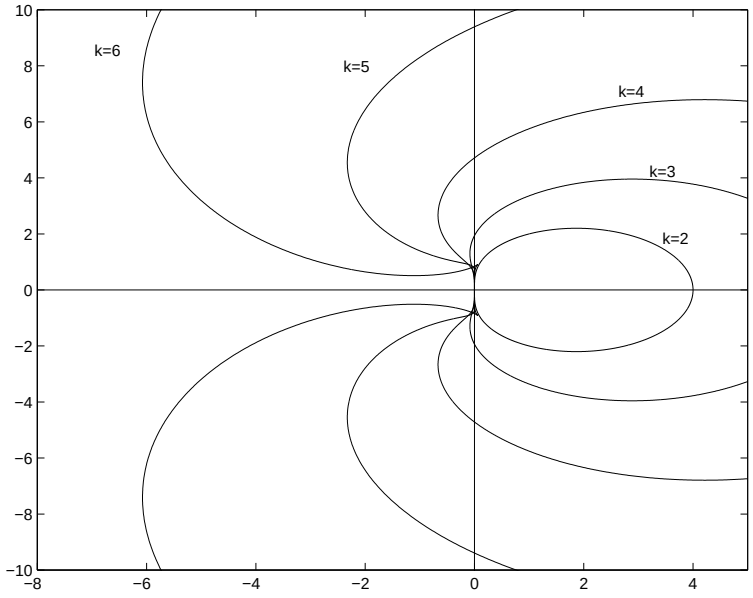

Abbildung 4.7: Stabilitätsgebiete von BDF-Verfahren

Bei vielen Anwendungen weiß man, dass die Eigenwerte der Linearisierung in einem festen Sektor der Gestalt (4.11) liegen, so dass man stabiles Verhalten der numerischen Lösung für alle Schrittweiten erwarten kann, wenn das Stabilitätsgebiet des benutzten Verfahrens diesen Sektor umfasst. Man betrachtet daher als Abschwächung der A–Stabilität

Definition 4.8. *Ein Verfahren heißt* **A(α)–stabil**, *wenn sein Stabilitätsgebiet den Sektor der Gestalt (4.11) enthält. Es heißt* **A(0)–stabil**, *wenn es A(α)–stabil für ein $\alpha > 0$ ist.*

Die BDF–Verfahren sind für $3 \leq k \leq 6$ A(α)–stabil mit den Öffnungswinkeln

k	1	2	3	4	5	6
α	90°	90°	86.0°	73.3°	51.8°	17.8°

Wegen dieser (gegenüber den Einschrittverfahren und den Adams–Bashforth und Adams–Moulton Verfahren) wesentlich besseren Stabilitätseigenschaften wurde die BDF-Verfahren schon früh für steife Probleme vorgeschlagen (vgl. *Curtiss, Hirschfelder* [12]). Eine Reihe von Codes basiert auf diesen Methoden.

Wir haben bisher als A–stabile Verfahren nur das implizite Euler Verfahren (der Ordnung 1) und die Trapezregel (der Ordnung 2) gefunden. Der nächste Satz von *Dahlquist* [14] zeigt, dass man keine besseren A–stabilen Mehrschrittverfahren finden kann. Einen Beweis findet man in *Hairer, Wanner* [30], p. 247 ff.

Satz 4.9. (Dahlquist)

(i) Explizite Mehrschrittverfahren sind niemals A–stabil.

(ii) Die Ordnung eines A–stabilen impliziten Mehrschrittverfahrens ist höchstens 2.

(iii) Die Trapezregel (4.10) ist das A–stabile Verfahren der Ordnung 2 *mit der kleinsten Fehlerkonstanten.*

Die Forderung der A–Stabilität ist zwar einerseits zu stark (und muss daher zur A(α)–Stabilität abgeschwächt werden), andererseits ist sie zu schwach und muss verschärft werden. Es gibt A–stabile Einschrittverfahren wie die Trapezregel, deren

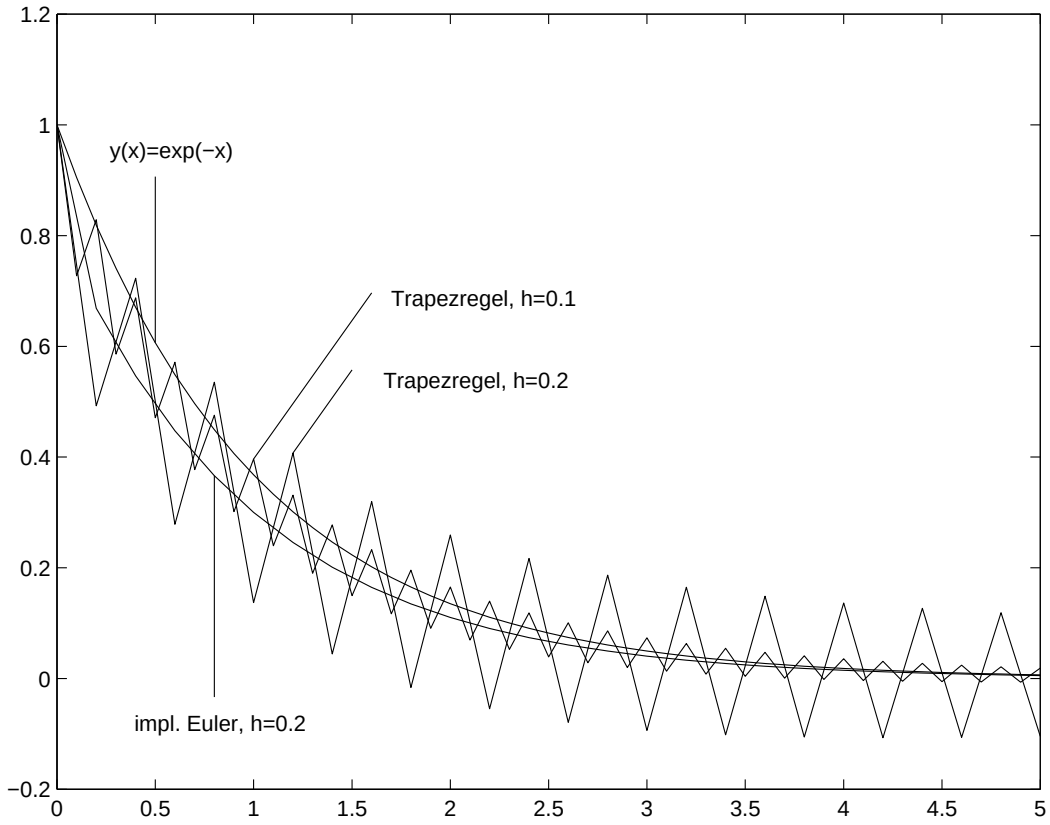

Abbildung 4.8: L–Stabilität

Stabilitätsbereich genau die linke Halbebene ist. Besitzt dieses für die Testgleichung die Gestalt $y_{n+1} = R(z)y_n$ mit einer rationalen Funktion R, so gilt $|R(iy)| = 1$ für alle $y \in \mathbb{R}$, und daher

$$\lim_{z\to\infty} |R(z)| = \lim_{y\to\infty,\ y\in\mathbb{R}} |R(iy)| = 1.$$

Dies bedeutet, dass zwar $|R(z)| < 1$ für alle $z \in \mathbb{C}_-$ gilt, aber für z, die großen negativen Realteil haben, $|R(z)|$ sehr nahe bei 1 liegt. Die Konsequenz ist, dass steife Komponenten nur sehr langsam ausgedämpft werden.

Beispiel 4.10. Wir betrachten wieder (vgl. Beispiel 4.1.) die Anfangswertaufgabe

$$y' = -\lambda(y - e^{-x}) - e^{-x}, \quad y(0) = 1,$$

für $\lambda = 1000$. Abbildung 4.8 enthält die Lösung $y(x) = e^{-x}$ und die numerischen Lösungen für das implizite Euler Verfahren mit der Schrittweite $h = 0.2$ und die Trapezregel für die Schrittweiten $h = 0.1$ und $h = 0.2$.

Dieses Verhalten gibt Anlass zu der folgenden

Definition 4.11. *Ein Einschrittverfahren heißt* **L–stabil**, *wenn es A–stabil ist und zusätzlich gilt*

$$\lim_{z\to\infty} R(z) = 0.$$

Offensichtlich ist das implizite Euler Verfahren mit der Funktion $R(z) = 1/(1+z)$ L–stabil, nicht aber die Trapezregel mit der Funktion $R(z) = (1-0.5z)/(1+0.5z)$.

4.3 Implizite Runge–Kutta Verfahren

A–stabile Verfahren (oder wenigstens Verfahren mit einem größeren Stabilitätsgebiet) kann man als Verallgemeinerung der expliziten Runge–Kutta Verfahren erhalten. Bei diesen verwenden wir zur Berechnung von k_j nur die Werte von y_n und k_i, $i = 1, \ldots, j-1$.

Definition 4.12. *Es seien $\beta_{ij} \in \mathbb{R}, \gamma_i \in \mathbb{R}$, $i, j = 1, \ldots, s$, gegeben und α_i mit*

$$\alpha_i = \sum_{j=1}^{s} \beta_{ij}, \ i = 1, \ldots, s. \tag{4.12}$$

Dann heißt das Verfahren

$$\begin{aligned} k_i &= f\Big(x_n + \alpha_i h, y_n + h \sum_{j=1}^{s} \beta_{ij} k_j\Big), \ i = 1, \ldots, s, \\ y_{n+1} &= y_n + h \sum_{i=1}^{s} \gamma_i k_i \end{aligned} \tag{4.13}$$

Runge–Kutta Verfahren *mit s Stufen.*

Bemerkung 4.13. Gilt $\beta_{ij} = 0$ für $i \le j$, so handelt es sich um ein explizites Runge–Kutta Verfahren. Gilt $\beta_{ij} \neq 0$ für ein $i \le j$, so nennen wir das Runge–Kutta Verfahren **implizit**. □

Beispiel 4.14. Das implizite Euler Verfahren

$$y_{n+1} = y_n + h f(x_{n+1}, y_{n+1})$$

ist offenbar ein einstufiges implizites Runge–Kutta Verfahren.

Beispiel 4.15. *Ein weiteres naheliegendes Verfahren ist die* **implizite Mittelpunktregel**

$$\begin{aligned} k_1 &= f\Big(x_n + \tfrac{h}{2}, y_n + \tfrac{1}{2}hk_1\Big) \\ y_{n+1} &= y_n + hk_1, \end{aligned} \tag{4.14}$$

die offenbar ebenfalls ein einstufiges implizites Runge–Kutta Verfahren ist.

Fr $f(x,y) = \lambda y$ gilt $k_1 = \lambda(y_n + \frac{1}{2}hk_1)$, d.h. $k_1 = \lambda y_n/(1 - 0.5h\lambda)$, und damit

$$y_{n+1} = y_n + \frac{\lambda h y_n}{1 - 0.5\lambda h} = \frac{2-z}{2+z} y_n, \ z = \lambda h.$$

Die A-Stabilitt folgt nun wie fr die Trapezregel. Wegen $\lim_{z\to\infty} R(z) = -1$ ist die implizite Mittelpunktregel aber nicht L-stabil.

Beispiel 4.16. *Die Trapezregel*

$$y_{n+1} = y_n + \frac{1}{2}h\Big(f(x_n, y_n) + f(x_{n+1}, y_{n+1})\Big)$$

kann als zweistufiges implizites Runge–Kutta Verfahren aufgefasst werden:

$$\begin{aligned} k_1 &= f(x_n, y_n), \\ k_2 &= f\Big(x_n + h, y_n + h(\frac{1}{2}k_1 + \frac{1}{2}k_2)\Big), \\ y_{n+1} &= y_n + h\Big(\frac{1}{2}k_1 + \frac{1}{2}k_2\Big). \end{aligned}$$

□

Wie in Kapitel 2 die expliziten Runge–Kutta Verfahren stellt man auch die impliziten Verfahren übersichtlich in einem Tableau dar:

$$\begin{array}{c|cccc} \alpha_1 & \beta_{11} & \beta_{12} & \dots & \beta_{1s} \\ \alpha_2 & \beta_{21} & \beta_{22} & \dots & \beta_{2s} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha_s & \beta_{s1} & \beta_{s2} & \dots & \beta_{ss} \\ \hline & \gamma_1 & \gamma_2 & \dots & \gamma_s. \end{array}$$

Hiermit erhalten die Verfahren aus Beispiel 4.14. die Gestalt

$$\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array} \qquad \begin{array}{c|c} 0.5 & 0.5 \\ \hline & 1 \end{array} \qquad \begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 0.5 & 0.5 \\ \hline & 0.5 & 0.5 \end{array}$$

implizites Euler V. Mittelpunktregel Trapezregel.

Nachteil der impliziten Runge–Kutta Verfahren ist, dass die k_j nicht nacheinander berechnet werden können, sondern dass in jedem Schritt ein (i.a. nichtlineares) Gleichungssystem von $s\cdot N$ Gleichungen in den $s\cdot N$ Unbekannten $k_1, \ldots, k_s$ gelöst werden muss, wobei N die Dimension des Differentialgleichungssystems (2.1) bezeichnet. Eine naheliegende Frage ist, unter welchen Bedingungen das System (eindeutig) lösbar ist.

Satz 4.17. *Es sei $f : [a,b]\times\mathbb{R}^N \to \mathbb{R}^N$ stetig und erfülle eine Lipschitz Bedingung bzgl. des zweiten Arguments mit der Lipschitz Konstante L. Erfüllt die Schrittweite $h > 0$ die Bedingung*

$$h < \frac{1}{L \max\limits_{i=1,\ldots,s} \sum\limits_{j=1}^{s} |\beta_{ij}|}, \tag{4.15}$$

so ist das Gleichungssystem

$$k_i = f\Big(x_n + \alpha_i h, y_n + h\sum_{j=1}^{s} \beta_{ij} k_j\Big),\ i = 1, \ldots, s, \tag{4.16}$$

eindeutig lösbar.

Beweis: Wir zeigen, dass die Abbildung

$$T : (k_i)_{i=1,\ldots,s} \mapsto \Big(f(x_n + \alpha_i h, y_n + h\sum_{j=1}^{s} \beta_{ij} k_j)\Big)_{i=1,\ldots,s}$$

kontrahierend auf $\mathbb{R}^{sN}$ ist und daher einen eindeutigen Fixpunkt besitzt.

Es ist

$$\begin{aligned}
&\|T((k_i)_{i=1,\ldots,s}) - T((\tilde{k}_i)_{i=1,\ldots,s})\|_\infty \\
&= \max_{i=1,\ldots,s} \Big\| f(x_n + \alpha_i h, y_n + h\sum_{j=1}^{s} \beta_{ij} k_j) - f(x_n + \alpha_i h, y_n + h\sum_{j=1}^{s} \beta_{ij} \tilde{k}_j)\Big\|_\infty \\
&\leq \max_{i=1,\ldots,s} L\Big\|\Big(y_n + h\sum_{j=1}^{s} \beta_{ij} k_j\Big) - \Big(y_n + h\sum_{j=1}^{s} \beta_{ij} \tilde{k}_j\Big)\Big\|_\infty \\
&= Lh \max_{i=1,\ldots,s} \Big\| \sum_{j=1}^{s} \beta_{ij}(k_j - \tilde{k}_j)\Big\|_\infty \\
&\leq Lh\|(k_j)_{j=1,\ldots,s} - (\tilde{k}_j)_{j=1,\ldots,s}\|_\infty \max_{i=1\ldots,s} \sum_{j=1}^{s} |\beta_{ij}|.
\end{aligned}$$

∎

Erfüllt die rechte Seite f nur eine Lipschitzbedingung in einer Umgebung von (x_n, y_n), so bleibt die Aussage des Satzes immer noch richtig, wobei die Schrittweite eventuell verkleinert werden muss, um zu sichern, dass das Argument von f in dieser Umgebung bleibt.

Aus dem Fixpunktsatz für kontrahierende Abbildungen folgt zugleich, dass die Fixpunktiteration gegen die Lösung $(k_i)_{i=1,\dots,s}$ von (4.16) konvergiert, wenn die Schrittweite h die Bedingung (4.15) erfüllt. Für steife Probleme ist diese Aussage aber wertlos, da hierfür i.a. die Lipschitzkonstante recht groß sein wird. Bei steifen Systemen muss das Gleichungssystem (4.16) immer mit dem Newton Verfahren oder einer verwandten Methode gelöst werden.

Der Aufwand zur Lösung des Systems (4.16) kann wesentlich verkleinert werden, wenn das System (4.16) in s (oder weniger) Systeme der Dimension N zerfällt. Dies ist dann der Fall, wenn die Berechnung von k_j unabhängig von den folgenden Werten $k_{j+1},\dots,k_s$ ist. Dies ist der Fall, wenn die Matrix (β_{ij}) eine untere Dreiecksmatrix ist.

Definition 4.18. *Das implizite Runge–Kutta Verfahren (4.13) heißt* **diagonal implizites Runge–Kutta Verfahren** *(***DIRK–Verfahren** *), wenn $\beta_{ij} = 0$ für alle $i < j$ gilt und $\beta_{ii} \neq 0$ für wenigstens ein $i \in \{1,\dots,s\}$.*

Stimmen zusätzlich alle Diagonalelemente überein, $\beta_{ii} = \beta$ für alle $i = 1,\dots,s$, so heißt das Verfahren **einfach diagonal implizit** *(***SDIRK** *= singly diagonal implicit Runge–Kutta method).*

Beispiel 4.19. Zweistufige Verfahren der Ordnung 3 müssen die Bedingungen

$$\gamma_1 + \gamma_2 = 1,\ \gamma_1\alpha_1 + \gamma_2\alpha_2 = \frac{1}{2},\ \gamma_1\alpha_1^2 + \gamma_2\alpha_2^2 = \frac{1}{3},$$

$$\gamma_1(\alpha_1\beta_{11} + \alpha_2\beta_{12}) + \gamma_2(\alpha_1\beta_{21} + \alpha_2\beta_{22}) = \frac{1}{6}$$

erfüllen.

Von Nørsett wurde 1974 gezeigt, dass dieses Gleichungssystem unter der zusätzlichen Voraussetzung, dass $\beta_{12} = 0$ ist (DIRK) und $\beta_{11} = \beta_{22}$ (SDIRK) gilt, zwei Lösungen besitzt, die durch das folgende Tableau gekennzeichnet sind:

$$\begin{array}{c|cc} \tau & \tau & 0 \\ 1-\tau & 1-2\tau & \tau \\ \hline & 0.5 & 0.5 \end{array}, \quad \tau = \frac{3 \pm \sqrt{3}}{6}$$

Die Stabilitätsgebiete sind in Abbildung 4.9 gegeben. Aus ihnen entnimmt man, dass das Verfahren für $\tau = (3+\sqrt{3})/6$ A–stabil ist, für $\tau = (3-\sqrt{3})/6$ aber nicht. □

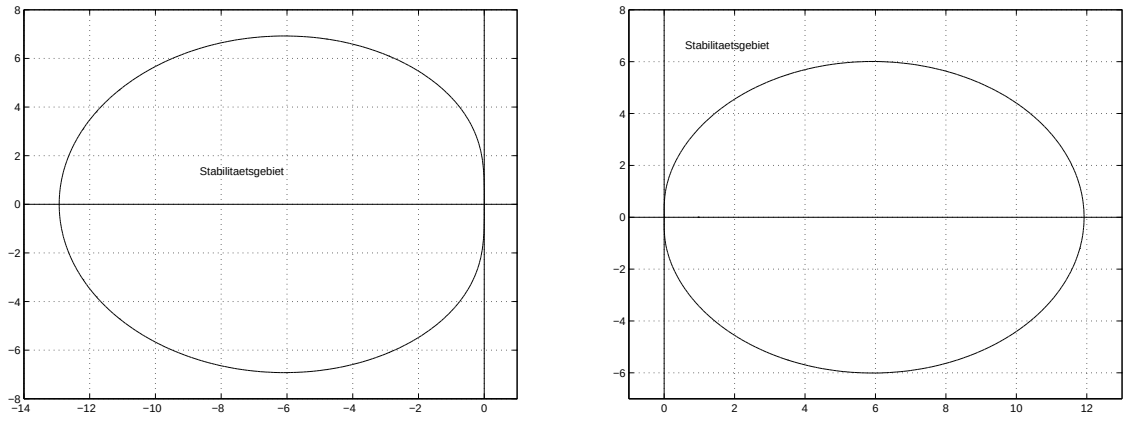

Abbildung 4.9: Stabilitätsgebiete des SDIRK Verfahrens der Ordnung 3

Das Stabilitätsgebiet haben wir unter Verwendung des folgenden Lemmas bestimmt.

Lemma 4.20. *Wird das s–stufige implizite Runge–Kutta Verfahren (4.13) auf die Testgleichung $y' = \lambda y$ angewandt, so gilt*

$$y_{n+1} = R(\lambda h) y_n$$

mit

$$R(z) = 1 + z\boldsymbol{\gamma}^T (\boldsymbol{I} - z\boldsymbol{B})^{-1} \boldsymbol{e}, \tag{4.17}$$

wobei $\boldsymbol{B} = (\beta_{ij})_{i,j=1,\ldots,s}$ und $\boldsymbol{e} = (1, \ldots, 1)^T$.

Beweis: Es gilt für $f(x, y) = \lambda y$

$$k_i = \lambda\Big(y_n + h \sum_{j=1}^{s} \beta_{ij} k_j\Big), \ i = 1, \ldots, s,$$

d.h. mit $z := \lambda h$

$$\boldsymbol{k} = \lambda y_n \boldsymbol{e} + z\boldsymbol{B}\boldsymbol{k} \quad \Longleftrightarrow \quad \boldsymbol{k} = \lambda y_n (\boldsymbol{I} - z\boldsymbol{B})^{-1} \boldsymbol{e},$$

und damit

$$y_{n+1} = y_n + h \sum_{i=1}^{s} \gamma_i k_i = y_n\Big(1 + z\boldsymbol{\gamma}^T (\boldsymbol{I} - z\boldsymbol{B})^{-1} \boldsymbol{e}\Big). \qquad \blacksquare$$

Beispiel 4.21. Für das SDIRK Verfahren aus Beispiel 4.19. erhält man hiermit

$$\begin{aligned} R(z) &= 1 + z(0.5, 0.5) \begin{pmatrix} 1 - \tau z & 0 \\ (2\tau - 1)z & 1 - \tau z \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\ &= 1 + \frac{z}{2(1 - \tau z)^2} (2 + z - 4\tau z) \\ &= \frac{1 + (1 - 2\tau)z + (0.5 - 2\tau + \tau^2) z^2}{(1 - \tau z)^2}. \end{aligned}$$

Eine andere Möglichkeit, die Funktion R zu bestimmen, liefert die Cramersche Regel:

Lemma 4.22. *Mit den Bezeichnungen von Lemma 4.20. gilt*

$$R(z) = \frac{\det(\boldsymbol{I} - z\boldsymbol{B} + z\boldsymbol{e}\boldsymbol{\gamma}^T)}{\det(\boldsymbol{I} - z\boldsymbol{B})}.$$

Beweis: Wendet man das s-stufige Runge Kutta Verfahren auf die Testgleichung an, so erhält man y_{n+1} aus y_n mit Hilfe das linearen Gleichungssystems

$$\begin{pmatrix} \boldsymbol{I} - z\boldsymbol{B} & \boldsymbol{0} \\ -z\boldsymbol{\gamma}^T & 1 \end{pmatrix} \begin{pmatrix} \boldsymbol{k}/\lambda \\ y_{n+1} \end{pmatrix} = \begin{pmatrix} \boldsymbol{e} \\ 1 \end{pmatrix} y_n.$$

Aus der Cramerschen Regel folgt

$$y_{n+1} = \frac{P(z)}{Q(z)} y_n$$

mit

$$P(z) = \det \begin{pmatrix} \boldsymbol{I} - z\boldsymbol{B} & \boldsymbol{e} \\ -z\boldsymbol{\gamma}^T & 1 \end{pmatrix} = \det \begin{pmatrix} \boldsymbol{I} - z\boldsymbol{B} + z\boldsymbol{e}\boldsymbol{\gamma}^T & \boldsymbol{0} \\ -z\boldsymbol{\gamma}^T & 1 \end{pmatrix} = \det(\boldsymbol{I} - z\boldsymbol{B} + z\boldsymbol{e}\boldsymbol{\gamma}^T)$$

und

$$Q(z) = \det(\boldsymbol{I} - z\boldsymbol{B}). \qquad \blacksquare$$

Besitzt der Nenner $Q(z)$ der rationalen Funktion $R(z)$ keine Nullstellen in der linken Halbebene $\mathbb{C}_-$, so ist $R(z)$ holomorph in $\mathbb{C}_-$, und das Maximumprinzip für holomorphe Funktionen liefert die A–Stabilität jedes zu $R(z)$ gehörenden Verfahrens, falls nur

$$|R(iy)| \leq 1 \quad \text{für alle } y \in \mathbb{R}. \tag{4.18}$$

Diese Bedingung wiederum ist äquivalent dazu, dass für das reelle Polynom

$$E(y) := |Q(iy)|^2 - |P(iy)|^2 = Q(iy)Q(-iy) - P(iy)P(-iy) \tag{4.19}$$

gilt

$$E(y) \geq 0 \quad \text{für alle } y \in \mathbb{R}.$$

Für das SDIRK Verfahren aus Beispiel 4.19. gilt

$$E(y) = (\tau - 0.5)^2(4\tau - 1)y^4.$$

Da für $\tau > 0$ der einzige Pol $z = 1/\tau$ in der rechten Halbebene liegt, ist dieses Verfahren A–stabil für $\tau \geq 0.25$. Insbesondere ist das Verfahren der Ordnung 3 A–stabil für $\tau = (3 + \sqrt{3})/6$ und nicht A–stabil für $\tau = (3 - \sqrt{3})/6$.

Lemma 4.23. *Ist die Matrix $\boldsymbol{B}$ in dem impliziten Runge-Kutta Verfahren (4.13) regulär und gilt eine der beiden Bedingungen*

$$\beta_{sj} = \gamma_j, \ j = 1, \ldots, s, \tag{4.20}$$

oder

$$\beta_{i1} = \gamma_1, \ i = 1, \ldots, s, \tag{4.21}$$

so gilt

$$\lim_{z\to\infty} R(z) = 0. \tag{4.22}$$

Ein A–stabiles Verfahren, das (4.20) oder (4.21) erfüllt, ist also sogar L–stabil.

Beweis: Die Bedingung (4.20) kann man schreiben als

$$\boldsymbol{B}^T \boldsymbol{e}^s = \boldsymbol{\gamma},$$

und daher gilt

$$\lim_{z\to\infty} R(z) = \lim_{z\to\infty} (1 + z\boldsymbol{\gamma}^T(\boldsymbol{I} - z\boldsymbol{B})^{-1}\boldsymbol{e}) = 1 - \boldsymbol{\gamma}^T\boldsymbol{B}^{-1}\boldsymbol{e} = 1 - (\boldsymbol{e}^s)^T\boldsymbol{e} = 0.$$

Die Bedingung (4.21) lautet $\boldsymbol{B}\boldsymbol{e}^1 = \gamma_1\boldsymbol{e}$. Da $\boldsymbol{B}$ regulär ist, ist $\gamma_1 \neq 0$, und es folgt

$$\lim_{z\to\infty} R(z) = 1 - \boldsymbol{\gamma}^T\boldsymbol{B}^{-1}\boldsymbol{e} = 1 - \frac{1}{\gamma_1}\boldsymbol{\gamma}^T\boldsymbol{e}^1 = 0. \qquad \blacksquare$$

Ein eleganter Weg zur Konstruktion impliziter Runge–Kutta Verfahren hoher Ordnung ist die **Kollokation**. Er besteht darin, ein Polynom $u(x)$ vom Grad s zu bestimmen, so dass die Ableitung (ein Polynom vom Grad $s-1$) an s gegebenen Stellen $x_n + \alpha_i h$, $i = 1, \ldots, s$, mit dem Vektorfeld der Differentialgleichung übereinstimmt, d.h.

$$u(x_n) = y_n \tag{4.23}$$

$$u'(x_n + \alpha_i h) = f(x_n + \alpha_i h, u(x_n + \alpha_i h)), \ i = 1, \ldots, s. \tag{4.24}$$

Die Näherung für die Lösung an der Stelle $x_n + h$ ist dann

$$y_{n+1} := u(x_n + h). \tag{4.25}$$

Dass dieses Vorgehen als implizites Runge–Kutta Verfahren gedeutet werden kann, sieht man so ein: Es sei

$$k_i := u'(x_n + \alpha_i h).$$

Dann gilt nach der Lagrangeschen Interpolationsformal

$$u'(x_n + th) = \sum_{j=1}^{s} k_j \ell_j(t), \quad \ell_j(t) = \prod_{k \neq j} \frac{t - \alpha_k}{\alpha_j - \alpha_k}. \tag{4.26}$$

Integriert man diese Gleichung, so erhält man mit

$$\begin{aligned} k_i &= u'(x_n + \alpha_i h) \\ \beta_{ij} &:= \int_0^{\alpha_i} \ell_j(t)\, dt \\ \gamma_j &:= \int_0^1 \ell_j(t)\, dt \end{aligned}$$

die Formel (4.13) des Runge–Kutta Verfahrens, denn es gilt

$$\begin{aligned} y_{n+1} &= u(x_n + h) = u(x_n) + h \int_0^1 u'(x_n + th)\, dt \\ &= y_n + h \sum_{j=1}^{s} \int_0^1 k_j \ell_j(t)\, dt = y_n + h \sum_{j=1}^{s} \gamma_j k_j. \end{aligned}$$

Zu zeigen bleibt

$$k_i = f(x_n + \alpha_i h, y_n + h \sum_{j=1}^{s} \beta_{ij} k_j).$$

Dies folgt aus

$$k_i = u'(x_n + \alpha_i h) = f(x_n + \alpha_i h, u(x_n + \alpha_i h))$$

und

$$\begin{aligned} u(x_n + \alpha_i h) &= y_n + h \int_0^{\alpha_i} u'(x_n + th)\, dt = y_n + h \int_0^{\alpha_i} \sum_{j=1}^{s} u'(x_n + \alpha_j h) \ell_j(t)\, dt \\ &= y_n + h \sum_{j=1}^{s} \beta_{ij} u'(x_n + \alpha_j h) = y_n + h \sum_{j=1}^{s} \beta_{ij} k_j. \end{aligned}$$

Für die Konsistenzordnung gilt

Satz 4.24. *Dem Kollokationsverfahren zu den Knoten α_j ist in natürlicher Weise eine Quadraturformel mit den Knoten α_j und den Gewichten $\gamma_j := \int_0^1 \ell_j(t)\, dt$ zugeordnet:*

$$\int_{x_n}^{x_n+h} f(x)\, dx = h \sum_{j=1}^{s} \gamma_j f(x_n + \alpha_j h) + O(h^{p+1}). \tag{4.27}$$

Besitzt diese die Ordnung p, so hat das Kollokationsverfahren (4.23), (4.24), (4.25) ebenfalls die Ordnung p.

Beweis: Wir beweisen diese Aussage nur für lineare Differentialgleichungssysteme

$$\boldsymbol{y}' = \boldsymbol{A}(x)\boldsymbol{y} + \boldsymbol{f}(x).$$

Den Beweis für den allgemeinen Fall findet man in *Deuflhard, Bornemann* [16] p. 244 ff.

Nach Konstruktion gilt

$$\boldsymbol{u}'(x) = \boldsymbol{A}(x)\boldsymbol{u} + \boldsymbol{f}(x) + \boldsymbol{\delta}(x)$$

mit

$$\boldsymbol{\delta}(x_n + \alpha_j h) = \boldsymbol{0},\ j = 1, \ldots, s.$$

Die Differenzfunktion

$$\boldsymbol{v}(x) := \boldsymbol{u}(x) - \boldsymbol{y}(x)$$

erfüllt die Anfangswertaufgabe

$$\boldsymbol{v}'(x) = \boldsymbol{A}(x)\boldsymbol{v}(x) + \boldsymbol{\delta}(x),\ \boldsymbol{v}(x_n) = \boldsymbol{0}.$$

Ist $\boldsymbol{V}$ eine Fundamentalmatrix des homogenen Problems $\boldsymbol{w}' = \boldsymbol{A}(x)\boldsymbol{w}$, so erhält man durch Variation der Konstanten

$$\boldsymbol{v}(x) = \boldsymbol{V}(x) \int_{x_n}^{x} \boldsymbol{V}^{-1}(\xi)\boldsymbol{\delta}(\xi)\, d\xi,$$

und damit

$$\begin{aligned}
\boldsymbol{u}(x_n + h) - \boldsymbol{y}(x_n + h) &= \boldsymbol{V}(x_n + h) \int_{x_n}^{x_n+h} \boldsymbol{V}^{-1}(\xi)\boldsymbol{\delta}(\xi)\, d\xi \\
&= h\boldsymbol{V}(x_n + h) \sum_{j=1}^{s} \gamma_j \boldsymbol{V}^{-1}(x_n + \alpha_j h)\boldsymbol{\delta}(x_n + \alpha_j h) + O(h^{p+1}) \\
&= O(h^{p+1}), \qquad \text{wegen } \boldsymbol{\delta}(x_n + \alpha_j h) = \boldsymbol{0},\ j = 1 \ldots, s.
\end{aligned}$$

∎

Nach diesem Ergebnis ist klar, welche Ordnung für implizite Runge–Kutta Verfahren man durch Kollokation maximal erreichen kann und wie man diese erreicht.

Sind die α_j die Knoten der Gaußschen Quadraturformel der Ordnung 2s, d.h. die Nullstellen des s–ten (geschifteten) Legendre Polynoms

$$\frac{d^s}{dx^s}\Big(x^s(1-x)^s\Big),$$

und sind γ_j die zugehörigen Gewichte, so heißt das durch (4.23), (4.24), (4.25) definierte implizite Runge–Kutta Verfahren s–**stufiges Gauß Verfahren** . Das s-stufige Gauß Verfahren hat die Ordnung $2s$.

Es kann keine impliziten Runge–Kutta Formeln der Stufe s mit höherer Konsistenzordnung als $2s$ geben, denn sonst hätte man zugleich durch (4.27) eine Quadraturformel mit s Knoten und einer höheren Ordnung als $2s$ gefunden.

Man kann ferner zeigen, dass das Gauß Verfahren der Ordnung $2s$ A–stabil ist (vgl. *Hairer, Wanner* [30] p. 72).

Das einstufige Gauß Verfahren ist die implizite Mittelpunktregel, das zweistufige Gauß Verfahren wird durch das Tableau

$$\begin{array}{c|cc} \frac{1}{2}-\frac{\sqrt{3}}{6} & \frac{1}{4} & \frac{1}{4}-\frac{\sqrt{3}}{6} \\ \frac{1}{2}+\frac{\sqrt{3}}{6} & \frac{1}{4}+\frac{\sqrt{3}}{6} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

gegeben und das dreistufige Verfahren durch

$$\begin{array}{c|ccc} \frac{1}{2}-\frac{\sqrt{15}}{10} & \frac{5}{36} & \frac{2}{9}-\frac{\sqrt{15}}{15} & \frac{5}{36}-\frac{\sqrt{15}}{30} \\ \frac{1}{2} & \frac{5}{36}+\frac{\sqrt{15}}{24} & \frac{2}{9} & \frac{5}{36}-\frac{\sqrt{15}}{24} \\ \frac{1}{2}+\frac{\sqrt{15}}{10} & \frac{5}{36}+\frac{\sqrt{15}}{30} & \frac{2}{9}+\frac{\sqrt{15}}{15} & \frac{5}{36} \\ \hline & \frac{5}{18} & \frac{4}{9} & \frac{5}{18} \end{array}$$

Entsprechend kann man mit den Quadraturformeln von Lobatto und von Radau implizite Runge–Kutta Formeln herleiten. Die **Radau Formeln** sind charakterisiert durch

$$\alpha_1 = 0,\ \beta_{11} = \beta_{12} = \ldots = \beta_{1s} = 0 \quad \text{oder} \quad \alpha_s = 1,\ \beta_{sj} = \gamma_j,\ j = 1, \ldots, s,$$

und die **Lobatto Formeln** durch

$$\alpha_1 = 0,\ \beta_{11} = \beta_{12} = \ldots = \beta_{1s} = 0 \quad \text{und} \quad \alpha_s = 1,\ \beta_{sj} = \gamma_j,\ j = 1, \ldots, s.$$

Die Radau Formeln haben den Vorteil, dass k_1 oder k_s explizit berechnet werden kann, und bei den Lobatto Formeln können sogar k_1 und k_s explizit berechnet werden. Die Dimension des nichtlinearen Systems, das in jedem Schritt zu lösen ist, sinkt also gegenüber den Gauß Formeln. Der Preis dafür ist, dass die Konsistenzordnung der Radau Formeln nur $2s-1$ und der Lobatto Formeln nur $2s-2$ ist.

Wir geben die wichtigsten impliziten Formeln vom Radau und Lobatto Typ an: Zweistufige Radau Formeln (der Ordnung 3) sind

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \frac{2}{3} & \frac{1}{3} & \frac{1}{3} \\ \hline & \frac{1}{4} & \frac{3}{4} \end{array} \quad \text{und} \quad \begin{array}{c|cc} \frac{1}{3} & \frac{5}{12} & -\frac{1}{12} \\ 1 & \frac{3}{4} & \frac{1}{4} \\ \hline & \frac{3}{4} & \frac{1}{4} \end{array} .$$

Die Lobatto Formel mit 2 Stufen (der Ordnung 2) ist

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

d.h. die Trapezregel, und mit 3 Stufen (der Ordnung 4)

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{5}{24} & \frac{8}{24} & -\frac{1}{24} \\ 1 & \frac{1}{6} & \frac{4}{6} & \frac{1}{6} \\ \hline & \frac{1}{6} & \frac{4}{6} & \frac{1}{6} \end{array},$$

4.4 Rosenbrock Verfahren

Wir haben bereits bemerkt, dass die nichtlinearen Gleichungssysteme, die in den impliziten Runge–Kutta Verfahren auftreten, nicht durch sukzessive Iteration gelöst werden können, sondern dass eine Linearisierung verwendet werden muss. Um dies zu umgehen, wurde von *Rosenbrock* [41] 1962 vorgeschlagen, nur den linearen Anteil der rechten Seite mit einem impliziten Verfahren zu behandeln und den Rest mit einem expliziten Verfahren. Wir erläutern das Vorgehen für das autonome System

$$\boldsymbol{y}' = f(\boldsymbol{y}). \tag{4.28}$$

Dies bedeutet keine Einschränkung, denn das nichtautonome System

$$\boldsymbol{y}' = f(x, \boldsymbol{y})$$

kann man durch Hinzunahme der Differentialgleichung

$$x' = 1$$

in ein autonomes System

$$\begin{aligned} \boldsymbol{y}' &= f(x, \boldsymbol{y}) \\ x' &= 1 \end{aligned} \tag{4.29}$$

transformieren.

(4.28) schreiben wir mit $\boldsymbol{J} := f'(\boldsymbol{y}_n)$ als

$$\boldsymbol{y}' = (f(\boldsymbol{y}_n) + \boldsymbol{J}(\boldsymbol{y} - \boldsymbol{y}_n)) + (f(\boldsymbol{y}) - f(\boldsymbol{y}_n) - \boldsymbol{J}(\boldsymbol{y} - \boldsymbol{y}_n)).$$

Wir zerlegen die rechte Seite also in die Linearisierung von f bei $\boldsymbol{y}_n$ (von der wir annehmen können, dass sie für das steife Verhalten verantwortlich ist) und (bei kleiner Schrittweite) einen kleinen Rest. Wir integrieren den linearen Anteil implizit und den kleinen Anteil explizit. Da dabei von beiden Formeln der konstante Anteil $f(\boldsymbol{y}_n) - \boldsymbol{J}\boldsymbol{y}_n$ exakt integriert werden wird, wenden wir die Idee auf die Zerlegung

$$\boldsymbol{y}' = \boldsymbol{J}\boldsymbol{y}(x) + (f(\boldsymbol{y}(x)) - \boldsymbol{J}\boldsymbol{y}(x)) \tag{4.30}$$

der rechten Seite von (4.28) an und berechnen hierfür für $i = 1, \ldots, s$ die "Steigungen"

$$\boldsymbol{k}_i = \boldsymbol{J}\Big(\boldsymbol{y} + h\sum_{j=1}^{i} \tilde{\beta}_{ij}\boldsymbol{k}_j\Big) + \Big(f(\boldsymbol{y} + h\sum_{j=1}^{i-1} \beta_{ij}\boldsymbol{k}_j) - \boldsymbol{J}(\boldsymbol{y} + h\sum_{j=1}^{i-1} \beta_{ij}\boldsymbol{k}_j)\Big). \tag{4.31}$$

Beachten Sie, dass das aus (4.31) zu berechnende $\boldsymbol{k}_i$ auf der rechten Seite nur linear auftritt und nur die bereits bekannten $\boldsymbol{k}_1, \ldots, \boldsymbol{k}_{i-1}$ auch im Argument der nichtlinearen Funktion f. Man kann also $\boldsymbol{k}_i$ aus einem linearen Gleichungssystem berechnen.

Definition 4.25. *Ein Schritt eines* **linear impliziten Runge–Kutta Verfahrens** *oder* **Rosenbrock Verfahrens** *mit s Stufen für das autonome System (4.28) besteht aus den folgenden Schritten*

$$\begin{aligned}
&(i) && \boldsymbol{J} &&= f'(\boldsymbol{y}_n) \\
&(ii) && (\boldsymbol{I} - h\tilde{\beta}_{ii}\boldsymbol{J})\boldsymbol{k}_i &&= h\sum_{j=1}^{i-1}(\tilde{\beta}_{ij} - \beta_{ij})\boldsymbol{J}\boldsymbol{k}_j + f(\boldsymbol{y}_n + h\sum_{j=1}^{i-1}\beta_{ij}\boldsymbol{k}_j), \\
& && && \quad i = 1, \ldots, s \\
&(iii) && \boldsymbol{y}_{n+1} &&= \boldsymbol{y}_n + h\sum_{j=1}^{s}\gamma_j\boldsymbol{k}_j
\end{aligned} \tag{4.32}$$

In jedem Schritt sind s lineare Gleichungssysteme zu lösen. Eine weitere Vereinfachung tritt ein, wenn man ähnlich wie bei den SDIRK Verfahren $\tilde{\beta}_{ii} = \tilde{\beta}$ fordert. Man benötigt dann nur eine LR–Zerlegung zur Lösung der s Systeme.

Bemerkung 4.26. Wendet man das linear implizite Verfahren (4.32) auf die Testgleichung $y' = \lambda y$ an, so fallen alle Terme, die β_{ij} enthalten, heraus und wie für die impliziten Verfahren sieht man, dass die Stabilitätsfunktion wieder die Gestalt

$$R(z) = 1 + z\boldsymbol{\gamma}^T(\boldsymbol{I} - z\boldsymbol{B})^{-1}\boldsymbol{e}$$

hat, wobei nun $\boldsymbol{B} := (\tilde{\beta}_{ij})_{i,j=1,\ldots,s}$ gesetzt ist. □

Beispiel 4.27. Das einfachste linear implizite Runge–Kutta Verfahren ist das **linear implizite Euler Verfahren**

$$\begin{aligned} (\boldsymbol{I} - hf'(\boldsymbol{y}_n))\boldsymbol{k}_1 &= f(\boldsymbol{y}_n) \\ \boldsymbol{y}_{n+1} &= \boldsymbol{y}_n + h\boldsymbol{k}_1 \end{aligned} \tag{4.33}$$

Dieses besitzt dieselbe Stabilitätsfunktion wie das implizite Euler Verfahren und ist daher ebenfalls L–stabil. Die Konsistenzordnung ist ebenfalls 1. □

Beispiel 4.28. Von Steihaug und Wolfbrandt [52] wurde das folgende linear implizite Runge–Kutta Verfahren zweiter Ordnung angegeben:

$$\begin{aligned} \boldsymbol{W}\boldsymbol{k}_1 &= f(\boldsymbol{y}_n) \\ \boldsymbol{W}\boldsymbol{k}_2 &= f\left(\boldsymbol{y}_n + \tfrac{2}{3}h\boldsymbol{k}_1\right) - \tfrac{4}{3}hd\boldsymbol{J}\boldsymbol{k}_1 \\ \boldsymbol{y}_{n+1} &= \boldsymbol{y}_n + \frac{h}{4}(\boldsymbol{k}_1 + 3\boldsymbol{k}_2), \end{aligned} \tag{4.34}$$

wobei $\boldsymbol{W} = \boldsymbol{I} - hd\boldsymbol{J}$, $\boldsymbol{J} = f'(\boldsymbol{y}_n)$ und $d = 1/(2+\sqrt{2})$ ist.

Bemerkung 4.29. Wendet man das linear implizite Verfahren (4.32) auf das "autonomisierte" System (4.29) an, so kann man die zu der Variablen x gehörige Gleichung explizit ausrechnen. Man erhält die folgende Form des Verfahrens

$$\begin{aligned} \boldsymbol{J} &= \frac{\partial}{\partial \boldsymbol{y}} f(x_n, \boldsymbol{y}_n), \\ (\boldsymbol{I} - h\tilde{\beta}_{ii}\boldsymbol{J})\boldsymbol{k}_i &= f(x_n + \alpha_i h, \boldsymbol{y}_n + h\sum_{j=1}^{i-1}\beta_{ij}\boldsymbol{k}_j) + \delta_i h\frac{\partial}{\partial x} f(x_n, \boldsymbol{y}_n) \\ &\quad + h\boldsymbol{J}\sum_{j=1}^{i-1}(\tilde{\beta}_{ij} - \beta_{ij})\boldsymbol{k}_j,\ i = 1, \ldots, s, \\ \boldsymbol{y}_{n+1} &= \boldsymbol{y}_n + h\sum_{j=1}^{s}\gamma_j\boldsymbol{k}_j \end{aligned} \tag{4.35}$$

Dabei ist

$$\alpha_i := \sum_{j=1}^{i-1}\beta_{ij},\ \delta_i := \tilde{\beta}_{ii} + \sum_{j=1}^{i-1}(\tilde{\beta}_{ij} - \beta_{ij}).$$

□

Beispiel 4.30. Für nichtautonome Anfangswertaufgaben erhält das linear implizite Euler Verfahren die Gestalt

$$\begin{aligned} \left(\boldsymbol{I} - h\frac{\partial}{\partial \boldsymbol{y}} f(x_n, \boldsymbol{y}_n)\right)\boldsymbol{k}_1 &= f(x_n, \boldsymbol{y}_n) + h\frac{\partial}{\partial x} f(x_n, \boldsymbol{y}_n) \\ \boldsymbol{y}_{n+1} &= \boldsymbol{y}_n + h\boldsymbol{k}_1 \end{aligned} \tag{4.36}$$

Rosenbrock Verfahren höherer Ordnung wurden von Kaps und Rentrop (1979), Shampine (1982) und van Veldhuizen (1984) konstruiert. Kaps und Ostermann (1989) konstruierten eingebettete linear implizite Runge–Kutta Verfahren. Sie können ähnlich leicht implementiert werden wie explizite Runge–Kutta Verfahren. Den Code ROS4 findet man in

http://www.unige.ch/math/folks/hairer/

Bemerkung 4.31. In der ODE–Suite von MATLAB ist als ODE23s das folgende eingebettete Paar von linear impliziten Verfahren der Ordnungen 2 und 3 mit der FSAL Eigenschaft implementiert:

$$\begin{aligned}
\boldsymbol{f}_0 &= f(x_n, \boldsymbol{y}_n) \\
\boldsymbol{W}\boldsymbol{k}_1 &= \boldsymbol{f}_0 + hd\boldsymbol{t} \\
\boldsymbol{f}_1 &= f(x_n + 0.5h, \boldsymbol{y}_n + 0.5h\boldsymbol{k}_1) \\
\boldsymbol{W}\boldsymbol{k}_2 &= \boldsymbol{f}_1 - \boldsymbol{k}_1 + \boldsymbol{W}\boldsymbol{k}_1 \\
\boldsymbol{y}_{n+1} &= \boldsymbol{y}_n + h\boldsymbol{k}_2 \\
\boldsymbol{f}_2 &= f(x_{n+1}, \boldsymbol{y}_{n+1}) \\
\boldsymbol{W}\boldsymbol{k}_3 &= \boldsymbol{f}_2 - (6+\sqrt{2})(\boldsymbol{k}_2 - \boldsymbol{f}_1) - 2(\boldsymbol{k}_1 - \boldsymbol{f}_0) + hd\boldsymbol{t} \\
\text{Fehler} &\approx h(\boldsymbol{k}_1 - 2\boldsymbol{k}_2 + \boldsymbol{k}_3)/6
\end{aligned}$$

mit $d = 1/(2+\sqrt{2})$, $\boldsymbol{W} = \boldsymbol{I} - hd\boldsymbol{J}$ und

$$\boldsymbol{J} \approx \frac{\partial}{\partial \boldsymbol{y}} f(x_n, \boldsymbol{y}_n) \quad \text{und} \quad \boldsymbol{t} \approx \frac{\partial}{\partial \boldsymbol{x}} f(x_n, \boldsymbol{y}_n).$$

War ein Schritt erfolgreich, so kann $\boldsymbol{f}_2$ dieses Schrittes offenbar als $\boldsymbol{f}_0$ des folgenden Schrittes verwendet werden. Wählt man $\boldsymbol{J}$ als die Jacobi Matrix von f und nicht nur als eine Approximation, so ist das Verfahren L–stabil.

Bemerkung 4.32. Eine weitere Verringerung des Rechenaufwands erzielt man, wenn man die Jacobi Matrix (und ihre LR–Zerlegung) nicht in jedem Schritt neu berechnet, sondern über einige Schritte fest lässt. Um hierfür eine Konsistenzordnung p zu sichern, muss man die Ordnung unabhängig von der verwendeten Matrix $\boldsymbol{J}$ bestimmen. Solche Verfahren wurden von Steihaug und Wolfbrandt (1979) konstruiert. Sie bezeichneten in ihrer Arbeit die Approximation der Jacobi Matrix mit $\boldsymbol{W}$. Diese Verfahren werden daher in der Literatur W **Methoden** genannt. □

4.5 Extrapolation

Auch Extrapolation kann mit Erfolg auf steife Systeme angewendet werden. Von Bader und Deuflhard (1983) wurden Extrapolationsverfahren auf der Grundlage der linear impliziten Mittelpunktregel untersucht.

Es sei

$$\boldsymbol{J} \approx \frac{\partial}{\partial \boldsymbol{y}} f(x_0, \boldsymbol{y}_0)$$

eine Approximation der Jacobi Matrix. Für $x_i := x_0 + ih$ und $x := x_0 + 2mh$ berechne man

$$\begin{aligned}
(\boldsymbol{I} - h\boldsymbol{J})(\boldsymbol{y}_1 - \boldsymbol{y}_0) &= hf(x_0, \boldsymbol{y}_0), && (4.37)\\
(\boldsymbol{I} - h\boldsymbol{J})(\boldsymbol{y}_{i+1} - \boldsymbol{y}_i) &= -(\boldsymbol{I} + h\boldsymbol{J})(\boldsymbol{y}_i - \boldsymbol{y}_{i-1}) + 2hf(x_i, \boldsymbol{y}_i), && (4.38)\\
& \qquad i = 0, \ldots, 2m &&\\
\boldsymbol{S}_h(x) &= \frac{1}{2}(\boldsymbol{y}_{2m-1} - \boldsymbol{y}_{2m+1}). && (4.39)
\end{aligned}$$

(4.38) ist die linear implizite Mittelpunktregel. Um diese zu starten benötigt man $\boldsymbol{y}_1$. Diesen Wert beschafft man sich in (4.37) mit einem Schritt des linear impliziten Euler Verfahrens.

Für $\boldsymbol{S}_h(x)$ gilt die folgende Entwicklung des Fehlers (vgl. *Hairer, Wanner* [30] p. 135)

Satz 4.33. *Es sei f hinreichend glatt, und es sei $\boldsymbol{J}$ eine beliebige Matrix. Dann gilt für das mit (4.37), (4.38), (4.39) bestimmte $\boldsymbol{S}_h(x)$ eine Fehlerentwicklung der Gestalt*

$$\boldsymbol{y}(x) - \boldsymbol{S}_h(x) = \sum_{j=1}^{\ell} \boldsymbol{e}_j(x) h^{2j} + h^{2\ell+2} C(x, h) \qquad (4.40)$$

mit einer beschränkten Funktion C.

Da in der Fehlerentwicklung nur gerade Potenzen von h auftreten, wird man zur Extrapolation ein Polynom in h^2 verwenden. Mit

$$\boldsymbol{T}_{j1} := \boldsymbol{S}_{h_j}(x_0 + H),\ h_j := \frac{H}{n_j}$$

setzt man

$$\boldsymbol{T}_{j,k+1} = \boldsymbol{T}_{j,k} + \frac{\boldsymbol{T}_{j,k} - \boldsymbol{T}_{j-1,k}}{(n_j/n_{j-k})^2 - 1}.$$

Dann erhält man durch die $\boldsymbol{T}_{j,k}$ für alle k Approximationen der Ordnung $2k-1$ unabhängig von der Wahl von $\boldsymbol{J}$.

Tabelle 4.3: Extrapolation der linear impliziten Mittelpunktregel

$j=1$	$j=2$	$j=3$	$j=4$	$j=5$	$j=6$	$j=7$
90°						
90°	90°					
90°	90°	90°				
90°	89.3°	87.6°	87.3°			
90°	88.8°	86.9°	86.1°	86.0°		
90°	88.5°	87.3°	86.6°	86.4°	86.3°	
90°	88.4°	87.4°	87.0°	86.8°	86.7°	86.7°

Bemerkung 4.34. Die Matrix $\boldsymbol{J}$ in Satz 4.33. ist beliebig. Der der Extrapolation zu Grunde liegende Prozess ist also eine W Methode. Die sich durch Extrapolation ergebenden Näherungen kann man (wie man leicht nachrechnet) als Ergebnis von W Methoden interpretieren. Satz 4.33. sagt daher, dass es W Verfahren beliebiger Ordnung gibt. □

Die erste Spalte des Extrapolationsschemas besteht offenbar aus A–stabilen Approximationen der Lösung und man kann zeigen, dass auch $\boldsymbol{T}_{22}$, $\boldsymbol{T}_{32}$ und $\boldsymbol{T}_{33}$ A–stabil sind, die übrigen $\boldsymbol{T}_{jk}$ aber nicht. Sie sind jedoch noch A(α)–stabil. Die Öffnungswinkel der zugehörigen Sektoren sind in Tabelle 4.3 enthalten.

Sehr gute Ergebnisse erhält man ebenfalls mit der Extrapolation mit dem linear impiliziten Euler Verfahren

$$(\boldsymbol{I} - h\boldsymbol{J})(\boldsymbol{y}_{i+1} - \boldsymbol{y}_i) = hf(x_{i+1}, \boldsymbol{y}_i).$$

Es werden die Werte

$$\boldsymbol{T}_{j1} = \boldsymbol{y}_{h_j}(x_0 + H)$$

mit den Schrittweiten $h_j := H/n_j$ und $n_j = 1, 2, 3, 4, \ldots$ extrapoliert gemäß

$$\boldsymbol{T}_{j,k+1} = \boldsymbol{T}_{j,k} + \frac{\boldsymbol{T}_{j,k} - \boldsymbol{T}_{j-1,k}}{(n_j/n_{j-k}) - 1}.$$

Auch die $\boldsymbol{T}_{jk}$ sind A(α)–stabil mit den Öffnungswinkeln der Tabelle 4.4.

Implementierungen der Extrapolationsverfahren einschließlich einer Ordnungs– und Schrittweitensteuerung findet man für die linear impliziten Mittelpunktregel in den Programmen METAN1 von Bader und Deuflhard und SODEX von Hairer und Wanner, und für das linear implizite Euler Verfahren in dem Programm LIMEX von Deuflhard und Nowak. SODEX findet man in

http://www.unige.ch/math/folks/hairer/

Tabelle 4.4: Extrapolation des linear impliziten Euler Verfahrens

$j=1$	$j=2$	$j=3$	$j=4$	$j=5$	$j=6$	$j=7$	$j=8$
90°							
90°	90°						
90°	90°	89.85°					
90°	90°	89.90°	89.77°				
90°	90°	89.93°	89.84°	89.77°			
90°	90°	89.95°	89.88°	89.82°	89.87°		
90°	90°	89.96°	89.91°	89.86°	89.82°	89.80°	
90°	90°	89.97°	89.93°	89.89°	89.85°	89.83°	89.81°

und eine Übertragung von D. Abendroth in MATLAB in

http://www.tu-harburg.de/mat/LEHRE/Scripte.html

Die Codes METAN1 und LIMEX sind in

ftp://elib.zib.de/pub/elib/codelib/

In MATLAB sind 2 Methoden zur Lösung steifer Systeme implementiert. Das Programm ode15s verwendet BDF-Formeln oder **NDF Verfahren** der Ordnung $k \in \{1,2,3,4,5\}$. NDF sind Modifikationen der BDF Methoden, die ebenfalls A(α)-stabil sind. Sie besitzen eine etwas größere Genauigkeit als die BDF-Methoden, wobei der Öffnungswinkel α des maximalen im Stabilitätsgebiet enthaltenen Sektors aber nur wenig verkleinert wird. Ihre Konstruktion ist in *Shampine, Reichel* [49] beschrieben.

Das Programm ode23s verwendet ein Rosenbrock Verfahren der Ordnung 2, wobei der Fehler mit einer Modifikation der Ordnung 3 geschätzt wird. Es ist geeignet, wenn die Genauigkeitsansprüche nicht zu hoch sind.

Beispiel 4.35. Das folgende Beispiel, das die Entwicklung der Konzentrationen dreier Komponenten in einer chemischen Reaktion beschreibt, wurde von *Robertson* [40] 1967 angegeben und wird häufig verwendet, um die Eigenschaften von steifen Lösern zu demonstrieren:

$$\begin{aligned} y_1' &= -0.04y_1 + 10^4 y_2 y_3, & y_1(0) &= 1 \\ y_2' &= 0.04y_1 - 10^4 y_2 y_3 - 3\cdot 10^7 y_2^2, & y_2(0) &= 0 \\ y_3' &= 3\cdot 10^7 y_2^2, & y_3(0) &= 0. \end{aligned}$$

Abbildung 4.10 zeigt die Lösungen dieses Systems in halblogarithmischer Darstellung.

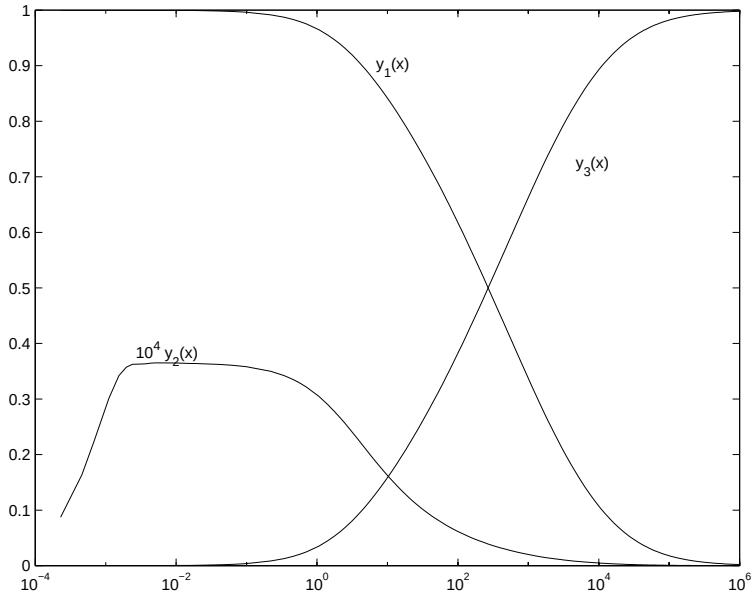

Abbildung 4.10: Lösungen von Beispiel 4.35. mit ode15s

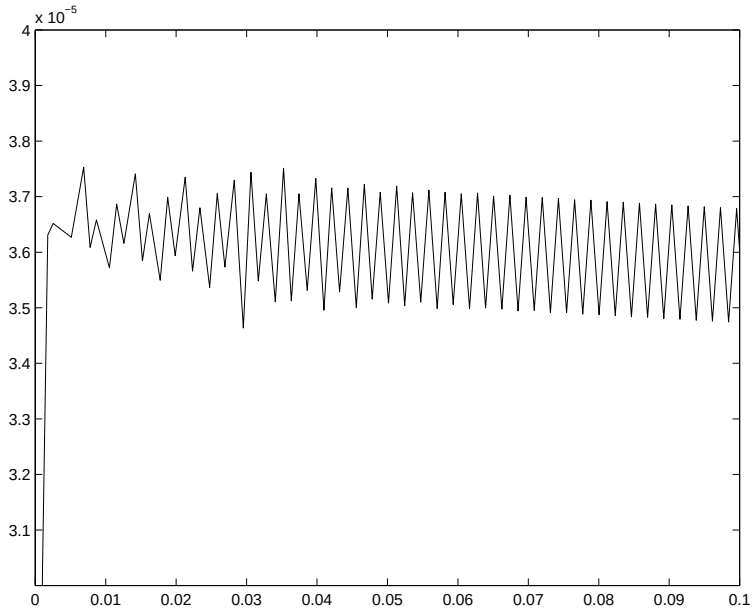

Abbildung 4.11: Lösungen von Beispiel 4.35. mit ode23

Zur Lösung dieses Systems im Intervall $[0, 10^6]$ (mit den voreingestellten Genauigkeitsschranken) benötigt das NDF Verfahren ode15s 146 Schritte. Mit dem impliziten Euler Verfahren benötigt man 310 Schritte. Qualitativ dasselbe Bild erhält man mit dem Rosenbrock Verfahren ode23s mit 61 Schritten.

Der Versuch, die Aufgabe mit dem nicht für steife Probleme geeigneten Verfahren ode23 zu lösen, benötigt in dem Intervall $[0, 1]$ schon 860 Schritte. Einen Ausschnitt des Graphen der zweiten Komponente der erzeugten Näherungslösung findet man in Abbildung 4.11. □

Es gibt viele (auch einander widersprechende) Definitionen der Steifheit. Häufig wird gesagt, dass ein Problem steif ist, wenn es verschieden schnell abklingende Lösungen besitzt, z.B. weil die Jacobi Matrix der rechten Seite Eigenwerte mit sehr unterschiedlichen (negativen) Realteilen besitzt.

Dies trifft jedoch nicht den Kern. Will man das schnelle Abklingen von Lösungskomponenten darstellen, so ist man gezwungen, die Lösungen mit sehr kleinen Schrittweiten zu approximieren, und hierzu kann man dann ein explizites Verfahren verwenden, denn diese sind in der Regel billiger als implizite Verfahren. Will man dagegen (wie in Beispiel 4.1.) eine sich langsam ändernde Lösung verfolgen, was eigentlich mit großen Schrittweiten möglich sein sollte, und wird ein explizites Verfahren durch sehr schnell abklingende Lösungsanteile zu sehr kleinen Schrittweiten gezwungen, so nennen wir ein Problem steif.

4.6 Abschließende Bemerkungen zur Wahl der Verfahren

Einer Anfangswertaufgabe sieht man nicht unmittelbar an, ob ihre Lösung steif ist. Es gibt einige Aufgabenklassen, bei denen man weiß, dass steife Lösungen zu erwarten sind wie z.B. Gleichungen, die chemische Reaktionen mit sehr unterschiedlichen Reaktionsgeschwindigkeiten beschreiben, Systeme, die sich mit der Linienmethode aus parabolischen oder hyperbolischen Anfangsrandwertaufgaben ergeben, oder singulär gestörte Probleme wie die van der Pol Gleichung mit sehr großem Parameter. In diesem Fall wird man sofort steife Löser verwenden.

Liegen keine guten Gründe dafür vor, dass eine steife Lösung zu erwarten ist, wird man zuerst versuchen, das gegebene Problem mit einem nicht-steifen Löser zu behandeln, denn explizite (eingebettete) Runge–Kutta Verfahren oder Mehrschrittverfahren vom Adams Typ sind wesentlich billiger als steife Löser. Bei steifen Lösern hat man ja in jedem Schritt ein (nichtlineares) Gleichungssystem zu lösen und hierzu die Jacobimatrix der rechten Seite oder eine Näherung davon in jedem Schritt zu bestimmen. Beobachtet man, dass der Lösungsprozess nur sehr langsam voranschreitet, wird man zu einem steifen Löser wechseln.

Runge–Kutta Verfahren ermöglichen eine einfache Schrittweitensteuerung, haben aber den Nachteil gegenüber den Adams Verfahren, dass in jedem Schritt die rechte Seite an mehreren Stellen ausgewertet werden muss (für das Verfahren von Dormand

und Prince der Ordnung 5 an 6 Stellen). Beim Prädiktor–Korrektor Verfahren kann man hohe Ordnungen mit 2 (im Fall PECE) oder 3 (im Fall PECECE) Auswertungen erreichen. Man wird daher ein Mehrschrittverfahren verwenden, wenn die Auswertung der rechten Seite der Differentialgleichung sehr teuer ist. In beiden Fällen wird man Verfahren hoher Ordnung nur dann verwenden, wenn die rechte Seite der Differentialgleichung sehr glatt ist. Man verwendet ja den Taylorschen Satz, um Methoden hoher Konsistenzordnung zu entwickeln.

Eine Regel für die Auswahl steifer Löser ist nicht so einfach zu formulieren. Einen Anhaltspunkt geben die Stabilitätsgebiete der Verfahren. BDF Formeln sind A(α)–stabil, wobei für große Ordnungen die Winkel α klein sind. Wenn man weiß, dass die Eigenwerte der Linearisierung der rechten Seite in der Nähe der negativen reellen Achse liegen (bei Linienmethoden für parabolische Aufgaben liegen sie sogar auf der negativen reellen Achse), so wird man BDF Formeln wählen. Weiß man, dass Eigenwerte der Jacobimatrix näher an der imaginären Achse als an der negativen reellen Achse liegen (bei Linienmethoden für hyperbolische Probleme liegen sie auf der imaginären Achse), so wird man Rosenbrockmethoden oder Extrapolationsverfahren verwenden.

Kapitel 5

Differentiell-algebraische Gleichungen vom Index 1

5.1 Einleitende Bemerkungen

Beispiel 5.1. Das ungedämpfte mathematische Pendel (der Länge 1 und der Masse 1) wird unter Verwendung des Auslenkwinkels φ bekanntlich beschrieben durch die gewöhnliche Differentialgleichung

$$\varphi'' = -g \sin \varphi.$$

Verwendet man Euklidische Koordinaten (y_1, y_2) der Endmasse, so liefert das Newtonsche Gesetz die Bewegungsgleichungen

$$\left.\begin{array}{rcl} y_1'' & = & -zy_1 \\ y_2'' & = & -zy_2 - g \end{array}\right\} \tag{5.1}$$

mit einem Lagrangeschen Parameter $z(t)$. Zusätzlich muss die Nebenbedingung

$$y_1^2 + y_2^2 = 1 \tag{5.2}$$

erfüllt sein. □

Bei der Simulation des dynamischen Verhaltens von Mehrkörpersystemen oder elektrischen Schaltkreisen treten häufig Systeme auf, die aus Differentialgleichungen und aus algebraischen Gleichungen bestehen. In diesem Fall spricht man von **differentiell-algebraischen Systemen** oder kurz **DAE** .

Man hat dann das Differentialgleichungssystem auf der Mannigfaltigkeit zu lösen, die durch die algebraischen Gleichungen gegeben it. Klar ist, dass dies nur möglich ist, wenn der Anfangspunkt auf der Mannigfaltigkeit liegt. Einen solchen Anfangsvektor nennt man **konsistent**.

Die Theorie der differentiell-algebraischen Gleichungen ist wesentlich schwieriger als die der gewöhnlichen Differentialgleichungen und ist Gegenstand aktiver Forschung. Einen Eindruck von der Theorie erhält man in den Büchern von *Ascher, Petzold* [4] oder *Hairer, Wanner* [30], eine ausführliche Darstellung in *Griepentrog, März* [23] oder *Hairer, Lubich, Roche* [28].

5.2 Der Index eines DAE Systems

Die allgemeinste Form eines (autonomen) DAE Systems ist

$$\boldsymbol{F}(\boldsymbol{u}, \boldsymbol{u}') = \boldsymbol{0}. \tag{5.3}$$

Dass wir ein von der unabhängigen Variable unabhängiges System betrachten, bedeutet wieder keine Einschränkung, denn wir können auch hier den allgemeinen Fall durch eine zusätzliche Gleichung $x' = 1$ auf den autonomen Fall zurückführen.

Ist

$$\frac{\partial}{\partial \boldsymbol{u}'}\boldsymbol{F}(\boldsymbol{u}, \boldsymbol{u}')$$

eine reguläre Matrix, so kann man nach dem Satz über implizite Funktionen die Gleichung $\boldsymbol{F}(\boldsymbol{u}, \boldsymbol{u}') = \boldsymbol{0}$ nach $\boldsymbol{u}'$ auflösen, d.h. es gibt eine Funktion $\boldsymbol{f}$ mit

$$\boldsymbol{F}(\boldsymbol{u}, \boldsymbol{u}') = \boldsymbol{0} \quad \Longleftrightarrow \quad \boldsymbol{u}' = \boldsymbol{f}(\boldsymbol{u}),$$

und das DAE System ist tatsächlich ein gewöhnliches Differentialgleichungssystem in impliziter Form.

Ist

$$\frac{\partial}{\partial \boldsymbol{u}'}\boldsymbol{F}(\boldsymbol{u}, \boldsymbol{u}')$$

singulär, so ist die Auflösung nach $\boldsymbol{u}'$ nicht (notwendig) möglich. Häufig kann man aber durch weiteres Differenzieren der DAE nach der unabhängigen Variable ein gewöhnliches Differentialgleichungssystem für $\boldsymbol{u}$ aufstellen.

Definition 5.2. *Das differentiell-algebraische System*

$$\boldsymbol{F}(\boldsymbol{u}, \boldsymbol{u}') = \boldsymbol{0}$$

besitzt den **Index** m*, wenn* $m \in \mathbb{N}$ *die minimale Zahl von Differentiationen ist, so dass das System*

$$\boldsymbol{F}(\boldsymbol{u}, \boldsymbol{u}') = \boldsymbol{0},\ \frac{d}{dx}\boldsymbol{F}(\boldsymbol{u}, \boldsymbol{u}') = \boldsymbol{0}, \ldots, \frac{d^m}{dx^m}\boldsymbol{F}(\boldsymbol{u}, \boldsymbol{u}') = \boldsymbol{0}$$

aufgelöst werden kann in ein Differentialgleichungssystem

$$\boldsymbol{u}' = \boldsymbol{\Phi}(\boldsymbol{u}).$$

Beispiel 5.3. *Ein gewöhnliches Differentialgleichungssystem*

$$\boldsymbol{y}' = \boldsymbol{f}(\boldsymbol{y})$$

hat als DAE den Index 0*.*

Beispiel 5.4. *Für das* **semi-explizite System differentiell-algebraischer Gleichungen**

$$\boldsymbol{y}' = \boldsymbol{f}(\boldsymbol{y}, \boldsymbol{z})$$

$$\boldsymbol{0} = \boldsymbol{g}(\boldsymbol{y}, \boldsymbol{z})$$

setzen wir voraus, dass

$$\frac{\partial}{\partial \boldsymbol{z}}\boldsymbol{g}(\boldsymbol{y}, \boldsymbol{z})$$

regulär ist. Wir differenzieren die zweite Gleichung:

$$\frac{\partial}{\partial \boldsymbol{y}}\boldsymbol{g}(\boldsymbol{y}, \boldsymbol{z})\boldsymbol{y}' + \frac{\partial}{\partial \boldsymbol{z}}\boldsymbol{g}(\boldsymbol{y}, \boldsymbol{z})\boldsymbol{z}' = \boldsymbol{0},$$

und erhalten als Differentialgleichung für $\boldsymbol{z}$

$$\boldsymbol{z}' = -\frac{\partial}{\partial \boldsymbol{z}}\boldsymbol{g}(\boldsymbol{y}, \boldsymbol{z})^{-1}\frac{\partial}{\partial \boldsymbol{y}}\boldsymbol{g}(\boldsymbol{y}, \boldsymbol{z})\boldsymbol{f}(\boldsymbol{y}, \boldsymbol{z}).$$

Das semi-explizite System hat also den Index 1.

Beispiel 5.5. *Wir betrachten das semi-explizite System*

$$\boldsymbol{y}' = \boldsymbol{f}(\boldsymbol{y}, \boldsymbol{z})$$

$$\boldsymbol{0} = \boldsymbol{g}(\boldsymbol{y}).$$

Die erste Differentiation der zweiten Gleichung liefert

$$\mathbf{0} = \frac{\partial}{\partial \boldsymbol{y}} \boldsymbol{g}(\boldsymbol{y})\boldsymbol{y}' = \frac{\partial}{\partial \boldsymbol{y}} \boldsymbol{g}(\boldsymbol{y})\boldsymbol{f}(\boldsymbol{y}, \boldsymbol{z}).$$

Die Lösung des DAE Systems liegt also nicht nur auf der Mannigfaltigkeit, die durch $\mathbf{0} = \boldsymbol{g}(\boldsymbol{y})$ *definiert ist, sondern (versteckt) auch auf der durch*

$$\mathbf{0} = \frac{\partial}{\partial \boldsymbol{y}} \boldsymbol{g}(\boldsymbol{y})\boldsymbol{f}(\boldsymbol{y}, \boldsymbol{z})$$

definierten.

Erneutes Differenzieren liefert

$$\frac{\partial^2 \boldsymbol{g}}{\partial \boldsymbol{y}^2}(\boldsymbol{f}, \boldsymbol{f}) + \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{y}} \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{y}} \boldsymbol{f} + \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{y}} \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{z}} \boldsymbol{z}' = \mathbf{0}.$$

Ist die Matrix $\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{y}} \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{z}}$ *regulär, so erhält man hieraus*

$$\boldsymbol{z}' = -\Big(\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{y}} \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{z}}\Big)^{-1} \Big(\frac{\partial^2 \boldsymbol{g}}{\partial \boldsymbol{y}^2}(\boldsymbol{f}, \boldsymbol{f}) + \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{y}} \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{y}} \boldsymbol{f}\Big),$$

und das System besitzt den Index 2.

Beispiel 5.6. *Das Pendelproblem in kartesischen Koordinaten besitzt den Index 3, denn als System 1. Ordnung hat es die Gestalt*

$$\left.\begin{aligned} y_1' &= y_2 \\ y_2' &= -y_5 y_1 \\ y_3' &= y_4 \\ y_4' &= -y_5 y_3 - g \\ 0 &= y_1^2 + y_3^2 - 1 \end{aligned}\right\} \tag{5.4}$$

Differenzieren der letzten Gleichung liefert

$$0 = 2y_1 y_1' + 2y_3 y_3',$$

d.h. unter Benutzung der 1. und 3. Gleichung des Systems (5.4)

$$0 = y_1 y_2 + y_3 y_4. \tag{5.5}$$

Durch erneutes Differenzieren erhält man

$$0 = y_1' y_2 + y_1 y_2' + y_3' y_4 + y_3 y_4' = y_2^2 - y_5 y_1^2 + y_4^2 - y_5 y_3^2 - g y_3 = y_2^2 + y_4^2 - g y_3 - y_5,$$

und hieraus folgt mit der dritten Differentiation

$$0 = 2y_2 y_2' + 2y_4 y_4' - g y_3' - y_5',$$

d.h. unter Benutzung von (5.4) und (5.5)

$$0 = -y_5' - 3g y_4.$$

5.3 Eine Einbettungsmethode

Wir betrachten der Einfachheit halber ein **semi-explizites System von differentiell-algebraischen Gleichungen**

$$\begin{aligned} \boldsymbol{y}' &= \boldsymbol{f}(\boldsymbol{y}, \boldsymbol{z}) && (5.6)\\ \boldsymbol{0} &= \boldsymbol{g}(\boldsymbol{y}, \boldsymbol{z}). && (5.7) \end{aligned}$$

Wir setzen voraus, dass die Jacobi Matrix

$$\frac{\partial}{\partial \boldsymbol{z}} \boldsymbol{g}(\boldsymbol{y}, \boldsymbol{z}) \tag{5.8}$$

in einer Umgebung der Lösung von (5.6), (5.7) regulär sei, so dass das System den Index 1 besitzt. Dann kann man nach dem Satz über implizite Funktionen Gleichung (5.7) lokal nach $\boldsymbol{z}$ auflösen, $\boldsymbol{z} = \boldsymbol{G}(\boldsymbol{y})$, und die DAE (5.6), (5.7) geht über in die gewöhnliche Differentialgleichung

$$\boldsymbol{y}' = \boldsymbol{f}(\boldsymbol{y}, \boldsymbol{G}(\boldsymbol{y})). \tag{5.9}$$

(5.9) heißt die zu (5.6), (5.7) gehörige **Zustandsraumgleichung**.

Wenn die Funktion $\boldsymbol{G}$ bekannt ist, liegt es natürlich nahe, die Differentialgleichung (5.9) numerisch mit einem der diskutierten Verfahren zu behandeln. I.a. ist aber nur die Existenz der Funktion $\boldsymbol{G}$ durch den Satz über implizite Funktionen gesichert, die explizite Auflösung von (5.7) aber nicht möglich. Man benötigt daher Verfahren zur Lösung von DAEs.

Eine Möglichkeit ist, anstatt (5.6), (5.7) die singuläre Störung

$$\begin{aligned} \boldsymbol{y}' &= \boldsymbol{f}(\boldsymbol{y}, \boldsymbol{z}) && (5.10)\\ \varepsilon \boldsymbol{z}' &= \boldsymbol{g}(\boldsymbol{y}, \boldsymbol{z}). && (5.11) \end{aligned}$$

für $\varepsilon > 0$ mit einem der bekannten Verfahren zu behandeln und in dem resultierenden Verfahren $\varepsilon = 0$ zu setzen.

Wir demonstrieren das Vorgehen für das (implizite) Runge–Kutta Verfahren. Wendet man dieses auf (5.10), (5.11) an, so erhält man im n-ten Schritt bei bekannten Näherungen $\boldsymbol{y}^n \approx \boldsymbol{y}(x_n)$ und $\boldsymbol{z}^n \approx \boldsymbol{z}(x_n)$

$$\boldsymbol{Y}_i = \boldsymbol{y}^n + h \sum_{j=1}^{s} \beta_{ij} \boldsymbol{f}(\boldsymbol{Y}_j, \boldsymbol{Z}_j) \tag{5.12}$$

$$\varepsilon \boldsymbol{Z}_i = \varepsilon \boldsymbol{z}^n + h \sum_{j=1}^{s} \beta_{ij} \boldsymbol{g}(\boldsymbol{Y}_j, \boldsymbol{Z}_j) \tag{5.13}$$

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}^n + h \sum_{i=1}^{s} \gamma_i \boldsymbol{f}(\boldsymbol{Y}_i, \boldsymbol{Z}_i) \tag{5.14}$$

$$\varepsilon \boldsymbol{z}^{n+1} = \varepsilon \boldsymbol{z}^n + h \sum_{i=1}^{s} \gamma_i \boldsymbol{g}(\boldsymbol{Y}_i, \boldsymbol{Z}_i). \tag{5.15}$$

Ist die Matrix $\boldsymbol{B} := (\beta_{ij})$ der Gewichte im Runge-Kutta Verfahren regulär, so erhält man aus (5.13)

$$h\boldsymbol{g}(\boldsymbol{Y}_i, \boldsymbol{Z}_i) = \varepsilon \sum_{j=1}^{s} \omega_{ij} (\boldsymbol{Z}_j - \boldsymbol{z}^n), \tag{5.16}$$

wobei $\boldsymbol{B}^{-1} =: (\omega_{ij})$ bezeichnet. Setzt man diesen Ausdruck in (5.15) ein, so kann man diese Gleichung durch ε kürzen. Setzt man schließlich noch $\varepsilon = 0$ in (5.13), so erhält man das Verfahren

$$\boldsymbol{Y}_i = \boldsymbol{y}^n + h \sum_{j=1}^{s} \beta_{ij} \boldsymbol{f}(\boldsymbol{Y}_j, \boldsymbol{Z}_j) \tag{5.17}$$

$$\boldsymbol{0} = \sum_{j=1}^{s} \beta_{ij} \boldsymbol{g}(\boldsymbol{Y}_j, \boldsymbol{Z}_j) \tag{5.18}$$

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}^n + h \sum_{i=1}^{s} \gamma_i \boldsymbol{f}(\boldsymbol{Y}_i, \boldsymbol{Z}_i) \tag{5.19}$$

$$\boldsymbol{z}^{n+1} = \Big(1 - \sum_{i,j=1}^{s} \gamma_i \omega_{ij}\Big) \boldsymbol{z}^n + \sum_{i,j=1}^{s} \gamma_i \omega_{ij} \boldsymbol{Z}_j. \tag{5.20}$$

Die numerische Lösung $(\boldsymbol{y}^{n+1}, \boldsymbol{z}^{n+1})$ wird i.a. nicht auf der Mannigfaltigkeit liegen, die durch $\boldsymbol{g}(\boldsymbol{y}, \boldsymbol{z}) = \boldsymbol{0}$ definiert ist. Dies kann man jedoch erreichen, wenn man in dem System (5.17), (5.18), (5.19), (5.20) die Gleichung (5.20) ersetzt durch

$$\boldsymbol{g}(\boldsymbol{y}^{n+1}, \boldsymbol{z}^{n+1}) = \boldsymbol{0}. \tag{5.21}$$

In diesem Fall gilt wegen (5.18) $\boldsymbol{Z}_j = \boldsymbol{G}(\boldsymbol{Y}_j)$ und wegen (5.21) auch $\boldsymbol{z}^{n+1} = \boldsymbol{G}(\boldsymbol{y}^{n+1})$. Das Verfahren (5.17), (5.18), (5.19), (5.21) ist daher äquivalent dem zu Grunde liegenden Runge-Kutta Verfahren für die Zustandsraumgleichung (5.9). Es heißt daher **Zustandsraumverfahren**.

Vorteil der Zustandsraummethode ist es, dass keine Konvergenztheorie nötig ist. Die Ergebnisse für gewöhnliche Anfangswertaufgaben übertragen sich sofort auf DAEs. Man kann ferner das Verfahren sofort mit einem expliziten Runge-Kutta Verfahren als Basis verwenden. Andererseit geben theoretische Ergebnisse über die Einbettungsmethode Einsichten über singulär gestörte Probleme.

5.4 Probleme mit Massenmatrizen

Viele differentiell-algebraische Probleme haben die Gestalt

$$\boldsymbol{M}\boldsymbol{u}' = \boldsymbol{F}(\boldsymbol{u}) \tag{5.22}$$

mit einer konstanten Matrix $\boldsymbol{M}$. Das semi-explizite Problem (5.6), (5.7) ist ein Spezialfall hiervon mit

$$\boldsymbol{M} = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} \end{pmatrix}, \ \boldsymbol{u} = \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{z} \end{pmatrix}.$$

Ist $\boldsymbol{M}$ regulär, so können wir jedes Verfahren der vorhergehenden Abschnitte auf das System

$$\boldsymbol{u}' = \boldsymbol{M}^{-1}\boldsymbol{F}(\boldsymbol{u})$$

anwenden und die entstehenden Formeln mit $\boldsymbol{M}$ multiplizieren, um zu einem Verfahren für die DAE (5.22) zu gelangen. Für ein implizites Runge-Kutta Verfahren erhält man auf diese Weise

$$\boldsymbol{M}(\boldsymbol{U}_i - \boldsymbol{u}^n) = h\sum_{j=1}^{s} \beta_{ij}\boldsymbol{F}(\boldsymbol{U}_j) \tag{5.23}$$

$$\boldsymbol{M}(\boldsymbol{u}^{n+1} - \boldsymbol{u}^n) = h\sum_{i=1}^{s} \gamma_i\boldsymbol{F}(\boldsymbol{U}_i). \tag{5.24}$$

Aus (5.23) erhält man wie im letzten Abschnitt mit der Matrix $\boldsymbol{B}^{-1} = (\omega_{ij})$

$$h\boldsymbol{F}(\boldsymbol{U}_i) = \sum_{j=1}^{s} \omega_{ij}\boldsymbol{M}(\boldsymbol{U}^j - \boldsymbol{u}^n),$$

und das Runge-Kutta Verfahren erhält die Gestalt

$$\boldsymbol{M}(\boldsymbol{U}_i - \boldsymbol{u}^n) = h\sum_{j=1}^{s} \beta_{ij}\boldsymbol{F}(\boldsymbol{U}_j) \tag{5.25}$$

$$\boldsymbol{u}^{n+1} = \Big(1 - \sum_{i,j=1}^{s} \gamma_i\omega_{ij}\Big)\boldsymbol{u}^n + \sum_{i,j=1}^{s} \gamma_i\omega_{ij}\boldsymbol{U}_j. \tag{5.26}$$

Dieses Verfahren ergibt auch einen Sinn, wenn die Matrix $\boldsymbol{M}$ singulär ist. In diesem Fall ist das System (5.22) äquivalent einem semi-expliziten System (5.6), (5.7), und das Verfahren (5.25), (5.26) enspricht der Einbettungsmethode (5.17) – (5.20). Dies sieht man so ein: Mit dem Gaußschen Eliminationsverfahren mit totaler Pivotsuche kann man reguläre Matrizen $\boldsymbol{S}$ und $\boldsymbol{T}$ bestimmen mit

$$\boldsymbol{M} = \boldsymbol{S}\begin{pmatrix} \boldsymbol{I}_r & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} \end{pmatrix}\boldsymbol{T}, \tag{5.27}$$

wobei r den Rang von $\boldsymbol{M}$ bezeichnet.

Setzt man (5.27) in (5.22) ein, multipliziert man mit $\boldsymbol{S}^{-1}$ und verwendet man die transformierten Variablen

$$\boldsymbol{T}\boldsymbol{u} = \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{z} \end{pmatrix},$$

so erhält man

$$\begin{pmatrix} \boldsymbol{y}' \\ \boldsymbol{0} \end{pmatrix} = \boldsymbol{S}^{-1}\boldsymbol{F}\Big(\boldsymbol{T}^{-1}\begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{z} \end{pmatrix}\Big) =: \begin{pmatrix} \boldsymbol{f}(\boldsymbol{y},\boldsymbol{z}) \\ \boldsymbol{g}(\boldsymbol{y},\boldsymbol{z}) \end{pmatrix}, \tag{5.28}$$

ein semi-explizites Problem (5.6), (5.7). Ein Anfangswert $\boldsymbol{u}^0$ ist konsistent, wenn $\boldsymbol{F}(\boldsymbol{u}^0)$ im Wertebereich der Matrix $\boldsymbol{M}$ liegt.

Setzt man (5.27) in (5.25), (5.26) ein und verwendet man die transformierten Variablen

$$\boldsymbol{T}\boldsymbol{U}_j =: \begin{pmatrix} \boldsymbol{Y}_j \\ \boldsymbol{Z}_j \end{pmatrix}, \ \boldsymbol{T}\boldsymbol{u}^n =: \begin{pmatrix} \boldsymbol{y}^n \\ \boldsymbol{z}^n \end{pmatrix},$$

so geht das Verfahren (5.25), (5.26) über in die Formeln (5.17) – (5.20). Damit gelten alle Resultate für das semi-explizite Problem (5.6), (5.7) und die Einbettungsmethode auch für Probleme des Typs (5.22).

5.5 Mehrschrittverfahren

Auch Mehrschrittverfahren lassen sich leicht auf semi-explizite DAEs übertragen. Wendet man ein Mehrschrittverfahren auf die Einbettung (5.10), (5.11) einer semi-expliziten DAE an, so liefert dies

$$\sum_{i=0}^{k} a_i \boldsymbol{y}^{n+i} = h\sum_{i=0}^{k} b_i \boldsymbol{f}(\boldsymbol{y}^{n+i}, \boldsymbol{z}^{n+i}) \tag{5.29}$$

$$\varepsilon\sum_{i=0}^{k} a_i \boldsymbol{z}^{n+i} = h\sum_{i=0}^{k} b_i \boldsymbol{g}(\boldsymbol{y}^{n+i}, \boldsymbol{z}^{n+i}). \tag{5.30}$$

Setzt man hier wieder $\varepsilon = 0$, so erhält man das eingebettete Verfahren

$$\sum_{i=0}^{k} a_i \boldsymbol{y}^{n+i} = h\sum_{i=0}^{k} b_i \boldsymbol{f}(\boldsymbol{y}^{n+1}, \boldsymbol{z}^{n+i}) \tag{5.31}$$

$$0 = \sum_{i=0}^{k} b_i \boldsymbol{g}(\boldsymbol{y}^{n+1}, \boldsymbol{z}^{n+i}). \tag{5.32}$$

Dieser Zugang wurde erstmals von Gear für BDF Methoden für differentiell-algebraische Systeme vorgeschlagen und in einem ersten (sehr erfolgreichen) Code für DAEs umgesetzt.

Wie vorher kann man (5.32) ersetzen durch

$$\boldsymbol{g}(\boldsymbol{y}^{n+k}, \boldsymbol{z}^{n+k}) = \boldsymbol{0}$$

und erhält die zugehörige Zustandsraummethode. Ferner kann man das Verfahren sofort auf das Problem (5.22) übertragen und erhält

$$\boldsymbol{M}\sum_{i=0}^{k} a_i \boldsymbol{u}^{n+i} = h\sum_{i=0}^{k} b_i \boldsymbol{F}(\boldsymbol{u}^{n+i}). \tag{5.33}$$

Für Mehrschrittverfahren gilt das folgende Konvergenzresultat:

Satz 5.7. *Das semi-explizite Problem (5.6), (5.7) erfülle die Voraussetzung (5.8). Das Mehrschrittverfahren, das (5.31), (5.32) zu Grunde liegt, besitze die Ordnung p, und es mögen die Punkte 0 und ∞ im Stabilitätsgebiet liegen. Ferner habe der Fehler des Anfangsfeldes $\boldsymbol{y}^j,\ \boldsymbol{z}^j,\ j = 0,\dots,k-1$, die Größe $O(h^p)$. Dann erfüllt der globale Fehler*

$$\boldsymbol{y}^n - \boldsymbol{y}(x_n) = O(h^p),\ \boldsymbol{z}^n - \boldsymbol{z}(x_n) = O(h^p),$$

für $x_n - x_0 = nh \le const.$

Beweis: s. *Hairer und Wanner* [30], p. 383

Beispiel 5.8. Wir betrachten erneut Beispiel 4.35..

$$\begin{aligned} y_1' &= -0.04y_1 + 10^4 y_2 y_3, & y_1(0) &= 1\\ y_2' &= 0.04y_1 - 10^4 y_2 y_3 - 3\cdot 10^7 y_2^2, & y_2(0) &= 0\\ y_3' &= 3\cdot 10^7 y_2^2, & y_3(0) &= 0. \end{aligned}$$

Summiert man die drei Gleichungen, so folgt

$$y_1' + y_2' + y_3' = 0,\ \text{d.h. } y_1 + y_2 + y_3 = \text{const},$$

und aus den Anfangsbedingungen folgt $y_1 + y_2 + y_3 = 1$.

Damit ist die Anfangswertaufgabe äquivalent dem differentiell-algebraischen Problem

$$\begin{aligned} y_1' &= -0.04y_1 + 10^4 y_2 z_1\\ y_2' &= 0.01y_1 + 10^4 y_2 z_1 - 3\cdot 10^7 y_2^2\ ,\\ 0 &= y_1 + y_2 + z_1 - 1 \end{aligned} \qquad \boldsymbol{y}(0) = \begin{pmatrix}1\\0\\0\end{pmatrix}. \tag{5.34}$$

Natürlich kann man hier die letzte Gleichung nach z_1 auflösen und erhält die Zustandsraumgleichung

$$\begin{aligned} y_1' &= -0.04y_1 + 10^4 y_2(1 - y_1 - y_2)\\ y_2' &= 0.04y_1 + 10^4 y_2(1 - y_1 - y_2) - 3\cdot 10^7 y_2^2 \end{aligned}\ , \qquad \boldsymbol{y}(0) = \begin{pmatrix}1\\0\end{pmatrix}. \tag{5.35}$$

Mit `ode15s` der ODE-Suite von MATLAB erhält man in allen drei Fällen die bereits bekannte Lösung. Dabei werden (mit `AbsTol=`10^{-12}) für das Originalproblem 83345 flops, für (5.34) 89355 und für die Zustandsraumgleichung (5.35) 59250 flops benötigt. □

Kapitel 6

Numerische Verfahren für Randwertaufgaben

6.1 Anfangswertmethoden

Wir haben in den vorhergehenden Abschnitten sehr leistungsfähige numerische Methoden zur Lösung von Anfangswertaufgaben zusammengetragen. Wir wollen diese Verfahren nun nutzen, um Randwertaufgaben numerisch zu lösen.

Wir beginnen mit der linearen Randwertaufgabe zweiter Ordnung

$$y'' + p(x)y' + q(x)y = f(x),\ a \le x \le b, \quad y(a) = \alpha,\ y(b) = \beta, \tag{6.1}$$

mit gegebenen stetigen Funktionen $p, q, f \in C[a, b]$ und gegebenen $\alpha, \beta \in \mathbb{R}$.

Wir schätzen die Anfangssteigung $s_1 \in \mathbb{R}$ der unbekannten Lösung der Randwertaufgabe und lösen die Anfangswertaufgabe

$$y'' + p(x)y' + q(x)y = f(x),\ a \le x \le b, \quad y(a) = \alpha,\ y'(a) = s_1. \tag{6.2}$$

Dann wird die Lösung $y(x; s_1)$ in der Regel nicht die zweite Randbedingung $y(b) = \beta$ erfüllen.

Mit einer zweiten geschätzten Anfangssteigung $s_2 \in \mathbb{R}$, $s_2 \neq s_1$, können wir die Lösung $y(x; s_2)$ der Anfangswertaufgabe

$$y'' + p(x)y' + q(x)y = f(x),\ a \le x \le b, \quad y(a) = \alpha,\ y'(a) = s_2. \tag{6.3}$$

bestimmen. Dann ist wegen der Linearität der Differentialgleichung

$$y_t(x) := ty(x; s_1) + (1 - t)y(x; s_2) \tag{6.4}$$

für jedes $t \in \mathbb{R}$ eine Lösung der Differentialgleichung in (6.1), und es ist auch die Anfangsbedingung

$$ty(a; s_1) + (1 - t)y(a; s_2) = \alpha$$

erfüllt.

Den freien Parameter t können wir nutzen, um auch die zweite Randbedingung zu erfüllen. Ist $y(b; s_1) = y(b; s_2)$, so löst

$$y(x) := y(x; s_1) - y(x; s_2) \not\equiv 0$$

die homogene Randwertaufgabe

$$y'' + p(x)y' + q(x)y = 0,\ a \le x \le b, \quad y(a) = 0,\ y(b) = 0,$$

und daher ist die vorgelegte Aufgabe nicht eindeutig lösbar. Sonst ist (6.1) eindeutig lösbar, und mit

$$\hat{t} = \frac{\beta - y(b; s_2)}{y(b; s_1) - y(b; s_2)}, \tag{6.5}$$

ist $y_{\hat{t}}(x)$ aus (6.4) die Lösung.

Die Anfangswertaufgaben (6.2) und (6.3) wird man in der Regel nicht exakt lösen können. Man kann jedoch mit den Methoden der vorhergehenden Abschnitte diese Aufgaben effizient numerisch lösen und mit den gewonnenen Approximationen den Parameter t aus (6.5) ermitteln. Insbesondere erhält man damit den (approximativen) Anfangswert

$$\hat{s} := \hat{t}s_1 + (1 - \hat{t})s_2$$

der Lösung der Randwertaufgabe (6.1). Bestimmt man nun noch die Lösung der Anfangswertaufgabe

$$y'' + p(x)y' + q(x)y = f(x),\ a \le x \le b, \quad y(a) = \alpha,\ y'(a) = \hat{s}, \tag{6.6}$$

so ist diese eine Approximation der Lösung von (6.1). Man hätte natürlich auch die Lösung gemäß

$$y_{\hat{t}}(x) := \hat{t}y(x; s_1) + (1 - \hat{t})y(x; s_2)$$

bestimmen können. Man benötigt dann aber die Lösungen der Anfangswertaufgaben (6.2) und (6.3) an denselben Knoten, so dass man diese Aufgaben nicht mit einer

Methode mit Schrittweitensteuerung lösen kann. Das vorgestellte Verfahren heißt **Schießverfahren** oder genauer **Einfach–Schießverfahren**.

Ähnlich können wir bei der allgemeineren linearen Randwertaufgabe

$$\boldsymbol{y}' = \boldsymbol{C}(x)\boldsymbol{y} + \boldsymbol{f} \tag{6.7}$$

$$\boldsymbol{A}\boldsymbol{y}(a) + \boldsymbol{B}\boldsymbol{y}(b) = \boldsymbol{c} \tag{6.8}$$

mit stetigen Funktionen $\boldsymbol{C} : [a,b] \to \mathbb{R}^{(n,n)}$ und $\boldsymbol{f} : [a,b] \to \mathbb{R}^n$ und gegeben $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{(n,n)}$ und $\boldsymbol{c} \in \mathbb{R}^n$ vorgehen.

Wir wissen bereits, dass wir die Lösung des Systems (6.7) schreiben können als

$$\boldsymbol{y}(x) = \boldsymbol{Y}(x)\boldsymbol{\alpha} + \tilde{\boldsymbol{y}}(x), \quad \boldsymbol{\alpha} \in \mathbb{R}^n \tag{6.9}$$

wobei $\boldsymbol{Y}(x)$ eine Fundamentalmatrix des homogenen Systems

$$\boldsymbol{y}' = \boldsymbol{C}(x)\boldsymbol{y} \tag{6.10}$$

und $\tilde{\boldsymbol{y}}(x)$ eine spezielle Lösung von (6.7) ist. Ferner ist die Randwertaufgabe (6.7), (6.8) genau dann eindeutig lösbar, wenn die Matrix

$$\boldsymbol{D} := \boldsymbol{A}\boldsymbol{Y}(a) + \boldsymbol{B}\boldsymbol{Y}(b) \tag{6.11}$$

regulär ist, und wir erhalten diese Lösung, wenn wir in (6.9)

$$\boldsymbol{\alpha} = (\boldsymbol{A}\boldsymbol{Y}(a) + \boldsymbol{B}\boldsymbol{Y}(b))^{-1}(\boldsymbol{c} - \boldsymbol{A}\tilde{\boldsymbol{y}}(a) - \boldsymbol{B}\tilde{\boldsymbol{y}}(b)) \tag{6.12}$$

setzen.

Mit diesem Vorgehen können wir die Lösung numerisch bestimmen: Wir wählen einen beliebigen Anfangsvektor $\boldsymbol{y}_0$ (z.B. $\boldsymbol{y}_0 = \boldsymbol{0}$) und bestimmen mit einem Verfahren zur Lösung von Anfangswertaufgaben die Lösung $\boldsymbol{y}_i$ der Anfangswertaufgabe

$$\boldsymbol{y}' = \boldsymbol{C}(x)\boldsymbol{y} + \boldsymbol{f}(x), \quad \boldsymbol{y}(a) = \boldsymbol{y}_0,$$

zum inhomogenen Problem, sowie die Lösungen $\boldsymbol{y}_{h,1},\ldots,\boldsymbol{y}_{h,n}$ der Anfangswertaufgaben

$$\boldsymbol{y}' = \boldsymbol{C}(x)\boldsymbol{y}, \quad \boldsymbol{y}(a) = \boldsymbol{e}^j,\ j = 1,\ldots,n,$$

zur homogenen Differentialgleichung. Dabei kann man natürlich $\boldsymbol{e}^1,\ldots,\boldsymbol{e}^n$ durch jede andere Basis des $\mathbb{R}^n$ ersetzen.

Es sei

$$\boldsymbol{Y}(x) := (\boldsymbol{y}_{h,1},\ldots,\boldsymbol{y}_{h,n})$$

die (numerische) Fundamentalmatrix des Systems (6.10), die durch $\boldsymbol{Y}(a) = \boldsymbol{E}$ normiert ist, und

$$\tilde{\boldsymbol{\alpha}} = (\boldsymbol{A} + \boldsymbol{B}\boldsymbol{Y}(b))^{-1}(\boldsymbol{c} - \boldsymbol{A}\boldsymbol{y}_i(a) - \boldsymbol{B}\boldsymbol{y}_i(b)).$$

Dann erhält man durch (numerische) Lösung der Anfangswertaufgabe

$$\boldsymbol{y}' = \boldsymbol{C}(x)\boldsymbol{y} + \boldsymbol{f}(x), \quad \boldsymbol{y}(a) = \boldsymbol{Y}(a)\tilde{\boldsymbol{\alpha}} + \boldsymbol{y}_0 \tag{6.13}$$

die Lösung der Randwertaufgabe (6.7), (6.8).

Natürlich kann man das Lösen der Anfangswertaufgabe (6.13) dadurch ersetzen, dass man die Lösung von (6.7), (6.8) bestimmt als

$$\boldsymbol{y}(x) = \sum_{j=1}^{n} \tilde{\alpha}_j \boldsymbol{y}_{h,j}(x) + \boldsymbol{y}_i(x).$$

Dies ist jedoch nur dann möglich, wenn alle Anfangswertaufgaben mit demselben Gitter gelöst worden sind. Damit ist die Verwendung eines Lösers mit Schrittweiten– oder Ordnungskontrolle ausgeschlossen. Ferner müssen die Lösungen $\boldsymbol{y}_{h,j}$ und $\boldsymbol{y}_i$ an allen Zwischenstellen gespeichert werden.

Bemerkung 6.1. Sind am linken Rand k Anfangsbedingungen vorgegeben, sind also die Randbedingungen gegeben durch

$$y_{j_i}(a) = c_i, \ i = 1, \ldots, k, \quad \tilde{\boldsymbol{A}}\boldsymbol{y}(a) + \tilde{\boldsymbol{B}}\boldsymbol{y}(b) = \tilde{\boldsymbol{c}} \tag{6.14}$$

mit $\tilde{\boldsymbol{A}}, \tilde{\boldsymbol{B}} \in \mathbb{R}^{(n-k,n)}$ und $\tilde{\boldsymbol{c}} \in \mathbb{R}^{n-k}$, so brauchen natürlich die gegeben Anfangsbedingungen nicht variiert zu werden. Nehmen wir ohne Einschränkung $j_i = i$, $i = 1, \ldots, k$ an und setzen wir $\eta := (y_1(a), \ldots, y_k(a), \tilde{\eta}_1, \ldots, \tilde{\eta}_{n-k})^T$, so haben wir das lineare Gleichungssystem

$$\tilde{\boldsymbol{A}}\eta + \tilde{\boldsymbol{B}}y(b;\eta) = \tilde{\boldsymbol{c}}$$

zu lösen, und daher nur $n - k + 1$ Anfangswertaufgaben zu behandeln. □

Wir übertragen die Vorgehensweise nun auf nichtlineare Randwertaufgaben. Wir betrachten das Differentialgleichungssystem

$$\boldsymbol{y}' = \boldsymbol{f}(x, \boldsymbol{y}), \ a \le x \le b, \tag{6.15}$$

mit den (nicht notwendig linearen) Zweipunkt–Randbedingungen

$$\boldsymbol{R}(\boldsymbol{y}(a), \boldsymbol{y}(b)) = \boldsymbol{0}. \tag{6.16}$$

Dabei seien

$$\boldsymbol{f} : \mathbb{R} \times \mathbb{R}^n \supset D_f \to \mathbb{R}^n, \quad \boldsymbol{R} : \mathbb{R}^{2n} \supset D_R \to \mathbb{R}^n$$

stetig, und es erfülle $\boldsymbol{f}$ eine Lipschitzbedingung bzgl. $\boldsymbol{y}$. Die Mengen D_f und D_R seien offen.

Nach dem Satz von Picard und Lindelöf besitzt die Anfangswertaufgabe

$$\boldsymbol{y}' = \boldsymbol{f}(x, \boldsymbol{y}), \quad \boldsymbol{y}(a) = \boldsymbol{\eta}, \tag{6.17}$$

für alle $\boldsymbol{\eta}$ mit $(a, \boldsymbol{\eta}) \in D$ eine eindeutige Lösung

$$\boldsymbol{y}(\cdot, \boldsymbol{\eta}) : [a, a + d(\boldsymbol{\eta})) \to \mathbb{R}^n, \tag{6.18}$$

wobei $d(\boldsymbol{\eta})$ die Länge des Existenzintervalls von $\boldsymbol{y}(\cdot, \boldsymbol{\eta})$ bezeichnet.

Die Randwertaufgabe kann man nun auf folgende Weise auf ein Gleichungssystem zurückführen. Besitzt (6.15), (6.16) eine Lösung $\hat{\boldsymbol{y}}$, so ist mit $\hat{\boldsymbol{\eta}} := \hat{\boldsymbol{y}}(a)$

$$\hat{\boldsymbol{y}} = \boldsymbol{y}(\cdot, \hat{\boldsymbol{\eta}}), \ d(\hat{\boldsymbol{\eta}}) > b - a,$$

und $\hat{\boldsymbol{\eta}}$ ist eine Nullstelle der Funktion

$$\boldsymbol{F} : \mathbb{R}^n \supset D_F \to \mathbb{R}^n, \ \boldsymbol{F}(\boldsymbol{\eta}) := R(\boldsymbol{\eta}, \boldsymbol{y}(b, \boldsymbol{\eta})). \tag{6.19}$$

Ist umgekehrt $d(\tilde{\boldsymbol{\eta}}) > b - a$ für eine Nullstelle $\tilde{\boldsymbol{\eta}}$ von $\boldsymbol{F}$, so ist $\boldsymbol{y}(\cdot, \tilde{\boldsymbol{\eta}})$ eine Lösung der Randwertaufgabe (6.15), (6.16). Damit ist gezeigt:

Satz 6.2. *Es sei*

$$D_F := \{\boldsymbol{\eta} \in \mathbb{R}^n : d(\boldsymbol{\eta}) > b - a, \ (\boldsymbol{\eta}, \boldsymbol{y}(b, \boldsymbol{\eta})) \in D_R\}.$$

dann gilt

$$\boldsymbol{\eta} \in D_F, \ \boldsymbol{F}(\boldsymbol{\eta}) = \boldsymbol{0} \qquad \Longleftrightarrow \qquad \boldsymbol{y}(\cdot, \boldsymbol{\eta}) \textit{ löst } (6.15), (6.16).$$

Damit ist die Randwertaufgabe (6.15), (6.16) äquivalent dem Nullstellenproblem (6.19) im $\mathbb{R}^n$. Eine Funktionsauswertung von $\boldsymbol{F}$ erfordert dabei das Lösen einer Anfangswertaufgabe (6.17). Dies wird mit einem der in den vorhergehenden Abschnitten besprochenen Verfahren getan.

Zur Lösung von $\boldsymbol{F}(\boldsymbol{\eta}) = \boldsymbol{0}$ verwendet man ein numerisches Verfahren. Das Newton Verfahren lautet z.B.

$$\boldsymbol{\eta}^{k+1} = \boldsymbol{\eta}^k - \boldsymbol{F}'(\boldsymbol{\eta}^k)^{-1} \boldsymbol{F}(\boldsymbol{\eta}^k). \tag{6.20}$$

Ist $\boldsymbol{f}$ stetig partiell differenzierbar nach den Komponenten von $\boldsymbol{y}$, so ist nach Korollar 1.8. $\boldsymbol{y}(\cdot, \boldsymbol{\eta})$ differenzierbar nach den Anfangswerten η_j, und die Funktionen

$$\boldsymbol{z}^j := \frac{\partial}{\partial \eta_j} \boldsymbol{y}(\cdot, \boldsymbol{\eta}),\ j = 1, \ldots, n,$$

sind die eindeutigen Lösungen der Anfangswertaufgaben

$$\frac{d}{dx}\boldsymbol{z}^j = \frac{\partial}{\partial \boldsymbol{y}} \boldsymbol{f}(x, \boldsymbol{y}(x, \eta))\boldsymbol{z}^j,\ \boldsymbol{z}^j(a) = \boldsymbol{e}^j,\ j = 1, \ldots, n. \tag{6.21}$$

Mit

$$\boldsymbol{Z}(x) := (\boldsymbol{z}^1(x), \ldots, \boldsymbol{z}^n(x))$$

erhält man die Ableitung

$$\boldsymbol{F}'(\boldsymbol{\eta}) = \frac{\partial}{\partial \boldsymbol{y}(a)} R(\boldsymbol{\eta}, \boldsymbol{y}(b, \boldsymbol{\eta})) + \frac{\partial}{\partial \boldsymbol{y}(b)} R(\boldsymbol{\eta}, \boldsymbol{y}(b, \boldsymbol{\eta}))\boldsymbol{Z}(b). \tag{6.22}$$

Um einen Schritt des Newton Verfahrens für das Nullstellenproblem $\boldsymbol{F}(\boldsymbol{\eta}) = \boldsymbol{0}$ ausführen zu können, muss man also (simultan) die Anfangswertaufgaben (6.17) und (6.21) lösen.

Da die Ableitungen in (6.21) sehr kompliziert sein können, ersetzt man i.a. die Matrix $\boldsymbol{F}'(\boldsymbol{\eta})$ durch eine Matrix von Differenzenquotienten:

$$\frac{\partial}{\partial \eta_j} \boldsymbol{y}(b; \boldsymbol{\eta}) \approx \frac{1}{\Delta \eta_j} \Big(\boldsymbol{y}(b; \eta_1, \ldots, \eta_j + \Delta\eta_j, \ldots, \eta_n) - \boldsymbol{y}(b; \boldsymbol{\eta}) \Big). \tag{6.23}$$

Bei dieser Variante benötigt man in jedem Schritt die Lösung von $n + 1$ Anfangswertaufgaben des Typs (6.17).

Ein Verfahren, das in jedem Schritt nur die Lösung einer Anfangswertaufgabe benötigt, ist das **Broyden Verfahren**:

Gegeben seien $\boldsymbol{\eta}^0$ und $\boldsymbol{S}_0 \in \mathbb{R}^{(n,n)}$
For $k = 0, 1, 2, \ldots$ do
  Löse $\boldsymbol{S}_k \boldsymbol{s}^k = -\boldsymbol{F}(\boldsymbol{\eta}^k)$;
  $\boldsymbol{\eta}^{k+1} := \boldsymbol{\eta}^k + \boldsymbol{s}^k$;
  $\boldsymbol{u}^k := \boldsymbol{F}(\boldsymbol{\eta}^{k+1}) - \boldsymbol{F}(\boldsymbol{\eta}^k)$;
  $\boldsymbol{S}_{k+1} := \boldsymbol{S}_k + \dfrac{(\boldsymbol{u}^k - \boldsymbol{S}_k \boldsymbol{s}^k)\boldsymbol{s}^{kT}}{\|\boldsymbol{s}^k\|_2^2}$.

Das Broyden Verfahren und weitere sog. **Quasi–Newton Verfahren** werden in *Dennis, Schnabel* [15], p. 168 ff, motiviert und untersucht. Man kann zeigen, dass das Broyden Verfahren lokal und superlinear gegen isolierte Nullstellen konvergiert.

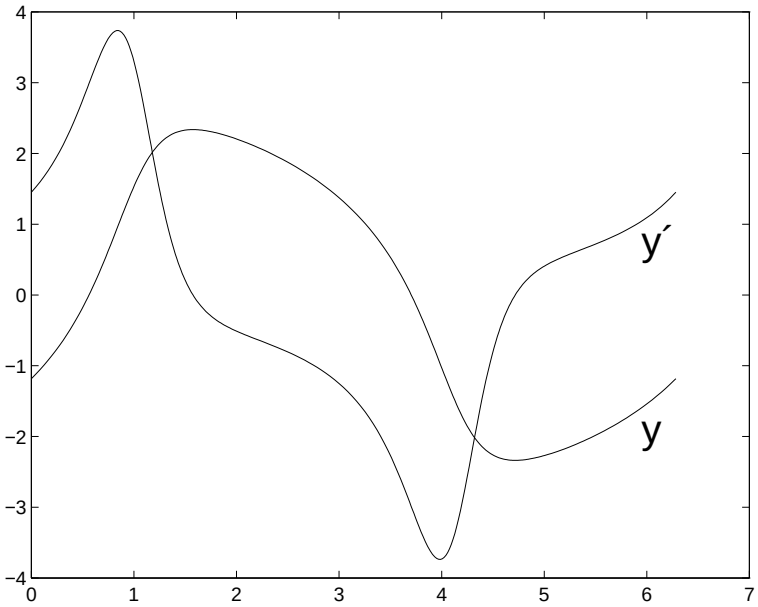

Abbildung 6.1: Lösung von Beispiel 6.3.

Beispiel 6.3. Wir betrachten die nicht–autonome van der Pol Gleichung

$$y'' - (1 - y^2)y' + y = \cos x \tag{6.24}$$

mit periodischen Randbedingungen

$$y(0) = y(2\pi),\ y'(0) = y'(2\pi). \tag{6.25}$$

Dieses Problem hat eine eindeutige Lösung, deren Graph in Abbildung 6.1 gegeben ist.

Mit dem Startvektor $(2, 0)$ erhält man mit dem approximativen Newton Verfahren, bei dem Näherungen für die Jacobi Matrix mit Differenzenquotienten berechnet werden (Schrittweite $\Delta_1 = \Delta_2 = 0.1$), und dem eingebetteten Verfahren von Dormand und Prince zur Lösung der Anfangswertaufgaben nach 4 Schritten eine Näherungslösung, für die $\|\boldsymbol{y}(0) - \boldsymbol{y}(2\pi)\| < 10^{-6}$ gilt.

Mit dem Broyden Verfahren mit dem unsinnigen Startwert $\boldsymbol{S}_0 = \boldsymbol{E}$ erhält man diese Genauigkeit nach 8 Schritten. Verwendet man als ersten einen approximativen Newton Schritt, um zugleich einen geeigneten Startwert für $\boldsymbol{A}_0$ zu erhalten, so benötigt man danach noch 5 Broyden Schritte.

Beachten Sie, dass für einen Newton Schritt in diesem Beispiel 3 Anfangswertaufgaben gelöst werden müssen, während das Broyden Verfahren nur das Lösen einer Anfangswertaufgabe pro Schritt erfordert. Das Broyden Verfahren führt also schneller zum Ziel als das approximative Newton Verfahren, wobei hier die Bestimmung des Startwerts $\boldsymbol{A}_0$ mit dem Newton Verfahren sich nicht auszahlt. □

Bei der Durchführung des Einfach–Schießverfahrens treten Schwierigkeiten auf, die wir nun an zwei Beispielen erläutern wollen.

Beispiel 6.4. Wir betrachten die Randwertaufgabe

$$y'' - 10y' = 0,\ y(0) = 0,\ y(10) = 1. \tag{6.26}$$

Dann ist offenbar

$$y(x; s) = \frac{s}{10}\Big(e^{10x} - 1\Big)$$

die Lösung der Anfangswertaufgabe

$$y'' - 10y' = 0,\ y(0) = 0,\ y'(0) = s,$$

und die Lösung der Randwertaufgabe (6.26) erhält man für

$$\bar{s} = \frac{10}{e^{100} - 1}.$$

Stört man die Anfangssteigung im Bereich der Maschinengenauigkeit $\varepsilon \approx 10^{-16}$, so erhält man

$$y(10; \bar{s} + \varepsilon) - y(10; \bar{s}) = \frac{\varepsilon}{10}\Big(e^{100} - 1\Big) \approx 2.69\varepsilon \cdot 10^{42} \approx 10^{26},$$

d.h. wegen $y(10; \bar{s}) = 1$

$$y(10; \bar{s} + \varepsilon) \approx 10^{26}.$$

Das Beispiel zeigt: Selbst wenn der Anfangswert $\bar{s}$ mit Maschinengenauigkeit bekannt ist, ist nicht gesichert, dass man $y(x; \bar{s})$ im betrachteten Intervall genau berechnen kann.

In diesem Beispiel gilt

$$|y(x; s_1) - y(x; s_2)| = O(e^{10x})|s_1 - s_2|.$$

Der Einfluss fehlerhafter Daten wächst also exponentiell mit x. Allgemeiner gilt (vgl. Satz 1.6.): Genügt die rechte Seite einer Anfangswertaufgabe

$$\boldsymbol{y}' = \boldsymbol{f}(x, \boldsymbol{y}),\ \boldsymbol{y}(a) = \boldsymbol{\eta}$$

einer Lipschitz Bedingung

$$\|\boldsymbol{f}(x, \boldsymbol{y}) - \boldsymbol{f}(x, \boldsymbol{z})\| \le L\|\boldsymbol{y} - \boldsymbol{z}\|,$$

so gilt für die Lösung $\boldsymbol{y}(x; \boldsymbol{\eta})$

$$\|\boldsymbol{y}(x; \boldsymbol{\eta}_1) - \boldsymbol{y}(x; \boldsymbol{\eta}_2)\| \le \|\boldsymbol{\eta}_1 - \boldsymbol{\eta}_2\| e^{L(x-a)}. \tag{6.27}$$

Diese Abschätzung zeigt, dass man durch Verkleinerung des Intervalls den Einfluss fehlerhafter Anfangsdaten klein halten kann. □

Beispiel 6.5. Das nächste Beispiel geht auf B.A. Troesch (1960) zurück.

$$y'' = 5\sinh(5y),\ y(0) = 0,\ y(1) = 1. \tag{6.28}$$

Man kann zeigen, dass diese Randwertaufgabe eine eindeutige Lösung y besitzt und dass $0 \le y(x) \le x$ für alle $x \in [0,1]$, und damit $0 \le y'(0) \le 1$ gilt.

Löst man die zugehörige Anfangswertaufgabe mit den Anfangssteigungen $y'(0) = 0.1 \cdot i$, $i = 1, \ldots, 10$, so erhält man in allen Fällen Exponentenüberlauf. Der Grund ist, dass für alle diese Werte für $y'(0)$ die Lösung eine singuläre Stelle besitzt, die im Intervall $(0,1)$ liegt.

Man kann für dieses Beispiel die Anfangswertaufgabe

$$y'' = 5\sinh(5y),\ y(0) = 0,\ y'(0) = s \tag{6.29}$$

mit Hilfe von elliptischen Funktionen exakt lösen. Die Lösung der Randwertaufgabe besitzt die Anfangssteigung

$$y'(0) = 4.57504614 \cdot 10^{-2}$$

und eine singuläre Stelle in

$$x_s \approx 1.0329.$$

Diese liegt in unmittelbarer Nähe des rechten Randes, und man kann elementar zeigen, dass schon für $y'(0) \ge 0.05$ die singuläre Stelle im Intervall $(0,1)$ liegt (vgl. *Stoer, Bulirsch* [54]). Man muss also die Anfangssteigung sehr genau kennen, um das Einfach–Schießverfahren überhaupt durchführen zu können. □

In (6.27) haben wir gesehen, dass Lösungen zu verschiedenen Anfangswerten exponentiell auseinander laufen können wie

$$\|\boldsymbol{y}(x;\boldsymbol{\eta}_1) - \boldsymbol{y}(x;\boldsymbol{\eta}_2)\| \le \|\boldsymbol{\eta}_1 - \boldsymbol{\eta}_2\| e^{L(x-a)}.$$

Man kann also erwarten, dass bei kleinen Intervalllängen das in den Beispielen angesprochene Verhalten noch nicht auftritt. Wir zerlegen daher das Integrationsintervall in Teilintervalle

$$a = x_1 < x_2 < \ldots < x_N < x_{N+1} = b,$$

schätzen in jedem Teilintervall einen Anfangswert $\boldsymbol{s}_j$, $j = 1, \ldots, N$, und lösen die Anfangswertaufgaben

$$\boldsymbol{y}_j' = f(x, \boldsymbol{y}_j),\ x_j < x < x_{j+1},\ \boldsymbol{y}_j(x_j) = \boldsymbol{s}_j.$$

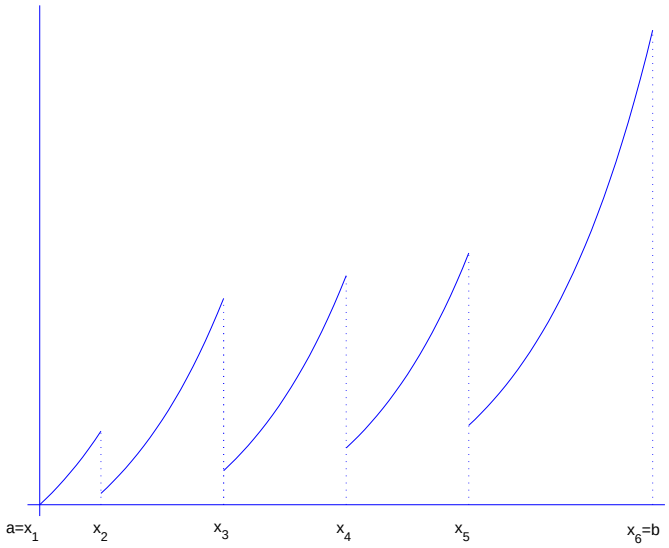

Abbildung 6.2: Mehrzielmethode

Kann man diese Lösungen $\boldsymbol{y}_j(\cdot; x_j, \boldsymbol{s}_j)$ stetig zusammensetzen,

$$\lim_{x \to x_j - 0} \boldsymbol{y}_{j-1}(x; x_{j-1}, \boldsymbol{s}_{j-1}) = \boldsymbol{s}_j = \lim_{x \to x_j + 0} \boldsymbol{y}_j(x; x_j, \boldsymbol{s}_j),$$

so gilt auch

$$\begin{aligned} \lim_{x \to x_j - 0} \boldsymbol{y}'_{j-1}(x; x_{j-1}, \boldsymbol{s}_{j-1}) &= \lim_{x \to x_j - 0} \boldsymbol{f}(x, \boldsymbol{y}_{j-1}(x; x_{j-1}, \boldsymbol{s}_{j-1})) \\ &= f(x_j, \boldsymbol{s}_j) = \lim_{x \to x_j + 0} \boldsymbol{y}'_j(x; x_j, \boldsymbol{s}_j). \end{aligned}$$

Es ist also die zusammengesetzte Funktion sogar stetig differenzierbar und erfüllt im ganzen Intervall $a < x < b$ die Differentialgleichung. Ist auch noch die Randbedingung erfüllt, so haben wir die Randwertaufgabe gelöst.

Wir betrachten nun das lineare System (6.7), (6.8)

$$\boldsymbol{y}' = \boldsymbol{C}(x)\boldsymbol{y} + \boldsymbol{f}(x), \;\; \boldsymbol{A}\boldsymbol{y}(a) + \boldsymbol{B}\boldsymbol{y}(b) = \boldsymbol{c}$$

genauer. Auf dem Teilintervall $[x_j, x_{j+1}]$ können wir die Lösung der Differentialgleichung (6.7) schreiben als

$$\boldsymbol{y}_j(x) = \boldsymbol{Y}_j(x)\boldsymbol{s}_j + \boldsymbol{v}_j(x), \;\; x_j \le x \le x_{j+1}. \tag{6.30}$$

Dabei ist $\boldsymbol{Y}_j(x)$ eine Fundamentalmatrix des zu (6.7) gehörigen homogenen Problems, $\boldsymbol{s}_j$ ist ein Parameter, und $\boldsymbol{v}_j(x)$ ist eine spezielle Lösung des Systems (6.7). Man erhält $\boldsymbol{Y}_j$ als (numerische) Lösung der Anfangswertaufgabe

$$\boldsymbol{Y}'_j(x) = \boldsymbol{C}(x)\boldsymbol{Y}_j, \;\; x_j < x < x_{j+1}, \;\; \boldsymbol{Y}_j(x_j) = \boldsymbol{F}_j, \tag{6.31}$$

mit einer vorgegebenen regulären Matrix $\boldsymbol{F}_j \in \mathbb{R}^{(n,n)}$ und $\boldsymbol{v}_j$ als (numerische) Lösung der Anfangswertaufgabe

$$\boldsymbol{v}_j'(x) = \boldsymbol{C}(x)\boldsymbol{v}_j(x) + \boldsymbol{f}(x),\ x_j < x < x_{j+1},\ \boldsymbol{v}_j(x_j) = \boldsymbol{\eta}_j \tag{6.32}$$

mit vorgegeben Vektoren $\boldsymbol{\eta}_j$.

Eine häufige Wahl ist $\boldsymbol{F}_j = \boldsymbol{E}$ und $\boldsymbol{\eta}_j = \boldsymbol{0}$ für alle j.

Unser Problem ist es, die $n \cdot N$ Parameter $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$ aus den Stetigkeitsbedingungen

$$\boldsymbol{Y}_j(x_{j+1})\boldsymbol{s}_j + \boldsymbol{v}_j(x_{j+1}) = \boldsymbol{Y}_{j+1}(x_{j+1})\boldsymbol{s}_{j+1} + \boldsymbol{v}_{j+1}(x_{j+1}),\ 1 \le j \le N-1,$$

und den Randbedingungen

$$\boldsymbol{A}(\boldsymbol{Y}_1(a)\boldsymbol{s}_1 + \boldsymbol{v}_1(a)) + \boldsymbol{B}(\boldsymbol{Y}_N(b)\boldsymbol{s}_N + \boldsymbol{v}_N(b)) = \boldsymbol{c}$$

zu bestimmen. Berücksichtigt man die Anfangsbedingungen, so ist dies das lineare Gleichungssystem

$$\begin{pmatrix} -\boldsymbol{Y}_1(x_2) & \boldsymbol{F}_2 & & & & & \\ & -\boldsymbol{Y}_2(x_3) & \boldsymbol{F}_3 & & & & \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \\ & & & & -\boldsymbol{Y}_{N-1}(x_N) & \boldsymbol{F}_N & \\ \boldsymbol{A}\boldsymbol{F}_1 & & & & & \boldsymbol{B}\boldsymbol{Y}_N(b) & \end{pmatrix} \begin{pmatrix} \boldsymbol{s}_1 \\ \boldsymbol{s}_2 \\ \cdots \\ \boldsymbol{s}_{N-1} \\ \boldsymbol{s}_N \end{pmatrix}$$

$$= \begin{pmatrix} \boldsymbol{v}_1(x_2) - \boldsymbol{\eta}_2 \\ \boldsymbol{v}_2(x_3) - \boldsymbol{\eta}_3 \\ \cdots \\ \boldsymbol{v}_{N-1}(x_N) - \boldsymbol{\eta}_N \\ \boldsymbol{c} - \boldsymbol{A}\boldsymbol{\eta}_1 - \boldsymbol{B}\boldsymbol{v}_N(b) \end{pmatrix}. \tag{6.33}$$

Es ist nun verlockend, die j–te Gleichung

$$-\boldsymbol{Y}_j(x_{j+1})\boldsymbol{s}_j + \boldsymbol{F}_j\boldsymbol{s}_{j+1} = \boldsymbol{v}_j(x_{j+1}) - \boldsymbol{\eta}_{j+1}$$

nach $\boldsymbol{s}_{j+1}$ aufzulösen, hiermit $\boldsymbol{s}_N$, $\boldsymbol{s}_{N-1}$, …, $\boldsymbol{s}_2$ aus der letzten Gleichung zu eliminieren, das entstehende lineare Gleichungssystem in $\boldsymbol{s}_1$ zu lösen und dann durch Vorwärtseinsetzen $\boldsymbol{s}_2$, …, $\boldsymbol{s}_N$ zu ermitteln. Dieses Vorgehen kann jedoch instabil sein. Für

$$\boldsymbol{F}_j = \boldsymbol{Y}_{j-1}(x_j),\ \boldsymbol{\eta}_j = \boldsymbol{v}_{j-1}(x_j)$$

ist es dem Einfach–Schießverfahren äquivalent.

Stabile Methoden zur Lösung von Systemen der Gestalt (6.33) (staircase system) werden in *Ascher, Mattheij, Russel* [3], p. 303 ff, diskutiert.

Die Systeme (6.30), (6.31) sind für verschiedene j unabhängig und können daher prinzipiell parallel gelöst werden. Dazu muss allerdings die Zerlegung $x_1 < \ldots < x_{N+1}$ vorher festgelegt werden. Die Notwendigkeit, die Mehrzielmethode anzuwenden, rührt häufig daher, dass die Lösungen der Anfangswertaufgaben bei gestörten Anfangswerten sehr stark wachsen oder dass verschiedene Lösungskomponenten der linearen Systeme (6.30) verschieden stark wachsen und daher die Fundamentalmatrix $\boldsymbol{Y}(x)$ nahezu linear abhängige Spalten besitzt. In diesem Fall ist es sinnvoll, das Wachstum der Lösungen zu beobachten, bei Bedarf einen neuen Zwischenpunkt x_j zu wählen, und eine neue Matrix $\boldsymbol{F}_j$ etwa durch QR–Zerlegung von $\boldsymbol{Y}_{j-1}(x_j)$ zu bestimmen.

Die Übertragung der Mehrzielmethode auf nichtlineare Randwertaufgaben

$$\boldsymbol{y}' = \boldsymbol{f}(x, \boldsymbol{y}),\ a \le x \le y,\ \boldsymbol{R}(\boldsymbol{y}(a), \boldsymbol{y}(b)) = \boldsymbol{0},$$

ist offensichtlich. Es sei $a = x_1 < x_2 < \ldots < x_N < x_{N+1} = b$ eine (vorgegebene oder während des Lösungsprozesses erzeugte) Zerlegung des Intervalls. Man bestimme für $j = 1, \ldots, N$ die (numerischen) Lösungen $\boldsymbol{y}_j(x; x_j, \boldsymbol{s}_j)$ der Anfangswertaufgaben

$$\boldsymbol{y}' = f(x, \boldsymbol{y}),\ \boldsymbol{y}(x_j) = \boldsymbol{s}_j,$$

und bestimme die $\boldsymbol{s}_j$ so, dass die zusammengesetzte Funktion stetig ist und die Randbedingungen erfüllt, d.h.

$$\boldsymbol{F}(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N) := \begin{pmatrix} \boldsymbol{s}_2 - \boldsymbol{y}_1(x_2; x_1, \boldsymbol{s}_1) \\ \boldsymbol{s}_3 - \boldsymbol{y}_2(x_3; x_2, \boldsymbol{s}_2) \\ \vdots \\ \boldsymbol{s}_N - \boldsymbol{y}_{N-1}(x_N; x_{N-1}, \boldsymbol{s}_{N-1}) \\ \boldsymbol{R}(\boldsymbol{s}_1, \boldsymbol{y}_N(b; x_N, \boldsymbol{s}_N)) \end{pmatrix} = \boldsymbol{0}. \tag{6.34}$$

Die Jacobi Matrix von $\boldsymbol{F}$ hat offenbar wieder Treppengestalt, so dass man die hierfür entwickelten Verfahren in den einzelnen Schritten des Newton Verfahrens verwenden kann. Genauso bleibt diese Struktur für Differenzenapproximationen der Jacobi Matrix erhalten. Das Broyden Verfahren läßt sich jedoch nicht unmittelbar übertragen.

6.2 Differenzenverfahren

Die Idee der Differenzenverfahren ist (ähnlich wie bei Anfangswertaufgaben), die Differentialgleichung nur auf einer endlichen Zerlegung

$$Z : a = x_0 < x_1 < \ldots < x_N = b$$

zu betrachten und die auftretenden Ableitungen durch Differenzenquotienten zu ersetzen.

Wir betrachten zunächst die lineare Differentialgleichung zweiter Ordnung

$$-y'' + p(x)y' + q(x)y = f(x),\ a < x < b, \tag{6.35}$$

wobei p, q, f gegebene stetige Funktionen auf $[a, b]$ sind, mit Dirichletschen Randbedingungen

$$y(a) = \gamma_1,\ y(b) = \gamma_2. \tag{6.36}$$

Wir nehmen an, dass

$$q(x) \geq 0 \quad \text{für alle } x \in [a, b] \tag{6.37}$$

gilt. Dann besitzt nach Satz 1.17. die Randwertaufgabe (6.35), (6.36) eine eindeutige Lösung $\tilde{y}$.

Nach dem Taylorschen Satz gilt

$$y'(x) = \frac{y(x+h) - y(x)}{h} + O(h) \quad \text{für } y \in C^2[a, b], \tag{6.38}$$

$$y'(x) = \frac{y(x+h) - y(x-h)}{2h} + O(h^2) \quad \text{für } y \in C^3[a, b], \tag{6.39}$$

$$y''(x) = \frac{y(x+h) - 2y(x) + y(x-h)}{h^2} + O(h^2) \quad \text{für } y \in C^4[a, b]. \tag{6.40}$$

Es liegt daher nahe, ein äquidistantes Gitter

$$x_j := a + jh,\ j = 0, \ldots, N,\ h := \frac{b-a}{N},$$

zu betrachten und die Ableitungen in (6.35) durch die **zentralen Differenzenquotienten** in (6.39) bzw. (6.40) zu ersetzen. Mit den Bezeichnungen $p_j := p(x_j)$, $q_j := q(x_j)$, $f_j := f(x_j)$ erhält man dann für die Näherungen y_j der Lösung $\tilde{y}(x_j)$ in den Knoten x_j das lineare Gleichungssystem

$$\frac{-y_{j+1} + 2y_j - y_{j-1}}{h^2} + p_j \frac{y_{j+1} - y_{j-1}}{2h} + q_j y_j = f_j,\ j = 1, \ldots, N-1, \tag{6.41}$$

$$y_0 = \gamma_1,\ y_N = \gamma_2. \tag{6.42}$$

Eliminiert man y_0 und y_N mit Hilfe der Randbedingungen (6.42) aus (6.41), so ergibt sich das lineare Gleichungssystem

$$\boldsymbol{A}_h \boldsymbol{y}_h = \boldsymbol{f}_h \tag{6.43}$$

zur Bestimmung von $\boldsymbol{y}_h := (y_1, \ldots, y_{N-1})^T$, wobei

$$\boldsymbol{A}_h := \frac{1}{h^2}\text{tridiag}\Big(-1 - \frac{h}{2}p_j, 2 + h^2 q_j, -1 + \frac{h}{2}p_j\Big), \tag{6.44}$$

$$\boldsymbol{f}_h := \Big(f_1 + h^{-2}(1 + 0.5hp_1)\gamma_1, f_2, \ldots, f_{N-2}, f_{N-1} + h^{-2}(1 - 0.5hp_{N-1})\gamma_2\Big)^T. \tag{6.45}$$

Um die Konvergenz nachzuweisen, vergleicht man die Lösung des diskretisierten Problems (6.43) mit der Restriktion der Lösung des kontinuierlichen Problems auf das Gitter.

Definition 6.6. *Es sei für* $y \in C[a,b]$

$$R_h \tilde{y} := (\tilde{y}(x_1), \ldots, \tilde{y}(x_{N-1}))^T$$

die Restriktion auf die inneren Gitterpunkte. Das Differenzenverfahren (6.43) heißt **konvergent**, *falls*

$$\lim_{h \to 0} \|R_h \tilde{y} - \boldsymbol{y}_h\|_\infty = 0. \tag{6.46}$$

Den Konvergenzbeweis kann man so führen:

$$\|R_h \tilde{y} - \boldsymbol{y}_h\|_\infty = \|\boldsymbol{A}_h^{-1}(\boldsymbol{A}_h R_h \tilde{y} - \boldsymbol{A}_h \boldsymbol{y}_h)\|_\infty \leq \|\boldsymbol{A}_h^{-1}\|_\infty \|\boldsymbol{A}_h R_h \tilde{y} - \boldsymbol{f}_h\|_\infty. \tag{6.47}$$

Definition 6.7. *Das Differenzenverfahren heißt* **konsistent**, *falls*

$$\lim_{h \to 0} \|\boldsymbol{A}_h R_h \tilde{y} - \boldsymbol{f}_h\| = 0$$

gilt, es heißt **konsistent von der Ordnung** p, *falls*

$$\|\boldsymbol{A}_h R_h \tilde{y} - \boldsymbol{f}_h\|_\infty = O(h^p),$$

und es heißt **stabil**, *falls es ein* $h_0 > 0$ *und ein* $C > 0$ *gibt mit*

$$\|\boldsymbol{A}_h^{-1}\| \leq C \quad \textit{für alle } h \in (0, h_0).$$

Ist $\tilde{y} \in C^4[a,b]$, so folgt die Konsistenz von der Ordnung 2 sofort aus (6.39) und (6.40), denn es gilt für $j = 1, \ldots, N-1$

$$\begin{aligned}
&\Big(\boldsymbol{A}_h R_h \tilde{y} - \boldsymbol{f}_h\Big)_j \\
&= -\frac{\tilde{y}(x_{j-1}) - 2\tilde{y}(x_j) + \tilde{y}(x_{j+1})}{h^2} + p(x_j)\frac{\tilde{y}(x_{j+1}) - \tilde{y}(x_{j-1})}{2h} + q(x_j)\tilde{y}(x_j) - f(x_j) \\
&= -y''(x_j) + p(x_j)y'(x_j) + q(x_j)y(x_j) - f(x_j) + O(h^2) \\
&= O(h^2).
\end{aligned}$$

Die Stabilität erhält man mit Hilfe der folgenden Eigenschaften der Matrizen $\boldsymbol{A}_h$.

Definition 6.8. *Die Matrix $\boldsymbol{B} \in \mathbb{R}^{(k,k)}$ heißt* **inversmonoton**, *falls für $\boldsymbol{y} \in \mathbb{R}^k$ gilt*

$$\boldsymbol{B}\boldsymbol{y} \geq \boldsymbol{0} \quad \Longrightarrow \quad \boldsymbol{y} \geq \boldsymbol{0}, \tag{6.48}$$

wobei die Ungleichungen komponentenweise gemeint sind.

Lemma 6.9. *Es sei $h_0 := 2/\max_{a\leq x\leq b} |p(x)|$. Dann ist die Matrix $\boldsymbol{A}_h$ inversmonoton für alle $h \in (0, h_0)$*

Beweis: Wir nehmen an, dass es ein $\boldsymbol{y} \in \mathbb{R}^{N-1}$ gibt mit $\boldsymbol{A}_h\boldsymbol{y} \geq \boldsymbol{0}$ und $\boldsymbol{y} \not\geq \boldsymbol{0}$. Es sei $j \in \{1, \ldots, N-1\}$ mit

$$y_j := \min_{i=1,\ldots,N-1} y_i < 0.$$

Aus

$$(\boldsymbol{A}_h\boldsymbol{y})_j \geq 0$$

folgt

$$\Big(-1 - \frac{h}{2}p_j\Big)y_{j-1} + 2y_j + \Big(-1 + \frac{h}{2}p_j\Big)y_{j+1} \geq -h^2 q_j y_j \geq 0,$$

und daher

$$y_j \geq \frac{1}{2}\Big(\Big(1 + \frac{h}{2}p_j\Big)y_{j-1} + \Big(1 - \frac{h}{2}p_j\Big)y_{j+1}\Big) =: \alpha_j y_{j-1} + \beta_j y_{j+1}.$$

Wegen $h \in (0, h_0)$ ist $\alpha_j \geq 0$, $\beta_j \geq 0$, und wegen $\alpha_j + \beta_j = 1$, $y_j \leq y_{j-1}$ und $y_j \leq y_{j+1}$ folgt

$$y_j = y_{j-1} = y_{j+1}.$$

Durch Wiederholung dieser Schlussweise erhalten wir, dass

$$y_1 = y_2 = \ldots = y_{N-1}$$

gilt. Dies ist aber nicht möglich, denn die erste Ungleichung lautet dann

$$(\boldsymbol{A}_h\boldsymbol{y})_1 = \frac{1}{h^2}\Big((2+h^2q_1)y_1 + (-1+0.5hp_1)y_1\Big) = \frac{1}{h^2}\Big(1+h^2q_1+0.5hp_1\Big)y_1 < 0.$$

∎

Inversmonotone Matrizen haben die folgenden Eigenschaften

Satz 6.10. *Es sei $\boldsymbol{A} \in \mathbb{R}^{(n,n)}$ inversmonoton. Dann gilt*

(i) $\boldsymbol{A}$ *ist regulär.*

(ii) Alle Elemente von $\boldsymbol{A}^{-1}$ *sind nichtnegativ.*

(iii) Ist $\boldsymbol{w} \in \mathbb{R}^n$ *mit* $(\boldsymbol{A}\boldsymbol{w})_j \geq 1$ *für alle* $j \in \{1,\ldots,n\}$*, so gilt*

$$\|\boldsymbol{A}^{-1}\|_\infty \leq \|\boldsymbol{w}\|_\infty.$$

Beweis: $\boldsymbol{A}$ ist regulär, denn aus $\boldsymbol{A}\boldsymbol{y} = \boldsymbol{0}$ folgt $\boldsymbol{A}\boldsymbol{y} \geq \boldsymbol{0}$, d.h. $\boldsymbol{y} \geq \boldsymbol{0}$, und zugleich $\boldsymbol{A}(-\boldsymbol{y}) \geq \boldsymbol{0}$, d.h. $-\boldsymbol{y} \geq \boldsymbol{0}$; zusammen also $\boldsymbol{y} = \boldsymbol{0}$.

Alle Elemente der inversen Matrix $\boldsymbol{A}^{-1}$ sind nichtnegativ, denn wenn die Matrix $\boldsymbol{A}^{-1} =: (\alpha_{ij})$ ein negatives Element $\alpha_{k\ell}$ besäße, so würde

$$(\boldsymbol{A}^{-1}\boldsymbol{e}^\ell)_k = \alpha_{k\ell} < 0,$$

gelten, und damit würde die Lösung von $\boldsymbol{A}\boldsymbol{y} = \boldsymbol{e}^\ell \geq \boldsymbol{0}$ eine negative Komponente besitzen.

Für die Zeilensummennorm gilt dann

$$\|\boldsymbol{A}^{-1}\|_\infty = \|\boldsymbol{A}^{-1}\boldsymbol{e}\|_\infty, \quad \boldsymbol{e} := (1,1,\ldots,1)^T.$$

Daher folgt aus $\boldsymbol{A}\boldsymbol{w} \geq \boldsymbol{e}$ wegen der Inversmonotonie von $\boldsymbol{A}$

$$\boldsymbol{w} \geq \boldsymbol{A}^{-1}\boldsymbol{e} \geq \boldsymbol{0},$$

und damit

$$\|\boldsymbol{w}\|_\infty \geq \|\boldsymbol{A}^{-1}\boldsymbol{e}\|_\infty = \|\boldsymbol{A}^{-1}\|_\infty.$$

∎

Um die Stabilität des Differenzenverfahrens nachzuweisen, haben wir Vektoren $\boldsymbol{w}_h$ zu konstruieren mit $\boldsymbol{w}_h > \boldsymbol{0}$ und $\boldsymbol{A}_h\boldsymbol{w}_h \geq \boldsymbol{e}$, deren Maximumnorm $\|\boldsymbol{w}_h\|_\infty$ nach oben (gleichmäßig bzgl. h) beschränkt ist.

Der Einfachheit halber betrachten wir nur den Fall $p(x) \equiv 0$. Die Lösung $\boldsymbol{v}_h$ des linearen Gleichungssystems $\boldsymbol{A}_h \boldsymbol{v}_h = \boldsymbol{e}$ ist die Lösung der diskreten Version der Randwertaufgabe

$$-v'' + q(x)v = 1,\ v(a) = 0,\ v(b) = 0.$$

Es sei

$$w(x) := \frac{1}{2}(x-a)(b-x)$$

die Lösung der Randwertaufgabe

$$-w'' = 1,\ w(a) = 0,\ w(b) = 0.$$

Dann gilt

$$-(w-v)'' = 1 - (1 - q(x)v(x)) = q(x)v(x) \geq 0,\ (w-v)(a) = 0,\ (w-v)(b) = 0,$$

und aus der Inversmonotonie der Randwertaufgabe erhalten wir

$$w(x) \geq v(x) \quad \text{für alle } x \in [a,b].$$

Für die Restriktion

$$(R_h w)_j = (w(x_j)) = \Big(\frac{1}{2}(x_j - a)(b - x_j)\Big) = \Big(\frac{h^2}{2} j(N-j)\Big)$$

gilt

$$(\boldsymbol{A}_h R_h w)_j = \frac{1}{2} h^2 q(x_j) j(N-j) + 1 \geq 1, \quad \text{also } \boldsymbol{A}_h R_h w \geq \boldsymbol{e}.$$

Für jede Komponente von $R_h w$ hat man

$$(R_h w)_j = \frac{1}{2} h^2 j(N-j) \leq \frac{1}{8} h^2 N^2 = \frac{1}{8}(b-a)^2.$$

Daher ist das Verfahren im Fall $p(x) \equiv 0$ stabil, und es gilt

$$\|\boldsymbol{A}_h^{-1}\|_\infty \leq \|R_h w\|_\infty \leq \frac{1}{8}(b-a)^2.$$

Für den allgemeinen Fall $p(x) \not\equiv 0$ kann man für

$$h_0 \leq \frac{1}{2 \max_{a \leq x \leq b} |p(x)|}$$

die Stabilitätsungleichung

$$\|\boldsymbol{A}_h^{-1}\|_\infty \leq \frac{1}{8}(b-a)^2 \exp\Big(\frac{(b-a)\max_{a\leq x\leq b}|p(x)|}{2 - h_0 \max_{a\leq x\leq b}|p(x)|}\Big) \quad \text{für alle } h \in (0, h_0)$$

beweisen.

Damit erhält man

Satz 6.11. *Unter der Voraussetzung (6.37) konvergieren die Lösungen der diskreten Aufgabe (6.43) von der Ordnung 2 gegen die Lösung der Randwertaufgabe (6.35), (6.36).*

Bemerkung 6.12. Da die Matrix $\boldsymbol{A}_h$ für $h < 1/(2\max_{a\le x\le b}|p(x)|)$ eine diagonaldominante Tridiagonalmatrix ist, kann man (6.43) problemlos durch Elimination ohne Pivotsuche lösen. □

Bemerkung 6.13. Wir haben nur die erste Randbedingung (6.36) betrachtet. Die allgemeine Sturmsche Randbedingung

$$\alpha_0 y(a) - \alpha_1 y'(a) = \gamma_1,\ \beta_0 y(b) + \beta_1 y'(b) = \gamma_2 \tag{6.49}$$

kann man diskretisieren durch einseitige Differenzenquotienten

$$\alpha_0 y_0 - \alpha_1 \frac{y_1 - y_0}{h} = \gamma_1,\ \beta_0 y_N + \beta_1 \frac{y_N - y_{N-1}}{h} = \gamma_2, \tag{6.50}$$

oder durch zentrale Differenzenquotienten

$$\alpha_0 y_0 - \alpha_1 \frac{y_1 - y_{-1}}{2h} = \gamma_1,\ \beta_0 y_N + \beta_1 \frac{y_{N+1} - y_{N-1}}{2h} = \gamma_2. \tag{6.51}$$

Im zweiten Fall nimmt man also zwei außerhalb des Intervalls liegende Hilfspunkte $y_{-1} := a - h$ und $y_{N+1} := b + h$ hinzu. In diesem Fall muss man die Differentialgleichung auch in den Randpunkten $x_0 = a$ und $x_N = b$ durch (6.41) diskretisieren und erhält ein lineares Gleichungssystem in $N + 3$ Variablen.

Gilt $\alpha_0, \alpha_1, \beta_0, \beta_1 \ge 0$ und $\alpha_0 + \alpha_1 > 0$, $\beta_0 + \beta_1 > 0$ sowie im Fall $q(x) \equiv 0$ auch $\alpha_0 + \beta_0 > 0$, so sind wieder die Randwertaufgabe und ihre Diskretisierung invers monoton, und man erhält wie oben die Stabilität. Bei der Diskretisierung (6.50) ist der lokale Fehler $O(h)$, und die Ordnung 1 überträgt sich auf den globalen Fehler, bei der Diskretisierung (6.51) erhält man für den lokalen Fehler und dann auch für den globalen Fehler die Ordnung 2. □

Bemerkung 6.14. Wesentlich für unsere Betrachtungen ist, dass die Inversmonotonie der Randwertaufgabe sich auf das diskrete Problem vererbt. Diskretisiert man die Randwertaufgabe

$$-y'' + q(x)y = f(x),\ y(a) = \gamma_1,\ y(b) = \gamma_2 \tag{6.52}$$

durch

$$\tilde{\boldsymbol{A}}_h \boldsymbol{y}_h = \tilde{\boldsymbol{f}}_h \tag{6.53}$$

mit

$$\tilde{\boldsymbol{A}}_h = \frac{1}{h^2}\text{tridiag}\Big(-1+\frac{h^2}{12}q_{j-1}, 2+\frac{10h^2}{12}q_j, -1+\frac{h^2}{12}q_{j+1}\Big), \tag{6.54}$$

$$\tilde{\boldsymbol{f}}_h = \Big(\frac{1}{12}(f_0+10f_1+f_2)+\frac{1}{h^2}(1-\frac{h^2}{12}q_0)\gamma_1, \ldots, \frac{1}{12}(f_{j-1}+10f_j+f_{j+1}), \ldots$$
$$\ldots, \frac{1}{12}(f_{N-2}+10f_{N-1}+f_N)+\frac{1}{h^2}(1-\frac{h^2}{12}q_N)\gamma_2\Big), \tag{6.55}$$

so ist $\tilde{\boldsymbol{A}}_h$ für alle

$$h \in (0, h_0), \quad h_0 := \sqrt{\frac{12}{\max_{a\le x\le b}|q(x)|}}$$

invers monoton. Daher kann man wie vorher eine Stabilitätsungleichung nachweisen, und da der lokale Fehler in diesem Fall durch Ch^4 beschränkt ist, erhält man eine Approximation der Ordnung 4.

Die Diskretisierung durch (6.54), (6.55) nennt man eine **Mehrstellenformel** (engl.: Hermite formula). Weitere Mehrstellenformeln findet man in *Collatz* [11]. □

Bemerkung 6.15. Wir haben vorausgesetzt, dass $q(x) \ge 0$ in $[a,b]$ gilt. Ohne diese Voraussetzung ist die Randwertaufgabe und dann auch die Matrix $\boldsymbol{A}_h$ nicht notwendig invers monoton. Ist die Randwertaufgabe (6.35), (6.36) eindeutig lösbar, so kann man (allerdings mit einem wesentlich aufwendigeren Beweis) eine Stabilitätsungleichung für genügend kleine $h > 0$ zeigen, und erhält so die Konvergenz. □

Für allgemeine lineare Systeme

$$\boldsymbol{y}' = \boldsymbol{C}(x)\boldsymbol{y} + \boldsymbol{f}(x),\ a \le x \le b, \tag{6.56}$$

$$\boldsymbol{A}\boldsymbol{y}(a) + \boldsymbol{B}\boldsymbol{y}(b) = \boldsymbol{\gamma} \tag{6.57}$$

können wir darauf verzichten, dass die Zerlegung äquidistant ist. Wir betrachten

$$a = x_0 < x_1 < \ldots < x_N = b.$$

Dann ist der Differenzenquotient

$$\frac{\boldsymbol{y}_j - \boldsymbol{y}_{j-1}}{h_j},\ h_j := x_j - x_{j-1},$$

eine Approximation von $\boldsymbol{y}'(x_{j-1/2})$, $x_{j-1/2} := 0.5(x_{j-1}+x_j)$ durch den zentralen Differenzenquotient, also eine Approximation der Ordnung 2.

Hiermit liegen die Diskretisierungen

$$\frac{1}{h_j}(\boldsymbol{y}_j - \boldsymbol{y}_{j-1}) = \frac{1}{2}(\boldsymbol{C}(x_j)\boldsymbol{y}_j + \boldsymbol{C}(x_{j-1})\boldsymbol{y}_{j-1}) + \frac{1}{2}(\boldsymbol{f}(x_j) + \boldsymbol{f}(x_{j-1})),\ 1 \le j \le N, \quad (6.58)$$

und

$$\frac{1}{h_j}(\boldsymbol{y}_j - \boldsymbol{y}_{j-1}) = \frac{1}{2}\boldsymbol{C}(x_{j-1/2})(\boldsymbol{y}_j + \boldsymbol{y}_{j-1}) + \boldsymbol{f}(x_{j-1/2}),\ 1 \le j \le N, \quad (6.59)$$

der Randwertaufgabe nahe. Ergänzt man diese um die Randbedingungen

$$\boldsymbol{A}\boldsymbol{y}_0 + \boldsymbol{B}\boldsymbol{y}_N = \boldsymbol{\gamma}, \quad (6.60)$$

so erhält man ein lineares System von $(N+1)n$ Gleichungen in den Unbekannten $\boldsymbol{y}_0, \ldots, \boldsymbol{y}_N$.

Die Diskretisierung (6.58) heißt **Trapezregel** und die Diskretisierung (6.59) **Mittelpunktregel** oder auch **Boxschema**.

In beiden Fällen hat das lineare System die Blockgestalt

$$\begin{pmatrix} \boldsymbol{S}_1 & \boldsymbol{R}_1 & \boldsymbol{O} & \ldots & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{S}_2 & \boldsymbol{R}_2 & \ldots & \boldsymbol{O} & \boldsymbol{O} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \ldots & \boldsymbol{S}_N & \boldsymbol{R}_N \\ \boldsymbol{A} & \boldsymbol{O} & \boldsymbol{O} & \ldots & \boldsymbol{O} & \boldsymbol{B} \end{pmatrix} \begin{pmatrix} \boldsymbol{y}_0 \\ \boldsymbol{y}_1 \\ \ldots \\ \boldsymbol{y}_{N-1} \\ \boldsymbol{y}_N \end{pmatrix} = \begin{pmatrix} \boldsymbol{f}_1 \\ \boldsymbol{f}_2 \\ \ldots \\ \boldsymbol{f}_N \\ \boldsymbol{\gamma} \end{pmatrix} \quad (6.61)$$

mit

$$\boldsymbol{S}_j = -\frac{1}{h_j}\boldsymbol{E} - \frac{1}{2}\boldsymbol{C}(x_{j-1}),\ \boldsymbol{R}_j = \frac{1}{h_j}\boldsymbol{E} - \frac{1}{2}\boldsymbol{C}(x_j),\ \boldsymbol{f}_j = \frac{1}{2}(\boldsymbol{f}(x_{j-1}) + \boldsymbol{f}(x_j))$$

im Falle der Trapezregel und

$$\boldsymbol{S}_j = -\frac{1}{h_j}\boldsymbol{E} - \frac{1}{2}\boldsymbol{C}(x_{j-1/2}),\ \boldsymbol{R}_j = \frac{1}{h_j}\boldsymbol{E} - \frac{1}{2}\boldsymbol{C}(x_{j-1/2}),\ \boldsymbol{f}_j = \boldsymbol{f}(x_{j-1/2})$$

im Falle der Mittelpunktregel.

In jedem Fall hat also die Systemmatrix dieselbe Besetzungsstruktur wie bei der Mehrzielmethode, und es können die dafür entwickelten Verfahren auch für diese Differenzenverfahren eingesetzt werden. Da man die Stabilität zeigen kann (vgl. *Ascher, Mattheij, Russel* [3] p. 201) und das Verfahren konsistent von der Ordnung 2 ist, konvergiert es von der Ordnung 2.

Die Übertragung auf nichtlineare Probleme ist offensichtlich. Wir verzichten darauf. Klar ist, dass die linearen Gleichungssysteme, die man in dem Newton Verfahren

für das diskrete Problem zu lösen hat, wieder die Besetzungsstruktur wie in (6.61) haben.

Verfahren höherer Ordnung kann man konstruieren, indem man die Differentialgleichung

$$\boldsymbol{y}' = \boldsymbol{f}(x, \boldsymbol{y}) \tag{6.62}$$

über das Teilintervall $[x_{j-1}, x_j]$ integriert

$$\boldsymbol{y}(x_j) - \boldsymbol{y}(x_{j-1}) = \int_{x_{j-1}}^{x_j} \boldsymbol{f}(t, \boldsymbol{y}(t))\, dt,$$

und das Integral der rechten Seite mit einer Quadraturformel behandelt.

Ersetzt man z.B. die Funktion

$$\boldsymbol{g}(x) := \boldsymbol{f}(x, \boldsymbol{y}(x))$$

(komponentenweise) durch das kubische Polynom $\boldsymbol{\phi}$, das die Bedingungen

$$\boldsymbol{\phi}(x_{j-1}) = \boldsymbol{g}(x_{j-1}),\ \boldsymbol{\phi}'(x_{j-1}) = \boldsymbol{g}'(x_{j-1}),\ \boldsymbol{\phi}(x_j) = \boldsymbol{g}(x_j),\ \boldsymbol{\phi}'(x_j) = \boldsymbol{g}'(x_j)$$

erfüllt, so erhält man nach Integration die Differenzenapproximation

$$\frac{\boldsymbol{y}_j - \boldsymbol{y}_{j-1}}{h_j} = \frac{1}{2}(\boldsymbol{f}(x_{j-1}, \boldsymbol{y}_{j-1}) + \boldsymbol{f}(x_j, \boldsymbol{y}_j)) + \frac{h_j}{12}(\boldsymbol{f}'(x_{j-1}, \boldsymbol{y}_{j-1}) + \boldsymbol{f}'(x_j, \boldsymbol{y}_j)), \tag{6.63}$$

wobei

$$\boldsymbol{f}'(x_j, \boldsymbol{y}_j) = \frac{\partial}{\partial x}\boldsymbol{f}(x_j, \boldsymbol{y}_j) + \frac{\partial}{\partial \boldsymbol{y}}\boldsymbol{f}(x_j, \boldsymbol{y}_j)\boldsymbol{f}(x_j, \boldsymbol{y}_j)$$

die totale Ableitung von $\boldsymbol{f}(x, \boldsymbol{y}(x))$ bezeichnet.

Das Differenzenverfahren (6.63), (6.60) hat die Ordnung 4. Nachteil ist aber, das man die totale Ableitung von $\boldsymbol{f}$ benötigt.

Eine andere Möglichkeit zur Konstruktion von Verfahren höherer Ordnung ist die Verwendung von (impliziten) Runge–Kutta Verfahren für die Differentialgleichung (6.62) in jedem der Intervalle $[x_{j-1}, x_j]$ (vgl. *Ascher, Mattheij, Russel* [3] p. 210 ff).

6.3 Variationsmethoden

Obwohl Variationsmethoden für gewöhnliche Randwertaufgaben praktisch kaum von Bedeutung sind, sollen sie hier wegen ihrer leichten theoretischen Zugänglichkeit als Vorbereitung für finite Elementmethoden für elliptische Probleme behandelt werden.

Wir betrachten die lineare Randwertaufgabe

$$Ly(x) := -(p(x)y')' + q(x)y = f(x),\ a < x < b, \quad y(a) = y(b) = 0. \tag{6.64}$$

Die in (6.35) behandelte Differentialgleichung kann man hierauf zurückführen, wenn man (6.35) mit $\exp(-\int^x p(t)\,dt)$ multipliziert.

Wir setzen voraus, dass gilt

$$p \in C^1[a,b],\ q \in C[a,b],\ p(x) > 0,\ q(x) \geq 0 \text{ für alle } x \in [a,b]. \tag{6.65}$$

Unter diesen Voraussetzungen besitzt (6.64) nach Satz 1.18. für alle stetigen rechten Seiten f eine eindeutige Lösung.

Ist y die Lösung von (6.64), so gilt für alle $v \in C^1[a,b]$ mit $v(a) = 0$, $v(b) = 0$

$$\int_a^b v(x)((-p(x)y'(x))' + q(x)y(x))\,dx = \int_a^b v(x)f(x)\,dx,$$

und durch partielle Integration erhält man

$$\int_a^b (p(x)y'(x)v'(x) + q(x)y(x)v(x))\,dx = \int_a^b v(x)f(x)\,dx$$
$$\text{für alle } v \in C^1[a,b]\ :\ v(a) = v(b) = 0. \tag{6.66}$$

(6.66) ist auch für $y \in C^1[a,b]$ mit $y(a) = y(b) = 0$ sinnvoll, so dass wir (6.64) in eine Variationsgleichung in diesem Raum überführt haben. $C^1[a,b]$ (mit den üblichen Normen) hat aber für die folgende Theorie noch nicht genügend schöne Eigenschaften.

Definition 6.16. *Die Funktion $v : [a,b] \to \mathbb{R}$ heißt* **absolut stetig** *in $[a,b]$, wenn für alle $\varepsilon > 0$ ein $\delta > 0$ existiert mit der folgenden Eigenschaft:*

Gilt $a \leq x_1 < x_2 < x_3 < \ldots < x_{2n-1} < x_{2n} \leq b$ mit $\sum_{j=1}^n (x_{2j} - x_{2j-1}) < \delta$, so folgt $\sum_{j=1}^n |v(x_{2j-1}) - v(x_{2j})| < \varepsilon$.

Offensichtlich ist jede absolut stetige Funktion auch gleichmäßig stetig und jede differenzierbare Funktion mit beschränkter Ableitung absolut stetig, denn nach dem Mittelwertsatz gibt es $\zeta_j \in (x_{2j-1}, x_{2j})$ mit

$$\sum_{j=1}^n |v(x_{2j-1}) - v(x_{2j})| = \sum_{j=1}^n |v'(\zeta_j)|(x_{2j} - x_{2j-1}) \leq \sup_{a \leq x \leq b} |v'(x)| \cdot \delta = \varepsilon,$$

wenn man nur $\delta := \varepsilon / \sup_{a \le x \le b} |v'(x)|$ wählt.

Tatsächlich kann man zeigen, dass jede absolut stetige Funktion fast überall in $[a, b]$ differenzierbar ist. Wir definieren hiermit den **Sobolev Raum**

$$H^1(a,b) := \{v \,:\, [a,b] \to \mathbb{R} \,:\, v \text{ absolut stetig}, \int_a^b v'(x)^2\,dx < \infty\}$$

und weiter den Raum

$$H_0^1(a,b) := \{v \in H^1(a,b) \,:\, v(a) = v(b) = 0\}.$$

Ersetzt man in den obigen Überlegungen $C^1[a,b]$ durch $H_0^1(a,b)$, so gelangt man zu der **Variationsaufgabe**

Bestimme $y \in H_0^1(a,b)$, so dass

$$\langle y, v\rangle_L := \int_a^b (py'v' + qyv)\,dx = \int_a^b fv\,dx \quad \text{für alle } v \in H_0^1(a,b). \quad (6.67)$$

Unabhängig von dem Existenzresultat in Satz 1.18. für die Randwertaufgabe kann man mit funktionalanalytischen Mitteln leicht zeigen, dass die Variationsaufgabe eine eindeutige Lösung besitzt. Es ist nämlich

$$\langle y, v\rangle_L := \int_a^b (p(x)y'(x)v'(x) + q(x)y(x)v(x))\,dx$$

ein inneres Produkt auf $H_0^1(a,b)$, $H_0^1(a,b)$ ist mit diesem inneren Produkt ein Hilbertraum, und

$$F(v) := \int_a^b f(x)v(x)\,dx$$

ist ein stetiges Funktional auf $H_0^1(a,b)$. Daher liefert der folgende Satz die eindeutige Lösbarkeit der Variationsaufgabe (6.67).

Satz 6.17. (Darstellungssatz von Riesz) *Es sei F ein stetiges, lineares Funktional auf dem Hilbertraum V mit dem inneren Produkt $\langle\cdot,\cdot\rangle_V$. Dann gibt es genau ein $y \in V$ mit*

$$F(v) = \langle y, v\rangle_V.$$

Beweis: In jedem Buch über Funktionalanalysis.

Dieses Vorgehen zeigt, dass die Variationsaufgabe (6.67) für wesentlich allgemeinere Funktionen f eindeutig lösbar ist. Ist z.B. $f : [a,b] \to \mathbb{R}$ stückweise stetig auf $[a,b]$, so ist wieder

$$F(v) := \int_a^b f(x)v(x)\,dx$$

ein stetiges lineares Funktional auf $H_0^1(a,b)$, und daher besitzt (6.67) eine eindeutige Lösung. Diese wird sicher an Unstetigkeitsstellen von f nicht zweimal differenzierbar sein, und daher erfüllt y die Differentialgleichung in (6.64) nicht in jedem Punkt von (a,b). Es ist y also sicher keine Lösung der Randwertaufgabe (6.64). Wir nennen jede Lösung von (6.67) eine **verallgemeinerte Lösung** oder **schwache Lösung** von (6.64). Im Gegensatz dazu heißt eine zweimal stetig differenzierbare Funktion, die in jedem Punkt $x \in (a,b)$ die Differentialgleichung in (6.64) und die Randbedingungen erfüllt, eine **klassische Lösung** oder **starke Lösung** der Randwertaufgabe (6.64).

Man kann sogar noch allgemeinere rechte Seiten zulassen. Die Diracsche Delta Distribution

$$\delta_t(v) := v(t)$$

definiert für $t \in (a,b)$ ein lineares, stetiges Funktional auf $H_0^1(a,b)$. Daher besitzt für jedes feste $t \in (a,b)$ die Variationsaufgabe

$$\langle \boldsymbol{y}, \boldsymbol{v} \rangle_L = \delta_t(v) \quad \text{für alle } v \in H_0^1(a,b)$$

eine eindeutige Lösung $y_t \in H_0^1(a,b)$. Diese ist nach Definition die verallgemeinerte Lösung der Randwertaufgabe

$$-(py')' + qy = \delta_t,\ y(a) = y(b) = 0.$$

Man kann zeigen, dass für die Greensche Funktion $g(x,t)$ der Randwertaufgabe (6.64) $y_t(x) = g(x,t)$ für alle $x,t \in (a,b)$ gilt.

Die Variationsaufgabe (6.67) bietet die folgende Möglichkeit zur Diskretisierung der Randwertaufgabe (6.64). Wir wählen einen endlich dimensionalen Teilraum V_h von $H_0^1(a,b)$ und bestimmen die Approximation y_h als Lösung der endlich dimensionalen Variationsaufgabe

Bestimme $y_h \in V_h$, so dass

$$\langle y_h, v \rangle_L = F(v) \quad \text{für alle } v \in V_h. \tag{6.68}$$

Ist $v_1, \ldots, v_n$ eine Basis von V_h, so besitzt $y_h \in V_h$ die Darstellung

$$y_h = \sum_{j=1}^{n} \xi_j v_j,$$

und die endlich dimensionale Variationsgleichung (6.68) ist äquivalent dem linearen Gleichungssystem

$$\sum_{j=1}^{n} \langle v_j, v_k \rangle_L \xi_j = F(v_k), \quad k = 1, \ldots, n, \tag{6.69}$$

für die Koeffizienten $\xi_1, \ldots, \xi_n$. Dieses ist eindeutig lösbar, denn die Gramsche Matrix $(\langle v_j, v_k \rangle_L)_{j,k=1,\ldots,n}$ ist regulär, da die v_j linear unabhängig sind.

Die eindeutige Lösung von (6.68) heißt **Ritz–Galerkin Lösung** bzgl. des Ansatzraumes V_h.

Der Fehler von y_h läßt sich bzgl. der durch $\langle \cdot, \cdot \rangle_L$ induzierten Norm $\|v\|_L := \sqrt{\langle v, v \rangle_L}$, der sogenannten **Energienorm**, leicht abschätzen. Für alle $v \in V_h$ gilt

$$\langle y - y_h, v \rangle_L = \langle y, v \rangle_L - \langle y_h, v \rangle_L = F(v) - F(v) = 0, \tag{6.70}$$

und damit wegen $y_h - v \in V_h$ und der Cauchy – Schwarzschen Ungleichung

$$\begin{aligned} \|y - y_h\|_L^2 &= \langle y - y_h, y - y_h \rangle_L = \langle y - y_h, y - y_h \rangle_L + \langle y - y_h, y_h - v \rangle_L \\ &= \langle y - y_h, y - v \rangle_L \leq \|y - y_h\|_L \|y - v\|_L. \end{aligned}$$

Daher folgt für $y \neq y_h$

$$\|y - y_h\|_L \leq \inf_{v \in V_h} \|y - v\|_L, \tag{6.71}$$

und für $y = y_L$ ist diese Abschätzung trivial.

Das Ritz – Galerkin Verfahren liefert also die beste Approximation für y im Raum V_h in der Energienorm $\| \cdot \|_L$.

Tatsächlich interessiert man sich nicht so sehr für den Fehler in der Energienorm als für den Fehler in der Maximumnorm.

Ist $w \in H_0^1(a, b)$, so gilt wegen $w(a) = 0$

$$w(x) = \int_a^x w'(t)\, dt \quad \text{für alle } x \in [a, b].$$

Die Cauchy – Schwarzsche Ungleichung liefert

$$w^2(x) \leq \int_a^x 1\, dt \int_a^x w'(t)^2\, dt = (x - a) \int_a^x w'(t)^2\, dt \leq (b - a) \int_a^b w'(t)^2\, dt, \tag{6.72}$$

und daher folgt

$$\begin{aligned} \|w\|_L^2 &= \int_a^b (p(t)w'(t)^2 + q(t)w(t)^2)\,dt \geq \min_{x\in[a,b]} p(x) \int_a^b w'(t)^2\,dt \\ &\geq \frac{1}{b-a} \min_{x\in[a,b]} p(x)\cdot w(x)^2. \end{aligned}$$

Damit ist gezeigt:

Satz 6.18. *Es existiert ein $C > 0$ mit*

$$\|w\|_\infty \leq C\|w\|_L \quad \textit{für alle } w \in H_0^1(a,b). \tag{6.73}$$

Bemerkung 6.19. Satz 6.18. bedeutet, dass die für die Norm $\|\cdot\|_L$ bewiesene Konvergenzgeschwindigkeit auch in der Maximumnorm eintritt. Wir weisen jedoch ausdrücklich darauf hin, dass diese Aussage nur für eindimensionale Grundgebiete (a,b), also bei gewöhnlichen Differentialgleichungen richtig ist. Schon für ebene Gebiete, also bei allen partiellen Randwertaufgaben, gilt (6.73) nicht mehr. □

Bemerkung 6.20. Ist $w \in C^1[a,b]$ mit $w(a) = w(b) = 0$, so folgt aus (6.72) sogar

$$\|w\|_\infty^2 \leq (b-a)^2\|w'\|_\infty^2,$$

und daher

$$\|w\|_L^2 \leq \Big(\|p\|_\infty\|w'\|_\infty^2 + \|q\|_\infty\|w\|_\infty^2\Big)(b-a) \leq C\|w'\|_\infty^2. \tag{6.74}$$

Ist also die Approximationsgüte für die Lösung y in V_n bzgl. der Norm $\|w\| := \|w'\|_\infty$ bekannt, so vererbt sich diese auf die Norm $\|\cdot\|_L$, und wegen Satz 6.18. gilt sie auch für $\|w\|_\infty$. □

Beispiel 6.21. Es sei

$$V_h := \{(x-a)(b-x)\sum_{j=0}^n a_j x^j \;:\; a_j \in \mathbb{R}\}.$$

Nach dem Satz von Jackson (vgl. G. Meinardus: Approximation of Functions: Theory and Numerical Methods) gilt für $g \in C^k[a,b]$

$$\inf_{b_j} \max_{x\in[a,b]} \Big|\sum_{j=0}^n b_j x^j - g(x)\Big| \leq C_0 n^{-k}$$

mit einer von n unabhängigen positiven Konstante C_0. Wendet man dieses Resultat auf

$$\frac{d}{dx}\Big(\frac{y(x)}{(b-x)(x-a)}\Big)$$

an, so liefert (6.74)

$$\inf_{v\in V_h} \|y-v\|_L \le C_1 n^{-k}, \tag{6.75}$$

falls

$$\frac{y(x)}{(b-x)(x-a)}$$

$k+1$ stetige Ableitungen in $[a,b]$ besitzt.

Es ist nämlich für alle $\varphi \in \Pi_{n+1}$

$$\begin{aligned}
\|y-(b-x)(x-a)\varphi(x)\|_L &\le C\|\frac{d}{dx}\Big(y(x)-(b-x)(x-a)\varphi(x)\Big)\|_\infty \\
&= C\left\|\frac{d}{dx}\Big\{(b-x)(x-a)\Big(\frac{y(x)}{(b-x)(x-a)}-\varphi(x)\Big)\Big\}\right\|_\infty \\
&= C\left\|(b+a-2x)\Big(\frac{y(x)}{(b-x)(x-a)}-\varphi(x)\Big)+(b-x)(x-a)\Big(\frac{d}{dx}\frac{y(x)}{(b-x)(x-a)}-\varphi'(x)\Big)\right\|_\infty \\
&\le C(b-a)\left\|\frac{y(x)}{(b-x)(x-a)}-\varphi(x)\right\|_\infty + \frac{1}{4}(b-a)^2 C\left\|\frac{d}{dx}\frac{y(x)}{(b-x)(x-a)}-\varphi'(x)\right\|_\infty .
\end{aligned}$$

Wählt man nun $\varphi' \in \Pi_n$ gemäß (6.75) mit

$$\left\|\frac{d}{dx}\frac{y(x)}{(b-x)(x-a)}-\varphi'(x)\right\|_\infty \le 2C_0 n^{-k},$$

$c := 0.5(a+b)$ und

$$\varphi(x) := \frac{y(c)}{4(b-a)^2} + \int_c^x \varphi'(t)\,dt,$$

so erhält man

$$\left|\frac{y(x)}{(b-x)(x-a)}-\varphi(x)\right| = \left|\int_c^x \left(\frac{d}{dx}\frac{y(x)}{(b-x)(x-a)}-\varphi'(x)\right)\,dx\right| \le (b-a)C_0 n^{-k},$$

und damit

$$\|y-(b-x)(x-a)\varphi(x)\|_L \le 2CC_0(b-a)n^{-k} + \frac{1}{4}CC_0(b-a)n^{-k} =: C_1 n^{-k}.$$

Nach (6.71) ist daher n^{-k} die Konvergenzordnung des Ritz – Galerkin Verfahrens mit Polynomansätzen in der Energienorm, und nach Satz 6.18. auch in der Maximumnorm. □

Tabelle 6.1: Fehler und Kondition zu Beispiel 6.22.

n	Fehler	Kondition
1	$1.39\ E-02$	$1.00\ E\ 00$
2	$8.71\ E-04$	$1.10\ E\ 01$
3	$4.13\ E-05$	$1.76\ E\ 02$
4	$1.72\ E-06$	$3.39\ E\ 03$
5	$6.01\ E-08$	$7.29\ E\ 04$
6	$1.88\ E-09$	$1.69\ E\ 06$
7	$4.49\ E-10$	$4.12\ E\ 07$
8	$1.24\ E-08$	$1.04\ E\ 09$
9	$3.98\ E-07$	$2.73\ E\ 10$
10	$1.44\ E-05$	$7.34\ E\ 11$
11	$5.86\ E-04$	$2.04\ E\ 13$
12	$1.00\ E-01$	$2.26\ E\ 15$

Trotz der guten Approximationseigenschaften von Polynomen sind die Ansatzfunktionen des letzten Beispiels nicht zu empfehlen, denn erstens ist die Matrix des diskretisierten Problems voll besetzt (man muss also $n(n+1)$ Integrale bestimmen zur Aufstellung des diskreten Problems mit n Ansatzfunktionen), und zweitens ist die Matrix in der Regel schlecht konditioniert.

Beispiel 6.22. Wir betrachten die Randwertaufgabe

$$-y'' = e^x,\ y(0) = 0,\ y(1) = 0$$

mit der Lösung

$$y(x) = 1 + (e-1)x - e^x.$$

Wir wenden das Ritz – Galerkin Verfahren an mit den Ansatzfunktionen

$$v_j(x) := x(1-x)x^{j-1},\ j = 1, 2, \ldots.$$

Tabelle 6.1 enthält die Fehler in der Maximumnorm und die Konditionen der Matrizen $\boldsymbol{A}_h$ für verschiedene Dimensionen n. Abbildung 6.3 enthält die Lösung und ihre Näherung für $n = 1$ und Abbildung 6.4 die Lösung und ihre Näherung für den Fall $n = 12$. □

Beispiel 6.23. Ein Funktionensystem, das bessere Eigenschaften als die Polynome in Beispiel 6.21. besitzt, erhält man auf folgende Weise. Es sei

$$a = x_0 < x_1 < \ldots < x_n = b$$

eine nicht notwendig äquidistante Zerlegung des Intervalls $[a, b]$. Dann sei V_h die Menge aller stückweise linearen Funktionen v, die in jedem der Teilintervalle $[x_{j-1}, x_j]$,

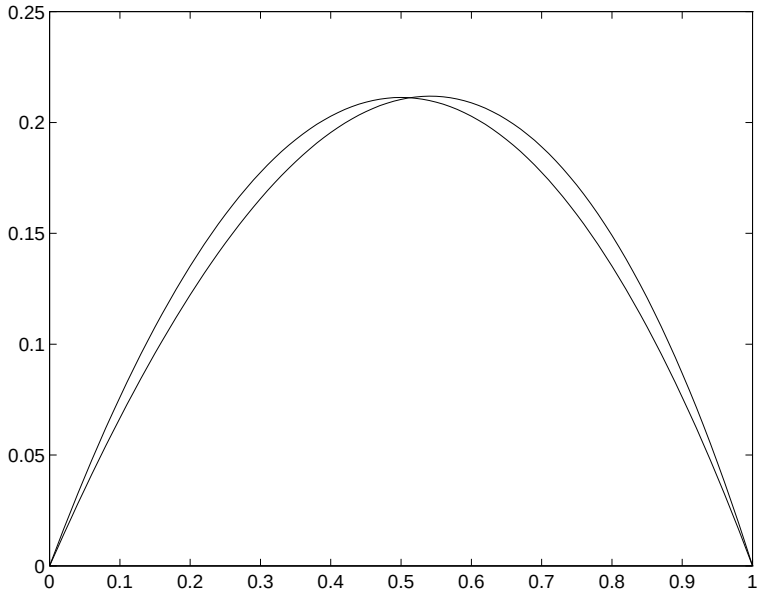

Abbildung 6.3: Ritz – Galerkin Verfahren zu Beispiel 6.22.; $n = 1$

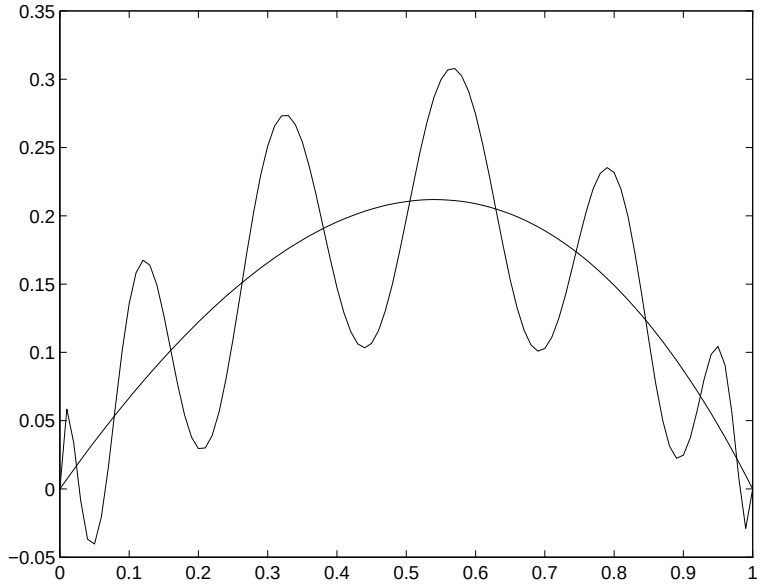

Abbildung 6.4: Ritz – Galerkin Verfahren zu Beispiel 6.22.; $n = 12$

$j = 1, \ldots, n$, linear sind und die die Randbedingungen $v(a) = 0$ und $v(b) = 0$ erfüllen.

Offenbar bilden die **Dachfunktionen** (engl.: hat functions)

$$v_j(x) := \begin{cases} \frac{1}{h_j}(x - x_{j-1}) & , \text{ für } x_{j-1} \leq x \leq x_j \\ \frac{1}{h_{j+1}}(x_{j+1} - x) & , \text{ für } x_j \leq x \leq x_{j+1} \\ 0 & , \text{ sonst} \end{cases} \quad , \quad j = 1, \ldots, n-1 \tag{6.76}$$

mit $h_j := x_j - x_{j-1}$ eine Basis von V_h (vgl. Abbildung 6.5).

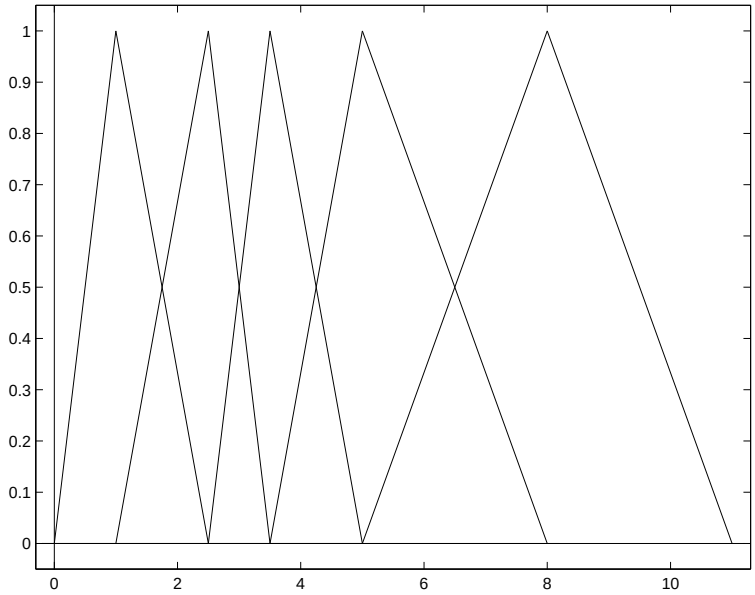

Abbildung 6.5: Dachfunktionen

Die Dachfunktion v_j hat den lokalen Träger $[x_{j-1}, x_{j+1}]$, und eine Funktion $v \in V_h$ besitzt die Darstellung

$$v(x) = \sum_{j=1}^{n-1} v(x_j) v_j(x).$$

Der Koordinatenvektor bzgl. der Basis der Dachfunktionen besteht also gerade aus den Funktionswerten von v an den Knoten.

Ein wesentlicher Vorteil der Basis der Dachfunktionen ist, dass $v_j(x)v_k(x) \equiv 0$ und $v_j'(x)v_k'(x) \equiv 0$ für alle j, k mit $|j-k| > 1$, gilt und dass damit die Steifigkeitsmatrix $\boldsymbol{A}_h$ tridiagonal ist. Bei der Aufstellung der diskreten Variationsaufgabe sind also nicht n^2 sondern nur $3n$ Integrale zu bestimmen. Zudem besteht der Träger nur aus zwei benachbarten Teilintervallen, während er bei Polynomen als Ansatzfunktionen das gesamte Intervall $[a, b]$ ist.

Ist $y \in C^2[a, b]$ die Lösung der Randwertaufgabe (6.64) und

$$v_y(x) = \sum_{j=1}^{n-1} y(x_j) v_j(x)$$

die stückweise lineare Funktion, die y in den Knoten x_j, $j = 0, \ldots, n$, interpoliert, so kann man leicht mit Hilfe des Taylorschen Satzes zeigen, dass

$$\|y - v_y\|_L = O(h), \quad h := \max_{j=1,\ldots,n} h_j,$$

gilt. Damit gilt auch

$$\inf_{v \in V_h} \|y - v\|_L = O(h).$$

Man kann zeigen, dass keine bessere Abschätzung erreichbar ist (auch nicht, wenn die Lösung y glatter ist), und daher ist das Ritz – Galerkin Verfahren mit stückweise linearen Ansatzfunktionen ein Verfahren der Ordnung 1 (auch bzgl. der Maximumnorm).

Speziell für die Randwertaufgabe

$$-y'' = f(x), a \le x \le b, \quad y(a) = 0,\ y(b) = 0, \tag{6.77}$$

erhält man für die Steifigkeitsmatrix $\boldsymbol{A}_h = (a_{jk})$

$$\begin{aligned} a_{j,j-1} &= \int_a^b v'_{j-1}(x) v'_j(x)\,dx = \int_{x_{j-1}}^{x_j} \frac{-1}{h_j} \cdot \frac{1}{h_j}\,dx = -\frac{1}{h_j}, \\ a_{jj} &= \int_{x_{j-1}}^{x_j} \frac{1}{h_j^2}\,dx + \int_{x_j}^{x_{j+1}} \frac{1}{h_{j+1}^2}\,dx = \frac{1}{h_j} + \frac{1}{h_{j+1}}, \\ a_{j,j+1} &= \int_{x_j}^{x_{j+1}} \frac{-1}{h_{j+1}} \cdot \frac{1}{h_{j+1}}\,dx = -\frac{1}{h_{j+1}}, \end{aligned}$$

und für die rechte Seite

$$f_j = \int_{x_{j-1}}^{x_{j+1}} f(x) v_j(x)\,dx = \frac{1}{h_j} \int_{x_{j-1}}^{x_j} f(x)(x - x_{j-1})\,dx + \frac{1}{h_{j+1}} \int_{x_j}^{x_{j+1}} f(x)(x_{j+1} - x)\,dx.$$

Ist die Zerlegung äquidistant $h_j = (b-a)/n$ für alle j, so erhält man als Steifigkeitsmatrix

$$\boldsymbol{A}_h = \frac{1}{h}\text{tridiag}(-1, 2, -1),$$

die wir auch schon beim Differenzenverfahren (bis auf einen Faktor h, der aber auch auf der rechten Seite auftritt) im letzten Abschnitt erhalten hatten. Die rechte Seiten stimmen jedoch i.a. nicht überein.

Dass die Konvergenzordnung nur 1 ist, scheint im Widerspruch zu dem Ergebnis in Abschnitt 6.2 zu stehen, dass das Differenzenverfahren die Konvergenzordnung 2 hat. Man beachte aber, dass Ch nur eine obere Schranke für den gleichmäßigen Fehler ist. Tatsächlich geht der Fehler an den Gitterpunkten in der Maximumnorm quadratisch gegen 0, wenn die Lösung glatt genug ist. □

Beispiel 6.24. Wendet man auf die Randwertaufgabe

$$-y'' = f(x),\ y(0) = 0,\ y(1) = 0$$

Tabelle 6.2: Kondition bei stückweise linearem Ansatz

n	Kondition
2	1.00 E 00
4	5.82 E 00
8	2.53 E 01
16	1.03 E 02
32	4.14 E 02
64	1.66 E 03
128	6.64 E 03
256	2.66 E 04
512	1.06 E 05
1024	4.25 E 05

Tabelle 6.3: Kondition und Fehler in Beispiel 6.25.

n	Kondition	Fehler in Knoten	Fehler
2	1.00 E 00	2.40 $E-03$	3.01 $E-02$
4	5.36 E 00	5.92 $E-04$	7.59 $E-03$
8	2.30 E 01	1.48 $E-04$	1.91 $E-03$
16	9.37 E 01	3.69 $E-05$	4.84 $E-04$
32	3.76 E 02	9.21 $E-06$	1.22 $E-04$
64	1.51 E 03	2.30 $E-06$	3.04 $E-05$
128	6.03 E 03	5.76 $E-07$	7.62 $E-06$
256	2.41 E 04	1.44 $E-07$	1.91 $E-06$
512	9.45 E 04	3.60 $E-08$	4.77 $E-07$

das Ritz – Galerkin Verfahren an mit stückweise linearen Ansatzfunktionen auf einem äquidistanten Gitter, so wächst die Kondition der Steifigkeitsmatrizen wesentlich langsamer als bei polynomialen Ansätzen (vgl. Tabelle 6.2).

□

Beispiel 6.25. Wir diskretisieren die Randwertaufgabe

$$-y'' + y = 1 + \frac{1}{2}x - \frac{1}{2}x^2,\ 0 < x < 1, \quad x(0) = 0,\ x(1) = 0,$$

mit dem Ritz – Galerkin Verfahren mit stückweise linearen Ansatzfunktionen. Dann erhält man als Steifigkeitsmatrix die tridiagonale Matrix

$$\boldsymbol{A}_h = \frac{1}{h}\text{tridiag}(-1, 2, -1) + h \cdot \text{tridiag}(\frac{1}{6}, \frac{2}{3}, \frac{1}{6}).$$

Tabelle 6.3 enthält die Konditionen der Steifigkeitsmatrizen, die maximalen Fehler in den Knoten und die Fehler in der Maximumnorm im Intervall $[0, 1]$. Man sieht, dass sowohl die maximalen Fehler in den Knoten als auch die Fehler in der Maximumnorm in $[0, 1]$ quadratisch gegen 0 konvergieren. □

Wir betrachten nun den Fall der allgemeinen Sturmschen Randwertaufgabe

$$Ly := -(p(x)y')' + q(x)y = f(x), \quad a < x < b, \tag{6.78}$$

$$\alpha_0 y(a) - \alpha_1 y'(a) = 0,\ \beta_0 y(b) + \beta_1 y'(b) = 0 \tag{6.79}$$

mit $p \in C^1[a,b]$, $q, f \in C[a,b]$, $p(x) > 0$ für alle $x \in [a,b]$, $q(x) \geq 0$ für alle $x \in [a,b]$, α_0, α_1, β_0, $\beta_1 \geq 0$, $(\alpha_0^2 + \alpha_1^2)(\beta_0^2 + \beta_1^2) > 0$.

Dann ist es sinnlos, im Falle $\alpha_1 \neq 0$ oder $\beta_1 \neq 0$ für $v \in H^1(a,b)$ das Erfülltsein der entsprechenden Randbedingungen zu fordern. $v'(a)$ bzw. $v'(b)$ muss ja nicht existieren (auch nicht als einseitige Ableitung).

Setzt man

$$V = \{v \in H^1(a,b)\ :\ v(a) = 0 \text{ falls } \alpha_1 = 0,\ v(b) = 0 \text{ falls } \beta_1 = 0\}, \tag{6.80}$$

so erhält man durch partielle Integration für alle $v \in V$ und die Lösung y von (6.78)

$$\begin{aligned}
\int_a^b v(x)Ly(x)\,dx &= -[p(x)y'(x)v(x)]_a^b + \int_a^b (p(x)y'(x)v'(x) + q(x)y(x)v(x))\,dx \\
&= \frac{\beta_0}{\beta_1}p(b)y(b)v(b) + \frac{\alpha_0}{\alpha_1}p(a)v(a)y(a) + \int_a^b (p(x)y'(x)v'(x) + q(x)y(x)v(x))\,dx \\
&= \int_a^b f(x)v(x)\,dx,
\end{aligned}$$

wobei der erste bzw. zweite Randterm weggelassen werden muss im Fall $\beta_1 = 0$ bzw. $\alpha_1 = 0$.

Damit ist y wieder Lösung einer Variationsaufgabe

Bestimme $y \in V$ mit

$$[y,v] := \langle y,v\rangle_L + \langle y,v\rangle_R = F(v) \quad \text{für alle } v \in V, \tag{6.81}$$

wobei V wie in (6.80) gewählt ist,

$$\langle y,v\rangle_L := \int_a^b (p(x)y'(x)v'(x) + q(x)y(x)v(x))\,dx,$$

$$\langle y,v\rangle_R := \begin{cases}
\frac{\beta_0}{\beta_1}p(b)y(b)v(b) + \frac{\alpha_0}{\alpha_1}p(a)y(a)v(a), & \text{falls } \alpha_1 \neq 0,\ \beta_1 \neq 0 \\
\frac{\beta_0}{\beta_1}p(b)y(b)v(b), & \text{falls } \alpha_1 = 0,\ \beta_1 \neq 0 \\
\frac{\alpha_0}{\alpha_1}p(a)y(a)v(a), & \text{falls } \alpha_1 \neq 0,\ \beta_1 = 0 \\
0 & \text{sonst}
\end{cases}$$

Man rechnet leicht nach, dass in jedem der Fälle $[\cdot,\cdot]$ ein inneres Produkt auf $C^1[a,b]$ mit den Randbedingungen in (6.80) ist (falls nicht $\alpha_0 = 0$, $\beta_0 = 0$ und $q(x) \equiv 0$ gilt), und dies gilt auch für den Raum V. Wie vorher ist V mit diesem inneren Produkt ein Hilbertraum, und nach dem Rieszschen Darstellungssatz ist (6.81) eindeutig lösbar.

Da die Lösung der Randwertaufgabe (6.79) die Variationsaufgabe (6.81) löst, muss die Lösung von (6.81) automatisch die Randbedingungen erfüllen, die nicht in V gefordert werden. Randbedingungen, in denen y' nicht auftritt, die also im Ansatzraum V erfüllt sein müssen, heißen **wesentliche Randbedingungen** ; Randbedingungen, in denen y' auftritt, die also nicht durch Elemente des Ansatzraumes erfüllt sein müssen, heißen **natürliche Randbedingungen** oder **restliche Randbedingungen**.

Die Approximation verläuft für diesen Typ von Aufgaben wie vorher. Man wählt einen endlich dimensionalen Teilraum V_h von V und eine Basis $\{v_1,\ldots,v_n\}$ von V_h und stellt hiermit die Ritz – Galerkin Gleichungen

$$\sum_{j=1}^{n} \xi_j [v_j, v_k] = F(v_k), \quad k = 1,\ldots,n,$$

auf. Diese sind eindeutig lösbar, und für die Näherungslösung $y = \sum_{j=1}^n \xi_j v_j$ gilt wieder die Fehlerabschätzung in der Energienorm

$$\sqrt{[y-y_h, y-y_h]} \le \inf_{v\in V_h} \sqrt{[y-v, y-v]}.$$

Die hier beschriebenen Verfahren wurden erstmals von Ritz (1908/09) und Galerkin (1915) in Spezialfällen verwendet. Dabei begründeten beide das Verfahren auf verschiedene Weisen.

Ritz beschrieb das folgende Vorgehen: Definiert man auf V (der Einfachheit halber im Fall $y(a) = y(b) = 0$) das Funktional

$$J(v) := \langle v, v\rangle_L - 2F(v),$$

so gilt für alle $v \neq y$

$$\begin{aligned} J(v) &= \langle v, v\rangle_L - 2F(v) = \langle v, v\rangle_L - 2\langle y, v\rangle_L = \langle v, v\rangle_L - 2\langle y, v\rangle_L + \langle y, y\rangle_L - \langle y, y\rangle_L \\ &= \langle v-y, v-y\rangle_L - \langle y, y\rangle_L > -\langle y, y\rangle_L = \langle y, y\rangle_L - 2F(y) = J(y). \end{aligned}$$

Das Variationsproblem

$$J(v) = \min!, \quad v \in V, \tag{6.82}$$

ist also eindeutig lösbar durch die Lösung der Randwertaufgabe (6.64).

Ersetzt man (6.82) durch das endlichdimensionale Problem

$$J(v) = \min!, \quad v \in V_h := \operatorname{span}\{v_1, \ldots, v_n\}, \tag{6.83}$$

so ist dies ein quadratisches Optimierungsproblem

$$\begin{aligned}\Phi(\xi_1, \ldots, \xi_n) &:= J(\sum_{j=1}^{n} \xi_j v_j) \\ &= \int_a^b \Big(p(x)(\sum_{j=1}^{n} \xi_j v_j'(x))^2 + q(x)(\sum_{j=1}^{n} \xi_j v_j(x))^2\Big)\, dx - 2\int_a^b f(x) \sum_{j=1}^{n} \xi_j v_j \, dx = \min!\end{aligned}$$

Notwendig für eine Lösung ist

$$\begin{aligned}0 &= \frac{\partial}{\partial \xi_k}\Phi(\xi_1, \ldots, \xi_n) \\ &= 2\int_a^b \Big(p(x) \sum_{j=1}^{n} \xi_j v_j'(x) v_k'(x) + q(x) \sum_{j=1}^{n} \xi_j v_j(x) v_k(x) - f(x) v_k(x)\Big)\, dx, \ k = 1, \ldots, n,\end{aligned}$$

d.h.

$$\sum_{j=1}^{n} \xi_j \int_a^b (p(x) v_j'(x) v_k'(x) + q(x) v_j(x) v_k(x))\, dx = \int_a^b f(x) v_k(x)\, dx, \ k = 1, \ldots, n. \tag{6.84}$$

Dies sind gerade die Ritz – Galerkin Gleichungen, die eindeutig lösbar sind. Da die Hessematrix auf $\mathbb{R}^n$ positiv definit ist, liegt tatsächlich ein Minimum vor, und dieses ist eindeutig.

Galerkin betrachtete

$$Ay := -(py')' + qy - f \tag{6.85}$$

als Abbildung von

$$D(A) := \{y \in C^2[a,b] \,:\, y(a) = y(b) = 0\} \subset L^2(a,b) \quad \text{in } L^2(a,b),$$

wobei $L^2(a,b)$ den Raum der (im Lebesgueschen Sinne) quadratisch integrierbaren Funktionen bezeichnet.

Ist $\{v_n \,:\, n \in \mathbb{N}\} \subset D(A)$ ein vollständiges Orthonormalsystem von $L^2(a,b)$, so ist y genau dann Lösung von $A(y) = 0$, wenn

$$\int_a^b Ay(x) v_n(x)\, dx = 0 \quad \text{für alle } n \in \mathbb{N} \tag{6.86}$$

gilt. Eine Näherungslösung v_h erhält man, wenn man sich auf eine Teilmenge $\{v_1, \ldots, v_n\}$ beschränkt, also Linearkombinationen der v_j für v_h betrachtet. Dann ist das System (6.86) überbestimmt. Beschränkt man sich auf

$$\int_a^b Ay(x)v_j(x)\,dx = 0 \quad \text{für alle } j \in \{1, \ldots, n\} \tag{6.87}$$

so erhält man die Ritz – Galerkin Gleichungen.

Beide Zugänge sind Anlass für Verallgemeinerungen.

Man kann nicht–quadratische Funktionale J in (6.82) zulassen und erhält auf dieselbe Weise wie oben Diskretisierungsmethoden für nichtlineare Differentialgleichungen (Euler – Lagrange Gleichungen), oder man kann Funktionale betrachten, die auch von y'' abhängen, und erhält so Verfahren für Differentialgleichungen vierter Ordnung.

Den Zugang von Galerkin kann man auf jede Gleichung im Hilbertraum, in dem eine Orthonormalbasis existiert, anwenden, man kann die Orthogonalität in (6.86) dadurch ersetzen, dass man fordert, dass die Projektion auf einen n–dimensionalen Teilraum verschwindet. Diese Formulierung (Projektionsverfahren) ist dann auch sinnvoll im Banachraum.

Kapitel 7

Differenzenverfahren für elliptische Randwertaufgaben

Bei der numerischen Behandlung elliptischer Randwertaufgaben mit Differenzenverfahren berechnet man wie bei den gewöhnlichen Randwertaufgaben Näherungen für die Lösung auf einem rechteckigen Gitter, wobei die auftretenden Ableitungen durch Differenzenquotienten ersetzt werden. Wir betrachten nur ebene Probleme. Die Übertragung auf räumliche Probleme oder Probleme höherer Dimension ist offensichtlich.

7.1 Das Modellproblem

Wir betrachten zunächst das Modellproblem

$$-\Delta u(x,y) \;=\; f(x,y) \quad \text{für } (x,y) \in \Omega := (0,1)\times(0,1) \tag{7.1}$$

$$u(x,y) \;=\; g(x,y) \quad \text{für } (x,y) \in \partial\Omega \tag{7.2}$$

wobei $f : \bar{\Omega} \to \mathbb{R}$ und $g : \partial\Omega \to \mathbb{R}$ gegebene Funktionen sind.

Nach dem Taylorschen Satz gilt für $u \in C^4$:

$$\begin{aligned} u(x+h,y) &= u(x,y) + u_x(x,y)h + \frac{1}{2}u_{xx}(x,y)h^2 + \frac{1}{6}u_{xxx}(x,y)h^3 + O(h^4) \\ u(x-h,y) &= u(x,y) - u_x(x,y)h + \frac{1}{2}u_{xx}(x,y)h^2 - \frac{1}{6}u_{xxx}(x,y)h^3 + O(h^4) \end{aligned}$$

Daher ist

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

und genauso

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

und es folgt

$$-\Delta u(x,y) = \frac{4u(x,y) - u(x,y-h) - u(x-h,y) - u(x+h,y) - u(x,y+h)}{h^2} + O(h^2) \tag{7.3}$$

falls $\{(\xi,\eta) \,:\, x-h \le \xi \le x+h, y-h \le \eta \le y+h\} \subset \Omega$.

Für $h = \frac{1}{n}$ ersetzen wir $\Omega = (0,1) \times (0,1)$ durch das Gitter

$$\Omega_h := \{(ih, jh) \,:\, i,j \in \mathbb{N},\ (ih,jh) \in \Omega\}$$

und den Rand $\partial\Omega$ durch das Randgitter

$$\partial\Omega_h := \{(ih, jh) \,:\, i,j \in \mathbb{N}_0,\ (ih,jh) \in \partial\Omega\}.$$

Dann erfordert die Auswertung des diskreten Laplace Operators

$$\Delta_h u(x,y) := \frac{1}{h^2}(u(x,y-h) + u(x-h,y) + u(x+h,y) + u(x,y+h) - 4u(x,y)) \tag{7.4}$$

für Punkte $(x,y) \in \Omega_h$ nur die Kenntnis von u in Punkten aus

$$\bar{\Omega}_h := \Omega_h \cup \partial\Omega_h.$$

Wir bestimmen daher Näherungen $U_{ij} \approx u(ih,jh)$, $i,j = 1,\ldots,n-1$, aus dem Gleichungssystem

$$\frac{1}{h^2}(4U_{ij} - U_{i,j-1} - U_{i-1,j} - U_{i+1,j} - U_{i,j+1}) = f(ih,jh),\ i,j = 1,\ldots,n-1,$$

wobei wegen der Randbedingung

$$U_{0j} = g(0,jh),\ U_{j0} = g(jh,0),\ U_{nj} = g(1,jh),\ U_{jn} = g(jh,1)$$

gesetzt wird.

Wählt man die lexikographische Anordnung der Variablen

$$U_h := (U_{11}, U_{12}, \ldots, U_{1,n-1}, U_{21}, U_{22}, \ldots, U_{n-1,n-2}, U_{n-1,n-1})^T,$$

so erhält man ein lineares Gleichungssystem

$$\boldsymbol{A}_h \boldsymbol{U}_h = \boldsymbol{f}_h,$$

mit der Koeffizientenmatrix

$$\boldsymbol{A}_h = \frac{1}{h^2}\begin{pmatrix} \boldsymbol{B} & -\boldsymbol{I} & \boldsymbol{O} & \dots & \boldsymbol{O} & \boldsymbol{O} \\ -\boldsymbol{I} & \boldsymbol{B} & -\boldsymbol{I} & \dots & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & -\boldsymbol{I} & \boldsymbol{B} & \dots & \boldsymbol{O} & \boldsymbol{O} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \dots & -\boldsymbol{I} & \boldsymbol{B} \end{pmatrix}$$

wobei $\boldsymbol{I}$ die Einheitsmatrix der Dimension $n-1$ ist und

$$\boldsymbol{B} = \begin{pmatrix} 4 & -1 & 0 & \dots & 0 & 0 \\ -1 & 4 & -1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & -1 & 4 \end{pmatrix} \in \mathbb{R}^{(n-1,n-1)}.$$

Um die Konvergenz zu untersuchen, betrachten wir wieder die Restriktion der Funktion u auf das Gitter

$$R_h u := (u(h,h), u(h,2h), \dots, u(h,(n-1)h), u(2h,h), \dots, u((n-1)h,(n-1)h))^T,$$

die dieselbe Dimension hat wie die Lösung U_h des diskreten Problems zur Schrittweite h.

Definition 7.1. *Das Differenzenverfahren heißt* **konvergent**, *wenn*

$$\lim_{h\to 0} \|U_h - R_h u\| = 0$$

gilt. Existiert ein $C > 0$ und ein $p > 0$ mit

$$\|U_h - R_h u\| \le C h^p,$$

so konvergiert das Verfahren mindestens **von der Ordnung** p.

Die Konvergenz zeigt man wie für gewöhnliche Randwertaufgaben, indem man die Konsistenz und Stabilität nachweist.

Definition 7.2. *Die Diskretisierung heißt* **stabil** *(bzgl. der Maximumnorm), wenn es eine Konstante $C > 0$ gibt mit*

$$\|\boldsymbol{A}_h^{-1}\|_\infty \le C \quad \textit{für alle Schrittweiten } h. \tag{7.5}$$

Sie heißt **konsistent** *von der Ordnung k (bzgl. der Maximumnorm), wenn mit einer Konstante K gilt*

$$\|\Delta_h R_h u - R_h \Delta u\|_\infty \le K h^k. \tag{7.6}$$

Ist $u \in C^4(\bar{\Omega})$, so ist die Diskretisierung nach dem Taylorschen Satz konsistent von der Ordnung 2, denn

$$\begin{aligned}
&(\Delta_h R_h u - R_h \Delta u)(x,y) \\
&= \frac{1}{h^2}(4u(x,y) - u(x+h,y) - u(x-h,y) - u(x,y-h) - u(x,y+h)) \\
&\quad -u_{xx}(x,y) - u_{yy}(x,y) \\
&= \frac{1}{12}h^2(u_{xxxx}(\xi,y) + u_{yyyy}(x,\eta))
\end{aligned}$$

mit $\xi \in (x-h, x+h)$ und $\eta \in (y-h, y+h)$, und damit gilt

$$\|\Delta_h R_h u - R_h \Delta u\|_\infty \leq \frac{1}{12}h^2 \|u\|_{C^4(\bar{\Omega})}. \tag{7.7}$$

Die Stabilität erhält man wieder aus der Inversmonotonie der Matrix $\boldsymbol{A}_h$.

Lemma 7.3. *Die Matrix $\boldsymbol{A}_h$ ist invers monoton, d.h.*

$$\boldsymbol{A}_h \boldsymbol{U}_h \geq \boldsymbol{0} \quad \Longrightarrow \quad \boldsymbol{U}_h \geq \boldsymbol{0}.$$

Beweis: Wir führen den Beweis indirekt. Wir nehmen an, dass es ein $\boldsymbol{U}_h$ gibt mit $\boldsymbol{A}_h \boldsymbol{U}_h \geq \boldsymbol{0}$ und $\boldsymbol{U}_h \not\geq \boldsymbol{0}$.

Es sei $(i,j) \in \{1,\ldots,n-1\}^2$ mit

$$\boldsymbol{U}_h(ih,jh) = \min_{(x,y)\in\Omega_h} U_h(x,y) < 0.$$

Dann folgt aus $(\boldsymbol{A}_h \boldsymbol{U}_h)_{ij} \geq 0$

$$\boldsymbol{U}_h(ih,jh) \geq \frac{1}{4}\Big(\boldsymbol{U}_h((i-1)h,jh) + \boldsymbol{U}_h((i+1)h,jh) + \boldsymbol{U}_h(ih,(j-1)h) + \boldsymbol{U}_h(ih,(j+1)h)\Big),$$

und dies ist nur möglich im Fall

$$\boldsymbol{U}_h((i-1)h,jh) = \boldsymbol{U}_h((i+1)h,jh) = \boldsymbol{U}_h(ih,(j-1)h) = \boldsymbol{U}_h(ih,(j+1)h) = \boldsymbol{U}_h(ih,jh).$$

Wiederholt man diese Argumentation mit den Nachbarn von $\boldsymbol{U}_h(ih,jh)$, so erhält man, dass $\boldsymbol{U}_h$ konstant auf Ω_h ist, und die erste Gleichung des Systems liefert den Widerspruch

$$4U_h(h,h) - U_h(2h,h) - U_h(h,2h) = 2U_h(h,h) \geq 0.$$

∎

Tabelle 7.1: Differenzenverfahren

h	$\|R_h u - U_h\|_\infty$
1/2	$3.77E-3$
1/4	$1.06E-3$
1/8	$2.76E-4$
1/16	$7.07E-5$
1/32	$1.77E-5$
1/64	$4.44E-6$

Satz 7.4. *Das Differenzenverfahren ist stabil. Es gilt*

$$\|\boldsymbol{A}_h^{-1}\|_\infty \leq \frac{1}{8}.$$

Beweis: Es sei $w(x,y) = 0.5(x - x^2)$ und $\boldsymbol{w}_h := R_h w$. Dann gilt für alle Gitterpunkte (ih, jh), für die auch die vier Nachbarn noch in Ω_h liegen, $(\boldsymbol{A}_h \boldsymbol{w}_h)_{ij} = 1$, und für die übrigen Gitterpunkte ist $(\boldsymbol{A}_h \boldsymbol{w}_h)_{ij} \geq 1$. Bezeichnet $\boldsymbol{v}_h$ die Lösung von $\boldsymbol{A}_h \boldsymbol{v}_h = \boldsymbol{e}$, so folgt $\boldsymbol{w}_h \geq \boldsymbol{v}_h$ wegen der Inversmonotonie von $\boldsymbol{A}_h$, und daher

$$\|\boldsymbol{A}_h^{-1}\|_\infty = \|\boldsymbol{v}_h\|_\infty \leq \|\boldsymbol{w}_h\|_\infty = \frac{1}{8}.$$

∎

Satz 7.5. *Ist die Lösung u von (7.1) viermal stetig differenzierbar in $\bar{\Omega}$, so konvergiert das Differenzenverfahren von der Ordnung 2.*

Beweis: Es gilt

$$\begin{aligned} \|R_h u - \boldsymbol{U}_h\|_\infty &= \|\boldsymbol{A}_h^{-1}(\boldsymbol{A}_h R_h u - \boldsymbol{A}_h \boldsymbol{U}_h)\|_\infty \leq \|\boldsymbol{A}_h^{-1}\|_\infty \|\Delta_h R_h u - R_h f\|_\infty \\ &\leq \frac{1}{8}\|\Delta_h R_h u - R_h \Delta u\|_\infty \leq \frac{1}{96} h^2 \|u\|_{C^4(\bar{\Omega})}. \end{aligned}$$

∎

Beispiel 7.6. Wir betrachten das Poisson Problem

$$\Delta u = 0 \quad \text{in } \Omega := (0,1) \times (0,1), \tag{7.8}$$

$$u(x,y) = e^x \cos y \quad \text{auf } \partial\Omega \tag{7.9}$$

mit der Lösung

$$u(x,y) = e^x \cos y,$$

die in $\bar{\Omega}$ beliebig oft differenzierbar ist.

Mit dem Differenzenverfahren erhält man die Fehler in Tabelle 7.1. Der Fehler wird offenbar geviertelt, wenn die Schrittweite halbiert wird. Dies demonstriert die Konvergenzordnung 2. □

7.2 Die Neumannsche Randwertaufgabe

Wir betrachten nun die Neumannsche Randwertaufgabe im Quadrat.

$$-\Delta u(x,y) \;=\; f(x,y) \quad \text{für } (x,y) \in \Omega := (0,1)\times(0,1) \tag{7.10}$$

$$\frac{\partial}{\partial n}u(x,y) \;=\; g(x,y) \quad \text{für } (x,y) \in \partial\Omega \tag{7.11}$$

wobei $f : \bar{\Omega} \to \mathbb{R}$ und $g : \partial\Omega \to \mathbb{R}$ gegebene Funktionen sind und n der äußere Normalenvektor auf dem Rand $\partial\Omega$ ist.

Für $\Omega = (0,1)\times(0,1)$ ist

$$\begin{aligned} \frac{\partial}{\partial n}u(x,0) = -u_y(x,0) \quad &, \quad \frac{\partial}{\partial n}u(x,1) = u_y(x,1) \\ \frac{\partial}{\partial n}u(0,y) = -u_x(0,y) \quad &, \quad \frac{\partial}{\partial n}u(1,y) = u_x(1,y). \end{aligned}$$

Es liegt nahe, die Normalableitungen durch einseitige Differenzen zu approximieren:

$$\begin{aligned} -u_y(jh,0) &\approx h^{-1}(U_{j0} - U_{j1}) = g(jh,0),\ j = 1,\dots,n-1 \\ u_y(jh,1) &\approx h^{-1}(U_{jn} - U_{j,(n-1)}) = g(jh,1),\ j = 1,\dots,n-1 \\ -u_x(0,jh) &\approx h^{-1}(U_{0j} - U_{1j}) = g(0,jh),\ j = 1,\dots,n-1 \\ u_x(1,jh) &\approx h^{-1}(U_{nj} - U_{(n-1),j}) = g(1,jh),\ j = 1,\dots,n-1. \end{aligned}$$

Man erhält (zusammen mit der Diskretisierung der Differentialgleichung) ein lineares Gleichungssystem von $(n+1)^2 - 4$ Gleichungen in den $(n+1)^2 - 4$ Unbekannten U_{ij}, $i,j = 0,1,\dots,n$,

$$(i,j) \notin \{(0,0),(0,n),(n,0),(n,n)\}.$$

Eliminiert man mit den Diskretisierungen der Randbedingungen die Funktionswerte in den Randknoten, so erhält man z.B. für die erste Gleichung

$$\begin{aligned} &\frac{1}{h^2}(4U_{11} - U_{01} - U_{10} - U_{12} - U_{21}) \\ &\quad = \frac{1}{h^2}(4U_{11} - (U_{11} + hg(0,h)) - (U_{11} + hg(h,0)) - U_{12} - U_{21}) = f(h,h), \end{aligned}$$

d.h.

$$\frac{1}{h^2}(2U_{11} - U_{12} - U_{21}) = f(h,h) + h^{-1}(g(h,0) + g(0,h)),$$

genauso für die nächste

$$\frac{1}{h^2}(3U_{12} - U_{11} - U_{13} - U_{22}) = f(h,2h) + h^{-1}g(0,2h),$$

u.s.w.

Insgesamt erhält man (bei lexikographischer Anordnung) ein lineares Gleichungssystem

$$\tilde{\boldsymbol{A}}_h \boldsymbol{U}_h = \tilde{\boldsymbol{f}}_h \tag{7.12}$$

mit

$$\tilde{\boldsymbol{A}}_h = \frac{1}{h^2} \begin{pmatrix} \tilde{\boldsymbol{B}} - \boldsymbol{I} & -\boldsymbol{I} & \boldsymbol{O} & \dots & \boldsymbol{O} & \boldsymbol{O} \\ -\boldsymbol{I} & \tilde{\boldsymbol{B}} & -\boldsymbol{I} & \dots & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & -\boldsymbol{I} & \tilde{\boldsymbol{B}} & \dots & \boldsymbol{O} & \boldsymbol{O} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \dots & \tilde{\boldsymbol{B}} & -\boldsymbol{I} \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \dots & -\boldsymbol{I} & \tilde{\boldsymbol{B}} - \boldsymbol{I} \end{pmatrix},$$

wobei $\boldsymbol{I}$ die Einheitsmatrix der Dimension $n-1$ ist und

$$\tilde{\boldsymbol{B}} = \begin{pmatrix} 3 & -1 & 0 & \dots & 0 & 0 \\ -1 & 4 & -1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 4 & -1 \\ 0 & 0 & 0 & \dots & -1 & 3 \end{pmatrix}$$

Die Matrix $\tilde{\boldsymbol{A}}_h$ ist symmetrisch und positiv semidefinit und singulär, denn für den Vektor $\boldsymbol{v}_h := (1, 1, \dots, 1)^T$ gilt

$$\tilde{\boldsymbol{A}}_h \boldsymbol{v}_h = 0.$$

Dies ist auch eine Basis des Nullraumes, denn ist $\boldsymbol{w}_h$ eine beliebige Lösung des homogenen Problems $\tilde{\boldsymbol{A}}_h \boldsymbol{w}_h = \boldsymbol{0}$ und nimmt $\boldsymbol{w}_h$ sein Maximum $w_{\max}$ in $(ih, jh) \in \Omega_h$ an, so hat $\boldsymbol{w}_h$ diesen Wert auch in allen Nachbarpunkten in Ω_h, denn da die Diagonalelemente von $\tilde{\boldsymbol{A}}_h$ positiv sind und die Nichtdiagonalelemente nicht positiv und da die Zeilensummen 0 sind, ist jede Komponente Mittelwert der benachbarten Komponenten. Wiederholt man diesen Schluss, so erhält man, dass der Vektor $\boldsymbol{w}_h$ konstant ist.

Das Gleichungssystem (7.12) besitzt also (wie das kontinuierliche Neumannsche Problem) entweder keine Lösungen oder, wenn es eine Lösung besitzt, so ist diese nur bis auf eine additive Konstante bestimmt.

(7.12) ist genau dann lösbar, wenn für die rechte Seite $\tilde{\boldsymbol{f}}_h$

$$\tilde{\boldsymbol{f}}_h^T \boldsymbol{w}_h = 0 \text{ für alle } \boldsymbol{w}_h \text{ mit } \tilde{\boldsymbol{A}}_h^T \boldsymbol{w}_h = \boldsymbol{0}$$

gilt, und wegen $\tilde{\boldsymbol{A}}_h^T = \tilde{\boldsymbol{A}}_h$ ist dies äquivalent zu

$$\sum_{(x,y) \in \Omega_h} \tilde{\boldsymbol{f}}_h(x, y) = 0,$$

d.h. zu

$$-h^2 \sum_{(x,y)\in\Omega_h} f(x,y) = h \sum_{(x,y)\in\partial\tilde{\Omega}_h} g(x,y)$$

mit

$$\partial\tilde{\Omega}_h = \partial\Omega_h \setminus \{(0,0),(0,1),(1,0),(1,1)\}.$$

Diese Bedingung kann man als diskrete Version der Lösbarkeitsbedingung

$$-\int_\Omega f(x,y)\,d(x,y) = \int_{\partial\Omega} g(x,y)\,ds$$

für das kontinuierliche Problem auffassen.

Wir fassen zusammen.

Lemma 7.7. *Das Gleichungssystem (7.12) ist genau dann lösbar, wenn*

$$-h^2 \sum_{(x,y)\in\Omega_h} f(x,y) = h \sum_{(x,y)\in\partial\tilde{\Omega}_h} g(x,y). \tag{7.13}$$

Sind $\boldsymbol{U}_h^1$ und $\boldsymbol{U}_h^2$ zwei Lösungen von (7.12), so gibt es eine Konstante c mit $\boldsymbol{U}_h^1 - \boldsymbol{U}_h^2 = c \cdot \boldsymbol{v}_h$.

Ist (7.12) lösbar, so kann man am einfachsten die Lösung auf folgende Weise bestimmen. Für einen festen Punkt $(x_0, y_0) \in \Omega$ normiere man die Lösung durch die zusätzliche Bedingung

$$u(x_0, y_0) = 0.$$

Dann ist das System (7.12) äquivalent dem System

$$\hat{\boldsymbol{A}}_h \hat{\boldsymbol{U}}_h = \hat{\boldsymbol{f}}_h, \tag{7.14}$$

wobei $\hat{\boldsymbol{A}}_h$ aus $\tilde{\boldsymbol{A}}_h$ dadurch entsteht, dass man die zu (x_0, y_0) gehörende Zeile und Spalte streicht, und $\hat{\boldsymbol{f}}_h$ aus $\tilde{\boldsymbol{f}}_h$ dadurch, dass man die zu (x_0, y_0) gehörende Komponente streicht.

$\hat{\boldsymbol{A}}_h$ ist dann regulär, denn ist $\hat{\boldsymbol{w}}_h$ ein Element des Nullraumes von $\hat{\boldsymbol{A}}_h$, und ergänzt man $\hat{\boldsymbol{w}}_h$ durch eine Null in der zu (x_0, y_0) gehörenden Komponente, so erhält man offenbar einen Vektor $\boldsymbol{w}_h$, der im Nullraum von $\tilde{\boldsymbol{A}}_h$ liegt. Dieser wird jedoch durch den Vektor $\boldsymbol{v}_h = (1, 1, \ldots, 1)^T$ aufgespannt, und daher folgt $\boldsymbol{w}_h = \boldsymbol{0}$.

Beachten Sie, dass das System (7.14) stets lösbar ist, auch wenn (7.12) keine Lösung besitzt. Es muss also die Lösbarkeitsbedingung (7.13) überprüft werden. In der PDE Toolbox zu MATLAB wird die Lösbarkeit (des mit finiten Elementen diskretisierten

Problems) nicht überprüft, und es wird in jedem Fall eine Lösung ausgegeben ohne jede Warnung.

Ein anderes Verfahren zur Lösung des Systems (7.12) erhält man, indem man die Dimension von (7.12) nicht reduziert, sondern das System erweitert.

Es sei weiterhin $\boldsymbol{v}_h := (1, 1, \ldots, 1)^T$. Wir betrachten für festes $\sigma \in \mathbb{R}$ das Gleichungssystem

$$\begin{pmatrix} \tilde{\boldsymbol{A}}_h & \boldsymbol{v}_h \\ \boldsymbol{v}_h^T & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{U}_h \\ \lambda \end{pmatrix} = \begin{pmatrix} \tilde{\boldsymbol{f}}_h \\ \sigma \end{pmatrix}. \tag{7.15}$$

Satz 7.8. *Das Gleichungssystem (7.15) ist stets lösbar.*

Ist $\lambda = 0$, so ist die Lösbarkeitsbedingung für das System (7.12) erfüllt, und $\boldsymbol{U}_h$ ist die durch $\boldsymbol{U}_h^T \boldsymbol{v}_h = \sigma$ normierte Lösung.

Ist $\lambda \neq 0$, so ist das Gleichungssystem $\tilde{\boldsymbol{A}}_h \boldsymbol{U}_h = \tilde{\boldsymbol{f}}_h$ nicht lösbar.

Beweis: Wir haben bereits gesehen, dass $\boldsymbol{v}_h$ orthogonal zu den Spalten von $\tilde{A}_h$ ist. Daher gilt

$$\mathrm{Rang}(\tilde{\boldsymbol{A}}_h, \boldsymbol{v}_h) = (n-1)^2.$$

Ferner ist $(\boldsymbol{v}_h^T, 0)$ linear unabhängig von den Zeilen der Matrix $(\tilde{\boldsymbol{A}}_h, \boldsymbol{v}_h)$, und daher besitzt die Systemmatrix von (7.15) den vollen Rang $(n-1)^2 + 1$.

Die weiteren Behauptungen liest man unmittelbar aus dem System (7.15) ab. ∎

Für die Diskretisierung (7.12) des Neumannschen Problems gilt der folgende Konvergenzsatz:

Satz 7.9. *Die Randwertaufgabe*

$$-\Delta u = f \text{ in } \Omega, \quad \frac{\partial}{\partial n} u = g \text{ auf } \partial\Omega$$

sei lösbar, und es sei $u \in C^4(\bar{\Omega})$ eine Lösung.

Es sei $\begin{pmatrix} \boldsymbol{U}_h \\ \lambda_h \end{pmatrix}$ die Lösung des diskreten Problems (7.15).

Dann gibt es ein $c \in \mathbb{R}$ und von u und h unabhängige Konstanten C und C', so dass gilt

$$\|R_h(u) - \boldsymbol{U}_h - c\boldsymbol{v}_h\| \leq C(h\|u\|_{C^2(\bar{\Omega})} + h^2\|u\|_{C^4(\bar{\Omega})}) \tag{7.16}$$

und

$$|\lambda_h| \leq C' h \|u\|_{C^1(\bar{\Omega})}. \tag{7.17}$$

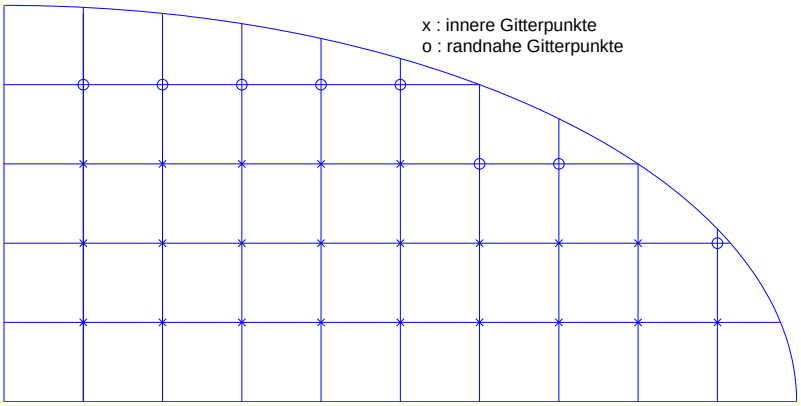

Abbildung 7.1: Innere und randhahe Gitterpunkte

Beweis: s. *Hackbusch* [27] p. 71. ∎

Bemerkung 7.10. Unter Verwendung von zentralen Differenzen für die Randbedingungen kann man auch für die Neumannsche Randwertaufgabe die Konvergenzordnung 2 erreichen (vgl. *Hackbusch* [27] p. 73).

7.3 Die Poisson Gleichung in allgemeinen Gebieten

Sei jetzt $\Omega \subset \mathbb{R}^2$ ein beliebiges Gebiet, (x_0, y_0) ein fester Punkt in $\mathbb{R}^2$ und für $h > 0$

$$G_h := \{(x_0 + jh, y_0 + kh) \ : \ j, k \in \mathbb{Z}\}$$

ein Gitter in $\mathbb{R}^2$.

Abweichend von den bisherigen Bezeichnungen zerlegen wir die Menge der Gitterpunkte in Ω in die Menge Ω_h der **inneren Gitterpunkte** (x, y), für die alle vier Nachbarn $(x-h, y)$, $(x+h, y)$, $(x, y-h)$ und $(x, y+h)$ und die Verbindungsgeraden zu diesen Nachbarpunkten in $\bar{\Omega}$ liegen, und mit

$$\partial\Omega_h := (\Omega \cap G_h) \setminus \Omega_h$$

die Menge der **randnahen Gitterpunkte**.

In den inneren Gitterpunkten diskretisieren wir die Differentialgleichung wie vorher.

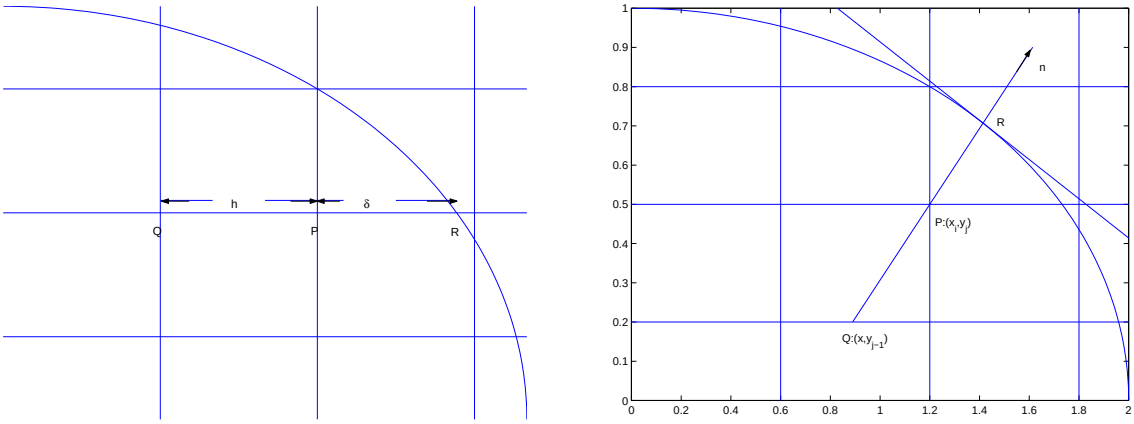

Abbildung 7.2: Differenzengleichung in randnahen Gitterpunkten

Ist $P \in \partial\Omega_h$, so gibt es einen Randpunkt $R \in \partial\Omega$, der auf einer Gitterlinie mit P liegt und dessen Abstand δ von P kleiner als h ist.

Wir nehmen an, dass man R so auswählen kann, dass der in Richtung $\vec{RP}$ gelegene Nachbar Q von P ein innerer Gitterpunkt ist (dies kann man meistens durch genügend kleine Wahl von h erreichen).

Dann wählen wir als Differenzengleichung in dem Punkt P

$$\boldsymbol{A}_h \boldsymbol{U}_h(P) := U(P) - \frac{\delta}{\delta + h} U(Q) = \frac{h}{\delta + h} g(R),$$

d.h. wir bestimmen $U(P)$ durch lineare Interpolation der Werte $U(Q)$ und $U(R) = g(R)$.

Zusammen haben wir damit für jeden Gitterpunkt (x, y) in Ω eine lineare Gleichung, insgesamt also ein lineares Gleichungssystem

$$\hat{\boldsymbol{A}}_h \boldsymbol{U}_h = \hat{\boldsymbol{f}}_h$$

zur Bestimmung von Näherungen $U(x, y) \approx u(x, y)$ in allen Gitterpunkten in Ω.

Ist $u \in C^4(\bar{\Omega})$, so geht der lokale Fehler der Differenzenapproximationen in inneren Gitterpunkten und der Interpolationen in randnahen Gitterpunkten quadratisch gegen 0. Wie in Abschnitt 7.1 kann man auch für dieses Problem eine Stabilitätsungleichung nachweisen (was hier aber wesentlich mehr Technik erfordert). Daraus erhält man (vgl. *Hackbusch* [27], p. 83)

Satz 7.11. *Die Dirichletsche Randwertaufgabe*

$$-\Delta u = f \text{ in } \Omega, \quad u = g \text{ auf } \partial\Omega$$

Tabelle 7.2: L–Bereich

h	$\|R_h u - U_h\|_\infty$
1/2	$3.62E-2$
1/4	$2.96E-2$
1/8	$2.01E-2$
1/16	$1.30E-2$
1/32	$8.30E-3$
1/64	$5.25E-3$

besitze eine Lösung $u \in C^4(\bar{\Omega})$, und es sei $h > 0$ so klein, dass die oben beschriebene Interpolation in randnahen Gitterpunkten möglich ist. Dann gibt es eine (von u und h unabhängige) Konstante C, so dass

$$\|U_h - R_h u\| \leq h^2 \|u\|_{C^2(\bar{\Omega})} + Ch^2 \|u\|_{C^4(\bar{\Omega})}$$

gilt.

Beispiel 7.12. Die Voraussetzung $u \in C^4(\bar{\Omega})$ ist wesentlich. Die Differentialgleichung

$$\Delta u = 0 \quad \text{in } \Omega := (-1,1) \times (-1,1) \setminus (-1,0] \times (-1,0] \tag{7.18}$$

mit den Randvorgaben (in Polarkoordinaten)

$$u(r,\varphi) = r^{2/3} \sin \frac{2\varphi + \pi}{3}$$

besitzt die Lösung

$$u(r,\varphi) = r^{2/3} \sin \frac{2\varphi + \pi}{3}.$$

Die ersten Ableitungen wachsen bei Annäherung an den Punkt $(0,0)$ über alle Grenzen, so dass nicht einmal $u \in C^1(\bar{\Omega})$ gilt.

Mit dem Differenzenverfahren erhält man die Fehler in Tabelle 7.2. Man liest ab, dass die Konvergenz wesentlich langsamer als quadratisch ist. □

Bemerkung 7.13. Eine andere Diskretisierung, die ebenfalls zu einem Verfahren der Ordnung 2 führt, bei dem die Fehlerkonstante allerdings kleiner ist als bei der Interpolation, ist die **Shortley–Weller Approximation**. Bei ihr wird die Interpolation in randnahen Gitterpunkten durch eine Differenzenapproximation ähnlich wie in inneren Gitterpunkten ersetzt.

Wir betrachten nun noch den Fall der Neumannschen oder der dritten Randbedingung

$$\alpha(x,y)u(x,y) + \beta(x,y)\frac{\partial}{\partial n}u(x,y) = g(x,y), \quad (x,y) \in \partial\Omega. \tag{7.19}$$

Es sei $P \in \partial\Omega_h$ ein randnaher Gitterpunkt. Wir fällen von P das Lot auf die Randkurve $\partial\Omega$. Dieses schneide im Punkt R den Rand und im Punkt Q das Gitterquadrat, in dem P liegt (vgl. Abbildung 7.2 rechts).

Gilt $P = (x_i, y_j)$ und $Q = (x, y_{j-1})$, so erhält man aus dem Taylorschen Satz

$$\frac{\partial}{\partial n}u(x_i,y_i) = \frac{u(x_i,y_j) - u(x,y_{j-1})}{\sqrt{h^2 + (x_i - x)^2}} + O(h) \ ,$$

sowie durch lineare Interpolation

$$u(x,y_{j-1}) = \frac{x - x_{i-1}}{h}u(x_{i-1},y_{j-1}) + \frac{x_i - x}{h}u(x_i,y_{j-1}) + O(h^2) \ .$$

Hiermit kann man die Randbedingung diskretisieren:

$$\begin{aligned}
0 &= \alpha(R)u(R) + \beta(R)\frac{\partial}{\partial n}u(R) - g(R) \\
&= \alpha(P)u(P) + \beta(P)\frac{\partial}{\partial n}u(P) - g(P) + O(h) \\
&= \alpha(P)u(P) - g(P) \\
&\quad + \frac{\beta(P)}{\sqrt{h^2 + (x_i - x)^2}}\left\{u(P) - \frac{x - x_{i-1}}{h}u(x_{i-1},y_{j-1}) - \frac{x_i - x}{h}u(x_i,y_{j-1})\right\} + O(h) \ .
\end{aligned}$$

Läßt man die Restglieder $O(h)$ fort, so erhält man die Differenzengleichung

$$0 = \alpha(x_i,y_j)U_{ij} - g_{ij} + \frac{\beta(x_i,y_j)}{\sqrt{h^2 + (x_i - x)^2}}\left(U_{ij} - \frac{x - x_{i-1}}{h}U_{i-1,j-1} - \frac{x_i - x}{h}U_{i,j-1}\right) \ . \tag{7.20}$$

Zusammen mit den Gleichungen für die inneren Gitterpunkte erhält man also wieder ein lineares Gleichungssystem zur Bestimmung von Näherungen für die Lösung in den Gitterpunkten.

Bemerkung 7.14. Wie für gewöhnliche Differentialgleichungen wurden auch für partielle Randwertaufgaben Mehrstellenformeln entwickelt, um höhere Konvergenzordnungen zu erreichen. Es gilt z.B. für $u \in C^6(\bar{\Omega})$

$$\begin{aligned}
&\frac{1}{h^2}\big(40u(x,y) - 8(u(x-h,y) + u(x+h,y) + u(x,y-h) + u(x,y+h)) \\
&\qquad -2(u(x-h,y-h) + u(x-h,y+h) + u(x+h,y-h) + u(x+h,y+h))\big) \\
&= \Delta u(x-h,y) + \Delta u(x+h,y) + \Delta u(x,y-h) + \Delta u(x,y+h) - 8\Delta u(x,y) + O(h^4)
\end{aligned}$$

Tabelle 7.3: Differenzenverfahren und Mehrstellenformel

h	5-Punkt-Diskretisierung	Mehrstellenformel
1/2	$3.77E-3$	$5.61E-07$
1/4	$1.06E-3$	$8.61E-09$
1/8	$2.76E-4$	$1.36E-10$
1/16	$7.07E-5$	$2.16E-12$
1/32	$1.77E-5$	$8.59E-14$

Diskretisiert man mit dieser Formel die Differentialgleichung

$$-\Delta u + c(x,y)u = f(x,y),$$

so kann man für die entstehenden Matrizen wieder eine Stabilitätsungleichung nachweisen und erhält Konvergenz von der Ordnung 4, falls $u \in C^6(\bar{\Omega})$ gilt. Weitere Mehrstellenformeln findet man in *Collatz* [11] □

Beispiel 7.15. Wir betrachten wie in Beispiel 7.6. das Poisson Problem

$$\Delta u = 0 \quad \text{in } \Omega := (0,1) \times (0,1), \tag{7.21}$$

$$u(x,y) = e^x \cos y \quad \text{auf } \partial\Omega \tag{7.22}$$

mit der Lösung

$$u(x,y) = e^x \cos y,$$

die in $\bar{\Omega}$ beliebig oft differenzierbar ist.

Tabelle 7.3 enthält die Fehler mit der Mehrstellenformel und zum Vergleich noch einmal die Fehler der Diskretisierung mit dem 5-Punkt-Stern. Bei Halbierung der Schrittweite wird der Fehler ungefähr mit dem Faktor 1/16 multipliziert. Dies demonstriert die Konvergenzordnung 4. □

7.4 Allgemeinere Differentialoperatoren

Wir betrachten nur selbstadjungierte Differentialgleichungen, auf die man die Ergebnisse für die Poissongleichung leicht übertragen kann. Allgemeinere Randwertaufgaben zweiter Ordnung werden in *Hackbusch* [27] oder in *Großmann, Roos* [26] diskutiert.

Es sei

$$Lu(x,y) := -\frac{\partial}{\partial x}\left(b\frac{\partial u}{\partial x}\right) - \frac{\partial}{\partial y}\left(c\frac{\partial u}{\partial y}\right) + au = f \text{ in } \Omega \tag{7.23}$$

in einem beschränkten Gebiet $\Omega \subset \mathbb{R}^2$ zusammen mit Dirichletschen oder Neumannschen Randbedingungen oder Randbedingungen dritter Art.

Dabei seien

$$b, c \in C^1(\bar{\Omega}), \quad a, f \in C(\bar{\Omega}),$$

und es seien die Vorzeichenbedingungen

$$b > 0,\ c > 0,\ a \geq 0 \text{ in } \bar{\Omega}$$

erfüllt.

Wir überziehen die Ebene $\mathbb{R}^2$ mit einem äquidistanten Gitter. In randnahen Gitterpunkten verwenden wir wieder die Diskretisierung der Randbedingungen durch Interpolation. Wir müssen also nur die Diskretisierung der Differentialgleichung in inneren Gitterpunkten beschreiben.

Für $u \in C^3$ gilt für den zentralen Differenzenquotienten

$$\begin{aligned} \frac{\partial}{\partial x}u(x,y) &= \frac{1}{2h}\Big(u(x+h,y) - u(x-h,y)\Big) + O(h^2) \\ \frac{\partial}{\partial y}u(x,y) &= \frac{1}{2h}\Big(u(x,y+h) - u(x,y-h)\Big) + O(h^2)\ , \end{aligned}$$

und daher erhält man für $u \in C^4$ mit der Schrittweite $0.5h$

$$\begin{aligned} & Lu(x,y) \\ & = -\frac{1}{h}\left(b(x+\frac{h}{2},y)\frac{\partial u}{\partial x}(x+\frac{h}{2},y) - b(x-\frac{h}{2},y)\frac{\partial u}{\partial x}(x-\frac{h}{2},y)\right) \\ & \quad -\frac{1}{h}\left(c(x,y+\frac{h}{2})\frac{\partial u}{\partial y}(x,y+\frac{h}{2}) - c(x,y-\frac{h}{2})\frac{\partial u}{\partial y}(x,y-\frac{h}{2})\right) \\ & \quad + a(x,y)u(x,y) + O(h^2) \\ = & -\frac{1}{h^2}\left(b(x+\frac{h}{2},y)\Big[u(x+h,y) - u(x,y)\Big] - b(x-\frac{h}{2},y)\Big[u(x,y) - u(x-h,y)\Big]\right. \\ & \quad \left. +c(x,y+\frac{h}{2})\Big[u(x,y+h) - u(x,y)\Big] - c(x,y-\frac{h}{2})\Big[u(x,y) - u(x,y-h)\Big]\right) \\ & \quad + a(x,y)u(x,y) + O(h). \\ = & \left\{\frac{1}{h^2}\left(b(x+\frac{h}{2},y) + b(x-\frac{h}{2},y) + c(x,y+\frac{h}{2}) + c(x,y-\frac{h}{2})\right) + a(x,y)\right\}u(x,y) \\ & \quad -\frac{1}{h^2}\left\{b(x+\frac{h}{2},y)u(x+h,y) + b(x-\frac{h}{2},y)u(x-h,y)\right. \\ & \qquad \left. +c(x,y+\frac{h}{2})u(x,y+h) + c(x,y-\frac{h}{2})u(x,y-h)\right\} + O(h)\ . \end{aligned}$$

Durch Taylorentwicklung der rechten Seite sieht man, dass man sogar $O(h)$ durch $O(h^2)$ ersetzen kann.

Hiermit können wir die Differentialgleichung in inneren Gitterpunkten diskretisieren. Mit

$$b_{i\pm1/2,j} := b(x_i \pm \frac{h}{2}, y_j),\ c_{i,j\pm1/2} := c(x_i, y_j \pm \frac{h}{2}),\ a_{ij} := a(x_i, y_j)$$

erhält man

$$\begin{aligned}(\boldsymbol{A}_h\boldsymbol{U})_{ij} := &\left\{\left(b_{i-1/2,j} + b_{i+1/2,j} + c_{i,j-1/2} + c_{i,j+1/2}\right) + h^2 a_{ij}\right\} U_{ij} \\ - &\left\{b_{i-1/2,j}U_{i-1,j} + b_{i+1/2,j}U_{i+1,j} + c_{i,j-1/2}U_{i,j-1} + c_{i,j+1/2}U_{i,j+1}\right\} \\ = &\ h^2 f_{ij}\ .\end{aligned}$$

Zusammen mit den Diskretisierungen der Randbedingungen in den randnahen Gitterpunkten erhält man ein lineares Gleichungssystem

$$\boldsymbol{A}_h\boldsymbol{U}_h = \boldsymbol{f}_h,$$

das die folgenden Besonderheiten aufweist.

- In jeder Zeile der Koeffizientenmatrix $\boldsymbol{A}_h$ sind außer dem Diagonalelement nur (höchstens) vier Elemente von Null verschieden.
- Die Diagonalelemente sind positiv, die übrigen Elemente sind nicht positiv, und die Zeilensumme in jeder Zeile von $\boldsymbol{A}_h$ ist nicht negativ.
- Verkleinert man die Schrittweite h, so wächst die Zahl der Variablen wie

  $$\text{const}\ \cdot h^{-2}.$$

  Man hat also bei Problemen der Praxis mit sehr großen Dimensionen zu rechnen. In Beispiel 7.12. haben wir bereits bei der Schrittweite $h = 1/64$ ein lineares Gleichungssystem der Dimension $n = 12033$.

  Für große und dünn besetzte lineare Gleichungssysteme wurden spezielle Varianten des Gaußschen Eliminationsverfahrens und iterative Verfahren entwickelt. Diese werden in den Vorlesungen “Numerische Lineare Algebra” und “Numerik großer Systeme” besprochen. Wir gehen in dieser Vorlesung nicht darauf ein.

Die Konvergenzeigenschaften werden durch den folgenden Satz beschrieben:

Satz 7.16. *Es gebe ein $h_0 > 0$, so dass für alle $h < h_0$ für $P \in \partial\Omega_h$ die Interpolation in der Diskretisierung der Randbedingungen mit einem inneren Gitterpunkt Q möglich ist und dass Ω_h gitterzusammenhängend ist (d.h. zwei Punkte aus Ω_h lassen sich durch eine Kette benachbarter Punkte von Ω_h verbinden). Dann ist das Gleichungssystem*

$$A_h U_h = f_h$$

für jede rechte Seite f_h eindeutig lösbar.

Ist $u \in C^4(\bar{\Omega})$ eine Lösung der Randwertaufgabe, so gilt

$$\|U_h - R_h u\| = O(h^2).$$

7.5 Idee der Methode der finiten Volumen

Finite Volumen Methoden (auch **Box–Schemata**) werden verwendet, um Differenzenverfahren auf unregelmäßigen Gittern für Erhaltungsgleichungen

$$\operatorname{div} w(x,y) = f(x,y),\ (x,y) \in \Omega \tag{7.24}$$

zusammen mit Randbedingungen zu konstruieren. Dabei ist w ein (ebenes) Vektorfeld auf Ω. Wir beschreiben das Konstruktionsprinzip für den folgenden Spezialfall:

$$w(x,y) = -c(x,y)\nabla u(x,y). \tag{7.25}$$

Dann ist

$$\operatorname{div} w(x,y) = -\frac{\partial}{\partial x}\Big(c(x,y)\frac{\partial}{\partial x}u(x,y)\Big) - \frac{\partial}{\partial y}\Big(c(x,y)\frac{\partial}{\partial y}u(x,y)\Big), \tag{7.26}$$

und speziell für $c(x,y) \equiv 1$ erhält man

$$\operatorname{div} w(x,y) = -\Delta u(x,y).$$

Es sei Ω ein polygonal begrenztes Gebiet, das in Dreiecke und Rechtecke zerlegt sei. Diese Zerlegung heißt **Primärzerlegung**.

Wir setzen voraus, dass die Zerlegung **zulässig** ist, d.h. dass der Durchschnitt zweier Dreiecke oder Rechtecke der Zerlegung entweder eine gemeinsame Begrenzungsgerade ist, ein Punkt ist oder leer ist. Abbildung 7.3 zeigt eine Zerlegung eines Gebiets in Dreiecke, die nur an der markierten Stelle unzulässig ist.

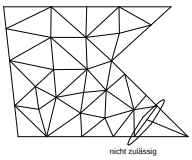

Abbildung 7.3: Zulässigkeit der Zerlegungen

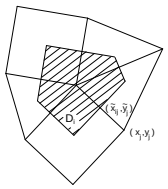

Abbildung 7.4: Sekundärzerlegungen

Jedem Eckpunkt (x_i, y_i) der Primärzerlegung wird ein Gebiet D_i einer **Sekundärzerlegung** zugeordnet durch

$$D_i := \{ \begin{pmatrix} x \\ y \end{pmatrix} \; : \; \| \begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} x_i \\ y_i \end{pmatrix} \|_2 < \| \begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} x_j \\ y_j \end{pmatrix} \|_2 \text{ für alle übrigen Knoten } \begin{pmatrix} x_j \\ y_j \end{pmatrix} \}.$$

Abbildung 7.4 zeigt ein Element einer Sekundärzerlegung.

Integration der Differentialgleichung über D_i liefert

$$-\int\limits_{D_i} \operatorname{div}\left(c(x,y)\nabla u(x,y)\right) d(x,y) = \int\limits_{D_i} f(x,y)\, d(x,y),$$

und mit dem Gaußschen Integralsatz folgt

$$-\int\limits_{\partial D_i} c(x,y)\frac{\partial}{\partial n}u(x,y)\,ds = \int\limits_{D_i} f(x,y)\,d(x,y).$$

Wir verwenden nun die folgenden Bezeichnungen: Es sei

I_i die Menge der Indizes (x_i, y_i) benachbarter Eckpunkte,

∂D_{ij} das geradlinige Begrenzungsstück von D_i, das auf der Mittelsenkrechten der Strecke von (x_i, y_i) nach (x_j, y_j), $j \in I_i$, liegt,

$(\tilde{x}_{ij}, \tilde{y}_{ij}) := 0.5((x_i, y_i) + (x_j, y_j))$ der Mittelpunkt der Strecke von (x_i, y_i) nach (x_j, y_j),

$\lambda_{ij} = \|(x_i, y_i) - (x_j, y_j)\|$ ihre Länge,

ℓ_{ij} die Länge der Strecke ∂D_{ij},

$\mu(D_i)$ der Flächeninhalt von D_i

Dann gilt

$$\int\limits_{\partial D_i} c(x,y)\frac{\partial}{\partial n}u(x,y)\,ds = \sum_{j\in I_i}\int\limits_{\partial D_{ij}} c(x,y)\frac{\partial}{\partial n}u(x,y)\,ds$$

Approximiert man die Integrale der rechten Seite mit der Mittelpunktregel:

$$\int\limits_{\partial D_{ij}} c(x,y)\frac{\partial}{\partial n}u(x,y)\,ds \approx \ell_{ij} c(\tilde{x}_{ij}, \tilde{y}_{ij})\frac{\partial}{\partial n}u(\tilde{x}_{ij}, \tilde{y}_{ij}),$$

und diskretisiert man die Normalableitung durch den zentralen Differenzenquotienten

$$\frac{\partial}{\partial n}u(\tilde{x}_{ij}, \tilde{y}_{ij}) \approx \frac{u(x_j, y_j) - u(x_i, y_i)}{\lambda_{ij}},$$

so erhält man

$$\int\limits_{\partial D_{ij}} c(x,y)\frac{\partial}{\partial n}u(x,y)\,ds \approx \ell_{ij} c(\tilde{x}_{ij}, \tilde{y}_{ij})\frac{u(x_j, y_j) - u(x_i, y_i)}{\lambda_{ij}}.$$

Approximiert man das Flächenintegral durch

$$\int\limits_{D_i} f(x,y)\,d(x,y) \approx f(x_i, y_i)\,\mu(D_i),$$

so erhält man schließlich die Diskretisierung

$$f(x_i, y_i)\,\mu(D_i) = \sum_{j\in I_i} \frac{\ell_{ij}}{\lambda_{ij}} c(\tilde{x}_{ij}, \tilde{y}_{ij})(U_i - U_j)$$

für alle inneren Knoten der Primärzerlegung. Hierzu kommen noch Diskretisierungen der Randbedingungen.

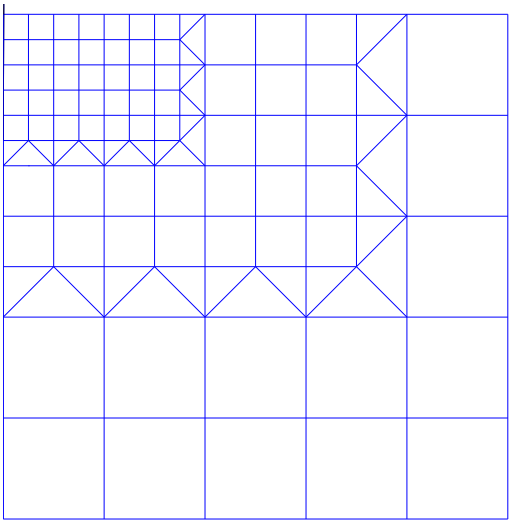

Abbildung 7.5: Gitterverfeinerung bei finiten Volumen

Beispiel 7.17.

$$-\Delta u = f \text{ in } \Omega := (0,1) \times (0,1), \quad u = g \text{ auf } \partial\Omega.$$

Zerlegt man Ω wie vorher in Quadrate der Seitenlänge $h := 1/n$, so gilt $\ell_{ij} = \lambda_{ij} = h$ und $\mu(D_i) = h^2$, und man erhält dieselbe Diskretisierung wie bei dem Differenzenverfahren.

Ein Vorteil des Verfahrens der finiten Volumen ist, dass man das Gitter (z.B. bei einspringenden Ecken des Grundbereichs) mühelos verfeinern kann. Abbildung 7.5 zeigt, wie man diese Verfeinerung vornehmen kann.

Kapitel 8

Finite Elemente für elliptische Randwertaufgaben

Die Idee der Methode der finiten Elemente besteht in einem bestimmten Konstruktionsprinzip von endlich dimensionalen Funktionenräumen als Ansatzräume für Ritz–Galerkin–Verfahren für Variationsprobleme. Da eine Vielzahl elliptischer Randwertaufgaben eine Formulierung als Variationsaufgaben besitzen, ist die Methode der finiten Elemente vor allem eine Methode zur Lösung elliptischer Randwertaufgaben, auch wenn es entsprechende Ansätze zur Lösung parabolischer und hyperbolischer Aufgaben gibt.

Mathematiker verweisen gern auf "ihren" Richard Courant, der wohl als erster 1943 die Idee formulierte und in einem Existenzsatz verwendete. Dennoch ist die Methode der finiten Elemente in wesentlichen Punkten ab 1956 von Ingenieuren zur Lösung von Problemen der konstruktiven Mechanik entwickelt worden.

8.1 Variationsmethoden

Wir betrachten in dem beschränkten Gebiet $\Omega \subset \mathbb{R}^m$ mit dem Rand $\partial\Omega$ die lineare Randwertaufgabe

$$-\Delta u(\boldsymbol{x}) = f(\boldsymbol{x}),\ \boldsymbol{x} \in \Omega, \quad u(\boldsymbol{x}) = 0,\ \boldsymbol{x} \in \partial\Omega, \tag{8.1}$$

wobei $f \in C(\bar{\Omega})$ gegeben ist.

Ist $u \in C^2(\Omega) \cap C(\bar{\Omega})$ eine Lösung von (8.1), so gilt für alle $v \in C^1(\Omega) \cap C(\bar{\Omega})$ mit $v(\boldsymbol{x}) = 0$ auf $\partial\Omega$

$$-\int_\Omega \Delta u(\boldsymbol{x}) \cdot v(\boldsymbol{x})\, d\boldsymbol{x} = \int_\Omega f(\boldsymbol{x}) v(\boldsymbol{x})\, d\boldsymbol{x}. \tag{8.2}$$

Ist Ω ein Standardbereich, so folgt aus der Greenschen Formel

$$\int_\Omega \Delta u(\boldsymbol{x}) \cdot v(\boldsymbol{x})\, d\boldsymbol{x} = -\int_\Omega \langle \nabla u(\boldsymbol{x}), \nabla v(\boldsymbol{x}) \rangle\, d\boldsymbol{x} + \int_{\partial\Omega} \frac{\partial u}{\partial n}(\boldsymbol{x}) v(\boldsymbol{x})\, do,$$

und wegen $v(\boldsymbol{x}) = 0$ für $\boldsymbol{x} \in \partial\Omega$

$$\begin{aligned} a(u,v) := \int_\Omega \langle \nabla u(\boldsymbol{x}), \nabla v(\boldsymbol{x}) \rangle\, d\boldsymbol{x} = \int_\Omega f(\boldsymbol{x}) v(\boldsymbol{x})\, d\boldsymbol{x} =: F(v) \\ \text{für alle } v \in \tilde{V} := \{ v \in C^1(\Omega) \cap C(\bar{\Omega}) \,:\, v = 0 \text{ auf } \partial\Omega \}. \end{aligned} \tag{8.3}$$

Genauso sind wir bei der Behandlung von gewöhnlichen Differentialgleichungen in Abschnitt 6.3 vorgegangen. Wir konnten dort sofort mit Hilfe der absolut stetigen Funktionen einen Funktionenraum angeben, der durch das innere Produkt $a(\cdot,\cdot)$ zum Hilbertraum wurde, und erhielten mit dem Darstellungssatz von Riesz für lineare stetige Funktionale die eindeutige Lösbarkeit der Variationsaufgabe (8.3). Bei partiellen Differentialgleichungen ist dies nicht mehr ganz so einfach.

Allgemeiner als in Abschnitt 6.3 gehen wir aus von einem reellen Hilbertraum V mit dem inneren Produkt $\langle \cdot, \cdot \rangle$. Es sei $a : V \times V \to \mathbb{R}$ eine **Bilinearform** auf V (d.h. $u \mapsto a(u,v)$ ist für jedes feste $v \in V$ ein lineares Funktional und $v \mapsto a(u,v)$ ist für jedes feste $u \in V$ ein lineares Funktional) und F ein stetiges lineares Funktional auf V. Dann liefert der folgende Satz eine hinreichende Bedingung für die eindeutige Lösbarkeit der Variationsaufgabe

$$\text{Bestimme } u \in V \,:\, a(u,v) = F(v) \quad \text{für alle } v \in V. \tag{8.4}$$

Satz 8.1. (Lax – Milgram) *Die Bilinearform a sei* **stetig** *, d.h. es existiere ein $M \geq 0$ mit*

$$|a(u,v)| \leq M \|u\| \cdot \|v\| \quad \textit{für alle } u, v \in V, \tag{8.5}$$

und **V–elliptisch**, *d.h. es gebe ein $\alpha > 0$ mit*

$$a(u,u) \geq \alpha \|u\|^2 \quad \textit{für alle } u \in V. \tag{8.6}$$

Dann besitzt das Variationsproblem (8.4) eine eindeutige Lösung $\tilde{u} \in V$.

Beweis: Wir zeigen den Satz von Lax–Milgram nur für symmetrisches a, d.h.

$$a(u,v) = a(v,u) \quad \text{für alle } u,v \in V, \tag{8.7}$$

den Beweis für den allgemeinen Fall findet man in *Großmann – Roos* [26], p. 91.

Unter der Voraussetzung (8.7) wird durch

$$\langle u,v\rangle_a := a(u,v), \quad u,v \in V,$$

ein inneres Produkt auf V definiert, und wegen (8.5) und (8.6) gilt für die durch $\|u\|_a := \sqrt{\langle u,u\rangle_a}$ induzierte **Energienorm**

$$\alpha\|u\|^2 \le \|u\|_a^2 \le M\|u\|^2 \quad \text{für alle } u \in V,$$

und damit sind $\|\cdot\|$ und $\|\cdot\|_a$ äquivalente Normen.

Der Rest folgt wieder aus dem Darstellungssatz von Riesz, da F auch bzgl. $\|\cdot\|_a$ stetig ist und daher eine eindeutige Darstellung

$$F(v) := \langle \tilde{u}, v\rangle_a \quad \text{für alle } v \in V$$

besitzt. ∎

Man kann für unser Problem leicht einen unitären Raum angeben, in dem die Voraussetzungen von Satz 8.1. erfüllt sind.

Satz 8.2. *Sei*

$$X := \{v \in C^1(\Omega) \cap C(\bar{\Omega}) \,:\, v = 0 \text{ auf } \partial\Omega\}$$

und für $u,v \in X$

$$(u,v) := \int_\Omega \Big(\langle \nabla u(\boldsymbol{x}), \nabla v(\boldsymbol{x})\rangle + u(\boldsymbol{x})v(\boldsymbol{x})\Big)\, d\boldsymbol{x}. \tag{8.8}$$

Dann ist X mit $(\cdot,\cdot)$ aus (8.8) ein (nicht vollständiger) unitärer Raum, und F ist stetig auf X. Ferner gelten die Bedingungen (8.5) und (8.6) mit X anstatt V.

Beweis: Dass $(\cdot,\cdot)$ ein Skalarprodukt auf X ist, rechnet man leicht nach.

Nach der Cauchy–Schwarzschen Ungleichung ist

$$\begin{aligned} |F(v)| &= \Big|\int_\Omega f(\boldsymbol{x})v(\boldsymbol{x})\,d\boldsymbol{x}\Big| \le \sqrt{\int_\Omega f(\boldsymbol{x})^2\,d\boldsymbol{x}}\sqrt{\int_\Omega v(\boldsymbol{x})^2\,d\boldsymbol{x}} \\ &\le \sqrt{\int_\Omega f(\boldsymbol{x})^2\,d\boldsymbol{x}}\sqrt{\int_\Omega \Big(\langle \nabla v(\boldsymbol{x}), \nabla v(\boldsymbol{x})\rangle + v(\boldsymbol{x})^2\Big)\,d\boldsymbol{x}} =: C\cdot\|v\|, \end{aligned}$$

und daher ist F beschränkt, also stetig.

Für $u, v \in X$ folgt aus der Cauchy–Schwarzschen Ungleichung

$$\begin{aligned} |a(u,v)| &= \int_\Omega \langle \nabla u(\boldsymbol{x}), \nabla v(\boldsymbol{x}) \rangle \, d\boldsymbol{x} \\ &\leq \sqrt{\int_\Omega \|\nabla u(\boldsymbol{x})\|_2^2 \, d\boldsymbol{x}} \sqrt{\int_\Omega \|\nabla v(\boldsymbol{x})\|_2^2 \, d\boldsymbol{x}} \\ &\leq \sqrt{\int_\Omega (\|\nabla u(\boldsymbol{x})\|_2^2 + u(\boldsymbol{x})^2) \, d\boldsymbol{x}} \sqrt{\int_\Omega (\|\nabla v(\boldsymbol{x})\|_2^2 + v(\boldsymbol{x})^2) \, d\boldsymbol{x}} \\ &= \|u\| \cdot \|v\|. \end{aligned}$$

Damit ist die Bilinearform a stetig.

Die Eliptizität erhält man aus der folgenden Ungleichung von Poincaré.

Lemma 8.3. (Ungleichung von Poincaré) *Es sei $\Omega \subset \mathbb{R}^m$ ein beschränktes Gebiet und für $u \in C^1(\Omega)$*

$$|u|_1 := \sqrt{\int_\Omega \|\nabla u(\boldsymbol{x})\|_2^2 \, d\boldsymbol{x}}. \tag{8.9}$$

Dann gibt es eine Konstante C, die nur von Ω abhängt, so dass

$$\int_\Omega u(\boldsymbol{x})^2 \, d\boldsymbol{x} \leq C|u|_1^2 \tag{8.10}$$

für alle

$$u \in X := \{u \in C^1(\Omega) \cap C(\bar{\Omega}) \, : \, u(\boldsymbol{x}) = 0 \text{ auf } \partial\Omega\}.$$

Beweis: Wegen der Beschränktheit von Ω gibt es ein $R > 0$ mit $\|\boldsymbol{x}\|_\infty \leq R$ für alle $\boldsymbol{x} \in \Omega$. Wir setzen die Funktion $u \in X$ auf $\tilde{\Omega} := [-R, R]^m$ fort zu

$$\tilde{u}(\boldsymbol{x}) := \begin{cases} u(\boldsymbol{x}) & \text{für} \quad \boldsymbol{x} \in \Omega \\ 0 & \text{für} \quad \boldsymbol{x} \in \tilde{\Omega} \setminus \Omega \end{cases}.$$

Dann ist $\tilde{u}$ stetig und stückweise stetig differenzierbar. Damit gilt für alle $\boldsymbol{x} \in \tilde{\Omega}$

$$\tilde{u}(\boldsymbol{x}) = \int_{-R}^{x_1} 1 \cdot \frac{\partial}{\partial x_1} \tilde{u}(t, x_2, \ldots, x_m) \, dt.$$

Mit der Cauchy–Schwarzschen Ungleichung folgt

$$\tilde{u}(\boldsymbol{x})^2 \leq \int_{-R}^{x_1} 1 \, dt \cdot \int_{-R}^{x_1} \Big(\frac{\partial}{\partial x_1} \tilde{u}(\boldsymbol{x})\Big)^2 \, dt \leq 2R \int_{-R}^{x_1} \Big(\frac{\partial}{\partial x_1} \tilde{u}(\boldsymbol{x})\Big)^2 \, dt,$$

und man erhält schließlich

$$\begin{aligned}\int_\Omega u(\boldsymbol{x})^2\,d\boldsymbol{x} &= \int_{\tilde\Omega} \tilde u(\boldsymbol{x})^2\,d\boldsymbol{x} = \int_{-R}^{R}\dots\int_{-R}^{R} \tilde u(\boldsymbol{x})\,dx_m\dots dx_1\\ &\le 2R\int_{-R}^{R}\dots\int_{-R}^{R}\int_{-R}^{R}\Big(\frac{\partial}{\partial x_1}\tilde u(t,x_2,\dots,x_m)\Big)^2\,dt\,dx_m\dots dx_1\\ &= 4R^2\int_{\tilde\Omega}\Big(\frac{\partial}{\partial x_1}\tilde u(\boldsymbol{x})\Big)^2\,d\boldsymbol{x} \le 4R^2\int_\Omega \|\nabla u(\boldsymbol{x})\|_2^2\,d\boldsymbol{x}.\end{aligned}$$

∎

Aus der Poincaréschen Ungleichung erhalten wir nun für $u \in X$

$$\begin{aligned}\|u\|^2 &= \int_\Omega \Big(\|\nabla u(\boldsymbol{x})\|_2^2 + u(\boldsymbol{x})^2\Big)\,d\boldsymbol{x} \le (1+4R^2)\int_\Omega \|\nabla u(\boldsymbol{x})\|_2^2\,d\boldsymbol{x}\\ &= (1+4R^2)a(u,u),\end{aligned}$$

d.h. (8.6) mit $\alpha := 1/(1+4R^2)$. ∎

Bemerkung 8.4. $|u|_1$ in (8.9) ist auf $C^1(\Omega)$ nur eine Seminorm, d.h. eine Abbildung $|\cdot|_1 : C^1(\Omega) \to \mathbb{R}_+$, die homogen ist und für die die Dreiecksungleichung gilt, die aber nicht notwendig definit ist. Aus der Poincaréschen Ungleichung folgt, dass auf dem Raum X aber $|\cdot|_1$ eine Norm ist, die der Norm $\|\cdot\|$ äquivalent ist:

$$C_1\|u\| \le |u|_1 \le C_2\|u\|.$$

Die erste dieser Ungleichungen folgt aus der Poincaréschen Ungleichung und die zweite ist trivial. □

Satz 8.2. legt nun nahe, als Hilbertraum die Vervollständigung von X bzgl. der durch $(\cdot,\cdot)$ induzierten Norm zu wählen. Man erhält dann den **Sobolevraum** $H_0^1(\Omega)$.

Damit erhält man die Variationsgleichung

$$\text{Bestimme } u \in H_0^1(\Omega) : a(u,v) = F(v) \text{ für alle } v \in H_0^1(\Omega). \tag{8.11}$$

Es ist klar, dass jede Lösung $u \in C^2(\Omega)\cap C(\bar\Omega)$ der Randwertaufgabe (8.1) auch eine Lösung der Variationsaufgabe (8.11) ist. Umgekehrt ist eine Lösung von (8.11) nur dann eine Lösung der Randwertaufgabe (8.1), wenn u glatt genug ist, nämlich $u \in C^2(\Omega)\cap C(\bar\Omega)$. Wir nennen jede Lösung von (8.11) eine **schwache Lösung**

der Randwertaufgabe (8.1) und nennen (zur Unterscheidung) eine Lösung $u \in C^2(\Omega) \cap C(\bar{\Omega})$ von (8.1) eine **starke Lösung** oder **klassische Lösung** von (8.1).

Einen anderen Zugang zum Sobolewraum $H_0^1(\Omega)$ erhält man über die folgende Verallgemeinerung der Differenzierbarkeit: Für $\phi : \Omega \to \mathbb{R}$ bezeichnen wir mit

$$\mathrm{supp}(\phi) := \overline{\{\boldsymbol{x} \in \Omega : \phi(\boldsymbol{x}) \neq 0\}}$$

den **Träger** der Funktion ϕ (engl.: support). Es sei hiermit

$$C_0^\infty(\Omega) := \{\phi \in C^\infty : \mathrm{supp}(\phi) \subset \Omega\}.$$

Beachten Sie, dass nach Voraussetzung Ω eine offene Menge ist und dass $\mathrm{supp}(\phi) \subset \Omega$ eine abgeschlossene Menge ist. Der Abstand von $\partial\Omega$ und $\mathrm{supp}(\phi)$ ist also positiv, und daher ist nicht nur ϕ auf dem Rand $\partial\Omega$ gleich 0, sondern dies gilt auch für alle Ableitungen von ϕ.

Ist $u \in C^1(\Omega) \cap C(\bar{\Omega})$, so gilt nach dem Integralsatz von Gauß mit dem äußeren Normalenvektor $\boldsymbol{n}$ auf $\partial\Omega$

$$\begin{aligned}\int_\Omega \frac{\partial}{\partial x_j} u(\boldsymbol{x})\phi(\boldsymbol{x})\, d\boldsymbol{x} &= -\int_\Omega \frac{\partial}{\partial x_j}\phi(\boldsymbol{x})u(\boldsymbol{x})\, d\boldsymbol{x} + \int_{\partial\Omega} u(\boldsymbol{x})\phi(\boldsymbol{x})\cos(\boldsymbol{n}(\boldsymbol{x}), \boldsymbol{e}^j)\, do \\ &= -\int_\Omega \frac{\partial}{\partial x_j}\phi(\boldsymbol{x})u(\boldsymbol{x})\, d\boldsymbol{x}.\end{aligned}$$

Man definiert daher

Definition 8.5. *Die Funktion $u : \Omega \to \mathbb{R}$ besitzt die* **verallgemeinerte Ableitung** *oder* **schwache Ableitung** *$v : \Omega \to \mathbb{R}$ nach x_j, wenn für alle $\phi \in C_0^\infty(\Omega)$ gilt*

$$\int_\Omega v(\boldsymbol{x})\phi(\boldsymbol{x})\, d\boldsymbol{x} = -\int_\Omega u(\boldsymbol{x})\frac{\partial}{\partial x_j}\phi(\boldsymbol{x})\, d\boldsymbol{x}.$$

Besitzt u verallgemeinerte Ableitungen nach allen Komponenten, so bezeichnet ∇u den Vektor dieser schwachen Ableitungen.

Höhere Ableitungen definiert man (wie für die klassischen Ableitungen) nun induktiv.

Wir bezeichnen mit $H^\ell(\Omega)$ die Menge aller Funktionen $u : \Omega \to \mathbb{R}$, für die alle verallgemeinerten Ableitungen

$$D^\alpha u := \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \dots \partial x_m^{\alpha_m}} u,\ \alpha \in {\mathbb{N}_0}^m,\ |\alpha| := \sum_{j=1}^m \alpha_j,$$

bis zur Ordnung ℓ existieren und die zusammen mit ihren Ableitungen (im Lebesgueschen Sinne) quadratisch integrierbar über Ω sind.

Man kann zeigen, dass $H^\ell(\Omega)$ mit dem Skalarprodukt

$$(u,v) := \int_\Omega \sum_{|\alpha|\le\ell} D^\alpha u(\boldsymbol{x}) D^\alpha v(\boldsymbol{x})\, d\boldsymbol{x}$$

ein Hilbertraum ist. Man kann ferner zeigen, dass $H_0^1(\Omega)$ ein Teilraum von $H^1(\Omega)$ ist, der aus denjenigen Funktionen in $H^1(\Omega)$ besteht, die (in einem verallgemeinerten Sinne) die Randbedingungen $u = 0$ auf $\partial\Omega$ erfüllen.

Wir geben nun eine für die numerische Behandlung von Differentialgleichungen besonders wichtige Menge von Elementen von $H_0^1(\Omega)$ an, die für die Verwendung in einem Ritz–Galerkin Verfahren in Frage kommen:

$$V^s := \{v \in C(\bar\Omega) \,:\, v \text{ stückweise stetig differenzierbar, } v(\boldsymbol{x}) = 0 \text{ auf } \partial\Omega\}. \quad (8.12)$$

Dabei heißt $v \in C(\bar\Omega)$ **stückweise stetig differenzierbar** auf Ω, wenn es zu v Normalgebiete Ω_j, $j = 1,\dots,N$, (d.h. Gebiete, in denen der Gaußsche Integralsatz anwendbar ist) gibt, so dass gilt

$$\bar\Omega = \bigcup_{j=1}^N \bar\Omega_j,\ \Omega_i \cap \Omega_j = \emptyset \text{ für } i \ne j,\ v \in C^1(\bar\Omega_j),\ j = 1,\dots,N.$$

Es ist klar, dass diese Funktionen in $H^1(\Omega)$ liegen, denn für die stückweise klassischen Ableitungen von v in den Teilgebieten Ω_j gilt

$$\begin{aligned}
\int_\Omega v(\boldsymbol{x}) \frac{\partial}{\partial x_k}\phi(\boldsymbol{x})\, d\boldsymbol{x} &= \sum_{j=1}^N \int_{\Omega_j} v(\boldsymbol{x}) \frac{\partial}{\partial x_k}\phi(\boldsymbol{x})\, d\boldsymbol{x} \\
&= -\sum_{j=1}^N \int_{\Omega_j} \frac{\partial}{\partial x_k} v(\boldsymbol{x})\phi(\boldsymbol{x})\, d\boldsymbol{x} + \sum_{j=1}^N \int_{\partial\Omega_j} v(\boldsymbol{x})\phi(\boldsymbol{x}) \cos(\boldsymbol{n}_j(\boldsymbol{x}), \boldsymbol{e}^k)\, do \\
&= -\sum_{j=1}^N \int_{\Omega_j} \frac{\partial}{\partial x_k} v(\boldsymbol{x})\phi(\boldsymbol{x})\, d\boldsymbol{x},
\end{aligned}$$

da die Oberflächenintergrale über die inneren Ränder der Ω_j sich gerade gegeneinander wegheben und $\phi = 0$ auf $\partial\Omega$ gilt. Damit ist die aus den klassischen Ableitungen auf den Teilgebieten zusammengesetzte Funktion verallgemeinerte Ableitung von v und wegen $v \in C^1(\bar\Omega_j)$ ist diese Funktion beschränkt und stückweise stetig, und daher quadratisch integrierbar. Wegen $v = 0$ auf $\partial\Omega$ liegt v sogar in $H_0^1(\Omega)$ (ohne Beweis).

Bemerkung 8.6. Wir haben nur den Laplaceoperator Δ betrachtet. Die Ergebnisse gelten genauso für lineare, gleichmäßig elliptische, formal selbstadjungierte Differentialoperatoren

$$\left.\begin{aligned} Lu(\boldsymbol{x}) := -\sum_{i,j=1}^{m} \frac{\partial}{\partial x_i}\Big(a_{ij}(\boldsymbol{x})\frac{\partial u}{\partial x_j}\Big) + c(\boldsymbol{x})u(\boldsymbol{x}) &= f(\boldsymbol{x}),\ \boldsymbol{x} \in \Omega \subset \mathbb{R}^m \\ u(\boldsymbol{x}) &= 0,\ \boldsymbol{x} \in \partial\Omega \end{aligned}\right\} \tag{8.13}$$

Dabei heißt der Operator **formal selbstadjungiert**, wenn er die Gestalt der linken Seite von (8.13) hat, und er heißt **gleichmäßig elliptisch** , wenn es eine Konstante $\alpha > 0$ gibt mit

$$\sum_{i,j=1}^{m} a_{ij}(\boldsymbol{x})\xi_i\xi_j > \alpha\|\boldsymbol{\xi}\|_2^2 \quad \text{für alle } \boldsymbol{x} \in \Omega \text{ und alle } \boldsymbol{\xi} \in \mathbb{R}^n.$$

Als Bilinearform der Variationsgleichung erhält man dann

$$a(u,v) := \sum_{i,j=1}^{m} \int_\Omega a_{ij}(\boldsymbol{x})\frac{\partial u}{\partial x_i}\frac{\partial v}{\partial x_j}\,d\boldsymbol{x} + \int_\Omega c(\boldsymbol{x})u(\boldsymbol{x})v(\boldsymbol{x})\,d\boldsymbol{x}. \tag{8.14}$$

□

Bemerkung 8.7. Wie in Abschnitt 6.3 kann man zeigen, dass die Lösung von (8.3) das quadratische Funktional

$$J(u) := a(u,u) - 2F(u)$$

auf V minimiert und dass umgekehrt die Lösung von

$$J(u) := \min!,\ u \in V, \tag{8.15}$$

die Variationsgleichung (8.3) löst. □

Bemerkung 8.8. Bei Problemen der Elastizitätstheorie ist $u(\boldsymbol{x})$ die Verschiebung eines elastischen Körpers unter Einwirkung einer Kraft mit der Kraftdichte f. Die Randbedingung $u(\boldsymbol{x}) = 0$ auf $\partial\Omega$ besagt dann, dass der Körper am Rande eingespannt ist.

$$a(u,u) := \sum_{i,j=1}^{m} \int_\Omega a_{ij}(\boldsymbol{x})\frac{\partial u}{\partial x_i}\frac{\partial u}{\partial x_j}\,d\boldsymbol{x} + \int_\Omega c(\boldsymbol{x})u(\boldsymbol{x})^2\,d\boldsymbol{x}$$

ist die innere Verformungsenergie, und das Funktional

$$J(u) = \frac{1}{2}a(u,u) - F(u)$$

beschreibt die potentielle Energie. Die Variationsaufgabe (8.15) lässt sich also als Energieminimierungsprinzip deuten. □

Bemerkung 8.9. Wir ersetzen die Dirichletschen Randbedingungen durch die Neumannsche Bedingung

$$\frac{\partial}{\partial n}u(\boldsymbol{x}) = 0 \quad \text{auf } \boldsymbol{x} \in \partial\Omega \tag{8.16}$$

und fordern $c(\boldsymbol{x}) > 0$ in $\bar{\Omega}$. Dann bleiben alle Resultate erhalten, wenn man in X und V^s die Randbedingungen ersatzlos streicht. Dies liegt wieder daran, dass $\frac{\partial}{\partial n}u$ keine wesentliche, sondern eine natürliche Randbedingung ist; sie muss also nicht von den Ansatzfunktionen erfüllt sein.

Bei der dritten Randbedingung oder bei gemischten Randbedingungen tritt in der Bilinearform (8.14) zusätzlich noch ein Randintegral auf, das die Ableitung in Richtung der Konormalen enthält. Für den Laplace Operator fallen die Konormale und die Normale zusammen, und man erhält für das Problem

$$-\Delta u = f \text{ in } \Omega,\ u = 0 \text{ auf } \Gamma_1,\ \frac{\partial u}{\partial n} + \alpha(\boldsymbol{x})u = 0 \text{ auf } \Gamma_2 \tag{8.17}$$

mit $\Gamma_1 \cup \Gamma_2 = \partial\Omega$, $\Gamma_1 \cap \Gamma_2 = \emptyset$, $\alpha : \Gamma_2 \to \mathbb{R}$, $\alpha \geq 0$, die Variationsaufgabe

$$a(u,v) := \int_\Omega \langle \nabla u, \nabla v \rangle \, d\boldsymbol{x} + \int_{\Gamma_2} \alpha(\boldsymbol{x})u(\boldsymbol{x})v(\boldsymbol{x}) \, do = \int_\Omega f(\boldsymbol{x})v(\boldsymbol{x}) \, d\boldsymbol{x} \text{ für alle } v \in V, \tag{8.18}$$

wobei V die Vervollständigung von

$$\{v \in C^2(\Omega) \cap C^1(\bar{\Omega}) \,:\, v(\boldsymbol{x}) = 0 \text{ auf } \Gamma_1\}$$

bzgl. des inneren Produktes (8.7) bezeichnet. □

Bemerkung 8.10. Ausgehend von (8.15) kann man auch gewisse nichtlineare Differentialgleichungen in Variationsprobleme überführen. Dem Problem

$$J(u) := \int_\Omega F(x,y,u,u_x,u_y) \, d(x,y) = \min!, \quad u \in C^1(\Omega) \cap C(\bar{\Omega}),\ u(x,y) = 0 \text{ auf } \partial\Omega$$

kann man als notwendige Bedingung die Euler–Lagrange Gleichung

$$-\frac{\partial}{\partial x}\frac{\partial}{\partial u_x}F - \frac{\partial}{\partial y}\frac{\partial}{\partial u_y}F + \frac{\partial}{\partial u}F = 0 \text{ in } \Omega,\ u = 0 \text{ auf } \partial\Omega,$$

gegenüberstellen, und diese mit dem Ritz–Galerkin Verfahren behandeln. □

8.2 Methode der finiten Elemente

Das Variationsproblem (8.3) (bzw. (8.14) bzw. (8.18)) kann man wieder mit dem Ritz–Galerkin Verfahren diskretisieren.

> Es sei V_h ein endlich dimensionaler Teilraum von $H_0^1(\Omega)$ (bzw. in Bemerkung 8.9. des Raumes V der Elemente von $H^1(\Omega)$, die die wesentlichen Randbedingungen $v = 0$ auf Γ_1 erfüllen).
>
> Bestimme als Näherung für eine (verallgemeinerte) Lösung der Randwertaufgabe (8.1) die Lösung der endlich dimensionalen Variationsaufgabe
>
> $$u_h \in V_h \,:\, a(u_h, v_h) = F(v_h) \quad \text{für alle } v_h \in V_h \tag{8.19}$$

Ist $\{v_1, \ldots, v_n\}$ eine Basis von V_h, so können wir jedes $v \in V_h$ darstellen als

$$v = \sum_{j=1}^{n} \xi_j v_j, \quad \xi_1, \ldots, \xi_n \in \mathbb{R},$$

und das Variationsproblem geht über in das lineare Gleichungssystem

$$\int_\Omega \sum_{j=1}^{n} \xi_j \langle \nabla v_j(\boldsymbol{x}), \nabla v_k(\boldsymbol{x}) \rangle \, d\boldsymbol{x} = \int_\Omega f(\boldsymbol{x}) v_k(\boldsymbol{x}) \, d\boldsymbol{x}, \quad k = 1, \ldots, n, \tag{8.20}$$

denn sicher gilt

$$a(u_h, v_h) = F(v_h) \quad \text{für alle } v_h \in V_h$$

genau dann, wenn für alle Elemente der Basis $v_1, \ldots, v_n$

$$a(u_h, v_j) = F(v_j)$$

gilt.

Das Gleichungssystem (8.20) ist eindeutig lösbar, wenn nur $v_1, \ldots, v_n$ linear unabhängig sind, denn die Koeffizientenmatrix

$$\boldsymbol{A} := \Big(\int_\Omega \langle \nabla v_j(\boldsymbol{x}), \nabla v_k(\boldsymbol{x}) \rangle \, d\boldsymbol{x} \Big)_{j,k=1,\ldots,n}$$

ist positiv definit. Es gilt nämlich nach der Ungleichung von Poincaré

$$\begin{aligned} \boldsymbol{\xi}^T \boldsymbol{A} \boldsymbol{\xi} &= \int_\Omega \langle \textstyle\sum_{j=1}^n \xi_j \nabla v_j(\boldsymbol{x}), \sum_{k=1}^n \xi_k \nabla v_k(\boldsymbol{x}) \rangle \, d\boldsymbol{x} \\ &\geq \frac{1}{1 + 4R^2} \| \sum_{j=1}^{n} \xi_j v_j \|^2 > 0, \quad \text{falls } \sum_{j=1}^{n} \xi_j v_j \not\equiv 0. \end{aligned}$$

Tabelle 8.1: Kondition in Beispiel 8.11.

n	Kondition
2	3.85 E 02
3	1.02 E 03
4	2.88 E 04
5	8.49 E 05
10	2.56 E 13
20	3.65 E 17

In der Praxis bereitet das Ritz–Galerkin Verfahren häufig große Schwierigkeiten:

(i) Es ist bei nicht ganz einfachen Gebieten Ω schwierig, geeignete Funktionen $v_1, \ldots, v_n$ zu finden, die die Randbedingungen erfüllen.

(ii) Die Aufstellung des Gleichungssystems erfordert wegen der dort auftretenden Integrationen über Ω großen Rechenaufwand.

(iii) Das Gleichungssystem ist sehr schlecht konditioniert, wenn man nicht sehr spezielle Basen $v_1, \ldots, v_n$ verwendet.

Beispiel 8.11.

$$\begin{aligned} -\Delta u &= f(x,y) \quad \text{in } \Omega := \{(x,y) \,:\, x^2+y^2<1\} \\ u &= 0 \quad \text{auf } \partial\Omega \end{aligned}$$

Die variationelle Form lautet:

$$u \in H_0^1(\Omega) \,:\, \int_\Omega \langle \nabla u, \nabla v\rangle \, d(x,y) = \int_\Omega f v \, d(x,y) \quad \text{für alle } v \in H_0^1(\Omega).$$

Mit den Ansatzfunktionen

$$v_j(x,y) := 1 - \sqrt{x^2+y^2}^{\,j}, \; j = 1, \ldots, n$$

lautet die Koeffizientenmatrix der Ritz – Galerkin Gleichungen

$$A = 2\pi\Big(\frac{ij}{i+j}\Big)_{i,j=1,\ldots,n},$$

und diese besitzt die Kondition in Tabelle 8.1. □

Die Schwierigkeiten werden weitgehend vermieden, wenn man V_h nach der **Methode der finiten Elemente** konstruiert. Wir beschreiben kurz den einfachsten Fall.

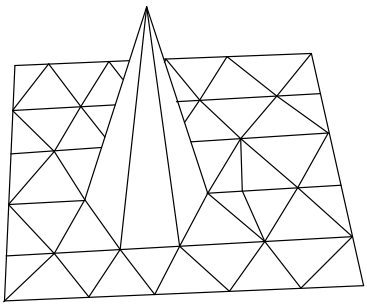

Abbildung 8.1: Dachfunktion

Es sei Ω ein polygonal berandetes Gebiet. Wir betrachten eine zulässige Zerlegung von Ω in Dreiecke (vgl. Abschnitt 7.5). Die Ecken der Dreiecke nennen wir Knoten.

Wir wählen V_h als die Menge aller auf $\bar{\Omega}$ stetigen Funktionen, die auf $\partial\Omega$ verschwinden und die auf jedem Dreieck mit einer in x und y linearen Funktion übereinstimmen.

Offenbar sind die Funktionen aus V_h eindeutig bestimmt, wenn man ihre Werte an den Knoten kennt (man erhält dann das Element aus V_h durch lineare Interpolation in jedem Dreieck).

Sind $P_1, \ldots, P_n$ die in $\bar{\Omega} \setminus \partial\Omega$ liegenden Knoten, so kann man in natürlicher Weise eine Basis $v_1, \ldots, v_n$ von V_h konstruieren durch die Forderung

$$v_k(P_j) = \delta_{kj}, \quad k, j = 1, \ldots, n \ .$$

Diese Basisfunktionen heißen wieder **Dachfunktionen** oder **hat functions**. Abbildung 8.1 zeigt den Graphen einer Dachfunktion.

Jedes $v \in V_h$ hat die Darstellung

$$v(x, y) = \sum_{k=1}^{n} v(P_k) v_k(x, y) \ .$$

Verwendet man diese Basis, so sind also die Entwicklungskoeffizienten gerade die Funktionswerte an den Knoten. Eine wichtige Eigenschaft dieser Basis ist, dass v_k nur in einer kleinen Umgebung von P_k von Null verschieden ist.

Weitere Vorteile dieser Basis sind

– v_j ist auf allen Dreiecken identisch Null, die P_k nicht als Ecke besitzen. Damit ist das Produkt $v_j v_k$ identisch Null, wenn es kein Dreieck gibt, in dem beide Knoten, P_j und P_k, liegen. Dasselbe gilt für $\langle \nabla v_j, \nabla v_k \rangle$.
Daher sind alle Elemente a_{jk} der Koeffizientenmatrix 0, für die es kein Dreieck gibt, in dem sowohl P_j als auch P_k liegt.
Die Matrix $\boldsymbol{A}$ ist also dünn besetzt. In jeder Zeile sind nur wenige Elemente von 0 verschieden. Das spart Speicherplatz und kann beim Lösungsprozess genutzt werden.

– Da $\boldsymbol{A}$ positiv definit ist und dünn besetzt, kann man iterative Methoden zur Lösung des linearen Systems verwenden (z.B. cg Verfahren mit Präkonditionierung).

– Da $\boldsymbol{A}$ dünn besetzt ist, hat man nur $O(n)$ (nicht n^2) Matrixelemente durch Integration zu berechnen. Zudem ist der Integrationsbereich einfach (Dreiecke), was die Integration erleichtert.

– Die Kondition wächst wesentlich langsamer mit der Dimension als für die meisten nicht lokalen Basen.

Beispiel 8.12. Wir betrachten die Randwertaufgabe

$$-\Delta u = f(x,y) \text{ in } D := (0,1) \times (0,1) \;, \quad u = 0 \text{ auf } \partial D. \tag{8.21}$$

Wir zerlegen Ω zunächst in Quadrate der Seitenlänge $h = 1/(m+1)$ und danach jedes Quadrat in zwei Dreiecke (vgl. Abbildung 8.2, links).

Wir können leicht das zugehörige Gleichungssystem aufstellen: Besitzt P_k die Koordinaten (ih, jh) und bezeichnet D_ℓ, $\ell = 1, \ldots, 6$, die Dreiecke, auf denen die zu P_k gehörende Basisfunktion v_k nicht verschwindet (vgl. Abbildung 8.2, rechts), so gilt z.B. in

$$D_1 : u(x,y) = \left(1 + i - \frac{x}{h}\right) u_{ij} + \left(\frac{x}{h} - i - \frac{y}{h} + j\right) u_{i+1,j} + \left(\frac{y}{h} - j\right) u_{i+1,j+1}.$$

Also ist

$$\begin{aligned} \nabla u^T \nabla v_k &= \left(-\frac{1}{h} u_{ij} + \frac{1}{h} u_{i+1,j} \,,\, -\frac{1}{h} u_{i+1,j} + \frac{1}{h} u_{i+1,j+1}\right) \begin{pmatrix} -1/h \\ 0 \end{pmatrix} \\ &= \frac{1}{h^2} \left(u_{ij} - u_{i+1,j}\right), \end{aligned}$$

und daher

$$\int_{D_1} (\nabla u)^T \nabla v_k \, dx \, dy = \frac{1}{2} \left(u_{ij} - u_{i+1,j}\right).$$

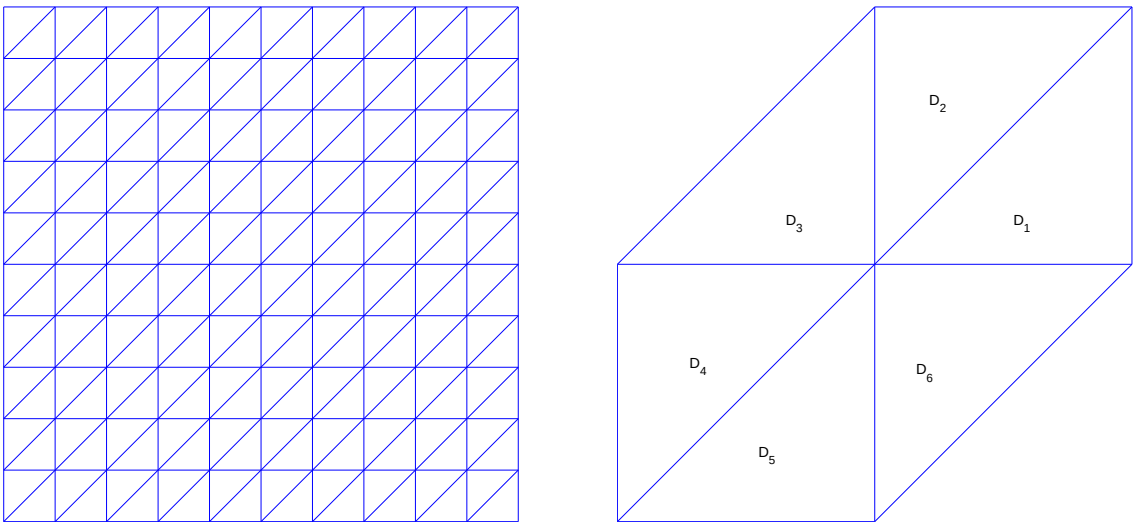

Abbildung 8.2: Zu Beispiel 8.12.

Behandelt man $D_2, \ldots, D_6$ entsprechend, so erhält man

$$\begin{aligned}\int_D (\nabla u)^T \nabla v_k \, dx\, dy &= \sum_{\ell=1}^{6} \int_{D_\ell} (\nabla u)^T \nabla v_k \, dx\, dy \\ &= \frac{1}{2}\Big\{ (u_{ij} - u_{i+1,j}) + (u_{ij} - u_{i,j+1}) + (u_{ij} - u_{i-1,j} - u_{i,j+1} + u_{ij}) \\ &\quad + (u_{ij} - u_{i-1,j}) + (u_{ij} - u_{i,j-1}) + (u_{ij} - u_{i,j-1} - u_{i+1,j} + u_{ij}) \Big\} \\ &= 4u_{ij} - u_{i+1,j} - u_{i,j+1} - u_{i-1,j} - u_{i,j-1}.\end{aligned}$$

Für die rechte Seite von (8.7) erhält man

$$\int_D f(x,y) v_k(x,y)\, d(x,y) = \sum_{\ell=1}^{6} \int_{D_\ell} f(x,y) v_k(x,y)\, d(x,y) =: f_k.$$

Die Ritz–Galerkin Gleichungen

$$4u_{ij} - u_{i-1,j} - u_{i+1,j} - u_{i,j-1} - u_{i,j+1} = f_k, \quad i,j = 1, \ldots, m,$$

haben also genau dieselbe Koeffizientenmatrix wie die Diskretisierung mit dem Differenzenverfahren.

Die Konditionen der Koeffizientenmatrizen (und die Dimensionen) für verschiedene Schrittweiten sind in Tabelle 8.2 enthalten. Sie wachsen viel langsamer als bei den üblichen globalen Ansatzfunktionen. □

8.3 Fehlerabschätzung

Bei den Differenzenverfahren lagen die Lösung der Randwertaufgabe und die Lösungen der diskretiserten Probleme in verschiedenen Räumen, und wir haben die Lösun-

Tabelle 8.2: Kondition in Beispiel 8.16.

m	Kondition	Dimension
1	1.00 E 00	1
3	5.83 E 00	9
7	2.53 E 01	49
15	1.03 E 02	225
31	4.14 E 02	961
63	1.66 E 03	3969

gen der diskreten Probleme mit den Restriktionen der Lösung des kontinuierlichen Problems verglichen. Im Fall des Ritz – Galerkin Verfahrens kann man die Lösung und die Näherungslösungen direkt vergleichen.

Grundlage der Fehlerabschätzungen ist der folgende Satz, der häufig als **Lemma von Cea** bezeichnet wird.

Satz 8.13. *Es sei $a(\cdot,\cdot)$ eine stetige und V-elliptische Bilinearform*

$$\alpha\|u\|^2 \le a(u,u),\ |a(u,v)| \le M\|u\|\cdot\|v\| \quad \textit{für alle } u,v \in V.$$

Dann sind für jedes stetige, lineare Funktional F auf V die Variationsgleichungen (8.4) und (8.19) eindeutig lösbar, und für die Lösungen $u \in V$ und $u_h \in V_h$ gilt

$$\|u-u_h\| \le \frac{M}{\alpha} \inf_{v_h\in V_h} \|u-v_h\|. \tag{8.22}$$

Beweis: Die eindeutige Lösbarkeit der beiden Variationsgleichungen folgt aus dem Satz von Lax – Milgram.

Wegen $V_h \subset V$ folgt aus (8.4)

$$a(u,v_h) = F(v_h) \quad \text{für alle } v_h \in V_h,$$

und unter Beachtung der Linearität von a erhält man aus (8.19)

$$a(u-u_h,v_h) = 0 \quad \text{für alle } v_h \in V_h,$$

und damit insbesondere

$$a(u-u_h,u_h) = 0.$$

Daher gilt

$$a(u-u_h,u-u_h) = a(u-u_h,u-v_h) \quad \text{für alle } v_h \in V_h,$$

und die Stetigkeit und V–Elliptizität von a liefert

$$\alpha\|u-u_h\|^2 \le M\|u-u_h\|\cdot\|u-v_h\| \quad \text{für alle } v_h \in V_h,$$

woraus man die Ungleichung (8.22) erhält. ∎

Werden die Ansatzräume V_h so gewählt, dass sie asymptotisch dicht in V liegen, gibt es zu jedem $z \in V$ und jedem $\varepsilon > 0$ ein V_h ($h > 0$ genügend klein) und ein $z_h \in V_h$, so dass $\|z - z_h\| < \varepsilon$, so folgt aus dem Lemma von Cea Konvergenz des zugehörigen Ritz–Galerkin Verfahrens:

$$\lim_{h \to 0+0} \|u - u_h\| = 0.$$

Der Abstand von u vom Ansatzraum V_h in (8.22) wird häufig mit Hilfe einer Projektion von u auf V_n, z.B. eines Interpolationsoperators $\Pi_h : V \to V_n$ abgeschätzt durch

$$\inf_{v_h \in V_h} \|u - v_h\| \leq \|u - \Pi_h u\|.$$

Hiermit erhält man z.B. das folgende Resultat

Satz 8.14. *Es sei $\Omega \subset \mathbb{R}^2$ ein polygonal berandetes Gebiet. Die (verallgemeinerte) Lösung u der Randwertaufgabe*

$$-\Delta u = f \text{ in } \Omega, \quad u = 0 \text{ auf } \partial\Omega$$

besitze (verallgemeinerte) Ableitungen zweiter Ordnung, die quadratisch integrierbar seien.

Dann gilt für die mit der Methode der finiten Elemente mit stückweise linearen C^0-Ansätzen über Dreieckszerlegungen der Feinheit h erzeugte Näherungslösung u_h die Abschätzung

$$\|u - u_h\|_{1,\Omega} \leq Ch|u|_{2,\Omega}$$

mit einer Konstante $C > 0$, falls die auftretenden Winkel in den beteiligten Dreiecken bei der Verfeinerung nicht beliebig spitz werden (eine solche Zerlegungsfolge heißt **quasi–gleichmäßig***).*

Dabei ist

$$\|u\|_{1,\Omega} = \left\{ \int_\Omega (u^2 + u_x^2 + u_y^2)\, d(x,y) \right\}^{1/2},$$

$$|u|_{2,\Omega} = \left\{ \int_\Omega (u_{xx}^2 + u_{xy}^2 + u_{yy}^2)\, d(x,y) \right\}^{1/2}.$$

Beweis: s. *Großmann, Roos* [26], p. 179. ∎

Tabelle 8.3: Fehler und Fehlerasymptotik

n	Fehler	Fehler$/(h^2\lvert\ln h\rvert)$
144	$5.241\ E-03$	0.700
541	$1.610\ E-03$	0.698
2097	$4.725\ E-04$	0.690
8257	$1.350\ E-04$	0.681
32769	$3.925\ E-05$	0.697

Bemerkung 8.15. Dass der Fehler nur wie $C \cdot h$ gegen 0 geht scheint im Widerspruch zu stehen zu Beispiel 8.16., nach dem die Koeffizientenmatrix des diskreten Problems mit der Matrix übereinstimmt, die man beim Differenzenverfahren erhält, und Satz 7.5. nach dem der Fehler quadratisch gegen 0 geht. Man beachte aber, dass die Norm in Satz 8.14. die Sobolewnorm in $H^1(\Omega)$ ist, in der auch die ersten Ableitungen von u berücksichtigt sind.

Ist das Gebiet Ω polygonal berandet und konvex, so kann man zeigen, dass

$$\|u-u_h\|_0 := \sqrt{\int_\Omega (u(x,y)-u_h(x,y))^2\, d(x,y)} \le Ch^2|u|_2, \tag{8.23}$$

mit einer von f unabhängigen Konstante C gilt, und hieraus erhält man mit Hilfe von inversen Ungleichungen die gleichmäßige Fehlerabschätzung

$$\|u-u_h\|_\infty \le Ch|u|_2. \tag{8.24}$$

Wesentlich aufwendiger ist es, die Konvergenzordnung

$$\|u-u_h\|_\infty \le Ch^2|\ln h| \tag{8.25}$$

zu beweisen. Diese ist nicht mehr verbesserbar. □

Beispiel 8.16. Wir betrachten die Randwertaufgabe

$$-\Delta u(x,y) = -4 \text{ in } \Omega := \{(x,y) : x^2+y^2<1\}, \quad u = 1 \text{ auf } \partial\Omega \tag{8.26}$$

mit der Lösung

$$u(x,y) = x^2+y^2.$$

Wir diskretisieren (8.26) mit linearen Elementen auf Dreiecken. Dann erhält man die Fehler in Tabelle 8.3. Die Fehlerasymptotik in (8.25) wird gut bestätigt. □

8.4 Realisierung von Verfahren der finiten Elemente

Ein Finite – Elemente –Programm besteht aus:

- Preprocessing:
  - Beschreibung der Geometrie
  - Eingabe der Koeffizientenfunktionen
  - Eingabe der Randbedingungen
  - Gittergenerierung
  - Bestimmung des endlich dimensionalen Problems (Steifigkeitsmatrix, rechte Seite)
- Lösung des endlich dimensionalen Problems
- Postprocessing
  - Aufbereitung der erhaltenen Ergebnisse
  - Ableitung mittelbarer Resultate
  - Graphische Darstellung der Lösung
  - Bewertung der Ergebnisse

Bei modernen Programmen sind die ersten beiden Schritte nicht mehr getrennt. Aus dem Lösungsprozess werden Informationen gezogen, um das Gitter lokal zu verfeinern.

Zur Erzeugung eines (Ausgangs–) Gitters wird das Grundgebiet Ω zunächst mit einem regelmäßigen Gitter überzogen, und dieses wird durch Verschiebung von randnahen Gitterpunkten lokal an die tatsächliche Geometrie angepasst.

Abbildung 8.3 zeigt ein regelmäßiges Startgitter und Abbildung 8.4 das daraus erzeugte Gitter auf dem konvexen Gebiet Ω.

Unter Verwendung von lokalen Fehlerschätzungen (vgl. *Großmann, Roos* [26], p. 234) wird im Lösungsprozess erkannt, wo das Gitter verfeinert werden muss, um die gewünschte Genauigkeit zu erzielen. Das Ziel ist es, mit festem Rechenaufwand eine

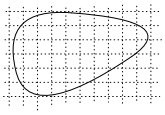

Abbildung 8.3: Regelmäßiges Gitter

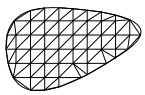

Abbildung 8.4: Startgitter

möglichst hohe Genauigkeit zu erzielen, oder umgekehrt eine gewünschte Genauigkeit mit möglich geringem Aufwand zu erreichen. Dazu werden die Zerlegungen so gesteuert, dass alle Dreiecke ungefähr gleichen Beitrag zum Gesamtfehler leisten.

Durch Halbierung aller Seiten wird ein Dreieck in vier kongruente Dreiecke unterteilt und es werden die drei Nachbardreiecke so halbiert, dass die gesamte Zerlegung zulässig ist. In Abbildung 8.5 ist das dick markierte Dreieck in vier kongruente Vierecke zerlegt.

Um zu vermeiden, dass bei weiterer Verfeinerung Dreiecke mit sehr spitzen Winkeln entstehen, werden zusätzlich die halbierten Dreiecke markiert. Ist im Verlauf der weiteren Verfeinerung die erneute Halbierung eines markierten Dreiecks erforderlich, so wird die vorhergehende Halbierung durch eine Zerlegung in vier kongruente Dreiecke ersetzt (vgl. Abbildung 8.6).

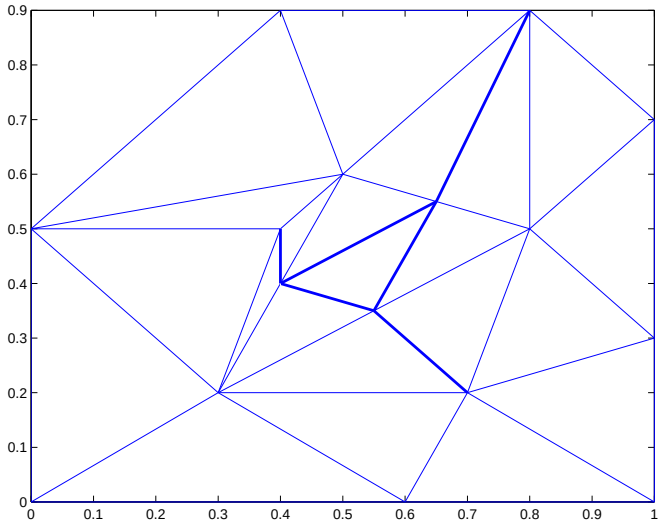

Abbildung 8.5:Gitterverfeinerung

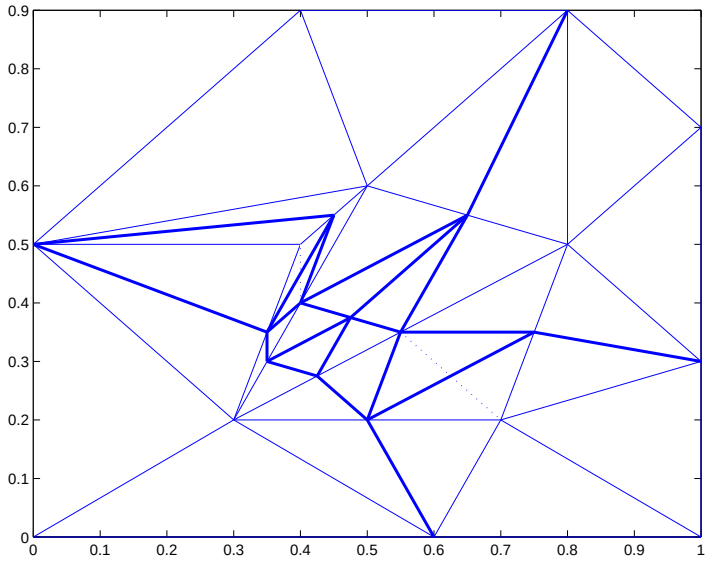

Abbildung 8.6: Regelmäßiges Gitter

Beispiel 8.17. Der Bereich Ω sei in Polarkoordinaten beschrieben durch

$$\Omega := \{(r, \varphi) \,:\, 0 < r < 1,\ -0.75\pi < \varphi < 0.75\pi\}.$$

Wir betrachten die Randwertaufgabe

$$-\Delta u = 0, \text{ in } \Omega,$$

$$u = 0, \text{ für } 0 \le r \le 1,\ \varphi \in \{-0.75\pi, 0.75\pi\},$$
$$u = \cos\frac{2\varphi}{3} \text{ für } r = 1,\ -0.75\pi \le \varphi \le 0.75\pi$$

mit der Lösung

$$u(r, \varphi) = r^{2/3} \cos\frac{2\varphi}{3}. \tag{8.27}$$

Tabelle 8.4: Fehler und Dimensionen in Beispiel 8.11.

Stufe	Fehler	Dimension
0	$1.72\ E-02$	118
1	$1.16\ E-02$	434
2	$7.46\ E-03$	1663
3	$4.74\ E-03$	6509

Löst man die Randwertaufgabe mit stückweise linearen finiten Elementen mit der PDE Toolbox von MATLAB, so erhält man bei gleichmäßiger Verfeinerung die maximalen Fehler in den Knoten in der Tabelle 8.4. Dabei werden diskrete Probleme der angegebenen Dimensionen gelöst. Wählt man eine adaptive Verfeinerung ausgehend von dem gröbsten der oben verwendeten Gitter, so erhält eine maximalen Fehler von $2.18\ E-03$ in den Knoten, wobei ein lineares Gleichungssystem mit 286 Unbekannten gelöst wird. Abbildung 8.7 enthält links das gleichmäßig verfeinerte Gitter der Stufe 1 und rechts das adaptiv verfeinerte Gitter.

□

8.5 Weitere ebene Elemente

Wir beschreiben in diesem Abschnitt weitere Elementtypen, die für ebene elliptische Probleme verwendet werden und die bessere Approximationseigenschaften haben als die stückweise linearen Elemente auf Dreieckszerlegungen.

8.5.1 Quadratische Ansätze in Dreiecken

Wir zerlege Ω wie vorher in Dreiecke T_j. Knoten sind die Eckpunkte und die Seitenmittelpunkte der Dreiecke, in denen die Funktionswerte vorgegeben werden.

Die Restriktionen der Ansatzfunktionen auf die Dreiecke sind quadratische Polynome

$$u\Big|_{D_j}(x,y) = \alpha_j + \beta_j x + \gamma_j y + \delta_j x^2 + \epsilon_j xy + \zeta_j y^2.$$

Die sechs Parameter $\alpha_j, \ldots, \zeta_j$ können bei gegeben Werten in den Knoten eindeutig durch Interpolation bestimmt werden.

Die zusammengesetzte Funktion ist stetig, denn über jeder Dreiecksseite ist u ein quadratisches Polynom von einer Variablen, und dieses ist eindeutig durch die Funktionswerte in den Endpunkten und dem Mittelpunkt der Seite bestimmt.

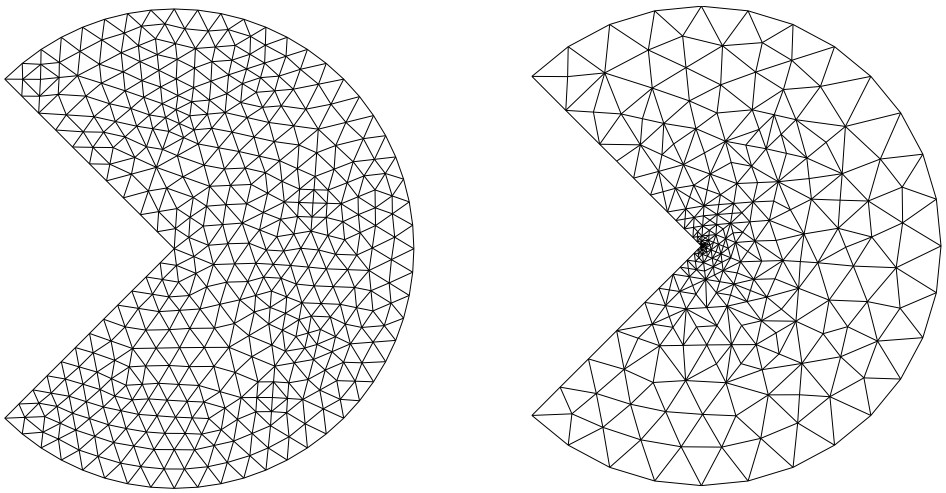

Abbildung 8.7: Verwendete Gitter in Beispiel 8.17.

8.5.2 Kubische Ansätze in Dreiecken

Wir zerlegen Ω wie vorher in Dreiecke T_j.

Knoten sind die Eckpunkte und der Schwerpunkt jedes Dreiecks. In den Eckpunkten werden der Funktionswert u und die beiden Ableitungen u_x und u_y vorgegeben, im Schwerpunkt nur der Funktionswert.

Die Restriktionen der Ansatzfunktionenen auf die Dreiecke sind kubische Polynome

$$u\Big|_{D_j}(x,y) = \alpha_0 + \alpha_1 x + \beta_1 y + \alpha_2 x^2 + \beta_2 xy + \gamma_2 y^2 + \alpha_3 x^3 + \beta_3 x^2 y + \gamma_3 x y^2 + \delta_3 y^3.$$

Die zehn Parameter $\alpha_0, \ldots, \delta_3$ können bei gegeben Werten in den Knoten eindeutig durch Interpolation bestimmt werden.

Die zusammengesetzte Funktion ist stetig, denn über jeder Dreiecksseite ist u ein kubisches Polynom von einer Variablen, und dieses ist eindeutig durch die Funktionswerte und Ableitungswerte (in Richtung der Seite) in den Endpunkten der Seite bestimmt.

8.5.3 Bilineare Ansätze in Rechtecken

Wir zerlegen Ω in Rechtecke Q_j.

Knoten sind die Eckpunkte jedes Rechtecks, in denen die Funktionswerte vorgegeben werden.

Die Restriktionen der Ansatzfunktionenen auf die Rechtecke sind bilineare Funktionen

$$u\Big|_{D_j}(x,y) = \alpha_j + \beta_j x + \gamma_j y + \delta_j xy.$$

Die vier Parameter $\alpha_j, \ldots, \delta_j$ können bei gegeben Werten in den Knoten eindeutig durch Interpolation bestimmt werden.

Die zusammengesetzte Funktion ist stetig, denn über jeder Rechtecksseite ist u ein lineares Polynom von einer Variablen, und dieses ist eindeutig durch die Funktionswerte in den Endpunkten der Seite bestimmt.

8.6 Software

Vom Rechenzentrum der TU Hamburg–Harburg werden die kommerziellen Software Pakete **MARC/MENTAT** und **ANSYS** und **I–DEAS** bereitgehalten. Die Beratung hierfür wird von Frau Bredehöft (Tel. 2526) durchgeführt. Eine Kurzbeschreibung findet man im WWW unter

http://www.tu-harburg.de/rzt/tuinfo/software/fem/

Wir erwähnen besonders ein Public Domain Projekt:

KASKADE wird am Konrad Zuse Zentrum, Berlin, entwickelt. Es ist ein in C (Version 2.1) bzw. C++ (Version 3.1) geschriebenes Paket zur Lösung elliptischer und parabolischer Differentialgleichungen unter Benutzung von räumlich adaptiven finiten Elementen und von Mehrgittermethoden. Die Software und Dokumentation können geladen werden von

http://www.zib.de/en/numerik/software/kaskade-7.html

Eine sehr schnelle Möglichkeit zur Lösung einfacher partieller Differentialgleichungsprobleme (elliptische Randwertaufgaben, Eigenwertaufgaben, parabolische und hyperbolische Anfangsrandwertaufgaben) mit linearen Elementen auf Dreiecken bietet die PDE Toolbox zu MATLAB. Sie bietet eine sehr bequeme graphische Benutzeroberfläche zur Beschreibung der Geometrie des Problems und zur Eingabe der Differentialgleichung und der Randbedingungen.

COMSOL ist ein Programmpaket zur numerischen Lösung gewöhnlicher und partieller Differentialgleichungen mit der Methode der finiten Elemente unter Verwendung von adaptiver Gitterverfeinerung und automatischer Fehlerkontrolle. Es gibt auch eine Version, die auf MATLAB aufsetzt. COMSOL ist im Rechenzentrum vorhanden.

Kapitel 9

Parabolische Anfangsrandwertaufgaben

9.1 Differenzenverfahren

Wir betrachten die Anfangsrandwertaufgabe der Wärmeleitungsgleichung

$$\left.\begin{array}{rl} u_t - u_{xx} = f(x,t), & 0 < x < 1,\ t > 0 \\ u(x,0) = u_0(x), & 0 < x < 1 \\ u(0,t) = u(1,t) = 0, & t > 0. \end{array}\right\} \tag{9.1}$$

Wir beschränken uns auf dieses Modellproblem. Es ist klar, wie man die Ergebnisse auf allgemeinere Differentialoperatoren und allgemeinere Randbedingungen übertragen kann.

Es ist naheliegend, die auftretenden partiellen Ableitungen durch Differenzenquotienten zu ersetzen. Approximiert man u_{xx} durch den zentralen Differenzenquotienten

$$u_{xx}(x,t) \approx \frac{1}{h^2}\Big(u(x+h,t) - 2u(x,t) + u(x-h,t)\Big)$$

und u_t durch den vorwärtsgenommenen Differenzenquotienten

$$u_t(x,t) \approx \frac{1}{k}\Big(u(x,t+k) - u(x,t)\Big)$$

so erhält man für die Näherungen

$$U_{ij} \approx u(ih, jk)\,,\ h := \frac{1}{n},\ k > 0,\ i = 1, \ldots, n-1,\ j = 0, 1, 2, \ldots$$

Tabelle 9.1: Vorwärts-Dz.-Verfahren, $h = 0.125, k = 0.1$

t	$u(t, 0.000)$	$u(t, 0.125)$	$u(t, 0.250)$	$u(t, 0.375)$	$u(t, 0.500)$
0.0	$0.0E+00$	$0.0E+00$	$0.0E+00$	$0.0E+00$	$1.0E-10$
0.1	$0.0E+00$	$0.0E+00$	$0.0E+00$	$6.4E-10$	$-1.2E-09$
0.2	$0.0E+00$	$0.0E+00$	$4.1E-09$	$-1.5E-08$	$2.2E-08$
0.3	$0.0E+00$	$2.6E-08$	$-1.4E-07$	$3.5E-07$	$-4.5E-07$
0.4	$0.0E+00$	$-1.2E-06$	$4.1E-06$	$-7.9E-06$	$9.8E-06$
...	.....	.....	.....	.....	.....
1.7	$0.0E+00$	$2.1E+12$	$-3.9E+12$	$5.1E+12$	$-5.6E+12$
1.8	$0.0E+00$	$-5.0E+13$	$9.3E+13$	$-1.2E+14$	$1.3E+14$
1.9	$0.0E+00$	$1.2E+15$	$-2.2E+15$	$2.9E+15$	$-3.1E+15$
2.0	$0.0E+00$	$-2.8E+16$	$5.2E+16$	$-6.8E+16$	$7.3E+16$

die Differenzengleichung

$$\frac{U_{i,j+1} - U_{ij}}{k} - \frac{U_{i+1,j} - 2U_{ij} + U_{i-1,j}}{h^2} = f_{ij},$$

oder mit $r := k/h^2$

$$U_{i,j+1} = rU_{i-1,j} + (1-2r)U_{ij} + rU_{i+1,j} + kf_{ij}. \tag{9.2}$$

Da aus den Anfangsbedingungen die Werte U_{i0}, $i = 0, \ldots, n$, bekannt sind, kann man U_{i1}, $i = 1, \ldots, n-1$, berechnen. Die Randbedingungen liefern $U_{01} = 0 = U_{n1}$. Damit hat man eine Näherung für u auf der ersten Zeitschicht $t = k$. Danach kann man genauso nacheinander Näherungen auf den Zeitschichten $t = jk$, $j = 2, 3, \ldots$, ermitteln. Das Verfahren (9.2) heißt **explizites Differenzenverfahren** .

Den lokalen Fehler des Verfahrens erhält man durch Taylorentwicklung. Ist u viermal stetig partiell differenzierbar nach x und zweimal stetig partiell differenzierbar nach t, so gilt

$$\begin{aligned}
&\varepsilon_{h,k}(x_i, t_j) \\
&:= \frac{u(x_i, t_j + k) - u(x_i, t_j)}{k} - \frac{u(x_i - h, t_j) - 2u(x_i, t_j) + u(x_i + h, t_j)}{h^2} - f(x_i, t_j) \\
&= u_t(x_i, t_j) + O(k) - u_{xx}(x_i, t_j) + O(h^2) + f(x_i, t_j) = O(k) + O(h^2).
\end{aligned}$$

Wir wenden das Verfahren (9.2) nun auf (9.1) an mit der rechten Seite $f(x,t) \equiv 0$ und der Anfangsfunktion $u_0(x) \equiv 0$ (und der Lösung $u(x,t) \equiv 0$). Dann liefert (9.2) offenbar $U_{ij} = 0$ für alle i, j. Mit $h = 0.125$, $k = 0.1$ und der Störung $U_{4,0} = 1E-10$ erhält man die (unbrauchbaren) Ergebnisse aus Tabelle 9.1. Der Anfangsfehler schaukelt sich auf, und das Verfahren ist offenbar instabil.

Tatsächlich gilt mit $r = k/h^2$

$$|U_{i,j+1}| = |rU_{i-1,j} + (1-2r)U_{ij} + rU_{i+1,j} + kf_{ij}|$$

Tabelle 9.2: Vorwärts-Dz.-Verfahren, $h = 0.125, k = 0.0075$

t	$u(t, 0.000)$	$u(t, 0.125)$	$u(t, 0.250)$	$u(t, 0.375)$	$u(t, 0.500)$
0.0000	$0.0E+00$	$0.0E+00$	$0.0E+00$	$0.0E+00$	$1.0E-10$
0.0075	$0.0E+00$	$0.0E+00$	$0.0E+00$	$4.6E-11$	$8.0E-12$
0.0150	$0.0E+00$	$0.0E+00$	$2.1E-11$	$7.4E-12$	$4.3E-11$
0.0225	$0.0E+00$	$9.7E-12$	$5.1E-12$	$3.0E-11$	$1.0E-11$
0.0300	$0.0E+00$	$3.1E-12$	$1.9E-11$	$9.4E-12$	$2.8E-11$
0.0375	$0.0E+00$	$8.9E-12$	$7.3E-12$	$2.2E-11$	$1.1E-11$
.....	.....	.....	.....	.....	.....
0.7200	$0.0E+00$	$9.0E-15$	$1.7E-14$	$2.2E-14$	$2.3E-14$
0.7275	$0.0E+00$	$8.4E-15$	$1.5E-14$	$2.0E-14$	$2.2E-14$
0.7350	$0.0E+00$	$7.8E-15$	$1.4E-14$	$1.9E-14$	$2.0E-14$
0.7425	$0.0E+00$	$7.2E-15$	$1.3E-14$	$1.7E-14$	$1.9E-14$
0.7500	$0.0E+00$	$6.7E-15$	$1.2E-14$	$1.6E-14$	$1.8E-14$

d.h. für $1 - 2r \geq 0$

$$\max_i |U_{i,j+1}| \leq \max_i |U_{i,j}| + k \max_i |f_{ij}|,$$

und durch mehrfache Anwendung dieser Ungleichung erhält man

$$\max_{1\leq j\leq m} \max_{1\leq i\leq n-1} |U_{i,j}| \leq \|u_0\|_\infty + k \sum_{j=1}^{m-1} \max_i |f_{ij}|.$$

Damit ist gezeigt, dass das explizite Differenzenverfahren unter der Bedingung $1 - 2r \geq 0$, d.h.

$$k \leq \frac{1}{2}h^2 \tag{9.3}$$

stabil ist. Genauer haben wir gezeigt: Fasst man die Näherungswerte der j–ten Zeitschicht zu dem Vektor

$$\boldsymbol{U}^j = (U_{1j}, \ldots, U_{n-1,j})^T$$

zusammen, so bleibt unter der Stabilitätsbedingung (9.3) die Maximumnorm dieser Vektoren beschränkt. Das Verfahren ist also stabil bzgl. der (diskreten) Maximumnorm.

In dem obigen Beispiel ist die Stabilitätsungleichung mit $h = 0.125$ und $k = 0.1$ verletzt. Für $h = 0.125$ und $k = 0.0075$ ist sie erfüllt. Man erhält die Ergebnisse aus Tabelle 9.2, und das Verfahren ist für diese Schrittweiten wie erwartet stabil.

Die Beschränkung (9.3) der Zeitschrittweite k ist bei kleinen Ortsschrittweiten h außerordentlich restriktiv und der Rechenaufwand, um eine feste Zeitschicht $t = T$ zu erreichen, wird unannehmbar groß. Um ein besseres Stabilitätsverhalten zu erreichen, diskretisieren wir nun u_t durch den rückwärtsgenommenen Differenzenquotienten:

$$u_t(x, t) \approx \frac{u(x, t) - u(x, t - k)}{k}.$$

Man erhält dann die Differenzengleichung (auf der Zeitschicht $t = (j+1)k$)

$$\frac{U_{i,j+1} - U_{ij}}{k} - \frac{U_{i-1,j+1} - 2U_{i,j+1} + U_{i+1,j+1}}{h^2} = f_{i,j+1},$$

d.h. mit der Matrix $\boldsymbol{B}_{h,k} = \text{tridiag}\,(-r, 1+2r, -r)$

$$\boldsymbol{B}_{h,k}\boldsymbol{U}^{j+1} = \boldsymbol{U}^j + k\boldsymbol{f}^{j+1},\ j = 0, 1, 2, \ldots, \tag{9.4}$$

wobei wie vorher $\boldsymbol{U}^j := (U_{1j}, \ldots, U_{n-1,j})^T$ ist und $\boldsymbol{f}^j$ entsprechend definiert ist.

Anders als beim Verfahren (9.2) kann man hier nicht die Approximationen $U_{i,j+1}$ auf der Zeitschicht $t = (j+1)k$ nacheinander auf den bekannten Werten U_{ij} der Zeitschicht $t = jk$ berechnen, sondern man hat auf jeder Zeitschicht ein lineares Gleichungssystem zu lösen. Ein Verfahren vom Typ (9.4) heißt **implizit** .

Genau wie für das Verfahren (9.2) gilt für den lokalen Fehler des Verfahrens (9.4)

$$\delta(h,k) = O(k) + O(h^2).$$

Die Matrix $\boldsymbol{B}_{h,k}$ ist (vgl. Kapitel 6) inversmonoton, und mit $\boldsymbol{e} = (1, \ldots, 1)^T$ gilt $\boldsymbol{B}_{h,k}\boldsymbol{e} \geq \boldsymbol{e}$. Daher gilt nach Satz 6.10.

$$\|\boldsymbol{B}_{h,k}^{-1}\|_\infty \leq 1,$$

und es folgt aus (9.4)

$$\|\boldsymbol{U}^{j+1}\|_\infty \leq \|\boldsymbol{U}^j\|_\infty + k\|\boldsymbol{f}^{j+1}\|_\infty \leq \|u_0\|_\infty + k\sum_{\ell=0}^{j}\|\boldsymbol{f}^{\ell+1}\|_\infty.$$

Das Verfahren (9.4) ist also für alle Schrittweiten $h > 0$ und $k > 0$ in der diskreten Maximumnorm stabil.

Für das Beispiel erhält man bei der Schrittweite $k = 0.1$ die Fehlerfortpflanzung aus Tabelle 9.3.

Bessere Konvergenzeigenschaften als die Verfahren (9.2) und (9.4) hat das **Crank–Nicholson Verfahren** , das man als Kombination der obigen Methoden auffassen kann. Wir approximieren im Punkt $(ih, (j+0.5)k)$ die Ableitung u_{xx} durch den Mittelwert der zentralen Differenzenquotienten auf den Zeitschichten $t = jk$ und $t = (j+1)k$:

$$u_{xx} \approx \frac{1}{2h^2}(U_{i+1,j} - 2U_{ij} + U_{i-1,j} + U_{i+1,j+1} - 2U_{i,j+1} + U_{i-1,j+1})$$

Tabelle 9.3: Rückwärts-Dz.-Verfahren, $h = 0.125, k = 0.1$

t	$u(t, 0.000)$	$u(t, 0.125)$	$u(t, 0.250)$	$u(t, 0.375)$	$u(t, 0.500)$
0.0	$0.0E+00$	$0.0E+00$	$0.0E+00$	$0.0E+00$	$1.0E-10$
0.1	$0.0E+00$	$3.1E-12$	$6.7E-12$	$1.1E-11$	$1.8E-11$
0.2	$0.0E+00$	$2.2E-12$	$4.3E-12$	$6.0E-12$	$6.8E-12$
0.3	$0.0E+00$	$1.2E-12$	$2.3E-12$	$3.0E-12$	$3.3E-12$
0.4	$0.0E+00$	$6.3E-13$	$1.2E-12$	$1.5E-12$	$1.6E-12$
0.5	$0.0E+00$	$3.2E-13$	$5.9E-13$	$7.7E-13$	$8.3E-13$
0.6	$0.0E+00$	$1.6E-13$	$3.0E-13$	$3.9E-13$	$4.2E-13$
...	.....	.....	.....	.....	.....
9.5	$0.0E+00$	$8.2E-40$	$1.5E-39$	$2.0E-39$	$2.2E-39$
9.6	$0.0E+00$	$4.2E-40$	$7.7E-40$	$1.0E-39$	$1.1E-39$
9.7	$0.0E+00$	$2.1E-40$	$3.9E-40$	$5.1E-40$	$5.5E-40$
9.8	$0.0E+00$	$1.1E-40$	$2.0E-40$	$2.6E-40$	$2.8E-40$
9.9	$0.0E+00$	$5.4E-41$	$1.0E-40$	$1.3E-40$	$1.4E-40$
10.0	$0.0E+00$	$2.7E-41$	$5.1E-41$	$6.6E-41$	$7.2E-41$

und u_t durch den zentralen Differenzenquotienten

$$u_t \approx \frac{1}{k}(U_{i,j+1} - U_{ij}).$$

Dann erhält man

$$-rU_{i-1,j+1} + (2+2r)U_{i,j+1} - rU_{i+1,j+1} = rU_{i-1,j} + (2-2r)U_{ij} + rU_{i+1,j} + k(\boldsymbol{f}^j + \boldsymbol{f}^{j+1})$$

d.h. mit der Matrix $\boldsymbol{J} = \text{tridiag}\,(-1, 2, -1)$

$$(2\boldsymbol{E} + r\boldsymbol{J})\boldsymbol{U}^{j+1} = (2\boldsymbol{E} - r\boldsymbol{J})\boldsymbol{U}^j + k(\boldsymbol{f}^j + \boldsymbol{f}^{j+1}),\ j = 0, 1, \ldots. \tag{9.5}$$

Für den lokalen Fehler von (9.5) erhält man mit dem Taylorschen Satz

$$\varepsilon(k, h) = O(k^2) + O(h^2),$$

falls u dreimal stetig partiell differenzierbar nach t und viermal stetig partiell differenzierbar nach x ist.

Die Stabilität bzgl. der diskreten Maximumnorm erhält man wieder aus Satz 6.10.: $2\boldsymbol{E} + r\boldsymbol{J}$ ist inversmonoton mit

$$\|(2\boldsymbol{E} + r\boldsymbol{J})^{-1}\|_\infty \le \frac{1}{2}$$

(man verwende in Satz 6.10. den Vektor $\boldsymbol{w} = 0.5\boldsymbol{e}$), und daher folgt

$$\|\boldsymbol{U}^{j+1}\|_\infty \le \frac{1}{2}\|(2\boldsymbol{E} - r\boldsymbol{J})\boldsymbol{U}^j\|_\infty + \frac{k}{2}(\|\boldsymbol{f}^j\|_\infty + \|\boldsymbol{f}^{j+1}\|_\infty).$$

Für $r \le 1$ besitzt die Matrix $2\boldsymbol{E} - r\boldsymbol{J}$ nichtnegative Elemente, und man erhält hierfür

$$\|\boldsymbol{U}^{j+1}\|_\infty \le \|\boldsymbol{U}^j\|_\infty + \frac{k}{2}(\|\boldsymbol{f}^j\|_\infty + \|\boldsymbol{f}^{j+1}\|_\infty),$$

und damit

$$\max_{j=1,\ldots,m-1} \|\boldsymbol{U}^{j+1}\|_\infty \le \|u_0\|_\infty + \frac{k}{2}\sum_{j=0}^{m-1}(\|\boldsymbol{f}^j\|_\infty + \|\boldsymbol{f}^{j+1}\|_\infty).$$

Das Crank–Nicholson Verfahren ist also bzgl. der diskreten Maximumnorm stabil, falls $r \le 1$, d.h. für $k \le h^2$.

Für allgemeine lineare parabolische Aufgaben mit variablen Koeffizienten des Typs

$$\frac{\partial u}{\partial t} - \frac{\partial}{\partial x}\Big(a(x,t)\frac{\partial u}{\partial x}\Big) - b(x,t)\frac{\partial u}{\partial x} - c(x,t)u(x,t) = f(x,t) \tag{9.6}$$

können die Methoden zur Stabilitätsuntersuchung in der diskreten Maximumnorm übertragen werden (vgl. *Samarskij* [42]).

Für die Wärmeleitungsgleichung kann man die Stabilitätsfrage bzgl. der Euklidischen Norm auf folgende Weise behandeln, wobei wir uns der Einfachheit halber auf den Fall $f(x,t) \equiv 0$ zurückziehen: Das explizite Differenzenverfahren kann man schreiben als

$$\boldsymbol{U}^{j+1} = (\boldsymbol{E} - r\boldsymbol{J})\boldsymbol{U}^j =: \boldsymbol{A}\boldsymbol{U}^j = \boldsymbol{A}^{j+1}\boldsymbol{U}^0,$$

und es folgt

$$\|\boldsymbol{U}^{j+1}\|_2 \le \|\boldsymbol{A}\|_2^{j+1}\|\boldsymbol{U}^0\|_2.$$

Hieraus liest man unmittelbar ab, dass das Verfahren genau dann stabil ist in der Euklidischen Norm, wenn $|\lambda_\nu| \le 1$ für alle Eigenwerte λ_ν von $\boldsymbol{A}$ gilt (Beachten Sie, dass $\boldsymbol{A}$ symmetrisch ist und damit diagonalisierbar ist).

Wegen $\boldsymbol{A} = \boldsymbol{E} - r \cdot \boldsymbol{J}$ sind die Eigenwerte von $\boldsymbol{A}$

$$\lambda_\ell = 1 - r\mu_\ell, \quad \text{wobei } \mu_\ell = 2\left(1 - \cos\frac{\ell\pi}{n}\right), \quad \ell = 1, \ldots, n-1,$$

die Eigenwerte von $\boldsymbol{J}$ sind. Wegen $0 < \mu_\ell < 4$ gilt $1 - 4r < \lambda_\ell < 1$, und das Verfahren ist stabil, wenn $1 - 4r \ge -1$, d.h. wie für die diskrete Maximumnorm

$$k \le \frac{1}{2}h^2. \tag{9.7}$$

Das implizite Verfahren (9.4) ist (wie in der diskreten Maximumnorm) für alle Schrittweitenverhältnisse stabil, denn für die Eigenwerte λ von $\boldsymbol{B}_{h,k} = \boldsymbol{E} + r\boldsymbol{J}$ gilt $\lambda > 1$, und daher liegen alle Eigenwerte von $\boldsymbol{B}_{h,k}^{-1}$ im Intervall $(0,1)$.

Das Crank–Nicholson Verfahren (9.5) kann mit der Matrix

$$\boldsymbol{C} := (2\boldsymbol{E} + r\boldsymbol{J})^{-1}(2\boldsymbol{E} - r\boldsymbol{J})$$

Tabelle 9.4: Crank-Nicholson-Verfahren, $h = 0.125, k = 0.1$

t	$u(t, 0.000)$	$u(t, 0.125)$	$u(t, 0.250)$	$u(t, 0.375)$	$u(t, 0.500)$
0.0	$0.0E+00$	$0.0E+00$	$0.0E+00$	$0.0E+00$	$1.0E-10$
0.1	$0.0E+00$	$6.8E-12$	$1.6E-11$	$3.0E-11$	$-4.7E-11$
0.2	$0.0E+00$	$6.9E-13$	$-2.6E-12$	$-1.7E-11$	$4.6E-11$
0.3	$0.0E+00$	$-5.4E-13$	$2.6E-12$	$1.8E-11$	$-3.2E-11$
0.4	$0.0E+00$	$1.5E-12$	$1.1E-13$	$-1.4E-11$	$2.7E-11$
0.5	$0.0E+00$	$-1.4E-12$	$-7.3E-13$	$1.3E-11$	$-2.1E-11$
0.6	$0.0E+00$	$1.3E-12$	$1.4E-12$	$-1.1E-11$	$1.7E-11$
...	.....	.....	.....	.....	.....
9.5	$0.0E+00$	$1.8E-18$	$-3.4E-18$	$4.4E-18$	$-4.8E-18$
9.6	$0.0E+00$	$-1.6E-18$	$2.9E-18$	$-3.8E-18$	$4.1E-18$
9.7	$0.0E+00$	$1.3E-18$	$-2.4E-18$	$3.2E-18$	$-3.5E-18$
9.8	$0.0E+00$	$-1.1E-18$	$2.1E-18$	$-2.7E-18$	$3.0E-18$
9.9	$0.0E+00$	$9.5E-19$	$-1.8E-18$	$2.3E-18$	$-2.5E-18$
10.0	$0.0E+00$	$-8.1E-19$	$1.5E-18$	$-2.0E-18$	$2.1E-18$

geschrieben werden kann als $\boldsymbol{U}^{j+1} = \boldsymbol{C}\boldsymbol{U}^j$, $j = 0, 1, \ldots$. Daher sind die Eigenwerte λ_ℓ von $\boldsymbol{C}$ gegeben durch

$$\lambda_\ell = \frac{2 - r\mu_\ell}{2 + r\mu_\ell},$$

und wegen $\mu_\ell > 0$ gilt $\lambda_\ell \in (-1, 1)$ für alle $r > 0$.

Das Verfahren von Crank und Nicholson ist also bzgl. der Euklidischen Norm für alle Schrittweitenverhältnisse r stabil. Für die diskrete Maximumnorm ist dies nur für $r \leq 1$ der Fall.

Die Tabelle 9.4 enthält die Fehlerfortpflanzung des Verfahrens von Crank und Nicholson für $k = 0.1$ und $h = 0.125$.

Für Probleme der Raumdimension 2 geht man prinzipiell wie in dem eben betrachteten, eindimensionalen Fall vor. Das Wärmeleitungsproblem lautet hier

$$\left.\begin{aligned} u_t = \Delta u, &\quad (x, y, t) \in \Omega \times \mathbb{R}^+ \\ u(x, y, 0) = u_0(x, y), &\quad (x, y) \in \Omega \\ Ru(x, y, t) = 0, &\quad (x, y) \in \partial\Omega,\ t > 0, \end{aligned}\right\} \tag{9.8}$$

wobei $\Omega \subset \mathbb{R}^2$ ein beschränktes Gebiet ist und der Operator R eine der in Kapitel 7 besprochenen Randbedingungen beschreibt.

Wir diskretisieren bei festgehaltener Zeitschicht $t = \ell k$ den Differentialoperator Δu in Ω unter Benutzung der Randbedingung $Ru(x, y, \ell k) = 0$, $(x, y)^T \in \partial\Omega$, wie in Kapitel 7 auf einem quadratischen Gitter der Maschenweite h mit der 5-Punkt-Differenzenformel in inneren Gitterpunkten und durch Interpolation in den randnahen Gitterpunkten:

$$-\frac{1}{h^2}\boldsymbol{A}_h\boldsymbol{U}^\ell.$$

Ersetzt man u_t durch den vorwärtsgenommenen Differenzenquotienten, so erhält man das explizite Verfahren

$$\boldsymbol{U}^{\ell+1} = (\boldsymbol{E} - r\boldsymbol{A}_h)\,\boldsymbol{U}^{\ell}, \tag{9.9}$$

wobei $r := k/h^2$ und $\boldsymbol{U}^{\ell} := (U(ih, jh, \ell k))_{i,j}$.

Das Verfahren (9.9) ist stabil in der Euklidischen Norm, wenn alle Eigenwerte λ_ν der Matrix $\boldsymbol{E} - r\boldsymbol{A}_h$ dem Betrage nach kleiner als 1 sind.

Ist $\boldsymbol{A}_h$ symmetrisch und positiv definit mit den Eigenwerten $\mu_\nu > 0$, so gilt

$$-1 < \lambda_\nu = 1 - r\mu_\nu < 1$$

genau dann, wenn $r < 2/\max_\nu \mu_\nu$. Treten in $\boldsymbol{A}_h$ keine Randterme auf (z.B. wenn Ω ein Rechteck und h passend ist), so gilt nach dem Satz von Gerschgorin $\max_\nu \mu_\nu \leq 8$, und man erhält Stabilität für $r \leq \dfrac{1}{4}$, d.h. für

$$k \leq \frac{1}{4}h^2.$$

Unter dieser Bedingung sind alle Elemente von $\boldsymbol{E} - r\boldsymbol{A}_h$ nichtnegativ, und man erhält wie im eindimensionalen Fall auch bzgl. der Maximumnorm Stabilität.

Die Beschränkung der Zeitschrittweite aus Stabilitätsgründen ist hier noch einschneidender als im eindimensionalen Fall.

Diskretisiert man u_t mit dem rückwärtsgenommenen Differenzenquotienten, so erhält man das implizite Verfahren

$$(\boldsymbol{E} + r\boldsymbol{A}_h)\,\boldsymbol{U}^{\ell+1} = \boldsymbol{U}^{\ell}, \tag{9.10}$$

bzw. durch Kombination von (9.9) und (9.10) das Verfahren von Crank und Nicholson

$$(2\boldsymbol{E} + r\boldsymbol{A}_h)\,\boldsymbol{U}^{\ell+1} = (2\boldsymbol{E} - r\boldsymbol{A}_h)\,\boldsymbol{U}^{\ell},\ \ell = 0, 1 \ldots. \tag{9.11}$$

Die Verfahren (9.9) und (9.10) sind wieder für alle Schrittweitenverhältnisse $r = k/h^2$ stabil in der Euklidischen Norm, sie erfordern aber, da hier die Matrix $\boldsymbol{A}_h$ zwar noch dünn besetzt aber nicht mehr tridiagonal ist, auf jeder Zeitschicht einen recht hohen Rechenaufwand beim Lösen des linearen Gleichungssystems (9.10) bzw. (9.11).

Um den Aufwand zu verkleinern, wurde von Peaceman und Rachford ein Verfahren vorgeschlagen, bei dem in jedem Zeitschritt nur einige tridiagonale Gleichungssysteme zu lösen sind. Die Idee ist, in jedem Schritt zwei verschiedene Differenzenapproximationen für die Ableitungen bzgl. x und y zu kombinieren.

Dazu wird zunächst ein halber Zeitschritt ausgeführt:

$$\begin{aligned}&\frac{2}{k}\left(U_{ij}^{\ell+1/2}-U_{ij}^{\ell}\right)\\&\quad=\frac{1}{h^2}\left(U_{i+1,j}^{\ell+1/2}-2U_{ij}^{\ell+1/2}+U_{i-1,j}^{\ell+1/2}\right)+\frac{1}{h^2}\left(U_{i,j-1}^{\ell}-2U_{ij}^{\ell}+U_{i,j+1}^{\ell}\right).\end{aligned}\tag{9.12}$$

u_{xx} wird also auf der Zeitschicht $t=(\ell+0.5)k$ und u_{yy} auf der Zeitschicht $t=\ell k$ mit dem zentralen Differenzenverfahren diskretisiert. Für jedes festgehaltene j ist dann (9.12) ein lineares, tridiagonales Gleichungssystem, das man effizient lösen kann.

Im zweiten Halbschritt wird u_{xx} auf der Zeitschicht $t=(\ell+0.5)k$ und u_{yy} auf der Schicht $t=(\ell+1)k$ diskretisiert:

$$\begin{aligned}&\frac{2}{k}\left(U_{ij}^{\ell+1}-U_{ij}^{\ell+1/2}\right)\\&\quad=\frac{1}{h^2}\left(U_{i+1,j}^{\ell+1/2}-2U_{ij}^{\ell+1/2}+U_{i-1,j}^{\ell+1/2}\right)+\frac{1}{h^2}\left(U_{i,j+1}^{\ell+1}-2U_{ij}^{\ell+1}+U_{i,j-1}^{\ell+1}\right)\end{aligned}\tag{9.13}$$

und dies ist für jedes feste i ein lineares Gleichungssystem mit tridiagonaler Koeffizientenmatrix.

Das Verfahren, das durch (9.12), (9.13) gegeben ist, heißt **Methode der alternierenden Richtungen**.

9.2 Linienmethode

Im letzten Abschnitt haben wir alle auftretenden partiellen Ableitungen durch Differenzenquotienten ersetzt und so die Anfangsrandwertaufgabe diskretisiert. Man kann die Differentialgleichung auch nur bzgl. der Ortsvariablen diskretisieren und so durch ein System von Anfangswertaufgaben ersetzen oder nur die Zeitvariable diskretisieren und so die parabolische Anfangsrandwertaufgabe durch eine Folge von elliptischen Randwertaufgaben ersetzen. In beiden Fällen nennt man das Vorgehen eine **Semidiskretisierung** der Aufgabe. Genauer spricht man im ersten Fall, den wir in diesem Abschnitt betrachten wollen, von einer **Linienmethode** oder manchmal auch vertikalen Linienmethode, im zweiten Fall von einem **Rothe Verfahren** oder einer horizontalen Linienmethode.

Wir betrachten als einführendes Beispiel die eindimensionale Wärmeleitungsgleichung

$$\left.\begin{aligned}u_t-u_{xx}&=f(x,t), && 0<x<1,\ t>0\\u(x,0)&=u_0(x), && 0<x<1\\u(0,t)&=u(1,t)=0, && t>0.\end{aligned}\right\}\tag{9.14}$$

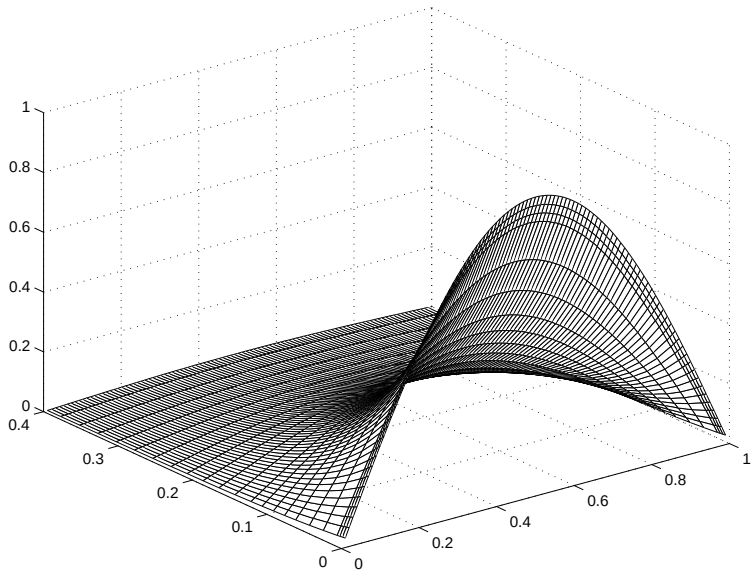

Abbildung 9.1: Lösungen für Beispiel 9.1.

Ersetzt man u_{xx} auf einem äquidistanten Gitter der Maschenweite $h = 1/n$ durch zentrale Differenzenquotienten und betrachtet man die Differentialgleichung nur auf den Linien $\{(x_j, t) : t \geq 0\}$, $j = 1, \ldots, n-1$, so erhält man das System gewöhnlicher Anfangswertaufgaben

$$\begin{aligned} \frac{d}{dt} v_j(t) &= \frac{1}{h^2}\Big(v_{j-1}(t) - 2v_j(t) + v_{j+1}(t)\Big) + f(x_j, t), \qquad (9.15) \\ v_j(0) &= u_0(x_j),\ j = 1, \ldots, n-1, \end{aligned}$$

wobei sich aus den Randbedingungen $v_0(t) \equiv 0$ und $v_n(t) \equiv 0$ ergibt. $v_j(t)$ ist dabei eine Näherung für $u(x_j, t)$.

Wendet man auf (9.15) das Eulersche Polygonzugverfahren an, so erhält man das explizite Differenzenverfahren, das implizite Euler Verfahren für (9.15) liefert das implizite Differenzenverfahren (9.4) und die Trapezregel das Crank–Nicholson Verfahren (9.5).

Die Differentialgleichung in (9.15) kann man schreiben als

$$\frac{d}{dt}\boldsymbol{v} = \boldsymbol{A}\boldsymbol{v} + \boldsymbol{f}(t),\ \boldsymbol{A} := -\frac{1}{h^2}\text{tridiag}\,(-1, 2, -1). \qquad (9.16)$$

Die Eigenwerte $\mu_j = -2n^2(1 - \cos(j\pi/n))$, $j = 1, \ldots, n-1$, von $\boldsymbol{A}$ zeigen, dass das System (9.15) für kleine Ortsschrittweiten h steif ist, so dass man es mit nichtsteifen Lösern nur mit sehr kleinen Zeitschritten behandeln kann. Die Eigenwerte liegen alle auf der negativen reellen Achse, so dass zur Lösung $A(0)$–stabile Verfahren geeignet sind.

Tabelle 9.5: Linienverfahren für Beispiel 9.1.

Löser	Schritte	rel. Fehler	flops
ode15s	23	$2.94e-3$	$1.24e+6$
ode23s	19	$9.40e-3$	$1.70e+6$
ode23	6224	$1.70e-3$	$3.35e+7$
ode45	19161	$1.38e-3$	$1.11e+8$
ode113	9572	$7.93e-3$	$4.81e+7$

Beispiel 9.1. Wir betrachten die Anfangswertaufgabe (9.14) mit $f(x,t) \equiv 0$ und $u_0(x) = \sin(\pi x)$. Die Lösung

$$u(x,t) = \exp(-\pi^2 t)\sin(\pi x)$$

ist in Abbildung 9.1 dargestellt.

Wendet man die 5 Löser der MATLAB 6.1 ODE-Suite mit der Ortschrittweite $h = 0.01$ an, so erhält man Ergebnisse der Tabelle 9.5 Man liest ab, dass die steifen Löser deutlich überlegen sind. □

Wir betrachten nun allgemeiner die parabolische Anfangsrandwertaufgabe

$$\left.\begin{aligned} \frac{\partial}{\partial t}u(\boldsymbol{x},t) + Lu(\boldsymbol{x},t) = f(\boldsymbol{x},t), &\quad \boldsymbol{x} \in \Omega,\ t > 0 \\ u(\boldsymbol{x},t) = 0, &\quad \boldsymbol{x} \in \partial\Omega,\ t > 0 \\ u(\boldsymbol{x},0) = u_0(\boldsymbol{x}) &\quad \boldsymbol{x} \in \Omega. \end{aligned}\right\} \tag{9.17}$$

Dabei sei $\Omega \subset \mathbb{R}^n$ ein beschränktes Gebiet mit glattem Rand und L ein gleichmäßig elliptischer Differentialoperator. Wir beschränken uns auf homogene Dirichletsche Randbedingungen. Die Übertragung auf andere Randbedingungen ist ohne Schwierigkeiten möglich.

Ist das Grundgebiet eindimensional, so diskretisiert man den Differentialoperator L durch ein Differenzenverfahren bzgl. x und erhält wie oben ein System gewöhnlicher Differentialgleichungen.

Ist $n \geq 2$ so verwendet man in der Regel eine finite Elementmethode oder ein finite Volumen Verfahren. Dazu muss (9.17) in schwacher Form geschrieben werden. Wir verzichten auf eine genaue Formulierung, da uns die benötigten Sobolevräume hier nicht zur Verfügung stehen. Gesucht ist eine Funktion u aus einem geeigneten Raum, so dass

$$\left.\begin{aligned} \frac{d}{dt}\langle u(\cdot,t), v\rangle + a(u(\cdot,t), v) = \langle f(\cdot,t), v\rangle &\quad \text{für alle } v \in H_0^1(\Omega) \\ u(\cdot,0) = u_0. \end{aligned}\right\} \tag{9.18}$$

wobei a die dem Operator L zugeordnete Bilinearform bezeichnet (vgl. (8.14)) und

$$\langle f, g\rangle := \int_\Omega f(\boldsymbol{x}) g(\boldsymbol{x})\, d\boldsymbol{x}.$$

Wir wählen einen finiten Element Raum $V_h \subset H_0^1(\Omega)$ und hierin eine Basis $\{\phi_1, \ldots, \phi_m\}$ und bestimmen eine Funktion

$$u_h(\boldsymbol{x}, t) = \sum_{j=1}^m U_j(t)\phi_j(\boldsymbol{x}) \tag{9.19}$$

mit

$$\left.\begin{aligned} \frac{d}{dt}\langle u_h(\cdot,t), v_h\rangle + a(u_h(\cdot,t), v_h) = \langle f(\cdot,t), v_h\rangle \quad \text{für alle } v_h \in V_h \\ u_h(\cdot,0) = u_{h,0}. \end{aligned}\right\} \tag{9.20}$$

mit einer Approximation $u_{h,0} \in V_h$ von u_0, d.h. wegen (9.19)

$$\left.\begin{aligned} \sum_{j=1}^m \langle \phi_i, \phi_j\rangle \frac{d}{dt} U_j + \sum_{j=1}^m a(\phi_i, \phi_j) U_j = \langle f(\cdot,t), \phi_i\rangle,\ i = 1, \ldots, m, \\ u_h(0) = u_{h,0}. \end{aligned}\right\} \tag{9.21}$$

Bezeichnet $\boldsymbol{M} := (\langle \phi_i, \phi_j\rangle)_{i,j=1,\ldots,m}$ die Massenmatrix, $\boldsymbol{K} := (a(\phi_i, \phi_j))_{i,j=1,\ldots,m}$ die Steifigkeitsmatrix des Systems und $\boldsymbol{f}(t) = (\langle f(\cdot,t), \phi_i\rangle)_{i=1,\ldots,m}$, so erhält man für die Vektorfunktion $\boldsymbol{U}(t) := (U_i(t))_{i=1,\ldots,m}$ das Differentialgleichungssystem

$$\boldsymbol{M}\frac{d}{dt}\boldsymbol{u} + \boldsymbol{K}\boldsymbol{u} = \boldsymbol{f}(t) \tag{9.22}$$

mit der Anfangsbedingung für $\boldsymbol{u}(0)$, die sich aus $u_h(0) = u_{h,0}$ ergibt.

Beispiel 9.2. Für die Anfangsrandwertaufgabe (9.14) liefert die Semidiskretisierung mit linearen Elementen auf einer äquidistanten Zerlegung der Schrittweite h die Anfangswertaufgabe (9.22) mit

$$\boldsymbol{M} = \frac{h}{6}\text{tridiag}\,(1, 4, 1), \quad \boldsymbol{K} = \frac{1}{h}\text{tridiag}\,(-1, 2, -1)$$

Dividiert man (9.22) durch h so erhält man für u_{xx} dieselbe Diskretisierung wie beim Differenzenverfahren. □

Bemerkung 9.3. Nachteil bei der Semidiskretisierung mit finiten Elementen ist das Auftreten der Massenmatrix $\boldsymbol{M}$, die nur dann diagonal ist, wenn die Ansatzfunktionen orthogonal sind bzgl. $\langle \cdot, \cdot\rangle$. In der Regel ist also Software zur Lösung von Anfangswertaufgaben des Typs

$$\frac{d}{dt}\boldsymbol{y} = f(t, \boldsymbol{y}),\ \boldsymbol{y}(0) = \boldsymbol{y}_0,$$

nicht unmittelbar anwendbar. Die Codes der MATLAB 6.1 ODE-Suite zur Lösung steifer Systeme sind so geschrieben, dass auch Anfangswertaufgaben der Gestalt

$$\boldsymbol{M}\frac{d}{dt}\boldsymbol{y} = f(t,\boldsymbol{y}), \quad \boldsymbol{y}(0) = \boldsymbol{y}_0,$$

behandelt werden können. □

Bemerkung 9.4. Wurde die Semidiskretisierung mit linearen Elementen vorgenommen, so kann man die in Bemerkung 9.3. angesprochene Schwierigkeit durch das sog. "**lumping**" umgehen. Die Massenmatrix wird ersetzt durch die Diagonalmatrix

$$\boldsymbol{D} = \operatorname{diag}(d_i), \quad d_i := \sum_j m_{ij}.$$

Dies lässt sich so interpretieren, dass die Elemente $\langle \phi_i, \phi_j \rangle$ der Massenmatrix nicht exakt ausgewertet wird, sondern mit Hilfe einer einfachen Quadraturformel (vgl. *Großmann, Roos* [26]. □

Wir haben hier die Idee der Linienmethode nur an linearen Anfangsrandwertaufgaben erläutert. Es ist klar, dass sie auch bei nichtlinearen Problemen oder bei Systemen von parabolischen Gleichungen zur Semidiskretisierung verwendet werden kann. Die entstehende gewöhnliche Anfangswertaufgabe ist immer mit steifen Lösern zu behandeln. Die angesprochene Verteilung der Eigenwerte der Matrix $\boldsymbol{A}$ in (9.15) ist typisch für parabolische Probleme. Ist der Operator L in (9.17) selbstadjungiert, so besitzt die Matrix $-\boldsymbol{M}^{-1}\boldsymbol{K}$ nur negative Eigenwerte, und bei Verfeinerung der Schrittweite h wächst der größte gegen eine negative Zahl und der kleinste fällt unter alle Grenzen. Die Anfangswertaufgabe (9.22) ist also bei kleinem h steif.

Linienmethoden zur Lösung von parabolischen Anfangsrandwertaufgaben sind in den meisten FEM Paketen enthalten.

In der PDE Toolbox von MATLAB werden ebene parabolische Probleme mit der Linienmethode mit linearen finiten Elementen gelöst.

Die Löser für steife Anfangswertaufgaben geben die Möglichkeit, die Schrittweite bzgl. der Variablen t zu steuern. Schwierig ist es aber, die Ortsschrittweite an die Lösung anzupassen. Für Probleme der Raumdimension 1 wurde von U. Nowak das Programmpaket PDEX1M erstellt, das eine lokale Anpassung der Schrittweiten in Raum und Zeit ermöglicht. Es ist fast public domain: Um es zu erhalten muss man den Autor kontakten (nowak@zib.de).

Neben den (vertikalen) Linienmethoden werden auch **Rothe Verfahren** oder horizontale Linienmethoden verwendet. Dabei wird für die Anfangsrandwertaufgabe (9.17) eine Semidiskretisierung von u_t mit dem impliziten Euler Verfahren

$$\left.\begin{aligned} \frac{u_{j+1}-u_j}{\Delta t} + Lu_{j+1}(\boldsymbol{x}) &= f(\boldsymbol{x}, t_{j+1}), & \boldsymbol{x} &\in \Omega \\ u_{j+1}(\boldsymbol{x}) &= 0, & \boldsymbol{x} &\in \partial\Omega \end{aligned}\right\} \tag{9.23}$$

vorgenommen, wobei $u_j : \Omega \to \mathbb{R}$ eine Näherung für $u(t_j, \boldsymbol{x})$ auf der Zeitschicht $t = t_j$ bezeichnet. Hierdurch wird die parabolische Aufgabe auf eine Folge von elliptischen Randwertaufgaben zurückgeführt. Auf diese wendet man nun eine der in Kapitel 7 oder Kapitel 8 besprochenen Methoden an. Es ist klar, dass man die für elliptische Probleme entwickelten räumlich adaptiven Verfahren auf jeder Zeitschicht verwenden kann.

Die Rothe Methode kann sofort auf Systeme (auch nichtlinearer) parabolischer Anfangsrandwertaufgaben übertragen werden. Auf dieser Basis wurde am Konrad Zuse Zentrum innerhalb des KASKADE Projekts (vgl. Kapitel 8) das Paket KARDOS (KAskade Reaction DiffusiOn System) zur Behandlung von (räumlich eindimensionalen) Reaktions–Diffusions Problemen entwickelt. Es kann geladen werden von

http://www.zib.de/de/numerik/software/kardos.html

Literaturverzeichnis

[1] *W.F. Ames*: Numerical Methods for Partial Differential Equations. 3^{rd} Ed., Academic Press, Boston 1992

[2] *E. Arge, A.M. Bruaset, H.-P. Lantangen*: Modern Software Tool for Scientific Computing. Birkhaeuser, Boston 1997

[3] *U.M. Ascher, R.M.M. Mattheij, R.D. Russel*: Numerical Solution of Boundary Value Problems for Ordinary Differential Equations. Prentice Hall, Englewood Cliffs 1988

[4] *U.M. Ascher, L. Petzold*: Computer Methods for Ordinary Differential Equations and Differential–Algebraic Equations. SIAM, Philadelphia 1998

[5] *K.-J. Bathe*: Finite Element Procedures. Prentice Hall, Englewood Cliffs 1996

[6] *P. Bogacki, L.F. Shampine*: A 3(2) pair of Runge–Kutta formulas. Appl. Math. Lett. 2, 1 – 9 (1989)

[7] *D. Braess*: Finite Elemente. Springer Verlag, Berlin 1992

[8] *K.E. Brenan, S.L. Campbell, L.R. Petzold*: Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations. North-Holland, New York 1989

[9] *J.C. Butcher*: Coefficients for the study of Runge-Kutta integration processes. J. Austral. Math. Soc. 3, 185-201 (1963)

[10] *J.C. Butcher*: The Numerical Analysis of Ordinary Differential Equations. Wiley, New York 1987

[11] *L. Collatz*: The Numerical Treatment of Differential Equations. 3^{rd} Ed., Springer Verlag, Berlin 1966

[12] *C.F. Curtiss, J.O. Hirschfelder*: Integration of stiff equations. Proc. Nat. Acad. Sci. 38, 235 – 243 (1952)

[13] *M. Daehlen, A. Tveito (eds.)*: Numerical Methods and Software Tools in Industrial Mathematics. Birkaeuser, Boston 1997

[14] *G. Dahlquist*: A special stability problem for linear multistep methods. BIT 3, 27 – 43 (1963)

[15] *J.E. Dennis, Jr., R.B. Schnabel*: Numerical Methods for Unconstrained Optimization and Nonlinear Equations. SIAM, Philadelphia 1996

[16] *P. Deuflhard, F. Bornemann*: Numerische Mathematik II. Integration gewöhnlicher Differentialgleichungen. De Gruyter, Berlin 1994

[17] *P. Deuflhard, A. Hohmann*: Numerische Mathematik I. Eine algorithmisch orientierte Einführung. 2. Aufl., De Gruyter, Berlin 1993

[18] *J.R. Dormand, P.J. Prince*: A family of embedded Runge–Kutta formulae. J. Comput. Appl. Math. 6, 19 – 26 (1980)

[19] *G. Engeln–Müllges, F. Reutter*: Numerik–Algorithmen. VDI Verlag, Düsseldorf 1996

[20] *K. Eriksson, D. Estep, P. Hansbo, C. Johnson*: Computational Differential Equations. Cambridge University Press, Cambridge 1996

[21] *I. Fried*: Numerical Solution of Differential Equations. Academic Press, New York 1979

[22] *C.W. Gear*: Numerical Initial Value Problems in Ordinary Differential Equations. Prentice Hall, Englewood Cliffs 1971

[23] *E. Griepentrog, R. März*: Differential-Algebraic Equations and Their Numerical Treatment. Teubner, Leipzig 1986

[24] *R.D. Grigorieff*: Numerik gewöhnlicher Differentialgleichungen I. Teubner Verlag, Stuttgart 1972

[25] R.D. Grigorieff: Numerik gewöhnlicher Differentialgleichungen II. Teubner Verlag, Stuttgart 1977

[26] C. Großmann, H.-G. Roos: Numerik partieller Differentialgleichungen. 2. Auflage. Teubner Verlag, Stuttgart 1994

[27] *W. Hackbusch*: Theorie und Numerik elliptischer Differentialgleichungen. Teubner Verlag, Stuttgart 1986

[28] *E. Hairer, C. Lubich, M. Roche*: The Numerical Solution of Differential-Algebraic Systems by Runge–Kutta Methods. Lecture Notes in Mathematics 1409, Springer, New York 1989

[29] *E. Hairer, S.P. Nørsett, G. Wanner*: Solving Ordinary Differential Equations I. Nonstiff Problems. 2. Auflage. Springer Verlag, Berlin 1993

[30] *E. Hairer, G. Wanner*: Solving Ordinary Differential Equations II. Stiff and Differential–Algebraic Problems. 2. Auflage. Springer Verlag, Berlin 1996

[31] *W.S. Hall*: The Boundary Element Method. Kluwer Academic Publishers, Dodrecht 1994

[32] *J.J.I.M. van Kan, A. Segal*: Numerik partieller Differentialgleichungen für Ingenieure. Teubner Verlag, Stuttgart 1995

[33] *P. Kaps, A. Ostermann*: Rosenbrock methods using few LU–decompositions. IMA J. Numer. Anal. 9, 15 – 27 (1984)

[34] *P. Kaps, P. Rentrop*: Generalized Runge–Kutta methods of order four with stepsize control for stiff ordinary differential equations. Numer. Math. 38, 279 – 298 (1979)

[35] *N. Köckler*: Numerical Methods and Scientific Computing Using Software Libraries for Problems Solving. Clarendon Press, Oxford 1994

[36] *J.D. Lambert*: Computational Methods in Ordinary Differential Equations. Wiley & Sons, Chichester 1973

[37] *A.R. Mitchell, D.F. Griffith*: The Finite Difference Method for Partial Differential Equations. Wiley & Sons, Chichester 1979

[38] Partial Differential Equation TOOLBOX. For Use with MATLAB. User's Guide. 2. Auflage. The Math Works, Natick 1995

[39] *M.H. Protter, H.F. Weinberger*: Maximum Principles in Differential Equations. Prentice Hall, Englewood Cliffs 1967

[40] *H.H. Robertson*: The Solution of a Set of Reaction Rate Equations. In J. Walsh (ed.), Numerical Analysis: An Introduction. Academic Press, London 1967, pp. 178 – 182

[41] *H.H. Rosenbrock*: Some general implicit processes for the numerical solution of differential equations. Computer J. 5, 329 – 330 (1962/63)

[42] *A.A. Samarskij*: Theorie der Differenzenverfahren. Geest & Portig, Leipzig 1984

[43] *H.R. Schwarz*: Methode der finiten Elemente. Teubner Verlag, Stuttgart 1980

[44] *H.R. Schwarz*: FORTRAN-Programme zur Methode der finiten Elemente. Teubner Verlag, Stuttgart 1981

[45] *H.R. Schwarz*: Numerische Mathematik. Teubner, Stuttgart 1993

[46] *L.F. Shampine*: Implementation of Rosenbrock methods. ACM Trans. Math. Soft. 8, 93 – 113 (1982)

[47] *L.F. Shampine*: Numerical Solution of Ordinary Differential Equations. Chapman & Hall, New York 1994

[48] *L. Shampine, M.K. Gordon*: Computer Lösung gewöhnlicher Differentialgleichungen. Vieweg Verlag, Braunschweig 1984

[49] *L.F. Shampine, M.W. Reichel*: The MATLAB ODE suite. SIAM J. Sci. Comput. 18, 1 – 22 (1997)

[50] *L.F. Shampine, M.W. Reichel, J.A. Kierzenka*: Solving Index-I DAEs in MATLAB and Simulink. SIAM Review 41, 538 – 552 (1999)

[51] SIMULINK Dynamic System Simulation Software. User's Guide. The Mathworks, Natick 1992

[52] *T. Steihaug, A. Wolfbrandt*: An attempt to avoid exact Jacobian and nonlinear equations in the numerical solution of stiff differential equations. Math. Comp. 33, 521 – 534 (1979)

[53] *H.J. Stetter*: Analysis of Discretization Methods for Ordinary Differential Equations. Springer Verlag, Berlin 1973

[54] *J. Stoer, R. Bulirsch*: Einführung in die Numerische Mathematik II. Springer Verlag, Berlin 1973

[55] *G. Strang, G.J. Fix*: An Analysis of the Finite Element Method. Prentice Hall, Englewood Cliffs 1973

[56] *K. Strehmel, R. Weiner*: Numerik gewöhnlicher Differentialgleichungen Teubner, Stuttgart 1995

[57] *W. Törnig, M. Gipser, B. Kaspar*: Numerische Lösung von partiellen Differentialgleichungen in der Technik. 3^{rd} Ed., Teubner Verlag, Stuttgart 1991

[58] *E.H. Twizell*: Computational Methods for Partial Differential Equations. John Wiley & Sons, New York 1984

[59] *M.V. van Veldhuizen*: D–stability and Kaps–Rentrop methods. Computing 32, 229 – 237 (1984)

Index

A(α)–stabil, 75
A(0)–stabil, 75
Abbruchfehler, 25
Ableitung
    schwache, 168
    verallgemeinerte, 168
absolut stetig, 128

BDF Formeln, 61
Bilinearform, 164
Bilinerform
    stetige, 164
Bogacki, 43
Boxschema, 126, 159
Broyden Verfahren, 112

Cea, Lemma von, 177
Crank–Nicholson Verfahren, 190

Dachfunktion, 135, 174
DAE, 97
Dahlquist, 75
diagonal implizites Runge–Kutta Verfahren, 80
Differentialgleichung
    elliptische, 10
    halblinear, 12
    hyperbolische, 10
    parabolische, 10
    quasilinear, 12
differentiell-algebraisches System, 97
Differenzenquotient
    zentraler, 119
Differenzenverfahren
    explizites, 188
    implizites, 190
Dirichletsche Randwertaufgabe, 13
DIRK–Verfahren, 80
Dormand, 43
drei achtel Regel, 40
dritte Randwertaufgabe, 14

einfach diagonal implizites Runge–Kutta Verfahren, 80
Einfach–Schießverfahren, 109
eingebette Runge–Kutta Formeln, 41
Einschrittverfahren, 31
elliptisch, 10, 164
    gleichmäßig, 13, 170
Energienorm, 131, 165
England, 42
erste Randwertaufgabe, 13
Euler Verfahren
    implizites, 60, 66
    linear implizites, 88
Eulersches Polygonzugverfahren, 24
explizit, 77
explizites Differenzenverfahren, 188
explizites Runge–Kutta Verfahren, 35

Fehlberg, 41
Fehler
    globaler, 25
    lokaler, 25, 31

finite Elemente, 173
finite Volumen Methode, 159
formal selbstadjungiert, 170
Formel von Kuntzmann, 40
Formelpaar von Bogacki und Shampine, 43
Formelpaar von Dormand und Prince, 43
Formelpaar von England, 42
Formelpaar von Fehlberg, 41
Formelpaar von Verner, 42
FSAL-Verfahren, 42
Fundamentallösung, 5
Fundamentalsystem, 5

Gauß Verfahren, 85
Gitterpunkt
  innerer, 152
  randnaher, 152
gleichmäßig elliptisch, 13, 170
globaler Fehler, 25
Gonwall, Lemma von, 3
Greensche Matrix, 8

halblinear, 12
Heun, Verfahren von, 34
hyperbolisch, 10

implizit, 77
implizite Mittelpunktregel, 77
implizites Differenzenverfahren, 190
implizites Euler Verfahren, 57, 60, 66
Index, 99
innerer Gitterpunkt, 152
invers monoton, 9
inversmonoton, 121

klassische Lösung, 10, 130, 168
klassisches Runge–Kutta Verfahren, 39
Kollokation, 83
konsistent, 31, 52, 120, 145
konsistenter Anfangswert, 98
konvergent, 54, 120, 145
Kuntzmann, Formel von, 40

L–stabil, 77
Lösung
  klassische, 130, 168
  schwache, 168
  starke, 168
  verallgemeinerte, 130
Lemma
  von Cea, 177
  von Gronwall, 3
linear implizites Euler Verfahren, 88
linear implizites Runge–Kutta Verfahren, 88
Linienmethode, 195
Lobatto Formel, 86
lokaler Fehler, 25, 31, 52
lumping, 199

Maximumprinzip, 20
  schwaches, 14
Mehrschrittverfahren, 51
Mehrstellenformel, 125
Methode der alternierenden Richtungen, 195
Methode der finiten Elemente, 173
Milne, 60
Milne Formel, 60
Milne–Simpson Formeln, 60
Mittelpunktregel, 59, 77, 126

natürliche Randbedingung, 140
NDF Verfahren, 93
Neumannsche Randwertaufgabe, 14

Nyström Formeln, 59

Ordnung, 52, 145
Ordnung eines Verfahrens, 31, 120

parabolisch, 10
Peano, Existenzsatz von, 2
Picard, Lindelöf, Satz von, 2
Poincarésche Ungleichung, 166
Poisson Gleichung, 11
Polygonzugverfahren, 24, 32
    verbessertes, 33
Potentialgleichung, 11
Prädiktor–Korrektor Verfahren, 57
Primärzerlegung, 159
Prince, 43

quasi–gleichmäßige Zerlegung, 178
Quasi–Newton Verfahren, 112
quasilinear, 12

rückwärtsgenommene Differenz, 60
Radau Formel, 86
Randbedingung
    natürliche, 140
    restliche, 140
    wesentliche, 140
randnaher Gitterpunkt, 152
Randwertaufgabe
    Dirichletsche, 13
    dritte, 14
    erste, 13
    Neumannsche, 14
    Robinsche, 14
    zweite, 14
restliche Randbedingung, 140
Ritz–Galerkin Lösung, 131
Robinsche Randwertaufgabe, 14
Rosenbrock Verfahren, 88
Rothe Verfahren, 195, 200
Runge–Kutta Verfahren, 35, 77
    diagonal implizites, 80
    eingebettetes, 41
    klassisches, 39
    linear implizites, 88

sachgemäß, 12
Satz
    von Peano, 2
    von Picard und Lindelöf, 2
Schießverfahren, 109
Schrittweitensteuerung, 28
schwache Ableitung, 168
schwache Lösung, 130, 168
schwaches Maximumprinzip, 14
SDIRK–Verfahren, 80
Sekundärzerlegung, 160
semi-explizite DAE, 101
Semidiskretisierung, 195
Shampine, 43
Shortley–Weller Approximation, 154
Simpson, 60
Sobolev Raum, 129
Sobolevraum, 167
stückweise stetig differenzierbar, 169
stabil, 54, 120, 145
Stabilitätsgebiet, 68
stark stabil, 54
starke Lösung, 130, 168
steif, 63
stetige Bilinearform, 164
Stufe eines Runge–Kutta Verfahrens, 35

Träger einer Funktion, 168
Trapezregel, 57, 72, 126

ultrahyperbolisch, 11

Variationsaufgabe, 129
verallgemeinerte Ableitung, 168
verallgemeinerte Lösung, 130
verbessertes Polygonzugverfahren, 33
Verfahren
    Prädiktor–Korrektor, 57
    von Crank–Nicholson, 190
    von Rothe, 200
Verfahren von Heun, 34
Verfahren von Runge–Kutta, 35
Verfahrensfunktion, 31
Verner, 42

W methoden, 90
Wärmeleitungsgleichung, 11
Wellengleichung, 11
wesentliche Randbedingung, 140

zentraler Differenzenquotient, 119
zulässige Zerlegung, 159
Zustandsraumgleichung, 101
Zustandsraumverfahren, 102
zweite Randwertaufgabe, 14