Numerische Verfahren für Differentialgleichungen

Steffen Börm

Stand 1. Juli 2014

Inhaltsverzeichnis

Vorwort

Differentialgleichungen sind ein wichtiges Werkzeug bei der Behandlung vieler mathematischer, physikalischer, biologischer, chemischer oder wirtschaftswissenschaftlicher Fragestellungen.

- In der Mathematik lassen sich beispielsweise mit ihrer Hilfe Geodäten beschreiben, die kürzesten Verbindungen zwischen zwei Punkten einer gekrümmten Oberfläche.
- In der Physik werden beispielsweise die Bewegung von Körpern, die Ausbreitung elektromagnetischer Wellen und die Ausbreitung von Wärme mit Hilfe von Differentialgleichungen beschrieben.
- In der Biologie werden Differentialgleichungen beispielsweise eingesetzt, um die zeitliche Entwicklung von Bakterienpopulationen zu modellieren.
- In der Chemie lässt sich mit Hilfe von Differentialgleichungen die Dynamik von Molekülen beschreiben.
- In den Wirtschaftswissenschaften kommen Differentialgleichungen bei der Simulation der Entwicklung von Kursen am Aktien- oder Geldmarkt zum Einsatz.

Den meisten dieser Gleichungen ist gemeinsam, dass sie sich in der Regel nicht „mit Papier und Bleistift“ lösen lassen, mit den Methoden der reinen Mathematik können häufig nur Rückschlüsse auf das Verhalten der Lösung ziehen, aber nur in seltenen Spezialfällen die Lösung angeben.

Deshalb kommen *numerische* Lösungsverfahren zum Einsatz. Mit Hilfe dieser Verfahren kann die exakte Lösung zwar auch nicht berechnet werden, aber sie lässt sich beliebig genau *approximieren*, und für die meisten praktischen Anwendungen genügt eine hinreichend gute Näherung der Lösung.

Diese Vorlesung gibt einen Überblick über einige der wichtigsten Verfahren für die Approximation der Lösungen von Differentialgleichungen. Besonders eingehend behandelt werden dabei *gewöhnliche Differentialgleichungen*, die vor allem bei der Simulation zeitabhängiger Phänomene eine wichtige Rolle spielen, und *elliptische partielle Differentialgleichungen*, mit denen sich Kraft- und Spannungsfelder beschreiben lassen, beispielsweise bei der Untersuchung elektromagnetischer Felder oder strukturmechanischer Fragestellungen.

Zwei weitere Typen partieller Differentialgleichungen, *hyperbolische* und *parabolische Gleichungen*, werden nur am Rande behandelt. Parabolische Gleichungen lassen sich allerdings als Kombination gewöhnlicher und elliptischer Gleichungen interpretieren und deshalb mit den hier beschriebenen Verfahren behandeln.

Die Behandlung hyperbolischer Differentialgleichungen erfordert dagegen häufig spezialisierte Lösungsverfahren, die den Rahmen dieser Vorlesung sprengen würden. Deshalb wird lediglich einer der einfachsten Lösungsansätze anhand eines einfachen Beispielproblems demonstriert und die Diskussion allgemeinerer Verfahren spezialisierten Vorlesungen überlassen, beispielsweise aus dem Bereich der Erhaltungsgleichungen oder der Strömungsmechanik.

Die Vorlesung setzt Kenntnisse der Analysis und der linearen Algebra voraus: Aus der Analysis sollten neben Grundbegriffen (Konvergenz von Folgen und Reihen, Stetigkeit) natürlich grundlegende Sätze der Differential- und Integralrechnung bekannt sein (Hauptsatz, Mittelwertsätze). Aus der linearen Algebra sollten Kenntnisse über das Rechnen in Vektorräumen und den Umgang mit Skalarprodukten mitgebracht werden.

Danksagung

Ich bedanke mich bei Janina Gnutzmann, Hendrik Felix Pohl und Sven Christophersen für Korrekturen und Verbesserungsvorschläge.

1 Einleitung

Bevor wir uns der Analyse und numerische Behandlung von gewöhnlichen Differentialgleichungen, insbesondere von Anfangswertproblemen, zuwenden, sollen zunächst einige mehr oder weniger einfache Probleme vorgestellt werden, die sich mit Hilfe derartiger Gleichungen beschreiben lassen.

1.1 Federpendel

Ein sehr einfaches Beispiel für ein Anfangswertproblem ist das abstrakte Federpendel: Es besteht aus einer Masse m, die mittels einer Feder mit einem festen Punkt verbunden ist und nach oben oder unten ausgelenkt werden kann. Die Auslenkung aus der Ruhelage zu einem bestimmten Zeitpunkt t bezeichnen wir mit $u(t)$.

Falls die Masse sich nicht im Nullpunkt befindet, ist die Feder angespannt und übt eine Kraft aus, die die Masse in den Nullpunkt zurückzieht. Im einfachten Fall ist diese Kraft $F(t)$ durch das *Hookesche Gesetz* [7]

$$F(t) = -\frac{\hat{c}}{\ell}u(t) \tag{1.1}$$

gegeben, wobei $\hat{c}$ eine vom Material der Feder abhängende Konstante und ℓ die Länge der Feder im Ruhezustand ist.

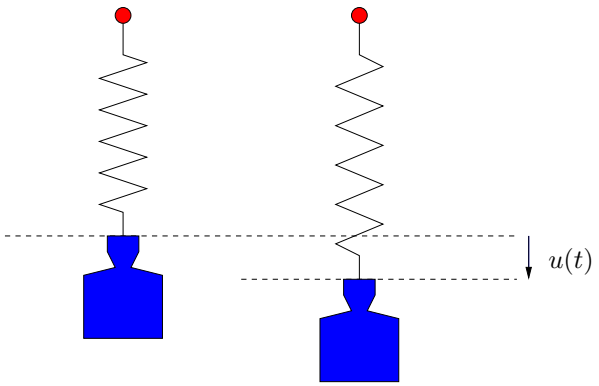

Abbildung 1.1: Modell eines Federpendels

Gemäß der Newtonschen Axiome [8] bewirkt die Kraft F eine Beschleunigung $a(t)$ der Masse, die proportional zum Kehrwert von m ist, es gilt also

$$F(t) = ma(t), \qquad a(t) = \frac{1}{m}F(t).$$

Die Beschleunigung ist die Ableitung der Geschwindigkeit $v(t)$ der Masse, und die Geschwindigkeit ist die Ableitung der Auslenkung $u(t)$, so dass wir die Gleichungen

$$u'(t) = v(t), \qquad v'(t) = a(t) = \frac{1}{m}F(t) = -\frac{\hat{c}}{m\ell}u(t)$$

erhalten. Indem wir $\lambda := \hat{c}/(m\ell)$ einführen und $u(t)$ und $v(t)$ zu einem Vektor

$$\mathbf{y}(t) = \begin{pmatrix} u(t) \\ v(t) \end{pmatrix}$$

zusammenfassen, erhalten wir die kompakte Schreibweise

$$\mathbf{y}'(t) = \begin{pmatrix} 0 & 1 \\ -\lambda & 0 \end{pmatrix} \mathbf{y}(t),$$

mit der die Bewegungen des abstrakten Pendels vollständig beschrieben werden können.

Wenn die Auslenkung $u(0)$ und die Geschwindigkeit $v(0)$ zum Startzeitpunkt bekannt sind, können wir Auslenkung und Geschwindigkeit zu jedem späteren Zeitpunkt $t \geq 0$ als Lösung des Systems

$$\mathbf{y}(0) = \begin{pmatrix} u(0) \\ v(0) \end{pmatrix}, \qquad \mathbf{y}'(t) = \begin{pmatrix} 0 & 1 \\ -\lambda & 0 \end{pmatrix} \mathbf{y}(t) \qquad \text{für alle } t \in \mathbb{R}_{\geq 0} \tag{1.2}$$

bestimmen. Ein derartiges Gleichungssystem, bei dem die Lösung zu einem Anfangszeitpunkt und die Ableitung der Lösung zu jedem Zeitpunkt t bekannt sind, nennt man *Anfangswertproblem*.

In unserem Fall haben wir es mit einem besonders einfachen System zu tun, dass sich analytisch lösen lässt: Wir führen die Matrix

$$\mathbf{A} := \begin{pmatrix} 0 & 1 \\ -\lambda & 0 \end{pmatrix}$$

ein und erhalten

$$\mathbf{y}'(t) = \mathbf{A}\mathbf{y} \qquad \text{für alle } t \in \mathbb{R}_{\geq 0},$$

ein *lineares* Anfangswertproblem. Wenn $\mathbf{A}$ eine Zahl wäre, könnten wir $t \mapsto \alpha \exp(At)$ als Lösungsansatz verwenden, wobei $\exp(x) = e^x$ die Exponentialfunktion bezeichnet. Dann müssen wir nur α so bestimmen, dass die Anfangsbedingung erfüllt ist.

Da $\mathbf{A}$ in unserem Fall eine Matrix ist, bietet es sich an, nach einer Verallgemeinerung der Exponentialfunktion zu suchen. Die naheliegende Definition

$$\exp(\mathbf{C}) := \sum_{j=0}^{\infty} \frac{\mathbf{C}^j}{j!} \tag{1.3}$$

für eine Matrix $\mathbf{C} \in \mathbb{R}^{n\times n}$ beruht auf der Exponentialreihe. Für eine beliebige induzierte Matrixnorm gilt

$$\|\exp(\mathbf{C})\| = \left\|\sum_{j=0}^{\infty} \frac{\mathbf{C}^j}{j!}\right\| \leq \sum_{j=0}^{\infty} \left\|\frac{\mathbf{C}^j}{j!}\right\| \leq \sum_{j=0}^{\infty} \frac{\|\mathbf{C}\|^j}{j!} = \exp(\|\mathbf{C}\|),$$

also ist die Reihe absolut konvergent, also insbesondere konvergent, und die Matrix-Exponentialfunktion damit durch (1.3) wohldefiniert.

Im Kontext der Anfangswertprobleme sind wir an der Ableitung der Funktion $t \mapsto \exp(t\mathbf{A})$ interessiert, die durch

$$\begin{aligned}\frac{\partial}{\partial t}\exp(t\mathbf{A}) &= \frac{\partial}{\partial t}\sum_{j=0}^{\infty}\frac{t^j\mathbf{A}^j}{j!} = \sum_{j=0}^{\infty}\frac{jt^{j-1}\mathbf{A}^j}{j!} = \sum_{j=1}^{\infty}\frac{t^{j-1}\mathbf{A}^j}{(j-1)!}\\ &= \mathbf{A}\sum_{j=1}^{\infty}\frac{t^{j-1}\mathbf{A}^{j-1}}{(j-1)!} = \mathbf{A}\exp(t\mathbf{A}) \qquad \text{für alle } t\in\mathbb{R}\end{aligned}$$

gegeben ist. Da aus (1.3) auch $\exp(\mathbf{0}) = \mathbf{I}$ folgt, erfüllt die durch

$$\mathbf{y}(t) := \exp(t\mathbf{A})\mathbf{y}(0) \qquad \text{für alle } t\in\mathbb{R}_{\geq 0}$$

definierte Funktion gerade

$$\begin{aligned}&\mathbf{y}(t_0) = \exp(0\mathbf{A})\mathbf{y}(0) = \mathbf{y}(0),\\ &\mathbf{y}'(t) = \mathbf{A}\exp(t\mathbf{A})\mathbf{y}(0) = \mathbf{A}\mathbf{y}(t) \qquad \text{für alle } t\in\mathbb{R}_{\geq 0},\end{aligned}$$

ist also eine Lösung des linearen Anfangswertproblems.

Statt durch direkte Auswertung der Exponentialsumme können wir die Exponentialfunktion auch berechnen, indem wir $\mathbf{A}$ mit einer Ähnlichkeitstransformation diagonalisieren: Mit der Matrix

$$\mathbf{T} := \begin{pmatrix} 1 & 1\\ i\sqrt{\lambda} & -i\sqrt{\lambda}\end{pmatrix}, \qquad \mathbf{T}^{-1} = \frac{1}{2}\begin{pmatrix} 1 & -i/\sqrt{\lambda}\\ 1 & i/\sqrt{\lambda}\end{pmatrix},$$

erhalten wir

$$\mathbf{T}^{-1}\mathbf{A}\mathbf{T} = \begin{pmatrix} i\sqrt{\lambda} & \\ & -i\sqrt{\lambda}\end{pmatrix} =: \mathbf{D}$$

und können die Exponentialfunktion durch

$$\begin{aligned}\exp(t\mathbf{A}) &= \sum_{i=0}^{\infty}\frac{t^i\mathbf{A}^i}{i!} = \sum_{i=0}^{\infty}\frac{t^i(\mathbf{T}\mathbf{D}\mathbf{T}^{-1})^i}{i!} = \mathbf{T}\left(\sum_{i=0}^{\infty}\frac{t^i\mathbf{D}^i}{i!}\right)\mathbf{T}^{-1}\\ &= \mathbf{T}\begin{pmatrix}\sum_{i=0}^{\infty}\frac{(ti\sqrt{\lambda})^i}{i!} & \\ & \sum_{i=0}^{\infty}\frac{(-ti\sqrt{\lambda})^i}{i!}\end{pmatrix}\mathbf{T}^{-1}\end{aligned}$$

$$= \mathbf{T} \begin{pmatrix} \exp(ti\sqrt{\lambda}) & \\ & \exp(-ti\sqrt{\lambda}) \end{pmatrix} \mathbf{T}^{-1}$$

auf die Berechnung der Exponentialfunktion für skalare Werte zurückführen.

Die Exponentialfunktion eines rein imaginären Werts steht in enger Beziehung zu Sinus- und Cosinus-Funktionen, deshalb überrascht es nicht, dass wir auch direkt den Ansatz

$$\begin{aligned} u(t) &= \alpha \sin(\omega t) + \beta \cos(\omega t), \\ v(t) &= \alpha\omega \cos(\omega t) - \beta\omega \sin(\omega t) \qquad \text{für alle } t \in \mathbb{R}_{\geq t_0} \end{aligned}$$

verwenden können. Der Parameter $\omega = \sqrt{\lambda} = \sqrt{\hat{c}/(m\ell)}$ beschreibt die Frequenz der Schwingung in Abhängigkeit von Masse, Materialeigenschaften und Federlänge, während die Parameter α und β verwendet werden können, um sicherzustellen, dass die Anfangsbedingungen erfüllt sind.

1.2 Mehrkörperprobleme

Das abstrakte Federpendel ist ein relativ einfaches Beispiel, weil die Ableitung $\mathbf{y}'$ und die Funktion $\mathbf{y}$ lediglich durch eine Matrix, also eine lineare Abbildung, gekoppelt sind und sich deshalb die Lösung analytisch angeben lässt.

Die in der Praxis auftretenden Probleme sind in der Regel nicht so einfach zu behandeln. Ein Beispiel ist das *Mehrkörperproblem*, bei dem n Massen $m_1, \ldots, m_n$ zu einem Zeitpunkt t an n verschiedenen Positionen $x_1(t), \ldots, x_n(t)$ im zwei- oder höherdimensionalen Raum liegen und mittels der Gravitation aufeinander einwirken.

In diesem Fall übt die Masse m_i auf die Masse m_j eine Kraft von

$$F_{ij}(t) = \varrho m_i m_j \frac{x_j(t) - x_i(t)}{\|x_j(t) - x_i(t)\|^3}$$

aus, wobei ϱ die Gravitationskonstante ist. Insgesamt wirkt also eine Kraft von

$$F_i(t) = \sum_{\substack{j=1 \\ j \neq i}} F_{ij}(t) = \varrho m_i \sum_{\substack{j=1 \\ j \neq i}}^{n} m_j \frac{x_j(t) - x_i(t)}{\|x_j(t) - x_i(t)\|^3}$$

auf die Masse m_i, und entsprechend der Newton-Axiome entsteht dadurch eine Beschleunigung von

$$a_i(t) = \varrho \sum_{\substack{j=1 \\ j \neq i}}^{n} m_j \frac{x_j(t) - x_i(t)}{\|x_j(t) - x_i(t)\|^3}. \tag{1.4}$$

Wie im Falle des Federpendels benutzen wir die Newtonschen Axiome um festzustellen, dass a_i die Ableitung der Geschwindigkeit v_i und v_i die Ableitung des Ortes x_i ist, also

$$x_i'(t) = v_i(t), \qquad v_i'(t) = a_i(t) = \varrho \sum_{\substack{j=1 \\ j \neq i}}^{n} m_j \frac{x_j(t) - x_i(t)}{\|x_j(t) - x_i(t)\|^3} \qquad \text{für alle } t \in \mathbb{R}_{\geq 0}$$

gilt. Wir fassen die Orte x_i, die Geschwindigkeiten v_i und die Beschleuigungen a_i zu Vektoren

$$\mathbf{x}(t) := \begin{pmatrix} x_1(t) \\ \vdots \\ x_n(t) \end{pmatrix}, \qquad \mathbf{v}(t) := \begin{pmatrix} v_1(t) \\ \vdots \\ v_n(t) \end{pmatrix}, \qquad \mathbf{a}(t) := \begin{pmatrix} a_1(t) \\ \vdots \\ a_n(t) \end{pmatrix}$$

zusammen und schreiben die Differentialgleichung in der Form

$$\begin{pmatrix} \mathbf{x}'(t) \\ \mathbf{v}'(t) \end{pmatrix} = \begin{pmatrix} \mathbf{v}(t) \\ \mathbf{a}(t) \end{pmatrix} \qquad \text{für alle } t \in \mathbb{R}_{\geq 0}.$$

Gemäß (1.4) können wir $\mathbf{a}(t)$ als Funktion $\mathbf{A}$ von $\mathbf{x}$ schreiben, erhalten also

$$\mathbf{a}(t) = \mathbf{A}(\mathbf{x}(t)) \qquad \text{für alle } t \in \mathbb{R}_{\geq 0}.$$

Zur Vereinheitlichung der Darstellung fassen wir $\mathbf{x}(t)$ und $\mathbf{v}(t)$ zu einem Vektor

$$\mathbf{y}(t) := \begin{pmatrix} \mathbf{x}(t) \\ \mathbf{v}(t) \end{pmatrix} \qquad \text{für alle } t \in \mathbb{R}_{\geq 0}$$

zusammen und führen die Funktion

$$\mathbf{f}(\mathbf{y}) := \begin{pmatrix} \mathbf{y}_2 \\ \mathbf{A}(\mathbf{y}_1) \end{pmatrix}$$

ein, um die kompakte Darstellung

$$\mathbf{y}'(t) = \mathbf{f}(\mathbf{y}(t)) \qquad \text{für alle } t \in \mathbb{R}_{\geq 0}$$

zu erhalten.

Im Falle des Federpendels war $\mathbf{f}$ lediglich eine lineare Abbildung, im Falle des Mehrkörperproblems ist $\mathbf{f}$ nicht linear, und die einfachen analytischen Lösungsansätze für lineare Probleme lassen sich nicht mehr verwenden.

Für $n \leq 3$ ist es noch möglich, die Lösung $\mathbf{y}$ wenigstens formal (etwa durch spezielle Reihenentwicklungen) darzustellen, für $n > 3$ dagegen sind numerische Approximationsverfahren das Mittel der Wahl.

1.3 Schwingende Saite

Wir können Mehrkörpersysteme auch verwenden, um Näherungen im Wesentlichen kontinuierlicher Phänomene zu gewinnen. Als Beispiel untersuchen wir eine Saite, die zwischen zwei Punkten eingespannt ist und die zwischen diesen Punkten frei schwingen kann.

Als Modell verwenden wir Punktmassen, die durch Federn aneinander gekoppelt sind. Sei dazu $n \in \mathbb{N}$. Wir verwenden $n+2$ Punktmassen, die wir mit $\{0, \ldots, n+1\}$ durchnumerieren. Dabei sollen die Punktmassen mit den Nummern 0 und $n+1$ den Endpunkten

der Saite entsprechen. Die Punktmasse mit der Nummer $i \in \{1, \dots, n\}$ soll mit den Punktmassen mit den Nummern $i-1$ und $i+1$ durch Federn verbunden sein.

Wenn wir die Länge der Saite im Ruhezustand mit $\ell \in \mathbb{R}_{>0}$ bezeichnen und wir sie in gleich lange Stücke einteilen, hat jede Feder die Ruhelänge

$$h := \frac{\ell}{n+1}.$$

Um die von den Federn ausgeübten Kräfte in Beschleunigungen umrechnen zu können, müssen wir den Punktmassen Massen zuordnen. Dazu nehmen wir an, dass die Masse einer Saite der Ruhelänge 1 durch $\hat{m}$ gegeben ist. Für $i \in \{1, \dots, n\}$ ersetzt die i-te Punktmasse ein Stück der Saite der Länge h, für die Endpunkte $i \in \{0, n+1\}$ dagegen nur ein Reststück der Länge $h/2$. Wenn wir davon ausgehen, dass die Masse gleichmäßig über die Saite verteilt ist, ergibt sich für die i-te Punktmasse die Masse

$$m_i := \begin{cases} \hat{m}h/2 & \text{falls } i \in \{0, n+1\}, \\ \hat{m}h & \text{ansonsten.} \end{cases}$$

Da wir uns für Schwingungen der Saite interessieren, bietet es sich an, die Positionen der einzelnen Punktmassen zu untersuchen. Wir bezeichnen mit $x_i(t)$ die Position der i-ten Masse zu einem Zeitpunkt $t \in \mathbb{R}$.

Die von der Feder zwischen der i-ten und der $(i+1)$-ten Punktmasse auf erstere ausgeübte Kraft ist nach dem Hooke'schen Gesetz (1.1) durch

$$\frac{\hat{c}}{h}(x_{i+1}(t) - x_i(t))$$

gegeben. Die von der Feder zwischen der $(i-1)$-ten und der i-ten Punktmasse ausgeübte Kraft addiert sich hinzu, so dass wir insgesamt die Kraft

$$\begin{aligned} F_i(t) &= \frac{\hat{c}}{h}(x_{i+1}(t) - x_i(t))\frac{\hat{c}}{h}(x_{i-1}(t) - x_i(t)) \\ &= \frac{\hat{c}}{h}(x_{i+1}(t) - 2x_i(t) + x_{i-1}(t)), \end{aligned}$$

erhalten. Die Beschleunigung ergibt sich wie zuvor, indem wir durch die Masse m_i dividieren:

$$\begin{aligned} a_i(t) &= \frac{1}{m_i}\frac{\hat{c}}{h}(x_{i+1}(t) - 2x_i(t) + x_{i-1}(t)) \\ &= \frac{\hat{c}}{\hat{m}}\frac{x_{i+1}(t) - 2x_i(t) + x_{i-1}(t)}{h^2}. \end{aligned}$$

Anders als im Fall der Gravitationskraft sind die einzelnen Komponenten der Vektoren $x_i(t)$, $F_i(t)$ und $a_i(t)$ voneinander völlig unabhängig, so dass wir sie getrennt voneinander untersuchen können.

Im Folgenden beschränken wir uns auf eine der Komponenten, verwenden aber weiterhin die bisher eingesetzte Notation. Mit den Newton-Axiomen erhalten wir das System

$$x_i'(t) = v_i(t), \qquad v_i'(t) = \frac{\hat{c}}{\hat{m}}\frac{x_{i+1}(t) - 2x_i(t) + x_{i-1}(t)}{h^2} \qquad \text{für alle } t \in \mathbb{R}. \tag{1.5}$$

Wie zuvor können wir das System kompakt schreiben, indem wir die Größen zu Vektoren

$$\mathbf{x}(t) := \begin{pmatrix} x_1(t) \\ \vdots \\ x_n(t) \end{pmatrix}, \qquad \mathbf{v}(t) := \begin{pmatrix} v_1(t) \\ \vdots \\ v_n(t) \end{pmatrix}$$

zusammenfassen. Die Berechnung der Beschleunigung aus $\mathbf{x}(t)$ lässt sich durch die Matrix

$$\mathbf{L} := \frac{\hat{c}}{\hat{m}h^2} \begin{pmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix}$$

kompakt in die Form

$$\mathbf{v}'(t) = -\mathbf{L}\mathbf{x}(t)$$

bringen, so dass sich insgesamt

$$\begin{pmatrix} \mathbf{x}'(t) \\ \mathbf{v}'(t) \end{pmatrix} = \begin{pmatrix} & \mathbf{I} \\ -\mathbf{L} & \end{pmatrix} \begin{pmatrix} \mathbf{x}(t) \\ \mathbf{v}(t) \end{pmatrix}$$

ergibt. Man erkennt eine gewisse Ähnlichkeit zu der Gleichung (1.2), durch die wir das Federpendel beschrieben haben: Die Matrix $\mathbf{L}$ tritt an die Stelle des Parameters λ.

Die durch Federn verbundenen Punktmassen sind lediglich als Näherung einer kontinuierlichen Saite gedacht. Um zu einer entsprechenden Gleichung zu gelangen, untersuchen wir den Quotienten in der Gleichung (1.5) etwas näher.

Lemma 1.1 (Differenzenquotient für die zweite Ableitung) *Sei $h \in \mathbb{R}_{>0}$, und sei $g : [-h, h] \to \mathbb{R}$ viermal stetig differenzierbar. Dann existiert ein $\eta \in [-h, h]$ mit*

$$\frac{g(h) - 2g(h) + g(-h)}{h^2} = g''(0) + \frac{h^2}{12} g^{(4)}(\eta).$$

Beweis. Mit dem Satz von Taylor erhalten wir

$$\begin{aligned} g(h) &= g(0) + hg'(0) + \frac{h^2}{2}g''(0) + \frac{h^3}{6}g^{(3)}(0) + \frac{h^4}{24}g^{(4)}(\eta_+), \\ g(-h) &= g(0) - hg'(0) + \frac{h^2}{2}g''(0) - \frac{h^3}{6}g^{(3)}(0) + \frac{h^4}{24}g^{(4)}(\eta_-) \end{aligned}$$

für geeignete Zwischenpunkte $\eta_+ \in [0, h]$ und $\eta_- \in [-h, 0]$. Indem wir beide Gleichungen addieren ergibt sich

$$g(h) + g(-h) = 2g(0) + h^2 g''(0) + \frac{h^4}{12} \frac{g^{(4)}(\eta_+) + g^{(4)}(\eta_-)}{2}.$$

Wir bringen $2g(0)$ auf die linke Seite und dividieren durch h^2, um zu

$$\frac{g(h) - 2g(0) + g(-h)}{h^2} = g''(0) + \frac{h^2}{12}\frac{g^{(4)}(\eta_+) + g^{(4)}(\eta_-)}{2}$$

zu gelangen. Da $g^{(4)}$ stetig ist, finden wir mit dem Zwischenwertsatz ein $\eta \in [\eta_-, \eta_+] \subseteq [-h, h]$, das

$$\frac{g^{(4)}(\eta_+) + g^{(4)}(\eta_-)}{2} = g^{(4)}(\eta)$$

erfüllt. Damit ist unsere Gleichung bewiesen. ∎

Um Lemma 1.1 auf die Gleichung (1.5) anwenden zu können, legen wir für jeden Zeitpunkt $t \in \mathbb{R}$ eine viermal stetig differenzierbare Kurve $s \mapsto x(t, s)$ derart durch die Punkte $x_i(t)$, das

$$x(t, s) = x_i(t) \qquad \text{für alle } i \in \{0, \dots, n+1\},\ t \in \mathbb{R},\ s = ih$$

gilt. Dann folgt mit dem Lemma

$$\begin{aligned}\frac{x_{i+1}(t) - 2x_i(t) + x_{i-1}(t)}{h^2} &= \frac{x(t, ih+h) - 2x(t, ih) + x(t, ih-h)}{h^2} \\ &= \frac{\partial^2 x}{\partial s^2}(t, s) + \frac{h^2}{12}\frac{\partial^4 x}{\partial s^4}(t, \eta_t)\end{aligned}$$

mit einem Parameter η_t. Wenn wir h gegen null streben lassen und annehmen, dass die vierten Ableitungen beschränkt bleiben, wird so aus der Gleichung (1.5) die partielle Differentialgleichung

$$\frac{\partial x}{\partial t}(t, s) = v(t, s), \qquad \frac{\partial v}{\partial t}(t, s) = \frac{\hat{c}}{\hat{m}}\frac{\partial^2 x}{\partial s^2}(t, s) \qquad \text{für alle } t \in \mathbb{R},\ s \in (0, \ell),$$

aus der sich durch Elimination der Geschwindigkeit $v(t, s)$ die *eindimensionale Wellengleichung*

$$\frac{\partial^2 x}{\partial t^2}(t, s) = \frac{\hat{c}}{\hat{m}}\frac{\partial^2 x}{\partial s^2}(t, s) \qquad \text{für alle } t \in \mathbb{R},\ s \in (0, \ell)$$

ergibt. Sie beschreibt die Schwingung einer kontinuierlichen Saite.

1.4 Wärmeleitung

Ein weiteres Beispiel für eine partielle Differentialgleichung ist die *eindimensionale Wärmeleitungsgleichung*

$$\frac{\partial u}{\partial t}(x, t) = \kappa\frac{\partial^2 u}{\partial x^2}(x, t) \qquad \text{für alle } t \in \mathbb{R}_{>0}, x \in [0, 1]. \tag{1.6}$$

Sie beschreibt die Erwärmung oder Abkühlung eines Drahtes der Länge 1: $x \in [0,1]$ gibt die Position auf dem Draht an, t den Zeitpunkt, und $u(x,t)$ ist die Temperatur im Punkt x zum Zeitpunkt t. Wir nehmen zur Vereinfachung an, dass die Randbedingungen

$$u(0,t) = u(1,t) = 0 \qquad \text{für alle } t \in \mathbb{R}_{>0}$$

gelten, dass also die Temperatur an den beiden Endpunkten des Drahts fixiert ist.

Um diese Gleichung numerisch behandeln zu können, kehren wir den Weg um, den wir bei der Wellengleichung gegangen sind: Mit Lemma 1.1 gilt

$$\frac{u(x-h,t) - 2u(x,t) + u(x+h,t)}{h^2} = \frac{\partial^2 u}{\partial x^2}(x,t) + \frac{h^2}{12}\frac{\partial^4 u}{\partial x^4}(\eta,t) \tag{1.7}$$

für ein geeignetes $\eta \in [0,1]$. Falls die vierte Ableitung gleichmäßig beschränkt ist, können wir also die zweite Ableitung durch den Differenzenquotienten auf der linken Seite approximieren, und die Approximation wird wie h^2 gegen die korrekte Ableitung konvergieren.

Für ein $n \in \mathbb{N}$ wählen wir eine Schrittweite

$$h := \frac{1}{n+1}$$

und ersetzen das kontinuierliche Intervall $[0,1]$ durch $(n+2) \in \mathbb{N}$ diskrete Punkte $0 = x_0 < x_1 < \ldots < x_n < x_{n+1} = 1$, die durch

$$x_i := ih \qquad \text{für alle } i \in \{0,\ldots,n+1\}$$

gegeben sind, und wir beschreiben entsprechend die Funktion $u(x,t)$ durch den Vektor $\mathbf{y}(t) = (y_i(t))_{i=1}^n$ mit

$$y_i(t) = u(x_i,t) \qquad \text{für alle } t \in \mathbb{R}_{>0}, i \in \{1,\ldots,n\}.$$

Indem wir die zweite Ableitung durch (1.7) approximieren, stellen wir fest, dass die Wärmeleitungsgleichung (1.6) durch die Gleichungen

$$y_i'(t) \approx \begin{cases} \frac{\kappa}{h^2}(y_{i-1}(t) - 2y_i(t) + y_{i+1}(t)) & \text{falls } 1 < i < n, \\ \frac{\kappa}{h^2}(-2y_i(t) + y_{i+1}(t)) & \text{falls } i = 1, \\ \frac{\kappa}{h^2}(y_{i-1}(t) - 2y_i(t)) & \text{falls } i = n, \end{cases} \qquad \begin{array}{c}\text{für alle } i \in \{1,\ldots,n\} \\ \text{und } t \in \mathbb{R}_{\geq 0}\end{array}$$

approximiert wird. Wenn wir also das System

$$\mathbf{y}'(t) = \mathbf{A}\mathbf{y}(t) \qquad \text{für alle } t \in \mathbb{R}_{\geq 0}$$

mit der Matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, gegeben durch

$$\mathbf{A} := -\frac{\kappa}{h^2}\begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix},$$

lösen, dürfen wir darauf hoffen, dass die Komponenten $y_i(t)$ des Vektors $\mathbf{y}(t)$ gute Näherungen für die Werte $u(x_i, t)$ der tatsächlichen Lösung sind. Es handelt sich also wieder um ein lineares Anfangswertproblem, allerdings steht diesmal die Approximation eines Differentialoperators hinter der Matrix $\mathbf{A}$. Eine genauere Analyse zeigt, dass sich bestimmte Eigenschaften dieses Operators auf die Matrix $\mathbf{A}$ übertragen und dazu führen, dass sich einfache Verfahren zur Behandlung von Anfangswertproblemen für dieses Problem nicht gut eignen. Ein wichtiges Ziel wird deshalb darin bestehen, Techniken zu entwickeln, mit denen sich auch dieses Problem effizient behandeln lässt.

2 Einige theoretische Aussagen über gewöhnliche Differentialgleichungen

Bevor wir numerische Lösungsverfahren für gewöhnliche Differentialgleichungen untersuchen können, müssen wir zunächst klären, unter welchen Bedingungen diese Gleichungen überhaupt eine Lösung besitzen. In Hinblick auf die numerische Behandlung ist ebenfalls wichtig, wie empfindlich die Lösung auf Störungen der Parameter, insbesondere des Startwerts, reagiert.

2.1 Allgemeine Problemstellung

Wir konzentrieren uns auf die Analyse des Anfangswertproblems

$$y(a) = y_0, \qquad y'(t) = f(t, y(t)) \qquad \text{für alle } t \in [a, b] \tag{2.1}$$

auf einem kompakten Intervall $[a, b]$ mit einem Startwert y_0 in einem Banachraum V und einer Funktion $f : [a, b] \times V \to V$. Gesucht ist eine mindestens einmal stetig differenzierbare Funktion $y : [a, b] \to V$.

Das allgemeinere Problem

$$\begin{aligned} y(a) = y_0, \quad y'(a) = y_1, \quad \ldots, \quad y^{(m-1)}(a) = y_{m-1}, \\ y^{(m)}(t) = f(t, y(t), y'(t), \ldots, y^{(m-1)}(t)) \qquad \text{für alle } t \in [a, b] \end{aligned}$$

lässt sich auf die Form (2.1) zurückführen, indem wir den Hilfsvektor

$$\mathbf{w}(t) := \begin{pmatrix} y(t) \\ y'(t) \\ \vdots \\ y^{(m-1)}(t) \end{pmatrix} \qquad \text{für alle } t \in [a, b]$$

einführen und das erweiterte System

$$\mathbf{w}(a) = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_{m-1} \end{pmatrix}, \qquad \mathbf{w}'(t) = \begin{pmatrix} w_2(t) \\ w_3(t) \\ \vdots \\ w_m(t) \\ f(t, w_1(t), w_2(t), \ldots, w_m(t)) \end{pmatrix}$$

für alle $t \in [a, b]$ lösen.

2.2 Existenz und Eindeutigkeit

Bei der Untersuchung der Eigenschaften eines Anfangswertproblems hat es sich als sehr nützlich erwiesen, anstelle der differentiellen Formulierung (2.1) eine Integralformulierung zu verwenden, die ohne die Forderung nach Differenzierbarkeit auskommt.

Lemma 2.1 (Integralformulierung) *Sei eine stetige Funktion $f \in C([a,b] \times V, V)$ gegeben. Falls eine Funktion $y \in C^1([a,b], V)$ das Anfangswertproblem (2.1) löst, gilt*

$$y(t) = y_0 + \int_a^t f(s, y(s))\, ds \qquad \text{für alle } t \in [a,b]. \tag{2.2}$$

Falls umgekehrt eine stetige Funktion $y \in C([a,b], V)$ die Integralgleichung (2.2) erfüllt, ist sie auch stetig differenzierbar und löst das Anfangswertproblem (2.1).

Beweis. Im ersten Schritt gehen wir davon aus, dass y das Anfangswertproblem löst. Nach Hauptsatz der Differential- und Integralrechnung gilt dann

$$y_0 + \int_a^t f(s, y(s))\, ds = y_0 + \int_a^t y'(s)\, ds = y(a) + y(t) - y(a) = y(t),$$

für alle $t \in [a,b]$, also die Integralgleichung (2.2).

Im zweiten Schritt gehen wir davon aus, dass $y \in C([a,b], V)$ die Gleichung (2.2) erfüllt. Für $t = a$ folgt aus ihr unmittelbar $y(a) = y_0$. Wir müssen nachweisen, dass y differenzierbar ist und seine Ableitung die gewünschte Eigenschaft aufweist.

Sei $t \in [a,b)$. Die Ableitung $y'(t)$ ist als Grenzwert des Quotienten

$$\frac{y(t+h) - y(t)}{h} = \frac{1}{h}\left(\int_a^{t+h} f(s, y(s))\, ds - \int_a^t f(s, y(s))\, ds\right) = \frac{1}{h}\int_t^{t+h} f(s, y(s))\, ds$$

für $h \to 0$ definiert. Wir müssen zeigen, dass dieser Grenzwert existiert und mit $f(t, y(t))$ übereinstimmt, also sollte

$$\begin{aligned}\left\| f(t, y(t)) - \frac{y(t+h) - y(t)}{h} \right\| &= \left\| \frac{1}{h}\int_t^{t+h} f(t, y(t)) - f(s, y(s))\, ds \right\| \\ &\le \frac{1}{h}\int_t^{t+h} \| f(t, y(t)) - f(s, y(s)) \|\, ds\end{aligned}$$

für $h \to 0$ gegen null konvergieren.

Mit dem Mittelwertsatz der Integralrechnung finden wir einen Zwischenpunkt $\eta \in [t, t+h]$ derart, dass

$$\int_t^{t+h} \| f(t, y(t)) - f(s, y(s)) \|\, ds = h \| f(t, y(t)) - f(\eta, y(\eta)) \|$$

gilt, also folgt

$$\left\| f(t, y(t)) - \frac{y(t+h) - y(t)}{h} \right\| \le \| f(t, y(t)) - f(\eta, y(\eta)) \|.$$

Wegen $\eta \in [t, t+h]$ impliziert $h \to 0$ auch $\eta \to t$, und aufgrund der Stetigkeit der Funktionen f und y dürfen wir auf $f(\eta, y(\eta)) \to f(t, y(t))$ schließen, so dass wir insgesamt

$$\lim_{h \to 0} \left\| f(t, y(t)) - \frac{y(t+h) - y(t)}{h} \right\| = 0$$

bewiesen haben, also $y'(t) = f(t, y(t))$.

Für $t = b$ folgt die Aussage, indem wir entsprechend den linksseitigen Differenzenquotienten zur Approximation der Ableitung einsetzen. ∎

Bei genauerer Betrachtung stellt sich heraus, dass die Integralgleichung (2.2) eine Fixpunktgleichung ist: Wir definieren auf dem Raum $U := C([a,b], V)$ den Operator $\Psi : U \to U$ durch

$$\Psi[y](t) = y_0 + \int_a^t f(s, y(s))\, ds \qquad \text{für alle } y \in U,\ t \in [a,b]$$

und halten fest, dass (2.2) sich dann kurz als

$$y = \Psi[y] \tag{2.3}$$

schreiben lässt. Nach Lemma 2.1 ist also das Lösen des Anfangswertproblems äquivalent dazu, einen Fixpunkt des Operators Ψ zu finden.

Ein zentrales Hilfsmittel für den Beweis von Existenz und Eindeutigkeit von Fixpunkten ist der folgende Fixpunktsatz von Banach.

Satz 2.2 (Banach) *Sei X eine vollständige Teilmenge eines normierten Raumes. Sei $\Psi : X \to X$ eine Abbildung, und sei $L \in [0,1)$ eine Zahl mit*

$$\|\Psi(u) - \Psi(v)\| \le L\|u - v\| \qquad \textit{für alle } u, v \in X. \tag{2.4}$$

Dann besitzt Ψ einen Fixpunkt in X, es existiert also ein $u^* \in X$ mit

$$\Psi(u^*) = u^*.$$

Dieser Fixpunkt ist eindeutig bestimmt.

Beweis. (vgl. [1, Théorème 6] und [4]) Sei $u^{(0)} \in X$. Wir definieren die Folge $(u^{(m)})_{m=0}^{\infty}$ durch

$$u^{(m+1)} = \Psi(u^{(m)}) \qquad \text{für alle } m \in \mathbb{N}_0.$$

Unser Ziel ist es, nachzuweisen, dass $(u^{(m)})_{m=0}^{\infty}$ eine Cauchy-Folge ist.

Zunächst beweisen wir

$$\|u^{(m+1)} - u^{(m)}\| \le L^m \|u^{(1)} - u^{(0)}\| \tag{2.5}$$

durch Induktion für alle $m \in \mathbb{N}_0$. Für $m = 0$ ist (2.5) trivial.

Gelte nun (2.5) für ein $m \in \mathbb{N}_0$. Nach Voraussetzung gilt dann

$$\begin{aligned}\|u^{(m+2)} - u^{(m+1)}\| = \|\Psi(u^{(m+1)}) - \Psi(u^{(m)})\| &\overset{(2.4)}{\leq} L\|u^{(m+1)} - u^{(m)}\| \\ &\leq LL^m\|u^{(1)} - u^{(0)}\| = L^{m+1}\|u^{(1)} - u^{(0)}\|,\end{aligned}$$

und der Induktionsschritt ist bewiesen.

Seien nun $m \in \mathbb{N}_0$ und $n \in \mathbb{N}_{\geq m}$ gegeben. Dann gilt

$$\begin{aligned}\|u^{(n)} - u^{(m)}\| = \left\| \sum_{j=0}^{n-m-1} u^{(m+j+1)} - u^{(m+j)} \right\| &\leq \sum_{j=0}^{m-n-1} \|u^{(m+j+1)} - u^{(m+j)}\| \\ &\overset{(2.5)}{\leq} \sum_{j=0}^{m-n-1} L^{m+j}\|u^{(1)} - u^{(0)}\| = \|u^{(1)} - u^{(0)}\|L^m \sum_{j=0}^{m-n-1} L^j \\ &\leq \|u^{(1)} - u^{(0)}\|L^m \sum_{j=0}^{\infty} L^j = \|u^{(1)} - u^{(0)}\|\frac{L^m}{1-L}\end{aligned}$$

dank der geometrischen Summenformel. Mit dieser Abschätzung können wir nachweisen, dass $(u^{(m)})_{m=0}^{\infty}$ eine Cauchy-Folge ist: Sei $\epsilon \in \mathbb{R}_{>0}$. Wir wählen $m_0 \in \mathbb{N}_0$ so, dass

$$\|u^{(1)} - u^{(0)}\|\frac{L^{m_0}}{1-L} \leq \epsilon$$

gilt. Für alle $m, n \in \mathbb{N}_0$ mit $m_0 \leq m \leq n$ gilt dann

$$\|u^{(n)} - u^{(m)}\| \leq \|u^{(1)} - u^{(0)}\|\frac{L^m}{1-L} \leq \|u^{(1)} - u^{(0)}\|\frac{L^{m_0}}{1-L} \leq \epsilon,$$

also ist $(u^{(m)})_{m=0}^{\infty}$ eine Cauchy-Folge.

Da X vollständig ist, muss es ein $u^* \in X$ mit

$$\lim_{m\to\infty} \|u^* - u^{(m)}\| = 0$$

geben, und wir müssen nur noch nachprüfen, dass u^* auch ein Fixpunkt von Ψ ist.

Sei dazu $\epsilon \in \mathbb{R}_{>0}$. Da $(u^{(m)})_{m=0}^{\infty}$ gegen u^* konvergiert, gibt es ein $m \in \mathbb{N}_0$ so, dass

$$\|u^* - u^{(m)}\| \leq \epsilon/2, \qquad \|u^* - u^{(m+1)}\| \leq \epsilon/2$$

gelten, und wir erhalten wegen $u^{(m+1)} = \Psi(u^{(m)})$ die Abschätzung

$$\begin{aligned}\|u^* - \Psi(u^*)\| &= \|u^* - u^{(m+1)} + \Psi(u^{(m)}) - \Psi(u^*)\| \\ &\leq \|u^* - u^{(m+1)}\| + \|\Psi(u^{(m)}) - \Psi(u^*)\| \\ &\leq \|u^* - u^{(m+1)}\| + L\|u^* - u^{(m)}\| \\ &\leq \epsilon/2 + L\epsilon/2 < \epsilon.\end{aligned}$$

Da ϵ beliebig gewählt werden kann, folgt $u^* = \Psi(u^*)$, also ist u^* in der Tat ein Fixpunkt.

Zum Nachweis der Eindeutigkeit wählen wir einen zweiten Fixpunkt $u^{**} \in X$ und erhalten

$$\|u^* - u^{**}\| = \|\Psi(u^*) - \Psi(u^{**})\| \leq L\|u^* - u^{**}\|,$$

also folgt aus $L < 1$ bereits $u^* = u^{**}$. ∎

Indem wir diesen Satz auf die alternative Formulierung (2.3) anwenden, erhalten wir die folgende für uns zentrale Aussage über Existenz und Eindeutigkeit der Lösung eines Anfangswertproblems.

Satz 2.3 (Picard-Lindelöf) *Die Funktion $f \in C([a,b] \times V, V)$ erfülle die* globale Lipschitz-Bedingung

$$\|f(t,x) - f(t,y)\| \leq L_f\|x - y\| \qquad \textit{für alle } t \in [a,b] \textit{ und } x, y \in V. \tag{2.6}$$

Dann besitzt das Anfangswertproblem (2.1) eine eindeutige Lösung $y \in C^1([a,b], V)$.

Beweis. (vgl. [10, Abschnitt 1.6]) Wir führen den Beweis mit Hilfe des Banachschen Fixpunktsatzes 2.2: Dazu führen wir den Operator Ψ durch

$$\Psi[u](t) := y_0 + \int_a^t f(s, u(s))\, ds \qquad \text{für alle } t \in [a,b], u \in C([a,b], V)$$

ein, der den Banachraum $X := C([a,b], V)$ in sich abbildet, und untersuchen die von ihm induzierte Fixpunktiteration. Nach Lemma 2.1 wissen wir nämlich, dass ein Fixpunkt des Operators Ψ gerade eine Lösung des Anfangswertproblems (2.1) ist.

Damit wir Satz 2.2 anwenden können, müssen wir eine geeignete Norm auf dem Raum $C([a,b], V)$ einführen. Wir verwenden die gewichtete Supremumsnorm

$$\|u\|_e := \sup\left\{e^{-2L_f x}\|u(x)\| \;:\; x \in [a,b]\right\}, \qquad \text{für alle } u \in C([a,b], V),$$

die wegen $0 < e^{-2L_f b} \leq e^{-2L_f a}$ äquivalent zu der üblichen Supremumsnorm ist, so dass $C([a,b], V)$ auch mit dieser Norm vollständig ist. Bezüglich dieser Norm gilt

$$\begin{aligned}
e^{-2L_f t}\|\Psi[u](t) - \Psi[v](t)\| &= e^{-2L_f t}\left\|\int_a^t f(s,u(s)) - f(s,v(s))\, ds\right\| \\
&\leq e^{-2L_f t}\int_a^t \|f(s,u(s)) - f(s,v(s))\|\, ds \\
&\leq L_f e^{-2L_f t}\int_a^t \|u(s) - v(s)\|\, ds \\
&= L_f e^{-2L_f t}\int_a^t e^{2L_f s}e^{-2L_f s}\|u(s) - v(s)\|\, ds \\
&\leq L_f e^{-2L_f t}\int_a^t e^{2L_f s}\|u - v\|_e\, ds
\end{aligned}$$

$$= \frac{1}{2} e^{-2L_f t} \|u - v\|_e \int_a^t 2L_f e^{2L_f s}\, ds$$
$$= \frac{1}{2} e^{-2L_f t} \|u - v\|_e (e^{2L_f t} - e^{2L_f a}) \leq \frac{1}{2} \|u - v\|_e$$

für alle $t \in [a, b]$ und alle $u, v \in C([a, b], V)$, wobei wir im vorletzten Schritt ausgenutzt haben, dass $s \mapsto 2L_f e^{2L_f s}$ die Ableitung der Funktion $s \mapsto e^{2L_f s}$ ist, so dass sich das Integral mit dem Hauptsatz berechnen lässt. Indem wir zu dem Maximum über alle $t \in [a, b]$ übergehen folgt

$$\|\Psi[u] - \Psi[v]\|_e \leq \frac{1}{2} \|u - v\|_e \qquad \text{für alle } u, v \in C([a, b], V),$$

so dass wir Satz 2.2 anwenden können, um zu folgern, dass ein eindeutig bestimmter Fixpunkt $y \in C([a, b], V)$ mit $\Psi[y] = y$ existiert.

Nach Lemma 2.1 ist diese Funktion y auch die eindeutig bestimmte Lösung des Anfangswertproblems. ∎

Satz 2.3 ist nicht nur ein Existenz- und Eindeutigkeitsresultat, er bietet uns auch ein Konstruktionsverfahren für die Lösung des Anfangswertproblems:

Bemerkung 2.4 (Picard-Iteration) *Ausgehend von einer beliebigen Funktion u_0 können wir, wie im Satz 2.2, die Folge $u_{n+1} := \Psi(u_n)$ konstruieren, und Satz 2.3 impliziert, dass diese Folge gegen die Lösung des Anfangswertproblems (2.1) konvergieren wird. Diese Konstruktion trägt den Namen* Picard-Iteration.

Für die Praxis ist diese Konstruktion nur dann anwendbar, wenn sich die einzelnen Iterierten u_n geeignet im Rechner darstellen lassen, etwa mit Hilfe einer Diskretisierung.

2.3 Störungen der Daten

Für die numerische Behandlung des Anfangswertproblems (2.1) ist neben der prinzipiellen Lösbarkeit auch der Einfluss von Störungen relevant, schließlich wird im praktischen Algorithmus in der Regel mit Gleitpunktarithmetik beschränkter Genauigkeit gearbeitet.

Ein wichtiges Hilfsmittel für die Analyse ist die *Grönwallsche Ungleichung*, von der wir hier nur die folgende vereinfachte Variante benötigen:

Lemma 2.5 (Grönwall) *Seien $[a, b] \subseteq \mathbb{R}$ ein Intervall, sei $\alpha \in C[a, b]$ eine monoton wachsende Funktion, sei $\beta \in \mathbb{R}_{\geq 0}$. Falls eine Funktion $u \in C[a, b]$ die Abschätzung*

$$u(t) \leq \alpha(t) + \beta \int_a^t u(s)\, ds \qquad \textit{für alle } t \in [a, b] \tag{2.7}$$

erfüllt, gilt die Ungleichung

$$u(t) \leq \alpha(t) e^{\beta(t-a)} \qquad \textit{für alle } t \in [a, b].$$

Beweis. (vgl. [2] und [5]) Sei $u \in C[a,b]$ eine Funktion, die (2.7) erfüllt.
Wir führen die Hilfsfunktion $v \in C([a,b))$ mit

$$v(t) := e^{-\beta(t-a)} \int_a^t \beta u(s)\, ds \qquad \text{für alle } t \in [a,b]$$

ein und erhalten mit der Produktregel und (2.7) die Abschätzung

$$\begin{aligned} v'(t) &= -\beta e^{-\beta(t-a)} \int_a^t \beta u(s)\, ds + e^{-\beta(t-a)} \beta u(t) \\ &= \beta e^{-\beta(t-a)} \left(u(t) - \int_a^t \beta u(s)\, ds \right) \overset{(2.7)}{\leq} \beta e^{-\beta(t-a)} \alpha(t) \end{aligned}$$

für alle $t \in [a,b]$. Aus $v(a) = 0$ folgt

$$\begin{aligned} e^{-\beta(t-a)} \int_a^t \beta u(s)\, ds = v(t) = v(t) - v(a) = \int_a^t v'(s)\, ds &\leq \beta \int_a^t \alpha(s) e^{-\beta(s-a)}\, ds \\ &\leq \beta \alpha(t) \int_a^t e^{-\beta(s-a)}\, ds \\ &= \beta \alpha(t) \left(-\frac{1}{\beta} \right) (e^{-\beta(t-a)} - e^{-\beta(a-a)}) = \alpha(t) - \alpha(t) e^{-\beta(t-a)} \end{aligned}$$

und indem wir mit $e^{\beta(t-a)}$ multiplizieren

$$\beta \int_a^t u(s)\, ds \leq \alpha(t) e^{\beta(t-a)} - \alpha(t).$$

Durch Einsetzen in (2.7) gelangen wir zu

$$u(t) \leq \alpha(t) + \int_a^t \beta u(s)\, ds \leq \alpha(t) + \alpha(t) e^{\beta(t-a)} - \alpha(t) = \alpha(t) e^{\beta(t-a)},$$

und das ist die zu beweisende Ungleichung. ∎

Mit Hilfe dieses Korollars und des Lemmas 2.1 können wir nun den Einfluss von Störungen der Anfangsdaten untersuchen:

Satz 2.6 (Störungen) *Sei $U \subseteq V$. Die Funktion $f \in C([a,b] \times U, U)$ erfülle die bereits aus Satz 2.3 bekannte globale Lipschitz-Bedingung (2.6). Sei $g \in C([a,b] \times U, U)$ eine weitere Funktion.*

Seien $y_0, z_0 \in U$, und seien $y, z \in C^1([a,b], U)$ Lösungen der Anfangswertprobleme

$$\begin{aligned} &y(a) = y_0, && y'(t) = f(t, y(t)), \\ &z(a) = z_0, && z'(t) = g(t, z(t)) && \text{für alle } t \in [a,b]. \end{aligned}$$

Dann gilt die Abschätzung

$$\|y(t) - z(t)\| \leq e^{L_f(t-a)} \left(\|y_0 - z_0\| + \int_a^t \|f(s, z(s)) - g(s, z(s))\|\, ds \right) \quad \text{für alle } t \in [a,b],$$

kleine Störungen der Anfangsdaten und der rechten Seite führen also auch nur zu kleinen Störungen der Lösung.

Beweis. Mit Lemma 2.1 erhalten wir

$$\begin{aligned} y(t) - z(t) &= y_0 - z_0 + \int_a^t f(s, y(s)) - g(s, z(s))\, ds, \\ \|y(t) - z(t)\| &\leq \|y_0 - z_0\| + \int_a^t \|f(s, y(s)) - g(s, z(s))\|\, ds \\ &\leq \|y_0 - z_0\| + \int_a^t \|f(s, y(s)) - f(s, z(s))\| + \|f(s, z(s)) - g(s, z(s))\|\, ds \\ &\overset{(2.6)}{\leq} \|y_0 - z_0\| + \int_a^t L_f \|y(s) - z(s)\|\, ds + \int_a^t \|f(s, z(s)) - g(s, z(s))\|\, ds. \end{aligned} \tag{2.8}$$

Wir definieren

$$\beta := L_f, \quad \alpha(t) := \|y_0 - z_0\| + \int_a^t \|f(s, z(s)) - g(s, z(s))\|\, ds, \quad u(t) := \|y(t) - z(t)\|$$

und stellen fest, dass (2.8) gerade

$$u(t) \leq \alpha(t) + \beta \int_a^t u(s)\, ds \qquad \text{für alle } t \in [a, b]$$

entspricht. Da β nicht-negativ und α monoton wachsen ist, können wir die Grönwall-Ungleichung aus Lemma 2.5 anwenden und erhalten

$$\begin{aligned} \|y(t) - z(t)\| = u(t) &\leq \alpha(t) e^{\beta(t-a)} \\ &= e^{L_f(t-a)} \left(\|y_0 - z_0\| + \int_a^t \|f(s, z(s)) - g(s, z(s))\|\, ds \right), \end{aligned}$$

also die gewünschte Abschätzung. ∎

Beispiel 2.7 (Entfernung vom Anfangswert) *Neben der offensichtlichen Anwendung auf gestörte Anfangswerte und rechte Seiten lässt sich Satz 2.6 auch anderweitig verwenden.*

Beispielsweise können wir $g = 0$ und $z_0 = y_0$ einsetzen. Dann gilt offenbar $z(t) = y_0$ und wir erhalten

$$\|y(t) - y_0\| \leq e^{L_f(t-a)} \int_a^t \|f(s, y_0)\|\, ds,$$

können also abschätzen, wie schnell sich die Lösung des Anfangswertproblems vom Anfangswert entfernt.

3 Einschrittverfahren

Die exakte Lösung eines Anfangswertproblems der Form (2.1) wird sich im allgemeinen Fall nicht exakt berechnen lassen. Stattdessen müssen wir auf eine Approximation zurückgreifen: Statt nach einer geschlossenen Formel für die Lösung zu suchen, beschränken wir uns darauf, sie nur in einzelnen Punkten $t_0, \dots, t_n \in [a, b]$ näherungsweise zu berechnen.

Wir sind natürlich an Verfahren interessiert, die uns eine möglichst genaue Näherung zur Verfügung stellen, und das bei möglichst geringem Rechen- und Speicheraufwand.

Ein möglicher Zugang wäre etwa die Picard-Iteration (vgl. Bemerkung 2.4): Wir könnten die Iterierten durch ihre Werte in Punkten des Intervalls approximieren und zur Berechnung der Integrale eine Quadraturformel verwenden, die nur diese Punktwerte benötigt. Der Nachteil dieses Zugangs besteht darin, dass *alle* Punktwerte gleichzeitig im Speicher gehalten werden müssen.

Wir suchen stattdessen nach einem Verfahren, bei dem wir die Werte zu den verschiedenen Zeitpunkten der Reihe nach berechnen können. Ein *Einschrittverfahren* versucht, den Wert zu einem Zeitpunkt t_{i+1} nur auf Grundlage des Wertes zum unmittelbar vorangehenden Zeipunkt t_i zu approximieren.

Um die Diskussion der Lösbarkeit zu vermeiden, setzen wir, sofern nicht gesondert erwähnt, im folgenden Kapitel voraus, dass die rechte Seite f des Anfangswertproblems (2.1) im zweiten Argument Lipschitz-stetig ist (vgl. Bedingung (2.6)). Satz 2.3 impliziert dann die eindeutige Lösbarkeit für beliebige Startwerte in V und Startpunkte in $[a, b]$.

3.1 Euler-Verfahren

Wir untersuchen zunächst ein besonders einfaches Einschrittverfahren: Das *Euler-Verfahren* lässt sich aus der in Lemma 2.1 eingeführten Integralformulierung gewinnen. Wir beschränken uns für den Moment auf den Fall $t = b$, für den

$$y(b) = y_0 + \int_a^b f(s, y(s))\, ds$$

gilt. Es bietet sich an, das Integral mit einer Quadraturformel mit Quadraturgewichten $w_0, \dots, w_m \in \mathbb{R}$ und Quadraturpunkten $s_0, \dots, s_m \in [a, b]$ zu approximieren:

$$y(b) \approx y_0 + \sum_{j=0}^{m} w_j f(s_j, y(s_j)).$$

Leider können wir die rechte Seite dieser Gleichung im Allgemeinen nicht auswerten, da uns die Werte $y(s_0), \dots, y(s_m)$ nicht zur Verfügung stehen.

Allerdings kennen wir $y(a) = y_0$, so dass wir immerhin eine Quadraturformel mit $m = 0$ und $s_0 = a$ verwenden könnten. Damit wenigstens konstante Funktionen von dieser Quadraturformel exakt integriert werden, müssen wir das Gewicht $w_0 = b - a$ verwenden und erhalten

$$y(b) \approx \tilde{y}(b) := y_0 + (b-a)f(a, y(a)). \tag{3.1}$$

Falls y zweimal stetig differenzierbar ist, können wir den Fehler wie folgt abschätzen:

Lemma 3.1 (Genauigkeit) *Sei $y \in C^2([a,b], V)$, und sei $\tilde{y}(b)$ wie in (3.1) definiert. Dann existiert ein $\eta \in [a,b]$ mit*

$$\|y(b) - \tilde{y}(b)\| \leq \frac{(b-a)^2}{2} \|y''(\eta)\|.$$

Beweis. Um die Aufgabe etwas besser zugänglich zu machen, führen wir sie auf das Einheitsintervall $[0,1]$ zurück, indem wir die Funktion

$$\hat{y} : [0,1] \to V, \qquad t \mapsto y(a + (b-a)t),$$

untersuchen. Sie erfüllt offenbar $y(a) = \hat{y}(0)$, $y(b) = \hat{y}(1)$ sowie nach Definition des Anfangswertproblems und Kettenregel

$$\tilde{y}(b) = y_0 + (b-a)f(a, y(a)) = y(a) + (b-a)y'(a) = \hat{y}(0) + \hat{y}'(0).$$

Mit dem Hauptsatz der Integral- und Differentialrechnung erhalten wir

$$y(b) - \tilde{y}(b) = \hat{y}(1) - \hat{y}(0) - \hat{y}'(0) = \int_0^1 \hat{y}'(t)\,dt - \hat{y}'(0) = \int_0^1 \hat{y}'(t) - \hat{y}'(0)\,dt.$$

Da mit y auch $\hat{y}$ zweimal stetig differenzierbar ist, können wir den Hauptsatz erneut anwenden, um zu

$$y(b) - \tilde{y}(b) = \int_0^1 \int_0^t \hat{y}''(s)\,ds\,dt$$

zu gelangen. Um die Integrationsgrenzen des inneren Integrals von t unabhängig zu machen, substituieren wir $s = tr$ und erhalten

$$y(b) - \tilde{y}(b) = \int_0^1 t \int_0^1 \hat{y}''(tr)\,dr\,dt.$$

Wir wollen den Beweis mit dem Mittelwertsatz der Integralrechnung abschließen, der nur für reellwertige Funktionen gilt. Also gehen wir zu der Norm über und erhalten

$$\|y(b) - \tilde{y}(b)\| \leq \int_0^1 t \int_0^1 \|\hat{y}''(tr)\|\,dr\,dt.$$

Nun können wir den Mittelwertsatz der Integralrechnung erst auf das äußere und dann auf das innere Integral anwenden, um $\eta_t, \eta_r \in [0,1]$ mit

$$\|y(b) - \tilde{y}(b)\| = \int_0^1 t\,dt \int_0^1 \|\hat{y}''(\eta_t r)\|\,dr = \frac{1}{2}\|\hat{y}''(\eta_t \eta_r)\|$$

zu finden. Per Kettenregel folgt

$$\|y(b) - \tilde{y}(b)\| = \|\frac{1}{2}\hat{y}''(\eta_t\eta_r)\| = \|\frac{(b-a)^2}{2}y''(a + (b-a)\eta_t\eta_r)\|,$$

also mit $\eta := a + (b-a)\eta_t\eta_r$ die Behauptung. ∎

Bemerkung 3.2 (Optimale Abschätzung) *Aus Lemma 3.1 folgt die Abschätzung*

$$\|y(b) - \tilde{y}(b)\| \leq \frac{(b-a)^2}{2}\|y''\|_{\infty,[a,b]}.$$

Für den Fall $V = \mathbb{R}$ können wir den Beweis des Lemmas 3.1 so modifizieren, dass wir

$$y(b) - \tilde{y}(b) = \frac{(b-a)^2}{2}y''(\eta)$$

für ein $\eta \in [a,b]$ erhalten. Wenn wir die Abschätzung auf $y(t) = (t-a)^2/2$ anwenden, gelten $y'(t) = t - a$, $y''(t) = 1$ und $\tilde{y}(b) = 0$, so dass wir

$$y(b) - \tilde{y}(b) = \frac{(b-a)^2}{2} - 0 = \frac{(b-a)^2}{2}\|y''\|_{\infty,[a,b]}$$

erhalten. Wir haben also ein Beispiel gefunden, in dem sich unsere Abschätzung nicht verbessern lässt.

Offenbar ist der Fehler um so kleiner, je kürzer das Intervall ist, auf dem die Näherung verwendet wird. Deshalb zerlegen wir das Intervall $[a,b]$ in Teilintervalle: Wir wählen $n \in \mathbb{N}$ sowie $t_0, \ldots, t_n \in [a,b]$ mit $a = t_0 < t_1 < \ldots < t_n = b$. Unser Ziel ist es, Näherungswerte der Lösung y in diesen Punkten zu berechnen.

Mit dem Hauptsatz der Integral- und Differentialrechnung finden wir

$$y(t_i) = y(t_{i-1}) + \int_{t_{i-1}}^{t_i} y'(s)\,ds = y(t_{i-1}) + \int_{t_{i-1}}^{t_i} f(s, y(s))\,ds \quad \text{für alle } i \in \{1, \ldots, n\},$$

und indem wir (3.1) auf die Intervalle $[t_{i-1}, t_i]$ anwenden, folgt

$$y(t_i) \approx y(t_{i-1}) + (t_i - t_{i-1})f(t_{i-1}, y(t_{i-1})) \qquad \text{für alle } i \in \{1, \ldots, n\}.$$

Um die Formel etwas zu verkürzen definieren wir die *Schrittweiten*

$$h_i := t_i - t_{i-1} \qquad \text{für alle } i \in \{1, \ldots, n\}$$

und schreiben die Gleichung in der Form

$$y(t_i) \approx y(t_{i-1}) + h_i f(t_{i-1}, y(t_{i-1})) \qquad \text{für alle } i \in \{1, \ldots, n\}.$$

Da wir $y(t_1), \dots, y(t_{n-1})$ nicht kennen, können wir diese Approximation nicht direkt einsetzen, wir können allerdings *der Reihe nach* Näherungslösungen berechnen, die die Stelle der exakten Werte annehmen. Damit erhalten wir die Rechenvorschrift

$$\tilde{y}(t_0) := y_0, \qquad \tilde{y}(t_i) := \tilde{y}(t_{i-1}) + h_i f(t_{i-1}, \tilde{y}(t_{i-1})) \qquad \text{für alle } i \in \{1, \dots, n\}.$$

Offenbar ist dieses Verfahren sehr effizient durchführbar: In jedem Schritt muss f einmal ausgewertet und eine Linearkombination berechnet werden, und es brauchen nur jeweils $\tilde{y}(t_i)$ und $\tilde{y}(t_{i-1})$ gleichzeitig im Speicher gehalten zu werden. Da der Wert $\tilde{y}(t_i)$ jeweils direkt berechnet werden kann, spricht man von einem *expliziten* Verfahren, nämlich von dem *expliziten Euler-Verfahren.*

Bemerkung 3.3 (Diskretisierung) *Das Euler-Verfahren ist das erste* Diskretisierungsverfahren*, das wir behandeln. Der Name stammt daher, dass das kontinuierliche Intervall $[a, b]$ durch die diskrete Punktmenge $\{t_0, \dots, t_n\}$ ersetzt wird. Im allgemeinen Fall spricht man schon von einer Diskretisierung, wenn ein unendlich-dimensionaler Funktionenraum durch einen endlich-dimensionalen Raum ersetzt wird. Ein derartiger Schritt ist fast immer erforderlich, wenn Differentialgleichungen mit Hilfe eines Computers gelöst werden sollen, da einem Computer nur endlich viel Speicher und seinem Benutzer nur endlich viel Zeit zur Verfügung steht.*

Wir können uns bei der Approximation der Gleichung

$$y(b) = y(a) + \int_a^b f(s, y(s))\, ds$$

auch auf eine Quadraturformel stützen, die $y(b)$ statt $y(a)$ verwendet. So erhalten wir

$$y(b) \approx y(a) + (b - a) f(b, y(b)),$$

also die Näherung einer Fixpunktgleichung. Dementsprechend können wir eine Näherung $\tilde{y}(b)$ des Werts $y(b)$ durch

$$\tilde{y}(b) = y(a) + (b - a) f(b, \tilde{y}(b)) \tag{3.2}$$

definieren, falls sich dieses, im allgemeinen nichtlineare, Gleichungssystem lösen lässt.

Ein brauchbarer Ansatz hierzu ist eine Fixpunkt-Iteration mit dem Operator

$$\Psi(x) := y(a) + (b - a) f(b, x) \qquad \text{für alle } x \in V.$$

Falls f Lipschitz-stetig im zweiten Argument ist, also

$$\|f(b, x_1) - f(b, x_2)\| \leq L_f \|x_1 - x_2\| \qquad \text{für alle } x_1, x_2 \in V$$

gilt, erhalten wir

$$\|\Psi(x_1) - \Psi(x_2)\| = \|(b - a) f(b, x_1) - (b - a) f(b, x_2)\|$$

$$\leq (b-a)L_f\|x_1 - x_2\| \qquad \text{für alle } x_1, x_2 \in V,$$

und der Satz 2.2 von Banach garantiert Konvergenz gegen einen eindeutig bestimmten Fixpunkt x^*, falls wir $b - a < 1/L_f$ sicherstellen können. Dieser Fixpunkt erfüllt

$$x^* = \Psi(x^*) = y(a) + (b-a)f(b, x^*),$$

ist also die gesuchte Lösung $\tilde{y}(b)$ der Gleichung (3.2). Je kleiner $b-a$ wird, desto schneller konvergiert die Fixpunktiteration. Falls f hinreichend oft differenzierbar ist und gute Startwerte bekannt sind, kann man natürlich statt der Fixpunkt-Iteration auch alternative Ansätze wie beispielsweise das Newton-Verfahren verwenden, um $\tilde{y}(b)$ zu berechnen.

Entsprechend der Vorgehensweise für das explizite Euler-Verfahren können wir die Näherungswerte wieder der Reihe nach berechnen und erhalten die Vorschrift

$$\tilde{y}(t_0) := y_0, \qquad \tilde{y}(t_i) = \tilde{y}(t_{i-1}) + hf(t_i, \tilde{y}(t_i)) \qquad \text{für alle } i \in \{1, \ldots, n\}$$

des *impliziten Euler-Verfahrens.*

Bemerkung 3.4 (Genauigkeit) *Die Analyse des Approximationsfehlers für das implizite Euler-Verfahren gestaltet sich etwas schwieriger als für die explizite Variante. Wir können von*

$$\begin{aligned} y(b) - \tilde{y}(b) &= y(b) - y(a) - (b-a)f(b, \tilde{y}(b)) \\ &= y(b) - y(a) - (b-a)f(b, y(b)) + (b-a)(f(b, y(b)) - f(b, \tilde{y}(b))) \end{aligned}$$

ausgehen, den ersten Term wie in Lemma 3.1 behandeln und den zweiten mit Hilfe der Lipschitz-Stetigkeit der Funktion f abschätzen, um

$$\|y(b) - \tilde{y}(b)\| \leq \frac{(b-a)^2}{2(1 - L_f(b-a))}\|y''(\eta)\| \tag{3.3}$$

für ein $\eta \in [a, b]$ zu erhalten.

3.2 Konvergenz

Natürlich ist das Euler-Verfahren nur dann nützlich, wenn es auch eine hinreichend gute Approximation der tatsächlichen Lösung berechnet. Wir müssen also untersuchen, ob und, falls ja, wie schnell die approximative Lösung gegen die echte Lösung konvergiert. Dazu gehen wir davon aus, dass die Voraussetzungen des Existenzsatzes 2.3 erfüllt sind.

Die Theorie basiert auf Vergleichen zwischen exakten und approximativen Lösungen zu verschiedenen Startwerten: Wir wollen auch Lösungen untersuchen können, die zu anderen Startzeitpunkten $t_* \in [a, b]$ von anderen Startwerten $y_* \in V$ ausgehen. Glücklicherweise ist auch deren Existenz durch den erwähnten Satz gesichert:

Lemma 3.5 (Partielle Lösungen) *Sei $t_* \in [a,b]$, und sei $y_* \in V$. Dann existiert eine Funktion $y(\cdot; t_*, y_*) \in C^1([t_*, b], V)$, die die Gleichungen*

$$y(t_*; t_*, y_*) = y_*, \qquad \frac{\partial}{\partial t} y(t; t_*, y_*) = f(t, y(t; t_*, y_*)) \qquad \text{für alle } t \in [t_*, b] \tag{3.4}$$

erfüllt. Die Funktion $y(\cdot; t_*, y_*)$ ist durch diese Gleichungen eindeutig bestimmt.

Beweis. Satz 2.3 angewendet auf dem Teilintervall $[t_*, b]$. ∎

Eine wichtige Konsequenz der Eindeutigkeit der Lösungen besteht darin, dass zwei partielle Lösungen, die in einem Punkt übereinstimmen, bereits auf dem gesamten Definitionsbereich identisch sein müssen:

Lemma 3.6 (Fortsetzung) *Seien $t_*, s_* \in [a,b]$ mit $t_* \leq s_*$ gegeben, und sei $y_* \in V$. Dann gilt*

$$y(t; t_*, y_*) = y(t; s_*, y(s_*; t_*, y_*)) \qquad \text{für alle } t \in [s_*, b]. \tag{3.5}$$

Beweis. Sei $x_* = y(s_*; t_*, y_*)$ der Wert der Funktion $y(\cdot; t_*, y_*)$ in dem Zwischenpunkt s_*. Nach Definition (3.4) gilt

$$y(s_*; s_*, x_*) = x_* = y(s_*; t_*, y_*),$$

also stimmen die Funktionen $y(\cdot; s_*, x_*)$ und $y(\cdot; t_*, y_*)$ im Punkt s_* überein. Dieselbe Definition beinhaltet auch

$$\frac{\partial}{\partial t} y(t; s_*, x_*) = f(t, y(t; s_*, x_*)) \qquad \text{für alle } t \in [s_*, b].$$

Wegen $t_* \leq s_*$ gilt außerdem

$$\frac{\partial}{\partial t} y(t; t_*, y_*) = f(t, y(t; t_*, y_*)) \qquad \text{für alle } t \in [s_*, b],$$

also erfüllen $y(\cdot; t_*, y_*)$ und $y(\cdot; s_*, x_*)$ auf $[s_*, b]$ dieselbe Differentialgleichung mit demselben Anfangswert. Nach Satz 2.3 müssen sie deshalb identisch sein. ∎

Nun benötigen wir eine ähnliche Aussage für die approximativen Lösungen. Wir untersuchen ein allgemeines Einschrittverfahren:

Definition 3.7 (Einschrittverfahren) *Sei $h_0 \in \mathbb{R}_{>0} \cup \{\infty\}$. Für die Menge*

$$\Delta := \{(t,h) \ : \ t \in [a,b], \ h \in [0, b-t] \cap [0, h_0]\} \tag{3.6}$$

sei eine Funktion

$$\Phi : \Delta \times V \to V$$

fixiert. Diese Funktion definiert ein Einschrittverfahren durch

$$\tilde{y}(t_i) = \tilde{y}(t_{i-1}) + h_i \Phi(t_{i-1}, h_i, \tilde{y}(t_{i-1})) \qquad \text{für alle } i \in \{1, \ldots, n\}, \tag{3.7}$$

falls

$$(t_{i-1}, h_i) \in \Delta \qquad \text{für alle } i \in \{1, \dots, n\}$$

gilt. Die Funktion Φ *bezeichnen wir in diesem Kontext als die* Verfahrensfunktion *des Einschrittverfahrens, während wir* h_0 *als die* maximale Schrittweite *bezeichnen.*

In der Definition verwenden wir die Menge Δ, um sicher zu stellen, dass für jedes Paar $(t, h) \in \Delta$ auch der nächste zu berechnende Punkt $t + h$ im Intervall $[a, b]$ enthalten ist. Damit ist sicher gestellt, dass unsere Algorithmen wohldefiniert sind. Die Schranke h_0 für die Schrittweite ist erforderlich, um beispielsweise bei dem impliziten Euler-Verfahren die Lösbarkeit der definierenden Fixpunktgleichung sicher zu stellen. Mit Hilfe einer Verfahrensfunktion können wir, ähnlich wie in Lemma 3.5, partielle diskrete Lösungen definieren:

Definition 3.8 (Diskrete partielle Lösungen) *Für beliebige* $j \in \{0, \dots, n\}$ *und* $y_j \in V$ *definieren wir analog zu (3.7) die Werte*

$$\tilde{y}(t_i; t_j, y_j) := \begin{cases} y_j & \text{falls } i = j, \\ \tilde{y}(t_{i-1}; t_j, y_j) & \\ \quad + h_i \Phi(t_{i-1}, h_i, \tilde{y}(t_{i-1}; t_j, y_j)) & \text{ansonsten} \end{cases} \qquad \text{für } i \in \{j, \dots, n\}.$$

Aus der Definition folgt insbesondere

$$\tilde{y}(t_i; t_{i-1}, x) - x = \tilde{y}(t_{i-1}; t_{i-1}, x) + h_i \Phi(t_{i-1}, h_i, \tilde{y}(t_{i-1}; t_{i-1}, x)) - x = h_i \Phi(t_{i-1}, h_i, x),$$

so dass es sich anbietet, die Gleichung

$$\Phi(t, h, x) = \frac{\tilde{y}(t + h; t, x) - x}{h} \qquad \text{für alle } (t, h) \in \Delta,\ x \in V \tag{3.8}$$

zu verwenden, um den bisher definierten Näherungsverfahren eine Verfahrensfunktion zuzuordnen.

Für das explizite Euler-Verfahren erhalten wir

$$\Phi(t, h, x) = \frac{x + hf(t, x) - x}{h} = f(t, x) \qquad \text{für alle } (t, h) \in \Delta,\ x \in V$$

und dürfen $h_0 = \infty$ wählen. Für das implizite Euler-Verfahren ist es etwas schwieriger, Φ zu definieren: Für $(t, h) \in \Delta$ und $x \in V$ gilt nach Definition

$$\tilde{y}(t + h; t, x) = x + hf(t + h, \tilde{y}(t + h; t, x)),$$

also ergibt sich

$$\begin{aligned} \Phi(t, h, x) &= \frac{x + hf(t + h, \tilde{y}(t + h; t, x)) - x}{h} \\ &= f(t + h, \tilde{y}(t + h; t, x)) = f(t + h, x + h\Phi(t, h, x)). \end{aligned}$$

Demnach ist $z = \Phi(t, h, x)$ Lösung der Fixpunktgleichung

$$z = f(t + h, x + hz). \tag{3.9}$$

Um den Satz 2.2 von Banach anwenden zu können, definieren wir

$$\Psi : V \to V, \qquad z \mapsto f(t + h, x + hz),$$

und stellen fest, dass

$$\begin{aligned}\|\Psi(z_1) - \Psi(z_2)\| &= \|f(t + h, x + hz_1) - f(t + h, x + hz_2)\| \\ &\leq L_f \|x + hz_1 - x - hz_2\| = L_f h \|z_1 - z_2\| \qquad \text{für alle } z_1, z_2 \in V\end{aligned}$$

gilt. Für $h < 1/L_f$ ist Ψ also eine Kontraktion und damit der Fixpunkt $z = \Phi(t, h, x)$ nach Satz 2.2 eindeutig definiert. Wir können sogar (3.9) so umformulieren, dass wir einen expliziten (wenn auch unhandlichen) Ausdruck für z erhalten:

$$\begin{aligned} z &= f(t + h, x + hz), \\ z - f(t + h, x + hz) &= 0, \\ (\bullet - f(t + h, x + h\bullet))(z) &= 0, \\ z &= (\bullet - f(t + h, x + h\bullet))^{-1}(0), \end{aligned}$$

wobei in der dritten Zeile eine Funktion $z \mapsto z - f(t+h, x+hz)$ definiert wird, indem mit „$\bullet$“ die Stellen bezeichnet werden, an denen das Argument eingesetzt werden soll. In der vierten Zeile wird dann ihre Umkehrfunktion verwendet, die aufgrund der eindeutigen Lösbarkeit der Fixpunktgleichung wohldefiniert ist. Für das implizite Euler-Verfahren können wir demnach ein $h_0 < 1/L_f$ wählen und

$$\Phi(t, h, x) = (\bullet - f(t + h, x + h\bullet))^{-1}(0) \qquad \text{für alle } (t, h) \in \Delta,\ x \in V$$

verwenden. Beide Varianten des Euler-Verfahrens lassen sich also in der beschriebenen Form darstellen.

Auch für die diskreten Näherungen der Lösung sind wir daran interessiert, eine Fortsetzungseigenschaft nachzuweisen:

Lemma 3.9 (Diskrete Fortsetzung) *Seien $i, j \in \{0, \ldots, n\}$ mit $i \leq j$ gegeben, und sei $y_i \in V$. Dann gilt*

$$\tilde{y}(t_k; t_i, y_*) = \tilde{y}(t_k; t_j, \tilde{y}(t_j; t_i, y_*)) \qquad \textit{für alle } k \in \{j, \ldots, n\}. \tag{3.10}$$

Beweis. Per Induktion über $m := k - j \in \mathbb{N}_0$.

Für $m = 0$ haben wir $k = j$, und es gilt

$$\tilde{y}(t_k; t_i, y_*) = \tilde{y}(t_k; t_k, \tilde{y}(t_j; t_i, y_*)) = \tilde{y}(t_k; t_j, \tilde{y}(t_j; t_i, y_*))$$

nach Definition.

Sei nun $m \in \mathbb{N}_0$ so gewählt, dass (3.10) für alle $k, j \in \{0, \dots, n\}$ mit $k - j \leq m$ gilt. Seien $k, j \in \{0, \dots, n\}$ mit $k - j = m + 1$ gewählt. Sei $y_j := \tilde{y}(t_j; t_i, y_i)$. Aus $k - 1 - j = m \geq 0$ folgt insbesondere $k > j$, und damit

$$\tilde{y}(t_k; t_i, y_i) = \tilde{y}(t_{k-1}; t_i, y_i) + h_k \Phi(t_{k-1}, h_k, \tilde{y}(t_{k-1}; t_i, y_i)).$$

Wegen $k - 1 - j = m$ können wir die Induktionsvoraussetzung anwenden und finden die Gleichung

$$\tilde{y}(t_{k-1}; t_i, y_i) = \tilde{y}(t_{k-1}; t_j, \tilde{y}(t_j; t_i, y_i)) = \tilde{y}(t_{k-1}; t_j, y_j),$$

die wir einsetzen können, um

$$\begin{aligned} \tilde{y}(t_k; t_i, y_i) &= \tilde{y}(t_{k-1}; t_j, y_j) + h_k \Phi(t_{k-1}, h_k, \tilde{y}(t_{k-1}; t_j, y_j)) \\ &= \tilde{y}(t_k; t_j, y_j) = \tilde{y}(t_k; t_j, \tilde{y}(t_j; t_i, y_i)) \end{aligned}$$

zu erhalten. Damit ist die Induktion vollständig. ∎

Mit Hilfe der Fortsetzungseigenschaften (3.5) und (3.10) können wir nun eine Darstellung für den Approximationsfehler finden: Wir wählen $i, j, k \in \{0, \dots, n\}$ mit $k \geq j \geq i$ und $y_i \in V$. Dank Lemma 3.6 wissen wir, dass

$$y(t_k; t_i, y_i) = y(t_k; t_j, y(t_j; t_i, y_i))$$

gilt. Aus Lemma 3.9 folgt, dass auch

$$\tilde{y}(t_k; t_i, y_i) = \tilde{y}(t_k; t_j, \tilde{y}(t_j; t_i, y_i))$$

gelten muss. Um eine Beziehung zwischen beiden Werten herzustellen, führen wir die diskrete Lösung ein, die ausgehend von dem Zeitpunkt t_j mit dem exakten Startwert $y(t_j; t_i, y_i)$ konstruiert wird, wir stellen den Fehler also in der Form

$$\begin{aligned} y(t_k; t_i, y_i) - \tilde{y}(t_k; t_i, y_i) &= y(t_k; t_j, y(t_j; t_i, y_i)) - \tilde{y}(t_k; t_j, \tilde{y}(t_j; t_i, y_i)) \\ &= y(t_k; t_j, y(t_j; t_i, y_i)) - \tilde{y}(t_k; t_j, y(t_j; t_i, y_i)) \\ &\quad + \tilde{y}(t_k; t_j, y(t_j; t_i, y_i)) - \tilde{y}(t_k; t_j, \tilde{y}(t_j; t_i, y_i)) \end{aligned}$$

dar. Die erste Zeile beschreibt dabei den Fehler, der auf dem Teilintervall $[t_j, t_k]$ durch die Diskretisierung entsteht, während die zweite Zeile beschreibt, wie sich die Störung im Anfangswert in t_j fortpflanzt.

Falls es uns gelingt, diesen zweiten Term unter Kontrolle zu bringen, können wir den Fehler mit Hilfe einer einfachen Induktion über die Länge des Teilintervalls abschätzen. Dazu verwenden wir ein diskretes Gegenstück der Stabilitätsaussage aus Satz 2.6.

Definition 3.10 (Stabilität) *Sei Φ eine Verfahrensfunktion. Sie heißt* stabil*, falls eine Konstante $L_\Phi \in \mathbb{R}_{\geq 0}$ so existiert, dass*

$$\|\Phi(t, h, x) - \Phi(t, h, z)\| \leq L_\Phi \|x - z\| \qquad \text{für alle } x, z \in V,\ (t, h) \in \Delta \tag{3.11}$$

gilt, also Lipschitz-Stetigkeit im letzten Argument. In diesem Fall nennen wir auch das zugehörige Einschrittverfahren stabil und bezeichnen L_Φ als die Stabilitätskonstante*.*

Bei dieser Definition ist zu beachten, dass die Stabilität eines Einschrittverfahrens von der rechten Seite f des Anfangswertproblems (2.1) abhängt.

Wir sind daran interessiert, ein diskretes Gegenstück des Störungssatzes 2.6 zu beweisen. Dazu fixieren wir eine zweite Verfahrensfunktion Ψ und führen analog zu Definition 3.8 die korrespondierenden diskreten Lösungen zu $j \in \{0, \dots, n\}$ und Anfangswerten $z_j \in V$ durch

$$\tilde{z}(t_i; t_j, z_j) := \begin{cases} z_j & \text{falls } i = j, \\ \tilde{z}(t_{i-1}; t_j, z_j) & \\ \quad + h_i \Psi(t_{i-1}, h_i, \tilde{z}(t_{i-1}; t_j, z_j)) & \text{ansonsten} \end{cases} \qquad \text{für } i \in \{j, \dots, n\}$$

ein. Der Unterschied zwischen den Näherungslösungen $\tilde{y}$ und $\tilde{z}$ lässt sich wie folgt abschätzen:

Lemma 3.11 (Diskrete Störungen) *Sei Φ stabil mit Stabilitätskonstante L_Φ. Sei $j \in \{0, \dots, n\}$, und seien $y_j, z_j \in V$ gegeben. Zur Abkürzung setzen wir*

$$\tilde{y}_i := \tilde{y}(t_i; t_j, y_j), \qquad \tilde{z}_i := \tilde{z}(t_i; t_j, z_j) \qquad \textit{für alle } i \in \{j, \dots, n\}.$$

Dann gilt die Abschätzung

$$\begin{aligned} &\|\tilde{y}(t_i; t_j, y_j) - \tilde{z}(t_i; t_j, z_j)\| \\ &\quad \le e^{L_\Phi (t_i - t_j)} \|y_j - z_j\| \qquad \textit{für alle } i \in \{j, \dots, n\} \\ &\qquad + \sum_{k=j}^{i-1} h_{k+1} e^{L_\Phi (t_i - t_{k+1})} \|\Phi(t_k, h_{k+1}, \tilde{z}_k) - \Psi(t_k, h_{k+1}, \tilde{z}_k)\|. \end{aligned}$$

Beweis. Wir beweisen die Abschätzung per Induktion über $i \in \{j, \dots, n\}$.

Induktionsanfang. Gelte $i = j$. Dann folgt die Abschätzung direkt aus der Definition.

Induktionsvoraussetzung. Sei $i \in \{j, \dots, n-1\}$ so gegeben, dass die Abschätzung gilt.

Induktionsschritt. Nach Definition erhalten wir

$$\begin{aligned} \|\tilde{y}_{i+1} - \tilde{z}_{i+1}\| &= \|\tilde{y}_i + h_{i+1}\Phi(t_i, h_{i+1}, \tilde{y}_i) - \tilde{z}_i - h_{i+1}\Psi(t_i, h_{i+1}, \tilde{z}_i)\| \\ &\le \|\tilde{y}_i - \tilde{z}_i\| + h_{i+1}\|\Phi(t_i, h_{i+1}, \tilde{y}_i) - \Psi(t_i, h_{i+1}, \tilde{z}_i)\| \\ &\le \|\tilde{y}_i - \tilde{z}_i\| + h_{i+1}\|\Phi(t_i, h_{i+1}, \tilde{y}_i) - \Phi(t_i, h_{i+1}, \tilde{z}_i)\| \\ &\qquad + h_{i+1}\|\Phi(t_i, h_{i+1}, \tilde{z}_i) - \Psi(t_i, h_{i+1}, \tilde{z}_i)\| \\ &\le \|\tilde{y}_i - \tilde{z}_i\| + h_{i+1}L_\Phi\|\tilde{y}_i - \tilde{z}_i\| \\ &\qquad + h_{i+1}\|\Phi(t_i, h_{i+1}, \tilde{z}_i) - \Psi(t_i, h_{i+1}, \tilde{z}_i)\| \\ &\le (1 + h_{i+1}L_\Phi)\|\tilde{y}_i - \tilde{z}_i\| \\ &\qquad + h_{i+1}\|\Phi(t_i, h_{i+1}, \tilde{z}_i) - \Psi(t_i, h_{i+1}, \tilde{z}_i)\| \\ &\le e^{L_\Phi h_{i+1}}\|\tilde{y}_i - \tilde{z}_i\| \\ &\qquad + h_{i+1}\|\Phi(t_i, h_{i+1}, \tilde{z}_i) - \Psi(t_i, h_{i+1}, \tilde{z}_i)\|. \end{aligned}$$

wobei wir im letzten Schritt $1 + s \leq e^s$ ausgenutzt haben. Mit der Induktionsvoraussetzung können wir nun auf

$$\begin{aligned}
\|\tilde{y}_{i+1} - \tilde{z}_{i+1}\| &\leq e^{L_\Phi h_{i+1}} \Bigg(e^{L_\Phi(t_i - t_j)} \|y_j - z_j\| \\
&\qquad + \sum_{k=j}^{i-1} h_{k+1} e^{L_\Phi(t_i - t_{k+1})} \|\Phi(t_k, h_{k+1}, \tilde{z}_k) - \Psi(t_k, h_{k+1}, \tilde{z}_k)\| \Bigg) \\
&\quad + h_{i+1} \|\Phi(t_i, h_{i+1}, \tilde{z}_i) - \Psi(t_i, h_{i+1}, \tilde{z}_i)\| \\
&= e^{L_\Phi(t_{i+1} - t_j)} \|y_j - z_j\| \\
&\quad + \sum_{k=j}^{i-1} h_{k+1} e^{L_\Phi(t_{i+1} - t_{k+1})} \|\Phi(t_k, h_{k+1}, \tilde{z}_k) - \Psi(t_k, h_{k+1}, \tilde{z}_k) \\
&\quad + h_{i+1} e^{L_\Phi(t_{i+1} - t_{i+1})} \|\Phi(t_i, h_{i+1}, \tilde{z}_i) - \Psi(t_i, h_{i+1}, \tilde{z}_i)\| \\
&= e^{L_\Phi(t_{i+1} - t_j)} \|y_j - z_j\| \\
&\quad + \sum_{k=j}^{i} h_{k+1} e^{L_\Phi(t_{i+1} - t_{k+1})} \|\Phi(t_k, h_{k+1}, \tilde{z}_k) - \Psi(t_k, h_{k+1}, \tilde{z}_k)\|
\end{aligned}$$

schließen. Damit ist der Induktionsbeweis vollständig. ∎

Bemerkung 3.12 (Vergleich mit Störungssatz) *Indem wir Lemma 3.11 auf $i = n$ und $j = 0$ anwenden, den exponentiellen Faktor in der Summe verschwenderisch durch $e^{L_\Phi(b-a)}$ abschätzen und ihn aus der Summe herausziehen erhalten wir*

$$\|\tilde{y}(t_i) - \tilde{z}(t_i)\| \leq e^{L_\Phi(b-a)} \left(\|y_0 - z_0\| + \sum_{k=0}^{n-1} h_{k+1} \|\Phi(t_k, h_{k+1}, \tilde{z}_k) - \Psi(t_k, h_{k+1}, \tilde{z}_k)\| \right).$$

Diese Abschätzung weist eine gewisse Ähnlichkeit zu der in Satz 2.6 gewonnenen auf: Die Stabilitätskonstante L_Φ tritt an die Stelle der Lipschitzkonstanten L_f, die Summe tritt an die Stelle des Integrals.

Anstelle der Lipschitz-Stetigkeit von f setzt diese Abschätzung die der Verfahrensfunktion Φ voraus. Im Falle des expliziten Euler-Verfahrens gilt $\Phi(t, h, x) = f(t, x)$, und da wir bereits vorausgesetzt haben, dass f im zweiten Argument Lipschitz-stetig im zweiten Argument mit der Lipschitz-Konstanten L_f ist, folgt direkt, dass Φ mit der Stabilitätskonstanten $L_\Phi = L_f$ stabil ist.

Für das implizite Euler-Verfahren ist die Situation wieder etwas komplizierter: Seien $(t, h) \in \Delta$ und $x_1, x_2 \in V$ gegeben. Dann sind $z_1 := \Phi(t, h, x_1)$ und $z_2 := \Phi(t, h, x_2)$ nach (3.9) gegeben durch die Fixpunktgleichungen

$$z_1 = f(t + h, x_1 + h z_1), \qquad z_2 = f(t + h, x_2 + h z_2),$$

also erhalten wir

$$\begin{aligned}\|z_1 - z_2\| &= \|f(t+h, x_1 + hz_1) - f(t+h, x_2 + hz_2)\| \leq L_f \|x_1 + hz_1 - x_2 - hz_2\| \\ &\leq L_f(\|x_1 - x_2\| + h\|z_1 - z_2\|) = L_f\|x_1 - x_2\| + L_f h\|z_1 - z_2\|.\end{aligned}$$

Indem wir den zweiten Summanden auf die linke Seite bringen folgt

$$(1 - L_f h)\|z_1 - z_2\| \leq L_f\|x_1 - x_2\|,$$

und da $h \leq h_0 < 1/L_f$ vorausgesetzt ist, dürfen wir durch $1 - L_f h$ dividieren, um

$$\|\Phi(t, h, x_1) - \Phi(t, h, x_2)\| = \|z_1 - z_2\| \leq \frac{L_f}{1 - L_f h}\|x_1 - x_2\| \leq \frac{L_f}{1 - L_f h_0}\|x_1 - x_2\|$$

zu erhalten. Also ist auch die Verfahrensfunktion des impliziten Euler-Verfahrens stabil, und die Stabilitätskonstante $L_\Phi = L_f/(1 - L_f h_0)$ konvergiert gegen L_f, falls die Schrittweiten gegen null gehen.

Mit Hilfe des soeben bewiesenen Stabilitätsresultats können wir uns nun dem Nachweis der Konvergenz des allgemeinen Einschrittverfahrens zuwenden.

Satz 3.13 (Konvergenz) *Sei Φ eine stabile Verfahrensfunktion mit der Stabilitätskonstanten L_Φ. Wir bezeichnen mit*

$$y_i := y(t_i), \qquad \tilde{y}_i := \tilde{y}(t_i), \qquad \text{für alle } i \in \{0, \dots, n\}$$

die Werte der exakten und diskreten Lösung des Anfangswertproblems (2.1) und mit

$$K_\Phi := \max\left\{ \frac{\|y(t_i; t_{i-1}, y_{i-1}) - \tilde{y}(t_i; t_{i-1}, y_{i-1})\|}{h_i} \;:\; i \in \{1, \dots, n\} \right\} \tag{3.12}$$

den maximalen Fehler, den das Einschrittverfahren in einem Schritt *relativ zur Schrittweite verursachen kann.*

Dann gilt die Abschätzung

$$\|y(t_k) - \tilde{y}(t_k)\| \leq \begin{cases} \frac{e^{L_\Phi(t_k - a)} - 1}{L_\Phi} K_\Phi & \text{falls } L_\Phi > 0, \\ (t_k - a) K_\Phi & \text{ansonsten} \end{cases} \qquad \text{für alle } k \in \{0, \dots, n\}. \tag{3.13}$$

Beweis. Falls Φ stabil mit der Konstanten $L_\Phi = 0$ ist, ist es nach Definition auch stabil für ein beliebiges $L_\Phi > 0$. Deshalb konzentrieren wir uns zunächst auf diesen Fall. Sei also $L_\Phi > 0$.

Wir beweisen (3.13) per Induktion über $k \in \{0, \dots, n\}$.

Induktionsanfang. Sei $k = 0$. Dann gilt nach Definition $y(t_k) = y(t_0) = \tilde{y}(t_0) = \tilde{y}(t_k)$ und wegen $e^{L_\Phi(t_k - a)} - 1 = e^0 - 1 = 0$ ist (3.13) erfüllt.

Induktionsvoraussetzung. Sei $k \in \{0, \dots, n-1\}$ so gegeben, dass (3.13) gilt.

Induktionsschritt. Nach Lemma 3.6 gilt

$$\|y(t_{k+1}) - \tilde{y}(t_{k+1})\| = \|y(t_{k+1}; t_k, y_k) - \tilde{y}(t_{k+1}; t_k, \tilde{y}_k)\|$$

$$\begin{aligned} &= \|y(t_{k+1}; t_k, y_k) - \tilde{y}(t_{k+1}; t_k, y_k) \\ &\quad + \tilde{y}(t_{k+1}; t_k, y_k) - \tilde{y}(t_{k+1}; t_k, \tilde{y}_k)\|, \end{aligned}$$

so dass wir mit der Dreiecksungleichung, (3.12) und Lemma 3.11 zu

$$\begin{aligned} \|y(t_{k+1}) - \tilde{y}(t_{k+1})\| &\leq \|y(t_{k+1}; t_k, y_k) - \tilde{y}(t_{k+1}; t_k, y_k)\| \\ &\quad + \|\tilde{y}(t_{k+1}; t_k, y_k) - \tilde{y}(t_{k+1}; t_k, \tilde{y}_k)\| \\ &\leq K_\Phi h_{k+1} + e^{L_\Phi h_{k+1}} \|y_k - \tilde{y}_k\| \\ &= K_\Phi h_{k+1} + e^{L_\Phi h_{k+1}} \|y(t_k) - \tilde{y}(t_k)\| \end{aligned}$$

gelangen. Mit der Induktionsvoraussetzung und $1 + s \leq e^s$ folgt daraus

$$\begin{aligned} \|y(t_{k+1}) - \tilde{y}(t_{k+1})\| &\leq K_\Phi h_{k+1} + e^{L_\Phi h_{k+1}} \frac{e^{L_\Phi (t_k - a)} - 1}{L_\Phi} K_\Phi \\ &= \frac{L_\Phi h_{k+1} + e^{L_\Phi h_{k+1}} e^{L_\Phi (t_k - a)} - e^{L_\Phi h_{k+1}}}{L_\Phi} K_\Phi \\ &\leq \frac{L_\Phi h_{k+1} + e^{L_\Phi (h_{k+1} + t_k - a)} - 1 - L_\Phi h_{k+1}}{L_\Phi} K_\Phi \\ &= \frac{e^{L_\Phi (t_{k+1} - a)} - 1}{L_\Phi} K_\Phi. \end{aligned}$$

Damit ist die Induktion vollständig und (3.13) für den Fall $L_\Phi > 0$ bewiesen.

Zum Abschluss widmen wir uns dem Fall $L_\Phi = 0$, den wir als Grenzfall für $L_\Phi \searrow 0$ interpretieren. Mit der Regel von l'Hôspital gilt

$$\lim_{L_\Phi \to 0} \frac{e^{L_\Phi (t_k - a)} - 1}{L_\Phi} = \lim_{L_\Phi \to 0} \frac{(t_k - a) e^{L_\Phi (t_k - a)}}{1} = t_k - a,$$

also folgt unsere Behauptung. ∎

3.3 Konsistenz

Aus Satz 3.13 folgt, dass für die Konvergenz des Näherungsverfahrens das Verhalten des Faktors K_Φ ausschlaggebend ist. Wenn wir in (3.12) die Definition von $\tilde{y}(t_{i+1}; t_i, y_i)$ einsetzen, erhalten wir

$$\begin{aligned} K_\Phi &= \max \left\{ \frac{\|y(t_i; t_{i-1}, y_{i-1}) - \tilde{y}(t_i; t_{i-1}, y_{i-1})\|}{h_i} \;:\; i \in \{1, \ldots, n\} \right\} \\ &= \max \left\{ \frac{\|y(t_i; t_{i-1}, y_{i-1}) - y_{i-1} - h_i \Phi(t_{i-1}, h_i, y_{i-1})\|}{h_i} \;:\; i \in \{1, \ldots, n\} \right\} \\ &= \max \left\{ \left\| \frac{y(t_{i-1} + h_i; t_{i-1}, y_{i-1}) - y_{i-1}}{h_i} - \Phi(t_{i-1}, h_i, y_{i-1}) \right\| \;:\; i \in \{1, \ldots, n\} \right\}. \end{aligned}$$

Für $h_i \to 0$ wird der linke Term gegen $y'(t_{i-1})$ konvergieren, also gegen $f(t_{i-1}, y(t_{i-1})) = f(t_{i-1}, y_{i-1})$. Offenbar kann also das Näherungsverfahren nur dann erfolgreich sein, wenn für $h \to 0$ die Verfahrensfunktion Φ gegen f konvergiert.

Definition 3.14 (Konsistenzfehler) *Sei Φ eine Verfahrensfunktion. Wir definieren den* Konsistenzfehler *zu dem durch Φ gegebenen expliziten Einschrittverfahren durch*

$$\tau(t,h,x) := \begin{cases} \frac{y(t+h;t,x)-x}{h} - \Phi(t,h,x) & \text{falls } h > 0, \\ f(t,x) - \Phi(t,h,x) & \text{ansonsten} \end{cases} \qquad \text{für alle } (t,h) \in \Delta,\ x \in V.$$

Bei dieser Definition ist zu beachten, dass wegen Satz 2.3 und wegen der Lipschitz-Stetigkeit von f die Funktion τ für alle $x \in V$ wohldefiniert ist. Die Behandlung des Sonderfalls $h = 0$ entspricht wegen

$$\lim_{h\to 0} \frac{y(t+h;t,x)-x}{h} = y'(t;t,x) = f(t,x) \qquad \text{für alle } t \in [a,b],\ x \in V$$

gerade der stetigen Fortsetzung.

Wie bereits gesehen gilt

$$K_\Phi = \max\{\|\tau(t_{i-1}, h_i, y_{i-1})\| \ :\ i \in \{1,\dots,n\}\},$$

nach Satz 3.13 ist es also für die Konvergenz der Näherungslösung sehr erstrebenswert, dass $\tau(t,h,y(t))$ für $h \to 0$ gleichmäßig gegen Null geht.

Definition 3.15 (Konsistenz) *Sei Φ eine Verfahrensfunktion. Das durch sie definierte explizite Einschrittverfahren heißt* konsistent mit dem Anfangswertproblem (2.1)*, falls*

$$\lim_{h\to 0} \sup\{\|\tau(t,h,y(t))\| \ :\ t \in [a,b-h]\} = 0 \tag{3.14}$$

gilt. Das Verfahren heißt von der Ordnung p konsistent mit dem Problem *für ein $p \in \mathbb{N}$, falls es eine Konstante $C_{\mathrm{ko}} \in \mathbb{R}_{\geq 0}$ so gibt, dass*

$$\|\tau(t,h,y(t))\| \leq C_{\mathrm{ko}} h^p \qquad \text{für alle } (t,h) \in \Delta \tag{3.15}$$

gilt. Offenbar impliziert diese Bedingung bereits, dass Φ auch konsistent ist.

Da bei der Konsistenzbedingung lediglich Ausgangswerte auf der Lösungskurve y verwendet werden, lässt sich der Konsistenzfehler besonders einfach darstellen: Dank Lemma 3.6 gilt

$$\begin{aligned} \tau(t,h,y(t)) &= \frac{y(t+h;t,y(t)) - y(t)}{h} - \Phi(t,h,y(t)) \\ &= \frac{y(t+h) - y(t)}{h} - \Phi(t,h,y(t)) \qquad \text{für alle } (t,h) \in \Delta. \end{aligned} \tag{3.16}$$

Für den Nachweis der Konsistenz müssen in der Regel sowohl die Eigenschaften des Einschrittverfahrens als auch des zu lösenden Anfangswertproblems berücksichtigt werden. Im Fall des Euler-Verfahrens genügt bereits die Lipschitz-Stetigkeit der Funktion f des Problems (2.1), aus der sich die Lipschitz-Stetigkeit der Ableitung y' gewinnen lässt, die direkt zu einer Abschätzung des Konsistenzfehlers führt.

Lemma 3.16 (Konsistenz Euler) *Sei f Lipschitz-stetig im zweiten Argument mit der Lipschitz-Konstanten L_f, und sei y' Lipschitz-stetig auf $[a,b]$ mit der Konstanten L_y.*

Das explizite Euler-Verfahren ist dann von erster Ordnung konsistent mit der Konsistenzkonstanten $C_{\mathrm{ko}} = L_y$. Das implizite Euler-Verfahren ist ebenfalls von erster Ordnung konsistent mit $C_{\mathrm{ko}} = L_y/(1 - L_f h_0)$.

Beweis. Seien $(t,h) \in \Delta$ gegeben. Mit dem Mittelwertsatz der Differentialrechnung finden wir ein $\eta \in [t, t+h]$ mit

$$\frac{y(t+h) - y(t)}{h} = y'(\eta). \tag{3.17}$$

Nach Definition gilt für das explizite Euler-Verfahren

$$\Phi(t,h,y(t)) = f(t,y(t)) = y'(t),$$

so dass wir mit (3.16) dank der Lipschitz-Stetigkeit von y' zu

$$\|\tau(t,h,y(t))\| = \left\| \frac{y(t+h) - y(t)}{h} - \Phi(t,h,y(t)) \right\| = \|y'(\eta) - y'(t)\| \le L_y|\eta - t| \le L_y h$$

gelangen. Also ist das explizite Euler-Verfahren von erster Ordnung konsistent.

Für die Untersuchung des impliziten Verfahrens setzen wir wieder $h \le h_0 < 1/L_f$ voraus und greifen auf (3.9) zurück, um den Vektor $z = \Phi(t,h,y(t))$ als Lösung der Fixpunktgleichung

$$z = f(t+h, y(t) + hz)$$

darzustellen. Mit (3.17) folgt

$$\begin{aligned}
\|\tau(t,h,y(t))\| &= \|y'(\eta) - f(t+h, y(t)+hz)\| \\
&= \|y'(\eta) - f(t+h, y(t+h)) + f(t+h, y(t+h)) - f(t+h, y(t)+hz)\| \\
&\le \|y'(\eta) - y'(t+h)\| + \|f(t+h, y(t+h)) - f(t+h, y(t)+hz)\| \\
&\le L_y|\eta - (t+h)| + L_f\|y(t+h) - y(t) - hz\| \\
&\le L_y h + L_f h \left\| \frac{y(t+h) - y(t)}{h} - \Phi(t,h,y(t)) \right\| \\
&= L_y h + L_f h \|\tau(t,h,y(t))\|,
\end{aligned}$$

so dass wir

$$\begin{aligned}
(1 - L_f h)\|\tau(t,h,y(t))\| &\le L_y h, \\
\|\tau(t,h,y(t))\| &\le \frac{L_y}{1 - L_f h} h \le \frac{L_y}{1 - L_f h_0} h
\end{aligned}$$

erhalten. Also ist auch das implizite Euler-Verfahren von erster Ordnung konsistent. ∎

Bemerkung 3.17 (Differenzierbarkeit) *Falls die Lösung y zweimal stetig differenzierbar ist, folgt die in Lemma 3.16 geforderte Lipschitz-Stetigkeit mit $L_y = \|y''\|$: Für $t, s \in [a, b]$ gilt aufgrund des Mittelwertsatzes der Differentialrechnung*

$$y'(t) - y'(s) = (t - s)y''(\eta)$$

mit einem $\eta \in [a, b]$, also folgt

$$\|y'(t) - y'(s)\| = |t - s| \, \|y''(\eta)\| \leq |t - s| \, \|y''\|_{\infty,[a,b]},$$

und damit die Behauptung.

Indem wir eine Konsistenzaussage mit dem Konvergenzsatz 3.13 kombinieren, erhalten wir eine Fehlerabschätzung für das Einschrittverfahren:

Satz 3.18 (Konsistenz und Konvergenz) *Sei Φ eine Verfahrensfunktion, sei das korrespondierende Einschrittverfahren stabil und konsistent mit (2.1). Dann gilt*

$$\lim_{\substack{n\to\infty \\ h=(b-a)/n}} \sup\{\|y(t_k; a, y_0) - \tilde{y}(t_k; a, y_0)\| \ : \ t_k = a + kh, \ k \in \{0, \ldots, n\}\} = 0,$$

die diskreten Näherungslösungen konvergieren also gegen die exakte Lösung.

Falls das Verfahren konsistent von Ordnung p ist, gibt es Konstanten $C_{\mathrm{kn}} \in \mathbb{R}_{\geq 0}$ und $n_0 \in \mathbb{N}$ mit

$$\sup\{\|y(t_k; a, y_0) - \tilde{y}(t_k; a, y_0)\| \ : \ t_k = a + kh, \ k \in \{0, \ldots, n\}\} \leq C_{\mathrm{kn}} h^p$$
$$\text{für alle } n \in \mathbb{N}, \ n \geq n_0 \text{ mit } h = (b - a)/n.$$

Beweis. Sei zunächst das Verfahren konsistent. Sei $\epsilon \in \mathbb{R}_{>0}$, und sei

$$\hat{\epsilon} := \frac{\epsilon}{(b - a)e^{L_\Phi (b-a)}}.$$

Nach Voraussetzung existiert ein $h_1 \in [0, h_0]$ mit

$$\sup\{\|\tau(t, h, y(t))\| \ : \ t \in [a, b - h]\} \leq \hat{\epsilon} \qquad \text{für alle } h \in [0, h_1].$$

Wir wählen ein $n_0 \in \mathbb{N}$ mit $(b - a)/n_0 \leq h_1$. Sei $n \in \mathbb{N}$ mit $n \geq n_0$ gegeben. Es gilt

$$h := \frac{b - a}{n} \leq \frac{b - a}{n_0} \leq h_1,$$

also erhalten wir

$$\begin{aligned} K_\Phi &:= \max\left\{ \left\| \frac{y(t_i; t_{i-1}, y_{i-1}) - y_{i-1}}{h_i} - \Phi(t_{i-1}, h_i, y_{i-1}) \right\| \ : \ i \in \{1, \ldots, n\} \right\} \\ &= \max\{\|\tau(t_{i-1}, h_i, y_{i-1})\| \ : \ i \in \{1, \ldots, n\}\} \leq \hat{\epsilon}. \end{aligned}$$

Indem wir diese Konstante in Satz 3.13 einsetzen, folgt

$$\|y(t_k) - \tilde{y}(t_k)\| \leq K_\Phi (t_k - a) e^{L_\Phi (t_k - a)} \leq \hat{\epsilon}(b-a) e^{L_\Phi (b-a)} = \epsilon \quad \text{für alle } k \in \{0, \dots, n\}.$$

Da wir diese Abschätzung für alle $n \in \mathbb{N}_{\geq n_0}$ und beliebiges $\epsilon \in \mathbb{R}_{>0}$ bewiesen haben, erhalten wir die gewünschte Konvergenzaussage.

Sei nun das Verfahren konsistent von Ordnung p, und sei $C_{\text{ko}} \in \mathbb{R}_{>0}$ die Konstante aus (3.15). Dann gilt wegen $h \leq h_0$ $(t,h) \in \Delta$ für alle $t \in [a, b-h]$, also dürfen wir (3.15) anwenden, um

$$K_\Phi \leq C_{\text{ko}} h^p$$

zu erhalten. Einsetzen in Satz 3.13 ergibt

$$\begin{aligned} \|y(t_k) - \tilde{y}(t_k)\| &\leq K_\Phi (t_k - a) e^{L_\Phi (t_k - a)} \leq C_{\text{ko}} h^p (t_k - a) e^{L_\Phi (t_k - a)} \\ &\leq C_{\text{ko}} h^p (b-a) e^{L_\Phi (b-a)} \qquad \text{für alle } k \in \{0, \dots, n\}. \end{aligned}$$

Mit der Konstanten

$$C_{\text{kn}} := C_{\text{ko}} (b-a) e^{L_\Phi (b-a)}$$

folgt daraus die gewünschte Aussage. ∎

Folgerung 3.19 (Konvergenz Euler) *Sei y' Lipschitz-stetig. Dann gibt es eine Konstante $C_{\text{eu}} \in \mathbb{R}_{>0}$ so, dass für alle $n \in \mathbb{N}$ die per explizitem Euler-Verfahren mit Schrittweite $h = (b-a)/n$ berechnete Näherungslösung die Abschätzung*

$$\|y(t_k) - \tilde{y}(t_k)\| \leq C_{\text{eu}} h \qquad \text{für alle } t_k = a + hk,\ k \in \{0, \dots, n\}$$

erfüllt. Insbesondere konvergiert die Näherung für $n \to \infty$ gegen die exakte Lösung.

Falls die Schrittweite klein genug ist, gilt dieselbe Aussage auch für das implizite Euler-Verfahren.

Beweis. Wir kombinieren Satz 3.18 mit Lemma 3.16. ∎

3.4 Lokalisierte Konvergenzaussagen

Satz 3.13 erfordert die Lipschitz-Stetigkeit der Inkrement-Funktion Φ auf dem gesamten Raum V und bietet eine Fehlerabschätzung ohne weitere Einschränkungen an K_Φ.

In der Praxis passiert es häufig, dass die Funktion f, und damit in der Regel auch die von ihr abhängende Verfahrensfunktion Φ, nur in einer Umgebung der exakten Lösung Lipschitz-stetig sind. In dieser Situation kann es sinnvoll sein, Φ Lipschitz-stetig auf den gesamten Raum V fortzusetzen und dann die bereits bewiesenen Aussagen auf die modifizierte Verfahrensfunktion anzuwenden.

Im Interesse der Einfachheit beschränken wir uns in diesem Abschnitt auf den Fall, dass V ein Hilbert-Raum ist. Die Grundlage unseres Fortsetzungsarguments ist die Projektion beliebiger Vektoren auf die Einheitskugel:

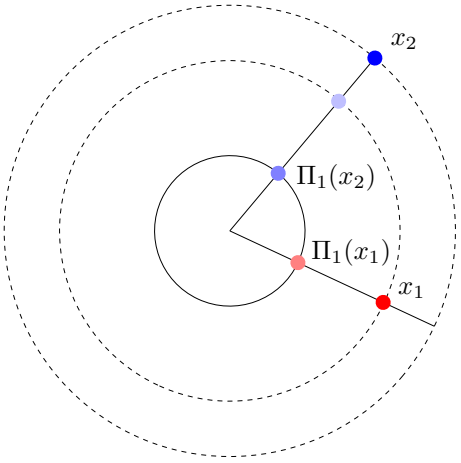

Abbildung 3.1: Beweisskizze für Lemma 3.20: In einem ersten Schritt wird die Länge des Vektors x_2 der des Vektors x_1 angeglichen, dann werden beide auf den Einheitskreis projiziert.

Lemma 3.20 (Projektion) *Wir definieren die Projektion*

$$\Pi_1(x) := \begin{cases} x & \text{falls } \|x\| < 1, \\ \frac{x}{\|x\|} & \text{ansonsten} \end{cases} \qquad \text{für alle } x \in V.$$

Dann gilt

$$\|\Pi_1(x_1) - \Pi_1(x_2)\| \leq \|x_1 - x_2\| \qquad \text{für alle } x_1, x_2 \in V.$$

Beweis. Seien $x_1, x_2 \in V$, und sei ohne Beschränkung der Allgemeinheit $\|x_1\| \leq \|x_2\|$ angenommen. Wir untersuchen zuerst den Fall, dass lediglich einer der beiden Vektoren skaliert wird, genauer gesagt wollen wir

$$\|x_1 - \alpha x_2\| \leq \|x_1 - x_2\| \qquad \text{für alle } \alpha \in [\|x_1\|/\|x_2\|, 1] \tag{3.18}$$

beweisen. Mit Hilfe dieser Abschätzung könnten wir beispielsweise beide Vektoren auf dieselbe Länge bringen, der Rest des Beweises wäre dann einfach.

Wir beweisen (3.18), indem wir zunächst beide Seiten der Abschätzung quadrieren und ausnutzen, dass die Norm durch das Skalarprodukt gegeben ist:

$$\begin{aligned} \|x_1 - \alpha x_2\|^2 &= \|x_1\|^2 - 2\alpha\langle x_1, x_2\rangle + \alpha^2\|x_2\|^2, \\ \|x_1 - x_2\|^2 &= \|x_1\|^2 - 2\langle x_1, x_2\rangle + \|x_2\|^2, \end{aligned}$$

also müssen wir lediglich

$$2(1-\alpha)\langle x_1, x_2\rangle \overset{!}{\leq} (1-\alpha^2)\|x_2\|^2 = (1-\alpha)(1+\alpha)\|x_2\|^2$$

beweisen. Da $\alpha \leq 1$ gilt, können wir durch $1 - \alpha$ dividieren und erhalten

$$2\langle x_1, x_2\rangle \overset{!}{\leq} (1+\alpha)\|x_2\|^2.$$

Diese Ungleichung folgt aus der Cauchy-Schwarz-Ungleichung:

$$\begin{aligned} 2\langle x_1, x_2\rangle &\leq 2\|x_1\|\,\|x_2\| = (\|x_1\| + \|x_1\|)\|x_2\| \leq (\|x_2\| + \|x_1\|)\|x_2\| \\ &= (1 + \|x_1\|/\|x_2\|)\|x_2\|^2 \leq (1+\alpha)\|x_2\|^2. \end{aligned}$$

Damit ist (3.18) bewiesen und wir können uns der eigentlich zu zeigenden Aussage zuwenden.

1. Fall: $1 \leq \|x_1\| \leq \|x_2\|$. Wir wenden (3.18) auf $\alpha = \|x_1\|/\|x_2\|$ an und erhalten

$$\|\Pi_1(x_1) - \Pi_1(x_2)\| = \left\| \frac{x_1}{\|x_1\|} - \frac{x_2}{\|x_2\|} \right\| = \frac{\|x_1 - \alpha x_2\|}{\|x_1\|} \leq \frac{\|x_1 - x_2\|}{\|x_1\|} \leq \|x_1 - x_2\|.$$

2. Fall: $\|x_1\| < 1 \leq \|x_2\|$. Diesmal wenden wir (3.18) auf $\alpha = 1/\|x_2\|$ an und finden

$$\|\Pi_1(x_1) - \Pi_1(x_2)\| = \left\| x_1 - \frac{x_2}{\|x_2\|} \right\| = \|x_1 - \alpha x_2\| \leq \|x_1 - x_2\|.$$

3. Fall: $\|x_1\| \leq \|x_2\| < 1$. Trivial wegen $\Pi_1(x_1) = x_1$ und $\Pi_2(x_2) = x_2$. ∎

Wir sind nicht an Projektionen auf die Einheitskugel interessiert, sondern auf potentiell kleinere Kugeln mit Radius $\gamma \in \mathbb{R}_{>0}$, die sich einfach per

$$\Pi_\gamma(x) := \gamma\Pi_1(x/\gamma) \qquad \text{für alle } x \in V$$

definieren lassen. Offenbar gilt auch hier

$$\begin{aligned} \|\Pi_\gamma(x) - \Pi_\gamma(z)\| &= \gamma\|\Pi_1(x/\gamma) - \Pi_1(z/\gamma)\| \\ &\leq \gamma\|x/\gamma - z/\gamma\| = \|x - z\| \qquad \text{für alle } x, z \in V, \end{aligned}$$

und wir können die Fortsetzung von Φ konstruieren:

Folgerung 3.21 (Lokalisierung) *Sei $y_0 \in V$, und sei $y : [a,b] \to V$ eine Lösung des Anfangswertproblems (2.1).*

Sei $\gamma \in \mathbb{R}_{>0}$. Wir definieren die Umgebungen

$$S(t) := \{x \in V \;:\; \|x - y(t)\| \leq \gamma\} \qquad \textit{für alle } t \in [a,b]$$

(vgl. Abbildung 3.2) und setzen voraus, dass die Verfahrensfunktion auf ihnen im zweiten Argument Lipschitz-stetig ist, dass also

$$\|\Phi(t,h,x_1) - \Phi(t,h,x_2)\| \leq L_\Phi\|x_1 - x_2\| \qquad \textit{für alle } (t,h) \in \Delta,\ x_1, x_2 \in S(t)$$

für ein $L_\Phi \in \mathbb{R}_{\geq 0}$ gilt. Wir gehen davon aus, dass für die in (3.12) definierte Konstante die Schranke

$$K_\Phi(b-a)e^{L_\Phi(b-a)} \leq \gamma \tag{3.19}$$

gilt. Dann folgt

$$\|y(t_k) - \tilde{y}(t_k)\| \leq K_\Phi(b-a)e^{L_\Phi(b-a)} \qquad \textit{für alle } k \in \{0,\ldots,n\},$$

wir erhalten also dieselbe Fehlerabschätzung wie in Satz 3.13.

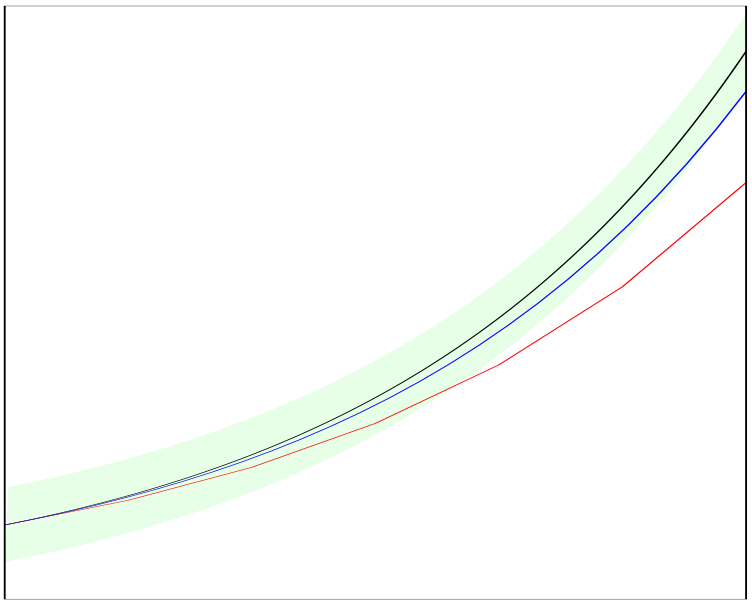

Abbildung 3.2: Ansatz zur Lokalisierung: Falls die Verfahrensfunktion auf einer Umgebung (grün) der Lösung (schwarz) Lipschitz-stetig ist, wird bei hinreichend kleiner Schrittweite auch die Näherung (blau) in diesem Bereich liegen.

Beweis. Da die Verfahrensfunktion Φ nur lokal Lipschitz-stetig ist, setzen wir sie zu einer global Lipschitz-stetigen Funktion Φ_* fort: Falls ein Vektor nicht in der durch $S(t)$ definierten Umgebung der Lösung liegt, wird er mit Hilfe der Abbildung

$$\pi_t : V \to V, \qquad x \mapsto y(t) + \Pi_\gamma(x - y(t)),$$

in diese Menge projiziert. Nach Definition von Π_γ folgt

$$\|\pi_t(x) - y(t)\| \leq \gamma,$$

also $\pi_t(x) \in S(t)$. Die Fortsetzung Φ_* der Verfahrensfunktion definieren wir durch

$$\Phi_*(t, h, x) := \Phi(t, h, \pi_t(x)) \qquad \text{für alle } (t, h) \in \Delta,\ x \in V,$$

und dank Lemma 3.20 erhalten wir

$$\begin{aligned} \|\Phi_*(t, h, x) - \Phi_*(t, h, z)\| &= \|\Phi(t, h, \pi_t(x)) - \Phi(t, h, \pi_t(z))\| \\ &\leq L_\Phi \|\pi_t(x) - \pi_t(z)\| \leq L_\Phi \|x - z\| \qquad \text{für alle } (t, h) \in \Delta,\ x, z \in V, \end{aligned}$$

die Verfahrensfunktion Φ_* ist also *global* Lipschitz-stetig im letzten Argument, also stabil.

Analog zu $\tilde{y}$ definieren wir Näherungslösungen $\tilde{y}_*$ für die fortgesetzte Verfahrensfunktion Φ_* durch

$$\tilde{y}_*(t_0) := y_0,$$

$$\tilde{y}_*(t_i) := \tilde{y}_*(t_{i-1}) + h_i \Phi_*(t_{i-1}, h_i, \tilde{y}_*(t_{i-1})) \qquad \text{für alle } i \in \{1, \dots, n\}$$

und bezeichnen die (3.12) entsprechende Konstante mit K_{Φ_*}. Wegen $y(t) \in S(t)$ folgt $\Phi_*(t, h, y(t)) = \Phi(t, h, y(t))$, also auch $K_\Phi = K_{\Phi_*}$, und aus Satz 3.13 und (3.19) erhalten wir die Abschätzung

$$\|y(t_k) - \tilde{y}_*(t_k)\| \leq K_\Phi (t_k - a) e^{L_\Phi (t_k - a)} \leq \gamma \qquad \text{für alle } k \in \{0, \dots, n\}. \tag{3.20}$$

Daraus folgt insbesondere

$$\tilde{y}_*(t_k) \in S(t_k), \qquad \Phi_*(t_k, h_k, \tilde{y}_*(t_k)) = \Phi(t_k, h_k, \tilde{y}_*(t_k)) \qquad \text{für alle } k \in \{0, \dots, n\},$$

und mit einer einfachen Induktion erhalten wir somit

$$\tilde{y}_*(t_k) = \tilde{y}(t_k) \qquad \text{für alle } k \in \{0, \dots, n\},$$

also überträgt sich die Fehlerabschätzung (3.20) auf $\tilde{y}$ und die Aussage ist bewiesen. ∎

Die Näherungslösung wird also auch dann gegen die Lösung konvergieren, wenn die Verfahrensfunktion nicht global Lipschitz-stetig ist. Für eine hinreichend kleine Schrittweite dürfen wir sogar dieselbe Fehlerabschätzung wie zuvor erwarten.

3.5 Konsistenzkriterium

Wie wir in Satz 3.18 gesehen haben, entscheidet die Konsistenzordnung p darüber, wie schnell sich der Fehler der Näherungslösung reduziert. Im Falle des Euler-Verfahrens bewirkt eine Halbierung der Schrittweite lediglich eine Halbierung des Fehlers, während bei einem Verfahren p-ter Ordnung der Fehler bereits um den Faktor 2^{-p} reduziert werden würde.

Wenn wir eine gewisse Genauigkeit $\epsilon \in \mathbb{R}_{>0}$ erreichen wollen, muss

$$\epsilon \sim h^p \sim n^{-p}$$

gelten, wir benötigen also

$$n \sim \epsilon^{-1/p}$$

Schritte des Einschrittverfahrens. Grob abgeschätzt bedeutet dass, das tausend Schritte eines Verfahrens zweiter Ordnung ungefähr denselben Fehler erzielen wie eine Million Schritte eines Verfahrens erster Ordnung, Verfahren hoher Ordnung sind deshalb in der Regel *wesentlich* effizienter als solche niedriger Ordnung.

Wir sind also daran interessiert, Verfahren möglichst hoher Ordnung zu konstruieren, die trotzdem möglichst effizient durchführbar sein sollten. Dazu müssen wir den Konsistenzfehler τ analysieren. Wir wählen $(t, h) \in \Delta$ mit $h > 0$ und setzen $y \in C^{p+1}([a, b], V)$ sowie $\Phi(t, \cdot, y(t)) \in C^p([0, h], V)$ voraus und erhalten mit (3.16) als Darstellung des Konsistenzfehlers unter Anwendung des Satzes von Taylor die Gleichung

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

$$
\begin{aligned}
&= \frac{1}{h}\left(\sum_{\nu=0}^{p}\frac{h^\nu}{\nu!}y^{(\nu)}(t) + \frac{h^{p+1}}{(p+1)!}y^{(p+1)}(\eta_t) - y(t)\right)\\
&\quad - \left(\sum_{\nu=0}^{p-1}\frac{h^\nu}{\nu!}\frac{\partial^\nu\Phi}{\partial h^\nu}(t,0,y(t)) + \frac{h^p}{p!}\frac{\partial^p\Phi}{\partial h^p}(t,\eta_h,y(t))\right)\\
&= \sum_{\nu=1}^{p}\frac{h^{\nu-1}}{\nu!}y^{(\nu)}(t) + \frac{h^p}{(p+1)!}y^{(p+1)}(\eta_t)\\
&\quad - \left(\sum_{\nu=0}^{p-1}\frac{h^\nu}{(\nu+1)!}(\nu+1)\frac{\partial^\nu\Phi}{\partial h^\nu}(t,0,y(t)) + \frac{h^p}{(p+1)!}(p+1)\frac{\partial^p\Phi}{\partial h^p}(t,\eta_h,y(t))\right)\\
&= \sum_{\nu=0}^{p-1}\frac{h^\nu}{(\nu+1)!}\left(y^{(\nu+1)}(t) - (\nu+1)\frac{\partial^\nu\Phi}{\partial h^\nu}(t,0,y(t))\right)\\
&\quad + \frac{h^p}{(p+1)!}\left(y^{(p+1)}(\eta_h) - (p+1)\frac{\partial^p\Phi}{\partial h^p}(t,\eta_h,y(t))\right) \qquad (3.21)
\end{aligned}
$$

für Zwischenstellen $\eta_t \in [t, t+h]$ und $\eta_h \in [0,h]$. Eine Konsistenzordnung von p können wir nur erwarten, falls die erste Summe verschwindet, falls also die Ableitungen $y^{(\nu+1)}$ und $(\nu+1)\partial^\nu\Phi/\partial h^\nu$ für $\nu \in \{0,\dots,p-1\}$ übereinstimmen.

Lemma 3.22 (Konsistenzkriterium) *Sei $p \in \mathbb{N}$, und sei $y \in C^{p+1}[a,b]$. Sei $h_0 \in \mathbb{R}_{>0}$ so gewählt, dass die Abbildung $(t,h) \mapsto \Phi(t,h,y(t))$ für $(t,h) \in \Delta$ p-mal partiell nach h differenzierbar mit stetiger p-ter partieller Ableitung ist. Gelte*

$$(\nu+1)\frac{\partial^\nu\Phi}{\partial h^\nu}(t,0,y(t)) = y^{(\nu+1)}(t) \qquad \text{für alle } t \in [a,b],\ \nu \in \{0,\dots,p-1\}.$$

Dann ist das durch Φ definierte Verfahren von p-ter Ordnung konsistent.

Beweis. (vgl. [4, Abschnitt 11.5.1]) Da $y^{(p+1)}$ stetig ist, ist

$$C_1 := \max\{\|y^{(p+1)}(t)\| \ :\ t \in [a,b]\} \in \mathbb{R}_{\geq 0}$$

als Maximum einer stetigen Funktion auf dem kompakten Intervall $[a,b]$ wohldefiniert. Da Φ im zweiten Argument p-mal stetig differenzierbar ist, ist

$$C_2 := \max\left\{\left\|\frac{\partial^p\Phi}{\partial h^p}(t,h,y(t))\right\| \ :\ (t,h) \in \Delta\right\} \in \mathbb{R}_{\geq 0}$$

als Maximum einer stetigen Funktion auf der (nach Heine-Borel) kompakten Menge Δ (vgl. (3.6)) ebenfalls wohldefiniert.

Sei $(t,h) \in \Delta$. Durch Einsetzen in (3.21) folgt sofort

$$\tau(t,h,y(t)) = \frac{h^p}{(p+1)!}\left(y^{(p+1)}(\eta_t) - (p+1)\frac{\partial^p\Phi}{\partial h^p}(t,\eta_h,y(t))\right)$$

mit Zwischenpunkten $\eta_t \in [t, t+h]$ und $\eta_h \in [0, h]$. Nach Definition von C_1 und C_2 erhalten wir also

$$\begin{aligned}\|\tau(t,h,y(t))\| &\leq \frac{h^p}{(p+1)!}\left(\|y^{(p+1)}(\eta_t)\| + (p+1)\left\|\frac{\partial^p \Phi}{\partial h^p}(t,\eta_h,y(t))\right\|\right)\\ &\leq \frac{h^p}{(p+1)!}(C_1 + (p+1)C_2) = C_{\mathrm{ko}} h^p\end{aligned}$$

für die Konstante

$$C_{\mathrm{ko}} := \frac{C_1}{(p+1)!} + \frac{C_2}{p!},$$

und damit ist (3.15) bewiesen. ∎

Dieses Resultat lässt sich als Verallgemeinerung von Lemma 3.16 interpretieren: Für das explizite Euler-Verfahren verwenden wir $\Phi(t,h,x) = f(t,x)$, also gilt immerhin

$$\Phi(t,0,y(t)) = f(t,y(t)) = y'(t),$$

so dass die Voraussetzungen des Lemmas 3.22 für $p = 1$ erfüllt sind. Für $p = 2$ gelten sie allerdings wegen

$$\frac{\partial \Phi}{\partial h}(t,0,y(t)) = 0$$

nur, falls $y'' = 0$ gilt, falls also y ein lineares Polynom ist.

Beispiel 3.23 (Konsistenz zweiter Ordnung) *Lemma 3.22 kann verwendet werden, um Einschrittverfahren beliebig hoher Konsistenzordnung zu entwickeln. Dazu schreiben wir die rechte Seite des Anfangswertproblems (2.1) in der Form*

$$f_y : [a,b] \to V, \qquad t \mapsto f(t,y(t)),$$

und erhalten durch Differenzieren der Gleichung $y'(t) = f_y(t)$ die Beziehung

$$\begin{aligned}y''(t) = f_y'(t) &= Df(t,y(t)) \cdot (1, y'(t)) = Df(t,y(t)) \cdot (1, f(t,y(t)))\\ &= \frac{\partial f}{\partial t}(t,y(t)) + \frac{\partial f}{\partial x}(t,y(t)) f(t,y(t)) \qquad \text{für alle } t \in [a,b],\end{aligned}$$

wir können also diese Größe berechnen, falls uns die Ableitungen von f zur Verfügung stehen. Um das Konsistenzkriterium zu erfüllen, müssen wir

$$\Phi(t,0,y(t)) = y'(t), \qquad 2\frac{\partial \Phi}{\partial h}(t,0,y(t)) = y''(t) \qquad \text{für alle } t \in [a,b]$$

sicherstellen, und dieses Ziel lässt sich nun leicht erreichen, da uns $y'(t) = f(t,y(t))$ und $y''(t)$ zur Verfügung stehen: Wir setzen

$$\Phi(t,h,x) := f(t,x) + \frac{h}{2}\left(\frac{\partial f}{\partial t}(t,x) + \frac{\partial f}{\partial x}(t,x) f(t,x)\right) \qquad \text{für alle } (t,h) \in \Delta,\ x \in V,$$

und Lemma 3.22 zeigt, dass das durch diese Verfahrensfunktion definierte Verfahren die Konsistenzordnung 2 *besitzt.*

Falls uns höhere Ableitungen der rechten Seite f zur Verfügung stehen, können wir in derselben Weise Verfahrensfunktionen konstruieren, die noch höhere Konsistenzordnungen erreichen.

In der Praxis stehen uns sehr oft die Ableitungen von f nicht zur Verfügung, so dass wir uns für Verfahren interessieren, die eine höhere Konsistenzordnung auch ohne diese zusätzliche Information erzielen (schließlich können wir eine Funktion statt per Taylor-Entwicklung auch durch Lagrange-Interpolation approximieren).

Beispiel 3.24 (Heun) *Es ist möglich, aus einem Quadraturverfahren höherer Ordnung eine Verfahrensfunktion höherer Konsistenzordnung zu konstruieren. Als Beispiel verwenden wir die Trapezregel*

$$\int_t^{t+h} g(s)\,ds \approx \frac{h}{2}(g(t) + g(t+h)),$$

die bei einem zweimal differenzierbaren Integranden einen Fehler der Ordnung h^3 aufweist.

Wir wenden diese Regel auf die Integraldarstellung aus Lemma 2.1 an und erhalten

$$y(t+h) = y(t) + \int_t^{t+h} f(s, y(s))\,ds \approx y(t) + \frac{h}{2}(f(t, y(t)) + f(t+h, y(t+h))),$$

also ein zunächst implizites Einschrittverfahren.

Falls h klein genug ist, dürfen wir erwarten, dass wir mit der Fixpunktiteration

$$y^{(i+1)}(t+h) := y(t) + \frac{h}{2}(f(t, y(t)) + f(t+h, y^{(i)}(t+h))) \qquad \text{für alle } i \in \mathbb{N}_0$$

schnell eine gute Näherung von $y(t+h)$ erhalten können.

Wenn wir zur Bestimmung des Startwerts das explizite Euler-Verfahren verwenden, erhalten wir $y^{(0)}(t+h) := y(t) + hf(t, y(t))$ und nach dem ersten Iterationsschritt

$$\begin{aligned} y^{(1)}(t+h) &:= y(t) + \frac{h}{2}(f(t, y(t)) + f(t+h, y^{(0)}(t+h)) \\ &= y(t) + \frac{h}{2}(f(t, y(t)) + f(t+h, y(t) + hf(t, y(t)))). \end{aligned}$$

Für eine hinreichend kleine Schrittweite h können wir davon ausgehen, dass dieser Wert bereits eine gute Näherung von $y(t+h)$ darstellt. Die entsprechende Verfahrensfunktion lautet

$$\Phi(t, h, x) := \frac{1}{2}\left(f(t, x) + f(t+h, x + hf(t, x))\right)$$

und definiert das Heun-Verfahren.

Falls f einmal stetig differenzierbar ist, folgt aus $\Phi(t,0,x) = f(t,x)$ nach Lemma 3.22 bereits, dass das Verfahren von erster Ordnung konsistent ist. Falls f zweimal stetig differenzierbar ist, erhalten wir per Kettenregel

$$\begin{aligned} 2\frac{\partial \Phi}{\partial h}(t,h,x) &= Df(t+h, x+hf(t,x)) \cdot (1, f(t,x)), \\ y''(t) &= f_y'(t) = Df(t,y(t)) \cdot (1, y'(t)) \\ &= Df(t,y(t)) \cdot (1, f(t,y(t))), \\ 2\frac{\partial \Phi}{\partial h}(t,0,y(t)) &= y''(t), \end{aligned}$$

also ist das Verfahren gemäß Lemma 3.22 in diesem Fall von zweiter Ordnung konsistent.

Statt der Trapezregel können wir auch mit der Mittelpunktregel arbeiten, die ebenfalls einen Fehler in der Größenordnung von h^3 erwarten lässt. Auch hier müssen wir den Wert im Mittelpunkt des Intervalls geeignet approximieren, beispielsweise durch das Euler-Verfahren:

Beispiel 3.25 (Runge) *Analog können wir auch die Mittelpunktregel verwenden, um eine Verfahrensfunktion zu konstruieren: Wir gehen wieder von Lemma 2.1 aus und setzen*

$$y(t+h) = y(t) + \int_t^{t+h} f(s, y(s))\, ds \approx y(t) + hf\left(t + \frac{h}{2}, y\left(t + \frac{h}{2}\right)\right).$$

Wie schon in Beispiel 3.24 verwenden wir das explizite Euler-Verfahren, um die Approximation

$$y\left(t + \frac{h}{2}\right) \approx y(t) + \frac{h}{2} f(t, y(t))$$

zu gewinnen und erhalten

$$y(t+h) \approx y(t) + hf\left(t + \frac{h}{2}, y(t) + \frac{h}{2} f(t, y(t))\right).$$

Die zugehörige Verfahrensfunktion

$$\Phi(t,h,x) := f\left(t + \frac{h}{2}, x + \frac{h}{2} f(t,x)\right)$$

definiert das Runge- *oder auch* Euler-Collatz-Verfahren.

Falls f einmal stetig differenzierbar ist, folgt aus $\Phi(t,0,x) = f(t,x)$ per Lemma 3.22, dass das Verfahren von erster Ordnung konsistent ist. Falls f zweimal stetig differenzierbar ist, impliziert die Kettenregel

$$2\frac{\partial \Phi}{\partial h}(t,h,x) = 2Df\left(t + \frac{h}{2}, x + \frac{h}{2} f(t,x)\right) \cdot \left(\frac{1}{2}, \frac{1}{2} f(t,x)\right)$$

$$= Df\left(t + \frac{h}{2}, x + \frac{h}{2} f(t,x)\right) \cdot (1, f(t,x)),$$

$$2\frac{\partial \Phi}{\partial h}(t,0,y(t)) = Df(t,y(t)) \cdot (1, f(t,y(t))) = f'_y(t),$$

und dank Lemma 3.22 können wir schließen, dass das Verfahren auch von zweiter Ordnung konsistent ist.

Es stellt sich die Frage, ob man durch Quadraturformeln höherer Ordnung auch zu Verfahrensfunktionen höherer Ordnung gelangen kann. Im Prinzip können wir eine Quadraturformel

$$\int_t^{t+h} f(s, y(s))\, ds \approx \sum_{i=1}^{m} \omega_i f(s_i, y(s_i))$$

verwenden, müssen dann aber brauchbare Näherungswerte für $y(s_i)$ in allen Quadraturpunkten zur Verfügung stellen. Eine Analyse der Fehlerfortpflanzung im Quadraturverfahren zeigt, dass eine Quadraturordnung von $p+1$, also eine Konsistenzordnung von p, nur dann zu erwarten ist, wenn die Näherungswerte für $y(s_i)$ genau bis auf einen Fehler der Ordnung $p-1$ sind. Diese Eigenschaft ist im Allgemeinen nur schwer sicherzustellen.

Im Falle des Heun- und Euler-Collatz-Verfahrens profitieren wir davon, dass das explizite Euler-Verfahren eine Näherung erster Ordnung für $y(t+h)$ bzw. $y(t+h/2)$ zur Verfügung stellt, für höhere Ordnungen ist dieses Ziel schwieriger zu erreichen.

3.6 Runge-Kutta-Verfahren

Sowohl das Heun- als auch das Euler-Collatz-Verfahren basieren darauf, die Funktion f in Punkten auszuwerten, die von vorangehenden Auswertungen der Funktion abhängen können. Allgemein haben wir also die Struktur

$$k_i = f\left(t + c_i h, x + h \sum_{j=1}^{i-1} a_{ij} k_j\right) \qquad \text{für } i \in \{1, \dots, s\}, \tag{3.22a}$$

$$\Phi(t,h,x) = \sum_{i=1}^{s} b_i k_i \tag{3.22b}$$

eines expliziten Verfahrens, das mit Hilfe von s Auswertungen der Funktion f eine Verfahrensfunktion definiert. Die entscheidenden Parameter sind die Vektoren $\mathbf{c} \in \mathbb{R}^s$, die die Zeitpunkte für die Auswertungen angeben, die Matrix $\mathbf{A} = (a_{ij})_{i,j=1}^s$, die angibt, wie die i-te Funktionsauswertung von den vorangehenden Auswertungen beeinflusst wird, und der Vektor $\mathbf{b} \in \mathbb{R}^s$, der beschreibt, wie die einzelnen Funktionsauswertungen kombiniert werden müssen, um die Verfahrensfunktion zu erhalten.

Die durch (3.22) beschriebenen Verfahren bezeichnen wir als *Runge-Kutta-Verfahren* der Stufe s. Die bisher betrachteten expliziten Einschrittverfahren lassen sich als Runge-Kutta-Verfahren interpretieren: Das explizite Euler-Verfahren ist einstufig mit den Parametern

$$\mathbf{A} = \begin{pmatrix} 0 \end{pmatrix}, \qquad \mathbf{c} = \begin{pmatrix} 0 \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} 1 \end{pmatrix},$$

das Heun-Verfahren ist zweistufig mit

$$\mathbf{A} = \begin{pmatrix} 0 & \\ 1 & 0 \end{pmatrix}, \qquad \mathbf{c} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix},$$

und das Runge-Verfahren ist ebenfalls zweistufig mit

$$\mathbf{A} = \begin{pmatrix} 0 & \\ 1/2 & 0 \end{pmatrix}, \qquad \mathbf{c} = \begin{pmatrix} 0 \\ 1/2 \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Kompakt lassen sich Runge-Kutta-Verfahren in Form des *Butcher-Tableaus*

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

schreiben, die drei oben erwähnten Verfahren nehmen dann die Form

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

an. Anschaulich entsprechen bei diesem Schema die ersten s Zeilen jeweils einer Auswertung von f: die $\mathbf{c}$-Spalte gibt den Zeitpunkt an, die restlichen Spalten beschreiben den Ort. Die unterste Zeile des Butcher-Schemas beschreibt, wie die einzelnen Zwischengrößen k_i zusammengesetzt werden müssen, um das Endergebnis zu berechnen.

Beispiel 3.26 (Quadratur) *Ein Spezialfall eines Anfangswertproblems ist die Berechnung eines Integrals: Für ein $g \in C([0,1], V)$ ist die Lösung der Gleichungen*

$$y(0) = 0, \qquad y'(t) = g(t) \qquad \text{für alle } t \in [0,1]$$

nach Lemma 2.1 auch eine Lösung der Gleichung

$$y(1) = \int_0^1 g(s)\,ds,$$

also kann jedes Lösungsverfahren für eine gewöhnliche Differentialgleichung auch verwendet werden, um Integrale zu approximieren.

Wenn wir einen Schritt eines Runge-Kutta-Verfahrens s-ter Stufe durchführen, um $y(b)$ zu berechnen, erhalten wir

$$k_i = g(c_i) \qquad \text{für alle } i \in \{1, \dots, s\},$$
$$y(1) \approx y(0) + \Phi(t, 1, y(0)) = \sum_{i=1}^{s} b_i k_i = \sum_{i=1}^{s} b_i g(c_i),$$

also eine Quadraturformel mit den Quadraturpunkten c_i und den zugehörigen Quadraturgewichten b_i.

Durch Taylor-Entwicklung lassen sich aus den in Lemma 3.22 gegebenen Bedingungen nichtlineare Gleichungssysteme herleiten, die zur Konstruktion von Runge-Kutta-Verfahren höherer Ordnung verwendet werden können. Unserem Beispiel 3.26 können wir entnehmen, dass Quadraturformeln einen guten Lösungsansatz bieten: Für den Vektor $\mathbf{c}$ lassen sich Quadraturpunkte auf dem Intervall $[0,1]$ verwenden, für den Vektor $\mathbf{b}$ die entsprechenden Quadraturgewichte.

Beispiel 3.27 (Klassisches Runge-Kutta-Verfahren) *Wir gehen von der Simpson-Quadraturformel*

$$\int_0^1 g(s)\,ds \approx \frac{1}{6}(g(0) + 4g(1/2) + g(1))$$

aus, die wir aus Symmetriegründen in der Form

$$\int_0^1 g(s)\,ds \approx \frac{1}{6}g(0) + \frac{1}{3}g(1/2) + \frac{1}{3}g(1/2) + \frac{1}{6}g(1)$$

schreiben. Wenn man die passenden Koeffizienten a_{ij} berechnet, erhält man das Schema

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

das das klassischen Runge-Kutta-Verfahren *vierter Stufe beschreibt. Es lässt sich nachweisen, dass dieses Verfahren die Konsistenzordnung vier besitzt.*

Entsprechend kann man auch die 3/8-Quadraturformel von Newton verwenden, die durch

$$\int_0^1 g(s)\,ds \approx \frac{1}{8}g(0) + \frac{3}{8}g(1/3) + \frac{3}{8}g(2/3) + \frac{1}{8}g(1)$$

gegeben ist und zu dem Butcher-Schema

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

führt. Man kann zeigen, dass auch das zu diesem Schema gehörende Runge-Kutta-Verfahren die Konsistenzordnung vier besitzt.

Im Interesse einer hohen Genauigkeit sind wir natürlich an Verfahren möglichst hoher Konsistenzordnung interessiert. Die Konstruktion derartiger Verfahren ist im Allgemeinen relativ schwierig, aber es ist immerhin möglich, eine obere Schranke für die maximal erreichbare Ordnung anzugeben, indem man ein einfaches Beispielproblem analysiert.

Lemma 3.28 (Exponentialfunktion) *Sei ein explizites Runge-Kutta-Verfahren der Stufe s durch $(\mathbf{A}, \mathbf{b}, \mathbf{c})$ gegeben.*

Wendet man es auf das für $\lambda \in \mathbb{R}$ gegebene einfache Anfangswertproblem[1]

$$y_\lambda(0) = 1, \qquad y'_\lambda(t) = \lambda y_\lambda(t) \qquad \text{für alle } t \in \mathbb{R}_{\geq 0} \tag{3.23}$$

an, dessen Lösung offenbar durch $y_\lambda(t) = e^{\lambda t}$ gegeben ist, so gilt für die durch das Verfahren definierte Näherungslösung

$$\tilde{y}_\lambda(t+h;t,x) := x + h\Phi(t,h,x) = g(\lambda h)x \qquad \text{für alle } (t,h) \in \Delta,\ x, \lambda \in \mathbb{R}$$

mit einem Polynom $g \in \Pi_s$, das nur von $\mathbf{A}$, $\mathbf{b}$ und $\mathbf{c}$ abhängt. Dieses Polynom wird manchmal als Stabilitätsfunktion *bezeichnet.*

Beweis. Das Problem (3.23) entspricht dem Anfangswertproblem (2.1) mit der rechten Seite $f(t,x) = \lambda x$. Im trivialen Fall $\lambda = 0$ setzen wir $g \equiv 1$ und sind fertig.

Sei nun $\lambda \neq 0$ angenommen.

Einsetzen der Differentialgleichung in (3.22a) führt auf die Gleichungen

$$k_i = \lambda \left(x + h \sum_{j=1}^{i-1} a_{ij} k_j \right) \qquad \text{für alle } i \in \{1, \dots, s\}. \tag{3.24}$$

Um den Faktor λ an die richtige Stelle zu bringen definieren wir

$$\hat{k}_i := \lambda^{-1} k_i \qquad \text{für alle } i \in \{1, \dots, s\}$$

und schreiben (3.24) in der Form

$$\hat{k}_i = \lambda^{-1} k_i = x + h \sum_{j=1}^{i-1} a_{ij} k_j = x + (\lambda h) \sum_{j=1}^{i-1} a_{ij} \hat{k}_j \qquad \text{für alle } i \in \{1, \dots, s\}, \tag{3.25}$$

die es nahelegt, die skalierten Hilfsvektorn $\hat{k}_i$ als Polynome in λh darzustellen. Konkret suchen wir $q_i \in \Pi_{i-1}$ mit

$$\hat{k}_i = q_i(\lambda h)x \qquad \text{für alle } i \in \{1, \dots, s\}. \tag{3.26}$$

Wir verwenden dazu eine abschnittsweise Induktion, zeigen also

$$\hat{k}_i = q_i(\lambda h)x \qquad \text{für alle } i \in \{1, \dots, \ell\} \tag{3.27}$$

für alle $\ell \in \{1, \dots, s\}$. Der Induktionsanfang $\ell = 1$ ist einfach: Aus (3.25) folgt unmittelbar $\hat{k}_i = x$, also gilt die Behauptung mit $q_1 = 1 \in \Pi_0$.

[1] Auch bekannt als das *Testproblem von Dahlquist.*

Gelte nun (3.27) für ein $\ell \in \{1, \dots, s-1\}$. Dann erhalten wir mit (3.25) und dank der Induktionsvoraussetzung

$$\hat{k}_{\ell+1} = x + (\lambda h) \sum_{j=1}^{\ell} a_{\ell+1,j} \hat{k}_j = x + (\lambda h) \sum_{j=1}^{\ell} a_{\ell+1,j} q_j(\lambda h) x = q_{\ell+1}(\lambda h) x$$

für das Polynom

$$q_{\ell+1}(\zeta) := 1 + \zeta \sum_{j=1}^{\ell} a_{\ell+1,j} q_j(\zeta),$$

das in Π_ℓ liegen muss, da die Polynome q_j für $j \leq \ell$ in $\Pi_{\ell-1}$ liegen. Damit ist die Induktion abgeschlossen und (3.26) bewiesen.

Die Näherungslösung ist nach (3.22b) gegeben durch

$$\begin{aligned} \tilde{y}(t+h; t, x) &= x + h\Phi(t, h, x) = x + h \sum_{i=1}^{s} b_i k_i = x + \lambda h \sum_{i=1}^{s} b_i \hat{k}_i \\ &= x + \lambda h \sum_{i=1}^{s} b_i q_i(\lambda h) x = \left(1 + \lambda h \sum_{i=1}^{s} b_i q_i(\lambda h)\right) x = g(\lambda h) x \end{aligned}$$

für das Polynom

$$g(\zeta) := 1 + \zeta \sum_{i=1}^{s} b_i q_i(\zeta).$$

Da $q_i \in \Pi_{i-1}$ für alle $i \in \{1, \dots, s\}$ gilt, folgt $g \in \Pi_s$. ∎

Offenbar steht die Stabilitätsfunktion g in enger Beziehung zum Approximationsfehler: Die Differenz zwischen exakter Lösung y_λ und approximativer Lösung $\tilde{y}_\lambda$ ist gerade durch

$$y_\lambda(t+h) - \tilde{y}_\lambda(t+h; t, y_\lambda(t)) = e^{\lambda(t+h)} - g(\lambda h) e^{\lambda t} = e^{\lambda t}(e^{\lambda h} - g(\lambda h))$$

gegeben, für eine hohe Konsistenzordnung muss also g eine möglichst gute Approximation der Exponentialfunktion sein. Aus dieser Beobachtung ergibt sich die folgende Schranke für die von einem s-stufigen expliziten Runge-Kutta-Verfahren erreichbare Konsistenzordnung:

Lemma 3.29 (Maximale Ordnung) *Die Konsistenzordnung p eines s-stufigen expliziten Runge-Kutta-Verfahrens beträgt höchstens s. Im Falle $p = s$ gilt*

$$g(\zeta) = \sum_{i=0}^{s} \frac{\zeta^i}{i!}. \tag{3.28}$$

Beweis. Sei ein s-stufiges explizites Runge-Kutta-Verfahren gegeben, und sei $g \in \Pi_s$ die entsprechende Stabilitätsfunktion nach Lemma 3.28. Wir untersuchen den Konsistenzfehler für das Problem (3.23) mit $\lambda \in \mathbb{R}_{>0}$. Mit $t = 0$ und $h \in \mathbb{R}_{>0}$ ist er wegen $y_\lambda(0) = 1$ durch

$$\tau(0, h, 1) = \frac{y_\lambda(h) - 1}{h} - \Phi(0, h, 1) = \frac{y_\lambda(h) - (1 + h\Phi(0, h, 1))}{h}$$

$$= \frac{y_\lambda(h) - \tilde{y}_\lambda(h;0,1)}{h} = \frac{e^{\lambda h} - g(\lambda h)}{h}. \tag{3.29}$$

Indem wir den Satz von Taylor auf das Polynom $g \in \Pi_s$ und die Exponentialfunktion anwenden, finden wir einen Zwischenwert $\eta_h \in [0,h]$ mit

$$g(\lambda h) = \sum_{i=0}^{s} \frac{g^{(i)}(0)(\lambda h)^i}{i!}, \qquad e^{\lambda h} = \sum_{i=0}^{s} \frac{(\lambda h)^i}{i!} + \frac{(\lambda h)^{s+1}}{(s+1)!} + \frac{(\lambda \eta_h)^{s+2}}{(s+2)!},$$

und durch Einsetzen in (3.29) folgt

$$\tau(0,h,1) = \frac{1}{h}\left(\sum_{i=0}^{s} \frac{(1-g^{(i)}(0))(\lambda h)^i}{i!} + \frac{(\lambda h)^{s+1}}{(s+1)!} + \frac{(\lambda h_+)^{s+2}}{(s+2)!}\right).$$

Das Verfahren kann nur konsistent von p-ter Ordnung sein, wenn

$$\frac{|\tau(t,h,x)|}{h^p} = \left|\sum_{i=0}^{s} \frac{(1-g^{(i)}(0))(\lambda h)^i}{i!\ h^{p+1}} + \frac{(\lambda h)^{s+1}}{(s+1)!\ h^{p+1}} + \frac{(\lambda h_+)^{s+2}}{(s+2)!\ h^{p+1}}\right|$$

für $h \to 0$ beschränkt bleibt. Für $p > s$ divergiert der zweite Summand, also kann die Konsistenzordnung nicht größer als p sein.

Damit die maximale Konsistenzordnung $p = s$ erreicht werden kann, muss $g^{(i)}(0) = 1$ für alle $i \in \{0,\dots,p\}$ gelten, denn sonst würde in der Summe ein Term auftreten, der wie $h^{i-(p+1)}$ divergiert. Diese Eigenschaft ist äquivalent zu (3.28). ∎

Die Umkehrung dieser Aussage gilt nicht: Es kann vorkommen, dass zu einer gegebenen Konsistenzordnung p kein p-stufiges Runge-Kutta-Verfahren existiert.

Bemerkung 3.30 (Implizite Verfahren) *Für ein explizites Runge-Kutta-Verfahren muss die zugehörige Matrix A eine strikte untere Dreiecksmatrix sein. Wäre sie es nicht, könnten die Größen $k_1,\dots,k_s$ nicht der Reihe nach explizit berechnet werden.*

Wenn wir beliebige Matrizen A zulassen, erhalten wir im Allgemeinen implizite *Runge-Kutta-Verfahren. Diese Verfahren unterliegen nicht der in Lemma 3.29 bewiesenen Schranke für die Konsistenzordnung, sondern können wesentlich höhere Genauigkeiten erzielen.*

Beispielsweise können auch in diesem Fall die Vektoren c und b entsprechend einer Quadraturformel gewählt werden, etwa entsprechend einer Gauß-Formel. Es ist bekannt, dass eine Gauß-Formel mit s Quadraturpunkten Polynome in Π_{2s-1} exakt integriert, und man kann zeigen, dass das mit Hilfe einer derartigen Formel definierte implizite Runge-Kutta-Verfahren die Konsistenzordnung $2s$ besitzt.

Es ist auch bekannt, dass eine Quadraturformel mit s Quadraturpunkten keine höhere Exaktheitsordnung erreichen kann, mit der Argumentation des Beispiels 3.26 folgt daraus, dass auch ein s-stufiges Runge-Kutta-Verfahren die Konsistenzordnung von $2s$ nicht überschreiten kann.

Ein so definiertes allgemeines Runge-Kutta-Verfahren hat den Nachteil, dass in jedem Schritt ein nichtlineares Gleichungssystem mit s unbekannten Vektoren k_i gelöst werden muss, wodurch ein hoher Rechenaufwand zustande kommen kann.

Einen Mittelweg beschreiten semi-implizite *Runge-Kutta-Verfahren, bei denen die Matrix A zwar eine untere Dreiecksmatrix ist, aber Diagonaleinträge ungleich Null zugelassen werden. Dann können die Vektoren $k_1, k_2, \ldots, k_s$ der Reihe nach bestimmt werden, indem eine Folge von s nichtlinearen Gleichungssystemen für jeweils einen einzelnen Vektor gelöst wird.*

4 Verfeinerte Techniken für gewöhnliche Differentialgleichungen

Die Einschrittverfahren, die wir bisher kennen gelernt haben, sind bereits ausreichend, um viele in der Praxis auftretende Probleme zu behandeln.

In diesem Abschnitt untersuchen wir Verfeinerungen der Technik, die die Genauigkeit verbessern, den Rechenaufwand reduzieren, oder die Stabilität steigern.

4.1 Extrapolation

Unser erstes Ziel besteht darin, Einschrittverfahren beliebig hoher Konsistenzordnung zu konstruieren. Da höhere Ordnungen mit expliziten Runge-Kutta-Verfahren nur schwierig zu erreichen sind, verfolgen wir einen allgemeineren Ansatz: Wir bezeichnen mit $\tilde{y}_h(b)$ die Näherungslösung, die das explizite Euler-Verfahren mit einer konstanten Schrittweite $h \in \mathbb{R}_{>0}$ berechnet. Das ist nur möglich, falls $b - a$ ein Vielfaches von h ist, falls also

$$h \in H_{a,b} := \{h \in \mathbb{R}_{>0} \ : \ (b-a)/h \in \mathbb{N}\}$$

gilt. Laut Korollar 3.19 gilt

$$\|y(b) - \tilde{y}(b)\| \leq C_{\mathrm{eu}} h \qquad \text{für alle } h \in H_{a,b},$$

die diskrete Lösung konvergiert also proportional zu der Schrittweite h gegen die exakte Lösung. Mit einigem Aufwand und unter zusätzlichen Voraussetzungen lässt sich dieses Resultat präzisieren zu

$$\|y(b) - \tilde{y}(b) + he_1\| \leq C_{\mathrm{ae}} h^2 \qquad \text{für alle } h \in H_{a,b} \tag{4.1}$$

wobei $e_1 \in V$ eine von h unabhängige Konstante ist.

Der Ansatz der *Extrapolation* besteht darin, zwei Lösungen zu unterschiedlichen Schrittweiten miteinander zu vergleichen: Wir berechnen die diskrete Lösung $\tilde{y}_h(b)$ mit einer Schrittweite von h und die diskrete Lösung $\tilde{y}_{h/2}(b)$ mit einer Schrittweite von $h/2$. Dann folgen aus (4.1) die Abschätzungen

$$\begin{aligned} \|y(b) - \tilde{y}_h(b) + he_1\| &\leq C_{\mathrm{ae}} h^2, \\ \left\| y(b) - \tilde{y}_{h/2}(b) + \frac{h}{2} e_1 \right\| &\leq \frac{C_{\mathrm{ae}}}{4} h^2, \end{aligned}$$

also anschaulich

$$y(b) \approx \tilde{y}_h(b) + he_1,$$

$$y(b) \approx \tilde{y}_{h/2}(b) + \frac{h}{2} e_1$$

mit einem Fehler, der sich wie h^2 verhält. Den unbekannten Vektor e_1 können wir eliminieren, indem wir eine Linearkombination der beiden Gleichunge bilden:

$$y(b) \approx \hat{y}(b) := 2\tilde{y}_{h/2}(b) - \tilde{y}_h(b).$$

Durch Einsetzen der bekannten Abschätzungen folgt

$$\begin{aligned}
\|y(b) - \hat{y}(b)\| &= \|y(b) - 2\tilde{y}_{h/2}(b) + \tilde{y}_h(b)\| = \|2y(b) - 2\tilde{y}_{h/2}(b) - y(b) + \tilde{y}_h(b)\| \\
&= \left\| 2\left(y(b) - \tilde{y}_{h/2}(b) + \frac{h}{2} e_1 \right) - (y(b) - \tilde{y}_h(b) + he_1) \right\| \\
&\leq 2 \left\| y(b) - \tilde{y}_{h/2}(b) + \frac{h}{2} e_1 \right\| + \|y(b) - \tilde{y}_h(b) + he_1\| \\
&\leq 2C_{\mathrm{ae}} \frac{h^2}{4} + C_{\mathrm{ae}} h^2 = C_{\mathrm{ae}} \left(\frac{1}{2} + 1 \right) h^2,
\end{aligned}$$

also können wir durch Kombination der beiden diskreten Lösungen $\tilde{y}_h$ und $\tilde{y}_{h/2}$ eine neue Lösung $\hat{y}$ berechnen, die eine höhere Genauigkeit erreicht.

Um diesen Ansatz zu verallgemeinern, untersuchen wir ihn aus einem etwas abstrakteren Blickwinkel: Wir untersuchen $\tilde{y}_h(b)$ als Funktion der Schrittweite h, interessieren uns also für

$$g : H_{a,b} \to V, \qquad h \mapsto \tilde{y}_h(b). \tag{4.2}$$

Die Voraussetzung (4.1) nimmt die Form

$$g(h) = \tilde{y}_h(b) = y(b) + e_1 h + r(h)h^2 \qquad \text{für alle } h \in H_{a,b} \tag{4.3}$$

an, wobei $r : H_{a,b} \to V$ eine Funktion ist, die die Abschätzung

$$\|r(h)\| \leq C_{\mathrm{ae}} \qquad \text{für alle } h \in H_{a,b}$$

erfüllt. Die Gleichung (4.3) legt nahe, dass sich g gut durch das lineare Polynom

$$q(h) := y(b) + e_1 h \qquad \text{für alle } h \in \mathbb{R}$$

approximieren lässt. Ideal wäre es natürlich, wenn wir $q(0) = y(b)$ berechnen könnten, allerdings ist das in der Regel nicht möglich.

Stattdessen approximieren wir q durch Interpolation: Wir konstruieren ein lineares Polynom $p \in \Pi_1$, das die Funktion g in den Punkten h und $h/2$ interpoliert, nämlich

$$p(\zeta) = \frac{\zeta - h/2}{h - h/2} \tilde{y}_h(b) + \frac{h - \zeta}{h - h/2} \tilde{y}_{h/2}(b) \qquad \text{für alle } \zeta \in \mathbb{R}.$$

Wir dürfen hoffen, dass p eine passable Approximation des Polynoms q ist, also sollte insbesondere $p(0)$ nicht allzu weit von $q(0) = y(b)$ entfernt liegen. Im Gegensatz zu $q(0)$ lässt sich $p(0)$ allerdings einfach ausrechnen:

$$p(0) = \frac{-h/2}{h - h/2}\tilde{y}_h(b) + \frac{h}{h - h/2}\tilde{y}_{h/2}(b) = 2\tilde{y}_{h/2}(b) - \tilde{y}_h(b) = \hat{y}(b).$$

Bei unserer Herleitung der verbesserten Approximation $\hat{y}(b)$ haben wir also lediglich das Polynom q durch ein Interpolationspolynom p der Funktion g ersetzt und den Wert $p(0)$ als Näherung der exakten Lösung $y(b) = q(0)$ verwendet. Da wir den Interpolanten nicht zwischen den Interpolationspunkten h und $h/2$ auswerten, sondern im Nullpunkt außerhalb des von ihnen begrenzten Intervalls, spricht man von *Extra*polation statt *Inter*polation.

Dieser Zugang lässt sich natürlich verallgemeinern, indem wir Polynome höherer Ordnung verwenden: Sei $m \in \mathbb{N}$, und seien $e_1, e_2, \ldots, e_m \in V$ so gegeben, dass

$$\|y(b) + he_1 + h^2e_2 + \ldots + h^me_m - \tilde{y}_h(b)\| \leq C_{\mathrm{ae}}h^{m+1} \qquad \text{für alle } h \in H_{a,b} \tag{4.4}$$

folgt. Wir verallgemeinern den bereits bekannten Ansatz: Unser Ziel ist es, das Polynom

$$q(h) = y(b) + he_1 + h^2e_2 + \ldots + h^me_m$$

zu approximieren, für das $q(0) = y(b)$ und dank (4.4) auch

$$\|q(h) - \tilde{y}_h(b)\| \leq C_{\mathrm{ae}}h^{m+1} \qquad \text{für alle } h \in H_{a,b} \tag{4.5}$$

gilt. Dazu verwenden wir verschiedene Schrittweiten $h_i = h/\alpha_i$ mit paarweise verschiedenen $\alpha_0, \ldots, \alpha_m \in \mathbb{N}$. Wir berechnen wie üblich Näherungswerte $\tilde{y}_{h_0}(b), \ldots, \tilde{y}_{h_m}(b) \in V$ der Lösung mit diesen Schrittweiten und konstruieren ein Polynom $p \in \Pi_m$, das

$$p(h_i) = \tilde{y}_{h_i}(b) \qquad \text{für alle } i \in \{0, \ldots, m\}$$

erfüllt. Die verbesserte Approximation der Lösung ergibt sich durch Auswertung dieses Polynoms in null, also als

$$\hat{y}(b) := p(0), \tag{4.6}$$

und kann beispielsweise mit dem Neville-Aitken-Verfahren effizient berechnet werden.

Das Polynom p ist lediglich eine Approximation des Polynoms q, die von gemäß (4.5) gestörten Werten ausgeht. Wir sollten also untersuchen, wie sehr derartige Störungen sich auswirken können.

Definition 4.1 (Stabilität der Extrapolation) *Wir bezeichnen mit*

$$\ell_i(\xi) := \prod_{\substack{j=0 \\ j \neq i}}^{m} \frac{\xi - 1/\alpha_j}{1/\alpha_i - 1/\alpha_j} \qquad \textit{für alle } i \in \{0, \ldots, m\},\ \xi \in \mathbb{R}$$

die Lagrange-Polynome zu den Interpolationspunkten $1/\alpha_0, \dots, 1/\alpha_m \in (0,1]$ *und bezeichnen*

$$\Lambda_0 := \sum_{i=0}^{m} |\ell_i(0)|$$

als die Stabilitätskonstante *der Extrapolation.*

Die Stabilitätskonstante erlaubt es uns, den Einfluss von Störungen auf die Qualität der Lösung abzuschätzen:

Lemma 4.2 (Stabilitätskonstante) *Seien* $h_0, \dots, h_m$ *paarweise verschiedene Stützstellen. Sei* $r \in \Pi_m$ *das Interpolationspolynom zu den Stützwerten* $g_0, \dots, g_m \in V$*. Dann gilt*

$$\|r(0)\| \leq \Lambda_0 \max\{\|g_i\| \ : \ i \in \{0, \dots, m\}\}.$$

Beweis. Mit den durch

$$\ell_{h,i}(\zeta) := \prod_{\substack{j=0 \\ j \neq i}}^{m} \frac{\zeta - h_j}{h_i - h_j} \qquad \text{für alle } i \in \{0, \dots, m\},\ \zeta \in \mathbb{R}$$

definierten Lagrange-Polynomen erhalten wir die Darstellung

$$r = \sum_{i=0}^{m} g_i \ell_{h,i}.$$

Wir stellen zunächst fest, dass

$$\begin{aligned} \ell_{h,i}(h\xi) &= \prod_{\substack{j=0 \\ j \neq i}}^{m} \frac{h\xi - h_j}{h_i - h_j} = \prod_{\substack{j=0 \\ j \neq i}}^{m} \frac{h\xi - h/\alpha_j}{h/\alpha_i - h/\alpha_j} \\ &= \prod_{\substack{j=0 \\ j \neq i}}^{m} \frac{\xi - 1/\alpha_j}{1/\alpha_i - 1/\alpha_j} = \ell_i(\xi) \qquad \text{für alle } \xi \in \mathbb{R} \end{aligned}$$

gilt. Daraus folgt mit der Dreiecksungleichung für $\xi = 0$ die Abschätzung

$$\begin{aligned} \|r(0)\| = \left\| \sum_{i=0}^{m} g_i \ell_{h,i}(0) \right\| &\leq \sum_{i=0}^{m} \|g_i\| \ |\ell_{h,i}(0)| = \sum_{i=0}^{m} \|g_i\| \ |\ell_i(0)| \\ &\leq \Lambda_0 \max\{\|g_i\| \ : \ i \in \{0, \dots, m\}\}, \end{aligned}$$

die zu beweisen war. ∎

Die Störung des Werts des Interpolationspolynoms in null lässt sich durch das Produkt aus der Stabilitätskonstanten und dem Maximum der Störungen der Werte in den Interpolationspunkten beschränken. Mit Hilfe dieser Eigenschaft können wir nun die gewünschte Fehlerabschätzung gewinnen:

Satz 4.3 (Extrapolationsfehler) *Wir setzen voraus, dass (4.4) gilt und dass wir $\hat{y}(b)$ gemäß (4.6) berechnen. Dann folgt*

$$\|y(b) - \hat{y}(b)\| \leq C_{\mathrm{ae}} \Lambda_0 h^{m+1}.$$

Beweis. Wir setzen

$$h_i := h/\alpha_i, \qquad g_i := q(h_i) - \tilde{y}_{h_i}(b) = q(h_i) - p(h_i) \qquad \text{für alle } i \in \{0, \ldots, m\}$$

und stellen fest, dass $r := q - p$ die Gleichungen

$$r(h_i) = g_i \qquad \text{für alle } i \in \{0, \ldots, m\}$$

erfüllt, also die Werte $g_0, \ldots, g_m$ in $h_0, \ldots, h_m$ interpoliert. Mit Lemma 4.2 folgt

$$\|y(b) - \hat{y}(b)\| = \|q(0) - p(0)\| \leq \Lambda_0 \max\{\|g_i\| \ : \ i \in \{0, \ldots, m\}\}.$$

Aus (4.5) erhalten wir

$$\|g_i\| = \|q(h_i) - \tilde{y}_{h_i}(b)\| \leq C_{\mathrm{ae}} h_i^{m+1} \leq C_{\mathrm{ae}} h^{m+1} \qquad \text{für alle } i \in \{0, \ldots, m\},$$

also insgesamt

$$\|y(b) - \hat{y}(b)\| \leq \Lambda_0 C_{\mathrm{ae}} h^{m+1}$$

und damit die gewünschte Aussage. ∎

Bemerkung 4.4 (Lokale Extrapolation) *In der Praxis wird Extrapolation in der Regel nicht für die Berechnung des Endergebnisses $y(b)$, sondern für die Durchführung eines einzelnen Schritts verwendet: Um von $y(t)$ zu $y(t+h)$ zu gelangen, werden Schrittweiten $h_i := h/\alpha_i$ eingefügt und Näherungslösungen $\tilde{y}_{h_i}(t + h; t, y(t))$ mit Hilfe von α_i Schritten des ursprünglichen Verfahrens bestimmt. Aus diesen Näherungen wird dann wie zuvor eine verbesserte Näherung $\hat{y}(t+h; t, y(t))$ berechnet, die als Grundlage für den nächsten Schritt dient. Auf diese Weise lassen sich Verfahren beliebig hoher Konsistenzordnung konstruieren, allerdings sind für jeden Gesamtschritt $\alpha_0 + \ldots + \alpha_m$ Schritte des ursprünglichen Verfahrens durchzuführen.*

4.2 Schrittweitensteuerung

Bei den bisher diskutierten Verfahren ist der Rechenaufwand für die Durchführung eines Schritts von einem Zeitpunkt t_i zu dem folgenden Zeitpunkt t_{i+1} immer derselbe, also ist der Aufwand für die Berechnung der gesamten Lösung proportional zu der Anzahl der Schritte. Um die Rechenzeit zu reduzieren ist es deshalb sehr erstrebenswert, nach Verfahren zu suchen, die eine vorgegebene Genauigkeit mit einer möglichst geringen Anzahl von Schritten erreichen.

Die Technik der *adaptiven Schrittweitensteuerung* verwendet Informationen über das lokale Verhalten der Lösung, um dieses Ziel zu erreichen: Zu jedem Zeitpunkt t_i wird

abgeschätzt, wie sich die Lösung auf dem Teilintervall $[t_i, t_{i+1}] = [t_i, t_i + h_i]$ verhalten wird, und dann wird h_i so klein gewählt, dass der Fehler unterhalb einer Schranke bleibt.

Als Beispiel verwenden wir wieder das explizite Euler-Verfahren: Falls $y \in C^2([a,b], V)$ gilt, erhalten wir mit dem Mittelwertsatz der Differentialrechnung

$$\tau(t,h,y(t)) = \frac{y(t+h)-y(t)}{h} - y'(t) = y'(\xi_t) - y'(t) = (\xi_t - t)y''(\eta_t)$$

für Zwischenpunkte $\xi_t \in [t, t+h]$ und $\eta_t \in [t, \xi_t] \subseteq [t, t+h]$, wir können also den Konsistenzfehler durch

$$\|\tau(t,h,y(t))\| \leq h\|y''\|_{\infty,[t,t+h]}$$

beschränken. Falls uns Abschätzungen für y'' zur Verfügung stehen, können wir wie folgt vorgehen: Wir fixieren ein $\hat{\epsilon} \in \mathbb{R}_{>0}$. Zunächst wählen wir $h_0 \in \mathbb{R}_{>0}$ so, dass

$$\|\tau(t_0,h_0,y(t_0))\| \leq h_0\|y''\|_{\infty,[t_0,t_0+h_0]} \leq \hat{\epsilon}$$

gilt. Nun können wir $t_1 = t_0 + h_0$ berechnen und nach einem $h_1 \in \mathbb{R}_{>0}$ suchen, das

$$\|\tau(t_1,h_1,y(t_1))\| \leq h_1\|y''\|_{\infty,[t_1,t_1+h_1]} \leq \hat{\epsilon}$$

erfüllt. Damit erhalten wir $t_2 = t_1 + h_1$ und fahren fort, bis wir den Endpunkt b des zu behandelnden Intervalls erreicht haben. Analog zu dem Beweis des Satzes 3.18 folgt aus Satz 3.13 dann, dass der Fehler sich in der Form

$$\|y(b) - \tilde{y}(b)\| \leq \hat{\epsilon}(b-a)e^{L_\Phi(b-a)}$$

beschränken lässt, bei geeigneter Wahl von $\hat{\epsilon}$ können wir also jede beliebige Genauigkeit erreichen. Im Gegensatz zu vorher werden allerdings diesmal die Schrittweiten an das Verhalten der Lösung angepasst: Falls sich die Lösung kaum ändert, genügen große Intervalle, anderenfalls werden automatisch kleinere Intervalle eingesetzt.

Dieses Verhalten wäre ideal, ist aber in der Praxis nicht zu erreichen, da unsere Konstruktion die Kenntnis der Lösung und ihrer Ableitungen auf dem gesamten Intervall voraussetzt. Das erste Problem lässt sich lösen, indem wir die Beweisführung des Satzes 3.13 etwas modifizieren: Als „Referenzlösung“ dient uns nun die diskrete Lösung, und die exakte Lösung wird als Störung behandelt. Wir erhalten so das folgende Resultat:

Satz 4.5 (Konvergenz) *Sei f Lipschitz-stetig im zweiten Argument mit der Lipschitz-Konstanten L_f. Wir bezeichnen mit*

$$\tilde{y}_i := \tilde{y}(t_i) \qquad \text{für alle } i \in \{0,\ldots,n\}$$

die Werte der diskreten Lösung des Anfangswertproblems (2.1) und mit

$$\widehat{K}_\Phi := \max\{\|\tau(t_{i-1},h_i,\tilde{y}_{i-1})\| \ : \ i \in \{1,\ldots,n\}\} \tag{4.7}$$

den maximalen Fehler, den das Einschrittverfahren in einem Schritt ausgehend von der Näherungslösung *verursachen kann.*

Dann folgt die Abschätzung

$$\|y(t_k) - \tilde{y}(t_k)\| \leq \begin{cases} \frac{e^{L_f(t_k-a)}-1}{L_f}\widehat{K}_\Phi & \text{falls } L_f > 0, \\ (t_k - a)\widehat{K}_\Phi & \text{ansonsten} \end{cases} \quad \text{für alle } k \in \{0,\dots,n\}.$$

Beweis. Wir gehen wie im Beweis des Satzes 3.13 vor, schieben diesmal allerdings den Wert der *exakten* Lösung ausgehend von der *diskreten* Lösung ein: Für $i \leq j \leq j+1 = k$ erhalten wir mit Hilfe der Fortsetzungseigenschaften aus den Lemmas 3.6 und 3.9 und der Dreiecksungleichung die Abschätzung

$$\begin{aligned}\|y(t_k;t_i,\tilde{y}_i) - \tilde{y}(t_k;t_i,\tilde{y}_i)\| &= \|y(t_k;t_j,y(t_j;t_i,\tilde{y}_i)) - \tilde{y}(t_k;t_j,\tilde{y}(t_j;t_i,\tilde{y}_i))\| \\ &\leq \|y(t_k;t_j,y(t_j;t_i,\tilde{y}_i)) - y(t_k;t_j,\tilde{y}(t_j;t_i,\tilde{y}_i))\| \\ &\quad + \|y(t_k;t_j,\tilde{y}(t_j;t_i,\tilde{y}_i)) - \tilde{y}(t_k;t_j,\tilde{y}(t_j;t_i,\tilde{y}_i))\|.\end{aligned}$$

Der erste Summand beschreibt das Verhalten der exakten Lösung bei Störung der Anfangswerte, lässt sich also mit Hilfe des Satzes 2.6 abschätzen:

$$\|y(t_k;t_j,y(t_j;t_i,\tilde{y}_i) - y(t_k;t_j,\tilde{y}(t_j;t_i,\tilde{y}_i)\| \leq e^{L_f(t_k-t_j)}\|y(t_j;t_i,\tilde{y}_i) - \tilde{y}(t_j;t_i,\tilde{y}_i)\|.$$

Der zweite Term ist dank $k = j+1$ und $\tilde{y}(t_j;t_i,\tilde{y}_i) = \tilde{y}_j$ durch $\widehat{K}_\Phi h_j$ beschränkt, so dass wir insgesamt

$$\|y(t_k;t_i,\tilde{y}_i) - \tilde{y}(t_k;t_i,\tilde{y}_i)\| \leq e^{L_f(t_k-t_j)}\|y(t_j;t_i,\tilde{y}_i) - \tilde{y}(t_j;t_i,\tilde{y}_i)\| + \widehat{K}_\Phi h_j$$

erhalten. Wie schon im Beweis des Satzes 3.13 können wir nun eine Induktion über $k - i \in \mathbb{N}_0$ durchführen, um die gewünschte Abschätzung zu erreichen. ∎

Um diesen Satz anwenden zu können, müssen wir sicherstellen, dass die Konsistenzfehler $\tau(t_{i-1}, h_i, \tilde{y}_{i-1})$ unter Kontrolle sind. Diese Fehler hängen lediglich von dem *lokalen* Verhalten der Lösung ab, sollten sich also auch lokal steuern lassen, indem die Schrittweite h_i geeignet gewählt wird.

Da wir den Fehler in der Regel nicht exakt kennen, sind wir auf Abschätzungen angewiesen, die sich praktisch berechnen lassen. Ein guter Ansatz besteht darin, zwei Einschrittverfahren Φ_1 und Φ_2 unterschiedlicher Ordnung zu verwenden.

Wir bezeichnen die zu Φ_1 und Φ_2 gehörenden Näherungslösungen mit $\tilde{y}_1$ und $\tilde{y}_2$ und die entsprechenden Konsistenzfehler mit τ_1 und τ_2 und fordern, dass die beiden Verfahren konsistent von p-ter beziehungsweise $(p+1)$-ter Ordnung sind, dass also Konstanten $h_0, C_1, C_2 \in \mathbb{R}_{>0}$ mit

$$\|\tau_1(t_i,h,\tilde{y}_i)\| \leq C_1 h^p, \quad \|\tau_2(t_i,h,\tilde{y}_i)\| \leq C_2 h^{p+1} \quad \text{für alle } h \in (0,h_0),\ i \in \{0,\dots,n-1\}$$

existieren. Durch einen Vergleich der beiden Lösungen $\tilde{y}_1$ und $\tilde{y}_2$ können wir den Konsistenzfehler abschätzen: Es gilt

$$\|\tau_1(t_i,h,\tilde{y}_i)\| = \frac{\|y(t_i+h;t_i,\tilde{y}_i) - \tilde{y}_1(t_i+h;t_i,\tilde{y}_i)\|}{h}$$

$$
\begin{aligned}
&= \frac{\|\tilde{y}_2(t_i+h;t_i,\tilde{y}_i) - \tilde{y}_1(t_i+h;t_i,\tilde{y}_i) + y(t_i+h;t_i,\tilde{y}_i) - \tilde{y}_2(t_i+h;t_i,\tilde{y}_i)\|}{h} \\
&\leq \frac{\|\tilde{y}_2(t_i+h;t_i,\tilde{y}_i) - \tilde{y}_1(t_i+h;t_i,\tilde{y}_i)\|}{h} + \|\tau_2(t_i,h,\tilde{y}_i)\| \\
&\leq \frac{\|\tilde{y}_2(t_i+h;t_i,\tilde{y}_i) - \tilde{y}_1(t_i+h;t_i,\tilde{y}_i)\|}{h} + C_2 h^{p+1},
\end{aligned}
$$

für hinreichend kleine Werte von h können wir also abschätzen, wie groß der Konsistenzfehler für eine gegebene Schrittweite ist. Ist er zu groß, reduzieren wir die Schrittweite und prüfen, ob der neue geschätzte Fehler unter der gegebenen Schranke $\hat{\epsilon}$ liegt. Entsprechend können wir bei einem zu kleinen Fehler die Schrittweite auch wieder erhöhen. Der Nachteil dieser Methode besteht darin, dass sie wiederholte Berechnungen der Näherungen $\tilde{y}_1$ und $\tilde{y}_2$ benötigt, um eine gute Schrittweite zu finden.

Eine elegantere Methode beruht auf der Idee der Extrapolation. Wir gehen davon aus, dass sich der Fehler τ_1 in Null in einer Taylor-Reihe der Ordnung $p+1$ entwickeln lässt, dass also ein Vektor $e_p \in V$ und eine Konstante $C_c \in \mathbb{R}_{>0}$ mit

$$
\|\tau_1(t,h,\tilde{y}_i) - h^p e_p\| \leq C_c h^{p+1} \qquad \text{für alle } h \in (0,h_0),\ i \in \{0,\ldots,n-1\} \tag{4.8}
$$

existieren. Aufgrund dieser Annahme gilt

$$
\begin{aligned}
&\tilde{y}_2(t_i+h;t_i,\tilde{y}_i) - \tilde{y}_1(t_i+h;t_i,\tilde{y}_i) \\
&\qquad = \tilde{y}_2(t_i+h;t_i,\tilde{y}_i) - y(t_i+h;t_i,\tilde{y}_i) + y(t_i+h;t_i,\tilde{y}_i) - \tilde{y}_1(t_i+h;t_i,\tilde{y}_i) \\
&\qquad = -h\tau_2(t_i,h,\tilde{y}_i) + h\tau_1(t_i,h,\tilde{y}_i) \\
&\qquad = hh^p e_p - h\tau_2(t_i,h,\tilde{y}_i) + h\tau_1(t_i,h,\tilde{y}_i) - hh^p e_p \\
&\qquad = h^{p+1} e_p - h\tau_2(t_i,h,\tilde{y}_i) + h(\tau_1(t_i,h,\tilde{y}_i) - h^p e_p),
\end{aligned}
$$

und indem wir $h_i^{p+1} e_p$ auf beiden Seiten der Gleichung subtrahieren, durch h^{p+1} dividieren und die Norm berechnen, erhalten wir

$$
\begin{aligned}
\left\| \frac{\tilde{y}_1(t_i+h;t_i,\tilde{y}_i) - \tilde{y}_2(t_i+h;t_i,\tilde{y}_i)}{h^{p+1}} - e_p \right\| &= h \frac{\|\tau_2(t_i,h,\tilde{y}_i)\| + \|\tau_1(t_i,h,\tilde{y}_i) - h^p e_p\|}{h^{p+1}} \\
&\leq (C_c + C_2)h,
\end{aligned}
$$

wir können also zumindest den führenden Term des lokalen Fehlers $\tau_1(t_i,h,\tilde{y}_i)$ approximativ bestimmen. Das bietet uns die Möglichkeit, die Schrittweite zu regeln: Wenn h hinreichend klein ist, ist

$$
\tilde{e}_p := \frac{\tilde{y}_1(t_i+h;t_i,\tilde{y}_i) - \tilde{y}_2(t_i+h;t_i,\tilde{y}_i)}{h^{p+1}}
$$

eine gute Approximation von e_p, und es gilt

$$
\begin{aligned}
\|\tau_1(t_i,\hat{h},\tilde{y}_i)\| &= \|\tau_1(t_i,\hat{h},\tilde{y}_i) - \hat{h}^p e_p + \hat{h}^p e_p\| \leq C_c \hat{h}^{p+1} + \hat{h}^p \|e_p\| \\
&\leq C_c \hat{h}^{p+1} + \hat{h}^p \|\tilde{e}_p\| + \hat{h}^p \|\tilde{e}_p - e_p\| \leq C_c \hat{h}^{p+1} + (C_c + C_2)\hat{h}^{p+1} + \hat{h}^p \|\tilde{e}_p\|
\end{aligned}
$$

$$= (2C_c + C_2)\hat{h}^{p+1} + \hat{h}^p \|\tilde{e}_p\| \qquad \text{für alle } \hat{h} \in (0, h_0).$$

Für hinreichend kleines $\hat{h}$ können wir den ersten Term vernachlässigen, müssen also lediglich

$$\hat{h}^p \|\tilde{e}_p\| \leq \hat{\epsilon}$$

sicherstellen. Das ist äquivalent zu

$$\hat{h}^p \leq \frac{\hat{\epsilon}}{\|\tilde{e}_p\|} = \frac{\hat{\epsilon} h^{p+1}}{\|\tilde{y}_1(t_i + h; t_i, \tilde{y}_i) - \tilde{y}_2(t_i + h; t_i, \tilde{y}_i)\|},$$
$$\hat{h} \leq h \left(\frac{\hat{\epsilon} h}{\|\tilde{y}_1(t_i + h; t_i, \tilde{y}_i) - \tilde{y}_2(t_i + h; t_i, \tilde{y}_i)\|} \right)^{1/p}.$$

Die praktische Vorgehensweise sieht nun wie folgt aus: Wir gehen von einer nicht zu großen Schrittweite h aus und führen jeweils einen Schritt mit dieser Schrittweite und beiden Verfahren durch. Aus der Differenz der Ergebnisse ermitteln wir dann mit Hilfe der obigen Formel eine verbesserte Schrittweite $\hat{h}$.

Ein praktischer Nachteil der Technik besteht darin, dass wir für jedes der beiden Näherungsverfahren jeweils einen Schritt durchführen müssen, es kann also bei naivem Vorgehen fast der doppelte Rechenaufwand entstehen. Dieser Nachteil lässt sich mit Hilfe eines *Runge-Kutta-Fehlberg-Verfahrens* abmildern: Wenn wir das explizite Euler-Verfahren

$$\tilde{y}_1(t_i + h; t_i, \tilde{y}_i) = \tilde{y}_i + h f(t_i, \tilde{y}_i)$$

mit dem Verfahren von Heun (siehe Beispiel 3.24)

$$\tilde{y}_2(t_i + h; t_i, \tilde{y}_i) = \tilde{y}_i + \frac{h}{2} \left(f(t_i, \tilde{y}_i) + f(t_i + h, \tilde{y}_i + h f(t_i, \tilde{y}_i)) \right)$$

vergleichen, stellen wir fest, dass beide durch die Hilfsgrößen

$$k_1 := f(t_i, \tilde{y}_i), \qquad k_2 := f(t_i + h, \tilde{y}_i + h k_1)$$

durch

$$\tilde{y}_1(t_i + h; t_i, \tilde{y}_i) = \tilde{y}_i + h k_1, \qquad \tilde{y}_2(t_i + h; t_i, \tilde{y}_i) = \tilde{y}_i + \frac{h}{2}(k_1 + k_2)$$

dargestellt werden können. Es genügen also zwei Auswertungen der rechten Seite f, um beide Näherungen zu berechnen, während es bei einer naiven Vorgehensweise drei wären. Wir haben bereits gesehen, dass das explizite Euler-Verfahren eine Konsistenzordnung von eins besitzt, während sie für das Heun-Verfahren zwei beträgt, also könnten wir die beiden Verfahren in der zuvor beschriebenen Weise kombinieren, um die Schrittweite des Euler-Verfahrens zu regeln.

Abstrakter gesehen sind beide Verfahren Runge-Kutta-Verfahren, die dieselben Hilfsgrößen k_1, k_2 verwenden. Die Idee der Runge-Kutta-Fehlberg-Verfahren besteht darin, Paare von Runge-Kutta-Verfahren höherer Ordnung zu konstruieren, die sich dieselben Hilfsgrößen $k_1, \ldots, k_s$ teilen.

Ein typischer Vertreter ist das Runge-Kutta-Fehlberg-Verfahren der Ordnungen 4 und 5, das durch das folgende verallgemeinerte Butcher-Schema beschrieben ist:

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

Das erweiterte Butcher-Tableau ist wie folgt zu interpretieren: Die beiden letzten Zeilen geben Vektoren $\mathbf{b}$ und $\widehat{\mathbf{b}}$ an, mit denen die beiden Näherungslösungen

$$\tilde{y}_1(t+h;t,x) = x + h\sum_{i=1}^{s-1} b_i k_i, \qquad \tilde{y}_2(t+h;t,x) = x + h\sum_{i=1}^{s} \hat{b}_i k_i$$

berechnet werden können. Die erste Näherung ist konsistent von vierter Ordnung, die zweite von fünfter. Wie wir sehen handelt es sich um ein siebenstufiges Verfahren. Wenn wir ein Runge-Kutta-Verfahren vierter Ordnung, das mindestens vier Stufen erfordert, und ein Verfahren fünfter Ordnung, für das mindestens sechs Stufen nötig sind, verwenden würden, wären insgesamt zehn Auswertungen der Funktion f erforderlich, die Fehlberg-Methode spart also ungefähr 30 Prozent des Rechenaufwands ein.

Ein genauerer Blick zeigt, dass wir im Mittel sogar mit ungefähr sechs Auswertungen von f pro Schritt auskommen können: Da die letzte Zeile der Matrix $\mathbf{A}$ mit dem Vektor $\mathbf{b}$ übereinstimmt, ist der Wert k_7 eines Schritts gleich dem Wert k_1 des nächsten Schritts, so dass wir uns in allen Schritten außer dem ersten die Berechnung von k_1 sparen können.

Bemerkung 4.6 (Warnung) *Alle hier vorgestellten Verfahren zur Steuerung der Schrittweite zur Schätzung des Fehlers beruhen auf Annahmen über die Struktur der Lösung: In die Konstanten C_1 und C_2 aus den Konsistenzbedingungen der Verfahren Φ_1 und Φ_2 gehen in der Regel die Ableitungen der* unbekannten *Lösung y ein, und dasselbe gilt für die Konstante C_c aus Abschätzung (4.8). Die Voraussetzung, dass die Schrittweiten „klein genug“ sein müssen, um den Fehler abschätzen zu können, lässt sich deshalb in der Praxis oft nicht nachprüfen, weil sie von unbekannten Größen abhängt. Im schlimmsten Fall kann es passieren, dass y gar nicht hinreichend glatt ist, um tatsächlich eine höhere Ordnung von Φ_2 feststellen zu können.*

Deshalb empfiehlt es sich, bei konkreten Implementierungen darauf zu achten, dass eine Möglichkeit vorgesehen wird, die Schrittweite unabhängig von allen anderen Parametern zu beschränken, so dass notfalls durch Reduktion der Schrittweite eine Konvergenz mit Hilfe des Satzes 3.13 erzwungen werden kann.

4.3 Steife Differentialgleichungen

Bisher waren die von uns analysierten numerischen Verfahren überwiegend explizit, und wir haben gesehen, dass bei einer hinreichend kleinen Schrittweite diese Verfahren brauchbare Approximationen der Lösung eines Anfangswertproblems bestimmen.

Die Bedeutung von „hinreichend klein“ hängt dabei im Wesentlichen davon ab, wie groß die Lipschitz-Konstante der rechten Seite f ist: Je größer sie ist, desto „weniger glatt“ ist die Lösung, und desto kleiner müssen wir die Schrittweite wählen, desto höher wird also der Rechenaufwand.

Es gibt Situationen, in denen eine große Lipschitz-Konstante nicht unbedingt eine geringe Schrittweite erforderlich macht, weil der Term, der die Konstante in die Höhe treibt, nur geringen Einfluss auf die exakte Lösung des Problems hat. In diesen Situationen sind wir daran interessiert, Verfahren zu entwickeln, die diese Eigenschaft erben, also mit einer größeren Schrittweite trotzdem eine gute Approximation bestimmen können.

Als Beispiel befassen wir uns mit dem linearen Anfangswertproblem

$$\mathbf{y}(0) = \mathbf{y}_0, \qquad \mathbf{y}'(t) = \mathbf{A}\mathbf{y}(t) \qquad \text{für alle } t \in \mathbb{R}_{\geq 0}$$

mit einem Startvektor $\mathbf{y}_0 \in \mathbb{R}^2$ und der Matrix

$$\mathbf{A} := \frac{1}{2}\begin{pmatrix} \alpha+\beta & \alpha-\beta \\ \alpha-\beta & \alpha+\beta \end{pmatrix}$$

mit Koeffizienten $\alpha, \beta \in \mathbb{R}$.

Zunächst bestimmen wir die bestmögliche Lipschitz-Konstante L_f für die korrespondierende rechte Seite $f(t, \mathbf{x}) = \mathbf{A}\mathbf{x}$. Da $\mathbf{A}$ symmetrisch ist, können wir diese Aufgabe durch eine Eigenwertbetrachtung lösen: Mit

$$\mathbf{Q} := \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$$

erhalten wir $\mathbf{Q}^*\mathbf{Q} = \mathbf{I}$, also $\mathbf{Q}^* = \mathbf{Q}^{-1}$, und

$$\mathbf{Q}^*\mathbf{A}\mathbf{Q} = \frac{1}{4}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} \alpha+\beta & \alpha-\beta \\ \alpha-\beta & \alpha+\beta \end{pmatrix}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} \alpha & \\ & \beta \end{pmatrix},$$

also folgt sofort $\|\mathbf{A}\|_2 = \max\{|\alpha|, |\beta|\}$ und

$$\|f(t, \mathbf{x}) - f(t, \mathbf{z})\|_2 = \|\mathbf{A}(\mathbf{x} - \mathbf{z})\|_2 \leq \|\mathbf{A}\|_2 \|\mathbf{x} - \mathbf{z}\|_2,$$

so dass $L_f := \|\mathbf{A}\|_2$ eine gute Wahl für die Lipschitz-Konstante ist. Da wir $\mathbf{x} - \mathbf{z}$ als Eigenvektor zu dem betragsgrößeren der beiden Eigenwerte wählen können, ist L_f sogar die bestmögliche Lipschitz-Konstante.

Zur Analyse des Verhaltens der Lösung $\mathbf{y}$ führen wir die Hilfsfunktion $\widehat{\mathbf{y}}$ mit

$$\widehat{\mathbf{y}}(t) := \mathbf{Q}^*\mathbf{y}(t) \qquad \text{für alle } t \in \mathbb{R}_{\geq 0}$$

ein. Wegen $\mathbf{y}(t) = \mathbf{Q}\widehat{\mathbf{y}}(t)$ muss

$$\widehat{\mathbf{y}}'(t) = \mathbf{Q}^*\mathbf{y}'(t) = \mathbf{Q}^*\mathbf{A}\mathbf{y}(t) = \mathbf{Q}^*\mathbf{A}\mathbf{Q}\widehat{\mathbf{y}}(t) = \begin{pmatrix} \alpha & \\ & \beta \end{pmatrix} \widehat{\mathbf{y}}(t) \qquad \text{für alle } t \in \mathbb{R}_{\geq 0}$$

gelten, die Hilfsfunktion löst also das Anfangswertproblem

$$\widehat{\mathbf{y}}(0) = \widehat{\mathbf{y}}_0 := \mathbf{Q}^*\mathbf{y}_0, \qquad \widehat{\mathbf{y}}'(t) = \begin{pmatrix} \alpha & \\ & \beta \end{pmatrix} \widehat{\mathbf{y}}(t) \qquad \text{für alle } t \in \mathbb{R}_{\geq 0}.$$

In diesem Problem sind die beiden Komponenten von $\widehat{\mathbf{y}}$ voneinander entkoppelt, so dass sich die Lösung explizit durch

$$\hat{y}_1(t) = \hat{y}_{0,1}e^{\alpha t}, \qquad \hat{y}_2(t) = \hat{y}_{0,2}e^{\beta t} \qquad \text{für alle } t \in \mathbb{R}_{\geq 0}$$

darstellen lässt. Aus $\mathbf{y}(t) = \mathbf{Q}\widehat{\mathbf{y}}(t)$ folgt direkt

$$\mathbf{y}(t) = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \hat{y}_{0,1}e^{\alpha t} \\ \hat{y}_{0,2}e^{\beta t} \end{pmatrix} = \mathbf{a}e^{\alpha t} + \mathbf{b}e^{\beta t} \qquad \text{für alle } t \in \mathbb{R}_{\geq 0}$$

mit den Hilfsvektoren

$$\mathbf{a} := \frac{\hat{y}_{0,1}}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad \mathbf{b} := \frac{\hat{y}_{0,2}}{\sqrt{2}} \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

Untersuchen wir nun das Verhalten eines Näherungsverfahrens. Der Einfachheit halber beschränken wir uns auf das explizite Euler-Verfahren, das bei fester Schrittweite h die Näherungslösung $\widetilde{\mathbf{y}}(t)$ mit

$$\widetilde{\mathbf{y}}(t_{j+1}) = \widetilde{\mathbf{y}}(t_j) + h\mathbf{A}\widetilde{\mathbf{y}}(t_j) = (\mathbf{I} + h\mathbf{A})\widetilde{\mathbf{y}}(t_j) \qquad \text{für alle } j \in \mathbb{N}_0$$

berechnet. Per Induktion erhalten wir

$$\widetilde{\mathbf{y}}(t_j) = (\mathbf{I} + h\mathbf{A})^j\widetilde{\mathbf{y}}_0 \qquad \text{für alle } j \in \mathbb{N}_0, \tag{4.9}$$

und dank

$$\mathbf{A} = \mathbf{Q} \begin{pmatrix} \alpha & \\ & \beta \end{pmatrix} \mathbf{Q}^*, \qquad \mathbf{I} + h\mathbf{A} = \mathbf{Q} \begin{pmatrix} 1+h\alpha & \\ & 1+h\beta \end{pmatrix} \mathbf{Q}^*,$$
$$(\mathbf{I} + h\mathbf{A})^j = \mathbf{Q} \begin{pmatrix} (1+h\alpha)^j & \\ & (1+h\beta)^j \end{pmatrix} \mathbf{Q}^*$$

können wir die diskrete Lösung explizit durch

$$\widetilde{\mathbf{y}}(t_j) = \mathbf{a}(1+h\alpha)^j + \mathbf{b}(1+h\beta)^j \qquad \text{für alle } j \in \mathbb{N}_0$$

darstellen. Wir vergleichen die exakte und die genäherte Lösung:

$$\begin{array}{rcccl} \mathbf{y}(t_j) & = & \mathbf{a}e^{\alpha h j} & + & \mathbf{b}e^{\beta h j}, \\ \widetilde{\mathbf{y}}(t_j) & = & \mathbf{a}(1+\alpha h)^j & + & \mathbf{b}(1+\beta h)^j \qquad \text{für alle } j \in \mathbb{N}_0 \end{array}$$

Falls $\alpha, \beta > 0$ gilt, verhalten sich die beiden Funktionen ähnlich: Für $t \to \infty$ streben sie exponentiell gegen unendlich.

Anders sieht es im Fall $\alpha, \beta < 0$ aus: Jetzt konvergiert $\mathbf{y}(t_j)$ gegen Null, wenn wir t_j gegen unendlich gehen lassen, aber für $\widetilde{\mathbf{y}}(t_j)$ gilt das nur, wenn die Schrittweite h klein genug ist, wenn also

$$|1 + h\alpha| < 1, \qquad |1 + h\beta| < 1$$

gilt. Wegen $\alpha, \beta < 0$ und $L_f = \max\{|\alpha|, |\beta|\}$ ist das gerade für $h < 2/L_f$ sichergestellt, die Schrittweite wird also tatsächlich durch den betragsgrößten Eigenwert bestimmt.

Falls $\beta \ll \alpha < 0$ gilt, ist für das langfristige Verhalten die Lösung nur der von α abhängige Term relevant, weil $e^{\beta t}$ sehr schnell gegen null streben wird. Trotzdem müssen wir in unserem Näherungsverfahren die Schrittweite so wählen, dass $h < 2/L_f$ mit $L = |\beta|$ gilt, wir müssen also etwas approximieren, das uns eigentlich gar nicht interessiert.

Ein Anfangswertproblem, bei dem zwar die Lösung für lange Zeiträume gegen null strebt, bei dem aber verschiedene Anteile der Lösung das mit sehr unterschiedlichen Geschwindigkeiten tun, bezeichnet man als *steif*. Bis zu einer gewissen Grenzschrittweite (in unserem Beispiel $2/L$) berechnet das Näherungsverfahren unbrauchbare Lösungen, sobald diese Schrittweite unterschritten wird, funktioniert plötzlich alles.

Da steife Anfangswertprobleme in vielen Anwendungen auftreten, stellt sich die Frage, ob man Verfahren finden kann, die nicht auf die sehr restriktive Bedingung $h < 2/L_f$ angewiesen sind.

In dieser Hinsicht sehr erfolgreich sind implizite Verfahren. Als Beispiel stellen wir dem expliziten Euler-Verfahren das implizite Euler-Verfahren gegenüber, das in unserem Beispiel die Form

$$\widetilde{\mathbf{y}}(t_{j+1}) = \widetilde{\mathbf{y}}(t_j) + h\mathbf{A}\widetilde{\mathbf{y}}(t_{j+1}) \qquad \text{für alle } j \in \mathbb{N}_0$$

annimmt. Da unsere Gleichung linear ist, können wir $\widetilde{\mathbf{y}}(t_{j+1})$ direkt berechnen, indem wir ein lineares Gleichungssystem lösen: Es gilt

$$(\mathbf{I} - h\mathbf{A})\widetilde{\mathbf{y}}(t_{j+1}) = \widetilde{\mathbf{y}}(t_j) \qquad \text{für alle } j \in \mathbb{N}_0. \tag{4.10}$$

Wie zuvor können wir die Matrix diagonalisieren, um eine explizite Darstellung der diskreten Lösung zu erhalten: Aus

$$\mathbf{A} = \mathbf{Q}\begin{pmatrix}\alpha & \\ & \beta\end{pmatrix}\mathbf{Q}^*, \qquad \mathbf{I} - h\mathbf{A} = \mathbf{Q}\begin{pmatrix}1-h\alpha & \\ & 1-h\beta\end{pmatrix}\mathbf{Q}^*,$$
$$(\mathbf{I} - h\mathbf{A})^{-1} = \mathbf{Q}\begin{pmatrix}\frac{1}{1-h\alpha} & \\ & \frac{1}{1+h\beta}\end{pmatrix}\mathbf{Q}^*$$

folgt die Darstellung

$$\widetilde{\mathbf{y}}(t_{j+1}) = \mathbf{Q}\begin{pmatrix}\frac{1}{1-h\alpha} & \\ & \frac{1}{1-h\beta}\end{pmatrix}\mathbf{Q}^*\widetilde{\mathbf{y}}(t_j) \qquad \text{für alle } j \in \mathbb{N}_0,$$

und per Induktion erhalten wir schließlich

$$\widetilde{\mathbf{y}}(t_j) = \mathbf{Q}\begin{pmatrix}(1-h\alpha)^{-j}\hat{y}_{0,1}\\(1-h\beta)^{-j}\hat{y}_{0,2}\end{pmatrix} = \mathbf{a}(1-\alpha h)^{-j} + \mathbf{b}(1-\beta h)^{-j} \qquad \text{für alle } j \in \mathbb{N}_0.$$

Diese Näherungslösung besitzt andere Eigenschaften als die, die wir im Falle des expliziten Verfahrens erhalten haben: Sie könnte nur dann divergieren, wenn $|1-\alpha h| < 1$ oder $|1-\beta h| < 1$ gilt, aber dank $\alpha, \beta < 0$ ist das ausgeschlossen.

Das implizite Euler-Verfahren wird also eine Lösung berechnen, die für *beliebige* Schrittweiten abklingt. Falls $\beta \ll \alpha < 0$ gilt, ist $1 - h\beta \gg 1 - h\alpha > 0$, der von β abhängige Anteil der Lösung wird also auch wesentlich schneller als der von α abhängige abklingen, die numerisch bestimmte Näherungslösung verhält sich also zumindest in dieser Hinsicht wie die exakte Lösung.

Insbesondere genügt es in dieser Situation, die Schrittweite h so zu wählen, dass

$$\frac{1}{1-\alpha h} \approx e^{\alpha h}$$

gilt, dass wir also die entscheidende, langsamer abklingende Komponente gut approximieren können, denn die schneller abklingende wird ohnehin für den längerfristigen Verlauf der Lösung keine Rolle spielen.

Zur näheren Analyse dieses Phänomens untersuchen wir wieder die in Lemma 3.28 eingeführte Stabilitätsfunktion: Wie schon bei der Untersuchung der maximalen Konsistenzordnung von expliziten Runge-Kutta-Verfahren beschränken wir uns auf das einfache Anfangswertproblem (3.23), das durch

$$y_\lambda(0) = 1, \qquad y_\lambda'(t) = \lambda y_\lambda(t) \qquad \text{für alle } t \in \mathbb{R}_{\geq 0}$$

für ein $\lambda \in \mathbb{C}$ gegeben ist.

Definition 4.7 (Stabilitätsfunktion) *Sei ein Näherungsverfahren für das Anfangswertproblem (3.23) gegeben, und sei $\tilde{y}_\lambda$ die von ihm berechnete Näherungslösung. Falls eine Funktion $g : \mathbb{C} \to \mathbb{C}$ existiert, die*

$$\tilde{y}_\lambda(t_{j+1}) = g(\lambda h)\tilde{y}_\lambda(t_j) \qquad \textit{für alle } h \in \mathbb{R}_{>0},\ \lambda \in \mathbb{C},\ j \in \mathbb{N}_0 \tag{4.11}$$

erfüllt, bezeichnen wir sie als Stabilitätsfunktion *des Verfahrens.*

Wie man an (4.9) leicht ablesen kann, besitzt das explizite Euler-Verfahren die Stabilitätsfunktion

$$g_{\mathrm{ex}}(z) = 1 + z, \qquad \text{für alle } z \in \mathbb{C},$$

während wir aus (4.10) folgern können, dass das implizite Euler-Verfahren die Stabilitätsfunktion

$$g_{\mathrm{im}}(z) = \frac{1}{1-z} \qquad \text{für alle } z \in \mathbb{C}$$

besitzt. Wir haben bereits in Lemma 3.29 gesehen, dass die Konsistenzordnung damit zusammenhängt, wie gut die Stabilitätsfunktion die Exponentialfunktion in $z = 0$ approximiert. Man kann die Stabilitätsfunktion aber auch verwenden, um zu charakterisieren, für welche Schrittweiten die Näherungslösung abklingen wird: Durch Induktion folgt aus (4.11) direkt

$$\tilde{y}_\lambda(t_j) = g(\lambda h)^j \qquad \text{für alle } j \in \mathbb{N}_0,$$

wir können also nur dann ein Abklingen erwarten, wenn $|g(\lambda h)| < 1$ gilt.

Definition 4.8 (Stabilitätsgebiet) *Die Menge*

$$S_g := \{z \in \mathbb{C} \ : \ |g(z)| < 1\}$$

heißt Stabilitätsgebiet *zu der Stabilitätsfunktion g.*

Das Näherungsverfahren wird also zu einem sinnvollen asymptotischen Verhalten der Lösung führen, wenn wir $\lambda h \in S_g$ sicherstellen können.

Für das explizite Euler-Verfahren erhalten wir

$$S_{g_{\text{ex}}} = \{|1 + z| < 1 \ : \ z \in \mathbb{C}\} = K(-1, 1),$$

das Stabilitätsgebiet ist also eine offene Kreisscheibe um -1, und in unserem Fall ist -2 der Punkt, an dem die reelle Achse in das Stabilitätsgebiet eintritt, wir müssen also $-2 < h\lambda < 0$ sicherstellen. Das entspricht dem Kriterium, dass wir für unser Modellproblem bewiesen haben.

Für das implizite Euler-Verfahren finden wir

$$S_{g_{\text{im}}} = \{1/|1 - z| < 1 \ : \ z \in \mathbb{C}\} = \{|1 - z| > 1 \ : \ z \in \mathbb{C}\} = \mathbb{C} \setminus \overline{K(1, 1)},$$

das Stabilitätsgebiet ist also die gesamte komplexe Ebene mit Ausnahme einer abgeschlossenen Kreisscheibe um 1. In unserem Modellproblem ist $h\lambda$ für alle Schrittweiten h negativ, also immer im Stabilitätsgebiet enthalten. Diese Eigenschaft ist natürlich besonders nützlich.

Definition 4.9 (A-Stabilität) *Ein Näherungsverfahren mit*

$$\{z \in \mathbb{C} \ : \ \operatorname{Re} z < 0\} \subseteq S_g$$

heißt A-stabil.

Bei einem A-stabilen Verfahren dürfen wir also erwarten, dass es sich besonders gut für steife Anfangswertprobleme eignet. Offensichtlich ist das implizite Euler-Verfahren A-stabil, während das explizite Euler-Verfahren es nicht ist.

Wir haben bereits gesehen, dass bei expliziten Runge-Kutta-Verfahren die Stabilitätsfunktion g ein Polynom ist, und da für alle nicht-konstanten Polynome

$$\lim_{z \to -\infty} |g(z)| = \infty$$

gilt, folgt sofort, dass derartige Verfahren niemals A-stabil sein können.

Eine Chance auf A-Stabilität haben wir also nur dann, wenn wir Verfahren mit nichtpolynomialer Stabilitätsfunktion untersuchen. Im Falle des impliziten Euler-Verfahrens beispielsweise ist die Stabilitätsfunktion rational.

Wir untersuchen allgemeine Runge-Kutta-Verfahren, die durch ein Gleichungssystem der Form

$$k_i = f\left(t + c_i h, x + h\sum_{j=1}^{s} a_{ij} k_j\right) \qquad \text{für alle } i \in \{1,\dots,s\}$$

beschrieben werden. Für unser Modellproblem (3.23) erhalten wir daraus

$$k_i = \lambda x + \lambda h \sum_{j=1}^{s} a_{ij} k_j, \qquad k_i - \lambda h \sum_{j=1}^{s} a_{ij} k_j = \lambda x \qquad \text{für alle } i \in \{1,\dots,s\},$$

und wenn wir die Zwischenergebnisse k_i in einem Vektor $\mathbf{k} \in \mathbb{R}^s$ zusammenfassen und den konstanten Vektor $\mathbf{1} := (1)_{i=1}^s$ einführen, können wir diese Gleichungen kompakt in der Form

$$(\mathbf{I} - \lambda h \mathbf{A})\mathbf{k} = \lambda \mathbf{1} x, \qquad \mathbf{k} = (\mathbf{I} - \lambda h \mathbf{A})^{-1} \mathbf{1} \lambda x$$

darstellen, sofern die Matrix invertierbar (und damit das Verfahren überhaupt durchführbar) ist. Die Verfahrensfunktion ist durch

$$\Phi(t,h,x) = \sum_{i=1}^{s} b_i k_i = \langle \mathbf{b}, \mathbf{k} \rangle_2 = \langle \mathbf{b}, (\mathbf{I} - \lambda h \mathbf{A})^{-1} \mathbf{1} \rangle_2 \lambda x$$

gegeben, die nächste Iterierte durch

$$\begin{aligned} \tilde{y}(t_{j+1}) &= \tilde{y}(t_j) + h\Phi(t,h,\tilde{y}(t_j)) = \tilde{y}(t_j) + \langle \mathbf{b}, (\mathbf{I} - \lambda h \mathbf{A})^{-1} \mathbf{1} \rangle_2 \lambda h \tilde{y}(t_j) \\ &= (1 + \langle \mathbf{b}, (\mathbf{I} - \lambda h \mathbf{A})^{-1} \mathbf{1} \rangle_2 \lambda h) \tilde{y}(t_j), \end{aligned}$$

also muss die Stabilitätsfunktion gerade durch

$$g(z) = 1 + \langle \mathbf{b}, (\mathbf{I} - z\mathbf{A})^{-1} \mathbf{1} \rangle_2 z \qquad \text{für alle } z \in \mathbb{C} \tag{4.12}$$

gegeben sein. Diese Funktion ist rational, und ihre Singularitäten sind gerade die Kehrwerte der Eigenwerte der Matrix $\mathbf{A}$.

Im Falle eines expliziten Verfahrens ist $\mathbf{I} - z\mathbf{A}$ eine untere Dreiecksmatrix, bei der alle Diagonaleinträge gleich eins sind, also immer invertierbar. Durch Vorwärtseinsetzen können wir direkt das Resultat aus Lemma 3.28 gewinnen.

Interessanter ist natürlich die Anwendung auf implizite Verfahren. Als Beispiel untersuchen wir das auf der Trapezregel basierende implizite Runge-Kutta-Verfahren, das durch das Butcher-Schema

0	0	0
1	1/2	1/2
	1/2	1/2

definiert ist. Einsetzen in (4.12) führt zu

$$\begin{aligned} g_{\mathrm{tr}}(z) &= 1 + \begin{pmatrix} 1/2 & 1/2 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -z/2 & 1 - z/2 \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ 1 \end{pmatrix} z \\ &= 1 + \begin{pmatrix} 1/2 & 1/2 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ \frac{z/2}{1-z/2} & \frac{1}{1-z/2} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} z \\ &= 1 + \begin{pmatrix} 1/2 & 1/2 \end{pmatrix} \begin{pmatrix} 1 \\ \frac{1+z/2}{1-z/2} \end{pmatrix} z = 1 + \frac{1/2 - z/4 + 1/2 + z/4}{1 - z/2} z \\ &= 1 + \frac{z}{1 - z/2} = \frac{1 + z/2}{1 - z/2}. \end{aligned}$$

Um das Stabilitätsgebiet zu bestimmen, müssen wir diejenigen $z \in \mathbb{C}$ finden, für die $|g_{\mathrm{tr}}(z)| < 1$ gilt. Wir wählen ein $z \in \mathbb{C}$ und stellen es durch seinen Realteil $z_r \in \mathbb{R}$ und seinen Imaginärteil $z_i \in \mathbb{R}$ dar, also durch $z = z_r + iz_i$. Es gilt

$$\begin{aligned} |g_{\mathrm{tr}}(z)| < 1 &\iff |1 + z/2| < |1 - z/2| \iff |2 + z| < |2 - z| \\ &\iff |2 + z|^2 < |2 - z|^2 \iff (2 + z_r)^2 + z_i^2 < (2 - z_r)^2 + z_i^2 \\ &\iff 4 + 4z_r + z_r^2 < 4 - 4z_r + z_r^2 \iff 4z_r < -4z_r \iff 8z_r < 0, \end{aligned}$$

also ist $|g_{\mathrm{tr}}(z)| < 1$ äquivalent zu $z_r < 0$ und das Stabilitätsgebiet ist gegeben durch

$$S_{g_{\mathrm{tr}}} = \{z \in \mathbb{C} \ : \ \mathrm{Re}\, z < 0\}.$$

Wir sehen, dass das Stabilitätsgebiet der impliziten Trapezregel deutlich kleiner als das des impliziten Euler-Verfahrens ist, dass es aber immer noch die linke komplexe Halbebene enthält, so dass auch die implizite Trapezregel A-stabil ist. Allerdings kann man unserer Rechnung auch entnehmen, dass g_{tr} für $z \to -\infty$ nicht gegen null, sondern gegen eins konvergieren wird, so dass für zu große Schrittweiten nicht das Verhalten der exakten Lösung reproduziert wird.

4.4 Differential-algebraische Gleichungen

Die Behandlung steifer Differentialgleichungen wird dadurch erschwert, dass sich unterschiedliche Komponenten der Lösung unterschiedlich verhalten. Noch komplizierter wird die Situation, wenn einzelne Komponenten gar nicht mehr über eine Differentialgleichung beschrieben werden, andere jedoch schon.

Als Beispiel untersuchen wir das *mathematische Pendel* (vgl. [11]): Wir gehen davon aus, dass das Pendel im Nullpunkt an einem nicht dehnbaren Faden der Länge $L \in \mathbb{R}_{>0}$ aufgehängt ist und dass seine Position zum Zeitpunkt $t \in \mathbb{R}$ durch einen Vektor $x(t) \in \mathbb{R}^2$ gegeben ist. Die Position genügt nicht, um die physikalischen Zusammenhänge zu beschreiben, wir benötigen außerdem die Geschwindigkeit $v(t) \in \mathbb{R}^2$.

Nach Newton [8] ist die Ableitung der Position gerade die Geschwindigkeit, es gilt also

$$x'(t) = v(t) \qquad \text{für alle } t \in \mathbb{R}.$$

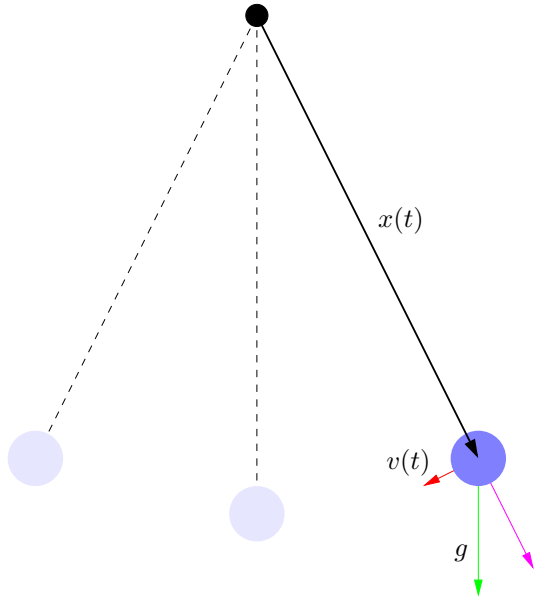

Abbildung 4.1: Mathematisches Pendel

Die Newton-Axiome besagen auch, dass die Ableitung der Geschwindigkeit die Beschleunigung ist, die wiederum durch eine Kraft bewirkt wird. In unserem Fall wirken zwei Kräfte: Einerseits die Gravitation, die die Masse des Pendels nach unten zieht, andererseits die „Rückstellkraft“ des Fadens, die dafür sorgt, dass sich der Abstand der Masse zum Nullpunkt nicht ändert. Diese Kraft wirkt immer in Richtung des Nullpunkts, also gerade in der Richtung $x(t)$. Ihre Stärke $\lambda(t)$ hängt davon ab, wie stark die Gravitationskraft ist, die gerade auf die Masse wirkt. Wir erhalten die Gleichung

$$v'(t) = \lambda(t)x(t) - \begin{pmatrix} 0 \\ g \end{pmatrix} \qquad \text{für alle } t \in \mathbb{R},$$

wobei g die Stärke der Gravitation angibt. Schließlich müssen wir noch eine Gleichung aufnehmen, mit der sich die soeben eingeführte Variable λ bestimmen lässt, und an dieser Stelle weichen wir von der Form einer gewöhnlichen Differentialgleichung ab: λ ist implizit dadurch bestimmt, dass der Abstand der Masse vom Nullpunkt konstant bleiben muss, dass also

$$\|x(t)\|_2 = L \qquad \text{für alle } t \in \mathbb{R} \tag{4.13}$$

gelten soll. Wie man sieht taucht λ in dieser Gleichung überhaupt nicht auf. Insgesamt erhalten wir also das System

$$x'(t) = v(t), \qquad v'(t) = \lambda(t)x(t) - \begin{pmatrix} 0 \\ g \end{pmatrix}, \qquad \|x(t)\|_2^2 = L \qquad \text{für alle } t \in \mathbb{R}, \tag{4.14}$$

und dieses System ist keine gewöhnliche Differentialgleichung, sondern beinhaltet eine algebraische Nebenbedingung in Gestalt der dritten Gleichung (4.13). Derartige Systeme bezeichnet man als *differential-algebraische Gleichungen*, häufig auch als *DAE* (aus dem Englischen: *differential algebraic equation*). Dabei ist es in der Regel nicht wichtig, dass die dritte Bedingung tatsächlich eine algebraische Gleichung ist, wichtig ist lediglich, dass es sich nicht um eine Differentialgleichung handelt, die wir direkt behandeln können.

Glücklicherweise lässt sich das System auf eine gewöhnliche Differentialgleichung zurückführen: Wir differenzieren die dritte Gleichung und erhalten

$$L = \|x(t)\|_2^2 = x_1^2(t) + x_2^2(t), \qquad 0 = 2x_1(t)x_1'(t) + 2x_2(t)x_2'(t) = 2\langle x(t), x'(t)\rangle_2.$$

Aus der ersten Gleichung folgt

$$0 = 2\langle x(t), v(t)\rangle_2 \qquad \text{für alle } t \in \mathbb{R}, \tag{4.15}$$

und diese Gleichung besagt, dass die Geschwindigkeit immer senkrecht auf dem Positionsvektor stehen muss. Wir differenzieren erneut und erhalten

$$0 = 2\langle x'(t), v(t)\rangle_2 + 2\langle x(t), v'(t)\rangle_2 = 2\|v(t)\|_2^2 + 2\langle x(t), v'(t)\rangle_2.$$

Nun können wir die zweite Gleichung einsetzen, um

$$0 = 2\|v(t)\|_2^2 + 2\lambda(t)\|x(t)\|_2^2 - 2gx_2(t) \qquad \text{für alle } t \in \mathbb{R}$$

zu erhalten. Dank der dritten Gleichung folgt

$$0 = 2\|v(t)\|_2^2 + 2\lambda(t)L^2 - 2gx_2(t) \qquad \text{für alle } t \in \mathbb{R},$$

die wir umstellen können, um einen Ausdruck für $\lambda(t)$ zu erhalten:

$$\lambda(t)L^2 = gx_2(t) - \|v(t)\|_2^2 \qquad \text{für alle } t \in \mathbb{R}.$$

Wir differenzieren ein drittes Mal und erhalten

$$\lambda'(t)L^2 = gx_2'(t) - 2\langle v(t), v'(t)\rangle_2 = gv_2(t) - 2\lambda(t)\langle v(t), x(t)\rangle_2 + 2gv_2(t),$$

und dank (4.15) fällt der zweite Term weg, so dass nur

$$\lambda'(t)L^2 = 3gv_2(t) \qquad \text{für alle } t \in \mathbb{R}$$

übrig bleibt. Damit haben wir die gesuchte Differentialgleichung für die letzte Variable λ gefunden und können das Gesamtsystem in der Form

$$x'(t) = v(t), \quad v'(t) = \lambda(t)x(t) - \begin{pmatrix} 0 \\ g \end{pmatrix}, \quad \lambda'(t) = \frac{3g}{L^2}v_2(t) \qquad \text{für alle } t \in \mathbb{R} \tag{4.16}$$

schreiben. Indem wir die Variablen zusammenfassen, erhalten wir

$$y(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \\ v_1(t) \\ v_2(t) \\ \lambda(t) \end{pmatrix}, \qquad y'(t) = \begin{pmatrix} v_1(t) \\ v_2(t) \\ \lambda(t)x_1(t) \\ \lambda(t)x_2(t) - g \\ 3gv_2(t)/L^2 \end{pmatrix} \qquad \text{für alle } t \in \mathbb{R}, \tag{4.17}$$

so dass wir die bisher entwickelten numerischen Verfahren zur Lösung der Differentialgleichung einsetzen können. Voraussetzung dabei ist natürlich, dass uns ein Anfangswert $y(0)$ zur Verfügung steht, der die algebraische Bedingung (4.13) erfüllt.

Bemerkung 4.10 (Differentiationsindex) *In unserem Beispiel war es erforderlich, die Gleichungen dreimal zu differenzieren, um die Gleichung (4.14) in die Form (4.16) einer gewöhnlichen Differentialgleichung zu bringen.*

Wenn sich eine differential-algebraische Gleichung durch m-maliges Differenzieren auf eine gewöhnliche Differentialgleichung reduzieren lässt, bezeichnet man m als den Differentiationsindex *(oder kurz* Index*) der Gleichung.*

Da das umgeformte System (4.16) bereits als gewöhnliche Differentialgleichung gegeben ist, hat es den Index null. Unsere Umformung hat also zu einer Indexreduktion *geführt.*

In allgemeinen Anwendungen ist es gelegentlich erforderlich, eine Indexreduktion mit Hilfe von Computer-Algebra-Systemen automatisiert durchführen zu lassen.

Für die Darstellung differential-algebraischer Gleichungen sind verschiedene Normalformen üblich. Der allgemeinste Fall ist die *vollständig implizite* Darstellung durch eine Funktion $F : [a, b] \times V \times V \to V$, bei der die Lösung $y \in C^1([a, b], V)$ die Gleichungen

$$y(a) = y_0, \qquad F(t, y(t), y'(t)) = 0 \qquad \text{für alle } t \in [a, b]$$

erfüllen muss. Im Fall des mathematischen Pendels könnten wir beispielsweise die Gleichungen (4.14) durch die Funktion

$$F(t, y(t), y'(t)) = \begin{pmatrix} y_1'(t) - y_3(t) \\ y_2'(t) - y_4(t) \\ y_3'(t) - y_5(t)y_1(t) \\ y_4'(t) - y_5(t)y_2(t) + g \\ y_1^2(t) + y_2^2(t) - L^2 \end{pmatrix} = 0$$

ausdrücken, wobei wir die Komponenten von y wie in (4.17) verwenden. Diese allgemeinste Form ist auch die am schwierigsten handzuhabende.

Glücklicherweise sind in unserem Fall zwei der drei Gleichungen des Systems (4.14) bereits differentiell, so dass wir zwischen den „differentiellen Unbekannten“ x und v und der „algebraischen Unbekannten“ λ unterscheiden können. Wir fassen sie in den Variablen

$$y(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \\ v_1(t) \\ v_2(t) \end{pmatrix}, \qquad z(t) = \lambda(t) \qquad \text{für alle } t \in \mathbb{R}$$

zusammen und erhalten die *semi-explizite Darstellung*

$$y'(t) = f(t, y(t), z(t)), \qquad 0 = h(t, y(t), z(t)) \qquad \text{für alle } t \in \mathbb{R}, \tag{4.18}$$

wobei wir

$$f(t, y(t), z(t)) = \begin{pmatrix} y_3(t) \\ y_4(t) \\ z(t)y_1(t) \\ z(t)y_2(t) - g \end{pmatrix}, \qquad h(t, y(t), z(t)) = y_1^2(t) + y_2^2(t) - L^2$$

verwenden. Diese Darstellung ermöglicht es uns, die differential-algebraische Gleichung als gewöhnliche Differentialgleichung auf der *Nullstellenmenge* der Funktion h zu interpretieren.

Dafür bräuchte man natürlich eine Verallgemeinerung des Begriffs der Ableitung, der auch auf solchen Mengen noch benutzt werden kann. Die nötigen Werkzeuge stellt die Theorie die *differenzierbaren Mannigfaltigkeiten* zur Verfügung: Unter gewissen Bedingungen ist die Nullstellenmenge eine solche Mannigfaltigkeit, und es lassen sich geeignete Ableitungen und damit auch Differentialgleichungen innerhalb der Mannigfaltigkeit definieren.

In unserem Fall ist die Situation besonders einfach: Die Position des Pendels muss auf dem Kreis mit Radius L liegen, und diesen Kreis können wir durch die *Parametrisierung*

$$\gamma : \mathbb{R} \to \mathbb{R}^2, \qquad \varphi \mapsto \begin{pmatrix} L\sin(\varphi) \\ -L\cos(\varphi) \end{pmatrix},$$

darstellen. Mit dem Ansatz $x(t) = \gamma(\varphi(t))$ nehmen unsere Gleichungen die Form

$$\begin{aligned} v(t) = x'(t) = (\gamma \circ \varphi)'(t) = \gamma'(\varphi(t))\varphi'(t) &= \begin{pmatrix} L\cos(\varphi(t))\varphi'(t) \\ L\sin(\varphi(t))\varphi'(t) \end{pmatrix}, \\ v'(t) = \lambda(t)\gamma(\varphi(t)) - \begin{pmatrix} 0 \\ g \end{pmatrix} &= \begin{pmatrix} \lambda(t)L\sin(\varphi(t)) \\ -\lambda(t)L\cos(\varphi(t)) - g \end{pmatrix}, \end{aligned}$$

an, so dass wir durch Einsetzen der ersten Gleichung in die zweite die Gleichung

$$\begin{pmatrix} -L\sin(\varphi(t))(\varphi'(t))^2 + L\cos(\varphi(t))\varphi''(t) \\ L\cos(\varphi(t))(\varphi'(t))^2 + L\sin(\varphi(t))\varphi''(t) \end{pmatrix} = \begin{pmatrix} \lambda(t)L\sin(\varphi(t)) \\ -\lambda(t)L\cos(\varphi(t)) - g \end{pmatrix}$$

erhalten. Wir multiplizieren die erste Zeile mit $\cos(\varphi(t))$, die zweite mit $\sin(\varphi(t))$, und addieren beide, um die Gleichung

$$\begin{aligned} -g\sin(\varphi(t)) &= -L\sin(\varphi(t))\cos(\varphi(t))(\varphi'(t))^2 + L\cos^2(\varphi(t))\varphi''(t) \\ &\quad + L\sin(\varphi(t))\cos(\varphi(t))(\varphi'(t))^2 + L\sin^2(\varphi(t))\varphi''(t) \\ &= L(\cos^2(\varphi(t)) + \sin^2(\varphi(t)))\varphi''(t) \\ &= L\varphi''(t) \end{aligned}$$

zu erhalten, also eine gewöhnliche Differentialgleichung, mit der wir φ, also auch die interessanten Größen x und v, bestimmen können.

In allgemeinen Anwendungen ist die Situation leider häufig wesentlich komplizierter, da sich die Nullstellenmenge der Funktion h nicht direkt an den Gleichungen ablesen und erst recht nicht einfach durch eine Parametrisierung darstellen lässt. In derartigen Fällen können, ähnlich wie im Fall steifer Differentialgleichungen, immerhin noch implizite Zeitschrittverfahren zum Einsatz kommen. Wenn wir beispielsweise das semi-explizite System (4.18) lösen wollen, können wir mit Hilfe des impliziten Euler-Verfahrens den differentiellen Anteil durch

$$y(t+h) - y(t) \approx hf(t, y(t+h), z(t+h))$$

approximieren und dann eine Näherung $(\tilde{y}(t+h), \tilde{z}(t+h))$ als Lösung des Systems

$$\tilde{y}(t+h) = y(t) + hf(t, \tilde{y}(t+h), \tilde{z}(t+h)), \qquad 0 = g(t, \tilde{y}(t+h), \tilde{z}(t+h))$$

suchen. Durch die Approximation haben wir das differential-algebraische System auf ein potentiell nichtlineares System reduziert, das sich hoffentlich mit Techniken wie der Newton-Iteration behandeln lassen wird.

5 Beispiele für partielle Differentialgleichungen

Während bei den bisher betrachteten gewöhnlichen Differentialgleichungen nur Ableitungen nach einer Variablen eine Rolle spielten und die Lösungen auf einem eindimensionalen Definitionsbereich gesucht wurden, sind bei partiellen Differentialgleichungen höherdimensionale Definitionsbereiche zugelassen, und deshalb auch partielle Ableitungen nach den einzelnen Koordinatenrichtungen.

Im Vergleich zu gewöhnlichen Differentialgleichungen treten dadurch zusätzliche Schwierigkeiten auf:

- Wenn die Lösung von mehreren Variablen abhängt, lässt sie sich nicht mehr durch eine einfache Integralgleichung der Form (2.2) beschreiben, so dass Existenzaussagen wie die des Satzes 2.3 von Picard-Lindeløf nicht mehr gelten.

- Der Definitionsbereich kann nicht mehr ein einfaches Intervall sein, schon im zweidimensionalen Fall sind wesentlich kompliziertere Formen möglich. Dadurch wird es im Allgemeinen schwierig, Randwerte für die Lösung festzulegen.

Aufgrund dieser Schwierigkeiten gibt es bis heute keine Theorie, mit der sich alle Arten partieller Differentialgleichungen einheitlich behandeln lassen, stattdessen gibt es angepasste Techniken für bestimmte Klassen von Gleichungen. Besonders wichtige Typen sind

- hyperbolische Differentialgleichungen, mit denen sich beispielsweise die Erhaltung physikalischer Größen wie der Masse oder der Energie beschreiben lassen und die in der Strömungsmechanik eine wichtige Rolle spielen,

- elliptische Differentialgleichungen, die unter anderem bei der Modellierung von Phänomenen aus der Elektrodynamik oder auch der Strukturmechanik Anwendung finden, und

- parabolische Differentialgleichungen, die vor allem für zeitabhängige Diffusionsprozesse verwendet werden, etwa für die Beschreibung der Wärmeausbreitung in Materialien.

In diesem Kapitel werden wir für jede Kategorie ein Beispiel untersuchen und jeweils ein einfaches numerisches Lösungsverfahren diskutieren.

5.1 Hyperbolische Gleichungen und das Verfahren der Charakteristiken

Bei der Beschreibung vieler physikalischer Vorgänge spielt die Erhaltung gewisser physikalischer Größen eine Rolle. Beispielsweise sollten in einem geschlossenen System keine Masse oder keine Energie verloren gehen.

Als ein einfaches Beispiel untersuchen wir die Massenerhaltung: Wir untersuchen die *Dichte* eines Materials, das sich in dem Intervall $[0,1]$ verteilt. Die Dichte zu einem Zeitpunkt $t \in \mathbb{R}_{\geq 0}$ in einem Punkt $x \in [0,1]$ bezeichnen wir mit $u(t,x)$, so dass sich der Zustand unseres Systems durch eine Funktion $u \in C^1(\mathbb{R}_{\geq 0} \times [0,1])$ beschreiben lässt.

Damit das System interessant ist, sollten wir zulassen, dass sich die Dichte verändert. Dieser Vorgang wird durch eine *Flussfunktion* $f \in C^1(\mathbb{R} \times [0,1] \times \mathbb{R})$ modelliert, die über die Gleichung

$$\frac{\partial u}{\partial t}(t,x) = -\frac{\partial}{\partial x} f(t,x,u(t,x)) \qquad \text{für alle } t \in \mathbb{R}_{\geq 0},\ x \in [0,1] \tag{5.1}$$

beschreibt, wie sich die räumliche Verteilung der Dichte im Laufe der Zeit verändert.

Der Ausgangspunkt unserer Betrachtung war die Massenerhaltung, also empfiehlt es sich, nachzuprüfen, in welcher Beziehung die Gleichung (5.1) zu diesem übergeordneten Konzept steht. Die in einem Teilintervall $[a,b]$ vorhandene Masse ist durch das Integral der Dichte definiert, also durch

$$m_{a,b}(t) := \int_a^b u(t,x)\,dx \qquad \text{für alle } t \in \mathbb{R}_{\geq 0}.$$

Die zeitliche Veränderung der Masse kann mit Hilfe der Gleichung (5.1) durch

$$\begin{aligned} m'_{a,b}(t) &= \frac{\partial}{\partial t} \int_a^b u(t,x)\,dx = \int_a^b \frac{\partial u}{\partial t}(t,x)\,dx \\ &= -\int_a^b \frac{\partial}{\partial x} f(t,x,u(t,x))\,dx = f(t,a,u(t,a)) - f(t,b,u(t,b)) \end{aligned} \tag{5.2}$$

ausgedrückt werden, und mit dieser Gleichung lässt sich die Bedeutung der Flussfunktion besser verstehen: Die Flussfunktion f beschreibt, wie angesichts ihres Namens nicht anders zu erwarten, wie die Masse fließt. Dabei bedeutet ein positiver Wert eine Fluss von links nach rechts und ein negativer einen Fluss in der entgegengesetzten Richtung.

Die Gleichung (5.2) besagt dann einfach, dass die Veränderung der Masse in dem Intervall $[a,b]$ sich als Differenz aus dem Zufluss am linken Rand und dem Abfluss am rechten Rand ergibt, entspricht also der anschaulichen Vorstellung. Aus diesem Grund bezeichnet man Differentialgleichungen der Form (5.1) als *Erhaltungsgleichungen*.

Wenn wir sicher stellen wollen, dass in unserem Intervall keine Masse verloren geht oder hinzu kommt, müssen wir also

$$f(t,0,u(t,0)) = f(t,1,u(t,1)) \qquad \text{für alle } t \in \mathbb{R}_{\geq 0}$$

sicherstellen. Besonders einfache Möglichkeiten sind etwa

$$f(t,0,z) = f(t,1,z) \qquad \text{für alle } t \in \mathbb{R}_{\geq 0},\ z \in \mathbb{R},$$

oder

$$f(t,0,z) = f(t,1,z) = 0 \qquad \text{für alle } t \in \mathbb{R}_{\geq 0},\ z \in \mathbb{R}.$$

Die erste Bedingung legt fest, dass Zu- und Abfluss gerade gleich groß sind, die zweite ist restriktiver und verbietet jeden Zu- und Abfluss: Der linke und rechte Rand des Intervalls sind undurchlässig.

Um die Gleichung (5.1) numerisch behandeln zu können, bietet es sich an, sie auf eine einfachere Form zu bringen. Insbesondere das implizite Auftreten von u auf der rechten Seite der Gleichung bereitet Schwierigkeiten, die wir allerdings relativ einfach lösen können, indem wir die Ableitung berechnen: Nach Kettenregel gilt

$$\frac{\partial}{\partial x} f(t,x,u(t,x)) = \frac{\partial f}{\partial x}(t,x,u(t,x)) + \frac{\partial f}{\partial u}(t,x,u(t,x))\frac{\partial u}{\partial x}(t,x)$$
$$\text{für alle } t \in \mathbb{R}_{\geq 0},\ x \in [0,1],$$

also lässt sich die Gleichung (5.1) auch in der Form

$$0 = \frac{\partial u}{\partial t}(t,x) + \frac{\partial f}{\partial u}(t,x,u(t,x))\frac{\partial u}{\partial x}(t,x) + \frac{\partial f}{\partial x}(t,x,u(t,x))$$
$$\text{für alle } t \in \mathbb{R}_{\geq 0},\ x \in [0,1]$$

darstellen. Da diese Gleichung von den Ableitungen von u nur noch linear abhängt, bezeichnet man sie als *quasilineare* Form der Erhaltungsgleichung.

Im Allgemeinen schreiben wir quasilineare Erhaltungsgleichungen in der Form

$$c(t,x,u(t,x)) = a(t,x,u(t,x))\frac{\partial u}{\partial t}(t,x) + b(t,x,u(t,x))\frac{\partial u}{\partial x}(t,x) \tag{5.3}$$
$$\text{für alle } t \in \mathbb{R}_{\geq 0},\ x \in [0,1],$$

und unsere Aufgabe ist es nun, derartige Gleichungen zu lösen.

Die Idee des *Verfahrens der Charakteristiken* besteht darin, die „festen“ Zeit- und Ortskoordinaten t und x durch bewegliche Koordinaten zu ersetzen, die der Bewegung der fließenden Masse angepasst ist. Die Bewegung (sowohl von Zeit- als auch Ortskoordinate) beschreiben wir durch eine Funktion

$$\gamma : [0,\beta] \to \mathbb{R}_{\geq 0} \times [0,1].$$

Aus der Kettenregel folgt

$$(u \circ \gamma)'(\tau) = \frac{\partial u}{\partial t}(\gamma(\tau))\gamma_1'(\tau) + \frac{\partial u}{\partial x}(\gamma(\tau))\gamma_2'(\tau) \qquad \text{für alle } \tau \in [0,\beta].$$

Wir vergleichen mit den Termen der Gleichung (5.3) und stellen Ähnlichkeiten fest: $\gamma_1'(\tau)$ nimmt den Platz von $a(t, x, u(t, x))$ ein, $\gamma_2'(\tau)$ den von $b(t, x, u(t, x))$, und $(u \circ \gamma)'(\tau)$ den von $c(t, x, u(t, x))$. Das bringt uns auf die Idee, die Funktionen γ_1, γ_2 und $u \circ \gamma$ zu einer vektorwertigen Funktion

$$y : [0, \beta] \to \mathbb{R}_{\geq 0} \times [0, 1] \times \mathbb{R}, \qquad \tau \mapsto \begin{pmatrix} \gamma_1(\tau) \\ \gamma_2(\tau) \\ u \circ \gamma(\tau) \end{pmatrix},$$

zusammenzufassen und eine gewöhnliche Differentialgleichung zu formulieren, die diese Funktion beschreibt:

$$y'(\tau) = \begin{pmatrix} \gamma_1'(\tau) \\ \gamma_2'(\tau) \\ (u \circ \gamma)'(\tau) \end{pmatrix} \stackrel{!}{=} \begin{pmatrix} a(y(\tau)) \\ b(y(\tau)) \\ c(y(\tau)) \end{pmatrix} \qquad \text{für alle } \tau \in [0, \beta]. \tag{5.4}$$

Falls die Funktionen a, b und c die entsprechenden Voraussetzungen erfüllen und falls geeignete Startwerte vorliegen, erhalten wir ein Anfangswertproblem, das wir mit den bereits behandelten Techniken bearbeiten können.

Die durch γ beschriebenen Kurven heißen *Charakteristiken* der Differentialgleichung, und wie wir gesehen haben, lässt sich mit ihrer Hilfe die Behandlung zumindest von Gleichungen des Typs (5.3) auf das Lösen gewöhnlicher Differentialgleichungen zurückführen.

Ein Nachteil des Verfahrens der Charakteristiken ist, dass wir im Allgemeinen nicht die Funktion u in einem beliebigen Punkt (t, x) auswerten können: Das Verfahren berechnet die Werte der Lösung nur an den Punkten auf der Charakteristik, und falls wir keine Charakteristik finden können, die den Punkt (t, x) trifft, können wir auch den Wert der Lösung in diesem Punkt nicht berechnen.

Beispiel 5.1 (Transportgleichung) *Ein besonders einfaches Beispiel ist die* Transportgleichung

$$0 = \frac{\partial u}{\partial t}(t, x) + \varrho \frac{\partial u}{\partial x}(t, x) \qquad \textit{für alle } t \in \mathbb{R}_{\geq 0},\ x \in \mathbb{R}.$$

Diese Gleichung ist von der Form (5.3) mit den Koeffizientenfunktionen $a = 1$, $b = \varrho$ und $c = 0$, so dass wir

$$y'(\tau) = \begin{pmatrix} 1 \\ \varrho \\ 0 \end{pmatrix} \qquad \textit{für alle } \tau \in \mathbb{R}_{\geq 0}$$

lösen müssen. Diese Aufgabe ist auch ohne Rechner in den Griff zu bekommen: Es gilt

$$y(\tau) = \begin{pmatrix} y_1(0) + \tau \\ y_2(0) + \varrho\tau \\ y_3(0) \end{pmatrix} \qquad \textit{für alle } \tau \in \mathbb{R}_{\geq 0}.$$

Die erste Komponente der Lösung ist die Zeit, hier können wir also durch Wahl des Anfangswerts $y_1(0) = 0$ dafür sorgen, dass τ und t übereinstimmen. Die zweite Komponente ist der Ort, hier können wir also ein $x_0 \in \mathbb{R}$ vorgeben und $y_2(0) = x_0$ setzen. Die dritte Komponente ist der Wert der Funktion u im Punkt $\gamma(0) = (y_1(0), y_2(0)) = (0, x_0)$, also gerade der Anfangswert von u im Punkt x_0.

Wenn wir nun $u(t, x)$ berechnen wollen, können wir den Anfangsort x_0 so wählen, dass $x_0 + \varrho t = x$ gilt, also als $x_0 = x - \varrho t$, denn dann folgt

$$\begin{aligned} u(t,x) &= u(t, x_0 + \varrho t) = u(y_1(\tau), y_2(\tau)) = y_3(\tau) \\ &= y_3(0) = u(y_1(0), y_2(0)) = u(0, x_0) = u(0, x - \varrho t). \end{aligned}$$

Wir können damit die Lösung der Gleichung in jedem beliebigen Punkt auswerten.

In diesem Fall sind die Charakteristiken besonders einfach: Es sind die Linien

$$\{(t, x_0 + \varrho t) \ : \ t \in \mathbb{R}_{\geq 0}\} \qquad \text{für } x_0 \in \mathbb{R},$$

und dank $c = 0$ ist die Lösung entlang einer Charakteristik konstant.

Beispiel 5.2 (Allgemeinere Transportgleichung) *Eine etwas interessantere Situation tritt auf, falls a und b unabhängig von ihren dritten Parametern sind, falls also*

$$c(t, x, u(t, x)) = a(t, x)\frac{\partial u}{\partial t}(t, x) + b(t, x)\frac{\partial u}{\partial x}(t, x) \qquad \text{für alle } t \in \mathbb{R}_{\geq 0},\ x \in \mathbb{R}$$

gilt, denn in diesem Fall sind die ersten beiden Komponenten der Lösungsfunktion y von c und u unabhängig. Wir können also γ einmal als Lösung der gewöhnlichen Differentialgleichung

$$\gamma'(\tau) = \begin{pmatrix} a(\gamma(\tau)) \\ b(\gamma(\tau)) \end{pmatrix} \qquad \text{für alle } \tau \in [0, \beta]$$

bestimmen und dann für beliebige Funktionen c die dritte Komponente von y als Lösung der Gleichung

$$y_3'(\tau) = c(\gamma_1(\tau), \gamma_2(\tau), y_3(\tau)) \qquad \text{für alle } \tau \in [0, \beta]$$

berechnen. Also lassen sich Lösungen in beliebigen Punkten berechnen, indem wir entlang der bekannten Charakteristiken integrieren.

Das Verfahren der Charakteristiken lässt sich relativ einfach auf höhere Dimensionen übertragen: Wenn wir statt im zweidimensionalen Raum im d-dimensionalen arbeiten, treten d Summanden auf der rechten Seite der quasilinearen Erhaltungsgleichung (5.3) auf, und die Funktion γ bildet in den d-dimensionalen Raum ab.

5.2 Elliptische Differentialgleichungen und das Finite-Differenzen-Verfahren

Als Beispiel für eine elliptische Differentialgleichung untersuchen wir die *Potentialgleichung* (auch bekannt als die *Poisson-Gleichung*) der Elektrostatik. Auf dem Einheitsquadrat nimmt sie die Form

$$-\frac{\partial^2 u}{\partial x^2}(x,y) - \frac{\partial^2 u}{\partial y^2}(x,y) = f(x,y) \qquad \text{für alle } (x,y) \in \Omega := (0,1)^2$$

an. Sie beschreibt das elektrostatische Potential, das von einer Ladungsverteilung hervorgerufen wird. Die Funktion $f \in C(\Omega)$ gibt dabei die Ladungsdichte in allen Punkten des Rechengebiets an.

Man kann sich überlegen, dass es bei dieser Gleichung nicht genügt, nur entlang einer Linie Randbedingungen zu formulieren, so wie wir es im Fall der Erhaltungsgleichung getan haben, stattdessen müssen wir auf dem gesamten Rand

$$\partial\Omega := \{0,1\} \times [0,1] \cup [0,1] \times \{0,1\}$$

des Gebiets Bedingungen stellen. Besonders einfach ist die *Dirichlet-Randbedingung*

$$u(x,y) = 0 \qquad \text{für alle } (x,y) \in \partial\Omega.$$

In der physikalischen Interpretation beschreibt sie, dass der Rand des Gebiets supraleitend ist, so dass keine Potentialunterschiede auftreten können.

Die Potentialgleichung können wir kürzer schreiben, indem wir den *Laplace-Operator*

$$\Delta u(x,y) = \frac{\partial^2 u}{\partial x^2}(x,y) + \frac{\partial^2 u}{\partial y^2}(x,y) \qquad \text{für alle } (x,y) \in \Omega$$

definieren und das Gesamtproblem in der folgenden Form kompakt notieren:

> Wir suchen eine Funktion $u \in C(\bar{\Omega})$ mit $u|_\Omega \in C^2(\Omega)$, die
>
> $$-\Delta u(x,y) = f(x,y) \qquad \text{für alle } (x,y) \in \Omega, \tag{5.5a}$$
> $$u(x,y) = 0 \qquad \text{für alle } (x,y) \in \partial\Omega \tag{5.5b}$$
>
> erfüllt. Dabei ist $\bar{\Omega} = \Omega \cup \partial\Omega$ der Abschluss des Gebiets.

Wie wir gesehen haben, entwickelt sich die Lösung bei einer hyperbolischen Differentialgleichung im Wesentlichen entlang der Charakteristiken, so dass die Lösung auf einer Charakteristik nicht von Lösungen auf anderen Charakteristiken abhängt. Bei elliptischen Differentialgleichungen beeinflussen alle Punkte des Gebiets alle anderen Punkte, so dass wir andere Lösungstechniken anwenden müssen.

Einen Möglichkeit haben wir bereits in Gestalt der Formel (1.7) kennen gelernt: Wir können Ableitungen durch Differenzenquotienten approximieren und versuchen, damit die Differentialgleichung durch ein lineares Gleichungssystem zu approximieren. Indem

wir die Formel (1.7) auf die partiellen Ableitungen nach x und y anwenden, erhalten wir die Näherung

$$\begin{aligned}\frac{2u(x,y)-u(x+h,y)-u(x-h,y)}{h^2}&\\ +\frac{2u(x,y)-u(x,y+h)-u(x,y-h)}{h^2} &= -\frac{\partial^2 u}{\partial x^2}(x,y)-\frac{\partial^2 u}{\partial y^2}(x,y)\\ &\quad+\frac{h^2}{12}\left(\frac{\partial^4 u}{\partial x^4}(\eta_x,y)+\frac{\partial^4 u}{\partial y^4}(x,\eta_y)\right)\end{aligned} \tag{5.6}$$

mit geeigneten Zwischenpunkten $\eta_x \in [x-h, x+h]$ und $\eta_y \in [y-h, y+h]$. Also können wir den Laplace-Operator durch den Differenzenquotienten

$$\Delta_h u(x,y) = \frac{u(x+h,y)+u(x-h,y)+u(x,y+h)+u(x,y-h)-4u(x,y)}{h^2} \tag{5.7}$$

$$\text{für alle } (x,y)\in\Omega,\ h\in H_{x,y} \tag{5.8}$$

approximieren, wobei die Menge

$$H_{x,y} := \{h\in\mathbb{R}_{>0} \ :\ x+h\in[0,1],\ x-h\in[0,1],\ y+h\in[0,1],\ y-h\in[0,1]\}$$

beschreibt, für welche Schrittweiten in einem Punkt $(x,y)\in\Omega$ der Differenzenquotient ausgewertet werden kann.

Damit können wir (5.6) in der kompakten Form

$$|\Delta_h u(x,y)-\Delta u(x,y)| \le \frac{h^2}{6}|u|_{4,\Omega} \qquad \text{für alle } (x,y)\in\Omega,\ h\in H_{x,y} \tag{5.9}$$

schreiben. Hier ist

$$|u|_{4,\Omega} := \max\left\{\left\|\frac{\partial^{\nu+\mu}u}{\partial x^\nu \partial y^\mu}\right\|_{\infty,\Omega^2} \ :\ \nu,\mu\in\mathbb{N}_0,\ \nu+\mu=4\right\}$$

eine aus einer Variante der Maximum-Norm entstandene Halbnorm, die die Ableitungen vierter Ordnung einbezieht und sich deshalb gut eignet, um die Restterme der Taylor-Entwicklung zu beschränken.

Gegenüber dem Differentialoperator Δ bietet der *Differenzenoperator* Δ_h den Vorteil, dass er lediglich Werte der Funktion in einzelnen Punkten benötigt. Unser Ziel ist es, diese Eigenschaft auszunutzen, um das Gebiet Ω durch eine endliche Punktmenge zu ersetzen, die sich für die Berechnung im Computer wesentlich besser eignet.

Definition 5.3 (Gitter) *Sei $N\in\mathbb{N}$, und sei*

$$\begin{aligned} h &:= \frac{1}{N+1},\\ \Omega_h &:= \{(ih,jh) \ :\ i,j\in\{1,\dots,N\}\}\subseteq\Omega,\\ \partial\Omega_h &:= \{(ih,0),(ih,1),(0,jh),(1,jh) \ :\ i,j\in\{0,\dots,N+1\}\}\subseteq\partial\Omega,\\ \bar\Omega_h &:= \Omega_h\cup\partial\Omega_h. \end{aligned}$$

Wir nennen Ω_h, $\partial\Omega_h$ und $\bar\Omega_h$ Gitter *für die Gebiete Ω, $\partial\Omega$ und $\bar\Omega$.*

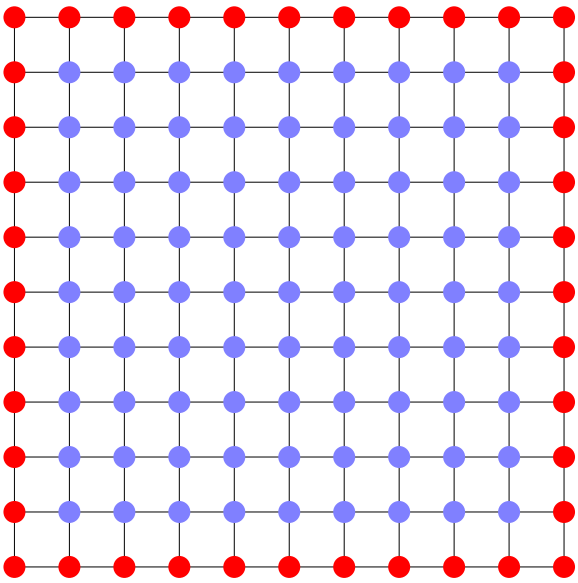

Abbildung 5.1: Gitter für $N = 9$

Wenn wir die Abschätzung (5.9) auf das Gitter Ω_h einschränken, erhalten wir

$$|-\Delta_h u(x,y) - f(x,y)| = |-\Delta_h u(x,y) + \Delta u(x,y)| \leq \frac{h^2}{6}\|u\|_{4,\bar{\Omega}} \quad \text{für alle } (x,y) \in \Omega,$$

also liegt es nahe, nach Lösungen der Gleichung $-\Delta_h u = f$ zu suchen, da wir hoffen dürfen, dass wir so die Lösung u approximieren können. Da bei der Auswertung des Differenzenoperators $\Delta_h u$ in einem Punkt $(x,y) \in \Omega_h$ nur Werte in Punkten auf $\bar{\Omega}_h$ verwendet werden, bietet es sich an, Funktionen zu untersuchen, die nur in diesen Punkten definiert sind:

Definition 5.4 (Gitterfunktionen) *Seien Ω_h und $\bar{\Omega}_h$ Gitter. Die Räume*

$$G(\Omega_h) := \{u_h \ : \ u_h \text{ ist eine Abbildung von } \Omega_h \text{ nach } \mathbb{R}\},$$
$$G(\bar{\Omega}_h) := \{u_h \ : \ u_h \text{ ist eine Abbildung von } \bar{\Omega}_h \text{ nach } \mathbb{R}\}$$

bezeichnen wir als die Räume der Gitterfunktionen *auf Ω_h beziehungsweise $\bar{\Omega}_h$. Den Raum*

$$G_0(\bar{\Omega}_h) := \{u_h \in G(\bar{\Omega}_h) \ : \ u_h(x,y) = 0 \text{ für alle } (x,y) \in \partial\Omega_h\}$$

bezeichnen wir als den Raum der Gitterfunktionen mit homogenen Dirichlet-Randwerten.

Offenbar ist Δ_h eine lineare Abbildung von $G(\bar{\Omega}_h)$ nach $G(\Omega_h)$, und wir können dem Gleichungssystem (5.5) die folgende Approximation gegenüberstellen:

Wir suchen eine Gitterfunktion $u_h \in G_0(\bar{\Omega}_h)$, die

$$-\Delta_h u_h(x,y) = f(x,y) \qquad \text{für alle } (x,y) \in \Omega_h, \tag{5.10}$$

erfüllt.

Da dieses Gleichungssystem statt auf einer kontinuierlichen Menge Ω auf einer diskreten Punktmenge Ω_h gegeben ist, bezeichnet man das System (5.10) als eine *Diskretisierung* der Potentialgleichung (5.5). Da bei dieser Technik alle Differentialoperatoren durch Differenzenquotienten endlich vieler Funktionswerte auf dem Gitter ersetzt werden, trägt sie den Namen *Finite-Differenzen-Verfahren.*

Es stellt sich natürlich die Frage nach der Lösbarkeit des diskreten Systems. Wir können leicht nachprüfen, dass $-\Delta_h$ eine lineare Abbildung von $G_0(\bar{\Omega}_h)$ nach $G(\Omega_h)$ ist und dass

$$\dim G_0(\bar{\Omega}_h) = \dim G(\Omega_h) = N^2$$

gilt. Für den Nachweis der eindeutigen Lösbarkeit des Gleichungssystems (5.10) genügt es deshalb, die Injektivität der Abbildung $-\Delta_h$ nachzuweisen.

Ein für diesen Zweck sehr nützliches Hilfsmittel ist das folgende Stabilitätsresultat für die Maximumnorm:

Lemma 5.5 (Maximumprinzip) *Sei $v_h \in G(\bar{\Omega}_h)$ eine Gitterfunktion, die*

$$-\Delta_h v_h(x,y) \le 0 \qquad \textit{für alle } (x,y) \in \Omega_h$$

erfüllt. Dann existiert ein Randpunkt $(x_0,y_0) \in \partial\Omega_h$ mit

$$v_h(x,y) \le v_h(x_0,y_0) \qquad \textit{für alle } (x,y) \in \bar{\Omega}_h,$$

die Gitterfunktion nimmt ihr Maximum also auf dem Rand des Gitters an.

Beweis. Wir definieren die Menge der Nachbarpunkte durch

$$N(x,y) := \{(x-h,y),(x+h,y),(x,y-h),(x,y+h)\} \qquad \text{für alle } (x,y) \in \Omega_h.$$

Den Abstand eines Gitterpunkts zum Rand des Gitters bezeichnen wir mit

$$\delta : \bar{\Omega}_h \to \mathbb{N}_0, \quad (x,y) \mapsto \begin{cases} 0 & \text{falls } (x,y) \in \partial\Omega_h, \\ 1 + \min\{\delta(x',y') \ : \ (x',y') \in N(x,y)\} & \text{ansonsten.} \end{cases}$$

Wir bezeichnen das Maximum der Funktion v_h mit

$$m := \max\{v_h(x,y) \ : \ (x,y) \in \bar{\Omega}_h\}$$

und werden nun per Induktion beweisen, dass

$$(v_h(x,y) = m \wedge \delta(x,y) \le d) \Rightarrow \exists (x_0,y_0) \in \partial\Omega_h : v_h(x_0,y_0) = m$$

für alle $d \in \mathbb{N}_0$ und $(x, y) \in \bar{\Omega}_h$ gilt. Daraus folgt offenbar unsere Behauptung.

Der Induktionsanfang $d = 0$ ist einfach: Falls ein Gitterpunkt $(x, y) \in \bar{\Omega}_h$ mit $v_h(x, y) = m$ und $\delta(x, y) = d = 0$ existiert, folgt aus der Definition der Abstandsfunktion bereits $(x, y) \in \partial\Omega_h$, wir haben also mit $(x_0, y_0) = (x, y)$ den gesuchten Randpunkt gefunden.

Sei nun $d \in \mathbb{N}_0$ so gewählt, dass die Behauptung gilt. Sei $(x, y) \in \bar{\Omega}_h$ ein Gitterpunkt mit $\delta(x, y) = d+1$ und $v_h(x, y) = m$. Dann gilt insbesondere $(x, y) \in \Omega_h$ und wir erhalten

$$4h^{-2}v_h(x, y) - \sum_{(x',y')\in N(x,y)} h^{-2}v_h(x', y') = -\Delta_h v_h(x, y) \leq 0,$$
$$\sum_{(x',y')\in N(x,y)} h^{-2}(v_h(x, y) - v_h(x', y')) \leq 0.$$

Da $m = v_h(x, y)$ das Maximum der Gitterfunktion v_h ist, kann jeder der Summanden auf der linken Seite dieser Ungleichung nicht negativ sein. Da die Summe nicht positiv ist, dürfen wir schließen, dass jeder der Summanden gleich null sein muss, also gilt

$$m = v_h(x, y) = v_h(x', y') \qquad \text{für alle } (x', y') \in N(x, y).$$

Da $\delta(x, y) = d + 1$ gilt, muss ein $(x', y') \in N(x, y)$ mit $\delta(x', y') = d$ existieren, und da wir gerade bewiesen haben, dass $v_h(x', y') = m$ für *alle* $(x', y) \in N(x, y)$ gilt, können wir die Induktionsvoraussetzung anwenden, um den Beweis abzuschließen. ∎

Das Maximumprinzip garantiert bereits die Injektivität des Differenzenoperator $-\Delta_h$: Falls für zwei Gitterfunktionen $u_h^{(1)}$ und $u_h^{(2)}$ die Gleichung $\Delta_h u_h^{(1)} = \Delta_h u_h^{(2)}$ gilt, können wir Lemma 5.5 auf $v_h = u_h^{(1)} - u_h^{(2)}$ anwenden und folgern, dass beide Funktionen sich nur unterscheiden können, falls sie sich auf dem Rand unterscheiden. Diesen Fall schließt die Dirichlet-Randbedingung gerade aus, also muss $u_h^{(1)} = u_h^{(2)}$ gelten. Damit ist Δ_h auf der Menge der Gitterfunktionen mit Dirichlet-Randbedingungen injektiv, also bijektiv, also ist das diskrete Problem (5.10) eindeutig lösbar.

Da Lemma 5.5 lediglich voraussetzt, dass $-\Delta_h u_h$ nicht echt positiv ist, können wir neben der Lösbarkeit des diskreten Problems sogar die folgende Stabilitätsaussage gewinnen, die garantiert, dass kleine Störungen der rechten Seite des Systems (5.10) nicht zu sehr verstärkt werden:

Lemma 5.6 (Stabilität) *Sei $u_h \in G_0(\bar{\Omega}_h)$ eine Gitterfunktion mit homogener Dirichlet-Randbedingung. Dann gilt*

$$\|u_h\|_{\infty,\Omega_h} \leq \frac{1}{8}\|\Delta_h u_h\|_{\infty,\Omega_h}.$$

Insbesondere ist Δ_h eine injektive Abbildung.

Beweis. (vgl. [6, Theorem 4.4.1]) Den Ausgangspunkt unseres Beweises bildet die Funktion

$$w : \bar{\Omega} \to \mathbb{R}_{\geq 0}, \qquad (x, y) \mapsto \frac{x}{2}(1 - x),$$

die $|w|_{4,\Omega} = 0$ erfüllt, so dass aus

$$-\Delta w(x,y) = 1 \qquad \text{für alle } (x,y) \in \Omega$$

dank (5.9) insbesondere auch

$$-\Delta_h w_h(x,y) = 1 \qquad \text{für alle } (x,y) \in \Omega_h$$

mit der Gitterfunktion $w_h := w|_{\bar{\Omega}_h} \in G(\bar{\Omega}_h)$ folgt.

Wir bezeichnen Minimum und Maximum der Funktion $-\Delta_h u_h$ mit

$$\alpha := \min\{-\Delta_h u_h(x,y) \ : \ (x,y) \in \Omega_h\},$$
$$\beta := \max\{-\Delta_h u_h(x,y) \ : \ (x,y) \in \Omega_h\}$$

und setzen

$$u_h^+ := w_h \beta.$$

Wir stellen fest, dass

$$-\Delta_h u_h^+(x,y) = -\Delta_h w_h(x,y)\beta = \beta \qquad \text{für alle } (x,y) \in \Omega_h$$

gilt, also muss insbesondere auch

$$-\Delta_h(u_h - u_h^+)(x,y) = -\Delta_h u_h(x,y) - \beta \le 0 \qquad \text{für alle } (x,y) \in \Omega_h$$

gelten. Dank Lemma 5.5 existiert ein Randpunkt $(x_0, y_0) \in \partial\Omega_h$ mit

$$u_h(x,y) - u_h^+(x,y) \le u_h(x_0,y_0) - u_h^+(x_0,y_0) \qquad \text{für alle } (x,y) \in \Omega_h.$$

Nach Voraussetzung gilt $u_h(x_0,y_0) = 0$, also folgt

$$u_h(x,y) \le u_h^+(x,y) - u_h^+(x_0,y_0) \qquad \text{für alle } (x,y) \in \Omega_h.$$

Aus einer einfachen Kurvendiskussion erhalten wir $0 \le w(x,y) \le 1/8$ für alle $(x,y) \in \bar{\Omega}_h$, also können wir

$$u_h(x,y) \le \frac{1}{8}\beta \qquad \text{für alle } (x,y) \in \Omega_h$$

folgern. Wir können dieselbe Argumentation auf $-u_h$ anwenden: Da $-u_h$ von oben durch $-\alpha$ beschränkt ist, erhalten wir

$$u_h(x,y) \ge \frac{1}{8}\alpha \qquad \text{für alle } (x,y) \in \Omega_h,$$

also insgesamt

$$\|u_h\|_{\infty,\Omega_h} \le \frac{1}{8}\max\{|\alpha|,|\beta|\} = \frac{1}{8}\|\Delta_h u_h\|_{\infty,\Omega_h},$$

und das ist die gewünschte Abschätzung. ∎

Wie schon bei der Fehleranalyse allgemeiner Einschrittverfahren (vgl. Satz 3.13) und der Extrapolationstechnik (vgl. Satz 4.3) können wir auch in diesem Fall die Konvergenz eines Näherungsverfahrens beweisen, indem wir die Stabilitätsaussage des Lemmas 5.6 mit einer *Konsistenzaussage* kombinieren. Im aktuellen Kontext erfüllt die Fehlerabschätzung (5.9) diesen Zweck: Wenn wir mit $\hat{u}_h := u|_{\bar{\Omega}_h}$ die Einschränkung der exakten Lösung auf das Gitter $\bar{\Omega}_h$ bezeichnen, folgt aus (5.9) und (5.5) die Abschätzung

$$\begin{aligned}\| - \Delta_h \hat{u}_h - f_h \|_{\infty,\Omega_h} &= \| - \Delta_h \hat{u}_h - f|_{\Omega_h} \|_{\infty,\Omega_h} \\ &= \| - \Delta_h \hat{u}_h + (\Delta u)|_{\Omega_h} \|_{\infty,\Omega_h} \leq \frac{h^2}{6} |u|_{4,\Omega}. \end{aligned} \tag{5.11}$$

Ähnlich wie bei Einschrittverfahren ist es angemessen, unsere Diskretisierung aufgrund dieser Abschätzung als *von zweiter Ordnung konsistent* zu bezeichnen. Indem wir Konsistenz und Stabilität kombinieren, können wir die Konvergenz der diskreten Lösung u_h gegen die Einschränkung der exakten Lösung $\hat{u}_h$ beweisen:

Satz 5.7 (Konvergenz) *Es gilt*

$$\|u_h - \hat{u}_h\|_{\infty,\Omega_h} \leq \frac{h^2}{48} |u|_{4,\Omega}.$$

Beweis. Wir kombinieren Lemma 5.6 mit (5.11) und erhalten

$$\begin{aligned}\|u_h - \hat{u}_h\|_{\infty,\Omega_h} &\leq \frac{1}{8} \|\Delta_h u_h - \Delta_h \hat{u}_h\|_{\infty,\Omega_h} \\ &= \frac{1}{8} \| - f_h - \Delta_h \hat{u}_h\|_{\infty,\Omega_h} \leq \frac{1}{8} \frac{h^2}{6} |u|_{4,\Omega}.\end{aligned}$$

Wenn doch alle Konvergenzbeweise so einfach wären. ∎

Wenn wir also das lineare Gleichungssystem (5.10) lösen können, dürfen wir auf eine Näherung der exakten Lösung hoffen, die wie h^2 konvergiert. Um das System praktisch zu lösen, bietet es sich an, geeignete Basen für die Räume $G_0(\bar{\Omega}_h)$ und $G(\Omega_h)$ zu wählen und $-\Delta_h$ in diesen Basen auszudrücken. Eine naheliegende Wahl ist die Basis $(b^{(v,w)})_{(v,w)\in\Omega_h}$ der Funktionen, die in (v,w) gleich eins und in allen anderen Punkten gleich null sind, denn diese Funktionen bilden offensichtlich eine Basis des Raums $G_0(\bar{\Omega}_h)$. Durch Einschränkung auf Ω_h erhalten wir auch eine Basis des Raums $G(\Omega_h)$, und in diesen Basen wird $-\Delta_h$ durch eine Matrix $\mathbf{L} \in \mathbb{R}^{\Omega_h \times \Omega_h}$ ausgedrückt, deren Einträge durch

$$(l_h)_{(x,y),(v,w)} := \begin{cases} 4h^{-2} & \text{falls } v = x,\ y = w, \\ -h^{-2} & \text{falls } |v - x| = h,\ y = w, \\ -h^{-2} & \text{falls } v = x,\ |y - w| = h, \\ 0 & \text{ansonsten} \end{cases} \qquad \text{für alle } (x,y),(v,w) \in \Omega_h$$

gegeben sind. Wenn wir die Gitterfunktionen u_h und f_h ebenfalls in den entsprechenden Basen ausdrücken, erhalten wir Vektoren $\mathbf{u}_h, \mathbf{f}_h \in \mathbb{R}^{\Omega_h}$, mit denen die diskrete Potentialgleichung (5.10) die Form

$$\mathbf{L}_h \mathbf{u}_h = \mathbf{f}_h \tag{5.12}$$

annimmt. Da (5.10) eindeutig lösbar ist, gilt dasselbe auch für (5.12).

Die Matrix $\mathbf{L}_h$ ist in diesem Fall besonders gutartig: Ein Blick auf die Koeffizienten zeigt, dass $\mathbf{L}_h = \mathbf{L}_h^*$ gilt, die Matrix ist also symmetrisch. Indem man Lemma 5.6 auf Teilmengen des Gitters Ω_h anwendet, lässt sich nachweisen, dass nicht nur die Matrix $\mathbf{L}_h$, sondern auch ihre sämtlichen Hauptuntermatrizen regulär sind, so dass die Existenz einer LR-Zerlegung gesichert ist, mit deren Hilfe sich das System (5.12) einfach lösen ließe. Es lässt sich sogar beweisen, dass $\mathbf{L}_h$ positiv definit ist, so dass auch eine effizientere Cholesky-Zerlegung zum Einsatz kommen könnte.

Für große Werte von N ist dieser Ansatz allerdings nicht sehr effizient, da er die besondere Struktur die Matrix $\mathbf{L}_h$ nicht ausnutzt: Jede Zeile oder Spalte der Matrix enthält nach Definition höchstens fünf von null verschiedene Einträge. Matrizen mit der Eigenschaft, dass nur wenige Einträge pro Zeile und Spalte von null abweichen, bezeichnet man als *schwachbesetzt*, und diese Eigenschaft lässt sich beispielsweise ausnutzen, um die Multiplikation einer Matrix mit einem Vektor besonders effizient durchzuführen oder das lineare Gleichungssystem besonders schnell zu lösen.

Finite-Differenzen-Verfahren eignen sich besonders gut für die Behandlung von Differentialgleichungen auf einfach geformten Gebieten wie dem hier untersuchten Einheitsquadrat. Die Behandlung komplizierterer Gebiete macht einerseits den Einsatz komplizierterer Techniken wie der *Shortley-Weller-Diskretisierung* erforderlich und erhöht andererseits die Komplexität der entstehenden Algorithmen beträchtlich.

5.3 Parabolische Differentialgleichungen und die Linienmethode

Parabolische Differentialgleichungen können als Kombination von elliptischen und gewöhnlichen Differentialgleichungen interpretiert werden: In einer Zeitvariablen verhält sie sich wie eine gewöhnliche Differentialgleichung, in den verbliebenen Ortsvariablen wie eine elliptische. Ein typisches Beispiel ist die *Wärmeleitungsgleichung*, die wir bereits in Abschnitt 1.4 kennen gelernt haben. Auf dem zweidimensionalen Gebiet $\Omega = (0,1)^2$, das wir bereits im vorigen Abschnitt verwendet haben, nimmt sie die Form

$$\frac{\partial u}{\partial t}(t,x,y) = f(t,x,y) + \frac{\partial^2 u}{\partial x^2}(t,x,y) + \frac{\partial^2 u}{\partial y^2}(t,x,y) \qquad \text{für alle } t \in [a,b], \quad (x,y) \in \Omega \tag{5.13}$$

an, hinzu kommen eine Anfangsbedingung

$$u(a,x,y) = u_0(x,y) \qquad \text{für alle } (x,y) \in \Omega$$

und der Einfachheit halber homogene Dirichlet-Randbedingungen

$$u(t,x,y) = 0 \qquad \text{für alle } t \in [a,b],\ (x,y) \in \partial\Omega$$

auf dem Rand $\partial\Omega$ des Gebiets Ω.

Die Idee der *Linienmethode* besteht darin, die Orts- und die Zeitvariable separat zu diskretisieren, in der Regel wird dabei zuerst die Ortsvariable behandelt. Dazu schreiben wir u und f nicht mehr als Funktionen in drei gleichberechtigten Variablen, sondern als Abbildungen, die jedem Zeitpunkt $t \in [a,b]$ Funktionen $u(t)$ und $f(t)$ in den Ortsvariablen zuordnen. Die Gleichung (5.13) lässt sich dann in der Form

$$u'(t)(x,y) = f(t)(x,y) + \Delta u(t)(x,y) \qquad \text{für alle } t \in [a,b],\ (x,y) \in \Omega$$

einer gewöhnlichen Differentialgleichung in dem Raum

$$C_0^2(\bar{\Omega}) := \{u \in C(\bar{\Omega}) \ : \ u|_\Omega \in C^2(\Omega),\ u|_{\partial\Omega} = 0\}$$

schreiben, der neben den Differenzierbarkeitsvoraussetzungen auch die Dirichlet-Randwerte der Lösung beschreibt.

Damit die rechte Seite diese Randbedingungen erfüllt, müssen wir

$$f(t) \in C_0^2(\bar{\Omega}) \qquad \text{für alle } t \in [a,b]$$

fordern und den Differentialoperator zu

$$\Delta_0 u(x,y) := \begin{cases} \Delta u(x,y) & \text{falls } (x,y) \in \Omega, \\ 0 & \text{ansonsten} \end{cases} \qquad \text{für alle } (x,y) \in \bar{\Omega}$$

fortsetzen. Die Wärmeleitungsgleichung (5.13) nimmt dann die Gestalt der gewöhnlichen Differentialgleichung

$$u'(t) = f(t) + \Delta_0 u(t), \qquad u(a) = u_0, \qquad \text{für alle } t \in [a,b] \tag{5.14}$$

mit Werten in dem Raum $C_0^2(\bar{\Omega})$ an.

Da dieser Raum unendlich-dimensional ist, müssen wir ihn durch einen endlich-dimensionalen Raum ersetzen, um ein praktisch durchführbares Verfahren zu erhalten. Dafür bietet sich der Raum $G_0(\bar{\Omega}_h)$ an, den wir im vorangehenden Abschnitt kennen gelernt haben. Wenn wir $C_0^2(\bar{\Omega})$ durch $G_0(\bar{\Omega}_h)$ ersetzen, liegt es nahe, auch Δ_h als Approximation des Differentialoperators Δ zu verwenden, allerdings müssen wir dann auch in diesem Fall dafür sorgen, dass geeignete Randbedingungen sichergestellt sind: Wir setzen

$$\Delta_{0,h} w_h(x,y) := \begin{cases} \Delta_h w_h(x,y) & \text{falls } (x,y) \in \Omega_h, \\ 0 & \text{anderenfalls} \end{cases} \qquad \text{für alle } w_h \in G_0(\bar{\Omega}_h),\ (x,y) \in \bar{\Omega}_h$$

und approximieren die Gleichung (5.14) durch die *semidiskrete Gleichung*

$$u_h'(t) = f_h(t) + \Delta_{0,h} u_h(t), \qquad u_h(a) = u_{0,h}, \qquad \text{für alle } t \in [a,b], \tag{5.15}$$

bei der die Gitterfunktionen

$$u_{0,h} := u_0|_{\bar{\Omega}_h} \in G_0(\bar{\Omega}_h), \qquad f_h(t) := f(t)|_{\bar{\Omega}_h} \in G_0(\bar{\Omega}_h) \qquad \text{für alle } t \in [a,b]$$

die Funktionen u_0 und $f(t)$ ersetzen und die Lösung u_h hoffentlich eine Näherung der exakten Lösung u in den Gitterpunkten sein wird.

Die Gleichung (5.15) ist eine gewöhnliche Differentialgleichung in dem endlichdimensionalen Raum $G_0(\bar{\Omega}_h)$, die wir mit den in den vorangehenden Kapiteln vorgestellten numerischen Verfahren behandeln können.

Zu klären bleibt dabei, ob sich sicherstellen lässt, dass die Lösung der semidiskreten Gleichung (5.15) eine Näherung der Lösung der ursprünglichen Gleichung (5.14) ist. Dazu untersuchen wir die Ableitung des Fehlers

$$e_h : [a,b] \to G_0(\bar{\Omega}_h), \qquad t \mapsto u_h(t) - u(t)|_{\bar{\Omega}_h},$$

und erhalten

$$\begin{aligned} e_h'(t) &= u_h'(t) - u'(t)|_{\bar{\Omega}_h} = f_h(t) + \Delta_{0,h} u_h(t) - f(t)|_{\bar{\Omega}_h} - \Delta_0 u(t)|_{\bar{\Omega}_h} \\ &= \Delta_{0,h} u_h(t) - \Delta_{0,h} u(t)|_{\bar{\Omega}_h} + \Delta_{0,h} u(t)|_{\bar{\Omega}_h} - \Delta_0 u(t)|_{\bar{\Omega}_h} \\ &= \Delta_{0,h}(u_h(t) - u(t)|_{\bar{\Omega}_h}) + \Delta_{0,h} u(t)|_{\bar{\Omega}_h} - (\Delta_0 u(t))|_{\bar{\Omega}_h} \\ &= \Delta_{0,h} e_h(t) + \gamma_h(t) \qquad \text{für alle } t \in [a,b], \end{aligned}$$

wobei die Gitterfunktion

$$\gamma_h : [a,b] \to G_0(\bar{\Omega}_h), \qquad t \mapsto \Delta_{0,h} u(t)|_{\bar{\Omega}_h} - (\Delta_0 u(t))|_{\bar{\Omega}_h},$$

den *räumlichen* Diskretisierungsfehler in jedem Gitterpunkt beschreibt. Da nach Voraussetzung

$$e_h(a) = u_h(a) - u(a)|_{\bar{\Omega}_h} = u_{0,h} - u_0|_{\bar{\Omega}_h} = 0$$

gilt, haben wir gezeigt, dass sich der Fehler e_h als Lösung der gewöhnlichen Differentialgleichung

$$e_h'(t) = \Delta_{0,h} e_h(t) + \gamma_h(t), \qquad e_h(a) = 0 \qquad \text{für alle } t \in [a,b] \tag{5.16}$$

darstellen lässt. Falls es uns also gelingt, diese Gleichung zu lösen und die Norm der Lösung abzuschätzen, haben wir auch eine Schranke für den Fehler gefunden.

Lemma 5.8 (Stabilität) *Sei $v_{h,0} \in G_0(\bar{\Omega})$, und sei $\gamma_h \in C([a,b], G_0(\bar{\Omega}))$. Die gewöhnliche Differentialgleichung*

$$v_h'(t) = \gamma_h(t) + \Delta_{0,h} v_h(t) \qquad v_h(a) = v_{h,0}, \qquad \textit{für alle } t \in [a,b] \tag{5.17}$$

besitzt genau eine Lösung, und diese Lösung erfüllt die Abschätzung

$$\|v_h(t)\|_{\infty,\Omega_h} \leq \|v_{h,0}\|_{\infty,\Omega_h} + \int_a^t \|\gamma_h(s)\|_{\infty,\Omega_h}\, ds \qquad \textit{für alle } t \in [a,b].$$

Beweis. Existenz und Eindeutigkeit der Lösung der Gleichung (5.17) folgen dank der Lipschitz-Stetigkeit der rechten Seite aus Satz 2.3.

Wir beweisen die Abschätzung für die Norm der Lösung, indem wir die gewöhnliche Differentialgleichung mit dem impliziten Euler-Verfahren approximieren und ausnutzen, dass die Näherungslösungen nach Folgerung 3.19 gegen die kontinuierliche Lösung konvergieren.

Sei $t \in [a, b]$. Um $v_h(t)$ mit Hilfe des impliziten Euler-Verfahrens zu approximieren, wählen wir $m \in \mathbb{N}$ und legen Zeitschrittweite und Zwischenpunkte durch

$$\delta := \frac{t-a}{m}, \qquad t_i := a + i\delta \qquad \text{für alle } i \in \{0, \dots, m\}$$

fest. Die Näherungslösung des impliziten Euler-Verfahrens für die Schrittweite δ ist dann gegeben durch

$$\tilde{v}_h(t_0) = v_{h,0}, \quad (I - \delta\Delta_{0,h})\tilde{v}_h(t_{i+1}) = \tilde{v}_h(t_i) + \delta\gamma_h(t_{i+1}) \quad \text{für alle } i \in \{0, \dots, m-1\}.$$

Zunächst beweisen wir eine Variante der Stabilitätsaussage des Lemma 5.6 für den Operator $I - \delta\Delta_{0,h}$. Sei $u_h \in G(\bar{\Omega}_h)$. Wir wählen $(x, y) \in \Omega_h$ so, dass $u_h(x, y)$ maximal ist. Dann folgt

$$(I - \delta\Delta_{0,h})u_h(x,y) = u_h(x,y) + \delta h^{-2} \sum_{(x',y') \in N(x,y)} \underbrace{(u_h(x,y) - u_h(x',y'))}_{\geq 0} \geq u_h(x,y),$$

also ist das Maximum der Gitterfunktion u_h beschränkt durch das Maximum der Gitterfunktion $(I - \delta\Delta_{0,h})u_h$. Indem wir dieses Resultat auch auf $-u_h$ anwenden, folgt

$$\|u_h\|_{\infty,\Omega_h} \leq \|(I - \delta\Delta_{0,h})u_h\|_{\infty,\Omega_h} \qquad \text{für alle } \delta \in \mathbb{R}_{\geq 0},\ u_h \in G(\bar{\Omega}_h). \tag{5.18}$$

Für das implizite Euler-Verfahren folgt daraus

$$\|\tilde{v}_h(t_{i+1})\|_{\infty,\Omega_h} \leq \|\tilde{v}_h(t_i)\|_{\infty,\Omega_h} + \delta\|\gamma_h(t_{i+1})\|_{\infty,\Omega_h} \qquad \text{für alle } i \in \{0, \dots, m-1\}.$$

Mit einer einfachen Induktion erhalten wir

$$\|\tilde{v}_h(t_i)\|_{\infty,\Omega_h} \leq \|v_{h,0}\|_{\infty,\Omega_h} + \delta \sum_{j=1}^{i} \|\gamma_h(t_j)\|_{\infty,\Omega_h} \qquad \text{für alle } i \in \{0, \dots, m\}.$$

Nach Mittelwertsatz der Integralrechnung finden wir für jedes $j \in \{1, \dots, m\}$ ein $\eta_j \in [t_{j-1}, t_j]$, das

$$\int_{t_{j-1}}^{t_j} \|\gamma_h(s)\|_{\infty,\Omega_h}\, ds = h\|\gamma_h(\eta_j)\|_{\infty,\Omega_h}$$

erfüllt, also folgt

$$\|\tilde{v}_h(t)\|_{\infty,\Omega_h} \leq \|v_{h,0}\|_{\infty,\Omega_h} + \sum_{j=1}^{m} \int_{t_{j-1}}^{t_j} \|\gamma_h(s)\|_{\infty,\Omega_h}\, ds + \delta(\|\gamma_h(t_j)\|_{\infty,\Omega_h} - \|\gamma_h(\eta_j)\|_{\infty,\Omega_h})$$

$$= \|v_{h,0}\|_{\infty,\Omega_h} + \int_a^t \|\gamma_h(s)\|_{\infty,\Omega_h}\, ds$$
$$+ (t-a)\max\{\|\gamma_h(t_j)\|_{\infty,\Omega_h} - \|\gamma_h(\eta_j)\|_{\infty,\Omega_h} \;:\; j \in \{1,\dots,m\}\}.$$

Für $m \to \infty$, also $\delta \to 0$, konvergiert die linke Seite dieser Ungleichung nach Folgerung 3.19 gegen $v_h(t)$. Da die Funktion $t \mapsto \|\gamma_h(t)\|_{\infty,\Omega_h}$ stetig auf dem kompakten Intervall $[a,b]$ ist, ist sie auch gleichmäßig stetig, wir können also für jedes $\epsilon > 0$ ein $m \in \mathbb{N}$ so finden, dass $\|\gamma_h(t_j)\|_{\infty,\Omega} - \|\gamma_h(\eta_j)\|_{\infty,\Omega_h} < \epsilon$ gilt. Damit konvergiert der letzte Summand der rechten Seite für $m \to \infty$ gegen null und die Aussage ist bewiesen. ∎

Mit Hilfe dieser Abschätzung können wir eine Konvergenzaussage für die Lösung $\hat{u}_h$ des semidiskreten Problems (5.15) entwickeln.

Satz 5.9 (Konvergenz) *Es gilt*

$$\|u_h(t) - u(t)|_{\bar{\Omega}_h}\|_{\infty,\bar{\Omega}_h} \le \frac{h^2}{6}\int_a^t |u(s)|_{4,\Omega}\, ds \qquad \text{für alle } t \in [a,b].$$

Beweis. Wir haben bereits gezeigt, dass die durch (5.16) definierte Gitterfunktionen $e_h(t)$ gerade den Unterschied zwischen der exakten Lösung u und der Näherungslösung u_h beschreibt.

Aus der Abschätzung (5.9) des räumlichen Diskretisierungsfehlers folgt

$$\|\gamma_h(t)\|_{\infty,\bar{\Omega}_h} \le \frac{h^2}{6}|\hat{u}(t)|_{4,\Omega} \qquad \text{für alle } t \in [a,b],$$

und dank Lemma 5.8 erhalten wir damit die gesuchte Fehlerschranke. ∎

Bemerkung 5.10 (Steife Differentialgleichung) *Es ist zu betonen, dass für den Beweis der zentralen Hilfsaussage des Lemmas 5.8 von entscheidender Bedeutung ist, dass ein* implizites *Euler-Verfahren verwendet wird. Man kann beweisen, dass die Eigenwerte des räumlichen Differenzenoperators $\Delta_{0,h}$ strikt negativ sind und dass der betragsgrößte von ihnen sich näherungsweise proportional zu h^{-2} verhält, während der betragskleinste näherungsweise konstant ist. Damit sind wir in der Situation einer* steifen Differentialgleichung *des in Abschnitt 4.3 besprochenen Typs, und wie wir bereits gesehen haben, empfehlen sich für diese Gleichungen implizite Verfahren.*

Im Fall des Beweises des Lemmas 5.8 profitieren wir davon, dass wir dank des impliziten Euler-Verfahrens die Stabilitätsabschätzung (5.18) verwenden können, um die Fortpflanzung des Fehlers aus vorangehenden Zeitschritten besonders elegant unter Kontrolle zu bringen.

6 Variationsformulierungen und das Finite-Elemente-Verfahren

Mit dem Finite-Differenzen-Verfahren haben wir in Abschnitt 5.2 bereits eine Möglichkeit kennen gelernt, um elliptische Differentialgleichungen zu behandeln. Diese Technik ist relativ einfach, aber auch nicht allzu flexibel: Die Behandlung allgemeiner Geometrien ist schwierig, und die theoretische Untersuchung führt in unserem Fall nur zu guten Abschätzungen, wenn die Lösung viermal differenzierbar mit gleichmäßig beschränkten Ableitungen ist.

Wir befassen uns in diesem Kapitel mit einem sehr viel flexibleren Ansatz für die Behandlung partieller Differentialgleichungen: Die Differentialgleichung wird in eine *Variationsformulierung* überführt, die mit einem Skalarprodukt in einem geeignet gewählten Hilbertraum korrespondiert. Die Idee der *Galerkin-Diskretisierung* besteht darin, diesen Hilbertraum durch einen endlich-dimensionalen Teilraum zu ersetzen, in dem die Variationsformulierung einem linearen Gleichungssystem entspricht. Die Lösung des Gleichungssystems definiert eine Approximation der Lösung der Variationsformulierung und damit auch der ursprünglichen Differentialgleichung. Bei der Wahl des endlich-dimensionalen Teilraums hat sich das *Finite-Elemente-Verfahren* bewährt, das das zugrundeliegende Gebiet in kleine Teilgebiete (die besagten finiten Elemente) zerlegt und auf jedem dieser Teilgebiete einen polynomiellen Ansatz verwendet. Diese Konstruktion hat zur Folge, dass einerseits das resultierende lineare Gleichungssystem eine günstige Struktur besitzt und andererseits die Analyse der Approximationseigenschaften relativ elegant durchgeführt werden kann.

6.1 Variationsformulierung

Wir untersuchen die neu zu entwickelnden Techniken wieder am Beispiel der Potentialgleichung, allerdings diesmal auf einem allgemeinen offenen zusammenhängenden Polygongebiet $\Omega \subseteq \mathbb{R}^d$ mit einer rechten Seite $f \in C(\bar{\Omega})$. Wir verwenden wieder homogene Dirichlet-Randbedingungen, suchen also eine Funktion

$$u \in C_0^2(\Omega) = \{u \in C(\bar{\Omega}) \ : \ u|_\Omega \in C^2(\Omega),\ u|_{\partial\Omega} = 0\},$$

die die Gleichung

$$-\Delta u(x) = f(x) \qquad \text{für alle } x \in \Omega \tag{6.1}$$

erfüllt. Der Laplace-Operator ist im d-dimensionalen Raum durch

$$\Delta = \sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2} = \frac{\partial^2}{\partial x_1^2} + \ldots + \frac{\partial^2}{\partial x_d^2}$$

gegeben, für $d = 2$ fällt diese Definition offenbar mit der aus Abschnitt 5.2 zusammen. Im Gegensatz zu diesem Abschnitt fassen wir jetzt die Koordinaten zu einem Vektor $x \in \Omega$ zusammen, um den Schreibaufwand zu reduzieren.

Um die Gleichung (6.1) analysieren zu können, überführen wir sie in eine *Variationsformulierung*, indem wir mit Funktionen $v \in C(\bar{\Omega})$ multiplizieren und integrieren. Falls (6.1) gilt, muss auch

$$-\int_\Omega v(x)\Delta u(x)\,dx = \int_\Omega v(x)f(x)\,dx \qquad \text{für alle } v \in C(\bar{\Omega}) \tag{6.2}$$

gelten. Statt also die Gültigkeit der Gleichung in allen Punkten des Gebiets zu fordern, multiplizieren wir sie mit *Testfunktionen* v und fordern, dass die Integrale auf beiden Seiten der Gleichung übereinstimmen. Die punktweise Identität beider Seiten wird also durch gewichtete Integralmittelwerte ersetzt.

Trotzdem können wir uns überlegen, dass beide Formulierungen gleichwertig sind, sofern uns „genügend viele“ Testfunktionen zur Verfügung stehen: (6.2) entspricht der Gleichung

$$\int_\Omega v(x)(f(x) + \Delta u(x))\,dx = 0 \qquad \text{für alle } v \in C(\bar{\Omega}),$$

und indem wir $v := f + \Delta u$ einsetzen, folgt

$$\int_\Omega (f(x) + \Delta u(x))^2\,dx = 0$$

und damit insbesondere (6.1).

Die Formulierung (6.2) ist nur wohldefiniert, falls u zweimal stetig differenzierbar ist. Unser Ziel ist es nun, diese Voraussetzung abzuschwächen, indem wir die linke Seite der Gleichung *partiell integrieren*. Dazu benötigen wir eine Verallgemeinerung der partiellen Integration für mehrdimensionale Polygongebiete Ω.

Lemma 6.1 (Greensche Formel) *Seien $u, v \in C(\bar{\Omega})$ mit $u|_\Omega, v|_\Omega \in C^1(\Omega)$ gegeben. Sei $n : \partial\Omega \to \mathbb{R}^d$ eine Abbildung, die jedem Randpunkt einen in das Äußere des Gebiets weisenden Normaleneinheitsvektor zuordnet. Sei $i \in \{1, \dots, d\}$. Dann gilt*

$$\int_\Omega v(x)\frac{\partial u}{\partial x_i}(x)\,dx = -\int_\Omega \frac{\partial v}{\partial x_i}(x)u(x)\,dx + \int_{\partial\Omega} v(s)u(s)n_i(s)\,ds.$$

Das rechte Integral ist dabei ein Kurvenintegral und im Sinne eines Lebesgue-Integrals zu interpretieren, da der Faktor n_i in unserem Fall nur stückweise konstant ist.

Beweis. Ein allgemeiner Beweis würde hier zu weit führen, deshalb sei nur kurz darauf verwiesen, dass sich die Aussage aus der Produktregel und dem Gaußschen Integralsatz ergibt, der beispielsweise in [3, 9] bewiesen wird.

Zur Veranschaulichung der Gleichung beschränken wir uns auf den Fall des Einheitsquadrats $\Omega = (0,1)^2$ und $i = 1$. Indem wir das zweidimensionale Integral durch eindimensionale Integrale ausdrücken und letztere partiell integrieren folgt

$$\begin{aligned}\int_\Omega v(x)\frac{\partial u}{\partial x_1}(x)\,dx &= \int_0^1\int_0^1 v(x_1,x_2)\frac{\partial u}{\partial x_1}(x_1,x_2)\,dx_1\,dx_2\\ &= \int_0^1\left(-\int_0^1 \frac{\partial v}{\partial x_1}(x_1,x_2)u(x_1,x_2)\,dx_1 + v(1,x_2)u(1,x_2) - v(0,x_2)u(0,x_2)\right)dx_2\\ &= -\int_0^1\int_0^1 \frac{\partial v}{\partial x_1}(x_1,x_2)u(x_1,x_2)\,dx_1\,dx_2\\ &\quad + \int_0^1 v(1,x_2)u(1,x_2)\,dx_2 - \int_0^1 v(0,x_2)u(0,x_2)\,dx_2.\end{aligned}$$

Auf dem linken Rand des Einheitsquadrats ist $(-1,0)$ der äußere Normaleneinheitsvektor, auf dem rechten ist es $(1,0)$, auf dem oberen $(0,1)$ und auf dem unteren $(0,-1)$, so dass sich

$$\begin{aligned}\int_\Omega v(x)\frac{\partial u}{\partial x_1}(x)\,dx &= -\int_\Omega \frac{\partial v}{\partial x_1}(x)u(x)\,dx\\ &\quad + \int_0^1 v(1,s)u(1,s)n_1(s)\,ds + \int_0^1 v(0,s)u(0,s)n_1(s)\,ds\\ &= -\int_\Omega \frac{\partial v}{\partial x_1}(x)u(x)\,dx + \int_{\partial\Omega} v(s)u(s)n_1(s)\,ds\end{aligned}$$

ergibt, und entsprechend können wir auch mit $i = 2$ verfahren. ∎

Mit Hilfe der Greenschen Formel können wir (6.2) partiell integrieren und erhalten

$$\begin{aligned}-\int_\Omega v(x)\Delta u(x)\,dx &= -\sum_{i=1}^d \int_\Omega v(x)\frac{\partial^2 u}{\partial x_i^2}(x)\,dx\\ &= \sum_{i=1}^d \int_\Omega \frac{\partial v}{\partial x_i}(x)\frac{\partial u}{\partial x_i}(x)\,dx - \int_{\partial\Omega} v(s)\frac{\partial u}{\partial s_i}(s)n_i(s)\,ds.\end{aligned}$$

Um die Randintegrale zu vermeiden, beschränken wir uns auf Funktionen

$$v \in C_0^1(\Omega) = \{v \in C(\bar\Omega)\ :\ v|_\Omega \in C^1(\Omega),\ v|_{\partial\Omega} = 0\}$$

und erhalten

$$-\int_\Omega v(x)\Delta u(x)\,dx = \sum_{i=1}^d \int_\Omega \frac{\partial v}{\partial x_i}(x)\frac{\partial u}{\partial x_i}(x)\,dx \qquad \text{für alle } v \in C_0^1(\Omega).$$

Diese Formel können wir etwas kompakter schreiben, indem wir die folgende Abkürzung einführen:

Definition 6.2 (Gradient) *Sei $u \in C^1(\Omega)$. Dann bezeichnen wir die Abbildung*

$$\nabla u : \Omega \to \mathbb{R}^d, \qquad x \mapsto \begin{pmatrix} \frac{\partial u}{\partial x_1}(x) \\ \vdots \\ \frac{\partial u}{\partial x_d}(x) \end{pmatrix},$$

als den Gradienten *der Funktion u.*

Das Ergebnis der partiellen Integration lässt sich mit Hilfe des Gradienten als

$$-\int_\Omega v(x)\Delta u(x)\,dx = \int_\Omega \langle \nabla v(x), \nabla u(x)\rangle_2\,dx \qquad \text{für alle } v \in C_0^1(\Omega)$$

zusammenfassen, und die Gleichung (6.2) nimmt die folgende Form an:

$$\int_\Omega \langle \nabla v(x), \nabla u(x)\rangle_2\,dx = \int_\Omega v(x) f(x)\,dx \qquad \text{für alle } v \in C_0^1(\Omega).$$

Da in dieser Gleichung nur noch erste Ableitungen der Funktion u auftreten, können wir die Voraussetzungen an deren Differenzierbarkeit reduzieren und gelangen zu der folgenden *schwachen Formulierung*:

Wir suchen eine Funktion $u \in C_0^1(\Omega)$, die

$$\int_\Omega \langle \nabla v(x), \nabla u(x)\rangle_2\,dx = \int_\Omega v(x) f(x)\,dx \qquad \text{für alle } v \in C_0^1(\Omega) \qquad (6.3)$$

erfüllt.

Aus unserer Herleitung folgt, dass jede Lösung der ursprünglichen Gleichung (6.1) auch eine Lösung der schwachen Formulierung des Problems ist. Die Umkehrung gilt in der Regel nicht, allerdings lässt sich nachweisen, dass Lösungen der schwachen Formulierung, sofern sie existieren, eindeutig sind, so dass wir die klassische Lösung erhalten, falls sie existiert, und anderenfalls eine verallgemeinerte Lösung.

Wir widmen unsere Aufmerksamkeit der linken Seite der Gleichung (6.3), die durch die Abbildung

$$a : C_0^1(\Omega) \times C_0^1(\Omega) \to \mathbb{R}, \qquad (v,u) \mapsto \int_\Omega \langle \nabla v(x), \nabla u(x)\rangle_2\,dx, \qquad (6.4)$$

beschrieben wird. Diese Abbildung besitzt einige besondere Eigenschaften: Es gelten

$$\begin{aligned} &a(v + \alpha w, u) = a(v,u) + \alpha a(w,u), \\ &a(v, u + \alpha w) = a(v,u) + \alpha a(v,w) \qquad \text{für alle } u,v,w \in C_0^1(\Omega),\ \alpha \in \mathbb{R}, \end{aligned}$$

also ist a eine *Bilinearform*, es gilt

$$a(v,u) = a(u,v) \qquad \text{für alle } u,v \in C_0^1(\Omega),$$

also ist *a symmetrisch*, und es gilt auch

$$a(u,u) > 0 \qquad \text{für alle } u \in C_0^1(\Omega) \setminus \{0\},$$

also ist *a positiv definit*. Dass $a(u,u) \geq 0$ für alle $u \in C_0^1(\Omega)$ gilt, folgt dabei aus der Positivität des Integrals. Falls ein $u \in C_0^1(\Omega)$ mit $a(u,u) = 0$ gegeben ist, folgt nach Definition $\nabla u = 0$, also muss u konstant sein. Da u die homogenen Dirichlet-Randbedingungen erfüllt, können wir auf $u = 0$ schließen.

Eine symmetrische positiv definite Bilinearform wird als *Skalarprodukt* bezeichnet, und Skalarprodukte sind in der Regel mit *Hilberträumen* assoziiert. Falls es uns also gelingen sollte, a als Skalarprodukt auf einem geeignet gewählten Hilbertraum zu identifizieren, können wir darauf hoffen, mit Hilfe der in diesen Räumen zur Verfügung stehenden Existenz- und Eindeutigkeitsaussagen die Lösbarkeit der Variationsformulierung untersuchen zu können. Die entscheidende Hürde dabei ist die Vollständigkeit des Raums: In dem Beweis für die Existenz einer Lösung wird diese Lösung als Grenzwert einer Cauchy-Folge konstruiert, also muss sichergestellt sein, dass wir in einem Raum arbeiten, in dem Cauchy-Folgen einen Grenzwert besitzen.

6.2 Sobolew-Räume

Es stellt sich die Frage, wie wir einen Hilbertraum konstruieren können, auf dem unsere Bilinearform a ein Skalarprodukt ist. Da a eine Bilinearform auf einem Funktionenraum ist, bietet es sich an, bei der Suche mit dem Raum der quadratintegrablen Funktionen zu beginnen, einem der grundlegenden aus Funktionen bestehenden Hilberträume.

Definition 6.3 ($L^2(\Omega)$) *Wir bezeichnen mit*

$$L^2(\Omega) := \{u : \Omega \to \mathbb{R} \ : \ u \text{ Lebesgue-messbar}, \ u^2 \text{ Lebesgue-integrierbar}\}$$

den Raum aller quadratintegrablen Funktionen. Wir versehen ihn mit dem Skalarprodukt

$$\langle u, v\rangle_{L^2} := \int_\Omega u(x)v(x)\,dx \qquad \text{für alle } u, v \in L^2(\Omega),$$

und der Norm

$$\|u\|_{L^2} := \left(\int_\Omega u(x)^2\,dx\right)^{1/2} = \sqrt{\langle u, u\rangle_{L^2}} \qquad \text{für alle } u \in L^2(\Omega).$$

Wie üblich werden dabei Funktionen miteinander identifiziert, die sich nur auf einer Nullmenge unterscheiden.

Wie man sieht lässt sich unsere Bilinearform a mit Hilfe des L^2-Skalarprodukts in der Form

$$a(v,u) = \langle \nabla v, \nabla u\rangle_{L^2} \qquad \text{für alle } u, v \in C_0^1(\Omega)$$

darstellen, wir müssten also „nur noch“ klären, unter welchen Bedingungen die partiellen Ableitungen einer Funktion u Elemente des Raums $L^2(\Omega)$ sind.

Um diese Frage allgemein untersuchen zu können, bietet es sich an, die folgende einheitliche (und kompaktere) Notation für partielle Ableitungen zu verwenden:

Definition 6.4 (Multiindizes) *Wir bezeichnen die Menge*

$$M_d := \mathbb{N}_0^d = \{(\nu_1, \ldots, \nu_d) \ : \ \nu_1, \ldots, \nu_d \in \mathbb{N}_0\}$$

als die Menge der d-dimensionalen Multiindizes, *eines ihrer Elemente* $\nu \in M_d$ *nennen wir einen* Multiindex. *Für jeden Multiindex* $\nu \in M_d$ *nennen wir*

$$|\nu| := \nu_1 + \ldots + \nu_d$$

seinen Betrag *und*

$$\partial^\nu := \frac{\partial^{|\nu|}}{\partial x^{\nu_1} \ldots \partial x^{\nu_d}}$$

den zugehörigen partiellen Ableitungsoperator.

Um die Ableitung einer Funktion aus $L^2(\Omega)$ zu definieren, greifen wir wieder auf die Idee zurück, mit einer Testfunktion zu multiplizieren und partiell zu integrieren. Damit wir Ableitungen beliebig hoher Ordnung definieren können, verwenden wir als Testfunktionen Elemente des Raums

$$C_0^\infty(\Omega) := \{u \in C^\infty(\bar{\Omega}) \ : \ \text{ der Träger von } u \text{ ist eine kompakte Teilmenge von } \Omega\}.$$

Dabei sei daran erinnert, dass der *Träger* einer Funktion der Abschluss der Teilmenge des Definitionsbereichs ist, auf dem sie nicht gleich null ist.

Für ein $\nu \in M_d$, eine Funktion $u \in C^{|\nu|}(\Omega)$ und eine Funktion $\varphi \in C_0^\infty(\Omega)$ erhalten wir durch wiederholtes partielles Integrieren die Gleichung

$$\int_\Omega \partial^\nu \varphi(x) u(x)\, dx = (-1)^{|\nu|} \int_\Omega \varphi(x) \partial^\nu u(x)\, dx,$$

und diese Gleichung können wir einsetzen, um eine Verallgemeinerung des Ableitungsbegriffs zu definieren: Falls eine Funktion $w \in L^2(\Omega)$ existiert, die für alle $\varphi \in C_0^\infty(\Omega)$ die Rolle der Ableitung $\partial_\nu u$ in der obigen Gleichung spielen kann, können wir sie als verallgemeinerte Ableitung verwenden.

Definition 6.5 (Schwache Ableitung) *Sei* $u \in L^2(\Omega)$ *und sei* $\nu \in M_d$. *Falls eine Funktion* $w \in L^2(\Omega)$ *existiert, die*

$$\int_\Omega \partial^\nu \varphi(x) u(x)\, dx = (-1)^{|\nu|} \int_\Omega \varphi(x) w(x)\, dx \qquad \text{für alle } \varphi \in C_0^\infty(\Omega)$$

erfüllt, nennen wir w *die* ν*-te* schwache Ableitung *der Funktion* u *und bezeichnen sie mit* $\partial^\nu u := w$.

Man kann nachweisen, dass der Raum $C_0^\infty(\Omega)$ eine dichte Teilmenge des Raums $L^2(\Omega)$ ist, und damit folgt aus dieser Definition bereits, dass die schwachen Ableitungen einer Funktion eindeutig definiert sind, sofern sie existieren. Aus dieser Eindeutigkeit folgt insbesondere, dass die schwache Ableitung mit der klassischen Ableitung übereinstimmt, falls die klassische Ableitung existiert, die schwache Ableitung kann also als Fortsetzung des Ableitungsoperators auf eine Teilmenge des Raums $L^2(\Omega)$ interpretiert werden.

Es liegt nahe, in Anlehnung an die Räume der differenzierbaren Funktionen nun Räume schwach differenzierbarer Funktionen zu definieren.

Definition 6.6 (Sobolew-Raum) *Sei $m \in \mathbb{N}_0$. Der Raum*

$$H^m(\Omega) := \{u \in L^2(\Omega) \ : \ \text{für alle } \nu \in M_d \text{ mit } |\nu| \leq m \text{ existiert } \partial^\nu u \in L^2(\Omega)\}$$

heißt Sobolew-Raum m-ter Ordnung. *Mit dem durch*

$$\langle v, u\rangle_{H^m} := \sum_{\substack{\nu \in M_d \\ |\nu| \leq m}} \langle \partial^\nu v, \partial^\nu u\rangle_{L^2} \qquad \text{für alle } v, u \in H^m(\Omega)$$

definierten Skalarprodukt und der durch

$$\|u\|_{H^m} := \langle u, u\rangle_{H^m}^{1/2} \qquad \text{für alle } u \in H^m(\Omega)$$

definierten Norm ist $H^m(\Omega)$ ein Hilbertraum.

Die Vollständigkeit des Sobolew-Raums $H^m(\Omega)$ folgt aus der des Raums $L^2(\Omega)$: Falls $(u_\ell)_{\ell\in\mathbb{N}}$ eine Cauchy-Folge in $H^m(\Omega)$ ist, muss $(\partial^\nu u_\ell)_{\ell\in\mathbb{N}}$ für jedes $\nu \in M_d$ mit $|\nu| \leq m$ eine Cauchy-Folge in $L^2(\Omega)$ sein, also einen Grenzwert besitzen. Dass diese Grenzwerte gegen die schwachen Ableitungen derselben Funktion konvergieren, folgt aus der Cauchy-Schwarz-Ungleichung und Definition 6.5.

Für die Analyse der Bilinearform a bietet sich der Sobolew-Raum $H^1(\Omega)$ an, allerdings wäre sie auf diesem Raum nicht positiv definit: Die konstante Funktion ist in $H^1(\Omega)$ enthalten, und wenn wir sie in a einsetzen, erhalten wir null. Also kann a auf diesem Raum nur positiv semidefinit sein.

Die Lösung dieses Problems kennen wir glücklicherweise schon: Wie schon im Fall des Problems (6.3) müssen wir die homogenen Dirichlet-Randbedingungen ausnutzen. Das „schwache Gegenstück“ des Raums $C_0^1(\Omega)$ müssen wir dabei, da uns die Einschränkung auf $\partial\Omega$ für Funktionen aus $H^1(\Omega)$ nicht zur Verfügung steht, indirekt definieren:

Definition 6.7 (Homogene Randbedingungen) *Sei $m \in \mathbb{N}_0$. Der Raum*

$$H_0^m(\Omega) := \{u \in H^m(\Omega) \ : \ u \text{ ist Grenzwert einer Folge in } C_0^\infty(\Omega)\}$$

heißt Sobolew-Raum m-ter Ordnung mit Dirichlet-Randwerten. *Als abgeschlossene Teilmenge des Hilbertraums $H^m(\Omega)$ ist er ebenfalls ein Hilbertraum.*

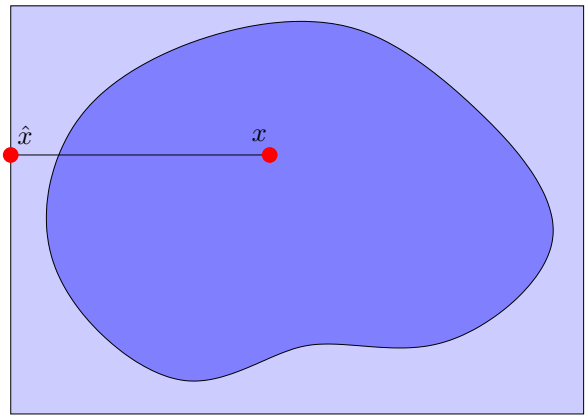

Abbildung 6.1: Illustration des Beweises der Friedrichs-Ungleichung

Auf Sobolew-Räumen mit Dirichlet-Randwerten können wir die $L^2(\Omega)$-Norm einer Funktion durch einen Term beschränken, der nur von ihren schwachen Ableitungen und der Form des Gebiets abhängt:

Lemma 6.8 (Friedrichs-Ungleichung) *Wir definieren die* H^1-Halbnorm *durch*

$$|u|_{H^1} := \left(\sum_{\substack{\nu \in M_d \\ |\nu|=1}} \|\partial^\nu u\|_{L^2}^2 \right)^{1/2} \qquad \text{für alle } u \in H^1(\Omega).$$

Es existiert eine Konstante $C_\Omega \in \mathbb{R}_{\geq 0}$*, die nur von dem Gebiet* Ω *abhängt und*

$$\|u\|_{L^2} \leq C_\Omega |u|_{H^1} \qquad \text{für alle } u \in H_0^1(\Omega)$$

erfüllt. Insbesondere sind auf dem Teilraum $H_0^1(\Omega)$ *die* H^1*-Halbnorm und die* H^1*-Norm äquivalent, denn es gilt*

$$|u|_{H^1} \leq \|u\|_{H^1} \leq (C_\Omega^2 + 1)^{1/2} |u|_{H^1} \qquad \text{für alle } u \in H_0^1(\Omega).$$

Beweis. Sei $r \in \mathbb{R}_{>0}$ so gewählt, dass

$$\Omega \subseteq \widehat{\Omega} := [-r, r]^d$$

gilt. Wir untersuchen zunächst ein beliebiges $u \in C_0^\infty(\Omega)$. Da u und seine Ableitungen auf dem gesamten Rand $\partial\Omega$ gleich null sein müssen, können wir die Funktion durch null auf den einschließenden Würfel $\widehat{\Omega}$ fortsetzen.

Für alle $x \in \Omega$ definieren wir $x_* \in \mathbb{R}^{d-1}$ so, dass

$$x = \begin{pmatrix} x_1 \\ x_* \end{pmatrix}$$

gilt, und wir setzen

$$\hat{x} := \begin{pmatrix} -r \\ x_* \end{pmatrix}$$

und folgern aus $x \in \Omega \subseteq \widehat{\Omega}$, dass auch $\hat{x} \in \widehat{\Omega}$ gelten muss. Aus unserer Voraussetzung folgt $u(\hat{x}) = 0$, und mit Hilfe des Hauptsatzes der Integral- und Differentialrechnung und der Cauchy-Schwarz-Ungleichung erhalten wir

$$\begin{aligned} u(x) &= u(\hat{x}) + \int_{-r}^{x_1} \frac{\partial u}{\partial y}(y, x_*)\,dy = \int_{-r}^{x_1} 1 \frac{\partial u}{\partial y}(y, x_*)\,dy \\ &\leq \left(\int_{-r}^{x_1} 1^2 \right)^{1/2} \left(\int_{-r}^{x_1} \left(\frac{\partial u}{\partial y}(y, x_*) \right)^2 dy \right)^{1/2} \\ &\leq \sqrt{2r} \left(\int_{-r}^{r} \left(\frac{\partial u}{\partial y}(y, x_*) \right)^2 dy \right)^{1/2}. \end{aligned}$$

Damit haben wir die Werte der Funktion u durch ihre Ableitung ausgedrückt, und indem wir über das gesamte Intervall integrieren folgt

$$\begin{aligned} \|u\|_{L^2}^2 &= \int_\Omega u(x)^2\,dx \leq 2r \int_\Omega \int_{-r}^{r} \left(\frac{\partial u}{\partial y}(y, x_*) \right)^2 dy\,dx \\ &= 2r \int_{\widehat{\Omega}} \int_{-r}^{r} \left(\frac{\partial u}{\partial y}(y, x_*) \right)^2 dy\,dx = 2r \int_{[-r,r]^{d-1}} \int_{-r}^{r} \int_{-r}^{r} \left(\frac{\partial u}{\partial y}(y, x_*) \right)^2 dy\,dx_1\,dx_* \\ &= (2r)^2 \int_{[-r,r]^{d-1}} \int_{-r}^{r} \left(\frac{\partial u}{\partial y}(y, x_*) \right)^2 dy\,dx_* = (2r)^2 \left\| \frac{\partial u}{\partial x_1} \right\|_{L^2}^2 \leq (2r)^2 |u|_{H^1}^2, \end{aligned}$$

so dass die Aussage für alle $u \in C_0^\infty(\Omega)$ mit der Konstanten $C_\Omega := 2r$ bewiesen ist.

Sei nun $u \in H_0^1(\Omega)$. Nach Definition existiert eine Folge $(u_\ell)_{\ell=1}^\infty$ in $C_0^\infty(\Omega)$, die in der H^1-Norm gegen u konvergiert. Sei $\epsilon \in \mathbb{R}_{>0}$, und sei $u_\ell \in C_0^\infty(\Omega)$ so gewählt, dass

$$\|u - u_\ell\|_{H^1} \leq \epsilon$$

gilt. Dann folgt

$$\begin{aligned} \|u\|_{L^2} &= \|u_\ell - u_\ell + u\|_{L^2} \leq \|u_\ell\|_{L^2} + \|u - u_\ell\|_{L^2} \leq 2r|u_\ell|_{H^1} + \epsilon \\ &= 2r|u - u + u_\ell|_{H^1} + \epsilon \leq 2r|u|_{H^1} + 2r|u - u_\ell|_{H^1} + \epsilon \leq 2r|u|_{H^1} + 2r\epsilon + \epsilon, \end{aligned}$$

und da ϵ beliebig gewählt war, muss die Behauptung auch für $u \in H_0^1(\Omega)$ gelten.

Die Definition der H^1-Norm besagte gerade

$$\|u\|_{H^1}^2 = \|u\|_{L^2}^2 + \sum_{\substack{\nu \in M_d \\ |\nu|=1}} \|\partial^\nu u\|_{L^2}^2 = \|u\|_{L^2}^2 + |u|_{H^1}^2 \qquad \text{für alle } u \in H^1(\Omega),$$

so dass die Normäquivalenz offensichtlich ist. ∎

Die H^1-Halbnorm ist für uns vor allem von Interesse, weil

$$|u|_{H^1}^2 = a(u,u) \qquad \text{für alle } u \in H_0^1(\Omega)$$

gilt, wir haben also soeben bewiesen, dass die von unserer Bilinearform a induzierte Norm zu der H^1-Norm äquivalent ist. Insbesondere muss dann $H_0^1(\Omega)$ auch bezüglich dieser Norm vollständig sein, also ein Hilbertraum.

6.3 Existenz und Eindeutigkeit

Wir haben gezeigt, dass $H_0^1(\Omega)$ mit dem durch a gegebenen Skalarprodukt ein Hilbertraum ist. Unsere Aufgabe besteht nun darin, zu zeigen, dass daraus bereits die Lösbarkeit unserer Variationsgleichung in diesem Raum folgt. Dazu benötigen wir einige allgemeine Aussagen über Hilberträume, die wir für einen beliebigen reellen Hilbertraum V mit dem Skalarprodukt $\langle\cdot,\cdot\rangle_V$ und der Norm $\|\cdot\|_V$ beweisen.

Lemma 6.9 (Grundlagen) *Seien $u, v \in V$ gegeben. Dann gelten die* Cauchy-Schwarz-Ungleichung

$$\langle u, v\rangle_V^2 \leq \|u\|_V^2 \|v\|_V^2 \tag{6.5}$$

und die Parallelogramm-Gleichung

$$\|u+v\|_V^2 + \|u-v\|_V^2 = 2(\|u\|_V^2 + \|v\|_V^2). \tag{6.6}$$

Beweis. Beide Aussagen ergeben sich aus der Beziehung zwischen Norm und Skalarprodukt. Im Falle der Cauchy-Schwarz-Ungleichung haben wir

$$\begin{aligned} 0 \leq \|u-\lambda v\|_V^2 &= \langle u-\lambda v, u-\lambda v\rangle_V = \langle u,u\rangle_V - \lambda\langle v,u\rangle_V - \lambda\langle u,v\rangle_V + \lambda^2\langle v,v\rangle_V \\ &= \|u\|_V^2 - 2\lambda\langle u,v\rangle_V + \lambda^2\|v\|_V^2. \end{aligned} \tag{6.7}$$

Da die Cauchy-Schwarz-Ungleichung für $v = 0$ trivial ist, brauchen wir nur den Fall $v \neq 0$ zu untersuchen. Wir wählen λ so, dass die rechte Seite der Ungleichung (6.7) minimal wird, also als

$$\lambda = \frac{\langle u,v\rangle_V}{\|v\|_V^2},$$

und erhalten

$$0 \leq \|u\|_V^2 - 2\frac{\langle u,v\rangle_V^2}{\|v\|_V^2} + \frac{\langle u,v\rangle_V^2}{\|v\|_V^2} = \|u\|_V^2 - \frac{\langle u,v\rangle_V^2}{\|v\|_V^2},$$

und durch Multiplikation mit $\|v\|_V^2$ folgt das gewünschte Ergebnis.

Die Parallelogramm-Gleichung lässt sich direkt nachrechnen: Es gilt

$$\begin{aligned} \|u+v\|_V^2 &= \langle u+v, u+v\rangle_V = \|u\|_V^2 + 2\langle u,v\rangle_V + \|v\|_V^2, \\ \|u-v\|_V^2 &= \langle u-v, u-v\rangle_V = \|u\|_V^2 - 2\langle u,v\rangle_V + \|v\|_V^2, \\ \|u+v\|_V^2 + \|u-v\|_V^2 &= 2\|u\|_V^2 + 2\|v\|_V^2. \end{aligned}$$

■

Definition 6.10 (Konvexe Menge) *Eine Menge $U \subseteq V$ heißt* konvex*, falls*

$$(1-\lambda)u + \lambda v \in U \qquad \text{für alle } u, v \in U,\ \lambda \in [0,1]$$

gilt, falls also die Verbindungslinie zweier Elemente der Menge vollständig in der Menge enthalten ist.

Lemma 6.11 (Bestapproximation) *Sei $U \subseteq V$ konvex, abgeschlossen und nicht leer, sei $w \in V$. Dann existiert genau ein $\widetilde{w} \in U$ mit*

$$\|w - \widetilde{w}\|_V \leq \|w - u\|_V \qquad \text{für alle } u \in U.$$

Damit ist $\widetilde{w}$ die beste Approximation des Vektors w in der Menge U.

Beweis. Da U nicht leer ist, ist

$$\delta := \inf\{\|w-u\|_V \ :\ u \in U\} \tag{6.8}$$

eine wohldefinierte reelle Zahl. Nach Definition des Infimums als größte untere Schranke der Menge können wir für jedes $\epsilon > 0$ ein $u \in U$ mit $\|w-u\|_V \leq \delta + \epsilon$ finden. Es stellt sich die Frage, ob wir einen Grenzwert für $\epsilon \to 0$ finden können.

Nehmen wir an, dass $u, v \in U$ Approximationen des Vektors w sind. Wir möchten den Abstand zwischen diesen beiden Vektoren abschätzen. Mit Hilfe der Parallelogramm-Gleichung erhalten wir

$$\begin{aligned}\|v-u\|_V^2 &= \|(w-u)-(w-v)\|_V^2 \\ &= 2(\|w-u\|_V^2 + \|w-v\|_V^2) - \|(w-u)+(w-v)\|_V^2.\end{aligned}$$

Für den letzten Term gilt

$$\|(w-u)+(w-v)\|_V^2 = \|2w - 2(u+v)/2\|_V^2 = 4\|w-(u+v)/2\|_V^2.$$

Da U konvex ist, muss $(u+v)/2 \in U$ gelten, also folgt mit Gleichung (6.8) die Abschätzung

$$\|(w-u)+(w-v)\|_V^2 = 4\|w-(u+v)/2\|_V^2 \geq 4\delta^2,$$

so dass wir insgesamt

$$\|v-u\|_V^2 \leq 2(\|w-u\|_V^2 + \|w-v\|_V^2) - 4\delta^2 \tag{6.9}$$

bewiesen haben.

Nach Definition des Infimums in (6.8) muss eine Folge $(w_n)_{n=0}^{\infty}$ mit

$$\|w-w_n\|_V^2 \leq \delta^2 + 2^{-n} \qquad \text{für alle } n \in \mathbb{N}_0 \tag{6.10}$$

existieren. Wir werden nun nachweisen, dass es sich dabei um eine Cauchy-Folge handelt. Sei $\epsilon \in \mathbb{R}_{>0}$ gegeben, und sei $n_0 \in \mathbb{N}_0$ so gewählt, dass $2^{-n_0} \leq \epsilon^2/4$ gilt. Dann folgt aus der Ungleichung (6.9)

$$\|w_n - w_m\|_V^2 \leq 2(\|w-w_n\|_V^2 + \|w-w_m\|_V^2) - 4\delta^2 \leq 2(\delta^2 + 2^{-n} + \delta^2 + 2^{-m}) - 4\delta^2$$

$$= 2(2^{-n} + 2^{-m}) \leq 2(2^{-n_0} + 2^{-n_0}) \leq 4\epsilon^2/4 = \epsilon^2 \qquad \text{für alle } n, m \in \mathbb{N}_{\geq n_0}.$$

Also ist $(w_n)_{n=0}^{\infty}$ eine Cauchy-Folge in der abgeschlossenen Teilmenge U des vollständigen Raums V und muss damit einen Grenzwert $\widetilde{w} \in U$ besitzen. Durch Grenzübergang $n \to \infty$ in (6.10) folgt

$$\|w - \widetilde{w}\|_V \leq \delta,$$

also nach (6.8) auch $\|w - \widetilde{w}\|_V = \delta$.

Zum Nachweis der Eindeutigkeit wählen wir ein $v \in U$ mit $\|w - v\|_V = \delta$ und folgern aus (6.9), dass

$$\|\widetilde{w} - v\|_V^2 = 2(\|w - \widetilde{w}\|_V^2 + \|w - v\|_V^2) - 4\delta^2 = 2(\delta^2 + \delta^2) - 4\delta^2 = 0$$

gelten muss, also $v = \widetilde{w}$. ∎

Lemma 6.12 (Approximation im Teilraum) *Sei $U \subseteq V$ ein Teilraum, sei $w \in V$ und $\widetilde{w} \in U$. Es gilt*

$$\|w - \widetilde{w}\|_V \leq \|w - u\|_V \qquad \textit{für alle } u \in U \tag{6.11}$$

genau dann, wenn

$$\langle v, w - \widetilde{w}\rangle_V = 0 \qquad \textit{für alle } v \in U \tag{6.12}$$

gilt.

Beweis. Bevor wir mit dem eigentlichen Beweis beginnen, stellen wir fest, dass für alle $v \in U$ und $\lambda \in \mathbb{R}$ die Gleichung

$$\begin{aligned} \|w - (\widetilde{w} + \lambda v)\|_V^2 &= \langle (w - \widetilde{w}) - \lambda v, (w - \widetilde{w}) - \lambda v\rangle_V \\ &= \|w - \widetilde{w}\|_V^2 - 2\lambda\langle v, w - \widetilde{w}\rangle_V + \lambda^2\|v\|_V^2 \end{aligned} \tag{6.13}$$

gilt. Diese Gleichung beschreibt, ob sich eine Näherung $\widetilde{w}$ verbessern lässt, indem man ein geeignetes Vielfaches eines Vektors $v \in U$ hinzuaddiert.

Gelte zunächst (6.12). Sei $u \in U$. Wir setzen in (6.13) $v := u - \widetilde{w}$ sowie $\lambda := 1$ ein, um

$$\begin{aligned} \|w - u\|_V^2 = \|w - (\widetilde{w} - v)\|_V^2 &= \|w - \widetilde{w}\|_V^2 - 2\langle v, w - \widetilde{w}\rangle_V + \|v\|_V^2 \\ &= \|w - \widetilde{w}\|_V^2 + \|v\|_V^2 \geq \|w - \widetilde{w}\|_V^2 \end{aligned}$$

zu erhalten. Daraus folgt (6.11).

Gelte nun (6.11). Sei $v \in U$. Für $v = 0$ ist (6.12) trivial, also beschränken wir uns auf den Fall $v \neq 0$. Wir minimieren die rechte Seite der Gleichung (6.13), indem wir

$$\lambda := \frac{\langle v, w - \widetilde{w}\rangle_V}{\|v\|_V^2}$$

setzen und folgern mit $u := \widetilde{w} + \lambda v \in U$ aus (6.11) die Ungleichung

$$\begin{aligned}\|w - \widetilde{w}\|_V^2 \leq \|w - u\|_V^2 &= \|w - (\widetilde{w} + \lambda v)\|_V^2 \\ &= \|w - \widetilde{w}\|_V^2 - 2\lambda\langle v, w - \widetilde{w}\rangle_V + \lambda^2\|v\|_V^2 \\ &= \|w - \widetilde{w}\|_V^2 - 2\frac{\langle v, w - \widetilde{w}\rangle_V^2}{\|v\|_V^2} + \frac{\langle v, w - \widetilde{w}\rangle_V^2}{\|v\|_V^4}\|v\|_V^2 \\ &= \|w - \widetilde{w}\|_V^2 - \frac{\langle v, w - \widetilde{w}\rangle_V^2}{\|v\|_V^2} \leq \|w - \widetilde{w}\|_V^2,\end{aligned}$$

also muss $\langle v, w - \widetilde{w}\rangle_V = 0$ gelten. ∎

Folgerung 6.13 (Lotfußpunkt) *Sei $U \subseteq V$ ein abgeschlossener Teilraum, sei $w \in V$. Dann existiert genau ein $\widetilde{w} \in U$ mit*

$$\langle v, w - \widetilde{w}\rangle_V = 0 \qquad \text{für alle } v \in U.$$

Beweis. Da U als Teilraum insbesondere konvex und nicht leer ist, existiert nach Lemma 6.11 genau ein $\widetilde{w} \in U$, das (6.11) erfüllt. Lemma 6.12 zufolge ist dieses $\widetilde{w}$ das einzige Element des Teilraums, das die Gleichung (6.12) erfüllt. ∎

Wenn wir uns die rechte Seite des Variationsproblems (6.3) ansehen, stellen wir fest, dass es sich um eine stetige Abbildung handelt, die jeder Testfunktion v einen Wert aus dem Körper zuordnet. Derartige Abbildungen bezeichnet man als *Funktionale*:

Definition 6.14 (Dualraum) *Eine stetige lineare Abbildung $\lambda : V \to \mathbb{R}$ bezeichnen wir als* Funktional*. Der Raum aller Funktionale*

$$V' := \{\lambda : V \to \mathbb{R} \ : \ \lambda \text{ stetig und linear}\}$$

heißt der Dualraum *des Raums V. Wir versehen ihn mit der durch*

$$\|\lambda\|_{V'} := \sup\left\{\frac{|\lambda(v)|}{\|v\|_V} \ : \ v \in V \setminus \{0\}\right\} \qquad \text{für alle } \lambda \in V'$$

definierten Dualnorm*.*

Für jedes beliebige $u \in V$ definiert

$$\lambda_u(v) := \langle v, u\rangle_V \qquad \text{für alle } v \in V$$

eine lineare Funktion, die dank der Cauchy-Schwarz-Ungleichung (6.5) auch

$$|\lambda_u(v)| = |\langle v, u\rangle_V| \leq \|v\|_V\|u\|_V \qquad \text{für alle } v \in V$$

erfüllt, also stetig ist. Damit gilt $\lambda_u \in V'$ mit $\|\lambda_u\|_{V'} \leq \|u\|_V$. Unser Ziel ist es nun, zu beweisen, dass *jedes* Funktional $\lambda \in V'$ auf diesem Weg konstruiert werden kann.

Satz 6.15 (Riesz) *Für jedes $\lambda \in V'$ existiert genau ein $u \in V$ mit*

$$\lambda(v) = \langle v, u\rangle_V \qquad \text{für alle } v \in V.$$

Für dieses u gilt $\|\lambda\|_{V'} = \|u\|_V$.

Diese Eigenschaft ist äquivalent dazu, dass der Riesz-Isomorphismus

$$\Psi_V : V \to V', \qquad u \mapsto \langle \cdot, u\rangle_V,$$

ein isometrischer Isomorphismus ist.

Beweis. Sei $\lambda \in V'$. Falls $\lambda = 0$ gilt, können wir $u = 0$ setzen und sind fertig.

Sei nun also $\lambda \neq 0$. Wir bezeichnen den Kern des Funktionals λ mit

$$U := \{v \in V \ : \ \lambda(v) = 0\} = \lambda^{-1}(\{0\})$$

und stellen fest, dass U als Urbild der abgeschlossenen Menge $\{0\}$ unter der stetigen Abbildung λ abgeschlossen sein muss.

Da $\lambda \neq 0$ gilt, können wir ein $w \in V$ mit $\lambda(w) \neq 0$ finden. Wie wir in Folgerung 6.13 gezeigt haben, gibt es eine Approximation $\widetilde{w} \in U$, die

$$\langle v, w - \widetilde{w}\rangle_V = 0 \qquad \text{für alle } v \in U \tag{6.14}$$

erfüllt. Den richtigen Kern scheint das Funktional $\langle \cdot, w - w_*\rangle_V$ also zu haben, jetzt müssen wir nur noch für die richtige Skalierung sorgen. Dazu setzen wir $u := \alpha(w - w_*)$ und wollen $\alpha \in \mathbb{R}$ so bestimmen, dass

$$\lambda(w) \overset{!}{=} \langle w, u\rangle_V = \alpha\langle w, w - \widetilde{w}\rangle_V = \alpha\langle w - \widetilde{w}, w - \widetilde{w}\rangle_V = \alpha\|w - \widetilde{w}\|_V^2$$

gilt, wobei die vorletzte Gleichung aus $\langle \widetilde{w}, w - \widetilde{w}\rangle_V = 0$ folgt. Wegen $\lambda(w) \neq 0$ gilt $w \notin U$, während wir $\widetilde{w} \in U$ nach Konstruktion haben, so dass insbesondere $w \neq \widetilde{w}$ und damit $\|w - \widetilde{w}\|_V \neq 0$ gilt. Demnach sind

$$\alpha := \frac{\lambda(w)}{\|w - \widetilde{w}\|_V^2}, \qquad u := \frac{\lambda(w)}{\|w - \widetilde{w}\|_V^2}(w - \widetilde{w}) \neq 0$$

wohldefiniert. Nun müssen wir lediglich nachprüfen, dass dieses u unseren Wünschen entspricht.

Wir stellen zunächst fest, dass aus $\widetilde{w} \in U$ bereits $\lambda(\widetilde{w}) = 0$ folgt und wir

$$\begin{aligned}\lambda(u) &= \frac{\lambda(w)}{\|w - \widetilde{w}\|_V^2}\lambda(w - \widetilde{w}) = \frac{\lambda(w)}{\|w - \widetilde{w}\|_V^2}\lambda(w) = \frac{\lambda(w)^2}{\|w - \widetilde{w}\|_V^2} \\ &= \frac{\lambda(w)^2}{\|w - \widetilde{w}\|_V^4}\|w - \widetilde{w}\|_V^2 = \left(\frac{\lambda(w)}{\|w - \widetilde{w}\|_V^2}\|w - \widetilde{w}\|_V\right)^2 = \|u\|_V^2 \neq 0\end{aligned} \tag{6.15}$$

erhalten. Sei nun ein $v \in V$ fixiert. Wir zerlegen es in einen Anteil in Richtung des Vektors u und einen Rest

$$v_0 := v - \frac{\lambda(v)}{\lambda(u)}u,$$

der wegen

$$\lambda(v_0) = \lambda(v) - \frac{\lambda(v)}{\lambda(u)}\lambda(u) = \lambda(v) - \lambda(v) = 0$$

im Kern U des Funktionals enthalten ist. Mit (6.14) sowie (6.15) folgt

$$\langle v, u\rangle_V = \langle v_0 + \frac{\lambda(v)}{\lambda(u)}u, u\rangle_V = \langle v_0, u\rangle_V + \frac{\lambda(v)}{\lambda(u)}\langle u, u\rangle_V = \frac{\lambda(v)}{\lambda(u)}\|u\|_V^2 = \frac{\lambda(v)}{\lambda(u)}\lambda(u) = \lambda(v).$$

Damit besitzt u die geforderte Eigenschaft.

Wie bereits gesehen folgt mit der Cauchy-Schwarz-Ungleichung (6.5) die Ungleichung

$$|\lambda(v)| = |\langle v, u\rangle_V| \leq \|v\|_V\|u\|_V \qquad \text{für alle } v \in V,$$

also insbesondere $\|\lambda\|_{V'} \leq \|u\|_V$. Aus (6.15) folgt $\|\lambda\|_{V'} \geq \|u\|_V$, also haben wir die Gleichung $\|\lambda\|_{V'} = \|u\|_V$ bewiesen.

Zu zeigen bleibt noch die Eindeutigkeit des Vektors u. Sei dazu ein Vektor $\widehat{u} \in V$ gegeben, der ebenfalls

$$\langle v, \widehat{u}\rangle_V = \lambda(v) \qquad \text{für alle } v \in V$$

erfüllt. Indem wir $v := u - \widehat{u}$ einsetzen, folgt

$$0 = \lambda(v) - \lambda(v) = \langle v, u\rangle_V - \langle v, \widehat{u}\rangle_V = \langle v, u - \widehat{u}\rangle_V = \langle u - \widehat{u}, u - \widehat{u}\rangle_V = \|u - \widehat{u}\|_V^2,$$

also $u = \widehat{u}$. ∎

Mit Hilfe des Darstellungssatzes 6.15 können wir nun Existenz und Eindeutigkeit einer Lösung des Variationsproblems untersuchen. Wir setzen dazu die in (6.4) definierte Bilinearform auf den Sobolew-Raum $H_0^1(\Omega)$ fort, indem wir die partiellen Ableitungen durch ihre schwachen Gegenstücke ersetzen:

$$a : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}, \qquad (v, u) \mapsto \int_\Omega \langle \nabla v(x), \nabla u(x)\rangle_2 \, dx. \tag{6.16}$$

Aus der Friedrichs-Ungleichung (6.8) folgt, dass

$$|u|_{H^1} = \sqrt{a(u, u)} \qquad \text{für alle } u \in H_0^1(\Omega)$$

eine Norm auf dem Raum $H_0^1(\Omega)$ ist, also bildet dieser Raum mit dem Skalarprodukt a einen Hilbertraum. Um Satz 6.15 anwenden zu können, schreiben wir die rechte Seite des Variationsproblems (6.3) in der Form eines Funktionals

$$\lambda : H_0^1(\Omega) \to \mathbb{R}, \qquad v \mapsto \int_\Omega v(x) f(x) \, dx,$$

und erhalten die folgende *schwache Formulierung* des Variationsproblems:

Wir suchen eine Funktion $u \in V := H_0^1(\Omega)$, die

$$a(v,u) = \lambda(v) \qquad \text{für alle } v \in V \tag{6.17}$$

erfüllt.

Der Darstellungssatz lässt sich direkt anwenden:

Folgerung 6.16 (Existenz und Eindeutigkeit) *Sei λ ein stetiges Funktional. Dann besitzt die schwache Formulierung (6.17) des Variationsproblems genau eine Lösung $u \in H_0^1(\Omega)$.*

Beweis. Aus Lemma 6.8 folgt, dass a auf dem Raum $V = H_0^1(\Omega)$ ein Skalarprodukt ist, dessen Norm zu der Sobolew-Norm äquivalent ist. Also ist V auch mit dem Skalarprodukt a ein Hilbertraum. Mit Satz 6.15 erhalten wir die Existenz und Eindeutigkeit einer Lösung $u \in V$ des Problems (6.17). ∎

In der Praxis ist man häufig daran interessiert, die Bilinearform a lediglich für die Definition des Variationsproblems einzusetzen, aber für Stabilitäts- und Fehlerabschätzungen mit der üblichen Sobolew-Norm $\|\cdot\|_{H^m}$ anstelle der von a induzierten Norm zu arbeiten. Die Beziehung zwischen beiden Normen beschreibt man dabei in der Regel durch die Begriffe der Stetigkeit und der *Elliptizität*:

Definition 6.17 (Stetigkeit und Elliptizität) *Eine Bilinearform a auf einem Hilbertraum V heißt* stetig, *falls eine Konstante $C_S \in \mathbb{R}_{\geq 0}$ existiert, die*

$$|a(v,u)| \leq C_S \|v\|_V \|u\|_V \qquad \textit{für alle } v,u \in V \tag{6.18}$$

erfüllt. Die Bilinearform heißt elliptisch, *falls sie stetig ist und eine Konstante $C_E \in \mathbb{R}_{\geq 0}$ existiert, die*

$$a(u,u) \geq C_E \|u\|_V^2 \qquad \textit{für alle } u \in V \tag{6.19}$$

erfüllt. Wir bezeichnen C_S und C_E als die Stetigkeits- *beziehungsweise* Elliptizitätskonstante *der Bilinearform.*

Beispiel 6.18 (Potentialgleichung) *Wir untersuchen die durch (6.16) definierte Bilinearform a, die bei der Behandlung der Potentialgleichung auftritt.*

Indem wir die Cauchy-Schwarz-Ungleichung erst auf das euklidische und dann auf das L^2-Skalarprodukt anwenden, folgt

$$\begin{aligned} |a(v,u)| &= \left| \int_\Omega \langle \nabla v(x), \nabla u(x) \rangle_2 \, dx \right| \leq \int_\Omega \|\nabla v(x)\|_2 \|\nabla u(x)\|_2 \, dx \\ &\leq \left(\int_\Omega \|\nabla v(x)\|_2^2 \, dx \right)^{1/2} \left(\int_\Omega \|\nabla u(x)\|_2^2 \, dx \right)^{1/2} \end{aligned}$$

$$
\begin{aligned}
&= \left(\sum_{\substack{\nu \in M_d \\ |\nu|=1}} \|\partial^\nu v(x)\|_2^2 \, dx \right)^{1/2} \left(\sum_{\substack{\nu \in M_d \\ |\nu|=1}} \|\partial^\nu u(x)\|_2^2 \, dx \right)^{1/2} \\
&\leq \left(\sum_{\substack{\nu \in M_d \\ |\nu|\leq 1}} \|\partial^\nu v(x)\|_2^2 \, dx \right)^{1/2} \left(\sum_{\substack{\nu \in M_d \\ |\nu|\leq 1}} \|\partial^\nu u(x)\|_2^2 \, dx \right)^{1/2} \\
&= \|v\|_{H^1} \|u\|_{H^1} \qquad \textit{für alle } v, u \in H_0^1(\Omega),
\end{aligned}
$$

also ist a stetig mit der Stetigkeitskonstanten $C_S = 1$.

Die Bilinearform besitzt auch die Eigenschaft, dass

$$
\begin{aligned}
a(u,u) &= \int_\Omega \langle \nabla u(x), \nabla u(x) \rangle_2 \, dx = \int_\Omega \|\nabla u(x)\|_2^2 \, dx \\
&= \sum_{\substack{\nu \in M_d \\ |\nu|=1}} \|\partial^\nu u(x)\|_2^2 \, dx = |u|_{H^1}^2 \qquad \textit{für alle } u \in H_0^1(\Omega)
\end{aligned}
$$

gilt. Mit der Konstanten $C_\Omega \in \mathbb{R}_{\geq 0}$ aus Lemma 6.8 folgt

$$
\begin{aligned}
\|u\|_{H^1}^2 &= \|u\|_{L^2}^2 + |u|_{H^1}^2 \leq C_\Omega^2 |u|_{H^1}^2 + |u|_{H^1}^2 \\
&= (C_\Omega^2 + 1)|u|_{H^1}^2 = (C_\Omega^2 + 1) a(u,u) \qquad \textit{für alle } u \in H_0^1(\Omega),
\end{aligned}
$$

also ist a auch elliptisch mit der Elliptizitätskonstanten $C_E = 1/(C_\Omega^2 + 1)$.

Lemma 6.19 (Lax-Milgram) *Sei $a : V \times V \to \mathbb{R}$ eine symmetrische elliptische Bilinearform, und sei $\lambda \in V'$ ein stetiges Funktional. Dann existiert genau ein $u \in V$ mit*

$$
a(v,u) = \lambda(v) \qquad \textit{für alle } v \in V, \tag{6.20}
$$

und dieses Element erfüllt die Stabilitätsabschätzung

$$
\|u\|_V \leq \frac{1}{C_E} \|\lambda\|_{V'}.
$$

Beweis. Wir definieren die *Energienorm* durch

$$
\|u\|_a := \sqrt{a(u,u)} \qquad \text{für alle } u \in V \tag{6.21}
$$

und stellen fest, dass aus der Stetigkeit

$$
\|u\|_a = \sqrt{a(u,u)} \leq \sqrt{C_S \|u\|_V^2} = C_S^{1/2} \|u\|_V \qquad \text{für alle } u \in V
$$

und aus der Elliptizität

$$
\|u\|_a = \sqrt{a(u,u)} \geq \sqrt{C_E \|u\|_V^2} = C_E^{1/2} \|u\|_V \qquad \text{für alle } u \in V \tag{6.22}
$$

folgen, also ist $\|\cdot\|_a$ äquivalent zu $\|\cdot\|_V$. Damit ist V auch mit dem Skalarprodukt a ein Hilbertraum, und wir können den Satz 6.15 von Riesz anwenden, um genau ein u mit

$$a(v,u) = \lambda(v) \qquad \text{für alle } v \in V \tag{6.23}$$

zu finden. Es bleibt nur noch die Stabilitätsabschätzung zu zeigen. Für $u = 0$ ist sie trivial, also konzentrieren wir uns auf den Fall $u \neq 0$. Wir setzen $v = u$ in die Gleichung (6.23) ein und erhalten dank (6.22)

$$\|u\|_V^2 \leq \frac{1}{C_E}\|u\|_a^2 = \frac{1}{C_E}a(u,u) = \frac{1}{C_E}\lambda(u) \leq \frac{1}{C_E}\frac{|\lambda(u)|}{\|u\|_V}\|u\|_V \leq \frac{1}{C_E}\|\lambda\|_{V'}\|u\|_V.$$

Da wir $u \neq 0$ vorausgesetzt haben, können wir auf beiden Seiten durch $\|u\|_V$ dividieren und erhalten die gewünschte Abschätzung. ∎

Bemerkung 6.20 (Ladyschenskaja-Babuška-Brezzi-Bedingung) *Die Voraussetzung der Symmetrie der Bilinearform a ist nicht für die Existenz einer Lösung des Variationsproblems erforderlich. Wir können allgemein den Operator*

$$A : V \to V', \qquad u \mapsto a(\cdot, u),$$

einführen und das Variationsproblem (6.20) in der äquivalenten Form einer Gleichung

$$Au = \lambda$$

in dem Dualraum V' formulieren. Falls a stetig ist, folgt aus (6.18) unmittelbar, dass der Operator A ebenfalls stetig ist. Wir interessieren uns für die Frage, ob A eine stetige Inverse besitzt.

Die Ladyschenskaja-Babuška-Brezzi-Bedingung (LBB-Bedingung) *(gelegentlich auch kurz als* inf-sup-Bedingung *bezeichnet)*

$$\inf_{u \in V\setminus\{0\}} \sup_{v \in V\setminus\{0\}} \frac{a(v,u)}{\|v\|_V\|u\|_V} \geq \gamma, \tag{6.24}$$

ist äquivalent zu

$$\|Au\|_{V'} \geq \gamma\|u\|_V \qquad \textit{für alle } u \in V, \tag{6.25}$$

impliziert also insbesondere die Injektivität des Operators A. Falls A stetig ist, kann man mit einem Cauchy-Folgen-Argument relativ einfach aus (6.25) folgern, dass das Bild von A ein abgeschlossener Teilraum des Dualraums V' ist.

Es ist noch zu klären, ob A auch surjektiv ist. Nach dem Satz 6.15 von Riesz ist das äquivalent dazu, dass $\Psi_V^{-1}A$ surjektiv ist. Da das Bild von A abgeschlossen und Ψ_V ein isometrischer Isomorphismus ist, ist auch das Bild von $\Psi_V^{-1}A$ abgeschlossen. Wäre dieses Bild nicht der gesamte Raum V, könnte man mit Folgerung 6.13 einen Vektor $v \in V \setminus \{0\}$ konstruieren, der senkrecht darauf steht. Um das auszuschließen fordern wir

$$\sup_{u \in V\setminus\{0\}} \frac{a(v,u)}{\|u\|_V} > 0 \qquad \textit{für alle } v \in V \setminus \{0\}.$$

In Kombination mit (6.24) folgt aus dieser Bedingung die Existenz des inversen Operators $A^{-1} : V' \to V$, und durch Einsetzen in (6.25) erhalten wir, dass er stetig ist mit $\|A^{-1}\|_{V\to V'} \leq 1/\gamma$.

Für elliptische Bilinearformen sind unsere Bedingungen mit $\gamma = C_E$ offenbar erfüllt.

6.4 Galerkin-Verfahren

Da wir nun wissen, dass die schwache Form (6.20) des Variationsproblems eine Lösung besitzt, stellt sich die Frage, wie sich diese Lösung, wenigstens approximativ, berechnen lässt.

Die Idee des *Galerkin-Verfahrens* besteht darin, einen endlich-dimensionalen Teilraum $V_h \subseteq V$ zu wählen und nach einer Lösung des folgenden Variationsproblems zu suchen:

> Wir suchen eine Funktion $u_h \in V_h$, die
>
> $$a(v_h, u_h) = \lambda(v_h) \qquad \text{für alle } v_h \in V_h \tag{6.26}$$
>
> erfüllt.

Die Schreibweise u_h für die Lösung ist dadurch begründet, dass in der Regel der Raum V_h mit Hilfe eines Gitters mit einer (geeignet verallgemeinerten) Schrittweite h konstruiert wird.

Falls a elliptisch ist, folgt aus $V_h \subseteq V$ und dem bereits bekannten Lax-Milgram-Lemma 6.19 direkt, dass auch das auf den Raum V_h eingeschränkte Variationsproblem genau eine Lösung besitzen muss.

Um diese Lösung zu berechnen, bietet es sich an, eine Basis $(\varphi_i)_{i\in\mathcal{I}}$ des Raums V_h zu wählen. Dabei ist $\mathcal{I}$ eine Indexmenge, deren Mächtigkeit gerade der Dimension des Raums entspricht. Für jedes $i \in \mathcal{I}$ ist $\varphi_i \in V_h$ dann eine Funktion aus einem passenden Sobolew-Raum, beispielsweise im Fall unseres Modellproblems aus $H_0^1(\Omega)$.

Die gesuchte Lösung $u_h \in V_h$ können wir in der Basis durch einen Koeffizientenvektor $\mathbf{x} \in \mathbb{R}^{\mathcal{I}}$ darstellen, also in der Form

$$u_h = \sum_{j\in\mathcal{I}} \varphi_j x_j. \tag{6.27}$$

Indem wir mit den Testfunktionen entsprechend verfahren, erhalten wir das folgende Resultat:

Lemma 6.21 (Lineares Gleichungssystem) *Der die Lösung $u_h \in V_h$ des Variationsproblems (6.26) gemäß (6.27) beschreibende Koeffizientenvektor $\mathbf{x} \in \mathbb{R}^{\mathcal{I}}$ ist die Lösung des linearen Gleichungssystems*

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{6.28}$$

für die durch

$$a_{ij} := a(\varphi_i, \varphi_j) \qquad \textit{für alle } i, j \in \mathcal{I} \tag{6.29}$$

definierte Matrix $\mathbf{A} \in \mathbb{R}^{\mathcal{I}\times\mathcal{I}}$ *und den durch*

$$b_i := \lambda(\varphi_i) \qquad \text{für alle } i \in \mathcal{I}$$

definierten Vektor $\mathbf{b} \in \mathbb{R}^{\mathcal{I}}$.

Beweis. Sei $\mathbf{x} \in \mathbb{R}^{\mathcal{I}}$ gemäß (6.27) definiert, und sei $i \in \mathcal{I}$. Dann gilt

$$(\mathbf{Ax})_i = \sum_{j\in\mathcal{I}} a_{ij}x_j = \sum_{j\in\mathcal{I}} a(\varphi_i,\varphi_j)x_j = a\left(\varphi_i, \sum_{j\in\mathcal{I}} \varphi_j x_j\right) = a(\varphi_i, u_h) = \lambda(\varphi_i) = b_i,$$

also ist $\mathbf{x}$ auch Lösung des linearen Gleichungssystems (6.28).

Sei nun $\mathbf{x} \in \mathbb{R}^{\mathcal{I}}$ eine Lösung dieses Gleichungssystems, und sei $u_h \in V_h$ gemäß (6.27) definiert. Sei $v_h \in V_h$. Da $(\varphi_i)_{i\in\mathcal{I}}$ eine Basis des Raums V_h ist, existiert ein Koeffizientenvektor $\mathbf{y} \in \mathbb{R}^{\mathcal{I}}$ mit

$$v_h = \sum_{i\in\mathcal{I}} \varphi_i y_i,$$

und wir erhalten

$$\begin{aligned} a(v_h,u_h) &= \sum_{i,j\in\mathcal{I}} y_i a(\varphi_i,\varphi_j)x_j = \sum_{i,j\in\mathcal{I}} y_i a_{ij}x_j = \sum_{i\in\mathcal{I}} y_i \sum_{j\in\mathcal{I}} a_{ij}x_j = \sum_{i\in\mathcal{I}} y_i(\mathbf{Ax})_i \\ &= \sum_{i\in\mathcal{I}} y_i b_i = \sum_{i\in\mathcal{I}} y_i\lambda(\varphi_i) = \lambda\left(\sum_{i\in\mathcal{I}} \varphi_i y_i\right) = \lambda(v_h). \end{aligned}$$

Da wir diese Gleichung für beliebige $v_h \in V_h$ bewiesen haben, muss u_h eine Lösung des diskreten Variationsproblems (6.26) sein. ∎

Durch die Wahl einer Basis haben wir also das Variationsproblem auf ein äquivalentes lineares Gleichungssystem zurückgeführt, und Existenz und Eindeutigkeit der Lösung beider Probleme sind direkt aneinander gekoppelt.

Bemerkung 6.22 (Positiv definit) *Falls a symmetrisch ist, gilt*

$$a_{ij} = a(\varphi_i,\varphi_j) = a(\varphi_j,\varphi_i) = a_{ji} \qquad \text{für alle } i,j \in \mathcal{I},$$

also ist die Matrix $\mathbf{A}$ *symmetrisch.*

Falls a elliptisch ist, gilt für jeden Vektor $\mathbf{x} \in \mathbb{R}^{\mathcal{I}}$ *die Gleichung*

$$\begin{aligned} \langle \mathbf{x}, \mathbf{Ax}\rangle_2 &= \sum_{i,j\in\mathcal{I}} x_i a_{ij} x_j = \sum_{i,j\in\mathcal{I}} x_i a(\varphi_i,\varphi_j)x_j \\ &= a\left(\sum_{i\in\mathcal{I}} \varphi_i x_i, \sum_{j\in\mathcal{I}} \varphi_j x_j\right) = a(u_h,u_h) \geq C_E\|u_h\|_V^2 \end{aligned}$$

mit der gemäß (6.27) definierten Funktion u_h. *Also ist* $\mathbf{A}$ *positiv definit und damit insbesondere regulär.*

Mit Hilfe des Gleichungssystems (6.28) können wir die Lösung des diskreten Variationsproblems (6.26) praktisch berechnen, sofern uns eine geeignete Basis $(\varphi_i)_{i\in\mathcal{I}}$ zur Verfügung steht, also bleibt nur noch die Frage zu klären, wie sich die Näherungslösung $u_h \in V_h$ des Problems (6.26) zu der Lösung $u \in V$ des ursprünglichen Problems (6.20) verhält. Diese Frage lässt sich beantworten, indem wir in (6.20) eine Testfunktion $v_h \in V_h \subseteq V$ einsetzen: Es gelten

$$a(v_h, u) = f(v_h), \qquad a(v_h, u_h) = f(v_h) \qquad \text{für alle } v_h \in V_h,$$

und indem wir beide Gleichungen subtrahieren folgt

$$a(v_h, u - u_h) = 0 \qquad \text{für alle } v_h \in V_h. \tag{6.30}$$

Diese Beziehung ist unter dem Namen *Galerkin-Orthogonalität* bekannt. Ein Vergleich mit Lemma 6.12 zeigt, dass u_h gerade die beste Approximation der Lösung u bezüglich des a-Skalarprodukts ist. Indem wir wieder die Sobolew-Norm anstelle der problemspezifischen Norm einsetzen, erhalten wir die folgende Abschätzung:

Lemma 6.23 (Céa) *Sei $a : V \times V \to \mathbb{R}$ eine symmetrische elliptische Bilinearform, und sei $\lambda \in V'$ ein stetiges Funktional. Sei $u \in V$ die Lösung des Variationsproblems (6.20), und sei $u_h \in V_h$ die Lösung des diskreten Variationsproblems (6.26). Dann gilt*

$$\|u - u_h\|_V \leq \frac{C_S}{C_E}\|u - v_h\|_V \qquad \textit{für alle } v_h \in V_h,$$

der Fehler der Lösung des diskreten Variationsproblems kann also durch den Fehler der bestmöglichen Approximation der Lösung u in dem Raum u_h abgeschätzt werden.

Beweis. Sei $v_h \in V_h$. Wir erhalten

$$\begin{aligned}\|u - u_h\|_V^2 &\leq \frac{1}{C_E}a(u - u_h, u - u_h) = \frac{1}{C_E}a(u - v_h + v_h - u_h, u - u_h)\\ &= \frac{1}{C_E}a(u - v_h, u - u_h) + \frac{1}{C_E}a(v_h - u_h, u - u_h).\end{aligned}$$

Da $v_h - u_h \in V_h$ gilt, ist dank der Galerkin-Orthogonalität (6.30) der zweite Summand gleich null, so dass nur

$$\|u - u_h\|_V^2 \leq \frac{1}{C_E}a(u - v_h, u - u_h)$$

bleibt. Aufgrund der Stetigkeit der Bilinearform folgt

$$\|u - u_h\|_V^2 \leq \frac{1}{C_E}a(u - v_h, u - u_h) \leq \frac{C_S}{C_E}\|u - v_h\|_V\|u - u_h\|_V,$$

und indem wir bei Bedarf durch $\|u - u_h\|_V$ dividieren, erhalten wir das gewünschte Ergebnis. ∎

6.5 Interpretation als Minimierungsproblem

Falls die Bilinearform symmetrisch und positiv definit ist, falls also

$$a(u,v) = a(v,u) \qquad \text{für alle } u,v \in V$$

sowie

$$a(u,u) > 0 \qquad \text{für alle } u \in V \setminus \{0\}$$

gelten, lässt sich das Galerkin-Verfahren auch als Minimierungsproblem interpretieren: Jedem Element des Hilbertraums V ordnen wir eine *Energie* zu:

$$g : V \to \mathbb{R}, \qquad u \mapsto \frac{1}{2}a(u,u) - \lambda(u).$$

Der Name „Energie“ liegt darin begründet, dass g im Fall der Potentialgleichung in Bezug zu der physikalischen Energie des elektrostatischen Felds steht.

Satz 6.24 (Energieminimierung) *Sei a symmetrisch und positiv definit. $u \in V$ ist genau dann Lösung der Variationsaufgabe*

$$a(v,u) = \lambda(v) \qquad \textit{für alle } v \in V, \tag{6.31}$$

wenn es ein globales Minimum der Energie ist, falls also

$$g(u) \leq g(w) \qquad \textit{für alle } w \in V \tag{6.32}$$

gilt.

Beweis. Der Beweis verläuft analog zu dem des Lemmas 6.9.

Seien $u, v \in V$ sowie $\alpha \in \mathbb{R}$ fixiert. Da a eine symmetrische Bilinearform ist, gilt

$$\begin{aligned} g(u-\alpha v) &= \frac{1}{2}a(u-\alpha v, u-\alpha v) - \lambda(u-\alpha v) \\ &= \frac{1}{2}a(u,u) - \frac{\alpha}{2}a(v,u) - \frac{\alpha}{2}a(u,v) + \frac{\alpha^2}{2}a(v,v) - \lambda(u) + \alpha\lambda(v) \\ &= g(u) - \alpha(a(v,u) - \lambda(v)) + \frac{\alpha^2}{2}a(v,v). \end{aligned} \tag{6.33}$$

Sei nun zunächst u Lösung der Variationsaufgabe (6.31). Sei $w \in V$. Wir wenden (6.33) auf $\alpha = 1$ und $v = u - w$ an und erhalten

$$g(w) = g(u-\alpha v) = g(u) + \frac{\alpha^2}{2}a(v,v) \geq g(u),$$

da a positiv definit ist und deshalb $a(v,v) \geq 0$ gilt.

Gelte nun umgekehrt (6.32). Sei $v \in V$. Für $v = 0$ folgt (6.31) sofort, sei also im Folgenden $v \neq 0$. Um den größtmöglichen Nutzen aus (6.33) ziehen zu können, wählen

wir α so, dass die rechte Seite möglichst klein wird. Durch Kurvendiskussion ergibt sich, dass das Minimum für

$$\alpha := \frac{a(v,u) - \lambda(v)}{a(v,v)}$$

angenommen wird. Indem wir (6.33) auf $w = u - \alpha v$ anwenden, erhalten wir

$$g(u) \leq g(u - \alpha v) = g(u) - \frac{(a(v,u) - \lambda(v))^2}{2a(v,v)} \leq g(u)$$

und folgern, dass $a(v,u) = \lambda(v)$ gelten muss. ∎

Die Idee des Galerkin-Verfahrens besteht einfach darin, das Minimum der Energie g nicht in V, sondern in dem Teilraum V_h zu suchen. Offenbar kann das Minimum der Energie auf $V_h \subseteq V$ nicht kleiner als das Minimum auf V sein.

Der Unterschied der Energien für u und u_h lässt sich als Maß für die Genauigkeit der Approximation verwenden, denn da beide Lösungen der entsprechenden Variationsprobleme sind, gilt

$$\begin{aligned} g(u_h) - g(u) &= \frac{1}{2}a(u_h,u_h) - \lambda(u_h) - \frac{1}{2}a(u,u) + \lambda(u) \\ &= \frac{1}{2}(a(u_h,u_h) - \lambda(u_h)) - \frac{1}{2}\lambda(u_h) - \frac{1}{2}a(u,u) + a(u,u) \\ &= -\frac{1}{2}a(u_h,u) + \frac{1}{2}a(u,u) = \frac{1}{2}a(u-u_h,u), \end{aligned}$$

und mit Hilfe der Galerkin-Orthogonalität (6.30) folgt

$$g(u_h) - g(u) = \frac{1}{2}a(u-u_h,u) = \frac{1}{2}a(u-u_h,u-u_h).$$

Die Differenz der Energien entspricht also gerade dem halben Quadrat der Energienorm $\|u-u_h\|_a$ des Fehlers (vgl. (6.21)).

6.6 Eindimensionale finite Elemente

Das Céa-Lemma besagt, dass wir auf eine gute Näherung der Lösung u des Variationsproblems (6.20) hoffen dürfen, falls wir einen endlich-dimensionalen Raum V_h finden können, in dem so eine Näherung existiert. Aus einer Existenzaussage erhalten wir also eine Approximationsaussage.

Unser Ziel ist es nun, einen geeigneten Raum V_h zu konstruieren. Zur Motivation untersuchen wir zunächst den eindimensionalen Fall: Für ein Intervall $\Omega = (a,b)$ konstruieren wir einen endlich-dimensionalen Teilraum V_h des Sobolew-Raums $V = H_0^1(\Omega)$. Ein einfacher Ansatz besteht darin, das Intervall (a,b) in gleich große Teilintervalle zu zerlegen: Wir wählen $n \in \mathbb{N}$ und setzen

$$h := \frac{b-a}{n+1}, \qquad x_i := a + ih \qquad \text{für alle } i \in \{0,\ldots,n+1\}.$$

Wir definieren eine Vorstufe $\widehat{V}_h$ des gesuchten Raums, indem wir fordern, dass eine Funktion $u_h \in \widehat{V}_h$ auf jedem Teilintervall (x_i, x_{i+1}) ein Polynom der Ordnung $m \in \mathbb{N}$ sein muss:

$$\widehat{V}_h := \{u \in L^2(\bar{\Omega}) \ : \ u|_{(x_i,x_{i+1})} \in \Pi_m \text{ für alle } i \in \{0,\dots,n\}\}.$$

Unser Ziel ist es, einen Teilraum $V_h \subseteq \widehat{V}_h \cap H_0^1(\Omega)$ zu konstruieren. Dazu müssen wir die Frage untersuchen, welche Funktionen des Raums $\widehat{V}_h$ eine schwache Ableitung besitzen.

Seien dazu ein $v_h \in \widehat{V}_h$ und eine Testfunktion $\varphi \in C_0^\infty(\Omega)$ gewählt. Sei $i \in \{0,\dots,n\}$. Da $v_h|_{(x_i,x_{i+1})}$ ein Polynom und damit insbesondere stetig differenzierbar ist, erhalten wir durch partielle Integration

$$\begin{aligned}-\int_{x_i+\epsilon}^{x_{i+1}-\epsilon} \varphi'(x)v_h(x)\,dx &= \varphi(x_{i+1}-\epsilon)v_h(x_{i+1}-\epsilon) - \varphi(x_i+\epsilon)v_h(x_i+\epsilon)\\ &\quad + \int_{x_i+\epsilon}^{x_{i+1}-\epsilon} \varphi(x)v_h'(x)\,dx \qquad \text{für alle } \epsilon \in (0,(x_{i+1}-x_i)/2).\end{aligned}$$

Durch Grenzübergang folgt

$$\begin{aligned}-\int_{x_i}^{x_{i+1}} \varphi'(x)v_h(x)\,dx &= \varphi(x_{i+1})\lim_{\epsilon\to 0} v_h(x_{i+1}-\epsilon) - \varphi(x_i)\lim_{\epsilon\to 0} v_h(x_i+\epsilon)\\ &\quad + \int_{x_i}^{x_{i+1}} \varphi(x)v_h'(x)\,dx.\end{aligned}$$

Falls wir die Randterme in den Griff bekommen könnten, wäre also eine Funktion $w \in L^2(\Omega)$ mit

$$w|_{(x_i,x_{i+1})} = v_h'|_{(x_i,x_{i+1})} \qquad \text{für alle } i \in \{1,\dots,n\}$$

eine naheliegende Wahl für die schwache Ableitung der Funktion v_h.

Wir erhalten

$$\begin{aligned}-\int_\Omega \varphi'(x)v_h(x)\,dx &= \sum_{i=1}^n -\int_{x_i}^{x_{i+1}} \varphi'(x)v_h(x)\,dx\\ &= \sum_{i=1}^n \varphi(x_{i+1})\lim_{\epsilon\to 0} v_h(x_{i+1}-\epsilon) - \varphi(x_i)\lim_{\epsilon\to 0} v_h(x_i+\epsilon)\\ &\quad + \sum_{i=1}^n \int_{x_i}^{x_{i+1}} \varphi(x)w(x)\,dx\\ &= \varphi(b)\lim_{\epsilon\to 0} v_h(b-\epsilon) - \varphi(a)\lim_{\epsilon\to 0} v_h(a+\epsilon)\\ &\quad + \sum_{i=1}^{n-1} \varphi(x_{i+1})(\lim_{\epsilon\to 0} v_h(x_{i+1}-\epsilon) - v_h(x_{i+1}+\epsilon)) \qquad (6.34)\\ &\quad + \int_\Omega \varphi(x)w(x)\,dx.\end{aligned}$$

Da der Träger der Testfunktion φ kompakt in der offenen Menge $\Omega = (a, b)$ ist, muss $\varphi(a) = 0 = \varphi(b)$ gelten. Wenn wir den *Sprung* der Funktion v_h mit

$$[v_h](x) := \lim_{\epsilon \to 0} v_h(x - \epsilon) - \lim_{\epsilon \to 0} v_h(x + \epsilon) \qquad \text{für alle } x \in \Omega$$

bezeichnen, nimmt die Gleichung (6.34) die Form

$$-\int_\Omega \varphi'(x) v_h(x)\, dx = \int_\Omega \varphi(x) w(x)\, dx + \sum_{i=1}^{n-1} \varphi(x_{i+1})[v_h](x_{i+1})$$

an. Also kann w nur dann die schwache Ableitung der Funktion v_h sein, wenn

$$[v_h](x_{i+1}) = 0 \qquad \text{für alle } i \in \{1, \ldots, n-1\}$$

gilt. Offenbar ist das genau dann der Fall, wenn v_h stetig ist. Indem wir die Dirichlet-Randbedingungen einbeziehen, erhalten wir den Raum

$$V_h := \{u \in L^2(\bar{\Omega}) \ : \ u|_{(x_i, x_{i+1})} \in \Pi_m \text{ für alle } i \in \{1, \ldots, n\}, \\ u \in C[a, b],\ u(a) = 0 = u(b)\}.$$

Wir haben bereits bewiesen, dass jede Funktion dieses Raums eine schwache Ableitung in $L^2(\Omega)$ besitzt, es gilt also $V_h \subseteq H^1(\Omega)$, und dass die schwache Ableitung in jedem Teilintervall mit der klassischen Ableitung übereinstimmt.

Unsere Aufgabe besteht nun darin, eine geeignete Basis $(\varphi_i)_{i \in \mathcal{I}}$ für den Raum V_h zu finden. Diese Basis sollte nicht nur den Raum V_h aufspannen, sondern sie sollte auch dazu führen, dass sich das resultierende Gleichungssystem (6.28) möglichst einfach lösen lässt. Entscheidend für dieses System ist die Matrix $\mathbf{A}$, deren Einträge im eindimensionalen Fall durch

$$a_{ij} = a(\varphi_i, \varphi_j) = \int_\Omega \varphi_i'(x) \varphi_j'(x)\, dx \qquad \text{für alle } i, j \in \mathcal{I}$$

gegeben sind. Wir stellen fest, dass a_{ij} nur dann einen von null verschiedenen Wert annehmen kann, wenn sich die *Träger* von φ_i und φ_j überschneiden. Falls es uns gelingt, die Basisfunktionen so zu wählen, dass ihre Träger möglichst klein sind, sich also mit den Trägern von möglichst wenigen anderen Basisfunktionen überschneiden, wird die Matrix $\mathbf{A}$ sehr viele Nulleinträge enthalten. Diese Eigenschaft ist sehr erstrebenswert:

- Nulleinträge brauchen wir nicht zu berechnen, also sparen wir Zeit.
- Nulleinträge brauchen wir auch nicht abzuspeichern, also sparen wir auch Speicherplatz.
- Eine genauere Untersuchung zeigt, dass auch Lösungsverfahren für das Gleichungssystem (6.28) davon profitieren, wenn nur wenige Matrixeinträge von null abweichen.

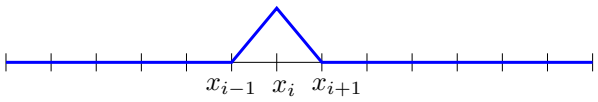

Abbildung 6.2: Basisfunktion φ_i

Wir beschränken uns bei der Konstruktion auf den Fall $m = 1$, unsere Funktionen sollen also stückweise linear auf jedem Teilintervall sein. In diesem Fall muss für beliebige Funktionen $v_h \in V_h$ die Gleichung

$$v_h(x) = \frac{x - x_i}{x_{i+1} - x_i} v_h(x_{i+1}) + \frac{x_{i+1} - x}{x_{i+1} - x_i} v_h(x_i) \qquad \text{für } i \in \{0, \dots, n\},\ x \in [x_i, x_{i+1}]$$

gelten, die Funktion ist also vollständig durch ihre Werte in den Punkten x_i bestimmt. Insbesondere ist die Funktion auf einem Intervall $[x_i, x_{i+1}]$ gleich null, wenn sie in dessen Endpunkten gleich null ist. Deshalb konstruieren wir die Basisfunktionen so, dass sie in genau einem Punkt x_i von null verschieden sind, denn dann besteht ihr Träger nur aus den beiden unmittelbar benachbarten Intervallen.

Also setzen wir $\mathcal{I} := \{1, \dots, n\}$ und verwenden die Basisfunktionen

$$\varphi_i(x) := \begin{cases} \frac{x - x_{i-1}}{x_i - x_{i-1}} & \text{für } x \in [x_{i-1}, x_i], \\ \frac{x_{i+1} - x}{x_{i+1} - x_i} & \text{für } x \in [x_i, x_{i+1}], \\ 0 & \text{ansonsten} \end{cases} \qquad \text{für alle } x \in \Omega,\ i \in \mathcal{I}.$$

Dabei ist durch die Wahl der Indexmenge $\mathcal{I}$ sicher gestellt, dass alle Basisfunktionen die Randbedingungen $\varphi_i(a) = \varphi_i(b) = 0$ erfüllen. Für diese Funktionen gelten

$$\varphi_i(x) \neq 0 \Rightarrow x \in (x_{i-1}, x_{i+1}) \qquad \text{für alle } i \in \mathcal{I},\ x \in \Omega,$$

$$\varphi_i(x_j) = \begin{cases} 1 & \text{falls } i = j, \\ 0 & \text{ansonsten} \end{cases} \qquad \text{für alle } i, j \in \mathcal{I}.$$

Insbesondere stellen wir fest, dass sich die Träger zweier Basisfunktionen φ_i und φ_j nur dann überschneiden können, wenn $|i - j| \leq 1$ gilt. Also ist $\mathbf{A}$ eine Tridiagonalmatrix und lässt sich deshalb besonders einfach handhaben.

Bei der Konstruktion der Einträge der Matrix bietet es sich an, auf die einzelnen Intervalle zurück zu greifen: Auf jedem Intervall $[x_i, x_{i+1}]$ sind höchstens die beiden Basisfunktionen φ_i und φ_{i+1} von null verschieden, so dass wir nur die Integrale

$$a_{\nu\mu}^{(i)} := \int_{x_i}^{x_{i+1}} \varphi'_{i+\nu}(x) \varphi'_{i+\mu}(x)\, dx \qquad \text{für alle } \nu, \mu \in \{0, 1\}$$

zu berechnen brauchen, um anschließend

$$a_{ij} = \begin{cases} a_{00}^{(i)} + a_{11}^{(i-1)} & \text{falls } j = i, \\ a_{01}^{(i)} & \text{falls } j = i+1, \\ a_{10}^{(i)} & \text{falls } j = i-1, \\ 0 & \text{ansonsten} \end{cases} \qquad \text{für alle } i, j \in \mathcal{I}$$

zu erhalten. Wir können also die Gesamtmatrix $\mathbf{A}$ aus den *Elementmatrizen* $\mathbf{A}^{(i)} \in \mathbb{R}^{2\times 2}$ zusammensetzen, indem wir die Beiträge der einzelnen Teilintervalle aufsummieren.

Um die Berechnung der Elementmatrizen weiter zu vereinfachen, führen wir sie auf Berechnungen auf dem Einheitsintervall zurück: Mit Hilfe der Transformation

$$\Phi_i : [0,1] \to [x_i, x_{i+1}], \qquad \hat{x} \mapsto x_i + (x_{i+1} - x_i)\hat{x},$$

erhalten wir

$$\begin{aligned} a_{\nu\mu}^{(i)} &= \int_{x_i}^{x_{i+1}} \varphi'_{i+\nu}(x)\varphi'_{j+\mu}(x)\,dx \\ &= \int_0^1 |\Phi_i'(\hat{x})|\ \varphi'_{i+\nu}(\Phi_i(\hat{x}))\varphi'_{j+\mu}(\Phi_i(\hat{x}))\,d\hat{x} \qquad \text{für alle } \nu, \mu \in \{0,1\}. \end{aligned}$$

Wir führen *lokale* Basisfunktionen ein, die durch

$$\begin{aligned} \hat{\varphi}_0 &: [0,1] \to \mathbb{R}, & \hat{x} &\mapsto 1 - \hat{x}, \\ \hat{\varphi}_1 &: [0,1] \to \mathbb{R}, & \hat{x} &\mapsto \hat{x}, \end{aligned}$$

gegeben sind und

$$\hat{\varphi}_\nu(\hat{x}) = \varphi_{i+\nu}(\Phi_i(\hat{x})) \qquad \text{für alle } \hat{x} \in [0,1],\ \nu \in \{0,1\}$$

erfüllen. Es gilt insbesondere

$$\hat{\varphi}'_\nu(\hat{x}) = \varphi'_{i+\nu}(\Phi_i(\hat{x}))\Phi_i'(\hat{x}) \qquad \text{für alle } \hat{x} \in [0,1],\ \nu \in \{0,1\},$$

so dass wir aus

$$\Phi_i' = x_{i+1} - x_i = h$$

unmittelbar

$$\frac{\hat{\varphi}'_\nu(\hat{x})}{h} = \varphi'_{i+\nu}(\Phi_i(\hat{x})) \qquad \text{für alle } \hat{x} \in [0,1],\ \nu \in \{0,1\}$$

erhalten und die Einträge der Elementmatrix in der Form

$$\begin{aligned} a_{\nu\mu}^{(i)} &= \int_0^1 |\Phi_i'(\hat{x})|\varphi'_{i+\nu}(\Phi_i(\hat{x}))\varphi'_{i+\mu}(\Phi_i(\hat{x}))\,d\hat{x} \\ &= \frac{1}{h}\int_0^1 \hat{\varphi}'_\nu(\hat{x})\hat{\varphi}'_\mu(\hat{x})\,d\hat{x} \qquad \text{für alle } \nu, \mu \in \{0,1\} \end{aligned}$$

darstellen können. Damit ist es uns gelungen, die Elementmatrix vollständig auf Größen auf dem *Referenzintervall* $[0,1]$ zurück zu führen. Da $\hat{\varphi}_0' = -1$ und $\hat{\varphi}_1' = 1$ gelten, folgt

$$\mathbf{A}^{(i)} = \frac{1}{h}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix},$$

und wir können die Gesamtmatrix aus den Elementmatrizen zusammensetzen, um

$$\mathbf{A} = \frac{1}{h}\begin{pmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix}$$

zu schließen. Selbstverständlich hätten wir die Einträge dieser Matrix auch direkt berechnen können, allerdings bietet der beschriebene Weg über Elementmatrizen und das Referenzintervall den großen Vorteil, dass er sich auf sehr viel allgemeinere Situationen übertragen lässt.

6.7 Mehrdimensionale finite Elemente

Sei nun $d \in \{2,3\}$. Wir wenden uns dem Fall zu, dass $\Omega \subseteq \mathbb{R}^d$ ein d-dimensionales Gebiet ist. Um die Darstellung übersichtlich zu halten, beschränken wir uns auf ein Polygon- beziehungsweise Polyeder-Gebiet, das sich als Vereinigung disjunkter Dreiecke oder Tetraeder darstellen lässt.

Definition 6.25 (Simplex) *Sei $\sigma \in \{0,\dots,d\}$. Eine Menge $\omega \subseteq \mathbb{R}^d$ bezeichnen wir als* σ-dimensionalen Simplex*, falls es Punkte $t_0,\dots,t_\sigma \in \mathbb{R}^d$ so gibt, dass ω die konvexe Hülle dieser Punkte ist, falls also*

$$\omega = \left\{ \sum_{i=0}^{\sigma} \alpha_i t_i \ : \ \alpha_0,\dots,\alpha_\sigma \in [0,1],\ \alpha_0 + \ldots + \alpha_\sigma \leq 1 \right\} \tag{6.35}$$

gilt. Offenbar spielt die Reihenfolge der Punkte $t_0,\dots,t_\sigma \in \mathbb{R}^d$ keine Rolle, so dass ω durch die Menge der Eckpunkte *$t = \{t_0,\dots,t_\sigma\}$ vollständig definiert ist.*

Die Menge aller Eckpunktmengen, die σ-dimensionale Simplizes beschreiben, bezeichnen wir mit

$$\mathcal{T}_\sigma := \{t \ : \ \#t = \sigma + 1,\ t \subseteq \mathbb{R}^d\}.$$

Für jede Menge $t = \{t_0,\dots,t_\sigma\} \in \mathcal{T}_\sigma$ schreiben wir den korrespondierenden Simplex als

$$\omega_t := \left\{ \sum_{p\in t} \alpha_p p \ : \ \alpha_p \in [0,1] \text{ für alle } p \in t,\ \sum_{p\in t} \alpha_p \leq 1 \right\}.$$

Um partiell integrieren zu können, benötigen wir eine Beschreibung des Randes eines Simplex. Wir stellen fest, dass für jedes $p \in t$ die Menge $t \setminus \{p\}$ die Eckpunkte der

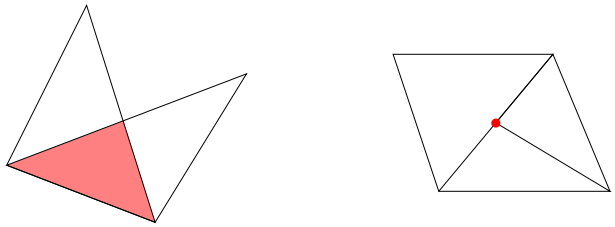

Abbildung 6.3: Beispiele für Simplizes, die keine Triangulationen bilden

Seitenfläche des Simplex ω_t beschreibt, die dem Punkt p gegenüber liegt. Diese Menge ist ein $(\sigma-1)$-dimensionaler Simplex, und wir können den vollständigen Rand des Simplex ω_t durch

$$\partial\omega_t = \bigcup_{p\in t} \omega_{t\setminus\{p\}}$$

beschreiben, und sein Inneres durch

$$\dot{\omega}_t = \omega_t \setminus \partial\omega_t.$$

Definition 6.26 (Triangulation) *Sei $\Omega \subseteq \mathbb{R}^d$ ein Gebiet, und sei $T \subseteq \mathcal{T}_d$ eine endliche Menge. Wir fordern, dass*

$$\bar{\Omega} = \bigcup_{t\in T} \omega_t, \tag{6.36}$$

gilt, dass also $\bar{\Omega}$ als Vereinigung der Simplizes dargestellt werden kann.

Wir fordern auch, dass

$$\dot{\omega}_t \cap \dot{\omega}_s = \emptyset \qquad \text{für alle } t,s\in T,\ t\neq s \tag{6.37}$$

gilt, dass also die Simplizes in T, abgesehen von ihren Rändern, disjunkt sind.

Wir fordern außerdem

$$\omega_t \cap \omega_s \neq \emptyset \quad\Longrightarrow\quad \omega_t \cap \omega_s = \omega_{t\cap s} \qquad \text{für alle } t,s\in T. \tag{6.38}$$

Zwei Simplizes aus T sollen also entweder identisch sein, über eine gemeinsame Seitenfläche, Kante oder einen gemeinsamen Punkt verfügen, oder disjunkt sein.

Eine Familie T, die diese Voraussetzungen erfüllt, bezeichnen wir als Triangulation *des Gebiets Ω.*

Die Triangulation des Gebiets Ω kann nun die Rolle der Zerlegung in Teilintervalle übernehmen, mit der wir im eindimensionalen Fall gearbeitet haben. Bevor wir stückweise polynomiale Funktionen definieren können, sollten wir zunächst Polynome auf $\mathbb{R}^d$ definieren. Dazu greifen wir wieder auf Multiindizes zurück: Wir setzen

$$x^\nu := x_1^{\nu_1}\dots x_d^{\nu_d} \qquad \text{für alle } x\in\mathbb{R}^d,\ \nu\in M_d$$

und definieren für $m \in \mathbb{N}_0$ den Raum der d-dimensionalen Polynome durch

$$\Pi_m^d := \operatorname{span}\{x \mapsto x^\nu \ : \ \nu \in M_d, \ |\nu| \leq m\}.$$

So gehört beispielsweise ein konstantes Polynom zu Π_0^d, ein lineares Polynom wie x_1 oder $2x_1 + x_3$ zu Π_1^d und ein kubisches Polynom wie $x_1 x_2^2 + x_3^3$ zu Π_3^d. Wir definieren den Raum der stückweise linearen Funktionen durch

$$V_h := \{u \in L^2(\bar{\Omega}) \ : \ u|_{\omega_t} \in \Pi_1^d, \ u \in C(\bar{\Omega}), \ u|_{\partial\Omega} = 0\}.$$

Wie schon im eindimensionalen Fall müssen wir sicher stellen, dass die Elemente des Raums V_h schwach differenzierbar sind.

Lemma 6.27 (Schwach differenzierbar) *Jede Funktion $v_h \in V_h$ erfüllt $v_h \in H^1(\Omega)$, und ihre schwachen Ableitungen sind gegeben durch*

$$\partial^\nu v_h|_{\omega_t} = \partial^\nu (v_h|_{\omega_t}) \qquad \text{für alle } t \in T, \ \nu \in M_d, \ |\nu| = 1.$$

Beweis. Für jedes Tupel $t \in T$ bezeichnen wir mit

$$n_t : \partial\omega_t \to \mathbb{R}^d$$

die Abbildung, die jedem Randpunkt den äußeren Normaleneinheitsvektor zuordnet.

Sei $v_h \in V_h$, und $\nu \in M_d$ mit $|\nu| = 1$ gegeben. Dann muss ein $i \in \{1, \ldots, d\}$ mit $\nu_i = 1$ existieren und somit

$$\partial^\nu = \frac{\partial}{\partial x_i}$$

gelten. Wir definieren $w_h \in L^2(\Omega)$ durch

$$w_h|_{\omega_t} = \frac{\partial}{\partial x_i}(v_h|_{\omega_t}) \qquad \text{für alle } t \in T.$$

Für eine Testfunktion $\varphi \in C_0^\infty(\Omega)$ erhalten wir per partieller Integration

$$\begin{aligned} -\int_\Omega \partial^\nu \varphi(x) v_h(x)\,dx &= \sum_{t\in T} -\int_{\omega_t} \frac{\partial}{\partial x_i}\varphi(x) v_h(x)\,dx \\ &= \sum_{t\in T} \int_{\omega_t} \varphi(x) \frac{\partial}{\partial x_i} v_h(x)\,dx - \sum_{t\in T} \int_{\partial\omega_t} \varphi(x) v_h(x) n_{t,i}(x)\,dx \\ &= \int_\Omega \varphi(x) w_h(x)\,dx - \sum_{t\in T}\sum_{p\in t} \int_{\omega_t\setminus\{p\}} \varphi(x) v_h(x) n_{t,i}(x)\,dx. \end{aligned}$$

Unsere Aufgabe besteht darin, nachzuweisen, dass der zweite Summand verschwindet. Dazu untersuchen wir die Menge der Seitenflächen (beziehungsweise im zweidimensionalen Fall Kanten)

$$E := \bigcup_{t\in T} \bigcup_{p\in t} (t \setminus \{p\}),$$

über die summiert wird. Wir zerlegen die Menge der Flächen in Randflächen

$$E_{\text{ext}} := \{e \in E \ : \ \omega_e \subseteq \partial\Omega\}$$

und innere Flächen

$$E_{\text{int}} := E \setminus E_{\text{ext}}.$$

Für eine Randfläche $e \in E_{\text{ext}}$ muss wegen $\omega_e \subseteq \partial\Omega$ die Funktion φ auf ω_e verschwinden, also verschwindet auch der entsprechende Summand.

Falls $e \in E_{\text{int}}$ eine innere Fläche ist, muss es Seitenfläche mindestens zweier Simplizes sein, und nach (6.37) können es auch nur genau zwei Simplizes sein. Für jedes $e \in E_{\text{int}}$ können wir deshalb $t_e, s_e \in T$ mit $t_e \neq s_e$ und $e \subseteq t_e, s_e$ fixieren. Der äußere Normalenvektor des Simplex ω_{t_e} auf der Fläche ω_e ist gerade der innere Normalenvektor des Simplex ω_{s_e} auf dieser Fläche, es gilt also

$$n_{t_e}(x) = -n_{s_e}(x) \qquad \text{für alle } x \in \omega_e.$$

Daraus folgt

$$\begin{aligned}
&\sum_{t\in T}\sum_{p\in t}\int_{\omega_{t\setminus\{p\}}} \varphi(x)v_h(x)n_{t,i}(x)\,dx \\
&\quad= \sum_{e\in E}\left(\int_{\omega_e} \varphi(x)v_h(x)n_{t_e,i}(x)\,dx + \int_{\omega_t} \varphi(x)v_h(x)n_{s_e,i}(x)\,dx\right) \\
&\quad= \sum_{e\in E_{\text{int}}}\int_{\omega_e} \varphi(x)v_h(x)(n_{t_e}(x)+n_{s_e}(x))\,dx = 0,
\end{aligned}$$

und unser Beweis ist vollständig. ∎

Also können wir das diskrete Variationsproblem (6.26) auf diesem Raum V_h formulieren und nach einer Lösung suchen. Um das Gleichungssystem (6.28) konstruieren zu können, brauchen wir geeignete Basisfunktionen. Wie schon im eindimensionalen Fall sind wir auch hier daran interessiert, diese Funktionen so zu wählen, dass ihre Träger möglichst klein sind, und wie im eindimensionalen Fall stellen wir fest, dass eine stückweise lineare Funktion $v_h \in V_h$ durch ihre Werte in den Eckpunkten jedes Simplex bereits eindeutig bestimmt ist. Wir bezeichnen die Menge der Eckpunkte mit

$$N := \bigcup_{t\in T} t$$

und zerlegen sie wieder in Randpunkte und innere Punkte

$$N_{\text{ext}} := \{p \in N \ : \ p \in \partial\Omega\}, \qquad N_{\text{int}} := N \setminus N_{\text{ext}} = \{p \in N \ : \ p \in \Omega\}.$$

In Randpunkten sind die Funktionen des Raums V_h nach Definition immer gleich null, also bietet es sich an, je eine Basisfunktion für jeden inneren Punkt zu definieren. Wir setzen

$$\mathcal{I} := N_{\text{int}}$$

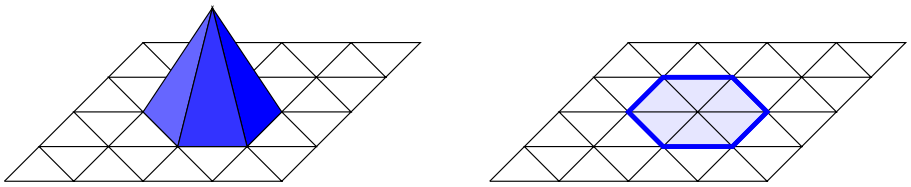

Abbildung 6.4: Stückweise lineare Basisfunktion im zweidimensionalen Fall

und definieren die Basisfunktionen durch

$$\varphi_i(p) = \begin{cases} 1 & \text{falls } i = p, \\ 0 & \text{ansonsten} \end{cases} \qquad \text{für alle } i \in \mathcal{I},\ p \in N.$$

Wieder stellt die Wahl der Indexmenge sicher, dass die homogenen Dirichlet-Randbedingungen erfüllt sind, und die Basisfunktion φ_i kann nur auf denjenigen Simplizes $t \in T$ von null abweichen, für die $i \in t$ gilt:

$$\varphi_i(x) \neq 0 \Rightarrow \text{ es existiert ein } t \in T \text{ mit } x \in \omega_t \qquad \text{für alle } i \in \mathcal{I},\ x \in \Omega$$

Nach Konstruktion muss auch

$$\varphi_i(j) = \begin{cases} 1 & \text{falls } i = j, \\ 0 & \text{ansonsten} \end{cases} \qquad \text{für alle } i, j \in \mathcal{I}$$

gelten, so dass die von uns definierten Funktionen φ_i linear unabhängig sein müssen.

Bei der Berechnung der Matrixeinträge haben wir im eindimensionalen Fall auf die einzelnen Intervalle zurückgegriffen, in die wir das Intervall Ω zerlegt hatten. Im allgemeinen Fall verwenden wir stattdessen die Simplizes: Auf einem Simplex $t \in T$ sind nur die Basisfunktionen φ_i mit $i \in t$ von null verschieden, so dass wir die Integrale

$$a_{ij}^{(t)} := \int_{\omega_t} \langle \nabla\varphi_i(x), \nabla\varphi_j(x) \rangle_2 \, dx \qquad \text{für alle } i, j \in t$$

berechnen und die Gesamtmatrix in der Form

$$a_{ij} = \sum_{\substack{t \in T \\ i,j \in t}} a_{ij}^{(t)} \qquad \text{für alle } i, j \in \mathcal{I}$$

darstellen zu können. Um den Eintrag a_{ij} zu bestimmen, genügt es also, die Beiträge aller Simplizes, die i und j enthalten, aufzusummieren.

Bemerkung 6.28 (Assemblierung) *In der Praxis geht man in der Regel so vor, dass man für jeden Simplex $t \in T$ der Triangulation die Elementmatrix $\mathbf{A}^{(t)}$ berechnet und ihre Einträge dann den korrespondierenden Einträgen der Gesamtmatrix $\mathbf{A}$ hinzufügt:*

for $t \in T$ **do**
  Berechne die Elementmatrix $\mathbf{A}^{(t)}$
  for $i, j \in t$ **do**
    if $i \in \mathcal{I}$ *und* $j \in \mathcal{I}$ **then** $a_{ij} \leftarrow a_{ij} + a_{ij}^{(t)}$

Dieser Zugang bietet den Vorteil, dass besonders einfache Datenstrukturen verwendet werden können und die für die Berechnung der Elementmatrizen erforderlichen Hilfsgrößen nur für jedes Element einmal berechnet werden müssen.

Wie im eindimensionalen Fall können wir die Berechnung der Elementmatrizen auf ein Einheitssimplex

$$\hat{\omega} := \{x \in \mathbb{R}^d \ : \ x_i \geq 0 \text{ für alle } i \in \{1, \ldots, d\}, \ x_1 + \ldots + x_d \leq 1\}$$

zurückführen, der von den Punkten

$$\hat{t}_0 := \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \qquad \hat{t}_1 := \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \qquad \ldots \qquad \hat{t}_d := \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}$$

aufgespannt wird. Für ein $t \in T$ numerieren wir dazu die Eckpunkte durch, indem wir $\{t_0, \ldots, t_d\} = t$ setzen, und definieren die Transformation

$$\Phi_t : \hat{\omega} \to \omega_t, \qquad \hat{x} \mapsto t_0 + \sum_{i=1}^{d} \hat{x}_i (t_i - t_0),$$

von $\hat{\omega}$ auf ω_t. verwenden. Nach Transformationsformel gilt

$$\begin{aligned} a_{ij}^{(t)} &= \int_{\omega_t} \langle \nabla \varphi_i(x), \nabla \varphi_j(x) \rangle_2 \, dx \\ &= \int_{\hat{\omega}} |\det D\Phi_t(\hat{x})| \langle \nabla \varphi_i(\Phi_t(\hat{x})), \nabla \varphi_j(\Phi_t(\hat{x})) \rangle_2 \, d\hat{x} \qquad \text{für alle } i, j \in t. \end{aligned} \tag{6.39}$$

Da Φ_t eine affine Transformation ist, ist $D\Phi_t$ und damit auch $\det D\Phi_t$ konstant, so dass wir die Berechnung dieser Größe aus dem Integral herausziehen können. Die Basisfunktionen führen wir wieder auf Basisfunktionen auf dem Referenzelement zurück, die durch

$$\begin{aligned} &\hat{\varphi}_0 : \hat{\omega} \to \mathbb{R}, && \hat{x} \mapsto 1 - \hat{x}_1 - \ldots - \hat{x}_d, \\ &\hat{\varphi}_\nu : \hat{\omega} \to \mathbb{R}, && \hat{x} \mapsto \hat{x}_\nu, && \text{für alle } \nu \in \{1, \ldots, d\} \end{aligned}$$

definiert sind und

$$\hat{\varphi}_\nu(\hat{t}_\mu) = \begin{cases} 1 & \text{falls } \nu = \mu, \\ 0 & \text{ansonsten} \end{cases} \qquad \text{für alle } \nu, \mu \in \{0, \ldots, d\}$$

erfüllen. Daraus folgt

$$\begin{aligned} \hat{\varphi}_\nu(\hat{t}_\mu) &= \varphi_{t_\nu}(t_\mu) = \varphi_{t,\nu} \circ \Phi_t(\hat{t}_\mu), \\ \hat{\varphi}_\nu \circ \Phi_t^{-1}(t_\mu) &= \varphi_{t_\nu}(t_\mu) \qquad \text{für alle } \nu, \mu \in \{0, \dots, d\}, \end{aligned}$$

und da die Funktionen auf ω_t affin sind, können wir aus der Identität in den Eckpunkten bereits

$$\varphi_{t_\nu} = \hat{\varphi}_\nu \circ \Phi_t^{-1} \qquad \text{für alle } \nu \in \{0, \dots, d\}$$

schließen. Mit Hilfe der Kettenregel erhalten wir

$$\begin{aligned} \nabla \varphi_{t_\nu}(x) = D\varphi_{t_\nu}(x)^* &= (D\hat{\varphi}_\nu \circ \Phi_t^{-1} D(\Phi_t^{-1})(x))^* = D(\Phi_t^{-1})(x)^* (D\hat{\varphi}_\nu \circ \Phi_t^{-1})^* \\ &= D(\Phi_t^{-1})(x)^* \nabla \hat{\varphi}_\nu \circ \Phi_t^{-1} \qquad \text{für alle } \nu \in \{0, \dots, d\},\ x \in \omega_t, \end{aligned}$$

und da uns dank Umkehrsatz

$$D(\Phi_t^{-1})(x) = (D\Phi_t)^{-1}(\Phi_t^{-1}(x)) \qquad \text{für alle } x \in \omega_t$$

zur Verfügung steht, folgt

$$\nabla \varphi_{t_\nu}(x) = D\Phi_t^{-*}(\Phi_t^{-1}(x))\ \nabla \hat{\varphi}_\nu(\Phi_t^{-1}(x)) \qquad \text{für alle } \nu \in \{0, \dots, d\},\ x \in \omega_t.$$

Hier verwenden wir die Abkürzung $(\mathbf{A}^{-1})^* = \mathbf{A}^{-*}$. Indem wir diese Gleichung in (6.39) einsetzen, ergibt sich die Darstellung

$$a_{t_\nu t_\mu}^{(t)} = \int_{\hat{\omega}} |\det D\Phi_t(\hat{x})|\ \langle D\Phi_t(\hat{x})^{-*} \nabla \hat{\varphi}_\nu(\hat{x}), D\Phi_t(\hat{x})^{-*} \nabla \hat{\varphi}_\mu(\hat{x}) \rangle_2 \, d\hat{x}$$
$$\text{für alle } \nu, \mu \in \{0, \dots, d\}.$$

In unserem Fall ist Φ_t eine affine Abbildung, also ist $D\Phi_t$ konstant, und damit sind es auch $\det D\Phi_t$ und $D\Phi_t^{-*}$. Für einen stückweise polynomialen Ansatzraum muss damit der Integrand ein Polynom sein, so dass sich die Einträge der Elementmatrix relativ einfach mit einer Quadraturformel berechnen lassen.

6.8 Analyse des Approximationsfehlers

Für die Abschätzung des Approximationsfehlers verwendet man in der Regel das Céa-Lemma 6.23, also die Ungleichung

$$\|u - u_h\|_V \leq \frac{C_S}{C_E} \|u - v_h\|_V \qquad \text{für alle } v_h \in V_h,$$

die die exakte Lösung u in Bezug zu der Näherungslösung u_h setzt.

In der Regel genügt es, die Existenz einer Funktion $v_h \in V_h$ nachzuweisen, die u hinreichend gut approximiert, um, abgesehen von der Konstanten, dieselbe Eigenschaft für die mit Hilfe des Gleichungssystems (6.28) und der Formel (6.27) praktisch berechenbare Funktion u_h zu erhalten.

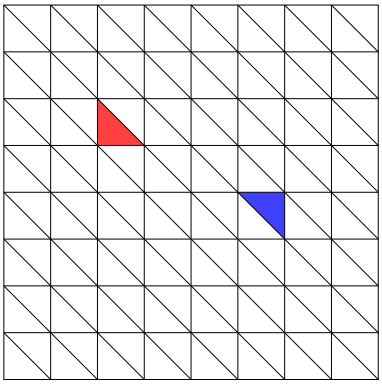

Abbildung 6.5: Regelmässige Triangulation. Je ein Dreieck aus T_1 und T_2 ist rot beziehungsweise blau markiert.

Reduktion auf das Referenzelement

Da wir an dieser Stelle nicht auf die Details der Approximationstheorie in Sobolew-Räumen eingehen können, beschränken wir uns darauf, einen besonders einfachen Fall zu behandeln: Wir untersuchen wieder das Einheitsquadrat $\Omega = [0,1]^2$ und legen ein $N \in \mathbb{N}$ fest, das die Auflösung der zu konstruierenden Triangulation bestimmt. Wir definieren die Menge der Punkte durch

$$h := \frac{1}{N+1}, \qquad x_{ij} := \begin{pmatrix} ih \\ jh \end{pmatrix} \qquad \text{für alle } i,j \in \{0,\dots,N+1\},$$
$$\mathcal{N} := \{x_{ij} \ : \ i,j \in \{0,\dots,N+1\}\}, \qquad \mathcal{N}_{\text{int}} := \{x_{ij} \ : \ i,j \in \{1,\dots,N\}\}.$$

Für die Triangulation des Gebiets Ω verwenden wir zwei Typen von Dreiecken:

$$T_1 := \{\{x_{ij}, x_{i+1,j}, x_{i,j+1}\} \ : \ i,j \in \{0,\dots,N\}\},$$
$$T_2 := \{\{x_{i+1,j+1}, x_{i,j+1}, x_{i+1,j}\} \ : \ i,j \in \{0,\dots,N\}\},$$
$$T := T_1 \cup T_2.$$

Die resultierende Triangulation ist in Abbildung 6.5 dargestellt. Sie hat den großen Vorteil, dass alle auftretenden Dreiecke zu dem Referenzdreieck

$$\hat{\omega} = \{x \in \mathbb{R}^2 \ : \ x_1, x_2 \geq 0, \ x_1 + x_2 \leq 1\}$$

kongruent sind, außerdem lassen sich die Abbildungen von $\hat{\omega}$ auf ein Element ω_t besonders einfach darstellen: Falls $t \in T_1$ gilt, existieren $i,j \in \{0,\dots,N\}$ mit

$$t = \{x_{ij}, x_{i+1,j}, x_{i,j+1}\},$$

und wir können die Abbildung

$$\Phi_t : \hat{\omega} \to \omega_t, \qquad \hat{x} \mapsto x_{ij} + h\hat{x},$$

verwenden. Für $t \in T_2$ finden wir nach Definition $i, j \in \{0, \dots, N\}$ mit

$$t = \{x_{i+1,j+1}, x_{i,j+1}, x_{i+1,j}\},$$

und wir können mit

$$\Phi_t : \hat{\omega} \to \omega_t, \qquad \hat{x} \mapsto x_{i+1,j+1} - h\hat{x},$$

arbeiten. In beiden Fällen ist die Berechnung der Ableitungen offenbar sehr einfach, es gilt

$$\begin{aligned} D\Phi_t &= h\mathbf{I}, & D\Phi_t^{-*} &= h^{-1}\mathbf{I}, & \det D\Phi_t &= h^2 & &\text{für alle } t \in T_1,\\ D\Phi_t &= -h\mathbf{I}, & D\Phi_t^{-*} &= -h^{-1}\mathbf{I}, & \det D\Phi_t &= h^2 & &\text{für alle } t \in T_2. \end{aligned}$$

Als nächstes müssen wir untersuchen, wie sich die Ableitungen von Funktionen unter diesen Transformationen verhalten. Seien also $w \in L^2(\Omega)$ und $\nu \in M_d$ so gegeben, dass $\partial^\nu w \in L^2(\Omega)$ existiert. Wir definieren

$$\hat{w}_t := w \circ \Phi_t \qquad \text{für alle } t \in T$$

und stellen mit Hilfe der partiellen Integration und der Transformationsformel fest, dass die schwachen Ableitungen $\partial^\nu \hat{w}_t \in L^2(\hat{\omega})$ für alle $t \in T$ existieren und die Gleichungen

$$\partial^\nu \hat{w}_t = \begin{cases} h^{|\nu|}(\partial^\nu w) \circ \Phi_t & \text{falls } t \in T_1,\\ (-h)^{|\nu|}(\partial^\nu w) \circ \Phi_t & \text{ansonsten.} \end{cases} \qquad \text{für alle } t \in T$$

erfüllen. Durch Einsetzen in die Definition folgt

$$\|w\|_{L^2}^2 = \sum_{t\in T} \|w|_{\omega_t}\|_{L^2}^2 = h^2 \sum_{t\in T} \|\hat{w}_t\|_{L^2}^2, \tag{6.40a}$$

$$|w|_{H^1}^2 = \sum_{t\in T} |w|_{\omega_t}|_{H^1}^2 = \sum_{t\in T} |\hat{w}_t|_{H^1}^2, \tag{6.40b}$$

$$|w|_{H^2}^2 = \sum_{t\in T} |w|_{\omega_t}|_{H^2}^2 = h^{-2} \sum_{t\in T} |\hat{w}_t|_{H^2}^2, \tag{6.40c}$$

wir können also alle Normen durch Summen von Normen auf dem Referenzelement darstellen.

Lokale Interpolation per Sobolew-Einbettungssatz

Wie bei Fehlerabschätzungen für die Taylor-Entwicklung oder die Interpolation müssen wir auch im Fall der finiten Elemente voraussetzen, dass die zu approximierende Funktion hinreichend oft differenzierbar ist. In unserem Fall lässt sich beweisen, dass eine Konstante $C_{\mathrm{rg}} \in \mathbb{R}_{>0}$ so existiert, dass für jede rechte Seite $f \in L^2(\Omega)$ die Lösung $u \in H_0^1(\Omega)$ der Aufgabe (6.17) auch $u \in H^2(\Omega)$ und die Abschätzung

$$\|u\|_{H^2} \le C_{\mathrm{rg}} \|f\|_{L^2} \tag{6.41}$$

erfüllt. Derartige *Regularitätsaussagen* gelten auf konvexen Gebieten, bei nicht-konvexen Gebieten gilt noch $u \in H^{1+\alpha}(\Omega)$, wobei $\alpha \in (0,1)$ ein von den Außenwinkeln abhängender Parameter ist.

Um die Approximation v_h konstruieren zu können, greifen wir auf ein Resultat der Theorie der Sobolew-Räume zurück: Der *Einbettungssatz von Sobolew* besagt, dass in unserem Fall $H^2(\Omega) \subseteq C(\Omega)$ gilt, dass also jede Funktion aus $H^2(\Omega)$ stetig ist (genauer gesagt sich höchstens auf einer Nullmenge von einer stetigen Funktion unterscheidet), und dass eine Konstante $C_{\mathrm{so}} \in \mathbb{R}_{>0}$ existiert, die

$$\|v\|_{\infty,\Omega} \leq C_{\mathrm{so}} \|v\|_{H^2} \qquad \text{für alle } v \in H^2(\Omega)$$

erfüllt. Der Raum $H^2(\Omega)$ ist demnach stetig in $C(\Omega)$ eingebettet. Also können wir Funktionen aus $H^2(\Omega)$ punktweise auswerten und so stückweise interpolieren, indem wir den Lagrange-Interpolationsoperator

$$\mathfrak{I}_h : H^2(\Omega) \to V_h, \qquad v \mapsto \sum_{i \in \mathcal{I}} v(i)\varphi_i,$$

definieren. Hierbei ist $v(i)$ so zu verstehen, dass der stetige Repräsentant der Äquivalenzklasse v im Punkt $i \in \mathbb{R}^2$ ausgewertet wird. Da $\mathfrak{I}_h$ in den Raum V_h abbildet, muss infolge des Céa-Lemmas

$$\|u - u_h\|_V \leq \frac{C_S}{C_E} \|u - \mathfrak{I}_h[u]\|_V$$

gelten, so dass wir „nur noch“ die rechte Seite dieser Abschätzung beschränken müssen.

Dazu greifen wir wieder auf das Referenzelement zurück: Wir definieren

$$e := u - \mathfrak{I}_h[u], \qquad \hat{e}_t := e \circ \Phi_t \qquad \text{für alle } t \in T$$

und erhalten dank (6.40a) und (6.40b)

$$\|u - \mathfrak{I}_h[u]\|_{H^1}^2 = \|e\|_{H^1}^2 = \|e\|_{L^2}^2 + |e|_{H^1}^2 = \sum_{t \in T} h^2 \|\hat{e}_t\|_{L^2}^2 + |\hat{e}_t|_{H^1}^2.$$

Für alle $t \in T$ ist $(\mathfrak{I}_h[u])|_{\omega_t}$ gerade das lineare Lagrange-Interpolationspolynom der Funktion u, also muss $\mathfrak{I}_h[u] \circ \Phi_t$ das Interpolationspolynom der Funktion $\hat{u}_t = u \circ \Phi_t$ auf dem Referenzelement $\hat{\omega}$ sein. Wenn wir den Interpolationsoperator auf $\hat{\omega}$ mit

$$\widehat{\mathfrak{I}} : H^2(\hat{\omega}) \to \Pi_1^2, \qquad v \mapsto v(\hat{t}_0)\hat{\varphi}_0 + v(\hat{t}_1)\hat{\varphi}_1 + v(\hat{t}_2)\hat{\varphi}_2,$$

bezeichnen, folgt

$$\hat{e}_t = e \circ \Phi_t = (u - \mathfrak{I}_h[u]) \circ \Phi_t = \hat{u}_t - \widehat{\mathfrak{I}}[\hat{u}_t],$$

wir müssen also lediglich die lineare Lagrange-Interpolation auf dem Referenzelement $\hat{\omega}$ analysieren, allerdings in der „Sprache“ der Sobolew-Räume.

Lokale Approximation per Bramble-Hilbert-Lemma

Dazu verwenden wir ein weiteres Hilfsmittel aus der Theorie der Sobolew-Räume: Das *Bramble-Hilbert-Lemma* ist eine Verallgemeinerung der Taylor-Entwicklung für schwach differenzierbare Funktionen und besagt in unserem Fall, dass eine Konstante $C_{\mathrm{bh}} \in \mathbb{R}_{>0}$ so existiert, dass für alle $v \in H^2(\hat{\omega})$ ein Polynom $v_0 \in \Pi_1^2$ existiert, das die Abschätzungen

$$\|v - v_0\|_{H^2} \leq C_{\mathrm{bh}} |v|_{H^2}$$

erfüllt. Dieses Resultat wenden wir nun für $t \in T$ auf $\hat{w}_t$ an: Wir finden $u_0 \in \Pi_1^2$ mit

$$\|\hat{u}_t - \hat{u}_0\|_{H^2} \leq C_{\mathrm{bh}} |\hat{u}_t|_{H^2} \tag{6.42}$$

und stellen fest, dass

$$\widehat{\mathfrak{I}}[\hat{u}_0] = \hat{u}_0$$

gilt, da die Lagrange-Interpolation eine Projektion auf den Raum der jeweiligen Polynome ist. Also folgt auch

$$\hat{u}_t - \widehat{\mathfrak{I}}[\hat{u}_t] = \hat{u}_t - \hat{u}_0 - \widehat{\mathfrak{I}}[\hat{u}_t - \hat{u}_0] = (I - \widehat{\mathfrak{I}})[\hat{u}_t - \hat{u}_0].$$

Mit dem Einbettungssatz von Sobolew finden wir eine Konstante $\widehat{C}_{\mathrm{so}} \in \mathbb{R}_{>0}$, die von t unabhängig ist und

$$\|(I - \widehat{\mathfrak{I}})[\hat{u}_t - \hat{u}_0]\|_{L^2} \leq \|I - \widehat{\mathfrak{I}}\|_{L^2 \leftarrow C(\hat{\omega})} \|\hat{u}_t - \hat{u}_0\|_{\infty,\hat{\omega}} \leq \|I - \widehat{\mathfrak{I}}\|_{L^2 \leftarrow C(\hat{\omega})} \widehat{C}_{\mathrm{so}} \|\hat{u}_t - \hat{u}_0\|_{H^2}$$

erfüllt. Mit Hilfe des Bramble-Hilbert-Lemmas folgt

$$\begin{aligned} \|\hat{u}_t - \widehat{\mathfrak{I}}[\hat{u}_t]\|_{L^2} = \|(I - \widehat{\mathfrak{I}})[\hat{u}_t - \hat{u}_0]\|_{L^2} &\leq \widehat{C}_{\mathrm{so}} \|I - \widehat{\mathfrak{I}}\|_{L^2 \leftarrow C(\hat{\omega})} \|\hat{u}_t - \hat{u}_0\|_{H^2} \\ &\leq \widehat{C}_{\mathrm{so}} C_{\mathrm{bh}} \|I - \widehat{\mathfrak{I}}\|_{L^2 \leftarrow C(\hat{\omega})} |\hat{u}_t - \hat{u}_0|_{H^2}, \end{aligned}$$

und da die zweiten Ableitungen des linearen Polynoms $\hat{u}_0$ gleich null sind, erhalten wir schließlich

$$\|\hat{u}_t - \widehat{\mathfrak{I}}[\hat{u}_t]\|_{L^2} \leq C_0 |\hat{u}_t|_{H^2}, \qquad C_0 := \widehat{C}_{\mathrm{so}} C_{\mathrm{bh}} \|I - \widehat{\mathfrak{I}}\|_{L^2 \leftarrow C(\hat{\omega})} \qquad \text{für alle } t \in T.$$

Entsprechend können wir auch

$$|\hat{u}_t - \widehat{\mathfrak{I}}[\hat{u}_t]|_{H^1} \leq C_1 |\hat{u}_t|_{H^2}, \qquad C_1 := \widehat{C}_{\mathrm{so}} C_{\mathrm{bh}} \|I - \widehat{\mathfrak{I}}\|_{H^1 \leftarrow C(\hat{\omega})} \qquad \text{für alle } t \in T$$

zeigen und erhalten mit (6.40a) und (6.40c) die Abschätzung

$$\|e\|_{L^2}^2 = h^2 \sum_{t \in T} \|\hat{e}_t\|_{L^2}^2 \leq C_0^2 h^2 \sum_{t \in T} |\hat{u}_t|_{H^2}^2 = C_0^2 h^4 h^{-2} \sum_{t \in T} |\hat{u}_t|_{H^2}^2 = C_0^2 h^4 |u|_{H^2}^2, \tag{6.43}$$

während sich mit (6.40b) und (6.40c) die Abschätzung

$$|e|_{H^1}^2 = \sum_{t \in T} |\hat{e}_t|_{H^1}^2 \leq C_1^2 \sum_{t \in T} |\hat{u}_t|_{H^2}^2 = C_1^2 h^2 h^{-2} \sum_{t \in T} |\hat{u}_t|_{H^2}^2 = C_1^2 h^2 |u|_{H^2}^2, \tag{6.44}$$

ergibt. Diese Schlusstechnik, bei der sich h-Potenzen durch das unterschiedliche Verhalten der unterschiedlichen Ableitungen unter Skalierung des Gebiets ergeben, spielt eine entscheidende Rolle in vielen Bereichen der Theorie der finiten Elemente und wird häufig als *Skalierungsargument* (engl. *scaling argument*) bezeichnet.

Durch Kombination der Teilergebnisse (6.43) und (6.44) folgt

$$\begin{aligned}\|u - \mathfrak{I}_h[u]\|_{H^1} &= \sqrt{\|e\|_{L^2}^2 + |e|_{H^1}^2} \leq \sqrt{C_0^2 h^4 + C_1^2 h^2}|u|_{H^2} \\ &= h\sqrt{C_0^2 h^2 + C_1^2}|u|_{H^2}.\end{aligned}$$

In unserem Fall gilt $h \leq 1$, also haben wir schließlich

$$\|u - \mathfrak{I}_h[u]\|_{H^1} \leq h\sqrt{C_0^2 + C_1^2}|u|_{H^2} \qquad \text{für alle } u \in H^2(\Omega)$$

bewiesen. Mit dem Céa-Lemma und (6.41) folgt

$$\|u - u_h\|_{H^1} \leq \frac{C_S}{C_E} h\sqrt{C_0^2 + C_1^2}|f|_{L^2}, \tag{6.45}$$

der Diskretisierungsfehler wird also proportional zu h fallen, wenn wir das Gitter verfeinern.

Fehlerabschätzung für die L^2-Norm per Aubin-Nitsche-Lemma

Ein zu h proportional fallender Fehler ist nicht befriedigend, wenn wir bedenken, dass der Fehler des einfachen Finite-Differenzen-Verfahren gemäß Satz 5.7 proportional zu h^2 fällt. Allerdings ist zu beachten, dass die Aussage (6.45) nicht nur eine Schranke für den Fehler der Funktion darstellt, sondern auch für ihre Ableitung.

Falls uns die Konvergenz der L^2-Norm genügt, können wir mit einem eleganten Trick eine wie h^2 fallende Fehlerschranke gewinnen.

Lemma 6.29 (Aubin-Nitsche) *Sei $C_{\text{apx}} \in \mathbb{R}_{>0}$ so gegeben, dass für jede rechte Seite $f \in L^2(\Omega)$ die Lösung $u \in H_0^1(\Omega)$ des Problems (6.17) und die Lösung $u_h \in V_h$ des Problems (6.26) die Fehlerabschätzung*

$$\|u - u_h\|_{H^1} \leq C_{\text{apx}} h \|f\|_{L^2} \tag{6.46}$$

erfüllen. Dann gilt auch

$$\|u - u_h\|_{L^2} \leq C_S C_{\text{apx}}^2 h^2 \|f\|_{L^2}.$$

Beweis. Die Idee dieses Beweises besteht darin, ein geschickt gewähltes Hilfsproblem zu untersuchen: Wir definieren das Funktional

$$\mu : L^2(\Omega) \to \mathbb{R}, \qquad v \mapsto \int_\Omega v(x)(u(x) - u_h(x))\,dx,$$

und suchen nach Funktionen $w \in H_0^1(\Omega)$ und $w_h \in V_h$, die die Variationsgleichungen

$$\begin{aligned} a(w, v) &= \mu(v) && \text{für alle } v \in H_0^1(\Omega), \\ a(w_h, v_h) &= \mu(v_h) && \text{für alle } v_h \in V_h \end{aligned}$$

lösen. Da a symmetrisch ist, erfüllen auch die Lösungen dieser Probleme

$$\|w - w_h\|_{H^1} \leq C_{\text{apx}} h \|u - u_h\|_{L^2}.$$

Infolge der geschickten Wahl des Funktionals μ erhalten wir

$$\|u - u_h\|_{L^2}^2 = \mu(u - u_h) = a(w, u - u_h),$$

mit der Galerkin-Orthogonalität (vgl. (6.30)) folgt

$$\|u - u_h\|_{L^2}^2 = a(w - w_h, u - u_h),$$

und dank der Stetigkeit (6.18) ergibt sich

$$\|u - u_h\|_{L^2}^2 \leq C_S \|w - w_h\|_{H^1} \|u - u_h\|_{H^1}.$$

Nun können wir die Fehlerabschätzung (6.46) für u und w einsetzen, um

$$\|u - u_h\|_{L^2}^2 \leq C_S C_{\text{apx}} h \|u - u_h\|_{L^2} C_{\text{apx}} h \|f\|_{L^2} = C_S C_{\text{apx}}^2 h^2 \|u - u_h\|_{L^2} \|f\|_{L^2}$$

zu erhalten, und daraus folgt die Behauptung. ∎

7 Lösungsverfahren für schwachbesetzte Matrizen

Sowohl bei der Finite-Differenzen- als auch bei der Finite-Elemente-Methode entstehen Matrizen $\mathbf{A} \in \mathbb{R}^{\mathcal{I}\times\mathcal{I}}$, bei denen jede Zeile nur wenige von null verschiedene Einträge enthält. Das hat zur Folge, dass sich das Matrix-Vektor-Produkt $\mathbf{Ay}$ für beliebige Vektoren $\mathbf{y} \in \mathbb{R}^{\mathcal{I}}$ sehr effizient berechnen lässt, und diese Eigenschaft können wir ausnutzen, um die linearen Gleichungssysteme, die sowohl bei elliptischen als auch parabolischen Problemen auftreten, schnell zu lösen.

Unsere Aufgabe besteht darin, das System

$$\mathbf{Ax} = \mathbf{b} \tag{7.1}$$

zu lösen. Wir beschränken uns auf den Fall, dass die Matrix $\mathbf{A}$ symmetrisch und positiv definit ist, dass also

$$\mathbf{A}^* = \mathbf{A}, \qquad \langle \mathbf{Ay}, \mathbf{y}\rangle_2 > 0 \qquad \text{für alle } \mathbf{y} \in \mathbb{R}^{\mathcal{I}} \setminus \{\mathbf{0}\}$$

gilt. In allen von uns betrachteten Modellproblemen sind diese Voraussetzungen erfüllt, und dasselbe gilt auch für viele in der Praxis auftretende Probleme.

7.1 Gradientenverfahren

Wir gehen wie in Abschnitt 6.5 vor und bringen das lineare Gleichungssystem (7.1) in die Form eines Minimierungproblems, indem wir die Funktion

$$f : \mathbb{R}^{\mathcal{I}} \to \mathbb{R}, \qquad \mathbf{x} \mapsto \frac{1}{2}\langle \mathbf{x}, \mathbf{Ax}\rangle_2 - \langle \mathbf{x}, \mathbf{b}\rangle_2,$$

definieren und feststellen, dass sie ihr Minimum gerade für die Lösung des Systems (7.1) annimmt:

Lemma 7.1 (Minimierungsproblem) *Seien $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{\mathcal{I}}$. Dann gilt*

$$f(\mathbf{x}) \leq f(\mathbf{x} + \lambda\mathbf{y}) \qquad \textit{für alle } \lambda \in \mathbb{R} \tag{7.2}$$

genau dann, wenn

$$\langle \mathbf{y}, \mathbf{b} - \mathbf{Ax}\rangle_2 = 0 \tag{7.3}$$

erfüllt ist.

Insbesondere ist $\mathbf{x}$ *genau dann eine Lösung des linearen Gleichungssystems (7.1), wenn*

$$f(\mathbf{x}) \leq f(\mathbf{z}) \qquad \text{für alle } \mathbf{z} \in \mathbb{R}^{\mathcal{I}} \tag{7.4}$$

gilt. Da (7.1) genau eine Lösung besitzt, muss somit f *genau ein globales Minimum besitzen.*

Beweis. Zunächst stellen wir fest, dass

$$\begin{aligned} f(\mathbf{x}+\lambda\mathbf{y}) &= \frac{1}{2}\langle \mathbf{x}+\lambda\mathbf{y}, \mathbf{A}(\mathbf{x}+\lambda\mathbf{y})\rangle_2 - \langle \mathbf{x}+\lambda\mathbf{y}, \mathbf{b}\rangle_2 \\ &= f(\mathbf{x}) - \lambda\langle \mathbf{y}, \mathbf{b}-\mathbf{A}\mathbf{x}\rangle_2 + \frac{\lambda^2}{2}\langle \mathbf{y}, \mathbf{A}\mathbf{y}\rangle_2 \end{aligned} \tag{7.5}$$

für alle $\lambda \in \mathbb{R}$ gilt.

Nehmen wir nun an, dass (7.2) gilt. Für $\mathbf{y} = \mathbf{0}$ ist die Aussage (7.3) trivial, sei also im Folgenden $\mathbf{y} \neq \mathbf{0}$ vorausgesetzt. Indem wir

$$\lambda := \frac{\langle \mathbf{y}, \mathbf{b}-\mathbf{A}\mathbf{x}\rangle_2}{\langle \mathbf{y}, \mathbf{A}\mathbf{y}\rangle_2}$$

in (7.5) einsetzen, erhalten wir

$$f(\mathbf{x}+\lambda\mathbf{y}) = f(\mathbf{x}) - \frac{\langle \mathbf{y}, \mathbf{b}-\mathbf{A}\mathbf{x}\rangle_2^2}{\langle \mathbf{y}, \mathbf{A}\mathbf{y}\rangle_2} + \frac{\langle \mathbf{y}, \mathbf{b}-\mathbf{A}\mathbf{x}\rangle_2^2}{2\langle \mathbf{y}, \mathbf{A}\mathbf{y}\rangle_2} = f(\mathbf{x}) - \frac{\langle \mathbf{y}, \mathbf{b}-\mathbf{A}\mathbf{x}\rangle_2^2}{2\langle \mathbf{y}, \mathbf{A}\mathbf{y}\rangle_2},$$

und da (7.2) vorausgesetzt ist, folgt

$$0 \leq \langle \mathbf{y}, \mathbf{b}-\mathbf{A}\mathbf{x}\rangle_2^2 \leq 0,$$

also gerade (7.3).

Gehen wir also nun davon aus, dass (7.3) gilt. Dann folgt aus (7.5) direkt

$$f(\mathbf{x}+\lambda\mathbf{y}) = f(\mathbf{x}) + \frac{\lambda^2}{2}\langle \mathbf{y}, \mathbf{A}\mathbf{y}\rangle_2 \geq f(\mathbf{x}),$$

da die Matrix $\mathbf{A}$ positiv definit ist. Also folgt (7.2).

Sei nun $\mathbf{x}$ eine Lösung des Gleichungssystems (7.1), und sei $\mathbf{z} \in \mathbb{R}^{\mathcal{I}}$ Dann gilt insbesondere (7.3) für den Vektor $\mathbf{y} := \mathbf{z} - \mathbf{x}$, und wie wir bereits gezeigt haben folgt daraus

$$f(\mathbf{x}) \leq f(\mathbf{x}+\mathbf{y}) = f(\mathbf{z}).$$

Um die letzte Folgerung zu beweisen, setzen wir voraus, dass $\mathbf{x}$ die Minimalitätsbedingung (7.4) erfüllt. Wir wählen $\mathbf{y} := \mathbf{b} - \mathbf{A}\mathbf{x}$ und folgern, dass auch (7.2) gelten muss. Aus der bereits bewiesenen Äquivalenz zu (7.3) folgt

$$\|\mathbf{b}-\mathbf{A}\mathbf{x}\|_2^2 = \langle \mathbf{y}, \mathbf{b}-\mathbf{A}\mathbf{x}\rangle_2 = 0,$$

also ist $\mathbf{x}$ Lösung des Gleichungssystems (7.1). ∎

Statt nach einer Lösung des linearen Gleichungssystems (7.1) zu suchen, können wir also auch das Minimierungsproblem (7.4) zu lösen versuchen.

Dazu gehen wir *iterativ* vor: Ausgehend von einer Näherung $\mathbf{x}^{(0)} \in \mathbb{R}^{\mathcal{I}}$ der Lösung konstruieren wir eine verbesserte Näherung $\mathbf{x}^{(1)} \in \mathbb{R}^{\mathcal{I}}$, daraus eine weitere Näherung $\mathbf{x}^{(2)} \in \mathbb{R}^{\mathcal{I}}$, bis wir eine ausreichend hohe Genauigkeit erreicht haben. Im Gegensatz zu Verfahren wie der Gauß-Elimination wird dieser Prozess im Allgemeinen nicht die exakte Lösung nach endlich vielen Rechenschritten bestimmen, aber er wird sie beliebig zu annähern. Das reicht für unsere Zwecke allerdings völlig aus: Die Lösung des linearen Gleichungssystems beschreibt in unserer Anwendung eine Näherung der Lösung der kontinuierlichen Differentialgleichung, und den durch die Diskretisierung eingeführten Fehler können wir auch durch exaktes Lösen des Gleichungssystems nicht reduzieren.

Wir bezeichnen die Folge der Näherungslösungen mit $(\mathbf{x}^{(m)})_{m\in\mathbb{N}_0}$ und den erwünschten Grenzwert mit $\mathbf{x}^* \in \mathbb{R}^{\mathcal{I}}$, definiert durch

$$\mathbf{A}\mathbf{x}^* = \mathbf{b}.$$

Unsere Aufgabe besteht darin, zu einer Näherung $\mathbf{x}^{(m)} \in \mathbb{R}^{\mathcal{I}}$ eine verbesserte Näherung $\mathbf{x}^{(m+1)} \in \mathbb{R}^{\mathcal{I}}$ zu konstruieren.

Wir lösen diese Aufgabe in zwei Schritten: Zunächst wählen wir eine *Suchrichtung* $\mathbf{p}^{(m)} \in \mathbb{R}^{\mathcal{I}}$ aus, entlang derer wir die Lösung verbessern wollen. Dann bestimmen wir die optimale *Schrittweite* $\lambda^{(m)} \in \mathbb{R}$. Beide zusammen definieren die nächste Näherung

$$\mathbf{x}^{(m+1)} = \mathbf{x}^{(m)} + \lambda^{(m)}\mathbf{p}^{(m)}.$$

Um die Wahl der Suchrichtung analysieren zu können, bietet es sich an, dass *Residuum*

$$\mathbf{r}^{(m)} := \mathbf{b} - \mathbf{A}\mathbf{x}^{(m)} \qquad \text{für alle } m \in \mathbb{N}_0$$

zu definieren. Indem wir (7.5) auf $\mathbf{x}^{(m)}$ und $\mathbf{p}^{(m)}$ anwenden, erhalten wir

$$\begin{aligned} f(\mathbf{x}^{(m)} + \lambda\mathbf{p}^{(m)}) &= f(\mathbf{x}^{(m)}) - \lambda\langle\mathbf{p}^{(m)}, \mathbf{b} - \mathbf{A}\mathbf{x}^{(m)}\rangle_2 + \frac{\lambda^2}{2}\langle\mathbf{p}^{(m)}, \mathbf{A}\mathbf{p}^{(m)}\rangle_2 \\ &= f(\mathbf{x}^{(m)}) - \lambda\langle\mathbf{p}^{(m)}, \mathbf{r}^{(m)}\rangle_2 + \frac{\lambda^2}{2}\langle\mathbf{p}^{(m)}, \mathbf{A}\mathbf{p}^{(m)}\rangle_2. \end{aligned}$$

Falls λ hinreichend klein ist, können wir den dritten Term vernachlässigen, müssen also lediglich dafür sorgen, dass

$$\langle\mathbf{p}^{(m)}, \mathbf{r}^{(m)}\rangle_2$$

möglichst groß wird. Aufgrund der Cauchy-Schwarz-Ungleichung wird das Skalarprodukt maximal, wenn seine Argumente linear abhängig sind, also dürfte

$$\mathbf{p}^{(m)} := \mathbf{r}^{(m)}$$

eine gute Wahl für die Suchrichtung sein.

Der nächste Schritt besteht darin, die optimale Schrittweite zu bestimmen. Damit

$$\mathbf{x}^{(m+1)} = \mathbf{x}^{(m)} + \lambda^{(m)}\mathbf{p}^{(m)}$$

optimal ist, muss

$$f(\mathbf{x}^{(m)} + \lambda^{(m)}\mathbf{p}^{(m)}) \leq f(\mathbf{x}^{(m)} + \lambda\mathbf{p}^{(m)}) \qquad \text{für alle } \lambda \in \mathbb{R} \tag{7.6}$$

gelten. Aus Lemma 7.1 folgt, dass das äquivalent zu

$$0 = \langle \mathbf{p}^{(m)}, \mathbf{b} - \mathbf{A}(\mathbf{x}^{(m)} + \lambda^{(m)}\mathbf{p}^{(m)})\rangle_2 = \langle \mathbf{p}^{(m)}, \mathbf{r}^{(m)}\rangle_2 - \langle \mathbf{p}^{(m)}, \lambda^{(m)}\mathbf{A}\mathbf{p}^{(m)}\rangle_2$$

ist, und indem wir nach $\lambda^{(m)}$ auflösen folgt

$$\lambda^{(m)} = \frac{\langle \mathbf{p}^{(m)}, \mathbf{r}^{(m)}\rangle_2}{\langle \mathbf{p}^{(m)}, \mathbf{A}\mathbf{p}^{(m)}\rangle_2}, \tag{7.7}$$

und für die von uns gewählte Suchrichtung schließlich

$$\lambda^{(m)} = \frac{\langle \mathbf{r}^{(m)}, \mathbf{r}^{(m)}\rangle_2}{\langle \mathbf{r}^{(m)}, \mathbf{A}\mathbf{r}^{(m)}\rangle_2} = \frac{\|\mathbf{r}^{(m)}\|_2^2}{\langle \mathbf{r}^{(m)}, \mathbf{A}\mathbf{r}^{(m)}\rangle_2}.$$

Damit ist unser ersters Iterationsverfahren vollständig definiert.

Definition 7.2 (Gradientenverfahren) *Die durch*

$$\lambda^{(m)} := \begin{cases} \frac{\|\mathbf{r}^{(m)}\|_2^2}{\langle \mathbf{r}^{(m)}, \mathbf{A}\mathbf{r}^{(m)}\rangle_2} & \textit{falls } \mathbf{r}^{(m)} \neq \mathbf{0}, \\ 0 & \textit{ansonsten}, \end{cases}$$
$$\mathbf{x}^{(m+1)} := \mathbf{x}^{(m)} + \lambda^{(m)}\mathbf{r}^{(m)} \qquad \textit{für alle } m \in \mathbb{N}_0$$

definierte Folge von Näherungslösungen bezeichnen wir als die Folge der Iterierten des Gradientenverfahrens*.*

Der Name des Verfahrens ist darauf zurück zu führen, dass $-\mathbf{r}^{(m)}$ gerade der Gradient $\nabla f(\mathbf{x}^{(m)})$ der zu minimierenden Funktion ist.

Falls $\mathbf{r}^{(m)} = \mathbf{0}$ gilt, muss $\mathbf{x}^{(m)}$ nach Definition bereits eine Lösung des Gleichungssystems (7.1) sein, also brauchen wir das Verfahren nicht weiter durchzuführen.

Anderenfalls impliziert die Optimalitätsbedingung (7.6), dass

$$f(\mathbf{x}^{(m+1)}) \leq f(\mathbf{x}^{(m)})$$

gelten muss, die Näherungslösungen können in diesem Sinne nur besser werden.

Um das Verfahren effizient durchführen zu können, bietet es sich an, darauf zu achten, dass die in der Regel zeitaufwendige Multiplikation mit der Matrix $\mathbf{A}$ nur einmal pro Iterationsschritt durchgeführt wird. Dieses Ziel lässt sich erreichen, indem wir die Hilfsvariable

$$\mathbf{a}^{(m)} := \mathbf{A}\mathbf{r}^{(m)} \qquad \text{für alle } m \in \mathbb{N}_0$$

einführen und

$$\begin{aligned} \mathbf{r}^{(m+1)} &= \mathbf{b} - \mathbf{A}\mathbf{x}^{(m+1)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(m)} - \lambda^{(m)}\mathbf{A}\mathbf{r}^{(m)} \\ &= \mathbf{r}^{(m)} - \lambda^{(m)}\mathbf{a}^{(m)} \qquad \text{für alle } m \in \mathbb{N}_0 \end{aligned}$$

für die Berechnung ausnutzen. Es ergibt sich der folgende Algorithmus:

$\mathbf{r} \leftarrow \mathbf{b} - \mathbf{A}\mathbf{x}$;
while $\|\mathbf{r}\|_2$ zu groß **do begin**
  $\mathbf{a} \leftarrow \mathbf{A}\mathbf{r}$;
  $\lambda \leftarrow \|\mathbf{r}\|_2^2 / \langle \mathbf{r}, \mathbf{a} \rangle_2$;
  $\mathbf{x} \leftarrow \mathbf{x} + \lambda \mathbf{r}$;
  $\mathbf{r} \leftarrow \mathbf{r} - \lambda \mathbf{a}$
end

Eine bemerkenswerte Eigenschaft dieses Verfahrens besteht darin, dass es lediglich Skalarprodukte, Linearkombinationen und die Auswertung der Matrix $\mathbf{A}$ für einen gegebenen Vektor erfordert. Etwa bei Finite-Differenzen-Verfahren sind alle Einträge der Matrix $\mathbf{A}$ a priori bekannt, so dass die Auswertung für einen beliebigen Vektor erfolgen kann, ohne die Matrix explizit speichern zu müssen. Dadurch wird nicht nur sehr viel Speicherplatz gespart, sondern auch die Geschwindigkeit des Algorithmus verbessert, weil auf modernen Computern Zugriffe auf den Hauptspeicher häufig sehr viel langsamer als Rechenoperationen ausgeführt werden.

Wir sind natürlich daran interessiert, eine quantitative Konvergenzaussage zu erhalten. Eine Norm, mit der sich die Konvergenz des Gradientenverfahrens besonders gut analysieren lässt, ist die *Energienorm*, die durch das *Energieskalarprodukt* definiert wird:

$$\langle \mathbf{y}, \mathbf{x} \rangle_A := \langle \mathbf{y}, \mathbf{A}\mathbf{x} \rangle_2, \qquad \|\mathbf{x}\|_A := \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_A} \qquad \text{für alle } \mathbf{x}, \mathbf{y} \in \mathbb{R}^{\mathcal{I}}.$$

Diese Norm passt besonders gut zu unserem Problem, weil

$$\begin{aligned} \|\mathbf{x} - \mathbf{x}^*\|_A^2 &= \langle \mathbf{x} - \mathbf{x}^*, \mathbf{A}(\mathbf{x} - \mathbf{x}^*) \rangle_2 = \langle \mathbf{x}, \mathbf{A}\mathbf{x} \rangle_2 - 2\langle \mathbf{x}, \mathbf{A}\mathbf{x}^* \rangle_2 + \langle \mathbf{x}^*, \mathbf{A}\mathbf{x}^* \rangle_2 \\ &= \langle \mathbf{x}, \mathbf{A}\mathbf{x} \rangle_2 - 2\langle \mathbf{x}, \mathbf{b} \rangle_2 + \langle \mathbf{x}^*, \mathbf{A}\mathbf{x}^* \rangle_2 = 2f(\mathbf{x}) + \|\mathbf{x}^*\|_A^2 \qquad \text{für alle } \mathbf{x} \in \mathbb{R}^{\mathcal{I}} \end{aligned}$$

gilt, bis auf den konstanten Term $\|\mathbf{x}^*\|_A^2$ stimmen also das Quadrat der Energienorm und die von uns minimierte Funktion überein, und insbesondere bedeutet eine Minimierung der Funktion f auch eine Minimierung der Energienorm.

Wir sind daran interessiert, die Entwicklung des Fehlers

$$\mathbf{e}^{(m)} := \mathbf{x}^* - \mathbf{x}^{(m)} \qquad \text{für alle } m \in \mathbb{N}_0 \tag{7.8}$$

zu untersuchen. Da $\mathbf{A}\mathbf{x}^* = \mathbf{b}$ gilt, erfüllt er die Gleichung

$$\mathbf{A}\mathbf{e}^{(m)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(m)} = \mathbf{r}^{(m)} \qquad \text{für alle } m \in \mathbb{N}_0,$$

und nach Definition der Iterationsvorschrift entwickelt er sich gemäß

$$\begin{aligned} \mathbf{e}^{(m+1)} &= \mathbf{x}^* - \mathbf{x}^{(m+1)} = \mathbf{x}^* - \mathbf{x}^{(m)} - \lambda^{(m)} \mathbf{r}^{(m)} \\ &= \mathbf{e}^{(m)} - \lambda^{(m)} \mathbf{A}\mathbf{e}^{(m)} \qquad \text{für alle } m \in \mathbb{N}_0. \end{aligned}$$

Wir haben $\lambda^{(m)}$ gerade so gewählt, dass

$$f(\mathbf{x}^{(m+1)}) = f(\mathbf{x}^{(m)} + \lambda^{(m)} \mathbf{r}^{(m)}) \leq f(\mathbf{x}^{(m)} + \lambda \mathbf{r}^{(m)}) \qquad \text{für alle } \lambda \in \mathbb{R}$$

gilt, und wie wir bereits gesehen haben, ist das äquivalent zu

$$\begin{aligned}\|\mathbf{e}^{(m+1)}\|_A^2 &= \|\mathbf{x}^{(m+1)} - \mathbf{x}^*\|_A^2 = 2f(\mathbf{x}^{(m+1)}) + \|\mathbf{x}^*\|_A^2 \\ &\leq 2f(\mathbf{x}^{(m)} + \lambda\mathbf{r}^{(m)}) + \|\mathbf{x}^*\|_A^2 = \|\mathbf{x}^* - \mathbf{x}^{(m)} - \lambda\mathbf{r}^{(m)}\|_A^2 \\ &= \|\mathbf{e}^{(m)} - \lambda\mathbf{A}\mathbf{e}^{(m)}\|_A^2 \qquad \text{für alle } \lambda \in \mathbb{R}. \end{aligned} \tag{7.9}$$

Mit Hilfe einer Eigenwertanalyse lässt sich aus dieser Abschätzung eine Aussage über die Konvergenz des Fehlers gewinnen:

Lemma 7.3 (Konvergenz) *Seien $\alpha, \beta \in \mathbb{R}_{>0}$ so gewählt, dass $\sigma(\mathbf{A}) \subseteq [\alpha, \beta]$ gilt, dass also alle Eigenwerte zwischen diesen Schranken liegen. Dann gilt*

$$\|\mathbf{e}^{(m+1)}\|_A \leq \varrho\|\mathbf{e}^{(m)}\|_A, \qquad \varrho := \frac{\beta - \alpha}{\beta + \alpha} < 1 \qquad \textit{für alle } m \in \mathbb{N}_0.$$

Insbesondere konvergieren die Näherungslösungen $\mathbf{x}^{(m)}$ gegen die exakte Lösung $\mathbf{x}^*$.

Beweis. Da $\mathbf{A}$ symmetrisch ist, ist die Matrix auch orthogonal diagonalisierbar, es existieren also eine orthogonale Matrix $\mathbf{Q} \in \mathbb{R}^{\mathcal{I}\times\mathcal{I}}$ und eine Diagonalmatrix $\mathbf{D} \in \mathbb{R}^{\mathcal{I}\times\mathcal{I}}$ mit

$$\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}^*, \qquad \mathbf{Q}^*\mathbf{Q} = \mathbf{I}.$$

Sei $m \in \mathbb{N}_0$. Wir bezeichnen mit

$$\widehat{\mathbf{e}}^{(m)} := \mathbf{Q}^*\mathbf{e}^{(m)}$$

den in die Eigenvektorbasis transformierten Fehler. Nach (7.9) gilt

$$\begin{aligned}\|\mathbf{e}^{(m+1)}\|_A^2 &\leq \|\mathbf{e}^{(m)} - \lambda\mathbf{A}\mathbf{e}^{(m)}\|_A^2 = \|(\mathbf{I} - \lambda\mathbf{A})\mathbf{e}^{(m)}\|_A^2 \\ &= \langle(\mathbf{I} - \lambda\mathbf{A})\mathbf{e}^{(m)}, \mathbf{A}(\mathbf{I} - \lambda\mathbf{A})\mathbf{e}^{(m)}\rangle_2 \\ &= \langle\mathbf{Q}(\mathbf{I} - \lambda\mathbf{D})\mathbf{Q}^*\mathbf{e}^{(m)}, \mathbf{Q}\mathbf{D}\mathbf{Q}^*\mathbf{Q}(\mathbf{I} - \lambda\mathbf{D})\mathbf{Q}^*\mathbf{e}^{(m)}\rangle_2 \\ &= \langle(\mathbf{I} - \lambda\mathbf{D})\widehat{\mathbf{e}}^{(m)}, \mathbf{D}(\mathbf{I} - \lambda\mathbf{D})\widehat{\mathbf{e}}^{(m)}\rangle_2.\end{aligned}$$

Da $\mathbf{D}$ eine Diagonalmatrix ist, müssen Eigenwerte $(\mu_i)_{i\in\mathcal{I}}$ mit

$$d_{ii} = \mu_i, \qquad \mu_i \in [\alpha, \beta] \qquad \text{für alle } i \in \mathcal{I}$$

existieren, und wir erhalten

$$\|\mathbf{e}^{(m+1)}\|_A^2 \leq \sum_{i\in\mathcal{I}} \mu_i(1 - \lambda\mu_i)^2(\hat{e}_i^{(m)})^2 = \sum_{i\in\mathcal{I}} \mu_i g_\lambda(\mu_i)(\hat{e}_i^{(m)})^2$$

für die Hilfsfunktion

$$g_\lambda : [\alpha, \beta] \to \mathbb{R}, \qquad \mu \mapsto (1 - \lambda\mu)^2.$$

Unsere Aufgabe besteht darin, den Parameter $\lambda \in \mathbb{R}_{>0}$ so zu wählen, dass g_λ auf $[\alpha, \beta]$ möglichst geringe Werte annimmt. Ein Blick auf die zweite Ableitung der Funktion legt nahe, dass sie ihr Maximum nur in den Randpunkten α und β annehmen kann. Aus

$$g_\lambda(\alpha) \leq g_\lambda(\beta) \iff 1 - 2\alpha\lambda + \alpha^2\lambda^2 \leq 1 - 2\beta\lambda + \beta^2\lambda^2 \iff -2\alpha + \alpha^2\lambda \leq -2\beta + \beta^2\lambda$$
$$\iff 2(\beta - \alpha) \leq (\beta^2 - \alpha^2)\lambda \iff \frac{2}{\beta + \alpha} = \frac{2(\beta - \alpha)}{\beta^2 - \alpha^2} \leq \lambda$$

folgt, dass das Maximum für $\lambda \leq \lambda_0 := 2/(\beta + \alpha)$ im linken und anderenfalls im rechten Randpunkt angenommen wird. Aufgrund der Abschätzungen

$$\frac{\partial}{\partial\lambda} g_\lambda(\alpha) = 2\alpha(\alpha\lambda - 1) \leq 2\alpha\frac{2\alpha - \alpha - \beta}{\beta + \alpha} = 2\alpha\frac{\alpha - \beta}{\beta + \alpha} \leq 0 \qquad \text{für alle } \lambda \leq \frac{2}{\beta + \alpha},$$
$$\frac{\partial}{\partial\lambda} g_\lambda(\beta) = 2\beta(\beta\lambda - 1) \geq 2\beta\frac{2\beta - \alpha - \beta}{\beta + \alpha} = 2\beta\frac{\beta - \alpha}{\beta + \alpha} \geq 0 \qquad \text{für alle } \lambda \geq \frac{2}{\beta + \alpha}$$

fällt das Maximum monoton, bis wir λ_0 erreichen, um dann wieder monoton zu wachsen. Also ist

$$\lambda := \frac{2}{\beta + \alpha}$$

die bestmögliche Wahl ist. So erhalten wir

$$g_\lambda(\beta) = g_\lambda(\alpha) = \left(1 - \frac{2\alpha}{\beta + \alpha}\right)^2 = \left(\frac{\beta + \alpha - 2\alpha}{\beta + \alpha}\right)^2 = \left(\frac{\beta - \alpha}{\beta + \alpha}\right)^2 = \varrho^2$$

und schließlich

$$\|\mathbf{e}^{(m+1)}\|_A^2 \leq \sum_{i\in\mathcal{I}} \mu_i g_\lambda(\mu_i)(\hat{e}_i^{(m)})^2 \leq \varrho^2 \sum_{i\in\mathcal{I}} \mu_i(\hat{e}_i^{(m)})^2 = \varrho^2\langle\widehat{\mathbf{e}}^{(m)}, \mathbf{D}\widehat{\mathbf{e}}^{(m)}\rangle_2$$
$$= \varrho^2\langle\mathbf{Q}^*\mathbf{e}^{(m)}, \mathbf{D}\mathbf{Q}^*\mathbf{e}^{(m)}\rangle_2 = \varrho^2\langle\mathbf{e}^{(m)}, \mathbf{A}\mathbf{e}^{(m)}\rangle_2 = \varrho^2\|\mathbf{e}^{(m)}\|_A^2.$$

Die gewünschte Konvergenzaussage folgt, indem wir die Wurzel aus dieser Abschätzung ziehen. ∎

Die bestmögliche Wahl für die Parameter α und β in dieser Konvergenzaussage sind offenbar der kleinste und größte Eigenwert der Matrix $\mathbf{A}$. Mit Hilfe der orthogonalen Diagonalisierung lassen sich diese optimalen Parameter durch

$$\alpha = \frac{1}{\|\mathbf{A}^{-1}\|_2}, \qquad \beta = \|\mathbf{A}\|_2$$

darstellen, und die Schranke der Fehlerreduktion nimmt so die Form

$$\varrho = \frac{\beta - \alpha}{\beta + \alpha} = \frac{\beta/\alpha - 1}{\beta/\alpha + 1} = \frac{\|\mathbf{A}\|_2\|\mathbf{A}^{-1}\|_2 - 1}{\|\mathbf{A}\|_2\|\mathbf{A}^{-1}\|_2 + 1} = \frac{\kappa_2(\mathbf{A}) - 1}{\kappa_2(\mathbf{A}) + 1} = 1 - \frac{2}{\kappa_2(\mathbf{A}) + 1}$$

an, das Gradientenverfahren wird also um so langsamer konvergieren, je größer die Konditionszahl $\kappa_2(\mathbf{A})$ der Matrix $\mathbf{A}$ ist. Das ist ein sehr ungünstiges Verhalten, da bei der Behandlung partieller Differentialgleichungen in der Regel die Konditionszahl wächst, wenn die Dimension des Gleichungssystems wächst, so dass wir bei größeren Matrizen langsamere Konvergenz erwarten müssen.

7.2 Verfahren der konjugierten Gradienten

Die relativ schlechten Konvergenzeigenschaften des Gradientenverfahrens sind auf schlecht gewählte Suchrichtungen zurückzuführen: Wir wählen $\lambda^{(m)}$ gerade so, dass die Optimalitätsbedingung

$$0 = \langle \mathbf{r}^{(m)}, \mathbf{b} - \mathbf{A}(\mathbf{x}^{(m)} + \lambda^{(m)}\mathbf{p}^{(m)})\rangle_2 = \langle \mathbf{r}^{(m)}, \mathbf{b} - \mathbf{A}\mathbf{x}^{(m+1)}\rangle_2 = \langle \mathbf{r}^{(m)}, \mathbf{r}^{(m+1)}\rangle_2$$

gilt. Da wir in jedem Schritt das Residuum als Suchrichtung wählen, stehen alle Suchrichtungen senkrecht aufeinander. Das bedeutet beispielsweise für ein zweidimensionales Problem, dass alle geradzahligen und alle ungeradzahligen Suchrichtungen jeweils voneinander linear abhängig sein müssen, wir werden also immer wieder in derselben Richtung zu optimieren versuchen.

Im folgenden nennen wir eine Iterierte $\mathbf{x}^{(m)}$ *optimal* bezüglich einer Richtung $\mathbf{p}^{(\ell)}$, falls

$$f(\mathbf{x}^{(m)}) \leq f(\mathbf{x}^{(m)} + \lambda\mathbf{p}^{(\ell)}) \qquad \text{für alle } \lambda \in \mathbb{R}$$

gilt, falls sich also $\mathbf{x}^{(m)}$ nicht durch zu Hinzuaddieren eines Vielfachen der Richtung $\mathbf{p}^{(\ell)}$ verbessern lässt. Nach Lemma 7.1 ist das genau dann der Fall, wenn

$$0 = \langle \mathbf{p}^{(\ell)}, \mathbf{b} - \mathbf{A}\mathbf{x}^{(m)}\rangle_2$$

gilt. Unser Ziel ist es nun, die Suchrichtungen so zu wählen, dass eine einmal bezüglich einer Richtung erreichte Optimalität nicht wieder verloren geht. Nach dem ersten Schritt des Gradientenverfahrens ist durch die Wahl der Schrittweite $\lambda^{(0)}$ sichergestellt, dass $\mathbf{x}^{(1)}$ optimal bezüglich der im vorangehenden Schritt verwendeten Richtung $\mathbf{p}^{(0)} = \mathbf{r}^{(0)}$ ist. Die nächster Iterierte ist von der Form

$$\mathbf{x}^{(2)} = \mathbf{x}^{(1)} + \lambda^{(1)}\mathbf{p}^{(1)},$$

und falls wir sicherstellen wollen, dass auch sie noch optimal bezüglich der Suchrichtung $\mathbf{p}^{(0)}$ ist, müssen wir dafür sorgen, dass

$$\begin{aligned} 0 &= \langle \mathbf{p}^{(0)}, \mathbf{b} - \mathbf{A}\mathbf{x}^{(2)}\rangle_2 = \langle \mathbf{p}^{(0)}, \mathbf{b} - \mathbf{A}(\mathbf{x}^{(1)} + \lambda^{(1)}\mathbf{p}^{(1)})\rangle_2 \\ &= \langle \mathbf{p}^{(0)}, \mathbf{b} - \mathbf{A}\mathbf{x}^{(1)}\rangle_2 - \lambda^{(1)}\langle \mathbf{p}^{(0)}, \mathbf{A}\mathbf{p}^{(1)}\rangle_2 \end{aligned}$$

erfüllt ist. Da $\mathbf{x}^{(1)}$ optimal bezüglich $\mathbf{p}^{(0)}$ ist, fällt der erste Term weg und wir erhalten

$$0 = \lambda^{(1)}\langle \mathbf{p}^{(0)}, \mathbf{A}\mathbf{p}^{(1)}\rangle_2.$$

Da $\lambda^{(1)} = 0$ zu keiner Verbesserung der Iterierten führen würde, muss also

$$0 = \langle \mathbf{p}^{(0)}, \mathbf{A}\mathbf{p}^{(1)}\rangle_2 = \langle \mathbf{p}^{(0)}, \mathbf{p}^{(1)}\rangle_A$$

gelten, die Suchrichtung $\mathbf{p}^{(1)}$ muss bezüglich des *Energieskalarprodukts* senkrecht auf $\mathbf{p}^{(0)}$ stehen. Solche Richtungen bezeichnet man als zueinander *konjugiert*.

Diese Beobachtung legt es nahe, dafür zu sorgen, dass alle Suchrichtungen konjugiert zueinander sind, denn dann ist sichergestellt, dass die Iterierte $\mathbf{x}^{(m)}$ bezüglich aller vorangehender Suchrichtungen $\mathbf{p}^{(0)}, \ldots, \mathbf{p}^{(m-1)}$ optimal ist. Indem wir das Orthogonalisierungsverfahren von Gram und Schmidt jeweils auf das Residuum $\mathbf{r}^{(m)}$ anwenden, erhalten wir

$$\mathbf{p}^{(m)} := \mathbf{r}^{(m)} - \sum_{\ell=0}^{m-1} \frac{\langle \mathbf{p}^{(\ell)}, \mathbf{r}^{(m)} \rangle_A}{\langle \mathbf{p}^{(\ell)}, \mathbf{p}^{(\ell)} \rangle_A} \mathbf{p}^{(\ell)} \qquad \text{für alle } m \in \{0, \ldots, m_0\}. \tag{7.10}$$

Hier gibt die Konstante

$$m_0 := \min\{m \in \mathbb{N}_0 \ : \ \mathbf{p}^{(m)} = \mathbf{0}\}$$

an, nach wievielen Schritten wir keine neue Suchrichtung mehr finden können. Aus (7.10) folgt aus $\mathbf{p}^{(m_0)} = \mathbf{0}$ allerdings

$$\mathbf{r}^{(m_0)} \in \operatorname{span}\{\mathbf{p}^{(0)}, \ldots, \mathbf{p}^{(m_0-1)}\},$$

und da $\mathbf{x}^{(m_0)}$ optimal bezüglich aller vorangehenden Suchrichtungen ist, muss es damit auch bezüglich der Richtung $\mathbf{r}^{(m_0)}$ optimal sein, so dass

$$\|\mathbf{r}^{(m_0)}\|_2^2 = \langle \mathbf{r}^{(m_0)}, \mathbf{r}^{(m_0)} \rangle_2 = \langle \mathbf{r}^{(m_0)}, \mathbf{b} - \mathbf{A}\mathbf{x}^{(m_0)} \rangle_2 = 0$$

folgt. Also muss $\mathbf{x}^{(m_0)} = \mathbf{x}^*$ gelten. Kurz gesagt: Falls wir keine neue Suchrichtung mehr konstruieren können, müssen wir auch keine weiteren Schritte mehr durchführen.

Für die praktische Umsetzung des Verfahrens ist die Definition (7.10) unpraktisch, denn ihre direkte Auswertung erfordert m Skalarprodukte und Linearkombinationen, so dass der Aufwand von Schritt zu Schritt wachsen würde. Glücklicherweise lässt sich die Orthogonalisierung etwas eleganter gestalten.

Definition 7.4 (Krylow-Raum) *Sei $\mathbf{z} \in \mathbb{R}^{\mathcal{I}}$. Wir bezeichnen mit*

$$\mathcal{K}_m(\mathbf{z} = := \operatorname{span}\{\mathbf{z}, \mathbf{A}\mathbf{z}, \ldots, \mathbf{A}^m\mathbf{z}\} \qquad \textit{für alle } m \in \mathbb{N}_0$$

den m-ten Krylow-Raum *zu dem Startvektor $\mathbf{z}$ und der Matrix $\mathbf{A}$.*

Lemma 7.5 (Teilräume) *Für alle $m \in \{0, \ldots, m_0 - 1\}$ gilt*

$$\operatorname{span}\{\mathbf{p}^{(0)}, \ldots, \mathbf{p}^{(m)}\} = \operatorname{span}\{\mathbf{r}^{(0)}, \ldots, \mathbf{r}^{(m)}\} = \mathcal{K}_m(\mathbf{r}^{(0)}).$$

Beweis. Aus der Definition der Suchrichtungen $\mathbf{p}^{(\ell)}$ folgt mit einer einfachen Induktion

$$\operatorname{span}\{\mathbf{p}^{(0)}, \ldots, \mathbf{p}^{(m-1)}\} \subseteq \operatorname{span}\{\mathbf{r}^{(0)}, \ldots, \mathbf{r}^{(m-1)}\} \quad \text{für alle } m \in \{0, \ldots, m_0 - 1\}. \tag{7.11}$$

Wie im Falle des Gradientenverfahrens gilt

$$\mathbf{r}^{(m+1)} = \mathbf{b} - \mathbf{A}(\mathbf{x}^{(m)} + \lambda^{(m)}\mathbf{p}^{(m)}) = \mathbf{r}^{(m)} - \lambda^{(m)}\mathbf{A}\mathbf{p}^{(m)} \qquad \text{für alle } m \in \mathbb{N}_0,$$

also erhalten wir mit (7.11) und einer weiteren einfachen Induktion auch

$$\operatorname{span}\{\mathbf{r}^{(0)},\dots,\mathbf{r}^{(m-1)}\} \subseteq \mathcal{K}_m(\mathbf{r}^{(0)}) \qquad \text{für alle } m \in \{0,\dots,m_0-1\}. \tag{7.12}$$

Sei nun $m \in \{0,\dots,m_0-1\}$. Da die Vektoren $\mathbf{p}^{(0)},\dots,\mathbf{p}^{(m)}$ bezüglich des Energieskalarprodukts paarweise senkrecht aufeinander stehen und nach Definition der Zahl m_0 keiner von ihnen gleich null ist, müssen sie linear unabhängig sein, also gilt

$$\dim \operatorname{span}\{\mathbf{p}^{(0)},\dots,\mathbf{p}^{(m)}\} = m+1,$$

und aus der Definition des Krylow-Raums erhalten wir

$$\dim \mathcal{K}_m(\mathbf{r}^{(0)}) \leq m+1,$$

so dass wir durch Kombination der Inklusionen (7.11) und (7.12) auf die gewünschte Identität der drei Teilräume schließen dürfen. ∎

Wir wählen $m \in \{0,\dots,m_0-1\}$ und $\ell \in \{0,\dots,m-2\}$. Aus Lemma 7.5 folgt

$$\mathbf{A}\mathbf{p}^{(\ell)} \in \{\mathbf{A}\mathbf{z} \;:\; \mathbf{z} \in \mathcal{K}_\ell(\mathbf{r}^{(0)})\} \subseteq \mathcal{K}_{\ell+1}(\mathbf{r}^{(0)}) = \operatorname{span}\{\mathbf{p}^{(0)},\dots,\mathbf{p}^{(\ell+1)}\}.$$

Da $\mathbf{x}^{(m)}$ optimal bezüglich der Suchrichtungen $\mathbf{p}^{(0)},\dots,\mathbf{p}^{(m-1)}$ ist und $\ell+1 \leq m-1$ vorausgesetzt ist, erhalten wir

$$0 = \langle \mathbf{A}\mathbf{p}^{(\ell)}, \mathbf{b}-\mathbf{A}\mathbf{x}^{(m)}\rangle_2 = \langle \mathbf{A}\mathbf{p}^{(\ell)}, \mathbf{r}^{(m)}\rangle_2 = \langle \mathbf{r}^{(m)}, \mathbf{p}^{(\ell)}\rangle_A = \langle \mathbf{p}^{(\ell)}, \mathbf{r}^{(m)}\rangle_A,$$

indem wir im letzten Schritt die Symmetrie der Matrix $\mathbf{A}$ ausnutzen. Dank dieser Gleichung verschwinden fast alle Summanden aus der Gleichung (7.10), so dass lediglich

$$\mathbf{p}^{(m)} := \begin{cases} \mathbf{r}^{(0)} & \text{falls } m=0, \\ \mathbf{r}^{(m)} - \frac{\langle \mathbf{p}^{(m-1)}, \mathbf{r}^{(m)}\rangle_A}{\langle \mathbf{p}^{(m-1)}, \mathbf{p}^{(m-1)}\rangle_A}\mathbf{p}^{(m-1)} & \text{ansonsten} \end{cases} \qquad \text{für alle } m \in \{0,\dots,m_0-1\}$$

übrig bleibt. Diese Formel lässt sich effizient verwenden.

Definition 7.6 (Verfahren der konjugierten Gradienten) *Die durch*

$$\begin{aligned} \mathbf{p}^{(m)} &:= \begin{cases} \mathbf{r}^{(0)} & \textit{falls } m=0, \\ \mathbf{r}^{(m)} - \frac{\langle \mathbf{r}^{(m)}, \mathbf{A}\mathbf{p}^{(m-1)}\rangle_2}{\langle \mathbf{p}^{(m-1)}, \mathbf{A}\mathbf{p}^{(m-1)}\rangle_2}\mathbf{p}^{(m-1)} & \textit{ansonsten} \end{cases}, \\ \lambda^{(m)} &:= \frac{\langle \mathbf{p}^{(m)}, \mathbf{r}^{(m)}\rangle_2}{\langle \mathbf{p}^{(m)}, \mathbf{A}\mathbf{p}^{(m)}\rangle_2}, \\ \mathbf{x}^{(m+1)} &:= \mathbf{x}^{(m)} + \lambda^{(m)}\mathbf{p}^{(m)} \qquad \textit{für alle } m \in \{0,\dots,m_0-1\} \end{aligned}$$

definierte Folge von Näherungslösungen bezeichnen wir als die Folge der Iterierten des Verfahrens der konjugierten Gradienten.

Das Verfahren der konjugierten Gradienten wird häufig auch als *cg-Verfahren* bezeichnet, nach dem englischen *conjugate gradients*.

Indem wir diesmal

$$\mathbf{a}^{(m)} := \mathbf{A}\mathbf{p}^{(m)} \qquad \text{für alle } m \in \{0, \dots, m_0 - 1\}$$

definieren, können wir auch das Verfahren der konjugierten Gradienten mit einer einzigen Matrix-Vektor-Multiplikation pro Iterationsschritt durchführen:

$\mathbf{r} \leftarrow \mathbf{b} - \mathbf{A}\mathbf{x}$; $\mathbf{p} \leftarrow \mathbf{r}$;
while $\|\mathbf{r}\|_2$ zu groß **do begin**
  $\mathbf{a} \leftarrow \mathbf{A}\mathbf{p}$;
  $\lambda \leftarrow \langle \mathbf{r}, \mathbf{a} \rangle_2 / \langle \mathbf{p}, \mathbf{a} \rangle_2$;
  $\mathbf{x} \leftarrow \mathbf{x} + \lambda \mathbf{p}$;
  $\mathbf{r} \leftarrow \mathbf{r} - \lambda \mathbf{a}$;
  $\mu \leftarrow \langle \mathbf{r}, \mathbf{a} \rangle_2 / \langle \mathbf{p}, \mathbf{a} \rangle_2$;
  $\mathbf{p} \leftarrow \mathbf{r} - \mu \mathbf{p}$
end

Bei der Analyse des Richardson-Verfahrens haben wir die Konvergenzrate abgeschätzt, indem wir uns die Optimalität der Schrittweite zunutze machten. Für das cg-Verfahren erhalten wir die folgende wesentlich stärkere Aussage:

Lemma 7.7 *Für die wieder gemäß (7.8) definierten Fehler des cg-Verfahrens gilt*

$$\|\mathbf{e}^{(m)}\|_A \leq \|p(\mathbf{A})\mathbf{e}^{(0)}\|_A \qquad \textit{für alle } m \in \mathbb{N}_0 \textit{ und } p \in \Pi_m \textit{ mit } p(0) = 1.$$

Beweis. Sei $m \in \mathbb{N}_0$, und sei $p \in \Pi_m$ mit $p(0) = 1$ gegeben. Dann existieren Koeffizienten $\gamma_0, \dots, \gamma_m \in \mathbb{R}$ mit

$$p(\xi) = \sum_{\ell=0}^{m} \gamma_\ell \xi^\ell \qquad \text{für alle } \xi \in \mathbb{R}.$$

Aus $p(0) = 1$ folgt durch Einsetzen in die Gleichung direkt $\gamma_0 = 1$, und wir erhalten

$$p(\mathbf{A})\mathbf{e}^{(0)} = \gamma_0 \mathbf{e}^{(0)} + \sum_{\ell=1}^{m} \gamma_\ell \mathbf{A}^\ell \mathbf{e}^{(0)} = \mathbf{e}^{(0)} + \sum_{\ell=0}^{m-1} \gamma_{\ell+1} \mathbf{A}^{\ell+1} \mathbf{e}^{(0)} = \mathbf{e}^{(0)} + \sum_{\ell=0}^{m-1} \gamma_{\ell+1} \mathbf{A}^\ell \mathbf{r}^{(0)}.$$

Wir setzen

$$\mathbf{c} := \mathbf{x}^{(m)} - \mathbf{x}^{(0)} - \sum_{\ell=0}^{m-1} \gamma_{\ell+1} \mathbf{A}^\ell \mathbf{r}^{(0)}.$$

Da $\mathbf{x}^{(m)} - \mathbf{x}^{(0)} \in \operatorname{span}\{\mathbf{p}^{(0)}, \dots, \mathbf{p}^{(m-1)}\}$ nach Konstruktion gilt, folgt mit Lemma 7.5 auch die Beziehung

$$\mathbf{c} \in \operatorname{span}\{\mathbf{p}^{(0)}, \dots, \mathbf{p}^{(m-1)}\}.$$

Da $\mathbf{x}^{(m)}$ optimal bezüglich der Richtungen $\mathbf{p}^{(0)}, \ldots, \mathbf{p}^{(m-1)}$ ist, muss es auch bezüglich $\mathbf{c}$ optimal sein, und es folgt

$$\begin{aligned}\|\mathbf{e}^{(m)}\|_A^2 &= 2f(\mathbf{x}^{(m)}) + \|\mathbf{x}^*\|_A^2 \leq 2f(\mathbf{x}^{(m)} + \mathbf{c}) + \|\mathbf{x}^*\|_A^2 = \|\mathbf{e}^{(m)} - \mathbf{c}\|_A^2 \\ &= \|\mathbf{x}^* - \mathbf{x}^{(m)} - \mathbf{c}\|_A^2 = \left\|\mathbf{x}^* - \mathbf{x}^{(0)} + \sum_{\ell=0}^{m-1} \gamma_{\ell+1} \mathbf{A}^\ell \mathbf{r}^{(0)}\right\|_A^2 = \|p(\mathbf{A})\mathbf{e}^{(0)}\|_A^2.\end{aligned}$$

∎

Aus der Optimalität des Fehlers können wir durch geschickte Wahl des Polynoms p eine explizite Konvergenzaussage erhalten:

Satz 7.8 (Konvergenz) *Seien $\alpha, \beta \in \mathbb{R}_{>0}$ so gewählt, dass $\sigma(\mathbf{A}) \subseteq [\alpha, \beta]$ gilt. Dann folgt*

$$\|\mathbf{e}^{(m)}\|_A \leq \frac{2c^m}{1 + c^{2m}} \|\mathbf{e}^{(0)}\|_A, \qquad c := \frac{\sqrt{\beta/\alpha} - 1}{\sqrt{\beta/\alpha} + 1} < 1 \qquad \text{für alle } m \in \mathbb{N}_0.$$

Beweis. Wir diskutieren hier nur die grundlegende Idee. Mit Hilfe der Abbildung

$$\Phi : [\alpha, \beta] \to [-1, 1], \qquad t \mapsto 2\frac{t - \alpha}{\beta - \alpha} - 1,$$

wird das m-te Tschebyscheff-Polynom $T_m \in \Pi_m$ so transformiert, dass

$$p := \frac{T_m \circ \Phi}{T_m \circ \Phi(0)}$$

auf $[\alpha, \beta]$ besonders kleine Wert annimmt. Die Skalierung sorgt dafür, dass $p(0) = 1$ gilt. Durch eine detaillierte Analyse des Ausdrucks $T_m \circ \Phi(0)$ folgt die Behauptung. ∎

Bemerkung 7.9 *Es lässt sich einfach nachprüfen, dass*

$$\frac{2c}{1 + c^2} = \frac{\beta - \alpha}{\beta + \alpha}$$

gilt, im ersten Schritt wird also das cg-Verfahren genauso gut beziehungsweise schlecht wie das Gradientenverfahren sein.

Für die bestmögliche Wahl der Parameter $\alpha, \beta \in \mathbb{R}$ folgt

$$\frac{\sqrt{\beta/\alpha} - 1}{\sqrt{\beta/\alpha} + 1} = \frac{\sqrt{\kappa_2(\mathbf{A})} - 1}{\sqrt{\kappa_2(\mathbf{A})} + 1},$$

die Konditionszahl geht also in die Konvergenzaussagen des cg-Verfahrens nur über die Quadratwurzel ein, so dass die Konvergenzgeschwindigkeit wesentlich weniger empfindlich auf eine Verschlechterung der Kondition des Gleichungssystems reagiert.

Index

Literaturverzeichnis

[1] S. Banach. Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. *Fund. Math.*, 3:133–181, 1922.

[2] R. Bellman. The stability of solutions of linear differential equations. *Duke Math J.*, 10(4):543–647, 1943.

[3] R. Courant. *Differential and Integral Calculus*, volume 2. Blackie & Son, 1936.

[4] W. Dahmen and A. Reusken. *Numerik für Ingenieure und Naturwissenschaftler.* Springer, 2006.

[5] T. H. Gronwall. Note on the derivatives with respect to a parameter of the solutions of a system of differential equations. *Ann. of Math.*, 20(2):292–296, 1919.

[6] W. Hackbusch. *Elliptic Differential Equations. Theory and Numerical Treatment.* Springer-Verlag Berlin, 1992.

[7] R. Hooke. *Lectures de potentia restitutiva, or, of Spring: explaining the power of spring bodies: to which are added some collections.* John Martyn, Printer to the Royal Society, 1678.

[8] I. Newton. *Philosophiæ Naturalis Principia Mathematica.* 1687.

[9] H. Triebel. *Higher Analysis.* Barth, Heidelberg, 1997.

[10] W. Walter. *Gewöhnliche Differentialgleichungen.* Springer, 1993.

[11] Wikipedia. Differential algebraic equation — wikipedia, the free encyclopedia, 2010. [Online; accessed 31-May-2010].