Numerik Partieller Differentialgleichungen
Eine elementare Einführung

Volker John

Sommersemester 2009

Inhaltsverzeichnis

Teil I

Lineare Zwei–Punkt–Randwertprobleme

Kapitel 1

Analytisches Verhalten der Lösung

1.1 Das Modellproblem

Definition 1.1 Lineares Zwei–Punkt–Randwertproblem. Ein lineares Zwei–Punkt–Randwertproblem besitzt die Gestalt

$$-\varepsilon u'' + b(x)u' + c(x)u = f(x), \quad \text{für } x \in (d, e), \tag{1.1}$$

mit den Randbedingungen

$$\begin{aligned} \alpha_d u(d) - \beta_d u'(d) &= \gamma_d, \\ \alpha_e u(e) - \beta_e u'(e) &= \gamma_e. \end{aligned} \tag{1.2}$$

Hierbei gelte $b, c, f \in C([e, d])$, $0 < \varepsilon \in \mathbb{R}$ und die Konstanten $\alpha_d, \alpha_e, \beta_d, \beta_e, \gamma_d, \gamma_e$ sind gegeben. □

Bemerkung 1.2 Bedeutung linearer Zwei–Punkt–Randwertprobleme. Das Randwertproblem (1.1), (1.2) ist das einfachste Modellproblem zur Beschreibung von Prozessen, welche Diffusion und Transport beinhalten.

Ein Beispiel aus [Goe77] ist wie folgt. Fließt einem Stömungsreaktor bei konstanter Temperatur kontinuierlich eine Reaktionsmasse zu und ein Produkt ab, so berechnet sich die Konzentrationsverteilung $c(t, x, y, z)$ im Reaktor gemäß der partiellen Differentialgleichung

$$\frac{\partial c}{\partial t} - \operatorname{div}(D \operatorname{grad} c) + \operatorname{div}(c\mathbf{u}) = r(c),$$

wobei $\mathbf{u}$ der Vektor der Strömungsgeschwindigkeit, $r(c)$ eine die Reaktion beschreibende Funktion und D der Diffusionskoeffizient sind. Bei einem stationären Reaktorbetrieb, das heißt die zeitliche Änderung ist sehr langsam und kann vernachlässigt werden, bei konstanten Parametern D, $\mathbf{u}$ und wenn die Konzentration sich nur in x–Richtung ändert, erhält man aus der partiellen Differentialgleichung eine gewöhnliche Differentialgleichung für $c(x)$

$$-Dc''(x) + uc'(x) = r(c(x)).$$

Sei $x \in [0, L]$, wobei L die Reaktorlänge bezeichne. Mit den dimensionslosen Größen

$$\xi := \frac{x}{L}, \quad \gamma := \frac{c}{c_0},$$

wobei c_0 eine Referenzkonzentration ist, gelangt man zu einer dimensionslosen gewöhnlichen Differentialgleichung. Es sind mit Kettenregel

$$\frac{d\gamma(\xi)}{d\xi} = \frac{d(c(x)/c_0)}{dx}\frac{dx}{d\xi} = L\frac{c'(x)}{c_0}, \quad \frac{d^2\gamma(\xi)}{d\xi^2} = L^2\frac{c''(x)}{c_0}.$$

Einsetzen in die Differentialgleichung ergibt, im Fall $u \neq 0$,

$$-\frac{1}{\mathrm{Pe}}\gamma''(\xi) + \gamma'(\xi) = \rho(\gamma(\xi)),\ \xi \in (0,1), \quad \text{mit} \quad \mathrm{Pe} := \frac{uL}{D},\ \rho = \frac{L}{uc_0}r.$$

Die Zahl Pe wird Péclet[1]–Zahl genannt. Zur Vervollständigung der Problemstellung sind jetzt noch Randbedingungen für $\xi \in \{0,1\}$ nötig.

Aus der eben beschriebenen Anwendung heraus werden die Terme in (1.1) wie folgt genannt:

- $-\varepsilon u''$ – Diffusionsterm,
- $b(x)u'$ – Konvektions–, Advektions– oder Transportterm,
- $c(x)u$ – Reaktionsterm.

Das Modellproblem (1.1), (1.2) wird Konvektions–Diffusions–Problem genannt, falls $b(x) \not\equiv 0$.

Die Péclet–Zahl gibt das Verhältnis von Konvektion und Diffusion an. Falls dieses Verhältnis groß ist, wird dies in der numerischen Lösung von (1.1), (1.2) zu erheblichen Schwierigkeiten führen. □

Definition 1.3 Randbedingungen. Seien $\gamma_d, \gamma_e \in \mathbb{R}$, $\alpha_d, \alpha_e \in \mathbb{R} \setminus \{0\}$. Randbedingungen der Gestalt:

1. $$u(d) = \gamma_d, \quad u(e) = \gamma_e$$
   heißen Randbedingungen erster Art oder Dirichlet[2]–Randbedingungen,
2. $$u'(d) = \gamma_d, \quad u'(e) = \gamma_e$$
   heißen Randbedingungen zweiter Art oder Neumann[3]–Randbedingungen,
3. $$\alpha_d u(d) + u'(d) = \gamma_d, \quad \alpha_e u(e) + u'(e) = \gamma_e$$
   heißen Randbedingungen dritter Art oder Robin[4]–Randbedingungen.

□

Bemerkung 1.4 Normierung eines linearen Zwei–Punkt–Randwertproblems.

- Man kann ohne Beschränkung der Allgemeinheit $x \in [0,1]$ annehmen. Das erreicht man durch die Transformation
$$x \mapsto \frac{x-d}{e-d}.$$
- Man kann ebenfalls ohne Beschränkung der Allgemeinheit homogene Randbedingungen $\gamma_d = \gamma_e = 0$ annehmen, indem man von $u(x)$ eine glatte Funktion $\psi(x)$, welche die ursprünglichen Randbedingungen erfüllt, subtrahiert. Sind beispielsweise Dirichlet–Randbedingungen
$$u(d) = \gamma_d, \quad u(e) = \gamma_e,$$

[1] Jean Claude Eugene Péclet (1793 – 1857)
[2] Johann Peter Gustav Lejeune Dirichlet (1805 – 1859)
[3] Carl Gottfried Neumann (1832 – 1925)
[4] Gustave Robin (1855 – 1897)

gegeben, dann setzt man

$$\psi(x) = \gamma_d \frac{x-e}{d-e} + \gamma_e \frac{x-d}{e-d}$$

und

$$u^*(x) = u(x) - \psi(x).$$

Dann ist $u^*(x)$ die Lösung eines linearen Zwei–Punkt–Randwertproblems mit homogenen Dirichlet–Randbedingungen.
Dirichlet–Randbedingungen sind

- in Anwendungen am wichtigsten,
- vom Standpunkt der Analysis am schwierigsten.

Deshalb werden wir uns auf diese konzentrieren.

□

Definition 1.5 Modellproblem. Das Modellproblem besitzt die Gestalt

$$Lu := -\varepsilon u'' + b(x)u' + c(x)u = f(x) \quad \text{für } x \in (0,1), \tag{1.3}$$

mit den Randbedingungen

$$u(0) = u(1) = 0. \tag{1.4}$$

Hierbei gelte $b, c, f \in C([0,1])$, $0 < \varepsilon \in \mathbb{R}$. □

Bemerkung 1.6 Differentialoperator. In (1.3) bezeichnet L einen Differentialoperator. Unter einem Operator versteht man eine Abbildung zwischen zwei (Funktionen–)Räumen. Insoweit ist der Begriff des Operators synonym zum Begriff der Abbildung. Ein linearer Operator ist eine lineare Abbildung A auf einem linearen Raum X, so dass

$$A(\alpha u + \beta v) = \alpha Au + \beta Av$$

für alle Skalare α, β und alle $u, v \in X$ ist. Ein Differentialoperator ist ein Operator, der, angewandt auf geeignete Funktionen, Ableitungen enthält. Zur vollständigen Definition eines Operators ist dessen Definitionsbereich anzugeben. □

Beispiel 1.7 Das Randwertproblem

$$-\varepsilon u'' + u' = 1 \quad \text{auf } (0,1), \quad u(0) = u(1) = 0$$

besitzt die Lösung (vorrechnen)

$$u(x) = x - \frac{\exp\left(-\frac{1-x}{\varepsilon}\right) - \exp\left(-\frac{1}{\varepsilon}\right)}{1 - \exp\left(-\frac{1}{\varepsilon}\right)}.$$

Je kleiner der Parameter ε ist, umso steiler wird die Lösung in der Nähe des rechten Randes, siehe Abbildung 1.1. Diesen Teil der Lösung nennt man Grenzschicht. Solche starken Änderungen der Lösung auf einem sehr kleinen Gebiet führen zu Komplikationen bei der numerischen Approximation der Lösung.

□

Bemerkung 1.8 Transformation des Modellproblems auf ein symmetrisches Problem. Sei $b(x)$ hinreichend glatt. Definiert man

$$\tilde{u}(x) := u(x) \exp\left(-\frac{1}{2\varepsilon} \int_0^x b(\xi)\, d\xi\right), \quad x \in [0,1],$$

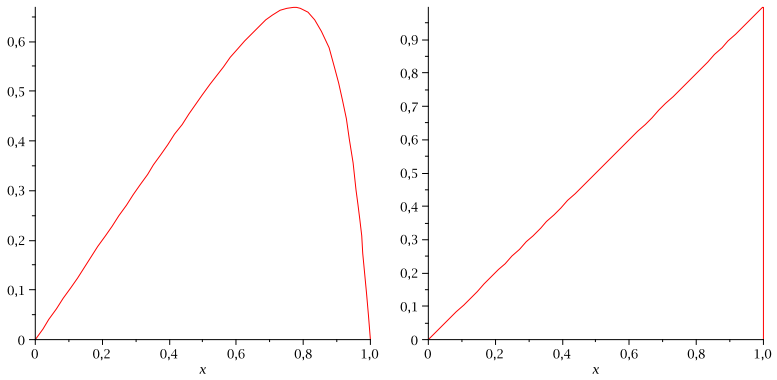


Abbildung 1.1: Lösung für $\varepsilon = 0.1$ links und $\varepsilon = 0.0001$ rechts.

so kann man (1.3), (1.4) in das symmetrische Problem

$$-\varepsilon\tilde{u}''(x) + \tilde{c}(x)\tilde{u}(x) = \tilde{f}(x),\ x \in (0,1), \quad \tilde{u}(0) = \tilde{u}(1) = 0,$$

transformieren, wobei

$$\tilde{c}(x) := \frac{1}{4}\frac{b}{\varepsilon}^2(x) - \frac{1}{2}b'(x) + c(x), \quad \tilde{f}(x) := f(x)\exp\left(-\frac{1}{2\varepsilon}\int_0^x b(\xi)\,d\xi\right),$$

sind. *Übungsaufgabe* □

Definition 1.9 Reduziertes Problem, reduzierte Lösung. Das reduzierte Problem erhält man, indem man formal $\varepsilon = 0$ setzt

$$L_0 u_0 := b(x)u_0' + c(x)u_0 = f(x), \quad \text{für } x \in (0,1).$$

Die Randbedingung muss immer dort gesetzt werden, wo die Konvektion herkommt. Im Fall $b(x) > 0$ für alle $x \in [0,1]$, besitzt dieses Problem also die Randbedingung

$$u_0(0) = 0,$$

im Fall $b(x) < 0$ für alle $x \in [0,1]$ die Randbedingung

$$u_0(1) = 0.$$

Die Lösung des reduzierten Problems wird reduzierte Lösung genannt. □

Beispiel 1.10 Das reduzierte Problem zu Beispiel 1.7 lautet

$$u_0' = 1 \quad \text{auf } (0,1), \quad u_0(0) = 0.$$

Seine Lösung ist $u_0(x) = x$.

Damit setzt sich die Lösung des nicht reduzierten Problems aus Beispiel 1.7 aus der Lösung des reduzierten Problems zusammen und einem Anteil, der dafür sorgt, dass die zweite Randbedingung erfüllt ist. □

1.2 Lösungsverhalten

Bemerkung 1.11 Für die Untersuchung der Lösbarkeit des Randwertproblems (1.3), (1.4) spielt die Größe von $\varepsilon > 0$ keine Rolle. Nach Division durch ε und Umbenennung der Daten betrachtet man das Problem

$$Lu := -u'' + b(x)u' + c(x)u = f(x), \quad \text{für } x \in (0,1), \tag{1.5}$$

mit den Randbedingungen

$$u(0) = u(1) = 0. \tag{1.6}$$

□

Definition 1.12 Klassische Lösung. Eine Funktion $u(x)$ wird klassische Lösung von (1.5), (1.6) genannt, falls

- $u \in C^2(0,1) \cap C([0,1])$,
- $u(x)$ erfüllt die Gleichung (1.5) identisch,
- $u(x)$ genügt den Randbedingungen (1.6).

□

Wir betrachten zuerst nur die Differentialgleichung (1.5). Eine klassische Lösung von (1.5) muss die ersten beiden Eigenschaften der obigen Definition besitzen.

Satz 1.13 Superpositionsprinzip. *Betrachte die homogene, lineare Differentialgleichung*

$$-u'' + b(x)u' + c(x)u = 0, \quad x \in (0,1),$$

mit Koeffizienten $b, c \in C([0,1])$. Dann gibt es zwei linear unabhängige Lösungen in $C^2([0,1])$ und jede klassische Lösung ist als Linearkombination dieser darstellbar.

Beweis: Dies wurde in der Vorlesung Theorie und Numerik Gewöhnlicher Differentialgleichungen bewiesen. ■

Satz 1.14 *Betrachte die inhomogene, lineare Differentialgleichung*

$$-u'' + b(x)u' + c(x)u = f(x), \quad x \in (0,1),$$

mit $b, c, f \in C([0,1])$. Dann gibt es eine klassische Lösung $u_p(x)$, die so genannte partikuläre Lösung, und jede klassische Lösung ist darstellbar als

$$u(x) = c_1 u_1(x) + c_2 u_2(x) + u_p(x), \quad c_1, c_2 \in \mathbb{R},$$

wobei $(u_1(x), u_2(x))$ ein System von zwei linear unabhängigen Lösungen (Fundamentalsystem) der zugehörigen homogenen Gleichung ist. Es gilt $u \in C^2([0,1])$.

Beweis: Mit dem globalen Existenz– und Eindeutigkeitssatz von Picard[5]–Lindelöf[6], Übungsaufgabe. ■

Nun wird das Randwertproblem (1.5), (1.6) betrachtet.

Beispiel 1.15 Nichteindeutigkeit der Lösung eines Dirichlet–Randwertproblems. Betrachte die Differentialgleichung

$$-u''(x) - u(x) = 0.$$

Die allgemeine Lösung dieser homogenen, linearen Differentialgleichung lautet

$$u(x) = c_1 \cos x + c_2 \sin x, \quad c_1, c_2 \in \mathbb{R}.$$

[5] Emile Picard (1856 – 1941)
[6] Ernst Lindelöf (1870 – 1946)

- Seien die Randbedingungen

$$u(0) = u(\pi/2) = 1$$

gegeben, dann lautet die eindeutig bestimmte Lösung $u(x) = \cos x + \sin x$.
- Sind die Randbedingungen

$$u(0) = u(\pi) = 1$$

gegeben, dann besitzt das Randwertproblem keine Lösung, da sowohl $c_1 = 1$ als auch $c_1 = -1$ gelten müssten.
- Seien die Randbedingungen

$$u(0) = 1, \quad u(\pi) = -1$$

vorgelegt, dann gibt es unendlich viele Lösungen, denn es folgt aus den Randbedingungen lediglich $c_1 = 1$. Der Wert c_2 kann beliebig gewählt werden.

Dieses Beispiel zeigt, dass selbst in einfachen Fällen keine eindeutige Lösung des Randwertproblems (1.5), (1.6) existieren muss. Es wird sich zeigen, dass die Koeffizientenfunktionen bestimmte Bedingungen erfüllen müssen, damit diese Eigenschaft gegeben ist. □

Satz 1.16 Existenz und Eindeutigkeit der Lösung des Modellproblems mit homogener rechter Seite. *Gegeben sei das Randwertproblem (1.5), (1.6) mit $b \in C^1([0,1])$, $c \in C([0,1])$ und $f(x) \equiv 0$. Gilt für alle $x \in (0,1)$*

$$\tilde{c}(x) := \frac{1}{4}b^2(x) - \frac{1}{2}b'(x) + c(x) \geq 0, \tag{1.7}$$

so besitzt das Problem (1.5), (1.6) nur die triviale Lösung.

Beweis: Zunächst ist offensichtlich, dass $u(x) \equiv 0$ eine Lösung des gestellten Problems ist.

Angenommen, $u(x) \not\equiv 0$ sei eine weitere klassische Lösung. Nach Satz 1.14 gilt $u \in C^2([0,1])$. Mit der Transformation von Bemerkung 1.8 erhält man das Problem

$$-\tilde{u}''(x) + \tilde{c}(x)\tilde{u}(x) = 0, \quad x \in (a,b),\ \tilde{u}(0) = \tilde{u}(1) = 0.$$

Eine Lösung dieses Problems ist $\tilde{u}(x) \equiv 0$. Sei $\tilde{u}(x)$ eine weitere Lösung. Multipliziere nun die Gleichung mit dieser Lösung und integriere partiell. Das führt auf

$$\begin{aligned} 0 &= \int_0^1 \left(-\tilde{u}''(x)\tilde{u}(x) + \tilde{c}(x)\tilde{u}^2(x)\right)\, dx \\ &= -\tilde{u}''(1)\tilde{u}(1) + \tilde{u}''(0)\tilde{u}(0) + \int_0^1 \left(\left(\tilde{u}'(x)\right)^2 + \tilde{c}(x)\tilde{u}^2(x)\right)\, dx \\ &= \int_0^1 \left(\left(\tilde{u}'(x)\right)^2 + \tilde{c}(x)\tilde{u}^2(x)\right)\, dx, \end{aligned}$$

da $\tilde{u}(x)$ an den Randpunkten verschwindet. Wegen $\tilde{c}(x) \geq 0$ ist der Integrand nichtnegativ, also muss er verschwinden. Daraus folgt insbesondere $\left(\tilde{u}'(x)\right)^2 = 0$, also $\tilde{u}'(x) = 0$, also ist $\tilde{u}(x)$ konstant. Wegen der Stetigkeit von $\tilde{u}(x)$ und wegen der Randbedingungen folgt $\tilde{u}(x) \equiv 0$. Daraus ergibt sich aber auch

$$u(x) = \tilde{u}(x) \exp\left(\frac{1}{2}\int_0^x b(\xi)\, d\xi\right) \equiv 0,$$

im Widerspruch zur Annahme. ■

Bemerkung 1.17 Konstante Koeffizienten. Im Spezialfall konstanter Koeffizienten reduziert sich die Bedingung (1.7) auf

$$D := \frac{b^2}{4} + c \geq 0.$$

Auch für den Fall $D < 0$ kann man das Lösungsverhalten des Randwertproblems genau beschreiben, Übungsaufgabe. □

Bemerkung 1.18 Anderes Kriterium zur Eindeutigkeit der Lösung des vollhomogenen Randwertproblem. Betrachte das Randwertproblem mit homogener rechter Seite (1.5), (1.6). Seien $u_1(x), u_2(x)$ zwei linear unabhängige Lösungen und bezeichne

$$R := \det \begin{pmatrix} u_1(0) & u_2(0) \\ u_1(1) & u_2(1) \end{pmatrix}.$$

Die allgemeine Lösung der homogenen Differentialgleichung lautet

$$u(x) = c_1 u_1(x) + c_2 u_2(x).$$

Die Koeffizienten bestimmen sich aus den Randbedingungen

$$0 = c_1 u_1(0) + c_2 u_2(0), \quad 0 = c_1 u_1(1) + c_2 u_2(1),$$

was zur Lösung des linearen Gleichungssystems

$$\begin{pmatrix} u_1(0) & u_2(0) \\ u_1(1) & u_2(1) \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

äquivalent ist. Diese Lösung ist genau dann eindeutig ($c_1 = c_2 = 0$), falls $R \neq 0$ gilt. Genau in diesem Fall besitzt das vollhomogene Randwertproblem nur die triviale Lösung. □

Bemerkung 1.19 Zum inhomogenen Randwertproblem. Betrachte nun das Randwertproblem mit inhomogener rechter Seite (1.5), (1.6). Seien $u_1(x), u_2(x)$ zwei linear unabhängige Lösungen der zugehörigen homogenen Differentialgleichung. Weiter bezeichne

$$W(x) := \det \begin{pmatrix} u_1(x) & u_2(x) \\ u_1'(x) & u_2'(x) \end{pmatrix}$$

die Wronski[7]–Determinante. Wegen der linearen Unabhängigkeit gilt $W(x) \neq 0$ für alle $x \in [0, 1]$, siehe Vorlesung Theorie und Numerik gewöhnlicher Differentialgleichungen. Weiter seien

$$A(x) := \det \begin{pmatrix} u_1(0) & u_2(0) \\ u_1(x) & u_2(x) \end{pmatrix}, \quad B(x) := \det \begin{pmatrix} u_1(x) & u_2(x) \\ u_1(1) & u_2(1) \end{pmatrix}.$$

Für die Betrachtung des Randwertproblems mit inhomogener rechter Seite, wird der Begriff der Greenschen Funktion wiederholt. □

Definition 1.20 Green[8]sche Funktion. Die Funktion $\Gamma(x, \xi)$ heißt Greensche Funktion für das homogene Randwertproblem $Lu = 0$, $u(0) = u(1) = 0$, wenn:

1. $\Gamma(x, \xi)$ ist stetig auf dem Quadrat $Q := \{(x, \xi) \ : \ x, \xi \in [0, 1]\}$.

[7] Joseph Marie Wronski (1758 – 1853)
[8] Georg Green (1793 – 1841)

2. In jedem der Dreiecke

$$Q_1 := \{(x,\xi) \ : \ 0 \le \xi \le x \le 1\}, \quad Q_2 := \{(x,\xi) \ : \ 0 \le x \le \xi \le 1\}$$

existieren stetige partielle Ableitungen $\Gamma_x(x,\xi)$ und $\Gamma_{xx}(x,\xi)$.
3. Bei festem $\xi \in I = (0,1)$ ist $\Gamma(x,\xi)$ als Funktion von x eine Lösung von $L\Gamma = 0$ für $x \neq \xi$, $x \in I$.
4. Auf der Diagonalen $x = \xi$ besitzt die erste Ableitung eine Sprung der Form

$$\Gamma_x(x+0,x) - \Gamma_x(x-0,x) = \frac{1}{p(x)}, \quad 0 < x < 1,$$

mit

$$p(x) = \exp\left(\frac{1}{\varepsilon}\int_0^x b(s)\ ds\right).$$

5. $\Gamma(0,\xi) = \Gamma(1,\xi) = 0$ für alle $\xi \in (0,1)$.

□

Satz 1.21 Existenz und Eindeutigkeit der Lösung des Modellproblems mit inhomogener rechter Seite. *Betrachte das Modellproblem (1.5), (1.6) mit $b, c, f \in C([0,1])$. Besitzt das zugehörige Randwertproblem für die Gleichung mit homogener rechter Seite nur die triviale Lösung, so besitzt das Randwertproblem (1.5), (1.6) genau eine klassische Lösung. Diese hat die Gestalt*

$$u(x) = \int_0^1 \Gamma(x,\xi) f(\xi)\ d\xi$$

mit der Greenschen Funktion

$$\Gamma(x,\xi) = \frac{1}{R\ W(\xi)} \begin{cases} A(\xi)B(x) & \text{für } 0 \le \xi \le x \le 1, \\ A(x)B(\xi) & \text{für } 0 \le x \le \xi \le 1. \end{cases}$$

Beweis: Dass $\Gamma(x,\xi)$ eine Greensche Funktion ist, rechnet man direkt. Die Existenz einer Lösung zeigt man, indem man nachrechnet, dass $u(x)$ Lösung des Randwertproblems (1.5), (1.6) ist. Die Eindeutigkeit folgt schließlich analog zu Bemerkung 1.18, indem man zeigt, dass sich die freien Parameter der allgemeinen Lösung der homogenen Gleichung unter den gegebenen Voraussetzungen eindeutig bestimmen lassen. ■

Dass das vollhomogene Problem nur die triviale Lösung besitzt, ist zum Beispiel gegeben, wenn (1.7) erfüllt ist. Die Umkehrung des vorstehenden Satzes gilt auch.

Satz 1.22 *Besitzt das inhomogene Randwertproblem (1.5), (1.6) genau eine klassische Lösung, so besitzt das zugehörige Randwertproblem mit homogener rechter Seite nur die triviale Lösung.*

Beweis: Sei $u(x)$ die eindeutige klassischen Lösung des inhomogenen Randwertproblems. Sei $u_h(x)$ eine nichttrivale Lösung des vollhomogenen Randwertproblems, dann folgt auf Grund der Linearität des Problems, dass dann $u(x) + u_h(x)$ eine klassischen Lösung des inhomogenen Randwertproblems ist, im Widerspruch zur vorausgesetzten Einzigkeit dieser Lösung. ■

Folgerung 1.23 Existenz und Eindeutigkeit der Lösung des Modellproblems mit beliebigen Dirichlet–Randdaten. *Betrachte das Modellproblem (1.5) mit $b \in C^1([0,1])$, $c, f \in C([0,1])$ und mit den Dirichlet–Randdaten $u(0) = a$, $u(1) = b$ mit $a, b \in \mathbb{R}$. Gilt für alle $x \in (0,1)$ die Beziehung (1.7), dann existiert genau eine klassische Lösung.*

Beweis: Inhomogene Dirichlet–Randdaten können in die rechte Seite transformiert werden, siehe Bemerkung 1.4. Diese Transformation ist zweimal stetig differenzierbar und sie kann so erfolgen, dass die neue rechte Seite stetig in $[0,1]$ ist. Für das so erhaltene Problem mit homogenen Dirichlet–Randbedingungen kann man die obigen Aussagen anwenden. Nach Satz 1.16 besitzt das vollhomogene Problem nur die triviale Lösung. Aus Satz 1.21 folgt, dass es genau eine klassische Lösung gibt. Da die Rücktransformation zweimal stetig differenzierbar ist, existiert damit genau eine klassische Lösung für das Problem mit inhomogenen Dirichlet–Randbedingungen. ■

Bemerkung 1.24 Es gibt auch andere hinreichende Bedingungen als (1.7) dafür, dass das vollhomogene Problem nur die triviale Lösung besitzt, siehe Folgerung 1.30. □

1.3 Maximumprinzip und Stabilität

Im Folgenden sei L der durch

$$(Lu)(x) := -u''(x) + b(x)u'(x) + c(x)u(x), \quad x \in (0,1)$$

definierte lineare Differentialoperator, der für $b, c \in C([0,1])$ offenbar $C^2(0,1)$ in $C(0,1)$ abbildet.

Lemma 1.25 *Seien $b \in C([0,1])$ und $c(x) = 0$ für alle $x \in [0,1]$. Dann gilt für jedes $u \in C^2(0,1) \cap C([0,1])$*

i) aus $(Lu)(x) \leq 0$ für $x \in (0,1)$ folgt $u(x) \leq \max\{u(0), u(1)\}$ für $x \in [0,1]$,
ii) aus $(Lu)(x) \geq 0$ für $x \in (0,1)$ folgt $u(x) \geq \min\{u(0), u(1)\}$ für $x \in [0,1]$.

Beweis: Es braucht nur i) gezeigt zu werden. Die Aussage ii) ergibt sich dann, wenn $u(x)$ durch $-u(x)$ ersetzt wird.

Es wird zuerst gezeigt, dass aus der schärferen Voraussetzung $(Lu)(x) < 0$ auf $(0,1)$ die Behauptung folgt. Angenommen, die Funktion $u(x)$ nimmt ihr Maximum nicht am Rand, sondern im Inneren des Intervalls an. Dann gibt es ein $x_0 \in (0,1)$, mit $u'(x_0) = 0$ (lokales Extremum) und $u''(x_0) \leq 0$ (lokales Maximum). Es folgt

$$-u''(x_0) + b(x_0)u'(x_0) = -u''(x_0) \geq 0,$$

was im Widerspruch zur Voraussetzung steht.

Nun wird i) gezeigt. Hierzu sei für $\delta, \lambda > 0$

$$w(x) = \delta e^{\lambda x}, \quad x \in [0,1].$$

Ist λ hinreichend groß, $\lambda > \max_{x\in[0,1]} b(x)$, so gilt für alle $x \in (0,1)$

$$(Lw)(x) = -\lambda^2 w(x) + b(x)\lambda w(x) = -\lambda\left(\lambda - b(x)\right) w(x) < 0.$$

Mit der Linearität des Differentialoperators folgt

$$\left(L(u+w)\right)(x) = (Lu)(x) + (Lw)(x) < 0.$$

Nach dem ersten Teil des Beweises gilt

$$u(x) + w(x) \leq \max\{u(0) + w(0) + u(1) + w(1)\}.$$

Für $\delta \to 0$ ergibt sich die Behauptung. ■

Satz 1.26 Maximumprinzip. *Seien $b, c \in C([0,1])$ und $c(x)$ auf $[0,1]$ nichtnegativ. Dann gilt für jedes $u \in C^2(0,1) \cap C([0,1])$*

i) aus $(Lu)(x) \leq 0$ *für alle* $x \in (0,1)$ *folgt* $u(x) \leq \max\{0, u(0), u(1)\}$ *für* $x \in [0,1]$,

ii) aus $(Lu)(x) \geq 0$ *für alle* $x \in (0,1)$ *folgt* $u(x) \geq \min\{0, u(0), u(1)\}$ *für* $x \in [0,1]$.

Beweis: Wiederum ergibt sich die zweite Aussage aus der ersten, wenn man dort $u(x)$ durch $-u(x)$ ersetzt.

Da $u(x)$ in $[0,1]$ stetig ist, ist die Menge

$$\mathcal{M}^+ := \{x \in (0,1) \ : \ u(x) > 0\}$$

entweder leer oder die Vereinigung offener Teilintervalle von $(0,1)$, Analyis I. Sei $\mathcal{M}^+ = \emptyset$, sei also $u(x)$ in $(0,1)$ nichtpositiv. Dann ist die die Behauptung trivialerweise erfüllt.

Sei $\mathcal{M}^+ = (0,1)$. Dann gilt für $x \in (0,1)$

$$-u''(x) + b(x)u'(x) \leq -u''(x) + b(x)u'(x) + c(x)u(x) = (Lu)(x) \leq 0.$$

Nach Lemma 1.25 folgt

$$u(x) \leq \max\{u(0), u(1)\},$$

was die Behauptung auch in diesem Falle zeigt.

Sei nun $\emptyset \neq \mathcal{M}^+ \neq (0,1)$. Es wird gezeigt, dass $\mathcal{M}^+$ an 0 oder 1 heranreichen muss. Sei $(a_0, b_0) \subseteq \mathcal{M}^+$. Gelten $a_0 \neq 0$ und $u(a_0) > 0$, so folgt, wegen der Stetigkeit von $u(x)$, dass entweder $u(0) > 0$ oder ein $0 \leq a_1 < a_0$ existiert mit $u(a_1) = 0$. Analoges gilt für b_0. Man kann daher $a_0 = 0$ oder $u(a_0) = 0$ sowie $b_0 = 1$ oder $u(b_0) = 0$ annehmen. Nach der Voraussetzung gilt für alle $x \in (a_0, b_0)$

$$(Lu)(x) \leq 0 \quad \Longrightarrow \quad -u''(x) + b(x)u'(x) \leq -c(x)u(x) \leq 0.$$

Damit kann man wieder Lemma 1.25 anwenden. Es folgt also für alle $x \in (a_0, b_0)$

$$0 < u(x) \leq \max\{u(a_0), u(b_0)\}. \tag{1.8}$$

Offenbar kann nicht zugleich $u(a_0) = u(b_0) = 0$ gelten, denn dies würde dieser Relation widersprechen. Der Fall $a_0 = 0, b_0 = 1$ wurde bereits betrachtet. Es bleiben die Fälle $a_0 = 0$ und $u(b_0) = 0$ sowie $u(a_0) = 0$ und $b_0 = 1$.

Damit ist gezeigt: Ist die Menge $\mathcal{M}^+$ nicht leer, so gibt es Zahlen $\hat{a}, \hat{b} \in [0,1]$ mit $\hat{a} \leq \hat{b}$, so dass

$$\mathcal{M}^+ = (0, \hat{a}) \cup (\hat{b}, 1),$$

wobei $u(\hat{a}) = 0$ wenn $\hat{a} \neq 0$, und $u(\hat{b}) = 0$, wenn $\hat{b} \neq 1$. Mit (1.8) gilt für $x \in (0,1)$

$$\begin{aligned} u(x) &\leq \max\left\{ \max_{x \in (0,\hat{a})} u(x), \max_{x \in (\hat{b},1)} u(x), 0 \right\} \\ &\leq \max\Big\{ \max\{u(0), u(\hat{a})\}, \max\{u(\hat{b}), u(1)\}, 0 \Big\} \\ &= \max\{0, u(0), u(1)\}. \end{aligned}$$

■

Folgerung 1.27 Inverse Monotonie, Isotonie, Vergleichsprinzip. *Unter den Voraussetzungen von Satz 1.26 folgt für zwei Funktionen* $u, v \in C^2(0,1) \cap C([0,1])$ *mit* $u(0) \leq v(0)$ *und* $u(1) \leq v(1)$ *aus* $(Lu)(x) \leq (Lv)(x)$ *für* $x \in (0,1)$, *dass* $u(x) \leq v(x)$ *für* $x \in [0,1]$.

Beweis: Satz 1.26 ist auf die Differenz $(u - v)(x)$ anzuwenden. ■

Mit Hilfe des Maximumprinzips kann man nun eine Stabilitätsabschätzung für Lösungen des inhomogenen Randwertproblems beweisen, welche die stetige Abhängigkeit der Lösung von den Daten zeigt.

Satz 1.28 Stabilität der Lösung und stetige Abhängigkeit von den Daten. *Vorgelegt sei das Randwertproblem (1.5), (1.6), mit $b, c, f \in C([0,1])$. Ist $c(x)$ auf $[0,1]$ nichtnegativ, so gilt für jede klassische Lösung $u(x)$ die Abschätzung*

$$\|u\|_{C([0,1])} \leq \Lambda \|f\|_{C([0,1])},$$

wobei die Konstante $\Lambda > 0$ von $b(x), c(x)$ abhängt, aber nicht von $f(x)$.

Beweis: Für ein noch zu spezifizierendes $\lambda > 0$ setzt man

$$w(x) := Be^{\lambda x} - A, \quad x \in (0,1),$$

mit

$$A := \Lambda B, \quad B := \|f\|_{C([0,1])}, \quad \Lambda := e^{\lambda} - 1 > 0.$$

Dann gilt für $x \in (0,1)$

$$\begin{aligned} (Lw)(x) &= -\left(\lambda^2 - \lambda b(x) - c(x)\right) Be^{\lambda x} - Ac(x) \\ &\leq -\left(\lambda^2 - \lambda b(x) - c(x)\right) Be^{\lambda x}. \end{aligned}$$

Nun wählt man λ derart, dass $\left(\lambda^2 - \lambda b(x) - c(x)\right) e^{\lambda x} \geq 1$. Diese Relation ist erfüllt, wenn λ groß genug ist, etwa

$$\lambda \geq \max_{x \in [0,1]} \left(\frac{b(x)}{2} + \sqrt{\frac{b^2(x)}{4} + c(x) + 1} \right),$$

da $e^{\lambda x} \geq 1$. Dann gilt für alle $x \in (0,1)$

$$(Lw)(x) \leq -B = -\|f\|_{C([0,1])}.$$

Es folgt für alle $x \in (0,1)$ mit Hilfe der Normdefinition im Raum der stetigen Funktionen

$$(L(\pm u + w))(x) = \pm f(x) + (Lw)(x) \leq |f(x)| - \|f\|_{C([0,1])} \leq 0.$$

Nach dem Maximumprinzip gilt

$$\pm u(x) + w(x) \leq \max\{0, \pm u(0) + w(0), \pm u(1) + w(1)\} = \max\{0, w(0), w(1)\}.$$

Damit folgt für alle $x \in (0,1)$

$$\pm u(x) \leq \max\{0, w(0), w(1)\} - w(x).$$

Aus $e^{\lambda x} \geq 1$ folgt

$$w(x) \geq B - A = w(0), \quad w(1) = Be^{\lambda} - A,$$

also

$$\begin{aligned} |u(x)| &\leq \max\{0, w(0), w(1)\} - w(x) \leq \max\{0, B - A, Be^{\lambda} - A\} + A - B \\ &= \max\{A - B, 0, B(e^{\lambda} - 1)\} = \max\{A - B, 0, \Lambda B\} = \max\{A - B, 0, A\} = A, \end{aligned}$$

was die Behauptung war. ■

Bemerkung 1.29 Nichtnormiertes Problem. Für das nichtnormierte Problem (1.1), (1.2) mit Dirichlet–Randbedingungen $u(d) = \alpha$, $u(e) = \beta$ erhält man analog

$$\|u\|_{C([e,d])} \leq \Lambda \|f\|_{C([e,d])} + \max\{|\alpha|, |\beta|\},$$

wobei Λ jetzt auch von $e - d$ abhängen kann, aber nicht von α, β abhängt, [Emm04, Satz 2.5.4], Übungsaufgabe.

Dass diese Abschätzung tatsächlich eine Stabilitätsabschätzung ist, sieht man ein, wenn man sie auf die Differenz $u(x) - \tilde{u}(x)$ anwendet. Dabei sei $u(x)$ Lösung des exakten Problems und $\tilde{u}(x)$ Lösung eines Problems mit gestörter rechter Seite $\tilde{f}$ oder gestörten Randbedingungen $\tilde{\alpha}, \tilde{\beta}$. Aus der Linearität des Problems folgt sofort

$$\|u - \tilde{u}\|_{C([e,d])} \leq \Lambda \left\| f - \tilde{f} \right\|_{C([e,d])} + \max\left\{ |\alpha - \tilde{\alpha}|, \left|\beta - \tilde{\beta}\right| \right\}.$$

Das bedeutet, kleine Störungen in den Daten führen auch nur zu kleinen Störungen in der Lösung. □

Folgerung 1.30 Eindeutigkeit der Lösung des homogenen Problems. *Gegeben sei das Randwertproblem (1.5), (1.6), mit $b, c \in C([0,1])$ und $f(x) \equiv 0$. Ist $c(x)$ auf $[0,1]$ nichtnegativ, so besitzt das Problem nur die triviale Lösung $u(x) \equiv 0$.*

Beweis: Das folgt unmittelbar aus der Abschätzung von Satz 1.28 ■

Bemerkung 1.31 Diese Aussage folgt auch schon aus dem Maximumprinzip, Satz 1.26, weil für ein homogenes Problem beide Teile i) und ii) dieses Satzes gelten. □

Folgerung 1.32 *Vorgelegt sei das Randwertproblem (1.5), (1.6) mit $b, c, f \in C([0,1])$. Ist $c(x)$ auf $[0,1]$ nichtnegativ, so besitzt das Randwertproblem genau eine klassischen Lösung.*

Beweis: Das folgt unmittelbar aus Folgerung 1.30 und Satz 1.21. ■

Jetzt wird noch das starke Maximumprinzip eingeführt.

Lemma 1.33 *Seien $b, c \in C([0,1])$, sei $c(x)$ auf $[0,1]$ nichtnegativ und gelte $u \in C^2(0,1) \cap C([0,1])$. Gilt $(Lu)(x) < 0$ für alle $x \in (0,1)$, so kann $u(x)$ kein nichtnegatives Maximum im Innern des Intervalls annehmen.*

Beweis: Übungsaufgabe. Angenommen, es gäbe ein nichtnegatives inneres Maximum $x_0 \in (0,1)$. Dann sind $u(x_0) \geq 0$, $u'(x_0) = 0$ und $u''(x_0) \leq 0$. Damit folgt

$$(Lu)(x_0) = -u''(x_0) + b(x_0)u'(x_0) + d(x_0)u(x_0) \geq 0,$$

im Widerspruch zur Voraussetzung. ■

Satz 1.34 Starkes Maximumprinzip. *Seien $b, c \in C([0,1])$ und sei $c(x)$ auf $[0,1]$ nichtnegativ. Nimmt $u \in C^2(0,1) \cap C([0,1])$ im Inneren des Intervalls ein nichtnegatives Maximum an und gilt $(Lu)(x) \leq 0$ für alle $x \in (0,1)$, so ist $u(x)$ konstant.*

Beweis: Die Funktion $u(x)$ nehme in $x_0 \in (0,1)$ das größte nichtnegative Maximum an, also gilt insbesondere $u(x_0) \geq 0$.

Angenommen, $u(x)$ sei nicht konstant. Dann gibt es ein $x_1 \in (0,1)$ mit $u(x_1) < u(x_0)$. Ohne Beschränkung der Allgemeinheit sei $x_1 > x_0$, der Fall $x_1 < x_0$ kann analog behandelt werden.

Es sei für $\delta, \lambda > 0$

$$w(x) := \delta \left(e^{\lambda(x-x_0)} - 1 \right), \quad x \in [0, x_1].$$

Offenbar gelten

$$w(x) \begin{cases} < 0 & \text{für } x < x_0, \\ = 0 & \text{für } x = x_0, \\ > 0 & \text{für } x > x_0. \end{cases}$$

Nun wählt man λ hinreichend groß, etwa

$$\lambda \geq \max_{x \in [0,x_1]} \left(\frac{b(x)}{2} + \sqrt{\frac{b^2(x)}{4} + c(x)} \right),$$

so dass für alle $x \in (0, x_1)$ gilt

$$(Lw)(x) = -\left(\lambda^2 - \lambda b(x) - c(x) \right) \delta e^{\lambda(x-x_0)} - c(x)\delta < 0.$$

Nun wird δ so klein gewählt, dass

$$u(x_1) + w(x_1) < u(x_0).$$

Dann gilt nach Voraussetzung auch

$$(L(u+w))(x) = (Lu)(x) + (Lw)(x) < 0, \quad x \in (0, x_1).$$

Wegen

$$\begin{aligned} u(x) + w(x) &< u(x_0), \quad \text{für } x \in (0, x_0), \\ u(x_0) + w(x_0) &= u(x_0), \\ u(x_1) + w(x_1) &< u(x_0), \end{aligned}$$

nimmt die Funktion $(u+w)(x)$ in $(0, x_1)$ ein nichtnegatives Maximum an. Das steht nach Lemma 1.33 im Widerspruch zu $(L(u+w))(x) < 0$. Demzufolge ist die Annahme, dass $u(x)$ nicht konstant ist, falsch. ■

Bemerkung 1.35 Durch Ersetzung von $u(x)$ mit $-u(x)$ erhält man aus den Maximumprinzipien entsprechende Minimumprinzipien. □

Kapitel 2

Finite–Differenzen–Verfahren

2.1 Finite Differenzen

Bemerkung 2.1 Idee. Die grundlegende Idee von Finite–Differenzen–Verfahren besteht darin, dass man die Ableitungen in der Differentialgleichung durch geeignete finite Differenzen ersetzt. Dazu wird das Intervall $[0, 1]$ mittels eines äquidistanten Gitters zerlegt:

$$\begin{aligned} x_i &= ih, \quad i = 0, \ldots, N, \quad h = 1/N, \\ \omega_h &= \{x_i \ : \ i = 0, \ldots, N\} \ - \text{Gitter.} \end{aligned}$$

□

Definition 2.2 Gitterfunktion. Ein Vektor $\mathbf{u}_h = (u_0, \ldots, u_N)^T \in \mathbb{R}^{N+1}$, der jedem Gitterpunkt einen Funktionswert zuordnet, heißt Gitterfunktion. Die Restriktion einer Funktion $u \in C([0, 1])$ auf eine Gitterfunktion wird mit $R_h u$ bezeichnet, das heißt

$$R_h u := (u(x_0), u(x_1), \ldots, u(x_N))^T .$$

□

Beispiel 2.3 Sei ein Gitter mit den Punkten $\{0, 0.25, 0.5, 0.75, 1\}$ gegeben. Dann ist die Gitterfunktion zu $u(x) = x^2$

$$R_h u = \left(0, \frac{1}{16}, \frac{1}{4}, \frac{9}{16}, 1\right)^T .$$

Unterschiedliche Funktionen können für ein gegebenes Gitter die gleiche Gitterfunktion haben. Betrachte beispielsweise $u(x) = \sin(4\pi x)$ auf dem obigen Gitter. Die zugehörige Gitterfunktion ist

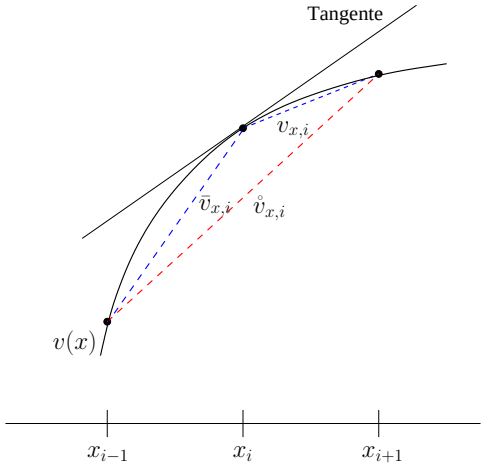
$$R_h u = (0, 0, 0, 0, 0)^T .$$

Dies ist offensichtlich auch die Gitterfunktion von $u(x) = 0$. Das obige Gitter ist zu grob, um die Funktion $u(x) = \sin(4\pi x)$ vernünftig auflösen zu können. □

Definition 2.4 Differenzenoperatoren. Sei $v(x)$ eine genügend glatte Funktion. Bezeichne $v_i = v(x_i)$, wobei x_i Knoten eines Gitters ist. Die folgenden Differenzen-

quotienten (finite Differenzen) nennt man

$$\begin{aligned}
D^+v(x_i) &= v_{x,i} &&= \frac{v_{i+1}-v_i}{h} && \text{– Vorwärtsdifferenz,}\\
D^-v(x_i) &= v_{\overline{x},i} &&= \frac{v_i-v_{i-1}}{h} && \text{– Rückwärtsdifferenz,}\\
D^0v(x_i) &= v_{\mathring{x},i} &&= \frac{v_{i+1}-v_{i-1}}{2h} && \text{– zentrale Differenz,}\\
D^+D^-(v)(x_i) &= v_{\overline{x}x,i} &&= \frac{v_{i+1}-2v_i+v_{i-1}}{h^2} && \text{– zweite Differenz.}
\end{aligned}$$



□

Bemerkung 2.5 Die Formel für $D^+D^-(v)(x_i)$ kontrolliert man durch direktes Nachrechnen. Weiter gilt

$$D^0v(x_i) = \frac{1}{2}(D^+v(x_i) + D^-v(x_i)).$$

□

Definition 2.6 Konsistenz eines Differenzenoperators, diskrete Maximumsnorm. Sei L ein Differentialoperator. Der Differenzenoperator $L_h : \mathbb{R}^{N+1} \to \mathbb{R}^{N+1}$ heißt mit L konsistent mit der Ordnung k, wenn

$$\max_{0\le i\le N} |(Lu)(x_i) - (L_hu_h)_i| =: \|(Lu)(x_i) - (L_hu_h)_i\|_{\infty,d} = \mathcal{O}(h^k)$$

gilt. Hierbei ist $\|\cdot\|_{\infty,d}$ die diskrete Maximumsnorm im Raum der Gitterfunktionen.
□

Die Konsistenz ist ein Maß für die Approximationsgüte von L_h.

Beispiel 2.7 Aus der Taylor[1]–Entwicklung für $v(x)$ an der Stelle x_i ergibt sich

$$\begin{aligned}
D^+v(x_i) &= v'(x_i) + \mathcal{O}(h),\\
D^-v(x_i) &= v'(x_i) + \mathcal{O}(h),\\
D^0v(x_i) &= v'(x_i) + \mathcal{O}(h^2),\\
D^+D^-(v)(x_i) &= v''(x_i) + \mathcal{O}(h^2).
\end{aligned}$$

Die Differenzenoperatoren $D^+v(x_i), D^-v(x_i), D^0v(x_i)$ sind damit konsistent zu $L = \frac{d}{dx}$ mit der Ordnung 1,1 beziehungsweise 2. Der Operator $D^+D^-(v)(x_i)$ ist von zweiter Ordnung konsistent mit $L = \frac{d^2}{dx^2}$. □

[1]Brook Taylor (1685 – 1731)

Beispiel 2.8 Betrachtet wird der Differentialoperator

$$Lu = \frac{d}{dx}\left(k(x)\frac{du}{dx}\right),$$

wobei $k(x)$ stetig differenzierbar ist. Wir definieren den Differenzenoperator L_h wie folgt

$$\begin{aligned}(L_h u_h)_i &= D^+(aD^-u(x_i)) = \frac{1}{h}\Big(a(x_{i+1})D^-u(x_{i+1}) - a(x_i)D^-u(x_i)\Big)\\ &= \frac{1}{h}\left(a_{i+1}\frac{u_{i+1}-u_i}{h} - a_i\frac{u_i-u_{i-1}}{h}\right),\end{aligned}$$

wobei a eine Gitterfunktion ist, die geeignet gewählt werden soll. Es folgt mit Produktregel beziehungsweise mit Taylor–Entwicklung

$$\begin{aligned}(Lu)_i &= k'(x_i)(u')_i + k(x_i)(u'')_i,\\ (L_h u_h)_i &= \frac{a_{i+1}-a_i}{h}(u')_i + \frac{a_{i+1}+a_i}{2}(u'')_i + \frac{h(a_{i+1}-a_i)}{6}(u''')_i + \mathcal{O}(h^2).\end{aligned}$$

Für die Differenz ergibt sich

$$\begin{aligned}(Lu)_i - (L_h u_h)_i &= \left(k'(x_i) - \frac{a_{i+1}-a_i}{h}\right)(u')_i + \left(k(x_i) - \frac{a_{i+1}+a_i}{2}\right)(u'')_i\\ &\quad -\frac{h(a_{i+1}-a_i)}{6}(u''')_i + \mathcal{O}(h^2).\end{aligned}$$

Damit L_h von zweiter Ordnung mit L konsistent ist, müssen somit gelten

$$\frac{a_{i+1}-a_i}{h} = k'(x_i) + \mathcal{O}(h^2), \quad \frac{a_{i+1}+a_i}{2} = k(x_i) + \mathcal{O}(h^2).$$

Aus der ersten Forderung folgt $a_{i+1} - a_i = \mathcal{O}(h)$, womit außerdem folgt, dass der dritte Summand in der Fehlergleichung von Ordnung $\mathcal{O}(h^2)$ wird. Mögliche Varianten sind (*Übungsaufgaben?*)

$$a_i = \frac{k_i + k_{i-1}}{2}, \quad a_i = k\left(x_i - \frac{h}{2}\right), \quad a_i = (k_i k_{i-1})^{1/2}.$$

Man beachte, die ,,natürliche" Wahl $a_i = k_i$ garantiert nur Konsistenz von erster Ordnung, siehe die Taylorentwicklung für $D^+v(x_i)$. □

2.2 Klassische Konvergenztheorie für zentrale Differenzen

Bemerkung 2.9 In diesem Abschnitt wird das 2–Punkt–Randwertproblem

$$Lu := -u'' + b(x)u' + c(x)u = f(x), \quad \text{für } x \in (0,1), \quad u(0) = u(1) = 0, \tag{2.1}$$

betrachtet, das heißt $\varepsilon = 1$, um die klassische Lösungstheorie darzustellen. Es wird angenommen, dass die Parameterfunktionen b, c, f hinreichend glatt sind und dass $c(x) \geq 0$ für alle $x \in [0,1]$ gilt. □

Definition 2.10 Zentrales Differenzenschema. Das zentrale Differenzenschema für (2.1) besitzt die Gestalt

$$\begin{aligned}-D^+D^-u_i + b_i D^0 u_i + c_i u_i &= f_i, \quad \text{für } i = 1, \ldots, N-1,\\ u_0 = u_N &= 0.\end{aligned} \tag{2.2}$$

□

Bemerkung 2.11

- Das zentrale Differenzenschema führt auf ein tridiagonales System linearer Gleichungen

$$r_i u_{i-1} + s_i u_i + t_i u_{i+1} = f_i, \quad i = 1, \ldots, N-1, \quad u_0 = u_N = 0,$$

mit

$$r_i = -\frac{1}{h^2} - \frac{1}{2h} b_i, \quad s_i = c_i + \frac{2}{h^2}, \quad t_i = -\frac{1}{h^2} + \frac{1}{2h} b_i.$$

- Die folgenden Fragen müssen beantwortet werden:
  - Welche Eigenschaften besitzt das diskrete Problem (2.2)?
  - Was kann man über den Fehler $|u(x_i) - u_i|$ aussagen?

  Dazu werden die Konzepte von Konsistenz und Stabilität verwendet.

□

Definition 2.12 Konsistenz eines Differenzenschemas und Konsistenzordnung. Betrachte ein Differenzenschema der Gestalt $L_h u_h = R_h(Lu)$. Dabei seien die Randbedingungen derart integriert, dass die erste und letzte Zeile von L_h identisch zur ersten und letzten Zeile der Einheitsmatrix sind und $R_h(Lu)_0 = u_0$ sowie $R_h(Lu)_N = u_N$ gelten. Das Schema wir konsistent von der Ordnung k in der diskreten Maximumsnorm genannt, falls

$$\|L_h R_h u - R_h(Lu)\|_{\infty,d} \leq c h^k$$

ist, wobei die positiven Konstanten c und k unabhängig von h sind. □

Lemma 2.13 Konsistenzordnung des zentralen Differenzenschemas. *Unter der Annahme, dass $u \in C^4([0,1])$ gilt, besitzt das zentrale Differenzenschema (2.2) die Konsistenzordnung* 2.

Beweis: Mit Taylor–Entwicklung, Übungsaufgabe. ■

Definition 2.14 Stabilität eines Differenzenschemas. Ein Differenzenschema $L_h u_h = f_h$ wird stabil in der diskreten Maximumsnorm genannt, wenn es eine Stabilitätskonstante c_S unabhängig von h gibt, so dass

$$\|u_h\|_{\infty,d} \leq c_S \|L_h u_h\|_{\infty,d}$$

für alle Gitterfunktionen u_h gilt. □

Definition 2.15 Konvergenz eines Differenzenschemas und Konvergenzordnung. Ein Differenzenschema für (2.1) ist konvergent von Ordnung k in der diskreten Maximumsnorm, falls es positive Konstanten c und k unabhängig von h gibt, so dass

$$\|u_h - R_h u\|_{\infty,d} \leq c h^k.$$

□

Satz 2.16 Konsistenz + Stabilität $\Longrightarrow$ Konvergenz. *Ein konsistentes und stabiles Differenzenschema ist konvergent. Konsistenz– und Konvergenzordnung sind gleich.*

Beweis: Es gilt

$$\begin{aligned}\|u_h - R_h u\|_{\infty,d} &\overset{\text{Stab.}}{\leq} c_S \|L_h (u_h - R_h u)\|_{\infty,d} \overset{\text{lin.}}{=} c_S \|L_h u_h - L_h R_h u\|_{\infty,d} \\ &= c_S \|f_h - L_h R_h u\|_{\infty,d} = c_S \|R_h f - L_h R_h u\|_{\infty,d} \\ &= c_S \|R_h L u - L_h R_h u\|_{\infty,d} \overset{\text{Kons.}}{\leq} K h^k,\end{aligned}$$

wobei die Konstante K das Produkt aus den Konstanten der Stabilitäts– und Konsistenzbedingung ist. ■

Bemerkung 2.17 Man muss also Konsistenz und Stabilität untersuchen.

- Konsistenzuntersuchungen basieren üblicherweise auf Taylor–Entwicklungen und sie laufen in der Regel nach dem gleichen Muster ab.
- Stabilitätsuntersuchungen werden nicht an Funktionen sondern mit Matrizen und Funktionen durchgeführt, siehe Definition 2.14. Sie sind nicht so einfach und es werden einige neue Begriffe benötigt, die im folgenden bereitgestellt werden.

□

Definition 2.18 Natürliche Ordnung von Vektoren und Matrizen, invers–monotone Matrix. Seien $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Dann schreibt man $\mathbf{x} \leq \mathbf{y}$ genau dann, wenn $x_i \leq y_i$ für alle $i = 1, \ldots, n$ gilt. Die Notation $\mathbf{x} \geq \mathbf{1}$ bedeutet $x_i \geq 1$ für alle $i = 1, \ldots, n$. Analog bedeutet für eine Matrix $A = (a_{ij}) \in \mathbb{R}^{n \times n}$ die Bezeichnung $A \geq 0$, dass $a_{ij} \geq 0$ für alle $i, j = 1, \ldots, n$ gilt.

Eine Matrix A, für welche A^{-1} existiert mit $A^{-1} \geq 0$, wird invers–monotone Matrix genannt. □

Das nächste Lemma gibt ein diskretes Analogon zum Vergleichsprinzip von Folgerung 1.27.

Lemma 2.19 Diskretes Vergleichsprinzip. *Sei $A \in \mathbb{R}^{n \times n}$ invers–monoton. Gilt $A\mathbf{v} \leq A\mathbf{w}$ für $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$, dann folgt $\mathbf{v} \leq \mathbf{w}$.*

Beweis: Nach Voraussetzung gilt

$$A(\mathbf{v} - \mathbf{w}) := \mathbf{b} \leq \mathbf{0}.$$

Multiplikation mit A^{-1} ergibt

$$\mathbf{v} - \mathbf{w} = A^{-1}\mathbf{b} \leq \mathbf{0}.$$

Die letzte Ungleichung folgt daraus, dass A invers–monoton ist. Nichtnegative Matrixeinträge werden mit nichtpositiven Vektoreinträgen von $\mathbf{b}$ multipliziert. Man erhält als Ergebnis einen Vektor mit nichtpositiven Komponenten. ■

Eine wichtige Teilklasse der Klasse der invers–monotonen Matrizen ist die folgende.

Definition 2.20 M–Matrix. Eine Matrix $A = (a_{ij}) \in \mathbb{R}^{n \times n}$ wird M–Matrix genannt, falls:

1. $a_{ij} \leq 0$ für $i \neq j$,
2. A^{-1} existiert mit $A^{-1} \geq 0$.

□

Lemma 2.21 *Sei $A = (a_{ij}) \in \mathbb{R}^{n \times n}$ eine M–Matrix. Dann gilt $a_{ii} > 0$, $i = 1, \ldots, n$.*

Beweis: Übungsaufgabe. ■

Die zweite Bedingung der Definition ist im allgemeinen schwer zu überprüfen. Es gibt aber auch handlichere Charakterisierungen von M–Matrizen.

Satz 2.22 M–Matrix–Kriterium. *Sei $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ mit $a_{ij} \leq 0$ für $i \neq j$. Dann ist A eine M–Matrix genau dann, wenn ein Vektor $\mathbf{e} \in \mathbb{R}^n$, $\mathbf{e} > 0$ existiert, so dass $A\mathbf{e} > \mathbf{0}$. Dann gilt für die Zeilensummennorm*

$$\left\|A^{-1}\right\|_\infty \leq \frac{\|\mathbf{e}\|_{\infty,d}}{\min_k (A\mathbf{e})_k}. \tag{2.3}$$

Der Vektor $\mathbf{e}$ wird majorisierendes Element genannt.

Beweis: Siehe Literatur [Boh81, AK90]. ■

Bemerkung 2.23 Zum M–Matrix–Kriterium.

- Das folgende Rezept ist oft erfolgreich, um ein majorisierendes Element zu konstruieren:
  - Finde eine Funktion $e(x) > 0$, so dass $(Le)(x) > 0$ für $x \in (0, 1)$. Das ist ein majorisierendes Element des Differentialoperators L.
  - Schränke $e(x)$ auf die Gitterfunktion e_h ein.

  Falls der erste Schritt in dieser Herangehensweise möglich ist, und die Diskretisierung konsistent ist, dann funktioniert die Herangehensweise im allgemeinen, zumindest für hinreichend kleine Gitterweite.
- Mit (2.3) kann man die Konstante c_S in der Stabilitätsdefinition abschätzen

  $$\|u_h\|_{\infty,d} = \left\|A^{-1}(f_h)\right\|_{\infty,d} \leq \left\|A^{-1}\right\|_\infty \|f_h\|_{\infty,d} = \left\|A^{-1}\right\|_\infty \|L_h u_h\|_{\infty,d},$$

  also gilt für diese Konstante

  $$c_S \leq \frac{\|\mathbf{e}\|_{\infty,d}}{\min_k (A\mathbf{e})_k}.$$

- Für Dirichletrandbedingungen eliminiert man die Variablen u_0 und u_N bevor man Satz 2.22 anwendet.

□

Beispiel 2.24 Zum M–Matrix–Kriterium. Betrachte (2.1) mit $b(x) \equiv 0$

$$Lu(x) = -u''(x) + c(x)u(x), \quad c(x) \geq 0 \text{ in } [0, 1].$$

Wähle $e(x) := \frac{1}{2}x(1 - x)$. Dann folgt

$$Le(x) = 1 + c(x)e(x) \geq 1.$$

Nun setzt man $e_h := R_h e$. Damit ergibt sich

$$(L_h e_h)_i = -D^+D^- e_{h,i} + c_i e_{h,i} = 1 + c_i e_{h,i} \geq 1,$$

weil der zweite Differenzenquotient die zweiten Ableitungen von quadratischen Funktionen in inneren Gitterpunkten exakt diskretisiert, siehe Beispiel 2.7. Das heißt

$$L_h e_h \geq (1, \ldots, 1)^T \quad \Longleftrightarrow \quad A\mathbf{e} \geq \mathbf{1}.$$

Für die Abschätzung der Stabilitätskonstanten erhält man

$$c_S \leq \frac{\|\mathbf{e}\|_{\infty,d}}{\min_k (A\mathbf{e})_k} \leq \frac{e_h(1/2)}{1} = \frac{1/8}{1} = \frac{1}{8}.$$

Dieses Beispiel zeigt, dass im Fall $b(x) \equiv 0$ die M–Matrix Eigenschaft ohne Einschränkungen an das Gitter gilt. □

Lemma 2.25 Stabilität des zentralen Differenzenschemas für hinreichend feines Gitter. *Für hinreichend kleine Gitterweite h ist das zentrale Differenzenschema (2.2) für das Randwertproblem (2.1) in der diskreten Maximumsnorm stabil. Die Koeffizientenmatrix ist eine M–Matrix.*

Beweis: Sei $e(x)$ die Lösung des Randwertproblems

$$-w'' + b(x)w' = 1, \quad w(0) = w(1) = 0.$$

Nach dem Maximumprinzip, Lemma 1.25, gilt $e(x) \geq 0$ für $x \in (0,1)$. Da $c(x) \equiv 0$ ist, folgt nach Folgerung 1.32, dass das obige Problem eindeutig lösbar ist und insbesondere $e \in C([0,1])$ gilt. Damit ist $e(x)$ beschränkt. Für innere Gitterpunkte gilt wegen $c(x) \geq 0$

$$\begin{aligned}
(L_h e_h)_i &= (R_h L e)_i + (L_h e_h - R_h L e)_i \\
&= \left(R_h \left(1 + c(x)e(x)\right)\right)_i + \left(-D^+D^- e_h + b_i D^0 e_h + c_i e_h - 1 - c_i e_h\right)_i \\
&\geq 1 + \left(-D^+D^- e_h + b_i D^0 e_h - 1\right)_i \\
&= \left(-D^+D^- e_h + b_i D^0 e_h\right)_i .
\end{aligned}$$

Da e_h die zu $e(x)$ gehörende Gitterfunktion ist, approximiert der Ausdruck in der letzten Zeile für hinreichend feines h den Term $-e''(x_i) + b(x_i)e'(x_i)(= 1)$ hinreichend gut, siehe Beispiel 2.7. Insbesondere gibt es ein $H > 0$, so dass für alle $h \in (0, H]$ gilt

$$(L_h e_h)_i \geq \frac{1}{2}.$$

Das M–Matrix–Kriterium beweist nun die Aussage des Satzes. ■

Folgerung 2.26 Konvergenz zweiter Ordnung des zentralen Differenzenschemas. *Unter der Annahme, dass $u \in C^4([0,1])$ gilt, konvergiert das zentrale Differenzenschema (2.2) von zweiter Ordnung.*

Beweis: Das folgt aus Satz 2.16, indem man Lemmata 2.13 und 2.25 kombiniert. ■

Beispiel 2.27 Betrachte das 2–Punkt–Randwertproblem

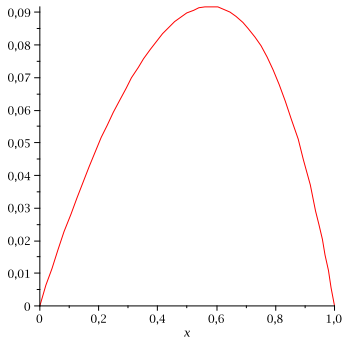
$$-u''(x) + 2u'(x) + 3u(x) = 1 \quad \text{in } (0,1), \quad u(0) = u(1) = 0.$$

Die Lösung dieses Problems lautet

$$u(x) = \frac{1}{3}\left(1 + \frac{1 - e^{-1}}{e^{-1} - e^3} e^{3x} + \frac{e^3 - 1}{e^{-1} - e^3} e^{-x}\right).$$

Man erhält die folgenden Fehler für unterschiedliche Gitterweiten:

Intervalle N	$\|u - u_h\|_{\infty,d}$
4	4.2388e-4
8	9.8811e-5
16	2.4529e-5
32	6.1537e-6
64	1.5368e-6
128	3.8440e-7
256	9.6093e-8
512	2.4023e-8
1024	6.0058e-9



Man erkennt, dass für hinreichend feine Gitter, sich der Fehler bei einer Halbierung der Gitterweite um den Faktor Vier verringert. Das ist zweite Ordnung Konvergenz. □

2.3 Upwind–Verfahren

Bemerkung 2.28 Von nun an werden Finite–Differenzen–Verfahren für das singulär gestörte Randwertproblem

$$Lu := -\varepsilon u'' + b(x)u' + c(x)u = f(x), \quad \text{für } x \in (0,1), \tag{2.4}$$

mit den Randbedingungen

$$u(0) = u(1) = 0, \tag{2.5}$$

unter den Voraussetzungen

$$\begin{aligned} \varepsilon &> 0, \\ b(x) &> 0 \quad \text{für alle } x \in [0,1], \\ c(x) &\geq 0 \quad \text{für alle } x \in [0,1], \end{aligned}$$

mit hinreichend glatten Funktion $b(x), c(x)$ und $f(x)$. Das Problem nennt man dann singulär gestört, wenn $\varepsilon \ll |b(x)|$ ist. Der Parameter ε wird singulärer Störungsparameter genannt. Für die Konvektion ist nur wichtig, dass $b(x) \neq 0$ für alle $x \in [0,1]$ gilt. Ist $b(x) < 0$ in $[0,1]$, gelangt man mit der Variablentransformation $x \mapsto 1 - x$ auf ein Problem mit den obigen Voraussetzungen.

Ist ε klein, so besitzt die Lösung von (2.4), (2.5) im allgemeinen eine Randgrenzschicht bei $x = 1$, vergleiche Beispiel 1.7. Diese Grenzschicht beeinflusst sowohl Stabilität als auch Konsistenz eines numerischen Verfahrens. Sind die Randbedingungen so gewählt, dass keine Grenzschicht auftritt, dann verbessert sich der Konsistenzfehler, aber die Stabilität des Verfahrens kann immer noch ein Problem sein. □

Beispiel 2.29 Zentrales Differenzenschema angewandt auf ein vereinfachtes singulär gestörtes Problem. Betrachte das Problem

$$-\varepsilon u'' + u' = 0 \text{ auf } (0,1), \quad u(0) = 0,\ u(1) = 1.$$

Die Lösung dieses Problems lautet

$$u(x) = \frac{e^{-(1-x)/\varepsilon} - e^{-1/\varepsilon}}{1 - e^{-1/\varepsilon}}.$$

Die Transformation $u(x) := x + v(x)$ würde auf ein Problem mit homogenen Randbedingungen führen. Man kann das Differenzenverfahren aber direkt auf das Problem mit inhomogenen Randbedingungen anwenden. Das wird in der Praxis auch so gemacht. Das diskrete Problem ist

$$-\varepsilon D^+ D^- u_i + D^0 u_i = 0, \quad u_0 = 0,\ u_N = 1.$$

Die Lösung dieses Gleichungssystems ist *Übungsaufgabe*

$$u_i = \frac{r^i - 1}{r^N - 1} \quad \text{mit} \quad r = \frac{2\varepsilon + h}{2\varepsilon - h}.$$

Es gilt insbesondere $|r| > 1$.

Ist $h \gg 2\varepsilon$, dann gilt $r \approx -1$ und es folgt

$$u_i \approx \frac{(-1)^i - 1}{(-1)^N - 1}.$$

Ist N gerade, so wird durch eine sehr kleine positive Zahl dividiert. Für gerade i, ist der Zähler ebenfalls klein und positiv, für ungerade i ist der Zähler negativ. Die diskrete Lösung oszilliert sehr stark, siehe Abbildung 2.1.

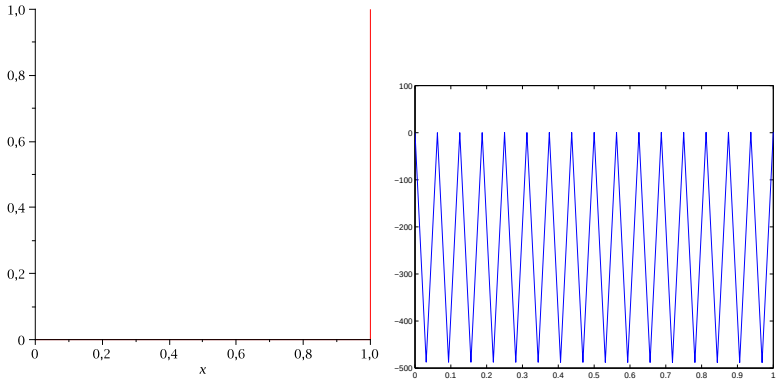


Abbildung 2.1: Lösung und diskrete Lösung mit dem zentralen Differenzenschema für $\varepsilon = 10^{-6}$ und $h = 1/32$.

Ist jedoch $h < 2\varepsilon$, dann erhält man mit dem zentralen Differenzenschema eine sinnvolle Approximation der Lösung. In Anwendungen ist jedoch oft $\varepsilon \leq 10^{-6}$, so dass man sehr feine Gitter braucht, um das zentrale Differenzenschema anwenden zu können. In einer Dimension ist das heute oft möglich, für Probleme in zwei oder drei Dimensionen jedoch nicht. □

Bemerkung 2.30 Zentrales Differenzenschema angewandt auf das allgemeine singulär gestörte Problem. Betrachte nun das singulär gestörte Problem (2.4), (2.5) und schreibe das Differenzenschema in der Form aus Bemerkung 2.11

$$r_i u_{i-1} + s_i u_i + t_i u_{i+1} = f_i, \quad i = 1, \ldots, N-1, \quad u_0 = u_N = 0,$$

mit

$$r_i = -\frac{\varepsilon}{h^2} - \frac{1}{2h} b_i, \quad s_i = c_i + \frac{2\varepsilon}{h^2}, \quad t_i = -\frac{\varepsilon}{h^2} + \frac{1}{2h} b_i, \quad b_i > 0.$$

Damit erhält man eine M–Matrix und somit Stabilität, falls man

$$t_i \leq 0 \quad \Longrightarrow \quad h \leq h_0(\varepsilon) = \frac{2\varepsilon}{\|b\|_\infty}$$

voraussetzt. Dies verallgemeinert die Beobachtung aus Beispiel 2.29. Man beachte, dass $h_0(\varepsilon) \to 0$ für $\varepsilon \to 0$ gilt. □

Bemerkung 2.31 Motivation für Upwind–Verfahren. Eine andere heuristische Erklärung für das Versagen des zentralen Differenzenverfahrens für $\varepsilon \ll h$ ist wie folgt. In diesem Fall besitzt das Verfahren für Beispiel 2.29 im wesentlichen die Gestalt

$$D^0 u_i = 0, \quad \Longleftrightarrow \quad \frac{u_{i+1} - u_{i-1}}{2h} = 0.$$

Daraus folgt insbesondere $u_{N-2} \approx u_N = 1$. Das ist eine schlechte Approximation des exakten Wertes $u(x_{N-2}) \approx 0$.

Diese Beobachtung sagt uns, dass es zur Approximation von $u'(x_{N-1})$ besser ist, den Wert an der Stelle u_N nicht zu verwenden. Der einfachste Kandidat, der diese Bedingung erfüllt, ist

$$u'(x_i) \approx \frac{u_i - u_{i-1}}{h}.$$

Verfolgt man das Ziel, die Matrixeinträge des diskreten Problems vom zentralen Differenzenschema so zu modifizieren, dass man eine M–Matrix erhält, so kann man ebenfalls diese Approximation motivieren. □

Definition 2.32 Einfaches Upwind–Verfahren. Das einfache Upwind–Verfahren für das singulär gestörte Randwertproblem (2.4), (2.5) besitzt die Gestalt

$$\begin{aligned} -\varepsilon D^+D^-u_i + b_i D^{\mathcal{N}} u_i + c_i u_i &= f_i, \quad \text{für } i = 1, \ldots, N-1, \\ u_0 = u_N &= 0 \end{aligned} \tag{2.6}$$

mit

$$D^{\mathcal{N}} := \begin{cases} D^+ & \text{für } b < 0, \\ D^- & \text{für } b > 0. \end{cases}$$

□

Bemerkung 2.33 Zum einfachen Upwind–Verfahren.

- Upwind, deutsch stromaufwärts, bedeutet, dass die finite Differenzenapproximation des konvektiven Termes mit Werten aus der Stromaufwärts–Richtung genommen. Bei konvektions–dominanten Problem erfolgt der Informationstransport in Richtung der Konvektion. Aus der Stromaufwärts–Richtung kommt daher die Information.
- Mit dem einfachen Upwind–Verfahren erhält man eine viel bessere Approximation der Lösung aus Beispiel 2.29, siehe Abbildung 2.2.

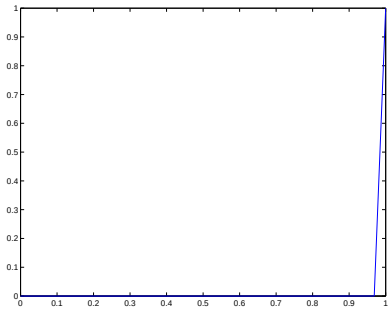


Abbildung 2.2: Diskrete Lösung mit dem einfachen Upwind–Verfahren für $\varepsilon = 10^{-6}$ und $h = 1/32$.

- Beim einfachen Upwind–Verfahren wird die Approximation zweiter Ordnung D^0 durch eine Approximation erster Ordnung, D^+ oder D^- ersetzt. Das wird sich natürlich in der Genauigkeit des Verfahrens bemerkbar machen.
- Sei L_h die Matrix des Upwind–Verfahrens, nachdem die Randwerte u_0 und u_N eliminiert wurden. In der Form von Bemerkung 2.11 besitzen die Matrixeinträge die Gestalt

  $$r_i = -\frac{\varepsilon}{h^2} - \frac{1}{h}\max\{0, b_i\}, \quad s_i = c_i + \frac{2\varepsilon}{h^2} + \frac{1}{h}\left|b_i\right|, \quad t_i = -\frac{\varepsilon}{h^2} + \frac{1}{h}\min\{0, b_i\}.$$

  Nun sind die Nichtdiagonaleinträge nichtpositiv, unabhängig von der Größe von ε und h.

□

Satz 2.34 Stabilität des einfachen Upwind–Verfahrens. *Unter den in Bemerkung 2.28 gemachten Voraussetzungen ist die Koeffizientenmatrix L_h des einfachen Upwind–Verfahrens (2.6) eine M–Matrix. Das einfache Upwind–Verfahren ist gleichmäßig stabil bezüglich des Parameters ε, das heißt es gilt*

$$\|u_h\|_{\infty,d} \leq c_S \|L_h u_h\|_{\infty,d}$$

mit einer von ε und h unabhängigen Stabilitätskonstanten $c_S > 0$.

Beweis: Betrachte nur den Fall $b(x) \geq \beta > 0$. Man konstruiert einen geeignetes majorisierendes Element. Wähle $e(x) = x$, dann gilt

$$Le(x) = -\varepsilon e''(x) + b(x)e'(x) + c(x)e(x) = b(x) + xc(x) \geq \beta.$$

Für das einfache Upwind–Verfahren und die Gitterfunktion e_h erhält man

$$\begin{aligned}
(L_h e_h)_i &= r_i x_{i-1} + s_i x_i + t_i x_{i+1} \\
&= \left(-\frac{\varepsilon}{h^2} - \frac{1}{h} b_i\right)(x_i - h) + \left(c_i + \frac{2\varepsilon}{h^2} + \frac{1}{h} b_i\right) x_i - \frac{\varepsilon}{h^2}(x_i + h) \\
&= \left(-\frac{\varepsilon}{h^2} - \frac{1}{h} b_i + c_i + \frac{2\varepsilon}{h^2} + \frac{1}{h} b_i - \frac{\varepsilon}{h^2}\right) x_i + \left(\frac{\varepsilon}{h^2} + \frac{1}{h} b_i - \frac{\varepsilon}{h^2}\right) h \\
&= c_i x_i + b_i \geq \beta.
\end{aligned}$$

Nach Satz 2.22 ist L_h eine M–Matrix. Mit der Abschätzung der Stabilitätskonstanten aus Bemerkung 2.23 erhält man schließlich

$$c_S \leq \frac{\|e_h\|_{\infty,d}}{\min_k (L_h e_h)_k} = \frac{1}{\beta}.$$

■

Zur Konsistenzuntersuchung benötigt man eine relativ genaue Abschätzung der Ableitungen der Lösung des stetigen Problems.

Lemma 2.35 *Seien $b(x) \geq \beta > 0$ und $b(x), c(x), f(x)$ hinreichend glatt. Dann erfüllt die Lösung $u(x)$ von (2.4), (2.5)*

$$\left|u^{(i)}(x)\right| \leq C\left[1 + \varepsilon^{-i} \exp\left(-\beta \frac{1-x}{\varepsilon}\right)\right], \quad i = 1, 2, \ldots, q,$$

für $x \in [0, 1]$. Die maximale Ordnung q hängt von der Glätte der Daten ab.

Beweis: Der Beweis erfolgte in [KT78], man findet ihn in [RST08, S. 21]. ■

Satz 2.36 Konsistenz des einfachen Upwind–Verfahrens. *Unter den in Bemerkung 2.28 gemachten Voraussetzungen mit $b(x) \geq \beta > 0$ existiert eine positive Konstante β^*, die nur von β abhängt, so dass der Fehler des einfachen Upwind–Verfahrens (2.6) in den inneren Gitterpunkten $\{x_i \ : \ i = 1, \ldots, N-1\}$*

$$|u(x_i) - u_i| \leq \begin{cases} Ch[1 + \varepsilon^{-1} \exp(-\beta^*(1 - x_i)/\varepsilon)] & \text{falls } h < \varepsilon, \\ Ch + C \exp(-\beta^*(1 - x_{i+1})/\varepsilon) & \text{falls } h \geq \varepsilon \end{cases}$$

erfüllt.

Beweis: Der Beweis folgt [KT78]. Hier wird nur der interessante Fall $h \geq \varepsilon$ betrachtet und auch für diesen wird der Beweis nicht ganz vollständig angegeben. Den vollständigen Beweis findet man in [RST08, S. 49f.].

Im Fall $h \geq \varepsilon$ zerlegt man die Lösung von (2.4), (2.5) in

$$u(x) = -u_0(1) \exp\left(-\frac{b(1)(1-x)}{\varepsilon}\right) + z(x) =: v(x) + z(x),$$

wobei $u_0(x)$ die reduzierte Lösung ist. Analog zu Lemma 2.35 findet man

$$\left|z^{(k)}(x)\right| \leq C\left[1 + \varepsilon^{1-k} \exp\left(-\frac{b(1)(1-x)}{\varepsilon}\right)\right], \quad k = 1, 2, 3.$$

Es gilt

$$L_h u_h = f_h = R_h(f) = R_h(Lu) = R_h\left(L(v+z)\right) = R_h(Lv) + R_h(Lz).$$

Damit hat man eine Zerlegung $u_h = v_h + z_h$, wobei die Gitterfunktionen durch

$$L_h v_h = R_h(Lv) \quad \text{und} \quad L_h z_h = R_h(Lz)$$

definiert sind, wobei v_h und z_h mit $v(x)$ beziehungsweise $z(x)$ in x_0 und x_N übereinstimmen. Mit Dreiecksungleichung gilt

$$|u(x_i) - u_i| = |v(x_i) + z(x_i) - (v_i + z_i)| \leq |v(x_i) - v_i| + |z(x_i) - z_i| \,.$$

Betrachte nun den Konsistenzfehler für $z(x)$. Dazu wird die Taylor–Entwicklung von $z(x)$ im Punkt x_i verwendet und man erhält im ersten Schritt *Übungsaufgabe*

$$|\tau_i| := |L_h z(x_i) - f(x_i) - v(x_i)| \leq C \int_{x_{i-1}}^{x_{i+1}} \left(\varepsilon \left| z^{(3)}(t) \right| + \left| z^{(2)}(t) \right| \right) \, dt.$$

Die rechte Seite kommt durch die Restglieder. Nun verwendet man die obige Abschätzung für die Ableitungen von $z(x)$

$$\begin{aligned}
|\tau_i| &\leq C \int_{x_{i-1}}^{x_{i+1}} \left(\varepsilon + \varepsilon^{-1} \exp\left(-b(1) \frac{1-t}{\varepsilon} \right) + 1 + \varepsilon^{-1} \exp\left(-b(1) \frac{1-t}{\varepsilon} \right) \right) \, dt \\
&\leq \int_{x_{i-1}}^{x_{i+1}} (\varepsilon + 1) \, dt + \varepsilon^{-1} \int_{x_{i-1}}^{x_{i+1}} \left(\exp\left(-\beta \frac{1-t}{\varepsilon} \right) + \exp\left(-\beta \frac{1-t}{\varepsilon} \right) \right) \, dt \\
&\leq Ch + C\varepsilon^{-1} \int_{x_{i-1}}^{x_{i+1}} \exp\left(-\beta \frac{1-t}{\varepsilon} \right) \, dt \\
&= Ch + C\varepsilon^{-1} \left(\frac{\varepsilon}{\beta} \exp\left(-\beta \frac{1-t}{\varepsilon} \right) \Big|_{x_i - h}^{x_i + h} \right) \\
&= Ch + C \left[\exp\left(-\beta \frac{1 - x_i - h}{\varepsilon} \right) - \exp\left(-\beta \frac{1 - x_i + h}{\varepsilon} \right) \right] \\
&= Ch + C \exp\left(-\beta \frac{1-x_i}{\varepsilon} \right) \left[\exp\left(\frac{\beta h}{\varepsilon} \right) - \exp\left(-\frac{\beta h}{\varepsilon} \right) \right] \\
&= Ch + C \sinh\left(\frac{\beta h}{\varepsilon} \right) \exp\left(-\beta \frac{1-x_i}{\varepsilon} \right).
\end{aligned}$$

Es gilt

$$\sinh(t) = \frac{e^t - e^{-t}}{2} \leq \frac{e^t}{2} = Ce^t.$$

Damit folgt

$$|\tau_i| \leq Ch + C \exp\left(-\beta \frac{1-x_i}{\varepsilon} + \frac{\beta h}{\varepsilon} \right) = Ch + C \exp\left(-\beta \frac{1 - x_{i+1}}{\varepsilon} \right).$$

Auch für den Konsistenzfehler des zweiten Anteiles $v(x)$ findet man

$$|v(x_i) - v_i| \leq Ch + C \exp\left(-\beta \frac{1 - x_{i+1}}{\varepsilon} \right).$$

Die Kombination beider Anteile ergibt die Gesamtabschätzung. ■

Folgerung 2.37 Konvergenz des einfachen Upwind–Verfahrens außerhalb von Grenzschichten. *Unter den Voraussetzungen von Satz 2.34 und Satz 2.36 konvergiert das einfache Upwind–Verfahren in einem Intervall $[0, 1-\delta]$ für festes $\delta > 0$ von erster Ordnung mit einer Konvergenzkonstante unabhängig von ε.*

Bemerkung 2.38 Verhalten innerhalb der Grenzschicht basierend auf der Abschätzung. Sei $\varepsilon < h$, dann erhält man im Punkt x_{N-2} die Abschätzung

$$\begin{aligned}
|u(x_{N-2}) - u_{N-2}| &\leq Ch + C \exp\left(-\beta^* \frac{1 - x_{N-1}}{\varepsilon} \right) = Ch + C \exp\left(-\beta^* \frac{h}{\varepsilon} \right) \\
&\leq Ch + Ch = \mathcal{O}(h),
\end{aligned}$$

da die Exponentialfunktion mit einem betragsmäßig großen negativen Argument gegen die lineare Funktion abgeschätzt werden kann. Für x_{N-1} erhält man jedoch

$$|u(x_{N-1}) - u_{N-1}| \leq Ch + C \exp\left(-\beta^* \frac{1 - x_N}{\varepsilon}\right) = Ch + C = \mathcal{O}(1),$$

da $x_N = 1$. □

Beispiel 2.39 Die obige Beobachtung ist kein Problem der erzielten Abschätzung. Betrachte

$$-\varepsilon u''(x) - u'(x) = 0, \quad u(0) = 0,\ u(1) = 1.$$

Die Lösung dieses Problems besitzt eine Grenzschicht bei $x = 0$. Mit dem einfachen Upwind–Verfahren erhält man

$$u_i = \frac{1 - r^i}{1 - r^N}, \quad \text{mit} \quad r = \frac{\varepsilon}{\varepsilon + h}.$$

Für $h = \varepsilon$ erhält man

$$u_1 = \frac{1 - r}{1 - r^N} = \frac{1 - 1/2}{1 - (1/2)^N} = \frac{1/2}{1 - (1/2)^N} \approx \frac{1}{2}.$$

Für die Lösung gilt jedoch

$$u(x_1) = \frac{1 - e^{-1}}{1 - e^{-1/\varepsilon}} \approx 0.63$$

für kleine ε. Damit ist der Fehler $\mathcal{O}(1)$. Somit kann man nicht erwarten, die Abschätzung aus Satz 2.36 wesentlich zu verbessern. □

Bemerkung 2.40 Typisches Verhalten innerhalb der Grenzschicht in numerischen Simulationen. Betrachte konstantes ε und variables h. Ist h groß genug, dann liegen alle Gitterpunkte außerhalb der Grenzschicht. Wird h verkleinert, erhöht sich der Fehler, weil dann der erste Gitterpunkt von außerhalb sich in die Grenzschicht hineinbewegt, siehe Abbildung 2.3. Wenn h dann hinreichend klein wird, dann fällt der Fehler wieder. In diesem Falle greift die erste Abschätzung von Satz 2.36.

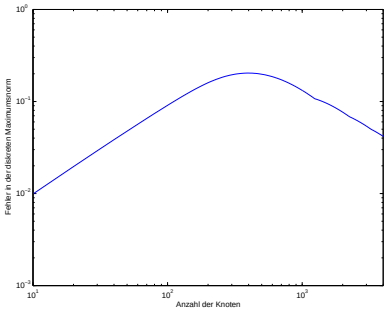


Abbildung 2.3: Fehler des einfachen Upwind–Verfahrens für Beispiel 1.7, $\varepsilon = 1e-3$ und unterschiedlicher Anzahl von Gitterpunkten.

□

Bemerkung 2.41 Interpretation des Upwind–Verfahrens als künstliche Diffusion. Die Schwierigkeiten der numerischen Lösung eines singulär gestörten Problems liegen in den unterschiedlichen Größenordnungen von Diffusion und Konvektion. Es ist klar, dass die numerische Lösung einfacher wird, je größer die Diffusion im Vergleich zur Konvektion ist.

Betrachte $b > 0$. Dann gilt

$$\begin{aligned} b_i D^{\mathcal{N}} u_i &= b_i D^- u_i = \frac{u_i - u_{i-1}}{h} = b_i \frac{u_{i+1} - u_{i-1}}{2h} + b_i \frac{-u_{i+1} + 2u_i - u_{i-1}}{2h} \\ &= b_i D^0 u_i - \frac{b_i h}{2} D^+ D^- u_i. \end{aligned}$$

Damit kann das einfache Upwind–Verfahren (2.6) in der Form

$$\begin{aligned} -\left(\varepsilon + \frac{b_i h}{2}\right) D^+ D^- u_i + b_i D^0 u_i + c_i u_i &= f_i, \quad \text{für } i = 1, \ldots, N-1, \\ u_0 = u_N &= 0, \end{aligned}$$

geschrieben werden.

Der Diffusionskoeffizient wird also künstlich erhöht und er besitzt die Größenordnung $\mathcal{O}(h)$. Das einfache Upwind–Verfahren ist also nichts anderes als das zentrale Differenzenverfahren angewandt auf ein Problem mit hinreichend großer, $\mathcal{O}(h)$, Diffusions. Man hat bereits in Beispiel 2.29 gesehen, dass das zentrale Differenzenverfahren für einen Diffusionskoeffizienten der Größenordnung $\mathcal{O}(h)$ vernünftige Ergebnisse liefert. □

Man kann Verfahren mit künstlicher Diffusion auch direkt definieren.

Definition 2.42 Verfahren mit künstlicher Diffusion, angepasstes Upwind–Verfahren. Ein Finite–Differenzen–Verfahren mit künstlicher Diffusion ist durch

$$\begin{aligned} -\varepsilon \sigma(q(x_i)) D^+ D^- u_i + b_i D^0 u_i + c_i u_i &= f_i, \quad \text{für } i = 1, \ldots, N-1, \\ u_0 = u_N &= 0, \\ q(x) &:= \frac{b(x)h}{2\varepsilon}, \end{aligned} \tag{2.7}$$

gegeben. Man nennt das Verfahren auch angepasstes Upwind–Verfahren. □

Bemerkung 2.43

- Das einfache Upwind–Verfahren (2.6) erhält man für $\sigma(q) = 1 + q$.
- Die Einführung künstlicher Diffusion verfälscht die ursprüngliche Aufgabe erheblich. Betrachte beispielsweise

  $$-\varepsilon u'' + u' = 1 \quad \text{auf } (0,1), \quad u(0) = u(1) = 0$$

  mit der Lösung

  $$u(x) = x - \frac{\exp\left(-\frac{1-x}{\varepsilon}\right) - \exp\left(-\frac{1}{\varepsilon}\right)}{1 - \exp\left(-\frac{1}{\varepsilon}\right)}, \tag{2.8}$$

  siehe Beispiel 1.7. Der zweite Term ist für das Erfüllen der Randbedingung bei $x = 1$ verantwortlich. Er ist nur wesentlich von Null verschieden im Intervall $[1 - \varepsilon, 1]$, siehe Abbildung 2.4. Führt man künstliche Diffusion ein, dann erhält man eine gestörte Lösung und der Term, der für das Erfülltsein der Randbedingung verantwortlich ist, ist im Intervall $[1 - \varepsilon\sigma(q(x_{N-1})), 1]$ wesentlich von Null verschieden. Das bedeutet, die Grenzschicht ist (weit) weniger steil. Man sagt, die Grenzschicht wird verschmiert.

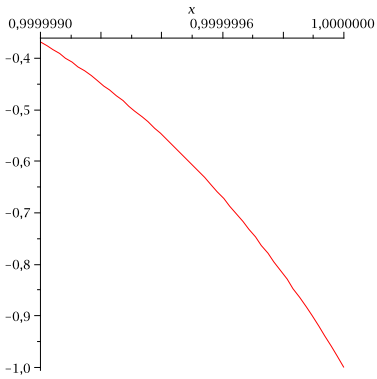


Abbildung 2.4: Zweiter Term der Lösung (2.8) für $\varepsilon = 10^{-6}$.

Beim einfachen Upwind–Verfahren ist

$$\varepsilon\sigma(q(x_{N-1})) = \varepsilon + \varepsilon\frac{b_{N-1}h}{2\varepsilon} = \varepsilon + \frac{b_{N-1}h}{2}.$$

Dieser Ausdruck ist in realistischen Situationen, das heißt für $\varepsilon \ll b_{N-1}$ und $\varepsilon \ll h$, um Ordnungen größer als ε.

□

Die Frage ist, ob man stabile Verfahren mit weniger Verschmierungen konstruieren kann.

Satz 2.44 Stabilität von Verfahren mit künstlicher Diffusion. *Seien $b(x) > \beta > 0$, $c(x) \geq 0$ und $\sigma(q) > q$. Dann ist die Koeffizientenmatrix des Verfahrens mit künstlicher Diffusion (2.7) eine M–Matrix und das Verfahren ist stabil in der diskrete Maximumsnorm. Die Stabilitätskonstante hängt nicht von ε ab.*

Beweis: Der Beweis geht im Prinzip wie der von Satz 2.34, Übungsaufgabe. ■

Satz 2.45 Konsistenz von Verfahren mit künstlicher Diffusion. *Seien die Voraussetzungen von Satz 2.44 gegeben, sei $u \in C^4([0,1])$ und sei*

$$|\sigma(q) - 1| \leq \min\{q, Mq^2\},$$

mit einer Konstanten $M > 0$. Dann ist für festes ε der Konsistenzfehler des Verfahrens mit künstlicher Diffusion (2.7) von zweiter Ordnung.

Beweis: Der Konsistenzfehler im Punkt x_i ist

$$\begin{aligned} |\tau_i| &= \big|\big[-\varepsilon\sigma(q_i)\, D^+D^-u(x_i) + b_i D^0 u(x_i) + c_i u(x_i)\big] \\ &\quad -\big[-\varepsilon u''(x_i) + b_i u'(x_i) + c_i u(x_i)\big]\big| \\ &= \big|\varepsilon\sigma(q_i)\left(u''(x_i) - D^+D^-u_i\right) + \varepsilon\left(1-\sigma(q_i)\right)u''(x_i) + b_i\left(D^0u(x_i) - u'(x_i)\right)\big|. \end{aligned}$$

Aus den Konsistenzfehlerabschätzungen aus Beispiel 2.7 folgt

$$|\tau_i| \leq C\left[\varepsilon\,|\sigma(q_i)|\,h^2\left\|u^{(4)}\right\|_\infty + \varepsilon\,|1-\sigma(q_i)|\left\|u^{(2)}\right\|_\infty + h^2\left\|u^{(3)}\right\|_\infty\right].$$

Mit der Voraussetzung des Satzes und der Definition von $q(x)$ ergeben sich

$$\begin{aligned} |\sigma(q_i)| &\leq |\sigma(q_i) - 1| + 1 \leq \min\left\{q_i, Mq_i^2\right\} + 1 \leq q_i + 1 \leq C\frac{h}{\varepsilon} + 1, \\ |1-\sigma(q_i)| &\leq Mq_i^2 \leq C\frac{h^2}{\varepsilon^2}. \end{aligned}$$

Durch Einsetzen folgt

$$\begin{aligned} |\tau_i| &\leq C\left[\left(\frac{h}{\varepsilon}+1\right)\varepsilon h^2\left\|u^{(4)}\right\|_\infty + \varepsilon C\frac{h^2}{\varepsilon^2}\left\|u^{(2)}\right\|_\infty + h^2\left\|u^{(3)}\right\|_\infty\right] \\ &\leq C(\varepsilon)h^2. \end{aligned} \tag{2.9}$$

■

Bemerkung 2.46

- Beispiel für Funktionen $\sigma(q)$, welche die Voraussetzungen von Satz 2.45 erfüllen sind *Übungsaufgabe*

$$\sigma(q) = \max\{1, q\}, \quad \sigma(q) = \sqrt{1+q^2}, \quad \sigma(q) = 1 + \frac{q^2}{1+q}.$$

  Die letzte Variante wird Samarskii[2]–Upwind–Verfahren genannt.
- Die Konsistenz ist nur für konstantes ε von zweiter Ordnung. Der Vorfaktor $C(\varepsilon)$ divergiert gegen Unendlich für $\varepsilon \to 0$, siehe mittlerer Summand in (2.9). Damit werden die Verfahren für $\varepsilon \to 0$ immer schlechter. Man kann zeigen, dass die Konsistenz unabhängig von ε außerhalb der Grenzschicht nur von erster Ordnung ist. Das typische Verhalten in der Grenzschicht ist wie beim einfachen Upwind–Verfahren, siehe Bemerkung 2.40.

□

Bemerkung 2.47 Fazit.

- Das zentrale Differenzenverfahren ist für singulär gestörte Probleme nicht geeignet.
- Das einfache Upwind–Verfahren ist stabil, aber zu ungenau (von erster Ordnung konsistent). Es verschmiert die Grenzschichten.
- Upwind–Verfahren lassen sich als Verfahren mit künstlicher Diffusion interpretieren.
- Angepasste Upwind–Verfahren können für festes ε von zweiter Ordnung konsistent sein. Diese Eigenschaft ist aber nicht gleichmäßig in ε.

Die bisher vorgestellten Upwind–Verfahren sind nicht befriedigend, da sie für kleine ε zu ungenau sind und die Konvergenz innerhalb der Grenzschicht von ε abhängt.

□

2.4 Gleichmäßig konvergente Verfahren

Bemerkung 2.48 Motivation. Ziel ist es, Verfahren zu entwickeln, die im gesamten Intervall $[0, 1]$ gleichmäßig konvergieren, also insbesondere auch innerhalb der Grenzschicht. Dazu werden zwei Wege vorgestellt:

- ein Verfahren, welches man durch eine geeignete Wahl der künstlichen Diffusion $\sigma(q)$ in (2.7) erhält,
- Verfahren, welche man durch die Wahl geeigneter Gitter definiert.

In der Praxis hat man oft sehr kleine Diffusionen. Deshalb ist es wichtig, dass numerische Verfahren auch für diese Fälle gute Ergebnisse liefern. Die Konstruktion solcher Verfahren ist nicht trivial. Das wird schon dadurch klar, dass der Grenzübergang $\varepsilon \to 0$ in gewisser Weise unstetig ist, weil sich dadurch die Ordnung der Differentialgleichung ändert. Damit ändert sich zum Beispiel die Anzahl der benötigten Randbedingungen, aber auch die Eigenschaften von Lösungen der Differentialgleichungen unterschiedlicher Ordnung sind unterschiedlich, zum Beispiel die Glätte.

[2] Alexander Andreewitsch Samarskii (1919 – 2008)

Ein gleichmäßig konvergentes numerisches Verfahren muss diesen Grenzübergang ohne Qualitätsverlust bewerkstelligen können.

Dieser Abschnitt folgt teilweise [GR05]. □

Definition 2.49 Gleichmäßige Konvergenz. Man nennt ein Verfahren zur Lösung von (2.4), (2.5) gleichmäßig konvergent von der Ordnung p bezüglich des singulären Störungsparameters ε in der diskreten Maximumsnorm, wenn eine Abschätzung der Form

$$\|u - u_h\|_{\infty,d} \leq Ch^p, \quad p > 0,$$

mit einer von ε unabhängigen Konstanten C gilt. □

2.4.1 Geeignete künstliche Diffusion

Die Wahl einer geeigneten künstlichen Diffusion $\sigma(q)$ lässt sich motivieren, indem man die Lösung von (2.4), (2.5) für $\varepsilon \to 0$ betrachtet.

Lemma 2.50 Konvergenz gegen reduzierte Lösung. *Sei $u(x,\varepsilon)$ die Lösung von (2.4), (2.5) mit $b(x) \geq \beta > 0$, $c(x) \geq 0$ und sei $u_0(x)$ die Lösung des reduzierten Problems. Dann gilt für alle $x \in [0, x_0)$ mit $x_0 < 1$*

$$\lim_{\varepsilon \to 0} u(x,\varepsilon) = u_0(x).$$

Beweis: Der Beweis beruht auf dem Vergleichsprinzip, Folgerung 1.27. Setze

$$v_1(x) := \gamma \exp(\beta x), \quad \gamma > 0,$$

dann folgt

$$(Lv_1)(x) = \gamma(-\varepsilon\beta^2 + b(x)\beta + c(x)) \exp(\beta x) \geq \gamma\beta^2(1-\varepsilon)\exp(\beta x) \geq 1$$

für hinreichend großes γ. Weiter setzt man

$$v_2(x) := \exp\left(-\beta\frac{1-x}{\varepsilon}\right).$$

Dann gilt

$$\begin{aligned}(Lv_2)(x) &= \left(-\varepsilon\frac{\beta^2}{\varepsilon^2} + b(x)\frac{\beta}{\varepsilon} + c(x)\right)\exp\left(-\beta\frac{1-x}{\varepsilon}\right)\\ &\geq \frac{\beta}{\varepsilon}\left(-\beta + b(x)\right)\exp\left(-\beta\frac{1-x}{\varepsilon}\right) \geq 0.\end{aligned}$$

Betrachte nun

$$v(x) := M_1\varepsilon v_1(x) + M_2 v_2(x).$$

Dann sind

$$\begin{aligned}(Lv)(x) &= M_1\varepsilon(Lv_1)(x) + M_2(Lv_2)(x) \geq M_1\varepsilon(Lv_1)(x) \geq M_1\varepsilon \geq \varepsilon\left|u_0''(x)\right|\\ &= |L(u-u_0)(x)|,\\ v(0) &= M_1\varepsilon v_1(0) + M_2 v_2(0) = M_1\varepsilon\gamma + M_2\exp(-\beta/\varepsilon) \geq 0 = |(u-u_0)(0)|,\\ v(1) &= M_1\varepsilon v_1(1) + M_2 v_2(1) = M_1\varepsilon\gamma\exp(\beta) + M_2 \geq M_2 \geq |u_0(1)|,\end{aligned}$$

für geeignet gewählte, von ε unabhängige, Konstanten M_1 und M_2. Die Konstanten müssen hinreichend groß sein und sie hängen nur von $u_0(x)$ ab. Nach dem Vergleichsprinzip folgt

$$|(u-u_0)(x)| \leq v(x) = M_1\varepsilon\gamma\exp(\beta x) + M_2\exp\left(-\beta\frac{1-x}{\varepsilon}\right).$$

Damit erhält man für $x < 1$

$$\lim_{\varepsilon\to 0}|(u-u_0)(x)| = 0.$$

■

Lemma 2.51 *Unter den Voraussetzungen von Lemma 2.50 existiert eine von x und ε unabhängige Konstante C, so dass für die Lösung von (2.4), (2.5) gilt*

$$\left| u(x,\varepsilon) - \left[u_0(x) - u_0(1)\exp\left(-b(1)\frac{1-x}{\varepsilon}\right)\right]\right| \leq C\varepsilon, \quad x \in [0,1].$$

Beweis: Der Beweis ist ähnlich wie der von Lemma 2.50. ∎

Bemerkung 2.52 Notwendige Bedingung für eine geeigneten Funktion $\sigma(q)$. Seien $\rho^* := h/\varepsilon$ fest und i fest. Das heißt, für $h \to 0$ gilt auch $\varepsilon \to 0$. Ziel ist es, in diesem Falle eine Bedingung für eine geeignete Funktion $\sigma(q)$ zu finden. Wegen $\varepsilon \to 0$ für $h \to 0$ folgt nach Lemma 2.51

$$\begin{aligned}\lim_{h\to 0} u(1-ih) &= \lim_{h\to 0} u((N-i)h) = \lim_{h\to 0}\left(u_0(1) - u_0(1)\exp\left(-b(1)\frac{ih}{\varepsilon}\right)\right)\\ &= u_0(1) - u_0(1)\exp\left(-ib(1)\rho^*\right)\\ &= u_0(1)\left(1-\exp(-2iq(1)\right). \end{aligned} \tag{2.10}$$

Das angepasste Upwind–Verfahren besitzt die Gestalt

$$-\varepsilon\sigma\left(q(b_i)\right)\frac{u_{i+1}-2u_i+u_{i-1}}{h^2} + b_i\frac{u_{i+1}-u_{i-1}}{2h} = f_i - c_iu_i$$

oder nach Erweiterung mit h^2/ε

$$\begin{aligned}-\sigma\left(q(b_i)\right)\left(u_{i+1}-2u_i+u_{i-1}\right) + q(b_i)\left(u_{i+1}-u_{i-1}\right) &= \frac{h^2}{\varepsilon}\left(f_i - c_iu_i\right)\\ &= h\rho^*\left(f_i - c_iu_i\right).\end{aligned}$$

Für den rechten Rand, das heißt $i = N-1$ gilt insbesondere

$$\begin{aligned}&\lim_{h\to 0}\left(-\sigma\left(q_{N-1}\right)\left(u_N - 2u_{N-1}+u_{N-2}\right) + q_{N-1}\left(u_N - u_{N-2}\right)\right)\\ &\quad= \lim_{h\to 0} h\rho^*\left(f_{N-1} - c_{N-1}u_{N-1}\right) \quad \Longrightarrow\\ 0 &= \lim_{h\to 0}\left(-\sigma\left(q_{N-1}\right)\left(u_N - 2u_{N-1}+u_{N-2}\right) + q_{N-1}\left(u_N - u_{N-2}\right)\right).\end{aligned}$$

Einsetzen von (2.10) liefert, wobei ohne Beschränkung der Allgemeinheit $u_0(1) \neq 0$ angenommen wird,

$$\begin{aligned}0 &= -\sigma\left(q(1)\right)\Big(-\exp(-2Nq(1)) + 2\exp(-2(N-1)q(1))\\ &\quad -\exp(-2(N-2)q(1))\Big) + q(1)\Big(-\exp(-2Nq(1)) + \exp(-2(N-2)q(1))\Big)\end{aligned}$$

und nach Division durch $-\exp(-2Nq(1)) \neq 0$

$$0 = -\sigma\left(q(1)\right)\Big(1 - 2\exp(2q(1)) + \exp(4q(1))\Big) + q(1)\Big(1-\exp(4q(1))\Big).$$

Nun gilt

$$\frac{1-e^{4x}}{1-2e^{2x}+e^{4x}} = \frac{e^{-2x}-e^{2x}}{e^{-2x}-2+e^{2x}} = \frac{\left(e^x - e^{-x}\right)\left(e^x+e^{-x}\right)}{\left(e^x-e^{-x}\right)^2} = \frac{e^x+e^{-x}}{e^x-e^{-x}} = \coth(x).$$

Damit folgt

$$\sigma\left(q(1)\right) = q(1)\frac{1-\exp(4q(1))}{1-2\exp(2q(1))+\exp(4q(1))} = q(1)\coth\left(q(1)\right).$$

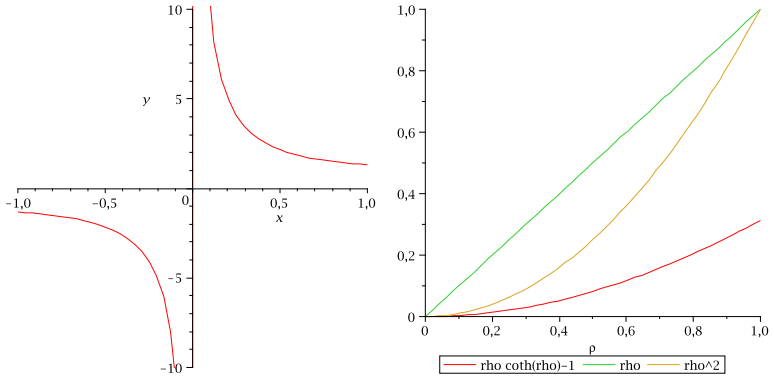


Abbildung 2.5: $\coth(x)$ und Vergleich zu den Bedingungen aus Satz 2.45.

Eine Wahl, die diesem Grenzwert genügt ist

$$\sigma(q) = q \coth(q).$$

Diese Funktion erfüllt auch die Bedingungen für die Konsistenz von Verfahren mit künstlicher Diffusion, Satz 2.45, siehe Abbildung 2.5.

□

Definition 2.53 Iljin[3]–Verfahren, Iljin–Allen[4]–Southwell[5]–Verfahren. Das Verfahren

$$\begin{aligned} -\frac{h}{2} b_i \coth\left(\frac{h}{2\varepsilon} b_i\right) D^+ D^- u_i + b_i D^0 u_i + c_i u_i &= f_i, \quad \text{für } i = 1, \ldots, N-1, \\ u_0 = u_N &= 0, \end{aligned}$$

wird Iljin–Verfahren oder Iljin–Allen–Southwell–Verfahren genannt. □

Satz 2.54 Gleichmäßige Konvergenz des Iljin–Allen–Southwell–Verfahrens. *Das Iljin–Allen–Southwell–Verfahren konvergiert auf* $[0,1]$ *gleichmäßig von erster Ordnung in der diskreten Maximumsnorm, das heißt*

$$\|u(x_i) - u_i\|_{\infty,d} \leq Ch$$

mit einer von ε *und* h *unabhängigen Konstanten* C*.*

Beweis: Der Beweis ist relativ rechenaufwändig, deshalb wird auf die Literatur, [RST08], verwiesen. ■

Beispiel 2.55 Betrachte

$$-\varepsilon u'' + u' = 1 \quad \text{auf } (0,1), \quad u(0) = u(1) = 0,$$

mit der Lösung

$$u(x) = x - \frac{\exp\left(-\frac{1-x}{\varepsilon}\right) - \exp\left(-\frac{1}{\varepsilon}\right)}{1 - \exp\left(-\frac{1}{\varepsilon}\right)}.$$

Für den Fehler in der diskreten Maximumsnorm im Fall $\varepsilon = 10^{-3}$ erhält man

[3] A.M. Iljin
[4] D.N. de G. Allen
[5] Richard V. Southwell

Intervalle	zentr. Diff.	einf. Upwind	IAS–Verfahren
2	124.5	0.00199	0
4	31.004	0.00398	0
8	7.715	0.00793	0
16	2.0235	0.01574	0
32	0.91132	0.03100	2.2204e-16
64	0.77305	0.06015	1.5543e-15
128	0.59276	0.11307	8.3598e-15
256	0.34287	0.18371	1.2388e-14
512	0.12997	0.19679	1.0976e-14
1024	0.03277	0.12933	5.8457e-14
2048	0.00750	0.07486	1.5675e-13
4096	0.00183	0.04076	2.8882e-13

Man sieht, dass das Iljin–Allen–Southwell–Verfahren immer die genauesten Ergebnisse liefert. Sind die Knoten hinreichend entfernt von der Grenzschicht, dann sind die Ergebnisse in den Knoten sogar exakt. □

2.4.2 Grenzschichtangepasste Gitter

Bemerkung 2.56 Motivation. Es wurde bereits gezeigt, dass die Lösung von singulär gestörten Problemen aus zwei Bestandteilen besteht:

- der Lösung des reduzierten Problems, diese ist im allgemeinen glatt und einfach zu approximieren,
- einem Korrekturterm, der das Erfülltsein der Randbedingung am Ausflussrand erzwingt. Dieser ist dafür verantwortlich, dass die Grenzschicht auftritt, dass sich die Lösung in einem sehr kleinen Intervall dramatisch verändert.

Betrachte als typisches Beispiel das 2–Punkt–Randwertproblem aus Beispiel 2.55. Im Intervall $[0, 1-\varepsilon]$ hat die Lösung praktisch die Gestalt $u(x) = x$, ist also sehr einfach auf einem groben Gitter zu approximieren. Der interessante Bestandteil der Lösung ist im Intervall $[1-\varepsilon, 1]$. Wählt man ein äquidistantes Gitter der Schrittweite h, dann gilt im allgemeinen $h > \varepsilon$ und das Intervall $[1-\varepsilon, 1]$ ist in $[x_{N-1}, x_N] = [1-h, 1]$ enthalten. Man kann nicht erwarten, damit das Verhalten der Lösung in $[1-\varepsilon, 1]$ zu auflösen zu können.

Die Idee von grenzschichtangepassten Gittern besteht darin, in der Grenzschicht ein (wesentlich) feineres Gitter zu wählen als außerhalb derselben. Damit besteht die Möglichkeit, die Lösung in der Grenzschicht gut zu approximieren. □

Bemerkung 2.57 Shishkin[6]–Gitter. Betrachte der einfacheren Notation halber ein Problem, bei welchem die Grenzschicht sich bei $x = 0$ befindet. Desweiteren sei $b = -\beta \in \mathbb{R}^+$ eine Konstante. Nun werden die Gitterpunkte gemäß

$$x_i = \phi(i/N),$$

verteilt, wobei die Funktion $\phi(\xi)$ so gewählt werden muss, dass bei $x = 0$ ein hinreichend feines Gitter entsteht. Die Anzahl N der Intervalle ist vorgegeben. Ein Gitter von Shishkin–Typ ist gegeben durch

$$\phi(\xi) = \begin{cases} \dfrac{\sigma\varepsilon}{\beta}\hat{\phi}(\xi) \text{ mit } \hat{\phi}(1/2) = \ln(N) & \text{für } \xi \in [0, 1/2], \\ 1 - 2\left(1 - \dfrac{\sigma\varepsilon}{\beta}\ln(N)\right)(1-\xi) & \text{für } \xi \in [1/2, 1], \end{cases}$$

[6] Grigory I. Shishkin

wobei $\sigma > 0$ ein Parameter ist. Das Shishkin–Gitter (1988) erhält man für

$$\hat{\phi}(\xi) = 2\ln(N)\xi.$$

Damit hat man für die Gitterpunkte $x_0, \ldots, x_{N/2}$

$$x_i - x_{i-1} = \phi\left(\frac{i}{N}\right) - \phi\left(\frac{i-1}{N}\right) = \frac{\sigma\varepsilon}{\beta} 2\ln(N)\left(\frac{i}{N} - \frac{i-1}{N}\right) = 2\frac{\sigma\varepsilon}{\beta}\frac{\ln(N)}{N}$$

unabhängig von i. Für die Gitterpunkte $x_{N/2+1}, \ldots, x_N$ gilt

$$\begin{aligned} x_i - x_{i-1} &= \phi\left(\frac{i}{N}\right) - \phi\left(\frac{i-1}{N}\right) \\ &= 1 - 2\left(1 - \frac{\sigma\varepsilon}{\beta}\ln(N)\right)\left(1 - \frac{i}{N}\right) - 1 + 2\left(1 - \frac{\sigma\varepsilon}{\beta}\ln(N)\right)\left(1 - \frac{i-1}{N}\right) \\ &= \frac{2}{N} - 2\frac{\sigma\varepsilon}{\beta}\frac{\ln(N)}{N}, \end{aligned}$$

unabhängig von i. Dies ist ein stückweise äquidistantes Gitter. Der Übergangspunkt vom sehr feinen auf das grobe Gitter ist bei

$$\tau = x_{N/2} = \frac{\sigma\varepsilon}{\beta}\ln(N).$$

□

Die Wahl des Shishkin–Gitters wird mit dem folgenden Satz gerechtfertigt.

Satz 2.58 Konvergenz des einfachen Upwind–Verfahrens auf einem Shishkin–Gitter. *Betrachte das einfache Upwind–Verfahren auf einem Shishkin–Gitter mit dem Übergangspunkt*

$$\tau = \min\left\{\frac{1}{2}, \frac{\varepsilon}{\beta}\ln(N)\right\},$$

also mit $\sigma = 1$. Dann gilt die Fehlerabschätzung

$$\|u(x_i) - u_i\|_{\infty,d} \leq CN^{-1}\ln(N),$$

mit einer von ε und N unabhängigen Konstanten C.

Beweis: Der Beweis basiert auf der Zerlegung der Lösung in den Anteil vom reduzierten Problem (glatter Anteil) und den Korrekturterm. Er ist relativ aufwändig, siehe [RST08]. ■

Bemerkung 2.59

- Die Konvergenz ist wegen des Faktors $\ln(N)$ leicht suboptimal. Man sieht aber in numerischen Beispielen, dass die obige Abschätzung scharf ist, dass dieser Faktor also nicht entfallen kann.
- Die Idee der Verwendung grenzschichtangepasster Gitter geht bereits auf Bachvalov[7] (1969) zurück. Bei Bachvalov–Gittern gibt es einen glatten Übergang vom feinem zum groben Gittern. Numerische Verfahren sind auf Bachvalov–Gittern schwieriger zu analysieren als auf Shishkin–Gittern.
- Die a priori (vor der numerischen Lösung) Konstruktion geeigneter grenzschichtangepasster Gitter erfordert im wesentlichen die Kenntnis der Lösung. Dies ist in der Praxis vollkommen unrealistisch, insbesondere bei Problemen in zwei oder drei Dimensionen. Man benötigt vielmehr eine a posteriori (während der numerischen Lösung) Konstruktion von angepassten Gittern. Auch dazu gibt es Wege.

[7] Nikolai Sergejewitsch Bachvalov (1934 – 2005)

- Die wesentliche Erkenntnis der Analysis von Verfahren auf a priori grenzschichtangepassten Gittern besteht darin, dass gezeigt wird, dass man auf einem geeigneten Gitter ein einfaches Verfahren verwenden kann und damit vernünftige Fehlerabschätzungen erhält.
- Bei der Nutzung von Shishkin–Gittern muss man jetzt Differenzenquotienten im Knoten $x_{N/2}$ erklären, für welchen die anliegenden Intervalle nicht gleich lang sind.
  Sei x_i ein Knoten und haben die Intervalle $[x_{i-1}, x_i]$ und $[x_i, x_{i+1}]$ die Längen h_i und h_{i+1}. Für den Rückwärts– und Vorwärtsquotienten ändert sich nichts zur Definition 2.4, da man dort immer nur eines der anliegenden Intervalle braucht. Ansonsten definiert man

  $$\tilde{h_i} := \frac{h_i + h_{i+1}}{2}.$$

  Der zentrale Differenzenquotient ist das gewichtete Mittel

  $$D^0 v(x_i) = \frac{1}{2\tilde{h_i}} \left(h_i D^+ v(x_i) + h_{i+1} D^- v(x_i) \right)$$

  Diese Approximation ist von zweiter Ordnung konsistent. Die zweite Ableitung wird wie folgt approximiert

  $$v''(x_i) \approx \delta^2 v_i := \frac{1}{\tilde{h_i}} \left(D^+ v(x_i) - D^- v(x_i) \right) = \frac{1}{\tilde{h_i}} \left(\frac{v_{i+1} - v_i}{h_{i+1}} - \frac{v_i - v_{i-1}}{h_i} \right).$$

  Diese Approximation ist nicht mehr von zweiter Ordnung konsistent. *Übungsaufgabe*
- Die Matrizen, die man bei der Nutzung von grenzschichtangepassten Gittern erhält, sind sehr schlecht konditioniert.

□

Beispiel 2.60 Betrachte wieder

$$-\varepsilon u'' + u' = 1 \quad \text{auf } (0,1), \quad u(0) = u(1) = 0,$$

mit der Lösung

$$u(x) = x - \frac{\exp\left(-\frac{1-x}{\varepsilon}\right) - \exp\left(-\frac{1}{\varepsilon}\right)}{1 - \exp\left(-\frac{1}{\varepsilon}\right)}.$$

Die Grenzschicht ist in diesem Beispiel bei $x = 1$. Deshalb wählt man den Übergangspunkt hier

$$\tau = x_{N/2} = 1 - \frac{\sigma\varepsilon}{\beta} \ln(N) = 1 - \sigma\varepsilon \ln(N).$$

Die Fehler in der diskreten Maximumsnorm für $\varepsilon = 10^{-6}$ und $\sigma = 2$ sind

Intervalle	$\|u - u_h\|_{\infty,d}$
4	0.25584
8	0.16455
16	0.10833
32	0.069125
64	0.043656
128	0.026335
256	0.015402
512	0.0087902
1024	0.0049257
2048	0.0027225
4096	0.0014891

Man stellt fest, das die Ergebnisse stark von der Wahl von σ abhängen, *Übungsaufgabe*. □

Bemerkung 2.61 Fazit. Gleichmäßig konvergente Verfahren kann man auf zwei Arten bekommen:

- Verwendung eines geeignet modifizierten Verfahrens auf einem einfachen Gitter,
- Verwendung eines einfachen Verfahrens auf einem geeignet gewählten Gitter.

□

Kapitel 3

Schwache Lösungstheorie

Bemerkung 3.1 Motivation. Dieses Kapitel stellt eine Erweiterung des Lösungsbegriffes von partiellen Differentialgleichungen vor – die schwache Lösung. Diese Erweiterung ist aus folgenden Gründen notwendig:

- Man kann im allgemeinen nicht erwarten, dass eine partielle Differentialgleichung eine klassische Lösung besitzt. Dazu müssen die Parameterfunktionen hinreichend oft differenzierbar sein und in höheren Dimensionen muss auch das Gebiet einige Forderungen erfüllen. Zum Beispiel darf es keine Ecke besitzen. Diese Forderungen sind aber in der Natur oder in Anwendungen oft nicht erfüllt. Trotzdem laufen die durch die partielle Differentialgleichung beschriebenen Prozesse ab und es gibt offensichtlich eine Lösung. Nur wird diese bestimmte (Differenzierbarkeits–)Eigenschaften der klassischen Lösung nicht besitzen und man benötigt einen erweiterten Lösungsbegriff.
- Die im Kapitel 4 vorgestellte Finite–Element–Methode beruht auf der schwachen Formulierung der zu Grunde liegenden Gleichung.

Die grundlegende Idee bei der Abschwächung des Lösungsbegriffes besteht in der Formel der partiellen Integration. Betrachte das Zwei–Punkt–Randwertproblem

$$-u''(x) = f(x), \quad x \in (0,1), \quad u(0) = u(1) = 0.$$

Multiplikation mit einer gewissen Funktion $v(x)$, mit $v(0) = v(1) = 0$, der so genannten Testfunktion, Integration über $(0,1)$ und anschließende partielle Integration führen auf

$$\begin{aligned} \int_0^1 -u''(x)v(x)\ dx &= -u'(1)v(1) + u'(0)v(0) + \int_0^1 u'(x)v'(x)\ dx \\ &= \int_0^1 u'(x)v'(x)\ dx = \int_0^1 f(x)v(x)\ dx. \end{aligned}$$

Um dieser Gleichung einen Sinn zu geben, benötigt man nur noch, dass die Produkte $u'(x)v'(x)$ und $f(x)v(x)$ integrierbar sind. Es wird nicht einmal die Existenz der zweiten Ableitung von $u(x)$ verlangt. Natürlich muss geklärt werden, welche Eigenschaften geeignete Testfunktionen besitzen müssen. Der schwache Lösungsbegriff wird verlangen, dass die obige Integralgleichung für alle geeigneten Testfunktionen erfüllt ist. □

3.1 Funktionenräume

Bereits aus Bemerkung 3.1 wird deutlich, dass man Räume von Funktionen benötigt, die geeignet integrierbar sind.

3.1.1 Lebesgue–Räume

Definition 3.2 Lebesgue[1]–Räume. Die Lebesgue–Räume oder Räume der Lebesgue–messbaren Funktionen $L^p(a,b)$ sind die Räume aller Funktionen, für die gilt

$$v \in L^p(a,b) \iff \int_a^b |v(x)|^p \; dx < \infty, \quad \text{für } p \in [1,\infty),$$

$$v \in L^\infty(a,b) \iff \operatorname{ess\,sup}_{x\in(a,b)} |v(x)| := \left(\inf_{\mu(N)=0} \sup_{(a,b)\setminus N} |v(x)| \right) < \infty.$$

Hierbei ist N eine beliebige Menge vom (Lebesgue–)Maß Null, $\mu(N) = 0$ und ess sup wird wesentliches Supremum genannt. Der Raum $L^\infty(a,b)$ wird auch Raum der wesentlich beschränkten Funktionen genannt. □

Bemerkung 3.3 Zu den Lebesgue–Räumen.

- Die Integrale sind im Lebesgue–Sinn zu verstehen. Die Definition des Lebesgue–Integrals beruht auf sogenannten einfachen Funktionen. Das sind nicht–negative messbare (siehe Vorlesung über Maßtheorie) Funktionen, welche nur endlich viele Funktionswerte annehmen dürfen. Treppenfunktionen, wie man sie bei der Definition des Riemann–Integrals verwendet, sind eine Teilmenge der Menge der einfachen Funktionen. Im Gegensatz zu Treppenfunktionen, dürfen einfache Funktionen die endlich vielen Funktionswerte jedoch in unendlich vielen verschiedenen Intervallen annehmen. Die bekannteste einfache Funktion, die keine Treppenfunktion ist, ist die Dirichletsche Funktion

  $$g : [0,1] \to \{0,1\}, \quad g(x) = \begin{cases} 1 & x \in \mathbb{Q}, \\ 0 & \text{sonst.} \end{cases}$$

  Diese Funktion ist nicht Riemann–integrierbar, aber Lebesgue–integrierbar mit dem Integralwert Null. Das Lebesgue–Integral für eine beliebige Funktion wird dann wie üblich über einen Grenzprozess definiert. Da die Treppenfunktionen eine Teilmenge der einfachen Funktionen sind ist klar, dass das Lebesgue–Integral allgemeiner als das Riemann–Integral ist. Es gilt, dass jede Riemann–integrierbare Funktion auch Lebesgue–integrierbar ist und der Integralwert ist derselbe.
- Die Funktionen aus $L^p(a,b)$ sind nur bis auf eine Menge vom (Lebesgue–)Maß Null eindeutig bestimmt. In diesem Sinne ist eine Funktion $v(x)$ eigentlich eine Äquivalenzklasse aller Funktionen, die sich von $v(x)$ nur auf einer Menge vom Maß Null unterscheiden. Man sagt, dass sie fast überall gleich sind oder für fast alle $x \in [a,b]$ übereinstimmen. Aus jeder Äquivalenzklasse kann man immer einen entsprechenden Vertreter wählen.
  Ein einfacher Vertreter aus der Klasse der Dirichletschen Funktion $g(x)$ ist die Funktion $v(x) = 0$ für alle $x \in [0,1]$.
- Da ein Punkt eine Menge vom Maß Null ist, macht es im allgemeinen keinen Sinn, nach dem Funktionswert einer Funktion $u \in L^p(a,b)$ in einem bestimmten Punkt $x \in [a,b]$ zu fragen. Das geht nur, wenn es in der Äquivalenzklasse von $u(x)$ einen stetigen Repräsentanten gibt. Falls $u(x)$ bestimmte Eigenschaften besitzt, ist dies erfüllt, siehe beispielsweise Bemerkung 3.22.
- Die Räume $L^p(a,b)$, $p \in [1,\infty)$, werden Banach[2]–Räume (vollständige normierte Räume) mit der Norm

  $$\|v\|_p := \left(\int_a^b |v(x)|^p \; dx \right)^{1/p}.$$

[1] Henri Lebesgue (1875 – 1941)
[2] Stefan Banach (1892 – 1945)

- Der Raum $L^\infty(a,b)$ wird zu einem Banach–Raum mit der Norm

$$\|v\|_\infty := \operatorname{ess}\sup_{x\in(a,b)} |v(x)|\,.$$

  Gibt es einen stetigen Vertreter aus der Äquivalenzklasse von $v(x)$, der sich auch stetig auf den Rand fortsetzen lässt, man sagt dann wie üblich $v \in C([a,b])$, dann ist die Normdefinition äquivalent zu

$$\|v\|_\infty := \max_{x\in[a,b]} |v(x)|\,.$$

- Mit dem Skalarprodukt

$$(u,v) := \int_a^b u(x)v(x)\ dx$$

  ist der $L^2(a,b)$ ein Hilbert–Raum. *Nachweis Skalarprodukt: Übungsaufgabe*

□

Beispiel 3.4 Betrachte $u(x) = 1/\sqrt{x}$ auf $(0,1)$. Dann ist für $p \neq 2$

$$\begin{aligned}\int_0^1 \left(\frac{1}{\sqrt{x}}\right)^p\ dx &= \int_0^1 \left(\frac{1}{x}\right)^{p/2}\ dx = \int_0^1 x^{-p/2}\ dx \\ &= \frac{1}{1-p/2}\left(1^{1-p/2} - \lim_{x\to 0} x^{1-p/2}\right).\end{aligned}$$

Für $p < 2$ ist $1 - p/2 > 0$ und der Grenzwert ist Null. Somit existiert das Integral. Für $p > 2$ ist $1 - p/2 < 0$, der Grenzwert exisiert nicht und somit auch das Integral. Für $p = 2$ rechnet man direkt nach, dass das Integral auch divergiert. Die Funktion gehört auch nicht zu $L^\infty(0,1)$, da sie für $x \to 0$ unbeschränkt wächst. Also ist $u(x) \in L^p(0,1)$ für $p \in [1,2)$. □

Definition 3.5 $L^1_{\text{loc}}(a,b)$**.** Der Raum $L^1_{\text{loc}}(a,b)$ ist der Raum der (Äquivalenzklassen von) auf jeder kompakten Teilmenge von (a,b) Lebesgue–integrierbaren Funktionen. Es gilt also $u \in L^1_{\text{loc}}(a,b)$ genau dann, wenn $u \in L^1(a',b')$ für alle Intervalle $[a',b'] \subset (a,b)$. □

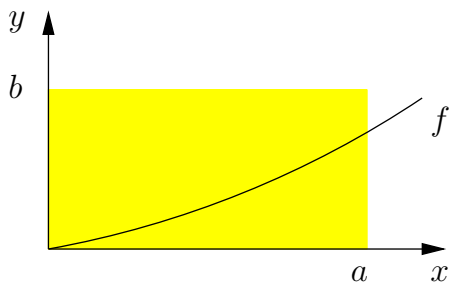
Beispiel 3.6 Die Funktion $u(x) = 1/x$ gehört zu $L^1_{\text{loc}}(0,1)$ aber nicht zu $L^1(0,1)$. □

Das nächste Ziel ist der Beweis einer wichtigen Ungleichung in Lebesgue–Räumen.

Lemma 3.7 *Sei* $f : \mathbb{R}^+ \cup \{0\} =: \mathbb{R}_0^+ \to \mathbb{R}$ *eine stetige und streng monoton wachsende Funktion mit* $f(0) = 0$ *und* $f(x) \to \infty$ *für* $x \to \infty$*. Dann gilt für alle* $a, b \in \mathbb{R}^+ \cup \{0\}$

$$ab \le \int_0^a f(x)\ dx + \int_0^b f^{-1}(y)\ dy.$$

Beweis:



Das Intervall $(0, a)$ wird auf der x–Achse abgetragen und das Intervall $(0, b)$ auf der y–Achse. Dann sind ab der Flächeninhalt des zugehörigen Rechtecks, $\int_0^a f(x)\,dx$ die Fläche unterhalb der Kurve und $\int_0^b f^{-1}(y)\,dy$ die Fläche zwischen der positiven y–Achse und der Kurve. Damit ist die Ungleichung bewiesen. Gleichheit tritt genau dann auf, wenn $f(a) = b$ gilt. ■

Beispiel 3.8 Young[3]sche Ungleichung. Die Youngsche Ungleichung

$$ab \leq \frac{\varepsilon}{2}a^2 + \frac{1}{2\varepsilon}b^2 \quad \forall\ a, b \in \mathbb{R}_0^+,\ \varepsilon \in \mathbb{R}^+$$

erhält man aus diesem Lemma mit $f(x) = \varepsilon x$, $f^{-1}(y) = \varepsilon^{-1}y$. Sie lässt sich auch direkt mit der Binomischen Formel beweisen. Zum Beweis der verallgemeinerten Youngschen Ungleichung

$$ab \leq \frac{\varepsilon^p}{p}a^p + \frac{1}{q\varepsilon^q}b^q, \quad \forall\ a, b \in \mathbb{R}_0^+, \varepsilon \in \mathbb{R}^+$$

mit $p^{-1} + q^{-1} = 1, p, q \in (1, \infty)$ wählt man $f(x) = x^{p-1}$, $f^{-1}(y) = y^{1/(p-1)}$ und wendet das obige Lemma auf die Intervalle mit den Grenzen εa und $\varepsilon^{-1}b$ an. □

Beispiel 3.9 Cauchy[4]–Schwarz[5]–Ungleichung. Die Cauchy–Schwarz–Ungleichung

$$|(\mathbf{x}, \mathbf{y})| \leq \|\mathbf{x}\|_2 \|\mathbf{y}\|_2 \ \forall\ \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$$

kann man mit Hilfe der Youngschen Ungleichung beweisen. Dazu stellt man zunächst fest, dass die Cauchy–Schwarz–Ungleichung richtig ist, falls einer der beiden Vektoren verschwindet. Seien $\mathbf{x}, \mathbf{y}$ mit $\|\mathbf{x}\|_2 = \|\mathbf{y}\|_2 = 1$. Man erhält aus der Youngschen Ungleichung

$$|(\mathbf{x}, \mathbf{y})| = \left|\sum_{i=1}^n x_i y_i\right| \leq \sum_{i=1}^n |x_i|\,|y_i| \leq \frac{1}{2}\sum_{i=1}^n |x_i|^2 + \frac{1}{2}\sum_{i=1}^n |y_i|^2 = 1.$$

Damit gilt die Cauchy–Schwarz–Ungleichung für $\mathbf{x}, \mathbf{y}$. Sind $\tilde{\mathbf{x}} \neq \mathbf{0}, \tilde{\mathbf{y}} \neq \mathbf{0}$ beliebig, nutzt man die Homogenität der Cauchy–Schwarz–Ungleichung aus. Aus der Gültigkeit der Cauchy–Schwarz–Ungleichung für $\mathbf{x}$ und $\mathbf{y}$ folgt durch Skalierung

$$\left|\left(\underbrace{\|\tilde{\mathbf{x}}\|_2^{-1}\tilde{\mathbf{x}}}_{\mathbf{x}}, \underbrace{\|\tilde{\mathbf{y}}\|_2^{-1}\tilde{\mathbf{y}}}_{\mathbf{y}}\right)\right| \leq 1$$

Die beiden Vektoren $\mathbf{x}, \mathbf{y}$ haben die Norm 1. Also

$$\frac{1}{\|\tilde{\mathbf{x}}\|_2 \|\tilde{\mathbf{y}}\|_2}\,|(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})| \leq 1.$$

Das war zu beweisen.

Die verallgemeinerte Cauchy–Schwarz–Ungleichung

$$|(\mathbf{x}, \mathbf{y})| \leq \left(\sum_{i=1}^n |x_i|^p\right)^{1/p} \left(\sum_{i=1}^n |y_i|^q\right)^{1/q}$$

mit $p^{-1} + q^{-1} = 1, p, q \in (1, \infty)$ beweist man auf dem gleichen Wege mit Hilfe der verallgemeinerten Youngschen Ungleichung. □

[3] Young
[4] Augustin Louis Cauchy (1789 – 1857)
[5] Hermann Amandus Schwarz (1843 – 1921)

Satz 3.10 ***Hölder[6]sche Ungleichung.*** *Sei* $p^{-1} + q^{-1} = 1, p, q \in (1, \infty)$. *Wenn* $u \in L^p((a,b))$ *und* $v \in L^q((a,b))$, *dann ist* $uv \in L^1((a,b))$ *und es gilt*

$$\|uv\|_1 \leq \|u\|_p \|v\|_q .$$

Für $p = q = 2$ *wird dies auch Cauchy–Schwarz–Ungleichung genannt.*

Beweis: Man muss zunächst zeigen, dass $|u(x)v(x)|$ durch eine integrierbare Funktion abgeschätzt werden kann. Man setzt in der verallgemeinerten Youngschen Ungleichung $\varepsilon = 1$, $a = |u(x)|$ und $b = |v(x)|$. Dann folgt

$$|u(x)v(x)| \leq \frac{1}{p} |u(x)|^p + \frac{1}{q} |v(x)|^q .$$

Da die rechte Seite dieser Ungleichung nach Voraussetzung integrierbar ist, ist $uv \in L^1((a,b))$ gezeigt. Auch die Höldersche Ungleichung ist bereits für den Fall $\|u\|_p = \|v\|_q = 1$ durch diese Ungleichung bewiesen

$$\int_(a,b) |u(x)v(x)| \; dx \leq \frac{1}{p} \int_(a,b) |u(x)|^p \; dx + \frac{1}{q} \int_(a,b) |v(x)|^q \; dx = 1.$$

Die allgemeine Ungleichung folgt nun durch ein Homogenitätsargument wie bei der Cauchy–Schwarz–Ungleichung für den Fall dass beide Funktionen nicht fast überall verschwinden. Im Fall, dass eine Funktion fast überall verschwindet, ist die Ungleichung trivialerweise erfüllt. ■

3.1.2 Verallgemeinerte Ableitung und Sobolev–Räume

Definition 3.11 $C_0^\infty(a,b)$**.** Der Raum $C_0^\infty(a,b) \subset C^\infty(a,b)$ ist der Raum der auf (a,b) beliebig oft differenzierbaren Funktionen $\varphi(x)$, deren Träger

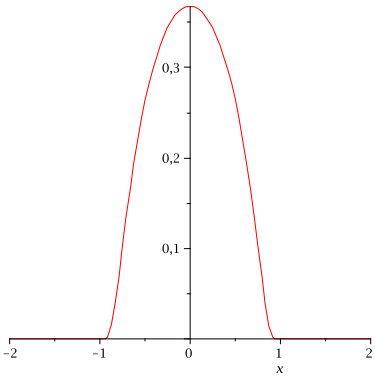
$$\text{supp } \varphi := \overline{\{x \in (a,b) \; : \; \varphi(x) \neq 0\}}$$

beschränkt und in (a,b) enthalten ist. Man sagt auch, $\varphi(x)$ hat einen kompakten Träger. □

Beispiel 3.12 Die Funktion

$$\varphi(x) = \begin{cases} 0 & \text{für } |x| \geq 1, \\ \exp\left(-\dfrac{1}{1-x^2}\right) & \text{für } |x| < 1, \end{cases}$$

ist beliebig in $\mathbb{R}$ differenzierbar *Übungsaufgabe* mit supp $\varphi = [-1, 1]$.



□

[6]Otto Ludwig Hölder (1859 – 1937)

Definition 3.13 Verallgemeinerte oder schwache Ableitung. Seien $u, v \in L^1_{\text{loc}}(a,b)$ und gelte für alle $\varphi \in C_0^\infty(a,b)$

$$\int_a^b u(x)\varphi'(x)\ dx = -\int_a^b v(x)\varphi(x)\ dx.$$

Dann heißt $v(x)$ verallgemeinerte oder schwache Ableitung von $u(x)$, im Symbol $v(x) = u'(x)$. □

Bemerkung 3.14

- In der Mathematik wird die Bezeichnung, dass etwas schwach gilt, im allgemeinen in dem Sinne gebraucht, dass eine Eigenschaft für alle geeigneten Testfunktionen erfüllt ist.
- Für im klassischen Sinne differenzierbare Funktionen stimmt die verallgemeinerte Ableitung mit der klassischen Ableitung überein.
- Für verallgemeinerte Ableitungen gelten die üblichen Differentiationsregeln. *Übungsaufgabe*

□

Beispiel 3.15 Betrachte $u(x) = |x|$ auf $(-1,1)$. Dann ist die verallgemeinerte Ableitung durch

$$u'(x) = \begin{cases} -1 & \text{für } x \in (-1,0), \\ 1 & \text{für } x \in (0,1) \end{cases}$$

gegeben. Man beachte, dass die Wahl des Funktionswertes von $u'(x)$ in $x = 0$ keine Rolle spielt, denn gesucht ist eine Funktion aus $L^1_{\text{loc}}(-1,1)$, genauer also eine Klasse von Funktionen, die fast überall auf $(-1,1)$ übereinstimmen.

Die Funktion $u'(x)$ besitzt keine schwache Ableitung. Das bedeutet, die schwache Ableitung muss nicht notwendig existieren. Man kann allerdings den Ableitungsbegriff noch weiter verallgemeinern, so dass jede (verallgemeinerte) Funktion unendlich oft in diesem Sinne differenzierbar ist. □

Lemma 3.16 Fundamentallemma der Variationsrechnung. *Sei $u \in L^1_{\text{loc}}(a,b)$ und gelte für alle $\varphi \in C_0^\infty(a,b)$*

$$\int_a^b u(x)\varphi(x)\ dx = 0.$$

Dann folgt $u(x) = 0$ für fast alle $x \in [a,b]$.

Beweis: Der Beweis erfordert eine Reihe technischer Vorbereitungen. Das sprengt den Rahmen dieser Vorlesung, deshalb sei auf [Emm04, p.61 ff.] verwiesen. ■

Folgerung 3.17 *Sei $u \in L^1_{\text{loc}}(a,b)$ und gelte für alle $\varphi \in C_0^\infty(a,b)$*

$$\int_a^b u(x)\varphi'(x)\ dx = 0.$$

Dann gibt es eine reelle Konstante C, so dass $u(x) = C$ für fast alle $x \in [a,b]$ gilt.

Beweis: Übungsaufgabe. ■

Lemma 3.18 *Sei $u \in L^1_{\text{loc}}(a,b)$. Dann ist die schwache Ableitung $u'(x)$ (als Äquivalenzklasse von Funktionen, die fast überall gleich sind) eindeutig bestimmt.*

Beweis: Indirekter Beweis. Sei $v(x)$ neben $u'(x)$ eine weitere schwache Ableitung. Dann gilt für alle $\varphi \in C_0^\infty(a,b)$

$$\begin{aligned}\int_a^b \big(u'(x) - v(x)\big)\,\varphi(x)\;dx &= \int_a^b u'(x)\varphi(x)\;dx - \int_a^b v(x)\varphi(x)\;dx \\ &= -\int_a^b u(x)\varphi'(x)\;dx + \int_a^b u(x)\varphi'(x)\;dx = 0.\end{aligned}$$

Dann muss aber nach dem Fundamentallemma der Variationsrechnung $u'(x) = v(x)$ für fast alle $x \in [a,b]$ gelten. ■

Bemerkung 3.19

- Sei $u \in L^1(a,b)$ und gelte $u' \in L^1(a,b)$, so kann man zeigen, dass es in der Äquivalenzklasse von $u(x)$ einen stetigen (sogar absolut stetigen) Repräsentanten auf $[a,b]$ gibt. In diesem Falle macht es also auch Sinn, der Funktion $u(x)$ Funktionswerte in Punkten aus $[a,b]$ zuzuweisen.
  Eine reelle Funktion $g(x)$ auf $[a,b]$ heißt absolut stetig, falls zu jedem $\varepsilon > 0$ ein $\delta > 0$ existiert, so dass für jedes endliche System disjunkter Teilintervalle (a_k, b_k), $k = 1, \ldots, n$, mit der Gesamtlänge $\sum_{k=1}^n (b_k - a_k) < \delta$ gilt $\sum_{k=1}^n |g(b_k) - g(a_k)| < \varepsilon$. Jede absolut stetige Funktion ist stetig ($n = 1$), nicht aber umgekehrt. Lipschitz–stetige Funktionen sind absolut stetig.
- Höhere verallgemeinerte Ableitungen können analog zu Definition 3.13 mittels der Forderung

  $$\int_a^b u(x)\varphi^{(n)}(x)\;dx = (-1)^n \int_a^b u(x)^{(n)}\varphi(x)\;dx$$

  für alle $\varphi \in C_0^\infty(a,b)$ definiert werden.

□

Definition 3.20 Sobolev[7]–Raum. Der Sobolev–Raum $W^{k,p}(a,b)$ mit $p \in [1,\infty]$ ist definiert durch

$$W^{k,p}(a,b) := \Big\{ u \in L^p(a,b) \;:\; u^{(i)} \in L^p(a,b),\; i = 1, \ldots, k \Big\}.$$

Oft wird $H^k(a,b)$ anstelle von $W^{k,2}(a,b)$ geschrieben. □

Lemma 3.21 *Auf $H^1(a,b)$ sind durch*

$$\|v\|_{1,2} := \left((v,v)\right)_{1,2}^{1/2}, \quad \left((u,v)\right)_{1,2} := \int_a^b \left(u(x)v(x) + u'(x)v'(x)\right)\;dx$$

eine Norm und ein Skalarprodukt definiert. Mit dieser Norm und diesem Skalarprodukt ist $H^1(a,b)$ ein Hilbert–Raum. Durch

$$|v|_{1,2} := \left(\int_a^b \left(v'(x)\right)^2\;dx\right)^{1/2}$$

ist eine Halbnorm definiert.

Beweis: Norm, Skalarprodukt und Halbnorm sind Übungsaufgaben. Für den Hilbert–Raum fehlt noch die Vollständigkeit, siehe dafür Literatur. ■

[7] Sergej Lwowitsch Sobolev (1908 – 1989)

Bemerkung 3.22 Wichtige Aussagen zum $H^1(a,b)$.

1. $H^1(a,b)$ besitzt eine abzählbare Basis. Diese Eigenschaft nennt man separabel. Also ist $H^1(a,b)$ ein separabler Hilbert–Raum. Das ist eine sehr reichhaltige Struktur.
2. Jede Funktion $u \in H^1(a,b)$ ist fast überall gleich einer (absolut) stetigen Funktion. Das bedeutet, man kann einen stetigen Repräsentanten auswählen. Es gibt eine Konstante $C > 0$, die nur vom Intervall (a,b) abhängt, so dass für alle $u \in C([a,b])$
$$\|u\|_{C([a,b])} \leq C \|u\|_{1,2}$$
gilt. Solch eine Eigenschaft und zugehörige Ungleichung gilt in höheren Dimensionen nicht mehr.
3. Der Raum $C^\infty([a,b])$ liegt dicht in $H^1(a,b)$, das heißt für jedes $u \in H^1(a,b)$ und jedes $\varepsilon > 0$ gibt es ein $\varphi \in C^\infty([a,b])$ mit
$$\|u - \varphi\|_{1,2} \leq \varepsilon.$$

Die Beweise findet man in der Literatur. □

Definition 3.23 $H_0^1(a,b)$**.** Der Raum $H_0^1(a,b)$ ist definiert durch
$$H_0^1(a,b) := \left\{ v \in H^1(a,b) \ : \ v(a) = v(b) = 0 \right\}.$$

□

Bemerkung 3.24 In einer Dimension macht die obige Definition Sinn, denn nach Bemerkung 3.22 besitzt jede Äquivalenzklasse in $H^1(a,b)$ einen stetigen Repräsentanten. In höheren Dimensionen muss man die Randwerte noch geeignet erklären. □

Satz 3.25 Poincaré[8]–Friedrichs[9]–Ungleichung. *Ist $u \in H_0^1(a,b)$, so gilt*
$$\|u\|_2 \leq \frac{b-a}{\sqrt{2}} |u|_{1,2}.$$

Beweis: Sei $u(x)$ ein absolut stetiger Repräsentant. Dann gilt wegen $u(a) = 0$
$$u(x) = \int_a^x u'(\xi) \, d\xi, \quad x \in (a,b),$$
wobei $u'(x)$ die verallgemeinerte Ableitung bezeichnet. Mit der Cauchy–Schwarz–Ungleichung folgt
$$|u(x)|^2 = \left(\int_a^x u'(\xi) \, d\xi \right)^2 \leq \left(\int_a^x 1 \, d\xi \right) \left(\int_a^x |u'(\xi)|^2 \, d\xi \right) \leq (x-a) \, |u|_{1,2}^2 .$$
Integration über (a,b) gibt
$$\|u\|_2^2 = \int_a^b |u(x)|^2 \, dx \leq \int_a^b (x-a) \, dx \ |u|_{1,2}^2 = \frac{(b-a)^2}{2} |u|_{1,2}^2 .$$

■

Bemerkung 3.26 Zum $H_0^1(a,b)$ und zur Poincaré–Friedrichs–Ungleichung.

[8] Jules Henri Poincaré (1854 – 1912)
[9] Kurt Otto Friedrichs (1901 – 1982)

- Die Poincaré–Friedrichs–Konstante C_{PF} kann man noch etwas verbessern. Die Proportionalität zu $b-a$ bleibt jedoch erhalten.
- Im Beweis sieht man, dass es ausreicht, wenn $u(x)$ in einem Randpunkt den Wert Null annimmt.
- Auf $H_0^1(a,b)$ werden durch $|\cdot|_{1,2}$ eine Norm und durch

$$(u,v)_{1,2} := \int_a^b u'(x)v'(x)\ dx$$

ein Skalarprodukt definiert. Damit ist $H_0^1(a,b)$ ein separabler Hilbert–Raum.
- Die Poincaré–Friedrichs–Ungleichung zeigt die Äquivalenz der Normen $\|\cdot\|_{1,2}$ und $|\cdot|_{1,2}$ auf $H_0^1(a,b)$

$$|u|_{1,2} \le \|u\|_{1,2} = \left(\|u\|_{0,2}^2 + |u|_{1,2}^2\right)^{1/2} \le \left(C_{\mathrm{PF}}^2\,|u|_{1,2}^2 + |u|_{1,2}^2\right)^{1/2} = C\,|u|_{1,2}\,.$$

- Es gilt für $u \in H_0^1(a,b)$

$$\|u\|_{C([a,b])} \le \sqrt{b-a}\,|u|_{1,2}\,.$$

- Der Raum $C_0^\infty(a,b)$ liegt dicht in $H_0^1(a,b)$.
- Der Raum $H_0^1(a,b)$ liegt dicht in $L^2(a,b)$.

□

Definition 3.27 Dualraum $H^{-1}(a,b)$. Es bezeichne $H^{-1}(a,b)$ den Raum aller stetigen linearen Funktionale auf $H_0^1(a,b)$, das heißt den Raum aller stetigen linearen Abbildungen $H_0^1(a,b) \to \mathbb{R}$. □

Bemerkung 3.28 Zum Dualraum.

- Zu jedem $f \in H^{-1}(a,b)$ gibt es ein nicht eindeutig bestimmtes $u_f \in L^2(a,b)$, so dass

$$\begin{aligned} \langle f, v\rangle \quad &= \quad -\int_a^b u_f(x)v'(x)\ dx \quad \text{für alle } v \in H_0^1(a,b), \\ &\overset{\text{formal}}{=} \quad \int_a^b u_f'(x)v(x)\ dx = \int_a^b f(x)v(x)\ dx. \end{aligned}$$

  Die Definition $\langle f, v\rangle$ bedeutet, dass man falls der eine Faktor aus $H_0^1(a,b)$ ist, einen zweiten Faktor wählen kann, der nicht einmal in $L^2(a,b)$ liegen muss, damit das Integral noch einen Sinn ergibt. Der Raum $H^{-1}(a,b)$ beschreibt konkret, was *nicht einmal in $L^2(a,b)$ liegen* bedeutet.
- Auf $H^{-1}(a,b)$ wird durch

$$\|f\|_{-1,2} := \sup_{v \in H_0^1(a,b)} \frac{|\langle f, v\rangle|}{|v|_{1,2}}$$

  eine Norm definiert. Mit dieser Norm ist $H^{-1}(a,b)$ ein Banach–Raum. Diese Norm ist im allgemeinen praktisch nicht berechenbar.
- Es gilt

$$H_0^1(a,b) \subset L^2(a,b) \subset H^{-1}(a,b).$$

  Das ist der sogenannte Gelfand[10]–Dreier.

[10] Israel Moissejewitsch Gelfand, geb. 1913

- Ein lineares Funktional aus dem $H^{-1}(a,b)$ heißt beschränkt, wenn es ein $c \in \mathbb{R}$ gibt, so dass
$$\langle f, v\rangle \leq c\,|v|_{1,2}$$
für alle $v \in H_0^1(a,b)$ gilt. Ein lineares Funktional heißt stetig, wenn aus $v_n(x) \to v(x)$ in $H_0^1(a,b)$ folgt $\langle f, v_n\rangle \to \langle f, v\rangle$ in $\mathbb{R}$. Man kann zeigen: Ein lineares Funktional ist genau dann stetig, wenn es beschränkt ist.
- Oft schreibt man $f(v)$ anstelle von $\langle f, v\rangle$. Der duale Raum eines Hilbert–Raumes V wird im allgemeinen mit V' bezeichnet.

□

3.2 Variationelle Formulierung

Bemerkung 3.29 Herleitung der variationellen oder schwachen Formulierung. Betrachte das Zwei–Punkt–Randwertproblem

$$-\varepsilon u''(x) + b(x)u'(x) + c(x)u(x) = f(x), \quad \text{für } x \in (0,1), \quad u(0) = u(1) = 0. \tag{3.1}$$

Multiplikation der Differentialgleichung mit einer geeigneten Funktion $v(x)$, mit $v(0) = v(1) = 0$, Integration der resultieren Gleichung über $(0,1)$ und anschließende partielle Integration ergibt

$$\begin{aligned}
&\int_0^1 \Big(-\varepsilon u''(x) + b(x)u'(x) + c(x)u(x)\Big) v(x)\; dx \\
&= \quad -\varepsilon u'(1)v(1) + \varepsilon u'(0)v(0) \\
&\qquad + \int_0^1 \Big(\varepsilon u'(x)v'(x) + b(x)u'(x)v(x) + c(x)u(x)v(x)\Big)\; dx \\
&= \quad \int_0^1 \Big(\varepsilon u'(x)v'(x) + b(x)u'(x)v(x) + c(x)u(x)v(x)\Big)\; dx \\
&= \quad \int_0^1 f(x)v(x)\; dx.
\end{aligned}$$

Man hat also von der höchsten Ableitung von $u(x)$ eine Ableitung auf die Funktion $v(x)$ übertragen. Damit die obige Schreibweise Sinn macht, müssen die Funktionen natürlich so beschaffen sein, dass die Integrale wohldefiniert sind. □

Definition 3.30 Variationelle oder schwache Formulierung. Seien $b, c \in L^\infty(0,1)$ und $f \in H^{-1}(0,1)$. Die schwache Formulierung des Zwei–Punkt–Randwertproblems (3.1) lautet: Finde $u \in H_0^1(0,1)$, so dass für alle $v \in H_0^1(0,1)$

$$\int_0^1 \Big(\varepsilon u'(x)v'(x) + b(x)u'(x)v(x) + c(x)u(x)v(x)\Big)\; dx = \int_0^1 f(x)v(x)\; dx \tag{3.2}$$

gilt. Die Lösung nennt man schwache oder verallgemeinerte Lösung. Der Raum in dem die Lösung gesucht wird heißt Lösungs– oder Ansatzraum. Die Funktionen $v(x)$ heißen Testfunktionen und der Raum aus dem sie stammen Testraum. □

Bemerkung 3.31 Zur schwachen Formulierung.

- Die Voraussetzungen sind so, dass alle Integrale wohldefiniert sind.
- Im Gegensatz zur klassischen Lösung muss die schwache Lösung nur noch einmal differenzierbar sein, und das auch nur schwach.
- Jede klassische Lösung ist auch schwache Lösung. Die Umkehrung gilt jedoch nur bei hinreichend glatten Koeffizienten und rechter Seite.

□

Beispiel 3.32 Betrachte ein Zwei–Punkt–Randwertproblem der Form (3.1) in $(-1,1)$ mit $\varepsilon = 1$, $b(x) = c(x) = 0$ für alle $x \in (-1,1)$,

$$f(x) = \begin{cases} 2 & \text{für } x < 0, \\ -2 & \text{für } x \geq 0 \end{cases}$$

und $u(-1) = u(1) = 0$. Die rechte Seite ist nicht stetig, deshalb kann dieses Problem keine klassische Lösung besitzen.

Dieses Problem kann als Modell der Wärmeleitung in einem eindimensionalen Stab der Länge Zwei aufgefasst werden. In $[-1,0)$ wird der Stab erhitzt, in $(0,1]$ abgekühlt. Was im Punkt $x = 0$ passiert ist für die schwache Formulierung unwichtig. Gesucht ist die Temperatur $u(x)$. An den Stabenden ist die Temperatur jeweils Null. Die Wärmeleitung findet nur durch Diffusion (Molekularbewegung) statt.

Das Problem besitzt die schwache Lösung

$$u(x) = \begin{cases} -x^2 - x & \text{für } x < 0, \\ x^2 - x & \text{für } x \geq 0, \end{cases}$$

sie Abbildung 3.1.

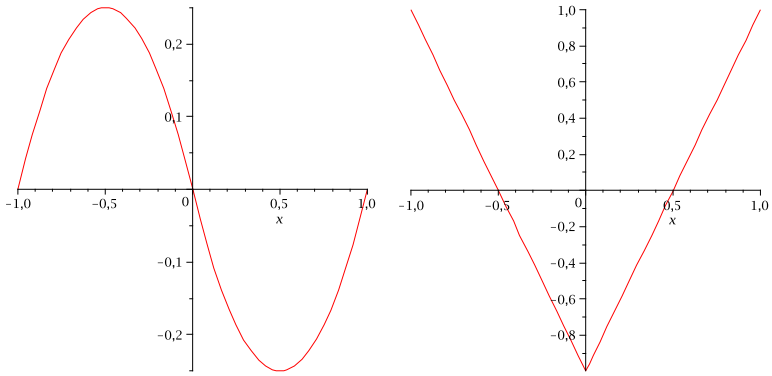


Abbildung 3.1: Beispiel 3.32: schwache Lösung und ihre Ableitung.

Die erste Ableitung von $u(x)$ ist stetig

$$u'(x) = \begin{cases} -2x - 1 & \text{für } x < 0, \\ 2x - 1 & \text{für } x \geq 0. \end{cases}$$

Die variationelle Formulierung: Finde $u \in H_0^1(-1,1)$, so dass

$$\int_{-1}^{1} u'(x)v(x)\ dx = \int_{-1}^{1} f(x)v(x)\ dx$$

für alle $v \in H_0^1(-1,1)$ ist also wohldefiniert. Es gilt für alle $v \in H_0^1(-1,1)$

$$\begin{aligned} \int_{-1}^{1} u'(x)v'(x)\ dx &= \int_{-1}^{0} u'(x)v'(x)\ dx + \int_{0}^{1} u'(x)v'(x)\ dx \\ &= \int_{-1}^{0} -u''(x)v(x)\ dx + \lim_{x\to 0-} u'(x)v(x) - u'(-1)v(-1) \\ &\quad + \int_{0}^{1} -u''(x)v(x)\ dx + u'(1)v(1) - \lim_{x\to 0+} u'(x)v(x) \\ &= \int_{-1}^{0} 2v(x)\ dx + \int_{0}^{1} (-2)v(x)\ dx = \int_{-1}^{1} f(x)v(x)\ dx. \end{aligned}$$

Die Terme an den Randpunkten verschwinden, weil $v(x)$ dort verschwindet. Die Terme im Punkt $x = 0$ sind gleich, weil $u'(x)$ und $v(x)$ stetig sind. □

Bemerkung 3.33 Andere Randbedingungen.

- Betrachte zunächst inhomogene Dirichlet–Bedingungen

$$u(0) = a, \quad u(1) = b.$$

Dann sieht die schwache Formulierung genauso aus wie (3.2), auch der Testraum bleibt $H_0^1(0,1)$, aber der Ansatzraum ändert sich zu

$$\begin{aligned} V_a &:= \left\{ v \in H^1(0,1) \ : \ v = g + w, \ w \in H_0^1(0,1) \right\} \\ &= \left\{ v \in H^1(0,1) \ : \ v(0) = a, \ v(1) = b \right\}, \end{aligned}$$

wobei $g \in H^1(0,1)$ eine beliebige, aber fixierte Funktion mit $g(0) = a, g(1) = b$ ist. Dirichlet–Bedingungen nennt man auch wesentliche Randbedingungen, da sie entscheidend in die Definition des Ansatzraumes eingehen.

- Dagegen können Neumann–Randbedingungen in natürlicher Weise eingearbeitet werden, weshalb sie auch natürliche Randbedingungen genannt werden. Seien $\varepsilon u'(0) = \alpha$, $\varepsilon u'(1) = \beta$, dann sucht man ein $u \in H^1(0,1)$, so dass

$$\begin{aligned} &\int_0^1 \Big(\varepsilon u'(x)v'(x) + b(x)u'(x)v(x) + c(x)u(x)v(x) \Big) \, dx \\ &= \int_0^1 f(x)v(x) \, dx - \beta v(1) + \alpha v(0) \end{aligned}$$

für alle $v \in H^1(0,1)$ erfüllt ist. Die Randterme fallen bei der partiellen Integration der klassischen Formulierung nicht weg.

□

Definition 3.34 Bilinearform. Sei $(V, \|\cdot\|)$ ein Banach–Raum. Eine Abbildung $a : V \times V \to \mathbb{R}$ heißt

1. bilinear, falls $a(\cdot,\cdot)$ in jedem Argument linear ist,
2. symmetrisch, falls $a(u,v) = a(v,u)$ für alle $u, v \in V$ gilt,
3. positiv, falls $a(v,v) \geq 0$ für alle $v \in V$ gilt,
4. stark positiv oder koerzitiv oder V–elliptisch oder positiv definit, falls es ein $\mu > 0$ gibt, so dass $a(v,v) \geq \mu \|v\|^2$ für alle $v \in V$ gilt.
5. Eine Bilinearform heißt beschränkt, falls es ein $\beta > 0$ gibt, so dass

$$|a(u,v)| \leq \beta \|u\| \, \|v\|$$

für alle $u, v \in V$ gilt.

□

Beispiel 3.35 Betrachte das Zwei–Punkt–Randwertproblem mit homogenen Dirichlet–Randbedingungen.

- Dann ist

$$a(u,v) := \int_0^1 \Big(\varepsilon u'(x)v'(x) + b(x)u'(x)v(x) + c(x)u(x)v(x) \Big) \, dx \qquad (3.3)$$

eine Bilinearform auf $V = H_0^1(0,1)$. Das folgt direkt aus der Linearität der Integration und der Linearität der Differentiation.

- Ist $b(x) = 0$ für alle $x \in (0,1)$, dann ist $a(u,v)$ symmetrisch.
- Seien $b \in C^1([0,1])$ und $c \in C([0,1])$. Es ist

$$\begin{aligned}\frac{1}{2}\int_0^1 b(x)v'(x)v(x)\ dx &= -\frac{1}{2}\int_0^1 (b(x)v(x))'v(x)\ dx \\ &= -\frac{1}{2}\int_0^1 b'(x)v(x)v(x)\ dx - \frac{1}{2}\int_0^1 b(x)v'(x)v(x)\ dx \\ \Longrightarrow \quad \int_0^1 b(x)v'(x)v(x)\ dx &= -\frac{1}{2}\int_0^1 b'(x)v(x)v(x)\ dx.\end{aligned}$$

  Einsetzen in (3.3) mit $u(x) = v(x)$ ergibt

$$a(v,v) = \int_0^1 \left(\varepsilon\left(v'(x)\right)^2 + \left(-\frac{b'(x)}{2} + c(x)\right)\left(v(x)\right)^2\right)\ dx$$

  Falls $-b'(x)/2 + c(x) \geq 0$ für alle $x \in [0,1]$ ist, folgt für alle $v \in H_0^1(0,1)$

$$a(v,v) \geq \varepsilon\,|v|_{1,2}^2$$

  und $a(\cdot,\cdot)$ ist koerzitiv, da nach Bemerkung 3.26 $|\cdot|_{1,2}$ eine Norm in $H_0^1(0,1)$ ist.
- Seien $b, c \in L^\infty(0,1)$. Mit Hilfe der Cauchy–Schwarz–Ungleichung und der Poincaré–Friedrichs–Ungleichung folgt

$$\begin{aligned}|a(u,v)| &\leq \varepsilon\,\|u'\|_0\,\|v'\|_0 + \|b\|_\infty\,\|u'\|_0\,\|v\|_0 + \|c\|_\infty\,\|u\|_0\,\|v\|_0 \\ &\leq \varepsilon\,\|u'\|_0\,\|v'\|_0 + C_{PF}\,\|b\|_\infty\,\|u'\|_0\,\|v'\|_0 + C_{PF}^2\,\|c\|_\infty\,\|u'\|_0\,\|v'\|_0 \\ &= C\,\|u'\|_0\,\|v'\|_0 = C\,|u|_{1,2}\,|v|_{1,2}\,.\end{aligned}$$

  Somit ist die Bilinearform beschränkt. *Übungsaufgabe, Bsp. S.94*

□

Satz 3.36 ***Riesz***[11]***scher Darstellungssatz.*** *Sei V ein Hilbert–Raum mit dem Skalarprodukt $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ und der Norm $\|v\|_V = a(v,v)^{1/2}$. Zu jedem stetigen linearen Funktional $f \in V'$ gibt es ein eindeutig bestimmtes $u \in V$ mit*

$$a(u,v) = f(v) \quad \forall\ v \in V.$$

Desweiteren ist $u(x)$ die eindeutig bestimmte Lösung des Variationsproblems

$$\min_{v\in V} F(v) := \min_{v\in V}\left(\frac{1}{2}a(v,v) - f(v)\right).$$

Beweis: Als erstes wird die Existenz einer Lösung $u(x)$ des Variationsproblems gezeigt. Wegen der Stetigkeit von f gilt die Abschätzung

$$|f(v)| \leq c\,\|v\|_V \quad \forall\ v \in V$$

und daher

$$F(v) \geq \frac{1}{2}\,\|v\|_V^2 - c\,\|v\|_V\,.$$

Der rechte Ausdruck ist eine nach oben geöffnete Parabel in $\|v\|_V$. Somit ist dieser Ausdruck nach unten beschränkt und mit dem notwendigen Kriterium für ein Minimum erhält man $0 = \|v\|_V - c$. Einsetzen ergibt

$$F(v) \geq -\frac{1}{2}c^2.$$

[11] Frigyes Riesz (1880 – 1956)

Da das Funktional F nach unten beschränkt ist, existiert

$$d = \inf_{v \in V} F(v)$$

Sei $\{v_k\}_{k\in\mathbb{N}}$ eine Minimalfolge, d.h, $F(v_k) \to d$ für $k \to \infty$. Im Hilbert–Raum gilt *Übungsaufgabe*

$$\|v_k - v_l\|_V^2 + \|v_k + v_l\|_V^2 = 2\,\|v_k\|_V^2 + 2\,\|v_l\|_V^2\,.$$

Es folgt, unter Nutzung der Linearität von f,

$$\begin{aligned}
&\|v_k - v_l\|_V^2 \\
&= 2\,\|v_k\|_V^2 + 2\,\|v_l\|_V^2 - 4\left\|\frac{v_k + v_l}{2}\right\|_V^2 - 4f(v_k) - 4f(v_l) + 8f\left(\frac{v_k + v_l}{2}\right) \\
&= 4F(v_k) + 4F(v_l) - 8F\left(\frac{v_k + v_l}{2}\right) \\
&\leq 4F(v_k) + 4F(v_l) - 8d \to 0
\end{aligned}$$

für $k, l \to \infty$. Damit ist $\{v_k\}_{k\in\mathbb{N}}$ eine Cauchy–Folge, die wegen der Vollständigkeit von V einen Grenzwert $u \in V$ besitzt. Da F stetig ist, ist $F(u) = d$ und $u(x)$ ist die Lösung des Variationsproblems.

Im nächsten Schritt wird gezeigt, dass jede Lösung des Variationsproblems auch eine Lösung der Gleichung ist. Es ist, unter Nutzung der Bilinearität und Symmetrie,

$$\begin{aligned}
\Phi(\varepsilon) &= F(u + \varepsilon v) = \frac{1}{2}a(u + \varepsilon v, u + \varepsilon v) - f(u + \varepsilon v) \\
&= \frac{1}{2}a(u, u) + \varepsilon a(u, v) + \frac{\varepsilon^2}{2}a(v, v) - f(u) - \varepsilon f(v)
\end{aligned}$$

für alle $v \in V$. Wenn $u(x)$ das Variationsproblem minimiert, dann besitzt die Funktion $\Phi(\varepsilon)$ an der Stelle $\varepsilon = 0$ ein Minimum. Das notwendige Kriterium führt auf die Bedingung

$$0 = \Phi'(0) = a(u, v) - f(v) \quad \text{für alle } v \in V.$$

Zum Schluss wird die Eindeutigkeit der Lösung gezeigt. Seien $u_1(x)$ und $u_2(x)$ zwei Lösungen der Gleichung. Aus der Differenz der beiden Gleichungen erhält man

$$a(u_1 - u_2, v) = 0 \quad \text{für alle } v \in V.$$

Diese Beziehung gilt speziell für $v(x) = (u_1 - u_2)(x)$ woraus $u_1(x) = u_2(x)$ folgt. Die Lösung des Variationsproblems ist eindeutig auf Grund der Eindeutigkeit der Lösung der Gleichung. ■

Beispiel 3.37 Seien $V = H_0^1(0,1)$ ausgestattet mit dem Skalarprodukt $a(u,v) = \int_0^1 u'(x)v'(x)\,dx$ und $V' = H^{-1}(0,1)$. Es ist $a(v,v)^{1/2} = |v|_{1,2}$ eine Norm in $H_0^1(0,1)$. Aus dem Rieszschen Darstellungssatz folgt, dass

$$a(u,v) = \int_0^1 u'(x)v'(x)\;dx = \int_0^1 f(x)v(x)\;dx$$

für jedes $f \in H^{-1}(0,1)$ eine eindeutige Lösung besitzt. Desweiteren minimiert diese Lösung das sogenannte Energiefunktional

$$\min_{v\in H_0^1(0,1)} F(v) := \min_{v\in H_0^1(0,1)} \left(\int_0^1 \frac{1}{2}\left(v'(x)\right)^2 - f(x)v(x)\;dx\right).$$

Den Ausdruck unter dem Integral kann man physikalisch als Energie interpretieren. □

Der Satz von Riesz kann wie folgt verallgemeinert werden.

Satz 3.38 ***Lemma von Lax*[12]*–Milgram*[13]*.*** *Sei $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ eine beschränkte und positiv definite Bilinearform auf dem Hilbert–Raum V. Zu jedem beschränkten linearen Funktional $f \in V'$ gibt es genau ein $u \in V$ mit*

$$a(u,v) = f(v) \quad \textit{für alle } v \in V. \tag{3.4}$$

Beweis: Der Beweis erfolgt mit Hilfe des Satzes von Riesz. Dazu werden gewisse Operatoren definiert. Auf die Einführung dieses Kalküls wurde aus Zeitgründen verzichtet.

Für Details des Beweises, siehe Literatur. ■

Folgerung 3.39 Lösung des schwachen Problems (3.3). *Seien $V = H_0^1(0,1)$, $f \in V'$, $b, b', c \in L^\infty(0,1)$ und gelte*

$$c(x) - \frac{b'(x)}{2} \geq 0 \quad \textit{fast überall in } (0,1).$$

Dann besitzt (3.3) genau eine Lösung.

Beweis: Der Beweis folgt mit dem Lemma von Lax–Milgram und den bereits bewiesenen Eigenschaften der Bilinearform $a(\cdot,\cdot)$. ■

Satz 3.40 Regularitätsaussage. *Seien die Voraussetzungen von Folgerung 3.39 erfüllt und gelte zusätzlich für ein $k \in \mathbb{N} \setminus \{0\}$, dass $b, c \in C^{k-1}([0,1])$ sowie $f, f', \ldots, f^{k-1} \in L^2(0,1)$. Dann gilt für die schwache Lösung von (3.3) $u, u', \ldots, u^{k+1} \in L^2(0,1)$.*

Beweis: Siehe Literatur, zum Beispiel [GT83]. ■

[12] Peter Lax, geb. 1926
[13] Arthur Norton Milgram

Kapitel 4

Finite–Elemente–Methoden (FEM)

4.1 Das Ritzsche Verfahren

Bemerkung 4.1 Grundidee von Finite–Elemente–Methoden, das Ritz[1]sche Verfahren. Sei V ein Hilbert–Raum mit dem Skalarprodukt $a(\cdot,\cdot)$. Wir betrachten das Problem

$$\min_{v\in V} F(v) = \min_{v\in V}\left(\frac{1}{2}a(v,v) - f(v)\right),$$

wobei $f(\cdot) : V \to \mathbb{R}$ ein beschränktes lineares Funktional ist. Wie bereits bewiesen ist, besitzt das Variationsproblem eine eindeutig bestimmte Lösung $u \in V$, die außerdem die Gleichung

$$a(u,v) = f(v) \quad \forall\, v \in V \tag{4.1}$$

löst, Satz 3.36 (Rieszscher Darstellungssatz).

Um die Lösung der obigen Probleme mit einem numerischen Verfahren zu approximieren, setzen wir voraus, dass V ein separabler Hilbert–Raum ist, das heißt V besitzt eine abzählbare Basis. Dann gibt es endlich–dimensionale Teilräume $V_1, V_2, \ldots \subset V$ mit $\dim V_k = k$, die folgende Eigenschaft besitzen: zu jedem $u \in V$ und $\varepsilon > 0$ gibt es ein $K \in \mathbb{N}$ und ein $u_k \in V_k$ mit

$$\|u - u_k\|_V \leq \varepsilon \quad \forall\, k \geq K.$$

Es wird dabei nicht verlangt, dass es eine Inklusion der Form $V_k \subset V_{k+1}$ gibt.

Die Ritz–Approximation von (4.1) ist wie folgt definiert. Gesucht ist $u_k \in V_k$ mit

$$a(u_k, v_k) = f(v_k) \quad \forall\, v_k \in V_k. \tag{4.2}$$

Die wesentliche Idee des Ritzschen Verfahrens besteht also darin, dass man den unendlich–dimensionalen Raum V durch einen endlich–dimensionalen Raum V_k ersetzt. □

Lemma 4.2 Eigenschaften der Ritzschen Approximation.

1. *Der Fehler ist orthogonal zum Raum V_k, das heißt es gilt*

$$a(u - u_k, v_k) = 0 \quad \forall\, v_k \in V_k. \tag{4.3}$$

2. *u_k ist die Bestapproximierende von u in V_k bezüglich der von $a(\cdot,\cdot)$ induzierten Norm.*

[1] Walter Ritz (1878 – 1909)

3. *Die Folge der Ritz–Approximierenden konvergiert gegen die Lösung von (4.1), das heißt $u_k \to u$ für $k \to \infty$.*

Beweis: Da endlich–dimensionale Teilräume von Hilbert–Räumen wiederum Hilbert–Räume sind, besitzt nach dem Rieszschen Darstellungssatz auch die Gleichung der Ritz–Approximation eine eindeutige Lösung, die ebenso ein Minimierungsproblem im Raum V_k löst. Aus der Differenz der Gleichungen (4.1) und (4.2) erhält man die Orthogonalitätsrelation

$$a(u - u_k, v_k) = 0 \quad \forall\, v_k \in V_k.$$

Das besagt, dass der Fehler $u - u_k$ senkrecht zum Raum V_k ist: $u - u_k \perp V_k$. Demnach ist u_k die orthogonale Projektion von u in den Raum V_k bezüglich des Skalarproduktes von V. Das heißt, u_k ist die Bestapproximierende von u in V_k

$$\|u - u_k\|_V = \inf_{v_k \in V_k} \|u - v_k\|_V .$$

Zum Beweis nutzt man die Orthogonalität (4.3) und die Cauchy–Schwarz–Ungleichung. Sei $w_k \in V_k$ beliebig, dann ist

$$\begin{aligned} \|u - u_k\|_V^2 &= a(u - u_k, u - u_k) = a(u - u_k, u - \underbrace{(u_k - w_k)}_{v_k}) = a(u - u_k, u - v_k) \\ &\leq \|u - u_k\|_V \|u - v_k\|_V . \end{aligned}$$

Da $w_k \in V_k$ beliebig ist, ist auch $v_k \in V_k$ beliebig.

Mit der Bestapproximationseigenschaft erhält man

$$\|u - u_k\|_V = \inf_{v_k \in V_k} \|u - v_k\|_V \leq \varepsilon,$$

woraus schließlich die Konvergenz der Ritz–Approximation $u_k \to u$ für $k \to \infty$ folgt. ■

Bemerkung 4.3 Formulierung als lineares Gleichungssystem. Für die Berechnung der u_k kann man eine beliebige Basis $\{\phi_i\}_{i=1}^k$ von V_k verwenden. Zunächst gilt, dass die Gleichung der Ritz–Approximation (4.2) genau dann für alle $v_k \in V_k$ erfüllt ist, wenn sie für jede Basisfunktion ϕ_i erfüllt ist. Das folgt aus der Linearität der Gleichung bezüglich der Testfunktion und daraus, dass man jede Funktion $v_k \in V_k$ als Linearkombination der Basisfunktionen darstellen kann. Man setzt auch die Lösung als Linearkombination der Basisfunktionen an

$$u_k = \sum_{j=1}^{k} u^j \phi_j$$

mit unbekannten Koeffizienten $\mathbf{u} = (u^1, \ldots, u^k)^T$ und erhält, indem man als Testfunktionen jetzt die Basisfunktionen nutzt,

$$\sum_{j=1}^{k} a(u^j \phi_j, \phi_i) = f(\phi_i), \quad i = 1, \ldots, k.$$

Das ist äquivalent zu einem Gleichungssystem $A\mathbf{u} = \mathbf{b}$, wobei

$$A = (a_{ij}) = a(\phi_j, \phi_i)$$

Steifigkeitsmatrix genannt wird. Man beachte die unterschiedliche Reihenfolge der Indizes bei den Matrixeinträgen und beim Skalarprodukt. Die rechte Seite ist ein Vektor der Länge k mit den Einträgen $b_i = f(\phi_i)$.

Mit der eineindeutigen Zuordnung zwischen dem Koordinatenvektor $(u^1, \ldots, u^k)^T$ und dem Element $u_k = \sum_{i=1}^k u^i \phi_i$ lässt sich zeigen, dass die Matrix A symmetrisch und positiv definit ist:

$$\begin{aligned} A = A^T &\iff a(v, w) = a(w, v)\ \forall v, w \in V_k, \\ x^T A x > 0 \text{ für } x \neq 0 &\iff a(v, v) > 0\ \forall v \in V_k, v \neq 0. \end{aligned}$$

Übungsaufgabe □

Bemerkung 4.4 Der Fall einer unsymmetrischen Bilinearform. Im nichtvariationellen Fall, also wenn $b(\cdot,\cdot)$ unsymmetrisch, aber äquivalent zum Skalarprodukt $a(\cdot,\cdot)$ ist, kann man: Finde $u \in V$ mit

$$b(u,v) = f(v) \quad \forall\, v \in V \tag{4.4}$$

auch mit dem Ritzschen Verfahren approximieren. Die Eigenschaften von $b(\cdot,\cdot)$ seien Beschränktheit

$$|b(u,v)| \le M \|u\|_V \|v\|_V \quad M \in \mathbb{R},$$

und Koerzitivität

$$m \|v\|_V^2 \le b(v,v), \quad m > 0.$$

Das diskrete Problem lautet: Finde $u_k \in V_k$, so dass

$$b(u_k, v_k) = f(v_k) \quad \forall\, v_k \in V_k. \tag{4.5}$$

Die diskrete Lösung existiert eindeutig nach Satz 3.38 (Lax–Milgram). Sie ist jedoch keine orthogonale Projektion in V_k mehr. Trotzdem kann man die gleiche Fehlerabschätzung wie im variationellen Fall beweisen. □

Lemma 4.5 Lemma von Cea[2]. *Sei die Bilinearform $b(\cdot,\cdot)$ beschränkt und koerzitiv. Dann gilt*

$$\|u - u_k\|_V \le \frac{M}{m} \inf_{v_k \in V_k} \|u - v_k\|_V . \tag{4.6}$$

Beweis: Aus der Differenz der stetigen Gleichung (4.4) und der diskreten Gleichung (4.5)

$$b(u - u_k, v_k) = 0 \quad \forall\, v_k \in V_k$$

und

$$m \|v\|_V^2 \le b(v,v) \text{ und } |b(u,v)| \le M \|u\|_V \|v\|_V$$

folgt sofort

$$\begin{aligned} \|u - u_k\|_V^2 &\le \frac{1}{m} b(u - u_k, u - u_k) = \frac{1}{m} b(u - u_k, u - v_k) \\ &\le \frac{M}{m} \|u - u_k\|_V \|u - v_k\|_V . \end{aligned}$$

■

Bemerkung 4.6 Galerkin[3]–Methode. Im unsymmetrischen Fall wird dieses Verfahren Galerkin–Methode genannt. Das lineare Gleichungssystem wird genauso wie im symmetrischen Fall hergeleitet. Betrachte dazu das Zwei–Punkt–Randwertproblem

$$-\varepsilon u''(x) + b(x)u'(x) + c(x)u(x) = f(x), \quad \text{für } x \in (0,1), \quad u(0) = u(1) = 0.$$

Die schwache Formulierung lautet: Finde $u \in H_0^1(0,1)$, so dass für alle $v \in H_0^1(0,1)$

$$\int_0^1 \Big(\varepsilon u'(x)v'(x) + b(x)u'(x)v(x) + c(x)u(x)v(x) \Big)\, dx = \int_0^1 f(x)v(x)\, dx$$

gilt. Falls $(\cdot,\cdot)$ das Skalarprodukt in $L^2(0,1)$ bezeichnet, kann die schwache Formulierung übersichtlicher geschrieben werden

$$b(u,v) := \varepsilon(u', v') + (bu', v) + (cu, v) = (f, v).$$

[2] Cea

[3] Boris Grigorievich Galerkin (1871 – 1945)

Sei $\{\phi_i\}_{i=1}^k$ eine beliebige Basis von V_k, dann macht man wieder den Ansatz

$$u_k = \sum_{j=1}^{k} u^j \phi_j$$

mit unbekannten Koeffizienten $\mathbf{u} = (u^1, \ldots, u^k)^T$. Auch im nichtsymmetrischen Fall ist die variationelle Formulierung genau dann erfüllt, wenn sie für alle Basisfunktionen erfüllt ist. Man erhält

$$\sum_{j=1}^{k} \Big[\varepsilon(\phi_j', \phi_i') + (b\phi_j', \phi_i) + (c\phi_j, \phi_i)\Big] u^j = (f, \phi_i), \quad i = 1, \ldots, k,$$

was äquivalent zu einem Gleichungssystem $A\mathbf{u} = \mathbf{b}$ ist. Die Einträge der Steifigkeitsmatrix sind

$$a_{ij} = \varepsilon(\phi_j', \phi_i') + (b\phi_j', \phi_i) + (c\phi_j, \phi_i).$$

Die Systemmatrix ist nicht mehr symmetrisch.

Die Eigenschaften der Bilinearform wurden im Beispiel 3.35 untersucht. Sind $b, c \in L^\infty(0,1)$, so ist die Bilinearform beschränkt und die Konstante M ist in der Größenordnung von $\max\{\|b\|_\infty, \|c\|_\infty\}$. Gilt $-b'(x)/2 + c(x) \geq 0$, so ist sie koerzitiv mit $m = \varepsilon$. Falls beide Bedingungen erfüllt sind, dann ist das Lemma von Cea anwendbar und für den Fehler gilt

$$\|u - u_k\|_{H_0^1} \leq C \frac{\max\{\|b\|_\infty, \|c\|_\infty\}}{\varepsilon} \inf_{v_k \in V_k} \|u - v_k\|_{H_0^1}, \quad C \in \mathbb{R}.$$

Im singulär gestörten Fall $\varepsilon \ll \|b\|_\infty$ ist der erste Faktor in dieser Fehlerabschätzung sehr groß. □

4.2 Finite–Element–Räume in 1D

Bemerkung 4.7 Motivation für die Wahl der Räume beim Ritzschen Verfahren und der Galerkin–Methode. Der wichtigste Punkt beim Ritzschen Verfahren und bei der Galerkin–Methode ist die Wahl der Räume V_k, oder genauer, die Wahl von geeigneten Basen $\{\phi_i\}_{i=1}^k$, die einen Raum V_k aufspannen. In dieser Vorlesung wird nur der Fall betrachtet, dass $V_k \subset V$ gilt. Es gibt auch Finite–Elemente–Methoden, bei denen diese Eigenschaft nicht erfüllt ist.

Vom numerischen Standpunkt aus sollten die Elemente a_{ij} der Steifigkeitsmatrix schnell zu berechnen sein und die Matrix A sollte nur schwach besetzt sein, das heißt sie sollte viele Nulleinträge besitzen. Das führt auf folgende Überlegungen:

- Die Einträge von A berechnen sich mit Hilfe von Integralen, welche die Ansatz– und Testfunktionen, sowie deren Ableitungen enthalten. Funktionen, für die sich solche Integrale besonders einfach berechnen lassen, sind Polynome.
- Falls man Basisfunktionen $\{\phi_i(x)\}_{i=1}^k$ wählt, die im gesamten Intervall $(0,1)$ ungleich Null sein können, so werden im allgemeinen nur sehr wenige Integrale verschwinden und nur wenige Einträge der Matrix A werden Null. Deshalb ist es zweckmäßig Funktionen zu verwenden, die nur auf einem kleinen Teil von $(0,1)$ nicht Null sind.
- Wegen $V_k \subset V (= H_0^1(a,b))$, müssen die Funktionen aus V_k stetig sein, vergleiche Bemerkung 3.22.

Aus diesen Gründen bietet es sich an, als Basis stetige Funktionen zu verwenden, die stückweise polynomial sind.

Analog zu den Finite–Differenzen–Verfahren wird $[0,1]$ mittels eines (zunächst) äquidistanten Gitters mit den Gitterpunkten

$$x_i = ih, \quad i = 0, \ldots, N, \quad h = 1/N,$$

zerlegt. Die Intervalle $K_i = (x_i, x_{i+1})$ werden Gitterzellen genannt. Ihre Vereinigung

$$\mathcal{T}_h = \bigcup_{i=0}^{N-1} \overline{K_i}$$

heißt Triangulierung. □

Bemerkung 4.8 Träger von Finite–Elemente–Funktionen. Ein Kriterium zur Konstruktion von geeigneten Basisfunktionen für Finite–Elemente ist, dass ihr Träger, siehe Definition 3.11, möglichst klein sein soll. Wenn das der Fall ist, dann ist es sehr wahrscheinlich, dass der gemeinsame Träger von zwei verschiedenen Basisfunktionen $\phi_i(x), \phi_j(x)$ das Maß Null hat, zum Beispiel leer ist oder nur ein Punkt ist. In diesem Fall sind alle Integrale für die betreffenden Komponenten a_{ij} und a_{ji} gleich Null, man hat also Nulleinträge in der Matrix.

In 1D muss es Basisfunktionen $\phi_i(x)$ geben, deren Träger aus zwei benachbarten Gitterzellen $[x_{i-1}, x_i] \cup [x_i, x_{i+1}]$ besteht. Ansonsten müssten wegen der Stetigkeit alle Basisfunktionen in den Gitterpunkten $x_0, \ldots, x_N$ verschwinden, womit man nur Funktionen approximieren könnte, die diese Eigenschaft haben. Das wird aber im allgemeinen für die Lösung einer Differentialgleichung nicht gelten. □

Beispiel 4.9 Stückweise lineare Basisfunktionen. Die einfachsten Basisfunktionen besitzen einen Träger aus zwei benachbarten Gitterzellen und sie sind stückweise linear. Sie sind eindeutig durch ihre Werte in den Gitterpunkten bestimmt

$$\phi_i(x_j) = \delta_{ij} = \begin{cases} 1 & \text{für } i = j, \\ 0 & \text{sonst.} \end{cases}, \quad i, j = 1, \ldots, N-1.$$

Die Darstellung als Formel ist

$$\phi_i(x) = \begin{cases} \dfrac{x - x_{i-1}}{x_i - x_{i-1}} & \text{für } x \in [x_{i-1}, x_i], \\ \dfrac{x_{i+1} - x}{x_{i+1} - x_i} & \text{für } x \in [x_i, x_{i+1}], \\ 0 & \text{sonst.} \end{cases}$$

Wegen ihrer charakteristischen Form werden diese Funktionen auch Hütchenfunktionen genannt, siehe Abbildung 4.1. In jeder Gitterzelle $[x_{i-1}, x_i]$ gibt es höchstens zwei (in den inneren Gitterzellen genau zwei) Basisfunktionen, bei welchen diese Gitterzelle eine Teilmenge ihres Trägers ist. Eine der Basisfunktionen nimmt den Wert Eins in x_{i-1} an und Null in x_i, bei der anderen Basisfunktion ist es genau umgekehrt. Somit erhält man für die Testfunktion $\phi_i(x)$ höchsten bei den Ansatzfunktionen $\phi_{i-1}(x), \phi_i(x), \phi_{i+1}(x)$ Nichtnulleinträge. Das bedeutet, in jeder Zeile der Matrix A gibt es höchstens drei Nichtnulleinträge.

Der aufgespannte Finite–Element–Raum $\text{span}\{\phi_i(x)\}_{i=1}^{N-1}$ wird P_1 genannt und er besitzt die Dimension $N-1$. □

Beispiel 4.10 Stückweise quadratische Basisfunktionen. Eine Erweiterung besteht nun darin, stückweise quadratische Basisfunktionen zu betrachten. Auch hier soll der Träger jeder Basisfunktion höchstens die Vereinigung zweier benachbarter Gitterzellen sein. Eine quadratische Funktion ist durch die Vorgabe von drei

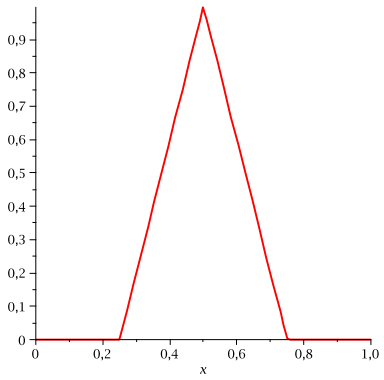


Abbildung 4.1: Hütchenfunktion in $[0.25, 0.5] \cup [0.5, 0.25]$.

Punkten eindeutig festgelegt. In einem Intervall $[x_{i-1}, x_i]$ werden dazu die Werte in den Gitterpunkten x_{i-1} und x_i sowie im Mittelpunkt $(x_{i-1} + x_i)/2$ vorgegeben. Die Gesamtheit der Punkte, in denen man Werte vorgibt, nennt man Knoten. Man hat also $2N - 1$ Knoten ξ_i. Die stückweise quadratische Basis wird so gewählt, dass gilt

$$\phi_i(\xi_j) = \delta_{ij}, \quad i, j = 1, \ldots, 2N - 1.$$

Darstellungsformeln: Übungsaufgabe Damit gibt es zwei Typen von Basisfunktionen, siehe Abbildung 4.2. Ist ξ_j ein Gitterpunkt, dann besteht der Träger der Basisfunktion aus zwei benachbarten Gitterzellen. Für die Testfunktion $\phi_j(x)$ wird man Nichtnulleinträge im allgemeinen dann bekommen, wenn die Ansatzfunktion aus der Menge $\{\phi_{j-2}(x), \phi_{j-1}(x), \phi_j(x), \phi_{j+1}(x), \phi_{j+2}(x)\}$ kommt. Ist ξ_j kein Gitterpunkt, dann ist der Träger sogar nur eine Gitterzelle. Man spricht auch von Blasenfunktionen. Hier wird es Nichtnulleinträge bei Ansatzfunktionen aus der Menge $\{\phi_{j-1}(x), \phi_j(x), \phi_{j+1}(x)\}$ geben.

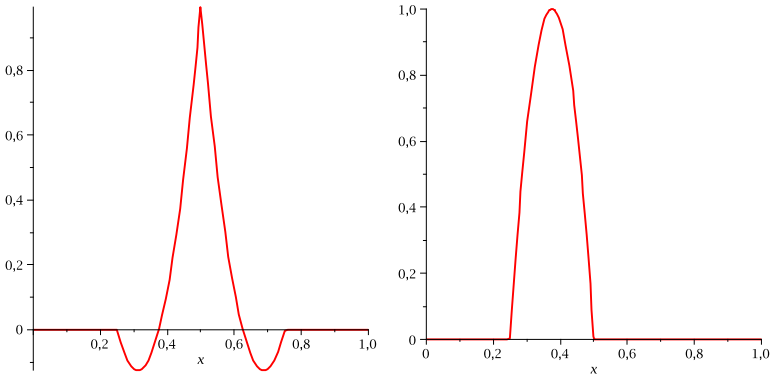


Abbildung 4.2: Quadratische Basisfunktionen in $[0.25, 0.5] \cup [0.5, 0.25]$ beziehungsweise in $[0.25, 0.5]$.

Der aufgespannte Finite–Element–Raum $\text{span}\{\phi_i(x)\}_{i=1}^{2N-1}$ wird P_2 genannt und er besitzt die Dimension $2N-1$. Man hat mit diesem Raum mehr Aufwand als mit P_1 (höhere Dimension des Gleichungssystems, mehr Nichtnulleinträge in der Matrix), aber man wird im allgemeinen auch genauere Ergebnisse erwarten. □

Bemerkung 4.11 Finite–Elemente höherer Ordnung. Diese Konstruktionen lassen sich natürlich fortsetzten. Bei finite Elementen höherer Ordnung gibt es je-

doch unterschiedliche Ansätze, um die Knoten im Inneren der Gitterzelle zu wählen, zumindest in 1D, vergleiche [Sol06]. □

Bemerkung 4.12 Affines Konzept, Referenzzelle, Referenzabbildung. Die Finite–Elemente–Räume in den vorangegangenen Beispielen wurden direkt auf den Zellen des Gitter definiert. Es gibt jedoch noch eine andere Herangehensweise. Bei dieser werden die Basisfunktionen mit ihren Eigenschaften auf einer Referenzzelle $\hat{K}$ definiert. Die Basisfunktionen auf dem Gitter ergeben sich dann durch Referenzabbildungen $F_K : \hat{K} \to K$ auf die Gitterzellen. Falls die Referenzabbildungen affin sind (lineare Abbildung plus konstante Verschiebung), dann sind beide Definitionen oft äquivalent. Die Referenzabbildungen hängen nur von der Gestalt der Gitterzellen ab, aber nicht vom Finite–Element–Raum.

Mit dem affinen Konzept ist zunächst nur definiert, was auf jeder Gitterzelle passiert. Um die Definition eines Finite–Elemente–Raumes zu vervollständigen, man muss noch zusätzliche Eigenschaften für den Übergang zwischen benachbarten Gitterzellen fordern, beispielsweise Stetigkeit für die Räume P_1 und P_2.

Dieses affine Konzept besitzt viele Vorteile bei der Implementierung von Finite–Element–Methoden, da man alle benötigten Informationen (Basisfunktionen, Quadraturformeln) nur auf der Referenzzelle zu programmieren braucht. □

Beispiel 4.13 Affines Konzept für P_1 in 1D. Man nimmt als Referenzgitterzelle beispielsweise $\hat{K} = [-1, 1]$. Die Referenzabbildung auf eine Gitterzelle $K = [x_i, x_{i+1}]$ wird so definiert, dass sie affin ist, das heißt es gilt

$$F_K(\hat{x}) = \alpha\hat{x} + \beta = x,$$

den Punkt -1 bildet man auf x_i sowie den Punkt 1 auf x_{i+1} ab. Das heißt

$$\begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} x_i \\ x_{i+1} \end{pmatrix} \quad \Longrightarrow \quad \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \frac{1}{2} \begin{pmatrix} x_{i+1} - x_i \\ x_{i+1} + x_i \end{pmatrix}.$$

Auf $\hat{K}$ definiert man nun zwei lineare Basisfunktionen

$$\hat{\phi}_1(\hat{x}) = \frac{1}{2}(-\hat{x} + 1), \quad \hat{\phi}_2(\hat{x}) = \frac{1}{2}(\hat{x} + 1).$$

Die erste Funktion ist im Punkt -1 gleich Eins und verschwindet im Punkt 1, bei der zweiten ist es genau umgekehrt.

Die Rücktransformation $F_K^{-1} : K \to \hat{K}$ von der Gitterzelle K auf die Referenzzelle $\hat{K}$ besitzt die Gestalt

$$\hat{x} = \frac{x - \beta}{\alpha} = \frac{2x - (x_{i+1} + x_i)}{x_{i+1} - x_i}.$$

Die Basisfunktionen auf K sind definiert durch

$$\phi_i(x) := \hat{\phi}_i \left(F_K^{-1}(x) \right), \quad i = 1, 2.$$

Damit erhält man

$$\begin{aligned} \phi_1(x) &= \hat{\phi}_1 \left(\frac{2x - (x_{i+1} + x_i)}{x_{i+1} - x_i} \right) = \frac{1}{2} \left(-\frac{2x - (x_{i+1} + x_i)}{x_{i+1} - x_i} + 1 \right) \\ &= -\frac{1}{2} \left(\frac{2x - 2x_{i+1}}{x_{i+1} - x_i} \right) = \frac{x_{i+1} - x}{x_{i+1} - x_i}, \\ \phi_2(x) &= \hat{\phi}_2 \left(\frac{2x - (x_{i+1} + x_i)}{x_{i+1} - x_i} \right) = \frac{x - x_i}{x_{i+1} - x_i}. \end{aligned}$$

Das sind gerade die beiden Basisfunktionen, die man mit direkter Definition auf der Zelle K erhält, Beispiel 4.9. □

Bemerkung 4.14 Assemblierung: Berechnung der Matrixeinträge und der rechten Seite. Die Matrixeinträge des Modellproblems besitzen die Form, siehe Bemerkung 4.6,

$$\begin{aligned} a_{ij} &= \int_0^1 \Big(\varepsilon\phi_j'(x)\phi_i'(x) + b(x)\phi_j'(x)\phi_i(x) + c(x)\phi_j(x)\phi_i(x)\Big)\ dx \\ &= \sum_{k=0}^{N-1} \int_{x_k}^{x_{k+1}} \Big(\varepsilon\phi_j'(x)\phi_i'(x) + b(x)\phi_j'(x)\phi_i(x) + c(x)\phi_j(x)\phi_i(x)\Big)\ dx. \end{aligned}$$

Das heißt, man kann die Integrale auf den einzelnen Gitterzellen berechnen und dann aufsummieren. Beim P_1–Finite–Element hat man für $i = j$ Integrale auf genau zwei Gitterzellen zu berechnen, für $i = j \pm 1$ auf genau einer Gitterzelle und sonst sind alle Integrale Null. Für die Assemblierung der rechten Seite gilt eine analoge Formel.

Es bestehen die Möglichkeiten, die Integrale direkt auf einer Gitterzelle K zu berechnen oder das Integral auf die Referenzzelle $\hat{K}$ zu transformieren. Der zweite Weg ist für die Implementierung von Finite–Element–Methoden günstiger. Nach der Transformation auf $\hat{K}$ kann man eine Quadraturformel anwenden, die für das Referenzelement implementiert ist. Man muss sich jedoch ansehen, wie sich die Terme in den Integralen transformieren.

Die Transformation der Integrale erfolgt natürlich mit der Substitutionsregel unter Verwendung der Referenzabbildung $F_K : [-1,1] \to [x_k, x_{k+1}]$

$$\int_{x_k}^{x_{k+1}} f(x)\ dx = \int_{-1}^{1} f\left(F_K(\hat{x})\right) F_K'(\hat{x})\ d\hat{x}.$$

Es gilt, siehe Beispiel 4.13,

$$F_K(\hat{x}) = \frac{1}{2}\big((x_{k+1} - x_k)\hat{x} + (x_{k+1} + x_k)\big) = x \quad \Longrightarrow \quad F_K'(\hat{x}) = \frac{x_{k+1} - x_k}{2}.$$

Das ist die halbe Länge der Gitterzelle $[x_k, x_{k+1}]$, woraus $F_K'(\hat{x}) > 0$ folgt. Für die Basisfunktion ist die Transformation auf die Referenzzelle durch

$$\phi_i(x) = \hat{\phi}_i\left(F_K^{-1}(x)\right) = \hat{\phi}_i(\hat{x})$$

gegeben. Zur Transformation der Ableitung verwendet man die Kettenregel

$$\phi_i'(x) = \frac{d\phi_i(x)}{dx} = \frac{d\hat{\phi}_i(\hat{x})}{dx} = \frac{d\hat{\phi}_i(\hat{x})}{d\hat{x}}\frac{d\hat{x}}{dx} = \frac{d\hat{\phi}_i(\hat{x})}{d\hat{x}}\frac{2}{x_{k+1} - x_k}.$$

Die Ableitungen der Basisfunktionen auf der Referenzzelle kann man vorher explizit ausrechnen und dann implementieren. Damit erhält man

$$\begin{aligned} \int_{x_k}^{x_{k+1}} \varepsilon\phi_j'(x)\phi_i'(x)\ dx &= \frac{2}{x_{k+1} - x_k}\int_{-1}^{1} \varepsilon\frac{d\hat{\phi}_j(\hat{x})}{d\hat{x}}\frac{d\hat{\phi}_i(\hat{x})}{d\hat{x}}\ d\hat{x}, \\ \int_{x_k}^{x_{k+1}} b(x)\phi_j'(x)\phi_i(x)\ dx &= \int_{-1}^{1} b(F_K(\hat{x}))\frac{d\hat{\phi}_j(\hat{x})}{d\hat{x}}\hat{\phi}_i(\hat{x})\ d\hat{x}, \\ \int_{x_k}^{x_{k+1}} c(x)\phi_j(x)\phi_i(x)dx &= \frac{x_{k+1} - x_k}{2}\int_{-1}^{1} c(F_K(\hat{x}))\hat{\phi}_j(\hat{x})\hat{\phi}_i(\hat{x})\ d\hat{x}, \\ \int_{x_k}^{x_{k+1}} f(x)\phi_i(x)dx &= \frac{x_{k+1} - x_k}{2}\int_{-1}^{1} f(F_K(\hat{x}))\hat{\phi}_i(\hat{x})\ d\hat{x}. \end{aligned}$$

Nun kann man hinreichend genaue Quadraturformeln auf der Referenzzelle zur Approximation der Integrale verwenden. Eine genaue Quadratur ist vor allem für finite Elemente höherer Ordnung wesentlich, damit die Genauigkeit nicht durch Quadraturfehler beeinträchtigt wird. □

Beispiel 4.15 Assemblierung: P_1 in 1D. Betrachte den Fall, dass die Parameterfunktionen konstant sind, $b(x) = b$, $c(x) = c$ und $f(x) = f$. Für das P_1–Finite–Element auf der Referenzzelle $[-1, 1]$ gelten

$$\hat{\phi}_1(\hat{x}) = \frac{1}{2}(-\hat{x} + 1), \quad \hat{\phi}_1'(\hat{x}) = -\frac{1}{2}, \qquad \hat{\phi}_2(\hat{x}) = \frac{1}{2}(\hat{x} + 1), \quad \hat{\phi}_2'(\hat{x}) = \frac{1}{2}.$$

Betrachte nun den Matrixeintrag $a_{i,i+1}$, der mit Hilfe der Testfunktion $\phi_i(x)$, die auf $\hat{\phi}_1(\hat{x})$ transformiert wird, und der Ansatzfunktion $\phi_{i+1}(x)$, die auf $\hat{\phi}_2(\hat{x})$ transformiert wird, berechnet wird. Der gemeinsame Träger ist die Gitterzelle $[x_i, x_{i+1}]$. Bezeichne h die Länge dieser Zelle. Dann folgt

$$\begin{aligned} a_{i,i+1} &= \frac{2\varepsilon}{h} \int_{-1}^{1} \frac{1}{2} \cdot \left(-\frac{1}{2}\right)\, d\hat{x} + b \int_{-1}^{1} \frac{1}{2} \cdot \frac{1}{2}(-\hat{x} + 1)\, d\hat{x} \\ &\quad + \frac{ch}{2} \int_{-1}^{1} \frac{1}{2}(\hat{x} + 1) \frac{1}{2}(-\hat{x} + 1)\, d\hat{x} \\ &= -\frac{\varepsilon}{h} + \frac{b}{2} + \frac{ch}{6}. \end{aligned}$$

Für die i–te Komponente der rechten Seite, erhält, man

$$f_i = f \int_0^1 \phi_i(x)\, dx = hf,$$

da die Fläche unter einer Hütchenfunktion das Maß h besitzt. *andere Einträge als Übungsaufgabe*

Verwendet man zur Approximation der Integrale die Trapezregel, so ergibt sich

$$\frac{ch}{2} \int_{-1}^{1} \frac{1}{2}(-\hat{x} + 1) \frac{1}{2}(\hat{x} + 1)\, d\hat{x} = \frac{ch}{2} 2(0 + 0) = 0.$$

In diesem Fall ist

$$a_{i,i+1} = -\frac{\varepsilon}{h} + \frac{b}{2}.$$

Für $c = 0$ (oder mit Trapezregel) sind das, bis auf den Faktor h, die gleichen Einträge wie beim zentralen Differenzenverfahren, siehe Bemerkung 2.11. Man erhält für $c = 0$

$$-h\varepsilon D^+ D^- u_i + h b_i D^0 u_i = h f_i \quad \Longleftrightarrow \quad -\varepsilon D^+ D^- u_i + b_i D^0 u_i = f_i.$$

Dieser Zusammenhang zwischen Finite–Differenzen–Methoden und Finite–Element–Methoden gilt im allgemeinen nicht mehr, wenn die Koeffizientenfunktionen nicht konstant sind. In höheren Dimensionen unterscheiden sich FDM und FEM im allgemeinen auch bei konstanten Koeffizienten. □

Bemerkung 4.16 Andere Randbedingungen. Hat man inhomogene Dirichlet–Randbedingungen

$$u(0) = a \quad u(1) = b$$

oder Neumann–Randbedingungen

$$\varepsilon u'(0) = \alpha, \quad \varepsilon u'(1) = \beta,$$

so nimmt man zur Gleichungsassemblierung auch die Basisfunktionen hinzu, die in den Randpunkten den Wert Eins haben und in allen anderen Knoten verschwinden. Bei inhomogenen Dirichlet–Randbedingungen ersetzt man dann die entsprechenden

Gleichungen (bei Numerierung von links nach rechts die erste und letzte Gleichung) durch

$$(1,0,\ldots,0)\begin{pmatrix} u^0 \\ \vdots \end{pmatrix} = \begin{pmatrix} a \\ \vdots \end{pmatrix}, \quad (0,0,\ldots,1)\begin{pmatrix} \vdots \\ u^k \end{pmatrix} = \begin{pmatrix} \vdots \\ b \end{pmatrix}.$$

Bei Neumann–Randbedingungen treten in natürlicher Art und Weise, siehe Bemerkung 3.33, zusätzliche Terme auf der rechten Seite der ersten und letzten Gleichung auf. □

Bemerkung 4.17 CSR–Speicherschema von schwach besetzten Matrizen. Von schwach besetzten Matrizen speichert man natürlich nur die Einträge, die nicht Null sind und zugehörige Informationen über die Position des Eintrags. Die am weitesten verbreiteste Herangehensweise ist das CSR–Speicherschema (condensed sparse row). Bei diesem Schema werden die Nichtnulleinträge zeilenweise abgespeichert. Innerhalb einer Zeile brauchen sie nicht bezüglich der Spaltenindizes angeordnet zu werden.

Sei eine schwach besetzte Matrix $A \in \mathbb{R}^{m\times n}$ mit nnz Nichtnullelementen zu speicher. Dann braucht man drei Arrays:

- `double`–Array `entries` der Länge nnz, darin werden die Einträge von A zeilenweise gespeichert,
- `int`–Array `col_ptr` der Länge nnz, darin stehen die Spaltenindizes der zugehörigen Einträge von `entries`.
- `int`–Array `row_ptr` der Länge $m+1$, darin wird abgespeichert, an welcher Stelle im Array `entries` die i–te Zeile beginnt, $i = 1,\ldots,m$; der letzte Eintrag von `row_ptr` verweist auf den ersten Speicherplatz nach dem Ende des Arrays `entries`,

□

Beispiel 4.18 Die Matrix

$$A = \begin{pmatrix} 1 & 0 & 0 & 2 & 0 \\ 3 & 4 & 0 & 5 & 0 \\ 6 & 0 & 7 & 8 & 9 \\ 0 & 0 & 10 & 11 & 0 \\ 0 & 0 & 0 & 0 & 12 \end{pmatrix}$$

kann wie folgt gespeichert werden (Numerierung beginnt bei 0):

`entries`	–	1	2	3	4	5	6	7	8	9	10	11	12
`col_ptr`	–	0	3	0	1	3	0	2	3	4	2	3	4 .
`row_ptr`	–	0	2	5	9	11	12						

Eine andere Möglichkeit ist

`entries`	–	2	1	4	5	3	7	9	8	6	11	10	12
`col_ptr`	–	3	0	1	3	0	2	4	3	0	3	2	4 .
`row_ptr`	–	0	2	5	9	11	12						

□

4.3 Polynominterpolation in Sobolov–Räumen und Konvergenzabschätzungen

Bemerkung 4.19 Motivation. Die variationelle Formulierung partieller Differentialgleichungen benutzt Funktionen aus Sobolev–Räumen. Die Lösung soll mit Hilfe

der Ritzschen Methode und endlich–dimensionaler Finite–Element–Räume approximiert werden. Der Fehler in der durch den Raum V induzierten Norm hängt davon ab, wie gut man Funktionen aus Sobolev–Räumen überhaupt mit Funktionen aus Finite–Element–Räumen annähern kann, siehe zum Beispiel das Lemma von Cea, Abschätzung (4.6). Die Approximationsgüte von Finite–Element–Räumen wird in diesem Abschnitt untersucht. □

4.3.1 Das Bramble–Hilbert–Lemma

Wir beginnen mit grundlegenden Prinzipien der Polynominterpolation in Sobolev–Räumen.

Lemma 4.20 *Sei $(a,b) \subset \mathbb{R}$. Für jeden Index γ mit $0 \leq \gamma \leq m$ sei ein $a_\gamma \in \mathbb{R}$ gegeben. Dann gibt es ein eindeutig bestimmtes Polynom $p \in P_m(a,b)$ mit*

$$\int_a^b p^{(\gamma)}(x)\ dx = a_\gamma,\ 0 \leq \gamma \leq m.$$

Beweis: Jedes Polynom aus $P_m(a,b)$ hat die Gestalt

$$p(x) = \sum_{\mu=0}^{m} b_\mu x^\mu.$$

Einsetzen dieser Darstellung in die Bedingungen ergibt ein lineares Gleichungssystem $M\mathbf{b} = \mathbf{a}$, mit

$$M = (M_{\gamma\mu}),\ M_{\gamma\mu} = \int_a^b (x^\mu)^{(\gamma)}\ dx,\ \mathbf{b} = (b_\mu),\ \mathbf{a} = (a_\gamma),$$

für $0 \leq \gamma, \mu \leq m$. Das ist ein quadratisches Gleichungssystem, welches genau dann eine eindeutige Lösung besitzt, wenn M regulär ist.

Angenommen, M ist singulär. Dann besitzt das zugehörige homogene Gleichungssystem eine nichttriviale Lösung. Dass heißt, es gibt ein Polynom $q \in P_m(a,b) \setminus \{0\}$ mit

$$\int_a^b q^{(\gamma)}(x)\ dx = 0 \text{ für alle } 0 \leq \gamma \leq m.$$

Das Polynom q besitzt die Darstellung $q(x) = \sum_{\mu=0}^m c_\mu x^\mu$. Wähle nun das $c_\mu \neq 0$ mit maximalem μ. Dann gilt $q^{(\mu)}(x) = \mu(\mu-1)\ldots\cdot 2\cdot 1\cdot c_\mu = const \neq 0$, woraus

$$\int_a^b q^{(\mu)}(x)\ dx = \int_a^b const\ dx = (b-a)const \neq 0$$

folgt. Das widerspricht dem Verschwinden des Integrals für $q^{(\mu)}(x)$. Somit ist die Annahme falsch und M ist nicht singulär. ■

Das Lemma besagt, dass ein Polynom eindeutig bestimmt ist, wenn man für jede Ableitung eine Bedingung an das Integral über (a,b) stellt.

Lemma 4.21 Ungleichung vom Poincaré–Typus. *Sei (a,b) mit $R = b-a$. Seien $k,l \in \mathbb{N}$ mit $0 \leq k \leq l$ und sei $p \in \mathbb{R}$ mit $p \in [1,\infty]$. Dann gilt für jedes $v \in W^{l,p}(a,b)$, welches*

$$\int_a^b v^{(\gamma)}(x)\ dx = 0 \text{ für alle } 0 \leq \gamma \leq l-1$$

erfüllt, die Abschätzung

$$\left\|v^{(k)}\right\|_{L^p(a,b)} \leq CR^{l-k}\left\|v^{(l)}\right\|_{L^p(a,b)},$$

wobei die Konstante c nicht von (a,b) und von $v(x)$ abhängt.

Beweis: Im Fall $k = l$ braucht man nichts zu beweisen. Des weiteren genügt es, das Lemma für $k = 0$ und $l = 1$ zu beweisen, da der allgemeine Fall folgt, wenn man das Resultat dann auf die γ–te Ableitung anwendet.

Zu zeigen ist also

$$\|v\|_{L^p(a,b)} \le CR \left\|v'\right\|_{L^p(a,b)} \quad \text{falls} \quad \int_a^b v(x)\, dx = 0. \tag{4.7}$$

Es gilt für $x, y \in (a, b)$

$$\begin{aligned} \int_0^1 v'(tx + (1-t)y)\, dt &= \int_0^1 v'(t(x-y) + y)\, dt \\ &= \frac{1}{x-y}\Big(v(t(x-y)+y)|_{t=1} - v(t(x-y)+y)|_{t=0}\Big) \\ &= \frac{v(x) - v(y)}{x-y}, \end{aligned}$$

was eine Form des Mittelwertsatzes ist. Multiplikation mit $(x - y)$ und anschließende Integration bezüglich y ergibt

$$v(x)\int_a^b dy - \underbrace{\int_a^b v(y)\, dy}_{=0} = \int_a^b \int_0^1 v'(tx + (1-t)y)(x-y)\, dt\, dy,$$

wobei das eine Integral auf der linken Seite nach Voraussetzung an $v(x)$ verschwindet. Damit wurde schon die Voraussetzung von (4.7) verwendet. Es folgt

$$v(x) = \frac{1}{R}\int_a^b \int_0^1 v'(tx + (1-t)y)(x-y)\, dt\, dy.$$

Nun muss man versuchen, die Terme der Behauptung in (4.7) zu bekommen. Man beginnt mit der linken Seite der Ungleichung. Es wird verwendet, dass der Betrag eines Integrals durch das Integral des Betrags abgeschätzt werden kann, sowie dass $|x - y| \le R$ gilt

$$|v(x)| \le \frac{1}{R}\int_a^b \int_0^1 \left|v'(tx + (1-t)y)\right| |x-y|\, dt\, dy \le \frac{R}{R}\int_a^b \int_0^1 \left|v'(tx + (1-t)y)\right|\, dt\, dy. \tag{4.8}$$

Für $p < \infty$ wird diese Abschätzung mit p potenziert und bezüglich x integriert. Man erhält durch Anwendung der Hölderschen Ungleichung mit $p^{-1} + q^{-1} = 1$

$$\begin{aligned} \int_a^b |v(x)|^p\, dx &\le \int_a^b \left(\int_a^b \int_0^1 \left|v'(tx + (1-t)y)\right|\, dt\, dy\right)^p dx \\ &\le \int_a^b \left[\underbrace{\left(\int_a^b \int_0^1 1^q\, dt\, dy\right)^{p/q}}_{R^{p/q}} \left(\int_a^b \int_0^1 \left|v'(tx + (1-t)y)\right|^p\, dt\, dy\right)\right] dx \\ &= R^{p/q} \int_a^b \left(\int_a^b \int_0^1 \left|v'(tx + (1-t)y)\right|^p\, dt\, dy\right) dx. \end{aligned}$$

Damit hat man die p–te Potenz der linken Seite der Ungleichung in (4.7). Nun braucht man noch die p–te Potzen der rechten Seite der Ungleichung. Es werden zunächst die Integrationen vertauscht (Satz von Fubini)

$$\int_a^b |v(x)|^p\, dx \le R^{p/q} \int_0^1 \int_a^b \left(\int_a^b \left|v'(tx + (1-t)y)\right|^p\, dy\right) dx\, dt.$$

Mit dem Mittelwertsatz der Integralrechnung findet man ein $t_0 \in [0, 1]$, so dass

$$\int_a^b |v(x)|^p\, dx \le R^{p/q}(1-0) \int_a^b \left(\int_a^b \left|v'(t_0 x + (1-t_0)y)\right|^p\, dy\right) dx.$$

Man setzt $|v'(x)|^p$ auf $\mathbb{R}$ durch Null fort und nennt die Fortsetzung ebenfalls $|v'(x)|^p$. Dann ist

$$\int_a^b |v(x)|^p \ dx \leq R^{p/q} \int_a^b \left(\int_{\mathbb{R}} |v'(t_0 x + (1-t_0)y)|^p \ dy \right) dx.$$

Sei $t_0 \in [0, 1/2]$. Da das Integrationsgebiet nun der ganze $\mathbb{R}$ ist, ergibt die Variablensubstitution $t_0 x + (1-t_0)y = z$

$$\int_{\mathbb{R}} |v'(t_0 x + (1-t_0)y)|^p \ dy = \frac{1}{1-t_0} \int_{\mathbb{R}} |v'(z)|^p \ dz \leq 2 \left\| v' \right\|^p_{L^p(a,b)},$$

da $1/(1-t_0) \leq 2$. Führt man nun noch die äußere Integration über x aus, erhält man insgesamt

$$\begin{aligned} \int_a^b |v(x)|^p \ dx &\leq R^{p/q} \int_a^b 2 \left\| v' \right\|^p_{L^p(a,b)} \ dx = 2R^{p/q} \left\| v' \right\|^p_{L^p(a,b)} \int_a^b \ dx \\ &= 2R^{p/q+1} \left\| v' \right\|^p_{L^p(a,b)} = 2R^p \left\| v' \right\|^p_{L^p(a,b)}. \end{aligned}$$

Im Fall $t_0 > 1/2$ vertauscht man die Rollen von x und y sowie die Integrationsreihenfolge mit dem Satz von Fubini und argumentiert analog.

Der Fall $p = \infty$ folgt aus (4.8). *Übungsaufgabe* ■

Bemerkung 4.22 Das Lemma besagt, dass man die $L^p(a,b)$–Norm einer niederen Ableitung von $v(x)$ durch dieselbe Norm einer Ableitung höherer Ordnung abschätzen kann, falls die Integralmittelwerte der niederen Ableitungen verschwinden. Eine wichtige Anwendung dieses Lemmas ist der Beweis des Bramble–Hilbert–Lemmas. Dieses besagt, dass der Wert eines stetigen linearen Funktionals, das auf einem Sobolev–Raum definiert ist und auf einem Polynomraum der Ordnung m verschwindet, durch die Lebesgue–Norm der $m+1$–ten Ableitung der Funktionen aus dem Sobolev–Raum abgeschätzt werden kann. □

Satz 4.23 ***Bramble***[4]***–Hilbert–Lemma.*** *Seien $m \in \mathbb{N}$, $m \geq 0$, $p \in [1, \infty]$ und $F : W^{m+1,p}(a,b) \to \mathbb{R}$ ein stetiges lineares Funktional und seien die Voraussetzungen der Lemmata 4.20 und 4.21 erfüllt. Weiter sei*

$$F(p) = 0 \quad \forall \ p \in P_m(a,b).$$

Dann gibt es eine Konstante $c(a,b)$, die unabhängig von $v(x)$ und F ist, mit

$$|F(v)| \leq c(a,b) \left\| v^{(m+1)} \right\|_{L^p(a,b)} \quad \forall \ v \in W^{m+1,p}(a,b).$$

Beweis: Sei $v \in W^{m+1,p}(a,b)$ mit

$$\int_a^b v^{(\gamma)}(x) \ dx = a_\gamma \text{ für } 0 \leq \gamma \leq m.$$

Wegen Lemma 4.20 gibt es ein Polynom aus $P_m(a,b)$ mit

$$\int_a^b p^{(\gamma)}(x) \ dx = -a_\gamma, \quad 0 \leq \gamma \leq m, \quad \Longrightarrow \quad \int_a^b (v+p)^{(\gamma)}(x) \ dx = 0, \quad 0 \leq \gamma \leq m.$$

Lemma 4.21 liefert, mit $l = m+1$, nun die Abschätzung

$$\begin{aligned} \|v+p\|_{W^{m+1,p}(a,b)} &= \left(\sum_{i=0}^{m+1} \left\| (v+p)^{(i)} \right\|^p_{L^p(a,b)} \right)^{1/p} \\ &\leq \left(\sum_{i=0}^{m+1} c_i(a,b) \left\| (v+p)^{(m+1)} \right\|^p_{L^p(a,b)} \right)^{1/p} \\ &= \left(\sum_{i=0}^{m+1} c_i(a,b) \right)^{1/p} \left\| (v+p)^{(m+1)} \right\|_{L^p(a,b)} \\ &= c(a,b) \left\| (v+p)^{(m+1)} \right\|_{L^p(a,b)} = c(a,b) \left\| v^{(m+1)} \right\|_{L^p(a,b)}. \end{aligned}$$

[4] James Bramble

Aus dem Verschwinden von F für $p \in P_m(a,b)$ und der Stetigkeit von F folgt nun

$$|F(v)| = |F(v) + F(p)| = |F(v+p)| \leq c \|v+p\|_{W^{m+1,p}(a,b)} \leq c(a,b) \left\| v^{(m+1)} \right\|_{L^p(a,b)}.$$

■

4.3.2 Interpolationsfehlerabschätzung

Bemerkung 4.24 Herangehensweise. Der Interpolationsfehler wird nun mit Hilfe des Bramble–Hilbert–Lemmas abgeschätzt. Die Strategie wird darin bestehen, dass man

- zuerst Abschätzungen auf einer Referenzgitterzelle zeigt,
- dann werden alle Abschätzungen über beliebige Gitterzellen K auf Abschätzungen über die Referenzgitterzelle überführt,
- die dort gezeigten Abschätzungen werden verwendet und
- schließlich wird auf K zurücktransformiert.

Dabei muss man auch untersuchen, was bei den beiden Transformationen geschieht.

□

Bemerkung 4.25 Interpolierende. Eine Interpolierende ist eine (vernünftige Approximation einer Funktion aus einem Sobolev–Raum durch eine Funktion aus dem Finite–Elemente–Raum.

Die analytische Formulierung auf einer Referenzgitterzelle $\hat{K}$ ist wie folgt. Seien $\hat{K} \subset \mathbb{R}$, zum Beispiel $\hat{K} = [-1,1]$, $\hat{P}(\hat{K})$ ein Polynomraum der Dimension N und $\hat{\Phi}_1, \ldots, \hat{\Phi}_N : C^s(\hat{K}) \to \mathbb{R}$ stetige lineare Funktionale und $\hat{\phi}_1(\hat{x}), \ldots, \hat{\phi}_N(\hat{x}) \in \hat{P}(\hat{K})$ eine lokale Basis. Das heißt, $\{\hat{\phi}_i(\hat{x})\}_{i=1}^N$ ist eine Basis von $\hat{P}(\hat{K})$ und es gilt

$$\hat{\Phi}_i(\hat{\phi}_j) = \delta_{ij}, \quad i,j = 1, \ldots, N.$$

Für $\hat{v} \in C^s(\hat{K})$ wird die Interpolierende $(I_{\hat{K}}\hat{v})(\hat{x})$ durch

$$I_{\hat{K}}\hat{v}(\hat{x}) = \sum_{i=1}^{N} \hat{\Phi}_i(\hat{v})\hat{\phi}_i(\hat{x})$$

definiert. Der Operator $I_{\hat{K}}$ ist ein stetiger und linearer Operator von $C^s(\hat{K})$ nach $\hat{P}(\hat{K})$. Aus der Linearität folgt, dass $I_{\hat{K}}$ die Identität auf $\hat{P}(\hat{K})$ ist *Übungsaufgabe*

$$\left(I_{\hat{K}}\hat{p}\right)(\hat{x}) = \hat{p}(\hat{x}) \quad \forall\, \hat{p} \in \hat{P}(\hat{K}).$$

□

Beispiel 4.26 Unterschiedliche konstante Interpolierende. Seien $\hat{K} \subset \mathbb{R}$ beliebig, $\hat{P}(\hat{K}) = P_0(\hat{K})$ und

$$\hat{\Phi}(\hat{v}) = \frac{1}{\left|\hat{K}\right|} \int_{\hat{K}} \hat{v}(\hat{x})\, d\hat{x}.$$

Das Funktional $\hat{\Phi}$ ist linear wegen der Linearität der Integration. Es ist stetig auf $C^0(\hat{K})$, da

$$\left|\hat{\Phi}(\hat{v})\right| \leq \frac{1}{\left|\hat{K}\right|} \int_{\hat{K}} |\hat{v}(\hat{x})|\, d\hat{x} \leq \frac{\left|\hat{K}\right|}{\left|\hat{K}\right|} \max_{\hat{x}\in\hat{K}} |\hat{v}(\hat{x})| = \|\hat{v}\|_{C^0(\hat{K})}.$$

Für die konstante Funktion $1 \in P_0(\hat{K})$ gilt $\hat{\Phi}(1) = 1 \neq 0$. Damit ist $\{1\}$ eine lokale Basis. Der Operator

$$I_{\hat{K}}\hat{v}(\hat{x}) = \hat{\Phi}(\hat{v})\hat{\phi}(\hat{x}) = \frac{1}{\left|\hat{K}\right|}\int_{\hat{K}} \hat{v}(\hat{x})\ d\hat{x}$$

ist der Mittelwertoperator, das heißt jede stetige Funktion auf $\hat{K}$ wird durch eine konstante Funktion interpoliert, deren Funktionswert gleich dem Integralmittelwert ist, siehe Abbildung 4.3 für ein konkretes Beispiel.

Man kann auch $\hat{\Phi}(\hat{v}) = \hat{v}(\hat{x}_0)$ für einen beliebigen Punkt $\hat{x}_0 \in \hat{K}$ setzen. Auch dieses Funktional ist linear

$$\hat{\Phi}(\alpha\hat{v} + \beta\hat{w}) = (\alpha\hat{v} + \beta\hat{w})(\hat{x}_0) = \alpha\hat{v}(\hat{x}_0) + \beta\hat{w}(\hat{x}_0)$$

für alle $\alpha, \beta \in \mathbb{R}$ und $\hat{v}, \hat{w} \in C^0(\hat{K})$ und stetig auf $C^0(\hat{K})$

$$\left|\hat{\Phi}(\hat{v})\right| = |\hat{v}(\hat{x}_0)| \leq \max_{\hat{x}\in\hat{K}} |\hat{v}(\hat{x})| = \|\hat{v}\|_{C^0(\hat{K})}\,.$$

Der damit definierte Interpolationsoperator $I_{\hat{K}}$ interpoliert jede stetige Funktion durch eine konstante Funktion, deren Funktionswert gleich dem Funktionswert in $\hat{x}_0$ ist.

Diese Beispiele zeigen, dass der Interpolationsoperator $I_{\hat{K}}$ von $\hat{P}(\hat{K})$ und von den gewählten Funktionalen $\hat{\Phi}_i$ abhängt.

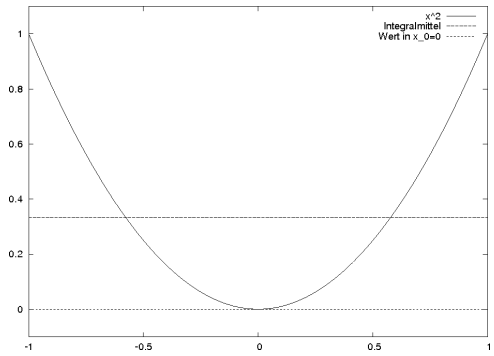


Abbildung 4.3: Interpolation von x^2 im Intervall $[-1, 1]$ in den P_0 mit Integralmittelwert und mit den Funktionswert in $x_0 = 0$.

Übungsaufgabe, Interpolationen für andere FE □

Satz 4.27 Interpolationsfehlerabschätzung auf der Referenzgitterzelle. *Seien $P_m(\hat{K}) \subset \hat{P}(\hat{K})$ und $p \in [1, \infty]$ mit $(m + 1 - s)p > 1$, wobei die Bezeichnungen von Bemerkung 4.25 verwendet werden. Dann gibt es eine von $\hat{v}(\hat{x})$ unabhängige Konstante c mit*

$$\left\|\hat{v} - I_{\hat{K}}\hat{v}\right\|_{W^{m+1,p}(\hat{K})} \leq c\left\|\hat{v}^{(m+1)}\right\|_{L^p(\hat{K})} \quad \forall\ \hat{v} \in W^{m+1,p}(\hat{K}).$$

Beweis: Mit der Bedingung $(m + 1 - s)p > 1$ kann man zeigen, dass

$$W^{m+1,p}(\hat{K}) \subset C^s(\hat{K}), \quad \|\hat{v}\|_{C^s(\hat{K})} \leq c\,\|\hat{v}\|_{W^{m+1,p}(\hat{K})}$$

gelten, siehe Literatur. Damit ist der Interpolationsoperator auf $W^{m+1,p}(\hat{K})$ wohldefiniert.

Aus der Identität des Interpolationsoperators auf $P_m(\hat{K})$, der Beschränktheit des Interpolationsoperators und der obigen Ungleichung erhält man für $\hat{q} \in P_m(\hat{K})$

$$\begin{aligned}
\|\hat{v} - I_{\hat{K}}\hat{v}\|_{W^{m+1,p}(\hat{K})} &= \|\hat{v} + \hat{q} - I_{\hat{K}}(\hat{v} + \hat{q})\|_{W^{m+1,p}(\hat{K})} \\
&\leq \|\hat{v} + \hat{q}\|_{W^{m+1,p}(\hat{K})} + \|I_{\hat{K}}(\hat{v} + \hat{q})\|_{W^{m+1,p}(\hat{K})} \\
&\leq \|\hat{v} + \hat{q}\|_{W^{m+1,p}(\hat{K})} + c\,\|\hat{v} + \hat{q}\|_{C^s(\hat{K})} \\
&\leq c\,\|\hat{v} + \hat{q}\|_{W^{m+1,p}(\hat{K})}\,.
\end{aligned}$$

In Lemma 4.20 wird $\hat{q}(\hat{x})$ nun so gewählt, dass

$$\int_{\hat{K}} (\hat{v} + \hat{q})^{(\gamma)}(\hat{x})\; d\hat{x} = 0 \quad 0 \leq \gamma \leq m.$$

Damit sind die Voraussetzungen von Lemma 4.21 erfüllt und es gilt

$$\|\hat{v} + \hat{q}\|_{W^{m+1,p}(\hat{K})} \leq c \left\|(\hat{v} + \hat{q})^{(m+1)}\right\|_{L^p(\hat{K})} = c \left\|\hat{v}^{(m+1)}\right\|_{L^p(\hat{K})}.$$

■

Bemerkung 4.28 Eigenschaften der Referenzabbildung. Die Eigenschaften der Referenzabbildung in einer Dimension

$$F_K(\hat{x}) = \alpha\hat{x} + \beta = x$$

lassen sich sehr leicht feststellen. Sie bildet ein Intervall $\hat{K}$ mit fester Länge auf ein Intervall K mit Länge h_K ab. Demzufolge gilt $\alpha = ch_K$ mit $c \in \mathbb{R}$, siehe Beispiel 4.13. Bei der Substitutionsregel der Integration erhält man beim Übergang von K zu $\hat{K}$ den Faktor α und bei der Umkehrsubstitution den Faktor $1/\alpha$.

In höheren Dimensionen ist die Untersuchung der Referenzabbildung weitaus komplizierter. □

Bemerkung 4.29 Transformation des Interpolationsoperators. Als nächstes soll sichergestellt werden, dass der transformierte Interpolationsoperator mit dem natürlichen Interpolationsoperator auf K übereinstimmt. Der letztere ist durch

$$I_K v(x) = \sum_{i=1}^{N} \Phi_{K,i}(v)\phi_{K,i}(x)$$

definiert, wobei $\{\phi_{K,i}(x)\}$ die Basis des Raums

$$P(K) = \{p \;:\; K \to \mathbb{R} \;:\; p = \hat{p} \circ F_K^{-1}, \hat{p} \in \hat{P}(\hat{K})\}$$

ist, die der Beziehung $\Phi_{K,i}(\phi_{K,j}) = \delta_{ij}$ genügt. Die Funktionale waren durch

$$\Phi_{K,i}(v) = \hat{\Phi}_i(v \circ F_K)$$

definiert. Daher folgt aus der Bedingung für die lokale Basis

$$\Phi_{K,i}(\hat{\phi}_j \circ F_K^{-1}) = \hat{\Phi}_{K,i}(\hat{\phi}_j \circ F_K^{-1} \circ F_K) = \hat{\Phi}_i(\hat{\phi}_j) = \delta_{ij},$$

also wegen $\Phi_{K,i}(\phi_{K,j}) = \delta_{ij}$ folgt $\phi_{K,j}(x) = \left(\hat{\phi}_j \circ F_K^{-1}\right)(x)$. Aus

$$\begin{aligned}
I_{\hat{K}}\hat{v}(\hat{x}) &= \sum_{i=1}^{N} \hat{\Phi}_i(\hat{v})\hat{\phi}_i(\hat{x}) = \sum_{i=1}^{N} \Phi_{K,i}(\underbrace{\hat{v} \circ F_K^{-1}}_{=v})\,(\phi_{K,i} \circ F_K)\,(\hat{x}) \\
&= \left(\sum_{i=1}^{N} \Phi_{K,i}(v)\phi_{K,i}(x)\right) \circ F_K = (I_K v \circ F_K)\,(x)
\end{aligned}$$

ergibt sich, dass $I_{\hat{K}}\hat{v}(\hat{x})$ sich richtig transformiert. □

Satz 4.30 Interpolationsabschätzung für eine beliebige Gitterzelle. *Seien eine Referenzgitterzelle $\hat{K}$, Funktionale $\{\hat{\Phi}_i\}$ und ein Polynomraum $\hat{P}(\hat{K})$ gegeben. Weiter seien alle Bedingungen aus dem Satz 4.27 erfüllt. Dann gibt es eine Konstante c unabhängig von $v \in W^{m+1,p}(K)$ mit*

$$\left\|(v - I_K v)^{(k)}\right\|_{L^p(K)} \leq c h_K^{m+1-k} \left\|v^{(m+1)}\right\|_{L^p(K)}, \quad 0 \leq k \leq m+1, \tag{4.9}$$

für alle $v \in W^{m+1,p}(K)$. Man beachte, dass die Potenz von h_K unabhängig von p ist.

Beweis: Seien $\hat{v}(\hat{x}) = v\left(F_K(\hat{x})\right)$ beziehungsweise $v(x) = \hat{v}\left(F_K^{-1}(x)\right)$. Mit der Kettenregel folgen

$$\frac{d\hat{v}(x)}{d\hat{x}} = \frac{dv(x)}{dx}\frac{dx}{d\hat{x}} = \alpha\frac{dv(x)}{dx} = c h_K \frac{dv(x)}{dx}, \quad \frac{dv(x)}{dx} = \frac{d\hat{v}(x)}{d\hat{x}}\frac{d\hat{x}}{dx} = \frac{1}{\alpha}\frac{d\hat{v}(x)}{d\hat{x}} = \frac{c}{h_K}\frac{d\hat{v}(x)}{d\hat{x}}.$$

Die Konstante c kann an unterschiedlichen Stellen verschiedene Werte annehmen. Daraus ergeben sich, mit jeder Ableitung erhält man einen weiteren Faktor ch_K beziehungsweise c/h_K,

$$\left|v^{(k)}(x)\right| \leq c h_K^{-k} \left|\hat{v}^{(k)}(\hat{x})\right|, \quad \left|\hat{v}^{(k)}(\hat{x})\right| \leq c h_K^{k} \left|v^{(k)}(x)\right|.$$

Man erhält mit Substitutionsregel für jede Funktion $v \in W^{k,p}(K)$

$$\begin{aligned}\int_K \left|v^{(k)}(x)\right|^p \, dx &\leq c h_K^{-kp} \int_{\hat{K}} \left|\hat{v}^{(k)}(\hat{x})\right|^p h_K \, d\hat{x} = c h_K^{-kp+1} \int_{\hat{K}} \left|\hat{v}^{(k)}(\hat{x})\right|^p \, d\hat{x} \\ &= c h_K^{-kp+1} \left\|\hat{v}^{(k)}\right\|^p_{L^p(\hat{K})}\end{aligned}$$

und

$$\begin{aligned}\int_{\hat{K}} \left|\hat{v}^{(k)}(\hat{x})\right|^p \, d\hat{x} &\leq c h_K^{kp} \int_K \left|v^{(k)}(x)\right|^p h_K^{-1} \, dx = c h_K^{kp-1} \int_K \left|v^{(k)}(x)\right|^p \, dx \\ &= c h_K^{kp-1} \left\|v^{(k)}\right\|^p_{L^p(K)}.\end{aligned}$$

Aus der Interpolationsfehlerabschätzung auf der Referenzzelle folgt

$$\left\|(\hat{v} - I_{\hat{K}}\hat{v})^{(k)}\right\|^p_{L^p(\hat{K})} \leq c \left\|\hat{v}^{(k)}\right\|^p_{L^p(\hat{K})}, \quad 0 \leq k \leq m+1.$$

Fasst man alle Abschätzungen zusammen, dann erhält man für den Interpolationsfehler

$$\begin{aligned}\left\|(v - I_K v)^{(k)}\right\|^p_{L^p(K)} &\leq c h_K^{-kp+1} \left\|(\hat{v} - I_{\hat{K}}\hat{v})^{(k)}\right\|^p_{L^p(\hat{K})} \\ &\leq c h_K^{-kp+1} \left\|\hat{v}^{(m+1)}\right\|^p_{L^p(\hat{K})} \\ &\leq c h_K^{(m+1-k)p} \left\|v^{(m+1)}\right\|^p_{L^p(K)}.\end{aligned}$$

Damit ist die Interpolationsfehlerabschätzung für eine beliebige Gitterzelle gezeigt. ■

Bemerkung 4.31 Uniforme Triangulierung. Sei eine uniforme Triangulierung mit $h_K = h$ für alle Gitterzellen gegeben. Dann erhält man durch Summation über die Gitterzellen die Interpolationsfehlerabschätzung für den globalen Finite–Element–Raum

$$\begin{aligned}\left\|(v - I_h v)^{(k)}\right\|_{L^p(a,b)} &= \left(\sum_{K \in \mathcal{T}_h} \left\|(v - I_K v)^{(k)}\right\|^p_{L^p(K)}\right)^{1/p} \\ &\leq \left(\sum_{K \in \mathcal{T}_h} c h_K^{(m+1-k)p} \left\|v^{(m+1)}\right\|^p_{L^p(K)}\right)^{1/p} \\ &\leq c h^{(m+1-k)} \left\|v^{(m+1)}\right\|_{L^p(a,b)}.\end{aligned}$$

Für lineare Finite–Elemente P_1 ($m = 1$) hat man beispielsweise die Abschätzungen

$$\|v - I_h v\|_{L^p(a,b)} \le ch^2 \|v''\|_{L^p(a,b)}, \quad \|(v - I_h v)'\|_{L^p(a,b)} \le ch \|v''\|_{L^p(a,b)},$$

für alle $v \in W^{2,p}(a,b)$. □

Bemerkung 4.32 Im singulär gestörten Fall ist die Interpolationsfehlerabschätzung noch nicht hinreichend um für große Gitterweiten gute Ergebnisse zu erhalten, weil der Faktor im Lemma von Cea sehr groß ist, siehe Bemerkung 4.6. Auf groben Gitter wird man deswegen große Fehler bekommen. □

Beispiel 4.33 Konvergenzordnung. Betrachte das Beispiel

$$-\varepsilon u'' + u' = 1 \quad \text{auf } (0,1), \quad u(0) = u(1) = 0.$$

Mit linearen Finite-Elementen erhält man folgende Fehler und Konvergenzordnungen im Fall $\varepsilon = 0.1$:

Int.	$\|(u-u_h)'\|_{L^2}$	Ord.	$\|u-u_h\|_{L^2}$	Ord.	$\|u-u_h\|_{L^\infty}$	Ord.
2	3.8323		0.53506		0.75669	
4	2.399	0.67576	0.096706	2.468	0.19332	1.9687
8	1.3309	0.8501	0.022568	2.0993	0.055709	1.795
16	0.68964	0.94844	0.0053014	2.0899	0.012119	2.2006
32	0.34823	0.98581	0.0012958	2.0325	0.0030185	2.0054
64	0.17455	0.99636	0.00032195	2.009	0.00074843	2.0119
128	0.087332	0.99908	8.0358e-5	2.0023	0.00018708	2.0003
256	0.043673	0.99977	2.0082e-5	2.0006	4.6746e-5	2.0007
512	0.021837	0.99994	5.0199e-6	2.0001	1.1686e-5	2
1024	0.010919	0.99999	1.2549e-6	2	2.9215e-6	2
2048	0.0054594	1	3.1373e-7	2	7.3038e-7	2
4096	0.0027297	1	7.8426e-8	2.0001	1.8259e-7	2

Für $\|u - u_h\|_{L^2}$ und $\|(u-u_h)'\|_{L^2}$ sind das genau die Ordnungen, die von der Theorie vorhergesagt werden. Im singulär gestörten Fall, $\varepsilon = 10^{-6}$, erhält man folgende Ergebnisse:

Int.	$\|(u-u_h)'\|_{L^2}$	Ord.	$\|u-u_h\|_{L^2}$	Ord.	$\|u-u_h\|_{L^\infty}$	Ord.
2	4.3301e+5		88388		1.25e+5	
4	3.3072e+5	0.38881	22097	2	31250	2
8	2.4206e+5	0.45024	5523.9	2.0001	7812.4	2
16	1.7399e+5	0.47636	1380.7	2.0003	1953.1	2
32	1.2402e+5	0.48848	344.91	2.0011	488.25	2
64	88037	0.49434	85.965	2.0044	122.06	2.0001
128	62370	0.49726	21.234	2.0174	30.52	1.9997
256	44139	0.49878	5.076	2.0646	7.6686	1.9927
512	31217	0.49972	1.132	2.1648	2.0758	1.8853
1024	22063	0.50073	0.35015	1.6928	1.0265	1.016
2048	15577	0.50219	0.17193	1.0262	0.99184	0.049515
4096	10981	0.5044	0.08561	1.0059	0.98375	0.011819

Diese schlechten Ergebnisse sind wegen des Zusammenhanges mit dem zentralen Differenzenverfahren, siehe Beispiel 4.15, zu erwarten gewesen. □

Bemerkung 4.34 Inverse Ungleichung. In diesem Abschnitt wird die Methode zum Beweis der Interpolationsfehlerabschätzung dazu verwendet, um sogenannte inverse Abschätzungen zu zeigen. Im Gegensatz zu Interpolationsfehlerabschätzungen wird dabei eine Norm einer höheren Ableitung einer Finite–Element–Funktion durch die Norm einer niederen Ableitung abgeschätzt. Man erhält als Faktor dann negative Potenzen des Durchmessers der Gitterzelle. □

Satz 4.35 Inverse Ungleichung. *Seien $0 \leq k \leq l$ natürliche Zahlen und $p, q \in [1, \infty]$. Dann gibt es eine Konstante c, die nur von $k, l, p, q, \hat{K}, \hat{P}(\hat{K})$ abhängt, mit*

$$\left\| v_h^{(l)} \right\|_{L^q(K)} \leq c h_K^{(k-l)-(p^{-1}-q^{-1})} \left\| v_h^{(k)} \right\|_{L^p(K)} \quad \forall\ v_h \in P(K).$$

Beweis: Zunächst wird die Abschätzung für $h_{\hat{K}} = c$ und $k = 0$ auf der Referenzzelle gezeigt. Da in einem endlichdimensionalen Raum alle Normen äquivalent sind, erhält man

$$\left\| \hat{v}_h^{(l)} \right\|_{L^q(\hat{K})} \leq \|\hat{v}_h\|_{W^{l,q}(\hat{K})} \leq c \|\hat{v}_h\|_{L^p(\hat{K})} \quad \forall\ \hat{v}_h \in \hat{P}(\hat{K}).$$

Im Falle $k > 0$ setzt man

$$\tilde{P}(\hat{K}) = \left\{ \hat{v}_h^{(k)} \ :\ \hat{v}_h \in \hat{P}(\hat{K}) \right\},$$

was gleichfalls ein Polynomraum ist. Wendet man die obige Abschätzung auf $\tilde{P}(\hat{K})$ an, erhält man

$$\left\| \hat{v}_h^{(l)} \right\|_{L^q(\hat{K})} = \left\| \left(\hat{v}_h^{(k)} \right)^{(l-k)} \right\|_{L^q(\hat{K})} \overset{\text{Normäquivalenz}}{\leq} c \left\| \hat{v}_h^{(k)} \right\|_{L^p(\hat{K})}.$$

Diese Abschätzung wird genauso wie in der Interpolationsfehlerabschätzung auf die Gitterzelle K transformiert. Aus den Abschätzungen für die Transformationen erhält man

$$\begin{aligned} \left\| v_h^{(l)} \right\|_{L^q(K)} &\leq c h_K^{-l+1/q} \left\| \hat{v}_h^{(l)} \right\|_{L^q(\hat{K})} \leq c h_K^{-l+1/q} \left\| \hat{v}_h^{(k)} \right\|_{L^p(\hat{K})} \\ &\leq c h_K^{k-l+1/q-1/p} \left\| v_h^{(k)} \right\|_{L^p(K)}. \end{aligned}$$

■

Übungsaufgabe: per Hand für gewisse FE nachrechnen.

Bemerkung 4.36

- Der springende Punkt im Beweis war die Äquivalenz aller Normen, eine Eigenschaft die bekanntlich bei unendlich–dimensionalen Räumen nicht gilt.
- Für $p = q$ überträgt sich die Abschätzung auf den globalen Finite–Element–Raum, sofern eine uniforme Triangulierung von (a, b) verwendet wird

$$\left\| v_h^{(l)} \right\|_{L_h^p(a,b)} \leq c h^{k-l} \left\| v_h^{(k)} \right\|_{L_h^p(a,b))}, \tag{4.10}$$

mit

$$\|\cdot\|_{L_h^p(a,b)} = \left(\sum_{K \in \mathcal{T}_h} \|\cdot\|_{L^p(K)}^p \right)^{1/p}.$$

Die zellenweise Normdefinition ist wichtig für $l \geq 2$, da dann die Finite–Element–Funktionen im allgemeinen nicht mehr die nötige Regularität für die globale Norm besitzen.

□

4.4 Stabilisierte Finite–Element–Methoden

Bemerkung 4.37 Zum Lemma von Lax–Milgram für singulär gestörte Probleme. Betrachte das Modellproblem: Finde $u \in V = H_0^1(0,1)$ so dass

$$a(u,v) = f(v) \quad \forall\, v \in V$$

mit

$$\begin{aligned} a(u,v) &:= \int_0^1 \Big(\varepsilon u'(x)v'(x) + b(x)u'(x)v(x) + c(x)u(x)v(x)\Big)\, dx, \\ f(v) &:= \int_0^1 f(x)v(x)\, dx. \end{aligned}$$

Sei

$$c(x) - \frac{b'(x)}{2} \geq \omega > 0 \quad \text{für alle } x \in [0,1].$$

Mit einer analogen Rechnung wie in Bemerkung 3.35 zeigt man, dass $a(\cdot,\cdot)$ koerzitiv bezüglich der von ε abhängigen Norm

$$\|v\|_\varepsilon^2 := \varepsilon\, |v|_{1,2}^2 + \|v\|_0^2 = \varepsilon\, \|v'\|_{L^2(0,1)}^2 + \|v\|_{L^2(0,1)}^2$$

ist. Das heißt, es existiert eine von ε unabhängige Konstante μ, so dass

$$a(v,v) \geq \mu\, \|v\|_\varepsilon^2 \quad \forall\, v \in V$$

gilt. Mit partieller Integration (*Übungsaufgabe*) zeigt man, dass es eine von ε unabhängige Konstante β gibt, so dass

$$|a(v,w)| \leq \beta\, \|v\|_\varepsilon\, \|w\|_{H^1} \quad \forall\, (v,w) \in V \times V.$$

Es gibt jedoch keine von ε unabhängige Konstante γ mit

$$|a(v,w)| \leq \gamma\, \|v\|_\varepsilon\, \|w\|_\varepsilon \quad \forall\, (v,w) \in V \times V.$$

Nutzt man die Abschätzungen mit Konstanten die unabhängig von ε sind, erhält man analog zum Beweis des Lemmas von Cea

$$\|u - u_h\|_\varepsilon \leq C \inf_{v_h \in V_h} \|u - v_h\|_{H^1}$$

mit C unabhängig von ε. Ist V^h ein Standard–Finite–Elemente–Raum (stückweise polynomial), dann kann man zeigen, dass in Grenzschichten

$$\inf_{v_h \in V_h} \|u - v_h\|_{H^1} \to \infty \quad \text{für} \quad \varepsilon \to 0$$

für festes h gilt. Deswegen hat man keine gleichmäßige Konvergenz $\|u - u_h\|_\varepsilon \to 0$ für $h \to 0$. □

4.4.1 Petrov–Galerkin–Methoden und Upwind–Verfahren

Bemerkung 4.38 Petrov[5]–Galerkin–Methode. Eine Finite–Element–Methode, bei welcher Ansatz– und Testraum unterschiedlich sind, wird Petrov–Galerkin–Methode genannt. Seien S_h der Ansatzraum und T_h der Testraum, mit $\dim(S_h) = \dim(T_h)$, dann lautet eine Petrov–Galerkin–Methode: Finde $u_h \in S_h$, so dass

$$a(u_h, v_h) = f(v_h) \quad \forall\, v_h \in T_h.$$

□

[5] Petrov

Beispiel 4.39 Petrov–Galerkin–Methode und Upwind–Verfahren. Betrachte

$$-\varepsilon u''(x) + bu'(x) = 0$$

mit $b \in \mathbb{R} \setminus \{0\}$. Nutze als Ansatzfunktionen stückweise lineare Funktionen

$$\phi_i(x) = \begin{cases} (x - x_{i-1})/h & \text{für } x \in [x_{i-1}, x_i], \\ (x_{i+1} - x)/h & \text{für } x \in [x_i, x_{i+1}], \\ 0 & \text{sonst,} \end{cases} \quad , \quad i = 1, \ldots, N-1.$$

Definiere die Blasenfunktion

$$\sigma_{i-1/2}(x) = \begin{cases} 4(x - x_{i-1})(x_i - x)/h^2 & \text{für } x \in [x_{i-1}, x_i], \\ 0 & \text{sonst.} \end{cases}$$

Die Testfunktionen werden nun als stückweise quadratische Funktionen definiert

$$\psi_i(x) = \phi_i(x) + \frac{3}{2}\kappa \left(\sigma_{i-1/2}(x) - \sigma_{i+1/2}(x)\right), \quad i = 1, \ldots, N-1,$$

wobei κ ein zu wählender Upwind–Parameter ist. Direktes Nachrechnen (*Übungsaufgabe*) zeigt, dass man damit das folgende Schema erhält

$$-\varepsilon D^+ D^- u_i + b\left[\left(\frac{1}{2} - \kappa\right) D^+ u_i + \left(\frac{1}{2} + \kappa\right) D^- u_i\right] = 0.$$

Wählt man $\kappa = \operatorname{sgn}(b)/2$, so erhält man das einfache Upwind–Finite–Differenzen–Verfahren, siehe Definition 2.32.

Eine Testfunktion $\psi_i(x)$, definiert in den Knoten $\{0, 0.5, 1\}$ für $\kappa = 1/2$ ist in Abbildung 4.4 dargestellt.

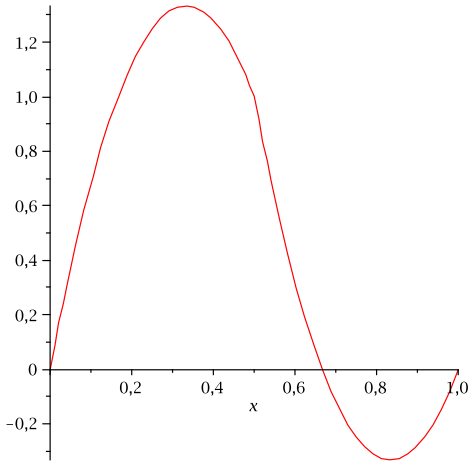


Abbildung 4.4: Stückweise quadratische Testfunktion.

□

Bemerkung 4.40 Man kann auch einige angepasste Upwind–Verfahren mit Hilfe von Petrov–Galerkin–Methoden gewinnen. Das funktioniert auch für nichtkonstante Koeffizientenfunktionen. Bessere Ergebnisse als etwa beim Iljin–Allen–Southwell–Verfahren, Satz 2.54, das heißt lineare Konvergenz, sind aber nicht zu erreichen. □

4.4.2 Die Stromlinien–Diffusions–Finite–Elemente–Methode

Bemerkung 4.41 Ziele. Das Ziel besteht darin, ein Verfahren zu konstruieren, welches stabiler als das Galerkin–Verfahren ist und welches für Finite–Elemente beliebiger Ordnung genutzt werden kann. Die Konvergenz dieses Verfahrens soll zudem von höherer Ordnung sein.

Es wird das Modellproblem

$$-\varepsilon u''(x) + b(x)u'(x) + c(x)u(x) = f(x) \text{ in } (0,1), \quad u(0) = u(1) = 0, \tag{4.11}$$

unter der Bedingung

$$c(x) - \frac{b'(x)}{2} \geq \omega > 0 \quad \text{für alle } x \in [0,1].$$

betrachtet. □

Bemerkung 4.42 Idee. Eine Idee zur Konstruktion eines stabileren Verfahrens besteht darin, zur Galerkin–Finite–Element–Methode gewichtete Residuen der starken Formulierung der Differentialgleichung (4.11) zu addieren. Dazu wird (4.11) mit $(bv')(x)$ multipliziert, über jedes Teilintervall (x_{i-1}, x_i), $i = 1, \ldots, N$, mit einem Gewicht versehen und integriert, und dann zur Galerkin–Methode addiert. □

Definition 4.43 Stromlinien–Diffusions–Finite–Elemente–Methode, SDFEM, Stromlinien–Upwind–Petrov–Galerkin FEM. Die Stromlinien–Diffusions–Finite–Elemente–Methode (SDFEM) oder Stromlinien–Upwind–Petrov–Galerkin (SUPG) FEM ist wie folgt definiert: Finde $u_h \in V_h$, so dass

$$a_h(u_h, v_h) = f_h(v_h) \quad \forall\ v_h \in V_h \tag{4.12}$$

gilt, mit

$$\begin{aligned} a_h(v,w) &:= \varepsilon(v', w') + (bv' + cv, w) \\ &\quad + \sum_{i=1}^{N} \int_{x_{i-1}}^{x_i} \delta_i \Big(-\varepsilon v''(x) + b(x)v'(x) + c(x)v(x) \Big) \Big(b(x)w'(x) \Big)\, dx, \\ f_h(w) &:= (f,w) + \sum_{i=1}^{N} \int_{x_{i-1}}^{x_i} \delta_i f(x) \Big(b(x)w'(x) \Big)\, dx. \end{aligned} \tag{4.13}$$

Dabei sind $\{\delta_i\}_{i=1}^N$ geeignet zu wählende Gewichte, welche SD–Parameter genannt werden. □

Bemerkung 4.44 Zur SDFEM.

- Der Name Stromlinien–Diffusion–FEM wird erst in höheren Dimensionen klar. Dort wird als Testfunktion für das Residuum die Ableitung in Konvektionsrichtung gewählt. Das ist die sogenannte Stromlinienrichtung.
- Für die Lösung der Galerkin–FEM wird sich das Residuum der starken Formulierung der Gleichung im allgemeinen stark von Null unterscheiden. Bei der SDFEM wird verlangt, dass dieses Residuum (in einem schwachen Sinne) nicht zu groß sein darf. Das Gewicht dieses Residuum in der SDFEM wird durch die SD–Parameter bestimmt.
- Für eine Finite–Elemente–Funktion ist die zweite Ableitung im allgemeinen nur stückweise definiert, und zwar innerhalb der Gitterzellen.
- Die SD–Parameter werden oft in jedem Intervall (x_{i-1}, x_i) konstant gewählt. Das Ziel der Analysis besteht darin, eine möglichst günstige Wahl dieser Parameter aufzuzeigen.

□

Beispiel 4.45 SDFEM für P_1. Betrachte $V_h = P_1$ auf einem äquidistanten Gitter mit $h_i = h$, $i = 1, \ldots, N$. Sind alle Koeffizientenfunktionen konstant, $c = 0$, und wählt man die SD–Parameter auch konstant, so reduziert sich die rechte Seite der SDFEM zu

$$\begin{aligned}\varepsilon(u_h', v_h') + (bu_h', v_h) + \sum_{i=1}^{N} \delta \int_{x_{i-1}}^{x_i} \Big(-\varepsilon \cdot 0 + bu_h'(x)\Big)\Big(bv_h'(x)\Big)\, dx \\ = \quad \varepsilon(u_h', v_h') + b(u_h', v_h) + \delta b^2 (u_h', v_h').\end{aligned}$$

Das entspricht der Galerkin FEM einer Gleichung mit rechter Seite

$$-\left(\varepsilon + \delta b^2\right) u''(x) + bu'(x).$$

Aus Beispiel 4.15 ist bekannt, dass die Galerkin FEM wiederum einem zentralen Differenzenverfahren entspricht. Die rechte Seite ist

$$(f, v_h) + \sum_{i=1}^{N} \delta \int_{x_{i-1}}^{x_i} fbv_h'(x)\, dx = (f, v_h) + \delta f b \underbrace{\sum_{i=1}^{N} \int_{x_{i-1}}^{x_i} v_h'(x)\, dx}_{=0} = (f, v_h).$$

Die Summe verschwindet, da sich jede Testfunktion $v_h'(x)$ als Linearkombination der Basisfunktionen $\{\phi_i(x)\}$ von P_1 schreiben lässt und das Integral über die Ableitung jeder Basisfunkion gleich Null ist.

Insgesamt entspricht die SDFEM unter den obigen Voraussetzungen dem angepassten Finite–Differenzen–Upwind–Verfahren (2.7)

$$-\varepsilon \left(1 + \delta \frac{b^2}{\varepsilon}\right) D^+ D^- u_i + bD^0 u_i, = f_i,$$

das heißt $\sigma(q) = 1 + \delta b^2/\varepsilon$, $q = bh/(2\varepsilon)$. Wählt man den SD–Parameter als

$$\delta(q) = \frac{h}{2b}\left(\coth(q) - \frac{1}{q}\right),$$

so ist

$$\sigma(q) = 1 + \frac{hb^2}{2b\varepsilon}\left(\coth(q) - \frac{1}{q}\right) = 1 + q\left(\coth(q) - \frac{1}{q}\right) = q\coth(q).$$

Damit erhält man das Iljin–Allen–Southwell–Verfahren.

Mit $\delta = h/(2b)$ erhält man das einfache Upwind–Verfahren.

Achtung: diese einfachen Zusammenhänge gelten in höheren Dimensionen nicht mehr! □

Definition 4.46 Konsistente Finite–Element–Methode. Sei $u(x)$ eine hinreichend glatte Lösung von (4.11). Eine Finite–Element–Methode: Finde $u_h \in V_h$, so dass

$$a_h(u_h, v_h) = f_h(v_h) \quad \forall\ v_h \in V_h,$$

wird konsistent genannt, wenn gilt

$$a_h(u, v_h) = f_h(v_h) \quad \forall\ v_h \in V_h. \tag{4.14}$$

□

Bemerkung 4.47 Konsistenz einer Finite–Element–Methode ist nicht das gleiche wie Konsistenz einer Finiten–Differenzen–Methode, siehe Definition 2.6. Für Finite–Element–Methoden bedeutet Konsistenz, dass eine hinreichend glatte Lösung auch die diskrete Gleichung erfüllt. □

Lemma 4.48 Galerkin–Orthogonalität. *Eine konsistente Finite–Element–Methode besitzt die Eigenschaft der Galerkin–Orthogonalität*

$$a_h(u - u_h, v_h) = 0 \quad \forall\ v_h \in V_h. \tag{4.15}$$

Man sagt auch, dass der Fehler „senkrecht“ auf dem Finite–Element–Raum steht.

Beweis: Die Aussage folgt sofort aus der Gültigkeit von (4.12) und (4.14) durch Subtraktion dieser beiden Gleichungen. ■

Lemma 4.49 Konsistenz der SDFEM. *Die SDFEM (4.12) – (4.13) ist konsistent.*

Beweis: Für eine hinreichend glatte Lösung $u(x)$ von (4.11) ist das Residuum der starken Form der Gleichung gleich Null. Damit verschwinden die SDFEM–Terme in (4.13). Durch partielle Integration erhält man aus den übrigen Termen, dass

$$\int_0^1 \Big(-\varepsilon u''(x) + b(x)u'(x) + c(x)u(x) - f(x)\Big) v_h(x)\ dx = 0 \quad \forall\ v_h \in V_h$$

gilt. Für eine hinreichend glatte Lösung verschwindet der Ausdruck in der Klammer und diese Aussage ist wahr. Damit ist die SDFEM konsistent. ■

Bei der Analysis stabilisierter FEM ist es wichtig, dass man geeignete Normen verwendet.

Definition 4.50 Stromlinien–Diffusions–Norm, SD–Norm. Auf V_h wird die Stromlinien–Diffusions–Norm

$$|||v_h|||_{SD} := \left(\varepsilon\, |v_h|_1^2 + \omega\, \|v_h\|_0^2 + \sum_{i=1}^{N} \left\| \sqrt{\delta_i} b v_h' \right\|_{0,I_i}^2 \right)^{1/2}$$

definiert. Hierbei ist $I_i := (x_{i-1}, x_i)$ und $\|\cdot\|_{0,I_i}$ ist die Norm in $L^2(I_i)$. □

Satz 4.51 Koerzitivität der SD–Bilinearform. *Sei*

$$0 < \delta_i \le \frac{1}{2} \min \left\{ \frac{h_i^2}{\varepsilon c_{\text{inv}}^2}, \frac{\omega}{\|c\|_{L^\infty(I_i)}^2} \right\}, \tag{4.16}$$

wobei c_{inv} die Konstante der inversen Ungleichung

$$\|v_h''\|_{0,I_i} \le c_{\text{inv}} h_i^{-1} \|v_h'\|_{0,I_i} \tag{4.17}$$

ist. Dann ist die SD–Bilinearform (4.13) koerzitiv bezüglich der SD–Norm, das heißt es gilt

$$a_h(v_h, v_h) \ge \frac{1}{2} |||v_h|||_{SD}^2 \quad \forall\ v_h \in V_h.$$

Beweis: Mit partieller Integration folgt, siehe Beispiel 3.35,

$$(bv_h' + cv_h, v_h) = \left(\left(-\frac{b'}{2} + c \right) v_h, v_h \right) \quad \forall\ v_h \in V_h.$$

Mit der Definition von ω ergibt sich

$$\begin{aligned} a_h(v_h,v_h) &= \varepsilon\,|v_h|_1^2 + \int_0^1 \underbrace{\left(c(x) - \frac{b'(x)}{2}\right)}_{\geq\omega>0} v_h^2(x)\;dx + \sum_{i=1}^N \left\|\sqrt{\delta_i}bv_h'\right\|_{0,I_i}^2 \\ &\quad + \sum_{i=1}^N \int_{x_{i-1}}^{x_i} \delta_i\left(-\varepsilon v_h''(x) + c(x)v_h(x)\right)\left(b(x)v_h'(x)\right)\;dx \\ &\geq |||v_h|||_{SD}^2 + \sum_{i=1}^N \int_{x_{i-1}}^{x_i} \delta_i\left(-\varepsilon v_h''(x) + c(x)v_h(x)\right)\left(b(x)v_h'(x)\right)\;dx. \end{aligned}$$

Nun wird der zweite Term nach oben abgeschätzt, womit man insgesamt eine Abschätzung nach unten erhält, wenn man die Abschätzung des zweiten Terms vom ersten Term subtrahiert. In der Abschätzung wird die Definition des SD–Parameters verwendet. Es ist

$$\begin{aligned} &\left|\int_{x_{i-1}}^{x_i} \delta_i\left(-\varepsilon v_h''(x) + c(x)v_h(x)\right)\left(b(x)v_h'(x)\right)\;dx\right| \\ &\leq \int_{x_{i-1}}^{x_i} \left(\delta_i^{1/2}\varepsilon\left|v_h''(x)\right|\right)\left(\delta_i^{1/2}\left|b(x)v_h'(x)\right|\right)\;dx \\ &\quad + \int_{x_{i-1}}^{x_i} \left(\delta_i^{1/2}\left|c(x)\right|\left|v_h(x)\right|\right)\left(\delta_i^{1/2}\left|b(x)v_h'(x)\right|\right)\;dx \\ &\overset{\text{CSU}}{\leq} \left(\delta_i^{1/2}\varepsilon\left\|v_h''\right\|_{0,I_i} + \delta_i^{1/2}\left\|c\right\|_{L^\infty(I_i)}\left\|v_h\right\|_{0,I_i}\right)\left\|\sqrt{\delta_i}bv_h'\right\|_{0,I_i} \\ &\overset{(4.17)}{\leq} \left(\delta_i^{1/2}\frac{\varepsilon c_{\text{inv}}}{h_i}\left\|v_h'\right\|_{0,I_i} + \delta_i^{1/2}\left\|c\right\|_{L^\infty(I_i)}\left\|v_h\right\|_{0,I_i}\right)\left\|\sqrt{\delta_i}bv_h'\right\|_{0,I_i} \\ &\overset{(4.16)}{\leq} \left(\frac{h_i}{\sqrt{2\varepsilon}c_{\text{inv}}}\frac{\varepsilon c_{\text{inv}}}{h_i}\left\|v_h'\right\|_{0,I_i} + \frac{\sqrt{\omega}}{\sqrt{2}\left\|c\right\|_{L^\infty(I_i)}}\left\|c\right\|_{L^\infty(I_i)}\left\|v_h\right\|_{0,I_i}\right)\left\|\sqrt{\delta_i}bv_h'\right\|_{0,I_i} \\ &= \left(\sqrt{\frac{\varepsilon}{2}}\left\|v_h'\right\|_{0,I_i} + \sqrt{\frac{\omega}{2}}\left\|v_h\right\|_{0,I_i}\right)\left\|\sqrt{\delta_i}bv_h'\right\|_{0,I_i} \\ &\overset{\text{Young Ugl.}}{\leq} \frac{\varepsilon}{2}\left\|v_h'\right\|_{0,I_i}^2 + \frac{1}{4}\left\|\sqrt{\delta_i}bv_h'\right\|_{0,I_i}^2 + \frac{\omega}{2}\left\|v_h\right\|_{0,I_i}^2 + \frac{1}{4}\left\|\sqrt{\delta_i}bv_h'\right\|_{0,I_i}^2 \\ &= \frac{1}{2}\left|||v_h|||_{SD,I_i}\right.. \end{aligned}$$

Summation über alle Gitterzellen und Einsetzen in die erste Abschätzung ergibt die Aussage des Satzes ■

Folgerung 4.52 Koerzitivität der SD–Bilinearform für lineare finite Elemente. *Für stückweise lineare finite Elemente ist die SD–Bilinearform (4.13) koerzitiv bezüglich der SD–Norm mit der Parameterwahl*

$$0 < \delta_i \leq \frac{\omega}{\left\|c\right\|_{L^\infty(I_i)}^2}. \tag{4.18}$$

Beweis: Der Beweis ist wie für Satz 4.51, wobei man ausnutzt, dass für stückweise lineare finite Elemente $v_h''(x) = 0$ in I_i, $i = 1,\ldots,N$, ist und die entsprechenden Terme im Beweis entfallen. ■

Bemerkung 4.53 Zur Koerzitivität.

- Der Beweis von Satz 4.51 ist typisch für die Untersuchung stabilisierter Finite–Element–Methoden. Man versucht die störenden Terme irgendwie mit der verwendeten Norm abzuschätzen. Das geht im allgemeinen nur, wenn man eine geeignete Norm verwendet. Insbesondere muss die Stabilisierung in dieser Norm irgendwie auftauchen.

- Aus Satz 4.51 folgt die Stabilität der SDFEM bezüglich der SD–Norm. Alle $v_h \in V_h$ erfüllen
$$|||v_h|||_{SD} \geq \min\{1, \omega\} \|v_h\|_\varepsilon .$$
Damit folgt, dass die SDFEM auch bezüglich der Norm $\|\cdot\|_\varepsilon$ stabil ist. Bezüglich $\|\cdot\|_\varepsilon$ ist auch die Galerkin–FEM stabil, jedoch nicht bezüglich $|||\cdot|||_{SD}$. Damit ist die Stabilitätsaussage von Satz 4.51 stärker als die Stabilitätsaussage für die Galerkin–FEM.

□

Beispiel 4.54 Fortsetzung: SDFEM für P_1. Im Beispiel 4.45 wurde gezeigt, dass man unter gewissen Bedingungen mit
$$\delta(q) = \frac{h}{2b}\left(\coth(q) - \frac{1}{q}\right), \quad q = \frac{bh}{2\varepsilon},$$
das Iljin–Allen–Southwell–Verfahren, also ein gleichmäßig konvergentes Verfahren, erhält. Man braucht aber auch Parameter im Falle von nichtkonstanten Koeffizientenfunktionen, Finite–Elementen höherer Ordnung und für Probleme in höheren Dimensionen. Dabei kann man versuchen, den Spezialfall zu verallgemeinern. Mit Taylor–Entwicklung, Übungsaufgabe, sieht man, dass
$$\begin{aligned} \coth q - \frac{1}{q} &= \frac{q}{3} + \mathcal{O}\left(q^3\right) \quad \text{für } q \to 0, \\ \coth q - \frac{1}{q} &= 1 + \mathcal{O}\left(\frac{1}{q}\right) \quad \text{für } q \to \infty. \end{aligned}$$
Ist ε konstant und geht $h \to 0$, so folgt $q \to 0$ und es ist $\delta(q) \approx hq/(6b)$. Für festes h und $\varepsilon \to 0$ folgt $q \to \infty$ und es ist $\delta(q) \approx h/(2b)$. Damit sind ist die folgende Wahl des SD–Parameters motiviert
$$\delta(q) = \begin{cases} \dfrac{h^2}{12\varepsilon} & \text{für } 0 < q \ll 1, \\ \dfrac{h}{2b} & \text{für } q \gg 1. \end{cases}$$
Falls das Gitter sehr grob im Vergleich zu ε ist, also $q \gg 1$, dann geht die SDFEM in 1D in das einfache Upwind–Verfahren über. □

Satz 4.55 Konvergenz des SDFEM. *Gelte für die Lösung von (4.11) $u \in H^{k+1}(0,1)$ und betrachte die SDFEM mit P_k–Finite–Elementen. Die SD–Parameter seien wie folgt gegeben*
$$\delta_i = \begin{cases} C_0 \dfrac{h_i^2}{\varepsilon} & \text{für } h_i < \varepsilon, \\ C_0 h_i & \text{für } \varepsilon \leq h_i, \end{cases} \tag{4.19}$$
wobei die Konstante $C_0 > 0$ klein genug ist, um (4.16) für $k \geq 2$ beziehungsweise (4.18) für $k = 1$ zu erfüllen. Dann erfüllt die Lösung $u_h \in P_k$ die Fehlerabschätzung
$$|||u - u_h|||_{SD} \leq C\left(\varepsilon^{1/2} h^k + h^{k+1/2}\right) |u|_{k+1}$$
mit einer von ε unabhängigen Konstanten C und $h = \max_{i=1,\ldots,N} h_i$.

Beweis: Sei $u^I \in V_h$ die Knoteninterpolierende von $u(x)$. Mit Dreiecksungleichung erhält man
$$|||u - u_h|||_{SD} \leq \left|\left|\left|u - u^I\right|\right|\right|_{SD} + \left|\left|\left|u^I - u_h\right|\right|\right|_{SD}.$$

Der erste Term auf der rechten Seite ist der Interpolationsfehler. Mit Hilfe der Interpolationsfehlerabschätzung (4.9), die man für jeden Term der SD–Norm anwendet, erhält man

$$\left|\left|\left|u-u^I\right|\right|\right|_{SD} \leq C\left(\varepsilon^{1/2}h^k + h^{k+1/2}\right)|u|_{k+1}\,.$$

Betrachte nun den zweiten Term auf der rechten Seite. Die Koerzitivität, Satz 4.51 und die Galerkin–Orthogonalität (4.15) ergeben

$$\frac{1}{2}\left|\left|\left|u^I-u_h\right|\right|\right|_{SD}^2 \leq a_h(u^I-u_h,u^I-u_h) = a_h(u^I-u,u^I-u_h).$$

Nun wird die Dreiecksungleichung auf $a_h(u^I-u,u^I-u_h)$ angewandt und dann jeder Term einzeln abgeschätzt. Wesentlich dabei ist die Interpolationsabschätzung (4.9). Sei $w_h = u^I - u_h$. Für den Diffusionsterm gilt

$$\begin{aligned}\left|\varepsilon\left((u^I-u)',w_h'\right)\right| &\overset{\text{CSU}}{\leq} \varepsilon\left\|(u^I-u)'\right\|_0\left\|w_h'\right\|_0 = \varepsilon^{1/2}\left\|(u^I-u)'\right\|_0\varepsilon^{1/2}\left\|w_h'\right\|_0\\ &\overset{(4.9)}{\leq} C\varepsilon^{1/2}h^k|u|_{k+1}\,ep^{1/2}\left\|w_h'\right\|_0 \leq C\varepsilon^{1/2}h^k|u|_{k+1}\,|||w_h|||_{SD}\,.\end{aligned}$$

Für den reaktiven Term erhält man auf ähnliche Art und Weise

$$\begin{aligned}\left|\left(c(u^I-u),w_h\right)\right| &\overset{\text{CSU}}{\leq} \|c\|_\infty\left\|u^I-u\right\|_0\|w_h\|_0 = \omega^{-1/2}\|c\|_\infty\left\|u^I-u\right\|_0\omega^{1/2}\|w_h\|_0\\ &\overset{(4.9)}{\leq} Ch^{k+1}|u|_{k+1}\,|||w_h|||_{SD}\,.\end{aligned}$$

Als nächstes werden die Terme betrachtet, die man bei der SDFEM–Stabilisierung erhält. Wegen $\varepsilon\delta_i \leq C_0h_i^2$ folgt

$$\begin{aligned}&\left|\sum_{i=1}^N\left(-\varepsilon(u^I-u)'',\delta_ibw_h'\right)\right|\\ &\overset{\text{CSU}}{\leq} \sum_{i=1}^N\varepsilon^{1/2}\left\|(u^I-u)''\right\|_{0,I_i}\varepsilon^{1/2}\delta_i^{1/2}\left\|\sqrt{\delta_i}bw_h'\right\|_{0,I_i}\\ &\leq C_0^{1/2}\sum_{i=1}^N h_i\varepsilon^{1/2}\left\|(u^I-u)''\right\|_{0,I_i}\left\|\sqrt{\delta_i}bw_h'\right\|_{0,I_i}\\ &\overset{\text{CSU}}{\leq} C_0^{1/2}\varepsilon^{1/2}h\left(\sum_{i=1}^N\left\|(u^I-u)''\right\|_{0,I_i}^2\right)^{1/2}\left(\sum_{i=1}^N\left\|\sqrt{\delta_i}bw_h'\right\|_{0,I_i}^2\right)^{1/2}\\ &\overset{(4.9)}{\leq} C\varepsilon^{1/2}h\left(\sum_{i=1}^N h_i^{2(k-1)}|u|_{k+1,I_i}^2\right)^{1/2}\left(\sum_{i=1}^N\left\|\sqrt{\delta_i}bw_h'\right\|_{0,I_i}^2\right)^{1/2}\\ &\leq C\varepsilon^{1/2}h^k|u|_{k+1}\,|||w_h|||_{SD}\,.\end{aligned}$$

Für die anderen Terme erhält man unter Nutzung von $\delta_i \leq C_0 h_i$

$$\begin{aligned}
&\left|\sum_{i=1}^{N}\left(b(u^I-u)'+c(u^I-u),\delta_i b w_h'\right)\right| \\
&\overset{\text{CSU}}{\leq} \quad \sum_{i=1}^{N}\|b\|_\infty \left\|(u^I-u)'\right\|_{0,I_i}\delta_i^{1/2}\left\|\sqrt{\delta_i}bw_h'\right\|_{0,I_i} ep^{1/2}\left\|w_h'\right\|_0 \\
&\qquad +\sum_{i=1}^{N}\|c\|_\infty \left\|(u^I-u)\right\|_{0,I_i}\delta_i^{1/2}\left\|\sqrt{\delta_i}bw_h'\right\|_{0,I_i} \\
&\leq \quad C\left(\sum_{i=1}^{N}h_i^{1/2}\left\|(u^I-u)'\right\|_{0,I_i}\left\|\sqrt{\delta_i}bw_h'\right\|_{0,I_i}\right. \\
&\qquad \left.+\sum_{i=1}^{N}h_i^{1/2}\left\|(u^I-u)\right\|_{0,I_i}\left\|\sqrt{\delta_i}bw_h'\right\|_{0,I_i}\right) \\
&\leq \quad Ch_i^{1/2}\left[\left(\sum_{i=1}^{N}\left\|(u^I-u)'\right\|_{0,I_i}^2\right)^{1/2}+\left(\sum_{i=1}^{N}\left\|(u^I-u)\right\|_{0,I_i}^2\right)^{1/2}\right]\left(\sum_{i=1}^{N}\left\|\sqrt{\delta_i}bw_h'\right\|_{0,I_i}^2\right)^{1/2} \\
&\overset{(4.9)}{\leq} \quad C\left(h^{k+1/2}+h^{k+3/2}\right)|u|_{k+1}\,|||w_h|||_{SD}\,.
\end{aligned}$$

Für eine optimale Abschätzung des konvektiven Terms muss man diesen erst partiell integrieren

$$\begin{aligned}
\left(b(u^I-u)',w_h\right) &= \left((u^I-u)',bw_h\right) = -\left((u^I-u),(bw_h)'\right) \\
&= -\left((u^I-u),b'w_h\right)-\left((u^I-u),bw_h'\right).
\end{aligned}$$

Nun schätzt man die letzten beiden Terme einzeln ab. Mit den gleichen Techniken wie bei den bisherigen Abschätzungen erhält man

$$\begin{aligned}
\left|\left((u^I-u),b'w_h\right)\right| &\leq \omega^{-1/2}\left\|b'\right\|_\infty\left(\sum_{i=1}^{N}\left\|u^I-u\right\|_{0,I_i}^2\right)^{1/2}\omega^{1/2}\left\|w_h\right\|_0 \\
&\leq Ch^{k+1}\,|u|_{k+1}\,|||w_h|||_{SD}\,.
\end{aligned}$$

Bei der Abschätzung des anderen Terms muss man unterscheiden, ob im Intervall I_i gilt $\varepsilon \leq h_i$ oder $\varepsilon > h_i$. Man erhält

$$\begin{aligned}
&\left|\left((u^I-u),bw_h'\right)\right| \\
&\overset{\text{CSU}}{\leq} \quad \sum_{\varepsilon\leq h_i}\delta_i^{-1/2}\left\|u^I-u\right\|_{0,I_i}\left\|\sqrt{\delta_i}bw_h'\right\|_{0,I_i}+\sum_{\varepsilon>h_i}\|b\|_\infty\left\|u^I-u\right\|_{0,I_i}\left\|w_h'\right\|_{0,I_i} \\
&\overset{(4.9)}{\leq} \quad C\left(\sum_{\varepsilon\leq h_i}\delta_i^{-1/2}h_i^{k+1}\,|u|_{k+1,I_i}\left\|\sqrt{\delta_i}bw_h'\right\|_{0,I_i}+\sum_{\varepsilon>h_i}h_i^{k+1}\,|u|_{k+1,I_i}\left\|w_h'\right\|_{0,I_i}\right) \\
&\overset{\delta_i\leq\ldots,\varepsilon>h_i}{\leq} \quad C\left(\sum_{\varepsilon\leq h_i}C_0^{-1/2}h_i^{-1/2}h_i^{k+1}\,|u|_{k+1,I_i}\left\|\sqrt{\delta_i}bw_h'\right\|_{0,I_i}+\sum_{\varepsilon>h_i}h_i^{k+1/2}\,|u|_{k+1,I_i}\,\varepsilon^{1/2}\left\|w_h'\right\|_{0,I_i}\right) \\
&\overset{\text{CSU}}{\leq} \quad Ch^{k+1/2}\,|u|_{k+1}\left[\left(\sum_{i=1}^{N}\left\|\sqrt{\delta_i}bw_h'\right\|_{0,I_i}^2\right)^{1/2}+\varepsilon\,|w_h|_1\right] \\
&\leq \quad Ch^{k+1/2}\,|u|_{k+1}\,|||w_h|||_{SD}\,.
\end{aligned}$$

Fasst man nun alle Abschätzungen zusammen, so erhält man die Aussage des Satzes. ■

Bemerkung 4.56 Zur Konvergenzabschätzung. Wesentlich für die Abschätzung mit einer von ε unabhängigen Konstanten C ist, dass der Term

$$\left(\sum_{i=1}^{N} \left\| \sqrt{\delta_i} b w_h' \right\|_{0,I_i}^2 \right)^{1/2}$$

Bestandteil der Norm ist, in der man den Fehler abschätzt. Eine solche Abschätzung gilt für die von ε abhängigen Norm $\|\cdot\|_\varepsilon$ nicht. □

Beispiel 4.57 SDFEM. Das Standardbeispiel

$$-\varepsilon u'' + u' = 1 \quad \text{auf } (0,1), \quad u(0) = u(1) = 0,$$

passt nicht in die Theorie der SDFEM, da $c(x) - \frac{b'(x)}{2} = 0$ ist. Trotzdem kann man auch auf dieses Beispiel die SDFEM–Stabilisierung anwenden. Man hat allerdings nicht die in Satz 4.55 bewiesenen Konvergenzordnungen.

Ein grundlegendes Problem der Anwendung der SDFEM ist die freie Konstante C_0 in der Parameterdefinition (4.19). Am Standardbeispiel sieht man sehr gut, dass man für verschiedene Konstanten stark unterschiedliche Ergebnisse erhält, siehe Abbildung 4.5. Ist C_0 zu groß, dann ist die Grenzschicht verschmiert, für ein geeignetes C_0 findet man eine Lösung die (fast) knotenexakt ist, und ist C_0 zu klein, dann entstehen an der Grenzschicht unphysikalische Oszillationen.

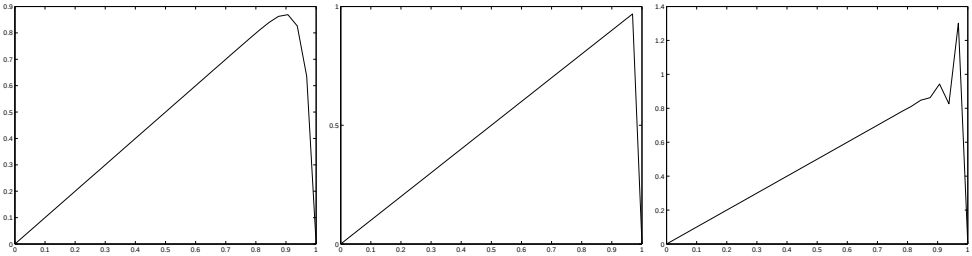


Abbildung 4.5: Mit der SDFEM berechnete Ergebnisse für das Standardbeispiel, $C_0 = 1$, $C_0 = 0.5$, $C_0 = 0.25$ von links nach rechts, $h = 1/32$, P_1 Finites–Element.

Für allgemeine Probleme ist es schwierig, C_0 geeignet zu wählen. In höheren Dimensionen wird man im allgemeinen auch kein C_0 mehr finden, so dass man eine (fast) knotenexakte Lösung erhält. Dafür besitzen mit der SDFEM berechnete Lösungen in höheren Dimensionen im allgemeinen unphysikalische Oszillationen an Grenzschichten. □

Bemerkung 4.58 Andere Wahl des SDFEM–Parameters. Man nimmt auch statt (4.19) den Parameter aus Beispiel 4.54

$$\delta_i = \frac{h_i}{2 \|b\|_{L^\infty(I_i)}} \left(\coth(\mathrm{Pe}_i) - \frac{1}{\mathrm{Pe}_i} \right), \quad \mathrm{Pe}_i = \frac{\|b\|_{L^\infty(I_i)} h_i}{2\varepsilon},$$

wobei Pe_K die lokale Péclet–Zahl ist. In dieser Definition hat man zwar keinen freien Parameter mehr, aber man stellt fest, dass bei Gleichungen in höheren Dimensionen unphysikalische Oszillationen in den berechneten Lösungen auftreten. □

Kapitel 5

Finite–Volumen–Methoden

Bemerkung 5.1 Grundlegende Idee. Finite–Volumen–Methoden (FVM) basieren auf Integralbilanzen über sogenannten Kontrollvolumen. Dazu wird das Intervall zunächst in kleine Gebiete, eben diese Kontrollvolumen, zerlegt und die Differentialgleichung wird über jedem Kontrollvolumen integriert. Im Anschluss wird partielle Integration (Gaußscher Satz in mehreren Dimensionen) angewandt, um die Integrale über den Kontrollvolumen, die Ableitungen enthalten, in Integrale auf dem Rand der Kontrollvolumen zu überführen. In einer Dimension, sind das Punktwerte in den Randpunkten. Dann verwendet man geeignete Approximationen für die Randintegrale, womit man ein Differenzenverfahren erhält.

Die Integralbilanzen können oft als Erhaltungsgesetze für physikalische Größen interpretiert werden. Deshalb werden Finite–Volumen–Methoden vor allem bei solchen Problemen mit Erfolg verwendet, bei denen die Erhaltung von Größen sehr wichtig ist, da diese Verfahren die Erhaltungseigenschaft bei der Approximation bewahren. Ein Beispiel sind inkompressible Strömungen, bei denen die Masse des Fluids in einem festen Strömungsgebiet konstant ist. □

Beispiel 5.2 Betrachte

$$-\varepsilon u''(x) + (b(x)u(x))' + c(x)u(x) = f(x) \text{ für } x \in (0,1), \quad u(0) = u(1) = 0,$$

mit $b(x) \geq \beta > 0$ und $c(x) \geq 0$. Das Intervall wird mit Hilfe eines Gitters mit den Gitterpunkten $0 = x_0, \ldots, x_N = 1$ zerlegt. Der Einfachheit halber sei das Gitter äquidistant mit Gitterweite h.

Finite–Volumen–Methoden benötigen ein Zweit–Gitter (secondary grid). Dieses wird in einer Dimension mit Hilfe der Mittelpunkte der Teilintervalle definiert. Setze

$$x_{i+1/2} := \frac{x_i + x_{i-1}}{2}, \quad i = 0, \ldots, N-1.$$

Die Kontrollvolumen werden mit Hilfe des Zweit–Gitters definiert

$$(0, x_{1/2}), (x_{1/2}, x_{3/2}), \ldots, (x_{N-1/2}, 1).$$

Integration der Gleichung über ein Kontrollvolumen ergibt

$$\begin{aligned}
&\int_{x_{i-1/2}}^{x_{i+1/2}} \Big(-\varepsilon u''(x) + (b(x)u(x))' + c(x)u(x) \Big)\, dx \\
&= \quad -\varepsilon u'(x)|_{x_{i+1/2}}^{x_{i-1/2}} + (bu)(x_{i+1/2}) - (bu)(x_{i-1/2}) + \int_{x_{i-1/2}}^{x_{i+1/2}} c(x)u(x)\, dx \quad (5.1) \\
&= \quad \int_{x_{i-1/2}}^{x_{i+1/2}} f(x)\, dx.
\end{aligned}$$

Die Terme werden nun durch Werte auf dem Originalgitter approximiert: die Ableitungen im ersten Term auf der linken Seite durch Differenzenquotienten, die Funktionswerte durch Mittelwerte und die Integrale durch Quadraturformeln. Mögliche Varianten sind

$$\begin{aligned} u'(x_{i+1/2}) &\approx \frac{u_{i+1}^N - u_i^N}{h}, & u'(x_{i-1/2}) &\approx \frac{u_i^N - u_{i-1}^N}{h}, \\ g(x_{i\pm 1/2}) &\approx \frac{g(x_i) + g(x_{i\pm 1})}{2} & \int_{x_{i-1/2}}^{x_{i+1/2}} g(x)\, dx &\approx g(x_i)h. \end{aligned}$$

Dafür braucht man vernünftige Approximationen für $u'(0)$ und $u'(1)$.

Für konstantes $b(x)$ erhält man mit diesen Approximationen

$$-\varepsilon \left(\frac{u_{i+1}^N - u_i^N}{h} - \frac{u_i^N - u_{i-1}^N}{h} \right) + b \left(\frac{u_i^N + u_{i+1}^N}{2} - \frac{u_i^N + u_{i-1}^N}{2} \right) + c_i h u_i^N = f_i h$$

mit $c_i = c(x_i), f_i = f(x_i)$. Das ist äquivalent zum zentralen Differenzenschema

$$-\varepsilon \frac{u_{i+1}^N - 2u_i^N + u_{i-1}^N}{h^2} + b \frac{u_{i+1}^N - u_{i-1}^N}{2h} + c_i u_i^N = f_i.$$

□

Bemerkung 5.3 Cell–centered Finite–Volumen–Methoden. Finite–Volumen–Methoden, welche ein Zweit–Grid nutzen um die Kontrollvolumen zu definieren, werden cell–centered Finite–Volumen–Methoden genannt. Man kann auch das Originalgitter zur Definition der Kontrollvolumen verwenden. Diese Methoden heißen dann cell–vertex Finite–Volumen–Methoden. Die letzteren Methoden sind aber nicht besonders populär, da sie instabil sind. □

Bemerkung 5.4 Finite–Volumen–Methoden für singulär gestörte Probleme. Um für singulär gestörte Probleme eine stabile Finite–Volumen–Methode zu erhalten, muss man den Konvektionsterm $(bu)(x_{i\pm 1/2})$ in (5.1) durch einen Upwind–Term approximieren, zum Beispiel durch

$$(bu)(x_{i+1/2}) \approx b(x_{i+1/2}) \left(\lambda_i u_{i+1}^N + (1-\lambda_i) u_i^N \right),$$

mit $\lambda_i \in [0, 1/2]$. Für $\lambda_i = 1/2$ erhält man das zentrale Differenzenschema und für $\lambda_i = 0$ das einfache Upwind–Verfahren aus Definition 2.32. Mit Werten zwischen 0 und 1/2 kann man die Größe des Upwindings variieren.

Seien $b(x)$ und $\lambda_i = \lambda$ konstant. Dann erhält man mit der Upwind–Approximation

$$\begin{aligned} &(bu)(x_{i+1/2}) - (bu)(x_{i-1/2}) \\ &\quad\approx b\left(\left(\lambda u_{i+1}^N + (1-\lambda) u_i^N \right) - \left(\lambda u_i^N + (1-\lambda) u_{i-1}^N \right) \right) \\ &\quad= b\left(\lambda u_{i+1}^N + (1-2\lambda) u_i^N - (1-\lambda) u_{i-1}^N \right) \\ &\quad= b\left(\frac{u_{i+1}^N - u_{i-1}^N}{2} + \left(\lambda - \frac{1}{2} \right) u_{i+1}^N + (1-2\lambda) u_i^N - \left(\frac{1}{2} - \lambda \right) u_{i-1}^N \right) \\ &\quad= b\left(\frac{u_{i+1}^N - u_{i-1}^N}{2} \right) - \frac{bh\,(1-2\lambda)}{2} \left(\frac{u_{i+1}^N - 2u_i^N + u_{i-1}^N}{h} \right). \end{aligned}$$

Nun kann das stabilisierte Finite–Volumen–Verfahren als angepasstes Upwind–Verfahren (2.7) mit

$$\sigma(q) = 1 + q\,(1-2\lambda), \quad q = \frac{bh}{2\varepsilon},$$

interpretiert werden. Damit übertragen sich auch alle Eigenschaften von angepassten Upwind–Verfahren auf diese stabilisierte Finite–Volumen–Methode.

Insbesondere ist es auch möglich, dass Iljin–Allen–Southwell–Verfahren aus Definition 2.53 mit Hilfe einer Finiten–Volumen–Methode zu generieren, siehe [RST08]. □

Bemerkung 5.5 Finite–Volumen–Methoden in höheren Dimensionen. Anders als in einer Dimension, sind Finite–Volumen–Methoden in höheren Dimensionen grundsätzlich von Finite–Differenzen–Methoden und Finite–Element–Methoden verschieden ! □

Kapitel 6

Zusammenfassung und Ausblick

Bemerkung 6.1 Verfahren. Zur Diskretisierung von partiellen Differentialgleichungen gibt es im wesentlichen drei Verfahren:

- Finite–Differenzen–Methoden:
  - approximieren die Ableitungen der starken Form der Gleichung mit Hilfe von Differenzenquotienten,
  - einfach zu verstehen und zu implementieren,
  - Taylor–Entwicklung wesentlich in der Analysis,
- Finite–Element–Methoden:
  - basieren auf der schwachen (variationellen) Formulierung der zu Grunde liegenden Gleichung in Sobolev–Räumen,
  - approximieren den unendlich–dimensionalen Sobolev–Raum durch einen endlich–dimensionalen Raum,
  - Analyis basiert auf Konzepten der Funktionalanalysis,
- Finite–Volumen–Methoden:
  - basiert auf Integration der zu Grunde liegenden Gleichung,
  - sichert die Erhaltung von Größen in Kontrollvolumen.

□

Bemerkung 6.2 Dimension.

- In einer Dimension lassen sich die Verfahren oft ineinander überführen.
- In höheren Dimensionen sind die drei Herangehensweisen grundsätzlich verschieden. Alle Verfahren besitzen Vor– und Nachteile, zum Beispiel:
  - in komplizierten Gebieten sind Finite–Elemente und Finite–Volumen flexibler als Finite–Differenzen,
  - die Implementierung von Finite–Differenzen ist wesentlich aufwändiger als die von Finite–Differenzen und Finite–Volumen,
  - Finite–Volumen–Methoden sind dort erfolgreich, wo man Erhaltungssätze erfüllen muss,
  - für Finite–Element–Methoden ist die Theorie am weitesten entwickelt.
- Ein neues Problem in höheren Dimensionen ist, dass komplizierte Gebiete auftreten können. Das hat sowohl Auswirkungen in der Analysis (Regularität der Lösung) als auch in der Praxis (Gittergenerierung).
- Die Gitterzellen in d Dimensionen sind d–dimensional. Diese Gitterzellen müssen geeignet angeordnet werden, damit ein vernünftiges Gitter entsteht. Das ist insbesondere bei komplizierten Gebieten nicht trivial. Gittergenerie-

rung, insbesondere in drei Dimensionen, ist ein wichtiges Forschungsgebiet.

□

Bemerkung 6.3 Singulär gestörte Probleme. Standard–Diskretisierungen berechnen nutzlose Lösungen für singulär gestörte Probleme, schon bei konstanten Koeffizienten. Man benötigt geeignete Stabilisierungen.

- In einer Dimension findet man Verfahren, um sehr gute Lösungen zu erhalten, zum Beispiel das Iljin–Allen–Southwell–Verfahren.
- In höheren Dimensionen ist die Entwicklung geeigneter stabiler Verfahren ein aktueller Forschungsgegenstand. Die bisher entwickelten Verfahren führen oft zu nicht zufriedenstellenden Lösungen (Grenzschichten zu stark verschmiert, unphysikalische Oszillationen).

□

Literaturverzeichnis

[AK90] O. Axelsson and L. Kolotilina. Monotonicity and discretization error estimates. *SIAM J. Numer. Anal.*, 27:1591 – 1611, 1990.

[Boh81] E. Bohl. *Finite Modelle gewöhnlicher Randwertaufgaben*. Teubner, Stuttgart, 1981.

[Emm04] E. Emmrich. *Gewöhnliche und Operator–Differentialgleichungen*. Vieweg, 2004.

[Goe77] H. Goering. *Asymptotische Methoden zur Lösung von Differentialgleichungsproblemen*, volume 144 of *Wissenschaftliche Taschenbücher, Reihe Mathematik und Physik*. Akademie–Verlag, Berlin, 1977.

[GR05] C. Großmann and H.-G. Roos. *Numerische Behandlung partieller Differentialgleichungen*. Teubner Studienbücher Mathematik. Teubner Verlag 2005, 3. edition, 2005.

[GT83] D. Gilbarg and N.S. Trudinger. *Elliptic Partial Differential Equations of Second Order*, volume 224 of *Grundlehren der mathematischen Wissenschaften*. Springer, 1983.

[KT78] B. Kellogg and A. Tsan. Analysis of some difference approximations for a singularly perturbed problem without turning points. *Math. Comp.*, 32:1025 – 1039, 1978.

[RST08] H.-G. Roos, M. Stynes, and L. Tobiska. *Robust Numerical Methods for Singularly Perturbed Differential Equations*, volume 24 of *Springer Series in Computational Mathematics*. Springer, 2nd edition, 2008.

[Sol06] P. Solin. *Partial Differential Equations and the Finite Element Method*. Pure and Applied Mathematics. Wiley – Interscience, 2006.

Index