Numerik partieller Differentialgleichungen

Herbert Egger, Stefan Takacs

Wintersemester 2023/24

Inhaltsverzeichnis

1. Differentialgleichungen als mathematische Modelle

Differentialgleichungen spielen eine große Rolle für die mathematische Modellierung verschiedener Sachverhalte, nicht nur in Technik und Naturwissenschaften. In den folgenden Abschnitten dieses Kapitels wollen wir einige Beispiele geben, die als kurze Motivation für die in dieser Vorlesung betrachteten Probleme dienen sollen.

Wachstumsprozesse

Aus Beobachtungen im Labor weiß man, dass sich Bakterien in einer Petri-Schale nach folgender Gesetzmäßigkeit vermehren

$$u(t+h) \approx (1+\alpha h)\, u(t). \tag{1.1}$$

Hierbei ist $u(t)$ die Anzahl der Bakterien zum Zeitpunkt t, α die Wachstumsrate und h die Zeit, die man gewartet hat. (Wir verwenden hier, wie auch sonst im Skriptum für die gesuchte *unbekannte* Funktion den Kleinbuchstaben u.) Im Grenzübergang $h \to 0$ erhält man bei Vernachlässigung der Modellierungsfehler

$$u'(t) = \alpha u(t), \quad t > 0. \tag{1.2}$$

Die *Anfangsbedingung*

$$u(0) = u_0 \tag{1.3}$$

beschreibt die Größe der Bakterienpopulation zum Zeitpunkt $t = 0$. Das *Anfangswertproblem* (1.2)–(1.3) ist eindeutig lösbar; die Lösung ist durch die Formel

$$u(t) = u_0 e^{\alpha t}$$

gegeben. Das Anfangswertproblem beschreibt also ein exponentielles Wachstum einer Bakterienpopulation in einer Petri-Schale.

Bemerkung: Aus dem Gesetz (1.1) erhält man sofort eine numerische Methode zur näherungsweisen Berechnung der Lösung $u(t)$, und zwar

$$u_{i+1} = u_i + h\alpha u_i, \qquad t_0 = 0,\ t_{i+1} = t_i + h, \tag{1.4}$$

wobei $h > 0$ und $u_i \approx u(t_i)$ eine Näherung für die Lösung zum Zeitpunkt t_i darstellt. Die Punkte (t_i, u_i) lassen sich zu einem Polygonzug verbinden, den wir mit u_h bezeichnen:

u_h stetig und auf linear auf allen Teilintervallen $[t_{i,i+1}]$, sodass $u_h(t_i) = u_i$.

Das Verfahren (1.4) ist ein Spezialfall des Euler'schen Polygonzugverfahrens (kurz: Eulerverfahren, explizites Eulerverfahren) und lässt sich sehr leicht am Computer implementieren. Aus numerischen Experimenten erkennt man, dass

$$\max_{t\in[0,T]} |u_h(t) - u(t)| \leq Ch = \mathcal{O}(h)$$

gilt. Das numerische Verfahren konvergiert hier also mit Ordnung 1 in h gegen die exakte Lösung.

Logistisches Wachstum. Genauere Beobachtungen zeigen, dass die Zahl der Bakterien nicht unbeschränkt wachsen kann. Aufgrund von Platz- und Nährstoffmangel geht das Wachstum bei großer Population zurück. Dies lässt sich durch eine Wachstumsrate $\alpha(n)$ modellieren, die von der Populationsgröße abhängt. Durch Vergleich mit experimentellen Daten kann man auf ein Gesetz der Form

$$\alpha(n) = \alpha_{max}\left(1 - \frac{n}{n_{max}}\right), \qquad \alpha_{max}, n_{max} > 0, \tag{1.5}$$

schließen. Für eine kleine Population gilt $\alpha(n) \approx \alpha_{max}$, wie beim exponentiellen Wachstum. Bei $n \approx n_{max}$ tritt praktisch kein Wachstum auf. Für sehr große Population ($n > n_{max}$) wird die Wachstumsrate sogar negativ.

Mit denselben Überlegungen wie beim exponentiellen Wachstum lässt sich die zeitliche Entwicklung der Bakterienpopulation nun durch das *Anfangswertproblem*

$$u'(t) = \alpha_{max}\left(1 - \frac{u(t)}{n_{max}}\right)u(t), \quad t > 0, \qquad u(0) = u_0 \tag{1.6}$$

beschreiben. Auch für diesen Fall können wir noch eine analytische Lösung bestimmen. Diese lautet

$$u(t) = \frac{u_0 n_{max} e^{\alpha_{max} t}}{n_{max} + (e^{\alpha_{max} t} - 1)u_0}.$$

Zur näherungsweisen Berechnung der Lösung lässt sich wieder das Euler-Verfahren anwenden. Wir ersetzen $u(t_i)$ durch die Annäherung u_i und $u'(t_i)$ durch den Vorwärtsdifferenzenquotienten $\frac{1}{h}(u_{i+1} - u_i)$ und erhalten:

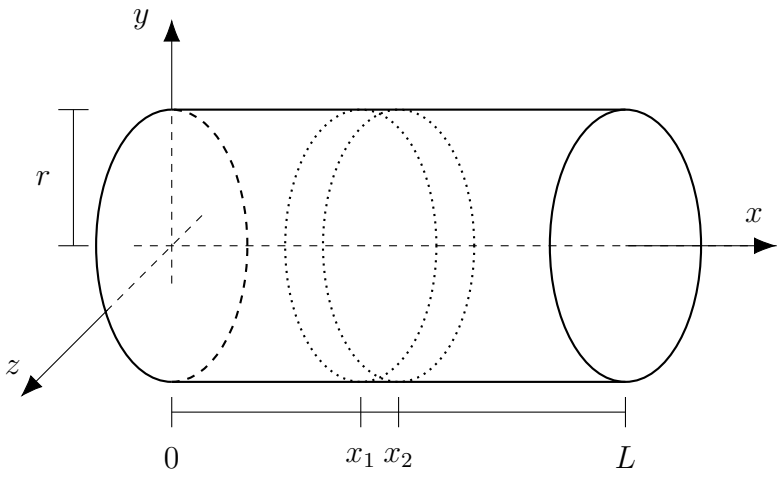
$$u_{i+1} = u_i + h\alpha_{max}\left(1 - \frac{u_i}{n_{max}}\right)u_i, \qquad t_0 = 0,\ t_{i+1} = t_i + h. \tag{1.7}$$

Wie beim exponentiellen Wachstum konvergieren die Näherungslösungen u_h auf beschränkten Intervallen $[0, T]$ wieder mit Rate $\mathcal{O}(h)$ gegen die exakte Lösung des Problems.

Im Verlauf der Vorlesung wollen wir herausfinden, unter welchen Voraussetzungen wir eine solche Konvergenz erzielen können. Dazu werden wir uns mit Konvergenztheorie auseinandersetzen.

Wärmeleitung

Wir sind an der Wärmeverteilung in einem zylindrischen Stab der Länge L und Radius r während eines Zeitintervalls $[0, T]$ interessiert.



Der Stab wird durch die Menge $\Omega := (0, L) \times S$, wobei $S := \{(y, z) : y^2 + z^2 < r\}$ der Querschnitt ist, modelliert. Wir interessieren uns für den Fall, dass der Radius r klein ist ($r \ll L$) und *nehmen der Einfachheit halber an*, dass die Temperatur u über den gesamten Querschnitt gleich ist:

$$u(x, y, z, t) = u(x, t).$$

Erhaltungsgleichung. Wir betrachten nun ein Stabsstück $(x_1, x_2) \times S$ der Länge $h := x_2 - x_1$ während der Zeitspanne (t_1, t_2) der Länge $\tau := t_2 - t_1$.

Die Änderung der Wärmemenge über das Zeitintervall (t_1, t_2) ist gleich der Wärmemenge, die im Stabsstück $\omega := (x_1, x_2) \times S$ während des Zeitintervalls (t_1, t_2) durch Aufheizung entsteht, abzüglich der Wärmemenge, die während des Zeitintervalls (t_1, t_2) über den Mantel $(x_1, x_2) \times \partial S$ und über die Seitenflächen

$\{x_1\} \times S$ und $\{x_2\} \times S$ abfließt (Wärmefluss). Wir erhalten also

$$\int_\omega c\rho(u(x,t_2) - u(x,t_1))\,\mathrm{d}\vec{x} = \int_{t_1}^{t_2} \int_\omega f_0(x,t)\mathrm{d}\vec{x}\mathrm{d}t$$
$$- \int_{t_1}^{t_2} \int_{x_1}^{x_2} \int_{\partial S} \sigma_0(x,t)\,\mathrm{d}(y,z)\,\mathrm{d}x\,\mathrm{d}t - \int_{t_1}^{t_2} \int_S \sigma_1(x_1,t) + \sigma_2(x_2,t)\,\mathrm{d}(y,z)\,\mathrm{d}t,$$

wobei c die Wärmekapazität, ρ die Dichte, f_0 die Energiedichte der Aufheizung und σ_0, σ_1 und σ_2 jeweils den Wärmefluss über Mantel sowie Seitenflächen beschreibt. Eine Vereinfachung und die Konvention, dass die wir die unbekannten Beiträge auf die linke Seite und die bekannten Beiträge auf die rechte Seite der Gleichung schreiben ergibt:

$$|S| \int_{x_1}^{x_2} c\rho(u(x,t_2) - u(x,t_1))\,\mathrm{d}x - |\partial S| \int_{t_1}^{t_2} \int_{x_1}^{x_2} \sigma_0(x,t)\,\mathrm{d}x\,\mathrm{d}t$$
$$- |S| \int_{t_1}^{t_2} \sigma_1(x_1,t) + \sigma_2(x_2,t)\,\mathrm{d}t = |S| \int_{t_1}^{t_2} \int_{x_1}^{x_2} f_0(x,t)\mathrm{d}x\mathrm{d}t.$$

Materialgesetz. Zur Bestimmung des Wärmeflusses benötigen wir ein *Materialgesetz*. Nach dem *Fourier'schen Gesetz* ist der Wärmefluss über die Seitenfläche (σ_1, σ_2) proportional zum Temperaturgradienten in Richtung des äußeren Normalvektors n. So erhalten wir für die Wärmemenge, die durch die linke Seitenflächen fließt, den Wärmefluss

$$\sigma_1(x_1,t) = \lambda\partial_n u(x_1,t) = -\lambda\partial_x u(x_1,t), \quad \sigma_2(x_2,t) = \lambda\partial_n u(x_2,t) = \lambda\partial_x u(x_2,t),$$

wobei λ der Wärmeleitkoeffizient ist und ∂_x und ∂_n die Richtungsableitungen bezeichnen.

Der Wärmefluss über den Mantel nehmen wir als proportional zur Temperaturdifferenz an:

$$\sigma_0 = -\alpha(u(x,t) - u_1(x,t)), \tag{1.8}$$

wobei α der Wärmetauschkoeffizient und $u_1(x,t)$ die Umgebungstemperatur ist.

Durch Einsetzen der Materialgesetze in die Erhaltungsgleichung erhalten wir:

$$|S| \int_{x_1}^{x_2} c\rho(u(x,t_2) - u(x,t_1))\,\mathrm{d}x - |S| \int_{t_1}^{t_2} \lambda\left(\partial_x u(x_2,t) - \partial_x u(x_1,t)\right)\,\mathrm{d}t$$
$$+ |\partial S| \int_{t_1}^{t_2} \int_{x_1}^{x_2} \alpha u(x,t)\,\mathrm{d}x\,\mathrm{d}t = \int_{t_1}^{t_2} \int_{x_1}^{x_2} \left(|S|\,f_0(x,t) + |\partial S|\,\alpha u_1(x,t)\right)\mathrm{d}x\,\mathrm{d}t.$$

Indem wir die gesamte Gleichung durch $h\tau$ teilen und den Grenzwert $h \to 0$ und $\tau \to 0$ bilden, erhalten wir für stetige Integranden

$$c\rho|S|\,\partial_t u(x,t) - \lambda|S|\,\partial_{xx} u(x,t) + \alpha|\partial S|\,u(x,t) = |S|\,f_0(x,t) + \alpha|\partial S|\,u_1(x,t).$$

Durch Vereinfachen erhalten wir die Differentialgleichung

$$\partial_t u(x,t) - \kappa \partial_{xx} u(x,t) + \beta u(x,t) = f(x,t), \quad x \in (0,L),\ t \in (0,T], \tag{1.9}$$

wobei $\kappa := \frac{\lambda}{c\rho} > 0$, $\beta := \frac{\alpha|\partial S|}{c\rho|S|} \geq 0$ und $f(x,t) := \frac{1}{c\rho} f_0(x,t) + \frac{|\partial S|}{c\rho|S|} u_1(x,t)$.

Zusätzlich zur Differentialgleichung benötigen wir Anfangsbedingungen und Randbedingungen, um ein *Anfangs-Randwertproblem* zu erhalten. Als Anfangsbedingung bietet sich die Temperaturverteilung zum Anfangszeitpunkt an:

$$u(x,0) = u_0(x), \quad x \in [0,L]. \tag{1.10}$$

Außerdem benötigen wir für jeden Teil des Rades jeweils eine Randbedingung, dies kann die Vorgabe der Temperatur sein, etwa

$$u(0,t) = g_D(0,t), \quad t \in (0,T], \tag{1.11}$$

oder die Vorgabe des Wärmeflusses, was sich in der Form

$$\kappa \partial_n u(L,t) = g_N(L,t), \quad t \in (0,T] \tag{1.12}$$

schreiben lässt. Randbedingungen der Form (1.11) nennen wir *Dirichlet*-Randbedingungen und die der Form (1.12) nennen wir *Neumann*-Randbedingungen. Es ist auch möglich, auf beiden Seiten Dirichlet-Randbedingungen oder auf beiden Seiten Neumann-Randbedingungen zu wählen.

Wärmeleitung im allgemeinen Fall. Nun wollen wir uns das Wärmeleitproblem genauer ansehen und zwar ohne die vereinfachende Annahme, wonach die Temperatur nur von einer der Ortsvariablen abhängt. Sei nun das Rechengebiet $\Omega \subset \mathbb{R}^d$ mit $d = 2, 3$ eine offene, beschränkte und zusammenhängende Menge. (Genauere Anforderungen an Ω müssen wir erst im Zusammenhang mit der Entwicklung von Theorie formulieren.)

Betrachten wir nun wieder ein offenes Kontrollgebiet $\omega \subseteq \Omega$, so können wir wieder eine Wärmebilanz aufstellen. Die Wärmeänderung über die Zeit abzüglich des Wärmeflusses in Außenrichtung muss der durch Aufheizung entsprechenden Wärmemenge entsprechen:

$$\int_\omega u(x,t_2) - u(x,t_1)\,\mathrm{d}x - \int_{t_1}^{t_2} \int_{\partial\omega} \underbrace{\lambda \partial_n u(x,t)}_{\text{Wärmefluss}}\,\mathrm{d}s(x)\,\mathrm{d}t = \int_{t_1}^{t_2} \int_\omega f(x,t)\,\mathrm{d}x\,\mathrm{d}t.$$

Auch hier ist der Wärmefluss durch das Fourier'sche Gesetz gegeben, wobei der Wärmefluss in Richtung des nach Außen zeigenden Normalvektors (Außennormalvektors) n an $\partial\omega$ in die Gleichung eingeht.

Mit dem Gauß'schen Integralsatz, $\int_\omega \operatorname{div}(\vec{q})\mathrm{d}x = \int_{\partial\omega} \vec{q} \cdot n\, ds(x)$, erhalten wir (wenn $\partial\omega$ hinreichend glatt ist)

$$\int_{\partial\omega} \lambda \partial_n u(x,t)\,\mathrm{d}s(x) = \int_\omega \operatorname{div}\big(\lambda \nabla u(x,t)\big)\,\mathrm{d}x,$$

wobei $\nabla u = (\text{grad } u)^\top = (\partial_{x_1} u, \dots, \partial_{x_d} u)^\top$ und $\text{div}\,(v_1, \dots, v_d)^\top = \partial_{x_1} v_1 + \cdots + \partial_{x_d} v_d$. Nach Einsetzen können wir die Bilanzgleichung wieder durch $\tau|\omega|$ dividieren. Wenn wir nun etwa $\omega := \omega(x,h) := \{y \in \mathbb{R}^d : \|y - x\|_{\ell^2} < h\}$ wählen, wobei $h > 0$ so klein sein muss, dass $\omega(x,h) \subseteq \Omega$ gilt, können wir wieder einen Grenzübergang $h \to 0$ und $\tau \to 0$ machen, um für stetige Integranden folgende Differentialgleichung zu erhalten:

$$\partial_t u - \text{div}\big(\lambda \nabla u\big) = f \quad \text{in } \Omega \times (0,T).$$

Wenn λ konstant ist, dann können wir stattdessen

$$\partial_t u - \lambda \Delta u = f \quad \text{in } \Omega \times (0,T),$$

schreiben, wobei $\Delta u := \text{div} \nabla u = \frac{\partial^2 u}{\partial x_1^2} + \cdots + \frac{\partial^2 u}{\partial x_d^2}$ der Laplace-Operator ist.

Wie schon oben besprochen, geben wir als *Anfangsbedingung* den Funktionswert für $t = 0$ vor:

$$u(x,t) = u_0(x), \quad x \in \Omega. \tag{1.13}$$

Wir benötigen wieder für jeden Teil des Randes $\partial\Omega$ jeweils genau eine *Randbedingung* ($\partial\Omega = \Gamma_D \cup \Gamma_N \cup \Gamma_R$ und $\Gamma_D \cap \Gamma_N = \Gamma_D \cap \Gamma_R = \Gamma_N \cap \Gamma_R = \emptyset$). Analog zu (1.10) können wir die Temperatur am Rand vorgeben (Dirichlet-Randbedingung):

$$u(x,t) = g_D(x,t), \quad x \in \Gamma_D,\ t \in (0,T]. \tag{1.14}$$

Analog zu (1.11) können wir den Wärmefluss am Rand vorgeben (Neumann-Randbedingung):

$$\lambda \partial_n u(x,t) = g_N(x,t), \quad x \in \Gamma_N,\ t \in (0,T]. \tag{1.15}$$

Für den Mantel hatten wir die Bedingung (1.8) gefordert, wonach der Wärmefluss proportional zur Differenz zur Temperatur u und der Umgebungstemperatur u_1 ist. Ein solcher Zusammenhang lässt sich einfach als Robin-Randbedingung schreiben:

$$\lambda \partial_n u(x,t) + \gamma u(x,t) = g_R(x,t) \quad x \in \Gamma_R,\ t \in (0,T]. \tag{1.16}$$

Die Kombination aus der Differentialgleichung, der Anfangsbedingung und den Randbedingungen ergibt das *Rand- und Anfangswertproblem*:

$$\begin{aligned}
\partial_t u - \text{div}\big(\lambda \nabla u\big) &= f && \text{in } \Omega \times (0,T],\\
u(\cdot,0) &= u_0 && \text{auf } \Omega,\\
u &= g_D && \text{auf } \Gamma_D \times (0,T],\\
\lambda \partial_n u &= g_N && \text{auf } \Gamma_N \times (0,T],\\
\lambda \partial_n u + \gamma u &= g_R && \text{auf } \Gamma_R \times (0,T].
\end{aligned}$$

Stationäre Wärmeverteilung

Wenn die Funktionen f, g_D und g_N nicht von der Zeit abhängen, dann kann sich ein stationärer Zustand einpendeln. Da sich bei einem stationären Zustand die Temperatur nach der Zeit nicht mehr ändert, gilt offensichtlich $\partial_t u = 0$. Da in diesem Fall ein eventueller ,,Anfangszustand" keine Rolle mehr spielt, gibt es auch keine Anfangsbedingungen. Ein stationärer Zustand ist durch das *Randwertproblem* charakterisiert:

$$\begin{aligned} -\mathrm{div}\big(\lambda \nabla u\big) &= f && \text{in } \Omega, \\ u &= g_D && \text{auf } \Gamma_D, \\ \lambda \partial_n u &= g_N && \text{auf } \Gamma_N, \\ \lambda \partial_n u + \gamma u &= g_R && \text{auf } \Gamma_R. \end{aligned}$$

Wir werden uns in Teil II mit solchen Problemen näher beschäftigen. Hier wird uns ganz besonders auch interessieren, ob das Problem (eindeutig) lösbar ist, also die Frage, ob das Wärmeleitproblem überhaupt einen stationären Zustand annehmen kann.

Chemische Reaktion

Wir haben, wie zuvor im Fall der Wärmeleitgleichungen, wieder ein beschränktes, offenes und zusammenhängendes Gebiet $\Omega \subset \mathbb{R}^d$, $d = 1, 2, 3$, wobei wir an der Konzentration $u(x,t)$ eines Stoffes interessiert sind, wie etwa der Salzkonzentration in einer Wasserlösung.

Wir können wieder eine Bilanzgleichung aufstellen. Für jedes offene Kontrollvolumen $\omega \subseteq \Omega$ betrachten wir die Gesamtmenge des Stoffes. Ihre Änderung über die Zeit entspricht dem über $\partial\omega$ strömenden Stoff und die Änderung, die durch eine chemische Reaktion hervorgerufen wird. Letztere wird durch eine Quelldichte modelliert. Damit haben wir

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_\omega u(x,t)\,\mathrm{d}x = -\int_{\partial\omega} \vec{q}(x,t) \cdot n(x)\,\mathrm{d}s(x) + \int_\omega s(x,t)\,\mathrm{d}x. \tag{1.17}$$

Dabei ist $\vec{q}$ der Fluss bzw. Mengenstrom, n der äußere Normalvektor am Rand und s die Quelldichte.

Durch Umstellen der Terme und mit dem Gauß'schen Integralsatz folgt sofort

$$\int_\omega \partial_t u(x,t) + \mathrm{div}(\vec{q}(x,t))\,\mathrm{d}x = \int_\omega s(x,t)\,\mathrm{d}x. \tag{1.18}$$

Da diese Integralbilanz für alle offenen $\omega \subseteq \Omega$ gilt, folgt unter der Annahme, dass die Integranden stetige Funktionen sind, punktweise

$$\partial_t u(x,t) + \mathrm{div}(\vec{q}(x,t)) = s(x,t) \quad x \in \Omega,\ t \in (0,T]. \tag{1.19}$$

Dieser Differentialgleichung liegt das Prinzip der Erhaltung der Stoffmenge zu Grunde; sie wird entsprechend eine Erhaltungsgleichung genannt.

Um zu einem abgeschlossenen Modell mit gleichen vielen Gleichungen wie unbekannten Funktionen zu kommen, müssen wir noch Relationen zwischen Konzentrationen und Flüssen bzw. Quellen vorgeben. Wir nehmen zunächst an, dass

$$\vec{q} = \underbrace{-D\nabla u}_{\text{Fick'sches Gesetz}} + \underbrace{\vec{b}u}_{\text{Transport}} \tag{1.20}$$

gilt, wobei $D > 0$ den Diffusionskoeffizienten und $\vec{b} = \vec{b}(x,t)$ das Geschwindigkeitsfeld der Strömung bezeichnet. Die Abhängigkeit der Funktionen von x und t wollen wir der Übersichtlichkeit halber im Folgenden nicht mehr explizit hervorheben. Der Quellterm sei weiters beschrieben durch

$$s = \underbrace{f}_{\text{externe Quelle}} - \underbrace{cu}_{\text{Zerfall}}, \tag{1.21}$$

wobei $c \geq 0$ eine Reaktions- oder Zerfallskonstante sei. Durch Einsetzen in die Erhaltungsgleichung erhält man dann die Differentialgleichung

$$\partial_t u - \operatorname{div}(D\nabla u - \vec{b}u) + cu = f, \quad x \in \Omega,\ t \in (0,T]. \tag{1.22}$$

Als *Anfangsbedingung* wählen wir einen Anfangszustand für die Konzentration

$$u(\cdot,0) = u_0 \qquad \text{in } \Omega. \tag{1.23}$$

Wir benötigen wieder für jeden Teil des Randes *Randbedingungen* ($\partial\Omega = \Gamma_D \cup \Gamma_N$ und $\Gamma_D \cap \Gamma_N = \emptyset$). Wir fordern

$$u = g_D \qquad \text{auf } \Gamma_D \times (0,T] \tag{1.24}$$

$$-n \cdot \vec{q} = g_N \qquad \text{auf } \Gamma_N \times (0,T]. \tag{1.25}$$

Im ersten Fall wird die Konzentration und im zweiten Fall der Fluss in Normalrichtung vorgegeben. Für den Fluss $\vec{q}$ kann wieder das Materialgesetz (1.20) eingesetzt werden.

Schwingende Saite

Wir betrachten die Auslenkung $u(x,t)$ einer Saite der Länge L ($x \in \Omega = (0,L)$, $t \in (0,T)$) mit kleinem Radius ($r \ll L$). Die wirkende Beschleunigung $\partial_{tt}u$ ist proportional zur wirkenden Kraft. Einerseits ergibt sich eine Kraft aus der Eigenschaft des Materials, sich einer Krümmung $\partial_{xx}u$ zu widersetzen (Hook'sches

Gesetz), andererseits kann sich eine solche durch eine äußere Anregung ergeben. Insgesamt erhalten wir also

$$\partial_{tt} u(x,t) - \lambda \partial_{xx} u(x,t) = f(x,t), \quad x \in (0,L),\ t \in (0,T],$$

wobei $\lambda > 0$ (Skizze zum Vorzeichen!) ein materialabhängige Proportionalitätsfaktor ist und f die äußere Anregung modelliert. Daneben benötigen wir wieder *Anfangs- und Randbedingungen*. Da die Differentialgleichung zweite Ableitungen in der Zeit enthält, benötigen wir zwei Anfangsbedingungen:

$$\begin{aligned} u(x,0) &= u_0(x), \quad x \in (0,L), \\ \partial_t u(x,0) &= u_1(x), \quad x \in (0,L). \end{aligned}$$

Daneben benötigen wir wieder jeweils genau eine Randbedingung für jeden Teil des Randes ($\{0,L\} = \Gamma_D \cup \Gamma_N \cup \Gamma_R$ und $\Gamma_D \cap \Gamma_N = \Gamma_D \cap \Gamma_R = \Gamma_N \cap \Gamma_R = \emptyset$): Durch Dirichlet-Randbedingungen der Art

$$u(x,t) = g_D(x,t), \quad x \in \Gamma_D,\ t \in (0,T]$$

können wir die jeweilige Auslenkung und durch Neumann-Randbedingungen der Art

$$\partial_n u(x,t) = g_N(x,t), \quad x \in \Gamma_N,\ t \in (0,T]$$

können wir die jeweils wirken Kräfte vorgeben. Werden Robin-Randbedingungen

$$\partial_n u(x,t) = g_R(x,t) - \gamma u(x,t), \quad x \in \Gamma_R,\ t \in (0,T]$$

vorgegeben, so hängt die Kraft linear von der Auslenkung ab.

Ausblick

Im Rahmen der Vorlesung werden wir die Existenz- und Eindeutigkeit von Lösungen u besprechen und uns dann insbesondere deren numerischer Approximation widmen. Geschlossene Formeln für die Lösung lassen sich nämlich in den meisten Fällen nicht finden.

Dabei werden wir mit den *Anfangswertproblemen* beginnen. Da es sich bei diesen um gewöhnliche Differentialgleichungen handelt, haben Sie einiges zu Existenz- und Eindeutigkeit von Lösungen bereits in der entsprechenden Vorlesung gesehen. Wir wollen das etwas auffrischen und dann numerische Methoden zu den Anfangswertproblemen besprechen, die wir später benötigen.

Im folgenden Teil werden wir *Randwertprobleme* behandeln. Randwertprobleme in einer einzigen Ortsdimension basieren zwar ebenfalls auf einer gewöhnlichen Differentialgleichung, wir fordern nun jedoch Randbedingungen statt den Anfangsbedingungen. Wir werden nicht nur Probleme in nur einer Ortsdimension

kennenlernen, sondern auch die Erweiterungen auf zwei- und dreidimensionale Gebiete diskutieren.

Die Behandlung der *Anfangs-Randwertprobleme* folgt dann als nächster Schritt. Hier werden wir zwischen den parabolischen Problemen mit nur erster Zeitableitung (wie in den Abschnitten zur Wärmeleitung und zur chemischen Reaktion gesehen) und den hyperbolischen Problemen mit einer zweiten Zeitableitung (wie im Abschnitt zur schwingenden Saite) unterscheiden.

Ergänzend werden wir uns auch ansehen, wie wir die sich aus der Diskretisierung unserer Probleme ergebenden linearen Gleichungssysteme effizient lösen werden. Außerdem werden wir uns noch die Finite Elemente Methode ansehen, diese aber auch als Ausblick auf die VL *Numerik elliptischer Probleme*.

Literatur

- W. Zulehner: *Numerische Mathematik. Eine Einführung anhand von Differentialgleichungsproblemen, Band 1: Stationäre Probleme.* Birkhäuser. 2008.
- W. Zulehner: *Numerische Mathematik. Eine Einführung anhand von Differentialgleichungsproblemen, Band 2: Instationäre Probleme.* Birkhäuser. 2011.
- M. Hanke-Bourgeois: *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens.* Teubner. 2006.
- A. Tveito, R. Winther: *Introduction to Partial Differential Equations: A Computational Approach.* Springer. 1991.
- M. Jung, U. Langer: *Methode der finiten Elemente für Ingenieure: Eine Einführung in die numerischen Grundlagen und Computersimulationen.* Springer Vieweg. 2013

Teil I.

Anfangswertprobleme

2. Theoretische Grundlagen und Einschrittverfahren

Modellproblem. In diesem Kapitel wollen wir uns mit der numerischen Lösung von Anfangswertproblemen der Form

$$u'(t) = f(t, u(t)), \qquad t \in (0, T], \tag{2.1}$$
$$u(0) = u_0 \tag{2.2}$$

beschäftigen. Dabei sind $f : \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}^N$ und $u_0 \in \mathbb{R}^N$ gegeben. Wir zitieren zunächst einige Sätze aus der Theorie gewöhnlicher Differentialgleichungen. Dann stellen wir numerische Verfahren zum Lösen gewöhnlicher Differentialgleichungen vor. Wir beschränken uns dabei auf die Einschrittverfahren. Im nächsten Kapitel werden wir dann eine Klasse von Einschrittverfahren, die Runge-Kutta Verfahren näher beleuchten.

Lösbarkeit und Stabilität der Lösung

Durch Integration von (2.1) und Verwendung von (2.2) erhält man

$$u(t) = u(0) + \int_0^t u'(s)\mathrm{d}s = u_0 + \int_0^t f(s, u(s))\mathrm{d}s, \qquad t \in (0, T]. \tag{2.3}$$

Nach Konstruktion erfüllt jede (stetige) Lösung des Anfangswertproblems auch diese Integralgleichung. Durch Differenzieren sieht man, dass jede stetig differenzierbare Lösung der Integralgleichung auch das Anfangswertproblem erfüllt. Die Gleichung (2.3) lässt sich auch kompakt als Fixpunktgleichung

$$u = \Phi(u) \tag{2.4}$$

schreiben, wobei die Fixpunktabbildung $\Phi : C([0, T]; \mathbb{R}^N) \to C([0, T]; \mathbb{R}^N)$ durch die Vorschrift $\Phi(u)(t) = u_0 + \int_0^t f(s, u(s))\mathrm{d}s$ definiert ist. Jede Lösung des Anfangswertproblems ist nach Konstruktion ein Fixpunkt der Abbildung Φ und umgekehrt. Mit Hilfe des Banach'schen Fixpunktsatzes erhalten wir

Satz 2.1 (Picard-Lindelöf). *Sei $u_0 \in \mathbb{R}^N$ und $f : [0,T] \times \mathbb{R}^N \to \mathbb{R}^N$ sei stetig und gleichmäßig Lipschitz-stetig bezüglich des zweiten Arguments, d.h., es gibt eine Konstante $L > 0$, sodass*

$$\|f(t,x) - f(t,y)\|_{\ell^2} \leq L\|x-y\|_{\ell^2} \quad \forall t \in [0,T],\ x,y \in \mathbb{R}^N. \tag{2.5}$$

Dann hat (2.1)–(2.2) eine eindeutige Lösung $u \in C^1([0,T];\mathbb{R}^N)$.

Beweis. Siehe Übung und VL *Gewöhnliche Differentialgleichungen und Dynamische Systeme*. □

Sind die Voraussetzungen nur in der Umgebung um u_0 erfüllt, so kann gezeigt werden, dass es ein Invtervall $[0,T_0]$ gibt, auf dem das Anfangswertproblem eindeutig lösbar ist, vgl. Übung.

Ist f sogar stetig differenzierbar auf $[0,T]\times\mathbb{R}^N$, dann folgt mit der Kettenregel

$$u''(t) = \partial_t f(t,u(t)) + \partial_u f(t,u(t))u'(t) \in C([a,b];\mathbb{R}^N),$$

d.h., die Lösung $u(t)$ ist sogar zweimal stetig differenzierbar. Mittels Induktion lässt sich dann folgende Aussage über die erhöhte Regularität zeigen.

Satz 2.2 (Regularität). *Die Funktion f sei m-mal stetig differenzierbar auf $[0,T] \times \mathbb{R}^N$ und $m \geq 1$. Dann sind die Voraussetzungen des Satzes 2.1 erfüllt und die Lösung u von (2.1)–(2.2) ist $m+1$-mal stetig differenzierbar.*

Als nächstes untersuchen wir die Frage, ob und unter welchen Voraussetzungen die Lösung von (2.1)–(2.2) stabil von den Daten u_0 und f abhängt. Zur Beantwortung dieser Frage ist das Gronwall-Lemma hilfreich.

Satz 2.3 (Gronwall-Lemma). *Sei $w : [0,T] \to \mathbb{R}$ eine stetige Funktion mit*

$$0 \leq w(t) \leq u(t) := a + \int_0^t bw(s) + c\,ds \qquad \forall\, t \in [0,T]$$

und Konstanten $a,b,c \geq 0$. Dann gilt $w(t) \leq v(t) := ae^{bt} + \frac{c}{b}(e^{bt}-1)$.

Beweis. Für $0 \leq t \leq T$ gilt offensichtlich

$$u'(t) = bw(t) + c \leq bu(t) + c \quad \text{und} \quad v'(t) = bv(t) + c$$

und weiters $u(0) = v(0) = a$. Wir zeigen, dass $u(t) \leq v(t)$, womit der Satz wegen $w(t) \leq u(t)$ bewiesen ist. Hierzu betrachten wir die Funktion $\phi(t) = e^{-bt}(u(t) - v(t))$. Für diese gilt $\phi(0) = 1 \cdot (a-a) = 0$ und

$$\phi'(t) = [e^{-bt}(u(t)-v(t))]' = e^{-bt}[(u'-bu) - (v'-bv)] \leq e^{-bt}[c-c] = 0.$$

Somit ist $\phi(t) \leq 0$, woraus $u(t) \leq v(t)$ für alle $0 \leq t \leq T$ folgt. □

Aus (2.3) und (2.5) (und der Stetigkeit von f) erhalten wir sofort

$$\|u(t) - u_0\|_{\ell^2} \leq \int_0^t L\|u(s) - u_0\|_{\ell^2} + \max_{\sigma\in[0,T]} \|f(\sigma, u_0)\|_{\ell^2} \,\mathrm{d}s.$$

Mit dem Gronwall-Lemma erhalten wir daraus die Beschränktheit der Lösung.

Lemma 2.4. *Unter den Voraussetzungen des Satzes von Picard-Lindelöf gilt:*

$$\|u(t) - u_0\|_{\ell^2} \leq \frac{e^{Lt} - 1}{L} \max_{\sigma\in[0,T]} \|f(\sigma, u_0)\|_{\ell^2}. \tag{2.6}$$

Nun können wir uns der Stabilitätsaussage näher widmen. Seien u und $\tilde{u}$ die Lösungen zu (2.1)–(2.2) mit Daten (u_0, f) und $(\tilde{u}_0, \tilde{f})$. Dann gilt

$$\tilde{u}(t) - u(t) = \tilde{u}_0 - u_0 + \int_0^t \tilde{f}(s, \tilde{u}(s)) - f(s, u(s))\mathrm{d}s.$$

Erweitern des Integranden mit $\pm f(s, \tilde{u}(s))$ und Dreiecksungleichung liefern

$$\begin{aligned}\|\tilde{u}(t) - u(t)\|_{\ell^2} &\leq \|\tilde{u}_0 - u_0\|_{\ell^2} \\ &\quad + \int_0^t \|\tilde{f}(s, \tilde{u}(s)) - f(s, \tilde{u}(s))\|_{\ell^2} + \|f(s, \tilde{u}(s)) - f(s, u(s))\|_{\ell^2}\mathrm{d}s.\end{aligned}$$

Wir nehmen nun an, dass auch $\tilde{u}(t)$ beschränkt ist, also

$$\|\tilde{u}(t) - \tilde{u}_0\|_{\ell^2} \leq \tilde{C} \qquad \forall t \in [0, T]. \tag{2.7}$$

Wenn $\tilde{f}$ die Lipschitz-Bedingung (2.5) erfüllt, dann gilt dies analog zu Lemma 2.4 mit $\tilde{C} = \frac{e^{L\,t}-1}{L}$. Mit der Stetigkeit von f und der Lipschitz-Bedingung (2.5) erhalten wir ferner

$$\begin{aligned}\|\tilde{u}(t) - u(t)\|_{\ell^2} &\leq \|\tilde{u}_0 - u_0\|_{\ell^2} \\ &\quad + \int_0^t \left(\max_{\sigma\in[0,T]} \max_{z:\|z-\tilde{u}_0\|_{\ell^2}\leq\tilde{C}} \|\tilde{f}(\sigma, z) - f(\sigma, z)\|_{\ell^2} + L\|\tilde{u}(s) - u(s)\|_{\ell^2} \right) \mathrm{d}s.\end{aligned}$$

Wir wenden nun das Gronwall-Lemma mit

- $w(t) := \|\tilde{u}(t) - u(t)\|_{\ell^2}$,
- $a := \|\tilde{u}_0 - u_0\|_{\ell^2}$,
- $b := L$,
- $c := \max_{\sigma\in[0,T]} \max_{z:\|z-\tilde{u}_0\|_{\ell^2}\leq\tilde{C}} \|\tilde{f}(\sigma, z) - f(\sigma, z)\|_{\ell^2}$

an, um folgenden Satz zu erhalten:

Satz 2.5 (Stabilitätssatz)**.** *Seien $u, \tilde{u} : [0,T] \to \mathbb{R}^N$ Lösungen zu* (2.1)–(2.2) *mit rechten Seiten $f, \tilde{f} \in C([0,T] \times \mathbb{R}^N; \mathbb{R}^N)$ und Anfangswerten $u_0, \tilde{u}_0 \in \mathbb{R}^N$. Weiters erfülle f die Lipchitz-Bedingung* (2.5) *mit Konstante L und die Lösung $\tilde{u}$ die Beschränktheitsbedingung* (2.7) *mit Konstante $\tilde{C}$. Dann gilt*

$$\|u(t) - \tilde{u}(t)\|_{\ell^2} \leq e^{Lt}\|u_0 - \tilde{u}_0\|_{\ell^2} + \frac{e^{Lt}-1}{L} \max_{s\in[0,T]} \max_{z:\|z-\tilde{u}_0\|_{\ell^2}\leq\tilde{C}} \|f(s,z) - \tilde{f}(s,z)\|_{\ell^2}.$$

Die Lösung des Anfangswertproblems hängt also stabil von den Daten ab.

Bemerkung 2.6. Wir sehen, dass die Lösung (Lemma 2.4) bzw. die Abweichung zwischen den Lösungen (Satz 2.5) auf allen beschränkten Zeithorizonten beschränkt bleibt, die obere Schranke aber mit der Länge des Zeithorizonts exponentiell ansteigt. Diese Stabilitätsaussage bildet auch die Grundlage für die Fehleranalyse numerischer Verfahren zum näherungsweisen Lösen. In Kapitel 4 werden Fälle kennenlernen, für die wir stärkere Aussagen zur Stabilität machen können.

Einschrittverfahren

Unser nächstes Ziel ist es, numerische Verfahren für die Lösung des Modellproblems zu formulieren. Wir nehmen im Folgenden an, dass – wie im Satz von Picard-Lindelöf (Satz 2.1) – die Funktion f stetig ist und die Lipschitz-Bedingung (2.5) erfüllt. Der Einfachheit halber werden wir unsere Herleitungen oftmals für den den skalaren Fall $N = 1$ motivieren und schreiben $C([0,T])$ auch für $C([0,T];\mathbb{R}^N)$.

In Kapitel 1 hatten wir bereits das explizite Eulerverfahren kennengelernt, das die numerische Lösung des Anfangswertproblems erlaubt. Wir erlauben nun unterschiedliche Schrittweiten. Sei

$$\mathcal{T}_h = \{t_0, t_1, \ldots, t_n\} \quad \text{mit} \quad 0 = t_0 < t_1 < \cdots < t_n = T$$

ein (zeitliches) *Gitter* und $h_i := t_{i+1} - t_i$ die jeweilige (lokale) Gitterweite. Dann hat das explizite Eulerverfahren die Form

$$u_{i+1} = u_i + h_i f(t_i, u_i). \tag{2.8}$$

Wir betrachten im Folgenden gleich eine ganze Klasse von Verfahren, die eine Verallgemeinerung des expliziten Eulerverfahrens (2.8) darstellen.

Definition 2.7 (Einschrittverfahren)**.** Ein numerisches Verfahren der Form

$$u_{i+1} = u_i + h_i \, \phi(t_i, u_i; h_i) \tag{2.9}$$

zur Konstruktion von Näherungen $u_i = u_h(t_i) \approx u(t_i)$ heißt Einschrittverfahren (ESV) mit Verfahrensfunktion ϕ und zeitlichem Gitter $\mathcal{T}_h$.

Bemerkung 2.8. Das explizite Eulerverfahren ist ein ESV mit Verfahrensfunktion $\phi(t, y; h) = f(t, y)$.

Wir nennen $u_h : \mathcal{T}_h \to \mathbb{R}$ mit $u_h(t_i) = u_i$ *Gitterfunktion* bzw. *Skelettlösung*. Nach den obigen Bemerkungen kann die Gitterfunktion u_h stets durch Interpolation wieder zu einer stetigen Funktionen auf $[0, T]$ fortgesetzt werden; als Beispiel hatten wir stetige und stückweise lineare Funktion (Polygonzüge) kennengelernt.

Nun wollen wir zwei Interpretationen für die Einschrittverfahren geben.

Bemerkung 2.9 (Interpretation über Integraldarstellung)**.** Durch Integration der Differentialgleichung (2.1) erhalten wir

$$u(t_{i+1}) = u(t_i) + \int_{t_i}^{t_{i+1}} f(t, u(t)) \, \mathrm{d}t = u(t_i) + h_i \int_0^1 f(t_i + sh_i, u(t_i + sh_i)) \, \mathrm{d}s. \tag{2.10}$$

Dementsprechend soll die Verfahrensfunktion ϕ im ESV das Integral $\int_0^1 f(t_i + sh_i, u(t_i + sh_i)) \, \mathrm{d}s$ durch bekannte (berechenbare) Terme annähern. Beim expliziten Eulerverfahren erfolgt die Annäherung dieses Integrals durch die linksseitige Rechtecksregel, also $\int_0^1 f(t_i + sh_i, u(t_i + sh_i)) \, \mathrm{d}s \approx f(t_i, u(t_i))$.

Bemerkung 2.10 (Interpretation über Differentialquotienten)**.** Alternativ lässt sich das Einschrittverfahren auch schreiben als

$$\frac{u_h(t_{i+1}) - u_h(t_i)}{h} = \phi(t_i, u_i; h_i). \tag{2.11}$$

Als Approximation für die Differentialgleichung $u'(t) = f(t, u(t))$ verwenden wir also eine Differenzengleichung, in der $u'(t)$ durch den Vorwärtsdifferenzenquotienten angenähert wird und die Verfahrensfunktion ϕ eine Approximation der rechten Seite f ist.

Konvergenz von Einschrittverfahren

Wir nehmen für die theoretischen Überlegungen ein *äquidistantes Gitter* an, d.h., wir nehmen an, dass alle Schrittweiten gleich sind: $h_0 = h_1 = \cdots = h_{n-1} =: h$.

Die Aussagen lassen sich leicht auf den nicht äquidistante Gitter verallgemeinern, wobei dann meist die maximale Schrittweite $h := \max_i h_i$ wesentlich ist.

Die folgenden Begriffe sind für die Analyse von Einschrittverfahren essentiell.

Definition 2.11 (Konsistenz)**.** Sei $u : [0,T] \to \mathbb{R}$ eine Lösung von $u' = f(t,u)$ und ϕ die Verfahrensfunktion eines entsprechenden Einschrittverfahrens.

(i) Die Größe

$$\tau_h(t_i + h, u) := \frac{u(t_i + h) - u(t_i)}{h} - \phi(t_i, u(t_i); h)$$

heißt *lokaler Abschneidefehler* (engl: truncation error) und

$$d_h(t_i + h, u) := u(t_i + h) - [u(t_i) + h\phi(t_i, u(t_i), h)] = h\tau_h(t_i + h, u)$$

heißt *lokaler Fehler* des Verfahrens bei u im Zeitschritt $t_i \to t_i + h$.

(ii) Das Einschrittverfahren (mit Verfahrensfunktion ϕ) heißt *konsistent* (mit der Differentialgleichung $u' = f(t,u)$), wenn

$$\|\tau_h\|_{\infty,h} := \max_{t_i \in \mathcal{T}_h} |\tau_h(t_i, u)| \to 0, \qquad h \to 0,$$

für alle Lösungen u von $u' = f(t,u)$.

(iii) Das Verfahren heißt *konsistent mit Ordnung p*, falls für $f \in C^p$ (und somit $u \in C^{p+1}$ wegen Regularitätssatz)

$$\|\tau_h\|_{\infty,h} \leq C(f)h^p,$$

mit einer Konstanten $C(f)$, welche nur von f und ihren Ableitungen abhängt.

Satz 2.12 (Konsistenz)**.** *Ein Einschrittverfahren ist genau dann konsistent mit $u' = f(t,u)$, falls für alle $t \in [0,T]$ und alle zulässigen Lösungen u folgendes gilt:*

$$\lim_{h \to 0} \phi(t, u(t); h) = f(t, u(t)).$$

Beweis. Folgt sofort aus der Definition der Begriffe und Stetigkeit. □

Beachte. Der lokale Fehler ist der Fehler, der bei der Annäherung der rechten Seite für ein fixes u bzw. ein fixes u_h entsteht. Diese Information reicht aber noch nicht aus, um den Fehler $\max_{t \in [0,T]} |u(t) - u_h(t)|$ aus den Daten (f und u_0) bestimmen (oder von oben abschätzen) zu können.

Wir betrachten ganz ähnlich wie bei Lösungen u von Anfangswertproblemen nun auch die Stabilität (=stetige Abhängigkeit) der Näherungen u_h aus dem Einschrittverfahren bezüglich Anfangsdaten und der rechten Seite.

Definition 2.13 (Diskrete Stabilität)**.** Ein Einschrittverfahren heißt (diskret) stabil, wenn für Gitterfunktionen u_h und $\tilde{u}_h$ mit

$$u_{i+1} = u_i + h\phi(t_i, u_i; h),$$
$$\tilde{u}_{i+1} = \tilde{u}_i + h(\phi(t_i, \tilde{u}_i; h) + \theta_i),$$

bei beliebigen Störungen $\theta_i = \theta_h(t_i)$ der rechten Seiten und der Anfangswerte $u_0 - \tilde{u}_0$ der Fehler beschränkt ist durch

$$\|u_h - \tilde{u}_h\|_{\infty,h} \leq C\big(\|u_0 - \tilde{u}_0\|_{\ell^2} + \|\theta_h\|_{\infty,h}\big)$$

mit einer Konstante $C \neq C(h)$. Zur Erinnerung: $\|u_h\|_{\infty,h} = \max_{t_i \in \mathcal{T}_h} \|u_i\|_{\ell^2}$.

Zum Beweis der Stabilität hatten wir im ersten Abschnitt dieses Kapitels das Gronwall-Lemma (Satz 2.3) verwendet. Zum Beweis der diskreten Stabilität benötigen wir nun ein diskretes Gronwall-Lemma.

Satz 2.14 (Diskretes Gronwall-Lemma)**.** *Sei $\{w_i\}_{i\geq 0}$ gegeben mit*

$$0 \leq w_{i+1} \leq (1 + hL)w_i + hc_i$$

und $c_i \geq 0$, $h \geq 0$, $L \geq 0$. Dann gilt

$$w_n \leq w_0 e^{Lhn} + \frac{e^{Lhn} - 1}{L} \max_{0\leq i\leq n-1} c_i.$$

Beweis. Mit $q := 1 + hL \geq 1$ erhält man durch Anwenden der Rekursion

$$\begin{aligned} w_n &\leq qw_{n-1} + hc_{n-1} \leq q(qw_{n-2} + hc_{n-2}) + hc_{n-1} \leq \dots \\ &\leq q^n w_0 + h\sum_{m=0}^{n-1} q^m c_{n-m-1} \leq q^n w_0 + h\sum_{m=0}^{n-1} q^m \max_{0\leq i\leq n-1} c_i \\ &= q^n w_0 + h\frac{q^n - 1}{q - 1} \max_{0\leq i\leq n-1} c_i = q^n w_0 + \frac{q^n - 1}{L} \max_{0\leq i\leq n-1} c_i. \end{aligned}$$

Die Aussage folgt dann mit $q = 1 + hL \leq 1 + hL + \frac{1}{2!}(hL)^2 + \dots = e^{hL}$. □

Satz 2.15 (Diskrete Stabilität)**.** *Die Verfahrensfunktion des Einschrittverfahrens sei gleichmäßig Lipschitz-stetig bezüglich des zweiten Arguments, d.h.,*

$$|\phi(t, x; h) - \phi(t, y; h)| \leq L|x - y|, \qquad \forall t \in [0, T],\ x, y \in \mathbb{R}, \tag{2.12}$$

mit Konstante $L \neq L(h)$. Dann ist das Verfahren diskret stabil *nach obiger Definition.*

Beweis. Wir gehen ganz analog zum Stabilitätssatz für Anfangswertprobleme vor: Die Differenz $w_h = u_h - \tilde{u}_h$ erfüllt

$$w_{i+1} = w_i + h[\phi(t_i, y_i; h) - \phi(t_i, \tilde{u}_i; h) - \theta_i], \qquad w_0 = u_0 - \tilde{u}_0.$$

Hieraus erhält man mit Dreiecksungleichung:

$$\|w_{i+1}\|_{\ell^2} \leq \|w_i\|_{\ell^2} + hL\|w_i\|_{\ell^2} + h\|\theta_i\|_{\ell^2}.$$

Mit dem diskreten Gronwall-Lemma und $nh = t_n \leq T$ erhalten wir:

$$\begin{aligned}\|u_n - \tilde{u}_n\|_{\ell^2} &\leq \max\left\{e^{L\,t_n}, \frac{e^{L\,t_n} - 1}{L}\right\}(\|u_0 - \tilde{u}_0\|_{\ell^2} + \|\theta_h\|_{\infty,h}) \\ &\leq \max\left\{e^{L\,T}, \frac{e^{L\,T} - 1}{L}\right\}(\|u_0 - \tilde{u}_0\|_{\ell^2} + \|\theta_h\|_{\infty,h})\end{aligned}$$

Wir sehen, dass wir $C \neq C(h)$ erhalten können. □

Bemerkung 2.16. Die Abweichung ist durch eine Konstante C beschränkt, die von der Gitterweite unabhängig ist. Diese Konstante kann jedoch mit der Länge des Zeithorizonts exponentiell ansteigen. Dieses Verhalten entspricht genau dem Verhalten, das wir für die Lösung des Anfangswertproblems kennengelernt haben.

Wir werden nun sehen, dass aus Konsistenz und Stabilität eines Einschrittverfahrens dessen Konvergenz folgt. Zunächst führen wir den Konvergenzbegriff ein.

Definition 2.17 (Konvergenz)**.** Sei u eine Lösung der Gleichung $u' = f(t, u)$ mit Startwert $u(0) = u_0$, und u_h die entsprechende Näherungslösung aus dem Einschrittverfahren $u_{i+1} = u_i + h\phi(t_i, y_i; h)$ mit $t_{i+1} = t_i + h$.

(i) Die Gitterfunktion E_h definiert durch

$$E_h(t_i) = u(t_i) - u_h(t_i), \qquad t_i \in \mathcal{T}_h$$

heißt *globaler Fehler* des Verfahrens.

(ii) Das Verfahren (die Näherungslösung) heißt *konvergent*, falls

$$\|E_h\|_{\infty,h} = \max_{t_i \in \mathcal{T}_h} |u(t_i) - u_h(t_i)| \to 0, \qquad h \to 0.$$

(iii) Das Verfahren heißt konvergent mit Ordnung p, falls

$$\|E_h\|_{\infty,h} \leq Ch^p,$$

wobei die Konstante C nur von der exakten Lösung u bzw. den Daten f, u_0 und den entsprechenden Ableitungen abhängen darf.

Wir bemerken, dass die Konstante C nicht von der jeweiligen Näherungslösung u_h und natürlich auch nicht von h abhängen darf.

Definition (Interpolationsoperator). Der Interpolationsoperator ist jener lineare Operator r_h, der jeder Funktion $u \in C[0,T]$ den interpolierenden Polygonzug $r_h\, u := \tilde{u}_h$ zuordnet, also mit $\tilde{u}_h(x_i) = u(x_i)$ für alle $x_i \in \mathcal{T}_h$.

Satz 2.18 (Konvergenz)**.** *Das Einschrittverfahren sei konsistent (mit Ordnung p) und diskret stabil. Dann ist es auch konvergent (mit Ordnung p). Die Aussagen gelten natürlich nur für hinreichend glatte Lösungen!*

Beweis. Wir haben die Näherungslösung u_h mit Werten $u_i = u_h(t_i)$ und definieren die interpolierende Gitterfunktion $\tilde{u}_h = r_h u$ (also mit $\tilde{u}_h(t_i) = \tilde{u}_i = u(t_i)$). Aus der Definition von τ_h folgt nun sofort

$$\tilde{u}_{i+1} = \tilde{u}_i + h\big(\phi(t_i, \tilde{u}_i; h) + \tau_h(t_{i+1})\big).$$

Aus der Stabilität des Einschrittverfahrens folgt sofort

$$\|E_h\|_{\infty,h} = \|u_h - r_h u\|_{\infty,h} = \|u_h - \tilde{u}_h\|_{\infty,h} \le C\|\tau_h\|_{\infty,h}.$$

Konvergenz und Konvergenzrate folgen nun aus der Konsistenz. □

Bemerkung 2.19. Die Aussage des vorigen Satzes lässt sich formulieren als

$$\text{Konsistenz} + \text{(diskrete) Stabilität} \implies \text{Konvergenz}.$$

Mit etwas Mühe lässt sich im Prinzip auch die Umkehrung zeigen. Diese überaus wichtige Einsicht geht auf Lax und Richtmeyer zurück.

Beispiele für Einschrittverfahren

Wir stellen kurz einige einfache Einschrittverfahren und die Prinzipien vor, mit denen sie hergeleitet werden, und weisen dann die benötigten Eigenschaften nach.

Beispiel 2.20 (Explizites Eulerverfahren)**.** Wir wissen bereits, dass dieses Verfahren die Form

$$u_{i+1} = u_i + hf(t_i, u_i)$$

hat, also $\phi(t_i, u_i; h) = f(t_i, u_i)$. Wir wissen auch, dass das explizite Eulerverfahren durch die Verwendung des Vorwärtsdifferenzenquotienten (Bemerkung 2.10) oder auch durch die Verwendung der linksseitigen Rechtecksregel (Bemerkung 2.9)

motiviert werden kann. Aus Lipschitz-Stetigkeit von f erhält man die Lipschitz-Stetigkeit von ϕ und mit Satz 2.15 die diskrete Stabilität. Für stetiges f ist auch die Konsistenz des Verfahrens gesichert. Falls f stetig differenziebar ist, dann können wir sogar eine Konvergenzordnung zeigen. Zuerst bemerken wir, dass eine Taylor-Entwicklung mit Lagrange-Restglied

$$u(t+h) = u(t) + hu'(t) + \tfrac{h^2}{2}u''(\tilde{t})$$

für ein $\tilde{t} \in [t, t+h]$ ergibt. Mit der Definition von d_h, dieser Taylor-Entwicklung, (2.1) in Verbindung mit der Kettenregel und nochmals (2.1) erhalten wir

$$\begin{aligned} d_h(t+h, u(t+h)) &= u(t+h) - u(t) - hf(t, u(t)) \\ &= \big(u(t) + hu'(t) + \tfrac{h^2}{2}u''(\tilde{t})\big) - u(t) - hu'(t) \\ &= \tfrac{h^2}{2}u''(\tilde{t}) \\ &= \tfrac{h^2}{2}\big(\partial_t f(\tilde{t}, u(\tilde{t})) + \partial_u f(\tilde{t}, u(\tilde{t}))u'(\tilde{t})\big). \\ &= \tfrac{h^2}{2}\big(\partial_t f(\tilde{t}, u(\tilde{t})) + \partial_u f(\tilde{t}, u(\tilde{t}))f(\tilde{t}, u(\tilde{t}))\big) \end{aligned}$$

Mit $\tau_h(t,u) = d_h(t,u)/h$ folgt dann

$$|\tau_h(u,t)| \leq \tfrac{h}{2}(\|\partial_t f\|_\infty + \|\partial_u f\|_\infty \|f\|_\infty),$$

wobei $\|f\|_\infty = \sup_{t,y} |f(t,y)|$ ist. Das explizite Eulerverfahren besitzt also mindestens Konsistenzordnung $p = 1$. Eine höhere Ordnung liegt tatsächlich nicht vor, was man durch Taylorentwicklung bis zur nächsten Ordnung überprüft.

Beispiel 2.21 (Implizites Eulerverfahren)**.** Das explizite Eulerverfahren war dadurch motiviert, dass das Integral in

$$u(t+h) = u(t) + \int_t^{t+h} f(s, u(s))\mathrm{d}s \tag{2.13}$$

durch die linksseitige Rechtecksregel approximiert wurde. Verwendet man stattdessen die rechtseitige Rechtecksregel

$$\int_{t_i}^{t_{i+1}} f(s, u(s))\mathrm{d}s \approx hf(t_{i+1}, u(t_{i+1}))$$

und setzen diese in die obige Formel ein, erhalten wir das implizite Eulerverfahren

$$u_{i+1} = u_i + hf(t_{i+1}, u_{i+1}).$$

Man beachte, dass der Wert u_{i+1} hier nur implizit, d.h. über die Lösung einer (nichtlinearen) Gleichung, bestimmt ist. Durch Anwenden des impliziten

Funktionensatzes lässt sich (für hinreichend kleine Schrittweite h) eine (eindeutige) Lösbarkeit zeigen. Durch Umstellen erhält man

$$\frac{u_{i+1} - u_i}{h} = f(t_{i+1}, u_{i+1}) = u'(t_{i+1}).$$

Das bedeutet, dass das implizite Eulerverfahren auch dadurch motiviert werden kann, dass $u'(t_{i+1})$ durch den Rückwärtsdifferentenquotienten approximiert wird.

Das implizite Eulerverfahren lässt sich durch ein Verfahrensfunktion $\phi(t_i, u_i; h)$ beschreiben, die aber nur implizit gegeben, also nicht wirklich bekannt, ist:

$$\phi(t_i, u_i; h) = hf(t_i + h, u_{i+1}) = hf(t_i + h, u_i + h\phi(t_i, u_i; h)).$$

Über den impliziten Funktionensatz und Taylorentwicklung kann man trotzdem wieder die diskrete Stabilität des Verfahrens und die Konsistenzordnung 1 nachweisen; mehr dazu im nächsten Kapitel.

Beispiel 2.22 (Verbessertes Eulerverfahren)**.** Gehen wir erneut von der Integralformel (2.13) aus, dann wissen wir, dass wir bei Anwendung der Mittelpunktregel einen kleineren Fehler machen, als bei Anwendung einer den beiden Rechtecksregeln. Ganz analog sieht man, mit Hilfe der Taylorentwicklung sieht man leicht, dass zwar nur

$$\frac{u(t+h) - u(t)}{h} = u'(t) + \mathcal{O}(h^2) \quad \text{und} \quad \frac{u(t+h) - u(t)}{h} = u'(t+h) + \mathcal{O}(h^2),$$

aber

$$\frac{u(t+h) - u(t)}{h} = u'(t + \tfrac{h}{2}) + \mathcal{O}(h^3),$$

falls u zumindest zweimal stetig differenzierbar ist. Dies führt auf den Ansatz

$$u_{i+1} = u_i + hf(t_i + \tfrac{h}{2}, \hat{u}_{i+1/2}).$$

Zur Bestimmung des noch unbekannten Wertes $\hat{u}_{i+1/2}$ bietet sich an, das explizite Eulerverfahren zu verwenden, also

$$\hat{u}_{i+1/2} = u_i + \tfrac{h}{2} f(t_i, u_i).$$

Setzt man diese Vorschrift in die erste ein, so erhält man ein Einschrittverfahren mit Verfahrensfunktion

$$\phi(t, u; h) = f(t + \tfrac{h}{2}, u + \tfrac{h}{2} f(t, u)).$$

Aus der Lipschitz-Stetigkeit von f folgt sofort wieder die entsprechende Eigenschaft für ϕ und mit Satz 2.15 die diskrete Stabilität. Auch die Konsistenz des Verfahrens ergibt sich wieder sofort aus der Stetigkeit von f. Mit Taylorentwicklung zeigt man weiters, ähnlich wie im vorhergehenden Beispiel, dass hier sogar Konsistenzordnung $p = 2$ vorliegt; siehe Übung.

3. Runge-Kutta-Verfahren

Wir beschäftigen uns in diesem Kapitel mit der systematischen Konstruktion stabiler Einschrittverfahren

$$u_{i+1} = u_i + h\phi(t_i, u_i; h), \qquad t_{i+1} = t_i + h \tag{3.1}$$

mit hoher Konsistenzordnung und somit auch Konvergenzordnung. Dazu betrachten wir eine wichtige Klasse von Methoden, die Runge-Kutta-Verfahren. Für weitere Verfahren (Extrapolationsmethoden, Taylormethode, Kollokationsverfahren) sei auf die einschlägige Literatur verwiesen.

Konstruktion und einfache Beispiele

Die Idee der Runge-Kutta ist, dass man in der Integraldarstellung der Lösung

$$u(t+h) = u(t) + \int_t^{t+h} f(s, u(s))\mathrm{d}s = u(t) + h\int_0^1 f(t+\gamma h, u(t+\gamma h))\mathrm{d}\gamma$$

das Integral durch numerische Integration ersetzt. Für die Näherung $u_i \approx u(t_i)$ erhält man hieraus die Vorschrift

$$u_{i+1} = u_i + h\sum_{j=1}^{s} \beta_j f(t_i + \gamma_j h, g_i^{(j)}), \tag{3.2}$$

mit Stützstellen γ_j und Integrationsgewichten β_j. Zur Approximation der noch unbekannten Zwischenwerte $g_i^{(j)} \approx u(t_i + \gamma_j h)$ verwenden wir wieder ein Schema derselben Form, also

$$g_i^{(j)} = u_i + h\sum_{k=1}^{s} \alpha_{jk} f(t_i + \gamma_k h, g_i^{(k)}). \tag{3.3}$$

Dabei sind γ_k die Stützstellen und α_{jk} Gewichte für eine Integrationsformel zur Approximation des Integrals $\int_0^{\gamma_j} \cdot\, \mathrm{d}\gamma$.

Definition 3.1 (Runge-Kutta-Verfahren)**.**

(i) Eine Methode der Form (3.2)–(3.3) heißt *Runge-Kutta-Verfahren* mit s Stufen.

(ii) Gilt $\alpha_{jk} = 0$ für $k \geq j$, dann heißt das Verfahren *explizit*, andernfalls *implizit*.

Bemerkung 3.2 (Butcher-Tableau)**.** Ein Runge-Kutta-Verfahren ist durch das Koeffizientenschema (Butcher-Tableau)

$$\begin{array}{c|ccc} \gamma_1 & \alpha_{11} & \cdots & \alpha_{1s} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_s & \alpha_{s1} & \dots & \alpha_{ss} \\ \hline & \beta_1 & \cdots & \beta_s \end{array}$$

bereits eindeutig bestimmt. Wir schreiben dafür auch zur Abkürzung $\begin{array}{c|c} \gamma & A \\ \hline & \beta^\top \end{array}$.

An der Matrix A erkennt man sofort, ob das Verfahren implizit oder explizit ist.

Bemerkung 3.3 (Alternative Darstellung)**.** Wir bezeichnen mit

$$k_i^{(j)} = f(t_i + \gamma_j h, g_i^{(k)})$$

die *Steigungen* an den Zwischenpunkten $t_i + \gamma_j h$. Dann lässt sich das RK-Verfahren alternativ schreiben als

$$\begin{aligned} u_{i+1} &= u_i + h \sum_{j=1}^{s} \beta_j k_i^{(j)} \\ k_i^{(j)} &= f\left(t_i + \gamma_j h, u_i + h \sum_{k=1}^{s} \alpha_{jk} k_i^{(k)} \right), \qquad j = 1, \dots, s \end{aligned}$$

Diese k-Form des Verfahrens ist äquivalent mit der g-Form (3.2)–(3.3), d.h., die erzeugten Näherungen u_i stimmen überein; siehe Übung.

Wie die folgenden Beispiele illustrieren, lassen sich die bereits besprochenen Einschrittverfahren allesamt auch als RK-Schemas darstellen. Wir geben dazu einfach die zugehörigen Butcher-Tableaus an.

Beispiel 3.4 (Explizites Eulerverfahren)**.** Das explizite Eulerverfahren lässt sich schreiben als

$$\begin{aligned} g_i^{(1)} &= u_i + h[0 \cdot f(t_i + 0 \cdot h, g_i^{(1)})] \\ u_{i+1} &= u_i + h[1 \cdot f(t_i + 0 \cdot h, g_i^{(1)})] \end{aligned} \qquad \begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}$$

Setzt man den Ausdruck für den Zwischenwert $g_i^{(1)}$ in die zweite Formel ein, so erhält man wieder $u_{i+1} = u_i + hf(t_i, u_i)$. Das explizite Eulerverfahren ist also ein explizites RK-Verfahren mit einer Stufe.

Beispiel 3.5 (Implizites Eulerverfahren)**.** Das implizite Eulerverfahren lässst sich analog schreiben als

$$\begin{aligned} g_i^{(1)} &= u_i + h[1 \cdot f(t_i + 1 \cdot h, g_i^{(1)})] \\ u_{i+1} &= u_i + h[1 \cdot f(t_i + 1 \cdot h, g_i^{(1)})] \end{aligned} \qquad \begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}$$

Man beachte, dass hier $g_i^{(1)} = u_{i+1}$ gilt und implizit, d.h. als Lösung einer nichtlinearen Gleichung, definiert ist. Setzt man $g_i^{(1)}$ in die zweite Formel ein, so erhält man die Verfahrensvorschrift $u_{i+1} = u_i + hf(t_{i+1}, u_{i+1})$. Das implizite Eulerverfahren ist also ein implizites RK-Verfahren mit einer Stufe.

Beispiel 3.6 (Verbessertes Eulerverfahren)**.** Das verbesserte Eulerverfahren kann geschrieben werden als

$$\begin{aligned} g_i^{(1)} &= u_i + h[0 \cdot f(t_i, g_i^{(1)}) + 0 \cdot f(t_i + \tfrac{1}{2}h, g_i^{(2)})] \\ g_i^{(2)} &= u_i + h[\tfrac{1}{2} \cdot f(t_i, g_i^{(1)}) + 0 \cdot f(t_i + \tfrac{1}{2}h, g_i^{(2)})] \\ u_{i+1} &= u_i + h[0 \cdot f(t_i, g_i^{(1)}) + 1 \cdot f(t_i + \tfrac{1}{2}h, g_i^{(2)})] \end{aligned} \qquad \begin{array}{c|cc} 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \hline & 0 & 1 \end{array}$$

Das verbesserte Eulerverfahren ist also ein explizites RK-Verfahren mit 2 Stufen. Durch Eliminieren der Zwischenwerte $g_i^{(j)}$ erhält man wieder die kompakte Verfahrensvorschrift $u_{i+1} = u_i + hf(t_i + \frac{1}{2}h, u_i + \frac{h}{2}f(t_i, u_i))$.

Ausblick. Im Falle der Durchführbarkeit sind RK-Verfahren stets Einschrittverfahren der Form $u_{i+1} = u_i + h\phi(t_i, u_i; h)$ und zur Analyse sind die Resultate des vorigen Abschnitts anwendbar. Im Weiteren gehen wir daher wie folgt vor:

(a) Wir zeigen, unter welchen Bedingungen RK-Verfahren durchführbar sind.
(b) Systematische Bestimmung der Konsistenzordnung von RK-Verfahren.
(c) Nachweis der diskreten Stabilität (als Einschrittverfahren).

Aus der Konvergenztheorie für Einschrittverfahren erhalten wir dann unmittelbar entsprechende Konvergenzresultate für RK-Verfahren.

Explizite Runge-Kutta-Verfahren

Wir betrachten zunächst im Speziellen explizite Runge-Kutta-Verfahren mit s Stufen. Diese besitzen die allgemeine Form

$$g_i^{(j)} = u_i + h \sum_{k=1}^{j-1} \alpha_{jk} f(t_i + \gamma_k h, g_i^{(k)}), \qquad i = j, \ldots, s$$
$$u_{i+1} = u_i + h \sum_{j=1}^{s} \beta_j f(t_i + \gamma_j h, g_i^{(j)}).$$

Beobachtung. Da in der Summe für die Bestimmung der Zwischenwerte $g_i^{(j)}$ nur mehr die Indizes $k < j$ Verwendung finden, können die Zwischenwerte $g_i^{(j)}$ hier einfach nacheinander berechnet werden. Anschließend kann damit u_{i+1} berechnet werden. Wir erhalten:

Satz 3.7 (Durchführbarkeit expliziter RK-Verfahren)**.**
Sei $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. Dann sind explizite RK-Verfahren stets durchführbar und Einschrittverfahren im Sinne von Definition 2.7. Die Verfahrensfunktion ϕ ergibt sich aus $\phi(t_i, u_i; h) = \sum_{j=1}^{s} \beta_j f(t_i + \gamma_j h, g_i^{(j)})$ durch (wiederholtes) Einsetzen der entsprechenden Ausdrücke für die Zwischenwerte $g_i^{(j)}$.

Aufgrund des einfachen Aufbaus überträgt sich die Lipschitzstetigkeit von f unmittelbar auf die Zwischenwerte $g_i^{(j)}$ und somit auf die Verfahrensfunktion ϕ.

Satz 3.8 (Stabilität expliziter RK-Verfahren)**.** *Sei $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ stetig und Lipschitz-stetig bzgl. des zweiten Arguments. Dann ist auch die Verfahrensfunktion ϕ des expliziten RK-Verfahrens Lipschitz-stetig. Explizite RK-Verfahren sind also stabile Einschrittverfahren.*

Beweis. Mit Induktion zeigt man, dass die Stützwerte $g_i^{(j)}$ und in der Folge auch die Verfahrensfunktion Lipschitz-stetig von u_i abhängen. □

Als nächstes betrachten wir die Konsistenzordnung expliziter RK-Verfahren. Wie die folgenden Beispiele zeigen, basiert dies immer auf Taylorentwicklung.

Beispiel 3.9 (Einstufige explizite RK Verfahren)**.** Wir betrachten allgemeine einstufige Runge-Kutta-Verfahren der Form

$$\begin{aligned} g_i^{(1)} &= u_i \\ u_{i+1} &= u_i + h\beta_1 f(t_i + \gamma_1 h, g_i^{(1)}) \end{aligned} \qquad \begin{array}{c|c} \gamma_1 & 0 \\ \hline & \beta_1 \end{array}$$

mit noch frei wählbaren Parametern γ_1 und β_1. Eliminieren des Zwischenwertes $g_i^{(1)}$ liefert die Verfahrensvorschrift

$$u_{i+1} = u_i + h\beta_1 f(t_i + \gamma_1 h, u_i).$$

Für den lokalen Fehler erhält man durch Taylorentwicklung sowie $u'(t) = f(t, u(t))$ und $u''(t) = \frac{\mathrm{d}}{\mathrm{d}t} f(t, u(t)) = \partial_t f(t, u(t)) + \partial_u f(t, u(t)) u'(t)$ und der Kettenregel

$$\begin{aligned} d_h(t,u) &= u(t+h) - u(t) - h\beta_1 f(t + \gamma_1 h, u(t)) \\ &= \big(u(t) + hu'(t) + \tfrac{h^2}{2} u''(t) + \mathcal{O}(h^3)\big) \\ &\qquad - u(t) - h\beta_1 \big(f(t,u(t)) + \gamma_1 h\, \partial_t f(t,u(t)) + \mathcal{O}(h^2)\big) \\ &= h(1-\beta_1) f(t,u(t)) \\ &\quad + h^2 \Big(\tfrac{1}{2} \partial_t f(t,u(t)) + \tfrac{1}{2} \partial_u f(t,u(t)) f(t,u(t)) - \beta_1 \gamma_1 \partial_t f(t,u(t)) \Big) + \mathcal{O}(h^3). \end{aligned}$$

Die Wahl $\beta_1 = 1$ ist nötig, um den ersten Term zu eliminieren. Der h^2-Term kann jedoch allgemeinen nicht ausgelöscht werden. Aus $\tau_h(t,y) = e_h(t,y)/h$ folgt also:

(i) Die Wahl $\beta_1 = 1$, γ_1 beliebig führt auf Konsistenzordnung $p = 1$.

(ii) Ordnung $p \geq 2$ ist mit einstufigen expliziten RK-Verfahren nicht erreichbar!

(iii) Für $\beta_1 = 1$, $\gamma_1 = 1$ erhält man das explizite Eulerverfahren. Dieses besitzt bereits die maximal erreichbare Konsistenzordnung $p = 1$.

Beispiel 3.10 (Zweistufige explizite RK-Verfahren)**.** Zur Erhöhung der Konsistenzordnung betrachten wir nun zweistufige Verfahren

$$\begin{aligned} g_i^{(1)} &= u_i \\ g_i^{(2)} &= u_i + h\alpha_{21} f(t_i + \gamma_1 h, g_i^{(1)}) \\ u_{i+1} &= u_i + h\beta_1 f(t_i + \gamma_1 h, g_i^{(1)}) + h\beta_2 f(t_i + \gamma_2 h, g_i^{(2)}). \end{aligned} \qquad \begin{array}{c|cc} \gamma_1 & 0 & 0 \\ \gamma_2 & \alpha_{21} & 0 \\ \hline & \beta_1 & \beta_2 \end{array}$$

mit fünf frei wählbaren Parametern $\gamma_1, \gamma_2,\ \beta_1, \beta_2$ und α_{21}. Das Verfahren lässt sich hier kompakt schreiben als

$$u_{i+1} = u_i + h[\beta_1 f(t_i + \gamma_1 h, u_i) + \beta_2 f(u_i + \gamma_2 h, u_i + h\alpha_{21} f(t_i + \gamma_1 h, u_i))].$$

Mit Taylorentwicklung erhält man für den lokalen Fehler

$$
\begin{aligned}
&d_h(t,u)\\
&= u(t+h) - u(t) - h\beta_1 f(t+\gamma_1 h, u) - h\beta_2 f(t+\gamma_2 h, y + h\alpha_{21} f(t+\gamma_1 h, u))\\
&= u(t) + hu'(t) + \tfrac{h^2}{2}u''(t) + \tfrac{h^3}{6}u'''(t) + \mathcal{O}(h^4)\\
&\quad - u(t)\\
&\quad - h\beta_1[f(t,u) + \partial_t f(t,u)\gamma_1 h + \tfrac{1}{2}\partial_{tt} f(t,y)\gamma_1^2 h^2 + \mathcal{O}(h^3)]\\
&\quad - h\beta_2[f(t,u) + h\gamma_2\partial_t f(t,u) + h\alpha_{21}\partial_u f(t,u)(f(t,u) + h\gamma_1\partial_t f(t,u) + \mathcal{O}(h^2))\\
&\qquad\quad + \tfrac{1}{2}h^2\gamma_2^2\partial_{tt} f(t,u) + h^2\gamma_2\alpha_{21}(f(t,u) + \mathcal{O}(h))\\
&\qquad\quad + \tfrac{1}{2}(h\alpha_{21})^2\partial_{uu} f(t,u)(f(t,u) + \mathcal{O}(h))]\\
&= h(1-\beta_1-\beta_2)f + h^2\big((\tfrac{1}{2} - \beta_1\gamma_1 - \beta_2\gamma_2)\partial_t f + (\tfrac{1}{2} - \beta_2\alpha_{21})\partial_u f\ f\big)\\
&\qquad + h^3(...) + \mathcal{O}(h^4).
\end{aligned}
$$

Die Bedingung $\beta_1 + \beta_2 = 1$ garantiert, dass die Terme der Ordnung h verschwinden. Für die h^2-Terme sind die Bedingungen $\beta_1\gamma_1 + \beta_2\gamma_2 = \frac{1}{2}$ sowie $\beta_2\alpha_{21} = \frac{1}{2}$ nötig. Die zusätzlichen Bedingungen, die zum Eliminieren der Terme mit h^3 nötig wären, sind nicht zu erfüllen. Mittels $\tau_h = e_h/h$ erhalten wir somit:

(i) Erfüllt das zweistufige explizite Verfahren die Bedingung

$$\beta_1 + \beta_2 = 1,$$

so besitzt es mindestens Ordnung 1.

(ii) Gilt zusätzlich

$$\beta_1\gamma_1 + \beta_2\gamma_2 = \tfrac{1}{2}, \qquad \beta_2\alpha_{21} = \tfrac{1}{2},$$

dann besitzt das Verfahren mindestens Ordnung 2.

(iii) Ordnung $p \geq 3$ ist mit einem expliziten zweistufigen expliziten Verfahren im Allgemeinen wiederum nicht zu erreichen.

Als zwei häufig verwendete zweistufige Verfahren der Ordnung 2 erwähnen wir

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

verbessertes Eulerverfahren Methode von Heun (=exp. Trapezregel)

Beobachtung. Über Taylorentwicklung lassen sich systematisch (wenn auch mit langwieriger Rechnung) notwendige und hinreichende Bedingungen an die Koeffizienten herleiten, welche zum Erreichen einer bestimmten Konsistenzordnung

nötig sind. Um die Anzahl der freien Parameter zu verringern, betrachten wir im folgenden nur Verfahren, welche der *Autonomieinvarianzbedingung*

$$\gamma_j = \sum_{k=1}^{s} \alpha_{j,k}, \qquad j = 1, \dots, s \tag{AI}$$

genügen. Dies hat folgende einfache Begründung.

Satz 3.11 (Autonomieinvarianz)**.** *Falls die Autonomieinvarieznbedingung* (AI) *gilt, dann erzeugt das RK-Verfahren mit Koeffizienten $\gamma_j, \alpha_{j,k}, \beta_j$ für das Anfangswertproblem*

$$u'(t) = f(t, u(t)), \quad u(0) = u_0$$

und das entsprechende autonome System

$$\begin{pmatrix} t'(s) \\ z'(s) \end{pmatrix} = \begin{pmatrix} 1 \\ f(t(s), z(s)) \end{pmatrix}, \qquad \begin{pmatrix} t(0) \\ z(0) \end{pmatrix} = \begin{pmatrix} 0 \\ u_0 \end{pmatrix},$$

dieselben Iterierten, d.h. $z_i = u_i$.

Beweis. Siehe Übung. □

Beachte. Die Koeffizienten γ_j sind durch die AI-Bedingung schon vollständig festgelegt und brauchen daher nicht weiter berücksichtigt zu werden. Für die Bedingungen bis zur Ordnung 4 erhält man dann aus der Taylorentwicklung folgenden Satz.

Satz 3.12 (Konsistenzbedingungen)**.** *Die Funktion f (und somit die Lösungen u zu $u' = f(t, u)$) sei hinreichend glatt.*

(i) Falls

$$\sum_j \beta_j = 1 \tag{O1}$$

gilt und $f \in C^1$, dann ist das Verfahren konsistent und besitzt Ordnung $p \geq 1$.

Für Verfahren höherer Ordnungen betrachten wir nur Verfahren mit (AI).
(ii) Ordnung 2 wird erreicht, wenn $f \in C^2$ und neben (AI) und (O1) gilt:

$$\sum_{j,k} \beta_j \alpha_{jk} = \tfrac{1}{2}. \tag{O2}$$

(iii) Ordnung 3 wird erreicht, wenn $f \in C^2$ und neben (AI), (O1) und (O2) gilt:

$$\sum_{j,k,l} \beta_j \alpha_{jk} \alpha_{jl} = \tfrac{1}{3}, \qquad \sum_{j,k,l} \beta_j \alpha_{jk} \alpha_{kl} = \tfrac{1}{6}. \tag{O3}$$

(iv) Ordnung 4 *wird erreicht, wenn* $f \in C^4$ *und neben (AI) und (O1)–(O3) gilt:*

$$\begin{aligned} \sum_{j,k,l,m} \beta_j \alpha_{jk} \alpha_{jl} \alpha_{jm} &= \tfrac{1}{4}, & \sum_{j,k,l,m} \beta_j \alpha_{jk} \alpha_{jl} \alpha_{lm} &= \tfrac{1}{8} \\ \sum_{j,k,l,m} \beta_j \alpha_{jk} \alpha_{kl} \alpha_{km} &= \tfrac{1}{12}, & \sum_{j,k,l,m} \beta_j \alpha_{jk} \alpha_{kl} \alpha_{lm} &= \tfrac{1}{24}. \end{aligned} \tag{O4}$$

Beweis. Über Taylorentwicklung. Es genügt hierzu, autonome Differentialgleichungen zu betrachten. □

In den folgenden Beispielen stellen wir überblicksmäßig einige der gängigen expliziten Runge-Kutta-Verfahren mit 1 bis 4 Stufen vor.

Beispiel 3.13 (einstufige explizite Verfahren: $s = 1$, $p = 1$)**.**

0	0
	1

explizites Eulerverfahren

Beispiel 3.14 (zweistufige explizite Verfahren: $s = 2$, $p = 2$)**.**

0	0	
1	1	0
	$\frac{1}{2}$	$\frac{1}{2}$

Verfahren von Heun

0	0	
$\frac{1}{2}$	$\frac{1}{2}$	0
	0	1

verbessertes Eulerverfahren

Beispiel 3.15 (dreistufige explizite Verfahren: $s = 3$, $p = 3$)**.**

0	0		
$\frac{1}{3}$	$\frac{1}{3}$	0	
$\frac{2}{3}$	0	$\frac{2}{3}$	0
	$\frac{1}{4}$	0	$\frac{3}{4}$

Verfahren von Heun 3. Ordnung

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

Verfahren von Kutta 3. Ordnung

Beispiel 3.16 (vierstufige explizite Verfahren: $s = 4$, $p = 4$)**.**

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

Runge-Kutta-Verfahren 4. Ord.

0	0			
$\frac{1}{3}$	$\frac{1}{3}$	0		
$\frac{2}{3}$	$-\frac{1}{3}$	1	0	
1	1	-1	1	0
	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

3/8-Regel

Zum Abschluss dieses Kapitels wollen wir noch anmerken, welche Ordnungen mit expliziten s-stufigen RK-Verfahren theoretisch erreichbar sind.

Bemerkung 3.17 (Maximal erreichbare Ordnung, Butcher-Barrieren)**.** Die Anzahl der Ordnungsbedingungen nimmt mit zunehmender Ordnung sehr schnell zu – deutlich schneller als die Anzahl der verfügbaren Parameter bei Erhöhung der Stufenzahl. Die mindestens erforderliche Stufenzahl zum Erreichen einer bestimmten Ordnung ist in folgender Tabelle zusammengefasst.

Stufenzahl (s)	1	2	3	4	5	6	7	8	9
max. Ordnung (p)	1	2	3	4	4	5	6	6	7

Maximal erreichbare Ordnung expliziter RK-Verfahren

Für ein Verfahren der Ordnung $p \geq 1$ ist also mindestens Stufenzahl $s \geq p$ erforderlich; für Ordnung $p \geq 5$ mindestens Stufe $s \geq p + 1$, für $p \geq 7$ benötigt man $s \geq p + 2$ usw. Diese Schranken sind als *Butcher-Barrieren* bekannt.

Implizite Runge-Kutta-Verfahren

Wir betrachten jetzt allgemeine, insbesondere implizite, Runge-Kutta-Verfahren, wobei wir jeweils von der k-Form ausgehen:

$$u_{i+1} = u_i + h \sum\nolimits_{j=1}^{s} \beta_j k^{(j)}, \tag{3.4}$$

$$k^{(j)} = f\big(t_i + \gamma_j h, u_i + h \sum\nolimits_{l=1}^{s} \alpha_{jl} k^{(l)}\big), \qquad j = 1, \ldots, s. \tag{3.5}$$

Im Gegensatz zu expliziten Verfahren lassen sich die Steigungen $k = (k^{(1)}, \ldots, k^{(s)})$ hier im Allgemeinen nur implizit über Lösung eines nichtlinearen Gleichungssystems bestimmen. Wir schreiben das System (3.5) auch kurz als

$$k = \psi(k), \tag{3.6}$$

wobei $\psi = (\psi_1, \dots, \psi_s)^\top : \mathbb{R}^s \to \mathbb{R}^s$ mit Komponenten

$$\psi_j(k) = f\big(t_i + \gamma_j h, u_i + h \sum\nolimits_{l=1}^{s} k^{(l)}\big).$$

Wir weisen zunächst nach, dass und unter welchen Bedingungen das Verfahren durchführbar ist.

Satz 3.18 (Durchführbarkeit impliziter RK-Verfahren)**.** *Sei $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ stetig, gleichmäßig beschränkt und Lipschitz-stetig bzgl. des zweiten Arguments, d.h.*

$$|f(t,y)| \le M, \qquad |f(t,x) - f(t,y)| \le L|x-y|, \qquad \forall x, y, t.$$

Weiters sei $hL\|A\|_\infty < 1$, wobei $A = [\alpha_{ij}]_{i,j=1}^s$ die Koeffizientenmatrix des Butcher-Tableaus und $\|A\|_\infty = \max_i \sum_j |\alpha_{i,j}|$ die Zeilensummennorm ist. Dann hat das nichtlineare Gleichungssystem (3.6) *eine eindeutige Lösung und die Iteration $k_{n+1} = \psi(k_n)$ konvergiert für $k_0 = \mathbf{0}$ gegen die Lösung $k = \psi(k)$.*

Beweis. Wir überprüfen die Voraussetzungen des Banach'schen Fixpunktsatzes.
(i) Die Menge $\mathcal{M} = \{k \in \mathbb{R}^s : \|k\|_\infty \le M\}$ ist eine abgeschlossene, beschränkte und nicht leere Teilmenge von $(\mathbb{R}^s, \|\cdot\|_\infty)$.
(ii) $\psi : \mathcal{M} \to \mathcal{M}$ ist eine Selbstabbildung.
(iii) ψ ist eine Kontraktion auf $\mathcal{M}$, denn

$$\begin{aligned}
&\|\psi(k) - \psi(\tilde{k})\|_\infty \\
&\quad = \max_j |f(t_i + \gamma_j h, u_i + h\textstyle\sum_l \alpha_{jl} k^{(l)}) - f(t_i + \gamma_j h, u_i + h\sum_l \alpha_{jl} \tilde{k}^{(l)})| \\
&\quad \le Lh \max_j \textstyle\sum_l |\alpha_{jl}||k^{(l)} - \tilde{k}^{(l)}| \le Lh\|A\|_\infty \|k - \tilde{k}\|_\infty.
\end{aligned}$$

Die Kontraktionseigenschaft folgt nun aus der Annahme, dass $Lh\|A\|_\infty < 1$ ist. Der B-FPS liefert nun die Existenz einer eindeutigen Lösung und darüber hinaus noch die Konvergenz des Iterationsverfahrens. □

Bemerkung 3.19. Das RK-Verfahren ist für hinreichend kleine Schrittweiten ($hL \preceq 1$) also stets durchführbar und die Berechnung der Steigungen k kann durch die Fixpunktiteration $k_{n+1} = \psi(k_n)$ erfolgen. Als Startwert ist $k_0 = \mathbf{0}$ möglich. Für $hL \ll 1$ konvergiert die Iteration sehr schnell! Für den obigen Beweis wird nur benötigt, dass f lokal, in einer kleinen Umgebung der aktuellen Näherung (t_i, u_i), stetig (und somit beschränkt) und gleichmäßig Lipschitz stetig ist. Analoge Überlegungen gelten natürlich auch für Systeme von DGLen.

Mit ähnlichen Überlegungen zeigt man folgende Aussage.

Satz 3.20 (Diskrete Stabilität)**.** *Sei f stetig und Lipschitz-stetig bzgl. des zweiten Arguments. Ist $hL\|A\|_\infty < 1$, so gilt*

$$\|k - \tilde{k}\|_\infty \leq \frac{L}{1 - hL\|A\|_\infty} |u_i - \tilde{u}_i|,$$

wobei k und $\tilde{k}$ die Lösungen der Gleichung (3.5) *mit Argument u_i bzw $\tilde{u}_i$ ist.*

Beweis. Siehe Übung. □

Als unmittelbare Folgerung dieser Abschätzungen und der Überlegungen zu allgemeinen Einschrittverfahren erhalten wir folgendes Resultat.

Satz 3.21 (Stabilität impliziter RK-Verfahren)**.** *Unter den Voraussetzungen der vorigen Sätze ist das implizite RK-Verfahren ein stabiles Einschrittverfahren im Sinne von Kapitel 2.*

Beweis. Das Verfahren hat die Form $u_{i+1} = u_i + h\sum_j \beta_j k^{(j)} =: u_i + h\phi(t_i, u_i; h)$. Wegen der Lipschitz-stetigen Abhängigkeit der Steigungen $k^{(j)}$ ist auch die Verfahrensfunktion ϕ Lipschitz-stetig bzgl. u. □

Als Nächstes zitieren wir noch Aussagen zur Konsistenz impliziter RK-Verfahren.

Satz 3.22 (Konsistenz impliziter RK-Verfahren)**.** *Die Aussagen über Konsistenz und Konsistenzordnung aus Satz 3.12 gelten analog auch für implizite RK-Verfahren.*

Beweis. Folgt mit implizitem Funktionensatz und Taylorentwicklung; für Details sei auf das Buch Deuflhard&Bornemann: *Numerische Mathematik 2. Gewöhnliche Differentialgleichungen* verwiesen. □

Wir präsentieren zu Abschluss der Konvergenzanalyse noch zwei weitere Sätze, welche die Überprüfung der Ordnung bei impliziten RK-Verfahren erheblich erleichtern. Zur Motivation der Aussagen erinnern wir daran, dass

$$\begin{aligned} u_{i+1} &= u_i + h\sum_k \beta_k f(t_i + \gamma_k h, g_i^{(k)}) \\ &\approx u(t_i + h) = u(t_i) + h\int_0^1 f(t_i + \gamma h, u(t_i + \gamma h))\mathrm{d}\gamma \\ g_i^{(j)} &= u_i + h\sum_k \alpha_{jk} f(t_i + \gamma_k h, g_i^{(k)}) \\ &\approx u(t_i) + h\int_0^{\gamma_j} f(t_i + \gamma h, y(t_i + \gamma h))\mathrm{d}\gamma. \end{aligned}$$

Man sieht, dass (γ_k, β_k) bzw (γ_k, α_{jk}) Stützstellen und Gewichte für numerische Integrationsformeln zur Approximation der Integrale $\int_0^1 \cdot \,\mathrm{d}\gamma$ bzw $\int_0^{\gamma_j} \cdot \,\mathrm{d}\gamma$ sind.

Satz 3.23 (Konsistenbedingung)**.** *Sei $f \in C^p$ und $hL\|A\|_\infty < 1$. Weiters gelte*

$$\sum_{k=1}^{s} \alpha_{jk}\gamma_k^l = \tfrac{\gamma_j^{l+1}}{l+1}, \qquad l = 0, \dots, q, \qquad j = 1, \dots, s, \tag{3.7}$$

$$\sum_{j=1}^{s} \beta_j\gamma_j^l = \tfrac{1}{l+1}, \qquad l = 0, \dots, p-1 \tag{3.8}$$

mit $q = p-2$. Dann besitzt die entsprechende RK-Formel mindestens Konsistenzordnung p.

Beweis. Siehe Übung. □

Bemerkung 3.24. Die Bedingung (3.8) ist äquivalent dazu, dass die Quadraturformel $Q(w) := \sum_j \beta_j w(\gamma_j)$ für alle Polynomfunktionen w vom Grad $p-1$ exakt ist ($Q(w) = \int_0^1 w(x)\,\mathrm{d}x$). Analog ist (3.7) äquivalent dazu, dass $Q_j(w) := \sum_k \alpha_{jk} w(\gamma_k) = \int_0^{\gamma_j} w(x)\,\mathrm{d}x$ für alle Polynomfunktionen w vom Grad q.

Wir stellen ferner fest, dass (3.7) mit Wahl $l = 0$ der Autonomieinvarianzbedingung (AI) gleicht.

Durch Wahl hinreichend guter Quadraturformeln kann man also mit impliziten Verfahren stets zumindest Ordnung $p = s$ erreichen. Wie der folgende Satz belegt, ist sogar eine noch höhere Ordnung möglich.

Satz 3.25 (Verallgemeinerte Konsistenzbedingungen)**.**
Sei $f \in C^p$ und $hL\|A\|_\infty < 1$. Falls

(i) die Bedingungen (3.7), (3.8) *mit $q = s-1$ und $s \le p \le 2s$ gelten,* oder

(ii) die Bedingungen (3.7), (3.8) *mit $q = s-2$ und $s-1 \le p \le 2s-1$ und $\gamma_s = 1$ und $\alpha_{1,s} = \alpha_{2,s} = \cdots = \alpha_{s,s} = 0$ gelten,*

dann besitzt die RK-Formel mindestens Konsistenzordnung p.

Beweis. Siehe *Grigorieff: Numerik gewöhnlicher Differentialgleichungen*, Teil 1, Satz (38). □

Mit diesem Satz lassen sich insbesondere die RK-Formeln von Gauß, Radau, und Lobatto motivieren und ihre Ordnungen bestimmen; siehe unten.

Bemerkung. Die Bedeutung obiger Sätze lässt sich wie folgt zusammenfassen:
(i) Gute Quadraturformeln (γ_k, β_k), (γ_k, α_{jk}) führen zu Verfahren hoher Ordnung.
(ii) Die maximal erreichbare Ordnung ist $p = 2s$ (vgl. Gauß-Quadraturformeln!).
(iii) Für explizite RK-Verfahren sind die Kriterien jedoch nicht erfüllbar.

Zum Abschluss führen wir noch wichtige Beispiele impliziter RK-Verfahren an.

Beispiel 3.26 (Gauß-Formeln: $p = 2s$)**.** Wir wählen (γ_k, β_k) als Stützstellen und Gewichte der Gauß-Quadraturformel. Somit ist (3.8) für $l = 0, \ldots, 2s-1$ erfüllt. Für jedes j hat man s Parameter α_{jk} mit denen man die Exaktheitsbedingungen (3.7) für $l = 0, \ldots, s-1$ erfüllen kann. Dies geschieht durch Lösen linearer Gleichungssysteme. Satz 3.25 liefert somit Ordnung $p = 2s$.

Gauß-1 (=implizite Mittelpunktsregel): $s = 1$, $p = 2s = 2$

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

Gauß-2: $s = 2$, $p = 4$

$$\begin{array}{c|cc} \frac{1}{2} - \frac{\sqrt{3}}{6} & \frac{1}{4} & \frac{1}{4} - \frac{\sqrt{3}}{6} \\ \frac{1}{2} + \frac{\sqrt{3}}{6} & \frac{1}{4} + \frac{\sqrt{3}}{6} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

Gauß-3: $s = 3$, $p = 6$

$$\begin{array}{c|ccc} \frac{1}{2} - \frac{\sqrt{15}}{10} & \frac{5}{36} & \frac{2}{9} - \frac{\sqrt{15}}{15} & \frac{5}{36} - \frac{\sqrt{15}}{30} \\ \frac{1}{2} & \frac{5}{36} + \frac{\sqrt{15}}{24} & \frac{2}{9} & \frac{5}{36} - \frac{\sqrt{15}}{24} \\ \frac{1}{2} + \frac{\sqrt{15}}{10} & \frac{5}{36} + \frac{\sqrt{15}}{30} & \frac{2}{9} + \frac{\sqrt{15}}{15} & \frac{5}{36} \\ \hline & \frac{5}{18} & \frac{4}{9} & \frac{5}{18} \end{array}$$

Beispiel 3.27 (Radau Formeln: $p = 2s - 1$)**.** Bei diesen Formeln wird entweder $\gamma_1 = 0$ oder $\gamma_s = 1$ fixiert. Damit reduziert sich die maximal erreichbare Ordnung der Quadraturformel (γ_k, β_k) um 1. Die restlichen Integrationsstellen γ_k werden optimal gewählt und die Integrationsgewichte β_j über die Bedingungen (3.8) für $l = 0, \ldots, s-1$ bestimmt. Bei den Gewichten α_{jk} kann man entweder $\alpha_{j1} = \beta_1$ oder $\alpha_{sj} = \beta_j$ und bestimmt die restlichen über die Bedingungen (3.7). Man erhält die Radau-I-A bzw. Radau-II-A Formeln.

Radau-IA: $s = 1$, $p = 1$

0	0
	1

Radau-IIA: $s = 1$, $p = 1$

1	1
	1

Radau-IA: $s = 2$, $p = 3$

0	$\frac{1}{4}$	$-\frac{1}{4}$
$\frac{2}{3}$	$\frac{1}{4}$	$\frac{5}{12}$
	$\frac{1}{4}$	$\frac{3}{4}$

Radau-IIA: $s = 2$, $p = 3$

$\frac{1}{3}$	$\frac{5}{12}$	$-\frac{1}{12}$
1	$\frac{3}{4}$	$\frac{1}{4}$
	$\frac{3}{4}$	$\frac{1}{4}$

Beispiel 3.28 (Lobatto Formeln: $p = 2s - 2$)**.** Hier wählt man $\gamma_1 = 0$ und $\gamma_s = 1$ und bestimmt die restlichen Stützstellen optimal. Der Exaktheitsgrad der Integrationsformel (γ_k, β_k) ist durch die Zusatzbedingungen um 2 reduziert. Durch weitere Einschänkungen an die Koeffizienten α_{jk} erhält man

$s = 2$, $p = 2$:

Lobatto-III-A

0	0	0
1	$\frac{1}{2}$	$\frac{1}{2}$
	$\frac{1}{2}$	$\frac{1}{2}$

Lobatto-III-B

0	$\frac{1}{2}$	0
1	$\frac{1}{2}$	0
	$\frac{1}{2}$	$\frac{1}{2}$

Lobatto-III-C

0	$\frac{1}{2}$	$-\frac{1}{2}$
1	$\frac{1}{2}$	$\frac{1}{2}$
	$\frac{1}{2}$	$\frac{1}{2}$

$s = 3$, $p = 4$:

Lobatto-III-A

0	0	0	0
$\frac{1}{2}$	$\frac{5}{24}$	$\frac{1}{3}$	$-\frac{1}{24}$
1	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$
	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$

Lobatto-III-B

0	$\frac{1}{6}$	$-\frac{1}{6}$	0
$\frac{1}{2}$	$\frac{1}{6}$	$\frac{1}{3}$	0
1	$\frac{1}{6}$	$\frac{5}{6}$	0
	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$

Lobatto-III-C

0	$\frac{1}{6}$	$-\frac{1}{3}$	$\frac{1}{6}$
$\frac{1}{2}$	$\frac{1}{6}$	$\frac{5}{12}$	$-\frac{1}{12}$
1	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$
	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$

Bemerkungen. Die zweistufige Lobatto-III-B und III-C Formel sind nicht autonomieinvariant. Die Lobatto-III-A Formel mit $s = 2$ heißt auch implizite-Trapezregel oder Cranck-Nicholson Verfahren. Für die Lobatto-III-A Formel ist $g_i^{(1)} = u_i$. Aufwand und Genauigkeit sind also ähnlich zum Gauß-Verfahren mit $(s-1)$ Stufen.

Im nächsten Abschnitt zeigen wir, dass implizite RK-Formeln neben den besseren Konsistenzordnungen noch weitere Vorzüge bezüglich ihrer Stabilität besitzen.

Abschließende Bemerkungen

Fehlerschätzung und adaptive Schrittweitensteuerung

Für die effiziente Implementierung eines RK-Verfahrens ist es ratsam, die Schrittweiten h_i adaptiv zu wählen, und zwar so, dass eine bestimmte Genauigkeit des globalen Fehlers E_h garantiert werden kann. Dies kann über Fehlerschätzung, z.B. durch Extrapolation oder eingebettete Runge-Kutta-Verfahren realisiert werden.

Effiziente Realisierung impliziter Verfahren

Bei impliziten RK-Verfahren hoher Ordnung müssen mitunter große gekoppelte nichtlineare Gleichungssysteme gelöst werden, die mit der Stufenzahl s größer werden. Um die Effizienz zu steigern, kann man diagonal-implizite Runge-Kutta (DIRK) Verfahren betrachten; diese haben $\alpha_{j,k} = 0$ für $k > j$. Die Stützwerte $g_i^{(k)}$ können somit nacheinander (durch Lösen einfacher nichtlineare Gleichungen) berechnet werden. Weitere Vereinfachungen sind die SDIRK Verfahren, welche dieselben Einträge $\alpha_{k,k}$ auf der Diagonalen besitzen. Alternativ dazu können verschiedene Linearisierungen zum Einsatz kommen, bei denen die nichtlinearen Gleichungssysteme nur näherungsweise gelöst. Da führt auf sogenannte linear-implizite RK-Verfahren, wie z.B. Rosenbrock und Wanner-Verfahren.

Mehrschrittverfahren

Eine andere Möglichkeit die Ordnung eines Verfahrens zu steigern besteht darin, alte Funktionswerte zur Extrapolation heranzuziehen. Die wird bei den linearen Mehrschrittverfahren verwendet. Jeder Schritt eines solchen Verfahrens ist im Prinzip in etwa so aufwändig wie ein Schritt des expliziten bzw. impliziten Eulerverfahrens. Gleichzeitig kann jedoch im Prinzip hohe Ordnung erreicht werden. Neben den Vorteilen bei der Effizienz haben die entsprechenden Verfahren jedoch auch Nachteile, z.B. bei Veränderung der Schrittweite oder bezüglich ihrer Stabilitätseigenschaften für steife Probleme; siehe nächster Abschnitt.

Umfassende weiterführende Information zu den angeführten Themen findet man in der, in diesem Abschnitt zitierten einschlägigen Literatur.

4. Stärkere Stabilitätsbegriffe

In den letzten Kapiteln hatten wir Stabilitätsresultate kennengelernt, wo die Schranke an die Lösung u bzw. an die Abweichung $u - \tilde{u}$ exponentiell mit der Zeit ansteigt (vgl. Bemerkung 2.16). Wenngleich diese Resultate im Allgemeinen scharf sind, sind wir an Resultaten interessiert, die für bestimmte relevante Fälle stärkere Aussagen treffen können.

Definition 4.1 (Stabilität)**.**

(i) Ein Differentialgleichungssystem $u'(t) = f(t, u(t)), t > 0$ heißt *stabil (im Sinne von Lyapunov)*, falls alle Lösungen $u, \tilde{u} : [0, \infty) \to \mathbb{R}^N$ des Systems

$$\forall \epsilon > 0 \; \exists \delta > 0 \; \forall t \geq 0 \; : \|u(0) - \tilde{u}(0)\|_{\ell^2} \leq \delta \implies \|u(t) - \tilde{u}(t)\|_{\ell^2} \leq \epsilon$$

erfüllen. (Beachte: Die Funktionen u und $\tilde{u}$ können sich bei den Anfangsbedingungen unterscheiden.)

(ii) Die Gleichung heißt *asymptotisch stabil*, falls sie stabil ist und

$$\|u(0) - \tilde{u}(0)\|_{\ell^2} \leq \delta_0 \implies \|u(t) - \tilde{u}(t)\|_{\ell^2} \overset{t\to\infty}{\to} 0,$$

für alle Lösungen $u, \tilde{u}$ der Differentialgleichung.

(iii) Die Gleichung heißt *exponentiell stabil*, wenn es C, α, δ_0 gibt, sodass

$$\|u(t) - \tilde{u}(t)\|_{\ell^2} \leq Ce^{-\alpha t}$$

für alle Lösungen $u, \tilde{u}$ mit $\|u(0) - \tilde{u}(0)\|_{\ell^2} \leq \delta_0$.

Bemerkung 4.2 (Bedeutung der Stabilitätsbegriffe)**.**

(i) Die Begriffe beschreiben, dass kleine Störungen in den Anfangswerten zu kleinen Störungen in der Lösung führen, und das für alle Zeiten $t \geq 0$.

(ii) Wie das Beispiel $u' = -u$ mit Lösung $u(t) = u_0 e^{-t}$ zeigt, sind die Begriffe für $t \geq 0$ schärfer als die Aussagen, die man aus dem Gronwall-Lemma (Satz 2.3) herleiten kann; dort wird die Zeitrichtung nicht berücksichtigt!

Stabilität bei skalaren linearen DGL

Wir betrachten im Folgenden das einfache lineare skalare Modellproblem

$$u' = \lambda u, \qquad t > 0, \tag{4.1}$$
$$u(0) = u_0, \tag{4.2}$$

mit $\lambda \in \mathbb{C}$ und u_0 gegeben. Wir werden am Ende des Kapitels sehen, dass sich alles Resultate mittels Diagonalisierung bzw. Jordan-Zerlegung sofort auch auf lineare Systeme $u' = Au + b$ und zum Teil sogar auch auf nichtlineare Systeme $u' = f(t, u)$ übertragen lassen.

Zunächst starten wir wieder mit einigen theoretischen Begriffen und Überlegungen.

Satz 4.3 (Stabilität der Lösung)**.**

(i) Die Lösung u von (4.1)–(4.2) ist genau dann beschränkt auf $[0, \infty)$ (und somit stabil im Sinne von Lyapunov), wenn $Re(\lambda) \leq 0$ gilt.

(ii) Gilt $Re(\lambda) < 0$, dann folgt $u(t) \to 0$ für $t \to \infty$ (die Lösung ist asymptotisch stabil). Man erhält sogar exponentielle Konvergenz zur 0.

Beweis. Die Aussage folgt sofort aus der Lösungsdarstellung

$$u(t) = u_0 e^{\lambda t} = u_0 e^{t\,\mathrm{Re}(\lambda)} e^{\mathrm{i}\,t\,\mathrm{Im}\lambda}$$

und der Tatsache, dass $|e^{\mathrm{i}\,r}| = 1$ für alle $r \in \mathbb{R}$ und somit $|u(t)| = |u_0|\, e^{t\,\mathrm{Re}(\lambda)}$. □

Bemerkung 4.4 (Linearität, Stabilität)**.** Die Differenz $w(t) = u_1(t) - u_2(t)$ zweier Lösungen der linearen DGLen $u'(t) = \lambda u(t) + f(t)$ erfüllt die homogene Gleichung $w'(t) = \lambda w(t)$ mit Anfangswert $w(0) = u_1(0) - u_2(0)$. Die Stabilitätsbegriffe aus obigem Satz machen also eine Aussage über die Fortpflanzung von Fehlern in den Anfangswerten bei homogenen und inhomogenen linearen Differentialgleichungen. Bei stabilien Problemen bleiben Anfangsfehler für alle Zeiten beschränkt; bei asymptotisch bzw. exponentiell stabilen Problemen klingen sie entsprechend ab.

Motivation und Ziel. Wir wollen natürlich gewährleisten, dass die numerischen Näherungslösungen zu (4.1)–(4.2) dieselben Stabilitätseigenschaften wie die analytische Lösung besitzen. Wie wir sehen werden, entstehen hieraus neue Anforderungen an die Verfahren. Wir beschränken uns dabei wieder auf Einschrittverfahren.

Achtung. Die nachfolgenden Begriffe machen vorderhand nur in Zusammenhang mit dem Modellproblem (4.1)–(4.2) wirklich Sinn!

Definition 4.5 (Stabilitätsbegriffe)**.**

(i) Falls sich die Näherungen des Einschrittverfahrens $u_{i+1} = u_i + h\phi(t_i, u_i; h)$ für das Modellproblem (4.1)–(4.2) in der Form

$$u_{i+1} = R(z)\, u_i, \qquad z := h\lambda, \tag{4.3}$$

darstellen lassen, dann heißt $R : \mathbb{C} \to \mathbb{C}$ *Stabilitätsfunktion* des Verfahrens.

(ii) Die Menge

$$S := \{z \in \mathbb{C} : |R(z)| \leq 1\} \tag{4.4}$$

heißt in diesem Fall *Stabilitätsgebiet* des Verfahrens.

(iii) Ein Einschrittverfahren heißt

- 0-*stabil*, falls $0 \in S$.
- *A-stabil*, falls $\mathbb{C}^- \subset S$, wobei $\mathbb{C}^- = \{z \in \mathbb{C} : \mathrm{Re}(z) \leq 0\}$.
- *L-stabil*, falls A-stabil und $R(z) \to 0$ für $\mathrm{Re}(z) \to -\infty$.

Bemerkung 4.6 (Stabile Näherungslösungen)**.** Die Größe der Näherungswerte $u_i \approx u(t_i)$ eines Einschrittverfahrens mit Stabilitätsfunktion $R(z)$ lässt sich leicht abschätzen durch

$$|u_{i+1}| = |R(h\lambda)||u_i| = \ldots = |R(h\lambda)|^{i+1}|u_0|.$$

Hieraus erhält man sofort folgende Aussagen:

(i) Wenn λh im Stabilitätsgebiet liegt, dann liefert das Einschrittverfahren beschränkte Näherungslösungen.

(ii) Ein 0-stabiles Einschrittverfahren liefert für das Problem $u' = 0$ (Modellproblem mit $\lambda = 0$) beschränkte Näherungslösungen.

(iii) Ein A-stabiles ESV erzeugt für das stabile Modellproblem $u' = \lambda u$ mit $\mathrm{Re}(\lambda) \leq 0$ eine beschränkte Näherung u_h, und zwar für alle Schrittweiten $h > 0$.

(iv) Bei einem L-stabilen Verfahren klingt die Näherungslösung u_i für $u' = \lambda u$ mit $\mathrm{Re}(\lambda) \ll 0$ auch bei großen Schrittweiten sehr schnell ab.

Wir skizzieren jetzt anhand einiger einfacher Beispiele die Berechnung der Stabilitätsfunktion und des Stabilitätsgebietes.

Beispiel 4.7 (Explizites Eulerverfahren)**.** Für die Lösung von $u' = \lambda u$ (d.h. $f(t, u) = \lambda u$) erhält man die Rekursionsformel

$$u_{i+1} = u_i + hf(t_i, u_i) = u_i + h\lambda u_i = (1 + h\lambda)u_i.$$

Mit $z = h\lambda$ erhält man also die Stabilitätsfunktion $R(z) = 1 + z$. Das Stabilitätsgebiet ist dann

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

Das ist ein Kreis mit Radius 1 und Mittelpunkt -1. Es gilt $0 \in S$: das Verfahren ist also 0-stabil. Aus einer Skizze sieht man jedoch sofort, dass $\mathbb{C}^- \not\subset S$. Das Verfahren ist also nicht A-stabil und somit auch nicht L-stabil.

Beispiel 4.8 (Implizites Eulerverfahren)**.** Für das Modellproblem $u' = \lambda u$ erhält man die Näherungswerte

$$u_{i+1} = u_i + hf(t_{i+1}, u_{i+1}) = u_i + h\lambda u_{i+1}.$$

Wir können dies nach u_{i+1} auflösen und erhalten die Iterationsvorschrift

$$u_{i+1} = \frac{1}{1 - h\lambda} u_i$$

und damit auch

$$|u_{i+1}| = \frac{1}{|1 - h\lambda|} |u_i|.$$

Die Stabilitätsfunktion ist also $R(z) = \frac{1}{1-z}$ und das Stabilitätsgebiet ist

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

Das ist die ganze komplexe Ebene mit Ausnahme eines Kreises mit Radius 1 und Mittelpunkt 1. Insbesondere ist $0 \in S$ und das Verfahren ist daher 0-stabil. Weiters sieht man sofort, dass $\mathbb{C}^- \subset S$; das Verahren ist also auch A-stabil. Aus

$$|R(z)| = \tfrac{1}{|z-1|} \to 0 \qquad \text{mit } |z| \to \infty$$

folgt auch sofort die L-Stabilität des impliziten Eulerverfahrens.

Wir können dies auch für eine etwas allgemeinere Klasse betrachten.

Beispiel (θ-Schema). Für jedes $\theta \in [0, 1]$ ergibt sich das θ-Schema via

$$u_{i+1} = u_i + h\big((1 - \theta)f(t_i, u_i) + \theta f(t_{i+1}, u_{i+1})\big);$$

wir erhalten also für $\theta = 0$ das explizite, für $\theta = 1$ das implizite Euler-Verfahren. Für $\theta = \frac{1}{2}$ erhalten wir die implizite-Trapezregel; diese Methode ist uns bereits bei den Lobatto-III-A Formeln untergekommen.

Beim θ-Schema handelt es sich auch um ein Runge-Kutta-Verfahren und zwar mit folgendem Tableau:

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & (1-\theta) & \theta \\ \hline & (1-\theta) & \theta \end{array}$$

Für das Modellproblem $u' = \lambda u$ erhalten wir also

$$u_{i+1} = R(h\lambda)u_i, \qquad R(z) = \frac{1+z(1-\theta)}{1-z\theta}u_i.$$

Auch für dieses Schema können wir das Stabilitätsgebiet bestimmen:

(i) Für $\theta < \frac{1}{2}$ ist das Stabilitätsgebiet $S = \{z \in \mathbb{C} : |z-1/(2\theta-1)| \leq 1/(2\theta-1)\}$ – wie beim expliziten Eulerverfahren – das Innere eines Kreises mit Mittelpunkt auf der negativen reellen Achse; das Verfahren ist also nicht A-stabil.

(ii) Für $\theta = \frac{1}{2}$ ist $S = \mathbb{C}_0^-$, also ist das Verfahren A-stabil.

(iii) Für $\theta > \frac{1}{2}$ ist das Stabilitätsgebiet $S = \{z \in \mathbb{C} : |z-1/(2\theta-1)| \geq 1/(2\theta-1)\}$ – wie beim impliziten Eulerverfahren – die gesamte Zahlenebene mit Ausnahme des Inneren eines Kreises mit Mittelpunkt auf der positiven reellen Achse; das Verfahren ist demnach ebenfalls A-stabil.

Beispiel 4.9 (Verbessertes Eulerverfahren)**.** Hier ist

$$\begin{aligned} u_{i+1} &= u_i + hf(t_i + \tfrac{h}{2}, u_i + \tfrac{h}{2}f(t_i, u_i)) \\ &= u_i + h[\lambda(u_i + \tfrac{h}{2}\lambda u_i)] = (1 + h\lambda + \tfrac{h^2\lambda^2}{2})u_i. \end{aligned}$$

Die Stabilitätsfunktion ist dann gegeben durch $R(z) = 1+z+\frac{z^2}{2}$. Es gilt $R(0) = 1$ und somit $0 \in S$; das Verfahren ist also 0-stabil. Für $|z| \to \infty$ folgt jedoch $|R(z)| \to \infty$. Das Verfahren ist somit nicht A-stabil und folglich auch nicht L-stabil.

Über die Stabilitätseigenschaften von RK-Verfahren lässt sich Folgendes sagen.

Satz 4.10 (Stabilitätsfunktion von RK-Methoden)**.**

(i) Die Stabilitätsfunktion einer RK-Methode mit Tableau $\begin{array}{c|c} c & A \\ \hline & b^\top \end{array}$ *lautet*

$$R(z) = 1 + z\, b^\top (I - zA)^{-1}\underline{1}, \qquad \textit{wobei } \underline{1} = (1, 1, \ldots)^\top.$$

Die Formel ist formal nur sinnvoll, wenn $1/z$ kein Eigenwert der Matrix A ist.

(ii) Für eine explizite RK-Methode gilt

$$R(z) = 1 + z \sum\nolimits_{j=0}^{s-1} b^\top (zA)^j \underline{1}.$$

Die Stabilitätsfunktion ist also ein Polynom vom Grad $\leq s$.

BEWEIS. (i) Siehe Übung.
(ii) Bei einem expliziten Verfahren ist die Matrix A nilpotent, denn es gilt $A^n = 0$ für alle $n \geq s$. Dies zeigt man relativ einfach durch Induktion. Hieraus folgt

$$(I - zA)^{-1} = \sum_{j=0}^{s-1} (zA)^j,$$

wovon man sich durch Multiplikation mit $(I - zA)$ leicht überzeugt. □

Aus der Darstellung der Stabilitätsfunktion erhält man sofort folgende Aussagen über die Stabilitätseigenschaften von Runge-Kutta Methoden.

Satz 4.11 (Stabilitätseigenschaften von RK-Methoden)**.**

(i) Runge-Kutta Methoden sind stets 0-stabil.

(ii) Explizite Runge-Kutta Methoden sind nie A-stabil und können folglich auch nicht L-stabil sein.

Im Gegensatz zu expliziten Runge-Kutta Verfahren haben implizite Verfahren oftmals deutlich bessere Stabilitätseigenschaften, wie der folgende Satz belegt.

Satz 4.12 (A-stabile RK-Methoden)**.**

(i) Gauß, Lobatto-III-A und Lobatto-III-B Verfahren sind A-stabil, nicht L-stabil.

(ii) Radau-II-A und Lobatto-III-C Verfahren sind A-stabil und L-stabil.

BEWEIS. Siehe *Grigorieff: Numerik gewöhnlicher Differentialgleichungen.* □

Bemerkung 4.13 (Konsequenzen fehlender A- bzw. L-Stabilität)**.**

(i) Für $u' = \lambda u$ mit $\mathrm{Re}(\lambda) \ll 0$ ist bei moderaten Schrittweiten $|z| = |h\lambda| > 1$. Explizite RK-Verfahren erzeugen dann im Allgemeinen instabile (d.h. betragsmäßig wachsende) Lösungen; erst bei hinreichend kleinen Schrittweiten (z.B. $h \leq |\lambda|$) erhält man stabile (beschränkte) Approximationen und kann dann auch das entsprechende Konvergenzverhalten sehen.

(ii) Im Gegensatz dazu liefern (A- bzw. L-stabile) implizite RK-Verfahren für alle Schrittweiten, also bereits für moderate Schrittweiten h stabile Näherungen. Sie sind stabil für alle Schrittweiten, engl: *unconditionally stable*.

Stabilität bei linearen DGL-Systemen

Im Folgenden skizzieren wir die Erweiterung der obigen Stabilitätsbegriffe für Einschrittverfahren auf Systeme von Differentialgleichungen im $\mathbb{R}^N$:

$$u' = Au + b, \qquad t > 0, \tag{4.5}$$
$$u(0) = u_0, \tag{4.6}$$

mit einer Matrix $A \in \mathbb{R}^{N\times N}$ und einem Vektor $b \in \mathbb{R}^N$.

Satz 4.14 (Stabilität linearer Systeme)**.**

(i) Für ein lineares System $u' = Au + b$ ist jede Lösung genau dann exponentiell stabil, wenn gilt

$$Re(\lambda) < 0, \quad \textit{für alle Eigenwerte } \lambda \textit{ von } A.$$

(ii) Falls $Re(\lambda) \leq 0$ und die Vielfachheit des Eigenwerts $\lambda = 0$ eins ist, dann ist jede Lösung immer noch stabil, aber nicht exponentiell stabil.

Da die Begriffe bei linearen Systemen unabhängig von der Lösung sind, nennt man gleich das ganze System stabil bzw. exponentiell stabil.

Beweis. Die Differenz zweier Lösungen $w := u - \tilde{u}$ erfüllt die homogene Gleichung $w' = Aw$. Seien nun $\lambda_1, \ldots, \lambda_k$ die Eigenwerte von A mit (algebraischen) Vielfachheiten $m_1, \ldots, m_j$. Dann besitzt die allgemeine Lösung zur homogenen linearen Differentialgleichung $w' = Aw$ die Form (siehe VL *Gewöhnliche Differentialgleichungen und Dynamische Systeme*)

$$w(t) = c_1(t)e^{\lambda_1 t} + \ldots + c_k(t)e^{\lambda_k t},$$

wobei $c_1, \ldots, c_k$ vektorwertige Polynome vom Grad $\leq m_k - 1$ sind. Für die Eigenwerte mit $\mathrm{Re}(\lambda) < 0$ ist $c(t)e^{\lambda t}$ beschränkt für alle $t \geq 0$, da die Exponentialfunktion schneller fällt als die Polynomfunktion c wächst. Falls $\mathrm{Re}(\lambda) = 0$ ist, dann besitzt das entsprechende Polynomgrad 0, und die Lösungskomponente bleibt zumindest beschränkt. □

Bemerkung 4.15 (Lösungsverhalten bei linearen Systemen)**.** Aus der Lösungsdarstellung für linear Systeme folgt unmittelbar:

(i) Ist $u' = Au$ stabil, d.h. $\mathrm{Re}(\lambda) \leq 0$ und $\mathrm{Re}(\lambda) = 0 \implies$ Vielfachheit von λ ist eins, dann ist die Lösung u beschränkt.

(ii) Für stabile lineare Probleme erzeugt ein A-stabiles Einschrittverfahren beschränkte Näherungen.

Bemerkung 4.16. Wir werden auf diese Bedingungen bei der Diskretisierung von Anfangs-Randwertproblemen zurückkommen.

Stabilität bei nichtlinearen DGL-Systemen

Zuletzt wollen wir noch kurz andeuten, wie sich Stabilitätsbegriffe zum Teil auch auf nichtlineare Gleichungen und Systeme erweitern lassen.

Definition 4.17. (i) f genügt einer *einseitigen Lipschitzbedingung*, falls

$$(f(t,u) - f(t,\tilde{u}), u - \tilde{u})_{\ell^2} \leq \lambda_0 \|u - \tilde{u}\|_{\ell^2}^2.$$

(ii) Gilt $\lambda_0 \leq 0$, so heißt die Differentialgleichung $u' = f(t,u)$ *dissipativ*.

Satz 4.18. *Für Lösungen u und $\tilde{u}$ eines dissipativen Systems gilt*

$$\|u(t) - \tilde{u}(t)\|_{\ell^2} \leq \|u(s) - \tilde{u}(s)\|_{\ell^2}, \qquad s \leq t.$$

Insbesondere sind dissipative Systeme stabil im Sinne von Lyapunov.

Beweis. Die Funktion f genüge einer einseitigen Lipschitzbedingung. Dann gilt

$$\begin{aligned} \tfrac{1}{2}\tfrac{\mathrm{d}}{\mathrm{d}t}\|u(t) - \tilde{u}(t)\|_{\ell^2}^2 &= (u'(t) - \tilde{u}'(t), u(t) - \tilde{u}(t))_{\ell^2} \\ &= (f(t,u(t)) - f(t,\tilde{u}(t)), u(t) - \tilde{u}(t))_{\ell^2} \leq \lambda_0 \|u(t) - \tilde{u}(t)\|_{\ell^2}^2. \end{aligned}$$

Durch Aufintegrieren nach der Zeit erhält man

$$\|u(t) - \tilde{u}(t)\|_{\ell^2}^2 \leq e^{2\lambda_0(t-s)} \|u(s) - \tilde{u}(s)\|_{\ell^2}^2, \qquad s \leq t.$$

Für $\lambda_0 \leq 0$ folgt bereits die Aussage des Satzes. □

Bemerkung 4.19. Falls f einer einseitigen Lipschitzbedingung genügt, ist die Fehlerverstärkung mit der Zeit nur wie $e^{\lambda_0 t}$ und nicht wie im Gronwall-Lemma mit e^{Lt}, wobei L die Lipschitzkonstante ist. Ist das System sogar dissipativ, dann werden Fehler eher gedämpft als verstärkt.

Es ist natürlich wünschenswert, dass sich die Eigenschaften wieder auf die numerische Lösung übertragen. Entsprechend führen wir folgende Begriffe ein.

Definition 4.20 (B-Stabilität)**.**

(i) Erfüllen die Näherungen eines Einschrittverfahrens $\|u_i - \tilde{u}_i\|_{\ell^2} \leq \|u_j - \tilde{u}_j\|_{\ell^2}$ für $j \leq i$, dann heißt das Verfahren kontraktiv (für das gegebene Problem).

(ii) Ein Einschrittverfahren, dass für dissipative Systeme stets kontraktive Näherungslösungen liefert heißt B-stabil.

Beispiel 4.21. Das implizite Eulerverfahren ist B-stabil. Es gilt nämlich

$$\begin{aligned}\|u_{i+1} - \tilde{u}_{i+1}\|_{\ell^2}^2 &= (u_i + f(t_{i+1}, u_{i+1}) - \tilde{u}_i - f(t_{i+1}, \tilde{u}_{i+1}), u_{i+1} - \tilde{u}_{i+1})_{\ell^2} \\ &= (u_i - \tilde{u}_i, u_{i+1} - \tilde{u}_{i+1})_{\ell^2} + (f(t_{i+1}, u_{i+1}) - f(t_{i+1}, \tilde{u}_{i+1}), u_{i+1} - \tilde{u}_{i+1})_{\ell^2} \\ &\leq \|u_i - \tilde{u}_i\|_{\ell^2} \|u_{i+1} - \tilde{u}_{i+1}\|_{\ell^2}.\end{aligned}$$

Nach Division durch $\|u_{i+1} - \tilde{u}_{i+1}\|_{\ell^2}$ folgt dann sofort $\|u_{i+1} - \tilde{u}_{i+1}\|_{\ell^2} \leq \|u_i - \tilde{u}_i\|_{\ell^2}$ und per Induktion dann die B-Stabilität.

Bemerkung 4.22. Ein lineares System $u' = Au + b$ mit konstanter Koeffizientenmatrix A ist genau dann dissipativ, wenn es stabil im Sinne von Lyapunov ist. In diesem Fall verhält sich jedes A-stabile Verfahren bereits kontraktiv. Umgekehrt sieht man daraus sofort, dass B-Stabilität stets die A-Stabilität impliziert. Für weitere Überlegungen verweisen wir wieder auf die einschlägige Literatur.

5. Differentialgleichungen höherer Ordnung

Bisher hatten wir uns nur numerische Methoden für Anfangswertprobleme für Differentialgleichungen erster Ordnung angesehen. Wir werden uns nun mit ihrer Erweiterung auf Differentialgleichungen (und Systeme) zweiter Ordnung näher beschäftigen, viele der Ideen lassen sich auch auf höhere Ordnungen verallgemeinern. Wir beschränken unsere Diskussionen auf Differentialgleichungen der Art

$$u''(t) = f(t, u(t)),\ t \in (0, T]. \tag{5.1}$$

Auch hier nehmen wir wieder an, dass $f : \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}^N$ stetig und in seiner zweiten Komponente Lipschitz-stetig ist:

$$\|f(t, x) - f(t, y)\|_{\ell^2} \leq L\|x - y\|_{\ell^2}.$$

Aus der VL *Gewöhnliche Differentialgleichungen und Dynamische Systeme* wissen wir noch, dass zwei *Anfangsbedingungen* erforderlich sind:

$$u(0) = u_0, \qquad u'(0) = v_0.$$

Wir können dieses *Anfangswertproblem zweiter Ordnung* in der Form eines Systems erster Ordnung schreiben:

$$\underbrace{\begin{pmatrix} u(t) \\ v(t) \end{pmatrix}'}_{w'(t)} = \underbrace{\begin{pmatrix} v(t) \\ f(t, u(t)) \end{pmatrix}}_{\tilde{f}(t, w(t))}, \qquad \underbrace{\begin{pmatrix} u(0) \\ v(0) \end{pmatrix}}_{w(0)} = \begin{pmatrix} u_0 \\ v_0 \end{pmatrix}. \tag{5.2}$$

Wir sehen sofort, dass sich die Stetigkeit und die Lipschitz-Stetigigekit in der zweiten Komponente von f auf $\tilde{f}$ übertragen:

$$\begin{aligned} \|\tilde{f}(t, w(t)) - \tilde{f}(t, \tilde{w}(t))\|_{\ell^2}^2 &= \|v(t) - \tilde{v}(t)\|_{\ell^2}^2 + \|f(t, u(t)) - f(t, \tilde{u}(t))\|_{\ell^2}^2 \\ &\leq \|v(t) - \tilde{v}(t)\|_{\ell^2}^2 + L^2\|u(t) - \tilde{u}(t)\|_{\ell^2}^2 \leq \max\{1, L^2\}\|w(t) - \tilde{w}(t)\|_{\ell^2}^2 \end{aligned}$$

gilt für alle $w(t) = (u(t), v(t))$, $\tilde{w}(t) = (\tilde{u}(t), \tilde{v}(t))$. Das bedeutet, wir können die Methoden und die Konvergenztheorie aus den letzten Kapiteln wieder anwenden.

Beispiel 5.1 (Explizites Eulerverfahren)**.** Wählen wir das explizite Eulerverfahren zur Lösung des Systems (5.2), erhalten wir

$$\begin{aligned} u_{i+1} &= u_i + hv_i \\ v_{i+1} &= v_i + hf(t_i, u_i). \end{aligned}$$

Durch Einsetzen erhalten wir

$$\begin{aligned} u_{i+1} &= u_i + hv_{i-1} + h^2 f(t_{i-1}, u_{i-1}) = u_i + h\tfrac{u_i - u_{i-1}}{h} + h^2 f(t_{i-1}, u_{i-1}) \\ &= 2u_i - u_{i-1} + h^2 f(t_{i-1}, u_{i-1}), \qquad i \geq 1. \end{aligned}$$

mit Anfangsbedingungen

$$u_0 = u(0), \qquad u_1 = u_0 + hu'(0).$$

Wir stellen fest, dass sich das Verfahren auch dadurch motivieren lässt, dass wir die zweite Ableitung in (5.1) durch einen Vorwärtsdifferenzenquotienten annäheren:

$$\frac{u_{i+2} - 2u_{i+1} + u_i}{h^2} \approx u''(t_i) = f(t_i, u(t_i)).$$

Wir können nun wieder die Konvergenztheorie aus Kapitel 2 anwenden und erhalten diskrete Stabilität (Definition 2.13) und Konsistenz (mit Ordnung 1) und damit auch eine entsprechende Konvergenz $\|u - u_h\|_{h,\infty} = \mathcal{O}(h)$.

Beispiel 5.2 (Implizites Eulerverfahren)**.** Wählen wir das implizite Eulerverfahren zur Lösung des Systems (5.2), erhalten wir

$$\begin{aligned} u_{i+1} &= u_i + hv_{i+1} \\ v_{i+1} &= v_i + hf(t_{i+1}, u_{i+1}). \end{aligned}$$

Durch Einsetzen erhalten wir

$$\begin{aligned} u_{i+1} &= u_i + hv_i + h^2 f(t_{i+1}, u_{i+1}) = u_i + h\tfrac{u_i - u_{i-1}}{h} + h^2 f(t_{i+1}, u_{i+1}) \\ &= 2u_i - u_{i-1} + h^2 f(t_{i+1}, u_{i+1}), \qquad i \geq 1. \end{aligned}$$

mit Anfangsbedingungen

$$u_0 = u(0), \qquad u_1 = u(0) + hu'(0) + h^2 f(u_1, t_1).$$

Wir stellen fest, dass sich das Verfahren auch dadurch motivieren lässt, dass wir die zweite Ableitung in (5.1) durch einen Rückwärtsdifferenzenquotienten annäheren:

$$\frac{u_i - 2u_{i-1} + u_{i-2}}{h^2} \approx u''(t_i) = f(t_i, u(t_i)).$$

Wir können nun wieder die Konvergenztheorie aus Kapitel 2 anwenden und erhalten diskrete Stabilität (Definition 2.13) und Konsistenz (mit Ordnung 1) und damit auch eine entsprechende Konvergenz $\|u - u_h\|_{h,\infty} = \mathcal{O}(h)$.

Eine bessere Konsistenzordnung können wir – wie bei der Herleitung des verbesserten Eulerverfahrens – durch die Annäherung der Integrale mit der Mittelpunktregel erzielen.

Beispiel 5.3. Zur Motivation wählen wir entsprechend geschachtelte Gitter für u und $\hat{v} \approx u'$:

$$\begin{aligned}\mathcal{T} &= (t_0, t_1, \dots, t_n) = (0, h, \dots, T),\\ \hat{\mathcal{T}} &= (t_{-1/2}, t_{1/2}, t_{3/2}, \dots, t_{n-1/2}) = (-h/2, h/2, 3h/2, \dots, T - h/2).\end{aligned}$$

Mit dieser Wahl können wir ohne weiteres die Mittelpunktregel anwenden:

$$u_{i+1} = u_i + h\hat{v}_{i+1/2}, \qquad \hat{v}_{i+1/2} = \hat{v}_{i-1/2} + hf(t_i, u_i), \qquad i \geq 0.$$

Durch Einsetzen erhalten wir

$$u_{i+1} = 2u_i - u_{i-1} + h^2 f(t_i, u_i), \qquad i \geq 1. \tag{5.3}$$

Wegen der Berechnung von u_1, siehe unten.

Wir stellen fest, dass sich das Verfahren auch dadurch motivieren lässt, dass wir die zweite Ableitung in (5.1) durch einen zentralen Differenzenquotienten annähern:

$$\frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} \approx u''(t_i) = f(t_i, u(t_i)).$$

Wir können auch dieses Verfahren mit $v_i = \widehat{v}_{i-1/2}$ in der uns bekannten Art schreiben:

$$\underbrace{\begin{pmatrix} u_{i+1} \\ v_{i+1} \end{pmatrix}}_{w_{i+1}} = \underbrace{\begin{pmatrix} u_i \\ v_i \end{pmatrix}}_{w_i} + h\phi(t, w_i; h)$$

mit Verfahrensfunktion

$$\phi(t, w_i; h) = \begin{pmatrix} v_{i+1} \\ f(t_i, u_i) \end{pmatrix} = \begin{pmatrix} v_i + hf(t_i, u_i) \\ f(t_i, u_i) \end{pmatrix},$$

die die Lip.-Bedingung erfüllt. Mit Satz 2.15 erhalten wir die diskrete Stabilität:

$$\|w_h - \tilde{w}_h\|_{\infty,h} \leq C(\|w_0 - \tilde{w}_0\|_{\ell^2} + \|\theta_h\|_{\infty,h}),$$

wobei $\tilde{w}_h$ die Gitterfunktion ist, die wir durch Interpolation der exakten Funktion erhalten. Es gilt also $\tilde{w}_h(t_i) = \tilde{w}_i = (\tilde{u}_i, \tilde{v}_i) = (u(t_i), v(t_i - h/2))^\top$. Dabei ist (vgl.

Definition 2.13)

$$\begin{aligned}
h\theta_i &= \tilde{w}_{i+1} - [\tilde{w}_i + h\phi(t, \tilde{w}_i; h)] \\
&= \begin{pmatrix} u(t_i + h) \\ v(t_i + h/2) \end{pmatrix} - \left[\begin{pmatrix} u(t_i) \\ v(t_i - h/2) \end{pmatrix} + h \begin{pmatrix} v(t_i + h/2) \\ f(t_i, u(t_i)) \end{pmatrix} \right] \\
&= \begin{pmatrix} u(t_i + h) - u(t_i) - hu'(t_i + h/2) \\ u'(t_i + h/2) - u'(t_i - h/2) - hu''(t_i) \end{pmatrix} \\
&= \begin{pmatrix} [u(t_i) + hu'(t_i) + \frac{h^2}{2}u''(t_i) + \mathcal{O}(h^3)] - u(t_i) - h[u'(t_i) + \frac{h}{2}u''(t_i) + \mathcal{O}(h^2)] \\ [u'(t_i) + \frac{h}{2}u''(t_i) + \frac{h^2}{8}u'''(t_i) + \mathcal{O}(h^3)] - [u'(t_i) - \frac{h}{2}u''(t_i) + \frac{h^2}{8}u'''(t_i) + \mathcal{O}(h^3)] - hu''(t_i) \end{pmatrix} \\
&= \mathcal{O}(h^3),
\end{aligned}$$

woraus wir $\|\theta_h\|_{\infty,h} = \mathcal{O}(h^2)$, also eine Konsistenzordnung von 2 für die Lösung der Differentialgleichung erhalten.

Schlussendlich müssen wir uns noch mit der Wahl der Anfangsbedingungen beschäftigen. Wenn wir uns das Schema (5.3) ansehen, dann benötigen wir $u_0 = u(t_0)$ und u_1. Wählen wir entsprechend der Taylor-Formel $u_1 = u(0) + hu'(0)$, erhalten wir für diesen Anfangswert nur eine Konsistenzordnung von 1 (siehe Übung), was das gesamte Verfahren auf eine Konvergenzordnung von 1 herabstuft!

Stattdessen müssen wir eine Approximation zweiter Ordnung wählen, die wir unter Verwendung der Differentialgleichung auflösen können:

$$u_0 = u(0), \quad u_1 = u(0) + hu'(0) + \tfrac{h^2}{2}u''(0) = u(0) + hu'(0) + \tfrac{h^2}{2}f(0, u(0)). \quad (5.4)$$

Diese Wahl entspricht $v_0 = \hat{v}_{-1/2} = u'(0) - \frac{h}{2}f(0, u(0))$. Für das Verfahren (5.4), (5.3) können wir nun Konsistenz- und Konvergenzordnung 2 zeigen; siehe Übung.

Teil II.

Randwertprobleme

6. Vorbemerkungen

Im Folgenden werden wir uns mit Randwertproblem zweiter Ordnung beschäftigen:

$$-u''(x) = f(x, u(x), u'(x)), \qquad x \in (0,1), \tag{6.1}$$
$$u(0) = 0, \tag{6.2}$$
$$u(1) = 0. \tag{6.3}$$

Wir stellen sofort fest, dass sich dieses Randwertproblem vom Problem aus Kapitel 5 durch die Wahl der zusätzlichen Bedingungen unterscheiden: In Kapitel 5 wurden zwei Anfangsbedingungen gefordert, nun fordern wie zwei (homogene) *Dirichlet*-Randbedingungen, und zwar je eine für jede Seite. (Die Behandlung anderer Randbedingungen werden wir noch besprechen.)

Zur Einstimmung wollen wir versuchen, das Randwertproblem mit den Methoden für die Lösung von Anfangswertproblemen zu lösen.

Das Schießverfahren

Wenn das Randwertproblem (6.1) – (6.3) (eventuell eindeutig) lösbar ist, dann erfüllt eine Lösung u auch folgendes Anfangswertproblem:

$$-v_\alpha''(x) = f(x, v_\alpha(x), v_\alpha'(x)), \qquad x \in (0,1), \tag{6.4}$$
$$v_\alpha(0) = 0, \tag{6.5}$$
$$v_\alpha'(0) = \alpha \tag{6.6}$$

mit $\alpha = u'(0)$. Dieses Anfangswertproblem können wir mit den in Teil I eingeführten Methoden lösen; unter milden Regularitätsannahmen an f ist es möglich, Existenz und Eindeutigkeit einer Lösung dieses Anfangswertproblems zu zeigen. Leider gibt es mit dieser Idee ein Problem: Wir kennen den Wert für α nicht.

Ein intuitiver Ansatz für die Lösung des Randwertproblems wäre folgendes Vorgehen, das teilweise unter dem Begriff „Schießverfahren" bekannt ist:

1. Wähle einen geeigneten Startwert für α.
2. Löse das Anfangswertproblem.
3. Wenn $|v_\alpha(1) - 0|$ klein genug ist, dann haben wir eine Lösung gefunden!

4. Andernfalls: Bestimme eine bessere Näherung für α und springe zu Punkt 2.

Wir können wir bessere Näherungen für α finden?

Sei $F : \mathbb{R} \to \mathbb{R}$ definiert über $F(\alpha) = v_\alpha(1)$; diese Funktion weist also jedem Startwert α den Funktionswert des Anfangswertproblems am rechten Rand (also für $x = 1$) zu. Die Lösungen des Randwertproblems erhalten wir über Nullstellen von F, also

$$F(\alpha) = 0.$$

Wir können nun alle Verfahren zur Nullstellensuche anwenden, die uns bekannt sind. Dabei gilt es natürlich zu bedenken, dass die Funktionsauswertung von F nicht ganz einfach ist, da jede Funktionsauswertung der Lösung eines Anfangswertproblems entspricht. Daher sind wir an schnellen Lösungsverfahren, etwa dem Newton-Verfahren

$$\alpha_{k+1} = \alpha_k - \frac{F(\alpha_k)}{F'(\alpha_k)}$$

interessiert.

Die für das Newton-Verfahren erforderliche Ableitung lässt sich aufgrund folgendes Lemmas bestimmen.

Lemma 6.1. *Für jedes α ist $F'(\alpha) = w_\alpha(1)$, wobei w_α die Lösung von folgendem linearen Anfangswertproblem ist:*

$$\begin{aligned} -w''_\alpha(x) &= \partial_u f(x, v_\alpha(x), v'_\alpha(x)) w_\alpha(x) + \partial_{u'} f(x, v_\alpha(x), v'_\alpha(x)) w'_\alpha(x), \qquad x \in (0,1), \\ w_\alpha(0) &= 0, \\ w'_\alpha(0) &= 1. \end{aligned}$$

Beweis. Es gilt $F(\alpha) = v_\alpha(1)$ und aufgrund der Definition der Ableitung gilt

$$F'(\alpha) = \lim_{\delta \to 0} \frac{F(\alpha+\delta) - F(\alpha)}{\delta} = \lim_{\delta \to 0} \frac{v_{\alpha+\delta}(1) - v_\alpha(1)}{\delta},$$

wobei $v_{\alpha+\delta}$ und v_α die jeweiligen Lösungen des Anfangswertproblems (6.4) – (6.6) sind. Für $W_{\alpha,\delta} := \frac{1}{\delta}(v_{\alpha+\delta} - v_\alpha)$ gilt entsprechend:

$$\begin{aligned} -W''_{\alpha,\delta}(x) &= \tfrac{1}{\delta}(f(x, v_{\alpha+\delta}(x), v'_{\alpha+\delta}(x)) - f(x, v_\alpha(x), v'_\alpha(x))), \qquad x \in (0,1), \\ W_{\alpha,\delta}(0) &= 0, \\ W'_{\alpha,\delta}(0) &= 1 \end{aligned}$$

und somit $F'(\alpha) = \lim_{\delta \to 0} W_{\alpha,\delta}(1)$. Bei entsprechender Glattheit der involvierten Funktionen erhalten wir mithilfe der Kettenregel das gewünschte Resultat; für Details siehe Martin Hanke-Borgeois: *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens*. 2002. □

Obwohl dieses Verfahren für Randwertprobleme (in einer Unbekannten!) grundsätzlich anwendbar ist, zeigt die numerische Praxis, dass man auch bei an sich einfachen Randwertproblemen mit numerischer Instabilität zu kämpfen hat, vgl. Übung.

Wir werden uns nun mit einem anderen Ansatz beschäftigen, der eine direkte Lösung von *linearen* Randwertproblemen erlaubt. (Wir werden uns dabei auf Randwertprobleme zweiter Ordnung beschränken.) Dies erlaubt uns, folgende Themen zu thematisieren, die bisher offen blieben:

- Existenz und Eindeutigkeit einer Lösung zum kontinuierlichen Randwertproblem,
- Existenz und Eindeutigkeit einer Lösung zum diskretisierten Randwertproblem,
- Diskretisierungsfehlerabschätzungen und Konvergenz,
- Verallgemeinerung auf mehrere Ortsdimensionen.

Schließlich soll noch erwähnt werden, dass nichtlineare Randwertprobleme oftmals nicht nach dem Vorbild des Schießverfahrens, sondern durch Linearisierung der Differentialgleichung selbst gelöst werden. Die im Folgenden behandelten Methoden zur Lösung linearer Randwertprobleme können oft auch (eventuell mit Modifikationen) zur Lösung der linearen Randwertprobleme verwendet werden, die man bei der Linearisierung erhält.

7. Lineare Dirichletprobleme

Wie angesprochen, beschränken wir uns auf lineare Randwertprobleme zweiter Ordnung. Wir suchen eine (hinreichend glatte) Lösung des Randwertproblems

$$-u''(x) + b(x)u'(x) + c(x)u(x) = f(x), \qquad x \in (0,1), \tag{7.1}$$
$$u(0) = 0 \tag{7.2}$$
$$u(1) = 0. \tag{7.3}$$

Wir erinnern uns: Die Randbedingungen (7.2) und (7.3) sind Dirichlet-Randbedingungen, daher nennen wir das Randwertproblem auch *Dirichletproblem*. Wir werden auf den Fall anderer Randbedingungen später wieder zu sprechen kommen.

Für $c < 0$ lassen sich leicht Fälle (etwa $b = 0, c = -\pi^2$) konstruieren, wo das Problem nicht eindeutig lösbar ist. Daher werden wir stets

$$c(x) \geq 0 \quad \forall x \in [0,1] \tag{7.4}$$

annehmen.

Das Modellproblem (7.1) – (7.3) lässt sich in eine lineare Operatorgleichung umschreiben. Sei $C_0^2(0,1) := \{v \in C^2(0,1) \cap C[0,1] : v(0) = v(1) = 0\}$. Der Differentialoperator $L : C_0^2(0,1) \to C(0,1)$ ist nun definiert über

$$L\,u := -u'' + bu' + cu.$$

Das Modellproblem lautet nun: Finde $u \in C_0^2(0,1)$ mit

$$L\,u = f.$$

Zuerst wollen wir uns nun mit dem Fall der Existenz und Eindeutigkeit einer Lösung befassen. Für den Fall $b, c = 0$ ist diese Frage leicht zu beantworten.

Konstruktion einer Lösung für den Fall $b, c = 0$

Im eindimensionalen Fall ist es noch möglich, die Lösung analytisch zu bestimmen. Seien nun $b = 0$ und $c = 0$. Durch zweimaliges Integrieren der Differential-

gleichung (7.1) und durch Tauschen der Integrale erhalten wir

$$\begin{aligned} u(x) &= -\int_0^x \int_0^t f(s)\mathrm{d}s\mathrm{d}t + c_0 + c_1 x = -\int_{0\le s\le t\le x} f(s)\mathrm{d}(s,t) + c_0 + c_1 x \\ &= \int_0^x (s-x) f(s)\mathrm{d}s + c_0 + c_1 x = \int_0^1 \min\{0, y-x\} f(y)\mathrm{d}y + c_0 + c_1 x. \end{aligned} \tag{7.5}$$

Durch Einsetzen der Randbedingungen erhalten wir

$$0 = u(0) = c_0$$

und

$$0 = u(1) = \int_0^1 (1-y) f(y)\mathrm{d}y + c_0 + c_1, \tag{7.6}$$

woraus wir die Konstanten c_0 und c_1 bestimmen können. Durch Einsetzen erhalten wir

$$u(x) = \int_0^1 \min\{0, y-x\} f(y)\mathrm{d}y - x \int_0^1 \min\{0, y-1\} f(y)\mathrm{d}y.$$

Es lässt sich leicht nachrechnen, dass sich die Lösung u in der Form

$$u(x) = \int_0^1 G(x,y) f(y)\mathrm{d}y \tag{7.7}$$

schreiben lässt, wobei

$$G(x,y) := \begin{cases} y(1-x) & \text{für } y \le x \\ x(1-y) & \text{für } y > x \end{cases} \tag{7.8}$$

die zugehörige Green'sche Funktion ist. Wir stellen fest, dass G eine symmetrische ($G(y,x) = G(x,y)$), nicht-negative ($G(x,y) \ge 0$) und stetige Funktion ist.

Durch die Konstruktion der Lösung erhalten wir sofort die Existenz einer Lösung, außerdem dass die Lösung zweimal stetig differenzierbar sein muss.

Satz 7.1 (Existenz und Regularität der Lösung)**.** *Für jedes $f \in C(0,1)$, $b = 0$ und $c = 0$, gibt es eine Lösung $u \in C^2(0,1) \cap C[0,1]$.*

Beweis. Existenz folgt aus der Konstruktion der Lösung. Die Regularitätsaussage folgt unmittelbar aus $-u''(x) = f(x)$. □

Bemerkung 7.2 (Maximumsprinzip und Monotonieeigenschaft)**.** Aus der Konstruktion der Lösung erhalten wir, dass aus $Lu = f \le 0$ die Abschätzung $u \le 0$

folgt, also das Maximum der (in diesem Abschnitt konstruierten) Lösungen am Rand angenommen wird. Eine solche Eigenschaft ist auch als Maximumsprinzip bekannt. Aus ihr folgt wegen Linearität auch die folgende Monotonieeigenschaft:

$$f \leq \tilde{f} \quad \Rightarrow \quad u \leq \tilde{u},$$

wobei $u, \tilde{u}$ die (in diesem Abschnitt konstruierten) Lösungen zu den Dirichletproblemen $-u'' = f$ und $-\tilde{u}'' = \tilde{f}$ sind.

Wir können außerdem zeigen, dass die Lösung stetig von den Daten abhängt. Dazu erinnern wir uns an die Supremumsnorm

$$\|u\|_\infty := \sup_{x \in [0,1]} |u(x)|.$$

Satz 7.3 (Stabilität)**.** *Sei $f \in C(0,1)$ und sei $u \in C^2(0,1) \cap C[0,1]$ die (in diesem Abschnitt konstruierte) Lösung des zugehörigen Randwertproblems* (7.1) – (7.3) *mit $b = 0$ und $c = 0$. Dann gilt*

$$\|u\|_\infty \leq \tfrac{1}{8} \|f\|_\infty.$$

Beweis. Sei $x \in [0,1]$. Da G nicht-negativ ist, erhalten wir

$$|u(x)| \leq \int_0^1 G(x,y)\ |f(y)| \mathrm{d}y \leq \int_0^1 G(x,y) \mathrm{d}y\ \|f\|_\infty.$$

Es gilt

$$\int_0^1 G(x,y) \mathrm{d}y = \int_0^x y(1-x) \mathrm{d}y + \int_x^1 x(1-y) \mathrm{d}y = \tfrac{1}{2} x(1-x) \leq \tfrac{1}{8},$$

woraus die Aussage unmittelbar folgt. □

Maximumsprinzip und Eindeutigkeit einer Lösung

Um die Eindeutigkeit der Lösung zu etablieren, dürfen wir nicht von der im letzten Abschnitt konstruierten Lösung ausgehen, wir müssen vielmehr Resultate angeben, die für alle in Frage kommenden Lösungen gelten.

Wir können ganz allgemein für Lösungen der Differentialgleichung ein Maximumsprinzip zeigen; wir nehmen wieder an, dass die Daten (f, b und $c \geq 0$) stetig sind und die Lösung u hinreichend glatt ist.

Satz 7.4 (Maximumsprinzip)**.** *Sei $u \in C^2(0,1)$ mit $Lu \leq 0$. Dann gilt*

$$u(x) \leq \max\{0, u(0), u(1)\}, \qquad \forall x \in [0,1].$$

Wenn $c = 0$, dann gilt

$$u(x) \leq \max\{u(0), u(1)\}, \qquad \forall x \in [0,1].$$

Beweis. Wir machen den Beweis in mehreren Schritten.

(i) Wir betrachten zuerst den Fall, dass $Lu < 0$. Wir machen einen Widerspruchsbeweis, nehmen also an, dass es im Inneren einen Punkt x_0 gibt, sodass $u(x_0) > \max\{0, u(0), u(1)\}$. Somit muss u im Inneren ein Maximum annahmen, also gibt es ein x^* mit

$$u'(x^*) = 0, \quad u''(x^*) \leq 0, \quad u(x^*) > 0.$$

Es gilt dort

$$\underbrace{-u''(x^*)}_{\geq 0} + \underbrace{b(x^*)u'(x^*)}_{=0} + \underbrace{c(x^*)}_{\geq 0}\underbrace{u(x^*)}_{>0} = Lu(x^*) < 0,$$

was offensichtlich ein Widerspruch ist.

(ii) Nun betrachten wir den Fall, dass $Lu \leq 0$. Wir beschränken uns der Einfachheit halber auf $b = 0$. Wir nehmen wieder an, dass es im Inneren einen Punkt x_0 gibt, sodass $u(x_0) > \max\{0, u(0), u(1)\}$. Wir definieren

$$v(x) := u(x) + \epsilon\phi(x)$$

mit

$$\epsilon := \tfrac{1}{2}(u(x_0) - \max\{0, u(0), u(1)\}) \quad \text{und} \quad \phi(x) := x(x-1).$$

Es gilt nun

$$v(x_0) > \max\{0, u(0), u(1)\} = \max\{0, v(0), v(1)\}. \tag{7.9}$$

Auf der anderen Seite gilt

$$Lv = \underbrace{Lu}_{\leq 0} + \underbrace{(-2\epsilon)}_{<0} + \underbrace{c\epsilon x(x-1)}_{\leq 0}.$$

Da nun $Lv < 0$ gilt, muss nach dem bereits behandelten Fall (i) ein Maximumsprinzip gelten. Dies steht im Widerspruch zu (7.9). Den Fall $b \neq 0$ kann man mit $\phi(x) = \int_0^x t \exp(\int_0^t b(s)\mathrm{d}s)\mathrm{d}t - \sup_z \int_0^z t \exp(\int_0^t b(s)\mathrm{d}s)\mathrm{d}t$ beweisen, vgl. Übung.

(iii) Wenn $c = 0$ gilt, dann können wir auf die Überlegungen zum Vorzeichen von u verzichten. □

Da wir uns in unserem Modellproblem auf beiden Rändern Dirichlet-Randbedingungen (7.2) und (7.3) gewählt haben, haben wir sofort

$$Lu \leq 0 \quad \Rightarrow \quad u \leq 0$$

sowie (da wir u durch $-u$ substituieren können) $Lu \geq 0 \Rightarrow u \geq 0$ und damit

$$Lu = 0 \quad \Rightarrow \quad u = 0. \tag{7.10}$$

Daraus folgt sofort

Satz 7.5 (Eindeutigkeit)**.** *Das Modellproblem* (7.1) – (7.3) *kann nur eine Lösung haben.*

Beweis. Wenn wir annehmen, dass es zwei Lösungen u und $\tilde{u}$ gibt, dann erfüllt die Differenz $w := u - \tilde{u}$ die Differentialgleichung $Lw = 0$ und zugleich die Randbedingungen $w(0) = w(1) = 0$. Mit (7.10) erhalten wir $w = 0$, also $u = \tilde{u}$, somit, dass alle Lösungen gleich sind. □

Bemerkung 7.6 (Monotonieeigenschaft)**.** Eine weitere Folge aus dem Maximumsprinzip ist folgende Monotonieeigenschaft. Seien u und $\tilde{u}$ Lösungen der Dirichlet-Probleme $Lu = f$ und $L\tilde{u} = \tilde{f}$. Dann gilt

$$\tilde{f} \leq f \quad \Rightarrow \quad \tilde{u} \leq u. \tag{7.11}$$

Verallgemeinerung des Existenzresultats

Nun wollen wir die Existenz einer Lösung für das Problem

$$Lu := -u'' + bu' + cu = f, \qquad u(0) = 0, \quad u(1) = 0$$

zeigen. Auch für diese Verallgemeinerung benutzen wir eine Technik, die nur im eindimensionalen Fall funktioniert; wir erinnern uns an das Schießverfahren. Wir betrachten folgende Hilfsprobleme:

$$Lv = f, \qquad v(0) = 0, \qquad v'(0) = 0, \tag{7.12}$$
$$Lw = 0, \qquad w(0) = 0, \qquad w'(0) = 1. \tag{7.13}$$

Wir wissen, dass solche Anfangswertprobleme in Systeme erster Ordnung umgeschrieben werden können (Kapitel 5) und dass dann mithilfe des Satzes von Picard-Lindelöf (Satz 2.1) Existenz und Eindeutigkeit bewiesen werden kann.

Als Ansatz für die Lösung für des Dirichletproblems wählen wir $v+\lambda w$. Wegen der Linearität von L haben wir

$$Lu = f, \qquad u(0) = 0, \qquad u(1) = v(1) + \lambda w(1).$$

Der Parameter λ ist über die Gleichung $v(1) + \lambda w(1) = 0$ gegeben, die genau dann eindeutig lösbar ist, wenn $w(1) \neq 0$.

Lemma 7.7. *Die Lösung w des Hilfsproblems* (7.13) *erfüllt $w(1) \neq 0$.*

Beweis. Widerspruchsbeweis. Wir nehmen an, dass $w(1) = 0$. Dann ist w offensichtlich eine Lösung des Randwertproblems

$$Lw = 0, \qquad w(0) = 0, \qquad w(1) = 0. \tag{7.14}$$

Da $w'(0) = 1$, ist $w \not\equiv 0$. Offensichtlich hat das RWP (7.14) mit der Nulllösung eine zweite Lösung. Dies widerspricht aber dem Satz 7.5, der bereits die Eindeutigkeit für die Lösung des Randwertproblems (7.14) garantiert. □

Da wir nun mit $\lambda := v(1)/w(1)$ eine Lösung konstruieren können und Satz 7.5 Eindeutigkeit garantiert, erhalten wir folgendes Resultat.

Satz 7.8 (Existenz und Eindeutigkeit)**.** *Das Dirichletproblem* (7.1) – (7.3) *ist eindeutig lösbar.*

Stabilität

In diesem Fall wollen wir ein Stabilitätsresultat wie in Satz 7.3 zeigen, jedoch nun für den allgemeinen Fall, wir wollen also zeigen, dass die Lösung u eine Abschätzung der Art

$$\|u\|_\infty \leq C\|f\|_\infty$$

erfüllt, wobei die Konstante C zwar von b und c abhängen darf, nicht jedoch von $f = Lu$. Da wir bereits die Existenz und Eindeutigkeit einer Lösung wissen, können wir auch den Lösungsoperator $L^{-1} : C[0,1] \to C_0^2(0,1)$ einführen.

Mit diesem Lösungsoperator lässt sich die Stabilität als

$$\|L^{-1}\|_\infty := \sup_{g \in C(0,1)} \frac{\|L^{-1}g\|_\infty}{\|g\|_\infty} \leq C \tag{7.15}$$

schreiben, wobei $\|\cdot\|_\infty$ bei Anwendung auf einen Operator (hier: L^{-1}) die zur Supremumsnorm gehörende Operatornorm ist.

Lemma 7.9. *Sei L ein linearer Differentialoperator, der die Montonieeigenschaft* (7.11) *erfüllt. Dann gilt*

$$\|L^{-1}\|_\infty = \|L^{-1}\mathbf{1}\|_\infty,$$

wobei $\mathbf{1}$ *die konstante Funktion mit Wert* 1 *ist.*

Beweis. Sei $g \in C(0,1)$ beliebig und $\gamma := \|g\|_\infty$. Mit dieser Wahl gilt offensichtlich

$$-\gamma\mathbf{1} \leq g \leq \gamma\mathbf{1}.$$

Mit der Linearität und der Monotonieeigenschaft folgt

$$-\gamma L^{-1}\mathbf{1} \leq L^{-1}g \leq \gamma L^{-1}\mathbf{1}.$$

Für die Supremumsnorm folgt daraus $\|L^{-1}g\|_\infty \leq \gamma\|L^{-1}\mathbf{1}\|_\infty$ für alle $g \in C(0,1)$, also $\|L^{-1}g\|_\infty/\|g\|_\infty \leq \|L^{-1}\mathbf{1}\|_\infty$. Da g beliebig gewählt war, gilt somit $\|L^{-1}\|_\infty \leq \|L^{-1}\mathbf{1}\|_\infty$. Die andere Richtung ist trivial. □

Mit diesem Lemma folgt sofort

$$\|L^{-1}f\|_\infty \leq \|L^{-1}\|_\infty\|f\|_\infty = \|L^{-1}\mathbf{1}\|_\infty\|f\|_\infty,$$

wodurch wir eine Konstante $C := \|L^{-1}\mathbf{1}\|_\infty$ gefunden haben, die zwar vom Differentialoperator L (und damit von b und c), nicht jedoch von f abhängt.

Bemerkung 7.10. Mit den bereits gezeigten Monotonieeigenschaften können wir zeigen, dass $\|L^{-1}\|_\infty$ durch eine von $c \geq 0$ unabhängige Stabilitätskonstante beschränkt werden kann. Seien dazu

$$\begin{aligned} Lu &:= -u'' + bu' + cu = \mathbf{1}, & u(0) = u(1) = 0, \\ \tilde{L}w &:= -w'' + bw' = \mathbf{1}, & w(0) = w(1) = 0. \end{aligned}$$

Aus der Monotonieeigenschaft erhalten wir $u \geq 0$ und wegen $c \geq 0$ gilt

$$\tilde{L}u = -u'' + bu' = \mathbf{1} - cu \leq \mathbf{1} = \tilde{L}w$$

und daher unter nochmaliger Anwendung der Monotonieeigenschaft $0 \leq L^{-1}\mathbf{1} = u \leq w = \tilde{L}^{-1}\mathbf{1}$. In Operatornotation haben wir daher

$$\|L^{-1}\|_\infty = \|L^{-1}\mathbf{1}\|_\infty \leq \|\tilde{L}^{-1}\mathbf{1}\|_\infty = \|\tilde{L}^{-1}\|_\infty,$$

wobei $\|\tilde{L}^{-1}\|_\infty$ eine Konstante ist, die nur von b, nicht aber von c oder f abhängt.

8. Differenzenverfahren für lineare Dirichletprobleme

Im Folgenden besprechen wir Methoden zur näherungsweisen Lösung des Randwertproblems

$$-u''(x) + b(x)u'(x) + c(x)u(x) = f(x), \qquad x \in (0,1), \tag{8.1}$$
$$u(0) = 0 \tag{8.2}$$
$$u(1) = 0. \tag{8.3}$$

Für die Diskretisierung des Randwertproblems können wir – wie für die Anfangswertprobleme – ein Gitter

$$\mathcal{T}_h = \{x_1, \cdots, x_n\} \quad \text{mit} \quad 0 = x_0 < x_1 < \ldots < x_{n+1} = 1$$

einführen. Der Einfachheit halber nehmen wir ein äquidistantes Gitter an, also

$$x_i = i\ h,$$

wobei $h = (n+1)^{-1}$ (mit $n \in \mathbb{N} = \{1, 2, \ldots\}$) die Gitterweite ist. Die Funktionswerte der zugehörigen Näherungslösung an den Gitterpunkten x_i bezeichnen wir mit u_i. Die Funktionswerte an den Gitterpunkten sollen die exakte Lösung u approximieren; es gilt also $u_i \approx u(x_i)$. Die Näherungslösungen u_i können wir wieder zu einem Polygonzug u_h mit $u_h(x_i) = u_i$ verbinden.

Aus den Dirichlet-Randbedingungen (8.2) und (8.3) ergibt sich unmittelbar

$$u_0 = 0, \qquad u_{n+1} = 0.$$

Für die Diskretisierung von u'' bietet sich insbesondere der zentrale Differenzenquotient an:

$$u''(x) \approx D_h^2 u(x) := \tfrac{1}{h^2}\left(u(x+h) - 2u(x) + u(x-h)\right).$$

Für die Diskretisierung von u' kommen uns sofort mehrere Alternativen in den Sinn, insbesondere der zentrale Differenzenquotient

$$u'(x) \approx D_{2h} u(x) := \tfrac{1}{2h}\left(u(x+h) - u(x-h)\right),$$

sowie Vorwärts- und Rückwärtsdifferenzenquotienten

$$D_h^+ u(x) := \tfrac{1}{h}\left(u(x+h) - u(x)\right), \qquad \text{und} \qquad D_h^- u(x) := \tfrac{1}{h}\left(u(x) - u(x-h)\right).$$

Soweit im Folgenden nichts anderes angenommen ist, nehmen wir an, dass die Annäherung durch den zentralen Differenzenquotienten erfolgt (vgl. auch Bemerkung 8.18. Durch Einsetzen in die Differenzialgleichung ergibt sich nun ein Differenzenschema

$$\begin{aligned} \tfrac{1}{h^2}(-u_{i-1} + 2u_i - u_{i+1}) + \tfrac{b(x_i)}{2h}(-u_{i-1} + u_{i+1}) + c(x_i)u_i &= f(x_i), \qquad i = 1, \dots, n, \\ u_0 &= 0, \\ u_{n+1} &= 0. \end{aligned} \tag{8.4}$$

Dieses Schema kann offensichtlich in der Form eines linearen Gleichungssystems geschrieben werden kann, wo wir die bereits bekannten Werte für u_0 und u_{n+1} eliminieren und $b_i := b(x_i)$, $c_i := c(x_i)$ und $f_i := f(x_i)$ nutzen:

$$\underbrace{\frac{1}{h^2}\begin{pmatrix} 2+c_1h^2 & -1+\frac{h}{2}b_1 & & & \\ -1-\frac{h}{2}b_2 & 2+c_2h^2 & -1+\frac{h}{2}b_2 & & \\ & \ddots & \ddots & \ddots & \\ & & -1-\frac{h}{2}b_{n-1} & 2+c_{n-1}h^2 & -1+\frac{h}{2}b_{n-1} \\ & & & -1-\frac{h}{2}b_n & 2+c_nh^2 \end{pmatrix}}_{A_h :=} \underbrace{\begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{n-1} \\ u_n \end{pmatrix}}_{\underline{u}_h :=} = \underbrace{\begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_{n-1} \\ f_n \end{pmatrix}}_{\underline{f}_h :=}. \tag{8.5}$$

Die Matrix $A_h \in \mathbb{R}^{n\times n}$ wird als *Steifigkeitsmatrix*, der Vektor $\underline{f}_h$ als *Lastvektor* bezeichnet. Dieses Gleichungssystem lässt sich durch $A_h := [a_{i,j}]_{i,j=1}^n$, $\underline{u}_h = [u_i]_{i=1}^n$, $\underline{f}_h = [f_i]_{i=1}^n$ und

$$a_{i,i-1} = -\frac{1}{h^2} - \frac{b_i}{2h}, \qquad a_{i,i} = \frac{2}{h^2} + c_i, \qquad a_{i,i+1} = -\frac{1}{h^2} + \frac{b_i}{2h},$$

sowie $a_{i,j} = 0$ für $|i-j| \geq 2$ eindeutig beschreiben.

Nun stellen sich einige Fragen. Ist die Matrix A_h invertierbar? Wie groß ist der Fehler $E_h(x_i) = u(x_i) - u_i$ bzw. seine Norm? Wie verhält sich die Norm des Fehlers für $h \to 0$?

Eindeutige Lösbarkeit und diskretes Maximumsprinzip

Zuerst wollen wir uns mit der eindeutigen Lösbarkeit des diskretisierten Problems beschäftigen. Da die Matrix A_h eine quadratische Matrix ist, ist das System (8.5) offensichtlich genau dann eindeutig lösbar, wenn die Matrix A_h invertierbar (=regulär) ist. Dies ist genau dann der Fall, wenn $A_h\underline{w}_h = 0$ nur eine einzige Lösung hat.

Für die weitere Analyse führen wir den diskreten Differentialoperator

$$L_h u(x) := -D_h^2 u(x) + b(x) D_{2h} u(x) + c(x) u(x)$$

ein. Außerdem führen wir das bewährte Konzept des Polygonzugs wieder ein. Die Räume der Polygonzüge sind gegeben durch

$$V_h := \{v \in C[0,1] : v|_{(x_i, x_{i+1})} \text{ linear}\} \quad \text{und} \quad V_{h,0} := C_0[0,1] \cap V_h.$$

Der dem Vektor $\underline{u}_h = (u_1, \ldots, u_n)^\top$ zugeordnete Polygonzug ist jene Funktion $u_h \in V_{h,0}$, die

$$u_h(x_i) = u_i \quad \forall x_i \in \mathcal{T}_h$$

erfüllt. Wir können nun das diskretisierte Problem (8.5) auch folgendermaßen schreiben: Finde ein $u_h \in V_{h,0}$, sodass

$$L_h u_h(x_i) = f(x_i) \quad \forall x_i \in \mathcal{T}_h.$$

Jede Lösung dieses Problems erfüllt das folgende diskrete Maximumsprinzip.

Satz 8.1 (Diskretes Maximumsprinzip)**.** *Sei* $L_h u_h(x_i) = a_{i,i-1} u_h(x_{i-1}) + a_{i,i} u_h(x_i) + a_{i,i+1} u_h(x_{i+1})$ *mit* $a_{i,i-1} < 0$ *und* $a_{i,i+1} < 0$ *und* $a_{i,i-1} + a_{i,i} + a_{i,i+1} \geq 0$ *für* $i = 1, \ldots, n$. *Sei* u_h *ein Polygonzug mit* $L_h u_h(x_i) \leq 0$ *für alle* $x_i \in \mathcal{T}_h$. *Dann gilt*

$$u_h(x_i) \leq \max\{0, u_h(0), u_h(1)\}, \qquad \forall x_i \in \mathcal{T}_h.$$

Wenn außerdem $a_{i,i-1} + a_{i,i} + a_{i,i+1} = 0$, *dann gilt*

$$u_h(x_i) \leq \max\{u_h(0), u_h(1)\}, \qquad \forall x_i \in \mathcal{T}_h.$$

Wenn die jeweilige Ungleichung mit Gleichheit erfüllt ist, dann ist u_h *konstant.*

Beweis. Wir verwenden $u_i = u_h(x_i)$. Wir nehmen an, dass es einen Gitterpunkt x_k im Inneren gibt, sodass $u_k \geq \max\{0, u_h(0), u_h(1)\}$. Offensichtlich muss es dann aber auch im Inneren ein lokales Maximum geben, also ein $i \in \{1, \ldots, n\}$ mit

$$u_i \geq 0 \qquad \text{und} \qquad u_i \geq u_j \quad \forall j \in \{0, 1, \ldots, n, n+1\}.$$

Wir haben nun

$$L_h u_h = \underbrace{(-a_{i,i-1})}_{>0} \underbrace{(u_i - u_{i-1})}_{\geq 0} + \underbrace{(-a_{i,i+1})}_{>0} \underbrace{(u_i - u_{i+1})}_{\geq 0} + \underbrace{(a_{i,i} + a_{i,i-1} + a_{i,i+1})}_{\geq 0} \underbrace{u_i}_{\geq 0}.$$

Da alle Summanden ≥ 0 sind, folgt aus $L_h u_h \leq 0$, dass alle Summanden verschwinden müssen, also insbesondere auch

$$u_{i+1} = u_{i-1} = u_i.$$

Durch wiederholte Anwendung des Arguments folgt daraus, dass u_h konstant ist, also auch $u_k \leq \max\{0, u_h(0), u_h(1)\}$ folgt.

Wenn $a_{i,i-1} + a_{i,i} + a_{i,i+1} = 0$ gilt, dann können wir auf die Überlegungen zum Vorzeichen von u_h verzichten. □

Für das Differenzenschema (8.4) gilt

$$a_{i,i-1} = -\tfrac{1}{h^2} - \tfrac{b_i}{2h}, \qquad a_{i,i+1} = -\tfrac{1}{h^2} + \tfrac{b_i}{2h}, \qquad a_{i,i} + a_{i,i-1} + a_{i,i+1} = c_i \geq 0$$

und daher ist das diskrete Maximumsprinzip anwendbar, wenn $h < 2/\|b\|_\infty$ gilt. (Beachte auch das Verhalten am Rand. Wegen $u_0 = u_{n+1} = 0$ ist die Wahl von $a_{1,0}$ und $a_{n,n+1}$ irrelevant.)

Satz 8.2. *Wenn $h < 2/\|b\|_\infty$ und $f \leq 0$ gilt, dann gilt für jede Lösung des Differenzenschemas* (8.4) *$u_h \leq 0$.*

Beweis. Verwende Maximumsprinzip und die Randbedingungen. □

Wie wir bereits im letzten Kapitel gesehen haben, können wir wieder f durch $-f$ ersetzen und erhalten dadurch, dass das Differenzenschema für $f = 0$ nur die Lösung $u_h = 0$ hat. In der Matrix-Vektor-Notation bedeutet das, dass das Gleichungssystem $A_h \underline{u}_h = 0$ eindeutig lösbar ist. Daraus folgt die Invertierbarkeit von A_h und damit auch die Lösbarkeit des Problems (Existenz und Eindeutigkeit).

Satz 8.3 (Existenz und Eindeutigkeit)**.** *Das Differenzenschema* (8.4) *ist eindeutig auflösbar, wenn $h < 2/\|b\|_\infty$.*

Bemerkung 8.4. Für $b = 0$ haben wir keine Einschränkung an die Schrittweite!

Bemerkung 8.5 (M-Matrix)**.** In diesem Abschnitt haben wir gezeigt, dass die Steifigkeitsmatrix A_h eine sogenannte M-Matrix ist. Eine M-Matrix $A = [a_{i,j}]_{i,j=1}^n$ ist definiert als eine Matrix,

- die invertierbar ist,
- die $a_{i,j} \leq 0$ für alle $i \neq j$ erfüllt, und
- deren Inverse komponentenweise nicht negativ ist, kurz $A^{-1} \geq 0$.

Diskrete Stabilität für $b = 0$

Da wir die Existenz und Eindeutigkeit einer Lösung zum Differenzenschema gezeigt haben, können wir auch den Lösungsoperator $L_h^{-1} : C(0,1) \to V_{h,0}$

einführen und verwenden, der jeder rechten Seite f die Lösung $u_h \in V_{h,0}$ mit $L_h\, u_h(x_i) = f(x_i)$ für alle $x_i \in \mathcal{T}_h$ zuordnet. Analog zum Begriff der Stabilität für das Dirichletproblem aus dem letzten Kapitel wollen wir nun

$$\|L_h^{-1}\|_{\infty,h} := \sup_{f \in C(0,1)} \frac{\|L_h^{-1} f\|_{\infty,h}}{\|f\|_{\infty,h}}, \qquad \|f\|_{\infty,h} := \max_{x_i \in \mathcal{T}_h} |f(x_i)|$$

von oben beschränken. Wie wir dies bereits für die Anfangswertprobleme kennen, ist es uns wichtig, dass $\|L_h^{-1}\|_{\infty,h}$ mit einer von der Gitterweite unabhängigen Konstante h beschränkt werden kann, wir also an einer gleichmäßigen Beschränktheit interessiert sind.

Definition 8.6 (Diskrete Stabilität)**.**
(i) Ein Differenzenverfahren für ein Randwertproblem heißt *diskret stabil für Gitterweite* h^* (bezüglich der Supremumsnorm), wenn ein $C > 0$ existiert, sodass der Operator L_h für alle $h \in (0, h^*]$ invertierbar ist (also L_h^{-1} existiert) und

$$\|L_h^{-1}\|_{\infty,h} \leq C.$$

(ii) Ein Differenzenverfahren für ein Randwertproblem heißt *diskret stabil*, wenn es eine Gitterweite $h^* > 0$ gibt, sodass es stabil im Sinne von (i) ist.

Spezialfall $b, c = 0$**.** In diesem Fall haben wir

$$L_h u_h(x) = -D_h^2 u_h(x) = \tfrac{1}{h^2}(-u_h(x-h) + 2u_h(x) - u_h(x+h)).$$

Analog zum Stabilitätsresultat von Satz 7.3 wollen wir nun ein diskretes Stabilitätsresultat zeigen. Der Beweis von Satz 7.3 basiert auf der Green'schen Funktion

$$G(x,y) = \begin{cases} y(1-x) & y \leq x \\ x(1-y) & y > x \end{cases}.$$

Satz 8.7. *Für jedes $f \in C(0,1)$, erfüllt jeder Polygonzug $u_h \in V_h$ mit*

$$u_h(x_i) = h \sum_{j=1}^{n} G(x_i, x_j) f(x_j) \tag{8.6}$$

die Dirichlet-Bedingungen sowie $L_h u_h(x_i) = f(x_i)$ für alle $x_i \in \mathcal{T}_h$.

Beweis. Sei $G_j \in V_{h,0}$ jener Polygonzug, der $G_i(x_j) = G(x_i, x_j)$ erfüllt. Man kann einfach ausrechnen, dass

$$L_h G_i(x_j) = h^{-2}\left(-G(x_i, x_{j+1}) + 2G(x_i, x_j) - G(x_i, x_{j-1})\right) = 0 \quad \text{für alle} \quad i \neq j,$$

siehe Übung. Für $i = j$ haben wir nun

$$\begin{aligned} L_h G_j(x_j) &= h^{-2}\left(-G(x_j, x_{j+1}) + 2G(x_j, x_j) - G(x_j, x_{j-1})\right) \\ &= h^{-2}\left(-x_j(1 - x_{j+1}) + 2x_j(1 - x_j) - x_{j-1}(1 - x_j)\right) \\ &= h^{-2}\left(-x_j(1 - x_j - h) + 2x_j(1 - x_j) - (x_j + h)(1 - x_j)\right) = h^{-1}. \end{aligned}$$

Daraus folgt nun $L_h G_i(x_j) = h^{-1}\delta_{i,j}$, wobei

$$\delta_{i,j} := \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

das Kronecker-delta ist.

Sei nun f gegeben. Wir zeigen, dass das durch (8.6) gegebene u_h die Operatorgleichung erfüllt, also eine Lösung ist. Es gilt

$$L_h u_h(x_i) = h\sum_{j=1}^{n} L_h G_j(x_i) f(x_j) = h\sum_{j=1}^{n} h^{-1}\delta_{i,j} f(x_j) = f(x_i)$$

für alle $x_i \in \mathcal{T}_h$, was zu zeigen war. □

Wie im kontinuierlichen Fall können wir nun ein diskretes Stabilitätsresultat zeigen.

Satz 8.8 (Diskrete Stabilität)**.** *Es gilt $\|L_h^{-1}\|_{\infty,h} \leq \frac{1}{8}$; damit liegt für alle Gitterweiten h^* diskrete Stabilität vor.*

Beweis. Sei $f \in C(0,1)$ beliebig und $u_h = L_h^{-1} f$. Dann gilt

$$\begin{aligned} |u_h(x_i)| &= \left| h\sum_{j=1}^{n} G(x_j, x_i) f(x_j) \right| \leq \sum_{j=1}^{n} hG(x_j, x_i) \max_j |f(x_j)| \\ &= \left(\sum_{j=1}^{i} x_j(1 - x_i) + \sum_{j=i+1}^{n} x_i(1 - x_j) \right) h\|f\|_{\infty,h} \\ &= \left(\tfrac{1}{2}(i+1)ih(1 - ih) + \tfrac{1}{2}ih(i - n)(ih + h + nh - 2)\right) h\|f\|_{\infty,h} \\ &= \tfrac{1}{2}\left(ih - i^2h^2\right)\|f\|_{\infty,h} = \tfrac{1}{2}\tfrac{i}{n}(1 - \tfrac{i}{n})\|f\|_{\infty,h} \leq \tfrac{1}{8}\|f\|_{\infty,h} \end{aligned}$$

für alle $x_i \in T_h$. Daraus folgt die Aussage. □

Erweiterung auf den Fall $b = 0$, $c \geq 0$. Analog zum letzten Kapitel erhalten wir, dass wir für die Bestimmung der Stabilitätskonstante nur die konstante Funktionen **1** als rechte Seite berücksichtigen müssen.

Lemma 8.9. *Sei L_h ein lineares Differenzenschema, das die Montonieeigenschaft*

$$f \leq 0 \quad \Rightarrow \quad L_h^{-1} f \leq 0$$

erfüllt. Dann gilt

$$\|L_h^{-1}\|_{\infty,h} = \|L_h^{-1}\mathbf{1}\|_{\infty,h}$$

Beweis. Ganz analog zum Beweis von Satz 7.9. □

Mit diesem Satz erhalten wir folgende Aussage für den Fall $b = 0$, $c \geq 0$, also

$$L_h u_h(x) = -D_h^2 u_h(x) + c(x)u_h(x) = \tfrac{1}{h^2}(-u_h(x-h) + 2u_h(x) - u_h(x+h)) + c(x)u_h(x).$$

Dabei benutzen wir, dass die geforderte Monotonieeigenschaft aus dem Maximumsprinzip folgt.

Satz 8.10 (Diskrete Stabilität 2)**.** *Es gilt $\|L_h^{-1}\|_{\infty,h} \leq \frac{1}{8}$; damit liegt für alle Gitterweiten h^* diskrete Stabilität vor.*

Beweis. Seien

$$\begin{aligned} L_h u_h &:= -D_h^2 u_h + c u_h = \mathbf{1}, & u_h(0) = u_h(1) = 0, \\ \tilde{L}_h w_h &:= -D_h^2 w_h = \mathbf{1}, & w_h(0) = w_h(1) = 0. \end{aligned}$$

Aus der Monotonieeigenschaft erhalten wir $u_h \geq 0$ und wegen $c \geq 0$ gilt

$$\tilde{L}_h u_h = -D_h^2 u_h = \mathbf{1} - c u_h \leq \mathbf{1} = \tilde{L}_h w_h$$

und daher unter nochmaliger Anwendung der Monotonieeigenschaft

$$0 \leq L_h^{-1}\mathbf{1} = u_h \leq w_h = \tilde{L}_h^{-1}\mathbf{1}.$$

In Operatornotation haben wir daher

$$\|L_h^{-1}\|_{\infty,h} = \|L_h^{-1}\mathbf{1}\|_{\infty,h} \leq \|\tilde{L}_h^{-1}\mathbf{1}\|_{\infty,h} = \|\tilde{L}_h^{-1}\|_{\infty,h} = \tfrac{1}{8},$$

wobei wir für die letzte Abschätzung Satz 8.8 verwendet haben. □

Diskrete Stabilität für den allgemeinen Fall

Folgender Satz (gültig für Dirichletprobleme) ist für die weitere Analyse hilfreich.

Satz 8.11. *Sei* $L_h u_h(x_i) = a_{i,i-1}u_h(x_{i-1}) + a_{i,i}u_h(x_i) + a_{i,i+1}u_h(x_{i+1})$ *für alle* $x_i \in \mathcal{T}_h$, *wobei* $a_{i,i-1}+a_{i,i}+a_{i,i+1} \geq 0$. *Wenn es Konstanten* $0 < \alpha \leq 1 \leq \beta$ *gibt, sodass* $\{-h^2 a_{i,i-1}, -h^2 a_{i,i+1}\} \subset [\alpha, \beta]$ *für alle* $i = 1, \ldots, n$, *dann gilt* $\|L_h^{-1}\| \leq (\beta/\alpha)^{n+1}$.

Beweis. Im Folgenden wollen wir nun motiviert durch Lemma 8.9 einen Polygonzug $v_h \in V_{h,0}$ finden, sodass $L_h v_h(x_i) \geq 1$ für alle $x_i \in \mathcal{T}_h$. Sei nun

$$Q_k := \prod_{\ell=1}^{k} \frac{a_{\ell,\ell-1}}{a_{\ell,\ell+1}} > 0, \qquad k = 0, \ldots, n.$$

Wir definieren

$$v_k := \tfrac{h^2}{\alpha} \sum_{i=1}^{k} \sum_{j=i}^{n} \frac{Q_{i-1}}{Q_j} \geq 0, \qquad k = 0, \ldots, n+1$$

und stellen die Ähnlichkeit zur Konstruktion der Green'schen Funktion fest. Mit $v_k - v_{k-1} = \frac{h^2}{\alpha} \sum_{j=k}^{n} \frac{Q_{k-1}}{Q_j}$ erhalten wir für $k = 1, \ldots, n$

$$\begin{aligned} f_k &:= a_{k,k-1}v_{k-1} + a_{k,k}v_k + a_{k,k+1}v_{k+1} \geq a_{k,k+1}(v_{k+1} - v_k) - a_{k,k-1}(v_k - v_{k-1}) \\ &= \tfrac{h^2}{\alpha} \left(a_{k,k+1} \sum_{j=k+1}^{n} \frac{Q_k}{Q_j} - a_{k,k-1} \sum_{j=k}^{n} \frac{Q_{k-1}}{Q_j} \right) \\ &= \tfrac{h^2}{\alpha} \left(a_{k,k+1} \sum_{j=k+1}^{n} \frac{Q_k}{Q_j} - a_{k,k+1} \sum_{j=k}^{n} \frac{Q_k}{Q_j} \right) = \tfrac{h^2}{\alpha} a_{k,k+1} \left(\sum_{j=k+1}^{n} \frac{Q_k}{Q_j} - \sum_{j=k}^{n} \frac{Q_k}{Q_j} \right) \\ &= -\tfrac{h^2}{\alpha} a_{k,k+1} \geq 1. \end{aligned}$$

Sei nun $v_h \in V_{h,0}$ mit $v_h(x_k) = v_k$ für $k = 1, \ldots, n$. (Wegen $v_h \in V_{h,0}$ gilt schon $v_h(x_0) = 0$ und $v_h(x_{n+1}) = 0$.) Für $k = 2, \ldots, n-1$ haben wir keine Randterme zu beachten und es gilt $L_h v_h(x_k) = f_k \geq 1$. Für $k = 1$ haben wir wegen $v_0 = v_h(x_0) = 0$, dass $L_h v_h(x_1) = f_1 - a_{1,0}v_0 = f_1 \geq 1$. Für $k = n$ haben wir wegen $v_{n+1} \geq 0 = v_h(x_{n+1})$, dass $L_h v_h(x_n) = f_n - a_{n,n+1}v_{n+1} \geq f_n \geq 1$. Damit haben wir $L_h v_h(x_k) \geq \mathbf{1}$ für alle $x_k \in \mathcal{T}_h$ gezeigt. Das bedeutet nun mit Lemma 8.9, dass wir

$$\|L_h^{-1}\|_{\infty,h} = \|L_h^{-1}\mathbf{1}\|_{\infty,h} \leq \|v_h\|_{\infty,h} = \max_{k=1,\ldots,n} |v_k|$$

wissen. Wir haben wegen $h^2 = (n+1)^{-2} \leq n^2$ für alle $k = 1, \ldots, n$, dass

$$|v_k| = \frac{h^2}{\alpha} \left| \sum_{i=0}^{k-1} \sum_{j=i+1}^{n} \frac{Q_i}{Q_j} \right| \leq \frac{1}{\alpha} \max_{0<i<j\leq n} \left| \frac{Q_i}{Q_j} \right| = \frac{1}{\alpha} \max_{0<i<j\leq n} \left| \frac{a_{i+1,i+2} \cdots a_{j,j+1}}{a_{i+1,i} \cdots a_{j,j-1}} \right| \leq \frac{\beta^n}{\alpha^{n+1}},$$

was wegen $\beta \geq 1$ den Beweis abschließt. □

Nun können wir die diskrete Stabilität zeigen.

Satz 8.12 (Diskrete Stabilität 3)**.** *Das Differenzenverfahren ist diskret stabil für Gitterweiten $h^* < 2/\|b\|_\infty$ im Sinne der Definition 8.6.*

Beweis. Bei Verwendung des zentralen Differenzenquotienten wissen wir, dass $a_{i,i-1} = -\frac{1}{h^2} - b_i \frac{1}{2h}$ und $a_{i,i+1} = -\frac{1}{h^2} + b_i \frac{1}{2h}$ und $a_{i,i-1} + a_{i,i} + a_{i,i+1} = c(x_i) \geq 0$ gilt. Damit sind die Voraussetzungen des letzten Satzes mit

$$0 < \alpha = 1 - \|b\|_\infty \tfrac{h^*}{2} < 1$$

und $\beta = 1 + \|b\|_\infty \frac{h^*}{2} > 1$ erfüllt, womit wir

$$\|L_h^{-1}\|_{\infty,h} \leq \left(\frac{1+\|b\|_\infty \frac{h}{2}}{1-\|b\|_\infty \frac{h}{2}} \right)^{1/h}$$

für jedes $h = (n+1)^{-1} \in [0, h^*)$ erhalten. Wir benötigen noch, dass dieser Term durch eine von h unabhängige Konstante C_{stab} (die aber von h^* abhängen darf) nach oben beschränkt ist. Sei $\psi(x) := \left(\frac{1+x}{1-x}\right)^{1/x}$. Durch elementare Rechnung erhält man, dass ψ auf $[0, 1)$ streng monoton steigend ist; vgl. Übung. Damit haben wir

$$\|L_h^{-1}\|_{\infty,h} \leq \psi\left(\tfrac{\|b\|_\infty h}{2}\right)^{\|b\|_\infty/2} \leq \underbrace{\psi\left(\tfrac{\|b\|_\infty h^*}{2}\right)^{\|b\|_\infty/2}}_{C_{stab}:=},$$

was den Beweis abschließt. □

Konsistenz und Konvergenz

Zur Abschätzung des Diskretisierungsfehlers verwenden wir – wie bei den Anfangswertproblemen – die Konsistenz, also Abschätzungen der Abschneidefehlers.

Definition 8.13 (Konsistenz)**.** Sei $f \in C(0,1)$ gegeben und $u \in C_0^2(0,1)$ mit $Lu = f$.

(i) Wir bezeichnen die Gitterfunktion $\tau_h \in V_{0,h}$ mit

$$\tau_h(x_i) = (L_h u)(x_i) - f(x_i)$$

als den Abschneidefehler (*truncation error*) oder das Residuum.
(ii) Wir bezeichnen das Verfahren als konsistent, wenn

$$\lim_{h\to 0} \|\tau_h\|_{\infty,h} = \lim_{h\to 0} \max_i |\tau_h(x_i)| = 0.$$

(iii) Wir bezeichnen das Verfahren als konsistent mit Ordnung p, wenn

$$\|\tau_h\|_{\infty,h} \le Ch^p$$

mit einer Konstante C, die nicht von h oder u_h abhängt.

Wir bemerken, dass für den Abschneidefehler der diskrete Differentialoperator L_h nicht auf die Näherungslösung u_h, sondern auf die exakte Lösung angewandt wird.

Satz 8.14 (Konsistenz)**.** *Wenn $u \in C_0^2(0,1) \cap C^4(0,1)$, dann ist das Differenzenschema* (8.4) *konsistent mit Ordnung* 2*; wenn nur $u \in C_0^2(0,1) \cap C^3(0,1)$, dann mit Ordnung* 1*.*

Beweis. Wir haben mit der Dreiecksungleichung

$$\begin{aligned}\|L_h u - f\|_{\infty,h} &= \| - D_h^2 u + bD_{2h}u + cu - (-u'' + bu' + cu)\|_{\infty,h} \\ &\le \max_{x_i\in\mathcal{T}_h} |u''(x_i) - D_h^2 u(x_i)| + \|b\|_\infty \max_{x_i\in\mathcal{T}_h} |u'(x_i) - D_{2h}u(x_i)|.\end{aligned}$$

Im ersten Fall erhalten wir mit der Taylor-Entwicklung $u(x_i \pm h) = u(x_i) \pm hu'(x_i) + \frac{h^2}{2}u''(x_i) \pm \frac{h^3}{6}u'''(x_i) + \mathcal{O}(h^4)$

$$\begin{aligned}u''(x_i) - D_h^2 u(x_i) &= u''(x_i) + \frac{1}{h^2}\left(-u(x_i - h) + 2u(x_i) - u(x_i + h)\right) \\ &= u''(x_i) - \frac{1}{h^2}\Big(\quad u(x_i) - hu'(x_i) + \tfrac{h^2}{2}u''(x_i) - \tfrac{h^3}{6}u'''(x_i) + \mathcal{O}(h^4) \\ &\qquad\qquad - 2u(x_i) \\ &\qquad\qquad + u(x_i) + hu'(x_i) + \tfrac{h^2}{2}u''(x_i) + \tfrac{h^3}{6}u'''(x_i) + \mathcal{O}(h^4)\Big) = \mathcal{O}(h^2).\end{aligned}$$

Im zweiten Fall müssen wir mit einem Taylorpolynom vom Grad 2 auskommen. Für die Abschätzungen der ersten Ableitungen nutzen wir $u(x_i \pm h) = u(x_i) \pm$

$hu'(x_i) + \frac{h^2}{2}u''(x_i) + \mathcal{O}(h^3)$ und erhalten

$$\begin{aligned} u'(x_i) - D_{2h}u(x_i) &= u'(x_i) - \frac{1}{2h}\left(u(x_i+h) - u(x_i-h)\right) \\ &= u'(x_i) - \frac{1}{2h}\Big(u(x_i) + hu'(x_i) + \tfrac{h^2}{2}u''(x_i) + \mathcal{O}(h^3) \\ &\qquad - u(x_i) + hu'(x_i) - \tfrac{h^2}{2}u''(x_i) + \mathcal{O}(h^3)\Big) = \mathcal{O}(h^2). \end{aligned}$$

Daraus folgt für den ersten Fall $\|\tau_h\|_{\infty,h} = \|L_h u - f\|_{\infty,h} = \mathcal{O}(h^2)$, was zu zeigen ist. Der zweite Fall kann ganz analog behandelt werden. □

Nun können wir auch den Konvergenzbegriff einführen.

Definition 8.15 (Konvergenz)**.** Sei $f \in C(0,1)$ gegeben. Seien $u \in C_0^2(0,1)$ und $u_h \in V_{0,h}$, sodass $Lu = f$ und $L_h\, u_h = f$.
(i) Wir bezeichnen den Polygonzug $E_h \in V_{0,h}$ mit

$$E_h(x_i) = u_h(x_i) - u(x_i)$$

als den Diskretisierungsfehler.
(ii) Wir bezeichnen das Verfahren als konvergent, wenn

$$\lim_{h\to 0} \|E_h\|_{\infty,h} = 0.$$

(iii) Wir bezeichnen das Verfahren als konvergent mit Ordnung p, wenn

$$\|E_h\|_{\infty,h} \le Ch^p$$

mit einer Konstante C, die nicht von h oder u_h abhängt.

Wir zeigen nun folgendes allgemeine Resultat.

Satz 8.16. *Wenn das Verfahren diskret stabil (also $\|L_h^{-1}\|_{\infty,h} \le C$ mit einer von h unabhängigen Konstante C) und konsistent ist (mit Ordnung p), dann ist es auch konvergent (mit Ordnung p).*

Beweis. Sei $f \in C(0,1)$ gegeben. Sei $u \in C_0^2(0,1)$, sodass $Lu = f$ auf $(0,1)$ und sei $u_h \in V_{0,h}$, sodass $L_h\, u_h = f_h$ auf $\mathcal{T}_h$. Es gilt nun

$$\begin{aligned} \|E_h\|_{\infty,h} = \|u - u_h\|_{\infty,h} &= \big\|L_h^{-1}L_h\big(u - u_h\big)\big\|_{\infty,h} \\ &\le \|L_h^{-1}\|_{\infty,h}\|L_h u - f\|_{\infty,h} = \|L_h^{-1}\|_{\infty,h}\|\tau_h\|_{\infty,h}. \end{aligned}$$

Mit Stabilität und Konsistenz erhalten wir nun die gewünschten Aussagen. □

Die Kombination der vorstehenden Sätze ergibt folgendes Resultat.

Satz 8.17 (Konvergenz)**.** *Wenn $u \in C_0^2(0,1) \cap C^4(0,1)$, dann ist das Differenzenschema* (8.4) *konvergent mit Ordnung* 2 *für alle Gitterweiten $h^* < 2/\|b\|_\infty$; ist nur $u \in C_0^2(0,1) \cap C^3(0,1)$, dann reduziert sich die Ordnung auf* 1*.*

Bemerkung 8.18. Wenn wir die ersten Ableitungen durch den Vorwärts- oder Rückwärtsdifferenzenquotienten approximieren, so bekommen wir ganz anloge Resultate. Für das Maximumsprinzip, die diskrete Stabilität und die Konvergenz benötigen wir nun $h^* < 1/\|b\|_\infty$. Außerdem kann man bei Verwendung von Vorwärts- oder Rückwärtsdifferenzenquotienten nur eine Kosistenz- und damit Konvergenzordnung von 1 erreichen.

Erweiterung auf andere Randbedingungen

Schließlich wollen wir noch die Resultate dieses Kapitels auf den Fall mit anderen Randbedingungen übertragen und dabei auftretende Probleme diskutieren.

Beispiel 8.19 (Inhomogene Dirichlet-Randbedingungen)**.**
Für die Lösung des Randwertproblems

$$Lu(x) := -u''(x) + b(x)u'(x) + c(x)u(x) = f(x), \qquad x \in (0,1), \tag{8.7}$$
$$u(0) = g_0, \tag{8.8}$$
$$u(1) = g_1 \tag{8.9}$$

mit inhomogenen Dirichtlet-Randbedingungen machen wir den Ansatz

$$u = v + w,$$

wobei $v \in C_0^2(0,1)$ und

$$w(x) = g_0(1-x) + g_1(x). \tag{8.10}$$

Offensichtlich erfüllt v das Randwertproblem mit homogenen Randbedingungen:

$$Lv(x) = f(x) - \underbrace{Lw(x)}_{=bw'+cw}, \qquad x \in (0,1),$$
$$v(0) = 0,$$
$$v(1) = 0,$$

auf das wir die Methoden dieses Kapitels anwenden können. Da wir hier ein homogenes Problem erhalten, sprechen wir auch von *Homogenisierung*.

Aus den nun für das homogene Problem bekannten Eigenschaften lassen sich auch entsprechende Eigenschaften für das Problem (8.7) – (8.9) herleiten, wie etwa Existenz und Eindeutigkeit einer Lösung oder die stetige Abhängigkeit von den Daten:

$$\begin{aligned}\|u\|_\infty &\le \|v\|_\infty + \|w\|_\infty \le C\|f - Lw\|_\infty + \|w\|_\infty \\ &\le C\|f\|_\infty + (1 + 2\|b\|_\infty + \|c\|_\infty)\max\{|g_0|, |g_1|\}.\end{aligned}$$

Für die praktische Realisierung der Differenzenschemata wählt man in der Regel nicht den Weg der Homogenisierung, da der Übergang von v auf u (also die Addition von w zur errechneten Lösung) eine nicht-lokale Operation ist, deren Erweiterung insbesondere auf zwei- und dreidimenisonale Randwertprobleme keineswegs trivial ist. Stattdessen setzt man das Differenzenschema für das inhomogene Problem an. Hier beschränken wir uns der Einfachheit halber auf den Fall $b = 0$:

$$\begin{aligned} u_0 &= g_0 \\ \tfrac{1}{h^2}(-u_{i-1} + 2u_i - u_{i+1}) + c_i u_i &= f_i := f(x_i), \qquad i = 1, \dots, n \\ u_{n+1} &= g_1, \end{aligned}$$

wo wir bei Elimination der bekannten Werte g_0 und g_1 folgendes System erhalten:

$$\frac{1}{h^2}\begin{pmatrix} 2+h^2c_1 & -1 & & & \\ -1 & 2+h^2c_2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2+h^2c_{n-1} & -1 \\ & & & -1 & 2+h^2c_n \end{pmatrix}\begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{n-1} \\ u_n \end{pmatrix} = \begin{pmatrix} f_1 + h^{-2}g_0 \\ f_2 \\ \vdots \\ f_{n-1} \\ f_n + h^{-2}g_1 \end{pmatrix}.$$

Die Steifigkeitsmatrix unterscheidet sich nicht von der im homogenen Fall erhaltenen Matrix, daher kann das System genau so gelöst werden, wie auch das System im homogenen Fall. Man kann sich leicht überlegen, dass dieser Ansatz genau die gleichen Lösungen liefert, wie der Ansatz der Homogenisierung. Daher erhalten wir auch für diese Formulierung dieselben Stabilitäts- und Konvergenzaussagen, wie für das homogenisierte Problem, insbesondere wieder Konvergenz $\|u - u_h\|_{\infty,h} \le Ch^p$ mit Ordnung $p = 2$ bzw. 1, wie in Satz 8.17.

Nun wollen wir uns kurz den allgemeineren Randbedingungen widmen.

Satz 8.20. *Das Problem*

$$Lu(x) := -u''(x) + b(x)u'(x) + c(x)u(x) = f(x), \qquad x \in (0,1) \tag{8.11}$$

$$Ru := \begin{pmatrix} \alpha_0 u(0) - \beta_0 u'(0) \\ \alpha_1 u(1) + \beta_1 u'(1) \end{pmatrix} = \begin{pmatrix} g_0 \\ g_1 \end{pmatrix} =: g \tag{8.12}$$

hat für jedes $f \in C^0(0,1)$ genau dann eine eindeutige Lösung u, wenn das Problem $Lw = 0$, $Rw = 0$ nur die Nulllösung hat.

Beweis. Der Beweis der Eindeutigkeit basiert darauf, dass die Differenz zweier Lösungen $u - \tilde{u}$ das homogene Problem erfüllen muss. Der Beweis der Existenz kann analog zu Satz 7.8 realisiert werden. Für Details, siehe Satz 35.2 in H. Heuser: *Gewöhnliche Differentialgleichungen.* Teubner. 1989. □

Wir wollen die Lösungsvefahren nun für einige wichtige Spezialfälle diskutieren und für den allgemeinen Fall auf die Literatur verweisen.

Beispiel 8.21 (Gemischte Randbedingungen)**.** Für die Lösung des Randwertproblems, finde $u \in C^2(0,1) \cap C^1(0,1] \cap C[0,1]$ mit

$$\begin{aligned} -u''(x) &= f(x), \qquad x \in (0,1), \quad f \in C(0,1) \\ u(0) &= g_0, \\ u'(1) &= g_1 \end{aligned}$$

verwenden wir den Ansatz (7.5), wobei die Konstanten über $g_0 = u(0) = c_0$ und über $g_1 = u'(1) = -\int_0^1 f(s)\mathrm{d}s + c_1$ gegeben sind. Daher erhalten wir, dass die Lösung durch

$$u(x) = \int_0^1 G(x,y) f(y) \mathrm{d}y + g_0 + g_1 x$$

mit Green'scher Funktion

$$G(x,y) = \begin{cases} y & \text{für } y \le x \\ x & \text{für } y > x \end{cases}$$

gegeben ist. Der Konstruktion folgend erhalten wir wieder Existenz einer Lösung. Mit der Green'schen Funktion können wir wieder ein Stabilitätsresultat zeigen (vgl. Satz 7.3), allerdings mit anderer Konstante.

Wir diskretisieren das Problem analog zum Dirichletproblem, statt der Dirichlet-Randbedingung für $x = 1$ müssen wir jedoch die Neumann-Randbedingung

vorgeben. Wir näheren $u(1)$ durch den nach innen gerichteten Differenzenquotienten an:

$$g_1 = u'(1) = u'(x_{n+1}) \approx \tfrac{1}{h}(u(x_{n+1}) - u(x_n)).$$

Damit erhalten wir das Differenzenschema

$$\begin{aligned} -\tfrac{1}{h^2}(u_{i-1} - 2u_i + u_{i+1}) &= f(x_i), \qquad i = 1, \dots, n \\ u_0 &= g_0, \\ \tfrac{1}{h}(u_{n+1} - u_n) &= g_1, \end{aligned}$$

was in Matrix-Vektorschreibweise folgendermaßen lautet:

$$\frac{1}{h^2}\begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -h & h \end{pmatrix}\begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \\ u_{n+1} \end{pmatrix} = \begin{pmatrix} f(x_1) + h^{-2}g_0 \\ f(x_2) \\ \vdots \\ f(x_n) \\ g_1 \end{pmatrix}.$$

Um eine symmetrische Matrix zu erhalten, können wir die letzte Zeile durch h teilen und erhalten folgendes äquivalentes System:

$$\underbrace{\frac{1}{h^2}\begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{pmatrix}}_{A_h :=}\underbrace{\begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \\ u_{n+1} \end{pmatrix}}_{\underline{u}_h :=} = \underbrace{\begin{pmatrix} f(x_1) + h^{-2}g_0 \\ f(x_2) \\ \vdots \\ f(x_n) \\ h^{-1}g_1 \end{pmatrix}}_{\underline{f}_h :=}.$$

Für die Analyse führen wir die Räume $C^2_D(0,1) = \{v \in C^2(0,1) : v(x) = 0 \text{ für } x \in \Gamma_D\}$ und $V_{D,h} = \{v_h \in V_h : v_h(x) = 0 \text{ für } x \in \Gamma_D\}$ ein, wobei $\Gamma_D = \{0\}$. Um wieder eine eine Konsitenzordnung von 2 erhalten zu können, bietet es sich an, $L_h : V_{D,h} \to V_{D,h}$ und die Gitterfunktion $f_h \in V_{D,h}$ wie folgt zu definieren:

$$\begin{aligned} (L_h u_h)(x_i) &= h^{-2}(-u_{i-1} + 2u_i - u_{i+1}), & i = 1, \dots, n, \\ (L_h u_h)(x_{n+1}) &= h^{-1}(u_{n+1} - u_n), & \\ f_h(x_i) &= f(x_i), & i = 1, \dots, n, \\ f_h(x_{n+1}) &= g_1. & \end{aligned}$$

Diese Definition entspricht dem oben angesprochenen Gleichungssystem (und liefert daher dieselbe Lösung).

Man kann zeigen, dass die Lösung u_h des diskreten Problems durch

$$u_h = h \sum_{j=1}^{n+1} G_j f_h(x_j)$$

gegeben ist, wobei $G_j \in V_{D,h}$ mit

$$\begin{aligned} G_j(x_i) &= G(x_i, x_j) \text{ für } i = 1, \dots, n+1 \text{ und } j = 1, \dots, n \\ G_{n+1}(x_i) &= h^{-1} G(x_i, x_{n+1}) \text{ für } i = 1, \dots, n+1, \end{aligned}$$

vgl. Übung. Daraus lässt sich diskrete Stabilität ableiten, vgl. Übung:

$$\|u_h\|_{\infty,h} \leq C \|f_h\|_{\infty,h}.$$

Für $u \in C^3(0,1)$ können wir wieder Konsistenz mit Ordnung 1 zeigen:

$$\|\tau_h\|_{\infty,h} \leq Ch.$$

Für den Beweis müssen der Fall

$$\tau_h(x_i) = h^{-2}(-u(x_{i-1}) + 2u(x_i) - u(x_{i+1})) - f(x_i) \quad \text{mit} \quad i = 1, \dots, n$$

und der Fall

$$\tau_h(x_{n+1}) = h^{-1}(-u(x_n) + u(x_{n+1})) - g_1$$

gesondert behandelt werden, vgl. Übung.

Eine Konsistenzordnung von 2 lässt sich erzielen, wenn man nicht $f_h(x_{n+1}) = g_1$, sondern

$$f_h(x_{n+1}) = g_1 + \tfrac{h}{2} f(x_{n+1})$$

wählt, siehe Übung.

Aus diskreter Stabilität und Konstistenz folgt wieder Konvergenz.

Beispiel 8.22 (Reine Neumann-Randbedingungen)**.** Für die Lösung des Randwertproblems, finde $u \in C^2(0,1) \cap C^1[0,1]$ mit

$$\begin{aligned} -u''(x) &= f(x), \qquad x \in (0,1), \qquad f \in C(0,1) \\ -u'(0) &= g_0, \\ u'(1) &= g_1 \end{aligned}$$

verwenden wir den Ansatz (7.5). Die Ableitungen sind:

$$u'(x) = -\int_0^x f(s) \mathrm{d}s + c_1.$$

Durch Einsetzen der Randbedingungen erhalten wir

$$g_0 = -u'(0) = -c_1, \qquad g_1 = u'(1) = -\int_0^1 f(s)\mathrm{d}s + c_1.$$

Dieses System lässt sich nur dann auflösen, wenn die *Kompatibilitätsbedingung*

$$\int_0^1 f(s)\mathrm{d}s + g_0 + g_1 = 0 \tag{8.13}$$

erfüllt ist. Ist die Kompatibilitätsbedingung erfüllt, so stellen wir fest, dass für alle $c_0 \in \mathbb{R}$

$$u(x) = \int_0^1 G(x,y)f(y)\mathrm{d}y + g_0 x + c_0 \tag{8.14}$$

mit Green'scher Funktion

$$G(x,y) = \begin{cases} y & \text{für } y \le x \\ x & \text{für } y > x \end{cases}$$

eine Lösung ist. Da wir $c_0 \in \mathbb{R}$ frei wählen können, können wir keine Schranke für $\|u\|_\infty$ angeben.

Eindeutigkeit können wir erzielen, indem wir den konstanten Anteil durch eine Zusatzbedingung fixieren. Oftmals lautet die Zusatzbedingung

$$\int_0^1 u(x)\mathrm{d}x = 0,$$

im 1D-Fall können wir aber auch einfach den Funktionswert von u an einem beliebigen Punkt festlegen. Wenn wir „zufälligerweise" den Funktionswert am linken Rand vorgeben, also etwa $u(0) = 0$ setzen, dann sind wir wieder beim Fall des Beispiels 8.21. Gilt die Kompatibilitätsbedingung (8.13), dann erfüllt die Lösung von Beispiel 8.21 auch die Neumann-Randbedingung $-u'(0) = g_0$. Haben wir eine Lösung u berechnet, wissen wir, dass $\{u + c : c \in \mathbb{R}\}$ der gesamte Lösungsraum ist.

Bemerkung 8.23. Ganz analog zu den obigen Beispielen können Robin-Randbedingungen der Art

$$-u'(0) + \gamma_0 u(0) = g_0 \qquad \text{mit } \gamma_0 > 0 \text{ und } g_0 \text{ gegeben}$$

bzw.

$$u'(1) + \gamma_1 u(1) = g_1 \qquad \text{mit } \gamma_1 > 0 \text{ und } g_1 \text{ gegeben}$$

behandelt werden. Probleme mit Existenz und Eindeutigkeit wie im Beispiel 8.22 ergeben sich nicht.

Auch der Fall $c > 0$ kann ganz analog behandelt werden; auch in diesem Fall ergeben sich keine Probleme mit Existenz und Eindeutigkeit.

Numerisches Experiment

Beispiel 8.24. In diesem Beispiel rechnen wir das Dirichletproblem

$$-u''(x) = -(3\pi)^2 \sin(3\pi x), \qquad x \in (0,1), \qquad u(0) = u(1) = 0.$$

Die exakte Lösung ist offensichtlich $u(x) = \sin(3\pi x)$. Damit können wir die Fehler exakt bestimmen. Wir nutzen die Verfahren, die wir kennengelernt haben. Die Fehler (und Raten) in der diskreten Supremumsnorm (für die wir die Theorie gemacht haben) sind in Tabelle 8.1 angegeben. Die Faktoren e_{2h}/e_h erreichen den Wert $4 = \frac{(2h)^2}{h^2}$, was bedeutet, dass die Konvergenzordnung 2 exakt erreicht wird.

Wir erinnern uns, dass $\|u - u_h\|_{\infty,h} = \sup_{x_i \in \mathcal{T}_h} |u(x_i) - u_i|$. Wenn wir die errechnete Lösung via Polygonzug interpretieren, ist sie nicht nur am Gitter, sondern auf ganz $\Omega = (0,1)$ definiert. Wenn wir den Fehler in der Supremumsnorm $\|u - u_h\|_\infty = \sup_{x \in \Omega} |u(x) - u_h(x)|$ bestimmen, dann erhalten wir $\|u - u_h\|_{\infty,h} \leq \|u - u_h\|_\infty$. Für das Modellproblem sieht man einen Unterschied erst nach der zweiten Nachkommastelle (damit gilt auch hier das Resultat von Tabelle 8.1).

h	Fehler e_h	Faktor e_{2h}/e_h
2^{-2}	$6.26\ 10^{-1}$	
2^{-3}	$1.24\ 10^{-1}$	5.04
2^{-4}	$2.94\ 10^{-2}$	4.22
2^{-5}	$7.26\ 10^{-3}$	4.05
2^{-6}	$1.81\ 10^{-3}$	4.01
2^{-7}	$4.52\ 10^{-4}$	4.00
2^{-8}	$1.13\ 10^{-4}$	4.00

Tabelle 8.1.: Fehler in der diskreten Supremumsnorm $e_h = \|u - u_h\|_{\infty,h}$.

9. Singulär gestörte Probleme

Die Schranke $h < 2/\|b\|_\infty$ aus Satz 8.12 bzw. $h < 1/\|b\|_\infty$ aus Bemerkung 8.18 zeigt, dass für große Werte von b eine sehr kleine Schrittweite gewählt werden muss.

Modellproblem

Wir behandeln zuerst das Modellproblem

$$-\epsilon u''(x) + u'(x) = 1, \qquad x \in (0,1), \tag{9.1}$$
$$u(0) = 0 \tag{9.2}$$
$$u(1) = 0, \tag{9.3}$$

mit einem Parameter $\epsilon > 0$. Nach Division durch ϵ wird dieses Modellproblem zu einem Problem der Art (8.1) – (8.3) mit $\|b\|_\infty = \epsilon^{-1}$.

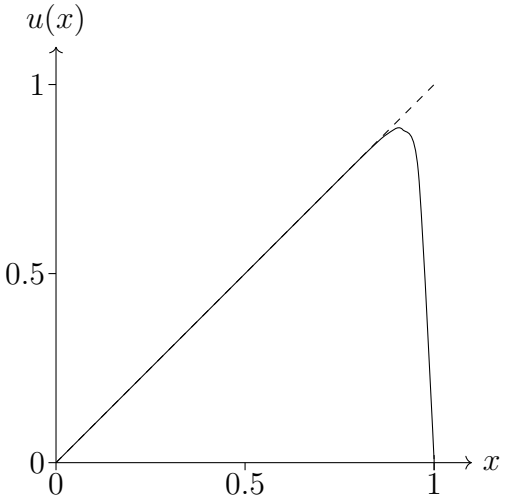


Abbildung 9.1.: Funktion $u(x)$ für $\epsilon = 0.025$

Bevor wir die Lösung des Modellproblems diskutieren, stellen wir fest, dass wir für den Fall $\epsilon = 0$ ein Problem erster Ordnung erhalten, für das wir nur eine

einzige Zusatzbedingung benötigen würden, etwa

$$v'(x) = 1, \qquad v(0) = 0, \tag{9.4}$$

was zur Lösung $v(x) = x$ führt. Würden wir für v zusätzlich $v(1) = 0$ fordern, dann wäre das Problem überbestimmt. Da das Modellproblem aber ein Problem zweiter Ordnung ist, benötigen wir für dieses natürlich 2 Randbedingungen, wie vorgegeben. Das Modellproblem ist noch einfach genug, sodass wir die Lösung analytisch bestimmen können:

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

Für $\epsilon > 0$ klein, vgl. Abbildung 9.1, verhält sich die Lösung $u(x)$ im Wesentlichen wie die Lösung von (9.4), also $u(x) \approx v(x) = x$. Nur nahe des rechten Randes muss die Lösungsfunktion noch die Kurve kratzen, um die geforderte Randbedingung zu erfüllen. Für kleines $\epsilon > 0$ bezeichnen wir dieses Problem als singulär gestört, da die zwar kleine Störung der Differentialgleichung doch die Art des Problems qualitativ ändert.

Nach der behandelten Theorie müssten wir $h < 2\epsilon$ wählen; diese Wahl garantiert eine gute Auflösung der Lösung an der Grenzschicht.

Wenn wir den zentralen Differenzenquotienten zur Diskretisierung verwenden, lautet die Steifigkeitsmatrix

$$A_h = \begin{pmatrix} a_{1,1} & a_{1,2} & & & \\ a_{2,1} & a_{2,2} & a_{2,3} & & \\ & \ddots & \ddots & \ddots & \\ & & a_{n-1,n-2} & a_{n-1,n-1} & a_{n-1,n} \\ & & & a_{n,n-1} & a_{n,n} \end{pmatrix}$$

mit

$$a_{i,i} = \frac{2\epsilon}{h^2}, \quad a_{i,i-1} = -\frac{\epsilon}{h^2} - \frac{1}{2h}, \quad a_{i,i+1} = -\frac{\epsilon}{h^2} + \frac{1}{2h}.$$

Um die Voraussetzungen für das Maximumsprinzip (Satz 8.1) zu erhalten, benötigen wir $a_{i,i} > 0$ sowie $a_{i,i-1} < 0$ und $a_{i,i+1} < 0$, was $h < 2\epsilon$ voraussetzt.

Löst man nun das Modellproblem mit zentralem Differenzenquotienten für die ersten Ableitungen und moderatem h, erhält man eine oszillierende Näherungslösung nahe der Grenzschicht. (Die Oszillationen verschwinden für sehr kleine h, wie von der Theorie vorhergesagt.)

Wenn wir den Vorwärtsdifferenzenquotienten wählen, haben wir

$$a_{i,i} = \frac{2\epsilon}{h^2} - \frac{1}{h}, \quad a_{i,i-1} = -\frac{\epsilon}{h^2}, \quad a_{i,i+1} = -\frac{\epsilon}{h^2} + \frac{1}{h}.$$

Um die Voraussetzungen für Maximumsprinzip zu erhalten, benötigen wir $a_{i,i} > 0$ sowie $a_{i,i-1} < 0$ und $a_{i,i+1} < 0$, was $h < \frac{\epsilon}{2}$ voraussetzt.

Wenn wir den Rückwärtsdifferenzenquotienten wählen, haben wir

$$a_{i,i} = \frac{2\epsilon}{h^2} + \frac{1}{h}, \quad a_{i,i-1} = -\frac{\epsilon}{h^2} - \frac{1}{h}, \quad a_{i,i+1} = -\frac{\epsilon}{h^2},$$

wo die Bedingungen $a_{i,i} > 0$ sowie $a_{i,i-1} < 0$ und $a_{i,i+1} < 0$ stets erfüllt sind. Somit erhalten wir das diskrete Maximumsprinzip. Wegen $L_h u_h = 1 > 0$ gilt es natürlich als Minimumsprinzip; da wir es auch lokal anwenden können, haben wir $u_k \geq \min\{u_{k-1}, u_{k+1}\}$ für alle $k = 1, \ldots, n$, was Oszillationen ausschließt. Wir haben auch diskrete Stabilität *für alle Gitterweiten*; Satz 8.11 gibt eine in ϵ nicht scharfe Abschätzung:

$$\|L_h^{-1}\|_{\infty,h} = \epsilon^{-1}\|(\epsilon^{-1}L_h)^{-1}\|_{\infty,h} \leq \epsilon^{-1}(1+\epsilon^{-1}h)^{n+1} \leq \epsilon^{-1}e^{h(n+1)/\epsilon} \leq e^{2/\epsilon}.$$

Das bedeutet, dass das diskretisierte Problem (ohne weitere Voraussetzungen an h und ϵ!) eindeutig lösbar ist. Wie im letzten Kapitel kann man wieder Konvergenz zeigen. Wegen Bemerkung 8.18 ist sie auf Ordnung 1 beschränkt.

Upwind-Verfahren

Nun wollen wir diese Aussagen noch für den allgemeinen Fall (8.1) – (8.3) formulieren. Wir betrachten also wieder

$$\begin{aligned} -u''(x) + b(x)u'(x) + c(x)u(x) &= f(x), \qquad x \in (0,1), \\ u(0) &= 0 \\ u(1) &= 0. \end{aligned}$$

Hier haben wir bei Verwendung des Rückwärtsdifferenzenquotienten nun

$$a_{i,i} = \frac{2}{h^2} + \frac{b(x_i)}{h} + c(x_i), \quad a_{i,i-1} = -\frac{1}{h^2} - \frac{b(x_i)}{h}, \quad a_{i,i+1} = -\frac{1}{h^2}.$$

Wenn $b(x_i) \geq 0$, dann sind bei dieser Wahl die Bedingungen $a_{i,i} > 0$ sowie $a_{i,i-1} < 0$ und $a_{i,i+1} < 0$ und die Voraussetzungen für das diskrete Maximumsprinzip ohne Voraussetzungen an h erfüllt. Bei $b(x_i) < 0$ sieht die Sache aber schon wieder anderes aus!

Bei Verwendung des Vorwärtsdifferenzenquotienten erhalten wir

$$a_{i,i} = \frac{2}{h^2} - \frac{b(x_i)}{h} + c(x_i), \quad a_{i,i-1} = -\frac{1}{h^2}, \quad a_{i,i+1} = -\frac{1}{h^2} + \frac{b(x_i)}{h}.$$

Hier sind für den Fall $b(x_i) \leq 0$, die Bedingungen $a_{i,i} > 0$ sowie $a_{i,i-1} < 0$ und $a_{i,i+1} < 0$ und die Voraussetzungen für das diskrete Maximumsprinzip ohne Voraussetzungen an h erfüllt.

Die Idee des Upwind-Verfahrens ist es nun, die Diskretisierung auf Basis einer Fallunterscheidung zu definieren: Wir wählen also den Rückwärtsdifferenzenquotienten bei $b(x_i) > 0$ und den Vorwärtsdifferenzenquotienten bei $b(x_i) < 0$.

Wir wählen also

$$a_{i,i} = \frac{2}{h^2} + \frac{|b(x_i)|}{h} + c(x_i), \quad a_{i,i-1} = -\frac{1}{h^2} - \frac{b^+(x_i)}{h}, \quad a_{i,i+1} = -\frac{1}{h^2} + \frac{b^-(x_i)}{h},$$

wobei

$$b^+(x_i) := \max\{0, b(x_i)\} \quad \text{und} \quad b^-(x_i) := \min\{0, b(x_i)\}.$$

Die Stabilität des Upwind-Verfahrens ist analog zur Stabilität des impliziten Euler-Verfahrens, wobei für $b > 0$ eine Zeitrichtung von links nach rechts, für $b < 0$ jedoch eine Zeitrichtung von rechts nach links angenommen wird. Das Upwind-Verfahren lässt sich auch physikalisch interpretieren: der Konvektionsterm $b(x)u'(x)$ modelliert die Wirkungen einer Strömung bzw. eines Windes. Die Diskretisierung wird dabei so gewählt dass die Diskretisierungsrichtung der Windrichtung entgegengesetzt ist („gegen den Wind" bzw. „upwind").

Anzumerken bleibt, dass das Upwind-Verfahren zwar die gewünschten Stabilitätseigenschaften (und damit auch etwa wieder ein diskretes Maximumsprinzip) aufweist, jedoch nur eine Konsistenzordnung von 1 und damit auch nur eine Konvergenzordnung von 1 aufweist, die Konvergenz also langsamer ist, als wir dies für Probleme mit $b = 0$ oder betragsmäßig kleinem b (und Verwendung des zentralen Differenzenquotienten) erwarten würden, vgl. Bemerkung 8.18.

Wegen $b^+ = (b+|b|)/2$ und $b^- = (b-|b|)/2$ können wir mit $\alpha(x) := 1 + \frac{h}{2}|b(x)|$ das Upwind-Verfahren auch als

$$a_{i,i} = \frac{2\alpha(x_i)}{h^2} + c(x_i), \quad a_{i,i-1} = -\frac{\alpha(x_i)}{h^2} - \frac{b(x_i)}{2h}, \quad a_{i,i+1} = -\frac{\alpha(x_i)}{h^2} + \frac{b(x_i)}{2h}, \tag{9.5}$$

schreiben, was der Verwendung des zentralen Differenzenquotienten für die erste Ableitung entspricht. Die Wahl $\alpha(x) > 1$ ist jedoch eine Form der Stabilisierung, die hier durch eine Verstärkung des Diffusionsterms ausgedrückt wird. Der Differentialoperator L_h mit Upwind-Diskretisierung kann also auch als Diskretisierung des Operators

$$\widetilde{L}u(x) = -(1 + \tfrac{h}{2}|b(x)|)u''(x) + b(x)u'(x) + c(x)u(x)$$

interpretiert werden.

Wendet man nun (9.5), jedoch mit der Wahl $\alpha(x) = \frac{h}{2}|b(x)| \coth\left(\frac{h}{2}|b(x)|\right)$, an, so erhält man das deutlich bessere Approximationseigenschaften zeigende Iljin-Verfahren; siehe etwa Kapitel 85 in Martin Hanke-Borgeois: *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens*. 2002.

10. Differenzenverfahren in mehreren Ortsdimensionen

Wir wollen nun ein Randwertproblem auf einem beschränkten und zusammenhängenden Gebiet $\Omega \subset \mathbb{R}^d$ mit $d = 2$ und Rand $\partial\Omega$ lösen. Wir nehmen, wie bisher, an, dass das Rechengebiet Ω eine offene Menge mit Abschluss $\overline{\Omega} = \Omega \cup \partial\Omega$ ist. Wir suchen nach einer hinreichend glatten Funktion, etwa $u \in C^2(\Omega) \cap C(\overline{\Omega})$, sodass

$$\underbrace{-\partial_{xx}u - \partial_{yy}u}_{-\Delta u :=} + cu = f \quad \text{in} \quad \Omega. \tag{10.1}$$

Wie in den letzten Kapiteln nehmen wir an, dass $c \geq 0$. Wir müssen außerdem Randbedingungen setzen; wir beschränken uns dabei auf Dirichlet-Bedingungen:

$$u = g \quad \text{auf} \quad \partial\Omega.$$

Wir schreiben das Problem wieder in folgender Form: Finde $u \in C_0^2(\Omega) := C^2(\Omega) \cap \{u \in C(\overline{\Omega}) : u|_{\partial\Omega} = 0\}$, sodass

$$Lu := -\Delta u + cu = f.$$

Die Analyse der Existenz und Eindeutigkeit von Lösungen zu diesem Problem ist im Gegensatz zum eindimensionalen Fall hier keineswegs einfach; wir werden später nochmals darauf zu sprechen kommen, jetzt aber von der Existenz einer Lösung ausgehen und nur kurz Fragen zur Diskretisierung mit Differenzenschemata ansprechen.

Diskretisierung

Im Folgenden wollen wir das Modellproblem wieder mit einem Differenzenschema diskretisieren. Wir nehmen zunächst an, dass das Rechengebiet das Einheitsquadrat ist:

$$\Omega = (0, 1)^2.$$

Wir führen nun auf dem Einheitsquadrat ein regelmäßiges Gitter mit Gitterweite h (mit $h^{-1} = n + 1 \in \mathbb{N}$) ein:

$$\mathcal{T}_h := \bar{\mathcal{T}}_h \cap \Omega \qquad \text{mit} \qquad \bar{\mathcal{T}}_h := \{\mathbf{x}_{i,j} = (ih, jh) : i, j = 0, \ldots, n + 1\}$$

Die diskrete Lösung u_h ist durch ihre Funktionswerte auf den Gitterpunkten $\mathbf{x}_{i,j}$ bestimmt:

$$u_{i,j} = u_h(\mathbf{x}_{i,j}).$$

Die Dirichlet-Randbedingungen geben wir auf den am Rand liegenden Gitterpunkten vor:

$$u_{i,j} = g(\mathbf{x}_{i,j}) \qquad \forall \mathbf{x}_{i,j} \in \partial\mathcal{T}_h := \bar{\mathcal{T}}_h \cap \partial\Omega. \tag{10.2}$$

Für die Diskretisierung von $-\Delta = -\partial_{xx} - \partial_{yy}$ wählen wir den bewährten zentralen Differenzenquotienten:

$$\begin{aligned} -\Delta u(\mathbf{x}_{i,j}) \approx \tfrac{1}{h^2}\big(& \qquad\qquad - u_h(\mathbf{x}_{i,j-1}) \\ & - u_h(\mathbf{x}_{i-1,j}) + 4u_h(\mathbf{x}_{i,j}) - u_h(\mathbf{x}_{i+1,j}) \\ & \qquad\qquad - u_h(\mathbf{x}_{i,j+1}) \qquad\qquad \big). \end{aligned}$$

Wir verwenden für die Diskretisierung des Laplace-Operators also 5 Gittterpunkte; daher spricht man auch vom 5-*point stencil*.

Durch Einsetzen in die Differentialgleichung erhalten wir

$$\tfrac{1}{h^2}\big(4u_{i,j} - u_{i-1,j} - u_{i+1,j} - u_{i,j-1} - u_{i,j+1}\big) + c_{i,j}u_{i,j} = f_{i,j} \quad \forall \mathbf{x}_{i,j} \in \mathcal{T}_h, \tag{10.3}$$

wobei $c_{i,j} := c(\mathbf{x}_{i,j})$ und $f_{i,j} := f(\mathbf{x}_{i,j})$.

Um die Steifigkeitsmatrix und den Lastvektor aufstellen zu können, müssen wir die Gitterpunkte durchnummerieren. Dafür verwenden wir eine lexikografische Nummerierung; es wird dem

$$\text{dem Gitterpunkt } \mathbf{x}_{i,j} \text{ der Index } \nu_{i,j} := (i-1) * n + j \text{ zugeordnet,} \tag{10.4}$$

wobei $i, j = 1, \ldots, n$. Die Steifigkeitsmatrix A_h und der Lastvektor $\underline{f}_h$ ergeben sich wieder aus der Darstellung der Bedingungen (10.2) und (10.3); für $n = 4$ und $c = 0$ haben wir etwa:

$$\frac{1}{h^2}\begin{pmatrix}
4 & -1 & & & -1 & & & & & & & & & & & \\
-1 & 4 & -1 & & & -1 & & & & & & & & & & \\
 & -1 & 4 & -1 & & & -1 & & & & & & & & & \\
 & & -1 & 4 & & & & -1 & & & & & & & & \\
\hline
-1 & & & & 4 & -1 & & & -1 & & & & & & & \\
 & -1 & & & -1 & 4 & -1 & & & -1 & & & & & & \\
 & & -1 & & & -1 & 4 & -1 & & & -1 & & & & & \\
 & & & -1 & & & -1 & 4 & & & & -1 & & & & \\
\hline
 & & & & -1 & & & & 4 & -1 & & & -1 & & & \\
 & & & & & -1 & & & -1 & 4 & -1 & & & -1 & & \\
 & & & & & & -1 & & & -1 & 4 & -1 & & & -1 & \\
 & & & & & & & -1 & & & -1 & 4 & & & & -1 \\
\hline
 & & & & & & & & -1 & & & & 4 & -1 & & \\
 & & & & & & & & & -1 & & & -1 & 4 & -1 & \\
 & & & & & & & & & & -1 & & & -1 & 4 & -1 \\
 & & & & & & & & & & & -1 & & & -1 & 4
\end{pmatrix}
\begin{pmatrix} u_{1,1} \\ u_{1,2} \\ u_{1,3} \\ u_{1,4} \\ u_{2,1} \\ u_{2,2} \\ u_{2,3} \\ u_{2,4} \\ u_{3,1} \\ u_{3,2} \\ u_{3,3} \\ u_{3,4} \\ u_{4,1} \\ u_{4,2} \\ u_{4,3} \\ u_{4,4} \end{pmatrix}
=
\begin{pmatrix} f_{1,1} \\ f_{1,2} \\ f_{1,3} \\ f_{1,4} \\ f_{2,1} \\ f_{2,2} \\ f_{2,3} \\ f_{2,4} \\ f_{3,1} \\ f_{3,2} \\ f_{3,3} \\ f_{3,4} \\ f_{4,1} \\ f_{4,2} \\ f_{4,3} \\ f_{4,4} \end{pmatrix}$$

Wenn $c \neq 0$, dann sind natürlich die entsprechenden Diagonaleinträge zu addieren. Die Steifigkeitsmatrix hat keine tridiagonale Struktur mehr; in jeder Zeile i gibt es nun in bis zu 5 Spalten nicht-verschwindende Einträge:

$$i, \quad i-1, \quad i+1, \quad i-(n+2), \quad i+(n+2). \tag{10.5}$$

Konvergenz

Wir definieren den Operator L_h über

$$L_h u(x,y) := D_h^2 u(x,y) + c(x,y)u(x,y),$$
$$D_h^2 u(x,y) = \tfrac{1}{h^2}\big(4u(x,y) - u(x-h,y) - u(x+h,y) - u(x,y-h) - u(x,y+h)\big).$$

Sei nun $V_{0,h}$ der Raum der (auf $\bar{\mathcal{T}}_h$ definierten) Gitterfunktionen, die die Dirichlet-Randbedingungen erfüllen. Dann lässt sich das Differenzenschema schreiben als:

$$\text{Gesucht } u_h \in V_{0,h}: \qquad L_h u_h = f \qquad \text{auf } \mathcal{T}_h.$$

Satz 10.1. *Wenn $u \in C^4(\Omega)$, ist das Verfahren konsistent mit Ordnung 2, es gibt also eine Konstante $C > 0$ mit*

$$\|L_h u - f\|_{\infty,h} = \max_{\mathbf{x}\in\mathcal{T}_h} |L_h u(\mathbf{x}) - f(\mathbf{x})| \le Ch^2,$$

Wenn $u \in C^3(\Omega)$, ist es konsistent mit Ordnung 1.

Beweis. Wir haben $\|L_h u - f\|_{\infty,h} = \|L_h u - Lu\|_{\infty,h} = \|D_h^2 u - u''\|_{\infty,h}$, was man mit Taylor-Entwicklung abschätzen kann. □

Nun wenden wir uns der Stabilität zu. Dazu zeigen wir zuerst wieder ein diskretes Maximumsprinzip, das bei Beschränkung auf Dirichlet-Randbedingungen wie folgt lautet:

Satz 10.2 (Diskretes Maximumsprinzip)**.** *Für $u_h \in V_{0,h}$ mit*

$$L_h u_h(\mathbf{x}_{i,j}) \le 0 \quad \forall \mathbf{x}_{i,j} \in \mathcal{T}_h,$$

gilt $u_h(\mathbf{x}_{i,j}) \le 0$ für alle $\mathbf{x}_{i,j} \in \mathcal{T}_h$.

Beweis. Wir nehmen an, dass $u_{i,j} \ge \max_{m,n} u_{m,n}$ mit $u_{i,j} \ge 0$ ein nicht-negatives Maximum (im Inneren) wäre. Dann gilt

$$\begin{aligned} 0 &\ge h^2 L_h u_h(\mathbf{x}_{i,j}) \\ &= \underbrace{(u_{i,j} - u_{i-1,j})}_{\ge 0} + \underbrace{(u_{i,j} - u_{i+1,j})}_{\ge 0} + \underbrace{(u_{i,j} - u_{i,j-1})}_{\ge 0} + \underbrace{(u_{i,j} - u_{i,j+1})}_{\ge 0} + \underbrace{h^2 c(\mathbf{x}_{i,j}) u_{i,j}}_{\ge 0}. \end{aligned}$$

Daraus folgt nun $u_{i,j} = u_{i-1,j} = u_{i+1,j} = u_{i,j-1} = u_{i,j+1}$. Durch wiederholte Anwendung dieses Arguments erhalten wir nun, dass u_h konstant sein muss, woraus $u_h \le 0$ folgt. □

Satz 10.3 (Existenz und Eindeutigkeit)**.** *Das Differenzenschema hat eine eindeutlige Lösung.*

Beweis. Aus dem Maximumsprinzip erhalten wir wieder – wie in Kapitel 8 – die Eindeutigkeit der Lösung, woraus unmittelbar die Invertierbarkeit der Steifigkeitsmatrix A_h folgt. Daraus folgt auch Existenz einer Lösung. □

Satz 10.4. *L_h ist für alle Gitterweiten stabil im Sinne von Definition 8.6.*

Beweis. Sei L_h^{-1} wieder der Lösungsoperator. Da das diskrete Maximumsprinzip gilt, können wir wieder Lemma 8.9 anwenden; daher ist $\|L_h^{-1}\|_{\infty,h} = \|L_h^{-1}\mathbf{1}\|_{\infty,h}$. Wir konstruieren eine Funktion $v_h \in V_{0,h}$ mit $L_h v_h \geq \mathbf{1}$. Wir wählen $v_h(x,y) = \frac{1}{2}x(1-x)$ für $(x,y) \in \mathcal{T}_h$ und erhalten mit

$$\begin{aligned}
&\tfrac{1}{h^2}\big(2v(\mathbf{x}_{i,j}) - v_h(\mathbf{x}_{i-1,j}) - v_h(\mathbf{x}_{i+1,j})\big) = 1, \\
&\tfrac{1}{h^2}\big(2v(\mathbf{x}_{i,j}) - v_h(\mathbf{x}_{i,j-1}) - v_h(\mathbf{x}_{i,j+1})\big) = 0, && j \in \{2,\dots,n-1\}, \\
&\tfrac{1}{h^2}\big(2v(\mathbf{x}_{i,j}) - v_h(\mathbf{x}_{i,j-1}) - v_h(\mathbf{x}_{i,j+1})\big) = \tfrac{1}{2h^2}x_i(1-x_i) \geq 0, && j \in \{1,n\},
\end{aligned}$$

dass $L_h v_h \geq \mathbf{1} + cv_h \geq \mathbf{1}$. Mit dem diskreten Maximumprinzip (Monotonie) folgt

$$\|L_h^{-1}\|_{\infty,h} = \|L_h^{-1}\mathbf{1}\|_{\infty,h} \leq \|v_h\|_{\infty,h} \leq \tfrac{1}{8},$$

wobei wir für die letzte Abschätzung Satz 8.8 verwendet haben. □

Wegen Satz 8.16 erhalten wir mit mit Konsistenz (Satz 10.1) und diskreter Stabilität (Satz 10.4) Konvergenz:

Satz 10.5 (Konvergenz)**.** *Wenn $u \in C^4(0,1)$, dann ist das Verfahren konvergent mit Ordnung 2, es gibt also eine Konstante $C > 0$ mit*

$$\|u - u_h\|_{\infty,h} \leq Ch^2.$$

Wenn $u \in C^3(\Omega)$, dann ist das Verfahren konvergent mit Ordnung 1.

Erweiterungen

Eine Erweiterung auf 3D ist kein Problem.

Behandlung allgemeiner Gebiete

Grundsätzlich lässt sich die besprochene Methode auch auf allgemeinere Gebiete verallgemeinern; wir wollen da eine etwas naive Herangehensweise diskutieren.

Ohne Beschränkung der Allgemeinheit nehmen wir an, dass das Rechengebiet Ω in das Einheitsquadrat eingebettet werden kann, also $\Omega \subseteq (0,1)^2$. Wir führen ohne Beachtung der Form von Ω für das gesamte Gebiet $[0,1]^2$ ein regelmäßiges Gitter ein:

$$\bar{\mathcal{T}}_h = \{(ih, jh) : i, j = 0, \ldots, n+1\}.$$

Die Gitterpunkte im Gebiet Ω bezeichnen wir mit Ω_h:

$$\mathcal{T}_h := \Omega \cap \bar{\mathcal{T}}_h.$$

Die Dirichlet-Randbedingungen setzen wir nun für die Gitterpunkte, die außerhalb der (offenen) Menge Ω liegen. Da für die Berechnung nur jene Gitterpunkte eine Rolle spielen, die Nachbarn eines Gitterpunktes in $\mathcal{T}_h$ sind, definieren wir

$$\partial\mathcal{T}_h := \{\mathbf{x}_{i,j} \in \bar{\mathcal{T}}_h : \mathbf{x}_{i,j} \notin \mathcal{T}_h, \{\mathbf{x}_{i-1,j}, \mathbf{x}_{i+1,j}, \mathbf{x}_{i,j-1}, \mathbf{x}_{i,j+1}\} \cap \mathcal{T}_h \neq \emptyset\}.$$

Je nach gewähltem Gebiet kann $\partial\mathcal{T}_h \subset \partial\Omega$ gelten oder eben auch nicht.

Indem wir nun auf $\mathcal{T}_h$ die Erfüllung der diskretisierten Differentialgleichung und auf $\partial\mathcal{T}_h$ die homogenen Dirichlet-Randbedingungen fordern, erhalten wir folgendes System von Gleichungen:

$$\begin{aligned} \tfrac{1}{h^2}\big(4u_{i,j} - u_{i-1,j} - u_{i+1,j} - u_{i,j-1} - u_{i,j+1}\big) + c(\mathbf{x}_{i,j})u_{i,j} &= f(\mathbf{x}_{i,j}) \quad \text{für} \quad x_{i,j} \in \mathcal{T}_h \\ u_{i,j} &= 0 \qquad\quad \text{für} \quad x_{i,j} \in \partial\mathcal{T}_h. \end{aligned}$$

Dieses System kann – nach Durchnummerieren aller Gitterpunkte in $\mathcal{T}_h$ – wieder als lineares Gleichungssystem geschrieben werden.

Bemerkung 10.6 (Inhomogene Dirichlet-Randbedingungen)**.** Für die Realisierung der inhomogenen Dirichlet-Randbedingungen

$$u(\mathbf{x}) = g_D(\mathbf{x}), \qquad \mathbf{x} \in \partial\Omega$$

müssten wir nun Funktionswerte für die Gitterpunkte $\mathbf{x}_{i,j} \in \partial\mathcal{T}_h$ vorgeben. Wenn $\partial\mathcal{T}_h \subset \partial\Omega$, dann ist das kein Problem.

Wenn dies jedoch nicht zutrifft, muss man mehr arbeiten, denn g_D ist im Allgemeinen nur auf $\partial\Omega$ definiert. Daher müssten wir die Randdaten irgendwie auf $\partial\mathcal{T}_h$ projizieren.

Beispiel 10.7. Wenn das Rechengebiet Ω etwa das L-förmige Gebiet

$$\Omega = (0,1)^2 \backslash [\tfrac{1}{2}, 1)^2$$

ist und wir sicherstellen, dass $n+1 = h^{-1}$ gerade ist, dann gilt $\partial \mathcal{T}_h \subset \partial\Omega$; siehe Abbildung 10.1.

Wir lösen nun ein Randwertproblem auf Ω. Als Modellproblem wählen wir

$$\begin{aligned} -\Delta u(x,y) &= f(x,y) := 8\pi^2 \sin(2\pi x)\sin(2\pi y) \qquad \text{für } (x,y) \in \Omega \\ u(x,y) &= 0 \qquad \text{für } (x,y) \in \partial\Omega. \end{aligned}$$

Es lässt sich leicht nachprüfen, dass

$$u(x,y) = 2\pi^2 \sin(2\pi x)\sin(2\pi y)$$

eine Lösung ist; siehe Abbildung 10.2 für eine Berechnung mit $h = 2^{-6}$.

Der Fehler in der Maximumsnorm $e_h = \|u - u_h\|_{\infty,h}$ verhält sich entsprechend Tabelle 10.1, woraus abgelesen werden kann, dass das Verfahren wie $\mathcal{O}(h^2)$ konvergiert. Die im Experiment erhaltene Rate stimmt somit exakt mit der oberen Schranke aus der Theorie überein.

Beispiel 10.8. Wenn wir als Rechengebiet Ω einen Kreis wählen, also

$$\Omega = \{\mathbf{x} \in \mathbb{R}^2 : \|\mathbf{x} - (\tfrac{1}{2}, \tfrac{1}{2})\|_{\ell^2} < \tfrac{1}{2}\}$$

dann ist im Allgemeinen $\partial \mathcal{T}_h \not\subset \partial\Omega$; siehe Abbildung 10.3.

Wir lösen nun ein Randwertproblem auf Ω. Als Modellproblem wählen wir

$$\begin{aligned} -\Delta u(x,y) &= f(x,y) := 16 \qquad \text{für } (x,y) \in \Omega \\ u(x,y) &= 0 \qquad \text{für } (x,y) \in \partial\Omega. \end{aligned}$$

Es lässt sich leicht nachprüfen, dass

$$u(x,y) = 1 - 4\left(\left(x - \tfrac{1}{2}\right)^2 + \left(y - \tfrac{1}{2}\right)^2\right)$$

eine Lösung ist; siehe Abbildung 10.4 für eine Berechnung mit $h = 2^{-6}$.

Der Fehler in der Maximumsnorm $e_h = \|u - u_h\|_{\infty,h}$ verhält sich entsprechend Tabelle 10.2, woraus abgelesen werden kann, dass zwar Konvergenz vorliegt, die aber langsamer ist, als $\mathcal{O}(h)$. Dies liegt daran, dass die besprochene Methode das Rechengebiet Ω relativ schlecht approximiert. (Da das Gebiet nicht exakt dargestellt wird, ist das Konvergenzresultat aus Satz 10.5 auf dieses Beispiel *nicht* anwendbar.) Eine Alternative, die sich hier als nützlich erweist, ist die Finite Elemente Methode (FEM).

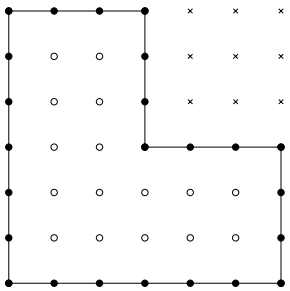


Abbildung 10.1.: Gitterpunkte in $\mathcal{T}_h$ sind nicht-gefüllte Kreise, Gitterpunkte in $\partial\mathcal{T}_h$ gefüllte Kreise und die ignorierten Gitterpunkte sind Kreuze.

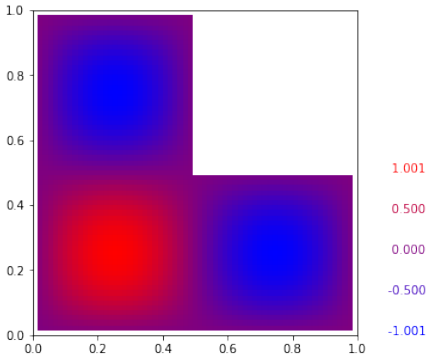


Abbildung 10.2.: Lösung am L-förmigen Gebiet.

h	Fehler e_h	Faktor e_{2h}/e_h
2^{-2}	$2.34\ 10^{-1}$	
2^{-3}	$5.30\ 10^{-2}$	4.41
2^{-4}	$1.30\ 10^{-2}$	4.09
2^{-5}	$3.22\ 10^{-3}$	4.02
2^{-6}	$8.04\ 10^{-4}$	4.01
2^{-7}	$2.01\ 10^{-4}$	4.00
2^{-8}	$5.02\ 10^{-5}$	4.00
2^{-9}	$1.25\ 10^{-5}$	4.00
2^{-10}	$3.14\ 10^{-6}$	4.00

Tabelle 10.1.: Konvergenz am L-förmigen Gebiet.

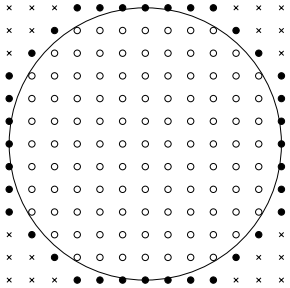


Abbildung 10.3.: Gitterpunkte in $\mathcal{T}_h$ sind nicht-gefüllte Kreise, Gitterpunkte in $\partial\mathcal{T}_h$ gefüllte Kreise und die ignorierten Gitterpunkte sind Kreuze.

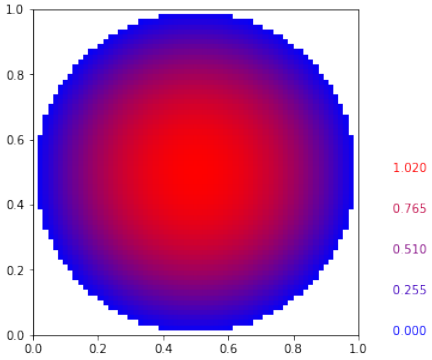


Abbildung 10.4.: Lösung am Kreis.

h	Fehler e_h	Faktor e_{2h}/e_h
2^{-2}	$1.88\ 10^{-1}$	
2^{-3}	$1.51\ 10^{-1}$	1.24
2^{-4}	$1.09\ 10^{-1}$	1.38
2^{-5}	$5.33\ 10^{-2}$	2.05
2^{-6}	$3.28\ 10^{-2}$	1.63
2^{-7}	$1.79\ 10^{-2}$	1.83
2^{-8}	$8.57\ 10^{-3}$	2.08
2^{-9}	$5.03\ 10^{-3}$	1.70
2^{-10}	$2.65\ 10^{-3}$	1.90

Tabelle 10.2.: Konvergenz am Kreis.

Neumann- und Robin-Randbedingungen

In zwei- und mehrdimensionalen Gebieten werden Neumann- und Robin-Randbedingungen die Ableitungen in Normalenrichtung vorgegeben:

$$-\partial_n u = g_N \qquad \text{bzw.} \qquad \partial_n u + \gamma u = g_R$$

mit $\gamma \geq 0$. Solange der Rand achsenparallel ist, ist die Diskretisierung dieser Randbedingungen analog zum eindimensionalen Fall möglich. Für alle anderen Fälle ist die Realisierung von Neumann- oder Robin-Randbedingungen nicht so einfach; auch hier wird die Finite Elemente Methode eine Alternative sein.

Bemerkung 10.9 (Reines Neumannproblem)**.** Wenn $c = 0$ und am ganzen Rand Neumann-Randbedingungen gefordert werden, dann ist das Problem nur lösbar, wenn die Komplementaritätsbedingung

$$\int_\Omega f \mathrm{d}\mathbf{x} + \int_{\partial\Omega} g_N \, \mathrm{d}s(\mathbf{x}) = 0$$

erfüllt ist. Wenn das Problem eine Lösung u besitzt, ist auch $u + c$ mit $c \in \mathbb{R}$ eine Lösung, also ist das Problem nicht eindeutig lösbar; die konstanten Funktionen sind also im Kern des Differentialoperators.

Die konstanten Funktionen sind *insbesondere* dann nicht im Kern des Differentialoperators, wenn

(i) $c \geq c_{min} > 0$ auf einer nicht-leeren Teilmenge von Ω,

(ii) $\gamma \geq \gamma_{min} > 0$ auf einer nicht-leeren Teilmenge des Robin-Randes oder

(iii) auf Teilen des Randes Dirichlet-Randbedingungen gefordert werden.

Behandlung eines Konvektionsterms

Die Behandlung eines Konvektionsterms ähnlich zum eindimensionalen Problem möglich. Wir hätten dann b_1, b_2, c und f gegeben. Gesucht wäre ein (hinreichend glattes) u mit

$$-\Delta u + \underbrace{b_1 \partial_x u + b_2 \partial_y u}_{= b \cdot \nabla u} + cu = f$$

und entsprechende Randbedingungen.

Die Richtungsableitungen $\partial_x u$ und $\partial_y u$ können wie im eindimensionalen Fall mit zentralen Differenzenquotienten oder mit einseitigen Differenzenquotienten realisiert werden. Letztere können wieder beispielsweise als Upwind-Schema realisiert werden. Man kann sich leicht überlegen, dass ein diskretes Maximumsprinzip erfüllt sein wird,

(i) wenn die Gitterweite klein genug ist oder

(ii) wenn etwa das folgendermaßen definierte Upwind-Schema verwendet wird:

$$\begin{aligned}(L_h u_h)(\mathbf{x}_{i,j}) = &\left(\tfrac{4}{h^2} + \tfrac{|b_1(\mathbf{x}_{i,j})|+|b_2(\mathbf{x}_{i,j})|}{h} + c(\mathbf{x}_{i,j})\right) u_{i,j} \\ &- \left(\tfrac{1}{h^2} + \tfrac{b_1^+(\mathbf{x}_{i,j})}{h}\right) u_{i-1,j} - \left(\tfrac{1}{h^2} - \tfrac{b_1^-(\mathbf{x}_{i,j})}{h}\right) u_{i+1,j} \\ &- \left(\tfrac{1}{h^2} + \tfrac{b_2^+(\mathbf{x}_{i,j})}{h}\right) u_{i,j-1} - \left(\tfrac{1}{h^2} - \tfrac{b_2^-(\mathbf{x}_{i,j})}{h}\right) u_{i,j+1},\end{aligned}$$

wobei

$$b_n^+(\mathbf{x}_{i,j}) := \max\{0, b_n(\mathbf{x}_{i,j})\} \quad \text{und} \quad b_n^-(\mathbf{x}_{i,j}) := \min\{0, b_n(\mathbf{x}_{i,j})\}.$$

Weitere Schritte zu Stabilitäts- und Konvergenzanalyse sind in der Literatur zu finden.

11. Konvergenz in anderen Normen

Wir kehren zum eindimensionalen Fall zurück. Bisher hatten wir die Konvergenz der Differenzenschemata in der Supremumsnorm (Maximumnorm) betrachtet und je nach Glattheit von f eine Konvergenzordnung von 1 (für $f \in C^1$) oder 2 (für $f \in C^2$) erreicht. Für manche Anwendungen ist es auch von Interesse, Informationen über etwas wie die mittlere Abweichung zu wissen. Außerdem kann es sein, dass wir nicht nur an der Funktion u, sondern auch der Ableitung u' interessiert sind. Konvergenzordnung 1 bezüglich der Supremumsnorm sagt nichts über die Konvergenz der Ableitungen aus; Skizze!

Wir betrachten also wieder das bereits bekannte Anfangswertproblem, in diesem Kapitel stets mit $b \equiv 0$, also

$$L\,u := -u''(x) + c(x)u(x) = f(x), \qquad u(0) = u(1) = 0 \tag{11.1}$$

mit $c(x) \geq 0$. Zur Diskretisierung betrachten wir wieder ein äquidistantes Gitter

$$\mathcal{T}_h = \{x_1, \cdots, x_n\} \quad \text{mit} \quad x_i = ih, \quad h = \tfrac{1}{n+1}$$

und das auf dem zentralen Differenzenquotienten basierende Differenzenschema

$$\begin{aligned} L_h u_h(x_i) &:= -D_h^2 u_h(x_i) + c(x_i)u_h(x_i) \\ &= \tfrac{1}{h^2}(-u_{i-1} + 2u_i - u_{i+1}) + c(x_i)u_i = f(x_i), \qquad x_i \in \mathcal{T}_h, \\ u_h(0) &= u_h(1) = 0. \end{aligned} \tag{11.2}$$

Symmetrie und positive Definitheit

Der Ausgangspunkt für unsere Überlegungen sind Eigenschaften des Differenzenschemas. Uns ist bereits aufgefallen, dass die Steifigkeitsmatrix im Fall $b \equiv 0$ symmetrisch ist. Wir können eine entsprechende Eigenschaft sofort auch für L_h herleiten. Für Polygonzüge $u_h, v_h \in V_h$ mit $u_i = u_h(x_i)$, $v_i = v_h(x_i)$ definieren wir

$$(u_h, v_h)_{0,h} := \tfrac{h}{2}u_0v_0 + h\sum_{i=1}^{n} u_iv_i + \tfrac{h}{2}u_{n+1}v_{n+1}$$

als Näherung für $(u_h, v_h)_0 = \int_0^1 u_h(x)v_h(x)\mathrm{d}x$. (Diese Näherung erhalten wir etwa, wenn wir die Integrale $\int_{x_i}^{x_{i+1}} u_h(x)v_h(x)\mathrm{d}x$ jeweils mit der Trapezregel auflösen.) Dazu passend führen wir auch eine Norm ein:

$$\|u_h\|_{0,h} := (u_h, u_h)_{0,h}^{1/2}.$$

Lemma 11.1. *Der Operator L_h ist symmetrisch in dem Sinn, dass*

$$(L_h u_h, v_h)_{0,h} = (L_h v_h, u_h)_{0,h} \quad \forall\, u_h, v_h \in V_{0,h}.$$

Beweis. Analog zum partiellen Integrieren können wir auch partiell summieren:

$$\sum_{i=1}^{n}(a_{i+1}-a_i)b_i = \sum_{i=1}^{n} a_{i+1}b_i - \sum_{i=1}^{n} a_i b_i = a_{n+1}b_n - a_1 b_0 + \sum_{i=1}^{n} a_i b_{i-1} - \sum_{i=1}^{n} a_i b_i$$
$$= a_{n+1}b_n - a_1 b_0 - \sum_{i=1}^{n} a_i(b_i - b_{i-1}) \quad \forall a_1,\dots,a_{n+1}, b_0,\dots,b_n \in \mathbb{R}.$$

Mit Definitionen und den Randdaten $u_0 = u_{n+1} = v_0 = v_{n+1} = 0$ erhalten wir

$$\begin{aligned}
(L_h u_h, v_h)_{0,h} &= h\sum_{i=1}^{n}\left(-\tfrac{1}{h^2}(u_{i+1}-2u_i+u_{i-1})v_i + c_i u_i v_i\right)\\
&= h\sum_{i=1}^{n}\left(\tfrac{1}{h^2}(u_{i+1}-u_i) - (u_i - u_{i-1})\right)v_i + c_i u_i v_i\Big)\\
&= \sum_{i=1}^{n+1}\left(\tfrac{1}{h}(u_i - u_{i-1})(v_i - v_{i-1}) + h c_i u_i v_i\right) \qquad (11.3)\\
&= h\sum_{i=1}^{n}\left(\tfrac{1}{h^2}u_i((v_{i+1}-v_i) - (v_i - v_{i-1})) + c_i u_i v_i\right) = (L_h v_h, u_h)_{0,h},
\end{aligned}$$

was zu zeigen war. □

Die Aussage von Lemma 11.1 ist äquivalent zur Symmetrie der Steifigkeitsmatrix. Für $u_h, v_h \in V_{h,0} = \{w_h \in V_h : w_h(0) = w_h(1) = 0\}$ und entsprechender Koordinatendarstellung $\underline{u}_h := (u_1, \cdots, u_n)^\top$ und $\underline{v}_h := (v_1, \cdots, v_n)^\top$ haben wir

$$\begin{aligned}
\underline{v}_h^\top \underline{u}_h &= \tfrac{1}{h}(u_h, v_h)_{0,h},\\
\underline{v}_h^\top A_h \underline{u}_h &= \tfrac{1}{h}(L_h u_h, v_h)_{0,h} = \tfrac{1}{h}(L_h v_h, u_h)_{0,h} = \underline{v}_h^\top A_h^\top \underline{u}_h.
\end{aligned}$$

Aus dem Beweis lässt sich jedoch noch mehr ablesen. Wir erinnern uns: eine Matrix ist positiv definit (semidefinit), wenn alle Eigenwerte größer (oder gleich) 0 sind. Dieses können wir über den Rayleigh-Quotient charakterisieren: A_h ist positiv (semi)definit, wenn $\underline{u}_h^\top A_h \underline{u}_h > 0$ (≥ 0) für alle Vektoren $\underline{u}_h \neq \mathbf{0}$. Mit (11.3) haben wir

$$h\,\underline{u}_h^\top A_h \underline{u}_h = (L_h u_h, u_h)_{0,h} = \sum_{i=1}^{n+1}\left(\tfrac{1}{h}(u_i - u_{i-1})^2 + h c_i u_i^2\right) \geq 0. \qquad (11.4)$$

Für $c \geq c_{min} > 0$ würden wir bereits sofort erhalten, dass $(L_h u_h, u_h)_{0,h} \geq c_{min}\|u_h\|_{0,h}^2 > 0$ für alle $u_h \in V_{0,h} \setminus \{0\}$, woraus sich positive Definitheit ergibt.

Für viele Anwendungen ist aber eine solche untere Schranke an c nicht vorhanden. Daher gehen wir davon aus, dass sie nicht zur Verfügung steht. Dies gibt uns den Anlass, eine Seminorm einzuführen:

$$|u_h|_{1,h}^2 := \sum_{i=1}^{n+1} \tfrac{1}{h}(u_i - u_{i-1})^2.$$

Man kann sich leicht überlegen, dass $|u_h|_{1,h}^2 = \sum_{i=1}^{n+1} \int_{x_{i-1}}^{x_i} (u_h'(x))^2 \mathrm{d}x$ für alle Polygonzüge $u_h \in V_h$ gilt. Da die Funktion u_h nicht stetig differenzierbar ist, wäre $\int_0^1 (u_h'(x))^2 \mathrm{d}x$ im klassischen Sinn nicht wohldefiniert; dieses Problem ergibt sich bei der stückweisen Definition nicht. Der Index 1 in der Definition der Norm deutet hier darauf hin, dass wir es mit ersten Ableitungen zu tun haben. Diese Seminorm erfüllt offensichtlich die Eigenschaften einer Seminorm (positive Semidefinitheit $|u_h|_{1,h} \geq 0$, Homogenität $|\alpha u_h|_{1,h} = |\alpha|\,|u_h|_{1,h}$ und Dreiecksungleichung $|u_h + w_h|_{1,h} \leq |u_h|_{1,h} + |w_h|_{1,h}$), jedoch sie wegen $|\mathbf{1}|_{1,h} = 0$ nicht Definit und daher Norm.

Nehmen wir eine Zusatzinformation dazu, können wir wieder postitive Definitheit garantieren.

Satz 11.2 (Diskrete Poincaré/Friedrichs-Ungleichung)**.**
Es gibt eine Konstante $c_P > 0$, sodass für alle Polygonzüge $u_h \in V_h$ gilt:

(i) $\|u_h\|_{0,h} \leq c_P\big(|u_h(0)|^2 + |u_h|_{1,h}^2\big)^{1/2}$,

(ii) $\|u_h\|_{0,h} \leq c_P\big(|u_h(1)|^2 + |u_h|_{1,h}^2\big)^{1/2}$,

(iii) $\|u_h\|_{0,h} \leq c_P\big(\frac{1}{(t-s)^2}|\int_s^t u_h(x)\mathrm{d}x|^2 + |u_h|_{1,h}^2\big)^{1/2}$ *mit* $0 \leq s < t \leq 1$.

Beweis. Wir beschränken uns auf den Fall (i). Wir erinnern uns an die Cauchy-Schwarz'sche Ungleichung $\sum_{i=1}^n a_i b_i \leq (\sum_{i=1}^n a_i^2)^{1/2}(\sum_{i=1}^n b_i^2)^{1/2}$ und stellen sofort fest, dass wir mit $b_i = 1$ daraus auch $(\sum_{i=1}^n a_i)^2 \leq n \sum_{i=1}^n a_i^2$ schließen können. Wegen $u_i = u_0 + \sum_{j=1}^i (u_j - u_{j-1})$ haben wir für $i \in \{0, \ldots, n+1\}$

$$u_i^2 = \left(u_0 + \sum_{j=1}^{i}(u_j - u_{j-1})\right)^2 \leq 2u_0^2 + 2(n+1)\sum_{j=1}^{i}(u_j - u_{j-1})^2 \leq 2u_0^2 + 2|u_h|_{1,h}^2.$$

Wegen $h = 1/(n+1)$ folgt daraus

$$\|u_h\|_{0,h}^2 = \tfrac{h}{2}u_0^2 + h\sum_{i=1}^{n} u_i^2 + \tfrac{h}{2}u_{n+1}^2 \leq \max_{i=0,\ldots,n+1} u_i^2 \leq 2u_0^2 + 2|u_h|_{1,h}^2,$$

also das gewünschte Resultat mit $c_P^2 = 2$. Der Fall (ii) ist komplett analog und (iii) ist ein Fall für die Übung. □

Wenn wir die Ungleichung umdrehen wollen, erhalten wir

Satz 11.3 (Inverse Ungleichung)**.** *Es gibt eine Konstante c_{inv}, sodass*

$$|u_h|_{1,h} \leq \frac{c_{inv}}{h}\|u_h\|_{0,h} \qquad \forall u_h \in V_h.$$

Beweis. Mit der Ungleichung $(a-b)^2 \leq 2a^2 + 2b^2$ haben wir

$$|u_h|_{1,h}^2 = \tfrac{1}{h}\sum_{i=1}^{n+1}(u_i - u_{i-1})^2 \leq \tfrac{2}{h}\sum_{i=1}^{n+1}(u_i^2 + u_{i-1}^2) \leq \tfrac{4}{h}\sum_{i=0}^{n+1} u_i^2 = \tfrac{4}{h^2}\|u_h\|_{0,h}^2,$$

also das gewünschte Resultat. □

Bemerkung 11.4 (Definitheit von $|\cdot|_{1,h}$ auf $V_{h,0}$)**.**
Mit der diskreten Poincaré/Friedrichs-Ungleichung erhalten wir für alle $u_h \in V_{h,0} \setminus \{0\}$, dass $|u_h|_{1,h} \geq c_P^{-1}\|u_h\|_{0,h} > 0$, also die Definitheit. Das bedeutet, dass $|\cdot|_{1,h}$ am Raum $V_{h,0}$ eine Norm ist.

Mit der diskreten Poincaré/Friedrichs-Ungleichung, der inversen Ungleichung und (11.4) erhalten wir

$$c_P^{-2}\|u_h\|_{0,h}^2 \leq |u_h|_{1,h}^2 \leq (L_h\, u_h, u_h)_{0,h} \leq (h^{-2}c_{inv}^2 + \|c\|_\infty)\|u_h\|_{0,h}^2 \tag{11.5}$$

für alle $u_h \in V_{h,0}$. Mit (11.2) und der Cauchy-Schwarz'schen Ungleichung erhalten wir

$$(L_h u_h, u_h)_{0,h} = (f, u_h)_{0,h} \leq \|f\|_{0,h}\|u_h\|_{0,h} \leq c_P\|f\|_{0,h}|u_h|_{1,h}$$

und durch Kombination dieser Abschätzung mit (11.5) auch diskrete Stabilität:

$$|u_h|_{1,h} \leq c_P\|f\|_{0,h} \quad \text{und} \quad \|u_h\|_{0,h} \leq c_P^2\|f\|_{0,h}. \tag{11.6}$$

Eigenschaften der Steifigkeitsmatrix

Wir wollen noch kurz die Eigenschaften der Steifigkeitsmatrix diskutieren. Zuerst stellen wir fest, dass für eine gute Approximation eine entsprechend kleine Gitterweite nötig ist und daher das asymptotische Verhalten für $h \to 0$ besonders interessant ist. Es gilt:

(i) Die Dimension der Matrix N verhält sich wie h^{-1} (bzw. h^{-d} für den Fall von d Ortsdimensionen, wie wir später sehen werden);

(ii) Die Anzahl der nichtverschwiendenden Einträge pro Zeile ist gleichmäßig beschränkt (in den von uns behandelten eindimensionalen Fällen mit 3 Einträgen); Familien von Matrizen mit dieser Eigenschaft bezeichnen wir als *dünnbesetzt*

(engl. *sparse*);
(iii) Wegen (11.5) gilt für den kleinsten Eigenwert der Steifigkeitsmatrix:

$$\lambda_{\min}(A_h) \geq c_P^{-2};$$

(iv) Wegen (11.5) gilt für den größten Eigenwert der Steifigkeitsmatrix:

$$\lambda_{\max}(A_h) \leq h^{-2} c_{inv}^2 + \|c\|_\infty;$$

(v) Daher gilt für die Konditionszahl (die Konstante c_κ hängt auch von $\|c\|_\infty$ ab):

$$\kappa(A_h) \leq c_\kappa h^{-2}. \tag{11.7}$$

Bemerkung 11.5. Den Exponenten -2 in der Konditionszahlabschätzung erhalten wir, da wir Differentialgleichungen der Ordnung 2 betrachtet haben.

Konvergenz

Mit der in diesem Kapitel eingeführten Technik können wir auch Konvergenz zeigen. Wir betrachten den Fehler zwischen der Lösung u des Randwertproblems (11.1) und dem Polygonzug u_h, der das Differenzenschema (11.2) löst. Wir werten diesen Unterschied in der diskreten Norm $|\cdot|_{1,h}$ aus und vergleichen dazu die den Polygonzug u_h mit dem Polynomzug, der u interpoliert. Dazu ist folgender Operator sinnvoll.

Definition 11.6 (Interpolationsoperator)**.** Der Interpolationsoperator ist jener lineare Operator $r_h : C[0,1] \to V_h$, der jeder Funktion w den interpolierenden Polygonzug $r_h w := w_h \in V_h$ zuordnet, also mit $w(x_i) = w_h(x_i), \forall\, i = 0, \ldots, n+1$.

Für den Interpolationsfehler ist unter anderem folgende Abschätzung möglich:

Satz 11.7 (Interpolationsfehler)**.** *Es gibt eine Konstante $c_{int} > 0$, sodass*

$$\int_0^1 |w - r_h w| \mathrm{d}x \leq c_{int} h \sum_{i=1}^{n+1} \int_{x_{i-1}}^{x_i} |w'| \,\mathrm{d}x$$

für alle $w \in C[0,1]$ mit $w|_{(x_{i-1},x_i)} \in C^1$.

Beweis. Sei $v \in C[0,1]$ und $v_h(x) := v(0)(1-x)+v(1)x$ die lineare Interpolation. Es gilt dann wegen $(1-y)+y = 1$ und mit dem Hauptsatz der Differential- und

Integralrechnung

$$\begin{aligned}\int_0^1 |v(y) - v_h(y)|\,\mathrm{d}y &= \int_0^1 |(v(y) - v(0))(1-y) + (v(y) - v(1))y|\,\mathrm{d}y \\ &= \int_0^1 \left| \int_0^y |v'(t)|\mathrm{d}t\,(1-y) + \int_x^1 |v'(t)|\mathrm{d}t\,y \right| \mathrm{d}y \\ &\le \int_0^1 \big((1-y)+y\big)\,\mathrm{d}y \int_0^1 |v'(t)|\mathrm{d}t = \int_0^1 |v'(y)|\mathrm{d}y.\end{aligned}$$

Mithilfe von Substitutions- und Kettenregel erhalten wir mit $\psi_k(y) = x_{k-1} + yh$, $v(y) = w(\psi_k(y))$ und $v_h(y) = r_h w(\psi_k(y))$

$$\int_{x_{k-1}}^{x_k} |w(x) - r_h w(x)|\,\mathrm{d}x \le h \int_{x_{k-1}}^{x_k} |w'(x)|\,\mathrm{d}x.$$

Das gesuchte Resultat erhalten wir nun durch Bildung der Summe über k. □

Satz 11.8 (Diskretisierungsfehlerabschätzung)**.**
Für $f, c \in C^1$ gilt

$$|r_h u - u_h|_{1,h} = \mathcal{O}(h),$$

wobei u und u_h die Lösungen von Randwertproblem (11.1) *und Differenzenschema* (11.2) *sind.*

Beweis. Sei $w_h := r_h u - u_h \in V_{h,0}$. Wir verwenden die Definitionen und partielle Integration. Wegen der Stetigkeit von $u' w_h$ und $w_h(0) = w_h(1) = 0$, gilt

$$\begin{aligned}(r_h u, w_h)_{1,h} &= \tfrac{1}{h} \sum_{i=1}^{n+1} (u(x_i) - u(x_{i-1}))(w_i - w_{i-1}) = \sum_{i=1}^{n+1} \int_{x_{i-1}}^{x_i} u'(x) w_h'(x)\mathrm{d}x \\ &= \sum_{i=1}^{n+1} \int_{x_{i-1}}^{x_i} -u''(x) w_h(x)\mathrm{d}x = \int_0^1 -u'' w_h \mathrm{d}x = \int_0^1 (f - cu) w_h \mathrm{d}x. \qquad (11.8)\end{aligned}$$

Analog zum Beweis von Lemma 11.1 und mit Einsetzen in das Differenzenschema erhalten wir unter Ausnutzung von $(a,b)_{0,h} = \int_0^1 r_h(ab)\mathrm{d}x$ (Zwischenrechnung!)

$$(u_h, w_h)_{1,h} = (-D_h^2 u_h, w_h)_{0,h} = (f - cu_h, w_h)_{0,h} = \int_0^1 r_h\big((f - cu_h) w_h\big)\mathrm{d}x. \quad (11.9)$$

Wegen $c \ge 0$, $w_h = r_h u - u_h$, erneut $(a,b)_{0,h} = \int_0^1 r_h(ab)\mathrm{d}x$ und $r_h r_h u = r_h u$ gilt

$$\begin{aligned}|w_h|_{1,h}^2 &\le |w_h|_{1,h}^2 + (cw_h, w_h)_{0,h} = (r_h u, w_h)_{1,h} - (u_h, w_h)_{1,h} + \int_0^1 r_h(cw_h^2)\mathrm{d}x \\ &= \int_0^1 (f - cu) w_h - r_h\big((f - cu) w_h\big)\mathrm{d}x.\end{aligned}$$

Wir schätzen das Integral mit Satz 11.7 ab und erhalten mit fundamentalen Umformungen und der Poincaré-Friedrichs-Ungleichung:

$$
\begin{aligned}
|w_h|_{1,h}^2 &\leq c_{int} h \sum_{i=1}^{n+1} \int_{x_{i-1}}^{x_i} |(f-cu)' w_h + (f-cu) w_h'| \mathrm{d}x \\
&\leq c_{int} h \sum_{i=1}^{n+1} \left(\int_{x_{i-1}}^{x_i} |(f-cu)'| \mathrm{d}x \, (|w_i| + |w_{i-1}|) + \int_{x_{i-1}}^{x_i} |f-cu| \mathrm{d}x \, \tfrac{1}{h} |w_i - w_{i-1}| \right) \\
&\leq \sqrt{2} c_{int} h \left(\int_0^1 |(f-cu)'|^2 \mathrm{d}x \right)^{1/2} \|w_h\|_{0,h} + c_{int} h \left(\int_0^1 |f-cu|^2 \mathrm{d}x \right)^{1/2} |w_h|_{1,h} \\
&\leq c_{int} h \left(\sqrt{2} c_P \left(\int_0^1 |(f-cu)'|^2 \mathrm{d}x \right)^{1/2} + \left(\int_0^1 |f-cu|^2 \mathrm{d}x \right)^{1/2} \right) |w_h|_{1,h},
\end{aligned}
$$

woraus das gesuchte Resultat unter Ausnutzung von $c, f, u \in C^1$ unmittelbar folgt. □

Bemerkung 11.9.

(i) Wir erinnern uns: Mit der Theorie aus dem Kapitel 8 konnten wir für $u \in C^4$ (und damit $f \in C^2$) Resultate der Art $\|r_h u - u_h\|_{\infty,h} = \mathcal{O}(h^2)$ zu zeigen. Daraus folgt wegen $\|w_h\|_{0,h} \leq \|w_h\|_{\infty,h}$ und Satz 11.3 auch $\|r_h u - u_h\|_{0,h} = \mathcal{O}(h^2)$ und $|r_h u - u_h|_{1,h} = \mathcal{O}(h)$. Durch die nun behandelte Analyse ist es gelungen, $|r_h u - u_h|_{1,h} = \mathcal{O}(h)$ auch für den Fall $f \in C^1$ zu zeigen, also mit einer schwächeren Regularitätsannahme.

(ii) Wenn man sich die Analyse etwas genauer ansieht, kann man feststellen, dass es ausreicht, wenn f und c jeweils stetig und nur *stückweise* stetig differenzierbar sind ($f \in C^0$ mit $f|_{(x_{i-1},x_i)} \in C^1$ und entsprechend für c).

(iii) Durch eine etwas ausgereiftere Analyse kann man für $f \in C^1$ auch Konvergenzresultate der Art $\|r_h\, u - u_h\|_{0,h} = \mathcal{O}(h^2)$ zeigen.

(iv) Für einfache Fälle, etwa c konstant, lassen sich die Eigenwerte der Steifigkeitsmatrix errechnen (vgl. Übung). Auch mit diesem Ansatz lässt sich diskrete Stabilität zeigen (vgl. Übung).

Numerisches Experiment

Beispiel 11.10. Wir betrachten wieder das Beispiel 8.24, nur betrachten wir nun den Fehler $e_{1,h} := |r_h u - u_h|_{1,h}$ in Tabelle 11.1. Die Faktoren $e_{1,2h}/e_{1,h}$ erreichen den Wert 2, was bedeutet, dass die Konvergenzordnung 1 erreicht wird. Der Fehler $e_{0,h} := \|r_h u - u_h\|_{0,h}$ ist in Tabelle 11.2 angegeben. Die Faktoren $e_{0,2h}/e_{0,h}$ erreichen nun den Wert 4, was bedeutet, dass – wie bei der Supremumsnorm – die Konvergenzordnung 2 erreicht wird. Beide Resultate stimmen perfekt mit der Theorie überein.

h	Fehler $e_{1,h}$	Faktor $e_{1,2h}/e_{1,h}$
2^{-2}	$5.27\ 10^{0}$	
2^{-3}	$2.35\ 10^{0}$	2.25
2^{-4}	$1.14\ 10^{0}$	2.05
2^{-5}	$5.68\ 10^{-1}$	2.01
2^{-6}	$2.83\ 10^{-1}$	2.00
2^{-7}	$1.42\ 10^{-1}$	2.00
2^{-8}	$7.08\ 10^{-2}$	2.00

Tabelle 11.1.: Fehler in der Norm $e_{1,h} = |r_h u - u_h|_{1,h}$.

h	Fehler $e_{0,h}$	Faktor $e_{0,2h}/e_{0,h}$
2^{-2}	$2.64\ 10^{-1}$	
2^{-3}	$4.29\ 10^{-2}$	6.16
2^{-4}	$9.53\ 10^{-3}$	4.50
2^{-5}	$2.31\ 10^{-3}$	4.12
2^{-6}	$5.73\ 10^{-4}$	4.03
2^{-7}	$1.43\ 10^{-4}$	4.01
2^{-8}	$3.57\ 10^{-5}$	4.00

Tabelle 11.2.: Fehler in der Norm $e_{0,h} = \|r_h u - u_h\|_{0,h}$.

12. Variationsformulierung und Finite Elemente Methode

Die im letzten Kapitel behandelte Theorie gibt uns die Idee für eine neue Art der Formulierung unseres Randwertproblems und seiner Diskretisierung. Wir betrachten dazu wieder ein Randwertproblem folgender Art: Finde $u \in C^2(0,1) \cap C^1(0,1] \cap C^0[0,1]$, sodass

$$-(a(x)u'(x))' + c(x)u(x) = f(x), \qquad x \in (0,1) \tag{12.1}$$

$$u(0) = g_0, \tag{12.2}$$

$$a(1)u'(1) + \gamma_1 u(1) = g_1, \tag{12.3}$$

wobei $a \geq \underline{a} > 0, c \geq 0, f, \gamma_1 \geq 0$, g_0 und g_1 gleichmäßig beschränkt und hinreichend glatt sind. Mit dieser Wahl von Randbedingungen wollen wir zeigen, dass die im folgenden vorgestellte Methodik für alle eingeführten Arten von Randbedingungen möglich ist. Neumann-Bedingungen sind durch diese Wahl miterfasst.

Wenn u das Randwertproblem erfüllt und außerdem die Integrale wohldefiniert sind, dann gilt wegen partieller Integration für jedes $v \in C^1[0,1]$ mit $v(0) = 0$:

$$\begin{aligned}\int_0^1 f(x)v(x)\mathrm{d}x &= \int_0^1 -(a(x)u'(x))'v(x)\mathrm{d}x + \int_0^1 c(x)u(x)v(x)\mathrm{d}x \\ &= \int_0^1 a(x)u'(x)v'(x)\mathrm{d}x + a(0)u'(0)v(0) - a(1)u'(1)v(1) + \int_0^1 c(x)u(x)v(x)\mathrm{d}x \\ &= \int_0^1 a(x)u'(x)v'(x)\mathrm{d}x + (\gamma_1 u(1) - g_1)v(1) + \int_0^1 c(x)u(x)v(x)\mathrm{d}x.\end{aligned}$$

Somit erhalten wie folgendes Variationsproblem: Finde $u \in C^1_{g,D}$, sodass

$$\underbrace{\int_0^1 \big(au'v' + cuv\big)\mathrm{d}x + \gamma_1 u(1)v(1)}_{a(u,v):=} = \underbrace{\int_0^1 fv\mathrm{d}x + g_1 v(1)}_{\ell(v):=} \quad \forall v \in C^1_{0,D}, \tag{12.4}$$

wobei $C^1_{0,D} := \{v \in C^1[0,1] : v(0) = 0\}$ und $C^1_{g,D} := \{u \in C^1[0,1] : u(0) = g_0\}$.

Der Name Variationsgleichung ergibt sich aus der Tatsache, dass die Testfunktion v ja noch variiert werden darf. Das Variationsproblem ist auch als *variationelle Formulierung* bekannt. Das Randwertproblem (12.1) – (12.3) ist es auch als *klassische* oder *starke Formulierung* bekannt.

Bemerkung 12.1 (Wesentliche und natürliche Randbedingungen)**.**
Die Robin-Randbedingung bei $x = 1$ ließ sich durch Einsezten in der Herleitung der Variationsgleichung unmittelbar verwenden; sie wird daher als *natürliche* Randbedingung bezeichnet. (Dasselbe gilt auch für Neumann-Randbedingungen.)

Die Dirichlet-Randbedingung $u(0) = g_0$ kann dagegen nicht direkt eingesetzt werden und muss noch zusätzlich gefordert werden; man spricht von einer *wesentlichen* (oder: *essentiellen*) Randbedingung. Die entsprechende homogene Bedingung $v(0) = 0$ musste auch an die Testfunktion gestellt werden, um den Randterm bei $x = 0$ loszuwerden. Dies entspricht dem Prinzip, dass als Testfunktionen die möglichen Variationen der Lösung zu wählen sind, also die möglichen Differenzen von Elementen in der Menge $C^1_{g,D}$, in der wir die Lösungen suchen.

Die Menge $C^1_{0,D}$ ist ein *Vektorraum*, währenddessen $C^1_{g,D} = g_0 + C^1_{0,D}$ dadurch erzeugt wird, dass der Vektorraum um g_0 verschoben wird.

Äquivalenz zwischen variationeller und klassicher Formulierung

Aus der Herleitung wissen wir, dass eine Lösung des Randwertproblems (12.1) – (12.3) mit $u \in C^1[0,1]$ auch eine Lösung der Variationsproblems (12.4) ist. Nun wollen wir uns überlegen, dass wir auch den umgekehrten Weg gehen können. Sei nun $u \in C^1_{g,D} \cap C^2(0,1)$ eine Lösung des Variationsproblems. Mit der Wahl $v \in C^1$ mit $v(0) = v(1) = 0$ erhalten wir unter Ausnutzung der Stetigkeit von uv', dass

$$0 = \int_0^1 \big(au'v' + cuv - fv\big)\,\mathrm{d}x = \int_0^1 \underbrace{\big(-(au')' + cu - f\big)}_{w :=} v\,\mathrm{d}x.$$

Es gilt nun $w \equiv 0$; wenn dies nicht der Fall ist, gibt es ein $x^* \in (0,1)$ mit $w(x^*) > 0$ (oder < 0, was analog behandelt werden kann). Wegen Stetigkeit gibt es ein $\epsilon > 0$ mit $[x^* - \epsilon, x^* + \epsilon] \subset (0,1)$ und $w|_{[x^*-\epsilon, x^*+\epsilon]} > 0$. Sei nun

$$v(x) := \big(\max\{\epsilon^2 - (x - x^*)^2, 0\}^2\big)^2,$$

dann ist $0 = \int_0^1 wv\mathrm{d}x = \int_{x^*-\epsilon}^{x^*+\epsilon} wv\mathrm{d}x > 0$, was ein Widerspruch ist. Daher gilt $w \equiv 0$, also (12.1). Wählen wir nun $v \in C^1$ mit $v(0) = 0$ und $v(1) = 1$, so erhalten wir mit partieller Integration und $-(au')' + cu = f$, die Geltung der Robin-Randbedingung (12.3). Die Dirichlet-Randbedingung (12.2) gilt schon wegen der Forderung $u \in C^1_{g,D}$. Somit haben wir gezeigt:

Satz 12.2. *Jede Lösung u des Variationsproblems* (12.4)*, welche die Regularitätsbedingung $u \in C^2(0,1) \cap C^1(0,1] \cap C[0,1]$ erfüllt, ist auch eine Lösung der klassischen Formulierung* (12.1) – (12.3)*.*

Es ist jedoch möglich, dass Lösungen eines Variationsproblems nicht in C^2 liegen und daher die starke Formulierung nicht erfüllen.

Eindeutigkeit der schwachen Lösung und Stabilität

Seien

$$(u,v)_0 := \int_0^1 u(x)v(x)\mathrm{d}x \quad \text{und} \quad \|u\|_0 := (u,u)_0^{1/2}$$

das Skalarprodukt (= bilinear, symmetrisch und positiv definit) und die zugehörige Norm, die Sie schon aus der VL *Funktionalanalysis* von der Definition des Lebesgue-Funktionenraums $L^2(0,1)$ kennen. $(\cdot,\cdot)_0$ und $\|\cdot\|_0$ sind aber auch auf $C^0[0,1]$ wohldefiniert und erfüllen dort die Eigenschaften von Skalarprodukt und Norm. Analog zum letzten Kapitel führen wir auch die Seminorm

$$|u|_1 := \|u'\|_0 = \left(\int_0^1 \big(u'(x)\big)^2 \mathrm{d}x\right)^{1/2}$$

sowie die Norm

$$\|u\|_1 := (\|u\|_0^2 + |u|_1^2)^{1/2} = \left(\int_0^1 u^2(x) + \big(u'(x)\big)^2 \mathrm{d}x\right)^{1/2}$$

ein, die Sie schon aus den VL *Funktionalanalysis* und *Partielle Differentialgleichungen* von der Definition des Sobolevraumes $H^1(0,1)$ kennen. Auch $\|\cdot\|_0$ und $|\cdot|_1$ erfüllen die Voraussetzungen an Normen und Seminormen.

Bemerkung 12.3. Die Idee der Räume $L^2(0,1)$ und $H^1(0,1)$ war es, dass auf diesen Räumen die jeweils zugehörige Norm ($\|\cdot\|_0$ bzw. $\|\cdot\|_1$) nicht nur die Normeigenschaften erfüllt, sondern dass auch alle Cauchy-Folgen in L^2 bzw. H^1 konvergieren (=der Raum L^2 bezüglich der Norm $\|\cdot\|_0$ und der Raum H^1 bezüglich der Norm $\|\cdot\|_1$ vollständig ist). Daher sind L^2 und H^1 Banach-Räume, ja sogar Hilbert-Räume! Im Gegensatz dazu ist $C^0[0,1]$ nicht vollständig bezüglich unserer Normen; dies soll uns jedoch *vorerst* nicht stören.

Wir leiten nun die Eigenschaften her, die uns bereits aus dem letzten Kapitel für das Differenzenschema bekannt sind. Wir haben

$$a(u,u) = \int_0^1 (au'(x))^2 + c(x)u(x)^2\,\mathrm{d}x \geq \underline{a}\int_0^1 (u'(x))^2\,\mathrm{d}x = \underline{a}|u|_1^2 \geq 0. \tag{12.5}$$

Analog zum letzten Kapitel gilt auch $a(u,u) \geq c_{min}\|u\|_0^2$, wenn $c \geq c_{min} > 0$ gilt, was wir aber nicht annehmen wollen. Stattdessen nutzen wir wieder

Satz 12.4 (Poincaré/Friedrichs-Ungleichung)**.**
Es gibt eine Konstante $c_P > 0$, sodass für alle $u \in C^1[0,1]$ gilt:
(i) $\|u\|_0 \le c_P\big(|u(0)|^2 + |u|_1^2\big)^{1/2}$,
(ii) $\|u\|_0 \le c_P\big(|u(1)|^2 + |u|_1^2\big)^{1/2}$,
(iii) $\|u\|_0 \le c_P\big(\frac{1}{(t-s)^2}|\int_s^t u(x)\mathrm{d}x|^2 + |u|_1^2\big)^{1/2}$ *mit* $0 \le s < t \le 1$.

Beweis. Wir beschränken uns auf den Fall (i). Wegen $u(x) = u(0) + \int_0^x u'(t)\mathrm{d}t$ haben wir mit der Cauchy-Schwarz'schen Ungleichung, die auch im kontinuierlichen Fall gilt, für $x \in [0,1]$

$$u(x)^2 \le 2|u(0)|^2 + 2\left(\int_0^x u'(t)\mathrm{d}t\right)^2 \le 2|u(0)|^2 + 2x\int_0^x (u'(t))^2\mathrm{d}t \le 2|u(0)|^2 + 2\|u'\|_0^2.$$

Durch Integrieren erhalten wir das gesuchte Resultat mit $c_P^2 = 2$. Die Fälle (ii) und (iii) lassen sich analog zeigen. □

Bemerkung 12.5. Das Verhältnis $|u|_1/\|u\|_0$ ist unbeschränkt ($u_k(x) := \sin(k\pi x)$ mit $k \in \mathbb{N}$), es gibt also keine inverse Ungleichung (vgl. Satz 11.3) für den kontinuierlichen Fall.

Für den Beweis der Beschränktheit benötigen wir auch folgende Ungleichung.

Satz 12.6 (Spurungleichung)**.** *Es gibt eine Konstante $c_S > 0$, sodass für alle $u \in C^1[0,1]$ und alle $x \in [0,1]$ die Abschätzung $|u(x)| \le c_S\|u\|_1$ gilt.*

Beweis. Mit dem Hauptsatz der Differential- und Integralrechnung erhalten wir $u(x) = u(y) + \int_y^x u'(t)\mathrm{d}t$, woraus mit der Dreiecksungleichung und der Cauchy-Schwarz'schen Ungleichung $|u(x)|^2 \le 2|u(y)|^2 + 2\|u'\|_0^2$ für alle $x, y \in [0,1]$ folgt. Die gewünschte Ungleichung erhält durch Integrieren über y. □

Nun können wir die wichtigsten Eigenschaften von $a(\cdot,\cdot)$ und $\ell(\cdot)$ zeigen:
(i) $a(\cdot,\cdot)$ ist *bilinear* (= linear in beiden Komponenten) und $\ell(\cdot)$ ist *linear*;
(ii) $a(\cdot,\cdot)$ ist *symmetrisch* ($a(u,v) = a(v,u)$);
(iii) $a(\cdot,\cdot)$ ist *stetig* (=*beschränkt*): Es gilt für alle $u, v \in C^1[0,1]$

$$\begin{aligned} a(u,v) &\le \|a\|_\infty\|u'\|_0\|v'\|_0 + \|c\|_\infty\|u\|_0\|v\|_0 + c_S|\gamma_1|\|u\|_1\|v\|_1 \\ &\le (\|a\|_\infty + \|c\|_\infty + c_S|\gamma_1|)\|u\|_1\|v\|_1; \end{aligned}$$

(iv) $\ell(\cdot)$ ist *stetig* (=*beschränkt*): Es gilt für alle $v \in C^1$

$$\ell(v) \le \|f\|_0\|v\|_0 + c_S|g_1|\|v\|_1 \le (\|f\|_0 + c_s|g_1|)\|v\|_1;$$

(v) $a(\cdot,\cdot)$ ist *elliptisch* (=*koerziv*): Es gilt für alle $u \in C^1_{0,D}$

$$a(u,u) \ge \underline{a}\|u'\|_0^2 + \gamma_1|u(1)|^2 \ge \underline{a}(c_P^2+1)^{-1}\|u\|_1^2.$$

Mit der Koerzivität erhalten wir

Satz 12.7 (Eindeutigkeit)**.** *Das Variationsproblem* (12.4) *kann nur eine Lösung haben.*

Beweis. Die Differenz $u_1 - u_2$ zweier Lösungen erfüllt $a(u_1 - u_2, v) = 0$ für alle $v \in C^1_{0,D}$. Wir stellen fest, dass wir wegen $u_1(0) = u_2(0) = u_0$ hier $v := u_1 - u_2 \in C^1_{0,D}$ wählen dürfen. Mit dieser Wahl und der Koerzivität erhalten wir $0 = a(u_1 - u_2, u_1 - u_2) \ge \underline{a}(c_P^2+1)^{-1}\|u_1 - u_2\|_1^2$, woraus $u_1 - u_2 = 0$ folgt. □

Bemerkung 12.8 (Andere Randbedingungen)**.**
Die Randbedingungen bereiten beim Beweis der Stetigkeit keine großen Probleme, sind jedoch beim Beweis der Elliptizität essentiell. Für den Beweis der Elliptizität ist jedenfalls erforderlich, dass $c \ge 0$ sowie die Koeffizienten γ_0 und γ_1 von Robin-Randbedingungen ≥ 0 sind; außerdem benötigen wir eine der weiteren Bedingungen (a) – (c). Wenn

(a) auf *einem* der beiden Ränder eine Dirichlet-Randbedingung oder

(b) auf *einem* der beiden Ränder eine Robin-Randbedingung mit γ_0 bzw. $\gamma_1 > 0$

gefordert ist, können wir mit den Fällen (i) und (ii) der Poincaré/Friedrichs-Ungleichung Elliptizität zeigen.

(c) Wenn es $0 \le s < t \le 1$ und $\underline{c} > 0$ gibt, sodass $c(x) \ge \underline{c}$ für alle $x \in [s,t]$, dann können wir Elliptizität mit Fall (iii) der Poincaré/Friedrichs-Ungleichung zeigen (vgl. Übung).

Wenn $c \equiv 0$ (also wir die Differentialgleichung $-u'' = f$ betrachten) und wir zwei Neumann-Randbedingungen ($-u'(0) = g_0$, $u'(1) = g_1$) fordern, dann liegen die konstanten Funktionen im Kern des Differentialoperators (vgl. Beispiel 8.22), Elliptizität können wir in diesem Fall nicht zeigen. Das Problem ist nur dann lösbar, wenn die Kompatibilitätsbedingung $g_0 + g_1 + \int_0^1 f\,\mathrm{d}x = 0$ erfüllt ist. Wenn u_0 eine Lösung ist, dann ist der Lösungsraum durch $\{u_0 + \alpha : \alpha \in \mathbb{R}\}$ gegeben. Um Eindeutigkeit zu erzielen, wird dann oft $\int_0^1 u(x)\,\mathrm{d}x = 0$ gefordert, womit wir wieder im (iii) der Poincaré/Friedrichs-Ungleichung sind.

Diskretisierung

Wir wollen das Problem diskretisieren. Dazu führen wir wieder ein Gitter

$$\mathcal{T}_h = \{x_1, \cdots, x_n\}, \quad 0 = x_0 < x_1 < \cdots < x_n < x_{n+1} = 1$$

ein. In der Finite Elemente Methode spielen auch die Elemente

$$T^{(k)} = (x_{k-1}, x_k)$$

eine große Rolle. Für die Diskretisierung verwenden wir in diesem Kapitel wieder Polygonzüge:

$$V_h := \{u \in C[0,1] : u|_{\overline{T^{(k)}}} \text{ linear für alle } k \in \{1, \ldots, n+1\}\}.$$

Außerdem definieren wir den Raum der stückweise stetig differenzierbaren Funktionen:

$$C^1_{pw} := \{u \in C[0,1] : u|_{\overline{T^{(k)}}} \in C^1(\overline{T^{(k)}}) \text{ für alle } k \in \{1, \ldots, n+1\}\}. \tag{12.6}$$

Unter Ausnutzung der Stetigkeit von u' und v können wir die Variationsform aus der klassischen Form auch für $v \in C^1_{pw}$ herleiten (Übung!) und erhalten aus der klassischen Formulierung (12.1) – (12.3), dass die Lösung u der klassischen Formulierung des Randwertproblems auch

$$\underbrace{\sum_{i=1}^{n+1} \int_{x_{i-1}}^{x_i} \big(au'v' + cuv\big)\mathrm{d}x + \gamma_1 u(1)v(1)}_{a(u,v):=} = \underbrace{\int_0^1 fv\mathrm{d}x + g_1 v(1)}_{\ell(v):=} \tag{12.7}$$

mit $v \in V_0 := \{v \in C^1_{pw} : v(0) = 0\}$ erfüllt. Dieses Variationsproblem ist für alle $u, v \in C^1_{pw}$ definiert. Für $u, v \in C^1$ stimmt diese Definition von $a(\cdot,\cdot)$ mit der Definition aus (12.4) überein. Daher ist (12.7) eine *Erweiterung* der bisherigen Definition.

Ganz analog gehen wir bei den Normen vor und definieren für $u \in C^1_{pw}$

$$|u|_1 := \left(\sum_{i=1}^{n+1} \int_{x_{i-1}}^{x_i} \big(u'(x)\big)^2 \mathrm{d}x\right)^{1/2}$$

in einer Weise, die mit der bestehenden Definition für $u \in C^1$ übereinstimmt. Die Norm $\|\cdot\|_0$ ist ohnehin bereits so definiert, dass sie für $u \in C^1_{pw}$ Sinn ergibt. Die Norm $\|\cdot\|_1$ ist wieder über $\|u\|_1 := (\|u\|_0^2 + |u|_1^2)^{1/2}$ gegeben.

Wenn man sich den Beweis der Spurungleichung (Satz 12.6) Schritt für Schritt durcharbeitet, wird man feststellen, dass sie auch für $u \in C^1_{pw}$ gilt:

Satz 12.9. *Es gibt eine Konstante $c_S > 0$, sodass für alle $u \in C^1_{pw}$ und alle $x \in [0,1]$ die Abschätzung $|u(x)| \leq c_S \|u\|_1$ gilt.*

Beweis. Als erster Schritt lässt sich die Abschätzung $|u(x_i)|^2 \leq \sum_{j=1}^{n+1} h|u(x_j)|^2 + \sum_{j=1}^{n+1} \int_{x_{i-1}}^{x_i} \big(u'(t)\big)\,\mathrm{d}t$ zeigen. Daraus folgt dann wie im Beweis von Satz 12.6 das gewünschte Resultat, vgl. Übung. □

Analog könnte man auch für den Beweis der Poincaré-Ungleichung (Satz 12.4) vorgehen. Stattdessen führen wir ein allgemeines Resultat ein und verwenden es.

Satz 12.10 (Dichtheit)**.** *Für jedes $v \in C^1_{pw}$ gibt es eine Folge $v_n \in C^1[0,1]$, sodass $\|v - v_n\|_1 \leq \frac{1}{n}$.*

Beweis. Die Idee für einen Beweis wäre, dass die Funktion an den ,,Ecken" entsprechend ,,abgerundet" wird. Für Resultate dieser Art, siehe etwa Adams und Fournier: *Sobolev Spaces.* 2003. Das Argument funktioniert nur, wenn die Funktion v stetig ist. □

Satz 12.11. *Die Poincaré-Ugleichung (Satz 12.4) gilt auch für $u \in C^1_{pw}$.*

Beweis. Wir beschränken uns auf den Fall (i) von Satz 12.4. Sei $u \in C^1_{pw}$ und u_n eine Folge im Sinne des Satzes 12.10. Mit der Dreiecksungleichung sowie Poincaré-Ungleichung (angewandt auf $u_n \in C^1$) und der Dreiecksungleichung erhalten wir

$$\begin{aligned}\|u\|_0 &\leq \|u - u_n\|_0 + \|u_n\|_0 \leq \|u - u_n\|_0 + c_P(|u_n(0)|^2 + |u_n|_1^2)^{1/2} \\ &\leq \|u - u_n\|_0 + c_P((|u(0) - u_n(0)| + |u(0)|)^2 + (|u - u_n|_1 + |u|_1)^2)^{1/2}\end{aligned}$$

Mit der Spurungleichung (Satz 12.9) und der Dichtheit haben wir nun

$$\|u\|_0 \leq \tfrac{1}{n} + c_P((\tfrac{c_S}{n} + |u(0)|)^2 + (\tfrac{1}{n} + |u|_1)^2)^{1/2},$$

also für $n \to \infty$ das gesuchte Resultat. □

Mit diesen Resultaten erhalten wir, dass auch auf C^1_{pw} wie im letzten Abschnitt (i) a bilinear, ℓ linear, (ii) a symmetrisch, (iii) a stetig, (iv) ℓ stetig und (v) a elliptisch ist. Daraus folgt auch wieder Eindeutigkeit einer Lösung. Wir haben nun die erforderliche Vorarbeit geleistet für die

Diskretisierung mit dem Galerkin-Prinzip

Unter Annahme von $u \in C^1[0,1]$ erfüllt die Lösung des Randwertproblems das Variationsproblem

$$\text{Gesucht } u \in V_g\text{:} \quad a(u,v) = \ell(v) \quad \text{für alle } v \in V_0. \tag{12.8}$$

wobei im Falle unseres Modellproblems $V_g := \{u \in C^1_{pw} : u(0) = g_0\}$ und $V_0 = \{v \in C^1_{pw} : v(0) = 0\}$.

Wir ersetzen nun den Raum, in dem wir die diskrete Lösung u_h suchen, durch den Raum der Polygonzüge V_h bzw. eigentlich auf $V_{g,h} := V_g \cap V_h$. Gleichzeitig müssen wir auch die Testfunktionen auf V_h bzw. eigentlich auf $V_{0,h} := V_0 \cap V_h$ beschränken (Galerkin-Prinzip). Dies führt auf das diskretisierte Problem:

$$\text{Gesucht } u_h \in V_{g,h}\text{:} \quad a(u_h,v_h) = \ell(v_h) \quad \text{für alle } v_h \in V_{0,h}. \tag{12.9}$$

Die Beschränkung der Testfunktionen auf den Raum der Polygonzüge lässt sich wieder dadurch erklären, dass die Testfunktionen die zulässigen Variationen von u_h realisieren. Außerdem wäre ohne eine Einschränkung des Raumes der Testfunktionen das diskretisierte Variationsproblem *überbestimmt* und daher im Allgemeinen unlösbar.

Wegen der Koerzivität erhalten wir wieder – wie im Satz 12.7 – die Eindeutigkeit einer Lösung des diskretisierten Variationsproblems (12.9). Für den Beweis der Existenz einer Lösung verwenden wir die Idee der Homogenisierung. Wir stellen wir fest, dass es seine Funktion $g \in V$ gibt, sodass $V_{g,h} = g + V_{0,h}$. Daher ist das diskrete Variationsproblem (12.9) äquivalent zu

$$\text{Gesucht } u_{0,h} \in V_{0,h}\text{:} \quad a(u_{0,h},v_h) = \ell(v_h) - a(g,v_h) \quad \text{für alle } v_h \in V_{0,h},$$

was sich als lineares Gleichungssystem mit gleicher Zahl von Gleichungen und Unbekannten schreiben lässt. Da wir bereits Eindeutigkeit einer Lösung gezeigt haben, wissen wir, dass die resultierende Steifigkeitsmatrix invertierbar sein muss, wir also auch die Existenz einer Lösung haben.

Satz 12.12. *Das diskretisierte Variationsproblem* (12.9) *ist eindeutig lösbar.*

Wegen $V_{0,h} \subset V_0$ erfüllt die Lösung des (kontinuierlichen) Variationsproblems auch $a(u,v_h) = \ell(v_h)$, womit wir sofort die *Galerkin-Orthogonalität* erhalten:

$$a(u - u_h, v_h) = 0 \quad \text{für alle } v_h \in V_{0,h}. \tag{12.10}$$

Wir erinnern uns, dass $a(\cdot,\cdot)$ in unserem Fall ein Skalarprodukt realisiert. Daher besagt die Galerkin-Orthogonalität, dass der Diskretisierungsfehler $u - u_h$ orthogonal auf den Raum $V_{0,h}$ steht, das Galerkin-Prinzip also garantiert, dass u_h die Orthogonalprojektion (bzgl. a) von u auf $V_{0,h}$ ist.

Bemerkung 12.13. Diese geometrische Interpretation setzt die Symmetrie der Bilinearform, also $a(u,v) = a(v,u)$ voraus, für die Herleitung der Formel (12.10) benötigen wir jedoch nur $V_{0,h} \subset V_0$ und die Bilinearität.

Diskretisierungsfehler

Nun ist die Bestimmung des Diskretisierungsfehlers (und damit der Konvergenz) ganz einfach. Wir nutzen dafür folgenden Satz

Satz 12.14 (Lemma von Céa, 1967)**.**
Sei V_0 ein reeller Vektorraum mit Norm $\|\cdot\|_{V_0}$. Wenn die Bilinearform $a(\cdot,\cdot)$

(i) V_0-elliptisch ist, also $a(u,u) \geq \mu_1 \|u\|_{V_0}^2$ für alle $u \in V_0$ gilt, und

(ii) V_0-stetig ist, also $a(u,v) \leq \mu_2 \|u\|_{V_0} \|v\|_{V_0}$ für alle $u, v \in V_0$ gilt,

dann erfüllen die Lösungen $u \in V_g = g + V_0$ von (12.8) und $u_h \in V_{g,h} \subset V_g$ von (12.9) die Abschätzung

$$\|u - u_h\|_{V_0} \leq \tfrac{\mu_2}{\mu_1} \inf_{v_h \in V_{g,h}} \|u - v_h\|_{V_0}. \tag{12.11}$$

Beweis. Aus Koerizivität und Beschränktheit der Bilinearform sowie der Galerkin-Orthogonalität folgt

$$\begin{aligned} \mu_1 \|u - u_h\|_{V_0}^2 &\leq a(u - u_h, u - u_h) \\ &= a(u - u_h, u - v_h) + a(u - u_h, v_h - u_h) \\ &= a(u - u_h, u - v_h) \leq \mu_2 \|u - u_h\|_{V_0} \|u - v_h\|_{V_0}. \end{aligned}$$

Die Abschätzung erhält man dann nach Division durch $\|u - u_h\|_{V_0}$ und Bilden des Infimums über alle $v_h \in V_{g,h}$. □

Für unser Variationsproblem erhalten wir also

$$\|u - u_h\|_1 \leq \frac{\|a\|_\infty + \|c\|_\infty + c_S|\gamma_1|}{\underline{a}(c_P^2 + 1)^{-1}} \inf_{v_h \in V_h} \|u - u_h\|_1.$$

Wir können demnach den Diskretisierungsfehler (=Fehler, den wir bei der Diskretisierung machen) $\|u - u_h\|_1$ mit dem Approximationsfehler (=Fehler der bestmöglichen Approximation) $\inf_{v_h \in V_h} \|u - u_h\|_1$ abschätzen. Die bestmögliche Approximation kann mit jeder erdenklichen Approximation abgeschätzt werden, etwa durch den Interpolationsfehler $\|u - r_h u\|_1$. Diesen können wir weiter abschätzen.

Satz 12.15 (Interpolationsfehler)**.** *Es gibt ein c_{int}, sodass für alle $u \in C^2[0,1]$*

(i) $|u - r_h\, u|_1 \le c_{int} h \|u''\|_0$ *und*

(ii) $\|u - r_h\, u\|_0 \le c_{int} h^2 \|u''\|_0$.

Beweis. Wir verwenden wieder die Transformation $\psi_k(y) = x_{k-1} + y(x_k - x_{k-1})$ von $\widehat{T} := (0,1)$ auf $T^{(k)} = (x_{k-1}, x_k)$. Wir beschränken uns auf (i). Sei $v \in C^2$ und $v_h(y) := v(0)(1-y) + v(1)y$. Die Taylorformel mit Integralrestglied $v(y) = v(0) + yv'(0) + \int_0^y tv''(t)\mathrm{d}t$ und die Cauchy-Schwarz'schen Ungleichung ergeben

$$(v(y) - v_h(y))^2 = \left(\int_0^y tv''(t)\mathrm{d}t - y \int_0^1 tv''(t)\mathrm{d}t \right)^2 \le \int_0^1 \big(v''(y)\big)^2 \mathrm{d}y$$

und daher $\int_0^1 (v(y) - v_h(y))^2 \mathrm{d}y \le \int_0^1 \big(v''(y)\big)^2 \mathrm{d}y$. Daraus erhalten wir mit Substitutions- und Kettenregel für $v(y) = u(\psi_k(y))$ und $v_h(y) = u_h(\psi_k(y))$

$$\int_{x_{k-1}}^{x_k} (u(x) - r_h u(x))^2 \mathrm{d}x \le \int_{x_{k-1}}^{x_k} \big(u''(x)\big)^2 \mathrm{d}x.$$

Das gewünschte Resultat erhalten wir durch Bilden der Summe. □

Mit diesem Satz erhalten wir sofort folgendes Konvergenzresultat:

$$\|u - u_h\|_1 \le \sqrt{2} c_{int} \frac{\|a\|_\infty + \|c\|_\infty + c_S |\gamma_1|}{\underline{a}(c_P^2 + 1)^{-1}}\, h \|u''\|_0 = \mathcal{O}(h).$$

Bemerkung 12.16. (i) Wir erhalten für die FEM nun ein Konvergenzresultat der Art $\|u - u_h\|_1 = \mathcal{O}(h)$, wie wir dies im letzten Kapitel für die FDM auch erhalten hatten. Als Regulariätsresultat benötigen wir nur, dass $u \in C^2$ (also $f \in C^0$), währenddessen wir im letzten Kapitel $f \in C^1$ benötigt hatten. Da der Beweis von Satz 12.15 jeweils lokal auf den einzelnen Elementen war, reicht es sogar aus, wenn u nur elementweise in C^2 liegt. Im nächsten Kapitel werden wir sehen, dass wir dafür auch einen Preis zahlen müssen (vgl. Bemerkung 13.6).

(ii) Mit dem *Aubin-Nitsche-Dualitätsargument*, das wir in der VL *Numerische Methoden für elliptische Probleme* näher behandeln werden lässt sich das verbesserte Konvergenzresultat $\|u - u_h\|_0 \le Ch^2 \|u''\|_0 = \mathcal{O}(h^2)$ zeigen.

(iii) Insbesondere bei der Behandlung der (hier: *natürlichen*) Robin- und Neumann-Randbedingungen hat sich die variationelle Herangehensweise besonders bezahlt gemacht; bei der FDM hatten uns gerade diese einige Anstrengungen gekostet.

(iv) Bisher hatten wir mit unserer ganz intuitiven Einführung der Variationsprobleme die Eindeutigkeit der Lösungen zeigen können, außerdem auch Diskretisierungsfehlerabschätzungen (Konvergenz). Für Existenzresultate müssen wir aber viel härter arbeiten, siehe übernächstes Kapitel.

13. Implementierung der Finiten Elemente Methode

Im Folgenden diskutieren wir in einigem Detail das generelle Vorgehen bei der Implementierung der Finite-Elemente Methode. Für die folgenden Überlegungen, betrachten wir das einfachere Modellproblem

$$-(au')' + cu = f \qquad \text{in } (0,1) \tag{13.1}$$

mit Robin Randbedingungen

$$-a(0)u'(0) + \gamma_0 u(0) = g_0 \tag{13.2}$$
$$a(1)u'(1) + \gamma_1 u(1) = g_1. \tag{13.3}$$

Eine Verallgemeinerung auf andere Randbedingungen ist ohne Probleme möglich. Für die Diskretisierung wählen wir wieder ein Gitter

$$\mathcal{T}_h = \{x_1, \cdots, x_n\}, \quad 0 = x_0 < x_1 < \cdots < x_n < x_{n+1} = 1,$$

die zugehörigen Elemente

$$T^{(k)} := (x_{k-1}, x_k) \quad \text{der Länge} \quad h_k := x_k - x_{k-1},$$

die wir von einem Referenzelement $\hat{T} := (0,1)$ mittels $\psi_k(y) := x_{k-1} + yh_k$ parameterisieren können. Für die Diskretisierung wollen wir wieder V_h, den Raum der Polygonzüge wählen. Dies entspricht der Wahl linearer Basisfunktionen auf jedem der Elemente; diese sind durch Vorgabe der Funktionswerte an den Gitterpunkten x_i eindeutig bestimmt. Die Gitterpunkte spielen hier die Rolle von Knoten ($t_i = x_i$); vgl. zum Unterschied zwischen den Gitterpunkte x_i und den Knoten t_i das Beispiel am Ende dieses Kapitels.

Der Galerkin-Isomorphismus

Wir betrachten ein diskretisiertes Variationsproblem

$$\text{Gesucht } u_h \in V_h\text{:} \qquad a(u_h, v_h) = \ell(v_h) \qquad \forall v_h \in V_h. \tag{13.4}$$

Sei nun $\{\phi_i : i = 0, \cdots, n+1\} \subset V_h$ eine Basis von V_h und folglich $\dim(V_h) = n+2$. Dann lässt sich jede Funktion $v_h \in V_h$ darstellen als

$$v_h = \sum_{i=0}^{n+1} v_i \phi_i$$

mit eindeutig bestimmten Koeffizientenvektor $\underline{v}_h = (v_0, \ldots, v_{n+1})^\top \in \mathbb{R}^{n+2}$. Die Abbildung

$$\mathbb{R}^{n+2} \to V_h, \qquad \underline{v}_h \mapsto v_h = \sum_{i=0}^{n+1} v_i \phi_i \tag{13.5}$$

ist linear und bijektiv und wird *Galerkin-Isomorphismus* genannt.

Satz 13.1 (Galerkin-Isomorphismus)**.**
Das diskrete Variationsproblem (13.4) *ist äquivalent zum Gleichungssystem*

$$A_h \, \underline{u}_h = \underline{b}_h \tag{13.6}$$

mit

(i) der Steifigkeitsmatrix $A_h = (a_{i,j})_{i,j=0}^{n+1}$ *mit Koeffizeinten* $a_{i,j} = a(\phi_j, \phi_i)$,

(ii) dem Lastvektor $\underline{b}_h = (b_i)_{i=0}^{n+1}$ *mit Koeffizienten* $b_i = \ell(\phi_i)$ *und*

(iii) dem Lösungsvektor $\underline{u}_h = (u_j)_{j=0}^{n+1}$, *der die Lösung* $u_h = \sum_{j=0}^{n+1} u_j \phi_j$ *bestimmt.*

Beweis. Die Gleichung für den Index i folgt durch Testen von (13.4) mit der speziellen Testfunktion $v_h = \phi_i$. Wegen Linearität reicht es aus, nur die Basisfunktionen ϕ_i als Testfunktionen zu wählen. □

Bemerkung 13.2.
(i) Die konkrete Form von A_h und $\underline{b}_h$ hängt von der Wahl der Basis ab!

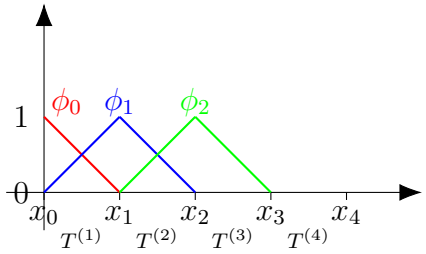
(ii) Wichtige Eigenschaften der Bilinearform $a(\cdot, \cdot)$ übertragen sich – wie wir dies bereits in Kapitel 11 gesehen haben – unmittelbar auf die Matrix A_h. So gilt etwa

- $a(\cdot, \cdot)$ koerziv $\Rightarrow$ A_h positiv definit, d.h. $\underline{v}_h^\top A_h \, \underline{v}_h > 0$ für alle $\underline{v}_h \neq 0$.
- $a(\cdot, \cdot)$ symmetrisch $\Rightarrow$ A_h symmetrisch, d.h., $A_h^\top = A_h$.

Als Basis wird typischerweise eine *Knotenbasis* verwendet, also $\phi_i \in V_h$ mit $\phi_i(x_j) = \delta_{i,j}$, wobei $\delta_{i,j}$ das Kronecker-Delta ist. Für den Raum der Polygonzüge ergeben sich dadurch *Hütchenfunktionen*:

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

Damit erhalten wir:



Da $\dim(V_h) = n+2$ und dies mit der Zahl der Hütchenfunktionen übereinstimmt und die Hütchenfunktionen linear unabhängig sind (siehe Übung), bilden die Hütchenfunktionen eine Basis von V_h (Hütchenbasis).

Aus der Skizze erhalten wir sofort

$$\phi_i(x) \equiv 0 \qquad \text{auf } T^{(k)} \qquad \text{für} \qquad i \notin \{k-1, k\}. \tag{13.7}$$

Assemblierung der Steifigkeitsmatrix

Nach den Vorüberlegungen ist das diskretisierte Variationsproblem äquivalent zu einem linearen Gleichungssystem (13.6). Wir diskutieren jetzt noch genauer, wie sich die Einträge $a_{i,j} = a(\phi_j, \phi_i)$ der Steifigkeitsmatrix und $b_i = \ell(\phi_i)$ des Lastvektors in der Praxis berechnen lassen. Für unser Modellproblem gilt etwa

$$\begin{aligned} a_{i,j} &= a(\phi_j, \phi_i) \\ &= \int_0^1 a(x)\phi_j'(x)\phi_i'(x)\,\mathrm{d}x + \int_0^1 c(x)\phi_j(x)\phi_i(x)\,\mathrm{d}x + \sum_{\delta=0}^{1} \gamma_\delta \phi_j(\delta)\phi_i(\delta) \\ &= \sum_{k=1}^{n+1} \underbrace{\int_{x_{k-1}}^{x_k} a(x)\phi_j'(x)\phi_i'(x)\,\mathrm{d}x}_{=:k_{i,j}^{(k)}} + \sum_{k=1}^{n+1} \underbrace{\int_{x_{k-1}}^{x_k} c(x)\phi_j(x)\phi_i(x)\,\mathrm{d}x}_{=:m_{i,j}^{(k)}} + \sum_{\delta=0}^{1} \underbrace{\gamma_\delta \phi_j(\delta)\phi_i(\delta)}_{=:r_{i,j}^{(\delta)}}. \end{aligned}$$

Die Einträge $a_{i,j}$ lassen sich also durch *Aufsummieren* (=*Assemblieren*) der lokalen Beiträge $k_{i,j}^{(k)}$, $m_{i,j}^{(k)}$ und $r_{i,j}^{(\delta)}$ berechnen.

Wegen (13.7) brauchen bei der Berechnung der Beiträge $k_{i,j}^{(k)}$ zum $k$ten Element $T^{(k)}$ nur die Basisfunktionen mit $i, j \in \{k-1, k\}$ berücksichtigt werden. Die entsprechenden Einträge lassen sich in einer 2×2 Matrix $K_h^{(k)} := (k_{k-2+\alpha,k-2+\beta}^{(k)})_{\alpha,\beta=1}^{2}$ speichern. Man erhält unter der Annahme, dass $a|_{T^{(k)}} \equiv a_k$ konstant ist, für $i, j \in \{k-1, k\}$ die Elementbeiträge

$$k_{i,j}^{(k)} = \int_{x_{k-1}}^{x_k} a(x)\partial_x\phi_m(x)\partial_x\phi_n(x)\,\mathrm{d}x = \frac{a_k}{h_k}\int_0^1 \partial_y\hat{\phi}_\alpha(y)\partial_y\hat{\phi}_\beta(y)\,dy.$$

Dabei haben wir im zweiten Schritt einfach mittels $x = \psi_k(y) = x_{k-1} + yh_k$ substitutiert; vgl. Abschätzung des Interpolationsfehlers. Weiters haben wir die globalen Knotennummern $i, j \in \{k-1, k\}$ mit den lokalen $\alpha, \beta \in \{1, 2\}$ identifiziert und

$$\hat{\phi}_1(y) = 1 - y \qquad \text{und} \qquad \hat{\phi}_2(y) = y$$

verwendet; siehe Skizze. Dabei handelt es sich um die Basisfunktionen am Referenzelement $\hat{T} = (0, 1)$. Die Einträge der Elementmatrix sind dann gegeben durch

$$K_h^{(k)} = \frac{a_k}{h_k} \begin{pmatrix} \int_0^1 \partial_y(1-y)\,\partial_y(1-y)\,\mathrm{d}y & \int_0^1 \partial_y y\,\partial_y(1-y)\,\mathrm{d}y \\ \int_0^1 \partial_y(1-y)\,\partial_y y\,\mathrm{d}y & \int_0^1 \partial_y y\,\partial_y y\,\mathrm{d}y \end{pmatrix} = \frac{a_k}{h_k} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}.$$

Diese müssen beim Assemblieren an der richtigen Stelle dazuaddiert werden. Nach Assemblieren erhalten wir:

$$K_h = \begin{pmatrix} \frac{a_1}{h_1} & -\frac{a_1}{h_1} & & & \\ -\frac{a_1}{h_1} & \frac{a_1}{h_1} + \frac{a_2}{h_2} & -\frac{a_2}{h_2} & & \\ & -\frac{a_2}{h_2} & \ddots & \ddots & \\ & & \ddots & \frac{a_n}{h_n} + \frac{a_{n+1}}{h_{n+1}} & -\frac{a_{n+1}}{h_{n+1}} \\ & & & -\frac{a_{n+1}}{h_{n+1}} & \frac{a_{n+1}}{h_{n+1}} \end{pmatrix}.$$

Diese Matrix ist symmetrisch und positiv semi-definit. Für $h_1 = \cdots = h_{n+1} = h$ and $a_1 = \cdots = a_{n+1} = a$ vereinfacht sich das zu

$$K_h = \frac{a}{h} \begin{pmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & \ddots & \ddots & \\ & & \ddots & 2 & -1 \\ & & & -1 & 1 \end{pmatrix}.$$

Ganz analog können wir $m_{i,j}^{(k)}$ unter der Annahme $c|_{T^{(k)}} \equiv c_k$ bestimmen; dabei erhalten wir als lokale Gram-Matrix bzw. lokale Masse-Matrix $M_h^{(k)} := (m_{k-2+\alpha,k-2+\beta}^{(k)})_{\alpha,\beta=1}^2$:

$$M_h^{(k)} = c_k h_k \begin{pmatrix} \int_0^1 (1-y)\,(1-y)\,\mathrm{d}y & \int_0^1 y\,(1-y)\,\mathrm{d}y \\ \int_0^1 (1-y)\,y\,\mathrm{d}y & \int_0^1 y\,y\,\mathrm{d}y \end{pmatrix} = \frac{c_k h_k}{6} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

Nach dem Assemblieren erhalten wir:

$$M_h = \begin{pmatrix} \frac{2c_1h_1}{6} & \frac{c_1h_1}{6} & & & \\ \frac{c_1h_1}{6} & \frac{2c_1h_1}{6} + \frac{2c_2h_2}{6} & \frac{c_2h_2}{6} & & \\ & \frac{c_2h_2}{6} & \ddots & \ddots & \\ & & \ddots & \frac{2c_nh_n}{6} + \frac{2c_{n+1}h_{n+1}}{6} & \frac{c_{n+1}h_{n+1}}{6} \\ & & & \frac{c_{n+1}h_{n+1}}{6} & \frac{2c_{n+1}h_{n+1}}{6} \end{pmatrix}.$$

Diese Matrix ist symmetrisch und (wenn $c \geq 0$) positiv semi-definit. Für $c(x) \equiv c$ konstant und $h_1 = \cdots = h_{n+1} = h$ vereinfacht sich das zu

$$M_h = \frac{ch}{6} \begin{pmatrix} 2 & 1 & & & \\ 1 & 4 & 1 & & \\ & 1 & \ddots & \ddots & \\ & & \ddots & 4 & 1 \\ & & & 1 & 2 \end{pmatrix}.$$

In Summe erhalten wir

$$A_h = K_h + M_h + \begin{pmatrix} \gamma_0 & & & & \\ & 0 & & & \\ & & \ddots & & \\ & & & 0 & \\ & & & & \gamma_1 \end{pmatrix}.$$

Bemerkung 13.3. Die hier beschriebene *elementweise Assemblierung* beschreibt das Verfahren, das üblicherweise für die Realisierung der Finiten Elemente Methode verwendet wird.

Bemerkung 13.4. Wenn a oder c nicht auf den einzelnen Elementen $T^{(k)}$ konstant ist, dann sind die Integrale nicht so leicht aufzulösen. In der Praxis nähert man aber dann das Integral durch eine geeignete Quadraturformel an oder approximiert a bzw. c auf $T^{(k)}$ einfach durch einen konstanten Wert (etwa nach der Mittelpunksregel). Damit führt man natürlich einen weiteren Fehlerterm ein. Wenn a bzw. c hinreichend glatt ist, kann man zeigen (Lemma von Strang), dass durch geignete Annäherunen die Konvergenzordnung des Verfahrens nicht zerstört wird. Mehr dazu wird in der VL *Numerische Methoden für elliptische Probleme* behandelt werden.

Bemerkung 13.5 (Andere Randbedingungen)**.**
(i) Die hier besprochenen Robin-Randbedingungen schließen den Fall von Neumann-Bedingungen ein ($\gamma_0 = 0$ bzw. $\gamma_1 = 0$).

(ii) Dirichlet-Randbedingungen lassen sich mit $\gamma_0 = \epsilon^{-1}$ bzw. $\gamma_1 = \epsilon^{-1}$ mit $\epsilon > 0$ klein anähern; dies ist programmiertechnisch sehr einfach, die Konditionszahl der Steifigkeitsmatrix steigt jedoch mit ϵ^{-1} an. In der Praxis ist dies kein Problem.

(iii) Alternativ können die Dirichlet-Randbedingungen auch durch Elimination der entsprechenden Freiheitsgrade (wie in der Definition der Räume) realisieren. Da die Werte der Lösung u_h am Dirichlet-Rand bekannt sind, entfallen die

zugehörigen Zeilen. Da die Testfunktion v_h am Dirichtlet-Rand verschwindet, entfallen die zugehörigen Spalten. Während dies im 1D-Fall noch leicht programmiertechnisch umgesetzt werden kann, ist dies bei mehr Dimensionen aufwändiger zu realisieren.

Assemblierung des Lastvektors

Die Einträge des Lastenvektors können wieder über eine elementweise Assemblierung errechnet werden. Wir erhalten

$$b_i = \sum_{k=1}^{n+1} \underbrace{\int_{x_{k-1}}^{x_k} f(x)\phi_i(x)\,dx}_{=:f_i^{(k)}} + \underbrace{g_0\phi_i(0)}_{=:g_i^{(0)}} + \underbrace{g_1\phi_i(1)}_{=:g_i^{(1)}}.$$

Bemerkung 13.6 (Numerische Integration)**.** Falls die Funktion f keine einfache Gestalt hat, lassen sich die Elementintegrale $f_i^{(k)}$ nicht so ohne weiteres berechnen. Wir können dann numerische Integration verwenden, z.B. die Trapezregel

$$f_i^{(k)} \approx h_k \left(\tfrac{1}{2} f(x_{k-1})\phi_i(x_{k-1}) + \frac{1}{2} f(x_k)\phi_i(x_k) \right);$$

mit dieser Wahl erhält man

$$\underline{b}_h = \begin{pmatrix} g_0 + \frac{h_1}{2} f(x_0) \\ \frac{h_1+h_2}{2} f(x_1) \\ \vdots \\ \frac{h_n+h_{n+1}}{2} f(x_n) \\ g_1 + \frac{h_{n+1}}{2} f(x_{n+1}) \end{pmatrix}, \text{ also für äquidistante Gitter } \underline{b}_h = \begin{pmatrix} g_0 + \frac{h}{2} f(x_0) \\ h f(x_1) \\ \vdots \\ h f(x_n) \\ g_1 + \frac{h}{2} f(x_{n+1}) \end{pmatrix}.$$

Wenn f elementweise stetig differenzierbar ist ($f|_{T^{(k)}} \in C^1$), kann zeigen (Lemma von Strang), dass dadurch die Konvergenzordnung des Verfahrens nicht zerstört wird; das kann man auch mit numerischen Tests überprüfen. Mehr dazu wird in der VL *Numerik elliptische Probleme* behandelt werden.

Abschließende Bemerkungen

Vergleich mit der Finite-Differenzen Methode

Zur Einordnung der Resultate diskutieren wir noch kurz einen Zusammenhang mit der Finite Differenzen Methode. Für das Modellproblem

$$-u'' = f \qquad \text{in } \Omega = (0,1),$$

mit Robin-Randbedingungen

$$\begin{aligned} -u'(0) + \gamma_0 u(0) &= g_0 \\ u'(1) + \gamma_1 u(1) &= g_1. \end{aligned}$$

erhalten wir mit der Finiten Elemente Methode (FEM) und der Finite Differenzen Methode (FDM) – bis auf die Skalierung – eine völlig gleich aussehende Steifigkeitsmatrix und einen völlig gleich aussehenden Lastvektor. (Dies setzt voraus, dass in der FDM die Robin-Randbedingungen analog zu den Ausführungen zu Beispiel 8.21 so umgesetzt werden, dass Konsistenzordnung 2 erreicht wird!) Damit geben uns beide Methoden auch dieselbe Lösung.

Beim Problem $-u'' + cu = f$ erhalten wir einen Unterschied, da die FEM den Term cu mit der Masse-Matrix, die FDM jedoch mit einer Diagonalmatrix realisiert. In Kapitel 11 hatten wir gesehen, dass wir auch für die FDM eine Fehleranalyse in den Normen $\|\cdot\|_0$ und $\|\cdot\|_1$ machen kann. Umgekehrt kann man aber auch auf die FEM-Matrizen die aus der FDM stammende, auf dem Maximumsprinzip aufbauende Konvergenztheorie anwenden, wodurch auch Konvergenzresultate in der Maximumsnorm erzielt werden können. Die Voraussetzungen für das diskrete Maximumsprinzip sind erfüllt, wenn $h_k < \sqrt{6/c_k}$, also die Gitterweite so klein ist, dass Nebendiagonaleinträge negativ sind.

Methoden höherer Ordnung

Im ersten Teil der Vorlesung hatten wir gesehen, dass auch eine Konstruktion von Methoden höherer Ordnung möglich und oft auch gewünscht ist. Mit der Finiten Elemente Methode lässt sich dies auch für Randwertprobleme leicht realisieren.

Wir verwenden dazu auf den einzelnen Elementen $T^{(k)}$ einfach Polynome vom Grad $p \geq 1$, wobei wir weiterhin benötigen, dass die konstruierten Funktionen in C^1_{pw} liegen, also stetig sind. Um ein Polynom vom Grad p eindeutig zu bestimmen, müssen wir den Funktionswert an $p+1$ Punkten (=Knoten) vorgeben. Auf dem Referenzelement $\widehat{T}$ können wir diese Knoten einfach äquidistant verteilen. Um globale Stetigkeit garantieren zu können, muss jeweils ein Knoten auf jedem der beiden Ränder liegen.

Um die Notation einfach zu halten, diskutieren wir den Rest nur für $p = 2$. Durch äquidistante Verteilung der Knoten erhalten wir $\hat{t}_1 = 0$, $\hat{t}_2 = \frac{1}{2}$, $\hat{t}_3 = 1$. (Global sind demnach die Knoten $0 = t_0, \ldots, t_{2n+1} = 1$ durch $t_{2i-1} = x_i$ und $t_{2i} = (x_i + x_{i+1})/2$ gegeben.) Dementsprechend haben wir bei Nutzung einer Knotenbasis 3 Basisfunktionen am Element, die wir mit der Lagrange'schen Interpolationsformel bestimmen können:

$$\hat{\phi}_1(y) = 2(y - \tfrac{1}{2})(y - 1), \quad \hat{\phi}_2(y) = -4(y - 0)(y - 1), \quad \hat{\phi}_3(y) = 2(y - 0)(y - \tfrac{1}{2}).$$

Für $a \equiv 1$, $c \equiv 1$ erhalten wir die lokalen Steifigkeits- und Massematrizen:

$$K_h^{(k)} = \tfrac{1}{3} \begin{pmatrix} 7 & -8 & 1 \\ -8 & 16 & -8 \\ 1 & -8 & 7 \end{pmatrix}, \qquad M_h^{(k)} = \tfrac{1}{30} \begin{pmatrix} 4 & 2 & -1 \\ 2 & 16 & 2 \\ -1 & 2 & 4 \end{pmatrix},$$

die wir wieder K_h und M_h assemblieren können; die resultierenden Matrizen sind Bandmatrizen mit bis zu $p + 1$ nichtverschwindenden Einträgen pro Zeile. Auch für den Lastvektor $\underline{b}_h$ können wir ganz analog vorgehen. Der Lösungsvektor $\underline{u}_h$ gibt uns dann die Funktionswerte an den einzelnen Knoten, die Lösung u_h ist eine stetige und elementweise quadratische Funktion.

Die allgemeinen theoretischen Aussagen gelten auch für diese Diskretisierungen, insbesondere auch die Aussagen zu Existenz und Eindeutigkeit einer Lösung zum diskretisierten Problem und das Lemma von Céa. Analog zu Satz 12.15 kann man Interpolationsfehlerabschätzungen der Art

$$\|u - r_h\, u\|_1 \leq c_{int} h^p \left(\sum\nolimits_{k=1}^{n+1} \int_{T^{(k)}} (\partial^{p+1} u)^2 \mathrm{d}x \right)^{1/2}$$

zeigen; man erhält also, wenn u auf jedem Element $p+1$ mal stetig differenzierbar ist, dass $\|u - u_h\|_1 = \mathcal{O}(h^p)$.

14. Existenz von Variationslösungen

Der Beweis der Existenz einer klassischen Lösung zu Randwertproblemen ist im allgemeinen Fall kein leichtes Unterfangen. Während man im eindimensionalen Fall noch mit dem Satz von Picard-Lindelöf argumentieren kann (was wir auch taten), sind vergleichbare Resultate für den mehrdimensionalen Fall nur in Spezialfällen machbar.

Alternativ kann man die Existenz einer Lösung eines Variationsproblems zeigen. Wenn man dann zusätzliche Regularitätsresultate hat, kann man eventuell auch zeigen, dass die Lösung der variationellen Formulierung eines Randwertprobelms auch eine Lösung der klassischen Formulierung ist. Es kann jedoch auch vorkommen, dass eine klassische Lösung gar nicht existiert. Auch in diesem Fall hat die variationelle Formulierung eine physikalische Bedeutung (etwa bei der Wärmeleitgleichung mit stückweise konstanten Wärmeleitkoeffizienten).

Satz von Lax-Milgram

Der Beweis der Existenz einer Variationslösung ist mit dem Satz von Lax-Milgram möglich. Dieser setzt vollständige normierte Räume voraus, die wir nun einführen.

Definition 14.1. (i) Ein normierter Raum $(V, \|\cdot\|_V)$ ist ein *Banach-Raum*, wenn der Raum vollständig ist, also alle Cauchy-Folgen konvergieren.

(ii) Ein Banach-Raum ist ein *Hilbert-Raum*, wenn die Norm durch ein Skalarprodukt induziert ($\|\cdot\|_V^2 = (\cdot,\cdot)_V$) wird.

Satz 14.2 (Lax-Milgram, 1957)**.** *Sei V ein reeller Hilbert-Raum mit Norm $\|\cdot\|_V$. Wenn die* Bilinearform $a(\cdot,\cdot) : V \times V \to \mathbb{R}$

(i) V-elliptisch ist, also $a(u,u) \geq \mu_1 \|u\|_V^2$ für alle $u \in V$ gilt, und

(ii) V-stetig ist, also $a(u,v) \leq \mu_2 \|u\|_V \|v\|_V$ für alle $u, v \in V$ gilt,

und das lineare *Funktional $\ell : V \to \mathbb{R}$*

(iii) V-stetig ist, also $\ell(v) \leq \mu_3$ für alle $v \in V$ gilt,

dann gibt es ein eindutiges $u \in V$ mit

$$a(u, v) = \ell(v) \qquad \forall v \in V,$$

und weiters gilt die a-priori Abschätzung $\|u\|_V \leq \frac{\mu_3}{\mu_1}$.

Der Beweis basiert am Banach'schen Fixpunktsatz, wobei für die Formulierung der Fixpunktiteration der Riesz-Isomorphismus verwendet wird (dessen Existenz aus dem Riesz'schen Darstellungssatz folgt). Daher benötigen wir die Vollständigkeit des Funktionenraumes!

Wie bereits das Lemma von Céa benötigt auch der Satz von Lax-Milgram nicht die Symmetrie von a.

Vollständige Funktionenräume

Die Herleitung der Räume und die Eigenschaften, die wir im Folgenden diskutieren, sollten Ihnen bereits aus den VL *Funktionalanalysis* und *Partielle Differentialgleichungen* bekannt sein. Für die Resultate sei auf das Buch Adams und Fournier: *Sobolev Spaces.* 2003. verwiesen.

Der Lebesgue-Raum L^2

Für die Konstruktion des Raumes L^2 sind folgende Überlegungen hilfreich:

(i) Der Raum $C[0, 1]$ ist gemeinsam mit der durch $\|u\|_0 = \left(\int_0^1 u(x)^2 \, \mathrm{d}x\right)^{1/2}$ gegebenen Norm nicht vollständig, vgl. etwa die Funktionenfolge

$$u_n(x) := \max\{-1, \min\{1, n(x - \tfrac{1}{2})\}\},$$

die punktweise gegen die unstetige Funktion $u(x) = \operatorname{sign}(x - \frac{1}{2})$ konvergiert (Skizze). Diese Funktionenfolge erfüllt auch $\|u_n - u_m\|_0 \to 0$ für $n, m \to 0$; sie ist also eine Cauchy-Folge. Da wir an einem vollständigen Raum interessiert sind, muss der Funktionenraum unstetige Funktionen zulassen.

(ii) Lässt man unstetige Funktionen zu, so ist nicht klar, dass $\|u\|_0$ wohldefiniert ist; daher schränkt man auf alle Funktionen ein, deren Quadrat Lebesgueingegrierbar mit endlichem Integral ist.

(iii) Mit dieser Wahl ist $\|\cdot\|_0$ keine Norm mehr, sondern nur mehr eine Seminorm. So gilt etwa für

$$w(x) := \begin{cases} 1 & \text{für } x = 1/2 \\ 0 & \text{sonst} \end{cases},$$

dass $\|w\|_0 = 0$, obwohl offensichtlich $w \neq 0$. Generell gilt, dass die Norm aller Funktionen verschwindet, die nur auf einer Menge mit Maß 0 (Nullmenge) einen von 0 verschiedenen Wert haben. Um die Norm-Eigenschaften wiederzuerlangen, definiert man den Funktionenraum dann als Menge von Äquivalenzklassen, wobei Funktionen assoziiert werden, die sich nur auf einer Nullmenge unterscheiden.

Definition 14.3. (i) Wir nennen eine Funktion $u : \Omega \to \mathbb{R}$ *quadratisch integrierbar*, wenn u^2 Lebesque-integrierbar mit endlichem Integral ($\int_\Omega u^2(x)\mathrm{d}x < \infty$) ist.

(ii) Der Raum $L^2(\Omega)$ ist der Raum von Äquvialenzklassen von quadratisch integrierbaren Funktionen, die sich jeweils nur auf einer Nullmenge unterscheiden.

Der Raum L^2 ist bezüglich der Norm $\|u\|_{L^2(0,1)} = \|u\|_0 = \left(\int_0^1 u^2\mathrm{d}x\right)^{1/2}$ vollständig, daher ist er ein Hilbert-Raum. Integrale von Termen, die eine L^2-Funktion enthalten, sind wohldefiniert, wenn sie für alle Elemente der Äquivalenzklasse denselben Wert haben; in der Regel reicht es aus, mit einem Repräsentanten zu rechnen.

Schwache Ableitungen und der Sobolev-Raum H^1

Die Idee des Sobolev-Raums H^1 ist es, einen Teilraum von L^2 zu finden, auf dem wir einen Ableitungsbegriff definieren können.

Sei $C_0^\infty(0,1)$ die Menge der unendlich oft differenzierbaren Funktionen, welche in einer Umgebung des Randes $\{0,1\}$ verschwinden. Für jedes $\phi \in C_0^\infty(0,1)$ gibt es dann eine kompakte Menge $K \subset (0,1)$, sodass $\phi(x) = 0$ für alle $x \in (0,1) \setminus K$. Diese Funktionen haben also kompakten Träger. Für jede Funktion $v \in C^1[0,1]$ gilt nun offensichtlich nach partieller Integration

$$\int_0^1 v'(x)\phi(x)\,\mathrm{d}x = -\int_0^1 v(x)\phi'(x)\,\mathrm{d}x \qquad \forall \phi \in C_0^\infty(0,1). \tag{14.1}$$

Die Randterme bei der partiellen Differentiation verschwinden, da ja ϕ samt seinen Ableitungen dort 0 ist. Weiters macht die rechte Seite obiger Gleichung Sinn, solange v zumindest lokal integrierbar ist. (Eine Funktion v heißt *lokal integrierbar*, kurz $v \in L^1_{loc}(0,1)$, falls sie auf allen kompakten Mengen $K \subset (0,1)$ integrierbar ist, also $v|_K \in L^1(K)$.) Dies führt auf folgenden Begriff.

Definition 14.4 (Schwache Ableitung)**.** Ein lokal integrierbares $v : (0,1) \to \mathbb{R}$ heißt *schwach differenzierbar*, falls es ein lokal integrierbares $w : (0,1) \to \mathbb{R}$ gibt mit

$$\int_0^1 w(x)\phi(x)\,\mathrm{d}x = -\int_0^1 v(x)\phi'(x)\,\mathrm{d}x \qquad \forall \phi \in C_0^\infty(0,1). \tag{14.2}$$

Die Funktion w heißt *schwache Ableitung* von v und wird mit $w = v'$ oder $w = \partial_x v$ bezeichnet.

Satz 14.5. *Eine stückweise stetig differenzierbare Funktion ist genau dann schwach differenzierbar, wenn sie stetig ist. Die schwache Ableitung entspricht dann der stückweisen. (Demnach sind Funktionen in C^1_{pw} im Sinne von* (12.6) *stets schwach differenzierbar.)*

Beweis. Ohne Beschränkung der Allgemeinheit genügt es, den Fall *einer* Sprungstelle, z.B. bei $x_1 \in (0,1)$ zu betrachten. Der einzig sinnvolle Kandidat für die schwache Ableitung ist dann offensichtlich

$$w(x) = \begin{cases} v'(x), & x \in (0, x_1), \\ v'(x), & x \in (x_1, 1). \end{cases}$$

Über die Werte von w an einzelnen Punkten (etwa x_1) müssen wir uns keine Gedanken machen. Das sieht man, indem man auf einem Teilintervall partiell integriert. Mit elementarer Rechnung folgt nun

$$\begin{aligned} -\int_0^1 v(x)\phi'(x)\,\mathrm{d}x &= -\int_0^{x_1} v(x)\phi'(x)\,\mathrm{d}x - \int_{x_1}^1 v(x)\phi'(x)\,\mathrm{d}x \\ &= \int_0^{x_1} v'(x)\phi'(x)\,\mathrm{d}x + \int_{x_1}^1 v'(x)\phi(x)\,\mathrm{d}x - v(x_1^-)\phi(x_1^-) + v(x_1^+)\phi(x_1^+), \end{aligned}$$

wobei wir mit $f(x_1^\pm) = \lim_{\epsilon\to 0} f(x_1 \pm \epsilon)$ wie üblich den Grenzwert von unten bzw. oben bezeichnen und wir $\phi(0) = \phi(1) = 0$ nutzen. Da ϕ stetig ist, folgt durch Zusammenziehen der Integrale

$$-\int_0^1 v(x)\phi'(x)\,\mathrm{d}x = \int_0^1 w(x)\phi(x)\,\mathrm{d}x + (v(x_1^+) - v(x_1^-))\phi(x_1).$$

Der störende Sprungterm verschwindet genau dann, wenn v stetig ist. □

Der Sobolev-Raum H^1 umfasst alle Funktionen (Äquivalenzklassen) in L^2, deren schwache Ableitung ebenfalls in L^2 liegt:

Definition 14.6 (Sobolev-Raum H^1)**.** Wir definieren $H^1 := \{u \in L^2 : u' \in L^2\}$, wobei u' die schwache Ableitung bezeichnet.

Dieses Resultat gilt jeweils gleichermaßen für alle Elemente der Äquivalenzklassen (die sich ja nur auf für das Integral irrelevanten Nullmengen unterscheiden!)

Auf dem Sobolev-Raum H^1 ist mit

$$\|v\|_{H^1(0,1)} := \|v\|_1 = \left(\int_0^1 \left(v^2 + (v')^2 \right) \mathrm{d}x \right)^{1/2}.$$

eine Norm definiert, die durch das Skalarprodukt

$$(u, v)_{H^1(0,1)} = \int_0^1 \left(uv + u'v' \right) \mathrm{d}x$$

induziert ist.

Man kann zeigen, dass der Sobolev-Raum H^1 bezüglich der genannten Norm vollständig ist, der Raum also ein Hilbert-Raum ist.

Dichthheitsargumente

Mit den Elementen in H^1, die ja Äquivalenzklassen von Funktionen sind, lässt es sich nicht so leicht arbeiten. Praktischerweise lassen sie sich als Folgen von Funktionen darstellen.

Satz 14.7 (Dichtheit)**.** *Die Mengen $C^k[0,1]$, $k \geq 1$ der k-mal setig differenzierbarer Funktionen sind dichte Teilräume von $H^1(0,1)$, d.h.*

$$\forall v \in H^1(0,1)\, \forall n \in \mathbb{N}\, \exists v_n \in C^k[0,1] \,:\, \|v - v_n\|_{H^1(0,1)} \leq \tfrac{1}{n}.$$

Siehe Theorem 3.17 in Adams und Fournier: *Sobolev Spaces.* 2003.

Jede Funktion $v \in H^1(0,1)$ lässt sich also beliebig gut in der Norm $\|\cdot\|_{H^1(0,1)}$ durch glatte Funktionen $v_n \in C^k[0,1]$ approximieren.

Spuroperator und Suprsatz

Da die Elemente $v \in H^1(0,1) \subset L^2(0,1)$ Äquivalenzklassen von Funktionen sind, die bis auf Nullmengen übereinstimmen, macht es vorerst keinen Sinn, nach dem Wert $v(a)$ für einzelne Punkte $a \in [0,1]$ zu fragen; bei ein-elementigen Mengen $\{a\}$ handelt es sich ja um Nullmengen bzgl. des Lebesgue-Maßes. Wir werden nun zeigen, wie sich dieses Dilemma beseitigen lässt, was unter anderem für die geeignete Formulierung von Randbedingungen wichtig ist.

Satz 14.8 (Spursatz)**.** *Es existiert eine Konstante $c_S > 0$ und einen eindeutig definierten stetigen linearen Operator $\gamma_0 : H^1(0,1) \to \mathbb{R}$ mit*

$$\begin{aligned} \gamma_0 v &= v(0) && \forall v \in C^1[0,1] \text{ und} \\ |\gamma_0 v| &\leq c_S \|v\|_{H^1(0,1)} && \forall v \in H^1(0,1). \end{aligned} \tag{14.3}$$

Beweis. Satz 12.6 besagt:

$$|v(0)| \leq c_S \|v\|_{H^1(0,1)} \qquad \forall v \in C^1[0,1]. \tag{14.4}$$

Zum Beweis verwenden wir nun ein Dichtheitsargument: Sei $v \in H^1(0,1)$ beliebig und $v_n \subset C^1[0,1]$ eine Folge so, dass $\|v - v_n\|_{H^1(0,1)} \leq 1/n$ (vgl. Satz 14.7). Wegen der Spurungleichung (14.4) ist $y_n := \gamma_0 v_n = v_n(0)$ eine Cauchy-Folge in $\mathbb{R}$, also konvergent mit Grenzwert $y \in \mathbb{R}$. Wir definieren einfach $\gamma_0 v := y = \lim_{n\to\infty} \gamma_0 v_n = \lim_{n\to\infty} v_n(0)$. Dieser Grenzwert ist unabhängig von der approximierenden Folge: Falls nämlich $v_n' \in C^1[0,1]$ eine weitere Folge mit $\|v - v_n\|_{H^1(0,1)} \leq 1/n$ ist, dann folgt mit (14.3) und Linearität des Spuroperators

$$|\gamma_0 v_n - \gamma_0 v_n'| = |\gamma_0(v_n - v_n')| \leq c_S \|v_n - v_n'\|_{H^1(0,1)} \leq c_S(\tfrac{1}{n} + \tfrac{1}{n}) \overset{n\to\infty}{\to} 0,$$

wobei wir im letzten Schritt v eingeschoben und Dreiecksungleichung verwendet haben. Somit ist $\gamma_0 v$ eindeutig definiert. Die Linearität von γ_0 sowie die Abschätzung (14.4) übertragen sich natürlich auf den Grenzwert; siehe Übung. □

Mit dem Spursatz wissen wir, dass für $u \in H^1$ die Punktauswertung $u(0)$ im Sinne des Spuroperators Sinn ergibt; dies gilt natürlich nicht nur für $x = 0$, sondern für alle $x \in [0,1]$. Dies erlaubt uns wieder die Formulierung der Randbedingungen.

Schwache Formulierung, Existenz und Eindeutigkeit

Nun können wir die variationelle Formulierung auf Basis des Sobolev-Raums formulieren. Dazu betrachten wir wieder das Problem in klassischer Formulierung:

$$-(a(x)u'(x))' + c(x)u(x) = f(x), \qquad x \in (0,1) \tag{14.5}$$

$$u(0) = g_0, \tag{14.6}$$

$$a(1)u'(1) + \gamma_1 u(1) = g_1. \tag{14.7}$$

Wir haben für dieses Problem bereits die Variationsform hergeleitet; die schwache Formulierung des Problems in nun folgende: Gesucht ist $u \in H^1(0,1)$ mit $u(0) = g_0$, sodass

$$\int_0^1 \big(au'v' + cuv\big)\,\mathrm{d}x + \gamma_1 u(1)v(1) = \int_0^1 fv\,\mathrm{d}x + g_1 v(1) \tag{14.8}$$

für alle Testfunktionen $v \in H^1(0,1)$ mit $v(0) = 0$, wobei $u(0)$, $u(1)$, $v(0)$, $v(1)$ jeweils im Sinne des Spuroperators zu verstehen sind. Lösungen dieser Formulierung heißen *schwache Lösungen*.

Bemerkung 14.9. (i) Nach Herleitung ist jede Lösung der klassischen Formulierung mit $u \in H^1(0,1)$ eine schwache Lösung. Umgekehrt kann man wieder zeigen: Falls eine schwache Lösung hinreichend regulär ist, z.B. $u \in C^2[0,1]$, dann ist sie auch klassische Lösung.

(ii) Die schwache Formulierung macht Sinn für $a, b, c, f \in L^\infty(0,1)$, also insbesondere für springende Koeffizienten. In diesem Fall existiert meist keine klassische Lösung; der Begriff der schwachen Lösung ist also tatsächlich allgemeiner.

Satz 14.10. *Das Problem* (14.8) *ist eindeutig lösbar.*

Beweis. Zum Beweis der Existenz und Eindeutigkeit einer Lösung in H^1 verwenden wir nun den Satz von Lax-Milgram auf dem Raum $V_0 := \{v \in H^1(0,1) : v(0) = 0\}$. Dieser Raum ist ein abgeschlossener Teilraum von H^1 und daher selbst auch ein Hilbert-Raum (mit derselben Norm).

Die Stetigkeit von a und ℓ lässt sich analog zum Kapitel 12 zeigen, wobei wir nun den Spursatz für H^1 verwenden müssen. Für den Beweis der Elliptizität von a auf V_0 verwendet man wieder die Poincaré-Friedrichs-Ungleichung (Satz 12.4), die man mit einem Dichtheitsargument auf H^1 erweitern kann. (Die Erweiterung erfolgt ganz analog zum Beweis von Satz 12.11, wo wir bereits ein Dichtheitsargument verwendet haben, um diese Poincaré-Friedrichs-Ungleichung von C^1 auf C^1_{pw} zu erweitern.)

Somit sind alle Voraussetzungen des Satzes von Lax-Milgram erfüllt; wir wissen also, dass für jedes $g \in H^1(0,1)$ folgendes Problem eine eindeutige Lösung hat:

$$\text{Gesucht } u_0 \in V_0 : a(u_0, v) = \ell(v) - a(g, v) \qquad \forall\, v \in V_0$$

Wählen wir g so, dass es die Dirichlet-Randbedingung erhält, wissen wir auch, dass (14.8) lösbar ist. Die Eindeutigkeit der Lösung erfolgt dann wie im Beweis des Satzes 12.7 direkt aus der Elliptizität. □

Unsere Diskussionen zur Wahl anderer Randbedingungen gelten analog.

15. Finite Elemente in mehreren Ortsdimensionen

Sei $\Omega \subset \mathbb{R}^d$ ein beschränktes Gebiet (=offene und zusammenhängende Menge) mit zumindest stückweise glattem Rand. Wir betrachten Randwertprobleme der Form

$$-\mathrm{div}(a\nabla u) + cu = f \qquad \text{in } \Omega, \tag{15.1}$$

$$\vec{n} \cdot (a\nabla u) + \gamma u = g \qquad \text{auf } \partial\Omega, \tag{15.2}$$

wobei a, c, f, γ und g hinreichend glatte, gegebene Funktionen sind. Diese Wahl umfasst auch die Wahl der Neumann-Randbedingungen. Dirichlet-Randbedingungen kann man mit $\gamma = \epsilon^{-1}$ mit $\epsilon > 0$ klein approximieren oder analog zum 1D-Fall direkt fordern. Wichtig ist, dass auf jedem Teil des Randes genau eine Randbedingung gefordert wird.

Variationsformulierung

Wir gehen völlig analog zum 1D-Fall vor, d.h., multilizieren mit einer Testfunktion integrieren über Ω und benutzen partielle Integration. Dies führt auf

$$\begin{aligned}\int_\Omega f(x)v(x)\,\mathrm{d}x &= \int_\Omega [-\mathrm{div}(a(x)\nabla u(x)) + c(x)u(x)]v(x)\,\mathrm{d}x \\ &= \int_\Omega a(x)\nabla u(x)\cdot\nabla v(x) + c(x)u(x)v(x)\,\mathrm{d}x \\ &\qquad - \int_{\partial\Omega} \vec{n}(x)\cdot a(x)\nabla u(x)\, v(x)\,\mathrm{d}s(x).\end{aligned}$$

Unter Benutzung der Robin-Randbedingung (15.2) kann man die Randterme ersetzen under erhält somit unmittelbar, dass jede reguläre Lösung von (15.1) – (15.2) folgende Variationsgleichung erfüllt:

$$\underbrace{\int_\Omega a\nabla u\cdot\nabla v + cuv\,\mathrm{d}x + \int_{\partial\Omega}\gamma uv\,\mathrm{d}s(x)}_{a(u,v) :=} = \underbrace{\int_\Omega fv\,dx + \int_{\partial\Omega} gv\,\mathrm{d}s(x)}_{\ell(v) :=}. \tag{15.3}$$

Wir nennen (15.3) wieder die *variationelle Formulierung* und jede Lösung davon eine *variationelle Lösung* des Randwertproblems (15.1) – (15.2).

Sobolevräume, Spuroperator, Poincaré-Ungleichungen

Völlig analog zum 1D-Fall definieren wir

$$H^1(\Omega) := \{u \in L^2(\Omega) : \nabla u \in L^2(\Omega)\}.$$

Dabei ist mit ∇u die schwache Ableitung gemeint; für glatte Funktionen stimmt diese natürlich wieder mit der klassischen Ableitung überein. Man beachte, dass $H^1(\Omega)$ wieder aus Äquivalenzklassen von Funktionen besteht, die bis auf Nullmengen übereinstimmen. Zusammen mit der Norm

$$\|u\|_{H^1(\Omega)} := \left(\|u\|^2_{L^2(\Omega)} + \|\nabla u\|^2_{L^2(\Omega)}\right)^{1/2}$$

und dem entsprechenden Skalarprodukt ist $H^1(\Omega)$ ein vollständiger normierter Raum, dessen Norm durch ein Skalarprodukt induziert wird, also ein Hilbert-Raum. Wir können wieder mit Dichtheit argumentieren:

Bemerkung 15.1. Sei $\Omega \subset \mathbb{R}^d$ offen und beschränkt. Dann gilt

$$C^k(\Omega) \cap H^1(\Omega) \subset H^1(\Omega) \qquad \text{dicht.}$$

Zu $u \in H^1(\Omega)$ existiert also eine Folge $u_n \in C^k(\Omega)$ mit beschränkter H^1-Norm, so dass $\|u - u_n\|_{H^1(\Omega)} \leq 1/n$ gilt. Falls Ω darüber hinaus einen stückweise glatten Rand besitzt (allgemeiner: ein Lipschitz-Gebiet ist), dann gilt sogar

$$C^k(\overline{\Omega}) \subset H^1(\Omega) \qquad \text{dicht.}$$

Die approximierenden Funktionen u_n sind also o.B.d.A. glatt bis zum Rand.

Mit derselben Argumentation wie in einer Dimension kann man dann auf Lipschitzgebieten wieder Randwerte für Funktionen $u \in H^1(\Omega)$ sinnvoll definieren.

Satz 15.2 (Spursatz)**.** *Für jedes Ω gibt es eine Konstante c_S, sodass*

$$\|\gamma_{\partial\Omega} u\|_{L^2(\partial\Omega)} \leq c_S \|u\|_{H^1(\Omega)}$$

mit $\gamma_{\partial\Omega} u := u|_{\partial\Omega}$ für alle $u \in C^1(\overline{\Omega})$. Weiters ist der Spuroperator $\gamma_{\partial\Omega}$ *linear, stetig und dicht-definiert und lässt sich daher stetig auf ganz $H^1(\Omega)$ erweitern. Die obige Spurungleichung bleibt dabei erhalten.*

Bemerkung 15.3. Wir können also jeder Funktion $u \in H^1(\Omega)$ eine Spur bzw. Randwert $u|_{\partial\Omega} := \gamma_{\partial\Omega} u \in L^2(\partial\Omega)$ zuweisen. Nicht jede Funktion $u \in L^2(\partial\Omega)$ ist allerdings auch Randwert einer Funktion $u \in H^1(\Omega)$. Wir definieren also

$$H^{1/2}(\partial\Omega) := \{\gamma_{\partial\Omega} u : u \in H^1(\Omega)\} \subset L^2(\partial\Omega)$$

als Raum möglicher Randwerte. Dieser lässt sich mithilfe der Norm

$$\|g\|_{H^{1/2}(\partial\Omega)} := \inf_{\substack{u\in H^1(\Omega)\\ \gamma_{\partial\Omega}u=g}} \|u\|_{H^1(\Omega)}$$

zu einem Hilbert-Raum machen: dem sogenannten Spurraum von H^1-Funktionen. Wenn wir also inhomogene Dirichlet-Randbedingungen fordern, können wir die Existenz einer Lösung nur zeigen, wenn die Randdaten in $H^{1/2}$ liegen.

Bemerkung 15.4. Wir können einer Funktion $u \in H^1(\Omega)$ zwar Randwerte zuordnen, aber keine Punktwerte. Eine Abbildung $\gamma_x : H^1 \to \mathbb{R}$ mit $\gamma_x u := u(x)$ ist in H^1 unbeschränkt. Es gibt also eine Folge von Funktionen $u_n \in H^1(\Omega)$ mit $\|u_n\|_{H^1} = 1$ und $\gamma_x u_n \to \infty$.

Zum Nachweis der Elliptizität der Bilinearform $a(\cdot,\cdot)$ werden wir, wie schon in einer Dimension, noch ein weiteres Resultat benötigen.

Satz 15.5 (Poincaré-Friedrichs-Ungleichung)**.** *Sei $\Omega \subset \mathbb{R}^d$ ein beschränktes Lipschitz-Gebiet.*

(i) Sei $\Omega_0 \subseteq \Omega$ eine Teilmenge mit positivem Maß. Dann gibt es eine Konstante $C_P > 0$, die nur von Ω und Ω_0 abhängt, sodass

$$\|u\|_{L^2(\Omega)} \le C_P \left(\|\nabla u\|_{L^2(\Omega)} + \left| \int_{\Omega_0} u(x)\mathrm{d}x \right| \right) \qquad \forall u \in H^1(\Omega).$$

(ii) Sei $\Gamma_0 \subseteq \partial\Omega$ eine Teilmenge mit positivem Maß. Dann gibt es eine Konstante $C_F > 0$, die nur von Ω und Γ_0 abhängt, sodass

$$\|u\|_{L^2(\Omega)} \le C_F(\|\nabla u\|_{L^2(\Omega)} + \|u\|_{L^2(\Gamma_0)}) \qquad \forall u \in H^1(\Omega).$$

Existenz und Eindeutigkeit schwacher Lösungen

Satz 15.6. *Sei $\Omega \subset \mathbb{R}^d$ ein beschränktes Lipschitz-Gebiet und seinen*

(i) $a, c \in L^\infty(\Omega)$, $\gamma \in L^\infty(\partial\Omega)$,

(ii) $a(x) \ge \underline{a} > 0$ f.ü. in Ω,

(iii) $c(x) \ge 0$ f.ü. in Ω und $\gamma(x) \ge 0$ f.ü. in $\partial\Omega$ und

(iv) zuminest eine der folgenden Bedingungen:

(a) $\gamma(x) > \underline{\gamma} > 0$ auf einer Teilmenge Γ_0 von $\partial\Omega$ mit positivem Maß,
(b) $c(x) > \underline{c} > 0$ auf einer Teilmenge Ω_0 von Ω mit positivem Maß.

Dann besitzt (15.3) *für alle* $f \in L^2(\Omega)$ *und* $g \in L^2(\partial\Omega)$ *eine eindeutige schwache Lösung* $u \in H^1(\Omega)$ *und es gilt*

$$\|u\|_{H^1(\Omega)} \le C(\|f\|_{L^2(\Omega)} + \|g\|_{L^2(\partial\Omega)}).$$

Dabei hängt C *nur von* Ω *und den Koeffizienten* a, c, γ *ab.*

Beweis. Wir überprüfen die Vorausetzungen des Satzes von Lax-Milgram.

(i) Der Raum $V = H^1(\Omega)$ ist ein Hilbert-Raum.

(ii) Die Abbildungen $a : V \times V \to \mathbb{R}$ und $\ell : V \to \mathbb{R}$ sind bilinear bzw. linear.

(iii) Mit Cauchy-Schwarz Ungleichung und der Spur-Ungleichung lässt sich problemlos die Beschränktheit von a und ℓ zeigen.

(iv) Für den Beweis der Elliptizität verwenden wir je nach vorhandener Bedingung

$$a(u,u) \ge \underline{a}\|\nabla u\|_{L^2(\Omega)} + \underline{c}\|u\|_{L^2(\Omega_0)} \quad \text{oder} \quad a(u,u) \ge \underline{a}\|\nabla u\|_{L^2(\Omega)} + \underline{\gamma}\|u\|_{L^2(\Gamma_0)}.$$

Um nun $a(u,u)$ von unter abzuschätzen, kann man entsprechende Varianten der Poincaré-Friedrichs-Ungleichungen verwenden.

Aus dem Satz von Lax-Milgram folgt nun sofort die Existenz und Eindeutigkeit einer Lösung $u \in V = H^1(\Omega)$ zur entsprechenden schwachen Formulierung und ebenso die a-priori Abschätzung für die Lösung. □

Zum Abschluss zitieren wir noch ein Satz über die zusätzliche Regularität von Lösungen für hinreichend glatte Daten. Dies macht zum einen die Annahmen plausibel, unter denen Konvergenzraten für die FE-Methode gezeigt werden.

Satz 15.7 (Regularität)**.** *Sei* $\Omega \subset \mathbb{R}^d$ *ein beschränktes Lipschitz-Gebiet und*

$$\Omega \textit{ konvex} \qquad \textit{oder} \qquad \partial\Omega \textit{ glatt.}$$

Dann gilt für jedes $f \in L^2(\Omega)$ *und* $g \in H^{3/2}(\partial\Omega)$ *für die Lösungen von*

$$\textit{Gesucht } u \in V_g: \quad \int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x = \int_\Omega f v \,\mathrm{d}x \quad \forall v \in V_0,$$

wobei $V_g := \{v \in H^1(\Omega) : \gamma_{\partial\Omega} u = g\}$*, sogar* $u \in H^2(\Omega)$*. Außerdem gibt es eine nur vom Gebiet abhängende Konstante* C_R*, sodass* $\|u\|_{H^2(\Omega)} \le C_R(\|f\|_{L^2(\Omega)} + \|g\|_{H^{3/2}(\partial\Omega)})$*.*

Für den nicht trivialen Beweis verweisen wir auf das Buch Grisvard: *Elliptic problems in non-smooth domains.* 1985.

Bemerkung 15.8. Unter Regularitätsannahmen an die Koeffizienten a, $\vec{b}$, c und d lässt sich die Regularitätsaussage auch auf das Problem (15.1)–(15.2) verallgemeinern, ganz in dem Sinn: ,,Für hinreichend reguläre Daten liegt die Lösung in H^2". Wichtig dabei ist jedoch, dass zu den ,,Daten" hier auch das Gebiet Ω zählt. Auf allgemeinen Lipschitz-Gebieten gilt die Aussage nämlich nicht, wie man durch Gegenbeispiele zeigen kann; siehe Übung.

Die Finite-Elemente Methode in 2D

Wir betrachten wieder den Fall, dass am gesamten Rand Robin-Randbedingen gefordert werden.

Gitter, Triangulierung

Wir zerlegen Ω in Dreiecke $T^{(k)}$ (=Elemente) und nennen $\{T^{(k)} : 1 \leq k \leq n_T\}$ eine Triangulierung von Ω; dabei garantieren wir, dass

(i) die Elemente $T^{(k)}$ ganz Ω abdecken, also der Abschluss von Ω gleich der Vereinigung der Abschlüsse der Elemente T_k ist,

(ii) die Elemente $T^{(k)}$ sich nicht überlappen, also $T^{(k)} \cap T^{(\ell)} = \emptyset$ für $k \neq \ell$,

(iii) es keine hängenden Knoten gibt, also kein Eckpunkt eines Dreiecks auf der Kante eines anderen Dreiecks liegt.

Die Information über das Gitter kann in Listen mit Knotenkoordinaten und Element-Knoten-Nummern bzw. Randelement-Knotennummern gespeichert werden. Für ein einfaches Gitter könnte das dann z.B. wie folgt aussehen:

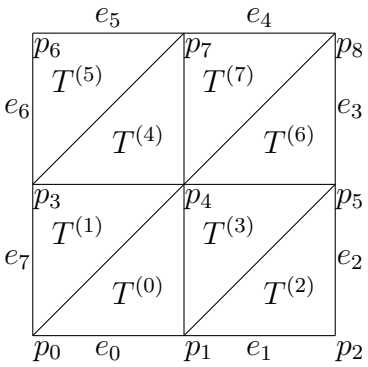
```
# Knotenkoordinaten
p = [ [ 0.0, 0.0 ], # p0
      [ 0.5, 0.0 ], # p1
      [ 1.0, 0.0 ], # p2
      [ 0.0, 0.5 ], # p3
      [ 0,5, 0.5 ], # p4
      [ 1.0, 0.5 ], # p5
      [ 0.0, 1.0 ], # p6
      [ 0,5, 1.0 ], # p7
      [ 1.0, 1.0 ] ]# p8

# Element-Knotennummern
t = [ [ 0, 1, 4 ], # T1
      [ 0, 4, 3 ], # T2
```

```
        [ 1, 2, 5 ], # T3
        [ 1, 5, 4 ], # T4
        [ 3, 4, 7 ], # T5
        [ 3, 7, 6 ], # T6
        [ 4, 5, 8 ], # T7
        [ 4, 8, 7 ] ]# T8

# Randelement-Knotennummern
e = [ [ 0, 1 ], [ 1, 2 ], [ 2, 5 ], [ 5, 8 ],
      [ 8, 7 ], [ 7, 6 ], [ 6, 3 ], [ 3, 0 ] ]
```

Dies würde folgendes Netz beschreiben:



Bei der Nummerierung der Knoten innerhalb der Elemente und des Randes behält man die Orientierung bei. Den Tabellen für die Element-Knotennummern und für die Randelement-Knotennummern können zusätzliche Daten angeschlossen werden (etwa elementweise konstante Werte der Konstanten oder Informationen zur Art der Randbedingung auf dem jeweiligen Randelement).

Finite-Elemente Approximation

Für die Approximation verwenden wir dann wieder den einfachsten Raum:

$$V_h = \{v \in C^0(\Omega) : u|_{T^{(k)}} \text{ linear } \forall\, T^{(k)}\}$$

ist der Raum der stetigen, stückweise linearen Funktionen und bemerken, dass $V_h \subset H^1(\Omega)$ (Courant-Element). Die Funktionen sind wieder eindeutig durch die Funktionswerte an den Knoten definiert; Knoten sind in unserem Fall nun die Ecken der Dreiecke.

Fehlerabschätzung

Mit dem Satz von Lax-Milgram folgt sofort, dass das diskretisierte Variationsproblem (Galerkin-Approximation, Finite-Elemente Methode) eindeutig lösbar ist, und mit dem Lemma von Céa erhält man weiters

$$\|u - u_h\|_{H^1(\Omega)} \leq \tfrac{\mu_2}{\mu_1} \inf_{v_h \in V_h} \|u - v_h\|_{H^1(\Omega)} \leq \tfrac{\mu_2}{\mu_1} \|u - r_h u\|_{H^1(\Omega)},$$

wobei $r_h : V \to V_h$ ein geeigneter Operator ist. Wenn $u \in C^0(\overline{\Omega})$, dann können wir für r_h die stückweise lineare Interpolation von u verwenden; diese ist durch folgende Interpolationsbedingungen eindeutig festgelegt:

$$r_h u(p_i) = u(p_i) \qquad \text{für alle Knoten } p_i.$$

Für Interpolation mit stückweise linearen Funktionen auf einer regulären Triangulierung gelten folgende Abschätzungen für den Interpolationsfehler

$$\|u - r_h u\|_{L^2(T)} \leq C_a h_T^2 \|u\|_{H^2(T)},$$
$$\|u - r_h u\|_{H^1(T)} \leq C_a h_T \|u\|_{H^2(T)},$$

wobei h_T den Durchmesser des Elements T bezeichnet, also die lokale Gitterweite. Das sind genau dieselben Abschätzungen wie in einer Dimension. Hiermit lassen sich dann folgende Fehlerabschätzungen für unser Modellproblem zeigen:

$$\|u - u_h\|_{H^1(\Omega)} \leq \frac{\mu_2}{\mu_1} C_a h \|u\|_{H^2(\Omega)}, \tag{15.4}$$

wobei $h = \max_T h_T$ die globale Gitterweite bezeichnet. Durch ein Dichtheitsargument lässt sich zeigen, dass (15.4) für alle $u \in H^2(\Omega)$ gilt.

Unter den Voraussetzungen des Satzes 15.7 über H^2-Regularität kann man zeigen, dass

$$\|u - u_h\|_{L^2(\Omega)} \leq Ch \|u\|_{H^1(\Omega)} \quad \text{und} \quad \|u - u_h\|_{L^2(\Omega)} \leq Ch^2 \|u\|_{H^2(\Omega)}; \tag{15.5}$$

für den Beweis wird das *Aubin-Nitsche-Dualitätsargument*, verwendet das wir in der VL *Numerik elliptischer Probleme* näher kennenlernen werden.

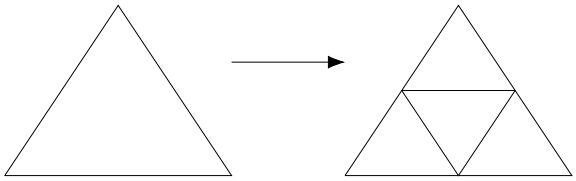
Bemerkungen zur Implementierung

Wie in einer Dimension können wir jedem Knoten p_i eine stückweise lineare und stetige Basisfunktion $\phi_i \in V_h$ zuordnen, und diese sind durch $\phi_i(p_j) = \delta_{ij}$ eindeutig definiert. Mit einer Skizze sieht man relativ leicht, dass die so gegebenen Basisfunktionen wieder die Form von *Hüten* haben und

$$\phi_i \equiv 0 \qquad \text{auf } T \qquad \text{falls } p_i \notin \overline{T}.$$

Auf jedem Element (=Dreieck) sind also genau drei Basisfunktionen nicht konstant gleich null, und zwar diejenigen, die zu den entsprechenden Knoten gehören.

Das Aufstellen der Steifigkeitsmatrix A_h und des Lastvektors $\underline{b}_h$ erfolgt dann wieder über elementweise *Assemblierung*; die für Neumann- und Robin-Randbedingungen erforderlichen Beiträge erhält man über randelementweise Assemblierung (dafür benötigen wir die auch die Randelement-Knotennummern). Zur Verfeinerung des Netzes können Verfeinerungsalgorithmen verwendet werden; so kann etwa jedes Dreieck in 3 deckungsgeliche Dreiecke untergliedert werden:



Auf Basis dieses Prinzips kann aus Netz (spezifiziert durch `p`, `t` und `e`) ein neues Netz bestimmt werden.

Details werden in der VL *Numerik elliptischer Probleme* im Sommersemester besprochen.

Numerische Experimente

Wir wenden nun die gelernten Methoden erneut auf die in Kapitel 10 besprochenen Beispiele an.

Beispiel 15.9. Zuerst betrachten wir als Rechengebiet Ω das L-förmige Gebiet

$$\Omega = (0,1)^2 \backslash [\tfrac{1}{2}, 1)^2,$$

siehe Abbildung 15.1. Wir wählen wieder

$$\begin{aligned} -\Delta u(x,y) = f(x,y) := 8\pi^2 \sin(2\pi x)\sin(2\pi y) \qquad & \text{für } (x,y) \in \Omega \\ u(x,y) = 0 \qquad & \text{für } (x,y) \in \partial\Omega. \end{aligned}$$

Die Lösung ist

$$u(x,y) = 2\pi^2 \sin(2\pi x)\sin(2\pi y),$$

wobei sich $u \in H^2(\Omega)$ leicht nachprüfen lässt. (Satz 15.7 lässt sich nicht anwenden!)

Zum Vergleich mit der FDM betrachten wir zuerst den Fehler in der Maximumsnorm $e_h^{(\infty)} = \|u - u_h\|_{\infty,h}$, siehe Tabelle 15.1, woraus abgelesen werden kann, dass das Verfahren wie $\mathcal{O}(h^2)$ konvergiert.

Die FEM-Konvergenztheorie gilt nun für die H^1-Norm. Da diese nicht ganz einfach auszurechnen ist, berechnen wir den Fehler auf einem Finite-Elemente-Netz, und zwar auf dem Netz, das bei einer weiteren Verfeinerungsstufe erreicht wird; wir wählen also $e_h^{(1)} = \|r_{h/2}u - u_h\|_{H^1(\Omega)}$, wobei $r_{h/2}u$ die Interpolation der exakten Lösung $u \in C^\infty(\Omega)$ an den Knotenpunkten ist. Diese Resultate finden sich in Tabelle 15.2 (links). Wir erzielen – wegen $u \in H^2(\Omega)$ entsprechend der Theorie – eine Konvergenzrate von $\mathcal{O}(h)$.

Schließlich betrachten wir in Tabelle 15.2 (rechts) die Konvergenz in der L^2-Norm, also $e_h^{(0)} = \|r_{h/2}u - u_h\|_{L^2(\Omega)}$, wo wir eine Rate von $\mathcal{O}(h^2)$ erhalten; die Gründe dafür werden wir in der VL *Numerik elliptischer Probleme* sehen.

Beispiel 15.10. Nun betrachten wir als Rechengebiet Ω wieder einen Kreis

$$\Omega = \{\mathbf{x} \in \mathbb{R}^2 : \|\mathbf{x}\|_{\ell^2} < 1\}.$$

Diesen können wir durch ein Dreiecknetz nur annähern, jedoch können wir bei der Verfeinerung darauf achten, das Netz immer besser zu approximieren; siehe Abbildung 15.2. (Der Unterschied zwischen dem Rechengebiet hier und dem aus dem Kapitel 10 ist nur der einfachen Implementierbarkeit geschuldet.) Als Randwertproblem wählen wir wieder

$$\begin{aligned} -\Delta u(x,y) = f(x,y) := 16 \qquad & \text{für } (x,y) \in \Omega \\ u(x,y) = 0 \qquad & \text{für } (x,y) \in \partial\Omega. \end{aligned}$$

Es lässt sich leicht nachprüfen, dass

$$u(x,y) = 4 - 4\left(x^2 + y^2\right).$$

Der Fehler in der Maximumsnorm verhält sich entsprechend Tabelle 15.3, woraus abgelesen werden kann, dass die Konvergenzrate nun besser ist, als $\mathcal{O}(h)$. (In der FDM war die Rate ja schlechter als $\mathcal{O}(h)$.)

In Tabelle 15.4 sehen wir wieder die Fehlerschätzungen in der H^1-Norm und der L^2-Norm (wie im letzten Beispiel). Die erzielten Raten sind wieder ähnlich zum letzten Beispiel; hier ist anzumerken, dass das Beispiel mit dem Kreis nicht vollständig von unserer Theorie abgedeckt ist, da wir bei der Theorie stets angenommen hatten, dass Ω durch das Dreiecksnetz exakt dargestellt wird.

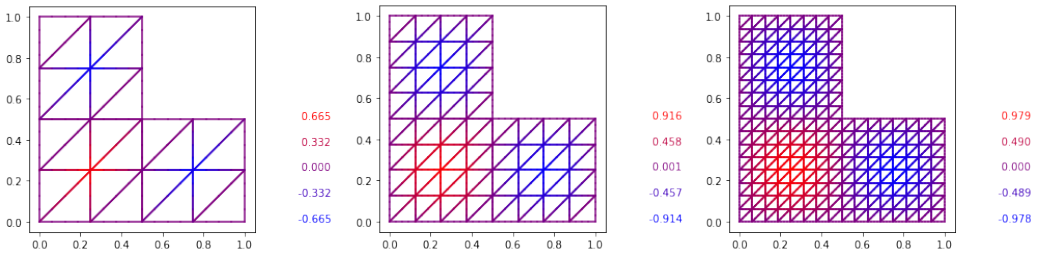


Abbildung 15.1.: Netz für das L-förmige Gebiet.

Verf.	Fehler $e_h^{(\infty)}$	Faktor $e_{2h}^{(\infty)}/e_h^{(\infty)}$
1	$5.00\ 10^{-1}$	
2	$2.08\ 10^{-1}$	3.40
3	$5.81\ 10^{-2}$	3.58
4	$1.49\ 10^{-2}$	3.89
5	$3.76\ 10^{-3}$	3.97
6	$9.41\ 10^{-4}$	3.99

Tabelle 15.1.: Konvergenz am L-förmigen Gebiet in Maximumsnorm.

Verf.	Fehler $e_h^{(1)}$	Faktor $e_{2h}^{(1)}/e_h^{(1)}$
1	$2.41\ 10^{0}$	
2	$1.28\ 10^{0}$	1.88
3	$6.51\ 10^{-1}$	1.97
4	$3.27\ 10^{-1}$	1.99
5	$1.64\ 10^{-1}$	2.00
6	$8.18\ 10^{-2}$	2.00

Verf.	Fehler $e_h^{(0)}$	Faktor $e_{2h}^{(0)}/e_h^{(0)}$
1	$2.09\ 10^{-1}$	
2	$6.98\ 10^{-2}$	2.99
3	$1.89\ 10^{-2}$	3.69
4	$4.84\ 10^{-3}$	3.92
5	$1.22\ 10^{-3}$	3.98
6	$3.04\ 10^{-4}$	3.99

Tabelle 15.2.: Konvergenz am L-förmigen Gebiet in H^1-Norm (links) und der L^2-Norm (rechts).

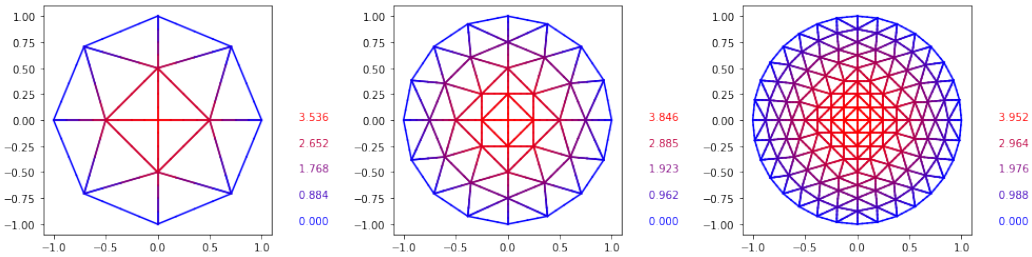


Abbildung 15.2.: Approximation des Kreises.

Verf.	Fehler $e_h^{(\infty)}$	Faktor $e_{2h}^{(\infty)}/e_h^{(\infty)}$
1	$6.31\ 10^{-1}$	
2	$1.95\ 10^{-1}$	3.23
3	$5.83\ 10^{-2}$	3.35
4	$1.69\ 10^{-2}$	3.45
5	$4.80\ 10^{-3}$	3.52
6	$1.34\ 10^{-3}$	3.57
7	$3.72\ 10^{-4}$	3.61

Tabelle 15.3.: Konvergenz am Kreis in Maximumsnorm.

Verf.	Fehler $e_h^{(1)}$	Faktor $e_{2h}^{(1)}/e_h^{(1)}$
1	$2.87\ 10^{0}$	
2	$1.42\ 10^{0}$	2.02
3	$6.95\ 10^{-1}$	2.04
4	$3.42\ 10^{-1}$	2.03
5	$1.69\ 10^{-1}$	2.02
6	$8.40\ 10^{-2}$	2.01
7	$4.19\ 10^{-2}$	2.01

Verf.	Fehler $e_h^{(0)}$	Faktor $e_{2h}^{(0)}/e_h^{(0)}$
1	$6.99\ 10^{-1}$	
2	$2.02\ 10^{-1}$	3.46
3	$5.32\ 10^{-2}$	3.79
4	$1.36\ 10^{-2}$	3.91
5	$3.43\ 10^{-3}$	3.96
6	$8.62\ 10^{-4}$	3.98
7	$2.16\ 10^{-4}$	3.99

Tabelle 15.4.: Konvergenz am Kreis in H^1-Norm (links) und der L^2-Norm (rechts).

Abschließende Bemerkungen

Wir haben nun gesehen, dass die Finite Elmente Methode insbesondere wegen der Flexibilität beim Aufsetzen der Netze vorteilhaft ist. Dies bezieht sich nicht nur auf nicht-triviale Gebiete, sondern auch auf die Möglichkeit adaptiver Netzverfeinerung. Außerdem hatten wir gesehen, dass die variationelle Herangehensweise beim Einbau von natürlichen Randbedingungen deutlich einfacher ist, sie außerdem mit deutlich reduzierten Glattheiten auskommt.

Teil III.

Anfangs-Randwertprobleme

16. Parabolische Differentialgleichungen

Wir wenden uns jetzt parabolischen Differentialgleichungen zu und betrachten als Modellprobem ein Anfangs-Randwertproblem für die Wärmeleitungsgleichung

$$\partial_t u(x,t) - \Delta u(x,t) = f(x,t) \qquad x \in \Omega,\ t > 0, \tag{16.1}$$

$$u(x,t) = 0 \qquad x \in \partial\Omega,\ t > 0, \tag{16.2}$$

$$u(x,0) = u_0(x) \qquad x \in \Omega. \tag{16.3}$$

Zur Diskretisierung werden wir die bereits bekannten Resultate für Randwertprobleme und für Anfangswertprobleme kombinieren. Die Resultate dieses Abschnitts können ganz einfach auf Robin- oder Neumann-Randbedingungen erweitert werden, außerdem kann der Laplace-Operator hier durch jeden anderen Differentialoperator im Ort ersetzt werden, für den die Theorie aus Teil II gilt. Wir wollen uns aber auf das Modellproblem beschränken.

Die Grundidee für die Vorgehensweise in diesem Teil ist:

(i) Zuerst wird das Problem im Ort diskretisiert, wobei wir die Techniken aus Teil II verwenden. Wir verwenden dazu die FEM, wenngleich auch die FDM verwendet werden könnte. Dabei erhalten wir eine *Semidiskretisierung im Ort.*

(ii) Dann wird das Problem in der Zeit mit einem Einschrittverfahren wie aus Teil I diskretisiert. Hier erhalten wir dann ein *volldiskretisiertes* Problem.

Bevor wir zu den ersten theoretischen Aussagen kommen, machen wir uns mit dem Problem vertraut. Für die Anwendung der FEM benötigen wir eine Variationsformulierung, die wir ganz analog zum Teil II herleiten können. Wir nehmen an, dass $\Omega \subset \mathbb{R}^d$ ein beschränktes Gebiet mit Lipschitz-Rand ist. Unter den bekannten Regularitätsannahmen erfüllt eine Lösung u die variationelle Identität

$$\underbrace{\int_\Omega \partial_t u(x,t) v(x)\,\mathrm{d}x}_{=(\partial_t u(\cdot,t),v)_{L^2(\Omega)}} + \underbrace{\int_\Omega \nabla u(x,t)\cdot\nabla v(x)\,\mathrm{d}x}_{=a(u(\cdot,t),v)} = \underbrace{\int_\Omega f(x,t)v(x)\,\mathrm{d}x}_{=\ell_t(v)} \tag{16.4}$$

für alle $v \in H_0^1(\Omega)$. Da $u(\cdot,t)$ in die Bilinearform $a(\cdot,\cdot)$ eingesetzt wird, soll u eine stetige Funktion sein, die von $[0,T]$ auf $H_0^1(\Omega)$ abbildet. Da die Zeitableitung von u in das L^2-Skalarprodukt eingesetzt wird, soll es außerdem eine stetig

differenzierbare Funktion sein, die von $[0,T]$ auf $L^2(\Omega)$ abbildet. Wir schreiben

$$u \in C([0,T]; H_0^1(\Omega)) \cap C^1([0,T]; L^2(\Omega)). \tag{16.5}$$

Im Folgenden schreiben wir statt $u(\cdot,t)$ nur kurz $u(t)$; analog verfahren wir auch mit anderen Funktionen, die von Ort und Zeit abhängen.

Für die Semidiskretisierung im Ort wählen wir nun ein $V_h \subset H_0^1(\Omega)$ mit endlicher Dimension. Die Galerkin-Lösung $u_h \in C([0,T]; V_h)$ erfüllt dann (neben der Anfangsbedingung) die Identität

$$(\partial_t u_h(t), v_h)_{L^2(\Omega)} + a(u_h(t), v_h) = \ell_t(v_h)$$

für alle $v_h \in V_h$. Zusätzlich haben wir eine Anfangsbedingung $u_h(0) = u_{0,h}$ mit geeignetem $u_{0,h}$. Mit dem Galerkin-Isomorphismus können wir dieses Problem in Matrix-Vektor-Schreibweise schreiben:

$$M_h \partial_t \underline{u}_h(t) + A_h \underline{u}_h(t) = \underline{b}_h(t), \qquad \underline{u}_h(0) = \underline{u}_{0,h}$$

wobei $\underline{u}_h, \underline{b}_h : [0,T] \to \mathbb{R}^N$ immer noch Funktionen in der Zeit sind. A_h ist hier die Steifigkeitsmatrix, die sich aus der Diskretisierung von $a(\cdot,\cdot)$ ergibt. M_h ist die Massematrix, die sich aus der Diskretisierung von $(\cdot,\cdot)_{L^2(\Omega)}$ ergibt.

Dieses System von gewöhnlichen Differentialgleichungen können wir mit den Methoden aus Teil I analysieren und lösen. Es lässt sich in

$$\partial_t \underline{u}_h(t) = f(t, \underline{u}_h(t)) := M_h^{-1} \underline{b}_h(t) - M_h^{-1} A_h \underline{u}_h(t)$$

umschreiben. Offensichtlich ist f in seiner zweiten Komponente Lipschitz-stetig. Als Norm bietet sich etwa $\|\underline{v}_h\|_{M_h} := (\underline{v}_h^\top M_h \underline{v}_h)^{1/2}$ an:

$$\begin{aligned}
\|f(t,\underline{u}_h(t)) - f(t,\underline{w}_h(t))\|_{M_h} &= \|A_h(\underline{u}_h(t) - \underline{w}_h(t))\|_{M_h^{-1}} \\
&= \sup_{0 \neq \underline{v}_h \in \mathbb{R}^N} \frac{\underline{v}_h^\top A_h(\underline{u}_h(t) - \underline{w}_h(t))}{\|\underline{v}_h\|_{M_h}} = \sup_{0 \neq v_h \in V_h} \frac{a(u_h(t) - w_h(t), v_h)}{\|v_h\|_{L^2(\Omega)}} \\
&\leq \sup_{0 \neq v_h \in V_h} \frac{\|u_h(t) - w_h(t)\|_{H^1(\Omega)} \|v_h\|_{H^1(\Omega)}}{\|v_h\|_{L^2(\Omega)}} \leq c_{inv}^2 h^{-2} \|\underline{u}_h(t) - \underline{w}_h(t)\|_{M_h},
\end{aligned}$$

wobei wir für die letzte Abschätzung die inverse Ungleichung verwendet haben.

Daraus folgt mit dem Satz von Pircard-Lindelöf (Satz 2.1) die Existenz und die Eindeutigkeit einer Lösung für das semidiskretisierte Problem und mit Gronwall (Satz 2.3) Stabilität. Für $h \to 0$ wird die Stabilität immer schlechter.

Hilfreich ist nun die Theorie aus Kapitel 4. Da $a(\cdot,\cdot)$ elliptisch ist, haben wir eine einseitige Lipschitz-Bedingung wie in Definition 4.17:

$$\begin{aligned}
(f(t,\underline{u}_h(t)) - f(t,\underline{w}_h(t)), \underline{u}_h(t) - \underline{w}_h(t))_{M_h} &= (\underline{u}_h(t) - \underline{w}_h(t), \underline{u}_h(t) - \underline{w}_h(t))_{A_h} \\
&= -a(u_h(t) - w_h(t), u_h(t) - w_h(t)) \leq 0.
\end{aligned}$$

Das Problem ist also *dissipativ*. Auf Grundlage dieser Überlegungen zeigen wir nun einige theoretische Aussagen. Auch für die Konstruktion stabiler Zeitdiskretisierungen hilft uns diese Erkenntnis.

Lösbarkeit und Stabilitätsabschätzungen

Wir betrachten nun die variationelle Identität (16.4) mit u wie in (16.5). Mit denselben Argumenten wie bei elliptischen Differentialgleichungen kann man zeigen, dass jede Lösung des Systems (16.1) – (16.3), die in (16.5) lebt, die variationelle Identität (16.4) erfüllt. Ähnlich zum elliptischen Fall kann man unter entsprechenden Regularitätsannahmen auch wieder den umgekehrten Weg gehen.

Die Regularitätsannahmen an die schwache Lösung ließen sich sogar noch weiter abschwächen; siehe z.B. das Buch von *Evans: Partial Differential Equations.*

Satz 16.1. *Für jedes $f \in C([0,T];L^2(\Omega))$ und $u_0 \in H_0^1(\Omega) \cap H^2(\Omega)$ existiert eine eindeutige schwache Lösung $u \in C([0,T];H_0^1(\Omega)) \cap C^1([0,T];L^2(\Omega))$ zum Anfangs-Randwertproblem* (16.1) – (16.3)*, welche für alle $t \in (0,T]$ die Stabilitätsabschätzungen*

$$\|u(t)\|^2_{L^2(\Omega)} + \int_0^t \|\nabla u(s)\|^2_{L^2(\Omega)} \mathrm{d}s \leq C \left(\|u_0\|^2_{L^2(\Omega)} + \int_0^t \|f(s)\|^2_{L^2(\Omega)} \mathrm{d}s \right) \quad (16.6)$$

und

$$\|u(t)\|^2_{H^1(\Omega)} + \int_0^t \|\partial_t u(s)\|^2_{L^2(\Omega)} \mathrm{d}s \leq C \left(\|u_0\|^2_{H^1(\Omega)} + \int_0^t \|f(s)\|^2_{L^2(\Omega)} \mathrm{d}s \right) \quad (16.7)$$

erfüllt.

Beweis. Für die Existenzaussage sei auf das Buch *Evans: Partial Differential Equations* verwiesen. Wir beweisen hier nur die Normabschätzungen für die Lösung. Mit elementarer Rechnung sieht man

$$\tfrac{\mathrm{d}}{\mathrm{d}t}\|u(t)\|^2_{L^2(\Omega)} = 2\int_\Omega \partial_t u(t) u(t)\,\mathrm{d}x.$$

Setzt man nun $v = u(t)$ in (16.4) ein, so folgt mit der Young'schen Ungleichung

$$\begin{aligned} \tfrac{1}{2}\tfrac{\mathrm{d}}{\mathrm{d}t}\|u(t)\|^2_{L^2(\Omega)} &= \int_\Omega f(t)u(t)\,dx - \int_\Omega |\nabla u(t)|^2 dx \\ &\leq \tfrac{\epsilon}{2}\|f(t)\|^2_{L^2(\Omega)} + \tfrac{1}{2\epsilon}\|u(t)\|^2_{L^2(\Omega)} - \|\nabla u(t)\|^2_{L^2(\Omega)} \end{aligned}$$

für jedes $\epsilon > 0$. Aufgrund der Friedrichs-Ungleichung gilt $\|u\|_{L^2(\Omega)} \leq C_F \|\nabla u\|_{L^2(\Omega)}$ für alle $u \in H_0^1(\Omega)$. Setzt man also in obiger Ungleichung $\epsilon = C_F^2$, so kann man die letzten beiden Terme zusammenziehen. Nach Umstellen und Multiplikation mit zwei folgt dann

$$\tfrac{\mathrm{d}}{\mathrm{d}t}\|u(t)\|^2_{L^2(\Omega)} + \|\nabla u(t)\|^2_{L^2(\Omega)} \leq C_F^2 \|f(t)\|^2_{L^2(\Omega)}.$$

Durch Integration nach der Zeit und Verwenden von (16.3) erhält man unmittelbar die erste Schranke (16.6). Testet man in Gleichung (16.4) hingegen mit $v = \partial_t u(t)$, so kommt man mit ähnlicher Rechnung auf die Ungleichung

$$\|\partial_t u(t)\|^2_{L^2(\Omega)} + \tfrac{\mathrm{d}}{\mathrm{d}t}\|\nabla u(t)\|^2_{L^2} \leq \|f(t)\|^2_{L^2(\Omega)}.$$

Durch Aufintegrieren in der Zeit und Komination mit der ersten Abschätzung erhält man dann die zweite Schranke (16.7), vgl. Übung. Die Eindeutigkeit folgt aus den Normabschätzungen. □

Bemerkung 16.2. Die Argumente, die zum Nachweis der Schranken für die Lösung verwendet wurde, werden *Energieabschätzungen* genannt. Anhand des Beweises kann man erahnen, dass sich die Regularitätsvoraussetzungen an die schwache Lösung und die Daten noch weiter abschwächen lassen. Mit ähnlichen Argumenten lassen sich auch allgemeinere Differentialgleichungen und Randbedingungen betrachten. Für Details verweisen wir wieder auf das Buch von Evans.

Semidiskretisierung im Ort mit FEM

Als nächstes wenden wir unsere Kenntnisse über elliptische Differentialgleichungen zur Diskretisierung im Ort an. Wie in der Einleitung zu diesem Kapitel angesprochen, sei $V_h \subset V := H^1_0(\Omega)$ ein endlich-dimensionaler Teilraum, z.B. der Finite-Elemente Raum der stückweise linearen stetigen Funktionen. Zur numerischen Approximation unseres Anfangs-Randwertproblems betrachten wir dann

Problem 16.3 (Semidiskretisierung)**.**
Gesucht ist $u_h \in C([0,T], V_h)$ mit $u_h(0) = u_{0,h} \in V_h$ und so, dass

$$\int_\Omega \partial_t u_h(t) v_h \,\mathrm{d}x + \int_\Omega \nabla u_h(t) \cdot \nabla v_h \,\mathrm{d}x = \int_\Omega f(t) v_h \,\mathrm{d}x \tag{16.8}$$

für alle $v_h \in V_h$ und alle $t \in (0,T]$ gilt.

Die Wahl einer geeigneten Approximation $u_{0,h}$ für den Anfangswert u_0 wird weiter unten noch diskutiert. Dessen ungeachtet können wir schon einige Eigenschaften des numerischen Verfahrens und seiner Lösung zeigen.

Wir haben bereits gesehen, dass sich das Problem mit dem Galerkin-Isomporphismus in Matrix-Vektor-Form schreiben lässt:

$$\begin{aligned} M_h \partial_t \underline{u}_h(t) + A_h \underline{u}_h(t) &= \underline{b}_h(t), \qquad t \in (0,T], \\ \underline{u}_h(0) &= \underline{u}_{0,h}. \end{aligned}$$

Existenz einer eindeutigen Lösung folgt sofort aus dem Satz von Picard-Lindelöf (Satz 2.1) bzw. aus der expliziten Lösungsdarstellung mittels Matrix-Exponentialfunktion und Duhamel-Formel (=Variation der Konstanten). Analog zum kontinuierlichen Fall lassen sich wieder Stabilitätsabschätzungen zigen. Wir erhalten daher sofort

Satz 16.4. *Sei $V_h \subset H_0^1(\Omega)$ mit $\dim(V_h) < \infty$. Dann besitzt Problem 16.3 eine eindeutige Lösung u_h, die für alle $0 \in (0,T]$ folgende Abschätzungen erfüllt:*

$$\|u_h(t)\|^2_{L^2(\Omega)} + \int_0^t \|\nabla u_h(s)\|^2_{L^2(\Omega)} \mathrm{d}s \leq C \left(\|u_{0,h}\|^2_{L^2(\Omega)} + \int_0^t \|f(s)\|^2_{L^2(\Omega)} \mathrm{d}s \right),$$
$$\|u_h(t)\|^2_{H^1(\Omega)} + \int_0^t \|\partial_t u_h(s)\|^2 \mathrm{d}s \leq C \left(\|u_{0,h}\|^2_{H^1(\Omega)} + \int_0^t \|f(s)\|^2_{L^2(\Omega)} \mathrm{d}s \right).$$

Fehlerabschätzung

Für die folgenden Überlegungen sei stets $V_h := P_1(T_h) \cap H_0^1(\Omega)$ der bereits betrachtete Finite-Elemente Raum mit stückweise linearen stetigen Funktionen über einer hinreichend regulären Triangulierung T_h des Gebiets Ω; siehe Teil II. Zur Fehlerabschätzung gehen wir dann wie folgt vor: Wie zerlegen

$$\|u(t) - u_h(t)\| \leq \|u(t) - \tilde{u}_h(t)\| + \|\tilde{u}_h(t) - u_h(t)\|, \tag{16.9}$$

wobei $\tilde{u}_h(t) \in V_h$ und die Normen noch frei gewählt werden können.

Wir schätzen also auf der linken Seite stehenden *Diskretisierungsfehler* durch die Summe eines *Approximationsfehlers*, der durch geeignete Wahl von $\tilde{u}_h$ kontrolliert werden kann, und einem *diskreter Fehler*, den wir mit Energieabschätzungen für das diskrete Problem behandeln werden, ab.

Approximationsfehler

Zunächst fixieren wir die Hilfsfunktion $\tilde{u}_h(t)$, die wir in der folgenden Analyse verwenden werden. Wir wählen sie so, dass $\tilde{u}_h(t) \in V_h$ die Funktion ist, die $u(t)$ in der H^1-Norm bestmöglich approximiert. Wir wissen bereits, dass die Bestapproximation durch die Galerkin-Approximation gegeben ist. Wir haben also

Satz 16.5. *Zu jedem $u(t) \in H_0^1(\Omega)$ existiert ein eindeutiges $\tilde{u}_h(t) \in V_h$ mit*

$$\int_\Omega \nabla \tilde{u}_h(t) \cdot \nabla v_h \,\mathrm{d}x = \int_\Omega \nabla u(t) \cdot \nabla v_h \,\mathrm{d}x \qquad \forall v_h \in V_h.$$

Die Abbildung $\tilde{\Pi}_h : H_0^1(\Omega) \to V_h$, $u \mapsto \tilde{\Pi}_h u := \tilde{u}_h$ ist linear und stetig und erfüllt

$$\|u - \tilde{\Pi}_h u\|_{H^1(\Omega)} \leq Ch|u|_{H^2(\Omega)}. \tag{16.10}$$

BEWEIS. Die Formel $\ell(v) := \int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x$ definiert ein stetiges lineares Funktional $H_0^1(\Omega) \to \mathbb{R}$, es gilt also $\ell \in [H_0^1(\Omega)]^* \subset V_h^*$. Die Existenz und Eindeutigkeit von $\tilde{u}_h$ folgen mit dem Satz von Lax-Milgram. Die Linearität der Abbildung $\tilde{\Pi}_h$ und die Stetigkeit sind offensichtlich. Die Fehlerabschätzung ergibt sich schließlich unmittelbar aus den Konvergenzaussagen zur Finite-Elemente Approximation bei elliptischen Problemen. □

Bemerkung 16.6. Die Abbildung $\tilde{\Pi}_h : V \to V_h$ ist über das numerische Lösen eines elliptischen Variationsproblems definiert und heißt auch *elliptische Projektion* bzw. *Ritz-Projektor* nach Walter Ritz.

Mit dem *Aubin-Nitsche-Dualitätsargument* erhält man im 1D-Fall immer, im 2D-Fall unter den Voraussetzungen des Satzes 15.7 (also wenn $\Omega \subseteq \mathbb{R}^2$ konvex ist *oder* glatten Rand hat) auch folgende Abschätzungen:

$$\|u - \tilde{\Pi}_h u\|_{L^2(\Omega)} \leq Ch|u|_{H^1(\Omega)} \qquad \text{und} \qquad \|u - \tilde{\Pi}_h u\|_{L^2(\Omega)} \leq Ch^2|u|_{H^2(\Omega)}. \tag{16.11}$$

Bemerkung 16.7. Wenn ein anderes Problem betrachtet werden soll, etwa $(\partial_t u(t), v)_{L^2} + a(u(t), v) = \ell_t(v)$ mit einer anderen H^1-elliptischen und H^1-stetigen Bilinearform a, dann ist $\tilde{\Pi}_h$ über die a-Orthogonalität zu definieren.

Diskreter Fehler

Zuerst stellen wir fest, dass aus der Kombination von (16.4) und (16.8) folgendes Galerkin-Orthogonalitätsresultat folgt:

$$\int_\Omega \partial_t(u(t) - u_h(t))v_h \, \mathrm{d}x + \int_\Omega \nabla(u(t) - u_h(t)) \cdot \nabla v_h \, \mathrm{d}x = 0 \tag{16.12}$$

gilt für alle $v_h \in V_h$ und alle $t \in (0, T]$. Der Operator $\tilde{\Pi}_h$ aus Satz 16.5 erfüllt nach seiner Definition das Galerkin-Orthogonalitätsresultat

$$\int_\Omega \nabla(u(t) - \tilde{\Pi}_h u(t)) \cdot \nabla v_h \, \mathrm{d}x = 0 \tag{16.13}$$

für alle $v_h \in V_h$ und alle $t \in (0, T]$. Durch Bildung der Differenz erhalten wir, dass der diskrete Fehler

$$d_h(t) := \tilde{u}_h(t) - u_h(t) = \tilde{\Pi}_h u(t) - u_h(t)$$

folgende Identität erfüllt:

$$\int_\Omega \partial_t d_h(t) v_h dx + \int_\Omega \nabla d_h(t) \cdot \nabla v_h dx = \int_\Omega \partial_t \underbrace{(\tilde{\Pi}_h u(t) - u(t))}_{\rho_h(t) :=} v_h dx.$$

Beobachtung: Der diskrete Fehler $d_h(t)$ ist Lösung eines diskreten Variationsproblems (16.8) mit spezieller rechter Seite $\partial_t \rho_h(t)$. Aus dieser Überlegung und der Energieabschätzung für die Lösung des diskreten Problems erhalten wir nun unmittelbar die folgende Aussage.

Satz 16.8 (Diskreter Fehler)**.**
Sei u_h die Lösung zum Problem 16.3 mit Startwert $u_{0,h} = \tilde{\Pi}_h u_0$ und weiters $\tilde{u}_h(t) = \tilde{\Pi}_h u(t)$ und $\rho_h(t) = \tilde{u}_h(t) - u(t)$. Dann gilt für alle $t \in (0, T]$

$$\|d_h(t)\|^2_{L^2(\Omega)} + \int_0^t \|\nabla d_h(s)\|^2_{L^2(\Omega)} \mathrm{d}s \leq C \int_0^t \|\partial_t \rho_h(s)\|^2_{L^2(\Omega)} \mathrm{d}s.$$

Durch Kombination der vorhergehenden Resultate und der Poincaré-Ungleichung folgt nun

Satz 16.9 (Konvergenz der Semidiskretisierung)**.**
Sei u eine hinreichend reguläre Lösung des Problems (16.1) – (16.3) *und u_h die Finite-Elemente Lösung zu Problem 16.3 mit $u_{h,0} = \tilde{\Pi}_h u_0$. Dann gilt für $t \in (0, T]$*

$$\|u(t) - u_h(t)\|^2_{L^2(\Omega)} + \int_0^t \|u(s) - u_h(s)\|^2_{H^1(\Omega)} \, \mathrm{d}s \leq C(u; T) h^2$$

mit einem $C(u; T)$, das von u und T abhängt, nicht jedoch von t oder h abhängt.

Beweis. siehe Übung. □

Bemerkung 16.10. Man beachte, dass in den Fehlerabschätzungen Quadrate von Normen auftauchen. Zieht man die Wurzeln, so sieht man, dass völlig analog zum elliptischen Problem der gesamte Fehler mit $\mathcal{O}(h)$ konvergiert. Zusammenfassend kann man sagen, dass sich die Konvergenzaussagen für elliptische Probleme auch auf den parabolischen Fall übertragen lassen.

Diskretisierung in der Zeit mit Einschrittverfahren

Um eine tatsächlich implementierbare Methode zu erhalten, betrachten wir noch die anschließende Zeitdiskretisierung des bereits im Ort diskretisierten Problems (Volldiskretisierung). Wir bezeichnen im Folgenden die Zeitschrittweite mit $\tau > 0$ und definieren mittels $t^{(n)} := n\tau$ die diskreten Zeitpunkte.

Für die Diskretisierung in der Zeit verwenden wir eine Klasse einfacher Runge-Kutta-Formeln. Konkret beschränken wir uns auf das θ-Schema, das das explizite

Euler-Verfahren ($\theta = 0$), das implizite Euler-Verfahren ($\theta = 1$) und die implizite Trapezregel ($\theta = 1/2$; diese entspricht dem Lobatto-III-A Verfahren mit $s = 2$ Stufen) als Spezialfälle enthält. Wir realisieren also die Zeitdiskretisierung eines Anfrangswertproblems $y'(t) = f(t, y)$ mit dem Schema

$$y^{(n+1)} = y^{(n)} + \tau\big((1-\theta)f(t^{(n)}, y^{(n)}) + \theta f(t^{(n+1)}, y^{(n+1)})\big). \tag{16.14}$$

Notation. Zur Abkürzung schreiben wir $u^{(n)} = u(t^{(n)})$ und bezeichnen mit

$$\mathrm{d}_\tau u^{(n)} := \tfrac{1}{\tau}(u^{(n)} - u^{(n-1)}) \qquad \text{und} \qquad u^{(n+\theta)} := (1-\theta)u^{(n)} + \theta u^{(n+1)}$$

den Rückwärtsdifferenzenquotienten und entsprechende Mittelwerte in der Zeit. Anwenden des θ-Schemas auf das semidiskrete Problem (16.8) führt dann auf

Problem 16.11 (Volldiskretisierung)**.** Sei $u_h^{(0)} := u_{h,0}$. Finde $u_h^{(n)}$, $n \in \{1, 2, \ldots\}$, sodass

$$\int_\Omega \mathrm{d}_\tau u_h^{(n)} v_h \,\mathrm{d}x + \int_\Omega \nabla u_h^{(n-1+\theta)} \cdot \nabla v_h \,\mathrm{d}x = \int_\Omega f^{(n-1+\theta)} v_h \,\mathrm{d}x \tag{16.15}$$

für alle $v_h \in V_h$ und alle $t^{(n)} \in \{0, \tau, \ldots, T\}$ gilt.

Wie man mit wenig Mühe überprüfen kann, besitzt das volldiskrete Problem für $\theta \in [0, 1]$ stets eine eindeutige Lösung. Als nächstes leiten wir eine diskrete Energieabschätzung her, die wieder für die Fehleranalyse essentiell ist.

Satz 16.12 (Stabilität)**.** *Sei $\theta \in [1/2, 1]$. Dann gilt für die Lösung von Problem 16.11*

$$\|u_h^{(n)}\|^2_{L^2(\Omega)} + \sum_{k=1}^{n} \tau \|\nabla u_h^{(k-1+\theta)}\|^2_{L^2(\Omega)} \le C \left(\|u_{0,h}\|^2_{L^2(\Omega)} + \sum_{k=1}^{n} \tau \|f^{(k-1+\theta)}\|^2_{L^2(\Omega)} \right)$$

mit C unabhängig von der Schrittweite τ, vom Raum V_h, und vom Index n.

Beweis. Aus der Definition des Mittelwertes $u_h^{(n-1+\theta)} = (1-\theta)u_h^{(n-1)} + \theta u_h^{(n)}$ erhält man durch einfaches Nachrechnen die elementare Formel

$$|u_h^{(n)}|^2 - |u_h^{(n-1)}|^2 = 2(u_h^{(n)} - u_h^{(n-1)}, u_h^{(n-1+\theta)}) - (2\theta - 1)|u_h^{(n)} - u_h^{(n-1)}|^2. \tag{16.16}$$

Wegen $\theta \ge 1/2$ haben wir

$$|u_h^{(n)}|^2 - |u_h^{(n-1)}|^2 \le 2(u_h^{(n)} - u_h^{(n-1)}, u_h^{(n-1+\theta)}) = 2\tau(\mathrm{d}_\tau u_h^{(n)}, u_h^{(n-1+\theta)}).$$

Integriert man über das Gebiet Ω und setzt für den letzten Term in das diskrete Schema (16.15) mit $v_h = u_h^{(n-1+\theta)}$ ein, so erhält man durch Umstellen

$$\begin{aligned}\tfrac{1}{2}\mathrm{d}_\tau\|u_h^{(n)}\|_{L^2}^2 &= \tfrac{1}{2\tau}\big(\|u_h^{(n)}\|_{L^2}^2 - \|u_h^{(n-1)}\|_{L^2}^2\big) \le (\mathrm{d}_\tau u_h^{(n)}, u_h^{(n-1+\theta)}) \\ &= (f^{(n-1+\theta)}, u_h^{(n-1+\theta)}) - (\nabla u_h^{(n-1+\theta)}, \nabla u_h^{(n-1+\theta)}) \\ &\le \tfrac{\epsilon}{2}\|f^{(n-1+\theta)}\|_{L^2} + \frac{1}{2\epsilon}\|u_h^{(n-1+\theta)}\|_{L^2}^2 - \|\nabla u_h^{(n-1+\theta)}\|_{L^2}^2.\end{aligned}$$

Mit der Friedrichs-Ungleichung folgt $\|u_h^{(n-1+\theta)}\|_{L^2} \le C_F\|\nabla u_h^{(n-1+\theta)}\|_{L^2}$, und mit der Wahl $\epsilon = C_F^2$ erhält man

$$\mathrm{d}_\tau\|u_h^{(n)}\|_{L^2}^2 + \|\nabla u_h^{(n-1+\theta)}\|_{L^2}^2 \le C_F^2\|f^{(n-1+\theta)}\|_{L^2}^2.$$

Die Energie-Abschätzung aus dem Lemma folgt dann unmittelbar durch Multiplikation mit τ und Aufsummieren bezüglich n. □

Bemerkung 16.13 (Stabilität insbesondere für das explizite Euler-Verfahren)**.** Für $\theta \in [0, 1/2)$, was den Fall des expliziten Euler-Verfahrens ($\theta = 0$) einschließt, besitzt der zweite Term in (16.16) ein ,,falsches" Vorzeichen und wirkt sich negativ auf die Stabilität aus. Mit der inversen Ungleichung und dem diskreten Gronwall-Lemma (Satz 2.14) kann man jedoch zeigen, dass die a-priori Schranke aus dem Lemma immer noch gilt, solange mit geeigneter Konstante c

$$\tau \le ch^2$$

gilt. Man erhält also eine stabile Lösung nur für hinreichend kleine Schrittweite τ. Das ist kein Wunder, denn das bei der Ortsdiskretisierung mit FEM erzielte DGL-System ist steif, das explitite Euler-Verfahren (wie auch das θ-Schema generell für $\theta \in [0, 1/2)$) ist nicht A-stabil.

Im Gegensatz dazu sind das implizite Euler-Verfahren und die implizite Trapezregel (wie auch generell das θ-Schema für $\theta \in [1/2, 1]$) A-stabil.

Fehleranalyse für das volldiskrete Verfahren

Zum Abschluss wenden wir uns wieder der Fehleranalyse zu. Hierzu zerlegen wir den Fehler zunächst mittels

$$\|u(t^{(n)}) - u_h^{(n)}\| \le \|u(t^{(n)}) - \tilde{u}_h^{(n)}\| + \|\tilde{u}_h^{(n)} - u_h^{(n)}\| \tag{16.17}$$

wieder in einen Approximationsfehler und einen diskreten Fehler. Wie bei der Semidiskretisierung, definieren wir $\tilde{u}_h^{(n)} := \tilde{\Pi}_h u(t^{(n)})$ über die elliptische Projektion. Der Approximationsfehler

$$\rho_h^{(n)} := \tilde{u}_h^{(n)} - u(t^{(n)}) = \tilde{\Pi}_h u(t^{(n)}) - u(t^{(n)})$$

lässt sich dann wieder mit den Aussagen über die elliptische Projektion ((16.10) und (16.11)) abschätzen.

Für den diskreten Fehler $d_h^{(n)} := \tilde{u}_h^{(n)} - u_h^{(n)}$ leiten wir erneut eine entsprechende diskrete Fehlergleichung her:

$$\begin{aligned}
&\int_\Omega \mathrm{d}_\tau d_h^{(n)} v_h \,\mathrm{d}x + \int_\Omega \nabla d_h^{(n-1+\theta)} \cdot \nabla v_h \,\mathrm{d}x \\
&= \int_\Omega \mathrm{d}_\tau \tilde{u}_h^{(n)} v_h \,dx + \int_\Omega \nabla \tilde{u}_h^{(n-1+\theta)} \cdot \nabla v_h \,\mathrm{d}x - \int_\Omega f^{(n-1+\theta)}\, v_h \,\mathrm{d}x \\
&= \int_\Omega \mathrm{d}_\tau \tilde{u}_h^{(n)} v_h \,dx + \int_\Omega \nabla (\tilde{u}_h^{(n-1+\theta)} - u^{(n-1+\theta)}) \cdot \nabla v_h \,\mathrm{d}x - \int_\Omega \partial_t u^{(n-1+\theta)}\, v_h \,\mathrm{d}x \\
&= \int_\Omega \mathrm{d}_\tau \rho_h^{(n)} v_h \,\mathrm{d}x + \int_\Omega \nabla \rho_h^{(n-1+\theta)} \cdot \nabla v_h \,\mathrm{d}x + \int_\Omega (\mathrm{d}_\tau u^{(n)} - \partial_t u^{(n-1+\theta)})\, v_h \,\mathrm{d}x \\
&= \int_\Omega \mathrm{d}_\tau \rho_h^{(n)} v_h \,\mathrm{d}x + \int_\Omega (\mathrm{d}_\tau u^{(n)} - \partial_t u^{(n-1+\theta)})\, v_h \,\mathrm{d}x.
\end{aligned}$$

Für den ersten Schritt nutzen wir neben der Definition von $d_h^{(n)}$, dass $u_h^{(n)}$ das diskrete Schema (16.15) erfüllt. Im zweiten Schritt wurde hier das gewichtete Mittel der variationellen Identität (16.4) bei $t^{(n)}$ und $t^{(n-1)}$ mit Testfunktion $v = v_h$ verwendet, wobei wir hier die Kurzschreibweisen $u^{(n-1+\theta)} = (1-\theta)u(t^{(n-1)}) + \theta u(t^{(n)})$ und $\partial_t u^{(n-1+\theta)} = (1-\theta)\partial_t u(t^{(n-1)}) + \theta \partial_t u(t^{(n)})$ verwenden. Im dritten Schritt verwenden wir die Definition des Approximationsfehlers $\rho_h^{(n)}$ und im lezten Schritt die Galerkin-Orthogonalität der elliptischen Projektion.

Wie zuvor ist der diskrete Fehler $d_h^{(n)}$ also wieder Lösung des diskreten Problems, hier (16.15), mit spezieller rechter Seite

$$\tilde{f}^{(n-1+\theta)} = \mathrm{d}_\tau \rho_h^{(n)} + (\mathrm{d}_\tau u^{(n)} - \partial_t u^{(n-1+\theta)}).$$

Mit Taylorentwicklung in der Zeit und Abschätzungen für die elliptische Projektion lässt sich die rechte Seite $\tilde{f}^{(n-1+\theta)}$ nun wie folgt abschätzen.

Satz 16.14. *Sei u hinreichend glatte Lösung von* (16.1) – (16.3). *Dann gilt*

$$\tau \|\tilde{f}^{(n-1+\theta)}\|^2_{L^2(\Omega)} \le C \left(h^{2\ell} \int_{t^{(n-1)}}^{t^{(n)}} \|\partial_t u(s)\|^2_{H^2(\Omega)} \,\mathrm{d}s + \tau^{2r} \int_{t^{(n-1)}}^{t^{(n)}} \|\partial_t^{r+1} u(s)\|^2_{L^2(\Omega)} \,\mathrm{d}s \right)$$

mit $r = 1$ für alle $\theta \in [0,1]$ und $r = 2$ nur im Fall $\theta = 1/2$ sowie $\ell = 1$ stets, $\ell = 2$ nur unter den Voraussetzungen für das Aubin-Nitsche-Resultat.

Beweis. Zunächst gilt $\|\tilde{f}^{(n-1+\theta)}\|^2_{L^2} \le 2\|\mathrm{d}_\tau \rho_h^{(n)}\|^2_{L^2} + 2\|\partial_t u^{(n-1+\theta)} - \mathrm{d}_\tau u^{(n)}\|^2_{L^2}$. Mit den Definitionen von d_τ und $\rho_h^{(n)}$, der Approximationsfehlerabschätzungen (16.10)

bzw. (16.11), dem Hauptsatz der Differential- und Integralrechnung sowie der Cauchy-Schwarz-Ungleichung folgt dann für den ersten Teil

$$\tau\|\mathrm{d}_\tau\rho_h^{(n)}\|_{L^2}^2 = \tfrac{1}{\tau}\|\rho_h^{(n)} - \rho_h^{(n-1)}\|_{L^2}^2 = \tfrac{1}{\tau}\left\|\int_{t^{(n-1)}}^{t^{(n)}} \partial_t u(t) - \tilde{\Pi}_h\partial_t u(t)\,\mathrm{d}t\right\|_{L^2}^2$$
$$\leq \int_{t^{(n-1)}}^{t^{(n)}} \|\partial_t u(t) - \tilde{\Pi}_h\partial_t u(t)\|_{L^2}^2\,\mathrm{d}t \leq Ch^{2\ell}\int_{t^{(n-1)}}^{t^{(n)}} \|\partial_t u(t)\|_{H^2(\Omega)}^2\,\mathrm{d}t.$$

Zur Abschätzung des zweiten Terms gehen nun wir wie folgt vor: Aus dem Hauptsatz der Differential- und Integralrechnung folgt zunächst

$$\partial_t u^{(n-1-\theta)} = \theta\partial_t u(t^{(n)}) + (1-\theta)\partial_t u(t^{(n-1)}) = \partial_t u(t^{(n-1)}) + \theta\int_{t^{(n-1)}}^{t^{(n)}} \partial_{tt}u(s)\,\mathrm{d}s.$$

Durch Entwicklung bei $t^{(n-1)}$ folgt weiters

$$\mathrm{d}_\tau u^{(n)} = \tfrac{1}{\tau}(u(t^{(n)}) - u(t^{(n-1)})) = \partial_t u(t^{(n-1)}) + \tfrac{1}{\tau}\tfrac{1}{2!}\int_{t^{(n-1)}}^{t^{(n)}} (t^{(n)} - s)\partial_{tt}u(s)\,\mathrm{d}s,$$

wobei wir für $u(t^{(n)})$ Taylorentwicklung um $t^{(n-1)}$ mit integraler Restglieddarstellung verwendet haben. Subtraktion der beiden Ausdrücke führt dann auf

$$\partial_t u^{(n-1+\theta)} - \mathrm{d}_\tau u^{(n)} = \int_{t^{(n-1)}}^{t^{(n)}} (\theta - \tfrac{t^{(n)}-s}{2\tau})\partial_{tt}u(s)\,\mathrm{d}s.$$

Mittels Cauchy-Schwarz Ungleichung und $|\theta - \frac{t^{(n)}-s}{2\tau}| \leq 1$ erhält man

$$\tau\|\partial_t u^{(n-1+\theta)} - \mathrm{d}_\tau u^{(n)}\|_{L^2(\Omega)}^2 \leq \tau^2\int_{t^{(n-1)}}^{t^{(n)}} \|\partial_{tt}u(s)\|_{L^2(\Omega)}^2\,\mathrm{d}s.$$

Durch Kombination der beiden Abschätzungen erhält man die erste Aussage des Lemmas. Die verbesserte Abschätzung für $\theta = 1/2$ zeigt man analog, wenn man in den Taylorentwicklungen um einen Grad weiter geht; siehe Übung. □

Zusammen mit der diskreten Stabilitätsabschätzung und der Fehlerabschätzungen für die elliptische Projektion erhält man das folgende Resultat.

Satz 16.15 (Konvergenz)**.** *Sei u glatte Lösung von* (16.1) – (16.2) *und $(u_h^{(n)})_n$ Lösung von* (16.15) *mit $\theta \in [1/2, 1]$ und $u_h^{(0)} = \tilde{\Pi}_h u_0$. Dann gilt*

$$\|u(t^{(n)}) - u_h^{(n)}\|_{L^2}^2 + \sum_{k=1}^{n} \tau\|\nabla u^{(k-1+\theta)} - \nabla u_h^{(k-1+\theta)}\|_{L^2}^2 \leq C(u;T)\,(h^2 + \tau^{2r})$$

mit $r = 1$ für alle $\theta \in [0,1]$ und $r = 2$ nur für $\theta = 1/2$.

Mit dem *Aubin-Nitsche-Dualitätsargument* erhält man im 1D-Fall immer, im 2D-Fall unter den Voraussetzungen des Satzes 15.7 (also wenn $\Omega \subseteq \mathbb{R}^2$ konvex ist *oder* glatten Rand hat) auch für die in diesem Kapitel besprochene Wärmeleitgleichung folgende Abschätzung zeigen:

$$\|u(t^{(n)}) - u_h^{(n)}\|_{L^2}^2 \leq C(u;T)\ (h^4 + \tau^{2r}).$$

Für $\theta = 1/2$ erhält man also eine Methode, die in der L^2-Norm bezüglich h und τ quadratisch konvergiert.

Wie wir gesehen haben, lassen sich die Konvergenzresultate für die Finite-Elemente Methode für elliptische Probleme ohne größere Probleme auf parabolische Probleme übertragen. Zur Zeitdiskretisierung sollten aufgrund der Steifheit implizite Diskretisierungsverfahren verwendet werden, z.B. das implizite Euler-Verfahren oder die implizite Trapezregel; dies sind gerade die Verfahren niderigster Ordnung vom Radau-IIA bzw. Lobatto-IIIA Typ.

Numerische Tests

Zuletzt wollen wir noch die Resultate in einem numerischen Test veranschaulichen. Dabei beschränken wir uns auf das den Fall $f \equiv 0$, das als Abkühlproblem bekannt ist. Dabei sei eine Anfangswärmeverteilung u_0 gegeben. Wir wollen dann folgendes Problem lösen:

$$\begin{aligned} \partial_t u(x,t) - \partial_{xx} u(x,t) &= 0, && x \in \Omega = (0,1),\ t \in (0,T], \\ u(0,t) = u(1,t) &= 0, && t \in (0,T], \\ u(x,0) &= u_0(x), && x \in \Omega. \end{aligned}$$

Wenn wir annehmen, dass sich u_0 als Fourier-Reihe

$$u_0(x) = \sum_{i=1}^{\infty} a_i \sin(i\pi x)$$

darstellen lässt, dann lässt sich leicht nachprüfen, dass (bei Annahme von Konvergenz) durch

$$u(x,t) = \sum_{i=1}^{\infty} a_i \sin(i\pi x) e^{-(i\pi)^2 t}$$

eine Lösung gegeben ist. Die Theorie garantiert die Eindeutigkeit.

In den folgenden Beispielen wählen wir $T = 1/2$ und $u_0(x) = \sin(\pi x)$. Die Folgenden Beispiele entsprechen den Werten $\theta = 0, 1/2, 1$.

Beispiel 16.16 (Implizites Euler-Verfahren; $\theta = 1$)**.** Wir diskretisieren das Ortsproblem mit der Finiten-Elemente-Methode und dem Courant-Element mit einer äquidistanten Schrittweite $h = n^{-1}$ mit $n \in \mathbb{N}$. Bei Diskretisierung mit dem impliziten Euler-Verfahren, erhalten wir als Näherungen $\underline{u}_h^{(j)}$:

$$\tfrac{1}{\tau} M_h(\underline{u}_h^{(j)} - \underline{u}_h^{(j-1)}) + K_h \underline{u}_h^{(j)} = 0, \quad \text{also} \quad (\tfrac{1}{\tau} M_h + K_h)\underline{u}_h^{(j)} = \tfrac{1}{\tau} M_h \underline{u}_h^{(j-1)} \tag{16.18}$$

wobei $M_h, K_h \in \mathbb{R}^{(n-1)\times(n-1)}$ Masse- und Steifigkeitsmatrix sind:

$$M_h = \frac{h}{6}\begin{pmatrix} 4 & 1 & & \\ 1 & 4 & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & 4 \end{pmatrix} \quad \text{und} \quad K_h = \frac{1}{h}\begin{pmatrix} 2 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix}.$$

Wir können das ARWP daher nach Setzen von $\underline{u}_h^{(0)}$ auf die Anfangsbedingungen dadurch lösen, dass wir schrittweise für $j = 1, 2, \ldots$ das System (16.18) auflösen. Der *relative* Fehler in der diskreten Maximumsnorm (Auswertung des Fehlers an allen Gitterpunkten $(ih, j\tau)$) ist in Tabelle 16.1 gegeben.[1]

	$\tau = 2^{-13}$	$\tau = 2^{-14}$	$\tau = 2^{-15}$	$\tau = 2^{-16}$	$\tau = 2^{-17}$	$\tau = 2^{-18}$
$h = 2^{-3}$	$5.89\,10^{-2}$ (4 / 1)	$6.03\,10^{-2}$ (4 / 1)	$6.10\,10^{-2}$ (4 / 1)	$6.14\,10^{-2}$ (4 / 1)	$6.16\,10^{-2}$ (4 / 1)	$6.17\,10^{-2}$ (4 / 1)
$h = 2^{-4}$	$1.28\,10^{-2}$ (5 / .8)	$1.43\,10^{-2}$ (4 / .9)	$1.50\,10^{-2}$ (4 / 1)	$1.54\,10^{-2}$ (4 / 1)	$1.56\,10^{-2}$ (4 / 1)	$1.57\,10^{-2}$ (4 / 1)
$h = 2^{-5}$	$9.89\,10^{-4}$ (13 / 2)	$2.47\,10^{-3}$ (6 / .4)	$3.22\,10^{-3}$ (5 / .8)	$3.59\,10^{-3}$ (4 / .9)	$3.77\,10^{-3}$ (4 / 1)	$3.86\,10^{-3}$ (4 / 1)
$h = 2^{-6}$	$1.98\,10^{-3}$ (.5 / 3)	$4.95\,10^{-4}$ (5 / 4)	$2.48\,10^{-4}$ (13 / 2)	$6.19\,10^{-4}$ (6 / .4)	$8.05\,10^{-4}$ (5 / .8)	$8.98\,10^{-4}$ (4 / .9)
$h = 2^{-7}$	$2.73\,10^{-3}$ (.7 / 2)	$1.24\,10^{-3}$ (.4 / 2)	$4.95\,10^{-4}$ (.5 / 3)	$1.24\,10^{-4}$ (5 / 4)	$6.19\,10^{-5}$ (13 / 2)	$1.55\,10^{-4}$ (6 / .4)
$h = 2^{-8}$	$2.91\,10^{-3}$ (1 / 2)	$1.42\,10^{-3}$ (.9 / 2)	$6.81\,10^{-4}$ (.7 / 2)	$3.10\,10^{-4}$ (.4 / 2)	$1.24\,10^{-4}$ (.5 / 3)	$3.10\,10^{-5}$ (5 / 4)
$h = 2^{-9}$	$2.96\,10^{-3}$ (1 / 2)	$1.47\,10^{-3}$ (1 / 2)	$7.28\,10^{-4}$ (1 / 2)	$3.56\,10^{-4}$ (.9 / 2)	$1.70\,10^{-4}$ (.7 / 2)	$7.74\,10^{-5}$ (.4 / 2)

Tabelle 16.1.: Konvergenz implizites Euler-Verfahren.

Neben den jeweiligen Fehlerwerten $e(h, \tau)$ sind in der zweiten Zeile in blauer Farbe (erster Zahlenwert) auch die Quotienten $e(2h, \tau)/e(h, \tau)$ und in violetter Farbe (zweiter Zahlenwert) die auch die Quotienten $e(h, 2\tau)/e(h, \tau)$ zu sehen.

[1] Dies ist nicht die Norm, in der wir die theoretischen Aussagen gemacht haben. Die diskrete Maximumsnorm wählen wir, weil sie sehr einfach zu bestimmen ist. Der Grund für die Bestimmung des relativen Fehlers ist, dass die exakte Lösung mit der Zeit gegen 0 konvergiert. Da die Lösung des diskreten Problems dies (bei Wahl eines stabilen Verfahrens) auch tut, würden wir bei Betrachtung des absoluten Fehlers im Wesentlichen nur das Verhalten des Verfahrens während den ersten paar Zeitschritte messen.

Hinsichtlich der Konvergenz in h betrachten wir die ,,blauen" Quotienten; der Wert 4 deutet auf eine Konvergenz wie $\mathcal{O}(h^2)$ hin, die wir erzielen, wenn τ/h^2 hinreichend klein ist (oben rechts). Wird τ/h^2 zu groß, wird der Fehler nicht mehr kleiner, teilweise sogar leicht größer. Beim Übergang zwischen den beiden Regimen haben wir teilweise eine unerwartet gute Verbesserung des Fehlers.

Hinsichtlich der Konvergenz in τ betrachten wir die ,,violetten" Quotienten; der Wert 2 deutet auf eine Konvergenz von $\mathcal{O}(\tau)$ hin, die wir erzielen, wenn h^2/τ hinreichend klein ist (unten links). Wird h^2/τ zu groß, wird der Fehler nicht mehr kleiner, teilweise sogar leicht größer. Zusammenfassend zeigt die Tabelle ein Bild, das mit einer Fehlerschranke $\mathcal{O}(h^2+\tau)$ vereinbar ist. Da sich der Fehler wie $\mathcal{O}(h^2+\tau)$ verhält, ist $\tau \approx h^2$ zu bevorzugen, um die Fehler aus Orts- und Zeitdiskretisierung in gleicher Größenordnung zu halten.

Beispiel 16.17 (Explizites Euler-Verfahren; $\theta = 0$)**.** Nach Ortsdiskretisierung wie im letzten Beispiel, haben wir nun

$$\tfrac{1}{\tau} M_h(\underline{u}_h^{(j+1)} - \underline{u}_h^{(j)}) + K_h \underline{u}_h^{(j)} = 0, \quad \text{also} \quad \tfrac{1}{\tau} M_h \underline{u}_h^{(j+1)} = (\tfrac{1}{\tau} M_h - K_h)\underline{u}_h^{(j)}, \tag{16.19}$$

wobei $M_h, K_h \in \mathbb{R}^{(n-1)\times(n-1)}$ Masse- und Steifigkeitsmatrix wieder wie oben sind. Der *relative* Fehler in der diskreten Maximumsnorm ist in Tabelle 16.2 gegeben, wobei die Fälle, in denen die Methode nicht konvergierte, mit ,,div" gekennzeichnet sind.

	$\tau = 2^{-13}$	$\tau = 2^{-14}$	$\tau = 2^{-15}$	$\tau = 2^{-16}$	$\tau = 2^{-17}$	$\tau = 2^{-18}$
$h = 2^{-3}$	$6.46\,10^{-2}$ (4 / 1)	$6.32\,10^{-2}$ (4 / 1)	$6.25\,10^{-2}$ (4 / 1)	$6.21\,10^{-2}$ (4 / 1)	$6.19\,10^{-2}$ (4 / 1)	$6.18\,10^{-2}$ (4 / 1)
$h = 2^{-4}$	$1.87\,10^{-2}$ (3 / 1)	$1.72\,10^{-2}$ (4 / 1)	$1.65\,10^{-2}$ (4 / 1)	$1.61\,10^{-2}$ (4 / 1)	$1.59\,10^{-2}$ (4 / 1)	$1.58\,10^{-2}$ (4 / 1)
$h = 2^{-5}$	$6.92\,10^{-3}$ (3 / –)	$5.44\,10^{-3}$ (3 / 1)	$4.70\,10^{-3}$ (4 / 1)	$4.33\,10^{-3}$ (4 / 1)	$4.14\,10^{-3}$ (4 / 1)	$4.05\,10^{-3}$ (4 / 1)
$h = 2^{-6}$	div (– / –)	div (– / –)	$1.73\,10^{-3}$ (3 / –)	$1.36\,10^{-3}$ (3 / 1)	$1.18\,10^{-3}$ (4 / 1)	$1.08\,10^{-3}$ (4 / 1)
$h = 2^{-7}$	div (– / –)	div (– / –)	div (– / –)	div (– / –)	$4.33\,10^{-4}$ (3 / –)	$3.41\,10^{-4}$ (3 / 1)
$h = 2^{-8}$	div (– / –)	div (– / –)	div (– / –)	div (– / –)	div (– / –)	div (– / –)

Tabelle 16.2.: Konvergenz explizites Euler-Verfahren.

Wir können also klar erkennen, dass die Konvergenz nur dann gegeben ist, wenn die in Bedingung aus Bemerkung 16.13 ($\tau \lesssim h^2$) erfüllt ist. Andernfalls divergiert das Verfahren tatsächlich.

Auch hier zeigen wir neben den Angaben zum Fehler wieder die beiden Quotienten $e(2h,\tau)/e(h,\tau)$ (erster Zahlenwert; blau) und $e(h,2\tau)/e(h,\tau)$ (zweiter Zahlenwert; violett). Hier können wir wieder klar die h-Konvergenzordnung erkennen, wie beim impliziten Verfahren. In den Bereichen, in denen das Verfahren konvergiert muss τ/h^2 klein genug sein. Hier ist der Fehlerterm $\mathcal{O}(h^2)$ dominant, daher sehen wir keine Verbesserung bei Verfeinerung von τ. Zusammenfassend zeigt die Tabelle ein Bild, das mit einer Fehlerschranke $\mathcal{O}(h^2+\tau)$ vereinbar ist.

Beispiel 16.18 (Vereinfachtes explizites Euler-Verfahren)**.** Auch bei Verwendung von (16.19) müssen wir immer noch ein lineares Gleichungssystem auflösen. Wenn wir aber das Problem mit der Finiten-Differenzen-Methode diskretisieren würden, oder wenn wir (was äquivalent ist!) $M_h := hI$ wählen, dann erhalten wir eine Vorschrift, die ohne Lösen eines linearen Gleichungssystems auskommt:

$$\underline{u}_h^{(j+1)} = (I - \tfrac{\tau}{h} K_h)\underline{u}_h^{(j)},$$

wobei K_h immer noch die FEM-Steifigkeitsmatrix wie oben ist. Der *relative* Fehler in der diskreten Maximumsnorm ist in Tabelle 16.3 gegeben. Qualitativ ändert sich hier nichts im Vergleich zum klassischen expliziten Euler-Verfahren.

	$\tau = 2^{-13}$	$\tau = 2^{-14}$	$\tau = 2^{-15}$	$\tau = 2^{-16}$	$\tau = 2^{-17}$	$\tau = 2^{-18}$
$h = 2^{-3}$	$6.20\,10^{-2}$ (4 / 1)	$6.36\,10^{-2}$ (4 / 1)	$6.44\,10^{-2}$ (4 / 1)	$6.47\,10^{-2}$ (4 / 1)	$6.49\,10^{-2}$ (4 / 1)	$6.50\,10^{-2}$ (4 / 1)
$h = 2^{-4}$	$1.30\,10^{-2}$ (5 / .8)	$1.45\,10^{-2}$ (4 / .9)	$1.52\,10^{-2}$ (4 / 1)	$1.56\,10^{-2}$ (4 / 1)	$1.58\,10^{-2}$ (4 / 1)	$1.59\,10^{-2}$ (4 / 1)
$h = 2^{-5}$	$9.93\,10^{-4}$ (13 / 2)	$2.48\,10^{-3}$ (6 / .4)	$3.23\,10^{-3}$ (5 / .8)	$3.60\,10^{-3}$ (4 / .9)	$3.78\,10^{-3}$ (4 / 1)	$3.88\,10^{-3}$ (4 / 1)
$h = 2^{-6}$	$1.98\,10^{-3}$ (.5 / –)	$4.95\,10^{-4}$ (5 / 4)	$2.48\,10^{-4}$ (13 / 2)	$6.20\,10^{-4}$ (6 / .4)	$8.05\,10^{-4}$ (5 / .8)	$8.98\,10^{-4}$ (4 / .9)
$h = 2^{-7}$	div (– / –)	div (– / –)	$4.95\,10^{-4}$ (.5 / –)	$1.24\,10^{-4}$ (5 / 4)	$6.19\,10^{-5}$ (13 / 2)	$1.55\,10^{-4}$ (6 / .4)
$h = 2^{-8}$	div (– / –)	div (– / –)	div (– / –)	div (– / –)	$1.24\,10^{-4}$ (.5 / –)	$3.10\,10^{-5}$ (5 / 4)

Tabelle 16.3.: Konvergenz vereinfachtes explizites Euler-Verfahren.

Beispiel 16.19 (Implizite Trapezregel / Crank-Nicolson-Verfahren; $\theta = 1/2$)**.** Für das θ-Verfahren mit $\theta = 1/2$ erhalten wir die implizite Trapezregel

$$(\tfrac{1}{\tau} M_h + \tfrac{1}{2} K_h)\underline{u}_h^{(j+1)} = (\tfrac{1}{\tau} M_h - \tfrac{1}{2} K_h)\underline{u}_h^{(j)}, \tag{16.20}$$

wobei $M_h, K_h \in \mathbb{R}^{(n-1)\times(n-1)}$ Masse- und Steifigkeitsmatrix wieder wie in Beispiel 16.16 sind. Dieses Verfahren ist (wenn die Ortsdiskretisierung mit der FDM

erfolgt) auch als Crank-Nicolson-Verfahren bekannt. Wir können das ARWP daher nach Setzen von $\underline{u}_h^{(0)}$ auf die Anfangsbedingungen dadurch lösen, dass wir schrittweise für $j = 0, 1, 2, \ldots$ das System (16.19) auflösen. Der *relative* Fehler in der diskreten Maximumsnorm ist in Tabelle 16.4 gegeben.

	$\tau = 2^{-6}$	$\tau = 2^{-7}$	$\tau = 2^{-8}$	$\tau = 2^{-9}$	$\tau = 2^{-10}$	$\tau = 2^{-11}$
$h = 2^{-3}$	$7.13\,10^{-2}$ (3 / 1)	$6.41\,10^{-2}$ (4 / 1)	$6.23\,10^{-2}$ (4 / 1)	$6.19\,10^{-2}$ (4 / 1)	$6.18\,10^{-2}$ (4 / 1)	$6.18\,10^{-2}$ (4 / 1)
$h = 2^{-4}$	$2.55\,10^{-2}$ (3 / 2)	$1.82\,10^{-2}$ (4 / 1)	$1.64\,10^{-2}$ (4 / 1)	$1.59\,10^{-2}$ (4 / 1)	$1.58\,10^{-2}$ (4 / 1)	$1.58\,10^{-2}$ (4 / 1)
$h = 2^{-5}$	$1.37\,10^{-2}$ (2 / 3)	$6.40\,10^{-3}$ (3 / 2)	$4.57\,10^{-3}$ (4 / 1)	$4.11\,10^{-3}$ (4 / 1)	$4.00\,10^{-3}$ (4 / 1)	$3.97\,10^{-3}$ (4 / 1)
$h = 2^{-6}$	$1.08\,10^{-2}$ (1 / 4)	$3.43\,10^{-3}$ (2 / 3)	$1.60\,10^{-3}$ (3 / 2)	$1.14\,10^{-3}$ (4 / 1)	$1.03\,10^{-3}$ (4 / 1)	$1.00\,10^{-3}$ (4 / 1)
$h = 2^{-7}$	$1.00\,10^{-2}$ (1 / 4)	$2.69\,10^{-3}$ (1 / 4)	$8.59\,10^{-4}$ (2 / 3)	$4.00\,10^{-4}$ (3 / 2)	$2.86\,10^{-4}$ (4 / 1)	$2.57\,10^{-4}$ (4 / 1)
$h = 2^{-8}$	$9.83\,10^{-3}$ (1 / 4)	$2.51\,10^{-3}$ (1 / 4)	$6.73\,10^{-4}$ (1 / 4)	$2.15\,10^{-4}$ (2 / 3)	$1.00\,10^{-4}$ (3 / 2)	$7.15\,10^{-5}$ (4 / 1)
$h = 2^{-9}$	$9.78\,10^{-3}$ (1 / 4)	$2.46\,10^{-3}$ (1 / 4)	$6.27\,10^{-4}$ (1 / 4)	$1.68\,10^{-4}$ (1 / 4)	$5.37\,10^{-5}$ (2 / 3)	$2.50\,10^{-5}$ (3 / 2)

Tabelle 16.4.: Konvergenz implizite Trapezregel.

Bei der impliziten Trapezregel, die auch für alle Kombinationen aus τ und h stabil ist, können wir beobachten, dass sich der Fehler wie $\mathcal{O}(h^2 + \tau^2)$ verhält. (Die Tabelle für die implizite Trapezregel gibt die Fehlerwerte für viel größere Zeitschrittweiten an, als die anderen Tabellen.)

Auch hier zeigen wir neben den Angaben zum Fehler wieder die beiden Quotienten $e(2h, \tau)/e(h, \tau)$ und $e(h, 2\tau)/e(h, \tau)$. Für den Fall τ/h klein (rechts oben) erhalten wir eine Konvergenz wie $\mathcal{O}(h^2)$ und für den Fall h/τ klein (links unten) eine Konvergenz wie $\mathcal{O}(\tau^2)$. Insgesamt passt das Verhalten zur vorhergesagten Rate $\mathcal{O}(\tau^2 + h^2)$. Hier ist nun $\tau \approx h$ zu bevorzugen.

Für das hier behandelte Problem benötigte die implizite Trapezregel, die gegenüber dem impliziten Euler-Verfahren praktisch keinen Zusatzaufwand verursacht, die geringste Rechenzeit.

17. Hyperbolische Differentialgleichungen

Zum Abschluss der Vorlesung besprechen wir jetzt noch die numerische Lösung von hyperbolischen Differentialgleichungen zweiter Ordnung, welche bei der Modellierung von Schwingungsphänomenen auftauchen. Als Modellproblem betrachten wir die lineare Wellengleichung mit homogenen Randbedingungen

$$\partial_{tt} u(x,t) - \Delta u(x,t) = f(x,t) \qquad x \in \Omega,\ t > 0, \tag{17.1}$$

$$u(x,t) = 0 \qquad x \in \partial\Omega,\ t > 0, \tag{17.2}$$

welche nun mit zwei Anfangsbedingungen zu komplettieren ist, nämlich

$$u(x,0) = u_0(x), \quad \partial_t u(x,0) = v_0(x) \qquad x \in \Omega. \tag{17.3}$$

Wie im vorhergehenden Kapitel stellen wir zunächst einige Resultate zur Analyse des Problems vor und wenden uns dann der Diskretisierung mittels Finite-Elemente Verfahren im Ort und geeigneter Zeitschrittverfahren zu.

Lösbarkeit und Energieabschätzung

Mit denselben Argumenten, wie sie bei der Behandlung der Poisson und Wärmeleitungsgleichung verwendet wurden, erhält man folgende Aussage.

Satz 17.1. *Sei u eine glatte Lösung von* (17.1) – (17.2)*. Dann gilt*

$$\int_\Omega \partial_{tt} u(x,t) v(x)\,\mathrm{d}x + \int_\Omega \nabla u(x,t) \nabla v(x)\,\mathrm{d}x = \int_\Omega f(x,t) v(x)\,\mathrm{d}x \tag{17.4}$$

für alle Testfunktionen $v \in H_0^1(\Omega)$ und alle $t \geq 0$.

Eine Funktion u hinreichender Regularität, welche die variationellen Identitäten (17.4) und auch noch die Anfangsbedingungen erfüllt, nennnen wir eine schwache Lösung des Anfangs-Randwertproblems (17.1) – (17.3). Der folgende Satz fasst wieder wichtige Aussagen zu solchen schwachen Lösungen zusammen.

Satz 17.2. *Sei $\Omega \subset \mathbb{R}^d$ ein beschränktes Lipschitz-Gebiet und f eine hinreichend glatte Funktion. Dann besitzt* (17.1) – (17.3) *für alle $u_0, v_0 \in H_0^1(\Omega)$ genau eine schwache Lösung $u \in C^2([0,T]; L^2(\Omega)) \cap C([0,T]; H_0^1(\Omega))$. Weiters gilt*

$$E(t) \leq e^{\alpha t} E(0) + \frac{1}{2\alpha} \int_0^t e^{\alpha(t-s)} \|f(s)\|^2_{L^2(\Omega)} \mathrm{d}s$$

für jedes $\alpha > 0$ und $t \geq 0$, wobei

$$E(t) = \tfrac{1}{2}\|\partial_t u(t)\|^2_{L^2(\Omega)} + \tfrac{1}{2}\|\nabla u(t)\|^2_{L^2(\Omega)}$$

die Energie der Lösung bezeichnet, bestehend aus kinetischem und potentiellem Anteil.

Beweis. Für die Existenzaussage verweisen wir auf das Buch von *Evans: Partial Differential Equations*. Durch Differenzieren der Energie erhält man

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} E(t) &= \int_\Omega \partial_{tt} u(t) \partial_t u(t) \,\mathrm{d}x + \int_\Omega \nabla u(t) \cdot \nabla \partial_t u(t) \,\mathrm{d}x \\ &= \int_\Omega f(t) \partial_t u(t) \,\mathrm{d}x \leq \frac{\alpha}{2} \|f(t)\|^2_{L^2(\Omega)} + \alpha E(t). \end{aligned} \tag{17.5}$$

Dabei haben wir im ersten Schritt der zweiten Zeile einfach Gleichung (17.4) mit $v = \partial_t u(t)$ verwendet, im letzten Schritt dann die Cauchy-Schwarz-Ungleichung, die Young'sche Ungleichung und im Anschluss $\frac{1}{2}\|\partial_t u(t)\|^2_{L^2(\Omega)} \leq E(t)$. Die Energieabschätzung folgt nun unmittelbar mit Hilfe des Gronwall-Lemmas (Satz 2.3). Aus der Energieabschätzung folgt unmittelbar die Eindeutigkeit. □

Bemerkung 17.3. Im wesentlichen Schritt des Beweises wurde die Variationsgleichung (17.4) mit $v = \partial_t u(t)$ verwendet. Wie bei der Wärmeleitungsgleichung kann man durch Testen mit anderen Funktionen auch wieder Abschätzungen in stärkeren Normen gewinnen.

Falls $f \equiv 0$ ist, dann erhält man aus (17.5) direkt, dass $\frac{\mathrm{d}}{\mathrm{d}t} E(t) = 0$, also die Energieerhaltungseigenschaft $E(t) = E(0)$. Zur Anschaung: In einer perfekt abgeschlossenen Box (Dirichlet-Randbedingung) wird eine Welle (ohne Dämpfung) ewig hin- und her oszillieren und entsprechend die Energie erhalten bleiben.

Ortsdiskretisierung mit FEM

Als ersten Schritt betrachten wir die numerische Approximation im Ort, wofür wir wie schon bei elliptischen und parabolischen Problemen wieder eine Galerkin-Methode verwenden wollen. Wir wählen dafür einen endlich-dimensionalen Teilraum $V_h \subset H_0^1(\Omega)$ und betrachten die folgende Methode.

Problem 17.4. Seien $u_{h,0}, v_{h,0} \in V_h$ gegebe Anfangswerte.
Finde $u_h : [0,T] \to V_h$ mit $u_h(0) = u_{h,0}$ und $\partial_t u_h(0) = v_{h,0}$ so, dass

$$\int_\Omega \partial_{tt} u_h(x,t) v_h(x)\,\mathrm{d}x + \int_\Omega \nabla u_h(x,t) \nabla v_h(x)\,\mathrm{d}x = \int_\Omega f(x,t) v_h(x)\,\mathrm{d}x \qquad (17.6)$$

für alle Testfunktionen $v_h \in V_h$ und alle $t \in (0,T]$ erfüllt ist.

Dass dies ein vernünftiges Verfahren ist, belegt folgendes Resultat.

Satz 17.5. *Sei $V_h \subset H_0^1(\Omega)$ mit $\dim(V_h) < \infty$ und $f \in C([0,T]; L^2(\Omega))$. Dann hat Problem 17.4 eine eindeutige Lösung $u_h \in C^2([0,T]; V_h)$ und es gilt*

$$E_h(t) \le e^{\alpha t} E_h(0) + \frac{1}{2\alpha} \int_0^t e^{\alpha(t-s)} \|f(s)\|^2_{L^2(\Omega)} \mathrm{d}s$$

für jedes $\alpha > 0$ und $t \ge 0$ mit diskreter Energie $E_h(t) := \frac{1}{2}\|\partial_t u_h(t)\|^2_{L^2(\Omega)} + \frac{1}{2}\|\nabla u_h(t)\|^2_{L^2(\Omega)}$.

Beweis. Die diskrete Variationsgleichung (17.6) aus der Definition des Problems ist äquivalent zu folgendem System gewöhnlicher Differentialgleichungen:

$$M_h \partial_{tt} \underline{u}_h(t) + A_h \underline{u}_h(t) = \underline{b}_h(t),$$

wobei $\underline{u}_h(t)$ wieder den Koordinatenvektor der Funktion $u_h(\cdot,t)$ bezüglich einer gewählten Basis von V_h darstellt. Entsprechend sind A_h und M_h die Steifigkeits- und Massenmatrix und $\underline{b}_h(t)$ der Lastvektor der die rechte Seite $f(\cdot,t)$ repräsentiert. Mittels $\underline{v}_h(t) := \partial_t \underline{u}_h(t)$ und der diskreten Anfangswerte erhalten wir dann

$$\begin{aligned} M_h \partial_t \underline{u}_h(t) &= M_h \underline{v}_h(t), & \underline{u}_h(0) &= \underline{u}_{h,0}, \\ M_h \partial_t \underline{v}_h(t) &= \underline{b}_h(t) - A_h \underline{u}_h(t), & \underline{v}_h(0) &= \underline{v}_{h,0}. \end{aligned}$$

Da M_h regulär ist, können wir beide Zeilen mit M_h^{-1} multiplizieren und erhalten ein lineares System gewöhnlicher Differentialgleichungen mit entsprechenden Anfangswerten. Existenz und Eindeutigkeit der Lösung ist durch den Satz von Picard-Lindelöf garantiert. Die Energieabschätzung zeigt man analog zum kontinuierlichen Level; siehe Übung. Sie garantiert auch wieder die Eindeutigkeit. □

Fehlerabschätzung

Als spezielle Approximation betrachten wir eine Finite-Elemente Methode mit stetigen und stückweise linearen Ansatzfunktionen, also $V_h = P_1(T_h) \cap H_0^1(\Omega)$.

Den Fehler können wir dann wieder mittels

$$\|u(t) - u_h(t)\| \leq \| \underbrace{\tilde{u}_h(t) - u(t)}_{=:\rho_h(t)} \| + \| \underbrace{\tilde{u}_h(t) - u_h(t)}_{=:d_h(t)} \|$$

in einen Approximationsfehler ρ_h und einen diskreten Fehler d_h zerlegen. Als Hilfsfunktion $\tilde{u}_h$ wählen wir wie im vorigen Kapitel $\tilde{u}_h(t) = \tilde{\Pi}_h u(t)$, wobei $\tilde{\Pi}_h$ die elliptische Projektion bezeichnet. Satz 16.4 gibt uns wieder die erforderlichen Aussagen über den Approximationsfehler.

Diskreter Fehler

Aus den Gleichungen (17.6) und (17.4) erhält man

$$\int_\Omega \partial_{tt}(u_h(t) - u(t))v_h \,\mathrm{d}x + \int_\Omega \nabla(u_h(t) - u(t)) \cdot \nabla v_h \,\mathrm{d}x = 0$$

für alle t und für $v_h \in V_h$, also die Galerkin-Orthogonalität. Mit dieser Aussage und der Definition von d_h folgt

$$\begin{aligned}\int_\Omega \partial_{tt} d_h(t) v_h \,\mathrm{d}x + \int_\Omega \nabla d_h(t) \cdot \nabla v_h \,\mathrm{d}x \\ &= \int_\Omega \partial_{tt}(\tilde{u}_h(t) - u_h(t))v_h \,\mathrm{d}x + \int_\Omega \nabla(\tilde{u}_h(t) - u_h(t)) \cdot \nabla v_h \,\mathrm{d}x \\ &= \int_\Omega \partial_{tt}(\tilde{u}_h(t) - u(t))v_h \,\mathrm{d}x + \int_\Omega \nabla(\tilde{u}_h(t) - u(t)) \cdot \nabla v_h \,\mathrm{d}x.\end{aligned}$$

Nach Definition (16.13) der elliptischen Projektion $\tilde{u}_h(t) = \tilde{\Pi}_h u(t)$ fällt der letzte Term weg. Wir erhalten somit die diskrete Fehlergleichung

$$\int_\Omega \partial_{tt} d_h(t) v_h \,\mathrm{d}x + \int_\Omega \nabla d_h(t) \cdot \nabla v_h \,\mathrm{d}x = \int_\Omega \partial_{tt} \rho_h(t) v_h \,\mathrm{d}x. \tag{17.7}$$

Der diskrete Fehler d_h ist also Lösung eines diskreten Problems mit spezieller rechter Seite $\tilde{f}(t) = \partial_{tt}\rho_h(t)$. Mit Hilfe der Energieabschätzung aus Satz 17.5 (mit der Wahl $\alpha = 1$) erhalten wir

$$\begin{aligned}&\|\partial_t d_h(t)\|^2_{L^2(\Omega)} + \|\nabla d_h(t)\|^2_{L^2(\Omega)} \\ &\quad \leq e^t \left(\|\partial_t d_h(0)\|^2_{L^2(\Omega)} + \|\nabla d_h(0)\|^2_{L^2(\Omega)} + \int_0^t \|\partial_{tt}\rho_h(s)\|^2_{L^2(\Omega)} \mathrm{d}s \right).\end{aligned} \tag{17.8}$$

Wir wählen nun die Startwerte im diskreten Problem mittels

$$u_{h,0} = \tilde{\Pi}_h u_0 \qquad \text{und} \qquad v_{h,0} = \tilde{\Pi}_h v_0 \tag{17.9}$$

und erhalten dadurch $d_h(0) = \partial_t d_h(0) = 0$. Zusammen mit dem Splitting in Approximationsfehler und diskreten Fehler und den Abschätzungen für den Approximationsfehler aus Satz 16.5 ergibt sich nun unmittelbar folgendes Resultat.

Satz 17.6. *Sei u eine hinreichend reguläre Lösung von* (17.1) – (17.3) *und u_h die entsprechende Finite-Elemente Lösung von Problem 17.4 mit $V_h = P_1(T_h) \cap H_0^1(\Omega)$ und diskreten Anfangswerten* (17.9)*. Dann gilt*

$$\max_{t\in[0,T]} \|\partial_t u(t) - \partial_t u_h(t)\|^2_{L^2(\Omega)} + \int_0^T \|\nabla u(t) - \nabla u_h(t)\|^2_{L^2(\Omega)} \,\mathrm{d}t \le C(u;T)h^2.$$

Die Konstante $C(u;T)$ hängt dabei nicht von t oder h ab.

Beweis. Wie im Fall des parablolischen Problems nutzen wir, dass den Fehler aufspalten können: $u - u_h = d_h - \rho_h$. Mit (17.8) und (17.11) folgt sofort

$$\begin{aligned}
&\max_{t\in[0,T]} \|\partial_t u(t) - \partial_t u_h(t)\|^2_{L^2(\Omega)} + \int_0^T \|\nabla u(t) - \nabla u_h(t)\|^2_{L^2(\Omega)} \,\mathrm{d}t \\
&\le 2 \max_{t\in[0,T]} \|\partial_t \rho_h(t)\|^2_{L^2(\Omega)} + 2\int_0^T \|\nabla \rho_h(t)\|^2_{L^2(\Omega)} \,\mathrm{d}t + 2e^T \int_0^T \|\partial_{tt} \rho_h(t)\|^2_{L^2(\Omega)} \,\mathrm{d}t
\end{aligned}$$

Mit $\|\cdot\|_{L^2(\Omega)} \le \|\cdot\|_{H^1(\Omega)}$ sowie $\|\nabla \cdot\|_{L^2(\Omega)} \le \|\cdot\|_{H^1(\Omega)}$ und den Abschätzungen in Satz 16.5 über den Fehler der elliptische Projektion erhalten wir nun sofort:

$$\begin{aligned}
&\|\partial_t d_h(t)\|^2_{L^2(\Omega)} + \|\nabla d_h(t)\|^2_{L^2(\Omega)} \le C(u;T) \\
&:= C(T)h^2 \left(\max_{t\in[0,T]} \|\partial_t u(t)\|^2_{H^2(\Omega)} + \int_0^T \|u(t)\|^2_{H^2(\Omega)} \,\mathrm{d}t + \int_0^T \|\partial_{tt} u(t)\|^2_{H^2(\Omega)} \,\mathrm{d}t \right)
\end{aligned}$$

mit einer Konstanten $C(T) > 0$, die nur von T abhängt. □

Bemerkung 17.7. Man beachte, dass hier wieder Quadrate der Fehlernormen abgeschätzt wurden. Für den Fehler in der $H^1(\Omega)$-Norm erhalten wir daher $\mathcal{O}(h)$-Konvergenz.

Für die Fehler in der L^2-Norm kann man – wie im parabolischen Fall – mit einem Aubin-Nitsche Resultat sogar $\mathcal{O}(h^2)$-Konvergenz bekommen kann, falls Ω konvex ist oder einen glatten Rand hat:

$$\max_{t\in[0,T]} \|\partial_t u(t) - \partial_t u_h(t)\|^2_{L^2(\Omega)} + \max_{t\in[0,T]} \|u(t) - u_h(t)\|^2_{L^2(\Omega)} \le C(u;T)h^4.$$

In beiden Fällen muss natürlich auch die exakte Lösung u hinreichend regulär sein. Zusammenfassend sieht man, dass sich die Fehlerabschätzungen für elliptische Probleme wieder sinngemäß auf die entsprechenden hyperbolischen Probleme übertragen lassen.

Zeitdiskretisierung mit Zweischrittverfahren

Zum Abschluss betrachten wir nun die Zeitdiskretisierung. Wie zuvor bezeichnen wir mit $t^{(n)} = n\tau$, $n \geq 0$ die Folge von diskreten Zeitpunkten und jetzt mit

$$\begin{aligned} \mathrm{d}_{\tau\tau} u_h^{(n)} &= \tfrac{1}{\tau^2}(u_h^{(n+1)} - 2u_h^{(n)} + u_h^{(n-1)}) \\ \mathrm{d}_{2\tau} u_h^{(n)} &= \tfrac{1}{2\tau}(u_h^{(n+1)} - u_h^{(n-1)}), \qquad u_h^{(n+1/2)} = \tfrac{1}{2}(u_h^{(n+1)} + u_h^{(n)}) \end{aligned} \tag{17.10}$$

den zentralen zweiten und ersten Differenzenquotienten sowie den Mittelwert. Zur Approximation unseres Problems betrachten wir folgendes Verfahren, das der in Beispiel 5.3 besprochenen Methode entspricht.

Problem 17.8. Seien $u_h^{(0)}, u_h^{(1)} \in V_h$ gegeben. Finde $u_h^{(n)} \in V_h$, $n \geq 1$, so dass

$$\int_\Omega \mathrm{d}_{\tau\tau} u_h^{(n)} v_h \, dx + \int_\Omega \nabla u_h^{(n)} \cdot \nabla v_h \, dx = \int_\Omega f(t^{(n)}) v_h \, dx \qquad \forall v_h \in V_h. \tag{17.11}$$

Bemerkung 17.9. Im $n$ten Schritt des Verfahrens wird der Wert $u_h^{(n+1)}$ mit Hilfe von bekannten Werten $u_h^{(n)}$, $u_h^{(n-1)}$ und $f(t^{(n)})$ bestimmt. Es handelt sich also um ein Zweischrittverfahren. Da wir für die Bestimmung von $u_h^{(n+1)}$ nur ein lineares Gleichungssystem mit der Massematrix (die wir aber durch eine skalierte Einheitsmatrix annähern könnten, vgl. Beispiel 16.18) lösen müssen, können wir von einem expliziten Verfahren sprechen.

Als nächstes fassen wir einige wichtige Eigenschaften des Schemas zusammen.

Satz 17.10. *Problem 17.8 besitzt eine eindeutige Lösung* $(u_h^{(n)})_{n=0,\dots,N}$. *Weiters gilt*

$$E_h^{(n+1/2)} = E_h^{(n-1/2)} + \tau \int_\Omega f(t^{(n)}) \, \mathrm{d}_{2\tau} u_h^{(n)} \, \mathrm{d}x, \tag{17.12}$$

wobei $E_h^{(n+1/2)} := \frac{1}{2\tau^2} \|u_h^{(n+1)} - u_h^{(n)}\|^2_{L^2(\Omega)} + \frac{1}{2} \int_\Omega \nabla u_h^{(n+1)} \cdot \nabla u_h^{(n)} \, \mathrm{d}x$ *die diskrete Energie bezeichnet.*

Beweis. Die Implementierung der Methode führt auf eine Iterationsvorschrift der Form $\frac{1}{\tau^2} M_h(\underline{u}_h^{(n+1)} - 2\underline{u}_h^{(n)} + \underline{u}_h^{(n-1)}) = -A_h \underline{u}_h^{(n)} + \underline{b}_h^{(n)}$ mit Massematrix M_h und Steifigkeitsmatrix A_h. Da die Massenmatrix M_h symmetrisch und positiv definit ist, kann $\underline{u}_h^{(n+1)}$ in jedem Schritt einfach durch Lösen eines linearen Gleichungssystems bestimmt werden. Daraus folgt Existenz und Eindeutigkeit der Lösung.

Zum Nachweis der Energieidentität benutzen wir die elementare algebraische Identität $(a-b)^2-(b-c)^2=(a-2b+c)(a-c)$. Aus dieser folgt sofort

$$\tfrac{1}{2\tau^2}\|u_h^{(n+1)}-u_h^{(n)}\|^2_{L^2(\Omega)}-\tfrac{1}{2\tau^2}\|u_h^{(n)}-u_h^{(n-1)}\|^2_{L^2(\Omega)}=\int_\Omega \mathrm{d}_{\tau\tau}u_h^{(n)}\tfrac{u_h^{(n+1)}-u_h^{(n-1)}}{2}\,\mathrm{d}x.$$

Für die Differenzen der zweiten Teile in der Energie gilt offensichtlich

$$\tfrac{1}{2}\int_\Omega \nabla u_h^{(n+1)}\cdot\nabla u_h^{(n)}\,\mathrm{d}x-\tfrac{1}{2}\int_\Omega \nabla u_h^{(n)}\cdot\nabla u_h^{(n-1)}\,\mathrm{d}x=\int_\Omega \nabla u_h^{(n)}\cdot\nabla\tfrac{u_h^{(n+1)}-u_h^{(n-1)}}{2}\,\mathrm{d}x.$$

Durch Addition der beiden Identitäten folgt nun unmittelbar

$$E_h^{(n+1/2)}-E_h^{(n-1/2)}=\int_\Omega \mathrm{d}_{\tau\tau}u_h^{(n)}v_h\,\mathrm{d}x+\int_\Omega \nabla u_h^{(n)}\cdot\nabla v_h\,\mathrm{d}x$$

mit Testfunktion $v_h=\frac{u_h^{(n+1)}-u_h^{(n-1)}}{2}=\tau\mathrm{d}_{2\tau}u_h^{(n)}$. Mit Einsetzen in das diskrete Variationsproblem (17.11) folgt nun sofort die Behauptung. □

Bemerkung 17.11. Aus der Definition sieht man relativ leicht, das $E_h^{(n+1/2)}$ eine Approximation der Energie zum Zeitpunkt $t^{(n+1/2)}$ darstellt. Für den Fall $f\equiv 0$ wissen wir bereits, dass die Lösung des Anfangsrandwertproblems eine Energieerhaltungseigenschaft erfüllt, durch Einsetzen von $f\equiv 0$ in (17.12) sehen wir, dass in diesem Fall auch die Lösung des Differenzenschemas die diskrete Energie erhält.

Wegen des Terms $\int_\Omega \nabla u_h^{(n+1)}\cdot\nabla u_h^{(n)}\,\mathrm{d}x$ können wir nicht auf den ersten Blick ausschließen, dass $E_h^{(n+1/2)}$ negativ werden kann. Wir wollen nun zeigen, dass das nicht der Fall ist. Für die folgenden Überlegungen nehmen wir an, dass eine inverse Ungleichung der Form

$$\|\nabla v_h\|_{L^2(\Omega)}\le c_{inv}h^{-2}\|v_h\|_{L^2(\Omega)}\qquad\forall v_h\in V_h \tag{17.13}$$

gilt. Wir wissen bereits, dass sich diese für Finite-Elemente Räume auf quasi-uniformen Netzen zeigen lässt. Weiters verlangen wir, dass die Zeitschrittweite hinreichend klein ist, d.h.

$$\tau\le c_{inv}^{-1}h. \tag{17.14}$$

Diese Bedingung wird nach Courant, Friedrichs und Levy in der Literatur häufig als eine CFL-Bedingung bezeichnet. Es gilt dann

$$\begin{aligned}\tfrac{1}{2}\int_\Omega \nabla u_h^{(n+1)}\cdot\nabla u_h^{(n)}\,\mathrm{d}x&=\tfrac{1}{8}\|\nabla(u_h^{(n+1)}+u_h^{(n)})\|^2_{L^2(\Omega)}-\tfrac{1}{8}\|\nabla(u_h^{(n+1)}-u_h^{(n)})\|^2_{L^2(\Omega)}\\&=\tfrac{1}{2}\|\nabla u_h^{(n+1/2)}\|^2_{L^2(\Omega)}-\tfrac{1}{8}\|\nabla(u_h^{(n+1)}-u_h^{(n)})\|^2_{L^2(\Omega)}\\&\ge\tfrac{1}{2}\|\nabla u_h^{(n+1/2)}\|^2_{L^2(\Omega)}-\tfrac{1}{4}(\tfrac{1}{2\tau^2}\|u_h^{(n+1)}-u_h^{(n)}\|^2_{L^2(\Omega)}),\end{aligned}$$

wobei wir im letzten Schritt (17.13) und (17.14) verwendet haben. Offensichtlich gilt aber auch $\frac{1}{2}\int_\Omega \nabla u_h^{(n+1)} \cdot \nabla u_h^{(n)}\,\mathrm{d}x \le \frac{1}{2}\|\nabla u_h^{(n+1/2)}\|^2_{L^2(\Omega)}$. Mit der Definition von $E_h^{(n+1/2)}$ folgt daraus $\frac{3}{4}\tilde{E}_h^{(n+1/2)} \le E_h^{(n+1/2)} \le \tilde{E}_h^{(n+1/2)}$, wobei

$$\tilde{E}_h^{(n+1/2)} := \tfrac{1}{2\tau^2}\|u_h^{(n+1)} - u_h^{(n)}\|^2_{L^2(\Omega)} + \tfrac{1}{2}\|\nabla u_h^{(n+1/2)}\|^2_{L^2(\Omega)} \ge 0. \tag{17.15}$$

Daher ist unter Annahme von (17.14) die diskrete Energie also tatsächlich nichtnegativ. Mit Hilfe dieser Vorüberlegungen können wir jetzt folgende Aussage zeigen.

Satz 17.12. *Wenn die Bedingungen* (17.13) *und* (17.14) *erfüllt sind, dann gilt für die numerische Lösung* $(u_h^{(n)})_n$ *von Problem 17.8 die Abschätzung*

$$E_h^{(n+1/2)} \le e^{\tau n}\left(E_h^{(1/2)} + \sum_{k=1}^{n} \tau\|f(t^{(k)})\|^2_{L^2(\Omega)}\right). \tag{17.16}$$

Beweis. Mit (17.12), (17.15) und $\tilde{E}_h^{(n+1/2)} \le \frac{4}{3}E_h^{(n+1/2)}$ folgt sofort

$$\begin{aligned}
E_h^{(n+1/2)} - E_h^{(n-1/2)} &\le \tau\|f(t^{(n)})\|_{L^2}\|\tfrac{1}{2\tau}(u_h^{(n+1)} - u_h^{(n-1)})\|_{L^2} \\
&\le \tau\left(\beta\|f(t^{(n)})\|^2 + \tfrac{1}{8\tau^2\beta}\|u_h^{(n+1)} - u_h^{(n)}\|^2_{L^2} + \tfrac{1}{8\tau^2\beta}\|u_h^{(n)} - u_h^{(n-1)}\|^2_{L^2}\right) \\
&\le \tau\beta\|f(t^{(n)})\|^2 + \tfrac{\tau}{4\beta}\tilde{E}_h^{(n+1/2)} + \tfrac{\tau}{4\beta}\tilde{E}_h^{(n-1/2)} \\
&\le \tau\beta\|f(t^{(n)})\|^2 + \tfrac{\tau}{3\beta}E_h^{(n+1/2)} + \tfrac{\tau}{3\beta}E_h^{(n-1/2)}
\end{aligned}$$

für alle $\beta > 0$. Durch Umstellen folgt für $\beta = 1$ sofort die Ungleichung

$$(1 - \tfrac{\tau}{3})E_h^{(n+1/2)} \le (1 + \tfrac{\tau}{3})E_h^{(n-1/2)} + \tau\|f(t^{(n)})\|^2_{L^2}.$$

Mit elementarer Rechnung lässt sich überprüfen, dass $1 + \frac{\tau}{3} \le e^{\tau/3}$ und $1 - \frac{\tau}{3} \ge e^{-2\tau/3}$ für alle $\tau \le 1$ gilt. Hiermit folgt

$$E_h^{(n+1/2)} \le e^{\tau}(E_h^{(n-1/2)} + \tau\|f(t^{(n)})\|^2_{L^2}).$$

Für den Fall $\tau > 1$ lässt sich dasselbe Resultat mit $\beta = \tau$ zeigen. Die Abschätzung aus dem Lemma folgt dann mit vollständiger Induktion. □

Fehlerabschätzung für das volldiskretisierte System

Wir betrachten wieder eine Zerlegung des Fehlers

$$u(t^{(n)}) - u_h^{(n)} = \underbrace{(u(t^{(n)}) - \tilde{u}_h^{(n)})}_{=:\rho^{(n)}} + \underbrace{(\tilde{u}_h^{(n)} - u_h^{(n)})}_{=:d_h^{(n)}} \tag{17.17}$$

in einen Approximationsfehler und einen diskreten Fehler. Wie zuvor wählen wir $\tilde{u}_h^{(n)} = \tilde{u}_h(t^{(n)}) = \tilde{\Pi}_h u(t^{(n)})$ über die elliptische Projektion. Den entsprechenden Fehler können wir mit

$$\|\rho_h^{(n)}\|_{L^2(\Omega)}^2 = \|u(t^{(n)}) - \tilde{u}_h^{(n)}\|_{L^2(\Omega)}^2 \leq Ch^{2k}\|u(t^{(n)})\|_{H^2(\Omega)}^2 \tag{17.18}$$

abschätzen, wobei die Konstante C von h, τ, T und u unabhängig ist. Dieses Resultat erhalten wir für $k = 1$ mit $\|\cdot\|_{L^2} \leq \|\cdot\|_{H^1}$. Mit einem Aubin-Nitsche-Argument kann man hier auch $k = 2$ erhalten, wenn das Gebiet Ω konvex ist oder einen glatten Rand hat. Durch Einsetzen in das diskrete Problem sieht man, dass

$$\begin{aligned}\int_\Omega \mathrm{d}_{\tau\tau} d_h^{(n)} v_h \,\mathrm{d}x + \int_\Omega \nabla d_h^{(n)} \cdot \nabla v_h \,\mathrm{d}x \\ = \int_\Omega (\mathrm{d}_{\tau\tau}\tilde{u}_h^{(n)} - \partial_{tt} u(t^{(n)})) v_h \,\mathrm{d}x + \int_\Omega \nabla(\tilde{u}_h^{(n)} - u(t^{(n)})) \cdot \nabla v_h \,\mathrm{d}x.\end{aligned}$$

Dabei wurde (17.11) für $u_h^{(n)}$ und (17.4) für $u(t^{(n)})$ mit $v = v_h$ verwendet. Nach Definition der elliptische Projektion (16.13) fällt der zweite Term weg. Der diskrete Fehler $d_h^{(n)}$ ist also Lösung eines diskreten Problems

$$\int_\Omega \mathrm{d}_{\tau\tau} d_h^{(n)} v_h \,\mathrm{d}x + \int_\Omega \nabla d_h^{(n)} \cdot \nabla v_h \,\mathrm{d}x = \int_\Omega \big(\tilde{f}_1(t^{(n)}) + \tilde{f}_2(t^{(n)})\big) v_h \,\mathrm{d}x$$

mit den speziellen rechten Seiten

$$\tilde{f}_1(t^{(n)}) := \mathrm{d}_{\tau\tau}\tilde{u}_h^{(n)} - \mathrm{d}_{\tau\tau} u(t^{(n)}), \qquad \tilde{f}_2(t^{(n)}) := \mathrm{d}_{\tau\tau} u(t^{(n)}) - \partial_{tt} u(t^{(n)}).$$

Der erste Teil lässt sich mit $\|\cdot\|_{L^2(\Omega)} \leq \|\cdot\|_{H^1(\Omega)}$ und der Fehlerabschätzungen für die elliptische Projektion beschränken durch

$$\tau\|\tilde{f}_1^{(n)}\|_{L^2(\Omega)}^2 \leq C\tau h^{2k}\|\mathrm{d}_{\tau\tau} u(s)\|_{H^2(\Omega)}^2 \leq Ch^{2k}\int_{t^{(n-1)}}^{t^{(n+1)}} \|\partial_{tt} u(s)\|_{H^2(\Omega)}^2 \,\mathrm{d}s,$$

wobei k wie oben ist. Mit Taylorentwicklung erhält man für den zweiten Teil

$$\tau\|\tilde{f}_2^{(n)}\|_{L^2(\Omega)}^2 \leq C\tau^{2q}\int_{t^{(n-1)}}^{t^{(n+1)}} \|\partial_t^{q+2} u(s)\|_{L^2(\Omega)}^2 \,\mathrm{d}s$$

für $q = 1$ und $q = 2$. Wir betrachten zuerst den Fall $u^{(0)} = v^{(0)} = 0$. Dann gilt auch $u_h^{(0)} = 0$, $v_h^{(0)} = 0$ und sinnvollerweise auch $u_h^{(1)} = 0$. Daraus folgt $d_h^{(0)} = 0$ und $d_h^{(1)} = \tilde{\Pi}_h u(t^{(1)})$. Mit Satz 17.12, (17.15) und den genannten Anfangsbedingungen erhalten wir

$$\frac{3}{8\tau^2}\|d_h^{(n+1)} - d_h^{(n)}\|_{L^2(\Omega)}^2 \leq e^{\tau n}\frac{1}{2\tau^2}\|\tilde{\Pi}_h u(t^{(1)})\|_{L^2(\Omega)}^2 + e^{\tau n}\sum_{k=1}^{n} \tau\|\tilde{f}(t^{(k)})\|_{L^2(\Omega)}^2$$

für alle n. Mit einem einfachen Teleskopargument, $n\tau \le T$ und $\sum_{j=0}^{n-1} e^{\tau j} \le (e^T - 1)/(e^\tau - 1)$, $\tau \le e^\tau - 1$, den Abschätzungen für $\tilde{f}_1$ und $\tilde{f}_2$ und $\|\tilde{\Pi}_h u(t^{(1)})\|_{L^2(\Omega)} \le \|\tilde{\Pi}_h u(t^{(1)})\|_{H^1(\Omega)} \le \|u(t^{(1)})\|_{H^1(\Omega)}$, erhalten wir

$$\begin{aligned}
\|d_h^{(n)}\|_{L^2(\Omega)}^2 &\le n \sum_{j=0}^{n-1} \|d_h^{(j+1)} - d_h^{(j)}\|_{L^2(\Omega)}^2 \\
&\le 2n\tau^2 \left(\sum_{j=0}^{n-1} e^{\tau j} \right) \left(\frac{1}{2\tau^2} \|\tilde{\Pi}_h u(t^{(1)})\|_{L^2(\Omega)}^2 + \sum_{j=1}^{n} \tau \|\tilde{f}(t^{(j)})\|_{L^2(\Omega)}^2 \right) \\
&\le CT(e^T - 1)\Big(\frac{1}{2\tau^2} \|u(t^{(1)})\|_{H^1(\Omega)}^2 + h^{2k} \int_0^T \|\partial_{tt} u(s)\|_{H^2(\Omega)}^2 \,\mathrm{d}s \\
&\quad + \tau^{2q} \int_0^T \|\partial_t^{q+2} u(s)\|_{L^2(\Omega)}^2 \,\mathrm{d}s \Big).
\end{aligned} \tag{17.19}$$

Mit einer Taylor-Abschätzung erhalten wir

$$\frac{1}{2\tau^2} \|u(t^{(1)})\|_{H^1(\Omega)}^2 \le C\tau^2 \max_{\zeta \in [0,\tau]} \|\partial_{tt} u(\zeta)\|_{H^1(\Omega)}^2.$$

Damit haben wir $\|d_h^{(n)}\|_{L^2(\Omega)}^2 \le \tilde{C}(u;T)(h^{2k} + \tau^{2q})$, wobei $\tilde{C}(u;T)$ nur von u und T, nicht jedoch von h und τ abhängt.

Satz 17.13. *Sei u hinreichend reguläre Lösung des Problems* (17.1) – (17.3) *mit $u_0 = v_0 = 0$ und $f(0) = 0$. Weiters sei $(u_h^{(n)})_{n=0}^N$ die entsprechende diskrete Lösung von Problem 17.8 mit $u_h^{(0)} = u_h^{(1)} = 0$ und $V_h = P_1(T_h) \cap H_0^1(\Omega)$. Schließlich gelte* (17.13) *und* (17.14). *Dann folgt*

$$\max_{n \in \{0,\dots,N\}} \|u(t^{(n)}) - u_h^{(n)}\|_{L^2(\Omega)} \le C(u;T)(h^k + \tau^2).$$

mit $k = 1$. Wenn das Gebiet Ω konvex ist oder einen glatten Rand hat, gilt dasselbe auch für $k = 2$.

Der Beweis der Aussage folgt sofort aus (17.17), (17.18) und (17.19).

Bemerkung 17.14. Für allgemeine Anfangswerte und rechte Seite muss man noch die geeignete Wahl der diskreten Startwerte $\underline{u}_h^{(0)}$ und $\underline{u}_h^{(1)}$ diskutieren. In (17.9) hatten wir bereits verwendet, dass $\underline{u}_h^{(0)}$ und $\underline{v}_h^{(0)}$ mit der elliptischen Projektion $\tilde{\Pi}_h$ definiert werden kann. Auf den ersten Blick bietet sich für die Wahl von $u_h^{(1)}$ eine Approximation auf Basis der Taylor-Approximation erster Ordnung an:

$$\underline{u}_h^{(1)} = \underline{u}_h^{(0)} + \tau \underline{v}_h^{(0)},$$

die jedoch zu einem Fehler der Art $\mathcal{O}(h^k + \tau)$ führt, somit der Schwachstelle des gesamten Verfahrens wäre. Stattdessen ist es sinnvoll, eine Taylor-Approximation zweiter Ordnung anzuwenden, wobei $\partial_{tt} u_h^{(0)}$ auf Basis der Differentialgleichung approximiert wird; damit erhalten wir:

$$\underline{u}_h^{(1)} = \underline{u}_h^{(0)} + \tau \underline{v}_h^{(0)} + \tfrac{\tau^2}{2}(\underline{b}^{(0)} - A_h \underline{u}_h^{(0)}),$$

womit wir für das gesamte Verfahren eine Konvergenz der Art $\mathcal{O}(h^k + \tau^2)$ erhalten.

Zusammenfassung

Wie obige Überlegungen zeigen, lassen sich die Konvergenzaussagen für Finite-Elemente-Verfahren aus dem elliptischen Fall und für einfachen Zeitschrittverfahren zur Diskretisierung gewöhnlicher Differentialgleichungen mit überschaubarem Aufwand auf den Fall hyperbolischer partieller Differentialgleichungen übertragen.

Numerische Tests

Zuletzt wollen wir noch die Resultate in einem numerischen Test veranschaulichen. Dabei beschränken wir uns auf das den Fall $f \equiv 0$. Seien Anfangsbedingungen u_0, u_1 gegeben. Wir wollen dann folgendes Problem lösen:

$$\begin{aligned} \partial_{tt} u(x,t) - \kappa^2 \partial_{xx} u(x,t) &= 0, && x \in \Omega = (0,1), t \in (0,T] \\ u(0,t) = u(1,t) &= 0, && t \in (0,T], \\ u(x,0) &= u_0(x), && x \in \Omega, \\ \partial_t u(x,0) &= v_0(x), && x \in \Omega. \end{aligned}$$

Wenn wir annehmen, dass sich u_0 und v_0 als Fourier-Reihen

$$u_0(x) = \sum_{i=1}^{\infty} a_i \sin(i\pi x), \qquad v_0(x) = \sum_{i=1}^{\infty} b_i \sin(i\pi x),$$

darstellen lassen, dann lässt sich leicht nachprüfen, dass (unter Annahme von Konvergenz) durch

$$u(x,t) = \sum_{i=1}^{\infty} (a_i \sin(i\pi x)\cos(i\kappa\pi t) + \tfrac{b_i}{i} \sin(i\pi x)\sin(i\kappa\pi t))$$

eine Lösung gegeben ist. Die Theorie garantiert die Eindeutigkeit.

Als erstes testen wir die in (17.10) – (17.11) vorgeschlagene Methode. Für beide Beispiele wählen wir $T = 1$, $\kappa = 5$, $u_0(x) = \sin(\pi x)$ und $v_0(x) = 0$ als Daten.

Beispiel 17.15 (Explizites Verfahren)**.** Wir diskretisieren das Ortsproblem mit dem Courant-Element mit einer äquidistanten Schrittweite $h = n^{-1}$ mit $n \in \mathbb{N}$. Wir nähern nun $\partial_{tt}u(x,t)$ (wie in (17.10) vorgegeben) mit $\frac{1}{\tau^2}(-u(x,t+\tau) + u(x,t) - u(x,t-\tau))$ für eine Zeitschrittweite $\tau = m^{-1}$ mit $m \in \mathbb{N}$ an. Daher erfüllen die Näherungen $\underline{u}_h^{(j)}$ an die Lösung zum Zeitpunkt $t^{(j)} = j\tau$:

$$\tfrac{1}{\tau^2} M_h(\underline{u}_h^{(j+1)} - 2\underline{u}_h^{(j)} + \underline{u}_h^{(j-1)}) + \kappa^2 K_h \underline{u}_h^{(j)} = 0 \tag{17.20}$$

wobei $M_h, K_h \in \mathbb{R}^{(n-1)\times(n-1)}$ Masse- und Steifigkeitsmatrix sind:

$$M_h = \frac{h}{6}\begin{pmatrix} 4 & 1 & & \\ 1 & 4 & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & 4 \end{pmatrix} \quad \text{und} \quad K_h = \frac{1}{h}\begin{pmatrix} 2 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix}.$$

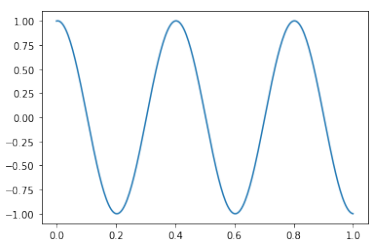
Zur Lösung des ARWP
(i) setzen wir zuerst $\underline{u}_h^{(0)}$ und $\underline{v}_h^{(0)}$ auf diskrete Werte der AB u_0 bzw. v_0 (hier durch die punktweise Interpolation; Alternative wären etwa $\tilde{\Pi}_h$ oder L^2-Projektion),
(ii) setzen $\underline{u}_h^{(1)} = \underline{u}_h^{(0)} + \tau\underline{v}_h^{(0)} - \frac{\tau^2}{2}\kappa^2 K_h \underline{u}_h^{(0)}$ auf Basis der Taylor-Approximation 2. Ordnung und
(iii) bestimmen die nächsten Näherungslösungen jeweils durch Auflösen des Systems (17.20) nach $\underline{u}_h^{(j+1)}$ für $j = 1, 2, \ldots$.

	$\tau = 2^{-8}$	$\tau = 2^{-9}$	$\tau = 2^{-10}$	$\tau = 2^{-11}$	$\tau = 2^{-12}$	$\tau = 2^{-13}$	$\tau = 2^{-14}$
$h = 2^{-4}$	$2.48\,10^{-2}$ (4 / –)	$2.32\,10^{-2}$ (4 / 1)	$2.28\,10^{-2}$ (4 / 1)	$2.28\,10^{-2}$ (4 / 1)	$2.28\,10^{-2}$ (4 / 1)	$2.28\,10^{-2}$ (4 / 1)	$2.28\,10^{-2}$ (4 / 1)
$h = 2^{-5}$	div (– / –)	$6.22\,10^{-3}$ (4 / –)	$5.82\,10^{-3}$ (4 / 1)	$5.72\,10^{-3}$ (4 / 1)	$5.70\,10^{-3}$ (4 / 1)	$5.69\,10^{-3}$ (4 / 1)	$5.69\,10^{-3}$ (4 / 1)
$h = 2^{-6}$	div (– / –)	div (– / –)	$1.56\,10^{-3}$ (4 / –)	$1.46\,10^{-3}$ (4 / 1)	$1.43\,10^{-3}$ (4 / 1)	$1.42\,10^{-3}$ (4 / 1)	$1.42\,10^{-3}$ (4 / 1)
$h = 2^{-7}$	div (– / –)	div (– / –)	div (– / –)	$3.90\,10^{-4}$ (4 / –)	$3.64\,10^{-4}$ (4 / 1)	$3.58\,10^{-4}$ (4 / 1)	$3.56\,10^{-4}$ (4 / 1)
$h = 2^{-8}$	div (– / –)	div (– / –)	div (– / –)	div (– / –)	$9.76\,10^{-5}$ (4 / –)	$9.11\,10^{-5}$ (4 / 1)	$8.95\,10^{-5}$ (4 / 1)
$h = 2^{-9}$	div (– / –)	div (– / –)	div (– / –)	div (– / –)	div (– / –)	$2.44\,10^{-5}$ (4 / –)	$2.28\,10^{-5}$ (4 / 1)

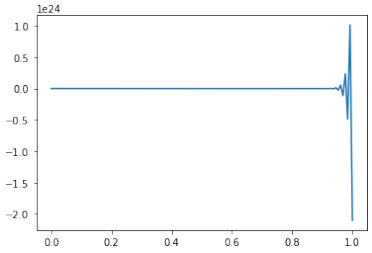
Tabelle 17.1.: Konvergenz explizites Verfahren.

Der *absolute* Fehler in der diskreten Maximumsnorm (Auswertung des Fehlers an allen Gitterpunkten $(ih, j\tau)$) ist in Tabelle 17.1 gegeben.[1] Wir zeigen nun

[1] Wir wählen erneut die diskrete Maximumsnorm wählen wir, weil sie sehr einfach zu bestimmen ist. Da die exakte Lösung nun nicht mehr gegen 0 konvergiert, bestimmen wir nun – wie auch sonst immer – den absoluten Fehler.



(a) $h = 2^{-4},\ \tau = 2^{-8}$



(b) $h = 2^{-4},\ \tau = 2^{-7}$

Abbildung 17.1.: Typische Verläufe von $u_h(1/2, t)$ für das explizite Verfahren.

auch noch $u_h(1/2, t)$. Zuerst stellen wir fest, dass die exakte Lösung $u(1/2, t) = \cos(5\pi t)$ sein sollte, also eine Schwingung mit gleichbleibender Ampiltude (das entspricht bereits unserem theoretischen Resultat, dass die Lösung für $f \equiv 0$ die Energieerhaltungseigenschaft erfüllt). Setzt man $f \equiv 0$ in (17.12) ein, erkennt man, dass auch die diskrete Energie eine Erhaltungsgröße ist. In Abbildung 17.2 (a) sehen wir einen Fall einer stablen Lösung, wo diese Eigenschaft erzielt wird. In Abbildung 17.2 (b) sehen wir einen Fall, wo die Stabilitätsbedingung (17.14), also $\tau \lesssim h$ nicht erfüllt ist und die Lösung aufschwingt (also aus dem Nichts Energie gewinnt). Solche Fälle sind in Tabelle 17.1 mit *div* gekennzeichnet. Wir erkennen, dass die Grenze zwischen Konvergenz und Divergenz auf einer Diagonale $\tau \sim h$ verläuft.

Wie im letzten Kapitel, betrachten wir auch wieder neben den Fehlern die Konvergenzraten. Die links angegebene Zahl in blauer Farbe gibt den Faktor an, um den sich der Fehler bei Halbierung von h ändert; der Wert 4 entspricht hier einer Konvergenz wie $\mathcal{O}(h^2)$. Die rechts angegebene Zahl in violetter Farbe gibt den Faktor an, um den sich der Fehler bei Halbierung von τ ändert; hier entspräche der Wert 4 einer Konvergenz wie $\mathcal{O}(\tau^2)$. Dieser Fall tritt hier aber nicht ein, da im Konvergenzbereich der Fehlerterm von h dominiert. Insgesamt ist das Verhalten mit der theoretischen Aussage $\mathcal{O}(h^2 + \tau^2)$ vereinbar.

Beispiel 17.16 (Implizites Verfahren)**.** Nach Ortsdiskretisierung wie im letzten Beispiel, nähern wir nun $\partial_{tt} u(x, t)$ mit $\frac{1}{\tau^2}(-u(x, t) + u(x, t - \tau) - u(x, t - 2\tau))$ an. Daher erfüllen die Näherungen $\underline{u}_h^{(j)}$ an die Lösung zum Zeitpunkt $t^{(j)} = j\tau$:

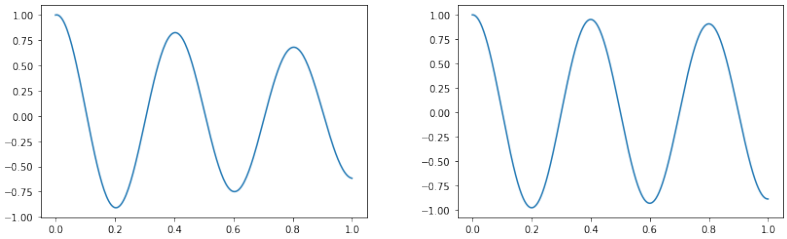
$$\tfrac{1}{\tau^2} M_h(\underline{u}_h^{(j)} - 2\underline{u}_h^{(j-1)} + \underline{u}_h^{(j-2)}) + \kappa^2 K_h \underline{u}_h^{(j)} = 0. \tag{17.21}$$

Zur Lösung des ARWP bestimmen wir wieder $\underline{u}_h^{(0)}$ und $\underline{u}_h^{(1)}$ wie im letzten Beispiel und bestimmen dann $\underline{u}_h^{(j)}$ für $j = 2, 3 \ldots$ jeweils durch Lösen des Systems (17.21).

Der *absolute* Fehler in der diskreten Maximumsnorm ist in Tabelle 17.2 gege-

	$\tau = 2^{-11}$	$\tau = 2^{-12}$	$\tau = 2^{-13}$	$\tau = 2^{-14}$	$\tau = 2^{-15}$	$\tau = 2^{-16}$	$\tau = 2^{-17}$
$h = 2^{-4}$	$5.88\,10^{-2}$ (2 / 2)	$3.15\,10^{-2}$ (3 / 2)	$2.45\,10^{-2}$ (4 / 1)	$2.27\,10^{-2}$ (4 / 1)	$2.25\,10^{-2}$ (4 / 1)	$2.26\,10^{-2}$ (4 / 1)	$2.26\,10^{-2}$ (4 / 1)
$h = 2^{-5}$	$5.85\,10^{-2}$ (1 / 2)	$2.97\,10^{-2}$ (1 / 2)	$1.50\,10^{-2}$ (2 / 2)	$7.96\,10^{-3}$ (3 / 2)	$6.15\,10^{-3}$ (4 / 1)	$5.70\,10^{-3}$ (4 / 1)	$5.63\,10^{-3}$ (4 / 1)
$h = 2^{-6}$	$5.85\,10^{-2}$ (1 / 2)	$2.97\,10^{-2}$ (1 / 2)	$1.49\,10^{-2}$ (1 / 2)	$7.50\,10^{-3}$ (1 / 2)	$3.76\,10^{-3}$ (2 / 2)	$2.00\,10^{-3}$ (3 / 2)	$1.54\,10^{-3}$ (4 / 1)
$h = 2^{-7}$	$5.85\,10^{-2}$ (1 / 2)	$2.97\,10^{-2}$ (1 / 2)	$1.49\,10^{-2}$ (1 / 2)	$7.50\,10^{-3}$ (1 / 2)	$3.76\,10^{-3}$ (1 / 2)	$1.88\,10^{-3}$ (1 / 2)	$9.41\,10^{-4}$ (2 / 2)
$h = 2^{-8}$	$5.85\,10^{-2}$ (1 / 2)	$2.97\,10^{-2}$ (1 / 2)	$1.49\,10^{-2}$ (1 / 2)	$7.50\,10^{-3}$ (1 / 2)	$3.76\,10^{-3}$ (1 / 2)	$1.88\,10^{-3}$ (1 / 2)	$9.41\,10^{-4}$ (1 / 2)
$h = 2^{-9}$	$5.85\,10^{-2}$ (1 / 2)	$2.97\,10^{-2}$ (1 / 2)	$1.49\,10^{-2}$ (1 / 2)	$7.50\,10^{-3}$ (1 / 2)	$3.76\,10^{-3}$ (1 / 2)	$1.88\,10^{-3}$ (1 / 2)	$9.41\,10^{-4}$ (1 / 2)

Tabelle 17.2.: Konvergenz implizites Verfahren.



(a) $h = 2^{-4},\ \tau = 2^{-8}$ (b) $h = 2^{-4},\ \tau = 2^{-10}$

Abbildung 17.2.: Typische Verläufe von $u_h(1/2, t)$ für das implizite Verfahren.

ben. Wir zeigen nun auch noch $u_h(1/2, t)$. In Abbildung 17.4 (a) sehen wir den stabilen Fall, wo jedoch die Schwingung abschwingt (also Energie verliert). In Abbildung 17.4 (b) erzielen wir mit kleinerem τ eine Lösung, wo dieses Phänomen deutlich abgeschwächt ist.

Wir können uns wieder die Angaben zu den Raten ansehen und stellen fest, dass wir nun ein Konvergenverhalten wie $\mathcal{O}(h^2 + \tau)$ erhalten, was man von der (von uns nicht behandelten Theorie) auch erwarten würde.

Wir sehen, dass das explizite Verfahren aus dem ersten Beispiel (weleches dem theoretisch behandelten Verfahren (17.10) – (17.11) entspricht) eine Konvergenz nur erzielt werden kann, wenn $\tau \lesssim h$. Das implizite Verfahren ist immer stabil, jedoch ist es tendentiell „zu stabil", da hier die Näherungslösungen abschwingen. Vergleicht man das Konvergenzverhalten – $\mathcal{O}(h^2 + \tau^2)$ vs. $\mathcal{O}(h^2 + \tau)$ – so sieht man, dass beim impliziten Verfahren viel kleinere Zeitschritte zu wählen sind, als beim expliziten Verfahren, vgl. Abbildung 17.4 (a) mit Abbildung 17.2 (a), wo in beiden Fällen $h = 2^{-4}$ und $\tau = 2^{-8}$ gewählt sind.

Prüfungfragen (WS 2023/24)

I. Anfangswertprobleme In der Prüfung wird die Theorie des Teils I in Verbindung mit einem konkreten AWP durchgegangen. Beispiele wären etwa $u'(t) = \sqrt{u(t)} + f(t)$, $u'(t) = a\,u(t) + f(t)$ mit $a > 0$ oder $a < 0$ oder $u''(t) = c(t)u(t) + u(t)$, etc.

1. **Theorie zu AWP:** Existenz und Eindeutigkeit von Lösungen; Stabilitätssatz und Lemma von Gronwall mit Beweis.
2. **Einschrittverfahren:** Formulierung und Eigenschaften: Durchführbarkeit, Konsistenz, Stabilität und Konvergenz; Konvergenzsatz mit Beweis; wichtigste Runge-Kutta Verfahren, Butcher-Tableau, maximal erreichbare Konvergenzraten.
3. **Stärkere Stabilitätskriterien:** Stabilitätsgebiet; Bedeutung: 0-, A- und L-Stabilität; Beispiele für entsprechende Verfahren; Bezug zum Modellproblem!

II. Randwertprobleme (FDM) In der Prüfung wird die Theorie des Teils II (FDM oder FEM) in Verbindung mit einem konkreten RWP zweiter Ordnung mit verschiedenen Randbedingungen durchgegangen.

1. **Theorie zu RWP in starker Form:** Existenz, Eindeutigkeit und Stabilität der Lösung; Greenfunktion; Maximumsprinzipien mit Beweisen; Erweiterbarkeit auf 2D.
2. **Diskretisierung mit FDM:** Differenzenschemata; diskretes Maximumsprinzipien mit Beweisen; Existenz, Eindeutigkeit und Stabilität der Lösung; Diskretisierungsfehler basierend auf Konsistenz und Stabilität; Erw. 2D.
3. **Singulär gestörte Probleme:** Probleme mit klassischem Ansatz; Upwind-Verfahren.
4. **Variationsformulierung:** Herleitung, Wohldefiniertheit des RWP (Spursätze); Existenz und Eindeutigkeit schwacher Lösungen; Warum benötigen wir Sobolevräume und schwache Ableitungen?; Regularität von schwachen Lösungen ($d = 1, 2$).
5. **Diskretisierung mit FEM:** Galerkin-Approximation; Galerkin-Isomorphismus, Matrix- und Vektorassemblierung; Satz von Céa mit Beweis; H^1-Fehlerabschätzung mit Beweis; Idee von Aubin-Nitsche; Eigenschaften der Steifigkeitsmatrix.

III. Anfangs-Randwertprobleme

1. **Modellprobleme:** Wärmeleitungs- und Wellengleichung, Variationsformulierung, Enerigeabschätzung: Herleitung und Bedeutung.
2. **Semidiskretisierung im Ort:** Formulierung, Fehlersplitting, diskrete Energieabschätzung, Fehlerabschätzung.
3. **Volldiskretisierung / Diskretisierung in der Zeit:** Formulierung, Klassifizierung der zugrundeliegenden Verfahren (implizit, explizit); diskrete Energieabschätzung, Fehlersplitting und Fehleranalyse; Energieerhaltung bei der Wellengleichung – Bedeutung?; welche Verfahren funktionieren?

Beantwortung mit Papier und Bleistift. Prüfungsdauer: ca 30 min. Anmeldung und Terminvergabe per e-Mail; siehe `https://numa.jku.at/team/takacs`

Weiterführende LVAs im Sommersemester 2024

- **VL/UE Numerical methods for elliptic equations** (A. Schafelner)
  FEM für elliptische RWP in 2D und 3D; inexakte Realisierung der FEM (zB mit numerischer Integration); a-posteriori Fehlerschätzung und adaptive Netzverfeinerung; discontinuous Galerkin-Verfahren; nichtlineare Probleme.
  VL ist Pflicht für Master Industrial Mathematics,
  UE gehört zu den Core Execises im Master Industrial Mathematics
- **Special Topics Numerical Analysis: Differential algebraic equations** (H. Egger)
- **Projektseminar Numerical Analysis (Thema: Electrical machine simulation)**
  mit Möglichkeit, auch Themen für Bachelorarbeiten (und dann auch Master- und Doktorarbeiten) zu erhalten.

Weiterführende LVAs im Wintersemester 2024/25

- **VL/UE Numerical methods in continuum mechanics**
  Gemischte Variationsprobleme; entsprechende Finite-Elemente Methoden; Anwendungen in der Kontinuumsmechanik (zB linearisierte Elastizität). Aufbauend auf Numerical methods for elliptic equations.
  VL ist Pflicht für Master Industrial Mathematics,
  UE gehört zu den Core Execises im Master Industrial Mathematics.

Weitere Informationen zu den LVAs und möglichen Bachelor- und Masterarbeiten sind über unsere Website verfügbar: `https://numa.jku.at/`