Skriptum zur Vorlesung

Numerik partieller Differentialgleichungen

Wintersemester 2006/07

Martin Burger

Institut für Numerische und Angewandte Mathematik

martin.burger@uni-muenster.de

http://www.math.uni-muenster.de/u/burger/

Inhaltsverzeichnis

Kapitel 1

Einleitung

Partielle Differentialgleichungen (*partial differential equations - PDEs*) gehören zu den am häufigsten auftretenden mathematischen Modellen realer Prozesse. Verschiedenste Prozesse wie Wärmeleitung, Ausbreitung von Wasser- oder Schallwellen, Strömungen, Dynamik von biologischen Populationen werden heute mit PDEs modelliert, und immer neue Anwendung wie die Preisbestimmung von Finanzprodukten oder Glättung von Bildern und Computergraphiken kommen dazu.

In den wenigsten Fällen ist die analytische Lösung dieser Gleichungen möglich, und eine numerische Lösung wird nötig. Dazu ersetzt man die Gleichungen durch Gleichungssysteme in $\mathbb{R}^n$ (*Diskretisierung*), mit möglichst grossem n um die (unendlichdimensionale) Differentialgleichungen sinnvoll approximieren können. Nach der Diskretisierung verbleibt noch die Aufgabe, das endlichdimensionale Problem numerisch zu lösen, was meist eine weitere Herausforderung wegen der Grösse der Gleichungssysteme darstellt. In dieser Vorlesung werden wir uns sowohl mit verschiedenen Diskretisierungsverfahren als auch mit der effizienten Lösung der diskretisierten Probleme befassen. Wie auch bei der Theorie der partiellen Differentialgleichungen ist meist eine Unterscheidung nach Typ notwendig, da sich die unterschiedlichen Eigenschaften elliptischer, parabolischer, und hyperbolischer Gleichungen auch in der Numerik niederschlagen. Im folgenden werden wir für drei einfache Beispiele die Grundprobleme darstellen.

1.1 Beispiele der Numerik partieller Differentialgleichungen

1.1.1 Elliptische Probleme: Poisson Gleichung

Wir beginnen mit der einfachsten Form einer elliptischen Differentialgleichung, nämlich einem Randwertproblem für die eindimensionale Poisson-Gleichung.

$$-\frac{\partial^2 u}{\partial x^2}(x) = f(x), \qquad x \in (0,1), u(0) = 0, u(1) = 0. \tag{1.1}$$

Streng genommen ist (1.1) nicht einmal eine partielle, sondern nur eine gewöhnliche Differentialgleichung, aber dennoch (oder gerade deswegen) ist dieses Problem gut geeignet um die Grundzüge der Numerik elliptischer Randwertprobleme darzustellen.

Der erste Schritt in den meisten Diskretisierungsverfahren (und wir werden hier nur solche behandeln) ist die Auswahl eines geeigneten Gitters auf dem gegebenen Gebiet, d.h. auf dem Intervall $(0,1)$ im Fall von (1.1). Die einfachste Wahl ist ein reguläres Gitter mit den Punkten

$x_j = j/(n+1)$, $j = 0, \ldots, n+1$. Als Gitterfeinheit h bezeichnen wir den maximalen Abstand zwischen benachbarten Punkten, d.h. $h = \frac{1}{n+1}$.

Die erste Diskretisierungsart, die wir diskutieren werden, sind finite Differenzen, d.h., wir versuchen u durch eine Funktion u^h zu approximieren, die wir in den Gitterpunkten berechnen, d.h, wir suchen einen Vektor $(u_j^h)_{j=0,\ldots,n+1}$ wobei $u_j^h = u^h(x_j)$. Wir sehen sofort, dass wir durch die Randbedingungen zwei Werte sofort berechnen können, nämlich $u_0^h = u^h(0) = 0$ und $u_{n+1}^h = u^h(1) = 0$. Also eliminieren wir diese zwei Unbekannten und suchen nur nach dem Vektor $U_h = (u_j^h)_{j=1,\ldots,n}$.

Um nun die Werte von u_j^h an den inneren Gitterpunkten zu berechnen, konstruieren wir finite Differenzen mittels Taylorentwicklung. Für eine glatte Funktion φ gilt ja (beachte $h = x_{j+1} - x_j = x_j - x_{j-1}$)

$$\begin{aligned}\varphi(x_{j+1}) &= \varphi(x_j) + \varphi'(x_j)h + \frac{1}{2}\varphi''(x_j)h^2 + \frac{1}{6}\varphi'''(x_j)h^3 + \mathcal{O}(h^4)\\ \varphi(x_{j-1}) &= \varphi(x_j) - \varphi'(x_j)h + \frac{1}{2}\varphi''(x_j)h^2 - \frac{1}{6}\varphi'''(x_j)h^3 + \mathcal{O}(h^4).\end{aligned}$$

Addieren wir diese Gleichungen und dividieren durch h^2, so erhalten wir die Formel

$$\varphi''(x_j) = \frac{1}{h^2}\left(\varphi(x_{j+1}) - 2\varphi(x_j) + \varphi(x_{j-1})\right) + \mathcal{O}(h^2).$$

Daraus erhalten wir den klassischen Differenzenquotienten zur Approximation der zweiten Ableitung, d.h.,

$$\frac{\partial^2 u^h}{\partial x^2}(x_j) \approx \frac{1}{h^2}(u_{j+1}^h - 2u_j^h + u_{j-1}^h)$$

und wir ersetzen (1.1) durch die diskretisierte Version

$$-\frac{1}{h^2}(u_{j+1}^h - 2u_j^h + u_{j-1}^h) = f(x_j) := f_j, \qquad j = 1, \ldots, n. \tag{1.2}$$

Mit der Matrix $\mathbf{K}_h \in \mathbb{R}^{n\times n}$

$$\mathbf{K}_h = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & 0 & 0 & \ldots & 0 & 0\\ -1 & 2 & -1 & 0 & \ldots & 0 & 0\\ 0 & -1 & 2 & -1 & \ldots & 0 & 0\\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots\\ 0 & 0 & 0 & 0 & \ldots & -1 & 0\\ 0 & 0 & 0 & 0 & \ldots & 2 & -1\\ 0 & 0 & 0 & 0 & \ldots & -1 & 2 \end{pmatrix} \tag{1.3}$$

und dem Vektor $F_h = (f(x_j))_{j=1,\ldots,n}$ können wir (1.2) in der Form

$$\mathbf{K}_h U_h = F_h. \tag{1.4}$$

schreiben.

Wir sehen sofort, dass n gross und damit h klein sein muss, damit die obige Taylorentwicklung und Approximation sinnvoll ist. Damit erhalten wir ein grosses lineares Gleichungssystem. Wir beobachten die folgenden Eigenschaften der Systemmatrix $\mathbf{K}_h$:

- Die Matrix $\mathbf{K}_h$ ist *dünnbesetzt*, d.h. nur ein kleiner Teil der Einträge ist von Null verschieden. Wie wir sehen werden, ist diese Eigenschaft kein Zufall, sondern tritt bei allen Verfahren die wir kennen lernen auf. Der Grund dafür ist die Lokalität der Differentialoperatoren, es ist intuitiv einleuchtend dass Funktionswerte in weiter entfernten Gitterpunkten für den Wert der Ableitung unbedeutend sind (und dies führt dann zu den Nulleinträgen in der Systemmatrix).

- Die Matrix $\mathbf{K}_h$ ist symmetrisch. Dies ist ein Resultat der Symmetrie des Differentialoperators (siehe unten).

- Die Matrix $\mathbf{K}_h$ ist positiv definit (siehe unten). Dies ist ebenfalls kein Zufall, sondern ein Resultat der Elliptizität der Gleichung.

- Die Matrix $\mathbf{K}_h$ ist monoton (siehe Kapitel 2), d.h. aus $F_h \geq 0$ folgt $U_h = \mathbf{K}_h^{-1} F_h \geq 0$. Dies ist eine Konsequenz aus dem Maximumprinzip (Monotonie) für elliptische Differentialgleichungen und wird in diesem Fall durch die Diskretisierung erhalten (was nicht für jede Diskretisierung der Fall ist).

Wir sehen also, dass das Gleichungssystem (1.4) viel Struktur aufweist, die wir auch bei der numerischen Lösung verwenden können. Die Dünnbesetztheit kann z.B. speichertechnisch ausgenutzt werden, man muss nur die Indizes und Werte der Nichtnullelemente speichern. Wir werden später sehen, dass auch andere strukturelle Eigenschaften für die effiziente Lösung von (1.4) wichtig sind.

Während die Lösung von (1.4) ein Problem der *numerischen linearen Algebra* ist, verbleiben noch die klassischen Probleme der *numerischen Analysis*:

- *Existenz und Eindeutigkeit:* Existiert die diskrete Lösung u^h (bzw. U_h) und ist sie eindeutig ?

- *Stabilität:* Bleibt die Lösung u^h für $h \to 0$ beschränkt (in einem noch zu klärenden Sinn) ?

- *Konsistenz:* Ergibt sich ein kleines Residuum, wenn man die Lösung u der Differentialgleichung in das diskretisierte Problem einsetzt, bzw. konvergiert dieses Residuum gegen Null für $h \to 0$?

- *Konvergenz:* Konvergiert für $h \to 0$ u^h gegen die kontinuierliche Lösung u (in einem noch zu klärenden Sinn) ?

- *Fehlerabschätzung:* Können wir die Fehler $u - u^h$ sinnvoll abschätzen als Funktion von h (in einer passenden Norm) ?

All diese Probleme werden wir im Laufe dieser Vorlesung behandeln. Für unser spezielles Beispiel ergibt sich natürlich Existenz und Eindeutigkeit sofort aus der positiven Definitheit der Systemmatrix. Weiter sehen wir aus dem obigen Argument für den Differenzenquotienten (angewandt auf $\varphi = u$) sofort die Konsistenz, falls u glatt genug ist. Dies ist im eindimensionalen Fall sofort durch die Eigenschaften von f nachprüfbar: $f \in C^k$ impliziert $u \in C^{k+2}$. Im mehrdimensionalen steckt hinter solchen Aussagen allerdings die komplizierte Regularitätstheorie für Lösungen partieller Differentialgleichungen.

Als alternative Methode zur Diskretisierung betrachten wir *finite Elemente* (FE). Die Grundidee einer FE Methode ist die Berechnung einer Näherungslösung als Linearkombination gegebener Basisfunktionen

$$u^h(x) = \sum_{j=1}^{n} u_j^h \phi_j^h(x)$$

wobei die Funktionen ϕ_j^h einen lokalen Träger haben. Die Basisfunktionen werden ebenfalls mit einem Gitter assoziert und erfüllen üblicherweise die Bedingung

$$\phi_j^h(x_k) = \delta_{jk}$$

mit dem Kronecker-Symbol δ_{jk}, das durch $\delta_{jj} = 1$ und $\delta_{jk} = 0$ für $k \neq j$ definiert ist. Man sieht leicht, dass unter dieser Bedingung die Werte u_j^h tatsächlich die Funktionswerte an den Gitterpunkten sind, d.h. $u^h(x_j) = u_j^h$. Um die Werte u_j^h durch direktes Einsetzen von u_h in die Differentialgleichung zu bestimmen, bräuchte man sehr glatte Funktionen ϕ_j^h und hätte ein sehr schlecht konditioniertes Gleichungssystem zu lösen. Deshalb geht man zur schwachen Formulierung der Differentialgleichung über, die man durch Multiplikation mit einer Testfunktion, Integration und anschliessender partieller Integraton erhält. Im Fall von (1.1) ist die schwache Formulierung gegeben durch die Variationsgleichung

$$\int_0^1 \frac{\partial u}{\partial x}(x) \frac{\partial \varphi}{\partial x}(x)\ dx = \int_0^1 f(x)\varphi(x) \tag{1.5}$$

für alle hinreichend glatten Testfunktionen φ. Zur Diskretisierung verwendet man nun einen Ansatz für u^h wie oben und wählt die Basisfunktionen φ_j^h als natürliche Testfunktionen. Damit erhält man die diskrete Variationsgleichung

$$\int_0^1 \frac{\partial u^h}{\partial x}(x) \frac{\partial \varphi_k^h}{\partial x}(x)\ dx = \int_0^1 f(x)\varphi_k^h(x)\ dx, \qquad k = 1, \ldots, n. \tag{1.6}$$

Durch Einsetzen der Linearkombination für u_j^h erhalten wir

$$\sum_j \int_0^1 \frac{\partial \varphi_j^h}{\partial x}(x) \frac{\partial \varphi_k^h}{\partial x}(x)\ dx u_j^h = \int_0^1 f(x)\varphi_k^h(x)\ dx, \qquad k = 1, \ldots, n.$$

Definieren wir wieder eine Systemmatrix $\mathbf{K}_h$ und einen Vektor F_h, in diesem Fall

$$\mathbf{K}_h = \left(\int_0^1 \frac{\partial \varphi_j^h}{\partial x}(x) \frac{\partial \varphi_k^h}{\partial x}(x)\ dx \right)_{j,k=1,\ldots,n}, \qquad F_h = \left(\int_0^1 f(x)\varphi_j^h(x)\ dx \right)_{j=1,\ldots,n},$$

dann lässt sich das diskrete Problem wieder in der Form (1.4) schreiben.

Wir sehen wieder, dass die Matrix $\mathbf{K}_h$ dünnbesetzt sein wird, denn für grosse Werte von $|j-k|$ werden sich die Träger von φ_j^h und φ_k^h nicht überschneiden und damit gilt entweder $\frac{\partial \varphi_j^h}{\partial x}(x) = 0$ oder $\frac{\partial \varphi_k^h}{\partial x}(x) = 0$. Dies wird noch deutlicher für die klassische Wahl stückweise linearer Ansatzfunktionen

$$\phi_j^h(x) = \begin{cases} \frac{x - x_{j-1}}{h} & \text{falls } x \in [x_{j-1}, x_j] \\ \frac{x_{j+1} - x}{h} & \text{falls } x \in [x_j, x_{j+1}] \\ 0 & \text{falls } |x - x_j| > h \end{cases} \tag{1.7}$$

Diese Ansatzfunktionen sind nicht C^1, wie wir aber noch sehen werden genügt es die Ableitungen stückweise in den Teilintervallen zu definieren. Wir erhalten dann

$$\frac{\partial \phi_j^h}{\partial x}(x) = \begin{cases} \frac{1}{h} & \text{falls } x \in [x_{j-1}, x_j] \\ -\frac{1}{h} & \text{falls } x \in [x_j, x_{j+1}] \\ 0 & \text{falls } |x - x_j| > h \end{cases}$$

Man berechnet leicht die Systemmatrix (Übung) als

$$\mathbf{K}_h = \frac{1}{h} \begin{pmatrix} 2 & -1 & 0 & 0 & \dots & 0 & 0 \\ -1 & 2 & -1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 2 & -1 & \dots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & -1 & 0 \\ 0 & 0 & 0 & 0 & \dots & 2 & -1 \\ 0 & 0 & 0 & 0 & \dots & -1 & 2 \end{pmatrix} \tag{1.8}$$

d.h. $\mathbf{K}_h$ aus (1.3) und (1.8) unterscheiden sich nur um einen Faktor h. Diese unterschiedliche Skalierung ist eine Konsequenz der Integration, die bei der Definition der schwachen bzw. FE Lösung verwendet wurde. Der Unterschied zur Diskretisierung mit finiten Differenzen wird deutlicher an der rechten Seite, die dort aus Punktauswertungen bestand, im Fall der FE Diskretisierung aber aus lokalen (gewichteten) Mittelwerten. Dadurch kann die FE Methode auch leicht für unstetige (oder sogar distributionelle) rechte Seiten angewandt werden.

Die Eigenschaften der Systemmatrizen resultierend aus finiten Differenzen bzw. finiten Elementen sind sehr ähnlich, manche Eigenschaften sind aber im FE Fall viel leichter nachzuprüfen. So sieht man sofort (auch ohne Berechnung) die Symmetrie von $\mathbf{K}_h$, da ja

$$(\mathbf{K}_h)_{jk} = \int_0^1 \frac{\partial \varphi_j^h}{\partial x}(x) \frac{\partial \varphi_k^h}{\partial x}(x)\, dx = \int_0^1 \frac{\partial \varphi_k^h}{\partial x}(x) \frac{\partial \varphi_j^h}{\partial x}(x)\, dx = (\mathbf{K}_h)_{kj}$$

gilt. Weiter sieht man sofort die positive Definitheit, da für einen Vektor $V \in \mathbb{R}^n \setminus \{0\}$ gilt

$$\begin{aligned} V \cdot (\mathbf{K}_h V) &= \sum_{j,k=1}^n v_j v_k \int_0^1 \frac{\partial \varphi_j^h}{\partial x}(x) \frac{\partial \varphi_k^h}{\partial x}(x)\, dx \\ &= \int_0^1 \left(\sum_{j=1}^n v_j \frac{\partial \varphi_j^h}{\partial x}(x) \right) \left(\sum_{j=1}^n v_k \frac{\partial \varphi_k^h}{\partial x}(x) \right) dx \\ &= \int_0^1 \left(\sum_{j=1}^n v_j \frac{\partial \varphi_j^h}{\partial x}(x) \right)^2 dx > 0. \end{aligned}$$

Auch die Stabilität der finiten Elemente Diskretisierung ist leicht nachprüfbar. Man be-

achte, dass

$$\begin{aligned}\int_0^1 \left(\frac{\partial u^h}{\partial x}\right)^2 \, dx &= \int_0^1 \left(\sum_{j=1}^n u_j^h \frac{\partial \varphi_j^h}{\partial x}(x)\right)^2 \, dx \\ &= U_h \cdot (\mathbf{K}_h U_h) = U_h \cdot F_h \\ &= \sum_{j=1}^n \int_0^1 u_j^h \varphi_j^h(x) f(x) \, dx \\ &= \int_0^1 u^h(x) f(x) \, dx \\ &\leq \sqrt{\int_0^1 |u^h(x)|^2 \, dx} \sqrt{\int_0^1 |f(x)|^2 \, dx},\end{aligned}$$

wobei wir die Cauchy-Schwarz Ungleichung in $L^2([0,1])$ für die letzte Zeile verwendet haben. Die Poincare-Ungleichung

$$\int_0^1 |\varphi(x)|^2 \, dx \leq \frac{1}{4} \int_0^1 \left(\frac{\partial \varphi}{\partial x}(x)\right)^2 \, dx$$

für Funktionen mit Randwerten $\varphi(0) = \varphi(1) = 0$ liefert dann die Stabilitätsabschätzung

$$\int_0^1 |u^h(x)|^2 \, dx \leq \frac{1}{4} \int_0^1 \left(\frac{\partial u^h}{\partial x}(x)\right)^2 \, dx \leq \frac{1}{16} \int_0^1 |f(x)|^2 \, dx.$$

Also erhalten wir eine von h unabhängige Schranke an die L^2-Norm sowohl von u^h als auch von $\frac{\partial u^h}{\partial x}$.

1.1.2 Parabolische Probleme: Die Wärmeleitungsgleichung

Als Beispiel für ein parabolisches Problem betrachten wir die eindimensionale Wärmeleitungsgleichung

$$\frac{\partial u}{\partial t}(x,t) - \frac{\partial^2 u}{\partial x^2}(x,t) = f(x,t), \quad u(x,0) = u_0(x), u(0,t) = u(1,t) = 0, \quad x \in (0,1), t \in (0,T). \tag{1.9}$$

Nun haben wir ein Anfangs-Randwertproblem im Raum-Zeit Zylinder zu lösen und benötigen neben der Orts- auch noch eine Zeitdiskretisierung. Dabei stellt sich sofort die Frage nach der Reihenfolge der Diskretisierung: Soll zuerst im Ort und dann in der Zeit diskretisiert werden oder umgekehrt (man nennt diese beiden Zugänge *horizontale* bzw. *vertikale Linienmethode*). Man könnte auch direkt in der Raum-Zeit diskretisieren, z.B. durch geeignete mehrdimensionale finite Elemente. In den meisten Fällen führen alle Zugänge aber auf ähnliche Diskretisierungen, so dass wir uns hier auf die vertikale Linienmethode beschränken, d.h. wir diskretisieren zuerst im Ort.

Bei einer finiten Differenzen Diskretisierung im Ort suchen wir nun die Werte $u_j^h(t)$ an den Gitterpunkten x_j für jeden Zeitpunkt. Da wir den selben Differentialoperator diskretisieren, erhalten wir sofort (mit der obigen Notation) das semi-diskrete Problem

$$\frac{dU_h}{dt}(t) + \mathbf{K}_h U_h(t) = F^h(t), \quad U_h(0) = (u_0(x_j))_{j=1,\ldots n} \tag{1.10}$$

d.h. ein Anfangswertproblem für ein System gewöhnlicher Differentialgleichungen.

Bei einer finiten Elemente Diskretisierung starten wir von der schwachen Form der Wärmeleitung

$$\int_0^1 \frac{\partial u}{\partial t}(x,t)\varphi(x) + \frac{\partial u}{\partial x}(x)\frac{\partial \varphi}{\partial x}(x)\ dx = \int_0^1 f(x,t)\varphi(x)\ dx$$

und machen für u^h wieder einen Ansatz als Linearkombination (mit zeitabhängigen Koeffizienten) der φ_j^h, die wir auch als Testfunktionen benützen. Damit erhalten wir ein diskretes System der Form

$$\mathbf{M}_h \frac{dU_h}{dt}(t) + \mathbf{K}_h U_h(t) = F^h(t), \quad U_h(0) = \left(\int_0^1 u_0(x)\varphi_j^h(x)\right)_{j=1,\dots n} \tag{1.11}$$

wobei wir nun zusätzlich eine Massenmatrix $\mathbf{M}_h$ erhalten, definiert durch

$$\mathbf{M}_h = \frac{h}{6}\begin{pmatrix} 4 & 1 & 0 & 0 & \dots & 0 & 0 \\ 1 & 4 & 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & 4 & 1 & \dots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & 0 & 0 & \dots & 4 & 1 \\ 0 & 0 & 0 & 0 & \dots & 1 & 4 \end{pmatrix}.$$

Man überprüft wieder leicht, dass $\mathbf{M}_h$ symmetrisch und positiv definit ist. Zumindest für theoretische Zwecke können wir deshalb $\mathbf{M}_h^{-1}$ anwenden und erhalten mit $\hat{\mathbf{K}}_h = \mathbf{M}_h^{-1}\mathbf{K}_h$ sowie $\hat{F}_h = \mathbf{M}_h^{-1}F_h$ wieder ein analoges System von gewöhnlichen Differentialgleichungen wie in (1.10) (durch Anwendung von $\mathbf{M}_h^{-1/2}$ von links und rechts kann auch die Symmetrie erhalten werden) . Deshalb beschränken wir uns im folgenden auf die Zeitdiskretisierung von (1.10).

Zur Zeitdiskretisierung führen wir ein Gitter auf dem Intervall $(0,T)$ ein, zur Vereinfachung wieder ein reguläres Gitter mit Punkten $t_k = k\tau, k = 0,\dots,m$, und Zeitschrittweite $\tau = \frac{T}{m}$. Nun approximieren wir U_h wieder durch Werte an den diskreten Zeitpunkten, d.h. wir suchen $U_{h,\tau}(t_k)$. Die Zeitableitung können wir wieder mit einem Differenzenquotienten berechnen. Dafür haben wir nun mehrere Möglichkeiten, wobei die einfachste ein Vorwärtsdifferenzenquotient ist, d.h.

$$\frac{dU_h}{dt}(t_k) \approx \frac{1}{\tau}\left(U_{h,\tau}(t_{k+1}) - U_{h,\tau}(t_k)\right). \tag{1.12}$$

Durch Einsetzen erhalten wir das *Vorwärts*-Euler Verfahren, eine *explizite Zeitdiskretisierung* der Form

$$U_{h,\tau}(t_{k+1}) = U_{h,\tau}(t_k) - \tau\left(\mathbf{K}_h U_{h,\tau}(t_k) - F_h(t_k)\right), \quad U_{h,\tau}(0) = U_h(0). \tag{1.13}$$

Bei der expliziten Diskretisierung ist keine Lösung eines Gleichungssystems nötig, in jedem Schritt können wir die diskrete Lösung direkt durch Anwenden der Matrix $\mathbf{K}_h$ aus dem vorherigen Zeitschritt bestimmen. Dadurch ist in diesem Fall die Existenz und Eindeutigkeit der diskreten Lösung klar. Nicht klar ist jedoch die Stabilität der diskreten Lösung. Zur

Vereinfachung untersuchen wir dabei den Fall $F_h = 0$. Seien λ_j, $j = 1, \ldots, n$ die Eigenwerte von $\mathbf{K}_h$ und sei Σ eine Orthogonalmatrix bestehend aus Eigenvektoren, sodass

$$\Sigma^T \mathbf{K}_h \Sigma = \text{ diag } (\lambda_j).$$

Definieren wir nun $V^k = \Sigma^T U_{h,\tau}(t_k)$, dann erhalten wir die Differenzengleichung

$$V^{k+1} = V^k - \tau \text{ diag } (\lambda_j) V^k,$$

oder komponentenweise

$$V_j^{k+1} = (1 - \tau\lambda_j) V_j^k.$$

Daraus können wir die Lösung als

$$V_k^j = (1 - \tau\lambda_j)^k V_j^0$$

berechnen. Stabilität erhalten wir nur für $|1 - \tau\lambda_j| \leq 1$, da sonst V_k^j geometrisch anwächst. Da $\mathbf{K}_h$ positiv definit ist, sind alle λ_j positiv und damit $1 - \tau\lambda_j < 1$. Weiter muss aber gelten $1 - \tau\lambda_j \geq -1$, oder als Schranke für die Zeitschrittweite $\tau \leq \frac{2}{\lambda_j}$ (für alle j). Aus der Skalierung in (1.3) erwarten wir, dass der grösste Eigenwert von $\mathbf{K}_h$ von der Ordnung h^{-2} ist (dies lässt sich auch beweisen). Also erhalten wir eine Schranke der Form $\tau = \mathcal{O}(h^2)$, d.h. die Zeitschrittweite muss sehr klein im Vergleich zur örtlichen Gittergrösse sein.

Eine Alternative zur expliziten Zeitdiskretisierung ist die Wahl eines Rückwärts-Differenzenquotienten

$$\frac{dU_h}{dt}(t_k) \approx \frac{1}{\tau} \left(U_{h,\tau}(t_k) - U_{h,\tau}(t_{k-1}) \right). \tag{1.14}$$

Mit dieser Wahl erhalten wir das *Rückwärts*-Euler Verfahren, eine *implizite Zeitdiskretisierung* der Form

$$U_{h,\tau}(t_k) + \tau \mathbf{K}_h U_{h,\tau}(t_k) = U_{h,\tau}(t_{k-1}) \tau F_h(t_k), \quad U_{h,\tau}(0) = U_h(0). \tag{1.15}$$

Im Gegensatz zu expliziten Verfahren erfordert die implizite Diskretisierung die Lösung eines linearen Gleichungssystems in jedem Zeitschritt. Die Systemmatrix $\mathbf{I} + \tau\mathbf{K}_h$ hat analoge Eigenschaften wie im elliptischen Fall. Würde man die obige Diskretisierung aus einer horizontalen Linienmethode herleiten, so sieht man, dass in jedem Zeitschritt eine Ortsdiskretisierung der elliptischen Gleichung

$$u^\tau(x, t_k) - \tau \frac{\partial^2 u^\tau}{\partial x^2}(x, t_k) = u^\tau(x, t_{k-1}) + \tau f(x, t_k)$$

gelöst wird. Damit ist natürlich der numerische Aufwand in jedem Schritt eines impliziten Verfahrens ungleich höher als in einem Schritt eines expliziten Verfahrens. Dies kann allerdings in den meisten Fällen durch eine grössere Zeitschrittweite kompensiert werden. Führen wir für $F_h = 0$ eine analoge Diagonalisierung wie im expliziten Fall durch, so erhalten wir die Rekursion

$$(1 + \tau\lambda_j) V_j^{k+1} = V_j^k$$

mit der Lösung

$$V_j^k = \frac{1}{(1 + \tau\lambda_j)^k} V_j^0.$$

Da nun $1 + \tau\lambda_j > 1$ gilt, erhalten wir sogar geometrischen Abfall der V_j^k (was die Wärmeleitung ohne Quelle natürlich besser approximiert) und damit insbesondere Stabilität unabhängig von der Zeitschrittweite.

1.1.3 Hyperbolische Probleme: Die Transportgleichung

Als einfaches Beispiel für hyperbolische Probleme (wie alle Differentialgleichungen erster Ordnung) betrachten wir die lineare eindimensionale Transportgleichung

$$\frac{\partial u}{\partial t}(x,t) + \frac{\partial u}{\partial x}(x,t) = 0, \quad u(x,0) = u_0(x), \quad u(0,t) = 0, \quad x \in (0,1), t \in (0,T). \tag{1.16}$$

Wegen der einfacheren Darstellung beschränken wir uns auf finite Differenzenverfahren zur Ortsdiskretisierung, diese sind auch die häufigst verwendeten für hyperbolische Probleme.

Wie schon zuvor bei der Zeitdiskretisierung haben wir verschiedene Möglichkeiten bei der Wahl der Differenzenquotienten für den Operator erster Ordnung $\frac{\partial u}{\partial x}(x,t)$. Bei der Wahl eines Vorwärtsdifferenzenquotienten erhalten wir das semidiskrete Verfahren

$$\frac{du_j^h}{dt}(t) = -\frac{1}{h}(u_{j+1}^h - u_j^h(t)) \tag{1.17}$$

bei einem Rückwärtsquotienten

$$\frac{du_j^h}{dt}(t) = -\frac{1}{h}(u_j^h - u_{j-1}^h(t)) \tag{1.18}$$

und bei einem zentralen Differenzenquotienten

$$\frac{du_j^h}{dt}(t) = -\frac{1}{2h}(u_{j+1}^h - u_{j-1}^h(t)). \tag{1.19}$$

In Matrixform erhalten wir in jedem Fall

$$\frac{dU_h}{dt}(t) = \mathbf{D}_h U_h(t)$$

mit den Matrizen

$$\mathbf{D}_h^+ = \frac{1}{h}\begin{pmatrix} 1 & -1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & -1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & -1 & \dots & 0 & 0 \\ 0 & 0 & 0 & 1 & \dots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & -1 \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix},$$

$$\mathbf{D}_h^- = \frac{1}{h}\begin{pmatrix} -1 & 0 & 0 & 0 & \dots & 0 & 0 \\ 1 & -1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & -1 & 0 & \dots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & -1 & 0 \\ 0 & 0 & 0 & 0 & \dots & 1 & -1 \end{pmatrix},$$

und

$$\mathbf{D}_h^c = \frac{1}{2h}\begin{pmatrix} 0 & -1 & 0 & 0 & \dots & 0 & 0 \\ 1 & 0 & -1 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & -1 & \dots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 0 & -1 \\ 0 & 0 & 0 & 0 & \dots & 1 & 0 \end{pmatrix}.$$

Man sieht sofort, dass beim Vorwärts- und beim zentralen Differenzenquotienten Probleme mit den Randbedingungen auftreten, da der Wert von u^h_{n+1}, d.h. von $u^h(1)$ benötigt würde. Beim Rückwärtsdifferenzenquotienten hingegen genügt es den Wert von $u^h_0 = u^h(0) = 0$ einzusetzen, was der gegebenen Randbedingung entspricht.

Im Fall des Rückwärtsdifferenzenquotienten können wir das semidiskrete Problem explizit lösen. Wegen $u^0_h = 0$ erhalten wir für u^1_h die Differentialgleichung

$$\frac{du^h_1}{dt}(t) = -\frac{1}{h}u^h_1, \qquad u^h_1(0) = u_0(x_1)$$

mit der Lösung $u^h_1(t) = e^{-\frac{t}{h}}u_0(x_1)$. Die weiteren Gleichungen können wir ebenfalls explizit lösen und erhalten induktiv

$$u^h_j(t) = e^{-\frac{t}{h}} \sum_{i=1}^{j} u_0(x_i) \frac{1}{(j-i)!} \left(\frac{t}{h}\right)^{j-i}.$$

Man sieht sofort die Stabilität des Verfahrens in der Supremum-Norm, es gilt

$$|u^h_j(t)| \leq e^{-\frac{t}{h}} \sum_{i=1}^{j} |u_0(x_i)| \frac{1}{(j-i)!} \left(\frac{t}{h}\right)^{j-i} \leq \max_i |u_0(x_i)| e^{-\frac{t}{h}} \sum_{i=0}^{j-1} \frac{1}{i!} \left(\frac{t}{h}\right)^{i} \leq \max_i |u_0(x_i)| \leq \|u_0\|_\infty.$$

Bei einem Vorwärtsdifferenzenquotienten erhalten wir hingegen

$$u^h_j(t) = \frac{1}{h} e^{\frac{t}{h}} \sum_{i=j}^{n} u_0(x_i) \frac{1}{i!} \left(-\frac{t}{h}\right)^{i-j} - \int_0^t (s-t)^{n-j} e^{\frac{t-s}{h}} u^h_{n+1}(s)\, ds$$

und sehen sofort dass wir durch den Faktor $e^{\frac{t}{h}}$ einen in der Zeit wachsenden Anteil erhalten, der für Instabilität des Verfahrens sorgt. Beim zentralen Differenzenquotienten erhält man in ähnlicher Weise Instabilität.

Der Grund für die Stabilität des Vorwärtsdifferenzenquotienten liegt in der Ausbreitung der Charakteristiken der hyperbolischen Gleichung (1.16). Die charakteristischen Gleichungen sind gegeben durch

$$\frac{dt}{d\tau} = 1, \qquad \frac{dx}{d\tau} = 1,$$

und somit erhält man Charakteristiken als Geraden der Form $x = t + c$. Entlang der Charakteristiken gilt in diesem Fall sogar $\frac{d}{d\tau}u(x(\tau), t(\tau)) = 0$, d.h. die Lösung ist konstant. Gehen wir vorwärts in der Zeit, dann wird die Information entlang der Charakteristiken also von links nach rechts ausgebreitet. Dieser Sichtweise entspricht der Rückwärtsdifferenzenquotient, da er zur Berechnung des Wertes am Gitterpunkt x_j nur Werte links dieses Gitterpunkts verwendet. Der Vorwärts- und zentrale Differenzenquotient verletzen hingegen die Kausalität, da sie zur Berechnung des Wertes in x_j auch auf den rechten Gitterpunkt x_{j+1} zugreifen. Man sieht in diesem Beispiel, dass die Charakteristiken bei hyperbolischen Problemen von grosser Bedeutung für die Konstruktion stabiler Verfahren sind. Wollen wir die Analysis also auf eine allgemeinere Gleichung, etwa

$$\frac{\partial u}{\partial t}(x,t) + v(x,t)\frac{\partial u}{\partial x}(x,t) = 0, \quad u(x,0) = u_0(x), \tag{1.20}$$

verallgemeinern, so berechnen wir zuerst die Charakteristiken

$$\frac{dt}{d\tau} = 1, \frac{dx}{d\tau} = v(x,t).$$

Hier können wir wieder $t = \tau$ setzen und erhalten also $\frac{dx}{dt} = v(x,t)$. Für die Ausbreitungsrichtung der Charakteristiken ist dann nur das Vorzeichen von v entscheidend. Ist v positiv verwenden wir wie oben den Rückwärtsdifferenzenquotienten, andernfalls den Vorwärtsquotienten. In Kurzform erhalten wir so das *Upwind-Verfahren*

$$\frac{du_j^h}{dt} + \max\{v(x_j,t),0\}\frac{u_j^h - u_{j-1}^h}{h} + \min\{v(x_j,t),0\}\frac{u_{j+1}^h - u_j^h}{h} = 0.$$

Im allgemeinen ist es nicht möglich, die semidiskreten Probleme wie oben zu lösen, weshalb auch eine Zeitdiskretisierung notwendig ist. Hierzu wählen wir wieder Gitterpunkte $t_k = k\tau$ auf der Zeitskala, mit kleinem Zeitschritt τ. Wir bezeichnen die Werte der diskreten Lösung am Ort x_j und zum Zeitpunkt t_k mit $u_{j,k}^h$. Nun haben wir wieder mehrere Möglichkeiten für die Wahl des Differenzenquotienten in der Zeit. Der Einfachheit halber wählen wir den Vorwärtsdifferenzenquotienten und erhalten wir das Vorwärts-Euler Verfahren

$$\frac{u_{j,k+1}^h - u_{j,k}^h}{\tau} + \frac{u_{j,k}^h - u_{j-1,k}^h}{h} = 0,$$

das die explizite Berechnung der Werte im nächsten Zeitschritt als

$$u_{j,k+1}^h = (1 - \frac{\tau}{h})u_{j,k}^h + \frac{\tau}{h}u_{j-1,k}^h$$

erlaubt. Aus dieser Formel sehen wir auch die Stabilität des Verfahrens abhängig von $\lambda = \frac{\tau}{h}$. Für $\lambda \leq 1$ ist die Lösung im nächsten Zeitschritt Konvexkombination von Werten im letzten Zeitschritt und Maximum und Minimum können deshalb im Verlauf der Zeit nicht grösser werden. Durch Rückeinsetzen der Zeitschritte erhalten wir

$$u_{j,k}^h = \sum_{i=0}^{k} \left(\begin{array}{c} n \\ i \end{array} \right) (1-\lambda)^{k-i}\lambda^i u_0(x_{j-i})$$

mit $u_0(x_\ell) = 0$ für $\ell < 0$. Um die Stabilität abzuschätzen verwenden wir für $\lambda \leq 1$

$$|u_{j,k}^h| \leq \sum_{i=0}^{k} \left(\begin{array}{c} n \\ i \end{array} \right) (1-\lambda)^{k-i}\lambda^i \max_j |u_0(x_{j-i})| = \|u_0\|_\infty.$$

Für $\lambda > 1$ können hingegen Instabilitäten auftreten, da man ein geometrisches Wachstum mit Faktor > 1 erhalten kann. Sei z.B. der Anfangswert so, dass $u_0(x_0) = 1$ und $u_0(x_j) = 0$ für $j > 0$ gilt. Dann ist $u_k^h = \lambda^k$, dieser Wert wächst also in der Zeit stark an und das Verfahren ist deshalb instabil. Man nennt die Beschränkung $\lambda \leq 1$ die CFL (Courant-Friedrichs-Levy) Bedingung. Für die allgemeinere Gleichung (1.20) wird Stabilität analog durch die CFL-Bedingung $\lambda \leq \|v\|_\infty$ erreicht.

Die CFL-Bedingung kann auch bezüglich der Charakteristiken interpretiert werden, da ja auch das diskrete Verfahren eine analoge Eigenschaft hat. Im diskreten Fall wird ja die Information entlang der Geraden mit Steigung λ fortgepflanzt, im stetigen Fall entlang der

Charakteristiken mit Steigung 1. Die CFL-Bedingung impliziert also, dass die "diskreten Charakteristiken" nicht steiler sind als die kontinuierlichen, d.h. die Information wird im numerischen Verfahren nicht schneller fortgepflanzt als in der Differentialgleichung.

Bezüglich des zentralen Differenzenquotienten kann das Upwind-Verfahren auch als Verfahren mit künstlicher Diffusion dargestellt werden

$$\frac{u^h_{j,k+1} - u^h_{j,k}}{\tau} + \frac{u^h_{j+1,k} - u^h_{j-1,k}}{2h} = \frac{h}{2}\frac{u^h_{j+1,k} - 2u^h_{j,k} + u^h_{j-1,k}}{h^2}.$$

Der Differenzenquotient auf der rechten Seite ist eine Diskretisierung der zweiten Ableitung und damit approximieren wir die Differentialgleichung

$$\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = \frac{h}{2}\frac{\partial^2 u}{\partial x^2},$$

d.h. wir erhalten künstliche Diffusion mit Koeffizienten $\frac{h}{2}$. Da der Diffusionskoeffizient von h abhängt, sollte dieser zusätzliche Effekt mit kleiner werdender Gittergrösse verschwinden.

Abschliessend können wir noch den Fehler bei der numerischen Approximation untersuchen und nehmen dazu an, dass die Lösung der Transportgleichung $u \in C^2$ erfüllt. Definieren wir den punktweisen Fehler als $e^h_{j,k} = u^h_{j,k} - u(x_j, t_k)$, so gilt

$$e^h_{j,k+1} = (1-\lambda)e^h_{j,k} + \lambda e^h_{j-1,k} + (1-\lambda)(u(x_j,t_k) - u(x_j,t_{k+1})) + \lambda(u(x_{j-1},t_k) - u(x_j,t_{k+1})).$$

Nun erhalten wir durch Taylor-Entwicklung

$$\begin{aligned} u(x_j,t_k) - u(x_j,t_{k+1}) &= -\frac{\partial u}{\partial t}(x_j,t_k)\tau + \mathcal{O}(\tau^2) \\ u(x_{j-1},t_k) - u(x_j,t_{k+1}) &= -\frac{\partial u}{\partial x}(x_j,t_k)h - \frac{\partial u}{\partial t}(x_j,t_k)\tau + \mathcal{O}(\tau^2+h^2) \end{aligned}$$

und damit

$$\begin{aligned} & (1-\lambda)(u(x_j,t_k) - u(x_j,t_{k+1})) + \lambda(u(x_{j-1},t_k) - u(x_j,t_{k+1})) \\ = \; & -(1-\lambda)\tau\frac{\partial u}{\partial t}(x_j,t_k) - \lambda\tau\frac{\partial u}{\partial t}(x_j,t_k) - \lambda h\frac{\partial u}{\partial x}(x_j,t_k) + \mathcal{O}(\tau^2+\lambda h^2) \\ = \; & -\tau\underbrace{\frac{\partial u}{\partial t}(x_j,t_k) + \frac{\partial u}{\partial x}(x_j,t_k)}_{=0} + \mathcal{O}(\tau^2+\lambda h^2). \end{aligned}$$

Für den maximalen Fehler in jedem Zeitschritt $e^h_k = \max_j |e^h_{j,k}|$ erhalten wir dann die Abschätzung

$$e^h_{k+1} \le e^h_k + C(\lambda h^2 + \tau^2)$$

und nach Rückeinsetzen

$$e^h_k \le e^h_0 + Ck\tau(h+\tau).$$

Da in jedem Fall $k = \mathcal{O}(\tau^{-1})$ gilt, folgt eine Fehlerabschätzung erster Ordnung

$$e^h_k \le e^h_0 + C(h+\tau).$$

Kapitel 2

Finite Differenzen

Im folgenden werden wir uns mit der Diskretisierung von Differentialoperatoren durch finite Differenzen (FD) beschäftigen. Um die Analysis einfach zu halten, werden wir die meisten Argumente nur im linearen Fall durchführen, d.h., der Differentialoperator hat die Form

$$Lu = \sum_{|\alpha| \leq k} a_\alpha(x) \frac{\partial^\alpha u}{\partial x^\alpha} \qquad x \in \Omega \tag{2.1}$$

wobei $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$ ein Multiindex ist, und wir die üblichen Schreibweisen

$$|\alpha| = \sum_{i=1}^{d} \alpha_i, \qquad x^\alpha = x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_d^{\alpha_d}$$

benutzen. Ω kann hier sowohl ein Ortsgebiet bei stationären oder ein Orts-Zeitgebiet bei instationären Problemen bezeichnen. Zur Erweiterung auf nichtlineare Probleme - falls möglich - werden wir an einigen Stellen kurz die wesentlichen Ideen erläutern. Wir werden im Rest der Vorlesung immer annehmen, dass $\Omega \subset \mathbb{R}^d$ ein Gebiet mit stückweise C^1-Rand ist.

2.1 Differenzen-Schema

Die Grundidee eines finiten Differenzen-Schemas ist die Approximation der Ableitung durch Differenzenbildung auf einem Gitter. Im Falle einer eindimensionalen Funktion kann man die erste Ableitung etwa durch

$$\begin{aligned}
\frac{\partial u}{\partial x} &\approx D^+ u(x) = \frac{u(x+h) - u(x)}{h} \\
\frac{\partial u}{\partial x} &\approx D^- u(x) = \frac{u(x) - u(x-h)}{h} \\
\frac{\partial u}{\partial x} &\approx D^c u(x) = \frac{u(x+h) - u(x-h)}{2h},
\end{aligned}$$

für kleines $h > 0$, approximieren. Man nennt D^+ Vorwärts- , D^- Rückwärts- und D^c zentralen Differenzenquotienten. Da alle drei Quotienten im Grenzwert $h \to 0$ gegen die Ableitung konvergieren, sollte man für h hinreichend klein eine gute Approximation erhalten.

Im Falle einer glatten Funktion u erhält man eine quantitative Aussage durch Betrachtung des Restglieds bei der Taylor-Entwicklung. Es gilt nach dem Mittelwertsatz für ein $\xi \in (x, x+h)$

$$u(x+h) - u(x) = \frac{\partial u}{\partial x}(x)h + \frac{1}{2}\frac{\partial^2 u}{\partial x^2}(\xi_+)h^2$$

und damit

$$|D^+u(x) - \frac{\partial u}{\partial x}(x)| = \frac{h}{2}|\frac{\partial^2 u}{\partial x^2}(\xi_+)| \leq \frac{h}{2}\sup_\xi |\frac{\partial^2 u}{\partial x^2}(\xi)| = \frac{h}{2}\|\frac{\partial^2 u}{\partial x^2}\|_\infty.$$

Da wir dieses Argument für beliebiges x anwenden können, gilt auch

$$\|D^+u(x) - \frac{\partial u}{\partial x}(x)\|_\infty \leq \frac{h}{2}\|\frac{\partial^2 u}{\partial x^2}\|_\infty.$$

Also machen wir bei der Approximation der ersten Ableitung mit einem Vorwärts-Differenzenquotienten einen Fehler erster Ordnung in h, man spricht deshalb von einer *Konsistenzordnung eins* (siehe Definition 2.2 unten).

Für den Rückwärtsdifferenzenquotienten erhalten wir durch völlig analoge Argumente ebenfalls Ordnung eins, für den zentralen Differenzenquotienten hingegen verwenden wir

$$u(x+h) - u(x) = \frac{\partial u}{\partial x}(x)h + \frac{1}{2}\frac{\partial^2 u}{\partial x^2}(x)h^2 + \frac{1}{6}\frac{\partial^3 u}{\partial x^3}(\xi_+)h^3$$

und

$$u(x-h) - u(x) = -\frac{\partial u}{\partial x}(x)h + \frac{1}{2}\frac{\partial^2 u}{\partial x^2}(x)h^2 - \frac{1}{6}\frac{\partial^3 u}{\partial x^3}(\xi_-)h^3$$

und erhalten

$$D^c u(x) - \frac{\partial u}{\partial x} = \frac{1}{12}\left(\frac{\partial^3 u}{\partial x^3}(\xi_+) + \frac{\partial^3 u}{\partial x^3}(\xi_-)\right)h^2.$$

D.h., der zentrale Differenzenquotient erreicht Konsistenzordnung zwei.

Die natürliche Approximation für die zweite Ableitung mit Werten an drei Gitterpunkten ist

$$D^2u(x) = \frac{u(x+h) - 2u(x) + u(x-h)}{h^2}.$$

In diesem Fall verwenden wir den Mittelwertsatz in der Form

$$u(x \pm h) - u(x) = \pm\frac{\partial u}{\partial x}(x)h + \frac{1}{2}\frac{\partial^2 u}{\partial x^2}(x)h^2 \pm \frac{1}{6}\frac{\partial^3 u}{\partial x^3}(x)h^3 + \frac{1}{24}\frac{\partial^4 u}{\partial x^4}(\xi_\pm)h^3$$

und erhalten

$$D^2u(x) - \frac{\partial^2 u}{\partial x^2} = \frac{1}{24}\left(\frac{\partial^4 u}{\partial x^4}(\xi_+) - \frac{\partial^4 u}{\partial x^4}(\xi_-)\right)h^2,$$

also wiederum Konsistenzordnung zwei.

Um allgemeine Differentialoperatoren durch finite Differenzen zu approximieren verwendet man im allgemeinen die Differenzenquotienten für erste, zweite, oder höhere Ableitungen als Grundzutaten. Dies passiert auf einem Gitter

$$G_h = \{\ x \in \Omega \mid x = (x^1_{j_1}, x^2_{j_2}, \ldots, x^d_{j_d}), \quad 1 \leq j_i \leq N_i\ \}, \tag{2.2}$$

im einfachsten Fall auf einem regulären Gitter $x^i_{j_i} = x^i_1 + (j_i - 1)h$.

2.2 Konsistenz, Stabilität und Konvergenz

Im Allgemeinen approximieren wir einen Differentialoperator L durch einen diskreten (finite Differenzen) Operator L_h. Ein Differentialoperator der Ordnung k ist dann eine Abbildung $L : C^k(\Omega) \to C^0(\Omega)$, während die diskrete Approximation nur auf einem Gitter G_h definiert ist. d.h. $L_h : G_h \to \mathbb{R}^N$ mit $N = N_1 N_2 \dots N_d$. Auf dem Gitter definieren wir eine Norm $\|.\|_h$, die optimalerweise die gewünschte kontinuierliche Norm approximiert für $h \to 0$. Durch Interpolation erhält man aus den Werten am Gitter auch eine Funktion $\tilde{u}^h \in C^k(\Omega)$ bzw. einen erweiterten diskreten Operator $\tilde{L}_h : C^k(\Omega) \to C^0(\Omega)$, sodass $(\tilde{L}_h \tilde{u})|_{G_h} = L_h(u|_{G_h})$ gilt.

Zur Definition von Konsistenz können wir nun entweder L_h oder $\tilde{L}_h$ verwenden. Im ersten Fall führt dies auf diskrete Konsistenz:

Definition 2.1 (Diskrete Konsistenz). Sei $L : C^k(\Omega) \to C^0(\Omega)$ ein Differentialoperator der Ordnung k und L_h eine diskrete Approximation auf einem Gitter G_h. Die Approximation heisst *diskret konsistent*, falls

$$\|L_h(u|_{G_h}) - (Lu)|_{G_h}\|_h \to 0 \tag{2.3}$$

gilt. Die *Konsistenzordnung* der Approximation ist m, falls

$$\|L_h(u|_{G_h}) - (Lu)|_{G_h}\|_h \leq Ch^m \tag{2.4}$$

für alle $u \in C^{k+m}(\Omega)$ gilt.

Definition 2.2 (Konsistenz). Sei $L : C^k(\Omega) \to C^0(\Omega)$ ein Differentialoperator der Ordnung k und $\tilde{L}_h : C^k(\Omega) \to C^0(\Omega)$ eine diskrete Approximation. Die Approximation heisst *konsistent* in der Norm $\|.\|$, falls

$$\|\tilde{L}_h u - Lu\| \to 0 \tag{2.5}$$

gilt. Die *Konsistenzordnung* der Approximation ist m, falls

$$\|\tilde{L}_h u - Lu\| \leq Ch^m \tag{2.6}$$

für alle $u \in C^{k+m}(\Omega)$ gilt.

Oben haben wir gesehen, dass Vorwärts- und Rückwärtsdifferenzenquotienten Konsistenzordnung eins, und der zentrale Differenzenquotient Konsistenzordnung zwei hat. Andererseits haben wir im Fall der Transportgleichung (1.16) gesehen, dass der zentrale Differenzenquotient kein stabiles Verfahren liefert und es günstiger sein kann, ein Verfahren niedrigerer Ordnung zu wählen. Neben der Konsistenz benötigen wir also noch ein Stabilitätskonzept um die Güte einer numerischen Approximation zu bewerten.

Definition 2.3 (Diskrete Stabilität). Sei $L_h : G_h \to \mathbb{R}^N$ die diskrete Approximation eines Differentialoperators. Dann heisst L_h *diskret stabil*, wenn L_h^{-1} existiert, für $h > 0$ hinreichend klein und $\|L_h^{-1}\|$ gleichmässig in h beschränkt ist.

Analog können wir kontinuierliche Stabilität definieren.

Definition 2.4 (Stabilität). Sei $\tilde{L}_h : C^k(\Omega) \to C^0(\Omega)$ die Approximation eines Differentialoperators. Dann heisst $\tilde{L}_h$ *stabil*, wenn $\tilde{L}_h^{-1}$ existiert für $h > 0$ hinreichend klein und $\|L_h^{-1}\|$ gleichmässig in h beschränkt ist.

Eine der groben Faustregeln in der numerischen Approximation ist, dass Konsistenz und Stabilität zusammen Konvergenz implizieren. Dies ist auch mit unserer Definition von Konsistenz und Stabilität der Fall.

Satz 2.5. *Sei $\tilde{L}_h : C^k(\Omega) \to C^0(\Omega)$ eine stabile und konsistente Approximation eines Differentialoperators $L : C^k(\Omega) \to C^0(\Omega)$. Sei u die Lösung der Differentialgleichung $Lu = f$ und $\tilde{u}_h$ die Lösung von $\tilde{L}_h\tilde{u}_h = \tilde{f}_h$, sodass $\tilde{f}_h \to f$ für $h \to 0$. Dann ist die Approximation konvergent, d.h. $\tilde{u}_h \to u$ für $h \to 0$.*

Proof. Durch Subtraktion der Gleichungen erhalten wir

$$\tilde{L}_h(u - \tilde{u}_h) = (\tilde{L}_h - L)u + (f - f_h)$$

und wegen der Stabilität folgt

$$\|u - \tilde{u}_h\| = \|\tilde{L}_h^{-1}((\tilde{L}_h - L)u + f - f_h)\| \leq \|\tilde{L}_h^{-1}\| \left(\|(\tilde{L}_h - L)u\| + \|f - f_h\|\right),$$

mit $\|\tilde{L}_h^{-1}\|$ gleichmässig beschränkt. Wegen der Konsistenz folgt $\|(\tilde{L}_h - L)u\| \to 0$ und da $\|f - f_h\| \to 0$ folgt die Konvergenz $\|u - \tilde{u}^h\| \to 0$. □

Eine ähnliche Aussage gilt auch bezüglich der Konsistenzordnung, die sich bei einer stabilen Approximation direkt in die Konvergenzordnung übersetzen lässt:

Korollar 2.6. *Sei $\tilde{L}_h : C^k(\Omega) \to C^0(\Omega)$ eine stabile und konsistente Approximation eines Differentialoperators $L : C^k(\Omega) \to C^0(\Omega)$ mit Konsistenzordnung m. Sei u die Lösung der Differentialgleichung $Lu = f$ und $\tilde{u}_h$ die Lösung von $\tilde{L}_h\tilde{u}_h = \tilde{f}_h$, sodass $\|f_h - f\| = \mathcal{O}(h^m)$ für $h \to 0$. Dann gilt eine Fehlerabschätzung der Form*

$$\|u - \tilde{u}_h\| \leq Ch^m$$

für eine Konstante $C > 0$.

Proof. Aus der obigen Abschätzung

$$\|u - \tilde{u}_h\| \leq \|\tilde{L}_h^{-1}\| \left(\|(\tilde{L}_h - L)u\| + \|f - f_h\|\right)$$

erhalten wir direkt die Fehlerabschätzung aus der Stabilität und Konsistenzordnung. □

Man sieht aus der Definition der Konsistenz sofort, dass eine direkte Übertragung auf nichtlineare Gleichungen möglich ist. Die Stabilität hingegen ändert sich stark, da wir keine lineare Operatornorm der Inversen mehr definieren können. Man ersetzt deshalb das obige Stabilitätskonzept meist durch a-priori Abschätzungen für die diskreten Lösungen.

Bei der Anwendung dieser Konvergenzaussagen auf spezifische Gleichungen ist vor allem die Wahl der richtigen Normen entscheidend. Bei finiten Differenzen wählt man meist die Supremumsnorm, da diese auch der punktweisen Approximation der Ableitungen entspricht. Im nächsten Kapitel werden wir dies im Fall elliptischer Differentialgleichungen zweiter Ordnung durchführen.

2.3 Approximation elliptischer Gleichungen zweiter Ordnung

Im folgenden diskutieren wir die Analysis von finite Differenzen Schemata für elliptische Differentialgleichungen zweiter Ordnung. Der Prototyp einer solchen Gleichung hat die Form

$$Lu = -\sum_{i,j=1}^{d} a_{ij}(x)\frac{\partial^2 u}{\partial x_i \partial x_j} + \sum_{i=1}^{d} b_i(x)\frac{\partial u}{\partial x_i} + c(x)u = f(x), \qquad x \in \Omega. \tag{2.7}$$

Die Gleichung ist elliptisch, wenn für alle $x \in \Omega$ gilt: $c(x) \geq 0$ und $A(x) = (a_{ij}(x))$ ist eine symmetrische positiv definite Matrix ist. Wir werden uns auf uniform elliptische Gleichungen beschränken, d.h. es gibt ein $a_0 \in \mathbb{R}_+$, sodass gilt:

$$A(x) - a_0 I \text{ ist positiv definit für alle } x \in \Omega.$$

In solchen Fällen ist der kleinste Eigenwert von $A(x)$ durch a_0 nach unten beschränkt. Zusätzlich zur Gleichung benötigen wir noch Randbedingungen. Auf disjunkten Teilen des Randes von Ω gelten entweder Dirichlet-Randbedingungen $u = g_D$, Neumann-Randbedingungen $\frac{\partial u}{\partial n} = g_N$ oder Robin-Randbedingungen $\frac{\partial u}{\partial n} + \alpha u = g_R$.

2.3.1 Finite Differenzen Schema

Zur Konstruktion eines finite Differenzen Schemas starten wir wieder mit einem Gitter, der Einfachheit halber nehmen wir an, dass $\Omega = (0,1)^d$ gilt und das Gitter regulär ist, d.h.

$$G_h = \{\ (i_1 h, i_2 h, \ldots, i_d h) \mid i_j \in (0, \ldots, n+1)\ \}$$

mit $n+1 = \frac{1}{h}$. Jedem Gitterpunkt ordnen wir einen eindeutigen Multiindex $(i_1, i_2, \ldots, i_d)$ zu. Die einfachste Approximation zweiter Ableitungen erhalten wir wie oben mit einem $(2d+1)$-*Punkte Stern*, d.h. zur Approximation der zweiten Ableitung im Punkt $(i_1, i_2, \ldots, i_d)$ verwenden wir den Punkt selbst, sowie alle Punkte der Form $(i_1, \ldots, i_j \pm 1, \ldots, i_d)$, d.h. alle Multiindizes in denen genau ein Index um den Wert eins geändert wurde. Die zweite Ableitung bezüglich der j-ten Variable können wir dann durch

$$\frac{\partial^2 u^h}{\partial x_j^2}(i_1 h, i_2 h, \ldots, i_d h) \approx \frac{1}{h^2}(u^h_{i_1,\ldots,i_j+1,\ldots,i_d} - 2u^h_{i_1,\ldots,i_j,\ldots,i_d} + u^h_{i_1,\ldots,i_j-1,\ldots,i_d})$$

approximieren. Analog können wir erste Ableitungen auf dem $(2d+1)$ Punkte Stern approximieren mit den drei Differenzenquotienten, die wir oben beschrieben haben. Wegen der höheren Konsistenzordnung ist die bevorzugte Wahl im allgemeinen der zentrale Differenzenquotient

$$\left(b_j \frac{\partial u^h}{\partial x_j}\right)(i_1 h, i_2 h, \ldots, i_d h) \approx \frac{1}{2h} B^j_{i_1,\ldots,i_d}\left(u^h_{i_1,\ldots,i_j+1,\ldots,i_d} - u^h_{i_1,\ldots,i_j-1,\ldots,i_d}\right)$$

Hier ist $B^j_{i_1,\ldots,i_d}$ eine geeignete Approximation von $b_j(i_1 h, i_2 h, \ldots, i_d h)$. Falls b_j keine glatte Funktion ist, kann die richtige Wahl von $B^j_{i_1,\ldots,i_d}$ ein nichttriviales Problem sein, das wir allerdings hier nicht im Detail diskutieren wollen. Bei konvektionsdominanten Problemen (d.h. relativ grossen Werten von b_j) ist aber wie bei der Transportgleichung auf die Stabilität

zu achten, und aus analogen Gründen sollten dann keine zentralen Differenzenquotienten verwendet werden, sondern eine Approximation der Form

$$\begin{aligned}\left(b_j \frac{\partial u^h}{\partial x_j}\right)(i_1 h, i_2 h, \ldots, i_d h) \quad &\approx \quad \max\{B^j_{i_1,\ldots,i_d}, 0\}\frac{1}{h}\left(u^h_{i_1,\ldots,i_j,\ldots,i_d} - u^h_{i_1,\ldots,i_j-1,\ldots,i_d}\right) + \\ &\quad \min\{B^j_{i_1,\ldots,i_d}, 0\}\frac{1}{h}\left(u^h_{i_1,\ldots,i_j+1,\ldots,i_d} - u^h_{i_1,\ldots,i_j,\ldots,i_d}\right).\end{aligned}$$

Den Term nullter Ordnung kann man einfach durch $c_{i_1,\ldots,i_d} u^h_{i_1,\ldots,i_d}$ approximieren.

Durch dieses Vorgehen erhält man an allen inneren Gitterpunkten eine Differenzengleichung an Stelle der ursprünglichen Differentialgleichung. Es verbleiben noch die Randpunkte, d.h. $i_j = 0$ oder $i_j = n+1$. Hier benötigt man eine geeignete Approximation der Randbedingung. Der einfachste Fall ist dabei die Dirichlet-Randbedingung, die wir exakt mit der Formel

$$u^h_{i_1,\ldots,i_j,\ldots,i_d} = g_D(i_1 h, i_2 h, \ldots, i_d h)$$

in den Randgitterpunkten (d.h. für zumindest ein j gilt $i_j = 1$ oder $i_j = d$) auswerten. Im Fall einer Neumann oder Robin Randbedingung muss zusätzlich die Normalableitung approximiert werden, und zwar durch einen geeigneten einseitigen Differenzenquotienten. Für $i_j = 0$ wählen wir dazu einen negativen Vorwärtsdifferenzenquotienten, d.h.

$$\frac{\partial u^h}{\partial n}(i_1 h, i_2 h, \ldots, 0, \ldots, i_d h) = -\frac{\partial u^h}{\partial x_j}(i_1 h, i_2 h, \ldots, 0, \ldots, i_d h) \approx \frac{1}{h}\left(u^h_{i_1,\ldots,0,\ldots,i_d} - u^h_{i_1,\ldots,1,\ldots,i_d}\right).$$

Diese Wahl ist natürlich, da wir für einen Rückwärts- oder zentralen Differenzenquotienten ja einen Wert bei $x_j = -h$ benötigen würden, der nicht zur Verfügung steht. Analog verwenden wir für $i_j = n+1$ einen Rückwärtsdifferenzenquotienten

$$\frac{\partial u^h}{\partial n}(i_1 h, i_2 h, \ldots, 1, \ldots, i_d h) = \frac{\partial u^h}{\partial x_j}(i_1 h, i_2 h, \ldots, 1, \ldots, i_d h) \approx \frac{1}{h}\left(u^h_{i_1,\ldots,n+1,\ldots,i_d} - u^h_{i_1,\ldots,n,\ldots,i_d}\right).$$

Abschliessend bemerken wir, dass Verallgemeinerungen der Differenzenverfahren auf allgemeinere Gebiete und Gitter möglich sind, allerdings mit erheblichen Komplikationen verbunden sind. So muss z.B. der Rand im Fall eines allgemeinen Gebiets entsprechend approximiert werden, was meist durch Wahl zusätzlicher Gitterpunkte passiert.

2.3.2 Maximumprinzipien und Monotonie

Elliptische und parabolische Differentialgleichungen zweiter Ordnung erfüllen sogenannte *Maximumprinzipien*, die implizieren, dass die Maxima bzw. Minima von Lösungen am Rand angenommen werden. Man unterscheidet zwischen schwachen (Maxima / Minima werden sicher am Rand angenommen) und starken (Maxima / Minima werden nur am Rand und nicht im Inneren angenommen).

Satz 2.7 (Starkes Maximumprinzip). *Sei $Lu < 0 (> 0)$ mit L wie in (2.7). Dann gilt $u \leq 0 (\geq 0)$ oder u hat kein lokales Maximum (Minimum) im Inneren von Ω.*

Proof. Wir nehmen an es existiert ein Maximum von u, dass in einem Punkt $\overline{x}$ im Inneren von Ω angenommen wird, mit $u(\overline{x}) > 0$. Dann gilt wegen der notwendigen Bedingungen für

lokale Maxima, dass $\nabla u(\overline{x}) = 0$ gilt und die Hessematrix $(\frac{\partial^2 u}{\partial x_i \partial x_j})$ negativ semidefinit ist. Also folgt

$$Lu(\overline{x}) \geq -\sum_{i,j=1}^{d} a_{ij}(x) \frac{\partial^2 u}{\partial x_i \partial x_j}.$$

Da die negative Hessematrix von u in $\overline{x}$ und auch $A(\overline{x})$ positiv semidefinit sind, folgt mit dem unten stehenden Lemma 2.8 die Ungleichung $Lu(\overline{x}) \geq 0$ und somit ein Widerspruch zu $Lu < 0$.

Im Fall $Lu > 0$ erhalten wir die entsprechende Aussage über Minima durch Anwendung des ersten Teils auf $-u$. □

Es bleibt noch das Lemma über positiv definite Matrizen zu beweisen:

Lemma 2.8. *Seien $A, B \in \mathbb{R}^{d\times d}$ symmetrisch und positiv semidefinit. Dann gilt*

$$A : B := \sum_{i,j=1}^{d} A_{ij} B_{ij} \geq 0$$

Proof. Für symmetrische Matrizen existiert eine Spektralzerlegung in der Form

$$B = \sum_{k=1}^{d} \lambda_k v_k v_k^T,$$

mit den Eigenvektoren $v_j \in \mathbb{R}^d$ und den Eigenwerten $\lambda_j \in \mathbb{R}$. Weiter gilt wegen der positiven Semidefinitheit $\lambda_j \geq 0$. Benutzen wir die Notation $v_j = (v_{jk})_{k=1,\ldots,d}$, dann ist

$$B_{ij} = \sum_{k=1}^{d} \lambda_k v_{ki} v_{kj}$$

und

$$\sum_{i,j=1}^{d} A_{ij} B_{ij} = \sum_{i,j,k=1}^{d} \lambda_k A_{ij} v_{ki} v_{kj} = \sum_{k=1}^{d} \lambda_k v_k^T A v_k.$$

Da A positiv semidefinit ist, folgt $v_k^T A v_k \geq 0$ für alle k, und mit $\lambda_k \geq 0$ folgt die Aussage. □

Im Weiteren wollen wir Maximumprinzipien eher für den Fall $Lu \geq 0$ oder $Lu = 0$ anwenden, als für den Fall strikter Positivität. Deshalb beweisen wir eine schwächere Version der obigen Aussage:

Satz 2.9 (Schwaches Maximumprinzip). *Sei $Lu \leq 0 (\geq 0)$ mit L wie in (2.7). Dann gilt $u \leq 0 (\geq 0)$ oder u nimmt sein globales Maximum (Minimum) am Rand von Ω an.*

Proof. Wir nehmen an, u nimmt sein globales Maximum in einem inneren Punkt $\overline{x} \in \Omega$ an und $u(\overline{x}) > 0$. Da A positiv semidefinit ist, gilt entweder $A \equiv 0$ oder es existiert ein Index j, sodass $A_{jj}(\overline{x}) = e_j^T A(\overline{x}) e_j > 0$ (da ja sonst $v^T A(\overline{x}) v \leq 0$ für alle $v \in \mathbb{R}^d$ gilt). Dann betrachten wir die Funktionen $u^\epsilon(x) = u(x) + \epsilon \exp(\lambda(x_j - \overline{x}_j))$. Dann gilt

$$\begin{aligned} (Lu^\epsilon)(x) &= (Lu)(x) - \epsilon(\lambda^2 a_{jj}(x) - \lambda b_j(x) - c(x)) \exp(\lambda(x_j - \overline{x}_j)) \\ &\leq -\epsilon(\lambda^2 a_{jj}(x) - \lambda b_j(x) - c(x)) \exp(\lambda(x_j - \overline{x}_j)) \end{aligned}$$

und bei geeigneter Wahl von λ (hinreichend gross) können wir erreichen, dass $(Lu^\epsilon)(x)$ in einer Umgebung von x (unabhängig von ϵ) negativ ist.

Man sieht sofort, dass u^ϵ gleichmässig gegen u konvergiert. Da bei gleichmässiger Konvergenz globale Maxima gegen globale Maxima konvergieren, gibt es $x^\epsilon \to \overline{x}$, sodass u^ϵ in x^ϵ ein Maximum annimmt und dort positiv ist. Dies ist aber ein Widerspruch zu Satz 2.7, da $Lu^\epsilon < 0$ in einer Umgebung von x^ϵ gilt. □

Aus dem Maximumprinzip folgt sofort die Eindeutigkeit der Lösung des Dirichlet-Problems, da ja für zwei Lösungen u_1 und u_2 die Differenz $u = u_1 - u_2$ die Gleichung $Lu = 0$ erfüllt sowie $u = 0$ am Rand. Damit folgt $0 \leq u \leq 0$ in Ω, d.h. $u \equiv 0$.

Das starke oder schwache Maximumprinzip kann in einigen Varianten bewiesen werden. Unter anderem sehen wir aus dem Beweis von Satz (2.7), dass im Fall $c \equiv 0$ die Lösung u kein Maximum bzw. Minimum im Inneren annehmen kann. Eine Variante existiert auch im Fall $c < 0$ (in dem sich die Gleichung eher hyperbolisch als elliptisch verhält). Dort gilt dann die Aussage, dass u an einem Maximum (Minimum) im Inneren nicht negativ (positiv) sein kann. Abschliessend können wir noch die ursprüngliche Annahme der strikten Elliptizität fallen lassen, wie wir sofort sehen genügt die positive Semidefinitheit von A. Damit können wir die Maximumprinzipien auch auf parabolische Gleichungen wie die Wärmeleitungsgleichung (eine Zeile und Spalte von A identisch null) oder Gleichungen erster Ordnung wie die Transportgleichung ($A \equiv 0$) anwenden.

Eine weitere interessante Folgerung aus dem Maximumprinzip ist Stabilität in der Supremum-Norm, die wir im folgenden Satz formulieren

Satz 2.10. *Sei L ein elliptischer Differentialoperator wie in* (2.7)*. Dann existiert eine Konstante $C > 0$, sodass für Lösungen u von*

$$Lu = f \quad \text{in } \Omega, \qquad u = g \quad \text{auf } \partial\Omega$$

die Stabilitätsabschätzung

$$\|u\|_\infty \leq C \max\left\{\|f\|_\infty, \|g\|_\infty\right\} \tag{2.8}$$

gilt.

Proof. Der Beweis benutzt wieder das Maximumprinzip. Sei v eine Funktion, sodass $Lv \geq 1$ in Ω und $v \geq 1$ auf $\partial\Omega$. Dann gelten für

$$u_\pm = \pm \max\left\{\|f\|_\infty, \|g\|_\infty\right\} v,$$

die Ungleichungen

$$L(u - u_+) \leq 0, \qquad L(u_- - u) \leq 0.$$

Da sowohl $u - u_+$ als auch $u_- - u$ am Rand nichtpositiv sind, gilt nach dem Maximumprinzip

$$u_- \leq u \leq u_+ \qquad \text{in } \overline{\Omega}.$$

Also folgern wir

$$\sup_{x\in\Omega} |u(x)| \leq C \max\left\{\|f\|_\infty, \|g\|_\infty\right\},$$

wobei $C = \sup_{x\in\Omega} |v(x)|$.

Um den Beweis abzuschliessen, müssen wir noch eine passende Funktion v finden. Sei

$$v(x) = \alpha - \beta \exp(\lambda \sum (x_j - \hat{x}_j)),$$

für ein $\hat{x} \in \Omega$. Dann gilt

$$Lv = \beta \sum_j (a_{jj}\lambda^2 + b_j\lambda) \exp(\lambda \sum (x_j - \hat{x}_j)) + cv.$$

Durch passende Wahl von α, β und λ (hinreichend gross) können wir erreichen, dass $v \geq 1$ und $Lv \geq 1$ gilt (wegen $a_{jj} = e_j^T A e_j \geq \lambda_{min}(A) \geq a_0 > 0$). □

2.3.3 M-Matrizen und diskrete Monotonie

Im folgenden betrachten wir die finite Differenzen Diskretisierung und ihre Analyse etwas genauer im vereinfachten Fall $A(x) = a(x)I$ mit einer skalaren Funktion a. Zur einfacheren Notation schränken wir uns auch auf den Fall $d = 2$ ein, alle Argumente sind aber nicht dimensionsabhängig und für beliebiges d analog (mit grösserer Schreibarbeit). Wir nehmen an, dass $\Omega = (0,1)^2$ gilt und verwenden ein reguläres Gitter

$$G_h = \{\ (ih, jh) \mid i,j = 0,1,\ldots,n+1, h = \frac{1}{n+1}\ \}.$$

Die Gesamtanzahl der Gitterpunkte ist dann $(n+1)^2$, bzw. der inneren Gitterpunkte ist $N = n^2$. Die äusseren Gitterpunkte $i,j \in \{0, n+1\}$ können wir aus der Dirichlet-Randbedingung sofort eliminieren.

Entsprechend der obigen Diskussion von Differenzen-Schema analysieren wir eine Diskretisierung auf einem Fünf-Punkte Stern der Form

$$\begin{aligned} \frac{a_{ij}}{h^2}(4u_{ij} - u_{ij+1} - u_{i+1j} - u_{ij-1} - u_{i-1j}) + \\ \frac{b_{1,ij}}{2h}(u_{i+1j} - u_{i-1j}) + \frac{b_{2,ij}}{2h}(u_{ij+1} - u_{ij-1}) + c_{ij}u_{ij} \quad = \quad f_{ij} \end{aligned} \tag{2.9}$$

oder nach Umordnung

$$\left(4\frac{a_{ij}}{h^2} + c_{ij}\right)u_{ij} - \left(\frac{a_{ij}}{h^2} - \frac{b_{1,ij}}{2h}\right)(u_{i+1j} + u_{ij+1}) - \left(\frac{a_{ij}}{h^2} + \frac{b_{1,ij}}{2h}\right)(u_{i-1j} - u_{ij-1}) = f_{ij}. \tag{2.10}$$

Sammeln wir die Werte u_{ij} und f_{ij} wieder in einem Vektor U_h bzw. F_h, z.B. in der Form

$$(U_h)_{i+(j-1)n} = u_{ij}, \qquad (F_h)_{i+(j-1)n} = f_{ij}$$

und definieren eine geeignete Matrix K_h, so können wir das System wieder in der Standardform

$$K_h U_h = F_h \tag{2.11}$$

schreiben. Für $k = i + (j-1)n$ erhalten wir die Diagonalelemente

$$(K_h)_{kk} = 4\frac{a_{ij}}{h^2} + c_{ij}$$

und die Nebendiagonalelemente

$$\begin{aligned}(K_h)_{kk+1} &= -\frac{a_{ij}}{h^2} + \frac{b_{1,ij}}{2h},\\ (K_h)_{kk+n} &= -\frac{a_{ij}}{h^2} + \frac{b_{2,ij}}{2h},\\ (K_h)_{kk-1} &= -\frac{a_{ij}}{h^2} - \frac{b_{1,ij}}{2h},\\ (K_h)_{kk-n} &= -\frac{a_{ij}}{h^2} - \frac{b_{2,ij}}{2h}.\end{aligned}$$

Wir sehen sofort, dass das Hauptdiagonalelement positiv ist, und für h hinreichend klein sind die Nebendiagonalelemente negativ. Weiter ist die Matrix (schwach) diagonaldominant, d.h. es gilt

$$(K_h)_{ii} \geq \sum_{j\neq i} |(K_h)_{ij}|.$$

Mit dieser Eigenschaft können wir ein diskretes Maximumprinzip herleiten:

Proposition 2.11. *Sei $A \in \mathbb{R}^{N\times N}$ so, dass $A_{ij} \leq 0$ für $i \neq j$ und*

$$0 \neq A_{ii} \geq -\sum_{j\neq i} A_{ij}$$

gilt, und sei $x \in \mathbb{R}^N$ die Lösung von $Ax = b$ mit $b < 0$. Dann gilt $x \leq 0$.

Proof. Wir nehmen an $x_j > 0$. ist das Maximum von x. Dann gilt

$$A_{jj}x_j = b_j - \sum_{k\neq j} A_{jk}x_k < -\sum_{k\neq j} A_{jk}x_j \leq A_{jj}x_j,$$

und diese Ungleichungskette liefert einen direkten Widerspruch, da $A_{jj} \neq 0$ ist. □

Wir können wiederum die Aussage auf den Fall $b_j = 0$ erweitern:

Satz 2.12. *Sei $A \in \mathbb{R}^{N\times N}$ so, dass $A_{ij} \leq 0$ für $i \neq j$ und*

$$0 \neq A_{ii} \geq -\sum_{j\neq i} A_{ij}$$

gilt, A^{-1} existiert, und sei $x \in \mathbb{R}^N$ die Lösung von $Ax = b$ mit $b \leq 0$. Dann gilt $x \leq 0$.

Proof. Wir wenden Proposition 2.11 auf $x^\epsilon = A^{-1}b^\epsilon$ an, mit $b^\epsilon_j = b_j - \epsilon < 0$. Dann gilt $x^\epsilon \leq 0$ oder x^ϵ nimmt sein Maximum am Rand an. Da x^ϵ für $\epsilon \to 0$ gegen x konvergiert, und die Eigenschaft sich im Grenzwert nicht verändert, folgt die Aussage. □

Das Maximumprinzip hat eine interessante Eigenschaft der inversen Matrix $G_h = K_h^{-1}$ zur Folge, diese hat nämlich nur nichtnegative Einträge. Um dies zu sehen, verwenden wir nichtnegative Randwerte für die diskrete Lösung. Damit gilt für $U_h = K_h^{-1}F_h$ automatisch $U_h \leq 0$ falls $F_h \leq 0$. Wenden wir das Maximumprinzip speziell für die rechte Seite $F_h = (f_j) = (-\delta_{jk})$ an, dann folgt

$$0 \geq (U^h)_i = \sum_k (G_h)_{ij}(F_h)_j = -(G_h)_{ik}.$$

Da wir i und k beliebig wählen können, folgt die Nichtnegativität von G_h. Eine Matrix mit nichtpositiven Nebendiagonalelementen und einer nichtnegativen Inverse nennt man auch *M-Matrix* (siehe [3]). Das M steht dabei für die Monotonie, denn eine M-Matrix A hat die Eigenschaft, dass aus $f \geq g$ auch $M^{-1}f \geq M^{-1}g$ folgt (wie wir durch Anwendung von M^{-1} auf $g - f$ sofort sehen). D.h. die Ordnung der Vektoren bleibt unter Anwendung von M^{-1} erhalten.

Wir sehen aus der Definition von K_h, dass die Nebendiagonalelemente nur unter der Bedingung

$$2a_{ij} \geq \max\{|b_{1,ij}|, |b_{2,ij}|\}h, \qquad \forall\, i, j. \tag{2.12}$$

nichtpositiv sind. Dies kann im konvektionsdominanten Fall ein Problem sein, d.h. falls a_{ij} relativ klein ist im Vergleich zu b, weil man dann sehr feine Gitter verwenden müsste. Wie schon erwähnt ist es dann günstiger einen einseitigen Differenzenquotienten analog zu verwenden, um Stabilität zu erreichen (um den Preis einer niedrigeren Konsistenzordnung).

Analog zum kontinuierlichen Fall (deshalb dieses Mal ohne Beweis) erhalten wir aus der M-Matrix Eigenschaft eine diskrete Stabilität:

Korollar 2.13. *Sei K_h die Systemmatrix der Differenzendiskretisierung, F_h die rechte Seite, und U_h die Lösung von $K_hU_h = F_h$. Weiters sei* (2.12) *erfüllt. Dann existiert eine Konstante $\tilde{C}$ unabhängig von h, sodass die Abschätzung*

$$\max_j |(U_h)_j| \leq \tilde{C} \max_j (F_h)_j \leq \tilde{C}\,(\|f\|_\infty + \|g\|_\infty) \tag{2.13}$$

gilt.

2.3.4 Fehleranalyse

Mit den Resultaten der vorangegangenen Kapitel 1 ist es nun relativ einfach eine Fehleranalyse bzw. Fehlerabschätzungen herzuleiten. Wie schon oben allgemein diskutiert sind die wichtigsten Zutaten dabei die Konsistenz und Stabilität, wobei wir in diesem Fall nur die diskreten Varianten verwenden müssen.

Wir beginnen mit der Konsistenz für die Approximation des Differentialoperators

$$(Lu)(x) = -a(x)\Delta u(x) + \sum_{i=1}^{d} b_i(x)\frac{\partial u}{\partial x_i} + c(x)u, \qquad x \in \Omega. \tag{2.14}$$

durch den Differenzenoperator

$$\begin{aligned}(L_hu)_{ij} = \frac{a_{ij}}{h^2}\,(4u_{ij} - u_{ij+1} - u_{i+1j} - u_{ij-1} - u_{i-1j}) + \\ \frac{b_{1,ij}}{2h}\,(u_{i+1j} - u_{i-1j}) + \frac{b_{2,ij}}{2h}\,(u_{ij+1} - u_{ij-1}) + c_{ij}u_{ij}\end{aligned} \tag{2.15}$$

mit $u_{ij} = u(ih, jh)$.

Die Konsistenzordnung können wir direkt abschätzen:

Proposition 2.14. *Sei $\varphi \in C^4(\overline{\Omega})$, und L, L_h definiert durch* (2.14), (2.15). *Dann existiert eine Konstante $C > 0$, nur abhängig von φ, sodass*

$$|(L\varphi)(ih, jh) - (L_h\varphi)_{ij}| \leq Ch^2$$

gilt.

Proof. Analog zum eindimensionalen Fall in Kapitel können wir den Fehler der Ordnung h^2 beim zentralen Differenzenquotienten für die ersten und zweiten Ableitungen durch Taylor-Entwicklung abschätzen. □

Nun haben wir die Stabilität aus Korollar 2.13 und die Konsistenzordnung aus Proposition 2.14, in Kombination erhalten wir daraus eine Fehlerabschätzung:

Satz 2.15. *Sei $u \in C^4(\overline{\Omega})$ die Lösung der Differentialgleichung $Lu = f$ mit dem Operator L definiert in* (2.14)*. Weiters sei u^h die Lösung der Differenzengleichung $L_h u^h = f^h$, wobei $f^h_{ij} = f(ih, jh)$, und die Diskretisierung erfülle* (2.12)*. Dann gilt eine Fehlerabschätzung der Form*

$$\max_{i,j} |u(ih, jh) - u^h(ih, jh)| \leq Ch^2, \tag{2.16}$$

mit einer Konstante C unabhängig von h.

Proof. Sei $V = (u_{ij}) = (u(ih, jh))$, dann gilt

$$L_h(u - u^h) = L_h u - f^h = L_h u - (Lu)(ih, jh) =: r_h(ih, jh).$$

Aus Proposition 2.14 folgt

$$\max_{i,j} |r_h(ih, jh)| \leq C_1 h^2$$

mit einer Konstante C_1 nur abhängig von u, und aus Korollar 2.13 folgt

$$\max_{i,j} |u(ih, jh) - u^h(ih, jh)| \leq C_2 \max_{i,j} |r_h(ih, jh)|.$$

Die Kombination dieser beiden Abschätzungen liefert (2.16). □

Kapitel 3

Finite Elemente

In diesem Kapitel befassen wir uns genauer mit der Diskretisierung von partiellen Differentialgleichungen mit Finite Elemente (FE) Methoden. Der Fokus liegt dabei wieder auf elliptischen Gleichungen, da finite Elemente meist zur Ortsdiskretisierung verwendet werden. Die Erweiterung auf den ortsabhängigen Teil einer parabolischen ist dann z.B. im Rahmen einer horizontalen Linienmethode völlig klar, da ja dort in jedem Zeitschritt elliptische Probleme gelöst werden müssen. Zumindest teilweise lassen sich die hier vorgestellten Konzepte auch auf hyperbolische Probleme übertragen, eine genauere Diskussion dieser Probleme würde aber den Rahmen dieser Vorlesung sprengen.

3.1 Schwache Formulierung elliptischer Randwertprobleme

Wir beginnen mit der schwachen Formulierung elliptischer Randwertprobleme in Divergenzform und den dazugehörigen funktionalanalytischen Grundlagen. Der Einfachheit halber betrachten wir vor allem Gleichungen zweiter Ordnung, werden an einigen Stellen aber auch die Erweiterung auf höhere Ordnung diskutieren. Das Randwertproblem in einem Gebiet $\Omega \subset \mathbb{R}^d$ besteht aus der Gleichung

$$-\nabla \cdot (A\nabla u) + cu = f - \nabla \cdot h \qquad \text{in } \Omega \tag{3.1}$$

mit den Randbedingungen

$$u = g_D \qquad \text{auf } \Gamma_D \tag{3.2}$$

$$(A\nabla u) \cdot n = g_N \qquad \text{auf } \Gamma_N, \tag{3.3}$$

wobei $\partial\Omega = \Gamma_D \cup \Gamma_N$ gelten soll. In der obigen Formulierung nehmen wir wieder an, dass $A(x)$ für alle $x \in \Omega$ positiv definit mit minimalem Eigenwert grösser gleich $a_0 > 0$ ist, sowie dass die skalare Funktion c nichtnegativ ist.

Wir leiten zunächst die schwache Formulierung des Randwertproblems (3.1)-(3.3) her. Dazu multiplizieren wir (3.1) mit einer Testfunktion v und integrieren über Ω. Die beiden Divergenzterme auf der linken Seite können wir mit Hilfe des Gauss'schen Integralsatzes umformen und erhalten so

$$\int_\Omega ((A\nabla)u \cdot \nabla v + cuv)\ dx - \int_{\partial\Omega} (A\nabla u) \cdot nv\ d\sigma = \int_\Omega (fv + h \cdot \nabla v)\ dx - \int_{\partial\Omega} h \cdot nv\ d\sigma.$$

Nun schränken wir uns auf Testfunktionen ein, die auf Γ_D verschwinden, und setzen auf Γ_N die Randbedingung ein, woraus wir die Identität

$$\int_\Omega ((A\nabla)u \cdot \nabla v + cuv)\ dx = \int_\Omega (fv + h \cdot \nabla v)\ dx + \int_{\Gamma_N} (g_n - h \cdot n)v\ d\sigma. \tag{3.4}$$

erhalten. Wir sehen, dass wir es auf der linken Seite mit einer *Bilinearform*

$$B(u, v) := \int_\Omega ((A\nabla)u \cdot \nabla v + cuv)\ dx \tag{3.5}$$

und auf der rechten Seite mit einem *linearen Funktional* der Testfunktion,

$$\ell(v) := \int_\Omega (fv + h \cdot \nabla v)\ dx + \int_{\Gamma_N} (g_n - h \cdot n)v\ d\sigma \tag{3.6}$$

zu tun haben.

Funktionalanalytisch machen Bilinearformen und lineare Funktionale nur auf geeigneten Vektorräumen Sinn. Deshalb drängt sich sofort die Frage nach einer geeigneten Wahl von Räumen für die schwache Formulierung der Gleichung auf. Wie wir im folgenden sehen werden, führt dies in natürlicher Weise auf die sogenannten Sobolev-Räume.

3.1.1 Sobolev-Räume

Wir beginnen diesen Abschnitt mit einer kleinen funktionalanalytischen Erinnerung:

E1 Ein *normierter Raum* X ist ein Vektorraum (d.h. für $x, y \in X$ und $\alpha, \beta \in \mathbb{R}$ ist $\alpha x + \beta y \in X$) mit einer Norm, d.h. einer Abbildung $\|.\| : X \to \mathbb{R}_+$, sodass $\|x\| > 0$ für $x \neq 0$ und die Dreiecksungleichung $\|x + y\| \leq \|x\| + \|y\|$ gilt.

E2 Ein *Banachraum* X ist ein vollständiger normierter Raum, d.h. die Häufungspunkte jeder Folge $(x_n) \subset X$ liegen wieder in X.

E3 Ein *Hilbertraum* X ist ein Banachraum, dessen Norm von einem Skalarprodukt erzeugt wird, d.h. $\|x\| = \sqrt{\langle x, x\rangle}$. Das Skalarprodukt $\langle ., .\rangle : X \times X \to \mathbb{R}$ ist eine bilineare und symmetrische Abbildung, sodass $\langle x, x\rangle > 0$ für $x \neq 0$ gilt.

E4 Eine *Bilinearform* $B : X \times X \to \mathbb{R}$ ist eine in jeder Komponente lineare Abbildung. Ein Skalarprodukt ist ein Beispiel einer Bilinearform. Die Bilinearform B ist stetig, wenn $B(x, y) \leq C\|x\|\|y\|$ für alle $x, y \in X$ gilt, mit einer fixen Konstante C.

E5 Ein *lineares Funktional* $\ell : X \to \mathbb{R}$ ist eine lineare Abbildung nach $\mathbb{R}$. Das lineare Funktional heisst stetig, wenn $\ell(x) \leq C\|x\|$ für alle $x \in X$ gilt, mit einer fixen Konstante C.

E6 Der *Dualraum* X^* eines normierten Raumes X ist der Raum aller stetigen linearen Funktionale auf X. Mit

$$\|\ell\| = \sup_{x \in X \setminus \{0\}} \frac{|\ell(x)|}{\|x\|} = \sup_{x \in X, \|x\| \leq 1} |\ell(x)| = \inf\{\ C > 0 \mid \ell(x) \leq C\|x\|, \forall\ x \in X\ \}$$

ist X^* wieder ein normierter Raum. Falls X Banach-(Hilbert-)raum ist, dann ist auch X^* Banach-(Hilbert-)raum.

E7 In einem Hilbertraum gilt der *Riesz'sche Darstellungssatz:* Jedem linearen Funktional $\ell \in X^*$ kann ein eindeutiges Elements $x_\ell \in X$ zugeordnet werden, sodass

$$\ell(x) = \langle x_\ell, x\rangle \qquad \forall\ x \in X$$

gilt. Natürlich gilt auch die Umkehr, für jedes $y \in X$ ist $\ell_y(x) = \langle y, x\rangle$ ein lineares Funktional. Damit lässt sich (durch $\ell \leftrightarrow x_\ell$) der Raum X^* mit X identifizieren.

Die ersten unendlichdimensionalen Banach- und Hilberträume, die man üblicherweise betrachtet, sind die *Lebesgue-Räume* $L^p(\Omega)$. Für $1 \le p < \infty$ gilt

$$L^p(\Omega) = \{\ u : \Omega \to \overline{\mathbb{R}} \mid u \text{ messbar }, \int_\Omega |u|^p\ dx < \infty\ \},$$

wobei hier das Lebesgue-Integral verwendet wird. Die Norm in $L^p(\Omega)$ ist gegeben durch

$$\|u\|_p = \left(\int_\Omega |u|^p\ dx\right)^{1/p}.$$

Auf einem beschränkten Gebiet gilt klarerweise $L^p(\Omega) \subset L^q(\Omega)$ für $p \ge q$, auf unbeschränkten Gebieten gilt keine solche Inklusion.

Der Dualraum von $L^p(\Omega)$ $(1 < p < \infty)$ kann mit $L^{p_*}(\Omega)$ identifiziert werden, wobei $\frac{1}{p} + \frac{1}{p_*} = 1$ gilt. Die linearen Funktionale sind dann von der Form

$$\ell(u) = \int_\Omega uv\ dx$$

für ein $v \in L^{p_*}$. Mit der Hölder-Ungleichung überzeugt man sich leicht von der Stetigkeit solcher linearer Funktionale, es gilt ja

$$|\ell(u)| = |\int_\Omega uv\ dx| \le \left(\int_\Omega |u|^p\ dx\right)^{1/p} \left(\int_\Omega |v|^{p_*}\ dx\right)^{1/p_*} = \|u\|_p \|v\|_{p_*}.$$

Im Fall $p = 1$ erhält man aus der obigen Rechnung $p_* = \infty$. Der entsprechende Lebesgue-Raum ist

$$L^p(\Omega) = \{\ u : \Omega \to \overline{\mathbb{R}} \mid u \text{ messbar }, \operatorname*{ess\,sup}_x |u(x)|\ \},$$

wobei dass *essentielle Supremum* definiert ist als

$$\operatorname*{ess\,sup}_x |u(x)| = \inf\{\ N \mid N \text{ Nullmenge }\ \} \sup_x |u(x)|.$$

Die Norm in $L^\infty(\Omega)$ ist definiert durch

$$\|u\|_\infty = \operatorname*{ess\,sup}_x |u(x)|.$$

Die Umkehrung des Dualraums gilt aber nicht, der Dualraum von $L^\infty(\Omega)$ ist echt grösser als $L^1(\Omega)$.

Für partielle Differentialgleichungen benötigt man Verallgemeinerungen der Lebesgue-Räume in denen auch Ableitungen vorkommen. Dazu definiert man zunächst die *distributionelle Ableitung*, wieder über lineare Funktionale. Die distributionelle Ableitung ist im allgemeinen ein Funktional auf $C_0^\infty(\Omega)$, dem Raum der unendlich oft differenzierbaren Funktionen mit kompaktem Träger in Ω. Man identifiziert w mit der Ableitung $\frac{u}{x_j}$, wenn

$$w(\varphi) = -\int_\Omega u \frac{\partial \varphi}{\partial x_j}\, dx, \qquad \forall\ \varphi \in C_0^\infty(\Omega)$$

gilt. Diese Definition ist konsistent mit der klassischen Definition einer Ableitung, denn in diesem Fall kann man durch partielle Integration (und Ausnutzen der Tatsache, dass φ kompakten Träger hat, also am Rand verschwindet) zeigen, dass $w(\varphi) = \int_\Omega \varphi \frac{\partial u}{\partial x_j}\, dx$ gilt. Nun kann man testen, ob w ein stetiges lineares Funktional auf einem Lebesgue-Raum $L^{p*}(\Omega)$ ist. Wenn ja, dann können wir das lineare Funktional als Element des Dualraums mit einem $v \in L^p(\Omega)$ identifizieren und man spricht von $v = \frac{\partial u}{\partial x_j}$ als schwache Ableitung. Diese Überlegung ist auch die Basis für die Definition von Sobolevräumen. Man definiert mittels der distributionellen Ableitung

$$W^{1,p}(\Omega) = \{\ u \in L^p(\Omega) \mid \frac{\partial u}{\partial x_j} \in L^p(\Omega), j = 1, \ldots, d\ \}. \tag{3.7}$$

$W^{1,p}$ ist ein normierter Raum mit Norm

$$\|u\|_{1,p} := \left(\|u\|_p + \sum_{j=1}^{d} \|\frac{\partial u}{\partial x_j}\|_p \right)^{1/p}. \tag{3.8}$$

Man verwendet üblicherweise die Notation

$$|u|_{1,p} := \left(\sum_{j=1}^{d} \|\frac{\partial u}{\partial x_j}\|_p \right)^{1/p}. \tag{3.9}$$

für die Halbnorm betreffend die ersten Ableitungen ($|u|_{1,p} = 0$ für u konstant). Die Konvergenz einer Folge $u_n \to u$ in der Norm von $W^{1,p}(\Omega)$ impliziert die Konvergenz $u_n \to u$ in $L^p(\Omega)$ und $\frac{\partial u}{\partial x_j} \to v_j$ in $L^p(\Omega)$. Weiters gilt für glatte Testfunktionen

$$-\int_\Omega u \frac{\partial \varphi}{\partial x_j}\, dx = \lim \left(-\int_\Omega u_n \frac{\partial \varphi}{\partial x_j}\, dx \right) = \lim \left(\int_\Omega \varphi \frac{\partial u_n}{\partial x_j}\, dx \right) = \int_\Omega \varphi v_j\, dx$$

und somit folgt $v_j = \frac{\partial u}{\partial x_j}$ im obigen Sinne. Damit gilt aber auch $u \in W^{1,p}(\Omega)$. Also ist $W^{1,p}(\Omega)$ vollständig und damit ein Banachraum.

Für $p = 2$ erhält man wie im Falle der Lebesgue-Räume sogar einen Hilbert-Raum, üblicherweise mit der Bezeichnung $H^1(\Omega) := W^{1,2}(\Omega)$. Das Skalarprodukt in $H^1(\Omega)$ ist gegeben durch

$$\langle u, v \rangle_1 := \int_\Omega \left(uv + \sum_{j=1}^{d} \frac{\partial u}{\partial x_j} \frac{\partial v}{\partial x_j} \right)\, dx,$$

also die Summe aus L^2-Skalarprodukten der Funktion und ihrer Ableitungen. Wie wir sehen werden, ist der Hilbert-Raum $H^1(\Omega)$ der wichtigste bei der Analysis linearer Gleichungen. Sobolevräume für $p \neq 2$ verwendet man meist bei nichtlinearen Gleichungen.

Durch Iteration der obigen Definitionen kann man analog höherere distributionelle Ableitungen als

$$w(\varphi) = (-1)^{|\alpha|} \int_\Omega u \frac{\partial^{|\alpha|}\varphi}{\partial x^\alpha}\, dx, \qquad \forall\ \varphi \in C_0^\infty(\Omega)$$

und Sobolevräume höherer Ordnung als

$$W^{k,p}(\Omega) = \{\ u \in L^p(\Omega \mid \frac{\partial^{|\alpha|}u}{\partial x^\alpha} \in L^p(\Omega),\ |\alpha| \leq k\ \}$$

definieren. Wiederum erhält man damit Banachräume mit der Norm

$$\|u\|_{k,p} := \left(\sum_{|\alpha|\leq k} \|\frac{\partial^{|\alpha|}u}{\partial x^\alpha}\|_p \right)^{1/p}$$

und definiert die Halbnorm

$$|u|_{k,p} := \left(\sum_{|\alpha|=k} \|\frac{\partial^{|\alpha|}u}{\partial x^\alpha}\|_p \right)^{1/p},$$

die auf Polynomen vom Grad kleiner oder gleich $k-1$ verschwindet. Im Fall $p = 2$ ist $H^k(\Omega) := W^{k,2}(\Omega)$ ein Hilbertraum mit dem Skalarprodukt

$$\langle u, v\rangle_k = \sum_{|\alpha|\leq k} \int_\Omega \frac{\partial^{|\alpha|}u}{\partial x^\alpha} \frac{\partial^{|\alpha|}v}{\partial x^\alpha}\, dx.$$

Die schwache Formulierung partieller Differentialgleichungen höherer als zweiter Ordnung führt in natürlicher Weise zu einem Sobolevraum mit $k > 1$.

Der Dualraum von $H^1(\Omega)$ wird üblicherweise mit $H^{-1}(\Omega)$ bezeichnet. Per Definition gilt ja die Einbettung $H^1(\Omega) \hookrightarrow L^2(\Omega)$. Damit definiert jedes Element $\psi \in L^2(\Omega)$ ein stetiges lineares Funktional $u \mapsto \int_\Omega u\psi\, dx$ auf $H^1(\Omega)$. Also folgt $H^{-1}(\Omega) \subset L^2(\Omega) \subset H^1(\Omega)$. Umgekehrt ist $H^1(\Omega)$ aber auch Hilbertraum und somit kann der Dualraum $H^{-1}(\Omega)$ wiederum mit $H^1(\Omega)$ identifiziert werden. Für ein lineares Funktional $w \in H^{-1}(\Omega)$ bedeutet diese Identifizierung ein Element $v \in H^1(\Omega)$ zu finden, sodass

$$\langle v, \varphi\rangle = w(\varphi) \qquad \forall\ \varphi \in H^1(\Omega)$$

gilt. Schreiben wir das Skalarprodukt aus, dann erhalten wir die Variationsgleichung

$$\int_\Omega (v\varphi + \nabla v \cdot \nabla\varphi)\, dx = w(\varphi) \qquad \forall\ \varphi \in H^1(\Omega).$$

Man erkennt, dass die Bilinearform auf der rechten Seite genau der schwachen Formulierung des Differentialoperators $-\Delta v + v$ entspricht. D.h., durch Anwendung des Riesz'schen Darstellungssatzes in $H^1(\Omega)$ lösen wir eigentlich eine elliptische Differentialgleichung zweiter

Ordnung mit rechter Seite in $H^{-1}(\Omega)$. Diese Sichtweise werden wir im nächsten Kapitel auf die Analysis schwacher Lösungen allgemeinerer Gleichungen übertragen.

Neben der Differentialgleichung in Ω haben wir immer auch Randbedingungen benötigt. Da Funktionen in Sobolev-Räumen über Lebesgue-Räume definiert werden, stellt sich sofort die Frage, ob bzw. in welchem Sinn die Auswertung von Funktionen aus $H^1(\Omega)$ definiert werden kann (der Rand ist ja eine Nullmenge). Dies geschieht durch sogenannte *Spursätze*, die jeder Funktion in einem Sobolevraum einen distributionellen Randwert zuordnen, obwohl die Punktauswertung am Rand keinen Sinn ergibt. Für glatte Funktionen u, ψ und φ mit $\frac{\partial\varphi}{\partial n} = \psi$ auf $\partial\Omega$ gilt ja nach dem Gauss'schen Integralsatz

$$\int_{\partial\Omega} u\psi \; d\sigma = \int_{\partial\Omega} u\frac{\partial\varphi}{\partial n} \; d\sigma = \int_\Omega \nabla\cdot(u\nabla\varphi) \; dx = \int_\Omega (u\Delta\varphi + \nabla u\cdot\nabla\varphi) \; dx.$$

Die rechte Seite kann auch sinnvoll auf Funktionen in $H^1(\Omega)$ erweitert werden, und damit erhält man sofort auch die distributionelle Definition eines Randwerts auf $\partial\Omega$, die sogenannte Spur (die man implizit auch immer mit dem Randwert gleichsetzt). Man kann zeigen, dass die Spur einer Funktion in $H^1(\Omega)$ immer einer Funktion in $L^2(\partial\Omega)$ entspricht. Es gilt sogar mehr, die Spur ist ein Element des (fraktionalen) Sobolevraums $H^{1/2}(\partial\Omega) \hookrightarrow L^2(\partial\Omega)$. Wie der Exponent $1/2$ schon nahelegt, ist $H^{1/2}(D)$ immer ein Zwischenraum zwischen $L^2(D)$ und $H^1(D)$, es gilt $L^2(D) \hookrightarrow H^{1/2}(D) \hookrightarrow H^1(D)$. Im Rahmen der Interpolationstheorie von Hilbert-Räumen (cf. [4]) kann man $H^{1/2}(D)$ tatsächlich als Interpolation von $L^2(D)$ und $H^1(D)$ betrachten. Wir gehen hier nicht weiter auf dieses fortgeschrittene Thema ein, erwähnen aber die Interpolationsgleichung

$$\|u\|_{H^{1/2}} = \sqrt{\|u\|_{L^2}\|u\|_{H^1}}.$$

Zum Abschluss dieser Diskussion liefern wir noch eine genaue Formulierung des Spursatzes:

Satz 3.1 (Spursatz). *Sei $\Gamma \subset \partial\Omega$ ein Teil des Randes mit positivem Maß und $\partial\Omega$ stückweise Lipschitz. Dann existiert ein stetiger linearer Operator $T : H^1(\Omega) \to H^{1/2}(\Gamma)$, sodass*

$$Tu = u|_\Gamma \qquad \forall u \in C(\overline{\Omega})$$

gilt. Der Operator T ist surjektiv, d.h. für jedes $v \in H^{1/2}(\Gamma)$ existiert ein $u \in H^1(\Omega)$ mit $Tu = v$.

Der Spursatz besagt, dass es eine Erweiterung der Randwerte von stetigen Funktionen auf Funktionen in $H^1(\Omega)$ gibt, und der Raum $H^{1/2}(\Gamma)$ genau aus diesen Randwerten besteht. Eine Erweiterung der Spur ist auch auf Sobolevräume $H^k(\Omega)$ möglich, es ist dann $T : H^k(\Omega) \to H^{k-1/2}(\Gamma)$ ein stetiger surjektiver Operator. Grob gesagt verliert man also beim Auswerten der Spur immer eine halbe Differentiationsordnung. Zur einfacheren Notation werden wir im Folgenden immer u für die Randwerte schreiben, damit aber eigentlich die Spur Tu assoziieren.

Neben den Randwerten von u (Dirichlet-Werte) haben wir auch die Normalableitungen $\frac{\partial u}{\partial n}$ (Neumann-Werte) in den Randbedingungen verwendet. Da nun eine Ableitung mehr auftritt, könnte man vermuten, dass $\frac{\partial u}{\partial n} \in H^{1/2-1}(\partial\Omega)$ definierbar ist. Dies ist auch tatsächlich der Fall, wenn man $H^{-1/2}(\Gamma)$ als Dualraum von $H^{1/2}(\Gamma)$ definiert (analog zur Definition von $H^{-1}(\Omega)$). Wir haben ja für glatte Funktionen nach dem Gauss'schen Integralsatz

$$\int_{\partial\Omega} \frac{\partial u}{\partial n} v \; dx = \int_\Omega (\nabla u\cdot\nabla v + \Delta uv) \; dx = \int_\Omega (\nabla u\cdot\nabla v + uv) \; dx - \int_\Omega (-\Delta u + u)v \; dx.$$

Die rechte Seite ist zunächst ein Skalarprodukt auf $H^1(\Omega)$, also auf Funktionen dieser Klasse auch sinnvoll definiert, und im zweiten Term kann man $-\Delta u + u \in H^{-1}(\Omega)$ wieder zu einem stetigen linearen Funktional auf $H^1(\Omega)$ erweitern. In dieser Sichtweise definiert also $\frac{\partial u}{\partial n}$ ein stetiges lineares Funktional auf den Spuren von Funktionen in $H^1(\Omega)$, nach dem Spursatz also genau in $H^{-1/2}(\Omega)$. Die Neumann-Randbedingung kann also in diesem Sobolev-Raum verstanden werden, und somit sollte man für Neumann-Daten auch $g_N \in H^{-1/2}(\Gamma_N)$ fordern.

Ein anderes interessantes Problem ist die Einbettung von Sobolev-Räumen in Räume stetiger Funktionen oder Hölder-Räume. Im eindimensionalen Fall haben wir ja gesehen, dass Funktionen in $H^1(\Omega)$ immer auch stetig sind, und sich die Supremumsnorm durch ein Vielfaches der H^1-Norm abschätzen lässt. In so einem Fall spricht man von einer Einbettung des Raumes $H^1(\Omega)$ in $C(\Omega)$. Solche Einbettungen sind aber immer dimensionsabhängig, was an einfachen Beispiel deutlich wird. Betrachten wir z.B. Funktionen der Form $u(x) = |x|^\alpha$ auf der Einheitskugel in $\mathbb{R}^d$ und mit $\alpha \in \mathbb{R}$. Dann gilt (in Polarkoordinaten)

$$\int_\Omega u(x)^2 \, dx = \int_\Omega |x|^{2\alpha} dx = S_d \int_0^1 r^{2\alpha} r^{d-1} dr,$$

wobei S_d die Oberfläche der Einheitssphäre in $\mathbb{R}^d$ ist. Da man Funktionen der Form r^β genau für $\beta > -1$ integrieren kann, folgt $u \in L^2(\Omega)$ für $\alpha > -\frac{d}{2}$. Der Gradient von u ist gegeben durch $\nabla u = \alpha x |x|^{\alpha-2}$ und es gilt

$$\int_\Omega |\nabla u|^2 \, dx = 2\alpha \int_\Omega |x|^{2\alpha-2} dx = \int_0^1 r^{2\alpha+d-3} dr.$$

Damit ist der Gradient integrierbar für $\alpha > \frac{1-d}{2}$. In Raumdimension 1 folgt damit $\alpha > 0$, was mit unserem vorigen Resultat über die Stetigkeit von H^1-Funktionen auch übereinstimmt. Für Raumdimensionen $d > 1$ ist auch $\alpha < 0$ möglich, und da $u(x)$ in diesem Fall in $x = 0$ unstetig ist, gibt es auch unstetige (sogar unbeschränkte) Elemente von $H^1(\Omega)$. Um eine Einbettung in $C(\Omega)$ zu erreichen, muss man dann zu einem $W^{1,p}(\Omega)$ mit höherem p oder zu einem $H^k(\Omega)$ mit höherem k übergehen. Allgemein gilt folgendes Resultat:

Satz 3.2 (Einbettungssatz für Sobolevräume). *Sei $\Omega \subset \mathbb{R}^d$ ein glattes, beschränktes Gebiet, und $kp > d$. Dann gilt die Einbettung $W^{k,p}(\Omega) \subset C(\overline{\Omega})$. Für $(k-j)p > n$ gilt weiters die Einbettung $W^{k,p}(\Omega) \subset C^j(\overline{\Omega})$.*

Der Einbettungssatz gibt nicht nur Auskunft über die Stetigkeit von Funktionen in Sobolevräumen, sondern auch über die Stetigkeit von Ableitungen. Man kann sich für eine Differentialgleichung zweiter (oder höherer) Ordnung anhand der jeweiligen Raumdimension ausrechnen, wie gross k bzw. p sein müssten, um aus einer Lösung in einem Sobolevraum durch Einbettung wieder eine stetige Lösung zu erhalten.

Aus den obigen Funktionen der Form $u(x) = |x|^\alpha$ sehen wir auch, dass Funktionen in Sobolevräume auch in Lebesgue-Räumen mit höherem Exponenten liegen. Es gilt z.B. $u \in H^1(\Omega)$ für $\alpha > \frac{1-d}{2}$. Umgekehrt ist $u \in L^p(\Omega)$ für $p\alpha + d > 0$. Setzen wir die untere Schranke für α ein, so folgt die Bedingung $p\alpha + d > 0$ sicher aus $p(1-d) + 2d > 0$, oder äquivalent $p < \frac{2d}{d-1}$. Man beachte, dass $\frac{2d}{d-1} = 2 + \frac{2}{d-1} > 2$ gilt, in Raumdimension 1 gelangt man sogar bis zu $p = \infty$. Allgemein gilt folgendes Resultat:

Satz 3.3 (Einbettung von Sobolev- in Lebesgueräume). *Sei $\Omega \subset \mathbb{R}^d$ ein beschränktes Gebiet, und $q \leq q_* = \frac{dp}{d-pk}$. Dann gilt die Einbettung $W^{k,p}(\Omega) \subset L^q(\Omega)$, diese ist sogar kompakt für $q < q_*$.*

Einbettung von Sobolev in Lebesgue-Räume lässt sich direkt (durch Einbettung der entsprechenden Ableitung) zur Einbettung von Sobolevräume $W^{k,p}(\Omega)$ in $W^{m,q}(\Omega)$ mit $m < k$ verallgemeinern. Die einfachste (und unten auch verwendete) Folgerung ist die kompakte Einbettung $W^{k,p}(\Omega) \hookrightarrow W^{m,q}(\Omega)$.

In manchen Fällen ist es wichtig äquivalente Normen zur obigen Sobolevraum-Norm zu verwenden, in dem man die Halbnorm mit anderen Halbnormen als der L^p-Norm der Funktion kombiniert. Solche Normen sind von der Form

$$|||u||| = \left(|u|_{k,p}^p + \sum_{j=1}^{m} f_i(u)^p \right)^{1/p}, \tag{3.10}$$

wobei f_i ein System von Halbnormen ist. Beispiele dafür sind

$$|||u||| = \left(|u|_{1,p}^p + |\int_\Omega u \, dx|^p \right)^{1/p},$$

oder

$$|||u||| = \left(|u|_{k,p}^p + \int_{\Gamma_d} |u(x)|^p \, dx \right)^{1/p}.$$

In diesen Fällen gilt der *Normierungssatz von Sobolev:*

Satz 3.4. *Sei $f_i : W^{k,p}(\Omega) \to \mathbb{R}$ ein System von Halbnormen, sodass*

$$0 \le f_i(u) \le C_i \|u\|_{k,p} \qquad \forall\ u \in W^{k,p}(\Omega)$$

gilt. Weiters soll für jedes Polynom $v \neq 0$ vom Grad kleiner oder gleich $k-1$ eine Halbnorm f_i existieren, sodass $f_i(v) \neq 0$. Dann sind die Normen $\|\|_{k,p}$ und $|||.|||$ wie in (3.10) äquivalent, d.h. es existieren Konstante $\alpha, \beta > 0$, sodass

$$\alpha |||u||| \le \|u\|_{k,p} \le \beta |||u||| \qquad \forall\ u \in W^{k,p}(\Omega).$$

Proof. Es gilt nach den obigen Voraussetzungen an f_i:

$$|||u|||^p = |u|_{k,p} + \sum f_i(u)^p \le \|u\|_{k,p} + \sum C_i^p \|u\|_{k,p}^p \le (1 + \sum C_i^p) \|u\|_{k,p}^p$$

und mit $\alpha := (1 + \sum C_i^p)^{-1/p}$ folgt die erste Ungleichung.

Die zweite Ungleichung beweisen wir indirekt durch Widerspruch. Wir nehmen an, es existiert keine solche Konstante β. Dann existiert eine Folge $u_n \in W^{k,p}(\Omega)$ sodass

$$\|u_n\|_{k,p} > n |||u_n|||.$$

Wir definieren nun $v_n := \frac{u_n}{\|u_n\|_{k,p}}$. Dann gilt $\|v_n\|_{k,p} = 1$ und $|||v_n||| < \frac{1}{n}$. Insbesonders folgt $|v_n|_{k,p} < \frac{1}{n}$. Da v_n uniform beschränkt in $W^{k,p}$ ist, existiert (wegen der Kompaktheit der Einbettung) eine konvergente Teilfolge $v_{n'}$ in $W^{k-1,p}(\Omega)$, deren Grenzwert wir mit v bezeichnen. Da aber die Ableitungen der Ordnung k gegen Null konvergieren ($|v_{n'}|_{k,p} < \frac{1}{n'}$) muss für den Grenzwert auch $|v|_{k,p} = 0$ gelten, d.h. alle Ableitungen der Ordnung k verschwinden. Damit ist v ein Polynom mit Grad kleiner gleich $k-1$. Wegen $f_i(v_{n'}) < \frac{1}{n'}$ folgt $f_i(v_{n'}) \to 0$. Es gilt aber auch wegen Dreiecksungleichung und Annahmen an f_i

$$|f_i(v_{n'}) - f_i(v)| \le f_i(v_{n'} - v) \le C_i \|v_{n'} - v\|_{k,p} \to 0,$$

und damit muss $f_i(v) = 0$ gelten für alle i. Aus den Annahmen über die f_i folgt dann aber sofort $v = 0$. Dies ist aber ein Widerspruch wegen

$$0 = \|v\|_{k,p} = \lim \|v_{n'}\|_{k,p} = 1.$$

□

Konsequenzen aus dem Normierungssatz von Sobolev sind unter anderem die Poincare-Ungleichung und die Friedrichs-Ungleichung. Die Poincare-Ungleichung ergibt sich für $k = 1$ mit der Seminorm $f_1(u) = |\int_\Omega u \, dx|$, d.h. es gilt

$$\|u\|_{1,p} \leq C \left(|\int_\Omega u \, dx|^p + |u|_{1,p}^p \right)^{1/p}.$$

Insbesonders folgt für Funktionen mit Mittelwert Null die Abschätzung

$$\|u\|_{1,p} \leq C|u|_{1,p}^p.$$

Friedrichs-Ungleichungen erhält man für $k = 1$ mit der Wahl $f_1(u)^p = \int_\Gamma |u|^p \, d\sigma$, und es folgt

$$\|u\|_{1,p} \leq C \left(\int_\Gamma |u|^p \, d\sigma + |u|_{1,p}^p \right)^{1/p}.$$

Für Funktionen mit homogenen Dirichlet-Randwerten auf Γ_D lässt sich wiederum die H^1-Norm durch die Halbnorm abschätzen. Diese Funktionen fasst man meist in einem eigenen Unterraum zusammen, mit der Notation

$$H_0^1(\Omega) := \{ \, u \in H^1(\Omega) \mid u = 0 \text{ auf } \Gamma_D \, \},$$

und verwendet darin wiederum das Skalarprodukt von $H^1(\Omega)$. Da die Spur stetig ist, kann eine Folge in $H_0^1(\Omega)$ nur Häufungspunkte mit verschwindender Spur auf Γ_D haben. Folglich ist $H_0^1(\Omega)$ abgeschlossen, also selbst ein Hilbertraum.

3.1.2 Schwache Lösungen

Aus der obigen Theorie der Sobolevräume können wir nun die schwache Formulierung in Funktionenräumen angeben. Dazu suchen wir uns zunächst eine beliebige Funktion in $H^1(\Omega)$ mit der Spur g_D auf Γ_D (so eine Funktion existiert nach dem Spursatz für $g_D \in H^{1/2}(\Gamma_D)$), die wir wieder mit g_D bezeichnen. Dann können wir eine Lösung $u \in g_D + H_0^1(\Omega)$ suchen, sodass

$$B(u, v) = \ell(v) \qquad \forall \, v \in H_0^1(\Omega) \tag{3.11}$$

gilt. Wir sehen sofort, dass es genügt $A \in L^\infty(\Omega; \mathbb{R}^{d \times d})$ und $c \in L^\infty(\Omega)$ zu fordern, um Stetigkeit der Bilinearform B zu erhalten. Es gilt ja

$$B(u, v) = |\int_\Omega (A \nabla u \nabla v + cuv) \, dx| \leq \max\{\|A\|_\infty, \|c\|_\infty\} \int_\Omega (|\nabla u \cdot \nabla v| + |uv|) \, dx$$

und das Integral können wir mit der Cauchy-Schwarz Ungleichung durch $\|u\|_{1,2}\|v\|_{1,2}$ abschätzen. Die Anforderung an c könnte sogar noch abgeschwächt werden auf $c \in L^q(\Omega)$ für dimensionsabhängiges $q < \infty$, da man ja durch Einbettungssätze $u, v \in L^p(\Omega)$ mit $p > 2$ erhält (und die entsprechenden Normen können durch die H^1-Norm abgeschätzt werden).

Unter den üblichen Elliptizitätsannahmen $A(x) \geq \alpha I$ und $c \geq 0$ kann man die *Koerzivität der Bilinearform* zeigen. Dazu zuerst die passende Definition:

Definition 3.5. Eine Bilinearform $B : X \times X \to \mathbb{R}$ auf einem Hilbertraum X heisst koerziv, wenn eine Konstante $\gamma > 0$ existiert, sodass

$$B(u,u) \geq \gamma \|u\|^2 \qquad \forall\ u \in X$$

gilt.

Wir sehen, dass

$$B(u,u) = \int_\Omega (A\nabla u \cdot \nabla u + cu^2)\ dx \geq \alpha |u|_{1,2} + \int_\Omega cu^2\ dx.$$

Gilt $c(x) \geq c_0$ auf einer Menge von positivem Maß, dann sehen wir sofort, dass $f_1(u) = \sqrt{\int_\Omega cu^2\ dx}$ die Bedingungen des Sobolev'schen Normierungssatzes erfüllt, und damit können wir Koerzivität mit einer Konstante γ abhängig von c und α zeigen.

Im Fall $c \equiv 0$ kann man immer noch Koerzivität erhalten, nämlich durch die Randbedingung. Ist Γ_D eine Menge von positivem Maß, so können wir mit $\tilde{u} = u - g_D$ das Problem als Variationsgleichung für $\tilde{u}$ in $H^1_0(\Omega)$ schreiben (mit gleicher Bilinearform und geänderter rechter Seite). Dann verwenden wir einfach die Friedrichs-Ungleichung, die auf $H^1_0(\Omega)$ impliziert, dass

$$|u|_{1,2} \geq C\|u\|_{1,2}$$

gilt, da ja das Randintegral verschwindet.

Um auf der rechten Seite ein stetiges lineares Funktional zu erhalten, würde es genügen, dass $h \in L^2(\Omega;\mathbb{R}^d)$, $g_N \in H^{-1/2}(\Gamma_N)$, und $f \in L^r(\Omega)$ gilt, mit dimensionsabhängigem $r < 2$ (wieder erhalten aus passenden Einbettungssätzen).

Wir finden nun folgende Situation: wegen der Stetigkeit und Koerzivität gilt

$$c\|u\|^2 \leq B(u,u) \leq C\|u\|^2,$$

d.h. B definiert ein äquivalentes Skalarprodukt auf $H^1(\Omega)$. In $H^1(\Omega)$ mit dem neuen Skalarprodukt B gibt es nach dem Riesz'schen Darstellungssatz wieder ein eindeutiges Element $u \in H^1(\Omega)$, dass das lineare Funktional darstellt, d.h. eine schwache Lösung $u \in H^1(\Omega)$ von (3.11).

Damit haben wir direkt die Existenz und Eindeutigkeit einer schwachen Lösung gezeigt, allerdings ist dieser Beweis über den Riesz'schen Darstellungssatz nicht konstruktiv. Deshalb geben wir noch einen zweiten Beweis an, der auch gleichzeitig die Konvergenz einer einfachen iterativen Methode zur Lösung der Variationsgleichung liefert.

Satz 3.6 (Lax-Milgram Lemma). *Sei $B : X \times X \to \mathbb{R}$ eine symmetrische, stetige und koerzive Bilinearform, d.h. es existieren Konstante $c, C \in \mathbb{R}^+$, sodass*

$$B(u,v) \leq C\|u\|\|v\| \qquad \forall\ u,v \in X \tag{3.12}$$

und

$$B(u,u) \geq c\|u\|^2 \qquad \forall\ u \in X. \tag{3.13}$$

Dann existiert für jedes $\ell \in X^*$ eine eindeutige Lösung $\hat{u} \in X$ der Variationsgleichung

$$B(u,v) = \ell(v) \qquad \forall\ v \in X. \tag{3.14}$$

Die Fixpunktiteration (Richardson-Verfahren)

$$\langle u^{k+1}, v\rangle = \langle u^k, v\rangle - \tau(B(u^k,v) - \ell(v)) \qquad \forall\ v \in X \tag{3.15}$$

liefert eine konvergente Folge $u^{k+1} \to \hat{u}$ für $\tau < \frac{1}{C}$.

Proof. Wir betrachten den Fixpunktoperator $A : X \to X$, in schwacher Form definiert durch

$$\langle Au, v\rangle := \langle u, v\rangle - \tau B(u, v) \qquad \forall\ v \in X.$$

Weiters erhalten wir eine rechte Seite $f \in X$, definiert durch

$$\langle f, v\rangle = \tau \ell(v) \qquad \forall\ v \in X.$$

Dann können wir die Fixpunktiteration auch als

$$u^{k+1} = Au^k - f$$

schreiben, und erhalten sofort die Lösung

$$u^k = A^k u^0 - \sum_{j=0}^{k-1} A^j f.$$

Für $\|A\| < 1$ konvergiert diese Neumann'sche Reihe gegen

$$\hat{u} = -\sum_{j=0}^{\infty} A^j f = (A - I)^{-1} f,$$

also folgt $(A - I)\hat{u} = f$, bzw. in schwacher Form

$$\langle \hat{u}, v\rangle - \tau B(\hat{u}, v) - \langle \hat{u}, v\rangle = -\tau \ell(v) \qquad \forall\ v \in X,$$

was äquivalent zu (3.14) ist.

Wir müssen also nur mehr die Operatornorm von A abschätzen. Dazu verwenden wir zunächst, dass A ein selbst-adjungierter Operator ist, d.h. es gilt

$$\langle Au, v\rangle = \langle Av, u\rangle \qquad \forall\ u, v \in X,$$

was man wegen der Symmetrie von B sofort sieht. Wegen der Stetigkeit der Bilinearform ist A beschränkt, und aus $\tau C < 1$ folgt

$$\langle Au, u\rangle = \langle u, u\rangle - \tau B(u, u) \geq \|u\|^2 (1 - \tau C) > 0.$$

Es gilt dann auch für $u, v \in X$, dass

$$0 \leq \langle A(u - v), A(u - v)\rangle = \langle Au, u\rangle + \langle Av, v\rangle - 2\langle Au, v\rangle.$$

Also folgt

$$\|A\| = \sup_{\|u\|\leq 1, \|v\|\leq 1} \langle Au, v\rangle \leq \frac{1}{2} \left(\sup_{\|u\|\leq 1} \langle Au, u\rangle + \sup_{\|v\|\leq 1} \langle Av, v\rangle \right) = \sup_{\|u\|\leq 1} \langle Au, u\rangle.$$

Wegen der Koerzivität folgt aber

$$\langle Au, u\rangle = \langle u, u\rangle - \tau B(u, u) \leq \|u\|^2 - \tau c \|u\|^2,$$

und damit $\|A\| \leq 1 - \tau c < 1$. □

Wir haben nun Existenz, Eindeutigkeit einer Lösung sowie bereits die Konvergenz einer iterativen Methode nachgewiesen, es fehlt nur noch die Stabilität der Lösung. Diese erhalten wir sehr direkt aus der Koerzivität:

Satz 3.7. *Es gelten die Voraussetzungen von Satz 3.6. Dann gilt die Stabilitätsabschätzung*

$$\|\hat{u}\| \leq \frac{1}{c}\|\ell\| \tag{3.16}$$

Proof. Sei $\hat{u} \in X$ die Lösung von (3.14). Dann erhalten wir mit der Testfunktion $v = \hat{u}$:

$$c\|\hat{u}\| \leq B(\hat{u}, \hat{u}) = \ell(\hat{u}) \leq \|\ell\|\|\hat{u}\|,$$

was direkt (3.16) impliziert. □

Wir sehen, dass die Stabilitätskonstante indirekt proportional zur Koerzivitätskonstante c und proportional zur Norm der rechten Seite ist. Die Konstante C aus der Stetigkeitsbedingung kommt hier nicht vor, sie ist aber in der Norm der rechten Seite versteckt, da man $\|\ell\|$ immer als relative Grösse zu B sehen sollte. Skaliert man die Bilinearform (durch Division durch C) so, dass die Norm von B eins ergibt, dann ist die effektive Koerzivitätskonstante $\frac{C}{c}$, diese Grösse ist analog zur Konditionszahl einer Matrix.

Da wir die Bedingungen des Lax-Milgram Lemmas für die schwache Formulierung von (3.1)-(3.3) oben unter vernünftigen Bedingungen an die Daten nachgeprüft haben, erhalten wir nun sofort die Existenz, Eindeutigkeit, und Stabilität für die elliptische Differentialgleichung zweiter Ordnung.

3.1.3 Variationsprinzip

Wir werden nun noch kurz ein Variationsprinzip diskutieren, das die Variationsgleichung mit einer Energieminimierung in Beziehung setzt. Dazu definieren wir die (quadratische) Energie

$$E(u) := \frac{1}{2}B(u,u) - \ell(u). \tag{3.17}$$

Nun berechnen wir die Richtungsableitungen des Energiefunktionals in eine beliebige Richtung $v \in X$, d.h.

$$E'(u)v = \lim_{\epsilon \downarrow 0} \frac{1}{\epsilon}\left(E(u+\epsilon v) - E(u)\right)$$

Wegen der Bilinearität von B und der Linearität von ℓ folgt

$$\begin{aligned} E(u+\epsilon v) &= \frac{1}{2}B(u+\epsilon v, u+\epsilon v) + \ell(u+\epsilon v) \\ &= \frac{1}{2}B(u,u) + \epsilon B(u,v) + \frac{\epsilon^2}{2}B(v,v) + \ell(u) + \epsilon\ell(v) \\ &= E(u) + \epsilon(B(u,v) - \ell(v)) + \frac{\epsilon^2}{2}B(v,v) \end{aligned}$$

und im Grenzwert erhalten wir

$$E'(u)v = B(u,v) - \ell(v).$$

Betrachten wir die Optimalitätsbedingung erster Ordnung für ein Minimum $E'(u)v = 0$ für alle $v \in X$, so erkennen wir wieder die Variationsgleichung, durch die wir also einen stationären Punkt des Energiefunktionals berechnen. Nun betrachten wir noch die zweite Ableitung, also

$$E''(u)(v,w) = \lim_{\epsilon \downarrow 0} \frac{1}{\epsilon} \left(E'(u+\epsilon w)v - E'(u)v \right).$$

Es gilt

$$E'(u+\epsilon w)v = B(u+\epsilon w, v) - \ell(v) = B(u,v) + \ell(v) + \epsilon B(w,v) = E'(u)v + \epsilon B(v,w).$$

Also erhalten wir $E''(u)(v,w) = B(v,w)$, und damit gilt wegen der Koerzivität

$$E''(u)(v,v) = B(v,v) \geq c\|v\|^2 > 0.$$

Damit erwarten wir natürlich, dass E konvex und die Lösung der Variationsgleichung damit ein Minimum des Energiefunktionals ist. Dies beweisen wir im nächsten Satz:

Satz 3.8. *Die Lösung $\hat{u} \in X$ der Variationsgleichung* (3.14) *ist das eindeutige Minimum des Energiefunktionals E in X und umgekehrt.*

Proof. Sei $\hat{u}$ die eindeutige Lösung von (3.14). Dann gilt

$$\begin{aligned} E(v) &= \frac{1}{2}B(\hat{u}+(v-\hat{u}), \hat{u}+(v-\hat{u})) - \ell(\hat{u}+(v-\hat{u})) \\ &= \frac{1}{2}B(\hat{u},\hat{u}) - \ell(\hat{u}) + \underbrace{B(\hat{u}, v-\hat{u}) - \ell(v-\hat{u})}_{=0} + \frac{1}{2}B(v-\hat{u}, v-\hat{u}) \\ &= E(\hat{u}) + \frac{1}{2}B(v-\hat{u}, v-\hat{u}) \\ &\geq E(\hat{u}) + \frac{c}{2}\|v-u\|^2. \end{aligned}$$

Daraus folgt für $v \neq \hat{u}$ die Ungleichung $E(v) > E(\hat{u})$, d.h. $\hat{u}$ ist das eindeutige Minimum.

Ist umgekehrt $\hat{u}$ Minimum von E, so ist $\hat{u}$ auch stationärer Punkt, d.h. $E'(u)v = 0$ für alle $v \in X$ und wie wir oben gesehen haben ist diese Bedingung genau (3.14). □

Wollen wir das ganze auch direkt auf konvexen Teilmengen $C \subset X$ und $g \in X$ anwenden, d.h. die Energie E über C minimieren, dann folgt aus der selben Rechnung wie oben, dass $\hat{u}$ ein Minimum ist, falls

$$B(\hat{u}, v-\hat{u}) - \ell(v-\hat{u}) \geq 0 \qquad \forall\ v \in C$$

gilt. Damit erhält man eine sogenannte *Variationsungleichung*. Dies können wir auch für den Fall inhomogener Dirichlet-Randbedingungen anwenden, d.h. für $C = g_D + H_0^1(\Omega)$. In diesem Fall ist das Minimum $\hat{u} \in g_D + H_0^1(\Omega)$ durch die obige Variationsgleichung für $v \in g_D + H_0^1(\Omega)$ charakterisiert. Es gilt dann $v - \hat{u} \in H_0^1(\Omega)$ und wir können offensichtlich jedes Element in $H_0^1(\Omega)$ durch geeignete Wahl von v erhalten. Damit erhalten wir

$$B(\hat{u}, w) - \ell(w) \geq 0 \qquad \forall\ w \in H_0^1(\Omega).$$

Da wir im Unterraum $H_0^1(\Omega)$ auch $-w$ als Testfunktion wählen können, wird aus der Ungleichung eine Gleichung, d.h. wir erhalten genau die Variationsgleichung mit Testfunktionen in $H_0^1(\Omega)$ wie auch oben gewählt.

Zum Abschluss geben wir noch den quadratischen Teil der Energie im Fall der partiellen Differentialgleichung (3.1) an, nämlich

$$B(u,u) = \int_\Omega (A\nabla u \cdot \nabla u + cu^2)\, dx.$$

Die Minimierung von E impliziert auch, dass eine gewichtete L^2-Norm des Gradienten und der Funktion klein werden, deshalb ist es auch nicht überraschend, dass genau diese Norm in der Stabilitätsabschätzung kontrolliert werden kann. Aus dem Variationsprinzip könnte man die Stabilität aus der Eigenschaft

$$\frac{1}{2}B(u,u) \le E(u) + \ell(u) \le E(0) + \ell(u) = \ell(u)$$

mit anschliessender Verwendung der Koerzivität und Stetigkeit von ℓ herleiten.

3.2 Galerkin-Approximation

Nachdem wir die Grundlage für die Theorie schwacher Lösungen nun detailliert diskutiert haben, wenden wir uns nun dem eigentlich Zweck dieses Kapitels zu, nämlich der Diskretisierung der Variationsgleichung. Dazu betrachten wir die so genannte *Galerkin-Methode*, d.h. wir wählen einen endlichdimensionalen Teilraum $X^h \subset X$ und suchen eine Lösung $u^h \in X^h$ der Variationsgleichung

$$B(u^h, v) = \ell(v) \qquad \forall\ v \in X^h. \tag{3.18}$$

Die Diskretisierung mit der Galerkin-Methode schränkt die gesamte Varationsgleichung (Lösung und Testfunktion) auf einen endlich-dimensionalen Teil ein. Es ist einfach zu zeigen, dass (3.18) äquivalent zur Minimierung der Energie E über dem Teilraum X^h ist.

Wir nehmen an die Dimension von X^h ist gleich N, und $\{\varphi_j\}_{j=1,\dots,N}$ ist eine Basis von X^h. Bezüglich der Basis können wir u^h in der Form

$$u^h = \sum_{j=1}^N U_j^h \varphi_j$$

entwickeln, mit einem Koeffizientenvektor $U^h \in \mathbb{R}^N$. Durch Verwendung der speziellen Testfunktionen $v = \varphi_k$ in (3.18) erhalten wir dann das $N \times N$ Gleichungssystem

$$\sum_{j=1}^N U_j^h B(\varphi_j, \varphi_k) = \ell(\varphi_k) \qquad k = 1, \dots, N,$$

für den Koeffizientenvektor U^h. Mit der Setzung

$$(K_h)_{jk} := B(\varphi_j, \varphi_k), \qquad (F_h)_j := \ell(\varphi_j) \tag{3.19}$$

für die Matrix $K_h \in \mathbb{R}^{N\times N}$ und die rechte Seite $F_h \in \mathbb{R}^N$ können wir die diskrete Variationsgleichung äquivalent als $K_h U^h = F_h$ schreiben. Für die spezielle Form von B und ℓ im Fall der schwachen Formulierung von (3.1)-(3.3) sehen wir, dass N^2 Gebietsintegrale berechnet werden müssen, um die Matrix K_h aufzustellen, sowie zumindest N Integrale um F_h zu berechnen. Dies kann einen enormen Rechenaufwand bedeuten, der das Verfahren ineffizient

machen würde. Deshalb entstand die Idee der finiten Elemente, d.h. von Ansatzfunktionen φ_j mit lokalem Träger. Durch die lokalen Träger verschwindet in den meisten Fällen eine der beiden Funktionen in der Integration von $B(\varphi_j, \varphi_k)$ und die Matrix K_h wird folglich dünnbesetzt. Bei den nicht verschwindenden Integralen muss nur über einen (kleinen) lokalen Träger integriert werden, was den Rechenaufwand weiter senkt. Wir werden die exakte Konstruktion (Knoten-basierter) finiter Elemente im nächsten Abschnitt genauer diskutieren. Zuvor sammeln wir noch einige allgemeine Resultate zur Analysis der Galerkin-Diskretisierung:

Satz 3.9 (Existenz, Eindeutigkeit, Stabilität). *Es gelten die Voraussetzungen von Satz 3.6 und $X^h \subset X$ sei ein endlichdimensionaler Teilraum. Dann besitzt die diskrete Variationsgleichung* (3.18) *eine eindeutige Lösung $\hat{u}^h \in X^h$ und es gilt die Stabilitätsabschätzung*

$$\|\hat{u}^h\| \leq \frac{1}{c}\|\ell\|. \tag{3.20}$$

Proof. Da X^h mit derselben Norm wie X wieder ein Hilbert-Raum ist (endlichdimensionale Teilräume sind immer abgeschlossen), können wir das Lax-Milgram Lemma auch in X^h anwenden um Existenz und Eindeutigkeit zu zeigen. Für die Stabilitätsabschätzung verwenden wir die Testfunktion $v = \hat{u}^h$ und die Koerzivität. □

Wir sehen, dass sich Existenz, Eindeutigkeit, und Stabilität direkt von der schwachen Formulierung der Differentialgleichung auf die Diskretisierung vererbt. Eine weitere interessante Eigenschaft ist die sogenannte *Galerkin-Orthogonalität*, die wir durch Wahl einer Testfunktion $v \in X^h \subset X$ sowohl in (3.14) und (3.18) erhalten. Subtrahieren wir die resultierenden Gleichungen, so folgt

$$B(\hat{u} - \hat{u}^h, v) = 0 \qquad \forall\, v \in X_h. \tag{3.21}$$

Dies bedeutet, dass der Fehler $\hat{u} - \hat{u}^h$ im durch B definierten Skalarprodukt auf den Teilraum X_h orthogonal steht. Dies ist äquivalent zur Aussage, dass $\hat{u}^h$ die Projektion von $\hat{u}$ auf X^h im durch B definierten Skalarprodukt ist. Um eine Abschätzung im ursprünglichen Skalarprodukt zu erhalten, verwenden wir in der Galerkin-Orthogonalität die Testfunktion $v - \hat{u}^h \in X^h$ und addieren links und rechts $B(\hat{u} - \hat{u}^h, \hat{u})$. Damit folgt

$$B(\hat{u} - \hat{u}^h, \hat{u} - \hat{u}^h) = B(\hat{u} - \hat{u}^h, \hat{u} - v).$$

Die linke Seite dieser Identität können wir unter Verwendung der Koerzivität durch $c\|\hat{u} - \hat{u}^h\|$ nach unten abschätzen, die rechte Seite unter Verwendung der Stetigkeit durch $C\|\hat{u} - \hat{u}^h\|\|\hat{u} - v\|$ nach oben. Insgesamt folgt also die Aussage

Lemma 3.10. *Es gelten die Voraussetzungen von Satz 3.9. Dann gilt*

$$\|\hat{u} - \hat{u}^h\| \leq \frac{C}{c}\|\hat{u} - v\| \qquad \forall\, v \in X^h \tag{3.22}$$

und damit auc

$$\|\hat{u} - \hat{u}^h\| \leq \frac{C}{c} \inf_{v \in X^h} \|\hat{u} - v\|. \tag{3.23}$$

Wir sehen also, dass bis auf eine multiplikative Konstante der Fehler bei der Galerkin-Diskretisierung der kleinstmögliche im gewählten endlichdimensionalen Teilraum ist. Man spricht in diesem Fall von einer Approximation *optimaler Ordnung*, der Fehler des Verfahrens lässt sich durch ein Vielfaches des Projektionsfehlers abzuschätzen. Um den Fehler nun weiter

und auch konkreter abzuschätzen, verwendet man geeignete Elemente v^h in Abhängigkeit von der jeweiligen Struktur der Basisfunktionen φ_j und der exakten Lösung $\hat{u}$. Wir werden später sehen, dass im Fall finiter Elemente zumeist eine geeignete Interpolierende gewählt wird, da der Interpolationsfehler am leichtesten abgeschätzt werden kann. Vom Standpunkt in Kapitel 2 gesehen liefert die Koerzivität genau die Stabilität des Verfahrens, während der noch abzuschätzende Teil $\inf_{v\in X^h} \|\hat{u}-v\|$ als Konsistenz bzw. Konsistenzordnung interpretiert werden kann.

3.3 Finite Elemente

Abstrakt definiert man ein finites Element als Tripel $(K, \mathcal{P}, \mathcal{X})$ mit den Eigenschaften:

(i) $K \subset \mathbb{R}^d$ ist ein kompakte Menge mit nichtleerem Inneren und stückweise stetigem Rand (das Elementgebiet bzw. Referenzelement).

(ii) $\mathcal{P}$ ist ein endlichdimensionaler Raum von Funktionen auf K (der Raum der Formfunktionen)

(iii) $\mathcal{X} = \{X_1, \ldots, X_N\}$ sei eine Basis von $\mathcal{P}'$ (die Knoten).

Die Knotenbasis $\{\varphi_1, \ldots, \varphi_N\}$ erhält man dann als Basis von $\mathcal{P}$ dual zu $\mathcal{P}'$, d.h. $X_j(\varphi_i) = \delta_{ij}$. Im oben diskutierten eindimensionalen Fall finiter Elemente kann man $K = (0, 1)$ als Referenzelement wählen, $\mathcal{P}$ als den Raum der affin-linearen Funktionen auf K und $\mathcal{X}$ entspricht den Randknoten $\{0, 1\}$ oder eigentlich genauer den Distributionen $X_1 = \delta$ und $X_2 = \delta(. - 1)$. Dann erhält man als Knotenbasis $\varphi_1(x) = 1-x$ und $\varphi_2(x) = x$. Durch (lineare) Transformation des Referenzelements $(0, 1)$ auf beliebige Teilintervalle lassen sich dann die Basisfunktionen auf einem Gitter konstruieren, so wie im ersten Kapitel verwendet.

Für das Verständnis finiter Elemente ist die obige Definition jedoch eher unhandlich. Wir werden uns deshalb im folgenden stark einschränken und eine einfachere und weniger allgemeine Sichtweise verwendet. Unter finiten Elementen versteht man normalerweise stückweise polynomiale Ansatzfunktionen auf einem Dreiecksgitter (Tetraedergitter in $3D$) mit lokalem Support. Wir werden uns im folgenden der Einfachheit halber auch auf den zweidimensionalen Fall einschränken, analoges Vorgehen ist aber auch in höheren Dimensionen möglich.

Der erste Schritt zur Konstruktion einer finite Elemente Methode ist dann eine Triangularisierung des Gebiets Ω (bzw. einer polygonalen Approximation $\hat{\Omega}$). Der Vorteil einer Triangularisierung gegenüber rechteckigen Gittern ist die grössere Flexibilität bei der Approximation krummliniger Ränder. Wir nehmen also an, es gilt

$$\overline{\Omega} = \bigcup_{j=1}^{M} T_j \tag{3.24}$$

wobei alle T_j Dreiecke sind, und $T_j \cap T_i$ für $i \neq j$ Mengen vom MaßNull sind. Die Schnittmenge $T_i \cap T_j$ soll entweder leer sein, genau einen Punkt (Knoten), oder genau eine Kante enthalten. Damit nehmen wir implizit an, dass es N Knoten P_i gibt, die durch Kanten zu Dreiecken verbunden sind. Jedem Knoten P_i lässt sich dann seine *Nachbarschaft* $N(P_i)$ zuordnen als

$$N(P_i) = \bigcup_{P_i \in \partial T_j} T_j.$$

Es ist in dieser Sichtweise (knotenbasierte Elemente) natürlich, die Freiheitsgrade in die Knoten P_i zu legen. Dies passiert in dem man für jeden Knoten eine Ansatzfunktion (finites Element) mit den folgenden Eigenschaften konstruiert:

$$\begin{array}{ll} \varphi \in C(\overline{\Omega}) & \\ \varphi_i|_{P_k} = \delta_{ik} & \\ \varphi_i(x) = 0 & x \notin N(P_i) \\ \varphi_i|T_j \in \mathcal{P}_K(T_j) & T_j \subset N(P_i) \end{array} \tag{3.25}$$

Hier bezeichnet $\mathcal{P}_\|(T_j)$ die Menge der Polynome vom Grad kleiner gleich k auf T_j. Die einfachste Wahl für knotenbasierte Elemente ist $k = 1$, was stückweise lineare Elemente liefert. Diese sind durch die Bedingungen in (3.25) eindeutig festgelegt, da es für eine lineare Funktion in zwei Dimensionen genügt die Funktionswerte in der Eckpunkten der Dreiecke zu spezifizieren.

In manchen Anwendungen werden auch Elemente vom Grad $k = 0$ verwendet, allerdings liefern diese keine stetigen Ansatzfunktionen mehr und sind deshalb als knotenbasierte Elemente ungeeignet. Man interpretiert solche unstetigen Ansatzfunktionen als volumsbasierte Elemente, der Freiheitsgrad wird dem Elementmittelpunkt zugeordnet. Eine dritte Klasse von finiten Elementen sind kantenbasierte Elemente, in denen üblicherweise der Freiheitsgrad den Kantenmittelpunkten des Dreiecks zugeordnet wird.

Wir werden jedes Dreieck T_j als Transformation des Einheitdreiecks

$$\hat{T} := \{\ (x_1, x_2) \in [0,1]^2 \mid x_1 + x_2 \leq 1\ \}$$

betrachten, d.h. $T_j = S_j^h(\hat{T})$, wobei die Transformation S_j^h natürlich von der Gittergrösse

$$h := \max_j \text{ diam } T_j \tag{3.26}$$

abhängt. Diese Transformation und ihrer Skalierung kann später vor allem zur Abschätzung des Interpolationsfehlers effektiv verwendet werden.

Der diskrete Teilraum $X^h \subset H^1(\Omega)$, den wir verwenden werden ist gegeben durch

$$X^h = \{\ \sum U_i^h \varphi_i \mid U^h \in \mathbb{R}^N\ \}. \tag{3.27}$$

Wir beginnen mit der Verifikation, dass X^h tatsächlich ein Teilraum von $H^1(\Omega)$ ist.

Lemma 3.11. *Sei $\{\varphi_i\}_{i=1,\ldots,N} \subset C(\Omega)$ eine Familie von Ansatzfunktionen, die* (3.25) *erfüllen. Dann ist X^h definiert durch* (3.27) *ein Teilraum von $H^1(\Omega)$ und $\{\varphi_i\}_{i=1,\ldots,N}$ ist eine Basis von X^h.*

Proof. Da Polynome beschränkt sind, gilt offensichtlich $\varphi_i \in L^\infty(\Omega) \subset L^2(\Omega)$. Um die distributionellen Ableitungen zu berechnen betrachten wir für $\psi \in C_0^\infty(\Omega)$ die Funktionale

$$w(\psi) = -\int_\Omega \varphi_i \frac{\partial \psi}{\partial x_k}\, dx.$$

Setzen wir die spezielle Form von ψ_j ein so folgt

$$w(\psi) = -\int_{N(P_i)} \varphi_i \frac{\partial \psi}{\partial x_k}\, dx = -\sum_{T_j \in N(P_i)} \int_{T_j} \varphi_i \nabla \cdot (\psi e_k)\, dx$$

mit dem Einheitsvektor e_k. Mit dem Gauss'schen Integralsatz erhalten wir daraus

$$w(\psi) = \sum_{T_j \in N(P_i)} \left(\int_{T_j} \frac{\partial \varphi_i|_{T_i}}{\partial x_k} \psi \; dx - \int_{\partial T_j} \varphi_i \psi e_k \cdot n \; d\sigma \right).$$

Für jede Kante $E \subset \partial T_j$ gilt entweder $E \subset \partial N(P_i)$ und damit $\varphi_i|_E = 0$ oder die Kante trennt zwei Dreiecke T_j und T_m. Da die Normalvektoren an ∂T_i und ∂T_j entgegengesetzt orientiert und φ_i, ψ stetig sind, erhalten wir zweimal das selbe Integral über E mit verschiedenen Vorzeichen. Aus diesen Argumenten sehen wir, dass

$$\sum_{T_j \in N(P_i)} \int_{\partial T_j} \varphi_i \psi e_k \cdot n \; d\sigma = 0$$

gilt. Damit folgt

$$w(\psi) = \sum_{T_j \in N(P_i)} \int_{T_j} \frac{\partial \varphi_i|_{T_i}}{\partial x_k} \psi \; dx = \int_\Omega \frac{\partial \varphi_i}{\partial x_k} \psi \; dx,$$

und damit stimmt die distributionelle Ableitung mit den stückweise Ableitungen in den Elementen überein. Da die stückweisen Ableitungen wieder Polynome sind, folgt auch deren Beschränktheit und damit $\varphi_i \in W^{1,\infty}(\Omega) \subset H^1(\Omega)$. Die lineare Hülle der φ_i ist folglich ein Teilraum von $H^1(\Omega)$.

Um die lineare Unabhängigkeit zu zeigen nehmen wir an es gilt

$$\sum \alpha_i \varphi_i = 0$$

für $\alpha_i \in \mathbb{R}$. Da alle Ansatzfunktionen stetig sind, folgt damit auch

$$\alpha_k = \sum \alpha_i \varphi_i(P_k) = 0, \qquad k = 1, \ldots, N.$$

Also sind die φ_j linear unabhängig. □

3.3.1 Assemblierung von Matrizen und Vektoren

Das erste Problem bei der praktischen Durchführung von finite Elemente Methoden ist die Triangulierung des Gebiets Ω (oder einer sinnvollen Approximation im Fall nichtpolygonaler Gebiete). Die Erzeugung eines Gitters ist ein nichttriviales informatisches Problem, vor allem da das Gitter gewisse Eigenschaften erfüllen sollte, wie wir bei der Analysis im nächsten Abschnitt sehen werden. In jedem Fall liefert ein typischer Gittergenerator Information über die Knoten, Kanten, und die Dreiecksflächen (oft als Elemente bezeichnet), in 3D über Knoten, Kanten, Flächen und Volumen. Diese werden normalerweise global indiziert, und durch entsprechende Zuordnungsvorschriften werden diese Indizes miteinander in Beziehung gesetzt, z.B. um zu klären bei welchen Dreieck ein Knoten als Eckpunkt auftritt.

Wir sehen, dass wir zum Aufstellen des diskreten Problems $K_h U^h = F_h$ Produkte von Testfunktionen und gegebenen Funktionen integrieren müssen. In den wenigsten Fällen sind diese Berechnungen analytisch möglich, weshalb man eine geeignete numerische Integrationsformel auf den Dreiecken verwenden sollte. Diese sollte zumindest eine so hohe Ordnung aufweisen wie die Approximationsordnung des benutzten Ansatzraumes um nicht den eigentlichen Fehler des finite Elemente Verfahrens durch den Integrationsfehler zu dominieren (den Approximationsfehler werden wir im nächsten Abschnitt abschätzen).

Neben der Frage der geeigneten numerischen Integration, stellt sich auch die Frage, wie die Matrix K_h und der Vektor F_h aufgebaut (assembliert werden sollen). Nach der obigen Definition scheint ein knotenbasiertes Vorgehen auf den ersten Blick naheliegend. Dabei würde man alle Knoten des Netzes einmal durchlaufen, und die entsprechende Basisfunktion φ_j (also jene die jeweils im aktuellen Knoten den Wert 1 annimmt) in Produkten mit allen anderen umgebenden bzw. mit der rechten Seite zu integrieren. Dies führt zu einem zeilenweisen Aufbau der Matrix K_h und des Vektors F_h. Dieser Zugang stellt sich jedoch nicht als der effizienteste heraus. Ein erstes Problem besteht darin, dass man sich zu jedem Knoten alle benachbarten suchen muss, um herauszufinden, welche der Terme $B(\varphi_i, \varphi_j)$ wirklich integriert werden müssen. Zweitens muss sich dann zu den Knoten auch noch die entsprechenden Elemente (und die darin liegenden Stützstellen für die numerische Integration suchen) um die Assemblierung wirklich durchzuführen.

Für eine effizientere Assemblierung verwendet man deshalb ein elementweises Vorgehen. Grundlage dafür ist die Aufspaltung

$$(K_h)_{ij} = B(\varphi_i, \varphi_j) = \sum_{T_k} \int_{T_k} (A\nabla\varphi_i \cdot \nabla\varphi_j + c\varphi_i\varphi_j).$$

Man kann also auch eine Schleife über die Dreiecke (in 3D Tetraeder) durchführen und die einzelnen Integrale über T berechnen. Diese können dann jeweils zum aktuellen Matrixeintrag ($(K_h)_{ij}$, initialisiert mit 0) addiert werden. In dieser Reihenfolge ist auch klar, welche Integrale jeweils für dieses Element berechnet werden, nämlich genau jene mit Basisfunktion in einem der Randknoten (9 pro Element für Dreiecke in $2D$). Ein völlig analoges Vorgehen ist natürlich auch für die rechte Seite F_h möglich.

Um diese Assemblierung durchzuführen ist es wichtig eine lokale Abbildung zwischen den Elementen und Knoten zu speichern, die jedem Element die Indizes seiner Randknoten und Kanten zuordnet.

3.3.2 Fehlerabschätzungen

Die Grundlage für jede finite Elemente Fehlerabschätzung ist die allgemeine Aussage für Galerkin-Approximationen in Lemma 3.10. Nach dieser Aussage genügt es den Projektionsfehler für die exakte Lösung abzuschätzen. Wir werden diese Abschätzung in diesem Abschnitt exemplarisch in der Dimension $d = 2$ für stückweise lineare Basisfunktionen durchführen, aber einige allgemeine Grundideen dabei betonen.

Wir bemerken auch, dass in der Praxis nicht wirklich die exakte Bilinearform und das wirkliche lineare Funktional ℓ verwendet werden, falls z.B. eine numerische Integration durchgeführt oder der Rand approximiert wird. In Wirklichkeit erhält man dann eine Bilinearform $B_h : X_h \times X_h$, sodass

$$B_h(u^h, v) = \langle \ell^h(v) \qquad \forall\ v \in X_h$$

gilt. Man kann aber dieses Variationsproblem als

$$B(u^h, v) = \ell(v) + r_h(v) \qquad \forall\ v \in X_h$$

schreiben, mit dem Restfunktional

$$r^h(v) = B(u^h, v) - B_h(u^h, v) + \ell_h(v) - \ell(v).$$

Nun kann man die Galerkin-Orthogonalität zu

$$B(\hat{u} - \hat{u}^h, \hat{u} - \hat{u}^h) = B(\hat{u} - \hat{u}^h, \hat{u} - v) + r_h(v)$$

modifizieren. Es genügt dann r^h quantitativ abzuschätzen (z.B. mit den bekannten Methoden für numerische Integration) um wieder eine Fehlerabschätzung für $\hat{u} - \hat{u}^h$ zu erhalten. Als obere Schranke erhält man dann die Summe aus Projektionsfehler und einer von r_h abhängigen Schranke.

Wir wenden uns nun also der Abschätzung des Interpolationsfehlers zu. Wir sehen sofort, dass der Interpolationsoperator bei stückweise linearen Basisfunktionen gegeben ist durch

$$I_h u = \sum_j u(x_j)\varphi_j(x) \qquad \forall\ u \in C(\Omega).$$

Der Fehler bei der Interpolation ist also $v = u - I_h u$. Wir nehmen in der Folge an, dass für die schwache Lösung der Variationsgleichung die Regularität $\hat{u} \in H^2(\Omega)$ gilt, und nach den oben diskutierten Einbettungssätzen folgt (in Dimension zwei) $\hat{u} \in C(\Omega)$. Der Interpolationsfehler ist dann natürlich auch sinnvoll definiert.

Im folgenden werden wir häufig die Transformation $\overset{h}{j}$ zwischen Dreiecken T_j und dem Referenzdreieck $\hat{T}$ verwenden. Nach der Transformationsregel für Integrale gilt

$$\int_{T_j} \varphi(x)\ dx = \frac{1}{\delta_j^h} \int_{\hat{T}} \varphi(S_j^h(y))\ dy,$$

wobei $\delta_j^h = |\det(\nabla S_j^h)|$. Wir berechnen leicht $\nabla S_j^h = \mathcal{O}(h)$ und damit $\delta_j^h = \mathcal{O}(h^2)$. Wir nehmen nun an, dass

$$c_1 h \leq \lambda_{\min}(\nabla S_j^h) \leq \lambda_{\max}(\nabla S_j^h) \leq c_2 h. \tag{3.28}$$

Damit folgt $c_1^2 h^2 \leq \delta_j^h \leq c_2^2 h^2$. Für Normen auf $\hat{T}$ erhalten wir folgende Transformation:

$$\begin{aligned} \int_{T_j} \varphi(x)^2\ dx &= \frac{1}{\delta_j^h} \int_{\hat{T}} \varphi(S_j^h(y))^2\ dy \\ \int_{T_j} |\nabla\varphi(x)|^2\ dx &= \frac{1}{\delta_j^h} \int_{\hat{T}} |(\nabla S_j^h)\nabla\varphi(S_j^h(y))|^2\ dy. \end{aligned}$$

Wir betrachten den Interpolationsfehler in T_j zurücktransformiert auf $\hat{T}$, d.h. $\varphi = (u - I_h u) \circ S_j^h$, für $u \in H^2(\Omega)$. Wegen $u|_{T_j} \in H^2(T_j)$ und da $(I_h u)|_{T_j}$ linear ist, folgt $\varphi \in H^2(\hat{)}$. Also können wir die H^2-Norm von φ betrachten, bzw. eine weitere Norm

$$|||\varphi||| = \sqrt{|\varphi|_{2,2}^2 + \varphi(P_1)^2 + \varphi(P_2)^2 + \varphi(P_3)^2},$$

wobei $P_1 = (0,0)$, $P_2 = (1,0)$ und $P_3 = (0,1)$ gilt. Sei $f_i(\varphi) = |\varphi(P_i)|$, dann ist f_i eine Halbnorm auf $H^2(\hat{T})$ und das System der Halbnormen (f_1, f_2, f_3) erfüllt die Bedingungen des Sobolevschen Normierungssatz, d.h. für eine lineare Funktion (Polynom vom Grad kleiner gleich eins) auf $\hat{T}$ gibt es immer ein P_i mit $\varphi(P_i) \neq 0$. Folglich ist die Norm $|||.|||$ äquivalent zur H^2-Norm auf $\hat{T}$, d.h. es existiert eine Konstante γ, sodass

$$\|\varphi\|_{1,2} \leq \|\varphi\|_{2,2} \leq \gamma |||\varphi|||$$

gilt.

Nun betrachten wir die Abschätzung näher für $\varphi = (u - I_h u) \circ S_j^h$. Da $(I_h u) \circ S_j^h$ immer noch linear auf $\hat{T}$ ist, verschwinden alle zweiten Ableitung. Also gilt $|\varphi|_{2,2} = |u \circ S_j^h|_{2,2}$. Weiter ist wegen der Interpolationseigenschaft $\varphi(P_k) = (u - I_h u)(S_j^h(P_i)) = 0$ (beachte $S_j^h(P_i)$ ist Eckpunkt des Dreiecks T_j und $I_h u$ interpoliert u in den Eckpunkten). Also folgt $|||\varphi||| = |u \circ S_j^h|_{2,2}$ und damit

$$\|\varphi\|_{1,2} \leq \gamma |u \circ S_j^h|_{2,2}.$$

Nun können wir die Rücktransformation der Normen auf das Dreieck T_j benutzen, es gilt ja

$$\|\varphi\|_{1,2}^2 = \int_{\hat{T}} (\varphi(x)^2 + |\nabla \varphi(x)|^2)\ dx = \frac{1}{\delta_j^h} \int_{T_j} ((u - I_h u)^2 + |(\nabla S_j^h)(\nabla u - \nabla(I_h u))|^2)\ dx.$$

Nun können wir die obigen Abschätzungen für die Eigenwerte von S_j^h benutzen, um weiter zu zeigen

$$\|\varphi\|_{1,2}^2 \geq \frac{1}{\delta_j^h} \int_{T_j} ((u - I_h u)^2 + c_1^2 h^2 |\nabla u - \nabla(I_h u)|^2)\ dx.$$

Analog können wir die Transformation für die zweiten Ableitungen ausrechnen. Es gilt

$$|u \circ S_j^h|_{2,2}^2 = \int_{\hat{T}} \sum_k |\nabla \partial_{x_k} (u \circ S_j^h)|^2\ dx = \frac{1}{\delta_j^h} \int_{T_j} \sum_k |(\nabla S_j^h) \nabla ((\nabla S_j^h) \nabla u)_k|^2\ dx,$$

und mit den obigen Abschätzungen folgt

$$\begin{aligned} |u \circ S_j^h|_{2,2}^2 &\leq c_2 h^2 \frac{1}{\delta_j^h} \int_{T_j} \sum_k |\nabla ((\nabla S_j^h) \nabla u)_k|^2\ dx \leq d c_2^2 h^4 \frac{1}{\delta_j^h} \int_{T_j} \sum_k |\nabla \partial_{x_k} u|^2\ dx \\ &= d c_2^2 h^4 \frac{1}{\delta_j^h} |u|_{H^2(T_j)}^2. \end{aligned}$$

Durch Kombination der obigen Abschätzungen erhalten wir also

$$\int_{T_j} ((u - I_h u)^2 + c_1^2 h^2 |\nabla u - \nabla(I_h u)|^2)\ dx \leq \gamma d c_2^2 h^4 |u|_{H^2(T_j)}^2.$$

Nun summieren wir noch über die Dreiecke T_j und folgern (für h hinreichend klein, sodass $c_1 h \geq 1$)

$$\begin{aligned} \|u - I_h u\|_{1,2}^2 &= \sum_j \int_{T_j} ((u - I_h u)^2 + c_1^2 h^2 |\nabla u - \nabla(I_h u)|^2)\ dx \\ &\leq \frac{1}{c_1^2 h^2} \sum_j \int_{T_j} ((u - I_h u)^2 + c_1^2 h^2 |\nabla u - \nabla(I_h u)|^2)\ dx \\ &\leq \gamma d \frac{c_2^2}{c_1^2} h^2 \sum_j |u|_{H^2(T_j)}^2 \\ &= \gamma d \frac{c_2^2}{c_1^2} h^2 |u|_{2,2}^2. \end{aligned}$$

Ziehen wir nun abschliessend noch die Wurzel und nennen $\tilde{C} = \sqrt{\gamma d} \frac{c_2}{c_1}$, dann haben wir folgendes Resultat bewiesen.

Proposition 3.12. *Seien B und ℓ wie oben und $u \in H^1(\Omega)$ die Lösung des Variationsproblems*

$$B(u, v) = \ell(v) \qquad \forall\ v \in H^1(\Omega),$$

mit der zusätzlichen Regularität $u \in H^2(\Omega)$. Weiter sei (3.28) mit Konstanten c_1, c_2 unabhängig von h erfüllt. Dann existiert eine von h unabhängige Konstante $C > 0$, sodass

$$\|u - I_h u\|_{1,2} \leq \tilde{C}|u|_{2,2}h \tag{3.29}$$

gilt.

Aus der Abschätzung des Interpolationsfehlers und der obigen Galerkin-Fehlerabschätzung können wir nun eine quantitative Fehlerabschätzung für die finite Elemente Diskretisierung angeben.

Satz 3.13. *Seien die Bedingungen von Proposition erfüllt und sei $u^h \in X^h$ die Lösung des diskretisierten Problems*

$$B(u^h, v) = \ell(v) \qquad \forall\ v \in X^h,$$

wobei X^h der Ansatzraum der stückweise linearen finiten Elemente ist. Dann gilt die Fehlerabschätzung

$$\|u - u^h\|_{1,2} \leq \frac{C}{c}\|u - I_h u\|_{1,2} \leq \frac{C\tilde{C}}{c}|u|_{2,2}h = \mathcal{O}(h). \tag{3.30}$$

Die obige Technik zur Herleitung von Fehlerabschätzungen kann sowohl bezüglich des Polynomgrads der Elemente, der Ordnung der Appproximation, als auch der Ordnung der Ableitungen verallgemeinert werden. Wichtig ist immer auf dem Referenzelement eine entsprechende Abschätzung einer Sobolev-Norm durch eine Halbnorm in einem Sobolev-Raum höherer Ableitungsordnung zu erhalten. Die unterschiedlichen Skalierungseigenschaften der Ableitungen bei der Rücktransformation auf die Dreiecke im Gitter liefern dann Abschätzungen bezüglich h.

Man sieht auch, dass die Eigenschaft (3.28) mit Konstanten c_1, c_2 unabhängig von h essentiell für die Güte der Fehlerabschätzung ist. Diese Bedingung lässt sich als eine Bedingung an die Dreiecke im Gitter auffassen, diese sollte (insbesonders mit kleiner werdendem h) nicht zu weit entarten. Wie wir in der Übung sehen werden, kann (3.28) auch als Bedingung an In- und Umkreisradius, sowie auch an den kleinsten und grössten Winkel im Dreieck interpretiert werden.

Die Abschätzung (3.30) ist ein Beispiel einer *a-priori* Abschätzung, d.h. es werden Eigenschaften über die Lösung u a-priori vorausgesetzt, und damit sind auch die Konstanten in der Abschätzung nicht explizit berechenbar (z.B. kennt man $|u|_{2,2}$ nicht, da man ja die Lösung u nicht kennt). Alternativ lassen sich sogenannte *a-posteriori* Abschätzungen herleiten, bei denen die Schranken nur von der diskreten Lösung u^h (die man wirklich berechnet) abhängen. Diese Abschätzungen werden dann vor allem zur adaptiven Verfeinerung des Gitters verwendet. Wir werden diese Aspekte in dieser Vorlesung nicht weiter diskutieren und verweisen auf weiterführende Lehrveranstaltungen zur Numerik partieller Differentialgleichungen.

3.3.3 Eigenwerte und Kondition von K_h

Zum Abschluss dieses Kapitels betrachten wir noch einmal die Eigenwerte bzw. Konditionszahl der Systemmatrix K_h. Da K_h symmetrisch und positiv definit ist, gilt nach dem

Rayleigh-Prinzip den grössten und kleinsten Eigenwert aus

$$\lambda_{\min} = \min_{V\in\mathbb{R}^N\setminus\{0\}} \frac{V^T K_h V}{V^T V} \tag{3.31}$$

$$\lambda_{\max} = \max_{V\in\mathbb{R}^N\setminus\{0\}} \frac{V^T K_h V}{V^T V} \tag{3.32}$$

berechnen, sowie die Konditionszahl als $\kappa = \frac{\lambda_{\max}}{\lambda_{\min}}$.

Wir definieren nun die Funktion $v = \sum V_j \varphi_j$, dann gilt

$$V^T K_h V = \sum_{j,k} V_j V_k B(\varphi_j, \varphi_k) = B(v, v).$$

Wie wir in der Übung sehen werden, folgt aus (3.28) die Existenz von Konstanten β_1 und β_2, sodass

$$\beta_1 \int_\Omega v^2 \, dx = \beta_1 \sum_j \int_{T_j}^2 v^2 \, dx \le h^2 \sum_i V_i^2 \le \beta_2 \sum_j \int_{T_j}^2 v^2 \, dx = \beta_2 \int_\Omega v^2 \, dx.$$

Verwenden wir nun noch die Koerzivität und Stetigkeit von B, so folgt

$$\frac{c}{\beta_2} h^2 \frac{\|v\|_{1,2}^2}{\|v\|_2^2} \le \frac{V^T K_h V}{V^T V} \le \frac{C}{\beta_1} h^2 \frac{\|v\|_{1,2}^2}{\|v\|_2^2}. \tag{3.33}$$

Wegen der Normabschätzung $\|v\|_{1,2} \ge \|v\|_2$ folgt

$$\lambda_{\min} = \min_{V\in\mathbb{R}^N\setminus\{0\}} \frac{V^T K_h V}{V^T V} \ge \min_{v\in X^h\setminus\{0\}} \frac{c}{\beta_2} h^2 \frac{\|v\|_{1,2}^2}{\|v\|_2^2} \ge \frac{c}{\beta_2} h^2.$$

Die Abschätzung des grössten Eigenwerts ist schwieriger, da wir im Allgemeinen die H^1-Norm nicht durch die L^2-Norm nach oben abschätzen können. Deswegen verwenden wir wieder die Transformation auf das Referenzelement und bezeichnen mit $\tilde{v}$ die transformierte Funktion, die wiederum linear auf dem Einheitsdreieck ist. Da der Raum der linearen Funktionen auf dem Einheitsdreieck endlichdimensional ist, und alle Normen auf einem endlichdimensionalen Raum äquivalent sind, existiert eine Konstante $\alpha > 0$, sodass

$$\|\varphi\|_{H^1(T)} \le \alpha \|\varphi\|_{L^2(T)}$$

für alle linearen Funktionen φ auf T gilt. Wie oben erhalten wir

$$\|\tilde{v}\|_{L^2(T)} = \frac{1}{\sqrt{\delta_j^h}} \|v\|_{L^2(T_j)}$$

und

$$\|\tilde{v}\|_{H^1(T)} \ge \frac{c_1 h}{\sqrt{\delta_j^h}} \|v\|_{H^1(T_j)},$$

und folglich

$$\|v\|_{H^1(T_j)} \le \frac{\alpha}{c_1 h} \|v\|_{L^2(T_j)}.$$

Nach Summation über die Dreiecke folgt die sogenannte *inverse Ungleichung*

$$\|v\|_{1,2} \leq \frac{\alpha}{c_1 h}\|v\|_2. \tag{3.34}$$

Nun können wir auch den grössten Eigenwert abschätzen, und zwar als

$$\lambda_{\max} = \max_{V\in\mathbb{R}^N\setminus\{0\}} \frac{V^T K_h V}{V^T V} \leq \max_{v\in X^h\setminus\{0\}} \frac{C}{\beta_1} h^2 \frac{\|v\|_{1,2}^2}{\|v\|_2^2} \leq \frac{C\alpha^2}{\beta_1 c_1^2}.$$

Abschliessend können wir nun auch die Konditionszahl abschätzen durch

$$\kappa \leq \frac{C\beta_2\alpha^2}{c\beta_1 c_1^2}\frac{1}{h^2}.$$

Neben den vom FE-Raum abhängigen Konstanten α, β_1 und c_1 sehen wir wieder den Quotienten $\frac{C}{c}$, der ein Maß für die Stabilität des ursprünglichen Problems ist (siehe Stabilitätsabschätzung nach Lax-Milgram). Zusätzlich tritt noch der Faktor $\frac{1}{h^2}$ auf, der bei kleinem h für eine sehr schlechte Kondition sorgt, und damit auch besondere Sorgfalt bei der Wahl iterativer Lösungsverfahren erfordert wie wir in Kapitel 5 noch sehen werden.

Kapitel 4

Zeitdiskretisierung

In diesem Kapitel werden wir kurz ein paar Aspekte der Zeitdiskretisierung parabolischer und hyperbolischer Probleme diskutieren. Bei einer vertikalen Linienmethode hat man nach der Diskretisierung im Ort (mit Methoden analog zu Kapitel 2) ein System gewöhnlicher Differentialgleichungen in der Zeit zu diskretisieren, und es werden tatsächlich dieselben Methoden wie bei gewöhnlichen Differentialgleichungen verwendet. Allerdings ist bei partiellen Differentialgleichungen besondere Vorsicht geboten wegen der Änderung mit der Diskretisierung h. Vor allem für feine Ortsdiskretisierung erhält man sogenannte *steife gewöhnliche Differentialgleichungen*, die besondere Vorsicht bezüglich der Stabilität erforden. Wir werden uns deshalb in diesem Kapitel vor allem mit der Stabilität der Verfahren befassen.

4.1 Parabolische Probleme

Ein abstraktes parabolisches Problem ist von der Form

$$\frac{\partial u}{\partial t} - Lu = f, \qquad \text{in } \Omega \times T \tag{4.1}$$

mit Anfangswerten $u(0) = u_0$, wobei L ein elliptischer Differentialoperator (mit Randbedingungen) ist. Wir nehmen an, dass L bzw. die zugehörige Bilinearform B in $H^m(\Omega)$ stabil ist. Im Falle eines Operators zweiter Ordnung gilt dann unter den typischen Voraussetzungen $L : H^1(\Omega) \to H^{-1}(\Omega)$. Damit sollten natürlich f und $\frac{\partial u}{\partial t}$ für fast alle t sinnvollerweise Elemente in $H^{-1}(\Omega)$ sein. Ein bisschen Vorsicht ist noch bezüglich der Abhängigkeit bezüglich t geboten. Dazu verwendet man normalerweise sogenannte *vektorwertige Sobolevräume*. Wir definieren

$$H^k(0,T;H^m(\Omega)) := \{\ u \in H^m(\Omega) \mid \frac{\partial u^j}{\partial t^j} \in H^m(\Omega), j = 1,\dots,k\ \}.$$

Analog schreiben wir $L^2(0,T;H^m(\Omega))$ im Fall $k = 0$ bzw. $H^k(0,T;L^2(\Omega))$ im Fall $m = 0$. Natürlich können wir den Raum auch für m negativ in der üblichen Weise definieren.

4.1.1 Schwache Formulierung

Um eine Idee für die schwache Formulierung zu erhalten, multiplizieren wir wieder mit einer Testfunktion v und integrieren, zunächts nur bezüglich x. Dann gilt

$$\langle \frac{\partial u}{\partial t}, v\rangle + \int_\Omega (-Lu)v\ dx = \int_\Omega fv\ dx.$$

Bis auf den Teil mit der Zeitableitung können wir alle Terme analog zur schwachen Formulierung im elliptischen Fall behandeln und erhalten

$$\langle \frac{\partial u}{\partial t}, v\rangle + B(u,v) = \ell(v).$$

Für $\frac{\partial u}{\partial t} \in H^{-m}(\Omega)$ und $v \in H^m(\Omega)$. Um eine Idee über die Zeitabhängigkeit zu erhalten, setzen wir wieder $v = u$ und integrieren über die Zeit. Dann gilt

$$\begin{aligned}\frac{1}{2}\|u(t)\|_2^2 + \int_0^t B(u(s),u(s))\ ds &= \frac{1}{2}\|u_0\|_2^2 + \int_0^t \langle(\frac{\partial u}{\partial t}, u\rangle + B(u(s),u(s)))\ ds \\ &= \int_0^t fu(s)\ ds + \frac{1}{2}\|u_0\|_2^2.\end{aligned}$$

Ist nun B koerziv in $H^m(\Omega)$, dann folgt

$$\frac{1}{2}\|u(t)\|_2^2 + c\int_0^t \|u(s)\|_{m,2}^2\ ds \le \frac{1}{2c}\int_0^t \|f\|_{-m,2}^2\ ds + \frac{c}{2}\int_0^t \|u(s)\|_{m,2}^2\ ds + \frac{1}{2}\|u_0\|_2^2,$$

und da $t \le T$ beliebig ist, folgt

$$\sup_t \|u(t)\|_2^2 \le \frac{1}{c}\int_0^T \|f\|_{-m,2}^2\ ds + \|u_0\|_2^2$$

und

$$\int_0^T \|u(s)\|_{m,2}^2\ ds \le \frac{1}{c^2}\int_0^T \|f\|_{-m,2}^2\ ds + \frac{1}{c}\|u_0\|_2^2.$$

Damit folgt $u \in L^\infty(0,T;L^2(\Omega)$ (mit der offensichtlichen Definition dieses Raums) und $u \in L^2(0,T;H^m(\Omega)$ mit jeweiliger Stabilitätsabschätzung.

Eine Abschätzung für die Zeitableitung erhält man mit der Testfunktion $v = (-L)^{-1}\frac{\partial u}{\partial t}$ (man beachte dass unter den üblichen Elliptizitätsvoraussetzungen $-L$ ein regulärer linearer Operator von $H^m(\Omega)$ nach $H^{-m}(\Omega)$ ist). Es folgt dann

$$B(v,v) + \langle u, \frac{\partial u}{\partial t}\rangle = \langle \frac{\partial u}{\partial t}, (-L)^{-1}\frac{\partial u}{\partial t}\rangle + \langle (-L)u, (-L)^{-1}\frac{\partial u}{\partial t}\rangle = \langle f, (-L)^{-1}\frac{\partial u}{\partial t}\rangle$$

und mit der Koerzivität folgt nach Zeitintegration

$$c\int_0^t \|v(s)\|_{m,2}^2\ ds + \frac{1}{2}\|u(t)\|_2^2 \le \frac{1}{2c}\int_0^t \|f\|_{-m,2}^2\ ds + \frac{c}{2}\int_0^t \|v(s)\|_{m,2}^2\ ds + \frac{1}{2}\|u_0\|_2^2.$$

Damit erhalten wir eine Stabilitätsabschätzung für $v = (-L)^{-1}\frac{\partial u}{\partial t} \in L^2(0,T;H^m(\Omega))$. Da $(-L)^{-1}$ ein stetiger Operator von $H-m(\Omega)$ nach $H^m(\Omega)$ ist, folgt also $\frac{\partial u}{\partial t} \in L^2(0,T;H^{-m}(\Omega))$.

Eine vernünftige schwache Lösung des parabolischen Problems mit Anfangswert $u_0 \in L^2(\Omega)$ sollte also

$$u \in L^\infty(0,T;L^2(\Omega)) \cap L^2(0,T;H^m(\Omega)) \cap H^1(0,T;H^{-1}(\Omega))$$

erfüllen. Eine analoge Stabilität sollte man dann auch vom diskretisierten Problem erwarten.

4.1.2 Maximumprinzip

Ist $u_0 \in L^\infty(\Omega)$ und der Operator L zweiter Ordnung, dann kann man auch wieder Maximumprinzipien anwenden und daraus eine Lösung $u \in L^\infty((0,T)\times\Omega)$ mit entsprechender Stabilitätsabschätzung erwarten. Das Maximumprinzip gilt analog zum elliptischen Fall, sogar noch mit einigen Erweiterungen:

Satz 4.1. *Sei L ein Operator zweiter Ordnung wie in Kapitel 2 mit $c = 0$. Weiter sei $u \in C(0,T;C^2(\Omega)) \cap C^1(0,T;C(\Omega))$ so dass*

$$\frac{\partial u}{\partial t} - Lu < 0 (> 0) \qquad \text{in } \Omega \times (0,T)$$

gilt. Dann nimmt u kein Maximum (Minimum) in $\Omega \times (0,T]$ an.

Proof. Angenommen ein Maximum wird in $(x,t) \in \Omega \times (0,T)$ angenommen. Dann gilt dort $\frac{\partial u}{\partial t}(x,t) = 0$ und wie in Kapitel 2 $Lu(x,t) \leq 0$. Dies ergibt bei Einsetzen in die Gleichung einen Widerspruch. Im Fall eines Maximums (x,T) würde ein Randmaximum bezüglich der Zeit vorliegen und damit gilt zumindest $\frac{\partial u}{\partial t}(x,t) \geq 0$ und wegen der Maximalität im Ort gilt $Lu(x,t) \leq 0$. Einsetzen in die Gleichung führt auch hier zu einem Widerspruch. □

Analoge Aussagen lassen sich natürlich auch für $c \geq 0$ zeigen, einige wie z.B. die Erhaltung von Positivität sind sogar für negatives c möglich. Dazu verwendet man die Transformation $v = e^{-\alpha t}u$ mit α hinreichend gross. Es gilt nämlich

$$\frac{\partial u}{\partial t} - Lu = e^{\alpha t}\left(\frac{\partial v}{\partial t} + \alpha v - Lv\right)$$

und für $\alpha > -\max c$ ist der Koeffizient niedrigster Ordnung für v sicher positiv. Damit kann man Maximumprinzipien für v anwenden, und da u dasselbe Vorzeichen wie v hat auch wieder aus Positivität oder Negativität von u schließen.

4.1.3 Ortsdiskretisierung

Wir betrachten nun ein im Ort diskretisiertes Problem. Dazu wenden wir entweder direkt finite Differenzen (mit zeitabhängigen Werten an den Gitterpunkten) oder finite Elemente (mit zeitabhängigen Koeffizienten) für die schwache Formulierung

$$\langle \frac{\partial u}{\partial t}, v\rangle + \int_\Omega (-Lu)v \; dx = \int_\Omega fv \; dx.$$

m Fall finiter Differenzen führt dies auf ein System gewöhnlicher Differentialgleichungen

$$\frac{dU^h}{dt} + K_h U^h = F_h, \qquad U^h(0) = U_0^h. \tag{4.2}$$

Bei einer finite Elemente Diskretisierung im Ort erhält man

$$M_h \frac{dU^h}{dt} + K_h U^h = F_h, \qquad U^h(0) = U_0^h \tag{4.3}$$

mit der Massenmatrix

$$(M_h)_{ij} = \langle \varphi_i, \varphi_j \rangle = \int_\Omega \varphi_i(x)\varphi_j(x) \; dx.$$

Die Matrix K_h (und auch M_h) ist unter den üblichen Annahmen symmetrisch positiv definit. Damit kann man (für theoretische Zwecke) die Matrix K_h diagonalisieren. Es existieren eine orthogonale Matrix S_h und eine Diagonalmatrix Σ_h, sodass $S_h^T K_h S_h = \Sigma_h$ gilt. Wir definieren nun $V^h = S_h^T U^h$, dann folgt

$$\frac{dV^h}{dt} + \Sigma_h V^h = G_h, \qquad V^h(0) = V_0^h \tag{4.4}$$

mit $G_h = S_h^T F_h$ und $V_0^h = S_h^T U_h^0$. Analog zu Kapitel 1 folgt $\|U^h\| = \|V^h\|$, d.h. L^2-Stabilität kann auch bezüglich V^h untersucht werden. Bei einer finite Elemente Diskretisierung kann man analog vorgehen, wenn man zuerst von links und rechts mit $(M_h^{-1/2})$ multipliziert (so eine Matrix existiert, da M_h symmetrisch positiv definit ist) und danach die Diagonalisierung anwendet.

4.1.4 Zeitdiskretisierungen: Explizit, Implizit und Mehrschritt

Wir führen nun drei verschiedene Zeitdiskretisierungen durch Differenzenquotienten in der Zeit ein. Wir bezeichnen dazu mit $U_k^{h,\tau}$ die diskrete Lösung zum Zeitpunk $t_k = k\tau$. Die erste Wahl ist das *explizite Euler-Verfahren*

$$U_{k+1}^{h,\tau} = U_k^{h,\tau} + \tau(-K_h U_k^{h,\tau} + F_h(t_k)). \tag{4.5}$$

Die erste Alternative dazu ist das *implizite Euler-Verfahren*

$$U_{k+1}^{h,\tau} + \tau K_h U_{k+1}^{h,\tau} = U_k^{h,\tau} + \tau F_h(t_{k+1}). \tag{4.6}$$

Ein drittes interessantes Verfahren erhält man aus dem Mittelwert der beiden obigen Formeln: das sogenannte *Crank-Nicholson Verfahren*

$$U_{k+1}^{h,\tau} + \frac{\tau}{2} K_h U_{k+1}^{h,\tau} = U_k^{h,\tau} - \frac{\tau}{2} K_h U_k^{h,\tau} + \frac{\tau}{2}(F_h(t_k) + F_h(t_{k+1})). \tag{4.7}$$

Bei einer finite Elemente Diskretisierung im Ort sehen alle drei Verfahren genauso aus, es werden nur die Terme, die aus der Zeitableitung resultieren, mit M_h multipliziert.

Wir sehen sofort, dass diese Verfahren eigentlich nur drei Beispiele eines allgemeineren Verfahrens mit Parameter $\theta \in [0,1]$ sind

$$U_{k+1}^{h,\tau} + \theta\tau K_h U_{k+1}^{h,\tau} = U_k^{h,\tau} - (1-\theta)\tau K_h U_k^{h,\tau} + \tau((1-\theta)F_h(t_k) + \theta F_h(t_{k+1}))). \tag{4.8}$$

Für $\theta = 0$ erhalten wir das explizite Euler-Verfahren, für $\theta = 1$ das implizite Euler-Verfahren, und für $\theta = \frac{1}{2}$ das Crank-Nicholson Verfahren. Andererseits kann man dieses Verfahren für $\theta \in (0,1)$ als Mehrschrittverfahren, wiederum zusammengesetzt aus explizitem und implizitem Euler-Verfahren interpretieren, d.h. man berechnet eigentlich einen Zwischenschritt $U_{k+1-\theta}^{h,\tau}$ mit Zeitschrittweite $(1-\theta)\tau$ explizit und danach $U_{k+1}^{h,\tau}$ implizit mit Zeitschrittweite $\theta\tau$

$$U_{k+1-\theta}^{h,\tau} = U_k^{h,\tau} + (1-\theta)\tau(-K_h U_k^{h,\tau} + F_h(t_k)) \tag{4.9}$$

$$U_{k+1}^{h,\tau} + \theta\tau K_h U_{k+1}^{h,\tau} = U_{k+1-\theta}^{h,\tau} + \theta\tau F_h(t_{k+1}). \tag{4.10}$$

4.1.5 Konsistenz

Wir beginnen unsere Untersuchungen mit der Konsistenz der Zeitdiskretisierung. Sei U^h Lösung des semidiskreten Problems (4.2) und $U_k^{h,\tau}$ Lösung des voll diskretisierten Problems (4.8). Aus einer Taylor-Entwicklung erhalten wir (man beachte in der zweiten Gleichung die zusätzliche Approximation $\frac{d^2U^h}{dt^2}(t_{k+1}) = \frac{d^2U^h}{dt^2}(t_k) + \mathcal{O}(\tau)$)

$$U^h(t_{k+1}) - U^h(t_k) = \tau \frac{dU^h}{dt}(t_k) + \frac{\tau^2}{2}\frac{d^2U^h}{dt^2}(t_k) + \mathcal{O}(\tau^3)$$
$$U^h(t_{k+1}) - U^h(t_k) = \tau \frac{dU^h}{dt}(t_{k+1}) - \frac{\tau^2}{2}\frac{d^2U^h}{dt^2}(t_k) + \mathcal{O}(\tau^3).$$

Nun multiplizieren wir die erste Gleichung mit $\frac{1-\theta}{\tau}$, die zweite mit $\frac{\theta}{\tau}$ und addieren sie. Damit erhalten wir

$$\frac{1}{\tau}(U^h(t_{k+1}) - U^h(t_k)) - (1-\theta)\frac{dU^h}{dt}(t_k) - \theta\frac{dU^h}{dt}(t_{k+1}) = \frac{\tau}{2}(1-2\theta)\frac{d^2U^h}{dt^2}(t_k) + \mathcal{O}(\tau^2).$$

Nun setzen wir noch (4.2) ein und erhalten

$$\frac{1}{\tau}(U^h(t_{k+1}) - U^h(t_k)) + (1-\theta)K_hU^h(t_k) + \theta K_hU^h(t_{k+1})$$
$$= (1-\theta)F_h(t_k) + \theta F_h((t_{k+1}) + \frac{\tau}{2}(1-2\theta)\frac{d^2U^h}{dt^2}(t_k) + \mathcal{O}(\tau^2).$$

Also ist der Fehler bei Einsetzen von U^h in das voll diskrete Verfahren

$$\frac{\tau}{2}(1-2\theta)\frac{d^2U^h}{dt^2}(t_k) + \mathcal{O}(\tau^2).$$

Wir sehen also, dass explizites und implizites Euler-Verfahren einen Fehler erster Ordnung $\mathcal{O}(\tau)$ liefern. Das Crank-Nicholson Verfahren (und jedes Verfahren mit der Wahl $\theta = \frac{1}{2} + \mathcal{O}(\tau)$) liefert eine Fehler zweiter Ordnung $\mathcal{O}(\tau^2)$, also eine höhere Konsistenzordnung.

4.1.6 Stabilität

Im folgenden werden wir uns der Stabilitätsanalyse der obigen zeitdiskretisierten Verfahren widmen. Dies werden wir in drei verschiedenen Versionen durchführen: bezüglich L^2-Stabilität mittels Diagonalisierung der Matrix K_h, bezüglich L^∞-Stabilität mittels M-Matrix Techniken für finite Differenzen und bezüglich H^1-Stabilität mittels Variationstechniken für finite Elemente.

L^2-Stabilität

Wie auch schon im Kapitel über finite Elemente Methoden diskutiert, ist die L^2-Norm einer Funktion u^h auf einem Gitter proportional zu h^dmultipliziert mit der euklidischen Norm des Vektors U^h von Funktionswerten am Gitter. Deshalb führen wir hier die Stabilitätsanalyse bezüglich der euklidischen Norm des Vektors $U_k^{h,\tau}$ durch.

Wie oben verwenden wir die Diagonalisierung $S_h^T K_h S_h = \Sigma_h$ gilt. Wir definieren wieder $V_k = S_h^T U_k^{h,\tau}$, und erhalten das diagonalisierte Verfahren

$$V_{k+1} + \theta\tau\Sigma_h V_k = V_k + (1-\theta)\tau\Sigma_h V_k + \theta\tau G_{k+1} + (1-\theta)\tau G_k, \tag{4.11}$$

mit $G_k = S_h^T F_h(t_k)$. Da Σ_h eine Diagonalmatrix aus den Eigenwerten λ_j. Damit können wir $(V_{k+1})_j$ aus der Iterationsformel

$$(V_{k+1})_j = \frac{1-\tau(1-\theta)\lambda_j}{1+\tau\theta\lambda_j}(V_k)_j + \frac{\tau}{1+\tau\theta\lambda_j}(\theta(G_{k+1}) + (1-\theta)(G_k)_j).$$

Entscheidend für die Stabilität oder Instabilität des Verfahrens ist der Faktor dieser geometrischen Reihe. Das Verfahren wird stabil genau dann, wenn

$$-1 \leq \frac{1-\tau(1-\theta)\lambda_j}{1+\tau\theta\lambda_j} \leq 1$$

für alle j gilt. Die rechte Ungleichung ist wegen der Positivität der Eigenwerte λ_j offensichtlich für beliebige Werte τ und $\theta \in [0,1]$ erfüllt. Die linke Ungleichung ist äquivalent zu

$$1 + \tau\theta\lambda_j \geq -1 + \tau(1-\theta)\lambda_j$$

bzw.

$$2 \geq (1-2\theta)\tau\lambda_j.$$

Wir sehen für $\theta \geq \frac{1}{2}$, dass die rechte Seite nichtpositiv und damit immer kleiner 2 ist. Für diese Werte von θ (und damit auch für das implizite Euler und Crank-Nicholson Verfahren) gilt also unbedingte Stabilität.

Für $\theta < \frac{1}{2}$ erhalten wir nur bedingte Stabilität mit der Einschränkung

$$\tau \leq \frac{2}{(1-2\theta)\max_j \lambda_j}.$$

Nun erinnern wir uns, dass für K_h resultierend aus der Ortsdiskretisierung eines elliptischen Differentialoperators zweiter Ordnung der maximale Eigenwert $\max_j \lambda_j$ von der Ordnung $\frac{1}{h^2}$ und damit ist die Zeitschrittbeschränkung von der Form $\tau = \mathcal{O}(h^2)$, eine meist zu starke Einschränkung. Damit wird auch das explizite Euler Verfahren weniger attraktiv, denn es ist zwar jeder Zeitschritt sehr einfach durchzuführen, man benötigt aber wegen der kleinen Wahl von τ sehr viele Zeitschritte im Vergleich zu einem impliziten oder zum Crank-Nicholson Verfahren.

L^∞-Stabilität

Eine speziell für Ortsdiskretisierungen mit finiten Differenzen sehr interessante Untersuchung ist die Stabilität in der Supremumsnorm, die wir ja auch im elliptischen Fall durchgeführt haben. Wir beschränken uns dabei der Einfachheit halber auf $\theta = 1$ und $\theta = 0$, d.h. explizites und implizites Euler-Verfahren.

Wir nehmen an, dass die Ortsdiskretisierung wie in Kapitel 2 eine M-Matrix K_h liefert. Damit ist natürlich auch die Matrix $A_h = I + \tau K_h$ eine M-Matrix (die Nebendiagonalterme bleiben gleich und negativ, die Diagonalterme werden sogar grösser). Also gilt für das implizite Euler Verfahren

$$U_{k+1}^{h,\tau} = A_h^{-1} U_k^{h,\tau} + \tau A_h^{-1} F_h(t_{k+1}).$$

Wegen der M-Matrix Eigenschaft gilt ein Maximumprinzip und damit Stabilität für A_h^{-1}:

$$\|U_{k+1}^{h,\tau}\|_\infty \leq \|A_h^{-1}\|(\|U_k^{h,\tau}\|_\infty + \tau\|F_h(t_{k+1})\|_\infty),$$

wobei für A_h^{-1} als passende Norm zur Supremumsnorm der Vektoren die Zeilensummennorm

$$\|B\| = \max_i \sum_j |B_{ij}|$$

verwendet wird. Die spezielle Gestalt der Matrix A_h als summe einer M-Matrix und der Einheitsmatrix liefert aber sogar noch mehr, nämlich die Konstante eins für $\|A_h^{-1}\|$ in der Stabilitätsabschätzung, wie wir im folgenden Lemma sehen werden:

Lemma 4.2. *Sei $B = (I + C)^{-1}$ für eine M-Matrix $C \in \mathbb{R}^{n\times n}$. Dann gilt*

$$\|B\| = \max_i \sum_j |B_{ij}| \leq 1.$$

Proof. Sei $G \in \mathbb{R}^n$ beliebig und $V = BG$, also $(I + C)V = G$. Wir nehmen an es gelte

$$V_\ell > G_i \quad \text{und} \quad V_\ell \geq V_i$$

für alle $i \in \{1, \ldots, n\}$. Dann folgt aus der $\ell-$ten Zeile des Gleichungssystems

$$(I + C_{\ell\ell})V_\ell = G_\ell - \sum_{j\neq\ell} C_{j\ell}V_j.$$

Wegen der Nichtpositivität der Nebendiagonalelemente in der M-Matrix $C_{j\ell}$ und der Eigenschaften der Zeilensummen einer M-Matrix folgt

$$(I + C_{\ell\ell})V_\ell \leq G_\ell - \sum_{j\neq\ell} C_{j\ell}V_\ell \leq G_\ell + C_{\ell\ell}V_\ell.$$

Also folgt $V_\ell \leq G_\ell$, ein Widerspruch zur Annahme $V_\ell > G_\ell$. Damit muss gelten

$$\max_i V_i \leq \max_i G_i, \qquad \min_i V_i \geq \min G_i,$$

wobei man die zweite Ungleichung durch Anwendung des selben Arguments auf $-G$ und $-V$ erhält. Insgesamt ist also

$$\|G\|_\infty = \max_i |G_i| \geq \max_i |V_i| = \max_i |(BG)_i| = \max_i |\sum_j B_{ij}G_j|.$$

Diese Ungleichung gilt für beliebiges $G \in \mathbb{R}^n$, also wählen wir G so, dass $G_k = \operatorname{sign}(B_{kj})$ gilt, wobei der Index k so gewählt ist, dass

$$\sum_j |B_{kj}| = \max_i \sum_j |B_{ij}|.$$

Somit gilt

$$1 \geq \|G\|_\infty \geq \max_i |\sum_j B_{ij}G_j| \geq |\sum_j B_{kj}G_j| = \sum_j |B_{kj}| = \max_i \sum_j |B_{ij}|$$

und wir erhalten die gewünschte Abschätzung. □

Aus dieser Eigenschaft der Matrix A_h folgt die Abschätzung

$$\|U_{k+1}^{h,\tau}\|_\infty \leq \|U_k^{h,\tau}\|_\infty + \tau \|F_h(t_{k+1})\|_\infty.$$

Summieren wir über k in dieser Ungleichung ($k = 0, \ldots, K-1$), dann folgt

$$\|U_K^{h,\tau}\|_\infty \leq \|U_0^{h,\tau}\|_\infty + \tau \sum_{k=1}^{K} \|F_h(t_k)\|_\infty.$$

Die Norm der Lösung lässt sich also wieder durch die Norm von der Daten (Anfangswert und rechte Seite) abschätzen. Für $h \to 0$ und $\tau \to 0$ sollte (unter geeigneten Annahmen and u_0 und f) gelten

$$\|U_0^{h,\tau}\|_\infty \to \|u_0\|\infty$$

und

$$\tau \sum_{k=1}^{K} \|F_h(t_k)\|_\infty \to \int_0^T \|f(t)\|_\infty \, dt,$$

mit $T = \lim K\tau$. Also wird die Stabilitätsabschätzung auch unabhängig von h und τ.

Für das explizite Euler Verfahren gilt

$$U_{k+1}^{h,\tau} = (I - \tau K_h) U_k^{h,\tau} + \tau F_h(t_k).$$

Die Nebendiagonalelemente der Matrix $I - \tau K_h$ sind nun nichtnegativ, die Diagonalelemente nur unter der Bedingung $\tau(K_h)_{jj} \leq 1$. In diesem Fall ist die Zeilensummennorm

$$\|I - \tau K_h\|_\infty = \max_i \sum_j |(I - \tau K_h)_{jj}| = 1 - \tau \min_i \sum_j (K_h)_{jj} \leq 1.$$

Also folgt Stabilität in der Supremumsnorm, es gilt

$$\|U_{k+1}^{h,\tau}\|_\infty \leq \|U_k^{h,\tau}\|_\infty + \tau \|F_h(t_k)\|_\infty.$$

Beachtet man, dass $(K_h)_{jj} = \mathcal{O}(\frac{1}{h^2})$ gilt, so folgt wieder bedingte Stabilität mit Zeitschritten $\tau = \mathcal{O}(h^2)$.

H^1-Stabilität

Wir erinnern noch einmal an das allgemeine Verfahren im Fall einer finite Elemente Diskretisierung im Ort.

$$M_h U_{k+1}^{h,\tau} + \theta\tau K_h U_{k+1}^{h,\tau} = M_h U_k^{h,\tau} - (1-\theta)\tau K_h U_k^{h,\tau} + \tau((1-\theta)F_h(t_k) + \theta F_h(t_{k+1}))). \quad (4.12)$$

Hier ist M_h die Massenmatrix und K_h die Steifigkeitsmatrix, d.h.

$$\begin{aligned} (M_h)_{ij} &= \langle \varphi_i, \varphi_j \rangle = \int_\Omega \varphi_i \varphi_j \, dx \\ (K_h)_{ij} &= B(\varphi_i, \varphi_j) \end{aligned}$$

mit der selben Bilinearform B resultierend aus der schwachen Formulierung wie im elliptischen Fall. Die rechte Seite berechnet sich wieder aus

$$F_h(t) = (\ell(t; \varphi))_{i=1,\ldots,n}.$$

Man beachte, dass man auch leicht zeitabhängigkeit von M_h und K_h hier einbauen könnte (etwa wegen zeitlich veränderlicher Koeffizienten in der parabolischen Differentialgleichunge), was wir der Einfachheit halber aber ignorieren.

Wie schon im Fall elliptischer Gleichungen kann man auch zeitdiskretisierte parabolische Probleme der Form (4.12) in eine schwache Formulierung überführen und Variationsprinzipien herleiten. Wir definieren dazu wieder die Funktion

$$u_k^{h,\tau} := \sum_j (U_k^{h,\tau})_j \varphi_j. \tag{4.13}$$

Dann gilt

$$(M_h U_k^{h,\tau})_i = \sum_j (M_h)_{ij} (U_k^{h,\tau})_j = \sum_j \langle \varphi_i, \varphi_j \rangle (U_k^{h,\tau})_j = \langle \varphi_i, u_k^{h,\tau} \rangle,$$

und

$$(K_h U_k^{h,\tau})_i = \sum_j (K_h)_{ij} (U_k^{h,\tau})_j = \sum_j B(\varphi_i, \varphi_j)(U_k^{h,\tau})_j = B(\varphi_i, u_k^{h,\tau}).$$

Mit der obigen Definition der rechten Seite (und unter Ausnutzung der Symmetrie des Skalarprodukts und der Bilinearform) können wir also (4.12) äquivalent als

$$\langle u_{k+1}^{h,\tau}, \varphi_i \rangle + \theta\tau B(u_{k+1}^{h,\tau}, \varphi_i) = \langle u_k^{h,\tau}, \varphi_i \rangle - (1-\theta)\tau B(u_k^{h,\tau}, \varphi_i) + \tau((1-\theta)\ell(t_k; \varphi_i) + \theta\ell(t_{k+1}; \varphi_i)),$$

für $i = 1, \ldots, n$ schreiben. Da die Funktionen φ_i eine Basis von X^h bilden, gilt dann wieder

$$\langle u_{k+1}^{h,\tau}, v \rangle + \theta\tau B(u_{k+1}^{h,\tau}, v) = \langle u_k^{h,\tau}, v \rangle - (1-\theta)\tau B(u_k^{h,\tau}, v) + \tau((1-\theta)\ell(t_k; v) + \theta\ell(t_{k+1}; v))$$

für alle $v \in X^h$.

Zur Vereinfachung werden wir bei der folgenden Stabilitätsuntersuchung annehmen, dass ℓ nicht von der Zeit abhängt. In diesem Fall hat die parabolische Gleichung eine dissipative Struktur, da das Energiefunktional

$$E(u) = \frac{1}{2} B(u, u) - \ell(u)$$

in der Zeit nicht zunimmt. Um dies (formal) zu sehen verwendet man in der schwachen Form $v = \frac{\partial u}{\partial t}$ und erhält

$$0 = \langle \frac{\partial u}{\partial t}, \frac{\partial u}{\partial t} \rangle + B(u, \frac{\partial u}{\partial t}) - \ell(\frac{\partial u}{\partial t}) = \| \frac{\partial u}{\partial t} \|^2 + \frac{dt}{dE}(u).$$

Mit Integration über die Zeit liefert

$$E(u(S)) \leq \int_s^S \| \frac{\partial u}{\partial t} \|^2 \, dt + E(u(S)) = E(u(s)),$$

d.h. das Energiefunktional nimmt in der Zeit ab, solange $\frac{\partial u}{\partial t} \neq 0$ gilt.

In Analogie zum kontinuierlichen Fall verwenden wir auch in der schwachen Formulierung des diskretisierten Problems die zeitliche Veränderung $v = u_{k+1}^{h,\tau} - u_k^{h,\tau}$ als Testfunktion (zur Vereinfachung der Notation lassen wir den Index h, τ im folgenden weg):

$$\langle v, v \rangle + \theta\tau B(u_{k+1}, v) = -(1-\theta)\tau B(u_k, v) + \tau\ell(v).$$

Einsetzen der Definition von v liefert

$$\|v\|^2 + \theta\tau B(u_{k+1}, u_{k+1}) - (1-\theta)\tau B(u_k, u_k)) + (2\theta - 1)B(u_{k+1}, u_k) - \tau\ell(u_{k+1}) + \tau\ell(u_k) = 0.$$

Eine einfache Umstellung der Terme und Division durch τ liefert

$$\begin{aligned}\frac{1}{\tau}\|v\|^2 + E(u_{k+1}) - E(u_k) &= -(\theta - \frac{1}{2})(B(u_{k+1}, u_{k+1}) - 2B(u_{k+1}, u_k) + B(u_k, u_k)) + 2(\theta - 1)B(u_k, u_k) \\ &= -(\theta - \frac{1}{2})B(v, v) + 2(\theta - 1)B(u_k, u_k).\end{aligned}$$

Wegen der Koerzivität folgt $B(v,v) \geq c\|v\|_H^2 \geq 0$ und $B(u_k, u_k) \geq c\|u_k\|_H^2 \geq 0$, wobei $\|.\|_H$ die Norm des entsprechenden Sobolev-Raumes bezeichnet (etwa $\|.\|_{1,2}$ für einen Differentialoperator zweiter Ordnung). Wegen $\theta \leq 1$ folgt

$$\frac{1}{\tau}\|v\|^2 + E(u_{k+1}) \leq E(u_k) - (\theta - \frac{1}{2})B(v, v)$$

und für $\theta \geq \frac{1}{2}$ erhalten wir direkt die "Energiestabilität"

$$E(u_{k+1}) \leq \|v\|^2 + E(u_{k+1}) \leq E(u_k).$$

Für $\theta < \frac{1}{2}$ sind wieder zusätzliche Bedingungen nötig, um Stabilität zu erhalten. In diesem Fall kann man die Stetigkeit der Bilinearform für die Abschätzung (mit $\eta := \frac{1}{2} - \theta > 0$)

$$\frac{1}{\tau}\|v\|^2 + E(u_{k+1}) \leq E(u_k) + \eta B(v, v) \leq E(u_k) + \eta C\|v\|_H^2.$$

Energiestabilität folgt, falls

$$\|v\|_H^2 \frac{1}{\eta C\tau}\|v\|^2,$$

also wieder eine inverse Ungleichung gilt. In dem in Kapitel 2 untersuchten Fall (Differentialoperator zweiter Ordnung, Sobolevraum $H_0^1(\Omega)$ und stückweise lineare Ansatzfunktionen) gilt eine inverse Ungleichung der Form

$$\|v\|_H^2 \frac{\beta^2}{h^2}\|v\|^2,$$

also folgt die Energiestabilität mit der schon bekannten Bedingung

$$\tau \leq \frac{1}{\eta C\beta^2}h^2 = \mathcal{O}(h^2).$$

Abschließend betrachten wir die Energiestabilität noch etwas näher. Aus dem Nichtanstieg des Zielfunktionals folgt

$$E(u_k) \leq E(u_{k-1}) \leq \ldots E(u_0),$$

und damit

$$c\|u_k\|_H^2 \leq B(u_k, u_k) \leq 2E(u_0) + 2\ell(u_k) \leq 2E(u_0) + 2\|\ell\|\|u_k\|_H$$

bzw. äquivalent

$$c(\|u_k\|_H - \frac{1}{c}\|\ell\|)^2 \leq 2E(u_0) + \frac{1}{c}\|\ell\|^2.$$

Damit erhalten wir wieder eine Stabilitätsabschätzung

$$\|u_k\|_H \leq \frac{1}{c}\|\ell\| + \sqrt{\frac{2c}{E}(u_0) + \frac{1}{c^2}\|\ell\|^2},$$

d.h. wieder eine uniforme Abschätzung der Norm von u_k abhängig von den Daten (Anfangswert und Linearform).

4.1.7 Hyperbolische Probleme

In diesem Abschnitt diskutieren wir noch kurz die Zeitdiskretisierung hyperbolischer Probleme. Der Einfachheit halber beschränken wir uns auf den örtlich eindimensionalen Fall. Wir untersuchen hyperbolische Systeme erster Ordnung in der Form

$$\frac{\partial u}{\partial t} = C\frac{\partial u}{\partial x},$$

wobei nun $u : \Omega \times \mathbb{R}_+ \to \mathbb{R}^p$ und $C : \Omega \times \mathbb{R}_+ \to \mathbb{R}^{p\times p}$ symmetrisch ist. Eine Transformation hyperbolischer Gleichungen auf eine solche Form ist in grosser Allgemeinheit möglich. Beispiele dafür sind die in Kapitel 1 untersuchte Transportgleichung, die einfach dem skalaren Fall des obigen Systems entspricht, oder die Wellengleichung

$$\frac{\partial^2 v}{\partial t^2} = c^2\frac{\partial^2 v}{\partial x^2},$$

mit einer positiven Wellenzahl c. Im letzeren Fall erhält man die Transformation auf ein System erster Ordnung mit den Variablen $u^1 = \frac{\partial v}{\partial t}$ und $u^2 = c\frac{\partial v}{\partial x}$, es gilt dann

$$\begin{aligned}\frac{\partial u^1}{\partial t} &= \frac{\partial^2 v}{\partial t^2} = c^2\frac{\partial^2 v}{\partial x^2} = c\frac{\partial u^2}{\partial x}\\ \frac{\partial u^2}{\partial t} &= c\frac{\partial^2 v}{\partial xt} = c\frac{\partial u^1}{\partial x}.\end{aligned}$$

Die Matrix C hat für die Wellengleichung also die Form

$$C = \begin{pmatrix} 0 & c\\ c & 0\end{pmatrix}.$$

C ist eine symmetrische Matrix mit den Eigenwerten $\pm c$.

Durch geeignete Wahl einseitiger Differenzenquotienten im Ort (siehe auch die Diskussion zur Transportgleichung in Kapitel 1) kann man eine Ortsdiskretisierung der Form

$$\frac{dU^h}{dt} = L_h U^h$$

mit Knotenwerten $U^h \in \mathbb{R}^{np}$ und einer schief-symmetrischen Matrix $L_h \in \mathbb{R}^{np\times np}$ erreichen, d.h. $(L_h)_{ij} = -(L_h)_{ji}$ bzw. $L_h^T = -L_h$. Wir illustrieren dies für den Fall der Wellengleichung und nehmen o.B.d.A. an, der Vektor U^h bestehe zuerst aus den n Gitterwerten von u^1 und danach den n Gitterwerten von u^2. Sei $v_j(t) = (U^h(t))_j$ und $w_j(t) = (U^h(t))_j$. Wir wählen für die Ortsableitung von v einen Vorwärts- und für jene von w einen Rückwärtsdifferenzenquotienten. Dann gilt

$$\frac{dv_j}{dt} = c(w_j - w_{j-1}), \qquad \frac{dw_j}{dt} = c(v_{j+1} - v_j).$$

Die Matrix L_h hat dann die Form

$$L_h = \begin{pmatrix} \mathbf{0} & c - D_+^T\\ cD_+ & \mathbf{0}\end{pmatrix},$$

wobei D_+ die Matrix für den Vorwärtsdifferenzenquotienten (siehe Kapitel 1) ist.

Schiefsymmetrische Matrizen liefern ein Gegenstück zu symmetrischen Matrizen: ihre Eigenwerte sind nicht reell, sondern rein imaginär. Dies sieht man aus der Identität

$$\lambda u^H u = u^H L_h u = -u^H L^H u = -(Lu)^H u = -\overline{\lambda} u^H u,$$

für einen Eigenwert λ mit zugehörigem Eigenvektor u. Eine schiefsymmetrische Matrix L_h kann komplex diagonalisiert werden. Mit einer Matrix Q (sodass $Q^H Q = QQ^H = I$) und einer Diagonalmatrix $i\Sigma$ bestehend aus den Imaginärwerten Eigenwerten λ_j gilt:

$$L_h = Q\Sigma Q^H.$$

Also können wir mit der Transformation $V^h = Q^H U^h$ das ortsdiskrete hyperbolische Problem zu

$$\frac{dV^h}{dt} = \Sigma V^h$$

diagonalisieren, und damit erfüllt jede Komponente die skalare Differentialgleichung

$$\frac{dV_j^h}{dt} = i\sigma_j V_j^h,$$

deren Lösung durch

$$V_j^h(t) = e^{i\sigma_j t})V_j^h(0)$$

gegeben ist. Wegen $|e^{i\sigma_j t}| = 1$ (dies folgt aus der Euler-Formel $e^{ia} = \cos a + i \sin a$) gilt $|V_j^h(t)| = |V_j^h(0)|$ und damit auch

$$|U^h(t)| = |V^h(t)| = |V^h(0)| = |U^h(0)|,$$

d.h. Erhaltung der Euklidischen Norm von U^h in der Zeit.

Zur Zeitdiskretisierung betrachten wir nun explizites und implizites Euler-Verfahren. Beim impliziten Euler-Verfahren berechnen wir einen Zeitschritt aus

$$U_{k+1}^h - \tau L_h U_{k+1}^h = U_k^h.$$

Die Diagonalisierung liefert dann

$$V_{k+1}^h - \tau i\Sigma V_{k+1}^h = V_k^h.$$

Komponentenweise erhalten wir

$$(V_{k+1}^h)_j = \frac{1}{1 - i\tau\sigma_j}(V_k^h)_j.$$

Wegen $|1 - i\tau\sigma_j| = \sqrt{1 + \tau^2\sigma_j^2} \geq 1$ folgt $|(V_{k+1}^h)_j| \leq |(V_k^h)_j|$ und daraus resultiert wieder die undbedingte Stabilität des Verfahrens.

Beim expliziten Euler-Verfahren berechnen wir einen Zeitschritt aus

$$U_{k+1}^h = U_k^h + \tau L_h U_k^h.$$

In diesem Fall ist nach der Diagonalisierung

$$V_{k+1}^h = V_k^h + \tau i\Sigma V_k^h,$$

komponentenweise erhalten wir

$$(V_{k+1}^h)_j = (1 + i\tau\sigma_j)(V_k^h)_j.$$

Der Verstärkungsfaktor ist in diesem Fall $1+i\tau\sigma_j$ und für $\sigma_j \neq 0$ gilt $|1+i\tau\sigma_j| = \sqrt{1+\sigma_j^2} > 1$ und damit wird dieses explizite Verfahren instabil. Bei einem System aus mehreren Gleichungen haben wir aber auch noch andere Möglichkeiten ein explizites Verfahren zu konstruieren. Wir schreiben dazu die Matrix L_h als Summe einer rechten oberen und einer linken unteren Dreiecksmatrix, d.h. (wegen der Schiefsymmetrie)

$$L_h = R_h - R_h^T.$$

Nun können wir ein explizites Verfahren auch als

$$U_{k+1}^h + \tau R_h^T U_{k+1}^h = U_k^h + \tau R_h U_k^h.$$

Man beachte, dass man in diesem Fall keine Matrix zu invertieren muss, denn R^T ist eine untere Dreiecksmatrix. Man kann also schrittweise die Komponenten von U_{k+1}^h explizit berechnen. Die Update-Formel kann als

$$U_{k+1}^h = (I + \tau R_h^T)^{-1}(I + \tau R_h)U_k^h$$

geschrieben werden, und für die Stabilität untersuchen wir wieder die Eigenwerte der Matrix $(I + \tau R_h^T)^{-1}(I + \tau R_h)$. Sei λ Eigenwert dieser Matrix mit Eigenvektor V, dann gilt

$$\lambda(V + \tau R_h^T V) = V + \tau R_h V.$$

Wir untersuchen nun der Einfachheit halber wieder den Fall der Wellengleichung, in dem wir $V = (V_1, V_2)$ mit $V_i \in \mathbb{R}^n$ schreiben können und

$$R_h = \begin{pmatrix} 0 & cD \\ 0 & 0 \end{pmatrix},$$

mit der Wellenzahl c und der Matrix D für einen einseitigen Differenzenquotienten. Zur einfacheren Notation schreiben wir $C = \tau c D$. Es gilt

$$\begin{pmatrix} I & 0 \\ C^T & I \end{pmatrix}^{-1} = \begin{pmatrix} I & 0 \\ -C^T & I \end{pmatrix}$$

und damit erfüllt die Lösung des Eigenwertproblems die Gleichung

$$\lambda V = \begin{pmatrix} I & C \\ -C^T & I - C^T C \end{pmatrix} V$$

bzw. mit $\mu = \lambda - 1$ folgt $\mu V_1 = C V_2$ und damit aus der zweiten Gleichung

$$\mu^2 V_2 = -(1+\mu)C^T C V_2 = -\lambda C^T C V_2.$$

Damit ist $\frac{\mu^2}{\lambda}$ Eigenwert der negativ definiten Matrix $-C^T C$ und es gilt

$$(\lambda - 1)^2 = -\gamma^2 \lambda$$

für ein $\gamma \in \mathbb{R}$ (Eigenwert von $C^T C$). Diese quadratische Gleichung hat die Lösungen

$$\lambda = \frac{\gamma^2}{2} - 1 \pm \frac{1}{2}\sqrt{\gamma^4 - 4\gamma^2}.$$

Für $\gamma^2 \leq 4$ ist die Wurzel imaginär und damit gilt

$$|\lambda|^2 = (1 - \frac{\gamma^2}{2}) + \gamma^2 - \frac{\gamma^4}{4} = 1,$$

d.h. das Verfahren wird stabil. Die Bedingung $\gamma \leq 2$ kann man wieder zu h und τ in Beziehung setzen, da ja γ^2 Eigenwert von $\tau^2 c^2 D^T D$ ist. Wegen der Skalierung der Matrix für einen Differenzenquotienten ist damit höchstens $\gamma \sim \frac{\tau c}{h}$. Die Stabilitätsbedingung lautet also

$$\tau = \mathcal{O}(\frac{h}{c}),$$

die sogenannte CFL-Bedingung (nach Courant, Friedrichs, Levy). Man erkennt dass diese Bedingung wesentlich weniger restriktiv ist als im parabolischen Fall, weshalb im hyperbolischen Fall bevorzugt explizite Verfahren der obigen Form benutzt werden.

Literaturverzeichnis

[1] D.Braess, *Finite Elemente. Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie*, Springer, Berlin, 1997.

[2] S.C.Brenner, L.R.Scott, *The Mathematical Theory of Finite Elements*, Springer, New York, 1994.

[3] C.Grossmann, H.G.Roos, *Numerik partieller Differentialgleichungen*, 3. Auflage, Teubner, Stuttgart, 2004.

[4] J.L.Lions, E.Magenes, *Non- Homogeneous Boundary Value Problems and Applications I, II*, Springer, Berlin, 1972.

[5] A.Quarteroni, A.Valli, *Numerical Approximation of Partial Differential Equations*, Springer, Berlin, 2002.