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 [1]). 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). □