12 Ausblick auf partielle Differentialgleichungen

12.1 Transportgleichung (hyperbolisches Problem)

Instationäre Transportvorgänge führen auf lineare partielle Differentialgleichungen erster Ordnung der Form

$$\partial_t u + c\partial_x u = 0\,. \tag{12.1.1}$$

Dabei ist $u = u(x,t)$ im einfachsten Fall eine skalare Funktion von Ort und Zeit, welche z. B. die Fortpflanzung einer Störung (Welle) entlang der x-Achse mit Fortpflanzungsgeschwindigkeit c beschreibt. Ein Anwendungsbeispiel ist etwa die Beschreibung von Wellenbewegungen auf der Wasseroberfläche. Die partiellen Ableitungen nach x sowie t werden im Folgenden abgekürzt als $\partial_t u := \partial u/\partial t$, $\partial_x u := \partial u/\partial x$ geschrieben. Analoge Bezeichnungen werden auch für höhere Ableitungen verwendet. Die allgemeine Lösung dieser Transportgleichung hat die Form

$$u(x,t) = \varphi(x - ct)$$

mit einer Funktion $\varphi(\cdot)$, welche zum Zeitpunkt $t = 0$ die Anfangsverteilung $u^0(x) = \varphi(x)$ beschreibt. In konkreten Anwendungen treten Modelle dieser Art in der Regel als nichtlineare (skalare) „Erhaltungsgleichungen“,

$$\partial_t u + \partial_x f(u) = 0\,, \tag{12.1.2}$$

mit konvexen Funktionen $f(\cdot)$ (z. B.: $f(u) = \frac{1}{2}u^2$), sowie als Systeme der Gestalt

$$\partial_t u - c\partial_x v = 0\,, \quad \partial_t v - c\partial_x u = 0\,,$$

auf. Kombination dieser Gleichungen erster Ordnung führt auf eine skalare Gleichung zweiter Ordnung,

$$\partial_t^2 \psi - c^2 \partial_x^2 \psi = 0\,, \tag{12.1.3}$$

der sog. „Wellengleichung“ für eine Funktion $\psi = \psi(x,t)$ mit $u = \partial_t \psi$, $v = \partial_x \psi$. Dies ist der Prototyp einer sog. „hyperbolischen“ Differentialgleichung. Bei einem (linearen) Transportproblem ist die Lösung offenbar durch Vorgabe von *Anfangswerten* zum Zeitpunkt $t = 0$ eindeutig festgelegt. Diese Werte werden entlang der sog. „Charakteristiken“ in der (x,t)-Ebene, $\{(x,t)|\ x - ct = konst.\}$, fortgepflanzt.

Wir unterscheiden die folgenden beiden Problemstellungen:

i) Anfangswertaufgabe (sog. „Cauchy-Problem“):

$u(x,0) = u^0(x)\,,\ -\infty < x < \infty$;

ii) Anfangs-Randwertaufgabe:

$u(x,0) = u^0(x)\,,\ 0 \le x < \infty, \quad u(0,t) = u^1(t)\,,\ 0 \le t < \infty\,.$

12.1.1 Differenzenverfahren

Wir überdecken die (x,t)-Ebene mit einem äquidistanten Punktgitter mit den Gitterweiten h in x-Richtung und k in t-Richtung. In den Gitterpunkten mit den Koordinaten $x_n := nh$ und $t_m := mk$ werden Näherungen $U_n^m \approx u_n^m := u(x_n, t_m)$ als Lösungen von Differenzengleichungen gesucht.

Zur Diskretisierung des reinen Anfangswertproblems kommen nur *explizite* Differenzenformeln in Frage.

i) Differenzenapproximation erster Ordnung:

$$\frac{1}{k}(U_n^m - U_n^{m-1}) + \frac{c}{h}(U_n^{m-1} - U_{n-1}^{m-1}) = 0\,, \tag{12.1.4}$$

mit Anfangswerten $U_n^0 := u^0(x_n)$. Der Abschneidefehler dieses Differenzenschemas hat für eine glatte Lösung (genauer: $u \in C^2(\mathbb{R}^n)$) die Ordnung

$$\tau_n^m := \frac{1}{k}(u_n^m - u_n^{m-1}) + \frac{c}{h}(u_n^{m-1} - u_{n-1}^{m-1}) = O(h+k)\,.$$

Zur Untersuchung der Stabilität dieses Schemas schreiben wir es in der Form

$$U_n^m = (1 - c\sigma)U_n^{m-1} + c\sigma U_{n-1}^{m-1}\,, \quad \sigma := \frac{k}{h}\,.$$

Genau unter der Bedingung $1 - c\sigma \geq 0$ gilt dann

$$\max_n |U_n^m| \leq \max_n |U_n^{m-1}| \leq \ \ldots\ \leq \max_n |U_n^0|\,, \tag{12.1.5}$$

d. h. ist das Verfahren stabil in der Maximumnorm. Die Bedingung für Stabilität lautet ausgeschrieben

$$k \leq c^{-1}h \tag{12.1.6}$$

und wird „Courant-Friedrich-Lewy-Bedingung“ oder auch kurz „CFL-Bedingung“ genannt. Sie ist typisch für explizite Differenzenverfahren speziell bei Transportproblemen bzw. allgemein bei hyperbolischen (und auch bei parabolischen) partiellen Differentialgleichungen. Die CFL-Bedingung bedeutet, dass der Informationsfluss im Differenzenschema, d. h. die Ausbreitungsgeschwindigkeit von Störungen auf dem Rechengitter, nicht schneller sein darf als diejenige im kontinuierlichen Problem.

Mit Hilfe der Stabilitätsaussage (12.1.5) lässt sich nun wieder das schon bekannte Argumentationsschema „Konsistenz + Stabilität $\Rightarrow$ Konvergenz“ realisieren. Die Fehlerfunktion $e_n^m := u_n^m - U_n^m$ genügt der Differenzengleichung

$$e_n^m = (1 - c\sigma)e_n^{m-1} + c\sigma e_{n-1}^{m-1} + k\tau_n^m\,.$$

Unter Annahme des Erfülltseins der CFL-Bedingung folgt dann durch Rekursion über $\mu = m, ..., 1$:

$$\max_n |e_n^m| \leq \max_n |e_n^0| + k\sum_{\mu=1}^{n} |\tau_n^\mu|\,.$$

Aus dieser Beziehung entnehmen wir die folgende a priori Fehlerabschätzung für das obige einfache Differenzenverfahren:

$$\max_n |e_n^m| \leq c(u) t_n \{h + k\}\,, \tag{12.1.7}$$

mit einer Konstante $c(u) \approx \max_{\mathbb{R}^2}\{|\partial_x^2 u| + |\partial_t^2 u|\}$. Man beachte das lineare Anwachsen der Fehlerkonstante mit der Zeit, was eine charakteristische (und unvermeidbare) Eigenschaft von Diskretisierungen von reinen Transportgleichungen ist.

ii) Lax-Wendroff-Schema: Das bisher betrachtete Differenzenschema ist nur von erster Ordnung genau und damit praktisch uninteressant. Eine Verbesserung erhält man durch Umformungen im Abschneidefehler,

$$\frac{1}{k}(u_n^m - u_n^{m-1}) = (\partial_t u)_n^m + \frac{k}{2}(\partial_t^2 u)_n^m + O(k^2)\,,$$

und nachfolgender Ausnutzung der Beziehungen $\partial_t u = -c\partial_x u$ und $\partial_t^2 u = c^2 \partial_x^2 u$,

$$\frac{1}{k}(u_n^m - u_n^{m-1}) = c(\partial_x u)_n^m + \frac{k}{2} c^2 (\partial_x^2 u)_n^m + O(k^2)\,.$$

Die Ortsableitungen werden nun durch zentrale Differenzenquotienten zweiter Ordnung diskretisiert:

$$\frac{1}{k}(u_n^m - u_n^{m-1}) = \frac{c}{2h}(u_{n+1}^m - u_{n-1}^m) + \frac{kc^2}{2h^2}(u_{n+1}^m - 2u_n^m + u_{n-1}^m) + O(h^2 + k^2)\,.$$

Diese Umformung setzt voraus, dass die Lösung beschränkte dritte Ableitungen in Ort und Zeit besitzt. Das so abgeleitete Differenzenschema (sog. „Lax-Wendroff-Schema“)

$$\frac{1}{k}(U_n^m - U_n^{m-1}) - \frac{c}{2h}(U_{n+1}^m - U_{n-1}^m) - \frac{kc^2}{2h^2}(U_{n+1}^m - 2U_n^m + U_{n-1}^m) = 0 \tag{12.1.8}$$

hat dann konstruktionsgemäß einen Abschneidefehler der Ordnung

$$\tau_n^m = O(h^2 + k^2)\,.$$

Umschreiben von (12.1.8) ergibt

$$U_n^m = (1 - c^2\sigma^2) U_n^{m-1} - \tfrac{1}{2} c\sigma(1 - c\sigma) U_{n+1}^{m-1} + \tfrac{1}{2} c\sigma(1 + c\sigma) U_{n-1}^{m-1}\,,$$

wieder mit der Abkürzung $\sigma := k/h$. Die Stabilität dieses Verfahrens lässt sich leider nicht mehr mit einem so einfachen Argument wie eben erschließen. Stattdessen bedient man sich einer Technik der Fourier-Analyse, die auf J. von Neumann zurückgeht. Diese liefert dann Stabilität im L^2-Sinne. Die Lösung der Differenzengleichung erlaubt zu jedem Zeitpunkt t_m auf dem äquidistanten Ortsgitter $\{x_n | \, n = 0, \pm 1, \pm 2, ...\}$ eine *diskrete* Fourier-Entwicklung der Form

$$U_n^m = \sum_{\nu=0}^{\pm\infty} A_\nu e^{a_\nu t_m} e^{i\nu x_n}\,,$$

mit Koeffizienten $a_\nu \in \mathbb{C}$, welche über die Differenzengleichung (12.1.8) bestimmt sind. Die Koeffizienten $A_\nu \in \mathbb{C}$ sind gerade die Fourier-Koeffizienten der Anfangswerte

$$U_n^0 = \sum_{\nu=0}^{\pm\infty} A_\nu e^{i\nu x_n} ,$$

für die angenommen wird, dass

$$\|U_h^0\|_h^2 := \sum_{\nu=0}^{\pm\infty} |A_\nu|^2 < \infty .$$

Hier steht allgemein $U_h^m = (U_n^m)_n$ für den (unendlichen) Vektor der Gitterwerte zum Zeitpunkt t_m. Da die Differenzengleichung linear ist, werden durch sie die einzelnen Summanden dieser Entwicklung separat von einem Zeitpunkt zum nächsten fortgepflanzt (mit der Setzung $r := c\sigma$),

$$e^{a_\nu t_m} e^{i\nu x_n} = (1-r^2) e^{a_\nu t_{m-1}} e^{i\nu x_n} - \tfrac{1}{2} r(1-r) e^{a_\nu t_{m-1}} e^{i\nu x_{n+1}} + \tfrac{1}{2} r(1+r) e^{a_\nu t_{m-1}} e^{i\nu x_{n-1}} ,$$

bzw. nach Kürzen von $e^{a_\nu t_{m-1}} e^{i\nu x_n}$,

$$\omega_\nu := e^{a_\nu k} = (1-r^2) - \tfrac{1}{2} r(1-r) e^{i\nu h} + \tfrac{1}{2} r(1+r) e^{-i\nu h} .$$

Unter Beachtung der Beziehungen $\cos 2x = \cos^2 x - \sin^2 x$ und $\sin^2 x + \cos^2 x = 1$ folgt

$$\begin{aligned} \omega_\nu &= (1-r^2) + \tfrac{1}{2} r^2 (\underbrace{e^{i\nu h} + e^{-i\nu h}}_{2\cos \nu h}) - \tfrac{1}{2} r (\underbrace{e^{i\nu h} - e^{-i\nu h}}_{2i \sin \nu h}) \\ &= \{1 - 2r^2 \sin^2(\tfrac{1}{2}\nu h)\} - i\{r \sin(\nu h)\} . \end{aligned}$$

Also ist

$$\begin{aligned} |\omega_\nu|^2 &= 1 - 4r^2 \sin^2(\tfrac{1}{2}\nu h) + 4r^4 \sin^4(\tfrac{1}{2}\nu h) + r^2 \sin^2(\nu h) \\ &= 1 - 4r^2 \sin^2(\tfrac{1}{2}\nu h) + 4r^4 \sin^4(\tfrac{1}{2}\nu h) + 4r^2 \{\sin^2(\tfrac{1}{2}\nu h) - \sin^4(\tfrac{1}{2}\nu h)\} \\ &= 1 - 4r^2(1-r^2) \sin^4(\tfrac{1}{2}\nu h) . \end{aligned}$$

Für $r = c\sigma \leq 1$, d. h. unter der CFL-Bedingung, gilt also $|\omega_\nu| \leq 1$ und damit

$$\|U_h^m\|_h^2 \leq \sum_{\nu=0}^{\pm\infty} |A_\nu|^2 |e^{\alpha t_{m-1}}|^2 = \|U_h^{m-1}\|_h^2 \leq \ldots \leq \sum_{\nu=0}^{\pm\infty} |A_\nu|^2 = \|U_h^0\|_h^2 .$$

Das Lax-Wendroff-Schema ist dann also L^2-stabil. Mit Hilfe einer Erweiterung dieses Arguments kann man damit noch das asymptotische Konvergenzverhalten $O(h^2 + k^2)$ bzgl. der diskreten L^2-Norm zeigen.

iii) Leap-Frog-Schema: Das folgende Differenzenschema

$$\frac{1}{2k}(U_n^m - U_n^{m-2}) + \frac{c}{2h}(U_{n+1}^{m-1} - U_{n-1}^{m-1}) = 0 \qquad (12.1.9)$$

wird aus geometrischen Gründen „Leap-Frog-Schema“ genannt. Es hat ebenfalls die Konsistenzordnung $O(h^2+k^2)$. Ausgehend von der Darstellung

$$U_n^m = U_n^{m-2} - c\sigma U_{n+1}^{m-1} + c\sigma U_{n-1}^{m-1}$$

erhält man mit dem Fourier-Ansatz die Beziehung

$$e^{a_\nu t_m} e^{i\nu x_n} = e^{a_\nu t_{m-2}} e^{i\nu x_n} - c\sigma e^{a_\nu t_{m-1}} e^{i\nu x_{n+1}} + c\sigma e^{a_\nu t_{m-1}} e^{i\nu x_{n-1}}$$

bzw. mit den Abkürzungen von eben:

$$\omega_\nu = \omega_\nu^{-1} - r(e^{i\nu h} - e^{-i\nu h}) = \omega_\nu^{-1} - 2ir\sin(\nu h)\,.$$

Es ergibt sich die quadratische Gleichung $\omega_\nu^2 + 2ir\sin(\nu h)\omega_\nu - 1 = 0$ mit den Wurzeln

$$\omega_\nu = -ir\sin(\nu h) \pm \sqrt{1 - r^2\sin^2(\nu h)}\,.$$

Unter der CFL-Bedingung $r \le 1$ gilt dann wieder $|\omega_\nu| \le 1$.

Zur Diskretisierung des Anfangs-Randwertproblems können auch *implizite* Differenzenformeln verwendet werden, um sich von der lästigen CFL-Bedingung an die Schrittweiten zu befreien.

i) Differenzenapproximation erster Ordnung:

$$\frac{1}{k}(U_n^m - U_n^{m-1}) + \frac{c}{h}(U_n^m - U_{n-1}^m) = 0\,, \tag{12.1.10}$$

mit Anfangswerten $U_n^0 := u^0(x_n)$, $U_0^m = u^1(t_m)$. Der Abschneidefehler dieses Differenzenschemas ist wieder von erster Ordnung in h und k,

$$\tau_n^m = O(h+k)\,.$$

Bei Verwendung der Randbedingungen $U_0^m = u^1(t_m)$ lassen sich alle Werte U_n^m aus der Differenzenformel sukzessiv (d. h. explizit) berechnen. Zur Untersuchung der Stabilität schreiben wir das Schema wieder in der Form

$$(1+c\sigma)U_n^m = c\sigma U_{n-1}^m + U_n^{m-1}\,.$$

Sei $M := \sup_{n\ge 0}|U_n^0|$. Unter der Annahme $|U_\nu^\mu| \le M$ für $0 \le \nu \le n-1$, $0 \le \mu \le m$ sowie für $0 \le \nu \le n$, $0 \le \mu \le m-1$ ist dann auch $|U_n^m| \le M$, und durch Induktion nach n folgt:

$$\sup_{n\ge 0}|U_n^m| \le \max\{\sup_{n\ge 0}|U_n^0|, \sup_{m\ge 0}|U_0^m|\}\,,$$

d. h. die Stabilität des Schemas ohne jede Schrittweitenrestriktion („unbedingte“ Stabilität).

ii) Wendroff-Schema: Ein implizites Differenzenschema *zweiter* Ordnung ist das sog. „Wendroff-Schema“

$$U_n^m = U_{n-1}^{m-1} + \frac{1-c\sigma}{1+c\sigma}(U_n^{m-1} - U_{n-1}^m)\,. \tag{12.1.11}$$

Auch dieses Verfahren ist unbedingt stabil.

12.1.2 Finite-Elemente-Galerkin-Verfahren

Die bisher betrachteten Differenzenapproximationen von Transportproblemen haben den Nachteil, dass sie nur auf sog. „Tensorprodukt-Gittern“ der (x,t)-Ebene definiert sind und somit eine dynamische Anpassung an lokale Lösungsstrukturen (z. B. wandernde Fronten) nur schwer möglich ist. Weiter gibt es für diese Verfahren keinen fundierten Ansatz zur a posteriori Fehlerschätzung und adaptiven Gittersteuerung. Diese Nachteile können durch Galerkin-Verfahren im Orts-Zeit-Raum mit Finite-Elemente-Ansatzfunktionen behoben werden. Dabei erhöht sich aber i. Allg. durch die globalere Kopplung der Unbekannten gegenüber den einfachen, expliziten Differenzenverfahren der rechnerische Aufwand. Wir beschreiben im Folgenden, wie die Idee des „unstetigen Galerkin-Verfahrens“ (hier speziell dG(0)-Verfahren) zur Lösung von Anfangswertaufgaben gewöhnlicher Differentialgleichungen auf Transportprobleme in höheren Dimensionen übertragen werden kann.

Wir schreiben die Transportgleichung (12.1.1) in etwas allgemeinerer Notation als (stationäre) Transportgleichung in der (x_1,x_2)-Ebene

$$b\cdot\nabla u(x)=0\,,\quad x:=(x_1,x_2)^T\in Q\,, \tag{12.1.12}$$

auf einem Quadrat $Q:=\{x\in\mathbb{R}^2|\ 0\le x_i\le 1\,(i=1,2)\}$, mit einem (festen) Richtungsvektor $b=(b_1,b_2)^T$ und dem Gradientenoperator $\nabla=(\partial_1,\partial_2)^T$. Die Transportgleichung (12.1.1) passt in diesen Rahmen mit $b=(1,c)^T$. Die Randbedingungen sind dann gestellt als sog. „Einströmrandbedingungen“

$$u=g\quad\text{auf}\quad\partial Q_-:=\{x\in\partial Q|\ b\cdot n(x)<0\}\,, \tag{12.1.13}$$

mit einer gegebenen Randbelegungsfunktion $g(x)$ und dem nach außen gerichteten Normaleneinheitsvektor $n=(n_1,n_2)^T$ entlang des Randes ∂Q von Q. Der andere Teil des Randes $\partial Q_+:=\partial Q\setminus\partial Q_-$ wird sinngemäß als „Ausströmrand“ bezeichnet.

Ausgangspunkt der Galerkin-Diskretisierung der Transportgleichung (12.1.12) ist wieder deren variationelle Formulierung. Zunächst wird das Lösungsgebiet Q zerlegt in Dreiecke (Zellen) K, wobei zwei Dreiecke dieser Triangulierung T_h jeweils nur eine ganze Seite oder einen Eckpunkt gemeinsam haben (d. h.: Die Triangulierung muss nicht *strukturiert* sein, aber sog. „hängende Knoten“ sind hier nicht erlaubt). Die Gitterweite wird beschrieben durch die Parameter $h_K:=\mathrm{diam}(K)$ sowie $h:=\max_K h_K$. Für jede Zelle K definieren wir ihren Ein- sowie Ausströmrand durch

$$\partial K_-:=\{x\in\partial K|\ b\cdot n(x)<0\}\,,\quad \partial K_+:=\partial K\setminus\partial K_-\,.$$

Bzgl. der Triangulierungen T_h führen wir Finite-Elemente-Ansatzräume bestehend aus stückweise konstanten Funktionen ein:

$$S_h^{(0)}:=\{v_h:Q\to\mathbb{R}|\ v_{h|K}\in P_0(K),\ K\in T_h\}\,.$$

Die Funktionen in $S_h^{(0)}$ sind i. Allg. unstetig über die Zellkanten. Für einen Punkt $x\in\partial K$ führen wir die folgenden Bezeichnungen ein:

$$v^-(x):=\lim_{s\to+0}v(x-sb),\quad v^+(x):=\lim_{s\to+0}v(x+sb),\quad [v]:=v^+-v^-\,.$$

Die diskreten Probleme lauten dann

$$u_h \in S_h^{(0)}: \quad B(u_h, \varphi_h) = 0 \quad \forall \varphi_h \in S_h^{(0)}\,, \tag{12.1.14}$$

mit der (affin)-bilinearen Form

$$B(v,w) := \sum_{K\in T_h} \Big\{ \int_K b\cdot\nabla v\, w\, dx - \int_{\partial K_-} n\cdot b\, [v] w^+\, ds \Big\}\,.$$

Man beachte, dass hier die Einströmrandbedingung so in das Verfahren eingebaut ist, dass implizit $u_h^- = g$ auf ∂Q_- realisiert wird. Die exakte Lösung erfüllt offensichtlich ebenfalls die Galerkin-Gleichung (12.1.14), so dass wir für den Fehler $e = u - u_h$ wieder die folgende Orthogonalitätsbeziehung haben:

$$B(e,\varphi_h) = 0\,, \quad \varphi_h \in S_h^{(0)}\,. \tag{12.1.15}$$

Die Galerkin-Diskretisierung ist stabil bzgl. der natürlichen „Energienorm“

$$\|v\|_b := \Big(\tfrac{1}{2} \sum_{K\in T_h} \int_{\partial K_-} |n\cdot b|\, |[v]|^2\, ds + \tfrac{1}{2} \int_{\partial Q_+} |n\cdot b|\, |v^-|^2\, ds \Big)^{1/2}.$$

Darüberhinaus gilt für jede (z. B. bzgl. T_h) stückweise differenzierbare Funktion v:

$$B(v,v) = \|v\|_b^2 - \tfrac{1}{2} \int_{\partial Q_-} |n\cdot b|\, |v^-|^2\, ds\,,$$

was man leicht durch zellweise partielle Integration erschließt. Da für den betrachteten Sonderfall stückweise konstanter Ansatzfunktionen auf jeder Zelle $b\cdot\nabla v_h \equiv 0$ ist, reduziert sich (12.1.14) auf eine Beziehung für die Zellwerte $U_K := u_{h|K}$

$$U_K = \Big(\int_{\partial K_-} n\cdot b\, ds \Big)^{-1} \int_{\partial K_-} n\cdot b\, u_h^-\, ds\,, \quad K\in T_h\,. \tag{12.1.16}$$

Dieses lokal gekoppelte System von Gleichungen kann wieder (wie beim impliziten Differenzenverfahren) ausgehend vom Einströmrand sukzessiv aufgelöst werden. Diese Galerkin-Diskretisierung stellt eine Verallgemeinerung des einfachen Upwind-Differenzenverfahrens (1. Ordnung) auf kartesischen Tensorproduktgittern für allgemeine, unstrukturierte Triangulierungen dar. Hierfür gilt die folgende Konvergenzaussage:

Satz 12.1 (DG-Verfahren): *Besitzt die Lösung des Transportproblems (12.1.12) quadratintegrable erste Ableitungen, so gilt für die unstetige Galerkin-Methode (12.1.14) die a priori Fehlerabschätzung*

$$\|u - u_h\|_b \le c(u)\, h^{1/2}\,, \tag{12.1.17}$$

mit einer Konstanten

$$c(u) \approx \|\nabla u\|_Q := \Big(\int_Q |\nabla u|^2\, dx \Big)^{1/2}.$$

Beweis: Zu der Lösung u definieren wir eine zellweise Interpolierende $\bar{u}_h \in S_h^{(0)}$ durch die Setzung $\bar{u}_{h|K} := |K|^{-1} \int_K u\, dx$. Für diese gilt die wohl bekannte Fehlerabschätzung

$$\|u - \bar{u}_h\|_K + \Big(\int_{\partial K_-} |b \cdot n|\, |(u - \bar{u}_h)^+|^2\, ds \Big)^{1/2} \le c_i\, h_K \|\nabla u\|_K\,. \tag{12.1.18}$$

Mit Hilfe der Galerkin-Orthogonalität (12.1.15) und unter Beachtung von $u_h^- = g$ auf ∂Q_- erschließen wir für den Fehler $e := u - u_h$:

$$\|e\|_b^2 = B(e,e) = B(e, u - \bar{u}_h)\,.$$

Da auf jeder Zelle $b \cdot \nabla u_h \equiv 0$, folgt mit Hilfe der Cauchyschen Ungleichung

$$\|e\|_b^2 \le \|b \cdot \nabla u\|_Q \|u - \bar{u}_h\|_Q + A \cdot B\,,$$

wobei

$$A := \Big(\sum_{K \in T_h} \int_{\partial K_-} |b \cdot n| [e]^2\, ds \Big)^{1/2}, \quad B := \Big(\sum_{K \in T_h} \int_{\partial K_-} |b \cdot n|\, |(u - \bar{u}_h)^+|^2\, ds \Big)^{1/2}.$$

Unter Beachtung von $A \le \|e\|_b$ und der Interpolationsabschätzung (12.1.18) ergibt sich hieraus die Behauptung. Q.E.D.

12.2 Wärmeleitungsgleichung (parabolisches Problem)

Wir betrachten die partielle Differentialgleichung

$$\partial_t u - a \partial_x^2 u = 0 \tag{12.2.19}$$

für Funktionen $u = u(x,t)$ mit Argumenten $x \in I := [0,1]$, $t \ge 0$. Diese Gleichung beschreibt z. B. die Ausbreitung von Temperatur in einem wärmeleitenden Draht (daher auch der Name „Wärmeleitungsgleichung"). Hierbei handelt es sich i. Allg. um ein Anfangs-Randwertproblem, d. h.: Es werden Anfangsbedingungen und Randbedingungen gestellt:

$$u(x,0) = u^0(x),\ x \in I, \quad u(0,t) = u(1,t) = 0,\ t \ge 0\,.$$

Die Anfangswerte stammen z. B. von einer vorgegebenen Temperaturverteilung im Draht, etwa ein plötzlicher Temperaturimpuls zum Zeitpunkt $t = 0$, während die Dirichletschen Randwerte der Vorgabe eines (unendlich großen) Wärmereservoirs entsprechen, an das die Enden des Drahtes angeschlossen sind. Realitätsnähere Randbedingungen sind die der perfekten Wärmeisolation, welche durch die Beziehungen $\partial_x u(0,t) = \partial_x u(1,t) = 0$ (sog. „Neumannsche Randbedingungen") beschrieben sind. Der Einfachheit halber bleiben wir im Folgenden aber bei den Dirichletschen Randbedingungen. Die Wärmeleitungsgleichung ist der Prototyp einer „parabolischen" Differentialgleichungen. Bei diesen treten im Gegensatz zu den *hyperbolischen* Gleichungen (z. B. der oben betrachteten Transportgleichung oder der Wellengleichung) als charakteristische Richtungen nur die Parallelen zur

x-Achse auf, .: Störungen breiten sich mit unendlich großer Geschwindigkeit $c = \infty$ im Ort aus.

Wir wollen kurz die Existenz von Lösungen der Wärmeleitungsgleichung und ihre Eindeutigkeit diskutieren. Zum Nachweis der Existenz von Lösungen wenden wir die sog. „Methode der Variablenseparation“ an. Für den Separationsansatz $u(x,t) = v(x)\psi(t)$ folgt durch Einsetzen in die Wärmeleitungsgleichung:

$$\psi'(t)v(x) = a\psi(t)v''(x) \quad \Rightarrow \quad \frac{\psi'(t)}{\psi(t)} = a\frac{v''(x)}{v(x)} \equiv konst.\,,$$

für alle Argumente $(x,t) \in Q$. Die Separationsfaktoren $v(x)$ und $\psi(t)$ sind also notwendig Lösungen der eindimensionalen Eigenwertprobleme

$$av''(x) + \lambda v(x) = 0\,,\ x \in I\,, \quad \psi'(t) + \lambda\psi(t) = 0\,,\ t \geq 0\,,$$

unter den Nebenbedingungen $v(0) = v(1) = 0$ bzw. $\psi(0) = 1$ mit Parametern $\lambda \in \mathbb{R}$. Die Eigenwertaufgabe für v besitzt die Lösungen

$$v_j(x) = a_j \sin(j\pi x)\,, \quad \lambda_j = aj^2\pi^2\,, \quad a_j = \Big(\int_I \sin^2(j\pi x)\,dx\Big)^{-1/2}\,, \quad j \in \mathbb{N}\,.$$

Die zugehörigen Lösungen für $\psi(t)$ sind $\psi_j(t) = e^{-\lambda_j t}$. Die Anfangsfunktion besitzt die (verallgemeinerte) Fourier-Entwicklung

$$u^0(x) = \sum_{j=0}^{\infty} u_j^0 v_j(x)\,, \quad u_j^0 = \int_I u^0(x)v_j(x)\,dx\,.$$

Durch Superposition der Einzellösungen für $j \in \mathbb{N}$,

$$u(x,t) := \sum_{j=1}^{\infty} u_j^0 v_j(x)e^{-\lambda_j t}\,,$$

erhalten wir folglich eine Lösung der Wärmeleitungsgleichung, welche den Randbedingungen und insbesondere den Anfangsbedingungen genügt. (Zum Nachweis überprüfe man die Konvergenz der Reihen der jeweils nach x sowie t abgeleiteten Einzellösungen.) dass diese Lösung die einzige ist, belegt das folgene Argument: Für eine reguläre Lösung u multiplizieren wir in (12.2.19) mit u, integrieren über I und danach partiell im Ort:

$$0 = \int_I \partial_t u\, u\, dx - a\int_I \partial_x^2 u\, u\, dx = \frac{1}{2}\frac{d}{dt}\int_I |u|^2\,dx + a\int_I |\partial_x u|^2\,dx\,.$$

Integration über die Zeit liefert

$$\int_I |u|^2\,dx \leq \int_I |u^0|^2\,dx\,, \quad t \geq 0\,.$$

Hieraus ersehen wir erstens, dass zwei (reguläre) Lösungen der Wärmeleitungsgleichung zu denselben Anfangswerten notwendig für alle Zeiten übereinstimmen, und zweitens, dass die somit (eindeutige) Lösung stetig bzgl. der L^2-Norm von den Anfangswerten abhängt.

12.2.1 Diskretisierungsverfahren

Bei der Diskretisierung von instationären, insbesondere parabolischen Problemen gibt es drei grundsätzliche Vorgehensweisen, die im Folgenden nacheinander diskutiert werden.

i) Linienmethode: Die Differentialgleichung wird zunächst im Ort und erst danach bzgl. der Zeit diskretisiert. Sei $0 = x_0 < ... < x_n < ... < x_{N+1} = 1$ wieder ein (zunächst als äquidistant angenommenes) Punktgitter auf dem Ortsbereich $I = [0,1]$ mit Gitterweite $h = (N+1)^{-1}$. Auf diesem Gitter werden Näherungen $U_n(t) \approx u(x_n,t)$ definiert durch Diskretisierung des Ortsoperators in (12.2.19) mit Hilfe des zentralen Differenzenquotienten 2. Ordnung,

$$a\partial_x^2 u(x_n,t) \approx \frac{a}{h^2}\{U_{n-1}(t) - 2U_n(t) + U_{n+1}(t)\}\,.$$

Die Vektorfunktion $U_h(t) = (U_n(t))_{n=1}^N$ genügt dann dem System gewöhnlicher Differentialgleichungen

$$\dot{U}_n(t) - \frac{a}{h^2}\{-U_{n-1}(t) + 2U_n(t) - U_{n+1}(t)\} = 0\,,$$

wobei bei Berücksichtigung der Randbedingungen $U_0 \equiv U_{N+1} \equiv 0$ gesetzt ist. Die Anfangswerte sind naturgemäß $U_n(0) = u^0(x_n)$. In kompakter Schreibweise lautet dies

$$\dot{U_h} + A_h U_h(t) = 0\,,\ t \geq 0, \quad U_h(0) = U^0\,, \tag{12.2.20}$$

mit der $(N \times N)$-Matrix

$$A_h = \frac{a}{h^2}\begin{bmatrix} -2 & 1 & & & 0 \\ 1 & -2 & & & \\ & \ddots & \ddots & \ddots & \\ & & & -2 & 1 \\ 0 & & & 1 & -2 \end{bmatrix}.$$

Diese Matrix hat die (positiven rellen) Eigenwerte und der zugehörigen Eigenvektoren

$$\lambda_n = 2ah^{-2}(1 - \cos(n\pi h))\,, \quad w^{(n)} = (\sin(jn\pi h))_{j=1}^N\,.$$

Für den kleinsten und größten Eigenwert gilt

$$\begin{aligned} \lambda_{\min} &= 2ah^{-2}(1 - \cos(\pi h)) = ah^{-2}(\pi^2 h^2 + O(h^4)) = a\pi^2 + O(h^2)\,, \\ \lambda_{\max} &= 2ah^{-2}(1 - \cos(\pi(1-h)) = 2ah^{-2}(1 + \cos(\pi h)) = 4ah^{-2} + O(1)\,. \end{aligned}$$

Die Spektralkondition von A_h hängt also wie folgt von der Gitterweite ab:

$$\mathrm{cond}_2(A_h) = 4\pi^{-2}h^{-2} \gg 1\,.$$

Das nach Ortsdiskretisierung entstandene System (12.2.20) wird für kleines h zunehmend *steif* mit Steifigkeitsrate $\kappa = O(h^{-2})$.

Bei der weiteren zeitlichen Diskretisierung werden explizite Schemata starken Schrittweitenbeschränkungen unterliegen. Beim expliziten Euler-Schema (Polygonzugmethode) wäre aus Stabilitätsgründen die Schrittweitenbedingung

$$-\lambda_{\max} k \in [-2, 0] \quad \Rightarrow \quad k \leq \frac{1}{2a} h^2$$

einzuhalten. Offensichtlich ist diese Bedingung viel restriktiver als die CFL-Bedingung $k \leq c^{-1}h$ bei der expliziten Zeitdiskretisierung der Transportgleichung (12.1.1). Der formale Vorteil der expliziten Verfahren, dass in den einzelnen Zeitschritten keine Gleichungssysteme zu lösen sind, wird durch die große Zahl von durchzuführenden Zeitschritten (besonders bei Verwendung lokal verfeinerter Ortsgitter) mehr als aufgehoben. Da die Eigenwerte der Systemmatrix A_h alle reell sind, käme zur stabilen Integration des Systems (12.2.20) jede der in Kapitel 4.3 betrachteten A(0)-stabilen Formeln in Frage. Dabei muss aber der hohe numerische Aufwand bei der Durchführung komplizierter impliziter Verfahren hoher Ordnung berücksichtigt werden. Auf der anderen Seite hat das einfache implizite Gegenstück zum expliziten Euler-Schema,

$$(I + kA_h)U_h^m = U_h^{m-1}\,,\ m \geq 1, \quad U_h^0 \approx u^0\,.$$

nur die Ordnung $O(k)$, so dass die zeitliche Genauigkeit i. Allg. nicht gut mit der örtlichen Genauigkeit $O(h^2)$ balanciert ist. Für Wärmeleitungsprobleme ist das Euler-Schema meist zu ungenau und zu stark dämpfend (Man beachte die extreme Struktur des Stabilitätsgebiets dieser Formel.). In der Praxis wird daher zur zeitlichen Diskretisierung solcher Probleme meist die A-stabile Trapezregel verwendet, welche in kompakter Form wie folgt lautet:

$$(I + \tfrac{1}{2}kA_h)U_h^m = (I - \tfrac{1}{2}kA_h)U_h^{m-1}\,,\ m \geq 1, \quad U_h^0 \approx u^0\,. \tag{12.2.21}$$

Dieses Schema findet man in der Literatur unter dem Namen „Crank-Nicolson-Verfahren“. Nach Konstruktion sollte die Konsistenzordnung des resultierenden Gesamtverfahrens $O(h^2 + k^2)$ sein, so dass örtliche und zeitliche Genauigkeit formal balanciert sind. Zur Konvergenzanalyse führen wir mit der Standardnotation $u_n^m := u(x_n, t_m)$ wieder den zugehörigen Abschneidefehler ein:

$$\tau_n^m := k^{-1}(u_n^m - u_n^{m-1}) - \tfrac{1}{2}ah^{-2}(u_{n-1}^m - 2u_n^m + u_{n+1}^m) - \tfrac{1}{2}ah^{-2}(u_{n-1}^{m-1} - 2u_n^{m-1} + u_{n+1}^{m-1})\,.$$

Durch Taylor-Entwicklung erhält man für die einzelnen Bestandteile des Abschneidefehlers die Darstellungen

$$\begin{aligned} k^{-1}(u_n^m - u_n^{m-1}) &= k^{-1}\int_{t_{m-1}}^{t_m} \partial_t u(x,t)\,dt\,, \\ \tfrac{1}{2}ah^{-2}(u_{n-1}^m - 2u_n^m + u_{n+1}^m) &= \tfrac{1}{2}a\partial_x^2 u(x_n,t_m) + \tfrac{1}{24}ah^2\partial_x^4 u(\xi_n,t_m)\,., \end{aligned}$$

und damit

$$\tau_n^m = k^{-1}\int_{t_{m-1}}^{t_m} \partial_t u(x_n,t)\,dt - \tfrac{1}{2}\{\partial_t u(x_n,t_m) + \partial_t u(x_n,t_{m-1})\} - \tfrac{1}{12}ah^2\partial_x^4 u(\xi_n,\eta_m)\,.$$

Auf jeder der Zellen $Q_n^m := [x_{n-1}, x_{n+1}] \times [t_{m-1}, t_m]$ gilt folglich

$$|\tau_n^m| \leq \tfrac{1}{12} k^2 \max_{Q_n^m} |\partial_t^3 u| + \tfrac{1}{12} a h^2 \max_{Q_n^m} |\partial_x^4 u| \,.$$

Zur Abschätzung des (globalen) Diskretisierungsfehlers führen wir für Gitterfunktionen $(v_n)_{n=1}^N$ diskrete Analoga des L^2-Skalarprodukts und der zugehörigen L^2-Norm ein:

$$(v, w)_h := h \sum_{n=1}^{N} v_n w_n \,, \quad \|v\|_h := (v, v)_h^{1/2} \,.$$

Satz 12.2 (Crank-Nicolson-Verfahren): *Das beschriebene Crank-Nicolson-Verfahren hat für hinreichend glatte Lösung u den globalen Diskretisierungsfehler*

$$\max_{1 \leq m \leq n} \|u^n - U_h^n\|_h \leq c(u)\, t_n \{h^2 + k^2\}, \tag{12.2.22}$$

mit einer Konstanten $c(u) \approx \max_Q \{|\partial_t^3 u| + a|\partial_x^4 u|\}$.

Beweis: Wir setzen $e_n^m := u_n^m - U_n^m$ und entsprechend $e^m := (e_n^m)_{n=1}^N$ sowie $\tau^m := (\tau_n^m)_{n=1}^N$. Mit dieser Notation gilt dann

$$k^{-1}(e^m - e^{m-1}) + \tfrac{1}{2} A_h (e^m + e^{m-1}) = \tau^m \,.$$

Multiplikation dieser Identität mit $e^m + e^{m-1}$ und Summation über m ergibt

$$k^{-1}\{\|e^m\|_h^2 - \|e^{m-1}\|_h^2\} + \tfrac{1}{2}(A_h(e^m + e^{m-1}), e^m + e^{m-1})_h = (\tau^m, e^m + e^{m-1})_h \,.$$

Der kleinste Eigenwert von A_h verhält sich wie $\lambda_1 = a\pi^2 + O(h^2)$. Damit erschließen wir

$$k^{-1}\{\|e^m\|_h^2 - \|e^{m-1}\|_h^2\} + \tfrac{1}{2}\lambda_1 \|e^m + e^{m-1}\|_h^2 \leq \tfrac{1}{2}\lambda_1 \|e^m + e^{m-1}\|_h^2 + \tfrac{1}{2}\lambda_1^{-1} \|\tau^m\|_h^2 \,,$$

bzw.

$$\|e^m\|_h^2 \leq \|e^{m-1}\|_h^2 + \tfrac{1}{2}\lambda_1^{-1} k \|\tau^m\|_h^2 \,.$$

Wir summieren nun über $m = n, ..., 1$ und erhalten

$$\|e^m\|_h^2 \leq \|e^0\|_h^2 + \tfrac{1}{2}\lambda_1^{-1} k \sum_{\mu=1}^{m} \|\tau^\mu\|_h^2 \,.$$

Mit $e^0 = 0$ und der obigen Abschätzung für den Abschneidefehler folgt schließlich die Behauptung. Q.E.D.

Die üblichen Einschrittformeln für das autonome System (12.2.20) (mit t-unabhängiger Matrix A_h) lassen sich in kompakter Form schreiben,

$$U_h^m = R(-kA_h) U_h^{m-1} \,,$$

mit rationalen Funktionen

$$R(z) = \frac{P(z)}{Q(z)}\,.$$

Zu den expliziten und impliziten Euler-Verfahren gehören die Funktionen $R(z) = 1 - z$ bzw. $R(z) = (1+z)^{-1}$. Das Crank-Nicolson-Verfahren wird beschrieben durch $R(z) = (1-\frac{1}{2}z)(1+\frac{1}{2}z)^{-1}$. Für die Brauchbarkeit dieser Verfahren für steife Anfangswertaufgaben sind die folgenden Eigenschaften wichtig:

(1) *A-stabil:* $|R(z)| \le 1\,, \quad Re\, z \le 0\,,$

(2) *stark A-stabil:* $|R(z)| < 1\,, \quad Re\, z \to -\infty\,,$

(3) *steif-stabil:* $|R(z)| \to 0\,, \quad Re\, z \to -\infty\,.$

Das von uns favorisierte Crank-Nicolson-Verfahren ist in diesem Sinne zwar A-stabil aber nicht *stark* A-stabil. Dies hat nachteilige Konsequenzen im Fall von irregulären Anfangswerten u^0 (z. B.: lokalen Temperaturspitzen). Die durch diese Anfangsdaten induzierten hochfrequenten Fehleranteile werden durch das Crank-Nicolson-Schema nur unzureichend ausgedämpft, so dass sich ein unphysikalisches Lösungsverhalten zeigen kann. Man beachte, dass der kontinuierliche Differentialoperator stark dämpfend ist:

$$\int_I |u(x,t)|^2\,dx \le e^{-\lambda t} \int_I |u^0(x)|^2\,dx\,, \quad t \ge 0\,,$$

mit dem kleinsten Eigenwert des Ortsoperators, $\lambda = a\pi^2$. Dieses unliebsame Verhalten des Crank-Nicolson-Verfahrens wird vermieden bei Verwendung einer Modifikation des Crank-Nicolson-Verfahrens als ein sog. „Teilschritt-θ-Verfahren“ bestehend aus jeweils drei sukzessiven Teilschritten vom Crank-Nicolson-Typ:

$$\begin{aligned}
(I + \alpha\theta k A_h) U^{m-1+\theta} &= (I - \beta\theta k A_h) U^{m-1} \\
(I + \beta\theta' k A_h) U^{m-\theta} &= (I - \alpha\theta' k A_h) U^{m-1+\theta} \\
(I + \alpha\theta k A_h) U^{m} &= (I - \beta\theta k A_h) U^{m-\theta}
\end{aligned}$$

mit den Parametern $\theta = 1 - \frac{1}{2}\sqrt{2} = 0,292893...$, $\theta' = 1 - 2\theta$ und beliebigen Werten $\alpha \in (\frac{1}{2}, 1]$, $\beta = 1 - \alpha$. Für den speziellen Wert $\alpha = (1-2\theta)(1-\theta)^{-1} = 0,585786...$ ist $\alpha\theta = \beta\theta'$, so dass die zu invertierenden Matrizen in den Teilschritten übereinstimmen. Dieses Verfahren wird beschrieben durch die rationale Funktion

$$R_\theta(z) = \frac{(1+\alpha\theta' z)(1+\beta\theta z)^2}{(1-\alpha\theta z)^2(1-\beta\theta' z)} = e^z + O(|z|^3)\,.$$

Aus dieser Beziehung liest man ab, dass das obige Teilschrittverfahren wie das einfache Crank-Nicolson-Schema von zweiter Ordnung genau ist, und insbesondere dass es *stark* A-stabil ist:

$$|R_\theta(z)| < 1\,,\ Re\, z < 0\,, \quad \lim_{Re\, z \to -\infty} |R_\theta(z)| = \frac{\beta}{\alpha} < 1\,.$$

Dieses Schema hat sich in der Praxis als besonders geeignet zur Behandlung von parabolischen Problemen mit nicht notwendig regulären Daten erwiesen.

ii) Rothe-Methode: Die Differentialgleichung wird zunächst mit einem A-stabilen Verfahren in der Zeit diskretisiert. Bei Verwendung z. B. des impliziten Euler-Schemas ergibt sich eine Folge von stationären Randwertaufgaben (vom Sturm-Liouville-Typ)

$$U^m - ak\, d_x^2 U^m = U^{m-1} + kf^m\,,\ m \geq 1\,,\ U^0 = u^0\,.$$

Diese Probleme werden nun nacheinander auf möglicherweise wechselnden, dem Lösungsverlauf angepassten Ortsgittern diskretisiert. Das Problem ist dabei der adäquate Transfer der jeweiligen Startlösung U^{m-1} vom alten auf das neue Ortsgitter. Hier zeigt sich wieder der systematische Vorteil einer Finite-Elemente-Galerkin-Methode, bei der sich ganz automatisch als *richtige* Wahl die L^2-Projektion von U^{m-1} auf das neue Gitter ergibt. Die theoretische Analyse der Rothe-Methode mit wechselnder Ortsdiskretisierung ist wesentlich schwieriger als die der einfachen Linienmethode und kann im Rahmen dieser kurzen Einführung nicht beschrieben werden.

iii) Globale Diskretisierung: Ähnlich wie bei den Transportproblemen könnte auch bei der Wärmeleitungsgleichung eine simultane Diskretisierung (etwa mit einem Finite-Elemente-Galerkin-Verfahren) auf einem unstrukturierten Gitter der ganzen (x,t)-Ebene erfolgen. Dieser theoretisch durchaus attraktive Ansatz wird aber bei höher dimensionalen Problemen wegen der globalen Kopplung aller Unbekannten zu rechenaufwendig und spielt daher in der Praxis keine Rolle.

12.3 Laplace-Gleichung (elliptisches Problem)

Wir verwenden wieder die Bezeichnung $x := (x_1, x_2)^T$ von oben für Punkte der (x_1, x_2)-Ebene und betrachten auf dem (offenen) Bereich $\Omega = \{x \in \mathbb{R}^2 |\, 0 < x_i < 1\ (i = 1, 2)\}$ die Differentialgleichung

$$-\Delta u(x) := -\partial_1^2 u(x) - \partial_2^2 u(x) = f(x)\,, \quad x \in \Omega\,, \tag{12.3.23}$$

mit dem sog. „Laplace-Operator“ Δ für gegebene rechte Seite $f(x)$. Diese „Poisson-Gleichung“ genannte Differentialgleichung 2. Ordnung ist der Prototyp eines „elliptischen“ Problems. Diese sind dadurch ausgezeichnet, dass es keine charakteristischen Richtungen gibt, d. h.: Störungen breiten sich in alle Richtungen mit unendlicher Geschwindigkeit aus. Entsprechend dürfen (und müssen) analog wie beim eindimensionalen Sturm-Liouville-Problem auch wieder entlang des ganzen Randes $\partial\Omega$ des betrachteten Lösungsgebiets Ω Vorgaben für die Lösung gemacht werden. Wir betrachten hier der Einfachheit halber nur homogene *Dirichletsche Randbedingungen* $u(x) = 0,\ x \in \partial\Omega$ (sog. „1. Randwertproblem des Laplace-Operators“). Die Lösung u beschreibt z. B. die Auslenkung einer (idealisierten) elastischen Membran, die über dem Gebiet Ω horizontal gespannt und mit einer Kraftdichte f vertikal belastet wird. Eine Lösung $u(x)$ ist i. Allg. nicht explizit angebbar, so dass man auf numerische Approximation angewiesen ist. Der Nachweis der Existenz von Lösungen der Poisson-Gleichung kann hier im Rahmen dieser Einführung nicht beschrieben werden. Er ist wesentlich aufwendiger als das entsprechende Argument bei den eindimensionalen Sturm-Liouville-Problemen. Die Eindeutigkeit von Lösungen folgt aber

wieder mit einem einfachen Variationsargument. Seien $u^{(i)}$ $(i = 1, 2)$ zwei Lösungen mit endlicher „potentielle Energie“:

$$E(u^{(i)}) := \tfrac{1}{2}(\nabla u^{(i)}, \nabla u^{(i)})_\Omega - (f, u^{(i)})_\Omega < \infty \,.$$

Hier bezeichnet $(v, w)_\Omega := \int_\Omega v(x)w(x)\,dx$ das L^2-Skalarprodukt über dem Gebiet Ω. Dann gilt für die Differenz $w = u^{(1)} - u^{(2)}$

$$(\nabla w, \nabla w)_\Omega = (-\Delta w, w)_\Omega = 0$$

und folglich wegen der Randbedingung notwendig $w \equiv konst. = 0$. Wir bemerken noch, dass die Lösungen elliptischer Probleme auf Gebieten mit nicht glatten Rändern (wie das hier betrachtete Quadrat) bei den Eckpunkten i. Allg. lokale Irregularitäten, d. h. Singularitäten in höheren Ableitungen, besitzen.

12.3.1 Differenzenverfahren

Die Differenzendiskretisierung des Poisson-Problems (12.3.23) erfolgt analog wie die des Sturm-Liouville-Problems in einer Dimension. Wir überdecken den Bereich $\bar{\Omega}$ wieder mit einem achsen-parallelen Tensorproduktgitter $\overline{\Omega}_h := \{x \in \overline{\Omega} \,|\, x = x_{ij} = (ih, jh)^T,\ (i, j = 0, ..., m + 1)$ mit der konstanten Gitterweite $h = (m + 1)^{-1}$. Die Verwendung derselben festen Gitterweite in alle Ortsrichtungen ist nicht zwingend und erfolgt hier nur der Einfachheit halber. Die $N = m^2$ *inneren* Gitterpunkte, bezeichnet als die Punktmenge Ω_h, werden zeilenweise durchnumeriert.

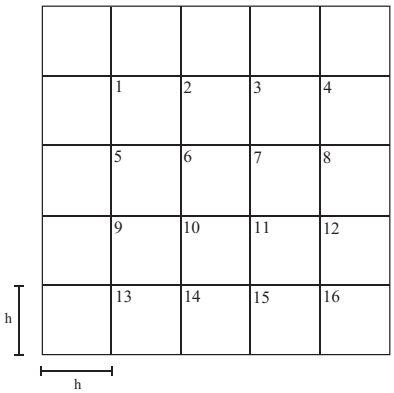

Abbildung 12.1: Diskretisierungsgitter des Modellproblems

Auf dem Gitter $\overline{\Omega}_h$ werden Gitterfunktionen $U_h := (U_{ij})_{i,j=0}^{m+1}$ gesucht als Lösungen der Differenzengleichungen

$$-\Delta_h U_h = F_h \,, \quad x_{ij} \in \Omega_h \,, \tag{12.3.24}$$

mit der Notation

$$(\Delta_h U_h)_{ij} := h^{-2}(4U_{ij} - U_{i+1,j} - U_{i-1,j} - U_{i,j-1} - U_{i,j+1})$$
$$(F_h)_{ij} := f(x_{ij})\,, \quad 1 \le i,j \le m\,.$$

Die geometrische Verteilung der Stützstellen für diesen Differenzenoperator begründen seinen Namen „5-Punkte-Operator“. Entsprechend den gegebenen Randbedingungen werden die Randwerte $U_{0,j} = 0$, $U_{i,0} = 0$ gesetzt. Die Gitterwerte sind dann Approximationen der exakten Lösung, $U_{ij} \approx u(x_{ij})$. Das Differenzenschema (12.3.24) ist äquivalent zu dem linearen Gleichungssystem

$$A_h U_h = b_h \tag{12.3.25}$$

für den Vektor $U_h = (U_{ij})_{i,j=1,\dots,m} \in \mathbb{R}^N$ der unbekannten Knotenwerte. Die Matrix hat die schon bekannte Gestalt (I_m die $m \times m$-Einheitsmatrix)

$$A_h = h^{-2}\left.\begin{bmatrix} B_m & -I_m & & \\ -I_m & B_m & -I_m & \\ & -I_m & B_m & \ddots \\ & & \ddots & \ddots \end{bmatrix}\right\} N \qquad B_m = \left.\begin{bmatrix} 4 & -1 & & \\ -1 & 4 & -1 & \\ & -1 & 4 & \ddots \\ & & \ddots & \ddots \end{bmatrix}\right\} m\,.$$

Die rechte Seite ist bestimmt durch $b_h = (f(x_{11}), \dots, f(x_{mm}))^T$. Die Matrix A_h ist

a) eine dünn besetzte Bandmatrix mit der Bandbreite $2m+1$;

b) symmetrisch, irreduzibel und schwach diagonal dominant;

c) positiv definit und M-Matrix.

Die M-Matrixeigenschaft von A_h erlaubt es, bei der Fehleranalyse für die Differenzenapproximation (12.3.24) analog vorzugehen wie vorher beim Sturm-Liouville-Problem in einer Dimension.

Satz 12.3 (5-Punkte-Schema): *Für die 5-Punkte-Differenzenapproximation (12.3.24) der Poisson-Gleichung (12.3.23) gilt im Falle einer hinreichend glatten Lösung die a priori Fehlerabschätzung*

$$\max_{x_{ij}} |u(x_{ij}) - U_{ij}| \le \tfrac{1}{96} M h^2\,, \tag{12.3.26}$$

mit der Konstante $M := \max_{\overline{\Omega}}\{|\partial_1^4 u| + |\partial_2^4 u|\}$.

Beweis: Der Abschneidefehler der 5-Punkte-Differenzenapproximation genügt der Abschätzung

$$\max_{\overline{\Omega}_h} |\tau_h| := \max_{x_{ij} \in \Omega_h} |f_{ij} - (\Delta_h u)_{ij}| \le \tfrac{1}{12} h^2 \max_{\overline{\Omega}} \{|\partial_1^4 u| + |\partial_2^4 u|\}\,.$$

Hieraus ersehen wir, daß die Differenzenapproximation exakt ist insbesondere für quadratische Polynome. Die Fehlerfunktion des Verfahrens sei mit $e_h := u_h - U_h$ bezeichnet. Aus der M-Matrix-Eigenschaft der zum Differenzenoperator $-\Delta_h$ korrespondierenden Matrix A_h folgt für deren Inverse komponentenweise $A_h^{-1} \geq 0$. Für jede Gitterfunktion v_h implizieren dann die Beziehungen $-\Delta_h v_h \leq 0$ in Ω und $v_h \leq 0$ auf $\partial\Omega_h$ notwendig, dass $v_h \leq 0$ in ganz Ω_h (diskretes „Maximumprinzip“). Wir definieren die quadratische Funktion $w(x) := \frac{1}{48} M h^2 (x_1(1-x_1) + x_2(1-x_2))$ sowie die zugehörige Gitterfunktion w_h. In Gitterpunkten auf dem Rand $\partial\Omega_h$ ist offensichtlich $w_h \geq 0$ und

$$-\Delta_h w_h = -\Delta w = -\tfrac{1}{12} M h^2 \,, \quad x_{ij} \in \Omega_h \,,$$

Dann ist weiter $\pm e_h - w_h \leq 0$ in Punkten auf $\partial\Omega_h$ und

$$-\Delta_h(\pm e_h - w_h) = \pm\tau_h - \tfrac{1}{12} M h^2 \leq 0 \,, \quad x_{ij} \in \Omega_h \,.$$

Folglich muss

$$-w_h \leq e_h \leq w_h$$

auf ganz Ω_h sein. Dies impliziert die behauptete Fehlerabschätzung. Q.E.D.

Das Hauptproblem bei der numerischen Approximation von elliptischen Randwertaufgaben vom Typ (12.3.23) ist die effiziente Lösung der auftretenden großen, linearen Gleichungssysteme. Um dies einzuschätzen, betrachten wir die folgende Modellsituation: Zu der speziellen rechten Seite $f(x) = 2\pi^2 \sin(\pi x_1) \sin(\pi x_2)$ gehört die exakte Lösung

$$u(x) = \sin(\pi x_1) \sin(\pi x_2) \,.$$

Aus der obigen a priori Fehlerabschätzung (12.3.26) entnehmen wir hierfür

$$\max_{x_{ij}} |u(x_{ij}) - U_{ij}| \leq \tfrac{1}{48} \pi^4 h^2 \,.$$

Zur Erzielung einer Genauigkeit von $\varepsilon = 10^{-4}$ (vier Stellen) ist also die Gitterweite

$$h \sim \sqrt{96}\pi^{-2} 10^{-2} \sim 10^{-2} \,.$$

erforderlich. Die Anzahl von Unbekannten im System (12.3.25) ist dann $N \sim 10^4$.

Zur Bestimmung der Konditionierung der zugehörigen Systemmatrix A_h berechnen wir wieder ihre Eigenwerte und Eigenvektoren:

$$\lambda_{kl} = h^{-2}\{4 - 2\,(\cos(kh\pi) + \cos(lh\pi))\} \,, \quad w^{kl} = (\sin(ikh\pi)\sin(jlh\pi))_{i,j=1,\ldots,m} \,.$$

Also ist

$$\begin{aligned} \lambda_{\max} &= h^{-2}\{4 - 4\cos(1-h)\pi\} \;=\; 8h^{-2} + O(1) \\ \lambda_{\min} &= h^{-2}\{4 - 4\cos(h\pi)\} = h^{-2}\{4 - 4(1 - \tfrac{1}{2}\pi^2 h^2)\} + O(h^2)) \;=\; 2\pi^2 + O(h^2) \end{aligned}$$

und somit

$$\mathrm{cond}_2(A_h) \;\approx\; 4\pi^{-2} h^{-2} \,.$$

Für die Gitterweite $h = 10^{-2}$ folgt also $\mathrm{cond}_2(A_h) \approx 4\pi^{-2}10^{-4} \approx 5.000$.

Zur Bestimmung der Lösung des durch Diskretisierung der Randwertaufgabe (12.3.23) entstehenden $(N \times N)$-Gleichungssystems $A_h U_h = b_h$ benötigen die einfachen Fixpunktiterationen (Jacobi- und Gauß-Seidel-Verfahren) $O(N^2)$ Operationen. Als direktes Verfahren erfordert das Cholesky-Verfahren bei Berücksichtigung der speziellen Struktur der Systemmatrix $O(m^2N) = O(N^2)$ Operationen zur Berechnung der Zerlegung $A_h = L_h L_h^T$ und das Vorwärts- und Rückwärtseinsetzen. Dabei ist jedoch zu berücksichtigen, dass letzteres $O(mN) = O(N^{3/2})$ Speicherplätze benötigt im Gegensatz zu den nur $O(N)$ der einfachen Iterationsverfahren. In den letzten Jahren wurden sehr effiziente Verfahren zur Lösung von Problemen des obigen Typs entwickelt (die sog. „Mehrgitter-Verfahren"), die in der Regel die N Unbekannten mit $O(N)$ Operationen auf Diskretisierungsfehlergenauigkeit berechnen.

12.3.2 Finite-Elemente-Galerkin-Verfahren

Das Finite-Elemente-Verfahren zur Approximation der Poisson-Gleichung (12.3.23) sieht formal genauso aus wie beim Sturm-Liouville-Problem in einer Dimension. Ausgangspunkt ist wieder eine variationellen Formulierung des Problems:

$$u \in \bar{V}: \quad (\nabla u, \nabla\varphi)_\Omega = (f, \varphi)_\Omega \quad \forall \varphi \in \bar{V}\,. \tag{12.3.27}$$

Dabei bezeichnet $\bar{V}$ den Raum der Funktionen mit endlicher Energie $E(\cdot)$. Für unsere Zwecke genügt es zu wissen, dass dieser Funktionenraum den folgenden Raum als Teilraum enthält:

$$V := \{v \in C(\bar{\Omega}) |\ v \text{ stückweise stetig differenzierbar, } v_{|\partial\Omega} = 0\}\,.$$

Für Funktionen aus V gilt die mehrdimensionale Poincarésche Ungleichung

$$\|\nabla v\|_\Omega \geq c_\Omega \|v\|_\Omega\,, \quad v \in V\,.$$

Zur Diskretisierung wird der Bereich $\bar{\Omega}$ wieder in Dreiecke K zerlegt, wobei je zwei Dreiecke nur eine ganze Seite oder einem Eckpunkt gemeinsam haben können, d. h. sog. „hängende" Knoten sind hier nicht erlaubt. Die Knotenpunkte dieser Triangulierung werden mit P_i bezeichnet, wobei die Art der Numerierung beliebig ist. Die Triangulierung T_h ist i. Allg. *unstrukturiert*, d. h.: Die Knotenpunkte brauchen keine zeilenweise bzw. spaltenweise Numerierung zu erlauben. Die Feinheit von T_h wird durch die lokale Zellweite $h_K := \mathrm{diam}(K)$ und die globale Gitterweite $h := \max_{K \in T_h} h_K$ beschrieben. Auf der Triangulierung T_h wird der folgende Finite-Elemente-Ansatzraum stückweise linearer Funktionen definiert:

$$V_h := \{v_h \in V |\ v_{h|K} \in P_1(K),\ K \in T_h\}\,.$$

Die approximierenden Probleme lauten dann

$$u_h \in V_h: \quad (\nabla u_h, \nabla\varphi_h)_\Omega = (f, \varphi_h)_\Omega \quad \forall \varphi_h \in V_h\,. \tag{12.3.28}$$

Wie im eindimensionalen Fall wird eine „Knotenbasis“ $\{\varphi_h^{(i)},\ i = 1, ..., N := \dim V_h\}$ von V_h eingeführt (sog. „Lagrange-Basis“), bzgl. derer jede Funktion $v_h \in V_h$ eine Darstellung der Form besitzt

$$v_h = \sum_{i=1}^{N} v_h(P_i)\varphi_h^{(i)} .$$

Bei Verwendung dieser Basis ist das diskrete Problem (12.3.28) wieder äquivalent zu einem linearen Gleichungssystem

$$A_h U_h = b_h$$

für den Vektor $U_h := (u_h(P_i))_{i=1}^N$ der Knotenwerte mit der Systemmatrix (sog. „Steifigkeitsmatrix“) $A_h = (a_{ij})_{ij}^N$ und der rechten Seite (sog. „Lastvektor“) $b_h = (b_j)_{j=1}^N$:

$$a_{ij} := (\nabla\varphi_h^{(j)}, \nabla\varphi_h^{(i)})_\Omega , \qquad b_j := (f, \nabla\varphi_h^{(j)})_\Omega .$$

Die Matrix A_h ist wieder symmetrisch und (im Hinblick auf die Poincarésche Ungleichung) auch positiv definit. Man kann zeigen, dass ihre Spektralkondition sich auch auf allgemeinen Triangulierungen stets wie die des 5-Punkte-Differenzenoperators verhält: $\mathrm{cond}_2(A_h) = O(h^{-2})$. Dabei ist die Potenz h^{-2} weder durch die Raumdimension noch durch den Polynomgrad der Ansatzfunktionen sondern allein durch die Ordnung des Differentialoperators Δ bestimmt.

Mit Hilfe der „Galerkin-Orthogonalität“ für die Fehlerfunktion $e := u - u_h$,

$$(\nabla e, \nabla\varphi_h)_\Omega = 0 , \quad \varphi_h \in V_h ,$$

erschließen wir wieder die Bestapproximationseigenschaft des Galerkin-Verfahrens:

$$\|\nabla e\|_\Omega \le \min_{\varphi_h \in V_h} \|\nabla(u - \varphi_h)\|_\Omega . \tag{12.3.29}$$

Für Funktionen $v \in V$ definieren wir wieder die Knoteninterpolierende $I_h v \in V_h$ durch $I_h v(P_i) = v(P_i),\ (i = 1, ..., N)$. Für den zugehörigen Interpolationsfehler gilt

$$\|\nabla(v - I_h v)\|_K \le c_i h_K \|\nabla^2 v\|_K , \tag{12.3.30}$$

wobei $\nabla^2 v$ den Tensor der zweiten partiellen Ableitungen von v bezeichnet. Durch Kombination der Abschätzungen (12.3.29) und (12.3.30) erhalten wir folgendes Resultat:

Satz 12.4 (FEM): *Für die Finite-Elemente-Galerkin-Methode (12.3.28) mit stückweise linearen Ansatzfunktionen gilt die Konvergenzabschätzung*

$$\|\nabla e\|_\Omega \le c_i h \, \|\nabla^2 u\|_\Omega . \tag{12.3.31}$$

Mit Hilfe eines Dualitätsarguments analog zu dem beim Sturm-Liouville-Problem verwendeten lässt sich auch eine verbesserte L^2-Fehlerabschätzung herleiten,

$$\|e\|_\Omega \le c_i c_s h^2 \, \|\nabla^2 u\|_\Omega . \tag{12.3.32}$$

Dabei ist diesmal allerdings die Abschätzung der Stabilitätskonstante c_s für das duale Problem

$$-\Delta z = \|e\|_\Omega^{-1} e \quad in\ \Omega , \qquad z_{|\partial\Omega} = 0, \tag{12.3.33}$$

wesentlich komplizierter (und i. Allg. auf nicht glattberandeten Gebieten in dieser Form auch gar nicht richtig).