FV/FD-Methoden zur numerischen Lösung partieller Differentialgleichungen

Günter Bärwolff

18. Juli 2017

Skript, geschrieben parallel zur Vorlesung FV/FD-Methoden zur numerischen Lösung partieller Differentialgleichungen im SS2017 an der TU Berlin,

Inhaltsverzeichnis

Kapitel 1

Vorwort

Diese Skript entsteht parallel zur Vorlesung im Sommersemester 2017 und enthält die wesentlichen Inhalte wie z.B. alle Definitionen und Sätze, wobei bei den Beweisen in der Regel nur Verweise auf Textbücher oder Beweisskizzen angegeben werden. Als Lehrbücher seien z.B.

- Grossmann, Roos, Numerische Lösung partieller Differentialgleichungen, Teubner/Springer
- Hackbusch, Theorie und Numerik elliptischer Differentialgleichungen, Selbstverlag
- Braess, Finite Elemente Methode, Springer
- Jüngel, Das kleine FEM-Skript
- Hans R. Schwarz, Finite Elemente Methode
- Dahmen, Reusken, Numerik für Ingenieure und Naturwissenschaftler, Springer
- Günter Bärwollf: Numerik für Ingenieure, Physiker und Informatiker

empfohlen.

Kapitel 2

Beispiele partieller Differentialgleichungen der math. Physik

1. Vorlesung am 18.05.2017

Im Ergebnis der mathematischen Modellierung bzw. Beschreibung von technischen Prozessen oder physikalischen Phänomenen entstehen partielle Differentialgleichungen. Als Beispiel seien hier die Kontinuitätsgleichung als Resultat einer Massenbilanz

$$\frac{\partial \rho}{\partial t} + \operatorname{div}(\rho\, v) = 0 \tag{2.1}$$

und die Navier-Stokes-Gleichung

$$\frac{\partial v}{\partial t} + (v \cdot \nabla)v = -\frac{1}{\rho}\nabla p + \nu[\frac{4}{3}\Delta v - \nabla \times (\nabla \times v)] + F \tag{2.2}$$

als Ergebnis der Bilanzierung des Impulses genannt. Die Differentiationen in der Gleichung sind dabei auf alle Komponenten des Vektorfeldes v anzuwenden und (2.2) besteht aus 3 skalaren Gleichungen für die 3 Geschwindigkeitskomponenten. Die Funktionen bzw. Vektorfelder

$$\rho : [0,T] \times \Omega \to \mathbb{R}, \quad p : [0,T] \times \Omega \to \mathbb{R}, \quad v : [0,T] \times \Omega \to \mathbb{R}^3$$

bezeichnen die Dichte, den Druck und das Geschwindigkeitsfeld. $\Omega \subset \mathbb{R}^3$ ist das räumliche Gebiet, in dem der jeweilige Prozess betrachtet wird, und $[0,T]$ ist das interessierende Zeitintervall. ν bezeichnet die kinematische Viskosität und F steht für ein äußeres Kraftfeld.
Im Fall eines inkompressiblen Fluids gilt ρ = const. und die Kontinuitätsgleichung (2.1) vereinfacht sich zu

$$\operatorname{div} v = 0\,. \tag{2.3}$$

Unter Nutzung von (2.3) vereinfacht sich die Navier-Stokes-Gleichung (2.2) zu

$$\frac{\partial v}{\partial t} + (v \cdot \nabla)v = -\frac{1}{\rho}\nabla p + \nu \Delta v + F \ . \tag{2.4}$$

Als Ergebnis der Energiebilanz erhält man für ein inkompressibles Medium als Spezialfall die parabolische Wärmeleitungsgleichung mit Berücksichtigung des konvektiven Transports

$$\frac{\partial \theta}{\partial t} + (v \cdot \nabla)\theta = a\Delta\theta + Q \tag{2.5}$$

für das Temperaturfeld $\theta : [0, T] \times \Omega \to \mathbb{R}$ (a ist die Temperaturleitzahl und Q beschreibt Wärmequellen oder -senken in Ω).
In der Navier-Stokes-Gleichung (2.2) beschreiben der Term

$$\rho[\frac{\partial v}{\partial t} + (v \cdot \nabla)v]$$

die Beschleunigungskräfte,

$$\nabla p$$

die Druckkraft und

$$\nu\rho[\frac{4}{3}\Delta v - \nabla \times (\nabla \times v)]$$

die Reibungskräfte. Z.B. bei der Modellierung der Umströmung eines Tragflügels spielen die Reibungskräfte nur eine untergeordnete Rolle, so dass bei diesem Strömungsproblem die Impulsbilanz als Spezialfall der Navier-Stokes-Gleichung (ohne Reibungsterme) durch die hyperbolische Euler-Gleichung

$$\frac{\partial v}{\partial t} + (v \cdot \nabla)v = -\frac{1}{\rho}\nabla p + F \tag{2.6}$$

beschrieben wird.
Bei den zeitabhängigen Problemen sind Anfangsbedingungen für die zu berechnenden Felder, z.B. für die Temperatur etwa

$$\theta(0, x) = \theta_0(x) \ , \quad x \in \Omega, \tag{2.7}$$

vorzugeben. Handelt es sich bei den beschreibenden Differentialgleichungen um Gleichungen mit räumlichen zweiten Ableitungen, sind Randbdingungen, als Beispiel

$$\theta(t, x) = \theta_r(t, x) \ , \quad x \in \Gamma = \partial\Omega, \tag{2.8}$$

zum Abschluss des jeweiligen Modells vorzugeben. Bei Vorgabe eines Geschwindigkeitsfeldes v sowie von a und Q ist durch (2.5), (2.7), (2.8)

ein Anfangs-Randwert-Problem zur Bestimmung des zeitlich veränderlichen Temperaturfeldes $\theta(t,x)$ in $[0,T]\times\Omega$ gegeben, dessen Lösung i.d.Regel numerische erfolgen muss.

Im Folgenden sollen noch 2 Randwertprobleme im Rahmen der Bestimmung des Minimums eines Funktionals bzw. der thermischen Kontrolle eines technologischen Prozesses angegeben werden.
Es soll das sogenannte Mumford-Shah-Funktional

$$E(f)=\int_\Omega [(f-d)^2+\alpha^2(R-I)^2]\,dF \tag{2.9}$$

minimiert werden. Dabei ist d ein gegebenes, i.d.Regel verrauschtes Datenfeld einer räumlichen Kontur (Fläche S im Raum), dass durch irgendwelche Sensoren generiert wurde. I beschreibt ein Intensitätsfeld R ist der Reflektionsgrad. Die gesuchte glatte Funktion f beschreibt die entrauschte geglättete Fläche S. Wenn l den Einheitsvektor in Richtung der Lichtquelle, die das zu erfassende Objekt mit der Oberfläche S beleuchtet, bezeichnet, und n den äußeren Normalvektor, ergibt sich für R

$$R=n\cdot l=\frac{(-f_x,-f_y,1)}{\sqrt{1+|\nabla f|^2}}\cdot(l_1,l_2,l_3)\ , \tag{2.10}$$

wobei f_x, f_y die partiellen Ableitunge von f bedeuten. Mit den Setzungen

$$\nabla_{f_x,f_y}R = -\frac{(l_1,l_2)}{\sqrt{1+|\nabla f|^2}}-\frac{n\cdot l}{\sqrt{1+|\nabla f|^2}}\nabla f\ , \tag{2.11}$$

$$V = \alpha^2(R-I)\nabla_{f_x,f_y}R \tag{2.12}$$

erhält man aus der notwendigen Extremalbedingung für die Variation $\delta E(f;v)=0$ für alle Richtungen v die Euler-Lagrange-Differentialgleichung

$$\nabla\cdot V+(d-f)=0\quad\text{auf}\quad\Omega \tag{2.13}$$

mit der Randbedingung

$$n\cdot V=0\ ,\quad \frac{\partial^2 f}{\partial n^2}=0\quad\text{auf}\quad\Gamma=\partial\Omega\ . \tag{2.14}$$

Bei genauerem Hinsehen erkennt man in (2.13) eine biharmonische Differentialgleichung mit Ableitungen von f bis zur Ordnung 4.

Im zweiten Beispiel zur Optimierung mit partiellen Differentialgleichungen

soll in einem Bereich Ω durch eine bestimmte Heiz- bzw. Kühlstrategie (realisiert durch eine vorzugebenden Wärmestromdichte am Rand) eine bestimmte vorgegebene Temperaturverteilung $\overline{T}$ eingestellt oder sehr gut angenähert werden. Denkbar wäre hier die Bearbeitung eines Stahlblockes oder das Aufschmelzen von Ausgangsstoffen zur Erzeugung eines homogenen Gemischs. Auf einem Teil des Randes Γ_d von Ω sei eine fixierte Temperatur vorgegeben und auf dem verbleibenden Rand Γ_c wird geheizt.

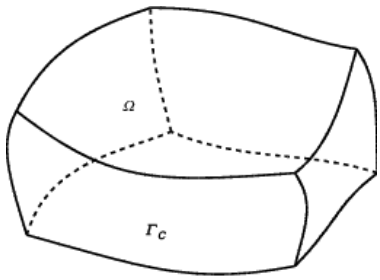

Abbildung 2.1: Bereich Ω und Heizungsrand Γ_c

Bemerkung 2.1. Für die nun folgenden Betrachtungen verabreden wir, dass wir von den beteiligten Funktionen soviel Regularität fordern, dass die vorkommenden Integrale existieren!

Es ist eine vorzugebende Wärmestromdichte (Heizstrategie) gesucht, die in Ω eine Temperaturverteilung zur Folge hat, die den um ein Kostenglied erweiterten quadratischen Abstand

$$J(T,q) = \frac{1}{2}\int_\Omega (T-\overline{T})^2\,dV + \frac{\alpha}{2}\int_{\Gamma_c} q^2\,dF \tag{2.15}$$

minimiert. Im Ergebnis der mathematischen Modellierung erhält man zur Berechnung der Temperaturverteilung T in Ω das elliptische Randwertproblem

$$-\Delta T = f \ \text{ in } \ \Omega, \qquad T = 0 \ \text{ auf } \ \Gamma_d, \ \frac{\partial T}{\partial \mathbf{n}} = q \ \text{ auf } \ \Gamma_c\,, \tag{2.16}$$

wobei f vorgegeben ist und q die gesuchte optimale Wärmestromdichte ist. Die Randbedingung $T = 0$ auf Γ_c stellt keine Einschränkung der Allgemeinheit dar, da man von Null verschiedene Randtemperaturen T_d auf Γ_d auf Ω zu T_0 fortsetzen kann, und für die Differenz $T - T_0$ auf Γ_d eine homogene Randbedingung erhält. Statt der Wärmeleitungsgleichung $-\Delta T = g$ würde

man dann für die Differenz die Gleichung $-\Delta(T - T_0) = g + \Delta T_0 =: f$ erhalten.
Wir definieren das LAGRANGE-Funktional

$$\begin{aligned} L(T,\kappa,q,\chi) &= \frac{1}{2}\int_\Omega (T-\overline{T})^2\,dV + \frac{\alpha}{2}\int_{\Gamma_c} q^2\,dF \\ &\quad - \int_\Omega (\Delta T + f)\kappa\,dV + \int_{\Gamma_c} (\frac{\partial T}{\partial \mathbf{n}} - q)\chi\,dF \end{aligned} \tag{2.17}$$

und man erkennt, dass für eine Lösung T von (2.16)

$$L(T,\kappa,q,\chi) = \frac{1}{2}\int_\Omega (T-\overline{T})^2\,dV + \frac{\alpha}{2}\int_{\Gamma_c} q^2\,dF = J(T,q)$$

gilt. Wir suchen das Minimum von L für auf Ω definierten Funktionen T und κ.
Für die FRÉCHET-Ableitung von L findet man an der Stelle $w = (T,\kappa,q,\chi)^T$ in Richtung $h = (\tilde{T},\tilde{\kappa},\tilde{q},\tilde{\chi})^T$

$$L'[w](h) = \begin{pmatrix} \int_\Omega (T-\overline{T})\tilde{T}dV - \int_\Omega \Delta\tilde{T}\,\kappa\,dV + \int_{\Gamma_c}\frac{\partial\tilde{T}}{\partial\mathbf{n}}\chi\,dF \\ -\int_\Omega(\Delta T + f)\tilde{\kappa}\,dV \\ \int_{\Gamma_c}\alpha q\tilde{q}\,dF - \int_{\Gamma_c}\tilde{q}\chi\,dF \\ \int_{\Gamma_c}(\frac{\partial T}{\partial\mathbf{n}} - q)\tilde{\chi}\,dF \end{pmatrix}. \tag{2.18}$$

Beachtet man, dass

$$\int_\Omega \Delta\tilde{T}\,\kappa\,dV = \int_\Omega \Delta\kappa\,\tilde{T}\,dV + \int_\Gamma \frac{\partial\tilde{T}}{\partial\mathbf{n}}\kappa\,dF - \int_\Gamma \frac{\partial\kappa}{\partial\mathbf{n}}\tilde{T}\,dF\,,$$

aufgrund der zweiten GREENschen Integralformel gilt, und variiert die Testfunktionen $\tilde{T},\tilde{\kappa},\tilde{q},\tilde{\chi}$, dann ergibt sich mit der speziellen Wahl $\chi = \kappa$ auf Γ_c, aus (2.18)

$$L'[w](h) = \begin{pmatrix} \int_\Omega[(T-\overline{T}) - \Delta\kappa]\tilde{T}dV + \int_{\Gamma_c}\frac{\partial\kappa}{\partial\mathbf{n}}\tilde{T}\,dF \\ -\int_\Omega(\Delta T + f)\tilde{\kappa}\,dV \\ \int_{\Gamma_c}[\alpha q - \kappa]\tilde{q}\,dF \\ \int_{\Gamma_c}(\frac{\partial T}{\partial\mathbf{n}} - q)\tilde{\chi}\,dF \end{pmatrix}. \tag{2.19}$$

Aus (2.19) wird deutlich, dass man mit der Lösung T des Randwertproblems (2.16) und der Lösung κ des dazu adjungierten Problems

$$-\Delta\kappa = -(T-\overline{T}) \text{ in } \Omega, \qquad \kappa = 0 \text{ auf } \Gamma_d,\ \frac{\partial\kappa}{\partial\mathbf{n}} = 0 \text{ auf } \Gamma_c\,, \tag{2.20}$$

sowie der Wärmestromdichte

$$q = \frac{1}{\alpha}\kappa \quad \text{auf} \quad \Gamma_c \tag{2.21}$$

einen stationären Punkt des Funktionals L gefunden hat, denn dann gilt

$$L'[w](h) = L'[T, \kappa, q, \chi](\tilde{T}, \tilde{\kappa}, \tilde{q}, \tilde{\chi}) = \mathbf{0} \, .$$

Für die Berechnung eines stationären Punktes sind damit zwei gekoppelte elliptische Randwertprobleme (2.16) und (2.20) zu lösen, und mit den Werten von κ auf Γ_c hat man letztendlich durch die Beziehung (2.21) eine optimale Heizstrategie gefunden. Die Diskussion der Existenz und Einzigkeit einer Lösung dieser Optimierungsaufgabe würde den Rahmen dieser Darstellung deutlich sprengen, da dazu umfassende funktionalanalytische Untersuchungen erforderlich werden. Deshalb wird darauf nicht eingegangen.

Abschließend sei mit der Wellengleichungen zweiter Ordnung

$$\frac{\partial^2 u}{\partial t^2} = a^2 \Delta u \tag{2.22}$$

bzw. Wellengleichunger erster Ordnung

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0 \, , \tag{2.23}$$

aus der die Gleichung (2.22) im räumlich eindimensionalen Fall folgt, auf die Klasse der hyperbolischen Differentialgleichungen hingewiesen. Die Gleichung (2.23) ist ein Spezialfall der Erhaltungsgleichung

$$\frac{\partial \vec{u}}{\partial t} + \nabla \cdot f(\vec{u}) = 0 \, , \tag{2.24}$$

die für

$$\vec{u} = \begin{pmatrix} \rho \\ \rho u \\ \rho v \end{pmatrix} \, , \quad f(\vec{u}) = \begin{pmatrix} f_1(\vec{u} \\ f_2(\vec{u}) \end{pmatrix}$$

mit

$$f_1 = \begin{pmatrix} \rho u \\ \rho u^2 + p \\ \rho uv \end{pmatrix} \text{ und } f_2 = \begin{pmatrix} \rho v \\ \rho uv \\ \rho v^2 + p \end{pmatrix} \, .$$

auch die Eulergleichungen umfasst.

Mit hyperbolischen Differentialgleichungen werden Wellenphänomene aus dem Gebiet der Akustik, der Elektromagnetik, der Seismik, der Optik bzw. der Strömungsmechanik beschrieben.

Kapitel 3

Klassifikation partieller Differentialgleichungen 2. Ordnung

Bei den oben diskutierten partiellen Differentialgleichungen handelte es sich Differentialgleichungen erster oder zweiter Ordnung, die man etwa in der Form

$$\mathrm{div} A \mathrm{grad} u + a^T \mathrm{grad} u + bu = f \tag{3.1}$$

aufschreiben kann, wobei u eine Funktion von n Veränderlichen ist, und A eine $(n \times n)$-Matrix, $a \in \mathbb{R}^n$ und $b \in \mathbb{R}$ reellwertige Koeffizienten darstellen. Im Falle von $n = 2$ ergibt die Differentialgleichung

$$u_{xx} + 4u_{xy} - 2u_{yy} + 3u_x - xyu_y + \sin x\, u = 3$$

die Darstellung

$$\mathrm{div} A \mathrm{grad} u + a^T \mathrm{grad} u + bu = f$$

mit

$$A = \begin{pmatrix} 1 & 2 \\ 2 & -2) \end{pmatrix} , \quad a = [3 \quad -xy]^T , \quad b = \sin x \, .$$

Für die Untersuchung oder die numerische Behandlung von Differentialgleichungen ist der **Typ** der Differentialgleichung von Bedeutung.

Definition 3.1. *Eine Differentialgleichung* (3.1) *heißt*

a) **elliptisch**, *wenn die Matrix A entweder positiv oder negativ definit ist, d.h. entweder nur positive oder nur negative Eigenwerte bestitzt,*

b) **hyperbolisch**, *wenn die Matrix A einen Eigenwert λ_1 und Eigenwerte $\lambda_2, \ldots, \lambda_n$ besitzt, mit den Eigenschaften, dass die Eigenwerte $\lambda_2, \ldots, \lambda_n$ das gleiche Vorzeichen besitzen (und ungleich Null sind), und der Eigenwert λ_1 das entgegengesetzte Vorzeichen hat,*

c) **parabolisch**, *wenn ein Eigenwert von A gleich Null ist, und die Matrix $[A|a]$ den vollen Rang besitzt, also $rg(A|a) = n$ gilt.*

Es ist offensichtlich, dass die instationäre Wärmeleitungsgleichung parabolisch ist, während die stationäre Wärmeleitungsgleichung offensichtlich elliptisch ist. Wellengleichungen sind hyperbolisch. Es ist zudem anzumerken, dass man nicht in jedem Fall einer partiellen Differentialgleichung zweiter Ordnung eine eben beschriebene Klassifikation durchführen kann.
Bei Differentialgleichungen mit variablen Koeffizienten A und a ist der Typ im Allg. auf ortsabhängig, so dass bei einer Aufgabenstellung Übergänge vom hyperbolischen zum elliptischen Typ etc. vorkommen können.

Beispiel 3.2. Die Differentialgleichung

$$x^2 u_{xx} + (y - x^2) u_{yy} + 4u_x - xyu_y + (x + y)u = 1$$

ändert ihren Typ in Abhängigkeit vom Ort, d.h. sie ist elliptisch, für $x \neq 0$ und $y > x^2$. Für $x \neq 0$ und $y < x^2$ ist sie hyperbolisch. Die Matrix A ergibt sich zu

$$A = \begin{pmatrix} x^2 & 0 \\ 0 & y - x^2 \end{pmatrix}$$

und der Vektor a ist in diesem Fall

$$a = [4 \quad -xy]^T \, .$$

Für Differentialgleichung

$$[\partial x \;\; \partial y] \begin{pmatrix} 1 & -x \\ -x & 3 \end{pmatrix} \begin{pmatrix} \frac{\partial u}{\partial x} \\ \frac{\partial u}{\partial y} \end{pmatrix} + [x \;\; 1 + y^2] \begin{pmatrix} \frac{\partial u}{\partial x} \\ \frac{\partial u}{\partial y} \end{pmatrix} - 4u = x^2 - y^2$$

ergibt sich für die Eigenwerte von A

$$\lambda_{1,2} = 2 \pm \sqrt{1 + x^2} \, ,$$

so dass Elliptizität für $|x| < 1$ vorliegt. Für $|x| > 1$ ist die Gleichung hyperbolisch. Für $|x| = 1$ ist die Gleichung parabolisch.

Kapitel 4

Finite-Differenzen zur Lösung partieller Differentialgleichungen

2. Vorlesung am 29.05.2017

Im Folgenden besprechen wir das klassische Finite Differenzen Verfahren (FDM) zur Lösung von Randwertproblemen. Bei der Finite-Differenzen Methode ersetzt man Ableitungen in der Differentialgleichung durch Differenzenquotienten. Dies führt dann zu einem linearen Gleichungssystem für Näherungswerte u_h, an die gesuchten Werte u der Lösung in vorgegebenen Knotenpunkten.

4.1 FDM für Elliptische Randwertprobleme, Grundlagen

Zur Illustration betrachten wir das Modelproblem

$$-\Delta u = f \quad \text{in} \quad \Omega =]0,1[^n \subset \mathbb{R}^n \,, \tag{4.1}$$

$$u = 0 \quad \text{auf} \quad \Gamma = \partial\Omega \,, \tag{4.2}$$

wobei f eine glatte Funktion sein soll. Mit einem Diskretisierungsparameter $h := 1/N,\ N \in \mathbb{N}$, wird die Menge aller Gitterpunkte

$$\overline{\Omega}_h = \{(x_1, \ldots, x_n) \in \mathbb{R}^n \mid x_1 = i_1 h, \ldots, x_n = i_n h, i_1, \ldots, i_n = 0, 1, \ldots, N\}$$

definiert. Für die Menge der **inneren** Gitterpunkte und der Randgitterpunkte ergibt sich

$$\Omega_h = \overline{\Omega}_h \cap \Omega \,, \quad \Gamma_h = \overline{\Omega}_h \cap \Gamma \,.$$

Die verwendete Diskretisierung von Ω ist mit dem konstanten Diskretsierungsparameter äquidistant.
Beim Finiten-Differenzenverfahren sucht man Gitterfunktionen $u_h : \overline{\Omega}_h \to \mathbb{R}$ als Näherungslösungen des ursprünglichen Problems (4.1), (4.2). Dazu betrachten wir die Funktionenräume

$$U_h = \{u_h : \overline{\Omega}_h \to \mathbb{R}\}, \quad U_h^0 = \{u_h : \overline{\Omega}_h \to \mathbb{R}, u_h|_{\Gamma_h} = 0\}, \quad V_h = \{u_h : \Omega_h \to \mathbb{R}\} .$$

Zur Approximation der Ableitungen in den partiellen Differentialgleichungen führen wir Differenzenquotienten für die partiellen Ableitungen nach x_j ein:

Definition 4.1. *Sei $x \in \mathbb{R}^n$ und $e^j = (0, \ldots, 0, 1, 0, \ldots, 0) \in \mathbb{R}^n$ der j-te kanonische Einheitsvektor und $h > 0$ ein Diskretisierungsparameter.*

- *Vorwärtsdifferenzenquotient:* $D_j^+ u(x) := \frac{u(x+he^j)-u(x)}{h}$
- *Rückwärtsdifferenzen-Quotient:* $D_j^- u(x) = \frac{u(x)-u(x-he^j)}{h}$
- *Zentraler Differenzen-Quotient:* $D_j^0 u(x) = \frac{1}{2}(D_j^+ u(x) + D_j^- u(x))$.

Zur Approximation von $u_{x_j x_j}(x)$ nutzen wir den zentralen Differenzenquotienten 2. Ordnung

$$D_j^+ D_j^- u(x) .$$

Zur Untersuchung von Konsistenz, Stabilität und Konvergenz von Finiten-Differenzen-Methoden müssen wir die Gitterfunktionsräume mit Normen ausstatten.

Definition 4.2. *Auf U_h^0 definieren wir mit*

$$||u_h||_{0,h}^2 := h^n \sum_{x_h \in \Omega_h} |u_h(x_h)|^2 , \quad u_h \in U_h^0 , \tag{4.3}$$

die diskrete L_2-Norm, mit

$$||u_h||_{1,h}^2 := h^n \sum_{x_h \in \Omega_h} \sum_{j=1}^{n} |[D_j^+ u_h](x_h)|^2 , \quad u_h \in U_h^0 , \tag{4.4}$$

die diskrete H^1-Norm, und mit

$$||u_h||_{\infty,h} := \max_{x_h \in \overline{\Omega}_h} |u_h(x_h)| , \quad u_h \in U_h^0 , \tag{4.5}$$

die diskrete Maximumnorm. Mit

$$||u_h||_{\infty,\Gamma_h} := \max_{x_h \in \Gamma_h} |u_h(x_h)|$$

bezeichnet man eine diskrete Randnorm. Mit

$$(u_h, v_h)_h := h^n \sum_{x_h \in \Omega_h} u_h(x_h) v_h(x_h) \ , \quad u_h, v_h \in U_h^0 \ , \tag{4.6}$$

wird in U_h^0 ein Skalarprodukt definiert.

Offensichtlich gilt

$$||u_h||_{0,h}^2 = (u_h, u_h)_h \quad \text{und} \quad ||u_h||_{1,h}^2 = \sum_{j=1}^{n} (D_j^+ u_h, D_j^+ u_h)_h \ , \ u_h \in U_h^0 \ .$$

Im Allgemeinen kann man ein Randwertproblem in der Form

$$Fu = f \tag{4.7}$$

mit geeignet zu wählenden Funktionenräumen U und V, einer Abbildung $F : U \to V$ und $f \in V$ beschreiben. Für das diskretisierte Problem kann man analog

$$F_h u_h = f_h \tag{4.8}$$

mit Gitterfunktionsräumen U_h und V_h, einer Abbildung $F_h : U_h \to V_h$ und $f_h \in V_h$ schreiben. $r_h : U \to U_h$ und $r_h v$ bezeichne die Einschränkung von $v \in U$ auf U_h (z.B. $U = C^2(\overline{\Omega})$ und U_h als Raum der Gitterfunktionen).

Definition 4.3. *(Konsistenz)*
Den Ausdruck

$$||F_h(r_h u) - f_h||_{V_h}$$

bezeichnet man als Konsistenzfehler bezüglich $u \in U$. Eine Diskretisierung von (4.7) heißt konsistent, wenn

$$||F_h(r_h u) - f_h||_{V_h} \to 0 \quad \textit{für} \quad h \to 0$$

gilt. Wenn für den Konsistenzfehler

$$||F_h(r_h u) - f_h||_{V_h} = O(h^p) \quad \textit{für} \quad h \to 0$$

gilt, dann spricht man von der Konsistenzordnung p.

Definition 4.4. *(Konvergenz)*
Eine Diskretisierungsmethode ist konvergent, wenn

$$||r_h u - u_h||_{U_h} \to 0 \quad \textit{für} \quad h \to 0$$

gilt, und man spricht von der Konvergenzordnung q, wenn

$$||r_h u - u_h||_{U_h} = O(h^q) \quad \textit{für} \quad h \to 0$$

gilt.

Definition 4.5. *(Stabilität)*
Eine Diskretisierungsmethode ist stabil, wenn für eine Konstante $S > 0$

$$||v_h - w_h||_{U_h} \leq S||F_h v_h - F_h w_h||_{V_h} \quad \textit{für} \quad v_h, w_h \in U_h$$

gilt.

Satz 4.6. *Vorausgesetzt das kontinuierliche* (4.7) *und das diskrete Problem* (4.8) *sind eindeutig lösbar und die Diskretisierungsmethode ist konsistent und stabil, dann ist die Methode konvergent. Die Konvergenzordnung ist dabei größer oder gleich der Konsistenzordnung.*

Beweis. Aus der Stabilität folgt

$$||u_h - r_h u_|||_{U_h} \leq S||F_h u_h - F_h r_h u||_{V_h} = S||f_h - F_h r_h u||_{V_h} = (*)$$

und aus der Konsistenz folgt

$$(*) \to 0 \quad \text{für} \quad h \to 0$$

und damit schließlich

$$||u_h - r_h u_|||_{U_h} \to 0 \quad \text{für} \quad h \to 0 \,,$$

also die Konvergenz. Hat man die Konsistenzordnung p, so folgt

$$||u_h - r_h u_|||_{U_h} = O(h^p) \quad \text{für} \quad h \to 0 \,,$$

d.h. man hat auch die Konvergenzordnung p. □

4.2 Konsistenz-, Stabilitäts- und Konvergenznachweise

Für das Modellproblem (4.1), (4.2)

$$-\Delta u = +c\,u = f \;\text{ in }\; \Omega =]0,1[^n\,, \quad u = 0 \;\text{ auf }\; \Gamma = \bar{\Omega} \setminus \Omega$$

soll im Folgenden ein Finite-Differenzenverfahren konstruiert und auf Konsistenz, Stabilität und Konvergenz bezüglich unterschiedlicher Normen untersucht werden.

4.2.1 Stabilität und Konvergenz in der L_2-Norm

Wir beginnen mit der diskreten L_2-Norm und verwenden die oben erklärten Diskretisierungen und Operatoren. Als wichtige Hilfsmittel werden nun einige Formeln bzw. Ungleichungen gezeigt.

Satz 4.7. *(diskrete Greensche Formel)*
Für alle $w_h, v_h \in U_h^0$ und $1 \le j \le n$ gilt

$$(D_j^- w_h, v_h)_h = -(w_h, D_j^+ v_h)_h\,,$$

also ist $-D_j^+$ der adjungierte Operator von D_j^- bezüglich des Skalarprodukts $(\cdot,\cdot)_h$.

Beweis. Mit der Definition von D_j^- erhält man

$$\begin{aligned}(D_j^- w_h, v_h)_h &= h^n \sum_{x_h \in \Omega_h} \frac{w_h(x_h) - w_h(x_h - he^j)}{h} v_h(x_h) \\ &= h^{n-1} \sum_{x_h \in \Omega_h} (w_h(x_h)v_h(x_h) - w_h(x_h - he^j)v_h(x_h))\end{aligned}$$

und wegen $w_h(x_h) = v_h(x_h) = 0$ auf Γ_h gilt weiter

$$\begin{aligned}(D_j^- w_h, v_h)_h &= h^{n-1} \sum_{x_h \in \Omega_h} w_h(x_h)(v_h(x_h) - v_h(x_h + he^j)) \\ &= h^n \sum_{x_h \in \Omega_h} w_h(x_h) \frac{v_h(x_h) - v_h(x_h + he^j)}{h} \\ &= -(w_h, D_j^+ v_h)_h\,.\end{aligned}$$

□

Als Folgerung aus dem Satz ergibt sich somit für

$$L_h w_h = -\sum_{j=1}^{n} D_j^- D_j^+ w_h$$

die Beziehung

$$(L_h w_h, v_h)_h = \sum_{j=1}^{n} a_j(x_h) D_j^+ w_h, D_j^+ v_h)_h \quad \text{für alle } \; w_h, v_h \in U_h^0 \tag{4.9}$$

und weiter

$$(L_h w_h, w_h)_h = \sum_{j=1}^{n} D_j^+ w_h, D_j^+ w_h)_h = ||w_h||_{1,h}^2 \quad \text{für alle } \; w_h, \in U_h^0 \; . \tag{4.10}$$

Es soll nun die **diskrete Friedrichssche Ungleichung**, die eine Beziehung zwischen den Normen $||\cdot||_{0,h}$ und $||\cdot||_{1,h}$ herstellt, bewiesen werden.

3. Vorlesung am 01.06.2017

Satz 4.8. *(Friedrichssche Ungleichung)*
Es existiert eine Konstante c, die nur von Ω abhängt, so dass

$$||w_h||_{0,h} \leq c||w_h||_{1,h} \quad \textit{für alle } \; w_h \in U_h^0 \tag{4.11}$$

gilt.

Beweis. Sei $x_h \in \Omega_h$ und $j \in \{1, \ldots, n\}$ beliebig aber fest. Da $w_h(x_h) = 0$ auf Γ_h gilt, gibt es einen Index $\hat{l} = \hat{l}(x_h) \leq N - 1$, so dass

$$|w_h(x_h)| = |\sum_{l=0}^{\hat{l}} (w_h(x_h + (l+1)he^j) - w_h(x_h + lhe^j))|$$

gilt. Mit der Dreiecksungleichung und der Cauchy-Schwarzschen Ungleichung erhält man

$$\begin{aligned}
|w_h(x_h)| &\leq \sum_{l=0}^{\hat{l}} |w_h(x_h + (l+1)he^j) - w_h(x_h + lhe^j)| \\
&= h \sum_{l=0}^{\hat{l}} |D_j^+ w_h(x_h + lhe^j)| \\
&\leq h(\sum_{l=0}^{\hat{l}} 1)^{1/2} (\sum_{l=0}^{\hat{l}} |D_j^+ w_h(x_h + lhe^j)|^2)^{1/2} \\
&\leq h\sqrt{N}(\sum_{l=0}^{\hat{l}} |D_j^+ w_h(x_h + lhe^j)|^2)^{1/2} \; .
\end{aligned}$$

Mit $h = 1/N$ erhält man daraus

$$\begin{aligned} ||w_h||_{0,h}^2 &= h^n \sum_{x_h \in \Omega_h} |w_h(x_h)|^2 \\ &\leq h^{n+2} N \sum_{x_h \in \Omega_h} \sum_{l=0}^{\hat{l}} |D_j^+ w_h(x_h + lhe^j)|^2 \\ &\leq h^{n+2} N^2 \sum_{x_h \in \Omega_h} |D_j^+ w_h(x_h + lhe^j)|^2 \\ &\leq h^n \sum_{j=1}^{n} \sum_{x_h \in \Omega_h} |D_j^+ w_h(x_h + lhe^j)|^2 = ||w_h||_{1,h}^2 \ . \end{aligned}$$

D.h., die Ungleichung gilt im Falle von $\Omega =]0,1[^n$ mit $c = 1$. □

Nun kann man unter Nutzung der Sätze 4.7 und 4.8 unter gewissen Glattheitsvoraussetzungen an die exakte Lösung Stabilität und Konvergenz in der L_2-Norm zeigen. Die im Folgenden vorausgesetzte Lösbarkeit der diskretisierten Aufgabenstellungen wird später im Zusammenhang mit dem diskreten Maximum-Prinzip gezeigt.

Satz 4.9. *Für die Lösung von* (4.1), (4.2) *gelte* $u \in C^4(\bar{\Omega})$. *Dann existieren positve Konstanten* c_1, c_2, *so dass für die Lösung des diskretiserten Problems*

$$L_h u_h = f_h \quad in\ \Omega_h \ , \quad u_h = 0 \quad auf\ \Gamma_h \ , \tag{4.12}$$

mit $f_h = r_h f$ *die Abschätzungen*

$$||u_h - r_h u||_{0,h} \leq c_1 ||u_h - r_h u||_{1,h} \leq c_2 ||u||_{C^4(\bar{\Omega})} h^2 \ ,$$

also Konsistenz, Stabilität und Konvergenz in der diskreten L_2*-Norm gelten.*

Beweis. Sei $e_h = u_h - r_h u$, dann gilt

$$(L_h e_h, e_h)_h = (d_h, e_h)_h$$

mit $||d_h||_{0,h} \leq c_2 ||u||_{C^4(\bar{\Omega})} h^2$ mit einer geeigneten Konstanten c_2, was man leicht durch Taylorentwicklungen zeigen kann. Aus der Gleichung (4.10) und der Cauchy-Schwarzschen Ungleichung folgt

$$||e_h||_{1,h}^2 \leq (L_h e_h, e_h)_h = (d_h, e_h)_h \leq ||d_h||_{0,h} ||e_h||_{0,h} \ . \tag{4.13}$$

Mit Satz 4.8 erhält man daraus

$$||e_h||_{1,h}^2 \leq ||d_h||_{0,h} ||e_h||_{1,h} \quad \text{bzw.} \quad ||e_h||_{1,h} \leq ||d_h||_{0,h} \ .$$

Die nochmalige Anwendung der Friedrichsschen Ungleichung ergibt mit

$$||u_h - r_h u||_{0,h} \leq c_1 ||u_h - r_h u||_{1,h} \leq c_2 ||u||_{C^4(\bar{\Omega})} h^2$$

die Behauptung des Satzes. □

Bemerkung 4.10. Aus der Gleichung 4.13 des eben durchgeführten Beweises ergibt sich für den Fall der Lösbarkeit von (4.12) die Eindeutigkeit einer Lösung des diskretisierten elliptischen Randwertproblems.
Die diskrete Friedrichssche Ungleichung (4.11) bedeutet die Stabilität der Diskretisierung in der diskreten L_2-Norm.
Die Aussage des Satzes 4.9 bedeutet Konvergenz sowohl in der diskreten L_2- als auch in der diskreten H^1-Norm.

4.2.2 Stabilität und Konvergenz in der Maximumnorm

Im Folgenden wollen wir für den Fall eines elliptischen Randwertproblems im Zweidimensionalen Finite-Differenzenverfahren auf Konvergenz in der Maximunorm untersuchen. Dabei lassen wir auch allgenmeinere Gebiete Ω als das Einheitsquadrat zu.
Wir wollen eine Richtungs-äquidistante Diskretierung von Ω erzeugen und diskretisieren dazu den $\mathbb{R}^2$ durch

$$\mathbb{R}_h^2 := \{(x_{1,i}, x_{2,j}) \mid x_{1,i} = ih_1, x_{2,j} = jh_2,\ h_1, h_2 > 0,\ i, j \in \mathbb{Z}\} .$$

Mit

$$G_{1,i} = \{(x_1, x_2) \mid x_1 = ih_1, x_2 \in \mathbb{R}\} , \quad G_{2,j} = \{(x_1, x_2) \mid x_2 = jh_2, x_1 \in \mathbb{R}\}$$

führen wir Gitterlinien ein. Durch

$$\Omega_h = \Omega \cap \mathbb{R}_h^2 \quad \text{und} \quad \Gamma_h = \Gamma \cap (\bigcup_{i \in \mathbb{Z}} G_{1,i} \cup \bigcup_{j \in \mathbb{Z}} G_{2,j})$$

diskretisieren wir Ω und der Rand $\Gamma = \partial\Omega$. Damit erhalten wir unterschiedliche Typen von Knoten/Gitterpunkten, und zwar

- Ω_h° die Menge der echt inneren Knoten, deren Nachbarknoten in $\bar{\Omega}$ liegen.
- Ω_h^* die Menge der randnahen Knoten, bei denen ein Nachbarknoten außerhalb von $\bar{\Omega}$ liegt.
- Γ_h die Menge der Randknoten.

- $\Omega_h = \Omega \cap \mathbb{R}_h^2 = \Omega_h^\circ \cup \Omega_h^*$ die Menge der inneren Knoten.

Als Gitter bezeichnen wir $\Omega_h \cup \Gamma_h$.
Die Diskretisierung des Laplace-Operators bei einer Richtungs-äquidistanten Diskretierung von Ω ergibt sich zu

$$\begin{aligned} -L_h u &:= (D_1^+ D_1^- + D_2^+ D_2^-) u(x_{1,i}, x_{2,j}) \\ &= \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h_1^2} + \frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{h_2^2} \end{aligned} \tag{4.14}$$

wobei wir die Notation

$$u_{i,j} = u(x_{1,i}, x_{2,j})$$

verwenden. (4.14) approximiert den Laplace-Operator mit der Ordnung $O(h_1^2 + h_2^2)$. Verzichtet man auf die Richtungsäquidistanz und bezeichnet mit h_1^+ den Abstand der Gitterpunktes $(x_{1,i}, x_{2,j})$ zum rechten Nachbarn, h_1^- zum linken Nachbarn usw. dann kann man an randnahen Knoten 2. Ableitungen durch

$$\frac{\partial^2 u}{\partial x_1^2}(x_1, x_2) \approx \frac{2}{h_1^+ + h_1^-}\left(\frac{u(x_{1,i} + h_1^+, x_{2,j}) - u(i,j))}{h_1^+} - \frac{u_{i,j} - u(x_{1,i} - h_1^-, x_{2,j})}{h_1^+}\right) \tag{4.15}$$

diskretieren und erhält einen Approximationsfehler von $O(\tilde{h}_1)$ mit $\tilde{h}_1 = \frac{h_1^+ + h_1^-}{2}$, so dass man an randnahen Knoten den Laplace-Operator im Allg. nur mit 1. Ordnung approximieren kann.

Definition 4.11. *Für einen Diskretierungsstern mit den Gitterpunkten*

$$(x_{1,i}, x_{2,j}),\ (x_{1,i} + h_1^+, x_{2,j}),\ (x_{1,i} - h_1^-, x_{2,j}),\ (x_{1,i}, x_{2,j} + h_2^+),\ (x_{1,i}, x_{2,j} - h_2^-)$$

bezeichnet $S(x_{1,i}, x_{2,j})$ die Menge der benachbarten Gitterpunkte, also

$$S(x) = \{\textit{Gitterpunkte } y \neq x \textit{ im Diskretisierungsstern um } x\} .$$

Betrachet man die Diskretisierung

$$D_1^+ D_1^- u + D_2^+ D_2^- u - cu = f$$

der elliptischen Differentialgleichung

$$\Delta u - cu = f \quad c \geq 0 ,$$

dann ergibt sich mit

$$a u_{i,j} = [b_o u_{i+1,j} + b_w u_{i-1,j} + b_n u_{i,j+1} + b_s u_{i,j-1}] + f \tag{4.16}$$

eine Differenzengleichung, wobei für die Koeffizienten

$$b_o = \frac{1}{(h_1^+ + h_1^-)h_1^+}\ , \quad b_w = \frac{1}{(h_1^+ + h_1^-)h_1^-}\ ,$$
$$b_n = \frac{1}{(h_2^+ + h_2^-)h_2^+}\ , \quad b_s = \frac{1}{(h_2^+ + h_2^-)h_2^-}\ ,$$
$$a = b_o + b_w + b_n + b_s + c \tag{4.17}$$

gilt. Die Gleichung (4.16) kann man auch für einen Gitterpunkt $x \in \Omega_h$ in der kompakten Form

4. Vorlesung am 08.06.2017

$$a(x)u(x) = \sum_{y \in S(x)} b(x,y)u(y) + f(x) \tag{4.18}$$

oder

$$d(x)u(x) = \sum_{y \in S(x)} b(x,y)(u(y) - u(x)) + f(x) \tag{4.19}$$

mit

$$d(x) = a(x) - \sum_{y \in S(x)} b(x,y)$$

schreiben.

Bemerkung 4.12. Für die Koeffizienten des Schemas (4.19) gilt

$$a(x) > 0, \quad b(x,y) > 0 \quad \forall x \in \Omega_h,\ \forall y \in S(x)\ ,$$
$$a(x) = 1, \quad b(x,y) = 0 \quad \forall x \in \Gamma_h\ .$$

Nach diesen vorbereitenden Überlegungen können wir ein diskretes Maximumprinzip formulieren. Es gilt der

Satz 4.13. *(diskretes Maximumprinzip)*
Sei $u(x) \neq$ const. auf $\Omega_h \cup \Gamma_h$ und $d(x) \geq 0$ für alle $x \in \Omega_h$. Dann folgt aus

$$L_h u(x) := d(x)u(x) - \sum_{y \in S(x)} b(x,y)(u(y) - u(x)) \leq 0$$

(bzw. $Lu(x) \geq 0$) auf Ω_h, dass $u(x)$ den maximalen positiven (bzw. minimalen negativen) Wert nicht auf Ω_h annehmen kann.

Beweis. (Widerspruchsbeweis)
Sei $L_h u(x) \leq 0$ für alle $x \in \Omega_h$ und

$$u(\bar{x}) = \max_{x \in \Omega_h \cup \Gamma_h} u(x) > 0$$

für ein $x \in \Omega_h$. Für $\bar{x}$ gilt

$$L_h u(\bar{x}) = d(\bar{x})u(\bar{x}) - \sum_{y \in S(\bar{x})} \overbrace{b(\bar{x},y)}^{>0} \overbrace{(u(y) - u(\bar{x}))}^{\leq 0} \geq \underbrace{d(\bar{x})}_{\geq 0} \underbrace{u(\bar{x})}_{>0} \geq 0 \ ,$$

d.h. $L_h u(\bar{x}) = 0$. Die Bedingung ist nur mit

$$d(\bar{x}) = 0 \quad \text{und} \quad u(y) = u(\bar{x}) \quad \text{für } y \in S(x)$$

erfüllt. Da u als nicht konstant auf $\Omega_h \cup \Gamma_h$ vorausgesetzt wurde, gibt es einen Punkt $\bar{\bar{x}} \in \Omega_h \cup \Gamma_h$ mit $u(\bar{x}) > u(\bar{\bar{x}})$ und eine Verbindung jeweils benachbarter Punkte von $\bar{x}$ zu $\bar{\bar{x}}$, also eine Punktfolge

$$\bar{x}, x_1, x_2, \ldots, x_m, \bar{\bar{x}}$$

mit

$$\begin{aligned} x_1 \in S(\bar{x}), &\quad u(x_1) = u(\bar{x}) = \max\nolimits_{x \in \Omega_h \cup \Gamma_h} u(x) \\ x_2 \in S(x_1), &\quad u(x_2) = u(x_1) = u(\bar{x}) \\ &\vdots \\ x_m \in S(x_{m-1}), &\quad u(x_m) = u(x_{m-1}) = u(\bar{x}) \\ \bar{\bar{x}} \in S(x_m) &\quad u(\bar{\bar{x}}) = u(x_m) = u(\bar{x}) \ , \end{aligned}$$

also ein Widerspruch zu $u(\bar{x}) > u(\bar{\bar{x}})$. □

Bemerkung 4.14. Die Gleichungen (4.18) und (4.19) mit den erfüllten Bedingungen

$$a(x) > 0, \quad b(x,y) > 0, \quad d(x) = a(x) - \sum_{y \in S(x)} b(x,y) \geq 0 \quad x \in \Omega_h$$

und $d(x) > 0$ für mindestens einen Punkt (randnaher Punkt) $x \in \Omega$ bedeuten für die Koeffizientenmatrix A des diskretisierten Problems bei der lexikographischen Nummerierung der Gitterpunkte gerade die Eigenschaft, **irreduzibel diagonal dominant** zu sein, was die Regularität von A bedeutet. Außerdem kann man zeigen, dass die Matrix A^{-1} beschränkt ist, wobei die Schranke unabhängig von h ist. Das bedeutet die Lösbarkeit des Problems (4.12).

Mit dem Maximumprinzip kann man nun ein Reihe von Eigenschaften des Differenzenverfahrens (4.16) einfach schlussfolgern.

Korollar 4.15.
Sei $u(x) \geq 0$ für alle $x \in \Gamma_h$ und $Lu(x) \geq 0$ auf Ω_h. Dann ist die Gitterfunktion u nichtnegativ auf $\Omega_h \cup \Gamma_h$.

Beweis. Angenommen es existiert ein $\bar{x} \in \Omega_h$ mit $u(\bar{x}) < 0$. Damit nimmt die Funktion u ein negatives Minimum auf Ω_h an. Das ist ein Widerspruch zum Maximumprinzip. □

Satz 4.16.
Das Problem

$$L_h u(x) = 0,\ x \in \Omega_h, \quad u(x) = 0,\ x \in \Gamma_h\ ,$$

hat nur die triviale Lösung $u(x) = 0,\ x \in \Omega_h \cup \Gamma_h$.
Damit folgt: Das Problem

$$L_h u(x) = f,\ x \in \Omega_h, \quad u(x) = \mu(x),\ x \in \Gamma_h\ ,$$

ist eindeutig lösbar (Fredholmsche Alternative).

Beweis.
Übung □

Satz 4.17. *(Vergleichsprinzip)*
Seien die Probleme

$$L_h u(x) = f,\ x \in \Omega_h, \quad u(x) = \mu(x),\ x \in \Gamma_h\ , \tag{4.20}$$
$$L_h \bar{u}(x) = \bar{f},\ x \in \Omega_h, \quad \bar{u}(x) = \bar{\mu}(x),\ x \in \Gamma_h\ , \tag{4.21}$$

mit $|f(x)| \leq \bar{f}(x),\ x \in \Omega_h$ und $|\mu(x)| \leq \bar{\mu}(x),\ x \in \Gamma_h$ gegeben.
Dann gilt

$$|u(x)| \leq \bar{u}(x)\ , \quad \text{für alle } x \in \Omega_h \cup \Gamma_h.$$

Beweis.
Die Aussage ergibt sich durch die Anwendung des Maximumprinzips für $w(x) := \bar{u}(x) - |u(x)|$. □

Korollar 4.18.
Für die Lösung des Problems

$$L_h u(x) = f,\ x \in \Omega_h, \quad u(x) = \mu(x),\ x \in \Gamma_h\ , \tag{4.22}$$

mit $d(x) = 0,\ x \in \Omega_h$ (das ist der Fall, wenn $c = 0$ ist) gilt

$$||u||_{\infty,h} \leq ||\mu||_{\infty,\Gamma_h}\ .$$

Beweis.
Mit $\bar{\mu}(x) := ||\mu||_{\infty,\Gamma_h} = \text{const.}$ und $\bar{f}(x) = 0,\ x \in \Omega_h$ ist offensichtlich $\bar{u}(x) = ||\mu||_{\infty,\Gamma_h}$ eine Lösung von (4.21) und aus dem Vergleichsprinzip 4.17 folgt die Aussage. □

.

Korollar 4.19.
Für die Lösung des Problems

$$L_h u(x) = f,\ x \in \Omega_h, \quad u(x) = 0,\ x \in \Gamma_h\ , \tag{4.23}$$

mit $d(x) > 0,\ x \in \Omega_h$ *gilt*

$$||u||_{\infty,h} \leq \max_{x\in\Omega} \frac{|f(x)|}{d(x)}\ .$$

Beweis.
Für die Funktionen

$$\bar{f}(x) := |f(x)| \geq f(x),\ x \in \Omega_h, \quad \text{und } \bar{\mu}(x) = \mu(x) = 0,\ x \in \Gamma_h,$$

betrachten wir das Problem

$$L_h \bar{u}(x) = \bar{f}(x),\ x \in \Omega_h, \quad \bar{u}(x) = 0,\ x \in \Gamma_h\ ,$$

dessen Lösung aufgrund von Folgerung 4.15 nichtnegativ ist, also $\bar{u}(x) \geq 0,\ x \in \Omega_h \cup \Gamma_h$.
Sei $\bar{x}$ der Punkt mit $\bar{u}(\bar{x}) = ||\bar{u}||_{\infty,h}$. In $\bar{x}$ gilt

$$\underbrace{d(\bar{x})}_{>0}\underbrace{\bar{u}(\bar{x})}_{\geq 0} - \sum_{y\in S(\bar{x})} \underbrace{b(\bar{x},y)}_{>0}\underbrace{(\bar{u}(y) - \bar{u}(\bar{x}))}_{\leq 0} = |f(\bar{x})|$$

und damit

$$\bar{u}(\bar{x}) \leq \frac{|f(\bar{x})|}{d(\bar{x})} \leq \max_{x\in\Omega_h} \frac{|f(x)|}{d(x)}\ .$$

Aus dem Vergleichsprinzip 4.17 folgt die Behauptung. □

Als Hilfsresultat für den Konvergenzbeweis des Finite-Differenzenverfahrens mit den Diskretisierungen (4.14) im Inneren und (4.15) und einer entsprechenden Diskretisierung für die 2. partielle Ableitung nach x_2 am Rand benötigt man

Korollar 4.20.
Für innere Gitterpunkte $x \in \Omega_h^\circ$ gelte $d(x) = f(x) = 0$ und für randnahe Punkte $x \in \Omega_h^*$ gelte $d(x) > 0$. Auf dem Rand Γ_h gelte $\mu(x) = 0$.
Für die Lösung von

$$L_h u(x) = f(x),\ x \in \Omega_h, \quad u(x) = 0,\ x \in \Gamma_h\ ,$$

gilt dann

$$||u(x)||_{\infty,h} \leq \max\{0, \max_{x \in \Omega_h^*} \frac{|f(x)|}{d(x)}\}\ .$$

Beweis.
Man setzt $\bar{f}(x) = |f(x)|$ in Ω_h und $\bar{\mu}(x) = 0$ auf Γ_h. Dann ist nach dem diskreten Maximum-Prinzip die Lösung $\bar{u}(x)$ von

$$L_h \bar{u}(x) = \bar{f}(x),\ x \in \Omega_h, \quad \bar{u}(x) = 0,\ x \in \Gamma_h\ ,$$

nichtnegativ auf $\bar{\Omega}_h$. Damit findet man einen Gitterpunkt $\bar{x}$ mit $\bar{u}(\bar{x}) = ||\bar{u}||_{\infty,h}$.
Falls $\bar{x} \in \Omega_h^*$ ein randnaher Punkt ist, setzt man $\bar{\bar{x}} = \bar{x}$.
Im anderen Fall $\bar{x} \in \Omega_h^\circ$ findet man wie im Beweis des diskreten Maximum-Prinzips

$$\bar{u}(\bar{x}) = \bar{u}(\bar{y}) \quad \text{für } y \in S(x)\ .$$

Ebenso findet man einen "Weg" $\bar{x}, x_1, \ldots, x_m, \bar{\bar{x}}$ mit $x_k \in \Omega_h^\circ$ (inneren Gitterpunkten) von $\bar{x}$ zu einem Punkt $\bar{\bar{x}} \in \Omega_h^*$, so dass letztendlich

$$\bar{u}(\bar{x}) = \bar{u}(x_1) = \cdots = \bar{u}(x_m) = \bar{u}(\bar{\bar{x}}) = ||\bar{u}||_{\infty,h}$$

gilt. Damit ergibt sich unter Nutzung von Satz 4.17 und Satz 4.19 mit

$$||u||_{\infty,h} \leq ||\bar{u}||_{\infty,h} = \bar{u}(\bar{\bar{x}}) \leq \frac{\bar{f}(\bar{\bar{x}})}{d(\bar{\bar{x}})} \leq \max\{0, \max_{x \in \Omega_h^*} \frac{|f(x)|}{d(x)}\}$$

die Aussage des Satzes. □

Zum Abschluss wird die zentrale Konvergenzaussage formuliert.

Satz 4.21.
Unter der Voraussetzung, dass die kontinuierliche Lösung u des Problems

$$-\Delta u = f, \quad in\ \Omega, \quad u = \mu, \quad auf\ \Gamma\ ,$$

4-mal stetig partiell differenzierbar ist konvergiert die Lösung des Problems (4.23)

$$L_h u_h = f_h, \quad in\ \Omega_h, \quad u_h = \mu, \quad auf\ \Gamma_h\ , \tag{4.24}$$

gegen u, d.h. es gilt

$$||u_h - r_h u||_{\infty,h} \leq h^2 K$$

mit einer von h unabhängigen Konstante $K > 0$, wobei das Gebiet Ω als zusammenhängend vorausgesetzt wird.

5. Vorlesung am 15.06.2017

Beweis. Die Idee des Beweises nach **Samarskij** besteht darin, die Lösung von (4.24) durch

$$u_h(x) = u_1(x) + u_2(x)$$

mit

$$L_h u_1 = f,\ \text{in}\ \Omega_h,\ u_1|_{\Gamma_h} = 0, \quad L_h u_2 = 0,\ \text{in}\ \Omega_h,\ u_2|_{\Gamma_h} = \mu,$$

aufzuspalten. Mit dem Satz 4.18 hat man für u_2 die Abschätzung

$$||u_2||_{\infty,h} \leq ||\mu||_{\infty,\Gamma_h}\ .$$

u_1 wird nun nochmal zerlegt in

$$u_1(x) = u_1^\circ(x) + u_1^*(x)\ ,$$

mit

$$L_h u_1^\circ = f^\circ,\ \text{in}\ \Omega_h,\ u_1^\circ|_{\Gamma_h} = 0, \quad L_h u_1^* = f^*,\ \text{in}\ \Omega_h,\ u_1^*|_{\Gamma_h} = 0,$$

und

$$f(x) = f^\circ(x) + f^*(x)$$

mit

$$f^\circ(x) = \begin{cases} f(x), & x \in \Omega_h^\circ \\ 0 & x \in \Omega_h^* \end{cases}, \quad f^*(x) = \begin{cases} 0, & x \in \Omega_h^\circ, \\ f(x), & x \in \Omega_h^* \end{cases}.$$

u_1° und u_1^* werden nun separat abgeschätzt. Zur Abschätzung von u_1° soll das Vergleichsprinzip 4.17 angewandt werden und dazu wird die Funktion

$$\bar{u} = \alpha(d^2 - x_1^2 - x_2^2)$$

mit $\alpha > 0$ und $d \geq |x| = \sqrt{x_1^2 + x_2^2}$ für alle $x \in \Omega$, eingeführt. Für innere Punkte $x \in \Omega_h^\circ$ ergibt sich mit der Diskretisierung (4.14)

$$L_h \bar{u}(x) = \cdots = -4\alpha =: \bar{f}(x) \ ,$$

und für randnahe Punkte $x \in \Omega_h^*$ erhält man mit der Diskretisierung (4.15) etc.

$$L_h \bar{u}(x) = \cdots = -\alpha[(h_1^+ + h_1^-)/\bar{h}_1 + (h_2^+ + h_2^-)/\bar{h}_2] =: \bar{f}(x) \ ,$$

und damit eine Vergleichsaufgabe

$$L_h \bar{u}(x) = \bar{f}(x), \text{ in } \Omega_h, \ \bar{u}(x) = \alpha(d^2 - x_1^2 - x_2^2), \text{ auf } \Gamma_h \ ,$$

mit der Lösung $\bar{u}(x)$. Wegen $d \geq |x| = \sqrt{x_1^2 + x_2^2}$ gilt $\bar{u}(x) \geq 0$ auf Γ_h. Für $\alpha := \frac{1}{4}||f^\circ||_{\infty,\Omega_h}$ gilt

$$\begin{aligned} \bar{f}(x) &\geq |f^\circ(x)| \ , \quad x \in \Omega_h, \\ \bar{u}(x) &\geq 0 \ , \qquad x \in \Gamma_h \ , \end{aligned}$$

und damit nach Satz 4.17

$$||u_1^\circ||_{\infty,h} \leq ||\bar{u}||_{\infty,h} \leq \alpha d^2 = \frac{d^2}{4}||f^\circ||_{\infty,\Omega_h} \ .$$

Für $x \in \Omega_h^* \neq \emptyset$ überlegt man sich, dass aufgrund der Randbedingung $u_1^*|_{\Gamma_h} = 0$

$$\bar{d}(x) = a(x) - \sum_{y \in \bar{S}(x)} b(x,y) = \sum_{y \in \bar{S}(x), y \in \Gamma_h} b(x,y) \geq \frac{1}{\bar{h}^2}$$

mit $\bar{h} = \max\{h_1, h_2\}$ gilt. Mit dem Satz 4.20 folgt die Abschätzung

$$||u_1^*||_{\infty,h} \leq \bar{h}^2 ||f^*||_{\infty,\Omega_h} \ .$$

Damit erhalten wir insgesamt die Abschätzung

$$||u_h||_{\infty,h} \leq ||\mu||_{\infty,\Gamma_h} + \frac{d^2}{4}||f||_{\infty,\Omega_h^\circ} + \bar{h}^2 ||f||_{\infty,\Omega_h^*} \ . \tag{4.25}$$

Die gewonnene Abschätzung bedeutet gerade die Stabilität des FD-Schemas in der diskreten Maximum-Norm bezügl. der Randbedingungen und der rechten Seite. Mit der Glattheitsvoraussetzung an die kontinuierliche Lösung und der damit verbundenen Konsistenzordnung im Inneren ($O(h^2)$) und am Rand ($O(h^\beta)$, $\beta \geq 1$), also für $e_h = u_h - r_h u$

$$\begin{aligned} L_h e_h &= O(h^2) \ , \quad \text{in } \Omega_h^\circ \ , \\ L_h e_h &= O(h^\beta) \ , \quad \text{in } \Omega_h^* \ , \\ e_h &= 0 \ , \quad \text{auf } \Gamma_h \ , \end{aligned}$$

folgt die Konvergenzaussage des Satzes. □

4.2.3 Approximation von gemischten zweiten Ableitungen und allgemeineren Randbedingungen

Approximation von gemischten zweiten Ableitungen

Im Allg. kann man eine elliptische Differentialgleichung im $\mathbb{R}^2$ in der Form

$$-[a_{11}u_{xx} + 2a_{12}u_{xy} + a_{22}u_{yy}] + b_1u_x + b_2u_y + cu = f$$

aufschreiben, wobei $A = (a_{ij})$ eine symmetrische und positiv definite Matrix ist. Die gemischten zweiten Ableitungen könnte man durch die Diagonalisierung von A, also eine Hauptachsentransformation beseitigen. Allerdings kann man sie auch durch Differenzenquotienten approximieren. Man muss dann allerdings den Differenzenstern vergrößern, man braucht insgesamt 9 Punkte. Wenn die Standard-Diskretisierung von $-\Delta$ (durch den 5-Punkt-Stern) mit der Matrix

$$\begin{pmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{pmatrix}$$

beschreibt, dann könnte man die Diskretisierung der zweiten gemischten Ableitung $\frac{\partial^2 u}{\partial x \partial y}$

$$\frac{1}{2\bar{h}_2}[\frac{u(x+h,y+h) - u(x-h,y+h)}{2\bar{h}_1} - \frac{u(x+h,y-h) - u(x-h,y-h)}{2\bar{h}_1}]$$

durch die Matrix

$$\begin{pmatrix} -1 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & -1 \end{pmatrix}$$

beschreiben. Eine solche FD-Approximation ist von zweiter Ordnung, allerdings gibt es ein Problem. Denn das Differenzenschema führt dazu, dass die insgesamt resultierende Koeffizientenmatrix keine M-Matrix mehr ist. Um eine M-Matrix zu erhalten und auch um das diskrete Maximumprinzip anwenden zu können, könnte man den gesamten Differentialoperator

$$-(a_{11}u_{xx} + 2a_{12}u_{xy} + a_{22}u_{yy}) \tag{4.26}$$

z.B. durch

$$\frac{1}{h^2}\begin{pmatrix} a_{12}^- & -(a_{22} - |a_{12}|) & -a_{12}^+ \\ -(a_{11} - |a_{12}|) & 2(a_{11} + a_{22} - |a_{12}|) & -(a_{11} - |a_{12}|) \\ -a_{12}^+ & -(a_{22} - |a_{12}|) & a_{12}^- \end{pmatrix}$$

mit $a_{12}^+ = \max\{a_{12}, 0\}$ und $a_{12}^- = \min\{a_{12}, 0\}$ bei einer äquidistanten Diskretisierung von Ω approximieren. Eine kanonische Diskretisierung des Operators (4.26) erhält man auch durch

$$-\sum_{i=1}^{2} D_i^-(\sum_{j=1}^{2} a_{ij} D_j^+ u)$$

mit dem Differenzenstern

$$\frac{1}{h^2}\begin{pmatrix} a_{12} & -(a_{22}-a_{12}) & 0 \\ -(a_{11}-a_{12}) & 2(a_{11}+a_{22}-a_{12}) & -(a_{11}-a_{12}) \\ 0 & -(a_{22}-a_{12}) & a_{12} \end{pmatrix}$$

als Ergebnis.
Die Terme niederer Ordnung $b_1 u_x + b_2 u_y - cu$ kann man mit der Ordnung 2 durch

$$\frac{1}{2h}\begin{pmatrix} 0 & b_2 & 0 \\ -b_1 & -2hc & b_1 \\ 0 & -b_2 & 0 \end{pmatrix}$$

approximieren.
Diese Diskretisierung führt unter der Voraussetzung

$$a_{ii} > |a_{12}| + \frac{h}{2}|b_i| \,, \quad i = 1, 2,$$

auf eine M-Matrix und die numerische Lösung konvergiert mit der Ordnung 2 gegen die kontinuierliche exakte Lösung $u \in C^4(\bar{\Omega})$.

Approximation von allgemeineren Randbedingungen

Um z.B. die Vorgabe von Wärmeflüssen am Rand oder den Einfluss der Umgebungstemperatur in der Wärmebilanz eines Körpers berücksichtigen zu können, werden neben Dirichlet-Randbedingungen $u = \mu$ auf einem Teil Γ_d des Randes Γ auch Randbedingungen der Form

$$\frac{\partial u}{\partial \nu} + \alpha u = g \tag{4.27}$$

auf Γ_n betrachtet. ν ist dabei der äußere Normalenvektor, so dass man statt (4.27) auch

$$\operatorname{grad} u \cdot \nu + \alpha u = g$$

schreiben kann. Auf achsenparallelen Randstücken, z.B. auf einem Rand parallel zur y-Achse gilt $\operatorname{grad} u \cdot \nu = \pm \frac{\partial u}{\partial x}$ und man kann die Ableitung durch $D_1^{\pm} u$

approximieren. Auf gekrümmten Rändern muss man durch eine geeignete Interpolation die Richtungsableitung in ν-Richtung durch eine entsprechende Wichtung der partiellen Ableitungen approximieren. Die Abbildung zeigt ein gekrümmtes Randstück. Zur Approximation von $\frac{\partial u}{\partial \nu}$ bestimmt man am Punkt P' einen Wert $u(P')$ durch Interpolation unter Nutzung der Werte $u(P_1)$ und $u(P_2)$, so dass

$$\frac{\partial u}{\partial \nu} \approx \frac{u(P) - u(P')}{|PP'|}$$

mit $|PP'|$ als Abstand der Punktes P' vom Randpunkt P.

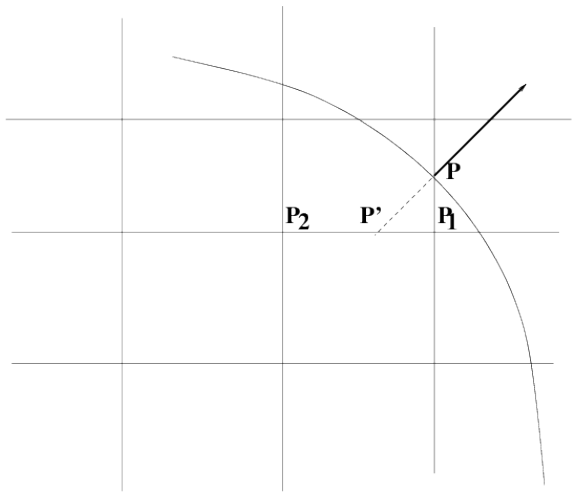

Abbildung 4.1: gekrümmter Rand mit äußerer Normalen ν

4.2.4 Entdimensionierung von partiellen Differentialgleichungen der mathematischen Physik

Dieses Thema bedeutet einen Einschub in die Thematik "Diskretiserung und numerische Lösungsverfahren" weil man bei der konkreten Behandlung von Anwendungsaufgaben mit echten physikalischen größen wie Geschwindigkeit, Temperatur, Zeit, konkreten räumlichen Abmessungen usw. zu tun hat.
Allerdings empfiehlt es sich, Größen geeignet zu skalieren oder auf charakteristische Größen zu beziehen, d.h. sie zu "entdimensionieren". Damit erreicht man auch, dass die dann dimensionslosen Größen bei der evtl. erforderlichen numerischen Untersuchung nicht sehr groß oder sehr klein werden, sondern

in einem Bereich liegen, der z.B. auf Computern genau abgebildet werden kann.
Wir wollen dies am Beispiel der Navier-Stokes-Gleichung demonstrieren. Die Navier-Stokes-Gleichung (Impulsbilanz eines inkompressiblen Mediums) lautet

$$\rho[\frac{\partial V}{\partial \tau} + (V \cdot \nabla)V] = -\nabla P + \eta \Delta V + F \;, \tag{4.28}$$

wobei V eine dimensionsbehaftete Geschwindigkeit ($[V] = \frac{m}{s}$) ist, und τ eine Zeit, gemessen in Sekunden ist. ρ ist die Dichte ($[\rho] = \frac{kg}{m^3}$ und η ist die dynamische Viskosität ($[\eta] = \frac{kg}{m \cdot s}$) und F ist eine äußere Volumenkraft, d.h. $[F] = \frac{kg}{m^2 \cdot s^2}$. Die räumlichen partiellen Ableitungen $\frac{\partial}{\partial X}$ beim Gradienten oder Laplace-Operator haben die physikalische Dimension $\frac{1}{m}$ und die zeitliche parielle Ableitung $\frac{\partial}{\partial \tau}$ hat die Dimension $\frac{1}{s}$. Entscheidend für die geeignete Dimensionierung oder Entdimensionierung sind charakteristische Größen des jeweiligen Problems, hier sind dies eine charakteristische bzw. typische Geschwindigkeit U_0 und eine charakteristische Länge L_0, z.B.

$$U_0 = 20\frac{m}{s} \;, \quad L_0 = 2\,m \;.$$

Wenn wir nun die dimensionslosen Geschwindikeiten und Längen mit

$$v = \frac{V}{U_0} \;, \quad x = \frac{X}{L_0}$$

und die mit U_0 und L_0 gebildete dimensionslose Zeit

$$t = \frac{\tau U_0}{L_0}$$

einführen, dann ergibts sich aus (4.28) und dem Ausklammern der Dimensionskonstanten

$$\rho\frac{U_0^2}{L_0}[\frac{\partial v}{\partial t} + (v \cdot \nabla)v] = -\frac{1}{L_0}\nabla P + \eta\frac{U_0}{L_0^2}\Delta v + F$$

und nach Division

$$\frac{\partial v}{\partial t} + (v \cdot \nabla)v = -\frac{1}{\rho U_0^2}\nabla P + \frac{\eta L_0}{\rho U_0}\Delta v + \frac{L_0}{\rho U_0^2}F \tag{4.29}$$

Wenn wir noch die Größen

$$p = \frac{1}{\rho U_0^2}P \;, \quad f = \frac{L_0}{\rho U_0^2}F$$

als dimensionslosen Druck bzw. dimensionslose Volumenkraft einführen, sowie mit

$$Re = \frac{\rho U_0}{\eta L_0} = \frac{U_0}{\nu L_0}$$

die Reynoldszahl als dimensionslose Kennzahl einführen ($\nu = \frac{\eta}{\rho}$ ist die kinematische Viskosität), dann erhält man schließlich die dimensionslose Navier-Stokes-Gleichung

$$\frac{\partial v}{\partial t} + (v \cdot \nabla)v = -\nabla p + \frac{1}{Re}\Delta v + f \ . \tag{4.30}$$

Dabei sind ∇ und Δ hier im Sinne von dimensionslosen Ableitungen $\frac{\partial}{\partial x}$ zu verstehen.
Die Reynoldszahl gibt bei Strömungsproblemen Auskunft über das Verhältnis von viskosen Gliedern zu den nichtlinearen Beschleunigungsgliedern. Kurz gesagt bedeuten große Reynoldszahlen, dass es sich um turbulente Strömungen handelt, während bei laminaren Strömungen kleine Reynoldsdzahlen auftreten.

4.3 Finite-Differenzen-Verfahren für parabolische Differentialgleichungen

6. Vorlesung am 22.06.2017

Im Folgenden wird die numerische Lösung von Rand-Anfangswert-Problemen mit Finite-Differenzen-Verfahren diskutiert. Dabei wird die Pipeline "Konsistenz-Stabilität-Konvergenz" für ein räumlich eindimensionales Problem vollständig dargestellt. Desweiteren wird auf mehrdimensionale Aufgaben und auf die gebräuchlichsten Stabilitätskonzepte eingegangen.

4.3.1 FD-Schemen für eindimensionale Rand-Anfangswertprobleme

Als Modellproblem betrachten wir das räumlich eindimensionale Rand-Anfangswertproblem

$$\begin{aligned} \frac{\partial u}{\partial t}(x,t) - \frac{\partial^2 u}{\partial x^2}(x,t) + cu(x,t) = f(x,t) \ , \quad x \in]0,1[, 0 < t \leq T, \\ u(0,t) = g_1(t),\ u(1,t) = g_2(t) \ , \quad 0 < t \leq T, \\ u(x,0) = u_0(x) \ , \quad x \in]0,1[\ . \end{aligned} \tag{4.31}$$

Gesucht ist eine diskrete Lösung, d.h. eine Gitterfunktion $u_{h,\tau}$

$$u_{h,\tau} : \bar{Q}_{h,\tau} \to \mathbb{R} \ ,$$

wobei

$$\bar{Q}_{h,\tau} = \{(x_i, t^k) | x_i = i\,h,\, t^k = k\,\tau\,, i = 0, \ldots, N,\; k = 0, \ldots, M\}\,,$$
$$h = 1/N,\; \tau = 1/M,\; N, M \in \mathbb{N}\,,$$

eine Diskretisierung des Raum-Zeit-Zylinders $\bar{Q} = [0,1] \times [0,T]$ ist. Die Werte der Gitterfunktion am Punkt (x_i, t^k) bezeichnet man mit u_i^k. Wenn man zur Approximation der Zeitableitung

$$D_t^+ u_i^k := \frac{u_i^{k+1} - u_i^k}{\tau}$$

einführt, diskretisieren wir das parabolische Problem 4.31 durch

$$\begin{aligned} D_t^+ u_i^k = \\ \sigma(D^- D^+ u_i^{k+1} - c u_i^{k+1} + f_i^{k+1}) + (1-\sigma)(D^- D^+ u_i^k) - c u_i^k + f_i^k)\,, \\ i = 1, \ldots, N-1,\; k = 0, \ldots, M-1 \end{aligned} \quad (4.32)$$

mit einem freien Parameter $\sigma \in [0,1]$, und den diskreten Anfangs- und Randbedingungen

$$u_i^0 = u_0(x_i)\,, \quad u_0^k = g_1(t^k),\; u_N^k = g_2(t^k)\,. \quad (4.33)$$

f_i^k sei dabei eine geeignete Approximation von $f(x_i, t^k)$. Mit der Wahl von σ legt man das konkrete Differenzen-Schema fest. Exemplarisch seien die folgenden in der Praxis oft genutzten Schemen genannt:

- Das explizite Euler-Schema für $\sigma = 0$

$$u_i^{k+1} = (1 - 2\gamma - \tau c) u_i^k + \gamma(u_{i-1}^k + u_{i+1}^k) + \tau f(x_i, t^k)\,; \quad (4.34)$$

- Das implizite Euler-Schema für $\sigma = 1$

$$(1 + 2\gamma + \tau c) u_i^{k+1} - \gamma(u_{i-1}^{k+1} + u_{i+1}^{k+1}) = u_i^k + \tau f(x_i, t^{k+1})\,; \quad (4.35)$$

- Das Crank-Nicolson-Verfahren für $\sigma = \frac{1}{2}$

$$\begin{aligned} & 2(\gamma + 1 + \tau c) u_i^{k+1} - \gamma(u_{i-1}^{k+1} + u_{i+1}^{k+1}) \\ = \;& 2(1 - \gamma - \tau c) u_i^k + \gamma(u_{i+1}^k + u_{i-1}^k) + \tau(f(x_i, t^k) + f(x_i + t^{k+1}))\,. \end{aligned} \quad (4.36)$$

Dabei wurde die Notation $\gamma = \frac{\tau}{h^2}$ verwendet. Zur Konsistenz des Differenzen-Schemas (4.32), (4.33) notieren wir den folgenden

Satz 4.22.
Das Differenzen-Schema (4.32), (4.33) *hat die folgende Konsistenzordnung in der Maximum-Norm:*
(a) $O(h^2+\tau)$ *für alle* $\sigma \in [0,1]$ *und* $f_i^k = f(x_i, t^k)$ *vorausgesetzt* $u \in C^{4,2}(\bar{Q})$
(b) $O(h^2+\tau^2)$ *für* $\sigma = \frac{1}{2}$ *und* $f_i^k = f(x_i, t^k + \frac{\tau}{2})$ *vorausgesetzt* $u \in C^{4,3}(\bar{Q})$ *,*
wobei $C^{l,m}(\bar{Q})$ *der Raum der* l*-mal nach* x *und* m*-mal nach* t *stetig differenzierbaren Funktionen auf* $\bar{Q}$ *bezeichnet.*

Beweis. Übung □

Um Konvergenzaussagen zu formulieren benötigt man einen Stabilitätsbegriff. Zur Erarbeitung eines geeigneten Stabilitätsbegriffs gehen wir von homogenen Randbedingungen aus (keine wirkliche Einschränkung der Allgemeinheit). Das Schema 4.32 schreibt man dazu in der Form

$$-\sigma\gamma u_{i-1}^{k+1} + (2\sigma\gamma + 1 + \tau c)u_i^{k+1} - \sigma\gamma u_{i+1}^{k+1} = F_i^k$$

mit

$$F_i^k = (1-\sigma)\gamma u_{i-1}^k + (1 - 2(1-\sigma)\gamma - \tau c)u_i^k + (1-\sigma)\gamma u_{i+1}^k + \tau f_i^k$$

auf. Wegen

$$(2\sigma\gamma + 1 + \tau c) > 2\sigma\gamma \Longleftrightarrow |a_{ii}| > \sum_{j=1, j\neq i}^{n} |a_{ij}| \,,$$

also der strikten Diagonaldominanz erhält man

$$\max_i |u_i^{k+1}| \leq \frac{1}{1+\tau c} \max_i |F_i^k| \leq \max_i |F_i^k| \,.$$

Mit $\sigma \in [0,1]$ und unter der Voraussetzung $1 - 2(1-\sigma)\gamma - \tau c \geq 0$ erhält man

$$\begin{aligned}
\max_i |F_i^k| &\leq |(1-\sigma)\gamma + (1 - 2(1-\sigma)\gamma - \tau c) + (1-\sigma)\gamma| \max_i |u_i^k| \\
&\qquad + \tau \max_i |f_i^k| \\
&= |1 - \tau c| \max_i |u_i^k| + \tau \max_i |f_i^k| \,, \quad \text{also} \\
\max_i |u_i^{k+1}| &\leq \max_i |F_i^k| \leq \max_i |u_i^k| + \tau \max_i |f_i^k| \,.
\end{aligned}$$

Unter der Voraussetzung $|1 - \tau c| \leq 1$ erhält man durch die sukzessive Anwendung der Ungleichung

$$\max_k \max_i |u_i^{k+1}| \leq \max |u_0(x)| + \tau \sum_{j=0}^{k} \max_i |f_i^j| \,. \tag{4.37}$$

Die durchgeführten Überlegungen rechtfertigen die

Definition 4.23. *(Stabilität in der Maximumnorm)*
Ein Differenzen-Schema 4.32 zur Lösung der Anfangs-Randwertaufgabe 4.31 heißt stabil in der Maximum-Norm, wenn die Ungleichung 4.37 gilt.

Damit stellt man fest, dass das Schema 4.32 stabil ist, wenn die Bedingung

$$1 - 2(1-\sigma)\gamma - \tau c \geq 0 \Longleftrightarrow (1-\sigma)\frac{\tau}{h^2} + \frac{\tau c}{2} \leq 1 \tag{4.38}$$

erfüllt ist. Zur Konvergenz der Lösung u_i^k gegen die exakte Lösung kann man nun folgende Aussage formulieren.

Satz 4.24. *(Konvergenz des Differenzen-Schemas 4.32)*
Sei $f_i^k = f(x_i, t^k)$ für alle i, k. Es wird

$$(1-\sigma)\frac{\tau}{h^2} + \frac{\tau c}{2} \leq 1$$

und $u \in C^{4,2}(\bar{Q})$ vorausgesetzt. Dann existiert eine Konstante $C > 0$ mit

$$\max_{i,k} |u(x_i, t^k) - u_i^k| \leq C(h^2 + \tau) \; .$$

Gilt darüberhinaus $\sigma = \frac{1}{2}$ und $u \in C^{4,3}(\bar{Q})$ sowie $f_i^k = f(x_i, t^k + \frac{\tau}{2})$, dann gilt

$$\max_{i,k} |u(x_i, t^k) - u_i^k| \leq C(h^2 + \tau^2) \; .$$

Der Beweis des Satzes 4.24 ergibt sich unmittelbar unter Nutzung der Konsistenz- und Stabilitätseigenschaft des Schema 4.32.

Bemerkung 4.25.
Für die Stabilität und Konvergenz des Differenzen-Schemas in der Maximum-Norm musste die Bedingung 4.38 erfüllt werden, d.h. auch im Falle des impliziten Eulerverfahrens $\sigma = 1$) und des Crank-Nicolson-Verfahrens ($\sigma = \frac{1}{2}$) hat man keine unbedingte Stabilität.
Anders ist es im Falle der L_2-Stabilität. Man kann dann für das Modellbeispiel 4.32 zeigen, dass

$$||u^{k+1}||_{0,h} \leq ||u^0||_{0,h} + K \sum_{j=0}^{k} ||f^j||_{0,h}$$

gilt, d.h. es liegt Stabilität bezügl. der Anfangsbedingung und der rechten Seite vor (für den Fall homogener Randbedingungen). Dabei erhält man unbedingte Stabilität für die Fälle $\sigma = 1$ (impl. Eulerverfahren) und $\sigma = \frac{1}{2}$ (Crank-Nicolson-Verfahren). Für das expl. Eulerverfahren ($\sigma = 0$) muss man

allerdings die Stabilitäts-Bedingung $\frac{\tau}{h^2} \leq \frac{1}{2}$ erfüllen.
Daraus ergeben sich auch entsprechende Konvergenzaussagen in der diskreten L_2-Norm. Die Ergebnisse in der L_2-Norm erfordern zwar weniger harte Bedingungen an die Diskretisierung, bedeuten aber nur Aussagen "im quadratischen Mittel", d.h. Konvergenz in der diskreten L_2-Norm schließt große punktuelle Unterschiede zwischen der exakten Lösung u und der numerischen Lösung u_i^k nicht aus, was für viele praktischen Aufgabenstellungen nicht akzeptabel ist.

4.3.2 Von Neumann-Stabilitätsanalyse

J. v. Neumann hat die Auswirkungen von harmonischen Anfangsstörungen bei der numerischen Lösung zeitabhängiger Differentialgleichungen untersucht. Die Technik wird auch **formale Fourier-Stabiltätsanalyse** oder **von Neumann-Stabilitätsanalyse** genannt.
Man betrachtet die Entwicklung einer Anfangsstörung in eine Fourierreihe

$$u_j^0 = \sum_l V_l^0 e^{\mathbf{i}\omega_l jh}$$

zum Beispiel bei einem expliziten Verfahren für die Lösung einer eindimensionalen Wärmeleitungsgleichung

$$\begin{aligned} D_t^+ u_j^k &= D_x^+ D_x^- u_j^k \\ u_j^{k+1} = (1-2\gamma)u_j^k + \gamma(u_{j+1}^k + u_{j-1}^k)\ , &\quad \gamma = \frac{\tau}{h^2}\ ., \end{aligned} \tag{4.39}$$

wobei man der Einfachheit halber von periodischen Randbedingungen ausgeht und untersucht die zeitliche Entwicklung der Anfangsstörung. Wegen der Linearität von (4.39) reicht es die zeitliche Entwicklung der einzelnen Summanden

$$V_l^k e^{\mathbf{i}\omega_l jh} \quad \text{oder} \quad V^k e^{\mathbf{i}\omega jh}$$

mit V als einem von der Wellenlänge ω abhängigen Verstärkungsfaktor zu untersuchen. Es interessiert also die Entwicklung des Summanden

$$V^k e^{\mathbf{i}\omega jh}\ . \tag{4.40}$$

für wachsendes k. Von Stabilität spricht man, wenn die Störung nicht explodiert, d.h. wenn $|V| \leq 1$ ist.

Setzt man (4.40) in (4.39) ein, dann findet man nach Division durch $V^k e^{\mathbf{i}\omega jh}$

$$\begin{aligned} V &= 1 - 2\gamma + 2\gamma(e^{\mathbf{i}\omega h} + e^{-\mathbf{i}\omega h}) \\ &= 1 - 2\gamma + 2\gamma\cos(\omega h) \\ &= 1 - 4\gamma\sin^2\frac{\omega h}{2} \quad \text{d.h.} \\ |V| &\leq 1 \Longleftrightarrow -1 \leq -4\gamma\sin^2\frac{\omega h}{2} \leq 1 \Longleftrightarrow \gamma\sin^2\frac{\omega h}{2} \leq \frac{1}{2} \end{aligned} \tag{4.41}$$

Damit liegt Stabilität vor, wenn $\gamma = \frac{\tau}{h^2} \leq \frac{1}{2}$ ist.
Für das vollständig implizite Schema

$$u_j^{k+1} = u_j^k + \gamma(u_{j+1}^{k+1} + u_{j-1}^{k+1} - 2u_j^{k+1})$$

erhält man nach dem Einsetzen von (4.40) und nach einer kurzen Rechnung

$$V = 1 + V\gamma(2\cos(\omega h) - 2)$$

bzw.

$$V(1 - \gamma(2\cos(\omega h) - 2)) = V(1 + \gamma 4\sin^2\frac{\omega h}{2}) = 1$$

und damit

$$V = \frac{1}{1 + \gamma 4\sin^2\frac{\omega h}{2}} \ ,$$

d.h. das implizite Schema ist unbedingt stabil.
Die von Neumann-Stabilitätsanalyse wird auch bei hyperbolischen Problemen zur Bewertung von Differenzenschemata benutzt. Außerdem ist die **von Neumann**-Stabilitätsanalyse nicht auf den räumlich eindimensionalen Fall beschränkt. Hat man es mit 2 oder 3 Raumdimensionen zu tun, dann muss man z.B. im zweidimensionalen Fall von einer Entwicklung der numerischen Lösung $u_{j,k}^n = u_h(x_j, y_k, t_n)$ in der Form

$$u_{j,k}^n = V^n e^{\mathbf{i}\theta j} e^{\mathbf{i}\kappa k} \tag{4.42}$$

ausgehen, wobei θ und κ die Wellenlängen in x- bzw. y-Richtung sind, und V ein von θ und κ abhängiger Verstärkungsfaktor ist.

Bemerkung 4.26. Obwohl die **von Neumann**-Stabilitätsanalyse nur für lineare Probleme gültig ist, wird sie auch oft auf nichtlineare Probleme angewandt. Das gleiche gilt für nicht-periodische Randbedingunen und oft reicht die lokale Analyse im Innern aus, um notwendige Bedingungen für die Stabilität zu erhalten oder Instabilität zu zeigen.
Probleme treten bei sehr kleinen und sehr großen Wellenlängen $\frac{b}{k}$ ($\theta \approx \pi$, $\theta \approx 0$) auf. Bei kleinen Wellenlängen "hilft" eine Dämpfung durch die Einführung einer künstlichen Viskosität, um Verfahren zu stabilisieren.

4.3.3 Parabolische Probleme in höheren Dimensionen

Die Ergebnisse bei der Diskretisierung mit Finiten Differenzen räumlich eindimensionaler zeitabhängiger Probleme kann man auf höhere Raumdimensionen übertragen. Für das Problem

$$\begin{aligned} \frac{\partial u}{\partial t} - \Delta\, u = f \quad \text{in} \quad \Omega\times]0,T], \\ u = 0 \quad \text{auf} \quad \Gamma\times]0,T], \\ u(t,0) = u_0(t) \quad \text{in} \quad \Omega\,, \end{aligned} \tag{4.43}$$

mit $\Omega \subset \mathbb{R}^2$ kann man das Differenzenschema

$$D_t^+ u_{ij}^k - [D_x^- D_x^+ + D_y^- D_y^+]((1-\sigma)u_{ij}^k + \sigma u_{ij}^{k+1}) = f_{ij}^k \quad (\sigma \in [0,1]), \tag{4.44}$$

bzw.

$$(E - \tau\sigma[D_x^- D_x^+ + D_y^- D_y^+])u_{ij}^{k+1} = (E + \tau(1-\sigma)[D_x^- D_x^+ + D_y^- D_y^+])u_{ij}^k + \tau f_{ij}^k$$

formulieren (u_{ij}^k steht für den Wert $u(x_i, y_j, t^k)$ einer Gitterfunktion). Für $\sigma > 0$ bedeutet (4.44) unter Hinzunahme der Randbedingungen pro Zeitschritt die Lösung eines linearen Gleichungssystems mit eine strikt diagonal dominanten Matrix $A = (E - \tau\sigma\Delta_h)$, die auch L- und M-Matrix ist, wobei

$$\Delta_h u = [D_x^- D_x^+ + D_y^- D_y^+]u$$

gilt. Die Matrix A ist schwach besetzt und hat bei geeigneter Nummerierung 5 Nichtnull-Diagonalen. Die folgende Rechnung

$$\begin{aligned} E - \tau\sigma\Delta_h &= E - \tau\sigma(D_x^- D_x^+ + D_y^- D_y^+) \\ &= (E - \tau\sigma D_x^- D_x^+)(E - \tau\sigma D_y^- D_y^+) - \tau^2\sigma^2 D_x^- D_x^+ D_y^- D_y^+ \end{aligned}$$

bedeutet, dass

$$(E - \tau\sigma D_x^- D_x^+)(E - \tau\sigma D_y^- D_y^+)u_{ij}^{k+1} = (E + \tau(1-\sigma)\Delta_h)u_{ij}^k + \tau f_{ij}^k \tag{4.45}$$

eine konsistente Finite-Differenzen-Approximation der Gleichung (4.43) ist. Die Gleichung (4.45) kann man wie folgt

$$(E - \tau D_x^- D_x^+)u_{ij}^{k+1/2} = (E + \tau(1-\sigma)\Delta_h)u_{ij}^k + \tau f_{ij}^k \tag{4.46}$$

$$(E - \tau D_y^- D_y^+)u_{ij}^{k+1} = u_{ij}^{k+1/2} \tag{4.47}$$

in zwei Schritten lösen, wobei jeweils tridiagonale Gleichungssysteme zu lösen sind. Die Methodik wird **ADI-Verfahren**, fractional-step method oder

Zwischenschritt-Methode genannt. ADI steht dabei für **alternating direction implicit**.
Wie im räumlich eindimensionale Fall kann man für das Differenzen-Schema (4.44) unter bestimmten Glattheitsannahmen der exakten Lösung des Problems (4.43) die Konsistenz, Stabilität und Konvergenz in der Maximum- und L_2-Norm zeigen, wobei beim expliziten Fall ($\sigma = 0$) Bedingungen zwischen zeitlichen und räumlichen Diskretisierungsparametern (τ, $\Delta x = h$, $\Delta_y = k$) zu berücksichtigen sind.

Kapitel 5

Finite-Volumen-Methode

7. Vorlesung am 29.06.2017

Bei der Finiten-Differenzen-Methode wurden die Ableitungen in der jeweiligen partiellen Differentialgleichung durch Differenzenquotienten approximiert und im Falle der Konsistenz und Stabilität konnten Konvergenzaussagen gemacht werden. Bei der Finite-Volumen-Methode bilanziert man die Differentialgleichung über das Gebiet Ω ihrer Gültigkeit bzw. über lokale Bilanzbereiche Ω_i, in die Ω unterteilt wird. Auf der Grundlage des Gaußschen Integralsatzes erhält man durch die Gesamtflussbilanz über den jeweiligen Rand der lokalen Bilanzbereiche Ω_i letztendlich Gleichungen zur Berechnung von Näherungslösungen an bestimmten Stützpunkten in Ω_i. Das soll im Folgenden präzisiert werden.

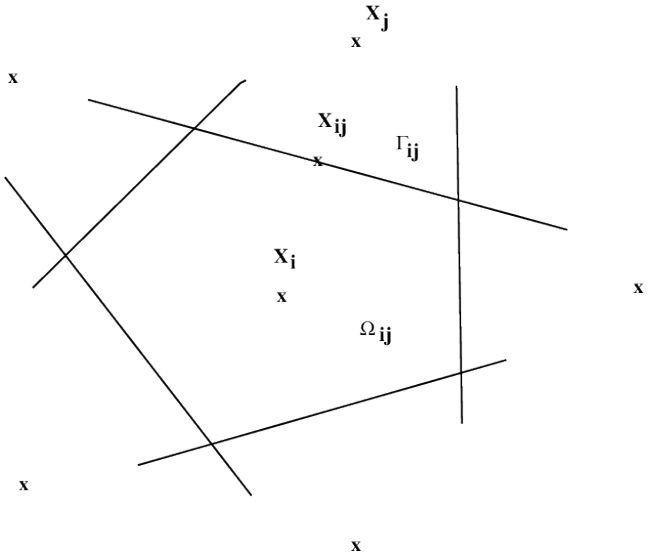

Abbildung 5.1: Voronoi-Box im $\mathbb{R}^2$

5.1 Grundlagen der FV-Methode und Definition

Zur Erläuterung der FVM betrachetn wir das elliptische Modellproblem

$$-\Delta u = f \ \text{ in } \ \Omega\,, \quad u = g \ \text{ auf } \ \Gamma = \partial\Omega\,, \tag{5.1}$$

wobei $\Omega \subset \mathbb{R}^n$ als beschränktes polygonales Gebiet angenommen wird. In Ω und auf Γ seien die Stützpunkte

$$x_i \in \Omega,\ i = 1, \ldots, N \quad \text{und} \quad x_i \in \Gamma,\ i = N+1, \ldots, M$$

gegeben und mit

$$J := \{1, \ldots, N\}\,, \quad \bar{J} := J \cup \{N+1, \ldots, M\}$$

seien die entsprechenden Indexmengen bezeichnet. Mit Hilfe von

$$B_{ij} = \{x \in \Omega \,|\, |x - x_i| < |x - x_j|\}\,, \quad J_i = \bar{J} \setminus \{i\}$$

definiert man mit

$$\Omega_i = \bigcap_{j \in J_i} B_{ij} \tag{5.2}$$

die **Voronoi-Box** Ω_i. In Abb. 5.1 ist eine solche Voronoi-Box skizziert. Es wird eine Stützpunktwahl vorausgesetzt, so dass $\mu_{n-1}(\bar{\Omega}_i \cap \Gamma) = 0$ f.a. $i \in J$ gilt (μ_k ist das Lebesgue-Maß im R^k). Durch

$$N_i = \{j \in J_i \,|\, \mu_{n-1}(\bar{\Omega}_j \cap \bar{\Omega}_i) > 0\}$$

ist die Index-Menge der maßgeblichen Nachbarpunkte von x_i bezeichnet. Durch

$$\Gamma_{ij} = \bar{\Omega}_i \cap \bar{\Omega}_j\,, \quad i \in J,\, j \in N_i,$$

wird die gemeinsame "Kante" der benachbarten Voronoi-Boxen Ω_i und Ω_j bezeichnet und

$$\Gamma_i = \bigcup_{j \in N_i} \Gamma_{ij}\,, \quad i \in J,$$

ergibt den Rand von Ω_i. Unter der Voraussetzung der eindeutigen Lösbarkeit von (5.1) mit der Lösung u gilt auch die Bilanz

$$-\int_{\Omega_i} \Delta\, u \, d\Omega_i = \int_{\Omega_i} f \, d\Omega_i\,, \quad i \in J\,.$$

Mit dem Gaußschen Integralsatz folgt für die Lösung u auch

$$-\int_{\Gamma_i} \frac{\partial u}{\partial n_i}\, d\Gamma_{ij} = \int_{\Omega_i} f\, d\Omega_i \quad \text{bzw.}$$
$$-\sum_{j\in N_i} \int_{\Gamma_{ij}} \frac{\partial u}{\partial n_{ij}}\, d\Gamma_{ij} = \int_{\Omega_i} f\, d\Omega_i\ , \quad i \in J\ , \tag{5.3}$$

wobei $\frac{\partial u}{\partial n_{ij}}$ Flüsse sind und n_i bzw. n_{ij} die äußeren Normalen auf Γ_i bzw. Γ_{ij} bezeichnen.
Der Schnittpunkt der Verbindungsstrecke der Punkte x_i und x_j, bezeichnet durch $[x_i, x_j]$, mit Γ_{ij} wird x_{ij} genannt. Mit dem Normalenvektor $n_{ij} = \frac{x_j - x_i}{|x_j - x_i|}$ erhält man durch

$$\frac{\partial u}{\partial n_{ij}}(x_{ij}) = \frac{u(x_j) - u(x_i)}{|x_j - x_i|} + O(|x_j - x_i|^2)\ , \quad j \in N_i,\ i \in J\ ,$$

eine kanonische Approximation zweiter Ordnung (es handelt sich um einen zentralen Differenzenquotienten) des Flusses durch Γ_{ij}. Darauf basierend erhält man mit

$$m_{ij} = \mu_{n-1}(\Gamma_{ij}) \quad \text{und} \quad d_{ij} = |x_j - x_i|$$

durch

$$-\sum_{j\in N_i} \frac{m_{ij}}{d_{ij}}(u_j - u_i) = \int_{\Omega_i} f\, d\Omega_i\ , \quad i \in J, \tag{5.4}$$

eine Approximation von (5.3). Mit den Randbedingungen

$$u_i = g(x_i)\ , \quad i \in \bar{J} \setminus J\ , \tag{5.5}$$

erhält man letztlich ein lineares Gleichungssystem

$$A_h u_h = f_h \tag{5.6}$$

zur Berechnung der Näherungslösung $u_h = (u_i)_{i\in J} \in \mathbb{R}^N$, für $u(x_i)$, $i \in J$. $A_h = (a_{ij})$ und $f_h = (f_i)$ ergeben sich aus (5.4) zu

$$a_{ij} = \begin{cases} \sum_{j\in N_i} \frac{m_{ij}}{d_{ij}} & i = j, \\ -\frac{m_{ij}}{d_{ij}} & j \in N_i \cap J, \\ 0 & \text{sonst} \end{cases} \quad \text{und} \quad f_i = \int_{\Omega_i} f\, d\Omega_i + \sum_{l\in N_i\setminus J} \frac{m_{il}}{d_{il}} g(x_i)\ .$$

Bemerkung 5.1. Für den Fall von rechteckigen Voronoi-Boxen Ω_i gleicher Größe evtl. eines Rechteckgebietes $\Omega \in \mathbb{R}^2$ bedeutet (5.4) bzw. (5.6) mit Ausnahme der Behandlung der rechten Seiten f_i ein klassisches FD-Schema, dass man aus (5.6) erhält, indem man mit dem reziproken Flächeninhalt $\frac{1}{h\cdot k}$ der Boxen Ω_i (h Breite, k Höhe) durchmultipliziert.

5.2 Existenz und Eindeutigkeit der FVM-Lösung

Das lineare Gleichungssystem (5.6) hat eine Koeffizientenmatrix $A_h = (a_{ij})$, die irreduzibel diagonal dominant ist, und damit ist die Existenz einer eindeutigen FVM-Lösung gesichert.

5.3 Konsistenz und Konvergenz der FV-Methode

Die Konsistenz- und Konvergenzuntersuchungen werden in der Maximumnorm durchgeführt. Grundlage ist dabei wie im Fall der FD-Methoden ein diskretes Maximumprinzip. Im Folgenden werden die entscheidenden Mittel zu den Nachweisen dargestellt.

Satz 5.2. *(diskretes Maximumprinzip)*
Das Problem (5.6) *genügt dem diskreten Maximumprinzip, d.h. es gilt*

$$\left.\begin{array}{rl} \sum_{j\in N_i} \frac{m_{ij}}{d_{ij}}(u_j - u_i) \le 0\ , & i \in J, \\ u_i \le 0\ , & i \in \bar{J} \setminus J \end{array}\right\} \Longrightarrow u_i \le 0\ , \quad i \in \bar{J}\ .$$

Beweis. Sei $k \in \bar{J}$ ein Index mit

$$u_k \ge u_j \quad \text{für alle } j \in \bar{J}\ . \tag{5.7}$$

Annahme: $u_k > 0$. Dann muss $k \in J$ wegen der vorausgesetzten diskreten Randbedingung $u_i \le 0\ , i \in \bar{J} \setminus J$ gelten. Weiterhin gibt es eine Kette von miteinander verbundenen Voronoi-Boxen Ω_i bis zum Rand Γ mit $u_i \le 0$ und man findet in dieser Folge auf jeden Fall einen Index $\bar{k}$ und einen Nachbarn $l \in N_{\bar{k}}$, so dass $u_{\bar{k}} > u_l$ gilt. Wegen (5.7) folgt dann

$$\sum_{j\in N_{\bar{k}}} \frac{m_{\bar{k}j}}{d_{\bar{k}j}} u_{\bar{k}} > \sum_{j\in N_{\bar{k}}} \frac{m_{\bar{k}j}}{d_{\bar{k}j}} u_j \Longleftrightarrow -\sum_{j\in N_{\bar{k}}} \frac{m_{\bar{k}j}}{d_{\bar{k}j}}(u_j - u_{\bar{k}}) > 0\ ,$$

was der Voraussetzung des Satzes widerspricht, d.h. unsere Annahme $u_k > 0$ trifft nicht zu, was den Beweis abschließt. □

Die Aussage des Satzes 5.2 wird im Folgenden benutzt, um durch eine geeignete Vergleichsfunktion eine Konsistenzaussage für das FV-Verfahren (5.6) nachzuweisen. Dazu dient das folgende Lemma.

Lemma 5.3.
Sei $v : \mathbb{R}^N \to \mathbb{R}$ mit Parametern $\alpha, \beta \in \mathbb{R}$ durch

$$v(x) = -\frac{\alpha}{2}|x|^2 + \beta$$

definiert ($|\cdot|$ ist hierbei die Euklidische Norm im $\mathbb{R}^N$, $N = card(J)$). Dann gilt

$$-\sum_{j \in N_i} \frac{m_{ij}}{d_{ij}}(v(x_j) - v(x_i)) = N\alpha \int_{\Omega_i} d\Omega_i \ , \quad i \in J \ .$$

Beweis. Für den Gradienten und den Laplace-Operator findet man für v

$$\nabla v = -\alpha x \quad \text{und} \quad -\Delta v = N\alpha \quad \text{für alle } x \in \mathbb{R}^n \ . \tag{5.8}$$

Mit dem Satz von Gauß folgt weiter

$$-\sum_{j \in N_i} \int_{\Gamma_{ij}} \frac{\partial v}{\partial n_{ij}} d\Gamma_{ij} = -\int_{\Omega_i} \Delta v \, d\Omega_i = N\alpha \int_{\Omega_i} d\Omega_i \ , \quad i \in J \ . \tag{5.9}$$

Für eine quadratische Funktion q gilt

$$\frac{q(b) - q(a)}{b - a} = q'(\frac{a+b}{2}) \ ,$$

also auch

$$-\frac{v(x_j) - v(x_i)}{d_{ij}} = -\frac{\partial v}{\partial n_{ij}}(x_{ij}) = -\nabla v(x_{ij}) \cdot n_{ij} \ ,$$

da x_{ij} der Mittelpunkt der Verbindungsstrecke von x_j nach x_i der Länge d_{ij} ist. Mit (5.8) und der Orthogonalität von n_{ij} und Γ_{ij} gilt dann

$$\begin{aligned} -\frac{v(x_j) - v(x_i)}{d_{ij}} &= \alpha x_{ij} \cdot n_{ij} = \alpha x \cdot n_{ij} + \alpha(x_{ij} - x) \cdot n_{ij} \\ &= \alpha x \cdot n_{ij} = -\frac{\partial v}{\partial n_{ij}}(x) \quad \text{für alle } x \in \Gamma_{ij} \ . \end{aligned}$$

Damit erhält man

$$\begin{aligned} -\sum_{j \in N_i} \frac{m_{ij}}{d_{ij}}(v(x_j) - v(x_i)) &= -\sum_{j \in N_i} \int_{\Gamma_{ij}} \frac{\partial v}{\partial n_{ij}} d\Gamma_{ij} \\ &= -\int_{\Omega_i} \Delta v \, d\Omega_i = N\alpha \int_{\Omega_i} d\Omega_i \ , \quad i \in J \ , \end{aligned}$$

wegen (5.9). □

8. Vorlesung am 06.07.2017

Für die Untersuchung der Konsistenz der FV-Diskretisierung in der Maximumnorm werden Diskretisierungsparameter

$$h_i = (\max_{j \in N_i} \mu_{n-1}(\Gamma_{ij}))^{1/(n-1)} , \quad h = \max_{i \in J} h_i \tag{5.10}$$

eingeführt. Zusätzlich zu den Voraussetzungen über eine geeignete Stützpunktverteilung fordern wir für $h \leq h_0$ (mit $h_0 > 0$) die Erfüllung der folgenden Bedingungen für die Voronoi-Boxen:

(B1) Die Zahl der maßgeblichen Nachbarn jedes Punktes x_i ist gleichmäßig beschränkt;

(B2) Jeder Punkt $x_{ij} = [x_i, x_j] \cap \Gamma_{ij}$ liegt auf dem Schwerpunkt von Γ_{ij}.

Die Bedingung (B1) ist keine wirkliche Einschränkung und ist problemlos erfüllbar. Die Bedingung (B2) bedeutet eine stärkere Einschränkung, denn damit fordert man mehr oder weniger homogene Voronoi-Boxen. Der Verzicht auf (B2) ist möglich, verkompliziert aber den nachfolgenden Beweis erheblich, und deshalb fordern wir aus Gründen einer besser nachvollziehbaren Beweisführung (B2).

Satz 5.4. *(Konsistenz der FVM)*
Seien (B1) und (B2) erfüllt. Die Lösung u von (5.1) liege in $C^4(\bar{\Omega})$. Dann gibt es eine Konstante c, so dass

$$|\sum_{j \in N_i} \frac{m_{ij}}{d_{ij}}(u(x_j) - u(x_i)) + \int_{\Omega_i} f \, d\Omega_i| \leq c \, h_i^{n+1} , \quad i \in J .$$

Beweis. Wir untersuchen zuerst den lokalen Diskretisierungsfehler auf den Kanten Γ_{ij}:

$$\sigma_{ij} = |\frac{m_{ij}}{d_{ij}}(u(x_j) - u(x_i)) - \int_{\Gamma_{ij}} \frac{\partial u}{\partial n_{ij}} \, d\Gamma_{ij}| .$$

Für $u \in C^4(\bar{\Omega})$ gilt offensichtlich für den Zentraldifferenzenquotienten

$$|\frac{u(x_j) - u(x_i)}{d_{ij}} - \frac{\partial u}{\partial n_{ij}}(x_{ij})| \leq c \, d_{ij}^2 \leq c \, h_i^2 , \quad i \in J, \, j \in N_i , \tag{5.11}$$

mit einer von hier ab generischen Konstante c, wobei die letzte Ungleichung aus der Definition von h_i und wegen (B1) folgt. Mit der Cauchy-Schwarzschen

Ungleichung folgt

$$
\begin{aligned}
&|\sum_{j\in N_i}\frac{m_{ij}}{d_{ij}}(u(x_j)-u(x_i))-\sum_{j\in N_i}m_{ij}\frac{\partial u}{\partial n_{ij}}(x_{ij})| \\
\le\ &(\sum_{j\in N_i}m_{ij}^2)^{1/2}(\sum_{j\in N_i}|\frac{u(x_j)-u(x_i)}{d_{ij}}-\frac{\partial u}{\partial n_{ij}}(x_{ij})|^2)^{1/2} \\
&\le c h_i^{n-1}\,h_i^2 = c h_i^{n+1}\ , \quad i\in J\ , \qquad (5.12)
\end{aligned}
$$

wobei (B1), (5.10) und (5.11) benutzt wurde. Für $i \in J$ wird nun das stetige lineare Funktional T_i auf $C^3(\bar{\Omega}_i)$ durch

$$
T_i u = \sum_{j\in N_i}(\int_{\Gamma_{ij}}\frac{\partial u}{\partial n_{ij}}d\Gamma_{ij} - m_{ij}\frac{\partial u}{\partial n_{ij}}(x_{ij}))
$$

definiert. Dann erhält man für $T_i u$ die Abschätzung

$$
|T_i u| \le c\,\mu_{n-1}(\Gamma_i)\max_{|\alpha|=1}\max_{x\in\Gamma_i}|[D^\alpha u](x)| \qquad (5.13)
$$

mit einer geigneten Konstante c. Wegen der Voraussetzung (B2) gilt

$$
\int_{\Gamma_{ij}} z\,d\Gamma_{ij} = m_{ij}z(x_{ij}) \quad \text{für alle } z\in\Pi_1\ , \qquad (5.14)
$$

wobei Π_k den Raum der Polynome bis zum Grad k bezeichnet. (5.14) bedeutet, dass man Polynome bis zum Grad 1 exakt mit der Mittelpunkts-Regel integrieren kann. Dann folgt

$$
T_i z = 0
$$

für alle Polynome $z \in \Pi_2$. Aus (5.13) und der Linearität von T_i folgt

$$
\begin{aligned}
|T_i u| &= |T_i(u-z)+T_i z| \le |T_i(u-z)| + |T_i z| \le |T_i(u-z)| \\
&\le c\,\mu_{n-1}(\Gamma_i)\max_{|\alpha|=1}\max_{x\in\Gamma_i}|[D^\alpha(u-z)](x)| \quad \text{für alle } z\in\Pi_2\ . \qquad (5.15)
\end{aligned}
$$

Nun betrachten wir das Taylorpolynom 2. Grades von u im Entwicklungspunkt x_i

$$
z_i(x) = u(x_i) + \nabla u(x_i)\cdot(x-x_i) + \frac{1}{2}(x-x_i)^T H_u(x_i)(x-x_i) \ \text{ für } \ x\in\bar{\Omega}_i\ ,
$$

wobei H_u die Hesse-Matrix von u ist. Aufgrund der vorausgesetzten Glattheit von u gibt es ein c mit

$$
|[D^\alpha(u-z)](x)| \le c|x-x_i|^2 \quad \text{für alle } x\in\bar{\Omega}_i,\ |\alpha|=1
$$

(Abschätzung für Ableitung des Restgliedes der Ordnung $O(|x-x_i|^3)$). Diese Ungleichung und (5.15) ergeben

$$|T_i u| \le c\,\mu_{n-1}(\Gamma_i) h_i^2 \le c h_i^{n+1}\,, \quad i \in J\,, \tag{5.16}$$

mit einer Konstanten $c > 0$. Die Nutzung von (5.12) und (5.16) ergibt

$$\begin{aligned}
&|\sum_{j\in N_i} \frac{m_{ij}}{d_{ij}}(u(x_j)-u(x_i)) + \int_{\Omega_i} f\,d\Omega_i| \\
&\le |\sum_{j\in N_i} \frac{m_{ij}}{d_{ij}}(u(x_j)-u(x_i)) - \sum_{j\in N_i} m_{ij}\frac{\partial u}{\partial n_{ij}}(x_{ij})| \\
&\quad +|\sum_{j\in N_i} m_{ij}\frac{\partial u}{\partial n_{ij}}(x_{ij}) - \int_{\Omega_i} \Delta u\,d\Omega_i| + |\int_{\Omega_i} \Delta u\,d\Omega_i + \int_{\Omega_i} f\,d\Omega_i| \\
&= |\sum_{j\in N_i} \frac{m_{ij}}{d_{ij}}(u(x_j)-u(x_i)) - \sum_{j\in N_i} m_{ij}\frac{\partial u}{\partial n_{ij}}(x_{ij}| \\
&\quad +|\sum_{j\in N_i} m_{ij}\frac{\partial u}{\partial n_{ij}}(x_{ij}) - \int_{\Gamma_i} \frac{\partial u}{\partial n_i} d\Gamma_i| + |\int_{\Omega_i} \Delta u\,d\Omega_i + \int_{\Omega_i} f\,d\Omega_i| \\
&\le c\,h_i^{n+1}\,, \quad i \in J\,.
\end{aligned}$$

□

Die Konvergenz der FV-Methode wird im folgenden Satz bewiesen.

9. Vorlesung am 13.07.2017

Satz 5.5. *(Konvergenz der FVM)*
Seien (B1) und (B2) erfüllt. Die Lösung u von (5.1) liege in $C^4(\bar{\Omega})$. Weiterhin setzen wir für die Voronoi-Boxen voraus, dass eine Konstante $c_0 > 0$ existiert, so dass

$$\mu_n(\Omega_i) \ge c_0 h_i^n \quad \textit{für alle } i \in J \tag{5.17}$$

gilt (Maß der Voronoi-Boxen ist nach unten durch $c_0 h_i^2$ beschränkt). Dann kann der Fehler der FVM in der Maximum-Norm durch

$$||u_h - r_h u||_{\infty,h} \le ch \tag{5.18}$$

mit einer Konstanten $c > 0$ abgeschätzt werden.

Beweis. Sei $w_h := u_h - r_h u$ mit

$$w_i := u_i - u(x_i)\,, \quad i \in \bar{J}\,,$$

Aufgrund der Dirichlet-Randbedingungen gilt

$$w_i := u_i - u(x_i) = 0 \ , \quad i \in \bar{J} \setminus J \ . \tag{5.19}$$

Die Definition der FVM und das Konsistenz-Lemma 5.4 ergeben

$$\begin{aligned} |\sum_{j\in N_i} \frac{m_{ij}}{d_{ij}}(w_j - w_i)| &= |\sum_{j\in N_i} \frac{m_{ij}}{d_{ij}}(u_j - u_i) - \sum_{j\in N_i} \frac{m_{ij}}{d_{ij}}(u(x_j) - u(x_i))| \\ &\le \underbrace{|\sum_{j\in N_i} \frac{m_{ij}}{d_{ij}}(u_j - u_i) + \int_{\Omega_i} f \, d\Omega_i \,|}_{=0, \text{weil } u_i \text{ FVM-Lösung ist}} \\ &\quad + |-\sum_{j\in N_i} \frac{m_{ij}}{d_{ij}}(u(x_j) - u(x_i)) - \int_{\Omega_i} f \, d\Omega_i| \\ &= |\sum_{j\in N_i} \frac{m_{ij}}{d_{ij}}(u(x_j) - u(x_i)) + \int_{\Omega_i} f \, d\Omega_i| \\ &\le ch_i^{n+1} \ . \end{aligned} \tag{5.20}$$

Mit der Vergleichsfunktion

$$v(x) := -\frac{\alpha}{2}|x|^2 + \beta$$

soll nun das diskrete Maximumprinzip 5.2 angewandt werden. Mit $z_i := w_i - v(x_i)$, $i \in \bar{J}$, erhält man mit der Aussage des Lemmas 5.3 und der Abschätzung (5.20)

$$-\sum_{j\in N_i} \frac{m_{ij}}{d_{ij}}(z_j - z_i) \le ch_i^{n+1} - n\alpha \underbrace{\int_{\Omega_i} d\Omega_i}_{=\mu_n(\Omega_i)} \ , \quad i \in J \ .$$

Nun wählt man $\alpha = c^* h$ mit einem genügend großem c^* und

$$\beta := \frac{\alpha}{2} \max_{i\in\bar{J}\setminus J} |x_i|^2 \ .$$

Bei Berücksichtigung der Voraussetzung (5.17) erhält man nun

$$-\sum_{j\in N_i} \frac{m_{ij}}{d_{ij}}(z_j - z_i) \le 0 \ , \quad i \in J \ , \text{ und } z_i \le 0 \ , \ i \in \bar{J} \setminus J \ .$$

Aus dem diskreten Maximumprinzip 5.2 folgt dann $z_i \le 0$ für alle $i \in \bar{J}$. Damit folgt

$$w_i \le v_i \ , \quad i \in \bar{J} \ ,$$

und da nach Konstruktion $0 \leq v \leq ch$ gilt

$$w_i \leq v_i \leq ch \ , \quad i \in \bar{J} \ .$$

Auf die gleiche Weise zeigt man, dass es ein $c > 0$ mit

$$w_i \geq -ch \ , \quad i \in \bar{J} \ .$$

gibt (man wählt $z_i = -w_i - v(x_i)$ und findet mit (5.17) und dem diskreten Maximumprinzip 5.2 $-w_i - v_i \leq 0$ bzw. $w_i \geq -v_i \geq -ch$). Die beiden Ungleichungen liefern

$$|w_i| = |u_i - u(x_i)| \leq ch \Longleftrightarrow ||u_h - r_h u||_{\infty,h} \leq ch \ .$$

□

5.4 Bilanzüberlegungen und Dirichlet-Randbedingungen

Bisher haben wir nur Dirichlet-Randbedingungen behandelt. Dabei haben wir mit Voronoi-Boxen gearbeitet, wobei

$$\mu_n(\Omega) - \sum_{i \in J} \mu_n(\Omega_i) > 0$$

galt, d.h. wir haben de facto einen Bilanzfehler gemacht. Positioniert man die Voronoi-Boxen wie in der Abb. 5.2, dann gilt

$$\bigcup_i \Omega_i = \Omega$$

und bilanziert die Differentialgleichung über die Voronoi-Boxen, dann stimmt die globale Bilanz mit der Summe der lokalen Bilanzen überein.
Es gibt allerdings ein Problem am Rand. Betrachtet man eine Voronoi-Box Ω_i am Rand mit der Randkante $\Gamma_{ij} \subset \Gamma$. Auf Γ_{ij} benötigt man eine Approximation von $\frac{\partial u}{\partial n_{ij}}$. Eine Approximation der Form

$$\frac{\partial u}{\partial n_{ij}}(x_{ij}) \approx \frac{u_j - u_i}{|x_j - x_i|} \tag{5.21}$$

ist nicht direkt möglich, denn x_j ist das Zentrum einer Voronoi-Box, die außerhalb von Ω liegt. Es gibt nun zwei Möglichkeiten:

(i) man verwendet die Approximation

$$\frac{\partial u}{\partial n_{ij}}(x_{ij}) \approx \frac{u_{ij} - u_i}{|x_{ij} - x_i|} \ , \tag{5.22}$$

die allerdings eine geringere Ordnung als (5.21) hat,

(ii) oder man führt mit Ω_j eine "Ghost"-Voronoi-Box ein, verwendet die Fluss-Approximation (5.21), und nutzt die Randbedingung $u_{ij} = r$, um durch eine Interpolation der Form

$$\frac{\alpha}{\alpha+\beta}u_i + \frac{\beta}{\alpha+\beta}u_j = u_{ij} = r \Longleftrightarrow u_j = \frac{\alpha+\beta}{\beta}r - \frac{\alpha}{\beta}u_i \tag{5.23}$$

die benötigte Approximation von u_j zu erhalten. Mit (5.21) und (5.23) hat man nun eine Realisierung der Dirichlet-Randbedingung bei der FVM.

5.5 Neumann-Randbedingungen

Die Behandlung von Neumann-Randbedingungen soll für das Randwertproblem

$$-\Delta u = f \ \text{ in } \Omega, \ u = r \ \text{ auf } \Gamma_d, \ \frac{\partial u}{\partial n} = q \ \text{ auf } \Gamma_n \ , \tag{5.24}$$

diskutiert werden.
Dabei soll $\Omega =]a, b[\times]c, d[$ das in der Abbildung 5.2 dargestellte Gebiet sein. Als Neumann-Rand betrachten wir $\Gamma_n = \{(b, y) | c < y < d\}$. Der Rest des Randes ist der Dirichlet-Rand $\Gamma_d = \partial\Omega \setminus \Gamma_n$.
Positioniert man die Voronoi-Boxen wie in der Abb. 5.2, dann hat man auf den Kanten $\Gamma_{ij} \subset \Gamma_n$ die für die FVM benötigten Flüsse $\frac{\partial u}{\partial n}$ durch die Neumann-Randbedingung mit q gegeben. Am Dirichlet-Rand kann man die erforderlichen Flüsse durch (5.21),(5.23) oder durch (5.22) bereitstellen.
Eine andere Möglichkeit zur Lösung des Randwertproblems (5.24) mit einer FVM kann man mit der in Abb. 5.3 skizzierten Diskretisierung erreichen. Dabei hat man am Neumann-Rand die Flüsse durch q direkt gegeben. Am Dirichlet-Rand werden die Flüsse $\frac{\partial u}{\partial n}$ an randnahen Kanten Γ_{ij} der randnahen Voronoi-Box Ω_i durch

$$\frac{\partial u}{\partial n_{ij}}(x_{ij}) \approx \frac{u_j - u_i}{|x_j - x_i|}$$

wie im obigen Abschnitt besprochen approximiert, wobei u_j durch die Dirichlet-Randbedingung gegeben ist. Am Dirichlet-Rand macht man allerdings einen Bilanzfehler.

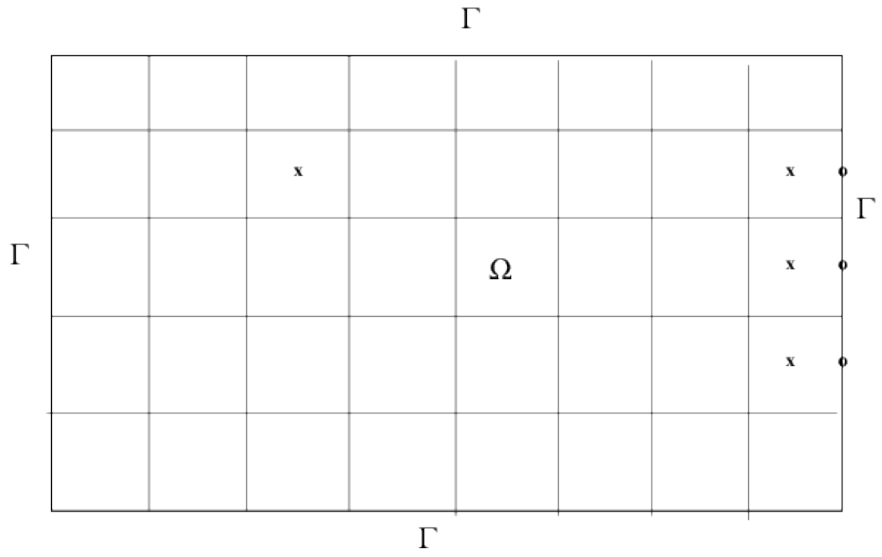

Abbildung 5.2: Diskretisierung mit Voronoi-Boxen (i)

5.6 Diskretisierung von Konvektions-Diffusionsgleichungen

Nachdem bisher hauptsächlich elliptische und parabolische Aufgabenstellungen behandelt wurden, sollen nun die sogenannten Konvektions-Diffusionsgleichungen, also Gleichungen der Form

$$\frac{\partial c}{\partial t} + \nabla \cdot (\vec{v}c) = D\Delta c + f \tag{5.25}$$

behandelt werden. c ist dabei z.B. eine Schadstoffkonzentration, $\vec{v}$ ein Geschwindigkeitsfeld und f ein Quell-Senken-Glied. Bei der Diskretisierung wollen wir den räumlich zweidimensionalen Fall betrachten, d.h. $\vec{v} = (u, v)^T$ und außerdem Inkompressiblität, d.h. $\nabla \cdot \vec{v} = 0$. Damit ergibt sich aus (5.25)

$$\frac{\partial c}{\partial t} + u\frac{\partial c}{\partial x} + v\frac{\partial c}{\partial y} = D\Delta c + f \ . \tag{5.26}$$

Die Diskretisierung soll am Beispiel einer FD-Diskretisierung besprochen werden. Mit c_{ij}^n soll die Gitterfunktion in der Zeitschicht $t = n\Delta t$ am Ortspunkt $(x, y) = (i\Delta x, j\Delta y)$ mit den Diskretisierungsparametern Δt, Δx, Δy einer Richtungs-äquidistanten räumlichen Diskretierung. Eine mögliche konsisten-

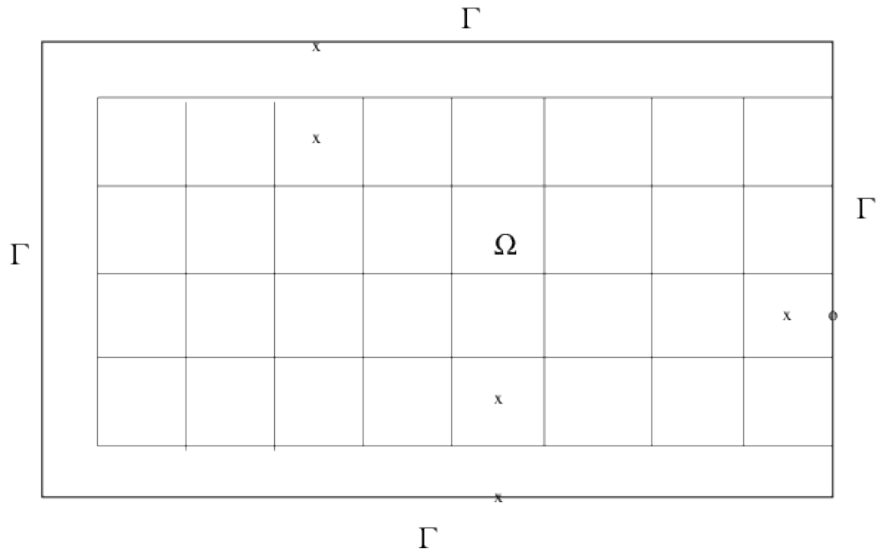

Abbildung 5.3: Diskretisierung mit Voronoi-Boxen (ii)

te Diskretisierung ist die folgende:

$$\begin{aligned}\frac{c_{ij}^{n+1}-c_{ij}^{n}}{\Delta t} &+ \sigma u_{ij}\frac{c_{i+1\,j}^{n+1}-c_{ij}^{n+1}}{\Delta x}+(1-\sigma)\frac{c_{ij}^{n+1}-c_{i-1\,j}^{n+1}}{\Delta x}\\ &+ \gamma v_{ij}\frac{c_{i\,j+1}^{n+1}-c_{ij}^{n+1}}{\Delta y}+(1-\gamma)\frac{c_{ij}^{n+1}-c_{i\,j-1}^{n+1}}{\Delta y}\\ &= D\frac{c_{i+1\,j}^{n+1}-2c_{ij}^{n+1}+c_{i-1\,j}^{n+1}}{\Delta x^2}+D\frac{c_{i\,j+1}^{n+1}-2c_{ij}^{n+1}+c_{i\,j-1}^{n+1}}{\Delta y^2}+f_{ij}^{n}\,.\end{aligned}$$

σ und γ sind dabei Wichtungsparameter aus $[0,1]$. Es entsteht damit pro Zeitschicht ein lineares Gleichungssystem mit einer 5-diagonalen Koeffizientenmatrix A_h. Aus (5.26) ergibt sich für die Hauptdiagonale der Eintrag

$$\frac{1}{\Delta t}-\sigma\frac{v_{ij}}{\Delta x}+(1-\sigma)\frac{v_{ij}}{\Delta x}-\gamma\frac{v_{ij}}{\Delta y}+(1-\gamma)\frac{v_{ij}}{\Delta y}+D\frac{2}{\Delta x^2}+D\frac{2}{\Delta y^2}$$

Für die 4 Nebendiagonalen ergeben sich die Einträge

$$\sigma\frac{u_{ij}}{\Delta x}-D\frac{1}{\Delta x^2}\,,$$

$$-(1-\sigma)\frac{u_{ij}}{\Delta x}-D\frac{1}{\Delta x^2}\,,$$

$$\gamma\frac{v_{ij}}{\Delta y}-D\frac{1}{\Delta x^2}\,,$$

$$-(1-\gamma)\frac{v_{ij}}{\Delta y} - D\frac{1}{\Delta y^2} \; .$$

Sieht man von dem Term $\frac{1}{\Delta t}$ ab (kommt bei stationären Aufgaben nicht vor), dann ergibt die Summe der sämtlichen Einträge pro Matrix-Zeile gerade Null. Eine für iterative Lösungsverfahren geeignete diagonal dominante Matrix erhält man z.B. bei der folgenden Wahl die Wichtungsparameter

$$\sigma = \begin{cases} 1 & \text{falls} \quad u_{ij} > 0 \\ 0 & \text{falls} \quad u_{ij} \leq 0 \end{cases}$$
$$\gamma = \begin{cases} 1 & \text{falls} \quad v_{ij} > 0 \\ 0 & \text{falls} \quad v_{ij} \leq 0 \end{cases}$$

Diese sogenannte upwind-Approximation ergibt natürlich eine nichtsymmetrische Matrix, die allerdings irreduzibel diagonal dominant ist. Es gilt dann

$$|a_{kk}| \geq \sum_{l=1, l\neq k}^{N} |a_{kl}| \quad \text{f.a.} \quad k$$

und

$$|a_{kk}| > \sum_{l=1, l\neq k}^{N} |a_{kl}| \quad \text{für mindestens ein} \quad k \; .$$

Die strikte Ungleichung ergibt sich durch die Berücksichtigung von Randbedingungen.

5.7 FVM für das Stokes-Problem

In der Strömungsmechanik wird bei der numerischen Lösung von Strömungsproblemen die FV-Methode gegenüber FD- oder FE-Methoden deutlich favorisiert. Die Gründe sollen im Folgenden am Beispiel des stationären Stokes-Problems erläutert werden. Das Stokes-Problem lautet

$$\nabla \cdot (\eta \nabla \vec{u}) + \nabla p = \vec{f} \; , \text{ in } \Omega \tag{5.27}$$
$$\nabla \cdot \vec{u} = 0 \; , \text{ in } \Omega \tag{5.28}$$
$$\vec{u} = \vec{r} \; , \text{ auf } \Gamma = \partial\Omega \; , \tag{5.29}$$

wobei $\vec{u} = (u, v)^T$ und p Geschwindigkeits- und Druckfeld bedeuten, $\vec{f} = (f_u, f_v)^T$ ist ein äußeres Kraftfeld und η ist die Viskosität.

Aus Darstellungsgründen verwenden wir ein Rechteckgebiet Ω. Zur Diskretisierung der Gleichung (5.28) verwenden wir rechteckige Voronoi-Boxen wie in der Abb. 5.3. Die Bilanzierung von (5.28) über Ω_{ij} ergibt

$$\int_{\Omega_{ij}} \nabla \cdot \vec{u}\, d\Omega_{ij} = \int_{\Gamma_{ij}} \vec{u} \cdot n\, d\Gamma_{ij} = 0\ . \tag{5.30}$$

Die Voronoi-Boxen versehen wir hier im Unterschied zur früheren Verfahrensweise mit 2 Indizes, ebenso den Rand, um die Position der Voronoi-Box in x- bzw. y-Richtung zu kennzeichnen (also handelt es sich bei Γ_{ij} nicht um eine Kante, sondern um den gesamten Rand von Ω_{ij}). Unterteilt man den Rand Γ_{ij} von Ω_{ij} in Ost/West- bzw. Nord/Süd-Rand, dann erhält man aus (5.30)

$$\int_{\Gamma_{ij}} \vec{u} \cdot n\, d\Gamma_{ij} = \int_{\gamma_o} u\, dy - \int_{\gamma_w} u\, dy + \int_{\gamma_n} v\, dx - \int_{\gamma_s} v\, dx = 0\ . \tag{5.31}$$

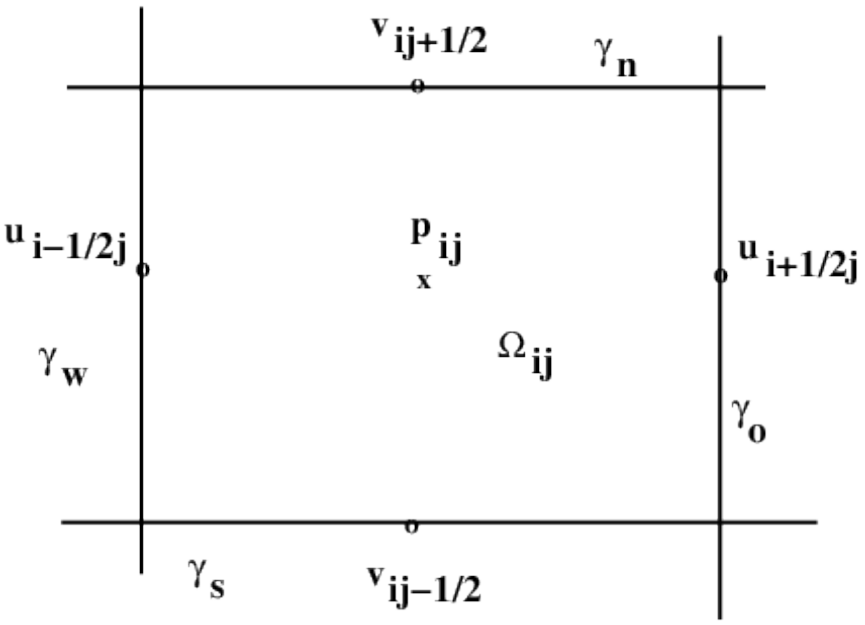

Abbildung 5.4: Voronoi-Box Ω_{ij}

Die Abb. 5.4 zeigt die Voronoi-Box Ω_{ij} und eine kanonische Wahl der Stützstellen für den Druck p_{ij} der Mittelpunkt x_{ij} von Ω_{ij} und die Geschwindigkeitskomponenten $u_{i+1/2\,j}$, $v_{i\,j+1/2}$ an den versetzten Punkten $x_{i+1/2\,j}$,

$x_{i\,j+1/2}$. Verwendet man z.B. die Approximationen

$$\int_{\gamma_o} u\,dy \approx u_{i+1/2\,j}k\ , \quad \int_{\gamma_n} v\,dx \approx v_{i\,j+1/2}h \quad \text{usw.}\ ,$$

wobei k und h Höhe und Breite von Ω_{ij} sind, dann erhält man für (5.28) die FV-Diskretisierung

$$\begin{aligned}[u_{i+1/2\,j} - u_{i-1/2\,j}]\,k + [v_{i\,j+1/2} - v_{i\,j-1/2}]\,h = 0 \\ \iff -\frac{u_{i+1/2\,j} - u_{i-1/2\,j}}{h} - \frac{v_{i\,j+1/2} - v_{i\,j-1/2}}{k} = 0\ . \end{aligned} \tag{5.32}$$

Durch die Wahl der Stützstellen der u- und der v-Komponenten des Geschwindigkeitsfeldes sind die Voronoi-Boxen und damit die Gitter für die Diskretisierung der beiden Komponenten der Impulsbilanz (5.27) vorgegeben, und zwar ergeben sich die Voronoi-Boxen $\Omega_{i+1/2\,j}$ mit den Zentren $x_{i+1/2\,j}$ für die Diskretisierung der u-Gleichung

$$\nabla\cdot(\eta\nabla u) + \frac{\partial p}{\partial x} = f_u \iff \nabla\cdot(\eta\nabla u) + \nabla\cdot\begin{pmatrix} p \\ 0 \end{pmatrix} = f_u \tag{5.33}$$

und die Voronoi-Boxen $\Omega_{i\,j+1/2}$ mit den Zentren $x_{i\,j+1/2}$ für die Diskretisierung der v-Gleichung

$$\nabla\cdot(\eta\nabla v) + \frac{\partial p}{\partial y} = f_v \iff \nabla\cdot(\eta\nabla v) + \nabla\cdot\begin{pmatrix} 0 \\ p \end{pmatrix} = f_v\ . \tag{5.34}$$

Die Gitter für die Diskretisierung der Gleichungen (5.33) und (5.34) sind damit jeweils um ein halbes Inkrement $\frac{h}{2}$ bzw. $\frac{k}{2}$ gegenüber dem Gitter für die Gleichung (5.28) versetzt. Man spricht deshalb auch von der "staggered grid"-Methode. Die Situation ist in der Abb. 5.5 dargestellt.
Die Bilanzierung der Gleichung (5.33) über $\Omega_{i+1/2\,j}$ ergibt nach Anwendung des Gaußschen Integralsatzes

$$\int_{\Gamma_{i+1/2\,j}} \nabla u\cdot n\,d\Gamma + \int_{\Gamma_{i+1/2\,j}} \begin{pmatrix} p \\ 0 \end{pmatrix}\cdot n\,d\Gamma = \int_{\Omega_{i+1/2\,j}} f_u\,d\Omega\ .$$

Mit der Approximation der Flüsse $\nabla u\cdot n$ durch entsprechende Differenzenquotienten erhält man die FV-Approximation

$$\begin{aligned} \eta\frac{u_{i+3/2\,j} - u_{i+1/2\,j}}{h}\,k - \eta\frac{u_{i+1/2\,j} - u_{i-1/2\,j}}{h}\,k \\ +\eta\frac{u_{i+1/2\,j+1} - u_{i+1/2\,j}}{k}\,h - \eta\frac{u_{i+1/2\,j} - u_{i+1/2\,j-1}}{k}\,h \\ +p_{i+1\,j}\,k - p_{ij}\,k = f_{u_{i+1/2\,j}}\,h\,k\ , \end{aligned}$$

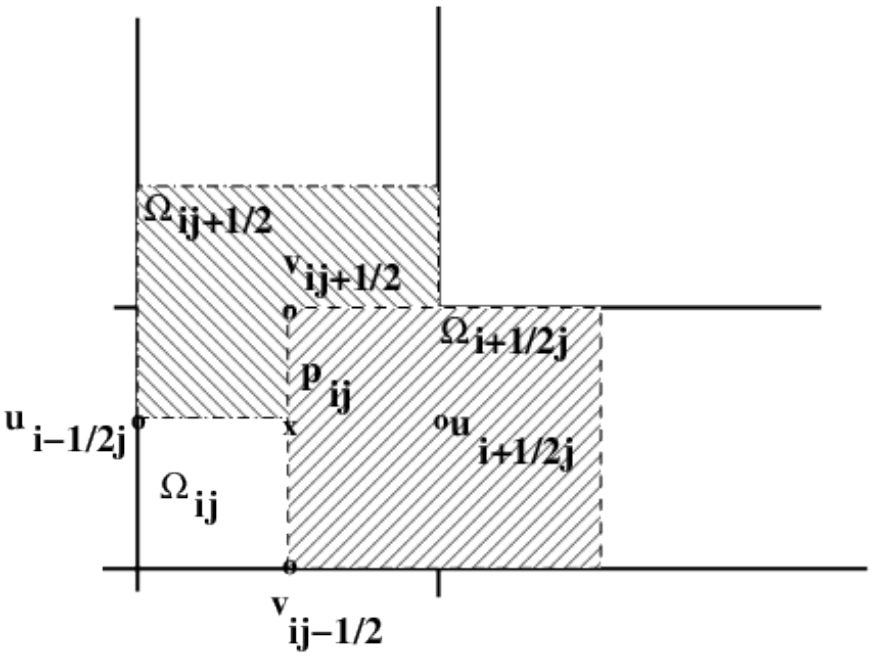

Abbildung 5.5: Voronoi-Boxen $\Omega_{i+1/2\,j}$, $\Omega_{i\,j+1/2}$

bzw. nach Division durch $\mu_2(\Omega_{i+1/2\,j}) = h\,k$

$$\eta\frac{u_{i+3/2\,j} - 2u_{i+1/2\,j} - u_{i-1/2\,j}}{h^2} + \eta\frac{u_{i+1/2\,j+1} - 2u_{i+1/2\,j} - u_{i+1/2\,j-1}}{k^2} + \frac{p_{i+1\,j} - p_{ij}}{h} = f_{u_{i+1/2\,j}}\,. \quad (5.35)$$

Für die FV-Diskretisierung der Gleichung (5.34) erhält man durch die Bilanzierung der Gleichung über $\Omega_{i\,j+1/2}$ analog

$$\eta\frac{v_{i\,j+3/2} - 2v_{i\,j+1/2} - v_{i\,j-1/2}}{k^2} + \eta\frac{v_{i+1\,j+1/2} - 2v_{i\,j+1/2} - v_{i-1\,j+1/2}}{h^2} + \frac{p_{i\,j+1} - p_{ij}}{k} = f_{v_{i\,j+1/2}}\,. \quad (5.36)$$

Schließt man die Gleichungen (5.35), (5.36) und (5.32) durch die Randbedingungen ab, erhält man letztendlich ein Blockgleichungssystem

$$\begin{pmatrix} L_u & \mathbf{0} & G_u \\ \mathbf{0} & L_v & G_v \\ G_u^T & G_v^T & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \\ \mathbf{p} \end{pmatrix} = \begin{pmatrix} \mathbf{r_u} \\ \mathbf{r_v} \\ \mathbf{r_p} \end{pmatrix}\,. \quad (5.37)$$

Hat man Ω durch ein Druckgitter der Dimension $N \times M$, also

$$\Omega = \bigcup_{1 \le i \le N, 1 \le j \le M} \Omega_{ij}$$

diskretisiert, dann sind $\mathbf{u} \in \mathbb{R}^{(N-1)*M}$, $\mathbf{v} \in \mathbb{R}^{N*(M-1)}$ und $\mathbf{p} \in \mathbb{R}^{N*M}$ die Vektoren mit den Näherungswerten $u_{i+1/2\,j}$, $v_{i\,j+1/2}$ und p_{ij}. (5.37) ist ein Sattelpunktproblem. In (5.37) bedeuten L_u, L_v Differenzenoperatoren der viskosen Operatoren der Impulsbilanz, G ist der Differenzenoperator des Gradienten und damit ist G^T der Differenzenoperator der negativen Divergenz als adjungierter Operator des Gradienten.

Bemerkung 5.6. Als Bedingung für die Lösbarkeit des Gleichungssystems (5.27) muss die globale Massenbilanz, also

$$\int_\Gamma \vec{u} \cdot n \, d\Gamma = 0$$

im Diskreten erfüllt sein. Die Koeffizientenmatrix

$$\mathcal{S} \in \mathbb{R}^{P \times P}, \quad P = (N-1) * M + N * (M-1) + N * M \;,$$

von (5.37) hat den Rang $P-1$. Die fehlende Eindeutigkeit liegt an dem Fakt, dass bei dem Stokes-Problem der Druck nur bis auf eine Konstante eindeutig ist (der Druckgradient ist eindeutig). Die Eindeutigkeit durch die Wahl einer Druck-Konstante könnte man z.B. durch die Forderung

$$\sum_{i,j} p_{i\,j} = 0$$

erreichen.

Kapitel 6

Eigenschaften von Matrizen im Ergebnis von FD-Schemen

Die Diskretisierung eines eindimensionalen elliptischen Randwertproblems führt bei einer geeigneten Nummerierung der Unbekannten (Funktionswerte der Gitterfunktionen u_h) auf ein lineares Gleichungssystem mit der Koeffizientenmatrix

$$A = (a_{ij}) = \frac{1}{h^2} \begin{pmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & \ddots & 0 \\ & \ddots & \ddots & \ddots & \\ 0 & \ddots & -1 & 2 & -1 \\ 0 & \dots & 0 & -1 & 2 \end{pmatrix} \tag{6.1}$$

Neben der sparsamen Besetztheit sind die Matrizen von FD-Schemen zur numerischen Lösung von elliptischen Randwertproblemen dadurch gekennzeichnet, dass

$$a_{ij} \leq 0\,, \quad i \neq j\,, \quad a_{ii} > 0\,, \quad \text{und} \quad |a_{ii}| \geq \sum_{j=1, i \neq j}^{n} |a_{ij}|\,, \quad i = 1, \dots, n\,,$$

gilt.

Definition 6.1.
Sei $A = ((a_{ij}) \in \mathbb{R}^{n \times n}$, dann heißt A

1) L_0-Matrix, wenn $a_{ij} \leq 0$, $i \neq j$ gilt,

2) L-Matrix, wenn A eine L_0-Matrix ist und $a_{ii} > 0$ gilt,

3) M-Matrix, wenn A eine L_0-Matrix ist, A^{-1} existiert und $A^{-1} \geq 0$ gilt,

4) stark (strikt) diagonal dominant, wenn

$$|a_{ii}| > \sum_{j=1, i \neq j}^{n} |a_{ij}| \quad \text{für } i = 1, \ldots, n \tag{6.2}$$

gilt, und

5) irreduzibel diagonal dominant, wenn A irreduzibel ist,

$$|a_{ii}| \geq \sum_{j=1, i \neq j}^{n} |a_{ij}| \quad \text{für } i = 1, \ldots, n$$

gilt, und mindestens für einen Index i_0 die strikte Ungleichung (6.2) *gilt.*

Bemerkung 6.2.
Die Relationen $\leq$, $\geq$, $<$ bzw. $>$ für Matrizen oder Vektoren sind so zu verstehen, dass dann jeweils die Relation für alle Elemente der Matrizen bzw. Komponenten der Vektoren gelten.
Eine Matrix $A = (a_{ij})$ heißt irreduzibel, wenn für alle $i, j \in I := \{1, 2, \ldots, n\}$ entweder $a_{ij} \neq 0$ ist, oder eine Indexfolge $i = i_0, i_1, \ldots, i_s = j \in I$ existiert mit $a_{i_{k-1} i_k} \neq 0,\ k = 1, \ldots, s$.

Es sollen nun im Folgenden die wichtigsten Aussagen für die in der Def. 6.1 erklärten Matrizen zusammen gefasst werden.

Satz 6.3. *Die Aussagen*
(a.1) A ist invertierbar und es gilt $A^{-1} \geq 0$,
(a.2) $Ax \leq 0 \Longrightarrow x \leq 0$,
(a.3) $Ax \leq Ay \Longrightarrow x \leq y$
sind äquivalent.
Die Eigenschaften (a.2) oder auch (a.3) nennt man auch Inversmonotonie von A.

Beweis.
Die Implikation (a.3) $\Longrightarrow$ (a.2) erhält man mit $y = 0$, und die Impliklation (a.2) $\Longrightarrow$ (a.3) erhält man durch Anwendung von (a.2) auf $z = x - y$.
Nachweis von (a.2) $\Longrightarrow$ (a.1):
Sei $Ax = 0$, dann ist $A(\pm x) = \pm Ax \leq 0$, woraus $\pm x \leq 0$, also $x = 0$ folgt, was die Injektivität von $l(x) := Ax$ bedeutet. Bijektivität folgt aus der

Endlichdimensionalität des $\mathbb{R}^n$ (bzw. $\mathbb{C}^n$). Damit existiert A^{-1}. (a.2) bedeutet mit $Ax = -y$

$$Ax \leq 0 \Longrightarrow x \leq 0 \quad \text{bzw.} \quad y \geq 0 \Longrightarrow A^{-1}y \geq 0 \; .$$

Setzen $y = (\delta_{ik})_{k \in I}$ für festes $k \in I$, damit folgt für alle $l \in I$

$$0 \leq (A^{-1}y)_l = \sum_{i=1}^{n} (A^{-1})_{li} y_i = (A^{-1})_{lk} \; .$$

Nachweis von (a.1) $\Longrightarrow$ (a.2):
Für $y \geq 0$ gilt

$$(A^{-1}y)_i = \sum_{j=1}^{n} (A^{-1})_{ij} y_j \geq 0 \; ,$$

also $A^{-1}y \geq 0$. Sei jetzt $Ax \leq 0$, dann ist $y = -Ax \geq 0$ und schließlich

$$A^{-1}y = -x \geq 0 \quad \text{bzw.} \quad x \leq 0 \; ,$$

und damit ist der Satz vollständig bewiesen. □

Aus der linearen Algebra sei an das Kriterium von Gerschgorin erinnert:

Satz 6.4. *(Gerschgorin-Kriterium)*
Sei $K_r(z) = \{\xi \in \mathbb{C} \; , \; |\xi - z| < r\}$. Dann gilt
(a.4) Alle Eigenwerte von A liegen in den Gerschgorin-Kreisen

$$\bigcup_{i=1}^{n} \bar{K}_{r_i}(a_{ii}) \quad \textit{mit} \;\; r_i = \sum_{j=1, j \neq i}^{n} |a_{ij}| \; .$$

(a.5) Ist A irreduzibel, dann liegen die Eigenwerte sogar in

$$[\bigcup_{i=1}^{n} K_{r_i}(a_{ii})] \cup [\bigcap_{i=1}^{n} \partial K_{r_i}(a_{ii})] \; .$$

Beweis. Sei λ ein EW von A mit dem EV x und (o.B.d.A.) $||x||_\infty = 1$ und

für $j \in I = \{1, 2, \ldots, n\}$ gelte $|x_j| = 1$. Aus $|x_j| = 1$ folgt

$$\begin{aligned}
\lambda x_j &= (Ax)_j = \sum_{k=1}^{n} a_{jk} x_k = a_{jj} + \sum_{k=1, k\neq j}^{n} a_{jk} x_k \\
\Longrightarrow \qquad & (\lambda - a_{jj}) x_j = \sum_{k=1, k\neq j}^{n} a_{jk} x_k \\
\Longrightarrow \qquad & |\lambda - a_{jj}|\, |x_j| \leq \sum_{k=1, k\neq j}^{n} |a_{jk}| \underbrace{|x_k|}_{\leq 1} \leq \sum_{k=1, k\neq j}^{n} |a_{jk}| \qquad (6.3) \\
\Longrightarrow \qquad & |\lambda - a_{jj}| \leq \sum_{k=1, k\neq j}^{n} |a_{jk}| = r_j \qquad (6.4)
\end{aligned}$$

Damit folgt $\lambda \in \bigcup_{i=1}^{n} \bar{K}_{r_i}(a_{ii})$, also (a.4).
Zum Nachweis von (a.5) zeigen wir für Eigenwerte λ, die nicht in $\bigcup_{i=1}^{n} K_{r_i}(a_{ii})$ liegen zuerst, dass im Falle $a_{ji} \neq 0$ gilt:

$$\text{aus } |x_j| = 1 \text{ und } |\lambda - a_{jj}| = r_j \implies |x_i| = 1 \text{ und } |\lambda - a_{ii}| = r_i \,. \qquad (6.5)$$

Nehmen wir an, dass $|x_i| < 1$ ist:
Es gilt nun

$$\begin{aligned}
r_j &= |\lambda - a_{jj}| \\
&= |(\lambda - a_{jj}) x_j| = |\sum_{k=1, k\neq j}^{n} a_{jk} x_k| \\
&\leq \sum_{k=1, k\neq j, i}^{n} |a_{jk}| \underbrace{|x_k|}_{\leq 1} + |a_{ji}| \underbrace{|x_i|}_{<1} \\
&< \sum_{k=1, k\neq j}^{n} |a_{jk}| = r_j, \;,
\end{aligned}$$

das ist ein Widerspruch, also war unsere Annahme falsch, und es gilt $|x_i| = 1$.
Aufgrund der Irreduzibilität folgt aus (6.5), dass

$$\lambda \in \bigcap_{k=1}^{n} \partial K_{r_k}(a_{kk})$$

gilt (man fängt in der j-ten Zeile von A (mit $||x||_\infty = |x_j|$) an, und kann sich wegen der Irreduzibilität durch die Matrix hangeln), so dass damit letztendlich für alle Indizes $k = 1, \ldots, n$ die Gültigkeit von $|\lambda - a_{kk}| = r_k$, also (a.5) folgt. □

Bemerkung 6.5. Eine Folgerung aus dem Satz ist, dass Eigenwerte von irreduziblen Matrizen, die nicht in der Vereinigung der offenen Gerschgorin-Kreise liegen, in der Schnittmenge aller Ränder der Gerschgorin-Kreise liegen.

Nun kann man den wichtigen Satz zur Regularität von irreduziblen diagonal dominanten Matrizen zeigen.

Satz 6.6. *Sei A eine L-Matrix und irreduzibel diagonal dominant. Dann gilt mit der Aufspaltung $A = D - B$ und D gleich dem Diagonalanteil von A sowie B dem negativen Außendiagonalanteil von A*

$$\rho(D^{-1}B) < 1 \, .$$

Beweis. Sei $C = D^{-1}B = (c)_{ij}$, dann gilt

$$c_{ij} = -\frac{a_{ij}}{a_{ii}} \quad \text{für } \; i \neq j, \quad \text{und} \quad c_{ii} = 0 \, .$$

Es gilt nun für

$$r_\alpha = \sum_{\beta=1, \beta \neq \alpha}^{n} |c_{\alpha\beta}| < 1$$

wegen der irreduziblen Diagonaldominanz für mindestens einen Index α. Für alle $\beta \in I$ gilt $r_\beta \leq 1$. Nach Gerschgorin liegen die EW von C in

$$\underbrace{[\bigcup_{i=1}^{n} K_{r_i}(0)]}_{\subset K_1(0)} \quad \cup \quad [\bigcap_{i=1}^{n} \partial K_{r_i}(0)] \, .$$

Es bleibt zu zeigen:

$$\bigcap_{i=1}^{n} \partial K_{r_i}(0) \subset K_1(0) \, .$$

Fall a):

$$r_\beta = r \quad \text{für alle } \beta \in I \, .$$

Wegen $r_\alpha = r < 1$ folgt

$$\bigcap_{\beta=1}^{n} \partial K_{r_\beta}(0) = \partial K_r(0) \subset K_1(0) \, ,$$

also sind alle EW von C betragsmäßig kleiner als 1, damit gilt $\rho(C) < 1$.
Fall b):
Die r_β sind nicht alle gleich. Dann gilt

$$\bigcap_{\beta=1}^{n} \partial K_{r_\beta}(0) = \emptyset \,,$$

d.h. nach Gerschgorin liegen die EW von C in $K_1(0)$, also gilt $|\lambda| < 1$ und damit auch $\rho(C) < 1$. □

Bemerkung 6.7.
Die soeben bewiesene Aussage für irreduzibel diagonal dominante Matrizen gilt auch für strikt diagonal dominante Matrizen (die nicht irreduzibel sein müssen).

Satz 6.8.
Sei $A = D - B$ eine L-Matrix, dann gilt mit D und C aus dem obigen Satz

$$A \;\; ist \;\; M\text{-}Matrix \Longleftrightarrow \rho(D^{-1}B) < 1 \,.$$

Beweis.
Richtung $\Longrightarrow$:
Sei A eine M-Matrix. Sei λ ein EW von $D^{-1}B$ mit dem EV $u \neq 0$. Dann gilt

$$|\lambda||u| = |\lambda u| = |D^{-1}Bu| \leq D^{-1}B|u| \,,$$

wegen $A^{-1}D \geq 0$ folgt

$$\begin{aligned} -A^{-1}DD^{-1}B|u| &\leq -A^{-1}D|\lambda||u| \quad \text{und} \\ |u| &= A^{-1}A|u| = A^{-1}(D-B)|u| \\ &= A^{-1}D(E - D^{-1}B)|u| \\ &\leq A^{-1}D|u| - A^{-1}D|\lambda||u| \\ &= (1-|\lambda|)\underbrace{A^{-1}D}_{\geq 0}|u| \,. \end{aligned}$$

Für $|\lambda| \geq 1$ folgt $|u| \leq 0$, d.h. $u = 0$, also muss $|\lambda| < 1$ für alle EW gelten, und damit gilt $\rho(D^{-1}B) < 1$.
Richtung $\Longleftarrow$:
Sei $\rho(D^{-1}B) < 1$. Damit konvergiert die Neumann-Reihe mit $C = D^{-1}B$ und es gilt

$$S = \sum_{\nu=0}^{\infty} C^\nu = (E - C)^{-1} \,.$$

Wegen $D^{-1} \geq 0$ und $B \geq 0$ folgt

$$C \geq 0\,, \quad C^{\nu} \geq 0\,, \quad S \geq 0\,.$$

Aus

$$E = S(E - C) = SD^{-1}(D - B) = SD^{-1}A$$

folgt

$$A^{-1} = SD^{-1} \implies A^{-1} \geq 0 \implies A \text{ ist } M\text{-Matrix}\,.$$

□

Aus den Sätzen 6.6, 6.8 folgt unmittelbar die wichtige Aussage

Satz 6.9.
Sei A eine L-Matrix. Ist A strikt diagonal dominant oder irreduzibel diagonal dominant, dann ist A eine M-Matrix.

Zum Schluss dieses Abschnittes soll noch das sogenannte M-Kriterium notiert werden.

Satz 6.10. *(M-Kriterium)*
Sei A eine M-Matrix und $e > 0$ ein Vektor mit $Ae > 0$. Dann gilt

$$||A^{-1}y||_{\infty} \leq C_e||y||_{\infty} \quad \textit{mit } C_e := \frac{\max_i e_i}{\min_i (Ae)_i}\,,$$

d.h.

$$||A^{-1}||_{\infty} \leq C_e\,.$$

Beweis.
Sei $x = A^{-1}y$, also $Ax = y$. Dann ist

$$\pm x_i = \sum_{j=1}^{n} (A^{-1})_{ij}(\pm y_j) \leq \sum_{j=1}^{n} (A^{-1})_{ij}||y||_{\infty}\,.$$

Sei $c = \min_i (Ae)_i$, d.h. $Ae \geq c(1, 1, \ldots, 1)^T$. Da A invers monoton ist (aus $Ax \leq 0$ folgt $x \leq 0$), gilt

$$e \geq cA^{-1}(1, 1, \ldots, 1)^T\,, \quad \text{d.h.} \quad e_i \geq c\sum_{j=1}^{n} (A^{-1})_{ij}\,.$$

Daraus folgt nun $\pm x_i \leq \frac{e_i}{c}||y||_{\infty}$ und damit

$$||x||_{\infty} \leq \frac{||e||_{\infty}}{c}||y||_{\infty} = C_e||y||_{\infty}\,,$$

also die Aussage des Satzes. □

Bemerkung 6.11.
Die Matrizen vom Typ (6.1) sind L-Matrizen und irreduzibel diagonal dominant, d.h. sie sind auch M-Matrizen. Damit sind sie invertierbar (regulär) und man kann in der Regel eine Schranke für die Maximumnorm der inversen Matrix A^{-1} angeben, was Stabilität bedeutet.
Allerdings muss man dazu einen Vektor e mit der geforderten Eigenschaft finden. Bei strikt diagonal dominanten L-Matrizen findet man mit

$$e = (1, 1, \ldots, 1)^T$$

diesen Vektor sofort.
Beim Beispiel des Poissonschen Randwertptroblems auf einem n-dimensionalen Einheitswürfel könnte man die Funktionswerte von

$$v(x) := x_1(1 - x_1) \quad \text{für} \quad x = (x_1, x_2, \ldots, x_n) \in \Omega_h$$

als Komponenten des Vektors $e > 0$ wählen, denn es gilt

$$-L_h v(x) = 2 \quad \text{bzw.} \quad Ae = (2, 2, \ldots, 2)^T > 0 \, .$$

Mit $||e||_\infty = \frac{1}{4}$ erhält man dann

$$||A^{-1}||_\infty \leq \frac{1}{8} \, .$$