Vorlesung *Modellierung und Simulation I*

Prof. Dr. Gillian Queisser

22. April 2015

Inhaltsverzeichnis

1 Einleitung

Die Einleitung in diesem Skript ist als Wiederholung aller Inhalte aus der Vektoranalysis, die für das Verständnis der Inhalte dieser Vorlesung wichtig sind, gedacht. Es werden spezielle lineare Abbildungen und die zugehörige Eigenwert- und Eigenvekortheorie behandelt sowie Anwendungen dieser auf die Berechnung von Matrixprodukten und die Lösung gewöhnlicher Differentialgleichungen. Der Abschluss dieses Kapitel erlaubt einen Übergang auf das erste Hauptkapitel über gewöhnliche Differentialgleichungen.

1.1 Lineare Abbildungen

Definition 1. *Lineare Abbildungen können in Matrixschreibweise definiert werden als*

$$\begin{array}{rcll} A : \mathbb{R}^n & \longrightarrow & \mathbb{R}^n & n \in \mathbb{N}^+ \\ x & \longmapsto & Ax & \\ \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} & \longmapsto & A \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} & \end{array}$$

mit $A := \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}$ *definiert eine lineare Abbildung.*

Beispiel 1.

$$A = \begin{pmatrix} 2 & 5 \\ 1 & 9 \end{pmatrix}$$

Dann ist

$$A \cdot x = \begin{pmatrix} 2 & 5 \\ 1 & 9 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2x_1 + 5x_2 \\ 1x_1 + 9x_2 \end{pmatrix}.$$

Frage: Hat die Gleichung

$$Ax = b$$

eine Lösung? Löse dazu das lineare Gleichungssystem

$$\begin{array}{rcl} 2x_1 + 5x_2 & = & b_1 \\ x_1 + 9x_2 & = & b_2. \end{array}$$

1.1.1 LU-Zerlegung

Ein allgemeiner Ansatz zur Lösung linearer Gleichungssysteme ist die Zerlegung der Matrix A in eine linke und rechte Dreiecksmatrix, die sogenannte LU-Zerlegung:

$$A = L \cdot U$$

Durch die Berechnung von Elementarmatrizen E_{ij}, welche den Eintrag a_{ij} in der Matrix A zu Null machen, kann ein algorithmisches Vorgehen der LU-Zerlegung hergeleitet werden.

Beispiel 2. *Betrachte*

$$A = \begin{pmatrix} 2 & 1 \\ 8 & 7 \end{pmatrix}.$$

Mit $E_{21} := \begin{pmatrix} 1 & 0 \\ -4 & 1 \end{pmatrix}$ *wird der Eintrag* a_{21} *zu Null:*

$$\underbrace{\begin{pmatrix} 1 & 0 \\ -4 & 1 \end{pmatrix}}_{E_{21}} \cdot \underbrace{\begin{pmatrix} 2 & 1 \\ 8 & 7 \end{pmatrix}}_{A} = \underbrace{\begin{pmatrix} 2 & 1 \\ 0 & 3 \end{pmatrix}}_{U}$$

Daraus folgt $A = L \cdot U$ *mit* $L = E_{21}^{-1} = \begin{pmatrix} 1 & 0 \\ 4 & 1 \end{pmatrix}$ *und ferner*

$$\begin{aligned} Ax &= LUx = E_{21}^{-1}Ux = b \\ \Rightarrow x &= U^{-1}E_{21}b. \end{aligned}$$

Beispiel 3. *Sei* $E_{32}E_{31}E_{21}A = U$ *mit*

$$E_{31} = Id, \quad E_{32} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -5 & 1 \end{pmatrix}, \quad E_{21} = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Es existieren zwei Möglichkeiten: Entweder das Produkt $\prod_{i,j} E_{i,j}$ *berechnen und dann invertieren oder erst* E_{ij}^{-1} *berechnen und dann das Produkt* $\prod_{i,j} E_{i,j}^{-1}$ *bilden.*
Fall 1:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -5 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 10 & -5 & 1 \end{pmatrix}$$

Hier müssen also, um L *zu erhalten, etliche Matrixprodukte sowie die Inverse einer Dreiecksmatrix berechnet werden.*
Fall 2:

$$\underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}}_{E_{21}^{-1}} \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 5 & 1 \end{pmatrix}}_{E_{32}^{-1}} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 0 & 5 & 1 \end{pmatrix}$$

Hier ergibt sich L auf natürliche Weise (ohne echte Rechnung) durch Übernahme der linken unteren Einträge aus den Elementarmatrix-Inversen.

Ein wichtiger Aspekt bei der numerischen Lösung von Matrixprodukten (zur Invertierung von Matrizen) ist der Aufwand.

1.1.2 Aufwandsberechung Matrixmultiplikation

Wie viele Operationen (definiert als eine Multiplikation und eine Addition) sind für die LU-Zerlegung einer $n \times n$-Matrix notwendig?

Beispiel 4. *Sei $n = 100$ und die Matrix A vollbesetzt. Dann werden für die Elimination der ersten Spalte etwa 100^2 Operationen benötigt, für die zweite Spalte nur mehr 99^2, für die dritte 98^2 usw. D.h. der Aufwand w beträgt*

$$w \approx 100^2 + 99^2 + 98^2 + \ldots + 1^2.$$

Im allgemeinen Fall gilt:

$$w \approx n^2 + (n-1)^2 + (n-2)^2 + \ldots + 1^2.$$

Die Größenordung des Aufwands w lässt sich berechnen durch

$$w \approx \sum_{x=0}^{n} x^2 \approx \int_0^n x^2 dx = \frac{1}{3}x^3|_0^n = \frac{1}{3}n^3.$$

Der Aufwand einer LU-Zerlegung beträgt also etwa $\mathcal{O}\left(\frac{1}{3}n^3\right)$.

Zusammenfassung:

- Systeme wie $Au = 0$ oder $Au = b$ können mit Hilfe der LU-Zerlegung berechnet werden.
- Nachteil: großer Aufwand!

Um verbesserte (schnellere) Verfahren für die Lösung linearer Gleichungssysteme zu entwickeln, benötigen wir genauere Informationen über die Matrix und ihre Eigenschaften.

1.2 Eigenwerte und Eigenvektoren

Eigenwerte und Eigenvektoren liefern wertvolle Informationen über eine Matrix.

Definition 2. *Wird ein Vektor $\vec{x} \neq 0$ von einer Matrix A auf $A\vec{x}$ abgebildet, sodass gilt*

$$A\vec{x} = \lambda\vec{x}, \tag{1.1}$$

so wird λ ein Eigenwert *und $\vec{x}$ ein zugehöriger* Eigenvektor *von A genannt.*

Wie finden wir nun Eigenwerte und Eigenvektoren einer Matrix?

Beispiel 5. *Was sind die Eigenwerte und -vektoren für eine Projektion* $P\colon \mathbb{R}^3 \to \mathbb{R}^3$*, die auf eine Ebene projiziert?*
Eigenvektoren (zum Eigenwert 1) sind jedenfalls alle $\vec{x}$ *in der Ebene* E*, auf die* P *projiziert. Für* P *erwarten wir also 2 linear unabhängige Eigenvektoren in* E*:*

$$\begin{aligned} Px_1 &= 1 \cdot x_1 \\ Px_2 &= 1 \cdot x_2 \end{aligned}$$

mit $x_1 \nparallel x_2$*. Ein weiterer Eigenvektor* $\vec{x}_3$ *(zum Eigenwert 0) steht senkrecht zu* E*, er kommt aus dem (eindimensionalen) Unterraum, der auf Null abgebildet wird d.h.*

$$Px_3 = 0 \cdot x.$$

Beispiel 6. *Man betrachte die Permutations-Matrix*

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

Die zugehörigen Eigenvektoren lauten:

$$\begin{aligned} x_1 &= \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \\ x_2 &= \begin{pmatrix} -1 \\ 1 \end{pmatrix} \end{aligned}$$

mit den Eigenwerten $\lambda_1 = 1$ *und* $\lambda_2 = -1$*.*

Nun stellt sich die Frage, wie die Gleichung

$$Ax = \lambda x \tag{1.2}$$

gelöst werden kann, in der λ sowie x Unbekannte sind. Wir führen dazu folgende Umformung durch

$$Ax = \lambda x \quad \Leftrightarrow \quad (A - \lambda \cdot Id)\, x = 0$$

Gilt obige Gleichung für $x \neq 0$, so muss $A - \lambda \cdot Id$ singulär sein. Dies ist äquivalent zu

$$\det(A - \lambda Id) = 0.$$

Definition 3. $\det(A - \lambda Id)$ *heißt* charakteristische Funktion.

Die Berechnung von Gleichung (1.2) kann also in zwei Schritten erfolgen: (1) Berechne λ als Nullstellen der charakteristischen Funktion und (2) berechne x.

Beispiel 7. *Sei* $A = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$. *Dann folgt*

$$\begin{aligned} \det(A - \lambda Id) &= \left| \begin{pmatrix} 3-\lambda & 1 \\ 1 & 3-\lambda \end{pmatrix} \right| = (3-\lambda)^2 - 1 = \\ &= \lambda^2 - 6\lambda + 8 \end{aligned}$$

Wir erhalten für A *die Eigenwerte* $\lambda_1 = 4$ *und* $\lambda_2 = 2$.

$$\Rightarrow A - 4 \cdot Id = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}$$

Wir lösen nun das System

$$\begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} \cdot x = 0$$

Daraus berechnet sich ein zu $\lambda_1 = 4$ *gehöriger Eigenvektor* $x_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ *und analog zu* $\lambda_2 = 2$ *ein Eigenvektor* $x_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$.

Bemerkung 1. *Natürlich sind die berechneten Eigenvektoren nicht die einzigen. Tatsächlich ist die Menge aller Eigenvektoren zu einem Eigenwert immer ein ganzer linearer Raum, denn für einen Eigenvektor* x *mit* $Ax = \lambda x$ *gilt immer auch* $A(\alpha x) = \lambda(\alpha x)$ *für jeden Skalar* α, *also sind (mindestens) alle* $y \neq 0$ *aus dem linearen Raum* $\langle x \rangle := \{\alpha x\}$ *Eigenvektor zu* λ. *Man spricht von einem Eigen*raum.

Bemerkung 2. *Die charakteristische Funktion zu einer* $n \times n$*-Matrix* A *mit Einträgen in* $\mathbb{C}$ *(man schreibt:* $A \in \mathbb{C}^{n \times n}$*) ist ein Polynom vom Grad* n. *Dieses hat über* $\mathbb{C}$ *genau* n *Nullstellen (mit Vielfachheiten), also existieren (mit Vielfachheiten gezählt) zu jeder solchen Matrix immer genau* n *Eigenwerte in* $\mathbb{C}$ *(nicht notwendig in* $\mathbb{R}$, *auch wenn* $A \in \mathbb{R}^{n \times n}$*!).*

Folgende Aussagen gelten für die Eigenwerte:

Satz 1. *Für die Eigenwerte* $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$ *einer Matrix* $A \in \mathbb{C}^{n \times n}$ *gilt*

$$\sum_{i=1}^{n} \lambda_i = \text{spur}(A) \tag{1.3}$$

und für das Produkt

$$\prod_{i=1}^{n} \lambda_i = \det(A). \tag{1.4}$$

1.2.1 Spezielle Eigenwerte und Eigenvektoren

Nicht immer besitzt eine reelle Matrix reelle Eigenwerte, wie dieses Beispiel zeigt:

Beispiel 8. *Man betrachte die Rotationsmatrix Q_α mit*

$$Q_\alpha = \begin{pmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{pmatrix}$$

Für eine Rotation um $90^o = \frac{\pi}{2}$ folgt

$$Q_{\frac{\pi}{2}} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

Was sind die Eigenwerte und Eigenvektoren von $Q_{\frac{\pi}{2}}$? Wir wissen:

- $\lambda_1 + \lambda_2 = \text{spur}\left(Q_{\frac{\pi}{2}}\right) = 0,$
- $\lambda_1 \cdot \lambda_2 = \det\left(Q_{\frac{\pi}{2}}\right) = 1.$

Dies sieht nach einem Widerspruch (in $\mathbb{R}$) aus. Wir untersuchen weiter:

$$\begin{aligned} \det\left(Q_{\frac{\pi}{2}} - \lambda Id\right) = \left|\begin{pmatrix} -\lambda & -1 \\ 1 & -\lambda \end{pmatrix}\right| &= \lambda^2 + 1 = 0 \\ &\Rightarrow \quad \lambda_{1/2} = \pm i. \end{aligned}$$

Reelle Matrizen können also unter Umständen komplexe Eigenwerte besitzen. Wir werden jedoch sehen, dass *symmetrische* Matrizen über reelle Eigenwerte verfügen – eine Eigenschaft, die wir häufig nutzen werden.
Ein weiterer Sonderfall ist die Tatsache, dass eine $n \times n$-Matrix nicht immer über n linear unabhängige Eigenvektoren verfügt, wie das folgende Beispiel zeigt:

Beispiel 9. *Man betrachte die Matrix $A = \begin{pmatrix} 3 & 1 \\ 0 & 3 \end{pmatrix}$. Was sind die zugehörigen Eigenwerte? Da A eine obere Dreiecksmatrix ist, können die Eigenwerte direkt abgelesen werden, $\lambda_{1/2} = 3$. Die zugehörigen Eigenvektoren lassen sich berechnen durch:*

$$(A - \lambda Id)x = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

A verfügt also nur über einen Eigenraum $x_1 = \left\langle \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right\rangle$.

Der Fall, dass eine $n \times n$-Matrix A über weniger als n linear unabhängige Eigenvektoren verfügt, muss gesondert behandelt werden. Wir widmen uns dem „gutartigen“ Fall, dass A über n Eigenvektoren verfügt.

1.2.2 Diagonalisierung von Matrizen

Die $n \times n$-Matrix A besitze n linear unabhängige Eigenvektoren $x_1, \ldots, x_n$. Wir nennen die Matrix S mit $x_1, \ldots, x_n$ als Spaltenvektoren die Eigenvektormatrix

$$S := \begin{pmatrix} x_1^1 & \cdots & x_n^1 \\ x_1^2 & \cdots & x_n^2 \\ \vdots & \ddots & \vdots \\ x_1^n & \cdots & x_n^n \end{pmatrix}. \tag{1.5}$$

Für das Produkt $A \cdot S$ gilt:

$$\begin{aligned} A \cdot S &= A \cdot \begin{pmatrix} x_1^1 & \cdots & x_n^1 \\ x_1^2 & \cdots & x_n^2 \\ \vdots & \ddots & \vdots \\ x_1^n & \cdots & x_n^n \end{pmatrix} = \begin{pmatrix} \lambda_1 x_1^1 & \cdots & \lambda_n x_n^1 \\ \lambda_1 x_1^2 & \cdots & \lambda_n x_n^2 \\ \vdots & \ddots & \vdots \\ \lambda_1 x_1^n & \cdots & \lambda_n x_n^n \end{pmatrix} \\ &= \begin{pmatrix} x_1^1 & \cdots & x_n^1 \\ x_1^2 & \cdots & x_n^2 \\ \vdots & \ddots & \vdots \\ x_1^n & \cdots & x_n^n \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \lambda_n \end{pmatrix} \end{aligned}$$

$\Rightarrow A \cdot S = S \cdot \Lambda$ mit

$$\Lambda := \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \lambda_n \end{pmatrix}$$

Λ bezeichnen wir als Eigenwert-Matrix. Durch Multiplikation beider Seiten mit S^{-1} folgt:

$$\begin{aligned} AS &= S\Lambda \\ \Leftrightarrow S^{-1}AS &= \Lambda \\ \Leftrightarrow A &= S\Lambda S^{-1} \end{aligned}$$

S^{-1} existiert nach der Voraussetzung der linearen Unabhängigkeit an die Eigenvektoren von A.

Bemerkung 3. *Matrix A lässt sich mit Hilfe der Eigenvektoren und Eigenwerte* diagonalisieren.

Beispiel 10. *Man betrachte $Ax = \lambda x$. Was gilt nun für A^2x?*

$$A^2x = A\lambda x = \lambda Ax = \lambda^2 x$$

$\Rightarrow$ *Die Eigenwerte von A^2 sind die quadrierten Eigenwerte von A.*

Damit können wir Potenzen von A leicht berechnen.

Bemerkung 4. *Mit $A = S\Lambda S^{-1}$ gilt:*

$$A^k = (S\Lambda S^{-1}) \cdots (S\Lambda S^{-1}) = S\Lambda^k S^{-1}.$$

Beispiel 11. *Gegeben sei die Matrix A. Man berechne A^{100}. Man kann hier 99 Matrix-Multiplikationen ausführen. Wesentlich effizienter ist der Ansatz über eine Eigenvektor-Hauptachsentransformation:*

$$A^{100} = (S\Lambda S^{-1})(S\Lambda S^{-1}) \cdots (S\Lambda S^{-1}) = S\Lambda^{100} S^{-1}.$$

Satz 2. *Es gilt*

$$A^k \to 0 \quad \text{für } k \to \infty,$$

falls $\|\lambda_i\| < 1 \quad \forall i = 1, \ldots, n$.

1.2.3 Diagonalisierbarkeit

Wie oben gesehen, sind Matrizen mit n linear unabhängigen Eigenvektoren durch eine Hauptachsentransformation diagonalisierbar. Falls n einfache Eigenwerte existieren (d.h. keine Eigenwerte mit Vielfachheiten), so existieren n unabhängige Eigenvektoren. Zwei Beispiele zeigen, dass bei Eigenwerten mit Vielfachheiten nicht eindeutig auf die Anzahl linear unabhängiger Eigenvektoren geschlossen werden kann.

Beispiel 12. *Sei A die Identität, also* $A = \begin{pmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{pmatrix}$. *Dann gilt:*

$$\lambda_1 = \lambda_2 = \ldots = \lambda_n = 1.$$

Trotz einem Eigenwert mit Vielfachheit n existieren n linear unabhängige Eigenvektoren

$$x_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \; x_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \; \ldots, \; x_n = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}.$$

Beispiel 13. *Sei nun* $A = \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}$. *Die Eigenwerte sind $\lambda_1 = \lambda_2 = 2$ (Vielfachheit 2). Die Eigenvektoren berechnen sich aus:*

$$(A - 2Id)\, x = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} x = 0.$$

Der Kern von $A - 2Id$ ist eindimensional mit Erzeuger $x_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, *es gibt also keine zwei linear unabhängigen Eigenvektoren.*

An diesen Beispielen sehen wir, dass bei Eigenwerten mit Vielfachheiten geprüft werden muss, ob die Matrix über Eigenvektoren diagonalisierbar ist. Andererseits wissen wir, dass Matrizen mit n unterschiedlichen Eigenwerten über n linear unabhängige Eigenvektoren verfügen. Diesen gutartigen Matrizen widmen wir uns nun im Kontext der evolutionären Prozesse.

1.3 Evolutionäre Prozesse

Dieser Typ von Problem wird uns später im Rahmen der gewöhnlichen Differentialgleichungen begegnen.

1.3.1 Eigenwerte und Eigenvektoren in evolutionären Prozessen

Sei $A \in \mathbb{R}^{n\times n}$. Man betrachte folgende Iterationsvorschrift:

$$u_{k+1} = Au_k, \tag{1.6}$$

wobei der Startvektor $u_0 \in \mathbb{R}^n$ gegeben sei. Zu berechnen sind nun

$$\begin{aligned} u_1 &= Au_0, \\ u_2 &= A^2u_0, \\ &\vdots \\ u_k &= A^ku_0. \end{aligned}$$

Wir können dazu folgenden Ansatz verfolgen: Man schreibe u_0 als Linearkombination der Eigenvektoren $x_1, x_2, \ldots, x_n$ von A.

$$\Rightarrow u_0 = c_1x_1 + c_2x_2 + \cdots + c_nx_n = Sc. \tag{1.7}$$

Für Au_0 gilt:

$$\begin{aligned} Au_0 &= Ac_1x_1 + Ac_2x_2 + \cdots + Ac_nx_n \\ &= c_1\lambda_1x_1 + c_2\lambda_2x_2 + \cdots + c_n\lambda_nx_n \end{aligned}$$

und für

$$A^ku_0 = c_1\lambda_1^kx_1 + c_2\lambda_2^kx_2 + \cdots + c_n\lambda_n^kx_n,$$

d.h. mit bekannten Eigenwerten und Eigenvektoren ist das iterative Lösen der Gleichung (1.6) erster Ordnung sehr einfach.
Wir betrachten nun als Beispiel einen Evolutionsprozess zweiter Ordnung:

Beispiel 14. *Man betrachte die Fibonacci-Reihe*

$$0, 1, 1, 2, 3, 5, 8, 13, \ldots$$

Fragen:

1. *Was ist F_{100}?*

2. *Wie schnell wächst die Fibonacci-Reihe?*

Das iterative Verfahren für die Fibonacci-Reihe lautet:

$$F_{k+2} = F_{k+1} + F_k$$

Dies ist ein Verfahren 2. Ordnung. Der Trick ist nun, dieses als System erster Ordnung zu formulieren.

$$\begin{aligned} F_{k+2} &= F_{k+1} + F_k \\ F_{k+1} &= F_{k+1}. \end{aligned}$$

Wir setzen $u_k := \begin{pmatrix} F_{k+1} \\ F_k \end{pmatrix}$. *Somit haben wir nun zu lösen:*

$$Au_k = u_{k+1}$$

mit $A = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$. *Wir berechnen nun*

$$\det(A - \lambda Id) = \left| \begin{pmatrix} 1-\lambda & 1 \\ 1 & -\lambda \end{pmatrix} \right| = -\lambda(1-\lambda) - 1 = 0$$

$$\begin{aligned} \Leftrightarrow \lambda^2 - \lambda - 1 &= 0 \\ \Leftrightarrow \lambda_{1/2} &= \frac{1 \pm \sqrt{1+4}}{2} \\ \Rightarrow \lambda_1 &\approx 1.618 \\ \lambda_2 &\approx -0.618. \end{aligned}$$

Da wir zwei unterschiedliche Eigenwerte für das 2×2 *System berechnet haben, wissen wir, dass es zwei linear unabhängige Eigenvektoren geben muss.*
Zu Frage 2: Es gilt

$$A^k = (S\Lambda S^{-1})^k = S \begin{pmatrix} 1.6^k & 0 \\ 0 & -0.6^k \end{pmatrix} S^{-1}$$

$$\Rightarrow \; F_{100} \approx c_1 \cdot 1.6^{100},\; F_k \approx c_1 \cdot 1.6^k$$

für $k \to \infty$. *Die Fibonacci-Reihe wächst somit etwa mit dem Faktor 1.6.*
Zu Frage 1: Berechne u_k *mit*

$$u_k = c_1 x_1 + c_2 x_2$$

Zu berechnen ist also x_1, x_2. *Dazu lösen wir*

$$\begin{aligned} (A - \lambda Id)x &= 0 \\ \Leftrightarrow \begin{pmatrix} 1-\lambda & 1 \\ 1 & -\lambda \end{pmatrix} \begin{pmatrix} x^1 \\ x^2 \end{pmatrix} &= \begin{pmatrix} 0 \\ 0 \end{pmatrix} \end{aligned}$$

Daraus ergibt sich etwa $x_1 = \begin{pmatrix}\lambda_1\\1\end{pmatrix}$ *und* $x_2 = \begin{pmatrix}\lambda_2\\1\end{pmatrix}$. *Um* c_1 *und* c_2 *zu bestimmen, verwenden wir* $u_0 = \begin{pmatrix}F_1\\F_0\end{pmatrix} = \begin{pmatrix}1\\0\end{pmatrix}$. *Daraus folgt:*

$$c_1 x_1 + c_2 x_2 = c_1 \begin{pmatrix}\lambda_1\\1\end{pmatrix} + c_2 \begin{pmatrix}\lambda_2\\1\end{pmatrix} = \begin{pmatrix}1\\0\end{pmatrix}.$$

Es bleibt somit nur noch das 2×2*-System zu lösen und dies ergibt* $c_{1/2} = \pm\frac{1}{\sqrt{5}} \approx \pm 0.447$. *Also* $F_{100} \approx 0.447 \cdot 1.6^{99} \approx 3.54 \times 10^{20}$.

1.3.2 Anwendungen für gewöhnliche Differentialgleichungen

Wir betrachten nun das Differentialgleichungssystem

$$\begin{aligned}\frac{du_1}{dt} &= -u_1 + 2u_2 \\ \frac{du_2}{dt} &= u_1 - 2u_2\end{aligned} \tag{1.8}$$

mit $u(0) = \begin{pmatrix}1\\0\end{pmatrix}$. Dieses System können wir schreiben als

$$\begin{pmatrix}\frac{du_1}{dt}\\ \frac{du_2}{dt}\end{pmatrix} = \begin{pmatrix}-1 & 2\\ 1 & -2\end{pmatrix}\begin{pmatrix}u_1\\u_2\end{pmatrix} = Au. \tag{1.9}$$

Die Eigenwerte von A sind $\lambda_1 = 0$ (da A singulär) und $\lambda_2 = -3$ (da $\sum \lambda_i = \text{spur}(A) = -3$). Eigenvektoren sind

$$\begin{aligned}x_1 &= \begin{pmatrix}2\\1\end{pmatrix}, \\ x_2 &= \begin{pmatrix}1\\-1\end{pmatrix}.\end{aligned}$$

Die allgemeine Lösung u von (1.8) lautet:

$$u(t) = c_1 \exp(\lambda_1 t)\, x_1 + c_2 \exp(\lambda_2 t)\, x_2. \tag{1.10}$$

Zur Bestimmung von c_1 und c_2 setzen wir Eigenwerte und Eigenvektoren ein:

$$c_1 \exp(\lambda_1 t)\, x_1 + c_2 \exp(\lambda_2 t)\, x_2 = c_1 \begin{pmatrix}2\\1\end{pmatrix} + c_2 \exp(-3t) \begin{pmatrix}1\\-1\end{pmatrix}$$

Mit $u(0) = \begin{pmatrix}1\\0\end{pmatrix}$ und für $t = 0$ folgt:

$$c_1 \begin{pmatrix}2\\1\end{pmatrix} + c_2 \begin{pmatrix}1\\-1\end{pmatrix} = \begin{pmatrix}1\\0\end{pmatrix} \quad \Leftrightarrow \quad c_1 = c_2 = \frac{1}{3}$$

$$\Rightarrow u(t) = \frac{1}{3}x_1 + \frac{1}{3}\exp(-3t)\,x_2.$$

Anhand der Lösung lässt sich nun beispielsweise direkt ablesen, was für $t \to \infty$ gilt:

$$\Rightarrow u(\infty) = \frac{1}{3}\begin{pmatrix}2\\1\end{pmatrix}.$$

1.3.3 Drei mögliche Entwicklungen

Basierend auf der Verteilung der Eigenwerte können wir drei allgemeine Fälle voraussagen:

1. $u(t) \to 0$: „Stabilität“
2. $u(t) \to v$: „Steady-State
3. $u(t) \to \pm\infty$: „Explosion“

Im Fall von $u(t) \to 0$ muss gelten:

$$e^{\lambda t} \to 0 \Rightarrow \lambda < 0.$$

An dieser Stelle kann man sich fragen, was für $\lambda \in \mathbb{C}$ gelten muss. Wir betrachten dazu folgendes Beispiel:

Beispiel 15. *Sei $\lambda = -3 + 6i$. Dann gilt:*

$$\left|e^{(-3+6i)t}\right| = \left|e^{-3t}\right|\left|e^{6it}\right|$$

Und mit

$$e^{6it} = \cos(6t) + i\sin(6t) = \begin{pmatrix}\cos(6t)\\\sin(6t)\end{pmatrix}\begin{pmatrix}1\\i\end{pmatrix}$$

folgt

$$\left|e^{6it}\right| = \left|\begin{pmatrix}\cos(6t)\\\sin(6t)\end{pmatrix}\right| = \sqrt{\cos^2(6t) + \sin^2(6t)} = 1.$$

D.h. entscheidend ist hier der Realteil des Eigenwertes. Für $Re(\lambda) < 0$ folgt Stabilität.

Der „Steady-State“-Fall tritt ein, wenn ein Eigenwert $\lambda = 0$ und $Re(\lambda) < 0$ für alle anderen Eigenwerte. Der dritte Fall „Explosion“ tritt ein, wenn mindestens ein $Re(\lambda) > 0$.

Beispiel 16. *Ein 2×2-System mit Matrix $A = \begin{pmatrix}a & b\\c & d\end{pmatrix}$ ist stabil, wenn gilt:*

$$\begin{aligned} Re\lambda_1 &< 0\\ \text{und } Re\lambda_2 &< 0. \end{aligned}$$

Die Spur von A ist $\operatorname{spur}(A) = a + d = \lambda_1 + \lambda_2 < 0$. Reicht $\operatorname{spur}(A) < 0$ für die Stabilität? Nein, denn die Spur, zum Beispiel, von $\begin{pmatrix}-2 & 0\\0 & 1\end{pmatrix}$ ist kleiner 0, aber $\lambda_2 = 1 > 0$ und somit „explodiert“ dieses System. Stabilität benötigt hier also noch eine weitere Bedingung, zum Beispiel

$$\det(A) > 0.$$

1.4 Eigenwerte und Eigenvektoren in Differentialgleichungen mit Exponentialfunktionen

Sei $A \in \mathbb{C}^{n\times n}$ diagonalisierbar. Wir betrachten in diesem Abschnitt das Differentialgleichungssystem

$$\frac{du}{dt} = Au.$$

Die Matrix A koppelt die Anteile u_i aus u. Mit Hilfe der Eigenvektoren lassen sich diese Kopplungen auflösen. Man setze dazu $u := Sv$, wobei S die Eigenvektormatrix sei. Damit erhalten wir:

$$\frac{dv}{dt} = S^{-1}ASv = \Lambda v$$

und schließlich

$$\begin{aligned} \Rightarrow v(t) &= e^{\Lambda t}v(0) \\ \Leftrightarrow S^{-1}u(t) &= e^{\Lambda t}S^{-1}u(0) \\ \Leftrightarrow u(t) &= Se^{\Lambda t}S^{-1}u(0). \end{aligned}$$

Es gilt ferner (s.u.)

$$e^{At} = Se^{\Lambda t}S^{-1},$$

und daher

$$u(t) = e^{At}u(0) = Se^{\Lambda t}S^{-1}u(0).$$

Was bedeutet eine Matrix im Exponenten? Wir erklären diese Schreibweise mit Hilfe der Taylorentwicklung von

$$e^{At} = Id + At + \frac{(At)^2}{2} + \frac{(At)^3}{6} + \ldots + \frac{(At)^n}{n!} + \ldots$$

Warum ist nun $e^{At} = Se^{\Lambda t}S^{-1}$? Verwende dazu die Taylorreihe. Es folgt

$$\begin{aligned} e^{At} &= Id + At + \frac{(At)^2}{2} + \frac{(At)^3}{6} + \ldots \\ &= Id + (S\Lambda S^{-1})t + \frac{(S\Lambda S^{-1})(S\Lambda S^{-1})t^2}{2 + \ldots} = \\ &= Id + (S\Lambda S^{-1})t + \frac{1}{2}t^2(S\Lambda^2 S^{-1}) + \frac{1}{6}t^3(S\Lambda^3 S^{-1}) + \ldots \\ &= S(S^{-1} + \Lambda S^{-1}t + \frac{1}{2}t^2\Lambda^2 S^{-1} + \ldots) = \\ &= S(Id + \Lambda t + \Lambda^2\frac{t^2}{2} + \frac{\Lambda^3 t^3}{6} + \ldots)S^{-1} = \\ &= Se^{\Lambda t}S^{-1} \end{aligned}$$

$$\Rightarrow e^{At} = e^{S\Lambda S^{-1}t} = Se^{\Lambda t}S^{-1}.$$

Dabei ist

$$e^{\Lambda t} = \begin{pmatrix} e^{\lambda_1 t} & 0 & \cdots \\ & \ddots & \\ \cdots & 0 & e^{\lambda_n t} \end{pmatrix}.$$

Daraus folgt zum Beispiel

$$e^{At} = S e^{\Lambda t} S^{-1} \to 0,$$

falls $e^{\Lambda t} \to 0$. Dies ist wiederum der Fall, wenn alle $Re(\lambda) < 0$.

1.4.1 Differentialgleichungen höherer Ordnung

Differentialgleichungen höherer Ordnung lassen sich in ein System von Gleichungen erster Ordnung umwandeln. Am Beispiel der Differentialgleichung 2. Ordnung vom Typ

$$y'' + by' + cy = 0 \tag{1.11}$$

wandelt man mit $u = \begin{pmatrix} y' \\ y \end{pmatrix}$ und

$$\begin{aligned} y'' + by' + cy &= 0 \\ y' &= y' \end{aligned}$$

die Differentialgleichung 2. Ordnung in ein System mit zwei Gleichungen 1. Ordnung um. Man erhält

$$u' = \begin{pmatrix} y'' \\ y' \end{pmatrix} = \begin{pmatrix} -b & -c \\ 1 & 0 \end{pmatrix} \begin{pmatrix} y' \\ y \end{pmatrix} = \begin{pmatrix} -b & -c \\ 1 & 0 \end{pmatrix} u. \tag{1.12}$$

2 Gewöhnliche Differentialgleichungen

Dieses Kapitel beschäftigt sich mit den Grundlagen zur Lösung von gewöhnlichen Differentialgleichungen. Die behandelten numerischen Verfahren lassen sich später im Rahmen der zeitlichen Diskretisierung bei partiellen Differentialgleichungen einsetzen.

2.1 Explizite und implizite Verfahren, Stabilität und Steifheit

Gegeben sei die Differentialgleichung

$$\begin{aligned} u' &= f(u,t) \\ u(t=0) &= u_0 \end{aligned} \tag{2.1}$$

oder das System

$$\begin{aligned} u_i' &= f_i(u,t), \quad i = 1,\ldots,n \\ u_i(t=0) &= u_{i,0}. \end{aligned} \tag{2.2}$$

Wir betrachten in diesem Kapitel lineare Differentialgleichungen mit konstanten Koeffizienten vom Typ

$$u' = au, \quad a \approx \frac{\partial f}{\partial u} \tag{2.3}$$

und

$$u' = Au, \quad A_{ij} \approx \frac{\partial f_i}{\partial u_j}.^{1} \tag{2.4}$$

Bei Verfahren zur Lösung von gewöhnlichen Differentialgleichungen unterscheidet man grob zwischen *expliziten* und *impliziten* Verfahren. Während bei expliziten Verfahren die neue Lösung u_{n+1} nur von $u_n, u_{n-1}, \ldots$ und $t_n, t_{n-1}, \ldots$ abhängt, existiert bei den impliziten Verfahren eine zusätzliche Abhängigkeit von u_{n+1} und t_{n+1}. Da implizite Verfahren vom neuen Schritt $n+1$ abhängen, müssen diese Kopplungen mit zusätzlichem numerischen Aufwand aufgelöst werden und sind deshalb typischerweise langsamer als explizite

[1] Für hinreichend reguläre rechte Seiten lässt sich das in (2.2) gegebene System gewöhnlicher Differentialgleichungen lokal durch das System

$$\vec{u}'(t) \;\approx\; D_u f\,(u(t),t)\,\vec{u}$$

wie in (2.4) approximieren. Ist die Jacobi-Matrix diagonalisierbar, so zerfällt dieses Näherungsproblem in ein System entkoppelter Gleichungen vom Typ wie in (2.3). Daher ist es oft ausreichend, sich in der Betrachtung etwa der Stabilität auf dieses Problem zurückzuziehen.

Verfahren. Warum würde man dann die langsamen, impliziten Verfahren einsetzen? Eine Klasse von gewöhnlichen Differentialgleichungen, die sogenannte „steife Probleme“ beschreiben, profitieren signifikant von impliziten Verfahren, da diese „Stabilität“ aufweisen. Um diese Begriffe zu veranschaulichen, betrachte man zunächst folgendes Beispiel:

Beispiel 17. *Sei*

$$u(t) = e^{-t} + e^{-99t}.$$

Isoliert stellen die einzelnen Beiträge von $u(t)$ kein Problem dar, in Kombination von unterschiedlich „schnellen“ Prozessen kontrolliert einerseits der langsame Anteil (in diesem Fall e^{-t}) die Lösung $u(t)$, andererseits kontrolliert der schnelle Anteil (hier e^{-99t}) den möglichen Zeitschritt Δt eines numerischen Verfahrens. Solche Gleichungen nennt man „steif“.

Steife Probleme treten beispielsweise auf bei

- chemischen Prozessen mit unterschiedlich schnellen Abbauraten,
- biologischen Prozessen,
- in Systemen von gewöhnlichen Differentialgleichungen.

Beispiel 18. *Sei $A = \begin{pmatrix} -50 & 49 \\ 49 & -50 \end{pmatrix}$. Die Lösung von $u = Au$ lautet wie im vorigen Beispiel*

$$u(t) = e^{-t} + e^{-99t}.$$

A hat die Eigenwerte $\lambda_1 = -1$ und $\lambda_2 = -99$. Der Größenunterschied der Eigenwerte ist ein Indiz für die Steifigkeit des Problems.

Die Kondition liefert ein Maß für die „Steifigkeit“:

Definition 4. *Die Kondition einer Matrix A wird definiert als:*

$$cond(A) := \frac{|\lambda|_{max}}{|\lambda|_{min}}. \tag{2.5}$$

In obigem Beispiel ist also $cond(A) = 99$.

Bemerkung 5. *Je größer die Kondition einer Matrix, desto „steifer“ ist das Problem. Für solche Probleme benötigt man stabile implizite Verfahren.*

Wir wollen uns nun an einer Definition des Begriffs der *Stabilität* versuchen. Im Grunde soll ein Verfahren *stabil* heißen, wenn kleine Störungen in den Problemdaten nur zu kleinen Störungen in der diskreten Lösung führen. Allerdings kann schon die exakte Lösung in diesem Sinne instabil sein. Man betrachte etwa die Differentialgleichung (2.3) mit $a > 0$. Wir konkretisieren also sinngemäß:

Definition 5. *Ein Verfahren zur Lösung einer Anfangswertaufgabe heißt* stabil*, wenn Störungen in den Problemdaten nicht zu wesentlich größeren Störungen in der diskreten Lösung als in der exakten Lösung führen.*

Im Folgenden soll untersucht werden, unter welchen Bedingungen Verfahren stabil bzw. instabil sein können. Dabei werden das explizite und implizite Eulerverfahren als Repräsentanten dieser Klassen vorgestellt. Dazu betrachtet man wieder die gewöhnliche Differentialgleichung

$$u' = f(u,t) = au. \tag{2.6}$$

2.1.1 Explizites Eulerverfahren

Das explizite Eulerverfahren approximiert den Differentialquotienten durch einen Differenzenquotienten der Form

$$u' \approx \frac{u_{n+1} - u_n}{\Delta t}. \tag{2.7}$$

Diese Approximation führt zu der Iterationsvorschrift

$$\begin{aligned} \frac{u_{n+1} - u_n}{\Delta t} &= au_n \\ \Leftrightarrow\ u_{n+1} &= u_n + \Delta t a u_n = (1 + a\Delta t)u_n \end{aligned}$$

$$\Rightarrow\ u_n = (1 + a\Delta t)^n u_0.$$

Hier sieht man, dass mit $1+a\Delta t > 1$ die Lösung u_n über alle Schranken wächst, während für $a < 0$ die korrekte Lösung gegen Null konvergiert. Damit das Verfahren stabil ist, müssen wir also fordern, dass

$$|1 + a\Delta t| \leq 1. \tag{2.8}$$

Da a problemspezifisch ist, kann diese Bedingung nur durch geeignete Wahl von Δt erfüllt werden. Für den Fall, dass $a < 0$ folgt:

$$0 < \Delta t < -\frac{2}{a}.$$

Man sieht: Explizite Verfahren können instabil werden.

2.1.2 Implizites Eulerverfahren

Das implizite Verfahren hängt von einer rechten Seite mit u_{n+1} (anstelle von u_n) ab:

$$\frac{u_{n+1} - u_n}{\Delta t} = f(u_{n+1}, t_{n+1}) = au_{n+1} \tag{2.9}$$

$$\Leftrightarrow\ u_{n+1} = \frac{1}{1 - a\Delta t} u_n \tag{2.10}$$

$$\Rightarrow\ u_n = \left(\frac{1}{1 - a\Delta t}\right)^n u_0.$$

Für $a < 0$ gilt:

$$\left|\frac{1}{1 - a\Delta t}\right| < 1 \quad \Rightarrow \text{ stabil unabhängig von } \Delta t.$$

2.1.3 Zusammenfassung

- Steife Probleme werden durch „langsame“ Anteile dominiert, der Zeitschritt jedoch durch „schnelle“ Anteile. Ob unterschiedliche Skalen in einem Problem existieren, kann durch die Berechnung der Eigenwerte der Systemmatrix geprüft werden.
- Die Kondition der Systemmatrix ist ein Maß für die Steifigkeit eines Problems.
- Explizite Verfahren eignen sich *nicht* für steife Probleme.
- Bei Senkentermen ($a < 0$) sind implizite Verfahren stabil, für Quellenterme ($a > 0$) auch explizite.

2.2 Fehleranalyse, Konvergenz und Konsistenz

Bei allen numerischen Verfahren stellt sich die Frage, wie gut die numerische Approximation an die (meist unbekannte) exakte Lösung ist. Es ist also einerseits entscheidend, zu wissen, wie *gut* die numerische Lösung im Vergleich zur exakten Lösung ist, andererseits entscheidet auch der Aufwand über die Brauchbarkeit eines numerischen Verfahrens. Dazu stellt sich die Frage, wie schnell für $\Delta t \to 0$ eine gute Approximation erreicht wird.
Wir widmen uns zunächst der Fehleranalyse und betrachten dazu die Iterationsvorschrift

$$u_{k+1} = u_k + \Delta t \Phi(u_{k+1}, u_k, t_k). \tag{2.11}$$

Φ wird üblicherweise *Verfahrensfunktion* genannt. In den folgenden Abschnitten gehen wir stets von einer konstanten Zeitschrittweite aus, d.h.

$$t_k = t_0 + k \cdot \Delta t,$$

und bezeichnen mit

$$\begin{array}{ll} u(t_k): & \text{die exakte Lösung in } t_k, \\ u_k: & \text{die approximierte Lösung in } t_k. \end{array}$$

2.2.1 Diskretisierungsfehler

Der Diskretisierungsfehler liefert Informationen darüber, welcher Fehler durch das numerische Diskretisierungsverfahren in einem Iterationsschritt gemacht wird.

Definition 6. *Unter dem lokalen Diskretisierungsfehler in t_{k+1} versteht man den Wert*

$$d_{k+1} := u\left(t_{k+1}\right) - u\left(t_k\right) - \Delta t \Phi\left(u\left(t_{k+1}\right), u\left(t_k\right), \ldots, t_k\right). \tag{2.12}$$

Beispiel 19. *Für das explizite Eulerverfahren für $u' = au$ gilt:*

$$\begin{aligned} u_{n+1} &= u_n + \Delta t a u_n \\ u(t_{n+1}) &= u(t_n) + \Delta t a u(t_n) + d_{n+t} \\ \Rightarrow d_{n+1} &= u(t_{n-1}) - u(t_n) - \Delta t a u(t_n) \end{aligned}$$

Während der Diskretisierungsfehler lokale Information über das Verfahren beinhaltet, ist es zusätzlich interessant, zu wissen, welchen Fehler man am Ende zum Zeitpunkt t_n gemacht hat. Was gilt also für

$$e_n := u(t_n) - u_n?$$

In obigem Beispiel ist

$$e_{n+1} = u(t_{n+1}) - u_{n+1} = u(t_n) + a\Delta t u\left(t_n\right) + d_{n+1} - \left(u_n + a\Delta t u_n\right)$$

$$\Rightarrow e_{n+1} = e_n + a\Delta t e_n + d_{n+1} = (1 + a\Delta t)^n d_1 + \ldots + (1 + a\Delta t)^{n+1-k} d_k + \ldots + d_{n+1}.$$

Falls $|1 + a\Delta t| \leq 1$ und $d_{k+1} = \frac{1}{2}\left(\Delta t\right)^2 u''\left(t_k + \theta\Delta t\right)$, $0 < \theta < 1$, so folgt:

$$e_{n+1} \leq (n+1) \cdot \frac{1}{2}\left(\Delta t\right)^2 \left\|u''\right\|_\infty = \frac{1}{2} T \cdot \Delta t \left\|u''\right\|_\infty$$

mit $T := (n+1) \cdot \Delta t$. Man sieht, der Fehler des expliziten (und auch des impliziten) Eulerverfahrens entwickelt sich linear in Δt.

2.2.2 Globaler Fehler

Wie oben schon beispielhaft angedeutet, ist der globale Fehler eines Verfahrens entscheidend für die Qualität der numerischen Lösung.

Definition 7. *Als* globalen Fehler g_k *in* t_k *bezeichnet man die Differenz*

$$g_k \coloneqq u(t_k) - u_k. \tag{2.13}$$

In Abschnitt 2.2.1 hat man gesehen, dass sich der globale Fehler aus den lokalen Fehlern in jedem Iterationsschritt zusammensetzt. Als nächstes wollen wir g_k durch die Diskretisierungsfehler d_k abschätzen. Um dies tun zu können, benötigt man die folgende Regularitätseigenschaft der Verfahrensfunktion:

Definition 8. *Die Verfahrensfunktion* $\Phi : B \to \mathbb{R}$ *erfüllt die* lokale Lipschitz-Bedingung, wenn ein $L \in \mathbb{R}$ mit $0 < L < \infty$ existiert, sodass

$$\begin{aligned} \left|\Phi(x, y_1, z, \Delta t) - \Phi(x, y_2, z, \Delta t)\right| &\leq L\left|y_1 - y_2\right|, \\ \left|\Phi(x, y, z_1, \Delta t) - \Phi(x, y, z_2, \Delta t)\right| &\leq L\left|z_1 - z_2\right| \end{aligned} \tag{2.14}$$

für alle Quadrupel $(x, y_1, z, \Delta t), (x, y_2, z, \Delta t)\,(x, y, z_1, \Delta t)\,(x, y, z_2, \Delta t) \in B$ *gilt.*

Damit nun lässt sich der globale Fehler aus dem lokalen Diskretisierungsfehler abschätzen:
Aus der Definition des lokalen Diskretisierungsfehlers (2.12) folgt:

$$u\left(t_{k+1}\right) = u\left(t_k\right) + \Delta t \cdot \Phi\left(u\left(t_{k+1}\right), u\left(t_k\right), t_k\right) + d_{k+1}. \tag{2.15}$$

Von dieser Gleichung substrahiert man die Definition des Einschrittverfahrens (2.11) und erhält (durch Einschieben eines Zwischenterms)

$$\begin{aligned} g_{k+1} &= g_k + \Delta t \left(\Phi\left(u\left(t_{k+1}\right), u\left(t_k\right), t_k\right) - \Phi\left(u\left(t_{k+1}\right), u_k, t_k\right)\right. \\ &\quad \left. + \Phi\left(u\left(t_{k+1}\right), u_k, t_k\right) - \Phi\left(u_{k+1}, u_k, t_k\right)\right) + d_{k+1}. \end{aligned} \tag{2.16}$$

Wegen der Lipschitzbedingung (2.14) und mit $\Delta t \cdot L < 1$ folgt

$$|g_{k+1}| \leq |g_k| + \Delta t \left(L\left|u\left(t_k\right) - u_k\right| + L\left|u\left(t_{k+1}\right) - u_{k+1}\right|\right) + |d_{k+1}| \tag{2.17}$$

$$\Rightarrow |g_{k+1}| \leq \frac{1 + \Delta tL}{1 - \Delta tL} |g_k| + \frac{|d_{k+1}|}{1 - \Delta tL}. \tag{2.18}$$

$$\tag{2.19}$$

Für den expliziten Fall (Φ hängt nicht von u_{k+1} ab) erhält man analog

$$|g_{k+1}| \leq (1 + \Delta tL) |g_k| + |d_{k+1}|. \tag{2.20}$$

Wegen $\Delta tL < 1$ existiert eine Konstante $K > 0$, sodass $\frac{1+\Delta tL}{1-\Delta tL} \leq 1 + \Delta tK$. Dann kann man in beiden Fällen (implizit und explitzit) schreiben

$$|g_{k+1}| \leq (1 + a) |g_k| + b \tag{2.21}$$

mit $a = \begin{cases} \Delta tK & \text{(implizit)} \\ \Delta tL & \text{(explizit)} \end{cases}$ und $b = \begin{cases} \frac{K}{2L} |d_{k+1}| & \text{(implizit)} \\ |d_{k+1}| & \text{(explizit)} \end{cases}$.
Um diese Aussage weiter zu verwerten, erweist sich das folgende Lemma als hilfreich.

Lemma 1. *(Gronwall)*
Erfüllt die Folge $(g_k)_{k\in\mathbb{N}}$ *die Bedingung*

$$|g_{k+1}| \leq (1 + a) |g_k| + b \quad \forall k \in \mathbb{N}_+, \tag{2.22}$$

so gilt

$$|g_k| \;\leq\; (1 + a)^k |g_0| + \frac{(1 + a)^k - 1}{a} b \;\leq\; e^{ka} |g_0| + \frac{b}{a}\left(e^{ka} - 1\right) \quad \forall k \in \mathbb{N}. \tag{2.23}$$

Beweis.

$$\begin{aligned} |g_k| &\leq (1 + a) |g_{k-1}| + b \;\leq\; (1 + a)^2 |g_{k-2}| + ((1 + a) + 1) b \\ &\vdots \\ &\leq (1 + a)^k |g_0| + \left((1 + a)^{k-1} + \ldots + (1 + a) + 1\right) b \\ &= (1 + a)^k |g_0| + \frac{(1 + a)^k - 1}{a} \cdot b. \end{aligned}$$

Ferner gilt $(1 + t) \leq e^t \; \forall t$, daher$(1 + a)^k \leq e^{ka}$.

□

Mit $g_0 = u(t_0) - u_0 = 0$ ergibt sich aus dem Lemma folgender

Satz 3. *Sei $D \coloneqq \max_k |d_k|$. Für den globalen Fehler g_n an der Stelle $t_n = t_0 + n\Delta t$ gilt für eine explizite Methode*

$$|g_n| \leq \frac{D}{\Delta t L} \left(e^{n\Delta t L} - 1\right) \leq \frac{D}{\Delta t L} \cdot e^{n\Delta t L}. \tag{2.24}$$

Für den impliziten Fall gilt

$$|g_n| \leq \frac{D}{2\Delta t L} \left(e^{n\Delta t K} - 1\right) \leq \frac{D}{2\Delta t L} \cdot e^{n\Delta t K}. \tag{2.25}$$

Der gobale Fehler hängt also proportional von D ab, außerdem von der Lipschitz-Konstanten L und natürlich der Schrittweite h.

Beispiel 20. *Für das explizite Eulerverfahren gilt*

$$d_{k+1} = u\left(t_{k+1}\right) - u\left(t_k\right) - \Delta t f\left(u\left(t_k\right), t_k\right).$$

Mithilfe der Taylorentwicklung ergibt sich

$$u\left(t_{k+1}\right) = u\left(t_k\right) + \Delta t u'\left(t_k\right) + \frac{1}{2}\left(\Delta t\right)^2 u''\left(t_k + \theta\Delta t\right)$$

mit $0 < \theta < 1$. Wegen $f\left(u\left(t_k\right), t_k\right) = u'\left(t_k\right)$ folgt

$$d_{k+1} \;=\; u\left(t_k\right) + \Delta t u'\left(t_k\right) + \frac{1}{2}\left(\Delta t\right)^2 u''\left(t_k + \theta\Delta t\right) - u\left(t_k\right) - \Delta t u'\left(t_k\right) = \frac{1}{2}\left(\Delta t\right)^2 u''\left(t_k + \theta\Delta t\right).$$

Sei $M \coloneqq \max_{t_0 \leq \xi \leq t_n} |u''(\xi)|$, dann gilt $\max |d_{k+1}| \leq \frac{1}{2}\left(\Delta t\right)^2 M$. Eingesetzt in (2.24):

$$|g_n| \leq \frac{\Delta t M}{2L} e^{n\Delta t L}.$$

D.h. g_n nimmt proportional zu Δt ab. Das Eulerverfahren besitzt die Fehlerordnung 1.

Definition 9. *Ein Einschrittverfahren besitzt die Fehlerordnung p, falls (mit einer Konstanten C) für seinen lokalen Diskretisierungsfehler d_k gilt:*

$$\max |d_k| \leq D = C\left(\Delta t\right)^{p+1} = \mathcal{O}\left(\left(\Delta t\right)^{p+1}\right). \tag{2.26}$$

Der globale Fehler g_n ist dadurch folgendermaßen beschränkt:

$$|g_n| \leq \frac{C}{L} e^{n\Delta t L} \left(\Delta t\right)^p = \mathcal{O}\left(\left(\Delta t\right)^p\right). \tag{2.27}$$

Definition 10. *Ein Einschrittverfahren heißt mit der Differentialgleichung* konsistent, *falls die Fehlerordnung mindestens 1 ist.*

2.3 Verbesserte Methoden

Wir wollen nun versuchen, die Fehlerordnung (Euler: 1) zu verbessern. Eine Idee ist beispielsweise, Terme höherer Ordnung aus der Taylorreihe zu verwenden.

2.3.1 Taylorreihenverfahren

Man verwende in der Taylorreihe Terme bis Ordnung p:

$$u\left(t_{k+1}\right)=u\left(t_k\right)+\frac{\Delta t}{1!}u'\left(t_k\right)+\frac{\left(\Delta t\right)^2}{2!}u''\left(t_k\right)+\frac{\left(\Delta t\right)^3}{3!}u^{(3)}\left(t_k\right)+\ldots+\frac{\left(\Delta t\right)^p}{p!}u^{(p)}\left(t_k\right)+R_{p+1} \tag{2.28}$$

und erreichen dadurch ein Verfahren der Ordnung p mit dem lokalen Diskretisierungsfehler

$$d_{k+1}=R_{p+1}=\frac{\left(\Delta t\right)^{p+1}}{(p+1)!}u^{(p+1)}\left(t_k+\theta\Delta t\right),\quad 0<\theta<1. \tag{2.29}$$

Der spätere Abbruch der Taylorreihe liefert eine bessere lokale Approximation der gesuchten Funktion, die Schwierigkeit besteht jedoch in der Berechnung zusätzlicher höherer Ableitung bis zur p-ten Ordnung.

Beispiel 21. *Man betrachte die Funktion*

$$u'=-2tu^2 \tag{2.30}$$

mit $u(0)=1$. Die Taylorentwicklung von u um t liefert

$$u\left(t_{k+1}\right)=u\left(t_k\right)+c_1\Delta t+c_2\left(\Delta t\right)^2+c_3\left(\Delta t\right)^3+c_4\left(\Delta t\right)^4+\ldots \tag{2.31}$$

mit unbekannten Koeffizienten c_i. Setze in $u'=-2tu^2$ ein, mit $t=t_k+\Delta t$:

$$\begin{aligned}
c_1+2c_2\Delta t \;+\; & 3c_3\left(\Delta t\right)^2+4c_4\left(\Delta t\right)^3+\ldots \\
= \; & -2(t_k+\Delta t)\left(u\left(t_k\right)+c_1\Delta t+c_2\left(\Delta t\right)^2+c_3\left(\Delta t\right)^3+c_4\left(\Delta t\right)^4+\ldots\right)^2 \\
= \; & -2(t_k+\Delta t)\left(u^2\left(t_k\right)+2c_1u\left(t_k\right)\Delta t+\left(c_1^2+2c_2u\left(t_k\right)\right)\left(\Delta t\right)^2+\right. \\
& \left.+\left(2c_1c_2+2c_3u\left(t_k\right)\right)\left(\Delta t\right)^3+\ldots\right) \\
= \; & -2t_ku^2\left(t_k\right)+\left(-2u^2\left(t_k\right)-4c_1t_ku\left(t_k\right)\right)\Delta t+ \\
& \left(-4c_1u\left(t_k\right)-2t_k\left(c_1^2+2c_2u\left(t_k\right)\right)\right)\left(\Delta t\right)^2+ \\
& +\left(-2\left(c_1^2+2c_2u\left(t_k\right)\right)-4t_k\left(c_1c_2+c_3u\left(t_k\right)\right)\right)\left(\Delta t\right)^3+\ldots
\end{aligned}$$

Ein Koeffizientenvergleich liefert

$$c_1=-2t_ku^2\left(t_k\right) \qquad \approx -2t_ku_k^2 \tag{2.32}$$

$$c_2=-\left(u\left(t_k\right)+2c_1t_k\right)u\left(t_k\right) \qquad \approx -\left(u_k-2c_1t_k\right)u_k \tag{2.33}$$

$$c_3=\frac{-\left(4c_1u\left(t_k\right)+2t_k\left(c_1^2+2c_2u\left(t_k\right)\right)\right)}{3} \qquad \approx -\frac{1}{3}\left(4c_1u_k+2t_k\left(c_1^2+2c_2u_k\right)\right) \tag{2.34}$$

$$c_4=-\frac{1}{2}c_1^2-c_2u\left(t_k\right)-t_k\left(c_1c_2+c_3u\left(t_k\right)\right) \qquad \approx -\frac{1}{2}c_1^2-c_2u_k-t_k\left(c_1c_2+c_3u_k\right) \tag{2.35}$$

Bei Betrachtung des Approximationsfehlers $e_k := u(t_k) - u_k$ weist dieses Verfahren eine bessere Näherung als das explizite Eulerverfahren auf.

2.3.2 Verbesserte Polygonzugmethode

Man kombiniert zwei Schrittweiten (Δt und $\frac{\Delta t}{2}$).

$$u_{k+1}^{(1)} = u_k + \Delta t f(u_k, t_k) \tag{2.36}$$

$$u_{k+\frac{1}{2}}^{(2)} = u_k + \frac{\Delta t}{2} f(u_k, t_k) \tag{2.37}$$

$$u_{k+1}^{(2)} = u_{k+\frac{1}{2}}^{(2)} + \frac{\Delta t}{2} f\left(u_{k+\frac{1}{2}}^{(2)}, t_k + \frac{\Delta t}{2}\right) \tag{2.38}$$

Man definiert dann (*Richardson-Extrapolation)*:

$$\begin{aligned} u_{k+1} &:= 2u_{k+1}^{(2)} - u_{k+1}^{(1)} = 2u_{k+\frac{1}{2}}^{(2)} + \Delta t f\left(u_{k+\frac{1}{2}}^{(2)}, t_k + \frac{\Delta t}{2}\right) - u_k - \Delta t f(u_k, t_k) \\ &= u_k + \Delta t f\left(u_k + \frac{\Delta t}{2} f(u_k, t_k), t_k + \frac{\Delta t}{2}\right). \end{aligned}$$

Algorithmisch berechnet man dazu:

$$k_1 := f(u_k, t_k), \tag{2.39}$$

$$k_2 := f\left(u_k + \frac{\Delta t}{2} k_1, t_k + \frac{\Delta t}{2}\right), \tag{2.40}$$

$$u_{k+1} = u_k + \Delta t k_2. \tag{2.41}$$

Diese Methode ist von der Ordnung 2.

2.3.3 Trapezmethode

Durch eine äquivalente Integraldarstellung des ursprünglichen ODE-Problems gelangt man zu der *Trapezmethode*.
Im Folgenden sei die Gleichung

$$u'(t) = f(u(t), t)$$

zugrunde gelegt. Auf dem Intervall $[t_k, t_{k+1}]$ wird auf beiden Seiten integriert:

$$\int_{t_k}^{t_{k+1}} u'(t)dt = \int_{t_k}^{t_{k+1}} f(u(t), t)\, dt \tag{2.42}$$

$$\Leftrightarrow \quad u(t_{k+1}) - u(t_k) = \int_{t_k}^{t_{k+1}} f(u(t), t)\, dt. \tag{2.43}$$

Da $u(t)$ unbekannt ist, wird das Integral auf der rechten Seite approximiert durch (Trapezregel):

$$\int_{t_k}^{t_{k+1}} f\left(u(t),t\right) dt \quad \approx \quad \frac{\Delta t}{2}\left(f\left(u_k,t_k\right)+f\left(u_{k+1},t_{k+1}\right)\right).$$

Die so konstruierte Trapezmethode ist also durch die Verfahrensfunktion

$$u_{k+1} = u_k + \frac{\Delta t}{2}\left(f\left(u_k,t_k\right)+f\left(u_{k+1},t_{k+1}\right)\right) \tag{2.44}$$

definiert.
Da die Trapezmethode ein implizites Verfahren ist, muss in jedem Schritt eine Näherung für die Lösung u_{k+1} von (2.44) gefunden werden. Dies kann durch eine Fixpunktiteration geschehen:

$$u_{k+1}^{(0)} = u_k + \Delta t f\left(u_k,t_k\right) \tag{2.45}$$

$$u_{k+1}^{(n+1)} = u_k + \frac{\Delta t}{2}\left(f\left(u_k,t_k\right)+f\left(u_{k+1}^{(n)},t_{k+1}\right)\right) \tag{2.46}$$

Diese Iteration konvergiert gegen den Fixpunkt u_{k+1}, falls f im ersten Argument Lipschitzstetig (mit Konstante L) ist und $\frac{\Delta t L}{2} < 1$, bzw. $\Delta t < \frac{2}{L}$ (wie man leicht durch Anwenden des Fixpunktsatzes von Banach nachprüft).

2.3.3.1 Fehlerordnung der Trapezmethode

Die Trapezmethode besitzt die Verfahrensfunktion

$$\Phi\left(u_k,u_{k+1},t_k\right) := \frac{1}{2}\left(f\left(u_k,t_k\right)+f\left(u_{k+1},t_{k+1}\right)\right).$$

Der lokale Diskretisierungsfehler lässt sich demnach berechnen durch

$$\begin{aligned}
d_{k+1} &= u\left(t_{k+1}\right) - u\left(t_k\right) - \frac{\Delta t}{2}\left(f\left(u\left(t_k\right),t_k\right)+f\left(u\left(t_{k+1}\right),t_{k+1}\right)\right)\\
&= u\left(t_{k+1}\right) - u\left(t_k\right) - \frac{\Delta t}{2}\left(u'\left(t_k\right)+u'\left(t_{k+1}\right)\right)\\
&= \Delta t u'\left(t_k\right) + \frac{\left(\Delta t\right)^2}{2}u''\left(t_k\right) + \frac{\left(\Delta t\right)^3}{6}u'''\left(t_k\right) + \mathcal{O}\left(\left(\Delta t\right)^4\right)\\
&\quad -\frac{\Delta t}{2}\left(u'\left(t_k\right)+u'\left(t_k\right)+\Delta t u''\left(t_k\right)+\frac{\left(\Delta t\right)^2}{2}u'''\left(t_k\right)+\mathcal{O}\left(\left(\Delta t\right)^3\right)\right)\\
&= -\frac{1}{12}\left(\Delta t\right)^3 u'''\left(t_k\right) + \mathcal{O}\left(\left(\Delta t\right)^4\right).
\end{aligned}$$

Das heißt, dass der Hauptteil des lokalen Diskretisierungsfehlers proportional zu $\left(\Delta t\right)^3$ ist und damit hat die Trapezmethode Fehlerordnung 2.
Diese ist identisch mit der Ordnung der verbesserten Polygonzugmethode. Wir werden aber später sehen, dass die Trapezmethode bessere *Stabilitätseigenschaften* hat.

2.3.4 Verfahren von Heun

In obigem Ansatz bleibt zu entscheiden, wie viele Iterationsschritte in der Fixpunktiteration ausgeführt werden. In der Praxis wird oft nur ein Schritt ausgeführt, aufgrund der Tatsache, dass durch das numerische Verfahren ohnehin Näherungen $u_{k+1} \approx u(t_{k+1})$ berechnet werden. Das Verfahren von Heun besteht aus einem Eulerschritt als *Prädiktor* und einem *Korrekturschritt* und wird deshalb auch als Prädiktor-Korrektor-Methode bezeichnet:

$$u_{k+1}^{(p)} = u_k + \Delta t f\left(u_k, t_k\right), \tag{2.47}$$

$$u_{k+1} = u_k + \frac{\Delta t}{2}\left(f\left(u_k, t_k\right) + f\left(u_{k+1}^{(p)}, t_{k+1}\right)\right). \tag{2.48}$$

Das algorithmische Vorgehen ist dabei

$$k_1 = f\left(u_k, t_k\right), \tag{2.49}$$

$$k_2 = f\left(u_k + \Delta t k_1, t_{k+1}\right), \tag{2.50}$$

$$u_{k+1} = u_k + \frac{\Delta t}{2}\left(k_1 + k_2\right). \tag{2.51}$$

Dies entspricht einer Mittelung der Steigungen k_1 und k_2 in den Punkten (t_k, u_k) und $\left(t_{k+1}, u_{k+1}^{(p)}\right)$. Die Fehlerordnung des Heun-Verfahrens ist 2, der Beweis ist analog zur Polygonzugmethode zu führen.

3 Partielle Differentialgleichungen

In diesem Kapitel werden Herleitungen für partielle Differentialgleichungen (PDEs) unterschiedlicher physikalischer Phänomene gezeigt. Die so motivierten PDEs werden in den späteren Kapiteln numerisch behandelt.

3.1 Gängige Operatoren der mehrdimensionalen Analysis

Im Folgenden werden einige grundlegende Begriffe und Operatoren definiert, die später in den klassischen Anwendungsfällen, d.h. für klassische partielle Differentialgleichungen, benötigt werden. Dieser Abschnitt liefert also nur einen marginalen Einblick in die mehrdimensionale Analysis, die ohnehin eine Grundvoraussetzung für die Analyse höherdimensionaler Probleme ist.

Definition 11. *Sei $U \subset \mathbb{R}^n$ eine offene Menge und $f : U \to \mathbb{R}$ eine reelle Funktion. f heißt im Punkt $x \in U$* partiell differenzierbar *bzgl. der i-ten Koordinatenrichtung, falls*

$$D_i f(x) := \lim_{h \to 0} \frac{f(x + he_i) - f(x)}{h} \tag{3.1}$$

existiert. Dabei ist $e_i \in \mathbb{R}^n$ der i-te Einheitsvektor, also $(e_i)_j = \delta_{ij}$, und wir schreiben auch $\partial_i f$ oder $\partial_{e_i} f$ oder $\frac{\partial f}{\partial x_i}$ für $D_i f$.

Definition 12. *Sei $U \subset \mathbb{R}^n$ offen, $f : U \to \mathbb{R}$ partiell differenzierbar (bzgl. jeder Koordinatenrichtung) und $x \in U$. Dann heißt*

$$\text{grad } f(x) := \left(\frac{\partial f}{\partial x_1}(x), \ldots, \frac{\partial f}{\partial x_n}(x) \right) \tag{3.2}$$

der Gradient *von f im Punkt x.*

Bemerkung 6. *Statt* grad(f) *wird häufig auch ∇f geschrieben mit*

$$\nabla := \left(\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_n} \right)$$

als vektorwertiger Differentialoperator.

Satz 4. *Seien $f, g : U \to \mathbb{R}$ zwei partiell differenzierbare Funktionen. Dann gilt*

$$\text{grad } (f \cdot g) = g \cdot \text{grad } f + f \cdot \text{grad } g \tag{3.3}$$

Definition 13. *Sei $U \subset \mathbb{R}^n$ eine offene Menge und*

$$v = (v_1, \ldots, v_n) : U \to \mathbb{R}^n$$

ein partiell differenzierbares Vektorfeld (d.h. alle v_i sind partiell differenzierbar). Dann heißt die Funktion

$$\text{div } v := \sum_{i=1}^{n} \frac{\partial v_i}{\partial x_i} \tag{3.4}$$

die Divergenz *des Vektorfeldes v.*

Bemerkung 7. *Die Divergenz schreibt man häufig auch als Skalarprodukt:*

$$\nabla \cdot v.$$

Definition 14. *Sei $U \subset \mathbb{R}^3$ offen. Für ein partiell differenzierbares Vektorfeld $v : U \to \mathbb{R}^3$ bezeichnet man*

$$\text{rot } v := \left(\frac{\partial v_3}{\partial x_2} - \frac{\partial v_2}{\partial x_3}, \frac{\partial v_1}{\partial x_3} - \frac{\partial v_3}{\partial x_1}, \frac{\partial v_2}{\partial x_1} - \frac{\partial v_1}{\partial x_2} \right) \tag{3.5}$$

als Rotation *von v.*

Bemerkung 8. *Die Rotation von v lässt sich auch als Vektorprodukt*

$$\text{rot } v = \nabla \times v$$

schreiben.

Definition 15. *Sei $U \to \mathbb{R}^n$ offen und $f : U \to \mathbb{R}$ zweimal stetig partiell differenzierbar. Dann ist der* Laplace-Operator *definiert als*

$$\Delta f := \text{divgrad } f = \text{div} \nabla f = \frac{\partial^2 f}{\partial x_1^2} + \ldots + \frac{\partial^2 f}{\partial x_n^2} \tag{3.6}$$

$$\Delta := \frac{\partial^2}{\partial x_1^2} + \ldots + \frac{\partial^2}{\partial x_n^2} = \sum_{i=1}^{n} \frac{\partial^2}{\partial x_i^2} \tag{3.7}$$

3.2 Beispiele partieller Differentialgleichungen

Im vorigen Kapitel wurden gewöhnliche Differentialgleichungen vom Typ

$$\frac{\mathrm{d}u}{\mathrm{d}t} = f(t, u)$$

betrachtet. Die numerische Behandlung dieser Gleichung führte zu unterschiedlichen Zeitschrittverfahren, d.h. der Operator $\frac{\mathrm{d}}{\mathrm{d}t}$ wurde diskret behandelt, wobei $f(t, u)$ eine kontinuierliche Funktion war. Im jetzigen Kontext ist f jedoch keine explizite Funktion, sondern nur über ihre Ortsableitungen definiert. Dies verlangt nach einer diskreten Beschreibung im Ort und führt somit zu Gleichungen mit Differentialoperatoren in Zeit und Ort, also partiellen Differentialgleichungen.

3.2.1 Die Diffusionsgleichung

Sei $c(x,t)$ eine Funktion in Raum und Zeit, welche die Konzentrationsverteilung einer Spezies, z.B. Kalziumionen in einer Nervenzelle, oder die Wärmeverteilung in einem Wärmeleiter beschreibt.

Ficksches Gesetz: Die Flussdichte F von $c(x,t)$ ist proportional zum negativen Gradienten der Konzentration:

$$F = -D \cdot \text{grad } c, \tag{3.8}$$

mit Diffusionskoeffizient D.

Durch die Forderung der *Massenerhaltung* und Anwendung des *gaußschen Integralsatzes* erhält man die Diffusionsgleichung:

Massenerhaltung. Sei V ein beliebiges Volumenelement mit hinreichend glattem Rand und c die Konzentration in V. Die Änderung von c in V ist beschrieben durch

$$\int_V \frac{\partial c}{\partial t} d\vec{x}.$$

Unter der Voraussetzung, dass keine Quellen oder Senken in V enthalten sind, gilt:

$$-\int_{\partial V} F \cdot \vec{n} dS = \int_V \frac{\partial c}{\partial t} d\vec{x}. \tag{3.9}$$

Satz 5. *(Integralsatz von Gauß). Für ein stetig differenzierbares Vektorfeld $F : \mathbb{R}^n \to \mathbb{R}^n$ und eine kompakte Menge $V \subset \mathbb{R}^n$ mit hinreichend glattem Rand gilt*

$$\int_V \text{div } F(\vec{x})\, d\vec{x} = \int_{\partial V} F(\vec{x}) \cdot \vec{n}\, dS. \tag{3.10}$$

Daraus folgt

$$-\int_V \text{div } F(\vec{x})\, d\vec{x} = -\int_{\partial V} F \cdot \vec{n}\, dS = \int_V \frac{\partial c}{\partial t}\, d\vec{x} \qquad \forall V \tag{3.11}$$

$$\Rightarrow -\text{div } F = \frac{\partial c}{\partial t} \tag{3.12}$$

$$\Rightarrow \frac{\partial c}{\partial t} = -\text{div } (D \nabla c) \quad \text{(Diffusionsgleichung)}. \tag{3.13}$$

3.2.2 Die Wellengleichung

Wir betrachten für die Wellengleichung die Größen *Geschwindigkeit* v, *Dichte* ρ und *Druck* p.

1. Es gilt
$$\frac{\partial \rho}{\partial t} = -\rho_0 \text{div } v \tag{3.14}$$
wobei ρ_0 eine feste Dichte definiert. Die Herleitung obiger Gleichung geschieht analog zur Diffusionsgleichung, also über die Annahme der Massenerhaltung, gefolgt von der Anwendung des gaußschen Integralsatzes.

2. **Newton'sches Gesetz:** Es gilt
$$\rho_0 \frac{\partial v}{\partial t} = -\text{grad } p. \tag{3.15}$$
Das bedeutet, dass eine räumliche Änderung des Druckfeldes eine Beschleunigung bewirkt.

3. **Zustandsgleichung:** Der Druck p ist bei konstanter Temperatur proportional zur Dichte
$$\Rightarrow p = c^2 \cdot \rho \tag{3.16}$$

$$\Rightarrow \frac{\partial^2}{\partial t^2}\rho \quad = -\rho_0 \text{div } \left(\tfrac{\partial v}{\partial t}\right) = \quad -\text{div } \left(\rho_0 \frac{\partial v}{\partial t}\right) \tag{3.17}$$
$$\Rightarrow \frac{1}{c^2}\frac{\partial^2}{\partial t^2} p \qquad\qquad = \qquad\qquad \text{div (grad } p) \tag{3.18}$$

$$\Leftrightarrow \frac{\partial^2}{\partial t^2} p = c^2 \cdot \text{div (grad } p) = c^2 \Delta p \tag{3.19}$$

Beispiel 22. *Wellengleichung in $\mathbb{R}^1$ und $\mathbb{R}^2$.*

1. *Im $\mathbb{R}^1$ beschreibt die Wellengleichung eine schwingende Saite:*
$$u_{tt} = u_{xx}.$$

2. *In $\mathbb{R}^2$ beschreibt die Wellengleichung eine schwingende Membran:*
$$u_{tt} = c^2 \Delta u.$$

3.2.3 Poisson- und Potential-Gleichung

Sei $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$ und $\rho : \Omega \to \mathbb{R}$ eine bekannte Ladungsdichteverteilung in Ω. Für das elektrische Potential Φ gilt
$$-\Delta \Phi = \rho \quad \text{in } \Omega. \tag{3.20}$$

Ein Spezialfall der Poissongleichung ist die *Potentialgleichung*

$$\Delta u = 0 \quad \text{in } \Omega \subset \mathbb{R}^d. \tag{3.21}$$

Sie beschreibt das elektrische Potential in einem ladungsfreien Raum.

Beispiel 23. *Lösung der Potentialgleichung auf einer Kreisscheibe mit Radius 1. Betrachte* $\Omega := \{(x, y) \in \mathbb{R}^2; x^2+y^2 < 1\}$ *und transformiere* x *und* y *in Polarkoordinaten:*

$$\begin{aligned} x &:= r \cdot \cos\phi \\ y &:= r \cdot \sin\phi \end{aligned}$$

In Polarkoordinaten hat der Laplace-Operator die Form:

$$\Delta u = \frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} + \frac{1}{r^2}\frac{\partial^2 u}{\partial \phi^2}. \tag{3.22}$$

Dann erfüllen $r^k \cos(k\phi)$ *und* $r^k \sin(k\phi)$ *die Potentialgleichung.*
Zu wählen ist noch eine Randbedingung *für Radius* $r = 1$*:*

$$u|_{\partial\Omega} = u(\cos\phi, \sin\phi) = a_0 + \sum_{k=1}^{\infty} \left(a_k \cos(k\phi) + b_k \sin(k\phi)\right). \tag{3.23}$$

Dann lautet die Lösung im Inneren ($r < 1$)*:*

$$u(x, y) = a_0 \sum_{k=1}^{\infty} r^k \cdot \left(a_k \cos(k\phi) + b_k \sin(k\phi)\right). \tag{3.24}$$

3.2.4 Poisson-Nernst-Planck Gleichungen

Als ein Beispiel für *gekoppelte* und *nichtlineare* PDEs können die Poisson-Nernst-Planck Gleichungen erwähnt werden. Diese Gleichungen beschreiben den Prozess der *Elektrodiffusion*, also die Diffusion von Ionenkonzentrationen c_i gekoppelt mit einem Konvektionsterm, der durch das zu berechnende elektrische Potential Φ bestimmt ist:

$$\frac{\partial c_i}{\partial t} = \nabla \cdot \left(D_i \nabla c_i + D_i \frac{z_i F}{RT} c_i \nabla \Phi \right) \tag{3.25}$$

$$-\nabla(\varepsilon_r \varepsilon_0 \nabla \Phi) = \rho_f + \sum_i z_i F c_i \tag{3.26}$$

Dabei sind D_i die je zu c_i gehörigen Diffusionstensoren, z_i die Teilchenladungen, ρ_f eine statische Ladungsdichte, ε_r und ε_0 die relative bzw. die Vakuum-Permettivität, F ist die Faraday-Konstante, R die universelle Gaskonstante und T die Temperatur.

4 Diskretisierung I: Differenzenverfahren für partielle Differentialgleichungen

In diesem Kapitel werden wir ein Approximationsverfahren für das kontinuierliche PDE-Problem herleiten, welches auf der Approximation der Ortsableitungen fundiert. Das als *Differenzenverfahren* bezeichnete Diskretisierungsverfahren wird für ein- und zweidimensionale Fälle hergeleitet. Weiter werden wir Eigenschaften der Systemmatrix des hergeleiteten Gleichungssystems analysieren, die uns am Ende des Kapitels zu einem Konvergenzbeweis für das Differenzenverfahren führen. Die sichtbar werdenden Vor- und Nachteile dieses Verfahrens leiten über in das folgende Kapitel alternativer und allgemeinerer Diskretisierungsverfahren.

Betrachte die Poissongleichung

$$-\Delta u = f \quad \text{auf } \Omega \tag{4.1}$$

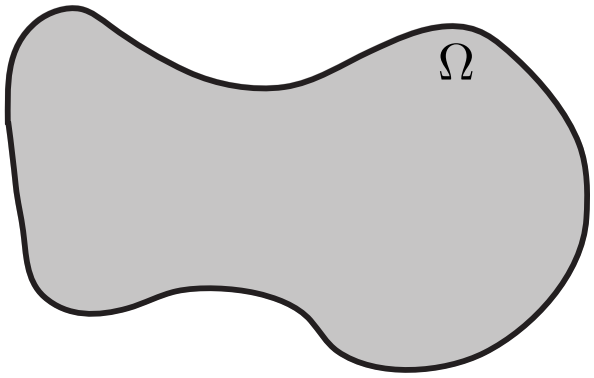

Abbildung 4.1: Kontinuierliches Rechengebiet Ω

Die Poissongleichung ist zunächst auf einem kontinuierlichen Gebiet Ω definiert, d.h. an unendlich vielen Punkten. Dies ist für ein numerisches Verfahren nicht zugänglich.

Idee 1. *Wähle endlich viele Punkte in Ω aus, in denen $-\Delta u = f$ erfüllt ist. Dies führt zu einer Approximation des kontinuierlichen Gebiets sowie zu der kontinuierlichen Gleichung.*

4.1 Gebietsdiskretisierung

Betrachte beispielsweise das Einheitsquadrat als kontinuierliches Gebiet

$$\Omega = \{(x, y) : 0 < x < 1, 0 < y < 1\}. \tag{4.2}$$

Das Vorgehen zur Approximation des kontinuierlichen Gebiets, d.h. *Diskretisierung* von Ω ist das Folgende:

1. Überziehe Ω mit einem gleichmäßigen Gitter. Die Menge der Gitterknoten wird als Ω_h bezeichnet. Eine feste Schrittweite h liefert

$$\Omega_h = \left\{(x, y) \in \Omega : \frac{x}{h}, \frac{y}{h} \in \mathbb{Z}\right\}.$$

2. Erfülle die Differentialgleichung in jedem Punkt aus Ω_h.

   - Ersetze $u(x)$ durch $u_h(x)$. Dabei ist $u(x)$ die kontinuierliche und $u_h(x)$ die diskrete Lösung.
   - Approximiere die Ableitungen:

$$\lim_{h \to 0} \frac{u(x+h) - u(x)}{h} \approx \frac{u(x+h) - u(x)}{h}$$

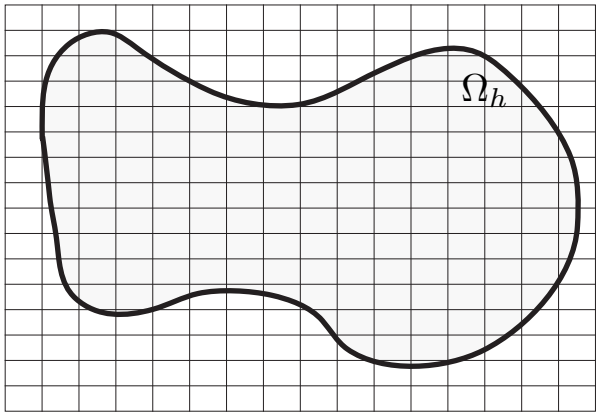

Abbildung 4.2: Diskretisiertes Gebiet Ω_h. Dieses entsteht durch Überziehen des kontinuierlichen Gebiets Ω mit einem Gitter. Die Gitterknoten definieren das endliche diskretisierte Gebiet.

4.2 Approximationseigenschaften im $\mathbb{R}^1$

Betrachte das eindimensionale Problem

$$\begin{aligned} u''(x) &= f(x) \quad \text{in } \Omega = (0,1), \\ u(0) &= \varphi_0, \\ u(1) &= \varphi_1. \end{aligned}$$

Die Approximation der Ableitung kann auf verschiedene Arten geschehen:

1. **rechtsseitig:** $\delta^+ u(x) = \frac{u(x+h)-u(x)}{h}$
2. **linksseitig:** $\delta^- u(x) = \frac{u(x)-u(x-h)}{h}$
3. **symmetrisch (zentral):** $\delta^0 u(x) = \frac{u(x+h)-u(x-h)}{2h}$

Für die zweite Ableitung können links- und rechtsseitige Differenzen δ^+ und δ^- kombiniert werden:

$$u''(x) \approx \delta^+\delta^- u(x) = \frac{\frac{u(x+h)-u(x)}{h} - \frac{u(x)-u(x-h)}{h}}{h} = \frac{u(x+h) - 2u(x) + u(x-h)}{h^2}. \quad (4.3)$$

Lemma 2. *Sei $[x-h, x+h] \subset \bar{\Omega}$. Es gilt*

(i) $\delta^\pm u(x) = u'(x) + hR$ *mit* $|R| \leq \frac{1}{2}\|u''\|_\infty$,

(ii) $\delta^0 u(x) = u'(x) + h^2 R$ *mit* $|R| \leq \frac{1}{6}\|u'''\|_\infty$,

(iii) $\delta^+\delta^- u(x) = u''(x) + h^2 R$ *mit* $|R| \leq \frac{1}{12}\|u^{(4)}\|_\infty$.

Beweis. Zu (i):

$$\begin{aligned} u(x \pm h) &= u(x) \pm hu'(x) + \frac{h^2}{2}u''(x) + \dots \\ &= u(x) \pm hu'(x) + \frac{h^2}{2}u''(\xi), \text{ mit } x \leq \xi \leq x+h \\ \Leftrightarrow \frac{u(x+h)-u(x)}{h} &= u'(x) + \frac{h}{2}u''(\xi) \\ &\leq u'(x) + \frac{h}{2}\|u''\|_\infty \end{aligned}$$

Zu (ii): Es gelten die Taylorentwicklungen um $x \pm h$:

$$\begin{aligned} u(x+h) &= u(x) + hu'(x) + \frac{h^2}{2}u''(x) + \frac{h^3}{6}u'''(\xi), \\ u(x-h) &= u(x) - hu'(x) + \frac{h^2}{2}u''(x) - \frac{h^3}{6}u'''(\tilde{\xi}). \end{aligned}$$

Subtraktion ergibt:

$$\begin{aligned}\delta^0 u(x) &= \frac{2hu'(x) + \frac{h^3}{6}(u'''(\xi) + u'''(\tilde{\xi}))}{2h} \\ &= u'(x) + \frac{h^2}{12}(u'''(\xi) + u'''(\tilde{\xi}) \\ &\leq u'(x) + \frac{h^2}{12} \cdot 2\|u'''\|_\infty = u'(x) + \frac{h^2}{6}\|u'''\|_\infty\end{aligned}$$

Zu (iii): Betrachte die Entwicklung um $x \pm h$ bis zur Ordnung 4:

$$\begin{aligned}u(x+h) &= u(x) + hu'(x) + \frac{h^2}{2}u''(x) + \frac{h^3}{6}u'''(x) + \frac{h^4}{24}u^{(4)}(\xi) \\ u(x-h) &= u(x) - hu'(x) + \frac{h^2}{2}u''(x) - \frac{h^3}{6}u'''(x) + \frac{h^4}{24}u^{(4)}(\tilde{\xi})\end{aligned}$$

Addition der obigen Gleichungen, Subtraktion von $2u(x)$ und Division durch h^2 liefert

$$\begin{aligned}\delta^+\delta^- u(x) &= \frac{h^2 u''(x) + \frac{h^4}{24}\left(u^{(4)}(\xi) + u^{(4)}(\tilde{\xi})\right)}{h^2} \\ \Rightarrow \delta^+\delta^- u(x) &= u''(x) + \frac{h^2}{24}\left(u^{(4)}(\xi) + u^{(4)}(\tilde{\xi})\right) \\ &\leq u''(x) + \frac{h^2}{12}\|u^{(4)}\|_\infty\end{aligned}$$

□

4.3 Erstellen eines linearen Gleichungssystems

Durch die Approximation der Ableitungen auf einem diskreten Gebiet entsteht ein endliches lineares Gleichungssystem, dass es schlussendlich zu lösen gilt. Wir betrachten weiter die Poisson-Gleichung

$$u''(x) = \Delta u(x) = f(x),$$

und approximieren $\Delta \approx \Delta_h = \delta^+\delta^-$. Wir erhalten also eine diskretisierte Gleichung

$$\delta^+\delta^- u(x) = f(x) + \mathcal{O}(h^2)$$

In Matrix-Vektor-Schreibweise lässt sich obige Gleichung zu dem Beispiel darstellen als

$$\delta^+\delta^- u(x) = \frac{1}{h^2}\begin{pmatrix} -2 & 1 & 0 \\ 1 & -2 & 1 \\ 0 & 1 & -2 \end{pmatrix} \cdot \begin{pmatrix} u_h(x_1) \\ u_h(x_2) \\ u_h(x_3) \end{pmatrix} = \begin{pmatrix} f(x_1) - \frac{1}{h^2}u(0) \\ f(x_2) \\ f(x_3) - \frac{1}{h^2}u(1) \end{pmatrix}.$$

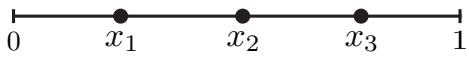

Abbildung 4.3: Beispieldiskretierung eines eindimensionalen Gebiets mit drei inneren Knoten

Im allgemeinen Fall erhalten wir

$$\frac{1}{h^2} \cdot \begin{pmatrix} -2 & 1 & 0 & \dots & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & \dots & 0 \\ & \ddots & & & \ddots & & \\ & & \ddots & & & \ddots & \\ 0 & 0 & \dots & 0 & 1 & -2 & 1 \\ 0 & 0 & \dots & 0 & 0 & 1 & -2 \end{pmatrix} \begin{pmatrix} u_1 \\ \\ \\ \vdots \\ \\ \\ u_n \end{pmatrix} = \begin{pmatrix} f_1 - \frac{u_0}{h^2} \\ f_2 \\ \\ \vdots \\ \\ f_{n-1} \\ f_n - \frac{u_n}{h^2} \end{pmatrix},$$

also ein Gleichungssystem von der Form

$$K_h \cdot u_h = f_h$$

Bemerkung 9. *Die Matrix K_h ist* dünn besetzt*, d.h.*

$$\#\{(i,j) \in \{1,\dots,n\}^2 : \ K_{i,j} \neq 0\} = \mathcal{O}(n).$$

4.4 Finite Differenzen in $\mathbb{R}^2$

Wir betrachten nun ein zweidimensionales Gebiet

$$\Omega := \left\{(x,y) \in \mathbb{R}^2 : \ 0 < x < 1, 0 < y < 1\right\}$$

und dessen Gebietsdiskretisierung

$$\Omega_h := \left\{(x,y) \in \Omega : \ \frac{x}{h}, \frac{y}{h} \in \mathbb{Z}\right\}$$

mit den Gebietsrändern

$$\begin{aligned} \Gamma &:= \left\{(x,y) \in \mathbb{R}^2 : \ x \in \{0,1\}, y \in \{0,1\}\right\}, \\ \Gamma_h &:= \left\{(x,y) \in \Gamma : \ \frac{x}{h}, \frac{y}{h} \in \mathbb{Z}\right\}. \end{aligned}$$

Wir betrachten nun das Randwertproblem

$$\begin{aligned} -\Delta u &= f \text{ in } \Omega, \\ u &= \varphi \text{ auf } \Gamma \end{aligned}$$

mit der diskreten Form

$$-\Delta_h u_h := (-\delta_x^- \delta_x^+ - \delta_y^- \delta_y^+) u_h(x). \tag{4.4}$$

Definition 16. *Mit u_h wird die* Gitterfunktion *von u bezeichnet, also die Reduktion von u auf das Gitter Ω_h.*

Wendet man den diskreten Laplace-Operator Δ_h auf die Gitterfunktion u_h an, so erhält man

$$\begin{aligned} -\Delta_h u_h &= (-\delta_x^- \delta_x^+ - \delta_y^- \delta_y^+) u_h(x) \\ &= -\frac{1}{h^2} \left(u_h(x-h,y) + u_h(x+h,y) + u_h(x,y-h) + u_h(x,y+h) - 4u_h(x,y) \right). \end{aligned}$$

Die Gitterfunktion u_h wird also an 5 Gitterpunkten ausgewertet. Deshalb wird obige Darstellung auch als *Fünfpunktformel* bezeichnet.

4.4.1 Matrix-Vektor-Schreibweise in $\mathbb{R}^2$

Im eindimensionalen Fall existiert eine natürliche Nummerierung der Knoten, welche die Matrixstruktur festlegt (diese Ordnung der Knoten wurde im vorigen Abschnitt stillschweigend verwendet). In $\mathbb{R}^2$ sind unterschiedliche Ordnungen der Knoten denkbar.

4.4.1.1 Lexikographische Knotennummerierung

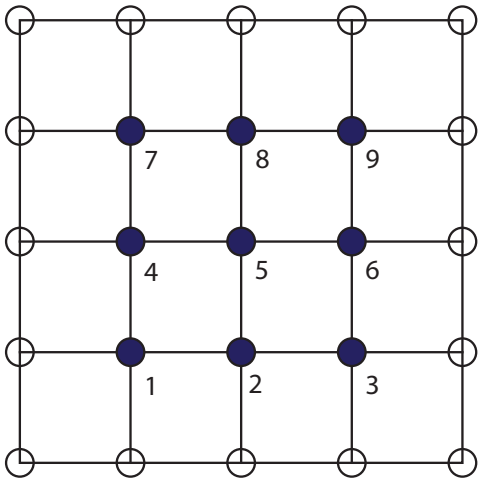

Abbildung 4.4: Lexikographische Nummerierung der inneren Knoten.

Die lexikographische Knotennummerierung ist eine zeilenweise Nummerierung der Knoten und definiert eine Matrix der Form

$$\frac{1}{h^2}\begin{pmatrix} 4 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 4 & -1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 4 & -1 & 0 & -1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 4 & -1 & 0 & -1 & 0 & 0 \\ 0 & -1 & 0 & -1 & 4 & -1 & 0 & -1 & 0 \\ 0 & 0 & -1 & 0 & -1 & 4 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 & 0 & 0 & 4 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & -1 & 4 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 4 \end{pmatrix} \cdot u_h = \tilde{f}_h \tag{4.5}$$

wobei mit $\tilde{f}_h$ die rechte Seite f versehen mit den Einträgen, die aus der Randbedingung entstehen, bezeichnet wird.
Obige Matrix hat eine Blocktridiagonalstruktur der Form

$$K_h = \frac{1}{h^2}\begin{pmatrix} D & -I & 0 & 0 \\ -I & D & -I & 0 \\ & & \ddots & \\ 0 & 0 & -I & D \end{pmatrix} \tag{4.6}$$

mit

$$D = \begin{pmatrix} 4 & -1 & 0 & \cdots & 0 \\ -1 & 4 & -1 & 0 & \cdots \\ & -1 & \ddots & \ddots & \\ & & \ddots & \ddots & -1 \\ 0 & \cdots & 0 & -1 & 4 \end{pmatrix},$$

$$-I = \begin{pmatrix} -1 & 0 & 0 & \cdots & 0 \\ 0 & -1 & 0 & \cdots & 0 \\ & & \ddots & & \\ & & & \ddots & \\ 0 & \cdots & 0 & 0 & -1 \end{pmatrix}.$$

4.4.1.2 Schachbrettnummerierung

Die Schachbrettnummerierung nummeriert die Knoten in der Schwarz/Weiß-Abfolge eines Schachbretts.
Daraus entsteht folgende Matrixstruktur

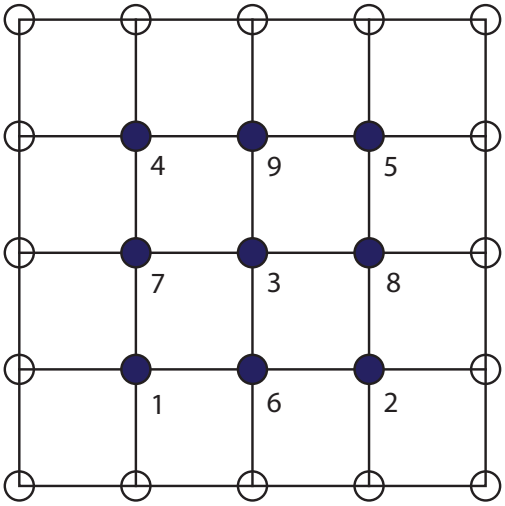

Abbildung 4.5: Schachbrettnummerierung der inneren Knoten

$$\frac{1}{h^2}\begin{pmatrix} 4 & 0 & 0 & 0 & 0 & -1 & -1 & 0 & 0 \\ 0 & 4 & 0 & 0 & 0 & -1 & 0 & -1 & 0 \\ 0 & 0 & 4 & 0 & 0 & -1 & -1 & -1 & -1 \\ 0 & 0 & 0 & 4 & 0 & 0 & -1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 4 & 0 & 0 & -1 & -1 \\ -1 & -1 & -1 & 0 & 0 & 4 & 0 & 0 & 0 \\ -1 & 0 & -1 & -1 & 0 & 0 & 4 & 0 & 0 \\ 0 & -1 & -1 & 0 & -1 & 0 & 0 & 4 & 0 \\ 0 & 0 & -1 & -1 & -1 & 0 & 0 & 0 & 4 \end{pmatrix} \cdot u_h = \tilde{f}_h \tag{4.7}$$

Dies definiert eine Matrix mit der Struktur

$$K_h = \begin{pmatrix} D_1 & L \\ L^T & D_2 \end{pmatrix} \tag{4.8}$$

mit den aus K_h ersichtlichen Submatrizen D_i, L und L^T.

Bemerkung 10. *Die Knotennummerierung ändert nichts an den algebraischen Eigenschaften des Systems, kann aber technische Aspekte bei der Implementierung beeinflussen und Das Konvergenzverhalten bestimmter iterativer Löser beeinflussen.*

4.4.2 Sternoperatoren

Die Fünfpunktformel definiert die Rechenvorschrift zur Approximation des Laplace-Operators in $\mathbb{R}^2$ über finite Differenzen. Eine vereinfachte Schreibweise hierfür ist die Definition eines Sternoperators.

Definition 17. *Die Fünfpunktformel wird über einen* Fünfpunktstern *folgendermaßen definiert*

$$\begin{aligned} -\Delta_h u_h &= \frac{1}{h^2}\left(-u(x-h,y)-u(x+h,y)-u(x,y-h)-u(x,y+h)+4u(x,y)\right) \\ &=: \frac{1}{h^2}\begin{bmatrix} & -1 & \\ -1 & 4 & -1 \\ & -1 & \end{bmatrix} = -\Delta_h. \end{aligned}$$

Bemerkung 11. *Der Fünfpunktstern ist keine Matrix, sondern lediglich eine Schablone, die auf Ω_h aufgelegt die Rechenvorschrift der Fünfpunktformel vorgibt und damit am Ende auch die Matrixstruktur von K_h. Die Nummerierung der Knoten geht nicht in den Fünfpunktstern ein, da er direkt auf das Gitter aufgelegt wird.*

4.4.2.1 Weitere Sternoperatoren und Rechenvorschriften

1. in $\mathbb{R}^1$
   a) $\delta^+ = \frac{1}{h} \cdot \begin{bmatrix} 0 & -1 & 1 \end{bmatrix}$
   b) $\delta^- = \frac{1}{h} \cdot \begin{bmatrix} -1 & 1 & 0 \end{bmatrix}$
   c) $\delta^0 = \frac{1}{2h} \cdot \begin{bmatrix} -1 & 0 & 1 \end{bmatrix}$
2. **Allgemeiner Differenzenstern:**

$$\frac{1}{h^k}\begin{bmatrix} & \vdots & & & \\ & c_{-1,1} & c_{0,1} & c_{1,1} & \\ \cdots & c_{-1,0} & c_{0,0} & c_{1,0} & \cdots \\ & c_{-1,-1} & c_{0,-1} & c_{1,-1} & \\ & \vdots & & & \end{bmatrix}(x,y)$$

$$= \frac{1}{h^k} \cdot \sum_{i,j} c_{ij} u_h(x+ih, y+jh)$$

3. **Multiplikation von Sternen:** Am Beispiel von zwei eindimensionalen Sternen wird die Verkettung (Faltung) von Sternen demonstriert.

$$\begin{aligned} \begin{bmatrix} a & \underline{b} & c \end{bmatrix}\begin{bmatrix} d & \underline{e} & f \end{bmatrix} u_h &= \begin{bmatrix} a & \underline{b} & c \end{bmatrix} \cdot (d \cdot u_h(x-h) + e \cdot u_h(x) + f \cdot u_h(x+h)) \\ &= a(du_h(x-2h) + eu_h(x-h) + fu_h(x)) \\ &\quad + b(du_h(x-h) + eu_h(x) + fu_h(x+h)) \\ &\quad + c(du_h(x) + eu_h(x+h) + fu_h(x+2h)) \\ &= \begin{bmatrix} ad & ae+bd & \underline{af+be+cd} & bf+ce & cf \end{bmatrix}. \end{aligned}$$

4.4.3 Eigenschaften von Differenzensternen

Wir wollen einige typische (aber nicht allgemein gültige) Eigenschaften von Differenzensternen zusammenstellen. Basierend darauf definieren wir danach eine Klasse von Matrizen, die abschließend wichtige Eigenschaften der Systemmatrix der diskretisierten Modellgleichung liefern wird.
Folgende gängige Eigenschaften von Differenzensternen, bzw. der Systemmatrix, werden betrachtet:

1. Die **Zeilensumme** ist Null, d.h.

$$\sum_{j=1}^{n} a_{ij} = 0 \quad \forall i = 1 \dots n.$$

2. **Vorzeichenmuster:** Für Δ_h gilt

$$a_{ii} > 0,\, a_{ij} \leq 0 \quad (i \neq j).$$

   Man beachte: Das Vorzeichenmuster gilt nicht immer, z.B. für Δ_h^2 ist das Muster nicht erfüllt.

3. **Diagonaldominanz:**

   a) *schwache* Diagonaldominanz

$$|a_{ii}| \geq \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| \quad \forall i = 1 \dots n$$

   b) *starke* Diagonaldominanz

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| \quad \forall i = 1 \dots n$$

4. K_h ist *symmetrisch.*

Aus diesen Eigenschaften lässt sich eine Klasse spezieller Matrizen, der sogenannten *M-Matrizen*, aufbauen.

4.5 M-Matrizen

In diesem Abschnitt wird die Definition einer M-Matrix eingeführt. Anschließend zeigen wir, dass die Systemmatrix K_h M-Matrix-Eigenschaften besitzt. Diese werden bei der Analyse von K_h nützlich sein, um abschließend wichtige Matrixeigenschaften zu spezifizieren.

4.5.1 Wiederholung von speziellen Matrixeigenschaften

Definition 18. *Eine Matrix $A \in \mathbb{K}^{n\times n}$ heißt* invertierbar*, wenn es eine Matrix $\tilde{A} \in \mathbb{K}^{n\times n}$ gibt mit*

$$A \cdot \tilde{A} = \tilde{A} \cdot A = E_n,$$

wobei E_n die Einheitsmatrix ist.

Definition 19. *Eine Matrix $A \in \mathbb{K}^{n\times n}$ heißt* regulär *(oder* nicht-singulär*), wenn*

$$\sum_{j=1}^{n} \lambda_j A_j = 0 \;\Leftrightarrow\; \lambda_j = 0 \;\forall j = 1 \dots n,$$

wobei A_j die Spaltenvektoren von A sind.

Lemma 3. *Eine reguläre Matrix $A \in \mathbb{K}^{n\times n}$ ist invertierbar und lässt sich als endliches Produkt von Elementarmatrizen darstellen.*

Lemma 4. *Für $A \in \mathbb{K}^{n\times n}$ sind folgende Bedingungen äquivalent:*

- *A ist invertierbar,*
- *A^T ist invertierbar,*
- *Spaltenrang von A ist n, Zeilenrang von A ist n.*

In jedem dieser Fälle gilt: $(A^T)^{-1} = (A^{-1})^T$.

4.5.2 Eigenschaften von M-Matrizen

Im Folgenden werden Matrizen $A, B \in \mathbb{K}^{n\times n}$ mit Einträgen $(a_{ij})_{i,j=1\dots n}$ bzw. $(b_{ij})_{i,j=1\dots n}$ betrachtet.

Definition 20. *Die Relation $A \geq B$ wird komponentenweise definiert, d.h. durch $a_{ij} \geq b_{ij}, \quad \forall i, j = 1 \dots n$. Analog lässt sich $A \leq B$, $A < B$ und $A > B$ definieren.*

Definition 21. *Eine $n \times n$-Matrix heißt* M-Matrix*, wenn folgende Eigenschaften erfüllt sind:*

1. *Vorzeichenbedingung: $a_{ii} > 0$, $a_{ij} \leq 0 \quad \forall i, j = 1 \dots n$, $i \neq j$,*
2. *A ist regulär und $A^{-1} \geq 0$.*

Frage 1. *Gelten für die Matrix K_h aus dem Modellproblem*

$$\begin{aligned} -\Delta u &= f, \\ K_h u_h &= f_h \end{aligned} \tag{4.9}$$

die M-Matrix Eigenschaften?

Die Vorzeichenbedingung lässt sich direkt an K_h ablesen. Zu zeigen bleibt, dass A regulär ist und $A^{-1} \geq 0$.

Definition 22. *Sei $A \in \mathbb{K}^{n\times n}$ eine Matrix und $i, j \in \{1, \ldots, n\}$ Indices.*

- *Index i ist mit Index j* direkt verbunden*, wenn $a_{ij} \neq 0$.*
- *Index i ist mit Index j* verbunden*, wenn eine Kette von Indizes $(i_k)_{k=1,\ldots,p} \subset \{1, \ldots, n\}$ derart existiert, dass $i = i_1$, $i_p = j$ und i_{k-1} mit i_k verbunden ist für alle $k = 2 \ldots n$.*
- *Eine Matrix A heißt* irreduzibel*, falls jeder Index $i \in \{1, \ldots, n\}$ mit jedem Index $j \in \{1, \ldots, n\}$* verbunden *ist.*
  Eine alternative Definition ist:
  A irreduzibel $\Leftrightarrow$ es existiert keine Permutation Π derart, dass

$$\Pi^T A \Pi = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}.$$

4.5.3 Abschätzen der Eigenwertbereiche einer Matrix

Wir werden das Kriterium von Gerschgorin in zwei Fassungen beweisen. Dieses Kriterium wird uns Auskunft über die Eigenwertbereiche geben.

4.5.3.1 Kriterium von Gerschgorin

Gegeben sei eine Matrix $A = (a_{ij}) \in \mathbb{C}^{n\times n}$.
Wir betrachten die abgeschlossenen Kreisscheiben

$$\overline{K}_i := \left\{ z \in \mathbb{C} : \ |z - a_{ii}| \leq \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \right\}, \quad i = 1 \ldots n$$

mit Zentrum im i-ten Diagonaleintrag und einem Radius, der durch die Summe der Beträge aller anderen Einträge der Zeile i gegeben ist.

Satz 6. *(Gerschgorin).*
Die n Eigenwerte von A liegen in der Vereinigung

$$\bigcup_{i=1}^{n} \overline{K}_i.$$

Beweis. Sei v Eigenvektor von A zum Eigenwert λ.
OBdA gilt $\|v\|_\infty := \max_{j=1}^{n} |v_j| = 1$ und wir nehmen an, dass das Maximum in der i-ten Zeile angenommen wird, also $|v_i| = 1$. Es gilt

$$(A - \lambda E_n)v = 0.$$

Für die i-te Zeile folgt

$$(a_{ii} - \lambda)v_i = -\sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}v_j.$$

Anwendung der Dreiecksungleichung ergibt

$$\begin{aligned}
|a_{ii} - \lambda| &= |(a_{ii} - \lambda)v_i| = \left|\sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}v_j\right| \\
&\leq \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}||v_j| \leq \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \\
\Rightarrow \qquad & |a_{ii} - \lambda| \leq \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \\
\Leftrightarrow \qquad & \lambda \in \overline{K}_i = \left\{ z \in \mathbb{C} : \ |z - a_{ii}| \leq \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \right\}.
\end{aligned}$$

□

Damit ist gezeigt, dass jeder Eigenwert in der Vereinigung aller Kreisscheiben $\overline{K}_i$ liegt. Wendet man das Kriterium von Gerschgorin auf die Systemmatrix K_h an, dann folgt:

1. $\sum_{j=1,j\neq i}^{n} |a_{ij}| = 4h^{-2}\, \forall j \ \Rightarrow$ Radius für alle Kreise ist $4h^{-2}$.
2. $a_{ii} = 4h^{-2}$ Zentrum aller Kreise ist $4h^{-2}$.

Zusammen ergibt sich, dass $\lambda \in \left[0, 8h^{-2}\right]$.

Im nächsten Schritt wollen wir zeigen, dass die Eigenwerte von K_h echt größer Null sind, d.h. wir wollen zeigen

$$\lambda \in \left(0, 8h^{-2}\right).$$

Dazu beweisen wir die

Satz 7. *(Verschärfung des Kriteriums von Gerschgorin).*
Unter der Voraussetzung, dass A irreduzibel ist, gilt:

$$\lambda \in \left(\bigcup_{i=1}^{n} K_i\right) \cup \left(\bigcap_{i=1}^{n} \partial K_i\right)$$

mit

$$\begin{aligned} K_i &:= \left\{ z \in \mathbb{C} : \ |z - a_{ii}| < \sum_{j=1, j\neq i}^{n} |a_{ij}| \right\}, \\ \partial K_i &:= \left\{ z \in \mathbb{C} : \ |z - a_{ii}| = \sum_{j=1, j\neq i}^{n} |a_{ij}| \right\}. \end{aligned} \tag{4.10}$$

Beweis. Es sei λ ein Eigenwert von A mit zugehörigen Eigenvektor v mit $\|v\|_\infty = 1$. Im Beweis zum (unverschärften) Gerschgorin-Kriterium wurde die Existenz eines $i \in \{1, \ldots, n\}$ mit $|v_i| = 1$ und $|\lambda - a_{ii}| \leq \sum_{j=1, j\neq i}^{n} |a_{ij}|$ bewiesen. Im Fall $|\lambda - a_{ii}| < \sum_{j=1, j\neq i}^{n} |a_{ij}|$ ist nichts zu zeigen, da dann bereits $\lambda \in K_i$ gilt. Dasselbe gilt analog für alle anderen Indices k mit $|v_k| = 1$. Sei also im Folgenden

$$|\lambda - a_{kk}| = \sum_{j=1, j\neq k}^{n} |a_{kj}| \ \forall k \in \{k \in \{1, \ldots, n\} : \ |v_k| = 1\}. \tag{4.11}$$

Man zeigt, dass dann bereits $|\lambda - a_{jj}| = \sum_{k=1, k\neq j}^{n} |a_{jk}| \ \forall j = 1 \ldots n$ und also $\lambda \in \bigcap_{i=1}^{n} \partial K_i$ gilt.
Sei also $j \in \{1, \ldots, n\} \setminus \{i\}$. Nach Voraussetzung über die Irreduzibilität existiert dann eine Kette von Verbindungen

$$i = i_0, i_1, \ldots, i_l = j \text{ mit } a_{i_{p-1} i_p} \neq 0.$$

Man zeigt, dass aus der Identität $\left|\lambda - a_{i_p i_p}\right| = \sum_{k=1, k\neq i_p}^{n} \left|a_{i_p k}\right|$ und $\left|v_{i_p}\right| = 1$ bereits folgt, dass auch $\left|\lambda - a_{i_{p+1} i_{p+1}}\right| = \sum_{k=1, k\neq i_{p+1}}^{n} \left|a_{i_{p+1} k}\right|$ und $\left|v_{i_{p+1}}\right| = 1$ gilt ($p = 0 \ldots l-1$). Induktiv folgt daraus die Behauptung.

Sei also $\left|\lambda - a_{i_p i_p}\right| = \sum_{j=1, j\neq i_p}^{n} \left|a_{i_p j}\right|$ und $\left|v_{i_p}\right| = 1$ für ein $p \in \{0, \ldots, l-1\}$. Aus der Dreiecksungleichung folgt wie im Beweis zum (unverschärften) Kriterium von Gerschgorin

$$\left|\lambda - a_{i_p i_p}\right| \leq \sum_{k=1, k\neq i_p}^{n} \left|a_{i_p k}\right| |v_k|$$

und beides zusammen liefert somit

$$\sum_{k=1, k\neq i_p}^{n} \left|a_{i_p k}\right| |v_k| \geq \sum_{k=1, k\neq i_p}^{n} \left|a_{i_p k}\right|.$$

Andererseits gilt wegen $\|v\|_\infty = 1$ auch die umgekehrte Ungleichung und damit also Gleichheit:

$$\sum_{k=1, k\neq i_p}^{n} \left|a_{i_p k}\right| |v_k| = \sum_{k=1, k\neq i_p}^{n} \left|a_{i_p k}\right|.$$

Daraus folgt wegen $|v_k| \leq 1 \ \forall k$ schon $|v_k| = 1 \ \forall k$ mit $a_{i_p k} \neq 0$, also insb. $\left|v_{i_{p+1}}\right| = 1$. Nach der Annahme in (4.11) gilt also $|\lambda - a_{i_{p+1} i_{p+1}}| = \sum_{k=1, k \neq i_{p+1}}^{n} |a_{i_{p+1} k}|$.

□

Frage 2. *Was folgt mit den Kriterien von Gerschgorin für die Matrix K_h?*

Es treten die folgenden 3 Fälle auf:

$$\begin{aligned} a_{ii} &= \frac{4}{h^2} \text{ und} \\ r_i &= \sum_{j=1, j \neq i}^{n} |a_{ij}| = \begin{cases} 2/h^2 & \text{für die Eckpunkte} \\ 3/h^2 & \text{für die Seitenpunkte} \\ 4/h^2 & \text{für die inneren Punkte} \end{cases} . \end{aligned}$$

Da K_h symmetrisch und reell ist, sind alle Eigenwerte von K_h reell. Die Kreisränder ∂K_j haben den gleichen Mittelpunkt mit unterschiedlichen Radien und sind deshalb disjunkt. Aus dem verschärften Kriterium von Gerschgorin folgt (da K_h irreduzibel ist):

$$\lambda \in \left(0, \frac{8}{h^2}\right).$$

4.5.4 Zusammenhang zwischen M-Matrix und Spektralradius

Zunächst definieren wir die Begriffe *Diagonaldominanz*, *Irreduzibilität* und den *Spektralradius*. Anschließend wird der Zusammenhang zwischen M-Matrix und Spektralradius analysiert.

Definition 23. *Eine Matrix $A \in \mathbb{K}^{n \times n}$ heißt*

- diagonaldominant*, falls*

$$\sum_{j, j \neq i} |a_{ij}| < |a_{ii}| \quad \forall i = 1, \ldots, n \tag{4.12}$$

- irreduzibel diagonaldominant*, falls A irreduzibel ist und*

$$\begin{aligned} \sum_{j, j \neq i} |a_{ij}| &< |a_{ii}| \quad \text{für ein } i & (4.13) \\ \sum_{j, j \neq i} |a_{ij}| &\leq |a_{ii}| \quad \forall i = 1, \ldots, n & (4.14) \end{aligned}$$

Bemerkung 12. *Aus Irreduzibilität und Diagonaldominanz folgt irreduzible Diagonaldominanz, die Rückrichtung gilt jedoch nicht.*

Definition 24. *Der* Spektralradius $\varrho(A)$ *einer Matrix $A \in \mathbb{K}^{n \times n}$ ist definiert durch*

$$\varrho(A) := \max\{|\lambda| : \ \lambda \text{ ist Eigenwert von } A\}. \tag{4.15}$$

Satz 8. *Folgende Aussagen gelten für die Matrix $D^{-1}B$ mit $D := \operatorname{diag}\{a_{ii} : i = 1, \ldots, n\}$ und $B := D - A$:*

(i) Ist A diagonaldominant oder irreduzibel diagonaldominant, so gilt $\varrho(D^{-1}B) < 1$.

(ii) Erfüllt A die Vorzeichenbedingung, so gilt: A ist M-Matrix $\Leftrightarrow \varrho(D^{-1}B) < 1$.

Beweis. (i) Betrachte $C := D^{-1}B$ mit

$$\begin{aligned} c_{ij} &= -\frac{a_{ij}}{a_{ii}}, \quad (i \neq j), \\ c_{ii} &= 0. \end{aligned}$$

a) Sei Diagonaldominanz vorausgesetzt. Dann gilt:

$$r_i \coloneqq \sum_{\substack{j=1 \\ j \neq i}}^{n} |c_{ij}| < 1 \quad \forall i = 1, \ldots, n.$$

Aus dem Kriterium von Gerschgorin folgt:

$$\begin{aligned} \lambda \in \bigcup_{i=1}^{n} \overline{K}_{r_i}(c_{ii}) &= \bigcup_{i=1}^{n} \overline{K}_{r_i}(0) \\ \Rightarrow |\lambda| &\leq \max_{i=1\ldots n} r_i < 1 \\ \Rightarrow \varrho(D^{-1}B) &< 1. \end{aligned}$$

b) Sei nun irreduzible Diagonaldominanz vorausgesetzt:

A irreduzibel diagonaldominant $\Rightarrow$ $\begin{array}{l} r_j \leq 1 \quad \forall j = 1, \ldots, n \\ r_i < 1 \quad \text{für mindestens ein } i \end{array}$

Aus der scharfen Version von Gerschgorin folgt

$$\lambda \in \bigcup_{j=1}^{n} K_{r_j}(0) \cup \left(\bigcap_{j=1}^{n} \partial K_{r_j}(0) \right).$$

Da es mindestens einen Index i gibt mit $r_i < 1$, also $\partial K_{r_i}(0) \subset K_1(0)$, folgt sofort $\bigcap_{j=1}^{n} \partial K_{r_j}(0) \subset K_1(0)$ und damit auch $\varrho(D^{-1}B) < 1$.

(ii) Für A gelte die Vorzeichenbedingung.
Zu zeigen ist: A ist nicht-singulär und $A^{-1} \geq 0 \;\Leftrightarrow\; \varrho(D^{-1}B) < 1$.
„$\Leftarrow$“: Sei $\varrho(D^{-1}B) = \varrho(C) < 1$. Dann konvergiert die Neumann-Reihe

$$S := \sum_{\nu=0}^{\infty} C^{\nu} = (I - C)^{-1}$$

$$\begin{aligned} \Leftrightarrow \quad & S(I - C) = I \\ \Leftrightarrow \quad & SD^{-1}(D - B) = SD^{-1}A = I \\ \Leftrightarrow \quad & A^{-1} = SD^{-1}. \end{aligned}$$

Es gilt $D^{-1} \geq 0$, $B \geq 0 \Rightarrow C \geq 0 \Rightarrow C^\nu \geq 0 \Rightarrow S \geq 0 \Rightarrow A^{-1} \geq 0$.
„⇒“: Sei A eine M-Matrix. Es sei $u \neq 0$ ein Eigenvektor von $D^{-1}B$ und λ der zugehörige Eigenwert. Mit $|u|$ sei der Vektor $(|u_i|)_{i=1}^n$ bezeichnet. Es gilt

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

Ferner gilt: $A^{-1} \geq 0$ und $D \geq 0 \Rightarrow A^{-1}D \geq 0$.

$$\begin{aligned} \Rightarrow \quad & -A^{-1}DD^{-1}B\,|u| \;\leq\; -A^{-1}D\,|\lambda|\,|u| \\ \Rightarrow \quad & |u| \;=\; A^{-1}(D-B)\,|u| \;=\; A^{-1}D(I-D^{-1}B)\,|u| \\ & \leq\; A^{-1}D\,|u| - A^{-1}D\,|\lambda|\,|u| \;=\; (1-|\lambda|)A^{-1}D\,|u| \end{aligned}$$

Wäre nun $|\lambda| \geq 1$, so folgte $|u| - (1-|\lambda|)A^{-1}D\,|u| \leq 0$. Da aber dann auch

$$I - (1-|\lambda|)A^{-1}D \geq 0,$$

folgte $|u| \leq 0 \;\Rightarrow\; u = 0$ im Widerspruch zur Annahme.

□

Satz 9. *Eine irreduzible M-Matrix A besitzt eine echt positive Inverse $A^{-1} > 0$.*

Beweis. Seien die Matrizen B, C und D wie oben definiert. Seien $\alpha, \beta \in \{1, \ldots, n\}$. Da A irreduzibel ist (und die Vorzeichenbedingung erfüllt), existiert eine Kette von Verbindungen $\alpha = \alpha_0, \alpha_1, \ldots, \alpha_k = \beta$ mit $a_{\alpha_p \alpha_{p+1}} < 0 \;\forall p \in \{0, \ldots, k-1\}$. Daher ist $c_{\alpha_p \alpha_{p+1}} > 0 \quad \forall p \in \{0, \ldots, k-1\}$ und daraus folgt

$$(C^k)_{\alpha\beta} \;=\; \sum_{\gamma_1, \ldots, \gamma_{k-1}} c_{\alpha\gamma_1} c_{\gamma_1\gamma_2} \ldots c_{\gamma_{k-1}\beta} \;\geq\; c_{\alpha\alpha_1} c_{\alpha_1\alpha_2} \cdot \ldots \cdot c_{\alpha_{k-1}\beta} \;>\; 0.$$

Wie oben bewiesen, gilt $\varrho(C) < 1$, d.h. die Neumann-Reihe $S \coloneqq \sum_{\nu=0}^{\infty} C^\nu$ konvergiert. Da $S_{\alpha\beta} \geq (C^k)_{\alpha\beta} > 0$, ist S durch $C^k > 0$ nach unten beschränkt. Also $S > 0$. Mit $A^{-1} = SD^{-1} > 0$ folgt $A^{-1} > 0$.

□

4.6 Eigenschaften der Systemmatrix der Poisson-Gleichung

Nachdem verschiedene Matrixnormen eingeführt werden, wenden wir uns in diesem Abschnitt den Eigenschaften der Matrix K_h aus der Poisson-Gleichung zu.

Definition 25. *(Vektornorm).*
V sei ein Vektorraum über $\mathbb{K}$ (d.h. $\mathbb{R}$ oder $\mathbb{C}$). Eine Abbildung $\|\cdot\| : V \to \mathbb{R}$ heißt Norm *auf V, falls für alle $u, v \in V$ und $\lambda \in \mathbb{K}$ gilt*

$$\|u\| \;=\; 0 \quad\Leftrightarrow\quad u = 0 \qquad \textit{(Definitheit)}, \tag{4.16}$$

$$\|\lambda u\| \;=\; |\lambda|\,\|u\| \qquad \textit{(Homogenität)}, \tag{4.17}$$

$$\|u+v\| \;\leq\; \|u\| + \|v\| \qquad \textit{(Dreicksungleichung)}. \tag{4.18}$$

Bemerkung 13. *Aus dieser Definition folgt sofort, dass eine Norm nur nicht-negative Werte annehmen kann.*

Definition 26. *(Matrixnorm).*
V sei ein Vektorraum versehen mit einer Vektornorm $\|\cdot\|$ und A ein linearer Operator auf V. Dann ist

$$\|A\|_M \coloneqq \sup\left\{\frac{\|Au\|}{\|u\|} : u \in V \setminus \{0\}\right\} = \sup\{\|Au\| : u \in V,\ \|u\| = 1\} \tag{4.19}$$

die von der Vektornorm $\|\cdot\|$ induzierte Operatornorm *(im endlich-dimensionalen Fall, wo sich A als Matrix darstellen lässt, auch* Matrixnorm *genannt).*

Lemma 5. *Es gilt*

$$\|A\|_M \geq \varrho(A). \tag{4.20}$$

Beweis. Mit $\|A\|_M = \sup\left\{\frac{\|Au\|}{\|u\|} : u \in V \setminus \{0\}\right\}$ und einem Eigenvektor v von A gilt:

$$\begin{aligned} \frac{\|Av\|}{\|v\|} &= \frac{\|\lambda v\|}{\|v\|} = |\lambda| \\ \Rightarrow \sup\left\{\frac{\|Au\|}{\|u\|}\right\} &\geq \max_{v \text{ EV}} \frac{\|Av\|}{\|v\|} = |\lambda^*| = \varrho(A). \end{aligned}$$

□

Bemerkung 14. *Für symmetrische Matrizen gilt auch die umgekehrte Ungleichung.*

4.6.1 Gebräuchliche Matrixnormen

Zeilensummennorm

Proposition 1. *Für die zur Maximumsnorm $\|\cdot\|_\infty$ zugehörige Matrixnorm gilt:*

$$\|A\|_\infty = \max_{i\in\{1,\dots,n\}} \left\{\sum_{j\in\{1,\dots,n\}} |a_{ij}|\right\}. \tag{4.21}$$

Man nennt diese Norm daher auch Zeilensummennorm.

Bemerkung 15. *Aus $A \geq B$ folgt $\|A\|_\infty \geq \|B\|_\infty$, da $A \geq B$ komponentenweise definiert ist.*

Für einen späteren Beweis werden wir folgenden Satz benötigen:

Satz 10. *Sei A eine M-Matrix. Existiert ein Vektor w mit $Aw \geq \mathbb{1}$, dann gilt*

$$\left\|A^{-1}\right\|_\infty \leq \|w\|_\infty. \tag{4.22}$$

Beweis. Bezeichnet man mit $|u|$ den Vektor, der komponentenweise die Beträge von u enthält, so gilt

$$|u| \leq \|u\|_\infty \cdot \mathbb{1} \leq \|u\|_\infty \cdot Aw.$$

Da A eine M-Matrix ist, gilt $A^{-1} \geq 0$, also

$$\begin{aligned} \left|A^{-1}u\right| &\leq A^{-1}|u| \leq A^{-1}\|u\|_\infty Aw \\ &= \|u\|_\infty A^{-1}Aw = \|u\|_\infty \cdot w \\ \Rightarrow \frac{\left|A^{-1}u\right|}{\|u\|_\infty} &\leq w \text{ und speziell } \frac{\left\|A^{-1}u\right\|_\infty}{\|u\|_\infty} \leq \|w\|_\infty \\ \Rightarrow \left\|A^{-1}\right\|_\infty &\leq \|w\|_\infty . \end{aligned}$$

□

Satz 11. *Seien A und B M-Matrizen mit $B \geq A$. Dann gilt:*

$$0 \leq B^{-1} \leq A^{-1} \text{ und } \left\|B^{-1}\right\|_\infty \leq \left\|A^{-1}\right\|_\infty \tag{4.23}$$

Beweis. Es gilt

$$A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}.$$

Da A, B M-Matrizen sind, folgt $A^{-1} \geq 0$ und $B^{-1} \geq 0$ und mit $B \geq A$ folgt $B - A \geq 0$.

$$\Rightarrow\ A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1} \geq 0$$

$$\Leftrightarrow\ A^{-1} \geq B^{-1}\ \Rightarrow\ \left\|B^{-1}\right\|_\infty \leq \left\|A^{-1}\right\|_\infty .$$

□

Spektralnorm

Zur euklidischen Vektornorm $\|u\|_2 = \sqrt{\sum_{i=1}^n |u_i|^2}$ lässt sich eine zugehörige Matrixnorm, die *Spektralnorm,* definieren.

Proposition 2. *Die auf einem endlich-dimensionalen Hilbertraum V von der euklidischen Vektornorm induzierte Matrixnorm $\|\cdot\|_2$ lässt sich schreiben als*

$$\|A\|_2 = \sqrt{\varrho\left(A^T A\right)} \tag{4.24}$$

und heißt Spektralnorm.

Beweis.

$$\|A\|_2 = \sup\left\{\frac{\|Au\|_2}{\|u\|_2} : u \in V \setminus \{0\}\right\} = \max_{u \in V, \|u\|_2 = 1} \|Au\|_2 .$$

Das Maximum wird angenommen, da auf endlich-dimensionalen Vektorräumen die Einheitskugel kompakt ist. Es folgt

$$\|A\|_2^2 = \max_{u \in V, \|u\|_2 = 1} \|Au\|_2^2 = \max_{\|u\|_2 = 1} \langle Au, Au\rangle = \max_{\|u\|_2 = 1} \langle A^T Au, u\rangle.$$

Da A^TA symmetrisch und positiv semi-definit ist, liefert eine Hauptachsentransformation

$$P^TA^TAP = \operatorname{diag}(\lambda_i) =: D$$

mit $PP^T = \mathbb{1}$, $\left\|P^Tu\right\| = \|u\|\ \forall u \in V$ und den Eigenwerten $\lambda_i > 0$ von A^TA. Es folgt:

$$\begin{aligned}
\|Au\|_2^2 &= \max_{\|u\|_2=1} \langle A^TAu, u\rangle = \max_{\|u\|_2=1} \langle A^TAPP^Tu, PP^Tu\rangle \\
&\underbrace{=}_{\tilde{u}=P^Tu} \max_{\|\tilde{u}\|_2=1} \langle A^TAP\tilde{u}, P\tilde{u}\rangle = \max_{\|\tilde{u}\|_2=1} \langle P^TA^TAP\tilde{u}, \tilde{u}\rangle = \max_{\|\tilde{u}\|_2=1} \langle D\tilde{u}, \tilde{u}\rangle \\
&= \max_{\|\tilde{u}\|_2=1} \left(\sum_{i=1}^{n} \lambda_i |\tilde{u}_i|^2\right) = \lambda_{max} = \varrho\left(A^TA\right) \\
&\Rightarrow \quad \|A\|_2 = \sqrt{\varrho\left(A^TA\right)}.
\end{aligned}$$

□

Bemerkung 16. *Für A symmetrisch ist $A^TA = A^2$. Da $\rho\left(A^2\right) = \rho^2(A)$, ist in diesem Fall $\|A\|_2 = \varrho(A)$.*

4.6.2 Positiv definite Matrizen

Bevor wir im folgenden Abschnitt die wichtigsten Eigenschaften der Systemmatrix zur diskreten Poisson-Gleichung zusammenstellen, benötigen wir noch den Begriff der *positiven Definitheit*.

Definition 27. *Eine Matrix $A \in \mathbb{K}^{n\times n}$ heißt* positiv definit*, wenn A symmetrisch ist und*

$$\langle Au, u\rangle > 0 \quad \forall u \in \mathbb{K}^n \setminus \{0\}. \tag{4.25}$$

Lemma 6. *A ist positiv definit genau dann, wenn alle Eigenwerte von A positiv sind.*

Beweis. A symmetrisch $\Rightarrow \exists P$: $P^TAP = \operatorname{diag}(\lambda_i)$. Dabei sind λ_i die Eigenwerte von A und P die Matrix zusammengesetzt aus den Eigenvektoren.

$$\Rightarrow \langle Au, u\rangle = \langle AP\tilde{u}, P\tilde{u}\rangle = \langle P^TAP\tilde{u}, \tilde{u}\rangle = \sum_{i=1}^{n} \lambda_i |u_i|^2$$

und daher gilt

$$\langle Au, u\rangle > 0\ \forall u \in \mathbb{K}^n \setminus \{0\} \ \Leftrightarrow\ \lambda_i > 0\ \forall i.$$

□

Daraus ergibt sich folgendes

Lemma 7. *Eine positiv definite Matrix ist regulär und besitzt eine positiv definite Inverse.*

Lemma 8. *Für A symmetrisch und diagonaldominant (oder irreduzibel diagonaldominant) mit $a_{ii} > 0$ ist A positiv definit.*

Beweis.

$$\sum_{j=1, j\neq i}^{n} |a_{ij}| < a_{ii} \Rightarrow \text{ Gerschgorinkreise liegen in } (0, \infty)$$

$\Rightarrow \lambda_i$ positiv $\Rightarrow A$ ist positiv definit.

□

Lemma 9. *$\lambda_{\min}$ und $\lambda_{\max}$ seien der minimale bzw. der maximale Eigenwert von A (positiv definit). Dann gilt:*

$$\begin{aligned} \|A\|_2 &= \lambda_{\max} \\ \|A^{-1}\|_2 &= \frac{1}{\lambda_{\min}} \end{aligned}$$

Beweis. A ist symmetrisch, daher gilt $\|A\|_2 = \varrho(A) = \lambda_{\max}$. Genauso ist A^{-1} symmetrisch und daher $\|A^{-1}\|_2 = \varrho\left(A^{-1}\right) = \frac{1}{\lambda_{\min}}$.

□

4.6.3 Matrixeigenschaften von K_h

Die in den vorigen Abschnitten zusammengefassten Sätze und Matrixeigenschaften ermöglichen den Beweis folgender Aussagen:

Satz 12. *Die Domain Ω sei durch $\Omega = (0,1) \times (0,1)$ gegeben. Die durch den Stern*

$$\begin{bmatrix} & 1 & \\ 1 & -4 & 1 \\ & 1 & \end{bmatrix},$$

definierte Matrix K_h besitzt dann folgende Eigenschaften:

1. *K_h ist eine M-Matrix.*
2. *K_h ist positiv definit.*
3. *$\|K_h\|_\infty \leq \frac{8}{h^2}$ und $\left\|K_h^{-1}\right\|_\infty \leq \frac{1}{8}$.*
4. *$\|K_h\|_2 \leq \frac{8}{h^2} \cos^2\left(\frac{\pi h}{2}\right) < \frac{8}{h^2}$ und*
5. *$\left\|K_h^{-1}\right\|_2 \leq \frac{1}{8} h^2 \sin^{-2}\left(\frac{\pi h}{2}\right) = \frac{1}{2\pi^2 + \mathcal{O}(h^2)} < \frac{1}{16}$ für h klein genug.*

Beweis.

1. K_h erfüllt die Vorzeichenbedingung und ist irreduzibel diagonaldominant. In Satz 8 haben wir gezeigt, dass K_h dann auch eine M-Matrix ist.
2. K_h ist symmetrisch und irreduzibel diagonaldominant, also ist K_h positiv definit (siehe Lemma 8).

3.

a) $\|K_h\|_\infty = \max_{i=1,\ldots,n} \left\{ \sum_{j=1}^n |K_{ij}| \right\} = \frac{1}{h^2} \max\{6,7,8\} = \frac{8}{h^2}$

b) Zu zeigen ist: $\left\|K_h^{-1}\right\|_\infty \leq \frac{1}{8}$. Nutze dazu Satz 10 mit

$$w(x,y) = \frac{x(1-x)}{2}$$

und betrachte

$$K_h w_h(x,y) = -\frac{(x-h)(1-(x-h))}{2h^2} - \frac{(x+h)(1-(x+h))}{2h^2} + \frac{2 \cdot x(1-x)}{2h^2} = 1$$

Es ist also $K_h w_h \geq \mathbb{1}$. Für w_h gilt offenbar $\|w_h\|_\infty \leq \max_{x,y} w(x,y) = \frac{1}{8}$. Wegen Satz 10 gilt also: $\left\|K_h^{-1}\right\|_\infty \leq \|w\|_\infty \leq \frac{1}{8}$.

4. Die Eigenvektoren von K_h sind $u^{\nu,\mu}$ $(1 \leq \nu, \mu \leq n-1)$ mit

$$u_{j,k}^{\nu,\mu} = \sin(\nu\pi jh)\sin(\mu\pi kh) \tag{4.26}$$

und den zugehörigen Eigenwerten

$$\lambda_{\nu,\mu} = \frac{2}{h^2}\left(\sin^2\left(\frac{\nu\pi h}{2}\right) + \sin^2\left(\frac{\mu\pi h}{2}\right)\right). \tag{4.27}$$

In der Tat: Für alle inneren Knoten indiziert durch (j,k) gilt:

$$\begin{aligned}(K_h u^{\nu,\mu})_{j,k} \;=\; & \frac{1}{h^2}\left(4\sin(\nu\pi jh)\sin(\mu\pi kh)\right. \\ & -\sin(\nu\pi(j-1)h)\sin(\mu\pi kh) - \sin(\nu\pi(j+1)h)\sin(\mu\pi kh) \\ & \left.-\sin(\nu\pi jh)\sin(\mu\pi(k-1)h) - \sin(\nu\pi jh)\sin(\mu\pi(k+1)h)\right).\end{aligned}$$

Man wendet das Additionstheorem

$$\sin(a \pm b) = \sin a \cos b \pm \sin b \cos a$$

an mit $a := \nu\pi jh$ und $b := \nu\pi h$ bzw. $a := \mu\pi kh$ und $b := \mu\pi h$ und erhält daraus

$$(K_h u^{\nu,\mu})_{j,k} = \frac{1}{h^2}\left(4 - 2\cos(\nu\pi h) - 2\cos(\mu\pi h)\right)\sin(\nu\pi jh)\sin(\mu\pi kh).$$

Aus der Identität $1 - \cos(a) = 2\sin^2\left(\frac{a}{2}\right)$ ergibt sich schließlich

$$(K_h u^{\nu,\mu})_{j,k} = \frac{4}{h^2}\left(\sin^2\left(\frac{\nu\pi h}{2}\right) + \sin^2\left(\frac{\mu\pi h}{2}\right)\right) u_{j,k}^{\nu,\mu}.$$

□

$\Rightarrow \|K_h\|_2 = \varrho(A) = \lambda_{max} \leq \frac{8}{h^2}\sin^2\left(\frac{\pi(n-1)h}{2}\right) = \frac{8}{h^2}\cos^2\left(\frac{\pi h}{2}\right) < \frac{8}{h^2}$.

□

5. Es gilt

$$\begin{aligned} \left\|K_h^{-1}\right\|_2 &= \varrho(K_h^{-1}) = \frac{1}{\lambda_{min}} \text{ mit} \\ \lambda_{min} &= \frac{8}{h^2}\sin^2\left(\frac{\pi h}{2}\right) = \frac{8}{h^2}\left(\frac{\pi h}{2} - \mathcal{O}(h^3)\right)^2 \\ &= \frac{8}{h^2}\left(\left(\frac{\pi h}{2}\right)^2 - \mathcal{O}(h^4)\right) = 2\pi^2 - \mathcal{O}(h^2) \\ \Rightarrow \quad \left\|K_h^{-1}\right\|_2 &\leq \frac{1}{2\pi^2 - \mathcal{O}(h^2)}. \end{aligned}$$

□

4.7 Konvergenzuntersuchung für das Finite-Differenzen-Verfahren

Dass ein numerisches Approximationsverfahren mit zunehmender Gitterfeinheit gegen die kontinuierliche Lösung konvergiert, ist Grundvoraussetzung für den effizienten Einsatz des Verfahrens. Wir werden im Folgenden zeigen, dass das behandelte Differenzenverfahren für die Poissongleichung *stabil* und *konvergent* ist.

4.7.1 Stetige Abhängigkeit von den Randdaten

4.7.1.1 Maximumsprizip

Lemma 10. *Sei u_h eine Lösung von $-\Delta_h u_h = f_h$ mit $f_h = 0$ und $u_h|_{\partial\Omega_h} = \varphi_h$. Dann nimmt u_h sein Maximum bzw. Minimum auf dem Rand $\partial\Omega_h$ an.*

Beweis. Für $-\Delta_h u_h = 0$ gilt

$$u_h(x,y) = \frac{1}{4}(u_h(x-h,y) + u_h(x+h,y) + u_h(x,y-h) + u_h(x,y+h)) \tag{4.28}$$

Angenommen, $u_h(x,y)$ ist maximal mit $(x,y) \in \Omega_h$.
Dann folgt aus (4.28) bereits, dass $u_h(x \pm h, y)$ und $u_h(x, y \pm h)$ maximal sein müssen. Da K_h aber irreduzibel ist, setzt sich diese Bedingung auf ganz $\overline{\Omega}_h$ fort. Also ist u_h konstant, insb. wird das Maximum (auch) auf dem Rand angenommen. Entsprechend für das Minimum.

□

4.7.1.2 Vergleichsprinzip

Seien u_h, v_h Lösungen von

$$-\Delta_h u_h = f_h \text{ mit } u_h|_{\partial\Omega_h} = \varphi_h^u$$
$$-\Delta_h v_h = f_h \text{ mit } v_h|_{\partial\Omega_h} = \varphi_h^v$$

Dann gilt:

1. $\|u_h - v_h\|_\infty \leq \max_{x\in\partial\Omega_h} |\varphi^u(x) - \varphi^v(x)|$
2. $u_h \leq v_h$, falls $\varphi^u \leq \varphi^v$ auf $\partial\Omega_h$.

Beweis. Betrachte $w_h := v_h - u_h$. Dann ist w_h Lösung der Poisson-Gleichung $-\Delta_h w_h = 0$ und $w_h \geq 0$ auf $\partial\Omega_h$, falls

$$\varphi^u \leq \varphi^v.$$

Aus dem Maximumsprinzip folgt: $w_h > 0$. $\Rightarrow$ Behauptung 2.
Mit $u_h = \varphi^u$ und $v_h = \varphi^v$ auf $\partial\Omega_h$ gilt

$$|w_h| \leq \max_{\partial\Omega_h} |\varphi^u - \varphi^v| \quad \text{auf } \partial\Omega_h.$$

Aufgrund des Maximumsprinzips gilt dies auf ganz $\Omega_h \Rightarrow$ Behauptung 1.

□

4.7.2 Konvergenz, Konsistenz und Stabilität

Wir wollen nun das kontinuierliche mit dem numerisch approximierten Problem vergleichen. Wir betrachten dazu

$$\begin{aligned} -\Delta u &= f \text{ in } \Omega, \\ u|_\Gamma &= \varphi \end{aligned}$$

und

$$\begin{aligned} -\Delta_h u_h &= f_h \text{ in } \Omega_h, \\ u_h|_{\Gamma_h} &= \varphi_h. \end{aligned}$$

Bemerkung 17. *Das kontinuierliche und das diskrete Problem existieren in unterschiedlichen Gebieten. Um beide Probleme vergleichen zu können, müssen sie in einem Raum definiert sein.*

Definition 28. *(Restriktion).*
Sei $h \in H \subset \mathbb{R}^+$ und H eine Menge ohne Häufungspunkt (in unserem Fall ist $H = \left\{\frac{1}{n} : n \in \mathbb{N}\right\}$). Sei U_h der Raum der Gitterfunktionen auf $\overline{\Omega}_h$. Die Abbildung

$$\begin{aligned} R_h : C\left(\overline{\Omega}\right) &\longrightarrow U_h \\ u &\mapsto R_h u \end{aligned}$$

mit $(R_h u)(x) = u(x)$ für alle $x \in \overline{\Omega}_h$ heißt die Restriktion *der Lösung in $\overline{\Omega}$ auf $\overline{\Omega}_h$.*

Bemerkung 18. *Im Folgenden is die genaue Norm, bezüglich derer die diskrete mit der kontinuierlichen Lösung verglichen wird, nicht näher spezifiziert, da sie aufgrund der Äquivalenz aller Normen auf endlich-dimensionalen $\mathbb{K}$-Vektorräumen ohnehin nicht releviert.*

4.7.2.1 Stabilität

Definition 29. *Die Diskretisierung K_h heißt* stabil *für $h \in H \subset \mathbb{R}^+$, falls*

$$\sup_{h \in H} \left\| K_h^{-1} \right\| \leq C < \infty \tag{4.29}$$

Bemerkung 19. *Man betrachte die beiden Systeme*

$$\begin{aligned} K_h(u_h) &= f_h, \\ K_h(\tilde{u}_h) &= f_h + \varepsilon. \end{aligned}$$

Dann folgt

$$\begin{aligned} u_h &= K_h^{-1}(f_h) \\ \tilde{u}_h &= K_h^{-1}(f_h + \epsilon) \\ \Rightarrow \|\tilde{u}_h - u_h\| &\leq C \cdot \|\varepsilon\| \end{aligned}$$

Stabilität bedeutet also, dass eine kleine Änderung auf der rechten Seite nur kleine Änderungen in der Lösung bewirken.

Beispiel 24. *Die Fünfpunktstern-Diskretisierungsmatrix K_h zur Poissongleichung hat die Eigenschaft*

$$\left\| K_h^{-1} \right\|_\infty \leq \frac{1}{8},$$

ist also stabil.

4.7.2.2 Konsistenz

Definition 30. *Sei $K_h u_h = f_h$ die Diskretisierung von $Ku = f$. K sei ein Differentialoperator der Ordnung m. Weiter seien R_h und $\tilde{R}_h$ Restriktionsoperatoren für u und f. Die Diskretisierung $(K_h, R_h, \tilde{R}_h)$ des Differetialoperators K hat die* Konsistenzordnung *k, falls gilt*

$$\left\| K_h R_h u - \tilde{R}_h K u \right\| \leq C \cdot h^k \cdot \|u\|_{C^{k+m}(\overline{\Omega})} \quad \forall u \in C^{k+m}(\overline{\Omega}). \tag{4.30}$$

Beispiel 25. *Sei $R_h = \tilde{R}_h$ gegeben durch*

$$(R_h u)(x) = u(x) \quad \forall x \in \Omega_h.$$

Dann ist $(K_h, R_h, \tilde{R}_h) = (\Delta_h, R_h, R_h)$ konsistent mit Ordnung 2.

Beweis. Man erinnere sich an die Abschätzung

$$(\partial^-\partial^+u)(x) = u''(x) + h^2R, \quad |R| \leq \frac{1}{12}\left\|u^{(4)}\right\|_{C^0(\bar{\Omega})}.$$

Im $\mathbb{R}^2$ wendet man diesen Ansatz in x und y-Richtung an:

$$\begin{aligned} -\Delta_h Ru(x,y) &= -\Delta u(x,y) + h^2(R_x + R_y) \\ \text{mit } |R_x|, |R_y| &\leq \frac{1}{12}\left\|u^{(4)}\right\|_{C^0(\bar{\Omega})} \leq \frac{1}{12}\|u\|_{C^4(\bar{\Omega})}. \end{aligned}$$

Daher gilt $\|K_hR_hu - R_hKu\| \leq C \cdot h^2\|u\|_{C^4(\bar{\Omega})}$ mit $C = \frac{1}{6}$.

□

4.7.2.3 Konvergenz

Definition 31. *Sei $K_hu_h = f_h$ die Diskretisierung von $Ku = f$. K sei ein Differentialoperator der Ordnung m. Die diskrete Lösung $u_h \in U_h$ $(h \in H)$ konvergiert mit der Ordnung k gegen u, falls*

$$\|u_h - R_hu\| \leq C \cdot h^k \cdot \|u\|_{C^{k+m}(\overline{\Omega})} \tag{4.31}$$

Definition 32. $u_h - R_hu$ *heißt* Diskretisierungsfehler *eines Diskretisierungsverfahrens.*

Satz 13. *(Satz über die Konvergenz).*
Sei K ein Differentialoperator der Ordnung m. Die Diskretisierung $(K_h, R_h, \tilde{R}_h)$ für K sei stabil und konsistent von der Ordnung k. Dann ist das Verfahren konvergent *von der Ordnung k, falls $u \in C^{k+m}(\overline{\Omega})$.*

Beweis. Man betrachte $w_h = u_h - R_hu$. Es soll gezeigt werden, dass $w_h \to 0$ für $h \to 0$. Es gilt:

$$K_hw_h = K_hu_h - K_hR_hu = f_h - K_hR_hu = \tilde{R}_hf - K_hR_hu = \tilde{R}_hKu - K_hR_hu$$

$$\begin{aligned} \Rightarrow w_h &= K_h^{-1}(\tilde{R}_hKu - K_hR_hu) \\ \Rightarrow \|w_h\| &\leq \left\|K_h^{-1}\right\| \left\|\tilde{R}_hKu - K_hR_hu\right\| \\ \Rightarrow \|u_h - R_hu\| &\leq Ch^k \|u\|_{C^{k+m}(\overline{\Omega})}. \end{aligned}$$

□

Beispiel 26. *Die Fünfpunktstern-Diskretisierung für den Laplace-Operator ist konvergent von der Ordnung 2. Sei $u \in C^4(\overline{\Omega})$. Dann gilt :*

$$\|u_h - R_hu\|_\infty \leq \frac{h^2}{48} \cdot \|u\|_{C^4(\overline{\Omega})}.$$

4.8 Das Neumann-Problem

Bisher wurden die Funktionswerte der gesuchten Funktion auf dem Rand des Gebietes vorgegeben mit $u|_\Gamma = \varphi$. Stattdessen können auch die Ableitungen von u in Normalenrichtung auf dem Rand vorgegeben werden, d.h. wir betrachten das Problem

$$\begin{aligned} -\Delta u &= f \text{ in } \Omega = (0,1) \times (0,1) \\ \frac{\partial u}{\partial n} &= \varphi \text{ auf } \Gamma \end{aligned} \tag{4.32}$$

Bemerkung 20. *Dadurch, dass $\frac{\partial u}{\partial n}$ vorgegeben wird, gibt es potentiell unendlich viele Lösungen. Denn falls u Lösung des Problems ist, so erfüllt auch $u + c$ das Problem. Für Eindeutigkeit der Lösung fordert man daher zusätzlich etwa $\int_\Omega u \, dx = 0$.*

Das Neumann-Problem (4.32) besitzt nicht für beliebige stetige Funktionen f und φ eine Lösung. Ein notwendiges Kriterium für die Lösbarkeit ist die sog. *Kompatibilitätsbedingung.*

Lemma 11. *(Kompatibilitätsbedingung)*
Existiert eine Lösung u des Neumann-Problems (4.32)*, so gilt:*

$$\int_\Omega f \, dx + \int_{\partial\Omega} \varphi \, ds \; = \; 0. \tag{4.33}$$

Beweis. Die Aussage folgt sofort aus dem Satz von Gauß:

$$\begin{aligned} -\int_\Omega f \, dx &= \int_\Omega \Delta u \, dx = \int_\Omega \operatorname{div}(\nabla u) \, dx \\ &= \int_{\partial\Omega} \nabla u \cdot \vec{n} \, ds = \int_{\partial\Omega} \frac{\partial u}{\partial \vec{n}} \, ds = \int_{\partial\Omega} \varphi \, ds. \end{aligned}$$

□

4.8.1 Diskretisierung des Neumann-Problems

Wir betrachten das Neumann-Problem auf dem Gebiet $\overline{\Omega} = [0,1] \times [0,1]$. Dabei treten im Inneren drei Typen von Knoten auf:

1. reine innere Knoten,
2. randnahe Knoten, also solche, die Nachbar zu genau einem Randknoten sind,
3. ecknahe Knoten, also solche, die zwei Randknoten zum Nachbar haben.

Die Diskretisierung des Laplace-Operators mit dem Fünfpunktstern lässt sich in den letzten beiden Fällen durch die mit einseitigen finiten Differenzen diskretisierte Randbedingung vereinfachen zu:

innere Punkte Für Punkte im Innern, die keine Randknoten zum Nachbarn haben, gilt:
$$-\Delta_h u_h(x,y) = \frac{1}{h^2}\left(4u_h(x,y) - u_h(x-h,y) - u_h(x+h,y) - u_h(x,y-h) - u_h(x,y+h)\right)$$

randnahe Punkte Für innere Knoten am rechten Rand von $\overline{\Omega}$ ergibt sich:
$$-\Delta_h u_h(x,y) = \frac{1}{h^2}\left(3u_h(x,y) - u_h(x-h,y) - u_h(x,y-h) - u_h(x,y+h)\right)$$

ecknahe Punkte Für den inneren Knoten rechts unten ergibt sich:
$$-\Delta_h u_h(x,y) = \frac{1}{h^2}\left(2u_h(x,y) - u_h(x-h,y) - u_h(x,y+h)\right)$$

4.8.1.1 Behandlung der Neumann-Randbedingung

Die Ableitung von u in Normalenrichtung $\vec{n}$ kann über einseitige Differenzenquotienten approximiert werden. Im eindimensionalen Fall lässt sich der rechte Rand folgendermaßen darstellen:

$$\frac{\partial u_h}{\partial \vec{n}}(x) \approx (\partial_n^- u_h)(x) = \frac{1}{h}(u_h(x) - u(x - h\vec{n})) = \varphi(x) \tag{4.34}$$

Für unser Modellbeispiel folgt damit:

$$\begin{aligned}
\frac{1}{h}\left(u_h(x,0) - u_h(x,h)\right) &= \varphi(x,0) \quad \text{(unten)},\\
\frac{1}{h}\left(u_h(x,1) - u_h(x,1-h)\right) &= \varphi(x,1) \quad \text{(oben)},\\
\frac{1}{h}\left(u_h(0,y) - u_h(h,y)\right) &= \varphi(0,y) \quad \text{(links)},\\
\frac{1}{h}\left(u_h(1,y) - u_h(1-h,y)\right) &= \varphi(1,y) \quad \text{(rechts)}.
\end{aligned}$$

Das resultierende diskrete Problem für die Gitterfunktion u_h *auf den inneren* Knoten

$$\begin{aligned}
K_h u_h &= \tilde{f}_h = f_h + \frac{1}{h}\varphi_h,\\
\varphi_h &= \sum_{\text{Nachbarrandknoten}} \varphi\,(x,y)
\end{aligned} \tag{4.35}$$

ist nicht notwendigerweise regulär. Deshalb muss zusätzlich die *diskrete Kompatibilitätsbedingung* erfüllt sein.

4.8.1.2 Diskrete Kompatibilitätsbedingung

Die diskrete Kompatibilitätsbedingung lautet

$$h^2 \sum_{x\in\Omega_h} f(x) + h \sum_{x\in\Gamma_h'} \varphi(x) \;=\; 0, \tag{4.36}$$

wobei Γ_h' gerade diejenigen Randknoten enthalten soll, die keine Eckknoten sind.

Satz 14. *Das Gleichungssystem $K_h u_h = \tilde{f}_h$ ist genau dann lösbar, wenn die diskrete Kompatibiltätsbedingung erfüllt ist. In diesem Fall unterscheiden sich zwei Lösungen nur um eine Konstante c.*

Beweis. Anwendung des diskreten Laplace-Operators K_h auf eine konstante Funktion ergibt Null, d.h.

$$K_h \cdot c \cdot \mathbb{1} = 0 \;\Rightarrow\; c \cdot \mathbb{1} \in \ker(K_h).$$

Man zeigt ferner mit wenig Aufwand, dass die konstanten Gitter-Funktionen bereits den ganzen Kern von K_h ausmachen, also $\dim(\ker(K_h)) = 1$.
Außerdem ist $K_h u_h = \tilde{f}_h$ lösbar genau dann, wenn $\tilde{f}_h \in \mathrm{im}(K_h)$ und wegen $\mathrm{im}(K_h) = \ker(K_h^T)^\perp = \ker(K_h)^\perp = span(\mathbb{1})^\perp$ gilt daher, dass $K_h u_h = \tilde{f}_h$ lösbar ist genau dann, wenn $\sum_{x\in\Omega_h} \tilde{f}_h(x) = 0$. Es ist aber wegen der Definition von $\tilde{f}_h$ und φ_h in (4.35): $\sum_{x\in\Omega_h} \tilde{f}_h(x) = \sum_{x\in\Omega_h} \tilde{f}_h(x) + \frac{1}{h}\sum_{x\in\Gamma'_h} \varphi(x)$.

□

4.8.2 Lösen des Neumann-Problems

Die Lösbarkeit des Neumann-Problems setzt, wie oben gezeigt, die Erfüllung der Kompatibilitätsbedingung voraus. Wir werden zur Vereinfachung der Notation im Folgenden mit f_h die rechte Seite mit den Beiträgen auf dem Rand bezeichnen, d.h. $f_h - \frac{1}{h}\varphi_h \rightsquigarrow f_h$.

Lemma 12. *Man wähle $x_0 \in \Omega_h$ beliebig und normiere u_h durch*

$$u_h(x_0) = 0.$$

Dann ist das um eine Zeile und eine Spalte reduzierte System

$$\hat{K}_h \hat{u}_h = \hat{f}_h$$

lösbar.

Beweis. $\hat{K}_h$ erfüllt die Vorzeichenbedingung, da bereits K_h sie erfüllt. Außerdem ist $\hat{K}_h$ (durch das Streichen einer Spalte) irreduzibel diagonaldominant, da in K_h die Betragssumme der Nebendiagonalelemente immer genau gleich dem Betrag des Diagonalelements ist. Damit ist $\hat{K}_h$ eine M-Matrix und also insb. invertierbar.

□

Bemerkung 21. *Man sieht schnell ein, dass das Setzen eines Dirichlet-0-Wertes im Punkt x_i äquivalent ist zum Lösen des korrigierten Systems*

$$K_h u_h = \check{f}_h$$
$$mit\ \left(\check{f}_h\right)_j = \begin{cases} (f_h)_j, & j \neq i \\ -\sum_{k\neq i} (f_h)_k, & j = i \end{cases}$$

unter der Nebenbedingung $(u_h)_i = 0$.

Man beachte, dass $\left(\check{f}_h\right)_i = (f_h)_i$ *genau dann gilt, wenn die diskrete Kompatibilitätsbedingung erfüllt ist. Ist dies nicht der Fall, löst man durch Streichen der i-ten Zeile und Spalte allerdings ein im i-ten Eintrag der rechten Seite gestörtes Problem. Da i.A. aus der kontinuierlichen Kompatibilitätsbedingung* (4.33) *noch nicht die diskrete folgt, ist dies ein Problem. Für glatte Funktionen* f *und* φ*, die* (4.33) *erfüllen, hat man etwa* $\left(\check{f}_h\right)_i - (f_h)_i = O\left(h^{-1}\right)$ *und erzeugt damit i.A. eine Singularität in* x_i*!*

4.8.2.1 Verteilung der Korrektur

Eine Alternative zum Setzen eines Dirichlet-Wertes ist es, die Korrektur durch Erweiterung des Systems auf alle Unbekannten zu verteilen.
Betrachte dazu

$$\bar{K}_h \bar{u}_h = \bar{f}_h$$

mit

$$\bar{K}_h = \begin{pmatrix} K_h & \mathbb{1} \\ \mathbb{1}^T & 0 \end{pmatrix}, \bar{u}_h = \begin{pmatrix} u_h \\ \lambda \end{pmatrix}, \bar{f}_h = \begin{pmatrix} f_h \\ \sigma \end{pmatrix}$$

mit beliebigem σ.

Lemma 13. *Das System* $\bar{K}_h \bar{u}_h = \bar{f}_h$ *ist eindeutig lösbar.*

Beweis. Da $\mathbb{1} \notin \operatorname{im}(K_h)$, ist $\operatorname{Rang}(K_h, \mathbb{1}) = \operatorname{Rang}(K_h) + 1$. Außerdem ist die letzte Zeile von $\bar{K}_h$ linear unabhängig von den anderen. Also ist $\bar{K}_h$ invertierbar.

□

Lemma 14. *Ist* $\bar{u}_h$ *Lösung von* $\bar{K}_h \bar{u}_h = \bar{f}_h$*, dann ist* u_h *Lösung des gestörten Problems* $K_h u_h = \check{f}_h$ *mit* $\check{f}_h = f_h - \lambda \cdot \mathbb{1}$*.*

Beweis. Die Aussage folgt sofort durch Subtraktion von $\lambda\mathbb{1}$ auf beiden Seiten des Gleichungssystems.

□

Bemerkung 22. *Im Fall* $\lambda = 0$ *gilt* (4.36) *und* u_h *ist diejenige Lösung von* (4.35)*, welche die Normierungsbedingung*

$$\mathbb{1}^T \cdot u_h = \sum_{x \in \Omega_h} u_h(x) = \sigma$$

erfüllt.
Andernfalls gilt

$$\lambda = \frac{\sum_{x \in \Omega_h} f_h(x)}{\mathbb{1}^T \mathbb{1}}$$

und für glatte Funktionen f *und* φ*, die* (4.33) *erfüllen, gilt daher* $\lambda = O(h)$*.*

4.9 Differenzenverfahren für allgemeine Probleme zweiter Ordnung

In diesem Abschnitt betrachten wir allgemeine Probleme zweiter Ordnung und die daraus resultierenden diskreten Gleichungen. Wir betrachten nun folgendes Problem:

$$\begin{aligned} Ku &= f \quad \text{in } \Omega, \\ K &= \sum_{i,j=1}^{n} a_{ij}(x)\frac{\partial^2}{\partial x_i x_j} + \sum_{i=1}^{n} b_i(x)\frac{\partial}{\partial x_i} + c(x). \end{aligned} \tag{4.37}$$

Bemerkung 23. *Man kann $a_{ij}(x) = a_{ji}(x)$ wählen, da $\frac{\partial^2}{\partial x_i x_j} = \frac{\partial^2}{\partial x_j x_i}$ für zweimal stetig differenzierbare Funktionen.*
Damit ist $A(x) = (a_{ij}(x))_{i,j=1\ldots n}$ symmetrisch. Man beachte, dass die Koeffizientenfunktionen $a_{ij}(x)$ ortsabhängig sein können. Der oben definierte Differentialoperator zweiter Ordnung ist in der allgemeinsten Operatorform notiert.

Definition 33. *(Elliptizität)*
Ein Differentialoperator heißt elliptisch*, falls alle Eigenwerte des Hauptteils (A) des Differentialoperators das gleiche Vorzeichen besitzen also wenn gilt:*

$$\sum_{i,j=1}^{n} a_{ij}(x)\xi_i\xi_j > 0 \quad \forall x \in \Omega, 0 \neq \xi \in \mathbb{R}^n \quad \text{oder}$$

$$\sum_{i,j=1}^{n} a_{ij}(x)\xi_i\xi_j < 0 \quad \forall x \in \Omega, 0 \neq \xi \in \mathbb{R}^n.$$

Bemerkung 24. *$-\Delta u = f$ ist eine elliptische partielle Differentialgleichung, da hier für den Hauptteil gilt:*

$$A = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Definition 34. *Ein Differentialoperator K heißt* gleichmäßig elliptisch *in Ω, falls gilt*

$$\sum_{i,j=1}^{n} a_{ij}(x)\xi_i\xi_j \geq c(x)\left\|\xi\right\|^2, \, c(x) > 0 \quad \forall x \in \Omega, \, 0 \neq \xi \in \mathbb{R}^n$$

Im zweidimensionalen Fall auf dem Gebiet $\Omega = (0,1) \times (0,1)$ erhalten wir für randferne Punkte den diskreten Operator

$$\begin{aligned} & a_{11}(x,y)\partial_x^+\partial_x^- + 2a_{12}(x,y)\,\partial_x^0\partial_y^0 + a_{22}(x,y)\partial_y^+\partial_y^- + b_1(x,y)\partial_x^0 + b_2(x,y)\partial_y^0 + c(x,y) \\ = \; & h^{-2}\begin{bmatrix} -\frac{1}{2}a_{12}(x,y) & a_{22}(x,y) & \frac{1}{2}a_{12}(x,y) \\ a_{11}(x,y) & -2(a_{11}(x,y)+a_{22}(x,y)) & a_{11}(x,y) \\ \frac{1}{2}a_{12}(x,y) & a_{22}(x,y) & -\frac{1}{2}a_{12}(x,y) \end{bmatrix} + \\ & + (2h)^{-1}\begin{bmatrix} 0 & b_2(x,y) & 0 \\ -b_1(x,y) & 0 & b_1(x,y) \\ 0 & b_2(x,y) & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & c(x,y) & 0 \\ 0 & 0 & 0 \end{bmatrix}. \end{aligned}$$

5 Diskretisierung II: Finite Elemente Verfahren

Während das Diskretisierungsverfahren der Finiten Differenzen einen einfachen mathematischen Zugang über die direkte Approximation der Differentialoperatoren bietet und auch in der Implementierung verhältnismäßig leicht zu handhaben ist, gibt es doch gewisse Einschränkungen aufgrund des grundsätzlichen Vorgehens in diesem Verfahren. Zum Beispiel erfordert die Behandlung komplexer Gebiete mit krummen Rändern besondere numerische Behandlung, was wiederum die Gesamtkomplexität beeinflusst. Ferner sahen wir in den Konvergenzbeweisen für das Finite-Differenzen-Verfahren sehr hohe Regularitätsannahmen für die Lösungsfunktion, die eine starke Einschränkung darstellen können. Um eine gebietsunabhängige Herangehensweise an die Diskretisierung aufzubauen sowie ein Verfahren, das substantiell schwächere Regularitätsvoraussetzungen stellt, werden wir einen funktionalanalytischen Weg finden. An die Stelle der direkten Approximation von Differentialoperatoren tritt die Approximation der Lösungsräume, in denen die Lösungsfunktion liegt, durch endlich-dimensionale, approximierende Funktionenräume. Ein daraus resultierendes numerisches Verfahren ist das Verfahren der *Finiten Elemente*, welchem wir uns in diesem Kapitel widmen werden.

5.1 Funktionalanalytische Grundlagen

Wie oben schon erwähnt, werden wir über die Funktionalanalysis eine neue Kategorie der Diskretisierung herleiten. Dazu sind einige funktionalanalytische Grundlagen notwendig, die in diesem Abschnitt zusammengestellt werden.

5.1.1 Normierte Räume

Definition 35. *Sei X ein Vektorraum über $\mathbb{R}$ oder $\mathbb{C}$ und $\|\cdot\|_X : X \to [0,\infty)$ eine Norm. Dann wird*

$$(X, \|\cdot\|_X)$$

normierter Raum *genannt.*

Beispiel 27. *Die stetigen Funktionen auf $\bar{\Omega}$ (mit $\Omega \subset \mathbb{R}^n$ offen und beschränkt) bilden den normierten Raum $C^0(\bar{\Omega})$ mit der Supremumsnorm $\|\cdot\|_\infty$.*

Definition 36. *Zwei Normen $\|\cdot\|^{(1)}$ und $\|\cdot\|^{(2)}$ auf X heißen* äquivalent*, wenn eine Konstante $0 < C < \infty$ existiert, mit*

$$\frac{1}{C}\|x\|^{(1)} \leq \|x\|^{(2)} \leq C\|x\|^{(1)} \quad \forall x \in X. \tag{5.1}$$

5.1.1.1 Operatoren

Definition 37. *(Operatornorm)*
Seien X und Y normierte Räume mit Normen $\|\cdot\|_X$ und $\|\cdot\|_Y$. Die Operatornorm *eines Operators $T\colon X \to Y$ ist definiert als*

$$\|T\| \coloneqq \sup_{x \in X} \left\{ \frac{\|Tx\|_Y}{\|x\|_X} \colon x \neq 0 \right\}. \tag{5.2}$$

Bemerkung 25. *Ist $\|T\|$ endlich, dann ist T beschränkt.*

Bemerkung 26. *Die beschränkten Operatoren bilden einen linearen Raum $L(X,Y)$ mit $(T_1 + T_2)\, x = T_1 x + T_2 x$.*

5.1.1.2 Offene Mengen

Definition 38. *$(X, \|\cdot\|)$ sein ein normierter Raum. $A \subset X$ heißt* offen*, falls für alle $x \in A$ ein $\varepsilon > 0$ existiert, sodass*

$$K_\varepsilon(x) \coloneqq \{y \in X \colon \|x - y\| < \varepsilon\}$$

in A enthalten ist.

5.1.2 Banach-Räume, Hilbert-Räume

Definition 39. *(Cauchy-Folge)*
Sei $(X, \|\cdot\|)$ ein normierter Raum. Eine Folge $\{x_n \in X \colon n \geq 1\}$ heißt Cauchy-Folge*, wenn gilt*

$$\sup\{\|x_n - x_m\| : n, m \geq k\} \to 0, \text{ für } k \to \infty \tag{5.3}$$

oder äquivalent:

$$\forall\, \varepsilon > 0\ \exists\, n_0(\varepsilon) \in \mathbb{N} \colon\ \forall\, n, m \geq n_0(\varepsilon) :\ \|x_n - x_m\| < \varepsilon. \tag{5.4}$$

Definition 40. *(Banach-Raum)*
Sei $(X, \|\cdot\|)$ ein normierter Raum. X heißt vollständig*, wenn jede Cauchy-Folge konvergiert. Ein normierter und vollständiger Raum heißt* Banach-Raum.

Definition 41. *(Skalarprodukt)*
Die Abbildung $(\cdot,\cdot) : X \times X \to \mathbb{K}$ heißt Skalarprodukt *auf X, falls gilt*

$$(x,x) > 0 \quad \forall x \in X, x \neq 0, \tag{5.5}$$

$$(\lambda x + y, z) = \lambda(x,z) + (y,z) \quad \forall \lambda \in \mathbb{K}, x, y, z \in X, \tag{5.6}$$

$$(x,y) = \overline{(y,x)} \quad \forall x, y \in X. \tag{5.7}$$

Das Skalarprodukt induziert eine Norm durch

$$\|x\| := \sqrt{(x,x)}.$$

Definition 42. *(Hilbertraum)*
Ein Banach-Raum X heißt Hilbert-Raum*, wenn ein Skalarprodukt $(\cdot,\cdot)$ auf X derart existiert, dass X bezüglich der durch $(\cdot,\cdot)$ induzierten Norm vollständig ist.*

5.1.3 Integrierbare Funktionen und das Lebesgue-Integral

Definition 43. *(Messraum)*
Sei X eine nicht-leere Menge. Man nennt eine Familie $\mathcal{A}$ von Teilmengen von X eine σ-Algebra *über X, wenn gilt:*

(i) $\emptyset \in \mathcal{A}$,

(ii) $A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}$,

(iii) $(A_n)_{n\in\mathbb{N}} \subset \mathcal{A} \Rightarrow \bigcup_{n\in\mathbb{N}} A_n \in \mathcal{A}$.

Das Paar $(X, \mathcal{A})$ heißt Messraum. *Elemente aus $\mathcal{A}$ heißen* messbar.

Beispiel 28. *Sei X ein topologischer Raum. Die Menge $\mathcal{O}$ aller offenen Mengen erzeugt die sogenannte Borelsche σ-Algebra (als die kleinste σ-Alegbra, die $\mathcal{O}$ enthält) und macht X damit auf natürliche Weise zu einem Messraum.*

Definition 44. *(Einfache Funktion)*
Sei $(X, \mathcal{A})$ ein Messraum. Eine Funktion $f\colon X \to \mathbb{R}$ heißt einfache Funtkion, *wenn es eine disjunkte Zerlegung $X = \dot{\bigcup}_{k=1}^{n} A_k$ mit $A_k \in \mathcal{A}$ $(k = 1, \dots n)$ gibt, sodass $f|_{A_k}$ für alle $k = 1 \dots n$ konstant ist.*

Definition 45. *(Messbare Abbildungen)*
Seien $(X, \mathcal{A})$ und $(Y, \mathcal{B})$ Messräume. Eine Abbildung $f\colon X \to Y$ heißt messbar, *wenn*

$$f^{-1}(B) \in \mathcal{A} \quad \forall\, B \in \mathcal{B}.$$

Ist $Y = \mathbb{R}$ versehen mit der Borelschen σ-Algebra, so ist f genau dann messbar, wenn eine Folge von einfachen Funktionen $(t_n)_{n\in\mathbb{N}}$ existiert, die punktweise gegen f konvergiert.

Definition 46. *(Maßraum)*
Sei $(X, \mathcal{A})$ ein Messraum. Eine Abbildung $\mu\colon \mathcal{A} \to [0, \infty]$ heißt Maß, *falls gilt*

(i) $\mu(\emptyset) = 0$,

(ii) Ist $(A_n)_{n\in\mathbb{N}} \subset \mathcal{A}$ mit $A_n \cap A_m = \emptyset$ für $n \neq m$, so gilt (σ-Additivität)

$$\mu\left(\dot{\bigcup_{n\in\mathbb{N}}} A_n\right) = \sum_{n\in\mathbb{N}} \mu(A_n).$$

Das Tripel $(X, \mathcal{A}, \mu)$ heißt dann Maßraum.

Beispiel 29. *Sei $\mathcal{I}^n$ die Menge (der Ring) der endlichen Vereinigungen paarweise disjunkter n-dimensionaler halboffener Quader $\prod_{k=1}^{n}(a_k, b_k] \subset \mathbb{R}^n$ mit $a_k \leq b_k$. Dann ist*

$$\mu_{\mathcal{I}^n}\colon \mathcal{I}^n \to [0, \infty] \qquad \text{mit } \mu_{\mathcal{I}^n}\left(\bigcup_{j=1}^{m}\prod_{k=1}^{n}(a_{jk}, b_{jk}]\right) = \sum_{j=1}^{m}\prod_{j=1}^{n}(b_{jk} - a_{jk})$$

ein „Maß“ auf $\mathcal{I}^n$.[1]
Die von $\mathcal{I}^n$ erzeugte σ-Algebra ist die Borelsche σ-Algebra $\mathcal{B}$ auf $\mathbb{R}^n$ und das „Maß“ $\mu_{\mathcal{I}^n}$ lässt sich eindeutig zu einem Maß μ auf $(\mathbb{R}^n, \mathcal{B})$ fortsetzen (Satz von Hahn), dem sog. Borel-Lebesgue-Maß.

Definition 47. *(Nullmenge)*
Sei $(X, \mathcal{A}, \mu)$ ein Maßraum. Eine Menge $A \in \mathcal{A}$ heißt Nullmenge, *wenn sie das Maß $\mu(A) = 0$ besitzt.*

Definition 48. *(fast überall)*
Sei X ein Maßraum und Y eine Menge. Zwei Funktionen $f, g \colon X \to Y$ heißen fast überall *gleich, falls eine Nullmenge N existiert, sodass*

$$f(x) \quad = \quad g(x) \ \forall\, x \in X \setminus N.$$

Definition 49. *(Integral über Treppenfunktionen)*
Sei $(X, \mathcal{A}, \mu)$ ein Maßraum und Y ein normierter Vektorraum. Eine Funktion $f \colon X \to Y$ heißt Treppenfunktion, *falls endlich viele disjunkte messbare Mengen $(A_k)_{k=1}^n \in \mathcal{A}$ mit $\mu(A_k) < \infty$ existieren, sodass $f|_{A_k}$ konstant ist und $f|_A = 0$ mit $A = \dot{\bigcup}_{k=1}^n A_k$. Der Raum der Treppenfunktionen von X nach Y wird mit $T(X, Y)$ bezeichnet.*
Man definiert dann das Integral für $t \in T(X, Y)$ als:

$$\int_X t \, d\mu \ := \ \sum_{k=1}^{n} t(A_k)\, \mu(A_k).$$

Definition 50. *(Integral)*
Sei $(X, \mathcal{A}, \mu)$ ein Maßraum. Man definiert den Raum der integrierbaren *Funktionen $\mathcal{L}^1(X)$ als die Menge aller Abbildungen $f \colon X \to \mathbb{R}$, für die eine Folge $(t_n)_{n\in\mathbb{N}} \subset T(X, \mathbb{R})$ von Treppenfunktionen existiert, die bezüglich der auf $T(X, \mathbb{R})$ definierten Seminorm*

$$|t|_1 \ := \ \int_X |t| \ d\mu$$

Cauchyfolge ist und fast überall punktweise gegen f konvergiert.
Man definiert dann das Integral für $f \in \mathcal{L}^1(X)$ als:

$$\int_X f \, d\mu \ := \ \lim_{k\to\infty} \int_X t_k \, d\mu.$$

Weiter definiert man $L^1(X)$ als den Raum der Äquivalenzklassen auf $\mathcal{L}^1(X)$ bezüglich der Äquivalenzrelation, in der zwei Elemente genau dann äquivalent sind, wenn sie fast

[1] $\mathcal{I}^n$ ist hier zwar kein Messraum, es sollen aber die gleichen Axiome für μ erfüllt sein wie für ein Maß auf einem Messraum.

überall gleiche Werte annehmen. D.h. für $f, g \in L^1(X)$*:* $f = g \Leftrightarrow f(x) = g(x)$ *fast überall. Der Raum* $L^1(X)$ *ist bezüglich der Norm*

$$\|f\|_1 := \int_X |f| \, d\mu$$

vollständig.

Bemerkung 27. *Ist* $\mu(X) < \infty$*, so sind die Treppenfunktionen genau die einfachen Funktionen und damit ist jede messbare Funktion integrierbar. Umgekehrt kann man zeigen, dass jede integrierbare Funktion auf einer Nullmenge so abgeändert werden kann, dass sie messbar ist. Also sind in diesem Fall die Elemente von* $L^1(X)$ *genau die messbaren Funktionen* $f : X \to \mathbb{R}$ *.*
Allgemeiner gilt: $f : X \to \mathbb{R}$ *ist integrierbar* $\Leftrightarrow$ f *ist messbar und* $|f|$ *ist integrierbar.*

Beispiel 30. *(Borel-Lebesgue-Integral)*
Der Integralbegriff, der hier im Folgenden verwendet wird, nutzt immer einen topologischen Raum $X \subset \mathbb{R}^n$*, versehen mit der* σ*-Algebra der Borel-Mengen* $\mathcal{B}$ *und dem Borel-Lebesge-Maß* μ*. Diesen Integralbegriff nennt man* Borel-Lebesgue-Integral*, oder auch einfach nur* Lebesgue-Integral*.*

5.1.4 Weitere Räume integrierbarer Funktionen

Neben dem Raum $L^1(X)$ der integrierbaren Funktionen gibt es noch weitere wichtige Funktionenräume integrierbarer Funktionen.

5.1.4.1 Der Raum $\mathbf{L}^\infty(\mathbf{D})$

Es sei $D \subset \mathbb{R}^n$ messbar (hier und im Folgenden will $(\mathbb{R}^n, \mathcal{B}, \mu)$ als Maßraum mit der Borelschen σ-Algebra und dem Borel-Lebesgue-Maß μ verstanden sein). $L^\infty(D)$ bezeichnet den Raum der auf D essentiell beschränkten integrierbaren Funktionen. $L^\infty(D)$ besteht aus Äquivalenzklassen, wobei

$$\begin{aligned} f &= g, \text{ falls } f = g \textit{ fast überall}; \\ \|u\|_{L^\infty(D)} &:= \inf_{\substack{A \in \mathcal{B} \\ \mu(A)=0}} \left\{ \sup_{x \in D \setminus A} \{|u(x)|\} \right\}. \end{aligned}$$

5.1.4.2 Der Hilbert-Raum $L^2(\Omega)$

Sei Ω eine offene Teilmenge von $\mathbb{R}^n$. $L^2(\Omega)$ definiert den Raum der Äquivalenzklassen *meßbarer* und *quadrat-integrabler* Funktionen:

$$L^2(\Omega) := \left\{ f : \Omega \to \mathbb{R} : f \text{ messbar}, |f|^2 \in L^1(\Omega) \right\}. \tag{5.8}$$

Zwei Funktionen f und g sind gleich, wenn sie bis auf in A mit $\mu(A) = 0$ gleich sind. Mit dem Skalarprodukt

$$(u, v)_0 = (u, v)_{L^2(\Omega)} \coloneqq \int_\Omega u\overline{v}\, d\mu \quad \forall u, v \in L^2(\Omega) \tag{5.9}$$

und der induzierten Norm

$$\|u\|_0 = \|u\|_{L^2(\Omega)} = \sqrt{\int_\Omega |u|^2 \, d\mu} \tag{5.10}$$

ist $L^2(\Omega)$ ein Hilbertraum und spielt damit eine besondere Rolle.

5.1.5 Schwache Differenzierbarkeit

In $L^2(\Omega)$ können keine Aussagen über die Differenzierbarkeit von $f \in L^2(\Omega)$ im klassischen Sinne gemacht werden. Aus der partiellen Integration folgernd gilt jedoch:

$$\int_\Omega f'(x)\varphi(x)dx = -\int_\Omega f(x)\varphi'(x)dx$$

für zwei stetig differenzierbare Funktionen f, φ mit $\varphi|_{\partial\Omega} = 0$. Mit Blick auf diese Eigenschaft erfolgt für (nicht notwendigerweise im klassischen Sinne differenzierbare) Funktionen in $L^2(\Omega)$ die folgende Definition:

Definition 51. *Sei $\Omega \subset \mathbb{R}^n$ offen.*

(i) Der Raum der Testfunktionen $C_c^\infty(\Omega)$ *sei definiert als der Raum aller unendlich oft stetig differenzierbaren Funktionen φ mit kompaktem Träger, d.h.*

$$C_c^\infty(\Omega) \coloneqq \left\{\varphi \in C^\infty(\Omega) : \overline{\{x \in \Omega \colon \varphi(x) \neq 0\}} \text{ ist kompakt}\right\}.$$

(ii) Falls für $f \in L^2(\Omega)$ eine Funktion $g \in L^2(\Omega)$ existiert, sodass gilt

$$\int_\Omega g(x)\varphi(x)\, dx = -\int_\Omega f(x)\varphi'(x)\, dx \quad \forall \varphi \in C_c^\infty(\Omega), \tag{5.11}$$

dann heißt g die schwache Ableitung *von f.*

Bemerkung 28. *Folgende Zusammenhänge bestehen zwischen klassischer und schwacher Differenzierbarkeit:*

- *Klassisch differenzierbare Funktionen sind auch schwach differenzierbar.*
- *Die schwache Ableitung ist durch die integrale Definition nicht punktweise definiert.*
- *Hinreichend oftmalige schwache Differenzierbarkeit impliziert klassische Differenzierbarkeit (s. Einbettungssätze von Sobolev).*

5.1.5.1 Höhere schwache Differenzierbarkeit

Sei $\alpha = (\alpha_1, \ldots, \alpha_n)$ ein Multiindex mit

$$\begin{aligned} |\alpha| &:= \sum_{i=1}^{n} \alpha_i, \\ D^\alpha &:= \frac{\partial^{|\alpha|}}{\partial^{\alpha_1} x_1 \ldots \partial^{\alpha_n} x_n} \end{aligned}$$

und $f \in L^2(\Omega)$. Dann heißt g die α-fache schwache Ableitung von f, wenn gilt

$$\int_\Omega g(x)\varphi(x)\,dx = (-1)^{|\alpha|} \int_\Omega f(x) D^\alpha \varphi(x)\,dx \quad \forall \varphi \in C_c^\infty(\Omega). \tag{5.12}$$

5.1.6 Die Hilbert-Räume $H^k(\Omega)$ und $H_0^k(\Omega)$

Die Menge aller Funktionen u aus $L^2(\Omega)$, die schwache Ableitungen $D^\alpha u \in L^2(\Omega)$ besitzen, bilden die *Sobolev-Räume*

$$H^k(\Omega) := \left\{ u \in L^2(\Omega) : D^\alpha u \in L^2(\Omega), |\alpha| \le k \right\} \tag{5.13}$$

mit $k \in \mathbb{N}_0$. $H^k(\Omega)$ wird in der Literatur auch als $W_2^k(\Omega)$ bzw. $W^{k,2}(\Omega)$ bezeichnet. $H^k(\Omega)$ ist ein Hilbert-Raum mit dem Skalarprodukt

$$(u,v)_k := (u,v)_{H^k(\Omega)} := \left(\sum_{|\alpha| \le k} \|D^\alpha u\|^2_{L^2(\Omega)} \right)^{\frac{1}{2}}. \tag{5.14}$$

Satz 15. *(Sobolevscher Einbettungssatz)*
Es gilt

$$H^s(\mathbb{R}^n) \subset C^k(\mathbb{R}^n), \text{ falls } s > k + \frac{n}{2},\ k \in \mathbb{N}_0. \tag{5.15}$$

5.1.7 Dualräume

Sei X ein normierter linearer Raum über $\mathbb{R}$. Mit X' wird der *Dualraum* bezeichnet, bestehend aus allen beschränkten linearen Abbildungen von X nach $\mathbb{R}$:

$$X' = L(X, \mathbb{R}).$$

Bemerkung 29. *Mit der Norm*

$$\|x'\| := \|x'\|_{\mathbb{R} \leftarrow X} := \sup \left\{ \frac{|x'(x)|}{\|x\|_X} : 0 \ne x \in X \right\}.$$

ist X' ein Banachraum. Elemente $x' \in X'$ heißen (stetige) lineare Funktionale *auf X und man schreibt auch:*

$$\langle x, x' \rangle_{X \times X'} := x'(x).$$

Lemma 15. *Seien X und Y normiert und $T \in L(X,Y)$. Für jedes $y' \in Y'$ definiert*

$$\langle Tx, y' \rangle_{Y \times Y'} = \langle x, x' \rangle_{X \times X'} \quad \forall x \in X \tag{5.16}$$

ein eindeutiges $x' \in X'$. Man kann also durch

$$\begin{aligned} T' \colon Y' &\longrightarrow X' \\ y' &\longmapsto x' \end{aligned}$$

den sog. dualen Operator definieren. Es gilt demnach $\langle Tx, y' \rangle_{Y \times Y'} = \langle x, T'y' \rangle_{X \times X'}$.

Lemma 16. *Es gilt*

$$\left\| T' \right\|_{X' \leftarrow Y'} = \|T\|_{Y \leftarrow X} . \tag{5.17}$$

Beweis. Einerseits folgt direkt aus der Definition der jeweiligen Normen

$$\begin{aligned} \left\| T' \right\|_{X' \leftarrow Y'} &= \sup_{y' \neq 0} \left\{ \frac{\|T'y'\|_{X'}}{\|y'\|_{Y'}} \right\} = \sup_{x, y' \neq 0} \left\{ \frac{\langle x, T'y' \rangle_{X \times X'}}{\|x\|_X \, \|y'\|_{Y'}} \right\} \\ &= \sup_{x, y' \neq 0} \left\{ \frac{\langle Tx, y' \rangle_{Y \times Y'}}{\|x\|_X \, \|y'\|_{Y'}} \right\} \leq \sup_{x, y' \neq 0} \left\{ \frac{\|T\|_{Y \leftarrow X} \, \|x\|_X \, \|y'\|_{Y'}}{\|x\|_X \, \|y'\|_{Y'}} \right\} = \|T\|_{Y \leftarrow X} \end{aligned}$$

und andererseits

$$\begin{aligned} \|T\|_{Y \leftarrow X} &= \sup_{x \neq 0} \left\{ \frac{\|Tx\|_Y}{\|x\|_X} \right\} \leq \sup_{x, y' \neq 0} \left\{ \frac{\langle Tx, y' \rangle_{Y \times Y'}}{\|x\|_X \, \|y'\|_{Y'}} \right\} \\ &= \sup_{x, y' \neq 0} \left\{ \frac{\langle x, T'y' \rangle_{X \times X'}}{\|x\|_X \, \|y'\|_{Y'}} \right\} \leq \sup_{x, y' \neq 0} \left\{ \frac{\|T'\|_{X' \leftarrow Y'} \, \|y'\|_{Y'} \, \|x\|_X}{\|x\|_X \, \|y'\|_{Y'}} \right\} = \|T'\|_{X' \leftarrow Y'}, \end{aligned}$$

wobei hier für die erste Ungleichung der Satz von Hahn-Banach benutzt wird, der die Existenz eines $y' \in Y'$ mit $\langle Tx, y' \rangle_{Y \times Y'} = \|Tx\|_Y$ und $\|y'\|_{Y'} = 1$ für ein beliebiges, aber festes $x \in X$ garantiert.

□

5.1.7.1 Adjungierte Operatoren

Sei X ein Hilbert-Raum über $\mathbb{R}$. Jedes $y \in X$ definiert durch

$$f_y(x) \coloneqq (x, y)_X \tag{5.18}$$

ein lineares Funktional $f_y \in X'$ mit $\|f_y\|_{X'} = \|y\|_X$. Umgekehrt definiert jedes Funktional f_y ein $y \in X$, wie folgender Satz zeigt:

Satz 16. *(Darstellungssatz von Riesz)*
X sei ein Hilbert-Raum und $f \in X'$ ein Funktional. Dann existiert genau ein $y_f \in X$, sodass

$$f(x) = (x, y_f)_X \quad \forall x \in X \text{ und es gilt } \|y_f\|_X = \|f\|_{X'} . \tag{5.19}$$

Beweis. **Existenz:**
Sei $N = \{x \in X \colon f(x) = 0\}$ der Kern von f. Ist $N = X$, folgt die Behauptung sofort mit $y_f = 0$. Wir können daher im Folgenden davon ausgehen, dass $N \neq X$.
Sei also $w \in X \setminus N$ und $d \coloneqq d(w, N) = \inf_{x \in N} \|w - x\|$der Abstand zwischen w und N. Dann gibt es eine Folge $(x_n)_{n \in \mathbb{N}}$ in N mit $d = \lim_{n \to \infty} \|w - x_n\|$. Wir zeigen zunächst, dass $(x_n)_{n \in \mathbb{N}}$ eine Cauchy-Folge ist.
Im Hilbert-Raum X gilt die Parallelogrammidentität:

$$\begin{aligned} &\|(w - x_m) + (w - x_n)\|^2 + \|(w - x_m) - (w - x_n)\|^2 = 2\left(\|w - x_m\|^2 + \|w - x_n\|^2\right) \\ \Leftrightarrow &\|x_m - x_n\|^2 = 2\left(\|w - x_m\|^2 + \|w - x_n\|^2\right) - 4\left\|w - \frac{1}{2}(x_m + x_n)\right\|^2. \end{aligned}$$

Da $\frac{1}{2}(x_m + x_n) \in N$, gilt $4\left\|w - \frac{1}{2}(x_m + x_n)\right\|^2 \geq 4d^2$. Sei $\varepsilon > 0$. Wähle m, n groß genug damit $2\left(\|w - x_m\|^2 + \|w - x_n\|^2\right) < 4d^2 + \varepsilon$. Dann ist

$$\|x_m - x_n\|^2 < 4d^2 + \varepsilon - 4d^2 = \varepsilon.$$

Also ist $(x_n)_{n \in \mathbb{N}}$ tatsächlich Cauchy. Da f stetig ist, ist N ein abgeschlossener Unterraum von X und damit konvergiert $(x_n)_{n \in \mathbb{N}}$ gegen ein $x^* \in N$ mit $\|w - x^*\| = d$. Seien nun $\lambda \in \mathbb{R}$ und $\tilde{x} \in N$ beliebig. Dann gilt

$$\begin{aligned} &d^2 \leq \|w - (x^* + \lambda\tilde{x})\|^2 = \|w - x^*\|^2 + \lambda^2\|\tilde{x}\|^2 - 2\lambda(w - x^*, \tilde{x}) \\ \Rightarrow &\lambda^2\|\tilde{x}\|^2 - 2\lambda(w - x^*, \tilde{x}) \geq 0. \end{aligned}$$

Da dies für alle $\lambda \in \mathbb{R}$ gilt, folgt

$$(w - x^*, \tilde{x}) = 0 \ \forall \tilde{x} \in N \tag{5.20}$$

(sonst Widerspruch mit $\lambda = \|\tilde{x}\|^{-2}(w - x^*, \tilde{x})$).
Setze $z = w - x^*$. Dann gilt für alle $x \in X$:

$$\begin{aligned} f\left(x - \frac{f(x)\,z}{f(z)}\right) &= f(x) - f\left(\frac{f(x)\,z}{f(z)}\right) \\ &= f(x) - \frac{f(x)}{f(z)} f(z) = 0 \\ &\Rightarrow \quad x - \frac{f(x)\,z}{f(z)} \in N. \end{aligned}$$

Eingesetzt für $\tilde{x}$ in (5.20) liefert dies

$$\begin{aligned} &\left(z, x - \frac{f(x)}{f(z)} z\right) = 0 \\ \Leftrightarrow &(z, x) - \frac{f(x)}{f(z)}(z, z) = 0 \\ \Leftrightarrow &f(x) = \frac{(x, z)}{\|z\|^2} f(z) = \left(x, \frac{f(z)\,z}{\|z\|^2}\right). \end{aligned}$$

Mit $y_f = \frac{f(z)}{\|z\|^2} z$ folgt also die Existenzbehauptung.

Eindeutigkeit: Sei $\tilde{y_f}$ ein weiteres Element mit

$$\begin{aligned} f(x) &= (x, \tilde{y_f}) = (x, y_f) \quad \forall x \in X \\ &\Rightarrow (x, \tilde{y_f} - y_f) = 0 \quad \forall x \in X \\ &\Rightarrow \tilde{y_f} - y_f = 0. \end{aligned}$$

□

5.1.7.2 Bilinearformen

Definition 52. *Sei V ein Hilbertraum. Die Abbildung $a(\cdot,\cdot) : V \times V \longrightarrow \mathbb{R}$ heißt Bilinearform, falls*

$$a(x + \lambda y, z) = a(x, z) + \lambda a(y, z), \tag{5.21}$$
$$a(x, y + \lambda z) = a(x, y) + \lambda a(x, z) \quad \forall \lambda \in \mathbb{R},\ x, y, z \in V. \tag{5.22}$$

Definition 53. *Die Bilinearform $a(\cdot,\cdot)$ heißt* stetig, *falls ein C_s existiert, sodass*

$$|a(x, y)| \leq C_s \|x\|_V \|y\|_V \quad \forall x, y \in V. \tag{5.23}$$

Lemma 17. *Zu einer stetigen Bilinearform existiert ein eindeutiger Operator $A \in L(V, V')$ mit*

$$a(x, y) = \langle Ax, y \rangle_{V' \times V} \quad \forall x, y \in V, \tag{5.24}$$
$$\|A\|_{V' \leftarrow V} \leq C_s. \tag{5.25}$$

Beweis. Man halte $x \in V$ fest.

$$\varphi_x(y) \coloneqq a(x, y)$$

definiert ein stetiges lineares Funktional $\varphi_x \in V'$ mit $\|\varphi_x\|_{V'} \leq C_s \|x\|_V$. Der Operator $A\colon V \to V'$ definiert durch

$$Ax \coloneqq \varphi_x$$

ist linear und es gilt

$$\begin{aligned} &\|Ax\|_{V'} \leq C_s \|x\|_V \\ \Rightarrow &\|A\|_{V' \leftarrow V} = \sup_{0 \neq x \in V} \left\{ \frac{\|Ax\|_{V'}}{\|x\|_V} \right\} \leq C_s. \end{aligned}$$

□

Definition 54. *(Elliptizität)*
Die Bilinearform $a(\cdot,\cdot)$ heißt elliptisch, *falls sie stetig ist und eine Konstante $C_E > 0$ existiert, sodass*

$$a(x, x) \geq C_E \|x\|_V^2 \quad \forall x \in V. \tag{5.26}$$

5.2 Variationsformulierung

Das Ziel in diesem Abschnitt ist es, über die Funktionalanalysis einen neuen Zugang zum Modellproblem zu finden. Ein äquivalentes Problem zum klassischen Problem definieren wir über eine Variationsformulierung. Diese wird Lösungen des Modellproblems in Hilberträumen angeben.

Satz 17. *(Charakterisierungssatz)*
Sei V ein Vektoraum und

$$a\colon V \times V \longrightarrow \mathbb{R}$$

eine symmetrische positiv definite Bilinearform. $f\colon V \longrightarrow \mathbb{R}$ sei ein lineares Funktional. Die Abbildung

$$J(v) \coloneqq \frac{1}{2}a(v,v) - f(v)$$

nimmt in V ihr Minimum genau dann bei u an, wenn gilt:

$$a(u,v) = f(v) \quad \forall v \in V.$$

Die Lösung u ist eindeutig.

Beweis. Seien $u, v \in V$, $t \in \mathbb{R}$. Betrachte

$$\begin{aligned} J(u+tv) &= \frac{1}{2}a(u+tv, u+tv) - f(u+tv) \\ &= \frac{1}{2}\left(a(u,u) + 2ta(u,v) + t^2 a(v,v)\right) - f(u) - tf(v) \\ &= J(u) + t\left(a(u,v) - f(v)\right) + \frac{1}{2}t^2 a(v,v) \end{aligned}$$

1. Gelte für $u \in V$: $a(u,v) = f(v)\ \forall v \in V$, dann:

$$\begin{aligned} \overset{t=1}{\Rightarrow}\ J(u+v) &= J(u) + \left(f(v) - f(v)\right) + \frac{1}{2}a(v,v) \\ &= J(u) + \frac{1}{2}a(v,v) > J(u)\ \ \forall v \in V. \end{aligned}$$

   Also ist u Minimalpunkt.

2. Sei $u \in V$ Minimalpunkt. Dann muss die Funktion $t \mapsto J(u+tv)$ für alle $v \in V$ in $t = 0$ ein Minimum annehmen. Aufgrund des notwendigen Minimalitätskriteriums gilt also:

$$\begin{aligned} &\Rightarrow 0 = \frac{dJ(u+tv)}{dt}|_{t=0} = a(u,v) - f(v) \\ &\Leftrightarrow a(u,v) = f(v). \end{aligned}$$

Die Eindeutigkeit der Lösung folgt sofort aus der positiven Definitheit. Sind nämlich u_1, u_2 zwei Lösungen, so gilt:

$$\begin{aligned} & a(u_1, v) = f(v) \ \wedge \ a(u_2, v) = f(v) \ \forall v \in V \\ \Rightarrow \ & a(u_1 - u_2, v) = 0 \ \forall v \in V \\ \Rightarrow \ & u_1 - u_2 = 0. \end{aligned}$$

□

5.2.1 Untersuchung des elliptischen Differentialoperators zweiter Ordnung

Wir betrachten den elliptischen Differentialoperator

$$Lu \coloneqq -\sum_{i,k=1}^{n} \partial_i \left(a_{ik} \partial_k u\right) + a_0 u \tag{5.27}$$

mit $a_0(x) \geq 0 \ (x \in \Omega)$ und $A = (a_{ik})_{i,k}$ symmetrisch positiv definit. Das zugehörige Problem

$$\begin{aligned} Lu &= f \quad \text{in } \Omega \\ u &= g \quad \text{auf } \partial\Omega \end{aligned}$$

mit $f \in L^2(\Omega)$ und $g \in H^{\frac{1}{2}}(\partial\Omega) \coloneqq \left\{v \in L^2(\partial\Omega) : \exists w \in H^1(\Omega), \gamma(w) = v\right\}$ (wobei hier γ der Spur-Operator ist) kann ohne Beschränkung der Allgemeinheit als homogenes Problem aufgefasst werden: Sei $\overline{g} \in H^1(\Omega)$ so gewählt, dass $\gamma(\overline{g}) = g$. Setze dann

$$\begin{aligned} w &\coloneqq u - \overline{g} \\ \Rightarrow \ Lw &= f - L\overline{g} =: f_1 \quad \text{in } \Omega \\ w &= 0 \quad \text{auf } \partial\Omega. \end{aligned}$$

Folgender Satz stellt den Zusammenhang zwischen Randwertaufgabe und Variationsproblem her:

Satz 18. *(Minimaleigenschaft)*
Jede klassische Lösung der Randwertaufgabe

$$-\sum_{i,k} \partial_i \left(a_{ik} \partial_k u\right) + a_0 u = f \text{ in } \Omega \tag{5.28}$$

$$u = 0 \text{ auf } \partial\Omega \tag{5.29}$$

ist Lösung des Minimierungsproblems

$$J(v) \coloneqq \int_{\Omega} \left(\frac{1}{2} \sum_{i,k} a_{ik} \partial_i v \partial_k v + \frac{1}{2} a_0 v^2 - fv \right) dx \to \min \tag{5.30}$$

unter allen Funktionen aus $C^2(\Omega) \cap C^0(\bar{\Omega})$ *mit Nullrandwerten.*

Beweis. Die Greensche Formel besagt

$$\int_\Omega v\,(\nabla\cdot\vec{w}) + \nabla v\cdot\vec{w}dx \;=\; \int_{\partial\Omega} v\,(\vec{w}\cdot\vec{n})\,ds.$$

Mit $v|_{\partial\Omega} = 0$ folgt

$$-\int_\Omega v\,(\nabla\cdot\vec{w})\,dx = \int_\Omega \nabla v\cdot\vec{w}dx$$

und mit $w_i = \sum_k a_{ik}\partial_k u$

$$-\int_\Omega v\sum_i \partial_i\left(\sum_k a_{ik}\partial_k u\right)dx = \int_\Omega \sum_{i,k} a_{ik}\partial_i v\partial_k u dx. \tag{5.31}$$

Setze

$$\begin{aligned} a\,(u,v) &:= \int_\Omega \sum_{i,k} a_{ik}\partial_i v\partial_k u + a_0 uv\,dx, \\ f\,(v) &:= \int_\Omega fv\,dx \end{aligned}$$

und erhalte so durch Multiplikation mit v und Integration aus (5.28) und (5.31) die Identität

$$\begin{aligned} a\,(u,v) - f\,(v) &= \int_\Omega v\left(-\sum \partial_i\,(a_{ik}\partial_k u) + a_0 u - f\right)dx \\ &= \int_\Omega v(Lu-f)dx \underset{\text{falls } Lu=f}{=} 0. \end{aligned}$$

Man zeigt leicht, dass die so definierte symmetrische Bilinearform $a\,(\cdot,\cdot)$ positiv definit ist und f linear. Aus dem Charakterisierungssatz folgt also die behauptete Minimalisierungseigenschaft.

□

Man erhält also folgende Aussage:

Ist u Lösung des Minimierungsproblems **und**

$$u \in C^2(\Omega)\cap C^0(\bar{\Omega})$$

$$\Downarrow$$

u ist klassische Lösung.

Wir klären nun die Frage nach der Existenz einer Lösung.

5.2.2 Existenz und Eindeutigkeit für das Variationsproblem

Betrachte das Dirichlet-Integral (Energiefunktional)

$$J(u) \coloneqq \int_{\Omega} |\nabla u|^2 \, dx. \tag{5.32}$$

Aus der Minimaleigenschaft folgt

$$\begin{aligned} J(u) \to \min \Longleftrightarrow -\Delta u &= 0 \text{ in } \Omega, \\ u &= \varphi \text{ auf } \Gamma. \end{aligned}$$

5.2.2.1 Dirichlet-Prinzip

Dirichlet argumentierte, da $J(u) \geq 0$, muss für ein u das Minimum angenommen werden.
$\Rightarrow$ Es existiert eine Lösung zu

$$\begin{aligned} -\Delta u &= 0 \text{ in } \Omega, \\ u &= \varphi \text{ auf } \Gamma. \end{aligned}$$

Einen Widerspruch zu dieser Aussage demonstrierte Weierstraß (1870):
Betrachte dazu

$$J(u) = \int_0^1 u^2(x) \, dx \longrightarrow \min \text{ für } u \in C^0([0,1])$$

und $u(0) = 1$, $u(1) = 0$.

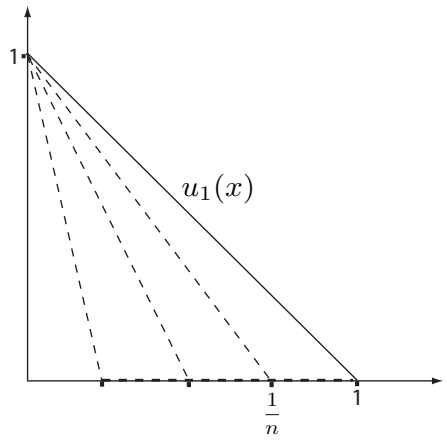

Abbildung 5.1: Folgenfunktionen $u_n(x)$

Wir definieren nun die Funktionenfolge

$$u_n(x) = \begin{cases} 1 - nx & 0 \leq x \leq \frac{1}{n}, \\ 0 & x > \frac{1}{n}. \end{cases} \tag{5.33}$$

Es gilt $\lim_{n\to 0} u_n = 0$, d.h. das Infimum von $J(u)$ ist

$$\inf J(u) = 0,$$

wird aber für stetige Funktionen nicht angenommen. Dies ist ein Widerspruch zum Dirichlet-Prinzip. Dieses Beispiel zeigt: Um eine eindeutige Lösung des Variationsproblems zu garantieren, muss der Funktionenraum richtig gewählt werden.

5.2.2.2 Existenz und Eindeutigkeit

Satz 19. *(Lax-Milgram)*
Sei V eine abgeschlossene, konvexe Menge im Hilbertraum H und $a\colon H \times H \longrightarrow \mathbb{R}$ eine stetige elliptische Bilinearform. Für jede stetige Linearform $l \in H'$ hat das Variationsproblem

$$J(v) \coloneqq \frac{1}{2} a(v,v) - \langle l, v\rangle \longrightarrow \min \tag{5.34}$$

genau eine Lösung in V.

Beweis. J ist nach unten beschränkt:

$$\begin{aligned} J(v) &\geq \frac{1}{2} C_E \|v\|^2 - \|l\| \|v\| \\ &= \frac{1}{2C_E}\left(C_E^2 \|v\|^2 - 2C_E \|l\| \|v\| + \|l\|^2\right) - \frac{1}{2C_E}\|l\|^2 \\ &= \frac{1}{2C_E}\left(C_E \|v\| - \|l\|\right)^2 - \frac{1}{2C_E}\|l\|^2 \geq -\frac{\|l\|^2}{2C_E}. \end{aligned}$$

Wir setzen nun $c_1 := \inf\{J(v) : v \in V\}$. Sei $(v_n)_{n\in\mathbb{N}}$ eine Minimalfolge, d.h.

$$\lim_{n\to\infty} J(v_n) = c_1.$$

Zunächst soll gezeigt werden, dass $(v_n)_{n\in\mathbb{N}}$ eine Cauchy-Folge ist. Es gilt aufgrund der Elliptizität und der Bilinearität von $a(\cdot,\cdot)$:

$$\begin{aligned} C_E \|v_n - v_m\|^2 \leq\ & a(v_n - v_m, v_n - v_m) \\ =\ & 2a(v_n, v_n) + 2a(v_m, v_m) - a(v_n + v_m, v_n + v_m). \end{aligned}$$

Setzt man im letzten Ausdruck für $a(v,v)$ jeweils $a(v,v) = 2J(v) + 2\langle l, v\rangle$, so erhält man

$$\begin{aligned} C_E \|v_n - v_m\|^2 \leq\ & 4J(v_n) + 4\langle l, v_n\rangle + 4J(v_m) + 4\langle l, v_m\rangle \\ & - \left(8J\left(\frac{v_n + v_m}{2}\right) + 4\langle l, v_n + v_m\rangle\right) \\ =\ & 4J(v_n) + 4J(v_m) - 8J\left(\frac{v_n + v_m}{2}\right) \\ \leq\ & 4J(v_n) + 4J(v_m) - 8c_1. \end{aligned}$$

Die letzte Ungleichung gilt, weil V konvex ist, also $\frac{v_n+v_m}{2} \in V$ und damit per Definition des Infimums $J\left(\frac{v_n+v_m}{2}\right) > c_1$.
Da aber sowohl $J(v_n) \to c_1 \quad (n \to \infty)$ als auch $J(v_m) \to c_1 \quad (m \to \infty)$, folgt

$$C_E \|v_n - v_m\|^2 \longrightarrow 0 \text{ für } n, m \to \infty,$$

also ist $(v_n)_{n\in\mathbb{N}}$ eine Cauchy-Folge in H. H ist ein Hilbertraum, deshalb existiert ein $u \in H$ (und da V abgeschlossen ist, auch $u \in V$) mit $\lim_{n\to\infty} v_n = u$ und $J(u) = \lim_{n\to\infty} J(v_n) = \inf_{v\in V} J(v)$.
$\Rightarrow J(v) = \frac{1}{2}a(v,v) - \langle l, v\rangle \longrightarrow \min$ hat eine Lösung $u \in V$.

Zu zeigen bleibt die Eindeutigkeit der Lösung: Seien u_1 und u_2 Lösungen, also Grenzwerte von Minimalfolgen. Dann ist auch $u_1, u_2, u_1, u_2, \ldots$ eine Minimalfolge. Daraus folgte wie oben: $u_1, u_2, u_1, u_2, \ldots$ ist Cauchy-Folge. Dies ist aber nur möglich, wenn $u_1 = u_2$.
□

Bemerkung 30. *Eine besondere Rolle spielen die folgenden beiden Spezialfälle:*

- *Mit $V = H$ folgt: Zu jedem $l \in H'$ gibt es genau ein $u \in H$ mit*

$$a(u,v) = \langle l, v\rangle \quad \forall v \in H.$$

- *Für $a(u,v) := (u,v)$ bekommt man gerade die Aussage des Rieszschen Darstellungssatzes: Zu jedem $l \in H'$ gibt es ein $u \in H$ mit*

$$(u,v) = \langle l, v\rangle \quad \forall v \in H.$$

  Damit erhält man eine Einbettung

$$\begin{aligned} H' &\longrightarrow H \\ l &\longmapsto u. \end{aligned}$$

5.2.3 Schwache Lösung des Randwertproblems

Definition 55. *Eine Funktion $u \in H_0^1(\Omega)$ heißt* schwache Lösung *der Randwertaufgabe 2. Ordnung*

$$\begin{aligned} Lu &= f \text{ in } \Omega, \\ u &= 0 \text{ auf } \Gamma \end{aligned}$$

mit L wie in (5.27)*, wenn gilt:*

$$a(u,v) = (f,v)_0 \quad \forall v \in H_0^1(\Omega) \tag{5.35}$$

$$\text{mit } a(u,v) := \int_\Omega \sum_{i,k} a_{ik}\partial_i u \partial_k v + a_0 uv \, dx. \tag{5.36}$$

Satz 20. *Sei L ein Differentialoperator 2. Ordnung wie in* (5.27)*. Dann hat das Randwertproblem*

$$\begin{aligned} Lu &= f \text{ in } \Omega, \\ u &= 0 \text{ auf } \Gamma \end{aligned}$$

mit $f \in L^2(\Omega)$ stets eine schwache Lösung in $H_0^1(\Omega)$ und diese ist das Minimum des Problems

$$\frac{1}{2} a(v,v) - (f,v)_0 \longrightarrow \min \text{ in } H_0^1(\Omega).$$

Beispiel 31. *Für das Modellproblem*

$$\begin{aligned} -\Delta u &= f \text{ in } \Omega, \\ u &= 0 \text{ auf } \Gamma \end{aligned}$$

ist die zugehörige Bilinearform

$$a(u,v) = \int_\Omega \nabla u \nabla v \, dx.$$

Das Variationsproblem lässt sich dann folgendermaßen stellen:
Finde $u \in H_0^1(\Omega)$ so, dass

$$(\nabla u, \nabla v)_0 = (f,v)_0 \quad \forall v \in H_0^1(\Omega).$$

Die Lösung u findet man auch als Lösung des Minimierungsproblems

$$\frac{1}{2} \int_\Omega \nabla u \nabla v \, dx - (f,v)_0 \longrightarrow \min \text{ in } H_0^1(\Omega).$$

5.2.4 Variationsproblem der Neumann-Randwertaufgabe

Wir betrachten das Problem

$$\begin{aligned} Lu &= f \text{ in } \Omega, \\ \sum_{i,k} n_i a_{ik} \partial_k u &= g \text{ auf } \Gamma, \end{aligned}$$

wobei n_i der i-te Anteil des lokalen Normalenvektors ist. Mit $f \in L_2(\Omega)$ und $g \in L_2(\Gamma)$ ist das lineare Funktional

$$\langle l, v \rangle := \int_\Omega f v \, dx + \int_\Gamma g v \, dx$$

definiert. Das Variationsproblem hat dann die Gestalt:
Finde $u \in H^1(\Omega)$ so, dass

$$\frac{1}{2} a(u,v) = (f,v)_{0,\Omega} + (g,v)_{0,\Gamma} \quad \forall v \in H^1(\Omega).$$

Satz 21. *Sie Ω ein beschränktes Gebiet mit stückweise glattem Rand, auf dem die Kegelbedingung*[2] *erfüllt ist, und sei ferner die* Kompatibilitätsbedingung

$$\int_{\Omega} f(x)\,dx + \int_{\partial\Omega} g(x)\,ds = 0$$

erfüllt. Man definiere

$$V := \left\{ v \in H^1(\Omega) : \int_{\Omega} v(x)\,dx = 0 \right\}.$$

Die Variationsaufgabe

$$J(v) := \frac{1}{2} a(u,v) - (f,v)_{0,\Omega} - (g,v)_{0,\Gamma} \longrightarrow \min \text{ in } V$$

ist eindeutig lösbar und die Lösung u erfüllt

$$\begin{aligned} Lu &= f \text{ in } \Omega, \\ \sum_{i,k} n_i a_{ik} \partial_k u &= g \text{ auf } \Gamma, \end{aligned}$$

falls $u \in C^2(\Omega) \cap C^1(\bar{\Omega})$.

Beachte: Die Variationsaufgabe hat in $H^1(\Omega)$ eine eindeutige Lösung, diese ist aber nicht notwendigerweise in $C^2(\Omega) \cap C^1(\bar{\Omega})$.

5.3 Galerkin-Verfahren

Beispiel 32. *Sei die Randwertaufgabe*

$$\begin{aligned} -\Delta u &= f \text{ in } \Omega, \\ u &= 0 \text{ auf } \Gamma \end{aligned}$$

gegeben. Die schwache Formulierung lautet:
Suche $u \in H_0^1(\Omega)$ mit

$$a(u,v) = (f,v) \quad \forall v \in H_0^1(\Omega)$$

mit

$$\begin{aligned} a(u,v) &:= \int_{\Omega} \nabla u \nabla v\,dx, \\ (f,v) &:= \int_{\Omega} f v\,dx. \end{aligned}$$

[2] Kegelbedingung: An jedem Punkt des Randes kann ein Kegel mit der Spitze so angelegt werden, dass er vollständig außerhalb des Gebietes ist, *und* so, dass er vollständig innerhalb ist.

Problem. *Hier muss ein Minimierungsproblem im unendlich-dimensionalen Raum $H_0^1(\Omega)$ gelöst werden. Für die numerische Behandlung des Minimierungsproblems benötigen wir einen endlich-dimensionalen Raum.*
Lösungsidee: *Ersetze den Lösungsraum durch einen endlich-dimensionalen (üblicherweise: Unter-) RaumV_h $\left(\subset H_0^1(\Omega)\right)$. Damit wird das obige Minimierungsproblem zu*

$$J(v) \coloneqq \frac{1}{2} a(v,v) - \langle l, v\rangle \longrightarrow \min \text{ in } V_h. \tag{5.37}$$

$u_h \in V_h$ ist genau dann Lösung des Minimierungsproblems, wenn gilt

$$a(u_h, v) = \langle l, v\rangle \quad \forall v \in V_h.$$

5.3.0.1 Das endliche Problem

Sei $\{\psi_1, \psi_2, \ldots, \psi_N\}$ eine Basis von V_h. Dann ist

$$a(u_h, v) = \langle l, v\rangle \quad \forall v \in V_h$$

aufgrund der Linearität von $a(\cdot,\cdot)$ und $l(\cdot)$ äquivalent zu

$$a(u_h, \psi_i) = \langle l, \psi_i\rangle \quad \forall i = 1, 2, \ldots, N. \tag{5.38}$$

Auch die diskrete Lösung $u_h \in V_h$ lässt sich als Linearkombination der ψ_i darstellen:

$$u_h = \sum_{k=1}^{N} z_k \psi_k \tag{5.39}$$

mit zu berechnenden Koeffizienten z_k. Setzt man (5.39) in (5.38) ein, ergibt sich ein Gleichungssystem:

$$\sum_{k=1}^{N} a(\psi_k, \psi_i) z_k = \langle l, \psi_i\rangle \quad i = 1, 2, \ldots N.$$

Mit $A_{ik} \coloneqq a(\psi_k, \psi_i)$ und $b_i \coloneqq \langle l, \psi_i\rangle$ lässt sich das Gleichungssystem schreiben als

$$Az = b.$$

Bemerkung 31. *Falls $a(\cdot,\cdot)$ elliptisch ist, so ist die Matrix A positiv definit:*

$$\begin{aligned} z^T A z &= \sum_{i,k} z_i A_{ik} z_k = a\left(\sum_k z_k \psi_k, \sum_i z_i \psi_i\right) \\ &= a(u_h, u_h) \geq C_E \|u_h\|_V^2 . \end{aligned}$$

Es stellt sich nun die Frage, wie gut die diskrete Lösung $u_h \in V_h$ die korrekte Lösung $u \in V$ approximiert. Eine erste Antwort darauf liefert das folgende Lemma:

Lemma 18. *(Céa-Lemma)*
Sei V ein Hilbertraum. Die Bilinearform $a\colon V \times V \to \mathbb{R}$ sei stetig (mit Konstante C_S) und elliptisch (mit Konstante C_E), $l \in V'$ und $V_h \subset V$ ein Unterraum.
Ferner sei $u \in V$ die eindeutige Lösung des Problems

$$a(u, v) = l(v) \ \forall v \in V \tag{5.40}$$

und $u_h \in V_h$ die eindeutige Lösung des Problems

$$a(u_h, v) = l(v) \ \forall v \in V_h. \tag{5.41}$$

Dann gilt:

$$\|u - u_h\| \leq \frac{C_S}{C_E} \inf_{v_h \in V_h} \|u - v_h\|. \tag{5.42}$$

Beweis. Da $V_h \subset V$, gilt (5.40) insbesondere auch für alle $v \in V_h$, daher kann man (5.40) und (5.41) voneinander abziehen:

$$a(u, v) - a(u_h, v) = a(u - u_h, v) = 0 \ \forall v \in V_h. \tag{5.43}$$

Diese Gleichung nennt man „Galerkin-Orthogonalität“.
Da $a(\cdot, \cdot)$ elliptisch ist, gilt

$$C_E \|u - u_h\|^2 \leq a(u - u_h, u - u_h).$$

Nun kann man auf der rechten Seite den Term $a(u - u_h, u_h - v_h)$ für in beliebiges $v_h \in V_h$ addieren, da dieser nach der Galerkin-Orthogonalität (5.43) immer Null ist:

$$\begin{aligned} C_E \|u - u_h\|^2 &\leq a(u - u_h, u - u_h) + a(u - u_h, u_h - v_h) \\ &= a(u - u_h, u - v_h) \\ &\leq C_S \|u - u_h\| \|u - v_h\| \quad \forall v_h \in V_h. \end{aligned}$$

Zuletzt wurde die Stetigkeit der Bilinearform benutzt. Daraus folgt direkt

$$\begin{aligned} \|u - u_h\| &\leq \frac{C_S}{C_E} \|u - v_h\| \quad \forall v_h \in V_h \\ \Rightarrow \|u - u_h\| &\leq \frac{C_S}{C_E} \inf_{v_h \in V_h} \|u - v_h\|. \end{aligned}$$

□

Bemerkung 32.

- *Die Galerkin-Orthogonalität* (5.43) *besagt, dass der Approximationsfehler $u - u_h$ a-orthogonal auf dem Ansatzraum V_h steht.*
- *Das Céa-Lemma garantiert, dass die approximative Lösung bis auf einen Faktor $\frac{C_S}{C_E}$ die bestmögliche Lösung im Ansatzraum V_h ist. Um die Güte der Lösung zu gewährleisten, muss man also nur noch einen Ansatzraum wählen, dessen Abstand zur Lösung möglichst gering ist.*
- *Ist die Bilinearform a symmetrisch, definiert $\|v\|_a \coloneqq (a(v, v))^{\frac{1}{2}}$ eine Norm auf V und bezüglich dieser Norm sind die Konstanten C_E und C_S jeweils 1. Bezüglich der Norm $\|\cdot\|_a$ ist u_h dann die garantiert beste Approximation für u in V_h.*

5.4 Finite Elemente Verfahren

Die Frage nach einer geeigneten Basis für V_h beantwortet das Finite Elemente Verfahren durch

1. Polynom-Basisfunktionen niedriger Ordnung

2. lokale Funktionen, d.h. Funktionen mit kompaktem Träger

Die Approximationseigenschaften von Funktionen mit niedriger Ordnung sind zwar nur in kleinen Bereichen gut, durch die Einfachheit der Funktionen kann eine bessere globale Approximation jedoch durch Gitterverfeinerung, auch in Kombination mit Erhöhung der Polynomordnung, erreicht werden. Durch Funktionen mit kompaktem Träger lassen sich die Einträge der Steifigkeitsmatrix und der rechten Seite durch lokale Integralapproximation einfach lösen und so das zu lösende lineare Gleichungssystem aufstellen. Dieses kann mitunter sehr groß werden, da die Gitterfeinheit aufgrund der obigen Grundannahmen für die Basisfunktionen hoch werden muss. Courant demonstrierte 1943 das folgende grundlegende Beispiel für das Vorgehen bei der Finite-Elemente-Methode.

5.4.1 Beispiel von Courant

Betrachte das Poisson-Problem

$$\begin{aligned} -\Delta u &= f \text{ in } \Omega = (0,1) \times (0,1), \\ u &= 0 \text{ auf } \Gamma. \end{aligned}$$

Das Gebiet Ω wird dabei in regelmäßiger Weise in Dreiecke zerlegt, sodass die Triangulierung lokal überall so geartet ist wie auf der linken Seite von Abbildung 5.2, mit den Knoten L, R, O, U, LO, RU und Z.

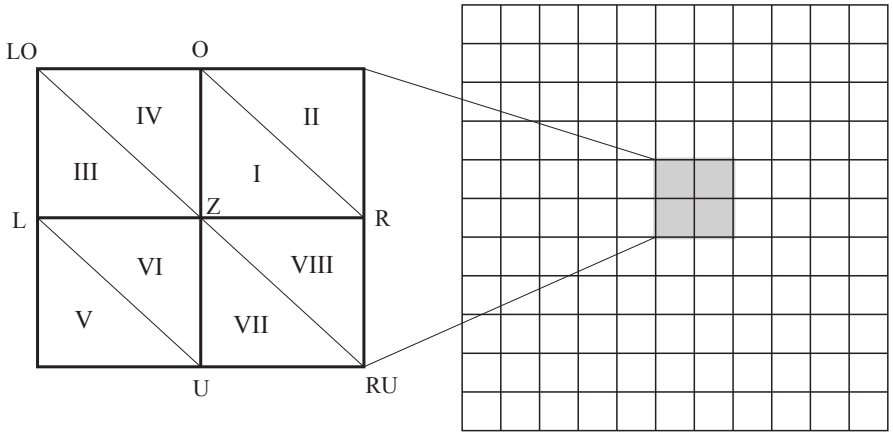

Abbildung 5.2: Unterteilung des Gebiets in Dreiecke

	I	II	III	IV	V	VI	VII	VIII
$\partial_1 \psi_Z$	$\frac{-1}{h}$	0	$\frac{1}{h}$	0	0	$\frac{1}{h}$	0	$\frac{-1}{h}$
$\partial_2 \psi_Z$	$\frac{-1}{h}$	0	0	$\frac{-1}{h}$	0	$\frac{1}{h}$	$\frac{1}{h}$	0

Tabelle 5.1: *Ableitungen von ψ_Z in die erste und zweite Raumrichtung*

Wir wählen V_h folgendermaßen:

$$V_h = \left\{ v \in C\left(\bar{\Omega}\right) : v \text{ linear auf jedem Dreieck der Triangulierung und } v|_\Gamma = 0 \right\}.$$

v lässt sich also in jedem Gitterdreieck darstellen als

$$v\left(x, y\right) = a + bx + cy.$$

Wenn man die Werte in den Gitterknoten kennt, lassen sich a, b und c eindeutig bestimmen. Bei N inneren Gitterknoten ist $\dim V_h = N$. Wir benötigen also N Basisfunktionen für V_h. Die Basisfunktionen $\{\psi_i\}_{i=1}^N$ seien definiert durch

$$\psi_i\left(K_j\right) = \delta_{ij}, \text{ wo } K_j \text{ der } j\text{-te Gitterknoten ist, } \forall i, j = 1, \ldots, N.$$

Die Basisfunktion über dem inneren Knoten Z ist damit

1. linear in jedem Dreieck,

2. Null auf den umliegenden Knoten,

erfüllt also die weist also die Merkmale „niedrige Ordnung“ und „Lokalität“ für Finite Elemente auf. Wir können nun die Gradienten der Basisfunktion ψ_Z in den Dreiecken I-VIII berechnen (siehe Tabelle 5.4.1).

Zu lösen ist das Gleichungssystem

$$Au = b$$

mit

$$A_{ij} = a\left(\psi_i, \psi_j\right).$$

In unserem Beispiel benötigen wir also $a\left(\psi_Z, \psi_Z\right)$, $a\left(\psi_Z, \psi_O\right)$, $a\left(\psi_Z, \psi_U\right)$, $a\left(\psi_Z, \psi_L\right)$, $a\left(\psi_Z, \psi_R\right)$, $a\left(\psi_Z, \psi_{LO}\right)$ und $a\left(\psi_Z, \psi_{RU}\right)$.

Für $a(\psi_Z, \psi_Z)$ gilt:

$$\begin{aligned} a(\psi_Z, \psi_Z) &= \int_{\Omega} (\nabla\psi_Z)^2 \, dxdy = \int_{\text{I–VIII}} (\nabla\psi_Z)^2 \, dxdy \\ &= 2 \int_{\text{I,III,IV}} \left((\partial_1\psi_Z)^2 + (\partial_2\psi_Z)^2 \right) dxdy \\ &= 2 \int_{\text{I,III}} (\partial_1\psi_Z)^2 \, dxdy + 2 \int_{\text{I,IV}} (\partial_2\psi_Z)^2 \, dxdy \\ &= \frac{2}{h^2} \int_{\text{I,III}} dxdy + \frac{2}{h^2} \int_{\text{I,IV}} dxdy \\ &= 4. \end{aligned}$$

Für $a(\psi_Z, \psi_O)$ gilt:

$$\begin{aligned} a(\psi_Z, \psi_O) &= \int_{\text{I–VIII}} \nabla\psi_Z \nabla\psi_O dxdy \\ &= \int_{\text{I,IV}} \nabla\psi_Z \nabla\psi_O dxdy = \int_{\text{I,IV}} \partial_1\psi_Z \partial_1\psi_O + \partial_2\psi_Z \partial_2\psi_O dxdy \\ &= \int_{\text{I,IV}} \partial_2\psi_Z \partial_2\psi_O dxdy = \int_{\text{I,IV}} -\frac{1}{h} \cdot \frac{1}{h} dxdy \\ &= -\frac{1}{h^2} \int_{\text{I,IV}} dxdy = -1. \end{aligned}$$

Aus Symmetriegründen folgt

$$a(\psi_Z, \psi_O) = a(\psi_Z, \psi_U) = a(\psi_Z, \psi_L) = a(\psi_Z, \psi_R) = -1.$$

Durch Nachrechnen erhält man zudem

$$a(\psi_Z, \psi_{RU}) = a(\psi_Z, \psi_{LO}) = 0.$$

Damit erhalten wir einen Stern der Form

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

genau wie bei der Diskretisierung durch Finite-Differenzen mit der Fünfpunkt-Formel (bis auf den Vorfaktor h^{-2}).

Bemerkung 33. *Die Identität der Sterne und damit der Systemmatrizen zwischen Finite-Differenzen- und Finite-Elemente-Verfahren gilt im Allgemeinen* nicht.

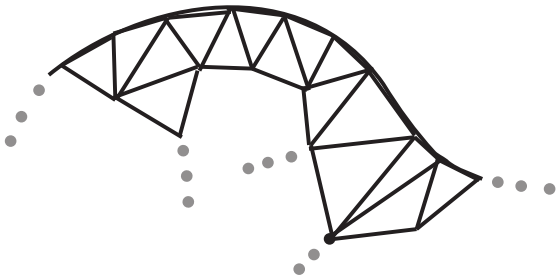

Abbildung 5.3: Triangulierung eines Gebiets

5.4.2 Triangulierung

Ein Gebiet mit gekrümmten Rändern kann lokal linear approximiert werden.

Definition 56. *(Zulässige Triangulierung)*
Eine Zerlegung $\mathcal{T} = \{T_1, T_2, \ldots, T_M\}$ *von* Ω *in (abgeschlossene) Dreiecks- bzw. Viereckselemente heißt* zulässig*, wenn folgende Eigenschaften erfüllt sind:*

(i) $\bar{\Omega} = \bigcup_{i=1}^{M} T_i$.

(ii) Besteht $T_i \cap T_j$ *aus genau einem Punkt, dann ist dieser ein Eckpunkt von* T_i *und* T_j.

(iii) Besteht die Schnittmenge $T_i \cap T_j$ $(i \neq j)$ *aus mehr als ein Punkt, dann ist* $T_i \cap T_j$ *eine Kante von* T_i *und* T_j.

Die Punkte 2 und 3 definieren *konforme Gitter*.

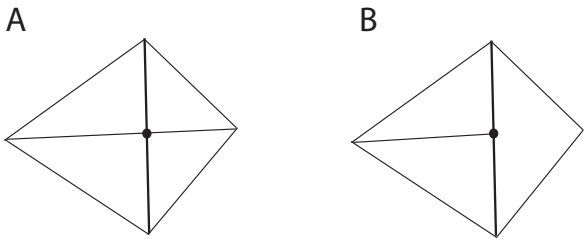

Abbildung 5.4: A: konformes Gitter, B: nicht konformes Gitter

In drei Raumdimensionen kommen verschiedene Gitterelemente zum Einsatz (Abb. 5.5).

Definition 57. *(Finite Elemente Raum)*
Für ein konformes Gitter Ω_h *über* $\Omega \subset \mathbb{R}^d$ *ist*

$$V_h^p(\mathcal{T}) = \left\{ u \in H^1 : \text{ für alle } T \in \mathcal{T} : u|_T \in \mathbb{P}_p \right\}$$

der konforme Finite-Elemente-Raum der Ordnung p.

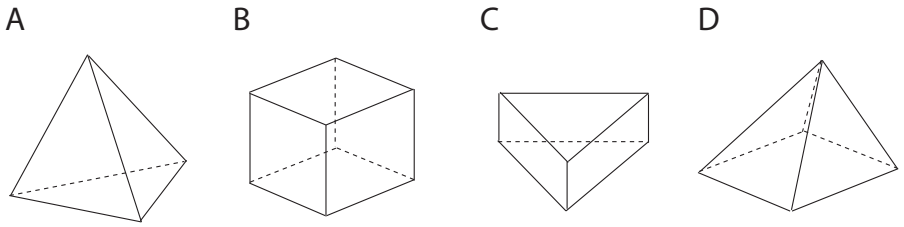

Abbildung 5.5: A: Tetraeder, B: Quader, C: Prisma, D: Pyramide

Aus dem Beispiel von Courant können wir ein allgemeines Verfahren im $\mathbb{R}^d\,(d = 1, 2, 3)$ ableiten.

5.4.3 Finite Elemente im $\mathbb{R}^1$

Wir betrachten als Modellproblem die Helmholtz-Gleichung

$$-\Delta u + u = f \text{ mit } a(u, v) = \int_\Omega (\nabla u \nabla v + uv)\, dx. \tag{5.44}$$

Sei $\mathcal{N} = \{a = x_0, x_1, x_2, \ldots, x_{N+1} = b\}$ die Diskretisierung des Gebiets $[a, b]$ mit Gitterweiten $h_i = x_{i+1} - x_i$ und lokalen Ansatzfunktionen φ_i, für die gilt

$$\varphi_i(x_j) = \delta_{ij}.$$

Damit hat die Lösung u_h die Gestalt

$$u_h = \sum_{i=1}^{N} a_i \varphi_i$$

mit $a_i = u_h(x_i)$.

5.4.3.1 Berechnung von u_h auf Referenzintervallen

Die Ansatzfunktionen φ_i können durch Formfunktionen Φ_i dargestellt werden.
Durch eine bijektive affine Transformation wird das Intervall $[x_i, x_{i+1}]$ auf das Referenzintervall $[0, 1]$ abgebildet. Für alle elementabhängigen Rechnungen (Auswertung von Funktionen, Gradienten, Integralen) kann man sich auf diese Weise auf das Intervall $[0, 1]$ mit den Formfunktionen Φ_i zurückziehen, dort vergleichsweise leicht auswerten und dann zurücktransformieren. Dies hat technische Vorteile gegenüber der Berechnung auf dem Originalelement.

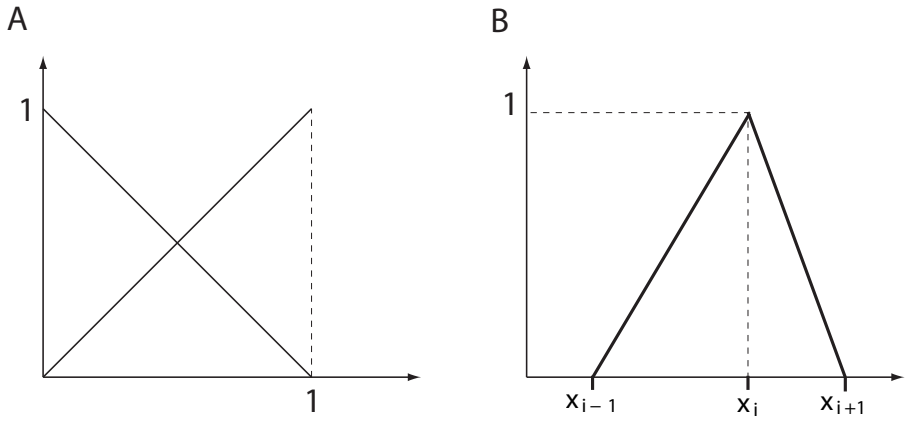

Abbildung 5.6: A: Formfunktionen über Referenzintervall, B: Ansatzfunktionen über Gitter

5.4.3.2 Transformationsabbildung

Sei $I_i = [x_i, x_{i+1}]$ und $\xi \in [0,1]$. Dann definieren

$$\begin{aligned} x_{I_i} \colon [0,1] &\longrightarrow I_i, \\ \xi &\longmapsto x_i + h_i \xi; \\ \xi_{I_i} \colon I_i &\longrightarrow [0,1], \\ x &\longmapsto \frac{(x - x_i)}{h_i} \end{aligned}$$

eine Bijektion zwischen $[0,1]$ und $[x_i, x_{i+1}]$. Auf dem Referenzintervall können wir unsere Lösung darstellen als

$$u_h(\xi) = \alpha_1 + \alpha_2 \xi.$$

Dabei gilt $u_i = u_h(0) = \alpha_1$ und $u_{i+1} = u_h(1) = \alpha_1 + \alpha_2$. Daraus folgt für $\xi \in [0,1]$

$$\begin{aligned} u_h(\xi) &= \alpha_1 + \alpha_2 \xi = u_i + (u_{i+1} - u_i)\,\xi \\ &= (1 - \xi)\, u_i + \xi u_{i+1} =: u_i \Phi_1(\xi) + u_{i+1} \Phi_2(\xi). \end{aligned}$$

Bemerkung 34. *Für die Formfunktionen gilt*

$$\forall \xi \in [0,1] : \Phi_1(\xi) + \Phi_2(\xi) = 1.$$

Für die Ansatzfunktionen gilt

$$\varphi_i(x) = \begin{cases} \Phi_2\left(\xi_{I_{i-1}}(x)\right), & x \in I_{i-1} \\ \Phi_1\left(\xi_{I_i}(x)\right), & x \in I_i \\ 0, & \text{sonst.} \end{cases}$$

Wir können nun auf dem Referenzintervall und abhängig von den Formfunktionen die Matrixeinträge der Systemmatrix berechnen.

5.4.3.3 Berechnung der Systemmatrix-Einträge

Zu berechnen sind die Einträge $a(\varphi_i, \varphi_j)$ der Systemmatrix A (in unserem Fall für das Helmholtz-Problem):

$$a(\varphi_i, \varphi_j) = \sum_{k=1}^{N} \int_{I_k} \nabla\varphi_i(x)\nabla\varphi_j(x) + \varphi_i(x)\varphi_j(x)\, dx.$$

Man berechnet die Matrix-Einträge elementweise: Die φ_i, φ_j können durch Φ_n und Φ_m ($n, m \in \{1, 2\}$) ersetzt werden. Man wendet die Kettenregel an, um den Gradienten bzgl. x in einen Gradienten bzgl. ξ umzuwandeln, und die Sustitutionsregel, um das Integral über I_k in ein Integral über $[0, 1]$ umzuwandeln:

$$\begin{aligned}
(A_{I_k})_{nm} &= \int_{I_k} \nabla_x \Phi_n\left(\xi_{I_k}(x)\right) \nabla_x \Phi_m\left(\xi_{I_k}(x)\right) + \Phi_n\left(\xi_{I_k}(x)\right) \Phi_m\left(\xi_{I_k}(x)\right) dx \\
&= h_k \int_0^1 \nabla_\xi \Phi_n(\xi)\, \xi'_{I_k}\left(x_{I_k}(\xi)\right) \nabla_\xi \Phi_m(\xi)\, \xi'_{I_k}\left(x_{I_k}(\xi)\right) + \Phi_n(\xi)\, \Phi_m(\xi)\, d\xi \\
&= h_k \int_0^1 \frac{1}{h_k^2} \nabla\Phi_n \nabla\Phi_m + \Phi_n \Phi_m d\xi.
\end{aligned}$$

$(A_{I_k})_{mn}$ bezeichnet hier den Eintrag in der Matrix, der auf dem Element I_k durch die m-ten und n-ten Formfunktionen erzeugt wird. Die Matrixeinträge von A_{I_k} lauten:

$$\begin{aligned}
(A_{I_k})_{11} &= \int_0^1 \frac{1}{h_k}\nabla\Phi_1\nabla\Phi_1 + \Phi_1\Phi_1 \cdot h_k d\xi \\
&= \int_0^1 \frac{1}{h_k} + h_k(1-\xi)^2 d\xi = \frac{1}{h_k} + \frac{1}{3}h_k \\
(A_{I_k})_{12} &= \int_0^1 \frac{1}{h_k}\nabla\Phi_1\nabla\Phi_2 + \Phi_1\Phi_2 \cdot h_k d\xi \\
&= \int_0^1 -\frac{1}{h_k} + \xi(1-\xi) \cdot h_k d\xi \\
&= -\frac{1}{h_k} + \frac{1}{6}h_k \\
(A_{I_k})_{21} &= -\frac{1}{h_k} + \frac{1}{6}h_k \\
(A_{I_k})_{22} &= \frac{1}{h_k} + \frac{1}{3}h_k \\
\Rightarrow \quad A_{I_k} &= \frac{1}{h_k}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} + h_k \begin{bmatrix} 1/3 & 1/6 \\ 1/6 & 1/3 \end{bmatrix}.
\end{aligned}$$

5.4.3.4 Quadratische Elemente

Statt einer linearen Approximation können wir auch einen quadratischen Ansatz wählen. Dies erhöht die lokale Approximationsgüte. Wir stellen die Lösung $u_h(\xi)$ deshalb dar als

$$u_h(\xi) = \alpha_1 + \alpha_2\xi + \alpha_3\xi^2 \text{ auf } I = [0,1].$$

Um dieses Polynom eindeutig zu bestimmen, benötigen wir 3 Stützstellen, zusätzlich zu u_i und u_{i+1} können wir z.B.

$$u_{i+\frac{1}{2}} = u_h\left(\frac{x_i + x_{i+1}}{2}\right).$$

verwenden. An den Stützstellen im Referenzelement gilt:

$$\begin{aligned}
u_i &= u_h(0) = \alpha_1, \\
u_{i+1} &= u_h(1) = \alpha_1 + \alpha_2 + \alpha_3, \\
u_{i+\frac{1}{2}} &= u_h\left(\frac{1}{2}\right) = \alpha_1 + \frac{1}{2}\alpha_2 + \frac{1}{4}\alpha_3.
\end{aligned}$$

Berechnung der α_i, $i = 1, \ldots, 3$ liefert

$$u_h(\xi) = u_i \Phi_1(\xi) + u_{i+1} \Phi_2(\xi) + u_{i+\frac{1}{2}} \Phi_3(\xi)$$

mit

$$\begin{aligned} \Phi_i(\xi) &= 2\left(\xi - \frac{1}{2}\right)(\xi - 1) \\ \Phi_2(\xi) &= 2\xi\left(\xi - \frac{1}{2}\right) \\ \Phi_3(\xi) &= 4\xi(1 - \xi) \end{aligned}$$

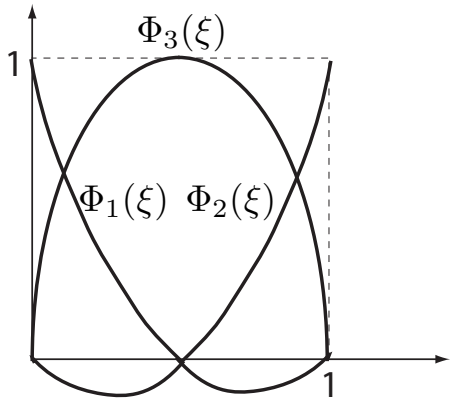

Abbildung 5.7: Ansatzfunktionen für quadratische Elemente

Die Submatrizen A_{I_k} sind jetzt 3×3-Matrizen und lassen sich analog zum linearen Fall ausrechnen.

Bemerkung 35. *Die Gradienten $\nabla \Phi_i(\xi)$ sind nun keine Konstanten über dem Referenzintervall. Man benötigt also noch ein numerischen Verfahren zur Berechnung der Integrale.*

5.4.4 Finite Elemente im $\mathbb{R}^2$

Wir übertragen nun die Vorgehensweise der Finiten Elemente im $\mathbb{R}^1$ auf den zweidimensionalen Fall. Gitterelemente, die im Eindimensionalen Intervalle waren, sind nun Dreiecke oder Vierecke.
Die affine Transformation vom Referenzdreieck in das Dreieck aus der Trangulierung ist

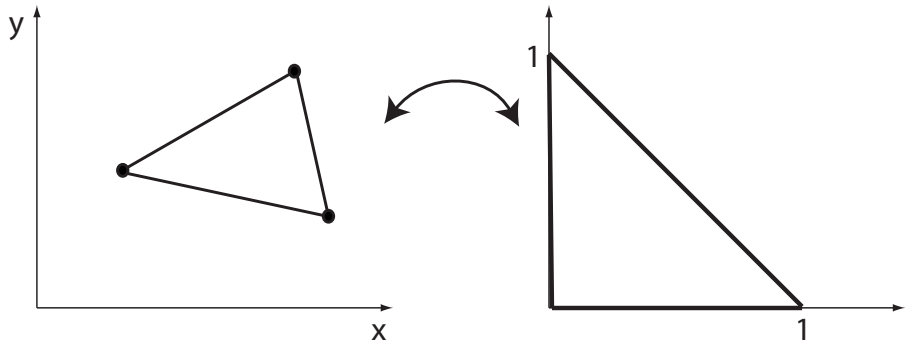

Abbildung 5.8: Affine Transformation von Dreiecken zwischen diskretisiertem Raum und Referenzelement

eine lineare, bijektive Abbildung mit der Eigenschaft:

$$\begin{aligned} x &= x_1 + (x_2 - x_1)\,\xi + (x_3 - x_1)\,\eta, \\ y &= y_1 + (y_2 - y_1)\,\xi + (y_3 - y_1)\,\eta \end{aligned}$$

$$\Leftrightarrow \quad \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} + \begin{pmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{pmatrix} \begin{pmatrix} \xi \\ \eta \end{pmatrix},$$

wobei (x_i, y_i) die Koordinaten des i-ten Eckpunkts des Dreiecks aus der Triangulierung sind. Die Rücktransformation hat damit die Gestalt:

$$\begin{aligned} \begin{pmatrix} \xi \\ \eta \end{pmatrix} &= \underbrace{\begin{pmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{pmatrix}^{-1}}_{A^{-1}} \begin{pmatrix} x - x_1 \\ y - y_1 \end{pmatrix} \\ &= \frac{1}{\det A} \begin{pmatrix} y_3 - y_1 & x_1 - x_3 \\ y_1 - y_2 & x_2 - x_1 \end{pmatrix} \begin{pmatrix} x - x_1 \\ y - y_1 \end{pmatrix} \end{aligned}$$

mit $\det A = (x_2 - x_1)\,(y_3 - y_1) - (x_3 - x_1)\,(y_2 - y_1)$. Für die partiellen Ableitungen gilt:

$$\begin{aligned} u_x &= u_\xi \xi_x + u_\eta \eta_x, \\ u_y &= u_\xi \xi_y + u_\eta \eta_y \end{aligned}$$

und

$$\begin{aligned}
\xi_x &= \frac{y_3 - y_1}{\det A}, \\
\eta_x &= \frac{y_1 - y_2}{\det A}, \\
\xi_y &= \frac{x_1 - x_2}{\det A}, \\
\eta_y &= \frac{x_2 - x_1}{\det A}.
\end{aligned}$$

Bemerkung 36. $\det A$ *beschreibt dabei die Volumenänderung des Dreiecks unter der vorgegebenen Transformation:*

$$dxdy = \det A d\xi d\eta.$$

Mit $\nabla\Phi_i$, $i = 1, 2, 3$ und $\xi_x, \xi_y, \eta_x, \eta_y$ sind die Komponenten der Matrix A_{I_k} vollständig bestimmbar. Dazu ist ein (einfaches) numerisches Integrationsverfahren notwendig.

5.4.4.1 Wahl der Φ_i: Lineare Elemente

Wir stellen die Lösung dar als

$$\begin{aligned}
u_h(\xi, \eta) &= \alpha_1 + \alpha_2\xi + \alpha_3\eta, \\
u_j &:= u_h\left(\overline{P_j}\right), \quad j = 1, 2, 3.
\end{aligned}$$

$\overline{P_j}$ sind die Eckpunkte des Referenzdreiecks. Es gilt

$$\begin{aligned}
u_1 &= u_h(0, 0) = \alpha_1, \\
u_2 &= u_h(1, 0) = \alpha_1 + \alpha_2, \\
u_3 &= u_h(0, 1) = \alpha_1 + \alpha_3.
\end{aligned}$$

Eingesetzt liefert dies

$$u_h(\xi, \eta) = u_1 + (u_2 - u_1)\xi + (u_3 - u_1)\eta = (1 - \xi - \eta)u_1 + \xi u_2 + \eta u_3.$$

Mit $\Phi_1 = 1 - \xi - \eta$, $\Phi_2 = \xi$, $\Phi_3 = \eta$ folgt

$$u_h(\xi, \eta) = u_1\Phi_1 + u_2\Phi_2 + u_3\Phi_3$$

mit

$$\begin{aligned}
\Phi_i\left(\overline{P_j}\right) &= \delta_{ij} \quad i, j = 1, 2, 3, \\
\sum_{i=1}^{3} \Phi_i(\xi, \eta) &= 1 \quad \xi, \eta \in \bar{T}.
\end{aligned}$$

5.4.4.2 Wahl der Φ_i: Quadratische Elemente

Wir benötigen für den Fall quadratischer Ansatzfunktionen 3 weitere Stützstellen, da wir die Lösung u_h nun schreiben als

$$u_h(\xi,\eta) = \alpha_1 + \alpha_2\xi + \alpha_3\eta + \alpha_4\xi^2 + \alpha_5\xi\eta + \alpha_6\eta^2.$$

Die Formfunktionen dazu lauten

$$\begin{aligned}
\Phi_1 &= (1-\xi-\eta)(1-2\xi+2\eta)\\
\Phi_2 &= \xi(2\xi-1)\\
\Phi_3 &= \eta(2\eta-1)\\
\Phi_4 &= 4\xi(1-\xi-\eta)\\
\Phi_5 &= 4\xi\eta\\
\Phi_6 &= 4\eta(1-\xi-\eta)
\end{aligned}$$

Bemerkung 37. *Im $\mathbb{R}^3$ geht man analog vor. Der lineare Fall definiert*

$$u_h(\xi,\eta,\zeta) = \alpha_1 + \alpha_2\xi + \alpha_3\eta + \alpha_4\zeta.$$

Das Tetraederelement im dreidimensionalen Fall liefert die nötigen 4 Knoten für die Bestimmung von $\alpha_1,\ldots,\alpha_4$.

5.4.5 Konvergenzaussagen zu Finite Elemente Verfahren

5.4.5.1 Abschätzung des Energiefehlers

Für die symmetrische positiv defintite Bilinearformen a mit $a(u,v) = \int\limits_\Omega \nabla u \nabla v dx$ ist die Energienorm definiert als

$$\|u\|_a \coloneqq (a(u,u))^{\frac{1}{2}}.$$

Wir wissen aus der Minimalitätseigenschaft (vgl. Bem. 32)

$$\|u-u_h\|_a = \min_{v_h\in V_h} \|u-v_h\|_a.$$

Satz 22. *Sei $\Omega\subset\mathbb{R}^d$, $d\le 3$ polygonal und konvex und $u\in H^2(\Omega)$ die schwache Lösung der Poisson-Gleichung. Sei $u_h\in V_h$ die Finite-Elemente-Lösung von*

$$a(u_h,v_h) = \langle f, v_h\rangle \ \forall v_h\in V_h \subset H_0^1(\Omega)$$

mit linearen, konformen Finiten Elementen. Dann gilt für den Diskretisierungsfehler

$$\|u-u_h\|_a \le c\cdot h\cdot \|f\|_{L^2(\Omega)}, \quad f\in L^2(\Omega). \tag{5.45}$$

Bemerkung 38. *Falls für alle $f\in L^2(\Omega)$ gilt*

$$\|u\|_{H^2(\Omega)} \le c\|f\|_{L^2(\Omega)},$$

dann heißt das Problem H^2-regulär.

5.4.5.2 Abschätzung des L^2-Fehlers

Satz 23. *Sei $\Omega \subset \mathbb{R}^d$, $d \leq 3$ und $u \in H^2(\Omega)$ schwache H^2-reguläre Lösung der Poisson-Gleichung. Dann gilt*

$$\|u - u_h\|_{L^2(\Omega)} \leq c \cdot h \|u - u_h\|_a , \tag{5.46}$$

$$\|u - u_h\|_{L^2(\Omega)} \leq c \cdot h^2 \|f\|_{L^2(\Omega)} . \tag{5.47}$$

6 Ausblick

6.1 Finite Volumen Verfahren

Die Idee der Finite-Volumen-Diskretisierung ist sehr ähnlich zu den Finite-Elemente-Verfahren. Das Finite-Elemente-Konzept wird jedoch nicht auf das bisherige Gitter Ω_h angewandt, sondern auf ein *duales* Gitter. Dies hat zur Folge, dass das Verfahren lokal flusserhaltend ist. Wir erinnern uns an Konzentrationsänderungen in einem beliebigen Volumenelement, die – ohne Quellen oder Senken im Volumen – gleich dem Fluss über den Elementrand sind:

$$\int_B \frac{\partial u}{\partial t}\, d\vec{x} = \int_B -\mathrm{div}\vec{F}\, d\vec{x} = \int_{\partial B} \vec{F}\cdot\vec{n}\, ds$$

Wählen wir ein solches Volumenelement über jedem Knoten in Ω_h, z.B. durch Verbinden von Kantenmittelpunkten, Flächenschwerpunkten und Volumenschwerpunkt, so erhält man sogenannte *Kontrollvolumen,* die zusammengesetzt ein konformes duales Gitter über dem Gebiet Ω erzeugen. Für jedes solches Kontrollvolumen B_i gilt die obige Gleichung. Macht man wie bei Finite-Elemente-Verfahren etwa den Ansatz einer stetigen und elementweise linearen diskreten Lösung u_h, kann man diese wieder als Linearkombination von Basisfunktionen darstellen und die Koeffizienten dieser Darstellung als Unbekannte behandeln. Numerische Integration in der obigen Gleichung führt dann zu einem Gleichungssystem

$$K_h^{FV} u_h^{FV} = f_h^{FV}.$$

6.2 Lösen von linearen Gleichungssystemen

Aus allen besprochenen Diskretisierungsverfahren gewinnen wir ein (mitunter sehr großes) lineares Gleichungssystem

$$Ku_h = b,$$

wobei wir am Ende u_h als Lösung des Problems berechnen wollen. D.h. wir benötigen die Darstellung

$$u_h = K^{-1}b,$$

müssen also K invertieren. Das Problem in realen Fällen ist jedoch, dass eine exakte Invertierung vom Rechen- und somit Zeitaufwand nicht vertretbar ist. Ein Ansatz zur Lösung des Gleichungssystems ist, die Inverse von A iterativ zu approximieren. Welche Vor- und Nachteile solche Verfahren haben und wie man effizient große (dünn besetzte) lineare Systeme löst, ist Thema einer Folgevorlesung.