Grundlagen zur Numerik partieller Differenzialgleichungen

Werner Vogt
Technische Universität Ilmenau
Institut für Mathematik
Postfach 100565
98684 Ilmenau

Ilmenau, den 7. März 2006

Zusammenfassung

Der Beitrag stellt die beiden wesentlichsten Verfahrenklassen zur numerischen Lösung von Randwertproblemen bei partiellen Differenzialgleichungen vor. Die Finite-Differenzen-Methode wird direkt auf kartesischen Gittern angewandt, wobei der globale Diskretisierungsfehler mittels asymptotischer Fehlerschätzung gewonnen wird. Nach Überführung des Randwertproblems in eine Variationsgleichung liefert andererseits die Finite-Elemente- Methode eine Lösungsapproximation auf allgemeineren Gebieten. Anwendungsbeispiele illustrieren das Konvergenzverhalten der Verfahren.

MSC 2000: 65N06, 65N30, 65N12

Keywords: Boundary value problems, finite difference methods, finite elements

1 Einleitung

Die Lösung partieller Differenzialgleichungen (PDGLn) auf komplizierten Geometrien lässt sich nur in speziellen Fällen in Termen elementarer Funktionen angeben. Auch Produktansätze sind meist wegen der Nichtlinearität der Gleichungen oder der Gestalt der Gebiete nicht anwendbar. Numerische Verfahren zeichnen sich dagegen durch hohe Adaptivität bezüglich

- der konkreten Form der linearen oder nichtlinearen Differenzialgleichungen,
- der Randbedingungen bzw. Anfangsbedingungen und
- der komplizierten Geometrie des Integrationsgebietes $\Omega \subset \mathbb{R}^n$

aus. Für allgemeine Klassen linearer oder nichtlinearer PDGLn auf einfachen kartesischen (rechteckigen achsenparallelen) Gebieten stellt die historisch ältere Finite-Differenzen-Methode (FDM) einen universellen Zugang dar und wird deshalb in Abschnitt 2 behandelt. Ihre Grenzen findet diese Methode allerdings bei komplizierteren, insbesondere krummlinig berandeten Gebieten. Im Ingenieurbereich wurde deshalb die Finite-Elemente-Methode (FEM) entwickelt, die auf komplizierten Geometrien wesentlich flexibler ist und zudem auch schwache (verallgemeinerte) Lösungen approximiert. Mit einem adaptiven Dreiecksnetz kann nun das Gebiet „trianguliert“ werden, womit viele praktische Probleme erst lösbar werden. Allerdings müssen die Differenzialgleichungen und Randbedingungen vorher in eine Variationsform überführt werden. Im Rahmen dieses Beitrages wollen wir die grundlegende Herangehensweise dieses leistungsfähigen Zuganges behandeln. Als einfaches Demonstrationsmodell nutzen wir in Abschnitt 3 die 2-dimensionale Poisson-Gleichung mit homogenen Dirichlet-Bedingungen

$$-\Delta u \;=\; f \quad \text{in} \quad \Omega, \qquad u \;=\; 0 \quad \text{auf} \quad \partial\Omega.$$

Die Verallgemeinerung der FEM auf andere Gleichungstypen, Randbedingungen und Gebiete findet man in den umfangreichen Darstellungen [6, 10, 11, 16].

Wir betrachten ein beschränktes Gebiet $\Omega \subset \mathbb{R}^n$ mit einem stückweise glatten Rand $\Gamma = \partial\Omega$. Dann lautet die allgemeine *fastlineare PDGL 2. Ordnung* für $u = u(x),\ x = (x_1, x_2, \ldots, x_n)$,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}(x)\frac{\partial^2 u}{\partial x_i \partial x_j} + f(x, u,\ \text{grad}\ u) \;=\; 0 \quad \text{mit} \quad \text{grad}\ u = (u_{x_1}, \ldots, u_{x_n}). \tag{1}$$

Die $\mathcal{C}^1$-Funktionen $a_{ij}(x)$ sollen die symmetrische Matrix $A(x)$ bilden, womit die Gleichung in der Kurzform

$$\operatorname{div}(A(x)\operatorname{grad} u) + f(x, u, \operatorname{grad} u) \;=\; 0 \tag{2}$$

darstellbar ist. Ist $A(x)$ (positiv oder negativ) definit, so ist die PDGL von elliptischem Typ.

Beispiel 1 Praktisch bedeutende elliptische Gleichungen mit dem *Laplace-Operator* $\Delta u \;:=\; u_{x_1x_1} + u_{x_2x_2} + \ldots + u_{x_nx_n}$ sind die in stationären Prozessen auftretende

- lineare *Helmholtz-Gleichung*

$$\Delta u + \lambda^2 u + f(x) \;=\; 0 \qquad \text{mit Parameter} \quad \lambda \in \mathbb{R}, \tag{3}$$

- linear-inhomogene *Poisson-Gleichung (auch: Potenzialgleichung)*

$$\Delta u + f(x) \;=\; 0 \qquad \text{mit Quellterm} \quad f = f(x), \tag{4}$$

- linear-homogene *Laplace-Gleichung*

$$\Delta u \;=\; 0, \tag{5}$$

- nichtlineare *Bratu-Gleichung (auch: Gelfand-Gleichung, vgl. [20])*

$$\Delta u + \lambda e^u \;=\; 0 \qquad \text{mit Parameter} \quad \lambda \in \mathbb{R}. \qquad \square \tag{6}$$

Um die Grundprinzipien numerischer Verfahren anschaulich darzustellen, betrachten wir als Modellproblem nachfolgend die zweidimensionale Poisson-Gleichung [1] auf dem Rechteck $\Omega = \{(x, y) \mid a < x < b,\ c < y < d\}$

$$-\Delta u \;=\; -u_{xx} - u_{yy} \;=\; f(x, y), \quad (x, y) \in \Omega \tag{7}$$

mit Dirichletschen Randbedingungen

$$u(x, y) \;=\; \varphi(x, y), \quad (x, y) \in \Gamma. \tag{8}$$

Für die weiteren Betrachtungen setzen wir voraus, dass f und φ stetige Funktionen sind.

Bemerkung 2 Die Poisson-Gleichung ist von erheblicher praktischer Bedeutung:

1. Beschreibt $f(x, y)$ die elektrische Ladungsdichte einer Platte Ω, so gibt die Lösung $u(x, y)$ das Potenzial im Punkt (x, y) an. $\varphi(x, y)$ ist das auf dem Rand vorgegebene Potenzial.
2. Belastet man eine am Rand Γ befestigte horizontale elastische Membran ($\varphi = 0$) transversal mit der Last $f(x, y)$, so beschreibt unter geeigneten Voraussetzungen die Lösung $u(x, y)$ die vertikale Ausdehnung dieser Membran.
3. Zudem ist leicht nachweisbar, dass die Poisson-Gleichung auch die stationäre Temperaturverteilung in einer Platte (bzw. im 3D-Fall in einem Körper) beschreibt.
4. Falls $f(x, y, z)$ die Massendichte im Raum darstellt, so gibt $u(x, y, z)$ das zugehörige Gravitationspotenzial des Gravitationsfeldes an.

[1] Siméon-Denis Poisson (1781–1840), französischer Physiker und Mathematiker mit zahlreichen Beiträgen zur mathematischen Physik und zur Wahrscheinlichkeitstheorie.

2 Finite-Differenzen-Methode

2.1 Rechteckgitter und finite Ausdrücke

Finite-Differenzenverfahren (engl.: finite difference methods, FDM) approximieren eine vorausgesetzte exakte Lösung $u(x, y)$ auf einem *Gitter* Ω_h der maximalen Schrittweite h. Dazu werden alle in der PDGL auftretenden Funktionswerte und Ableitungen durch *finite Ausdrücke* ersetzt, womit die bekannten Differenzenquotienten verallgemeinert werden. Unterteilen wir das Intervall $[a, b]$ äquidistant in N_x Teilintervalle der Länge h_x mit

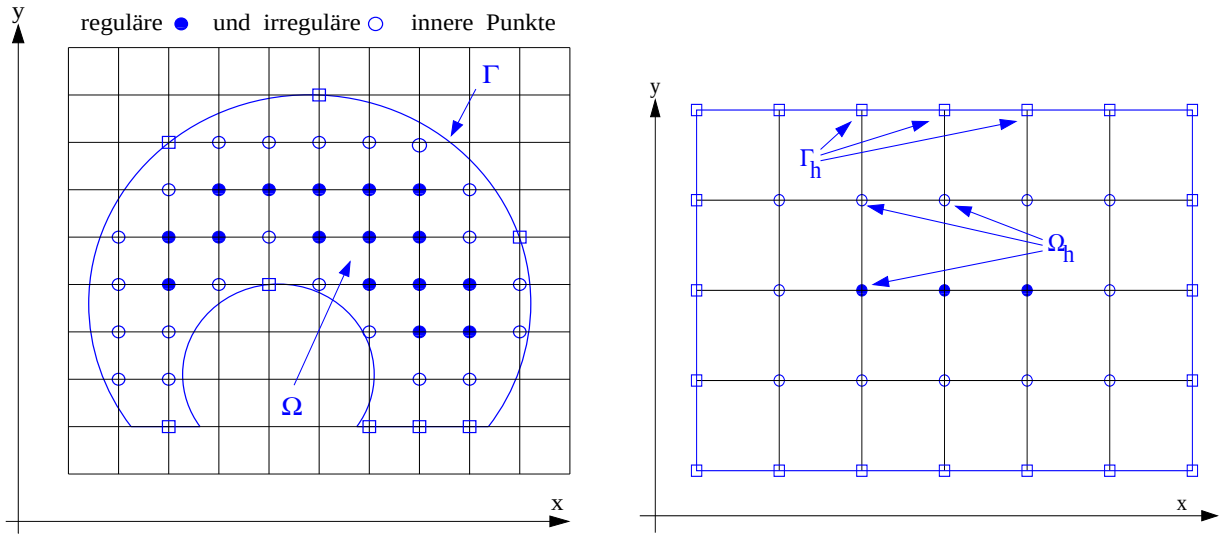

Abbildung 1: Gitter Ω_h auf Ω mit irregulärem Rand (a) und regulärem Rand (b)

$h_x = (b-a)/N_x > 0$ und analog dazu das Intervall $[c, d]$ äquidistant in N_y Intervalle der Länge $h_y = (d-c)/N_y > 0$, so definiert

$$\Omega_h \;=\; \{(x_i, y_j) \in \Omega \mid x_i = a + ih_x,\; y_k = c + jh_y,\; i, j \in \mathbb{Z}\} \;\subset\; \Omega \tag{9}$$

ein koordinatenweise äquidistantes *Gitter* mit den Schrittweiten $h_x,\; h_y$ und der *Maximalschrittweite* $h := \max(h_x, h_y)$. Alle Gitterpunkte (Knoten) $(x_i, y_j) \in \Omega$ nennt man innere Punkte. Auf dem Rand Γ liegende Punkte der Form (x_i, y_j) bilden das Gitter der Randpunkte

$$\Gamma_h \;=\; \{(x_i, y_j) \in \Gamma \mid x_i = a + ih_x,\; y_k = c + jh_y,\; i, j \in \mathbb{Z}\} \;\subset\; \Gamma, \tag{10}$$

die in Abb. 1 mittels □ dargestellt werden. Das abgeschlossene Gitter ist dann $\overline{\Omega}_h = \Omega_h \cup \Gamma_h$. Wir unterscheiden weiterhin randferne (auch: reguläre) innere Gitterpunkte, bei denen die 4 benachbarten Knoten (in östlicher, südlicher, westlicher und nördlicher Richtung) zu Ω gehören, von randnahen (auch: irregulären) inneren Knoten [13]. In Abb. 1 (a) sind für ein krummlinig berandetes Gebiet (mit irregulärem Rand) die 3 Knotenarten dargestellt. Für das Rechteckgitter in Abb. 1b sind nur 3 innere Gitterpunkte randfern, während an den Randpunkten die Funktionswerte durch die Dirichletschen Bedingungen $\varphi(x_i, y_j)$ vorgegeben sind. Alle weiteren Betrachtungen lassen sich auch auf nichtäquidistante Gitter verallgemeinern.

Um partielle Ableitungen $u_x, u_y, u_{xx}, u_{xy}, u_{yy}, \ldots$ auf dem Gitter zu approximieren, gehen wir

von einfachen Differenzenquotienten

$$\frac{\partial u}{\partial x} \approx \frac{1}{h_x}[u(x+h_x,y)-u(x,y)] \qquad \text{bzw.} \qquad \frac{\partial u}{\partial x} \approx \frac{1}{h_x}[u(x,y)-u(x-h_x,y)]$$

aus, womit die 2. Ableitung durch

$$\begin{aligned}\frac{\partial^2 u}{\partial x^2} &\approx \frac{1}{h_x}\left\{\frac{1}{h_x}[u(x+h_x,y)-u(x,y)]-\frac{1}{h_x}[u(x,y)-u(x-h_x,y)]\right\}\\ &\approx \frac{1}{h_x^2}[u(x+h_x,y)-2u(x,y)+u(x-h_x,y)]\end{aligned}$$

angenähert werden kann. Setzen wir allgemein die zu approximierende k-te partielle Ableitung als Linearkombination von Funktionswerten aus der Umgebung des Knotens (x,y) an

$$\frac{\partial^k u}{\partial x^k} \approx \frac{1}{h_x^k}\sum_{i=-p}^{q} c_i\, u(x+ih_x,y), \quad k=1,2,3,\ldots \qquad \text{mit} \quad c_i \in \mathbb{R}, \tag{11}$$

so erhalten wir *finite Ausdrücke* für diese Ableitungen. Die konstanten Gewichte c_i und die Indexgrenzen p und q gewinnen wir mittels Taylor-Entwicklung am Knoten (x,y) mit dem Ziel einer Fehlerordnung $\mathcal{O}(h_x^s)$. Wir betrachten besonders zentrale Differenzen mit $p=q$, Vorwärts-Differenzen ($p=0$) und Rückwärts-Differenzen ($q=0$). Hier einige Beispiele:

1. Zentrale Differenzen mit Fehler $\mathcal{O}(h_x^2)$

$$\begin{aligned}u_x &= \frac{1}{2h_x}[u(x+h_x,y)-u(x-h_x,y)]-\frac{1}{6}\frac{\partial^3 u}{\partial x^3}(x,y)h_x^2+\mathcal{O}(h_x^4)\\ u_{xx} &= \frac{1}{h_x^2}[u(x+h_x,y)-2u(x,y)+u(x-h_x,y)]-\frac{1}{12}\frac{\partial^4 u}{\partial x^4}(x,y)h_x^2+\mathcal{O}(h_x^4)\end{aligned}$$

2. Vorwärts-Differenzen mit Fehler $\mathcal{O}(h_x^2)$

$$\begin{aligned}u_x &= \frac{1}{2h_x}[-u(x+2h_x,y)+4u(x+h_x,y)-3u(x,y)]+\mathcal{O}(h_x^2)\\ u_{xx} &= \frac{1}{h_x^2}[-u(x+3h_x,y)+4u(x+2h_x,y)-5u(x+h_x,y)+2u(x,y)]+\mathcal{O}(h_x^2)\end{aligned}$$

3. Rückwärts-Differenzen mit Fehler $\mathcal{O}(h_x^2)$

$$\begin{aligned}u_x &= \frac{1}{2h_x}[3u(x,y)-4u(x-h_x,y)+u(x-2h_x,y)]+\mathcal{O}(h_x^2)\\ u_{xx} &= \frac{1}{h_x^2}[2u(x,y)-5u(x-h_x,y)+4u(x-2h_x,y)-u(x-3h_x,y)]+\mathcal{O}(h_x^2)\end{aligned}$$

4. Zentrale Differenzen mit Fehler $\mathcal{O}(h_x^4)$

$$\begin{aligned}u_x &= \frac{1}{12h_x}[-u(x+2h_x,y)+8u(x+h_x,y)-8u(x-h_x,y)+u(x-2h_x,y)]\\ &\quad +\frac{1}{30}\frac{\partial^5 u}{\partial x^5}(x,y)h_x^4+\mathcal{O}(h_x^6)\\ u_{xx} &= \frac{1}{12h_x^2}[-u(x+2h_x,y)+16u(x+h_x,y)-30u(x,y)+16u(x-h_x,y)\\ &\qquad -u(x-2h_x,y)]+\frac{1}{90}\frac{\partial^6 u}{\partial x^6}(x,y)h_x^4+\mathcal{O}(h_x^6).\end{aligned}$$

Analog ergeben sich die finiten Ausdrücke für die partiellen Ableitungen nach y, z.B. die zentralen Differenzen mit Fehler $\mathcal{O}(h_y^2)$

$$\begin{aligned} u_y &= \frac{1}{2h_y}[u(x,y+h_y) - u(x,y-h_y)] - \frac{1}{6}\frac{\partial^3 u}{\partial y^3}(x,y)h_y^2 + \mathcal{O}(h_y^4) \\ u_{yy} &= \frac{1}{h_y^2}[u(x,y+h_y) - 2u(x,y) + u(x,y-h_y)] - \frac{1}{12}\frac{\partial^4 u}{\partial y^4}(x,y)h_y^2 + \mathcal{O}(h_y^4). \end{aligned}$$

Die gemischte Ableitung $u_{xy} = \frac{\partial}{\partial y}\left(\frac{\partial u}{\partial x}\right)$ approximieren wir durch zentrale Differenzen 2. Ordnung nach y, angewandt auf die zentrale Differenz bezüglich x, womit wir

$$\begin{aligned} u_{xy} &= \frac{1}{4h_x h_y}[u(x+h_x,y+h_y) - u(x-h_x,y+h_y) - u(x+h_x,y-h_y) \\ &\quad + u(x-h_x,y-h_y)] - \frac{1}{6}\frac{\partial^4 u}{\partial x^3 \partial y}(x,y)h_x^2 - \frac{1}{6}\frac{\partial^4 u}{\partial x \partial y^3}(x,y)h_y^2 + \mathcal{O}(h_x^2 h_y^2) \end{aligned}$$

erhalten. Zur Konstruktion von Approximationen höherer Ableitungen sollte ein Computeralgebra-System, z.B. MAPLE genutzt werden. Die finiten Ausdrücke sind besonders einfach,

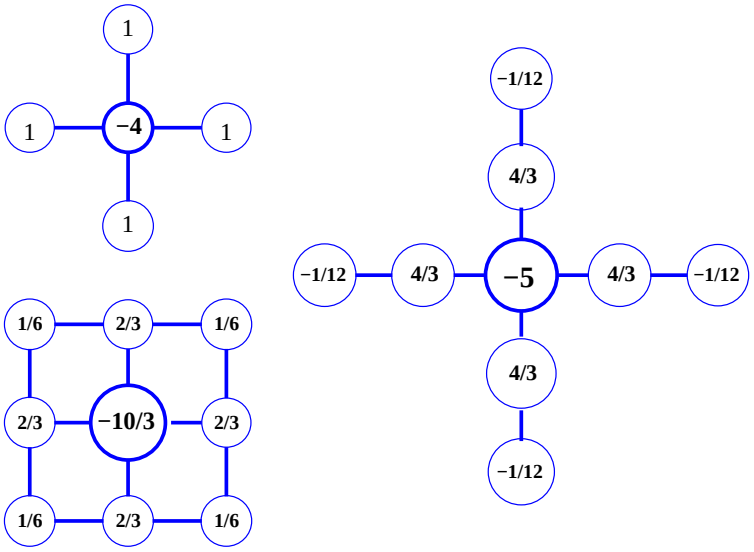

Abbildung 2: Differenzensterne (a) der Ordnung $\mathcal{O}(h^2)$ und (b) der Ordnung $\mathcal{O}(h^4)$

wenn alle Schrittweiten $h_x = h_y = h$ gleich sind. Für den Laplace-Operator ergibt dann die Formel 2. Ordnung

$$\begin{aligned} \Delta u &= \frac{1}{h^2}[u(x+h,y) + u(x-h,y+h) + u(x,y+h) + u(x,y-h) \\ &\quad -4u(x,y)] - \frac{1}{12}\left(\frac{\partial^4 u}{\partial x^4}(x,y) + \frac{\partial^4 u}{\partial y^4}(x,y)\right)h^2 + \mathcal{O}(h^4), \end{aligned} \tag{12}$$

deren Gewichte wir mit dem *5-Punkt-Differenzenstern (engl: computational stencil)* der Abb. 2a veranschaulichen. Approximieren wir jedoch die 2. Ableitungen mit den Formeln der Ordnung $\mathcal{O}(h^4)$, so gewinnen wir den 9-Punkt-Stern der Abb. 2b. Der in Randnähe geeigneter erscheinende quadratische 9-Punkt-Stern in Abb. 2a liefert jedoch nur eine Näherung des Laplace-Operators der Ordnung $\mathcal{O}(h^2)$ und bringt deshalb keinen Genauigkeitsgewinn gegenüber dem 5-Punkt-Stern.[2]

[2] Mit einer Mittelung der Werte $f(x,y)$ lässt sich jedoch die Ordnung 4 erreichen; vgl. [9], S. 127.

2.2 Lösung der finiten Gleichungssysteme

Wir ersetzen in der PDGL (7) an jedem inneren Knoten (x_i, y_j) sämtliche auftretenden Ableitungen durch finite Ausdrücke. Dabei sollten möglichst Formeln ein- und derselben Fehlerordnung benutzt werden. Vernachlässigung aller Fehlerterme ergibt mit der Approximation des Laplace-Operators durch die 5-Punkt-Formel (12) das endliche Gleichungssystem

$$-(\Delta_h u)_{ij} \;:=\; \frac{1}{h^2}(4u_{ij} - u_{i-1,j} - u_{i+1,j} - u_{i,j-1} - u_{i,j+1}) \;=\; f(x_i, y_j), \qquad (13)$$
$$i = 1, 2, \ldots, N_x - 1,\; j = 1, 2, \ldots, N_y - 1$$

an den Gitterpunkten (x_i, y_j). Mit Δ_h bezeichnen wir den diskretisierten Laplace-Operator. Dabei ist $u_{ij} \approx u(x_i, y_j)$ die zu bestimmende Approximation der Lösung am Knoten (x_i, y_j), die wir *diskrete Lösung* nennen wollen. Zusammen mit den vorgegebenen Randwerten

$$u_{ij} \;=\; \varphi(x_i, y_j) \qquad \text{für} \quad i = 0, N_x \quad \text{oder} \quad j = 0, N_y \qquad (14)$$

entsteht ein lineares Gleichungssystem mit den $(N_x - 1)(N_y - 1)$ reellen Unbekannten u_{ij}. Um die Standardform linearer Gleichungssysteme $Av = a$ mit einem Lösungsvektor $v = (v_1, v_2, \ldots, v_n)^T$ zu erreichen, sind die inneren Knoten sequenziell anzuordnen. Zwei übliche Nummerierungen entstehen durch die *zeilen- oder spaltenweise (allgemein: lexikografische) Anordnung* oder die *Schachbrett-Anordnung* in Abb. 3. Wir wollen zeilenweise lexikografisch

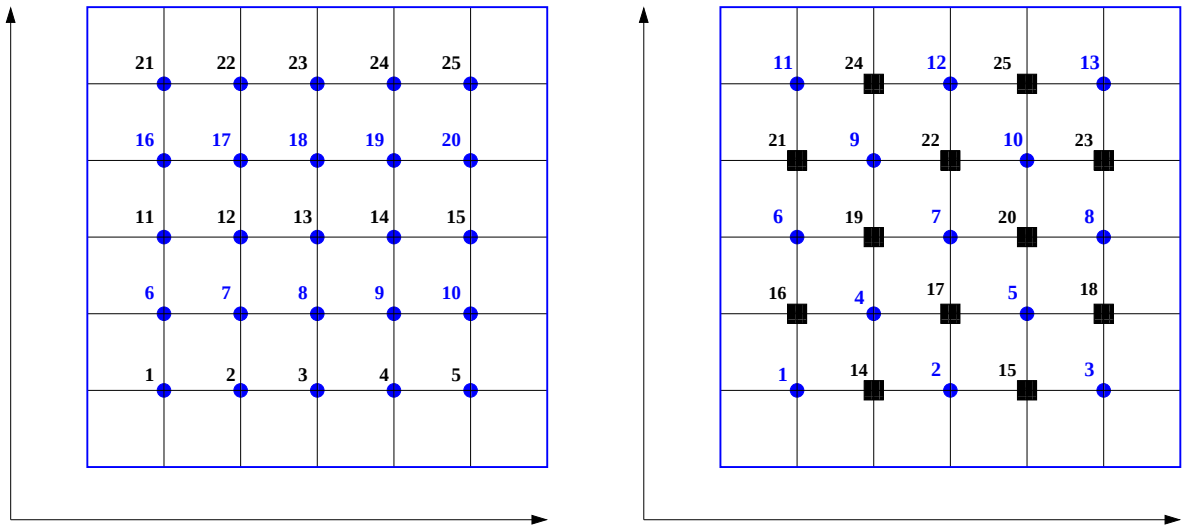

Abbildung 3: Lexikografische Anordnung (a) und Schachbrett-Anordnung (b)

anordnen, womit der Lösungsvektor $v = (v_1, v_2, \ldots, v_n)^T$ mit

$$v_k = u_{ij}, \quad k = i + (j-1)(N_x - 1), \quad i = 1, 2, \ldots, N_x - 1,\; j = 1, 2, \ldots, N_y - 1 \qquad (15)$$

und $n = (N_x - 1)(N_y - 1)$ entsteht. Entsprechend wird der Vektor der rechten Gleichungsseiten $a = (a_1, a_2, \ldots, a_n)^T$ mit $k = i + (j-1)(N_x - 1)$ durch

$$a_k \;=\; f(x_i, y_j) \;+\; \frac{1}{h^2} \sum_{(x_\mu, y_\nu) \in \Gamma} \varphi(x_\mu, y_\nu) \qquad (16)$$

aufgebaut. Unter dem Summenzeichen stehen alle zum Knoten (x_i, y_j) gehörenden Randwerte. Das reelle Gleichungsystem

$$Av \;=\; a \qquad \text{mit} \quad v, a \in \mathbb{R}^n, \quad n = (N_x - 1)(N_y - 1) \tag{17}$$

hat eine blocktridiagonale Koeffizientenmatrix A mit $N_y - 1$ Blöcken T der Dimension $(N_x - 1) \times (N_x - 1)$

$$A \;=\; \frac{1}{h^2} \begin{pmatrix} T & -I & & \cdots & O & O \\ -I & T & -I & \cdots & O & O \\ O & -I & T & \cdots & O & O \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ O & O & O & \cdots & T & -I \\ O & O & O & \cdots & -I & T \end{pmatrix}, \quad T \;=\; \begin{pmatrix} 4 & -1 & & \cdots & 0 & 0 \\ -1 & 4 & -1 & \cdots & 0 & 0 \\ 0 & -1 & 4 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 4 & -1 \\ 0 & 0 & 0 & \cdots & -1 & 4 \end{pmatrix}$$

und entsprechenden $(N_x - 1) \times (N_x - 1)$-Einheitsmatrizen I. In [7, S. 673] wird nachgewiesen, dass die Koeffizientenmatrix A diagonaldominant und irreduzibel ist, womit die eindeutige Lösbarkeit des Systems garantiert ist:

Satz 3 *Das diskretisierte Dirichlet-Problem (13), (14) mit gegebenen Funktionen $\varphi(x, y)$ und $f(x, y)$ besitzt für jedes $N_x > 1$, $N_y > 1$ eine eindeutige diskrete Lösung (u_{ij}), $i = 0, 1, \ldots, N_x$, $j = 0, 1, \ldots, N_y$.*

Die erste sich an diese Existenzaussage anschließende Frage betrifft geeignete Lösungsverfahren für die linearen Gleichungsysteme. Vereinfachend wollen wir dazu $\Omega = (0, 1) \times (0, 1)$ und $h = 1/N$ mit $N := N_x = N_y$ annehmen. Dann gilt (vgl. [9], S. 117f) folgendes

Lemma 4 *Das Spektrum von A hat die reellen Werte*

$$\lambda_{ij} \;=\; \frac{4}{h^2} \left[\sin^2 \left(\frac{i\pi h}{2} \right) + \sin^2 \left(\frac{j\pi h}{2} \right) \right], \quad i, j = 1(1)N - 1. \tag{18}$$

Wegen $\lambda_{ij} > 0$ ist die symmetrische Matrix A positiv definit, womit ebenfalls ihre Regularität folgt. Wir betrachten die Konsequenzen dieses Lemmas für den Einsatz verschiedener Lösungsverfahren.

1. *Direkte Verfahren:* Entscheidend ist die Kondition des Problems für kleines h. Mit

$$\lambda_{min} \;=\; \min_{i,j} \lambda_{ij} \;=\; \frac{8}{h^2} \sin^2 \left(\frac{\pi h}{2} \right), \qquad \lambda_{max} \;=\; \max_{i,j} \lambda_{ij} \;=\; \frac{8}{h^2} \cos^2 \left(\frac{\pi h}{2} \right)$$

berechnen wir die spektrale Konditionszahl (vgl. dazu [7, S. 646])

$$cond_2(A) \;=\; \frac{\lambda_{max}}{\lambda_{min}} \;=\; \frac{4}{\pi^2 h^2} - \frac{2}{3} + \mathcal{O}(h^2) \;\gg 1 \quad \text{für} \quad h \ll 1,$$

so dass direkte Verfahren bei kleinen Schrittweiten h Lösungen liefern, die große Rundungsfehler aufweisen.

2. *Klassische iterative Verfahren:* Das Richardson-Verfahren mit optimalem Parameter

$$x_{k+1} \;=\; x_k - \theta_{opt}(Ax_k - b), \qquad \theta_{opt} \;=\; 2/(\lambda_{max} + \lambda_{min})$$

besitzt die Iterationsmatrix $M = I - \theta_{opt} A$ mit Spektralradius

$$\varrho_{opt}(M) \;=\; (\lambda_{max} - \lambda_{min})/(\lambda_{max} + \lambda_{min}) \;=\; \cos \pi h \;=\; 1 - |\mathcal{O}(h^2)|.$$

Die Iterationsmatrix $M = I - \frac{1}{4}\omega h^2 A$ des Jacobi- und JOR-Verfahrens mit Relaxationsparameter ω hat den Spektralradius ebenfalls nahe 1 wegen

$$\varrho(M) \;=\; 1 - 2\omega \sin^2(\pi h/2) \;=\; 1 \;-\; |\mathcal{O}(h^2)|.$$

Dasselbe ungünstige asymptotische Verhalten zeigen Gauß-Seidel, SOR- und Gradientenverfahren (vgl. [7, S. 674]).

3. *Krylov-Unterraumverfahren:* SOR-Verfahren mit optimalem Relaxationsparameter ω_{opt} und Konjugierte-Gradienten (CG)-Verfahren besitzen Iterationsmatrizen mit Spektralradius $\varrho(M) = 1 \;-\; |\mathcal{O}(h)|$, der im Allgemeinen weiter von 1 entfernt ist. Damit konvergiert das CG-Verfahren bei kleinerer Diskretisierungsschrittweite h schneller als die klassischen Iterationsverfahren. Dieses Verhalten bestätigen auch Rechnungen mit anderen Krylov-

N	$n = (N-1)^2$	t in sec	*iter*	*relres*
31	900	0.951	120	7.33e-13
41	1600	1.763	156	7.61e-13
51	2500	5.728	193	9.38e-13
61	3600	40.99	233	9.70e-13
71	4900	142.1	269	8.95e-13

Tabelle 1: Iterationen des GMRES-Verfahrens für System (17)

Unterraumverfahren, insbesondere dem GMRES-Verfahren aus [7]. Lösen wir das System (17) mit dem GMRES-Verfahren bei einer vorgegebenen Genauigkeit $tol = 10^{-12}$ durch den MATLAB-Aufruf

```
[x,flag,relres,iter,resvec] = gmres(A,a,[],tol);
```

so liefert Tabelle 1 die Entwicklung der Rechenzeit[3] t und die Gesamtzahl *iter* der Iterationen, um das angegebene relative Residuum *relres* zu erreichen.

4. *Mehrgitterverfahren (MGM):* Diese leistungsfähige Verfahrensklasse garantiert eine Normabschätzung der Form $\|M\| \leq C$ für die Iterationsmatrix M mit einer problemabhängigen Konstanten $C < 1$ gleichmäßig für alle Schrittweiten $h > 0$. Für großdimensionale Systeme (17) ist sie oft den anderen Zugängen überlegen. Ihre nicht unkomplizierte Theorie findet man in [5].

Bemerkung 5 Schnelle Lösungsverfahren für die finiten Gleichungssysteme sind auch bei anderen PDGLn erforderlich, das betrifft insbesondere 3D-Gleichungen

$$-\Delta u \;=\; -u_{xx} - u_{yy} - u_{zz} \;=\; f(x,y,z), \quad (x,y,z) \in \Omega \subset \mathbb{R}^3, \tag{19}$$

deren Diskretisierung mit einem (13) entsprechenden 3D-Differenzenoperator Δ_h leicht möglich ist und dem Leser überlassen wird.

[3] CPU Intel Pentium 4, 1.6 GHz.

2.3 Diskrete Konvergenz

Die nach der Lösbarkeit auftretende zweite Frage ist diejenige nach der Konvergenz der diskreten Näherungslösungen u_h gegen die exakte Lösung $u(x,y)$, falls die Schrittweite h gegen Null strebt. Wir führen die Konvergenzanalyse mit der allgemeinen Theorie (vgl. [17, 19]) in 4 Schritten durch:

1. *Konstruktion des Diskretisierungsverfahrens:* Das Ausgangsproblem (7), (8) notieren wir als Operatorgleichung $F(u) = 0$ mit dem Operator $F : D \subset B \to B^0$

$$F(u) := \begin{cases} \Delta\, u(x,y) + f(x,y), & (x,y) \in \Omega \\ u(x,y) - \varphi(x,y), & (x,y) \in \Gamma \end{cases} \tag{20}$$

und den Banach-Räumen $B = B^0 = \mathcal{C}^0(\overline{\Omega})$. Wir nehmen hinreichende Glattheit aller Funktionen an und betrachten F auf $D = \mathcal{C}^s(\overline{\Omega}) \subset B,\ s \geq 4$. Das diskrete Problem (13), (14) stellen wir analog dazu mit dem Operator $F_h : B_h \to B_h^0$, definiert durch

$$F_h(u_h) := \begin{cases} \Delta_h\, u_h(x,y) + f(x,y), & (x,y) \in \Omega_h \\ u_h(x,y) - \varphi(x,y), & (x,y) \in \Gamma_h \end{cases}, \tag{21}$$

als Operatorgleichung $F_h(u_h) = 0$ dar. Die endlich dimensionalen Banach-Räume $B_h = B_h^0$ von Gitterfunktionen haben die Norm

$$\|u_h\|_{B_h} := \max_{\overline{\Omega}_h} |u_h| = \max(\max_{\Gamma_h} |u_h|, \max_{\Omega_h} |u_h|).$$

Schließlich vervollständigen die Restriktionsoperatoren p_h und p_h^0 mit

$$\{p_h u(x,y)\}_{ij} := u(x_i,y_j), \quad \{p_h^0 v(x,y)\}_{ij} := v(x_i,y_j), \quad (x_i,y_j) \in \overline{\Omega}_h$$

das Diskretisierungsverfahren mit Diskretisierungsparameter $h \to 0$.

2. *Konsistenz:* Für $u \in D$ betrachten wir die Abbildung G_h des Diskretisierungsfehlers

$$\begin{aligned} G_h u &= F_h p_h u - p_h^0 F u \\ &= (\Delta_h\, p_h u(x,y) + p_h f(x,y)) - (p_h^0 \Delta\, u(x,y) + p_h^0 f(x,y)), \end{aligned}$$

woraus am Punkt (x_i,y_j) durch Taylor-Entwicklung analog zu (12)

$$\{G_h u\}_{ij} = \Delta_h\, u(x_i,y_j) - \Delta\, u(x_i,y_j) = \frac{1}{12}\left(\frac{\partial^4 u}{\partial x^4}(\xi_i,y_j) + \frac{\partial^4 u}{\partial y^4}(x_i,\eta_j)\right) h^2$$

mit $\xi_i, \eta_j \in (0,1)$ entsteht. Betragsabschätzung liefert

$$|\{G_h u\}_{ij}| \leq \frac{1}{12}(M_x + M_y)h^2 \qquad \text{mit} \quad M_x := \max_{\overline{\Omega}} \left|\frac{\partial^4 u}{\partial x^4}(x,y)\right|, \quad M_y := \max_{\overline{\Omega}} \left|\frac{\partial^4 u}{\partial y^4}(x,y)\right|.$$

Aus (20) und (21) folgt zudem, dass der Diskretisierungsfehler $G_h u$ auf dem Rand Γ verschwindet, womit für die Norm die Abschätzung

$$\|G_h u\|_{B_h} := \max_{(x_i,y_j) \in \overline{\Omega}_h} |\{G_h u\}_{ij}| \leq K\, h^2$$

mit einer Konstanten $K > 0$ erfüllt ist. Damit gilt

Satz 6 *Falls $u \in D = \mathcal{C}^4(\overline{\Omega})$, so ist das Verfahren (13), (14) konsistent mit Ordnung 2. Mit der exakten Lösung $u^* \in D$ genügt der lokale Diskretisierungsfehler der Abschätzung*

$$\|\tau_h\|_{B_h} \;=\; \|F_h p_h u^* - p_h^0 F u^*\|_{B_h} \;\leq\; K\,h^2, \quad K > 0. \tag{22}$$

3. *Diskrete Stabilität:* Zum Nachweis der diskreten Stabilitätseigenschaft benutzen wir einen Hilfssatz, dessen umfangreicher Beweis mit dem so genannten diskreten Maximumprinzip z.B. in [8, S. 461–465] zu finden ist.

Lemma 7 *v_h sei eine beliebige Gitterfunktion, definiert auf den Mengen Ω_h und Γ_h. Dann gilt mit dem diskretisierten Laplace-Operator Δ_h*

$$\max_{\Omega_h} |v_h| \;\leq\; \max_{\Gamma_h} |v_h| + \frac{1}{2} \max_{\Omega_h} |\Delta_h\, v_h|. \tag{23}$$

Betrachten wir zwei Gitterfunktionen u_h^1, u_h^2 und setzen sie für $(x, y) \in \Omega_h$ in (21) ein

$$\begin{aligned} \Delta_h\, u_h^1 + f &= F_h(u_h^1) \;=\; \delta_1 \\ \Delta_h\, u_h^2 + f &= F_h(u_h^2) \;=\; \delta_2, \end{aligned}$$

so ergibt Subtraktion die Gleichung für die Abweichung

$$\Delta_h(u_h^1 - u_h^2) \;=\; \delta_1 - \delta_2.$$

Wir wenden nun Lemma 7 auf die Gitterfunktion $v_h := u_h^1 - u_h^2$ an und erhalten

$$\begin{aligned} \max_{\Omega_h} |u_h^1 - u_h^2| &\leq \max_{\Gamma_h} |u_h^1 - u_h^2| + \frac{1}{2} \max_{\Omega_h} |\Delta_h\,(u_h^1 - u_h^2)| \\ &= \max_{\Gamma_h} |u_h^1 - u_h^2| + \frac{1}{2} \max_{\Omega_h} |\delta_1 - \delta_2| \\ &= \max_{\Gamma_h} |F_h(u_h^1) - F_h(u_h^2)| + \frac{1}{2} \max_{\Omega_h} |F_h(u_h^1) - F_h(u_h^2)|, \end{aligned}$$

da bei Subtraktion der Randbedingungen in (21) die Randwerte φ herausfallen. Übergang zur Norm und Abschätzung liefert die Stabilitätsungleichung

$$\|u_h^1 - u_h^2\|_{B_h} \;\leq\; S\,\|F_h(u_h^1) - F_h(u_h^2)\|_{B_h^0} \tag{24}$$

mit einer Stabilitätskonstanten $S = 3/2$. Damit ist die Diskretisierung stabil auf dem gesamten Raum B_h von Gitterfunktionen.

4. *Diskrete Konvergenz:* $u^* \in D$ sei die exakte Lösung des Dirichlet-Problems und $u_h \in B_h$ die nach Satz 3 garantierte Lösung des diskretisierten Problems (13), (14). Dann können wir die Konvergenz der Näherungslösung schlussfolgern:

Satz 9 *Falls $u \in \mathcal{C}^4(\overline{\Omega})$, so ist das Verfahren (13), (14) konvergent mit Ordnung 2. Es existiert eine von h unabhängige Konstante $C > 0$, mit der sich der globale Diskretisierungsfehler*

$$\|u_h - u^*\|_{B_h} \;=\; \max_{\overline{\Omega}_h} |u_{ij} - u^*(x_i, y_j)| \;\leq\; C\,h^2 \quad \textit{für alle} \quad h \in (0, 1] \tag{25}$$

abschätzen lässt.

Algorithmus 8 (FDM für Poisson-Gleichung)

Function $[u, n] =$ PoissonFDM (f, φ, N_x)

1. Bestimme $h = 1/N_x, n = (N_x - 1)^2$ und Gitterpunkte $(x_i, y_j) \in \Omega_h$.
2. Generiere die $n \times n$-Matrix A gemäß (17).
3. Generiere den n-Vektor a gemäß (16).
4. Löse das System $Av = a$ direkt oder iterativ nach v.
5. Transformiere v gemäß (15) in die $(N_x - 1) \times (N_x - 1)$-Matrix u.
6. Return u und n

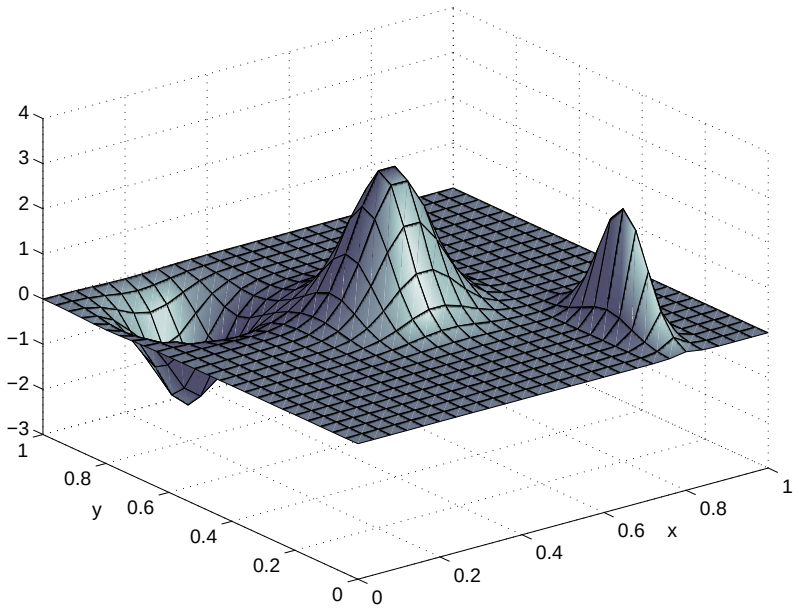

Abbildung 4: Quellterm $f(x, y)$ zu Beispiel 26 (1)

Beweis: Die Voraussetzungen des allgemeinen Konvergenzsatzes sind zu verifizieren. Wegen Ungleichung (24) ist dessen Stabilitätsvoraussetzung erfüllt, während Satz 6 die Konsistenzordnung 2 garantiert. Mit der Existenzvoraussetzung der Lösung folgt nun die Behauptung des Satzes. □

Der Algorithmus zur FDM-Lösung der Poisson-Gleichung mit Dirichletschen Randbedingungen ist nun recht einfach. Vorzugeben ist außer den Funktionen f und φ nur die Zahl $N = N_x = N_y$ der Teilintervalle, in die $[0, 1]$ unterteilt wird. Algorithmus 8 liefert dann die Näherungslösung u_h auf dem quadratischen $(N_x - 1) \times (N_y - 1)$-Gitter Ω_h.

Beispiel 10 Wir lösen die Poisson-Gleichung auf dem Einheitsquadrat $\Omega = (0, 1) \times (0, 1)$ mit homogenen Dirichlet-Bedingungen

$$-\Delta u \;=\; f(x, y), \quad (x, y) \in \Omega \qquad \text{und} \qquad u(x, y) \;=\; 0, \quad (x, y) \in \Gamma. \tag{26}$$

1. Der Quellterm f werde durch das MATLAB-File `f.m` mit

```
function u = f(x,y)
% Poisson-System - Inhomogenitaet
p1 = (x-0.2).^2 +(y-0.8).^2;
p2 = 12.*(x-0.8).^2 +(y-0.2).^2;
p3 = (x-0.5).^2 +(y-0.55).^2;
u = - 2.2*exp(-100*p1) + 2.6*exp(-90*p2) + 3.2*exp(-80*p3);
```

beschrieben und in Abb. 4 näherungsweise grafisch dargestellt. Wir wenden Algorithmus 8

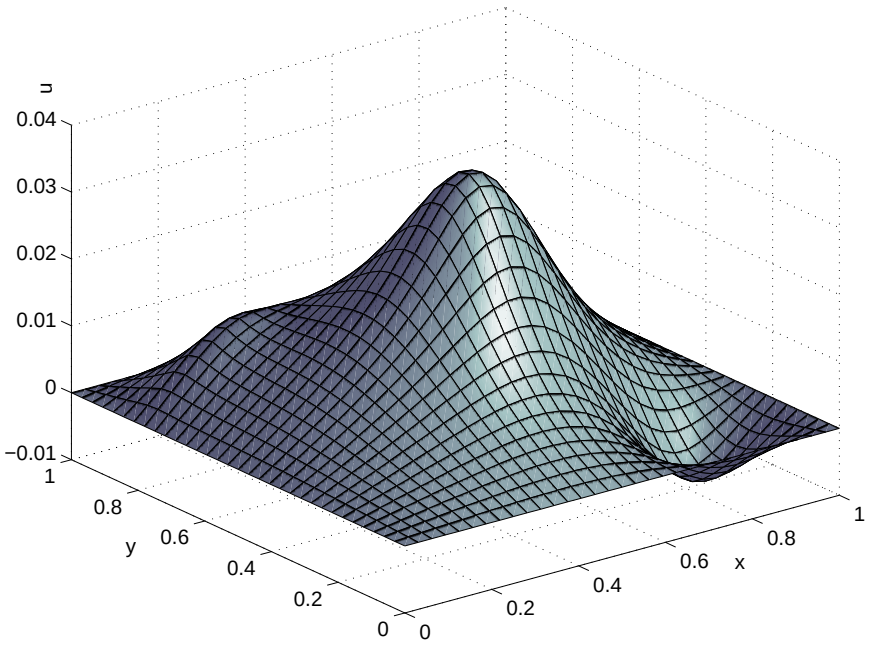

Abbildung 5: Näherungslösung u_h zu Beispiel 26(1) mit $N_x = N_y = 32$

mit $N_x = N_y = 32$ Teilintervallen an und erhalten damit auf dem 31×31-Gitter Ω_h die in Abb. 5 angegebene Näherungslösung u_h, über deren globalen Diskretisierungsfehler $u_h - u^*$ wir allerdings nichts aussagen können.

2. Wird die glatte Funktion f nun durch das MATLAB-File `f.m` mit

```
function u = f(x,y)
% Poisson-System - Inhomogenitaet
p1 = (x-0.5).^2 +(y-0.5).^2;
u =  100./(10.*sin(24.*p1).^2+0.01);
```

beschrieben, so erkennt man in Abb. 6 die scharf begrenzte ringförmige Struktur der Ladungsdichte.[4] Algorithmus 8 liefert mit $N_x = N_y = 64$ Teilintervallen die in Abb. 7 dargestellte Näherungslösung auf dem 63×63-Gitter Ω_h.

2.4 Asymptotische Fehlerschätzung

Kann höhere Glattheit der exakten Lösung $u(x, y)$ vorausgesetzt werden, so gewinnen wir mittels Taylor-Entwicklung des Differenzenoperators Δ_h analog zu (12) eine *asymptotische*

[4] Auch mit sehr feiner Auflösung wird f in der MATLAB-Grafik nur ungenau dargestellt.

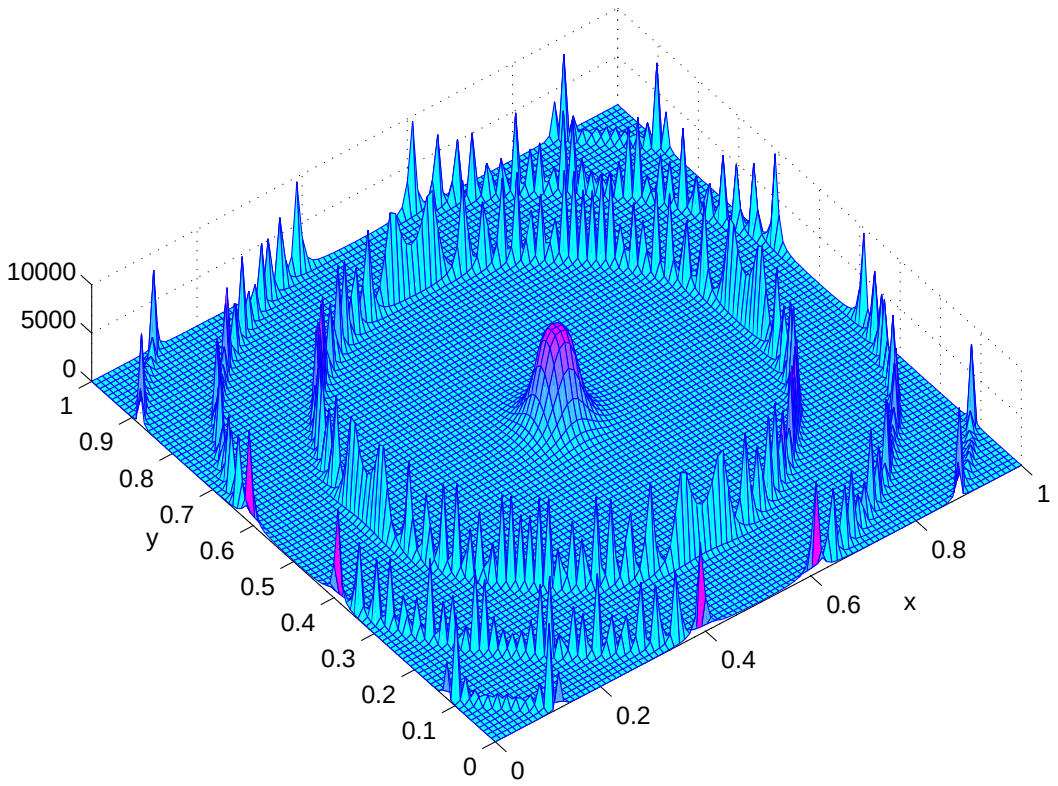

Abbildung 6: Ladungsdichte $f(x, y)$ zu Beispiel 26(2)

Entwicklung des lokalen Diskretisierungsfehlers

$$\tau_h \;=\; \Delta_h u - \Delta u \;=\; \sum_{k=2}^{K} \frac{2}{(2k)!} \left\{ \frac{\partial^{2k} u}{\partial x^{2k}} + \frac{\partial^{2k} u}{\partial y^{2k}} \right\} h^{2k-2} + \mathcal{O}(h^{2K}) \tag{27}$$

in geradzahligen Potenzen von h. In [17] wird allgemein nachgewiesen, dass dann auch der globale Diskretisierungsfehler e_h auf dem Gitter Ω_h eine derartige Entwicklung

$$e_h(x_i, y_j) \;=\; u_{ij} - u(x_i, y_j) \;=\; \sum_{k=2}^{K} e_k(x_i, y_j)\, h^{2k-2} + \mathcal{O}(h^{2K}) \tag{28}$$

mit den von h unabhängigen Funktionen e_k besitzt. Nach dem Extrapolationsprinzip lassen sich damit Verfahren der Konvergenzordnung 4 konstruieren: Außer der „Feinrechnung“ u_{ij}^F mit N Teilintervallen der Weite h bestimmen wir die „Grobwerte“ u_{ij}^G auf $N/2$ Intervallen der Weite $2h$. Mittels des Extrapolationsprinzips ergibt sich dann unmittelbar

Satz 11 *Falls $u \in \mathcal{C}^6(\overline{\Omega})$, so liefert die mit Verfahren (13), (14) gewonnene Größe*

$$error \;:=\; \frac{1}{3}(u_{ij}^G - u_j^F) \tag{29}$$

auf dem Grobgitter den asymptotischen Fehlerschätzer $error = e_h + \mathcal{O}(h^4)$. Mit seiner Hilfe kann die Feinlösung u_{ij}^F auf dem Grobgitter weiter zur Näherung

$$u_{ij}^{FG} \;=\; u_j^F - \frac{1}{3}(u_j^G - u_j^F) \tag{30}$$

mit dem globalen Diskretisierungsfehler $\mathcal{O}(h^4)$ verbessert werden.

N	h	$u_{\frac{N}{2},\frac{N}{2}}$	$error$	Q	$u^{FG}_{\frac{N}{2},\frac{N}{2}}$
8	1/8	3.557 460 e-2			
16	1/16	3.410 501 e-2	0.048 986 e-2		3.361 515 e-2
32	1/32	3.359 619 e-2	0.016 961 e-2	2.89	3.342 658 e-2
64	1/64	3.347 816 e-2	0.003 934 e-2	4.31	3.343 882 e-2
128	1/128	3.344 887 e-2	0.000 976 e-2	4.03	3.343 911 e-2
256	1/256	3.344 157 e-2	0.000 243 e-2	4.02	3.343 914 e-2

Tabelle 2: Poisson-Gleichung mit FDM 2. Ordnung und 4. Ordnung

Beispiel 12 Wir nutzen für die Poisson-Gleichung 6 verschiedene Gitter. In Tabelle 2 sind die am Mittelpunkt $(\frac{1}{2},\frac{1}{2})$ von Ω erhaltenen Näherungen $u_{\frac{N}{2},\frac{N}{2}}$ und die mit (29) berechneten Fehlerschätzungen dargestellt. Die Quotienten Q zweier untereinander stehender Fehler *error* zeigen die Fehlerordnung 2. Die Näherungen u^{FG}_{ij} gemäß (30) belegen, dass die Extrapolation zu einer spürbaren Verbesserung der Näherungswerte führt, ohne aufwändige Approximationen in Randnähe vornehmen zu müssen. □

Finite-Differenzenverfahren zeichnen sich durch einfache Näherungsformeln und universelle Anwendbarkeit aus. So ist das behandelte Verfahren 2. Ordnung prinzipiell auch auf nicht-

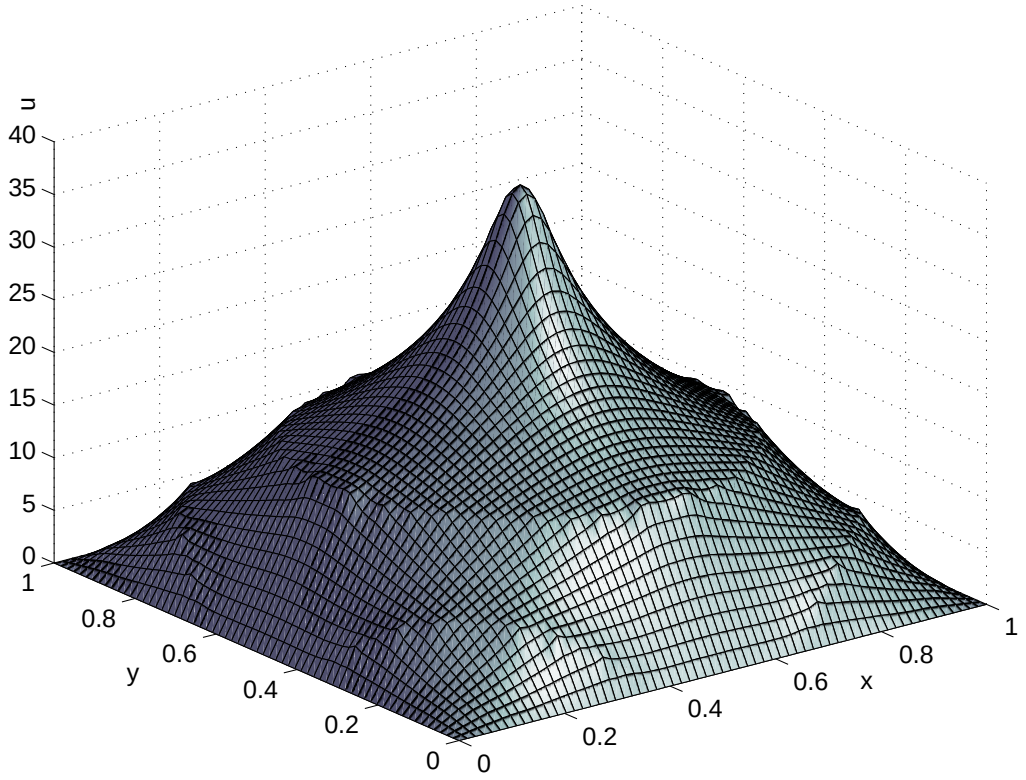

Abbildung 7: Näherungslösung u_h zu Beispiel 26 (2) mit $N_x = N_y = 64$

lineare PDGLn

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}(x)\frac{\partial^2 u}{\partial x_i \partial x_j} \;+\; f\left(x, u, \frac{\partial u}{\partial x_1}, \ldots, \frac{\partial u}{\partial x_n}\right) \;=\; 0 \tag{31}$$

über dem Einheitsintervall $\Omega = (0,1)^n \subset \mathbb{R}^n$ übertragbar, indem alle partiellen Ableitungen durch die beschriebenen zentralen Differenzen 2. Ordnung approximiert werden. Existenz- und Konvergenznachweise der Näherungslösungen sind dann oft aufwändig; die effiziente Lösung der entstehenden finiten nichtlinearen Gleichungssysteme wird z.B. in [7] behandelt.

Problematischer als Dirichlet-Bedingungen erweisen sich allgemeinere Randbedingungen. Ist $\Gamma = \Gamma_1 \cup \Gamma_2 \cup \Gamma_3$ mit abgeschlossener Menge Γ_3 eine disjunkte Zerlegung des Randes von $\Omega(0,1)^n$, so klassifiziert man die *Randbedingungen* bei gegebenen stetigen Funktionen $\varphi_i : \Gamma_i \to \mathbb{R},\ i = 1,2,3$ und $\alpha : \Gamma_3 \to \mathbb{R}$ als[5]

$$\begin{array}{lll}
\text{(1) Dirichlet-Bedingungen:} & u(x) = \varphi_1(x), & x \in \Gamma_1 \\
\text{(2) Neumann-Bedingungen:} & \frac{\partial u(x)}{\partial \mathbf{n}} = \varphi_2(x), & x \in \Gamma_2 \\
\text{(3) Gemischte Bedingungen:} & \alpha u(x) + \frac{\partial u(x)}{\partial \mathbf{n}} = \varphi_3(x), & x \in \Gamma_3
\end{array} \tag{32}$$

Darin bedeutet $\partial u(x)/\partial \mathbf{n}$ die Ableitung von u in Richtung der äußeren Normalen $\mathbf{n}$. Um die Konvergenzordnung 2 zu garantieren, müssen die Richtungsableitungen ebenfalls mit Ordnung 2 approximiert werden. Schwieriger gestalten sich Approximationen an randnahen Gitterpunkten bei krummlinig berandeten Gebieten (vgl. Abb. 1a). Hier sind geeignete Interpolationen erforderlich, die z.B. in [2, 18] behandelt werden. Bei PDGLn auf komplizierten Gebieten haben sich deshalb Finite-Elemente-Methoden besser bewährt.

3 Finite-Elemente-Methode

Anders als die im vorigen Abschnitt behandelten Finite-Differenzen-Methoden gehören die Finite-Elemente-Methoden (FEM) zur Klasse der Projektionsverfahren. Dabei wird das gegebene Problem in eine schwache (Variations-)Form überführt und mittels eines Galerkin- oder Kollokations-Ansatzes in endlichdimensionale Unterräume projiziert. Besonders geeeignet erweisen sich Unterräume stückweise polynomialer Funktionen, die auf „finiten Elementen“ definiert sind. Bei der Herleitung der teilweise abstrakten Methode orientieren wir uns an den gründlichen und vertiefenden Darstellungen in [3, 9, 10, 16].

3.1 Variationsgleichung und schwache Lösung

Wir beginnen mit der Variationsformulierung von Randwertproblemen. Ω sei ein beschränktes Gebiet in $\mathbb{R}^n$ $(n = 2,3)$ mit Lipschitz-stetigem Rand $\Gamma = \partial\Omega$. Wir betrachten Randwertprobleme für eine gesuchte Funktion u

$$Lu \;=\; f \quad \text{in } \Omega, \qquad Bu \;=\; 0 \quad \text{auf} \quad \partial\Omega, \tag{33}$$

[5] Die zugehörigen Probleme werden auch als 1., 2. und 3. Randwertproblem bezeichnet.

wobei f eine gegebene Funktion, L ein linearer Differenzialoperator und B ein affiner Rand-Operator[6] ist. Meistens ist L ein unbeschränkter Operator im Hilbert-Raum $H = \mathcal{L}^2(\Omega)$. Die Lösung u wird in einem Unterraum $U \subset H$ gesucht. Als konkretes Modell wollen wir elliptische Probleme 2. Ordnung betrachten. Der Differenzialoperator L sei deshalb stets durch

$$Lu \;:=\; -\sum_{i,j=1}^{n} D_i(a_{ij}D_j u) + \sum_{i=1}^{n} [D_i(b_i u) + c_i D_i u] + a_0 u \tag{34}$$

definiert. Wir bezeichnen die Ableitungen abkürzend mit dem Operator $D_i := \frac{\partial}{\partial x_i}$; die Funktionen $a_{ij} = a_{ij}(x),\ b_i = b_i(x),\ c_i = c_i(x),\ a_0 = a_0(x)$ sind gegeben.

Definition 13 *Der Differenzialoperator L heißt elliptisch in Ω, wenn eine Konstante $\alpha > 0$ existiert, so dass*

$$\sum_{i,j=1}^{n} a_{ij}(x)\xi_i\xi_j \;\geq\; \alpha\|\xi\|^2$$

für jedes $\xi \in \mathbb{R}^n$ und fast jedes $x \in \Omega$ gilt.

Das Problem (33) kann stets in eine Variationsgleichung (auch: schwache Form) umformuliert werden, wodurch wir schwache Lösungen erhalten, die im Unterschied zu den bisher vorausgesetzten klassischen Lösungen die Gleichungen (33) nicht notwendige punktweise erfüllen. Damit werden jedoch auch Anwendungen mit nicht glatten Daten möglich.

Wir leiten die schwache Form her, indem wir die Differenzialgleichungen $Lu = f$ mit so genannten *Testfunktionen* $v \in V$ multiplizieren und dann über Ω integrieren. Mittels partieller Integration gelingt es, die Differentiationsordnung der Lösung u zu reduzieren. Dabei wird die Formel

$$\int_\Omega \frac{\partial u}{\partial x_i} v\, dx \;=\; -\int_\Omega u\frac{\partial v}{\partial x_i}\, dx + \int_\Gamma u\, v\, n_i\, d\gamma, \qquad i = 1(1)n \tag{35}$$

genutzt, wobei $\mathbf{n} = (n_1, \ldots, n_n)$ der nach außen gerichtete Normalenvektor (die äußere Einheitsnormale) auf Γ ist. Im Ergebnis erhalten wir nach Einbeziehung der Randbedingungen $Bu = 0$ *eine Variationsgleichung* zu (33):

Definition 14 (Variationsgleichung, schwache Lösung) *U sei der Raum der zulässigen Lösungen (Grundraum) und V der Raum der Testfunktionen (Testraum).*

(i) Das Variationsproblem zu (33) lautet: Gesucht ist ein $u \in U$, das die Variationsgleichung

$$\mathcal{A}(u, v) \;=\; \mathcal{F}(v) \qquad \forall\, v \in V \tag{36}$$

erfüllt. Dabei ist $\mathcal{A} : U \times V \to \mathbb{R}$ eine zum Differenzialoperator L gehörende Bilinearform und $\mathcal{F} : V \to \mathbb{R}$ ein lineares Funktional, das durch die Funktion f und mögliche inhomogene Randbedingungen definiert wird.

(ii) $u \in U$ heißt schwache Lösung von (33), wenn es die Variationsgleichung (36) (auch: schwache Formulierung des Randwertproblems) erfüllt.

[6] Mitunter sind die Randbedingungen nur auf einer Teilmenge $\partial\Omega^*$ des Randes $\partial\Omega$ gegeben.

Für den in (34) definierten elliptischen Operator L liefert dieser Ansatz die Bilinearform

$$\mathcal{A}(u,v) \;:=\; \int\limits_{\Omega} \left\{ \sum_{i,j=1}^{n} a_{ij}\, D_j u\, D_i v - \sum_{i=1}^{n} (b_i\, u\, D_i v - c_i u\, D_i u) + a_0 uv \right\} \partial\Omega, \tag{37}$$

wobei u, v auf Ω definierte Funktionen sind. Werden die Randbedingungen $Bu = 0$ bereits bei der Definition des Raumes U berücksichtigt, so spricht man von *notwendigen (wesentlichen) Randbedingungen*. Andernfalls müssen *natürliche Randbedingungen* durch geeignete Wahl der Bilinearform $\mathcal{A}$ und des Funktionals $\mathcal{F}$ indirekt erzeugt werden. Die Räume U und V sind stets so zu wählen, dass die Formulierung (36) mathematisch gesehen Sinn macht. Oft ist $U = V \subset \mathcal{L}^2(\Omega)$ möglich, was anhand der Poisson-Gleichung demonstriert werden soll. Zuvor wollen wir diese Funktionenräume konkretisieren.

Der Raum $\mathcal{L}^2(\Omega)$ der auf $\Omega \subset \mathbb{R}^n$ messbaren Funktionen $v : \Omega \to \mathbb{R}$ mit $\int_\Omega |v(x)|^2\, dx < \infty$ ist ein Hilbert-Raum mit Skalarprodukt

$$\langle u, v\rangle_0 \;=\; \int\limits_{\Omega} u(x)v(x)\, dx \qquad \text{und Norm} \qquad \|v\|_0 \;=\; \left\{ \int\limits_{\Omega} |v(x)|^2\, dx \right\}^{1/2}. \tag{38}$$

Um geeignete Funktionenräume U und V differenzierbarer Funktionen anzugeben, führen wir die Multiindex-Notation ein: Ein Vektor $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ mit nichtnegativen ganzen Zahlen α_i heißt *Multiindex der Länge (Ordnung)* $|\alpha| = \alpha_1 + \alpha_2 + \ldots + \alpha_n$. Für Ableitungen der Ordnung $|\alpha|$ nach den Variablen $x = (x_1, x_2, \ldots, x_n)$ notieren wir damit abkürzend

$$D^\alpha u \;:=\; \frac{\partial^{\alpha_1+\alpha_2+\ldots+\alpha_n} u}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \cdots \partial x_n^{\alpha_n}} \tag{39}$$

mit dem Differenzialoperator D. Setzen wir $u \in \mathcal{C}^s(\Omega)$ mit $s = |\alpha|$ voraus, so ist diese Ableitung $D^\alpha u$ definiert und stetig. Mit Funktionen $\varphi \in \mathcal{C}_0^\infty(\Omega)$ aus dem Testraum

$$\mathcal{C}_0^\infty(\Omega) \;=\; \{\varphi \in \mathcal{C}^\infty(\Omega) \mid D^\beta \varphi = 0 \text{ auf } \partial\Omega \text{ für alle Multiindizes } \beta\}$$

liefert dann partielle Integration wegen der verschwindenden Integrale auf $\partial\Omega$

$$\int\limits_{\Omega} D^\alpha u\, \varphi\, dx \;=\; (-1)^s \int\limits_{\Omega} u\, D^\alpha \varphi\, dx. \tag{40}$$

Die Ausdehnung dieser Beziehung auf beliebige Funktionen $u \in \mathcal{L}^2(\Omega)$ führt uns zum Begriff der *verallgemeinerten (schwachen) Ableitung*:

Definition 15 (Verallgemeinerte Ableitung) *Zu gegebener Funktion $u \in \mathcal{L}^2(\Omega)$ heißt $v \in \mathcal{L}^2(\Omega)$ verallgemeinerte oder schwache Ableitung zum Multiindex α, wenn*

$$\int\limits_{\Omega} v\, \varphi\, dx \;=\; (-1)^{|\alpha|} \int\limits_{\Omega} u\, D^\alpha \varphi\, dx \qquad \textit{für alle} \quad \varphi \in \mathcal{C}_0^\infty(\Omega) \tag{41}$$

gilt. Schreibweise: $v = D^\alpha u$

Die verallgemeinerte Ableitung ist eindeutig bestimmt, d.h. ist $v_1 = D^\alpha u$ und $v_2 = D^\alpha u$, so $v_1 = v_2$ fast überall. Jede s-mal stetig differenzierbare Funktion $u \in \mathcal{C}^s(\Omega)$ besitzt wegen (40) alle verallgemeinerten Ableitungen $D^\alpha u$ für $|\alpha| \leq s$, die mit den klassischen (punktweisen) Ableitungen übereinstimmen. Mit dieser Verallgemeinerung des Ableitungsbegriffes sind wir nun imstande, geeignete Unterräume $U = V \subset \mathcal{L}^2(\Omega)$ zu definieren.

Definition 16 (Sobolev-Raum) *Zu gegebener nichtnegativer ganzer Zahl s enthält der Sobolev-Raum*[7]

$$\mathcal{H}^s(\Omega) \;=\; \{u \in \mathcal{L}^2(\Omega) \mid D^\alpha u \in \mathcal{L}^2(\Omega) \;\; \textit{für} \;\; |\alpha| \leq s\} \tag{42}$$

alle Funktionen aus $\mathcal{L}^2(\Omega)$, deren verallgemeinerte Ableitungen $D^\alpha u$ bis einschließlich zur Ordnung s ebenfalls zu $\mathcal{L}^2(\Omega)$ gehören[8]*. Damit ist $\mathcal{H}^0(\Omega) = \mathcal{L}^2(\Omega)$ und $\mathcal{H}^{s_1}(\Omega) \supset \mathcal{H}^{s_2}(\Omega)$ für $s_1 < s_2$.*

Bezüglich der Struktureigenschaften der $\mathcal{H}^s$-Räume verweisen wir auf die Literatur [12]. Wir wollen lediglich feststellen, dass $\mathcal{H}^s(\Omega)$ ein Banach-Raum bezüglich der Norm

$$\|u\|_s \;=\; \left\{ \sum_{|\alpha| \leq s} \|D^\alpha u\|_0^2 \right\}^{1/2} \tag{43}$$

mit der in (38) definierten $\mathcal{L}^2$-Norm $\|u\|_0$ ist. Wie sind diese Räume konkret in den Anwendungen zu wählen?

1. *Das Dirichlet-Problem:* Wir betrachten die n-dimensionale Poisson-Gleichung mit homogenen Dirichlet-Bedingungen

$$-\Delta u \;=\; f \quad \text{in} \quad \Omega, \qquad u \;=\; 0 \quad \text{auf} \quad \partial\Omega, \tag{44}$$

wobei $f \in \mathcal{L}^2(\Omega)$ eine gegebene Funktion und $\Delta := \sum_{i=1}^n \frac{\partial^2}{\partial x_i^2}$ der Laplace-Operator ist. Um die Variationsform zu finden, nutzen wir die Greensche Formel für diesen Operator

$$-\int_\Omega \Delta u \; v \; dx \;=\; \int_\Omega \nabla u \cdot \nabla v \; dx - \int_{\partial\Omega} \frac{\partial u}{\partial \mathbf{n}} \; v \; d\gamma$$

und gewinnen wegen $u = 0$ auf Γ unter Beachtung des verschwindenden Randintegrals

$$\int_\Omega \nabla u \cdot \nabla v \; dx \;=\; \int_\Omega f v \; dx.$$

Wir berücksichtigen nun die wesentlichen (notwendigen) Randbedingungen bereits in der Definition der beiden Funktionenräume U und V, indem wir $U = V = \mathcal{H}_0^1(\Omega)$ mit

$$\mathcal{H}_0^1(\Omega) \;=\; \{u \in \mathcal{L}^2(\Omega) \mid D^\alpha u \in \mathcal{L}^2(\Omega) \text{ für } |\alpha| \leq 1 \text{ und } u = 0 \text{ auf } \partial\Omega\} \tag{45}$$

[7] Sergej Lvovitsch Sobolev (1908–1989), russischer Mathematiker.

[8] Wegen $\mathcal{H}^s(\Omega) = \mathcal{W}^{s,2}(\Omega)$ bilden diese Räume einen Spezialfall der Räume $\mathcal{W}^{s,p}$.

wählen. Der Sobolev-Raum $\mathcal{H}_0^1(\Omega)$ enthält somit alle Funktionen aus $\mathcal{H}^1(\Omega)$, die die homogenen Randbedingungen erfüllen. Das abstrakte Variationsproblem für (44) kann leicht konkretisiert werden, wenn wir

$$\mathcal{A}(u,v) \equiv a(u,v) = \int_\Omega \nabla u \cdot \nabla v \, dx \quad \text{und} \quad \mathcal{F}(v) \equiv \langle u, v\rangle_0 = \int_\Omega f v \, dx \tag{46}$$

mit dem Skalarprodukt $\langle u, v\rangle_0$ von $\mathcal{L}^2(\Omega)$ und der Bilinearform $a(u,v)$ setzen. Zu gegebenem $f \in \mathcal{L}^2(\Omega)$ können wir die Variationsgleichung nun in der üblichen Form notieren:

$$\text{Finde } u \in \mathcal{H}_0^1(\Omega) \; : \quad a(u,v) = \langle f, v\rangle_0 \qquad \forall\, v \in \mathcal{H}_0^1(\Omega). \tag{47}$$

2. *Das Neumann-Problem*[9]: Die Poisson-Gleichung habe nun Neumannsche Randbedingungen

$$-\Delta u = f \quad \text{in} \quad \Omega, \qquad \frac{\partial u}{\partial n_L} = g \quad \text{auf} \quad \partial\Omega \tag{48}$$

mit den gegebenen Funktionen $f \in \mathcal{L}^2(\Omega)$ und $g \in \mathcal{L}^2(\partial\Omega)$, wobei

$$\frac{\partial u}{\partial n_L} := \sum_{i,j=1}^{n} a_{ij} D_j u \, n_i - \mathbf{b} \cdot \mathbf{n} \, u \tag{49}$$

die konormale Ableitung von u bezeichnet. Falls $\mathbf{b} = 0$ und $a_{ij} = \delta_{ij}$ ist, so gibt (49) die Ableitung von u in Richtung der äußeren Einheitsnormalen $\mathbf{n}$ an. Man überlegt sich, dass die schwache Formulierung hierzu

$$\text{Finde } u \in \mathcal{H}^1(\Omega) \; : \; a(u,v) = \langle f, v\rangle_0 + \langle g, v\rangle_{\partial\Omega} \qquad \forall\, v \in \mathcal{H}^1(\Omega) \tag{50}$$

lautet, wobei $\langle g, v\rangle_{\partial\Omega}$ das Skalarprodukt auf $\mathcal{L}^2(\partial\Omega)$ ist. Da der Sobolev-Raum

$$\mathcal{H}^1(\Omega) = \{u \in \mathcal{L}^2(\Omega) \mid D^\alpha u \in \mathcal{L}^2(\Omega) \text{ für } |\alpha| \le 1\} \tag{51}$$

nun die Randbedingungen nicht beinhaltet, wurden diese (natürlichen) Bedingungen in der schwachen Formulierung berücksichtigt.

3. *Das gemischte Problem:* Falls sich der Rand Γ aus $\Gamma = \overline{\Gamma}_D \cup \overline{\Gamma}_N$ mit offenen Teilmengen Γ_D und Γ_N zusammensetzt, so kann für das gemischte Problem

$$-\Delta u = f \quad \text{in} \quad \Omega, \qquad u = 0 \quad \text{auf} \quad \Gamma_D, \qquad \frac{\partial u}{\partial n_L} = g \quad \text{auf} \quad \Gamma_N \tag{52}$$

die schwache Formulierung ebenfalls leicht bestimmt werden:

$$\text{Finde } u \in \mathcal{H}^1_{\Gamma_D}(\Omega) \; : \; a(u,v) = \langle f, v\rangle_\Omega + \langle g, v\rangle_{\Gamma_N} \qquad \forall\, v \in \mathcal{H}^1_{\Gamma_D}(\Omega). \tag{53}$$

Darin beinhaltet der Sobolev-Raum $\mathcal{H}^1_{\Gamma_D}(\Omega)$ bereits die Randbedingung $u = 0$ auf Γ_D.

[9] John von Neumann (1903–1957), ungarisch stämmiger Mathematiker und Physiker, einer der vielseitigsten Wissenschaftler des 20 Jahrhunderts mit herausragenden Beiträgen zur Numerik partieller Differenzialgleichungn, Mengentheorie, Spieltheorie und Computerarchitektur.

Kann eine eindeutige klassische Lösung u des gegebenen Randwertproblems (33) vorausgesetzt werden, so folgt unter geeigneten Annahmen über U und V, dass u auch eine schwache Lösung darstellt. Der Existenznachweis einer eindeutigen schwachen Lösung hingegen erfordert weitere Annahmen an das Problem. Wir geben für den Fall $U = V$ das bekannteste Theorem an, dessen Beweis man u.a. in [1] finden kann.

Satz 17 (Lax-Milgram) *V sei ein Hilbert-Raum mit der Norm $\|v\|$. Gegeben sind eine Bilinearform $\mathcal{A}(u,v) : V \times V \to \mathbb{R}$ und ein lineares stetiges Funktional*[10] *$\mathcal{F}(v) : V \to \mathbb{R}$. Die Bilinearform $\mathcal{A}$ sei stetig, d.h. es existiert ein $\gamma > 0$ mit*

$$|\mathcal{A}(u,v)| \;\leq\; \gamma \|u\| \|v\| \qquad \forall\, u, v \in V \tag{54}$$

und koerziv, d.h. es existiert ein $\alpha > 0$ mit

$$\mathcal{A}(v,v) \;\geq\; \alpha \, \|v\|^2 \qquad \forall\, v \in V. \tag{55}$$

Dann existiert eine eindeutige schwache Lösung $u \in V$ von (36) mit der Abschätzung

$$\|u\| \;\leq\; \frac{1}{\alpha} \, \|\mathcal{F}\|_{V'}. \tag{56}$$

Falls die Bilinearform $\mathcal{A}$ symmetrisch ist, d.h. $\mathcal{A}(u,v) = \mathcal{A}(v,u)\ \forall\, u, v \in V$, so definiert $\mathcal{A}$ ein Skalarprodukt auf V und wir können folgendes *Minimierungsproblem* formulieren:

$$\text{Finde } u \in V \;:\; J(u) \;\leq\; J(v) \qquad \forall\, v \in V, \tag{57}$$

wobei J das quadratische Funktional (auch *Energiefunktional* genannt)

$$J(v) \;:=\; \frac{1}{2} \, \mathcal{A}(v,v) - \mathcal{F}(v) \tag{58}$$

ist. Unter den Voraussetzungen des Satzes 17 hat diese Minimierungsaufgabe dieselben Lösungen wie die Variationsgleichung (36) (vgl. [10]).

Beispiel 18 Das Energiefunktional für das Dirichlet-Problem lautet

$$J(v) \;:=\; \frac{1}{2} \, a(u,v) - \langle f, v \rangle_0 \;=\; \frac{1}{2} \int\limits_\Omega \nabla u \cdot \nabla v \, dx - \int\limits_\Omega f v \, dx,$$

für das Neumann-Problem erhalten wir

$$J(v) \;:=\; \frac{1}{2} \, a(u,v) - \langle f, v \rangle_0 - \langle g, v \rangle_{\partial\Omega},$$

während sich im gemischten Fall

$$J(v) \;:=\; \frac{1}{2} \, a(u,v) - \langle f, v \rangle_0 - \langle g, v \rangle_{\Gamma_N}$$

ergibt. □

[10] $\mathcal{F} \in V'$ mit dem Dualraum V' von V

3.2 Galerkin-Verfahren

Wir wollen nun ein Diskretisierungsverfahren zur Lösung der Variationsgleichung (36) im Falle übereinstimmender Hilbert-Räume $U = V$ konstruieren[11]. Der Diskretisierungsparameter soll stets mit $h > 0$ bezeichnet und als klein angenommen werden. h wird später durch den maximalen Durchmesser der finiten Elemente beschrieben. Mit $V_h \subset V,\ h > 0$ kennzeichnen wir eine Familie endlichdimensionaler Unterräume von V. Wir setzen voraus, dass die Beziehung

$$\inf_{v_h \in V_h} \|v - v_h\| \to 0, \quad \text{falls } h \to 0 \qquad \text{für alle } v \in V \tag{59}$$

zwischen den Räumen V_h und V erfüllt ist und ersetzen die (unendlichdimensionale) Variationsgleichung in V durch das endlichdimensionale Problem in V_h:

Definition 19 (Galerkin-Verfahren, Ritz-Verfahren) *Zu gegebenem $\mathcal{F} \in V'$ lautet das Galerkin-Verfahren[12]: Gesucht ist ein $u_h \in V_h$, das die endlichdimensionale Variationsgleichung*

$$\mathcal{A}(u_h, v_h) \;=\; \mathcal{F}(v_h) \qquad \forall\, v_h \in V_h \tag{60}$$

erfüllt. V_h heißt auch Ansatzraum. Das Ritz-Verfahren betimmt ein $u_h \in V_h$, das das endlichdimensionale Minimierungsproblem löst:

$$J(u_h) \;=\; \min_{v_h \in V_h} J(v_h) \qquad \textit{mit} \qquad J(v_h) \;:=\; \frac{1}{2}\,\mathcal{A}(v_h, v_h) - \mathcal{F}(v_h). \tag{61}$$

Bevor wir konkrete Darstellungen der Ansatzräume betrachten, wollen wir die Frage nach der Existenz von approximierenden Lösungen u_h klären und deren Konvergenz nachweisen. Dazu beweisen wir

Satz 20 (Céa) *Mit den Voraussetzungen des Lax-Milgram-Lemmas 17 existiert eine eindeutige Lösung u_h von (60) mit*

$$\|u_h\| \;\le\; \frac{1}{\alpha}\,\|\mathcal{F}\|_{V'}, \tag{62}$$

die damit numerisch stabil ist. Mit der Lösung u von (36) gilt die Fehlerschätzung

$$\|u - u_h\| \;\le\; \frac{\gamma}{\alpha} \inf_{v_h \in V_h} \|u - v_h\|, \tag{63}$$

woraus wegen Voraussetzung (59) die Konvergenz von u_h gegen u folgt.

Beweis: Da V_h ein Unterraum von V ist, gelten die Voraussetzungen des Satzes 17 auch in V_h, woraus die Existenz und Eindeutigkeit von $u_h \in V_h$ folgt. Nutzen wir die Koerzivität von $\mathcal{A}$ mit $v_h = u_h$, so erhalten wir

$$\alpha\,\|u_h\|^2 \;\le\; \mathcal{A}(u_h, u_h) \;=\; \mathcal{F}(u_h) \;\le\; \|\mathcal{F}\|_{V'}\,\|u_h\|$$

[11] Für den allgemeineren Fall mit $U \neq V$ findet der Leser einen Existenzsatz in [15].

[12] Boris Grigorjevich Galerkin (1871–1945), sowjetischer Ingenieur und Mathematiker, veröffentlichte 1915 seine Methode, die die Tradition der Variationsmethoden (Euler, Lagrange, Hamilton) fortsetzt.

und damit die Ungleichung (63). Hieraus ergibt sich die Stabilität der Diskretisierung, da $\|\mathcal{F}\|_{V'}$ und α unabhängig von h sind. Betrachten wir die Variationsgleichung (36) für eine Testfunktion $v_h = w_h - u_h, w_h \in V_h$, und subtrahieren davon die Galerkin-Gleichung (60), so erhalten wir

$$\mathcal{A}(u - u_h, w_h - u_h) \;=\; 0 \qquad \forall\, v_h = w_h - u_h \in V_h. \tag{64}$$

Mit den Voraussetzungen (54) und (55) folgen dann die Abschätzungen

$$\begin{aligned} \alpha \,\|u - u_h\|^2 \;&\leq\; \mathcal{A}(u - u_h, u - u_h) \;=\; \mathcal{A}(u - u_h, (u - w_h) + (w_h - u_h)) \\ &\leq\; \gamma \,\|u - u_h\| \,\|u - w_h\| \quad \forall\, w_h \in V_h, \end{aligned}$$

womit die Fehlerschätzung (63) gezeigt wurde. □

Für eine symmetrische Bilinearform $\mathcal{A}$ stellt (64) eine Orthogonalitätsbedingung dar. Gesucht ist ein $u_h \in V_h$, so dass der Fehler $u - u_h$ senkrecht auf allen Elementen des Unterraumes V_h steht, d.h.

$$\mathcal{A}(u - u_h, v_h) \;=\; 0 \qquad \forall\, v_h \in V_h \tag{65}$$

im Sinne des durch $\mathcal{A}$ erzeugten Skalarproduktes. Damit ist u_h die orthogonale Projektion von u auf V_h mit diesem Skalarprodukt.

Wir wollen nun die N_h-dimensionalen Vektorräume V_h konkretisieren. Die linear unabhängigen Elemente $\varphi_j \in V_h$ bilden eine Basis $B_h = \{\varphi_j \mid j = 1(1)N_h\}$, mit der wir $u_h \in V_h$ durch

$$u_h = \sum_{i=1}^{N_h} \xi_i \varphi_i \tag{66}$$

darstellen können. Die Verfahrensgleichung (60) ist genau dann für jedes Element $v_h \in V_h$ gültig, wenn sie durch jede Basisfunktion $v_h = \varphi_j,\ j = 1(1)N_h$, erfüllt wird. Damit liefert Einsetzen des Ansatzes (66) mit der Bilinearität von $\mathcal{A}$ das lineare algebraische Gleichungssystem der Dimension N_h

$$\sum_{i=1}^{N_h} \mathcal{A}(\varphi_i, \varphi_j)\, \xi_i \;=\; \mathcal{F}(\varphi_j), \qquad j = 1(1)N_h \tag{67}$$

für die Koeffizienten ξ_i. Mit den Abkürzungen $b_i = \mathcal{F}(\varphi_i)$ und $A_{ij} = \mathcal{A}(\varphi_i, \varphi_j)$ für $i, j = 1(1)N_h$ können wir dieses System vektoriell als

$$A\,\xi \;=\; b \qquad \text{mit} \qquad A \;=\; (A_{ij}),\ b \;=\; (b_i),\ \xi \;=\; (\xi_i) \tag{68}$$

notieren. Aus der Mechanik wurden die Bezeichnungen *Steifigkeitsmatrix* für A und *Lastvektor* für den Vektor b übernommen. Welche Eigenschaften besitzt die Matrix A? Bezeichne $\langle \xi, \eta \rangle$ das Euklidische Skalarprodukt in $\mathbb{R}^{N_h}$. Mit der Darstellung (66) gilt für jedes Element $u_h \in V_h$

$$\begin{aligned} \mathcal{A}(u_h, u_h) \;&=\; \mathcal{A}\left(\sum_{i=1}^{N_h} \xi_i \varphi_i, \sum_{j=1}^{N_h} \xi_j \varphi_j\right) \;=\; \sum_{i=1}^{N_h} \sum_{j=1}^{N_h} \xi_i \mathcal{A}(\varphi_i, \varphi_j) \xi_j \\ &=\; \sum_{i=1}^{N_h} \sum_{j=1}^{N_h} \xi_i A_{ij} \xi_j \;=\; \langle A\xi, \xi \rangle. \end{aligned}$$

Aus der Koerzitivität von $\mathcal{A}$ folgt damit für jedes $\xi \in \mathbb{R}^{N_h}, \xi \neq 0$ wegen $\langle A\xi, \xi\rangle > 0$ die positive Definitheit der Steifigkeitsmatrix A. Insbesondere haben alle Eigenwerte von A einen positiven Realteil, womit die Existenz und Eindeutigkeit einer Lösung von (60) auch rein algebraisch begründet wird. Wir fassen zusammen zu

Folgerung 21 *Unter den Voraussetzungen des Lax-Milgram-Lemmas 17 gilt mit dem Euklidischen Skalarprodukt in $\mathbb{R}^{N_h}$:*

(i) *Die Steifigkeitsmatrix A ist positiv definit.*

(ii) *System $A\xi = F$ ist eindeutig lösbar.*

(iii) *Falls die Bilinearform $\mathcal{A}$ symmetrisch ist, so ist auch die Matrix A symmetrisch.*

3.3 Triangulierung und finite Elemente

Nun bleibt die Frage zu klären, wie im konkreten Fall des Differenzialoperators L oder speziell für die Poisson-Gleichung die Räume V_h und deren Basen zu wählen sind, damit die Approximationseigenschaft (59) erfüllt ist.

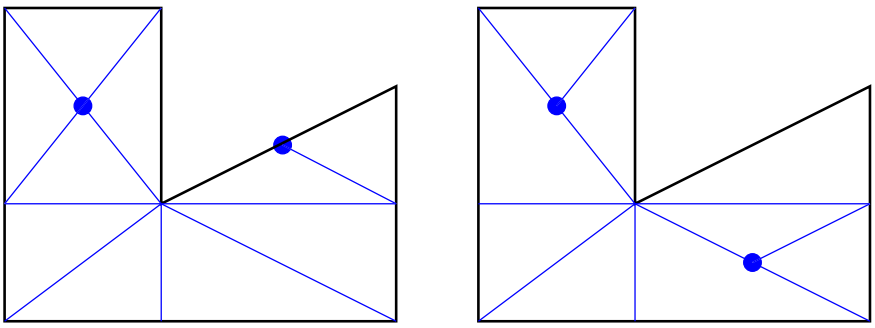

Abbildung 8: (a) Zulässige Triangulierung und (b) unzulässige Triangulierung

Triangulierung Der Bereich $\Omega \subset \mathbb{R}^n\,(n = 2, 3)$ soll wie in Abb. 8 stets eine *polygonale Menge* sein, d.h. eine offene beschränkte zusammenhängende Menge, so dass $\overline{\Omega}$ die Vereinigung endlich vieler Polyeder ist. Die endliche Zerlegung

$$\overline{\Omega} = \bigcup_{K \in \mathcal{T}_h} K \tag{69}$$

mit den *Elementen* $K = K_i$, $i = 1(1)I_h$ erfülle zu $h > 0$ folgende Bedingungen:

1. Jedes K_i ist ein Polyeder mit $\operatorname{int}(K_i) \neq \emptyset$.
2. $\operatorname{int}(K_i) \cap \operatorname{int}(K_j) = \emptyset$ für jedes $i \neq j$.
3. Falls $F = K_1 \cap K_2 \neq \emptyset$, so ist F eine gemeinsame Fläche, Seite oder ein gemeinsamer Eckpunkt von K_1 und K_2.
4. $h = \max\limits_{K \in \mathcal{T}_h} \operatorname{diam}(K)$ mit Durchmesser $\operatorname{diam}(K) := \sup\{|x - y| \mid x, y \in K\}$

Dann heißt $\mathcal{T}_h = \{K_1, K_2, \ldots, K_{I_h}\}$ eine *zulässige Triangulierung* von $\overline{\Omega}$ der Feinheit h. In Abb. 8 sind zulässige und unzulässige Triangulierungen dargestellt. Wir wollen nur zulässige Triangulierungen mit Dreiecken in $\mathbb{R}^2$ bzw. Tetraedern in $\mathbb{R}^3$ betrachten, deren Feinheit h

dann durch die längste Seite aller Dreiecke (bzw. die längste Kante aller Tetraeder) bestimmt ist. Bezüglich der Zerlegung in Parallelepipede (speziell in Rechtecke bzw. Quader) verweisen wir auf die eingangs genannte Literatur. Die Ecken der Elemente K werden *Knoten*

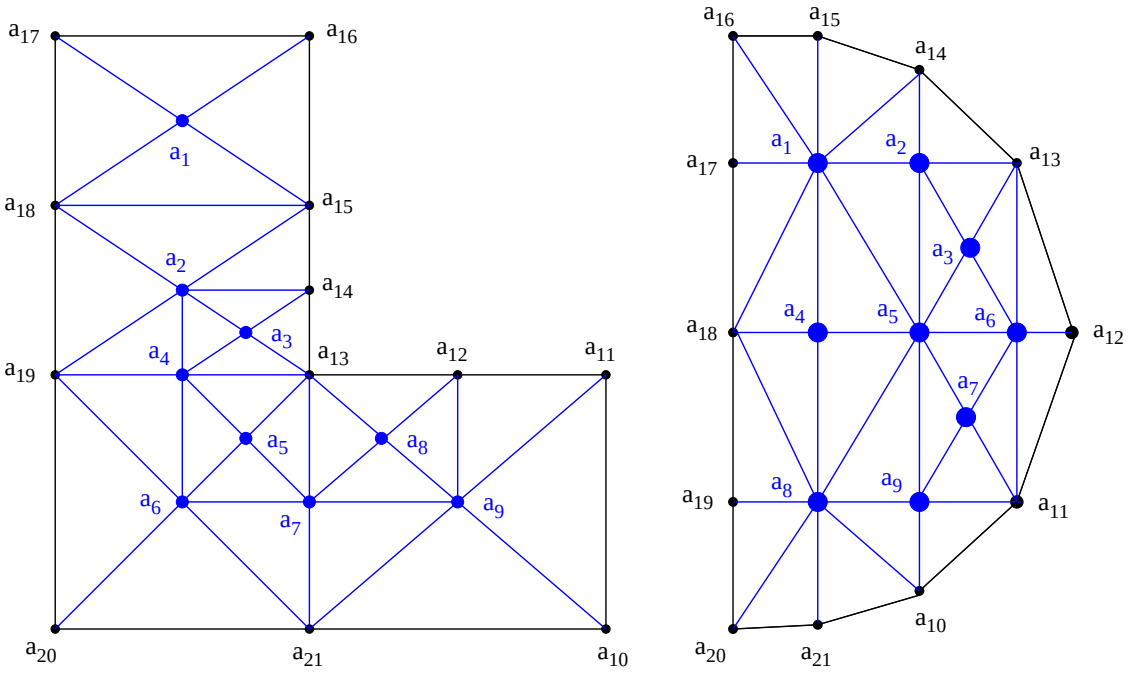

Abbildung 9: Triangulierung mit 9 inneren Knoten (blau) und 12 Randknoten (schwarz)

genannt und können wie z.B. in Abb. 9 mit $a_1, a_2, \ldots, a_{N_h}$ durchnummeriert werden. Wir unterscheiden dabei innere Knoten $a_i \in \Omega$ und Randknoten $a_i \in \partial\Omega$. Über der Triangulierung führen wir Räume X_h stückweise polynomialer Funktionen ein, d.h. für jedes $K \in \mathcal{T}_h$ ist $v_h \in X_h$ ein algebraisches Polynom. Bezeichnen wir mit $\mathcal{P}_k$, $k \geq 0$ den Raum der Polynome mit Grad kleiner oder gleich k in den Variablen $x_1, \ldots, x_n$, so hat dessen Dimension den Wert

$$\dim \mathcal{P}_k \;=\; \binom{n+k}{k}. \tag{70}$$

Der Raum X_h der *dreieckigen finiten Elemente (FE)* lässt sich durch

$$X_h \;=\; X_h^k \;=\; \big\{v_h \in \mathcal{C}^0(\overline{\Omega}) \mid v_{h|K} \in \mathcal{P}_k \;\; \forall K \in \mathcal{T}_h\big\}, \quad k \geq 1 \tag{71}$$

definieren[13]. Mit der nur für den $\mathbb{R}^2$ korrekten Ausdrucksweise „dreieckige finite Elemente“ sind allgemein $(n+1)$-eckige Elemente $K \subset \mathbb{R}^n$ gemeint, d.h. Tetraeder im 3D-Fall. Der Raum X_h wird als *Finite-Elemente-Raum (FE-Raum)* bezeichnet. Mit der Wahl (71) kann gezeigt werden, dass $X_h^k \subset \mathcal{H}^1(\Omega)$ für alle $k \geq 1$ mit dem durch (51) definierten Sobolev-Raum $\mathcal{H}^1(\Omega)$ gilt. Ob X_h^k bereits der geeignete Ansatzraum V_h ist oder weiter eingeschränkt werden muss, hängt allerdings von den konkreten Randbedingungen ab.

Beispiel 22 (Poisson-Gleichung) Für das Dirichlet-Problem (44) wählen wir

$$V_h \;=\; X_h^k \cap \mathcal{H}_0^1(\Omega) \;=\; \{v_h \in X_h^k \mid v_h = 0 \text{ auf } \partial\Omega\}, \quad k \geq 1, \tag{72}$$

[13] Räume parallelepepidaler finiter Elemente wollen wir hier nicht betrachten.

während im Falle des Neumann-Problems (48) der FE-Raum

$$V_h \;=\; X_h^k, \quad k \geq 1 \tag{73}$$

und für gemischte Probleme (52) schließlich

$$V_h \;=\; X_h^k \cap \mathcal{H}^1_{\Gamma_D}(\Omega) \;=\; \{v_h \in X_h^k \mid v_h = 0 \text{ auf } \Gamma_D\}, \quad k \geq 1 \tag{74}$$

problemadäquat ist. Im letzten Fall muss die Triangulierung so ausgeführt werden, dass kein Element $K \in \mathcal{T}_h$ beide Ränder Γ_D und Γ_N schneidet. □

Formfunktionen Wir wollen nun eine Basis $B_h = \{\varphi_i(x) \mid i = 1(1)N_h\}$ des Raumes X_h^k mit dem Ziel konstruieren, dass die Basisfunktionen $\varphi_i(x)$ möglichst einfach beschrieben werden können. Dazu wählen wir sie so, dass an den Knoten $a_1, a_2, \ldots, a_{N_h}$ der Triangulierung von $\overline{\Omega}$ die Bedingung

$$\varphi_i(a_j) \;=\; \delta_{ij}, \quad i,j = 1,\ldots,N_h \qquad \text{mit Kronecker-Symbol } \delta_{ij} \tag{75}$$

erfüllt ist. Diese Basisfunktionen werden *Formfunktionen* genannt. Wesentlich ist dabei, dass

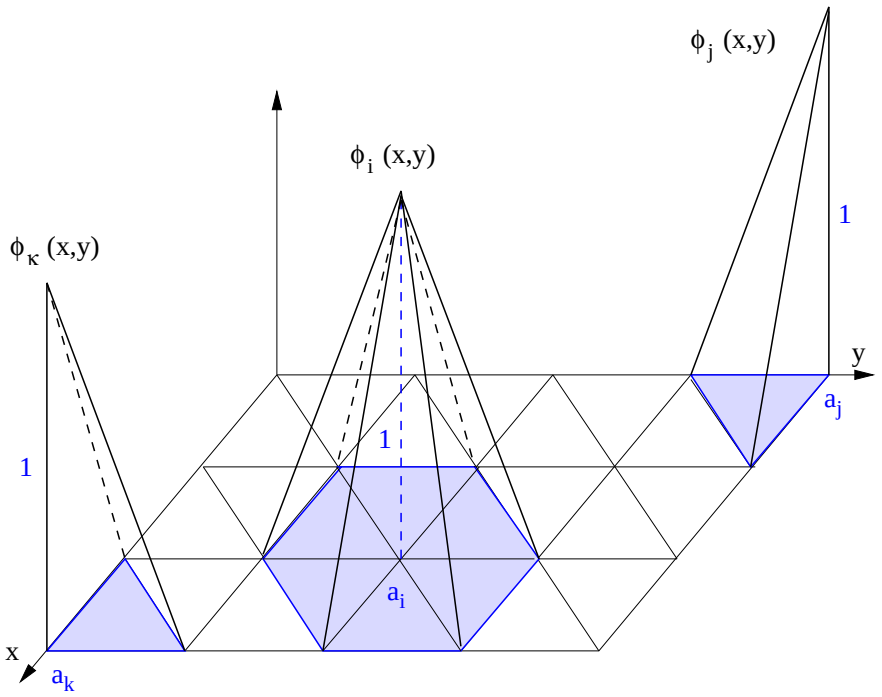

Abbildung 10: Formfunktionen φ_i und deren Träger $\text{supp}(\varphi_i)$ im Falle $n = 2,\ k = 1$

der *Träger (engl.: support)* $\text{supp}(\varphi_i)$ jeder Formfunktion klein ist, d.h. durch wenige Elemente der Triangulierung beschrieben wird. In Abb. 10 wird der Träger linearer Formfunktionen im ebenen Fall, abhängig von der Position des entsprechenden Knotens, dargestellt. Die Zahl der *Freiheitsgrade* auf jedem Element K wird durch die Parameterzahl bestimmt, die eine eindeutige Identifizierung einer Funktion aus $\mathcal{P}_k$ gestatten. Setzt man z.B. Formfunktionen 1. Grades in $\mathbb{R}^n$

$$\varphi_i(x) \;=\; \alpha_i^T x + \beta_i,\ x \in K \qquad \text{mit} \quad \alpha_i \in \mathbb{R}^n,\ \beta_i \in \mathbb{R}$$

an, so sind deren $n+1$ Koeffizienten zu ermitteln. Allgemein ist die Zahl der Freiheitsgrade auf einem Element K durch die Dimension $\dim \mathcal{P}_k = \binom{n+k}{k}$ aus Formel (70) bestimmt.

1. Betrachten wir den ebenen Fall $n = 2$. Um φ_i für $k = 1$ zu berechnen, müssen wir die 3

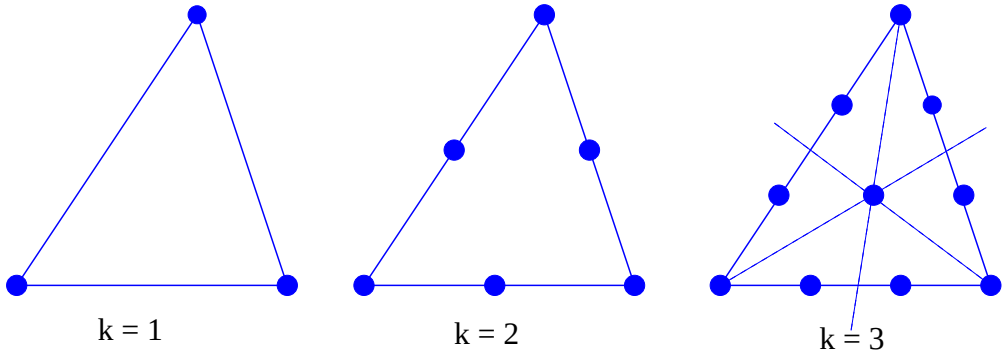

Abbildung 11: Freiheitsgrade im Fall $n = 2$

Freiheitsgrade auf jedem Element K festlegen, wobei wir zusätzlich fordern, dass $\varphi_i \in \mathcal{C}^0(\overline{\Omega})$ ist. Als einfachste Wahl bieten sich die Werte an den Eckpunkten von K an. Falls $k = 2$ ist, so können wir die 6 Freiheitsgrade der einzelnen Elemente durch die Werte an den Eckpunkten und den Seitenmittelpunkten festlegen. Man kann leicht beweisen, dass durch diese 6 Punkte eine Funktion $p \in \mathcal{P}_2$ stets eindeutig bestimmt ist. Analog kann man nachweisen, dass die 10 Freiheitsgrade einer kubischen Formfunktion für $k = 3$ durch die Werte an den folgenden Dreieckspunkten in Abb. 11 festgelegt sind: (i) die Eckpunkte, (ii) zwei Punkte auf jeder Seite, die diese in drei gleichlange Intervalle unterteilt und (iii) dem Schwerpunkt des Dreiecks.

2. Im 3-dimensionalen Fall $n = 3$ kann diese Wahl auf jedes Seitendreieck des Tetraeders

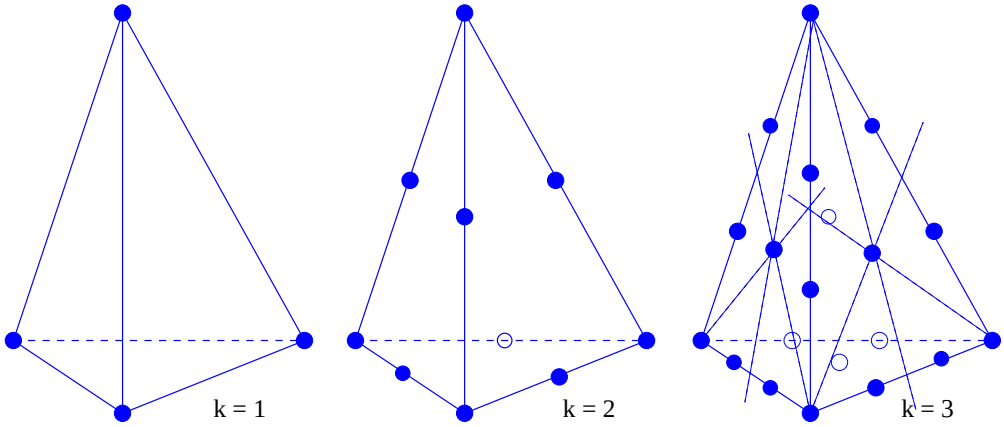

Abbildung 12: Freiheitsgrade im Fall $n = 3$

in Abb. 12 übertragen werden. So erhalten wir für $k = 1$ die 4 Eckpunkte entsprechend den 4 Freiheitsgraden, für $k = 2$ die 4 Eckpunkte und 6 Kantenmittelpunkte, passend zu den 10 Freiheitsgraden und schließlich für $k = 3$ die in Abb. 12 dargestellten 20 Punkte. Das Problem, ein Polynom $p \in \mathcal{P}_k$ in 3 Variablen zu bestimmen, wird damit auf den zweidimensionalen Fall reduziert. Auf jeder Tetraederseite ist der Freiheitsgrad gleich dem des entsprechenden zweidimensionalen Dreiecks. Falls also ein Polynom an den dargestellten

Punkten verschwindet, so auch auf jeder Seite. Wenn aber ein Polynom $p \in \mathcal{P}_k$ mit $k \leq 3$ auf 4 verschiedenen Ebenen verschwindet, so ist $p \equiv 0$.

Im Ergebnis lässt sich somit in den Fällen $k = 1, 2, 3$ eine *nodale Basis* $(\varphi_i(x)),\ i = 1(1)N_h$, konstruieren, mit deren Lagrange-Eigenschaft (75) jede auf $\overline{\Omega}$ definierte Funktion interpoliert werden kann. Denn mit der Basiseigenschaft gilt

Lemma 23 *Bei gegebener zulässiger Triangulierung $\mathcal{T}_h$ mit den Knoten $a_1, a_2, \ldots, a_{N_h}$ kann jede Funktion $u_h \in X_h$ eindeutig durch die Basisfunktionen $(\varphi_i(x)),\ i = 1(1)N_h$, mittels*

$$u_h(x) \;=\; \sum_{i=1}^{N_h} u_h(a_i)\,\varphi_i(x) \tag{76}$$

dargestellt werden. Dabei ist die Dimension N_h von X_h gleich der Zahl der Freiheitsgrade.

Der prinzipielle Weg zur Konstruktion einer Basis des FE-Raumes X_h und damit der Steifigkeitsmatrix A mit $A_{ij} = \mathcal{A}(\varphi_i, \varphi_j)$ und des Lastvektors b mit $b_i = \mathcal{F}(\varphi_i)$ des algebraischen Systems (68) ist damit beschrieben und muss nun konkret umgesetzt werden.

3.4 Das Dirichlet-Problem

Wir wollen einen Lösungsalgorithmus analog zu Abschnitt 2 für das Modell der zweidimensionalen Poisson-Gleichung auf dem Einheitsquadrat $\Omega = (0,1) \times (0,1)$ konstruieren. Dabei beschränken wir uns auf homogene Dirichlet-Bedingungen

$$-\Delta u \;=\; f \quad \text{in} \quad \Omega, \qquad u \;=\; 0 \quad \text{auf} \quad \Gamma = \partial\Omega \tag{77}$$

mit einer gegebenen Funktion $f \in \mathcal{L}^2(\Omega)$. Wie in (46) definieren wir mit dem Skalarprodukt $\langle u, v\rangle$ von $\mathcal{L}^2(\Omega)$

$$\begin{aligned} \mathcal{A}(u,v) \;&=\; \langle \nabla u, \nabla v\rangle \;=\; \int\limits_{\Omega} \nabla u \cdot \nabla v \, d(x,y) \quad \text{und} \\ \mathcal{F}(v) \;&=\; \langle u, v\rangle \;=\; \int\limits_{\Omega} f v \, d(x,y), \end{aligned}$$

womit die Galerkin-Methode aus Definition 19 die endlichdimensionale Variationsgleichung

$$\langle \nabla u_h, \nabla v_h\rangle \;=\; \langle f, v_h\rangle \qquad \forall\, v_h \in V_h \tag{78}$$

für ein $u_h \in V_h$ liefert. Wie bereits in Beispiel 22 besprochen, ist die Wahl

$$V_h \;=\; X_h^k \cap \mathcal{H}_0^1(\Omega) \;=\; \{v_h \in X_h^k \mid v_h = 0 \text{ auf } \partial\Omega\}, \quad k \geq 1$$

des Ansatzraumes sinnvoll, d.h. wir fordern das Verschwinden der Funktionen v_h auf dem Rand. Damit müssen nur innere Knoten $a_i \in \Omega$ bei der Konstruktion der FE-Basis berücksichtigt werden. Mit dem Ansatz (76) entsteht schließlich das Gleichungssystem der Dimension N_h

$$A\,\xi \;=\; b \qquad \text{mit} \quad A_{ij} \;=\; \langle \nabla\varphi_i, \nabla\varphi_j\rangle \quad \text{und} \quad b_i \;=\; \langle f, \varphi_i\rangle \tag{79}$$

für die gesuchten Knotenwerte $\xi_i = u_h(a_i)$. Betrachten wir die einzelnen Lösungsschritte:

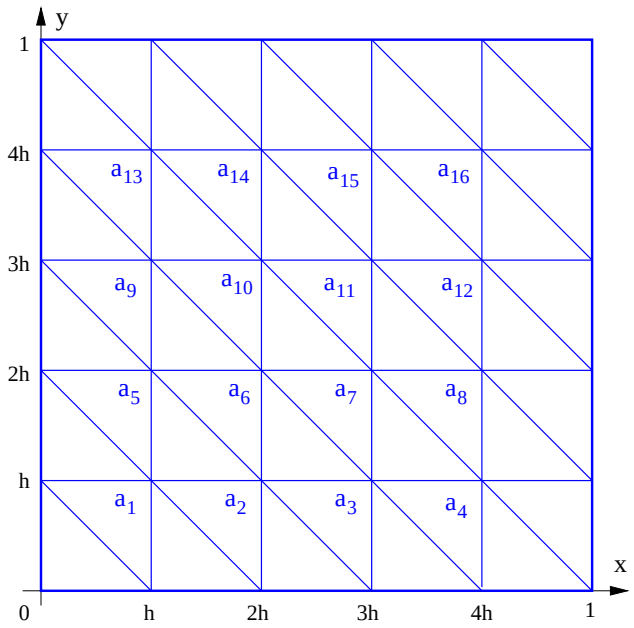

Abbildung 13: Standard-Triangulierung $\mathcal{T}_h$ des Einheitsquadrates $(m = 4)$

1. Triangulierung. Das Quadrat Ω wird mit der so genannten *Standard-Triangulierung* wie in Abb. 13 zerlegt. Dazu wählen wir ein $m \in \mathbb{N}, m \geq 1$, legen ein Gitter der Maschenweite $h = 1/(m+1)$ über Ω und zeichnen die in der Abbildung für $m = 4$ dargestellten Diagonalen ein. Offenbar ist diese Triangulierung $\mathcal{T}_h$ zulässig mit Durchmesser $\max_{K\in\mathcal{T}_h} \operatorname{diam}(K) = h\sqrt{2}$. Wir nummerieren die $N_h = m^2$ inneren Knoten $a_1, a_2, \ldots, a_{N_h}$ zeilenweise in der dargestellten Anordnung.

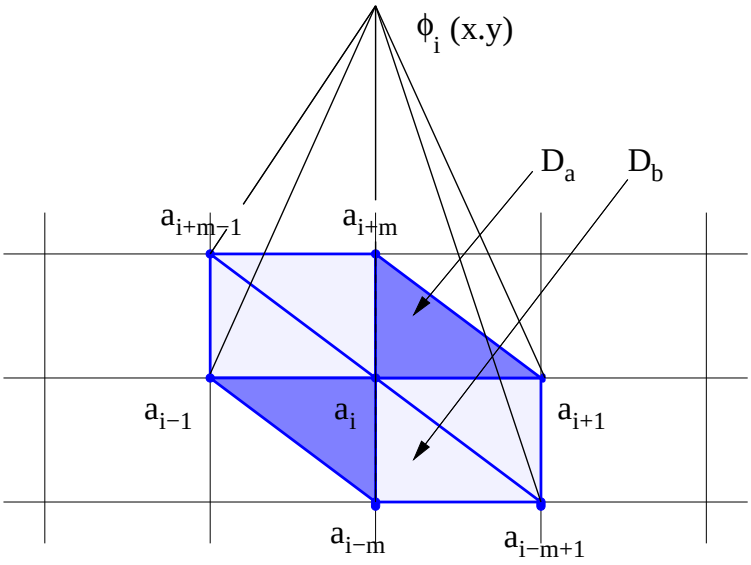

Abbildung 14: Basisfunktion φ_i und ihr Träger $\operatorname{supp}(\varphi_i)$ mit Nachbarknoten

2. Hutfunktionen. Die N_h Formfunktionen $\varphi_1, \varphi_2, \ldots, \varphi_{N_h}$ haben die in Abb. 14 dargestellte „Hutform“ und einen kleinen Träger, d.h. außerhalb des durch die Nachbarknoten beschriebenen Sechsecks verschwinden sie. Daraus folgt, dass 2 Basisfunktionen φ_i und φ_j nur dann

einen von Null verschiedenen Wert des Skalarproduktes $\langle \nabla\varphi_i, \nabla\varphi_j \rangle$ liefern, wenn sich ihre Träger überlappen. Offenbar trifft dies nur auf benachbarte Knoten $\varphi_i \neq \varphi_j$ zu, deren Träger sich in genau 2 Dreiecken überschneiden.[14] Zur Berechnung von $\langle \nabla\varphi_i, \nabla\varphi_i \rangle$ müssen dagegen alle 6 Dreiecke des Trägers ausgewertet werden.

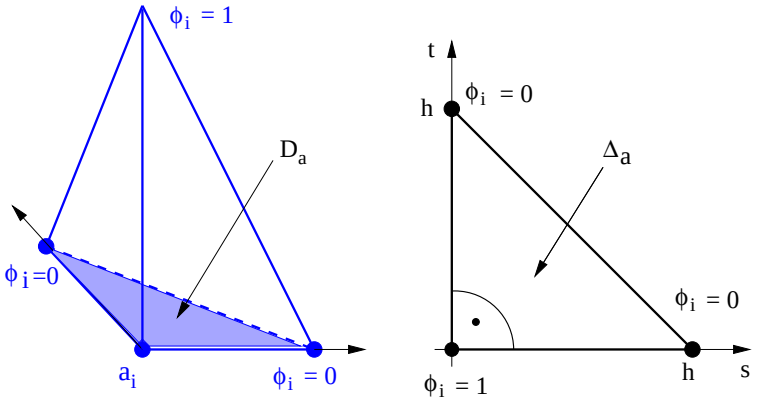

Abbildung 15: Basisfunktion φ_i im Fall D_a mit Referenzdreieck Δ_a

3. Referenzdreiecke und Steifigkeitsmatrix. Wir bestimmen zuerst die Matrixelemente

$$A_{ii} \;=\; \langle \nabla\varphi_i, \nabla\varphi_i \rangle \;=\; \int\limits_{\Omega} \|\nabla\varphi_i(x,y)\|^2 \, d(x,y) \;=\; \int\limits_{\mathrm{supp}(\varphi_i)} \|\nabla\varphi_i(x,y)\|^2 \, d(x,y)$$

über den 6 Dreiecken in Abb. 14. Offenbar treten wegen des nichtnegativen Integranden nur die beiden mit D_a und D_b gekennzeichneten Fälle auf.

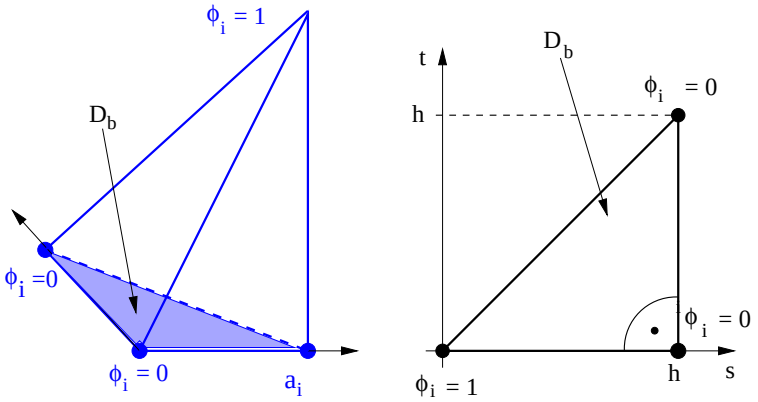

Abbildung 16: Basisfunktion φ_i im Fall D_b mit Referenzdreieck Δ_b

(D_a) Wir führen eine lineare Transformation des Dreiecks D_a auf das in Abb. 15 gezeichnete *Referenzdreieck* Δ_a mit den Ecken $(s,t) = (0,0), (h,0), (0,h)$ aus. Wegen $\varphi_i(0,0) = 1$ und $\varphi_i(h,0) = \varphi_i(0,h) = 0$ erhält man

- die transformierte Formfunktion $\varphi_i(s,t) = 1 - \frac{s}{h} - \frac{t}{h}$,
- den Gradienten $\nabla\varphi_i \;=\; (-\frac{1}{h}, -\frac{1}{h})$ und damit

[14] Die Basis $\varphi_1, \varphi_2, \ldots, \varphi_{N_h}$ ist also „fast“ orthogonal.

- das Integral $\displaystyle\int_{D_a} \|\nabla\varphi_i(x,y)\|^2\, d(x,y) \;=\; 2\int_0^h \int_0^{h-s} \frac{1}{h^2}\, dt\, ds \;=\; 1.$

(D_b) In Abb. 16 ist das zugehörige *Referenzdreieck* Δ_b mit den Eckwerten $\varphi_i(0,0) = 1$ und $\varphi_i(h,0) = \varphi_i(h,h) = 0$ dargestellt. Wir berechnen

- die transformierte Formfunktion $\varphi_i(s,t) = 1 - \frac{s}{h}$,
- den Gradienten $\nabla\varphi_i \;=\; (-\frac{1}{h}, 0)$ und damit
- das Integral $\displaystyle\int_{D_a} \|\nabla\varphi_i(x,y)\|^2\, d(x,y) \;=\; \int_0^h \int_0^{s} \frac{1}{h^2}\, dt\, ds \;=\; \frac{1}{2}.$

Summation über die 6 Dreiecke des Trägers ergibt den Gesamtwert $A_{ii} \;=\; \langle\nabla\varphi_i, \nabla\varphi_i\rangle \;=\; 4$. Nun ermitteln wir die Matrixelemente $A_{ij} \;=\; \langle\nabla\varphi_i, \nabla\varphi_j\rangle$ mit $i \neq j$. Nach Abb. 14 treten 2 Fälle für das Skalarprodukt auf:

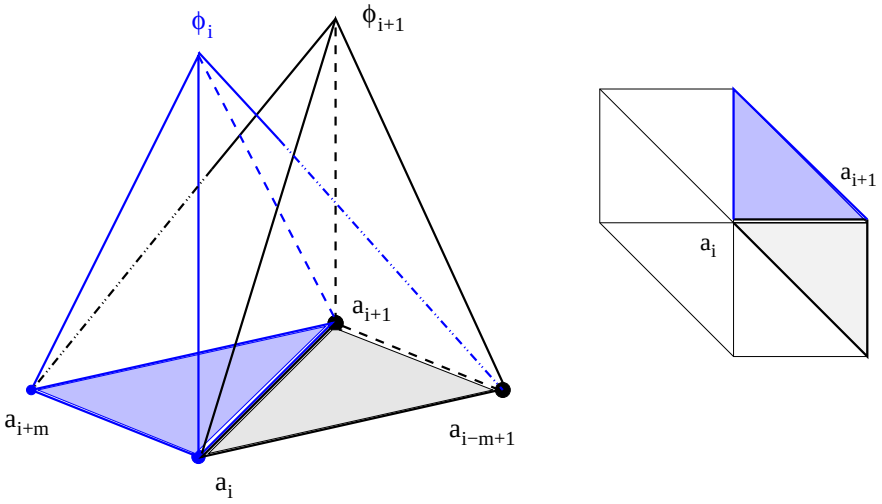

Abbildung 17: Basisfunktionen φ_i und φ_{i+1} im Fall S_a

(S_a) Wir berechnen $\langle\nabla\varphi_i, \nabla\varphi_{i+1}\rangle$, wie in Abb. 17 dargestellt, wegen

$$\varphi_i(s,t) = 1 - \frac{s}{h} - \frac{t}{h} \qquad \text{und} \qquad \varphi_{i+1}(s,t) = 1 - \frac{s}{h}$$

auf jedem Dreieck zu $-1/2$, womit sich insgesamt $\langle\nabla\varphi_i, \nabla\varphi_{i+1}\rangle = -1$ ergibt. Derselbe Wert wird auch für die folgenden 3 Skalarprodukte geliefert:

$$\langle\nabla\varphi_i, \nabla\varphi_{i-1}\rangle \;=\; \langle\nabla\varphi_i, \nabla\varphi_{i+m}\rangle \;=\; \langle\nabla\varphi_i, \nabla\varphi_{i-m}\rangle \;=\; -1. \tag{80}$$

(S_b) Für die verbleibenden 2 Skalarprodukte aus Abb. 14 berechnet man analog

$$\langle\nabla\varphi_i, \nabla\varphi_{i+m-1}\rangle \;=\; \langle\nabla\varphi_i, \nabla\varphi_{i-m+1}\rangle \;=\; 0. \tag{81}$$

Da alle weiteren Skalarprodukte verschwinden, entsteht eine *schwach besetzte (engl.: sparse) Steifigkeitsmatrix* mit den dargestellten $m \times m$-Tridiagonalblöcken:

$$A = \begin{pmatrix} \begin{matrix} 4 & -1 & \cdots & 0 \\ -1 & 4 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 4 \end{matrix} & \begin{matrix} -1 & 0 & \cdots & 0 \\ 0 & -1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & -1 \end{matrix} & \cdots & O \\ \begin{matrix} -1 & 0 & \cdots & 0 \\ 0 & -1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & -1 \end{matrix} & \begin{matrix} 4 & -1 & \cdots & 0 \\ -1 & 4 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 4 \end{matrix} & \cdots & O \\ \cdots & \cdots & \cdots & \cdots \\ O & O & \cdots & \begin{matrix} 4 & -1 & \cdots & 0 \\ -1 & 4 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 4 \end{matrix} \end{pmatrix}$$

Die knotenweise Aufstellung dieser Steifigkeitsmatrix wird als *knotenorientierte Assemblierung* bezeichnet. Wir stellen fest, dass diese Matrix bis auf den fehlenden Vorfaktor $1/h^2$ mit der blocktridiagonalen Koeffizientenmatrix A der Finite-Differenzen-Methode (17) des Abschnittes 2 übereinstimmt. Andere Nummerierungen der Knoten bzw. andere Triangulie-

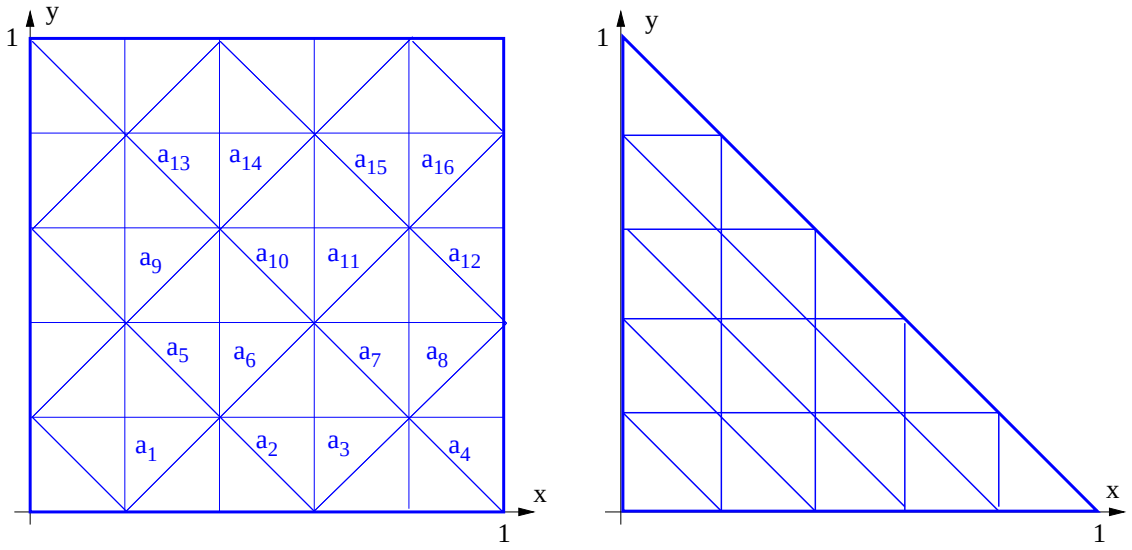

Abbildung 18: (a) „Union-Jack“-Triangulierung und (b) Standard-Dreiecks-Triangulierung

rungen, wie die „Union-Jack“-Triangulierung in Abb. 18a, führen jedoch auf davon abweichende Steifigkeitsmatrizen.

3. Lastvektor. Zur Berechnung von b_i müssen wir nur über die Elemente K des Trägers

$\text{supp}(\varphi_i)$ integrieren:

$$b_i \;=\; \langle f, \varphi_i \rangle \;=\; \int\limits_{\text{supp}(\varphi_i)} f(x,y)\,\varphi_i(x,y)\,d(x,y) \;=\; \sum_{K \in \text{supp}(\varphi_i)} \int\limits_K f(x,y)\,\varphi_i(x,y)\,d(x,y).$$

Hier bieten sich die Kubaturformeln über Dreiecksbereichen an. So liefert beispielsweise die *Prismenregel* über dem Dreieck $K \in \text{supp}(\varphi_i)$ aus Abb. 14 mit den Knoten a_i, a_{i+1}, a_{i+m} und dem Inhalt $|K|$ den Näherungswert

$$P_K \;=\; \frac{|K|}{3}[f(x_i,y_i)\cdot 1 + f(x_{i+1},y_{i+1})\cdot 0 + f(x_{i+m},y_{i+m})\cdot 0] \;=\; \frac{|K|}{3} f(x_i,y_i),$$

womit sich die rechten Seiten zu

$$b_i \;=\; \left(\frac{1}{3} \sum_{K \in \text{supp}(\varphi_i)} |K| \right) f(x_i,y_i), \quad i = 1(1)N_h \tag{82}$$

ergeben. Für die Standard-Triangulierung $\mathcal{T}_h$ des Einheitsquadrates gemäß Abb. 13 ist stets $|K| = h^2/2$, weshalb Formel (82) die Werte

$$b_i \;=\; h^2\, f(x_i,y_i), \quad i = 1(1)N_h$$

liefert. Hier ist also der fehlende Faktor $1/h^2$ geblieben! Wir stellen fest, dass unser FEM-Ansatz mit Standard-Triangulierung und linearen Elementen ($k = 1$) auf dasselbe endliche Gleichungssystem geführt hat wie die Finite-Differenzen-Methode mit der 5-Punkte-Formel (13). Für allgemeinere Bereiche Ω erweist sich dagegen die Finite-Elemente-Methode überlegen. Damit lässt sich das Dirichlet-Problem auf rechtwinkligen Dreiecksbereichen mit der in Abb. 18b dargestellten Standard-Triangulierung leicht lösen.

3.5 Implementierung der FEM

Um die Steifigkeitsmatrix A auch bei allgemeineren Triangulierungen $\mathcal{T}_h$ systematisch zu berechnen, betrachten wir für jedes einzelne Dreieck K die „elementeweisen“ Integrale

$$A_{ij}^K \;=\; \int\limits_K \nabla\varphi_i(x,y) \cdot \nabla\varphi_j(x,y)\,d(x,y), \qquad i,j = 1(1)N_h. \tag{83}$$

Sind a_i und a_j je zwei Knoten von K, so können wir die zugehörigen Werte zu einer Matrix $A^K = (A_{ij}^K)$ zusammenfassen. Wenn wir die 3 Knoten von K lokal mit $i = 1,2,3$ nummerieren, so hat A^K mit unserer nodalen Basis höchstens 3×3 nicht verschwindende Elemente. A^K heißt deshalb *Element-Steifigkeitsmatrix* von K.

Beispiel 24 Für das Referenzdreieck Δ_a aus Abb. 15 mit den Knoten $a_1 = (0,0)$, $a_2 = (h,0)$, $a_3 = (0,h)$ errechnen wir die Hutfunktionen und Gradienten auf Δ_a

$$\begin{aligned}
\varphi_1(s,t) &= -s/h - t/h + 1, & \nabla\varphi_1 &= (-1/h, -1/h)\\
\varphi_2(s,t) &= s/h, & \nabla\varphi_2 &= (1/h, 0)\\
\varphi_3(s,t) &= t/h, & \nabla\varphi_3 &= (0, 1/h),
\end{aligned}$$

womit die Koeffizienten der Element-Steifigkeitsmatrix durch

$$A_{ij}^{\Delta_a} \;=\; \int_0^h \int_0^{h-s} \nabla\varphi_i(s,t)\cdot\nabla\varphi_j(s,t)\,dt\,ds \;=\; \frac{h^2}{2}\,\nabla\varphi_i\cdot\nabla\varphi_j$$

berechnet werden können. Analog bestimmen wir für das Referenzdreieck Δ_b aus Abb. 16 mit den Knoten $a_1=(0,0),\ a_2=(h,0),\ a_3=(h,h)$ die Einträge

$$A_{ij}^{\Delta_b} \;=\; \int_0^h \int_0^{s} \nabla\varphi_i(s,t)\cdot\nabla\varphi_j(s,t)\,dt\,ds \;=\; \frac{h^2}{2}\,\nabla\varphi_i\cdot\nabla\varphi_j$$

und erhalten so die beiden Element-Steifigkeitsmatrizen

$$A^{\Delta_a} \;=\; \begin{pmatrix} 1 & -1/2 & -1/2 \\ -1/2 & 1/2 & 0 \\ -1/2 & 0 & 1/2 \end{pmatrix} \qquad \text{und} \quad A^{\Delta_b} \;=\; \begin{pmatrix} 1/2 & -1/2 & 0 \\ -1/2 & 1 & -1/2 \\ 0 & -1/2 & 1/2 \end{pmatrix}. \qquad \square$$

Mit passender Umnummerierung der Knoten lassen sich aus diesen Matrizen die Element-Steifigkeitsmatrizen A^K für alle Elemente K der Triangulierung gewinnen. Die *globale Steifigkeitsmatrix* A erhalten wir dann formal durch Summation über alle Elemente

$$A \;=\; \sum_{K\in\mathcal{T}_h} A^K. \tag{84}$$

Dieser Prozess der *elementorientierten Assemblierung* der Steifigkeitsmatrix lässt sich gut algorithmieren, wenn passende Datenstrukturen wie verkettete Listen genutzt werden, in denen nur die von Null verschiedenen Einträge gespeichert werden [6]. Die kompakte Form

Algorithmus 25 (FEM für Poisson-Gleichung)

Function $[a,u,N_h] =$ PoissonFEM (f,Ω,h)

1. Bestimme eine zulässige Triangulierung $\mathcal{T}_h$ der Feinheit h von $\overline{\Omega}$ mit den N_h Knoten a_i und Elementen K.
2. Ermittle die Element-Steifigkeitsmatrizen A^K für $K\in\mathcal{T}_h$.
3. Assembliere die globale Steifigkeitsmatrix A gemäß (84).
4. Assembliere den Lastvektor b gemäß (82).
5. Löse das System $A\xi = b$ direkt oder iterativ nach den Knotenwerten $\xi_i = u(a_i),\ i=1(1)N_h$.
6. Return $a=(a_1,\ldots,a_{N_h}),\ u=(u(a_1),\ldots,u(a_{N_h}))$ und N_h

des Algorithmus 25 zur Lösung der Poisson-Gleichung (77) mit homogenen Dirichletschen Randbedingungen auf einer polygonalen Menge $\Omega\subset\mathbb{R}^2$ benötigt außer der Funktion f und den Ecken von Ω nur die gewünschte Feinheit h der Triangulierung. Im Allgemeinen muss die

Geometrie komplizierterer Gebiete Ω und der Randbedingungen jedoch detailliert beschrieben werden. Algorithmus 25 liefert die N_h Knoten a_i und die Näherungslösung $u(a_i)$. Mit der Basisdarstellung (76) kann daraus die Lösung für jeden Punkt $x \in \Omega$ leicht interpoliert werden.

MATLAB verfügt über eine Toolbox `pdetool`[14] zur Lösung von 2D-Problemen der Form

$$-\nabla \cdot (a(x,y)\nabla u) + b(x,y)u \;=\; f(x,y) \qquad \text{mit} \quad a, b, f : \Omega \to \mathbb{C} \tag{85}$$

über einem beschränkten Gebiet $\Omega \subset \mathbb{R}^2$. a kann auch eine auf Ω definierte (reell- oder komplexwertige) 2×2-Matrixfunktion sein. Damit ist die Poisson-Gleichung (77) mit $a = 1$

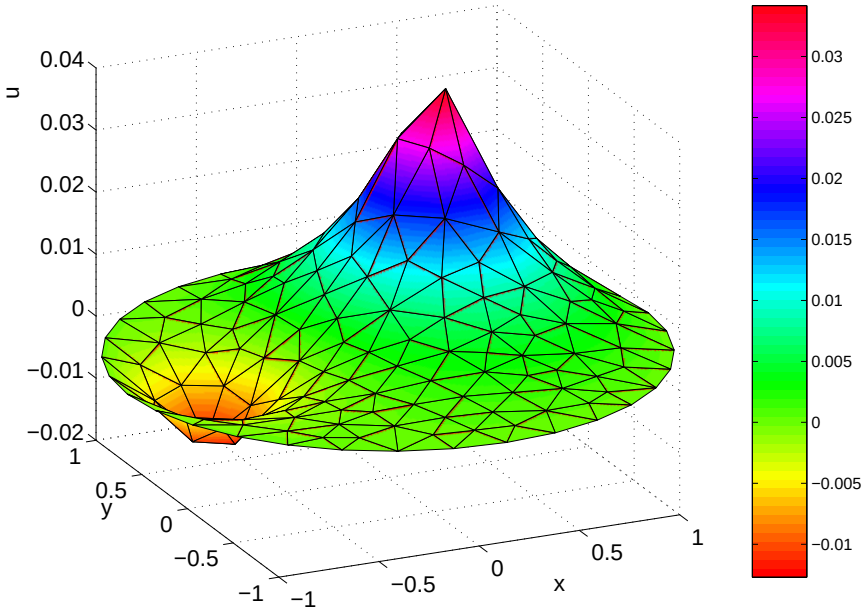

Abbildung 19: Lösung von Beispiel 26 (1) auf grobem Gitter

und $b = 0$ als Spezialfall enthalten. Als Randbedingungen sind

- Dirichlet-Bedingungen $\qquad h(x,y)u \;=\; r(x,y) \quad$ und
- verallgemeinerte Neumann-Bedingungen [15] $\qquad \mathbf{n} \cdot (c(x,y)\nabla u) + h(x,y)u \;=\; r(x,y)$

zugelassen. Darin ist $\mathbf{n}$ der nach außen gerichtete Einheits-Normalenvektor, während h und r auf $\partial\Omega$ definierte Funktionen sind. Als Lösungsverfahren ist die hier beschriebene FEM mit linearen ($k = 1$) Polynomen implementiert.

Beispiel 26 Wir lösen das Dirichlet-Problem mit homogenen Randbedingungen auf Gebieten mit komplizierterer Geometrie

$$-\Delta u \;=\; f(x,y), \quad (x,y) \in \Omega \qquad \text{und} \qquad u(x,y) \;=\; 0, \quad (x,y) \in \partial\Omega. \tag{86}$$

1. Das kreisförmige Gebiet um $(0,0)$ approximieren wir polygonal, wie in Abb. 20 dargestellt, mittels der implementierten Delaunay-Triangulierung [14]. Der Quellterm f werde ähnlich

[15] Diese Form wird auch oft als *Robin-Bedingungen* bezeichnet.

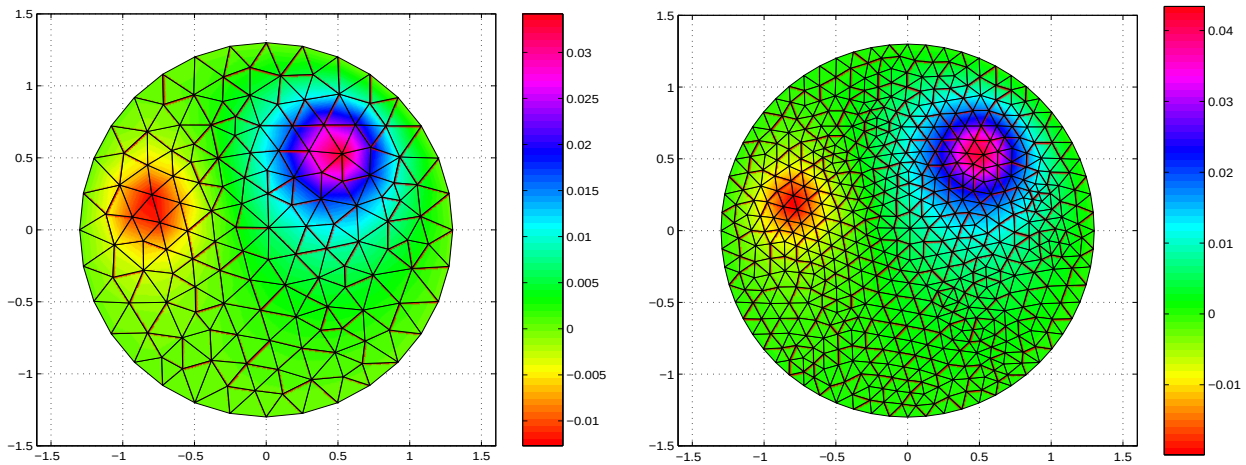

Abbildung 20: Ausgangs-Triangulierung (a) mit 155 Knoten und verfeinerte Triangulierung (b) mit 585 Knoten zu Bsp. 26 (1)

wie im FDM-Beispiel 10(1) durch

$$f(x,y) \;=\; 3.2\,e^{-80((x-0.5)^2+(y-0.55)^2)} - 2.6\,e^{-90(2(x+0.8)^2+(y-0.2)^2)}$$

definiert. Die 155-Knoten-Lösung in Abb. 19 kann durch ein gleichmäßig verfeinertes Gitter

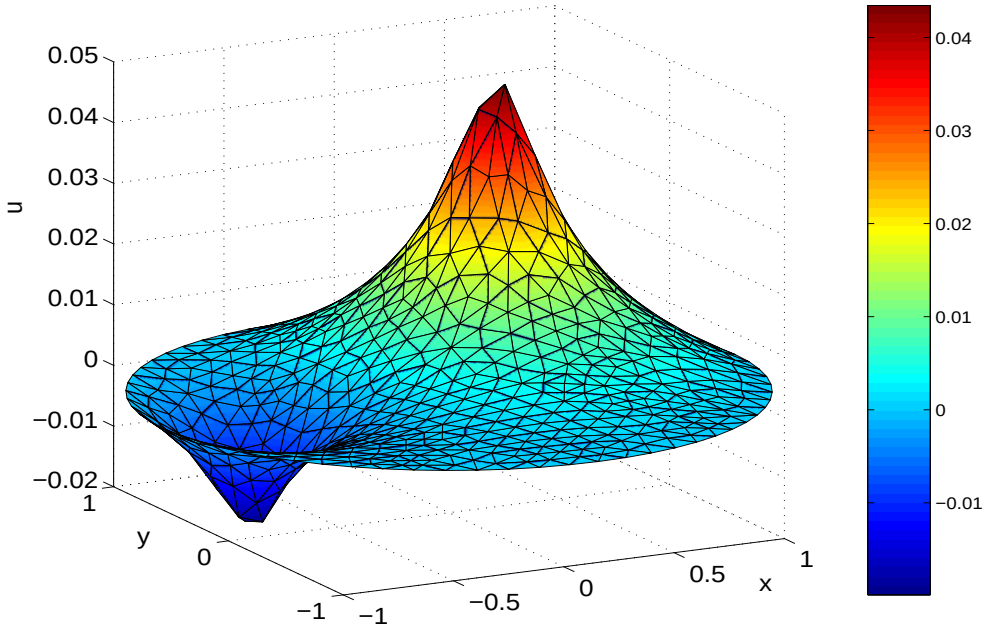

Abbildung 21: Lösung von Beispiel 26 (1) auf verfeinertem Gitter

mit 585 Knoten in Abb. 20 weiter verbessert werden. Bis auf die „Peaks“ wird in Abb. 21 die Lösung gut approximiert. Eine weitere Verfeinerung der Triangulierung sollte deshalb lokal erfolgen.

2. Nun lösen wir das Dirichlet-Problem auf dem in Abb. 22 dargestellten polygonalen Gebiet mit Ausgangs-Triangulierung nach Delaunay [14]. Die Ladungsdichte stimmt mit derjenigen

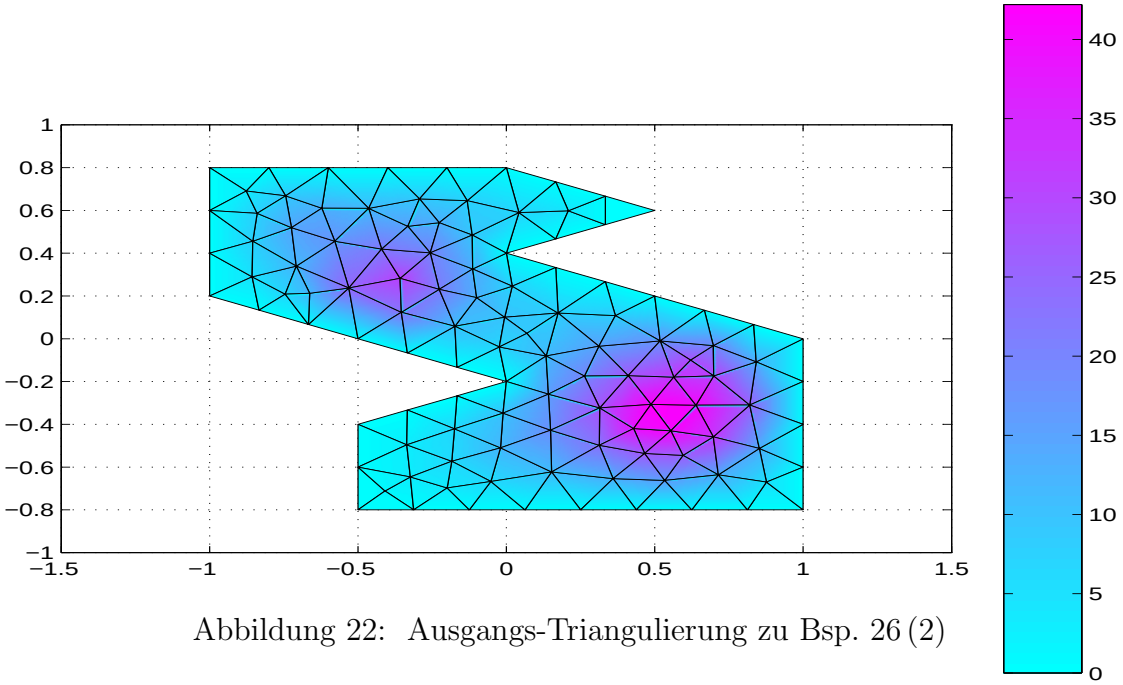

Abbildung 22: Ausgangs-Triangulierung zu Bsp. 26 (2)

des Beispiels 10 (2) zur FDM

$$f(x,y) \;=\; \frac{100}{10\cdot\sin^2(24\cdot((x-0.5)^2+(y-0.5)^2))+0.01}$$

überein und ist in Abb. 6 gezeichnet. Die in Abb. 23 dargestellte Näherungslösung stellt eine

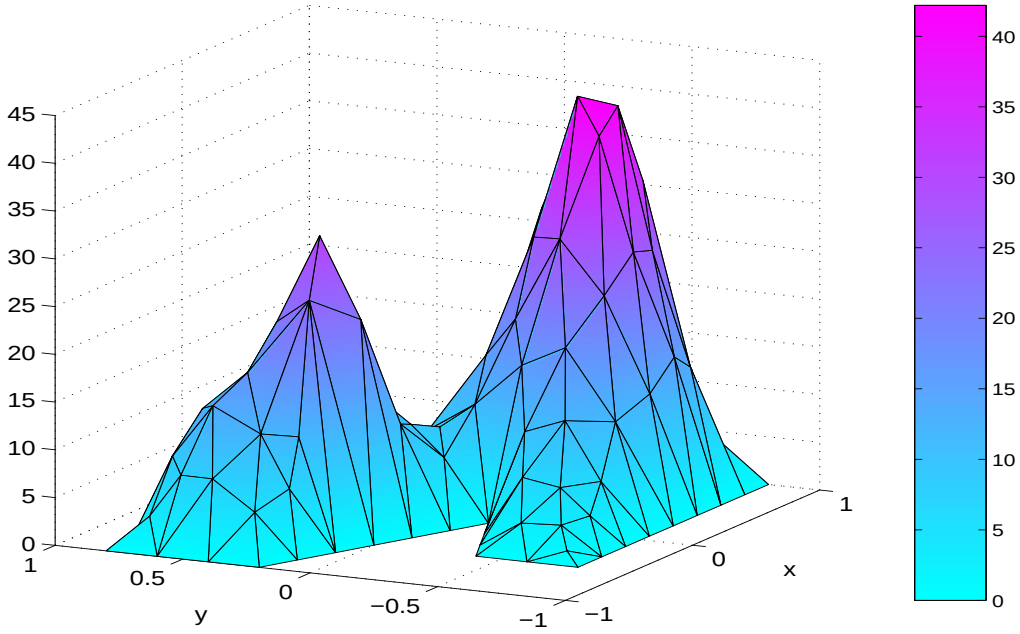

Abbildung 23: Lösung von Bsp. 26 (2) auf grobem Gitter

sehr grobe Approximation dar und ist stark verbesserungsbedürftig. Nach einer zweimaligen gleichmäßigen Gitterverfeinerung erhalten wir auf dem Gitter der Abb. 24 mit `pdetool` die in Abb. 25 dargestellte Lösung. □

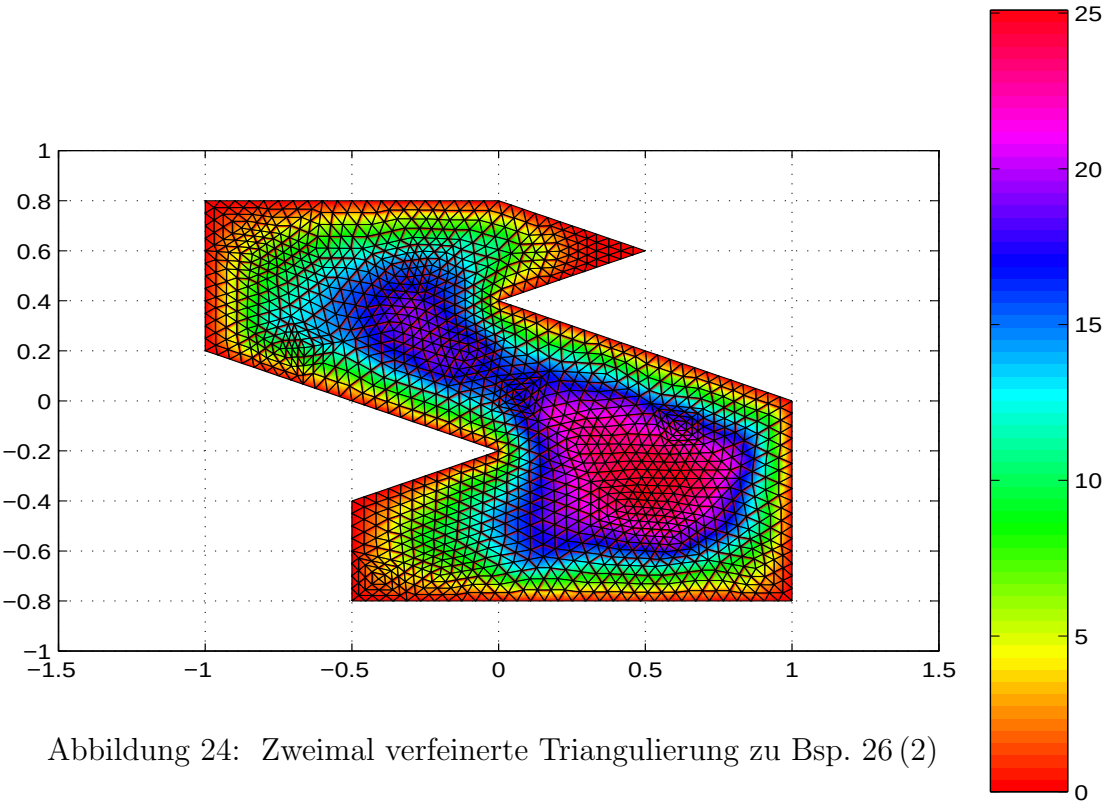

Abbildung 24: Zweimal verfeinerte Triangulierung zu Bsp. 26 (2)

Die MATLAB-Toolbox `pdetool` gestattet außer der FEM-Lösung linearer elliptischer Randwertprobleme[16] der Form (85) auch die Behandlung

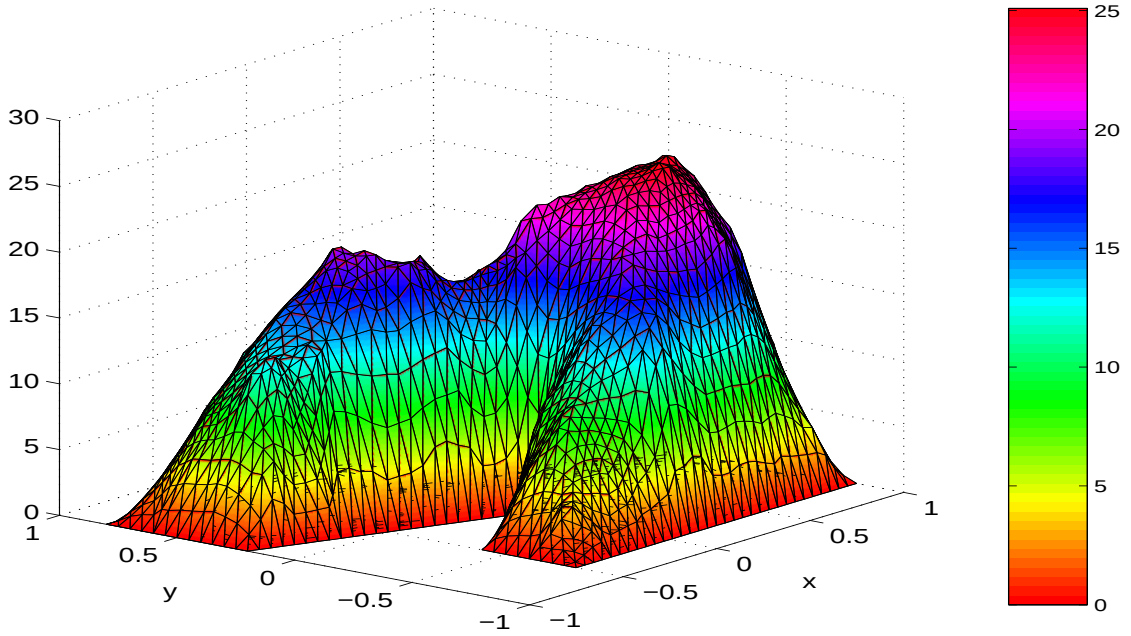

Abbildung 25: Lösung von Beispiel 26 (2) auf zweimal verfeinertem Gitter

[16] Die Termini „elliptisch“, „parabolisch“ und „hyperbolisch“ werden hier nur zur Grobklassifikation genutzt. Im konkreten Fall ist stets zu verifizieren, ob einer der 3 Fälle im Sinne der Definition vorliegt.

- linearer Eigenwertprobleme für $\lambda \in \mathbb{C}$

$$-\nabla \cdot (a(x,y)\nabla u) + b(x,y)u \;=\; \lambda d(x,y)u \qquad \text{mit} \quad a,b,d : \Omega \to \mathbb{C}, \tag{87}$$

- nichtlinearer elliptischer PDGLn

$$-\nabla \cdot (a(x,y,u)\nabla u) + b(x,y,u)u \;=\; f(x,y,u), \quad a,b,f : \Omega \times \mathbb{C} \to \mathbb{C}, \tag{88}$$

- instationärer PDGLn vom parabolischen Typ

$$d(x,y)\frac{\partial u}{\partial t} - \nabla \cdot (a(x,y)\nabla u) + b(x,y)u \;=\; f(x,y), \quad a,b,d,f : \Omega \to \mathbb{C}, \tag{89}$$

- instationärer PDGLn vom hyperbolischen Typ

$$d(x,y)\frac{\partial^2 u}{\partial t^2} - \nabla \cdot (a(x,y)\nabla u) + b(x,y)u \;=\; f(x,y), \quad a,b,d,f : \Omega \to \mathbb{C} \tag{90}$$

- sowie von Systemen partieller DGLn der Form (85), (87), (89) und (90).

Um beispielsweise hyperbolische Systeme mit n Gleichungen für die n gesuchten Funktionen $u = (u_1, u_2, \ldots, u_n)^T$ der Form

$$d(x,y)\frac{\partial^2 u}{\partial t^2} - \nabla \cdot (a(x,y) \otimes \nabla u) + b(x,y)u \;=\; f(x,y) \tag{91}$$

zu lösen, müssen die Funktionen f und u Spaltenvektoren der Länge n sein, während b und d nun $n \times n$-Matrizen sind. a ist ein $n \times n \times 2 \times 2$-Tensor, mit dem das Produkt $\nabla \cdot (a(x,y) \otimes \nabla u)$ einen n-Vektor mit der i-ten Komponente

$$\sum_{j=1}^{n} \left(\frac{\partial}{\partial x} a_{ij11} \frac{\partial}{\partial x} + \frac{\partial}{\partial x} a_{ij12} \frac{\partial}{\partial y} + \frac{\partial}{\partial y} a_{ij21} \frac{\partial}{\partial x} + \frac{\partial}{\partial y} a_{ij22} \frac{\partial}{\partial y} \right) u_j$$

liefert.

3.6 Konvergenz der FEM

Während wir bisher zu gegebenem Parameterwert h mit Algorithmus 25 eine einzelne Näherungslösung u_h berechnet haben, wollen wir nun eine Familie von Galerkin-Approximationen

$$\text{Finde } \; u_h \in V_h \; : \quad \mathcal{A}(u_h, v_h) \;=\; \mathcal{F}(v_h) \qquad \forall\, v_h \in V_h \tag{92}$$

für $h \to 0$ betrachten und das Konvergenzverhalten der Näherungen u_h untersuchen. Dazu konstruieren wir eine Familie von Triangulierungen $\mathcal{T}_h, h > 0$ mit folgender Eigenschaft:

Definition 27 (Reguläre Triangulierung) *Eine Familie zulässiger Triangulierungen $\mathcal{T}_h$, $h > 0$, von $\overline{\Omega}$ heißt regulär, wenn eine von h unabhängige Konstante $\sigma \geq 1$ existiert, mit der*

$$\max_{K \in \mathcal{T}_h} \frac{h_K}{\rho_K} \;\leq\; \sigma \qquad \forall\, h > 0 \tag{93}$$

gilt. Darin sind für jedes Element K die beiden Durchmesser $h_K := diam(K)$ und $\rho_K := sup\{diam(S) \mid S \text{ ist eine Kugel mit } S \subset K\}$ definiert.

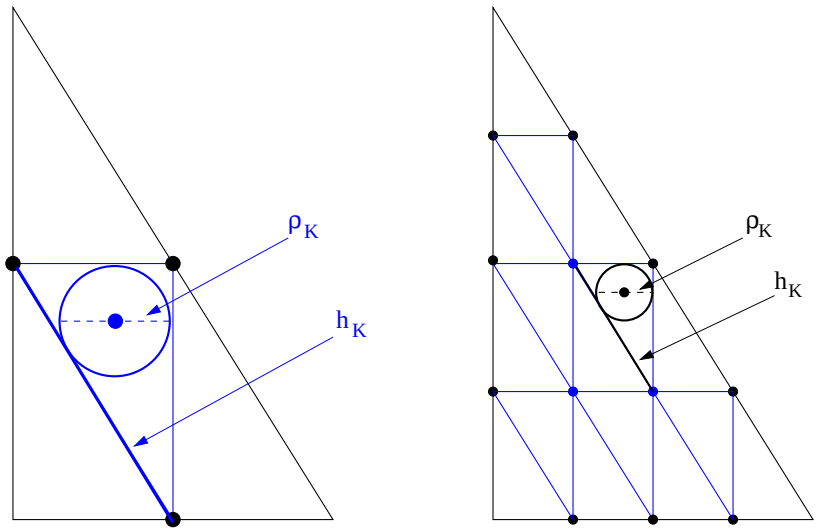

Abbildung 26: Anfangs-Triangulierung (a) und verfeinerte Triangulierung (b)

Für dreieckige finite Elemente ist offenbar h_K die längste Seite (bzw. Kante), während ρ_K den Durchmesser des Inkreises (bzw. der Inkugel) liefert. (93) fordert die gleichmäßige Beschränktheit des Verhältnisses dieser Größen, um Entartungen der Elemente K zu vermeiden. Verfeinert man eine Anfangs-Triangulierung wie in Abb. 26, indem jede Dreiecksseite halbiert und somit jedes Element in 4 kongruente Teildreiecke zerlegt wird, so bleibt das Verhältnis h_K/ρ_K konstant[17]. Sukzessive Wiederholung dieser *gleichmäßigen Gitterverfeinerung* liefert dann eine reguläre Triangulierung.

Den FE-Raum V_h wählen wir wie in Bsp. 22 dargestellt, also für das Dirichlet-Problem (44)

$$V_h \;=\; X_h^k \cap \mathcal{H}_0^1(\Omega) \qquad \text{mit} \quad X_h^k \;=\; \{v_h \in \mathcal{C}^0(\overline{\Omega}) \mid v_{h|K} \in \mathcal{P}_k \;\; \forall K \in \mathcal{T}_h\},\; k \geq 1,$$

während für das Neumann-Problem (48) der FE-Raum

$$V_h \;=\; X_h^k, \quad k \geq 1$$

und für gemischte Probleme (52)

$$V_h \;=\; X_h^k \cap \mathcal{H}_{\Gamma_D}^1(\Omega) \;=\; \{v_h \in X_h^k \mid v_h = 0 \text{ auf } \Gamma_D\}, \quad k \geq 1$$

geeignet ist. Weiterhin nehmen wir an, dass dreieckige Elemente mit den beschriebenen Freiheitsgraden und Formfunktionen gewählt wurden. Die Konvergenz der FEM wird in [16, Theorem 6.2.1] nachgewiesen, wozu die Sätze 17 und 20 benutzt werden und die Approximationseigenschaft (59)

$$\inf_{v_h \in V_h} \|v - v_h\| \to 0, \quad \text{falls } h \to 0 \qquad \text{für alle } v \in V$$

verifiziert wird. Für unsere Darstellung notieren wir (vgl. [16, S. 172])

Satz 28 (Konvergenz der FEM) *Der Bereich Ω sei eine polygonale Menge in $\mathbb{R}^n$, $n = 2, 3$ und $\mathcal{T}_h$ sei eine reguläre Familie von Triangulierungen mit dreieckigen Elementen. Angenommen, die Bilinearform $\mathcal{A}(u, v)$ ist stetig und koerziv auf V und das lineare*

[17] Die MATLAB-Toolbox `pdetool` nutzt diese Vorgehensweise.

Funktional $\mathcal{F}(v)$ ist stetig auf V. Der FE-Raum V_h sei durch (72), (73) oder (74) definiert. Unter diesen Annahmen konvergiert die Finite-Elemente-Galerkin-Methode (60). Wenn die exakte Lösung $u \in \mathcal{H}^s(\Omega)$ mit einem $s \geq 2$ genügt, so ist die Fehlerschätzung

$$\|u - u_h\|_1 \leq Ch^l \|u\|_{l+1} \qquad mit \quad l = min(k, s-1) \tag{94}$$

und einer Konstanten $C > 0$ erfüllt.

Bei hinreichender Glattheit der Lösung ($s > k$) liefert der Grad k der Formfunktionen die Konvergenzordnung der FEM. Mit den eingeführten Formfunktionen 1. Grades hat der Fehler $u - u_h$ in der $\mathcal{H}^1(\Omega)$-Norm die Ordnung $\mathcal{O}(h)$. Ist die Glattheit von u gering, so macht ein hoher Grad k der Ansatzfunktionen keinen Sinn, weshalb l auch *Regularitätsschranke* genannt wird.

Bemerkung 29 Unser FEM-Zugang wird mitunter als die *h-Version* der FEM bezeichnet, da sie den Grad k der Formfunktionen festhält und die Genauigkeit der Approximation durch Verfeinern der Gitterweite h erreicht. Bei höherer Glattheit von u kann man die Triangulierung $\mathcal{T}_h$ beibehalten und den Grad k der Polynome erhöhen, wodurch die so genannte *p-Version* der FEM entsteht. Vereint man beide Zugänge, so ergibt sich die *h-p-Version*.

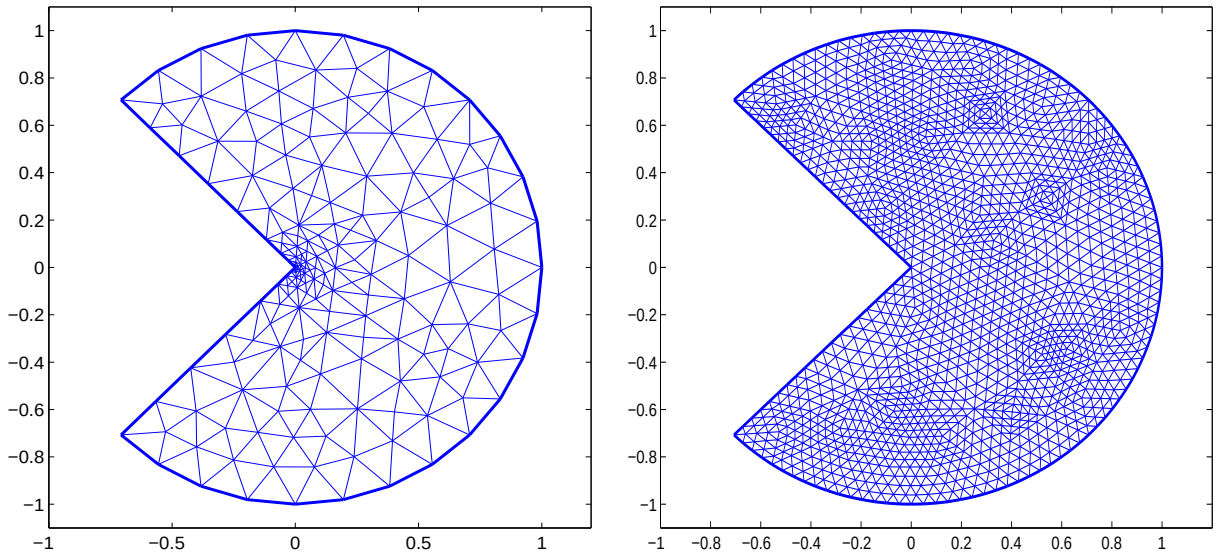

Abbildung 27: Adaptive Triangulierung (a) mit 316 Elementen und gleichmäßige Triangulierung (b) mit 3 152 Elementen zu Bsp. 30

Schließlich lässt sich auch eine Fehlerschätzung der Ordnung $l + 1$ in der $\mathcal{L}^2(\Omega)$-Norm

$$\|u - u_h\|_0 \leq Ch^{l+1} \|u\|_{l+1} \qquad \text{mit} \quad l = min(k, s-1) \tag{95}$$

für das Dirichlet-Problem und das Neumann-Problem über polygonalen konvexen Mengen Ω nachweisen (vgl. [16], S. 173).

Adaptive Triangulierung Besitzt die Lösung große Gradienten, z.B. an „Peaks“ und an Ecken von Ω, so ist an Stelle einer gleichmäßigen Verfeinerung eine lokale Anpassung der Triangulierung in der Umgebung der Punkte mit großem Fehler sinnvoll. Die Gewinnung

einer dazu erforderlichen Fehlerschätzung findet der interessierte Leser in [3]. Werden Elemente K adaptiv verfeinert, so entstehen neue Knoten an den Seitenmittelpunkten, die keine Ecke im Nachbardreieck bilden. Zwei derartige *hängende Knoten* sind in Abb. 8b dargestellt. Damit dennoch eine zulässige Triangulierung entsteht, werden diese Nachbardreiecke aufgespalten. Um bei sukzessiven Verfeinerungen eine Entartung der Elemente, d.h. einen großen Wert σ zu vermeiden, benutzt `pdetool` eine ausgefeilte Methode zur Halbierung der jeweils längsten Dreieckseiten [14].

Beispiel 30 Wir lösen die Poisson-Gleichung mit Dirichlet-Bedingungen auf dem in Abb. 27 dargestellten Kreissektor (vgl. [14])

$$-\Delta u \;=\; f(x,y), \quad (x,y) \in \Omega \qquad \text{und} \qquad u(x,y) \;=\; \varphi(x,y), \quad (x,y) \in \partial\Omega. \tag{96}$$

Dabei ist $f(x,y) = 0$ und $\varphi(x,y) = \cos(2/3 * \arctan(y/x))$ auf dem Kreisbogen bzw. $\varphi(x,y) = 0$ längs der beiden Radien.

1. Mit der MATLAB-Funktion `adaptmesh` wird iterativ eine automatische Anpassung der Triangulierung durchgeführt, bis nach 10 adaptiven Verfeinerungen das Gitter in Abb. 27a mit 316 Dreiecken erreicht wird. Die Lösung ist in Abb. 28a dargestellt, während Abb. 28b den Diskretisierungsfehler $u - u_h$ mit der exakten Lösung

$$u(x,y) \;=\; (x^2+y^2)^{1/3}\cos(2/3 * \arctan(y/x)) \tag{97}$$

zeigt. Der maximale absolute Fehler auf $\mathcal{T}_h$ ist $error = \max|u_h - u| = 0.0049$.

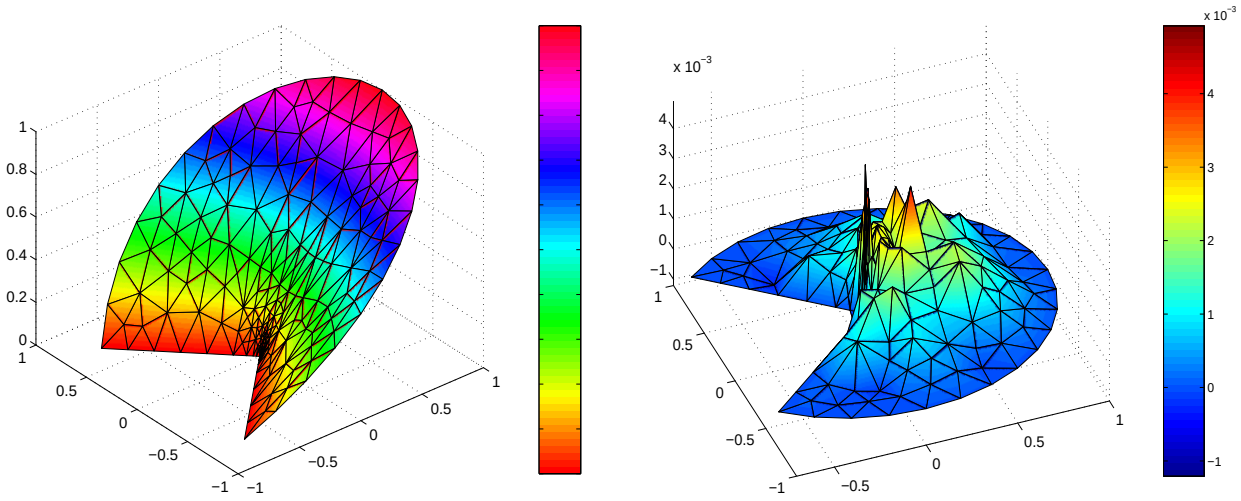

Abbildung 28: Näherungslösung (a) und Fehler (b) zu Bsp. 30 (1)

2. Die MATLAB-Funktionen `refinemesh` führt wie in Beispiel 26 eine gleichmäßige Verfeinerung der mittels `initmesh` erzeugten Delaunay-Triangulierung durch. Die erste Verfeinerung liefert den maximalen Fehler $error = 0.0121$ mit 788 Dreiecken, während eine zweite Verfeinerung das in Abb. 27b gezeichnete Gitter mit 3 152 Dreiecken und einem maximalen Fehler $error = 0.0078$ ergibt. Erst nach einer dritten Verfeinerung wird mit 12 608 Elementen der gewünschte Fehler $error = 0.0050$ erreicht! Eine adaptive Triangulierung ist in vielen praktischen Anwendungen der gleichmäßigen Verfeinerung vorzuziehen.

3. Wird durch $f(x,y) = \delta(x-0.5, y-0.5)$ eine zusätzliche Punktquelle bei $(0.5, 0.5)$ simuliert, so liefert eine adaptive Triangulierung mit 1 017 Elementen in Abb. 29a eine wesentlich

bessere Approximation als die in Abb. 29b gezeichnete Lösung über einem gleichmäßig verfeinerten Gitter mit 3 152 Dreiecken. □

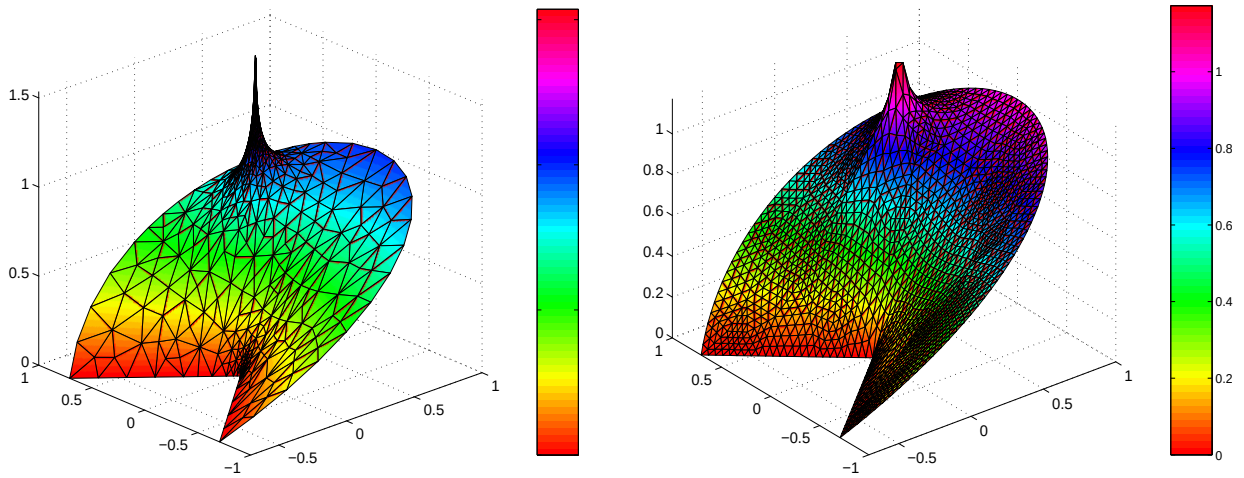

Abbildung 29: 1 017 Elemente (a) und 3 152 Elemente (b) bei einer Punktquelle

Verallgemeinerte Galerkin-Verfahren Zur Berechnung des Lastvektors b in (82) wurde eine Quadraturformel, die sogenannte Prismenregel benutzt, die eine zusätzliche Approximation $\mathcal{F}_h(v_h)$ des Funktionals $\mathcal{F}$ im Galerkin-Ansatz (60) bedeutet. Häufig müssen wir bei allgemeineren Funktionen a_{ij} des Differenzialoperators L die exakten Integrale der Steifigkeitsmatrix ebenfalls mit Quadraturformeln annähern. Allgemein entstehen wegen der Ersetzung der Bilinearform $\mathcal{A}$ durch h-abhängige Bilinearformen $\mathcal{A}_h$ auf diese Weise *verallgemeinerte Galerkin-Verfahren*

$$\mathcal{A}_h(u_h, v_h) \;=\; \mathcal{F}_h(v_h) \qquad \forall\, v_h \in V_h \tag{98}$$

für ein gesuchtes $u_h \in V_h$. Zu dieser Klasse gehören auch die praktisch bedeutsamen *Kollokationsmethoden*, auch *Pseudo-Spektralmethoden* genannt [4, 16].

Lassen wir in der Variationsgleichung (36) zu, dass der Raum der Testfunktionen (Testraum) V verschieden vom Grundraum U ist, so erhalten wir *Petrov-Galerkin-Verfahren*: Gesucht ist ein $u_h \in U_h$, das die endlichdimensionale Variationsgleichung

$$\mathcal{A}_h(u_h, v_h) \;=\; \mathcal{F}_h(v_h) \qquad \forall\, v_h \in V_h \tag{99}$$

erfüllt. U_h ist der Ansatzraum, während V_h der endlichdimensionale Testraum ist. Allgemeine Darstellungen zur Theorie dieser Verfahren findet man in [16].

4 Fazit

Die numerische Lösung partieller Differenzialgleichungen erfordert neben der Approximation der Differenzialgleichungen und Nebenbedingungen auch eine geeignete Darstellung des Integrationsgebietes $\Omega \subset \mathbb{R}^n$. Für komplizierte lineare oder nichtlineare PDGLn auf einfachen kartesischen Gebieten stellt die historisch ältere Finite-Differenzen-Methode einen sehr allgemeinen Zugang dar, während auf komplizierten Geometrien die Finite-Elemente-Methode flexibler ist und zudem schwache Lösungen approximiert.

Die *Finite-Differenzen-Methode (FDM)* ersetzt das Gebiet Ω durch ein achsenparalleles Gitter Ω_h der maximalen Schrittweite h und approximiert die Ableitungen der DGLn und Randbedingungen durch *finite Ausdrücke*. Als Verallgemeinerung elementarer Differenzenquotienten lassen sich damit die partiellen Ableitungen mit einer vorgegebenen Ordnung $\mathcal{O}(h^s)$ auf dem Gitter darstellen.

Nach Ersetzung aller Ableitungen und Funktionswerte entsteht ein lineares bzw. nichtlineares finites Gleichungssystem für die gesuchte *diskrete Lösung, die Gitterfunktion* u_h. Nach linearer Nummerierung der Gitterpunkte sollten für die großdimensionalen schwach besetzten (sparsen) Systeme *moderne iterative Löser* wie Newton-Krylov-Verfahren und Mehrgitter-Verfahren eingesetzt werden.

Die *diskrete Konvergenz* der Gitterfunktionen u_h gegen die exakte Lösung u lässt sich nachweisen, wenn man die *Konsistenz* der Approximation und die *diskrete Stabilität* der FDM verifiziert hat. Wesentliche Voraussetzung für die Konvergenztheorie ist eine hinreichende $\mathcal{C}^s$-Glattheit der im Problem auftretenden Funktionen. Gestattet der lokale Diskretisierungsfehler eine *asymptotische Entwicklung* nach Potenzen der Gitterweite h, so folgt unter allgemeinen Voraussetzungen auch eine entsprechende Entwicklung des globalen Fehlers. Damit ist das *Extrapolationsprinzip* anwendbar, mit dem eine effektive Fehlerschätzung und eine Verbesserung der Gitterfunktion möglich wird.

Die historisch neuere *Finite-Elemente-Methode (FEM)* basiert auf der *schwachen Form, der Variationsgleichung* des betrachteten Randwertproblems. Nach Multiplikation mit Testfunktionen und Integration über Ω gelingt es häufig, die Differentiationsordnung der Lösung u zu reduzieren. Damit sind auch Lösungen in *Sobolev-Räumen* $\mathcal{H}^s(\Omega)$ verallgemeinerter Ableitungen approximierbar.

Ein *Projektionsverfahren* ersetzt die lineare Variationsgleichung mittels Projektion durch eine Gleichung in endlichdimensionalen Unterräumen. Als bekannteste Zugänge liefern das *Galerkin-Verfahren*, das *Ritz-Verfahren* und das *Petrov-Galerkin-Verfahren* ein lineares Gleichungssystem in $\mathbb{R}^N$, das oft wesentliche Problemeigenschaften wie die Symmetrie und positive Definitheit erbt.

Die FEM approximiert eine Lösung u der Variationsgleichung durch stückweise polynomiale Funktionen des Grades k. Dazu wird das Gebiet $\overline{\Omega}$ mittels einer *zulässigen Triangulierung* der Feinheit h mit finiten Elementen überdeckt und eine nodale Basis mit *Formfunktionen* φ_i konstruiert.

Neben den beiden behandelten „klassischen“ Methoden sind moderne Zugänge wie Spektralmethoden und Pseudo-Spektralmethoden (Kollokationsmethoden), Finite-Volumen-Methoden und Randelement-Methoden von besonderer praktischer Bedeutung [4, 11, 16].

Literatur

[1] Alt, H. W.: *Lineare Funktionalanalysis*, Springer-Verlag Berlin 1992

[2] Chung-Yau Lam: *Applied Numerical Methods for Partial Differential Equations*, Prentice Hall New York 1994

[3] Eriksson, K.; Estep, D.; Hansbo, P.; Johnson, C.: *Computational Differential Equations*, Cambridge University Press 1996

[4] Fornberg, B.: *A Practical Guide to Pseudospectral Methods*, Cambridge University Press 1996

[5] Hackbusch, W.: *Multi-Grid Methods and Applications*, Springer-Verlag Berlin 1985

[6] Hanke-Bourgeois, M.: *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens*, B. G. Teubner-Verlag Stuttgart 2002

[7] Hoffmann, A.; Marx, B.; Vogt, W.: *Mathematik für Ingenieure 1. Lineare Algebra, Analysis – Theorie und Numerik*, Pearson Studium München 2005

[8] Isaacson, E.; Keller, H. B.: *Analyse numerischer Verfahren*, Verlag Harry Deutsch Frankfurt 1972

[9] Iserles, A.: *A First Course in the Numerical Analysis of Differential Equations*, Cambridge University Press 1996

[10] Knabner, P.; Angermann, L.: *Numerik partieller Differentialgleichungen*, Springer-Verlag Berlin 2000

[11] Larsson, S.; Thomee, V.: *Partielle Differentialgleichungen und numerische Methoden*, Springer-Verlag Berlin 2005

[12] Mangoldt, H. v.; Lösch, F.: *Einführung in die Höhere Mathematik*, Bd. IV, S. Hirzel Verlag Leipzig 1973

[13] Meis, T.; Marcowitz, U.: *Numerische Behandlung partieller Differentialgleichungen*, Springer-Verlag Berlin 1978

[14] Partial Differential Equation Toolbox. User's Guide. The Math Works Inc., 2002 `http://www.mathworks.com`

[15] Quarteroni, A.; Sacco, R.; Saleri, F.: *Numerische Mathematik, Band 1 und 2*, Springer-Verlag Berlin 2002

[16] Quarteroni, A.; Valli, A.: *Numerical Approximation of Partial Differential Equations*, 2. Auflage, Springer-Verlag Berlin 1997

[17] Stetter, H. J.: *Analysis of Discretization Methods for Ordinary Differential Equations*, Springer-Verlag Berlin 1973

[18] Törnig, W.; Spellucci, P.: *Numerische Mathematik für Ingenieure und Physiker. Band 1 und 2*, 2. Auflage, Springer-Verlag Berlin 1988

[19] Vogt, W.: *Zwei Anwendungen von Diskretisierungsverfahren für nichtlineare Operatorgleichungen*, Inst. f. Mathematik, Preprint No. M 12/02, TU Ilmenau 2002

[20] Walker, H. F.: *An Adaptation of Krylov Subspace Methods to Path Following Problems*, SIAM Journal on Scientific Computing, Vol. 21, No. 3, 1999, S. 1191-1198