2 Differenzen-Verfahren für elliptische Probleme

In diesem Kapitel werden wir zunächst die klassischen Differenzenapproximationen zur Lösung elliptischer Randwertaufgaben (RWAn) diskutieren. Der Übersichtlichkeit halber beschränken wir uns dabei auf das Modellproblem der Poisson-Gleichung in zwei Raumdimensionen mit Dirichletschen Randbedingungen, d. h. auf die 1. RWA:

$$Lu := -\Delta u = f \quad \text{in } \Omega, \quad u = g \text{ auf } \partial\Omega. \tag{2.0.1}$$

Das Definitionsgebiet $\Omega \subset \mathbb{R}^2$ wird zunächst wieder als glatt berandet oder als konvexes Polygongebiet vorausgesetzt. Die Problemdaten f, g sind ebenfalls glatt, so dass die im vorigen Kapitel beschriebenen Resultate anwendbar sind. Erweiterungen für Probleme mit variablen Koeffizienten oder anderen Randbedingungen sowie auf den dreidimensionalen Fall werden gegebenenfalls in Bemerkungen diskutiert.

2.1 Allgemeine Differenzenapproximationen

Zur Definition einer sog. „Differenzenapproximation“ der RWA wird das Lösungsgebiet Ω durch ein endliches (nicht notwendig äquidistantes oder kartesisches) Punktgitter überdeckt. Wir definieren disjunkte Punktmengen

$$\Omega_h := \{\text{„innere” Gitterpunkte in } \Omega\}, \quad \partial\Omega_h := \{\text{„Rand-Gitterpunkte” nahe bei } \partial\Omega\},$$

und setzen $\overline{\Omega}_h := \Omega_h \cup \partial\Omega_h$. Welche Punkte zu $\partial\Omega_h$ gehören, hängt von der Eigenart der gewählten Differenzenapproximation ab; im folgenden werden verschiedene Beispiele betrachtet. Der Parameter $h > 0$ beschreibt wie üblich die „Feinheit” des Gitters $\overline{\Omega}_h$, d. h. so etwas wie den maximalen (horizontalen oder vertikalen) Abstand benachbarter Gitterpunkte. Die Gitterpunkte in $\partial\Omega_h$ seien ähnlich dicht verteilt wie die in Ω_h. Die Differentialgleichung wird nun in den Punkten in Ω_h betrachtet und jede Ableitung durch einen Differenzenquotient ersetzt. Diese Differenzenappoximation greift dabei auch auf Randpunkte in $\partial\Omega_h$ zurück. So ergibt sich ein Differenzenschema zur Bestimmung einer diskreten Lösung $u_h : \overline{\Omega}_h \to \mathbb{R}$ der allgemeinen Gestalt

$$L_h u_h(P) = f_h(P) \text{ für } P \in \Omega_h, \quad u_h(P) = g_h(P) \text{ für } P \in \partial\Omega_h, \tag{2.1.2}$$

wobei f_h und g_h geeignete Approximationen der rechten Seite f bzw. der Randwerte g sind (im einfachsten Fall etwa $f_h(P) = f(P)$ und $g_h(P) = g(P)$). Auf natürliche Weise verstehen wir die Anwendung des Operators L_h auch auf die kontinuierliche Lösung $u \in C(\overline{\Omega})$. Zur Konstruktion konkreter Differenzenoperatoren definieren wir zu jedem Punkt eine Umgebung von (verschiedenen) Punkten

$$N(P) := \{Q_i,\, i = 0, ..., r_P\} \quad \text{(Konvention: } Q_0 := P),$$

auf denen eine Differenzenapproximation des Laplace-Operators definiert ist. Der Differenzenoperator hat dann die folgende Form:

$$L_h u_h(P) = \sum_{Q \in N(P)} \sigma(P,Q) u_h(Q), \tag{2.1.3}$$

mit gewissen Koeffizienten $\sigma(P,Q)$. Diese sind so zu bestimmen, dass die folgenden Forderungen erfüllt sind:

i) *Konsistenz:* Das Schema ist „konsistent", d. h.: Für den „Abschneidefehler"

$$\tau_h(P) := L_h u(P) - f_h(P), \quad P \in \Omega_h,$$

gilt

$$\max_{P \in \Omega_h} |\tau_h(P)| \to 0 \quad (h \to 0). \tag{2.1.4}$$

Wünschenswert wäre eine möglichst hohe „Konsistenzordnung" $m \geq 1$, d. h.:

$$\max_{P \in \Omega_h} |\tau_h(P)| = O(h^m). \tag{2.1.5}$$

Wir werden uns im Folgenden üblicherweise meist mit $m = 2$ begnügen. Aus Ökonomiegründen wird man r_P von moderater Größe wählen ($r_P \approx 4 - 24$ in zwei Raumdimensionen).

ii) *Stabilität:* Das Differenzenschema ist wohl-gestellt, d. h.: Es bestimmt eindeutige diskrete Lösungen $(u_h(P))_{P \in \overline{\Omega}_h}$, welche gleichmäßig bzgl. h stetig von den Daten des Problems abhängen. Dazu wäre z. B. eine Stabilitätsabschätzung der folgenden Form zu beweisen:

$$\max_{P \in \overline{\Omega}_h} |u_h(P)| \leq c_{\text{stab}} \Big\{ \max_{P \in \Omega_h} |L_h u_h(P)| + \max_{P \in \partial\Omega_h} |u_h(P)| \Big\}, \tag{2.1.6}$$

mit einer von h unabhängigen Konstante $c_{\text{stab}} > 0$.

iii) *Verträglichkeit:* Der Differenzenoperator L_h soll analoge charakteristische Eigenschaften wie der kontinuierliche L besitzen; z. B.: Symmetrie, Definitheit, Maximumprinzip etc..

Das Hauptziel der Konvergenzanalyse von Differenzenschemata ist der Nachweis, dass eine Konsistenzordnung m auch eine Konvergenzordnung m des Fehlers $e_h := u - u_h$ impliziert:

$$\max_{P \in \overline{\Omega}_h} |e_h(P)| = O(h^m), \tag{2.1.7}$$

vorausgesetzt, die Lösung u ist ausreichend regulär.

2.1.1 Konsistenz

Für den Fall $f_h(P) = f(P)$ betrachten wir o.B.d.A. den Punkt $P = Q_0 = 0$ und die Punktumgebung $N(P) = \{Q_i = (x_i, y_i),\ i = 0, ..., r_P\}$. Taylorentwicklung ergibt:

$$u(Q_i) = u(P) + x_i\partial_x u(P) + y_i\partial_y u(P) + \tfrac{1}{2}x_i^2\partial_x^2 u(P) + x_i y_i\partial_x\partial_y u(P) + \tfrac{1}{2}y_i^2\partial_y^2 u(P) + \cdots .$$

Wir wollen aus diesem Ansatz die Koeffizienten $\sigma_i := \sigma(P, Q_i)$ im Differenzenoperator so bestimmen, dass eine vorgegebene Konsistenzordnung garantiert ist. Mit dem Ansatz $\tau_h(P) = O(h^m)$ erhält man durch Koeffizientenvergleich als notwendige und hinreichende Bedingung für Konsistenz:

(B0) *Konsistenz:* Für die Koeffizienten des Differenzenschemas gilt:

$$\sum_{i=0}^{r_P}\sigma_i = \sum_{i=0}^{r_P} x_i\sigma_i = \sum_{i=0}^{r_P} y_i\sigma_i = \sum_{i=0}^{r_P} x_i y_i\sigma_i = 0, \qquad \tfrac{1}{2}\sum_{i=0}^{r_P} x_i^2\sigma_i = \tfrac{1}{2}\sum_{i=0}^{r_P} y_i^2\sigma_i = 1. \quad (2.1.8)$$

Aus diesen Beziehungen sind die Koeffizienten σ_i, $i = 0, ..., r_P$, im Differenzenoperator L_h zu bestimmen. Für Konsistenz sind also im allgemeinen Fall eines völlig unstrukturierten Gitters mindestens 6 Punkte in $N(P)$ erforderlich. Die Konsistenz des Differenzenschemas ist offenbar äquivalent dazu, daß der Differenzenoperator „exakt" ist für quadratische Polynome:

$$u \in P_2: \quad L_h u(P) = Lu(P), \quad P \in \Omega_h. \quad (2.1.9)$$

Da man in der Regel Diskretisierungen auf Folgen von Gittern mit Gitterweiten $h \to 0$ betrachtet, führt man skalierte Parameter $\xi_i = h^{-1}x_i$, $\eta_i = h^{-1}y_i$ ein, um sich von der h-Abhängigkeit in den Koeffizienten zu befreien. Wegen

$$\tfrac{1}{2}h^2\sum_{i=0}^{r_P}\xi_i^2\sigma_i = \tfrac{1}{2}h^2\sum_{i=0}^{r_P}\eta_i^2\sigma_i = 1,$$

ist $\sigma_i = 0$ oder $\sigma_i \sim h^{-2}$. Wenn also eine Lösung für $(\sigma_0, \ldots, \sigma_{r_P})$ existiert, dann ergibt sich für den Konsistenzfehler

$$\tau_h(P) = (L_h u - Lu)(P) = O(h^3\xi_i^3\sigma_i + \ldots) = O(h). \quad (2.1.10)$$

Auch bei ganz irregulärer Gitterpunktsanordnung erhält man somit schon mindestens eine Konsistenzordnung $m = 1$. Es gibt mehrere Möglichkeiten, die Ordnung zu erhöhen:

- Man nimmt mehr Gitterpunkte in die Menge $N(P)$ auf.
- Man ordnet das Gitter regulär an, um in der Taylor-Entwicklung des Abschneidefehlers Weghebeffekte zu erzielen.

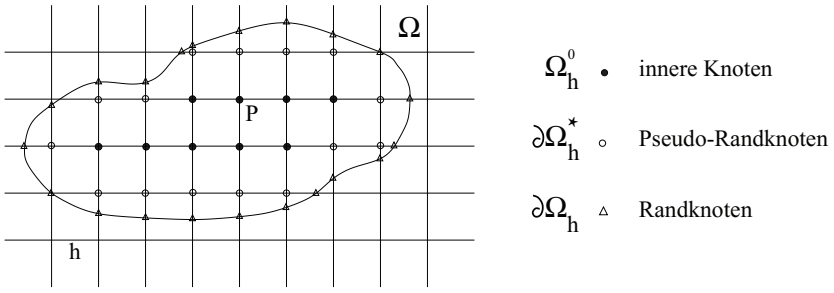


Abbildung 2.1: Gitter für Differenzenapproximation

Wir werden uns nun im Folgenden mit regulären, speziell äquidistanten, kartesischen Gittern beschäftigen. Der Einfachheit halber sollen die Gitterpunkte äquidistant entlang von Parallelen zu den Koordinatenachsen angeordnet sein. Dabei bezeichnet $\mathbb{R}^2_h$ das gesamte den $\mathbb{R}^2$ überdeckende Punktgitter. Die Schnittpunkte der Gitterlinien mit dem Rand $\partial\Omega$ bilden dabei natürliche Stützpunkte für die Approximation der Randwerte. Die „Gitterweite“ h hat auf solchen Gittern eine natürliche Bedeutung.

Die einfachste Approximation $\Delta_h \approx \Delta$ des Laplace-Operators verwendet zentrale Differenzenquotienten 2. Ordnung in jeder der Koordinatenrichtungen. Man erhält den sog. „5-Punkte-Operator“:

$$\Delta_h^{(5)} u_h(x,y) := \frac{1}{h^2}\Big\{u(x\pm h,y) + u(x,y\pm h) - 4u(x,y)\Big\}$$

für innere Gitterpunkte $P = (x,y) \in \Omega_h$. Wir verwenden hier die Konvention, dass „$\pm h$“ abkürzend für die Summe der beiden Terme „$+h$“ und „$-h$“ steht. Die Konsistenzordnung dieser Differenzenapproximation ist wegen der Äquidistanz des Gitters und der symmetrischen Plazierung der Stützpunkte $m = 2$:

$$|\Delta_h^{(5)} u(P) - \Delta u(P)| \le \tfrac{1}{6} M_4(u) h^2, \tag{2.1.11}$$

wobei $M_4(u) := \max_{\overline{\Omega}}\{|\partial_x^i \partial_y^j u|,\ i+j=4\}$. Diese Aussage bleibt sinngemäß gültig, wenn das Gitter nur in jeder einzelnen der Koordinatenrichtungen äquidistant ist. Dagegen ginge die Konsistenzordnung auf $m = 1$ zurück, wenn das Gitter zwar kartesisch, aber innerhalb einer Koordinatenrichtung nicht äquidistant wäre. Höhere Approximationsordnungen lassen sich z. B. auf einem gleichförmigen Gitter durch Hinzunahme von weiteren Punkten in der Umgebung des Auswertungspunkts (x,y) erzielen:

i) Approximation der zweiten Ableitungen im Laplace-Operator durch zentrale Differenzenquotienten auf jeweils 5 Punkten ergibt den „gestreckten“ 9-Punkte-Operator:

$$\Delta_h^{(9)} u_h(x,y) = \frac{1}{12h^2}\Big\{-u(x\pm 2h,y) + 16u(x\pm h,y) - u(x,y\pm 2h) + 16u(x,y\pm h) - 60u(x,y)\Big\}.$$

Dieser Differenzenoperator hat offensichtlich die Konsistenzordnung $m = 4$, doch hat die zugehörige Koeffizientenmatrix wegen des Vorzeichenwechsels ungünstige Eigenschaften.

ii) Der „kompakte“ 9-Punkte-Operator verwendet neben dem Auswertungspunkt (x, y) die 8 direkten Nachbarpunkte $(x \pm h, y)$, $(x \pm h, y \pm h)$, $(x, y \pm h)$:

$$\bar{\Delta}_h^{(9)} u_h(x,y) = \frac{1}{6h^2}\Big\{4u(x\pm h, y) + 4u(x, y\pm h) + u(x\pm h, y\pm h) - 20u(x,y)\Big\}.$$

Dieses Schema hat zunächst auch nur die Ordnung $m = 2$, doch kann man es durch eine Modifikation bei der Auswertung der rechten Seite auf die Ordnung $m = 4$ bringen (Übungsaufgabe):

$$f(x,y) \to f_h(x,y) := f(x,y) + \tfrac{1}{12} h^2 \Delta_h^{(5)} f(x,y). \tag{2.1.12}$$

Approximation entlang des Randes: Bei der Approximation am (möglicherweise gekrümmten) Rand des Gebiets gibt es verschiedene Möglichkeiten, die, wie wir später sehen werden, durchaus auf unterschiedliche Approximationsordnungen führen. Wir betrachten wieder den 5-Punkte-Operator und definieren zunächst die folgenden Gitterpunktmengen:

$$\Omega_h := \{P \in \mathbb{R}_h^2 |\ N(P) \subset \overline{\Omega}\}, \quad \partial\Omega_h := \bigcup_{P \in \Omega_h} N(P) \setminus \Omega_h.$$

i) *Konstante Randwertextrapolation:* In Punkten von Ω_h wird der 5-Punkte-Operator angesetzt, und in Punkten $P \in \partial\Omega_h$ werden die Randwerte $u_h(P) = g(P^0)$ verwendet. Dabei ist P^0 der P entlang einer der Koordinatenachsen am nächsten gelegene Punkt auf $\partial\Omega$. Dies ergibt entlang des Randes nur eine Approximationsordnung $m = 1$.

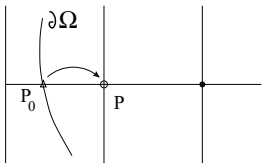


Abbildung 2.2: Schema der konstanten Randapproximation

ii) *Lineare Randwertinterpolation:* Jeder Punkt $P \in \partial\Omega_h$ liegt in x- und/oder y-Richtung zwischen zwei Punkten $P_0 \in \partial\Omega$ und $P_1 \in \Omega_h$ mit Abständen $0 \le \alpha h < h$ zu P_0 und h zu P_1. Man setzt dann (lineare Interpolation):

$$u_h(P) := \frac{1}{1+\alpha}\big\{g(P_0) + \alpha u_h(P_1)\big\}. \tag{2.1.13}$$

Damit wird eine implizite Kopplung der Randwerte an die „inneren“ Lösungswerte bewirkt. Diese Randwertapproximation hat die Ordnung $m = 2$.

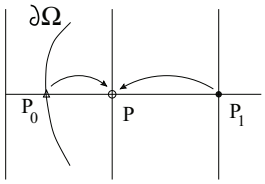


Abbildung 2.3: Schema der linearen Randapproximation

iii) *Shortley*[1]*-Weller*[2]*-Approximation:* Wir definieren die folgenden Punktmengen:

$$\Omega_h^0 := \{P \in \mathbb{R}_h^2 |\ N(P) \subset \Omega\}, \quad \partial\Omega_h^* := \bigcup_{P\in\Omega_h^0} N(P) \setminus \Omega_h,$$
$$\Omega_h := \Omega_h^0 \cup \partial\Omega_h^*, \quad \partial\Omega_h := \{\text{Schnittpunkte der Gitterlinien mit } \partial\Omega\}.$$

In Punkten $P \in \Omega_h$ wird wieder der normale 5-Punkte-Operator und in den „fiktiven" Randpunkten $P \in \partial\Omega_h^*$ der modifizierte 5-Punkte-Operator verwendet (siehe die schematische Darstellung):

$$-\Delta_h^* u_h(x,y) := h^{-2}\Big\{\Big(\frac{2}{\alpha}+\frac{2}{\beta}\Big)u_h(x,y) - \frac{2}{1+\alpha}u_h(x+h,y) - \frac{2}{\alpha(1+\alpha)}u_h(x-\alpha h,y) - \frac{2}{1+\beta}u_h(x,y+h) - \frac{2}{\beta(1+\beta)}u_h(x,y-\beta h)\Big\} = f(x,y)\,,$$

gemäß

$$-\Delta_h^* u_h = f \quad \text{auf } \partial\Omega_h^*, \tag{2.1.14}$$

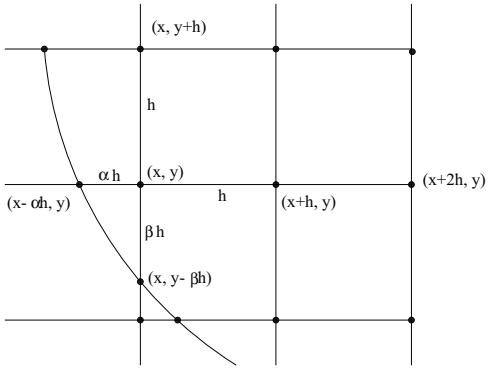
mit den Werten $u_h(P) := g(P)$ auf $\partial\Omega_h$. Für diesen Differenzenoperator gilt

$$|\Delta_h^* u(P) - \Delta u(P)| \le \tfrac{2}{3} M_3(u) h, \tag{2.1.15}$$

wobei $M_3(u) := \max_{\overline{\Omega}}\{|\partial_x^i \partial_y^j u|,\ i+j=3\}$.

[1]George H. Shortley (1910–????): US-Amerikanischer Astro-Physiker; wirkte 1935-2011 als Prof. an der Johns Hopkins University.

[2]Royal Weller (????–????): US-Amerikanischer Physiker und Ingenieur; die nach ihm und G. H. Shortley benannte Methods findet sich in „The numerical solution of Laplace's equation", J. Appl. Physics 9, 334 (1938); Mitherausgeber zweier Lehrbücher „Modern Physics for the Engineer" (1954) und „Modern Mathematics for the Engineer" (2013).



h Gitterweite
(x, y) Stützpunkt
$0 < \alpha,\ \beta \leq 1,\ \alpha\beta < 1$

Abbildung 2.4: Schema der Shortley-Weller-Approximation

Die Verwendung dieser Differenzenapproximationen führt auf Schemata der Form

$$L_h u_h = f_h \quad \text{in } \Omega_h, \qquad R_h u_h = g_h \quad \text{auf } \partial\Omega_h^*, \tag{2.1.16}$$

wobei der „Randoperator“ R_h die jeweilige Randwertapproximation beschreibt. Dabei werden die echten Randwerte von u_h auf $\partial\Omega_h$ unter Umständen implizit mit Werten im Innern verkoppelt. In diesem Fall sind alle Werte $u_h(P),\ P \in \overline{\Omega}_h$, zu bestimmen. Der Differenzenoperator L_h steht für den normalen 5-Punkte-Operator in Ω_h und gegebenfalls den modifizierten Shortley-Weller-Operator auf $\partial\Omega_h^*$. Im einfachsten Fall der trivialen Randwertapproximation $R_h u_h = u_h$ können die Randwerte $u_h(P) = g_h(P),\ P \in \partial\Omega_h$, direkt eliminiert werden. Wir werden im folgenden nur diesen Fall weiter diskutieren.

Bemerkung: Die obigen Differenzenschemata haben natürliche Analoga in drei Raumdimensionen. Man spricht dann aus naheliegendem Grund vom „7-Punkte-Operator“. Dessen Abschneidefehler genügt der Abschätzung

$$|\Delta_h^{(7)} u(P) - \Delta u(P)| \leq \tfrac{1}{4} M_4(u) h^2. \tag{2.1.17}$$

2.2 Eigenschaften der Differenzengleichungen

Ausgangspunkt der folgenden Untersuchungen ist das Differenzenschema der Gestalt

$$L_h u_h(P) := \sum_{Q \in N(P)} \sigma(P, Q) u_h(Q) = f_h(P), \quad P \in \Omega_h, \tag{2.2.18}$$

$$u_h(P) = g_h(P), \quad P \in \partial\Omega_h, \tag{2.2.19}$$

mit geeigneten Approximationen $f_h(\cdot)$ zu f und $g_h(\cdot)$ zu g. Wir definieren die Koeffizenten $\sigma(P,Q) := 0$ für Punkte $Q \notin N(P)$. Für $P \in \Omega_h$ gilt dann

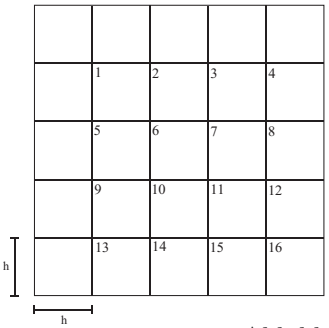
$$\sum_{Q\in\Omega_h} \sigma(P,Q)u_h(Q) = f_h(P) - \sum_{Q\in\partial\Omega_h} \sigma(P,Q)g_h(Q). \tag{2.2.20}$$

Bei (beliebiger) Numerierung der Gitterpunkte etwa gemäß $\Omega_h = \{P_n,\, n = 1, ..., N\}$, $\partial\Omega_h = \{P_n,\, n = N+1, ..., N+M\}$ ergibt sich ein quadratisches Gleichungssystem für den Vektor der (inneren) Knotenwerte $U = (U_n)_{n=1}^N$, $U_n := u_h(P_n)$:

$$AU = F, \tag{2.2.21}$$

mit $A = (A_{nm})_{n,m=1}^N$, $F = (F_n)_{n=1}^N$, wobei

$$A_{nm} := \sigma(P_n, P_m), \quad F_n := f_h(P_n) - \sum_{m=N+1}^{N+M} \sigma(P_n, P_m)g_h(P_m).$$



$h = \frac{1}{m+1}$ Gitterweite

$N = m^2$ „innere“ Gitterpunkte

Abbildung 2.5: Differenzengitter

Beispiel („Modellproblem”): Im Fall des Einheitsquadrats $\Omega = (0,1)^2$ ergibt sich für den 5-Punkte-Operator bei zeilenweiser Durchnumerierung des Gitters $\overline{\Omega}_h = \{P_{ij}\}_{i,j=0}^{m+1}$ die folgende dünn-besetzte Matrix der Dimension $N = m^2$:

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

mit der $m\times m$-Einheitsmatrix I_m. Die rechte Seite ist $F := (f(x_{11}), \ldots, f(x_{mm}))^T$. Die Matrix A ist eine dünn besetzte Bandmatrix mit der Halbbandbreite m, symmetrisch und (irreduzibel) diagonal-dominant. Damit ist sie auch regulär und positiv definit.

Bemerkung: In drei Raumdimensionen hat die entsprechende Matrix die Dimension $N = m^3$ und die Halbbandbreite m^2. Ansonsten hat sie dieselben Eigenschaften wie in zwei Dimensionen.

Im allgemeinen Fall ist für moderates r_P die Matrix A dünn besetzt, aber häufig wegen der Randwertapproximation *nicht* symmetrisch. Dies wäre ein schwerwiegender Nachteil, etwa bei der Approximation von Eigenwertaufgaben zum Laplace-Operator. Um die Regularität von A bzw. die Lösbarkeit der Differenzengleichung garantieren zu können, formulieren wir die folgenden Strukturbedingungen in numerierungs-unabhängiger Form:

(B1) *Erweiterte Diagonaldominanz:* Für Punkte $P \in \Omega_h$ gilt:

$$\text{(i)} \qquad \sum_{Q \in \overline{\Omega}_h,\, Q \neq P} |\sigma(P,Q)| \leq |\sigma(P,P)|\,, \tag{2.2.22}$$

und für Punkte $P \in \Omega_h^* := \{Q \in \Omega_h :\ N(Q) \cap \partial\Omega_h \neq \emptyset\}$:

$$\text{(ii)} \qquad \sum_{Q \in \Omega_h,\, Q \neq P} |\sigma(P,Q)| < |\sigma(P,P)|\,. \tag{2.2.23}$$

(B2) *Nicht-negativer Typ:* Für Punkte $P \in \Omega_h$ und $Q \in \overline{\Omega}_h \setminus \{P\}$ gilt:

$$\sigma(P,P) > 0, \qquad \sigma(P,Q) \leq 0\,. \tag{2.2.24}$$

(B3) *Zusammenhang:* Es ist $\partial\Omega_h \neq \emptyset$, und mit $N(P) := \{Q \in \overline{\Omega}_h,\, \sigma(P,Q) \neq 0\}$ gilt für jede echte Teilmenge $S_h \subset \Omega_h$:

$$\Big(\bigcup_{P \in S_h} N(P) \Big) \cap (\Omega_h \setminus S_h) \neq \emptyset\,. \tag{2.2.25}$$

Die ersten beiden Bedingungen (B1) und (B2) werden von vielen Differenzenschemata, insbesondere solchen höherer als zweiter Ordnung (z. B.: der „gestreckte" 9-Punkte-Operator), nicht erfüllt. Die hier betrachteten 5-Punkte-Schemata mit Randapproximation genügen ihnen aber. Wir zeigen im Folgenden, wie bei den einzelnen Randapproximationen die Bedingung (B1ii) erfüllt ist.

i) Konstante Randwertextrapolation: Für $P \in \Omega_h^*$ gilt

$$\sum_{Q \in \Omega_h,\, Q \neq P} |\sigma(P,Q)| \leq 3h^{-2} = \tfrac{3}{4}\,|\sigma(P,P)|\,,$$

ii) Lineare Randwertinterpolation: Für $P \in \Omega_h^*$ gilt

$$\sum_{Q \in \Omega_h,\, Q \neq P} |\sigma(P,Q)| \leq \big(4 - \tfrac{\alpha}{1+\alpha}\big)h^{-2} \leq 4h^{-2} = |\sigma(P,P)|\,,$$

iii) Shortley-Weller-Randwertapproximation: Für $P \in \Omega_h^*$ gilt

$$\sum_{Q \in \Omega_h,\, Q \neq P} |\sigma(P,Q)| \leq \big(\tfrac{2}{1+\alpha} + \tfrac{2}{1+\beta}\big)h^{-2} \leq \tfrac{1}{2}\,\big(\tfrac{2}{\alpha} + \tfrac{2}{\beta}\big)h^{-2} = \tfrac{1}{2}\,|\sigma(P,P)|\,.$$

Die dritte Bedingung (B3) sichert die Kopplung eines jeden inneren Gitterpunktes $P \in \overline{\Omega}_h$ mit allen anderen, insbesondere mit den Randpunkten. Die Eigenschaft des kontinuierlichen Problems, dass sich eine kleine Störung in einem Punkt global bemerkbar macht, spiegelt sich hier wider.

Die obigen Eigenschaften des Differenzenschemas korrespondieren zu den bereits bekannten der zugehörigen Koeffizientenmatrix A. unabhängig von der gewählten Numerierung der Gitterpunkte implizieren die Bedingungen (B1) und (B3) zunächst die einfache Diagonaldominanz von A, darüber hinaus aber auch die stärkere „irreduzible Diagonaldominanz" (analog dem „schwachen Zeilensummenkriterium" für die Konvergenz des bekannten Jacobi- oder des Gauß-Seidel-Verfahrens). Die Bedingung (B2) entspricht einer analogen für A.

Um Existenz und Eindeutigkeit einer Lösung für das allgemeine Differenzenschema garantieren zu können, geht man ähnlich wie im Kontinuierlichen vor und nutzt ein diskretes Analogon des „Maximumprinzips".

Satz 2.1 (Diskretes Maximumprinzip): *Unter den Voraussetzungen (B1i), (B2) und (B3) genügt das zugehörige Differenzenschema einem „diskreten Maximumprinzip", d. h.: Gitterfunktionen u_h mit der Eigenschaft*

$$L_h u_h(P) \le 0, \quad P \in \Omega_h, \tag{2.2.26}$$

haben in Ω_h kein positives Maximum. Genauer gilt $u_h \le 0$ auf $\overline{\Omega}_h$ oder

$$\max_{\overline{\Omega}_h} u_h \le \max_{\partial\Omega_h} u_h. \tag{2.2.27}$$

Beweis: Wir nehmen an, dass (2.2.26) für ein u_h erfüllt ist, und führen den Beweis indirekt. Es gebe also einen Punkt $P_0 \in \Omega_h$, so dass

$$M := u_h(P_0) = \max_{\overline{\Omega}_h} u_h > 0, \quad \max_{\partial\Omega_h} u_h < M.$$

Ausgehend von $L_h u_h(P) \le 0$, folgt mit Bedingung (B2):

$$\begin{aligned} u_h(P_0) &\le -\sum_{Q\neq P_0} \frac{\sigma(P_0,Q)}{\sigma(P_0,P_0)}\, u_h(Q) = \sum_{Q\neq P_0} \Big|\frac{\sigma(P_0,Q)}{\sigma(P_0,P_0)}\Big|\, u_h(Q) \\ &= \sum_{Q\neq P_0} \Big|\frac{\sigma(P_0,Q)}{\sigma(P_0,P_0)}\Big|\, u_h(P_0) + \sum_{Q\neq P_0} \Big|\frac{\sigma(P_0,Q)}{\sigma(P_0,P_0)}\Big|\, \{u_h(Q) - u_h(P_0)\}. \end{aligned}$$

Weiter gilt wegen Bedingung (B1i)

$$\sum_{Q\neq P_0} \Big|\frac{\sigma(P_0,Q)}{\sigma(P_0,P_0)}\Big| \le 1,$$

und folglich ($u_h(P_0) = M$)

$$M \le M + \sum_{Q \in N(P_0) \setminus \{P_0\}} \Big| \frac{\sigma(P_0, Q)}{\sigma(P_0, P_0)} \Big| \{u_h(Q) - M\}.$$

Da nach Voraussetzung $u_h(Q) \le M$ ist, muss also $u_h(Q) = M$ in allen Punkten $Q \in N(P_0)$ (d. h. solchen mit $\sigma(P_0, Q) \neq 0$) sein, da sonst ein Widerspruch entsteht. Anwendung derselben Schlussweise für alle Punkte in $N(P_0)$ liefert

$$u_h(Q) = M, \quad Q \in \bigcup_{P \in N(P_0)} N(P).$$

Mit Hilfe der Bedingung (B3) erschließen wir dann durch sukzessive Fortsetzung dieses Arguments, dass $u_h \equiv M$ auf Ω_h. Nach Voraussetzung ist $\partial\Omega_h \neq \emptyset$, so dass es wegen (B3) einen Punkt $Q \in \partial\Omega_h \cap N(P)$ mit einem „inneren“ Punkt $P \in \Omega_h$ geben muss. Für diesen folgt dann $u_h(Q) = M$, was im Widerspruch zur Annahme $\max_{\partial\Omega_h} u_h < M$ steht. Q.E.D.

Das diskrete Maximumprinzip hat eine Reihe von wichtigen Folgerungen für die obigen Differenzenschemata.

Korollar 2.1 (Eindeutigkeit): *Unter den Voraussetzungen (B1), (B2) und (B3) besitzt das Differenzenschema (2.2.18), (2.2.19) genau eine Lösung* u_h. *Im Falle* $f_h \ge 0$ *folgt aus* $g_h \ge 0$ *auch* $u_h \ge 0$.

Beweis: i) Wegen der Äquivalenz von (2.2.18), (2.2.19) zu dem Gleichungssystem $A_h U = F_h$ genügt es, die Eindeutigkeit zu zeigen. Seien also $u_h^{(1)}$ und $u_h^{(2)}$ zwei Lösungen:

$$L_h u_h^{(i)} = f_h \quad \text{in } \Omega_h, \quad u_h^{(i)} = g_h \quad \text{auf } \partial\Omega_h.$$

Für die Differenz $w_h := u_h^{(1)} - u_h^{(2)}$ gilt dann

$$L_h w_h = 0 \quad \text{in } \Omega_h, \quad w_h = 0 \quad \text{auf } \partial\Omega_h.$$

Nun wird das diskrete Maximumprinzip auf $L_h w_h \le 0$ sowie auf $L_h(-w_h) \le 0$ angewendet. Ersteres ergibt $w_h \le 0$ und letzteres $w_h \ge 0$ und folglich $w_h \equiv 0$.

ii) Sei nun $f_h \ge 0$ und $g_h \ge 0$. Dann impliziert das diskrete Maximumprinzip angewendet für $-u_h$, dass entweder $u_h \ge 0$ oder $\min_{\Omega_h} u_h = -\max_{\Omega_h}(-u_h) \ge -\max_{\partial\Omega_h}(-g_h) = \min_{\partial\Omega_h} g_h$, was zu beweisen war. Q.E.D.

Im nächsten Abschnitt werden wir das diskrete Maximumprinzip verwenden, um die stetige Abhängigkeit der Lösungen von den Problemdaten zu zeigen. Dies wird sich als Nebenprodukt einer sehr viel stärkeren Stabilitätsungleichung ergeben.

Korollar 2.2 (Inverse Monotonie): *Unter den Voraussetzungen (B1), (B2) und (B3) ist die zum Differenzenoperator L_h bei beliebiger Numerierung der Gitterpunkte gehörende Koeffizientenmatrix A eine sog. „M-Matrix" (invers-monotone Matrix), d. h.: Ihre Inverse A^{-1} ist elementweise nicht negativ:*

$$A^{-1} \geq 0. \tag{2.2.28}$$

Beweis: Die Matrix A wird wie üblich geschrieben als Summe $A = L + D + R$ einer linken unteren Dreiecksmatrix L, einer Diagonalmatrix D und einer rechten oberen Dreiecksmatrix R. Die Bedingungen (B1), (B2) und (B3) implizieren, wie oben bereits bemerkt, die (irreduzible) Diagonaldominanz von A und damit 1) die Regularität von A und D und 2) die Kontraktivität $\mathrm{spr}(J) < 1$ der Jacobi-Matrix $J := -D^{-1}(L + R)$. Dann existiert eine (natürliche) Matrizennorm $\|\cdot\|$, bzgl. derer $\|J\| < 1$ ist. Hiermit folgt die Existenz der Reihe (im Sinne der Matrizenkonvergenz)

$$\sum_{k=0}^{\infty} J^k = (I - J)^{-1}.$$

Bedingung (B2) garantiert, dass (elementweise) $J = -D^{-1}(L+R) \geq 0$ und folglich auch

$$A^{-1}D = (D^{-1}A)^{-1} = (I + D^{-1}(L + R))^{-1} = (I - J)^{-1} = \sum_{k=0}^{\infty} J^k \geq 0.$$

Wegen $D^{-1} \geq 0$ impliziert dies $A^{-1} \geq 0$. Q.E.D.

Die Matrix A^{-1} ordnet der rechten Seite f_h und den Randwerten g_h eindeutig den Lösungsvektor U der Differenzengleichung zu:

$$A^{-1} : \{f_h, g_h\} \to U = A^{-1}F(f_h, g_h).$$

Damit ist A^{-1} das algebraische Äquivalent der kontinuierlichen Greenschen Funktion $G(\cdot, \cdot)$:

$$G : \{f, g\} \to u(x) = \int_{\Omega} G(x, y)f(y)\, dy - \int_{\partial\Omega} \partial_n G(x, y)g(y)\, dy.$$

Als Folgerung des Maximumprinzips ist $G(x, y) \geq 0$, $x \neq y$, was analog zur gerade gezeigten Eigenschaft $A^{-1} \geq 0$ ist.

2.2.1 Das Konvergenzverhalten von Differenzenverfahren

Die Grundlage der Fehleranalyse für die oben eingeführten Differenzenschemata ist wie im Fall von Differenzenverfahren für gewöhnliche Differentialgleichungen eine asymptotische Stabilitätsabschätzung. Wir formulieren diese im folgenden Satz für ein allgemeines Differenzenschema

$$L_h u_h = f_h \quad \text{in } \Omega_h, \qquad u_h = g_h \quad \text{auf } \partial\Omega_h. \tag{2.2.29}$$

Wir setzen im folgenden stets voraus, dass dieses Schema konsistent mit der RWA ist.

Satz 2.2 (Stabilität): *Das Differenzenschema (2.2.29) sei konsistent und genüge den Bedingungen (B1), (B2) und (B3). Dann gilt für jede Gitterfunktion u_h die Stabilitätsabschätzung*

$$\max_{P\in\overline{\Omega}_h} |u_h(P)| \le \tfrac{1}{4} d_\Omega^2 \max_{\Omega_h} |L_h u_h(P)| + \max_{P\in\partial\Omega_h} |u_h(P)|\,, \tag{2.2.30}$$

wobei $d_\Omega := \operatorname{diam}(\Omega)$.

Beweis: Wir argumentieren im folgenden ähnlich wie auf der kontinuierlichen Ebene. Durch die folgende Vorschrift für beliebiges, festes $Q \in \overline{\Omega}_h$

$$L_h G_h(\cdot, Q) = h^{-2}\delta(\cdot, Q) \quad \text{in } \Omega_h, \quad G_h(\cdot, Q) = \delta(\cdot, Q) \quad \text{auf } \partial\Omega_h, \tag{2.2.31}$$

wird ein diskretes Analogon $G_h(P,Q) : \overline{\Omega}_h \times \overline{\Omega}_h \to \mathbb{R}$ zur kontinuierlichen Greenschen Funktion (des Laplace-Operators) definiert. Dabei ist $\delta(P,Q)$ das übliche „Kronecker[3]-Symbol". Aus dem diskreten Maximumprinzip folgt, dass $G_h(P,Q) \ge 0$. Für Gitterfunktionen v_h gilt dann die diskrete „Greensche Identität"

$$v_h(P) = h^2 \sum_{Q\in\Omega_h} G_h(P,Q) L_h v_h(Q) + \sum_{Q\in\partial\Omega_h} G_h(P,Q) v_h(Q), \quad P \in \overline{\Omega}_h. \tag{2.2.32}$$

Um dies zu sehen, bezeichnen wir die rechte Seite von (2.2.32) mit w_h und erhalten mit den Eigenschaften der Greenschen Funktion

$$L_h w_h = L_h v_h \quad \text{in } \Omega_h, \quad w_h = v_h \quad \text{auf } \partial\Omega_h.$$

Die Eindeutigkeit der Lösungen des Diffenzenschemas impliziert dann $w_h \equiv v_h$. Aus der diskreten Greenschen Identität folgt für jede Gitterfunktion v_h

$$|v_h(P)| \le h^2 \sum_{Q\in\Omega_h} G_h(P,Q) \max_{\Omega_h} |L_h v_h| + \sum_{Q\in\partial\Omega_h} G_h(P,Q) \max_{\partial\Omega_h} |v_h|, \quad P \in \overline{\Omega}_h. \tag{2.2.33}$$

i) Wegen der Konsistenz des Schemas gilt für die Gitterfunktion $w_h \equiv 1$ notwendig $L_h w_h \equiv 0$, so dass aus der Greenschen Identität folgt:

$$1 = h^2 \sum_{Q\in\Omega_h} G_h(P,Q) L_h w_h(Q) + \sum_{Q\in\partial\Omega_h} G_h(P,Q) w_h(Q) = \sum_{Q\in\partial\Omega_h} G_h(P,Q)\,. \tag{2.2.34}$$

Dies impliziert eine Schranke für die zweite Summe in (2.2.33).

[3]Leopold Kronecker (1823–1891): Deutscher Mathematiker; wirkte in Berlin als „Privatgelehrter"; betrieb die Arithmetisierung der Mathematik; wichtiger Vertreter des „Konstruktivismus", welcher die generelle Verwendung des Widerspruchsbeweises und des „aktual Unendlichen" in Form z. B. der allgemeinen reellen Zahlen ablehnt.

ii) Jeder Punkt $P_0 \in \overline{\Omega}_h$ ist Mittelpunkt eines $\overline{\Omega}$ enthaltenden Kreises mit Radius d_Ω. O.b.d.A. sei hier $P_0 = 0$. Für die Funktion $w(P) := \frac{1}{4}|P|^2$ gilt in Punkten $P \in \Omega_h$:

$$L_h w(P) = L_h w(P) + \Delta w(P) - \Delta w(P) = M_3(w)O(h) - \Delta w(P) = -1,$$

da $M_3(w) = 0$. Wir definieren nun die Gitterfunktion

$$v_h(P) := h^2 \sum_{Q \in \Omega_h} G_h(P, Q).$$

Durch elementares Nachrechnen verifiziert man, dass

$$\begin{aligned} L_h v_h = 1 \quad &\text{in } \Omega_h, \quad v_h = 0 \quad \text{auf } \partial\Omega_h, \\ L_h(v_h + w) = 0 \quad &\text{in } \Omega_h, \quad v_h + w \le \tfrac{1}{4} d_\Omega^2 \quad \text{auf } \partial\Omega_h. \end{aligned}$$

Aus dem diskreten Maximumprinzip folgt dann wegen $w \ge 0$ notwendig

$$\max_{\overline{\Omega}_h} v_h \le \max_{\overline{\Omega}_h} (v_h + w) \le \max_{\partial\Omega_h} (v_h + w) = \max_{\partial\Omega_h} w \le \tfrac{1}{4} d_\Omega^2,$$

und somit

$$h^2 \sum_{Q \in \Omega_h} G_h(P, Q) \le \tfrac{1}{4} d_\Omega^2, \quad P \in \overline{\Omega}_h. \tag{2.2.35}$$

Die Abschätzungen (2.2.34) und (2.2.35) ergeben zusammen mit (2.2.33) die Behauptung. Q.E.D.

Als unmittelbare Folgerung aus Satz 2.2 erhalten wir (in Analogie zum kontinuierlichen Fall) die stetige Abhängigkeit der Lösung von den Daten. Das wichtigste Resultat ist eine a priori Konvergenzabschätzung für das Differenzenschema (2.2.29).

Korollar 2.3 (Allgemeines Konvergenzresultat): *Unter den Voraussetungen von Satz 2.2 gilt für das Differenzenschema (2.2.29) die a priori Konvergenzabschätzung*

$$\max_{P \in \overline{\Omega}_h} |e_h(P)| \le \tfrac{1}{4} d_\Omega^2 \max_{P \in \Omega_h} |\tau_h(P)| + \max_{P \in \partial\Omega_h} |e_h(P)|, \tag{2.2.36}$$

mit dem Fehler $e_h = u - u_h$ *und dem Abschneidefehler* $\tau_h(P) := L_h u(P) + \Delta u(P)$.

Beweis: Für den Fehler e_h gilt die Differenzengleichung

$$L_h e_h(P) = \tau_h(P) \quad P \in \Omega_h,$$

so dass die Stabilitätsungleichung (2.2.30) unmittelbar die Behauptung liefert. Q.E.D.

Die Abschätzung (2.2.36) garantiert wieder, dass für ein „gutes“ Differenzenschema der globale Diskretisierungsfehler mit derselben Ordnung wie der lokale Abschneidefehler konvergiert. Wir wollen diese allgemeinen Resultate für die obigen speziellen Diskretisierungen

anwenden. Dabei wird wieder die folgende Bezeichnung für Normschranken der exakten Lösung verwendet:

$$M_m(u) := \max_{\overline{\Omega}}\{|\partial_x^i \partial_y^j u|,\ i+j=m\}.$$

Korollar 2.4 (Konvergenz des 5-Punkte-Operators): *Unter den Voraussetungen von Satz 2.2 gilt für den Fehler den 5-Punkte-Operator mit konstanter Randwertextrapolation die a priori Fehlerabschätzung*

$$\max_{P\in\overline{\Omega}_h} |e_h(P)| \le \tfrac{1}{24} d_\Omega^2 M_4(u) h^2 + M_1(u) h. \tag{2.2.37}$$

Beweis: Für den Abschneidefehler gilt

$$\max_{\Omega_h} |\tau_h| \le \tfrac{1}{6} M_4(u) h^2,$$

und in Randgitterpunkten

$$\max_{\partial\Omega_h} |e_h| = \max_{\partial\Omega_h} |u(P) - u(P^\star)| \le M_1(u) h.$$

Die Stabilitätsabschätzung (2.2.30) impliziert damit die Behauptung. Q.E.D.

Der ordnungsreduzierende Effekt der mangelhaften Randwertapproximation kann durch die oben beschriebene „lineare Randwertinterpolation“ behoben werden. Für das so modifizierte 5-Punkte-Schema erhält man ähnlich wie eben die a priori Fehlerabschätzung

$$\max_{P\in\overline{\Omega}_h} |e_h(P)| \le \tfrac{1}{12} d_\Omega^2 M_4(u) h^2 + \tfrac{1}{2} M_2(u) h^2. \tag{2.2.38}$$

Korollar 2.5 (Konvergenz des Shortley-Weller-Operators): *Unter den Voraussetungen von Satz 2.2 gilt für den modifizierten 5-Punkte-Operator nach Shortley-Weller die a priori Fehlerabschätzung*

$$\max_{P\in\overline{\Omega}_h} |e_h(P)| \le \tfrac{1}{24} d_\Omega^2 M_4(u) h^2 + \tfrac{1}{3} M_3(u) h^3. \tag{2.2.39}$$

Beweis: Für den Fehler e_h gelten die Beziehungen

$$-\Delta_h e_h = \tau_h \ \text{ in } \ \Omega_h, \quad -\Delta_h^* e_h = \tau_h^* \ \text{ auf } \ \partial\Omega_h^*, \quad e_h = 0 \ \text{ auf } \ \partial\Omega_h.$$

Für die Abschneidefehler gilt

$$\max_{\Omega_h} |\tau_h| \le \tfrac{1}{6} M_4(u) h^2, \qquad \max_{\partial\Omega_h^*} |\tau_h^*| \le \tfrac{2}{3} M_3(u) h.$$

In diesem Fall können wir nicht direkt die Stabilitätsungleichung (2.2.30) anwenden, da sie nur auf eine reduzierte Konvergenzordnung $O(h)$ führen würde. Statt dessen modifizieren wir in geeigneter Weise den Beweis dieser Abschätzung. Mit den Bezeichnungen

des Beweises von Satz 2.2 ergibt die Greensche Identität ($e_h = 0$ auf $\partial\Omega_h$)

$$e_h(P) = h^2 \sum_{Q\in\Omega_h^0} G_h(P,Q)\tau_h(Q) + h^2 \sum_{Q\in\partial\Omega_h^*} G_h(P,Q)\tau_h^*(Q)$$

und folglich

$$|e_h(P)| \le h^2 \sum_{Q\in\Omega_h^0} G_h(P,Q) \max_{\Omega_h^0} |\tau_h| + h^2 \sum_{Q\in\partial\Omega_h^*} G_h(P,Q) \max_{\partial\Omega_h^*} |\tau_h^*| \,. \tag{2.2.40}$$

Wie im Beweis von Satz 2.2 folgt

$$h^2 \sum_{Q\in\Omega_h^0} G_h(P,Q) \le h^2 \sum_{Q\in\Omega_h} G_h(P,Q) \le \tfrac{1}{4} d_\Omega^2 \,. \tag{2.2.41}$$

Wir wollen jetzt zeigen, dass

$$\sum_{Q\in\partial\Omega_h^*} G_h(P,Q) \le \tfrac{1}{2} \,. \tag{2.2.42}$$

Dazu definieren wir die Gitterfunktion w_h durch

$$w_h = 1 \ \text{ in } \Omega_h, \quad w_h = 0 \ \text{ auf } \partial\Omega_h \,.$$

Für diese verifiziert man durch Nachrechnen

$$-\Delta_h w_h = 0 \ \text{ in } \Omega_h^0, \quad -\Delta_h^* w_h \ge 2h^{-2} \ \text{ auf } \partial\Omega_h^* \,.$$

Mit Hilfe der Greenschen Identität folgt damit

$$1 = h^2 \sum_{Q\in\partial\Omega_h^*} G_h(P,Q)(-\Delta_h^*) w_h(Q) \ge 2 \sum_{Q\in\partial\Omega_h^*} G_h(P,Q),$$

was (2.2.42) impliziert. Dies vervollständigt den Beweis. Q.E.D.

Bemerkung: In drei Raumdimensionen gelten Satz 2.2 und Korollar 2.3 mit der Konstante $\frac{1}{6} d_\Omega^2$. Hiermit und der zugehörigen Abschätzung (2.1.17) für den Konsistenzfehler ergeben sich dann auch die a priori Fehlerabschätzungen der Korollare 2.4 und 2.4 mit der führenden Konstante $\frac{1}{24} d_\Omega^2$.

Bemerkung: Wir betonen, dass die Konvergenz der betrachteten Differenzenschemata eine vergleichsweise hohe Regularität der zu approximierenden Lösung erfordert. Das Shortley-Weller-Schema erfordert z. B. für seine „maximale“ Konvergenzordnung $m = 2$, dass u beschränkte vierte Ableitungen auf ganz $\overline{\Omega}$ besitzt ($M_4(u) < \infty$). Dies ist eine sehr einschneidende Forderung, wie wir im vorigen Kapitel gesehen haben. Sie kann i. Allg. nur für glatt berandete Gebiete sowie unter gewissen Zusatzbedingungen für spezielle Geometrien wie z. B. Rechtecke garantiert werden. Weiter erfordert sie auch hohe Glattheit der Daten f und g . Auf Gebieten mit *einspringenden* Ecken oder sonstwie

reduzierter Regularität von u können diese Sätze nicht verwendet werden. In solchen realitätsnäheren Situationen werden sich die im nächsten Abschnitt behandelten „Finite-Elemente-Verfahren“ als flexibler erweisen.

2.3 Lösungsaspekte

Wir diskutieren nun die Lösung der durch eine Differenzendiskretisierung entstehenden algebraischen Gleichungssysteme. Zugrunde gelegt wird wieder die Situation des Modellproblems der 1. RWA des Laplace-Operators

$$Lu := -\Delta u = f \quad \text{in } \Omega, \quad u = g \text{ auf } \partial\Omega, \tag{2.3.43}$$

auf einem Gebiet $\Omega \subset \mathbf{R}^2$. Erweiterungen für Probleme mit variablen Koeffizienten, anderen Randbedingungen, Unsymmetrien sowie auf drei Raumdimensionen werden wieder in Bemerkungen berücksichtigt. Die zugehörigen algebraischen Systeme haben die Gestalt

$$Ax = b, \tag{2.3.44}$$

Die Matrix $A = (a_{nm})_{n,m=1}^N$ und der Vektor $b = (b_n)_{n=1}^N$ haben die Dimension $N :=$ *Anzahl der inneren Gitterpunkte bzw. Knotenfreiheitsgrade.* In der Praxis ist meist $N \gg 1000$, so dass neben dem Rechenaufwand auch der Speicherbedarf ein wichtiger Aspekt ist. Die Matrix ist extrem dünn besetzt und besitzt abhängig von der gewählten Numerierung der Gitterpunkte bzw. Knoten eine Bandstruktur. Für die Wahl eines geeigneten (d. h.: möglichst sparsamen) Lösungsverfahrens ist die zu erwartende Dimension N entscheidend.

Wir orientieren die folgende Diskussion an der Modellsituation der Poisson-Gleichung auf dem Einheitsquadrat $\Omega = (0,1)^2$ und der Diskretisierung mit dem üblichen 5-Punkte-Schema auf einem äquidistanten, kartesischen Gitter $\overline{\Omega}_h = \Omega_h \cup \partial\Omega_h$ der Gitterweite h. Das Gebiet $\Omega = (0,1)^2$ hat den Durchmesser $d_\Omega = \sqrt{2}$. Wir wählen die Funktion $u(x,y) = \sin(\pi x)\sin(\pi y)$ als Lösung der Randwertaufgabe

$$-\Delta u = 2\pi^2 u =: f \quad \text{in } \Omega, \qquad u = 0 \quad \text{auf } \partial\Omega. \tag{2.3.45}$$

Es ist $M_4(u) = \pi^4$. Die a priori Fehlerschranke (2.2.39) für die Shortley-Weller-Approximation liefert damit die (pessimistische) Fehlerschätzung

$$\max_{\overline{\Omega}_h} |u - u_h| \approx \tfrac{1}{24} d_\Omega^2 \pi^4 h^2 \approx 8h^2. \tag{2.3.46}$$

Zur Erreichung garantierter 3-stelliger (relativer) Genauigkeit, $\mathrm{TOL} = 10^{-3}$, ist in diesem Fall also eine Gitterweite $h \approx 10^{-2}$ erforderlich. Dies führt auf ein (symmetrisches) Gleichungssystem der Dimension $N \approx 10^4$.

Bemerkung 2.1: Die Fehlerabschätzung (2.3.46) ist in der Tat sehr pessimistisch. Der wirkliche Fehler ist ungefähr um einen Faktor 10^{-1} kleiner. Dies zeigt, dass unsere *a*

priori Fehleranalyse selbst für ein so einfaches Modellproblem zu unscharf und für praktische Zwecke nur bedingt brauchbar ist.

Bemerkung 2.2: In drei Raumdimensionen ist $d_\Omega = \sqrt{3}$, und die a priori Fehlerabschätzung (2.3.46) liefert den Wert $\approx 12h^2$, so dass hier mindestens auch $h = 10^{-2}$ und damit $N \approx 10^6$ erforderlich wäre.

Die Struktur der Matrix A hängt von der gewählten Nummerierung der Gitterpunkte ab. Die gängigen Alternativen sind:

1) *Zeilenweise Nummerierung:* Die lexikographische Anordnung der Gitterpunkte,

$$(x_i, y_j) \leq (x_p, y_q) \quad \text{wenn} \quad j \leq q \ \text{ oder } \ j = q,\ i \leq p$$

führt auf eine Bandmatrix mit Bandbreite $2m + 1 \approx h^{-1}$.

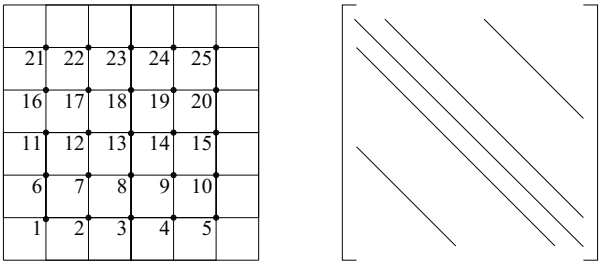


Abbildung 2.6: Lexikographische Nummerierung

2) *Diagonale Nummerierung:* Die sukzessive Numerierung diagonal zu den Koordinatenrichtungen führt auf eine Bandmatrix mit geringem „Bandinhalt“ ($\Rightarrow$ Speicherersparnis beim Gaußschen Eliminationsverfahren).

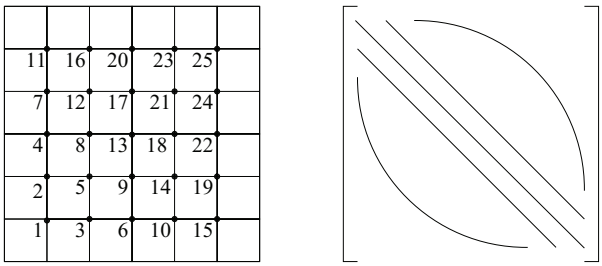


Abbildung 2.7: Diagonale Nummerierung

3) *Schachbrett-Nummerierung:* Die versetzte zeilenweise- sowie spaltenweise Nummerierung führt auf eine 2×2-Blockmatrix mit diagonalen Hauptblöcken.

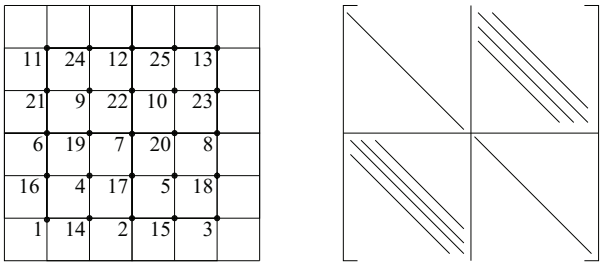


Abbildung 2.8: Schachbrett-Nummerierung

a) Direkte Lösung: Prinzipiell wäre das klassische Gaußsche Eliminationsverfahren zur Lösung des Systems (2.3.44) geeignet. Zu seiner Durchführung benötigt man bei zeilenweiser Nummerierung und Ausnutzung der Symmetrie etwa mN Speicherplätze (im Folgenden abgekürzt als „SP“) und m^2N arithmetische Operationen (im Folgenden abgekürzt als „OP“). Im Falle $N \approx 10^4$ sind dies etwa 10^6 SP und 10^8 OP. Die kurze Überschlagsrechnung zeigt, dass direkte Lösungsmethoden für das Gleichungssystem (2.3.44) nur in speziellen Situationen in Frage kommen. Sie spielen eine Rolle bei geringer Problemgröße ($N \leq 10^3$) oder im Falle sehr uniformer Matrixeinträge (bei der Approximation von Differentialoperatoren mit konstanten Koeffizienten) ($N \leq 10^5$). Speziell für die 5-Punkte-Approximation des Laplace-Operators auf Rechtecken gibt es sehr effiziente direkte Lösungstechniken auf Grundlage der sog. „schnellen Fourier-Transformation (FFT)“ ($N \leq 10^7$). Diese sind aber bei allgemeineren Problemstellungen meist nicht anwendbar und werden daher hier nicht weiter diskutiert.

Bemerkung 2.3: Zur Illustration machen wir dieselbe Überschlagsrechnung auch für die entsprechende dreidimensionale Situation $\Omega = (0,1)^3$. Hier hat die Matrix A die Dimension $N = m^3 \approx 10^6$ und die Bandbreite $2m^2 + 1 \approx 10^4$. Folglich betrüge der benötigte Lösungsaufwand $\approx 10^{10}$ SP und $\approx 10^{14}$ OP, was noch jenseits der Kapazität von Arbeitsplatzrechnern liegt.

b) Iterative Lösung: Wie bereits erwähnt, übertragen sich die wichtigen Eigenschaften (B1), (B2) und (B3) des Differenzenschemas unabangig von der gewählten Numerierung auf die jeweilige Matrix A. Diese ist (irreduzibel) diagonaldominant und von nichtnegativem Typ und folglich eine M-Matrix, d. h.: Es gilt elementweise $A^{-1} \geq 0$. Zusätzlich ist A oft auch symmetrisch und positiv definit. In diesem Falle konvergieren die meisten gängigen iterativen Verfahren. Ausgehend von einem Defektkorrekturansatz

$$d^t = F - Ax^t, \quad Cv^t = d^t, \quad x^{t+1} = x^t + v^t, \tag{2.3.47}$$

mit einer regulären Matrix C („Vorkonditionierer“) wird die Iteration zunächst als Fixpunktiteration geschrieben

$$Cx^{t+1} = Cx^t + b - Ax^t \quad \Leftrightarrow \quad x^{t+1} = (I - C^{-1}A)x^t + C^{-1}b. \tag{2.3.48}$$

Für den Fehler $e^{(t)} := x^{(t)} - x$ gilt

$$\begin{aligned} e^{(t)} &= (I - C^{-1}A)x^{(t-1)} + C^{-1}b - x \\ &= (I - C^{-1}A)x^{(t-1)} + C^{-1}b - (I - C^{-1}A)x - C^{-1}b \\ &= (I - C^{-1}A)e^{(t-1)} = \cdots = (I - C^{-1}A)^t e^{(0)}. \end{aligned} \tag{2.3.49}$$

Dies zeigt, dass die Iteration konvergiert, wenn die „Iterationsmatrix“ $B := I - C^{-1}A$ eine Kontraktion ist, d. h.: $\rho(B) := \max\{|\lambda|,\ \lambda \text{ Eigenwert von } B\} < 1$. Die „Konvergenzrate“ der Iteration ist dann gegeben durch

$$\rho := \sup_{x^{(0)} \in \mathbf{R}^N} \lim_{t\to\infty} \Big(\frac{\|x^{(t)} - x\|}{\|x^{(0)} - x\|} \Big)^{1/t} = \rho(B). \tag{2.3.50}$$

Es ist dann näherungsweise

$$\|x^{(t)} - x\| \le \rho^t \|x^{(0)} - x\|. \tag{2.3.51}$$

Die Anzahl der zur Erreichung einer Reduktion des Anfangsfehlers um ε erforderlichen Iterationsschritte ergibt sich damit zu

$$t_\varepsilon \approx \frac{\ln(\varepsilon)}{\ln(\rho)}. \tag{2.3.52}$$

Der „Vorkonditionierer“ C sollte folgende Bedingungen erfüllen:

- einfache Invertierung (mit $O(N)$ OP und Speicherbedarf $O(N)$ SP);
- möglichst kleines $\rho(I - C^{-1}A)$.

Dies sind leider gegenläufige Zielsetzungen. Die einfachsten Verfahren dieses Typs verwenden ausgehend von der natürlichen Aufspaltung $A = L + D + R$ die Vorkonditionierer:

1. *(Gedämpftes) Richardson*[4]*-Verfahren:* $0 < \theta \le 2\lambda_{max}(A)^{-1}$

$$C = \theta I, \qquad B = I - \theta A. \tag{2.3.53}$$

[4]Lewis Fry Richardson (1881–1953): Englischer Mathematiker und Physiker; wirkte an verschiedenen Institutionen in England und Schottland; typischer “angewandter Mathematiker”; leistete Pionierbeiträge zur Modellierung und Numerik in der Wettervorhersage.

2. *Jacobi* [5]*-Verfahren (Gesamtschrittverfahren):*

$$C = D, \qquad B = -D^{-1}(L + R). \tag{2.3.54}$$

3. *Gauß-Seidel* [6]*-Verfahren (Einzelschrittverfahren):*

$$C = D + L, \qquad B = -(D + L)^{-1}R. \tag{2.3.55}$$

4. *SOR-Verfahren („Successive Over-Relaxation"):* $\omega = \omega_{opt} \in (0, 2)$

$$C = D + \omega L, \qquad B = (D + \omega L)^{-1}\{(1 - \omega)D - \omega R\}. \tag{2.3.56}$$

5. *ILU-Verfahren („Incomplete LU Decomposition"):*

$$C = \tilde{L}\tilde{R}, \qquad B = I - \tilde{R}^{-1}\tilde{L}^{-1}A. \tag{2.3.57}$$

Für eine symmetrische, positiv definite Matrix wird das ILU- zum ILL^T-Verfahren („Incomplete Cholesky[7] Decomposition"). Die ILU-Zerlegung erhält man mit Hilfe des üblichen, rekursiven Prozesses zur Bestimmung der LU-Zerlegung aus der Gleichung $LU = A$ durch Nullsetzung aller Matrixeinträge zu Indexpaaren $\{n, m\}$ mit $a_{nm} = 0$:

$$\begin{aligned} n = 1, ..., N: \quad & \tilde{u}_{nm} = a_{nm} - \sum_{i=1}^{n-1} \tilde{l}_{ni}\tilde{u}_{im} \quad (m = 1, ..., N) \\ & \tilde{l}_{nn} = 1, \quad \tilde{l}_{in} = \tilde{u}_{nn}^{-1}\Big\{a_{in} - \sum_{j=1}^{n-1} \tilde{l}_{ij}\tilde{u}_{jn}\Big\} \quad (i = n + 1, ..., N) \\ & \tilde{l}_{nm} = 0, \ \tilde{u}_{nm} = 0, \ \text{wenn } a_{nm} = 0\,. \end{aligned}$$

6. *ADI-Verfahren („Alternating-Direction Implicit Iteration"):*

$$\begin{aligned} C &= (A_x + \omega I)(A_y + \omega I), \\ B &= (A_y + \omega I)^{-1}(\omega I - A_x)(A_x + \omega I)^{-1}(\omega I - A_y). \end{aligned} \tag{2.3.58}$$

Das ADI-Verfahren ist für beliebige Wahl des Parameters $\omega > 0$ konvergent. Wenn die Struktur der Matrix seine Anwendung zulässt, ist es bei optimaler Wahl von ω mindestens so schnell wie das optimale SOR-Verfahren.

[5]Carl Gustav Jakob Jacobi (1804-1851): Deutscher Mathematiker; schon als Kind. h.chbegabt; wirkte in Königsberg und Berlin; Beiträge zu vielen Bereichen der Mathematik: zur Zahlentheorie, zu elliptischer Funktionen, zu partiellen Differentialgleichungen, zu Funktionaldeterminanten und zur theoretischen Mechanik.

[6]Philipp Ludwig von Seidel (1821–1896): Deutscher Mathematiker; Prof. in München; Beiträge zur Analysis (u. a. Methode der kleinsten Fehlerquadrate) owie Himmelsmechanik und Astronomie.

[7]Andrè Louis Cholesky (1975–1918): Französischer Mathematiker; Militärkarriere; Beiträge zur Numerischen Linearen Algebra.

Die Konvergenzraten dieser einfachen Iterationsverfahren verhalten sich in Abhängigkeit von der (gleichförmigen) Gitterweite wie

$$\rho = \rho(I - C^{-1}A) = 1 - \mathcal{O}(h^r),$$

in Abhängigkeit von der Gitterweite h der Diskretisierung (bei fester Problemkonfiguration) mit einem geeigneten $r \geq 0$. Die Anzahl T der zur Gewinnung einer Dezimalstelle Genauigeit erforderlichen Iterationsschritte ist also ungefähr bestimmt durch

$$\rho^T \approx 10^{-1} \quad \Rightarrow \quad \mathrm{T} \approx -\frac{\ln(10)}{\ln(\rho)} \approx h^{-r}. \tag{2.3.59}$$

Hierzu beachte man, dass $\ln(1 - ch^r) = -ch^r + O(h^{2r})$. Da die Durchführung eines Iterationsschritts approximativ $N \approx h^{-2}$ OP (in zwei Raumdimensionen) kostet, ergibt sich ein Gesamtaufwand pro Dezimalstelle an Genauigkeit von $\approx h^{-2-r}$ OP. Der für die Durchführung dieser Iterationsverfahren benötigte Speicherplatz entspricht etwa dem zur Speicherung der wesentlichen (d. h. von Null verschiedenen) Elemente der Matrix A erforderlichen. Für das Jacobi- und Gauß-Seidel-Verfahren ist $r = 2$ und für das (optimale) SOR-Verfahren ist $r = 1$. Das ILU- und das ADI-Verfahren liegen bei speziellen Konfigu. a.ionen etwa gleich auf zum SOR-Verfahren. Damit ergibt sich ein Gesamtlösungsaufwand von jeweils $\mathcal{O}(N^2)$ bzw. $\mathcal{O}(N^{3/2})$ OP für die einzelnen Verfahren. Ein wirklich „effizientes“ Verfahren sollte ein möglichst kleines r aufweisen; optimal wäre $r = 0$. Dies lässt sich durch den Einsatz von sog. „Multi-Level-Techniken“ („Mehrgitterverfahren“) erreichen, welche später im Zusammenhang mit den Finite-Elemente-Diskretisierungen diskutiert werden.

2.3.1 Aufwandsanalyse: ein Beispiel

Wir wollen das Konvergenzverhalten der bisher betrachteten Iterationsverfahren und deren sich daraus ergebende Effizienz anhand des obigen Modellproblems eingehender diskutieren. Dabei soll insbesondere die Bedeutung der Formel (2.3.59) für die Konvergenzrate illustriert werden.

Für die obige Modellsituation (5-Punkte-Differenzenoperator auf einem äquidistanden, kartesischen Gitter des Einheitsquadrats) lassen sich die Eigenwerte und zugehörigen Eigenvektoren der Systemmatrix A explizit angeben. Für $k, l = 1, \ldots, m$ ergibt sich mit der Bezeichnung $Aw^{kl} = \lambda_{kl} w^{kl}$:

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

Also ist (für $h \ll 1$)

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

und somit

$$\kappa := \mathrm{cond}_2(A) \approx \frac{4}{\pi^2 h^2}\,. \tag{2.3.60}$$

Die Eigenwerte der Jacobi-Matrix $J = -D^{-1}(L+R)$ sind

$$\mu_{kl} = \tfrac{1}{2}\left(\cos(kh\pi) + \cos(lh\pi)\right) \quad (k,l = 1,\ldots,m)$$

Folglich wird

$$\rho_J = \mu_{\max} = \cos(h\pi) = 1 - \frac{\pi^2}{2}\,h^2 + O(h^4)\,. \tag{2.3.61}$$

Für die Iterationsmatrizen des Gauß-Seidel- und des (optimalen) SOR-Verfahrens folgt:

$$\begin{aligned}
\rho_{GS} = \rho_{GS}^2 &= 1 - \pi^2 h^2 + O(h^4)\,,\\
\rho_{SOR} = \frac{1-\sqrt{1-\rho_J^2}}{1+\sqrt{1-\rho_J^2}} &= \frac{1-\pi h + O(h^2)}{1+\pi h + O(h^2)} = 1 - 2\,\pi h + O(h^2)\,.
\end{aligned}$$

Für die Iterationszahlen (pro Dezimalstelle Fehlerreduktion) ergibt sich also asymptotisch:

$$\begin{aligned}
\mathrm{T}_J &\approx -\frac{\ln(10)}{\ln(1-\frac{\pi^2}{2}h^2)} \approx \frac{4,6}{\pi^2 h^2} \approx \tfrac{1}{2}\,N\,,\\
\mathrm{T}_{GS} &\approx -\frac{\ln(10)}{\ln(1-\pi^2 h^2)} \approx \frac{2,3}{\pi^2 h^2} \approx \tfrac{1}{4}\,N\,,\\
\mathrm{T}_{SOR} &\approx -\frac{\ln(10)}{\ln(1-2\pi h)} \approx \frac{2,3}{2\pi h} \approx \tfrac{1}{3}\sqrt{N}\,.
\end{aligned}$$

Der Vollständigkeit halber geben wir hier auch die entsprechenden Werte für das CG-Verfahren („Verfahren der konjugierten Richtungen“):

$$\mathrm{T}_{CG} = \tfrac{1}{2}\sqrt{\kappa}\,\ln(20) \approx \frac{3}{\pi h} \approx \sqrt{N}\,.$$

Das CG-Verfahren ist zwar langsamer als das „optimale“ SOR-Verfahren, erfordert aber nicht die Bestimmung eines Iterationsparameters.

Zum Vergleich der Effizienz der Iterationsverfahren muss natürlich auch der Aufwand pro Iterationsschritt berücksichtigt werden. Für die Anzahl OP der arithmetischen Operationen pro Iterationsschritt gilt (optimistische Schätzung)

$$\mathrm{OP_J},\ \mathrm{OP_{GS}},\ \mathrm{OP_{SOR}} \approx 6\,\mathrm{N}\,, \qquad \mathrm{OP_{CG}} \approx 10\,\mathrm{N}\,.$$

Als Endresultat finden wir, dass zur Bestimmung der Lösung des diskretisierten Modellproblems das Jacobi-Verfahren, das Gauß-Seidel-Verfahren und das Gradientenverfahren $O(N^2)$ OP benötigen. Das direkte Cholesky-Verfahren würde für die Berechnung der „exakten“ Lösung (auf Rundungsfehlergenauigkeit) ebenfalls $O(m^2 N) = O(N^2)$ OP benöti-

gen, erscheint also dem Gauß-Seidel-Verfahren überlegen zu sein. Es ist jedoch zu berücksichtigen, dass letzteres nur $O(N)$ SP benötigt, im Gegensatz zu den $O(mN) = O(N^{3/2})$ SP für das Cholesky-Verfahren. Die schnelleren, auf Mehrgitterkonzepten basierenden Iterationsverfahren sind dagegen dem einfachen direkten Löser asymptotisch klar überlegen.

Für das Beispiel mit $N = 10^4$ ergibt sich überschlagsmäßig der folgende Gesamtaufwand „GA“ zur sicheren Lösung des Systems (2.3.44) unter die Diskretisierungsgenauigkeit ($\mathrm{TOL} = 10^{-3}$, $h = 10^{-2}$, $N = 10^4$):

$$\begin{aligned} \mathrm{GA}_J(\mathrm{TOL}) &\approx 4 \cdot 3N^2 \approx 1,2 \cdot 10^9 \,\mathrm{OP}\,, \\ \mathrm{GA}_{GS}(\mathrm{TOL}) &\approx 4 \cdot 1,5N^2 \approx 6 \cdot 10^8 \,\mathrm{OP}\,, \\ \mathrm{GA}_{SOR}(\mathrm{TOL}) &\approx 4 \cdot 2N^{3/2} \approx 8 \cdot 10^6 \,\mathrm{OP}\,, \\ \mathrm{GA}_{CG}(\mathrm{TOL}) &\approx 4 \cdot 10N^{3/2} \approx 4 \cdot 10^7 \,\mathrm{OP}\,. \end{aligned}$$

Für ein „optimales“ Verfahren wie z. B. das Mehrgitterverfahren „MG“ würde man hier wesentlich bessere Werte erwarten: $\mathrm{GA}_{MG}(\mathrm{TOL}) \approx 4 \cdot 25N \approx 10^6 \,\mathrm{OP}$.

Bemerkung: In drei Raumdimensionen erhalten wir näherungsweise

$$\lambda_{max} \approx 12h^{-2}, \quad \lambda_{min} \approx 3\pi^2, \quad \kappa \approx \frac{8}{3\pi^2 h^2},$$

und folglich im wesentlichen dieselben Abschätzungen für ρ_J, ρ_{GS} und ρ_{SOR} sowie für die Iterationszahlen T_J, T_{GS}, T_{SOR} und T_{CG} wie im zweidimensionalen Fall. Der Aufwand für die Durchführung eines Iterationsschritts ist hier $\mathrm{OP_J}$, $\mathrm{OP_{GS}}$, $\mathrm{OP_{SOR}} \approx 8\,\mathrm{N}$ bzw. $\mathrm{OP_{CG}} \approx 12\,\mathrm{N}$. Für den Gesamtlösungsaufwand ergibt dies dann:

$$\begin{aligned} \mathrm{GA}_J(\mathrm{TOL}) &\approx 4 \cdot 4N^2 \approx 1,6 \cdot 10^{13} \,\mathrm{OP}\,, \\ \mathrm{GA}_{GS}(\mathrm{TOL}) &\approx 4 \cdot 2N^2 \approx 8 \cdot 10^{12} \,\mathrm{OP}\,, \\ \mathrm{GA}_{SOR}(\mathrm{TOL}) &\approx 4 \cdot 3N^{3/2} \approx 1,2 \cdot 10^{10} \,\mathrm{OP}\,, \\ \mathrm{GA}_{CG}(\mathrm{TOL}) &\approx 4 \cdot 12N^{3/2} \approx 4,8 \cdot 10^{10} \,\mathrm{OP}\,. \end{aligned}$$

Der Aufwand des Mehrgitterverfahrens erhöht sich aber nur vergleichsweise unwesentlich auf $\mathrm{GA}_{MG}(\varepsilon) \approx 4 \cdot 50N \approx 2 \cdot 10^8 \,\mathrm{OP}$. Zur Bewertung dieser Komplexitätsschätzungen muss man sich die Leistung verfügbarer Rechner vergegenwärtigen. Ein Arbeitsplatzrechner leiste real etwa 200 MFlops (200 Millionen „Floating-Point“ Operationen pro Sekunde). Das SOR-Verfahren braucht dann auf einem solchen Rechner zur Lösung des Gleichungssystems (auf Diskretisierungsfehlergenauigkeit) etwa $1,5$ Minuten, wogegen das Mehrgitterverfahren „nur“ 1 Sekunde benötigt.

2.4 Übungen

Übung 2.1: Man betrachte die Diskretisierung der 1. RWA des Laplace-Operators auf dem Einheitsquadrat des $\mathbb{R}^2$ mit dem 9-Punkte-Differenzenschema (sog. „Mehrstellenfor-

mel“)

$$-\Delta_h^{(9)} u_h(x,y) = f(x,y), \quad (x,y) \in \Omega_h,$$

mit dem „gestreckten“ Differenzenoperator

$$\Delta_h^{(9)} u(x,y) := \frac{1}{12h^2}\Big\{-u(x\pm 2h,y)+16u(x\pm h,y)-u(x,y\pm 2h)+16u(x,y\pm h)-60u(x,y)\Big\}$$

und dem 9-Punkte-Differenzenschema

$$-\Delta_h^{(9)} u_h(x,y) = f(x,y) + \tfrac{1}{12}h^2 \Delta f(x,y), \quad (x,y) \in \Omega_h$$

mit dem „kompakten“ Differenzenoperator

$$\bar{\Delta}_h^{(9)} u(x,y) = \frac{1}{6h^2}\Big\{4u(x\pm h,y) + 4u(x,y\pm h) + u(x\pm h, y\pm h) - 20u(x,y)\Big\}.$$

Man zeige, dass dies Approximationen mit der Konsistenzordnung $m = 4$ sind.

Übung 2.2: In vielen Fällen kann die asymptotische Konvergenzordnung eines Differenzenverfahrens nur experimentell bestimmt werden. Dazu werden bei bekannter exakter Lösung für zwei Schrittweiten h und $h/2$ die Fehler $e_h := u - u_h$ und $e_{h/2} := u - u_{h/2}$ berechnet und dann die Ordnung α über den formalen Ansatz $\|u - u_h\|_h = h^\alpha$ mit einer geeigneten Gitternorm $\|\cdot\|_h$ aus der folgenden Formel ermittelt:

$$\alpha = \frac{\log(\|e_h\|_h / \|e_{h/2}\|_h)}{\log(2)}.$$

a) Man rechtfertige diese Formel und überlege, wie man vorgehen kann, wenn keine exakte Lösung u bekannt ist.

b) Man bestimme die inhärente Konvergenzordnungen der folgenden Zahlenfolgen:

$h = 2^{-1}$	33.627	26.570
$h = 2^{-2}$	30.318	27.008
$h = 2^{-3}$	29.100	27.883
$h = 2^{-4}$	28.586	28.072
$h = 2^{-5}$	28.351	28.117

Übung 2.3: Man betrachte die Diskretisierung der 1. RWA des Laplace-Operators auf dem Einheitsquadrat des $\mathbb{R}^2$ mit dem 9-Punkte-Differenzenschema

$$\Delta_h^{(9)} u_h(x,y) = f_h(x,y) := f(x,y) + \tfrac{1}{12}h^2 \Delta f(x,y), \quad (x,y) \in \Omega_h,$$

mit dem „kompakten“ Differenzenoperator

$$\Delta_h^{(9)} u(x,y) = \frac{1}{6h^2}\Big\{4u(x\pm h,y) + 4u(x,y\pm h) + u(x\pm h, y\pm h) - 20u(x,y)\Big\}.$$

Diese Approximationen hat nach Aufgabe 2.1 die Konsistenzordnung $m = 4$.

a) Man zeige mit den Mitteln der Vorlesung die Fehlerabschätzung

$$\max_{P\in\overline{\Omega}_h} |u(P) - u_h(P)| \le cM_6(u)h^4.$$

b) (Zusatzaufgabe für Leser mit Ergeiz und Zeit) Im Falle eines allgemeinen, glattberandeten Gebiets $\Omega \subset \mathbb{R}^2$ werde entlang der gekrümmten Randabschnitte die Shortley-Weller-Approximation betrachtet:

$$-\Delta_h^{(9)} u_h = f + \tfrac{1}{12}h^2\Delta f \ \text{ in } \Omega_h, \quad -\Delta_h^* u_h = f \ \text{ in } \Omega_h^*, \quad u_h = g \ \text{ auf } \partial\Omega_h.$$

Man zeige hierfür mit den Mitteln des Textes die Fehlerabschätzung

$$\max_{P\in\overline{\Omega}_h} |u(P) - u_h(P)| \le c\,\{M_6(u)h^4 + M_3(u)h^3\}.$$

Übung 2.4: Sei A_h die zum 5-Punkte-Operator auf dem Einheitsquadrat gehörende $N \times N$-Matrix (bei zeilenweiser Nummerierung der Gitterpunkte mit m Punkten in jeder Zeile). Die $N = m^2$ Eigenvektoren $w^{\nu\mu}$, $\nu, \mu = 1, \ldots, m$, und die zugehörigen Eigenwerte $\lambda_{\nu\mu}$ von A_h sind gegeben durch:

$$w^{\nu\mu}(x, y) = \sin(\nu\pi x)\sin(\mu\pi y), \ \ (x, y) \in \Omega_h, \quad \lambda_{\nu\mu} = \frac{1}{h^2}\big(4 - 2(\cos(\nu h\pi) + \cos(\mu h\pi)\big).$$

Man zeige, dass für die Spektralkondition von A_h gilt:

$$\mathrm{cond}_2(A_h) =: \frac{\lambda_{\max}(A_h)}{\lambda_{\min}(A_h)} = \frac{4}{\pi^2 h^2} + \mathcal{O}(1).$$

Übung 2.5: Eine Matrix $A \in \mathbb{R}^{N\times N}$ heißt „M-Matrix“, wenn sie von nichtnegativem Typ und regulär ist und wenn ihre Inverse $A^{-1} = (a_{ij}^{(-1)})_{i,j=1}^N$ elementweise nichtnegativ ist: $a_{ij}^{(-1)} \ge 0$. Das 5-Punkte-Differenzenschema (bzw. das Shortley-Weller-Schema) zur Approximation der 1. RWA des Laplace-Operators führt z. B. auf eine solche M-Matrix.

a) Man zeige, dass M-Matrizen „invers-monoton“ sind, d. h.: Für Vektoren $v, w \in \mathbb{R}^N$ gilt komponentenweise:

$$Av \ge Aw \quad \Rightarrow \quad v \ge w.$$

b) Ist ferner $A_h w \ge (1, \ldots, 1)^T$ für einen Vektor $w \in \mathbb{R}^N$, so folgt bzgl. der Maximumnorm bzw. Maximalen-Zeilensummen-Norm:

$$\|A_h^{-1}\|_\infty \le \|w\|_\infty.$$

c) Man zeige mit Hilfe von (b), dass für die Systemmatrix A_h des 5-Punkte-Schemas auf dem Einheitsquadrat die Abschätzung

$$\|A_h^{-1}\|_\infty \le 1/8$$

gilt, und folgere hiermit für die l_∞-Kondition von A_h:

$$\mathrm{cond}_\infty(A_h) := \|A_h\|_\infty \|A_h^{-1}\|_\infty \leq h^{-2}.$$

Die l_∞-Kondition von A_h verhält sich also in Abhängigkeit von der Gitterweite h genauso wie die Spektralkondition. (Hinweis: Man versuche es mit der mit $w(x,y) = x(1-x)/2 + y(1-y)/2$ gebildeten Gitterfunktion.)

Übung 2.6: Gegeben sei die 1. RWA des Laplace-Operators

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

auf dem Dreiecksgebiet $\Omega = \{(x,y) \in \mathbb{R}^2 |\, x, y > 0, x+y < 1\}$. Man stelle zu einem äquidistanten, kartesischen Gitter das Gleichungssystem der 5-Punkte-Differenzenapproximation auf und vergleiche (a) die zeilenweise, (b) die diagonale und (c) die schachbrettartige Gitterpunktnummerierung in Bezug auf Matrixstruktur, Speicherplatzbedarf und Rechenaufwand bei der Lösung mit der LR-Zerlegung unter Ausnutzung der Bandstruktur.

Übung 2.7: *Zur Auffrischung der Kenntnise über iterative Lösungsverfahren:* Eine Vereinfachung des Jacobi-Verfahrens zur Lösung eines linearen $N \times N$-Gleichungssystems $Ax = b$ ist das sog. „Richardson-Verfahren". Dabei wird ausgehend von einem beliebigen Startvektor $x^0 \in \mathbb{R}^N$ mit einem Dämpfungsparameter $\theta \in \mathbb{R}$ wie folgt iteriert:

$$x^{t+1} = x^t - \theta(Ax^t - b), \quad t = 0, 1, 2, \ldots .$$

a) Im Falle, dass A nur reelle Eigenwerte $\lambda_{\min} \leq \cdots \leq \lambda \leq \cdots \leq \lambda_{\max}$ besitzt, zeige man für den Spektralradius $\rho(B_\theta)$ der zugehörigen Iterationsmatrix $B_\theta = I - \theta A$ die Gleichung

$$\rho(B_\theta) = \max\big\{|1 - \theta\lambda_{\min}|, |1 - \theta\lambda_{\max}|\big\}.$$

b) Im Falle, dass zusätzlich alle Eigenwerte positiv sind, zeige man

$$\rho(B_\theta) < 1 \quad \Leftrightarrow \quad 0 < \theta < \frac{2}{\lambda_{\max}}.$$

c) Für welchen Wert von θ wird $\rho(B_\theta)$ in dieem Falle minimal?

Übung 2.8: Betrachtet werde wieder das Modellproblem

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

auf dem Einheitsquadrat $\Omega \subset \mathbb{R}^2$. Die Systemmatrix des 5-Punkte-Diffrenzenoperators auf einem äquidistanten, kartesischen Punktgitter lässt sich schreiben als Summe der Anteile der beiden Differenzenquotienten in x- und y-Richtung: $A = A_x + A_y$. Bei lexikographischer Numerierung sind dabei A_x und A_y reguläre Tridiagonalmatrizen mit den Einträgen $2h^{-2}$ auf den Hauptdiagonalen und $-h^{-2}$ verteilt auf den Nebendiagonalen.

Das zugehörige Gleichungssystems $AU = F$ besitzt damit die äquivalenten Formen

$$(\sigma I + A_x)U = (\sigma I - A_y)U + F, \quad (\sigma I + A_y)U = (\sigma I - A_x)U + F$$

mit einem beliebigen Parmeterwert $\sigma > 0$. Dies legt das folgende zweistufige Iterationsverfahren (sog. „ADI-Verfahren“ = „**A**lternating **D**irection **I**mplicit Iteration“) nahe:

$$(\sigma I + A_x)U^{t+1/2} = (\sigma I - A_y)U^t + F, \quad (\sigma I + A_y)U^{t+1} = (\sigma I - A_x)U^{t+1/2} + F.$$

Man zeige die Konvergenz dieses Verfahrens. Für welche Wahl von σ wird die Konvergenz am schnellsten? (Hinweis: Man überlege sich, dass die Zerlegungsmatrizen A_x und A_y ein gemeinsames System von Eigenvektoren besitzen und folglich vertauschbar sind.)