4 Lösung der FE-Gleichungen

In diesem Kapitel werden iterative Lösungsverfahren für die durch Anwendung einer Finite-Differenzen- oder Finite-Elemente-Diskretisierung entstehenden (linearen) Gleichungssysteme diskutiert. Dies sind neben den traditionellen Fixpunktiterationen vor allem sog. „PCG-Verfahren“ (preconditioned conjugate gradient methods) und die modernen „Mehrgittermethoden“. Zugrunde gelegt wird dabei meist wieder das Modellproblem der 1. RWA des Laplace-Operators

$$Lu := -\Delta u = f \quad \text{in } \Omega, \quad u = g \text{ auf } \partial\Omega, \tag{4.0.1}$$

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

$$Ax = b, \tag{4.0.2}$$

mit Matrizen $A = (a_{nm})_{n,m=1}^N$ und Vektoren $b = (b_n)_{n=1}^N$ der Dimension N. In der Praxis ist meist $N \gg 1000$, so dass neben dem Rechenaufwand auch der Speicherbedarf ein wichtiger Aspekt ist. Je nach der gewählten Numerierung der Gitterpunkte bzw. Knoten haben die Matrizen in der Regel Bandstruktur und sind extrem dünn besetzt. Ist das kontinuierliche Problem selbstadjungiert sowie definit, wie z. B. im Fall der 1. RWA des Laplace-Operators, so übertragen sich diese Eigenschaften bei FE-Diskretisierungen direkt auf die Systemmatrizen A.

4.1 Krylow-Raum-Methoden

Wir diskutieren jetzt sog. „Krylow[1]-Raum-Methoden“, zu denen auch das klassische Verfahren der konjugierten Gradienten („CG-Verfahren“) gehört. Im folgenden werden euklidisches Skalarprodukt und Norm auf $\mathbb{R}^N$ mit $\langle x, y\rangle$ bzw. $|x|$ bezeichnet. Die Koeffizientenmatrix $A \in \mathbb{R}^{N\times N}$ sei zunächst als symmetrisch und positiv definit angenommen. Dann lässt sich die Gleichung (4.0.2) äquivalent charakterisieren durch ein quadratisches Minimierungsproblem:

$$Ax = b \quad \Leftrightarrow \quad Q(x) = \min_{y\in\mathbf{R}^N} Q(y) \tag{4.1.3}$$

mit

$$Q(y) := \tfrac{1}{2}\langle Ay, y\rangle - \langle b, y\rangle.$$

[1] Aleksei Nikolaevich Krylov (1863–1945): Russischer Mathematiker; Prof. an der Sov. Akademie der Wissensch. in St. Petersburg; Beiträge zu Fourier-Analysis und Differentialgleichungen, Anwendungen in der Schiffstechnik.

Wegen $\nabla^2 Q \equiv A$ folgt aus der Positiv-Definitheit von A die Existenz eines eindeutig bestimmten Minimums von $Q(\cdot)$, welches notwendig Lösung von (4.0.2) ist. Diese Konstruktion ist analog zu der beim Nachweis von „schwachen“ Lösungen der 1. RWA des Laplace-Operators. Wir halten fest, dass der Gradient von Q in einem Punkt $y \in \mathbb{R}^n$ gegeben ist durch

$$\nabla Q(y) = \tfrac{1}{2}(A + A^T)y - b = Ay - b\,. \tag{4.1.4}$$

Dies ist gerade der „Defekt“ im Punkt y. Für jede symmetrische, positiv definite Matrix $B \in \mathbb{R}^{N\times N}$ ist durch $\|y\|_B := \langle By, y\rangle^{1/2}$ eine sog. „Energie-Norm“ definiert. Mit dieser Notation gilt dann

$$\begin{aligned}
2Q(y) &= \langle Ay, y\rangle - 2\langle b, y\rangle \\
&= \langle Ay, y\rangle - \langle b, y\rangle - \langle Ay, A^{-1}b\rangle + \langle b, A^{-1}b\rangle - \langle b, A^{-1}b\rangle \\
&= \langle Ay - b, y - A^{-1}b\rangle - \langle b, A^{-1}b\rangle \\
&= \langle A^{-1}(Ay - b), Ay - b\rangle - \langle b, A^{-1}b\rangle = |Ay - b|^2_{A^{-1}} - |b|^2_{A^{-1}} \\
&= \langle y - A^{-1}b, A(y - A^{-1}b)\rangle - \langle AA^{-1}b, A^{-1}b\rangle = |y - x|^2_A - |x|^2_A,
\end{aligned}$$

d. h.: Die Minimierung des Funktionals $Q(\cdot)$ ist äquivalent zur Minimierung der Defektnorm $|Ay - b|_{A^{-1}}$ bzw. der Energie-Norm $|y - x|_A$.

Die sog. „Abstiegsverfahren“ bestimmen nun ausgehend von einem geeigneten Startvektor $x^{(0)} \in \mathbb{R}^n$ eine Folge von Iterierten $x^{(t)}$, $t \in \mathbb{N}$, durch

$$x^{(t+1)} = x^{(t)} + \alpha_t d^{(t)}\,. \tag{4.1.5}$$

Dabei sind die $d^{(t)}$ vorgegebene oder auch erst im Verlauf der Iteration berechnete „Abstiegsrichtungen“, und die „Schrittweiten“ $\alpha_t \in \mathbb{R}$ sind durch die Vorschrift bestimmt (sog. “line search”):

$$Q(x^{(t+1)}) = \min_{\alpha\in\mathbb{R}} Q(x^{(t)} + \alpha d^{(t)})\,. \tag{4.1.6}$$

Die notwendige Optimalitätsbedingung

$$\frac{d}{d\alpha} Q(x^{(t)} + \alpha d^{(t)}) = \nabla Q(x^{(t)} + \alpha d^{(t)}) \cdot d^{(t)} = \langle Ax^{(t)} - b, d^{(t)}\rangle + \alpha\langle Ad^{(t)}, d^{(t)}\rangle = 0$$

ergibt mit dem Residuum $r^{(t)} := b - Ax^{(t)} = -\nabla Q(x^{(t)})$:

$$\alpha_t = \frac{\langle r^{(t)}, d^{(t)}\rangle}{\langle Ad^{(t)}, d^{(t)}\rangle}\,.$$

Das allgemeine Abstiegsverfahren lautet also wie folgt:

Startwert: $x^{(0)} \in \mathbb{R}^N$,

für $t \geq 0$: Iterierte $x^{(t)}$, Residuum $r^{(t)} = b - Ax^{(t)}$, Abstiegsrichtung $d^{(t)}$,

$$\alpha_t = \frac{\langle r^{(t)}, d^{(t)}\rangle}{\langle Ad^{(t)}, d^{(t)}\rangle}, \quad x^{(t+1)} = x^{(t)} + \alpha_t d^{(t)}.$$

Praktisch günstiger ist die folgende Schreibweise, bei der man eine Matrix-Vektor-Multiplikation spart, wenn man den Vektor $Ad^{(t)}$ abspeichert:

$$\begin{aligned} &\text{Startwert:} && x^{(0)} \in \mathbb{R}^n, \quad r^{(0)} := b - Ax^{(0)}, \\ &\text{für } t \geq 0: && \text{Iterierte } x^{(t)}, \quad \text{Residuum } r^{(t)}, \quad \text{Abstiegsrichtung } d^{(t)}, \\ & && \alpha_t = \frac{\langle r^{(t)}, d^{(t)} \rangle}{\langle Ad^{(t)}, d^{(t)} \rangle}, \quad x^{(t+1)} = x^{(t)} + \alpha_t d^{(t)}, \quad r^{(t+1)} = r^{(t)} - \alpha_t Ad^{(t)}. \end{aligned}$$

Die verschiedenen Abstiegsverfahren unterscheiden sich im wesentlichen durch die jeweilige Wahl der Abstiegsrichtungen $d^{(t)}$. Die einfachste Möglichkeit wäre, die Richtungen $d^{(t)}$ zyklisch die kartesischen Einheitsvektoren $\{e^{(1)}, \ldots, e^{(n)}\}$ durchlaufen zu lassen. Die so erhaltene iterative Methode wird „Koordinatenrelaxation“ genannt. Ein voller Relaxationszyklus ist äquivalent zum Gauß-Seidel-Verfahren (Übungsaufgabe).

Naheliegender ist die Wahl der Richtung des stärksten Abfalls von $Q(\cdot)$ im Punkt $x^{(t)}$, d. h. des Gradienten bzw. Residuums, als Suchrichtung $d^{(t)} = -g^{(t)} = -\nabla Q(x^{(t)}) = r^{(t)}$. Diese „Gradientenverfahren“ ist aber relativ langsam (vergleichbar mit dem Jacobi-Verfahren). Je zwei aufeinander folgende Abstiegsrichtungen sind dabei orthogonal zu einander, $(d^{(t+1)}, d^{(t)}) = 0$, aber $d^{(t+2)}$ braucht nicht einmal annähernd orthogonal zu $d^{(t)}$ zu sein. Dies führt zu einem stark oszillatorischen Konvergenzverhalten des Gradientenverfahrens besonders bei Matrizen A mit weit auseinander liegenden Eigenwerten. Dies bedeutet etwa in Fall $N = 2$, dass das Funktional $Q(\cdot)$ stark exzentrische Niveaulinien hat und sich die Iterierten in einem Zickzackkurs der Lösung annähern (s. Abb. 4.1).

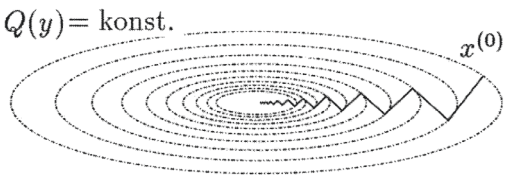

Abbildung 4.1: Niveaulinien des quadratischen Funktionals in zwei Dimensionen und „Zickzacking“ des Gradientenverfahrens

4.1.1 Verfahren der konjugierten Richtungen (CG-Verfahren)

Das Gradientenverfahren nutzt die Struktur des quadratischen Funktionals $Q(\cdot)$, d. h. die Verteilung der Eigenwerte der Matrix A, nur lokal von einem Schritt zum nächsten aus. Es wäre günstiger, wenn bei der Wahl der Abstiegsrichtungen auch die bereits gewonnenen Informationen über die globale Struktur von $Q(\cdot)$ berücksichtigt würden, d. h. wenn etwa die Abstiegsrichtungen paarweise orthogonal wären. Dies ist die Grundidee des sog. „Ver-

fahrens der konjugierten Gradienten“ nach Hestenes[2] und Stiefel[3] (“conjugate gradient method” oder kurz „CG-Verfahren“), welches sukzessive eine Folge von Abstiegsrichtungen $d^{(t)}$ erzeugt, die bzgl. des Skalarprodukts $(\cdot,\cdot)_A$ orthogonal sind („A-orthogonal“). Zur Konstruktion dieser Folge macht man den Ansatz

$$K_t = \text{span}\{d^{(0)}, ..., d^{(t-1)}\}$$

und sucht, Iterierte in der Form

$$x^{(t)} = x^{(0)} + \sum_{i=0}^{t-1} \alpha_i d^{(i)} \in x^{(0)} + K_t \tag{4.1.7}$$

zu bestimmen, so dass

$$Q(x^{(t)}) = \min_{y \in K_t} Q(x^{(0)} + y)\,.$$

Dies ist äquivalent zu den „Galerkin-Gleichungen“

$$\langle x^{(t)} - x, d^{(t)} \rangle_A = \langle Ax^{(t)} - b, d^{(j)} \rangle = 0, \quad j = 0, ..., t-1, \tag{4.1.8}$$

bzw. zu $r^{(t)} = b - Ax^{(t)} \perp K_t$. Setzt man den Ansatz (4.1.7) in (4.1.8) ein, so erhält man das Gleichungssystem

$$\sum_{i=0}^{t-1} \alpha_i \langle Ad^{(i)}, d^{(j)} \rangle = \langle b - Ax^{(0)}, d^j \rangle, \quad j = 0, ..., t-1, \tag{4.1.9}$$

mit der regulären Koeffizientenmatrix $M_A = (\langle Ad^{(k)}, d^{(j)} \rangle)_{j,k=0}^{t-1}$.

Eine natürliche Wahl der Ansatzräume K_t sind die sog. „Krylow-Räume“

$$K_t(r^{(0)}; A) := \text{span}\{r^{(0)}, Ar^{(0)}, ..., A^{t-1}r^{(0)}\}$$

zum Residuum $r^{(0)} = b - Ax^{(0)}$ des Startvektors $x^{(0)}$. Nach Konstruktion ist stets

$$\begin{aligned} r^{(t)} &= b - Ax^{(t)} = r^{(0)} - r^{(0)} + b - Ax^{(t)} \\ &= r^{(0)} + A(x^{(0)} - x^{(t)}) \in r^{(0)} + AK_t(r^{(0)}; A) \subset K_{t+1}(r^{(0)}; A). \end{aligned}$$

Da nach Konstruktion $r^{(t)} \perp K_t$, ist also stets

$$\langle r^{(t)}, r^{(i)} \rangle = 0, \quad i = 0, \ldots, t-1.$$

Ferner folgt im Fall $A^t r^{(0)} \in K_t(r^{(0)}; A)$ notwendig $r^{(t)} = 0$ bzw. $Ax^{(t)} = b$.

Ausgehend von der Formulierung (4.1.8) konstruiert das CG-Verfahren eine A-orthogo-

[2]Magnus R. Hestenes (1906-1991): US-amerikanischer Mathematiker; Prof. an der UCLA, USA, fundamentale Beiträge u.a. zur Numerischen Linearen Algebra.

[3]Eduard Stiefel (1909–1978): Schweizer Mathematiker; Prof. an der ETH Zürich; fundamentale Beiträge u.a. zur Numerischen Linearen Algebra.

nale Folge von Abstiegsrichtungen $d^{(i)}$, die eine Basis des Krylow-Raumes $K_t(r^{(0)}; A)$ bildet. Dies ließe sich etwa mit Hilfe des klassischen Gram[4] -Schmidt[5] -Algorithmus leisten, was aber numerisch sehr instabil ist. In Analogie zur vergleichbaren Situation bei der Konstruktion orthogonaler Polynome (z.B. die Legendre[6]-Polynome) ist aber zu erwarten, dass dasselbe durch eine zweistufige Rekursion erreichbar ist.

Ausgehend von einem Startpunkt $x^{(0)}$ mit Residuum (negativer Gradient) $r^{(0)} = b - Ax^{(0)}$ seien Iterierte $x^{(i)}$ und zugehörige Abstiegsrichtungen $d^{(i)} (i = 0, ..., t-1)$ bestimmt, so dass $\{d^{(0)}, ..., d^{(t-1)}\}$ eine A-orthogonale Basis von $K_t(d^{(0)}; A)$ ist. Zur Konstruktion des nächsten $d^{(t)} \in K_{t+1}(d^{(0)}; A)$ mit der Eigenschaft $d^{(t)} \perp_A K_t(d^{(0)}; A)$ machen wir den folgenden Ansatz:

$$d^{(t)} = r^{(t)} + \sum_{j=0}^{t-1} \beta_j^{t-1} d^{(j)} \ \in \ K_{t+1}(d^{(0)}; A). \tag{4.1.10}$$

Dabei wird o.B.d.A. angenommen, dass $r^{(t)} = b - Ax^{(t)} \notin K_t(d^{(0)}; A)$ ist, da andernfalls $r^{(t)} = 0$ bzw. $x^{(t)} = x$ wäre. Zur Bestimmung der Koeffizienten β_j^{t-1} beachten wir für $i = 0, ..., t-1$:

$$0 = \langle d^{(t)}, Ad^{(i)} \rangle = \langle r^{(t)}, Ad^{(i)} \rangle + \sum_{j=0}^{t-1} \beta_j^{t-1} \langle d^{(j)}, Ad^{(i)} \rangle = \langle r^{(t)} + \beta_i^{t-1}\, d^{(i)}, Ad^{(i)} \rangle \,.$$

Für $i < t-1$ ist $\langle r^{(t)}, Ad^{(i)} \rangle = 0$ wegen $Ad^{(i)} \in K_t(d^{(0)}; A)$ und demnach $\beta_i^{t-1} = 0$. Für $i = t-1$ führt die Bedingung

$$0 = \langle r^{(t)}, Ad^{(t-1)} \rangle + \beta_{t-1}^{t-1} \langle d^{(t-1)}, Ad^{(t-1)} \rangle \tag{4.1.11}$$

zu den Formeln

$$\beta_{t-1} := \beta_{t-1}^{t-1} = -\frac{\langle r^{(t)}, Ad^{(t-1)} \rangle}{\langle d^{(t-1)}, Ad^{(t-1)} \rangle}, \quad d^{(t)} = r^{(t)} + \beta_{t-1} d^{(t-1)}. \tag{4.1.12}$$

Die nächsten Iterierten $x^{(t+1)}$ und $r^{(t+1)} = b - Ax^{(t+1)}$ sind dann bestimmt durch

$$\alpha_t = \frac{\langle r^{(t)}, d^{(t)} \rangle}{\langle d^{(t)}, Ad^{(t)} \rangle}, \quad x^{(t+1)} = x^{(t)} + \alpha_t d^{(t)}, \quad r^{(t+1)} = r^{(t)} - \alpha_t Ad^{(t)}. \tag{4.1.13}$$

[4]Jørgen Pedersen Gram (1850–1916): Dänischer Mathematiker, Mitarbeiter und später Eigentümer einer Versicherungsgesellschaft, Beiträge zur Algebra (Invariantentheorie), Wahrscheinlichkeitstheorie, Numerik und Forstwissenschaft; das u.a. nach ihm benannte Orthogonalisierungsverfahren geht aber wohl auf Laplace zurück und wurde bereits von Cauchy 1836 verwendet.

[5]Erhard Schmidt (1876–1959): Deutscher Mathematiker, Prof. in Berlin, Gründer des dortigen Instituts für Angewandte Mathematik 1920, nach dem Krieg Direktor des Mathematischen Instituts der Akademie der Wissenschaften der DDR; Beiträge zur Theorie der Integralgleichungen und der Hilbert-Räume sowie später zur Topologie.

[6]Adrien-Marie Legendre (1752–1833): Französischer Mathematiker; Mitglied der Pariser Akademie der Wissensch.; Beiträge zur Himmelsmechanik, Zahlentheorie und Geometrie.

Dies sind die Rekursionsformeln des klassischen CG-Verfahrens. Wegen $r^{(t)} \perp K_t(r^{(0)}; A)$ und $K_t(r^{(0)}; A) = \text{span}\{d^{(0)}, \ldots, d^{(t-1)}\}$ ist

$$\langle r^{(t)}, d^{(i)} \rangle = 0, \quad i = 0, \ldots, t-1.$$

Damit lassen sich die Formeln für die Koeeffizienten α_t und β_t vereinfachen. Mit

$$|r^{(t)}|^2 = \langle d^{(t)} - \beta_{t-1} d^{(t-1)}, r^{(t+1)} + \alpha_t A d^{(t)} \rangle = \alpha_t \langle d^{(t)}, A d^{(t)} \rangle \,, \tag{4.1.14}$$

$$|r^{(t+1)}|^2 = \langle r^{(t)} - \alpha_t A d^{(t)}, r^{(t+1)} \rangle = -\alpha_t \langle A d^{(t)}, r^{(t+1)} \rangle \,. \tag{4.1.15}$$

ergibt sich

$$\alpha_t = \frac{|r^{(t)}|^2}{\langle d^{(t)}, A d^{(t)} \rangle}, \quad \beta_t = \frac{|r^{(t+1)}|^2}{|r^{(t)}|^2}, \tag{4.1.16}$$

solange die Iteration nicht mit $r^{(t)} = 0$ abbricht. Diese Konstruktion führt auf den folgenden „CG-Algorithmus“:

$$\begin{aligned}
&\text{Startwert:} && x^{(0)} \in \mathbb{R}^n, \quad r^{(0)} := b - A x^{(0)} \,, \\
&\text{für } t \geq 0: && \alpha_t = \frac{|r^{(t)}|^2}{\langle A d^{(t)}, d^{(t)} \rangle}, \\
& && x^{(t+1)} = x^{(t)} + \alpha_t d^{(t)}, \quad r^{(t+1)} = r^{(t)} - \alpha_t A d^{(t)}, \\
& && \beta_t = \frac{|r^{(t+1)}|^2}{|r^{(t)}|^2}, \quad d^{(t+1)} = r^{(t+1)} + \beta_t d^{(t)} \,.
\end{aligned}$$

In jedem Iterationsschritt ist dabei eine Matrix-Vektor-Multiplikation und fünf Operationen der Mächtigkeit eines Skalarproduktbildung durchzuführen. Im Falle einer dünn besetzten Matrix vom Typ der Modellmatrix sind das etwa $10N$ arithmetische Operationen. Das CG-Verfahren erzeugt eine Basis von Abstiegsrichtungen, so dass es zwangsläufig nach spätestens $t = N - 1$ Schritten mit der Lösung x des Gleichungssystems (4.0.2) abbricht. Es handelt sich hierbei also formal um ein „direktes“ Lösungsverfahren. Bei großer Dimension $N \geq 10^3$ geht die A-Orthogonalität der Folge $\{d^{(0)}, ..., d^{(t-1)}\}$ wegen des unvermeidlichen Rundungsfehlereinflusses schnell verloren, und das Verfahren wird zu einem nicht terminierenden, iterativen Prozess. Außerdem wäre die Durchführung von nahezu $N - 1$ Iterationsschritten wegen der damit verbundenen Zahl von $O(N^2)$ arithmetischen Operationen viel zu hoch. Man wird also mit einer wesentlich geringeren Anzahl von $t \ll N$ Schritten auskommen müssen. Wir fassen die wichtigsten Eigenschaften des CG-Verfahrens in folgendem Satz zusammen.

Satz 4.1 (CG-Konvergenz): *Das CG-Verfahren bricht für jeden Startvektor $x^{(0)} \in \mathbb{R}^N$ nach spätestens $N-1$ Schritten mit $x^{(t)} = x$ ab. Für $0 \leq t < N-1$ gilt die Fehlerabschätzung*

$$|e^{(t)}|_A \leq 2 \Big(\frac{1 - 1/\sqrt{\kappa}}{1 + 1/\sqrt{\kappa}} \Big)^t |e^{(0)}|_A, \quad t \geq 1, \tag{4.1.17}$$

mit der Spektralkondition $\kappa := \kappa_2(A) = \lambda_{\max}(A)/\lambda_{\min}(A)$ *von* A*. Zur Reduzierung des Anfangsfehlers um den Faktor* ε *sind höchstens*

$$t(\varepsilon) \leq \frac{1}{2}\sqrt{\kappa}\,\ln\left(\frac{2}{\varepsilon}\right) + 1 \tag{4.1.18}$$

Iterationsschritte erforderlich.

Dieses Resultat zeigt die Wichtigkeit der Kondition $\kappa(A)$ für die Konvergenzgeschwindigkeit des CG-Verfahrens. Für das Modellproblem ist $\kappa(A) = \mathcal{O}(h^{-2})$, was eine Gesamtlösungskomplexität von $O(N^{3/2})$ bedeutet. Das CG-Verfahren ist daher i.a. ähnlich effizient wie das „optimale“ SOR-Verfahren, allerdings mit einer größeren Fehlerkonstanten. Im Gegensatz zu letzterem erfordert das CG-Verfahren aber nicht die Bestimmung eines optimalen Relaxationsparameters. Dafür ist das Resultat (4.1.17) auf den Fall einer symmetrischen, positiv definiten Matrix A beschränkt.

Beweis: i) Unter Beachtung der Beziehung

$$|x^{(t)} - x|_A = \min_{y \in x^{(0)} + K_t} |y - x|_A\,,$$

$$K_t := K_t(r^{(0)}; A) = \operatorname{span}\{d^{(0)}, \ldots, d^{(t-1)}\} = \operatorname{span}\{A^0 r^{(0)}, \ldots, A^{t-1} r^{(0)}\}$$

finden wir

$$|x^{(t)} - x|_A = \min_{p \in P_{t-1}} |x^{(0)} - x + p(A) r^{(0)}|_A\,.$$

Wegen $r^{(0)} = b - Ax^{(0)} = A(x - x^{(0)})$ folgt weiter

$$\begin{aligned} |x^{(t)} - x|_A &= \min_{p \in P_{t-1}} |[I - p(A)A](x^{(0)} - x)|_A \\ &\leq \min_{p \in P_{t-1}} |I + Ap(A)|_A\, |x^{(0)} - x|_A \;\leq\; \min_{p \in P_t,\, p(0)=1} |p(A)|_A\, |x^{(0)} - x|_A\,, \end{aligned}$$

wobei die zu der Vektornorm $|\cdot|_A$ assoziierte natürliche Matrizennorm der Einfachheit halber ebenfalls mit $|\cdot|_A$ bezeichnet ist. Für beliebiges $y \in \mathbb{R}^N$ haben wir mit einer Orthonormalbasis $\{w^{(1)}, \ldots, w^{(N)}\}$ aus Eigenvektoren von A die Entwicklung

$$y = \sum_{i=1}^{N} \gamma_i w^{(i)}\,, \quad \gamma_i = \langle y, w^{(i)} \rangle\,,$$

und folglich

$$|p(A)y|_A^2 = \sum_{i=1}^{N} \lambda_i p(\lambda_i)^2 \gamma_i^2 \;\leq\; M^2 \sum_{i=1}^{N} \lambda_i \gamma_i^2 = M^2\, |y|_A^2\,,$$

wobei

$$M := \sup_{\lambda \leq \mu \leq \Lambda} |p(\mu)|\,, \quad \lambda := \lambda_{\min}(A),\;\; \Lambda := \lambda_{\max}(A)\,.$$

Dies impliziert dann

$$|p(A)|_A = \sup_{y\in\mathbb{R}^n,\, y\neq 0} \frac{|p(A)y|_A}{|y|_A} \leq M.$$

ii) Wir haben gefunden, dass

$$|x^{(t)} - x|_A \leq \min_{p\in P_t,\, p(0)=1} \Big\{ \sup_{\lambda\leq\mu\leq\Lambda} |p(\mu)| \Big\}\, |x^{(0)} - x|_A\,.$$

Dies ergibt die Behauptung, wenn wir zeigen können, dass

$$\min_{p\in P_t,\, p(0)=1} \Big\{ \sup_{\lambda\leq\mu\leq\Lambda} |p(\mu)| \Big\} \leq 2\Big(\frac{1-\sqrt{\lambda/\Lambda}}{1+\sqrt{\lambda/\Lambda}}\Big)^t.$$

Dabei handelt es sich um ein Problem der Bestapproximation mit Polynomen bzgl. der Maximumnorm (Tschebyscheff[7]-Approximation). Die Lösung $\bar{p}$ ist gegeben durch

$$\bar{p}(\mu) = T_t\Big(\frac{\Lambda+\lambda-2\mu}{\Lambda-\lambda}\Big)\; T_t\Big(\frac{\Lambda+\lambda}{\Lambda-\lambda}\Big)^{-1},$$

mit dem t-ten Tschebyscheff-Polynom T_t auf $[-1,1]$. Dabei ist

$$\sup_{\lambda\leq\mu\leq\Lambda} \bar{p}(\mu) = T_t\Big(\frac{\Lambda+\lambda}{\Lambda-\lambda}\Big)^{-1}.$$

Aus der Darstellung

$$T_t(\mu) = \tfrac{1}{2}\left[\big(\mu+\sqrt{\mu^2-1}\big)^t + \big(\mu-\sqrt{\mu^2-1}\big)^t\right], \quad \mu\in(-\infty,\infty),$$

für die Tschebyscheff-Polynome folgt über die Identität

$$\frac{\kappa+1}{\kappa-1} \pm \sqrt{\Big(\frac{\kappa+1}{\kappa-1}\Big)^2 - 1} = \frac{\kappa+1}{\kappa-1} \pm \frac{2\sqrt{\kappa}}{\kappa-1} = \frac{(\sqrt{\kappa}\pm 1)^2}{\kappa-1} = \frac{\sqrt{\kappa}\pm 1}{\sqrt{\kappa}\mp 1}$$

die Abschätzung nach unten

$$T_t\Big(\frac{\Lambda+\lambda}{\Lambda-\lambda}\Big) = T_t\Big(\frac{\kappa+1}{\kappa-1}\Big) = \frac{1}{2}\left[\Big(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\Big)^t + \Big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\Big)^t\right] \geq \frac{1}{2}\Big(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\Big)^t.$$

Also wird

$$\sup_{\lambda\leq\mu\leq\Lambda} \bar{p}(\mu) \leq 2\Big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\Big)^t,$$

was (4.1.17) impliziert.

[7]Pafnuty Lvovich Tschebyscheff (russ.: Chebyshev) (1821–1894): Russischer Mathematiker; Prof. in St. Petersburg; Beiträge zur Zahlentheorie, Wahrscjheinlichkeitstheorie und vor allem zur Approximationstheorie; entwickelte allgemeine Theorie orthogonaler Polsynome.

iii) Zur Herleitung von (4.1.18) fordern wir

$$2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{t(\varepsilon)} < \varepsilon \quad \Rightarrow \quad t(\varepsilon) > \ln\left(\frac{2}{\varepsilon}\right) \ln\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^{-1}.$$

Wegen

$$\ln\left(\frac{x+1}{x-1}\right) = 2\left\{\frac{1}{x} + \frac{1}{3}\frac{1}{x^3} + \frac{1}{5}\frac{1}{x^5} + \dots\right\} \geq \frac{2}{x}$$

ist dies erfüllt für $t(\varepsilon) \geq \frac{1}{2}\sqrt{\kappa}\,\ln(2/\varepsilon)$. Q.E.D.

4.1.2 CG-Verfahren für unsymmetrische und indefinite Probleme

Zur Lösung allgemeiner Gleichungssysteme $Ax = b$ mit einer regulären, aber nicht notwendig symmetrisch und positiv definiten Matrix $A \in \mathbb{R}^n$ mit Hilfe des CG-Verfahrens kann man etwa zu dem äquivalenten System

$$A^T A x \;=\; A^T b \tag{4.1.19}$$

mit der positiv definiten Matrix $A^T A$ übergehen. Hierauf angewendet, schreibt sich das CG-Verfahren wie folgt:

$$\begin{aligned}
\text{Startwerte:} \quad & x^{(0)} \in \mathbb{R}^N\,, \quad d^{(0)} = r^{(0)} = A^T(b - Ax^{(0)})\,, \\
\text{für } t \geq 0: \quad & \alpha_t = \frac{|r^{(t)}|^2}{|Ad^{(t)}|^2}\,, \\
& x^{(t+1)} = x^{(t)} + \alpha_t d^{(t)}\,, \quad r^{(t+1)} = r^{(t)} - \alpha_t A^T A d^{(t)}\,, \\
& \beta_t = \frac{|r^{(t+1)}|^2}{|r^{(t)}|^2}\,, \quad d^{(t+1)} = r^{(t+1)} + \beta_t d^{(t)}\,.
\end{aligned}$$

Die Konvergenzgeschwindigkeit ist dabei charakterisiert durch $\kappa(A^T A)$. Das ganze Verfahren beruht offenbar auf der Minimierung des Funktionals

$$Q(y) := \tfrac{1}{2}\,\langle A^T A y, y\rangle - \langle A^T b, y\rangle = \tfrac{1}{2}|Ay - b|^2 - \tfrac{1}{2}\,|b|^2\,. \tag{4.1.20}$$

Da $\kappa(A^T A) \sim \kappa(A)^2$ ist, muss man mit einer recht langsamen Konvergenz dieses „quadrierten“ CG-Verfahrens für nicht symmetrische Systeme rechnen.

Auf der Basis der Charakterisierung (4.1.3) ist das beschriebene CG-Verfahren auf symmetrische, positiv definite Matrizen beschränkt. Geht man allerdings von der „notwendigen Optimalitätsbedingung“ (4.1.8) aus, so ist dieser Ansatz auch für allgemeine Matrizen sinnvoll. Tatsächlich lassen sich auf diesem Wege leistungsfähige Verallgemeinerungen des CG-Verfahrens auch für unsymmetrische und indefinite Matrizen ableiten. Dabei werden in der Galerkin-Formulierung (4.1.8) als Ansatzräume meist wieder die Krylow-Räume

$$K_t = \text{span}\{r^{(0)}, Ar^{(0)}, ..., A^{t-1}r^{(0)}\}$$

verwendet. Als „Testräume“ treten gleichfalls $K_t^* = K_t$ oder auch

$$K_t^* = \text{span}\{r^{(0)}, A^T r^{(0)}, ..., (A^T)^{t-1} r^{(0)}\}$$

auf. Die resultierenden Verfahren GMRES („Generalized Minimal Residual“ von Y. Saad und M. H. Schultz, 1986), ORTHOMIN („Orthogonalization-Minimization“ nach P. K. W. Vinsome, 1976, und Eisenstat et al., 1983), CRS („Conjugate Residual Squared“ nach P. Sonneveld, 1989), BiCGSTAB („Biconjugate Gradient Stabilized“ nach H. A. Van der Vorst, 1992) u.s.w., haben dann jeweils die eine oder die andere Eigenschaft des normalen CG-Verfahrens, lassen aber keine so vollständige Konvergenzanalyse zu.

4.1.3 Vorkonditionierung (PCG-Verfahren)

Die Fehlerabschätzuung für das CG-Verfahren garantiert eine besonders gute Konvergenz, wenn die Kondition der Matrix A nahe bei Eins liegt. Daher wird eine „Vorkonditionierung“ vorgenommen, d.h.: Das System $Ax = b$ wird in ein äquivalentes umgeformt, $\tilde{A}\tilde{x} = \tilde{b}$, dessen Matrix $\tilde{A}$ besser konditioniert ist. Sei C eine symmetrische, positiv definite Matrix, welche explizit in Produktform

$$C = KK^T \tag{4.1.21}$$

gegeben ist mit einer regulären Matrix K. Das System $Ax = b$ wird dann in der äquivalenten Form geschrieben

$$\underbrace{K^{-1}A\,(K^T)^{-1}}_{\tilde{A}}\ \underbrace{K^T x}_{\tilde{x}} \;=\; \underbrace{K^{-1}b}_{\tilde{b}}\,. \tag{4.1.22}$$

Das CG-Verfahren wird nun auf das System $\tilde{A}\tilde{x} = \tilde{b}$ angewendet. Die Beziehung

$$(K^T)^{-1}\tilde{A}K^T \;=\; (K^T)^{-1}K^{-1}A(K^T)^{-1}K^T \;=\; C^{-1}A \tag{4.1.23}$$

zeigt, dass für $C \equiv A$ die Matrix $\tilde{A}$ ähnlich zu I, d. h. $\kappa(\tilde{A}) = \kappa(I) = 1$ wäre. Folglich wird man C^{-1} als möglichst gute Approximation von A^{-1} wählen, wobei natürlich die Zerlegung $C = KK^T$ bekannt sein muss. Das CG-Verfahren für das transformierte System $\tilde{A}\tilde{x} = \tilde{b}$ kann in den ursprünglichen Größen A, b und x als „PCG-Verfahren“ geschrieben werden:

$$\begin{aligned}
&\text{Startwert:} \qquad x^{(0)} \in \mathbb{R}^N,\ \ d^{(0)} = r^{(0)} = b - Ax^{(0)}\ \ \rho^{(0)} = C^{-1}r^{(0)}\,,\\
&\text{für } t \geq 0: \qquad \alpha_t = \frac{\langle r^{(t)}, \rho^{(t)}\rangle}{\langle Ad^{(t)}, d^{(t)}\rangle},\\
&\qquad\qquad x^{(t+1)} = x^{(t)} + \alpha_t d^{(t)},\ \ r^{(t+1)} = r^{(t)} - \alpha_t Ad^{(t)},\\
&\qquad\qquad \rho^{(t+1)} = C^{-1}r^{(t+1)},\\
&\qquad\qquad \beta_t = \frac{\langle r^{(t+1)}, \rho^{(t+1)}\rangle}{\langle r^{(t)}, \rho^{(t)}\rangle}\,, \quad d^{(t+1)} = r^{(t+1)} + \beta_t d^{(t)}\,.
\end{aligned}$$

Verglichen mit der einfachen CG-Iteration erfordert das PCG-Verfahren in jedem Schritt zusätzlich die Lösung des Systems $C\rho^{(t+1)} = r^{(t+1)}$, was unter Ausnutzung der Zerlegung $C = KK^T$ erfolgt. Zur Erzielung einer Komplexität von $\mathcal{O}(N)$ a.Op. pro Schritt sollte die Dreiecksmatrix K eine ähnliche Besetzungsstruktur wie der untere Dreiecksanteil L von A haben. Ausgehend von den oben betrachteten einfachen Fixpunktiterationen werden in der Praxis die folgenden Vorkonditionierer verwendet:

1) *Diagonal-Vorkonditionierung (Skalierung):* $C = D^{1/2}D^{1/2}$.
Die Skalierung bewirkt, dass die Elemente von A auf etwa gleiche Größenordnung gebracht werden, insbesondere wird $\tilde{a}_{ii} = 1$. Dies reduziert die Kondition, denn es gilt:

$$\kappa(A) \geq \frac{\max_{1\leq i\leq N}\, a_{ii}}{\min_{1\leq i\leq N}\, a_{ii}}\,. \tag{4.1.24}$$

Beispiel: Die Matrix $A = \mathrm{diag}\{\lambda_1 = ... = \lambda_{N-1} = 1, \lambda_N = 10^k\}$ hat die Kondition $\mathrm{cond}_2(A) = 10^k$. Die skalierte Matrix $\tilde{A} = D^{-1/2}AD^{-1/2}$ hat dagegen die optimale Kondition $\mathrm{cond}_2(\tilde{A}) = 1$.

2) *SSOR-Vorkonditionierung:* Mit einem Parameter ω wird gesetzt

$$C = (D + \omega L)D^{-1}(D + \omega R) = (\underbrace{D^{1/2} + \omega LD^{-1/2}}_{K})(\underbrace{D^{1/2} + \omega D^{-1/2}R}_{K^T})\,.$$

Offenbar besitzt die Dreiecksmatrix K dieselbe schwache Besetzung wie A. Pro Iterationsschritt erfordert das so vorkonditionierte Verfahren etwa doppelt so viel Aufwand wie das einfache Verfahren. Dagegen gilt für die Modellmatrix bei optimaler Wahl des Parameters ω (i. Allg. nicht leicht zu bestimmen!)

$$\kappa(\tilde{A}) = \sqrt{\kappa(A)}\,.$$

3) *ICCG-Verfahren (Incomplete Cholesky Conjugate Gradient):*
Die symmetrische, positiv definite Matrix A besitzt eine Cholesky-Zerlegung $A = LL^T$ mit einer unteren Dreiecksmatrix $L = (l_{ij})_{i,j=1}^N$. Die Elemente von L sind bestimmt durch die folgenden Rekursionsformeln:

$$l_{ii} = \Big(a_{ii} - \sum_{k=1}^{i-1} l_{ik}^2\Big)^{1/2}, \quad i = 1, \ldots, N\,,$$

$$l_{ji} = \frac{1}{l_{ii}}\Big(a_{ji} - \sum_{k=1}^{i-1} l_{jk}l_{ik}\Big), \quad j = i+1, \ldots, N\,.$$

Die Matrix L hat i. Allg. innerhalb der Hülle von A von Null verschiedene Elemente, erfordert also in der Regel weit mehr Speicherplatz als A selbst. Dies wird jedoch dadurch ausgeglichen, dass man nur eine "unvollständige Cholesky-Zerlegung" vornimmt, d. h.: Im Cholesky-Algorithmus werden einige der l_{ji} Null gesetzt, z. B.: $\tilde{l}_{ji} = 0$, wenn $a_{ji} = 0$.

Dies ergibt dann eine Zerlegung

$$A = \tilde{L}\tilde{L}^T + E \tag{4.1.25}$$

mit einer unteren Dreiecksmatrix $\tilde{L} = (\tilde{l}_{ij})_{i,j=1,\ldots,N}$, welche eine ähnliche (dünne) Besetzungsstruktur wie A besitzt. Man spricht von der $ICCG(0)$-Variante, wenn (4.1.25) gefordert wird. Werden im Fall einer Bandmatrix A weitere p Nebendiagonalen mit von Null verschiedenen Elementen in $\tilde{L}$ hinzugefügt bzw. weggestrichen, so nennt man dies die $ICCG(+p)$ bzw. $ICCG(-p)$-Variante. Zur Vorkonditionierung verwendet die ICCG-Methode die Matrix

$$C = KK^T = \tilde{L}\tilde{L}^T . \tag{4.1.26}$$

Obwohl keine strenge theoretische Begründung für den Erfolg dieses Ansatzes vorliegt, so zeigen doch numerische Tests an Modellproblemen, welchen Einfluss diese Konditionierung auf die Verteilung der Eigenwerte der Matrix $\tilde{A}$ hat. Zwar wird die Konditionszahl $\kappa(\tilde{A})$ nicht deutlich kleiner als $\kappa(A)$, doch die Eigenwerte von $\tilde{A}$ häufen sich im Gegensatz zu denen von A stark bei $\lambda = 1$. Dies bewirkt, wie eine feinere Analyse zeigt, eine deutliche Beschleunigung der Konvergenz.

4) *ADI-Vorkonditionierung:* $C = (A_x + \omega I)(A_y + \omega I)(A_y^T + \omega I)(A_x^T + \omega I)$.
Auch hier muss zur Bewahrung der Symmetrie von $\tilde{A}$ eine symmetrisierte Variante des ADI-Matrix verwendet werden.

Eine gute Vorkonditionierung der Modellmatrix bewirkt eine Verbesserung der Konvergenzrate von $\rho(A) = 1 - \mathcal{O}(h)$ auf $\rho(\tilde{A}) = 1 - \mathcal{O}(h^{1/2})$ und damit auf die Lösungskomplexität $\mathcal{O}(N^{5/4})$. Besonders die ILU-Vorkonditionierung hat sich in der Praxis bei vielen Problemen als effizient und robust erwiesen. Sie reduziert zwar nicht die Kondition, doch bewirkt eine Konzentration der Eigenwerte um den Wert $\lambda = 1$, was ebenfalls eine deutliche Beschleunigung der CG-Konvergenz mit sich bringt.

4.2 Mehrgitterverfahren

Mehrgitterverfahren gehören zum Typ der (verallgemeinerten) Defektkorrekturiterationen und verwenden eine Folge von Subproblemen ähnlicher Struktur, aber sukzessive kleiner werdender Dimension. Sie sind speziell zugeschnitten auf die Lösung der Gleichungssysteme, wie sie bei der Diskretisierung partieller Differentialgleichungen mit Differenzen- oder Finite-Elemente-Verfahren entstehen. Die Idee ist die eines allen diskreten Problemen zugrunde liegenden übergeordneten, kontinuierlichen Problems und der fortgesetzten Aufspaltung von Fehlern und Defekten auf den verschiedenen Gittern in „niedrig- und hochfrequente“ Anteile, die separat behandelt werden. Bei richtiger Zusammenstellung der einzelnen Verfahrenskomponenten erhält man Idealfall das gewünschte „optimale“ Lösungsverfahren mit arithmetischem Aufwand $\mathcal{O}(N)$ für die Berechnung der N Unbekannten auf dem feinsten Gitter. Die Grundidee des Mehrgitterverfahrens geht auf die

russischen Mathematiker R. Fedorenko[8] und N. Bachwalow[9] in den 1960-er Jahren zurück. Danach wurde der prinzipielle Ansatz in den späten 1970-er Jahren unabhngig voneinander von A. Brandt[10] und W. Hackbusch[11] zu einem allgemein anwendbaren Verfahren entwickelt. Die im Folgenden dargestellte Konvergenz- und Komplexitätsanalyse ist in ihrer rigorosen Form für Differenzenverfahren in Hackbusch [19] beschrieben.

Zum Einstieg betrachten wir das Gleichungssystem auf dem Gitter $\mathbb{T}_h$:

$$A_h x_h = b_h \tag{4.2.27}$$

und approximieren die Lösung mit dem Richardson-Verfahren

$$x_h^{(t+1)} = x_h^{(t)} + \theta_h(b_h - A_h x_h^{(t)}) = (I_h - \theta_h A_h)x_h^{(t)} + \theta_h b_h \tag{4.2.28}$$

mit einem Dämpfungsfaktor $0 < \theta_h \leq 1$. Die symmetrische, positiv definite Matrix A_h besitzt ein Orthonormalsystem von Eigenvektoren $\{w_h^{(i)}, i = 1, ..., N_h\}$ zu den geordneten Eigenwerten $\lambda_{\min}(A_h) = \lambda_1 \leq ... \leq \lambda_N = \lambda_{\max}(A_h) =: \Lambda_h$. Entwickelt man den Anfangsfehler in der Form

$$e_h^{(0)} := x_h^{(0)} - x_h = \sum_{i=1}^{N_h} \varepsilon_i w_h^{(i)},$$

so gilt für die iterierten Fehler entsprechend

$$e_h^{(t)} = (I_h - \theta_h A_h)^t e_h^{(0)} = \sum_{i=1}^{N_h} \varepsilon_i (I_h - \theta_h A_h)^t w_h^{(i)} = \sum_{i=1}^{N_h} \varepsilon_i (1 - \theta_h \lambda_i)^t w_h^{(i)}.$$

[8]Radi Petrowitsch Fedorenko (1930–2009): Russischer Mathematiker; arbeitete ab 1953 am Keldysh-Institut für Angewandte Mathematik der Russischen Akademie der Wissenschaften; numerische Berechnungen fr das sowjetische Kernwaffen- und Kerntechnikprojekt sowie fr Luft- und Raumfahrt; erste Arbeiten unterlagen der Geheimhaltung, erste Veröffentlichung 1958 über ein Problem der Magnetohydrodynamik; Pionier der Mehrgittermethode mit Arbeiten Anfang der 1960-er Jahre in Zusammenhang mit der numerischen Lsung der Poisson-Gleichung in der Wettervorhersage.

[9]Nikolai Sergejewitsch Bachwalow (1934–2005): Russischer Mathematiker; Arbeiten zur Numerik; Promotion 1958 in Moskau bei A. Kolmogorow; ab 1966 Prof. an der Lomonossow-Universität und 1981 Abteilung Numerische Mathematik; Pionier der Mehrgitterverfahren und Beiträge zur Numerik von Wellenphnomenen und Verbundmaterialien (Methode der Homogenisierung und der „Fictitious Domain"-Methode); Autor verbreiteter russischer Lehrbcher ber Numerik.

[10]Achi Brandt (1938–): Israelischer Mathematiker; Arbeiten über partielle Differentialgleichungen und Numerik; Prof. am Weizmann-Institut in Rechovot (Israel) und an der Univ. of California (Los Angeles, USA); einer der Pioniere der Mehrgittermethode (sog. „Full Approximation Scheme", FAS, 1977); behauptet, jede partielle Differentialgleichung sei durch Mehrgitterverfahren effizient und schnell lösbar; Mitgrnder der Softwarefirma VideoSurf.

[11]Wolfgang Hackbusch (1948–): Deutscher Mathematiker; Studium in Marburg; Promotion 1973 und Habilitation 1979 in Köln R. Bulirsch; Professuren für Praktische Mathematik in Bochum und Kiel; 1999/2000–2014 Direktor am MPI für fr Mathematik in den Naturwissenschaften in Leipzig; wichtige Beiträge zur Numerik von partiellen Differentialgleichungen und Integralgleichungen; am besten bekannt durch seine Arbeiten zur „Mehrgittermethode", dem sog. „Panel Clustering" und der „H-Matrizen".

Folglich ist

$$|e_h^{(t)}|^2 = \sum_{i=1}^{N_h} \varepsilon_i^2 (1 - \theta_h \lambda_i)^{2t}. \tag{4.2.29}$$

Die Bedingung $0 < \theta_h \leq \Lambda_h^{-1}$ ist hinreichend für die Konvergenz der Richardson-Iteration. Wegen $|1 - \theta_h \lambda_i| \ll 1$ für große λ_i und $|1 - \theta_h \lambda_1| \approx 1$ werden offenbar „hoch-frequente“ Komponenten des Fehlers sehr schnell, aber „niedrig-frequente“ nur sehr langsam gedämpft. Dasselbe gilt auch für das Residuum $r_h^{(t)} = b_h - A_h x_h^{(t)} = A_h e_h^{(t)}$, d.h.: Bereits nach wenigen Iterationen gilt:

$$|r_h^{(t)}|^2 \approx \sum_{i=1}^{[N/2]} \varepsilon_i^2 \lambda_i^2 (1 - \theta_h \lambda_i)^{2t}, \tag{4.2.30}$$

wobei $[N/2] := \max\{n \in \mathbf{N} | \, n \leq N/2\}$ ist. Dies kann so interpretiert werden, dass der iterierte Defekt $r_h^{(t)}$ auf dem Gitter $\mathbb{T}_h$ glatt ist. Daher sollte er auf einem gröberen Gitter $\mathbb{T}_{2h}$ mit Gitterweite $2h$ gut approximierbar sein. Die resultierende Defektgleichung zur Berechnung der Korrektur zur Näherung $x_h^{(t)}$ auf $\mathbb{T}_h$ würde dann wegen ihrer geringeren Dimension $N_{2h} \approx N_h/4$ auch weniger Aufwand kosten. Dieser Defektkorrekturprozess in Verbindung mit sukzessiver Vergröberung kann weitergeführt werden bis zu einem gröbsten Gitter, auf dem die Defektgleichung dann exakt gelöst wird. Die wichtigsten Bestandteile eines solchen Mehrgitterprozesses sind die „Glättungsiteration“ $x_h^{(\nu)} = S_h^\nu(x_h^{(0)})$ sowie geeignete Transferoperationen zwischen den Finite-Elemente-Räumen auf den verschiedenen Gittern. Die Glättungsoperation $S_h(\cdot)$ ist gewöhnlich gegeben in Form einer Fixpunktiteration (z.B. der Richardson-Iteration)

$$x_h^{(\nu+1)} = S_h(x_h^{(\nu)}) := (I_h - C_h^{-1} A_h) x_h^{(\nu)} + C_h^{-1} b_h,$$

mit der Iterationsmatrix $S_h := I_h - C_h^{-1} A_h$.

4.2.1 Mehrgitteralgorithmus im Finite-Elemente-Kontext

Zur Formalisierung des Mehrgitterprozesses betrachten wir nun eine Folge von Gittern $\mathbb{T}_l = \mathbb{T}_{h_l}$, $l = 0, ..., L$, zunehmender Feinheit $h_0 > ... > h_l > ... > h_L$ sowie zugehörige FE-Räume $V_l := V_{h_l} \subset V$. Der Einfachheit halber sei angenommen, dass die FE-Räume hierarchisch geordnet sind, d.h.: $V_0 \subset V_1 \subset ... \subset V_l \subset ... \subset V_L$. Diese Voraussetzung erleichtert die Analyse des Mehrgitterprozesses, ist aber nicht entscheidend für sein Funktionieren. Zwischen den Funktionen $v_l \in V_l$ und den zugehörigen Knotenwertevektoren $y_l \in \mathbb{R}^{N_l}$ gilt der übliche Zusammenhang $v_l(a_n) = y_{l,n}$, $n = 1, ..., N_l$. Wie üblich schreiben wir das kontinuierliche Problem und sein FE-Analogon in variationeller Form als

$$a(u, \varphi) = (f, \varphi) \quad \forall \varphi \in V, \tag{4.2.31}$$

bzw. auf dem feinsten Gitter $\mathbb{T}_L$ als

$$a(u_L, \varphi_L) = (f, \varphi_L) \quad \forall \varphi_L \in V_L. \tag{4.2.32}$$

Dabei sind $a(u, \varphi) := (Lu, \varphi)$ die zum (elliptischen) Operator L gehörende "Energie-Form" und (f, φ) das L^2-Skalarprodukt auf dem Lösungsgebiet Ω. Die „exakte" diskrete Lösung $u_L \in V_L$ genügt der a priori Fehlerabschätzung

$$\|u - u_L\| \leq c\, h_L^2\, \|f\|. \tag{4.2.33}$$

Ziel ist es, einen Lösungsprozess zu finden, der eine Approximation $\tilde{u}_L \approx u_L$ liefert mit

$$\|u_L - \tilde{u}_L\| \leq c\, h_L^2\, \|f\|. \tag{4.2.34}$$

Ist der dazu erforderliche Aufwand $\mathcal{O}(N_L)$ und zwar gleichmäßig bzgl. L, so nennt man diesen Prozess „komplexitäts-optimal". Wir werden sehen, dass der Mehrgitteralgorithmus bei richtiger Wahl der Verfahrenskomponenten in diesem Sinne „optimal" ist.

Sei $u_L^{(0)} \in V_L$ eine Schätzung für die exakte Lösung $u_L \in V_L$ auf Gitterlevel L. Zunächst wird $u_L^{(0)}$ „geglättet". Dazu werden ausgehend von $\bar{u}_L^{(0)} := u_L^{(0)}$ z. B. ν Schritte des Richardson-Verfahrens durchgeführt. In variationeller Schreibweise lautet dies:

$$(\bar{u}_L^{(k)}, \varphi_L) = (\bar{u}_L^{(k-1)}, \varphi_L) + \theta_L \big\{ (f, \varphi_L) - a(\bar{u}_L^{(k-1)}, \varphi_L) \big\} \quad \forall \varphi_L \in V_L, \tag{4.2.35}$$

wobei $\theta_L = \lambda_{max}(A_h)^{-1}$. Mit der geglätteten Approximation wird der Defekt $d_L \in V_L$ gebildet (ohne ihn wirklich zu berechnen):

$$(d_L, \varphi_L) := (f, \varphi_L) - a(\bar{u}_L^{(\nu)}, \varphi_L), \quad \varphi_L \in V_L. \tag{4.2.36}$$

Wegen $V_{L-1} \subset V_L$ erhält man auf dem nächst gröberen Gitter $\mathbb{T}_{L-1}$ die Defektgleichung („Grobgittergleichung")

$$a(q_{L-1}, \varphi_{L-1}) = (d_L, \varphi_{L-1}) = (f, \varphi_{L-1}) - a(\bar{u}_L^{(\nu)}, \varphi_{L-1}) \quad \forall \varphi_{L-1}. \tag{4.2.37}$$

Die Korrektur $q_{L-1} \in V_{L-1}$ wird nun entweder exakt berechnet (etwa mit einem „direkten" Löser) oder nur näherungsweise mit Hilfe einer Defektkorrekturiteration $q_{L-1}^{(0)} \to q_{L-1}^{(R)}$ unter Verwendung der noch gröberen Gitter $\mathbb{T}_{L-2}, ..., \mathbb{T}_0$. Das Ergebnis $q_{L-1}^{(R)} \in V_{L-1}$ wird dann als Element von V_L interpretiert und zur Korrektur von $\bar{u}_L^{(\nu)}$ verwendet:

$$\bar{\bar{u}}_L^{(0)} := \bar{u}_L^{(\nu)} + \omega_L q_{L-1}^{(R)}. \tag{4.2.38}$$

Dabei wird der Dämpfungsparameter $\omega_L \in (0, 1)$ verwendet, um das Residuum von $\bar{\bar{u}}_L$ zu minimieren. Auf diesen in der Praxis sehr nützlichen Trick wollen wir hier nicht weiter eingehen. Die erhaltene korrigierte Näherung $\bar{\bar{u}}_L$ wird nun nochmals μ-mal „nachgeglättet". Ausgehend von $\bar{\bar{u}}_L^{(0)} := \bar{\bar{u}}_L$ wird etwa wieder mit dem Richardson-Verfahren iteriert:

$$(\bar{\bar{u}}_L^{(k)}, \varphi_L) = (\bar{\bar{u}}_L^{(k-1)}, \varphi_L) + \theta_L \big\{ (f, \varphi_L) - a(\bar{\bar{u}}_L^{(k-1)}, \varphi_L) \big\} \quad \forall \varphi_L \in V_L. \tag{4.2.39}$$

Das Ergebnis wird schließlich als die nächste Mehrgitteriterierte $u_L^{(1)} := \bar{\bar{u}}_L^{(\mu)}$ akzeptiert. Damit haben wir einen Schritt des Mehrgitterverfahrens (einen „Zyklus“) auf dem Gitterlevel L beschrieben. Jeder solche Zyklus beinhaltet also neben $\nu+\mu$ Richardson-Schritten (auf Level L), welche jeweils eine Inversion der Massematrix erfordern, die Lösung des „Grobgitterproblems“ (4.2.37).

Wir wollen nun den beschriebenen Mehrgitteralgorithmus in etwas abstrakterer Form darstellen, um seine Struktur besser zu verstehen und ihn auch leichter analysieren zu können. Zu den Matrizen $A_l = A_{h_l}$ auf den Gittern $\mathbb{T}_l$ sind Operatoren $\mathcal{A}_l : V_l \to V_l$ assoziiert durch

$$(\mathcal{A}_l v_l, w_l) = a(v_l, w_l) = \langle A_l y_l, z_l \rangle \quad \forall v_l, w_l \in V_l \,. \tag{4.2.40}$$

Weiter seien $\mathcal{S}_l(\cdot)$ die zugehörigen Glättungsoperationen mit (linearen) Iterationsoperatoren $\mathcal{S}_l$. Beim Richardson-Verfahren ist der Iterationsoperator $\mathcal{S}_l = \mathcal{I}_l - \theta_l \mathcal{A}_l$. Schließlich führen wir noch Transferoperatoren zwischen aufeinander folgenden Räumen ein:

$$r_l^{l-1} : V_l \to V_{l-1} \ \text{(Restriktion)}, \quad p_{l-1}^l : V_{l-1} \to V_l \ \text{(Prolongation)}.$$

Im Finite-Elemente-Kontext ist natürlicherweise $r_l^{l-1} = P_{l-1}$ die L^2-Projektion und $p_{l-1}^l = id.$ die natürliche Einbettung. Wir beschreiben nun den Mehrgitterprozeß zur Berechnung der Lösung des Sytems

$$\mathcal{A}_L u_L = f_L \tag{4.2.41}$$

auf dem „feinsten“ Gitter $\mathbb{T}_L$.

Mehrgitterprozess: Ausgehend von einem Startwert $u_L^{(0)} \in V_L$ werden Iterierte $u_L^{(t)}$ durch den folgenden rekursiven Prozess

$$u_L^{(t+1)} = MG(L, u_L^{(t)}, f_L) \tag{4.2.42}$$

erzeugt. Sei also die t-te Mehrgitteriterierte $u_L^{(t)}$ bestimmt.

Grobgitterlösung: Für $l = 0$ bedeutet $MG(0, \cdot, g_0)$ stets die exakte Lösung des Systems $\mathcal{A}_0 v_0 = g_0$ (z.B. mit Hilfe eines direkten Lösungsverfahrens), d. h.:

$$v_0 = MG(0, \cdot, g_0) = \mathcal{A}_0^{-1} g_0 \,. \tag{4.2.43}$$

Rekursion: Sei für ein $1 \le l \le L$ das System $\mathcal{A}_l v_l = g_l$ zu lösen. Mit Parameterwerten ν, $\mu \ge 1$ ist dann

$$MG(l, v_l^{(0)}, g_l) := v_l^{(1)} \approx v_l \tag{4.2.44}$$

rekursiv definiert durch die folgenden Schritte:

i) *Vorglättung:*

$$\bar{v}_l := \mathcal{S}_l^{\nu}(v_l^{(0)}) \,; \tag{4.2.45}$$

ii) *Defektbildung:*

$$d_l := g_l - \mathcal{A}_l \bar{v}_l, \; ; \tag{4.2.46}$$

iii) *Restriktion:*

$$\tilde{d}_{l-1} := r_l^{l-1} d_l \, ; \tag{4.2.47}$$

iv) *Defektgleichung:* Ausgehend von $q_{l-1}^{(0)} := 0$ wird für $1 \le r \le R$ iteriert:

$$q_{l-1}^{(r)} := MG(l-1, q_{l-1}^{(r-1)}, \tilde{d}_{l-1}) \, ; \tag{4.2.48}$$

v) *Prolongation:*

$$q_l := p_{l-1}^l q_{l-1}^{(R)} ; \tag{4.2.49}$$

vi) *Korrektur:* Mit einem Dämpfungsparameter $\omega_l \in (0, 1]$ wird gesetzt:

$$\bar{\bar{v}}_l := \bar{v}_l + \omega_l q_l \, ; \tag{4.2.50}$$

vii) *Nachglättung:*

$$v_l^{(1)} := \mathcal{S}_l^{\mu}(\bar{\bar{v}}_l) \, ; \tag{4.2.51}$$

Im Falle $l = L$ wird schließlich gesetzt:

$$u_L^{(t+1)} := v_l^{(1)} \, . \tag{4.2.52}$$

Schematische Darstellung des Mehrgitterschritts $u_L^{(t)} \to u_L^{(t+1)}$:

$$\begin{aligned} u_L^{(t)} \;\to\; \bar{u}_L^{(t)} = S_L^{\nu}(u_L^{(t)}) \;\to\; d_L &= f_L - \mathcal{A}_L \bar{u}_L^{(t)} \\ \downarrow \;\; \tilde{d}_{L-1} &= r_L^{L-1} d_{L-1} \quad \text{(Restriktion)} \\ q_{L-1} = \tilde{\mathcal{A}}_{L-1}^{-1} \tilde{d}_{L-1} &\quad (R\text{-malige Defektkorrektur}) \\ \downarrow \;\; \tilde{q}_L &= p_{L-1}^L q_{L-1} \quad \text{(Prolongation)} \\ \bar{\bar{u}}_L^{(t)} = \bar{u}_L^{(t)} + \omega_L \tilde{q}_L \;\to\; u_L^{(t+1)} &= S_L^{\mu}(\bar{\bar{u}}_L^{(t)}) \end{aligned}$$

Wenn die Defektgleichung $\mathcal{A}_{L-1} q_{L-1} = \tilde{d}_{L-1}$ auf dem gröberen Gitter $\mathbb{T}_{L-1}$ „exakt“ gelöst wird (z. B. durch Gauß-Elimination), spricht man von einer „Zweigittermethode“. In der Regel wird der Prozess aber rekursiv zum Mehrgitterverfahren bis zum gröbsten Gitter fortgesetzt. Dabei kann der vollständige Mehrgitterzyklus auf verschiedene Art organisiert werden. Seine Struktur ist im wesentlichen durch den Parameter R bestimmt, der festlegt, wie oft der Defektkorrekturprozess auf jedem Gitterlevel durchgeführt wird. In der Praxis spielen nur die Fälle $R = 1$ oder $R = 2$ eine Rolle. Dem entsprechen der im schematischen Bild gezeigte sog. „V-Zyklus“ und der sog. „W-Zyklus“. Dabei stehen die Punkte „$\bullet$“ für Glättung und Defektkorrektur auf den Gittern $\mathbb{T}_l$, und die Linie „$-$“ für den Transfer zwischen aufeinander folgenden Gitterniveaus.

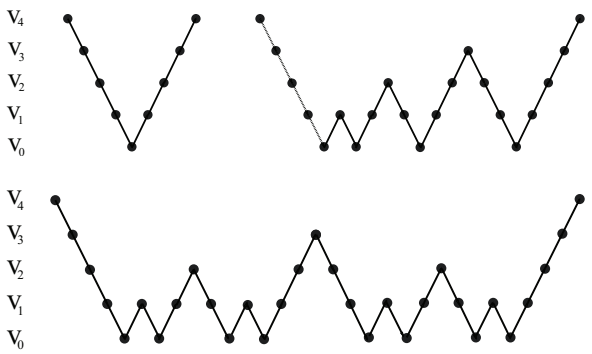

Abbildung 4.2: Schema eines Mehrgitterverfahrens mit V- (oben links) F- (oben rechts) und W-Zyklus (unten).

Der V-Zyklus ist sehr effizient (wenn er funktioniert), krankt aber oft an Instabilitäten, welche durch Irregularitäten im Problem (starke Unsymmetrien, Eckensingularitäten, Gitterunregelmäßigkeiten u.s.w.) hervorgerufen werden. Der W-Zyklus ist dagegen sehr robust, aber auch teurer. Methoden mit $R \geq 3$ sind gewöhnlich zu ineffizient. Ein guter Kompromiss zwischen V-Zyklus und W-Zyklus stellt der sog. „F-Zyklus“ dar. Dieser wird gewöhnlich auf Gitter $\mathbb{T}_L$ gestartet mit einem beliebigen Startvektor $u_L^{(0)}$ (meist $u_L^{(0)} = 0$). Wird dieser Prozess allerdings zur Lösung eines nicht linearen Problems im Rahmen einer Newton-Iteration angewendet, so kann dieser Startwert zu ungenau sein, und die ganze Iteration divergiert. In einem solchen Fall startet man zur Generierung einer hinreichend genauen Anfangsapproximation den Mehrgitterprozess gewöhnlich auf dem gröbsten Gitter $\mathbb{T}_0$. Wir beschreiben dieses sog. „geschachtelte“ Mehrgitter-Schema (“nested multigrid”) für das *lineare* Problem.

Geschachteltes MG: Ausgehend von dem Startwert $u_0 := \mathcal{A}_0^{-1} f_0$ auf dem gröbsten Gitter $\mathbb{T}_0$ werden für $l = 1, ..., L$ rekursiv Näherungen $\tilde{u}_l \approx u_l$ berechnet nach der Vorschrift:

$$
\begin{aligned}
u_l^{(0)} &= p_{l-1}^l \tilde{u}_{l-1} \\
u_l^{(t)} &= MG(l, u_l^{(t-1)}, f_l), \quad t = 1, ..., t_l, \quad \|u_l^{(t_l)} - u_l\| \leq \hat{c}\, h_l^2 \, \|f\|, \\
\tilde{u}_l &= u_l^{(t_l)}.
\end{aligned}
$$

Es gibt nicht „den Mehrgitteralgorithmus“. Die erfolgreiche Realisierung des Mehrgitterkonzepts erfordert eine sorgfältige Balance der verschiedenen Bestandteile wie Glätter $\mathcal{S}_l$ und Gitteroperatoren $\mathcal{A}_l$ sowie der Gittertransfers r_l^{l-1} und p_{l-1}^l jeweils für das zu lösende Problem. Im folgenden werden wir diese Verfahrenskomponenten im Rahmen des Finite-Elemente-Kontexts diskutieren.

i) *Glätter:* „Glätter“ sind üblicherweise einfache Fixpunktiterationen, die auch als „Löser“ verwendet werden können, aber mit einer sehr schlechten Konvergenzrate. Sie werden auf jedem Gitterniveau nur ein paarmal angewendet ($\nu, \mu \sim 1 - 4$), um die hochfrequenten Fehleranteile auszudämpfen. Wir betrachten im folgenden nur das klassische Richardson-

Verfahren,

$$\mathcal{S}_l := \mathcal{I}_l - \theta_l \mathcal{A}_l\,, \quad \theta_l = \lambda_{max}(\mathcal{A}_l)^{-1}\,, \tag{4.2.53}$$

welches aber nur bei sehr „gutartigen“ Problemen funktioniert. Leistungsfähiger und robuster sind das Gauß-Seidel- und das ILU-Verfahren. Diese funktionieren auch noch gut, wenn das Problem gewisse Pathologien beinhaltet. Im Fall eines starken Advektionsterms besitzt die Systemmatrix bei Numerierung der Knotenpunkte in Transportrichtung einen dominanten unteren Dreicksanteil L, für den die Gauß-Seidel-Methode „exakt“ ist. Für Probleme mit degenerierten Koeffizienten in einer Raumrichtung sowie auf stark anisotropen Gittern besitzt die Systemmatrix einen dominaten Tridiagonalanteil, für den wiederum die ILU-Iteration „exakt“ ist. Für echt indefinite Probleme werden spezielle, der jeweiligen Struktur des Problems angepasste Glätter verwendet, deren Diskussion aber außerhalb des Rahmens dieses einführenden Textes liegt. Auf lokal verfeinerten Gittern darf die Glättung im Wesentlichen nur auf den jeweils neu hinzugekommenen Zellen operieren, da sonst der arithmetische Aufwand pro Gitterlevel zu groß wird.

ii) *Gittertransfers:* Im Kontext einer Finite-Elemente-Diskretisierung mit geschachtelten Ansatzräumen $V_0 \subset V_1 \subset ... \subset V_l \subset ... \subset V_L$ ist die generische Wahl für die Prolongation $p_{l-1}^l : V_{l-1} \to V_l$ die zellweise Einbettung und für die Restriktion $r_l^{l-1} : V_l \to V_{l-1}$ die L^2-Projektion. Bei anderen Diskretisierungen (z. B. Differenzenschemata) verwendet man geeignete Interpolationsprozesse (z. B. bilineare Interpolation).

iii) *Grobgitteroperatoren:* Die Operatoren $\mathcal{A}_l$ auf den verschiedenen Gitterniveaus müssen nicht notwendig zur selben Diskretisierung des Ausgangsproblems gehören. Dies wird z. B. wichtig bei der Berücksichtigung von gitterweitenabhängiger künstlicher Diffusion („upwinding“) zur Behandlung von Transporttermen. Wir beschränken uns hier aber auf den Idealfall, dass alle $\mathcal{A}_l$ durch dieselben FE-Diskretisierungen auf der Gitterhierarchie $\{\mathbb{T}_l\}_{l=0,...,L}$ erzeugt sind. In diesem Fall gilt die für die theoretische Analyse nützliche Beziehung

$$\begin{aligned}(\mathcal{A}_{l-1} v_{l-1}, w_{l-1}) &= a(v_{l-1}, w_{l-1}) \\ &= a(p_{l-1}^l v_{l-1}, p_{l-1}^l w_{l-1}) \\ &= (\mathcal{A}_l p_{l-1}^l v_{l-1}, p_{l-1}^l w_{l-1}) = (r_l^{l-1} \mathcal{A}_l p_{l-1}^l v_{l-1}, w_{l-1})\,,\end{aligned}$$

d. h.: $\mathcal{A}_{l-1} = r_l^{l-1} \mathcal{A}_l p_{l-1}^l$.

iv) *Korrekturschritt:* Im Korrekturschritt wird ein Dämpfungsparameter $\omega_l \in (0, 1]$ verwendet, der im einfachsten Fall $\omega_l = 1$ gesetzt ist. Es hat sich als sehr wirksam erwiesen, ihn so zu wählen, dass der Defekt $\mathcal{A}_l \bar{\bar{v}}_l - \tilde{d}_{l-1}$ minimal wird. Dies führt auf die Formel

$$\omega_l = \frac{(\mathcal{A}_l \bar{\bar{v}}_l, \tilde{d}_{l-1} - \mathcal{A}_l \bar{\bar{v}}_l)}{\|\mathcal{A}_l \bar{\bar{v}}_l\|^2}. \tag{4.2.54}$$

In der folgenden Analyse werden wir der Einfachheit halber stets $\omega_l = 1$ setzen.

4.2.2 Konvergenz- und Aufwandsanalyse

Die klassische Analyse des Mehrgitteralgorithmus basiert auf seiner Interpretation als eine Defektkorrekturiteration und dem Konzept einer rekursiven Anwendung des Zweigitterverfahrens. Zur Vereinfachung nehmen wir an, dass nur Vorglättung angewendet wird (d. h.: $\nu > 0$, $\mu = 0$) und dass im Korrekturschritt keine Dämpfung erfolgt (d. h.: $\omega_l = 1$). Der Zweigitterprozess lässt sich dann in der folgenden Form schreiben:

$$\begin{aligned} u_L^{(t+1)} &= S_L^\nu(u_L^{(t)}) + p_{L-1}^L \mathcal{A}_{L-1}^{-1} r_L^{L-1}\big(f_L - \mathcal{A}_L S_L^\nu(u_L^{(t)})\big) \\ &= S_L^\nu(u_L^{(t)}) + p_{L-1}^L \mathcal{A}_{L-1}^{-1} r_L^{L-1} \mathcal{A}_L\big(u_L - S_L^\nu(u_L^{(t)})\big). \end{aligned}$$

Für den Iterationsfehler $e_L^{(t)} := u_L^{(t)} - u_L$ gilt daher

$$e_L^{(t+1)} = \big(\mathcal{I}_L - p_{L-1}^L \mathcal{A}_{L-1}^{-1} r_L^{L-1} \mathcal{A}_L\big)\big(S_L^\nu(u_L^{(t)}) - u_L\big). \tag{4.2.55}$$

Die Glättungsoperation ist gegeben in der (affin-linearen) Form

$$S_L(v_L) := S_L v_L + g_L$$

und erfüllt als Fixpunktiteration die Bedingung $S_L(u_L) = u_L$. Daraus erschließt man rekursiv, dass

$$S_L^\nu(u_L^{(t)}) - u_L = S_L\big(S_L^{\nu-1}(u_L^{(t)}) - u_L\big) = ... = S_L^\nu e_L^{(t)}.$$

Mit dem sog. „Zweigitteroperator“

$$ZG_L(\nu) := (\mathcal{I}_L - p_{L-1}^L \mathcal{A}_{L-1}^{-1} r_L^{L-1} \mathcal{A}_L)\mathcal{S}_L^\nu$$

gilt daher

$$e_L^{(t+1)} = ZG_L(\nu) e_L^{(t)}\,. \tag{4.2.56}$$

Satz 4.2 (Zweigitterkonvergenz): *Für hinreichend häufige Glättung, $\nu > 0$, ist der Zweigitteralgorithmus konvergent mit einer bzgl. L gleichmäßigen L^2-Konvergenzrate:*

$$\|ZG_L(\nu)\| \le \rho_{ZG}(\nu) = c\,\nu^{-1} < 1\,. \tag{4.2.57}$$

Beweis: Wir schreiben

$$ZG_L(\nu) = (\mathcal{A}_L^{-1} - p_{L-1}^L \mathcal{A}_{L-1}^{-1} r_L^{L-1}) \mathcal{A}_L \mathcal{S}_L^\nu \tag{4.2.58}$$

und schätzen ab:

$$\|ZG_L(\nu)\| \le \|\mathcal{A}_L^{-1} - p_{L-1}^L \mathcal{A}_{L-1}^{-1} r_L^{L-1}\| \|\mathcal{A}_L \mathcal{S}_L^\nu\|\,. \tag{4.2.59}$$

Der erste Term rechts beschreibt die Qualität der Approximation der Feingitterlösung auf dem gröberen Gitter, während der zweite Term den Glättungseffekt enthält. Die Idee für die weitere Analyse ist nun, zu zeigen, dass der Glätter $S_L(\cdot)$ die sog. „Glättungseigen-

schaft“,

$$\|\mathcal{A}_L\mathcal{S}_L^\nu v_L\| \le c_s\nu^{-1}h_L^{-2}\|v_L\|\,, \quad v_L \in V_L\,, \tag{4.2.60}$$

und die Grobgitterkorrektur die sog. „Approximationseigenschaft“ besitzt,

$$\|(\mathcal{A}_L^{-1} - p_{L-1}^L\mathcal{A}_{L-1}^{-1}r_L^{L-1})v_L\| \le c_a h_L^2\|v_L\|\,, \quad v_L \in V_L\,, \tag{4.2.61}$$

mit positiven Konstanten c_s, c_a gleichmäßig bzgl. L. Kombination dieser beiden Abschätzungen ergibt dann die behauptete Ungleichung (4.2.57). Für hinreichend häufige Glättung ist $\rho_{ZG} := c\nu^{-1} < 1$, und der Zweigitteralgorithmus konvergiert gleichmäßig bzgl. L. Alle im Folgenden auftretenden Konstanten sind unabhängig von L.

i) *Glättungseigenschaft:* Der selbstadjungierte Operator $\mathcal{A}_l$ besitzt reelle, positive Eigenwerte $0 < \lambda_1 \le ... \le \lambda_i \le ... \le \lambda_{N_L} =: \Lambda_L$ mit einem zugehörigen L^2-Orthonormalsystem von Eigenfunktionen $\{w^{(1)}, ..., w^{(N_L)}\}$, so dass sich jedes $v_L \in V_L$ in der Form

$$v_L = \sum_{i=1}^{N_L} \gamma_i w^{(i)}, \quad \gamma_i = (v_L, w^{(i)}) \tag{4.2.62}$$

darstellen lässt. Für den Richardson-Iterationsoperator

$$\mathcal{S}_L := \mathcal{I}_L - \theta_L\mathcal{A}_L : V_L \to V_L, \quad \theta_L = \Lambda_L^{-1}, \tag{4.2.63}$$

gilt dann

$$\mathcal{A}_L\mathcal{S}_L^\nu v_L = \sum_{i=1}^{N_L} \gamma_i\lambda_i\Big(1 - \frac{\lambda_i}{\Lambda_L}\Big)^\nu w^{(i)}\,, \tag{4.2.64}$$

und folglich:

$$\begin{aligned}
\|\mathcal{A}_L\mathcal{S}_L^\nu v_L\|^2 &= \sum_{i=1}^{N_L} \gamma_i^2\lambda_i^2\Big(1 - \frac{\lambda_i}{\Lambda_L}\Big)^{2\nu} \\
&\le \Lambda_L^2 \max_{1\le i\le N_L}\Big\{\Big(\frac{\lambda_i}{\Lambda_L}\Big)^2\Big(1 - \frac{\lambda_i}{\Lambda_L}\Big)^{2\nu}\Big\} \sum_{i=1}^{N_L}\gamma_i^2 \\
&= \Lambda_L^2 \max_{1\le i\le N_L}\Big\{\Big(\frac{\lambda_i}{\Lambda_L}\Big)^2\Big(1 - \frac{\lambda_i}{\Lambda_L}\Big)^{2\nu}\Big\} \|v_L\|^2\,.
\end{aligned}$$

Mit Hilfe der Beziehung (Übungsaufgabe)

$$\max_{0\le x\le 1}\{x^2(1-x)^{2\nu}\} \le (1+\nu)^{-2} \tag{4.2.65}$$

ergibt sich

$$\|\mathcal{A}_L\mathcal{S}_L^\nu v_L\|^2 \le \Lambda_L^2(1+\nu)^{-2}\|v_L\|^2\,. \tag{4.2.66}$$

Die Beziehung $\Lambda_L \leq ch_L^{-2}$ liefert dann schließlich die behauptete Ungleichung für den Richardson-Iterationsoperator

$$\|\mathcal{A}_L \mathcal{S}_L^\nu\| \leq c_s \nu^{-1} h_L^{-2}, \quad \nu \geq 1. \tag{4.2.67}$$

ii) *Approximationseigenschaft:* Wir erinnern daran, dass im vorliegenden Kontext geschachtelter FE-Räume Prolongationen und Restriktionen gegeben sind durch

$$p_{L-1}^L = id. \text{ (Identität)}, \quad r_L^{L-1} = P_{L-1} \ (L^2\text{-Projektion}).$$

Ferner erfüllt der Operator $\mathcal{A}_L : V_L \to V_L$ definitionsgemäß

$$(\mathcal{A}_L v_L, \varphi_L) = a(v_L, \varphi_L), \quad v_L, \varphi_L \in V_L.$$

Für ein beliebiges, aber fest gewähltes $f_L \in V_L$ gilt demnach für die Funktionen $v_L := \mathcal{A}_L^{-1} f_L$ und $v_{L-1} := \mathcal{A}_{L-1}^{-1} r_L^{L-1} f_L$:

$$\begin{aligned} a(v_L, \varphi_L) &= (f_L, \varphi_L) \quad \forall \varphi_L \in V_L, \\ a(v_{L-1}, \varphi_{L-1}) &= (f_L, \varphi_{L-1}) \quad \forall \varphi_{L-1} \in V_{L-1}. \end{aligned}$$

Der Funktion $v_L \in V_L$ ordnen wir eine Funktion $v \in V \cap H^2(\Omega)$ zu als Lösung der Randwertaufgabe

$$Lv = f_L \ \text{ in } \Omega, \quad v = 0 \ \text{ auf } \partial\Omega, \tag{4.2.68}$$

bzw. in „schwacher“ Formulierung

$$a(v, \varphi) = (f_L, \varphi) \quad \forall \varphi \in V\,. \tag{4.2.69}$$

Dafür gilt die a priori Abschätzung

$$\|\nabla^2 v\| \leq c\|f_L\|. \tag{4.2.70}$$

Dann ist

$$\begin{aligned} a(v_L, \varphi_L) &= (f_L, \varphi_L) = a(v, \varphi_L), \quad \varphi_L \in V_L, \\ a(v_{L-1}, \varphi_{L-1}) &= (f_L, \varphi_{L-1}) = a(v, \varphi_{L-1}), \quad \varphi_{L-1} \in V_{L-1}, \end{aligned}$$

d. h.: v_L und v_{L-1} sind gerade die Ritz-Projektionen von v auf V_L bzw. V_{L-1}. Für diese gelten die L^2-Fehlerabschätzungen

$$\|v_L - v\| \leq ch_L^2 \|\nabla^2 v\|, \quad \|v_{L-1} - v\| \leq ch_{L-1}^2 \|\nabla^2 v\|. \tag{4.2.71}$$

Damit erhalten wir wegen $h_{L-1} \leq 4h_L$ und der a priori Abschätzung (4.2.70):

$$\|v_L - v_{L-1}\| \leq ch_L^2 \|\nabla^2 v\| \leq ch_L^2 \|f_L\|\,. \tag{4.2.72}$$

Dies bedeutet mit der obigen Setzung, dass

$$\|\mathcal{A}_L^{-1} f_L - p_{L-1}^L \mathcal{A}_{L-1}^{-1} r_L^{L-1} f_L\| \leq c h_L^2 \|f_L\| \,. \tag{4.2.73}$$

Damit folgt die gewünschte Abschätztung

$$\|\mathcal{A}_L^{-1} - p_{L-1}^L \mathcal{A}_{L-1}^{-1} r_L^{L-1}\| \leq c h_L^2 \,. \tag{4.2.74}$$

Dies vervollständigt den Beweis. Q.E.D.

Das Resultat für den Zweigitteralgorithmus wird nun verwendet zum Nachweis der Konvergenz des vollen Mehrgitteralgorithmus.

Satz 4.3 (Mehrgitterkonvergenz): *Es sei angenommen, dass der Zweigitteralgorithmus konvergiert mit einer L^2-Konvergenzrate $\rho_{ZG}(\nu) \to 0$ für $\nu \to \infty$, gleichmäßig bzgl. L. Dann konvergiert für hinreichend häufige Glättung der Mehrgitteralgorithmus mit $R \geq 2$ (W-Zyklus) mit einer von L unabhängigen L^2-Konvergenzrate $\rho_{MG} < 1$,*

$$\|u_L - MG(L, u_L^{(t)}, f_L)\| \leq \rho_{MG} \, \|u_L - u_L^{(t)}\|. \tag{4.2.75}$$

Beweis: Der Beweis wird durch Induktion nach dem Gitterlevel L geführt. Wir betrachten nur den relevanten Fall $R = 2$ (W-Zyklus) und werden uns der Einfachheit halber nicht bemühen, die auftretenden Konstanten zu optimieren. Sei ν so groß, dass die Konvergenzrate des Zweigitteralgorithmus $\rho_{ZG} \leq \frac{1}{8}$ ist. Wir wollen zeigen, dass dann die Konvergenzrate des Mehrgitteralgorithmus $\rho_{MG} \leq \frac{1}{4}$ ist, gleichmäßig bzgl. L. Für $L = 1$ ist dies dann offenbar richtig. Sei nun auch $\rho_{MG} \leq \frac{1}{4}$ für Gitterlevel $L-1$. Auf Gitterlevel L gilt dann ausgehend von der Iterierten $u_L^{(t)}$ mit der approximativen Lösung $q_{L-1}^{(2)}$ (nach 2-maliger Anwendung der Grobgitterkorrektur) und der exakten Lösung $\hat{q}_{L-1}$ der Defektgleichung auf Level $L-1$:

$$\begin{aligned} u_L^{(t+1)} &= MG(L, u_L^{(t)}, f_L) = S_L^\nu(u_L^{(t)}) + p_{L-1}^L q_{L-1}^{(2)} \\ &= S_L^\nu(u_L^{(t)}) + p_{L-1}^L \hat{q}_{L-1} + p_{L-1}^L (q_{L-1}^{(2)} - \hat{q}_{L-1}) \\ &= ZG(L, u_L^{(t)}, f_L) + p_{L-1}^L (q_{L-1}^{(2)} - \hat{q}_{L-1}) \end{aligned} \tag{4.2.76}$$

Nach Induktionsvoraussetzung ist (Man beachte, dass der Startwert der Mehrgitteriteration auf Level $L-1$ gleich Null ist und $\hat{\rho}_{L-1} = \mathcal{A}_{L-1}^{-1} r_L^{L-1} d_L$):

$$\|\hat{q}_{L-1} - q_{L-1}^{(2)}\| \leq \rho_{MG}^2 \, \|\hat{q}_{L-1}\| = \rho_{MG}^2 \, \|\mathcal{A}_{L-1}^{-1} r_L^{L-1} \mathcal{A}_L \mathcal{S}_L^\nu (u_L - u_L^{(t)})\| \,. \tag{4.2.77}$$

Kombination der letzten Beziehungen ergibt für den Iterationsfehler $e_L^{(t)} := u_L^{(t)} - u_L$:

$$\|e_L^{(t+1)}\| \leq \left(\rho_{ZG} + \rho_{MG}^2 \, \|\mathcal{A}_{L-1}^{-1} r_L^{L-1} \mathcal{A}_L \mathcal{S}_L^\nu\|\right) \|e_L^{(t)}\| \,. \tag{4.2.78}$$

Die Norm rechts ist bereits im Zusammenhang mit der Konvergenz des Zweigitteralgorithmus abgeschätzt worden. Mit dem Zweigitteroperator $ZG_L = (\mathcal{A}_L^{-1} - p_{L-1}^L \mathcal{A}_{L-1}^{-1} r_L^{L-1}) \mathcal{A}_L \mathcal{S}_L^\nu$

gilt

$$\mathcal{A}_{L-1}^{-1} r_L^{L-1} \mathcal{A}_L \mathcal{S}_L^\nu = S_L^\nu - (\mathcal{A}_L^{-1} - p_{L-1}^L \mathcal{A}_{L-1}^{-1} r_L^{L-1}) \mathcal{A}_L \mathcal{S}_L^\nu = S_L^\nu - Z G_L$$

und somit

$$\|\mathcal{A}_{L-1}^{-1} r_L^{L-1} \mathcal{A}_L \mathcal{S}_L^\nu\| \leq \|S_L^\nu\| + \|Z G_L\| \leq 1 + \rho_{ZG} \leq 2\,. \tag{4.2.79}$$

Damit erhalten wir schließlich

$$\|e_L^{(t+1)}\| \leq \left(\rho_{ZG} + 2\rho_{MG}^2\right) \|e_L^{(t)}\|\,. \tag{4.2.80}$$

Mit Hilfe der Annahme über ρ_{ZG} und der Induktionsannahme folgt

$$\|e_L^{(t+1)}\| \leq \left(\tfrac{1}{8} + 2\tfrac{1}{16}\right) \|e_L^{(t)}\| \leq \tfrac{1}{4} \|e_L^{(t)}\|\,, \tag{4.2.81}$$

was den Induktionsbeweis vervollständigt. Q.E.D.

Für „gutartige“ Probleme (symmetrischer, positiv definiter Operator, glatte Koeffizienten, quasi-gleichförmige Gitter u.s.w.) erreicht man in der Regel Mehrgitterkonvergenzraten im Bereich $\rho_{MG} = 0,1 - 0,3$. Die obige Analyse ist nur für den W-Zyklus gültig, da im Beweisteil (ii) $R \geq 2$ benötigt wird. Der V-Zyklus kann nicht auf Basis nur einer Zweigitteranalyse behandelt werden. In der Literatur finden sich allgemeinere Ansätze, die Konvergenz von Mehrgitterverfahren auch in weniger regulären Situationen garantieren.

Als nächstes diskutieren wir die numerische Komplexität des Mehrgitteralgorithmus. Dabei werden die folgenden Bezeichnungen verwendet:

$$\begin{aligned}
OP(T) &:= \text{Anzahl der a.Op. zur Durchführung einer Operation } T,\\
R &:= \text{Anzahl der Defektkorrekturschritte auf den einzelnen Gitterniveaus,}\\
N_l &:= \dim V_l \approx h_l^{-d} \ (d = \text{Raumdimension}),\\
\kappa &:= \max_{1\leq l\leq L} N_{l-1}/N_l < 1,\\
C_0 &:= OP(\mathcal{A}_0^{-1})/N_0,\\
C_s &:= \max_{1\leq l\leq L}\{OP(\mathcal{S}_l)/N_l\}, \quad C_d := \max_{1\leq l\leq L}\{OP(d_l)/N_l\},\\
C_r &:= \max_{1\leq l\leq L}\{OP(r_l)/N_l\}, \quad C_p := \max_{1\leq l\leq L}\{OP(p_l)/N_l\}.
\end{aligned}$$

In der Praxis ist meist $\kappa \approx 2^{-d}$, $C_s \approx C_d \approx C_r \approx C_p \approx \#\{a_{nm} \neq 0\}$ und $C_0 N_0 \ll N_L$.

Satz 4.4 (Mehrgitterkomplexität): *Unter der Bedingung $q := R\kappa < 1$ gilt für einen Mehrgitterzyklus MG_L:*

$$OP(MG_L) \leq C_L N_L \tag{4.2.82}$$

mit

$$C_L = \frac{(\nu+\mu)C_s + C_d + C_r + C_p}{1-q} + C_0 q^L,$$

Der Mehrgitteralgorithmus liefert die N_L-dimensionale diskrete Lösung $u_L \in V_L$ auf dem Gitter $\mathbb{T}_L$ im Rahmen der Diskretisierungsgenauigkeit $\mathcal{O}(h_L^2)$ bzgl. der L^2-Norm mit $\mathcal{O}(N_L \ln(N_L))$ a.Op. und hat damit (fast) optimale Komplexität.

Beweis: Wir setzen $C_l := OP(MG_l)/N_l$. Ein Mehrgitterschritt beinhaltet die R-fache Anwendung desselben Algorithmus auf dem nächst gröberen Gitter. Bei Beachtung von $N_{l-1} \leq \kappa N_l$ gilt mit $\hat{C} := (\nu + \mu)C_s + C_d + C_r + C_p$:

$$C_L N_L = OP(MG_L) \leq \hat{C} N_L + R \cdot OP(MG_{L-1}) = \hat{C} N_L + R \cdot C_{L-1} N_{L-1} \leq \hat{C} N_L + q C_{L-1} N_L,$$

und folglich $C_L \leq \hat{C} + q C_{L-1}$. Rekursive Anwendung dieser Beziehung liefert

$$C_L \leq \hat{C}(1 + q + q^2 + ... + q^{L-1}) + q^L C_0 \leq \frac{\hat{C}}{1-q} + q^L C_0.$$

Dies impliziert die behauptete Abschätzung (4.2.82). Die Komplexität des Gesamtalgorithmus ergibt sich dann aus den Beziehungen

$$\rho_{MG}^t \approx h_L^2 \approx N_L^{-2/d}, \qquad t \approx -\frac{\ln(N_L)}{\ln(\rho_{MG})}.$$

Dies vervollständigt den Beweis. Q.E.D.

Wir bemerken, dass im Beweis der Aussage (4.2.82) die Bedingung

$$q := R\kappa = R \max_{1 \leq l \leq L} N_{l-1}/N_l < 1$$

wesentlich ist. Dies besagt für den W-Zyklus ($R = 2$), dass sich beim Übergang vom Gitter $\mathbb{T}_{l-1}$ zum nächst feineren $\mathbb{T}_l$ die Anzahl der Gitterpunkte (bzw. Freiheitsgrade) hinreichend stark erhöhen muss, etwa wie bei einer gleichförmigen Verfeinerung

$$N_l \approx 4 N_{l-1}.$$

Bei einem adaptiv gesteuerten Verfeinerungsprozess mit teilweise nur lokaler Gitterverfeinerung ist dies meist nicht erfüllt; selbst bei Verwendung der „Fest-Raten"-Strategie ist z. B. oft nur $N_l \approx 2N_{l-1}$. In solchen Fällen muss der Mehrgitterprozess zur Aufwandsersparnis modifiziert werden. Dies geschieht dadurch, dass die kostenintensive Glättung sowie die anderen Operationen nur jeweils auf den beim Übergang von $\mathbb{T}_{l-1}$ zu $\mathbb{T}_l$ neu hinzugekommenen Gitterpunkten durchgeführt werden. Bei der Implementierung eines Mehrgitteralgorithmus auf lokal verfeinerten Gittern ist viel Fingerspitzengefühl erforderlich, wenn der resultierende Gesamtalgorithmus komplexitäts-optimal sein soll.

Für das geschachtelte MG-Schema erhält man sogar im strengen Sinne „optimale" Lösungskomplexität $O(N_L)$, da auf jedem Gitterniveau bestmögliche Startwerte verwendet werden.

Satz 4.5 (Geschachteltes Mehrgitterverfahren): *Das geschachtelte MG-Schema ist*

komplexitäts-optimal, d.h.: Es liefert die diskrete Lösung $u_L \in V_L$ *auf dem feinsten Gitter* $\mathbb{T}_L$ *im Rahmen der Diskretisierungsgenauigkeit* $\mathcal{O}(h_L^2)$ *bzgl. der* L^2*-Norm mit einem Aufwand von* $\mathcal{O}(N_L)$ *a.Op.*

Beweis: Die Genauigkeitsanforderung für die Mehrgitteriteration auf Gitterlevel $\mathbb{T}_L$ ist

$$\|e_L^{(t)}\| \le \hat{c} h_L^2 \|f\|. \tag{4.2.83}$$

i) Wir wollen zunächst zeigen, dass (4.2.83) beim geschachtelten MG-Schema unter den Voraussetzungen des Mehrgitterkonvergenzsatzes 4.3 auf jedem Level L mit einer (hinreichend großen) festen Zahl t_* von Mehrgitterschritten erreichbar ist. Sei $e_L^{(t)} := u_L^{(t)} - u_L$ wieder der Iterationsfehler auf Level L. Nach Annahme ist $e_0^{(t)} = 0, t \ge 1$. Im Fall $u_L^{(0)} := u_{L-1}^{(t)}$ gilt dann

$$\begin{aligned}
\|e_L^{(t)}\| &\le \rho_{MG}^t \|e_L^{(0)}\| = \rho_{MG}^t \|u_{L-1}^{(t)} - u_L\| \\
&\le \rho_{MG}^t \big(\|u_{L-1}^{(t)} - u_{L-1}\| + \|u_{L-1} - u\| + \|u - u_L\| \big) \\
&\le \rho_{MG}^t \big(\|e_{L-1}^{(t)}\| + c h_L^2 \|f\| \big).
\end{aligned}$$

Rekursive Anwendung dieser Beziehung für $L \ge l \ge 1$ ergibt dann (wegen $h_l \le \kappa^{l-L} h_L$)

$$\begin{aligned}
\|e_L^{(t)}\| &\le \rho_{MG}^t \big(\rho_{MG}^t \big(\|e_{L-2}^{(t)}\| + c h_{L-1}^2 \|f\| \big) + c h_L^2 \|f\| \big) \\
&\quad \vdots \\
&\le \rho_{MG}^{Lt} \|e_0^{(t)}\| + \big(c \rho_{MG}^t h_L^2 + c \rho_{MG}^{2t} h_{L-1}^2 + ... + c \rho_{MG}^{Lt} h_1^2 \big) \|f\| \\
&= c h_L^2 \kappa^2 \left(\rho_{MG}^t \kappa^{-2\cdot 1} + \rho_{MG}^{2t} \kappa^{-2\cdot 2} + ... + \rho_{MG}^{Lt} \kappa^{-2L} \right) \|f\| \\
&\le c h_L^2 \kappa^2 \|f\| \, \frac{\kappa^{-2} \rho_{MG}^t}{1 - \kappa^{-2} \rho_{MG}^t},
\end{aligned}$$

vorausgesetzt $\kappa^{-2} \rho_{MG}^t < 1$. Offenbar gibt es also ein t_*, so dass (4.2.83) für $t \ge t_*$ erfüllt ist, und zwar gleichmäßig bzgl. L.

ii) Wir kommen nun zur Aufwandsanalyse. Satz 4.4 besagt, dass ein Zyklus des „einfachen“ Mehrgitteralgorithmus $MG(l, \cdot, \cdot)$ auf dem l-ten Level $W_l \le c_* N_l$ a.Op. benötigt (gleichmäßig bzgl. l). Sei nun $\hat{W}_l$ die Anzahl der a.Op. des geschachtelten Schemas auf Gitterlevel l. Dann gilt konstruktionsgemäß:

$$\hat{W}_L \le \hat{W}_{L-1} + t_* W_L.$$

Durch Iteration dieser Beziehung erhalten wir mit $\kappa := \max_{1 \le l \le L} N_{l-1}/N_l < 1$:

$$\hat{W}_L \le t_* c_* \{ N_L + ... + N_0 \} \le c t_* c_* N_L \{ 1 + ... + \kappa^L \} \le \frac{c t_* c_*}{1 - \kappa} N_L,$$

was zu beweisen war. Q.E.D.

4.3 Übungen

Übung 4.1: Das allgemeine „Abstiegsverfahren“ zur iterativen Lösung des Gleichungssystems $Ax = b$ mit symmetrischer, positiv-definiter Matrix $A \in \mathbb{R}^{N\times N}$ lautet:

$$\begin{aligned}
&\text{Startwert:} && x^{(0)} \in \mathbb{R}^n, \quad r^{(0)} := b - Ax^{(0)}\,,\\
&\text{für } t \geq 0: && \text{Abstiegsrichtung} \quad d^{(t)}\,,\\
& && \alpha_t = \frac{\langle r^{(t)}, d^{(t)}\rangle}{\langle Ad^{(t)}, d^{(t)}\rangle}\,,\\
& && x^{(t+1)} = x^{(t)} + \alpha_t d^{(t)}\,, \quad r^{(t+1)} = r^{(t)} - \alpha_t A d^{(t)}\,.
\end{aligned}$$

Die sog. „Koordinatenrelaxation“ erhält man durch zyklische Wahl der Abstiegsrichtungen $d^{(t)}$ aus den kartesischen Einheitsvektoren $\{e^{(1)}, \ldots, e^{(N)}\}$. Man zeige, dass jeder N-Zyklus der Koordinatenrelaxition äquivalent ist zum üblichen Gauß-Seidel-Verfahren.
Bemerkung: Für eine typische FE-Matrix hat die zyklische Koordinatenrelaxation also das Konvergenzverhalten:

$$|x^{(tN)} - x| \leq cq^t, \qquad q \approx 1 - \mathrm{cond}_2(A)^{-1} \approx 1 - h^2.$$

Übung 4.2: Die erste RWA des Laplace-Operators

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

auf einem „regulären” Gebiet $\Omega \subset \mathbb{R}^2$ werde mit einem FE-Verfahren mit stückweise linearen Ansatzfunktionen auf einer Folge von quasi-gleichförmigen Gittern (d. h. größen- und form-regulär) der Weite h diskretisiert. Dies führt auf lineare Gleichungssysteme $Ax = b$ mit symmetrischen, positiv definiten $(N\times N)$-Matrizen A, wobei N die Anzahl der Knotenpunkte ist.

Welchen arithmetischen Aufwand (ausgedrückt in Potenzen von h) erfordert dabei die Lösung dieser Gleichungssysteme mit dem CG-Verfahren mit der Genauigkeit des Diskretisierungsfehlers gemessen in der „Energie-Norm” $\|\nabla(u - u_h)\|$? Dazu verwende man die folgende bekannte Fehlerabschätzung für das CG-Verfahren:

$$|x - x^t|_A \leq 2q^t |x - x^0|_A, \qquad q := \frac{1 - 1/\sqrt{\kappa}}{1 + 1/\sqrt{\kappa}},$$

mit der diskreten „Energienorm” $\|x\|_A := (Ax, x)^{1/2}$ und der Spektralkondition $\kappa := \kappa_2(A)$ von A.
Hinweis: Man verwende die bekannte Beziehung für die Spektralkondition von A sowie die aus der obigen Fehlerabschätzung abgeleitete Abschätzung für die Anzahl der Iterationsschritte. Der Aufwand pro CG-Schritt entspricht etwas der zweimaligen Defektberechnung $x \to d := Ax - b$.

Übung 4.3: Man versuche, den Beweis der Konvergenz des Zweigitterverfahrens ZG aus dem Text für den Fall zu verallgemeinern, dass die Restriktion $r_l^{l-1} : V_l \to V_{l-1}$ mit Hilfe

lokaler, bilinearer Interpolation (anstelle der L^2-Projektion) auf dem Gitter $\mathbb{T}_{l-1}$ definiert ist. Wo ist dabei das Problem, und wie kann man damit fertig werden?

Übung 4.4: Die FE-Diskretisierung des Konvektions-Diffusionsproblems

$$-\Delta u + \partial_1 u = f \quad \text{in } \Omega, \quad u = 0 \text{ auf } \partial\Omega,$$

führt auf unsymmetrische Systemmatrizen A_h. In diesem Fall erfordert die Analyse des Mehrgitterverfahrens einige Modifikationen. Man übertrage den Beweis aus dem Text für die Konvergenz des Zweigitteralgorithmus, wenn als Glätter wieder das Richardson-Verfahren

$$x_h^{t+1} = x_h^t - \theta_t(A_h x_h^t - b_h), \quad t = 0, 1, 2, \ldots,$$

mit den Dämpfungsparametern $\theta_t := \frac{1}{2}\|A_h\|^{-1}$ verwendet wird.

Übung 4.5: Zur Lösung der 1. RWA der Laplace-Gleichung

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

auf dem Einheitsquadrat $\Omega = (0,1)^2$ werde auf einer Folge äquidistanter, kartesischer Gitter $\mathbb{T}_l$ mit Gitterweiten $h_l = 2^{-l}$ mit Hilfe bilinearer finiter Elemente approximiert. Die diskrete Gleichung auf Gitterlevel l werde dabei mit einem MG-Verfahren gelöst, wobei das Richardson-Verfahren zur Glättung, die natürliche Einbettung zur Prolongation und die lokale bilineare Interpolation zur Restriktion verwendet werden. Die Anzahl der Vor- und Nachglättungsschritte sei $\nu = 2$ und $\mu = 0$. Wieviele a. Op. kosten dann ungefähr ein V-Zyklus und ein W-Zyklus ausgedrückt in Vielfachen der Dimension $N_l = \dim V_l$?

Übung 4.6: Die erste RWA des Laplace-Operators

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

auf einem „regulären" Gebiet $\Omega \subset \mathbb{R}^2$ werde mit einem FE-Verfahren mit stückweise linearen Ansatzfunktionen auf einer Folge von quasi-gleichförmigen Gittern (d. h. größen- und form-regulär) der Weite h diskretisiert. Dies führt auf lineare Gleichungssysteme $Ax = b$ mit symmetrischen, positiv definiten $(N \times N)$-Matrizen A, wobei N die Anzahl der Knotenpunkte ist. Das „Richardson-Verfahren" iteriert zur Lösung des Gleichungssystems $Ax = b$ ausgehend von einem Startwert $x^0 \in \mathbb{R}^N$ mit einem Dämpfungsparameter $\theta \in \mathbb{R}$ gemäß

$$x^{t+1} = x^t - \theta(Ax^t - b), \quad t \in \mathbb{N}_0.$$

Im Falle, dass A nur reelle positive Eigenwerte $0 < \lambda_{\min} \leq \cdots \leq \lambda_{\max}$ besitzt, ist der Spektralradius der Iterationsmatrix $B_\theta = I - \theta A$ gegeben durch

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

Für welches θ wird $\rho(B_\theta)$ minimal, d. h. konvergiert die Iteration am besten, und für welches θ hat die Iteration die beste Glättungseigenschaft?

Übung 4.7: Man beschäftige sich mit den folgenden Fragen:

a) Wie lautet die variationelle („schwache“) Formulierung der Randwertaufgabe

$$-\nabla \cdot (a\nabla u) + bu = f \text{ in } \Omega, \quad u|_{\partial\Omega} = 0,$$

auf einem konvexen Polyeder $\Omega \subset \mathbb{R}^3$ und unter welchen Bedingungen an die Koeffizientenfunktionen $a \in C^1(\overline{\Omega})$ und $b \in C(\overline{\Omega})$ ist diese „wohl gestellt“.

b) Die Randwertaufgabe in a) werde durch einen konformen „quadratischen“ Finite-Elemente-Ansatz mit Gitterweite $h \in \mathbb{R}_+$ diskretisiert. Man beschreibe die einzelnen Schritte (Gitter, Knotenbasis, Systemmatrix) zur Aufstellung der zugehörigen algebraischen Gleichungssysteme

$$A_h x_h = b_h.$$

c) Man gebe für die Diskretisierung in b) optimale a priori Fehlerabschätzungen in der Energie- und der L^2-Norm an. Wie hängt in diesem Fall die Kondition der Systemmatrix A_h von der Gitterweite h ab?

d) Man formuliere i) das Gauß-Seidel-Verfahren und ii) das Gradienten-Verfahren zur iterativen Lösung des Gleichungssystems in b). Wieviele Iterationsschritte sind mit diesen Verfahren in Abhängigkeit von der Anzahl der Unbekannten $N_h := \dim V_h$ notwendig, um den Anfangsfehler um 10^{-3} zu reduzieren?