Kurzzusammenfassung zur Vorlesung „Numerik 1“

M. Schedensack

Stand: 5. September 2023

1 Einleitung

In diesem Kapitel wird die partielle Differentialgleichungen $-\Delta u + u^2 = f$ in $\Omega := (0,1)^2$ mit Randbedingungen $u|_{\partial\Omega} = 0$ eingeführt und die Galerkin-Approximation in

$$V_h := \left\{ v_h \in C(\Omega) \;\middle|\; \begin{array}{l} \forall j,k \in \{0,\dots,n-1\}\ \exists a_{jk}, b_{jk}, c_{jk}, d_{jk} \in \mathbb{R} : \\ v_h|_{T_{jk}}(x,y) = a_{jk}xy + b_{jk}x + c_{jk}y + d_{jk} \end{array} \right\},$$

wobei die Quadrate T_{jk} definiert sind als $T_{jk} := [x_j, x_{j+1}] \times [x_k, x_{k+1}]$ mit $x_j = j/n$ für $0 \le j \le n$. Zu Basisfunktionen φ_{jk} von V_h, die $\varphi_{jk}(x_\ell, x_m) = \delta_{j\ell}\delta_{km}$ erfüllen, wurde gezeigt, dass $u_h \in V_h$ die Galerkin-Approximation ist, genau dann, wenn der Koeffizientenvektor ein Gleichungssystem $Ax + B(x) = b$ löst. Die Steifigkeitsmatrix A wurde in der Vorlesung berechnet.

2 Lineare Gleichungssysteme: Gauß-Elimination und LR-Zerlegung

In diesem Kapitel wird für eine reguläre Matrix $A \in \mathbb{K}^{n\times n}$ und einen Vektor $b \in \mathbb{K}^n$ das lineare Gleichungssystem $Ax = b$ betrachtet. Durch Zeilenvertauschung (Permutationen) und das Subtrahieren von Spalten kann das lineare Gleichungssystem gelöst werden. Das Subtrahieren von ℓ_{ik} mal der k-ten Zeile von der i-ten Zeile kann über Frobenius-Matrizen dargestellt werden. Insgesamt zeigt dies, dass es eine Permutationsmatrix Q gibt, eine untere (linke) Dreiecksmatrix L mit 1en auf der Diagonalen und eine obere rechte Dreiecksmatrix R derart, dass

$$QA = LR.$$

Sind L und R berechnet, kann das lineare Gleichungssystem einfach durch Vorwärts- bzw. Rückwärtssubstitution gelöst werden. Der benötigte Speicherplatz für die LR-Zerlegung ist nur einer mehr als für die Speicherung einer vollbesetzten Matrix A. Der Rechenaufwand ist $(2/3)n^3$ (bis auf höhere Terme).

Begriffe und Sätze, die bekannt sein sollten: LR-Zerlegegung, Pivot-Element, Permutationsmatrix, Frobenius-Matrix, Vorwärts- und Rückwärtssubstitution, Rechenaufwand der LR-Zerlegegung

Literatur: [DH19, S. 3–12], [Ran17, S. 109–115], [Her20, S. 37–59]

3 Fehleranalyse und Grundlagen

Gleitkommazahlen. Im endlichen Speicherplatz des Computers werden reelle Zahlen durch Gleitkommazahlen angenähert. Die Eingabe von Daten, sowie die Ausführung von Maschinenoperationen (als Approximation der Grundoperationen $+$, $-$, $\cdot$, $/$) erfordern Rundungen auf Maschinenzahlen. Der relative Fehler der Rundung ist (wenn kein Exponentenunterlauf vorliegt) durch die Maschinengenauigkeit beschränkt.

Begriffe und Sätze, die bekannt sein sollten: Gleitkommazahlen, Mantisse, Mantissenlänge, Maschinenzahlen, Rundung, Maschinengenauigkeit, Maschinenoperationen,

Literatur: [Her20, Kap. 1.3], [Ran17, Kap. 1.1]

Normierte Räume. Im zweiten Teil des Kapitels werden Normen auf dem Raum der Vektoren und der Matrizen definiert. Eine wichtige Rolle spielen Operatornormen von Matrizen. Es wird gezeigt, wie die Operatornormen zur ℓ^1-, ℓ^2- und ℓ^∞-Norm aussehen.

Begriffe und Sätze, die bekannt sein sollten: Norm, ℓ^p-Norm, Frobenius-Norm, Matrizennorm, allgemeine Operatornorm von Matrizen, Operatornormen bzgl. der ℓ^1-, ℓ^2- und ℓ^∞-Normen, Skalarprodukt,

Literatur: [Ran17, Kap. 4.1.1–4.1.2]

Konditionierung numerischer Aufgaben. Im dritten Teil des Kapitels wird die Kondition eines Problems betrachet. Die (relativen) Konditionszahlen beschreiben die Gutartigkeit des Problems für kleine Abweichungen der Eingangsdaten. Die Subtraktion zweier fast gleich großer Zahlen führt zur Auslöschung und ist schlecht konditionert. Der Begriff der Kondition einer Matrix wird definiert und es wird ein Störungssatz für Matrizen bewiesen, der besagt, dass der relative Fehler beim Lösen eines linearen Gleichungssystems im wesentlichen durch die Kondition mal den relativen Fehlern in der Matrix und der rechten Seite beschränkt ist.

Begriffe und Sätze, die bekannt sein sollten: absoluter und relativer Fehler, Konditionszahl von Funktionen, gut und schlecht konditionierte Probleme, Kondition einer Matrix, Störungssatz für lineare Gleichungssysteme

Literatur: [DH19, Kap. 2.2.1] [Ran17, Satz 4.1]

Stabilität numerischer Algorithmen. Die Stabilität eines numerischen Algorithmus gibt an, wie Rundungsfehler innerhalb eines numerischen Algorithmus verstärkt werden. Die Maschinenoperationen sind stabil. Werden zwei Operationen hintereinander ausgeführt, hängt die Stabilität des gesamten Algorithmus von der Stabilität der beiden Operationen und auch von der Kondition der ersten Operation ab. Dies führt auf die Regel, dass schlecht konditionierte Operationen möglichst frühzeitig ausgeführt werden sollten. Für Polynome sollte die Auswertung möglichst mit dem Horner-Schema erfolgen.

Begriffe und Sätze, die bekannt sein sollten: Stabilitätsindikator, Stabilität von hintereinander ausgeführten Operationen, Horner-Schema

Literatur: [DH19, Kap. 2.3.2], [Ran17, Kap 1.3.3]

4 Lineare Gleichungssysteme, Teil 2

Lineare Ausgleichsrechnung und QR-Zerlegung. Die Kleinste-Quadrate-Lösung einer im Allgemeinen überbestimmten Gleichungssystems ist genau die Lösung der Gaußschen Normalgleichung. Da die Matrix $A^\star A$ allerdings im Allgemeinen eine große Kondition hat, ist es häufig nicht sinnvoll, diese Lösung mit der LR-Zerlegung der Matrix $A^\star A$ zu berechnen. Stattdessen kann die Matrix A durch Householder-Transformationen in eine verallgemeinerte rechte obere Dreiecksmatrix überführt werden. Diese Zerlegung heißt QR-Zerlegegung. Der Rechenaufwand hierfür beträgt $(4/3)n^3$ (bis auf höhere Ordnungsterme) für quadratische Matrizen. Mithilfe der QR-Zerlegung kann die Kleinste-Quadrate-Lösung leicht berechnet werden. Außerdem kann auch das Residuum berechnet werden.

Begriffe und Sätze, die bekannt sein sollten: Kleinste-Quadrate-Lösung, Gaußsche Normalgleichung, Kondition der Matrix $A^\star A$, orthogonale Matrizen, Householder-Transformation, QR-Zerlegegung, Rechenaufwand der QR-Zerlegegung, Residuum

Literatur: [Ran17, S. 131–], [Bar16, Kap. 5.2–5.4]

Klassische iterative Verfahren. Der Spektralradius einer Matrix A ist ein Maß dafür, ob und wie schnell A^k für alle Startvektoren gegen 0 konvergiert. Für die klassischen iterativen Verfahren benannt nach Richardson, Jacobi und Gauß-Seidel, kann das lineare Gleichungssystem $Ax = b$ als Fixpunktgleichung umgeschrieben werden. Falls die zugehörigen Iterationsmatrizen einen Spektralradius kleiner als 1 haben, konvergiert das jeweilige Verfahren. Ist A strikt diagonaldominant oder ist A diagonaldominant und irreduzibel, dann konvergieren das Jacobi und das Gauß-Seidel-Verfahren.

Begriffe und Sätze, die bekannt sein sollten: Spektralradius, Zusammenhang von Spektralradius und Operatornormen, Spektralradius als Maß der Konvergenzgeschwindigkeit, Richardson-Verfahren, Jacobi-Verfahren, Gauß-Seidel-Verfahren, Diagonaldominanz und strikte Diagonaldominanz, Irreduziblität, Gerschgorinscher Kreissatz, Konvergenz der klassischen Iterationsverfahren,

Literatur: [Her20, Kap. 2.6.2], [Bar16, Kap. 9.3], [Bar16, Satz 8.1]

Das cg-Verfahren. Ist die Matrix A in einem linearen Gleichungssystem symmetrisch und positiv definit, ist das Lösen von $Ax = b$ äquivalent zur Minimierung des Energiefunktionals

$$f(z) = \frac{1}{2} z^\top A z - z^\top b.$$

Dies motiviert Abstiegsverfahren, die in jedem Schritt entlang von einer Abstiegsrichtung das Funktional minimieren. Im Gradientenverfahren wird als Abstiegsrichtung einfach der Gradient der Funktion f gewählt. Sind die Abstiegsrichtungen paarweise A-orthogonal, so minimiert die Folge im Abstiegsverfahren das Energiefunktional schon über dem ganzen aufgespannten Raum und ist damit ein Galerkin-Verfahren. Insbesondere berechnet ein solches Verfahren nach höchstens n Schritten die exakte Lösung (unter exakter Rechenarithmetik). Die Abstiegsrichtungen, die durch das cg-Verfahren berechnet werden, sind A-orthogonal und spannen die Krylov-Räume auf. Da das cg-Verfahren ein Galerkin-Verfahren ist, folgt eine Fehlerabschätzung des

Fehlers im k-ten Schritt gegen die beste Approximation im entsprechenden Krylov-Raum. Dieser Fehler kann (mithilfe von Resultaten aus dem folgenden Kapitel) durch

$$2\left(\frac{\sqrt{\kappa_A}-1}{\sqrt{\kappa_A}+1}\right)^k \|x^\star - x_0\|$$

beschränk werden.

Begriffe und Sätze, die bekannt sein sollten: Minimierung des Energiefunktionals, Abstiegsverfahren, Gradientenverfahren, A-orthogonale Abstiegsrichtungen, Galerkin-Verfahren, cg-Verfahren, Krylov-Räume, Fehlerabschätzung für das cg-Verfahren

Literatur: [Bar16, Kap. 16]

Tschebyscheff-Polynome. Die Tschebyscheff-Knoten sind die Nullstellen der Tschebyscheff-Polynome. Sie lösen das minmax-Problem

$$\min_{(x_0,\dots,x_{n-1})\in\mathbb{R}^n} \max_{x\in[-1,1]} \Pi_{j=0}^{n-1}(x-x_j) = \max_{x\in[-1,1]} \Pi_{j=0}^{n-1}(x-t_j),$$

wobei hier die t_j die Tschebyscheff-Knoten bezeichnen. Ähnlich kann gezeigt werden, dass die transformierten Tschebyscheff-Polynome $\hat{T}_n \in P_n([a,b])$ die Eigenschaft

$$\min_{p_n\in P_n, p_n(\eta)=1} \max_{x\in[a,b]} |p_n(x)| = \max_{x\in[a,b]} |\hat{T}_n(x)|$$

haben. Mit dieser Eigenschaft kann eine Konvergenzrate für das cg-Verfahren hergeleitet werden.

Begriffe und Sätze, die bekannt sein sollten: Tschebyscheff-Polynome, Tschebyscheff-Knoten als Lösung des minmax-Problems, Konvergenzrate des cg-Verfahrens

Literatur: [Bar16, Kap. 11.4], [Bar16, Satz 16.1]

5 Interpolation

Polynom-Interpolation. Für paarweise verschiedene Stützstellen ist die Interpolationsaufgabe im Raum der Polynome eindeutig lösbar. Das interpolierende Polynome kann in der Vandermonde-Darstellung, in der Lagrange-Darstellung und in der Newton-Darstellung dargestellt werden. In der Lagrange-Darstellung kann das Interpolationspolynom sehr einfach aufgeschrieben werden. Allerdings müssen alle Basisfunktionen neu berechnet werden, wenn eine neue Stützstelle hinzukommt. In der Newton-Basis heißen die Koeffizienten dividierte Differenzen. Sie erfüllen eine Rekursionsformel, mit deren Hilfe die Koeffizienten berechnet werden können. Eine Rekursionsformel erlaubt außerdem die Berechnung des Interpolationspolynoms ausgewertet an einer Stelle ohne das Interpolationspolynom aufstellen zu müssen. Diese Vorschrift heißt Neville-Schema.

Die absolute Kondition der Interpolationsaufgabe ist gegeben durch die Lebesgue-Konstante, die wesentliche von der Wahl der Stützstellen abhängt. Für die Approximation einer Funktion durch die Interpolierende wird gezeigt, dass der Fehler in der Maximumsnorm beschränkt ist durch

$$\frac{\|f^{(n+1)}\|_{C^0([a,b])}}{(n+1)!} \max_{x\in C([a,b])} |\Pi_{j=0}^n (x-x_j)|.$$

Der letzte Term wird minimiert durch die Tschebyscheff-Knoten.

Begriffe und Sätze, die bekannt sein sollten: Interpolationsaufgabe, Vandermonde-Darstellung, Lagrange-Darstellung, Newton-Basis, Newton-Darstellung, dividierte Differenzen, Berechnung der dividierten Differenzen über Rekursionsformel, Neville-Schema, Kondition der Interpolationsaufgabe

Literatur: [Bar16, Kap. 11.1–11.3], [DH19, Kap. 7.1.1, 7.2.3]

Spline-Interpolation Splines sind Funktionen, die bezüglich einer Partitionierung stückweise polynomiell sind und global eine gewisse Glattheit besitzen. Der Raum der Splines $S^{m,m-1}(\mathcal{T}_n)$ hat die Dimension $m+n$. Der interpolierende natürliche kubische Spline ist eindeutig und minimiert eine linearisierte Biegeenergie. Die Interpolations-, Stetigkeits- und Randbedingungen ergeben ein lineares Gleichungssystem, das gelöst werden muss, um die Interpolation zu berechnen.

Begriffe und Sätze, die bekannt sein sollten: Spline vom Grad m und von der Ordnung n, natürlicher kubischer Spline, Existenz und Eindeutigkeit des interpolierenden, natürlichen Splines

Literatur: [Bar16, Kap. 12]

6 Numerische Integration

Newton-Cotes-Formeln. Eine Quadraturformel approximiert das Integral über eine Funktion über ein Intervall $[a, b]$. Der Exaktheitsgrad beschreibt, wür welche Polynomgrade die Quadratur exakt ist. Hierüber kann der Fehler, der durch die Quadratur gemacht wird, abgeschätzt werden. Die Gewichte in den Newton-Cotes-Formeln können über die Integration der Lagrange-Polynome berechnet werden. Als gebräuchliche Quadraturformeln ergeben sich so die Mittelpunktsregel, die Trapezregel und die Simpson-Regel. Summierte Quadraturformeln unterteilen das Intervall $[a, b]$ in N gleich große Teilintervalle und benutzen auf jedem der Teilintervalle eine Quadraturformel. Dies hat den Vorteil, dass die summierte Quadraturformel auch für nicht glatte Funktionen konvergiert.

Begriffe und Sätze, die bekannt sein sollten: Quadraturformel, Quadraturpunkte, Gewichte, Stabilitätsindikator, Exaktheitsgrad einer Quadraturformel, Fehlerabschätzung für Quadratur mit Exaktheitsgrad r, Newton-Cotes-Formeln, Mittelpunktsregel, Trapezregel, Simpson-Regel, summierte Quadraturformeln, Fehlerabschätzung und Konvergenz und Konvergenzordnungen für summierte Quadraturformeln,

Literatur: [Bar16, Kap. 14.1–14.3]

Gauß-Quadratur. Eine Newton-Cotes-Formel mit $n + 1$ Quadraturpunkten hat immer den Exaktheitsgrad n. Außerdem hat jede Quadraturformel höchstens den Exaktheitsgrad $2n + 1$. Tatsächlich kann dieser erreicht werden, indem die Quadraturpunkte als Nullstellen des $n + 1$-ten Orthogonalpolynoms gewählt werden. Dieses Vorgehen ist auch anwendbar, wenn die Approximation von gewichteten Integralen das Ziel ist. Eine wichtige Feststellung ist, dass die Nullstellen eines Orthogonalpolynoms immer einfach und reell sind und im Intervall liegen, über das integriert werden soll. Ansonsten wäre es nicht möglich, diese Nullstellen als Quadraturpunkte zu wählen. Die über diese Quadraturpunkte definierte Quadratur heißt Gauß-Quadratur und

ist tatsächlich exakt vom Grad $2n + 1$. Außerdem ist sie durch den Exaktheitsgrad schon eindeutig bestimmt.

Begriffe und Sätze, die bekannt sein sollten: maximaler Exaktheitsgrad einer Quadraturformel, Orthogonalpolynome, Legendre-Polynome, Nullstellen eines Orthogonalpolynoms, Gauß-Quadratur, Eindeutigkeit der Gauß-Quadratur

Literatur: [Bar16, Kap. 14.4]

Extrapolation und Romberg-Quadratur. Eine summierte Quadraturformel, die mit Rate h^γ konvergiert, kann für festes (und ausreichend glattes) f als eine Funktion in h^γ angesehen werden, die für $h = 0$ mit dem exakten Integral übereinstimmt. Legt man ein Interpolationspolynom durch die für $h, h/2, \ldots, h/2^k$ berechneten Werte, dann ist die Auswertung dieses Interpolationspolynoms bei $h = 0$ (dies nennt man Extrapolation) im Allgemeinen eine bessere Approximation an das Integral als der Wert für $h/2^k$. Für $k = 1$ ergibt sich beispielsweise eine Approximation der Ordnung $h^{2\gamma}$ statt h^γ. Diese Art von extrapolierten Quadraturregeln werden auch Romberg-Quadratur genannt.

Literatur: [Bar16, Kap. 14.5]

7 Nichtlineare Gleichungssysteme

In diesem Kapitel ist die Lösung eines Problems der Form $f(x) = 0$ (Nullstelle) beziehungsweise $g(x) = x$ (Fixpunkt) gesucht. Das Bisektionsverfahren für Funktionen $f : [a, b] \subseteq \mathbb{R} \to \mathbb{R}$ mit $f(a)f(b) < 0$ konvergiert für alle stetigen f. Das Fixpunktverfahren basiert auf dem Banachschen Fixpunktsatz und konvergiert linear, wenn g den Definitionsbereich in sich selbst abbildet und eine Kontraktion ist. Das Newton-Verfahren approximiert die Funktion f durch ihr erstes Taylor-Polynom. Als Approximation an eine Nullstelle von f wird dann die Nullstelle des Taylor-Polynoms berechnet. Der Satz von Newton-Kantorovich gibt Kriterien dafür an, dass eine Nullstelle von f existiert und das Newton-Verfahren gegen diese Nullstelle konvergiert. Für zwei mal stetig differenzierbares f mit einer Nullstelle, für die die Ableitung von f nicht verschwindet, konvergiert das Newton-Verfahren lokal quadratisch.

Begriffe und Sätze, die bekannt sein sollten: globale und lokale Konvergenz, lineare, quadratische und kubische Konvergenz, Bisektionsverfahren, Konvergenz des Bisektionsverfahrens, Banachscher Fixpunktsatz, Fixpunktverfahren, Konvergenz des Fixpunktverfahrens, Newton-Verfahren, Satz von Newton-Kantorovich, lokal quadratische Konvergenz des Newton-Verfahrens

Literatur: [Bar16, Kap. 15.1–15.2], [Bar16, Kap. 9.1], [Ran17, Kap. 5.5]

8 Eigenwertprobleme

Eine grobe Eingrenzung der Eigenwerte einer Matrix liefert der Gerschgorinsche Kreissatz. Die Potenzmethode (oder Vektoriteration oder von-Mises-Iteration) approximiert den (betragsmäßig) größten Eigenwert einer Matrix dadurch, dass ein Startvektor immer wieder mit der Matrix multipliziert wird. Für allgemeine Matrizen konvergiert die zugehörige Eigenwertapproximation mit Rate q^k, wobei q der

spectral gap ist. Für symmetrische Matrizen kann sogar eine Konvergenz von q^{2k} erreicht werden. Der kleinste Eigenwert kann durch die Inverse Iteration approximiert werden und die Approximation anderer Eigenwerte der Matrix kann durch eine Verschiebung erreicht werden.

Begriffe und Sätze, die bekannt sein sollten: Potenzmethode, Konvergenz der Potenzmethode, Inverse Iteration

Literatur: [Bar16, Kap. 8.3]

Literatur

[Bar16] Sören Bartels. *Numerik* 3 × 9. *Drei Themengebiete in jeweils neun kurzen Kapiteln.* Springer-Lehrb. Heidelberg: Springer Spektrum, 2016.

[DH19] Peter Deuflhard and Andreas Hohmann. *Numerische Mathematik 1. Eine algorithmisch orientierte Einführung.* De Gruyter Stud. Berlin: de Gruyter, 5th revised and expanded edition edition, 2019.

[Her20] Martin Hermann. *Numerische Mathematik. Band 1: Algebraische Probleme.* De Gruyter Stud. Berlin: De Gruyter, 4th revised and enlarged edition edition, 2020.

[Ran17] Rolf Rannacher. *Numerik 0: Einführung in die Numerische Mathematik.* Heidelberg University Publishing, 2017.