Vorlesungsmitschrift
Einführung in die Numerische Mathematik

von Michael Schaefer

6. Juli 2007

Zusammenfassung

Bei dem vorliegenden Dokument handelt es sich um die von mir mit LaTeXgesetzte Version meiner Vorlesungsmitschriften aus der Vorlesung **Einführung in die Numerische Mathematik** im Wintersemester 2006/2007, gelesen von **Prof. Dr. Helmut Maurer** an der **Westfälischen Wilhelms-Universität Münster**. Ich tue dies, damit ich beim Wiederholen des Stoffs, aber insbesondere auch zur Vorbereitung der Klausur, nicht auf eine lose Blattsammlung von handgeschriebenen und mehr oder weniger lesbaren Zetteln angewiesen bin. Wer Fehler in diesem Dokument findet, den bitte ich, mir diese per E-Mail mitzuteilen:

michael.schaefer@uni-muenster.de

Den Stand des Dokumentes entnehme man bitte dem weiter oben angeführten Datum.

Inhaltsverzeichnis

I Zahldarstellung und Fehlerrechnung **3**

§ 1 Zahldarstellung und Rundungsfehler . . . 3
1.1 Zahldarstellung . . . 3
1.2 Rundung . . . 4
1.3 Gleitpunktarithmetik . . . 4
§ 2 Fehleranalyse . . . 6
2.1 Fehlertypen . . . 6
2.2 Kondition und Gutartigkeit . . . 6

II Lineare Gleichungssysteme **10**

§ 3 Gauß-Elimination und LR-Zerlegung einer Matrix . . . 10
§ 4 Spezielle Matrizen. Das Cholesky-Verfahren. . . . 17
§ 5 Fehlerabschätzungen bei linearen Gleichungssystemen . . . 20
§ 6 Die QR-Zerlegung einer Matrix, Verfahren von HOUSEHOLDER . . . 25
§ 7 Überbestimmte lineare Gleichungssysteme, Lineare Ausgleichsprobleme, Diskrete Approximation . . . 28

III Iterationsverfahren zur Lösung von Gleichungen **33**

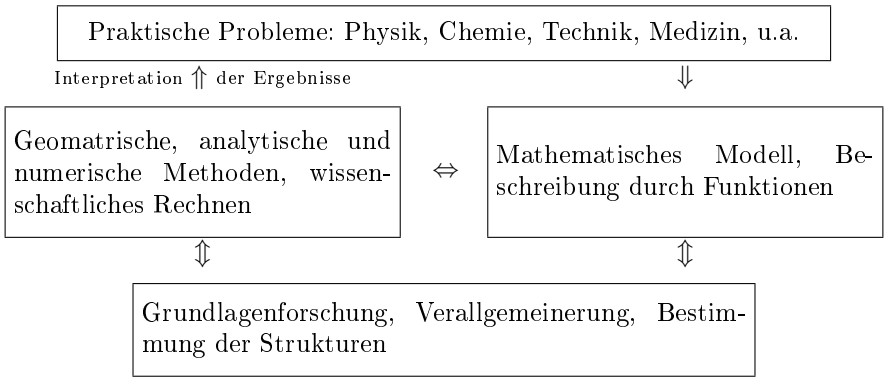
§ 8 Definitionen und Grundbegriffe . . . 33
8.1 Nullstellen . . . 33
8.2 Fixpunkte . . . 33
8.3 Konvergenzgeschwindigkeit . . . 34
§ 9 Nullstellen reeller Funktionen . . . 35
9.1 Intervallhalbierung, Bisektionsverfahren . . . 35
9.2 NEWTON-Verfahren . . . 35
9.3 Das Sekantenverfahren (Regular falsi) . . . 36
§ 10 Konvergenzsätze für Iterationsverfahren . . . 37
§ 11 Das NEWTON-Verfahren im $\mathbb{R}^n$ . . . 41
§ 12 Iterationsverfahren für lineare Gleichungssysteme . . . 43

IV Interpolation **48**

§ 13 Interpolation durch Polynome . . . 49
13.1 Die Interpolationsformel von LAGRANGE . . . 49
13.2 Der Algorithmus von AITKEN und NEVILLE. . . . 50
13.3 Die NEWTONsche Interpolation, Dividierte Differenzen . . . 51
13.4 Der Interpolationsfehler, Konvergenzfragen . . . 52
§ 14 Spline-Interpolation . . . 54
14.1 Polynom-Splines . . . 54
14.2 Kubische Splines ($l = 3$) . . . 54
14.3 Berechnung kubischer Splines . . . 56
14.4 Konvergenzeigenschaften . . . 58

Ein paar Vorbemerkungen

Stellung der Angewandten Mathematik



Ein sehr einfaches Beispiel

Praktisches Problem: Bestimme ein Rechteck mit gegebenem Umfang, welches maximalen Flächeninhalt aufweist.
Mathematisches Modell: Umfang $U = 2x + 2y > 0$. Maximiere $F = F(x, y) = xy$ unter der Nebenbedingung $U = 2x + 2y$.
Analytische Lösung: $2x + 2y = U \Rightarrow y = \frac{1}{2}(U - 2x)$. Folglich erhält man $F = xy = \frac{1}{2}x(U - 2x) = \frac{1}{2}xU - x^2 =: f(x)$.
Notwendige Bedingung für ein Maximum von $f(x)$: es gilt $0 = f'(x)$, also $\frac{1}{2}U - 2x = 0 \Rightarrow \underline{x = \frac{U}{4}}$. Damit erhält man dann $\underline{y = \frac{U}{4}}$.
Ergebnis: Die Lösung ist ein Quadrat mit $x = y = \frac{U}{4}$.
Grundlagenforschung, Verallgemeinerung: Sei $x = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$, seien $f : \mathbb{R}^n \to \mathbb{R}$ und $g : \mathbb{R}^n \to \mathbb{R}^m$.
Optimierungsproblem: Minimiere (Maximiere) die Funktion $f(x_1, \ldots, x_n)$ unter der Nebenbedingung $g(x_1, \ldots, x_n) = 0$.
Ziel: Notwendige und hinreichende Optimalitätsbedingung, Numerische Verfahren.

Grundaufgaben der Numerischen Mathematik

- Lösung linearer Gleichungssysteme
- Nichtlineare Gleichungen und Fixpunkte
  Sei $D \subset \mathbb{R}^n$ und sei $g : D \to \mathbb{R}^n$ stetig. Gesucht: Fixpunkt $\overline{x}$ mit $g(\overline{x}) = \overline{x}$.
  Sei $D \subset \mathbb{R}^n$ und $f : D \to \mathbb{R}^n$. Gesucht: Nullstelle $\overline{x} \in D$ von f, d.h. $f(\overline{x}) = 0$.
- Interpolation
  Gegeben: Messpunkte $x_0 < x_1 < \ldots < x_n$ und Messwerte $f_i, i = 0, \ldots, n$. Gesucht: Polynom $p_n(x) := a_0 + a_1 x + \ldots + a_n x^n$ mit $p_n(x_i) = f_i$ für $i = 0, \ldots, n$.
- Integration
  Sei $[a, b] \subset \mathbb{R}$ und $f : [a, b] \to \mathbb{R}$ stetig. Berechne $I = \int_a^b f(x)dx$.

I Zahldarstellung und Fehlerrechnung

§ 1 Zahldarstellung und Rundungsfehler

1.1 Zahldarstellung

Es ist bekannt, wie eine Zahl im Dezimalsystem dargestellt und eine Darstellung interpretiert wird:

$$\begin{aligned} 272.5 &= 2\cdot 10^2 + 7\cdot 10^1 + 2\cdot 10^0 + 5\cdot 10^{-1} \\ 0.0307 &= 3\cdot 10^{-2} + 7\cdot 10^{-4} \\ \pi &= 3.14159265358979 \\ 1 &= 1.000\ldots = 0.999\ldots \qquad \text{(nicht eindeutig)} \end{aligned}$$

Wie man allgemein eine Zahl $x \in \mathbb{R}$ zur Basis $d \in \mathbb{N}, n \geq 2$ darstellt, klärt der nächste Satz.

(1.1) Satz. Sei $d \in \mathbb{N}, d \geq 2$, und sei $x \in \mathbb{R}, x \neq 0$. Dann gibt es eine Darstellung

$$x = \pm \sum_{i=k}^{-\infty} \alpha_i d^i, \quad k \in \mathbb{Z}, \alpha_i \in \{0,\ldots,d-1\}, \alpha_k \neq 0.$$

Diese Darstellung ist eindeutig, wenn zusätzlich gilt: Zu jedem $n \in \mathbb{N}$ gibt es einen Index $i \leq -n$ mit $\alpha_i \neq d-1$ (Nichtzulassung einer Periode).
Schreibweise: $x = \pm\alpha_k\alpha_{k-1}\ldots\alpha_0.\alpha_{-1}\ldots, 0 \leq \alpha_i \leq d-1, \alpha_i \in \mathbb{N}$. Die am häufigsten verwendeten Basen sind 2,8,10 und 16:

Name des System	Basis d	Ziffern
Dual, Binär	2	$0,1$ oder $0,L$
Oktal	8	0,1,...,7
Dezimal	10	0,1,...,9
Hexadezimal	16	0,1,...,9,A,B,...,F

Beispiele:

$$\begin{aligned} 18.5 &= 1\cdot 2^4 + 0\cdot 2^3 + 0\cdot 2^2 + 1\cdot 2^1 + 0\cdot 2^0 + 1\cdot 2^{-2} = L00L0.L \\ 3.2 &= LL.\overline{00LL} \\ 29 &= 1D \end{aligned}$$

Problem: Auf dem Rechner können nur endlich viele (rationale) Zahlen dargestellt werden. Wir definieren daher die Menge der Maschinenzahlen

$$A := \left\{ x = \pm\left(\sum_{k=1}^{m} a_k d^{-k}\right)\cdot d^l \;\middle|\; a_k \in \{0,\ldots,d-1\} \right\} \tag{1.2}$$

Die Darstellung einer solchen Maschinenzahl ist $x = \pm 0.a_1a_2\ldots a_m \cdot d^l$. Dabei bezeichnen wir $\sum_{k=1}^{m} a_k d^{-k}$ als Mantisse der Länge m, wobei $m \in \mathbb{N}$ fest gewählt wird. Weiterhin nennen wir l den Exponenten, mit $l \in \tilde{\mathbb{Z}} \subset \mathbb{Z}$, $\tilde{\mathbb{Z}}$ endlich, etwa $l \in [l_{\min},\ldots,l_{\max}]$.

Mit dieser Definition sind verschiedene Datentypen für Zahlen möglich:

- Integerzahlen: $x = \pm\left(\sum_{k=1}^{m} a_k d^{-k}\right)\cdot d^m = \sum_{k=1}^{m} a_k d^{m-k}$.
- Festkommazahlen: $x = \pm\left(\sum_{k=1}^{m} a_k d^{-k}\right)\cdot d^q$, $q \in \mathbb{Z}$ fest gewählt.
  Diese Darstellung ist in vielen Fällen ungeeignet, wenn es z.B. eine große Bandbreite an Exponenten gibt. Als Beispiele aus der Physik seien die Ruhemasse eines Elektrons mit $9.11\cdot 10^{-28}g$ und die Lichtgeschwindigkeit mit $2.998\cdot 10^9 m/s$ genannt.
- (normierte) Gleitkommezahlen: $x = \pm\left(\sum_{k=1}^{m} a_k d^{-k}\right)\cdot d^l, a_1 \neq 0$ für $x \neq 0$.
  Hierbei gilt für den Exponenten $l_{min} \leq l \leq l_{max}$ mit geeigneten $l_{min}, l_{max} \in \mathbb{Z}$. Alle darstellbaren Zahlen liegen im Bereich $d^{l_{min}} \leq |x| \leq d^{l_{max}}$. Die Wahl von l_{min} und l_{max} hängt vom Rechner ab.
  Exponentenunterlauf: $|x| < d^{l_{min}}$, x wird durch 0 ersetzt.
  Exponentenüberlauf: $|x| > d^{l_{max}}$.

1.2 Rundung

Eine Abbildung $rd : \mathbb{R} \to A$ heißt Rundung, wenn gilt

$$|rd(x) - x| = \min_{a \in A} |a - x|. \tag{1.3}$$

Eine solche Rundungsfunktion, die eine Zahl $x = \pm \left(\sum_{k=1}^{\infty} a_k d^{-k}\right) \cdot d^l$ auf m (Mantissenlänge) Stellen rundet (für die (normierte) Gleitkommadarstellung), lautet

$$rd(x) = \begin{cases} \pm \left(\sum_{k=1}^{m} a_k d^{-k}\right) \cdot d^l, & \text{falls } a_{m+1} < \frac{d}{2} \\ \pm \left(\sum_{k=1}^{m} a_k d^{-k} + d^{-m}\right) \cdot d^l, & \text{falls } a_{m+1} \geq \frac{d}{2} \end{cases} \tag{1.4}$$

Beispiele: Sei $d = 10$ und $m = 4$.

$$\begin{aligned} \pi &= 3.14159\ldots, \quad rd(\pi) = 0.3142 \cdot 10^1 \\ \sqrt{57} &= 7.5498\ldots, \quad rd(\sqrt{57}) = 0.7550 \cdot 10^1 \end{aligned}$$

Die Zahl

$$\text{eps} = \frac{1}{2} d^{m-1}$$

heißt Maschinengenauigkeit einer normierten Gleitkommamaschine mit Mantissenlänge m zur Basis d.

(1.5) Definition. Sei $x \in \mathbb{R}$ und sei $\tilde{x} \in \mathbb{R}$ eine Näherung für x.

(i) $|x - \tilde{x}|$ heißt der absolute Fehler von $\tilde{x}$,

(ii) für $x \neq 0$ heißt $\left|\frac{x - \tilde{x}}{x}\right|$ der relative Fehler von $\tilde{x}$.

(1.6) Satz. Sei $rd : \mathbb{R} \to A$ die Rundung aus (1.4). Dann gilt für $x \in \mathbb{R} \setminus \{0\}$

(i) $\left|\frac{x - rd(x)}{x}\right| \leq eps = \frac{1}{2} d^{-m+1}$,

(ii) $\left|\frac{x - rd(x)}{rd(x)}\right| \leq eps$.

Beweis für (i). Sei $x = \pm 0.a_1 a_2 \ldots a_m a_{m+1} \ldots \cdot d^e$, $a_1 \neq 0$. Nach (1.4) gilt $|rd(x) - x| \leq \frac{d}{2} \cdot d^{-(m+1)} \cdot d^e$. Zusammen mit $|x| \geq d^{-1} \cdot d^e$ folgt

$$\left|\frac{rd(x) - x}{x}\right| \leq \frac{d}{2} \cdot d^{-m-1} \cdot d = \frac{1}{2} d^{-m+1}.$$

Beweis für (ii) analog. □

Damit erhalten wir also die Darstellung

$$rd(x) = x(1 + \varepsilon),\ |\varepsilon| \leq eps = \frac{1}{2} d^{-m+1}. \tag{1.7}$$

1.3 Gleitpunktarithmetik

Das Symbol $\square$ bezeichne eine der Rechenoperationen $+, -, \cdot, /$; genauer: $\boxplus, \boxminus, \boxdot, \boxed{/}$. Die Operation $\square$ ist nicht abgeschlossen in A.

Beispiele: $d = 10, m = 2$.

$$\begin{aligned} 0.11 \boxplus 0.0011 &= 0.1111 \notin A \\ 1.1 \boxdot 1.1 &= 1.21 \notin A \end{aligned}$$

Folgerung: Nach einer Operation $\square$ muss im Allgemeinen gerundet werden.

Ausführung auf dem Rechner:

a) möglichst genaue Berechnung von $z = x \square y$, $x, y \in A$, etwa auf $2m$ Ziffern

b) Normalisierung von z (Gleitkommazahl) und anschließende Rundung auf m Stellen.

Die Arithmetik auf dem Rechner kann so organisiert werden, dass gilt

$$gl(x\Box y) = rd(x\Box y) = (x\Box y)(1+\varepsilon) \text{ für } x, y \in A, |\varepsilon| \leq eps \tag{1.8}$$

Beachte: Die Gleitpunktoperationen sind weder assoziativ noch distributiv!

Beispiele: $d = 10, m = 2$. Berechne $0.75 + 0.055 - 0.8 = 0.005$ in Gleitpunktarithmetik:

1) $gl(0.75 + 0.055) = rd(0.805) = 0.81 \Rightarrow gl(0.81 - 0.8) = rd(0.01) = 0.01$.

2) $gl(0.75 + gl(0.055 - 0.8)) = gl(0.75 + rd(-0.745)) = gl(0.75 - 0.75) = 0$.

§ 2 Fehleranalyse

2.1 Fehlertypen

Gegeben: $D \subset \mathbb{R}^n$, $f : D \to \mathbb{R}^m$.

Problem: Berechne

$$\begin{cases} y = f(x),\ x \in D \\ y_i = f_i(x_1, \ldots, x_n),\ i = 1, \ldots, m \\ x = (x_1 \ldots, x_n)^t \to y = (y_1, \ldots, y_m)^t \end{cases} \tag{2.1}$$

Fehlertypen, die die Genauigkeit der Berechnung von $y = f(x)$ begrenzen, sind:

1. Fehler in den Eingabedaten $x \in \mathbb{R}^n$
2. Abbrech- und Diskretisierungsfehler
3. Rundungsfehler bei den Rechnungen

Beispiel: $m = 3$. Zu berechnen sei

$$e^x = \sum_{k=0,\infty} \frac{x^k}{k!},\ x \in \mathbb{R},$$

hier für $x = 1.2345$.

1. Führe die Rechnung mit der Approximation $\tilde{x} = rd(x) = 1,23 \in A$ durch.
2. Approximiere die unendliche Summe $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$ durch die endliche Summe $S_N := \sum_{k=0}^{N} \frac{x^k}{k!}$. Abbrechfehler: $e^x - S_N = \sum_{k=N+1}^{\infty} \frac{x^k}{k!}$.
3. Rundungsfehler: Berechne mit $\tilde{x}$ die Ausdrücke $gl\left(\frac{1.23^k}{k!}\right)$ für $k = 1, \ldots, N$.

Statt $y = f(x)$ werden die Approximationen berechnet:

$$\begin{aligned} &x \to y = f(x) \\ &\tilde{x} \to \tilde{y} = f(\tilde{x}) \\ &\tilde{x} \to \tilde{\tilde{y}} = \tilde{f}(\tilde{x}),\ \tilde{f} \text{ ist Approximation von } f. \end{aligned}$$

2.2 Kondition und Gutartigkeit

Zunächst betrachten wir den Fehlertyp 1.

Sei $\tilde{x} \in \mathbb{R}^n$ eine Näherung von $x \in D \subset \mathbb{R}^n$. Wir wollen den Fehler im Ergebnis $\tilde{y} = f(\tilde{x})$ zu $y = f(x)$ betrachten. Wir definieren

(i) $\Delta x = \tilde{x} - x$

(ii) $\Delta_{x_i} = \tilde{x_i} - x_i\, i = 1, \ldots, n$

(iii) $\Delta y = \tilde{y} - y$.

Dann gilt $\Delta y = f(x + \Delta x) - f(x)$. Mit der Taylorentwicklung erhalten wir

$$\Delta y_i = f_i(x + \Delta x) - f_i(x) \approx \sum_{j=1}^{n} \frac{\partial f_i(x)}{\partial x_j} \Delta x_j \tag{2.2}$$

Hierbei sei $f : D \to \mathbb{R}^m$ eine C^1-Funktion. Wir nennen hierbei $\frac{\partial f_i(x)}{\partial x_j}$ den Verstärkungsfaktor für den absoluten Fehler $\Delta x_j \to \Delta y_i$. Für den relativen Fehler gilt

$$\frac{\Delta y_i}{y_i} \approx \sum_{j=1}^{n} \left[\left(\frac{\partial f_i(x)}{\partial x_j} \cdot \frac{x_j}{y_i} \right) \cdot \frac{\Delta x_j}{x_j} \right] \text{ mit } y_i = f_i(x) \tag{2.3}$$

(2.4) Definition.

(i) Die Zahlen $k_{ij}(x) := \left| \frac{\partial f_i(x)}{\partial x_j} \cdot \frac{x_j}{f_i(x)} \right|$, $i = 1, \ldots, m$, $j = 1, \ldots, n$ heißen (relative) Verstärkungsfaktoren oder Konditionszahlen.

(ii) Das Problem "Berechne $y = f(x)$" heißt gut konditioniert, falls alle Konditionszahlen $k_{ij}(x)$ die Größenordnung 1 haben. Andernfalls heißt das Problem schlecht konditioniert.

Untersuchung der arithmetischen Operationen +,-,·,/

(i) **Multiplikation**

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2,\ x \to y = f(x_1, x_2) = x_1 x_2$$

Also $n = 2, m = 1$.

$$\begin{aligned} k_{11}(x) &= \left| \frac{\partial f(x)}{\partial x_1} \cdot \frac{x_1}{f(x)} \right| = \left| x_2 \cdot \frac{x_1}{x_1 x_2} \right| = 1 \\ k_{12}(x) &= \left| \frac{\partial f(x)}{\partial x_2} \cdot \frac{x_2}{f(x)} \right| = \left| x_1 \cdot \frac{x_2}{x_1 x_2} \right| = 1 \end{aligned}$$

Die Multiplikation ist also gut konditioniert bzw. gutartig.

(ii) **Division**

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2,\ x \to y = f(x_1, x_2) = \frac{x_1}{x_2}$$

Also $n = 2, m = 1$. Wir berechnen

$$\begin{aligned} k_{11}(x) &= 1 \\ k_{12}(x) &= 1 \end{aligned}$$

Also ist auch die Division gutartig.

(iii) **Addition und Subtraktion**

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2,\ x \to y = f(x_1, x_2) = x_1 + x_2$$

$$\begin{aligned} k_{11}(x) &= \left| \frac{\partial f(x)}{\partial x_1} \cdot \frac{x_1}{f(x)} \right| = \left| \frac{x_1}{x_1 + x_2} \right| \\ k_{12}(x) &= \left| \frac{x_2}{x_1 + x_2} \right| \end{aligned}$$

Das Problem ist schlecht konditioniert, falls $x_1 \approx -x_2$ (Addition) bzw. $x_1 \approx x_2$ (Subtraktion). Dieses Phänomen heißt Auslöschung. Dazu ein Beispiel:

$$\begin{aligned} 1.31 - 1.25 &= 0.06 \\ 1.32 - 1.24 &= 0.08 \end{aligned}$$

Der relative Fehler in x_1 und x_2 liegt bei etwa 0.8%, der Ergebnisfehler jedoch bei 30%! Genauer gilt:

$$x = \begin{pmatrix} 1.31 \\ -1.25 \end{pmatrix} \to y = x_1 + x_2 = 0.06$$

$$\Delta x = \begin{pmatrix} 0.01 \\ 0.01 \end{pmatrix},\ \left| \frac{\Delta x_i}{x_i} \right| \leq 0.08,\ i = 1, 2$$

$$k_{1i}(x) = \left| \frac{x_i}{x_1 + x_2} \right| \leq 22,\ i = 1, 2$$

Ergebnis: Der relative Fehler im Ergebnis ist ca. 40mal größer als der relative Fehler in den Daten!

(iv) **Wurzel**

$$x \in \mathbb{R}^+,\ x \to y = f(x) = \sqrt{x},\ y \in \mathbb{R}$$

Also $n = m = 1$. Wir berechnen

$$k(x) = \frac{\partial f(x)}{\partial x} \cdot \frac{x}{f(x)} = \frac{1}{2}\frac{1}{\sqrt{x}} \cdot \frac{x}{\sqrt{x}} = \frac{1}{2}$$

Die Wurzel ist also gutartig.

(2.5) Definition. Ein Algorithmus zur Berechnung einer Lösung $y = f(x)$ ist eine Sequenz von endlich vielen "elementaren Operationen" wie z.B. $+, -, \cdot, /, \sqrt{x}, \cos(x)$.
Es gibt im Allgemeinen mehrere Anordnungen der Rechenschritte, die zum gleichen Ergebnis $y = f(x)$ führen. In jedem dieser Schritte fallen Rundungsfehler an. Dabei kann eine ungünstige Anordnung der Rechenschritte zum Aufschaukeln der Rundungsfehler führen, obwohl das eigentliche Problem $y = f(x)$ gut konditioniert ist.

(2.6) Beispiel. Gegeben sei die Quadratische Gleichung $y^2 + 2py - q = 0$ mit $p \gg q > 0$. Zu berechnen sei die größere der beiden Nullstellen. Es gilt

$$y = -p + \sqrt{p^2+q} = \frac{q}{p + \sqrt{p^2+q}} = f(p,q) \text{ (zweiter Ausdruck mittels quadratischer Ergänzung).}$$

Diese Ausdrücke sind mathematisch äquivalent, jedoch nicht numerisch.

Das Problem ist gut konditioniert, denn es gilt

$$\begin{aligned} k_p(p,q) &= \left|\frac{\partial f(p,q)}{\partial p} \cdot \frac{p}{f(p,q)}\right| = \frac{p}{\sqrt{p^2+q}} < 1, \\ k_q(p,q) &= \left|\frac{\partial f(p,q)}{\partial q} \cdot \frac{q}{f(p,q)}\right| = \frac{p+\sqrt{p^2+q}}{2\sqrt{p^2+q}} < 1. \end{aligned}$$

1. Berechnungsmethode.

Wir berechnen in den Schritten $(p,q) \to p^2+q \to \sqrt{p^2+q} \overset{(*)}{\to} -p + \sqrt{p^2+q}$. An der Stelle $(*)$ kann Auslöschung auftreten! Zum Beispiel für $m = 3, p = 100, q = 10$ erhalten wir

$$p^2 + q = 10010 \Rightarrow rd(p^2+q) = 10000 \Rightarrow gl(-p + \sqrt{p^2+q}) = rd(-100+100) = 0.$$

2. Berechnungsmethode.

Wir erhalten hier $gl(y) = 0.05$.

Das exakte Ergebnis ist $y = 0.004999875\ldots$.

(2.7) Beispiel. Zu berechnen sei das Integral $I_N := \int_0^1 \frac{x^n}{x+5} dx$ für $n = 1, 2, \ldots$. Als Abschätzung erhalten wir

$$|I_N| \le \int_0^1 \frac{x^n}{5} dx = \frac{1}{5(n+1)} \to 0 \text{ für } n \to \infty.$$

1. Methode: Vorwärtsrekursion für I_n. Es gilt

$$I_n + 5I_{n-1} = \int_0^1 \frac{x^n + 5x^{n-1}}{x+5} dx = \int_0^1 x^{n-1} dx = \frac{1}{n} \Rightarrow I_n = \frac{1}{n} - 5I_{n-1}$$

Als Startwert verwenden wir $I_0 = \ln\left(\frac{6}{5}\right) = 0.1823215\ldots$. Hierbei wird jedoch ein Rundungsfehler in den Eingabedaten in jedem Schritt mit dem Faktor 5 multipliziert. Für den Fehler in I_n gilt also

$$|\varepsilon_n| \le 5^n \cdot \underbrace{\frac{1}{2} \cdot 10^{-m+1}}_{\text{Maschinengenauigkeit}}.$$

Für $n = 20$ und $m = 9$ folgt also bereits $|\varepsilon_{20}| \le 5 \cdot 10^5$.

2. Methode: Rückwärtsrekursion für I_n. Es gilt

$$I_{n-1} = \frac{1}{5}\left(-I_n + \frac{1}{n}\right)$$

Da wir wissen, dass $|I_n| \overset{n\to\infty}{\longrightarrow} 0$, setzen wir willkürlich $I_{50} = 0$ (eigentlich richtig wäre $I_{50} = 0.00327851462\ldots$) als Startwert fest. Der hierbei auftretende Anfangsfehler wird in jedem Schritt mit dem Faktor $\frac{1}{5}$ multipliziert, daher wird das Ergebnis schnell sehr exakt!

Übersicht:

n	$I_n = \frac{1}{n} - 5I_{n-1}$, $I_0 = \ln\left(\frac{6}{5}\right)$	$I_{n-1} = \frac{1}{5}\left(-I_n + \frac{1}{n}\right)$
1	0.088392216	0.088392216
5	0.028468364	0.028468352
10	0.015329188	0.015367550
15	0.130402734	0.010520733
20	$-3.746232 \cdot 10^{1}$	$7.997523028 \cdot 10^{-3}$

II Lineare Gleichungssysteme

§ 3 Gauß-Elimination und LR-Zerlegung einer Matrix

Sei $A = (a_{ik})_{1 \le i,k \le n}$ eine (n,n)-Matrix und sei $b = (b_1, \dots, b_n)^t \in \mathbb{R}^n$. Gesucht ist eine Lösung $x = (x_1, \dots, x_n)^t \in \mathbb{R}^n$ des linearen Gleichungssystems (LGS)

$$Ax = b \tag{3.1}$$

Motivationsbeispiel für die Gauß-Elimination

$$\begin{array}{rcrcrcl|l} 2x_1 & + & 4x_2 & + & 6x_3 & = & 2 & \\ x_1 & + & x_2 & & & = & 1 & -\frac{1}{2}I \\ x_1 & & & + & x_3 & = & 2 & -\frac{1}{2}I \end{array} \Rightarrow \begin{array}{rcrcrcl|l} 2x_1 & + & 4x_2 & + & 6x_3 & = & 2 & \\ & & -x_2 & - & 3x_3 & = & 0 & \\ & & -2x_2 & - & 2x_3 & = & 1 & -2I \end{array} \Rightarrow$$

$$\left.\begin{array}{rcrcrcl} 2x_1 & + & 4x_2 & + & 6x_3 & = & 2 \\ & & -x_2 & - & 3x_3 & = & 0 \\ & & & & 4x_3 & = & 1 \end{array}\right\} \underbrace{\begin{pmatrix} 2 & 4 & 6 \\ 0 & -1 & -3 \\ 0 & 0 & 4 \end{pmatrix}}_{\text{R}} \underbrace{\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}}_{\text{x}} = \underbrace{\begin{pmatrix} 2 \\ 0 \\ 1 \end{pmatrix}}_{\text{c}}$$

Man erhält $x_3 = \frac{1}{4}, x_2 = -\frac{3}{4}, x_1 = \frac{7}{4}$.

Idee der Gauß-Elimination

Führe das LGS $Ax = b$ durch eine geeignete Linearkombination von Gleichungen in ein gestaffeltes LGS über:

$$Rx = c \in \mathbb{R}^n,\ R = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ & r_{22} & & r_{2n} \\ & & \ddots & \vdots \\ 0 & & & r_{nn} \end{pmatrix}$$

R heißt rechte obere Dreiecksmatrix. Falls $r_{ii} \neq 0 \, \forall i = 1, \dots, n$, so kann $Rx = c$ rückwärts gelöst werden:

$$x_i = \frac{1}{r_{ii}} \left(c_i - \sum_{k=i+1}^{n} r_{ik} x_k \right) \text{ für } i = n, \dots, 1 \tag{3.2}$$

Durchführung der Gauß-Elimination

Es werden Elementarmatrizen L_j benötigt:

$$L_j = \begin{pmatrix} 1 & & & & 0 \\ & \ddots & & & \\ & & 1 & & \\ & & -l_{j+1,j} & \ddots & \\ & & \vdots & & \\ 0 & & -l_{n,j} & & 1 \end{pmatrix}$$

Rechenregeln für Elementarmatrizen

$$L_j A = \begin{pmatrix} a_1 \\ \vdots \\ a_j \\ a_{j+1} - l_{j+1,j} \cdot a_j \\ \vdots \\ a_n - l_{nj} \cdot a_j \end{pmatrix} \tag{3.3}$$

$$L_j^{-1} = \begin{pmatrix} 1 & & & & 0 \\ & \ddots & & & \\ & & 1 & & \\ & & l_{j+1,j} & \ddots & \\ & & \vdots & & \\ 0 & & l_{n,j} & & 1 \end{pmatrix} \tag{3.4}$$

Für $j < k$ gilt

$$L_j^{-1} L_k^{-1} = \begin{pmatrix} 1 & & & & & & 0 \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & l_{j+1,j} & \ddots & & & \\ & & \vdots & & 1 & & \\ & & \vdots & & l_{k+1,k} & \ddots & \\ & & \vdots & & \vdots & & \ddots \\ 0 & & l_{n,j} & & l_{n,k} & & 1 \end{pmatrix} \tag{3.5}$$

Gauß-Elimination und LR-Zerlegung ohne Pivotsuche

Setze

$$A = A^{(1)} = \begin{pmatrix} a_{11}^{(1)} & \cdots & a_{1n}^{(1)} \\ \vdots & & \vdots \\ a_{n1}^{(1)} & \cdots & a_{nn}^{(1)} \end{pmatrix}$$

1. Schritt: Sei $a_{11}^{(1)} \neq 0$. Bestimme $l_{i1}, i = 2, \ldots, n$ so, dass gilt

$$L_1 A = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} \\ \vdots & & & \vdots \\ 0 & a_{n2}^{(2)} & \cdots & a_{nn}^{(2)} \end{pmatrix}, \text{vgl. (3.3) mit } j = 1$$

Die Forderung $a_{i1}^{(1)} - l_{i1} \cdot a_{11}^{(1)} = 0$ führt auf

$$\begin{aligned} l_{i1} &= \frac{a_{i1}^{(1)}}{a_{11}^{(1)}}, i = 2, \ldots, n \text{ und} \\ a_{ik}^{(2)} &= a_{ik}^{(1)} - l_{i1} \cdot a_{1k}^{(1)}, i, k = 2, \ldots, n. \end{aligned}$$

Ausgangspunkt vor dem j-ten Schritt:

$$L_{j-1} \cdot \ldots \cdot L_2 L_1 A^{(1)} = A^{(j)} = \begin{pmatrix} a_{11}^{(1)} & & & & a_{1n}^{(1)} \\ & \ddots & & & \vdots \\ & & a_{jj}^{(j)} & \cdots & a_{jn}^{(j)} \\ & & \vdots & & \vdots \\ 0 & & a_{nj}^{(j)} & \cdots & a_{nn}^{(j)} \end{pmatrix}$$

j-ter Schritt: Sei $a_{jj}^{(j)} \neq 0$. Bestimme $l_{ij}, i = j+1, \ldots, n$ so, dass gilt

$$L_j A^{(j)} = A^{(j+1)} = \begin{pmatrix} a_{11}^{(1)} & \cdots & \cdots & \cdots & \cdots & a_{1n}^{(1)} \\ & \ddots & & & & \vdots \\ & & a_{jj}^{(j)} & \cdots & \cdots & a_{jn}^{(j)} \\ & & 0 & a_{j+1,j+1}^{(j+1)} & \cdots & a_{j+1,n}^{(j+1)} \\ & & \vdots & \vdots & \ddots & \vdots \\ 0 & & 0 & a_{n,j+1}^{(j+1)} & \cdots & a_{nn}^{(j+1)} \end{pmatrix}$$

Die Forderung $a_{ij}^{(j)} - l_{ij} a_{jj}^{(j)} = 0, i = j+1, \ldots, n$ führt auf

$$\begin{aligned} l_{ij} &= \frac{a_{ij}^{(j)}}{a_{jj}^{(j)}}, i = j+1, \ldots, n, \text{ woraus folgt} \\ a_{ik}^{(j+1)} &= a_{ik}^{(j)} - l_{ij} a_{jk}^{(j)}, i, k = j+1, \ldots, n \end{aligned}$$

Nach $n-1$ Schritten erhält man

$$\begin{pmatrix} a_{11}^{(1)} & \cdots & \cdots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} \\ \vdots & & \ddots & \vdots \\ 0 & & & a_{nn}^{(n)} \end{pmatrix} =: R = (r_{ik}) \tag{3.6}$$

Anwendung der Matrizen L_j auf die erweiterte Matrix $(A|b)$:

$$L_{n-1} \cdot \ldots \cdot L_1 \cdot (A|b) = (R|c), c \in \mathbb{R}^n.$$

Das LGS $Ax = b$ ist daher äquivalent zum LGS $Rx = c$ und lässt sich gemäß (3.2) lösen.

Aus (3.6) folgt nun mit (3.4) und (3.5) die LR-Zerlegung von A:

$$A = L_1^{-1} L_2^{-1} \cdot \ldots \cdot L_{n-1}^{-1} \cdot R =: LR, \tag{3.7}$$

$$L = L_1^{-1} \cdot \ldots \cdot L_{n-1}^{-1} = \begin{pmatrix} 1 & & & 0 \\ L_{21} & 1 & & \\ \vdots & l_{32} & \ddots & \\ \vdots & \vdots & 1 & \\ l_{n1} & l_{n2} & l_{n,n-1} & 1 \end{pmatrix} \text{ (linke (untere) Dreiecksmatrix)}$$

Bei gegebener Zerlegung $A = LR$ ist das LGS $Ax = b$ äquivalent zu den LGS

$$\begin{aligned} &Lc = b, && c_i = b_i - \textstyle\sum_{k=1}^{i-1} l_{ik} c_k,\ i = 1, \ldots, n && \text{(Vorwärtsrekursion)} \\ &Rx = c, && x_i = \tfrac{1}{r_{ii}} \left(c_i - \textstyle\sum_{k=i+1}^{n} r_{ik} x_k \right),\ i = n, \ldots, 1 && \text{(Rückwärtsrekursion)} \end{aligned}$$

Insbesondere folgt aus (3.7)

$$\det(A) = \det(L) \cdot \det(R) = \prod_{i=1}^{n} r_{ii}.$$

Problem: Wann gilt $a_{jj}^{(j)} \neq 0$ in $A^{(j)}$?

(3.8) Satz. Sei A eine (n,n)-Matrix, deren Hauptabschnittsmatrizen $A_j := (a_{ik})_{1 \leq i,k \leq j}$ regulär (d.h. $\det(A_j) \neq 0$) sind für $j = 1, \ldots, n$. Dann gibt es eine eindeutige Zerlegung $A = LR$, wobei L eine linke Dreiecksmatrix mit $l_{ii} = 1 \, \forall i = 1, \ldots, n$ und R eine reguläre rechts Dreiecksmatrix ist.

Beweis. Durch Induktion über n.
Für $n = 1$ ist die Behauptung klar. Sei daher die Behauptung für ein $n-1$ richtig. Dann ist für eine (n,n)-Matrix die folgende Zerlegung zu zeigen:

$$A = \left(\begin{array}{c|c} A_{n-1} & d \\ \hline a^t & a_{nn} \end{array} \right) = \left(\begin{array}{c|c} L_{n-1} & 0 \\ \hline l^t & 1 \end{array} \right) \cdot \left(\begin{array}{c|c} R_{n-1} & r \\ \hline 0 & r_{nn} \end{array} \right)$$

Nach Induktionsvoraussetzung gibt es eine Zerlegun

$$A_{n-1} = L_{n-1} \cdot R_{n-1}$$

mit R_{n-1} regulär. For die gesuchten Vektoren $l, r \in \mathbb{R}^{n-1}$ und $r_{nn} \in \mathbb{R}$ gelten die Gleichungen

(i) $d = L_{n-1} \cdot r \Rightarrow r = L_{n-1}^{-1} \cdot d$

(ii) $a^t = l^t \cdot R_{n-1} \Leftrightarrow a = R_{n-1}^t \cdot l \Rightarrow l = \left(R_{n-1}^t \right)^{-1} \cdot a$. Dies lässt sich ausrechnen, da R_{n-1} regulär ist.

(iii) $a_{nn} = l^t \cdot r + r_{nn} \Rightarrow r_{nn} = a_{nn} - l^t \cdot r$ □

Der Satz (3.8) garantiert uns $a_{jj}^{(j)} \neq 0$. Behandeln wir daher nun den Fall $a_{jj}^{(j)} = 0$. In diesem Fall wenden wir die sogenannte Pivotsuche an.

Dazu benötigen wir Permutationsmatrizen. Sei

$$e_i := \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \; i\text{ter kanonischer Einheitsvektor}$$

$e_i^t := (0, \ldots, 0, 1, 0, \ldots, 0)$ Zeilenvektor

Eine (n,n)-Matrix P heißt Permutationsmatrix, falls eine Permutation $(i_1, \ldots, i_n)$ von $(1, \ldots, n)$ existiert mit

$$P = \begin{pmatrix} e_{i_1}^t \\ \vdots \\ e_{i_n}^t \end{pmatrix}$$

Beachte:

a) Die Linksmultiplikation PA permutiert die Zeilen von A

b) Die Rechtsmultiplikation AP permutiert die Spalten von A

Beispiel:

$$\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} = \begin{pmatrix} 4 & 5 & 6 \\ 7 & 8 & 9 \\ 1 & 2 & 3 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} \cdot \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 3 & 1 & 2 \\ 6 & 4 & 5 \\ 9 & 7 & 8 \end{pmatrix}$$

Insbesondere benötigen wir

$$P = \begin{pmatrix} 1 & & & & & & & & 0 \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & 0 & 0 & \ldots & 0 & 1 & \\ & & & \vdots & 1 & & & \vdots & \\ & & & \vdots & & \ddots & & \vdots & \\ & & & \vdots & & & 1 & \vdots & \\ & & & 1 & 0 & \ldots & \ldots & 0 & \\ & & & & & & & & 1 \\ & & & & & & & & & \ddots \\ 0 & & & & & & & & & & 1 \end{pmatrix}$$

PA vertauscht die Zeilen j und r von A. Es gilt $\boxed{P^2 = E_n.}$

Gauß-Elimination und LR-Zerlegung mit Pivotsuche

Ausgangsmatrix vor dem $j$ten Schritt:

$$A^{(j)} = \begin{pmatrix} a_{11}^{(1)} & \dots & \dots & \dots & a_{1n}^{(1)} \\ & \ddots & & & \vdots \\ & & a_{jj}^{(j)} & \dots & a_{jn}^{(j)} \\ & & \vdots & & \vdots \\ 0 & & a_{nj}^{(j)} & \dots & a_{nn}^{(j)} \end{pmatrix} \tag{3.9}$$

Spaltenpivotsuche: Wähle $r \geq j$ mit $|a_{rj}^{(j)}| = \max\limits_{j \leq i \leq n} |a_{ij}^{(j)}|$.

1. **Fall:** $a_{rj}^{(j)} = 0 \Rightarrow A$ ist singulär (d.h. $\det(A) = 0$), da $A^{(j)}$ singulär ist. Setze hier $L_j = E_n$.
2. **Fall:** $a_{rj}^{(j)} \neq 0 \Rightarrow$ Vertausche die $j$te Teile mit der $r$ten Zeile. Dies entspricht einer Linksmultiplikation $P_j A^{(j)}$, wobei P_j definiert ist wie weiter oben.

Nach $n-1$ Schritten erhält man

$$L_{n-1}P_{n-1} \cdot \ldots \cdot L_2 P_2 L_1 P_1 A = R \tag{3.10}$$

Folgerung: Man beachte $|l_{ij} \leq 1|$, denn $l_{ij} = \frac{a_{ij}^{(j)}}{a_{rj}^{(j)}}$.

Ziel: $PA = LR$ mit $P := P_{n-1} \cdot \ldots \cdot P_2 P_1$.

(3.11) Hilfssatz. Sei $j < k$. Die Permutationsmatrix $P_k$vertausche die Zeilen k und $r \geq k$. Dann gilt

$$P_k L_j = L_j' P_k,$$

wobei L_j' aus L_j dadurch hervorgeht, dass in der $j$ten Spalte von L_j das $k$te und das $r$te Element vertauscht werden.

Beweis.

$$L_j = \begin{pmatrix} 1 & & & & & & 0 \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & \vdots & & & & \\ & & -l_{kj} & 1 & & & \\ & & \vdots & & \ddots & & \\ & & -l_{rj} & & & 1 & \\ & & \vdots & & & & \ddots \\ 0 & & -l_{nj} & & & & 1 \end{pmatrix}, \; P_k L_j = \begin{pmatrix} 1 & & & & & & 0 \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & \vdots & & \ddots & & \\ & & -l_{rj} & & 0 & & 1 \\ & & \vdots & & & \ddots & \\ & & -l_{kj} & & 1 & & 0 \\ & & \vdots & & & & \ddots \\ 0 & & -l_{nj} & & & & 1 \end{pmatrix}$$

(Achtung: In der rechten Matrix stehen zwischen den beiden Nullen neben $-l_{rj}$ und $-l_{kj}$ natürlich Einsen auf der Diagonalen!)

Wir erkennen

$$P_k L_j P_k = L_j' \text{ (beachte: Rechtsmultiplikation mit } P_k \text{ vertauscht Spalten } r \text{ und } k)$$

Wegen $P^2 = E_n \Leftrightarrow P_k^{-1} = P_k$ folgt dann

$$P_k L_j = \underbrace{P_k L_j P_k P_k}_{=L_j'} = L_j' P_k.$$

Ziel: $PA = LR$ mit $P = P_{n-1} \cdots P_2 P_1$.

Anwendung von (3.11) für $n = 4$

$$\begin{aligned} & L_3P_3L_2P_2L_1P_1A &&= R \\ \Leftrightarrow\ & L_3P_3L_2L_1'P_2P_1A &&= R \\ \Leftrightarrow\ & L_3L_2'L_1''P_3P_2P_1A &&= R \\ \Leftrightarrow\ & PA &&= LR, \end{aligned}$$

wobei im letzten Schritt $P := P_3P_2P_1$ und $L := (L_1'')^{-1}(L_2')^{-1}L_3^{-1}$ gilt.
Die Anwendung auf die erweiterte Matrix $(A|b)$ führt auf die Matrix $(R|c)$. Dann ist R regulär, wenn A regulär ist.

(3.12) Satz (LR-Zerlegung und Gauß-Elimination). Zu jeder (n,n)-Matrix A gibt es eine Permutationsmatrix P, eine linke Dreiecksmatrix L und eine rechte Dreiecksmatrix R mit $PA = LR, l_{jj} = 1 \forall j = 1, \ldots, n, |l_{ij}| \leq 1 \forall 1 \leq j \leq i \leq n$. Ist A regulär, dann ist auch R regulär.

Praktische Durchführung: Die wesentlichen Elemente von L, d.h. die Elemente $l_{ik}, i > k, k < j$, können auf den Nullelementen der Matrix A gespeichert werden.

Beispiel:

$$\begin{pmatrix} 2 & 1 & 3 \\ 3 & 1 & 6 \\ 1 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 7 \\ 2 \\ 4 \end{pmatrix}$$ (Pivotelement für die erste Spalte: 3)

1. Schritt: Vertausche 1. und 2. Zeile:

$$P_1(A|b) = \left(\begin{array}{ccc|c} 3 & 1 & 6 & 2 \\ 2 & 1 & 3 & 7 \\ 1 & 1 & 1 & 4 \end{array}\right), P_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Anwendung von L_1 und Abspeicherung von l_{21} und l_{31}

$$\left(\begin{array}{ccc|c} 3 & 1 & 6 & 2 \\ \boxed{\frac{2}{3}} & \frac{1}{3} & -1 & \frac{17}{3} \\ \boxed{\frac{1}{3}} & \frac{2}{3} & -1 & \frac{10}{3} \end{array}\right), L_1 = \begin{pmatrix} 1 & 0 & 0 \\ -\frac{2}{3} & 1 & 0 \\ -\frac{1}{3} & 0 & 1 \end{pmatrix}.$$ (Pivotelement für die zweite Spalte: $\frac{2}{3}$

2. Schritt: Vertausche 2. und 3. Zeile:

$$P_2(L_1(P_1(A|b))) = \left(\begin{array}{ccc|c} 3 & 1 & 6 & 2 \\ \boxed{\frac{1}{3}} & \frac{2}{3} & -1 & \frac{10}{3} \\ \boxed{\frac{2}{3}} & \frac{1}{3} & -1 & \frac{17}{3} \end{array}\right), P_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

Anwendung von L_2 und Abspeicherung von l_{32}

$$\left(\begin{array}{ccc|c} 3 & 1 & 6 & 2 \\ \boxed{\frac{1}{3}} & \frac{2}{3} & -1 & \frac{10}{3} \\ \frac{2}{3} & \boxed{\frac{1}{2}} & -\frac{1}{2} & 4 \end{array}\right), L_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -\frac{1}{2} & 1 \end{pmatrix}.$$

Insgesamt erhalten wir damit

$$\begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{3} & 1 & 0 \\ \frac{2}{3} & \frac{1}{2} & 1 \end{pmatrix}, P = P_2P_1 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}.$$

Es gilt nun $PA = LR$. Wir haben ein gestaffeltes LGS, welches wir lösen können:

$$\begin{pmatrix} 3 & 1 & 6 \\ 0 & \frac{2}{3} & -1 \\ 0 & 0 & -\frac{1}{2} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 2 \\ \frac{10}{3} \\ 4 \end{pmatrix} \Rightarrow \begin{array}{rcl} x_3 &=& -8 \\ x_2 &=& -7 \\ x_1 &=& 19 \end{array}$$

Programm (Pseudocode) zur LR-Zerlegung $PA = LR$ *(Die Permutationen werden im Vektor* $p = (p_1, \ldots, p_n)$ *gespeichert)*

```
für j = 1,...,n : p_j = j
für j = 1,...,n-1 :

  Pivotsuche:
  amax := |a_jj|, r := j
  für i = j+1,...,n :
    falls |a_ij| > amax :
      amax := |a_ij|, r := j
  falls amax = 0 : STOP, A singulär!

  Zeilentausch: (j ↔ r)
  falls r > j
    für k = 1,...,n
      hr := a_jk, a_jk := a_jr, a_jr := hr
      hp := p_j, p_j := p_r, p_r := hp

  Tranformation:
  für i = j+1,...,n:
    a_ij := a_ij / a_jj (Berechnung von l_ij)
      für k = j+1,...,n
        a_ik := a_ik - a_ij · a_jk
```

Beim Zeilentausch und der Transformation lässt man k von 1 bzw. $j+1$ bis $n+1$ laufen, um das ganze LGS umzuformen.

Gauß-Elimination zur Lösung der Gleichung $Ax = b$	**Anzahl benötigter Rechenoperationen (1 Flop $\hat{=}$ 1 Multiplik. + 1 Addition)**
$PA = LR, p = (p_1, \ldots, p_n)$ Permutationsvektor	$\sum_{j=1}^{n-1} \left[(n-j) + (n-j)^2\right] = \frac{1}{2}n(n-1) + \frac{1}{6}n(n-1)(2n-1) = \frac{1}{3}(n^3 - n)$
$Lc = Pb, c_i = b_{p_i} - \sum_{k=1}^{i-1} l_{ik}c_k \, \forall i = 1, \ldots, n$ (Vorwärtseinsetzen)	$1 + 2 + \ldots + (n-1) = \frac{1}{2}(n^2 - n)$
$Rx = c, x_i = \frac{1}{r_{ii}}\left(c_i - \sum_{k=i+1}^{n} r_{ik}x_k\right) \, \forall i = n, \ldots, 1$ (Rückwärtseinsetzen)	$1 + 2 + \ldots + n = \frac{1}{2}(n^2 + n)$

In der Gesamtsumme benötigt die Gauß-Elimination also

$$\frac{1}{3}(n^3 - n) + \frac{1}{2}(n^2 - n) + \frac{1}{2}(n^2 + n) = \frac{1}{3}n^3 + n^2 - \frac{1}{3}n = O(n^3)$$

Flops.

§ 4 Spezielle Matrizen. Das Cholesky-Verfahren.

Sei A eine (n,n)-Matrix, deren Hauptabschnittsmatrizen $A_j = (a_{ik})_{1\le i,k\le j}$ regulär sind für $j = 1,\dots,n$. Nach Satz (3.8) existiert die LR-Zerlegung (ohne Pivoting)

$$A = LR.$$

Setze $D := \operatorname{diag}(r_{ii})$. Dann gilt

$$A = LDR, \tag{4.1}$$

wobei L eine linke und R eine rechte Dreiecksmatrix ist, mit $l_{jj} = 1, r_{jj} = 1$ für $j = 1,\dots,n$.

(4.2) Definition. Eine (n,n)-Matrix heißt diagonaldominant, wenn gilt

$$\sum_{k=1,k\neq i}^{n} |a_{ik}| < |a_{ii}|$$

für $i = 1,\dots,n$. Man sagt, A erfüllt das starke Zeilensummenkriterium.

(4.3) Satz. Bei einer diagonaldominanten Matrix A sind alle Hauptabschnittsmatrizen A_j regulär für $j = 1,\dots,n$. Also existiert die LR-Zerlegung $A = LR$.

Beweis. Für A_j gelte $A_j x = 0$ für $x \in \mathbb{R}^j$. Zu zeigen ist dann $x = 0$.
Annahme: Es gelte $|x_r| = \max_{1\le k\le j} |x_k| > 0$.
Betrachte die $r$te Gleichung

$$\sum_{k=1}^{j} a_{rk}x_k = 0 \Leftrightarrow a_{rr}x_r = -\sum_{k=1,k\neq r}^{j} a_{rk}x_k.$$

Es folgt daraus nun

$$|a_{rr}||x_r| = \left|\sum_{k=1,k\neq r}^{j} a_{rk}x_k\right| \le \sum_{k=1,k\neq r}^{j} |a_{rk}| \underbrace{|x_k|}_{\le |x_r|} \le \left(\sum_{k=1,k\neq r}^{j} |a_{rk}|\right)\cdot |x_r|.$$

Wegen $|x_r| > 0$ kann man die letzte Gleichung durch $|x_r|$ dividieren und erhält

$$|a_{rr}| \le \sum_{k=1,k\neq r}^{j} |a_{rk}|.$$

Da aber nach Voraussetzung gerade A diagonaldominant ist, ist dies ein Widerspruch! □

Beispiele:

$$A = \begin{pmatrix} 4 & -2 & -1 \\ 3 & 5 & 1 \\ 0 & 1 & 2 \end{pmatrix},\quad A = \begin{pmatrix} 4 & 1 & & & 0 \\ 1 & 4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 4 & 1 \\ 0 & & & 1 & 4 \end{pmatrix} \text{ (Tridiagonalmatrix)}$$

Tridiagonalmatrizen treten auf bei der Spline-Interpolation. A ist diagonaldominant und daher LR-zerlegbar.

Spezielle Matrizen A, die die Voraussetzungen von Satz (3.8) erfüllen, sind positiv definite Matrizen. Eine Matrix A heißt positiv definit, wenn gilt

(i) $A = A^t$ (Symmetrie)

(ii) $x^t Ax > 0$ für alle $x \in \mathbb{R}^n \setminus \{0\}$.

(4.4) Satz. Sei A positiv definit.

(i) Alle Hauptabschnittsmatrizen $A_j = (a_{ik})_{1 \le i,k \le n}$ von A sind positiv definit und regulär.

(ii) Es gibt genau eine linke Dreiecksmatrx L mit $l_{ii} > 0$ für $i = 1, \ldots, n$, so dass gilt

$$A = LL^t.$$

(Achtung: Es wird nicht $l_{ii} = 1$ gefordert!)

Beweis zu (i). Setze $x^t = (y^t, 0, \ldots, 0), y \in \mathbb{R}^j, y \neq 0$. Dann gilt

$$0 < x^t A x = y^t A_j y \Rightarrow A_j \text{ ist positiv definit.}$$

Beweis zu (ii). Nach Satz (3.8) gibt es genau eine Zerlegung

$$A = U \cdot V$$

mit $U = (u_{ik})$ linke Dreiecksmatrix, $u_{ii} = 1$ für $i = 1, \ldots, n$ und $V = (v_{ik})$ reguläre rechte Dreiecksmatrix. Setze $D = \operatorname{diag}(v_{ii}), v_{ii} \neq 0$ für $i = 1, \ldots, n$ und setze $R := D^{-1}V$, rechte Dreiecksmatrix mit $r_{ii} = 1$ für $i = 1, \ldots, n$. Daraus folgt

$$A = U \cdot D \cdot R, A = A^t = R^t D U^t.$$

Wegen der Eindeutigkeit der Zerlegung (Satz (3.8)) folgt schon $\boxed{R^t = U}$. Es gilt also

$$A = U \cdot D \cdot U^t = R^t \cdot D \cdot R.$$

Behauptung. D ist positiv definit, d.h. es gilt $v_{ii} > 0$ für $i = 1, \ldots, n$.
Beweis. Nach Voraussetzung gilt für alle $x \in \mathbb{R}^n \setminus \{0\}$

$$0 < x^t A x = x^t R^t D R x = (Rx)^t D (Rx).$$

Daraus folgt $0 < y^t D y$ für alle $y \in \mathbb{R}^n \setminus \{0\}$, da R regulär ist. Also ist D positiv definit. q.e.d.

Setze nun $D^{\frac{1}{2}} = \operatorname{diag}(\sqrt{v_{ii}})$ für $i = 1, \ldots, n$ und setze $L := U \cdot D^{\frac{1}{2}}$ linke Dreiecksmatrix mit $l_{ii} > 0$ für $i = 1, \ldots, n$. Dann gilt insgesamt

$$A = LL^t$$

□

Berechnung der Elemente l_{ik} von L in der Zerlegung $A = LL^t$

Durch Ausmultiplizieren von $A = (a_{ik}) = LL^t$ erhält man für $k = 1, \ldots, n$:

$$\begin{aligned} a_{kk} &= l_{k1}^2 + \ldots + l_{kk}^2 \\ a_{ik} &= \sum_{j=1}^{k-1} l_{ij} \cdot l_{kj} \text{ für } i = k+1, \ldots, n \end{aligned}$$

Nach Satz (4.4) ist die Auflösung nach l_{ik} möglich:

$$\left\{ \begin{array}{l} \text{für } k = 1, \ldots, n: \\ l_{kk} = \sqrt{a_{kk} - \sum_{j=1}^{k-1} l_{kj}^2} \\ l_{ik} = \frac{1}{l_{kk}} \cdot \left(a_{ik} - \sum_{j=1}^{k-1} l_{ij} l_{kj} \right) \text{ für } i = k+1, \ldots, n \end{array} \right\} \tag{4.5}$$

Insbesondere erhält man die Abschätzung $|l_{kj}| \le \sqrt{a_{kk}}$ für $j = 1, \ldots, k$, $k = 1, \ldots, n$.

Rechenoperationen: n Wurzeln, $\boxed{\frac{1}{6}n^3 + \frac{n^2}{2} - \frac{2}{3}n}$.

Programm (Pseudocode) zur Cholesky-Zerlegung

```
für k=1,...,n :  (spaltenweise)
```

$$l_{kk} = \sqrt{a_{kk} - \sum_{j=1}^{k-1} l_{kj}^2}$$

```
    für i=k+1,...,n :
```

$$l_{ik} = \frac{1}{l_{kk}} \left(a_{ik} - \sum_{j=1}^{k-1} l_{ij} l^{kj} \right)$$

```
    end
end
```

Die Lösung des LGS $Ax = b$ erfolgt in 3 Schritten:

$$\left\{ \begin{array}{ll} 1.\ A = L \cdot L^t & \text{Cholesky-Zerlegung} \\ 2.\ Lc = b & \text{Vorwärtsrekursion} \\ 3.\ L^t x = c & \text{Rückwärtsrekursion} \end{array} \right\} \tag{4.6}$$

Für positiv definite Matrizen A gilt

$$a_{ii} = e_i^t A e_i > 0.$$

Man kann zeigen: jede diagonaldominante Matrix A mit $a_{ii} > 0$ für $i = 1, \ldots, n$, d.h. $a_{ii} > \sum_{k=1, k \neq i}^{n} |a_{ik}|$, ist bereits positiv definit.

Für positiv definite Matrizen A gilt

$$x^t A x = x^t L L^t x = (L^t x)^t \cdot (L^t x) = \sum_{j=1}^{n} \left(\sum_{k=j}^{n} l_{kj} x_k \right)^2,$$

d.h. die quadratische Form $x^t A x$ kann als Summe von Quadraten geschrieben werden. Für $A_j = (a_{ik})_{i \leq i, k \leq n}$ gilt $\det(A_j) = \prod_{i=1}^{j} l_{ii}^2$.

(4.7) Folgerung. Eine symmetrische Matrix A ist genau dann positiv definit, wenn $\det(A_j) > 0$ für $j = 1, \ldots, n$

Beispiel: (vgl. Aufgabe 11) Diskretisierung von Randwertaufgaben bei gewöhnlichen DGL. Sei

$$A_n = \begin{pmatrix} 2 & -1 & & 0 \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 2 \end{pmatrix}.$$

A_n ist tridiagonal, aber nicht diagonaldominant. A_n erfüllt das schwache Zeilensummenkriterium.

Behauptung: A_n ist positiv definit. Man überlegt sich mit Hilfe des Determinantenentwicklungssatzes, dass folgende 3-Term-Rekursion gilt:

$$\det(A_{n+1}) = 2 \det(A_n) - \det(A_{n-1}),\ n \geq 2$$

$$\det(A_1) = 2,\ \det(A_2) = 3$$

Lösung: Man erkennt $\det(A_n) = n + 1$ für alle $n \in \mathbb{N}_+$. Nach (4.7) ist damit A_n positiv definit.

§ 5 Fehlerabschätzungen bei linearen Gleichungssystemen

Seien A eine reguläre $n \times n$-Matrix und $b \in \mathbb{R}^n$. Sei $x = A^{-1}b \in \mathbb{R}^n$ die eindeutige Lösung des LGS $Ax = b$. Es stellt sich die Frage, wie sich Fehler in A und b, d.h. Änderungen der Form

(i) $b \to b + \Delta b$

(ii) $A \to A + \Delta A = A \cdot (E_n + F)\,,\ F = A^{-1} \cdot \Delta A$

auf die Lösung $x \in \mathbb{R}^n$ auswirken.

Beispiel: Löse $Ax = b$ mit $A = \begin{pmatrix} 1 & 1 \\ 1 & 0.99 \end{pmatrix}$, $b = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \Rightarrow x = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$. Die Spalten von A sind nun "fast" linear abhängig. Wir führen nun eine Störung in A ein:

$$\tilde{A} = \begin{pmatrix} 1.01 & 1.01 \\ 1 & 0.99 \end{pmatrix} \Rightarrow \Delta A = \begin{pmatrix} 0.01 & 0.01 \\ 0 & 0 \end{pmatrix}.$$

Die exakte Lösung von $\tilde{A}\tilde{x} = b$ ist $\tilde{x} = \begin{pmatrix} \frac{200}{101} \\ -\frac{100}{101} \end{pmatrix}$. Wir erkennen: Der (absolute) Fehler von in den Daten beträgt 1%, der Fehler in der Lösung jedoch etwa 100%!

Begründung: Die Matrix A hat eine hohe Kondition. Um diese formal zu definieren, benötigen wir als Fehlermaß zunächst Normen für Matrizen und Vektoren.

(5.1) Definition. Sei $\mathbb{K}$ der Körper $\mathbb{R}$ oder $\mathbb{C}$. Sei V ein Vektorraum über $\mathbb{K}$. Eine (Vektor-)Norm in V ist eine Abbildung $||\cdot|| : V \to \mathbb{R}$ mit

(i) $||x|| \geq 0$ für alle $x \in V$ oder $||x|| > 0$ für alle $x \in V$ mit $x \neq 0$.

(ii) $||\lambda x|| = |\lambda| \cdot ||x||$ für alle $x \in V$ und $\lambda \in \mathbb{K}$. (**Homogenität**)

(iii) $||x + y|| \leq ||x|| + ||y||$ für alle $x, y \in V$. (**Dreiecksungleichung**)

Beispiele: Sei $V = \mathbb{R}^n$ oder $V = \mathbb{C}^n$. Normen für V sind zum Beispiel

- $||x||_2 := \sqrt{x^t x} = \left(\sum_{k=1}^n x_k^2\right)^{\frac{1}{2}}$ für $x \in \mathbb{R}^n$ bzw. $||x||_2 = \left(\sum_{k=1}^n |x_k|^2\right)^{\frac{1}{2}}$ mit $|x_k|^2 = \overline{x_k} x_k$ für $x_k \in \mathbb{C}$. (**Euklidische oder L_2-Norm**)
- $||x||_1 := \sum_{k=1}^n |x_k|$ (L_1-**Norm**)
- $||x||_\infty := \max_{k=1,\ldots,n} |x_k|$ (**Maximums- oder L_∞-Norm**)

Für die Eigenschaften von Normen ist bekannt:

(i) Jede Norm im $\mathbb{R}^n$ ist gleichmäßig stetig bezüglich der üblichen Topologie.

(ii) Je zwei Normen $||\cdot||$ und $||\cdot||'$ in $\mathbb{R}^n$ sind äquivalent, d.h. es gibt $m, M \in \mathbb{R}$ mit $m||x||' \leq ||x|| \leq M||x||'$ für alle $x \in \mathbb{R}^n$. Dies gilt jedoch nur für endlich-dimensionale Vektorräume.

Eine Matrixnorm $||A||$ für eine $n \times n$-Matrix A ist eine Vektornorm im $\mathbb{R}^{n^2}$, d.h. es gilt

(i) $||A|| > 0$ für alle $A \neq 0$.

(ii) $||\lambda A|| = |\lambda| \cdot ||A||$ für alle $A \in \mathbb{R}^{n^2}$ und $\lambda \in \mathbb{R}$.

(iii) $||A + B|| \leq ||A|| + ||B||$ für alle $A, B \in \mathbb{R}^{n^2}$.

Beispiel. $||A||_F := \left(\sum_{i,k=1}^n a_{ik}^2\right)^{\frac{1}{2}}$ ist eine Matrix-Norm, die sogenannte FROBENIUS-Norm. Diese Norm ist verträglich mit der L_2-Norm, denn es gilt

$$||Ax||_2 \leq ||A||_F \cdot ||x||_2 \text{ für alle } x \in \mathbb{R}^n.$$

Wir wollen im Folgenden jedoch nicht diese, sondern die Norm der folgenden Definition verwenden:

(5.2) Definition. Sei $||\cdot||$ eine Norm im $\mathbb{R}^n$ und sei A eine $n \times n$-Matrix. Die Zahl

$$||A|| := \max_{x \neq 0} \frac{||Ax||}{||x||} = \max_{||x||=1} ||Ax||$$

heißt die der Vektornorm $||\cdot||$ zugeordnete Matrix-Norm oder Operator-Norm.

Beispiele.

(i) Die der Norm $||\cdot||_\infty$ zugeordnete Norm $||A||_\infty$ ist die Zeilensummen-Norm

$$||A||_\infty := \max_{i=1,\dots,n} \sum_{k=1}^{n} |a_{ik}|.$$

Beweis. Sei $x \in \mathbb{R}^n$ mit $||x||_\infty = \max_{k=1,\dots,n} |x_k| = 1$, d.h. $|x_k| \leq 1$, $k = 1, \dots, n$. Zunächst gilt nun

$$||Ax||_\infty = \max_{i=1,\dots,n} \left| \sum_{k=1}^{n} a_{ik} x_k \right| \overset{|x_k| \leq 1}{\leq} \max_{i=1,\dots,n} \sum_{k=1}^{n} |a_{ik}|.$$

Dieses Maximum werde nun für ein $j \in \{1, \dots, n\}$ angenommen. Wir definieren nun $\tilde{x} \in \mathbb{R}^n$ durch

$$\tilde{x}_k = \begin{cases} \frac{a_{jk}}{|a_{jk}|}, & \text{falls } a_{jk} \neq 0 \\ 0, & \text{sonst} \end{cases}.$$

Dann gilt $||\tilde{x}||_\infty = 1$ und

$$||A\tilde{x}||_\infty = \max_{i=1,\dots,n} \left| \sum_{k=1}^{n} a_{ik} \tilde{x}_k \right| \geq \left| \sum_{k=1}^{n} a_{jk} \tilde{x}_k \right| \overset{Def.\ \tilde{x}_k}{=} \sum_{k=1}^{n} |a_{jk}| = \max_{i=1,\dots,n} \sum_{k=1}^{n} |a_{ik}|$$

□

(ii) Die der L_1-Norm zugeordnete Norm ist $||A||_1 = \max_{k=1,\dots,n} \sum_{i=1}^{n} |a_{ik}|$ und heißt Spaltensummen-Norm.

(iii) Die der L_2- oder euklidischen Norm zugeordnete Norm heißt Spektralnorm und ist definiert als

$$||A||_2 = \sqrt{\rho(A^t A)}.$$

Hierbei nennt man $\rho(B)$ für eine $n \times n$-Matrix B den Spektralradius, der definiert ist als

$$\rho(B) := \max\{|\lambda| \,\big|\, \lambda \text{ ist Eigenwert (EW) von } B\}.$$

Dies sieht man wegen

$$||A||_2 = \max_{||x||=1} ||Ax||_2 = \max_{||x||=1} \sqrt{x^t A A^t x} = \max\{\sqrt{|\lambda|} \,\big|\, \lambda \text{ ist EW von } A^t A\}.$$

(iv) Sei T eine reguläre $n \times n$-Matrix (Transformation). Die T-Norm von A ist definiert als

$$||A||_T := \max_{x \neq 0} \frac{||TAx||}{Tx}.$$

Da T regulär ist, existiert zu jedem $y \in \mathbb{R}^n$ ein $x \in \mathbb{R}^n$ mit $y = Tx$. Wir betrachten daher $y = Tx \Rightarrow x = T^{-1}y$ und schreiben

$$||A||_T = \max_{x \neq 0} \frac{||TAx||}{||Tx||} = \max_{y \neq 0} \frac{||TAT^{-1}y||}{||y||} = ||TAT^{-1}||.$$

(5.3) Satz. Zugeordnete Matrix-Normen haben die folgenden Eigenschaften:

(i) Die Abbildung $A \to ||A||$ ist eine Norm im Vektorraum der Matrizen

(ii) $||AB|| \leq ||A|| \cdot ||B||$ für zwei $n \times n$-Matrizen A und B. Insbesondere folgt $||A^k|| \leq ||A||^k$ für $k \in \mathbb{N}$.

(iii) $||E|| = 1$ für die $n \times n$-Einheitsmatrix.

(iv) $||A|| = \min\{c \,\big|\, ||Ax|| \leq c \cdot ||x|| \text{ für alle } x \in \mathbb{R}^n\}$.

Beweis. Übungsaufgabe.

(5.4) Satz.

(i) Für jeden Eigenwert λ von A und jede Matrixnorm gilt $|\lambda| \leq ||A||$.

(ii) Zu jeder Matrix A und zu jedem $\varepsilon > 0$ gibt es eine Norm $||\cdot||_{A,\varepsilon}$ in $\mathbb{R}^n$, so dass für die zugeordnete Matrixnorm gilt

$$||A||_{A,\varepsilon} \leq \rho(A) + \varepsilon$$

mit $\rho(A) = \max\{|\lambda| \,\big|\, \lambda \text{ ist EW von } A\}$.

Beweis.

(i) Sei $x \in \mathbb{R}^n$ ein Eigenvektor zum EW λ, d.h. es gilt $Ax = \lambda x$. Weiterhin sei $||x|| = 1$. Dann folgt

$$|\lambda| = |\lambda| \cdot ||x|| = ||\lambda x|| = ||Ax|| \leq ||A||.$$

q.e.d.

(ii) Sei $D = \operatorname{diag}(1, \varepsilon, \varepsilon^2, \ldots, \varepsilon^{n-1})$. *Beachte:* Für eine Matrix B gilt, dass die Rechtsmultiplikation BD die $k$te Spalte von B mit ε^{k-1} multipliziert, die Linksmultiplikation DB ultipliziert die $k$te Zeile von B mit ε^{k-1}.
Sei nun $J = P^{-1}AP$ die JORDANsche Normalform von A mit regulärer Matrix P. Dann gilt

$$J = \begin{pmatrix} \lambda_1 & \mu_1 & & & 0 \\ & \lambda_2 & \mu_2 & & \\ & & \ddots & \ddots & \\ & & & \ddots & \mu_{n-1} \\ 0 & & & & \lambda_n \end{pmatrix}$$

mit $\mu_1, \ldots, \mu_{n-1} \in \{0, 1\}$. Dann gilt

$$C := D^{-1}JD = \begin{pmatrix} \lambda_1 & \varepsilon\mu_1 & & & 0 \\ & \lambda_2 & \varepsilon\mu_2 & & \\ & & \ddots & \ddots & \\ & & & \ddots & \varepsilon\mu_{n-1} \\ 0 & & & & \lambda_n \end{pmatrix}.$$

Sei $||x|| = ||x||_\infty$. Setze $T := PD$. Definiere nun

$$||A||_{A,\varepsilon} := ||T^{-1}AT||_\infty = ||(PD)^{-1}APD||_\infty = ||D^{-1}\underbrace{P^{-1}AP}_{=J}D||_\infty = ||C||_\infty \leq \rho(A) + \varepsilon.$$

Damit erfüllt $||\cdot||_{A,\varepsilon}$ die geforderten Eigenschaften. □

(5.5) Satz. Ist F eine $n \times n$-Matrix mit $||F|| < 1$, so gibt es $(E+F)^{-1}$ und es gilt

(i) $(E+F)^{-1} = \sum_{k=0}^{\infty}(-F)^k$. (NEUMANNsche Reihe)

(ii) $||(E+F)^{-1}|| \leq \frac{1}{1-||F||}$.

Beweis. Es gilt die Abschätzung

$$||\sum_{k=0}^{N}(-F)^k|| \leq \sum_{k=0}^{N}||F^k|| \leq \sum_{k=0}^{\infty}||F||^k = \frac{1}{1-||F||} < \infty.$$

Daher existiert die Reihe $\sum_{k=0}^{\infty}(-F)^k$. Durch gliedweise Multiplikation erhält man

$$(E+F)\sum_{k=0}^{\infty}(-F)^k = E.$$

Daraus folgt die Darstellung (i). Aus $||\sum_{k=0}^{\infty}(-F)^k \leq \frac{1}{1-||F||}||$ folgt damit auch die Abschätzung (ii).

(5.6) Definition. Die Kondition einer regulären $n \times n$-Matrix A bezüglich der gewählten Norm ist die Zahl

$$cond(A) = ||A|| \cdot ||A^{-1}||.$$

Wir wollen im Folgenden Störungen der Form $A \to A + \Delta A = A \cdot (E + F)$ mit $F = A^{-1} \cdot \Delta A$ betrachten.

(5.7) Hilfssatz. Sei A eine reguläre Matrix und sei ΔA eine Matrix mit

$$q := ||A^{-1}|| \cdot ||\Delta A|| = cond(A) \cdot \frac{||\Delta A||}{||A||} < 1.$$

Dann ist $A + \Delta A$ invertierbar und es gilt

$$||(A + \Delta A)^{-1}|| \leq \frac{||A^{-1}||}{1 - q}.$$

Beweis. Es gilt $A + \Delta A = A(E + F)$, $F = A^{-1}\Delta A$ und $||F|| \leq ||A^{-1}|| \cdot ||\Delta A|| = q < 1$. Nach Satz (5.5) existiert das Inverse von $(E + F)$ und es gilt

$$||(A + \Delta A)^{-1}|| = ||(E + F)^{-1}A^{-1}|| \leq ||(E + F)^{-1}|| \cdot ||A^{-1}|| \leq \frac{||A^{-1}||}{1 - q}.$$

□

(5.8) Satz. Sei A eine reguläre Matrix und x eine Lösung von $Ax = b$, $b \in \mathbb{R}^n$. Seien weiter $\Delta A, \Delta b$ Störungen von A und b mit

$$q = cond(A) \cdot \frac{||\Delta A||}{||A||} < 1.$$

Dann ist das gestörte System $(A + \Delta A) \cdot (x + \Delta x) = (b + \Delta b)$ eindeutig lösbar und es gilt

$$\frac{||\Delta x||}{||x||} \leq \frac{cond(A)}{1 - q} \left(\frac{||\Delta A||}{||A||} + \frac{||\Delta b||}{||b||} \right).$$

Beweis. Nach Hilfssatz (5.7) ist $A + \Delta A$ invertierbar. Daher ist $(A + \Delta A)(x + \Delta x) = (b + \Delta b)$ eindeutig lösbar. Es gilt

$$(A + \Delta A)\Delta x = \Delta b - \Delta A \cdot x.$$

Mit (5.7) folgt

$$||\Delta x|| \leq ||(A \cdot \Delta A)^{-1}|| \cdot ||\Delta b - \Delta A \cdot x|| \leq \frac{||A^{-1}||}{1 - q} \cdot (||\Delta b|| + ||\Delta A|| \cdot ||x||).$$

Wir folgern

$$\frac{||\Delta x||}{||x||} \leq \frac{||A^{-1}||}{1 - q} \cdot \left(\frac{||\Delta b||}{||x||} + ||\Delta A|| \right) = \frac{||A^{-1}||}{1 - q} \cdot \left(\frac{||\Delta b||}{||b||} \cdot \frac{||b||}{||x||} + \frac{||\Delta A||}{||A||} \cdot ||A|| \right).$$

Mit $||b|| = ||Ax|| \leq ||A|| \cdot ||x||$ folgt

$$\frac{||\Delta x||}{||x||} \leq \frac{||A|| \cdot ||A^{-1}||}{1 - q} \cdot \left(\frac{||\Delta b||}{||b||} + \frac{||\Delta A||}{||A||} \right).$$

Daraus folgt die Behauptung. □

Beispiele.

(i) $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$, $A^{-1} = \begin{pmatrix} -2 & 1 \\ \frac{3}{2} & -\frac{1}{2} \end{pmatrix}$. Wir erhalten $cond_\infty(A) = ||A||_\infty \cdot ||A^{-1}||_\infty = 7 \cdot 3 = 21$.

(ii) Wir betrachten Elementarmatrizen $L_j = \begin{pmatrix} 1 & & & & 0 \\ & \ddots & & & \\ & & 1 & & \\ & & -l_{j+1,j} & 1 & \\ & & \vdots & & \ddots \\ 0 & & -l_{n,j} & & 1 \end{pmatrix}$. Mit Pivotsuche gilt $|l_{ik}| \leq 1$. Wir erhalten daher $cond_\infty(L_j) = ||L_j||_\infty \cdot ||L_j^{-1}||_\infty \leq 2 \cdot 2 = 4$.

(iii) (vgl. Eingangsbeispiel) Man Löse $Ax = b$ für

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 0.99 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad x = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Wir erhalten zunächst

$$A^{-1} = \begin{pmatrix} -99 & 100 \\ 100 & -100 \end{pmatrix} \Rightarrow cond_\infty(A) = ||A||_\infty \cdot ||A^{-1}||_\infty = 2 \cdot 200 = 400.$$

Sei nun

$$\Delta A = \begin{pmatrix} 0.001 & 0.001 \\ 0 & 0 \end{pmatrix}$$

eine Störung von A, so gilt

$$q = cond_\infty(A) \cdot \frac{||\Delta A||_\infty}{||A||_\infty} = 400 \cdot 0.001 = 0.4 < 1.$$

Nun ist

$$x + \Delta x = \tilde{x} = \begin{pmatrix} \frac{1100}{1001} \\ \frac{-100}{1001} \end{pmatrix}$$

die exakte Lösung des gestörten LGS, es folgt also

$$\frac{||\Delta x||_\infty}{||x||_\infty} \approx 0.1.$$

Da - wie oben berechnet - $q < 1$ gilt, erhalten wir aus Satz (5.8) die Abschätzung

$$\frac{||\Delta x||_\infty}{||x||_\infty} \leq \frac{cond_\infty(A)}{1-q} \cdot \frac{||\Delta A||_\infty}{||A||_\infty} = \frac{400}{1-0.4} \cdot 0.001 = \frac{2}{3}.$$

Man sieht also, dass die Abschätzung in (5.8) deutlich zu hoch sein kann.

Motivation zur Kondition: Der schleifende Schnitt.

Sei $A = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$ eine 2×2-Matrix mit den Zeilen a_1 und a_2. Man löse das Gleichungssystem $Ax = b$, $b \in \mathbb{R}^2$. Als Lösung erhalten wir zwei Geradegleichungen

$$a_1 x = b_1 \text{ und } a_2 x = b_2.$$

Stehen diese beiden Geraden nahezu senkrecht aufeinander, so wirkt sich ein Fehler in den Eingabedaten (gekennzeichnet durch ein Verschieben der beiden Geraden) nicht sehr stark auf das Ergebnis aus. Sind jedoch a_1 und a_2 fast linear abhängig, so liegen die beiden Geraden beinahe aufeinander. Fehler in den Eingabedaten wirken sich nun sehr stark auf das Ergebnis aus.

§ 6 Die QR-Zerlegung einer Matrix, Verfahren von HOUSEHOLDER

Sei A eine $n \times n$-Matrix. A sei reell, aber nicht notwendig regulär.

- Die LR-Zerlegung lieferte uns $\boxed{A = LR}$, wobei L eine linke und R eine rechte Dreiecksmatrix ist.
- Die QR-Zerlegung von A lautet nun $\boxed{A = QR}$, wobei Q eine orthogonale Matrix ist, d.h. $Q^*Q = E$ [$*\hat{=}$ Transposition], und R eine rechte Dreiecksmatrix.

Motivation zur QR-Zerlegung.

Zur Lösung des LGS $Ax = b$ erzeugt man bei der LR-Zerlegung eine Sequenz $(A, b) = (A^{(1)}, b^{(1)}) \to (A^{(2)}, b^{(2)}) \to \ldots \to (A^{(j)}, b^{(j)}) \to \ldots \to (A^{(n)}, b^{(n)}) = (R, c)$. Dabei gilt $(A^{(j+1)}, b^{(j+1)}) = L_j \cdot (A^{(j)}, b^{(j)})$. Es sei $\varepsilon^{(j)}$ der Rundungsfehler bei der Berechnung von $(A^{(j)}, b^{(j)})$. Nach (5.8) gilt nun die Abschätzung

$$\frac{||\Delta x||}{||x||} \leq \sum_{j=1}^{n-1} \varepsilon^{(j)} \cdot cond(A^{(j)}).$$

Die GAUSS-Elimination ist nicht gutartig, falls $cond(A^{(j)}) \gg cond(A^{(1)}) = cond(A)$ gilt.

Idee: Wähle eine Matrix Q_j im Übergang $(A^{(j+1)}, b^{(j+1)}) = Q_j \cdot (A^{(j)}, b^{(j)})$, so dass gilt

$$cond(A^{(j+1)}) = cond(A^{(j)}) = \ldots = cond(A).$$

Von nun an sei mit $||\cdot||$ stets die euklidische Norm $||\cdot||_2$ bezeichnet. Es gilt

$$||x|| = ||x||_2 = \sqrt{x^*x},$$

woraus folgt, dass $||A|| = ||A||_2$ (Spektralnorm) die zugeordnete Matrixnorm ist.

(6.1) Hilfssatz. Für orthogonale Matrizen Q gilt

(i) $||Q||_2 = 1$.

(ii) $||QA||_2 = ||A||_2$ für alle Matrizen A.

(iii) Für reguläre Matrizen A gilt $cond_2(QA) = cond_2(A)$.

Beweis. Als Übungsaufgabe.

Das Verfahren von HOUSEHOLDER.

Sei $w \in \mathbb{R}^n$ mit $w^*w = 1$, d.h. $||w||_2 = 1$ gegeben. Sei

$$Q := E - 2ww^*.$$

Dann ist $ww^* = (w_i w_k)_{1 \leq i,k} \leq n$ eine $n \times n$-Matrix, die man auch dyadisches Produkt nennt. Dann ist Q orthogonal, denn es gilt

$$\begin{aligned} Q^*Q &= (E - 2ww^*)(E - 2ww^*) \\ &= E - 4ww^* + 4w \underbrace{w^*w}_{=1} w^* \\ &= E - 4ww^* + 4ww^* \\ &= E. \end{aligned}$$

Zusätzlich ist Q - wie man leicht sieht - symmetrisch. Nun bedeutet für $x \in \mathbb{R}^n$

$$Qx = (E - 2ww^*)x = x - 2(w^*x)w$$

eine Spiegelung an der Hyperebene

$$H := \{z \in \mathbb{R}^n \mid w^*z = 0\}.$$

Problem: Sei $x = (x_1, \ldots, x_n)^* \in \mathbb{R}^n, x \neq 0$, vorgegeben. Bestimme nun ein $w \in \mathbb{R}^n$ mit $w^*w = 1$ und

$$Qx = (E - 2ww^*)x = ce_1,$$

wobei $c \in \mathbb{R}$ und $e_1 = (1, 0, \ldots, 0)^* \in \mathbb{R}^n$. Wir wollen nun $w \in \mathbb{R}^n$ und $c \in \mathbb{R}$ berechnen.

Analytische Berechnung von Q.

Es soll gelten $Qx = ce_1$ mit $||x||_2 \overset{Q \text{ orth.}}{=} ||Qx||_2 = ||ce_1||_2 = |c|$, also $c = \pm||x||_2$. Mit

$$Qx = x - 2(w^*x)w = ce_1 \Rightarrow 2(w^*x)w = x - ce_1$$

erhalten wir, da ja $||w|| = 1$ gelten soll,

$$w = \frac{x - ce_1}{||x - ce_1||_2}.$$

Um Auslöschung zu vermeiden, wählen wir $c = -sign(x_1) \cdot ||x||_2$. Wir erhalten

$$||x - ce_1||_2^2 = (|x_1| + ||x||_2)^2 + x_2^2 + \ldots + x_n^2 = 2||x||_2 \cdot (|x_1| + ||x||_2)\,.$$

Zusammenfassung.

$$\begin{cases} Q = E - 2ww^* = E - 2\frac{(x-ce_1)(x-ce_1)^*}{||x-ce_1||_2^2} = E - \beta uu^* \\ c = -sign(x_1) \cdot ||x||_2, \beta := \frac{2}{||x-ce_1||_2^2} = \frac{1}{||x||_2 \cdot (|x_1|||x||_2)} \\ u = x - ce_1 = \begin{pmatrix} sign(x_1) \cdot (|x_1| + ||x||_2) \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \end{cases} \tag{6.2}$$

Beispiel.

Sei $x = \begin{pmatrix} 3 \\ 4 \end{pmatrix}$, $c = -sign(x_1) \cdot ||x||_2 = -5$. Es gilt $x - ce_1 = \begin{pmatrix} 3 \\ 4 \end{pmatrix} - (-5)\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 8 \\ 4 \end{pmatrix} = u$ und $||x - ce_1||_2^2 = 80$. Damit gilt

$$2 \cdot \frac{(x - ce_1)(x - ce_1)^*}{||x - ce_1||_2^2} = \frac{2}{80} \cdot \begin{pmatrix} 64 & 32 \\ 32 & 16 \end{pmatrix} = \begin{pmatrix} \frac{8}{5} & \frac{4}{5} \\ \frac{4}{5} & \frac{2}{5} \end{pmatrix}.$$

Damit errechnen wir nun

$$Q = E - 2 \cdot \frac{(x - ce_1)(x - ce_1)^*}{||x - ce_1||_2^2} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} \frac{8}{5} & \frac{4}{5} \\ \frac{4}{5} & \frac{2}{5} \end{pmatrix} = \begin{pmatrix} -\frac{3}{5} & -\frac{4}{5} \\ -\frac{4}{5} & \frac{3}{5} \end{pmatrix}.$$

Es sei nun $A := \begin{pmatrix} 3 & 1 \\ 4 & 2 \end{pmatrix}$. Dann ist

$$QA = \begin{pmatrix} -5 & -\frac{11}{5} \\ 0 & \frac{2}{5} \end{pmatrix} =: R.$$

QR-Zerlegung von A.

Man bilde eine Sequenz $A = A^{(1)} \to \ldots \to A^{(n)} = R$ mit $A^{(j+1)} = Q_j A^{(j)}$, wobei Q_j orthogonal ist.

$\underline{j\text{ter Schritt } (j = 1, \ldots, n-1)}$

$$A^{(j)} = \left(\begin{array}{ccc|ccc} * & \ldots & * & * & \ldots & * \\ & \ddots & \vdots & \vdots & & \vdots \\ & & * & * & \ldots & * \\ \hline & & & a_{jj}^{(j)} & \ldots & a_{jn}^{(j)} \\ & & & \vdots & & \vdots \\ 0 & & & a_{nj}^{(j)} & \ldots & a_{nn}^{(j)} \end{array}\right). \text{ Setze } x = \begin{pmatrix} a_{jj}^{(j)} \\ \vdots \\ a_{nj}^{(j)} \end{pmatrix} \in \mathbb{R}^{n-j+1}.$$

1. Fall $x = 0 \in \mathbb{R}^{n-j+1}$. Dann ist A singulär. Setze $Q_j = E$.

2. Fall $x \neq 0 \in \mathbb{R}^{n-j+1}$. Berechne nach (6.2) die orthogonale $(n-j+1) \times (n-j+1)$-Matrix $\tilde{Q}_j$ mit

$$\tilde{Q}_j \cdot \begin{pmatrix} a_{jj}^{(j)} \\ \vdots \\ a_{nj}^{(j)} \end{pmatrix} = c_j \cdot \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{R}^{n-j+1}.$$

Setze dann $Q_j := \left(\begin{array}{c|c} E_{j-1} & 0 \\ \hline 0 & \tilde{Q}_j \end{array}\right)$. Dann ist Q_j eine symmetrische, orthogonale $n \times n$-Matrix. Nach $n-1$ Schritten erhält man

$$Q_{n-1} \dots Q_2 Q_1 A = R = A^{(n)}. \tag{6.3}$$

Definiere die orthogonale $n \times n$-Matrix $Q := (Q_{n-1} \dots Q_2 Q_2)^{-1}$. Dann gilt $Q^2 = E$, d.h. $Q = Q^{-1}$. Es folgt die QR-Zerlegung

$$A = QR.$$

(6.4) Satz (QR-Zerlegung). Zu jeder $n \times n$-Matrix A existiert eine orthogonale $n \times n$-Matrix Q und eine rechte Dreiecksmatrix R mit

$$A = QR$$

und $rang(A) = rang(R)$. Insbesondere ist R regulär, wenn A regulär ist.

Beweis. Siehe obiges Konstruktionsverfahren. □

Erweiterung der QR-Zerlegung auf nicht-quadratische Matrizen.

Sei A eine $m \times n$-Matrix mit $m > n$. Hierzu bildet man eine Sequenz mit $A^{(j+1)} = Q_j A^{(j)}, A(1) = A$, wobei Q_j eine orthogonale $m \times m$-Matrix ist. Wegen $m > n$ erhält man nach n Schritten (statt $n-1$ im quadratischen Fall)

$$A^{(n+1)} = Q_n Q_{n-1} \dots Q_2 Q_1 A = \left(\frac{R}{0}\right) \in \mathbb{R}^{m \times n}. \tag{6.5}$$

Dabei ist $Q := Q_n \dots Q_1$ orthogonal.

§ 7 Überbestimmte lineare Gleichungssysteme, Lineare Ausgleichsprobleme, Diskrete Approximation

(7.1) Motivation: Ausgleichsgerade. Gegeben seien Messpunkte $t_1 < t_2 < \ldots < t_m$ und Messdaten $y_1, \ldots, y_m$. Gesucht ist eine Gerade $y = \alpha + \beta t$ mit $y_i = \alpha + \beta t_i$ für $i = 1, \ldots, m$.
Für $m > 2$ ist dies nicht möglich. Als Ersatzproblem betrachtet man die Methode der kleinsten Quadrate: Man sucht

$$\min_{\alpha,\beta} \sum_{i=1}^{m} (y_i - (\alpha + \beta t_i))^2 .$$

Dazu setzt man

$$A = \begin{pmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_m \end{pmatrix}, \quad b = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix}, \quad x = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \in \mathbb{R}^2.$$

Dann optimiert man

$$\min_{\binom{\alpha}{\beta}} ||b - Ax||_2^2.$$

Beispiel.

Gegeben seien die Daten $(t_1, t_2, t_3, t_4) = (0, 3, 4, 7) \in \mathbb{R}^4 \Rightarrow m = 4$ sowie die zugehörigen Messwerte $(b_1, b_2, b_3, b_4) = (1, 2, 6, 4) \in \mathbb{R}^4$. Als Ergebnis werden wir $x = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \frac{3}{2} \\ \frac{1}{2} \end{pmatrix}$ erhalten.

Allgemeine Aufgabenstellung.

Sei A eine $m \times n$-Matrix mit $m > n$ und den Zeilen $a_i \in \mathbb{R}^n$ mit $i = 1, \ldots, m$. Sei $b \in \mathbb{R}^m$. Das überbestimmte LGS

$$Ax = b \tag{7.2}$$

besitzt im Allgemeinen keine Lösung. Als Ersatzproblem betrachten wir das Optimierungsproblem

$$\min_{x \in \mathbb{R}^n} ||b - Ax||_2.$$

Dieses ist äquivalent zum Problem

$$\min_{x \in \mathbb{R}^n} ||b - Ax||_2^2 = \min_{x \in \mathbb{R}^n} \sum_{i=1}^{m} (b_i - a_i x)^2 \textbf{(Methode der kleinsten Quadrate)}. \tag{7.3}$$

Notwendige Optimierungbedingung für ein Minimum der Funktion $f(x) := ||b - Ax||_2^2$ ist

$$\Delta f(x) = 0 \in \mathbb{R}^n.$$

Wir berechnen

$$f(x) = (b - Ax)^*(b - Ax) = b^*b - b^+Ax - x^*A^*b + x^*A^*Ax.$$

Dann gilt

$$\Delta f(x) = -A^*b - A^*b + 2A^*Ax = 2\left(-A^*b + A^*Ax\right).$$

Wir definieren: Eine Normalgleichung ist eine Gleichung der Form

$$A^*Ax = A^*b. \tag{7.4}$$

Hinreichende Optimierungsbedingung für ein Minimum der Funktion $f(x)$ ist

$$\Delta f(x) = 0 \in \mathbb{R}^n \text{ und } D_x^2 f = f''(x) = 2A^*A > 0 \text{ positiv definit.}$$

Dies ist erfüllt, wenn $rang(A) = n$. Dann sind die Normalengleichungen (7.4) eindeutig lösbar:

$$\boxed{x = (A^*A)^{-1} \cdot A^*b}$$

Untersuchung des allgemeinen Falles.

Wir führen die Bezeichnungen

- $Kern(A) = ker(A) = \mathcal{N}(A) = \{x \in \mathbb{R}^n \mid Ax = 0\}$
- $Bild(A) = img(A) = range(A) = R(A) = \{Ax \in \mathbb{R}^m \mid x \in \mathbb{R}^n\}$

ein.

(7.5) Lemma.

(i) A^*A ist positiv semidefinit und genau dann positiv definit, falls $Kern(A) = \{0\}$.

(ii) Es gelten die Beziehungen $Kern(A^*A) = Kern(A)$ und $Bild(A^*A) = Bild(A^*)$.

(iii) Es gelten die orthogonalen Zerlegungen

- $Bild(A) \oplus Kern(A^*) = \mathbb{R}^m$
- $Bild(A^*) \oplus Kern(A) = \mathbb{R}^n$.

Beweis.

(i) Zunächst gilt $x^*A^*Ax = (Ax)^*(Ax) \geq 0$, d.h. A^*A ist positiv semidefinit. Nun ist $(Ax)^*(Ax) = 0 \Leftrightarrow Ax = 0$, d.h. $x \in Kern(A)$. Es folgt $x^*A^*Ax > 0$ für alle $x \in \mathbb{R}^n$, $x \neq 0$, falls $Kern(A) = \{0\}$.

(ii) Die Aussage $Kern(A) \subset Kern(A^*A)$ ist klar.
Es gilt $A^*Ax = 0 \Rightarrow x^*A^*Ax = 0 = (Ax)^*(Ax) \Rightarrow Ax = 0$, also $x \in Kern(A)$, d.h. $Kern(A^*A) \subset Kern(A)$. Also gilt schon

$$Kern(A) = Kern(A^*A).$$

Weiter gilt

$$\dim Bild(A^*A) = n - \dim Kern(A^*A) = n - \dim Kern(A) = rang(A) = \dim Bild(A^*).$$

Daraus folgt über ein Dimensionsargument dann $Bild(A^*A) = Bild(A^*)$.

(iii) Seien $y = Ax \in Bild(A)$ und $z \in Kern(A^*)$ gegeben. Es gilt $A^*z = 0$ und

$$\langle y, z\rangle = \langle Ax, z\rangle = \langle x, \underbrace{A^*z}_{=0}\rangle = 0,$$

also gilt $y \perp z$. Somit stehen $Bild(A)$ und $Kern(A^*)$ aufeinander senkrecht. Da außerdem gilt

$$\dim Bild(A) = m - \dim Kern(A^*),$$

folgt insgesamt

$$Bild(A) \oplus Kern(A^*) = \mathbb{R}^m.$$

Analog zeigt man $Bild(A^*) \oplus Kern(A) = \mathbb{R}^n$.

(7.6) Satz. Das lineare Ausgleichsproblem (7.3) besitzt mindestens eine Lösung $x_0 \in \mathbb{R}^n$, für die gilt $A^*Ax_0 = A^*b$.

Die Gesamtheit der Lösungen ist der affine Unterraum $x_0 + Kern(A)$.

Beweis. Wegen $Bild(A) \oplus Kern(A^*)$ gibt es zu $b \in \mathbb{R}^m$ zwei Vektoren $s = Ax_0 \in Bild(A)$ und $r \in Kern(A^*)$, d.h. $A^*r = 0$, mit $b = s + r = Ax_0 + r$, $x_0 \in \mathbb{R}^n$. Dann gilt

$$A^*b = A^*(Ax_0 + r) = A^*Ax_0 + \underbrace{A^*r}_{=0} = A^*Ax_0.$$

x_0 ist also eine Lösung von (7.3).

Behauptung. Jede Lösung $x_0 \in \mathbb{R}^n$ der Normalgleichung $A^*Ax_0 = A^*b$ löst das Ausgleichsproblem.

Beweis. Sei $x \in \mathbb{R}^n$ beliebig. Setze $z := Ax - Ax_0 \in Bild(A)$, $r := b - Ax_0$. Dann gilt $r \perp Bild(A)$, denn $A^*r = A^*b - A^*Ax_0 = 0 \Rightarrow r \in Kern(A^*)$. Daher gilt $z^*r = 0$. Wir erhalten

$$||b - Ax||_2^2 = ||b - Ax_0 + Ax_0 - Ax||_2^2 = ||r - z||_2^2 \overset{(*)}{=} ||r||_2^2 + ||z||_2^2 \geq ||r||_2^2 = ||b - Ax_0||_2^2.$$

Es gilt Gleichheit genau dann, wenn gilt

$$||z||_2^2 = 0 \Leftrightarrow A(x - x_0) = 0 \Leftrightarrow x - x_0 \in Kern(A) \Leftrightarrow x \in x_0 + Kern(A).$$

(An der Stelle $(*)$ gilt Gleichheit wegen $r \perp z$)

□

(7.1) Beispiel. Wir können das Beispiel (7.1) nun rechnerisch lösen. Es seien wieder Messpunkte $t_1 < t_2 < \ldots < t_m$ und Messdaten $y_1, \ldots, y_m$ gegeben. Gesucht ist eine Gerade $y = \alpha + \beta t$ mit $\sum_{i=1}^{4}(y_i - (\alpha + \beta t_t))^2$ minimal.
Wir haben

$$A = \begin{pmatrix} 1 & 0 \\ 1 & 3 \\ 1 & 4 \\ 1 & 7 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 2 \\ 6 \\ 4 \end{pmatrix}, \quad x = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \in \mathbb{R}^2.$$

Wir erhalten

$$A^*A = \begin{pmatrix} 4 & 14 \\ 14 & 74 \end{pmatrix}, \quad A^*b = \begin{pmatrix} 13 \\ 58 \end{pmatrix}.$$

Man löst nun $A^*A \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 13 \\ 58 \end{pmatrix}$ und erhält als Lösung $x^* = (\alpha, \beta) = (\frac{3}{2}, \frac{1}{2})$.

(7.7) Definition. $x^\dagger \in \mathbb{R}^n$ heißt verallgemeinerte (Moore-Penrose-)Lösung von $Ax = b$, wenn gilt

(i) $A^*Ax^\dagger = A^*b$

(ii) $x^\dagger$ hat minimale Norm $||x^\dagger||_2$ unter allen Lösungen x von $A^*Ax = A^*b$.

(7.8) Satz. Die verallgemeinerte Lösung von $Ax = b$ existiert und ist eindeutig bestimmt durch

$$A^*Ax^\dagger = A^*b, \quad x^\dagger \in Bild(A^*).$$

Beweis. Es gilt $\mathbb{R}^n = Kern(A) \oplus Bild(A^*)$. Sei x_0 Lösung von $A^*Ax_0 = A^*b$. Dann existiert eine Zerlegung

$$x_0 = x_1 + x_2 \text{ mit } x_1 \in Bild(A^*), x_2 \in Kern(A).$$

Es folgt

$$A^*Ax_0 = A^*Ax_1 + A^* \underbrace{Ax_2}_{=0} = A^*Ax_1 = A^*b,$$

d.h. x_1 genügt den Normalgleichungen.

Behauptung. $x^\dagger := x_1$ hat minimale Norm.
Beweis. Jede Lösung $x \in \mathbb{R}^n$ von $A^*Ax = A^*b$ hat die Darstellung

$$x = x_1 + z \text{ mit } z \in Kern(A), x_1 \perp z.$$

Es folgt

$$||x||_2^2 = ||x_1||_2^2 + ||z||_2^2 \geq ||x_1||_2^2.$$

Damit ist der Satz bewiesen. □

Folgerung.

Die Zuordnung $b \mapsto x^\dagger$ ist linear! Also gibt es eine $n \times m$-Matrix $A^\dagger$ mit

$$x^\dagger = A^\dagger b.$$

$A^\dagger$ heißt Pseudo-Inverse oder verallgemeinerte (Moore-Penrose-) Inverse.

Berechnung der Pseudo-Inversen.

1. Fall $rang(A) = n \leq m$. Dann ist A^*A regulär. Es gilt

$$A^*Ax^\dagger = A^*b \Rightarrow x^\dagger = (A^*A)^{-1}A^*b.$$

Daher ist $A^\dagger := (A^*A)^{-1}A^*$.

2. Fall $rang(A) \leq n \leq m$. Berechnung mit Singulärwertzerlegung.

Beispiel.

Es sei $A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}$, so ist $rang(A) = 1$ und es gilt $A^*A = \begin{pmatrix} 4 & 4 \\ 4 & 4 \end{pmatrix}$. Wählt man nun $b = e_i \in \mathbb{R}^4$ für $i = 1, \ldots, 4$, so erhält man

$$\begin{pmatrix} 4 & 4 \\ 4 & 4 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = A^*b = A^*e_i = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Als eine spezielle Lösung erhalten wir $x_0 = \begin{pmatrix} \frac{1}{4} \\ 0 \end{pmatrix}$. Wir berechnen nun

$$Kern(A) = \{(x_1, x_2) \in \mathbb{R}^2 \,|\, x_1 + x_2 = 0\} = \mathbb{R} \cdot \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

und

$$Bild(A^*) = \langle \{A^*e_i \,|\, i = 1, \ldots, 4\} \rangle = \mathbb{R} \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Es gilt $x_0 + Kern(A) = \begin{pmatrix} \frac{1}{4} \\ 0 \end{pmatrix} + \mathbb{R} \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \left\{ \begin{pmatrix} \frac{1}{4} + t \\ -t \end{pmatrix} \,|\, t \in \mathbb{R} \right\}$. Für ein $x \in x_0 + Kern(A)$ erhalten wir nun

$$||x||_2^2 = \left(\frac{1}{4} + t\right)^2 + t^2.$$

Dieser Ausdruck ist minimal für $t = \frac{1}{8}$. Damit ergibt sich die verallgemeinerte Lösung $x^\dagger$ zu

$$x^\dagger = \begin{pmatrix} \frac{1}{8} \\ \frac{1}{8} \end{pmatrix} = \frac{1}{8} \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Damit überlegt man sich

$$A^\dagger = \frac{1}{8} \cdot \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}.$$

Wir nehmen ab sofort stets $rang(A) = n < m$ *an.*

Nachteile bei der Lösung von $A^*Ax = A^*b$**.**

1. A^*A ist "schwierig" zu berechnen. Denn sei $A = \begin{pmatrix} 1 & 1 \\ \varepsilon & 0 \\ 0 & \varepsilon \end{pmatrix}$, so ist $A^*A = \begin{pmatrix} 1e\varepsilon^2 & 1 \\ 1 & 1 + \varepsilon^2 \end{pmatrix}$. Für $\varepsilon = \frac{1}{2}\sqrt{eps}$ ist dann aber $gl(A^*A) = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ singulär!
2. $cond(A^*A)$ ist im Allgemeinen zu groß.

Man vermeidet diese Nachteile mit der QR-Zerlegung aus §6. Denn es gibt eine Zerlegung

$$QA = \left(\frac{R}{0}\right)$$

mit R rechte Dreiecks- und Q orthogonale $m \times m$-Matrix. Setzt man nun

$$Qb = \begin{pmatrix} h_1 \\ h_2 \end{pmatrix}$$

mit $h_1 \in \mathbb{R}^n$ und $h_2 \in \mathbb{R}^{m-n}$, so erhält man

$$\begin{aligned} ||b - Ax||_2^2 &= ||Q \cdot (b - Ax)||_2^2 = ||Qb - QAx||_2^2 = || \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} - \begin{pmatrix} R \\ 0 \end{pmatrix} ||_2^2 \\ &= || \begin{pmatrix} h_1 - R \\ h_2 \end{pmatrix} ||_2^2 = ||h_1 \cdot Rx||_2^2 + ||h_2||_2^2. \end{aligned}$$

Dieser Ausdruck wird minimal, falls $Rx = h_1$, d.h. $x_0 = R^{-1}h_1$. Dann ist

$$||b - Ax||_2^2 = ||h_2||_2^2$$

die Länge des Residuums $r = b - Ax$.

Anwendung auf die diskrete Approximation.

Gegeben seien Messpunkte $(t_i, y_i) \in \mathbb{R}^2$ mit $i = 1, \ldots, m$ und Basisfunktionen $u_0(t), \ldots, u_n(t)$ mit $m \geq n + 1$. Beispielsweise seien $u_i(t) = t^i$ für $i = 0, \ldots, n$. Gesucht ist dann eine Linearkombination $u(t) = \sum_{k=0}^{n} \alpha_k u_k(t)$ mit $u(t_i) = y_i$ für alle $i = 1, \ldots, m$. Wir betrachten das Optimierungsproblem

$$\min_{\alpha_0, \ldots, \alpha_n} \sum_{i=0}^{m} (y_i - u(t_i))^2,$$

d.h. wir berechnen

$$\min_{\alpha_0, \ldots, \alpha_n} \sum_{i=0}^{m} (y_i - \sum_{k=0}^{n} \alpha_k u_k(t_i))^2.$$

Diese Problem hat die Form $\min_{x \in \mathbb{R}^{n+1}} ||b - Ax||_2^2$ mit

$$A = \begin{pmatrix} u_0(t_1) & u_1(t_1) & \ldots & u_n(t_1) \\ \vdots & & & \vdots \\ u_0(t_m) & u_1(t_m) & \ldots & u_n(t_m) \end{pmatrix}, \quad x = \begin{pmatrix} \alpha_0 \\ \vdots \\ \alpha_n \end{pmatrix} \in \mathbb{R}^{n+1}, \quad b = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} \in \mathbb{R}^m.$$

Beispiel.

Die Menge der Basisfunktionen sei $u_k(t) = t^k$ für $k = 0, \ldots, n$. Dann erhalten wir

$$A = \begin{pmatrix} 1 & t_1 & t_1^2 & \ldots & t_1^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & t_m & t_m^2 & \ldots & t_m^n \end{pmatrix},$$

die sogenannte Vandermonde-Matrix. Es ist $rang(A) = n + 1 \leq m$, falls $t_i \neq t_j$ für $i \neq j$. Der Fall $m = n + 1$ heißt Interpolation und wird von uns später behandelt. In diesem Fall ist $Ax = b$ eindeutig lösbar.

Anwendungen.

$n = 1$ Ausgleichsgerade (Regressionsanalyse in der Statistik). Es ist

$$A^*A = \begin{pmatrix} m & \sum_{i=1}^m t_i \\ \sum_{i=1}^m t_i & \sum_{i=1}^m t_i^2 \end{pmatrix}, \quad A^*b = \begin{pmatrix} \sum_{i=1}^m y_i \\ \sum_{i=1}^m y_i t_i \end{pmatrix}.$$

Die explizite Lösung von $A^*A \cdot \begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix} = A^*b$ erhält man durch Ausrechnen.

$n = 2$ Ausgleichsparabel (vgl. Übungsaufgabe 22).

$n = 3$ vgl. Aufgabe 23.

III Iterationsverfahren zur Lösung von Gleichungen

§ 8 Definitionen und Grundbegriffe

8.1 Nullstellen

Sei $D \subset \mathbb{R}^n$ und $f : D \to \mathbb{R}^n$, $f = (f_1, \ldots, f_n)^t$. Zu lösen sei die Gleichung

$$f(x) = 0, \quad x \in D, \text{ d.h. } f_i(x_1, \ldots, x_n) = 0 \, \forall i = 1, \ldots, n \tag{8.1}$$

Ein solcher Punkt $\overline{x} \in D$ mit $f(\overline{x})$ heißt Nullstelle von f.

Beispiele

(i) Polynome. $D \subset \mathbb{R}$ $(n = 1)$. Sei $f(x) = a_0 + a_1 x + \ldots + a_n x^n$, $a_i \in \mathbb{R}$ gegeben. Dieses Problem taucht auf bei der Berechnung von Eigenwerten (Nullstellen von charakteristischen Polynomen).

(ii) $f(x) = x - \tan(x) = 0$. Dieses Problem taucht auf bei der Berechnung von Schwingungen eines Balkens. Eine Lösung: $\overline{x} = 0$. In den Intervallen $(k\pi, (k + \frac{1}{2})\pi)$ liegen Nullstellen $\overline{x}_k$ für $k = 1, 2, \ldots$. Mit $\overline{x}_k$ ist auch $-\overline{x}_k$ Nullstelle.

(iii) Optimierungsprobleme. Sei $h : \mathbb{R}^n \to \mathbb{R}$ differenzierbar (Zielfunktion). Die Lösung des Optimierungsproblems

$$\min h(x)$$

ist eine Nullstelle der Funktion $f(x) = \nabla h(x) = \left(\frac{\partial h}{\partial x_1}(x), \ldots, \frac{\partial h}{\partial x_n}\right)^t$. Weiterhin gibt es sogenannte Minimierungsprobleme mit Nebenbedingungen. Man sucht dann $\min h(x)$ unter den Bedingungen

$$g_i(x) = 0 \text{ für } i = 1, \ldots, k \text{ und } g_i(x) \leq 0 \text{ für } i = k + 1, \ldots, m.$$

Die resultierenden Gleichungen nennt man Optimalitätsbedingungen von Kuhn, Tucker.

8.2 Fixpunkte

Sei $D \subset \mathbb{R}^n$ und $g : D \to \mathbb{R}^n$. Gesucht sind Lösungen der Fixpunktgleichung

$$x = g(x) \tag{8.2}$$

Ein Punkt $\overline{x} \in D$ mit $g(\overline{x}) = \overline{x}$ heißt Fixpunkt von g.

Sei $A(x)$ eine reguläre $n \times n$-Matrix, $x \in D$ und $f : D \to \mathbb{R}^n$, dann ist die Nullstellenbestimmung $f(x) = 0$ äquivalent zu einer Fixpunktgleichung

$$x = g(x) := x + A(x) f(x) \tag{8.3}$$

Anwendung: Zum Beispiel NEWTON-Verfahren mit $A(x) = f'(x)^{-1}$.

Iterationsverfahren.

Voraussetzung: $g(D) \subset D$, Startwert $x^0 \in D$.
Iteration:

$$x^{k+1} = g(x^k), \; k = 0, 1, \ldots \text{ (falls } n = 1 : x_{k+1} = g(x_k)) \tag{8.4}$$

Falls $\overline{x} := \lim_{k \to \infty} x^k$, g stetig und $\lim_{k \to \infty} x^{k+1} = \lim_{k \to \infty} g(x^k)$, so folgt $\overline{x} = g(\overline{x})$. Wir unterscheiden die folgenden Fälle:

1. Fall $0 \leq g'(\overline{x}) < 1$: anziehender Fixpunkt (Attraktor).

2. Fall $g'(\overline{x}) > 1$: abstoßender Fixpunkt.

3. Fall $-1 < g'(\overline{x}) \leq 0$: anziehender Fixpunkt

4. Fall $g'(\overline{x}) < -1$: abstoßender Fixpunkt.

Die Chaos-Abbildung.
Sei $g(x) := rx(1-x)$ mit $1 < r \leq 4$ und $D = [0,1]$. Es gilt $g(D) \subset D$. Diese Funktion besitzt zwei Fixpunkt $\overline{x}_1 = 0$ und $\overline{x}_2 = 1 - \frac{1}{r}$. Die Iteration lautet

$$x_{k+1} = rx_k(1-x_k)$$

Es ist $g'(x) = r - 2rx$, also $g'(\overline{x}_2) = r - 2r\overline{x}_2 = 2 - r$. Wir unterscheiden zwei Fälle:

$1 < r < 3$**:** $|g'(\overline{x}_2)| < 1 \Rightarrow x_k \overset{k\to\infty}{\to} 1 - \frac{1}{r}$.

$r > 3$**:** Es ist $g'(\overline{x}_2) < -1$. Die Funktion $g^2 = g \circ g$ hat zwei weitere Fixpunkte $\overline{x}_3, \overline{x}_4$ mit $x_4 = g(x_3)$ und $x_3 = g(x_4)$. Ist r nur wenig größer als 3, so hat man eine 2-periodische Folge mit $x_{k+2} \approx x_k$. Danach kommt es zu wiederkehrenden Periodenverdopplungen und Übergang ins Chaos.

Allgemein betrachtet man Startwerte $x^0, x^1, \ldots, x^s$ für ein $s \in \mathbb{N}$. Also Iteration wählt man

$$x^{k+1} = \Phi(x^k, x^{k-1}, \ldots, x^{k-s}),\ k = s, s+1, \ldots$$

Die Abbildung $\Phi : \underbrace{\mathbb{R}^n \times \ldots \mathbb{R}^n}_{s+1-\text{mal}} \to \mathbb{R}^n$ heißt Iterationsabbildung.

Beispiele: $s = 0, \Phi = g$. $x^{k+1} = \Phi(x^k) = g(x^k)$ oder für $s = 1$: Regula Falsi, Sekantenverfahren.

8.3 Konvergenzgeschwindigkeit

Beispiel: Sei $0 < q < 1$. Betrachte einen Abschnitt der geometrischen Reihe

$$x^k = \sum_{i=0}^{k} q^i = \frac{1-q^{k+1}}{1-q} \overset{k\to\infty}{\to} \frac{1}{1-q} =: \overline{x}.$$

Wir betrachten nun den Quotienten

$$\left| \frac{\overline{x} - x^{k+1}}{\overline{x} - x^k} \right| = q < 1,$$

d.h. $|\overline{x} - x^{k+1}| = q|\overline{x} - x^k|$.
Es sei nun $||\cdot||$ eine Norm im $\mathbb{R}^n$ und es sei $\{x^k\} \subset \mathbb{R}^n$ eine Folge mit $\overline{x} = \lim_{k\to\infty} x^k$. Falls $p \geq 1$ existiert, so dass gilt

$$c := \limsup_{k\to\infty} \frac{||\overline{x} - x^{k+1}||}{||\overline{x} - x^k||^p} \begin{cases} < 1 & \text{für } p = 1 \\ < \infty & \text{für } p > 1 \end{cases} \tag{8.5}$$

so heißt $\{x^k\}$ konvergent vom Grade p. Die Zahl p heißt der Konvergenzgrad. Man unterscheiden nun verschiedene Konvergenztypen, so z.B. lineare Konvergenz ($p = 1$) oder quadratische Konvergenz ($p = 2$).

Interpretation von (8.5): Es gibt ein

$$0 \leq c \begin{cases} < 1 & \text{falls } p = 1 \\ < \infty & \text{falls } p > 1 \end{cases}$$

und $k_0 \in \mathbb{N}$, so dass

$$||\overline{x} - x^{k+1}|| \leq c||\overline{x} - x^k||^p$$

für $k \geq k_0$. Die Konstante c in (8.5) heißt asymptotische Fehlerkonstante. Für den Fehler $e_k := ||\overline{x} - x^{k+1}||$ gilt

$$e_{k+1} \leq c \cdot e_k^p \text{ für alle } k \geq k_0 \Rightarrow e_{k+1} \in O(e_k^p).$$

§ 9 Nullstellen reeller Funktionen

Sei $D \subset \mathbb{R}$, $f : D \to \mathbb{R}$ und $\{x^k\} \subset \mathbb{R}$ eine reelle Iterationsfolge.

9.1 Intervallhalbierung, Bisektionsverfahren

Sei $D = [a, b]$ und es gelte $f(a)f(b) < 0$. f sei stetig. Dann gibt es ein $\overline{x} \in D$ mit $f(\overline{x}) = 0$. Wir berechnen $f(m)$ mit $m = \frac{a+b}{2}$ und wählen das Intervall, in dem die Nullstelle $\overline{x}$ liegen muss, d.h. entweder $[a, m]$, wenn $f(a)f(m) < 0$ oder $[m, b]$, wenn $f(m)f(b) < 0$. Die Mittelpunkte definieren eine Folge $\{x_k\} \subset \mathbb{R}$ mit

$$|x_k - \overline{x}| \leq \frac{b-a}{2^{k+1}}.$$

Beispiel: $f(x) = x - \tan(x)$, $a = 2$, $b = 4.6$. Es gilt $f(a) = 4.18$ und $f(b) = -4.26$. Wir berechnen

$$x_5 = 4.47812, \quad x_{10} = 4.49341, \quad \overline{x} \approx x_{100} = 4.49340946.$$

9.2 Newton-Verfahren

Sei : $[a, b] \to \mathbb{R}$ eine C^2-Funktion. Sei x_k eine Näherung für $\overline{x}$ mit $f(\overline{x}) = 0$. Approximiere

$$f(x) \approx T(x) = f(x_k) + f'(x_k)(x - x_k) \quad \text{(Taylorentwicklung erster Ordnung)}$$

Berechne nun x_{k+1} mit $0 = T(x_{k+1}) = f(x_k) + f'(x_k)(x_{k+1} - x_k)$. Wir erhalten

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} \tag{9.1}$$

Das Newton-Verfahren ist eine Fixpunktiteration:

$$x_{k+1} = g(x_k) \text{ mit } g(x) = x - \frac{f(x)}{f'(x)}.$$

Beispiele.

(i) $f(x) = x^2 - 2$, $\overline{x} = \sqrt{2}$, $g(x) = x - \frac{f(x)}{f'(x)} = \frac{1}{2}\left(x + \frac{2}{x}\right)$. Die Iteration lautet dann

$$x_{k+1} = \frac{1}{2}\left(x_k + \frac{2}{x_k}\right).$$

Als Ergebnis erhalten wir

k	x_k
0	1
1	1.5
2	1.417
3	1.414216
4	1.414213562

(ii) $f(x) = x - \tan(x)$, $\overline{x} = 4.49340945$. Für die Startwerte $x_0 = 2$ oder $x_0 = 4$ divergiert die Folge $\{x_k\}$. Für den Startwert $x_0 = 4.5$ erhält man $x_3 = 4.49340945 (\approx \overline{x})$.

Konvergenzverhalten von (9.1)

Es gilt $x_{k+1} - \overline{x} = x_k - \overline{x} - \frac{f(x_k)}{f'(x_k)}$. Wir betrachten die Taylor-Entwicklung um x_k:

$$0 = f(\overline{x}) = f(x_k) + f'(x_k)(\overline{x} - x_k) + \frac{1}{2}f''(x_k + \alpha(\overline{x} - x_k))(\overline{x} - x_k)^2,$$

wobei $0 \leq \alpha \leq 1$. Wir erhalten

$$x_k - \overline{x} - \frac{f(x_k)}{f'(x_k)} = \frac{1}{2}f''(x_k + \alpha(\overline{x} - x_k))(\overline{x} - x_k)^2 \frac{1}{f'(x_k)}.$$

Daher gilt nun

$$x_{k+1} - \overline{x} = \frac{1}{2}\frac{f''(x_k + \alpha(\overline{x} - x_k)}{f'(x_k)}(\overline{x} - x_k)^2.$$

Unter der Annahme, dass gilt $\overline{x} = \lim_{k\to\infty} x_k$ (Beweis folgt später), erhalten wir

$$\lim_{k\to\infty} \frac{|\overline{x} - x_{k+1}|}{|\overline{x} - x_k|^2} = \frac{1}{2}\left|\frac{f''(\overline{x})}{f'(\overline{x})}\right| =: c < \infty.$$

Das NEWTON-Verfahren hat also Konvergenzgrad 2, wenn es konvergiert. Den Konvergenzgrad des Verfahrens im $\mathbb{R}^n$ betrachten wir in §11.

9.3 Das Sekantenverfahren (Regular falsi)

Startwerte: x_0 und x_1. Iteration: Berechne die Verbindungsgerade (Sekante) zwischen den Punkten $(x_{k-1}, f(x_{k-1}))$ und $(x_k, f(x_k))$ und bestimme deren Nullstelle.

$$x_{k+1} = \Phi(x_k, x_{k-1}) = x_k - f(x_k)\frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} \tag{9.2}$$

Beachte: Es werden nur Funktionswerte benutzt, aber keine Ableitungen.

(9.3) Satz. Es sei $f : \mathbb{R} \to \mathbb{R}$ in einer Umgebung von $\overline{x}$ zweimal stetig differenzierbar. Es gelte $f(\overline{x}) = 0$ und $f'(\overline{x}) \neq 0$. Dann konvergiert die Iteration (9.2) gegen $\overline{x}$, falls die Startwerte x_1 und x_0 hinreichend nahe bei $\overline{x}$ liegen. Der Konvergenzgrad ist $p = \frac{1}{2}(1 + \sqrt{5})$.

Beweis. Siehe z.B. das Buch von Schwarz, Seite 201.

Beispiel: $f(x) = x - \tan(x)$.

k	x_k
0	4
1	4.6
5	4.6587
12	4.49340824
13	Division durch Null

§ 10 Konvergenzsätze für Iterationsverfahren

Sei $D \subset \mathbb{R}^n$ und $g : D \to \mathbb{R}^n$ mit $g(D) \subset D$. Gesucht ist ein Fixpunkt $\overline{x} \in D$ mit $\overline{x} = g(\overline{x})$. Wir erhalten die Fixpunktiteration

$$x^{k+1} = g(x^k) \text{ für } k = 0, 1, \ldots \text{ mit Startwert } x^0 \in D. \tag{10.1}$$

Sei $||\cdot||$ eine Norm im $\mathbb{R}^n$.

(10.2) Definition. $g : D \to \mathbb{R}^n$ heißt kontrahierend in D, falls $q \in \mathbb{R}_+$ existiert mit $0 \leq q < 1$ und

$$||g(x) - g(y)|| \leq q||x - y|| \, \forall x, y \in D.$$

Voraussetzung: g sei stetig differenzierbar. Unser Ziel ist die Berechnung von q. Wir definieren: Eine Menge $D \subset \mathbb{R}^n$ heißt konvex, falls $\alpha x + (1 - \alpha)y \in D$ für alle $x, y \in D$ und $0 \leq \alpha \leq 1$.

(10.3) Satz. Sei $D \subset \mathbb{R}^n$ konvex und sei $g : D \to \mathbb{R}^n$ stetig diffbar. Es gelte $\sup_{x \in D} ||g'(x)||_\infty < 1$. Dann ist g kontrahierend in D mit $q = \sup_{x \in D} ||g'(x)||_\infty$. Hierbei ist $g'(x) = Dg(x) = \left(\frac{\partial g_i}{\partial x_j}\right)_{i,j}(x)$.

Beweis. Seien $x, y \in D$ beliebig. Wir betrachten die Funktion $\varphi : [0, 1] \to \mathbb{R}^n$, gegeben durch

$$\varphi(t) := g(tx + (1 - t)y) \quad \text{für } t \in [0, 1].$$

Es gilt dann

$$\begin{aligned} \varphi(0) &= g(y) \\ \varphi(1) &= g(x) \\ \varphi'(t) &= g'(tx + (1 - t)y)(x - y) \end{aligned}$$

Aus dem Mittelwertsatz erhält man nun

$$\varphi(t) = (\varphi_1(t), \ldots, \varphi_n(t))^* \text{ mit } |\varphi_i(1) - \varphi_i(0)| \leq \max_{0 \leq t \leq 1} |\varphi_i'(t)|.$$

Damit berechnen wir

$$\begin{aligned} ||g(x) - g(y)||_\infty &= ||\varphi(1) - \varphi(0)||_\infty = \max_{i=1,\ldots,n} |\varphi_i(1) - \varphi_i(0)| \\ &\leq \max_{0 \leq t \leq 1} ||\varphi'(t)||_\infty = \max_{0 \leq t \leq 1} ||g'(tx + (1 - t)y)(x - y)||_\infty \\ &\leq \max_{0 \leq t \leq 1} ||g'(tx + (1 - t)y)||_\infty \cdot ||x - y||_\infty \\ &\leq \left(\sup_{z \in D} ||g'(z)||_\infty\right) \cdot ||x - y||_\infty \\ &= q \cdot ||x - y||_\infty \end{aligned}$$

für ein $q < 1$. □

Anwendung: Für $n = 1$ ist $D = [a, b] \subset \mathbb{R}$ konvex. $g : [a, b] \to \mathbb{R}$ ist kontrahierend, falls $\max_{x \in [a,b]} |g'(x)| < 1$ und falls g stetig differenzierbar ist.

(10.4) Satz (Fixpunktsatz von Banach) Sei $D \subset \mathbb{R}^n$ abgeschlossen und sei $g : D \to \mathbb{R}^n$ kontrahierend auf D mit $g(D) \subset D$. Dann konvergiert die Folge $x^{k+1} = g(x^k), k \geq 0$ für jeden Startwert $x^0 \in D$ gegen den eindeutig bestimmten Fixpunkt $\overline{x} \in D$ von g. Weiterhin gelten die Abschätzungen

(i) $||\overline{x} - x^k|| \leq \frac{q}{1-q}||x^k - x^{k-1}|| \leq \frac{q^k}{1-q}||x^1 - x^0||$ und

(ii) $||\overline{x} - x^{k+1}|| \leq q||\overline{x} - x^k||$ (Lineare Konvergenz)

Beweis. Wir zeigen, dass $\{x^k\}_{k \in \mathbb{N}} \subset \mathbb{R}^n$ eine Cauchy-Folge ist. Für $k \geq 1$ gilt

$$||x^{k+1} - x^k|| = ||g(x^k) - g(x^{k-1})|| \leq q||x^k - x^{k-1}|| \leq q^2||x^{k-1} - x^{k-2}|| \leq \ldots \leq q^k||x^1 - x^0||$$

Für alle $m \in \mathbb{N}$ gilt also

$$\begin{aligned} ||x^{k+m} - x^k|| &= || \sum_{j=0}^{m-1} \left(x^{k+j+1} - x^{k+j}\right) || \\ &\leq \sum_{j=0}^{m-1} ||x^{k+j+1} - x^{k+j}|| \leq \sum_{j=0}^{m-1} q^{k+j} ||x^1 - x^0|| \\ &= q^k \left(\sum_{j=0}^{m-1} q^j \right) ||x^1 - x^0|| \leq \frac{q^k}{1-q} ||x^1 - x^0||. \end{aligned} \tag{10.5}$$

Daher ist $\{x^k\}$ eine CAUCHY-Folge. Sei nun $\overline{x} := \lim_{k\to\infty} x^k$ der Grenzwert. Dann ist $\overline{x} \in D$, da D abgeschlossen ist. Es folgt

$$\overline{x} = \lim_{k\to\infty} x^{k+1} = \lim_{k\to\infty} g(x^k) = g(\overline{x}),$$

d.h. $\overline{x}$ ist ein Fixpunkt von g. Weiterhin gilt für $m \to \infty$ in (10.5) bei festem k die Abschätzung

$$||\overline{x} - x^k|| \leq \frac{q^k}{1-q} ||x^1 - x^0||.$$

Führt man die Ersetzungen $x^1 \to x^k$ und $x^0 \to x^{k-1}$ durch, so folgt wegen

$$||\overline{x} - x^k|| \leq \frac{q}{1-q} ||x^k - x^{k-1}||$$

insgesamt die Abschätzung (i). Die Abschätzung (ii) sieht man wegen

$$||\overline{x} - x^{k+1}|| = ||g(\overline{x}) - g(x^k)|| \leq q||\overline{x} - x^k||.$$

Für die Eindeutigkeit des Fixpunktes betrachten wir einen weiteren Fixpunkt $\tilde{x} \in D$ mit $\tilde{x} = g(\tilde{x})$. Es gilt

$$||\overline{x} - \tilde{x}|| = ||g(\overline{x}) - g(\tilde{x})|| \leq q||\overline{x} - \tilde{x}||.$$

Wegen $q < 1$ folgt daraus schon $||\overline{x} - \tilde{x}|| = 0$, also $\overline{x} = \tilde{x}$. □

Für die Anwendung des Satzes ergeben sich einige Schwierigkeiten:

- Man muss eine kontrahierende Funktion $g : D \to \mathbb{R}^n$ finden.
- Man muss $D \subset \mathbb{R}^n$ mit $g(D) \subset D$ bestimmen.

Beispiel. Bestimme $\overline{x} \in \mathbb{R}$ mit $\overline{x} = g(\overline{x})$ und $g(x) = e^{-x}$. Dazu wähle man $D = [0.5, 0.69]$. Dann gilt $g(D) \subset D$. Man berechnet die Kontraktionszahl q zu

$$q = \max_{x\in D} |g'(x)| = \max_{x\in D} e^{-x} = e^{-0.5} \approx 0.606531 < 1.$$

Mit dem Startwert $x_0 = 0.55 \in D$ erhält man

k	x_k
0	0.55
1	0.57694981
11	0.56717695
12	0.56712420
20	0.56714309
24	0.56714327

Man benutze nun die a-priori Abschätzung $|\overline{x} - x_k| \leq \frac{q^k}{1-q}|x_1 - x_0|$. Man kann sich nun für die benötigte Zahl k der Iterationen interessieren, so dass $|\overline{x} - x_k| \leq \frac{q^k}{1-q}|x_1 - x_0| \leq \varepsilon$ für beispielsweise $\varepsilon = 10^{-6}$ gilt. Wir berechnen

$$q^k \leq \frac{\varepsilon(1-q)}{|x_1 - x_0|} \overset{\log(q)<0}{\Rightarrow} k \geq \frac{\log\left(\frac{\varepsilon(1-q)}{|x_1-x_0|}\right)}{\log(q)}.$$

Im Beispiel gilt $\varepsilon = 10^{-6}$ und $q = 0.606531$. Dann folgt $k \geq 22.3$, d.h. wir haben mit $k = 23$ eine leichte Überschätzung gegenüber der Tabelle. Die a-posteriori Abschätzung liefert uns

$$|\overline{x} - x_{12}| \leq \frac{q}{1-q}|x_{12} - x_{11}| \approx 8.3 \cdot 10^{-5}.$$

Beispiel. Es sei $g(x) = x^2$. Aus $g(\overline{x}) = \overline{x}$ folgt dann $\overline{x} = 0$ oder $\overline{x} = 1$. Für $\overline{x} = 1$ ist $g'(\overline{x}) = 2 > 1$ und somit der Fixpunktsatz nicht anwendbar. Die Folge $\{x_k\}$ divergiert für alle $x_0 > 1$. Als Idee betrachtet man hier die Umkehrabbildung. Dazu bestimmen wir $h(x)$ so, dass gilt

$$h(x) = g^{-1}(x) \Rightarrow h(g(x)) = id(x) = x.$$

Dann hat h die gleichen Fixpunkte wie g, d.h. es gilt $h(\overline{x}) = \overline{x}$. Insbesondere gilt nun

$$h'(x) = \frac{1}{g'(x)} \Rightarrow |h'(x)| < 1 \text{ falls } |g'(x)| > 1.$$

Im konkreten Fall setze man $h(x) = \sqrt{x}$ und erhält $h'(1) = \frac{1}{2}$.

(10.6) Satz. (Lokaler Konvergenzsatz) Sei $g : \mathbb{R}^n \to \mathbb{R}^n$ mit $g(\overline{x}) = \overline{x}$ für ein $\overline{x} \in \mathbb{R}^n$ gegeben. Ist g stetig differenzierbar in einer Umgebung von $\overline{x}$ und gilt $||g'(\overline{x})||_\infty < 1$, dann gibt es eine Umgebung $D \subset \mathbb{R}^n$ von $\overline{x}$, so dass das Iterationsverfahren

$$x^{k+1} = g(x^k), \quad k \geq 0,$$

für alle $x^0 \in D$ gegen $\overline{x}$ konvergiert.

Beweis. Es sei $D := B_r(\overline{x}) = \{x \in \mathbb{R}^n \mid ||x - \overline{x}||_\infty \leq r\}$ die (abgeschlossene) Kugel um $\overline{x}$ mit Radius $r > 0$. Wähle r so, dass $||g'(x)||_\infty \leq q < 1$ für alle $x \in B_r(\overline{x})$. Dies ist möglich, da g stetig differenzierbar in einer Umgebung von $\overline{x}$ ist. Wir zeigen nun $g(D) \subset D$. Für $x \in B_r(\overline{x})$ gilt

$$||g(x) - \overline{x}||_\infty = ||g(x) - g(\overline{x})||_\infty \leq q||x - \overline{x}||_\infty \leq qr < r.$$

Außerdem ist g kontrahierend auf D wegen $\sup_{x \in D} ||g'(x)||_\infty \leq q < 1$. Damit folgt nun die Behauptung aus dem Fixpunktsatz von Banach (10.4). □

Anwendung im Fall $n = 1$: Fixpunktverfahren ist ein Verfahren $p$ter Ordnung.

(10.7) Satz. Sei $g : \mathbb{R} \to \mathbb{R}$ eine C^p-Funktion mit $p \in \mathbb{N}_+$. Sei $\overline{x} \in \mathbb{R}$ ein Fixpunkt von g mit den Eigenschaften

(i) $|g'(\overline{x})| < 1$ für $p = 1$ und

(ii) $g^{(i)}(\overline{x}) = 0$ für $i = 1, \ldots, p-1$, falls $p \geq 2$.

Dann gibt es ein Intervall $D = [\overline{x} - r, \overline{x} + r]$, $r > 0$, so dass für alle $x_0 \in D$ die Iteration $x_{k+1} = g(x_k)$, $k \geq 0$, konvergent gegen $\overline{x}$ vom Grad p ist mit

$$\lim_{k \to \infty} \frac{|x_{k+1} - \overline{x}}{|x_k - \overline{x}|^p} = \frac{1}{p!}|g^{(p)}(\overline{x})|.$$

Beweis. Mit den Voraussetzungen (i) und (ii) folgt insbesondere $|g'(\overline{x})| < 1$ für alle p. Daher folgt die Konvergenz der Folge $\{x_k\}_{k \in \mathbb{N}}$ für alle $x_0 \in D = [\overline{x} - r, \overline{x} + r]$ mit $r > 0$ geeignet aus (10.6). Die TAYLOR-Entwicklung von $g(x_k)$ um $\overline{x}$ liefert

$$\begin{aligned} x_{k+1} &= g(x_k) = g(\overline{x}) + \sum_{i=1}^{p-1} \frac{1}{i!} g^{(i)}(\overline{x})(x_k - \overline{x})^i + \frac{1}{p!} g^{(p)}(\overline{x} + \alpha(x_k - \overline{x}))(x_k - \overline{x})^p \\ &= \overline{x} + \frac{1}{p!} g^{(p)}(\overline{x} + \alpha(x_k - \overline{x}))(x_k - \overline{x})^p \end{aligned}$$

für ein geeignetes $0 \leq \alpha \leq 1$. Es folgt

$$\lim_{k \to \infty} \frac{|x_{k+1} - \overline{x}|}{|x_k - \overline{x}|^p} = \lim_{k \to \infty} \frac{1}{p!}|g^{(p)}(\overline{x} + \alpha(x_k - \overline{x}))| = \frac{1}{p!}|g^{(p)}(\overline{x})|.$$

□

Anwendung: NEWTON-Verfahren. Sei $\overline{x} \in \mathbb{R}$ eine Nullstelle von $f : D \to \mathbb{R}$ mit $D \subset \mathbb{R}$. Man betrachte die Iteration

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} \text{ für } k \geq 0.$$

Unter der Voraussetzung, dass $\overline{x}$ eine einfache Nullstelle von f ist, d.h. es gilt $f'(\overline{x}) \neq 0$, erhalt man für die Funktion g,

$$g(x) := x - \frac{f(x)}{f'(x)}$$

die Ableitungen

$$g'(x) = \frac{f(x)f''(x)}{f'(x)^2} \text{ und } g''(x) = \frac{f''(x)}{f'(x)}.$$

Insbesondere gilt $g'(\overline{x}) = 0$ wegen $f(\overline{x}) = 0$. Wendet man Satz (10.7) an mit $p = 2$, so erhält man

$$\lim_{k\to\infty} \frac{|x_{k+1} - \overline{x}|}{|x_k - \overline{x}|^2} = \frac{1}{2}\left|\frac{f''(\overline{x})}{f'(\overline{x})}\right|.$$

§ 11 Das NEWTON-Verfahren im $\mathbb{R}^n$

Sei $D \subset \mathbb{R}^n$ und sei $f : D \to \mathbb{R}^n$ eine C^1-Funktion. Gesucht ist eine Nullstelle $\overline{x} \in D$ von f mit $f(\overline{x}) = 0 \in \mathbb{R}^n$. Wir erhalten die verallgemeinerte Fixpunktiteration

$$x^{k+1} = x^k - f'(x^k)^{-1} \cdot f(x^k), \quad k \geq 0, \quad x^0 \in D. \tag{11.1}$$

Hierbei ist

$$f'(x^k) = \left(\frac{\partial f_i}{\partial x_j}(x^k) \right)_{1 \leq i,j \leq n}$$

die JACOBI-Matrix.

(11.2) Satz. Sei $D \subset \mathbb{R}^n$ offen und konvex. Sei $f : D \to \mathbb{R}^n$ eine C^1-Funktion, so dass $f'(x)^{-1}$ existiert für alle $x \in D$. Für eine Konstante $c > 0$ gelte die (affin-variante) Lipschitz-Bedingung für $f'(x)$:

$$||f'(x)^{-1} \left(f'(x - tv) - f'(x)\right) v|| \leq ct||v||^2 \forall\, t \in [0,1], x \in D \text{ und } v \in \mathbb{R}^n \text{ mit } x + v \in D \tag{11.3}$$

Es gelte $f(\overline{x}) = 0$ für ein $\overline{x} \in D$. Wählt man nun einen Startwert $x^0 \in D$ mit $r := ||\overline{x} - x^0|| < \frac{2}{c}$, $B_r(\overline{x}) \subset D$, so gilt für die durch (11.1) definierte Folge $\{x_k\}$:

(i) $||\overline{x} - x^k|| \leq r$, $k \geq 0$, $\overline{x} = \lim_{k \to \infty} x^k$

(ii) $||\overline{x} - x^{k+1}|| \leq \frac{c}{2}||\overline{x} - x^k||^2$ für $k \geq 0$

(iii) $\overline{x}$ ist die eindeutig bestimmte Lösung von $f(\overline{x}) = 0$ in $B_{\frac{2}{c}}(\overline{x})$.

Beweis. Falls $f : D \to \mathbb{R}^n$ eine C^2-Funktion ist, gilt

$$||f'(x + tv) - f'(x)|| = ||f''(x + stv)tv|| \leq c_0 t||v||$$

mit $0 \leq s \leq 1$ und $c_0 > 0$ geeignet. Außerdem gilt

$$||f'(x)^{-1}|| \leq c_1$$

für $c_1 > 0$ geeignet. Damit folgt die Abschätzung 11.3.

Hilfsbehauptung: Es gilt

$$||f'(x)^{-1} \left(f(y) - f(x) - f'(x)(y - x)\right)|| \leq \frac{c}{2}||y - x||^2 \forall x, y \in D \tag{11.4}$$

Beweis: Setze $\varphi(t) := f(x + t(y - x))$ für $0 \leq t \leq 1$. Dann gilt

$$f(y) - f(x) = \varphi(1) - \varphi(0) = \int_0^1 \varphi'(t)dt = \int_0^1 f'(x + t(y - x))(y - x)dt.$$

Daher gilt

$$f(y) - f(x) - f'(x)(y - x) = \int_0^1 \left(f'(x + t(y - x)) - f'(x)\right)(y - x)dt.$$

Insgesamt erhalten wir

$$\begin{aligned} ||f'(x)^{-1} \left(f(y) - f(x) - f'(x)(y - x)\right)|| &= ||\int_0^1 f'(x)^{-1} \left(f'(x + t(y - x)) - f'(x)\right)(y - x)dt|| \\ &\overset{(11.3)}{\leq} \int_0^1 ct||y - x||^2 dt = \frac{c}{2}||y - x||^2, \end{aligned}$$

d.h. es gilt (11.4).

Damit haben wir eine Abschätzung wie in §9.2 für $n = 1$:

$$\begin{aligned} x^{k+1} - \overline{x} &= x^k - \overline{x} - f'(x^k)^{-1} f(x^k) = x^k - \overline{x} - f'(x^k)^{-1} \left(f(x^k) - f(\overline{x})\right) \\ &= -f'(x^k)^{-1} \left(f(x^k) - f(\overline{x}) - f'(x^k)(x^k - \overline{x})\right). \end{aligned}$$

Mit (11.4) folgt dann

$$||x^{k+1} - \overline{x}|| \leq \frac{c}{2}||x^k - \overline{x}||^2.$$

Falls $||\overline{x} - x^k|| \leq r$, so folgt mit $q := \frac{c}{2}r < 1$

$$||x^{k+1} - \overline{x}|| \leq \underbrace{\frac{c}{2}||x^k - \overline{x}||}_{\leq \frac{c}{2}r = q < 1} ||x^k - \overline{x}|| \leq q||x^k - \overline{x}||.$$

Wegen $r = ||\overline{x} - x^0||$ folgt insbesondere $||x^k - \overline{x}|| < r$ für $k = 1, 2, \ldots$. Damit gilt dann

$$\overline{x} = \lim_{k \to \infty} x^k.$$

Dies zeigt (i) und (ii).

Zur Eindeutigkeit: Sei $\tilde{x} \in B_{\frac{2}{c}}$ mit $f(\tilde{x}) = 0$ eine weitere Nullstelle. Zu zeigen ist nun $\tilde{x} = \overline{x}$. Die Abschätzung (11.4) ergibt

$$||\tilde{x} - \overline{x}|| = ||f'(\overline{x})^{-1} \left(\underbrace{f(\tilde{x})}_{=0} - \underbrace{f(\overline{x})}_{=0} \cdot (\tilde{x} - \overline{x}) \right) || \leq \frac{c}{2} \underbrace{||\tilde{x} - \overline{x}||}_{< \frac{2}{c}} ||\tilde{x} - \overline{x}|| < ||\tilde{x} - \overline{x}||$$

Dies ist offenbar ein Widerspruch, also muss schon $\tilde{x} = \overline{x}$ gelten. □

Praktische Durchführung des NEWTON-Verfahrens

Schreibe (11.1) als das LGS

$$f'(x^k)(x^{k+1} - x^k) = -f(x^k).$$

Löse das LGS

$$f'(x^k)d^k = -f(x^k), \text{ setze } x^{k+1} = x^k + d^k \tag{11.5}$$

Falls $f'(x^k)$ bekannt ist, kann man für $n = 2$ das Inverse leicht ausrechnen und (11.5) explizit lösen.

Beispiele:

(i) Sei $z = x + iy \in \mathbb{C}$. Die Gleichung $e^z - z = 0$ in $\mathbb{C}$ führt auf

$$e^x \cos(y) - x = 0,\ e^x \sin(y) - y = 0$$

Iteration: $(x_0, y_0) = (1, 1)$, $k = 5$, $(x_5, y_5) = (0.318131505204, 1.337235570143)$.

(ii) $3x_1 - \cos(x_1 x_2) - \frac{1}{2} = 0$, $x_1^2 - 81(x_2 + 0.1)^2 + \sin(x_3) + 1.06 = 0$, $e^{-x_1 x_2} + 20x_3 + \frac{10\pi - 3}{3} = 0$. Wir erhalten $\overline{x} = (0.5, 0.0, -0.5235877)^*$.

Varianten des NEWTON-Verfahrens

(i) Approximation von $f'(x)$ durch Differenzenquotienten.

$$\frac{\partial f_i}{\partial x_j}(x^k) \approx \frac{f_i(x^k + he_j) - f_i(x^k)}{h} \text{ mit Schrittweite } h > 0.$$

Eine bessere Strategie ist das quasi-NEWTON-Verfahren: Dazu approximiert man $f'(x^k)$ durch eine $n \times n$-Matrix B_k und macht einen (leicht zu berechnenden) Übergang $B_k \to B_{k+1}$.

(ii) Modifiziertes NEWTON-Verfahren; λ-Strategie.
Mit einem Konvergenzerzeugenden Faktor $0 < \lambda_k \leq 1$ betrachten wir

$$x^{k+1} = x^k + \lambda_k d^k \text{ mit } d^k := -f'(x^k)^{-1} f(x^k)$$

Die optimale Wahl für λ_k erhält man durch "Merrit-Funktionen".

§ 12 Iterationsverfahren für lineare Gleichungssysteme

Sei A eine reguläre $n \times n$-Matrix. Zu lösen sei das LGS $Ax = b$ für $b \in \mathbb{R}^n$. Die GAUSS-Elimination benötigt dazu $\frac{1}{3}n^3 + O(n^2)$ Operationen. Dies ist für sehr große n oder schwach besetzte Matrizen ungünstig.

Beispiel

Sei $f \in C[a,b]$. Gesucht ist eine Lösung $u \in C^2[a,b]$ der Randwertaufgabe (RWP)

$$u''(x) = f(x), u(a) = 0, u(b) = 0.$$

Betrachte dazu folgende Diskretisierung: Wähle $n \in \mathbb{N}^+$ und setze die Schrittweite $h := \frac{b-a}{n+1}$. Nun definiere

$$x_i = a + ih$$

für $i = 0, \ldots, n+1$. Dann gilt $x_0 = a$ und $x_{n+1} = b$.
Nun Approximiere

$$u'(x_i) \approx \frac{1}{h}\left(u(x_i) - u(x_{i-1})\right) \text{ für } i = 1, \ldots, n+1 \text{ sowie}$$

$$\begin{aligned} f(x_i) &= u''(x_i) \approx \frac{1}{h}\left(u'(x_{i+1}) - u'(x_i)\right) \\ &\approx \frac{1}{h}\left(\frac{1}{h}\left(u(x_{i+1}) - u(x_i)\right) - \frac{1}{h}\left(u(x_i) - u(x_{i-1})\right)\right) \\ &= \frac{1}{h^2}\left(u(x_{i+1}) - 2u(x_i) + u(x_{i-1})\right) \text{ für } i = 1, \ldots, n \end{aligned}$$

Approximiere nun weiter $u(x_i)$ durch $u_i \in \mathbb{R}$. Wegen $u(a) = u(x_0) = 0$ und $u(b) = u(x_{n+1}) = 0$ fordert man $u_0 = u_{n+1} = 0$ und betrachtet das LGS für $(u_1, \ldots, u_n)$:

$$\left\{ \begin{array}{ccccccc} u_2 & - & 2u_1 & & & = & h^2 f(x_1) \\ u_{i+1} & - & 2u_i & + & u_{i-1} & = & h^2 f(x_i) \quad \text{für } i = 2, \ldots, n-1 \\ & - & 2u_n & + & u_{n-1} & = & h^2 f(x_n) \end{array} \right\}$$

Wir erhalten in Matrixschreibweise

$$\underbrace{\begin{pmatrix} -2 & 1 & & & 0 \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ 0 & & & 1 & -2 \end{pmatrix}}_{=:A \text{(neg. def. nach §4)}} \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} = h^2 \begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix}$$

Beachte, dass A nicht das starke, sondern nur das schwache Zeilensummenkriterium erfüllt.

Idee für ein Iterationsverfahren

Betrachte eine Äquivalenzumformung des LGS $Ax = b$ in eine Fixpunktgleichung

$$x = Cx + d$$

mit C geeignete $n \times n$-Matrix und $d \in \mathbb{R}^n$. Wir erhalten die Fixpunktiteration

$$x^{(k+1)} = Cxx^{(k)} + d,\ k \geq 0,\ x^{(0)} \in \mathbb{R}^n \text{ beliebige} \tag{12.1}$$

Betrachten wir einige Spezialfälle. Dazu zerlegen wir zunächst A in $A = L + D + R$ (dies ist keine LR-Zerlegung!) mit

$$L = \begin{pmatrix} 0 & & & 0 \\ a_{21} & \ddots & & \\ \vdots & \ddots & \ddots & \\ a_{n1} & \ldots & a_{n,n-1} & 0 \end{pmatrix}, D = \begin{pmatrix} a_{11} & & 0 \\ & \ddots & \\ 0 & & a_{nn} \end{pmatrix}, R = \begin{pmatrix} 0 & a_{12} & \ldots & a_{1n} \\ & \ddots & \ddots & \vdots \\ & & \ddots & a_{n-1,n} \\ 0 & & & 0 \end{pmatrix}$$

(i) Gesamtschritt- oder JACOBI-Verfahren (GS-Verfahren).

$$Ax = (L + D + R)x = b$$

Iteration für $a_{ii} \neq 0$ für alle $i = 1, \ldots, n$:

$$Dx^{(k+1)} + (L + R)x^{(k)} = b \tag{12.2}$$

$$x^{(k+1)} = Cx^{(k)} + d,\ k \geq 0,\ C := -D^{-1}(L + R),\ d := D^{-1}b \tag{12.3}$$

Explizit lautet die Iteration (12.2)

$$a_{ii}x_i^{(k+1)} + \sum_{j \neq i} a_{ij}x_j^{(k)} = b_i,\ i = 1, \ldots, n \tag{12.4}$$

(ii) Einzelschritt- oder GAUSS-SEIDEL-Verfahren (ES-Verfahren).
Idee: Ersetze für $j < i$ die Werte $x_j^{(k)}$ in (12.4) durch die bereits berechneten Werte $x_j^{(k+1)}$.

$$\sum_{j<i} a_{ij}x_j^{(k+1)} + a_{ii}x_i^{k+1} + \sum_{j>i} a_{ij}x_j^{(k)} = b_i,\ i = 1, \ldots, n \tag{12.5}$$

In Matrixform erhält man

$$(L + D)x^{(k+1)} + Rx^{(k)} = b, \tag{12.6}$$

wobei gilt $x^{(k+1)} = Cx^{(k)} + d$ mit $C = -(L + D)^{-1}R$ und $d = (L + D)^{-1}b$.

(12.7) Definition (Konvergenz). Sei C eine $n \times n$-Matrix. Die Iteration

$$x^{(k+1)} = Cx^{(k)} + d,\ k \geq 0,\ x^{(0)}, d \in \mathbb{R}^n$$

heißt konvergent, wenn die Folge $\{x^{(k)}\}$ für alle $x^{(0)}$ und d konvergiert.

(12.8) Satz (Konvergenz). Gegeben sei die Iteration $x^{(k+1)} = Cx^{(k)} + d$ mit $k \geq 0$ und $x^{(0)} \in \mathbb{R}^n$.

(i) Hinreichend für die Konvergenz ist die Bedingung $||C|| < 1$ für eine geeignete Norm $||\cdot||$.

(ii) Die Iteration konvergiert genau dann, wenn der Spektralradius $\rho(C) < 1$ ist.

Beweis.

(i) Es gilt $x^{(k+1)} = g(x^{(k)})$ mit $g(x) = Cx + d$. Die Abbildung g ist kontrahierend, da $||g(y) - g(x)|| = ||C(y - x)|| \leq ||C|| \cdot ||y - x||$ für alle $x, y \in \mathbb{R}^n$. Sei $q := ||C|| < 1$, so folgt nach dem Fixpunktsatz (10.4) von BANACH mit $D = \mathbb{R}^n$ die Konvergenz für alle $x^{(0)} \in D = \mathbb{R}^n$.

(ii) Sei $\rho(C) < 1$. Nach Satz (5.4) gibt es eine Norm $||\cdot||$ mit $||C|| < 1$. Nach Teil (i) folgt die Konvergenz.
Sei nun die Iteration konvergent. Sei $Cx = \lambda x$ mit $|\lambda| = \rho(C)$. Setze nun

$$x^{(0)} = d = x.$$

Die Iteration ergibt

$$x^{(1)} = Cx^{(0)} + d = Cx + x = \lambda x + x = (\lambda + 1)x.$$

Man verifiziert leicht, dass gilt

$$x^{(k+1)} = \left(\lambda^k + \lambda^{k-1} + \ldots + \lambda + 1\right) x.$$

Aus der Konvergenz von $\{x^{(k)}\}$ folgt $|\lambda| = \rho(C) < 1$. □

Anwendung auf das GS-Verfahren

(12.9) Satz.

(i) Das GS-Verfahren (12.3) konvergiert für alle diagonaldominanten Matrizen A.

(ii) Das GS-Verfahren (12.3) konvergiert für alle Matrizen A mit $\sum_{i\neq k}|a_{ik}| < |a_{kk}|$ für alle $k = 1,\dots,n$. (starkes Spaltensummenkriterium)

Beweis.

(i) Es gilt $C = -D^{-1}(L+R)$. Für $\sum_{k\neq i}|a_{ik}| < a_{ii}$ für alle $i = 1,\dots,n$ folgt $||C||_\infty < 1$ und damit die Konvergenz nach (12.8(i)).

(ii) Wende (i) auf A^* an. □

Im GS-Verfahren gilt $A = L + D + R$. Mit $C_G = -D^{-1}(L+R)$ und $d = D^{-1}b$ wird daraus die Vorschrift

$$x^{(k+1)} = C_G x^{(k)} + d.$$

Es gilt weiterhin

$$||C_G||_\infty = \max_{i=1,\dots,n}\frac{1}{a_{ii}}\sum_{k\neq i}|a_{ik}|,\ ||C_G||_\infty < 1 \Leftrightarrow \sum_{k\neq i}|a_{ik}| < |a_{ii}|\,\forall i = 1,\dots,n.$$

Als Folgerung erhalten wir die Abschätzung $||\overline{x} - x^{(k)}||_\infty \leq \frac{q^k}{1-q}||x^1 - x^0||_\infty$ mit $q := ||C_G||_\infty$.

(12.10) Definition. Eine Matrix $A = (a_{ik})$ heißt zerlegbar, wenn es nichtleere Teilmengen $N_1, N_2 \subset \{1,\dots,n\}$ gibt mit

(i) $N_1 \cup N_2 = \{1,\dots,n\}$ und $N_1 \cap N_2 = \emptyset$

(ii) Für $i \in N_1$ und $k \in N_2$ gilt $a_{ik} = 0$.

Das bedeutet: Ist $N_1 = \{1,\dots,q\}$, so gibt es eine Permutationsmatrix P mit

$$P^*AP = \begin{pmatrix} \tilde{A}_{11} & 0 \\ \tilde{A}_{21} & A_{22} \end{pmatrix},$$

wobei $\tilde{A}_{11}$ q Spalten enthält. A heißt unzerlegbar, wenn A nicht zerlegbar ist.

(12.11) Satz (Schwaches Zeilensummenkriterium). Sei A unzerlegbar und es gelte

$$\sum_{k\neq i}|a_{ik}| \leq |a_{ii}| \text{ für } i = 1,\dots,n$$

sowie

$$\sum_{k\neq i_0}|a_{ik}| < |a_{i_0,i_0}| \text{ für mindestens ein } i_0 \in \{1,\dots,n\}.$$

Dann konvergiert das GS-Verfahren.

Beweis. Zu zeigen ist $\rho(C_G) < 1$. Wegen $||C_G||_\infty \leq 1$ gilt $\rho(C_G) \leq 1$.
Annahme: $\rho(C_G) = 1 = |\lambda|$ für $\lambda \in \mathbb{C}$ ist EW. Sei $x \in \mathbb{C}^n$ mit $C_G x = \lambda x$, $||x||_\infty = 1$. Setze oBdA. $N_1 = \{i \in \{1,\dots,n\} \mid |x_i| = 1\}$ und $N_2 = \{1,\dots,n\} \setminus N_1$.
Es gilt $(C_G x)_i = \sum_{k\neq i} c_{ik}x_k$ wegen $C_G = (c_{ik})$ mit $c_{ii} = 0$ für $i = 1,\dots,n$. Es folgt

$$\sum_{k\neq i} c_{ik}x_k = \lambda x_i.$$

Wegen $|\lambda| = 1$ erhalten wir

$$|x_i| = |\lambda x_i| \leq \sum_{k\neq i}|c_{ik}|\cdot|x_k| = \sum_{k\neq i}\frac{|a_{ik}|}{|a_{ii}|}\underbrace{|x_k|}_{\leq 1} \leq \sum_{k\neq i}\frac{|a_{ik}|}{|a_{ii}|} \leq 1.$$

Für alle $i \in N_1$ gilt $|x_i| = 1$ und daher mit der Abschätzung $\sum_{k\neq i}|a_{ik}| = |a_{ii}|$. Wegen $\sum_{k\neq i_0}|a_{i_0,K}| < |a_{i_0,i_0}|$ für ein i_0 folgt $i_0 \in N_2$, d.h. $N_2 \neq \emptyset$. Da A unzerlegbar ist, gibt es $i_1 \in N_1$ und $k_1 \in N_2$ mit $a_{i_1,k_1} \neq 0$. Dies ergibt einen Widerspruch wegen

$$1 = |x_{i_1}| \leq \sum_{k\neq i_1}\frac{|a_{i_1,k}|}{|a_{i_1,i_1}|}|x_k| \overset{(*)}{<} \sum_{k\neq i_1}\frac{|a_{i_1},k}{|a_{i_1,i_1}|} \leq 1.$$

Die Abschätzung $(*)$ gilt wegen $|x_{k_1}| < 1$ und $|a_{i_1,k_1} > 0$. □

(12.12) Satz. Falls $||C_G|| \leq 1$, so folgt $||C_E||_\infty \leq ||C_G|| \leq 1$.
Beweis. Als Übungsaufgabe.

Beispiel.

Es sei $A = \begin{pmatrix} 1 & 0.1 \\ 6 & 1 \end{pmatrix}$. Dann gilt $C_G = \begin{pmatrix} 0 & -0.1 \\ -6 & 0 \end{pmatrix}$ und $C_E = \begin{pmatrix} 0 & -0.1 \\ 0 & 0.6 \end{pmatrix}$. Wir erhalten

$$||C_G||_\infty = 6 \geq 1,\ ||C_E||_\infty = 0.6 < 1.$$

Es gilt aber

$$\rho(C_G) = \sqrt{0.6} < 1\, \rho(C_E) = 0.6 < 1,$$

d.h. GS- und ES-Verfahren konvergieren beide.

Konvergenzverbesserung durch Relaxationsverfahren: Multipliziere (12.5) mit einem Parameter $\omega > 0$:

$$\omega \sum_{j<i} a_{ij} x_j^{(k+1)} + \underbrace{\omega a_{ii} x_i^{(k+1)}}_{(*)} + \omega \sum_{j>i} a_{ij} x_j^{(k)} = \omega b_i \,\forall i = 1, \ldots, n.$$

Ersetze $(*)$ durch eine "Mittelung" $a_{ii}\left(x_i^{(k+1)} - x_i^{(k)}\right) + \omega a_{ii} x_i^{(k)}$. Dies ergibt das Verfahren

$$a_{ii} x_i^{(k+1)} = a_{ii} x_i^{(k)} (1-\omega) + \omega \left(-\sum_{j<i} a_{ij} x_j^{(k+1)} - \sum_{j>i} a_{ij} x_j^{(k)} + b_i \right) \forall i = 1, \ldots, n \tag{12.13}$$

Programm (Pseudocode) für N Schritte:

```
für k=1,...,N :

    für i=1,...,n :

      x_i  :=  x_i(1-ω)  +  ω  ( b_i  -  Σ_{j ≠ i}  a_ij x_j )  /  a_ii

    end

end
```

Relaxationsverfahren in Matrixform

$$x^{(k+1)} = C(\omega) x^{(k)} + d(\omega),\ C(\omega) := -(\omega L + D)^{-1}\left((\omega - 1)D + \omega R\right)\ d(\omega) = \omega(\omega L + D)^{-1} b \tag{12.14}$$

Für $\omega = 1$ erhält man das ES-Verfahren. Für $\omega < 1$ nennt man es Unter- und für $\omega > 1$ Überrelaxation (auch SOR-Verfahren, Successive Overrelaxation).

(12.15) Satz. Für beliebige Matrizen A gilt $\rho(C(\omega)) \geq |\omega - 1|$ für alle ω. Für die Konvergenz muss $\rho(C(\omega)) < 1$ gelten $\Rightarrow |\omega - 1| < 1 \Rightarrow 0 < \omega < 2$.

(12.16) Satz. Für positiv definite Matrizen A gilt $\rho(C(\omega)) < 1$ für alle $0 < \omega < 2$.

Beweis. Zu zeigen ist $\rho(C(\omega)) < 1$. Sei $C(\omega)x = \lambda x$ mit $|\lambda| = \rho(C(\omega))$. Nach Definition gilt

$$\left((1-\omega)D - \omega R\right) x = \lambda (D + \omega L) x.$$

Bildet man auf beiden Seiten das Skalarprodukt mit x, so erhält man aus der Linearität des Skalarprodukten

$$(1-\omega)\langle Dx, x\rangle - \omega\langle Rx, x\rangle = \lambda(\langle Dx, x\rangle + \omega\langle Lx, x\rangle)$$

Da A positiv definit und insbesondere symmetrisch ist, gilt $R = L^*$. Wir erhalten

$$0 < \langle Dx, x\rangle,\ 0 < \langle Ax, x\rangle = \langle Dx, x\rangle + 2\langle Lx, x\rangle.$$

Wir definieren $q := \frac{\langle Lx,x\rangle}{\langle Dx,x\rangle} > -\frac{1}{2}$. Aus obiger Gleichung folgt dann

$$\lambda = \frac{1-\omega-\omega q}{1+\omega q}.$$

Wegen $0 < \omega < 2$ und $q > -\frac{1}{2}$ gilt

$$1+\omega q > 0,\ -1-q < q,\ -1 < 1-\omega.$$

Es folgt

$$-1-\omega q < 1-\omega-\omega q = 1+\omega(-1-q) < 1+\omega q,$$

also $|1-\omega-\omega q| < 1+\omega$. Also gilt $\rho(C(\omega)) = |\lambda| < 1$. □

Den optimalen Wert des Parameters ω^* bestimmt man durch $\rho(C(\omega^*)) = \min_{0<\omega<2} \rho(C(\omega))$. Qualitativ gilt für "konsistent geordnete" Matrizen A: $1 < \omega^* < 2$.

IV Interpolation

Einführung

Gegeben seien

(i) $n+1$ Paare reeller oder komplexer Zahlen

$$(x_j, f_j), j = 0, \ldots, n,$$

die sog. Stützstellen. Die x_j heißen Knoten: Falls $x_j \in \mathbb{R}$, so hat man die Anordnung

$$x_0 < x_1 < \ldots < x_n.$$

Zum Beispiel sind $f_j = f(x_j)$ die Messdaten einer unbekannten Funktion $f : \mathbb{R} \to \mathbb{R}$.

(ii) Eine durch $n+1$ Parameter $a_0, \ldots, a_n$ bestimmte Familie von Funktionen

$$\Phi(x; a_0, \ldots, a_n).$$

Gesucht sind Parameter $a_0, \ldots, a_n$ mit $\Phi(x_j; a_0, \ldots, a_n) = f_j$ für $j = 1, \ldots, n$. Dann ist $\Phi(x; a_0, \ldots, a_n)$ ein Näherungswert für $f(x)$ im Fall $x \neq x_j$.

Spezialfall: Lineare Interpolationsprobleme.

$\Phi(x; a_0, \ldots, a_n)$ hängt linear von $a_0, \ldots, a_n$ ab:

$$\Phi(x; a_0, \ldots, a_n) = a_0\Phi_0(x) + \ldots + a_n\Phi_n(x)$$

mit Basisfunktionen $\Phi_j(x)$ für $j = 0, \ldots, n$.

Beispiele.

(i) Interpolation durch Polynome: $\Phi_j(x) = x^j$, $\Phi(x; a_0, \ldots, a_n) = a_0 + a_1x + \ldots + a_nx^n$.

(ii) Trigonometrische Interpolation: Es sei

$$\Phi(x; a_0, \ldots, a_n) = a_0 + a_1e^{xi} + \ldots + a_{n-1}e^{(n-1)xi} = a_0 + a_1\omega + \ldots + a_{n-1}\omega^{n-1}$$

mit $\omega = e^{xi} = \cos(x) + i\sin(x)$, $i^2 = -1$. Die Basisfunktionen lauten dann $\Phi_k(x) = e^{kxi} = \omega^k$. Als Knoten wählt man $\omega_k = e^{x_k i}$, $x_k = \frac{2\pi k}{n}$ für $k = 0, \ldots, n-1$.

(Vergleiche komplexe Einheitswurzeln, etwa Forster I)

(iii) Spline-Interpolation: Betrachte z.B. kubische Splines mit

(a) $\Phi(\cdot; a_0, \ldots, a_n) \in C^2[x_0, x_n]$, $x_j \in \mathbb{R}$

(b) $\Phi(\cdot; a_0, \ldots, a_n)$ ist ein kubisches Polynom auf $[x_j, x_{j+1}]$ für $j = 0, \ldots, n-1$.

Nichtlineare Interpolationsprobleme.

(i) Interpolation durch rationale Funktionen:

$$\Phi(x; a_0, \ldots, a_n, b_1, \ldots, b_m) = \frac{a_0 + a_1x + \ldots + a_nx^n}{b_0 + b_1x + \ldots + b_mx^m}.$$

(ii) Interpolation durch Exponentialsummen;

$$\Phi(x; a_0, \ldots, a_n, \lambda_0, \ldots, \lambda_n) = a_0e^{\lambda_0 x} + \ldots + a_ne^{\lambda_n x}.$$

§ 13 Interpolation durch Polynome

Bezeichne Π_n die Menge aller reellen oder komplexen Polynome vom Grad $\leq n$, d.h.

$$\Pi_n := \{a_0 + a_1 x + \ldots + a_n x^n \mid a_0, \ldots, a_n, x \in \mathbb{R} \text{ oder } \mathbb{C}\}.$$

(13.1) Satz. Zu beliebigen $n+1$ Stützstellen (x_j, f_j) mit $j = 0, \ldots, n$ und $x_j \neq x_i \, \forall j \neq i$ gibt es genau ein Polynom $p \in \Pi_n$ mit $p(x_j) = f_j$ für $j = 0, \ldots, n$.

Beweis. Die Bedingungen $p(x_j) = f_j$ für $j = 0, \ldots, n$ ergeben das LGS

$$\sum_{k=0}^{n} a_k x_j^k = f_j, \; j = 0, \ldots, n.$$

Dies lässt sich schreiben in der Form

$$\underbrace{\begin{pmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^n \\ 1 & x_1 & x_1^2 & \ldots & x_1^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^n \end{pmatrix}}_{\text{VANDERMONDE-Matrix } X} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} f_0 \\ f_1 \\ \vdots \\ f_n \end{pmatrix}.$$

Es gilt $\det(X) = \det(x_j^k) = \prod_{j,k=0, j\neq k}^{n} (x_j - x_k) \neq 0$, da $x_j \neq x_k$ für $j \neq k$. □

Bemerkung: Die Berechnung der $a_0, \ldots, a_n$ mittels der CRAMERschen Regel ist numerisch ungünstig.

13.1 Die Interpolationsformel von LAGRANGE

Für das Polynom

$$L_i(x) := \prod_{k=0, k\neq i}^{n} \frac{x - x_k}{x_i - x_k}$$

gilt $L_i \in \Pi_n$ und $L_i(x_j) = \delta_{ij}$. Das interpolierende Polynom ist gegeben durch

$$p(x) = \sum_{i=0}^{n} L_i(x) f_i = \sum_{i=0}^{n} f_i \prod_{k=0, k\neq i}^{n} \frac{x - x_k}{x_i - x_k} \tag{13.2}$$

Dies sieht man, da $p(x_j) = \sum_{i=0}^{n} L_i(x_j) f_i = \sum_{i=0}^{n} \delta_{ij} f_i = f_j$. Die Behauptung folgt dann aus der Eindeutigkeit nach Satz (13.1).

Beispiel

j	x_j	f_j
0	0	1
1	1	3
2	3	2

Wir berechnen

$$\begin{aligned} L_0(x) &= \frac{(x-1)(x-3)}{(0-1)(0-3)} = \frac{1}{3}(x^2 - 4x + 3) \\ L_1(x) &= \frac{(x-0)(x-3)}{(1-0)(1-3)} = -\frac{1}{2}(x^2 - 3x) \\ L_2(x) &= \frac{(x-0)(x-1)}{(3-0)(3-1)} = \frac{1}{6}(x^2 - x) \end{aligned}$$

Das interpolierende Polynom berechnet sich dann zu

$$p(x) = 1 \cdot L_0(x) + 3 \cdot L_2(x) + 2 \cdot L_3(x) = -\frac{5}{6}x^2 + \frac{17}{6}x + 1.$$

Nachteile der LAGRANGE-Interpolation: Für das Hinzufügen eines weitern Knotens x_j ist deine komplette Neuberechnung erforderlich.

13.2 Der Algorithmus von Aitken und Neville.

Dies ist ein numerisch sparsamer Algorithmus zur Berechnung von $p(x) \in \Pi_n$ an einigen wenigen Stellen x.
Für Indizes $i_0, \dots, i_k \in \{0, \dots, n\}$ sei $p_{i_0,\dots,i_k} \in \Pi_k$ das Interpolationspolynom zu den Knoten $x_{i_0}, \dots, x_{i_k}$. Insbesondere gilt $p_i(x) \equiv f_i$ und $p_{0,\dots,n}(x) = p(x)$. Wir erhalten die folgende Rekursionsformal von Aitken:

$$p_{i_0,\dots,i_k}(x) = \frac{(x - x_{i_0})p_{i_1,\dots,i_k}(x) - (x - x_{i_k})p_{i_0,\dots,i_{k-1}}(x)}{x_{i_k} - x_{i_0}} \tag{13.3}$$

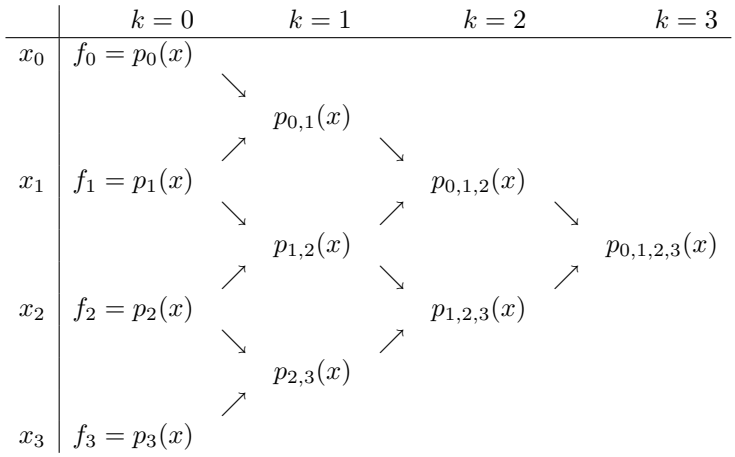
Beweis. Das Polynom $Q(x) \in \Pi_k$ auf der rechten Seite von (13.3) erfüllt

$$\begin{aligned} Q(x_{i_0}) &= p_{i_0,\dots,i_{k-1}}(x_{i_0}) = f_{i_0}, \\ Q(x_{i_k}) &= p_{i_1,\dots,i_k}(x_{i_k}) = f_{i_k}, \\ Q(x_{i_j}) &= f_{i_j} \text{ für } j = 1, \dots, k-1 \end{aligned}$$

Damit folgt (13.3) wegen der Eindeutigkeit der Interpolation (13.1).

Variante von Neville.

Sei x fest gewählt. Man erzeuge die Werte $p_{i-k,\dots,i}(x)$ zu den Werten $x_{i-k}, \dots, x_i$ nach dem Schema



Neue Berechnung (x sei fest):

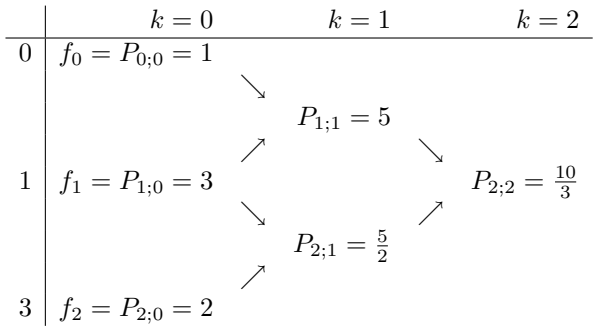
$$P_{i;k} := p_{i-k,\dots,i}(x).$$

Dies führt zu der Rekursion

$$\begin{aligned} P_{i;0} &= f_i \text{ für } i = 0, \dots, n \text{(Dies entspricht der Spalte } k = 0) \\ P_{i;k} &= \frac{(x - x_{i-k})P_{i;k-1} - (x - x_i)P_{i-1;k-1}}{x_i - x_{i-k}} \text{ für } k = 1, \dots, n \text{ und jeweils } i = 1, 2, \dots \end{aligned} \tag{13.4}$$

Es gilt dann insbesondere $P_{n;n} = p(x)$.

Beispiel.



Diese Variante von Neville wird speziell bei Extrapolationsalgorithmen benutzt.

13.3 Die NEWTONsche Interpolation, Dividierte Differenzen

Ziel: Man berechne die Koeffizienten des Interpolationspolynoms $p(x) = p_{0,1,\ldots,n}(x)$.

$$p(x) = a_0 + a_1(x - x_0) + \ldots + a_n(x - x_0)(x - x_1) \cdot \ldots \cdot (x - x_{n-1}) \tag{13.5}$$

Die Berechnung erfolgt nach dem HORNERschen Schema:

$$p(x) = (\ldots(a_n(x - x_{n-1}) + a_{n-1})(x - x_{n-2}) + \ldots + a_1)(x - x_0) + a_0.$$

Man erhält folgende Rückwärtsrekursion:

$$\begin{aligned} b_n &:= a_n \\ b_i &:= b_{i+1}(x - x_i) + a_i \text{ für } i = n-1, \ldots, 0 \end{aligned}$$

Dann ist $b_0 = p(x)$.

Wir definieren Abschnittspolynome $Q_k(x) := \sum_{i=0}^{k} a_i \prod_{j=0}^{i-1}(x - x_j)$. Es gilt

(i) $Q_k(x) = p_{0,1,\ldots,k}(x)$

(ii) Q_k ist der Koeffizient von x^k in $p_{0,1,\ldots,k}(x)$.

Die Koeffizienten a_i werden mit dem Differenzschema berechnet:

$$\begin{array}{c|ccccc} x_0 & [f_0] = f_0 & & & & \\ & & \searrow & & & \\ & & & [f_0, f_1] & & \\ & & \nearrow & & \searrow & \\ x_1 & [f_1] = f_1 & & & & [f_0, f_1, f_2] \\ & & \searrow & & \nearrow & \\ & & & [f_1, f_2] & & \\ & & \nearrow & & & \\ x_2 & [f_2] = f_2 & & & & \end{array}$$

Dafür gilt folgende Rekursion:

$$\begin{aligned} [f_i] &= f_i \text{ für } i = 0, \ldots, n \\ [f_i, \ldots, f_k] &= \frac{[f_{i+1}, \ldots, f_k] - [f_i, \ldots, f_{k-1}]}{x_k - x_i} \end{aligned}$$

(13.6) Satz. Es gilt $p_{0,\ldots,k} = \sum_{i=0}^{k} [f_0, \ldots, f_i] \prod_{j=0}^{i-1}(x - x_j)$, also insbesondere $a_i = [f_0, \ldots, f_i]$.

Beweis. Durch Induktion nach k. Für $k = 0$ ist die Aussage klar, es gilt $a_0 = [f_0]$. Die Aussage gelte daher für ein festes, aber beliebiges $k - 1 \geq 0$. Wir erhalten dann eine Darstellung

$$p_{0,\ldots,k}(x) = p_{0,\ldots,k-1}(x) + a(x - x_0) \cdot \ldots \cdot (x - x_{k-1}).$$

Zu zeigen ist nun $a = [f_0, \ldots, f_k]$. Es gilt

Der Koeffizient von x^k in $p_{0,\ldots,k}(x)$ lautet a.

Der Koeffizient von x^{k-1} in $p_{0,\ldots,k-1}(x)$ lautet $[f_0, \ldots, f_{k-1}]$ nach I.V.

Der Koeffizient von x^{k-1} in $p_{1,\ldots,k}(x)$ lautet $[f_1, \ldots, f_k]$ nach I.V.

Nach AITKEN gilt

$$p_{0,\ldots,k}(x) = \frac{(x - x_0)p_{1,\ldots,k}(x) - (x - x_k)p_{0,\ldots,k-1}(x)}{x_k - x_0}$$

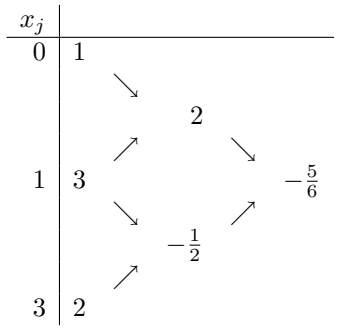
Der Koeffizient von x^k auf der rechten Seite ist

$$a = \frac{[f_1, \ldots, f_k] - [f_0, \ldots, f_{k-1}]}{x_k - x_0} \overset{Def.}{=} [f_0, \ldots, f_k]$$

□

Beispiel

Wir betrachten die Daten aus einem vorherigen Beispiel. Es gilt



Damit erhalten wir

$$p(x) = 1 + 2(x-0) - \frac{5}{6}(x-0)(x-1) = -\frac{5}{6}x^2 + \frac{17}{6}x + 1.$$

(13.7) Lemma. Für eine beliebige Permutation $i_0, \ldots, i_n$ von $0, \ldots, n$ gilt $[f_{i_0}, \ldots, f_{i_n}] = [f_0, \ldots, f_n]$.

Beweis. Übung.

13.4 Der Interpolationsfehler, Konvergenzfragen

Sei f eine $C^{n+1}[a,b]$-Funktion und seien $x_0, \ldots, x_n \in [a,b]$. Betrachte den Fehler $f(x) - p(x)$ für $x \in [a,b]$, wobei $p(x)$ das Interpolationspolynom zu den Stützstellen (x_j, f_j), $j = 0, \ldots, n$, $f_j = f(x_j)$ ist.

(13.8) Satz. Sei $f \in C^{n+1}[a,b]$ und seien $\overline{x}, x_j \in [a,b]$ für $j = 0, \ldots, n$. Dann gibt es ein $\xi \in [a,b]$ mit

$$f(\overline{x}) - p(\overline{x}) = L(\overline{x}) \frac{f^{(n+1)}(\xi)}{(n+1)!}, \quad L(x) = \prod_{j=0}^{n} (x - x_j).$$

Beweis. Betrachte für $\overline{x} \neq x_j$, $j = 0, \ldots, n$ die Funktion

$$F(x) := f(x) - p(x) - \frac{f(\overline{x}) - p(\overline{x})}{L(\overline{x})} L(x)$$

für $x \in [a,b]$. Dann ist $F(x) \in C^{n+1}[a,b]$. $F(x)$ hat die $n+2$ Nullstellen $\overline{x}, x_0, \ldots, x_n$ in $[a,b]$. Nach dem Satz von Rolle hat $F^{(n+1)}$ mindestens eine Nullstelle $\xi \in [a,b]$. Es gilt also

$$0 = F^{(n+1)}(\xi) = f^{(n+1)}(\xi) - \frac{f(\overline{x}) - p(\overline{x})}{L(\overline{x})} (n+1)!.$$

Damit folgt die Behauptung. □

Beispiele

(i) $f(x) = e^x$, $x_0 = 0$, $x_1 = 1$, $n = 1$. Die Approximation ist dann einer Gerade. Es gilt $f''(x) = e^x$ sowie $p(x) = (e-1)x + 1$ mit $p(0) = 1 = f(0)$ und $p(1) = e = f(1)$. Es gilt nun $L(x) = x(x-1)$ und wir erhalten nach (13.8) die Abschätzung

$$\max_{0 \le x \le 1} |e^x - p(x)| \le \underbrace{\left(\max_{0 \le x \le 1} |x(x-1)| \right)}_{= \frac{1}{4} \text{ für } x = \frac{1}{2}} \cdot \underbrace{\left(\max_{0 \le x \le 1} \frac{e^x}{2} \right)}_{= \frac{e}{2}} = \frac{e}{8} \approx 0.34.$$

Für den exakten Fehler definieren wir zunächst

$$d(x) := p(x) - e^x = (e-1)x + 1 - e^x \Rightarrow d'(x) = e - 1 - e^x.$$

Es folgt dann $x_{max} = \ln(e-1)$ und wir erhalten

$$d(x_{max}) = (e-1)\ln(e-1) + 1 - e^{x_{max}} \approx 0.211867.$$

(ii) Allgemeine Abschätzung von $\max_{a\le x\le b}|L(x)|$: Sei $h := \max_{j=0,\dots,n-1}|x_{j+1}-x_j|$. Dann gilt

$$\max_{x_0\le x\le x_n}|(x-x_0)\cdot\ldots\cdot(x-x_n)| \le \frac{n!}{4}h^{n+1}.$$

Für $n=1$ folgt

$$|(x-x_0)(x-x_1)| \le \frac{1}{4}(x_1-x_0)^2 \text{ mit } x_{max} = \frac{x_0+x_1}{2}.$$

(iii) $f(x)=\sin(x)$, $x_j = j\frac{\pi}{n}$ für $j=0,\dots,n$. Es gilt $h=\max|x_{j+1}-x_j| = \frac{\pi}{n}$ und nach (ii)

$$|\sin(x)-p(x)| \le \underbrace{\left(\max_{x\in[0,\pi]}|(x-x_0)\cdot\ldots\cdot(x-x_n)|\right)}_{\le \frac{n!}{4}\left(\frac{\pi}{n}\right)^{n+1}} \underbrace{\max_{x\in[0,\pi]}\frac{f^{(n+1)}(x)}{(n+1)!}}_{=\frac{1}{(n+1)!}} \le \frac{1}{4(n+1)}\left(\frac{\pi}{n}\right)^{n+1}.$$

Für $n=4$ ist dies $\le \frac{1}{4\cdot 5}\left(\frac{\pi}{4}\right)^5 \approx 0.015$, für $n=8$ ist es $\le \frac{1}{4\cdot 9}\left(\frac{\pi}{8}\right)^9 \approx 6.2\cdot 10^{-6}$.

Für $g\in C[a,b]$ bezeichne $||g||_\infty := \max_{a\le x\le b}|g(x)|$. Als Fehlerabschätzung erhalten wir

$$||f-p||_\infty \le ||L||_\infty \cdot ||f^{(n+1)}||_\infty \cdot (n+1)!^{-1}.$$

Es stellt sich die Frage, für welche Knoten $x_0,\dots,x_n\in[a,b]$ der Wert $||L||_\infty$ minimal wird. Dies ist der Fall für die sogenannten Chebychev-Knoten

$$x_j = \frac{1}{2}(a+b+(b-a)\cos\left(\frac{2j+1}{2n+2}\pi\right), \; j=0,\dots,n.$$

Für die Konvergenz betrachte man eine Folge von Intervallteilungen für $N\in\mathbb{N}$:

$$\Delta_N = \{a = x_0^{(N)} < x_1^{(N)} < \ldots < x_{n(N)}^{(N)} = b\}$$

mit

$$||\Delta_N|| := \max_i |x_{i+1}^{(N)} - x_i^{(N)}|.$$

Es sei nun $p_{\Delta_N}(x)$ das interpolierende Polynom von f bezüglich Δ_N, d.h. $p_{\Delta_N}(x_i) = f(x_i)$ für $i=0,\dots,n(N)$. Es stellt sich die Frage, ob

$$\lim_{N\to\infty} p_{\Delta_N}(x) = f(x)\,\forall x\in[a,b]$$

gilt, falls $\lim_{N\to\infty}||\Delta_N|| = 0$ gilt. Dies ist jedoch im Allgemeinen nicht richtig! Als Gegenbeispiel betrachten wir das

Beispiel von Runge.

Es sei $f(x) := \frac{1}{1+25x^2}$ im Bereich $-1\le x\le 1$. Interpolieren wir f in $x_j = -1+j\cdot\frac{2}{n}$ durch $p_n(x)$, so erhalten wir

n	1	5	13	19
$\|\|f-p_n\|\|_\infty$	0.96	0.43	1.07	8.57

Wählt man statt dessen die Chebychev-Knoten, so erhält man

n	1	5	13	19
$\|\|f-p_n\|\|_\infty$	0.93	0.56	0.12	0.04

Trotzdem gilt <u>nicht</u> $\lim_{N\to\infty}||f-p_{\Delta_N}||_\infty = 0$.

Ohne Beweis nun die folgenden Sätze:

(13.9) Satz. Zu jeder Folge von Intervallteilungen $\{\Delta_N\}_{N\in\mathbb{N}}$ mit $\lim_{N\to\infty}||\Delta_N||_\infty = 0$ gibt es $f\in C[a,b]$, so dass $\{p_{\Delta_N}(\cdot)\}$ nicht gleichmäßig gegen f konvergiert.

(13.10) Satz. Sei f eine ganze Funktion auf $\mathbb{C}$, d.h. holomorph auf $\mathbb{C}$, dann gilt $\lim_{N\to\infty}||f-p_{\Delta_N}||_\infty = 0$ für alle $\{\Delta_N\}$ mit $\lim_{N\to\infty}||\Delta_N||_\infty = 0$.

§ 14 Spline-Interpolation

14.1 Polynom-Splines

Sei $\Delta := \{a = x_0 < x_1 < \ldots < x_n = b\}$ eine Zerlegung von $[a,b]$. Man nennt dann $x_1, \ldots, x_{n-1}$ innere Knoten und $x_0 = a, x_n = b$ Randknoten.

(14.1) Definition. Eine Funktion $s : [a,b] \to \mathbb{R}$ heißt Polynom-Spline vom Grad l ($l = 0, \ldots$) zur Zerlegung Δ, wenn folgende Eigenschaften erfüllt sind:

(i) $s(\cdot) \in C^{l-1}[a,b]$

(ii) s ist ein Polynom vom Grad l auf $x_j \le x \le x_{j+1}$ für $j = 0, \ldots, n-1$

Für $l = 0$ ist $C^{-1}[a,b]$ der Raum der auf $[a,b]$ stückweise stetigen Funktionen.

Man bezeichnet mit $S_l(\Delta)$ die Menge aller Polynom-Splines vom Grad l zur Zerlegung Δ.

Beispiele

(i) Lineare Splines. Gegeben sind $n+1$ Punkte $(x_0, f_0), \ldots, (x_n, f_n)$. Dann legt man einen Polygonzug durch die Punkte (x_i, f_i), $i = 0, \ldots, n$.

(ii) Quadratische Splines. Man hat äquidistante Knoten $x_j = a + jh$ mit $j = 0, \ldots, n$ und $h = \frac{b-a}{n}$. Dann definiert man einen quadratischen Spline durch

$$s(x) = \frac{1}{2h^2} \cdot \begin{cases} (x - x_j)^2, & x_j \le x \le x_{j+1} \\ h^2 + 2h(x - x_{j+1}) - 2(x - x_{j+1})^2, & x_{j+1} \le x \le x_{j+2} \\ (x_{j+3} - x)^2, & x_{j+2} \le x \le x_{j+3} \end{cases}.$$

(iii) Die Funktionen $q_{lj} : [a,b] \to \mathbb{R}$, $j = 0, \ldots, n-1$ mit

$$q_{lj}(x) = (x - x_j)_+^l = \begin{cases} (x - x_j)^l, & \text{falls } x \ge x_j \\ 0, & \text{sonst} \end{cases}$$

sind Splines vom Grad l zu Δ

(14.2) Satz (Basis von $S_l(\Delta)$). Die Menge $S_l(\Delta)$ ist ein linearer Teilraum der Dimension $n + l$. Die Elemente $p_k(x) = x^k$, $k = 0, \ldots, l$ und $q_{lj}(x) = (x - x_j)_+^l$, $j = 1, \ldots, n-1$ bilden eine Basis von $S_l(\Delta)$.

Beweis. Mit dem Existenzsatz für sogenannte gewöhnliche Differentialgleichungen. □

Wir erhalten die Basisdarstellung

$$s(x) = \sum_{k=0}^{l} a_k x^k + \sum_{j=1}^{n-1} b_k (x - x_j)_+^l \tag{14.3}$$

für gewisse $a_k, b_k \in \mathbb{R}$.

Sei beispielsweise $s(x_j) = f_j$, $f_j = f(x_j)$ für ein f, das genügend oft diffbar ist. Wegen $dim(S_l(\Delta)) = n + l$ gibt es noch $l - 1$ Freiheitsgrade. Sei $l = 2m + 1$ ungerade, so ist $l - 1$ gerade. Wir können nun jeweils $\frac{l-1}{2}$ Bedingungen in $x_0 = a$ und $x_n = b$ fordern.

14.2 Kubische Splines ($l = 3$)

Es gilt

$$s(x_j) = f(x_j) = f_j \tag{14.4}$$

für $j = 0, \ldots, n$. Wir haben noch $l - 1 = 2$ Freiheitsgrade. Es gibt nun 3 Typen von Randbedingungen:

$$\left.\begin{array}{lll} (a) & \text{Natürliche Endbedingungen:} & s''(a) = s''(b) = 0, \\ (b) & \text{HERMITE Endbedingungen:} & s'(a) = f'(a), s'(b) = f'(b), \\ (c) & \text{Periodische Endbedingungen:} & s^{(i)}(a) = s^{(i)}(b),\ i = 0,1,2, \\ & & \text{falls } f \text{ periodisch}, f^{(i)}(a) = f^{(i)}(b) \text{ für } i = 0,1,2 \end{array}\right\} \tag{14.5}$$

Minimum-Norm-Eigenschaft kubischer Splines

(14.6) Definition. Wir definieren eine (Semi-)Norm durch

$$||f||_2 := \sqrt{\int_a^b f''(x)^2 dx}.$$

Für diese Norm gilt: $||f||_2 = 0 \Leftrightarrow f''(x) \equiv 0 \Leftrightarrow f$ ist linear in $[a,b]$.

(14.7) Satz (Existenz, Eindeutigkeit und Extremaleigenschaft kubischer Splines). Sei $f \in C^2[a,b]$. Dann gibt es genau einen Spline $s \in S_3(\Delta)$, der (14.4) und eine der Interpolationsbedingungen in (14.5) erfüllt. Dieser interpolierende Spline genügt der Minimum-Norm-Bedingung

(i) $||s||_2^2 \leq ||f||_2^2$

(ii) $0 \leq ||f-s||_2^2 = ||f||_2^2 - ||s_2^2||$.

Beweis. Zu $s \in S_3(\Delta)$ sei $d(x) := f(x) - s(x)$. Es gilt

$$||f-s||_2^2 = \int_a^b (f''(x)-s''(x))^2 dx = \int_a^b (f''(x)^2 - s''(x)^2 - 2d''(x)s''(x))dx = ||f||_2^2 - ||s||_2^2 - 2\int_a^b d''(x)s''(x)dx.$$

Wegen $s \in C^2[a,b]$ muss die partielle Integration zweimal angewandt werden. Wir erhalten

$$\begin{aligned}\int_a^b d''(x)s''(x)dx &= \sum_{j=1}^n \int_{x_{j-1}}^{x_j} d''(x)s''(x)dx \\ &= \sum_{j=1}^n \left([d'(x)s''(x) - d(x)s'''(x)]_{x_{j-1}}^{x_j} + \int_{x_{j-1}}^{x_j} d(x)s^{(4)}(x)dx \right)\end{aligned}$$

Da $s \in S_3(\Delta)$, folgt $s^{(4)}(x) \equiv 0$ in $[a,b]$. Aus (14.4) und (14.5) folgt

$$\sum_{j=1}^n [d'(x)s''(x) - d(x)s'''(x)]_{x_{j-1}}^{x_j} = [d'(x)s''(x) - d(x)s'''(x)]_a^b = 0.$$

Als Ergebnis erhalten wir

$$0 \leq ||f-s||_2^2 = ||f||_2^2 - ||s||_2^2,$$

also $||s||_2^2 \leq ||f||_2^2$.
Die Existenz folgt aus der Basisdarstellung (14.3). Für die Eindeutigkeit sei $\tilde{s}$ ein weiterer interpolierender Spline. Es gilt dann

$$||\tilde{s}-s||_2^2 = ||\tilde{s}||_2^2 - ||s||_2^2.$$

Durch Vertauschen von $\tilde{s}$ und s erhält man $||\tilde{s}-s||_2^2 = 0$. Also ist $\tilde{s}-s$ eine lineare Funktion in $[a,b]$. Da $\tilde{s}(a) = s(a)$ und $\tilde{s}(b) = s(b)$, folgt schon $\tilde{s} = s$. □

Geometrische Interpretation von $||s||_2^2 \leq ||f||_2^2$.

Unter der Krümmung $k(x)$ einer Kurve $y = f(x)$ in der $x-y$-Ebene versteht man die Funktion

$$k(x) = \frac{f''(x)}{\sqrt{1+f'(x)^2}}.$$

Unter der Annahme $|f'(x)| \ll 1$ erhält man

$$||k||_2^2 \approx \int_a^b f''(x)^2 dx.$$

Also minimiert der interpolierende Spline $s(x)$ in erster Näherung die mittlere Gesamtkrümmung.

14.3 Berechnung kubischer Splines

Zu berechnen sei der Spline $s \in S_3(\Delta)$, d.h. $s \in C^2[a,b]$ mit $s(x_j) = f_j = f(x_j)$ für $j = 0, \ldots, n$, welcher zusätzlich eine der Eigenschaften (14.5) a,b oder c erfüllt.
Wir setzen $h_j := x_j - x_{j-1}$ für $j = 1, \ldots, n$ und definieren die Momente $M_j := s''(x_j)$ für $j = 0, \ldots, n$. Dann ist $x''(x)$ linear in $[x_{j-1}, x_j]$ für $j = 1, \ldots, n$ und es gilt

$$s''(x) = \frac{1}{h_j}\left(M_j(x - x_{j-1}) + M_{j-1}(x_j - x)\right), \quad x_{j-1} \le x \le x_j.$$

Durch Integration von $s''(x)$ in $[x_{j-1}, x_j]$ erhalten wir

$$s'(x) = \frac{1}{2h_j}\left(M_j(x - x_{j-1})^2 - M_{j-1}(x_j - x)\right) + a_j \text{ und}$$

$$s(x) = \frac{1}{6h_j}\left(M_j(x - x_{j-1})^3 + M_{j-1}(x_j - x)^3\right) a_j(x - x_{j-1}) + b_j.$$

Dass man statt $a_j x$ hier $a_j(x - x_{j-1})$ benutzt, liegt an der Wahlfreiheit für die Konstante. Die getroffene Wahl erweist sich als rechentechnisch günstig.

Die Berechnung von a_j und b_j erfolgt aus den Gleichungen $s(x_j) = f_j$ für $j = 0, \ldots, n$:

$$M_{j-1}\frac{h_j^2}{6} + b_j = f_{j-1},\ M_j\frac{h_j^2}{6} + a_j h_j + b_j = f_j.$$

Als Lösungen erhalten wir

$$b_j = f_{j-1} - M_{j-1}\frac{h_j^2}{6}, \quad a_j = \frac{f_j - f_{j-1}}{h_j} - \frac{h_j}{6}(M_j - M_{j-1}).$$

Für $s'(x)$ in $[x_{j-1}, x_j]$ erhält man für $x = x_j^-$

$$s'(x_j^- = \frac{1}{2h_j}M_j h_j^2 + a_j = \frac{f_j - f_{j-1}}{h_j} + \frac{h_j}{3}M_j + \frac{h_j}{6}M_{j-1}.$$

Analog ergibt sich für $s'(x)$ in $[x_j, x_{j+1}]$ mit $x = x_j^+$

$$s'(x_j^+) = \frac{f_{j+1} - f_j}{h_{j+1}} + \frac{h_{j+1}}{3}M_j - \frac{h_{j+1}}{6}M_{j+1}.$$

Wegen $s'(x_j^-) = s'(x_j^+)$ erhält man das LGS

$$\mu_j M_{j-1} + M_j + \lambda_j M_{j+1} = d_j,\ j = 1, \ldots, n-1, \tag{14.8}$$

wobei gilt

$$\mu_j = \frac{h_j}{2(h_j + h_{j+1}}, \lambda_j = \frac{h_{j+1}}{2(h_j + h_{j+1}}, d_j = \frac{3}{h_j + h_{j+1}}\left(\frac{f_{j+1} - f_j}{h_{j+1}}\frac{f_j - f_{j-1}}{h_j}\right).$$

Für äquidistante Knoten gilt $h = h_j = \frac{b-a}{n}$ für $j = 1, \ldots, n$, und daher

$$\mu_j = \lambda_j = \frac{1}{4},\ j = 1, \ldots, n-1.$$

Fall (14.5) a) Natürliche Endbedingungen. Vorgegeben sind $M_0 = s''(a)$ und $M_n = s''(b)$. Für $M_1, \ldots, M_{n-1}$ erhalten wir das LGS

$$\begin{pmatrix} 1 & \lambda_1 & & & 0 \\ \mu_2 & 1 & \lambda_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \mu_{n-2} & 1 & \lambda_{n-2} \\ 0 & & & \mu_{n-1} & 1 \end{pmatrix} \cdot \begin{pmatrix} M_1 \\ \vdots \\ M_{n-1} \end{pmatrix} = \begin{pmatrix} d_1 - \mu_1 M_0 \\ d_2 \\ \vdots \\ d_{n-2} \\ d_{n-1} - \lambda_{n-1} M_n \end{pmatrix}.$$

Diese Matrix ist tridiagonal und diagonaldominant wegen $\mu_j + \lambda_j = \frac{1}{2}$. Die LR-Zerlegung existiert nach Satz (..). Für äquidistante Knoten ist die Matrix symmetrisch positiv definit.

Fall (14.5) b) HERMIT-Endbedingung. Vorgegeben sind $s'(a) = f'(a) = f'_0$ und $s'(b) = f'(b) = f'_n$. Aus der Darstellung von $s'(x)$ in $[a = x_0, x_1]$ und $[x_{n-1}, x_n = b]$ erhält man

$$\frac{h_1}{3}M_0 + \frac{h_1}{6}M_1 = \frac{f_1 - f_0}{h_1} - f'_0 \text{ und } \frac{h_n}{6}M_{n-1} + \frac{h_n}{3}M_n = f'_n - \frac{f_n - f_{n-1}}{h_n}.$$

Man setze $\lambda_0 = \frac{1}{2}$, $d_0 = \frac{3}{h_1}\left(\frac{f_1 - f_0}{h_1} - f'_0\right)$, $\mu_n = \frac{1}{2}$ und $d_n = \frac{3}{h_n}\left(f'_n - \frac{f_n - f_{n-1}}{h_n}\right)$. Dies ergibt dann ein LGS für $M_0, \ldots, M_n$:

$$\begin{pmatrix} 1 & \lambda_0 & & & 0 \\ \mu_1 & 1 & \lambda_1 & & \\ & \ddots & \ddots & \ddots & \\ & & \mu_{n-1} & 1 & \lambda_{n-1} \\ 0 & & & \mu_n & 1 \end{pmatrix} \cdot \begin{pmatrix} M_0 \\ \vdots \\ M_n \end{pmatrix} = \begin{pmatrix} d_0 \\ d_1 \\ \vdots \\ d_{n-1} \\ d_n \end{pmatrix}.$$

Diese Matrix ist tridiagonal und diagonaldominant, also LR-zerlegbar.

Fall (14.5) c) Periodische Endbedingungen. Es gilt $M_0 = M_n$ wegen $s''(a) = s''(b)$ und $f_0 = f_n$ wegen $s(a) = s(b)$. Die Gleichung $s'(a) = s'(b)$ ergibt

$$\mu_n M_{n-1} + M_n + \lambda_n M_{n+1} = d_n,$$

wobei man $h_{n+1} = h_1$, $M_{n+1} = M_1$ und $f_{n+1} = f_1$ setzt. Dies ergibt ein Gleichungssystem für die Variablen $M_1, \ldots, M_n$:

$$\begin{pmatrix} 1 & \lambda_1 & & & \mu_1 \\ \mu_2 & 1 & \lambda_2 & & 0 \\ & \ddots & \ddots & \ddots & \\ 0 & & \mu_{n-1} & 1 & \lambda_{n-1} \\ \lambda_n & & & \mu_n & 1 \end{pmatrix} \cdot \begin{pmatrix} M_1 \\ \vdots \\ M_n \end{pmatrix} = \begin{pmatrix} d_1 \\ \vdots \\ d_n \end{pmatrix}.$$

Diese Matrix ist nicht tridiagonal, aber dennoch diagonaldominant.

14.4 Konvergenzeigenschaften

Gegeben sei eine Folge von Zerlegungen von $[a,b]$, $\Delta_N = \{a = x_0^{(N)} < \ldots < x_{n_N}^{(N)} = b\}$. Es sei $||\Delta_N|| = \max_j \left(x_{j+1}^{(N)} - x_j^{(N)}\right)$.

(14.9) Satz. Sei $f \in C^4[a,b]$ mit $L := ||f^{(4)}||_\infty$. Sei weiterhin $\{\Delta_N\}_{N\in\mathbb{N}}$ eine Folge mit

$$\sup_j \frac{||\Delta_N||}{x_{j+1}^{(N)} - x_j^{(N)}} \leq K < \infty$$

und seien $s_N(\cdot)$ die zu f gehörigen Splines mit $s_N(x) = f(x)$ für $x \in \Delta_N$ sowie $s_N'(x) = f'(x)$ für $x = a$ und $x = b$.
Dann gibt es von Δ_N unabhängige Konstanten $c_i \leq 2$ für $i = 0,1,2,3$ mit $\left|f^{(i)}(x) - s_N^{(i)}(x)\right| \leq c_i \cdot K \cdot L \cdot ||\Delta_N||^{4-i}$ für $i = 0,1,2,3$ und für alle $x \in [a,b]$.

Beweis. Im Seminar zur Numerik (SS2007) oder beispielsweise im Stoer, Teil 1, §2.4.3. □