Numerische Mathematik 1

Prof. Dr. Lars Diening

16. Januar 2012

Inhaltsverzeichnis

Kapitel 1

Einführung

Ein Hauptziel der Numerik ist die Simulation von komplexen Naturvorgängen auf Computern. Dadurch werden kostspielige Experimente wie beispielsweise Windkanalversuche oder Feuchtigkeitstests von Betonkonstruktionen durch günstige, beliebig wiederholbare ersetzt. Auch das Wetter wird auf ähnliche Weise voraus berechnet. Die verwendeten Verfahren beruhen auf einfachen Bausteinen (z.B. Integralberechnungen, Lösen linearer Gleichungssysteme, Berechnung von Nullstellen), die meist schon aus der Vor-Computer-Zeit kommen. Ziel dieses Skripts und dieser Vorlesung ist es, diese Bausteine vorzustellen. Vorallem aber gehört zur numerischen Lösung eines Problems unbedingt auch eine *Analyse des Fehlers* - hierbei unterscheiden wir die folgenden Fehlerarten:

- *Modellfehler*
  - *Idealisierungsfehler:* Zur Beschreibung des Problems wird ein mathematisches Modell gebildet. Oft müssen Vereinfachungen - etwa Linearisierungen - vorgenommen werden.
  - *Datenfehler:* Die Daten des mathematischen Modells sind oft nur ungenau bekannt. Das betrifft z.B. Koeffizienten oder Materialeigenschaften.
- *Numerische Fehler*
  - *Diskretisierungsfehler:* Kontinuierliche Prozesse werden durch endliche Prozesse ersetzt. Ein Beispiel hierfür sind die Treppenfunktionen beim Integral.
  - *Abbruchfehler:* Unendliche Algorithmen (z.B. Limesbildung) werden nach endlich vielen Schritten abgebrochen.
  - *Rundungsfehler:* Computer haben einen endlichen Zahlenbereich (etwa $\frac{1}{3} \approx 0.3333$)

Dies soll an einem Beispiel demonstriert werden: Ein Stahlsteil der Länge $L = 1$ hängt an zwei Masten. Es soll die Auslenkung des Seils unter einer Last (Schwerkraft, Trapezkünstler) berechnet werden.

[Graphik-1: Stahlseil]

Die Auslenkung sei hierbei $y(x)$. Im *physikalischen Modell* gibt es zwei *Energien*:

(a) die *potentielle Energie*.

(b) die *Spannungsenergie*.

Zu (a): Die Belastung bei x sei $f(x)$. Damit ist die potentielle Energie $E_{pot} = -\int_0^1 f(x)\, y(x)\, dx$.

Zu (b): Nach dem *Hookeschen Gesetz* (lineare Elastizität) gilt: Spannungsenergie $\propto$ Längenänderung. Die Änderung beträgt dann $\Delta x \cdot \left(\sqrt{1 + |y'(x)|^2} - 1\right)$. Es folgt mit einer Materialkonstanten c:

$$\text{Spannungsenergie} \;=\; c \cdot \int_0^1 \sqrt{1 + |y'(x)|} - 1\, dx = \int_0^1 \frac{|y'(x)|^2}{\sqrt{1 + |y'(x)|^2} + 1}\, dx$$

Für kleine Auslenkungen ($|y| << 1$) gilt

$$\frac{|y'(x)|^2}{\sqrt{1+|y'(x)|^2}+1} \approx \frac{1}{2}|y'(x)|$$

und mit der Taylorentwicklung, $z = |y'(x)|^2$: $\sqrt{1+z}-1 \approx \frac{1}{2}z \pm \cdots$ Damit erhalten wir das mathematische Modell:

$$J(y) = \int_0^1 \frac{1}{2}|y'(x)|^2\,dx - \int_0^1 y(x)\,f(x)\,dx \to \min!$$

Nun soll die Funktion y bestimmt werden. Hierzu nützen wir die *1.Variation*. Sei $w \in C_0^\infty((0,\infty))$. Die 1. Variation berechnet sich zu

$$\delta J(y)(w) = \frac{d}{d\varepsilon}J(y+\varepsilon\cdot w)|_{\varepsilon=0} = \int_0^1 c\cdot y'(x)\,w'(x)\,dx - \int_0^1 w(x)\,f(x)\,dx =$$

$$= \int_0^1 \left(-cy''(x) - f(x)\right) w(x)\,dx$$

Wir erhalten damit die Bedingung

$$\boxed{-cy'' = f}$$

auf $(0,1)$ mit den Randwerten $y(0) = y(1) = 0$. Zur Lösung des Problems wird nun eine Diskretisierung vorgenommen:

$$t_i = i\cdot h\ ,\ i = 0,\cdots,N+1\ ,\ h = \frac{1}{N+1} \tag{1.1}$$

$$f_i = f(t_i) \tag{1.2}$$

$$y''(t_i) = \frac{1}{h^2}\left(y(t_{i+1}) - 2y(t_i) + y(t_{i-1})\right) \tag{1.3}$$

Dies führt auf ein lineares Gleichungssystem:

$$\frac{c}{h^2}\begin{pmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ \vdots & 0 & \ddots & \ddots & \vdots \end{pmatrix}\cdot\begin{pmatrix}\eta_1 \\ \vdots \\ \eta_n\end{pmatrix} = \begin{pmatrix} f_1 \\ \vdots \\ f_n\end{pmatrix}$$

bzw. in Kurzform $A\eta = b$. Dieses Gleichungssystem besitzt wegen $\det(A) \neq 0$ eine eindeutige Lösung. Diese kann wegen $D\eta = D\eta - A\eta + b$ für

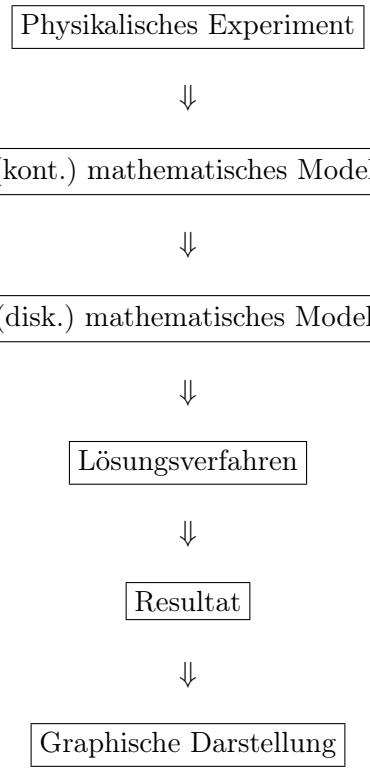
$$D = \begin{pmatrix} 2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & 2\end{pmatrix}$$

ausgehend vom Startvektor $\eta^0 \in \mathbb{R}^N$ durch Iteration $\eta^{k+1} = \eta^k - D^{-1}\left[A\eta^k - b\right]$ angenähert werden. (Bemerkung: $Id - D^{-1}A$ ist Kontraktion.) Man kann zeigen, dass $eta^k \to \eta,\ k \to \infty$. In der Praxis muss die Iteration jedoch abgerochen werden und es treten zusätzlich Rundungsfehler auf. Damit haben wir:

- Diskretisierungsfehler: $\max_i |\eta_i - y(t_i)|$
- Abbruchfehler: $\max_i |\eta_i^{(k)} - \eta_i|$

- Rundungsfehler: $\max_i |\tilde{\eta_i}^{(k)} - \eta_i^{(k)}|$

Zusammengefasst erhalten wir das folgende Diagramm:



1.1 Fehleranalyse

1.1.1 Gleitkommazahlen und Rundungsfehler

Bei der Abbildung der reellen Zahlen $\mathbb{R}$ auf den endlichen Zahlbereich des Computers entstehen zwangsweise Fehler. Damit stellt sich die Frage: *Wie bildet man reelle Zahlen ab?*

Eine *noramlisierte Gleitkommazahl* zur Basis $b \in \mathbb{N}, b \geq 2$ (meist $b = 2$ oder $b = 10$) ist eine Zahl der Form

$$x = \pm m \cdot b^{\pm e}$$

mit Mantisse $m = m_{-1}b^{-1} + m_{-2}b^{-2} + \cdots + m_{-r}b^{-r}$ und Exponent $e = e_{s-1}b^{s-1} + \cdots + e_1 b^1 + e_0 b^0$, wobei $m_i, e_i \in \{0, \cdots, b-1\}$ ist. Für $x \neq 0$ ist die Darstellung durch $m_1 \neq 0$ eindeutig.

Aus technischen Gründen bietet sicha auf dem Rechner das Binärsystem $b = 2$ an. Die in der obigen Form darstellbaren Zahlen heißen *Maschinenzahlen*. Sie bilden das *numerische Gleitkommagitter* $A(b, r, s)$. Da $A(b, r, s)$ endlich ist, gibt es eine größte und eine kleinste Zahl.

- (vom Betrag her) größte Zahl: $\pm(b-1)\left(b^{-1} + \cdots + b^{-r}\right) \cdot b^{(b-1)\left(b^{s-1}+\cdots+b^0\right)}$
- (vom Betrag her) kleinste Zahl: $\pm b^{-1} \cdot b^{-(b-1)\left(b^{s-1}+\cdots+b^0\right)}$

Beispiel: $b = 10, r = 4$, Lichtgeschwindigkeit $= 0.2998 \cdot 10^9 \frac{m}{s}$.

IEEE-Format: (double presicion, REAL 8 in Fortran) Zur Darstellung haben wir 64 Bit (=8 Byte) zur Verfügung. Als Basis wird $b = 2$ gewählt. Für $x \neq 0$ beginnt die Darstellung wegen $m_1 \neq 0$ immer mit $m_1 = 1$. Daher muss man für m_1 kein Bit zur Darstellung opfern und man gewinnt ein Bit Genauigkeit. Die Zahlen mit $m = 11\cdots1b$ und $m = 00\cdots00b$ haben eine Sonderrolle im IEEE Format und dienen zur Darstellung von $\pm 0, \pm\infty$ (NaN - Not a Number).

1.1.2 Rundung

Bei der Darstellung von Maschinenzahlen soll so gerundet werden, dass

$$|x - \mathrm{rd}\,(x)\,| = \min_{y \in A} |x - y|$$

wobei $x \mapsto \mathrm{rd}\,(x)$ die Rundungsabbildung ist. Bei IEEE ($b = 2$) ist das die natürliche Rundung:

$$\mathrm{rd}\,(\pm 0, m_1, \cdots, m_{53} m_{54} \cdots) = \left\{ \left(\begin{array}{c} \pm 0, m_1 ... m_{53} \text{ falls } m_{54} = 0, \\ \mathrm{rd}\,(x) = \pm 0, m_1 ... m_{53} + 2^{-53} \text{ falls} m_{54} = 1 \end{array} \right) \right\}$$

für $x \in [x_{negmin}, x_{negmax}] \cup \{0\} \cup [x_{posmin}, x_{posmax}]$. Bei der Rundung entsteht der maxmale Fehler

$$|x - \mathrm{rd}\,(x)\,| \leq \underbrace{\frac{1}{2} b^{-r}}_{\text{letzte halbe Ziffer}} \cdot b^e$$

Das ergibt den relativen Rundungsfehler

$$\frac{|x - \mathrm{rd}\,(x)\,|}{|x|} \leq \frac{1/2 b^{-r}}{|m| \cdot b^e} \leq \frac{1}{2} b^{1-r} = \text{ Maschinengenauigkeit} eps$$

Die arithmetischen Grundoperationen $\circ \in \{+, -, \cdot, /\}$ werden durch Maschinenoperationen ersetzt: $\otimes \in \{\oplus, \ominus, \odot, \oslash\}$. Es gilt

$$x \otimes y = \mathrm{rd}\,(x \circ y) = (x \circ y)\,(1 + \varepsilon) \text{ mit } |\varepsilon| \leq \varepsilon$$

Beachte, dass die Kommutativ- und Assoziativgesetze Im Allgemeinen *nicht* mehr gelten:

$$(x \oplus y) \oplus z \neq x \oplus (y \oplus z)$$

$$(x \oplus y) \odot z \neq (x \odot z) \oplus (y \odot z)$$

Insbesondere gilt

$$x \oplus y = x \text{ falls } |y| \leq \frac{|x|}{b} eps$$

1.1.3 Auslöschung

Wir betrachten einführend ein *Beispiel* für $b = 10$ und $r = 3$: Für $x = 0.9995$ und $y = -0.9984$ ist $x + y = 0.0011$. Das Ergebnis sollte laso nach Rundung $\mathrm{rd}\,(x + y) = 0.001$ sein. Addieren wir jedoch die gerundeten Werte $\mathrm{rd}\,(x) = 1.000$ und $\mathrm{rd}\,(y) = -0.998$ und runden anschließend wieder, so erhalten wir $\mathrm{rd}\,(x) \oplus \mathrm{rd}\,(y) = \mathrm{rd}\,(0.002) = 0.002 = 0.200 \cdot 10^{-2}$. D.h. wir haben einen größeren relativen Fehler von $\frac{0.002}{0.0011} \approx 1.81 \cdots$, obwohl wir mit 4 Stellen Genauigkeit gerechnete haben. Insofern ist Vorsicht geboten bei der Subtraktion ähnlich großer Zahlen. Durch geschickte Algorithmen kann man dieses Problem jedoch oft umgehen. Dies wollen wir am Beispiel der quadratischen Gleichung

$$x^2 + px + q = 0 \tag{1.4}$$

demonstrieren. Die Nullstellen sind gegeben durch

$$x_{1,2} = -\frac{p}{2} \pm \sqrt{\frac{p^2}{4} - q}$$

Konzentrieren wir uns auf den Fall $p > 0$ mit $|q| << \frac{p^2}{4}$. Dann kommt es nämlich bei x_1 zur Auslöschung, denn dann ist

$$\sqrt{\frac{p^2}{4} - q} \approx \frac{p}{2}$$

Dadurch erhält man für x_1 mit dieser Formel nur eine kleine relative Genauigkeit. Die Formel macht jedoch für x_2 keine Probleme. Aus dem Satz von *Vieta* folgt

$$x_1 x_2 = q$$

Somit können wir x_1 auch mittels der Beziehungen

$$x_2 = \frac{p}{2} + \sqrt{\frac{p^2}{4} - q}$$

$$x_1 = \frac{q}{x_2}$$

berechnen. Diese Berechnung für x_2 liefert deutlich bessere relative Fehler.

1.1.4 Kondition

Ein numerisches Problem heißt *gut konditioniert*, wenn kleine Ströungen der Daten nur kleine Änderungen der Ergebnisse zur Folge haben. (Beachte: Diese Definition wird meist auf den *relativen Fehler* angewandt, da dieser in der Numerik eine besonders wichtige Rolle spielt.)

1.1.5 Rundungsfehler

Wir betrachten nun das lineare Gleichungssystem

$$\begin{pmatrix} 1.2969 & 0.8648 \\ 0.2162 & 0.1441 \end{pmatrix} \cdot \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 0.86419999 \\ 0.14400001 \end{pmatrix} \tag{1.5}$$

mit der Kurzschreibweise $Ay = x$. Die Lösung ist gegeben durch

$$y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 0.86419999 \\ 0.14400001 \end{pmatrix}$$

Betrachten wir jedoch das gerundete Gleichungssystem $A\hat{y} = \hat{x}$ (wobei wir x auf 4 Stellen runden):

$$\begin{pmatrix} 1.2969 & 0.8648 \\ 0.2161 & 0.1441 \end{pmatrix} \cdot \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \end{pmatrix} = \begin{pmatrix} 0.8642 \\ 0.1440 \end{pmatrix} \tag{1.6}$$

so ergibt sich die Lösung

$$\begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \end{pmatrix} = \begin{pmatrix} 2 \\ -2 \end{pmatrix}$$

Wir bemerken, dass *kleine* Änderungen von x *große* Änderungen von y zur Folge haben. Berechnen wir dies genauer, so erhalten wir

$$\Delta x = \hat{x} - x = \begin{pmatrix} 0.00000001 \\ -0.00000001 \end{pmatrix}$$

und

$$\Delta y = \hat{y} - y = \begin{pmatrix} 1.0089 \\ 1.5130 \end{pmatrix}$$

Dadurch ergeben sich die *relativen Fehler*

$$\frac{|\Delta x|}{|x|} \approx 1.1 \cdot 10^{-8} \text{ und } \frac{|\Delta y|}{|y|} \approx 0.56 \tag{1.7}$$

Der relative Fehler von x hat sich also beim Lösen des Gleichungssystems um den Faktor $5 \cdot 10^7$ verstärkt. Folglich ist das Problem $x \mapsto A^{-1}x$ *schlecht konditioniert*. Die Konditionierung linearer Gleichungssysteme werden wir später noch genauer untersuchen.

1.1.6 Konditionszahlen

Zu gegebenen Daten $x \in \mathbb{R}^m$ und Resultat $y \in \mathbb{R}^n$ seien $x + \Delta x$ und $y + \Delta y$ gestörte Daten. Dann heißen $|\Delta x$ und $|\Delta y|$ *(absolute) Fehler* und $\frac{|\Delta x|}{|x|}, \frac{|\Delta y|}{|y|}$ *relative Fehler*. Der relative Fehler ist in der Anwendung meist wichtiger als der Fehler selbst: So ist beispielsweise ein Fehler von ±50km klein bei der Bestimmung des Abstands Mond-Erde, jedoch sehr groß beim Abstand Augsburg-München. Seien nun $f\colon \mathbb{R}^m \to \mathbb{R}^n$ eine glatte Funktion und sei $y = f(x)$. Für ein gestörtes Urbild $x + \Delta x$ definieren wir das gestörte Ergebnis $y + \Delta y = f(x + \Delta x)$. Wir wollen im Folgenden die Veränderung des relativen Fehlers komponentenweise untersuchen. Hierzu berechnen wir

$$\Delta y_i = f_i(x + \Delta x) - f_i(x) = \sum_{j=1}^{m} \partial_j f_i(x)\, \Delta x_j + R_i^f(x, \Delta x) \tag{1.8}$$

mit

$$\frac{|R_i^f(x, \Delta x)|}{|\Delta x|} \to 0,\ |\Delta x| \to 0 \tag{1.9}$$

Es gilt sogar $|R_i^f(x, \Delta x)| \in O\left(|\Delta x|^2\right)$, falls $f \in C^2$ ist. Wir benutzen im Folgenden *Landausche Symbole*:

$$g(t) = O(h(t)),\ t \to 0 \text{ falls } |g(t)| \le c \cdot |h(t)| \text{für ein } c > 0 \text{ und alle} |t| < t_0 \tag{1.10}$$

$$g(t) = o(h(t)),\ t \to 0 \text{ falls } |g(t)| \le c(t)\,|h(t)| \text{ mit } c(t) \to 0 \text{ für } t \to 0 \tag{1.11}$$

Sei nun $g \in C^2$. Dann gilt für einen Wert τ zwischen t und $t + \Delta t$

$$g(t + \Delta t) = g(t) + g'(t)\,\Delta t + g''(\tau)\,\frac{(\Delta t)^2}{2}$$

Dies können wir umschreiben zu

$$g(t + \Delta t) = g(t) + g'(t)\,\Delta t + O\left((\Delta t)^2\right)$$

und

$$\frac{g(t + \Delta t) + g(t)}{\Delta t} = g'(t) + O(\Delta t)$$

Für unsere obige Funktion f ergibt sich damit für $i = 1, \cdots, n$:

$$\Delta y_i = \sum_{j=1}^{m} \delta_j f_i(x)\,\Delta x_j + O\left(|\Delta x|^2\right) \tag{1.12}$$

Dies ergibt den relativen Fehler in der *i-ten* Komponente:

$$\frac{\Delta y_i}{y_i} = \sum_{j=1}^{m} \underbrace{\left(\partial_j f_i(x) \frac{x_j}{f_i(x)}\right)}_{|\cdot|=k_{ij}} \frac{\Delta x_j}{x_j} + O\left(\frac{|\Delta x|^2}{|y_i|}\right) \tag{1.13}$$

Dabei ist der letzte Summand vernachlässigbar gegenüber dem ersten Teil, falls $|\Delta x| = o(|y_i|)$. Die k_{ij} heißen *(relative) Konditionszahlen* von f im Punkt x. Mann nennt $f(x) = y$ *schlecht konditioniert*, falls $k_{ij} >> 1$ und sonst *gut konditioniert*. Bei $|k_{ij}| < 1$ sprechen wir von *Fehlerdämpfung*, bei $|k_{ij}| > 1$ von *Fehlerverstärkung*. BEISPIEL. (Summe) Wie wir schon gesehen haben, ist $x \oplus y$ schlecht konditioniert, da es zur Auslöschung kommen kann. BEISPIEL. (Produkt) Sei $f(x_1, x_2) = x_1 \cdot x_2$. Dann ist

$$\partial_1 f(x_1, x_2) = x_2 \ , \ \partial_2 f(x_1, x_2) = x_1$$

Dies ergibt für die *relativen Konditionszahlen*

$$k_{11} = \partial_1 f(x) \frac{x_1}{f(x_1, x_2)} = 1$$

$$k_{12} = \partial_2 f(x) \frac{x_2}{f(x_1, x_2)} = 1$$

Damit ist das Problem gut konditioniert.

Die relativen Konditionszahlen sind komponentenweise definiert. D.h. k_{ij} hängt mit dem relativen Fehler von y_i bezüglich x_j zusammen. Man kann alternativ auch vektoriell rechnen und den relativen Fehler von dem gesamten Vektor y in Bezug auf den relativen Fehler von x analysieren. Dadurch erhält man

$$\frac{|\Delta y|}{|y|} = \underbrace{\left(|\nabla f| \cdot \frac{|x|}{|f(x)|}\right)}_{=k} \cdot \frac{|\Delta x|}{|x|} + O\left(\frac{|\Delta x|^2}{|y|}\right) \tag{1.14}$$

Im Falle von $f(x_1, x_2) = x_1 x_2$ ist dann

$$k = \frac{|x| \cdot |x|}{|x_1 x_2|} = \frac{x_1^2 + x_2^2}{x_1 x_2} \geq 2$$

sehr groß für $|x_1| << |x_2|$ und $|x_2| << |x_1|$. Dies führt zu falschen Einschätzungen. Dies ist z.B. auch der Fall bei der Funktion

$$g\colon \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} 100 \cdot x_1 \\ x_2 \end{pmatrix}$$

Wir rechnen

$$\nabla g = \begin{pmatrix} 100 & 0 \\ 0 & 1 \end{pmatrix}, \ |\nabla g| \approx 100, \ k = \frac{|\nabla g| \cdot |x|}{|(100 \cdot x_1, x_2)^t|} \approx 100$$

Für die relativen Konditionszahlen haben wir jedoch $k_{11} = k_{22} = 1$ und $k_{12} = k_{21} = 0$. BEISPIEL. (Division) Sei $f(x_1, x_2) = x_1/x_2$. Dann ist

$$k_{11} = |\partial_1 f(x) \frac{x_1}{f(x)}| = 1$$

$$k_{22} = |\partial_2 f(x) \frac{x_2}{f(x)}| = 1$$

Aber wir wissen schon, dass f schlecht konditioniert ist für $x_2 \approx 0$. Wie kann das sein - mit anderen Worten - wo ist der Widerspruch? Das Problem liegt im vernachlässigten Term,

$$f''(x) O\left(\frac{|\Delta x|^2}{|y_i|}\right)$$

der für $x_2 \approx 0$ eben *nicht* vernachlässigt werden kann.

1.1.7 Stabilität eines Verfahrens

Im Folgenden wollen wir zeigen, dass verschiedene Algorithmen zur Lösung eines Problems sehr unterschiedliche *Stabilitätsverhalten* haben. Dies wollen wir am Beispiel der quadratischen Gleichung verdeutlichen. Betrachten wir hierzu

$$x^2 + px + q = 0 \tag{1.15}$$

Durch neue Variablen können wir dies in die für uns einfachere Form

$$x^2 - 2bx + c = 0$$

bringen. Dann sind die Lösungen gegeben durch

$$x_1 = b + \sqrt{b^2 - c}\,,\ x_2 = b\sqrt{b^2 - c}$$

Wir wollen im Folgenden speziell den Fall $b >> c, b > 0$ betrachten, denn in diesem Fall ist $\sqrt{b^2 - c} \approx b$ und es wird bei x_2 möglicherweise zur Auslöschung kommen.

- 1. Algorithmus. Wir setzen

$$y_1 = b \cdot b$$

$$y_2 = y_1 - c$$

$$y_3 = \sqrt{y_2}$$

$$x_1 = y_4 = b + y_3$$

$$x_2 = y_5 = b - y_3$$

  Die ersten 4 Schritte sind stabil. Beim 5. Schritt ergibt sich allerdings folgendes Problem:

$$k_{rel} = |\frac{\partial y_5}{\partial y_4}\frac{y_4}{y_5}| = |-\frac{y_4}{y_5}| =$$

$$= |\frac{-\sqrt{b^2 - c}}{b - \sqrt{b^2 - c}}| = |\frac{\left(b - \sqrt{b^2 - c}\right)\sqrt{b^2 - c}}{c}| \approx \frac{2|b|^2}{|c|} >> 1$$

- 2. Algorithmus. Wir benutzen in diesm Algorithmus, dass $x_1 x_2 = x$ und setzen

$$y_1 = b \cdot b$$

$$y_2 = y_1 - c$$

$$y_3 = \sqrt{y_2}$$

$$x_1 = y_4 = b + y_3$$

$$x_2 = y_5 = \frac{c}{y_4}$$

$$\Rightarrow k_{rel}\,(3 \to 4) = |\frac{\partial y_4}{\partial y_3}\frac{y_3}{y_4}| = |\frac{y_3}{y_4}| = |\frac{\sqrt{b^2 - c}}{b + \sqrt{b^2 - c}}| \leq 1$$

$$k_{rel}\,(4 \to 5) = |\frac{\partial y_5}{\partial y_4}\frac{y_4}{y_5}| = |\frac{-c}{y_4^2} \cdot \frac{y_4}{y_5}| = 1$$

  Dieses Problem ist also gut konditioniert.

1.1.8 Lineare Gleichungssysteme

Kommen wir nun zur Fehleranalyse des Problems $Ay = x$, d.h. $A^{-1}x = y$ mit $x \in \mathbb{R}^m, y \in \mathbb{R}^n, A \in \mathbb{R}^{m\times n}$ sowie $A^{-1} \in \mathbb{R}^{n\times n}$ (falls invertierbar). Aus der linearen Algebra benötigen wir die folgenden Begriffe:

- $\mathbb{K}$-Vektorraum
- Norm
- Hilbertraum
- Skalarprodukt
- Cauchy-Schwarz-Ungleichung ($|\langle x, y\rangle| \leq \|x\| \cdot \|y\|$)
- Eigenvektor
- Determinante
- charakteristisches Polynom

Die Eigenschaften einer Norm $\|\cdot\|$ auf einem $\mathbb{K}$-Vektorraum V sind:

(N1) $\|x\| \geq 0$,

(N2) $\|x\| = 0 \Leftrightarrow x = 0$

(N3) $\forall \alpha \in \mathbb{K}, x \in V\colon\ \|\alpha \cdot x\| = |\alpha| \cdot \|x\|$

(N4) $\forall x, y \in V\colon\ \|x + y\| \leq \|x\| + \|y\|$

Beispiel. Die folgenden Ausdrücke definieren Normen auf dem $\mathbb{K}^n$.

(a) $\|x\|_2 = \sqrt{\langle x, x\rangle} = \sqrt{\sum_{i=1}^n |x_i|^2}$

(b) $\|x\|_1 = \sum_{i=1}^n |x_i|$

(c) $\|x\|_\infty = \max_{i=1}^n |x_i|$

(d) $\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$

Ferner gilt $\|x\|_p \to \|x\|_\infty$, $p \to \infty$. Wir erinnern uns an die Normäquivalenz auf $\mathbb{K}^n$:

Theorem 1.16

Auf $\mathbb{K}^n$ sind alle Normen äquivalent. *D.h. zu zwei Normen $\|\cdot\|$ und $|||\cdot|||$ gibt es Zahlen $k, K > 0$ mit*

$$k\|x\| \leq |||x||| \leq K\|x\| \tag{1.17}$$

Beweis: Es genügt, die Aussage für $|||\cdot||| = \|\cdot\|_\infty$ zu zeigen. Sei nun $\|\cdot\|$ eine beliebige andere Norm und $e_1, \cdots, e_n$ eine Basis von $\mathbb{K}^n$. Dann gilt:

$$\|x\| = \left\|\sum_{i=1}^n x_i e_i\right\| \leq \sum_{i=1}^n |x_i| \cdot \|e_i\| \leq \|x\|_\infty \cdot \underbrace{\sum_{i=1}^n \|e_i\|}_{=C_0} = C_0 \|x\|_\infty$$

Das klärt $|||x|||| \leq K \|x\|_\infty$. Die Identität $Id\colon\ (\mathbb{K}^n, \|x\|_\infty) \to (\mathbb{K}^n, \|\cdot\|)$ ist stetig. Sei nun $S = \{x \in \mathbb{K}^n\colon\ \|x\|_\infty = 1\}$ die Einheitssphäre bzgl. der $\|\cdot\|_\infty$-Norm. Da S beschränkt und abgeschlossen in $(\mathbb{K}^n, \|\cdot\|_\infty)$, ist S kompakt in $(\mathbb{K}^n, \|\cdot\|_\infty)$. Daher nimmt $Id\colon\ (\mathbb{K}^n, \|x\|_\infty) \to (\mathbb{K}^n, \|\cdot\|)$ sein Minimum und Maximum auf S an. Sei daher

$$k = \min\{\|x\|\colon\ x \in S\} \tag{1.18}$$

$$K = \max\{\|x\| : \ x \in S\} \tag{1.19}$$

Da $\|\cdot\|$ positiv definit ist, gilt $0 < a \leq b$ und damit

$$\forall x \in S: \quad k \leq \|x\| \leq K \tag{1.20}$$

$$\Leftrightarrow \ \forall x \in \mathbb{K}^n \setminus \{0\}: \quad k \leq \left\| \frac{x}{\|x\|_\infty} \right\| \leq K$$

$$\Leftrightarrow \ \forall x \in \mathbb{K}^n \setminus \{0\}: \quad k \|x\|_\infty \leq \|x\| \leq \|x\|_\infty$$

Da für $0 \in \mathbb{K}^n$ ohnehin $\|0\| = \|0\|_\infty = 0$ gilt, erhalten wir für alle $x \in \mathbb{K}^n$:

$$\Leftrightarrow \ \forall x \in \mathbb{K}^n: \quad k \|x\|_\infty \leq \|x\| \leq \|x\|_\infty \tag{1.21}$$

■

Eine Norm *induziert* eine Topologie bzw. einen Konvergenzbegriff durch $x_n \to x$ genau dann, wenn $\|x_n - x\| \to 0, n \to \infty$. Die durch $\|\cdot\|_\infty$ erzeugte Konvergenz entspricht der komponentenweisen Konvergenz. Folglich gilt dies nach dem Normäquivalenzsatz für jede Norm. Da $\mathbb{K}^{n\times n} \cong \mathbb{K}^{n^2}$ gilt, kann man analog Normen für Matrizen betrachten. Besonders wichtig sind *Operatornormen* und *verträgliche Normen.*

Definition 1.22

Eine Norm $|||\cdot|||$ *auf* $\mathbb{K}^{n\times n}$ *heitß* verträglich *mit* $\|\cdot\|$ *auf* $\mathbb{K}^n$*, falls*

$$\forall x \in \mathbb{K}^{n\times n}, A \in \mathbb{K}^{n\times n}: \ \|Ax\| \leq |||A||| \cdot \|x\| \tag{1.23}$$

Sie heißt Matrixnorm*, falls sie zusätzlich* submultiplikativ *ist, d.h.*

$$\forall A, B \in \mathbb{K}^{n\times n}: \ |||AB||| \leq |||A||| \cdot normmB \tag{1.24}$$

Bemerkung: Eine Matrixnorm macht $\mathbb{K}^{n\times n}$ zu einer *Banachalgebra.* Die Abbildung $(x, y) \mapsto \langle x, y\rangle = \sum_{i=1}^n x_i \overline{y_i}$ definiert ein Skalarprodukt auf $\mathbb{K}^n$, wobei $\overline{y_i}$ die komplex-konjugierte Zahl von y_i ist. Für eine Matrix $A \in \mathbb{K}^{n\times n}$ sei A^T die transponierte Matrix und $\overline{A}^T = A^*$ die adjungierte Matrix. Es gilt

$$\langle Ax, y\rangle = \langle x, A^* y\rangle$$

Theorem 1.25

Sei $\|\cdot\|$ *eine Norm auf* $\mathbb{K}^n$*, so wird durch*

$$\|A\| = \sup_{x\neq 0} \frac{\|Ax\|}{\|x\|} = \sup_{\|x\|\leq 1} \|Ax\| \tag{1.26}$$

eine zu $\|\cdot\|$ *verträgliche Matrixnorm auf* $\mathbb{K}n \times n$ *definiert. Diese heißt auch* Operatornorm *oder auch* induzierte Matrixnorm*. Sie ist die kleinste verträgliche Norm.*

Beispiel: *Die Frobeniusnorm.* Für $A, B \in \mathbb{K}^{n\times n}$ definiert

$$\langle A, B\rangle = \operatorname{tr}(B^* A) \tag{1.27}$$

ein Skalarprodukt auf $\mathbb{K}^{n\times n}$, wobei $\operatorname{tr}(B) = \sum_{i=1}^n B_{ii}$ die *Spur* der Matrix B ist. Daher wird durch

$$\|A\|_{\text{Frob}} = \sqrt{\langle A, A\rangle} = \sqrt{\operatorname{tr}(A^* A)} = \left(\sum_{i,j} |a_{ij}|^2 \right)^{1/2} \tag{1.28}$$

eine Norm auf $\mathbb{K}^{n\times n}$ definiert. Damit ist $\|\cdot\|_{\text{Frob}}$ eine Hilbertraumnorm. Es gilt $\|Ax\|_2 \leq \|A\|_{\text{Frob}}\|x\|_2$, d.h. $\|\cdot\|_{\text{Frob}}$ ist verträglich mit $\|\cdot\|_2$. Weiterhin gilt

$$\|AB\|_{\text{Frob}} \leq \|A\|_{\text{Frob}}\|B\|_{\text{Frob}}$$

Damit ist $\|\cdot\|_{\text{Frob}}$ eine mit $\|\cdot\|_2$ verträgliche Matrixnorm. *Beachte aber:* $\|\cdot\|_{\text{Frob}}$ ist *nicht* die durch $\|\cdot\|_2$ induzierte Matrixnorm. BEMERKUNG: Für induzierte Matrixnormen gilt stets $\|E_n\| = 1$. Wegen $\|E_n\|_{\text{Frob}} = \sqrt{n}$ ist $\|\cdot\|_{\text{Frob}}$ keine induzierte Matrixnorm. Auch $\frac{1}{\sqrt{n}}\|\cdot\|_{\text{Frob}}$ kann dies nicht beheben, da dann im Allgemeinen $\|AB\| \leq \|A\| \cdot \|B\|$ verletzt ist, z.B. durch die Matrix

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Wir bestimmen nun die durch $\|\cdot\|_1, \|\cdot\|_2, \|\cdot\|_\infty$ induzierten Matrixnormen, welche wir ebenso mit $\|\cdot\|_1, \|\cdot\|_2, \|\cdot\|_\infty$ bezeichnen.

Theorem 1.29

Es gilt

$$\|A\|_1 = \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}| \text{ (maximale Spaltensumme)} \tag{1.30}$$

$$\|A\|_\infty = \max_{1\leq k\leq n} \sum_{k=1}^{n} |a_{ik}| \text{ (maximale Zeilensumme)} \tag{1.31}$$

Beweis : Wir beschränken uns auf den Fall $\|\cdot\|_\infty$. der Fall $\|\cdot\|_1$ ist analog und wird in den Übungen behandelt. Sei also $|||\cdot|||_\infty = \max\limits_{1\leq k\leq n} \sum\limits_{k=1}^{n} |a_{ik}|$, so gilt

$$\|Ax\|_\infty = \|\sum_k a_{jk}x_k\|_\infty = \max_j |\sum_k a_{jk}x_k| \leq \max_j \sum_k |a_{ik}| \cdot \|x\|_\infty = |||A||| \cdot \|x\|_\infty$$

Hieraus folgt schon $\|A\|_\infty \leq |||A|||_\infty$. Es bleibt noch zu zeigen, dass $\|A\|_\infty \geq |||A|||_\infty$. Wir können ohne Einschränkung $A \neq 0$ annehmen. Dann gilt $|||A|||_\infty > 0$. Sei nun j_0 so, dass $|||A|||_\infty = \sum_{k=1}^{n} |a_{j_0,k}|$, d.h. das Maximum wird bei j_0 angenommen. Sei nun $z_k = \text{sgn}\,(a_{j_0,k})$, so ist $a_{j_0,k}z_k = |a_{j_0,k}|$ und $\|z\|_\infty = 1$. Es folgt

$$|\,(Az)_{j_0}\,| = \sum_k a_{j_0,k}z_k = \sum_k |a_{j_0,k}| = |||A|||_\infty$$

Dies ergibt

$$\|Az\|_\infty \geq |||A|||_\infty$$

Da $\|z\|_\infty = 1$, folgt hieraus wie gewünscht die Behauptung. ■

Zu der von $\|\cdot\|_2$ induzierten Matrixnorm kommen wir später. Da $\|\cdot\|_{\text{Frob}}$ mit $\|\cdot\|_2$ verträglich ist, wissen wir aber schon mal, dass

$$\|A\|_2 \leq \|A\|_{\text{Frob}}$$

Die Frobeniusnorm ist leicht zu berechnen, aber nur eine ungenaue obere Schranke für $\|A\|_2$.

1.1.9 Eigenwerte und Skalarprodukt

Die Eigenwerte einer Matrix A sind die Nullstellen des charakteristischen Polynoms $\chi_A(\lambda) = \det(\lambda I - A)$ (bis auf Vorzeichen). Sei $\sigma(A)$ die Menge der Eigenwerte von A (das sog. *Spektrum*). Zu jedem Eigenwert λ existieren Eigenvektoren $w \in \mathbb{K}^n \setminus \{0\}$ mit

$$Aw = \lambda w$$

Sei nun $\|\cdot\|$ eine Norm auf $\mathbb{K}^n$, so gilt

$$|\lambda| \cdot \|w\| = \|\lambda w\| = \|Aw\| \leq \|A\| \cdot \|w\|$$

Hieraus folgt $|\lambda| \leq \|A\|$ für jeden Eigenwert λ von A. Damit gilt

$$\rho(A) = \max_{\lambda \in \sigma(A)} |\lambda| \leq \|A\|$$

Der abgeschlossene Ball um Null in $\mathbb{C}$ mit Radius $\rho(A)$ ist der kleinste in Null zentrierte Ball, der das Spektrum $\sigma(A)$ enthält. Aus diesem Grund heißt $\rho(A)$ der *Spektralradius* von A.

Definition 1.32

Eine Matrix $A \in \mathbb{K}^{n\times n}$ heißt hermitesch *oder* selbstadjungiert, *falls*

$$A = A^* \tag{1.33}$$

Reelle, hermitesche Matrizen heißen symmetrisch.

Theorem 1.34

Die von $\|\cdot\|_2$ induzierte Matrixnorm heißt Spektralnorm.

Theorem 1.35

Es gilt

$$\|A\|_2 = \sqrt{\max\{|\lambda| \colon \lambda \in \sigma(A^*A)\}} = \sqrt{\rho(A^*A)} \tag{1.36}$$

Ist A hermitesch, so gilt sogar

$$\|A\|_2 = \max\{|\lambda| \colon \lambda \in \sigma(A)\} = \rho(A) \tag{1.37}$$

Beweis: Sei A hermitesch. Wir wissen schon, dass $\rho(A) \leq \|A\|_2$. Es bleibt also zu zeigen, dass $\|A\|_2 \leq \rho(A)$. Aus der linearen Algebra wissen wir, dass dann A nur reelle Eigenwerte besitzt. Ferner existiert ein Orthonormalsystem aus Eigenvektoren $w^1, \cdots, w^n$ mit

$$Aw^i = \lambda_i w^i \text{ und } \langle w^i, w^j \rangle = \delta_{ij}$$

Wir zeigen nun $\|A\|_2 = \max_i |\lambda_i|$. Jedes $x \in \mathbb{K}^n$ besitzt eine Darstellung

$$x = \sum_i \alpha_i w^i$$

Hieraus folgt

$$\|x\|_2^2 = \langle x, x \rangle = \sum_{i,j} \alpha_i \overline{\alpha}_j \langle w^i, w^j \rangle = \sum_i |\alpha_i|^2$$

$$\|Ax\|_2^2 = \langle Ax, Ax \rangle \sum_{i,j} \lambda_i \alpha_i \overline{\lambda_j \alpha_j} \langle w^i, w^j \rangle = \sum_i |\lambda_i \alpha_i|^2$$

Damit erhalten wir die Abschätzung

$$\|Ax\|_2 \leq \left(\max_i |\lambda_i|\right) \|x\|_2 = \rho(A) \|x\|_2$$

Damit haben wir gezeigt, dass $\|A\|_2 \leq \rho(A)$. Dies ergibt die Behauptung für hermitesche A. Kommen wir nun zum allgemeinen Fall. Sei $B \in \mathbb{K}^{n\times n}$ beliebig. Dann ist $A = B^*B$ hermitesch und positiv semidefinit. Es gilt

$$\|B\|_2^2 = \sup_{\|x\|\leq 1} \|Bx\|_2^2 = \sup_{\|x\|\leq 1} \langle Bx, Bx\rangle \leq \sup_{\|x\|\leq 1} \langle B^*Bx, x\rangle$$

$$\leq \sup_{\|x\|\leq 1} \|B^*Bx\| \cdot \|x\| \leq \|B^*B\| = \rho(B^*B)$$

nach dem ersten Teil. Es gilt also $\|B\|_2 \leq \rho(B^*B)$. Sei nun w_j ein Orthonormalsystem von A wie oben zu Eigenwerten λ_j. Dann gilt

$$\|Bw_j\|_2^2 = \langle Bw_j, Bw_j\rangle = \langle B^*Bw_j, w_j\rangle = \langle \lambda_j w_j, w_j\rangle = \lambda_j \|w_j\|_2^2 = |\lambda_j| \cdot \|w_j\|_2^2$$

Hierbei haben wir im letzten Schritt benutzt, dass die Eigenwerte nicht-negativ sind, da A positiv semidefinit ist. Also gilt

$$\|B\|^2 \geq \max_j |\lambda_j| = \rho(A) = \rho(B^*B)$$

Dies liefert die fehlende Abschätzung $\|B\|^2 \geq \sqrt{\rho(B^*B)}$. ■

Definition 1.38

Eine Matrix $A \in \mathbb{K}^{n\times n}$ heißt positiv semidefinit, *falls*

$$\forall x \in \mathbb{K}^n : \ \langle Ax, x\rangle \geq 0 \tag{1.39}$$

Sie heißt positiv definit, *falls*

$$\forall x \in \mathbb{K}^n \setminus \{0\} : \ \langle Ax, x\rangle > 0 \tag{1.40}$$

Theorem 1.41

Eine symmetrische Matrix $A \in \mathbb{K}^{n\times n}$ ist positiv semidefinit genau dann, wenn alle Eigenwerte nicht-negativ sind. Eine symmetrische Matrix ist positiv definit genau dann, wenn alle Eigenwerte positiv sind. In diesem Fall sind alle Diagonalemente positiv und das betragsmäßig größte Element liegt auf der Hauptdiagonalen.

Beweis : In der linearen Algebra. ■

Bemerkung: Ist $A \in \mathbb{K}^{n\times n}$ positiv semidefinit, so kann man analog zu dem vorherigen Lemma zeigen, dass es unter allen betragsmäßig größten Elementen von A eines auf der Hauptdiagonalen gibt. Im Folgenden heißt positiv (semi)definit immer positiv (semi)definit und hermitesch (symmetrisch). Nun aber zurück zur Fehleranalyse von $Ax = x$ bzw$y = A^{-1}x$ mit $A \in \mathbb{K}^{n\times n}$ invertierbar.

Theorem 1.42

Sei $B \in \mathbb{K}^{n\times n}$ mit $\|B\| < 1$, so ist $I + B$ invertierbar und

$$(I + B)^{-1} = \sum_{k\geq 0} (-1)^k B^k \tag{1.43}$$

$$\| (I + B)^{-1} \| \leq \frac{1}{1 - \|B\|} \tag{1.44}$$

Beweis: Sei $A = \sum_{k\geq 0} (-1)^k B^k$. Wegen $\| (-1)^k B^k \| \leq \|B\|^k$ konvergiert die Reihe in $\mathbb{K}^{n\times n}$ absolut. Es gilt

$$(I+B)A = (I+B)\sum_{k\geq 0}(-1)^k B^k = \lim_{K\to\infty}\sum_{k=0}^{K}(I+B)(-1)^k B^k =$$

$$= \lim_{K\to\infty}\sum_{k=0}^{K}(-1)^k B^k + \sum_{k=0}^{K+1}(-1)^k B^k =$$

$$= \lim_{K\to\infty}\left((-1)^{K+1}B^{K+1} + B^0\right) = E_n$$

wobei $(-1)^{K+1}B^{K+1}$ für $K\to\infty$. Für $\|A\|$ erhalten wir die Abschätzung:

$$\|(I+B)^{-1}\| \leq \sum_{k\geq 0}\|B\|^k = \frac{1}{1-\|B\|}$$

■

Wir kommen nun zu einem Hauptresultat dieses Kapitels, dem *Störungssatz*:

Theorem 1.45

Sei $A \in \mathbb{K}^{n\times n}$ regulär und $\|\delta A\|\|A^{-1}\| = \frac{\|\delta A\|}{\|A\|}\mathrm{cond}(A) < 1$, so ist $A+\delta A$ regulär und es gilt für $(A+\delta A)(y+\delta y) = x+\delta x$ die Abschätzung

$$\frac{\|\delta y\|}{\|y\|} \leq \frac{cond(A)}{1 - cond(A)\frac{\|\delta A\|}{\|A\|}}\left(\frac{\|\delta x\|}{\|x\|} + \frac{\|\delta A\|}{\|A\|}\right) \tag{1.46}$$

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

Beweis: Es gilt

$$\|A^{-1}\delta A\| \leq \|A^{-1}\|\cdot\|\delta A\| < 1 \Rightarrow A+\delta A = A\left(E_n + A^{-1}\delta A\right) \text{ ist regulär}$$

$$\Rightarrow (A+\delta A)^{-1} = \left(E_n + A^{-1}\delta A\right)^{-1}A^{-1}$$

Aus $(A+\delta A)(y+\delta y) = x+\delta x$ folgt dann

$$(A+\delta A)\delta y = \delta x - (\delta A)y \Rightarrow \delta y = (A+\delta A)^{-1}(\delta x - \delta A y)$$

$$\Rightarrow \|\delta y\| \leq \|(A+\delta A)^{-1}\|\cdot\|\delta x - (\delta A)y\| \leq$$

$$\leq \frac{1}{1-\|A^{-1}\|\delta A\|}\|A^{-1}\|(\|\delta x\| + \|\delta A\|\cdot\|y\|) \leq \frac{\|A^{-1}\|}{1-\|A^{-1}\|\cdot\|\delta A\|}(\|\delta x\| + \|\delta A\|\cdot\|y\|)$$

$$\Rightarrow \frac{\|\delta y\|}{\|y\|} \leq \frac{\|A^{-1}\|\cdot\|A\|}{1-\|A^{-1}\|\cdot\|\delta A\|}\left(\frac{\|\delta x\|}{\|A\|\cdot\|y\|} + \frac{\|\delta A\|}{\|A\|}\right) \leq$$

$$\leq \frac{\mathrm{cond}(A)}{1-\mathrm{cond}(A)\frac{\|\delta A\|}{\|A\|}}\left(\frac{\|\delta x\|}{\|x\|} + \frac{\|\delta A\|}{\|A\|}\right)$$

■

Bemerkung: $\mathrm{cond}(A)$ hängt von $\|\cdot\|$ ab.

(a) Für $\|\cdot\|_\infty$ gilt $\text{cond}(A) = \|A\|_\infty \|A^{-1}\|_\infty$, wobei $\|A\|_\infty$ die maximale Zeilensumme von A ist.

(b) Für $\|\cdot\|_2$ und A hermitesch gilt

$$\text{cond}(A) = \frac{\lambda_{max}}{\lambda_{min}} \tag{1.47}$$

Ist

$$\text{cond}(A)\,\frac{\|\delta A\|}{\|A\|} << 1$$

so ist

$$\frac{\|\delta y\|}{\|y\|} \leq 2\text{cond}(A)\left(\frac{\|\delta x\|}{\|x\|} + \frac{\|\delta A\|}{\|A\|}\right)$$

mit Versträkungsfaktor $\text{cond}(A)$. Wir haben die folgende *Regel*: sei $\text{cond}(A) \approx 10^s$ und

$$\left(\frac{\|\delta x\|}{\|x\|} + \frac{\|\delta A\|}{\|A\|}\right) \approx 10^{-k}, k > s$$

so gilt

$$\frac{\|\delta y\|}{\|y\|} \approx 10^{s-k}$$

D.h. man verliert s Stellen Genauigkeit. Hierzu ein bereits besprochenes Beispiel:

$$A = \begin{pmatrix} 1.2969 & 0.8648 \\ 0.2161 & 0.1441 \end{pmatrix},\ A^{-1} = 10^8 \cdot \begin{pmatrix} 0.1441 & -0.8648 \\ 0.2161 & 12.969 \end{pmatrix} \Rightarrow \|A\|_\infty = 2.1617$$

$$\|A^{-1}\|_\infty = 1.530 \cdot 10^8 \Rightarrow \text{cond}(A) \approx 3.3 \cdot 10^8$$

Es gehen also schlimmstenfalls 8 Stellen verloren. Das Problem ist demnach sehr schlecht konditioniert.

Bemerkung: Die Abschätzung ist im Wesentlichen scharf. Sei hierzu λ_1 der kleinste und λ_n der größte Eigenwert. sei weiter $w_1, \cdots, w_n$ ein Orthonormalsystem von Eigenvektoren und $\delta A = 0$ sowie $x = w_n + \varepsilon w_1$. Dann folgt

$$\frac{\|\delta x\|}{\|x\|} = \frac{\varepsilon\|w_1\|}{\|w_n\|} = \varepsilon \text{ und } Ay = w_n \Rightarrow y = \frac{1}{\lambda_n} w_n$$

$$A(\delta y) = \varepsilon w_n \Rightarrow \delta y = \varepsilon \frac{w_1}{\lambda_1}$$

$$\Rightarrow \frac{\|\delta y\|}{\|y\|} = \varepsilon \frac{\lambda_n}{\lambda_1} = \frac{\|\delta x\|}{\|x\|}\frac{\lambda_n}{\lambda_1}$$

$$\Rightarrow \frac{\|\delta y\|}{\|y\|} = \text{cond}(A)\,\frac{\|\delta x\|}{\|x\|}$$

Kapitel 2

Interpolation, Approximation und Projektion

Die Ziele dieses Kapitels sind

- Rekonstruiere eine Funktion f, die nur an endlich vielen Stellen $x_0, \ldots, x_n$ bekannt ist.
- Finde eine einfach auszuwertende Darstellung einer analytischen Funktion.

Hierbei nützen wir die folgenden Funktionenklassen:

- Polynome P:

$$P = a_0 + \cdots + a_n x^n$$

- rationale Funktionen:

$$\frac{a_0 + \cdots + a_n x^n}{b_0 + \cdots + b_m x^m}$$

- trigonometrische Polynome:

$$\frac{1}{2}a_0 + \sum_{k=1}^{\infty} \big(a_k \cos(kx) + b_k \sin(kx)\big)$$

Wir benutzen dabei die folgende Terminologie:

- *Interpolation:* Fixiere Funktion durch Funktionswerte
- *Approximation:* Minimiere Abstand zu $g \in \mathcal{P}$, z.B.
  (a) $\max\limits_{a \le x \le b} |f(x) - g(x)|$
  (b) $\|f - g\|_2$
  (c) $\max\limits_{i=0,\ldots,n} |f(x_i) - g(x_i)|$

2.1 Polynominterpolation

Wir definieren

$$\mathcal{P}_n = \big\{p(x) = a_0 + \cdots + a_n x^n \,:\, a_i \in \mathbb{R},\ i = 0, \ldots, n\big\}$$

Das sind die Polynome vom Grad n.

2.1.1 Die Lagrange-Interpolationsaufgabe

Wir wollen nun zu $n+1$ verschiedenen Stützstellen (*Knoten*) $x_0, \ldots, x_n \in \mathbb{R}$ und Knotenwerten $y_0, \ldots, y_n \in \mathbb{R}$ ein Polynom $p \in \mathcal{P}_n$ mit $p(x_i) = y_i,\ i = 1, \ldots, n$ finden.

Theorem 2.1

Die Lagrange-Interpolationsaufgabe *ist eindeutig lösbar.*

Beweis: Zur Eindeutigkeit: Es seien $p, \tilde{p} \in \mathcal{P}_n$ mit $p(x_i) = \tilde{p}(x_i) = y_i$. Dann gilt für $q = p - \tilde{p} \in \mathcal{P}_n$ für alle $i = 0, \ldots, n$: $q(x_i) = 0$. Damit hat g mindestens $n+1$ Nullstellen. Mit dem Satz von Rolle ergibt sich für die i-te Ableitung von q: $q^{(i)}$ hat $n-i$ Nullstellen. Insbesondere hat dann aber auch $q^{(n)}$ eine Nullstelle, also $q^{(n)} = 0$. Zur Existenz: Für festes $x_0, \ldots, x_n$ ist die Abbildung $A\colon \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$,

$$(a_0, \ldots, a_n) \mapsto (y_0, \ldots, y_n)$$

mit $y_j = a_0 + \cdots + a_n x_j^n$ linear und wegen Eindeutigkeit injektiv. Somit ist A auch surjektiv und es existiert eine Lösung. ■

2.1.2 Lagrangesche Basispolynome

Wir suchen nun Polynome, die an vorgegebenen Knoten x_j den Wert 1 und sonst den Wert 0 haben. Hierzu sei

$$L_i^n(x) = \prod_{j=0, j\neq i}^{n} \frac{x - x_j}{x_i - x_j} \in \mathcal{P}_n,\ i = 0, \ldots, n$$

Dann gilt

$$L_i^n(x_k) = \delta_{ik}$$

mit dem Kronecker-Delta δ_{ik}. Diese sog. *Lagrangeschen Basispolynome* sind also tatsächlich linear unabhängig.

2.1.3 Lagrangesche Darstellung

Das Polynom

$$p(x) = \sum_{i=0}^{n} y_i L_i^n(x)$$

erfüllt $p(x_k) = y_k$ für alle $k = 0, \ldots, n$. Dennoch haben wir ein *Problem*: Nehmen wir einen Punkt hinzu, so ändern sich alle Koeffizienten.

Die *Lösung* sind die *Newtonschen Basispolynome*:

$$N_0(x) = 1,\ N_i(x) = \prod_{j=0}^{i-1} (x - x_j)$$

Sie sind linear unabhängig, da sie verschiedenen Grad haben. Wir machen den folgenden Ansatz:

$$p(x) = \sum_{i=0}^{n} a_i N_i(x)$$

$$\Rightarrow\ y_0 = p(x_0) = a_0$$

$$\Rightarrow\ y_1 = p(x_1) = a_0 + a_1(x_1 - x_0)$$

$$\vdots$$

$$y_n = p(x_n) = a_0 + a_1(x_1 - x_0) + \cdots + a_n(x_n - x_0)\dots(x_n - x_{n+1})$$

Hierdurch lassen sich die a_n bestimmen. Besser (d.h. numerisch stabiler) sind allerdings die *dividierten Differenzen*:

Theorem 2.2

Das Lagrangesche Interpolationspolynom *ist gegeben durch*

$$p(x) = \sum_{i=1}^{n} y[x_0, \dots, x_n] N_i(x)$$

mit den rekursiv definierten dividierten Differenzen

$$y[x_i] = y_i,\ y[x_i, \dots, x_{i+k}] = \frac{y[x_{i+1}, \dots, y_{i+k}] - y[x_i, \dots, x_{i+k-1}]}{x_{i+k} - x_i}$$

Beweis: Sei $p_{i,i+k}$ mit $i \le k$ das Lagrangesche Interpolationspolynom zu $(x_i, y_i), \dots, (x_k, y_k)$. Damit gilt also $p = p_{0,n}$. Wir machen nun eine Induktion nach k:

Induktionsanfang: $k = 0$: $p_{i,i} = y_i = y[x_i]$.

Induktionsschritt: $(k-1) \to k$:

$$p_{i,i+k}(x) = p_{i,i+k-1}(x) + a(x - x_i)\dots(x - x_{i+k-1})$$

und

$$p_{i,i+k}(x) = \frac{(x - x_i)p_{i+1,i+k-1}(x) - (x - x_i)p_{i,k-1}(x)}{x_{i+k} - x_i}$$

Bestimme nun a durch Koeffizientenvergleich des Monoms x^n:

$$p_{i+1,i+k}(x) \underbrace{=}_{IV} y[x_{i+1}, \dots, x_{i+k}]x^{k-1} + O(x^{n-1})$$

$$p_{i,i+k-1}(x) \underbrace{=}_{IV} y[x_i, \dots, x_{i+k-1}]x^{k-1} + O(x^{n-1})$$

$$\Rightarrow\ a = \frac{y[x_{i+1}, \dots, y_{i+k}] - y[x_i, \dots, x_{i+k-1}]}{x_{i+k} - x_i} = y[x_i, \dots, x_{i+k}]$$

■

Folgerung 2.3

Für jede Permutation $\sigma \in S_n$ gilt:

$$y[x_0, \dots, x_n] = y[x_{\sigma(0)}, \dots, x_{\sigma(n)}]$$

Beweis: p ist unabhängig von σ. Daher sind die Koeffizienten von x^n unabhängig. Es folgt: $y[x_0, \dots, x_n]$ sind unabhängig. ■

2.1.4 Interpolation von Funktionen

Gegeben seien $f(x_i), i = 0, \dots, N$ mit $x_0, \dots, x_N$ paarweise verschieden. Wir erhalten dazu das Interpolationspolynom

$$p(x) = \sum_{i=0}^{N} f[x_0, \dots, x_i](x - x_0) \dots (x - x_{i-1}) = N_i(x)$$

wobei wir jetzt $f[x_0, \dots, x_i]$ statt $y[x_0, \dots, x_i]$ schreiben.

Theorem 2.4

Sei $f \in C[a, b]$, so gilt

$$f(x) - p(x) = f[x_0, \dots, x_n, x] \underbrace{\prod_{j=0}^{n} (x - x_j)}_{=N_n(x)}$$

(zunächst) mit $x \in [a, b] \setminus \{x_0, \dots, x_n\}$.

Beweis : Die Polynome vom Grad $n + 1$

$$\alpha(t) = f(x),\ \beta(t) = p(t) + f[x_0, \dots, x_n, x]N_n(t)$$

stimmen auf den $N + 2$ Stellen $x_0, \dots, x_n, x$ überein. Damit gilt $\alpha = \beta$, da $p(t) + f[x_0, \dots, x_n, x]N_n(t)$ gleich $f(x)$ bei $t = x$ ist. ■

Theorem 2.5

Sei $f \in C^{n+1}[a, b]$ und $x \in [a, b] \setminus \{x_0, \dots, x_n\}$, so gilt (zunächst für paarweise verschiedene $x_0, \dots, x_n, x$)

$$f[x_0, \dots, x_n, x] = \int_0^1 \int_0^{t_n} \dots \int_0^{t_{n-1}} f(t_0 x_0 + \dots + t_n x_n + (1 - t_0 - \dots - t_n)x) dt_n \dots dt_0$$

Beweis : Der Beweis ist eine Übungsaufgabe und kann etwa mit vollständiger Induktion geführt werden. ■

Die Formel aus Theorem 2.5 erlaubt es, $f[x_0, \dots, x_n, x]$ auch für *nicht paarweise verschiedene* $x_0, \dots, x_n, x$ zu definieren. Insbesondere gilt

$$f[\underbrace{x_0, \dots, x_0}_{\text{(k+1)-mal}}] = \frac{f^{(k)}(x)}{k!}$$

Allgemein gilt

$$f[x_0, \dots, x_k] = \frac{f^{(k)}(\zeta)}{k!}$$

für ein $\zeta \in \text{conv}(\{x_0, \dots, x_n\})$.

Die Formel

$$f(x) = p(x) + f[x_0, \dots, x_n, x]N_n(x) = \sum_{i=0}^{n} f[x_0, \dots, x_i]N_i(x) + f[x_0, \dots, x_n, x]N_n(x)$$

bleibt aus Stetigkeitsgründen korrekt.

Folgerung 2.6

Sei $f \in C^{n+1}[a,b]$, so gilt

$$f(x) - p(x) = \frac{f^{(n+1)}(\zeta)}{(n+1)!} N_n(x)$$

für ein $\zeta \in \text{conv}(x_0, \ldots, x_n)$.

Beweis : Die Behauptung folgt sofort aus dem Mittelwertsatz der Integralrechnung und

$$\int_0^1 \int_0^{t_n} \ldots \int_0^{t_{n-1}} 1 \, dt_n \ldots dt_0 = \frac{1}{(n+1)!}$$

■

Also gilt für den Interpolationsfehler auf $[a,b]$:

$$|f(x) - p(x)| \leq \frac{1}{(n+1)!} \max_{\zeta \in [a,b]} |f^{(n+1)}(\zeta)|$$

Der erste Faktor auf der rechten Seite deutet auf eine gute Fehlerkontrolle hin, doch haben wir keine Garantie, dass der zweite Faktor nicht zu groß wird. Hierzu ein

Beispiel: Sei

$$f(x) = \frac{1}{1+x^2} \in C^\infty(\mathbb{R})$$

Dann ist

$$f'(x) = \frac{2x}{(1+x^2)^2}$$

$$\Rightarrow \; |f^{(n)}(x) \approx 2^n n! \, O(|x|^{-2-n})$$

für große x. Zwar kürzt sich der Faktor $n!$ bei der Abchätzung

$$|f(x) - p(x)| \leq \frac{1}{(n+1)!} \max_{\zeta \in [a,b]} |f^{(n+1)}(\zeta)|$$

heraus, doch der Faktor 2^n bleibt. Deshalb haben wir hier keine gleichmäßige Konvergenz (am Rand schlechter).

Bemerkung: Der *Weierstraßsche Approximationssatz* besagt, dass jedes $f \in C[a,b]$ gleichmäßig durch Polynome approximiert werden kann. Die Vermutung, dass dies Lagrangesche Interpolationspolynome sind, ist jedoch *falsch.*

Beispiele:

- $f(x) = |x|, \; x \in [-1,1]$. Hier bleibt der Fehler an den Rändern groß.
- $\exp(x), \cos(x), \sin(x)$ sind in Ordnung, da die Ableitung durch eine von n unabhängige Konstante K abgeschätzt werden kann:

$$\max_{a \leq \zeta \leq b} \|f^{(n)}(\zeta)\| \leq K$$

Bemerkung: *Richardsonsche Extrapolation.* Hier wollen wir einen Grenzwert $a(h), h \to 0$ berechnen, bei dem $a(0)$ nicht direkt berechenbar ist, aber die Entwicklung

$$a(x) = a_0 + \sum_{j=1}^{n} a_j x^j + a_{n+1}(h) h^{n+1} + O(h^{n+1})$$

Sei dann p_n das Interpolationspolynom für die Stützstellen $\rho^k, \ldots, \rho^{k+n}$ mit $\rho < 1$. Dann gilt (ohne Beweis)

$$a(0) - p_n(0) = O(h^{k(n+1)})$$

2.1.5 Spline-Interpolation

Lagrangesche Interpolationspolynome sind insbesondere schlecht für nicht glatte Funktionen. Dies liegt am geforderten C^∞-Übergang an den Knoten. Bei *Splines* fordert man weniger. Wir betrachten die Vektorräume

$$S_n^{(k,r)}([a,b]) = \{p \in C^r([a,b]) \,:\, p|_{I_i} \in \mathcal{P}_k(I_i)\}$$

wobei wir wieder von den Stützstellen $a = x_0, \ldots, x_n$ ausgehen mit $I_i = [x_i - 10x_i]$, $i = 1, \ldots, n$. Die Funktionen sind also lokal Polynome vom Grad k und global r-mal stetig differenzierbar.

Beispiel: Kubischer Spline.

Definition 2.7

$s_n \in S_n^{(3,2)}$ heißt kubischer Spline *der Zerlegung $a = x_0 < \cdots < x_n = b$. falls $s''(a) = s''(b) = 0$, so heißt s* natürlich.

Theorem 2.8

Zu (x_i, y_i) mit paarweise verschiedenen $x_0, \ldots, x_n$ existieren Splines s_n mit

$$s_n(x_i) = y_i$$

Unter zusätzlicher Vorgabe von ${s_n}''(a), {s_n}''(b)$ ist s_n eindeutig.

Beweis : Die Abbildung

$$\text{Koeff}(s_n) \mapsto (y_0, \ldots, y_n)$$

ist linear. Wir zeigen nun, dass die Abbildung Koeff auch injektiv ist. Seien hierzu $s_n^{(1)}, s_n^{(2)}$ Splines und $s = s_n^{(1)} - s_n^{(2)}$. Dann ist $s \in S_h^{(3,2)}$ und $s(x_i) = y_i$ sowie $s''(a) = s''(b) = 0$. Für $w \in C^2[a,b]$ mit $w(x_i) = 0$ gilt:

$$\int_a^b s''w''dx = \sum_i \int_{x_i}^{x_{i+1}} s''w''dx =$$

$$\sum_i (-\int_{x_i}^{x_{i+1}} s^{(3)}w'dx + [s''w']_{x_i}^{x_{i+1}}) =$$

$$\sum_i (\int_{x_i}^{x_{i+1}} s^{(3)}w'dx) + [s''w']_{x_0}^{x_n} =$$

$$\sum_i (\int_{x_i}^{x_{i+1}} s^{(4)}wdx - [s^{(3)}w]_{x_i}^{x_{i+1}}) = 0$$

wobei der letzte Schritt analog zum Vorherigen durchgeführt wird. Sei nun $w = s$. Dann erhalten wir

$$\int_a^b |s''|^2 dx = 0 \Rightarrow s'' = 0$$

Somit ist s linear und da s an den Punkten a und b gleich 0 ist, folgt $s \equiv 0$. Damit haben ist Koeff injektiv und der Spline eindeutig. Ferner ist Koeff auch linear und eindlichdimensional, damit auch surjektiv. Es folgt die Lösbarkeit. ■

Theorem 2.9

Die interpolierenden natürlichen Splines minimieren das Integral

$$\mathcal{J}(f) = \int_0^\infty |f''(x)|^2 dx$$

bezüglich $f \in C^2[a,b], f(x_i) = y_i, i = 0, \ldots, n.$

Beweis: Sei $N = \{w \in C^2[a,b] \,:\, w(x_i) = 0, i = 0, \ldots, n\}$. Dann gilt nach obigem Beweis:

$$\int_a^b s'' w'' dx = 0$$

d.h.$(\delta\mathcal{J})(s_n)(w) = 0$. Somit folgt

$$\mathcal{J}(f) = \mathcal{J}(s_n + \tilde{w}) \text{ für ein} \tilde{w} \in \mathbb{N}$$

$$= \mathcal{J}(s_n) + 2\int_a^b s''\tilde{w}'' dx + \mathcal{J}(\tilde{w}) \geq \mathcal{J}(s_n)$$

■

Explizite Berechnung: Auf den Intervallen $[x_i, x_{i+1}]$ gilt

$$s_n|_{[x_{i-1},x_i]} = p_i(x) = a_1^{(i)}(x - x_i) + a_2^{(i)}(x - x_i)^2 + a_3^{(i)}(x - x_i)^3$$

Wir erhalten somit

(a)

$$p_j(x_j) = y_j,\ p_i(x_{i-1}) = y_{i-1} \ \Rightarrow$$

$$a_0^{(i)} = y_i$$

$$y_{i-1} - y_i = -a_1^{(i)} h_i + a_2^{(i)} h_i^2 - a_3^{(i)} h_i^3$$

(b) Die Randbedingung $p_1''(x_0) = p_n''(x_n) = 0$ ergibt

$$a_2^{(1)} - 3a_3^{(1)} h_1 = 0,\ a_2^{(1)} = 0$$

(c) Die C^1-Forderung $p_i'(x_i) = p_{i+1}'(x_i)$ ergibt

$$a_1^{(i)} = a_1^{(i+1)} - 2a_2^{(i+1)} h_{i+1} + a_3^{(i+1)} h_{i+1}^2, i = 1, \ldots, n-1$$

(d) Die C^2-Forderung $p_i''(x_i) = p_{i+1}''(x_i)$ ergibt

$$a_2^{(i)} = a_2^{(i+1)} - 3a_3^{(i+1)} h_{i+1}$$

Wir erhalten insgesamt $4n$ Gleichungen. Drücke nun $a_3^{(i)}$ durch $a_2^{(i)}$ aus sowie $a_1^{(i)}$ durch $a_2^{(i)}$:

$$a_3^{(i)} = \frac{a_2^{(i)} - a_2^{(i+1)}}{3h_i},\ i = 1, \ldots, n$$

$$a_1^{(i)} = \frac{y_i - y_{i-1}}{h_i} + \frac{h_i}{3}(2a_2^{(i)} + a_2^{(i-1)})$$

Das führt zu folgender Matrix

$$A = \begin{pmatrix} 2(h_1+h_2) & h_2 & 0 & & & \\ h_2 & 2(h_2+h_3) & h_3 & & & \\ 0 & h_3 & & & & \\ & & & \ddots & & \\ & & \ddots & \ddots & h_{n-1} & \\ & & & h_{n-1} & 2(h_{n-1}+h_n) & \end{pmatrix}$$

Wir stellen fest, dass die Matrix A symmetrisch und strikt diagonal dominant ist. Daraus folgern wir, dass A auch invertierbar ist, denn

$$A = D + N = D(I + D^{-1}N)$$

wobei D die Diagonalmatrix von A ist. Da $D^{-1}N$ auch symmetrisch ist, stimmen auch dessen maximale Zeilensumme und maximale Spaltensumme überein. Diese sind dann < 1, weil A diagonaldominant ist. Also

$$\|D^{-1}N\|_\infty = \|D^{-1}\|_1 < 1$$

woraus die Invertierbarkeit von A folgt.

Theorem 2.10

Sei $f \in C^4([a,b])$, so erfüllt der kubische Spline mit

$$s_n''(a) = f''(a),\ s_n''(b) = f''(b)$$

die Abschätzung

$$\|f(x) - s_n(x)\|_\infty \leq \frac{1}{2}h^4 \|f^{(4)}(s)\|_\infty$$

Beweis: Wird in der Übung besprochen. ■

Bemerkung: Splines approximieren besser. Es gilt sogar $\|s_n - f\| \to 0$ für $f \in W^{1,2}((a,b))$, wobei dies der Raum der absolut stetigen Funktionen ist mit $\int |f'|^2 dx < \infty$.

2.1.6 Trigonometrische Interpolation

Diese Methode wird vor allem für periodische Funktionen $f(x) = f(x+\omega)$ angewendet. Interpoliert wird durch Linearkombinationen von $1, \cos(xk\frac{2\pi}{\omega})$ und $\sin(xk\frac{2\pi}{\omega})$. Wir nehmen folgende Vereinfachungen vor:

- $\omega = 2\pi$
- komplexe Schreibweise

Wir interpolieren also $f(x + 2\pi) = f(x)$ durch Linearkombinationen von e^{ikx}. In diesem Abschnitt sei also $i = \sqrt{-1}, k \in \mathbb{N}_0$, insbesondere i also *kein* Index. Sei entsprechend

$$\tau_{\mathbb{C}}^n = \left\{ \varphi(t) = \sum_{j=0}^{n} c_j e^{ijt} \ : \ c_j \in \mathbb{C}, j = 0, \ldots, n \right\}$$

der $n+1$-dimensionale Vektorraum dazu.

Theorem 2.11

Zu $y_0, \ldots, y_n \in \mathbb{C}$ existiert genau ein $\varphi \in \tau_{\mathbb{C}}^n$ mit

$$\varphi(x_j) = y_j,\ j = 0, \ldots, n$$

wobei $x_j = j\frac{2\pi}{N+1}$. Es gilt $\varphi(x) = \sum_{j=0}^{n} e^{ijx}$ mit

$$c_k = \frac{1}{n+1} \sum_{j=0}^{n} e^{-ijx_k} y_j,\ k = 0, \ldots, n$$

Beweis : Wir müssen das Gleichungssystem

$$c_n \omega_j^n + \cdots + c_1 \omega_j^1 + c_0 = y_j,\ j = 0, \ldots, n$$

mit $\omega_j = e^{ix_j} = e^{\frac{j2\pi \cdot i}{n+1}}$ lösen. In Matrixform heißt das

$$\begin{pmatrix} 1 & \omega_0 & \ldots & \omega_0^n \\ \vdots & & \vdots & \\ 1 & \omega_n & \ldots & \omega_n^n \end{pmatrix} \cdot \begin{pmatrix} c_0 \\ \vdots \\ c_n \end{pmatrix} = \begin{pmatrix} y_0 \\ \vdots \\ y_n \end{pmatrix}$$

Die Matrix ist eine Vandermonde-Matrix V und wir schreiben kurz $Vc = y$. Da die ω_j paarweise verschieden sind, ist V auch invertierbar. (Wir erinnern uns, dass für die Vandermonde-Matrix gilt: $\det(v) = \prod_{0 \le i,j \le n}(\omega_i - \omega_j)$) Die Existenz und Eindeutigkeit ist damit gezeigt. Für die ω_j gilt $|\omega_j| = 1$, $\omega_j = \omega_1^j, \overline{\omega_j} = \omega_j^{-1}$. Unter Verwendung von $\sum_{j=0}^{n} \omega_1^k = 0$ erhält man

$$V^*V = (n+1)I$$

Die Matrix $(n+1)^{-1}V^*$ ist also die Inverse zu V (Übung) und wir erhalten aus $Vc = y$ die Beziehung

$$c = \frac{1}{n+1} V^* y$$

mit $(V^*)_{jk} = \overline{V_{kj}} = \overline{\omega_k^j} = e^{-ijx_k}$, woraus sich die Koeffizienten c_k wie gewünscht ergeben. ■

Kapitel 3

Numerische Integration

In diesem Kapitel widmen wir uns der Berechnung von Integralen $\int_a^b f(x)\,dx$. Viele Funktionen f besitzen keine explizit darstellbare Stammfunktion. In der Praxis müssen die meisten Integrale daher durch endliche Summen approximiert werden. Wir kennen bereits die Approximation durch Rechtecke mit Obersummen bzw. Untersummen sowie die Rechteckregel:

$$\int_a^b f(x)\,dx \approx \sum_{i=0}^{n-1} (x_{i+1} - x_i)\, f(x_i) \tag{3.1}$$

Allerdings gibt es bessere Verfahren, Integrale zu approximieren. Diese wollen wir nun besprechen.

3.1 Interpolatorische Quadraturformeln

Wir approximieren die Funktion f durch ein Interpolationspolynom

$$p_n(x) = \sum_{i=0}^{n} f(x_i) L_i^n(x) \tag{3.2}$$

und erhalten

$$I^{(n)}(f) = \int_a^b p_n(x)\,dx = \sum_{i=0}^{n} f(x_i) \underbrace{\int_a^b L_i^n(x)\,dx}_{=\alpha_i} \tag{3.3}$$

Hierbei notiert $L_i^{(n)}$ das *Lagrangesche Interpolationspolynom*. Insbesondere hängen die α_i nur von den $x_0, \cdots, x_n$ ab. Wir präzisieren dies nun:

Theorem 3.4

Sei

$$I(f) - I^{(n)}(f) = \int_a^b f[x_0, \cdots, x_n, x] \prod_{j=0}^{n} (x - x_j)\,dx \tag{3.5}$$

Dann gilt die folgende Abschätzung:

$$|I(f) - I^{(n)}(f)| \leq \frac{\|f^{(n+1)}\|_\infty}{(n+1)!} \int_a^b |\prod_{j=0}^{n} (x - x_j)|\,dx \tag{3.6}$$

Bemerkung: Für ein Polynom $f \in \mathcal{P}^n$ ist die Interpolation sogar exakt.

Definition 3.7

Eine Quadraturformel *heißt (mindestens) von der Ordnung n, falls sie alle Polynome $f \in \mathcal{P}^n$ exakt integriert.*

Die $I^{(n)}$ sind also von der Ordnung $n+1$.

3.1.1 Abgeschlossene Newton-Cotes-Formeln

Wir wählen nun äquidistante Stützstellen $a = x_0, \cdots, x_n = b$, d.h.

$$x_i = a + i\frac{b-a}{n},\ i = 0, \cdots, n$$

Mit der Variablentransformation

$$t = \frac{x-a}{b-a} \cdots n$$

erhalten wir mit $t \in [0, n]$ und der Schrittweite $H = \frac{b-a}{n}$

$$x = a + t \cdot H$$

Die Polynome $L_i^{(n)}$ lauten damit

$$L_i^{(n)}(x) = \prod_{j=0, j\neq i}^{n} \frac{x - x_j}{x_i - x_j} = \prod_{j=0, j\neq i}^{n} \frac{a + tH - ajH}{a + iH - a - jH} = \prod_{j=0, j\neq i}^{n} \frac{t-j}{i-j} \tag{3.8}$$

Damit erhalten wir

$$\alpha_i = \int_a^b L_i^{(n)}(x)\, dx = H \cdot \int_0^n \prod_{j=0}^{n} \frac{t-j}{i-j} dt \tag{3.9}$$

Für $n = 1, 2, 3$ erhalten wir so die wichtigen folgenden Regeln:

Name	Formel
Mittelpunktregel	$I^{(1)}(f) = \frac{b-a}{2}(f(a) + f(b))$
Simpsonregel	$I^{(2)}(f) = \frac{b-a}{6}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right)$
$\frac{3}{8}$-Regel	$I^{(3)}(f) = \frac{b-a}{8}(f(a) + 3f(a+H) + 3f(b-H) + f(b))$

3.1.2 Offene Newton-Cotes-Formeln

Im Folgenden sei $x_0 = a + H,\ x_i = x_0 + iH$ und $b = x_n + H$. Wir betrachten also $n+2$ Intervalle der Länge H, wobei a und b *keine* Stützstellen sind.

$$\begin{array}{l} \hat{I}^{(0)}(f) = (b-a) f\left(\frac{a+b}{2}\right) \text{ (Mittelpunktsregel)} \\ \hat{I}^{(1)}(f) = \frac{b-a}{2}(f(a+H) + f(b-H)) \\ \hat{I}^{(2)} = \frac{b-a}{3}\left(2f(a+H) - f\left(\frac{a+b}{2}\right) + f(b-H)\right) \end{array}$$

Wir wollen Im Fogenden die *Trapezregel, Simpsonregel* und die *Mittelpunktsregel* untersuchen.

Theorem 3.10

Für die Restglieder der Trapezregel, Simpsonregel *und der* Mittelpunktsregel *gelten die folgenden Beziehungen:*

- *Mittelpunktsregel:*

$$I(f) - (b-a) f\left(\frac{a+b}{2}\right) = \frac{(b-a)^3}{24} f''(\zeta) \text{ für } f \in C^2[a,b] \tag{3.11}$$

- *Trapezregel:*

$$I(f) = I(f) - \frac{a+b}{2}(f(a) + f(b)) = -\frac{(b-a)^3}{12} f''(\zeta) \text{ für } f \in C^2[a,b] \tag{3.12}$$

- *Simpsonregel:*

$$I(f) - \frac{b-a}{6}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right) = -\frac{(b-a)^5}{2880} f^{(4)}(\zeta) \text{ für } f \in C^4[a,b] \tag{3.13}$$

für gewisse Zwischenstellen $\zeta \in [a,b]$.

BEWEIS: Für die *Trapezregel* erhalten wir beispielsweise durch Anwendung der Zwischnewertsätze

$$I(f) - I^{(1)}(f) = \int_a^b f[a,b,x](x-a)(x-b)\,dx =$$

$$f[a,b\zeta_1] \int_a^b (x-a)(x-b)\,dx =$$

$$\frac{1}{2} f''(\zeta_2) \cdot \frac{-(b-a)^3}{12}$$

Dabei notieren $f[a,b,x]$ die dividierten Differenzen. In der letzten Zeilen verwendeten wir den Zwischensatz der Differentialrechnung, hingegen in der vorherigen den Zwischenwertsatz der Integralrechnung:

Sei f stetig, g integrierbar und $g \geq 0$(bzw. $g \leq 0$). Dann existiert ein $\zeta \in [a,b]$ mit

$$\int_a^b f(x) g(x)\,dx = f(\zeta) \int_a^b g(x)\,dx$$

Man beachte, dass hier die Voraussetzungen erfüllt sind, da $(x-a)(x-b) \leq 0$ gilt. Bei der *Simpsonregel* können wir den Zwischenwertsatz nicht direkt anwenden, da g das Vorzeichen wechselt:

$$I(f) - I^{(2)}(f) = \int_a^b f\left[a, \frac{a+b}{2}, b, x\right](x-a)\left(x - \frac{a+b}{2}\right)(x-b)\,dx$$

Um dennoch den Zwischenwertsatz anwenden zu können, schreiben wir

$$I(f) - I^{(2)}(f) = \int_a^b \frac{f\left[a, \frac{a+b}{2}, b, x\right] - f\left[a, \frac{a+b}{2}, b, \frac{a+b}{2}\right]}{x - \frac{a+b}{2}}(x-a)\left(x - \frac{a+b}{2}\right)^2 (x-b)\,dx -$$

$$f\left[a, \frac{a+b}{2}, b, \frac{a+b}{2}\right] \cdot \int_a^b f\,(x-a)\left(x-\frac{a+b}{2}\right)(x-b)\,dx$$

Beachte nun, dass $(x-a)\left(x-\frac{a+b}{2}\right)^2(x-b) \geq 0$ und $\int_a^b f\,(x-a)\left(x-\frac{a+b}{2}\right)(x-b)\,dx$ punktsymmetrisch bzgl. $\frac{a+b}{2}$ ist. Damit gilt

$$I\,(f) - I^{(2)}\,(f) = \int_a^b \frac{f\left[a, \frac{a+b}{2}, b, x\right] - f\left[a, \frac{a+b}{2}, b\frac{a+b}{2}\right]}{x - \frac{a+b}{2}} \underbrace{(x-a)\left(x-\frac{a+b}{2}\right)^2(x-b)}_{\leq 0}\,dx \underbrace{=}_{ZWS}$$

$$f\left[a, \frac{a+b}{2}, \frac{a+b}{2}, b, \zeta_1\right] \cdot \int_a^b (x-a)\left(x-\frac{a+b}{2}\right)^2(x-b)\,dx =$$

$$= \frac{1}{4!} f^{(4)}\,(\zeta_2)\,\frac{1}{120}\,(b-a)^5$$

Die Mittelpunktsregel wird in den Übungen behandelt. □
BEMERKUNG:

- Bei $I^{(n)}$ mit $n \geq 7$ und $\hat{I}^{(n)}$ mit $n \geq 2$ treten negative Gewichte auf. Dadurch kann es möglicherweise zur Auslöschung kommen.
- Im Allgemeinen gilt *nicht* $I^{(n)}\,(f) \to I\,(f)$ für $n \to \infty$, da die Lagrange-Interpolation im Allgemeinen kein konvergenter Prozess ist. Man wendet daher $I^{(n)}$ nur auf Teilintervalle an. Das heißt

$$I_N^{(n)}\,(f) = \sum_{i=0}^{N-1} I_{[x_i, x_{i+1}]}^{(n)}\,(f) \text{ für } h = \frac{b-a}{N} \tag{3.14}$$

$$\Rightarrow I\,(f) - I_N^{(n)}\,(f) = \sum_{i=0}^{N-1} \omega_n h^{m+2} f^{(m+1)}\,(\zeta_i) =$$

$$= \omega_n h^{m+2} N f^{(m+1)}\,(\zeta) = \omega_n\,(a-b)\,h^{m+1} f^{(m+1)}\,(\zeta)$$

Dabei sei ω_n ein Faktor für ein $m \geq n$. Die letzte Zeile stellt dann eine Fehlerformel dar.

3.1.3 Gaußsche Quadraturformeln

$I^{(2)}$ integriert nicht nur $p \in \mathcal{P}^2$, sondern auch $p \in \mathcal{P}^4$ exakt. Kann die Wahl der $x_0, \cdots, x_n$ noch verbessert werden, dass sogar $p \in \mathcal{P}^k$ mit $k \geq 5$ exakt integriert werden?
BEMERKUNG: Der Fehler ist für n gerade $I - I^{(n)}\,(f) = O\left((b-a)^{n+1}\right)$ für n ungerade dagegen $O\left((b-a)^{n+1}\right)$.

Lemma 3.15

$I^{(n)}$ kann für $q \in \mathcal{P}^k$ mit $k \geq 2n+2$ nicht exakt sein.

BEWEIS: Sei

$$p(x) = \prod_{i=0}^{n} (x - x - i)^2 \in \mathcal{P}^{2n+2} \Rightarrow 0 < \int_a^b p(x)\, dx = I^{(n)}(p) = 0$$

Das ist ein Widerspruch und somit kann $I(f) - I^{(n)}(f)$ höchstens von der Ordnung $(a-b)^{n+2} cf^{(n+2)}(\zeta)$ sein. □

Wir wollen nun eine geschickte Wahl der $x_0, \cdots, x_n$ herleiten. Hierzu ergänzen wir $x_0, \cdots, x_n$ durch $x_{n+1}, \cdots, x_{2n+1}$. Dann gilt

$$I(f) - I^{(2n+1)}(f) = I(f) - \sum_{i=0}^{2n+1} f[x_0, \cdots, x_n] \int_a^b \prod_{j=0}^{i-1} (x - x_j)\, dx =$$

$$I(f) - I^{(n)}(f) - \sum_{i=n+1}^{2n+1} f[x_0, \cdots, x_i] \int_a^b \prod_{j=0}^{i-1} (x - x_j)\, dx$$

Ziel ist es nun, den letzten Faktor

$$\int_a^b \prod_{j=0}^{i-1} (x - x_j)\, dx = 0$$

zu setzen, denn dann wäre $I^{(n)}(f)$ so genau wie $I^{(2n+1)}(f)$. Für $i = n+1, \cdots, 2n+1$ schreiben wir

$$\int_a^b \prod_{j=0}^{i-1} (x - x_j)\, dx = \int_a^b \prod_{j=0}^{n} (x - x_j)\, q_i(x)\, dx = 0 \tag{3.16}$$

für ein Polynom $q_i \in \mathcal{P}^{i-n-1} \subset \mathcal{P}^n$. Es soll also für alle Polynome $q \in \mathcal{P}^n$ gelten:

$$\int_a^b \prod_{j=0}^{n} (x - x_j)\, q(x)\, dx = 0 \tag{3.17}$$

oder äquivalent dazu

$$\left\langle \prod_{j=0}^{n} (x - x_j), q \right\rangle = 0 \Leftrightarrow \prod_{j=0}^{n} (x - x_j) \perp \mathcal{P}^n \tag{3.18}$$

Da $\prod_{j=0}^{n} (x - x_j) \in \mathcal{P}^{n+1}$, wäre für die Lösung der Gleichung z.B. das Legendrepolynom L_{n+1} ein Kandidat. Mit dem folgenden Lemma zeigen wir, dass es tatsächlich eine Lösung ist:

Lemma 3.19

$L_m(x)$ hat paarweise verschiedene Nullstellen.

BEWEIS: Sei $N_m = \{\lambda \in (a,b) : \lambda \text{ ist Nullstelle ungerader Ordnung von } L_m\}$, d.h. N_m ist die Menge der Nullstellen, bei denen L_m das Vorzeichen wechselt. Wir behaupten nun $|N_m| = m$. Hierzu ein Widerspruchsbeweis. Wir nehmen an, $|N_m| < m$. Dann liegt $q(x) = \prod_{\lambda \in N_m} (x - \lambda)$ in $\mathcal{P}^m$ und $\langle q, L_m \rangle = 0$. Aber qL_m hat nur Nullstellen gerader Ordnung und damit *keinen* Vorzeichenwechsel, also wäre

$$\int_a^b q(x)\, L_m(x)\, dx \neq 0$$

Und dies ist ein Widerspruch. □.
Wählen wir also $x_0, \cdots, x_n$ als Nullstellen des Legendre-Polynoms L_{n+1}, so gilt

$$I(f) - I^{(n)}(f) = I(f) - I^{(2n+1)}(f) = c_{2n+1}(b-a)^{n+2} f^{(2n+2)}(\zeta) \tag{3.20}$$

d.h. $I^{(n)}(f)$ auf $\mathcal{P}^{2n+1}$.
<!– Local IspellDict: german-new8 –>

Kapitel 4

Lineare Gleichungssysteme - direkte Verfahren

Das Ziel dieses Kapitels ist es, das Gleichungssystem

$$Ax = b$$

für eine (zunächst) reguläre Matrix A zu lösen.

4.0.4 Eliminationsverfahren

Im einfachsten Fall ist A eine obere Dreiecksmatrix:

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ & \ddots & \vdots \\ 0 & & a_{nn} \end{pmatrix}$$

Dann gilt

$$Ax = b \Rightarrow x_n = \frac{1}{a_{nn}} b_n$$

und

$$x_j = \frac{1}{a_{jj}} \left(b_j - \sum_{k=j+1}^{n} a_{jk} x_k \right), \; j = 1, \cdots, n+1$$

Im allgemeinen Fall benutzt man z.B. das Gaußsche Eliminationsverfahren mit den elementaren Umformungen

(i) Vertauschen zweier Gleichungen

(ii) Addition eines vielfachen einer Zeile (Gleichung) zu einer anderen Zeile

(iii) Multiplizieren einer Zeile mit einer Zahl $\neq 0$.

In der Praxis wendet man das Verfahren auf die Matrix $(A|b)$ oder $(A|E_n)$ an.

Beschreibung des Verfahrens

1.Schritt:

- *(Pivotsuche)* Finde $a_{m1} \neq 0, m \in \{1, \cdots, n\}$ So ein m existiert, da A regulär ist.

- *(Pivotisierung)* Vertausche 1. und m-te Zeile. Das Resultat sei $\tilde{A}(0), \tilde{b}^{(0)}$.
- Ziehe für jedes $r \in \{2, \cdots, n\}$ das $\frac{\tilde{a}_{r1}^{(0)}}{\tilde{a}_{11}^{(0)}}$-fache der 1.Zeile von der r-ten Zeile ab. Das Resultat sei $\tilde{A}^{(1)}, \tilde{b}^{(1)}$.

Dies enstpricht der Multiplikation mit einer Permutationsmatrix P und einer Matrix G,also $\left[A^1|b^1\right] = GP\,[A|b]$. mit

$$P = \begin{pmatrix} 0 & 0 & 0 & & 1 & & \\ 0 & 1 & 0 & & & & \\ 0 & 0 & \ddots & & & & \\ & & & 1 & & & \\ 1 & & & & 0 & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{pmatrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \end{matrix} \text{1. und m-te Spalte} \tag{4.1}$$

$$G = \begin{pmatrix} 1 & & & 0 \\ -q_{21} & 1 & & \\ \vdots & 0 & \ddots & \\ -q_{n1} & & & 1 \end{pmatrix} \tag{4.2}$$

2. bis n-ter Schritt: Wende den ersten Schritt auf die um eine Zeile und eine Spalte verkleinerte Matrix $\left[A^{(1)}|b^{(1)}\right]$ an. Am Ende erhält man

$$[R|c] = G_{n-1}P_{n-1}\cdots G_1P_1\,[A|b] \tag{4.3}$$

mit einer oberen Dreiecksmatrix R.
Bemerkung: Man kann die wesentlichen Elemente der G_j in den frei werdenden Stellen von A speichern.

Definition 4.4

Eine Matrix A heißt nilpotent, *falls es ein $k \in \mathbb{N}$ gibt, so dass $A^n = 0$ gilt. Eine Matrix A heißt* unipotent, *falls $A - I$ nilpotent ist.*

Bemerkung: Obere Dreiecksmatrizen mit allen Diagonalelementen $= 0$ sind nilpotent, obere Dreiecksmatrizen mit Diagonalelementen $= 1$ sind unipotent.

Theorem 4.5

Mit Hilfe des Gaußschen Eliminationsalgortihmus erhält man die Zerlegung

$$PA = LR \tag{4.6}$$

mit der Permutationsmatrix $P = P_{n-1}\cdots P_1$, der unipotenten unteren Dreiecksmatrix L und der oberen Dreiecksmatrix R. Für $P = 1$ ist die Zerlegung eindeutig.

Bemerkung: Die G_j haben die Form

$$\begin{pmatrix} 1 & & 0 & & \\ & \cdots & 0 & & \\ 0 & & 1 & & \\ & & -q_{21} & 1 & \\ & & \vdots & & \ddots \\ 0 & & -q_{n1} & 0 & 1 \end{pmatrix}$$

Theorem 4.7

Sei A regulär und diagonaldominant, d.h.

$$\sum_{k=1,k\neq j}^{n} |a_{jk}| \leq |a_{jj}| \tag{4.8}$$

Dann kann die Gaußelimination ohne Pivotisierung durchgeführt werden.

Für positiv definite Matrizen haben wir also die Zerlegung $A = LR$. Es gilt

$$D = \begin{pmatrix} r_{11} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & r_{nn} \end{pmatrix} \text{ und } \tilde{R} = D^{-1}R$$

Dann ist $\tilde{R}$ unipotent. Da A symmetrisch ist, gilt

$$LR = A = A^T = (LR)^T = \left(LD\tilde{R}\right)^T = \tilde{R}^T \left(DL^T\right)$$

Da die LR-Zerlegung eindeutig ist, gilt

$$L = \tilde{R}^T \text{ bzw. } R = DL^T \Rightarrow A = LR = LDL^T = \tilde{L}\tilde{L}^T \text{ mit}$$

$$\tilde{L} = LD^{1/2} = L \cdot \begin{pmatrix} d_{11}^{1/2} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & d_{nn}^{1/2} \end{pmatrix}$$

Theorem 4.9

Jede symmetrische, positiv definite Matrix A hat eine Cholesky-Zerlegung $A = \tilde{L}\tilde{L}^T$ *mit einer unteren Dreiecksmatrix $\tilde{L}$.*

Bemerkung: Wählt man die Diagonalelemente von $\tilde{L}$ positiv, so ist die Cholesky-Zerlegung eindeutig.

Bandmatrizen

Definition 4.10

Eine Matrix $A = (a_{jk})_{j,k}$ heißt Bandmatrix *vom Typ (m_l, m_r) mit $0 \leq m_l, m_r \leq n-1$, falls*

$$a_{jk} = 0 \text{ für } j + m_r < k < j - m_l \tag{4.11}$$

Das heißt, A ist auf der Hauptdiagonale und höchstens auf $m_r + m_l$ Nebendiagonalen ungleich Null. Die Größe $m = m_r + m_l$ heißt Bandbreite.

Beispiel: Bandmatrizen vom Typ

- $(n-1, 0)$ heißen *untere Dreiecksmatrizen*
- $(0, n-1)$ heißen *obere Dreiecksmatrizen*
- $(1, 1)$ heißen *Trigiagonalmatrizen*

Beispiel: Die $(4,4)$-Bandmatrix

$$A = \begin{pmatrix} B & -I & 0 & 0 \\ -I & B & -I & 0 \\ 0 & -I & B & -I \\ 0 & 0 & -I & B \end{pmatrix} \text{ und } B = \begin{pmatrix} 4 & -1 & 0 & 0 \\ -1 & 4 & -1 & 0 \\ 0 & -1 & 4 & -1 \\ 0 & 0 & -1 & 4 \end{pmatrix}$$

und der 4×4-Einheitsmatrix ist diagonaldominant, aber nicht strikt.

Theorem 4.12

Ist die LR-Zerlegung von A ohne Zeilenvertauschung durchführbar und A vom Typ (m_l, m_r), so sind L und R vom Typ $(m_l, 0)$ und $(0, m_r)$. Der Aufwand der Zerlegung ist

$$\frac{1}{3} n m_l m_r \tag{4.13}$$

BEMERKUNG: Es genügt das Band zu speichern. Dadurh wird das Problem reduziert und man kann im schnellem Speicher rechnen.

Nachiteration

Sei $LR = PA$ eine Zerlegung. Durch Rechenfehler hat man nur $\tilde{L}\tilde{R} = PA$. Dadurch ergibt sich die fehlerhafte Lösung

$$\tilde{L}\tilde{R}x_0 = b$$

Das heißt für den *Defekt* d^0:

$$d^0 = b - Ax_0 \neq 0 \Rightarrow d^0 = A\left(x - x_0\right) \Rightarrow A^{-1}d^0 = x - x_0$$

Man müsste also x_0 um $A^{-1}d$ korrigieren. Stattdessen definiert man die Iteration

$$x^1 = x^0 + \left(\tilde{L}\tilde{R}\right)^{-1} d^0 \text{ und } x^{j+1} = x^j + \left(\tilde{L}\tilde{R}\right)^{-1} d^j$$

Es gilt

$$x^{j+1} = \underbrace{x^j + \left(\tilde{L}\tilde{R}\right)^{-1}\left(b - Ax^j\right)}_{=F(x^j)}$$

Also $F\left(x\right) = x + \left(\tilde{L}\tilde{R}\right)^{-1}\left(b - Ax\right)$. Damit erhalten wir

$$F\left(x\right) - F\left(y\right) = x - y + \left(\tilde{L}\tilde{R}\right)^{-1} A\left(y - x\right) = \left(I - \left(\tilde{L}\tilde{R}\right)^{-1} A\right)\left(x - y\right)$$

Die Iteration konvergiert, falls $\|I - \left(\tilde{L}\tilde{R}\right) A\| < 1$. Dies verbessert die Lösung. Es gilt

$$I - \left(\tilde{L}\tilde{R}\right)^{-1} A = \left(\tilde{L}\tilde{R}\right)^{-1}\left(\tilde{L}\tilde{R} - A\right) \text{ und} \tilde{L}\tilde{R} = A\left(I - A^{-1}\left(A - \tilde{L}\tilde{R}\right)\right)$$

Das heißt

$$\left(\tilde{L}\tilde{R}\right)^{-1} = \left(I - A^{-1}\left(A - \tilde{L}\tilde{R}\right)\right)^{-1} A^{-1} \Rightarrow$$

$$\| \left(\tilde{L}\tilde{R}\right)^{-1} \| \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\| \cdot \|A - \tilde{L}\tilde{R}\|} = \frac{\|A^{-1}\|}{1 - \operatorname{cond}\left(A\right) \frac{\|A - \tilde{L}\tilde{R}\|}{\|A\|}}$$

$$\Rightarrow \|I - \left(\tilde{L}\tilde{R}\right)^{-1} A\| \leq \frac{\operatorname{cond}\left(A\right)}{1 - \operatorname{cond}\left(A\right) \frac{\|A - \tilde{L}\tilde{R}\|}{\|A\|}} \cdot \frac{\|A - \tilde{L}\tilde{R}\|}{\|A\|}$$

BEMERKUNG: Man kann $\tilde{L}\tilde{R}$ mit geringerer Genauigkeit berechnen und dann mit hoher Genauigkeit nachiterieren.

Nichtreguläre Systeme

Sei $A \in \mathbb{R}^{m\times n}$ und $b \in \mathbb{R}^m$. Unser Ziel ist nun, $Ax = b$ zu lösen. Allerdings haben wir das folgende Problem: Falls $\operatorname{rang}(A) < \operatorname{rang}[A|b]$, so ist $Ax = b$ nicht im eigentlichen Sinne lösbar. Stattdessen suchen wir eine Lösung x, die das *Residuum* $\|Ax - b\|_2$ minimiert. Hierzu müssen wir

$$J(x) = \frac{1}{2}\|Ax - b\|_2^2 = \frac{1}{2}\langle Ax - b, Ax - b\rangle$$

minimieren. Es gilt

$$\delta J(x)(y) = \frac{d}{dt}J(x + ty)\,|_{t=0} =$$

$$\frac{d}{dt}\left(\frac{1}{2}\langle Ax - b, Ax - b\rangle + t\langle Ax - b, Ay\rangle + \frac{1}{2}t^2\langle Ay, Ay\rangle\right)|_{t=0} = \langle Ax - b, Ay\rangle$$

Also muss das Minimum $\overline{x}$

$$\langle A\overline{x} - b, Ay\rangle = 0 \forall y$$

erfüllen. Dies bedeutet

$$\left\langle A^T(A\overline{x} - Ab), y\right\rangle = 0 \Rightarrow$$

$$A^TA\overline{x} = A^Tb \tag{4.14}$$

Dies ist die *Normalgleichung.*

Theorem 4.15

Es existiert stets ein $\overline{x} \in \mathbb{R}^n$ welches

$$x \mapsto \|Ax - b\|_2 \tag{4.16}$$

*minimiert. Es gilt $A^TA\overline{x} = A^Tb$ (*Normalgleichung*). Für $\operatorname{rang}(A) = n$ ist $\overline{x}$ eindeutig. Die allgemeinen Lösungen sind gegeben durch $\overline{x} + \operatorname{Kern}(A)$.*

Beweis: Zunächst zur Existenz einer Lösung der Normalengleichung. Es gilt $\mathbb{R}^n = \operatorname{Bild}(A) \oplus \operatorname{Kern}\left(A^T\right)$ (orthogonale Summe). Beachte: $\langle Ax, y\rangle = \left\langle x, A^Ty\right\rangle$ Somit ist die Darstellung $b = \alpha + \beta$ mit $\alpha \in \operatorname{Bild}(A)$ und $\beta \in \operatorname{Kern}\left(A^T\right)$ eindeutig. Da $\alpha \in \operatorname{Bild}(A)$, existiert ein $\overline{x} \in \mathbb{R}^n$ mit $A\overline{x} = \alpha$. Damit folgt

$$A^TA\overline{x} = A^T\alpha = A^T(\alpha + \beta) = A^T\beta$$

D.h., $\overline{x}$ löst die Gleichung. Weiterhin gilt für alle x:

$$\|Ax - b\|_2^2 = \|A\overline{x} - b + A(x - \overline{x})\|_2^2 = \|A\overline{x} - b\|_2^2 + 2\langle A\overline{x} - b, A(x - \overline{x})\rangle + \|A(x - \overline{x})\|_2^2 \geq \|A\overline{x} - b\|_2^2$$

Das heißt, $\overline{x}$ minimiert $x \mapsto \|Ax - b\|$. Umgekehrt wissen wir schon, dass jedes Minimum die Normalengleichung erfüllt. Seien nun $\overline{x}, \hat{x}$ Lösungen, so gilt

$$b = \underbrace{A\overline{x}}_{\in \operatorname{Bild}(A)} + \underbrace{(b - A\overline{x})}_{\in \operatorname{Kern}(A^T)} = \underbrace{A\hat{x}}_{\in \operatorname{Bild}(A)} + \underbrace{(b - A\hat{x})}_{\in \operatorname{Kern}(A^T)}$$

Da aber $\mathbb{R}^n = \operatorname{Bild}(A) \oplus \operatorname{Kern}\left(A^T\right)$, gilt

$$A\overline{x} = A\hat{x} \Rightarrow \hat{x} - \overline{x} \in \operatorname{Kern}(A) \Rightarrow \hat{x} =$$

$$\overline{x} + (\hat{x} - \overline{x}) \in \overline{x} + \operatorname{Kern}\left(A^T\right)$$

Andererseits gilt $\|A(x + y) - b\| = \|Ax - b\|$ für jedes $y \in \operatorname{Kern}(A)$. Damit folgt die Behauptung. □

Gaußsche Ausgleichsrechnung

Gegeben seien Messpunkte $(x_j, y_j)\,,\ j = 1, \cdots, n$. Unser Ziel ist nun: Finde zu gegebenen Funktionen $u_1, \cdots, u_n$, $n < m$ eine Linearkombination

$$u = \sum_{k=1}^{n} c_k u_k$$

derart, dass

$$\Delta_2 = \left(\sum_{j=1}^{m} |u(x_j) - y_j|^2 \right)^{1/2} \tag{4.17}$$

minimiert wird. Setze hierzu

$$c = (c_1, \cdots, c_n)^T\,,\ y = (y_1, \cdots, y_m)^T\,,\ a_k = (u_k(x_1), \cdots, u_k(x_m))\,,\ k = 1, \cdots, n,\ A = [a_1| \cdots |a_n]$$

Dann müssen wir

$$F(x) = \frac{1}{2}\|Ac - y\|_2^2$$

minimieren. ($\Rightarrow A^T Ac = A^T y=$). Wir unterscheiden mehrere Fälle.

- Fall 1: $u_k(x) = x^{k-1},\ k = 1, \cdots, n$. Dann nennen wir u *Ausgleichsparabel*.
- Fall 2: $u_1(x) = 1,\ u_2(x) = x$ (linearer Fall).

x_i	-2	-1	0	1	2
y_i	0.5	0.5	2	3.5	3.5

Dies führt zu einem überbestimmten System

$$\begin{pmatrix} 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix} \cdot \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 0.5 \\ 0.5 \\ 2 \\ 3.5 \\ 3.5 \end{pmatrix}$$

Löse nun

$$\underbrace{\begin{pmatrix} 5 & 0 \\ 0 & 10 \end{pmatrix}}_{=A^T A} \cdot \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 10 \\ 9 \end{pmatrix}$$

Dies ergibt

$$\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 2 \\ 0.9 \end{pmatrix}$$

Somit ist $u(x) = 2 + 0.9x$ die *Ausgleichsgerade*.

BEMERKUNG: Man kann auch nichtlineare Größen durch Transfomrationen approximieren. Betrachte beispielsweise ein physikalisches Gesetz der Form

$$y(x) = \frac{a}{1 + bx}$$

Durch Umformen erhält man $\frac{1}{a} + \frac{b}{a}x = \frac{1}{y(x)}$ (dies ist linear in x mit den neuen Größen $\tilde{a} = \frac{1}{a}, \tilde{b} = \frac{b}{a}$.

Bei der Gaußschen Ausgleichsrechnung ist $\overline{A}^T A = A^* A$ wesentlich. Diese Matrix hat eine spezielle Struktur.

Lemma 4.18

Sei $A \in \mathbb{K}^{m\times n}$. Dann ist A^*A hermitesch *(bzw. symmetrisch) und* positiv semidefinit. *Ist rang$(A) = n$, so ist A^*A* positiv definit.

Beweis: Es gilt

$$\langle A^*A\zeta, \zeta\rangle = \langle A\zeta, A\zeta\rangle = \|A\zeta\|_2^2 \geq 0$$

und insbesondere auch $(A^*A)^* = A^*A$. Ist Rang$(A) = n$ (dies benötigt $m \geq n$!), so ist $A\colon \mathbb{R}^n \to \mathbb{R}^n$ injektiv. Damit folgt aus $Ax = 0$ schon $x = 0$. Also folgt aus $\langle A^*A\zeta, \zeta\rangle = 0 \Rightarrow \|A\zeta\|_2^2 = 0 \Rightarrow \zeta = 0$. Also ist A^*A positiv definit.□

Bemerkung: In der Regel ist A^*A schlecht konditioniert.

Bemerkung: Bei symmetrischen, positiv definiten Matrizen kann man die LR-Zerlegung durch die Cholesky-Zerlegung vereinfachen:

$$\begin{pmatrix} \tilde{l}_{11} & 0 & 0 \\ \vdots & \ddots & 0 \\ \tilde{l}_{n1} & \cdots & \tilde{l}_{nn} \end{pmatrix} \cdot \begin{pmatrix} \tilde{l}_{11} & \cdots & \tilde{l}_{n1} \\ 0 & \ddots & \vdots \\ 0 & 0 & \tilde{l}_{nn} \end{pmatrix} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}$$

Dann gilt

$$l_{11}^2 = a_{11} \Rightarrow l_{11} = \sqrt{a_{11}}$$

$$\tilde{l}_{j1} \cdot \tilde{l}_{11} = a_{j1},\ j \geq 2 \Rightarrow \tilde{l}_{j1} = \frac{\tilde{l}_{11}}{\sqrt{a_{11}}}$$

Weiters folgt

$$\underbrace{\sum_{k=1}^{i} \tilde{l}_{ik}^2}_{\text{bekannt}} = a_{ii} \Rightarrow l_{ii} = \sqrt{a_{ii} - \tilde{l}_{i1}^2 - \cdots - \tilde{l}_{i,i-1}^2}$$

$$\tilde{l}_{ji} = \frac{1}{\tilde{l}_{ii}} a_{ji} - \tilde{l}_{j1}\tilde{l}_{i1} - \cdots - \tilde{l}_{j,i-1}\tilde{l}_{i,i-1},\ i = i+1, \cdots, n$$

Bemerkung: Rundung kann dieses Vefahren jedoch verhindern. Betrachte z.B. bei dreistelliger Rundung

$$A \begin{pmatrix} 1.03 & 1.10 \\ 1.03 & 1.11 \\ 1.07 & 1.15 \end{pmatrix} \Rightarrow A^TA = \begin{pmatrix} 3.43 & 3.60 \\ 3.60 & 3.76 \end{pmatrix}$$

aber

$$\left\langle A^TA \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ 1 \end{pmatrix} \right\rangle = -0.01 < 0$$

Wir wollen deshalb eine neue Methode entwickeln, die ohne die Normalengleichung auskommt.

Theorem 4.19

Sei $A \in \mathbb{K}^{m\times n}$ mit $m \geq n$ und Rang$(A) = n$. Dann existiert eine eindeutige Matrix $Q \in \mathbb{R}^{m\times n}$ mit $Q^*Q = E_n$ und eine eindeutige obere Dreiecksmatrix $R \in \mathbb{R}^{n\times n}$ mit positiven Diagonalementen, so

dass

$$A = QR \tag{4.20}$$

BEWEIS: Zunächst zur Eindeutigkeit: $A = QR = Q_2R_2 \Rightarrow$

$$\underbrace{Q_2^T Q}_{\text{orthogonal}} = R_2R^{-1} = S$$

Zur Orthogonalität von $Q_2^T Q$:

$$\left(Q_2^T Q\right)^T Q_2^T Q = Q^T Q_2 Q_2^T Q = E_n$$

S ist dann eine obere Dreiecksmatrix mit positiven Diagonaleinträgen, denn

$$Q_2^T Q = E_n = R_2R^{-1} \text{ und}$$

$$S = R_2R^{-1} = (s_1|\cdots|s_n) \ , \ \langle s_1, s_1\rangle = 1 \Rightarrow s_{11} = 1 \ , \ \langle s_1, s_j\rangle = 0 \Rightarrow s_{1j} = 0 \text{ für } j \geq 2$$

Wir erhalten $R_2 = R$ und $Q = AR^{-1} = AR_2^{-1} = Q_2$. Jetzt zur Existenz: Wir wenden das *Gram-Schmidt-Verfahren* auf die Spaltenvektoren $a_1, \cdots, a_n$ von A an. Also

$$q_1 = \frac{a_1}{\|a_1\|}$$

$$\tilde{q}_k = a_k - \sum_{i=1}^{k-1} \langle a_k, q_i\rangle \, q_i$$

$$q_k = \frac{\tilde{q}_k}{\|\tilde{q}_k\|}$$

Da $\text{Rang}\,(A) = n$ gilt, sind $a_1, \cdots, a_n$ linear unabhängig und das Verfahren bricht somit nicht ab ($\|\tilde{q}_k\| \neq 0$). Ferner gilt $Q^*Q = E_n$ nach Konstruktion. Weiterhin ist

$$a_k = \tilde{q}_k + \sum_{i=1}^{k-1} \langle a_k, q_i\rangle \, q_i = \underbrace{\|\tilde{q}_k\|_2}_{=r_{kk}} q_k + \sum_{i=1}^{k-1} \underbrace{\langle a_k, q_i\rangle}_{=r_{ik}} q_i \ , \ i < k$$

Setze nun noch $r_{ik} = 0$ für $i > k$, so folgt $A = QR$ und die Behauptung ist bewiesen. □
Doch nun zurück zur Normalengleichung. Aus $A^TAx = A^Tb$ und $A = QR$ wird $R^TRx = A^Tb$, d.h. $R^TRx = A^Tb$. Dies ist durch Rückwärtseinsetzen lösbar.
BEMERKUNG: Für den Aufwand der Verfahren gilt die Beziehung

$$\text{Aufwand}\,(QR) \approx 2 \cdot \text{Aufwand}\,(LR)$$

Dennoch haben wir das folgende Problem: Das Gram-Schmidt-Verfahren ist bei Rundungsfehlern ungeeignet, da die Orthonormalität der Spalten rasch verloren geht.

Householder-Verfahren

Für $v \in \mathbb{K}^m$ sei $v \otimes \overline{v} = v\overline{v}^T = \left(\overline{v}_i, v_j\right)_{i,j}$ das sog. *dyadische Produkt.*

Definition 4.21

Für $v \in \mathbb{K}^n$ mit $\|v\|_2 = 1$ heißt

$$S = I - 2v \otimes \overline{v} \tag{4.22}$$

Householder-Transformation.

Es gilt

$$S = S^* = S^{-1} \tag{4.23}$$

Denn es ist

$$S^* = I - 2\left(v \otimes \overline{v}\right)^* = I - vv^* = S$$

und

$$SS = \left(I - 2vv^*\right)\left(I - 2vv^*\right) = I - 2vv^* - 2vv^* + 4vv^* = I, \text{ d.h. } S^{-1} = S$$

Damit ist S hermitesch und unitär.
Die Eigenwerte von S sind ± 1.

(i) Ist $w \in \operatorname{orth}(v)$, so gilt $Pw = w \Rightarrow$ der Eigenwert ist 1. Nehme dann $n-1$ linear unabhängige Vektoren aus $\operatorname{orth}(v)$.

(ii) $Pv = v - 2vv^T v = -v \Rightarrow$ der Eigenwerte ist -1 und der Eigenvektor ist v. Für die Determinante von S gilt dann $\det(S) = 1$.

Algorithmus-Idee: Wähle $S_1, \cdots, S_n$ so, dass iterativ

$$A^{(i)} = S_i A^{(i-1)} \text{ mit } A^{(0)} = A$$

die Form

$$A^{(i)} = \begin{pmatrix} \star & \cdots & \star & \star & \star \\ 0 & \ddots & \star & & \star \\ 0 & 0 & \star & & \star \\ 0 & 0 & 0 & \star & \star \end{pmatrix}$$

hat. Nach n Schritten ist

$$\tilde{R} = A^{(n)} = \underbrace{S_n \cdots S_1}_{\text{unitär}} A = \tilde{Q}^* A \Rightarrow \tilde{Q}\tilde{R} = A$$

mit $R \in \mathbb{K}^{n \times n}$:

$$\tilde{R} = \begin{pmatrix} R \\ - \\ 0 \end{pmatrix}$$

$$\Rightarrow A = \tilde{Q}\tilde{R} = \left[Q | \underbrace{?}_{\text{egal}} \right] \cdot \begin{pmatrix} R \\ - \\ 0 \end{pmatrix} \Rightarrow A = QR$$

Bestimme nun die Matrizen $S_1, \cdots, S_n$. Zu S_1: Wir müssen a_1 an $\operatorname{span}(e_1)$ spiegeln. Wir erhalten

$$v = \frac{a + \|a\| e_1}{\|a + \|a\| e_1\|}$$

Nach dem ersten Schritt ist dann

$$\begin{pmatrix} 1 & & & \\ 0 & \star & \cdots & \star \\ \vdots & \vdots & & \vdots \\ 0 & \star & \cdots & \star \end{pmatrix}$$

Im i-ten Schritt wende das Verfahren auf die reduzierte Matrix an:

$$S_2 = \begin{pmatrix} I & 0 \\ 0 & I - 2\tilde{v}\tilde{v}^* \end{pmatrix} = I - 2vv^*$$

mit

$$v = \begin{pmatrix} \underbrace{0 \cdots 0}_{i-1 \text{ Nullen}} & |\tilde{v} \end{pmatrix}$$

Bemerkung: Die durch die Householder-Transformation erhaltene QR-Zerlegung ist nicht die bekannte QR-Zerlegung. Dies liegt daran, dass die Diagonalemente r_{ii} nicht positiv sein müssen.
Beispiel: Wir wollen nun die QR-Zerlegung der Matrix

$$\begin{pmatrix} 1 & 2 & 3 \\ 0 & 3 & 2 \\ 2 & 0 & 1 \end{pmatrix}$$

mittels Householder-Transformationen bestimmen. Im ersten Schritt ist

$$a_1 = \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix} \Rightarrow \|a_1\| = \sqrt{5}$$

$$\Rightarrow \tilde{v}_1 = a_1 \pm \|a_1\| e_1, \; v_1 = \frac{\tilde{v}_1}{\|\tilde{v}_1\|}$$

Da $\operatorname{sgn}(a_{11}) > 0$ ist, nehme $+$. Wir erhalten damit

$$\tilde{v}_1 = \begin{pmatrix} 1 + \sqrt{5} \\ 0 \\ 2 \end{pmatrix}, \|\tilde{v}_1\|^2 = 10 + 2\sqrt{5}$$

Somit ist

$$S_1 = I - 2v_1 v_1^* = I - 2\frac{\tilde{v}_1 \tilde{v}_1^*}{\|v_1\|^2} = \begin{pmatrix} -1/\sqrt{5} & 0 & -2/\sqrt{5} \\ 0 & 1 & 0 \\ -2/\sqrt{5} & 0 & 1/\sqrt{5} \end{pmatrix}$$

$$\Rightarrow A^{(1)} = S_1 A \approx \begin{pmatrix} -2.236 & -0.894 & -2.236 \\ 0 & 3 & -2 \\ 0 & -1.788 & -2.236 \end{pmatrix}$$

$$\tilde{a}_2 = \begin{pmatrix} 3 \\ -1.788 \end{pmatrix}, \operatorname{sgn}(\tilde{a}_{22}) > 0$$

$$\Rightarrow \tilde{v}_2 = \tilde{a}_{22} + \|\tilde{a}_{22}\| e_2 \; , \; v_2 = \frac{\tilde{v}_2}{\|\tilde{v}_2\|} \; \Rightarrow \; \tilde{v}_2 = \begin{pmatrix} 6.492 \\ 1.788 \end{pmatrix}$$

Dann ist

$$\|\tilde{v}_2\|^2 \approx 45.346$$

und

$$\tilde{S}_2 = I_2 - \frac{\tilde{v}_2 \tilde{v}_2^*}{\|v_2\|^2} = \begin{pmatrix} -0.859 & 0.512 \\ 0.512 & 0.859 \end{pmatrix}$$

$$S_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \star & \star \\ 0 & \star & \tilde{S}_2 \end{pmatrix}$$

$$\Rightarrow A = S_2 A^{(1)} = \begin{pmatrix} -2.236 & -0.894 & -2.236 \\ 0 & -3.492 & -2.862 \\ 0 & 0 & -0.896 \end{pmatrix} = R$$

Da S_j unitär und hermitesch, folgt $S_2 S_1 A = R \Rightarrow A = S_1 S_2 R = QR$ und

$$Q = S_1 S_2 = \begin{pmatrix} -0.447 & -0.458 & -0.768 \\ 0 & -0.859 & 0.512 \\ -0.894 & 0.229 & 0.384 \end{pmatrix}$$

Möchte man $R_{ii} > 0$, so kann mann QR durch $(QD)(DR)$ ersetzen mit

$$D = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix}$$

Dann gilt $(QD)^* (QD) = D^* Q^* Q D = D^* D = E_3$.

Singulärwertzerlegung

Die bisher vorgestellten Methoden zur Lösung linearer Gleichungssysteme oder Ausgleichsprobleme werden instabil, falls A schlecht konditioniert ist. Dadurch ist die Bestimmung der Ranges mittels LR oder QR-Zerlegung nicht mit genügender Sicherheit möglich. Die beste aktuelle Methode ist die *Singulärwertzerlegung* (englisch: *svd- singular value decomposition*). Hierbei wird A von links und rechts orthogonal transformiert. Sei nun $A \in \mathbb{R}^{m \times n}$ und $Q \in \mathbb{R}^{m \times m}, Z \in \mathbb{R}^{n \times n}$ orthonormal, so gilt

$$\|QAZ\|_2 = \|A\|_2 \text{ und } \| (QAZ)^{-1} \|_2 = \|Z^* A^{-1} Q^*\|_2 = \|A^{-1}\|_2$$

Somit folgt

$$\text{cond}_2 (A) = \text{cond}_2 (QAZ)$$

Die Kondition ändert sich also *nicht.*

Theorem 4.24

Sei $A \in \mathbb{R}^{m \times n}$. Dann existieren orthonormale *Matrizen $U \in \mathbb{R}^{m \times n}$ und $V \in \mathbb{R}^{n \times n}$ so, dass*

$$U^T A V = diag(\sigma_1, \cdots, \sigma_p) = D \in \mathbb{R}^{m \times n} \tag{4.25}$$

mit $p = \min\{m, n\}$, $\sigma_1 \geq \cdots \geq \sigma_p$ und

$$diag(\sigma_1, \cdots, \sigma_p) = \begin{pmatrix} \sigma_1 & & & 0 & \cdots & 0 \\ & \ddots & & \vdots & & \vdots \\ & & \sigma_p & 0 & \cdots & 0 \end{pmatrix} \text{ bzw. } \begin{pmatrix} \sigma_1 & & \\ & \ddots & 0 \\ & & \sigma_p \\ 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{pmatrix} \tag{4.26}$$

Man nennt die σ_i die Singulärwerte *von A.*

BEMERKUNG: Die so definierten σ_i sind eindeutig.
Aus dem Satz folgt

$$Av_i = \sigma_i u_i \ , \ A^T u_i = \sigma_i v_i \text{ mit } \|u_i\| = \|v_i\| = 1$$

Man nennt die u_i die *linkssginulären* und die v_i die *rechtssingulären* Werte. Aus $U^T AV = D$ folgt

$$D^2 = \left(U^T AV\right)^T U^T AV = V^T A^T UU^T AV = V^T A^T AV \Rightarrow VD^2 = VV^T A^T AV = A^T AV$$

$$\Rightarrow \sigma_i^2 v_i = A^T Av_i$$

Somit sind die v_i Eigenvektoren von $A^T A$ zu Eigenwerten σ_i^2. Analog sind die u_i die Eigenvektoren von AA^T zu Eigenwerten σ_i^2. Hieraus lässt sich ein Existenzbeweis machen:
BEWEIS: Sei $B = A^T A \in \mathbb{R}^{n\times n}$. Dann ist B symmetrisch. Demnach existiert eine orthnormale Matrix $V = (v_1, \cdots, v_n) \in \mathbb{R}^{n\times n}$ mit

$$V^T BQ = \begin{pmatrix} \lambda_1 & & \\ 0 & \ddots & \\ 0 & 0 & \lambda_n \end{pmatrix}$$

und $\lambda_1 \geq \cdots \geq \lambda_n$ Eigenwerten von B. Da $\text{Rang}\left(A^T A\right) = \text{Rang}\,(A)$ und die λ_i positiv sind, können wir für $i = 1, \cdots, r$

$$u_i = \frac{1}{\sqrt{u_i}} Av_i \in \mathbb{R}^m$$

setzen. Es folgt

$$\langle u_i, u_j \rangle = \frac{1}{\sqrt{\lambda_i \lambda_j}} \langle v_i, A^T Av_j \rangle = \frac{\sqrt{v_j}}{\sqrt{\lambda_i}} = \delta_{ij}$$

Ergänze nun $u_1, \cdots, u_r$ zu einem vollständigen Orthonormalsystem $u_1, \cdots, u_m$ und setze

$$D = \begin{pmatrix} \sqrt{\lambda_1} & & & & \\ & \ddots & & & \\ & & \sqrt{\lambda_r} & & \\ & & & 0 & \\ & & & & \ddots \end{pmatrix} \in \mathbb{R}^{m\times n}$$

Es gilt

$$(u_1, \cdots, u_n) \begin{pmatrix} \sqrt{\lambda_1} & & \\ & \ddots & \\ & & \sqrt{\lambda_r} \end{pmatrix} = A\,(v_1, \cdots, v_r) \Rightarrow UD = AV$$

Es kommen auf beiden Seiten nur Nullen dazu, und somit folgt $D = U^T AV$. □
Betrachte nun $A = UDV^T$ bzw. $U^T AV = D$ mit

$$\text{diag}\,(\sigma_1, \cdots, \sigma_p) = \begin{pmatrix} \sigma_1 & & & 0 & \cdots & 0 \\ & \ddots & & \vdots & & \vdots \\ & & \sigma_p & 0 & \cdots & 0 \end{pmatrix} \text{ bzw. } \begin{pmatrix} \sigma_1 & & & \\ & \ddots & 0 & \\ & & \sigma_p & \\ 0 & \cdots & 0 & \\ \vdots & & \vdots & \\ 0 & \cdots & 0 & \end{pmatrix}$$

Dann ist

(i) $\text{Rang}\,(A) = r$

(ii) $\text{Kern}\,(A) = \text{span}\,\{v_{r+1}, \cdots, v_n\}$

(iii) $\text{Bild}\,\{u_1, \cdots, u_r\}$

(iv) $A = UDV^T = \sum_{i=1}^{r} \sigma_i u_i v_i = \sum_{i=1}^{r} \sigma_i u_i v_i^T$

Betrachten wir nun das Problem zur Bestimmung des *numerischen Ranges*

$$\text{Rang}\,(A, \varepsilon) = \min_{\|A-B\|_2 \leq \varepsilon} \text{Rang}\,(B)$$

Es heißt A *numerisch rang-defizient*, falls

$$\text{Rang}\,(A, \varepsilon) < \min\{m, n\}$$

mit $\varepsilon = \text{eps}\|A\|_2$. Bei Messreihen ist ε an die Genauigkeit der Messwerte gekoppelt.

Theorem 4.27

(Fehlerabschätzung) Sei $A = UDV^T$ wie oben, so gilt für

$$A_k = \sum_{i=1}^{r} \sigma_i u_i \otimes v_i \; k < r \tag{4.28}$$

die Abschätzung

$$\min_{Rang(B)=k} \|A - B\|_2 = \|A - A_k\|_2 = \sigma_{k+1} \tag{4.29}$$

Daraus ergibt sich für $r_\varepsilon = Rang\,(A, \varepsilon)$:

$$\sigma_1 \geq \cdots \geq \sigma_{r_\varepsilon} > \varepsilon > \sigma_{r_\varepsilon - 1} \geq \cdots \geq \sigma_{\min\{m,n\}} \tag{4.30}$$

Beweis: Es sit $U^T A V = \text{diag}\,(\sigma_1, \cdots, \sigma_r)$ und $U^T A_k V = \text{diag}\,(\sigma_1, \cdots, \sigma_k, 0 \cdots, 0)$. Damit folgt $U^T (A - A_k) V = \text{diag}\,(0, \cdots, 0, \sigma_{k+1}, 0 \cdots, 0)$ und daher

$$\|A - A_k\|_2 = \|\text{diag}\,(0, \cdots, 0, \sigma_{k+1}, \cdots, \sigma_r)\,\|_2 = \sigma_{k+1} \text{ bzw. } \|A - B\|_2 \geq \sigma_{k+1} \text{ für } \text{Rang}\,(B) = k$$

Dazu wähle eine Orthonormalbasis für $\text{Ker}\,(B) = \text{span}\,\{x_1, \cdots, x_{n-k}\}$. Aus Dimensionsgründen folgt dann

$$\text{Ker}\,(B) \cap \text{span}\,\{v_1, \cdots, v_{k+1}\} \neq \emptyset$$

Sei z aus dieser Menge mit $\|z\|_2 = 1$. Dann gilt

$$Bz = 0 \text{ und } Az = \sum_{i=1}^{r} \sigma_i \,(u_i \otimes v_i)\, z = \sum_{i=1}^{k+1} \sigma_i u_i v_i^T z$$

$$\Rightarrow \|A - B\|_2^2 \geq \|\,(A - B)\, z\|_2^2 = \|Az\|_2^2 = \sum_{i,j=1}^{k} |\sigma_i|^2 \left(v_i^T z\right)^2 \geq \sigma_{k+1}^2 \sum_{i=1}^{k+1} \left(v_i^T z\right)^2 = \sigma_{k+1}^2$$

□

Mit der Singulärwertzerlegung kann man auch das Ausgleichsproblem lösen. $\|Ax - b\|_2 = \min!$ führte zu $A^T Ax = A^T b$. Ist $\text{Rang}\,(A) = n$, so gilt $x = \left(A^T A\right)^{-1} A^T b$. Ansonsten gibt es ∞-viele Lösungen. Diejenige mit kleinster Norm $\|x\|_2$ heißt *Minimallösung* des Ausgleichsproblems.

Theorem 4.31

Sei $A = UDV^T$ die (eine) Singulärwertzerlegung von $A \in \mathbb{R}^{m\times n}$. Dann ist

$$\overline{x} = \sum_{i=1}^{r} \frac{u_i^T b}{\sigma_i} v_i \tag{4.32}$$

di eindeutige Minimallösung des Ausgleichsproblems. Der Fehler genügt der Beziehung

$$\|A\overline{x} - b\|_2^2 = \sum_{i=r+1}^{m} \left(u_i^T b\right)^2 \tag{4.33}$$

BEWEIS: Da $x \mapsto \|x\|$ strikt konvex ist, ist die Minimallösung eindeutig. Für jedes $x \in \mathbb{R}^n$ gilt

$$\|Ax - b\|_2^2 = \|AVV^T x - b\|_2^2 = \| \left(U^T AV\right) V^T x - U^T b\|_2^2 = \|DV^T x - U^T b\|_2^2$$

Sei $z = V^T x$ (d.h. $x = Vz$), so gilt

$$\|Ax - b\|_2^2 = \|Dz - U^T b\|_2^2 = \sum_{i=1}^{r} \left(\sigma_i z_i - u_i^T b\right)^2 + \sum_{i=r+1}^{m} \left(u_i^T b\right)^2$$

Die Minimallösung muss zusätzlich noch $\|x\|_2 = \|z\|_2$ minimieren. Das Minimum bezüglich z erfüllt

$$\sigma_i z_i = u_i^T b \Rightarrow z_i = \frac{1}{\sigma_i} u_i^T b \underbrace{\Rightarrow}_{x=Vz} x = \sum_i v_i z_i = \sum_i \frac{u_i^T b}{\sigma_i} z_i$$

Der minimale Fehler ist dann

$$\|A\overline{x} - b\|_2^2 = \sum_{i=r+1}^{m} \left(u_i^T b\right)^2$$

□

BEMERKUNG: Sei $D^+ = \operatorname{diag}\left(\frac{1}{\sigma_1}, \cdots, \frac{1}{\sigma_r}, 0, \cdots, 0\right) \in \mathbb{R}^{n\times m}$, so gilt $\overline{x} = VD^+U^T b = A^+$. Man nennt A auch die *Pseudoinverse* von A. Der Satz sagt dann aus, dass

$$\overline{x} = A^+ b \text{ und } \|A\overline{x} - b\|_2^2 = \| \left(I - AA^+\right) b\|_2^2$$

Die Pseudoinverse ist die eindeutige Lösung von

$$\min_{X\in\mathbb{R}^{n\times m}} \|AX - I\|_{\text{Frob}}$$

Es gilt

$$\operatorname{Rang}(A) = n \Rightarrow A^+ = \left(A^T A\right)^{-1} A^T \,,\ \operatorname{Rang}(A) = m = n \Rightarrow A^+ = A^{-1}$$

BEMERKUNG: In der Praxis muss man $\operatorname{Rang}(A, \varepsilon)$ bei der Definition von A^+ nehmen.
BEMERKUNG: Eine numerisch stabile Berechnung der Singulärwertzerlegung ist sehr aufwendig.

Kapitel 5

Lineare Gleichungssysteme (Iterative Verfahren)

Große Gleichungssysteme, $n \gg 1000 = 10^3$, benötigen für das Gaußverfahren viel Speicher $n^2 = 10^6$ und hohen Aufwand ($n^3 = 10^9$ Multiplikationen). Auch Bandssysteme mit $n = 10^6$ und $m = 10^2$ benötigen schon $n \cdot m = 10^8$ Speicherplatz. In der Praxis sind Matrizen jedoch dünn besetzt, d.h. nur $5 - 20$ Einträge pro Zeile. Die folgendeniterativen Verfahren benötigen nur so viel Speicher, wie man für A (und *nicht* A^{-1}) braucht. Sie sind nicht exakt, sondern nähern sich iterativ der Lösung. Wir wollen im Folgenden das

(a) *Jacobi-Verfahren* (Gesamtschrittverfahren)

(b) *Gauß-Seidel-Verfahren* (Einzelschrittverfahren)

vorstellen.

Sei $A \in \mathbb{R}^{n \times n}$. Wir zerlegen A ind $A = D + L + R$:

$$\begin{pmatrix} a_{11} & & & & R \\ & \ddots & & & \\ & & D & & \\ & & & \ddots & \\ L & & & & a_{nn} \end{pmatrix}$$

Dabei sind D die Diagonale, L die Einträge unter der Diagonalen und R die Einträge über der Diagonalen. Das Gleichungssystem $Ax = b$ können wir umschreiben als

$$a_{jj}x_j + \sum_{k=1, k\neq j}^{n} a_{jk}x_k = b_j \ , j = 1, \cdots, n \quad \text{bzw.} \quad Dx + (L+R)x = b$$

Ist D diagonaldominant, so kann man die sukzessive Iteration mit $C = D$ versuchen. Das heißt:

$$f(x) = b - Ax = 0 \ \Rightarrow \ x^{k+1} = x^k + Cf(x^k) = x^k + D^{-1}(b - Ax^k) =$$

$$x^k + D^{-1}(b - Ax^k) = x^k + D^{-1}(b - (L + D + R)x^k) = D^{-1}(b - (L + R)x^k) = -D^{-1}(L + R)x^k + D^{-1}b$$

Bemerkung: $f'(x) = A$, $C = D^{-1} \approx A^{-1}$

Hieraus folgt das *Jacobi-Verfahren*:

$$x^{k+1} = -\underbrace{D^{-1}(L+R)}_{\text{Jacobimatrix}} x^k + D^{-1}b \tag{5.1}$$

Im *Gauß-Seidel-Verfahren* benutzt man, dass $D + L$ eine untere Dreiecksmatrix ist und mittels Vorwärtseinsetzen lösbar ist. Sei also $C = L + D$. Dann folgt

$$x^{k+1} = x^k + C(b - Ax^k) = (I - CA)x^k + Cb = (I - (L + D)^{-1}(L + D + R))x^k + (L + D)b =$$

$$- \underbrace{(L + D)^{-1}R}_{\text{Gauß-Seidel-Matrix}} x^k + (L + D)^{-1}b$$

Dies ist das *Gauß-Seidel-Verfahren.*

Man spricht allgemein von einem *Splittingverfahren*, wenn man A in $A = B + (A - B)$ zerlegt und statt A nun B invertiert. Das entspricht der sukzessiven Approximation mit $C = -B^{-1}$, d.h.

$$x^{k+1} = x^k + B^{-1}(b - Ax^k) = (I - B^{-1}A)x^k + B^{-1}b = B^{-1}(B - A)x^k + B^{-1}b$$

Dabei gilt dann $f'(x) = b - Ax,\ f'(x) = -A\ , B^{-1} \approx -A^{-1}$. Sei nun

$$g(x) = \underbrace{B^{-1}(B - A)}_{=M} x + \underbrace{B^{-1}b}_{=c} \Rightarrow\ g(x) - g(y) = M(x - y) \tag{5.2}$$

Dies ist eine Kontraktion, falls $\|M\| < 1$. Die Konvergenz folgt dann aus dem Banachschen Fixpunktsatz. Wenden wir zwei Schritte an, so erhalten wir

$$g^2(x) = g(g(x)) = Mg(x) + c = M(Mx + c) + c = M^2x + Mc + c$$

Dies konvergiert für $\|M^2\| < 1$. Für die Matrix

$$M = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

ist

$$M^2 = \begin{pmatrix} cc0 & 0 \\ 0 & 0 \end{pmatrix}$$

sodass $\|M^2\| < 1$ ein schwächeres Kriterium ist. Dies führt schließlich zu der *Idee*:

$$x^k \mapsto g(x^k) \text{ konvergiert, falls } \lim_{n\to\infty} (\|M^n\|)^{1/n} < 1 \tag{5.3}$$

Bestimmen wir nun $\lim\limits_{n\to\infty} (\|M^n\|)^{1/n}$.

Lemma 5.4

Es gilt $\|M^n\|^{1/n} \geq \rho(M)$ *und*

$$\lim_{n\to\infty} (\|M^n\|)^{1/n} = \rho(M) \tag{5.5}$$

für jede *Matrixnorm, wobei* $\rho(M)$ *der Spektralradius ist.*

Beweis: Falls $\rho(M) = 0$, so ist $M = 0$ - daraus folgt dann die Behauptung. Sei also $\rho(M) > 0$ und $\lambda_1, \cdots, \lambda_m$ die Eigenwerte von M mit Eigenvektoren $x_1, \cdots, x_m$. Dann gilt

$$\|M^n\| \cdot \|x_i\| \geq \|M^n x_i\| = \|\lambda_i^n x_i\| = |\lambda_i|^n \|x_i\|,\ i = 1, \cdots, m$$

$$\Rightarrow \|M^n\|^{1/n} \geq |\lambda_i|,\ i = 1, \cdots, m \ \Rightarrow\ \|M^n\|^{1/n} \geq \rho(M)$$

$$\Rightarrow \liminf_{n\to\infty} \|M^n\|^{1/n} \geq \rho(M)$$

beziehungsweise

$$\limsup_{n\to\infty} \|M^n\|^{1/n} \leq \rho(M)$$

Sei $M_\delta = \frac{1}{\rho(M)+\delta} \cdot M \Rightarrow \rho(M_\delta) < 1$. Wir behaupten, dass

$$\lim_{n\to\infty} \|M_\delta^n\|^{1/n} \leq 1 \tag{5.6}$$

Sei hierzu $M_\delta = U^{-1} J_\delta U$ die Jordanzerlegung von M_δ, also

$$J_\delta = \begin{pmatrix} J_1 & & 0 \\ & \ddots & \\ 0 & & J_r \end{pmatrix}, \quad J_i = \begin{pmatrix} \alpha_i & 1 & 0 \\ & \ddots & 1 \\ 0 & & \alpha_i \end{pmatrix}$$

mit $|\alpha_i| < 1$. Sei $q = \max_i |\alpha_i| = \frac{\rho(M)}{\rho(M)+\delta} < 1$. Dann gilt

$$\lim_{n\to\infty} J_i^n = 0$$

komponentenweise und auch bzgl. $\|\cdot\|$. Daher gilt $J_\delta^n \to 0$ bezüglich $\|\cdot\|$ und somit $\|M_\delta^n\| \leq \|U^{-1}\| \cdot \|J_\delta^n\| \cdot \|U\| \to 0,\ n \to \infty$. Damit folgt $\|M_\delta^n\| < 1/2$ für alle $n \geq N$. Somit

$$\limsup_{n\to\infty} \|M_\delta^n\|^{1/n} \leq \limsup_{n\to\infty} (1/2)^{1/n} = 1$$

Das ergibt die Zwischenbehauptung. Weiter folgt

$$\limsup_{n\to\infty} \|M^n\|^{1/n} = (\rho(M)+\delta) \limsup_{n\to\infty} \|M_\delta^n\|^{1/n} \leq \rho(M) + \delta$$

Da dies für alle δ gilt, folgt

$$\limsup_{n\to\infty} \|M^n\|^{1/n} \leq \rho(M)$$

Damit folgt die Behauptung. ■

Bemerkung: $\rho(B)$ ist *keine* Matrixnorm. Beispielsweise ist

$$\rho\left(\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}\right) = 0 \quad \text{obwohl} \quad \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \neq \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

Für symmetrische $B \in \mathbb{R}^{n\times n}$ ist aber

$$\rho(B) = \|B\|_2 = \sup_{x\neq 0} \frac{\|Bx\|_2}{\|x\|_2} \tag{5.7}$$

Lemma 5.8

Zu jedem $M \in \mathbb{R}^{n\times n}$ und $\delta > 0$ existiert eine Norm $|||\cdot|||$, so dass die induzierte Matrixnorm

$$\rho(M) \leq |||M||| \leq \rho(M) + \delta \tag{5.9}$$

erfüllt.

Beweis: Die Beziehung $\rho(M) \leq |||M|||$ wurde bereits in Lemma 6.1. gezeigt. Für jede reguläre Matrix $C \in \mathbb{R}^{n\times n}$ ist $|||x||| = \|Cx\|_2$ eine Norm. Die induzierte Matrixnorm ist

$$|||A||| = \sup_{x\neq 0} \frac{|||Ax|||}{|||x|||} = \sup_{x\neq 0} \frac{|||CAx|||_2}{\|Cx\|_2} = \sup_{x\neq 0} \frac{\|CAC^{-1}x\|_2}{\|x\|_2} = \|CAC^{-1}\|_2$$

Also gilt

$$|||M||| = \|CAC^{-1}\|_2$$

Sei nun $M = U^{-1}JU$ die Jordanzerlegung.

Rest des Beweises fehlt noch.

■

Theorem 5.10

Für $M \in \mathbb{R}^{n\times n}, x_0 \in \mathbb{R}^n$ betrachte die Iteration

$$x^{k+1} = Mx^k + c,\ k \in \mathbb{N}_0 \tag{5.11}$$

Die Folge x^k konvergiert genau dann für jeden Startwert $x^0 \in \mathbb{R}^n$ gegen den Fixpunkt $\overline{x}$ von $\overline{x} = M\overline{x} + c$, falls $\rho(M) < 1$. Im Falle der Konvergenz gilt

$$\sup_{x^0\in\mathbb{R}^n} \limsup_{k\to\infty} \left(\frac{\|x^k - \overline{x}\|}{\|x^0 - \overline{x}\|}\right)^{1/k} = \rho(M) \tag{5.12}$$

Beweis: Sei $e^k = x^k - \overline{x}$. Dann gilt $e^{k+1} = x^{k+1} - \overline{x} = Mx^k - M\overline{x} = Me^k$. Nun unterscheiden wir mehrere Fälle.

(a) Ist $\rho(M) < 1$, so existiert nach Lemma 6.2. eine Norm $|||\cdot|||$ mit $|||M||| < 1$. Dann gilt $|||e^k||| \le |||M|||^k \cdot |||e^0||| \to 0,\ k \to \infty$. Die Folge $(e^k)_k$ konvergiert dann bezüglich $|||\cdot|||$, und da die Normen äquivalent sind, auch bezüglich $\|\cdot\|$.

(b) Aus der Konvergenz für jeden Startwert folgt für $x^0 = w + x$ mit einem Eigenvektor x zum größten Eigenwert λ, dass $e^0 = w$ und $e^k = M^k w = \lambda^k w$. Dann gilt

$$\|\lambda^k w\| \to 0,\ k \to \infty \ \Rightarrow\ |\lambda| < 1 \ \Rightarrow\ \rho(M) < 1$$

Außerdem erhalten wir

$$\left(\frac{\|e^k\|}{\|e^0\|}\right)^{1/k} = |\lambda|$$

Für den zweiten Teil des Beweises sei nun $\rho(M) < 1$. Dann konvergiert das Verfahren. Sei weiter $|||\cdot|||$ eine Norm mit

$$\rho(M) \le |||M||| \le \rho(M) + \delta$$

Aus der Äquivalenz der Normen erhalten wir dann $c_0|||x||| \le \|x\| \le c_1|||x|||$ und damit

$$\|e^k\| \le c_1|||e^k||| \le c_1|||M|||^k|||e^0||| \le \frac{c_1}{c_0}(\rho(M) + \delta)^k\|e^0\|$$

Daraus erhalten wir

$$\left(\frac{\|e^k\|}{\|e^0\|}\right)^{1/k} \le \left(\frac{c_1}{c_0}\right)^{1/k}(\rho(M) + \delta)$$

$$\Rightarrow\ \limsup_{k\to\infty}\left(\frac{\|e^k\|}{\|e^0\|}\right)^{1/k} \le \rho(M) + \delta$$

Da $\delta > 0$ beliebig war, folgt damit die Behauptung. ■

Bemerkung: Falls $\rho(M) \approx 0.99$, so braucht man immer noch ≈ 230 Schritte bis $(\rho(M))^k \le 10^{-1}$.
Wir diskutieren nun einige

Abbruchkriterien

(a) Nach dem Banachschen Fixpunktsatz gilt

$$\|x^k - \overline{x}\| \leq \frac{\|B\|}{1-\|B\|} \underbrace{\|x^k - \overline{x}\|}_{=\delta x^k \text{ Update}}$$

$$\Rightarrow \frac{\|x^k - \overline{x}\|}{\|x^k\|} \leq \frac{\|B\|}{1-\|B\|} \cdot \frac{\|\delta x^k\|}{\|x^k\|}$$

Daraus erhalten wir ein mögliches Abbruchkriterium:

$$\frac{\|B\|}{1-\|B\|} \cdot \frac{\|\delta x^k\|}{\|x^k\|} \leq \varepsilon$$

Dennoch bleibt das Problem, dass $\rho(B)$ bzw. $\|B\|$ abgeschätzt werden muss.

(b) Betrachte das *Residuum* $Ax^k - b$. Zur Erinnerung: Die Gleichung $Ax = b$ führte zur Iteration

$$x^{k+1} = \underbrace{B^{-1}(B-A)}_{=M} x^k + \underbrace{B^{-1}b}_{=c}$$

$$\Rightarrow e^k = x^k - \overline{x} = A^{-1}(Ax^k - A\overline{x}) = A^{-1}(Ax^k - b)$$

$$\Rightarrow \|e^k\| \leq \|A^{-1}\| \cdot \|Ax^k - b\| \leq \text{cond}(A)\frac{1}{\|A\|}\|Ax^k - b\|$$

Daraus folgt

$$\frac{\|e^k\|}{\|x\|} \leq \text{cond}(A)\frac{1}{\|A\| \cdot \|x\|}\|Ax^k - b\| \leq \text{cond}(A)\frac{\|Ax^k - b\|}{\|b\|}$$

Allerdings haben wir hier zwei Nachteile:

- Ax^k muss berechnet werden.
- cond(A) unbekannt.

Bemerkung: $Ax = b$ führte zum iterativen Verfahren

$$x^{k+1} = \underbrace{B^{-1}(B-A)}_{=M} x^k + \underbrace{B^{-1}b}_{=c}$$

Dabei haben wir zwei *Ziele*:

(a) *gute Konvergenz*. Hierfür brauchen wir $\rho(M) << 1$.

(b) $Mx^k = (I - B^{-1}A)x^k$ soll einfach berechenbar sein, d.h. $B^{-1}Ax^k$ soll einfach berechenbar sein.

Diese zwei Ziele sind konträr.

Zu (a): Optimal ist $B = A \Rightarrow M = 0 \Rightarrow \rho(M) = 0$.

Zu (b): Einfach ist $B = D$, jedoch ist oft $\rho(I - D^{-1}A) \approx 1$. Dann konvergiert das Verfahren aber langsam.

Beim Jacobi- und Gauß-Seidel-Verfahren ist (b) gut erfüllt. Dies geht aber auf Kosten von (a) - der Konvergenzgeschwindigkeit. Nun jedoch zurück zum Jacobi- und Gauß-Seidel-Verfahren:

Theorem 5.13

Sei $A \in \mathbb{R}^{n\times n}$ (bzgl. der Zeilen) strikt diagonaldominant, *d.h.*

$$\sum_{k=1,k\neq j}^{n} |a_{jk}| < |a_{jj}| \, , j = 1, \cdots, n \tag{5.14}$$

so ist mit der Jacobimatrix $J =^{-} D^{-1}(L+R)$ und der Gauß-Seidel-Matrix $H = -(D+R)^{-1}R$

$$\rho(J) < 1 \text{ und } \rho(H) < 1$$

Folglich konvergieren das Jacobi- und das Gauß-Seidel-Verfahren.

Beweis: Sei λ der größte Eigenwert von J mit dem Eigenvektor v, $\|v\|_\infty = 1$. Sei μ der größte Eigenwert von H mit Eigenvektor w mit $\|w\|_\infty = 1$. Dann gilt

$$|\lambda| \leq \|J\|_\infty = \|D^{-1}(L+R)\|_\infty \leq \max_j \frac{1}{|a_{jj}|} \sum_{k=1,k\neq j}^{n} |a_{jk}| < 1$$

nach Voraussetzung und

$$\mu w = Hw = -(D+L)^{-1}Rw \;\Rightarrow\; (\mu D + \mu L)w = -Rw \Rightarrow \mu w = -D^{-1}(muL + R)w$$

Wir erhalten

$$|\mu| \leq \|D^{-1}(\mu L + R)\| \infty \leq \max_j \frac{1}{|a_{jj}|} \sum_{k=1,k\neq j}^{n} \max\{|\mu|, 1\}|a_{jk}|$$

$$< \max\{|\mu|, 1\} \text{ nach Voraussetzung } \Rightarrow\; |\mu| < 1$$

Damit ist der Satz bewiesen. ■

Sei nun A die oben betrachtete Bandmatrix. A ist *nicht* strikt diagonaldominant, aber diagonaldominant. Sie ist jedoch in den ersten und letzten 4 Zeilen strikt diagonaldominant. Dies werden wir nun nutzen:

Definition 5.15

Eine Matrix heißt irreduzibel, *falls es* keine *Permutationsmatrix P gibt mit*

$$PAP^T = \begin{pmatrix} \tilde{A_{11}} & 0 \\ \tilde{A_{21}} & \tilde{A_{22}} \end{pmatrix} \tag{5.16}$$

mit $\tilde{A_{11}} \in \mathbb{R}^{p\times p}$, $\tilde{A_{22}} \in \mathbb{R}^{(n-p)\times(n-p)}$ und $0 < p < n$.

Lemma 5.17

$A \in \mathbb{R}^{n\times n}$ ist irreduzibel genau dann, wenn der Graph

$$G(A) = \{\text{Knoten } K_1, \cdots, K_n, \text{Kante } \overline{KjK_k} \Leftrightarrow\; a_{jk} \neq 0,\; j,k = 1, \cdots, n\} \tag{5.18}$$

Beweis: Die Reduzibilität ist äquivalent zur Existenz von $J, K \subset \{1, \cdots, n\}, J, K \neq \emptyset, J \cap K = \emptyset$ mit $a_{jk} = 0$ für alle $j \in J, k \in K$. ■

Theorem 5.19

(Schwaches Zeilensummenkriterium). Ist $A \in \mathbb{R}^{n\times n}$ irreduzibel, diagonaldominant und gilt

$$\sum_{k=1,k\neq r}^{n} |a_{rk}| < |a_{rr}| \tag{5.20}$$

für (mindestens) ein $r \in \{1, \cdots, n\}$, so gilt $\rho(J), \rho(H) < 1$.

Beweis : Wir beschränken uns auf $\rho(J) < 1$. Sei λ der größte Eigenwert von J mit Eigenvektor v mit $\|v\|_\infty = 1$.

$$|\lambda| \leq \max_j \frac{1}{|a_{jj}|} \sum_{k=1, k\neq j}^{n} |a_{j,k}| \leq 1, \ |\mu| \leq 1 \text{ für} H$$

Genauer gilt

$$|v_j| = |\lambda \| v_j| = |[D^{-1}(L+R)v]_j| \leq \frac{1}{a_{jj}} \sum_{k\neq j} |a_{jk}\| v_k| \underbrace{\Rightarrow}_{j=r} |v_r| < \|v\|_\infty = 1$$

Sei nun $i_1, \cdots, i_m$ Kette mit $i_1 = 1$ und $\{i_1, \cdots, i_m\} = \{1, \cdots, n\}$ (auch doppelte möglich).

Induktionsanfang: $|v_{i1}| < 1$

Induktionsschritt:

$$|v_{im}| \leq \frac{1}{|a_{jj}|} (\sum_{k\neq j, i_{m-1}} |a_{jk}\| v_k| + |a_{j,i_{m-1}}\| v_{i_{m-1}}|) < 1$$

$$\Rightarrow |v_k| < 1 \text{ für alle} k = 1, \cdots, n \ \Rightarrow \|v\|_\infty < 1$$

Damit erhalten wir einen Widerspruch und somit $\rho(J) < 1$. ■

SOR-Verfahren (successivce overrelaxation method)

Da oft $\rho(J), \rho(H) \approx 1 - \varepsilon$, sodass das Jacobi- und Gauß-Seidel-Verfahren zu langsam konvergieren, versucht man, das Verfahren durch *Relaxation* zu verbessern. Bei Gauß-Seidel gilt:

$$Dx^{k+1} = -Lx^{k+1} - Rx^k + b$$

Hieraus wird durch Relaxation der neue Algorithmus des SOR-Verfahrens:

$$D\tilde{x}^{k+1} = -Lx^{k+1} - Rx^k + b \ \text{ und } \ x^{k+1} = \omega\tilde{x}^{k+1} + (1-\omega)x^k \tag{5.21}$$

Beachte jedoch: *Dies ist eine Zirkeldefinition!* Durch Umformen erhalten wir

$$Dx^{k+1} = \omega D\tilde{x}^{k+1} + (1-\omega)Dx^k = -\omega Lx^{k+1} - \omega Rx^k + \omega b + (1-\omega)Dx^k$$

und somit

$$(D - \omega L)x^{k+1} = ((1-\omega)D - \omega R)x^k + \omega b \tag{5.22}$$

und durch Multiplikation mit $\frac{1}{\omega}$:

$$\left(\frac{1}{\omega}D + L\right)x^{k+1} = \left(\frac{1}{\omega}D - D - R\right)x^k + b$$

Damit ist das SOR-Verfahren ein *Splitting-Verfahren* mit

$$B_\omega := \frac{1}{\omega}D + L \ \text{ mit Iterationsmatrix } \ M_\omega := B^{-1}(B - A) \tag{5.23}$$

Bemerkung: Dies ist auch für $\omega = 0$ wohldefiniert.

Lemma 5.24

Für $A \in \mathbb{R}^{n\times n}$ mit regulärem Diagonalanteil D gilt

$$\rho(M_\omega) \geq |\omega - 1|,\ \omega \in \mathbb{R} \setminus \{0\} \tag{5.25}$$

Beweis: Wir berechnen die Determinante von M:

$$\det(M) = \det(B^{-1}(B-A)) = \det\left(\left(\frac{1}{\omega}D + L\right)^{-1}\right)\det(\frac{1}{\omega}D - D - R)$$

$$= \omega^n\left(1 - \frac{1}{\omega}\right)^n = (\omega - 1)^n$$

Damit folgt

$$\rho(M) = \max_i |\lambda_i| \geq \left(\prod_{i=1}^n |\lambda_i|\right)^{1/n} = |\det(M)|^{1/n} = |\omega - 1|$$

■

Bemerkung: Für $\rho(M) < 1$ muss also $\omega \in (0, 2)$ sein.

Theorem 5.26

Für positiv definite (symmetrische) Matrizen $A \in \mathbb{R}^{n\times n}$ gilt $\rho(M_\omega) < 1$ für $\omega \in (0, 2)$. Insbesondere ist das Gauß-Seidel-Verfahren konvergent.

Beweis: Es gilt $A = L + D + L^T$, da A symmetrisch ist. Sei im Folgenden $\omega \in (0, 2)$ sowie $\lambda \in \rho(M_\omega)$ mit Eigenvektor v, d.h. $M_\omega v = \lambda v$. Da $M_\omega = (D + \omega L)^{-1}((1-\omega)D - \omega L^T)$ folgt

$$((1-\omega)D - \omega L^T)v = \lambda(D + \omega L)v$$

und damit

$$\omega(D + L^T)v = (1-\lambda)Dv - \lambda\omega Lv$$

Hieraus erhalten wir

$$\omega Av = \omega(D + L^T)v + \omega Lv = (1-\lambda)Dv - \lambda\omega Lv + \omega Lv = (1-\lambda)Dv + \omega(1-\lambda)Lv$$

und

$$\lambda\omega Av = \lambda\omega(D + L^T)v + \lambda\omega Lv = \lambda\omega(D + L^T)v + (1-\lambda)Dv - \omega(D + L^T)v =$$

$$(1-\lambda)(1-\omega)Dv - \omega(1-\lambda)L^T v$$

Damit ergibt sich

$$wv^T Av = (1-\lambda)v^T Dv + \omega(1-\lambda)v^T Lv$$

und

$$\lambda w v^T Av = (1-\lambda)(1-\omega)v^T Dv - w(1-\lambda)\underbrace{v^T L^T v}_{=v^T Lv}$$

Letztlich folgt

$$\omega(1+\lambda)v^T Av = (1-\lambda)(2-\omega)v^T Dv$$

Dabei ist $v^T Av > 0$, da A positiv definit. Dann ist aber auch D positiv definit, also $v^T Dv > 0$. Es folgt, dass $\lambda \neq \pm 1$, denn sonst ist genau eine Seite gleich 0. Umstellen liefert

$$\frac{1+\lambda}{1-\lambda} = \frac{2-\omega}{\omega}\frac{v^T Dv}{v^T Av} > 0 \Rightarrow \mu = \frac{1+\lambda}{1-\lambda} > 0$$

$$\Rightarrow\ \lambda = \frac{\mu - 1}{\mu + 1} \in (-1, +1) \Rightarrow |\lambda| < 1$$

■

Die qualitativen Aussagen lassen sich für spezielle Matrizen verschärfen:

Definition 5.27

Eine Matrix $A \in \mathbb{R}^{n\times n}$ mit $A = L + D + R$ heißt konsistent geordnet, *falls für alle $\alpha \in \mathbb{C}$*

$$\sigma(D^{-1}(\alpha L + \alpha^{-1}R)) = \sigma(D^{-1}(L + R)) \tag{5.28}$$

gilt (unabhängig von α mit der Jacobi-Matrix $J = D^{-1}(L + R)$.

Bemerkung:

(a) Tridiagonalmatrizen sind stets konsistent geordnet.

(b) Die Bandmatrix des obigen Beispiels ist ebenso konsistent geordnet.

Theorem 5.29

Sei $A \in \mathbb{R}^{n\times n}$ konsistent geordnet und $\omega \in (0, 2)$, so gilt

(a) $\lambda \in \sigma(J) \Rightarrow \lambda \in \sigma(J)$

(b) Gilt $\mu \in \sigma(J)$, so ist jedes λ mit

$$(\lambda + \omega - 1) = (\pm)\lambda^{1/2}\omega\mu$$

ein Eigenwert von M_ω.

(c) Ist $\lambda \neq 0$ mit $\lambda \in \sigma(M_\omega)$, so ist $\mu \in \sigma(J)$.

Beweis : Zu (a): Wir rechnen

$$\sigma(J) = \sigma(D^{-1}(L + R)) = \sigma(D^{-1}(-L - U)) = \sigma(-J) = -\sigma(J)$$

Zu (c): Ist $\lambda \in \sigma(H_\omega), \lambda \neq 0$, so ist $H_\omega v = \lambda v$ äquivalent zu

$$((1 - \omega)I - wD^{-1}R)v = \lambda(I - \omega D^{-1}L)v$$

bzw.

$$(\lambda + \omega - 1)v = -\lambda^{-1/2}\omega J(\lambda^{1/2})v$$

Hier ist $\lambda^{1/2}$ *eine* Wurzel von λ. Dies ist allerdings äquivalent dazu, dass v ein Eigenvektor von $J(\lambda^{1/2})$ ist zum Eigenwert

$$\mu = \frac{\lambda + \omega - 1}{\lambda^{1/2}\omega}$$

(b) folgt völlig analog. ■

Folgerung 5.30

Sei A konsistent geordnet und positiv definit. Dann gilt

$$\rho(H) = \rho(J)^2 \tag{5.31}$$

d.h. das Gauß-Seidel-Verfahren ist doppelt so gut wie das Jacobi-Verfahren.

Beweis : Er folgt einfach aus $\lambda = \lambda^{1/2}\mu \Rightarrow \lambda = \mu^2$. ■

Hieraus lässt sich der optimale Relaxationsparameter ablesen. Es gilt

$$\omega_{\text{opt}} = \text{argmin}_{w\in(0,2)}\rho(H_\omega) \tag{5.32}$$

Und zwar gilt für konsistent geordnete, positiv definite A:

$$\rho(H_\omega) = \begin{cases} \omega - 1 & \text{für } \omega_{\text{opt}} \leq \omega \\ \frac{1}{2}(\rho(J)^2\omega \pm \sqrt{(\rho(J))^2\omega^2 - 4(\omega-1)})^2, & \text{für } \omega \leq \omega_{\text{opt}} \end{cases}$$

[Graphik_para_gera]

Da die Parabel bei ω_{opt} die Steigung $-\infty$ hat, sollte man lieber ω_{opt} leicht *größer* wählen.

Theorem 5.33

Sei A konsistent geordnet und positiv definit. Seien ferner die Eigenwerte von J reell mit $\rho(J) < 1$. Dann gilt:

$$\omega_{opt} = \frac{2}{q + \sqrt{1-(\rho(J))^2}}\ ,\ \rho(H_{\omega_{opt}}) = \omega_{opt} - 1 = \frac{1-\sqrt{1-(\rho(J))^2}}{1+\sqrt{1-\rho(J)^2}} \tag{5.34}$$

BEISPIELE:

(a) $(\rho(J))^2 = \rho(H_1) = 0.99 \Rightarrow \sqrt{1-(\rho(J))^2} = 0.1 \Rightarrow \rho(H_{\omega_{opt}}) = \frac{1-0.1}{1+0.1} \approx 0.81$

(b) $(\rho(J))^2 = \rho(H_1) = 0.84 \Rightarrow \sqrt{1-(\rho(J))^2} = 0.4 \Rightarrow \rho(H_{\omega_{opt}}) = \frac{0.6}{1.4} = 0.43$

Abstiegsverfahren

Im Folgenden betrachten wir primär (symmetrische) positiv definite Matrizen.

Lemma 5.35

Ist $A \in \mathbb{R}^{n\times n}$ positiv definit und symmetrisch, so ist die Lösung von $Ax = b$ gleich dem eindeutigen Minimierer von

$$\mathcal{E}\colon \mathbb{R}^{n\times n} \to \mathbb{R}\ ,\ \mathcal{E}(y) = \frac{1}{2}\langle Ay, y\rangle - \langle b, y\rangle \tag{5.36}$$

Beweis: $\mathcal{E}$ ist konvex. Es gilt

$$\mathcal{E}(y) \geq c_0\|y\|_2^2 - \|b\|_2\|y\|_2 \geq (c_0\|y\|_2 - \|b\|_2)\|y\|_2 \geq c_1$$

$\mathcal{E}$ ist somit nach unten beschränkt. Damit existiert ein Minimierer. Es ist $\mathcal{E} \in C^1$ und weiters

$$(\delta\mathcal{E})(y)(z) = \frac{1}{2}\langle Ay, z\rangle + \frac{1}{2}\langle Az, y\rangle - \langle b, z\rangle = \langle Ay, z\rangle - \langle b, z\rangle$$

da A symmetrisch ist. Beim Minimum $\overline{x}$ gilt:

$$\langle A\overline{x}, z\rangle - \langle b, z\rangle = 0\ \forall z$$

Damit ergibt sich $A\overline{x} = b$, und da A regulär ist, $x = \overline{x}$. Das ist aber die Behauptung. ■

BEMERKUNG: Sei x eine Lösung von $Ax = b$ (bzw. Minimierer von $\mathcal{E}$). Dann gilt für alle y:

$$\mathcal{E}(y) - \mathcal{E}(x) = 1/2\,\langle ay, y\rangle - 1/2\,\langle Ax, x\rangle - \langle b, y\rangle + \langle b, x\rangle = 1/2\,\langle Ay, y\rangle - 1/2\,\langle Ax, x\rangle + \langle b, x - y\rangle \tag{5.37}$$

$$= \tfrac{1}{2}\Big(\langle Ay, Ay\rangle + \langle Ax, x\rangle - 2\langle Ax, y\rangle\Big) = \tfrac{1}{2}\langle A(y-x), x-y\rangle.$$

Der Gradient von $\mathcal{E}$ ist gegeben durch

$$\langle \nabla \mathcal{E}(y), z \rangle = \delta \mathcal{E}(y)(z) =$$

$$\langle Ay, z \rangle - \langle b, z \rangle \Rightarrow \nabla \mathcal{E}(y) = Ay - b$$

Dies ist die Richtung des *steilsten Abstiegs*. Die *Idee von Abstiegsverfahren* ist die Folgende:

(a) Bestimme die Abstiegsrichtung r^k so, dass $\mathcal{E}$ in Richtung r^k kleiner wird.

(b) Setze $x^{k+1} = x^k + \alpha_r r^k$ mit einem Faktor $\alpha_r \in \mathbb{R}$ so, dass $\alpha \mapsto \mathcal{E}(x^k + \alpha r^k)$ minimiert wird.

Anhand von (b) bestimmen wir nun das *Abstiegsverfahren*. Sei $h(\alpha) = \mathcal{E}(x^k + \alpha r^k)$. Das Mininum erfüllt $h'(\alpha) = 0$, und daher gilt

$$0 = h'(\alpha) = \delta \mathcal{E}(x^k + \alpha r^k)(r^k) = \left\langle A(x^k + \alpha r^k) - b, r^k \right\rangle = \left\langle Ax^k - b, r^k \right\rangle + \alpha \left\langle Ar^k, r^k \right\rangle$$

$$\Rightarrow \alpha = -\frac{\left\langle Ax^k - b, r^k \right\rangle}{\langle Ar^k, r^k \rangle} = -\frac{\left\langle g^k, r^k \right\rangle}{\langle Ar^k, r^k \rangle}$$

mit Gradient $g^k = Ax^k - b$. Also erhalten wir für das allgemeine *Abstiegsverfahren*:

$$g^k = Ax^k - b \text{ und } \alpha_k = -\frac{\left\langle g^k, r^k \right\rangle}{\langle Ar^k, r^k \rangle} \text{ und } x^{k+1} = x^k - \alpha_k r^k \tag{5.38}$$

Wir diskutieren nun Gradientenverfahren.

Theorem 5.39

Sei $A \in \mathbb{R}^{n \times n}$ symmetrisch und positiv definit, so konvergiert das Gradientenverfahren, *d.h. Startwert $x_0 \in \mathbb{R}^n$,*

$$g^k = Ax^k - b \text{ und } \alpha_k = -\frac{\left\langle g^k, r^k \right\rangle}{\langle Ar^k, r^k \rangle} \text{ und } x^{k+1} = x^k - \alpha_k r^k \tag{5.40}$$

gegen die *Lösung von $Ax = b$.*

Bemerkung: Falls $\left\langle Ag^k, g^k \right\rangle = 0$, so gilt $g^k = 0$, also $Ax^k - b = 0$. Damit ist die Lösung schon gefunden.

Beweis Beweis des Satzes. : Sei x die Lösung von $Ax = b$. Dann gilt

$$\mathcal{E}(x^k) - \mathcal{E}(x) = 1/2 \left\langle A(x^k - x), x^k - x \right\rangle = 1/2 \left\langle e^k, e^k \right\rangle$$

$$\Rightarrow \mathcal{E}(x^k) - \mathcal{E}(x^{k+1}) = (\mathcal{E}(x^k) - \mathcal{E}(x)) - (\mathcal{E}(x^{k+1}) - \mathcal{E}(x)) = -1/2\alpha_k^2 \left\langle Ag^k, g^k \right\rangle + \alpha_k \left\langle Ae^k, g^k \right\rangle$$

Nach der Definition von α_k gilt

$$\mathcal{E}(x^k) - \mathcal{E}(x^{k+1}) = 1/2 \frac{\left\langle g^k, g^k \right\rangle^2}{\langle Ag^k, g^k \rangle} \Rightarrow$$

$$\frac{\mathcal{E}(x^k) - \mathcal{E}(x^{k+1})}{\mathcal{E}(x^k) - \mathcal{E}(x)} = \frac{\left\langle g^k, g^k \right\rangle^2}{\langle Ag^k, g^k \rangle \cdot \langle Ae^k, e^k \rangle} = \frac{\left\langle g^k, g^k \right\rangle^2}{\langle Ag^k, g^k \rangle \cdot \langle A^{-1}g^k, g^k \rangle}$$

A ist positiv definit. Damit gilt $\lambda \langle \zeta, \zeta \rangle \leq \langle A\zeta, \zeta \rangle \leq \Lambda \langle \zeta, \zeta \rangle$. Somit haben wir

$$\frac{1}{\Lambda} \langle \zeta, \zeta \rangle \leq \left\langle A^{-1}\zeta, \zeta \right\rangle \leq \frac{1}{\lambda} \langle \zeta, \zeta \rangle$$

Es folgt hieraus

$$\frac{\mathcal{E}(x^k) - \mathcal{E}(x^{k+1})}{\mathcal{E}(x^k) - \mathcal{E}(x)} \geq \frac{\lambda}{\Lambda}$$

$$\Rightarrow \mathcal{E}(x^{k+1}) - \mathcal{E}(x) \leq -\frac{\lambda}{\Lambda}(\mathcal{E}(x^k) - \mathcal{E}(x)) \leq \underbrace{(1 - \frac{\lambda}{\Lambda})}_{=q<1}(\mathcal{E}(x^k) - \mathcal{E}(x))$$

$$\Rightarrow \mathcal{E}(x^k) - \mathcal{E}(x) \leq q^k(\mathcal{E}(x^0) - \mathcal{E}(x))$$

Dies bedeutet Konvergenz linearer Ordnung. Weiters gilt

$$\mathcal{E}(x^k) - \mathcal{E}(x) \to 0,\ k \to \infty\ ,\ 1/2 \left\langle Ae^k, e^k \right\rangle \geq \frac{\lambda}{2} \|e^k\|_2^2$$

$$\Rightarrow \|e^k\|_2^2 \leq \frac{2}{\lambda} q^k(\mathcal{E}(x^0) - \mathcal{E}(x)) \to 0$$

■

BEMERKUNG: Sei $\langle x, y \rangle_A = \langle x, Ay \rangle$ mit symmetrischem, positiv definiten A. Dann ist $\langle \cdot, \cdot \rangle_A$ ein Skalarprodukt mit Norm $\|x\|_A = \sqrt{\langle x, x \rangle_A} = \sqrt{\langle x, Ax \rangle}$. Wir haben gezeigt, dass

$$\|e^k\|_A^2 = \left\langle e^k, Ae^k \right\rangle = 2(\mathcal{E}(x^k) - \mathcal{E}(x^0))$$

Hieraus folgt $\|e^k\|_A^2 \leq q^k \|e^0\|_A^2$.

Wir wollen im Folgenden die Konvergenzgeschwindigkeit verbessern.

Lemma 5.41

(Lemma von Kantorowitsch) Sei A symmetrisch und positiv definit mit $\lambda = \min_i \lambda_i$ und $\Lambda = \max_i \lambda_i$. Dann gilt für alle $\zeta \neq 0$:

$$\frac{\langle \zeta, \zeta \rangle^2}{\langle A\zeta, \zeta \rangle \langle A^{-1}, \zeta, \zeta \rangle} \geq \frac{4\lambda\Lambda}{(\Lambda + \lambda)^2} \tag{5.42}$$

BEMERKUNG: Dies ist besser als $\frac{\lambda}{\Lambda}$.

Beweis : Da A positiv definit ist, existiert eine orthogonale Matrix Q mit $A = Q^T DQ$ und $D = \text{diag}(\lambda_1, \cdots, \lambda_n)$.

Somit gilt

$$I = \frac{\langle \zeta, \zeta \rangle^2}{\langle A\zeta, \zeta \rangle \langle A^{-1}\zeta, \zeta \rangle} \underbrace{=}_{Z=Q\zeta, A^{-1}=Q^T D^{-1} Q} \frac{\langle z, z \rangle^2}{\langle Dz, z \rangle \langle D^{-1}z, z \rangle} \text{ für } z \neq 0$$

Definiere nun $\alpha_i = \frac{z_i^2}{\|z\|_2^2}$, so gilt $\sum_i \alpha_i = 1,\ \alpha_i \geq 0$ und

$$F(z) = \frac{1}{(\sum_i \lambda_i \alpha_i)(\sum_i \frac{1}{\lambda_i} \alpha_i)}$$

Betrachte die Punkte $P_i = (\lambda_i, \frac{1}{\lambda_i})$.

[Graphik Cucumber_Kurve]

Setze $\overline{\lambda} = \sum_i \lambda_i \alpha_i$. Da $\lambda \mapsto \frac{1}{\lambda}$, liegen alle $P_1, \cdots, P_n$ unterhalb der Geraden $\overline{P_1 P_n}$, die beschrieben wird durch

$$\lambda \mapsto \frac{1}{\lambda_1} + \frac{\frac{1}{\lambda_n} - \frac{1}{\lambda_1}}{\lambda_n - \lambda_1}(\lambda - \lambda_1) = \frac{\lambda_1 + \lambda_n - \lambda}{\lambda_1 \lambda_n} = g(\lambda)$$

Sei $Q = (\sum_i \lambda_i \alpha_i, \sum_i \frac{1}{\lambda_i}\alpha_i) = \sum \alpha_i P_i$. Dann gilt $Q \in \text{conv}(P_1, \cdots, P_n)$ und damit

$$\sum_i \frac{1}{\lambda_i}\alpha_i \leq g(\underbrace{\sum_i \lambda_i \alpha_i}_{=\overline{\lambda}}) \leq \frac{\lambda_1 + \lambda_n - \overline{\lambda}}{\lambda_1 \lambda_n}$$

$$\Rightarrow F(z) = \frac{1}{\overline{\lambda} \sum \frac{1}{\lambda_i}\alpha_i} \geq \frac{\lambda_1 \lambda_n}{\overline{\lambda}(\lambda_1 + \lambda_n - \overline{\lambda})}$$

Schließlich erhalten wir

$$F(z) \geq \frac{\lambda_1 \lambda_n}{(\frac{\lambda_1 + \lambda_n}{2})^2} = \frac{4\lambda_1 \lambda_n}{(\lambda_1 + \lambda_n)^2}.$$

■

Theorem 5.43

Für das Gradientenverfahren gilt die Fehlerabschätzung

$$\|e^k\|_A^2 = \|x^k - x\|_A \leq (\frac{\kappa - 1}{\kappa + 1})^k \|e^0\|_A \tag{5.44}$$

für $k \in \mathbb{N}$. Dabei ist $\kappa = \frac{\Lambda}{\lambda} = cond_2(A)$.

Beweis : Im vorherigen Satz haben wir gezeigt, dass

$$\frac{\mathcal{E}(x^{k+1}) - \mathcal{E}(x)}{\mathcal{E}(x^k) - \mathcal{E}(x)} = (-\frac{\langle g^k, g^k \rangle^2}{\langle Ag^k, g^k \rangle \langle A^{-1}g^k, g^k \rangle} + 1)\frac{\mathcal{E}(x^k) - \mathcal{E}(x)}{\mathcal{E}(x^k) - \mathcal{E}(x)}$$

Mit

$$\mathcal{E}(x^k) - \mathcal{E}(x) = \frac{1}{2}\langle e^k, Ax^k \rangle = \frac{1}{2}\|e^k\|_A^2$$

folgt

$$\|e^{k+1}\|_A^2 \leq (1 - \frac{\langle g^k, g^k \rangle^2}{\langle Ag^k, g^k \rangle \langle A^{-1}g^k, g^k \rangle})\|e^k\|_A^2 \underbrace{\leq}_{\text{Kantor.}}$$

$$(1 - 4\frac{\lambda\Lambda}{(\lambda + \Lambda)^2})\|e^k\|_A^2 = (\frac{\lambda - \Lambda}{\lambda + \Lambda})^2 \|e^k\|_A^2$$

$$\Rightarrow \|e^{k+1}\|_A \leq \frac{\Lambda - \lambda}{\Lambda + \lambda}\|e^k\|_A \Rightarrow \|e^k\|_A \leq (\frac{\Lambda - \lambda}{\Lambda + \lambda})^k \|e^0\|_A = (\frac{\kappa - 1}{\kappa + 1})^k \|e^0\|_A$$

■

Im Gradientenverfahren gilt

$$g^k = Ax^k - b \text{ und } \alpha_k = -\frac{\langle g^k, r^k \rangle}{\langle Ar^k, r^k \rangle} \text{ und } x^{k+1} = x^k - \alpha_k r^k$$

$$\Rightarrow g^{k+1} = Ax^{k+1} - b = Ax^k - \alpha_k Ag^k - b = g^k - \alpha_k Ag^k \Rightarrow \left\langle g^{k+1}, g^k \right\rangle = \left\langle g^k, g^k \right\rangle - \alpha_k \left\langle Ag^k, g^k \right\rangle = 0$$

Das bedeutet: *Aufeinanderfolgende Richtungen sind orthogonal.*

[Graphik energy_sketch_niveau_ellipse]

Hier wird als ein Zickzackkurs beschrieben.

CG- Verfahren (conjugate gradient mathod)

Das Gradientenverfahren *denkt* nur von Schritt zu Schritt. Die Grundidee der CG-Verfahren ist die Folgende:

Finde sukzessive Abstiegsrichtungen d^k, welche $\langle\cdot,\cdot\rangle_A$-orthogonal sind.

Genauer heißt das: Sei $\{d^1,\cdots,d^n\}$ Basis von $\mathbb{R}^n$, die $\langle\cdot,\cdot\rangle_A$-orthogonal ist, d.h.

$$\left\langle d^j,d^k\right\rangle_A=\delta_{jk}\left\langle Ad^j,d^k\right\rangle,\ j,k=1,\cdots,n$$

Dann gilt für alle

$$y=\sum_{k=1}^{n}t_kd^k$$

$$\mathcal{E}(z+y)=\frac{1}{2}\langle A(z+y),z+y\rangle-\langle z+y,b\rangle=\frac{1}{2}\langle Az,z\rangle-\langle z,b\rangle+\frac{1}{2}\langle Ay,y\rangle+\langle Az,y\rangle-\langle b,y\rangle=$$

$$=\mathcal{E}(z)+\frac{1}{2}\langle Ay,y\rangle+\langle Az-b,y\rangle=$$

$$=\mathcal{E}(z)+\frac{1}{2}\sum_{j,k=1}^{n}t_jt_k\left\langle Ad^j,d^k\right\rangle+\sum_{j=1}^{n}t_j\left\langle Az-b,d^j\right\rangle=\mathcal{E}(z)+\sum_{j=1}^{n}(\frac{1}{2}\left\langle Ad^j,d^j\right\rangle t_j^2-\left\langle Az-b,d^j\right\rangle)$$

Sei

$$F_j(t_j)=\frac{1}{2}\left\langle Ad^j,d^j\right\rangle t_j^2-\left\langle Az-b,d^j\right\rangle,$$

so ist

$$\mathcal{E}(z+y)=\mathcal{E}(z)+\sum_{j=1}^{n}F_j(t_j)$$

Damit ist das Problem entkoppelt bzgl. der t_j und man kann nacheinander $F_j(t_j)$ minimieren, um $\mathcal{E}(z+y)$ zu minimieren. Die $d^1,\cdots,d^n$ kann man sukzessive aus den Richtungen $g^1,\cdots,g^n$ bestimmen (nutze Gram-Schmidt). Wir formulieren nun den *Modellalgorithmus*:

- $x^1\in\mathbb{R}^n$ Startwert
- $g^1=\nabla\mathcal{E}(x^1)$
- $d^1=-g^1$ (steilste Abstiegsrichtung)
- Für $k\geq 2$: $g^k=Ax^k-b=\nabla\mathcal{E}(x^k)$. Falls $g^k=0\Rightarrow$ Stopp (bzw. $\|g^k\|\leq\varepsilon$)
- Sonst für $k\geq 2$: Bestimme d^k aus $\operatorname{span}\{g^1,\cdots,g^k\}$ mit $\left\langle d^k,d^j\right\rangle_A=0$ für $j=1,\cdots,k-1$. Wir minimieren in Richtung d^k, also $t_k=\operatorname{argmin}\mathcal{E}(x^k+t_kd^k)$. Somit $x^{k+1}=x^k+t_kd^k$.

Da $d^k \in \text{span}\{g^1, \cdots, g^k\} = V_k$, gilt $x^{k+1} \in x^1 + V_k$. Da $t_k = \text{argmin}\mathcal{E}(x^k + t_k d^k)$ folgt

$$(\delta\mathcal{E})(x^k + t_k d^k)(d^k) = 0$$

gilt

$$\left\langle \underbrace{A(x^k + t_k d^k) - b}_{=g^{k+1}}, d^k \right\rangle = 0$$

$$\Rightarrow \left\langle g^k, d^k \right\rangle = \left\langle Ax^k - b, d^k \right\rangle = -t_k \left\langle Ad^k, d^k \right\rangle = 0$$

$$\Rightarrow t_k = -\frac{\langle g^k, d^k \rangle}{\langle Ad^k, d^k \rangle} \tag{5.45}$$

Ferner ist

$$g^{k+1} = A(x^k + t_k d^k) - b = g^k + t_k A d^k \tag{5.46}$$

Bemerkung: Das CG-Verfahren bricht spätestens nach n Schritten ab, da dann die $\{d^1, \cdots, d^n\}$ den ganzen Raum $\mathbb{R}^n$ aufspannen. Dies gilt nur *ohne* Rundungsfehler. Dennoch haben wir schon nach wenigen Schritten $<< n$ eine gute Approximationslösung.

Konvergenzgeschwindigkeit des CG-Verfahrens

Da $V_k = \text{span}\{d^1, \cdots, d^k\} = \text{span}\{g^1, \cdots, g^k\}$ und $g^{k+1} = g^k + \alpha_t A d^k \in \text{span}\{g^k, A\text{span}\{g^1, \cdots, g^k\}\}$, folgt leicht (z.B. mit Induktion)

$$V_k = \text{span}\{g^1, Ag1, \cdots, A^{k-1} g^k\} = \{p(A)g^1 \colon p \in \mathcal{P}\}$$

Dies folgt aber auch aus dem Wissen, dass $x^k + t_k d^k$ die Energie auf $x^0 + V_k$ minimiert.