Einführung in die Numerische Mathematik

Vorlesung vom SS 2010

Mario Ohlberger

Institut für Numerische und Angewandte Mathematik
Fachbereich Mathematik und Informatik
Westfälische Wilhelms-Universität Münster

Dieses Skript beruht auf meinen Vorlesungen *Einführung in die Numerische Mathematik* und *Höhere Numerische Mathematik* vom Wintersemester 2007/2008 und Sommersemester 2008 an der Westfälische Wilhelms-Universität Münster.
Es besteht keine Garantie auf Richtigkeit und/oder Vollständigkeit des Manuskripts.

Mario Ohlberger

Inhaltsverzeichnis

Abbildungsverzeichnis

Kapitel 0

Einleitung

Die Numerische Mathematik, oder auch Numerik genannt, beschäftigt sich mit der numerischen Lösung endlichdimensionaler Probleme, sowie mit der Approximation unendlichdimensionaler Probleme durch endlichdimensionale. Die Numerik ist somit eine mathematische Schlüsseldisziplin zur Behandlung von Anwendungsproblemen mit Hilfe des Computers.
Der Numerik geht stehts die Modellierung voraus, deren Ziel es ist ein Anwendungsproblem in der mathematischen Sprache zu formulieren. Dieses Prozedere wird in der Abbildung 1 verdeutlicht.

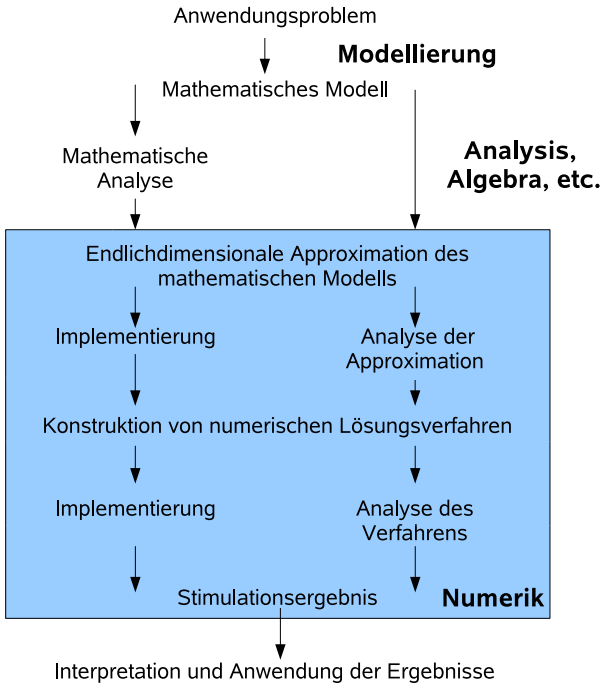

Abbildung 1: Illustration der Vorgehensweise zur Lösung eines Anwendungsproblems

Im folgenden wollen wir an einem einfachen Anwendungsbeispiel dieses Vorgehen skizzieren.

Beispiel 0.1 (Berechnung des Wärmetransports in einem Draht)

Schritt 1: Modellierung

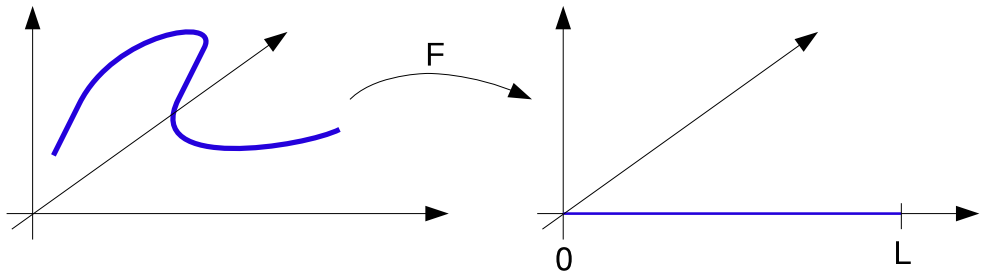

Abbildung 2: Koordinatentranformation zur mathematischen Betrachtung des Wärmetransports in einem Draht.

Wir betrachten den Wärmetransport in einem Draht. Nach einer Koordinatentransformationen, wie sie in Abbildung 2 skizziert ist, können wir den Draht eindimensional durch ein Intervall $I = [0, L]$ repräsentieren. Nach dem Fick'schen Gesetz gilt für die Wärmeleitung, dass der Wärmefluß q proportional zum Gradienten der Temperatur T ist, d.h.

$$q(x,t) = -\sigma \partial_x T(x,t), \forall x \in I, \forall t \in [0, T_{\max}].$$

Dabei bezeichnet $\sigma > 0$ die Wärmeleitfähigkeit und ist eine Materialkonstante. Außerdem gilt für jedes Teilintervall $[a, b] \in I$ und jeden Zeitabschnitt $[t_1, t_2] \in [0, T_{\max}]$:

$$\int_{[a,b]} T(x,t_2)dx - \int_{[a,b]} T(x,t_1)dx = \int_{[t_1,t_2]} q(a,t)dt - \int_{[t_1,t_2]} q(b,t)dt.$$

Sind die Temperatur und der Wärmefluß genügend oft differenzierbar, so folgt mit dem Hauptsatz der Differential- und Integralrechnung

$$\int_{[a,b]} \int_{[t_1,t_2]} \partial_t T(x,t)dxdt = -\int_{[a,b]} \int_{[t_1,t_2]} \partial_x q(x,t)dxdt.$$

Da dies für beliebige a, b, t_1, t_2 gilt, folgt mit dem Hauptsatz der Variationsrechnung

$$\partial_t T(x,t) = -\partial_x q(x,t) = \sigma\ \partial_{xx} T(x,t), \qquad \forall (x,t) \in I \times [0, T_{\max}].$$

Wir erhalten so als mathematisches Modell für die Wärmeleitung in einem Draht eine partielle Differentialgleichung, d.h. eine Gleichung, die die partiellen Ableitungen einer Funktion miteinander in Beziehung setzt. Eine mathematische Analyse zeigt, dass diese sogenannte Wärmeleitungsgleichung eine eindeutige Lösung T besitzt, falls man beispielsweise die Temperatur zum Zeitpunkt $t = 0$ und an den Endpunkten $x = 0, L$ vorgibt. Da die gesuchte Temperaturverteilung T eine differenzierbare Funktion in zwei Veränderlichen darstellt, handelt es sich hierbei um ein unendlichdimensionales Problem. Im nächsten Schritt wollen wir mit Hilfe von sogenannten Finite Differenzenverfahren eine endlichdimensionale Approximation angeben.

Schritt 2: Endlichdimensionale Approximation

Zur Approximation der Wärmeleitungsgleichungen führen wir Zerlegungen $T_h := \{x_i | x_i = hi, i = 0, ..., N+1\}$ und $J_k := \{t_n | t_n = kn, n = 0, ..., M\}$ des Ortsintervalls $[0, L]$ und des Zeitintervalls $[0, T_{\max}]$ ein. Dabei ist $h := L/(N+1)$ die Ortsschrittweite und $k := T_{\max}/M$ die Zeitschrittweite der jeweiligen Zerlegung.
Die Idee der Finite Differenzen besteht darin, alle Ableitungen in der Wärmeleitungsgleichungen durch Differezenquotienten zu ersetzen. Verwenden wir z.B.

$$\partial_t T(x,t_n) \approx (T(x,t_n) - T(x,t_{n-1}))/k$$

und

$$\partial_{xx} T(x_i, t) \approx (T(x_{i+1}, t) - 2T(x_i, t) + T(x_{i-1}, t))/h^2,$$

so erhalten wir für die Approximation T_i^n von $T(x_i, t_n)$ die Gleichung

$$(T_i^n - T_i^{n-1})/k = \sigma(T_{i+1}^n - 2T_i^n + T_{i-1}^n)/h^2, \quad \forall i = 1, ..., N, n = 1, ..., M.$$

Dies ist eine lineare Gleichung für T_i^n, die jedoch mit den Linearen Gleichungen des selben Typs für die ebenfalls unbekannten Werte T_i^{n-1} und T_{i+1}^n, T_{i-1}^n gekoppelt ist. Für die Anfangs- und Randwerte verwenden wir die vorgegebenen Werte der Temperatur, d.h.

$$T_i^0 := T(x_i, 0), \qquad T_0^n := T(0, t_n), \qquad T_{N+1}^n := T(L, t_n).$$

Berechnet man sukzessive zunächst die Lösung zu den Zeitpunkten $t_1, t_2, t_3, ...$, so erhält man für jeden dieser Zeitschritte t_n ein lineares Gleichungssystem mit N Gleichungen für die Unbekannten T_i^n, $i = 1, ..., N$. Wir haben das unendlichdimensionale Problem also durch das sukzessive Lösen von linearen Gleichungssystemen approximiert. Verfahren zur Approximation von Differentialgleichungen werden in Kapitel 6 vorgestellt und in der Vorlesung *Höhere Numerische Mathematik* im Wintersemester detailliert analysiert.

Schritt 3: Numerische Lösung des endlichdimensionalen Problems

Im letzten Lösungsschritt geht es darum die linearen Gleichungssysteme numerisch zu lösen. Hierzu kann z.B. die Gaußelimination verwendet werden. Falls jedoch die Dimension N des Systems sehr groß wird, sind andere Verfahren besser geeignet. Im zweiten Kapitel der Vorlesung wenden wir uns daher der numerischen Lösung von linearen Gleichungssystemen zu.

Wäre in unserem Beispiel der Wärmeleitfähigkeitskoeffizient temperaturabhängig, d.h. $\sigma = \hat{\sigma}(T(x, t))$, mit einer nichtlinearen Funktion $\hat{\sigma}$, so wäre das resultierende Gleichungssystem nichtlinear. Solchen Problemen ist das Kapitel 3 gewidmet.
Wählt man zur Approximation der Differentialgleichungen andere Verfahren, wie z.B. Finite Elemente Verfahren, so ist auch die numerische Approximation von Funktionen durch Polynome und die numerische Berechnung von Integralen Bestandteil der Verfahren. Diese Themen werden in der Vorlesung *Höhere Numerische Mathematik* im Wintersemester detailliert analysiert.

Das Beispiel der Wärmeleitung in einem Draht verdeutlicht die Notwendigkeit von numerischen Methoden, wie z.B. die Lösung linearer, oder nichtlinearer Gleichungssysteme, oder die Approximation von Differentialgleichungen. Bevor wir uns diesen Methoden zuwenden, werden im nächsten Kapitel jedoch einige Grundlagen dargestellt. Hierbei handelt es sich zum einen um eine Zusammenstellung wichtiger Begriffe auf der Analysis und Linearen Algebra und zum anderen wird auf die numerische Approximation reeller Zahlen eingegangen und daraus resultierende Fehlerquellen diskutiert.

Kapitel 1

Grundlagen

Im folgenden Abschnitt werden wir Definitionen angeben, die wir in den weiteren Kapiteln benötigen.

1.1 Normierte Räume

Seien $\mathbb{K} = \mathbb{R}$ oder $\mathbb{K} = \mathbb{C}$ ein Körper und V ein Vektorraum über $\mathbb{K}$.

Definition 1.1 (Norm)
Eine Abbildung $\|\cdot\| : V \longrightarrow \mathbb{R}$ *heißt* **Norm**, *falls gilt*

(i) $\|v\| > 0 \quad \forall v \in V \backslash \{0\}$,

(ii) $\|\lambda v\| = |\lambda| \, \|v\| \quad \forall \lambda \in \mathbb{K}, \ \forall v \in V$,

(iii) $\|v + w\| \leq \|v\| + \|w\| \quad \forall v, w \in V$.

Beispiel 1.2
Sei $V = \mathbb{R}^n, v = (v_1, \ldots, v_n) \in \mathbb{R}^n$. Dann ist
$\|v\|_\infty := \max_{1 \leq i \leq n} |v_i|, \quad \|v\|_1 := \sum_{i=1}^{n} |v_i|,$
$\|v\|_2 = \left(\sum_{i=1}^{n} |v_i|^2 \right)^{\frac{1}{2}}, \quad \|v\|_p := \left(\sum_{i=0}^{n} |v_i|^p \right)^{\frac{1}{p}} \quad (1 \leq p < \infty).$

Beispiel 1.3
Sei $V = C^0(I), I = [a, b] \subset \mathbb{R}$. Dann ist
$\|v\|_\infty := \sup \{|v(x)| \mid x \in I\},$
$\|v\|_p := \left(\int_a^b |v(x)|^p \, \mathrm{d}x \right)^{\frac{1}{p}}.$

Definition 1.4 (Normierter Raum)
Ein Vektorraum V *zusammen mit einer Norm* $\|\cdot\|$, *geschrieben* $(V, \|\cdot\|)$, *heißt* **normierter Raum**.

Definition 1.5 (Banachraum)
Eine Folge $(u_n)_{n\in\mathbb{N}} \subset V$ *konvergiert gegen* $u \in V \quad :\Longleftrightarrow$
$\forall \varepsilon > 0\ \exists N\ \forall\ n > N : \|u_n - u\| < \varepsilon.$

Eine Folge $(u_n)_{n\in\mathbb{N}} \subset V$ *heißt* **Cauchy Folge** $:\Longleftrightarrow$
$\forall\ \varepsilon > 0\ \exists\ N\ \forall\ m, l > N : \|u_l - u_m\| < \varepsilon.$

Ein normierter Raum $(V, \|\cdot\|)$ *heißt vollständig, falls alle Cauchy-Folgen in* V *bzgl.* $\|\cdot\|$ *in* V *konvergieren. Ein vollständiger normierte Raum heißt* **Banachraum**.

Beispiel 1.6
$(\mathbb{R}^n, \|\cdot\|)$ ist ein Banachraum für alle $\|\cdot\|$,
$\left(C^0(I), \|\cdot\|_\infty\right)$ ist ein Banachraum, $\left(C^0(I), \|\cdot\|_p\right)$ ist dagegen nicht vollständig

Satz 1.7
Sei $\dim V < \infty$, $\|\cdot\|_a$ und $\|\cdot\|_b$ zwei Normen. Dann existieren $m, M \in \mathbb{R} : m\|v\|_a \leq \|v\|_b \leq M\|v\|_a\ \forall\ v \in V$, d.h. $\|\cdot\|_a$ und $\|\cdot\|_b$ sind **äquivalente Normen**.

Definition 1.8 (Skalarprodukt)
Eine Abbildung $\langle\cdot,\cdot\rangle : V \times V \longrightarrow \mathbb{C}$ *heißt* **Skalarprodukt**, *falls gilt*

(i) $\forall\ v \in V\backslash\{0\} :\ \langle v, v\rangle \geq 0,$

(ii) $\forall\ u, v \in V\ \langle u, v\rangle = \overline{\langle v, u\rangle},$

(iii) $\forall\ u, v, w \in V \forall\ \alpha \in \mathbb{K} :$
$\langle \alpha u, v\rangle = \alpha\,\langle u, v\rangle,$
$\langle u + v, w\rangle = \langle u, w\rangle + \langle v, w\rangle.$

Folgerung: $\langle u, \alpha v\rangle = \overline{\alpha}\,\langle u, v\rangle\ ,\ \langle u, v + w\rangle = \langle u, v\rangle + \langle u, w\rangle.$

Satz 1.9 (Induzierte Norm)
Sei $\langle\cdot,\cdot\rangle$ ein Skalarprodukt, dann wird durch $\|v\| := \sqrt{\langle v, v\rangle}$ eine Norm induziert.

Definition 1.10 (Hilbertraum)
Ein Vektorraum mit Skalarprodukt heißt **Prähilbertraum** *, falls* V *mit der induzierten Norm* ***nicht*** *vollständig ist, sonst bezeichnet* V *einen* **Hilbertraum**.

Beispiel 1.11 (Cauchy-Schwarz-Ungleichung)
$\forall\ u, v \in V :\ |\langle u, v\rangle| \leq \sqrt{\langle u, u\rangle\,\langle v, v\rangle}$, Gleichheit $\iff$ u, v linear abhängig.

Beispiel 1.12

Sei $V = \mathbb{R}^n$, $\langle u, v\rangle := \sum_{i=1}^{n} u_i v_i$ ist ein Skalarprodukt und induziert die **euklidische Norm**

$$\|v\|_2 := \left(\sum_{i=1}^{n} |v_i|^2\right)^{\frac{1}{2}}.$$

1.2 Operatoren

Definition 1.13
U, V normierte Vektorräume, $D \subseteq U$. Wir bezeichnen eine Abbildung $T : D \longrightarrow V$ als **Operator**. *Dabei gilt:*

(i) T *heißt* **stetig** *in* $u \in D \iff$
$\forall\, \varepsilon > 0\ \exists\, \delta > 0\ \forall\, v \in D : \|u - v\|_U < \delta \implies \|T(u) - T(v)\|_V < \varepsilon.$

(ii) T *heißt stetig in* $D \iff$
T *ist stetig für alle* $u \in D$.

(iii) T *heißt* **Lipschitz-stetig** $\iff$
es existiert ein $L > 0\ \forall\, u, v \in D :\ \|T(u) - T(v)\|_V \leq L \|u - v\|_U$.

Bemerkung 1.14
Es ist leicht zu sehen, dass aus **(iii)** **(ii)** folgt und aus **(ii)** folgt **(i)**.

Definition 1.15
T *heißt* **linearer Operator** *(oder einfach linear), falls* $\forall\, u, v \in V, \alpha \in \mathbb{K}$:

(i) $T(u + v) = T(u) + T(v)$,

(ii) $T(\alpha v) = \alpha T(v)$.

Bemerkung 1.16
Ist T linear, so schreibt man häufig Tu statt $T(u)$.

Beispiel 1.17
$V = W = \mathbb{R}^n, A \in \mathbb{R}^{n\times n} :\ Tu = Au$ ist ein linearer Operator.
$V = C^0(I), W = \mathbb{R} :\ Tu := \int_a^b u(x)\mathrm{d}x$ ist ein linearer Operator.

Definition 1.18
$T : U \longrightarrow V$ *sei ein Operator.* T *heißt* **beschränkt**, *falls es ein* $C > 0$ *gibt, so dass* $\forall\, u \in U :\ \|T(u)\|_V \leq C \|u\|_U$.

Satz 1.19

Für einen linearen Operator $T : U \longrightarrow V$ sind äquivalent:

(i) T ist beschränkt,

(ii) T ist Lipschitz-stetig,

(iii) T ist stetig in 0.

Beweis: Siehe Übungsblatt 1 □

Bemerkung 1.20

(i) $\dim U < \infty, \dim V < \infty$, dann sind alle linearen Operatoren beschränkt und damit stetig.

(ii) Auf unendlich-dimensionalen Vektorräumen existieren auch unbeschränkte lineare Operatoren.

(iii) Die Aussage von Satz 1.19 (Seite 7) ist nur richtig für lineare Operatoren.
Bsp: $T : \mathbb{R} \longrightarrow \mathbb{R},\ x \longmapsto x^2$: es existiert keine Konstante C mit $\left|x^2\right| < C\,|x|\ \forall\, x \in \mathbb{R}$.

Definition 1.21
Mit $\mathrm{B}(U,V)$ *bezeichnen wir den Raum der beschränkten linearen Operatoren.* $\mathrm{B}(U,V)$ *ist ein Vektorraum. Durch*

$$\|T\|_{U,V} := \sup_{u \in U \setminus \{0\}} \frac{\|Tu\|_V}{\|u\|_U}$$

wird eine Norm auf $\mathrm{B}(U,V)$ *definiert. Diese wird als die durch* $\|\cdot\|_U, \|\cdot\|_V$ **induzierte Operatornorm** *bezeichnet.*

Folgerung 1.22

(i) $\|T\|_{U,V} = \sup\limits_{\substack{u \in U \\ \|u\|_U = 1}} \|Tu\|_V$ (wegen Linearität von T).

(ii) $\|Tu\|_V \leq \|T\|_{U,V} \|u\|_U$ und $\|T\|_{U,V}$ ist die kleinste Konstante mit dieser Eigenschaft für alle $u \in U$ (folgt aus der Definition).

(iii) $\|id\|_{U,U} = 1$, dabei ist $id \in \mathrm{B}(U,U)$ mit $id : u \longmapsto u$.

Beispiel 1.23
$U, V = \mathbb{R}^n$, dann entspricht $\mathrm{B}(U,V)$ dem Raum der $n \times n$ Matrizen. Daher wird die Operatornorm auch häufig Matrixnorm genannt.

Satz 1.24

Die induzierte (Matrix-) Operatornorm ist submultiplikativ, d.h. $\|A \circ B\| \leq \|A\| \cdot \|B\|$

Beweis: (gilt nur für die induzierte Matrixnorm)
$\|(A \circ B)x\| = \|A(B(x))\| \overset{1.22ii}{\leq} \|A\| \, \|Bx\| \overset{1.22ii}{\leq} \|A\| \, \|B\| \, \|x\|.$
Sei $x \neq 0 \implies \frac{\|ABx\|}{\|x\|} \leq \|A\| \, \|B\|$ □ □

Bemerkung 1.25
Die induzierten Operatornormen ergeben nicht alle Normen auf $\mathrm{B}(U, V)$. Sei etwa $A \in \mathbb{R}^{n \times n}, A = (a_{ij})$, dann wird durch $\|A\| = \sup\limits_{1 \leq i,j \leq n} |a_{ij}|$ eine Norm definiert, die nicht induziert ist.

Beispiel 1.26
Die durch $\|\cdot\|_1$ und $\|\cdot\|_\infty$ induzierte Operatornormen werden in den Übungen behandelt.

Sei $A : (\mathbb{R}^n, \|\cdot\|_2) \longrightarrow (\mathbb{R}^n, \|\cdot\|_2)$, dann gilt $\|A\|_{2,2} = \sqrt{\lambda_{\max}(A^*A)}$, wobei $\lambda_{\max}(B)$ für $B \in \mathbb{R}^{n \times n}$ den betragsmäßig größten **Eigenwert (EW)** bezeichnet. Sei $A = (a_{ij})$, dann ist $A^* = \overline{A}^\top$. Diese Norm wird als **Spektralnorm** bezeichnet. Ist $A \in \mathbb{R}^{n \times n}$, dann ist $\overline{A}^\top = A^\top$.

Beweis: **Bemerkungen:** $(A^*A)^* = A^*A \implies A^*A$ ist hermitesch $\implies$ alle Eigenwerte sind reell.
Es gilt $x^* (AA^*) x = (Ax)^* Ax = \langle Ax, Ax \rangle \geq 0$. Also ist A^*A positiv definitund somit alle EW positiv.

Da A^*A hermitesch ist, existiert ein Matrix $U \in \mathbb{C}^{n \times n}$ mit $U^*U = id$ (d.h. U ist unitär) und

$$U^*(A^*A)U = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} =: D \qquad (*)$$

Sei u_i die i-te Spalte von U, d.h. $U = (u_1, \ldots, u_n)$ und $\|u_i\|_2 = 1 \; \forall i \in \{1, \ldots, n\}$, dann ist

$$A^*AU = UD,$$

da $U^{-1} = U^*$ und somit $A^*Au_i = \lambda_i u_i$. Also sind die Vektoren u_i Eigenvektoren (EV) von A^*A zu den Eigenwerten λ_i. Aus $(*)$ folgt dann weiter $u_i^* A^* A u_i = \lambda_i$.
Sei $x \in \mathbb{C}^n$ mit $\|x\|_2 = 1$. Setze $y = U^*x$, so folgt wegen $x = Uy$:

$$\begin{aligned} \|Ax\|_2^2 &= \langle Ax, Ax \rangle = \langle x, A^*Ax \rangle \overset{(*)}{=} \langle x, UDU^*x \rangle \\ &= \langle x, UDy \rangle = \langle U^*x, Dy \rangle = \langle y, Dy \rangle \\ &= \sum_{i=1}^n \overline{y_i} \lambda_i y_i \overset{\lambda_i > 0}{\leq} \max_{1 \leq i \leq n} \lambda_i \sum_{i=1}^n y_i^2 \\ &= \lambda_{\max}(A^*A) \, \|y\|_2^2 = \lambda_{\max}(A^*A). \end{aligned}$$

da $\|y\|_2 = \|U^*x\| = \|x\| = 1$ (U^* unitär).

Also gilt $\|Ax\|_2 \leq \sqrt{\lambda_{\max}(A^*A)}$ für alle x mit $\|x\|_2 = 1$ und somit folgt

$$\|A\|_{2,2} \leq \sqrt{\lambda_{\max}(A^*A)}.$$

Sei nun $\lambda_i = \lambda_{\max}(A^*A)$ der größte Eigenwert und u_i der zugehörige Eigenvektor mit $\|u_i\|_2 = 1$. Da $\|A\|_{2,2} = \sup_{\|x\|_2 = 1} \|Ax\|_2$ folgt $\|A\|_{2,2} \geq \|Au_i\|_2$ und somit

$$\begin{aligned} \|A\|_{2,2}^2 &\geq \|Au_i\|_2^2 = \langle Au_i, Au_i\rangle = \langle u_i, A^*Au_i\rangle \\ &= \langle u_i, \lambda_i u_i\rangle = \lambda_i \langle u_i, u_i\rangle = \lambda_i \|u_i\|_2^2 = \lambda_i. \end{aligned}$$

Also folgt die Behauptung. □

1.3 Banachscher Fixpunktsatz

Definition 1.27 (Kontraktion)
Sei $D \subset X$, X normierter Verktorraum, Y normierter Vektorraum. Dann heißt ein Operator $T : D \longrightarrow Y$ eine **Kontraktion**, *falls T Lipschitz-stetig mit Lipschitz-Konstante $0 < L < 1$ ist, d.h. $\forall\, u, v \in D : \|T(u) - T(v)\|_Y \leq L \|u - v\|_X$.*

Definition 1.28 (Fixpunkt)
Sei $T : D \longrightarrow D$ ein Operator, der D auf sich selbst abbildet. Dann heißt $\overline{u} \in D$ **Fixpunkt** *von T in D, falls $T(\overline{u}) = \overline{u}$.*

Satz 1.29 (Banachscher Fixpunktsatz)
Sei X ein Banachraum, $D \subseteq X$ abgeschlossen, $T : D \longrightarrow D$ eine Kontraktion. Dann gilt:

(i) T hat genau einen Fixpunkt $\overline{u} \in D$.

(ii) Sei $u_o \in D$ beliebig und $u_{k+1} := T(u_k), k = 0, 1, \ldots \implies u_k \longrightarrow \overline{u}$.

(iii) $\|\overline{u} - u_k\| \leq L \|\overline{u} - u_{k-1}\|$ $(k \geq 1)$, d.h. der Fehler nimmt monoton ab.

(iv) $\|\overline{u} - u_k\| \leq \frac{L^k}{1-L} \|T(u_0) - u_0\|$ $(k \geq 1)$. (**a-priori Abschätzung**)

(v) $\|\overline{u} - u_k\| \leq \frac{L}{1-L} \|u_k - u_{k-1}\|$ $(k \geq 1)$. (**a-posteriori Abschätzung**)

Beweis: zu a): Wir zeigen zunächst, dass T höchstens einen Fixpunkt hat. Dazu nehmen wir zunächst an, dass T zwei Fixpunkte $\bar{u}, \bar{v}$ hätte. Dann folgt aus $T\bar{u} = \bar{u}$ und $T\bar{v} = \bar{v}$:

$$||\bar{u} - \bar{v}|| = ||T\bar{u} - T\bar{v}|| \leq L||\bar{u} - \bar{v}||.$$

Da nach Voraussetzung $L < 1$ ist, folgt $||\bar{u} - \bar{v}|| = 0$ und somit $\bar{u} = \bar{v}$.
Um die Existenz eines Fixpunktes zu beweisen, reicht es Teil b) zu zeigen.

zu b): Die Folge $(u_k)_{k\in\mathbb{N}}$ sei für beliebiges $u_0 \in D$ definiert durch die Fixpunktiteration

$$u_{k+1} = Tu_k.$$

Wir zeigen, dass $(u_k)_{k\in\mathbb{N}}$ Cauchy-Folge ist. Die Konvergenz gegen ein $\bar{u}$ folgt dann aus der Vollständigkeit von X. Mit vollständiger Induktion zeigen wir zunächts, dass für $k \geq 1$ gilt

$$||u_{k+1} - u_k|| \leq L^k||T(u_0) - u_0||. \qquad (*)$$

Induktionsanfang $(k = 1)$:
Es ist $||u_1 - u_0|| = ||Tu_0 - u_0|| = L^0||Tu_0 - u_0||$.

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

$$\begin{aligned} ||u_{k+2} - u_{k+1}|| &= ||Tu_{k+1} - Tu_k|| \\ &\leq L||u_{k+1} - u_k|| \\ &\overset{\text{Ind. Vor. } (*)}{\leq} LL^k||T(u_0) - u_0|| \\ &= L^{k+1}||T(u_0) - u_0||. \end{aligned}$$

Sei nun $m < n$ so folgt hieraus

$$\begin{aligned} ||u_n - u_m|| &= ||\sum_{k=m}^{n-1}(u_{k+1} - u_k)|| \\ &\leq \sum_{k=m}^{n-1}||(u_{k+1} - u_k)|| \\ &\leq \sum_{k=m}^{n-1} L^k||T(u_0) - u_0|| \\ &\leq L^m||T(u_0) - u_0|| \sum_{k=0}^{n-1-m} L^k \\ &\overset{\text{geom. Reihe}}{\leq} \frac{L^m}{1-L}||T(u_0) - u_0||. \end{aligned}$$

Da $L < 1$ ist, konvergiert die rechte Seite gegen Null für $m \to \infty$ und wir haben somit gezeigt, dass $(u_k)_{k\in\mathbb{N}}$ Cauchy-Folge ist. Dies beweist b) und somit auch a).

Die Teile c)-d) werden in den Übungen behandelt. □

Bemerkung 1.30
Satz 1.29 (iii) (Seite 10) gibt eine a-priori Schranke, die man nutzen kann, um einen Index k_0 zu bestimmen mit $\|\overline{u} - u_{k_0}\| \leq TOL$ für eine gegebene Toleranz $TOL > 0$:

Sei TOL gegeben, O.B.d.A $TOL < 1$

$$\begin{aligned} \|u_{k_0} - \overline{u}\| &\leq \tfrac{L^{k_0}}{1-L}\|T(u_0) - u_0\| \leq TOL \\ &\iff L^{k_0} \leq (1-L)\tfrac{TOL}{\|T(u_0)-u_0\|} \\ &\iff k_0 \log L \leq \log(1-L) + \log TOL - \log\left(\|T(u_0) - u_0\|\right) \\ &\iff k_0 \geq \tfrac{\log(1-L)+\log TOL - log(\|T(u_0)-u_0\|)}{\log L}, \end{aligned}$$

da $0 < L < 1$ und somit $\log L < 0$. Meistens ist dies eine Überschätzung des Aufwands.

Satz 1.29 (iv) (Seite 10) kann als Abbruchkriterium während der Iteration benutzt werden, d.h. man bricht ab, falls $\frac{L}{1-L}\|u_k - u_{k-1}\| < TOL$ ist.

1.4 Taylorreihe

Definition 1.31
Sei $C^0(I), I = (a.b)$ *der Raum der* **stetigen Funktionen auf I** . *Mit* $C^m(I) := \{f : I \longrightarrow \mathbb{R} \mid f, f', f'', \ldots, f^{(m)}$ ex. und sind stetig$\}$ *bezeichnen wir den Raum der* ***m*-mal stetig differenzierbaren Funktionen**.

Kurzschreibweise: $C^m(a,b)$ *statt* $C^m((a,b))$. *Mit der Definition* $C^\infty(I) := \bigcap_{m\in\mathbb{N}} C^m(I)$ *folgt dann*

$$C^\infty(I) \subset \ldots \subset C^m(I) \subset \ldots \subset C^0(I).$$

Satz 1.32 (Taylorreihe mit Lagrange Restglied)
Seien $f \in C^{m+1}(a,b)$ und $x_0 \in (a,b)$ fest. Dann existiert für jedes $x \in (a,b)$ ein ξ zwischen x_0 und x mit

$$f(x) = \sum_{k=0}^{m} \frac{1}{k!} f^{(k)}(x_0)(x-x_0)^k + R_m(x),$$

mit $R_m(x) := \frac{1}{(m+1)!} f^{(m+1)}(\xi)(x-x_0)^{m+1}$.

Satz 1.33 (Taylorreihe mit Integralrestterm)
Seien $f \in C^{m+1}(a,b),\ x_0 \in (a,b)$ fest. Dann gilt für jedes $x \in (a,b)$

$$f(x) = \sum_{k=0}^{m} \frac{1}{k!} f^{(k)}(x_0)(x-x_0)^k + R_m(x),$$

mit $R_m(x) := \frac{1}{m!} \int_{x_0}^{x} f^{(m+1)}(t)(x-t)^m \mathrm{d}t$.

Beweis: Für beide Sätze siehe Analysis I. □

Folgerung 1.34 (Häufig verwendete Form)
Seien $f \in C^{m+1}(x_0 - h_0, x_0 + h_0),\ x_0 \in \mathbb{R}, h_0 > 0$. Sei $|h| \leq h_0$, dann existiert eine Abbildung $\omega_m : (-h_0, h_0) \longrightarrow \mathbb{R}$ mit $\lim_{h \longrightarrow 0} \omega_m(h) = 0$, so dass gilt

$$f(x_0+h) = f(x_0) + \sum_{k=1}^{m} \frac{f^{(k)}(x_0)}{k!} h^k + \omega_m(h) h^m.$$

Beweis: Wende Satz 1.32 (Seite 12) an mit $x = x_0 + h$, d.h. es existiert ein ξ mit $|\xi| < |h|$ und

$$\begin{aligned} f(x_0+h) - \sum_{k=0}^{m-1} \tfrac{f^{(k)}(x_0)}{k!} h^k &= \tfrac{f^{(m)}(x_0+\xi)}{m!} h^m \\ &= \tfrac{f^{(m)}(x_0)}{m!} h^m - \tfrac{f^{(m)}(x_0)}{m!} h^m + \tfrac{f^{(m)}(x_0+\xi)}{m!} h^m \\ &= \tfrac{f^{(m)}(x_0)}{m!} h^m + \omega_m(h) h^m \end{aligned}$$

mit $\omega_m(h) = \frac{f^{(m)}(x_0+\xi)-f^{(m)}(x_0)}{m!}$.
Da $|\xi| < |h|$ und $f^{(m)}$ stetig, folgt $\lim_{h \longrightarrow 0} \frac{f^{(m)}(x_0+\xi)-f^{(m)}(x_0)}{m!} = 0$.

□

Definition 1.35
Die Funktion $f \in C^1(x_0 - h_0, x_0 + h_0)$ ist in **erster Näherung** *gleich $f(x_0) + f'(x_0)h$ in einer Umgebung um x_0, d.h. es existiert ein $\overline{\omega} : (-h_0, h_0) \longrightarrow \mathbb{R}$ mit $\frac{|\overline{\omega}(h)|}{|h|} \longrightarrow 0$ und $f(x_0 + h) = f(x_0) + f'(x_0)h + \overline{\omega}(h)$.*

Notation: $f(x_0 + h) \stackrel{\bullet}{=} f(x_0) + f'(x_0)h$.

Definition 1.36 (Landau Symbole)
Seien $g, h : \mathbb{R} \longrightarrow \mathbb{R}$. Dann schreiben wir:

(i) $g(t) = O(h(t))$ für $t \longrightarrow 0 \iff$ es eine Konstante $C > 0$ und ein $\delta > 0$ gibt, so dass $|g(t)| \leq C\,|h(t)| \ \forall\ |t| < \delta$.

(ii) $g(t) = o(h(t))$ für $t \longrightarrow 0 \iff$ es ein $\delta > 0$ und ein $c : (0, \delta) \longrightarrow \mathbb{R}$ gibt, so dass $|g(t)| \leq c\,(|t|)\,|h(t)| \ \forall\ |t| < \delta$ und $c(t) \longrightarrow 0$ für $t \longrightarrow 0$.

Beispiel: $f \in C^1(\mathbb{R})$, dann ist $f(x) - (f(x_0) + f'(x_0)(x - x_0)) = o\,(|x - x_0|)$ wegen Folgerung 1.34 (Seite 12) mit $h = x - x_0$ und $m = 1$.

Ist $f \in C^2(\mathbb{R})$, dann ist $f(x) - (f(x_0) + f'(x_0)(x - x_0)) = O\left(|x - x_0|^2\right)$ wegen Satz 1.32 (Seite 12), da f'' beschränkt in einer Umgebung von x_0, d.h. $|f''(\xi)| < C$.

1.5 Approximationsfehler und Fehleranalyse

Problem: Ein Stahlseil der Länge $L = 1$ sei an seinen Endpunkten so befestigt, dass es (fast) straff gespannt erscheint. Nun soll die Ausleknung des Seils berechnet werden, wenn sich in der Mitte des Seils ein Seiltänzer befindet.

1. **Modellfehler:** Wir gehen davon aus, dass sich das Seil als Graph einer Funktion $y : (0, 1) \longrightarrow \mathbb{R}$ beschreiben lässt, welche die sogenannte **potentielle Gesamtenergie**:

$$E(y) = \frac{c}{2} \int_0^1 \frac{y'(t)^2}{\sqrt{1 + y'(t)}} \mathrm{d}t - \int_0^1 f(t)y(t)\mathrm{d}t$$

   minimiert.

   Dabei ist c eine Materialkonstante und f die Belastungsdichte.

2. Zur Vereinfachung (Abb 1.1) nehmen wir an, dass $|y'(t)| \ll 1$. Dann können wir das Funktional E vereinfachen zu:

$$\overline{E}(y) = \frac{c}{2} \int_0^1 y'(t)^2 \mathrm{d}t - \int_0^1 f(t)y(t)\mathrm{d}t.$$

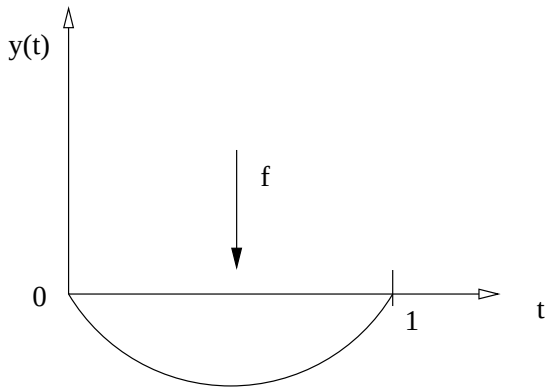

Abbildung 1.1: Modellfehler

Dabei sind eine Reihe von Effekten vernachlässigt worden. Dies führt zu Modellfehlern, die jedoch in dieser Vorlesung nicht weiter betrachtet werden. Wir nehmen an, dass die Minimierung von $\overline{E}$ das zu lössende Problem sei: Als notwendige und hinreichende Bedingung für die Minimierung von $\overline{E}$ erhält man durch Variation

$$\left(\frac{d}{d\alpha}\overline{E}\,(y+\alpha\varphi)\,|_{k=0} = 0 \;\forall\; \text{“zul”assige“ } \varphi\right)$$

die Differentialgleichung

$$-cy''(t) = f(t),\ \forall t \in (0,1)$$

mit Randwerten $y(0) = y(1) = 0$.

3. **Datenfehler:** c ist eine Materialkonstante, die vom Material des Seils abhängt (aber auch von Temperatur und Luftfeuchtigkeit). Der Wert für c kann nur durch Experimente bestimmt werden, und das ist zwangsläufig fehlerbehaftet. Daher muss sichergestellt werden, dass sowohl y als auch das numerische Verfahren nicht sensitiv vom konkreten Wert für c abhängen.

   Beispiel: Betrachten wir die Differentialgleichung $u'(t) = (c-u(t))^2,\ u(0) = 1\ c > 0$, so folgt durch Substitution

$$v(t) = \frac{1}{c-u(t)} \implies v'(t) = \frac{u'(t)}{(c-u(t))^2} = 1 \implies u(t) = \frac{1+tc(c-1)}{1+t(c-1)}.$$

   Studieren wir das **Verhalten von u in Abhängigkeit von c**, so sehen wir:

   $c = 1:\ u'(t) = 0 \implies u \equiv 1.$

   $c > 1:\ u' > 0$, d.h. u monoton wachsend und $\lim\limits_{t \longrightarrow \infty} u(t) = c.$

   $c < 1:\ u' > 0, \lim\limits_{t \longrightarrow t_0} u(t) = \infty$ für $t_0 = \frac{1}{1-c} > 0.$

   Durch Messfehler oder auch Approximationsfehler kann leicht $c > 1$ oder $c < 1$ eintreten, und man erhält qualitativ unterschiedliche Ergebnisse.

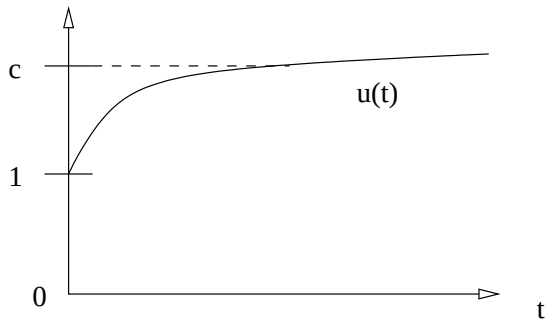

Abbildung 1.2: Auswirkung des Datenfehlers.

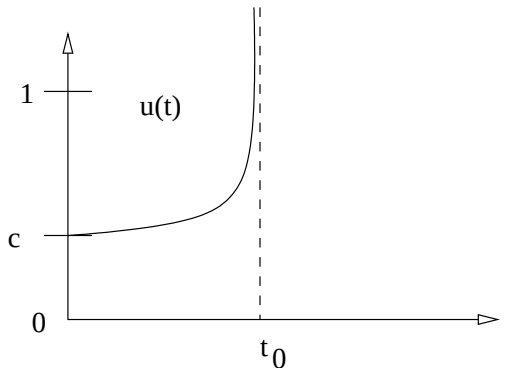

Abbildung 1.3: Auswirkung des Datenfehlers.

4. **Diskretisierungsfehler**: Zurück zu unserem Seiltänzerproblem. In der Numerik müssen wir Ableitungen durch etwas Berechenbares ersetzen, auch können wir $y(t)$ nicht für alle t bestimmen. Sei $N \in \mathbb{N}, x_i := ih, i = 0, \ldots, N+1$ mit $h = \frac{1}{N+1}$, so approximieren wir auch hier

$$y''(x_i) \approx \frac{1}{h^2} \left(y(x_{i+1}) - 2y(x_i) + y(x_{i-1}) \right).$$

Setze: $f_i \equiv f(x_i)$ und sei $y_i \approx y(x_i)$, dann ist eine Finite-Diffenrenzen Approximation von $-cy''(t) = f(t),\ t \in (0,1),\ y(0) = y(1) = 0$ gegeben durch

$$-\frac{c}{h^2} \left(y_{i+1} - 2y_i + y_{i-1} \right) = f_i,\ i = 1, \ldots, N, \quad y_0 = y_{N+1} = 0.$$

Die Differenz $|y_i - y(x_i)|$ ist der Diskretisierungsfehler im Punkt x_i. Es muss untersucht werden, wie sich dieser Fehler verhält wenn die Gitterweite h gegen 0 geht.

5. **Lösungsfehler/Abbruchfehler:**
Setze $A = \begin{pmatrix} 2 & -1 & & 0 \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 2 \end{pmatrix} \in \mathbb{R}^{N \times N}$ und $F = (f_i)_{i=1}^N \in \mathbb{R}^N$, so ergibt sich aus der Finite Differenzen Diskretisierung das diskretes Problem: Finde $y_h \in \mathbb{R}^N$ mit

$$\frac{c}{h^2} A y_h = F.$$

Zur Lösung verwenden wir die Identität

$$Dy_h = Dy_h - Ay_h + \frac{h^2}{c}F, \text{ mit } D = diag(2) = \begin{pmatrix} 2 & & 0 \\ & \ddots & \\ 0 & & 2 \end{pmatrix}.$$

Sei y_h^0 ein beliebiger Startwert (z.B: $y_h^0 = 0$). Zur Berechnung von y_h betrachten wir folgende Iterationsvorschrift:

$$Dy_h{}^{n+1} := Dy_h^n - Ay_h^n + \frac{h^2}{c}F$$

bzw.

$$y_h^{n+1} = y_h^n - D^{-1}\left(Ay_h^n + \frac{h^2}{c}F\right).$$

Es muss gezeigt werden, dass $y_h{}^n \to y_h$ für $n \to \infty$.

In der Praxis können wir nur bis zu einem endlichen Wert $n_0 \in \mathbb{N}$ rechnen. Das heißt die Lösung eines Problems wird $y_h^{n_0}$ sein. Der Abbruchfehler in der Norm $\|\cdot\|$ ist $\left\|y_h^{n_0} - y_h\right\|$.

6. **Rundungsfehler:** Auf einem Rechner kann nur eine endliche Teilmenge von $\mathbb{R}$ bearbeitet werden. Daher wird nicht $y_h^{n_0}$ berechnet, sondern die Approximation in dieser endlichen Menge.

Definition 1.37 (Gleitkommazahl)
Eine Gleitkommazahl zur Basis $b \in \mathbb{N}$ ist eine Zahl $a \in \mathbb{R}$ der Form

$$a = \pm\left[m_1 b^{-1} + \ldots + m_r b^{-r}\right] b^{\pm\left[e_{s-1}b^{s-1}+\ldots+e_0 b^0\right]}. \qquad (*)$$

Man schreibt $\pm a = 0, m_1 \ldots m_r b^{\pm E}$ mit $E = [e_{s-1}b^{s-1} + \ldots + e_0 b^0]$ und $m_i \in \{0, \ldots, b-1\}, E \in \mathbb{N}, r, s \in \mathbb{N}$ abhängig von der Rechnerarchitektur.

Bemerkung:

1. Diese Darstellung ermöglicht die gleichzeitige Speicherung sehr unterschiedlich großer Zahlen, wie etwa die Lichtgeschwindigkeit $c \approx 0.29998 \cdot 10^9 \frac{m}{s}$ oder Elektronenruhemasse $m_0 \approx 0.911 \cdot 10^{-30}$ kg.

2. Als Normierung nimmt man für $a \neq 0$ an, dass $m_1 \neq 0$ ist.

3. Für Computer ist $b = 2$ üblich, für Menschen $b = 10$.

Definition 1.38 (Maschinenzahlen)
Zu geg. (b, r, s) sei $A = A(b, r, s)$ die Menge der $a \in \mathbb{R}$ mit einer Darstellung $(*)$.

$A(b, r, s)$ ist endlich mit größtem und kleinstem positiven Element $a_{\max} = (1 - b^{-r}) \cdot b^{b^s - 1}$, $a_{\min} = b^{-b^s}$.

Zur Speicherung einer Zahl $a \in D = [-a_{\max}, -a_{\min}] \cup [a_{\min}, a_{\max}]$ wird eine Rundungsfunktion $rd : D \longrightarrow A$ mit $rd(a) = \min\limits_{\overline{a} \in A} |\overline{a} - a|$ definiert.

$rd(a)$ wird gespeichert als:[1]

$\pm$	m_1	$\cdots$	m_r	$\pm$	e_0	$\cdots$	e_{s-1}

$rd(a) = 0, \underbrace{m_1, \ldots, m_r}_{\text{Mantisse } M}, b^{\pm E}$ mit E als Exponent.

Die heutigen PC benutzen 52 Bits für die Mantisse und 11 Bits für den Exponent; die $\pm$ werden mit 1 (negativ) und 0 (positiv) dargestellt.

Für $a \in (-a_{\min}, a_{\min})$ wird in der Regel $rd(a) = 0$ gesetzt ("underflow").

Für $|a| > a_{\max}$ wird von "overflow" geredet. Viele Compiler setzen $a = NaN$ (not a number) und die Rechnung muss abgebrochen werden.

Satz 1.39 (Rundungsfehler)
Der absolute Rundungsfehler, der durch Rundung verursacht wird, kann abgeschätzt werden durch

$$|a - rd(a)| \leq \frac{1}{2} b^{-r} \cdot b^E,$$

wobei E der Exponent von a ist (in der $(*)$ Darstellung). Für den relativen Rundungsfehler gilt für $a \neq 0$

$$\frac{|rd(a) - a|}{|a|} \leq \frac{1}{2} b^{-r+1}.$$

Die Zahl $eps := \frac{1}{2} b^{-r+1}$ heißt Maschinengenauigkeit.

Beweis: $rd(a)$ weicht maximal eine halbe Einheit in der letzten Mantissenstelle von a ab.
Also $|a - rd(a)| \leq \frac{1}{2} b^{-r} b^E$.

Aufgrund der Normalisierung $m_1 \neq 0$ folgt $|a| \geq b^{-1} b^E$ und weiter

$$\frac{|rd(a) - a|}{|a|} \leq \frac{\frac{1}{2} b^{-r} b^E}{b^{-1} b^E} = \frac{1}{2} b^{-r+1}. \quad \square$$

Setzt man $\varepsilon := \frac{rd(a)-a}{a}$, so folgt $|\varepsilon| \leq eps$ und $rd(a) = \varepsilon a + a = a(1 + \varepsilon)$. $\square$

Definition 1.40 (Maschinenoperation)
Die Grundoperation $\star \in \{+, -, \times, /\}$ wird ersetzt durch $\circledast$. In der Regel gilt:

$$a \circledast b = rd(a \star b) = (a \star b)(1 + \varepsilon)$$

mit $|\varepsilon| \leq eps$.

Bemerkung: Die Verknüpfungen $\circledast$ erfüllen **nicht** das Assoziativ- bzw. Distributivgesetz.

Beispiel 1.41

Berechne das Integral $I_k := \int_0^1 \frac{x^k}{x+5} \mathrm{d}x$.

[1] Jedes Kästchen entspricht einem Bit

(A) Es gilt

$$I_0 = \ln(6) - \ln(5)$$

und

$$I_k + 5I_{k-1} = \frac{1}{k} \ (k \geq 1), \text{ da}$$

$$\int_0^1 \frac{x^k}{x+5} + 5\frac{x^{k-1}-1}{x+5} = \int_0^1 x^{k-1}\mathrm{d}x = \frac{1}{k}.$$

Bei einer Berechnung mit nur 3 Dezimalstellen ($r = 3, b = 10$) ergibt sich:

$$\begin{aligned}
\bar{I}_0 &= 0.182 \cdot 10^0 \\
\bar{I}_1 &= 0.900 \cdot 10^{-1} \\
\bar{I}_2 &= 0.500 \cdot 10^{-1} \\
\bar{I}_3 &= 0.833 \cdot 10^{-1} \\
\bar{I}_4 &= -0.166 \cdot 10^0
\end{aligned}$$

Dabei bezeichnet $\bar{I}_k$ den berechneten Wert unter Berücksichtigung der Rundungsfehler. Die Berechnung ist fehlerhaft. Offensichtlich sind die I_k monoton fallend, da $I_k \searrow 0$ ($k \longrightarrow \infty$), aber es gibt widersprüchliche Ergebnisse (siehe I_3). Auf einem Standard PC ergab: $\bar{I}_{21} = -0.158 \cdot 10^{-1}$ und $\bar{I}_{39} = 8.960 \cdot 10^{10}$.

Dies ist ein Beispiel für **Fehlerfortpflanzung**, da der Fehler in I_{k-1} mit 5 multipliziert wird, um I_k zu berechnen.

(B) Berechnet man die Werte I_k exakt, so ergibt sich bei einer Rundung auf drei Dezimalstellen $I_9 = I_{10}$ und eine Rückwärtsiteration $I_{k-1} = \frac{1}{5}\left(\frac{1}{k} - I_k\right)$ ergibt:

$$\begin{aligned}
\bar{I}_4 &= 0.343 \cdot 10^{-1} \\
\bar{I}_3 &= 0.431 \cdot 10^{-1} \\
\bar{I}_2 &= 0.500 \cdot 10^{-1} \\
\bar{I}_1 &= 0.884 \cdot 10^{-1} \\
\bar{I}_0 &= 0.182 \cdot 10^0
\end{aligned}$$

Hier tritt **Fehlerdämpfung** auf.

Beispiel 1.42
Zu lösen ist das LGS

$$\begin{pmatrix} 1.2969 & 0.8648 \\ 0.2161 & 0.1441 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0.86419999 \\ 0.14400001 \end{pmatrix} =: b.$$

Die exakte Lösung ist $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0.9911 \\ -0.4870 \end{pmatrix}$.

Durch Messfehler oder auch Rundung erhalten wir eine rechte Seite

$$\bar{b} = \begin{pmatrix} 0.8642 \\ 0.1440 \end{pmatrix}$$

Dann ist die Lösung $\begin{pmatrix} \widetilde{x}_1 \\ \widetilde{x}_2 \end{pmatrix} = \begin{pmatrix} 2 \\ -2 \end{pmatrix}$, d.h. wir erhalten ca. 100% Abweichung.

Dies bedeutet, dass kleine Änderungen der Eingabedaten zu großen Änderungen der Lösungen führen können. In diesem Kontext führen wir den Begriff der Kondition eines Problems ein.

Definition
Eine numerische Aufgabe (z.B. effizientes Lösen eines LGS oder Integrals) heißt **gut konditioniert**, *falls kleine Änderungen der Eingabedaten zu kleinen Änderungen der Lösung führen; sonst heißt das Problem* **schlecht konditioniert**.

Präzisieren wir: Was ist eine numerische Aufgabe? Was heißt klein?
Die Matrix

$$A := \begin{pmatrix} 1.2969 & 0.8648 \\ 0.2161 & 0.1441 \end{pmatrix}$$

sollte schlecht konditioniert sein.
Im folgenden 2 Ansätze:

1. Für einfache Probleme.
2. Für etwas komplexere Probleme.

Definition 1.43
Sei $f : U \longrightarrow \mathbb{R}^n$ mit $U \subset \mathbb{R}^m$ und sei $x_0 \in U$ vorgegeben. Dann versteht man unter der Aufgabe (f, x_0) die effektive Berechnung von f an der Stelle x_0. Dabei sind x_0 die Eingabedaten.

Beispiel: $Ax = b$, (f, b) mit $f(b) = A^{-1}b$

Satz 1.44
Sei $x_0 = (x_1, \ldots, x_m)$ und $x_0 + \Delta_x \in U$ eine Störung der Eingabedaten mit $\|\Delta x\| \ll 1$. Falls $f : U \longrightarrow \mathbb{R}$ (d.h. $n = 1$) einmal stetig differenzierbar, so ist der Ergebnisfehler $\Delta f(x_0) = f(x_0) - f(x_0 + \Delta x)$ in erster Nährung gleich

$$\sum_{j=1}^{m} \frac{\partial f}{\partial x_j}(x_0)\Delta x_j = \nabla f(x_0)\Delta x.$$

Für den relativen Fehler gilt in erster Nährung

$$\frac{\Delta f(x_0)}{f(x_0)} \doteq \sum_{j=1}^{m} \left(\frac{\partial f}{\partial x_j}(x_0) \frac{x_j}{f(x_0)} \right) \frac{\Delta x_j}{x_j}.$$

Definition 1.45 (Konditionszahlen I)
Wir nennen den Faktor $k_j := \frac{\partial f}{\partial x_j}(x_0)\frac{x_j}{f(x_0)}$ **(relative) Konditionszahl**.

Beweis: (Beweis von Satz 1.44)

Wie in der Folgerung 1.34 (Seite 12) kann man hier den Satz von Taylor anwenden:

$$f(x_0 + \Delta x) = f(x_0) + \bigtriangledown f(x_0)\Delta_x + \overline{\omega}\left(\|\Delta x\|\right)$$

mit $\overline{\omega}\left(\|\Delta x\|\right) = o\left(\|\Delta x\|\right) \implies$ Behauptung für den absoluten Fehler. □

Bemerkung: k_j beschreibt, wie der relative Fehler in den Eingabedaten x_j verstärkt bzw. abgeschwächt wird. □

Definition 1.46
Wir nennen das Problem (f, x_0) **gut konditioniert**, *falls alle* k_j $(j = 1, \ldots, m)$ *klein sind, sonst* **schlecht konditioniert**.

Beispiel 1.47 (Arithmetische Operationen)

(i) $f(x_1, x_2) = x_1 x_2,\ k_1 = \frac{\partial f}{\partial x_1}(x_1, x_2)\frac{x_1}{f(x_1,x_2)} = \frac{x_2 x_1}{x_1 x_2} = 1$

Analog für k_2 ergibt sich ebenfalls 1 $\implies$ Multiplikation ist gut konditioniert.

(ii) Division ist gut konditioniert.

(iii) Addition $f(x_1, x_2) = x_1 + x_2$:

$$k_j = 1\frac{x_j}{x_1 + x_2} = \frac{x_j}{x_1 + x_2}.$$

k_j wird beliebig groß, wenn $x_1 x_2 < 0$ und x_1 und x_2 betragsmäßig gleich groß sind. Das heißt, in diesem Fall ist die Addition schlecht konditioniert, ansonsten ist sie gut konditioniert.

(iv) Subtraktion ist schlecht konditioniert, falls $x_1 x_2 > 0$ und x_1 und x_2 betragsmäßig gleich groß sind.

Beispiel: $(n = 3) x = 0.9995 \quad y = 0.9984 \quad rd(x) = 0.1 \cdot 10^1 \quad rd(y) = 0.998 \cdot 10^0$ Dann gilt für $\circledast = -$

$$x \circledast y = rd(1 - 0.998) - rd(0.2 \cdot 10^{-2}) = 0.2 \cdot 10^{-2}$$

Der absolute Fehler beträgt $x \circledast y - (x - y) = 0.0001$
Der relative Fehler beträgt $\frac{x \circledast y - (x-y)}{(x-y)} = 0.82$

Das Problem wird als Auslöschung bezeichnet.

Bei komplexeren Problemen (etwa $n > 1$) betrachten wir einen anderen Ansatz:

Definition 1.48
Das Problem (f, x_0) *ist* **wohlgestellt** *in*

$$B_\delta(x_0) := \{x \in U \mid \|x - x_0\| < \delta\}$$

falls es eine Konstante $L_{abs} \geq 0$ *gibt, mit*

$$\|f(x) - f(x_0)\| \leq L_{abs} \|x - x_0\| \qquad (*)$$

für alle $x \in B_\delta(x_0)$*. Gibt es keine solche Konstante, so heißt das Problem* **schlecht gestellt**.

Sei im folgenden $L_{abs}(\delta)$ *die kleinste Zahl mit der Eigenschaft* $(*)$.
Analog sei $L_{rel}(\delta)$ *die kleinste Zahl mit*

$$\frac{\|f(x) - f(x_0)\|}{\|f(x_0)\|} \leq L_{rel}(\delta) \frac{\|x - x_0\|}{\|x_0\|}.$$

Definition 1.49 (Konditionszahlen II)
Wir definieren $K_{abs} := \lim\limits_{\delta \searrow 0} L_{abs}(\delta)$ *die* **absolute Konditionszahl** *und* $K_{rel} := \lim\limits_{\delta \searrow 0} L_{rel}(\delta)$ *die* **relative Konditionszahl**.

Bemerkung: Falls f differenzierbar, so gilt

$$K_{rel} = \left\|f'(x_0)\right\| \frac{\|x_0\|}{\|f(x_0)\|}.$$

Beachte: $f'(x_0)$ ist eine Matrix und $\|f'(x_0)\|$ eine Matrixnorm. K_{rel} hängt von der Wahl der Normen ab.

Beispiel 1.50 (Konditionierung eines LGS)
Zu lösen ist $Ax = b$, d.h. $f(b) = A^{-1}b$ und $f'(x) = A^{-1}$.
Damit folgt $K_{abs} = \left\|A^{-1}\right\|$; und hieraus mit $Ax = b$ und der Submultiplikativität der zugeordneten Norm:

$$K_{rel} = \left\|A^{-1}\right\| \frac{\|b\|}{\|A^{-1}b\|} = \left\|A^{-1}\right\| \frac{\|Ax\|}{\|x\|} \leq \frac{\left\|A^{-1}\right\| \cdot \|A\| \cdot \|x\|}{\|x\|} = \left\|A^{-1}\right\| \cdot \|A\|.$$

Wir definieren entsprechend die Kondition der Matrix A durch

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

Beachte, dass ein $x \in \mathbb{R}^m$ existiert mit $\|Ax\| = \|A\| \|x\|$, d.h. cond(A) ist eine gute Abschätzung für die Konditionierung vom Problem (f, b)

Mit A wie in Beispil 1.42 gilt: $cond(A) = \left\|A^{-1}\right\| \|A\| \approx 10^9$. Das Problem ist also schlechte konditioniert.

Kapitel 2

Lineare Gleichungssysteme

Wir werden in diesem Kapitel Probleme der Form

$$Ax = b$$

betrachten, wobei $A \in \mathbb{R}^{n \times n}$ und $x, b \in \mathbb{R}^n$. Es gibt im Wesentlichen 2 Klassen von Verfahren

1. Direkte Verfahren
2. Iterative Verfahren

Aus der Schule (und den Lineare Algebra-Vorlesungen) ist uns ein direkte Verfahren bekannt, das Gaußsche Eliminationsverfahren. Für kleine Gleichungssysteme eignet sich dieses Verfahren, jedoch kann das Verfahren für $n >> 1000$ sehr ineffizient werden, da das Verfahren einen Rechenaufwand der Ordnung n^3 hat. Aus diesem Grund werden wir andere Verfahren kennenlernen, mit denen man schneller ans Ziel kommen kann.

Wir werden Probleme folgender Art behandeln:

(A) Geg: $A \in \mathbb{R}^{n \times n}, b \in \mathbb{R}^n$
Ges: $x \in \mathbb{R}^n$ mit $Ax = b$ (falls eine Lösung existiert).

(B) Geg: $A \in \mathbb{R}^{n \times n}, b_1, \ldots, b_l \in \mathbb{R}^n$
Ges: $x_i \in \mathbb{R}^n$ mit $Ax_i = b_i$ $(i = 1, \ldots, l)$ (falls Lösungen existieren).

(C) Geg: $A \in \mathbb{R}^{n \times n}$
Ges: A^{-1} (falls die Inverse existiert).

Es sind äquivalent

(i) $\exists!\ x \in \mathbb{R}^n :\ Ax = b.$

(ii) $Ax = 0 \iff x = 0.$

(iii) $\det(A) \neq 0.$

(iv) 0 ist kein Eigenwert von A.

(v) A ist regulär, d.h. $\exists\ B \in \mathbb{R}^{n \times n}$ mit $AB = BA = E_n$. Dabei ist $B = A^{-1}$ und $x = A^{-1}b$ ist die eindeutige Lösung von $Ax = b$.

Alle diese Probleme sind äquivalent, aber es existieren Verfahren, die besonders geeignet für eines dieser Probleme sind.

Verfahren

1. Direkte Verfahren liefern die exakte Lösung x nach endlich vielen Schritten (bis auf Rundungsfehler). Beispiele dafür sind der *Gaußalgorithmus* mit Aufwand $O(n^3)$ und die *Cramersche Regel* mit Aufwand $O(n!)$. Der minimale theoretische Aufwand liegt bei $O(n^2)$, aber es existiert kein direktes Verfahren mit dieser Komplexität.

   Der Vorteil direkter Verfahren ist, dass A^{-1} in der Regel mitbestimmt wird und somit der Aufwand für (A), (B) und (C) ungefähr gleich groß ist.

   Ein Nachteil ist, dass während der laufenden Berechnung keine Näherung vorliegt, d.h. das Resultat steht erst nach Abarbeitung des Algorithmus, also erst nach n Schritten, fest. Je nach Anwendung sind diese Verfahren viel zu aufwändig und deshalb besonders ungeeignet für Problem (A), wenn n sehr groß ist.

2. Iterative Verfahren liefern nach endlich vielen Schritten eine beliebig genaue Approximation der Lösung (bis auf Rundungsfehler).

   Der Vorteil liegt darin, dass man in der Lage ist, die Lösung so genau zu bestimmen, wie es nötig ist. Häufig hat man bereits eine brauchbare Lösung nach $k \ll n$ Schritten

Satz 2.1 (Störungssatz für lineare Gleichungssysteme)
Sei $A \in \mathbb{R}^{n\times n}$ regulär und $\|\cdot\|$ die induzierte Matrixnorm. Sei $\Delta A \in \mathbb{R}^{n\times n}$ gegeben mit $\|\Delta A\| < \frac{1}{\|A^{-1}\|}$ und sei $b \in \mathbb{R}^n$ und $\Delta b \in \mathbb{R}^n$. Dann ist $A + \Delta A$ regulär und es gilt

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

Dabei ist $Ax = b$ und $\overline{x}$ die Lösung des von $(A + \Delta A)\,\overline{x} = b + \Delta b$.

Bemerkung: $cond(A) := \|A\| \left\|A^{-1}\right\|$ ist der entscheidende Verstärkungsfaktor für den relativen Fehler.

2.1 Direkte Verfahren

Idee: Hat A eine einfache Gestalt, so lässt sich x leicht bestimmen.

Beispiel 2.2 (Dreiecksmatrizen)
Sei $A \in \mathbb{R}^{n\times n}$ eine obere Dreiecksmatrix ($\triangle$-Matrix) , d.h. $a_{ij} = 0$ für $i > j$, oder

$$A = \begin{pmatrix} * & \cdots & * \\ & \ddots & \vdots \\ 0 & & * \end{pmatrix}.$$

Dann gilt $\det(A) = \prod_{i=1}^{n} a_{ii}$, d.h. A ist regulär $\iff$ $a_{ii} \neq 0 \; \forall \; i \in \{1, \ldots, n\}$. Ist A regulär, so ist $Ax = b$ lösbar. Aus

$$b_i = \sum_{u=1}^{n} a_{iu} x_u = \sum_{u=i}^{n} a_{iu} x_u$$

erhalten wir den Algorithmus:

$$\begin{aligned} i = n: \quad & x_n = \tfrac{b_n}{a_{nn}}, \\ i < n: \quad & x_i = \tfrac{1}{a_{ii}} \left(b_i - \sum_{u=i+1}^{n} a_{iu} x_n \right). \end{aligned}$$

Frage: Kann eine beliebige reguläre Matrix A so umgeformt werden, dass sie obere $\triangle$-Gestalt hat? D.h gesucht ist $\widetilde{A} \in \mathbb{R}^{n \times n}$ mit oberer $\triangle$-Gestalt, $\widetilde{b} \in \mathbb{R}^n$, so dass $\widetilde{A}x = \widetilde{b}$ dieselbe Lösung hat wie $Ax = b$.

Eine Lösung dieses Problems liefert der Gaußalgorithmus:

2.1.1 Gaußalgorithmus/LR-Zerlegung

Der Algorithmus startet mit

$$(A, b) = \left(A^{(0)}, b^{(0)}\right) = \left(\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} & b_n \end{array}\right)$$

und führt durch sukzessive Manipulation auf $\left(A^{(p)}, b^{(p)}\right), p = 1, \ldots, n-1$, so dass aus $A^{(p-1)}x = b^{(p-1)}$ folgt $A^{(p)}x = b^{(p)}$. Um $\left(A^{(1)}, b^{(1)}\right)$ zu berechnen, wird zur i-ten Zeile für $i = 2, \ldots, n$ das $a_{i1}^{(0)}/a_{11}^{(0)}$-fache der ersten Zeile hinzuaddiert. Wir erhalten somit

$$\left(A^{(1)}, b^{(1)}\right) = \left(\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ 0 & a_{22}^{(1)} & \cdots & a_{2n}^{(1)} & b_2^{(1)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & a_{n2}^{(1)} & \cdots & a_{nn}^{(1)} & b_n^{(1)} \end{array}\right)$$

$$\downarrow$$

$$\left(A^{(p-1)}, b^{(p-1)}\right) = \left(\begin{array}{ccccccc|c} a_{11} & \cdots & \cdots & \cdots & \cdots & \cdots & a_{1n} & b_1 \\ 0 & a_{22}^{(1)} & \cdots & \cdots & \cdots & \cdots & a_{2n}^{(1)} & b_2^{(1)} \\ \vdots & 0 & \ddots & \cdots & \cdots & \cdots & a_{in}^{(i-1)} & b_i^{(i-1)} \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & a_{pp}^{(p-1)} & \cdots & a_{pn}^{(p-1)} & b_p^{(p-1)} \\ \vdots & \vdots & \vdots & \vdots & a_{(p+1)(p+1)}^{(p-1)} & \cdots & a_{(p+1)n}^{(p-1)} & b_{(p+1)}^{(p-1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & a_{n(p+1)}^{(p-1)} & \cdots & a_{nn}^{(p-1)} & b_n^{(p-1)} \end{array}\right)$$

Wir erhalten schließlich $\left(A^{(n-1)}, b^{(n-1)}\right)$, wobei $A^{(n-1)}$ eine obere $\triangle$-Matrix ist und $Ax = b \iff A^{(n-1)}x = b^{(n-1)}$.

Unsere Rechnung setzt voraus, dass stets gilt $a_{pp}^{(p-1)} \neq 0$ gilt, ansonsten müssen zuerst Zeilen vertauscht werden. Weder das Vertauschen von Zeilen noch der Eliminationsschritt verändern die Lösung. Kann in einem Schritt $a_{pp}^{(p-1)} \neq 0$ nicht erreicht werden, nachdem man sämtliche Zeilen vertauscht hat, so bedeuted dies, dass A singulär ist. Das Zeilenvertauschen wird als **Pivotisierung** bezeichnet. Häufig wird die Zeile ausgesucht mit

$$\left|a_{kp}^{(p-1)}\right| = \max_{p \leq i \leq n} \left|a_{ip}^{(p-1)}\right|$$

und wird als **Teilpivotisierung** oder **Spaltenpivotisierung** bezeichnet.

Beispiel 2.3
Sei $A = \begin{pmatrix} \varepsilon & 1 \\ 1 & 1 \end{pmatrix}$ und $b = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \implies x = \begin{pmatrix} \frac{1}{1-\varepsilon} \\ \frac{1-2\varepsilon}{1-\varepsilon} \end{pmatrix} \approx \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ für $\varepsilon \ll 1$.

Ohne Pivotisierung folgt: $\left(A^{(1)}, b^{(1)}\right) = \left(\begin{array}{cc|c} \varepsilon & 1 & 1 \\ 0 & 1 - \frac{1}{\varepsilon} & 2 - \frac{1}{\varepsilon} \end{array}\right)$

Es folgt $x_2 = \frac{2-\varepsilon^{-1}}{1-\varepsilon^{-1}} \approx 1$ und $x_1 = (1 - x_2)\varepsilon^{-1} \approx 0$, da auf einem Computer $rd(2 - \varepsilon^{-1}) = -\varepsilon^{-1}, rd(1 - \varepsilon^{-1}) = -\varepsilon^{-1}$ berechnet werden.

Mit Pivotisierung folgt hingegen nach Zeilentausch:

$$\left(\begin{array}{cc|c} 1 & 1 & 2 \\ \varepsilon & 1 & 1 \end{array}\right)$$

und schließlich nach der Elimination

$$\left(\begin{array}{cc|c} 1 & 1 & 2 \\ 0 & 1 - \varepsilon & 1 - 2\varepsilon \end{array}\right)$$

Hier folgt also $x_2 = \frac{1-2\varepsilon}{1-\varepsilon} \approx 1$ und $x_1 = 2 - x_2 \approx 1$, da auf einem Computer $rd(1 - 2\varepsilon) = 1$ für sehr kleines ε gilt.

Das äquivalente Problem

$$\begin{pmatrix} 1 & \varepsilon^{-1} \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \varepsilon^{-1} \\ 2 \end{pmatrix}$$

kann durch Spaltenpivotisierung nicht gelöst werden. Hier muss **total pivoting** benutzt werden, d.h. sind die Matrixeinträge sehr unterschiedlich groß, so müssen auch die Spalten vertauscht werden. Das Vertauschen der Spalten ist jedoch umständlich und wird selten angewandt. Man vertauscht die Spalten nur dann, wenn man keine andere Wahl hat.

Wir fassen das Gaußverfahren, wie folgt zusammen.

Algorithmus 2.4 (Gaußverfahren)
Setze $q_i = i \ \ (i = 1, \ldots, n)$.

Für $p = 1, \ldots, n-1$:

Wähle $j \in \{p, \ldots, n\}$ mit $\left|a_{q_j p}\right| = \max\limits_{k=p,\ldots,n} \left|a_{q_k p}\right|$.

Vertausche die Zeilen $q_j \longleftrightarrow q_p$ [Spaltenpivotisierung]

Für $k = p+1, \ldots, n$:

Falls $a_{q_p p} = 0 \implies$ Abbruch.

Setze $l = \frac{a_{q_k p}}{a_{q_p p}}$ [Multiplikationsfaktor]

Setze $a_{q_k p} = l$ [Speichere l statt $a_{q_k p} = 0$]

Für $j = p+1, \ldots, n$:

Setze $a_{q_k j} = a_{q_k j} - l \cdot a_{q_p j}$ [Matrix $A^{(p)}$]

Setze $b_{q_k} = b_{q_k} - l \cdot b_{q_p}$ [Vektor $b^{(p)}$]

Die Lösung von $Ax = b$ wird anschließend durch Rückwärtseinsetzen wie folgt gelöst:

Setze $x_n = b_{q_n} / a_{q_n n}$.
Für $k = n-1, \ldots, 1$:
$$x_k = \left(b_{q_k} - \sum_{i=k+1}^{n} a_{q_k i} x_i \right) \Big/ a_{q_k k}.$$

Bemerkungen:

(i) Der Aufwand des Algorithmus liegt bei $\frac{1}{3}n^3 + O(n^2)$.

(ii) Anstelle der entstehenden Nullen wird der Multiplikationsfaktor l gespeichert.

(iii) Die Matrix A und der Vektor b werden überschrieben. Es ist deshalb ratsam, eine Kopie der Vektoren zu machen.

(iv) Die Pivotisierung wird als Vektor gespeichert, und die Zeilenvertauschung im Speicher wird nicht durchgeführt.

Formaler Zugang:

1) Uminterpretation der Pivotisierung:
Im i-ten Schritt des Gaußalgorithmus werden Zeilen vertauscht $i \longleftrightarrow k, \ k > i$. Zur Umformulierung

betrachten wir folgende Matrix.

$$P_{ik} := \begin{pmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & 0 & \cdots & 1 & & \leftarrow & i \\ & & & \vdots & \ddots & \vdots & & & \\ & & & 1 & \cdots & 0 & & \leftarrow & k \\ & & & & & & 1 & & \\ & & & & & & & \ddots & \\ & & & & & & & & 1 \end{pmatrix}$$

Lemma 2.5
Für die Matrizen P_{ik} gilt:

(i) $B = P_{ik}A$ entspricht der Matrix nach der Vertauschung der i-ten und k-ten Zeile.

(ii) $B = AP_{ik}$ entspricht der Matrix nach der Vertauschung der i-ten und k-ten Spalte.

(iii) $P_{ik}^2 = E_n$, d.h. $P_{ik}^{-1} = P_{ik}$.

Beweis: Durch nachrechnen. □

Definition 2.6
Eine Matrix $P \in \mathbb{R}^{n\times n}$ heißt **Permutationsmatrix**, *falls P durch Zeilenvertauschungen aus der Einheitsmatrix E_n entsteht.*

Bemerkung: P_{ik} ist eine Permutationsmatrix. Ist $P = P_{l_m,k_m}, \dots P_{l_1,k_1}$, eine Permutationsmatrix, so gilt $P^{-1} = P_{l_1,k_1}, \dots P_{l_m,k_m}$.

2) Berechnung im Gaußverfahren:
Wir betrachten die Matrix:

$$L_i := \left(\begin{array}{ccc|ccc} 1 & & & 0 & & 0 \\ & \ddots & & \vdots & & \\ & & & 1 & & \\ & & & l_{i+1,i} & \ddots & \\ & & & \vdots & & \\ 0 & & & l_{n,i} & 0 & 1 \end{array}\right)$$

mit $l_{ji} := \frac{a_{ji}}{a_{ii}},\ j = i+1, \dots, n$.

Lemma 2.7

(i) Sei $B = L_iA = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$, wobei $b_i \in \mathbb{R}^n$ Zeilenvktoren sind. Dann gilt

$$b_j = a_j \ \ (j = 1, \dots, i) \qquad \text{und} \qquad b_j = a_j + l_{ji}a_i \ \ (j = i+1, \dots, n).$$

(ii) $L_i^{-1} = 2E_n - L_i$, also

$$L_i^{-1} := \left(\begin{array}{ccc|c|cc} 1 & & & 0 & & 0 \\ & \ddots & & \vdots & & \\ & & & 1 & & \\ & & & -l_{i+1,i} & \ddots & \\ & & & \vdots & & \\ 0 & & & -l_{n,i} & 0 & 1 \end{array}\right)$$

d.h. für $B = L_i^{-1}A = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$ gilt

$b_j = a_j, \quad (j = 1, \ldots, i)$ und $\qquad b_j = a_j - l_{ji}a_i, \quad (j = i+1, \ldots, n).$

Beweis: Durch nachrechnen. □

Folgerung 2.8
Die Transformation von A auf obere $\triangle$-Gestalt kann geschrieben werden als

$$R = L_{n-1}^{-1} P_{n-1} \cdots L_1^{-1} P_1 A.$$

Dabei it $P_i = P_{ij}$ für ein $j \geq i$ und R eine obere $\triangle$-Matrix ist.

Satz 2.9 (LR-Zerlegung)
Sei $A \in \mathbb{R}^{n\times n}$ eine reguläre Matrix. Dann gilt:

- Es existiert eine Permutationsmatrix P, eine untere $\triangle$-Matrix L mit Diagonalelementen 1 und eine obere $\triangle$-Matrix R mit
$$PA = LR.$$
- Es gilt: Ist $A = LR = MS$, wobei L, M untere $\triangle$-Matrizen mit Diagonalelementen 1 sind, und R, S obere $\triangle$-Matrizen sind, so folgt $L = M,\ R = S$.

Bemerkung: Ist $PA = LR$ gegeben, so kann man $Ax = b$ lösen, indem man in zwei Schritten durch Vorwärts-, bzw. Rückwärtseinsetzen löst:

(a) Löse $Lz = Pb$,

(b) löse $Rx = z$.

Dies gilt, da

$$\begin{aligned} Ax = b \quad &\Longleftrightarrow \quad PAx = Pb, \\ &\Longleftrightarrow \quad LRx = Pb, \\ &\Longleftrightarrow \quad LRx = Lz, \\ &\Longleftrightarrow \quad Rx = z. \end{aligned}$$

I.A. wird L in den frei werdenden Stellen von A gespeichert (die 1 auf der Diagonalen muss man nicht speichern). Also

$$A^{(n-1)} := \begin{pmatrix} & \tilde{R} \\ \tilde{L} & \end{pmatrix}.$$

Dabei ist $\tilde{R}$ die obere Dreiecksmatrix von R und $\tilde{L}$ die untere Dreiecksmatrix von L ohne die Diagonale. L und R benötigen also zusammen n^2 Speicherstellen.

Beweis: (von 2.9)
Nach dem Gaußalgorithmus gilt: $R = L_{n-1}^{-1} P_{n-1}, \dots L_1^{-1} P_1 A$, wobei R obere $\triangle$-Matrix ist $\implies$ $P_1 L_1 \dots P_{n-1} L_{n-1} R = A$.

Definiere Permutationsmatrix $P := P_{n-1} \dots P_1$ und $L := P P_1 L_1 \dots P_{n-1} L_{n-1}$

$\implies P^{-1} LR = A \implies LR = PA$

Noch zu zeigen: L ist untere $\triangle$-Matrix mit Diagonaelementen 1.

Es ist:

$$\begin{aligned} L &= P P_1 L_1 \dots P_{n-1} L_{n-1} = P_{n-1} \dots \underbrace{P_1 P_1}_{=E_n} L_1 P_2 \dots P_{n-1} L_{n-1} \\ &= P_{n-1} \dots P_2 L_1 P_2 \dots P_{n-1} L_{n-1}. \end{aligned}$$

Wir setzen:

$$\tilde{L}_1 := P_2 L_1 P_2 \text{ und für} p = 2, ..., n-1 : \ \tilde{L}_p := P_{p+1} \tilde{L}_{p-1} L_p P_{p+1}.$$

Dann hat $\tilde{L}_p$ die Gestalt:

$$\tilde{L}_p = \begin{pmatrix} 1 & & & & & & 0 \\ * & \ddots & & & & & \\ \vdots & \ddots & \ddots & & & & \\ \vdots & & * & \ddots & & & \\ \vdots & & \vdots & 0 & \ddots & & \\ \vdots & & \vdots & \vdots & \ddots & \ddots & \\ * & \cdots & * & 0 & \cdots & 0 & 1 \end{pmatrix}$$

Außerdem gilt, dass $P_q \bar{L}_p P_q$ (mit $q > p$) ebenfalls diese Gestalt, da $P_q = P_{qk},\ q < k$.

Also folgt

$$\begin{aligned} L &= P_{n-1} \dots P_3 P_2 L_1 P_2 L_2 P_3 \dots P_{n-1} L_{n-1} \\ &= P_{n-1} \dots P_3 \widetilde{L}_1 L_2 P_3 \dots P_{n-1} L_{n-1} \\ &= P_{n-1} \dots P_4 \widetilde{L}_2 L_3 P_4 \dots P_{n-1} L_{n-1} \quad . \\ &\vdots \\ &= \widetilde{L}_{n-1} \end{aligned}$$

Somit ist L untere $\triangle$-Matrix.

Zur Eindeutigkeit:
Es gilt: L^{-1} hat ebenfalls untere $\triangle$-Gestalt mit Diagonalelementen 1 (betrachte $L^{-1} L = E_n$). Analog folgt, dass S^{-1} eine obere $\triangle$-Matrix ist.

Also ist $L^{-1} M$ untere $\triangle$-Matrix mit Diagonalelementen 1 und RS^{-1} hat obere $\triangle$-Gestalt. Aus $LR = MS$ folgt $RS^{-1} = L^{-1} M = E_n$ und hieraus mit der Eindeutigkeit der Inversen: $R = S \ \wedge \ L = M$.

□

Weitere Anwendungen der LR-Zerlegung

(a) Determinantenberechnung einer Matrix A.
Hat R obere/untere $\triangle$-Gestalt, so gilt

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

Aus der LR Zerlegung folgt

$$R = L_{n-1}^{-1} P_{n-1} \dots L_1^{-1} P_1 A$$

und somit

$$\det R = \det(L_{n-1}^{-1}) \det(P_{n-1}) \dots \det(L_1^{-1}) \det(P_1) \det(A).$$

Weiter gilt:

$$\det\left(L_i^{-1}\right) = 1 \quad \text{und} \quad \det(P_i) = \det(P_{ik}) = \begin{cases} 1 & : \quad i = k \\ -1 & : \quad i \neq k \end{cases}.$$

Also folgt:

$$\det(A) = \begin{cases} \det(R) & : \text{ gerade Anzahl von Zeilenvertauschungen} \\ -\det(R) & : \text{ ungerade Anzahl von Zeilenvertauschungen} \end{cases}$$

$$= \begin{cases} \prod_{i=1}^{n} r_{ii} & : \text{ gerade Anzahl von Zeilenvertauschungen} \\ -\prod_{i=1}^{n} r_{ii} & : \text{ ungerade Anzahl von Zeilenvertauschungen} \end{cases}$$

(b) Bestimmung von $\text{Rang}(A) = \sharp$ der linear unabhängigen Zeilenvektoren bei einer nicht unbedingt quadratischen Matrix.

Ist im p. Schritt $a_{pp}^{(p)} = 0$, so müssen Zeilen und eventuell auch Spalten vertauscht werden, was den Rang der Matrix nicht verändert. Ist dies nicht möglich, so hat $A^{(p)}$ die Gestalt

$$A^{(p)} = \left(\begin{array}{c|c} * & * \\ \hline 0 & 0 \end{array}\right)$$

Dabei sind die ersten p Zeilenvektoren linear unabhängig, aber alle weiteren Zeilenvektoren sind linear abhängig. Es folgt $\text{Rang}\,(A) = \text{Rang}\left(A^{(p)}\right) = p$.

Achtung: Aufgrund von Rundungsfehlern kann dieses Verfahren numerisch zu falschen Ergebnissen führen:

(c) Berechnung der Umkehrmatrix A^{-1} einer Matrix A.

1. Ansatz: Sei e_i der i. Einheitsvektor. Löse $Ax^{(i)} = e_i$ für $i = 1, \ldots, n \quad \Longrightarrow \quad A^{-1} = (x_1, \ldots, x_n)$ mittels LR-Zerlegung mit $Lz^{(i)} = Pe_i$ und $Rx^{(i)} = z^{(i)}$

2. Berechnung durch simultane Elimination

$$\begin{array}{c|ccc} & 1 & & 0 \\ A & & \ddots & \\ & 0 & & 1 \end{array} \longrightarrow \quad \overset{\text{Vorwärtselimination}}{\begin{array}{ccc|ccc} r_{11} & \cdots & * & * & & 0 \\ & \ddots & \vdots & \vdots & \ddots & \\ 0 & & r_{nn} & * & \cdots & * \end{array}} \longrightarrow$$

$$\overset{\text{Rückwärtselimination}}{\begin{array}{ccc|c} r_{11} & & 0 & \\ & \ddots & & * \\ 0 & & r_{nn} & \end{array}} \longrightarrow \quad \begin{array}{ccc|c} 1 & & 0 & \\ & \ddots & & A^{-1} \\ 0 & & 1 & \end{array}$$

2.1.2 Gauß-Jordan Verfahren

Diese Methode beruht darauf, durch Matrixumformungen von $Ax = b$ zu $Bb = x$ mit $B = A^{-1}$ überzugehen. Die Idee des Verfahren ist folgende: Ist $a_{pq} \neq 0$, so kann die p-te Gleichung nach x_q aufgelöst werden:

$$x_q = -\frac{a_{p1}}{a_{pq}}x_1 - \ldots - \frac{a_{pq-1}}{a_{pq}}x_{q-1} + \frac{1}{a_{pq}}b_q - \frac{a_{pq+1}}{a_{pq}}x_{q+1} - \ldots - \frac{a_{pn}}{a_{pq}}x_n.$$

Durch Einsetzung von x_q in die anderen Gleichungen ($j \neq p$)

$$\sum_{k=1}^{q-1}\left[a_{jk} - \frac{a_{jq}a_{pk}}{a_{pq}}\right]x_k + \frac{a_{jq}}{a_{pq}}b_q + \sum_{k=q+1}^{n}\left[a_{jk} - \frac{a_{jq}a_{jk}}{a_{pq}}\right]x_k = b_j.$$

Man erhält also eine Matrix $\widetilde{A}$ mit

$$\widetilde{A}\begin{pmatrix} x_1 \\ \vdots \\ b_q \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ x_q \\ \vdots \\ b_n \end{pmatrix}.$$

Kann dieser Schritt z.B. mit $p = q$ n-mal durchgeführt werden, so ergibt sich

$$\widetilde{A}\begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \quad \Longrightarrow \quad \widetilde{A} = A^{-1}.$$

Dies entspricht einem Algorithmus ohne Pivotisierung, d.h. $a_{ii} \neq 0$, Varianten mit Pivotisierung sind ebenfalls möglich (siehe z.B. Stoer, Numerische Mathematik I, Springer, 1989, Abschnitt 4.2).

2.1.3 Cholesky Verfahren für SPD-Matrizen

Sei $A \in \mathbb{R}^{n\times n}$ eine symmetrische positive definite Matrix. Dann existiert ein unterer $\triangle$-Matrix $Q \in \mathbb{R}^{n\times n}$, so dass gilt

$$A = QQ^T.$$

Dabei ist $Q = LD^{1/2}$ mit $D^{1/2} = \text{diag}(\sqrt{r_{11}}, \ldots, \sqrt{r_{nn}})$.
Die Existenz einer solchen Zerlegung sieht man wie folgt ein:

$$A = LR \implies A = LD\tilde{R} = LDL^T = LD^{1/2}(LD^{1/2})^T = QQ^T,$$

wobei $\tilde{R} = L^T$ aus der Symmetrie von A folgt. Die positive Definitheit der Matrix A ist notwendig, damit $D^{1/2}$ wohldefiniert ist.
Unter Ausnutzung der Symmetrie erhält man folgenden Algorithmus, der die untere Dreiecksmatrix von Q anstelle der unteren Dreiecksmatrix von A abspeichert und $D^{-1/2}$ in einem Vektor d:

Algorithmus 2.10 (Cholesky Verfahren)

```
Für i = 1,...,n:
    Für j = i,...,n:
        Setze u := a_ij
        Für k = i − 1,...,1:
            Setze u := u − a_jk a_ik
        Falls i = j,
            Setze d_i := 1/√u   (Abbruch, falls u ≤ 0)
        Sonst
            Setze a_ji := d_i u
```

Dieser Algorithmus hat aufgrund der Symmetrie den halben Aufwand im Vergeleich zum Gauß-Algorithmus.

2.1.4 LR-Zerlegung für Tridiagonalmatrizen

Sei A eine Tridiagonalmatrix

$$A = \begin{pmatrix} \alpha_1 & \gamma_1 & & 0 \\ \beta_1 & \ddots & \ddots & \\ & \ddots & \ddots & \gamma_{n-1} \\ 0 & & \beta_{n-1} & \alpha_n \end{pmatrix}$$

A kann mittels LR-Zerlegung zerlegt werden. Diese Zerlegung kann explizit in Abhängigkeit von α, β und γ hingeschrieben werden (siehe Übungsaufgabe).

2.2 Überbestimmte Gleichungssysteme/Ausgleichsrechnung

Problem: Gegeben sind m Messdaten (zum Beispiel Zeit und Konzentration) $(x_1, y_1), \ldots, (x_m, y_m)$ und Funktionen $u_1, \ldots, u_n,\ (n, m \in \mathbb{N},\ n \leq m)$.

Gesucht: Linearkombination $u(x) = \sum_{i=1}^{n} c_i u_i(x)$, welche die mittlere Abweichung minimier, also:

$$\triangle_2 := \left(\sum_{j=1}^{m} (u(x_j) - y_j)^2 \right)^{\frac{1}{2}} = \inf_{c_1,\ldots,c_n \in \mathbb{R}} \left(\sum_{j=1}^{m} \left(\sum_{i=1}^{n} (c_i u_i(x_j)) - y_j \right)^2 \right)^{\frac{1}{2}}$$

Dieses Problem wird als das **Gaußsche Ausgleichsproblem** oder als die Methode der kleinsten Quadrate (least squares) bezeichnet.

Bemerkung: Das **Tschebyscheffsche Ausgleichsproblem**, bei dem bezüglich der Maximumsnorm minimiert wird, d.h.

$$\triangle_\infty := \inf_{c_1,\ldots,c_n \in \mathbb{R}} \max_{i=1,\ldots,m} |c_i u_i(x_j) - y_j|,$$

ist deutlich schwieriger.

Sei:
$c = (c_1, \ldots, c_n)^\top \in \mathbb{R}^n$ (der gesuchte Lösungsvektor),
$x = (x_1, \ldots, x_m)^\top \in \mathbb{R}^m,\ y = (y_1, \ldots, y_m)^\top \in \mathbb{R}^m$,
$A = (a_{ij}) \in \mathbb{R}^{m \times n}$ mit $a_{ij} = u_i(x_j)$.

Dann ist das Ausgleichsproblem äquivalent zur Minimierung des Funktionals

$$F(c) := \|Ac - y\|_2 . \qquad (AGP)$$

Bemerkung: Sind $m = n,\ u_1, \ldots, u_n$ linear unabhängig und $x_1, \ldots, x_m$ paarweise verschieden, so ist A regulär und $c = A^{-1}y$ ist das gesuchte Minimum. Im Allgemeinen ist jedoch $n << m$, so dass $\text{Rang}(A) \leq n$ folgt. In solchen Fällen erwarten wir, dass (AGP) einen Minimierer hat, jedoch $Ac = y$ entweder keine oder sehr viele Lösungen hat.

Satz 2.11 (Normalengleichung)
Sei $A \in \mathbb{R}^{m \times n},\ b \in \mathbb{R}^m\ (n \leq m)$ gegeben. Dann existiert mindestens eine Lösung $\overline{x} \in \mathbb{R}^n$ des Ausgleichsproblems $(Ax = b)$ mit kleinstem Fehlerquadrat, d.h. $\overline{x}$ minimiert $F(x) = \|Ax - b\|_2$. Dies ist äquvalent dazu, dass $\overline{x}$ die **Normalengleichung**

$$A^\top A x = A^\top b$$

löst. ist $\text{Rang}(A) = n$ (d.h. maximal), so ist die Lösung eindeutig bestimmt, andernfalls ist jede weitere Lösung von der Form

$$x = \overline{x} + y$$

mit $y \in \text{Kern}(A)$. In diesem Fall wird meistens die Lösung x_A mit minimaler 2-Norm gesucht, d.h.

$$\|x_A\| = \inf \left\{ \|x\|_2 \;\middle|\; x \text{ L"osung des Ausgleichsproblems} \right\}.$$

Diese Lösung ist eindeutig.

Lemma 2.12
Sei $A \in \mathbb{R}^{m\times n}$, dann gelten für $A^\top A \in \mathbb{R}^{n\times n}$ folgende Eigenschaften:

(i) $A^\top A$ ist symmetrisch.

(ii) $A^\top A$ ist positiv semidefinit. Falls $\text{Rang}(A) = n$ ist, so ist $A^\top A$ positiv definit.

(iii) $\text{Kern}(A^\top A) = \text{Kern}(A)$.

(iv) $\mathbb{R}^m = \text{Bld}(A) \oplus \text{Kern}(A^\top)$.

(v) $A^\top A$ und $AA^\top$ haben dieselben (positiven, reellen) Eigenwerte und es gilt dimKern$(A\top A - \lambda I)$ = dimKern$(AA\top - \lambda I)$ für alle Eigenwerte $\lambda > 0$.

(vi) $r = \text{Rang}(A) = \text{Rang}(A^\top A) = \text{Rang}(AA^\top) = \text{Rang}(A^\top)$
$= \left|\left\{\lambda > 0 \;\middle|\; \lambda \text{ EW von } A^\top A\right\}\right|$.

Beweis: (Lemma 2.12(iv))

Es gilt $\mathbb{R}^m = \text{Bld}(A) \oplus \text{Bld}(A)^\perp$, d.h. es ist zu zeigen: $\text{Bld}(A)^\perp = \text{Kern}(A^\top)$

Sei $y \in \text{Bld}(A)^\perp$, d.h. $\forall\, z \in \text{Bld}(A) :\ \langle y, z\rangle = 0$
$\iff \forall\, x \in \mathbb{R}^n :\ \langle y, Ax\rangle = 0 \iff \forall\, x \in \mathbb{R}^n :\ \langle A^\top y, x\rangle = 0$
$\iff A^\top y = 0 \iff y \in \text{Kern}(A^\top)$. □

Beweis: (Satz 2.11)

Wir zeigen zunächst die Äquivalenz des Minimierungsproblems mit dem Lösen der Normalengleichung. Sei $\overline{x}$ Lösung von $A^\top Ax = A^\top b$. Dann folgt

$$\begin{aligned}
\|b - Ax\|_2^2 &= \|b - A\overline{x} + A(\overline{x} - x)\|_2^2 \\
&= \langle b - A\overline{x} + A(\overline{x} - x), b - A\overline{x} + A(\overline{x} - x)\rangle \\
&= \langle b - A\overline{x}, b - A\overline{x}\rangle + 2\langle b - A\overline{x}, A(\overline{x} - x)\rangle + \langle A(\overline{x} - x), A(\overline{x} - x)\rangle \\
&= \|b - A\overline{x}\|_2^2 + \|A(\overline{x} - x)\|_2^2 + 2\left\langle A^\top(b - A\overline{x}), \overline{x} - x\right\rangle \\
&\geq \|b - A\overline{x}\|_2^2
\end{aligned}$$

Also gilt für alle $x \in \mathbb{R}^n :\ \|b - A\overline{x}\|_2 \leq \|b - Ax\|_2$.

Sei nun umgekehrt $\overline{x}$ eine Lösung des Minimierungsproblems, so folgt

$$\begin{aligned}
0 &= \tfrac{\partial}{\partial x_i}(F(x))^2_{|x=\overline{x}} = \tfrac{\partial}{\partial x_i}\left(\sum_{j=1}^m\left(\sum_{k=1}^n a_{jk}x_k - b_j\right)^2\right)_{|x=\overline{x}} \\
&= \sum_{j=1}^n a_{ji}2\left(\sum_{k=1}^n a_{jk}\overline{x}_k - b_j\right) = 2\Big(\sum_{j=1}^m a_{ji}\sum_{k=1}^n a_{jk}\overline{x}_k - \underbrace{\sum_{j=1}^m a_{ji}b_j}_{a_{ij}^\top}\Big) \\
&= 2(A^\top A\overline{x} - A^\top b)_i.
\end{aligned}$$

Also folgt $A^\top A\overline{x} = A^\top b$ und somit ist $\overline{x}$ L"osung der Normalengleichung.

Insbesondere kann also die Existenz einer Lösung des AGPs durch das Lösen der Normalengleichung gezeigt werden.
Zur Lösung der Normalengleichung: Es ist $b \in \mathbb{R}^m = \mathrm{Bld}(A) \oplus \mathrm{Kern}(A^\top)$. Also können wir $b = s + r$ zerlegen mit $s \in \mathrm{Bld}(A),\ r \in \mathrm{Kern}(A^\top)$. Zu $s \in \mathrm{Bld}(A)$ existiert ein $\overline{x} \in \mathbb{R}^n$ mit $A\overline{x} = s$. Es folgt:

$$A^\top A\overline{x} = A^\top s + A^\top r = A^\top(r+s) = A^\top b,$$

d.h. $\overline{x}$ ist Lösung der Normalengleichung.

Ist $\mathrm{Rang}(A) = n$, so folgt, dass $A^\top A$ positiv definit (Lemma 2.12(ii)) und somit auch regulär ist. Insbesondere folgt daraus, dass $\overline{x} = (A^\top A)^{-1} A^\top b$ die eindeutige Lösung der Normalengleichung ist.

Ist $\mathrm{Rang}(A) < n$ und seien x_1, x_2 Lösungen der Normalengleichung. Dann gilt $b = Ax_i + (b - Ax_i) \in \mathrm{Bld}(A) \oplus \mathrm{Kern}(A^\top)$.

Da die Zerlegung $\mathbb{R}^m = \mathrm{Bld}(A) \oplus \mathrm{Kern}(A^\top)$ eindeutig ist, folgt $Ax_i = s = A\overline{x}$, also $A(x_i - \overline{x}) = 0$, d.h. $x_i - \overline{x} \in \mathrm{Kern}(A)$.

Wir definieren die Lösungsmenge K durch

$$K := \left\{ x \in \mathbb{R}^n \;\middle|\; x \text{ L"osung des AGPs und } \|x\|_2 \le \|\overline{x}\|_2 \right\}.$$

Dann ist K kompakt und da die Norm $\|\cdot\|_2$ stetig ist, nimmt sie ihr Minimun auf K an. Sei also x_A die Lösung des AGPs mit

$$\|x_A\|_2 = \inf\left\{ \|x_2\| \;\middle|\; x \in K \right\} =: \rho.$$

Sind nun $x_1, x_2 \in K$ Lösungen mit $\|x_1\|_2 = \|x_2\|_2 = \rho$, so folgt $\frac{x_1+x_2}{2} \in K$ und daher

$$\rho \le \left\| \frac{x_1 + x_2}{2} \right\|_2 \le \frac{1}{2}\|x_1\|_2 + \frac{1}{2}\|x_2\|_2 = \rho.$$

Wir erhalten somit $\left\|\frac{x_1+x_2}{2}\right\|_2 = \rho$ und es folgt

$$\begin{aligned}
\rho^2 &= \left\|\tfrac{x_1+x_2}{2}\right\|_2^2 = \tfrac{1}{4}\langle x_1 + x_2, x_1 + x_2\rangle \\
&= \tfrac{1}{4}\left(\|x_1\|_2^2 + 2\langle x_1, x_2\rangle + \|x_2\|_2^2\right) \\
&= \tfrac{1}{4}\left(\rho^2 + 2\langle x_1, x_2\rangle + \rho^2\right) = \tfrac{1}{2}\rho^2 + \tfrac{1}{2}\langle x_1, x_2\rangle, \\
&\Longrightarrow \langle x_1, x_2\rangle = \rho^2, \\
&\Longrightarrow \|x_1 - x_2\|_2^2 = \|x_1\|_2^2 - 2\langle x_1, x_2\rangle + \|x_2\|_2^2 = 2\rho^2 - 2\rho^2 = 0, \\
&\Longrightarrow x_1 = x_2.
\end{aligned}$$

□

Beispiel 2.13 (Ausgleichsgerade)
Gegeben: Messdaten:

x_i	-2	-1	0	1	2
y_i	1/2	1/2	2	7/2	7/2

Gesucht: Ausgleichungsgerade (linear fit[1]) $u(x) = bx + a$ mit

$$\left(\sum_{j=1}^{5} (bx_j + a - y_i)^2\right)^{\frac{1}{2}} = \min_{(\hat{a},\hat{b})\in\mathbb{R}^2} \left(\sum_{j=1}^{5} \left(\hat{b}x_j + \hat{a} - y_i\right)^2\right)^{\frac{1}{2}}.$$

[1] englischer Ausdruck der Ausgleichungsgerade

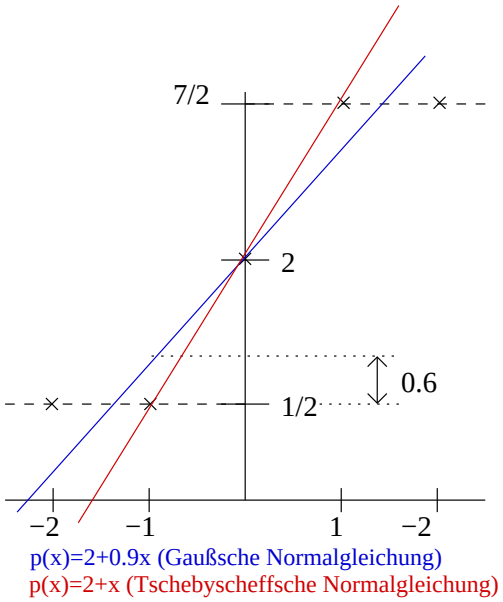

Abbildung 2.1: Ausgleichsgerade

Nach Satz 2.12 ist dann $(a,b)^\top$ die Lösung der Normalengleichung mit

$$A = \begin{pmatrix} 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix};\ c = \begin{pmatrix} 1/2 \\ 1/2 \\ 2 \\ 7/2 \\ 7/2 \end{pmatrix}$$

und folglich $A^\top A = \begin{pmatrix} 5 & 0 \\ 0 & 10 \end{pmatrix}$ und $A^\top c = \begin{pmatrix} 10 \\ 9 \end{pmatrix}$

Da $\text{Rang}(A) = 2$ ist, ist die Normalengleichung eindeutig lösbar. Aus $A^\top A \begin{pmatrix} a \\ b \end{pmatrix} = A^\top c$ folgt $\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 2 \\ 0.9 \end{pmatrix}$, d.h. $u(x) = 2 + 0.9x$ ist die Ausgleichsgerade.

Berechnet man die Abweichung, so folgt $\triangle_2 = \sqrt{0.9} < 1$, $\triangle_\infty = 0.6$.

Die Lösung des Tschebyscheffs-Problems ist gegeben durch $u(x) = 2 + x$. Hier erhält man $\triangle_2 = 1$; $\triangle_\infty = \frac{1}{2}$.

Bemerkung: Physikalisch könnte z.B. ein nichtlinearer Zusammenhang der Form $u(x) = \frac{a}{1+bx}$ sinnvoller sein.

Mithilfe der Transformation $\hat{u}(x) = \frac{1}{u(x)} = \frac{1}{a} + \frac{b}{a}x = \hat{a} + \hat{b}x$ lassen sich jedoch solche Probleme oft auf die Berechnung der linearen Ausgleichungsgeraden zurückführen.

Bemerkung: Das Ausgleichsproblem kann durch Lösen der Normalengleichung (etwa LR-Zelegung) behandelt werden. Allerdings ist dies numerisch nicht unbedingt der beste Zugang, da $cond(A^\top A)$ sehr viel größer als $cond(A)$ ist. Beispiel: gilt $\text{Rang}(A) = n$, $A \in \mathbb{R}^{n\times n}$, dann ist $cond(A^\top A) \sim cond(A)^2$.

2.2.1 QR-Zerlegung nach Householder

Idee: Sei $\text{Rang}(A) = n,\ A \in \mathbb{R}^{m\times n}$. Anstelle einer LR-Zerlegung sei

$$A = QR$$

mit einer oberen $\triangle$-Matrix $R \in \mathbb{R}^{m\times n}$ und einer orthogonalen Matrix $Q \in \mathbb{R}^{m\times m}$ gegeben, d.h. $Q^{-1} = Q^\top$. Dann folgt

$$A = QR = Q \left(\begin{array}{ccc} * & & * \\ & \ddots & \\ 0 & & * \\ \hline & 0 & \end{array}\right) = Q \begin{pmatrix} \widetilde{R} \\ \hline 0 \end{pmatrix},$$

wobei $\widetilde{R}$ eine reguläre obere $\triangle$-Matrix ist. Durch Einsetzen erhalten wir

$$A^\top A = (QR)^\top QR = R^\top Q^\top QR = R^\top R,$$

$$A^\top b = R^\top Q^\top b,$$

d.h. es gilt $A^\top A\overline{x} = A^\top b \iff R^\top R\overline{x} = R^\top Q^\top b$.

Definiere $\tilde{c}$ durch $c := Q^\top b = \begin{pmatrix} c_1 \\ \vdots \\ c_n \\ c_{n+1} \\ \vdots \\ c_m \end{pmatrix}$; $\widetilde{c} = \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix} \in \mathbb{R}^n$,

so folgt aus $R^\top = \Big(\underbrace{\widetilde{R}^\top}_{n} \;\Big|\; \underbrace{0}_{m-n}\Big)$: $R^\top c = \begin{pmatrix} \widetilde{R}^\top \widetilde{c} \\ 0 \end{pmatrix}$.

$\widetilde{R}^\top$ ist regulär. Sei also $\overline{x} \in \mathbb{R}^n$ die Lösung von $\widetilde{R}\overline{x} = \widetilde{c}$, so ist $\overline{x}$ leicht zu berechnen, da $\widetilde{R}$ obere $\triangle$-Matrix ist.

Da $R\overline{x} = \begin{pmatrix} \widetilde{c} \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ folgt $R^\top R\overline{x} = R^\top c = R^\top Q^\top b$ und somit ist $\overline{x}$ die Lösung der Normalengleichung.
Konzentrieren wir uns also auf die Berechnung einer QR Zerlegung.

QR-Zerlegung nach Householder

Sei $A \in \mathbb{R}^{m\times n}$ $(m \geq n)$ mit $\text{Rang}(A) = n$.

Ziel: Finde obere $\triangle$-Matrix $R \in \mathbb{R}^{m\times n}$ und eine orthogonale Matrix $Q \in \mathbb{R}^{m\times m}$ mit $A = QR$.

Definition 2.14 (Dyadisches Produkt)
Seien $u, v \in \mathbb{R}^m$ (Spaltenvektoren). Dann heißt die Matrix

$$A = uv^\top = \begin{pmatrix} u_1 \\ \vdots \\ u_m \end{pmatrix} (v_1, \dots, v_m)$$

das **dyadische Produkt** *von u, v.*
Es gilt $A \in \mathbb{R}^{m\times m}$ und $a_{ij} = u_i v_j$ $(1 \le i, j \le m)$.

Beachte: $\langle u, v\rangle = u^\top v = (u_1, \dots, u_n) \begin{pmatrix} v_1 \\ \vdots \\ v_m \end{pmatrix} \in \mathbb{R}$.

Folgerung 2.15
Seien $u, v \in \mathbb{R}^m$, $A = uv^\top$ und $w \in \mathbb{R}^m$. Dann gilt

(i) $Aw = \langle v, w\rangle\, u$,

(ii) $A^2 = \langle u, v\rangle\, A$.

Beweis:

(i) $(Aw)_i = \sum_{k=1}^{m} u_i v_k w_k = \langle v, w\rangle\, u_i$.

(ii) $(A^2)_{ij} = (AA)_{ij} = \sum_{k=1}^{m} u_i v_k u_k v_j = \left(\sum_{k=1}^{m} v_k u_k\right) u_i v_j = \langle v, u\rangle\, (A)_{ij}$.

□

Definition 2.16 (Householder Matrix)
Sei $v \in \mathbb{R}^m$, $v \neq 0$. Die Matrix $H(v) = \mathbb{I} - 2\frac{vv^\top}{\|v\|_2^2}$ heißt **Householder Matrix**.
Wir setzen $H(0) = \mathbb{I}$.

Folgerung 2.17
Sei $v \in \mathbb{R}^m$, dann gilt:

(i) $H(v)$ ist symmetrisch.

(ii) $H(v)$ ist orthogonal, d.h. $H(v) = H(v)^\top = H(v)^{-1}$.

Beweis:

(i) Es ist $H(v)_{ij} = \delta_{ij} - 2\frac{v_i v_j}{\|v\|_2^2} = \delta_{ji} - 2\frac{v_j v_i}{\|v\|_2^2} = H(v)_{ij}^\top$. Dabei bezeichnet δ_{ij} das Kronecker Symbol.

(ii) Wegen (i) bleibt zu zeigen, dass $H(v)^2 = \mathbb{I}$ gilt. Es ist

$$\begin{aligned} H(v)^2 &= \left(\mathbb{I} - 2\frac{vv^\top}{\|v\|_2^2}\right)\left(\mathbb{I} - 2\frac{vv^\top}{\|v\|_2^2}\right) = \mathbb{I} - 4\frac{vv^\top}{\|v\|_2^2} + 4\frac{(vv^\top)^2}{\|v\|_2^4} \\ &= \mathbb{I} - \frac{4}{\|v\|_2^2}\left(vv^\top - \frac{(vv^\top)^2}{\|v\|_2^2}\right) \\ &= \mathbb{I} - \frac{4}{\|v\|_2^2}\left(vv^\top - \frac{\langle v,v\rangle vv^\top}{\|v\|_2^2}\right) \text{ (wegen 2.15(ii))} \\ &= \mathbb{I}. \end{aligned}$$

□

Satz 2.18
Sei $a \in \mathbb{R}^m$ und $u := a \pm \|a\|_2 e_k \in \mathbb{R}^m$ $(1 \le k \le m)$. Dann gilt

$$H(u)a = \mp \|a\|_2 e_k.$$

Beweis: Im Fall $u = 0$ gilt aufgrund der Definition von u $a = \mp\|a\|_2 e_k$. Also folgt $H(u)a = \mathbb{I}a = \mp\|a\|_2 e_k$.

Sei also $u \neq 0$, etwa $u = a - \|a\|_2 e_k$. Dann folgt

$$H(u) = \mathbb{I} - \frac{uu^\top}{h} \quad \text{mit} \quad h = \frac{1}{2}\|u\|_2^2.$$

Und weiter

$$\begin{aligned} h &= \tfrac{1}{2}\langle a - \|a\|_2 e_k, a - \|a\|_2 e_k\rangle = \tfrac{1}{2}\left(\|a\|_2^2 - 2\|a\|_2\langle a, e_k\rangle + \|a\|_2^2\right) \\ &= \|a\|_2^2 - \|a\|_2\langle a, e_k\rangle, \end{aligned}$$

$$\begin{aligned} H(u)a &= \left(\mathbb{I} - \tfrac{1}{h}uu^\top\right)a = a - \tfrac{1}{h}(uu^\top)a \\ &= a - \tfrac{1}{h}\langle u, a\rangle u \qquad \text{(Folgerung 2.15(i))} \\ &= a - \tfrac{1}{h}\langle a - \|a\|_2 e_k, a\rangle u \\ &= a - \tfrac{1}{h}\left(\|a\|_2^2 - \|a\|_2\langle e_k, a\rangle\right)u \\ &= a - u = \|a\|_2 e_k \qquad \text{(Definition von } u\text{)}. \end{aligned}$$

□

Verfahren: QR Zerlegung

Sei $A \in \mathbb{R}^{m\times n}$ mit $A = (a_1, \ldots, a_n) = (a_1^{(0)}, \ldots, a_n^{(0)})$ gegeben.

Schritt 1:

Setze $u^{(0)} = (a_1^{(0)}) - \left\|a_1^{(0)}\right\|_2 e_1 \in \mathbb{R}^m$ und $Q_1 = H(u^{(0)})$, dann folgt

$$Q_1 A = R^{(1)} = \begin{pmatrix} * & \cdots & * \\ 0 & & \\ \vdots & A^{(1)} & \\ 0 & & \end{pmatrix} \text{ mit } A^{(1)} \in \mathbb{R}^{m-1\times n-1} = \left(a_2^{(1)}, \ldots, a_n^{(1)}\right).$$

Schritt 2:

Setze $u^{(1)} = a_2^{(1)} - \left\| a_2^{(1)} \right\| e_1 \in \mathbb{R}^{m-1}$ und $Q_2 := \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & H(u^{(1)}) & \\ 0 & & & \end{pmatrix}$.

Dann ist $Q_2 \in \mathbb{R}^{m \times m}$, $H(u^{(1)}) \in \mathbb{R}^{m-1 \times m-1}$ und es folgt

$$Q_2 Q_1 A = Q_2 R^{(1)} = \begin{pmatrix} * & \cdots & \cdots & 0 \\ 0 & * & \cdots & * \\ \vdots & 0 & & \\ \vdots & \vdots & A^{(2)} & \\ 0 & 0 & & \end{pmatrix}.$$

Iterativ erhalten wir so nach n Schritten:

$$Q_n \cdots Q_1 A = R^{(n)} = \left(\begin{array}{ccc} * & \cdots & * \\ & \ddots & \\ 0 & & * \\ \hline & 0 & \end{array} \right) = R$$

und $Q := Q_1 \cdots Q_n$ ist orthogonal, da alle Q_i orthogonal sind. Ausserdem gilt $A = QR$, da $Q_i^2 = \mathbb{I}$ und somit $QQ_n \cdots Q_1 = \mathbb{I}$.

Wir fassen die bisherigen Ergebnisse zusammen:

$$\begin{array}{lll} \text{Ausgleichsproblem} & \Longleftrightarrow & \text{Normalengleichung} \\ & \Longleftrightarrow & \tilde{R}\overline{x} = \widetilde{c} \text{ mit } A = QR = \left(\dfrac{\widetilde{R}}{0} \right), \ R \\ & & \text{und } R \text{ ist reguläre obere } \triangle\text{-Matrix,} \\ & & \text{falls } \mathrm{Rang}(A) = n \text{ gilt.} \end{array}$$

Betrachten wir nun die Kondition der Matrix R. Falls $A \in \mathbb{R}^{n \times n}$ und $\mathrm{Rang}(A) = n$ ist, gilt

$$\begin{array}{rcl} cond_2(A) & = & \|A\|_2 \left\|A^{-1}\right\|_2 \\ & = & \|QR\|_2 \left\|R^{-1}Q^{-1}\right\|_2 \\ & = & \|QR\|_2 \left\|R^{-1}Q^{\top}\right\|_2 \\ & = & \|R\|_2 \cdot \left\|R^{-1}\right\|_2 \end{array}.$$

Also folgt $\Longrightarrow$ $cond_2(A) = cond_2(R)$.

2.2.2 Singulärwertzerlegung einer Matrix

Die QR-Zerlegung liefert eine Möglichkeit, (AGP) numerisch zu lösen, falls $\mathrm{Rang}(A) = n$ ist. Für Probleme mit $\mathrm{Rang}(A) \leq n$ betrachten wir nun die Singulärwertzerlegung einer Matrix.

Satz 2.19 (Singulärwertzerlegung)

Sei $A \in \mathbb{R}^{m\times n}$, $\text{Rang}(A) = r$, $p = \min\{m,n\}$. Dann existieren orthogonale Matrizen $U = (u_1,\ldots,u_m) \in \mathbb{R}^{m\times m}$ und $V = (v_1,\ldots,v_n) \in \mathbb{R}^{n\times n}$ mit $U^\top AV = \Sigma \in \mathbb{R}^{m\times n}$ mit

$$\Sigma = \text{diag}(\sigma_1,\ldots,\sigma_p),$$

wobei $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > \sigma_{r+1} = \ldots = \sigma_p = 0$.

D. h. Σ hat die Form

$$\Sigma = \left(\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_r & \\ \hline & 0 & & 0 \end{array}\right),$$

mit $\text{diag}(\sigma_1,\ldots,\sigma_r) \in \mathbb{R}^{r\times r}$.

Die Werte $\sigma_1,\ldots,\sigma_r$ heißen **singuläre Werte** von A. Sie entsprechen gerade den Wurzeln aus den Eigenwerten von $A^\top A$ bzw. $AA^\top$ (Bem: nach 2.12(v) haben $A^\top A$ und $AA^\top$ dieselben positven und reellen Eigenwerte).

Beweis: Eindeutigkeit: Seien $U^\top AV = \Sigma$ und U, V orthogonal, dann gelten $Av_i = \sigma_i u_i$, da $AV = U\Sigma$ und $A^\top u_i = \sigma_i v_i$, da $A^\top U = V\Sigma$. Daraus ergibt sich

$$A^\top Av_i = \sigma_i A^\top u_i = \sigma_i^2 v_i \implies \sigma_i^2 \text{ ist Eigenwert von } A^\top A.$$

Analog folgt

$$AA^\top u_i = \sigma_i^2 u_i \implies \sigma_i^2 \text{ ist Eigenwert von } AA^\top.$$

Da nach 2.12(v) die Eigenwerte von $A^\top A$ und $AA^\top$ übereinstimmen, folgt die Eindeutigkeit.

Existenz: Sei $\sigma_1 := \|A\|_2 = \max\limits_{\|x\|_2=1} \|Ax\|_2$ (Bemerkung: Da $\|A\|_2 = \sqrt{\lambda_{\max}(A^\top A)}$, ist σ_1 ein guter Kandidat).

Dann existieren $x_1 \in \mathbb{R}^n, y_1 \in \mathbb{R}^m$ mit $\|x_1\|_2 = 1$, $\|y_1\|_2 = 1$ und $Ax_1 = \sigma_1 y_1$. Sei $V = (x_1,\ldots,x_n) \in \mathbb{R}^{n\times n}$ eine orthonormale Basis (ONB) des $\mathbb{R}^n$ und $U_1 = (y_1,\ldots,y_m) \in \mathbb{R}^{m\times m}$ eine ONB des $\mathbb{R}^m$.

Dann folgt:

$$A_1 := U_1^\top AV_1 = \left(\begin{array}{c|c} \sigma_1 & w^\top \\ \hline 0 & B \end{array}\right),\ B \in \mathbb{R}^{m-1\times n-1},\ w \in \mathbb{R}^{n-1}.$$

Da U_1, V_1 orthogonal, gilt:

$$\|A_1\|_2 = \|A\|_2 = \sigma_1.$$

Desweiteren gilt

$$A_1 \begin{pmatrix} \sigma_1 \\ w \end{pmatrix} = \begin{pmatrix} \sigma_1^2 + w^\top w \\ Bw \end{pmatrix} = \begin{pmatrix} \sigma_1^2 + \|w\|_2^2 \\ Bw \end{pmatrix}$$

und daher

$$\begin{aligned} \sigma_1^2 = \|A_1\|_2^2 &= \left(\max_x \tfrac{\|A_1x\|_2}{\|x\|_2}\right)^2 \geq \tfrac{1}{\|(\sigma_1,w)^\top\|_2^2} \left\|A_1 \begin{pmatrix} \sigma_1 \\ w \end{pmatrix}\right\|_2^2 \\ &= \tfrac{1}{(\sigma_1^2+\|w\|_2^2)} \left(\left(\sigma_1^2 + \|w\|_2^2\right)^2 + \underbrace{\|Bw\|_2^2}_{\geq 0}\right) \\ &\geq \tfrac{1}{\sigma_1^2+\|w\|_2^2} \left(\sigma_1^2 + \|w\|_2^2\right)^2 = \sigma_1^2 + \|w\|_2^2 \end{aligned}.$$

Insgesamt folgt also

$$\sigma_1^2 \geq \sigma_1^2 + \|w\|_2^2 \implies \|w\|_2^2 = 0 \implies w = 0$$

und wir erhalten

$$U_1^\top A V_1 = \left(\begin{array}{c|c} \sigma & 0 \\ \hline 0 & B \end{array}\right).$$

Die Aussage des Satzes folgt nun durch Induktion. □

Lösung des Ausgleichsproblems mit Singulärwertzerlegung:

Gesucht: $x \in \mathbb{R}^n$ mit $\|Ax - b\|_2 = \inf_{z \in \mathbb{R}^n} \|Az - b\|_2$, $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $m \geq n \geq r = \text{Rang}(A)$.

Sei $U^\top A V = \Sigma = diag(\sigma_1, \ldots, \sigma_n)$, so folgt

$$\begin{array}{rcl} \|Ax - b\|_2^2 & = & \langle Ax - b, Ax - b\rangle \overset{\text{da } U^\top = \text{orth.}}{} \langle U^\top(Ax - b), U^\top(Ax - b)\rangle \\ & \overset{VV^\top = \mathbb{I}}{=} & \langle U^\top A V(V^\top x) - U^\top b, U^\top A V(V^\top x) - U^\top b\rangle \\ & = & \langle \Sigma V^\top x - U^\top b, \Sigma V^\top x - U^\top b\rangle \\ & = & \left\|\Sigma V^\top x - U^\top b\right\|_2^2 \\ & = & \sum_{i=1}^{r} \left(\sigma_i (V^\top x)_i - u_i^\top b\right)^2 + \sum_{i=r+1}^{m} \left(u_i^\top b\right)^2 \\ & \geq & \sum_{i=r+1}^{m} \left(u_i^\top b\right) \end{array}.$$

Folgerung 2.20
$x \in \mathbb{R}^n$ ist genau dann Lösung des Ausgleichsproblems, wenn

$$V^\top x = \left(\frac{u_1^\top b}{\sigma_1}, \ldots, \frac{u_r^\top b}{\sigma_r}, \alpha_{r+1}, \ldots, \alpha_n\right), \text{ mit } \alpha_i \in \mathbb{R}\text{beliebig.}$$

Ist x Lösung des AGPs, so ist $\|x\|_2^2 \overset{V^\top \text{ orth.}}{=} \left\|V^\top x\right\|_2^2 = \sum_{i=1}^{r} \left(\frac{u_i^\top b}{\sigma_i}\right)^2 + \sum_{i=r+1}^{n} \alpha_i^2$.

Also ist $\|x\|_2$ minimal, g.d.w. $\alpha_{r+1} = \ldots = \alpha_n = 0$ ist. D.h. die eindeutige Lösung des AGPs mit minimaler 2-Norm ist gegeben durch

$$x = V\left(\frac{u_i^\top b}{\sigma_1}, \ldots, \frac{u_r^\top b}{\sigma_r}, 0, \ldots, 0\right)^\top = \sum_{i=1}^{r} \frac{u_i^\top b}{\sigma_i} v_i.$$

2.2.3 Pseudoinverse einer Matrix

Ziel: Verallgemeinerung der Inversen einer regulären Matrix auf beliebige Matrizen $A \in \mathbb{R}^{m \times n}$ mit

Hilfe der Singulärwertzerlegung.

Definition 2.21 (Pseudoinverse)
Zu $A \in \mathbb{R}^{m\times n}$ mit $Rang(A) = r$ sei

$$\Sigma := U^\top AV = \operatorname{diag}(\sigma_1, \ldots, \sigma_{\min\{m,n\}})$$

eine Singulärwertzerlegung von A. Wir definieren die $(n \times m)$-Matrix Σ^+ durch

$$\Sigma^+ := \left(\begin{array}{ccc|c} 1/\sigma_1 & & & \\ & \ddots & & 0 \\ & & 1/\sigma_r & \\ \hline & 0 & & 0 \end{array}\right),$$

*dann heißt $A^+ := V\Sigma^+U^\top \in \mathbb{R}^{n\times m}$ **Pseudoinverse** oder Penrose Inverse von A.*

Bemerkung: Die singulären Werte einer Matrix sind eindeutig bestimmt, die orthogonalen Matrizen U und V jedoch nicht. Die Eindeutigkeit der Pseudoinversen muß also noch gezeigt werden.

Satz 2.22
Sei $A \in \mathbb{R}^{m\times n}$ mit $Rang(A) = r$ gegeben. Dann gilt

(i) Ist $A^+ \in \mathbb{R}^{n\times m}$ eine Pseudoinverse zu A, so ist
$AA^+ = (AA^+)^\top, \quad A^+A = (A^+A)^\top, \quad AA^+A = A, \quad A^+AA^+ = A^+.$

(ii) Durch die "Penrose Bedingung"
$AB = (AB)^\top, \quad BA = (BA)^\top, \quad ABA = A, \quad BAB = B \qquad (*)$
ist eine Matrix $B \in \mathbb{R}^{n\times m}$ eindeutig bestimmt.
Insbesondere ist die Pseudoinverse wohldefiniert.

(iii) Sind $m \geq n$, $b \in \mathbb{R}^m$ und x_A die eindeutige lösung des Ausgleichsproblems, so ist $x_A = A^+b$.

(iv) Ist $m = n$ und $Rang(A) = n$, so ist $A^+ = A^{-1}$.

(v) Ist $m \geq n$ und $Rang(A) = n$, so ist $A^+ = (A^\top A)^{-1}A^\top$.

(vi) Sind $\sigma_1 \geq \ldots \sigma_r$ die singulären Werte von A, so ist $||A||_2 = \sigma_1$ und $||A^+||_2 = \frac{1}{\sigma_r}$.

Beweis: (siehe Übungsaufgabe) □

Bemerkung: Für die Pseudoinverse gilt auch $(A^+)^+ = A$, sowie $(A^\top)^+ = (A^+)^\top$, jedoch gilt im Allgemeinen nicht $(AB)^+ = B^+A^+$.

Definition 2.23 (Kondition einer singulären Matrix)
Für eine beliebige Matrix $A \in \mathbb{R}^{m\times n}$ definieren wir die Kondition von A bezüglich der Spektralnorm durch

$$cond_2(A) = ||A||_2||A^+||_2 = \frac{\sigma_1}{\sigma_r}.$$

2.3 Iterative Verfahren

Wir wollen uns nun iterativen Verfahren zur Lösung linearer Gleichungssysteme zuwenden.

Idee: Formuliere $Ax = b$ äquivalent als Fixpunktgleichung, so dass die Kontraktionsbedingung (vgl. Satz 1.29 auf Seite 10) erfüllt ist.

Ansatz: Wir zerlegen $A = M - N$ mit einer regulären Matrix M (i.A. ist M vorgegeben und $N := M - A$). Wir erhalten:

$$\begin{aligned} Ax = b \quad &\iff \quad (M-N)x = b \quad \iff \quad Mx - Nx = b \quad \iff \quad Mx = Nx + b \\ &\iff \quad x = M^{-1}Nx + M^{-1}b. \end{aligned}$$

Wir definieren $T := M^{-1}N = \mathbb{I} - M^{-1}A$, $c := M^{-1}b$ sowie eine Abbildung $F : \mathbb{R}^n \to \mathbb{R}^n$ durch $F(x) = Tx + c$.

Dann gilt: x ist die Lösung von $Ax = b$, g.d.w. x ein Fixpunkt von F ist, d.h. wenn gilt $x = F(x) = Tx + c$.

Iterationsverfahren: Sei ein Startwert $x^0 \in \mathbb{R}^n$ gegeben, dann definieren wir die Iteration für $k \in \mathbb{N}$ durch

$$x^{k+1} = F(x^k),$$

d.h. x^{k+1} wird berechnet durch folgende Schritte:

$$\begin{aligned} &1) \quad My^k &&= r^k, \quad \text{mit dem Residuum } r^k := b - Ax^k, \\ &2) \quad x^{k+1} &&= x^k + y^k. \end{aligned}$$

Bemerkung: Aus den Lösungsschritten erkennt man, dass solche Iterationsverfahren nur dann von Interesse sind, wenn die Defektgleichung $My^k = r^k$ leicht zu lösen ist. Dies ist z.B. der Fall, falls M eine Diagonalmatrix oder eine obere $\triangle$-Matrix ist.

Satz 2.24
Die Folge $(x^k)_{k\in\mathbb{N}}$ sei durch eine Fixpunktiteration zu $x = F(x) = Tx + c$ gegeben. Sei $\|\cdot\|$ eine Norm auf dem $\mathbb{R}^n$, so dass für die induzierte Matrixnorm gilt

$$q := \|T\| < 1.$$

Dann konvergiert x^k gegen ein x mit $Ax = b$ und es gelten die Fehlerabschätzungen

$$\left\|x - x^k\right\| \leq \frac{q^k}{1-q} \left\|x^1 - x^0\right\|,$$

bzw.

$$\left\|x - x^k\right\| \leq \frac{q}{1-q} \left\|x^k - x^{k-1}\right\|.$$

Beweis: Der Beweis folgt mit dem Banachschen Fixpunktsatz 1.29 da gilt

$$\begin{aligned} \|F(y_1) - F(y_2)\| &= \|Ty_1 + c - (Ty_2 + c)\| \\ &= \|T(y_1 - y_2)\| \leq \|T\| \, \|y_1 - y_2\| = q \, \|y_1 - y_2\| \end{aligned}.$$

Aus $q < 1$ folgt also, dass F eine Kontraktion ist.

Problem: Die Kontraktionsbedingung ist abhängig von der Norm, die Konvergenz nicht, da alle Normen auf $\mathbb{R}^n$ äquivalent sind.
Gesucht: Notwendige und hinreichende Bedingung für die Konvergenz.

Definition 2.25 (Spektralradius)
Für eine Matrix $B \in \mathbb{R}^{n\times n}$ definieren wir den **Spektralradius** $\rho(B)$ *durch*

$$\rho(B) := \max\{|\lambda| \mid \lambda \in \mathbb{C} \text{ ist Eigenwert von } B\}.$$

Bemerkung: Eine Matrix B hat in $\mathbb{C}$ die Eigenwerte $\lambda_1, \ldots, \lambda_n$ (falls Vielfachheit zugelassen wird). Es existiert eine reguläre Matrix $U \in \mathbb{C}^{n\times n}$ mit

$$U^{-1}BU = \begin{pmatrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix},$$

also eine Ähnlichkeitstransformation (Jordansche Normalform).

Lemma 2.26
Für $B \in \mathbb{R}^{n\times n}$ gilt

(i) $\rho(B) \le \|B\|$ für jede induzierte Matrixnorm.

(ii) $\forall\, \varepsilon > 0$ gibt es eine induzierte Norm $\|\cdot\|$ auf $\mathbb{C}^{n\times n}$ mit $\|B\| \le \rho(B) + \varepsilon$.

Beweis:

(i) Seien λ ein Eigenwert von B, $u \in \mathbb{C}^n\backslash\{0\}$ der zugehörige Eigenvektor und $\|\cdot\|$ eine Norm auf C^n. Dann folgt

$$\begin{aligned} Bu = \lambda u \quad &\Longrightarrow \quad |\lambda|\,\|u\| = \|\lambda u\| = \|Bu\| \le \|B\|\,\|u\| \\ &\Longrightarrow \quad |\lambda| \le \|B\| \text{ für alle EWe} \implies \rho(B) \le \|B\|. \end{aligned}$$

(ii) Sei $U \in \mathbb{C}^{n\times n}$ regulär mit $U^{-1}BU = \begin{pmatrix} \lambda_1 & & r_{ij} \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$.

Für $\delta > 0$ sei $D_\delta = diag(\delta^0, \ldots, \delta^{n-1})$. Für $G \in \mathbb{C}^{n\times n}$ gilt dann $\left(D_\delta^{-1} G D_\delta\right) = g_{ij}\delta^{j-i}$. Also folgt

$$(UD_\delta)^{-1} B (UD_\delta) = D_\delta^{-1}\left(U^{-1}BU\right) D_\delta = \begin{pmatrix} \lambda_1 & \delta r_{12} & \delta^2 r_{13} & \cdots & \delta^{n-1} r_{1n} \\ 0 & \lambda_2 & \delta r_{23} & \cdots & \delta^{n-2} r_{2n} \\ \vdots & & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & \delta r_{n-1,n} \\ 0 & \cdots & \cdots & \cdots & \lambda_n \end{pmatrix}.$$

Sei $\varepsilon > 0$ vorgegeben. Dann existiert ein $\delta > 0$ mit $\sum_{k=i+1}^{n} \delta^{k-i} |r_{ik}| \leq \varepsilon$ für alle $i \in \{1, \ldots, n-1\}$.

Setze $\|x\|_\delta = \left\| (UD_\delta)^{-1} x \right\|_\infty$. Diese ist eine Norm auf $\mathbb{C}^n$, da $(UD_\delta)^{-1}$ regulär. Es gilt:

$$\|B\|_\delta = \sup_{x \in \mathbb{C}^n \setminus \{0\}} \frac{\|Bx\|_\delta}{\|x\|_\delta} = \sup_{x \in \mathbb{C}^n \setminus \{0\}} \frac{\left\| (UD_\delta)^{-1} Bx \right\|_\infty}{\left\| (UD_\delta)^{-1} x \right\|_\infty}.$$

Mit $w := (UD_\delta)^{-1} x$ bzw. $x = (UD_\delta)\, w$ gilt:

Wenn x über $\mathbb{C}^n$ läuft, dann läuft auch w über den ganzen $\mathbb{C}^n$, da $(UD_\delta)^{-1}$ regulär, d.h. $\sup_{x \neq 0} \cdots = \sup_{w \neq 0} \cdots$. Also folgt

$$\begin{aligned}
\|B\|_\delta &= \sup_{w \neq 0} \frac{\left\| (UD_\delta)^{-1} B (UD_\delta) w \right\|_\infty}{\|w\|_\infty} \\
&\leq \left\| (UD_\delta)^{-1} B\, (UD_\delta) \right\|_\infty \\
&= \max_i \left\{ |\lambda_i| + \underbrace{\sum_{k=i+1}^{n} \delta^{k-1} \cdot |r_{ik}|}_{\leq \varepsilon} \right\} \text{ (Zeilensummennorm)} \\
&= \max_i |\lambda_i| + \varepsilon = \rho(B) + \varepsilon
\end{aligned}$$

Aus Lemma 2.26 folgt folgender Satz über die Konvergenz geometrischer Matrixfolgen. □

Satz 2.27
Sei $T \in \mathbb{C}^{n \times n}$ eine reguläre Matrix. Dann sind äquivalent

(i) $T^\nu \longrightarrow 0$ für $\nu \longrightarrow \infty$.

(ii) Für $u \in \mathbb{C}^n$ gilt: $T^\nu u \longrightarrow 0$ für $\nu \longrightarrow \infty$.

(iii) $\rho(T) < 1$.

(iv) Es existiert eine Norm auf $\mathbb{C}^n$, so dass für die induzierte Matrixnorm $\|T\| < 1$ gilt.

Beweis:
(i) $\Longrightarrow$ *(ii)*: Es gilt $\|T^\nu u\| \leq \|T^\nu\| \, \|u\| \longrightarrow 0$, da $T^\nu \longrightarrow 0$. Also folgt $T^\nu u \longrightarrow 0$ (für $\nu \longrightarrow \infty$).

(ii) $\Longrightarrow$ *(iii)*: Annahme: $\rho(T) \geq 1$, d.h. es existiert ein EW $\lambda \in \mathbb{C}$ mit $|\lambda| \geq 1$. Sei $u \in \mathbb{C}^n \setminus \{0\}$ der zugehörige EV, dann gilt $T^\nu u = T^{\nu-1}(Tu) = T^{\nu-1}(\lambda u) = \lambda T^{\nu-1} u = \cdots = \lambda^\nu u$. Weiter folgt $\|T^\nu u\| = \|\lambda^\nu u\| = |\lambda|^\nu \, \|u\| \not\to 0$, da $|\lambda| \geq 1$. Dies ist ein Widerspruch zu *(ii)*.
(iii) $\Longrightarrow$ *(iv)*: Lemma 2.26.

(iv) $\Longrightarrow$ *(i)*: Es gilt $\|T^\nu\| \overset{\text{Submultipl.}}{\leq} \|T\|^\nu \longrightarrow 0$, da $\|T\| < 1$. □

Folgerung 2.28
Das Iterationsverfahren konvergiert genau dann wenn $\rho(T) < 1$.

Bemerkung: (Ohne Beweis)

Aus den bisher gezeigten Sätzen folgt:

Um eine Dezimalstelle im Fehler zu gewinnen, müssen $K \sim -\frac{\ln 10}{\ln(\rho(T))}$ Schritte durchgeführt werden.

Das heißt für $\rho(T) \sim 1$ ist $-\ln(\rho(T)) \sim 0$ und K sehr groß. Somit muß das Ziel bei der Wahl der regulären Matrix M sein:

(a) $\rho(T) = \rho(\mathbb{I} - M^{-1}A)$ möglichst klein.

(b) Das Gleichungssystem $My^k = r^k$ muss leicht zu lösen sein.

Dies sind widersprüchliche Forderungen: Optimal für Bedingung (a) wäre $M = A \implies \rho(T) = 0$, aber dann ist die Bediengung (b) nicht erfüllt.
Auf der anderen Seite ist (b) erfüllt, falls M eine Diagonalmatrix ist. Dies führt uns zu folgendem Verfahren.

2.3.1 Gesamtschritt Verfahren (GSV)/ Jacobi Verfahren

Sei A regulär und $a_{ii} \neq 0$ für alle $i = 1, \ldots, n$. Setze

$$M := D = diag(a_{11}, \ldots, a_{nn}) \implies T = \mathbb{I} - D^{-1}A.$$

Dies führt zu folgender Iterationsvorschrift:

Sei ein Startvektor $x^0 \in \mathbb{R}^n$ gegeben.
Iteration:

$$\begin{aligned} x^{k+1} &= (\mathbb{I} - D^{-1}A)x^k + D^{-1}b \\ &= D^{-1}(Dx^k - Ax^k + b), \\ \implies x_i^{k+1} &= \frac{1}{a_{ii}}\left(b_i - \sum_{l \neq i} a_{il}x_l^k\right), \ i = 1, \ldots, n. \end{aligned}$$

Satz 2.29 (Hinreichende Bediengung für die Konvergenz des Jacobi-Verfahrens)
Falls entweder

$$\text{(a)} \ \max_i \left(\sum_{k \neq i} \frac{|a_{ik}|}{|a_{ii}|}\right) < 1 \qquad \textbf{(starkes Zeilensummenkriterium)}$$

oder

$$\text{(b)} \ \max_i \left(\sum_{k \neq i} \frac{|a_{ki}|}{|a_{ii}|}\right) < 1 \qquad \textbf{(starkes Spaltensummenkriterium)}$$

gilt, so konvergiert das Jacobi-Verfahren.

Falls (a) oder (b) erfüllt ist, heißt die Matrix A **stark diagonal dominant**, da in diesem Fall die Beträge der Diagonaleinträge größer sind als die Summe der Zeilen bzw. Spalten der Matrix.

Beweis: Gelte (a), so folgt

$$\begin{aligned}\rho(T) \le \|T\|_\infty &= \left\|\mathbb{I} - D^{-1}A\right\|_\infty \\ &= \max_i \left(\sum_{k=1}^{n} \left| \delta_{ik} - \frac{|a_{ik}|}{|a_{ii}|} \right| \right) \\ &= \max_i \left(\sum_{k \neq i} \frac{|a_{ik}|}{|a_{ii}|} \right) < 1 \text{ (wegen (a)).}\end{aligned}$$

Gelte (b), so folgt

$$\rho(T) \le \|T\|_1 = \max_i \left(\sum_{k \neq i} \frac{|a_{ki}|}{|a_{ii}|} \right) < 1 \text{ (wegen (b)).}$$ □

Die starke Diagonaldominanz ist nur eine hinreichende Bedingung. Betrachten wir folgendes Beispiel.

Beispiel 2.30
Bei der Diskretisierung $-\partial_{xx} u = f$ in §1.5 musste ein LGS $Ax = b$ gelöst werden mit

$$A = \begin{pmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix}$$

Für A gilt $\sum_{k \neq i} \frac{|a_{ik}|}{|a_{ii}|} = 1$ für $i = 2, \ldots, n-1$ und für $i = 1, n$ gilt $\sum_{k \neq i} \frac{|a_{ik}|}{|a_{ii}|} < 1$. Für dieses Beispiel ist also die starke Diagonaldominanz nicht erfüllt. Wir wollen nun eine schwächere Bedingung herleiten, die auch Matrizen eines solchen Typs mit beinhaltet.

Definition 2.31 (Zerlegbare Matrizen)
Eine Matrix $A = (a_{ik})$ heißt **zerlegbar**, *falls es eine Zerlegung von $N := \{1, \ldots, n\}$ in zwei Teilmengen $N_1, N_2 \subset N$ gibt mit $a_{ik} = 0 \ \forall (i,k) \in N_1 \times N_2$.*
(Zerlegung heißt: $N_1 \neq \emptyset$, $N_2 \neq \emptyset$, $N = N_1 \cup N_2$, $N_1 \cap N_2 = \emptyset$).

Lemma 2.32
Für $A \in \mathbb{R}^{n \times n}$ sind äquivalent

(i) A ist zerlegbar.

(ii) Der zugehörige gerichtete Graph
$$G(A) := \big(\text{Knoten } P_1, \ldots, P_n, \text{ gerichtete Kanten } \overline{P_j P_k} \iff a_{jk} \neq 0\big)$$
ist nicht zusammenhängend, d.h. es existieren Knoten P_j und P_k, so dass kein Pfad zwischen ihnen existiert. Es gibt also keine Folge $l_0, \ldots, l_L \in \{1, \ldots, N\}$ mit $l_0 = j, l_L = k$ und $a_{l_i l_{i+1}} \neq 0$ für alle $i = 0, \ldots, L$.

(iii) Es existiert eine Permutationsmatrix $P \in \mathbb{R}^{n \times n}$ mit

$$PAP^\top = \begin{pmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{pmatrix}$$

mit $A_{11} \in \mathbb{R}^{p \times p}$, $A_{22} \in \mathbb{R}^{q \times q}$, $A_{21} \in \mathbb{R}^{q \times p}$ und $p + q = n$.

Beweis: Wir zeigen $(i) \iff (ii)$. Die Äquivalenz mit (iii) wird in den Übungen behandelt.

$(i) \implies (ii)$: Sei A zerlegbar, d.h. es existieren $N_1 \neq \emptyset$, $N_2 \neq \emptyset$, $N = N_1 \cup N_2$, $N_1 \cap N_2 = \emptyset$ und es gilt $a_{ik} = 0 \; \forall (i,k) \in N_1 \times N_2$. Sei $(i,k) \in N_1 \times N_2$. Annahme: $G(A)$ ist zusammenhängend. Dann existiert eine Folge $l_0, \ldots, l_L \in \{1, \ldots, N\}$ mit $l_0 = i$ und $l_L = k$ und $a_{l_i l_{i+1}} \neq 0$ für alle $i = 0, \ldots, L$. Nach Voraussetzung ist $l_0 = i \in N_1$. Da $a_{l_0 l_1} \neq 0$, ist $l_1 \notin N_2 \implies l_1 \in N_1$ und induktiv folgt somit $l_L = k \in N_1$. Dies ist ein Widerspruch zur Annahme $k \in N_2$.

$(ii) \implies (i)$: Sei nun $G(A)$ nicht zusammenhängend, existiere etwa kein Pfad zwischen P_j und P_k. Setze $N_1 := \{l |$ es existiert ein Pfad zwischen P_j und $P_l\} \cup \{j\}$, $N_2 := N \setminus N_2$. Dann gilt $k \in N_2$. Also folgt $N_1 \neq \emptyset$, $N_2 \neq \emptyset$, $N = N_1 \cup N_2$, $N_1 \cap N_2 = \emptyset$. Sei $(i,l) \in N_1 \times N_2$. Wäre $a_{il} \neq 0$, so würde ein Pfad von P_j zu P_l existieren und somit wäre $l \in N_1$. Dies ist ein Widerspruch und folglich gilt $a_{il} = 0$ für alle $(i,l) \in N_1 \times N_2$.

□

Beispiel 2.33
Wir betrachten die Matrix

$$A = \begin{pmatrix} 2 & 0 & 2 \\ 2 & 2 & 1 \\ 0 & 0 & 1 \end{pmatrix}.$$

Dann ist der Graph $G(A)$ wie in Abb. 2.2 gegeben. Es gelten:

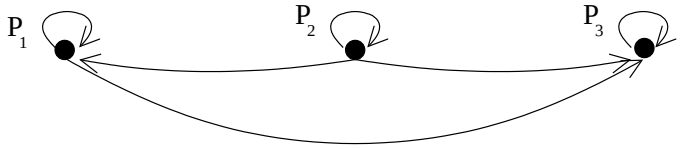

Abbildung 2.2: Graph, Beispiel 2.33

(i) $N_1 = \{1, 3\}$, $N_2 = \{2\}$ ist eine geeignete Zerlegung.

(ii) $G(A)$ nicht zusammenhängend, da es keine Verbindung zwischen P_1 und P_2 existiert. (Siehe Abbildung 2.2).

(iii)

$$P = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}; \; PAP^\top = \left(\begin{array}{c|cc} 2 & 0 & 0 \\ \hline 1 & 2 & 2 \\ 2 & 0 & 2 \end{array}\right).$$

Beispiel 2.34
Sei A eine Tridiagonalmatrix ohne Nullen auf der Diagonalen und den Nebendiagonalen, d.h.

$$A = \begin{pmatrix} * & * & & 0 \\ * & \ddots & \ddots & \\ & \ddots & \ddots & * \\ 0 & & * & * \end{pmatrix}$$

Dann ist A unzerlegbar, da der Graph zusammenhängend ist. (Man kommt von jedem inneren Knoten zu den Nachbarn rechts und links. Siehe Übungsaufgabe)

Satz 2.35 (Schwaches Zeilensummenkriterium)
Sei $A \in \mathbb{R}^{n\times n}$ unzerlegbar und erfülle das schwache Zeilensummenkriterium, d.h.

$$\max_i \left(\sum_{k\neq i} \frac{|a_{ik}|}{|a_{ii}|} \right) \leq 1$$

und es existiert ein $r \in \{1,\dots,n\}$ mit

$$\sum_{k\neq r} \frac{|a_{rk}|}{|a_{rr}|} < 1.$$

Dann kann das Jacobi-Verfahren angewendet werden und es konvergiert für alle Startvektoren $x^0 \in \mathbb{R}^n$.

Beweis: Wir zeigen zunächst, dass gilt $|a_{ii}| > 0$. Dazu nehmen wir an, dass gilt $\sum_{k\neq i} |a_{ik}| \leq |a_{ii}|$. Da A unzerlegbar ist, folgt $\sum_{k\neq i} |a_{ik}| > 0$, und somit $|a_{ii}| > 0$. Insbesondere kann also das Jacobi-Verfahren angewendet werden.
Analog zum Beweis von Satz 2.29 können wir zeigen: $\rho(T) \leq 1$ mit $T = M^{-1}N$. Wir müssen also noch zeigen, dass $\rho(T) \neq 1$ ist.
Annahme: Es existiert ein Eigenwert $\lambda \in \mathbb{C}$ mit $|\lambda| = 1$. Sei $v \in \mathbb{C}^n$ der zugehörige Eigenvektor mit $\|v\|_\infty = 1$, d.h. $\exists\, s \in \{1,\dots,n\}$ mit $|v_s| = 1$. Aus $Tv = \lambda v$ folgt für $i = 1,\dots,n$:

$$\lambda v_i = \sum_{k=1}^n t_{ik} v_k = \sum_{k=1}^n \left(\delta_{ik} - \frac{a_{ik}}{a_{ii}} \right) v_k = \sum_{k\neq i} \frac{a_{ik}}{a_{ii}} v_k.$$

Also folgt

$$|v_i| = |\lambda|\,|v_i| \leq \frac{1}{|a_{ii}|} \sum_{k\neq i} |a_{ik}|\,|v_k|\,. \qquad (*)$$

Da $G(A)$ zusammenhängend ist, existiert ein Pfad zwischen P_s und P_r, d.h. $l_0 = s,\dots,l_L = r$ und $a_{l_i,l_{i+1}} \neq 0$ $(i = 0,\dots,L-1)$. Mit $(*)$ folgt also:

$$|v_r| \leq \frac{1}{|a_{rr}|} \sum_{k\neq r} |a_{rk}|\,|v_k| \leq \frac{\|v\|_\infty}{|a_{rr}|} \sum_{k\neq r} |a_{rk}| < \|v\|_\infty\,,$$

$$\begin{aligned}
\left|v_{l_{L-1}}\right| &\overset{*}{\leq} \frac{1}{\left|a_{l_{L-1},l_{L-1}}\right|} \left(\sum_{k\neq l_{L-1}; k\neq l_L} \left|a_{l_{L-1},k}\right| |v_k| + \left|a_{l_{L-1},l_L}\right| |v_{l_L}| \right) \\
&< \frac{\|v\|_\infty}{\left|a_{l_{L-1},l_{L-1}}\right|} \left(\sum_{k\neq l_{L-1}} \left|a_{l_{L-1},k}\right| \right) \leq \|v\|_\infty\,, \\
&\vdots \\
\|v\|_\infty = |v_s| &= |v_{l_0}| < \|v\|_\infty\,.
\end{aligned}$$

Dies ist ein Widerspruch und somit muss $\rho(T) < 1$ sein. Mit Folgerung 2.28 folgt dann die Behauptung.

Als nächstes Iterationsverfahren wählen wir M als die untere Dreiecksmatrix von A. Wir erhalten das Einzelschritt Verfahren.

2.3.2 Einzelschritt Verfahren (ESV) / Gauß-Seidel-Verfahren

Sei A regulär mit $a_{ii} \neq 0$. Wir zerlegen A additiv in $A = L + D + R$ und setzen

$$M = L + D = \begin{pmatrix} a_{11} & & 0 \\ \vdots & \ddots & \\ a_{1n} & \cdots & a_{nn} \end{pmatrix}.$$

Da $a_{ii} \neq 0$ ist, ist M regulär und $N = M - A = -R$.
Dies führt zu folgender Iterationsvorschrift:

Sei ein Startvektor $x^0 \in \mathbb{R}^n$ gegeben.
Iteration (ESV):

$$\begin{aligned} x^{k+1} &= (L+D)^{-1}\left(b - Rx^{(k)}\right) \\ \Longrightarrow \quad x_i^{k+1} &= \frac{1}{a_{ii}}\left(b_i - \sum_{l=1}^{i-1} a_{il}x_l^{(k+1)} - \sum_{l=i+1}^{n} a_{il}x_l^{(k)}\right),\ i = 1,\dots,n. \end{aligned}$$

Vergleiche mit (GSV):

$$x_i^{k+1} = \frac{1}{a_{ii}}\left(b_i - \sum_{l=1}^{i-1} a_{il}x_l^{(k)} - \sum_{l=i+1}^{n} a_{il}x_l^{(k)}\right),\ i = 1,\dots,n.$$

Satz 2.36
Die Matrix A erfülle das starkes Zeilensummenkriterium. Dann konvergiert das Einzelschrittverfahren.

Beweis: Setze $T := M^{-1}N$ und $q = \max_i \left(\sum_{k\neq i} \frac{|a_{ik}|}{|a_{ii}|}\right) < 1$.

Wir zeigen: $\|T\|_\infty \leq q < 1$ und folglich ist T Kontraktion.

Es ist $\|T\|_\infty = \sup_{\|x\|_\infty = 1} \|Tx\|_\infty$. Sei also $x \in \mathbb{R}^n$ mit $\|x\|_\infty = 1$. Setze $y := Tx$. Zu zeigen $\|y\|_\infty \leq q$, d.h. $|y_i| \leq q$ $(1 \leq i \leq n)$.

Induktion:
I.A. $y_1 = -\frac{1}{a_{11}}\left(\sum_{k=2}^{n} a_{1k}x_k\right)$

$$\Longrightarrow \quad |y_1| \leq \frac{1}{|a_{11}|}\sum_{k=2}^{n} |a_{1k}| \underbrace{|x_k|}_{\leq 1} \leq \frac{1}{|a_{11}|}\sum_{k\neq 1} |a_{1k}| \leq q.$$

I.S. $y_i = -\frac{1}{a_{ii}}\left(\sum_{k=i+1}^{n} a_{ik}x_k - \sum_{k=1}^{i-1} a_{ik}y_k\right)$

$$\Longrightarrow \quad |y_i| \leq \frac{1}{|a_{ii}|}\left(\sum_{k=i+1}^{n} |a_{ik}| \underbrace{|x_k|}_{\leq 1} + \sum_{k=1}^{i-1} |a_{ik}| \underbrace{|y_k|}_{\leq q}\right) \leq q.$$

Wie für das Gesamtschrittverfahren erhalten wir auch hier folgenden Konvergenzsatz. □

Satz 2.37 (Ohne Beweis)
Sei $A \in \mathbb{R}^{n\times n}$ unzerlegbar und erfülle das schwache Zeilensummenkriterium, dann konvergiert das Gauß-Seidel-Verfahren.

Darüberhinaus können wir für das Gauß-Seidel-Verfahren auch folgenden Satz zeigen, der für das Jacobi-Verfahren nicht gilt.

Satz 2.38
A sei symmetrisch und positiv definit, dann konvergiert das Gauß-Seidel-Verfahren.

Beweis: Sei $A = L + D + R$. Da A symmetrisch ist gilt: $R = L^\top$ bzw. $R^\top = L$, $M = L + D$, $N = -R$, $T = M^{-1}N = -(L + D)^{-1}R$.

Sei $\lambda \neq 0$ ein Eigenwert von T in $\mathbb{C}$. Sei $x \in \mathbb{C}^n$ der zugehörige Eigenvektor mit $\|x\|_2 = 1$. Dann gilt: $-(L + D)^{-1}Rx = \lambda x$, bzw. $-Rx = \lambda Dx + \lambda Lx = \lambda Dx + \lambda R^\top x$. Es folgt

$$D = A - L - R = A - R^\top - R \implies -Rx = \lambda(A - R^\top - R)x + \lambda R^\top x = \lambda Ax - \lambda Rx.$$

Setze $\alpha := \langle Ax, x\rangle > 0$, da A positiv definit, $\sigma := \langle Rx, x\rangle = \sigma_1 + i\sigma_2 \in \mathbb{C}$, $\delta := \langle Dx, x\rangle$. Dann folgt

$$-\sigma = \lambda\alpha - \lambda\sigma = \lambda(\alpha - \sigma)$$

Hieraus folgt $\alpha - \sigma \neq 0$, da sonst $-\sigma = 0$ und somit $\alpha = \sigma = 0$ wäre. Dies ist ein Widerspruch zur positiven Definitheit. Also folgt

$\lambda = -\frac{\sigma}{\alpha-\sigma} \implies |\lambda|^2 = \frac{\sigma\overline{\sigma}}{(\alpha-\sigma)\overline{(\alpha-\sigma)}} = \frac{\sigma_1^2+\sigma_2^2}{(\alpha-\sigma_1)^2+\sigma_2^2}$, da $\alpha \in \mathbb{R}$ und $\alpha - \sigma = \alpha - \sigma_1 - i\sigma_2$. Weiter haben wir

$$\begin{aligned} \alpha = \langle Ax, x\rangle &= \left\langle R^\top x, x\right\rangle + \langle Dx, x\rangle + \langle Rx, x\rangle \\ &= \delta + 2\sigma_1 \implies \delta = \alpha - 2\sigma_1, \end{aligned}$$

$$\begin{aligned} (\alpha - \sigma_1)^2 &= (\delta + \sigma_1)^2 = \delta^2 + 2\delta\sigma_1 + \sigma_1^2 = \delta(\alpha - 2\sigma_1) + 2\delta\sigma_1 + \sigma_1^2 \\ &= \delta\alpha + \sigma_1^2 \geq \mu\delta + \sigma_1^2, \end{aligned}$$

wobei $\mu > 0$ der kleinste Eigenwert von A ist.

$$\delta = \langle Dx, x\rangle = \sum_{i=1}^{n} a_{ii}x_i\overline{x}_i \geq \min_i a_{ii} \|x\|_2^2 = \min_i a_{ii} =: \underline{\delta}$$

$$\implies |\lambda|^2 = \frac{\sigma_1^2 + \sigma_2^2}{(\alpha - \sigma_1)^2 + \sigma_2^2} \leq \frac{\sigma_1^2 + \sigma_2^2}{\mu\underline{\delta} + \sigma_1^2 + \sigma_2^2} < 1.$$

da $\mu\underline{\delta} > 0$. Also folgt insgesamt $|\lambda| < 1$ und somit $\rho(T) < 1$.

□

2.4 Gradientenverfahren

Generalvereinbarung 2.39
In diesem Abschnitt gelte stets

1) $A \in \mathbb{R}^{m\times m}$ mit $A = A^T$,

2) A sei positiv definit und

3) $b \in \mathbb{R}^m$.

Ziel: Löse das lineare Gleichungssystem $Ax = b$.

Dazu wollen wir das Problem zunächst in ein Minimierungsproblem überführen.

Definition 2.40 (Minimierungsaufgabe)
Sei $F(x) := \langle A^{-1}(Ax - b), Ax - b\rangle\,,\ x \in \mathbb{R}^m$.
x heißt Lösung der Minimierungsaufgabe (M), g.d.w.

$$(M) \qquad F(x) = \min_{y\in\mathbb{R}^m} F(y).$$

Lemma 2.41
A, b seineen mit den Eigenschaften der Generalvereinbarung 2.39 gegeben. Dann sind äquivalent:
1) $x \in \mathbb{R}^m$ löst $Ax = b$,

2) $x \in \mathbb{R}^m$ löst (M).

Beweis: A positiv definit $\implies$ A^{-1} positiv definit. Also gilt $F(y) \geq 0\ \ \forall y \in \mathbb{R}^m$.

1) $\Rightarrow$ 2): Gilt $Ax = b$, so folgt $F(x) = \left\langle A^{-1}0, 0\right\rangle = 0$.
Folglich löst x die Minimierungsaufgabe (M).

2) $\Rightarrow$ 1): x löst (M) $\implies$ $\nabla F(x) = 0$ $\iff$ $2(Ax - b) = 0$ $\iff$ $Ax - b = 0$ $\implies$ 1).

□

Idee der Gradientenverfahren 2.42
Aus Lemma 2.41 folgt: $Ax = b \iff F(x) = 0.$

Idee: Sei $(z_n)_{n\in\mathbb{N}}$ eine Folge im $\mathbb{R}^m$ definiert durch

$$z_{n+1} := z_n + \alpha_n t_n \quad n = 1, 2, \ldots$$

mit Koeffizienten $\alpha_n \in \mathbb{R}$ und Richtungsvektoren $t_n \in \mathbb{R}^m \setminus \{0\}$.
Wähle α_n, t_n so, dass $F(z_n) \longrightarrow 0\ \ (n \to \infty)$.

Ansatz für die Wahl von α_n:

$$\alpha_n := \beta_n \frac{\langle t_n, r_n\rangle}{\langle At_n, t_n\rangle}$$

mit $\beta_n \in \mathbb{R}$ und $r_n := b - Az_n$ der "Residuenvektor".

Dann gilt:

$$(*)\quad \left|\begin{aligned} F(z_n) - F(z_{n+1}) &= \left\langle A^{-1}r_n, r_n\right\rangle - \left\langle A^{-1}(r_n - A\alpha_n t_n), r_n - A\alpha_n t_n\right\rangle \\ &= 2\alpha_n \left\langle t_n, r_n\right\rangle - \alpha_n^2 \left\langle At_n, t_n\right\rangle \\ &= 2\beta_n \tfrac{\langle t_n, r_n\rangle^2}{\langle At_n, t_n\rangle} - \beta_n^2 \tfrac{\langle t_n, r_n\rangle^2}{\langle At_n, t_n\rangle} \\ &= (2-\beta_n)\beta_n \underbrace{\frac{\langle t_n, r_n\rangle^2}{\langle At_n, t_n\rangle}}_{\geq 0} \overset{0\leq\beta_n\leq 2}{\geq} 0. \end{aligned}\right.$$

$\Longrightarrow$ Für $0 \leq \beta_n \leq 2$ ist $F(z_n) \leq F(z_{n+1})$.
$\Longrightarrow$ $(F(z_n))_{n\in\mathbb{N}}$ ist monoton fallend.

Also folgt: $\lim\limits_{n\ to\infty} F(z_n) = \lambda \geq 0$ existiert.
$\Longrightarrow$ $(F(z_n) - F(z_{n-1}))_{n\in\mathbb{N}}$ ist Nullfolge.

Man nennt β_n den Relaxationsparameter der Gradientenverfahren.
Die Gleichung (*) zeigt:
Für $\beta_n = 1$ wird $F(z_n) - F(z_{n+1})$ maximal und F nimmt auf der Geraden $z_n + \alpha_n t_n$ ein Minimum an.

Definition 2.43 (Allgemeines Gradientenverfahren)
Seien $(\beta_n)_{n\in\mathbb{N}}$ und $(t_n)_{n\in\mathbb{N}}$ gegeben mit $\beta_n \in [0,2]$ und $t_n \in \mathbb{R}^m$. Dann heißt die Folge $(z_n)_{n\in\mathbb{N}}$, $z_n \in \mathbb{R}^m$ Lösung des Gradientenverfahren mit Startwert $z_1 \in \mathbb{R}^m$ wenn gilt:

$$r_1 = b - Az_1$$

und für $n = 1, 2, \ldots$ gilt

$$\alpha_n = \beta_n \frac{\langle t_n, r_n\rangle}{\langle At_n, t_n\rangle},$$
$$z_{n+1} = z_n + \alpha_n t_n,$$
$$r_{n+1} = b - Az_{n+1} = r_n - \alpha_n A t_n.$$

Definition 2.44 (Konvergenz)
Ein Gradientenverfahren heißt konvergent, falls gilt

$$r_n \longrightarrow 0 \quad (n \to \infty)$$

Dies ist äquivalent zu

$$\begin{aligned} & z_n \longrightarrow A^{-1}b \quad (n\to\infty) \\ \Longleftrightarrow\quad & z_n \longrightarrow x \quad\ \ (n\to\infty) \text{ mit } Ax = b. \end{aligned}$$

2.4.1 Eigentliches Gradientenverfahren

Definition 2.45 (Eigentliches Gradientenverfahren)
Das Gradientenverfahren 2.43 mit $t_n = r_n$ und $\beta_n = 1$ heißt eigentliches Gradientenverfahren. *Die Richtungsvektoren werden in Richtung des Gradienten von F gewählt:*

$$r_n = b - Az_n = -\frac{1}{2}\nabla F(z_n).$$

Satz 2.46 (Konvergenz)
Das eigentliche Gradientenverfahren 2.45 ist konvergent.

Beweis: Es gilt $\forall x \in \mathbb{R}^m : \langle Ax, x\rangle \overset{\text{Schwarzsche Ungleichung}}{\leq} \left(\sum_{i,j=1}^{m} a_{ij}^2\right)^{\frac{1}{2}} \langle x, x\rangle .$

Setze $k := \left(\sum_{i,j=1}^{m} a_{ij}^2\right)^{\frac{1}{2}}$. Dann gilt $\forall x \in \mathbb{R}^m \setminus \{0\}$:

$$\frac{\langle Ax, x\rangle}{\langle x, x\rangle} \leq k.$$

Mit $(*)$ aus 2.42 folgt:

$$F(z_n) - F(z_{n+1}) = \frac{\langle r_n, r_n\rangle^2}{\langle Ar_n, r_n\rangle} \geq \frac{1}{k}\langle r_n, r_n\rangle .$$

Aus 2.42 wissen wir, dass $F(z_n) - F(z_{n+1}) \to 0$ für $(n \to \infty)$. Also folgt für $k \neq 0$: $r_n \to 0 \ \ (n \to \infty)$.

□

Bemerkung 2.47
Je zwei aufeinanderfolgende Residuenvektoren des eigentlichen Gradientenverfahrens stehen senkrecht aufeinander:

$$\begin{aligned}\langle r_n, r_{n+1}\rangle &= \langle r_n, r_n - \alpha_n A t_n\rangle \\ &= \left\langle r_n, r_n - \tfrac{\langle r_n, r_n\rangle}{\langle Ar_n, r_n\rangle} Ar_n \right\rangle \\ &= \langle r_n, r_n\rangle - \langle r_n, r_n\rangle = 0.\end{aligned}$$

Definition 2.48 (Gradientenverfahren bezüglich der kanonischen ON-Basis des $\mathbb{R}^m$)
Wähle $\beta_n = 1$ und $t_{i+jm} = e_i$ für $i = 1, \ldots, m;\ j = 0, 1, 2, \ldots$, wobei $e_i \in \mathbb{R}^m$ der i-te Einheitsvektor ist. Dann folgt mit $n = i + jm$:

$$\begin{aligned}\alpha_n = \tfrac{\langle e_i, r_n\rangle}{\langle Ae_i, e_i\rangle} &= \tfrac{1}{a_{ii}}\langle e_i, b - Az_n\rangle \\ &= \tfrac{1}{a_{ii}}(b_i - \sum_{l=1}^{m} a_{il} z_{n,l})\end{aligned}$$

$$\Longrightarrow \quad z_{n+1} = z_n + \frac{1}{a_{ii}}(b_i - \sum_{l=1}^{m} a_{il} z_{n,l}) e_i.$$

D.h. z_{n+1} und z_n unterscheiden sich nur in der i-ten Komponente.

Das ist das Einschrittverfahren (siehe Numerik I).

Satz 2.49 (Konvergenz)
Das Gradientenverfahren 2.48 ist konvergent.

Beweis: Setze $\gamma := \max_{i=1,\dots,m} a_{ii}$. Dann ist mit $(*)$ aus 2.4

$$F(z_n) - F(z_{n+1}) \overset{n=i+jm}{=} \frac{\langle e_i, r_n\rangle^2}{\langle Ae_i, e_i\rangle} \geq \frac{1}{\gamma}(r_{n,i})^2 \geq 0.$$

Da nach 2.4 $F(z_n) - F(z_{n+1}) \to 0$, folgt $r_{n,i} \longrightarrow 0\ \forall i = 1, \dots, m$ und somit $r_n \to 0$ $(n \to \infty)$. □

2.4.2 *Conjugate Direction* Verfahren (CD)

Definition 2.50 (A-orthogonal)
Ein System von k Vektoren $q_1, \dots, q_k \in \mathbb{R}^m$ $(k \leq m)$ heißt A-orthogonal *oder* A-konjugiert, *wenn gilt:*

$$\langle Aq_i, q_j\rangle = 0 \quad i \neq j;\ i, j = 1, \dots, k,$$
$$\langle Aq_i, q_i\rangle \neq 0 \quad \forall i = 1, \dots k.$$

Für $k = m$ bilden A-orthogonale Systeme eine Basis des $\mathbb{R}^m$. Wir setzen $\langle x, y\rangle_A = \langle Ax, y\rangle$, $\|x\|_A = \sqrt{\langle x, x\rangle_A} = \sqrt{\langle Ax, x\rangle}$.

Definition 2.51 (cd-Verfahren) "conjugate direction"
Sei $\{q_i\}_{i=1}^m$ A-orthogonal. Wähle im allgemeinen Gradientenverfahren 2.43:

$$t_n = q_n \text{ und } \beta_n = 1 \text{ für } n = 1, \dots, m.$$

Satz 2.52 (Konvergenz der cd-Methode)
Das cd-Verfahren ist für beliebige Startvektoren $z_1 \in \mathbb{R}^m$ ein endliches Verfahren. Es gilt:

$$r_{m+1} = 0$$

und somit

$$z_{m+1} = A^{-1}b = x.$$

Beweis: Laut Gradientenverfahren folgt induktiv für $1 \leq n \leq m+1$:

$$r_n = r_{n-1} - \alpha_{n-1}At_{n-1} = r_{n-2} - \alpha_{n-2}At_{n-2} - \alpha_{n-1}At_{n-1} = \dots = r_{n-l} - \sum_{i=n-l}^{n-1} \alpha_i At_i.$$

Für $l = n - j$, $1 \leq j \leq n$ folgt

$$r_n = r_j - \sum_{i=j}^{n-1} \alpha_i Aq_i.$$

$$\Longrightarrow \quad \langle q_j, r_n \rangle \overset{\text{A-orthogonal}}{=} \langle q_j, r_j \rangle - \alpha_j \langle Aq_j, q_j \rangle$$
$$\overset{\text{Def. von } a_j}{=} \langle q_j, r_j \rangle - \langle q_j, r_j \rangle = 0.$$

$\Longrightarrow \quad \langle q_j, r_{m+1} \rangle = 0 \ \ \forall j = 1, \ldots, m.$
$\Longrightarrow \quad r_{m+1} = 0$, da $\{q_i\}_{i=1}^m$ Basis des $\mathbb{R}^m$.

□

2.4.3 *Conjugate Gradient* Verfahren (CG)

Definition 2.53 (cg-Verfahren) "conjugate gradient"
Wähle $\beta_n = 1 \ \ \forall n = 1, \dots, m+1, \ t_1 = r_1$ *und*

$$t_n = r_n + \gamma_{n-1} t_{n-1}$$

mit $\gamma_{n-1} := -\frac{\langle Ar_n, t_{n-1}\rangle}{\langle At_{n-1}, t_{n-1}\rangle}$ *für* $n = 2, 3, \dots$. *Idee: Das A-orthogonale System wird mit Hilfe des Schmitd'schen Orthogonalisierungsverfahrens bzgl.* $\langle \cdot, \cdot \rangle_A$ *schrittweise aufgebaut.*

Satz 2.54 (Konvergenz des cg-Verfahrens)
Sei $z_1 \in \mathbb{R}^m$ und $t_1 = r_1 = b - Az_1$. Dann lautet das cg-Verfahrenfür $n = 1, \dots, l$ mit $l \leq m$:

$$\left|\begin{array}{l} a_n = \frac{\langle t_n, r_n\rangle}{\langle At_n, t_n\rangle} \\ z_{n+1} = z_n + \alpha_n t_n \\ r_{n+1} = b - Az_{n+1} \\ \gamma_n = -\frac{\langle Ar_{n+1}, t_n\rangle}{\langle At_n, t_n\rangle} \\ t_{n+1} = r_{n+1} + \gamma_n t_n \end{array}\right.$$

Es erfüllt die Gleichungen
a) $\langle At_i, t_j\rangle = 0 \ \ 1 \leq j \leq i-1$

b) $\langle r_i, r_j\rangle = 0 \ \ 1 \leq j \leq i-1$

c) $\langle t_i, r_j\rangle = \langle r_j, r_j\rangle \ \ \ 1 \leq j \leq i$

und $l \leq m$ ist so gewählt, dass gilt $r_{l+1} = 0$.

Beweis: Folgt aus linearer Algebra und Satz 2.52. □

Bemerkung 2.55
Für Gleichungssysteme resultierend aus Diskretisierungsverfahren, wie z.B. der Finit Elemente Methode, ist m so groß, dass man in der Praxis weniger als m Schritte iterieren wird. Dass dies Sinn mach zeigt die folgende Fehlerabschätzung.

Satz 2.56
Sei $\kappa(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$ die Kondition von $A \in \mathbb{R}^{n\times n}$. Dann gilt für das cg-Verfahren für $Ax = b$:

$$\|z_n - x\|_A \leq 2\left(\frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1}\right)^n \|z_1 - x\|_A$$

wobei $\|y\|_A := \sqrt{\langle Ay, y\rangle}$ definiert ist.
(ohne Beweis)

Bemerkung 2.57

1) Vorkonditionierung: Anstelle von $Ax = b$ löst man $C^{-1}A(C^{-1})^T y = C^{-1}b$, wobei $y = C^T x$ ist und C auch folgende Anforderungen erfüllt

   1) $C \in \mathbb{R}^{n\times n}$ regulär.
   2) Gleichungsysteme mit C sollen einfach zu lösen sein.
   3) $\kappa(C^{-1}A(C^{-1})^T$ sollt möglich nahe an 1 liegen.

2) Ist A nicht symmetrisch und positiv definit, so kann man z.B. $A^TAx = A^Tb$ anstelle von $Ax = b$ lösen, denn A^TA ist positiv definit und symmetrisch.
   Dieser Ansatz führt auf das bicg-Verfahren.

3) Für eine gegbene Toleranz TOL, kann

$$\langle r_n, r_n\rangle \leq TOL$$

   als Abbruchbedingung verwendet werden.

2.5 Zusammenfassung

Wir haben in diesem Kapitel verschiedene Methoden zur Lösung linearer Gleichungssysteme der Form

$$Ax = b$$

kennengelernt und näher betrachtet,

(a) für $A \in \mathbb{R}^{n\times n}$ regulär

(b) für $A \in \mathbb{R}^{m\times n}$ mit $m \geq n$ (d.h. überstimmt)

Verfahren:

(a) Direkte Verfahren oder Direktlöser (LR-, Cholesky, QR-Zerlegung)

(b) Iterative Verfahren oder Iterativenlöser (Jacobi-, Gauß-Seidel-Verfahren)

Vor-/Nachteile

- **Direkte Verfahren:** Die Lösung wird bis auf Rundungsfehler exakt berechnet. Sie sind für große Gleichungssysteme sehr langsam (die Anzahl der arithmetischen Operationen liegen in $O(n^3)$); es tritt ein *fill-in* Problem auf: In Anwendungen ist A häufig dünn besetzt, d.h. pro Zeile sind nur k viele Einträge ungleich Null, wobei k unabhängig von n, und diese Einträge sind unstrukturiert verteilt. Eine typische Zeile sieht dann so aus:

$$\left(\begin{array}{ccccccc} * & & & & & & 0 \\ 0 & \ddots & & & & & \\ \hline * & 0\cdots 0 & * & * & 0\cdots 0 & * & \\ \hline & & & & \ddots & & \\ & & & & & * & \end{array}\right)$$

  Bei Zerlegungsverfahren werden Nulleinträge u.a. durch Einträge ungleich Null ersetzt, der Speicheraufwand zum Speichern von A liegt in der Regel bei $O(NK) = O(N)$; nach der Zerlegung wächst der Speicheraufwand bis auf $O(N^2)$.

- **Iterative Verfahren:** Die Lösung wird nur näherungsweise bestimmt und die Geschwindigkeit der Konvergenz hängt von $\rho(T)$ (Spektralradius) ab.

  Allerdings ist der zusätzliche Speicheraufwand sehr klein (das Gauß-Seidel-Verfahren hat gar kein zusätzlicher Speicheraufwand). Iterative Verfahren benötigen ein gutes Abbruchkriterium, z.B. $\left\|Ax^{(k)} - b\right\| < TOL$.

Beschleunigung/Stabilisierung

Das Hauptproblem: Einfluss durch Rundungsfehler und ihre Reduzierung.

Beispiel: Die Pivotisierung bei der LR-Zerlegung. Die Kondition des Problems ist proportional zur Kondition von A: $cond(A) = \|A\| \left\|A^{-1}\right\|$. Durch **Vorkonditionierung** kann versucht werden, die Kondition des Problems zu verkleinern.

Beispiel: Wähle C_1, C_2 reguläre Matrizen. Dann gilt:

$$Ax = b \iff \underbrace{C_1 A C_2}_{:=\tilde{A}} \underbrace{C_2^{-1} x}_{:=\tilde{x}} = \underbrace{C_1 b}_{:=\tilde{b}}$$

mit $\tilde{A} := C_1 A C_2,\ \tilde{x} := C_2^{-1} x,\ \tilde{b} := C_1 b$ und C_1, C_2 so gewählt, dass $cond(\tilde{A}) < cond(A)$ gilt.

Beschleunigung durch Relaxation:

Ein Iterationsverfahren hat die Gestalt

$$r^k = b - Ax^k,\ My^k = r^k,\ x^{k+1} = x^k + y^k$$

Statt y^k zu korrigieren, wählt man einen Relaxationsparameter $w > 0$ und setzt

$$x^{k+1} = x^k + wy^k.$$

Dies führt auf das **SOR-Verfahren** [2].

[2] **S**uccesive **O**ver **R**elaxation.

Kapitel 3

Nichtlineare Gleichungen/ Nullstellensuche

In diesem Kapitel wollen wir uns mit der Berechnung von Nullstellen einer gegebenen Funktion

$$f : \mathbb{R}^n \longrightarrow \mathbb{R}^n$$

beschäftigen. Da man die Lösung linearer und nichtlinearer Gleichungen auch als Nullstellensuche einer geeigneten Funktion f auffassen kann, ist diese Fragestellung eine direkte Verallgemeinerung der Lösung linearer Gleichungsprobleme, die in Kapitel 2 behandelt wurde. Bei der Bestimmung von Nullstellen unterscheiden wir zwei Problemstellungen.

Problem A: Gesucht ist ein $x^* \in \mathbb{R}^n$ mit $f(x^*) = 0$.
Problem B: Gesucht sind alle (größtes, kleinstes) $x^* \in \mathbb{R}^n$ mit $f(x^*) = 0$.

Beispiel:
$f(x) = Ax - b,\ A \in \mathbb{R}^{m\times n}$ (siehe Kap. 2).
$f(x) = ax^2 + bx + c$. Hier sind alle Nullstellen explizit berechenbar.
$f(x) = \cos(x)$. Gesucht $x^* \in [1, 2] \implies x^* = \frac{\pi}{2}$.

Wir beschäftigen uns im wesentlichen mit Verfahren zur Lösung vom Problem **A** für $n = 1$ und

$$f : [a, b] \longrightarrow \mathbb{R}$$

glatt.

Wir nehmen an, dass es ein $x^* \in [a, b]$ gibt mit $f(x^*) = 0$, etwa $f \in C^0(a, b)$ und $f(a)f(b) < 0$. Alle Verfahren konstruieren eine Folge $\left(x^{(k)}\right)_{k\in\mathbb{N}}$ mit $x^{(k)} \longrightarrow x^*,\ f(x^*) = 0$. Direkte Verfahren existieren nur in Spezialfällen.

3.1 Verfahren in einer Raumdimension

3.1.1 Intervallschachtelungsverfahren (ISV)

Voraussetzungen: Sei $f \in C^0(a, b)$ mit $a < b$ und $f(a)f(b) < 0$.

Verfahren: Setze $a_0 := a, b_0 := b$ und $x^{(0)} := \frac{1}{2}(a_0 + b_0)$.

Für $n = 0, \ldots, N$:

Falls $f(a_n)f(x^{(n)}) = 0$, dann Abbruch.

Falls $f(a_n)f(x^{(n)}) < 0 \implies a_{n+1} := a_n;\ b_{n+1} := x^{(n)}$

Falls $f(a_n)f(x^{(n)}) > 0 \implies a_{n+1} := x^{(n)};\ b_{n+1} := b_n$

Setze $x^{(n+1)} := \frac{1}{2}(a_{n+1} + b_{n+1})$.

Satz 3.1 (Konvergenz von ISV)
Seien $(a_n)_{n\in\mathbb{N}}$, $(b_n)_{n\in\mathbb{N}}$, $\left(x^{(n)}\right)_{n\in\mathbb{N}}$ durch das Intervallschachtelungsverfahren definiert. Dann gilt

$$\lim_{n\longrightarrow\infty} a_n = \lim_{n\longrightarrow\infty} b_n = \lim_{n\longrightarrow\infty} x^{(n)} = x^* \text{ und } f(x^*) = 0.$$

Es gilt die Fehlerabschätzung:

$$\left|x^{(n)} - x^*\right| \leq 2^{-(n+1)} \left|b - a\right|.$$

Beweis: Es gilt $(a_n)_{n\in\mathbb{N}}$ monoton wachsend und $(b_n)_{n\in\mathbb{N}}$ monoton fallend und $0 < b_n - a_n = 2^{-n}(b-a)$, $a_n < b, a < b_n$. Daraus folgt $(a_n)_{n\in\mathbb{N}}$, $(b_n)_{n\in\mathbb{N}}$ konvergieren und $\lim_{n\longrightarrow\infty} a_n = \lim_{n\longrightarrow\infty} b_n =: x^*$ Da $f \in C^0$ folgt weiter $f(x^*) = \lim_{n\longrightarrow\infty} f(a_n) = \lim_{n\longrightarrow\infty} f(b_n)$.

Nach Konstruktion gilt stets $f(a_n)f(b_n) < 0$, also folgt

$$f(x^*)^2 = \lim_{n\longrightarrow\infty} (f(a_n)f(b_n)) \leq 0 \implies f(x^*) = 0.$$

Da $x^* \in \left(x^{(n)}, b_n\right)$ oder $x^* \in \left(a_n, x^{(n)}\right)$ folgt auch

$$\left|x^{(n)} - x^*\right| \leq \min\left\{\left|x^{(n)} - b_n\right|, \left|x^{(n)} - a_n\right|\right\} = \frac{1}{2}\left|b_n - a_n\right| \leq 2^{-n-1}(b-a).$$

Bemerkung: Das Verfahren ist sehr robust aber auch sehr langsam: Man benötigt etwa 3 Schritte, um eine Dezimalzahlstelle Genauigkeit zu gewinnen.

Aufwand: Eine Auswertung von f pro Schritt.

Das Verfahren wird in der Regel eingesetzt, um grob ein Intervall $[a, b]$ zu bestimmen, in dem eine Nullstelle liegt. Zur genaueren Berechnung der Nullstellen werden dann effizientere Verfahren eingesetzt.

3.1.2 Newton Verfahren

Idee: Sei $x^{(k)}$ eine gegebene Approximation von x^*, d.h. $h := x^* - x^{(k)} \neq 0$ ist klein. Dann gilt mit der Taylorentwicklung:

$$0 = f(x^*) = f(x^{(k)} + h) = f(x^{(k)}) + f'(x^{(k)})h + \mathcal{O}(h^2).$$

Unter Vernachlässigung der Terme höherer Ordnung folgt

$$h = -\frac{f(x^{(k)})}{f'(x^{(k)})}$$

Daher sollte $x^{(k+1)} = x^{(k)} + h = x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})}$ eine bessere Approximierung von x^* sein als $x^{(k)}$.

Verfahren: Sie Startwert $x^{(0)}$ gegeben. Setze iterativ:

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

Aufwand: Eine Auswertung von f und f' pro Schritt, d.h. es muß gelten $f \in C^1$ und f' muss bekannt sein.

Geometrische Interpretation

Sei $l(x)$ die Linearisierung von f an der Stelle $x^{(k)}$, d.h. $l(x^{(k)}) = f(x^{(k)})$, $l'(x^{(k)}) = f'(x^{(k)})$ und $l(x) = ax + b$. Dann folgt aus der Taylorentwicklung $l(x) = f(x^{(k)}) + f'(x^{(k)})(x - x^{(k)})$.

Statt Nullstellen von f zu suchen, definieren wir $x^{(n+1)}$ Nullstelle von l

$$0 = f(x^{(k)}) + f'(x^{(k)})(x^{(k+1)} - x^{(k)}).$$

Dies ist gerade das Newton Verfahren.

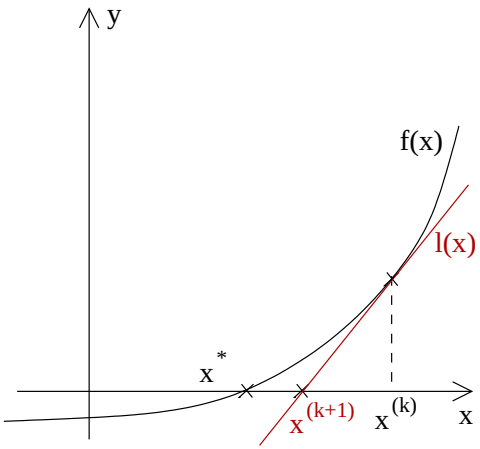

Abbildung 3.1: Newton Verfahren, Beispiel 1

Abbildung 3.1 veranschaulischt das Newton Verfahren. Mit dieser Anschauung sehen wir direkt ein, dass das Newton Verfahren nicht für alle Startwerte $x^{(0)}$ konvergieren wird. Abbildung 3.2 zeigt eine solche Situation. Hier gilt $\left|x^{(k)}\right| \longrightarrow \infty$. In Abbildung 3.3 gibt es zwei Nullstellen, aber es ist nicht klar, welche der beiden Nullstellen gefunden wird.

Satz 3.2 (Konvergenz des Newton-Verfahrens)
Sei $f \in C^2(a,b)$ und es es existiere ein $x^* \in (a,b)$ mit $f(x^*) = 0$. Sei $m := \min\limits_{a \le x \le b} |f'(x)| > 0$ und $M := \max\limits_{a \le x \le b} |f''(x)|$. Sei $\rho > 0$ so gewählt, dass $B_\rho(x^*) := \{x \mid |x - x^*| < \rho\} \subset [a,b]$ und $q := \frac{M}{2m}\rho < 1$. Dann konvergiert das Newton-Verfahren für jeden Startwert $x^{(0)} \in B_\rho(x^*)$. Es gilt die a-priori Fehlerschranke

(a) $$\left|x^{(k)} - x^*\right| \le \frac{M}{2m}\left|x^{(k-1)} - x^*\right| \le \frac{2m}{M} q^{2^k}$$

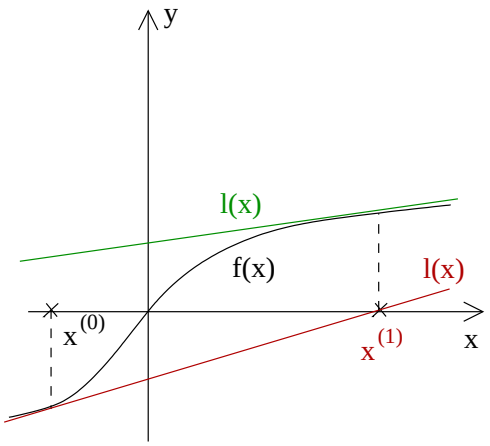

Abbildung 3.2: Newton Verfahren, Beispiel 2

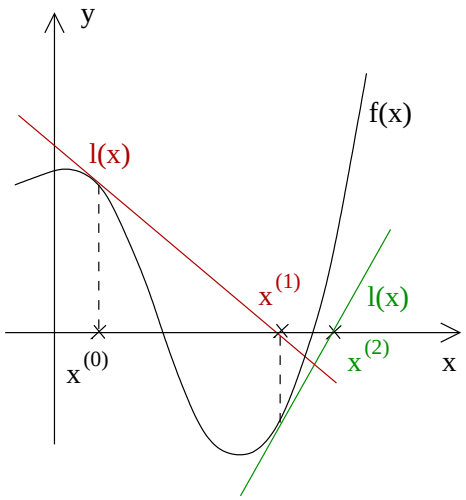

Abbildung 3.3: Newton Verfahren, Beispiel 3

und die a-posteriori Fehlerschranke

(b) $$\left|x^{(k)} - x^*\right| \leq \frac{1}{m}\left|f(x^{(k)})\right| \leq \frac{M}{2m}\left|x^{(k)} - x^{(k-1)}\right|^2$$

Bemerkung: Aus dem Mittelwertsatz folgt $\left|\frac{f(x)-f(y)}{x-y}\right| = |f'(\xi)| \geq m \;\forall\; x, y \in B_\rho(x^*);\; x \neq y \implies |x-y| \leq \frac{1}{m}|f(x)-f(y)|$. Folglich ist x^* ist die einzige Nullstelle in $B_\rho(x^*)$ und x^* ist **einfache** Nullstelle, d.h. $f(x^*) = 0$ und $f'(x^*) \neq 0$.

Beweis: Nach Taylorentwicklung (Satz 1.33) gilt

(1) $f(y) = f(x) + f'(x)(y-x) + R(y,x)$ mit

$$R(y,x) = \int_x^y f''(\xi)(y-\xi)\mathrm{d}\xi = (y-x)^2 \int_0^1 f''(x+s(y-x))(1-s)\mathrm{d}s.$$

Also folgt für alle $x, y \in B_\rho(x^*)$

$$(2)\quad |R(y,x)| \le |y-x|^2 M \int_0^1 (1-s)\mathrm{d}s = \tfrac{M}{2}|y-x|^2$$

Für $x \in B_\rho(x^*)$ setze $\Phi(x) := x - \frac{f(x)}{f'(x)}$. Dann folgt

$$\begin{aligned} |\Phi(x) - x^*| &= \left|(x-x^*) - \tfrac{f(x)}{f'(x)}\right| = \left|-\tfrac{1}{f'(x)}\left[f(x) + (x^*-x)f'(x)\right]\right| \\ &\overset{(1)}{=} \left|\tfrac{1}{f'(x)} R(x^*,x)\right| = \tfrac{1}{|f'(x)|}|R(x^*,x)| \overset{(2)}{\le} \tfrac{1}{m}|x-x^*|^2 \tfrac{M}{2} \end{aligned}$$

Also folgt für $x \in B_\rho(x^*)$

$$(3)\quad \begin{aligned} |\Phi(x) - x^*| &\le \tfrac{M}{2m}|x-x^*|^2, \\ &\le \tfrac{M}{2m}\rho^2 =: q\rho < \rho, \text{ da } q = \tfrac{M}{2m}\rho < 1. \end{aligned}$$

Insbesondere folgt $x^{(k)} \in B_\rho(x^*)$, falls $x^{(0)} \in B_\rho(x^*)$.

Sei $\rho^{(k)} := \frac{M}{2m}\left|x^{(k)} - x^*\right|$, so gilt wegen (3)

$$\begin{aligned} \rho^{(k)} &= \tfrac{M}{2m}\left|\Phi\left(x^{(k-1)}\right) - x^*\right| \le \tfrac{M}{2m}\left(\tfrac{M}{2m}\left|x^{(k-1)} - x^*\right|^2\right) \\ &= \left(\rho^{(k-1)}\right)^2 \le \ldots, \le \left(\rho^{(0)}\right)^{2^k} \\ \Longrightarrow \left|x^{(k)} - x^*\right| &\le \tfrac{2m}{M}\rho^{(k)} \le \tfrac{2m}{M}\left(\rho^{(0)}\right)^{2^k} = \tfrac{2m}{M}\left(\tfrac{M}{2m}\left|x^{(0)} - x^*\right|\right)^{2^k} \\ &\le \tfrac{2m}{M}\Big(\underbrace{\tfrac{M}{2m}\rho}_{=q}\Big)^{2^k} = \tfrac{2m}{M}q^{2^k}. \end{aligned}$$

Dies ist die a-priori Abschätzung. Da $q < 1$ ist, folgt $q^{(2^k)} \longrightarrow 0 \implies x^{(k)} \longrightarrow x^*$.
Für die a-posteriori Abschätzung benutzen wir nochmal (1) mit $y = x_{(k)}$ und $x = x^{(k-1)}$. Es folgt

$$\begin{aligned} f(x^{(k)}) &= f(x^{(k-1)}) + \left(x^{(k)} - x^{(k-1)}\right) f'(x^{(k-1)}) + R(x^{(k)}, x^{(k-1)}) \\ &= R(x^{(k)}, x^{(k-1)}), \text{ da } x^{(k)} = x^{(k-1)} - \tfrac{f(x^{(k-1)})}{f'(x^{(k+1)})} \end{aligned}$$

und somit

$$\begin{aligned} \left|x^{(k)} - x^*\right| &\overset{MWS}{\le} \tfrac{1}{m}\left|f(x^{(k)}) - f(x^*)\right| = \tfrac{1}{m}\left|f(x^{(k)})\right|, \\ &\overset{(2)}{=} \tfrac{1}{m}R(x^{(k)}, x^{(k-1)}) \le \tfrac{M}{2m}\left|x^{(k)} - x^{(k-1)}\right|^2. \end{aligned}$$

□

Bemerkung: Falls $x^{(0)} \in B_\rho(x^*)$, so konvergiert das Newton-Verfahren sehr schnell. Sei zum Beispiel $q = \frac{1}{2}$, so gilt nach 10 Iterationen $\left|x^{(10)} - x^*\right| \le \frac{2m}{M}q^{1024} \sim \frac{2m}{M}10^{-303}$.

Vergleichen wir dies mit dem Intervallschachtelungsverfahren, so folgt mit dem selben Startintervall: $|b-a| = \rho = q\frac{2m}{M} = \frac{m}{M}$, falls $q = \frac{1}{2}$ gilt. Nach 10 Schritten gilt also
$\left|x^{(10)} - x^*\right| = 2^{-11}|b-a| = 2^{-11}\frac{m}{M} \sim 10^{-4}\frac{2m}{M}$.

Folgerung 3.3
Für $f \in C^2(\mathbb{R})$ existiert für jede einfache Nullstelle x^* eine Umgebung U um den Wert x^*, so dass das Newton-Verfahren für alle $x^{(0)} \in U$ konvergiert und $\left|x^{(k)} - x^*\right| \leq q^{2^k}$ für $q < 1$.

Beweis: Übungsaufgabe. □

Offene Fragen:

(a) Wie kann ρ effektiv bestimmt werden?

(b) Kann die Berechnung von f' umgangen werden?

(c) Was passiert, falls $f'(x^*) = 0$ ist, d.h falls x^* mehrfache Nullstelle ist?

(d) Gibt es noch schnellere Verfahren?

Kombination von Newton Verfahren und Intervallschachtelung

Idee: ISV einsetzen, um ausreichend nahe an eine Nullstelle x^* zu kommen, so dass das Newton Verfahren schnell konvergiert.

Verfahren: Seien $a < b$ gegeben mit $f(a)f(b) < 0$.

Setze $x := \frac{1}{2}(a+b),\ \widetilde{a} := a,\ \widetilde{b} := b,\ f_0 := f(x),\ f_a := f(a)$.

`Solange` $|f_0| > TOL$:
$\quad$ `Falls` $f_a f_0 < 0$, `dann` $\widetilde{b} := x$, `sonst` $(\widetilde{a} := x;\ f_a := f_0)$
$\quad x := x - \frac{f_0}{f'(x)}$
$\quad f_1 := f(x)$
$\quad$ `Falls` $(|f_1| > |f_0|$ `oder` $x \notin (a,b))$, `dann`
$\quad\quad \left[a := \widetilde{a},\ b := \widetilde{b},\ x := \frac{1}{2}(a+b),\ f_1 := f(x)\right]$
$\quad f_0 := f_1$

3.1.3 Sekantenverfahren

Ein Nachteil des Newton-Verfahrens ist die Auswertung von $f'(x^{(k)})$.

Idee: Ersetze die Ableitung durch einen Differenzenquotient

$$f'(x^{(k)}) \sim \frac{f(x^{(k)}) - f(x^{(k-1)})}{x^{(k)} - x^{(k-1)}}.$$

Sekantenverfahren:

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

Geometrische Interpretation:

Die Sekante an f durch die Punkte $x^{(k)}, x^{(k-1)}$ ist gegeben durch

$$\frac{y - f(x^{(k)})}{x - x^{(k)}} = \frac{f(x^{(k-1)}) - f(x^{(k)})}{x^{(k-1)} - x^{(k)}},$$

wobei $y = y(x)$ die Gerade durch die Punkte $\left(x^{(k-1)}, f(x^{(k-1)})\right)$, $\left(x^{(k)}, f(x^{(k)})\right)$ ist, d.h.

$$y(x) = \frac{f(x^{(k-1)}) - f(x^{(k)})}{x^{(k-1)} - x^{(k)}} \left(x - x^{(k)}\right) + f(x^{(k)}).$$

$x^{(k+1)}$ wird also als Nullstelle der Sekante wählen.

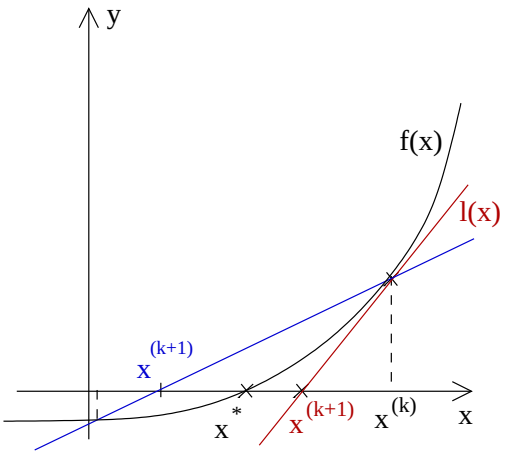

Abbildung 3.4: Sekantenverfahren, geometrische Interpretation

Bemerkung: Das Verfahren erfordert 2 Startwerte $x^{(0)}$, $x^{(1)}$ (siehe Abb. 3.4). Das Verfahren erfordert eine Auswertung von f pro Iterationschritt (speichern von $f(x^{(k-1)})$).

Satz 3.4 (Konvergenz des Sekantenverfahrens)
Sei $f \in C^2(a,b)$ mit $x^* \in [a,b]$, $f(x^*) = 0$ und $m := \min\limits_{a \le x \le b} |f'(x)| > 0$, $M := \max\limits_{a \le x \le b} |f''(x)|$. Sei $q := \frac{M}{2m}\rho < 1$ für $\rho > 0$. Seien $x^{(0)}$, $x^{(1)} \in B_\rho(x^*)$, $x^{(0)} \neq x^{(1)}$
und $\left(x^{(k)}\right)_{k\in\mathbb{N}}$ die Folge definiert durch das Sekantenverfahren.
Dann gilt $x^{(k)} \in B_\rho(x^*)$ und $x^{(k)} \longrightarrow x^*$ für $k \longrightarrow \infty$. Es gelten folgende Fehlerabschätzungen.

(a) A-priori Fehlerschranke:

$$\left|x^{(k)} - x^*\right| \le \frac{2m}{M} q^{\gamma_k},$$

wobei $(\gamma_k)_{k\in\mathbb{N}}$ die Folge der Fibonacci-Zahlen ist, d.h. $\gamma_0 = \gamma_1 = 1$; $\gamma_{k+1} = \gamma_k + \gamma_{k-1}$.

(b) A-posteriori Fehlerschranke:

$$\left|x^{(k)} - x^*\right| \le \frac{1}{m}\left|f(x^{(k)})\right| \le \frac{M}{2m}\left|x^{(k)} - x^{(k-1)}\right| \cdot \left|x^{(k)} - x^{(k-2)}\right|.$$

Beweis: Übungsaufgabe. □

Folgerung 3.5 (Konvergenz des Sekantenverfahrens)
Für $f \in C^2(\mathbb{R})$ existiert eine Umgebung U um jede einfache Nullstelle, so dass $x^{(k)} \longrightarrow x^*$ für $x^{(0)}$, $x^{(1)} \in U$ und es gilt $\left|x^{(k)} - x^*\right| \le \frac{2m}{M}\widetilde{q}^{\alpha^k}$ mit $\widetilde{q} < 1$ und $\alpha = \frac{1}{2}(1 + \sqrt{5}) \approx 1.618$.

Beweis: Zunächst zeigt man analog zu Satz 3.3, dass es ein $\rho > 0$ gibt, so dass die Voraussetzungen von Satz 3.4 auf $U = B_\rho(x^*)$ erfüllt sind. Mit der A-priori-Abschätzung von Satz 3.4 ist noch zu zeigen, dass gilt $q^{\gamma_k} \leq \widetilde{q}^{\alpha^k}$ mit $\widetilde{q} < 1$ geeignet gewählt.

Ansatz: $\gamma_k = \lambda^k$ mit $\lambda > 0$. Aus $\gamma_k - \gamma_{k-1} - \gamma_{k-2} = 0$ folgt

$$\lambda^{k-2}(\lambda^2 - \lambda - 1) = 0 \iff \lambda^2 - \lambda - 1 = 0 \implies \lambda_{1/2} = \frac{1}{2}(1 \pm \sqrt{5}).$$

D.h. die allgemeine Lösung der Differenzengleichung $\gamma_k - \gamma_{k-1} - \gamma_{k-2} = 0$ ist

$$\gamma_k = c_1\lambda_1^k + c_2\lambda_2^k$$

für $c_1, c_2 \in \mathbb{R}$.

Mit $\gamma_0 = \gamma_1 = 1$ folgt $c_1 = \frac{\lambda_1}{\sqrt{5}}$, $c_2 = -\frac{\lambda_2}{\sqrt{5}}$. Also $\gamma_k = \frac{1}{\sqrt{5}}\left(\lambda_1^{k+1} - \lambda_2^{k+1}\right)$. Aus $|\lambda_2| < |\lambda_1|$ folgt $\gamma_k \geq \frac{\lambda_1}{2\sqrt{5}}\lambda_1^k$ und somit

$$\begin{aligned} q^{\gamma_k} &\leq q^{\frac{\lambda_1}{2\sqrt{5}}(\lambda_1^k)}, \text{ da } q < 1 \\ &= \widetilde{q}^{(\alpha^k)} \end{aligned}$$

mit $\widetilde{q} := q^{\frac{\lambda_1}{2\sqrt{5}}} < 1$ und $\alpha = \lambda_1$. □

3.1.4 Zusammenfassung

Für die betrachteten Verfahren gilt:

ISV: $\left|x^{(k)} - x^*\right| \leq \frac{(b-a)}{2}\left(\frac{1}{2}\right)^k$, eine Auswertung von f pro Schritt.

Newton: $\left|x^{(k)} - x^*\right| \leq \frac{2m}{M}q^{2^k}$, je eine Auswertung von f und f' pro Schritt.

Sekanten: $\left|x^{(k)} - x^*\right| \leq \frac{2m}{M}q^{1.618^k}$, eine Auswertung von f pro Schritt.

Annahme: Die Auswertungen von f und f' seien gleich aufwändig. Dann hat das Newton-Verfahren pro Schritt den doppelten Aufwand.

Vergleich der Verfahren:
Definiere $z^{(k)} := x^{2k}$ beim ISV und Sekanten-Verfahren. Dann ist auch hier der Aufwand in einem Schritt durch zwei Auswertung von f gegeben. Es gilt:

ISV: $\left|z^{(k)} - x^*\right| \leq \frac{(b-a)}{2}\left(\frac{1}{2}\right)^{2k} \leq \frac{(b-a)}{2}\left(\frac{1}{4}\right)^k$.

Sekanten: $\left|z^{(k)} - x^*\right| \leq \frac{2m}{M}\widetilde{q}^{(\lambda_1^{2k})} = \frac{2m}{m}\widetilde{q}^{(2.618^k)}$.

Bei gleichen Aufwand konvergiert das Sekanten-Verfahren also schneller als das Newton- oder IS-Verfahren.

Problem beim Sekanten-Verfahren

Die Fehleranalyse wurde ohne Berücksichtigung von Rundungsfehlern gemacht. Seien $x^{(k)}$, $x^{(k-1)}$ sehr nah an x^*, dann liegen auch $f(x^{(k)})$, $f(x^{(k-1)})$ nahe zusammen und haben gleiches Vorzeichen. Da die Differenz in diesem Fall schlecht konditioniert ist, kann es hier zu Auslöschungen kommen.

3.2 Konvergenzordnung von Iterationsverfahren

Wir betrachten allgemeine Iterationsverfahren auf einem Banach-Raum X der Form

$$x^{(k+1)} = \Phi\left(x^{(k)}, \ldots, x^{(k-j)}\right), \; k \geq j \geq 0 \qquad (1)$$

mit Startwerten $x^{(0)}, \ldots, x^{(j)}$ und Iterationsfunktion $\Phi : X^j \longrightarrow X$.

Definition 3.6 Lokale Konvergenz
Das Iterationsverfahren (1) **konvergiert lokal** *gegen ein $x^* \in X$, falls es eine Umgebung U von x^* gibt, so dass für alle $x^{(0)}, \ldots, x^{(j)} \in U$ die Folge $(x^{(k)})_{k\in\mathbb{N}}$ gegen x^* konvergiert.*

Definition 3.7 (Konvergenzordnung)
Sei X ein Banach-Raum und $\left(x^{(k)}\right)_{k\in\mathbb{N}}$ Folge in X, die gegen ein $x^* \in X$ konvergiert. Die Folge hat mindestens die **Konvergenzordnung** *$p \geq 1$ falls*

$$\limsup_{k \longrightarrow \infty} \frac{\left\|x^{(k+1)} - x^*\right\|}{\left\|x^{(k)} - x^*\right\|^p} = c$$

mit $c < 1$ für $p = 1$ und $c < \infty$ für $p > 1$.

Die ***Ordnung ist genau*** **p**, *falls $c \neq 0$.*
Der Fall $p = 1$ heißt **lineare Konvergenz** *und falls für ein $p \geq 1$ gilt*

$$\limsup_{k \longrightarrow \infty} \frac{\left\|x^{(k+1)} - x^*\right\|}{\left\|x^{(k)} - x^*\right\|^p} = 0,$$

so spricht man von **superlinearer Konvergenz**.

Beispiel: Das Newton-Verfahren konvergiert quadratisch ($p = 2$), da $\left|x^{(k+1)} - x^*\right| \leq c \cdot \left|x^{(k)} - x^*\right|^2$. Das ISV Verfahren konvergiert linear, da $\left|x^{(k+1)} - x^*\right| \leq \frac{1}{2}\left|x^{(k)} - x^*\right|$.

Satz 3.8 (Konvergenzordnung von Iterationsverfahren)
Sei $I = [a, b]$, $\Phi : I \longrightarrow I$, $\Phi \in C^p(I)$. Sei $x^{(0)} \in I$, $x^{(k+1)} = \Phi(x^{(k)})$, $k \geq 0$ und es gelte $x^{(k)} \longrightarrow x^*$ mit $x^* = \Phi(x^*)$. Dann gilt:

(i) $p = 1 :$ $x^{(k)} \longrightarrow x^*$ mind. linear $\iff$ $|\Phi'(x^*)| < 1$.

(ii) $p \geq 2 :$ $x^{(k)} \longrightarrow x^*$ mind. mit der Ordnung p, g.d.w.

$$\Phi^{(\nu)}(x^*) = 0, \; (1 \leq \nu \leq p - 1).$$

Beweis: (i) Es existiert eine Zwischenstelle ξ_k zwischen x^k und x^*, so dass gilt

$$\lim_{k\to\infty} \frac{\left\|\Phi(x^{(k)}) - x^*\right\|}{\left\|x^{(k)} - x^*\right\|} = \lim_{k\to\infty} \left|\Phi'(\xi_k)\right| = \left|\Phi'(x^*)\right|.$$

Also folgt die Behauptung.

(ii) "$\Longrightarrow$": Annahme: $\exists\, j < p$ mit $\Phi^{(j)}(x^*) \neq 0$ (j minimal gewählt). Dann folgt mit Taylorentwicklung:

$$\begin{aligned} x^{(k+1)} - x^* &= \Phi(x^{(k)}) - \Phi(x^*) \\ &= \sum_{\nu=1}^{p-1} \frac{\Phi^{(\nu)}(x^*)}{\nu!} \left(x^{(k)} - x^*\right)^\nu + \Phi^{(p)}(\xi_k) \frac{(x^{(k)} - x^*)^p}{p!} \end{aligned}$$

mit ξ_k zwischen $x^{(k)}$ und x^*. Mit der Annahme folgt

$$x^{(k+1)} - x^* = \frac{\Phi^{(j)}(x^*)}{j!} \left(x^{(k)} - x^*\right)^j \cdot \underbrace{\left[1 + \left(x^{(k)} - x^*\right) \sum_{\nu=j+1}^{p} a_\nu \left(x^{(k)} - x^*\right)^{\nu-j-1}\right]}_{:=b}$$

mit $a_\nu := \frac{j!\Phi^{(\nu)}(x^*)}{\nu!\Phi^{(j)}(x^*)}$. Für große k gilt $|b| \geq \frac{1}{2}$, da $x^{(k)} \longrightarrow x^*$ und a_ν unabhängig von k beschränkt. Wir erhalten also

$$\left|x^{(k+1)} - x^*\right| \geq \frac{1}{2} \left|x^{(k)} - x^*\right|^j \frac{\left|\Phi^{(j)}(x^*)\right|}{j!}.$$

Da $x^{(k)} \longrightarrow x^*$ mit Ordnung p, gilt

$$\infty > c \geq \frac{\left|x^{(k+1)} - x^*\right|}{\left|x^{(k)} - x^*\right|^p} \geq \frac{1}{2j!} \left|\Phi^{(j)}(x^*)\right| \cdot \left|x^{(k)} - x^*\right|^{j-p} \longrightarrow \infty.$$

Dies ist ein Widerspruch, da nach Annahme $j - p < 0$.

"$\Longleftarrow$": Es gilt

$$\begin{aligned} x^{(k+1)} - x^* &= \sum_{\nu=1}^{p-1} \frac{\Phi^{\nu}(x^*)}{\nu!} \left(x^{(k)} - x^*\right)^\nu + \frac{\Phi^{(p)}(\xi_k)}{p!} \left(x^{(k)} - x^*\right)^p \\ &= \frac{\Phi^{(p)}(\xi_k)}{p!} \left(x^{(k)} - x^*\right)^p, \text{ da } \Phi^{(\nu)}(x^*) = 0 \ (1 \leq \nu \leq p) \\ \Longrightarrow \quad \frac{\left|x^{(k+1)} - x^*\right|}{\left|x^{(k)} - x^*\right|^p} &= \frac{1}{p!}\Phi^{(p)}(\xi_k) \overset{k\to\infty}{\to} \frac{1}{p!}\Phi^{(p)}(x^*), \end{aligned}$$

da ξ_k zwischen $x^{(k)}$ und x^* und $x^{(k)} \overset{k\to\infty}{\to} x^*$. □

3.2.1 Verfahren höher Ordnung ($p = 3$)

1. Idee: Konstruiere Verfahren 3. Ordnung aus einer Linearkombination von 2 Verfahren zweiter Ordnung. Seien $\Phi_0,\ \Phi_1$ Iterationsverfahren, die quadratisch konvergieren. Betrachte

$$\Phi_s(x) = (1 - s)\Phi_0(x) + s\Phi_1(x).$$

Dann gilt $\Phi_s'(x^*) = (1 - s)\Phi_0'(x^*) + s\Phi_1'(x^*) = 0$, da nach Satz 3.7 $\Phi_0'(x^*) = \Phi_1'(x^*) = 0$.

Bestimmt man s so, dass $\Phi_s''(x^*) = 0$ gilt, so konvergiert nach Satz 3.7 Φ_s mindestens mit Ordnung 3. (Beispiel: siehe Übungsblatt).

2. Idee: (Verbessertes Newton-Verfahren)

Ansatz: $\Phi(x) = x - g(x)f(x) - h(x)f(x)^2$ mit $g(x) \neq 0,\ h(x) \neq 0$ in einer Umgebung von x^*. Dann gilt $\Phi(x^*) = x^* \iff f(x^*) = 0$.

Idee: Bestimme $g,\ h$, so dass $\Phi'(x^*) = \Phi''(x^*) = 0$ für eine einfache Nullstelle x^*.

Sei x^* einfache Nullstelle, so gilt

$$\Phi'(x) = 1 - f'(x)g(x) - f(x)g'(x) - 2f(x)f'(x)h(x) - h'(x)f^2(x).$$

Also folgt $\Phi'(x^*) = 0 \overset{f(x^*)=0}{\iff} 1 - f'(x^*)g(x^*) = 0 \implies g(x) = \frac{1}{f'(x)}$.

Analog folgt für Φ'' mit dieser Wahl von g:

$$\begin{aligned}\Phi''(x) &= -f'(x)\left(g'(x) + h(x)f(x) + 2h(x)f'(x)\right) - f(x)(\cdots) - \underbrace{\Big(\underbrace{g(x)f'(x)}_{=1}\Big)'}_{=0} \\ &= \frac{f''(x)}{f'(x)} - 2h(x)f'(x) - f(x)(\cdots).\end{aligned}$$

Aus

$$0 = \Phi''(x^*) = \frac{f''(x^*)}{f'(x^*)} - 2h(x^*)f'(x^*)^2$$

folgt also für die Wahl von h:

$$h(x) = \frac{1}{2f'(x)^2}\frac{f''(x)}{f'(x)} = \frac{f''(x)}{2f'(x)^3}.$$

Folgerung 3.9 (Verbessertes Newton-Verfahren)
Das Verfahren, definiert durch

$$\Phi(x) = x - \frac{f(x)}{f'(x)} - \frac{1}{2}\frac{f(x)^2 f''(x)}{f'(x)^3}$$

konvergiert in der Umgebung einer einfachen Nullstelle x^* von f mit mindestens dritter Ordnung.

3.2.2 Newton-Verfahren für mehrfache Nullstellen

Definition 3.10 (Ordnung und Vielfachheit einer Nullstelle)
Sei $f \in C^n(I)$ $(n \geq 1)$. f hat eine **Nullstelle x^* der Ordnung $(\mathbf{n-1})$** *bzw. der* **Vielfachheit n** *gdw. $f^{(\nu)}(x^*) = 0,\ 0 \leq \nu \leq n-1$ und $f^{(n)}(x^*) \neq 0$.*

Satz 3.11
Sei $f \in C^{n+1}(a,b),\ n \geq 1$ und $x^* \in (a,b)$ eine n-fache Nullstelle von f mit $f^{(k)}(x) \neq 0$ $(x \neq x^*)$ $0 \leq k \leq n-1$ und $f^{(n)}(x) \neq 0$ für $x \in (a,b)$. Dann existiert eine Umgebung von x^*, so dass das Newton-Verfahren mindestens linear konvergiert.

Beweis: Setze $\Phi(x) = \begin{cases} x - \frac{f(x)}{f'(x)} & : \ x \neq x^* \\ x^* & : \ x = x^* \end{cases}$.

Zu zeigen: $\Phi \in C^1(a,b)$ und $|\Phi(x)| < 1$. Zusammen mit Satz 3.8 folgt dann die Behauptung des Satzes. Es gilt

$$\lim_{\substack{x\to x^*\\ x\neq x^*}} \Phi(x) = x^* - \lim_{x\to x^*} \frac{f(x)}{f'(x)} \overset{\text{L'Hôpital}}{=} x^* - \lim_{x\to x^*} \frac{f^{(n-1)}(x)}{f^{(n)}(x)} = x^* + \frac{0}{f^{(n)}(x^*)} = x^*.$$

Weiter folgt mit L'Hopital und $\Phi'(x) = 1 - \frac{f'(x)^2 - f(x)f''(x)}{f'(x)^2} = \frac{f(x)f''(x)}{f'(x)^2}$

$$\lim_{x\to x^*} \Phi'(x) = 1 - \frac{1}{n} \implies |\Phi'(x^*)| < 1.$$

□

Idee: Modifiziere das Verfahren je nach Vielfachheit:

Wähle $\Phi_\alpha(x) = x - \alpha\frac{f(x)}{f'(x)},\ \alpha \in \mathbb{R}$ fest

$$\implies \Phi'_\alpha(x^*) = 1 - \frac{\alpha}{n} = 0 \iff \alpha = n$$

Folgerung 3.12
Sei $f \in C^{n+1}(a,b),\ n \geq 1$ und x^* eine n-fache Nullstelle, dann konvergiert das Verfahren

$$x^{(k+1)} = x^{(k)} - n\frac{f(x^{(k)})}{f'(x^{(k)})}$$

quadratisch gegen x^*.

3.3 Nichtlineare Gleichungssysteme

Problem: $D \subset \mathbb{R}^n$ abgeschlossen und konvex ($n \geq 2$),
$f : D \longrightarrow \mathbb{R}^n$, d.h. $f(x) = (f_1(x), \ldots, f_n(x))^\top$, $x \in D$.

Gesucht: $x^* \in D$ mit $f(x^*) = 0$, d.h. $f_i(x^*) = 0$ $(i = 1, \ldots, n)$.
Annahme: Es existiere ein $x^* \in D$ mit $f(x^*) = 0$.

3.3.1 Newton-Verfahren für nichtlineare Systeme

Idee: Iteriere analog zum skalaren Fall mit

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

Dabei ist $Df(x)$ die Jaccobi-Matrix von f und $Df(x)^{-1}$ die Inverse.

Algorithmus: (Newton-Verfahren für Systeme)

`Sei` $x^{(0)} \in \mathbb{R}^n$ `gegeben. Für` $k \geq 0$ `iteriere`

`1) Löse Defektgleichung:` $Df(x^{(k)})\ y^{(k)} = -f(x^{(k)})$

`2) Setze neu Iterierte:` $x^{(k+1)} = x^{(k)} + y^{(k)}$

Problem: In jedem Schritt muss ein $n \times n$ LGS gelöst werden. Häufig wird Df auch für einige Schritte festgehalten. Mit der LR-Zerlegung können diese Schritte dann sehr effizient gelöst werden.

Satz 3.13
Sei $f_i \in C^2(\mathbb{R})$ und $Df(x^*)$ regulär (d.h. x^* einfache Nullstelle).
Dann existiert ein $\rho > 0$, so dass für alle $x^{(0)} \in B_\rho(x^*) := \{x \mid \|x - x^*\| < \rho\}$ gilt: $x^{(k)} \in B_\rho(x^*)$ und $x^{(k)} \overset{k\to\infty}{\longrightarrow} x^*$ mindestens quadratisch.

Beweis: Analog zum Satz 3.2 □

Kapitel 4

Eigenwertprobleme

Definition 4.1 (Eigenwertproblem)
Sei $A \in \mathbb{K}^{n\times n}$ mit $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$. Dann heißt $\lambda \in \mathbb{K}$ Eigenwert von A, wenn ein Eigenvektor $x \in \mathbb{K}^n$ mit $x \neq 0$ existiert mit der Eigenschaft $Ax = \lambda x$.

Vollständiges Eigenwertproblem: Finde alle EW (und EV) von A

Partielles Eigenwertproblem: Finde einzelne EW (und EV) von A (z.B. den größten und kleinsten EW)

4.1 Grundbegriffe der linearen Algebra und theoretische Grundlagen

Notation und Grundlagen 4.2

1) Für $x, y \in \mathbb{K}^n$ bezeichne $\langle x, y\rangle$ und $\|x\|$ das euklidische Skalarprodukt und die euklidische Norm.

2) $\|A\|$ bezeichnet die Spektralnorm von A. Bemerke: $\rho(A) \leq \|A\|_*$ für alle Normen $\|\cdot\|_*$.

3) Für $\mathbb{K} = \mathbb{C}$ ist $A^H := \bar{A}^T$. A heißt hermitesch, falls $A^H = A$.

4) Die EWe von A sind die Nullstellen des charakteristischen Polynoms

$$\rho_A(\lambda) := \det(A - \lambda I).$$

5) Ist ein EW λ bekannt, so findet man alle EVen zu λ durch Lösung des singulären homogenen Gleichungssystems

$$(A - \lambda I)x = 0$$

Umgekehrt bestimmt ein EV $x \neq 0$ den zugehörigen EW λ durch den Rayleigh Quotienten

$$R(x) = \frac{\langle Ax, x\rangle}{\|x\|^2},$$

d.h. $\lambda = R(x)$.

6) $\sigma(A) := \{\lambda \,|\, \lambda \text{ EW von } A\}$ heißt Spektrum von A.

7) Das charakteristische Polynom besitzt die Darstellung

$$\rho_A(x) = \prod_{i=1}^{s} (x - \lambda_i)^{\sigma_i},$$

wobei λ_i die paarweise verschiedenen Eigenwerte von A sind. Es gilt $\sum_{i=1}^{s} \sigma_i = n$ und σ_i heißt algebraische Vielfachheit von λ_i.
Die Eigenvektoren zu λ_i (Vereinigt mit dem Nullvektor) bilden den sogenannten Eigenraum $E_i := \text{Kern}(A - \lambda_i I)$. Ist $\rho_i := \dim(E_i)$, so heißt ρ_i geometrische Vielfachheit von λ_i.

Ähnlichkeitstransformationen 4.3
Ist $T \in \mathbb{K}^{n\times n}$ regulär, so heißte $B := T^{-1}AT$ Ähnlichkeitstransformation und B ähnlich zu A.

1) Ähnliche Matrizen besitzen dieselben EWe λ, denn mit $y := T^{-1}x$ folgt:

$$T^{-1}ATy = T^{-1}Ax = \lambda T^{-1}x = \lambda y.$$

D.h. λ ist EW zu B mit EV $y = T^{-1}x$.

2) Ähnliche Matrizen besitzen dasselbe char. Polynom, denn

$$\begin{aligned} \det(T^{-1}AT - \lambda I) &= \det(T^{-1}(A - \lambda I)T) \\ &= \det(T^{-1})\det(A - \lambda I)\det(T) \\ &= \det(A - \lambda I). \end{aligned}$$

Idee einer numerischen Methode 4.4
Wende eine Folge von Ähnlichkeitstransformationenen an, um A in einfachen Gestalt zu transformieren, d.h.

$$A^{(0)} := A,$$
$$A^{(i)} := T_i^{-1}A^{(i-1)}T_i, \quad i = 1, 2, 3, \ldots$$

und geeignete Matrizen T_i.

Satz 4.5 (Satz von Schur)
Zu jeder Matrix $A \in \mathbb{C}^{n\times n}$ existiert eine unitäre Matrix $U \in \mathbb{C}^{n\times n}$ (d.h. $U^H U = I$) mit

$$U^H AU = \begin{pmatrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}.$$

Beweis: (Induktion über n)
Ind. Anf: $n = 1$ $\surd$
Ind. Vor: Für $A \in \mathbb{C}^{(n-1)\times(n-1)}$ gilt die Behauptung.
Ind. Beh: Sei $A \in \mathbb{C}^{n\times n}$. Zeige: $\exists U$ unitär mit $U^H AU = \begin{pmatrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$.

Sei $\lambda \in \mathbb{C}$ EW von A und $z \in \mathbb{C}^n \setminus \{0\}$ ein EV mit $\|z\| = 1$.
Wir können z zu einer Orthonormalbasis auf $\mathbb{C}^n$ ergänzen, d.h. wir können eine unitäre Matrix $\hat{V}$ finden, so dass

$$V = (z, \hat{V}) \text{ mit } V^H V = I.$$

Sei $B = V^H A V$. Wegen $VBe_1 = AVe_1 = Az = \lambda z = \lambda V e_1$
$\Longrightarrow$ $Be_1 = \lambda e_1$, d.h. die erste Spalte von B ist ein λ-faches des ersten Einheitsvektors.
Also $B = V^H A V = \left(\begin{array}{c|c} \lambda & b \\ \hline 0 & C \end{array}\right)$.
Nach Ind. Vor. ex. eine unitäre Matrix W mit $W^H C W = T$ und T hat obere Dreiecksgestalt.
$\Longrightarrow$ $A = V \left(\begin{array}{c|c} \lambda & b \\ \hline 0 & W^H C W \end{array}\right) V^H = \underbrace{V \left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & W \end{array}\right)}_{=:U} \left(\begin{array}{c|c} \lambda & b \\ \hline 0 & T \end{array}\right) \underbrace{\left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & W^H \end{array}\right) V^H}_{:=U^H}$.
Daraus folgt die Behauptung. □

Folgerung 4.6 (Schur für hermitesche Matrix)
Sei $A \in \mathbb{C}^{n \times n}$ hermitesch. Dann ex. eine unitäre Matrix $U = (u_1, \ldots, u_n)$ mit

$$U^H A U = diag(\lambda_1, \ldots, \lambda_n).$$

$\lambda_1, \ldots, \lambda_n$ sind EW von A, die reell sind und u_i die EVen zu λ_i.
Insbesondere ist A eine Matrix mit n linear unabhängigen zu einander orthogonalen Eigenvektoren.

Beweis: A hermitesch $\Longrightarrow$ $A^H = A$ und somit

$$(U^H A U)^H = U^H A (U^H)^H = U^H A U.$$

$\Longrightarrow$ $U^H A U$ ist selbst wieder hermitesch.
$\Longrightarrow$ Beh. mit 4.5. □

4.2 Kondition des Eigenwertproblems

Satz 4.7 (Gerschgorinscher Kreissatz)
Sei $A \in \mathbb{C}^{n\times n}$. Für $i = 1, \ldots, n$ definiere die sogenannten Gerschgorin Kreise

$$G_i := \{z \in \mathbb{C} \,|\, |z - a_{ii}| \le r_i\}, \quad r_i := \sum_{j=1, j\neq i}^{n} |a_{ij}| \,.$$

Dann gilt:

(i) Ist λ EW von A, so ist $\lambda \in \bigcup_{i=1}^{n} G_i$, d.h. $\sigma(A) \subset \bigcup_{i=1}^{n} G_i$.

(ii) Hat die Vereinigung $\hat{G}$ von m Kreisen G_i einen leeren Schnitt mit den restlichen $n-m$ Kreisen, so enthält $\hat{G}$ genau m EWe von A (gezählt mit ihren algebraischen Vielfachheiten).

Beweis:

zu (i): Sei λ EW von A mit EV $x \neq 0$.
Aus $Ax = \lambda x$ folgt:

$$(\lambda - a_{ii})x_i = \sum_{j=1, j\neq i}^{n} a_{ij}x_j \quad \forall i = 1, \ldots, n.$$

Für $i \in \{1, \ldots, n\}$mit $|x_i| = \max_{i=1,\ldots,n} |x_j|$ folgt:

$$|\lambda - a_{ii}| = \left| \sum_{j=1, j\neq i}^{n} \frac{a_{ij}x_j}{x_i} \right| \le \sum_{j=1, j\neq i}^{n} |a_{ij}|$$

$\Longrightarrow \quad \lambda \in G_i \subset \bigcup_{j=1}^{n} G_j$.

zu (ii): Wir setzen $D = (a_{ii}\delta_{ij})_{i,j=1,\ldots,m}$ und betrachten

$$B(t) := D + t(A - D), \quad 0 \le t \le 1$$

mit Gerschgorin Kreisen

$$G_i(t) := \left\{ z \in \mathbb{C} \,|\, |z - a_{ii}| \le t \cdot \sum_{j=1, j\neq i}^{n} |a_{ij}| \right\} \quad i = 1, \ldots, n.$$

Es ist $B(0) = D$ und $B(1) = A$. Die EW von $B(t)$ sind die Nullstellen von $\rho_{B(t)}$ und hängen daher stetig von t ab. Wende (i) auf $B(t)$ an und lasse t von 0 nach 1 laufen.
Dabei wird der Radius der Kreise bei festem Mittelpunkt immer größer.
Die Anzahl der EWe in einem Kreis $G_i(t)$ kann sich erst dann ändern, wenn dieser einen anderen Kreis trifft. Daraus folgt die Beh.

□

Folgerung 4.8

Da A und A^H dieseleben Eigenwerte besitzen, gilt der Satz 4.7 auch mit

$$G_i' := \{z \in \mathbb{C} \mid |z - a_{ii}| \leq \sum_{j=1, j\neq i}^{n} |a_{ji}|$$

$$\Longrightarrow \quad \sigma(A) \subset \left(\bigcup_{i=1}^{n} G_i\right) \cap \left(\bigcup_{i=1}^{n} G_i'\right).$$

Lemma 4.9
Seien $A, B \in \mathbb{C}^{n\times n}$. Ist λ ein Eigenwert von A, aber kein Eigenwert von B, so gilt

$$1 \leq \left\|(\lambda I - B)^{-1}(A - B)\right\|.$$

Beweis: Sei $Ax = \lambda x,\ x \neq 0 \quad \Longrightarrow \quad \lambda x - Bx = (A - B)x$
$\Longrightarrow \quad x = (\lambda I - B)^{-1}(A - B)x.$

Also ist 1 ein EW von $(\lambda I - B)^{-1}(A - B)$, d.h.

$$1 \leq \rho((\lambda I - B)^{-1}(A - B)) \leq \left\|(\lambda I - B)^{-1}(A - B)\right\|.$$

□

Folgerung 4.10
Lemma 4.9 gilt insbesondere für $B = D = (a_{ij}\delta_{ij})_{i,j}$. Da die Spektralnorm kleiner als z.B. die Zeilensummennorm ist, folgt dann

$$1 \leq \max_{i=1,\ldots,n} \sum_{k\neq j} \left|\frac{a_{jk}}{\lambda - a_{ii}}\right|$$

$\Longrightarrow \quad \exists j$ mit $|\lambda - a_{jj}| = \sum_{k\neq j} |a_{jk}| \quad \Longrightarrow \quad \lambda \in G_j \subset \bigcup_{i=1}^{n} G_i$

Satz 4.11 (Kondition des Eigenwertproblems)
Sei $B \in \mathbb{C}^{n\times N}$ diagonalisierbar, d.h. es ex eine reguläre Matrix $P \in \mathbb{C}^{n\times n}$ mit

$$P^{-1}BP = D = diag(\lambda_1(B), \ldots, \lambda_n(B)).$$

Dann gibt es zu jedem EW $\lambda_j(B)$ einen EW $\lambda_j(A)$ von $A \in \mathbb{C}^{n\times n}$ mit

$$|\lambda_j(A) - \lambda_j(B)| \leq \kappa(P) \|A - B\|.$$

Dabei ist $\kappa(P) = \frac{\lambda_{\max}(P)}{\lambda_{\min}(P)}$ die Kondition von P.

Beweis:

$$\begin{aligned}
\left\|(\lambda I - B)^{-1}\right\| &= \left\|P(\lambda I - D)^{-1}P^{-1}\right\| \leq \left\|(\lambda I - D)^{-1}\right\| \kappa(P) \\
&\leq \max_j \frac{1}{|\lambda - \lambda_j(B)|}\kappa(P) \\
&= \frac{1}{\min_j |\lambda - \lambda_j(B)|}\kappa(P).
\end{aligned}$$

Aus Lemma 4.7 folgt:

$$1 \leq \left\|(\lambda I - B)^{-1}\right\| \|A - B\| .$$

Also folgt:

$$1 \leq \frac{1}{\min_j |\lambda - \lambda_j(B)|} \kappa(P) \|A - B\|$$

$$\Longrightarrow \quad \min_j |\lambda - \lambda_j(B)| \leq \kappa(P) \|A - B\|$$
$$\Longrightarrow \quad \exists \lambda = \lambda_j(A) \text{ mit } |\lambda_j(A) - \lambda_j(B)| \leq \kappa(P) \|A - B\| .$$

□

Für unitäre Matrizen U gilt $\kappa(U) = 1$. Also ist nach Folgerung 4.6 das Eigenwertproblem für hermitesche Matrizen gut konitioniert.

Beispiel 4.12 (Anwendung der Gerschgorin-Kreise)

$$A = \begin{pmatrix} 0,9 & 0 & \\ 0 & 0,4 & 0 \\ 0 & 0 & 0,2 \end{pmatrix} + 10^{-5} \begin{pmatrix} 0,1 & 0,4 & -0,2 \\ -0,1 & 0,5 & 0,1 \\ 0,2 & 0,1 & 0,3 \end{pmatrix}$$

Nach Satz 4.11 erwartet man, dass die EW einer Matrix A mit kleinen Außendiagonalelementen ungefähr mit dem Diagonalelementen übereinstimmen.
Die 3 Gerschgorinkreise sind disjunkt, daher besitzt A die EW $\lambda_1, \lambda_2, \lambda_3$ mit

$$\begin{aligned} \left|\lambda_1 - (0,9 + 0,1 \cdot 10^{-5})\right| &\leq 0,6 \cdot 10^{-5}, \\ \left|\lambda_2 - (0,4 + 0,5 \cdot 10^{-5})\right| &\leq 0,2 \cdot 10^{-5}, \\ \left|\lambda_3 - (0,2 + 0,3 \cdot 10^{-5})\right| &\leq 0,3 \cdot 10^{-5}. \end{aligned}$$

Diese Abschätzungen können noch wesentlich verbessert werden.
Sei $P := diag(10^5, 1, 1)$. Dann ist

$$P^{-1}AP = \begin{pmatrix} 0,9 & 0 & 0 \\ 0 & 0,4 & 0 \\ 0 & 0 & 0,2 \end{pmatrix} + \begin{pmatrix} 0,1 \cdot 10^{-5} & 0,1 \cdot 10^{-10} & -0,2 \cdot 10^{-10} \\ -0,1 & 0,5 \cdot 10^{-5} & 0,1 \cdot 10^{-5} \\ 0,2 & 0,1 \cdot 10^{-5} & 0,3 \cdot 10^{-5} \end{pmatrix} .$$

Der erste Gerschgorin Kreis ist noch disjunkt zu den beiden anderen, die nicht mehr disjunkt sind. Also Folgt für λ_1:

$$\left|\lambda_1 - (0,9 + 0,1 \cdot 10^{-5})\right| \leq 0,6 \cdot 10^{-10}.$$

Entsprechend kann die Abschätzung für λ_2, λ_3 verbessert werden (siehe Abbildung 4.1).

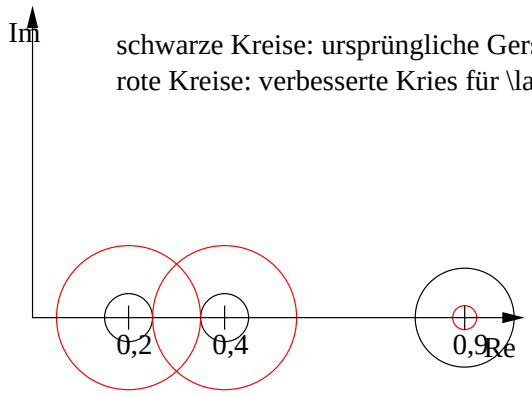

Abbildung 4.1: Gerschgorinkreise: Beispiel, Radien nicht maßstabsgetreu!!

4.3 Variationsprinzip für Eigenwerte hermitescher Matrizen

Satz 4.13 (Rayleighsches Maximumsprinzip)
Sei $A \in \mathbb{K}^{n\times n}$ hermitesch. Die EW von A seien $\lambda_1 \geq \ldots \geq \lambda_n$. Sei $U = (u_1, \ldots, u_n)$ unitäre Matrix mit $U^H AU = diag(\lambda_1, \ldots, \lambda_n) =: \Lambda$.
Für $j = 1, \ldots, n$ definiere den $(n+1-j)$-dimensionalen Teilraum

$$M_j := \{x \in \mathbb{K}^n \mid \langle u_i, x\rangle = 0 \,\forall i = 1, \ldots, j-1\} = (\text{span}(u_1, \ldots, u_n))^{\perp}.$$

Dann gilt:

$$\lambda_j = \max_{x\in M_j\setminus\{0\}} R(x)$$

mit dem Rayleigh-Quotienten $R(x) := \frac{\langle Ax,x\rangle}{\langle x,x\rangle}$.

Beweis: Sei $j \in \{1, \ldots, n\}$ und $x \in M_j \setminus \{0\}$ beliebig.
setze $y := U^H x$. Dann folgt:

$$y_i = \langle u_i, x\rangle = 0, \quad i = 1, \ldots, j-1.$$

Wegen $\lambda_i \leq \lambda_j$ für $i = j, \ldots, n$ gilt:

$$R(x) = \frac{\langle Ax,x\rangle}{\langle x,x\rangle} = \frac{\langle U\Lambda U^H x,x\rangle}{\langle x,x\rangle} = \frac{\langle \Lambda U^H x, U^H x\rangle}{\langle x,x\rangle} = \frac{\langle \Lambda y, y\rangle}{\langle y,y\rangle} = \frac{\sum_{i=1}^n \lambda_i |y_i|^2}{\sum_{i=1}^n |y_i|^2} \leq \lambda_j$$

und somit

$$\sup_{x\in M_j\setminus\{0\}} R(x) \leq \lambda_j.$$

Andererseits ist $u_j \in M_j \setminus \{0\}$ und $R(u_j) = \lambda_j$.
$\Longrightarrow$ $\max_{x\in M_j\setminus\{0\}} R(x) = \lambda_j$. □

Bemerkung 4.14
Definiert man $f : \mathbb{R} \longrightarrow \mathbb{R}$ bei festem $x \in \mathbb{K}^n \setminus \{0\}$ durch

$$f(\lambda) = \frac{1}{2}\|Ax - \lambda x\|^2 = \frac{1}{2}\lambda^2 \|x\|^2 - \lambda \langle Ax, x\rangle + \frac{1}{2}\|Ax\|^2$$

so nimmt f in $\lambda = R(x)$ sein Minimum an.
Ist daher x nähreungsweise ein EV von A, so ist $R(x)$ eine gute Näherung des zugehörigen Eigenwertes.

Satz 4.15 (Courantsches Minimum-Maximum Prinzip)
Sei $A \in \mathbb{K}^{n\times n}$ hermitesch mit EWen $\lambda_1 \geq \ldots \geq \lambda_n$. Für $j = 1, \ldots, n$ definiere

$$\mathcal{N}_j := \{N_j \subset \mathbb{K}^n \,|\, N_j \text{ ist linearer Teilraum der Dimension } n+1-j\}$$

Dann gilt:

$$\lambda_j = \min_{N_j \in \mathcal{N}_j} \max_{x \in N_j \setminus \{0\}} R(x), \quad j = 1, \ldots, n.$$

Beweis: Sei $U = (u_1, \ldots, u_n)$ die unitäre Matrix von EVen zu den EWen $\lambda_1, \ldots, \lambda_n$ von A. Sei $j \in \{1, \ldots, n\}$. Definiere $L_j = \text{span}(u_1, \ldots, u_j)$ und wähle $N_j \in \mathcal{N}_j$ beliebig.
Wegen

$$\begin{aligned} \dim(L_j \cap N_j) &= \dim(L_j) + \dim(N_j) - \dim(L_j \cup N_j) \\ &= n + 1 - \underbrace{\dim(L_j \cup N_j)}_{\leq n} \geq 1 \end{aligned}$$

existiert ein $x \in L_j \cap N_j$ mit $x \neq 0$.
Da $x \in L_j$, folgt: $x = \sum_{i=1}^{j} \alpha_i u_i$.
$\Longrightarrow \quad R(x) = \frac{\sum_{i=1}^{j} \lambda_i |\alpha_i|^2}{\sum_{i=1}^{j} |\alpha_i|^2} \geq \lambda_j$, da $\lambda_i \geq \lambda_j$ für $i = 1, \ldots, j$

$$\Longrightarrow \quad \min_{N_j \in \mathcal{N}_j} \max_{x \in N_j \setminus \{0\}} R(x) \geq \lambda_j$$

Wählt man andererseits $N_j = M_j$, so gilt nach Satz 4.15

$$\max_{x \in M_j \setminus \{0\}} R(x) = \lambda_j.$$

□

Folgerung 4.16
Seien $A, B \in \mathbb{K}^{n\times n}$ hermitesch und gelte $\lambda_1(A) \geq \ldots \geq \lambda_n(A)$ sowie $\lambda_1(B) \geq \ldots \geq \lambda_n(B)$.
Dann gilt die Abschätzung:

$$|\lambda_j(A) - \lambda_j(B)| \leq \|A - B\|, \quad j = 1, \ldots, n.$$

Beweis: A, B hermitesch $\Longrightarrow$ $E = A - B$ hermitesch. Sei $x \in \mathbb{K}^n \setminus \{0\}$.
Dann gilt:

$$R_E(x) = \frac{\langle (A-B)x, x\rangle}{\langle x, x\rangle} \leq \|A - B\|.$$

Somit folgt

$$(*) \qquad R_A(x) \leq R_B(x) + \|A - B\|.$$

Sei $j \in \{1, \ldots, n\}$ und $N_j \in \mathcal{N}_j$ (vgl. 4.15), so folgt aus $(*)$:

$$\min_{N_j \in \mathcal{N}_j} \max_{x \in N_j \setminus \{0\}} R_A(x) \leq \min_{N_j \in \mathcal{N}_j} \max_{x \in N_j \setminus \{0\}} R_B(x) + \|A - B\| .$$

Mit dem Courantschen Min-Max Prinzip folgt $\lambda_j(A) \leq \lambda_j(B) + \|A - B\|$.
Vertauscht man die Rollen von A und B, so folgt auch

$$\lambda_j(B) \leq \lambda_j(A) + \|A - B\|$$

$$\Longrightarrow \quad |\lambda_j(B) - \lambda_j(A)| \leq \|A - B\| .$$

□

4.4 Transformation auf Hessenberg-Form

Definition 4.17 (Hessenberg-Matrizen)
Eine Matrix $A = (a_{ij})_{i,j} \in \mathbb{R}^{n\times n}$ *heißt eine (obere)* Hessenberg-Matrix*, wenn* $a_{ij} = 0$ *für* $1 \le j \le i-2$*, d.h.*

$$A = \begin{pmatrix} * & \dots & \dots & * \\ * & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ 0 & & * & * \end{pmatrix}$$

Eine Hessenberg-Matrix heißt unreduziert, falls sämtliche Subdiagonalelemente $a_{i+1,i}$*,* $i = 1, \dots, n-1$ *ungleich 0 sind.*
Eine symmetrische Hessenberg-Matrix ist somit eine symmetrische Tridiagonalmatrix.

Erinnerung an die Numerik I 4.18 (Householder-Matrix)
Ist $x \in \mathbb{R}^n \backslash \{0\}$ und definiert man $u := x + \text{sign}(x_1)\,\|x\|\,e_1 \in \mathbb{R}^n$, so ist durch $P = I - \frac{2uu^T}{\langle u,u\rangle} = I - \beta uu^T$ mit $\beta := \frac{2}{\langle u,u\rangle} = \frac{1}{\|x\|(\|x\|+|x_1|)}$ eine Householder Matrix definiert mit $Px = -sgn(x_1)\,\|x\|\,e_1$.

Satz 4.19 (Householder-Transformationen auf Hessenberg-Form)
Zu einer Matrix $A \in \mathbb{R}^{n\times n}$ existieren $n-2$ Hoseholder-Matrizen $P_1, \dots, P_{n-2}$, so dass

$$P_{n-2} \dots P_1 A P_1 \dots P_{n-2}$$

eine zu A orthogonal-ähnliche Hessenberg-Matrix ist.

Beweis: (Induktion über $k = 1, \dots, n-2$)

Angenommen es seien schon $k-1$ Householder-Matrizen $P_1, \dots, P_{k-1}$ bestimmt, so dass

$$P_{k-1} \dots P_1 A P_1 \dots P_{k-1} = \left(\begin{array}{c|c|c} H_k & & B_k \\ \hline 0 & a_k & C_k \end{array}\right) \begin{matrix} \}k \\ \}n-k \end{matrix},$$

wobei $H \in \mathbb{R}^{k\times k}$ eine Hessenberg-Matrix, $B_k \in \mathbb{R}^{k\times k}, C \in \mathbb{R}^{(n-k)\times(n-k)}$ und $a_k \in \mathbb{R}^{n-k}$ ist.

Für P_k machen wir folgenden Ansatz:

$$P_k = \text{diag}(I_k, \tilde{P}_k) = \left(\begin{array}{c|c} I_k & 0 \\ \hline 0 & \tilde{P}_k \end{array}\right)$$

mit der Identität $I_k \in \mathbb{R}^{k\times k}$ und einer $(n-k)\times(n-k)$ Householder-Matrix $\tilde{P}_k$. Dann folgt:

$$P_k P_{k-1} \dots P_1 A P_1 \dots P_{k-1} P_k = \left(\begin{array}{c|c} I_k & 0 \\ \hline 0 & \tilde{P}_k \end{array}\right) \left(\begin{array}{c|c|c} H_k & & B_k \\ \hline 0 & a_k & C_k \end{array}\right) \left(\begin{array}{c|c} I_k & 0 \\ \hline 0 & \tilde{P}_k \end{array}\right)$$

$$= \left(\begin{array}{c|c|c} H_k & & B_k\tilde{P}_k \\ \hline 0 & \tilde{P}_k a_k & \tilde{P}_k C_k \tilde{P}_k \end{array}\right) = \left(\begin{array}{c|c|c} H_{k+1} & & B_{k+1} \\ \hline 0 & a_{k+1} & C_{k+1} \end{array}\right) \begin{matrix} \}k \\ \}n-k \end{matrix}$$

mit einer Hessenberg-Matrix $H_{k+1} \in \mathbb{R}^{(k+1)\times(k+1)}$, wenn die Householder-Matrix $\tilde{P}_k$ so gewählt wird, dass $\tilde{P}_k a_k$ ein Vielfaches des ersten Einheitsvektors in $\mathbb{R}^{n-k}$ ist (siehe 4.18). □

Algorithmus 4.20 (Householder-Transformation auf Hessenberg-Form)
Input: $A \in \mathbb{R}^{n\times n}$
Für $k = 1, \ldots, n-2$

Falls $a_k = (a_{k+1,k}, \ldots, a_{n,k})^T \neq 0$

dann Berechne Householder Matrix $\tilde{P}_k$ durch
$u_k := (a_{k+1,k} + \mathrm{sign}(a_{k+1,k}) \|a_k\|, a_{k+2,k}, \ldots, a_{n,k})^T$
$\beta_k := \|a_k\|^{-1} (\|a_k\| + |a_{k+1,k}|)^{-1}$
$\tilde{P}_k := I_{n-k} - \beta_k u_k (u_k)^T$ und berechne $A := P_k A P_k$

sonst setze $P_k := I$

Output: Die Ausgangsmatrix A wird in $n-2$ Schritten mit orthogonal-ähnlichen Transformationen in eine Hessenberg-Matrix $P^T AP, P = P_1 \cdot \ldots \cdot P_{n-2}$ überführt.

Bemerkung:
Da orthogonale Ähnlichkeitstransformationen die Symmetrie erhalten, wird eine symmetrische Matrix A durch $n-2$ Schritte auf eine Tridiagonalmatrix transformiert.

4.5 Eigenwertbestimmung für Hessenberg-Matrizen

Grundlegende Idee 4.21
Bestimme das charakteristische Polynom $\rho_A(\lambda)$ einer Hessenberg-Matrix A, sowie die Ableitung $\rho'_A(\lambda)$, so dass man die Eigenwerte durch Nullstellensuchen, etwa mit dem Newtonverfahren, berechnen kann.

Satz 4.22 (Berechnung von $\rho_A(\lambda), \rho'_A(\lambda)$ für Tridiagonalmatrizen)
Sei $T \in \mathbb{R}^{n\times n}$ eine symmetrische Tridiagonalmatrix, d.h.

$$T = \begin{pmatrix} b_1 & c_1 & & & 0 \\ c_1 & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & c_{n-1} \\ 0 & & & c_{n-1} & b_n \end{pmatrix}$$

und $\rho_A(\lambda) = det(T - \lambda I)$.
Für $k = n, \ldots, 0$ seien $f_k(\lambda), g_k(\lambda)$ definiert durch:

$$\begin{aligned} f_n(\lambda) &= 1, & g_n(\lambda) &= 0, \\ f_{n-1}(\lambda) &= b_1 - \lambda, & g_{n-1}(\lambda) &= -1, \\ f_{n-i-1}(\lambda) &= (b_{i+1} - \lambda) f_{n-i}(\lambda) - |c_i|^2 f_{n-i+1}(\lambda), \\ g_{n-i-1}(\lambda) &= -f_{n-i}(\lambda) + (b_{i+1} - \lambda) g_{n-i}(\lambda) - |c_i|^2 g_{n-i+1}(\lambda). \end{aligned}$$

Dann gilt:

$$\begin{aligned} f_0(\lambda) &= \rho_A(\lambda), \\ g_0(\lambda) &= \rho'_A(\lambda). \end{aligned}$$

Beweis: Wir zeigen durch vollst. Induktion über i, dass $f_{n-i}(\lambda)$ die Determinate des i-ten Hauptminors von $T - \lambda I$ ist:

Für $i = 0, 1$ ist dei Aussage richtig.
Sei die Aussage richtig für alle j mit $1 \le j \le i$, $2 \le i$.
Dann folgt mit Determinanten-Entwicklungssatz

$$\det \left(\begin{array}{ccccc|c|c} b_1 - \lambda & c_1 & & & & & 0 \\ c_1 & \ddots & \ddots & & & & \\ & \ddots & \ddots & \ddots & & & \\ & & \ddots & \ddots & & c_{i-1} & \\ \hline & & & & c_{i-1} & b_i - \lambda & c_i \\ \hline 0 & & & & & c_i & b_{i+1} - \lambda \end{array} \right) =$$

$$\begin{aligned} &= (b_{i+1} - \lambda) f_{n-i}(\lambda) - c_i c_i f_{n-(i-1)}(\lambda) \\ &= f_{n-i-1}(\lambda). \end{aligned}$$

Da $f'_{n-i}(\lambda) = g_{n-i}(\lambda)$ für $i = 0, \ldots, n$, folgt die Behauptung. □

Bemerkung 4.23

Das Verfahren lässt sich einfach erweitern auf allgemeine Tridiagonalmatrizen $T = \text{tridiag}(a, b, c)$, falls im Algorithmus $|c_i|^2$ durch $a_i c_i$ ersetzt wird.

Beispiel 4.24
Sei T gegeben durch

$$T = \begin{pmatrix} 1 & 2 & 0 \\ 2 & 3 & 4 \\ 0 & 4 & 5 \end{pmatrix}.$$

Dann ist
$f_3(\lambda) = 1,\ f_2(\lambda) = 1 - \lambda,\ f_1(\lambda) = \lambda^2 - 4\lambda - 1,\ f_0(\lambda) = -\lambda^3 + 9\lambda^2 - 3\lambda - 21.$
Aus $f_0(\lambda) = 0$ ergeben sich die Eigenwerte.
Weiter ist $g_3(\lambda) = 0,\ g_2(\lambda) = -1,\ g_1(\lambda) = 2\lambda - 4,\ g_0(\lambda) = -3\lambda^2 + 18\lambda - 3.$

Bemerkung 4.25
Gilt $c_k \neq 0$ für $1 \leq k \leq n-1$, so haben die Polynome f_{n-i} i reelle einfache Nullstellen $\lambda_j^{(i)}, j = 1, \ldots, i$ und die Nullstellen von f_{n-i} trennen die Nullstellen von f_{n-i-1}. Daher bilden die Polynome $(f_n, \ldots, f_0)$ eine Sturmsche Kette.
Dadurch gilt für ein beliebiges Intervall $[a, b]$ mit $f_0(a) f_0(b) \neq 0$:

Ist n_a die Anzahl von Vorzeichenwechsel der Sequenz

$$(f_n(a), \ldots, f_0(a)$$

und n_b die Anzahl von Vorzeichenwechsel der Sequenz

$$(f_n(b), \ldots, f_0(b),$$

so besitzt f_0 auf $[a, b]$ genau $n_a - n_b$ Nullstellen. Mit Intervallhalbierung kann man dann alle Nullstellen in $[a, b]$ finden.

Motivaton 4.26 (Bestimmung von $\rho_A(\lambda)$ für unreduzierte Hessenberg-Matrizen)
Sei $H = (h_{ij})_{i,j} \in \mathbb{K}^{n \times n}$ unreduzierte Hessenberg-Matrix, d.h. $h_{i+1,i} \neq 0\ \forall i = 1, \ldots, n-1$.

Vorschlag von Hyman (1957)
Betrachte das lineare Gleichnugssystem

$$(H - \lambda I)x = -c\ e_1, c \in \mathbb{K}.$$

Setzte man $x_n = 1$, so kann man durch Rückwärtseinsetzen nacheinander $x_{n-1}, x_{n-2}, \ldots, x_1$ berechnen und schließlich c bestimmen.

Andererseits kann x_n mit der Cramerschen Regel berechnet werden durch:

$$1 = x_n = \frac{(-1)^n c h_{2,1} h_{3,2} \ldots h_{n,n-1}}{\det(H - \lambda I)}.$$

Also folgt

$$\rho_H(\lambda) = \det(H - \lambda I) = (-1)^n c h_{2,1} h_{3,2} \ldots h_{n,n-1}.$$

Mit diesem Vorgehen erhalten wir folgenden Algorithmus.

Satz 4.27 (Verfahren nach Hyman)
Sei $H \in \mathbb{K}^{n\times n}$ unreduzierte Hessenberg-Matrix. Für H mit charakteristischem Polynom

$$\rho_H(\lambda) = (-1)^n h_{2,1} h_{3,2} \dots h_{n,n-1} \varphi(\lambda)$$

liefert der Algorithmus

$$\begin{aligned}
&\underline{\text{Input:}}\ \lambda \in \mathbb{C}\\
&h_{1,0} = 1, x_n = 1, y_n = 0\\
&x_{n-i} = \tfrac{1}{h_{n-i+1,n-i}}\Big(\lambda x_{n-i+1} - \sum_{j=n-i+1}^{n} h_{n-i+1,j} x_j\Big)\\
&y_{n-i} = \tfrac{1}{h_{n-i+1,n-i}}\Big(x_{n-i+1} + \lambda y_{n-i+1} - \sum_{j=n-i+1}^{n} h_{n-1+i,j} y_j\Big)
\end{aligned}$$

für $1 \le i \le n$ das Ergebnis

$$\begin{aligned}
x_0 &= \varphi(\lambda)\ \ (= c),\\
y_0 &= \varphi'(\lambda).
\end{aligned}$$

Ist λ ein Eigenwert von H, so ist $x = (x_1, \dots, x_n)^T$ ein zugehöriger Eigenvektor.

Beweis: Folgt aus 4.26 und $y_{n-i} = x'_{n-i}$. □

4.6 Vektoriteration für partielle Eigenwertprobleme

Definition 4.28 (Vektoriteration nach von Mises)
Sei $A \in \mathbb{C}^{n\times n}$ eine diagonalisierbare Matrix mit dominaten Eigenwert λ_1, d.h.

$$(D-EW) \qquad |\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n|.$$

Dann erhält man eine Folge von Approximationen $\lambda_k, k = 1, 2, \ldots$ von λ_1 durch folgenden Algorithmus:

Input: $z^0 \in \mathbb{C}^n$ mit $\|z^0\| = 1$ und $l \in \{1, \ldots, n\}$

Für $k = 1, 2, \ldots$ berechne

$$\begin{aligned} \tilde{z}^k &= Az^{k-1} \\ z^k &= \frac{1}{\|\tilde{z}^k\|}\tilde{z}^k \\ \lambda^k &= \frac{(Az^k)_l}{z^k_l} \end{aligned}$$

Satz 4.29 (Konvergenz der Vektoriteration nach von Mises)
Sei $A \in \mathbb{C}^{n\times n}$ diagonalisierbar mit einer Basis $\{u_1, \ldots, u_n\}$ von normierten Eigenvektoren zu den Eigenwerten $\lambda_1, \ldots, \lambda_n$. Sei λ_1 ein dominanter Eigenwert von A, d.h. es gelte $(D-EW)$. Der Startwert $z^0 \in \mathbb{C}$ habe eine nichttriviale Komponente in Richtung u_1, d.h.

$$z^0 = \sum_{i=1}^{n} \alpha_i u_i \quad \text{mit } \alpha_1 \neq 0.$$

Dann gilt für z^k, λ^k aus der Vektoriteration nach von Mises (Def. 4.28), falls $u_{1,l} \neq 0$ ist:

1) $\|z^k - \sigma_k u_1\| = \mathcal{O}\Big(\big|\frac{\lambda_2}{\lambda_1}\big|^k\Big) \quad (k \to \infty)$ mit $\sigma_k := \frac{\lambda_1^k \alpha_1}{|\lambda_1^k \alpha_1|}$, also $|\sigma_k| = 1$.

2) $\lambda^k - \lambda_1 = \mathcal{O}\Big(\big|\frac{\lambda_2}{\lambda_1}\big|^k\Big) \quad (k \to \infty)$.

Beweis: Für die Iterierten z^k zeigt man durch vollständige Induktion, dass

$$z^k = \frac{A^k z^0}{\|A^k z^0\|}.$$

Mit $z^0 = \sum_{i=1}^{n} \alpha_i u_i$ folgt weiter

$$A^k z^0 = \sum_{i=1}^{n} \alpha_i \lambda_i^k u_i = \lambda_1^k \alpha_1 \Big(u_1 + \sum_{i=2}^{n} \frac{\alpha_i}{\alpha_1}\Big(\frac{\lambda_i}{\lambda_1}\Big)^k u_i\Big).$$

Wegen $(D-EW)$ folgt dann

$$A^k z^0 = \lambda_1^k \alpha_1 \Big(u_1 + \mathcal{O}\Big(\big|\frac{\lambda_2}{\lambda_1}\big|^k\Big) \quad (k \to \infty).$$

und hieraus

$$z^k = \frac{\lambda_1^k \alpha_1 \Big(u_1 + \mathcal{O}\Big(\big|\frac{\lambda_2}{\lambda_1}\big|^k\Big)\Big)}{|\lambda_1^k \alpha_1| \, \Big\|u_1 + \mathcal{O}\Big(\big|\frac{\lambda_2}{\lambda_1}\big|^k\Big)\Big\|} = \sigma_k u_1 + \mathcal{O}\Big(\big|\frac{\lambda_2}{\lambda_1}\big|^k\Big).$$

Also ist 1) gezeigt. Weiter gilt

$$\begin{aligned}
\lambda^k &= \frac{(Az^k)_l}{z_l^k} \\
&= \frac{(A^{k+1}z^0)_l}{\|A^k z^0\|} \frac{\|A^k z^0\|}{(A^k z^0)_l} \\
&= \frac{\lambda_1^{k+1}\Big(\alpha_1 u_{1,l} + \sum_{i=2}^{n} \alpha_i \Big(\frac{\lambda_i}{\lambda_1}\Big)^{k+1} u_{i,l}\Big)}{\lambda_1^k \Big(\alpha_1 u_{1,l} + \sum_{i=2}^{n} \alpha_i \Big(\frac{\lambda_i}{\lambda_1}\Big)^{k} u_{i,l}\Big)} \\
&\overset{u_{1,l}\neq 0}{=} \lambda_1 + \mathcal{O}\Big(\big|\frac{\lambda_2}{\lambda_1}\big|^k\Big) \quad (k \to \infty).
\end{aligned}$$

□

Bemerkung 4.30

1) Die Konvergenz der Vektoriteration nach von Mises ist also umso besser, je weiter der dominante Eigenwert λ_1 von den anderen Eigenwerten entfernt ist.

2) Varainte für $A \in \mathbb{C}^{n\times n}$ hermitesch:
Ist A hermitesch, so erhält man eine bessere Näherung von λ_1, wenn man zur Berechnung von λ^k den Rayleigh-Quotienten verwendet, d.h.

$$\lambda^k := R_A(z^k).$$

In diesem Fall gilt:

$$\lambda^k - \lambda_1 = \mathcal{O}\Big(\big|\frac{\lambda_2}{\lambda_1}\big|^{2k}\Big) \quad (k \to \infty).$$

Definition 4.31 (Inverse Iteration nach Wielandt)
Sei $A \in \mathbb{C}^{n\times n}$ eine diagonalisierbare Matrix mit

$$|\lambda_1| \geq |\lambda_2| \geq \ldots > |\lambda_n|,$$

so erhält man eine Näherung des Kehrwertes $\frac{1}{\lambda_n}$ des kleinsten Eigenwertes von A, indem man in der Vektoriteration nach von Mises A durch A^{-1} ersetzt.

Bemerkung 4.32 (Inverse Iteration mit Diagonal-Shift)
Ist $A \in \mathbb{C}^{n\times n}$ diagonalisierbar und $\mu \neq \lambda_i$ für alle $i = 1, \ldots, n$ eine gute Näherung eines Eigenwertes λ_j mit

$$|\lambda_j - \mu| \ll |\lambda_i - \mu| \qquad \forall i \neq j,$$

so kann die Näherung μ durch inverse Iteration für die Matrix $B := A - \mu I$ verbessert werden (Diagonal-Shift).

4.7 Das QR-Verfahren

Ziel: Verwende Ähnlichkeitstransformationen, um A sukzessive auf obere Dreiecksform zu transformieren:

$$\begin{aligned} A^0 &:= A, \\ A^i &:= T_i^{-1} A^{i-1} T_i, \quad i = 1, 2, ldots \end{aligned}$$

Beim QR-Verfahren wird T_i unitär gewählt!

Definition 4.33 (QR-Verfahren)
Sei $A \in \mathbb{C}^{n\times n}$. Dann ist das QR-Verfahren definiert durch:

$$\begin{aligned} A^0 &:= A, \\ A^i &:= Q^i R^i, && (QR\text{-Zerlegung von } A^i) \\ A^{i+1} &:= R^i Q^i. \end{aligned}$$

Hierbei ist Q^i unitär und R^i obere Dreiecksmatrix (siehe Numerik I, Kapitel 2.2). Wegen

$$A^{i+1} = R^i Q^i = (Q^i)^H A^i Q^i$$

sind alle Iterierten A^i ähnlich zu A.

Satz 4.34 (Konvergenz des QR-Verfahrens)
Die Eigenwerte von $A \in \mathbb{C}^{n\times n}$ seien betragsmäßig getrennt, d.h.

$$|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n|\,.$$

Dann gilt für die Diagonalelemente a_{jj}^i der Matrizen A^i des QR-Verfahrens 4.33:

$$\{\lim_{i\to\infty} a_{jj}^i |\ j = 1, \ldots, n\} = \{\lambda_1, \ldots, \lambda_n\}.$$

Weiter gilt

$$\lim_{i\to\infty} a_{jk}^i = 0 \text{ für } j > k.$$

Beweis: Siehe z.B. Stoer, Bulirsch [**?**]. □

Bemerkung 4.35
Da die Berechnung der QR-Zerlegung für allgemeine Matrizen A sehr aufwendig ist, bringt man in den Anwendungen A zunächst durch Housholder-Transformation auf Hessenberg-Form. Die Berechnung der QR-Zerlegung einer Hessenberg-Matrix kann dann mit Hilfe der sogenannten Givens-Rotation durchgeführt werden.

Algorithmus 4.36 (QR-Verfahren für Hessenberg-Matrizen)
Sei $A \in \mathbb{R}^{n\times n}$ eine Hessenberg-Matrix. Der folgende Algorithmus bestimmt in Schritt 1 (implizit) die QR-Zerlegung $A = QR$ und überschreibt in Schritt 2 A mit $A = RQ$.

1. Für $k = 1, \ldots, n-1$:

   Bestimme $c_k = \cos\Phi_k$ und $s_k = \sin\Phi_k$ mit

$$\begin{pmatrix} c_k & -s_k \\ s_k & c_k \end{pmatrix} \begin{pmatrix} a_{k,k} \\ a_{k+1,k} \end{pmatrix} = \begin{pmatrix} * \\ 0 \end{pmatrix}.$$

   Für $j = k, \ldots, n$ setze

$$\begin{pmatrix} a_{k,j} \\ a_{k+1,j} \end{pmatrix} := \begin{pmatrix} c_k & -s_k \\ s_k & c_k \end{pmatrix} \begin{pmatrix} a_{k,j} \\ a_{k+1,j} \end{pmatrix}.$$

2. Für $k = 1, \ldots, n-1$:
   Für $j = k, \ldots, n$ setze

$$(a_{j,k}, a_{j,k+1}) = (a_{j,k}, a_{j,k+1}) \begin{pmatrix} c_k & s_k \\ -s_k & c_k \end{pmatrix}.$$

Bemerkung 4.37 (LR-Verfahren)
Ersetzt man in Definition 4.33 die QR-Zerlegung durch die LR-Zerlegung, so erhält man das LR-Verfahren. Unter geeigneten Vorausetzungen gilt

$$\lim_{i\to\infty} A^i = \lim_{i\to\infty} R^i = \begin{pmatrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

und

$$\lim_{i\to\infty} L^i = I.$$

Nachteil:

1) Eventuell ist Pivotisierung notwendig und gilt nur $P^iA^i = L^iR^i$ mit einer Permutationsmatrix P^i, so ist die Konvergenz nicht gesichert.

2) L^i ist nicht unitär und das Verfahren konvergiert schlechter als das QR-Verfahren.

Voteil:

Für Hessenberg-Matrizen $A \in \mathbb{R}^{n\times n}$ ist die Berechnung der LR-Zerlegung mit dem Gauß-Algorithmus etwa doppelt so schnell wie die Berechnung der QR-Zerlegung.

Kapitel 5

Approximation

5.1 Allgemeine Approximation in normierten Räumen

Definition 5.1 (Beste Approximation/Proximum)
Sei $(X, \|\cdot\|)$ ein normierter Raum und $T \subset X$ eine beliebige Teilmenge. Zu einem $v \in X$ definiere

$$J_v(u) := \|v - u\| .$$

Dann heißt ein $u \in T$ beste Approximation *oder* Proximum *von v in T, g.d.w.*

$$J_v(u) = \inf_{w \in T} J_v(w).$$

Die Zahl $E_v(T) := \inf_{w \in T} J_v(w)$ heißt Minimalabstand von $v \in X$ zur Teilmenge T.

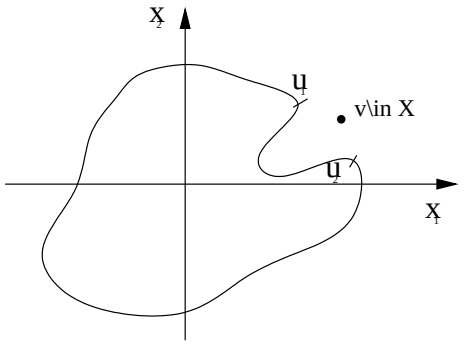

Abbildung 5.1: Proximum: Beispiel

Beispiel 5.2

1) Sei $X = \mathbb{R}^2$ und $\|\cdot\| = \|\cdot\|_2$ die euklidische Norm. Sei $T := \{x \in \mathbb{R}^2 \mid \|x\| \leq 1\}$. Dann existiert zu jedem $v \in X$ eine beste Approximation $u \in T$:
$\Longrightarrow$ Hier existiert zu jedem $v \in X$ genau ein Proximum.

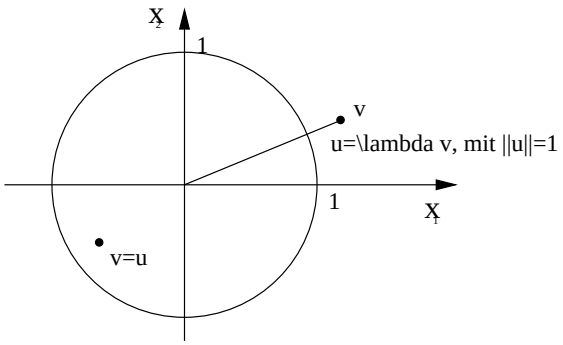

Abbildung 5.2: Proximum: Beispiel 1)

2) Sei $X = \mathbb{R}^2, \|\cdot\| = \|\cdot\|_2$ und $T := \{x \in \mathbb{R}^2 \mid \|x\| < 1\}$.
Ist $v \in X \setminus T$, so existiert keine beste Approximation $u \in T$ von V, denn $E_v(T) = \|v\| - 1$ und für alle $w \in T$ gilt $\|w - v\| > \|v\| - 1$
$\Longrightarrow$ Hier existiert für $x \in X \setminus T$ kein Proximum.

3) Sei $X = \mathbb{R}^2, \|\cdot\| = \|\cdot\|_\infty$, d.h. $\|x\|_\infty = \max\{|x_1|, |x_2|\}$ und $T := \{(x_1, 0) \mid x_1 \in \mathbb{R}\}$. Sei $v = (0, 1) \in X$. Dann gilt $u \in [-1, 1] \times \{0\}$ $\Longrightarrow$ u ist Proximum, da $J_v(u) = 1 = E_v(T)$
$\Longrightarrow$ Hier existieren unendlich viele beste Approximationen.

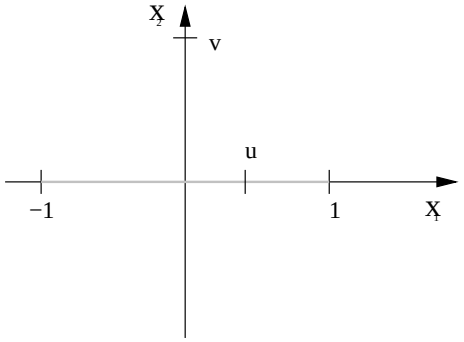

Abbildung 5.3: Proximum: Beispiel 3)

Satz 5.3 (Existenz eines Proximums)
Sei $T \subset X$ eine kompakte Teilmenge. Dann existiert zu jedem $v \in X$ ein Proximum $u \in T$.

Beweis: Sei $(u_n)_{n\in\mathbb{N}}$ eine Minimalfolge in T für $v \in X$, d.h. $\lim_{n \longrightarrow \infty} J_v(u_n) = E_v(T)$. Da T kompakt ist, enthält $(u_n)_{n\in\mathbb{N}}$ eine Teilfolge, die in T konvergiert, d.h. $\exists u \in T$ mit

$$\lim_{j \longrightarrow \infty} u_{n_j} = u.$$

Zu zeigen: u ist Proximum, d.h. $J_v(u) = E_v(T)$.

Es gilt:

$$\begin{array}{rlll} \|v-u\| \leq & \|v-u_{n_j}\| + & \|u_{n_j}-u\| \\ & \downarrow (j\to\infty) & \downarrow (j\to\infty) \\ & E_v(T) & 0 \end{array}$$

$\Longrightarrow \quad \|v-u\| \leq E_v(T)$.

Da $E_v(T) = \inf_{w\in T} J_v(w) = \inf_{w\in T} \|v-w\|$, folgt:

$$J_v(u) = \|v-u\| = E_v(T).$$

□

Definition 5.4 (Konvexe Teilmengen)
$T \subset X$ heißt konvex, g.d.w.

$$K_{u_1,u_2} := \{\lambda u_1 + (1-\lambda)u_2 \,|\, \lambda \in (0,1)\} \subset T$$

für alle $u_1, u_2 \in T$.

$T \subset X$ heißt streng konvex, g.d.w.

$$K_{u_1,u_2} \subset \mathring{T}\ \forall u_1, u_2 \in T$$

mit $u_1 \neq u_2$ (vgl. Abbildung 5.4).

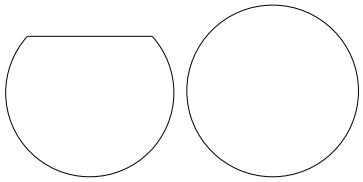

Abbildung 5.4: Konvexe und streng konvexe Mengen.

Satz 5.5 (Eindeutigkeit von Proxima)
Sei $T \subset X$ kompakt und streng konvexe Teilmenge des normierten Raumes X. Dann gibt es zu jedem $v \in X$ genau ein Proximum $u \in T$.

Beweis: Die Existenz ist klar nach Satz 5.3.
Eindeutigkeit: Seien u_1, u_2 mit $u_1 \neq u_2$ Proxima von $v \in X$ in T. Dann gilt:

$$\left\| \frac{1}{2}(u_1 + u_2) - v \right\| \leq \frac{1}{2} \underbrace{\|u_1 - v\|}_{=E_v(T)} + \frac{1}{2} \underbrace{\|u_2 - v\|}_{=E_v(T)} = E_V(T).$$

T konvex $\implies \left\| \frac{1}{2}(u_1 + u_2) - v \right\| = E_v(T)$.

Da T streng konvex ist, folgt $\frac{1}{2}(u_1 + u_2) \in \mathring{T}$. Also existiert ein $\tilde{\lambda} \in (0, 1)$, so dass

$$\tilde{u} = \frac{1}{2}(u_1 + u_2) + \tilde{\lambda}(v - \frac{1}{2}(u_1 + u_2)) \in T$$

ist. Dann gilt:

$$\begin{aligned} \|\tilde{u} - v\| &= \left\| \tfrac{1}{2}(1 - \tilde{\lambda})(u_1 + u_2) - (1 - \tilde{\lambda})v \right\| \\ &= (1 - \tilde{\lambda}) \left\| \tfrac{1}{2}(u_1 + u_2) - v \right\| \\ &= (1 - \tilde{\lambda}) E_v(T) \\ &< E_v(T). \end{aligned}$$

Dies ist ein Widerspruch zur Definition von $E_v(T) \implies u_1 = u_2$. □

Für Anwendungen ist vor allem der Fall wichtig, dass T ein endlich dimensionaler Teilraum von X ist.

Satz 5.6 (Fundamentalsatz der Approximationstheorie in normierten Räumen)
Sei $T \subset X$ ein endlichdimensionaler linearer Teilraum des normierten Vektorraums X. Dann existiert zu jedem $v \in X$ ein Proximum $u \in T$.

Beweis: Sei $(u_n)_{n \in \mathbb{N}}$ Minimalfolge für $v \in X$.
Wir zeigen zunächst: $(u_n)_{n \in \mathbb{N}}$ ist beschränkt.

Es gilt:

$$E_v(T) \leq \|v - u_n\| \leq E_v(T) + 1 \ \ \forall n \geq N.$$

Also ist $\|u_n\| \leq \|v - u_n\| + \|v\| \leq E_v(T) + 1 + \|v\| =: K_1 \ \ \forall n \geq N$.

Sei nun $K_2 \geq \|u_n\|$ für $n < N$ und setzte $K = \max\{K_1, K_2\}$, so folgt $\|u_n\| \leq K \ \forall n \in \mathbb{N}$.

Da T endlich dimensionaler linearer Teilraum ist, existiert also eine Teilfolge $(u_{n_j})_{j \in \mathbb{N}}$, die gegen ein $u \in T$ konvergiert.
Analog zum Beweis von Satz 5.3 zeigt man, dass u ein Proximum ist. □

Definition 5.7 (Streng normierter Raum)
Sei $(X, \|\cdot\|)$ ein normierter Raum. $\|\cdot\|$ *heißt* strenge Norm *und* X streng normiert*, g.d.w.*

$$(\|f + g\| = \|f\| + \|g\| \ \textit{für}\ f, g \in X \ \textit{mit}\ f, g \neq 0) \implies (\exists \lambda \in \mathbb{C} \ \textit{mit}\ g = \lambda f)\,.$$

Satz 5.8 (Eindeutigkeit in streng normierten Räumen)
Ist $(X, \|\cdot\|)$ ein streng normierter Raum und $T \subset X$ ein endlichdimensionaler linearer Teilraum, so existiert zu jedem $v \in X$ genau ein Proximum $u \in T$.

Beweis: Ist $v \in T$, so ist $u = v$ das eindeutige Proximum.

Sei also $v \in X \setminus T$. Sind u_1, u_2 verschiedene Proxima zu v, so gilt wie in Beweis von Satz 5.5:

$$E_v(T) \leq \left\| v - \frac{1}{2}(u_1 + u_2) \right\| \leq \frac{1}{2}\|v - u_1\| + \frac{1}{2}\|v - u_2\| = E_v(T)$$

also $\|(v - u_1) + (v - u_2)\| = \|v - u_1\| + \|v - u_2\|$

Da $\|\cdot\|$ strenge Norm ist, $\exists \lambda \in \mathbb{C}$ mit $v - u_1 = \lambda(v - u_2)$
$\Longrightarrow \quad (1 - \lambda)v = u_1 - \lambda u_2$.
Da $v \notin T$, folgt $\lambda = 1$ und somit $0 = u_1 - u_2$. Dies ist ein Widerspruch zur Annahme. □

5.2 Der Satz von Weierstraß: Approximation durch Polynome

Motivation 5.9
Bei der Approximation durch Polynome wurde in der Analysis immer von der Regularität der Funktion f gebrauch gemacht. So z.B. bei der Approximation von f durch die Taylorreihe.
In diesem Kapitel stellen wir uns die Frage, ob eine beliebig gute Polynomapproximation auch für Funktiontn $f \in C([a,b])$ möglich ist.

Satz 5.10 (Approximationssatz von Weierstraß)
Sei $X = C([a,b])$ und $\|\cdot\| = \|\cdot\|_\infty$ für $|a|, |b| < \infty$.
Dann gibt es zu jedem $f \in X$ und $\varepsilon > 0$ ein $n \in \mathbb{N}$ und ein $p \in \mathbb{P}_n$, so dass $\|f - p\|_\infty < \varepsilon$ ist.

Beweis: (Konstruktiv!)
Ohne Einschränkung sei $[a,b] = [0,1]$ (sonst Transformation).
Wir zeigen, dass die Folge der Bernstein-Polynome

$$(B_n f)(x) := \sum_{i=0}^{n} f(\frac{i}{n}) \binom{n}{i} x^i (1-x)^{n-i}$$

für $n \longrightarrow \infty$ auf $[0,1]$ gleichmäßig gegen f konvergieren.

Definiere $q_{ni} := \binom{n}{i} x^i (1-x)^{n-i}$, dann ist

$$1 = (x + (1-x))^n = \sum_{i=0}^{n} \binom{n}{i} x^i (1-x)^{n-i} = \sum_{i=0}^{n} q_{ni}.$$

Also folgt:

$$f(x) - (B_n f)(x) = \sum_{i=0}^{n} (f(x) - f(\frac{i}{n})) q_{ni}(x)$$

$$\implies |f(x) - (B_n f)(x)| \leq \sum_{i=0}^{n} \left| f(x) - f(\frac{i}{n}) \right| q_{ni}(x) \quad \forall x \in [0,1].$$

Da f gleichmäßig stetig ist, existiert zu jedem $\varepsilon > 0$ ein δ, so dass $\forall x, i$

$$\left| x - \frac{i}{n} \right| < \delta \implies \left| f(x) - f(\frac{i}{n}) \right| < \frac{\varepsilon}{2}.$$

Sei $x \in [0,1]$. Setze

$$N_< := \{i \in \{0, \ldots, n\} \mid \left| x - \frac{i}{n} \right| < \delta\},$$

$$N_\geq <:= \{i \in \{0, \ldots, n\} \mid \left| x - \frac{i}{n} \right| \geq \delta\}.$$

Wir erhalten:

$$\sum_{i \in N_<} \left| f(x) - f(\frac{i}{n}) \right| q_{ni}(x) \leq \frac{\varepsilon}{2} \sum_{i \in N_<} q_{ni}(x) \leq \frac{\varepsilon}{2}.$$

und

$$\begin{aligned}\sum_{i\in N_{\geq}} \left|f(x)-f(\tfrac{i}{n})\right| q_{ni}(x) &\leq \sum_{i\in N_{\geq}} \left|f(x)-f(\tfrac{i}{n})\right| q_{ni}(x)\frac{(x-\frac{i}{n})^2}{\delta^2}\\ &\leq \frac{2\,\|f\|_\infty}{\delta^2}\sum_{i\in N_{\geq}} q_{ni}(x)(x-\frac{i}{n})^2\\ &\leq \frac{2\,\|f\|_\infty}{\delta^2}\sum_{i=0}^{n} q_{ni}(x)(x^2-2x\frac{i}{n}+(\frac{i}{n})^2).\end{aligned}$$

Es ist

1) $\sum\limits_{i=0}^{n} q_{ni}(x)x^2 = x^2$,

2) $\sum\limits_{i=0}^{n} q_{ni}(x)2x\frac{i}{n} = 2x\cdot x\underbrace{\sum_{i=1}^{n}\binom{n-1}{i-1}x^{i-1}(1-x)^{(n-1)-(i-1)}}_{=1} = 2x^2$,

3)

$$\begin{aligned}\sum_{i=0}^{n} q_{ni}(x)(\frac{i}{n})^2 &= \frac{x}{n}\sum_{i=1}^{n}(i-1)\binom{n-1}{i-1}x^{i-1}(1-x)^{(n-1)-(i-1)}+\frac{x}{n}\\ &= \frac{x^2}{n}(n-1)\underbrace{\sum_{i=2}^{n}\binom{n-2}{i-2}x^{i-2}(1-x)^{(n-2)-(i-2)}}_{=1}+\frac{x}{n}\\ &= x^2(1-\frac{1}{n})+\frac{x}{n} = x^2+\frac{x}{n}(1-x).\end{aligned}$$

Mit 1), 2) und 3) folgt:

$$\begin{aligned}\sum_{i\in N_{\geq}} \left|f(x)-f(\tfrac{i}{n})\right| q_{ni}(x) &\leq \frac{2\,\|f\|_\infty}{\delta^2}(\underbrace{x^2-2x+x^2}_{=0}+\underbrace{\frac{x}{n}(1-x)}_{\leq\frac{1}{4n}}\\ &\leq \frac{2\,\|f\|_\infty}{\delta^2}\frac{1}{4n}\\ &< \frac{\varepsilon}{2}\quad \text{f"ur } n>\frac{M}{\delta^2\varepsilon}.\end{aligned}$$

Insgesamt folgt also

$$|f(x)-(B_nf)(x)| \leq \frac{\varepsilon}{2}+\frac{\varepsilon}{2} = \varepsilon\ \ \forall x\in[0,1].$$

□

Bemerkung 5.11
Satz 5.10 zeigt, dass sich $f\in C([a,b])$ in folgender Weise entwickeln lässt:

$$f(x) = (B_1f)(x)+[(B_2f)(x)-(B_1f)(x)]+\ldots+[(B_nf)(x)-(B_{n-1}f)(x)]+\ldots$$

Die Reihe konvergiert gleichmäßig, lässt sich aber im allgemeinen nicht zu einer Potenzreihe umordnen!

Satz 5.12 (Fehlerabschätzung für die Approximation mit Bernsteinpolynomen)
Seien die Voraussetzungen von Satz 5.10 erfüllt und sei

$$w_f(\delta) := \sup_{|x'-x''|\le\delta, x', x''\in[a,b]} \big|f(x') - f(x'')\big|$$

das Stetigkeitsmodul von f bezgl. δ. Dann gilt die Abschätung:

$$|f(x) - (B_n f)(x)| \le \frac{5}{4} w_f(\frac{1}{\sqrt{n}}).$$

Beweis: Sei $\lambda = \lambda(x', x'', \delta) := \left\lceil \frac{|x'-x''|}{\delta} \right\rceil$ das größte Ganze.
Mit der Definition von $w_f(\delta)$ folgt:

$$\delta_1 \le \delta_2 \implies w_f(\delta_1) \le w_f(\delta_2).$$

Damit folgt:

$$\big|f(x') - f(x'')\big| \le w_f(\big|x' - x''\big|) \le w_f((\lambda+1)\delta).$$

Aus $w_f(\mu\delta) \le \mu\, w_f(\delta)$ für $\mu \in \mathbb{N}$ folgt:

$$\big|f(x') - f(x'')\big| \le (\lambda+1) w_f(\delta).$$

Setze $N_{\ge} := \{i \in \{0,\dots,n\} \,|\, \lambda(x, \frac{i}{n}, \delta) \ge 1\}$ und $N_<$ entsprechend. Jetzt gehen wir analog zum Beweis von Satz 5.10 vor:

$$\begin{aligned} |f(x) - (B_n f)(x)| &\le \sum_{i=0}^{n} \left| f(x) - f(\frac{i}{n}) \right| q_{ni} \\ &\le w_f(\delta) \sum_{i=0}^{n} (1 + \lambda(x, \frac{i}{n}, \delta)) q_{ni}(x). \end{aligned}$$

Da $\lambda(x, \frac{i}{n}, \delta) = 0$ für alle $i \in N_<$ gilt, folgt weiter:

$$\begin{aligned} |f(x) - (B_n f)(x)| &\le w_f(\delta) \left(1 + \sum_{i\in N_{\ge}} \lambda(x, \frac{i}{n}, \delta) q_{ni}(x)\right) \\ &\le w_f(\delta) \left(1 + \frac{1}{\delta} \sum_{i\in N_{\ge}} \left| x - \frac{i}{n} \right| q_{ni}(x)\right) \\ &\overset{\frac{|x-\frac{i}{n}|}{\delta} \ge 1 \ \forall i\in N_{\ge}}{\le} w_f(\delta) \left(1 + \frac{1}{\delta^2} \sum_{i\in N_{\ge}} (x - \frac{i}{n})^2 q_{ni}(x)\right) \\ &\overset{\text{Analog zum Beweis von 5.10}}{\le} w_f(\delta) \left(1 + \frac{1}{4n\delta^2}\right). \end{aligned}$$

Wählen wir $\delta = \frac{1}{\sqrt{n}}$, so folgt die Abschätzung des Satzes. □

Bemerkung 5.13

1) Abhängig vom Stetigkeitsmodul kann die Schranke in Satz 5.12 beliebig langsam konvergieren. Bei höheren Anforderungen an die Stetigkeit von f kann andererseits eine schnellere Konvergenz erwartet werden.

2) In der Praxis ist die Approximation durch Bernstein-Polynome nicht von Bedeutung. Im nächsten Abschnitt werden wir wirkungsvollere Verfahren kennen lernen!

5.3 Gleichmäßige Approximation / Tschebyschev Approximation

Motivation 5.14
In §5.2 haben wir gesehen, dass sich jede stetige Funktion $f \in C^0([a,b])$ durch Polynome $p_n \in \mathbb{P}_n$ approximieren lassen. Die Frage, welches Polynom $p \in \mathbb{P}_n$ das Proximum zu f in $T = \mathbb{P}_n$ ist, wurde jedoch nicht beantwortet. Ist $\|\cdot\| = \|\cdot\|_\infty$, so beantworten wir diese Frage in diesem Abschnitt. $\|\cdot\|_\infty$ wird auch Tschebyschev-Norm genannt.

Definition 5.15 (Best Approximating Polynomial (BAP))
Sei $X = C^0(I)$, $I \subset \mathbb{R}$ beschränktes Intervall ausgestattet mit der ∞-Norm $\|\cdot\| = \|\cdot\|_\infty$.

Dann heißt $p_n \in \mathbb{P}_n$ Best Approximating Polynomial (BAP) vom Grad $\leq n$ von $f \in C^0(I)$, g.d.w. p_n Proximum von f in $\mathbb{P}_n$ ist.

Bemerkung 5.16
1) Der Fundamentalsatz der Approximationstheorie 5.6 liefert die Existenz eines (BAP), da $\mathbb{P}_n$ ein endlichdimensionaler linearer Teilraum von $C^0(I)$ ist.

2) Eindeutigkeit des (BAP) können wir zunächst nicht erwarten, da $\|\cdot\|_\infty$ keine strenge Norm ist.

Satz 5.17 (Charakterisierung von BAP)
Sei $f \in C^0(I), I = [a,b]$ beschränkt und $p \in \mathbb{P}_n$. Es gebe $n+2$ Punkte $a \leq x_0 < x_1 < \ldots < x_n < x_{n+1} \leq b$, so dass

1) $|f(x_i) - p(x_i)| = J_f(p) = \|f - p\|_\infty$ für $i = 0, \ldots, n+1$,

2) $f(x_{i+1}) - p(x_{i+1}) = -(f(x_i) - p(x_i))$ für $i = 0, \ldots, n$.

Dann ist p BAP vom Grad $\leq n$ an f.

Beweis: Sei $p^* \in \mathbb{P}_n$ und $M := \{x \in I \mid |f(x) - p^*(x)| = J_f(p^*)\}$.
Ist p^* kein BAP, so existiert ein $\bar{p} \in \mathbb{P}_n, \bar{p} \neq 0$, so dass $p^* + \bar{p}$ BAP ist (nach Satz 5.6). Dann gilt:

$$|(f(x) - p^*(x)) - \bar{p}(x)| = |f(x) - (p^*(x) + \bar{p}(x))| < |f(x) - p^*(x)| \quad \forall x \in M$$

$$\Longrightarrow \quad (f(x) - p^*(x))\bar{p}(x) > 0 \ \ \forall x \in M \qquad (*)$$

Sei nun $p \in \mathbb{P}_n$ und es gelten die Voraussetzungen 1) und 2).
Dann kann es kein $\bar{p} \neq 0, \bar{p} \in \mathbb{P}_n$ geben, so dass $(*)$ erfüllt ist. Denn dazu müsste $\bar{p}$ in $[a,b]$ mindestens $(n+1)$-mal das Vorzeichen wechseln (wegen 2)), also mindestens $n+1$ Nullstellen besitzen. Nach dem Fundamentalsatz der Algebra ist das nicht möglich.
Da es zu p kein $\bar{p}$ gibt, so dass $(*)$ gilt, muß p bereits BAP sein.

□

Satz 5.18 (Eindeutigkeit eines Proximums in $\mathbb{P}_n$)
Sei $U := \mathbb{P}_n$ Unterraum von $C^0([a,b])$. Dann ist das Proximum $p \in U$ an ein $f \in C^0([a,b])$ eindeutig bestimmt.

Beweis: Seien p_1 und p_2 Proxima aus U an $f \in C^0([a,b])$.
Dann ist auch $\frac{1}{2}(p_1+p_2)$ Proximum (Satz 5.5) und nach dem Characterisierungssatz 5.17 exisitert eine sogenannte Alternante der Länge $n+2$, , d.h.

$$f(x_i) - \frac{1}{2}(p_1(x_i)+p_2(x_i)) = \sigma(-1)^i E_f(U).$$

Also ist $\frac{1}{2}(f(x_i)-p_1(x_i)) + \frac{1}{2}(f(x_i)-p_2(x_i)) = \sigma(-1)^i E_f(U)$.
Da p_1, p_2 Proxima sind, gilt $|f(x_i)-p_k(x_i)| \leq E_f(U)$ $k=1,2$ und $i=0,\ldots,n+1$. Daraus folgt

$$f(x_i)-p_1(x_i) = f(x_i)-p_2(x_i) \;\; \forall i = 0,\ldots,n+1$$

$$\Longrightarrow \;\; p_1(x_i) = p_2(x_i) \;\; \forall i = 0,\ldots,n+1.$$

Da U Polynome vom maximalen Grad n sind, folgt hieraus $p_1 - p_2 \equiv 0$, da $p_1 - p_2$ mindestens $n+2$ Nullstellen hat.

□

Der Charakterisierungssatz 5.17 bildet die Grundlage für ein Verfahren zur Approximation des BAP zu einem $f \in C^0([a,b])$.

(Austauschalgorithmus von Remez) 5.19
Sei $f \in C^0([a,b])$. Dann erzeugt der folgende Algorithmus von Remez eine Folge von Polynomen $p^{(k)} \in \mathbb{P}_n$ mit

$$\lim_{k \longrightarrow \infty} \left\| p^{(k)} - p_n \right\|_\infty = 0,$$

wobei p_n das eindeutig bestimmte BAP von f in $\mathbb{P}_n$ ist.
Als Abbruchktiterium kann die Bedingung

$$\Delta - \delta \leq \varepsilon$$

gewählt werden, wobei δ, Δ bezüglich $p^{(k)}$ wie folgt definiert sind

$$\delta := \min_{i=0,\ldots,n+1} |f(x_i)-p(x_i)|,$$

$$\Delta := \max_{i=0,\ldots,n+1} |f(x_i)-p(x_i)|.$$

(Tatsächlich kann gezeigt werden, dass gilt $\delta \leq E_f(U) \leq \Delta$.)
Remez-Algorithmus:

Input: Startzerlegung $I^{(0)}$ von $[a,b]$, d.h.

$$I^{(0)} := \{x_0, \ldots, x_{n+1}\}$$

mit $a \leq x_0 < x_1 < \ldots < x_{n+1} \leq b$.

Iteration für $k = 0, 1, 2, \ldots$

1. Schritt: Zu $I^{(k)}$ bestimme Polynom $p^{(k)} \in \mathbb{P}_n$, so dass $I^{(k)}$ Alternante für $f - p^{(k)}$ ist und für alle $x_i \in I^{(k)}$ gilt:

$$\left| f(x_i) - p^{(k)}(x_i) \right| = \left| \xi^{(k)} \right|$$

mit einer Konstante $\xi^{(k)} \in \mathbb{R}$.
Setze man $p^{(k)}(x) := \sum_{j=0}^{n} \alpha_j^{(k)} x^j$, so führen diese Forderungen auf das Gleichungssystem:

$$(-1)^i \xi^{(k)} + \sum_{j=0}^{n} \alpha_j^{(k)} x_i^j = f(x_i)\,, \quad i = 0, \ldots, n+1$$

für die Unbekannten $\xi^{(k)}, \alpha_0^{(k)}, \ldots, \alpha_n^{(k)}$.

2. Schritt: Setze $r^{(k)}(x) := f(x) - p^{(k)}(x)$ und bestimme $y \in [a, b]$ mit der Eigenschaft:

$$r^{(k)}(y) = \max_{x \in [a,b]} r^{(k)}(x).$$

Ersetze einen Punkt $x_i \in I^{(k)}$ durch y und erhalte $I^{(k+1)}$ gemäß folgender Austauschvorschrift:

1. Fall: $x_j < y < x_{j+1}$ für eine $j \in \{0, \ldots, n\}$.
   Falls $\operatorname{sign}(r^{(k)}(x_j)) = \operatorname{sign}(r^{(k)}(y))$, ersetze x_j durch y, sonst erzetze x_{j+1} durch y.

2. Fall: $y < x_0$.
   Falls $\operatorname{sign}(r^{(k)}(x)) = \operatorname{sign}(r^{(k)}(y))$, ersetze x_0 durch y, sonst ersetze x_{n+1} durch y.

3. Fall: $y > x_{n+1}$.
   Falls $\operatorname{sign}(r^{(k)}(x_{n+1})) = \operatorname{sign}(r^{(k)}(y))$, ersetze x_{n+1} durch y, sonst ersetze x_0 durch y.

Bemerkung: Gilt $y = x_i$ für ein $x_i \in I^{(k)}$, so ist $p^{(k)}$ bereits BAP. (Charakterisierungssatz 5.17)

Beweis: Skizze: Man zeigt:

1) Das Gleichungssystem in Schritt 1 ist stets eindeutig lösbar.

2) Es gilt $\xi^{(k+1)} > \xi^{(k)} \;\; \forall k = 0, 1, 2, \ldots$.

3) $\lim_{m \longrightarrow \infty} I^{(k_m)} = \{\bar{x}_0, \ldots, \bar{x}_{n+1}\}$ für eine Teilfolge.

4) $\lim_{m \longrightarrow \infty} p^{(k_m)} = p$ für diese Teilfolge.

5) p ist BAP an f.

□

Im folgenden wollen wir der Frage nachgehen, wie für $f(x) = x^n$ in $[-1, 1]$ das BAP in $\mathbb{P}_{n-1}$ aussieht. Dies führt uns zu den Tschebyschev Polynomen 1. Art.

Definition 5.20 (Tschebyschev Polynome 1. Art)
Durch die Rekursion

$$T_0(x) = 1, T_1(x) = x$$

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x), n \geq 1$$

werden die sogenannten Tschebyschev Polynome 1. Art *definiert.*

Satz 5.21
Die Tschebyschev-Polynome 1. Art haben die Darstellung $T_n(x) = \cos(n \arccos(x))$, $x \in [-1, 1], n \in \mathbb{N}$. Es gelten die Eigenschaften

1) $|T_n(x)| < 1 \ \ \forall x \in [-1, 1]$,

2) $T_n(x)$ hat in $[-1, 1]$ die Extrempunkte

$$x_i^{(n)} = \cos(\frac{i\pi}{n}), T_n(x_i^{(n)}) = (-1)^i \ \ i = 0, \ldots, n,$$

3) T_n hat n einfache Nullstellen in $[-1, 1]$

$$\tilde{x}_i^{(n)} = \cos(\frac{2i-1}{2n}\pi) \ \ i = 1, \ldots, n,$$

4) Zwischen je zwei Nullstellen von T_{n+1} liegt eine Nullstelle von T_n.

Beweis: (ohne Beweis) □

Satz 5.22
Sei $f(x) = x^n \in C^0([-1, 1])$. Dann ist

$$p_{n-1} := x^n - \frac{1}{2^{n-1}} T_n(x)$$

BAP zu f in $\mathbb{P}_{n-1}$.

Beweis: 1.) Zeige $p_{n-1} \in \mathbb{P}_{n-1}$.
Unter Verwendung der Rekursionsformel zeigt man induktiv, dass $T_n(x)$ die Form

$$T_n(x) = 2^{n-1}x^n + \sum_{i=1}^{n-1} \alpha_i x^i$$

hat. Dann folgt:

$$p_{n-1} = x^n - \frac{1}{2^{n-1}} T_n(x) = -\frac{1}{2^{n-1}} \sum_{i=1}^{n-1} \alpha_i x^i \implies p_{n-1} \in \mathbb{P}_{n-1}.$$

2.) Zeige p_{n-1} ist BAP zu f.
Es gilt $x^n - p_{n-1}(x) = \frac{1}{2^{n-1}} T_n(x)$. Nach Satz 5.21 (2) ist $\{x_i^{(n)} := cos(\frac{i\pi}{n}) \,|\, i = 0, \ldots, n\}$ eine Alternante der Länge $(n-1)+2$ für $p_{n-1}(x)$ und es gilt nach 5.21 1), 2)

$$\left| \frac{1}{2^{n-1}} T_n(x_n^{(i)}) \right| = \frac{1}{2^{n-1}} = \max_{x \in [-1,1]} \left| \frac{1}{2^{n-1}} T_n(x) \right|.$$

Nach dem Charakterisierungssatz 5.17 muß p_{n-1} BAP zu f sein. □

Satz 5.23
Die Tschebyschev-Polynome bilden bezüglich der Gewichtsfunktion $w(x) := \frac{1}{\sqrt{1-x^2}}$ ein Orthogonalsystem. Es gilt:

$$\int_{-1}^{1} T_n(x) T_m(x) w(x) dx = \begin{cases} \pi & f"ur \quad n = m = 0 \\ \frac{\pi}{2} & \quad\quad\; n = m \neq 0 \\ 0 & \quad\quad\; n \neq m \end{cases}.$$

(ohne Beweis)

Definition 5.24 (Tschebyschev-Entwicklung)
Sei f stetig in $[-1,1]$. Dann heißen

$$a_k(f) := \frac{2}{\pi} \int_{-1}^{1} f(x) T_k(x) \frac{1}{\sqrt{1-x^2}} dx, \; k = 0, 1, 2, \ldots$$

Tschebyschev-Koeffizienten *von f und die formal gebildete Reihe*

$$S_f^n(x) = \frac{a_0(f)}{2} + \sum_{k=1}^{n} a_k(f) \cdot T_k(x)$$

heißt Tschebyschev-Entwicklung *von f.*

Entwicklungssatz 5.25
Sei $f \in C^2([-1,1])$. Dann konvergiert die Tschebyscheventwicklung für $x \in [-1,1]$ gleichmäßig gegen f, d.h.

$$S_f^\infty := \lim_{n \longrightarrow \infty} S_f^n = f.$$

Für die Tschebyschev-Koeffinzienten gilt die Abschätzung:

$$|a_k(f)| \leq \frac{c}{k^2}, \quad k = 1, 2, 3, \ldots.$$

Beweis: (ohne Beweis) □

Folgerung 5.26
Die Koeffizientenabschätzung in Satz 5.25 zeigt, dass für $f \in C^2([-1,1])$ eine gute N durch S_f^n gegeben ist.
Diese Möglichkeit der Approximation bietet sich dann an, wenn die Koeffizienten $a_k(f)$ einfach zu berechnen sind.

Index