Prof. Dr. Irwin Yousept Universität Duisburg-Essen

Numerische Mathematik 1

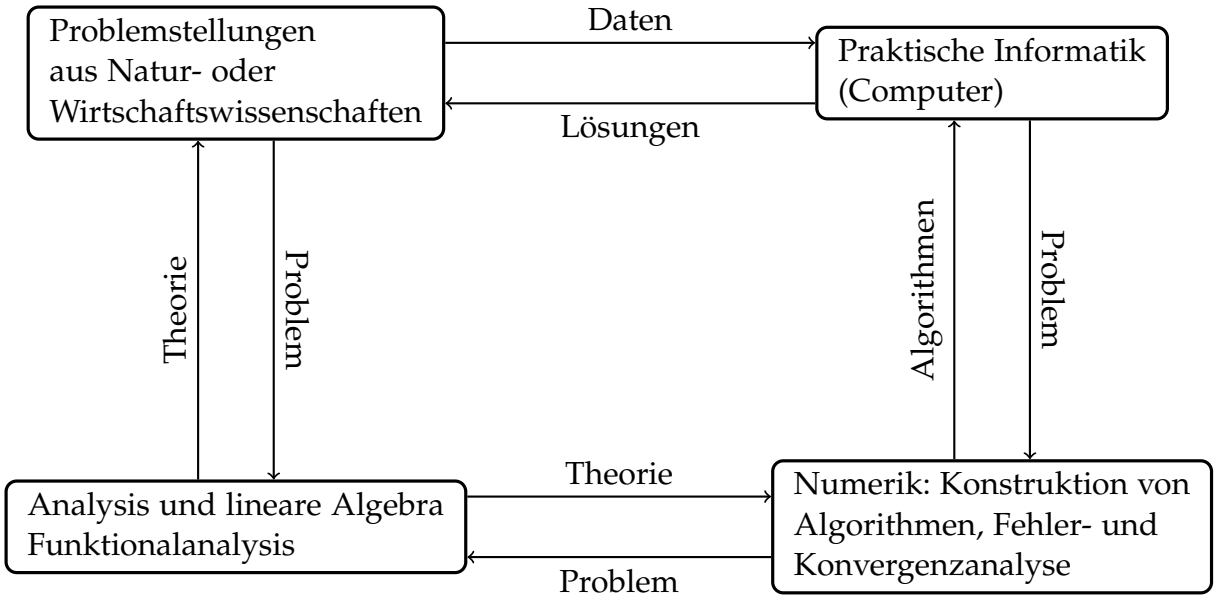
Typeset und Layout: Roman Händler
Fassung vom 4. März 2017

Inhaltsverzeichnis

Kapitel 1

Einleitung

Die Aufgabe der Numerik ist die Analyse und Konstruktion von Lösungsverfahren mathematischer Aufgaben. Das folgende Diagramm gibt einen kurzen Einblick in die Arbeitsweise der Numerik und zeigt den Bezug zu anderen wissenschaftlichen Bereichen auf.



Kapitel 2

Grundlagen

2.1 Zahlendarstellung und Rundungsfehler

Von einem Rechner bzw. Taschenrechner können nur endlich viele verschiedene Zahlen dargestellt und gespeichert werden, diese heißen *Maschinenzahlen*. Die Menge aller Maschinenzahlen wird mit M bezeichnet, also

$$\mathrm{M} = \text{Menge aller Maschinenzahlen.}$$

Zur Speicherung von Maschinenzahlen verwenden die Rechner die *Gleitpunktdarstellung*:

$$\pm\, a_0.a_1a_2\ldots a_{t-1}\cdot b^e,$$

wobei

$t \in \mathbb{N}$: Mantissenlänge,
$a_i \in \mathbb{N}\cup\{0\} : 0 \le a_i \le b-1,\ a_0 \neq 0$
$b \in \mathbb{N}$: die Basis des Zahlensystems,
$e \in \mathbb{Z}$: der Exponent mit $e \in [-n, n]$ mit $n \in \mathbb{N}$ die Exponentenlänge.

Eine Ausnahme ist die Zahl Null, die als Maschinenzahl angesehen wird und für die alle $a_i = 0$ sind.

Bemerkung 2.1. Im Falle $b = 2$ spricht man vom Binärsystem. Im Falle $b = 10$ spricht man vom Dezimalsystem.

Beispiel 2.2. Betrachte $b = 2$, $t = 3$, $e \in [-4, 4]$. Es gilt

Zahl	Gleitpunktdarstellung auf dem Rechner
4	$+1.00 \cdot 2^2$
-0.0625	$-1.00 \cdot 2^{-4}$
3.5	$+1.11 \cdot 2^1$
24	$+1.10 \cdot 2^4$

Beispiel 2.3. Betrachte $b = 10$, $t = 6$, $e \in [-9, 9]$. Es gilt

Zahl	Gleitpunktdarstellung auf dem Rechner
3	$+3.00000 \cdot 10^0$
-26.4	$-2.64000 \cdot 10^1$
0.005	$+5.00000 \cdot 10^{-3}$
1234567	$+1.23457 \cdot 10^6$ (Nicht exakt, wird also gerundet)
$\frac{1}{3} = 0.33\ldots$	$+3.33333 \cdot 10^{-1}$(gerundet)

Ist $x \in \mathbb{R}$ eine beliebige reelle Zahl, so wird x auf dem Rechner durch die nächstgelegene Maschinenzahl ersetzt (gerundet, bei Doppeldeutigkeit wird die betragsmäßig größere Zahl gewählt), die wir mit

$$rd: \ \mathbb{R} \to M, \quad \mathbb{R} \ni x \mapsto rd(x) \in M$$

bezeichnen. Die Rundung $rd : \mathbb{R} \to M$ erfüllt

1. $x \in M \quad \Rightarrow \quad rd(x) \in M$ und
2. $|x - rd(x)| = \min\limits_{y \in M} |x - y|$ ist die Rundung zur nächstgelegenen Maschinenzahl.

Es gilt also

$$|x - rd(x)| \leq |x - y| \quad \text{für alle } y \in M.$$

Bemerkung 2.4. In der Gleitpunktarithmetik gelten die üblichen Rechenregeln, wie das Assoziativ- oder auch das Distributivgesetz, *nicht* mehr.

Beispiel 2.5. Bei Maschinenzahlen mit $b = 10$, $t = 3$, $e \in [-1, 1]$ ergeben sich

$$(1.17 \cdot 10^1 \oplus 1.84 \cdot 10^0) \oplus 2.43 \cdot 10^0 = 1.35 \cdot 10^1 \oplus 2.43 \cdot 10^0 = 1.59 \cdot 10^1,$$

$$1.17 \cdot 10^1 \oplus (1.84 \cdot 10^0 \oplus 2.43 \cdot 10^0) = 1.17 \cdot 10^1 \oplus 4.27 \cdot 10^0 = 1.60 \cdot 10^1,$$

wobei wir mit dem Symbol $\oplus$ die Addition auf dem Rechner kennzeichnen. Der exakte Wert ist 15.97. Die zweite Rechnung liefert eine bessere Approximation.

Definition 2.6 (Exakte und relative Fehler)**.** Sei $x \in \mathbb{R}$ eine reelle Zahl und $rd(x) \in M$ die zugehörige Maschinenzahl. Dann definieren wir

$$e_{abs}(x) := x - rd(x), \quad e_{rel}(x) := \frac{x - rd(x)}{x}.$$

Hier heißt $e_{abs}(x)$ *exakter Fehler* und $e_{rel}(x)$ *relativer Fehler* von x.

Definition 2.7 (Maschinengenauigkeit)**.** Mit

$$\text{eps} := \min\{y \in M \mid 1 \oplus y > 1 \text{ und } y > 0\}$$

bezeichnen wir die *Maschinengenauigkeit*.

Bemerkung 2.8. Für $rd(x) = x(1+\varepsilon)$ gilt nach Definition $|\varepsilon| \leq \text{eps}$ und somit

$$|e_{rel}(x)| = \left|\frac{x - rd(x)}{x}\right| = |\varepsilon| \leq \text{eps}.$$

Im Folgenden untersuchen wir, welchen Einfluss der relative Fehler bei der Addition und der Multiplikation hat.

Lemma 2.9. *Seien $x, y \in \mathbb{R} \setminus \{0\}$ zwei reelle Zahlen, $rd(x)$ und $rd(y)$ die zugehörigen Maschinenzahlen sowie $e_{rel}(x)$ und $e_{rel}(y)$ die relativen Fehler. Dann gilt:*

(i) $\frac{x+y-(rd(x)+rd(y))}{x+y} = \frac{x}{x+y} \cdot e_{rel}(x) + \frac{y}{x+y} \cdot e_{rel}(y)$.

(ii) $\frac{xy - rd(x)rd(y)}{xy} = e_{rel}(x) + e_{rel}(y) - e_{rel}(x)e_{rel}(y)$.

Beweis.

Zu (i):
$$\frac{x}{x+y} \cdot e_{rel}(x) + \frac{y}{x+y} \cdot e_{rel}(y) \overset{\text{Def.}}{=} \frac{x}{x+y}\frac{x - rd(x)}{x} + \frac{y}{x+y}\frac{y - rd(y)}{y}$$
$$= \frac{x - rd(x) + y - rd(y)}{x+y}.$$

Zu (ii):
$$e_{rel}(x) + e_{rel}(y) - e_{rel}(x)e_{rel}(y) = \frac{x - rd(x)}{x} + \frac{y - rd(y)}{y} - \frac{(x - rd(x))(y - rd(y))}{xy}$$
$$= \frac{xy - rd(x)rd(y)}{xy}.$$

□

Bemerkung 2.10. Da $e_{rel}(x)$ und $e_{rel}(y)$ sehr klein sind, ist das Produkt $e_{rel}(x)e_{rel}(y)$ im Vergleich zu $e_{rel}(x) + e_{rel}(y)$ vernachlässigbar klein, so dass

$$\frac{xy - rd(x)rd(y)}{xy} = e_{rel}(x) + e_{rel}(y) - e_{rel}(x)e_{rel}(y) \approx e_{rel}(x) + e_{rel}(y)$$

gilt. Der relative Fehler bei der Multiplikation ist also klein!
Hingegen kann es bei der Addition zu einem großen relativen Fehler kommen, falls $x + y \approx 0$. In diesem Fall spricht man von *Auslöschung*.

Beispiel 2.11. Seien $x = \frac{99}{70} \approx 1.41428571$ und $y = \sqrt{2} \approx 1.41421356$. Bei Maschinenzahlen mit $b = 10$, $t = 6$, $e \in [-1, 1]$ gilt dann $rd(x) = 1.41429 \cdot 10^0$ und $rd(y) = 1.41421 \cdot 10^0$. Dann ist

$$e_{abs}(x) = x - rd(x) \approx -4.3 \cdot 10^{-6},$$
$$e_{abs}(y) = y - rd(y) \approx \quad 3.6 \cdot 10^{-6}$$

sowie

$$e_{rel}(x) = \frac{x - rd(x)}{x} \approx -3.0 \cdot 10^{-6},$$
$$e_{rel}(y) = \frac{y - rd(y)}{y} \approx \quad 2.5 \cdot 10^{-6}.$$

Also folgt für die Addition

$$\frac{x + y - (rd(x) + rd(y))}{x + y} \overset{\text{Lemma}}{=} \frac{x}{x+y} \cdot e_{rel}(x) + \frac{y}{x+y} \cdot e_{rel}(y)$$
$$\approx -1.5 \cdot 10^{-6} + 1.25 \cdot 10^{-6}.$$

Für die Multiplikation gilt

$$\frac{xy - rd(x)rd(y)}{xy} \overset{\text{Lemma}}{=} e_{rel}(x) + e_{rel}(y) - e_{rel}(x)e_{rel}(y) \approx -3.0 \cdot 10^{-6} + 2.5 \cdot 10^{-6}.$$

Bei der Subtraktion jedoch folgt

$$\frac{x - y - (rd(x) - rd(y))}{x - y} \approx 0.11,$$

wir können also den Effekt der Auslöschung feststellen.

Fazit 2.12. Vermeide die Subtraktion annähernd gleichgroßer Zahlen in der Gleitpunktarithmetik.

2.2 Vektor- und Matrixnormen

Definition 2.13 (Norm)**.** Sei X ein linearer Raum (Vektorraum) über $\mathbb{K}$ ($= \mathbb{R}$ oder $= \mathbb{C}$). Eine Abbildung $\|\cdot\| : X \to \mathbb{R}$ heißt *Norm*, falls sie den folgenden Bedingungen genügt:

(N1) $\|x\| \geq 0 \quad \forall x \in X \quad$ und $\quad \|x\| = 0 \quad \Leftrightarrow \quad x = 0,$

(N2) $\|\lambda x\| = |\lambda| \, \|x\| \quad \forall \lambda \in \mathbb{K} \quad \forall x \in X$ (positive Homogenität),

(N3) $\|x + y\| \leq \|x\| + \|y\| \quad \forall x, y \in X$ (Dreiecksungleichung).

Bemerkung 2.14. Da $x = x - 0$ gilt, kann man $\|x\|$ als Abstand von x zum Nullpunkt in X interpretieren.
(N1) besagt, dass alle Abstände positiv sind, sofern es sich nicht um den Nullpunkt handelt.
(N2) besagt, dass die Länge eines Vielfachen gleich das Vielfache der Länge ist.
(N3) besagt, dass die Strecke die kürzeste Verbindung zweier Punkte ist.

In dieser Vorlesung betrachten wir hauptsächlich endlich-dimensionale Vektorräume über $\mathbb{K} = \mathbb{R}$.

Bemerkung 2.15. Eine Norm auf $\mathbb{R}^n$ heißt auch *Vektornorm*. Im Falle $X = \mathbb{R}^{m\times n}$ spricht man auch von einer *Matrixnorm*.

Definition 2.16 (l_p-Norm)**.** Die l_p-Norm auf $\mathbb{R}^n$ ist definiert durch

$$\|x\|_p := (\sum_{i=1}^{n} |x_i|^p)^{\frac{1}{p}} \text{ für } 1 \le p < \infty$$
$$\|x\|_\infty := \max_{i\in\{1,\dots,n\}} |x_i| \text{ für } p = \infty.$$

Drei wichtige Spezialfälle sind:

$$p = 1 : \quad \|x\|_1 = \sum_{i=1}^{n} |x_i| \qquad \text{(Betragssummennorm)}$$
$$p = 2 : \quad \|x\|_2 = (\sum_{i=1}^{n} |x_i|^2)^{\frac{1}{2}} \qquad \text{(euklidische Norm)}$$
$$p = \infty : \quad \|x\|_\infty = \max_{i\in\{1,\dots,n\}} |x_i| \qquad \text{(Maximumnorm).}$$

Definition 2.17 (Äquivalenz von Normen)**.** Sei X ein reeller Vektorraum und seien $\|\cdot\|_\alpha : X \to \mathbb{R}$ sowie $\|\cdot\|_\beta : X \to \mathbb{R}$ zwei Normen auf X. Dann heißen $\|\cdot\|_\alpha$ und $\|\cdot\|_\beta$ äquivalent, falls es positive reelle Konstanten $0 < c_1 \le c_2$ existieren, so dass

$$c_1 \|x\|_\alpha \le \|x\|_\beta \le c_2 \|x\|_\alpha \text{ für alle } x \in X$$

gilt.

Lemma 2.18. *Alle Normen auf $\mathbb{R}^n$ sind äquivalent.*

Beweis. Sei $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ eine beliebig aber fest gewählte Norm auf $\mathbb{R}^n$. Wir zeigen, dass $\|\cdot\|$ und $\|\cdot\|_2$ äquivalent sind.
Die endlich-dimensionale Einheitssphäre

$$K = \{x \in \mathbb{R}^n \mid \|x\|_2 = 1\}$$

ist kompakt (Heine-Borel). Da jede Norm stetig ist, nimmt $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ ihr Minimum und Maximum auf dem Kompaktum K an (Weierstraß). Somit existieren die Konstanten

$$c_1 := \min_{x\in K} \|x\| > 0 \text{ und } c_2 := \max_{x\in K} \|x\| \ge c_1.$$

Für $x \in \mathbb{R}^n \setminus \{0\}$ ist $y := \frac{x}{\|x\|_2} \in K$. Dann gilt

$$c_1 = \min_{x\in K} \|x\| \overset{y\in K}{\le} \|y\| \overset{y\in K}{\le} \max_{x\in K} \|x\| = c_2 \quad \Leftrightarrow \quad c_1 \le \left\| \frac{x}{\|x\|_2} \right\| \le c_2$$
$$\Leftrightarrow \quad c_1 \le \frac{1}{\|x\|_2} \|x\| \le c_2$$
$$\Leftrightarrow \quad c_1 \|x\|_2 \le \|x\| \le c_2 \|x\|_2 .$$

Somit haben wir gezeigt, dass $\|\cdot\|$ und $\|\cdot\|_2$ äquivalent sind. Folglich sind alle Normen auf $\mathbb{R}^n$ äquivalent. □

Bemerkung 2.19.

(i) Alle Normen auf einem *endlich-dimensionalen* Vektorraum sind äquivalent.

(ii) Auf einem *unendlich-dimensionalen* Vektorraum sind zwei beliebige Normen im Allgemeinen nicht mehr äquivalent.

Im folgenden Satz zeigen wir, wie man eine Matrixnorm aus Vektornormen konstruieren kann.

Satz 2.20. *Seien* $\|\cdot\|_X : \mathbb{R}^n \to \mathbb{R}$ *und* $\|\cdot\|_Y : \mathbb{R}^m \to \mathbb{R}$ *Normen auf* $\mathbb{R}^n$ *und* $\mathbb{R}^m$*. Dann wird durch*

$$\|A\|_{YX} := \sup_{x\neq 0} \frac{\|Ax\|_Y}{\|x\|_X} = \sup_{\|x\|_X=1} \|Ax\|_Y,\ A \in \mathbb{R}^{m\times n}$$

eine Matrixnorm auf $\mathbb{R}^{m\times n}$ *definiert.*

Beweis. Da die Einheitssphäre $K = \{x \in \mathbb{R}^n \mid \|x\|_X = 1\}$ kompakt ist und die Abbildung $f : \mathbb{R}^n \to \mathbb{R}$, $x \mapsto \|Ax\|_Y$ stetig ist, existiert

$$\|A\|_{YX} = \max_{\|x\|_X=1} \|Ax\|_Y = \max_{x\neq 0} \frac{\|Ax\|_Y}{\|x\|_X} \in \mathbb{R}$$

für alle $A \in \mathbb{R}^{m\times n}$. Nun zeigen wir, dass $\|\cdot\|_{YX} : \mathbb{R}^{m\times n} \to \mathbb{R}$ eine Norm ist:

Zu (N1): Auf Grund der Konstruktion gilt $\|A\|_{YX} \geq 0$ für alle $A \in \mathbb{R}^{m\times n}$ und

$$\begin{aligned} 0 = \|A\|_{YX} \overset{\text{Def.}}{=} \sup_{x\neq 0} \frac{\|Ax\|_Y}{\|x\|_X} \quad &\Leftrightarrow \quad \|Ax\|_Y = 0 \quad \forall x \in \mathbb{R}^n \\ &\Leftrightarrow \quad Ax = 0 \quad \forall x \in \mathbb{R}^n \\ &\Leftrightarrow \quad A = 0. \end{aligned}$$

Zu (N2): Seien $\lambda \in \mathbb{R}$ und $A \in \mathbb{R}^{m\times n}$. Dann gilt laut Definition

$$\|\lambda A\|_{YX} = \sup_{\|x\|_X=1} \|\lambda Ax\|_Y = |\lambda| \sup_{\|x\|_X=1} \|Ax\|_Y = |\lambda|\, \|A\|_{YX}\,.$$

Zu (N3): Seien $A, B \in \mathbb{R}^{m\times n}$. Dann gilt

$$\begin{aligned} \|A+B\|_{YX} = \sup_{\|x\|_X=1} \|(A+B)x\|_Y &= \sup_{\|x\|_X=1} \|Ax+Bx\|_Y \\ &\leq \sup_{\|x\|_X=1} (\|Ax\|_Y + \|Bx\|_Y) \\ &\leq \sup_{\|x\|_X=1} \|Ax\|_Y + \sup_{\|x\|_X=1} \|Bx\|_Y \\ &= \|A\|_{YX} + \|B\|_{YX}\,. \end{aligned}$$

Somit ist $\|\cdot\|_{YX} : \mathbb{R}^{m\times n} \to \mathbb{R}$ eine Matrixnorm. □

Lemma 2.21. *Seien $\|\cdot\|_X$, $\|\cdot\|_Y$ und $\|\cdot\|_Z$ Vektornormen auf $\mathbb{R}^n$, $\mathbb{R}^m$ und $\mathbb{R}^p$. Dann gilt:*

(i) $\|Ax\|_Y \leq \|A\|_{YX} \|x\|_X$ *für alle* $A \in \mathbb{R}^{m\times n}$ *und alle* $x \in \mathbb{R}^n$,

(ii) $\|AB\|_{YZ} \leq \|A\|_{YX} \|B\|_{XZ}$ *für alle* $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{n\times p}$.

Beweis.

Zu (i): Nach Definition gilt für $A \in \mathbb{R}^{m\times n}$

$$\|A\|_{YX} \overset{\text{Def.}}{=} \sup_{x\neq 0} \frac{\|Ax\|_Y}{\|x\|_X} \geq \frac{\|Ax\|_Y}{\|x\|_X} \quad \forall x \in \mathbb{R}^n \setminus \{0\}.$$

Also ist

$$\|Ax\|_Y \leq \|A\|_{YX} \|x\|_X \quad \forall x \in \mathbb{R}^n \setminus \{0\}$$

und daraus folgt die Aussage (i).

Zu (ii): Seien $A \in \mathbb{R}^{m\times n}$ und $B \in \mathbb{R}^{n\times p}$ gegeben. Dann gilt $AB \in \mathbb{R}^{m\times p}$ und

$$\begin{aligned}\|AB\|_{YZ} \overset{\text{Def.}}{=} \sup_{\substack{x\neq 0\\ x\in\mathbb{R}^p}} \frac{\|A(Bx)\|_Y}{\|x\|_Z} &\leq \sup_{\substack{x\neq 0\\ x\in\mathbb{R}^p}} \frac{\|A\|_{YX}\|Bx\|_X}{\|x\|_Z} \\ &= \|A\|_{YX} \sup_{\substack{x\neq 0\\ x\in\mathbb{R}^p}} \frac{\|Bx\|_X}{\|x\|_Z} \\ &= \|A\|_{YX}\|B\|_{XZ},\end{aligned}$$

wodurch die Aussage (ii) gezeigt ist.

□

Bemerkung 2.22. Die Eigenschaft (i) bezeichnet man als *Verträglichkeit von Vektor- und induzierter Matrixnorm*, die Eigenschaft (ii) als *Submultiplikativität von induzierten Matrixnormen*.

Einige wichtige Matrixnormen auf $\mathbb{R}^{m\times n}$ sind gegeben durch:

$$\begin{aligned}\|A\|_1 &:= \max_{j=1,\dots,n} \sum_{i=1}^{m} |a_{ij}| \quad \text{(Spaltensummennorm)}\\ \|A\|_2 &:= \sqrt{\lambda_{max}(A^TA)} \quad \text{(Spektralnorm)}\\ \|A\|_\infty &:= \max_{i=1,\dots,m} \sum_{j=1}^{n} |a_{ij}| \quad \text{(Zeilensummennorm).}\end{aligned}$$

Hierbei bezeichnet $\lambda_{max}(A^TA)$ den größten Eigenwert von der Matrix A^TA.

Notation 2.23. Für eine Matrix A führen wir folgende Bezeichnung ein:

$$\mathbb{R}^{m\times n} \ni A = (a_{ij}) = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}.$$

Beispiel 2.24. Betrachte

$$A = \begin{pmatrix} -1 & 2 \\ 0 & 2 \\ 1 & 2 \end{pmatrix} \in \mathbb{R}^{3\times 2}.$$

Dann ist

$$\begin{aligned} \|A\|_1 &= \max\{|-1|+0+1,\ 2+2+2\} = 6, \\ \|A\|_\infty &= \max\{|-1|+2,\ 0+2,\ 1+2\} = 3. \end{aligned}$$

Für die Spektralnorm berechnen wir zuerst A^TA:

$$A^TA = \begin{pmatrix} -1 & 0 & 1 \\ 2 & 2 & 2 \end{pmatrix} \begin{pmatrix} -1 & 2 \\ 0 & 2 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 12 \end{pmatrix}.$$

Auf Grund der Diagonalgestalt der Matrix können wir die Eigenwerte der Matrix A^TA direkt ablesen: $\lambda_1 = 2$, $\lambda_2 = 12$. Somit ist

$$\|A\|_2 = \sqrt{\lambda_{max}(A^TA)} = \sqrt{12}.$$

Definition 2.25. Sei $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ eine quadratische Matrix. Dann heißt A

- symmetrisch, falls gilt: $A^T = A$, d.h. $a_{ij} = a_{ji}$ für alle $i, j = 1, \ldots, n$
- positiv semidefinit, falls gilt: $x^TAx \geq 0$ für alle $x \in \mathbb{R}^n$
- positiv definit, falls gilt: $x^TAx > 0$ für alle $x \in \mathbb{R}^n \setminus \{0\}$
- orthogonal, falls gilt: $AA^T = I$, wobei $I \in \mathbb{R}^{n\times n}$ die Einheitsmatrix bezeichnet.

Aus der linearen Algebra verwenden wir folgenden Hilfssatz ohne Beweis.

Hilfssatz 2.26. *Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch. Dann existiert eine orthogonale Matrix $Q \in \mathbb{R}^{n\times n}$, so dass*

$$Q^TAQ = D = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^{n\times n}$$

gilt. Die Diagonalelemente $\lambda_1, \ldots, \lambda_n$ von D sind gerade die Eigenwerte von A.

Korollar 2.27. *Jede symmetrische Matrix ist genau dann positiv definit (positiv semidefinit), wenn alle Eigenwerte positiv (nichtnegativ) sind.*

Korollar 2.28. *Sei $B \in \mathbb{R}^{n\times n}$ eine symmetrische Matrix. Dann gilt*

$$\lambda_{min} x^T x \leq x^T B x \leq \lambda_{max} x^T x \quad \forall x \in \mathbb{R}^n,$$

wobei λ_{min} bzw. λ_{max} den kleinsten bzw. den größten Eigenwert von B bezeichnen.

Beweis. Da $B \in \mathbb{R}^{n\times n}$ symmetrisch ist, existiert eine orthogonale Matrix $Q \in \mathbb{R}^{n\times n}$ mit

$$Q^T B Q = D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n).$$

Sei $x \in \mathbb{R}^n$. Dann gilt mit $y := Q^T x$:

$$x^T B x \overset{QQ^T=I}{=} x^T Q \underbrace{Q^T B Q}_{D} Q^T x = x^T Q D Q^T x = y^T D y = \sum_{i=1}^{n} \lambda_i y_i^2.$$

Insgesamt folgt:

$$x^T B x = \sum_{i=1}^{n} \lambda_i y_i^2 \geq \lambda_{min} y^T y \overset{y=Q^T x}{=} \lambda_{min} x^T Q^T Q x = \lambda_{min} x^T x.$$

Analog gilt:

$$x^T B x = \sum_{i=1}^{n} \lambda_i y_i^2 \leq \lambda_{max} y^T y \overset{y=Q^T x}{=} \lambda_{max} x^T Q^T Q x = \lambda_{max} x^T x.$$

Somit gilt die Behauptung. □

Satz 2.29. *Wählen wir sowohl im Raum $\mathbb{R}^n$ als auch im Raum $\mathbb{R}^m$ die Betragssummennorm $\|\cdot\|_1$, die euklidische Norm $\|\cdot\|_2$ und die Maximumnorm $\|\cdot\|_\infty$, so sind die jeweils induzierten Matrixnormen gegeben durch:*

$$\text{(i)}\ \|A\|_1 := \max_{1\leq j\leq n} \sum_{i=1}^{m} |a_{ij}| = \sup_{x\neq 0} \frac{\|Ax\|_1}{\|x\|_1} \quad \forall A \in \mathbb{R}^{m\times n} \quad \textit{(Spaltensummennorm)}$$

$$\text{(ii)}\ \|A\|_2 := \sqrt{\lambda_{max}(A^T A)} = \sup_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2} \quad \forall A \in \mathbb{R}^{m\times n} \quad \textit{(Spektralnorm)}$$

$$\text{(iii)}\ \|A\|_\infty := \max_{1\leq i\leq m} \sum_{j=1}^{n} |a_{ij}| = \sup_{x\neq 0} \frac{\|Ax\|_\infty}{\|x\|_\infty} \quad \forall A \in \mathbb{R}^{m\times n} \quad \textit{(Zeilensummennorm)}$$

Beweis.

Zu (i): Für alle $x \in \mathbb{R}^n$ gilt

$$\|Ax\|_1 = \sum_{i=1}^{m} \left| \sum_{j=1}^{n} a_{ij} x_j \right| \leq \sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}||x_j| = \sum_{j=1}^{n} \sum_{i=1}^{m} |a_{ij}||x_j| \leq \sum_{j=1}^{n} \max_{1 \leq j \leq n} \sum_{i=1}^{m} |a_{ij}||x_j|$$
$$= \max_{1 \leq j \leq n} \sum_{i=1}^{m} |a_{ij}| \sum_{j=1}^{n} |x_j|$$
$$= \|A\|_1 \|x\|_1 .$$

Daraus folgt

$$\sup_{x \neq 0} \frac{\|Ax\|_1}{\|x\|_1} \leq \|A\|_1 .$$

Umgekehrt wähle ein j mit $\|A\|_1 = \sum_{i=1}^{m} |a_{ij}|$, $x = e_j$. Dann ist

$$\|Ax\|_1 = \|Ae_j\|_1 = \sum_{i=1}^{m} |a_{ij}| = \|A\|_1 \|x\|_1 .$$

Also folgt für $x = e_j$

$$\frac{\|Ax\|_1}{\|x\|_1} = \|A\|_1 .$$

Somit ist

$$\sup_{x \neq 0} \frac{\|Ax\|_1}{\|x\|_1} \geq \|A\|_1$$

und insgesamt ergibt sich

$$\|A\|_1 \leq \sup_{x \neq 0} \frac{\|Ax\|_1}{\|x\|_1} \leq \|A\|_1$$

und daraus ergibt sich die Behauptung.

Zu (ii): Für alle $x \in \mathbb{R}^n$ gilt

$$\|Ax\|_2^2 \overset{\text{Def.}}{=} (Ax)^T Ax = x^T A^T Ax \overset{\text{Kor.}}{\leq} \lambda_{max}(A^T A) x^T x = \lambda_{max}(A^T A) \|x\|_2^2 .$$

Daraus folgt

$$\|Ax\|_2 \leq \sqrt{\lambda_{max}(A^T A)} \|x\|_2$$

und somit

$$\sup_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} \leq \sqrt{\lambda_{max}(A^T A)} = \|A\|_2 .$$

Umgekehrt sei $v \in \mathbb{R}^n$ ein Eigenvektor von A^TA zum Eigenwert $\lambda_{max}(A^TA)$. Dann gilt

$$\|Av\|_2^2 = v^TA^TAv = \lambda_{max}(A^TA)v^Tv = \lambda_{max}(A^TA)\,\|v\|_2^2\,.$$

Folglich ist

$$\|Av\|_2 = \sqrt{\lambda_{max}(A^TA)}\,\|v\|_2 = \|A\|_2\,\|v\|_2$$

und somit

$$\|A\|_2 = \frac{\|Av\|_2}{\|v\|_2} \leq \sup_{x\neq 0}\frac{\|Ax\|_2}{\|x\|_2}.$$

Insgesamt gilt also

$$\|A\|_2 \leq \sup_{x\neq 0}\frac{\|Ax\|_2}{\|x\|_2} \leq \|A\|_2$$

und daraus folgt die Behauptung.

Zu (iii): Es gilt für alle $x \in \mathbb{R}^n$

$$\|Ax\|_\infty = \max_{1\leq i\leq m}|\sum_{j=1}^n a_{ij}x_j| \leq \max_{1\leq i\leq m}\sum_{j=1}^n |a_{ij}|\max_{1\leq j\leq n}|x_j| = \|A\|_\infty\,\|x\|_\infty\,.$$

Daraus folgt

$$\sup_{x\neq 0}\frac{\|Ax\|_\infty}{\|x\|_\infty} \leq \|A\|_\infty\,.$$

Sei nun $k \in \{1,\ldots,m\}$ beliebig aber fest. Wir definieren den Vektor $v \in \mathbb{R}^n$ mit

$$v_j := \begin{cases} \frac{|a_{kj}|}{a_{kj}} & \text{falls } a_{kj} \neq 0 \\ 1 & \text{falls } a_{kj} = 0 \end{cases}.$$

Offenbar ist $\|v\|_\infty = 1$. Somit gilt

$$\begin{aligned}\sup_{x\neq 0}\frac{\|Ax\|_\infty}{\|x\|_\infty} \geq \frac{\|Av\|_\infty}{\|v\|_\infty} = \|Av\|_\infty = \max_{1\leq i\leq m}|(Av)_i| \geq |(Av)_k| \\ = |\sum_{j=1}^n a_{kj}v_j| \\ \overset{\text{Def.}}{=} |\sum_{j=1}^n |a_{kj}||.\end{aligned}$$

Daher ist

$$\sup_{x\neq 0}\frac{\|Ax\|_\infty}{\|x\|_\infty} \geq \sum_{j=1}^n |a_{kj}|.$$

Da aber der Zeilenindex k frei gewählt wurde, so folgt

$$\sup_{x\neq 0} \frac{\|Ax\|_\infty}{\|x\|_\infty} \geq \max_{1\leq i\leq m} \sum_{j=1}^{n} |a_{ij}| = \|A\|_\infty .$$

Insgesamt haben wir also gezeigt, dass

$$\|A\|_\infty \leq \sup_{x\neq 0} \frac{\|Ax\|_\infty}{\|x\|_\infty} \leq \|A\|_\infty$$

gilt und daraus folgt die Behauptung.

□

Folgerung 2.30. Es gilt für $\mu \in \{1, 2, \infty\}$:

(i): $\|Ax\|_\mu \leq \|A\|_\mu \|x\|_\mu$ (Verträglichkeit),

(ii): $\|AB\|_\mu \leq \|A\|_\mu \|B\|_\mu$ (Submultiplikativität).

2.3 Kondition und Fehlerabschätzung

Definition 2.31. Sei $A \in \mathbb{R}^{m\times n}$ eine Matrix. Dann definieren wir:

$$\begin{aligned} \text{Kern}(A) &:= \{x \in \mathbb{R}^n \mid Ax = 0\} \\ \text{Bild}(A) &:= \{y \in \mathbb{R}^m \mid \exists x \in \mathbb{R}^n : y = Ax\} \\ \text{Rang}(A) &:= \dim(\text{Bild}(A)). \end{aligned}$$

Aus der linearen Algebra verwenden wir ohne Beweis den folgenden Satz.

Satz 2.32 (Dimensionssatz)**.** *Sei $A \in \mathbb{R}^{m\times n}$ eine Matrix. Dann gilt*

$$n = \dim(\text{Kern}(A)) + \dim(\text{Bild}(A)).$$

Lemma 2.33 (Lösbarkeitskriterium für $Ax = b$)**.** *Sei $A \in \mathbb{R}^{n\times n}$ und $b \in \mathbb{R}^n$. Dann sind die folgenden Aussagen äquivalent:*

(i) Das lineare Gleichungssystem $Ax = b$ hat mindestens eine Lösung.

(ii) Es gilt $b \in \text{Bild}(A)$.

(iii) $\text{Rang}(A) = \text{Rang}(A, b)$, wobei $(A, b) \in \mathbb{R}^{n\times(n+1)}$ diejenige Matrix ist, die aus A durch Hinzunahme von b als weitere Spalte entsteht.

Beweis. Die Äquivalenz zwischen (i) und (ii) folgt direkt aus der Definition des Bildes. Aus (ii) folgt wieder mit der Definition des Bildes $\text{Bild}(A) = \text{Bild}(A, b)$ und daraus wieder (iii). Umgekehrt folgt aus (iii), dass b als Linearkombination von A darstellbar ist. Dies liefert $b \in \text{Bild}(A)$, woraus sich (ii) ergibt. Damit ist insgesamt die Äquivalenz aller Aussagen gezeigt. □

Lemma 2.34 (Lösungsmenge)**.** *Seien $A \in \mathbb{R}^{n\times n}$ und $b \in \mathbb{R}^n$. Es existiere $x^* \in \mathbb{R}^n$, so dass $Ax^* = b$ ist. Dann gilt für die Lösungsmenge:*

$$\mathbb{L} := \{x \in \mathbb{R}^n \mid Ax = b\} = x^* + \text{Kern}(A).$$

Beweis.

$\subseteq$: Sei $x \in \mathbb{L}$. Dann gilt:

$$A(x - x^*) = Ax - Ax^* = b - b = 0 \quad \overset{\text{Def.}}{\Rightarrow} \quad x - x^* \in \text{Kern}(A).$$

Damit ergibt sich:

$$x = x^* + \underbrace{x - x^*}_{\in \text{Kern}(A)} \in x^* + \text{Kern}(A).$$

$\supseteq$: Sei $x = x^* + z$ mit $z \in \text{Kern}(A)$. Dann gilt:

$$Ax = A(x^* + z) = Ax^* + Az = b + 0.$$

Damit ergibt sich $x \in \mathbb{L}$.

□

Definition 2.35. Sei $A \in \mathbb{R}^{n\times n}$ eine quadratische Matrix. Dann heißt A *regulär*, falls

$$\text{Kern}(A) = \{0\}.$$

Folgerung 2.36. Sei $A \in \mathbb{R}^{n\times n}$ eine quadratische Matrix. Dann sind die folgenden Aussagen äquivalent:

(i) A ist regulär ($\text{Kern}(A) = \{0\}$).

(ii) $\text{Bild}(A) = \mathbb{R}^n$.

(iii) Das lineare Gleichungssystem $Ax = b$ hat für jedes $b \in \mathbb{R}^n$ genau eine Lösung $x \in \mathbb{R}^n$.

Bemerkung 2.37. Aus der linearen Algebra ist auch bekannt:

$$\text{A ist regulär} \quad \Leftrightarrow \quad \det(A) \neq 0.$$

Jede reguläre Matrix $A \in \mathbb{R}^{n\times n}$ besitzt eine inverse Matrix $A^{-1} \in \mathbb{R}^{n\times n}$, die durch

$$A^{-1}A = I \quad \Leftrightarrow \quad AA^{-1} = I$$

charakterisiert ist.
Das Produkt zweier regulärer Matrizen $A, B \in \mathbb{R}^{n\times n}$ ist regulär mit

$$(AB)^{-1} = B^{-1}A^{-1},$$

denn es ist

$$(B^{-1}A^{-1})AB = B^{-1}(A^{-1}A)B = B^{-1}IB = I.$$

Ist $A \in \mathbb{R}^{n\times n}$ regulär, so ist die transponierte Matrix A^T auch regulär mit

$$(A^T)^{-1} = (A^{-1})^T,$$

denn es gilt

$$(A^{-1})^T A^T = (A^{-1}A)^T = I^T = I.$$

Definition 2.38 (Orthogonalraum)**.** Sei $U \subseteq \mathbb{R}^n$ nichtleer. Mit

$$U^\perp := \{y \in \mathbb{R}^n \mid \langle y, u\rangle = 0 \quad \forall u \in U\}$$

bezeichnen wir den *Orthogonalraum* von U. Hierbei ist

$$\langle\cdot,\cdot\rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$$

das euklidische Skalarprodukt $\langle x, y\rangle = x^T y$.

Bemerkung 2.39. Ist $U \subseteq \mathbb{R}^n$ nichtleer, so gilt:

(i) $U^\perp \subseteq \mathbb{R}^n$ ist ein Untervektorraum.

(ii) $U \subseteq (U^\perp)^\perp$.

Satz 2.40. *Sei $U \subseteq \mathbb{R}^n$ ein Untervektorraum (Unterraum) und $U^\perp \subseteq \mathbb{R}^n$ der Orthogonalraum von U. Dann gilt*

$$\mathbb{R}^n = U \oplus U^\perp.$$

Mit anderen Worten gibt es zu jedem $x \in \mathbb{R}^n$ genau ein $u \in U$ und genau ein $u^\perp \in U^\perp$, so dass

$$x = u + u^\perp.$$

Beweis. Sei $\{u_1, \ldots, u_m\}$ eine Orthonormalbasis von U. Definiere

$$P : \mathbb{R}^n \to U, \quad P(x) = \sum_{i=1}^{m} \langle x, u_i\rangle u_i.$$

Sei $x \in \mathbb{R}^n$. Wir zeigen:

$$x - P(x) \in U^\perp \quad \overset{\text{Def.}}{\Leftrightarrow} \quad \langle x - P(x), u\rangle = 0 \quad \forall u \in U.$$

Sei also $u \in U$ beliebig aber fest. Dann hat u die Darstellung

$$u = \sum_{i=1}^{m} \lambda_i u_i, \quad \lambda_i \in \mathbb{R},$$

da $\{u_1, \ldots, u_m\}$ eine Basis ist. Folglich gilt

$$\begin{aligned}
\langle x - P(x), u\rangle &= \langle x, u\rangle - \langle P(x), u\rangle \\
&= \sum_{i=1}^{m} (\lambda_i \langle x, u_i\rangle - \lambda_i \langle P(x), u_i\rangle) \\
&= \sum_{i=1}^{m} \left(\lambda_i \langle x, u_i\rangle - \lambda_i \sum_{j=1}^{m} \langle x, u_j\rangle \langle u_j, u_i\rangle \right) \\
&\overset{\langle u_j, u_i\rangle = \delta_{ij}}{=} \sum_{i=1}^{m} (\lambda_i \langle x, u_i\rangle - \lambda_i \langle x, u_i\rangle) \\
&= 0.
\end{aligned}$$

Daher ist $\langle x - P(x), u\rangle = 0$ für alle $u \in U$. Mit anderen Worten: $x - P(x) \in U^\perp$. Setzen wir

$$u := P(x) \in U \text{ und } u^\perp := x - P(x) \in U^\perp,$$

so hat x die Darstellung

$$x = P(x) + x - P(x) = u + u^\perp.$$

Zur Eindeutigkeit: Gibt es ein $\tilde{u} \in U$ und ein $\tilde{u}^\perp \in U^\perp$ mit

$$x = \tilde{u} + \tilde{u}^\perp,$$

so müssen wir nun zeigen, dass $u = \tilde{u}$ und $u^\perp = \tilde{u}^\perp$ ist. Es gilt

$$0 = u - \tilde{u} + u^\perp - \tilde{u}^\perp,$$

woraus

$$0 = \langle u - \tilde{u}, u - \tilde{u}\rangle + \langle u^\perp - \tilde{u}^\perp, u - \tilde{u}\rangle$$

folgt. Jetzt ist aber per Definition $\langle u^\perp - \tilde{u}^\perp, u - \tilde{u}\rangle = 0$, denn es ist $u^\perp - \tilde{u}^\perp \in U^\perp$ und $u - \tilde{u} \in U$. Also muss auch

$$\langle u - \tilde{u}, u - \tilde{u}\rangle = 0$$

gelten. Aus der Definition der Norm folgt aus $0 = \|u - \tilde{u}\|_2^2$, dass $u = \tilde{u}$ gilt und somit auch $u^\perp = \tilde{u}^\perp$. □

Folgerung 2.41. Ist $U \subseteq \mathbb{R}^n$ ein Untervektorraum, dann gilt

$$U = (U^\perp)^\perp.$$

Satz 2.42. *Sei $A \in \mathbb{R}^{m\times n}$ eine Matrix. Dann gilt:*

(i) $\text{Kern}(A) = \text{Bild}(A^T)^\perp$ *und* $\text{Kern}(A)^\perp = \text{Bild}(A^T)$.

(ii) $\text{Kern}(A) = \text{Kern}(A^T A)$.

(iii) $\text{Bild}(A^T) = \text{Bild}(A^T A)$.

Beweis.

Zu (i):

- $\subseteq$: Ist $v \in \text{Kern}(A)$, so ist per Definition $Av = 0$. Daraus folgt $\langle v, A^T x\rangle = \langle Av, x\rangle = 0$ für alle $x \in \mathbb{R}^m$. Mit anderen Worten ist also $v \in \text{Bild}(A^T)^\perp$.
- $\supseteq$: Ist $v \in \text{Bild}(A^T)^\perp$, so ist per Definition $\langle v, A^T x\rangle = 0$ für alle $x \in \mathbb{R}^m$. Dies ist gleichbedeutend mit $\langle Av, x\rangle = 0$ für alle $x \in \mathbb{R}^m$. Mit der Substitution $x = Av$ folgt $\langle Av, Av\rangle = 0$. Daher ist $Av = 0$ bzw. $v \in \text{Kern}(A)$.

Zu (ii):

- $\subseteq$: Ist $v \in \text{Kern}(A)$, so ist $Av = 0$. Daraus folgt $A^T Av = 0$. Das bedeutet $v \in \text{Kern}(A^T A)$.
- $\supseteq$: Ist $v \in \text{Kern}(A^T A)$, so ist $A^T Av = 0$. Daraus folgt $v^T A^T Av = 0$ bzw. $(Av)^T Av = 0$, woraus sich $Av = 0$ ergibt. Dies bedeutet $v \in \text{Kern}(A)$.

Zu (iii): Es gilt:

$$\begin{aligned}\text{Bild}(A^T) = (\text{Bild}(A^T)^\perp)^\perp \overset{\text{(i)}}{=} \text{Kern}(A)^\perp = \text{Kern}(A^T A)^\perp \overset{\text{(i)}}{=} (\text{Bild}(A^T A)^\perp)^\perp \\ = \text{Bild}(A^T A).\end{aligned}$$

□

Wir wollen nun die Stabilität der Lösung des linearen Gleichungssystems

$$Ax = b$$

mit einer regulären Matrix $A \in \mathbb{R}^{n\times n}$ und einem vorgegebenen Vektor $b \in \mathbb{R}^n$ untersuchen. Diese Aufgabe hat genau eine Lösung $x \in \mathbb{R}^n$, da A regulär ist. Betrachten wir den Vektor b mit einer „kleinen" Störung, also $\tilde{b} \in \mathbb{R}^n$, dann hat

$$A\tilde{x} = \tilde{b}$$

genau eine Lösung $\tilde{x} \in \mathbb{R}^n$. Wir stellen also in diesem Zusammenhang die Frage, wie sich

$$\|x - \tilde{x}\|$$

verhält.

Satz 2.43. *Es sei $A \in \mathbb{R}^{n\times n}$ eine reguläre Matrix sowie $b, \tilde{b} \in \mathbb{R}^n$ Vektoren mit $b \neq 0$. Weiter seien x und $\tilde{x}$ Lösungen der linearen Gleichungssysteme*

$$Ax = b \quad \textit{und} \quad A\tilde{x} = \tilde{b}$$

($\tilde{b}$ ist die gestörte rechte Seite). Dann gilt

$$\frac{\|x - \tilde{x}\|}{\|x\|} \leq \|A\| \, \|A^{-1}\| \frac{\|b - \tilde{b}\|}{\|b\|}.$$

Hierbei bezeichnet $\|\cdot\|$ eine (beliebige) Vektornorm bzw. die hierdurch induzierte Matrixnorm.

Beweis. Sei $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ eine Vektornorm und die induzierte Matrixnorm ist gegeben durch

$$\|B\| = \sup_{x \neq 0} \frac{\|Bx\|}{\|x\|} \quad \forall B \in \mathbb{R}^{n \times n}.$$

Da A regulär ist, gilt

$$\|x - \tilde{x}\| = \|A^{-1}(b - \tilde{b})\| \leq \|A^{-1}\| \left\|b - \tilde{b}\right\|.$$

Außerdem gilt

$$\|b\| = \|Ax\| \leq \|A\| \, \|x\| \overset{b \neq 0,\, x \neq 0}{\Rightarrow} \frac{1}{\|x\|} \leq \frac{\|A\|}{\|b\|}.$$

Insgesamt ergibt sich also

$$\frac{\|x - \tilde{x}\|}{\|x\|} \leq \|A\| \, \|A^{-1}\| \frac{\|b - \tilde{b}\|}{\|b\|}.$$

□

Bemerkung 2.44. Der Satz besagt

$$\underbrace{\frac{\|x - \tilde{x}\|}{\|x\|}}_{\text{Relativer Fehler in } x} \leq \underbrace{\|A\| \, \|A^{-1}\|}_{\text{Verstärkungsfaktor}} \underbrace{\frac{\|b - \tilde{b}\|}{\|b\|}}_{\text{Relativer Fehler in } b}.$$

Ist der Verstärkungsfaktor klein, so ist der relative Fehler in x auch klein. Problematisch ist, wenn $\|A\| \, \|A^{-1}\|$ groß ist. In diesem Fall führt eine kleine Störung der rechten Seite b auf eine große Änderung der Lösung.

Definition 2.45 (Kondition)**.** Sei $A \in \mathbb{R}^{n \times n}$ regulär und $\|\cdot\| : \mathbb{R}^{n \times n} \to \mathbb{R}$ eine Matrixnorm. Dann heißt

$$\text{cond}(A) := \|A\| \, \|A^{-1}\|$$

Kondition von A. Im Falle der Spektralnorm

$$\|A\|_2 = \sqrt{\lambda_{max}(A^T A)}$$

wird die Kondition

$$\text{cond}_2(A) := \|A\|_2 \, \|A^{-1}\|_2$$

auch als *Spektralkondition* bezeichnet.

Bemerkung 2.46. Ist $\|\cdot\| : \mathbb{R}^{n\times n} \to \mathbb{R}$ eine induzierte Matrixnorm, so lässt sich der Satz wie folgt formulieren:

$$\frac{\|x - \tilde{x}\|}{\|x\|} \leq \text{cond}(A) \frac{\|b - \tilde{b}\|}{\|b\|}.$$

Es gilt für $\text{cond}(A)$

$$1 = \|I\| = \|AA^{-1}\| \overset{\text{Submultiplikativität}}{\leq} \|A\| \, \|A^{-1}\| = \text{cond}(A).$$

Definition 2.47. Eine reguläre Matrix $A \in \mathbb{R}^{n\times n}$ heißt *schlecht konditioniert*, falls die Kondition $\text{cond}(A)$ groß ist ($\text{cond}(A) \ggg 1$).

Lemma 2.48. *Ist $Q \in \mathbb{R}^{n\times n}$ orthogonal (d.h. $Q^TQ = QQ^T = I$), so gilt für die Spektralkondition von Q:*

$$\text{cond}_2(Q) = 1.$$

Beweis. Sei $Q \in \mathbb{R}^{n\times n}$ orthogonal. Dann gilt:

$$\|Qx\|_2^2 = (Qx)^TQx = x^TQ^TQx = x^Tx = \|x\|_2^2.$$

Daraus folgt:

$$\|Q\|_2 = \sup_{x\neq 0} \frac{\|Qx\|_2}{\|x\|_2} = 1.$$

Analog gilt:

$$\|Q^{-1}x\|_2^2 = (Q^{-1}x)^TQ^{-1}x = x^T(Q^{-1})^TQ^{-1}x = x^T(Q^T)^{-1}Q^{-1}x = x^T(QQ^T)^{-1}x = \|x\|_2^2.$$

Daraus folgt

$$\|Q^{-1}\|_2 = \sup_{x\neq 0} \frac{\|Q^{-1}x\|_2}{\|x\|_2} = 1$$

und insgesamt gilt

$$\text{cond}_2(Q) = \|Q\|_2 \, \|Q^{-1}\|_2 = 1.$$

□

Lemma 2.49. *Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv semidefinit. Dann gilt*

$$\lambda_{max}(AA) = \lambda_{max}(A)^2.$$

Beweis. Da $A \in \mathbb{R}^{n\times n}$ symmetrisch ist, existiert eine orthogonale Matrix $Q \in \mathbb{R}^{n\times n}$, so dass

$$Q^TAQ = D = \text{diag}(\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^{n\times n},$$

wobei $\lambda_1, \ldots, \lambda_n$ die Eigenwerte von A sind. Weiter sind

$$\lambda_1, \ldots, \lambda_n \geq 0,$$

weil A positiv semidefinit ist. Wir zeigen, dass

$$\lambda_1^2, \ldots, \lambda_n^2$$

die Eigenwerte von AA sind.
Es gilt

$$AA = QQ^T AQQ^T AQQ^T = QD^2Q^T$$

mit

$$D^2 = \operatorname{diag}(\lambda_1^2, \ldots, \lambda_n^2).$$

Folglich gilt für jedes $i = 1, \ldots, n$

$$AA \underbrace{Qe_i}_{=w} = QD^2Q^TQe_i = QD^2e_i = Q\lambda_i^2 e_i = \lambda_i^2 Qe_i = \lambda_i^2 w,$$

was bedeutet, dass λ_i^2 Eigenwert von AA mit Eigenvektor $w = Qe_i \neq 0$ ist. Somit sind $\lambda_1^2, \ldots, \lambda_n^2$ Eigenwerte von AA. Folglich ist die Aussage richtig. □

Beispiel 2.50. Die Aussage gilt nicht für allgemeine Matrizen.

$$A = \begin{pmatrix} 1 & 0 \\ 0 & -10 \end{pmatrix}, \quad AA = \begin{pmatrix} 1 & 0 \\ 0 & 100 \end{pmatrix}.$$

Hier ist $\lambda_{max}(AA) = 100 \neq 1 = \lambda_{max}(A)^2$.

Lemma 2.51. *Es gilt*

$$\|B^TB\|_2 = \|B\|_2^2 \quad \forall B \in \mathbb{R}^{n\times n}.$$

Beweis. Es ist

$$\|B^TB\|_2 = \sqrt{\lambda_{max}((B^TB)^T(B^TB))} = \sqrt{\lambda_{max}((B^TB)(B^TB))}.$$

Die Matrix $B^TB \in \mathbb{R}^{n\times n}$ ist symmetrisch und positiv semidefinit. Somit liefert das Lemma 2.49

$$\lambda_{max}((B^TB)^T(B^TB)) = \lambda_{max}(B^TB)^2.$$

Daher gilt

$$\|B^TB\|_2 = \sqrt{\lambda_{max}((B^TB)^T(B^TB))} = \lambda_{max}(B^TB) = \|B\|_2^2.$$

□

Satz 2.52. *Sei $A \in \mathbb{R}^{n\times n}$ regulär. Dann gilt für die Spektralkondition:*

$$\operatorname{cond}_2(A^TA) = \operatorname{cond}_2(A)^2.$$

Beweis. Zunächst zeigen wir, dass die Eigenwerte von A^TA und AA^T gleich sind:

$$\begin{aligned}
&\lambda \text{ ist ein Eigenwert von } A^TA \text{ mit Eigenvektor } v \in \mathbb{R}^n \setminus \{0\} \\
\Leftrightarrow\ & A^TAv = \lambda v \\
\overset{\text{A regulär}}{\Leftrightarrow}\ & AA^T \underbrace{Av}_{w} = \lambda Av \\
\Leftrightarrow\ & AA^Tw = \lambda w \\
\Leftrightarrow\ & \lambda \text{ ist ein Eigenwert von } AA^T \text{ mit Eigenvektor } w \in \mathbb{R}^n \setminus \{0\}.
\end{aligned}$$

Daraus folgt

$$\lambda_{max}(A^TA) = \lambda_{max}(AA^T).$$

Daher ist

$$\|A\|_2 = \sqrt{\lambda_{max}(A^TA)} = \sqrt{\lambda_{max}(AA^T)} = \|A^T\|_2.$$

Die gleiche Aussage gilt auch für A^{-1}

$$\|A^{-1}\|_2 = \|(A^{-1})^T\|_2.$$

Das obige Lemma liefert

$$\|A^TA\|_2 = \|A\|_2^2$$

und

$$\|A^{-1}(A^{-1})^T\|_2 \overset{B=(A^{-1})^T}{=} \|(A^{-1})^T\|_2^2 \overset{\text{s.o.}}{=} \|A^{-1}\|_2^2.$$

Daraus folgt

$$\begin{aligned}
\text{cond}_2(A^TA) = \|A^TA\|_2\|(A^TA)^{-1}\|_2 = \|A\|_2^2\|(A^TA)^{-1}\|_2 &= \|A\|_2^2\,\|A^{-1}(A^{-1})^T\|_2 \\
&= \|A\|_2^2\,\|A^{-1}\|_2^2 \\
&= \text{cond}_2(A)^2.
\end{aligned}$$

□

Bemerkung 2.53. Ist A schlecht konditioniert, dann ist A^TA noch schlechter konditioniert.

Satz 2.54. *Ist $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit, so gilt*

$$\text{cond}_2(A) = \frac{\lambda_{max}(A)}{\lambda_{min}(A)}.$$

Beweis. Es gilt

$$\|A\|_2^2 = \lambda_{max}(A^TA) \overset{\text{A symm.}}{=} \lambda_{max}(AA) \overset{\text{Lemma + A s.p.d.}}{=} \lambda_{max}(A)^2.$$

Ist λ Eigenwert von A, so ist λ^{-1} Eigenwert von A^{-1}. Folglich ist

$$\|A^{-1}\|_2^2 = \lambda_{max}((A^{-1})^T A^{-1}) = \lambda_{max}(A^{-1}A^{-1}) = \lambda_{max}(A^{-1})^2 = \frac{1}{\lambda_{min}(A)^2}.$$

Daraus folgt

$$\text{cond}_2(A) = \|A\|_2 \|A^{-1}\|_2 = \frac{\lambda_{max}(A)}{\lambda_{min}(A)}.$$

□

Kapitel 3

Direkte Verfahren

Dieses Kapitel befasst sich mit direkten Verfahren zur Lösung von linearen Gleichungssystemen der Gestalt $Ax = b$.

3.1 Lineare Gleichungssysteme einfacher Strukturen

Wir betrachten zunächst ein lineares Gleichungssystem der Form

$$Dx = b$$

mit einer Diagonalmatrix

$$D = \operatorname{diag}(d_1, \ldots, d_n) = \begin{pmatrix} d_1 & & \\ & \ddots & \\ & & d_n \end{pmatrix} \in \mathbb{R}^{n \times n}.$$

Die Lösung x ist offenbar gegeben durch

$$d_i x_i = b_i \quad \forall i = 1, \ldots, n.$$

Hieraus ergibt sich:

Algorithmus 3.1 Lösung von $Dx = b$ mit einer Diagonalmatrix $D \in \mathbb{R}^{n \times n}$

```
1: for i = 1 : n do
2:     x_i := b_i / d_i
3: end for
```

Der Algorithmus 3.1 ist genau dann durchführbar, wenn $d_i \neq 0$ für alle $i = 1, \ldots, n$, also wenn die Matrix D regulär ist.
Sei nun $L \in \mathbb{R}^{n \times n}$ eine untere Dreiecksmatrix

$$L = \begin{pmatrix} l_{11} & & 0 \\ \vdots & \ddots & \\ l_{n1} & \cdots & l_{nn} \end{pmatrix}.$$

Wir betrachten das Gleichungssystem $Lx = b$. Das ist äquivalent zu:

$$\begin{aligned} l_{11}x_1 &= b_1 \\ l_{21}x_1 + l_{22}x_2 &= b_2 \\ \vdots &\quad \vdots \\ l_{n1}x_1 + l_{n2}x_2 + \cdots + l_{nn}x_n &= b_n. \end{aligned}$$

Aus der ersten Zeile lässt sich x_1 berechnen, danach ergibt sich x_2 aus der zweiten Zeile, usw. Diese Vorgehensweise heißt *Vorwärtseinsetzen* oder *Vorwärtssubstitution*.

Algorithmus 3.2 Lösung von $Lx = b$ mit einer unteren Dreiecksmatrix $L \in \mathbb{R}^{n\times n}$

1: $x_1 := b_1/l_{11}$
2: **for** $i = 2 : n$ **do**
3: $\quad x_i := (b_i - \sum_{j=1}^{i-1} l_{ij}x_j)/l_{ii}$
4: **end for**

Der Algorithmus 3.2 ist genau dann durchführbar, wenn $l_{ii} \neq 0$ für alle $i = 1, \ldots, n$, also wenn L regulär ist. Zur Berechnung von x_i benötigen wir:

- Für $i = 1$: Eine Division.
- Für $i > 1$: Eine Division, $i - 1$ Multiplikationen und $i - 1$ Subtraktionen.

Zur Durchführung des Algorithmus 3.2 benötigen wir also

$$1 + \sum_{i=2}^{n} (2(i-1) + 1) = n^2$$

Rechenoperationen.
Sei nun $R \in \mathbb{R}^{n\times n}$ eine obere Dreiecksmatrix:

$$R = \begin{pmatrix} r_{11} & \cdots & r_{1n} \\ & \ddots & \vdots \\ 0 & & r_{nn} \end{pmatrix}.$$

Das lineare Gleichungssystem lautet

$$Rx = b,$$

was äquivalent ist zu

$$\begin{aligned} r_{11}x_1 + r_{12}x_2 + \cdots + r_{1n}x_n &= b_1 \\ r_{22}x_2 + \cdots + r_{2n}x_n &= b_2 \\ \ddots &\quad \vdots \\ r_{nn}x_n &= b_n. \end{aligned}$$

Dieses System lässt sich explizit lösen, indem man zunächst x_n aus der n-ten Gleichung bestimmt, danach ergibt sich x_{n-1} aus der vorletzten Gleichung.

Algorithmus 3.3 Rückwärtssubstitution für $Rx = b$ mit einer oberen Dreiecksmatrix $R \in \mathbb{R}^{n\times n}$

1: $x_n := b_n / r_{nn}$
2: **for** $i = (n-1) : -1 : 1$ **do**
3: $\quad x_i := (b_i - \sum_{j=i+1}^{n} r_{ij} x_j) / r_{ii}$
4: **end for**

Zur Durchführung des Algorithmus 3.3 benötigen wir ebenso n^2 Rechenoperationen. Dieser Algorithmus ist durchführbar, wenn $r_{ii} \neq 0 \quad \forall i = 1, \ldots, n$.
Zum Abschluss betrachten wir ein lineares Gleichungssystem mit einer Permutationsmatrix.

Definition 3.1 (Permutationsmatrix)**.** Mit $e_i \in \mathbb{R}^n$ bezeichnen wir den i-ten Einheitsvektor. Eine Matrix $P \in \mathbb{R}^{n\times n}$ heißt *Permutationsmatrix*, falls P die folgende Gestalt hat:

$$P = \begin{pmatrix} \text{——} & (e_{\pi(1)})^T & \text{——} \\ \text{——} & (e_{\pi(2)})^T & \text{——} \\ & \vdots & \\ \text{——} & (e_{\pi(n)})^T & \text{——} \end{pmatrix}$$

mit einer bijektiven Abbildung (Permutation)

$$\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}.$$

Notation: $P = P(\pi)$.

Beispiel 3.2. Die folgenden zwei Matrizen sind Beispiele für Permutationsmatrizen im $\mathbb{R}^{3\times 3}$:

$$P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{oder} \quad P = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Bemerkung 3.3. Jede Permutationsmatrix ist orthogonal, also $P^T P = P P^T = I$.

Betrachte das Gleichungssystem

$$Px = b$$

mit einer Permutationsmatrix $\mathbb{R}^{n\times n} \ni P = P(\pi)$ und

$$\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$$

ist eine bijektive Abbildung. Dieses System ist äquivalent zu

$$\begin{pmatrix} \text{——} & (e_{\pi(1)})^T & \text{——} \\ \text{——} & (e_{\pi(2)})^T & \text{——} \\ & \vdots & \\ \text{——} & (e_{\pi(n)})^T & \text{——} \end{pmatrix} x = b.$$

Äquivalent lässt sich das schreiben als:

$$\begin{aligned} e_{\pi(i)}^T x &= b_i \\ x_{\pi(i)} &= b_i \quad \forall i = 1, \ldots, n. \end{aligned}$$

Hieraus erhalten wir:

Algorithmus 3.4 Lösung von $Px = b$ mit einer Permutationsmatrix $P = P(\pi)$

1: **for** $i = 1 : n$ **do**
2: $\quad x_{\pi(i)} := b_i$
3: **end for**

Analog lässt sich die Lösung von

$$P^T x = b$$

wie folgt berechnen:

$$P^T x = b \quad \Rightarrow \quad PP^T x = Pb \quad \Rightarrow \quad x = Pb.$$

Algorithmus 3.5 Lösung von $P^T x = b$ mit einer Permutationsmatrix $P = P(\pi)$

1: **for** $i = 1 : n$ **do**
2: $\quad x_i := b_{\pi(i)}$
3: **end for**

3.2 LR-Zerlegung ohne Pivotisierung

Im Folgenden sei $A \in \mathbb{R}^{n \times n}$ regulär und $b \in \mathbb{R}^n$. Gesucht sei die eindeutige Lösung von

$$Ax = b,$$

mit Hilfe des Eliminationsverfahrens von Gauß. Diese Methode führt auf eine Zerlegung von A der Form

$$A = LR$$

mit einer normierten unteren Dreiecksmatrix

$$L = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ * & & 1 \end{pmatrix} \in \mathbb{R}^{n \times n}$$

und einer oberen Dreiecksmatrix

$$R = \begin{pmatrix} * & \cdots & * \\ & \ddots & \vdots \\ 0 & & * \end{pmatrix} \in \mathbb{R}^{n \times n}.$$

Definition 3.4 (LR-Zerlegung)**.** Unter einer *LR-Zerlegung* einer Matrix versteht man eine Zerlegung der Form

$$A = LR$$

mit einer normierten unteren Dreiecksmatrix $L \in \mathbb{R}^{n \times n}$ und einer oberen Dreiecksmatrix $R \in \mathbb{R}^{n \times n}$.

Mit Hilfe der LR-Zerlegung lässt sich $Ax = b$ sehr einfach lösen. Denn:

$$Ax = b \quad \Leftrightarrow \quad L(Rx) = b.$$

Dazu löst man

$$Ly = b$$

mittels Vorwärtseinsetzen und dann

$$Rx = y$$

durch Rückwärtssubstitution.

Algorithmus 3.6 Lösung von $Ax = b$ mittels LR-Zerlegung

(S1) Bestimme LR-Zerlegung
(S2) Löse $Ly = b$ durch Vorwärtseinsetzen (Algorithmus 2.2)
(S3) Löse $Rx = y$ durch Rückwärtssubstitution (Algorithmus 2.3)

Definition 3.5 (Frobenius-Matrix)**.** Eine Matrix $L_i \in \mathbb{R}^{n \times n}$ heißt *Frobenius-Matrix*, wenn L_i die folgende Form hat:

$$L_i = \begin{pmatrix} 1 & & & 0 \\ & \ddots & & \\ & & 1 & \\ & & -l_{i+1,i} & \ddots \\ & & \vdots & \\ & & -l_{n,i} & 1 \end{pmatrix}.$$

Beispiel 3.6. Betrachte das lineare Gleichungssystem $Ax = b$ mit

$$\begin{pmatrix} 2 & 1 & 1 & 0 \\ 4 & 3 & 3 & 1 \\ 8 & 7 & 9 & 5 \\ 6 & 7 & 9 & 8 \end{pmatrix} \quad \text{und} \quad b \in \mathbb{R}^n.$$

Das System $Ax = b$ ist äquivalent zu

$$\begin{aligned} 2x_1 + 1x_2 + 1x_3 + 0x_4 &= b_1 \\ 4x_1 + 3x_2 + 3x_3 + 1x_4 &= b_2 \\ 8x_1 + 7x_2 + 9x_3 + 5x_4 &= b_3 \\ 6x_1 + 7x_2 + 9x_3 + 8x_4 &= b_4. \end{aligned}$$

Definiere

$$L_1 := \begin{pmatrix} 1 & & & \\ -2 & 1 & & \\ -4 & & 1 & \\ -3 & & & 1 \end{pmatrix}.$$

Dann ist

$$\begin{aligned} L_1 A &= \begin{pmatrix} 1 & & & \\ -2 & 1 & & \\ -4 & & 1 & \\ -3 & & & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 & 1 & 0 \\ 4 & 3 & 3 & 1 \\ 8 & 7 & 9 & 5 \\ 6 & 7 & 9 & 8 \end{pmatrix} = \begin{pmatrix} 2 & 1 & 1 & 0 \\ & 1 & 1 & 1 \\ & 3 & 5 & 5 \\ & 4 & 6 & 8 \end{pmatrix} =: A^{(2)} \\ L_2 A^{(2)} &= \underbrace{\begin{pmatrix} 1 & & & \\ & 1 & & \\ & -3 & 1 & \\ & -4 & & 1 \end{pmatrix}}_{=:L_2} \begin{pmatrix} 2 & 1 & 1 & 0 \\ & 1 & 1 & 1 \\ & 3 & 5 & 5 \\ & 4 & 6 & 8 \end{pmatrix} = \begin{pmatrix} 2 & 1 & 1 & 0 \\ & 1 & 1 & 1 \\ & & 2 & 2 \\ & & 2 & 4 \end{pmatrix} =: A^{(3)} \\ L_3 A^{(3)} &= \underbrace{\begin{pmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & -1 & 1 \end{pmatrix}}_{=:L_3} \begin{pmatrix} 2 & 1 & 1 & 0 \\ & 1 & 1 & 1 \\ & & 2 & 2 \\ & & 2 & 4 \end{pmatrix} = \begin{pmatrix} 2 & 1 & 1 & 0 \\ & 1 & 1 & 1 \\ & & 2 & 2 \\ & & & 2 \end{pmatrix} = R. \end{aligned}$$

Daraus folgt $A = L_1^{-1} L_2^{-1} L_3^{-1} R$ mit Frobeniusmatrizen L_1, L_2 und L_3. Die Inversen von L_i lauten:

$$L_1^{-1} = \begin{pmatrix} 1 & & & \\ 2 & 1 & & \\ 4 & & 1 & \\ 3 & & & 1 \end{pmatrix}, \quad L_2^{-1} = \begin{pmatrix} 1 & & & \\ & 1 & & \\ & 3 & 1 & \\ & 4 & & 1 \end{pmatrix}, \quad L_3^{-1} = \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & 1 & 1 \end{pmatrix}.$$

Damit ergibt sich:

$$L = L_1^{-1}L_2^{-1}L_3^{-1} = \begin{pmatrix} 1 & & & \\ 2 & 1 & & \\ 4 & 3 & 1 & \\ 3 & 4 & 1 & 1 \end{pmatrix}$$

und

$$A = LR.$$

Lemma 3.7. *Es sei $l_i := (0, \ldots, 0, l_{i+1,i}, \ldots, l_{n,i}) \in \mathbb{R}^n$. Definiere die Frobenius-Matrix wie folgt:*

$$L_i := I - l_i e_i^T = \begin{pmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{pmatrix} - \begin{pmatrix} 0 \\ \vdots \\ 0 \\ l_{i+1,i} \\ \vdots \\ l_{n,i} \end{pmatrix} (0, \ldots, 1, \ldots, 0)$$

$$= \begin{pmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ & & -l_{i+1,i} & \ddots \\ & & \vdots & \\ & & -l_{n,i} & & 1 \end{pmatrix}.$$

Dann gilt:

(i) $L_i^{-1} = I + l_i e_i^T = \begin{pmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ & & l_{i+1,i} & \ddots \\ & & \vdots & \\ & & l_{n,i} & & 1 \end{pmatrix}$,

(ii) $L_1^{-1} \cdot L_2^{-1} \cdot \ldots \cdot L_k^{-1} = I + \sum_{i=1}^{k} l_i e_i^T = \begin{pmatrix} 1 & & & & \\ l_{i,1} & \ddots & & & \\ \vdots & & 1 & & \\ l_{k+1,1} & \cdots & l_{k+1,k} & \ddots & \\ \vdots & & \vdots & & \ddots \\ l_{n,1} & \cdots & l_{n,k} & & 1 \end{pmatrix}$

für $k = 1, \ldots, n$.

Beweis.

Zu (i): $\underbrace{(I - l_i e_i^T)}_{=L_i}(I + l_i e_i^T) = I - l_i e_i^T + l_i e_i^T - l_i \underbrace{e_i^T l_i}_{=0} e_i^T = I.$

Zu (ii): Beweis durch vollständige Induktion.

Induktionsanfang k=1: Wurde bereits in (i) gezeigt.

Induktionsannahme: Die Aussage gelte für ein festes aber beliebiges k mit $1 \leq k \leq n-2$. Wir zeigen, dass die Aussage für $k+1$ gilt.

$$\begin{aligned} L_1^{-1} \cdot L_2^{-1} \cdot \ldots \cdot L_{k+1}^{-1} &= (I + \sum_{i=1}^{k} l_i e_i^T) L_{k+1}^{-1} \\ &\overset{(i)}{=} (I + \sum_{i=1}^{k} l_i e_i^T)(I + l_{k+1} e_{k+1}^T) \\ &= I + l_{k+1} e_{k+1}^T + \sum_{i=1}^{k} l_i e_i^T + \sum_{i=1}^{k} l_i \underbrace{e_i^T l_{k+1}}_{=0} e_{k+1}^T \\ &= I + \sum_{i=1}^{k+1} l_i e_i^T. \end{aligned}$$

□

Algorithmus 3.7 Grundversion einer LR-Zerlegung

1: Setze $A^{(1)} := A$
2: **for** $i = 1 : (n-1)$ **do**
3: Definiere eine Frobeniusmatrix $L_i = \begin{pmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ & & -l_{i+1,i} & \ddots \\ & & \vdots & \\ & & -l_{n,i} & 1 \end{pmatrix}$

mit $l_{j,i} := a_{j,i}^{(i)} / a_{i,i}^{(i)}$ für $j = i+1, \ldots, n$.
Setze $A^{(i+1)} := L_i A^{(i)}$
4: **end for**
5: Setze $R := A^{(n)}$, $L := L_1^{-1} \cdot \ldots \cdot L_{n-1}^{-1}$

Bemerkung 3.8. Der Algorithmus 3.7 liefert eine LR-Zerlegung, wenn er durchführbar ist, denn

$$L_{n-1} \cdot \ldots \cdot L_1 A = R \quad \Leftrightarrow \quad A = L_1^{-1} \cdot \ddots \cdot L_{n-1}^{-1} R = LR.$$

Dabei ist L eine untere normierte Dreiecksmatrix und R eine obere Dreiecksmatrix. Der Algorithmus ist durchführbar, wenn

$$a_{ii}^{(i)} \neq 0 \quad \forall i = 1, \ldots, n-1.$$

Definition 3.9. Das i-te Diagonalelement $a_{ii}^{(i)}$ der Matrix

$$A^{(i)} = L_{i-1} \cdot \ldots \cdot L_1 A$$

heißt **Pivot-Element**.

Definition 3.10. Für eine quadratische Matrix $A \in \mathbb{R}^{n\times n}$ heißt

$$A[k] := \begin{pmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & \ddots & \vdots \\ a_{k1} & \cdots & a_{kk} \end{pmatrix} \in \mathbb{R}^{k\times k}, \quad k \in \{1, \ldots, n\}$$

die führende $k-k-$Hauptachsenabschnittsmatrix (Hauptabschnittsmatrix) von A. Die Determinante $\det(A[k])$ heißt die führende $k-$te Hauptachsenabschnittsdeterminante von A.

Beispiel 3.11. Es sei

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

Dann ist

$$A[1] = (1), \quad A[2] = \begin{pmatrix} 1 & 2 \\ 4 & 5 \end{pmatrix}, \quad A[3] = A.$$

Lemma 3.12. *Es sei $A \in \mathbb{R}^{n\times n}$ und $L \in \mathbb{R}^{n\times n}$ eine untere (nicht notwendigerweise normierte) Dreiecksmatrix. Dann gilt*

$$(LA)[k] = L[k]A[k] \quad \forall k = 1, \ldots, n.$$

Beweis. Sei $k \in \{1, \ldots, n\}$ beliebig aber fest gewählt. Für $i, j \in \{1, \ldots, k\}$ gilt

$$((LA)[k])_{i,j} = \sum_{m=1}^{n} l_{i,m} a_{m,j} = \sum_{m=1}^{k} l_{i,m} a_{m,j} + \sum_{m=k+1}^{n} \underbrace{l_{i,m}}_{=0} a_{m,j},$$

denn $L \in \mathbb{R}^{n\times n}$ ist eine untere Dreiecksmatrix, also $l_{i,j} = 0 \quad \forall j > i$. Also folgt

$$((LA)[k])_{i,j} = \sum_{m=1}^{k} l_{i,m} a_{m,j} = (L[k]A[k])_{i,j}.$$

□

Satz 3.13 (Existenz einer LR-Zerlegung). *Es sei $A \in \mathbb{R}^{n\times n}$ eine reguläre Matrix. Dann besitzt A genau dann eine LR-Zerlegung, wenn*

$$\det(A[k]) \neq 0 \quad \forall k = 1, \ldots, n.$$

Beweis.

„⇒“: A habe eine LR-Zerlegung, das heißt

$$A = LR$$

mit einer normierten unteren Dreiecksmatrix $L \in \mathbb{R}^{n\times n}$ und einer oberen Dreiecksmatrix $R \in \mathbb{R}^{n\times n}$. Dann gilt

$$\det(A) = \det(L)\det(R) \quad \text{(Determinantenmultiplikationssatz).}$$

Da A regulär ist, gilt

$$0 \neq \det(A) \quad \Rightarrow \quad \det(L) \neq 0 \text{ und } \det(R) \neq 0.$$

Hieraus folgt

$$\det(L[k]) \neq 0 \quad \text{und} \quad \det(R[k]) \neq 0 \quad \forall k = 1, \ldots, n,$$

denn

$$\det \begin{pmatrix} \lambda_1 & & * \\ & \ddots & \\ & & \lambda_n \end{pmatrix} = \lambda_1 \cdot \ldots \cdot \lambda_n \quad \text{und} \quad \det \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ * & & \lambda_n \end{pmatrix} = \lambda_1 \cdot \ldots \cdot \lambda_n.$$

Das Lemma 3.12 liefert

$$\det(A[k]) \overset{\text{Lemma 3.12}}{=} \det(L[k]R[k]) = \det(L[k])\det(R[k]) \neq 0 \quad \forall k = 1, \ldots, n.$$

„⇐“: Es gelte

$$\det(A[k]) \neq 0 \quad \forall k = 1, \ldots, n.$$

Wir zeigen nun, dass A eine LR-Zerlegung besitzt. Dazu zeigen wir, dass der Algorithmus 3.7 durchführbar ist, also dass

$$a_{ii}^{(i)} \neq 0 \quad \forall i = 1, \ldots, n$$

gilt. Beweis durch vollständige Induktion:

Induktionsanfang i=1: $a_{11}^{(1)} \overset{Def}{=} \det(A[1]) \neq 0$ (Voraussetzung)

Induktionsannahme: Es gelte $a_{11}^{(1)} \cdot \ldots \cdot a_{ii}^{(i)} \neq 0$ für ein festes aber beliebiges i mit $1 \leq i \leq n-2$. Nun zeigen wir:

$$a_{i+1,i+1}^{(i+1)} \neq 0.$$

Aufgrund der Induktionsannahme existiert

$$A^{(i+1)} = L_i \cdot \ldots \cdot L_1 A \quad \text{(siehe Algorithmus 3.7).}$$

Hieraus folgt

$$A^{(i+1)}[i+1] \overset{\text{Lemma } 3.12}{=} L_i[i+1] \cdot \ldots \cdot L_1[i+1]A[i+1].$$

Daher ist

$$\det(A^{(i+1)}[i+1]) = \underbrace{\det(L_i[i+1])}_{=1} \cdot \ldots \cdot \underbrace{\det(L_1[i+1])}_{=1} \underbrace{\det(A[i+1])}_{\neq 0} \neq 0.$$

Laut Konstruktion (Algorithmus 3.7) ist

$$A^{i+1}[i+1] = \begin{pmatrix} a_{11}^{(i+1)} & & * \\ & \ddots & \\ & & a_{i+1,i+1}^{(i+1)} \end{pmatrix}.$$

Damit folgt

$$0 \neq \det(A^{(i+1)}[i+1]) = a_{11}^{(i+1)} \cdot \ldots \cdot a_{i+1,i+1}^{(i+1)}.$$

Insgesamt ergibt sich

$$a_{i+1,i+1}^{(i+1)} \neq 0$$

und daraus folgt die Behauptung.

□

Wir zeigen nun, dass es genau eine LR-Zerlegung gibt, falls diese existiert. Dazu verwenden wir folgende Lemmata.

Lemma 3.14. *Es gelten die folgenden Aussagen:*

(i) Sind $L^{(1)}, L^{(2)} \in \mathbb{R}^{n\times n}$ zwei untere (normierte) Dreiecksmatrizen, so ist das Produkt $L := L^{(1)}L^{(2)}$ eine untere (normierte) Dreiecksmatrix.

(ii) Sind $R^{(1)}, R^{(2)} \in \mathbb{R}^{n\times n}$ zwei obere Dreiecksmatrizen, so ist das Produkt $R := R^{(1)}R^{(2)}$ eine obere Dreiecksmatrix.

Beweis. Seien $L^{(1)}, L^{(2)} \in \mathbb{R}^{n\times n}$ untere Dreiecksmatrizen, setze $L := L^{(1)}L^{(2)}$. Sei $i \in \{1, \ldots, n-1\}$ beliebig aber fest. Dann gilt für $j > i$:

$$\begin{aligned} l_{im}^{(1)} &= 0 \quad \forall m = j, \ldots, n \\ l_{mj}^{(2)} &= 0 \quad \forall m = 1, \ldots, j-1, \end{aligned}$$

da $L^{(1)}$ und $L^{(2)}$ untere Dreiecksmatrizen sind. Somit gilt für $j > i$:

$$l_{ij} = \sum_{m=1}^{n} l_{im}^{(1)} l_{mj}^{(2)} = \sum_{m=1}^{j-1} l_{im}^{(1)} \underbrace{l_{mj}^{(2)}}_{=0} + \sum_{m=j}^{n} \underbrace{l_{im}^{(1)}}_{=0} l_{mj}^{(2)}.$$

Daraus folgt

$$l_{ij} = 0 \quad \forall j > i.$$

Also ist L eine untere Dreiecksmatrix.
Sind $L^{(1)}$ und $L^{(2)}$ zusätzlich normiert, d.h. $l_{ii}^{(1)} = l_{ii}^{(2)} = 1 \quad \forall i = 1, \ldots, n$, so folgt

$$\begin{aligned} l_{ii} = \sum_{m=1}^{n} l_{im}^{(1)} l_{mj}^{(2)} &= \sum_{m=1}^{i-1} l_{im}^{(1)} \underbrace{l_{mj}^{(2)}}_{=0} + l_{ii}^{(1)} l_{ii}^{(2)} + \sum_{m=i+1}^{n} \underbrace{l_{im}^{(1)}}_{=0} l_{mj}^{(2)} \\ &= l_{ii}^{(1)} l_{ii}^{(2)} = 1 \quad \forall i = 1, \ldots, n. \end{aligned}$$

Der Beweis der Aussage (ii) erfolgt analog. □

Lemma 3.15. *Es gelten die folgenden Aussagen:*

(i) Ist $L \in \mathbb{R}^{n \times n}$ eine reguläre untere Dreiecksmatrix, so ist L^{-1} auch eine untere Dreiecksmatrix. Ist L zusätzlich normiert, so ist auch L^{-1} normiert.

(ii) Ist $R \in \mathbb{R}^{n \times n}$ eine reguläre obere Dreiecksmatrix, so ist R^{-1} auch eine obere Dreiecksmatrix. Ist R zusätzlich normiert, so ist auch R^{-1} normiert.

Beweis. Da L regulär ist, existiert die Inverse

$$L^{-1} := X = \begin{pmatrix} | & | & & | \\ x_1 & x_2 & \ldots & x_n \\ | & | & & | \end{pmatrix}.$$

Dabei ist $x_i \in \mathbb{R}^n$ die i-te Spalte von X. Laut Definition gilt

$$LL^{-1} = LX = I \quad \Leftrightarrow \quad Lx_i = e_i \quad \forall i = 1, \ldots, n. \tag{3.1}$$

Wir zerlegen die Matrix L und die Vektoren x_i und e_i wie folgt:

$$L = \begin{pmatrix} L_{11}^{(i)} & 0 \\ L_{21}^{(i)} & L_{22}^{(i)} \end{pmatrix}, x_i = \begin{pmatrix} x_1^{(i)} \\ x_2^{(i)} \end{pmatrix}, e_i = \begin{pmatrix} 0 \\ e_1^{(i)} \end{pmatrix},$$

mit

$$L_{11}^{(i)} \in \mathbb{R}^{(i-1)\times(i-1)}, \quad L_{21}^{(i)} \in \mathbb{R}^{(n-i+1)\times(i-1)}, \quad L_{22}^{(i)} \in \mathbb{R}^{(n-i+1)\times(n-i+1)}$$

sowie

$$x_1^{(i)} \in \mathbb{R}^{i-1}, \quad x_2^{(i)} \in \mathbb{R}^{n-i+1} \quad \text{und} \quad e_i \in \mathbb{R}^{n-i+1}.$$

Dabei ist also e_i der 1-te Einheitsvektor des $\mathbb{R}^{n-i+1}$. Auf diese Weise lässt sich (3.1) wie folgt darstellen:

$$\begin{pmatrix} L_{11}^{(i)} & 0 \\ L_{21}^{(i)} & L_{22}^{(i)} \end{pmatrix} \begin{pmatrix} x_1^{(i)} \\ x_2^{(i)} \end{pmatrix} = \begin{pmatrix} 0 \\ e_1^{(i)} \end{pmatrix}.$$

Hieraus folgt

$$L_{11}^{(i)} x_1^{(i)} = 0 \quad \overset{L_{11}^{(i)} \text{ ist regulär}}{\Rightarrow} \quad x_1^{(i)} = 0.$$

Das heißt: Jede i-te Spalte von $X = L^{-1}$ besitzt in den ersten $i-1$ Einträgen lauter Nullen. Folglich ist

$$L^{-1} = X = \begin{pmatrix} * & & 0 \\ \vdots & \ddots & \\ * & \cdots & * \end{pmatrix}$$

eine untere Dreiecksmatrix. Ist L zusätzlich normiert, das heißt $l_{ii} = 1 \quad \forall i = 1, \ldots, n$, so ergibt sich aus der ersten Gleichung von

$$L_{22}^{(i)} x_2^{(i)} = e_1^{(i)},$$

dass das i-te Element von x_i eine Eins sein muss:

$$x_i = \begin{pmatrix} 0 \\ x_2^{(i)} \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ * \\ \vdots \\ * \end{pmatrix} \leftarrow i\text{-te Zeile} \quad .$$

Der Beweis der Aussage (ii) folgt analog. □

Satz 3.16 (Eindeutigkeit der LR-Zerlegung)**.** *Sei $A \in \mathbb{R}^{n \times n}$ regulär mit*

$$\det(A[k]) \neq 0 \quad \forall k = 1, \ldots, n.$$

Dann existiert genau eine LR-Zerlegung von A, d.h.

$$A = LR$$

mit einer eindeutigen unteren normierten Dreiecksmatrix $L \in \mathbb{R}^{n \times n}$ und einer eindeutigen oberen Dreiecksmatrix $R \in \mathbb{R}^{n \times n}$.

Beweis. Die Existenz wurde bereits gezeigt. Sei nun

$$A = L_1 R_1 \quad \text{und} \quad A = L_2 R_2$$

mit unteren normierten Dreiecksmatrizen $L_1, L_2 \in \mathbb{R}^{n \times n}$ und oberen Dreiecksmatrizen $R_1, R_2 \in \mathbb{R}^{n \times n}$. Dann gilt

$$\begin{aligned} L_1 R_1 &= L_2 R_2 \\ \Rightarrow \quad L_1 &= L_2 R_2 R_1^{-1} \\ \Rightarrow L_2^{-1} L_1 &= R_2 R_1^{-1}. \end{aligned}$$

Die vorherigen Lemmata liefern:

- $L_2^{-1}L_1$ ist eine untere normierte Dreiecksmatrix.
- $R_2R_1^{-1}$ ist eine obere Dreiecksmatrix.

Damit ist

$$L_2^{-1}L_1 = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ * & & 1 \end{pmatrix} = \begin{pmatrix} * & \cdots & * \\ & \ddots & \vdots \\ 0 & & * \end{pmatrix} = R_2R_1^{-1}$$

und daraus ergibt sich

$$L_2^{-1}L_1 = I \quad \text{und} \quad R_2R_1^{-1} = I.$$

Das ist gleichbedeutend mit

$$L_2^{-1} = L_1^{-1}, \text{ also } L_2 = L_1 \quad \text{und} \quad R_1 = R_2.$$

Insgesamt folgt also die Behauptung. □

Bemerkung 3.17. Der Beweis zeigt, dass die LR-Zerlegung immer eindeutig ist, sofern diese für eine reguläre Matrix existiert.

Bemerkung 3.18. Beachte, dass man in jeder Iteration des Algorithmus 3.7 eine Frobenius-Matrix L_i abspeichern muss. Zur Vermeidung der Abspeicherung von Frobenius-Matrizen betrachten wir das folgende Verfahren.

Algorithmus 3.8 Gauß-Elimination ohne Pivotisierung

```
1: for k = 1 : (n − 1) do
2:     for i = (k + 1) : n do
3:         a_ik := a_ik / a_kk
4:         for j = (k + 1) : n do
5:             a_ij := a_ij − a_ik a_kj
6:         end for
7:     end for
8: end for
```

Ist Algorithmus 3.8 durchführbar, so erhalten wir am Ende des Prozesses die folgende Matrix:

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} \longrightarrow \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{nn} \\ l_{21} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ l_{n1} & \cdots & l_{n,m-1} & r_{nn} \end{pmatrix}.$$

Dann ist

$$A = LR$$

mit

$$L = \begin{pmatrix} 1 & & 0 \\ l_{21} & \ddots & \\ \vdots & & \ddots \\ l_{n1} & \cdots & l_{n,m-1} \quad 1 \end{pmatrix} \quad \text{und} \quad R = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{nn} \\ & \ddots & & \vdots \\ & & \ddots & \vdots \\ & & & r_{nn} \end{pmatrix}.$$

Zur Durchführung des Algorithmus 3.8 benötigt man $\approx \frac{2}{3}n^3$ Rechenoperationen.

3.3 LR-Zerlegung mit Pivotisierung

Im vorherigen Abschnitt haben wir eine notwendige und hinreichende Bedingung zur Existenz und Eindeutigkeit der LR-Zerlegung für eine reguläre Matrix $A \in \mathbb{R}^{n\times n}$ hergeleitet. Diese lautet

$$\det(A[k]) \neq 0 \quad \forall k = 1, \ldots, n.$$

Diese Bedingung ist aber zu stark und wird oft verletzt. Ein einfaches Beispiel dafür ist

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

Diese Matrix hat keine LR-Zerlegung, ist jedoch regulär. Also gibt es zu jedem Vektor $b \in \mathbb{R}^2$ genau eine Lösung $x \in \mathbb{R}^2$ von

$$Ax = b.$$

Unsere Idee besteht darin, die Zeilen von A durch eine Permutationsmatrix P zu vertauschen, so dass PA eine LR-Zerlegung besitzt, d.h.

$$PA = LR$$

mit einer unteren normierten Dreiecksmatrix $L \in \mathbb{R}^{n\times n}$ und einer oberen Dreiecksmatrix $R \in \mathbb{R}^{n\times n}$. Für unser Beispiel ist

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

Dann ergibt sich

$$PA = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}}_{=L} \underbrace{\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}}_{=R}$$

als LR-Zerlegung von PA.

Satz 3.19. *Es sei $A \in \mathbb{R}^{n\times n}$ regulär. Dann existiert eine Permutationsmatrix $P \in \mathbb{R}^{n\times n}$, so dass gilt*

$$PA = LR,$$

mit einer unteren normierten Dreiecksmatrix $L \in \mathbb{R}^{n\times n}$ und einer oberen Dreiecksmatrix $R \in \mathbb{R}^{n\times n}$.

Beweis durch Induktion nach n.

Induktionsanfang $n = 1$: Wähle $P = L = I = \begin{pmatrix}1\end{pmatrix}$ und $R = A$. Dann gilt

$$PA = LR.$$

Induktionsannahme: Die Behauptung gelte für alle regulären Matrizen der Dimension $(n-1)\times(n-1)$ für ein festes $n \geq 2$.

Wir zeigen nun, dass die Aussage für alle regulären Matrizen der Dimension $n \times n$ gilt. Sei also $A \in \mathbb{R}^{n\times n}$ regulär. Die erste Spalte von A enthält mindestens ein von Null verschiedenes Element, da A regulär ist. Somit existiert eine Permutationsmatrix $P_n \in \mathbb{R}^{n\times n}$, so dass

$$P_n A = \begin{pmatrix}\alpha & r^T\\ l & C\end{pmatrix}$$

mit $\alpha \in \mathbb{R}\setminus\{0\}$, $r, l \in \mathbb{R}^{n-1}$ und $C \in \mathbb{R}^{(n-1)\times(n-1)}$. Wir definieren nun

$$B = C - \frac{1}{\alpha} l r^T \in \mathbb{R}^{(n-1)\times(n-1)}.$$

Dann ist

$$P_n A = \begin{pmatrix}\alpha & r^T\\ l & C\end{pmatrix} = \begin{pmatrix}1 & 0\\ \frac{1}{\alpha}l & I\end{pmatrix}\begin{pmatrix}\alpha & r^T\\ 0 & B\end{pmatrix}.$$

Ferner gilt

$$\det(P_n A) = \underbrace{\det(P_n)}_{\neq 0}\underbrace{\det(A)}_{\neq 0} \neq 0.$$

Also ist

$$\begin{aligned} 0 \neq \det(P_n A) = \det\begin{pmatrix}1 & 0\\ \frac{1}{\alpha}l & I\end{pmatrix}\det\begin{pmatrix}\alpha & r^T\\ 0 & B\end{pmatrix} &= 1\det\begin{pmatrix}\alpha & r^T\\ 0 & B\end{pmatrix}\\ &= \alpha\det(B).\end{aligned}$$

Damit ist auch

$$\det(B) \neq 0.$$

Also ist $B \in \mathbb{R}^{(n-1)\times(n-1)}$ regulär. Laut Induktionsannahme existieren eine Permutationsmatrix $P_{n-1} \in \mathbb{R}^{(n-1)\times(n-1)}$, eine normierte untere Dreiecksmatrix

$L_{n-1} \in \mathbb{R}^{(n-1)\times(n-1)}$ sowie eine obere Dreiecksmatrix $R_{n-1} \in \mathbb{R}^{(n-1)\times(n-1)}$, so dass

$$P_{n-1}B = L_{n-1}R_{n-1}.$$

Auf Grund der Orthogonalität der Permutationsmatrix P_{n-1} ist dies äquivalent zu

$$B = P_{n-1}^T L_{n-1} R_{n-1}.$$

Hieraus folgt

$$\begin{aligned} P_n A = \begin{pmatrix} 1 & 0 \\ \frac{1}{\alpha} l & I \end{pmatrix} \begin{pmatrix} \alpha & r^T \\ 0 & P_{n-1}^T L_{n-1} R_{n-1} \end{pmatrix} &= \begin{pmatrix} 1 & 0 \\ \frac{1}{\alpha} l & I \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & P_{n-1}^T L_{n-1} \end{pmatrix} \begin{pmatrix} \alpha & r^T \\ 0 & R_{n-1} \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 \\ \frac{1}{\alpha} l & P_{n-1}^T L_{n-1} \end{pmatrix} \begin{pmatrix} \alpha & r^T \\ 0 & R_{n-1} \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 \\ 0 & P_{n-1}^T \end{pmatrix} \begin{pmatrix} 1 & 0 \\ \frac{1}{\alpha} P_{n-1} l & L_{n-1} \end{pmatrix} \begin{pmatrix} \alpha & r^T \\ 0 & R_{n-1} \end{pmatrix}. \end{aligned}$$

Somit ist

$$P_n A = \begin{pmatrix} 1 & 0 \\ 0 & P_{n-1}^T \end{pmatrix} \begin{pmatrix} 1 & 0 \\ \frac{1}{\alpha} P_{n-1} l & L_{n-1} \end{pmatrix} \begin{pmatrix} \alpha & r^T \\ 0 & R_{n-1} \end{pmatrix},$$

und daher folgt

$$\begin{pmatrix} 1 & 0 \\ 0 & P_{n-1}^T \end{pmatrix}^T P_n A = \begin{pmatrix} 1 & 0 \\ \frac{1}{\alpha} P_{n-1} l & L_{n-1} \end{pmatrix} \begin{pmatrix} \alpha & r^T \\ 0 & R_{n-1} \end{pmatrix}.$$

Setzen wir nun

$$\begin{aligned} P &:= \begin{pmatrix} 1 & 0 \\ 0 & P_{n-1}^T \end{pmatrix}^T P_n && \text{(Permutationsmatrix)} \\ L &:= \begin{pmatrix} 1 & 0 \\ \frac{1}{\alpha} P_{n-1} l & L_{n-1} \end{pmatrix} && \text{(untere normierte Dreiecksmatrix)} \\ R &:= \begin{pmatrix} \alpha & r^T \\ 0 & R_{n-1} \end{pmatrix} && \text{(obere Dreiecksmatrix),} \end{aligned}$$

so erhalten wir die Zerlegung

$$PA = LR.$$

□

Mit der Zerlegung $PA = LR$ lässt sich das lineare Gleichungssystem

$$Ax = b$$

einfach lösen:

$$Ax = b \quad \Leftrightarrow \quad PAx = Pb \quad \Leftrightarrow \quad LRx = Pb.$$

Algorithmus 3.9 Lösung von $Ax = b$ durch LR-Zerlegung mit Pivotisierung

(S1) Bestimme eine Zerlegung $PA = LR$ von A.
(S2) Setze $c := Pb$ (Algorithmus 2.5).
(S3) Bestimme y als Lösung von $Ly = c$ (Vorwärtseinsetzen).
(S4) Bestimme x als Lösung von $Rx = y$ (Rückwärtseinsetzen).

Definition 3.20. Für $i, j \in \{1, \ldots, n\}$ mit $j \geq i$ definieren wir die Matrix

$$P_{ij} := \begin{pmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & 0 & \cdots & \cdots & 0 & 1 & \\ & & & \vdots & 1 & & & 0 & \\ & & & \vdots & & \ddots & & \vdots & \\ & & & 0 & & & 1 & \vdots & \\ & & & 1 & 0 & \cdots & \cdots & 0 & \\ & & & & & & & & 1 \\ & & & & & & & & & \ddots \\ & & & & & & & & & & 1 \end{pmatrix} \in \mathbb{R}^{n \times n}.$$

Die Matrix P_{ij} entsteht aus der Einheitsmatrix $I \in \mathbb{R}^{n \times n}$ durch Vertauschung der i-ten und j-ten Zeile. Das Produkt $P_{ij}A$ liefert die Matrix, die aus A durch Vertauschung der i-ten und j-ten Zeile hervorgeht.

Beispiel 3.21. Wir betrachten die Matrix

$$A = \begin{pmatrix} 0 & 2 & -1 & -2 \\ 2 & -2 & 4 & -1 \\ 1 & 1 & 1 & 1 \\ -2 & 1 & -2 & 1 \end{pmatrix}.$$

Das erste Element der ersten Zeile der Matrix A ist Null. Also müssen wir zuerst Zeilen vertauschen, um eine LR-Zerlegung finden zu können. In unserem Fall tauschen wir nun die erste mit der zweiten Zeile:

$$P_{12}A = \begin{pmatrix} 2 & -2 & 4 & -1 \\ 0 & 2 & -1 & -2 \\ 1 & 1 & 1 & 1 \\ -2 & 1 & -2 & 1 \end{pmatrix}.$$

Definiere nun eine Frobenius-Matrix

$$L_1 = \begin{pmatrix} 1 & & & \\ 0 & 1 & & \\ -\frac{1}{2} & & 1 & \\ 1 & & & 1 \end{pmatrix},$$

dann ist

$$L_1 P_{12} A = \begin{pmatrix} 2 & -2 & 4 & -1 \\ & 2 & -1 & -2 \\ & 2 & -1 & 1.5 \\ & -1 & 2 & 0 \end{pmatrix}.$$

Das Pivot-Element ist nun 2, also insbesondere nicht 0. Wählen also als Permutationsmatrix $P_{22} = I$ und erhalten

$$P_{22} L_1 P_{12} A = \begin{pmatrix} 2 & -2 & 4 & -1 \\ & 2 & -1 & -2 \\ & 2 & -1 & 1.5 \\ & -1 & 2 & 0 \end{pmatrix}.$$

Definieren wieder eine Frobenius-Matrix

$$L_2 = \begin{pmatrix} 1 & & & \\ & 1 & & \\ & -1 & 1 & \\ & \frac{1}{2} & & 1 \end{pmatrix}.$$

Es ergibt sich also

$$L_2 P_{22} L_1 P_{12} A = \begin{pmatrix} 2 & -2 & 4 & -1 \\ & 2 & -1 & -2 \\ & & 0 & 3.5 \\ & & 1.5 & -1 \end{pmatrix}.$$

Wieder ist das Pivot-Element Null, vertausche also die dritte mit der vierten Zeile:

$$P_{34} L_2 P_{22} L_1 P_{12} A = \begin{pmatrix} 2 & -2 & 4 & -1 \\ & 2 & -1 & -2 \\ & & 1.5 & -1 \\ & & 0 & 3.5 \end{pmatrix}.$$

Abschließend wählen wir

$$L_3 = \begin{pmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{pmatrix}$$

und erhalten das Ergebnis

$$L_3 P_{34} L_2 P_{22} L_1 P_{12} A = \begin{pmatrix} 2 & -2 & 4 & -1 \\ & 2 & -1 & -2 \\ & & 1.5 & -1 \\ & & & 3.5 \end{pmatrix} =: R.$$

Das lineare Gleichungssystem $Ax = b$ für dieses Beispiel lösen wir wie folgt:

$$Ax = b \quad \Rightarrow \quad L_3 P_{34} L_2 P_{22} L_1 P_{12} Ax = L_3 P_{34} L_2 P_{22} L_1 P_{12} b =: c.$$

Daraus folgt $Rx = c$ und wir können x bestimmen.

Algorithmus 3.10 Grundversion der LR-Zerlegung mit Pivotisierung

1: Setze $A^{(1)} := A$.
2: **for** $i = 1 : (n-1)$ **do**
3: Wähle aus der i-ten Spalte von $A^{(i)}$ ein Element $a_{ji}^{(i)} \neq 0$ mit $j \geq i$.
4: Setze $\tilde{A}^{(i)} = P_{ij} A^{(i)}$.
5: Bestimme die Frobenius-Matrix $L_i \in \mathbb{R}^{n \times n}$

$$L_i = \begin{pmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ & & -l_{i+1,i} & \ddots \\ & & \vdots & \\ & & -l_{n,i} & 1 \end{pmatrix}$$

mit $l_{m,i} = \tilde{a}_{mi}^{(i)} / \tilde{a}_{ii}^{(i)}$ für $m = i+1, \ldots, n$.
6: Setze $A^{(i+1)} := L_i \tilde{A}^{(i)}$
7: **end for**
8: Setze $R := \tilde{A}^{(n)}$.

Bemerkung 3.22. Algorithmus 3.10 ist durchführbar, so lange $A \in \mathbb{R}^{n \times n}$ regulär ist. Dieser Algorithmus liefert Frobenius-Matrizen

$$L_1, \ldots, L_{n-1} \in \mathbb{R}^{n \times n}$$

sowie Permutationsmatrizen

$$P_1, \ldots, P_{n-1} \in \mathbb{R}^{n \times n}, \quad P_i = P_{ij(i)} \text{ mit } j(i) \geq i,$$

so dass

$$L_{n-1} P_{n-1} L_{n-2} P_{n-2} \cdot \ldots \cdot L_2 P_2 L_1 P_1 = R.$$

Wir wollen nun die Darstellung

$$PA = LR$$

bestimmen, und zwar aus

$$L_{n-1} P_{n-1} L_{n-2} P_{n-2} \cdot \ldots \cdot L_2 P_2 L_1 P_1 = R.$$

Für $i = 1, \ldots, n-1$ definieren wir

$$L_i' = P_{n-1} \cdot \ldots \cdot P_{i+1} L_i P_{i+1}^T \cdot \ldots \cdot P_{n-1}^T \quad (\text{wobei } P_n = I).$$

Laut Konstruktion ist

$$L_i = \begin{pmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ & & * & \ddots \\ & & \vdots & \\ & & * & 1 \end{pmatrix} = I - l_i e_i^T$$

mit einem geeigneten Vektor $l_i \in \mathbb{R}^n$. Somit ist

$$\begin{aligned} L_i' &= P_{n-1} \cdot \ldots \cdot P_{i+1}(I - l_i e_i^T) P_{i+1}^T \cdot \ldots \cdot P_{n-1}^T \\ &= \underbrace{P_{n-1} \cdot \ldots \cdot P_{i+1} P_{i+1}^T \cdot \ldots \cdot P_{n-1}^T}_{=I} - P_{n-1} \cdot \ldots \cdot P_{i+1} - l_i e_i^T P_{i+1}^T \cdot \ldots \cdot P_{n-1}^T \\ &= I - P_{n-1} \cdot \ldots \cdot P_{i+1} l_i e_i^T P_{i+1}^T \cdot \ldots \cdot P_{n-1}^T \\ &= I - l_i' e_i^T \end{aligned}$$

mit $l_i' = P_{n-1} \cdot \ldots \cdot P_{i+1} l_i$.
Also ist L_i' eine Frobenius-Matrix. Außerdem gilt nach Definition:

$$L_{n-1}' L_{n-2}' \cdot \ldots \cdot L_2' L_1' P_{n-1} \cdot \ldots \cdot P_2 P_1 = L_{n-1} P_{n-1} L_{n-2} P_{n-2} \cdot \ldots \cdot P_2 L_2 P_1 L_1.$$

Dies folgt aus

$$L_{n-1}' L_{n-2}' = \underbrace{P_n}_{=I} L_{n-1} \underbrace{P_n^T}_{=I} P_{n-1} L_{n-2} P_{n-1}^T.$$

Insgesamt ergibt sich also

$$(L_{n-1}' L_{n-2}' \cdot \ldots \cdot L_2' L_1') P_{n-1} \cdot \ldots \cdot P_2 P_1 A = R.$$

Daraus folgt

$$P_{n-1} \cdot \ldots \cdot P_2 P_1 A = (L_{n-1}' L_{n-2}' \cdot \ldots \cdot L_2' L_1')^{-1} R$$

und damit

$$PA = LR$$

mit

$$P = P_{n-1} \cdot \ldots \cdot P_2 P_1 \quad \text{und} \quad L = (L_{n-1}' L_{n-2}' \cdot \ldots \cdot L_2' L_1')^{-1}.$$

In jeder Iteration des Algorithmus 3.10 werden sowohl Frobenius-Matrizen als auch eine Permutationsmatrix abgespeichert. Zur Vermeidung der Abspeicherung von diesen Matrizen verwenden wir das Gauß-Verfahren mit Pivotisierung.

Algorithmus 3.11 Gauß-Elimination mit Pivotisierung (vgl. Algorithmus 3.8)

1: **for** $k = 1 : (n-1)$ **do**
▷ Pivotsuche
2: $z = 0$
3: **for** $i = k : n$ **do**
4: **if** $|a_{ik}| > z$ **then**
5: $z = |a_{ik}|$
6: $s = i$
7: **end if**
8: **end for**
9: $P(k) = s$
10: **if** $z = 0$ **then**
11: Warnung! (Matrix A ist singulär)
12: **end if**
▷ Vertauschung der k-ten und s-ten Zeile
13: **if** $k < s$ **then**
14: **for** $j = 1 : n$ **do**
15: $z = a_{kj}$
16: $a_{kj} = a_{sj}$
17: $a_{sj} = z$
18: **end for**
19: **end if**
▷ Berechnung der l_{ik}
20: **for** $i = (k+1) : n$ **do**
21: $a_{ik} = a_{ik}/a_{kk}$
22: **end for**
▷ Spaltenweise Berechnung von aufdatierten A
23: **for** $j = (k+1) : n$ **do**
24: **for** $i = (k+1) : n$ **do**
25: $a_{ij} = a_{ij} - a_{ik}a_{kj}$
26: **end for**
27: **end for**
28: **if** $a_{nn} = 0$ **then**
29: Warnung! (Matrix A ist singulär)
30: **end if**
31: **end for**

Algorithmus 3.12 Lösung von $Ax = b$ nach erfolgter Gauß-Elimination mit Pivotisierung

IMPORT $P(1), \ldots, P(n-1)$ und A aus Algorithmus 3.11.

▷ Berechne $b := Pb$ (vgl. Algorithmus 3.5)

1: **for** $k = 1 : (n-1)$ **do**
2: **if** $k < P(k)$ **then**
3: $z = b_k$
4: $b_k = b_{P(k)}$
5: $b_{P(k)} = z$
6: **end if**
7: **end for**

▷ Berechne $b := L^{-1}b$ (vgl. Algorithmus 3.2)

8: **for** $k = 1 : (n-1)$ **do**
9: **for** $i = (k+1) : n$ **do**
10: $b_i = b_i - a_{ik}b_k$
11: **end for**
12: **end for**

▷ Berechne $b := R^{-1}b$ (vgl. Algorithmus 3.3)

13: **for** $k = n : -1 : 1$ **do**
14: $b_k = b_k / a_{kk}$
15: **for** $i = 1 : (k-1)$ **do**
16: $b_i = b_i - a_{ik}b_k$
17: **end for**
18: **end for**

3.4 Cholesky-Zerlegung

In diesem Abschnitt betrachten wir ein lineares Gleichungssystem

$$Ax = b$$

mit einer symmetrischen und positiv definiten (s.p.d.) Matrix $A \in \mathbb{R}^{n\times n}$. Insbesondere ist A regulär und somit existiert die Zerlegung

$$PA = LR.$$

Nun zeigen wir, dass A auch die folgende Zerlegung

$$A = LL^T$$

mit einer unteren Dreiecksmatrix $L \in \mathbb{R}^{n\times n}$ besitzt.

Bemerkung 3.23. Die Matrix L muss nicht unbedingt normiert sein.

Definition 3.24. Sei $A \in \mathbb{R}^{n\times n}$ regulär. Eine Zerlegung der Gestalt

$$A = LL^T$$

mit einer regulären unteren Dreiecksmatrix $L \in \mathbb{R}^{n\times n}$ heißt *Cholesky-Zerlegung*.

Lemma 3.25 (Notwendige Bedingung für Cholesky-Zerlegung)**.** *Sei $A \in \mathbb{R}^{n\times n}$ regulär. Es existiere eine Cholesky-Zerlegung für A, d.h.*

$$A = LL^T$$

mit einer regulären unteren Dreiecksmatrix $L \in \mathbb{R}^{n\times n}$. Dann ist A s.p.d.

Beweis.

- $A^T = (LL^T)^T = LL^T = A$.
- $A = LL^T \quad \Rightarrow \quad x^TAx = x^TLL^Tx = (L^Tx)^TL^Tx = \left\|L^Tx\right\|_2^2 > 0$ $\forall x \in \mathbb{R}^{n\times n} \setminus \{0\}$, denn L^T ist regulär.

□

Nun wollen wir zeigen, dass die Bedingung

$$A \in \mathbb{R}^{n\times n} \text{ ist s.p.d.}$$

hinreichend für die Cholesky-Zerlegung ist.

Lemma 3.26. *Sei $A \in \mathbb{R}^{n\times n}$ eine symmetrische und positiv definite Matrix. Dann gilt:*

(i) $a_{ii} > 0$ *für alle* $i = 1, \ldots, n$.

(ii) $A[k]$ *ist symmetrisch und positiv definit für alle* $k = 1, \ldots, n$.

Beweis.

Zu (i): $a_{ii} = e_i^T A e_i \overset{\text{A p.d.}}{>} 0 \quad \forall i = 1, \ldots, n.$

Zu (ii): $A = A^T \quad \Leftrightarrow \quad A[k] = A[k]^T$ und

$$y^T A[k] y = \begin{pmatrix} y \\ 0 \end{pmatrix}^T A \begin{pmatrix} y \\ 0 \end{pmatrix} \overset{\text{A p.d.}}{>} 0 \quad \forall y \in \mathbb{R}^k \setminus \{0\} \quad \forall k = 1, \ldots, n.$$

□

Satz 3.27 (Cholesky-Zerlegung)**.** *Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit. Dann existiert genau eine untere Dreiecksmatrix $L \in \mathbb{R}^{n\times n}$ mit positiven Diagonalelementen, so dass*

$$A = LL^T.$$

Beweis durch Induktion nach $n \in \mathbb{N}$.

Induktionsanfang n=1:

$$A = (a_{11}) \quad \Rightarrow \quad a_{11} > 0 \quad \Rightarrow \quad L = +\sqrt{a_{11}} \quad \Rightarrow \quad A = LL^T.$$

Induktionsannahme: Die Aussage gelte für alle s.p.d. Matrizen der Dimension $(n-1) \times (n-1)$.

Wir zeigen die Behauptung für alle s.p.d. Matrizen der Dimension $n \times n$. Sei $A \in \mathbb{R}^{n\times n}$ s.p.d. Wir zerlegen A wie folgt:

$$A = \begin{pmatrix} A_{n-1} & b \\ b^T & a_{nn} \end{pmatrix}$$

mit $A_{n-1} \in \mathbb{R}^{(n-1)\times(n-1)}$, $b \in \mathbb{R}^{n-1}$ und $a_{nn} \in \mathbb{R}$. Da A s.p.d. ist, gilt

- $a_{nn} > 0$ und
- $\mathbb{R}^{(n-1)\times(n-1)} \ni A_{n-1} = A[n-1]$ s.p.d. (Lemma).

Nach Induktionsannahme existiert genau eine untere Dreiecksmatrix $L_{n-1} \in \mathbb{R}^{(n-1)\times(n-1)}$ mit positiven Diagonalelementen, so dass

$$A_{n-1} = L_{n-1} L_{n-1}^T.$$

Wir machen den Ansatz

$$L := \begin{pmatrix} L'_{n-1} & 0 \\ c^T & \alpha \end{pmatrix}$$

mit $L'_{n-1} \in \mathbb{R}^{(n-1)\times(n-1)}$, $c \in \mathbb{R}^{n-1}$ und $\alpha \in \mathbb{R}$. Wir suchen L'_{n-1}, c sowie α, so dass die Zerlegung

$$A = LL^T$$

gilt. Diese ist äquivalent zu

$$A = \begin{pmatrix} A_{n-1} & b \\ b^T & a_{nn} \end{pmatrix} = \begin{pmatrix} L'_{n-1} & 0 \\ c^T & \alpha \end{pmatrix} \begin{pmatrix} (L'_{n-1})^T & c \\ 0 & \alpha \end{pmatrix},$$

was dem Gleichungssystem

$$\begin{aligned} A_{n-1} &= L'_{n-1}(L'_{n-1})^T \\ L'_{n-1}c &= b \\ c^T c + \alpha^2 &= a_{nn} \end{aligned}$$

entspricht. Aus diesem ergibt sich nach Induktionsannahme

$$L'_{n-1} = L_{n-1}$$

sowie

$$c = (L_{n-1})^{-1} b$$

und

$$\alpha^2 = a_{nn} - c^T c.$$

Es bleibt zu zeigen, dass $\alpha > 0$ ist. Dazu betrachte

$$0 \neq \det(A) = \det(LL^T) = \det(L)^2 \overset{\text{Entwicklungssatz}}{=} (\alpha \det(L_{n-1}))^2 = \alpha^2 \det(L_{n-1})^2.$$

Hieraus folgt

$$0 < \alpha^2 = a_{nn} - c^T c,$$

also

$$\alpha = +\sqrt{a_{nn} - c^T c} > 0.$$

Insgesamt gilt

$$A = LL^T$$

mit

$$L = \begin{pmatrix} L_{n-1} & 0 \\ L'_{n-1} b & +\sqrt{a_{nn} - ((L'_{n-1})^{-1} b)^T (L'_{n-1})^{-1} b} \end{pmatrix}.$$

□

Bemerkung 3.28. Die Eindeutigkeit der Cholesky-Zerlegung gilt nicht mehr, wenn man untere Dreiecksmatrizen mit negativen Diagonalelementen zulässt.

Algorithmus 3.13 Lösung von $Ax = b$ mit $A \in \mathbb{R}^{n\times n}$ s.p.d.: Cholesky-Zerlegung

(S1) Bestimme eine Cholesky-Zerlegung $A = LL^T$.
(S2) Bestimme y als Lösung von $Ly = b$ (Vorwärtseinsetzen).
(S3) Bestimme x als Lösung von $L^T x = y$ (Rückwärtseinsetzen).

Die Matrix $L = (l_{ij})$ aus der Cholesky-Zerlegung bestimmt man wie folgt:

$$A = LL^T$$

$$\Leftrightarrow \quad \begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} \end{pmatrix} = \begin{pmatrix} l_{11} & & & & \\ l_{21} & l_{22} & & & \\ l_{31} & l_{32} & l_{33} & & \\ \vdots & \vdots & \vdots & \ddots & \\ l_{n1} & l_{n2} & l_{n3} & \cdots & l_{nn} \end{pmatrix} \begin{pmatrix} l_{11} & l_{12} & l_{13} & \cdots & l_{1n} \\ & l_{22} & l_{23} & \cdots & l_{2n} \\ & & l_{33} & \cdots & l_{3n} \\ & & & \ddots & \vdots \\ & & & & l_{nn} \end{pmatrix}.$$

Hieraus folgt:

1. Spalte von L:

$$\begin{aligned} a_{11} &= l_{11}^2 && \Rightarrow && l_{11} = +\sqrt{a_{11}} \\ a_{21} &= l_{21}l_{11} && \Rightarrow && l_{21} = a_{21}/l_{11} \\ &\vdots \\ a_{n1} &= l_{n1}l_{11} && \Rightarrow && l_{n1} = a_{n1}/l_{11} \end{aligned}$$

2. Spalte von L:

$$\begin{aligned} a_{22} &= l_{21}^2 + l_{22}^2 && \Rightarrow && l_{22} = +\sqrt{a_{22} - l_{21}^2} \\ a_{32} &= l_{31}l_{21} + l_{32}l_{22} && \Rightarrow && l_{32} = (a_{32} - l_{31}l_{21})/l_{22} \\ &\vdots \\ a_{n2} &= l_{n1}l_{21} + l_{n2}l_{22} && \Rightarrow && l_{n2} = (a_{n2} - l_{n1}l_{21})/l_{22} \end{aligned}$$

n. Spalte von L:

$$a_{nn} = l_{n1}^2 + l_{n2}^2 + \ldots + l_{nn}^2 \quad \Rightarrow \quad l_{nn} = \sqrt{a_{nn} - l_{n1}^2 - l_{n2}^2 - \ldots l_{n,n-1}^2}$$

Algorithmus 3.14 Cholesky-Zerlegung für s.p.d. Matrizen $A \in \mathbb{R}^{n\times n}$

1: **for** $j = 1 : n$ **do**
2: $\quad l_{jj} := \sqrt{a_{jj} - \sum_{m=1}^{j-1} l_{jm}^2}$ $\quad (\sum_{m=1}^{0} l_{jm}^2 = 0 \text{ für } j = 1)$
3: $\quad$ **for** $i = (j+1) : n$ **do**
4: $\quad\quad l_{ij} := (a_{ij} - \sum_{m=1}^{j-1} l_{im} l_{jm})/l_{jj}$
5: $\quad$ **end for**
6: **end for**

Zusammenfassung 3.29. Sei $A \in \mathbb{R}^{n\times n}$ regulär:

- A besitzt eine LR-Zerlegung genau dann, wenn $\det(A[k]) \neq 0 \quad \forall k = 1, \ldots, n$.
- Die LR-Zerlegung ist eindeutig, sofern diese existiert.
- Es existiert immer eine Permutationsmatrix $P \in \mathbb{R}^{n\times n}$, so dass PA eine LR-Zerlegung besitzt.
- A besitzt eine Cholesky-Zerlegung ($A = LL^T$ mit einer regulären unteren Dreiecksmatrix $L \in \mathbb{R}^{n\times n}$) genau dann, wenn A s.p.d. ist.
- Die Cholesky-Zerlegung ist eindeutig, falls man nur untere Dreiecksmatrizen L mit positiven Diagonalelementen zulässt.
- Die Cholesky-Zerlegung (Algorithmus 3.14) benötigt ca. $\frac{1}{3}n^3$ Rechenoperationen, während die LR-Zerlegung ca. $\frac{2}{3}n^3$ Rechenoperationen benötigt.

Kapitel 4

Lineares Ausgleichsproblem

Seien $A \in \mathbb{R}^{m\times n}$ und $b \in \mathbb{R}^m$ mit $m \geq n$ (oft $m \gg n$) gegeben. Wir betrachten die Aufgabe

$$Ax = b. \tag{4.1}$$

Im Falle einer quadratischen Matrix $A \in \mathbb{R}^{n\times n}$ ($m = n$) lässt sich (4.1) zum Beispiel mit Hilfe der LR-Zerlegung mit Pivotisierung lösen, sofern A regulär ist.
Ist $m > n$, so hat (4.1) mehr Gleichungen als Unbekannte und somit hat (4.1) im Allgemeinen keine Lösung mehr.

Beispiel 4.1. Es seien

$$A = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \in \mathbb{R}^{2\times 1} \quad \text{und} \quad b = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

gegeben. Die Aufgabe

$$\begin{pmatrix} 1 \\ 1 \end{pmatrix} x = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

besitzt dann keine Lösung.

Beim linearen Ausgleichsproblem suchen wir eine optimale Approximation $x \in \mathbb{R}^n$, so dass

$$Ax \approx b.$$

Mathematisch formulieren wir diese Problemstellung als ein Minimierungsproblem

$$(P) \quad \min_{x\in\mathbb{R}^n} \|Ax - b\|,$$

wobei $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ eine Vektornorm bezeichnet. In diesem Kapitel setzen wir

$$\|\cdot\| = \|\cdot\|_2 .$$

Beispiel 4.2 (Hasenpopulation)**.** Zu verschiedenen Zeitpunkten

$$t_1, \ldots, t_m$$

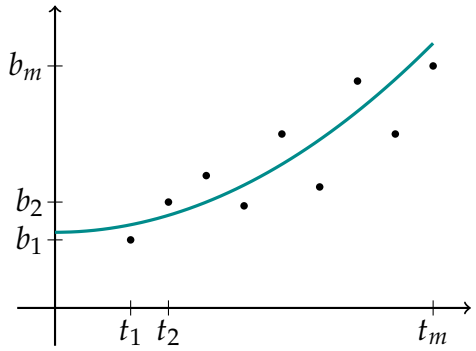
werden die Daten zur Hasenpopulation in Essen

$$b_1, \ldots, b_m$$

gesammelt. Gesucht ist eine Funktion $p : \mathbb{R} \to \mathbb{R}$, so dass

$$\left\| \begin{pmatrix} p(t_1) \\ p(t_2) \\ \vdots \\ p(t_m) \end{pmatrix} - \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix} \right\|$$

minimiert wird. Die Funktion $p : \mathbb{R} \to \mathbb{R}$ liefert eine kontinuierliche Approximation zur Hasenpopulation in Essen.



Für dieses Beispiel ist es sinnvoll (siehe Bild), eine quadratische Approximation

$$p(t) \approx c_0 + c_1 t + c_2 t^2$$

zu wählen. Gesucht sind $c_0, c_1, c_2 \in \mathbb{R}$. Wir kommen auf das folgende Minimierungsproblem:

$$\min \left\| \begin{pmatrix} c_0 + c_1 t_1 + c_2 t_1^2 \\ c_0 + c_1 t_2 + c_2 t_2^2 \\ \vdots \\ c_0 + c_1 t_m + c_2 t_m^2 \end{pmatrix} - \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix} \right\|.$$

Dies entspricht gerade dem Problem

$$\min \|Ax - b\|$$

mit

$$A = \begin{pmatrix} 1 & t_1 & t_1^2 \\ 1 & t_2 & t_2^2 \\ \vdots & \vdots & \vdots \\ 1 & t_m & t_m^2 \end{pmatrix} \in \mathbb{R}^{m \times 3}, \quad b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}, \quad x = \begin{pmatrix} c_0 \\ c_1 \\ c_2 \end{pmatrix}.$$

Definition 4.3. Seien $A \in \mathbb{R}^{m\times n}$ und $b \in \mathbb{R}^m$ mit $m \geq n$. Die Minimierungsaufgabe

$$\min_{x\in\mathbb{R}^n} \|Ax - b\| \tag{P}$$

heißt *lineares Ausgleichsproblem*. Ein Vektor $x^* \in \mathbb{R}^n$ heißt (optimale) Lösung zu (P), falls gilt

$$\|Ax^* - b\| \leq \|Ax - b\| \quad \forall x \in \mathbb{R}^n.$$

Bemerkung 4.4. Die Optimierungsaufgabe (P) ist endlich-dimensional und nicht linear, da die Normabbildung nicht linear ist.

Bei der Untersuchung des linearen Ausgleichsproblems wollen wir uns folgende Fragen stellen:

- Hat (P) eine Lösung?
- Ist die Lösung eindeutig?
- Gibt es eine Charakterisierung für die Lösung (Optimalitätsbedingung)?
- Numerische Approximation der Lösung?

Satz 4.5 (Notwendige und hinreichende Optimalitätsbedingung für (P))**.** *Ein Vektor $x^* \in \mathbb{R}^n$ ist genau dann eine Lösung zu* (P)*, wenn x^* den folgenden Normalgleichungen genügt:*

$$A^TAx^* = A^Tb.$$

Beweis.

„⇒“: Es sei $x^* \in \mathbb{R}^n$ eine Lösung zu (P). Annahme:

$$r := A^T(Ax^* - b) \neq 0.$$

Mit $x_t := x^* - tr \in \mathbb{R}^n$ folgt:

$$\begin{aligned}
\|Ax_t - b\|^2 &= \|Ax_t - b + Ax^* - Ax^*\|^2 \\
&= \|Ax^* - b + A(x_t - x^*)\|^2 \\
&= \|Ax^* - b - tAr\|^2 \\
&= \langle Ax^* - b - tAr, Ax^* - b - tAr\rangle \\
&= \langle Ax^* - b, Ax^* - b\rangle - 2\langle tAr, Ax^* - b\rangle + \langle tAr, tAr\rangle \\
&= \|Ax^* - b\|^2 - 2t\langle Ar, Ax^* - b\rangle + t^2\|Ar\|^2 \\
&= \|Ax^* - b\|^2 - 2t\langle r, \underbrace{A^T(Ax^* - b)}_{=r}\rangle + t^2\|Ar\|^2 \\
&= \|Ax^* - b\|^2 - 2t\|r\|^2 + t^2\|Ar\|^2.
\end{aligned}$$

Da $r \neq 0$ ist, gibt es ein $\bar{t} \in \mathbb{R}^+$, so dass

$$-2t \|r\|^2 + t^2 \|Ar\|^2 < 0 \quad \forall t \in (0, \bar{t}).$$

Somit gilt

$$\|Ax_t - b\|^2 < \|Ax^* - b\|^2 \quad \forall t \in (0, \bar{t}).$$

Daraus ergibt sich der Widerspruch

$$\|Ax^* - b\|^2 \leq \|Ax_t - b\|^2 < \|Ax^* - b\|^2 \quad \forall t \in (0, \bar{t})$$

und es folgt die Behauptung.

„⇐“: Es gelte $A^T Ax^* = A^T b$ für ein $x^* \in \mathbb{R}^n$. Zu zeigen: x^* ist eine Lösung zu (P). Dazu sei $x \in \mathbb{R}^n$ beliebig aber fest. Dann gilt:

$$\begin{aligned}
\|Ax - b\|^2 &= \|A(x - x^*) + Ax^* - b\|^2 \\
&= \|A(x - x^*)\|^2 + 2\langle A(x - x^*), Ax^* - b\rangle + \|Ax^* - b\|^2 \\
&= \|A(x - x^*)\|^2 + 2\langle x - x^*, \underbrace{A^T(Ax^* - b)}_{=0}\rangle + \|Ax^* - b\|^2 \\
&= \|A(x - x^*)\|^2 + \|Ax^* - b\|^2 \\
&\geq \|Ax^* - b\|^2 .
\end{aligned}$$

Somit gilt

$$\|Ax^* - b\| \leq \|Ax - b\| \quad \forall x \in \mathbb{R}^n.$$

Es folgt also, dass x^* optimal ist.

□

Bemerkung 4.6. Der Satz besagt, dass (P) und $A^T Ax = A^T b$ äquivalent sind.

Satz 4.7 (Existenz einer optimalen Lösung zu (P))**.** *Das lineare Ausgleichsproblem*

$$(\mathrm{P}) \quad \min_{x \in \mathbb{R}^n} \|Ax - b\|$$

besitzt eine Lösung.

Beweis. Wir haben bereits gezeigt:

$$x^* \in \mathbb{R}^n \text{ ist eine Lösung zu (P) genau dann, wenn } A^T Ax = A^T b \text{ gilt.}$$

Also genügt es zu zeigen, dass $A^T Ax = A^T b$ eine Lösung $x \in \mathbb{R}^n$ hat. Laut Definition ist $A^T b \in \mathrm{Bild}(A^T)$. In Satz 2.42 haben wir gezeigt, dass

$$\mathrm{Bild}(A^T) = \mathrm{Bild}(A^T A)$$

gilt. Also folgt insgesamt

$$A^T b \in \mathrm{Bild}(A^T) = \mathrm{Bild}(A^T A) \quad \Rightarrow \quad \exists x \in \mathbb{R}^n : A^T Ax = A^T b.$$

□

Bemerkung 4.8.

(i) Obwohl die Aufgabe $Ax = b$ im Allgemeinen keine Lösung besitzt, hat die Aufgabe (P) stets eine Lösung.

(ii) Die Lösung von (P) ist im Allgemeinen nicht eindeutig, wie folgendes Beispiel zeigt.

Beispiel 4.9. Es seien

$$A = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \in \mathbb{R}^{3\times 1} \quad \text{und} \quad b = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \in \mathbb{R}^{3\times 1}$$

gegeben. Dann ist

$$A^T A = \begin{pmatrix} 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} = 0 \quad \text{und} \quad A^T b = \begin{pmatrix} 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = 0.$$

Es gilt

$$A^T A x = A^T b \quad \forall x \in \mathbb{R}^n$$

und somit löst jedes $x \in \mathbb{R}^n$ die Aufgabe (P).

Satz 4.10 (Eindeutigkeit der Lösung)**.** *Das lineare Ausgleichsproblem hat genau dann eine eindeutige Lösung, wenn*

$$\operatorname{Rang}(A) = n.$$

Beweis.

„⇐“: Es sei $\operatorname{Rang}(A) = n$.

Daraus folgt, dass $A^T A$ regulär ist (Übungsaufgabe). Aus der Regularität von $A^T A$ ergibt sich, dass die Normalgleichungen $A^T A x = A^T b$ genau eine Lösung besitzen. Dies ist gleichbedeutend mit der Eindeutigkeit des Problems (P).

„⇒“: (P) habe genau eine Lösung. Annahme: $\operatorname{Rang}(A) < n$.

Daraus folgt, dass die Normalgleichungen $A^T A x = A^T b$ unendlich viele Lösungen besitzen. Dies ist äquivalent dazu, dass das Problem (P) unendlich viele Lösungen besitzt. Der Widerspruch zur Voraussetzung liefert die Behauptung.

□

Fazit 4.11. Das lineare Ausgleichsproblem

$$\text{(P)} \quad \min \|Ax - b\|, \quad A \in \mathbb{R}^{m\times n}, \quad b \in \mathbb{R}^m$$

hat stets eine Lösung und die Lösung ist genau dann eindeutig, wenn $\operatorname{Rang}(A) = n$. Die Lösung ist gegeben durch

$$A^T A x = A^T b.$$

4.1 QR-Zerlegung

Im Folgenden seien $A \in \mathbb{R}^{m\times n}$ und $b \in \mathbb{R}^m$ mit $m \geq n$ und $\text{Rang}(A) = n$. Somit hat (P) genau eine Lösung $x^* \in \mathbb{R}^n$. Wir wollen eine numerische Lösung zu (P) mittels einer QR-Zerlegung finden.

Definition 4.12 (QR-Zerlegung)**.**

(i) Die Zerlegung

$$A = \hat{Q}\hat{R}, \quad \hat{Q} \in \mathbb{R}^{m\times n}, \quad \hat{R} \in \mathbb{R}^{n\times n}$$

mit den Eigenschaften

- $\hat{Q} = \begin{pmatrix} | & & | \\ \hat{q}_1 & \cdots & \hat{q}_n \\ | & & | \end{pmatrix}$, $\hat{q}_i \in \mathbb{R}^m$, $\langle \hat{q}_i, \hat{q}_j \rangle = \delta_{ij}$ (Spaltenvektoren sind orthonormiert)
- $\hat{R} = \begin{pmatrix} r_{11} & \cdots & r_{1n} \\ & \ddots & \vdots \\ & & r_{nn} \end{pmatrix} \in \mathbb{R}^{n\times n}$

heißt *reduzierte QR-Zerlegung* von A.

(ii) Die Zerlegung

$$A = QR, \quad Q \in \mathbb{R}^{m\times m}, \quad R \in \mathbb{R}^{m\times n}$$

mit den Eigenschaften

- $Q^TQ = I \in \mathbb{R}^{m\times m}$ (Q ist orthogonal)
- $R = \left(\begin{array}{ccc} r_{11} & \cdots & r_{1n} \\ & \ddots & \vdots \\ & & r_{nn} \\ \hline & 0 & \end{array}\right)$

heißt *(volle) QR-Zerlegung* von A.

Bemerkung 4.13.

(i): Existiert eine QR-Zerlegung von A, so existiert auch eine reduzierte QR-Zerlegung. Denn laut Definition sind

$$Q = \begin{bmatrix} \hat{Q} & \tilde{Q} \end{bmatrix}$$
$$R = \begin{bmatrix} \hat{R} \\ 0 \end{bmatrix}$$

und somit

$$A = QR = \begin{bmatrix} \hat{Q} & \tilde{Q} \end{bmatrix} \begin{bmatrix} \hat{R} \\ 0 \end{bmatrix} = \hat{Q}\hat{R}.$$

(ii): Existiert eine reduzierte QR-Zerlegung von A, so existiert auch eine volle QR-Zerlegung von A, indem man die Spalten von $\hat{Q}$ zu einer Orthonormalbasis des $\mathbb{R}^m$ erweitert.

Satz 4.14 (Eindeutigkeit der reduzierten QR-Zerlegung). *Sei $A \in \mathbb{R}^{m\times n}$ mit $\operatorname{Rang}(A) = n \leq m$. Ferner seien*

$$A = \hat{Q}_1\hat{R}_1 \quad \textit{und} \quad A = \hat{Q}_2\hat{R}_2$$

zwei reduzierte QR-Zerlegungen von A. Dann existiert eine Diagonalmatrix

$$\mathbb{R}^{n\times n} \ni D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$$

mit $\lambda_i = \pm 1$, so dass

$$\hat{Q}_1 = \hat{Q}_2 D \quad \textit{und} \quad \hat{R}_2 = D\hat{R}_1.$$

Beweis. Es gilt

$$\hat{Q}_1\hat{R}_1 = \hat{Q}_2\hat{R}_2. \tag{4.2}$$

Laut Definition sind die Spaltenvektoren von $\hat{Q}_1$ orthonormiert. Die gleiche Aussage gilt auch für $\hat{Q}_2$. Folglich ist

$$\hat{Q}_2^T\hat{Q}_2 = \hat{Q}_1^T\hat{Q}_1 = I \in \mathbb{R}^{n\times n}.$$

Daraus folgt mit (4.2)

$$\begin{aligned}\hat{R}_1 &= \hat{Q}_1^T(\hat{Q}_1\hat{R}_1) = \hat{Q}_1^T\hat{Q}_2\hat{R}_2\\ \hat{R}_2 &= \hat{Q}_2^T(\hat{Q}_2\hat{R}_2) = \hat{Q}_2^T\hat{Q}_1\hat{R}_1\end{aligned}$$

und

$$\left.\begin{aligned}\hat{R}_1\hat{R}_2^{-1} &= \hat{Q}_1^T\hat{Q}_2\\ \hat{R}_2\hat{R}_1^{-1} &= \hat{Q}_2^T\hat{Q}_1,\end{aligned}\right\} \tag{4.3}$$

da $\hat{R}_1$ und $\hat{R}_2$ regulär sind ($\operatorname{Rang}(A) = n$). Wir setzen

$$D := \hat{R}_2\hat{R}_1^{-1} \in \mathbb{R}^{n\times n}.$$

Dann ist D eine obere Dreiecksmatrix, da $\hat{R}_2$ und $\hat{R}_1^{-1}$ obere Dreiecksmatrizen sind. Andererseits gilt

$$D^T \overset{\text{Def.}}{=} (\hat{R}_2\hat{R}_1^{-1})^T \overset{(4.3)}{=} (\hat{Q}_2^T\hat{Q}_1)^T = \hat{Q}_1^T\hat{Q}_2 \overset{(4.3)}{=} \hat{R}_1\hat{R}_2^{-1}. \tag{4.4}$$

Also ist D^T eine obere Dreiecksmatrix. Damit ist D sowohl eine obere als auch eine untere Dreiecksmatrix. Es gilt

$$D = \operatorname{diag}\lambda_1, \ldots, \lambda_n$$

mit $\lambda_i \in \mathbb{R}$. Es bleibt zu zeigen, dass $\lambda_i = \pm 1$ für alle $i = 1, \ldots, n$. Hierzu verwenden wir (4.4):

$$D^T = \hat{R}_1 \hat{R}_2^{-1} = (\hat{R}_2 \hat{R}_1^{-1})^{-1} \overset{\text{Def.}}{=} D^{-1}.$$

Also gilt

$$I = DD^T = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} = \begin{pmatrix} \lambda_1^2 & & \\ & \ddots & \\ & & \lambda_n^2 \end{pmatrix}$$

und somit folgt

$$\lambda_i = \pm 1 \quad \forall i = 1, \ldots, n.$$

□

Wir wollen nun das lineare Ausgleichsproblem

$$\min \|Ax - b\|, \quad A \in \mathbb{R}^{n \times n}, \quad b \in \mathbb{R}^n \tag{P}$$

mittels reduzierter QR-Zerlegung lösen. Auf Grund der Voraussetzung $\text{Rang}(A) = n$ hat (P) genau eine Lösung und die Lösung erfüllt

$$A^T Ax = A^T b.$$

Existiert eine reduzierte QR-Zerlegung von A

$$A = \hat{Q}\hat{R}, \quad \hat{Q} \in \mathbb{R}^{m \times n}, \quad \hat{R} \in \mathbb{R}^{n \times n},$$

so folgt

$$A^T Ax = (\hat{Q}\hat{R})^T \hat{Q}\hat{R}x = \hat{R}^T \hat{Q}^T \hat{Q}\hat{R}x = \hat{R}^T \hat{R}x$$

und

$$A^T b = (\hat{Q}\hat{R})^T b = \hat{R}^T \hat{Q}^T b.$$

Folglich ist

$$A^T Ax = b \quad \Leftrightarrow \quad \hat{R}^T \hat{R}x = \hat{R}^T \hat{Q}^T b.$$

Da $\text{Rang}(A) = n$ ist, ist $\hat{R}$ regulär. Somit ist auch $\hat{R}^T$ regulär. Insgesamt gilt

$$A^T Ax = A^T b \quad \Leftrightarrow \quad \hat{R}x = \hat{Q}^T b.$$

Algorithmus 4.1 Lösung von (P) mittels einer reduzierten QR-Zerlegung

(S1) Bestimme eine reduzierte QR-Zerlegung von A.
(S2) Setze $\hat{c} := \hat{Q}^T b$.
(S3) Löse $\hat{R}x = \hat{c}$ mittels Rückwärtseinsetzen.

Alternativ lässt sich (P) mittels voller QR-Zerlegung lösen. Sei $A = QR$ mit $Q \in \mathbb{R}^{m \times m}$ und $R \in \mathbb{R}^{m \times n}$ eine volle QR-Zerlegung. Das heißt:

(i) $Q^T Q = I \in \mathbb{R}^{m \times m}$

(ii) $R = \begin{pmatrix} \hat{R} \\ 0 \end{pmatrix}$ mit $\hat{R} = \begin{pmatrix} r_{11} & \cdots & r_{1n} \\ & \ddots & \vdots \\ & & r_{nn} \end{pmatrix} \in \mathbb{R}^{n \times n}$.

Wir setzen

$$c := Q^T b \in \mathbb{R}^m$$

und zerlegen

$$c = \begin{pmatrix} \hat{c} \\ \tilde{c} \end{pmatrix} \text{ mit } \hat{c} \in \mathbb{R}^n \text{ und } \tilde{c} \in \mathbb{R}^{m-n}.$$

Folglich gilt

$$\|Ax - b\| = \|QRx - b\| = \left\|QRx - QQ^T b\right\| = \left\|Q(Rx - Q^T b)\right\|.$$

Wegen

$$\|Qy\|^2 = (Qy)^T Qy = y^T Q^T Qy = y^T y = \|y\|^2 \quad \forall y \in \mathbb{R}^n,$$

folgt

$$\|Ax - b\| = \left\|Rx - Q^T b\right\| = \left\| \begin{pmatrix} \hat{R} \\ 0 \end{pmatrix} x - \begin{pmatrix} \hat{c} \\ \tilde{c} \end{pmatrix} \right\|.$$

Insgesamt gilt

$$\min \|Ax - b\| \quad \Leftrightarrow \quad \min \left\| \begin{pmatrix} \hat{R} \\ 0 \end{pmatrix} x - \begin{pmatrix} \hat{c} \\ \tilde{c} \end{pmatrix} \right\|.$$

Da aber

$$\left\| \begin{pmatrix} \hat{R} \\ 0 \end{pmatrix} x - \begin{pmatrix} \hat{c} \\ \tilde{c} \end{pmatrix} \right\|^2 \overset{\|\cdot\| = \|\cdot\|_2}{=} \left\|\hat{R}x - \hat{c}\right\|^2 + \|\tilde{c}\|^2$$

ist, kommen wir zur Folgerung:

$$x \text{ löst } (P) \quad \Leftrightarrow \quad x \text{ löst } \min \left\| \begin{pmatrix} \hat{R} \\ 0 \end{pmatrix} x - \begin{pmatrix} \hat{c} \\ \tilde{c} \end{pmatrix} \right\| \quad \Leftrightarrow \quad x \text{ löst } \min \left\|\hat{R}x - \hat{c}\right\|^2 \quad \Leftrightarrow \quad \hat{R}x = \hat{c}.$$

Das ist genau Schritt 3 im Algorithmus 4.1.

4.2 Gram-Schmidt-Orthogonalisierung

Im Folgenden sei $A \in \mathbb{R}^{m\times n}$ mit $\text{Rang}(A) = n \leq m$. Folglich hat (P) genau eine Lösung, die wir mit Hilfe des Algorithmus 4.1 bestimmen wollen.
In diesem Abschnitt zeigen wir die Existenz einer (reduzierten) QR-Zerlegung durch die Gram-Schmidt-Orthogonalisierungsmethode.
Aus $A = \hat{Q}\hat{R}$ mit $\hat{Q} \in \mathbb{R}^{m\times n}$ und $\hat{R} \in \mathbb{R}^{n\times n}$ erhalten wir (für eine bessere Lesbarkeit schreiben wir im Folgenden $q^{(j)} = \hat{q}^{(j)}$ für die Spaltenvektoren von $\hat{Q}$)

$$A = \begin{pmatrix} | & & | \\ a^{(1)} & \cdots & a^{(n)} \\ | & & | \end{pmatrix} = \begin{pmatrix} | & & | \\ q^{(1)} & \cdots & q^{(n)} \\ | & & | \end{pmatrix} \begin{pmatrix} r_{11} & \cdots & r_{1n} \\ & \ddots & \vdots \\ & & r_{nn} \end{pmatrix}.$$

Dies ist äquivalent zu dem System

$$\begin{aligned} a^{(1)} &= r_{11}q^{(1)} \\ a^{(2)} &= r_{12}q^{(1)} + r_{22}q^{(2)} \\ a^{(3)} &= r_{13}q^{(1)} + r_{23}q^{(2)} + r_{33}q^{(3)} \\ &\vdots \\ a^{(n)} &= r_{1n}q^{(1)} + \ldots + r_{nn}q^{(n)}, \end{aligned}$$

also

$$a^{(j)} = \sum_{i=1}^{j} r_{ij}q^{(i)} \quad \forall j = 1, \ldots, n.$$

Hieraus ergibt sich eine Formel für die orthonormalen Vektoren $q^{(j)}$ und die Zahlen r_{ij}.

Für $j = 1$:

$$a^{(1)} = r_{11}q^{(1)} \quad \Rightarrow \quad q^{(1)} = \frac{a^{(1)}}{r_{11}} \quad \text{und} \quad r_{11} = \pm \left\| a^{(1)} \right\|,$$

damit $\left\| q^{(1)} \right\| = 1$.

Für $j = 2$:

$$q^{(2)} = \frac{1}{r_{22}}(a^{(2)} - r_{12}q^{(1)}) \quad \text{und} \quad r_{22} = \pm \left\| a^{(2)} - r_{12}q^{(1)} \right\|,$$

damit $\left\| q^{(2)} \right\| = 1$ und r_{12} bestimmen wir aus der Orthogonalitätsbedingung

$$0 = \langle q^{(1)}, q^{(2)} \rangle = \frac{1}{r_{22}}(\langle q^{(1)}, a^{(2)} \rangle - r_{12} \underbrace{\langle q^{(1)}, q^{(2)} \rangle}_{=1}) = \frac{1}{r_{22}}(\langle q^{(1)}, a^{(2)} \rangle - r_{12}).$$

Also ist

$$r_{12} = \langle q^{(1)}, a^{(2)} \rangle.$$

Für $j \geq 3$:

$$q^{(j)} = \frac{1}{r_{jj}} \left(a^{(j)} - \sum_{i=1}^{j-1} r_{ij} q^{(i)} \right) \quad \text{und}$$
$$r_{jj} = \pm \left\| a^{(j)} - \sum_{i=1}^{j-1} r_{ij} q^{(i)} \right\| \quad \text{und}$$
$$r_{ij} = \langle q^{(i)}, a^{(j)} \rangle \quad \forall i = 1, \ldots, j-1.$$

Algorithmus 4.2 Gram-Schmidt-Verfahren

1: $r_{11} := \left\| a^{(1)} \right\|$
2: $q^{(1)} := a^{(1)} / r_{11}$
3: **for** $j = 2 : n$ **do**
4: $\quad q^{(j)} := a^{(j)}$
5: $\quad$ **for** $i = 1 : (j-1)$ **do**
6: $\quad\quad r_{ij} := \langle q^{(i)}, a^{(j)} \rangle$
7: $\quad\quad q^{(j)} := q^{(j)} - r_{ij} q^{(i)}$
8: $\quad$ **end for**
9: $\quad r_{jj} := \left\| q^{(j)} \right\|$
10: $\quad q^{(j)} := q^{(j)} / r_{jj}$
11: **end for**

Satz 4.15 (Existenz der QR-Zerlegung). *Es sei $A \in \mathbb{R}^{m \times n}$ mit $\text{Rang}(A) = n \leq m$. Dann existiert eine QR-Zerlegung von A*

$$A = \hat{Q} \hat{R}$$

mit $\hat{Q} \in \mathbb{R}^{m \times n}$ und $\hat{R} \in \mathbb{R}^{n \times n}$ (siehe Definition). Somit existiert auch eine volle QR-Zerlegung von A.

Beweis. Es bleibt zu zeigen, dass das Gram-Schmidt-Verfahren durchführbar ist. Wir müssen also zeigen:

$$r_{jj} \neq 0 \quad \forall j = 1, \ldots, n.$$

Annahme: Es gibt ein $j \in \{1, \ldots, n\}$, so dass $r_{ii} \neq 0 \quad \forall i = 1, \ldots, j-1$, aber $r_{jj} = 0$. Folglich

$$0 = r_{jj} = \left\| a^{(j)} - \sum_{i=1}^{j-1} r_{ij} q^{(i)} \right\|$$
$$\Leftrightarrow a^{(j)} = \sum_{i=1}^{j-1} r_{ij} q^{(i)}$$
$$\Leftrightarrow a^{(j)} \in \text{Span}\{q^{(1)}, \ldots, q^{(j-1)}\}.$$

Laut Konstruktion ist

$$\text{Span}\{q^{(1)}, \ldots, q^{(j-1)}\} = \text{Span}\{a^{(1)}, \ldots, a^{(j-1)}\}.$$

Somit gilt

$$a^{(j)} \in \text{Span}\{q^{(1)}, \ldots, q^{(j-1)}\} = \text{Span}\{a^{(1)}, \ldots, a^{(j-1)}\}.$$

Mit anderen Worten lässt sich der j-te Spaltenvektor von A als Linearkombination von den Spaltenvektoren $a^{(1)}, \ldots, a^{(j-1)}$ darstellen. Dies ist ein Widerspruch zu $\text{Rang}(A) = n$. Also folgt die Behauptung. □

Bemerkung 4.16. Das Gram-Schmidt-Verfahren ist numerisch instabil (vgl. Übung). Aufgrund der endlichen Rechengenauigkeit des Computers geht die Orthgonalität von $q^{(j)}$ schnell verloren.

Wir wollen nun eine modifizierte (stabile) Variante des Gram-Schmidt-Verfahrens herleiten, indem wir weitere Hilfsvektoren $p^{(j)}$ einführen, um das Überschreiben der Vektoren $q^{(j)}$ im Algorithmus 4.2 zu vermeiden.

$p^{(1)} = a^{(1)}$
$r_{11} := \left\| p^{(1)} \right\|$
$q^{(1)} := p^{(1)} / r_{11}$
for $j = 2 : n$ **do**
    for $i = 1 : (j-1)$ **do**
        $r_{ij} := \langle q^{(i)}, a^{(j)} \rangle$
    end for
    $p^{(j)} := a^{(j)} - \sum_{i=1}^{j-1} r_{ij} q^{(i)}$
    $r_{jj} := \left\| p^{(j)} \right\|$
    $q^{(j)} := p^{(j)} / r_{jj}$
end for

Die Berechnung für $p^{(j)}$ lautet:

$$p^{(j)} = a^{(j)} - \sum_{i=1}^{j-1} r_{ij} q^{(i)} = a^{(j)} - \sum_{i=1}^{j-1} \langle q^{(i)}, a^{(j)} \rangle q^{(i)} = a^{(j)} - \sum_{i=1}^{j-1} q^{(i)} \underbrace{\langle q^{(i)}, a^{(j)} \rangle}_{= (q^{(i)})^T a^{(j)}}$$

$$= \left(I - \sum_{i=1}^{j-1} q^{(i)} (q^{(i)})^T \right) a^{(j)}.$$

Insgesamt gilt

$$p^{(j)} = \left(I - \sum_{i=1}^{j-1} q^{(i)} (q^{(i)})^T \right) a^{(j)}.$$

Wegen der Orthogonalitätsbedingung für $q^{(i)}$ gilt

$$\left(I - \sum_{i=1}^{j-1} q^{(i)}(q^{(i)})^T\right) = (I - q^{(j-1)}(q^{(j-1)})^T) \cdot \ldots \cdot (I - q^{(2)}(q^{(2)})^T)(I - q^{(1)}(q^{(1)})^T)$$

und somit lässt sich $p^{(j)}$ auch wie folgt berechnen:

1: $p^{j,1} := a^{(j)}$
2: **for** $i = 1 : (j-1)$ **do**
3: $\quad r_{ij} := \langle q^{(i)}, p^{j,i} \rangle$
4: $\quad p^{j,i+1} := p^{j,i} - r_{ij} q^{(i)}$
5: **end for**
6: $p^{(j)} = p^{j,j}$

Hier gilt:

$$p^{j,1} = a^{(j)}$$

$$\begin{aligned} i = 1: \qquad p^{j,2} &= a^{(j)} - \langle q^{(1)}, a^{(j)} \rangle q^{(1)} \\ &= a^{(j)} - q^{(1)}(q^{(1)})^T a^{(j)} \\ &= (I - q^{(1)}(q^{(1)})^T) a^{(j)} \end{aligned}$$

$$i = 2: \qquad p^{j,3} = (I - q^{(2)}(q^{(2)})^T)(I - q^{(1)}(q^{(1)})^T) a^{(j)}$$

$$i = j-1: \quad p^{j,j} = (I - q^{(j-1)}(q^{(j-1)})^T) \cdot \ldots \cdot (I - q^{(2)}(q^{(2)})^T)(I - q^{(1)}(q^{(1)})^T) a^{(j)}.$$

Speichert man nun $p^{j,i}$ nicht explizit ab und überschreibt stattdessen einen Vektor $q^{(j)}$, so kommen wir auf das modifizierte Gram-Schmidt-Verfahren.

Algorithmus 4.3 Modifiziertes Gram-Schmidt-Verfahren

1: $r_{11} := \left\| a^{(1)} \right\|$
2: $q^{(1)} := a^{(1)} / r_{11}$
3: **for** $j = 2 : n$ **do**
4: $\quad q^{(j)} := a^{(j)}$
5: $\quad$ **for** $i = 1 : (j-1)$ **do**
6: $\quad\quad r_{ij} := \langle q^{(i)}, q^{(j)} \rangle$
7: $\quad\quad q^{(j)} := q^{(j)} - r_{ij} q^{(i)}$
8: $\quad$ **end for**
9: $\quad r_{jj} := \left\| q^{(j)} \right\|$
10: $\quad q^{(j)} := q^{(j)} / r_{jj}$
11: **end for**

Bemerkung 4.17. Algorithmen 4.2 und 4.3 benötigen

$$4m\frac{(n-1)n}{2} \approx 2mn^2$$

Rechenoperationen.

4.3 Householder-Spiegelungen

In diesem Abschnitt befassen wir uns mit Householder-Spiegelungen zur Konstruktion einer vollen QR-Zerlegung von A, das heißt

$$A = QR$$

mit einer Orthogonalmatrix $Q \in \mathbb{R}^{m\times m}$ und einer Matrix $R \in \mathbb{R}^{m\times n}$ der Gestalt

$$R = \begin{pmatrix} r_{11} & \cdots & r_{1n} \\ & \ddots & \vdots \\ & & r_{nn} \\ \hline & 0 & \end{pmatrix}.$$

Im Folgenden nehmen wir wieder an, dass

$$\text{Rang}(A) = n \leq m.$$

Somit existiert eine volle QR-Zerlegung

$$A = QR \quad \Leftrightarrow \quad Q^T A = R.$$

Idee: Die Householder-Spiegelung liefert ein Produkt von einfachen Orthogonalmatrizen $Q_i \in \mathbb{R}^{m\times m}$ mit

$$Q^T = Q_l \cdot \ldots \cdot Q_2 Q_1, \quad l = \min\{n, m-1\}.$$

Jedes Q_i sorgt dafür, dass in der i-ten Spalte von A geeignete Nullen erzeugt werden (vgl. Frobenius-Matrizen L_i).

$$A = \begin{pmatrix} * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \end{pmatrix} \quad \rightarrow \quad Q_1 A = \begin{pmatrix} * & * & * \\ 0 & * & * \\ 0 & * & * \\ 0 & * & * \end{pmatrix}$$

$$\rightarrow \quad Q_2 Q_1 A = \begin{pmatrix} * & * & * \\ 0 & * & * \\ 0 & 0 & * \\ 0 & 0 & * \end{pmatrix} \quad \rightarrow \quad Q_3 Q_2 Q_1 A = \begin{pmatrix} * & * & * \\ 0 & * & * \\ 0 & 0 & * \\ \hline 0 & 0 & 0 \end{pmatrix} =: R.$$

Definition 4.18 (Householder-Spiegelung). Eine Matrix $H \in \mathbb{R}^{n\times n}$ heißt *Householder-Spiegelung*, falls gilt

$$H = I - 2uu^T$$

für einen Vektor $u \in \mathbb{R}^n$ mit $\|u\| = 1$.

Lemma 4.19. *Jede Householder-Spiegelung $H \in \mathbb{R}^{n\times n}$ ist symmetrisch und orthogonal.*

Beweis.

- $H^T = (I - 2uu^T)^T = I - 2uu^T = H$.
- $H^T H = HH = (I - 2uu^T)(I - 2uu^T) = I - 4uu^T + 4uu^T uu^T = I - 4uu^T + 4uu^T = I$, da $u^T u = \|u\|^2 = 1$ ist.

□

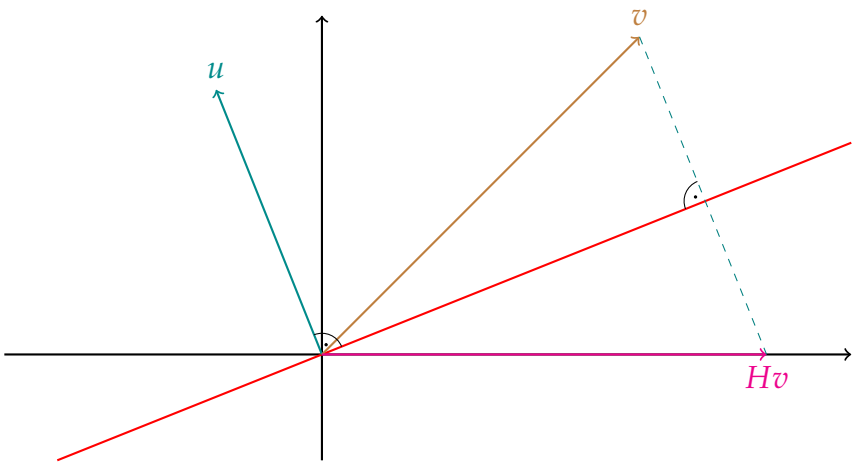
Bemerkung 4.20. Sei $u \in \mathbb{R}^n$ mit $\|u\| = 1$ und

$$H = I - 2uu^T \in \mathbb{R}^{n\times n} \quad \text{(Householder-Spiegelung).}$$

Geometrisch beschreibt die Abbildung

$$v \mapsto Hv, \quad \mathbb{R}^n \to \mathbb{R}^n$$

eine Spiegelung des Vektors v an einer durch den Nullpunkt gehenden Ebene mit dem Normalenvektor u.



Wir wollen zunächst untersuchen, wie man durch eine Householder-Spiegelung einen Vektor $v \in \mathbb{R}^n \setminus \{0\}$ in ein Vielfaches von $e_1 \in \mathbb{R}^n$ abbilden kann:

$$Hv = \rho e_1.$$

Sei also

$$H = I - 2uu^T \in \mathbb{R}^{n\times n}$$

mit $u \in \mathbb{R}^n$, so dass $\|u\| = 1$. Weiter sei $v \in \mathbb{R}^n$. Wir bestimmen nun $u \in \mathbb{R}^n$ und $\rho \in \mathbb{R}$, so dass

$$Hv = (I - 2uu^T)v = v - 2uu^T v \stackrel{!}{=} \rho e_1. \tag{4.5}$$

Wegen

$$\|v\| = \|Hv\| = \|\rho e_1\| = |\rho|$$

ist

$$\rho = \pm \|v\| .$$

Wir setzen nun

$$\gamma = 2\langle u, v\rangle,$$

und nehmen an, dass $\gamma \neq 0$ ist. Aus (4.5) erhalten wir

$$\gamma u \stackrel{\text{Def.}}{=} 2u\langle u, v\rangle = 2uu^T v = v - \rho e_1$$

und daher

$$u = \frac{1}{\gamma}(v - \rho e_1)$$

sowie

$$\gamma = \pm \|v - \rho e_1\| , \quad \text{da } \|u\| = 1.$$

Im Folgenden wählen wir $\gamma = + \|v - \rho e_1\|$. Insgesamt haben wir das folgende Resultat.

Lemma 4.21. *Sei $v \in \mathbb{R}^n \setminus \{0\}$ und*

$$\rho := \pm \|v\| , \quad \gamma = + \|v - \rho e_1\| , \quad u = \frac{1}{\gamma}(v - \rho e_1).$$

Ist $\gamma \neq 0$, so gilt für die Householder-Matrix $H = I - 2uu^T \in \mathbb{R}^{n\times n}$:

$$Hv = \rho e_1.$$

Bemerkung 4.22. Aus numerischen Gründen ist die Wahl

$$\rho = \begin{cases} - \|v\| & \text{falls } v_1 \geq 0 \\ + \|v\| & \text{falls } v_1 < 0 \end{cases}$$

besser geeignet, um den Effekt der Auslöschung bei der Berechnung von

$$v - \rho e_1$$

zu vermeiden. Bei dieser Wahl ist die Bedingung $\gamma \neq 0$ stets gesichert.

Das obige Lemma ist eine wichtige Grundlage für die volle QR-Zerlegung mittels Householder-Spiegelungen. Dazu betrachten wir ein Beispiel.

Beispiel 4.23. Es sei

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

1. Konstruktion von Q_1 (vgl. Lemma 4.21):

$$v = \begin{pmatrix} 0 \\ 0 \\ 2 \\ 0 \end{pmatrix}.$$

Damit ergibt sich

$$\begin{cases} \|v\| = 2 \\ \rho = -\|v\| = -2 \quad \text{(siehe Bemerkung 4.22)} \\ \gamma = \|v - \rho e_1\| = \left\| \begin{pmatrix} 0 \\ 0 \\ 2 \\ 0 \end{pmatrix} + \begin{pmatrix} 2 \\ 0 \\ 0 \\ 0 \end{pmatrix} \right\| = \left\| \begin{pmatrix} 2 \\ 0 \\ 2 \\ 0 \end{pmatrix} \right\| = \sqrt{8} \end{cases}$$

Es folgt

$$u = \frac{1}{\gamma}(v - \rho e_1) = \frac{1}{\sqrt{8}} \begin{pmatrix} 2 \\ 0 \\ 2 \\ 0 \end{pmatrix}$$

und

$$H_1 = I - 2uu^T = I - 2\frac{1}{8} \begin{pmatrix} 2 \\ 0 \\ 2 \\ 0 \end{pmatrix} \begin{pmatrix} 2 & 0 & 2 & 0 \end{pmatrix}$$

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

Insgesamt ergibt sich also

$$H_1 = \begin{pmatrix} 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Setze $Q_1 = H_1$ und erhalte

$$Q_1 A = \begin{pmatrix} 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & -\sqrt{2} \\ 2 & 1 & 0 \\ 0 & 1 & 2 \end{pmatrix} = \begin{pmatrix} -2 & -1 & 0 \\ 0 & 0 & -\sqrt{2} \\ 0 & -1 & 0 \\ 0 & 1 & 2 \end{pmatrix}.$$

Kontrolle:

$$Hv = \rho e_1 = \begin{pmatrix} -2 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

2. Konstruktion von Q_2:

$$v = \begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix}.$$

Damit ergibt sich

$$\begin{cases} \|v\| = \sqrt{2} \\ \rho = -\|v\| = -\sqrt{2} \\ \gamma = \|v - \rho e_1\| = \left\| \begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix} + \begin{pmatrix} \sqrt{2} \\ 0 \\ 0 \end{pmatrix} \right\| = \left\| \begin{pmatrix} \sqrt{2} \\ -1 \\ 1 \end{pmatrix} \right\| = 2 \end{cases}$$

Es folgt

$$u = \frac{1}{\gamma}(v - \rho e_1) = \frac{1}{2} \begin{pmatrix} \sqrt{2} \\ -1 \\ 1 \end{pmatrix}$$

und

$$\begin{aligned} H_2 = I - 2uu^T &= I - 2\frac{1}{4} \begin{pmatrix} \sqrt{2} \\ -1 \\ 1 \end{pmatrix} \begin{pmatrix} \sqrt{2} & -1 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & & \\ & 1 & \\ & & 1 \end{pmatrix} - \frac{1}{2} \begin{pmatrix} 2 & -\sqrt{2} & \sqrt{2} \\ -\sqrt{2} & 1 & -1 \\ \sqrt{2} & -1 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 0 & \frac{1}{2}\sqrt{2} & -\frac{1}{2}\sqrt{2} \\ \frac{1}{2}\sqrt{2} & \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2}\sqrt{2} & \frac{1}{2} & \frac{1}{2} \end{pmatrix}. \end{aligned}$$

Setze

$$Q_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & & & \\ 0 & & H_2 & \\ 0 & & & \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2}\sqrt{2} & -\frac{1}{2}\sqrt{2} \\ 0 & \frac{1}{2}\sqrt{2} & \frac{1}{2} & \frac{1}{2} \\ 0 & -\frac{1}{2}\sqrt{2} & \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

und

$$Q_2 Q_1 A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2}\sqrt{2} & -\frac{1}{2}\sqrt{2} \\ 0 & \frac{1}{2}\sqrt{2} & \frac{1}{2} & \frac{1}{2} \\ 0 & -\frac{1}{2}\sqrt{2} & \frac{1}{2} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 2 & -1 & 0 \\ 0 & 0 & -\sqrt{2} \\ 0 & -1 & 0 \\ 0 & 1 & 2 \end{pmatrix} = \begin{pmatrix} -2 & -1 & 0 \\ 0 & -\sqrt{2} & -\sqrt{2} \\ 0 & 0 & 0 \\ 0 & 0 & 2 \end{pmatrix}.$$

Kontrolle:

$$Hv = \rho e_1 = \begin{pmatrix} -\sqrt{2} \\ 0 \\ 0 \end{pmatrix}.$$

3. Konstruktion von Q_3:

$$v = \begin{pmatrix} 0 \\ 2 \end{pmatrix}.$$

Damit ergibt sich

$$\begin{cases} \|v\| = 2 \\ \rho = -\|v\| = -2 \\ \gamma = \|v - \rho e_1\| = \left\| \binom{2}{2} \right\| = \sqrt{8} \end{cases}$$

Es folgt

$$u = \frac{1}{\gamma}(v - \rho e_1) = \frac{1}{\sqrt{8}} \begin{pmatrix} 2 \\ 2 \end{pmatrix}$$

und

$$\begin{aligned} H_3 = I - 2uu^T &= I - 2\frac{1}{8} \begin{pmatrix} 2 \\ 2 \end{pmatrix} \begin{pmatrix} 2 & 2 \end{pmatrix} \\ &= \begin{pmatrix} 1 & \\ & 1 \end{pmatrix} - \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}. \end{aligned}$$

Setze

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & & \\ 0 & 0 & & H_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 0 \end{pmatrix}.$$

Somit gilt:

$$\begin{aligned} Q_3 Q_2 Q_1 A &= \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 0 \end{pmatrix} \begin{pmatrix} -2 & -1 & 0 \\ 0 & -\sqrt{2} & -\sqrt{2} \\ 0 & 0 & 0 \\ 0 & 0 & 2 \end{pmatrix} \\ &= \begin{pmatrix} -2 & -1 & 0 \\ 0 & -\sqrt{2} & -\sqrt{2} \\ 0 & 0 & -2 \\ 0 & 0 & 0 \end{pmatrix} \\ &= \begin{pmatrix} \hat{R} \\ 0 \end{pmatrix} =: R. \end{aligned}$$

Kontrolle:

$$Hv = \rho e_1 = \begin{pmatrix} -2 \\ 0 \end{pmatrix}.$$

Insgesamt haben wir eine QR-Zerlegung gefunden:

$$\begin{aligned} Q^T &:= Q_3 Q_2 Q_1 \\ A = QR \quad &\Leftrightarrow \quad Q^T A = R. \end{aligned}$$

Die Orthogonalmatrix lautet

$$Q = \begin{pmatrix} 0 & -\frac{1}{2}\sqrt{2} & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & \frac{1}{2}\sqrt{2} & -\frac{1}{2}\sqrt{2} \\ -1 & 0 & 0 & 0 \\ 0 & -\frac{1}{2}\sqrt{2} & -\frac{1}{2} & -\frac{1}{2} \end{pmatrix}.$$

Kontrolle:

$$QR = \begin{pmatrix} 0 & -\frac{1}{2}\sqrt{2} & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & \frac{1}{2}\sqrt{2} & -\frac{1}{2}\sqrt{2} \\ -1 & 0 & 0 & 0 \\ 0 & -\frac{1}{2}\sqrt{2} & -\frac{1}{2} & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} -2 & -1 & 0 \\ 0 & -\sqrt{2} & -\sqrt{2} \\ 0 & 0 & 0 \\ 0 & 0 & 2 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & -\sqrt{2} \\ 2 & 1 & 0 \\ 0 & 1 & 2 \end{pmatrix} = A.$$

Das Verfahren verläuft wie folgt:
Seien Orthogonalmatrizen $Q_i, \ldots, Q_1 \in \mathbb{R}^{m \times n}$ bereits berechnet mit

$$Q_i Q_{i-1} \cdot \ldots \cdot Q_1 A = \begin{pmatrix} R_i & C_i \\ 0 & M_i \end{pmatrix}$$

für eine obere Dreiecksmatrix $R_i \in \mathbb{R}^{i \times i}$ und geeignete Matrizen $C_i \in \mathbb{R}^{i \times (n-i)}$ und $M_i \in \mathbb{R}^{(m-i) \times (m-i)}$.
Mit $v \in \mathbb{R}^{m-i}$ bezeichnen wir den ersten Spaltenvektor von M_i. Hieraus bestimmen wir $H_{i+1} \in \mathbb{R}^{(m-i) \times (m-i)}$ und setze

$$Q_{i+1} = \begin{pmatrix} I_i & 0 \\ 0 & H_{i+1} \end{pmatrix} \in \mathbb{R}^{m \times m}$$

mit $I_i \in \mathbb{R}^{i \times i}$. Dann ist

$$Q_{i+1} Q_i \cdot \ldots \cdot Q_1 A = \begin{pmatrix} I_i & 0 \\ 0 & H_{i+1} \end{pmatrix} \begin{pmatrix} R_i & C_i \\ 0 & M_i \end{pmatrix} = \begin{pmatrix} R_{i+1} & C_{i+1} \\ 0 & M_{i+1} \end{pmatrix}$$

mit einer neuen oberen Dreiecksmatrix $R_{i+1} \in \mathbb{R}^{(i+1) \times (i+1)}$ und geeigneten Matrizen C_{i+1} und M_{i+1}.
Nach spätestens $l = \min\{m-1, n\}$ Schritten ist dann

$$Q_l Q_{l-1} \cdot \ldots \cdot Q_1 A = \begin{pmatrix} \hat{R} \\ 0 \end{pmatrix} = R$$

mit $\hat{R} \in \mathbb{R}^{n \times n}$ obere Dreicksmatrix.

Algorithmus 4.4 Householder-Spiegelung

1: Setze $l := \min\{m-1, n\}$.
2: **for** $i = 1 : l$ **do**
3: $\quad v := A(i:m, i)$
4: $\quad$ **if** $v_1 \geq 0$ **then**
5: $\quad\quad \rho = -\|v\|$
6: $\quad$ **else**
7: $\quad\quad \rho = +\|v\|$
8: $\quad$ **end if**
9: $\quad u^{(i)} := v - \rho e_1$
10: $\quad u^{(i)} := u^{(i)} / \left\|u^{(i)}\right\|$
11: $\quad A(i:m, i:m) := A(i:m, i:m) - 2u^{(i)}(u^{(i)})^T A(i:m, i:m)$
12: **end for**

Bemerkung 4.24. Algorithmus 4.4 überschreibt die Matrix A mit der oberen Dreiecksmatrix $\hat{R}$ und speichert $u^{(i)}$ zur Berechnung von Q.

Kapitel 5

CG-Verfahren

Dieses Kapitel befasst sich mit dem bekannten CG-Verfahren (CG = Conjugate Gradient) zur Lösung von

$$Ax = b$$

mit einer symmetrischen und positiv definiten Matrix $A \in \mathbb{R}^{n\times n}$.

5.1 Herleitung des CG-Verfahrens

Das CG-Verfahren ist eine iterative Methode zur Lösung von $Ax = b$. Die Herleitung des CG-Verfahrens erfolgt über ein quadratisches Optimierungsproblem. Dazu betrachten wir ein quadratisches Funktional

$$f : \mathbb{R}^n \to \mathbb{R}, \quad f(x) := \frac{1}{2}x^T Ax - b^T x$$

und das Optimierungsproblem

$$\min_{x\in\mathbb{R}^n} f(x).$$

Lemma 5.1. *Es sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit, $b \in \mathbb{R}^n$, sowie*

$$f : \mathbb{R}^n \to \mathbb{R}, \quad f(x) := \frac{1}{2}x^T Ax - b^T x.$$

Dann ist $x^* \in \mathbb{R}^n$ genau dann eine Lösung des linearen Gleichungssystems

$$Ax = b,$$

wenn x^* das Minimierungsproblem

$$\min_{x\in\mathbb{R}^n} f(x)$$

löst.

Beweis. Da $A \in \mathbb{R}^{n\times n}$ positiv definit ist, ist die Inverse A^{-1} ebenso positiv definit. Wir definieren ein Hilfsfunktional

$$g : \mathbb{R}^n \to \mathbb{R}, \quad g(x) := f(x) + \frac{1}{2}b^T A^{-1} b.$$

Die Funktion g unterscheidet sich von f nur um einen konstanten Term $\frac{1}{2}b^TA^{-1}b$ und somit gilt

$$x^* \text{ löst } \min_{x\in\mathbb{R}^n} f(x) \quad \Leftrightarrow \quad x^* \text{ löst } \min_{x\in\mathbb{R}^n} g(x).$$

Ferner ist

$$\begin{aligned}\frac{1}{2}(Ax-b)^TA^{-1}(Ax-b) &= \frac{1}{2}(Ax-b)^T(x-A^{-1}b)\\ &= \frac{1}{2}(x^TAx - x^TAA^{-1}b - b^Tx + b^TA^{-1}b)\\ &= \frac{1}{2}(x^TAx - 2b^Tx + b^TA^{-1}b)\\ &= \frac{1}{2}x^TAx - b^Tx + \frac{1}{2}b^TA^{-1}b\\ &= f(x) + \frac{1}{2}b^TA^{-1}b = g(x).\end{aligned}$$

Folglich ist

$$x^* \text{ löst } \min_{x\in\mathbb{R}^n} f(x) \quad \Leftrightarrow \quad x^* \text{ löst } \min_{x\in\mathbb{R}^n} \frac{1}{2}(Ax-b)^TA^{-1}(Ax-b).$$

Da A^{-1} positiv definit ist, gilt

$$\frac{1}{2}(Ax-b)^TA^{-1}(Ax-b) \geq 0 \quad \forall x \in \mathbb{R}^n$$

und

$$\frac{1}{2}(Ax-b)^TA^{-1}(Ax-b) = 0 \quad \Leftrightarrow \quad Ax = b.$$

Insgesamt gilt also

$$Ax^* = b \quad \Leftrightarrow \quad x^* \text{ löst } \min_{x\in\mathbb{R}^n} f(x).$$

□

Bemerkung 5.2. Die Aussage

$$Ax^* = b \quad \Leftrightarrow \quad x^* \text{ löst } \min_{x\in\mathbb{R}^n} f(x)$$

folgt auch unmittelbar aus den klassischen notwendigen und hinreichenden Optimalitätsbedingungen für differenzierbare Minimierungsaufgaben.

Definition 5.3. Es sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit.

(i): Zwei Vektoren $x, y \in \mathbb{R}^n \setminus \{0\}$ heißen *A-orthogonal* (A-konjugiert), wenn gilt

$$x^TAy = 0.$$

(ii): Es seien $d^0, d^1, \ldots, d^{n-1} \in \mathbb{R}^n \setminus \{0\}$ paarweise A-orthogonale Vektoren, d.h.

$$(d^i)^T A d^j = 0 \quad \forall i, j \in \{0, \ldots, n-1\} : i \neq j.$$

Das Verfahren der *sukzessiven 1D-Minimierung* entlang $d^0, d^1, \ldots, d^{n-1}$ ist wie folgt definiert:

$$\begin{cases} x^{k+1} = x^k + t_k d^k \\ f(x^k + t_k d^k) = \min_{t \in \mathbb{R}} f(x^k + t d^k) \\ k = 0, \ldots, n-1 \end{cases}$$

mit $x^0 \in \mathbb{R}^n$ Startvektor.

Bemerkung 5.4. Die paarweise A-orthogonalen Vektoren $d^0, d^1, \ldots, d^{n-1} \in \mathbb{R}^n \setminus \{0\}$ lassen sich zum Beispiel mittels Gram-Schmidt bestimmen.

Lemma 5.5. *Es sei $A \in \mathbb{R}^{n \times n}$ symmetrisch und positiv definit. Ferner seien $d^0, d^1, \ldots, d^{n-1} \in \mathbb{R}^n \setminus \{0\}$ paarweise A-orthogonale Vektoren. Dann sind $d^0, d^1, \ldots, d^{n-1}$ linear unabhängig.*

Beweis. Annahme: $d^0, d^1, \ldots, d^{n-1} \in \mathbb{R}^n \setminus \{0\}$ wären nicht linear unabhängig. Dann existiert ein Index $m \in \{0, \ldots, n-1\}$, so dass

$$d^m = \sum_{\substack{j=0 \\ j \neq m}}^{n-1} \lambda_j d^j$$

mit $\lambda_j \in \mathbb{R}$, $j \in \{0, \ldots, n-1\}$ und es gibt (mindestens) einen Index $i \neq m$ mit $\lambda_i \neq 0$. Folglich:

$$\begin{aligned} 0 = (d^i)^T A d^m = (d^i)^T A \sum_{\substack{j=0 \\ j \neq m}}^{n-1} \lambda_j d^j &= \sum_{\substack{j=0 \\ j \neq m}}^{n-1} \lambda_j \underbrace{(d^i)^T A d^j}_{=0 \text{ für } i \neq j} \\ &= \lambda_i (d^i)^T A d^i, \end{aligned}$$

also

$$0 = (d^i)^T A d^i \quad \Leftrightarrow \quad d^i = 0.$$

Dies ist ein Widerspruch zur Annahme, also folgt die Behauptung. □

Satz 5.6. *Es seien $A \in \mathbb{R}^{n \times n}$ symmetrisch und positiv definit, $b \in \mathbb{R}^n$, und $d^0, d^1, \ldots, d^{n-1} \in \mathbb{R}^n \setminus \{0\}$ paarweise A-orthogonale Vektoren. Dann konvergiert das Verfahren der sukzessiven 1D-Minimierung entlang $d^0, d^1, \ldots, d^{n-1}$ nach spätestens n Schritten gegen die Lösung $x^* \in \mathbb{R}^n$ von*

$$Ax = b.$$

Für jedes $k = 0, \ldots, n-1$ gilt mit $g^k = Ax^k - b$

$$t_k = -\frac{(g^k)^T d^k}{(d^k)^T A d^k}$$

und

$$(g^{k+1})^T d^j = 0 \quad \forall j = 0, \dots, k.$$

Beweis. Laut Konstruktion löst jedes $t_k \in \mathbb{R}$ die Minimierungsaufgabe

$$\min_{t \in \mathbb{R}} \varphi(t) := f(x^k + t d^k).$$

Aus der notwendigen und hinreichenden Optimalitätsbedingung

$$\varphi'(t_k) = 0$$

folgt

$$t_k = -\frac{(g^k)^T d^k}{(d^k)^T A d^k}. \tag{5.1}$$

Beachte, dass $(d^k)^T A d^k \neq 0$ ist, da $d^k \neq 0$ und A positiv definit ist. Für alle $k = 0, \dots, n-1$ gilt:

$$\begin{aligned}(g^{k+1})^T d^k = (Ax^{k+1} - b)^T d^k &= (Ax^k + t_k A d^k - b)^T d^k \\ &= (Ax^k - b + t_k A d^k)^T d^k \\ &= (g^k)^T d^k + t_k (d^k)^T A d^k.\end{aligned}$$

Also folgt

$$(g^{k+1})^T d^k \overset{(5.1)}{=} 0. \tag{5.2}$$

Laut Voraussetzung ist $(d^j)^T A d^i = 0$ für alle $i \neq j$ und somit

$$\begin{aligned}(g^{i+1} - g^i)^T d^j &= (Ax^{i+1} - b - (Ax^i - b))^T d^j \\ &= (Ax^{i+1} - Ax^i)^T d^j \\ &= (Ax^i - Ax^i + t_i A d^i)^T d^j \\ &= t_i (d^i)^T A d^j = 0 \quad \forall i \neq j.\end{aligned} \tag{5.3}$$

Aus (5.2) und (5.3) erhalten wir

$$(g^{k+1})^T d^j = \underbrace{(g^{j+1})^T d^j}_{=0 \ (5.2)} + \sum_{i=j+1}^{k} \underbrace{(g^{i+1} - g^i)^T d^j}_{=0 \ (5.3)} = 0 \quad \forall j = 0, \dots, k.$$

Es bleibt zu zeigen, dass das Verfahren nach spätestens n Schritten die Lösung von $Ax = b$ liefert. Wir wissen aber

$$(g^n)^T d^j = 0 \quad \forall j = 0, \dots, n-1.$$

Daher ist

$$g^n = 0,$$

da $d^0, d^1, \ldots, d^{n-1}$ linear unabhängig sind. Es folgt

$$0 = g^n = Ax^n - b$$

und insgesamt

$$Ax^n = b.$$

□

Bemerkung 5.7. Das Verfahren der sukzessiven 1D-Minimierung entlang $d^0, d^1, \ldots, d^{n-1}$ konvergiert stets gegen die Lösung von $Ax = b$. Der Nachteil dieses Verfahrens besteht darin, dass man vorab die paarweisen A-orthogonalen Vektoren $d^0, d^1, \ldots, d^{n-1} \in \mathbb{R}^n \setminus \{0\}$ bestimmen muss. Das kann numerisch teuer sein.

Beim CG-Verfahren bestimmt man die Vektoren $d^0, d^1, \ldots, d^{n-1} \in \mathbb{R}^n \setminus \{0\}$ nicht vorab. Diese werden sukzessiv im Verfahren mit der Bedingung

$$\nabla f(x^k)^T d^k < 0$$

erzeugt, d.h. $d^k \in \mathbb{R}^n \setminus \{0\}$ ist eine Abstiegsrichtung für f in x^k. Beachte:

$$\nabla f(x^k) = Ax^k - b = g(k).$$

Wir starten mit

$$d^0 := -\nabla f(x^0) = -(Ax^0 - b) = -g^0.$$

Nun gehen wir davon aus, dass $d^0, d^1, \ldots, d^l$ ($l \in \{0, \ldots, n-2\}$) mit

$$(d^i)^T A d^j = 0 \quad \forall i \neq j$$

vorliegen. Analog zum Verfahren der sukzessiven 1D-Minimierung setzen wir

$$x^{l+1} = x^l + t_l d^l,$$
$$t_l = -\frac{(g^l)^T d^l}{(d^l)^T A d^l}.$$

Ist

$$g^{l+1} \overset{\text{Def.}}{=} Ax^{l+1} - b = 0,$$

so sind wir fertig. Ist $g^{l+1} \neq 0$, dann betrachten wir den Ansatz

$$d^{l+1} := -g^{l+1} + \sum_{i=0}^{l} \beta_i^l d^i \tag{A}$$

mit geeigneten Koeffizienten $\beta_i^l \in \mathbb{R}$, so dass

$$(d^{l+1})^T A d^j = 0 \quad \forall j = 0, \ldots, l.$$

Aus

$$(d^i)^T A d^j = 0 \quad \forall i \neq j = 0, \ldots, l$$

folgt unmittelbar

$$\beta_j^l = \frac{(g^{l+1})^T A d^j}{(d^j)^T A d^j} \quad \forall j = 0, \ldots, l.$$

Bemerkung 5.8. Wir sehen leicht, dass d^{l+1} aus dem Ansatz (A) eine Abstiegsrichtung für f in x^{l+1} ist, denn:

$$\begin{aligned}\nabla f(x^{l+1})^T d^{l+1} = (Ax^{l+1} - b)d^{l+1} &\overset{\text{Def.}}{=} (g^{l+1})^T d^{l+1} \\ &\overset{\text{(A)}}{=} (g^{l+1})^T(-g^{l+1} + \sum_{i=0}^{l} \beta_i^l d^i) \\ &= -\|g^{l+1}\|_2^2 + \sum_{i=0}^{l} \beta_i^l \underbrace{(g^{l+1})^T d^i}_{=0} \\ &= -\|g^{l+1}\|_2^2\end{aligned}$$

Die gleiche Aussage gilt auch für die alten Iterationen:

$$\underbrace{(g^j)^T}_{=\nabla f(x^j)} d^j = -\|g^j\|_2^2 \quad \forall j = 0, \ldots, l+1.$$

Lemma 5.9. *Es gilt:*

$$\begin{aligned}\beta_j^l &= 0 \quad \forall j = 0, \ldots, l-1 \quad \textit{und} \\ \beta_l^l &= \frac{\|g^{l+1}\|_2^2}{\|g^l\|_2^2}.\end{aligned}$$

Beweis. Für $j = 0, \ldots, l$ ist

$$(g^{l+1})^T g^j \overset{\text{(A)}}{=} (g^{l+1})^T(\sum_{i=0}^{j-1} \beta_i^{j-1} d^i - d^j) = \sum_{i=0}^{j-1} \beta_i^{j-1} (g^{l+1})^T d^i - (g^{l+1})^T d^j = 0. \tag{5.4}$$

Weiter ist

$$\beta_j^l = \frac{(g^{l+1})^T A d^j}{(d^j)^T A d^j} = \frac{(g^{l+1})^T (g^{j+1} - g^j)}{t_j (d^j)^T A d^j}, \tag{5.5}$$

weil:

$$(g^{j+1} - g^j) \overset{\text{Def.}}{=} (Ax^{j+1} - b - Ax^j + b) = Ax^j + t_j A d^j - Ax^j = t_j A d^j.$$

Aus (5.4) und (5.5) folgt

$$\beta_j^l = 0 \quad \forall j = 0, \ldots, l-1$$

und

$$\beta_l^l = \frac{(g^{l+1})^T A d^l}{(d^l)^T A d^l} = \frac{\|g^{l+1}\|_2^2}{\|g^l\|_2^2},$$

denn:

$$t_l (d^l)^T A d^l = (g^{l+1} - g^l)^T d^l = \underbrace{(g^{l+1})^T d^l}_{=0} - (g^l)^T d^l = -(g^l)^T d^l \overset{\text{Bem.}}{=} \|g^l\|_2^2.$$

□

Algorithmus 5.1 CG-Verfahren zur Lösung von $Ax = b$ mit A s.p.d.

1: Wähle $x^0 \in \mathbb{R}^n$.
2: Setze $g^0 = Ax^0 - b$ und $d^0 = -g^0$.
3: **for** $k = 0, 1, 2, \ldots$ **do**
4: $\quad t_k = -\frac{(g^k)^T d^k}{(d^k)^T A d^k} = \frac{\|g^k\|_2^2}{(d^k)^T A d^k}$
5: $\quad x^{k+1} = x^k + t_k d^k$
6: $\quad g^{k+1} = Ax^{k+1} - b = g^k + t_k A d^k$
7: $\quad \beta_k = \frac{\|g^{k+1}\|_2^2}{\|g^k\|_2^2}$
8: $\quad d^{k+1} = -g^{k+1} + \beta_k d^k$
9: $\quad$ **if** $\|g^{k+1}\|_2 = 0$ **then**
10: $\quad\quad$ STOP
11: $\quad$ **end if**
12: **end for**

Bemerkung 5.10. In der Praxis ersetzt man das Abbruchkriterium

$$\|g^{k+1}\|_2 = 0$$

durch

$$\frac{\|g^{k+1}\|_2}{\|g^0\|_2} \leq \varepsilon$$

mit $\varepsilon > 0$ „klein“.

Satz 5.11. *Es sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit, $b \in \mathbb{R}^n$, und*

$$f : \mathbb{R}^n \to \mathbb{R}, \quad f(x) := \frac{1}{2} x^T A x - b^T x.$$

Dann liefert der Algorithmus 5.1 (CG-Verfahren) nach spätestens n Schritten die Lösung $x^* \in \mathbb{R}^n$ von

$$Ax = b \quad \Leftrightarrow \quad \min_{x\in\mathbb{R}^n} f(x).$$

Ist $m \in \{0, \ldots, n\}$ die kleinste Zahl mit

$$x^m = x^*,$$

dann gilt

$$\begin{aligned}
(d^k)^T A d^j &= 0 && \forall k = 1, \ldots, m \quad \forall j = 0, \ldots, k-1 \quad (\textit{A-Orthogonalität}) \\
(g^k)^T g^j &= 0 && \forall k = 1, \ldots, m \quad \forall j = 0, \ldots, k-1 \\
(g^k)^T d^j &= 0 && \forall k = 1, \ldots, m \quad \forall j = 0, \ldots, k-1 \\
(g^k)^T d^k &= -\|g^k\|_2^2 && \forall k = 1, \ldots, m \quad (\textit{Abstiegsrichtung, siehe Bemerkung}).
\end{aligned}$$

Beweis. Es bleibt nur noch zu zeigen, dass Algorithmus 4.1 nach spätestens n Schritten konvergiert. Ist $g^m = 0$ für $m < n$, dann sind wir fertig. Andernfalls erzeugt der Algorithmus paarweise A-orthogonale Vektoren $d^0, d^1, \ldots, d^{n-1} \in \mathbb{R}^n \setminus \{0\}$, und somit ist der Algorithmus 5.1 äquivalent zum Verfahren der sukzessiven 1D-Minimierungsaufgabe entlang $d^0, d^1, \ldots, d^{n-1}$. Folglich ist $x^n = x^*$ wegen der Konvergenz der sukzessiven 1D-Minimierung entlang $d^0, d^1, \ldots, d^{n-1}$. □

5.2 Charakterisierung des CG-Verfahrens

Im Folgenden sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit und $b \in \mathbb{R}^n$. Wir untersuchen weiter das CG-Verfahren.

Definition 5.12. Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit. Dann definieren wir die *A-gewichtete Norm* auf $\mathbb{R}^n$ wie folgt:

$$\|\cdot\|_{\mathbb{R}^n_A} : \mathbb{R}^n \to \mathbb{R}, \quad \|x\|_{\mathbb{R}^n_A} := \sqrt{x^T A x}.$$

Bemerkung 5.13. Da A symmetrisch und positiv definit ist, definiert

$$\langle\cdot,\cdot\rangle_{\mathbb{R}^n_A} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n, \quad \langle x, y\rangle_{\mathbb{R}^n_A} := x^T A y$$

ein Skalarprodukt auf $\mathbb{R}^n$ und

$$\|x\|_{\mathbb{R}^n_A} = \sqrt{\langle x, x\rangle_{\mathbb{R}^n_A}}$$

ist eine Norm auf $\mathbb{R}^n$.

Wir verwenden folgende Notation:

$$\begin{aligned} g^k &= Ax^k - b && \text{(Gradient)} \\ r^k &= b - Ax^k = -g^k && \text{(Residuum)}. \end{aligned}$$

Definition 5.14 (Krylov-Raum)**.** Der folgende Raum

$$\mathcal{K}_k := \mathcal{K}_k(r^0, A) := \operatorname{Span}\{r^0, Ar^0, \ldots, A^{k-1}r^0\}$$

heißt der von dem Anfangsresiduum r^0 und der Matrix A aufgespannte k-te *Krylov-Raum*.

Lemma 5.15. *Bricht der Algorithmus 5.1 nicht vorzeitig ab, so gelten die folgenden Identitäten:*

$$\begin{aligned} \operatorname{Span}\{d^0, d^1, \ldots, d^{k-1}\} = \operatorname{Span}\{g^0, g^1, \ldots, g^{k-1}\} &= \operatorname{Span}\{r^0, r^1, \ldots, r^{k-1}\} \\ &= \operatorname{Span}\{r^0, Ar^0, \ldots, A^{k-1}r^0\} \\ &= \mathcal{K}_k(r^0, A) = \mathcal{K}_k, \end{aligned}$$

für $k = 1, 2, \ldots$.

Aufgrund der Definition ist

$$\begin{aligned} x^k = x^{k-1} + t_{k-1}d^{k-1} &= (x^{k-2} + t_{k-2}d^{k-2}) + t_{k-1}d^{k-1} \\ &= \ldots = x^0 + \sum_{i=0}^{k-1} t_i d^i. \end{aligned}$$

Daher ist

$$\begin{aligned} x^k = x^0 + \sum_{i=0}^{k-1} t_i d^i &\in x^0 + \mathrm{Span}\{d^0, \ldots, d^{k-1}\} \\ &\overset{\text{Lemma 5.15}}{=} x^0 + \mathcal{K}_k. \end{aligned}$$

Mit anderen Worten ist die k-te Iterierte $x^k \in \mathbb{R}^n$ des CG-Verfahrens stets ein Element des affinen Raums $x^0 + \mathcal{K}_k$.

Lemma 5.16. *Die k-te Iterierte $x^k \in \mathbb{R}^n$ des CG-Verfahrens löst das folgende Minimierungsproblem*

$$\min_{x \in x^0 + \mathcal{K}_k} f(x)$$

mit dem Zielfunktional $f(x) = \frac{1}{2}x^T Ax - b^T x$.

Satz 5.17. *Es sei $A \in \mathbb{R}^{n \times n}$ symmetrisch und positiv definit, $b \in \mathbb{R}^n$, und $x^* := A^{-1}b$. Dann sind die folgenden Aussagen äquivalent:*

(i) x^k *löst* $\min_{x \in x^0 + \mathcal{K}_k} f(x)$,

(ii) x^k *löst* $\min_{x \in x^0 + \mathcal{K}_k} \|x - x^*\|^2_{\mathbb{R}^n_A}$ *(x^k minimert den Fehler $\|x - x^*\|^2_{\mathbb{R}^n_A}$ auf $x^0 + \mathcal{K}_k$.),*

(iii) x^k *löst* $\min_{x \in x^0 + \mathcal{K}_k} \|Ax - b\|^2_{\mathbb{R}^n_{A^{-1}}}$ *(x^k minimert das Residuum $\|Ax - b\|^2_{\mathbb{R}^n_{A^{-1}}}$ auf $x^0 + \mathcal{K}_k$.),*

wobei $\|y\|_{\mathbb{R}^n_A} = \sqrt{y^T Ay}$ und $\|y\|_{\mathbb{R}^n_{A^{-1}}} = \sqrt{y^T A^{-1} y}$.

Beweis.

$(i) \Leftrightarrow (ii)$: Laut Definition gilt

$$\begin{aligned} \|x - x^*\|^2_{\mathbb{R}^n_A} = (x - x^*)^T A(x - x^*) &= x^T Ax - 2x^T Ax^* + (x^*)^T Ax^* \\ &\overset{Ax^* = b}{=} x^T Ax - 2b^T x + b^T x^* \\ &\overset{Def.}{=} 2f(x) + b^T x^*. \end{aligned}$$

Somit gilt:

$$x^k \text{ löst } \min_{x \in x^0 + \mathcal{K}_k} f(x) \quad \Leftrightarrow \quad x^k \text{ löst } \min_{x \in x^0 + \mathcal{K}_k} \|x^k - x^*\|^2_{\mathbb{R}^n_A}.$$

$(ii) \Leftrightarrow (iii)$: Laut Definition gilt

$$\begin{aligned}\|x - x^*\|^2_{\mathbb{R}^n_A} = (x - x^*)^T A(x - x^*) = (x - x^*)^T AA^{-1}A(x - x^*) \\ \overset{A=A^T}{=} (A(x - x^*))^T A^{-1}(A(x - x^*)) \\ \overset{Ax^*=b}{=} (Ax - b)^T A^{-1}(Ax - b) \\ \overset{Def.}{=} \|Ax - b\|^2_{\mathbb{R}^n_{A^{-1}}}.\end{aligned}$$

Somit gilt:

$$x^k \text{ löst } \min_{x \in x^0 + \mathcal{K}_k} \|x^k - x^*\|^2_{\mathbb{R}^n_A} \quad \Leftrightarrow \quad x^k \text{ löst } \min_{x \in x^0 + \mathcal{K}_k} \|Ax - b\|^2_{\mathbb{R}^n_{A^{-1}}}.$$

□

Folgerung 5.18. Jede k-te Iterierte x^k des CG-Verfahrens minimiert sowohl den Fehler $\|x - x^*\|_{\mathbb{R}^n_A}$ als auch das Residuum $\|Ax - b\|_{\mathbb{R}^n_{A^{-1}}}$ auf $x^0 + \mathcal{K}_k$.

5.3 Konvergenzanalyse des CG-Verfahrens

Ist $A \in \mathbb{R}^{n \times n}$ symmetrisch und positiv definit, so wissen wir, dass das CG-Verfahren nach spätestens n Schritten die exakte Lösung $x^* \in \mathbb{R}^n$ von

$$Ax = b$$

liefert. In diesem Abschnitt wollen wir den Fehler

$$\|x - x^*\|_{\mathbb{R}^n_A}$$

für jede CG-Iterierte $x^k \in \mathbb{R}^n$ analysieren. Dazu betrachten wir zuerst die *Tschebyscheff-Polynome*.

Definition 5.19 (Tschebyscheff-Polynom)**.** Es sei $n \in \mathbb{N} \cup \{0\}$. Dann heißt

$$T_n : [-1, 1] \to \mathbb{R}, \quad T_n(x) := \cos(n \arccos(x))$$

n-tes *Tschebyscheff-Polynom* auf dem Intervall $[-1, 1]$.

Satz 5.20 (Eigenschaften der Tschebyscheff-Polynome)**.** *Es gelten die folgenden Aussagen:*

(i) $T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$ *für* $n = 1, 2, \ldots$, *wobei* $T_0 \equiv 1$ *und* $T_1(x) = x$.

(ii) T_n *ist ein Polynom vom Grad n.*

(iii) T_{n+1} *besitzt n+1 Nullstellen bei*

$$x_i = \cos\left(\frac{2i + 1}{2n + 2}\pi\right), \quad i = 0, 1, \ldots, n.$$

(iv) T_{n+1} *besitzt die folgende Darstellung:*

$$T_{n+1}(x) = 2^n \prod_{i=0}^{n} (x - x_i), \quad n = 0, 1, \ldots.$$

(v) Es gilt $|T_{n+1}(x)| \leq 1$ *für alle* $x \in [-1, 1]$ *und*

$$T_{n+1}(\bar{x}_i) = (-1)^i \quad \textit{bei} \quad \bar{x}_i = \cos\left(\frac{i\pi}{n+1}\right) \quad \textit{für} \quad i = 0, 1, \ldots, n+1.$$

Beweis.

zu (i): $T_0(x) \overset{Def.}{=} \cos(0 \arccos(x)) = \cos(0) = 1 \quad \forall x \in [-1, 1]$. Außerdem gilt

$$T_1(x) \overset{Def.}{=} \cos(1 \arccos(x)) = x \quad \forall x \in [-1, 1].$$

Wir benutzen das Additionstheorem

$$\cos(a + b) = \cos(a)\cos(b) - \sin(a)\sin(b)$$

mit $a := n \arccos(x)$ und $b := \arccos(x)$, und somit erhalten wir

$$\begin{aligned} T_{n+1}(x) &\overset{Def.}{=} \cos((n+1)\arccos(x)) \\ &= \cos(n \arccos(x))\cos(\arccos(x)) - \sin(n \arccos(x))\sin(\arccos(x)) \\ &= T_n(x)x - \sin(n \arccos(x))\sin(\arccos(x)). \end{aligned}$$

Nun benutzen wir ein anderes Additionstheorem

$$\sin\left(\frac{a+b}{2}\right)\sin\left(\frac{a-b}{2}\right) = \frac{1}{2}(\cos(b) - \cos(a))$$

mit $a := (n+1)\arccos(x)$ und $b := (n-1)\arccos(x)$, und somit:

$$\begin{aligned} \sin(n \arccos(x))\sin(\arccos(x)) &= \frac{1}{2}(\cos((n-1)\arccos(x)) \\ &\quad - \cos((n+1)\arccos(x))) \\ &= \frac{1}{2}\left(T_{n-1}(x) - T_{n+1}(x)\right). \end{aligned}$$

Somit erhalten wir

$$T_{n+1}(x) = xT_n(x) - \frac{1}{2}(T_{n-1}(x) - T_{n+1}(x))$$

und insgesamt ergibt sich

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x).$$

zu (ii): Dies folgt unmittelbar aus (i).

zu (iii): Dies folgt unmittelbar aus der Definition

$$T_{n+1}(x) = \cos((n+1)\arccos(x)).$$

zu (iv): Dies folgt aus (iii) und (i).

zu (v): Dies folgt unmittelbar aus der Definition.

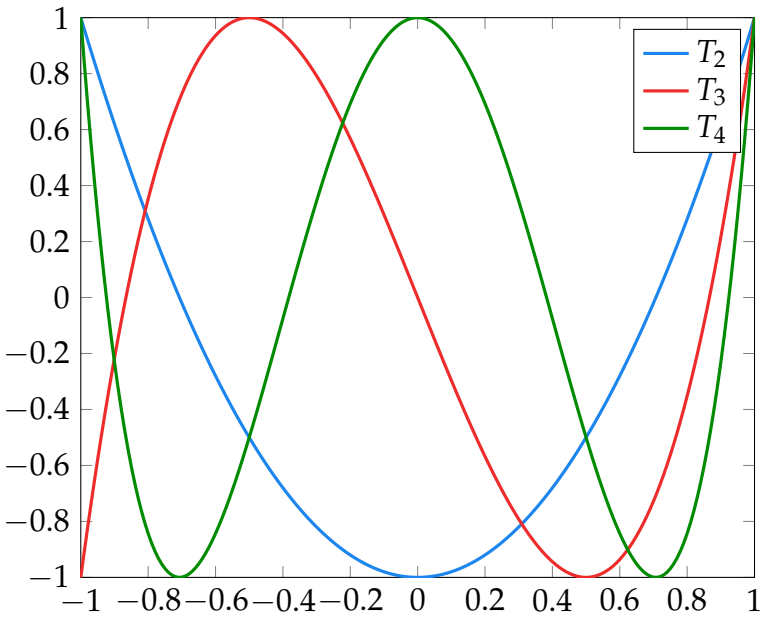
□

Aus der Rekursivformel

$$\begin{cases} T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x), & x \in [-1,1], \quad n = 1,2,\dots \\ T_0 \equiv 1, \quad T_1(x) = x \end{cases}$$

ergibt sich

$$\begin{aligned} T_2(x) &= 2x^2 - 1 \\ T_3(x) &= 4x^3 - 3x \\ T_4(x) &= 8x^4 - 8x^2 + 1. \end{aligned}$$



Definition 5.21. Das Tschebyscheff-Polynom auf $\mathbb{R}$ ist definiert durch die Rekursivformel

$$\begin{cases} T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x) \\ T_0 \equiv 1, \quad T_1(x) = x \\ \text{für } x \in \mathbb{R}, \quad n = 1,2,\dots \end{cases}$$

Lemma 5.22. *Das Tschebyscheff-Polynom auf* $\mathbb{R} \setminus [-1,1]$ *besitzt die folgende Darstellung*

$$T_n(x) = \frac{1}{2}\left[(x+\sqrt{x^2-1})^n + (x-\sqrt{x^2-1})^n\right]$$

für alle $x \in \mathbb{R}$ *mit* $|x| > 1$.

Beweis. Wir verwenden

$$\begin{aligned}\cos(n\theta) + i\sin(n\theta) &= (\cos(\theta) + i\sin(\theta))^n,\\ \cos(n\theta) - i\sin(n\theta) &= (\cos(\theta) - i\sin(\theta))^n.\end{aligned}$$

Addition beider Gleichungen liefert:

$$\cos(n\theta) = \frac{1}{2}\left[(\cos(\theta) + i\sin(\theta))^n + (\cos(\theta) - i\sin(\theta))^n\right].$$

Wir setzen $\theta := \arccos(x)$ für $x \in [-1,1]$ und erhalten mit $\sin^2(\theta) + \cos^2(\theta) = 1$:

$$\begin{aligned}\cos(n\arccos(x)) &= \frac{1}{2}\left[(\cos(\theta) + i\sin(\theta))^n + (\cos(\theta) - i\sin(\theta))^n\right]\\ &= \frac{1}{2}\left[(x+i\sqrt{1-x^2})^n + (x-i\sqrt{1-x^2})^n\right].\end{aligned}$$

Daher ist

$$T_n(x) = \frac{1}{2}\left[(x+i\sqrt{1-x^2})^n + (x-i\sqrt{1-x^2})^n\right] \quad \forall x \in [-1,1].$$

Da sowohl T_n als auch die rechte Seite ein Polynom vom Grad n ist und diese beiden Polynome für alle $x \in [-1,1]$ übereinstimmen, so gilt diese Gleichung auch für alle $x \in \mathbb{R}$. Mit

$$i\sqrt{1-x^2} = \sqrt{x^2-1} \quad \forall x \in \mathbb{R} \text{ mit } |x| > 1$$

folgt

$$T_n(x) = \frac{1}{2}\left[(x+\sqrt{x^2-1})^n + (x-\sqrt{x^2-1})^n\right] \quad \forall x \in [-1,1].$$

□

Folgerung 5.23. Das Tschebyscheff-Polynom genügt der folgenden Ungleichung

$$\left|T_n\left(\frac{\alpha+1}{\alpha-1}\right)\right| \geq \frac{1}{2}\left(\frac{\sqrt{\alpha}+1}{\sqrt{\alpha}-1}\right)^n$$

für alle $\alpha > 1$.

Beweis. Sei $\alpha > 1$. Dann gilt für $x := \frac{\alpha+1}{\alpha-1} > 1$

$$x \pm \sqrt{x^2-1} = \frac{\sqrt{\alpha} \pm 1}{\sqrt{\alpha} \mp 1}$$

und somit folgt

$$\begin{aligned} T_n\left(\frac{\alpha+1}{\alpha-1}\right) &\overset{Lemma}{=} \frac{1}{2}\left[\left(\frac{\sqrt{\alpha}+1}{\sqrt{\alpha}-1}\right)^n + \left(\frac{\sqrt{\alpha}-1}{\sqrt{\alpha}+1}\right)^n\right] \\ &\geq \frac{1}{2}\left(\frac{\sqrt{\alpha}+1}{\sqrt{\alpha}-1}\right)^n. \end{aligned}$$

□

Folgerung 5.24. Es gilt

$$T_n(-x) = (-1)^n T_n(x) \quad \text{für alle} \quad n \in \mathbb{N}.$$

Beweis mittels vollständiger Induktion:

Induktionsanfang $n = 1$:

$$T_1(-x) \overset{\text{Def.}}{=} -x \overset{\text{Def.}}{=} (-1)^1 T_1(x).$$

Induktionsannahme: Die Aussage gelte für alle Indizes $k = 1, \ldots, n$ mit $n \in \mathbb{N}$.
Nun zeigen wir, dass die Aussage auch für $n+1$ gilt. Lauf Definition ist aber

$$\begin{aligned} T_{n+1}(-x) &\overset{\text{Def.}}{=} (-2x)T_n(-x) - T_{n-1}(-x) \\ &\overset{\text{IA}}{=} (-2x)(-1)^n T_n(x) - (-1)^{n-1} T_{n-1}(x) \\ &= (-1)^{n+1} 2x T_n(x) - (-1)^{n-1}(-1)^2 T_{n-1}(x) \\ &= (-1)^{n+1}\left[2x T_n(x) - T_{n-1}(x)\right] \\ &= (-1)^{n+1} T_{n+1}(x). \end{aligned}$$

□

Satz 5.25 (Fehlerabschätzung für das CG-Verfahren)**.** *Es sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit, $b \in \mathbb{R}^n$, sowie $x^* = A^{-1}b$. Dann erfüllt die durch Algorithmus 5.1 (CG-Verfahren) erzeugte k-te Iterierte die folgende Fehlerabschätzung:*

$$\|x^k - x^*\|_{\mathbb{R}^n_A} \leq 2\left(\frac{\sqrt{\alpha}-1}{\sqrt{\alpha}+1}\right)^k \|x^0 - x^*\|_{\mathbb{R}^n_A} \quad \textit{für alle} \quad k = 1, 2, \ldots,$$

mit $\frac{\sqrt{\alpha}-1}{\sqrt{\alpha}+1} \in [0,1)$, *und* $\alpha := \operatorname{cond}_2(A) \in [1,\infty)$ *der Spektralkondition von A.*

Bemerkung 5.26. Es ist

$$\operatorname{cond}_2(A) = \|A\|_2 \, \|A^{-1}\|_2 \quad \left(= \frac{\lambda_{max}(A)}{\lambda_{min}(A)}, \text{ falls } A \text{ symmetrisch und positiv definit} \right).$$

Beweis. Wir beweisen die Aussage für $\alpha > 1$. Die Aussage für $\alpha = 1$ erhalten wir später als Folgerung.
Wir setzen

$$e^k := x^k - x^* \quad (e: \text{ error})$$

für den Fehler in der k-ten Iteration. Aus

$$r^0 \overset{Def.}{=} b - Ax^0 \overset{b=Ax^*}{=} A(x^* - x^0) = -Ae^0$$

folgt

$$A^j r^0 = -A^{j+1} e^0.$$

Also

$$\mathcal{K}_k \overset{Def.}{=} \operatorname{Span}\{r^0, Ar^0, \ldots, A^{k-1}r^0\} = \operatorname{Span}\{Ae^0, \ldots, A^k e^0\}.$$

Somit gilt:

$$\begin{aligned}
\|e^k\|_{\mathbb{R}^n_A} = \|x^k - x^*\|_{\mathbb{R}^n_A} \overset{Satz}{=} \min_{x \in x^0 + \mathcal{K}_k} \|x - x^*\|_{\mathbb{R}^n_A} &= \min_{z \in \mathcal{K}_k} \|x^0 - x^* + z\|_{\mathbb{R}^n_A} \\
&= \min_{\mu_1, \ldots, \mu_k \in \mathbb{R}} \|x^0 - x^* + \sum_{j=1}^{k} \mu_j A^j e^0\|_{\mathbb{R}^n_A} \\
&= \min_{\mu_1, \ldots, \mu_k \in \mathbb{R}} \|e^0 + \sum_{j=1}^{k} \mu_j A^j e^0\|_{\mathbb{R}^n_A} \\
&= \min_{\substack{p \in \prod_k \\ p(0)=1}} \|p(A)e^0\|_{\mathbb{R}^n_A}, \qquad (5.6)
\end{aligned}$$

wobei $\prod_k$ den Raum aller Polynome vom Grad k bezeichnet. Da A symmetrisch ist, existiert eine Orthonormalbasis $\{v^1, \ldots, v^n\}$ aus Eigenvektoren von A zu Eigenwerten $\lambda_i \in \mathbb{R}^+$, $i = 1, \ldots, n$. Beachte, dass $\lambda_i > 0$ für alle $i = 1, \ldots, n$ gilt, da A symmetrisch und positiv definit ist. Somit hat e^0 die Darstellung

$$e^0 = \sum_{i=1}^{n} \gamma_i v^i$$

und folglich

$$\begin{aligned}
\|e^0\|^2_{\mathbb{R}^n_A} \overset{Def.}{=} \langle e^0, Ae^0 \rangle = \left\langle \sum_{i=1}^{n} \gamma_i v^i, A\left(\sum_{j=1}^{n} \gamma_j v^j \right) \right\rangle \\
\overset{Av^j = \lambda_j v^j}{=} \left\langle \sum_{i=1}^{n} \gamma_i v^i, \sum_{j=1}^{n} \gamma_j \lambda_j v^j \right\rangle \\
\overset{\langle v^i, v^j \rangle = \delta ij}{=} \sum_{i=1}^{n} \gamma_i^2 \lambda_i. \qquad (5.7)
\end{aligned}$$

Analog gilt:

$$\begin{aligned}
\|p(A)e^0\|^2_{\mathbb{R}^n_A} \overset{Def.}{=} \left\langle p(A)e^0, A(p(A)e^0)\right\rangle = & \left\langle \sum_{i=1}^n \gamma_i p(A)v^i, A\left(\sum_{j=1}^n \gamma_j p(A)v^j\right)\right\rangle \\
\overset{p(A)v^j=p(\lambda_j)v^j}{=} & \left\langle \sum_{i=1}^n \gamma_i p(\lambda_i)v^i, A\left(\sum_{j=1}^n \gamma_j p(\lambda_j)v^j\right)\right\rangle \\
\overset{Av^j=\lambda_j v^j}{=} & \left\langle \sum_{i=1}^n \gamma_i p(\lambda_i)v^i, \sum_{j=1}^n \gamma_j p(\lambda_j)\lambda_j v^j\right\rangle \\
\overset{\langle v^i,v^j\rangle=\delta ij}{=} & \sum_{i=1}^n \gamma_i^2 p(\lambda_i)^2\lambda_i. \qquad (5.8)
\end{aligned}$$

Zusammen ergibt sich

$$\begin{aligned}
\|e^k\|_{\mathbb{R}^n_A} \overset{(5.6)}{=} \min_{\substack{p\in\prod_k\\ p(0)=1}} \|p(A)e^0\|_{\mathbb{R}^n_A} \overset{(5.8)}{=} & \min_{\substack{p\in\prod_k\\ p(0)=1}} \left(\sum_{i=1}^n \gamma_i^2 p(\lambda_i)^2\lambda_i\right)^{\frac{1}{2}} \\
\leq & \min_{\substack{p\in\prod_k\\ p(0)=1}} \max_{1\leq j\leq n} |p(\lambda_j)| \left(\sum_{i=1}^n \gamma_i^2\lambda_i\right)^{\frac{1}{2}} \\
\overset{(5.7)}{=} & \min_{\substack{p\in\prod_k\\ p(0)=1}} \max_{1\leq j\leq n} |p(\lambda_j)| \|e^0\|\mathbb{R}^n A.
\end{aligned}$$

Es bleibt nur noch zu zeigen, dass

$$\min_{\substack{p\in\prod_k\\ p(0)=1}} \max_{1\leq j\leq n} |p(\lambda_j)| \leq 2\left(\frac{\sqrt{2}-1}{\sqrt{2}+1}\right)^k$$

gilt. Dazu betrachten wir ein spezielles Polynom:

$$p_k(\lambda) := \frac{T_k(F(\lambda))}{T_k(F(0))},$$

mit

$$F(\lambda) = \frac{2\lambda - (\lambda_{max}+\lambda_{min})}{\lambda_{max}-\lambda_{min}}.$$

Da T_k ein Polynom vom Grad k ist, gilt $p_k \in \prod_k$ und auf Grund der Konstruktion gilt

$p(0) = 1$. Somit gilt

$$\begin{aligned}
\min_{\substack{p\in\prod_k \\ p(0)=1}} \max_{1\le j\le n} |p(\lambda_j)| \le \max_{1\le j\le n} |p_k(\lambda_j)| = & \max_{1\le j\le n} \frac{|T_k(F(\lambda_j))|}{|T_k(F(0))|} \\
\overset{1}{\le} & \frac{1}{|T_k(F(0))|} \\
= & \left| T_k\left(-\frac{\lambda_{max}+\lambda_{min}}{\lambda_{max}-\lambda_{min}}\right)\right|^{-1} \\
\overset{Folgerung}{=} & \left| T_k\left(\frac{\lambda_{max}+\lambda_{min}}{\lambda_{max}-\lambda_{min}}\right)\right|^{-1} \\
\overset{\alpha=\frac{\lambda_{max}}{\lambda_{min}}}{=} & \left| T_k\left(\frac{\alpha+1}{\alpha-1}\right)\right|^{-1} \\
\overset{Folgerung}{\le} & \left(\frac{1}{2}\left(\frac{\sqrt{\alpha}+1}{\sqrt{\alpha}-1}\right)^k\right)^{-1} \\
= & 2\left(\frac{\sqrt{\alpha}-1}{\sqrt{\alpha}+1}\right)^k.
\end{aligned}$$

Insgesamt gilt also

$$\|e^k\|\mathbb{R}^n A \le \min_{\substack{p\in\prod_k \\ p(0)=1}} \max_{1\le j\le n} |p(\lambda_j)|\|e^0\|\mathbb{R}^n A \le \max_{1\le j\le n} |p_k(\lambda_j)|\|e^0\|\mathbb{R}^n A \le 2\left(\frac{\sqrt{\alpha}-1}{\sqrt{\alpha}+1}\right)^k \|e^0\|\mathbb{R}^n A.$$

□

Bemerkung 5.27. Ist A schlecht konditioniert ($\text{cond}_2(A) \gg 1$), so ist

$$\left(\frac{\sqrt{\alpha}-1}{\sqrt{\alpha}+1}\right) \approx 1-\varepsilon$$

mit $\varepsilon > 0$ „klein“. In diesem Fall kann das CG-Verfahren sehr langsam konvergieren. Ist umgekehrt

$$\text{cond}_2(A) \approx 1+\varepsilon,$$

so ist $\left(\frac{\sqrt{\alpha}-1}{\sqrt{\alpha}+1}\right)$ nahe Null und somit konvergiert das CG-Verfahren in diesem Fall recht schnell.

Folgerung 5.28. Besitzt $A \in \mathbb{R}^{n\times n}$ insgesamt $m \le n$ verschiedene Eigenwerte, so bricht das CG-Verfahren nach spätestens m Schritten mit der Lösung von $Ax = b$ ab.

[1] $|T_k(x)| \le 1 \quad \forall x \in [-1,1]$

Beweis. Im Beweis vom obigen Satz haben wir bereits gezeigt, dass

$$\|e^k\|_{\mathbb{R}^n A} \leq \min_{\substack{p \in \prod_k \\ p(0)=1}} \max_{1 \leq j \leq n} |p(\lambda_j)| \|e^0\|_{\mathbb{R}^n A}$$

mit $e^k = x^k - x^*$ und $\lambda_1, \ldots, \lambda_n$ Eigenwerte von A gilt. Hat A insgesamt nur $m \leq n$ verschiedene Eigenwerte, so definieren wir

$$p_m(x) := \kappa \prod_{i=1}^{m} (x - \lambda_i)$$

mit $\kappa \in \mathbb{R}$, so dass $p_m(0) = 1$ gilt. Folglich ist

$$\|e^k\|_{\mathbb{R}^n A} \leq \min_{\substack{p \in \prod_k \\ p(0)=1}} \max_{1 \leq j \leq n} |p(\lambda_j)| \|e^0\|_{\mathbb{R}^n A} \overset{p=p_m}{\leq} 0.$$

Dies bedeutet

$$\|x^m - x^*\| = 0,$$

also

$$x^m = x^*.$$

□

Fazit 5.29. Die Konditionszahl der Matrix A ist wichtig für die Konvergenzgeschwindigkeit des CG-Verfahrens.

5.4 Das präkonditionierte CG-Verfahren

Im Folgenden sei $A \in \mathbb{R}^{n \times n}$ symmetrisch und positiv definit, und $b \in \mathbb{R}^n$. Unsere Konvergenzanalyse zeigt, dass das CG-Verfahren „schneller" konvergiert, wenn die Matrix A gut konditioniert ist. Aus diesem Grund ist es naheliegend, das lineare Gleichungssystem $Ax = b$ umzuschreiben in der folgenden Gestalt:

$$C^{-1}Ax = C^{-1}b \tag{5.9}$$

mit einer regulären Matrix $C \in \mathbb{R}^{n \times n}$, so dass

$$\text{cond}_2(C^{-1}A) \overset{Def.}{=} \|C^{-1}A\|_2 \|A^{-1}C\|_2 < \|A\|_2 \|A^{-1}\|_2 = \text{cond}_2(A).$$

Beachte, dass $C^{-1}A$ im Allgemeinen nicht symmetrisch ist, so dass das CG-Verfahren nicht direkt auf (5.9) angewendet werden kann. Deshalb benutzen wir die Umformulierung

$$\begin{aligned} C^{-1}AC^{-T}C^Tx = C^{-1}b, \quad &\text{mit } C^{-T} := (C^{-1})^T = (C^T)^{-1} \\ \Leftrightarrow \quad \tilde{A}\tilde{x} = \tilde{b}, \end{aligned} \tag{5.10}$$

mit

$$\left.\begin{array}{lcll} \tilde{A} & := & C^{-1}AC^{-T} & \text{symmetrisch und positiv definit,} \\ \tilde{x} & := & C^T x, & \\ \tilde{b} & := & C^{-1}b. & \end{array}\right\} \tag{5.11}$$

Nun ist $\tilde{A} \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit, und das auf (5.10) angewandte CG-Verfahren lautet dann wie folgt (siehe Algorithmus 5.1):

1: Wähle $\tilde{x}^0 \in \mathbb{R}^n$.
2: Setze $\tilde{g}^0 = \tilde{A}\tilde{x}^0 - \tilde{b}$ und $\tilde{d}^0 = -\tilde{g}^0$.
3: **for** $k = 0, 1, 2, \ldots$ **do**
4: $\quad \tilde{t}_k = \frac{\|\tilde{g}^k\|_2^2}{(\tilde{d}^k)^T\tilde{A}\tilde{d}^k}$
5: $\quad \tilde{x}^{k+1} = \tilde{x}^k + \tilde{t}_k\tilde{d}^k$
6: $\quad \tilde{g}^{k+1} = \tilde{A}\tilde{x}^{k+1} - \tilde{b} = \tilde{g}^k + \tilde{t}_k\tilde{A}\tilde{d}^k$
7: $\quad \tilde{\beta}_k = \frac{\|\tilde{g}^{k+1}\|_2^2}{\|\tilde{g}^k\|_2^2}$
8: $\quad \tilde{d}^{k+1} = -\tilde{g}^{k+1} + \tilde{\beta}_k\tilde{d}^k$
9: **end for**

Setzen wir

$$\begin{aligned} x^k &:= C^{-T}\tilde{x}^k \quad \text{und} \\ d^k &:= C^{-T}\tilde{d}^k, \end{aligned}$$

so ergibt sich aus (5.11):

(i): $\tilde{g}^k = \tilde{A}\tilde{x}^k - \tilde{b} = C^{-1}AC^{-T}C^Tx^k - C^{-1}b = C^{-1}(Ax^k - b) = C^{-1}g^k$,

(ii): $\tilde{t}_k \overset{(i)}{=} \frac{(\tilde{g}^k)^T\tilde{g}^k}{(\tilde{d}^k)^T\tilde{A}\tilde{d}^k} = \frac{(g^k)^T(C^{-T}C^{-1}g^k)}{(C^Td^k)^TC^{-1}AC^{-T}C^Td^k} = \frac{(g^k)^T(C^{-T}C^{-1}g^k)}{(d^k)^TAd^k}$,

(iii): $C^Tx^{k+1} = \tilde{x}^{k+1} = \tilde{x}^k + \tilde{t}_k\tilde{d}^k = C^Tx^k + \tilde{t}_kC^Td^k$,

(iv): $C^{-1}g^{k+1} \overset{(i)}{=} \tilde{g}^{k+1} = \tilde{g}^k + \tilde{t}_k\tilde{A}\tilde{d}^k \overset{(i)}{=} C^{-1}g^k + \tilde{t}_kC^{-1}AC^{-T}C^Td^k = C^{-1}(g^k + \tilde{t}_kAd^k)$,

(v): $\tilde{\beta}_k = \frac{\|\tilde{g}^{k+1}\|_2^2}{\|\tilde{g}^k\|_2^2} \overset{(i)}{=} \frac{(g^{k+1})^T(C^{-T}C^{-1}g^{k+1})}{(g^k)^T(C^{-T}C^{-1}g^k)}$,

(vi): $C^Td^{k+1} = \tilde{d}^{k+1} = -\tilde{g}^{k+1} + \tilde{\beta}_k\tilde{d}^k = -C^{-1}g^{k+1} + \tilde{\beta}_kC^Td^k$.

Multiplizieren wir die Aufdatierungsformeln (iii) und (vi) von links mit C^{-T} und die Formel (iv) von links mit C, so kommen wir auf das folgende CG-Verfahren für (5.10):

1: Wähle $x^0 \in \mathbb{R}^n$.

2: Setze $g^0 = Ax^0 - \tilde{b}$ und $\tilde{d}^0 = -C^{-T}C^{-1}g^0$.
3: **for** $k = 0, 1, 2, \ldots$ **do**
4: $\quad t_k = \frac{(g^k)^T(C^{-T}C^{-1}g^k)}{(d^k)^T A d^k}$ ▷ (vgl. (ii))
5: $\quad x^{k+1} = x^k + t_k d^k$ ▷ (vgl. (iii))
6: $\quad g^{k+1} = g^k + t_k A d^k$ ▷ (vgl. (iv))
7: $\quad \beta_k = \frac{(g^{k+1})^T(C^{-T}C^{-1}g^{k+1})}{(g^k)^T(C^{-T}C^{-1}g^k)}$ ▷ (vgl. (v))
8: $\quad d^{k+1} = -C^{-T}C^{-1}g^{k+1} + \beta_k d^k$ ▷ (vgl. (vi))
9: **end for**

Setzen wir $P := CC^T$, so erhalten wir das präkonditionierte CG-Verfahren.

Algorithmus 5.2 Präkonditioniertes CG-Verfahren

(S1) Wähle $x^0 \in \mathbb{R}^n$ und $P \in \mathbb{R}^{n \times n}$ symmetrisch und positiv definit.
(S2) Setze $g^0 = Ax^0 - b$ und bestimme d^0 aus $Pd^0 = -g^0$.
(S3) Setze $z^0 = d^0$.
1: **for** $k = 0, 1, 2, \ldots$ **do**
2: $\quad t_k = \frac{(g^k)^T z^k}{(d^k)^T A d^k}$
3: $\quad x^{k+1} = x^k + t_k d^k$
4: $\quad g^{k+1} = g^k + t_k A d^k$
5: $\quad$ Bestimme z^{k+1} aus $Pz^{k+1} = -g^{k+1}$.
6: $\quad \beta_k = \frac{(g^{k+1})^T z^{k+1}}{(g^k)^T z^k}$
7: $\quad d^{k+1} = z^k + \beta_k d^k$
8: $\quad$ **if** $\left\|g^{k+1}\right\|_2 = 0$ **then**
9: $\quad\quad$ STOP
10: $\quad$ **end if**
11: **end for**

Bemerkung 5.30. Die Matrix P wird als Präkonditionierer bezeichnet. Verschiedene Strategien zur effizienten Konstruktion von P findet man in der Literatur. Beispiele dafür sind

- die unvollständige Cholesky-Zerlegung,
- die unvollständige LR-Zerlegung,
- die splittingbasierte Zerlegung.

Kapitel 6

Nichtlineares Gleichungssystem

In diesem Kapitel beschäftigen wir uns mit einem nichtlinearen Gleichungssystem der Gestalt

$$F(x) = 0$$

mit einer nichtlinearen Funktion $F : \mathbb{R}^n \to \mathbb{R}^n$.

6.1 Differentialrechnung

Definition 6.1. Es sei $U \subset \mathbb{R}^n$ offen und $F : \mathbb{R}^n \supset U \to \mathbb{R}^m$.

(i) F heißt in $x \in U$ differenzierbar, wenn eine lineare Abbildung $F'(x) : \mathbb{R}^n \to \mathbb{R}^m$ existiert, so dass für alle $y \in U$ gilt:

$$F(y) = F(x) + F'(x)(y - x) + r(x, y),$$

und das Restglied $r(x, y)$ genügt der Bedingung:

$$\lim_{y \to x} \frac{r(x, y)}{\|y - x\|_2} = 0.$$

(ii) Die lineare Abbildung $F'(x) : \mathbb{R}^n \to \mathbb{R}^m$ heißt Ableitung von F in $x \in U$.

(iii) F heißt (auf U) differenzierbar, falls F in allen Punkten $x \in U$ differenzierbar ist.

Bemerkung 6.2.

(i) Ist $F : \mathbb{R}^n \supset U \to \mathbb{R}^m$ in $x \in U$ differenzierbar, so ist die Ableitung $F'(x) : \mathbb{R}^n \to \mathbb{R}^m$ eindeutig bestimmt.

(ii) Die Ableitung $F'(x) : \mathbb{R}^n \to \mathbb{R}^m$ ist laut Definition linear und somit stetig, da jede lineare Abbildung auf endlichdimensionalen Vektorräumen stetig ist.

(iii) Das Restglied ist wiederum eine Funktion $r(x, \cdot) : \mathbb{R}^n \supset U \to \mathbb{R}^m$. Diese hat die explizite Darstellung

$$r(x, y) = F(y) - F(x) - F'(x)(y - x) \quad \forall y \in U,$$

und

$$\lim_{y\to x}\frac{r(x,y)}{\|y-x\|_2}=0 \quad\Leftrightarrow\quad \lim_{y\to x}\frac{\|r(x,y)\|_2}{\|y-x\|_2}=0 \quad\Leftrightarrow\quad \lim_{y\to x}\frac{r_i(x,y)}{\|y-x\|_2}=0$$

für alle $i=1,\dots,m$ mit $r(x,y)=\begin{pmatrix} r_1(x,y)\\ \vdots\\ r_m(x,y)\end{pmatrix}$.

Geometrische Interpretation: Im Falle $F:\mathbb{R}^2\to\mathbb{R}$ ist der Graph von

$$y\mapsto F(x)+F'(x)(y-x)$$

die Tangentialebene im Punkt $F(x)$ an den Graphen von F.

Bemerkung 6.3 (Komponentenweise Differentiation)**.**

$$F=\begin{pmatrix}F_1\\ \vdots\\ F_m\end{pmatrix}:\mathbb{R}^n\supset U\to\mathbb{R}^m$$

mit Komponentenfunktionen $F_i:\mathbb{R}^n\supset U\to\mathbb{R}$. Dann ist F genau dann in $x\in U$ differenzierbar, wenn alle F_i in x differenzierbar sind. In diesem Fall gilt

$$F'(x)h=\begin{pmatrix}F_1'(x)h\\ \vdots\\ F_m'(x)h\end{pmatrix}\quad\forall h\in\mathbb{R}^n.$$

Definition 6.4 (Richtungsableitung)**.** Es sei $U\subset\mathbb{R}^n$ offen und $F:\mathbb{R}^n\supset U\to\mathbb{R}^m$. Dann heißt F in $x\in U$ in Richtung $h\in\mathbb{R}^n$ richtungsdifferenzierbar, wenn der Limes

$$\frac{\partial F}{\partial h}(x):=\lim_{t\to 0}\frac{F(x+th)-F(x)}{t}\in\mathbb{R}^m$$

existiert. Der Limes $\frac{\partial F}{\partial h}(x)\in\mathbb{R}^m$ heißt Richtungsableitung von F in x in Richtung h. Existiert $\frac{\partial F}{\partial h}(x)\in\mathbb{R}^m$ für alle Richtungen $h\in\mathbb{R}^n$, so heißt F in x richtungsdifferenzierbar.

Bemerkung 6.5. Im Falle $h=e_i$ heißt

$$\frac{\partial F}{\partial e_i}(x)$$

die i-te partielle Ableitung von F in x. Andere Notationen:

$$\frac{\partial F}{\partial e_i}(x)=\frac{\partial F}{\partial x_i}(x)=\partial_i F(x).$$

Lemma 6.6. *Es sei $U \subset \mathbb{R}^n$ offen und $F : \mathbb{R}^n \supset U \to \mathbb{R}^m$. Ist F in $x \in U$ differenzierbar, so ist F in x richtungsdifferenzierbar mit*

$$\frac{\partial F}{\partial h}(x) = F'(x)h \quad \forall h \in \mathbb{R}^n.$$

Beweis. Sei $h \in \mathbb{R}^n$ beliebig aber fest. Dann ist

$$x + th \in U$$

für hinreichend kleines $|t|$, weil U offen ist. Somit gilt für hinreichend kleines $|t|$:

$$\frac{F(x+th) - F(x)}{t} = \frac{F(x) + F'(x)(th) + r(x, x+th) - F(x)}{t} = F'(x)h + \frac{r(x, x+th)}{t}.$$

Aber

$$\frac{r(x, x+th)}{t} = \underbrace{\frac{r(x, x+th)}{\|x+th-x\|_2}}_{\to 0 \text{ für } t \to 0} \underbrace{\frac{\|th\|_2}{t}}_{=\pm\|h\|_2} \to 0 \quad \text{für } t \to 0.$$

Und somit

$$\lim_{t \to 0} \frac{F(x+th) - F(x)}{t} = F'(x)h.$$

□

Lemma 6.7. *Ist $F : \mathbb{R}^n \supset U \to \mathbb{R}^m$ ($U \subset \mathbb{R}^n$ offen) in $x \in U$ differenzierbar, so ist F in x stetig.*

Beweis. Sei $\{x_k\}_{k=1}^{\infty} \subset \mathbb{R}^n$ mit

$$\lim_{k \to \infty} \|x_k - x\|_2 = 0.$$

Dann gilt:

$$\|F(x_k) - F(x)\|_2 = \left\|F'(x)(x_k - x) + r(x, x_k)\right\|_2 \to 0 \text{ für } k \to \infty,$$

denn $F'(x) : \mathbb{R}^n \to \mathbb{R}^m$ ist stetig und $\lim_{k \to \infty} \|r(x, x_k)\|_2 = 0$. □

Fazit 6.8. Ist $F : \mathbb{R}^n \supset U \to \mathbb{R}^m$ in $x \in U$ differenzierbar, so ist

- F in x stetig und
- F in x richtungsdifferenzierbar mit $\frac{\partial F}{\partial h}(x) = F'(x)h \quad \forall h \in \mathbb{R}^n$.

Bemerkung 6.9. Ist $F : \mathbb{R}^n \supset U \to \mathbb{R}^m$ in $x \in U$ richtungsdifferenzierbar, so muss F in x nicht notwendigerweise differenzierbar sein (F muss auch nicht unbedingt in x stetig sein). Ein Beispiel dazu ist:

$$F : \mathbb{R}^2 \to \mathbb{R}, \quad F(x_1, x_2) = \begin{cases} 1 & \text{falls } x_1 = x_2^2 \text{ und } x_2 \neq 0 \\ 0 & \text{sonst} \end{cases}.$$

Diese Abbildung ist im Nullpunkt richtungsdifferenzierbar, ist dort aber nicht stetig.

Beispiel 6.10. Jede lineare Abbildung $F : \mathbb{R}^n \to \mathbb{R}^m$ ist differenzierbar mit der Ableitung
$F'(x) = F \quad \forall x \in \mathbb{R}^n$, denn

$$F(y) = F(x) + \underbrace{F(y-x)}_{=F'(x)(y-x)} + \underbrace{0}_{=r(x,y)} \qquad \forall x, y \in \mathbb{R}^n.$$

Beispiel 6.11. Betrachte

$$F : \mathbb{R}^n \to \mathbb{R}^{n-1}, \quad F(x) = x_1 \begin{pmatrix} x_2 \\ \vdots \\ x_n \end{pmatrix}.$$

Diese Abbildung ist auf $\mathbb{R}^n$ differenzierbar und für jedes $x \in \mathbb{R}^n$ gilt

$$F'(x)y = x_1 \begin{pmatrix} y_2 \\ \vdots \\ y_n \end{pmatrix} + y_1 \begin{pmatrix} x_2 \\ \vdots \\ x_n \end{pmatrix} \qquad \forall y \in \mathbb{R}^n.$$

Beweis. Offenbar ist $F'(x) : \mathbb{R}^n \to \mathbb{R}^{n-1}$, wie oben definiert, linear. Es bleibt also zu zeigen, dass das Restglied

$$r(x,y) = F(y) - F(x) - F'(x)(y-x)$$

die Bedingung

$$\lim_{y \to x} \frac{r(x,y)}{\|y-x\|_2} = 0$$

erfüllt (siehe Definition). Laut Definition gilt aber:

$$\begin{aligned}
\frac{r(x,y)}{\|y-x\|_2} &= \frac{F(y) - F(x) - F'(x)(y-x)}{\|y-x\|_2} \\
&= \frac{y_1 \begin{pmatrix} y_2 \\ \vdots \\ y_n \end{pmatrix} - x_1 \begin{pmatrix} x_2 \\ \vdots \\ x_n \end{pmatrix} - x_1 \begin{pmatrix} y_2 - x_2 \\ \vdots \\ y_n - x_n \end{pmatrix} - (y_1 - x_1) \begin{pmatrix} x_2 \\ \vdots \\ x_n \end{pmatrix}}{\|y-x\|_2} \\
&= \frac{y_1 \begin{pmatrix} y_2 - x_2 \\ \vdots \\ y_n - x_n \end{pmatrix} - x_1 \begin{pmatrix} y_2 - x_2 \\ \vdots \\ y_n - x_n \end{pmatrix}}{\|y-x\|_2} \\
&= \frac{(y_1 - x_1) \begin{pmatrix} y_2 - x_2 \\ \vdots \\ y_n - x_n \end{pmatrix}}{\|y-x\|_2}.
\end{aligned}$$

Daraus folgt

$$\frac{\|r(x,y)\|_2}{\|y-x\|_2} = \frac{\left\|(y_1-x_1)\begin{pmatrix} y_2-x_2 \\ \vdots \\ y_n-x_n \end{pmatrix}\right\|_2}{\|y-x\|_2} \leq \frac{\|y-x\|_2 \|y-x\|_2}{\|y-x\|_2} = \|y-x\|_2$$

und insgesamt ergibt sich

$$\lim_{y\to x} \frac{r(x,y)}{\|y-x\|_2} = 0.$$

□

Bemerkung 6.12. Ist $F : \mathbb{R}^n \supset U \to \mathbb{R}^m$ in $x \in U$ differenzierbar, so ist laut Definition $F'(x) : \mathbb{R}^n \to \mathbb{R}^m$ linear. Nach linearer Algebra kann somit $F'(x) : \mathbb{R}^n \to \mathbb{R}^m$ eindeutig durch eine Darstellungsmatrix dargestellt werden:

$$\begin{aligned} F'(x)y &= \left[F'(x)e_1 \cdots F'(x)e_n\right] y \\ &\overset{\text{Lemma}}{=} \left[\frac{\partial F}{\partial e_1}(x) \cdots \frac{\partial F}{\partial e_n}(x)\right] y \\ &= \begin{pmatrix} \frac{\partial F_1}{\partial e_1}(x) & \cdots & \frac{\partial F_1}{\partial e_n}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial F_m}{\partial e_1}(x) & \cdots & \frac{\partial F_m}{\partial e_n}(x) \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \forall y \in \mathbb{R}^n. \end{aligned}$$

Definition 6.13. Ist $F : \mathbb{R}^n \supset U \to \mathbb{R}^m$ in $x \in U$ differenzierbar, so heißt die Darstellungsmatrix

$$\mathbb{R}^{m\times n} \ni \left(\frac{\partial F_i}{\partial e_j}(x)\right) \overset{\text{Andere Notation}}{=} \left(\frac{\partial F_i}{\partial x_j}(x)\right) = \left(\partial_j F_i(x)\right)$$

Jacobi-Matrix bzw. Funktionalmatrix von F in x.

Beispiel 6.14. Ist $F : \mathbb{R}^n \supset U \to \mathbb{R}$ ($m = 1$) in $x \in U$ differenzierbar, so ist

$$F'(x)y = \nabla F(x)^T y = (\partial_1 F(x), \ldots, \partial_n F(x)) \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}.$$

Mit anderen Worten: Im Falle $m = 1$ stimmt die Jacobi-Matrix mit $\nabla F(x)^T$ überein.

Der folgende Satz liefert ein überaus wichtiges Kriterium für die Differenzierbarkeit einer Funktion $F : \mathbb{R}^n \supset U \to \mathbb{R}^m$.

Lemma 6.15 (Differenzierbarkeit und partielle Ableitungen). *Es sei $U \subset \mathbb{R}^n$ offen und $F : \mathbb{R}^n \supset U \to \mathbb{R}^m$. Existieren alle partiellen Ableitungen von F auf U und sind die Abbildungen*

$$\partial_i F : \mathbb{R}^n \supset U \to \mathbb{R}^m, \quad x \mapsto \partial_i F(x) \quad \forall i = 1, \ldots, n$$

stetig, so ist F auf U differenzierbar, und die Jacobi-Matrix von F in jedem $x \in U$ ist gegeben durch $(\partial_j F_i(x)) \in \mathbb{R}^{m \times n}$.

Beweis. Wir zeigen die Aussage für $m = 1$, d.h. $F : \mathbb{R}^n \supset U \to \mathbb{R}$. Die Aussage für $m > 1$ folgt danach mittels komponentenweiser Differentiation. Sei $x \in U$ beliebig aber fest. Wir definieren:

$$F'(x) : \mathbb{R}^n \to \mathbb{R}^m, \quad F'(x)y := \sum_{i=1}^{n} y_i \underbrace{\partial_i F(x)}_{\in \mathbb{R}}.$$

Da $U \subset \mathbb{R}^n$ offen ist, existiert ein $\varepsilon > 0$ mit

$$B_\varepsilon(x) := \{y \in \mathbb{R}^n \mid \|y - x\|_2 < \varepsilon\} \subset U.$$

Sei $y \in B_\varepsilon(x)$ mit $y \neq x$. Dann gilt:

$$\begin{aligned} F(y) - F(x) &= F(y_1, \ldots, y_n) - F(x_1, \ldots, x_n) \\ &= F(y_1, \ldots, y_n) - F(x_1, y_2, \ldots, y_n) \\ &+ F(x_1, y_2, \ldots, y_n) - F(x_1, x_2, y_3 \ldots, y_n) \\ &+ F(x_1, x_2, y_3 \ldots, y_n) - F(x_1, x_2, x_3, y_4 \ldots, y_n) \\ &\pm \ldots \\ &+ F(x_1, x_2, \ldots, x_{n-1}, y_n) - F(x_1, \ldots, x_n). \end{aligned}$$

In jeder Zeile verwenden wir nun den Mittelwertsatz aus der Analysis I:

$$\begin{aligned} F(y) - F(x) &= \partial_1 F(\xi_1, y_2, \ldots, y_n)(y_1 - x_1) \\ &+ \partial_2 F(x_1, \xi_2, y_3, \ldots, y_n)(y_2 - x_2) \\ &+ \ldots \\ &+ \partial_n F(x_1, x_2, \ldots, x_{n-1}, \xi_n)(y_n - x_n) \end{aligned}$$

mit ξ_i zwischen y_i und x_i. Folglich:

$$\begin{aligned} \frac{r(x,y)}{\|y - x\|_2} &= \frac{F(y) - F(x) - F'(x)(y - x)}{\|y - x\|_2} \\ &= \frac{\sum_{i=1}^{n} (y_i - x_i) \left[\partial_i F(x_1, \ldots, x_{i-1}, \xi_i, y_{i+1}, \ldots, y_n) - \partial_i F(x)\right]}{\|y - x\|_2}. \end{aligned}$$

Daraus folgt

$$\frac{r(x,y)}{\|y-x\|_2} = \sum_{i=1}^{n} \frac{y_i - x_i}{\|y-x\|_2} \underbrace{[\partial_i F(x_1,\dots,x_{i-1},\xi_i,y_{i+1},\dots,y_n - \partial_i F(x_1,\dots,x_n)]}_{\to 0 \text{ für } y\to x}$$
$$\to 0 \text{ für } y \to x,$$

da die Abbildungen $\partial_i F : \mathbb{R}^n \to \mathbb{R}$ laut Voraussetzung stetig sind. Beachte dabei, dass aus der Konvergenz von y gegen x die Konvergenz von ξ gegen x folgt. Insgesamt ergibt sich damit die Behauptung. □

Beispiel 6.16. Betrachte

$$F : \mathbb{R}^2 \to \mathbb{R}^2, \quad F(x) = \begin{pmatrix} \sin(x_1)\sin(x_2) \\ \cos(x_2) \end{pmatrix}.$$

Alle partiellen Ableitungen existieren:

$$\begin{aligned} \partial_1 F_1(x) &= \cos(x_1)\sin(x_2), \\ \partial_1 F_2(x) &= 0, \\ \partial_2 F_1(x) &= \sin(x_1)\cos(x_2), \\ \partial_2 F_2(x) &= -\sin(x_2). \end{aligned}$$

Die Abbildungen

$$x \mapsto \partial_1 F(x), \quad x \mapsto \partial_2 F(x)$$

sind stetig. Also ist $F : \mathbb{R}^2 \to \mathbb{R}^2$ differenzierbar mit

$$F'(x)y = \begin{pmatrix} \cos(x_1)\sin(x_2) & \sin(x_1)\cos(x_2) \\ 0 & -\sin(x_2) \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \quad \forall y \in \mathbb{R}^2.$$

Definition 6.17. Mit $\mathscr{L}(\mathbb{R}^n, \mathbb{R}^m)$ bezeichnen wir den Raum aller linearen Funktionen zwischen $\mathbb{R}^n$ und $\mathbb{R}^m$, d.h.

$$\mathscr{L}(\mathbb{R}^n, \mathbb{R}^m) = \{f : \mathbb{R}^n \to \mathbb{R}^m \text{ linear}\}.$$

Eine differenzierbare Abbildung $F : \mathbb{R}^n \supset U \to \mathbb{R}^m$ heißt (auf U) stetig differenzierbar, falls die Abbildung

$$F' : \mathbb{R}^n \supset U \to \mathscr{L}(\mathbb{R}^n, \mathbb{R}^m), \quad x \mapsto F'(x)$$

stetig ist. Dies ist äquivalent dazu, dass die Abbildung

$$(\partial_j F_i) : \mathbb{R}^n \supset U \to \mathbb{R}^{m\times n}, \quad x \mapsto (\partial_j F_i(x))$$

stetig ist.

Korollar 6.18. *Es sei U offen und $F : \mathbb{R}^n \supset U \to \mathbb{R}^m$. Existieren alle partiellen Ableitungen von F auf U und sind die Ableitungen*

$$\partial_i F : \mathbb{R}^n \supset U \to \mathbb{R}^m, \quad x \mapsto \partial_i F(x) \quad \forall i = 1, \ldots, n$$

stetig, so ist F stetig differenzierbar.

Beweis. Die Aussage folgt aus dem vorherigen Satz und der Definition der stetigen Differnzierbarkeit. □

6.2 Newton-Verfahren

Im Folgenden untersuchen wir das nichtlineare Gleichungssystem

$$F(x) = 0$$

mit einer stetig differenzierbaren Funktion $F : \mathbb{R}^n \supset U \to \mathbb{R}^n$ ($m = n$, $U = \mathbb{R}^n$).
Idee des Newton-Verfahrens: Angenommen es sei $x^k \in \mathbb{R}^n$ bereits berechnet. Dann gilt

$$F(x) = F(x^k) + F'(x^k)(x - x^k) + r(x^k, x).$$

Ist $x^k \approx x$, dann ist $r(x^k, x)$ sehr klein, so dass

$$F(x) \approx F(x^k) + F'(x^k)(x - x^k)$$

gilt. Es ist also naheliegend, das lineare Gleichungssystem

$$F(x^k) + F'(x^k)(x - x^k) = 0$$

zu lösen. Diese Aufgabe ist äquivalent zu:
Finde $x^{k+1} \in \mathbb{R}^n$ als Lösung von

$$\begin{aligned} & F'(x^k)(x - x^k) = -F(x^k) \\ \Leftrightarrow \quad & F'(x^k)d^k = -F(x^k) \quad \text{mit} \\ & x^{k+1} = x^k + d^k. \end{aligned}$$

Algorithmus 6.1 Newton-Verfahren

(S0) Wähle einen Startwert $x^0 \in \mathbb{R}^n$ und setze $k = 0$.
(S1) **if** $F(x^K) = 0$ **then** STOP **end if**
(S2) Bestimme $d^k \in \mathbb{R}^n$ als Lösung von

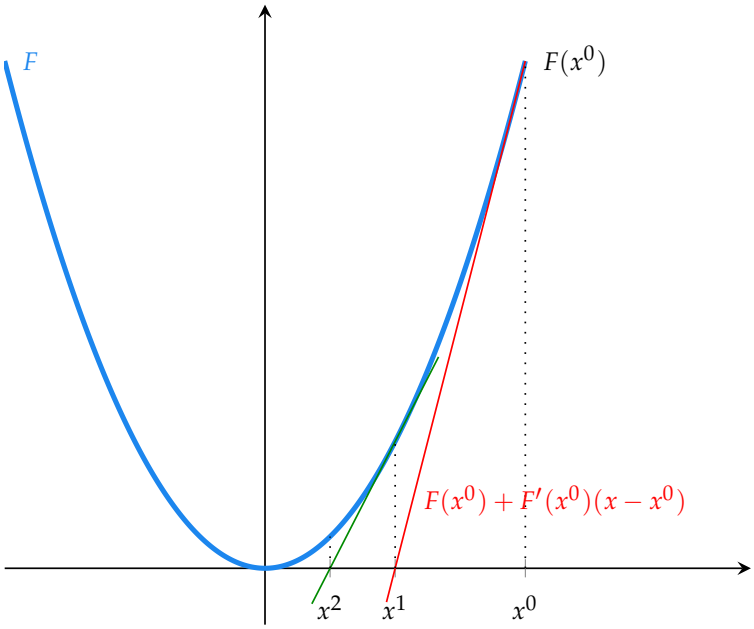
$$F'(x^k)d^k = -F(x^k).$$

(S3) Setze $x^{k+1} = x^k + d^k$, $k = k + 1$ und gehe zu (S1).

Bemerkung 6.19. In der Praxis ersetzen wir das Abbruchkriterium $F(x^k) = 0$ durch

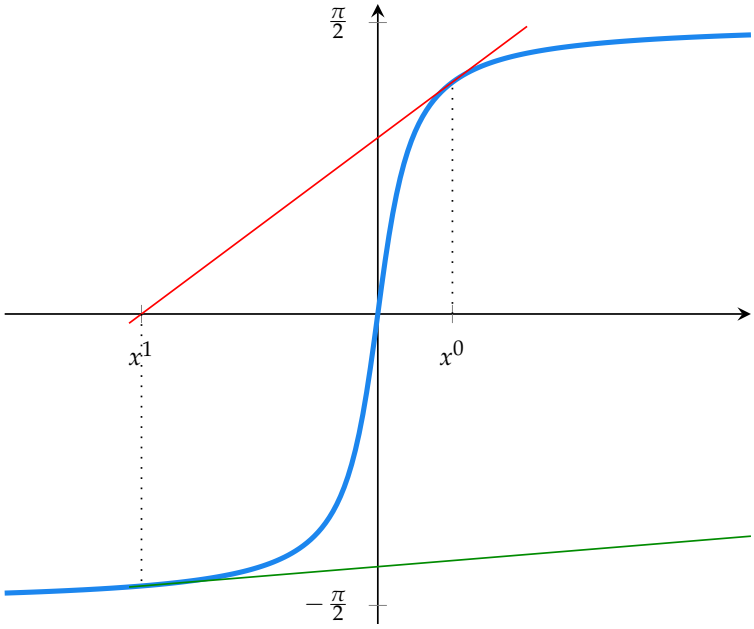
$$\|F(x^k)\| < \varepsilon \text{ mit } \varepsilon \text{ „klein“.}$$

Illustration des Newton-Verfahrens: $(F(x) = x^2)$



Bemerkung 6.20. Die Konvergenz des Newton-Verfahrens hängt im wesentlichen von dem Startwert ab! Ist der Startwert ungünstig gewählt, so kann das Newton-Verfahren unter Umständen nicht konvergieren! Dazu betrachten wir das Beispiel

$$F : \mathbb{R} \to \mathbb{R}, \quad F(x) = \arctan(x).$$



Unser Ziel ist es, die lokale Konvergenz des Newton-Verfahrens zu analysieren.

6.3 Lokale Konvergenz

Definition 6.21. Es sei $\{x_k\}_{k=1}^{\infty}$ eine gegen $\bar{x} \in \mathbb{R}^n$ konvergente Folge.

(i) Die Folge $\{x_k\}_{k=1}^{\infty}$ konvergiert *linear* mit Rate $\alpha \in (0,1)$ gegen $\bar{x}$, falls ein $n \in \mathbb{N}$ existiert, so dass
$$\|x^{k+1} - \bar{x}\|_2 \leq \alpha \|x^k - \bar{x}\|_2 \quad \forall k \geq n.$$

(ii) Die Folge $\{x_k\}_{k=1}^{\infty}$ konvergiert *superlinear* gegen $\bar{x}$, falls
$$\lim_{k\to\infty} \frac{\left\|x^{k+1} - \bar{x}\right\|_2}{\left\|x^k - \bar{x}\right\|_2} = 0.$$
Andere Notation: $\|x^{k+1} - \bar{x}\|_2 = o\left(\|x^k - \bar{x}\|_2\right)$ für $k \to \infty$.

(iii) Die Folge $\{x_k\}_{k=1}^{\infty}$ konvergiert *quadratisch* gegen $\bar{x}$, falls es eine Konstante $c > 0$ gibt, so dass
$$\|x^{k+1} - \bar{x}\|_2 \leq c\|x^k - \bar{x}\|_2^2 \quad \forall k \in \mathbb{N}.$$
Andere Notation: $\|x^{k+1} - \bar{x}\|_2 = \mathcal{O}\left(\|x^k - \bar{x}\|_2^2\right) \quad \forall k \in \mathbb{N}.$

Lemma 6.22. *Es sei $F : \mathbb{R}^n \to \mathbb{R}^n$ stetig differenzierbar und $\bar{x} \in \mathbb{R}^n$ eine Nullstelle von F. Ferner sei die Jacobi-Matrix von F in $\bar{x}$*
$$(\partial_j F_i(\bar{x})) \in \mathbb{R}^{n\times n}$$
regulär. Dann gibt es ein $\varepsilon > 0$ und ein $\beta > 0$ mit
$$\beta \|x - \bar{x}\|_2 \leq \|F(x)\|_2 \quad \forall x \in B_\varepsilon(\bar{x}).$$
Insbesondere ist $\bar{x} \in \mathbb{R}^n$ die einzige Nullstelle von F in $B_\varepsilon(\bar{x})$.

Beweis. Laut Voraussetzung gilt
$$\begin{aligned}\|x - \bar{x}\|_2 = \|(\partial_j F_i(\bar{x}))^{-1}(\partial_j F_i(\bar{x}))(x - \bar{x})\|_2 &\leq \|(\partial_j F_i(\bar{x}))^{-1}\|_2 \|(\partial_j F_i(\bar{x}))(x - \bar{x})\|_2 \\ &= \|(\partial_j F_i(\bar{x}))^{-1}\|_2 \|F'(\bar{x})(x - \bar{x})\|_2.\end{aligned}$$
Wir setzen
$$\beta := \frac{1}{2\left\|(\partial_j F_i(\bar{x}))^{-1}\right\|_2} \in \mathbb{R}^+.$$
Dann gilt
$$2\beta \|x - \bar{x}\|_2 \leq \left\|F'(\bar{x})(x - \bar{x})\right\|_2. \tag{6.1}$$
Laut Definition der Differenzierbarkeit gilt
$$\frac{\|r(\bar{x}, x)\|_2}{\|x - \bar{x}\|_2} = \frac{\|F(x) - F(\bar{x}) - F'(\bar{x})(x - \bar{x})\|_2}{\|x - \bar{x}\|_2} \to 0 \quad \text{für } x \to \bar{x}.$$

Folglich gibt es ein $\varepsilon > 0$, so dass

$$\frac{\|F(x) - F(\bar{x}) - F'(\bar{x})(x - \bar{x})\|_2}{\|x - \bar{x}\|_2} \leq \beta \quad \forall x \in B_\varepsilon(\bar{x}). \tag{6.2}$$

Mit (6.1), (6.2) und $F(\bar{x}) = 0$ folgt

$$\begin{aligned} 2\beta \|x - \bar{x}\|_2 \leq \left\|F'(\bar{x})(x - \bar{x})\right\|_2 &= \left\|F(x) - (F(x) - F(\bar{x}) - F'(\bar{x})(x - \bar{x}))\right\|_2 \|x - \bar{x}\|_2 \\ &\leq \|F(x)\|_2 + \left\|F(x) - F(\bar{x}) - F'(\bar{x})(x - \bar{x})\right\|_2 \\ &\leq \|F(x)\|_2 + \beta \|x - \bar{x}\|_2 \quad \forall x \in B_\varepsilon(\bar{x}). \end{aligned}$$

Insgesamt ergibt sich also

$$\beta \|x - \bar{x}\|_2 \leq \|F(x)\|_2 \quad \forall x \in B_\varepsilon(\bar{x}).$$

□

Lemma 6.23. *Es seien $A, B \in \mathbb{R}^{n \times n}$ mit $\|I - BA\|_2 < 1$. Dann sind A und B regulär und es gilt*

$$\|A^{-1}\|_2 \leq \frac{\|B\|_2}{1 - \|I - BA\|_2}.$$

Lemma 6.24. *Es sei $F : \mathbb{R}^n \to \mathbb{R}^n$ stetig differenzierbar und $\bar{x} \in \mathbb{R}^n$ eine Nullstelle von F. Ferner sei die Jacobi-Matrix von F in $\bar{x}$*

$$(\partial_j F_i(\bar{x})) \in \mathbb{R}^{n \times n}$$

regulär. Dann gibt es wieder ein $\varepsilon > 0$, so dass die Jacobi-Matrizen

$$(\partial_j F_i(x)) \in \mathbb{R}^{n \times n} \quad \forall x \in B_\varepsilon(\bar{x})$$

regulär sind, und es gilt

$$\|(\partial_j F_i(x))^{-1}\|_2 \leq 2\|(\partial_j F_i(\bar{x}))^{-1}\|_2 \quad \forall x \in B_\varepsilon(\bar{x}).$$

Beweis. Da $F : \mathbb{R}^n \to \mathbb{R}^n$ stetig differenzierbar ist, gibt es ein $\varepsilon > 0$, so dass

$$\left\|(\partial_j F_i(x)) - (\partial_j F_i(\bar{x}))\right\|_2 \leq \frac{1}{2\left\|(\partial_j F_i(\bar{x}))^{-1}\right\|_2} \quad \forall x \in B_\varepsilon(\bar{x}).$$

Also ist

$$\begin{aligned} \|I - (\partial_j F_i(\bar{x}))^{-1}(\partial_j F_i(x))\|_2 &= \|(\partial_j F_i(\bar{x}))^{-1}\left[(\partial_j F_i(\bar{x})) - (\partial_j F_i(x))\right]\|_2 \\ &\leq \|(\partial_j F_i(\bar{x}))^{-1}\|_2 \|(\partial_j F_i(\bar{x})) - (\partial_j F_i(x))\|_2 \\ &\leq \frac{1}{2} \quad \forall x \in B_\varepsilon(\bar{x}). \end{aligned}$$

Das vorherige Lemma mit

$$B := (\partial_j F_i(\bar{x}))^{-1} \quad \text{und} \quad A := (\partial_j F_i(x))$$

impliziert, dass $(\partial_j F_i(x)) \in \mathbb{R}^{n\times n}$ für alle $x \in B_\varepsilon(\bar{x})$ regulär ist und es gilt

$$\|(\partial_j F_i(x))^{-1}\|_2 \leq \frac{\left\|(\partial_j F_i(\bar{x}))^{-1}\right\|_2}{1 - \left\|I - (\partial_j F_i(\bar{x}))^{-1}(\partial_j F_i(x))\right\|_2} \leq 2\|(\partial_j F_i(\bar{x}))^{-1}\|_2 \quad \forall x \in B_\varepsilon(\bar{x}).$$

□

Satz 6.25 (Lokale superlineare Konvergenz des Newton-Verfahrens)**.** *Es sei $F : \mathbb{R}^n \to \mathbb{R}^n$ stetig differenzierbar und $\bar{x} \in \mathbb{R}^n$ eine Nullstelle von F. Ferner sei die Jacobi-Matrix*

$$(\partial_j F_i(\bar{x})) \in \mathbb{R}^{n\times n}$$

regulär. Dann gibt es ein $\varepsilon > 0$, so dass für jeden Startwert $x^0 \in B_\varepsilon(\bar{x})$ das Newton-Verfahren superlinear *gegen $\bar{x}$ konvergiert.*

Beweis. Laut Definition ist

$$r(x,y) = F(y) - F(x) - F'(x)(y-x) \quad \forall x, y \in \mathbb{R}^n.$$

Wir zeigen zunächst, dass das Restglied r auch die folgende Integraldarstellung besitzt:

$$r(x,y) = \int_0^1 F'(x + t(y-x))(y-x) - F'(x)(y-x)\,\mathrm{d}t \quad \forall x, y \in \mathbb{R}^n.$$

Seien $x, y \in \mathbb{R}^n$ beliebig aber fest. Wir definieren die Hilfsfunktion

$$G : \mathbb{R} \to \mathbb{R}^n, \quad G(t) = F(x + t(y-x)).$$

Nach Kettenregel ist

$$G'(t) = F'(x + t(y-x))(y-x).$$

Somit gilt

$$\int_0^1 F'(x + t(y-x))(y-x)\,\mathrm{d}t = \int_0^1 G'(t)\,\mathrm{d}t = G(1) - G(0) = F(y) - F(x).$$

Daraus folgt

$$\int_0^1 F'(x + t(y-x))(y-x)\,\mathrm{d}t = F(y) - F(x)$$

und wir erhalten

$$\begin{aligned} r(x,y) &= \int_0^1 F'(x + t(y-x))(y-x)\,\mathrm{d}t - F'(x)(y-x) \\ &= \int_0^1 F'(x + t(y-x))(y-x) - F'(x)(y-x)\,\mathrm{d}t. \end{aligned}$$

Aus der Integraldarstellung folgt

$$\begin{aligned}
\|r(x,\bar{x})\|_2 &= \left\| \int_0^1 F'(x+t(\bar{x}-x))(\bar{x}-x) - F'(x)(\bar{x}-x)\,\mathrm{d}t \right\|_2 \\
&\leq \int_0^1 \left\| F'(x+t(\bar{x}-x))(\bar{x}-x) - F'(x)(\bar{x}-x) \right\|_2 \,\mathrm{d}t \\
&= \int_0^1 \left\| \left[(\partial_j F_i(x+t(\bar{x}-x))) - (\partial_j F_i(x)) \right] (\bar{x}-x) \right\|_2 \,\mathrm{d}t \\
&\leq \int_0^1 \left\| (\partial_j F_i(x+t(\bar{x}-x))) - (\partial_j F_i(x)) \right\|_2 \|\bar{x}-x\|_2 \,\mathrm{d}t \\
&= \int_0^1 \left\| (\partial_j F_i(x+t(\bar{x}-x))) - (\partial_j F_i(x)) \right\|_2 \mathrm{d}t \, \|\bar{x}-x\|_2 \qquad (6.3)
\end{aligned}$$

Sei $\alpha \in (0,1)$ beliebig aber fest. Da $F : \mathbb{R}^n \to \mathbb{R}^n$ stetig differenzierbar ist, existiert ein $\varepsilon > 0$, so dass

$$\int_0^1 \left\| (\partial_j F_i(x+t(\bar{x}-x))) - (\partial_j F_i(x)) \right\|_2 \mathrm{d}t \leq \frac{\alpha}{2\left\| (\partial_j F_i(\bar{x}))^{-1} \right\|_2} \quad \forall x \in B_\varepsilon(\bar{x}).$$

Folglich

$$\|r(x,\bar{x})\|_2 = \frac{\alpha}{2\left\| (\partial_j F_i(\bar{x}))^{-1} \right\|_2} \|\bar{x}-x\|_2 \quad \forall x \in B_\varepsilon(\bar{x}). \qquad (6.4)$$

Wir können nun $\varepsilon > 0$ verkleinern, so dass gilt:

$$\|(\partial_j F_i(x))^{-1}\|_2 \leq 2\|(\partial_j F_i(\bar{x}))^{-1}\|_2 \quad \forall x \in B_\varepsilon(\bar{x}). \qquad (6.5)$$

Sei nun $x^k \in B_\varepsilon(\bar{x})$ und $x^{k+1} \in \mathbb{R}^n$ die durch das Newton-Verfahren erzeugte neue Iterierte. Dann gilt:

$$\begin{aligned}
x^{k+1} - \bar{x} &= x^{k+1} - x^k + x^k - \bar{x} \\
&= (\partial_j F_i(x^k))^{-1} (\partial_j F_i(x^k))(x^{k+1} - x^k) + x^k - \bar{x} \\
&= (\partial_j F_i(x^k))^{-1} \underbrace{F'(x^k)(x^{k+1} - x^k)}_{=-F(x^k)} + x^k - \bar{x} \\
&= -(\partial_j F_i(x^k))^{-1} F(x^k) + x^k - \bar{x} \\
&= (\partial_j F_i(x^k))^{-1} (-F(x^k) + (\partial_j F_i(x^k))(x^k - \bar{x})) \\
&\overset{F(\bar{x})=0}{=} (\partial_j F_i(x^k))^{-1} (F(\bar{x}) - F(x^k) + F'(x^k)(x^k - \bar{x})) \\
&= (\partial_j F_i(x^k))^{-1} (F(\bar{x}) - F(x^k) - F'(x^k)(\bar{x} - x^k)) \\
&= (\partial_j F_i(x^k))^{-1} r(x^k, \bar{x}).
\end{aligned}$$

Aus (6.4) und (6.5) folgt

$$\begin{aligned}\|x^{k+1}-\bar{x}\|_2 &= \|(\partial_j F_i(x^k))^{-1} r(x^k,\bar{x})\|_2 \\ &\leq \|(\partial_j F_i(x^k))^{-1}\|_2 \|r(x^k,\bar{x})\|_2 \\ &\overset{(6.4)}{\leq} \|(\partial_j F_i(x^k))^{-1}\|_2 \frac{\alpha}{2\|(\partial_j F_i(\bar{x}))^{-1}\|_2} \|\bar{x}-x\|_2 \\ &\overset{(6.5)}{\leq} \alpha \|x^k-\bar{x}\|_2.\end{aligned}$$

Ist $x^0 \in B_\varepsilon(\bar{x})$, so folgt aus der obigen Abschätzung:

$$x^k \in B_\varepsilon(\bar{x}) \quad \forall k \in \mathbb{N} \quad (\text{denn } \|x^k-\bar{x}\|_2 \leq \alpha^k \|x^0-\bar{x}\|_2 < \varepsilon)$$

und

$$\|x^{k+1}-\bar{x}\|_2 \leq \alpha \|x^k-\bar{x}\|_2 \quad \forall k \in \mathbb{N}.$$

Mit anderen Worten: Ist $x^0 \in B_\varepsilon(\bar{x})$, so konvergiert die durch das Newton-Verfahren erzeugte Folge $\{x^k\}_{k=1}^\infty$ linear (mit Rate α) gegen $\bar{x}$. Nun zeigen wir die superlineare Konvergenz:

$$\begin{aligned}\|x^{k+1}-\bar{x}\|_2 &= \|(\partial_j F_i(x^k))^{-1} r(x^k,\bar{x})\|_2 \\ &\overset{(6.5)}{\leq} 2\|(\partial_j F_i(x^k))^{-1}\|_2 \|r(x^k,\bar{x})\|_2 \\ &\overset{(6.3)}{\leq} 2\|(\partial_j F_i(x^k))^{-1}\|_2 \int_0^1 \left\| (\partial_j F_i(x^k + t(\bar{x}-x^k))) - (\partial_j F_i(x^k)) \right\|_2 \mathrm{d}t \|x^k-\bar{x}\|_2\end{aligned}$$

mit

$$\int_0^1 \left\| (\partial_j F_i(x^k + t(\bar{x}-x^k))) - (\partial_j F_i(x^k)) \right\|_2 \mathrm{d}t \to 0 \text{ für } k \to \infty,$$

da $F : \mathbb{R}^n \to \mathbb{R}^n$ stetig differenzierbar ist und $\{x_k\}_{k=1}^\infty$ gegen $\bar{x}$ konvergiert. Insgesamt erhalten wir:

$$\frac{\|x^{k+1}-\bar{x}\|_2}{\|x^k-\bar{x}\|_2} \to 0 \text{ für } k \to \infty.$$

□

Satz 6.26 (Lokale quadratische Konvergenz des Newton-Verfahrens). *Es sei $F : \mathbb{R}^n \to \mathbb{R}^n$ stetig differenzierbar und $\bar{x} \in \mathbb{R}^n$ eine Nullstelle von F. Ferner sei die Jacobi-Matrix*

$$(\partial_j F_i(\bar{x})) \in \mathbb{R}^{n\times n}$$

regulär. Es existiere ein $L > 0$, so dass

$$\left\| (\partial_j F_i(x)) - (\partial_j F_i(y)) \right\|_2 \leq L \|x-y\|_2 \quad \forall x,y \in \mathbb{R}^n,$$

d.h. $F' : \mathbb{R}^n \to \mathscr{L}(\mathbb{R}^n,\mathbb{R}^n)$ ist Lipschitz-stetig. Dann existiert ein $\varepsilon > 0$, so dass für jeden Startwert $x^0 \in B_\varepsilon(\bar{x})$ das Newton-Verfahren quadratisch *gegen $\bar{x}$ konvergiert.*

Beweis. Im Beweis des vorherigen Satzes haben wir gezeigt, dass ein $\varepsilon > 0$ existiert, so dass für jeden Startwert $x^0 \in B_\varepsilon(\bar{x})$ gilt:

(i): $x^k \in B_\varepsilon(\bar{x}) \quad \forall k \in \mathbb{N}$.

(ii): $\{x^k\}_{k=1}^\infty$ konvergiert superlinear gegen $\bar{x}$.

Außerdem gilt

$$\begin{aligned}
\|x^{k+1} - \bar{x}\|_2 &\leq 2 \left\| (\partial_j F_i(x^k))^{-1} \right\|_2 \int_0^1 \left\| (\partial_j F_i(x^k + t(\bar{x} - x^k))) - (\partial_j F_i(x^k)) \right\|_2 \mathrm{d}t \|x^k - \bar{x}\|_2 \\
&\leq 2 \left\| (\partial_j F_i(x^k))^{-1} \right\|_2 \int_0^1 L \left\| x^k + t(\bar{x} - x^k) - x^k \right\|_2 \mathrm{d}t \|x^k - \bar{x}\|_2 \\
&\leq 2 \left\| (\partial_j F_i(x^k))^{-1} \right\|_2 L \int_0^1 t \, \mathrm{d}t \|x^k - \bar{x}\|_2^2 \\
&= c \|x^k - \bar{x}\|_2^2
\end{aligned}$$

mit $c := 2 \left\| (\partial_j F_i(x^k))^{-1} \right\|_2 L \frac{1}{2} = L \left\| (\partial_j F_i(x^k))^{-1} \right\|_2$. Folglich gilt

$$\|x^{k+1} - \bar{x}\|_2 \leq c \|x^k - \bar{x}\|_2^2 \quad \forall k \in \mathbb{N}.$$

□

Bemerkung 6.27. Die Aussage bleibt auch richtig, falls die Abbildung $F' : \mathbb{R}^n \to \mathscr{L}(\mathbb{R}^n, \mathbb{R}^n)$ nur lokal Lipschitz-stetig ist.

6.4 Das vereinfachte Newton-Verfahren

Es sei $F : \mathbb{R}^n \to \mathbb{R}^n$ stetig differenzierbar. Beim Newton-Verfahren zur Lösung von $F(x) = 0$ lösen wir in jeder Iteration das lineare Gleichungssystem

$$F'(x^k) d^k = -F(x^k).$$

Für jedes $k \in \mathbb{N}$ müssen wir also:

(i) Die Jacobi-Matrix von F in x^k berechnen.

(ii) Das lineare Gleichungssystem

$$(\partial_j F_i(x^k)) d^k = -F(x^k)$$

zum Beispiel mittels LR-Zerlegung lösen.

Die Berechnungen von (i) und (ii) müssen in jedem Schritt des Newton-Verfahrens durchgeführt werden. Das kann aber manchmal sehr teuer sein.
Beim vereinfachten Newton-Verfahren bestimmen wir die $d^k \in \mathbb{R}^n$ als Lösung von

$$F'(x^0)d^k = -F(x^k) \quad \text{(Newton-Verfahren: } F'(x^k)d^k = -F(x^k)\text{)}.$$

Diese Strategie hat den Vorteil, dass man im gesamten Algorithmus nur einmal die Jacobi-Matrix von F in x^0 und deren LR-Zerlegung bestimmen muss. Der Rechenaufwand beim vereinfachten Newton-Verfahren ist somit meist überaus geringer als beim Newton-Verfahren.

Algorithmus 6.2 Vereinfachtes Newton-Verfahren

(S0) Wähle einen Startwert $x^0 \in \mathbb{R}^n$ und setze $k = 0$.
(S1) **if** $F(x^K) = 0$ **then** STOP **end if**
(S2) Bestimme $d^k \in \mathbb{R}^n$ als Lösung von

$$F'(x^0)d^k = -F(x^k).$$

(S3) Setze $x^{k+1} = x^k + d^k$, $k = k + 1$ und gehe zu (S1).

Der Nachteil des Verfahrens besteht darin, dass dieses nur linear lokal konvergiert (also insbesondere nicht superlinear oder quadratisch).

Satz 6.28 (Lokale lineare Konvergenz des vereinfachten Newton-Verfahrens)**.** *Es sei $F : \mathbb{R}^n \to \mathbb{R}^n$ stetig differenzierbar und $\bar{x} \in \mathbb{R}^n$ eine Nullstelle von F. Ferner sei die Jacobi-Matrix von F in $\bar{x}$*

$$(\partial_j F_i(\bar{x})) \in \mathbb{R}^{n\times n}$$

regulär. Dann gibt es ein $\varepsilon > 0$, so dass für jeden Startwert $x^0 \in B_\varepsilon(\bar{x})$ das vereinfachte Newton-Verfahren linear *gegen $\bar{x}$ konvergiert.*

Beweis. Da die Jacobi-Matrix $(\partial_j F_i(\bar{x})) \in \mathbb{R}^{n\times n}$ regulär ist, gibt es ein $\varepsilon_1 > 0$, so dass gilt

$$\|(\partial_j F_i(x))^{-1}\|_2 \le c \quad \forall x \in B_{\varepsilon_1}(\bar{x}) \tag{6.6}$$

mit $c := 2\left\|(\partial_j F_i(\bar{x}))^{-1}\right\|_2$ (siehe Lemma). Weiter existiert ein $\varepsilon_2 > 0$, so dass gilt

$$\|r(x,\bar{x})\|_2 \le \frac{1}{4}c\,\|x-\bar{x}\|_2 \quad \forall x \in B_{\varepsilon_2}(\bar{x}), \tag{6.7}$$

siehe Beweis der lokalen superlinearen Konvergenz des Newton-Verfahrens (Integraldarstellung von r). Außerdem gibt es ein $\varepsilon_3 > 0$, so dass

$$\left\|(\partial_j F_i(x)) - (\partial_j F_i(y))\right\|_2 \le \frac{1}{4c} \quad \forall x, y \in B_{\varepsilon_3}(\bar{x}), \tag{6.8}$$

da $F : \mathbb{R}^n \to \mathbb{R}^n$ stetig differenzierbar ist. Setze nun $\varepsilon = \min\{\varepsilon_1, \varepsilon_2, \varepsilon_3\}$ und wähle $x^0 \in B_\varepsilon(\bar{x})$. Dann gilt für jedes $x^k \in B_\varepsilon(\bar{x})$ und die durch das vereinfachte Newton-Verfahren erzeugte neue Iterierte $x^{k+1} \in \mathbb{R}^n$:

$$\begin{aligned}
\|x^{k+1} - \bar{x}\|_2 &= \|x^{k+1} - x^k + x^k - \bar{x}\|_2 \\
&= \|F'(x^0)^{-1}F'(x^0)(x^{k+1} - x^k) + x^k - \bar{x}\|_2 \\
&= \|-F'(x^0)^{-1}F(x^k) + x^k - \bar{x}\|_2 \\
&= \|(\partial_j F_i(x^0))^{-1}(-F(x^k) + (\partial_j F_i(x^0))(x^k - \bar{x}))\|_2 \\
&\leq \|(\partial_j F_i(x^0))^{-1}\|_2 \|-F(x^k) + F'(x^0)(x^k - \bar{x})\|_2 \\
&\overset{(6.6)}{\leq} c\|-F(x^k) + F'(x^0)(x^k - \bar{x})\|_2 \\
&= c\|-F(x^k) + F(\bar{x}) - F'(x^0)(\bar{x} - x^k)\|_2 \\
&\leq c\|F(\bar{x}) - F(x^k) - F'(x^k)(\bar{x} - x^k)\|_2 + c\|F'(x^k)(\bar{x} - x^k) - F'(x^0)(\bar{x} - x^k)\|_2 \\
&\overset{(6.7)}{\leq} \frac{1}{4}\|\bar{x} - x^k\|_2 + c\|\left[(\partial_j F_i(x^k)) - (\partial_j F_i(x^0))\right](\bar{x} - x^k)\|_2 \\
&\leq \frac{1}{4}\|\bar{x} - x^k\|_2 + c\|(\partial_j F_i(x^k)) - (\partial_j F_i(x^0))\|_2 \|(\bar{x} - x^k)\|_2 \\
&\overset{(6.8)}{\leq} \frac{1}{4}\|\bar{x} - x^k\|_2 + \frac{1}{4}\|\bar{x} - x^k\|_2 \\
&= \frac{1}{2}\|\bar{x} - x^k\|_2.
\end{aligned}$$

Folglich gilt für jeden Startwert $x^0 \in B_\varepsilon(\bar{x})$:

$$\begin{cases} x^k \in B_\varepsilon(\bar{x}) \quad \forall k \in \mathbb{N} \quad (\text{denn } \left\|x^k - \bar{x}\right\|_2 \leq (\frac{1}{2})^k \left\|x^0 - \bar{x}\right\|_2 < \varepsilon), \\ \left\|x^{k+1} - \bar{x}\right\|_2 \leq \frac{1}{2}\left\|x^k - \bar{x}\right\|_2 \quad \forall k \in \mathbb{N}. \end{cases}$$

Daraus ergibt sich die Behauptung. □

Kapitel 7

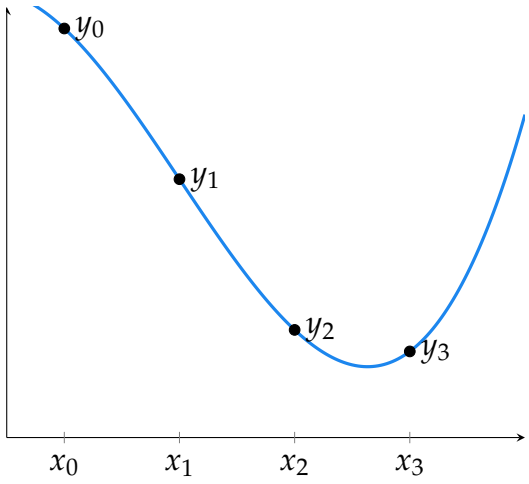
Interpolation

Die Interpolationsaufgabe besteht darin, zu gegebenen Punktepaaren $(x_k, y_k) \in \mathbb{R}^2$, $k = 0, \ldots, n$ eine Funktion $p : \mathbb{R} \to \mathbb{R}$ zu finden, so dass

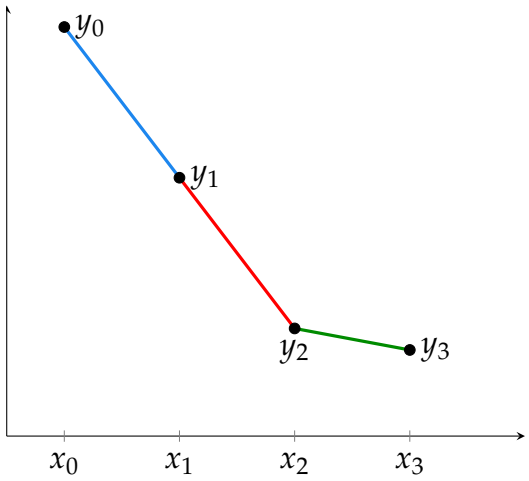
$$p(x_k) = y_k \quad \forall k = 0, \ldots, n$$

gilt. Mögliche Ansätze für $p : \mathbb{R} \to \mathbb{R}$ sind:

(i) Polynome $p(x) = a_0 + a_1 x + \ldots + a_n x^n$, also z.B.



(ii) Splines (stückweise Polynome), also z.B.



(iii) Trigonometrische Polynome.

7.1 Polynominterpolation

Definition 7.1. Mit Π_n bezeichnen wir die Menge aller reeller Polynome vom Grad $\leq n$, das heißt

$$\Pi_n := \{p : \mathbb{R} \to \mathbb{R} \mid \exists a_0, \ldots, a_n \in \mathbb{R} : p(x) = a_0 + a_1 x + \ldots + a_n x^n \quad \forall x \in \mathbb{R}\}.$$

Satz 7.2. *Es sei $n \in \mathbb{N}$ und $(x_k, y_k) \in \mathbb{R}^2$, $k = 0, \ldots, n$ mit $x_i \neq x_j \quad \forall i \neq j$. Dann gibt es genau ein Polynom $p \in \Pi_n$ mit*

$$p(x_k) = y_k \quad \forall k = 0, \ldots, n.$$

Beweis. Wir machen den Ansatz

$$p(x) = a_0 + a_1 x + \ldots + a_n x^n.$$

Gesucht sind $a_0, \ldots, a_n \in \mathbb{R}$, so dass gilt

$$p(x_k) = y_k \quad \forall k = 0, \ldots, n.$$

Dies ist äquivalent zu:

$$\begin{aligned} a_0 + a_1 x_0 + a_2 x_0^2 + \ldots + a_n x_0^n &= y_0 \\ a_0 + a_1 x_1 + a_2 x_1^2 + \ldots + a_n x_1^n &= y_1 \\ &\vdots \\ a_0 + a_1 x_n + a_2 x_n^2 + \ldots + a_n x_n^n &= y_n. \end{aligned}$$

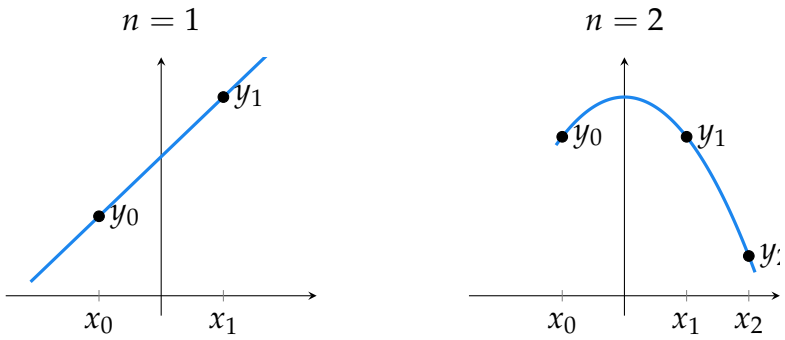
Dieses Gleichungssystem entspricht

$$\begin{pmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix}.$$

Da die x_i paarweise verschieden sind, ist die obige Matrix regulär. Somit gibt es genau $a_0, \ldots, a_n \in \mathbb{R}$ mit

$$p(x_k) = y_k \quad \forall k = 0, \ldots, n.$$

□



Definition 7.3 (Interpolationspolynom)**.** Das nach dem Satz eindeutig bestimmte Polynom $p \in \Pi_n$ mit der Eigenschaft

$$p(x_k) = y_k \quad \forall k = 0, \ldots, n$$

für die vorgegebenen Punktepaare

$$(x_k, y_k) \in \mathbb{R}^2, \quad k = 0, \ldots, n \text{ mit } x_i \neq x_j \quad \forall i \neq j$$

heißt *Interpolationspolynom*.

Eine kanonische Basis für Π_n ist gegeben durch

$$\{1, x, x^2, \ldots, x^n\} \quad \text{(Monombasis)}.$$

Wir wollen nun weitere Basis-Ansätze für Π_n untersuchen.

7.1.0.1 Lagrangesche Interpolation

Im Folgenden seien Punktepaare

$$(x_k, y_k) \in \mathbb{R}^2, \quad k = 0, \ldots, n$$

mit paarweise verschiedenen x_k gegeben. Die Interpolationsaufgabe lautet:

$$\text{Finde } p \in \Pi_n, \text{ so dass } p(x_k) = y_k \quad \forall k = 0, \ldots, n.$$

Definition 7.4 (Lagrange-Basispolynome)**.** Die Polynome

$$L_k(x) = \frac{(x-x_0)\cdot\ldots\cdot(x-x_{k-1})(x-x_{k+1})\cdot\ldots\cdot(x-x_n)}{(x_k-x_0)\cdot\ldots\cdot(x_k-x_{k-1})(x_k-x_{k+1})\cdot\ldots\cdot(x_k-x_n)}$$

heißen *Lagrange-Basispolynome*.

Beispiel 7.5 (n=2)**.**

$$\begin{aligned} L_0(x) &= \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}, \\ L_1(x) &= \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}, \\ L_2(x) &= \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}. \end{aligned}$$

und

$$\begin{aligned} &L_0(x_0) = 1, \quad L_0(x_1) = 0, \quad L_0(x_2) = 0, \\ &L_1(x_0) = 0, \quad L_1(x_1) = 1, \quad L_1(x_2) = 0, \\ &L_2(x_0) = 0, \quad L_2(x_1) = 0, \quad L_2(x_2) = 1. \end{aligned}$$

Lemma 7.6. *Es gilt*

$$L_k(x_j) = \delta_{kj} \quad \forall k, j = 0, \ldots, n.$$

Beweis. Klar nach Definition der Lagrange-Basispolynome. □

Satz 7.7 (Lagrangesche Interpolationsformel)**.** *Gegeben seien Punktepaare $(x_k, y_k) \in \mathbb{R}^2$, $k = 0, \ldots, n$ mit paarweise verschiedenen x_k. Dann ist das Lagrangesche Interpolationspolynom*

$$\bar{p}_L(x) := \sum_{k=0}^{n} y_k L_k(x)$$

die eindeutige Lösung der Interpolationsaufgabe:

$$\textit{Finde } p \in \Pi_n, \textit{ so dass } p(x_k) = y_k \quad \forall k = 0, \ldots, n.$$

Beweis. Da $L_j(x_i) = \delta_{ij} \quad \forall i, j = 0, \ldots, n$ gilt, so ist

$$\bar{p}_L(x_k) = \sum_{j=0}^{n} y_j L_j(x_k) = \sum_{j=0}^{n} y_j \delta_{jk} = y_k$$

für alle $k = 0, \ldots, n$. □

7.1.0.2 Newtonsche Interpolation

Definition 7.8 (Newton-Basispolynome)**.** Die Polynome

$$N_k(x) := (x - x_0) \cdot \ldots \cdot (x - x_{k-1}) \quad \forall k = 1, \ldots, n$$
$$N_0(x) \equiv 1$$

heißen *Newton-Basispolynome*.

Wir wollen nun die Interpolationsaufgabe mittels Newton-Basispolynomen lösen. Gesucht sind

$$a_0, \ldots, a_n \in \mathbb{R},$$

so dass gilt:

$$y_k = \sum_{j=0}^{n} a_j N_j(x_k) \quad \forall k = 0, \ldots, n.$$

Lemma 7.9 (Aitken)**.** *Es seien Punktepaare $(x_k, y_k) \in \mathbb{R}^2$ mit paarweise verschiedenen x_k. Ferner seien $P_{[0]}, P_{[n]} \in \Pi_{n-1}$ gegeben mit*

$$P_{[0]}(x_k) = y_k \quad \forall k = 0, \ldots, n-1,$$
$$P_{[n]}(x_k) = y_k \quad \forall k = 1, \ldots, n.$$

Dann löst das Polynom

$$p(x) := \frac{P_{[0]}(x)(x - x_n) - P_{[n]}(x)(x - x_0)}{x_0 - x_n}$$

die Interpolationsaufgabe

$$\textit{Finde } p \in \Pi_n \textit{ mit } p(x_k) = y_k \quad \forall k = 0, \ldots, n.$$

Beweis. Laut Definition ist

$$p(x) = \frac{P_{[0]}(x)(x - x_n) - P_{[n]}(x)(x - x_0)}{x_0 - x_n}$$

ein Polynom vom Grad $\leq n$, denn

$$P_{[0]}, P_{[n]} \in \Pi_n.$$

Für $k \in \{1, \ldots, n-1\}$ gilt

$$p(x_k) = \frac{P_{[0]}(x_k)(x - x_n) - P_{[n]}(x_k)(x - x_0)}{x_0 - x_n} = \frac{y_k(x - x_n) - y_k(x - x_0)}{x_0 - x_n}$$
$$= \frac{y_k(x_0 - x_n)}{x_0 - x_n}$$
$$= y_k.$$

Für $k \in \{0, n\}$ gilt

$$p(x_0) = \frac{P_{[0]}(x_0)(x - x_n) - P_{[n]}(x_0)(x - x_0)}{x_0 - x_n} = \frac{y_0(x_0 - x_n)}{x_0 - x_n} = y_0,$$
$$p(x_n) = \frac{P_{[0]}(x_n)(x - x_n) - P_{[n]}(x_n)(x - x_0)}{x_0 - x_n} = \frac{y_n(x_0 - x_n)}{x_0 - x_n} = y_n,$$

und somit folgt die Behauptung. □

Wir setzen

$$P_{00}(x) \equiv y_0,$$
$$P_{11}(x) \equiv y_1,$$
$$\vdots$$
$$P_{nn}(x) \equiv y_n.$$

Wir definieren $P_ij \in \Pi_{j-i}$, $0 \leq i < j \leq n$ rekursiv wie folgt:

$$P_{ij}(x) := \frac{P_{i,j-1}(x)(x - x_j) - P_{i+1,j}(x)(x - x_i)}{x_i - x_j}, \quad 0 \leq i < j \leq n.$$

Aus dem Lemma von Aitken wird die Interpolationsaufgabe

$$\text{Finde } p \in \Pi_n \text{ mit } p(x_k) = y_k \quad \forall k = 0, \ldots, n$$

eindeutig durch P_{0n} gelöst. Das Interpolationspolynom an der Stelle x lässt sich durch das sogenannte *Neville-Aitken-Schema* berechnen:

$$\begin{array}{c|ccccccc}
x_0 & y_0 = P_{00}(x) & & & & & & \\
 & & \searrow & & & & & \\
x_1 & y_1 = P_{11}(x) & \to & P_{01}(x) & \searrow & & & \\
 & & \searrow & & & & & \\
x_2 & y_2 = P_{22}(x) & \to & P_{12}(x) & \to & P_{02}(x) & & \\
\vdots & \vdots & & \vdots & & & \ddots & \\
x_n & y_n = P_{nn}(x) & \to & P_{n-1,n}(x) & \cdots & \cdots & \cdots & P_{0n}(x)
\end{array}$$

Satz 7.10 (Newtonsche Interpolationsformel)**.** *Gegeben seien Punktepaare*

$$(x_k, y_k) \in \mathbb{R}^2, \quad k = 0, \ldots, n$$

mit paarweise verschiedenen x_k. Dann ist das Newtonsche Interpolationspolynom

$$\bar{p}_N(x) := \sum_{j=0}^{n} [x_0, \ldots, x_n] N_j(x),$$

wobei

$$[x_0, \ldots, x_n] \in \mathbb{R}$$

rekursiv definiert sind durch

$$[x_j] := y_j \quad \forall j = 0, \ldots, n$$
$$[x_k, \ldots, x_j] := \frac{[x_{k+1}, \ldots, x_j] - [x_k, \ldots, x_{j-1}]}{x_j - x_k} \quad \forall j > k \geq 0,$$

die eindeutige Lösung der Interpolationsaufgabe.

Beweis. Wir haben oben bereits gezeigt, dass $P_{0n} \in \Pi_n$ die eindeutige Lösung der Interpolationsaufgabe ist, und P_{0n} ist definiert durch

$$\begin{aligned} P_{ij}(x) &= \frac{P_{i,j-1}(x)(x - x_j) - P_{i+1,j}(x)(x - x_i)}{x_i - x_j} \\ &= P_{i+1,j}(x) + \left(P_{i+1,j}(x) - P_{i,j-1}(x)\right) \frac{x - x_j}{x_j - x_i}, \quad 0 \leq i < j \leq n \end{aligned}$$

und

$$P_{kk}(x) \equiv y_k \quad \forall k = 0, \ldots, n.$$

Durch Koeffizientenvergleich folgt

$$P_{0n}(x) = \bar{p}_N(x).$$

□

7.1.0.3 Interpolationsfehler

Definition 7.11. Es seien $a, b \in \mathbb{R}$ mit $a < b$ und $n \in \mathbb{N}$. Wir definieren

$$C[a,b] := \{f : [a,b] \to \mathbb{R} \text{ stetig}\}$$
$$C^n[a,b] := \left\{f : (a,b) \to \mathbb{R} \text{ n-mal differenzierbar} \mid f^{(k)} \in C[a,b] \quad \forall k = 0, \ldots, n\right\},$$

wobei $f^{(k)}$ die k-te Ableitung von f bezeichnet, und $f^{(0)} = f$.

Lemma 7.12 (Rolle)**.** *Es seien $a, b \in \mathbb{R}$ mit $a < b$ und $f \in C^1[a,b]$. Ferner seien $x_1, x_2 \in [a,b]$ mit $x_1 \neq x_2$ und $f(x_1) = f(x_2)$. Dann gibt es ein $\xi \in (x_1, x_2)$ mit $f'(\xi) = 0$.*

Satz 7.13 (Interpolationsfehler)**.** *Seien $a, b \in \mathbb{R}$ mit $a < b$ und*

$$a = x_0 < x_1 < \ldots < x_n = b.$$

Ferner seien $f \in C^{n+1}[a,b]$, $n \in \mathbb{N}$ und $p \in \Pi_n$ mit

$$p(x_k) = f(x_k) \quad \forall k = 0, \ldots, n.$$

Dann gibt es zu jedem $\tilde{x} \in [a,b]$ ein $\xi \in (a,b)$ mit

$$f(\tilde{x}) - p(\tilde{x}) = \frac{w(\tilde{x})}{(n+1)!} f^{(n+1)}(\xi)$$

mit $w(\tilde{x}) = (\tilde{x} - x_0) \cdot \ldots \cdot (\tilde{x} - x_n)$.

Beweis. Die Aussage ist bereits richtig für $\tilde{x} \in \{x_0, \ldots, x_n\}$, denn

$$f(x_k) = p(x_k) \quad \forall k = 0, \ldots, n$$

und

$$w(x_k) = 0 \quad \forall k = 0, \ldots, n.$$

Sei nun $\tilde{x} \in [a,b] \setminus \{x_0, \ldots, x_n\}$. Dann gilt

$$w(\tilde{x}) \overset{\text{Def.}}{=} (\tilde{x} - x_0) \cdot \ldots \cdot (\tilde{x} - x_n) \neq 0$$

und wir setzen

$$\alpha := \frac{f(\tilde{x}) - p(\tilde{x})}{w(\tilde{x})} \in \mathbb{R}.$$

Dann ist

$$f(\tilde{x}) - p(\tilde{x}) - \alpha w(\tilde{x}) = 0. \tag{7.1}$$

Nun sei $F : [a,b] \to \mathbb{R}, \quad F(x) := f(x) - p(x) - \alpha w(x)$ mit $w(x) = (x - x_0) \cdot \ldots \cdot (x - x_n)$. Dann ist $F \in C^{n+1}[a,b]$, da $f \in C^{n+1}[a,b]$, $p \in \Pi_n$ und $w \in \Pi_{n+1}$. Außerdem hat F gerade $n+2$ Nullstellen

$$\tilde{x}, x_0, x_1, \ldots, x_n.$$

Nach dem Lemma von Rolle hat die Ableitung F' also $n+1$ Nullstellen, $F^{(2)}$ n Nullstelle, $\ldots$, $F^{(n+1)}$ hat eine Nullstelle $\xi \in (a,b)$. Also

$$0 = F^{(n+1)}(\xi) = f^{(n+1)}(\xi) - p^{(n+1)}(\xi) - \alpha w^{(n+1)}(\xi) = f^{(n+1)}(\xi) - 0 - \alpha(n+1)!.$$

Daraus ergibt sich

$$\alpha = \frac{1}{(n+1)!} f^{(n+1)}(\xi).$$

Einsetzen in (7.1) liefert

$$f(\tilde{x}) - p(\tilde{x}) = \frac{w(\tilde{x})}{(n+1)!} f^{(n+1)}(\tilde{x}).$$

□

7.2 Hermite-Interpolation

Im Folgenden sei $f \in C^1[a,b]$ mit $-\infty < a < b < \infty$. Ferner seien $x_0, \ldots, x_n \in [a,b]$ paarweise verschieden. Die Hermitesche Interpolationsaufgabe lautet:
Finde $p \in \Pi_{2n+1}$ mit

$$\begin{aligned} p(x_k) &= f(x_k) \quad \forall k = 0, \ldots, n, \\ p'(x_k) &= f'(x_k) \quad \forall k = 0, \ldots, n. \end{aligned}$$

Wir suchen Basis-Funktionen

$$\Phi_k, \Psi_k \in \Pi_{2n+1} \quad \forall k = 0, \ldots, n$$

mit den Eigenschaften

$$\begin{aligned} \Phi_k(x_j) &= \delta_{kj} \quad \forall k, j = 0, \ldots, n, \\ \Phi_k'(x_j) &= 0 \quad \forall k, j = 0, \ldots, n, \\ \Psi_k(x_j) &= 0 \quad \forall k, j = 0, \ldots, n, \\ \Psi_k(x_j) &= \delta_{kj} \quad \forall k, j = 0, \ldots, n. \end{aligned}$$

Wir wissen bereits, dass die Lagrange-Basisfunktion $L_k \in \Pi_n$ mit

$$L_k(x) = \frac{(x - x_0) \cdot \ldots \cdot (x - x_{k-1})(x - x_{k+1}) \cdot \ldots \cdot (x - x_n)}{(x_k - x_0) \cdot \ldots \cdot (x_k - x_{k-1})(x_k - x_{k+1}) \cdot \ldots \cdot (x_k - x_n)}$$

die Eigenschaft

$$L_k(x_j) = \delta_{kj} \quad \forall k, j = 0, \ldots, n$$

erfüllt. Nun definieren wir

$$\begin{aligned} \Phi_k(x) &:= (1 - 2L_k'(x_k)(x - x_k))L_k^2(x) \\ \Psi_k(x) &:= (x - x_k)L_k^2(x) \end{aligned}$$

für alle $k = 0, \ldots, n$. Nun gilt

(i) $\Phi_k(x_j) = (1 - 2L_k'(x_k)(x_j - x_k)) \underbrace{L_k^2(x_j)}_{=\delta_{kj}} = \delta_{kj} \quad \forall k, j = 0, \ldots, n.$

(ii)
$$\begin{aligned} \Phi_k'(x_j) &= -2L_k'(x_k)L_k^2(x_j) + (1 - 2L_k'(x_k)(x_j - x_k))2L_k(x_j)L_k'(x_j) \\ &= -2L_k'(x_k)\delta_{kj} + (1 - 2L_k'(x_k)(x_j - x_k))2\delta_{kj}L_k'(x_j) \\ &= -2L_k'(x_k)\delta_{kj} + 2\delta_{kj}L_k'(x_j) \\ &= 0 \quad \forall k, j = 0, \ldots, n. \end{aligned}$$

(iii) $\Psi_k(x_j) = (x_j - x_k)L_k^2(x_j) = (x_j - x_k)\delta_{kj} = 0 \quad \forall k, j = 0, \ldots, n.$

(iv) $\Psi_k'(x_j) = \underbrace{L_k^2(x_j)}_{=\delta_{kj}} + (x_j - x_k)2\underbrace{= L_k(x_j)}_{\delta_{kj}} L_k'(x_j)$
$= \delta_{kj} + 2(x_j - x_k)\delta_{kj} L_k'(x_j)$
$= \delta_{kj} \quad \forall k, j = 0, \ldots, n.$

Die Lösung der Hermiteschen Interpolationsaufgabe ist dann gegeben durch

$$\bar{p}_H(x) := \sum_{j=0}^{n} f(x_j)\Phi_j(x) + \sum_{j=0}^{n} f'(x_j)\Psi_j(x),$$

denn laut Konstruktion sind

$$\Phi_j, \Psi_j \in \Pi_{2n+1} \quad \forall j = 0, \ldots, n$$

und

$$\bar{p}_H(x_k) = \sum_{j=0}^{n} f(x_j) \underbrace{\Phi_j(x_k)}_{=\delta_{kj}} + \sum_{j=0}^{n} f'(x_j) \underbrace{\Psi_j(x_k)}_{=0} = f(x_k),$$

$$\bar{p}_H'(x_k) = \sum_{j=0}^{n} f(x_j) \underbrace{\Phi_j'(x_k)}_{=0} + \sum_{j=0}^{n} f'(x_j) \underbrace{\Psi_j'(x_k)}_{=\delta_{kj}} = f'(x_k).$$

7.3 Splines

Der Nachteil der Polynominterpolation besteht darin, dass man bei der Verfeinerung der Zerlegung keine gleichmäßige Konvergenz erwarten kann. Deshalb untersuchen wir nun ein anderes Verfahren mit einer besseren Konvergenzeigenschaft. Im Folgenden betrachten wir die Zerlegung

$$\Delta = \{a = x_0 < x_1 < \ldots < x_n = b\}.$$

Definition 7.14. Seien $p, q \in \mathbb{N} \cup \{0\}$ mit der Eigenschaft $0 \leq q < p$. Wir definieren den Raum

$$S(\Delta, p, q) := \left\{ s \in C^q[a, b] \mid s|_{[x_{k-1}, x_k]} \in \Pi_p \quad \forall k = 1, \ldots, n \right\}.$$

Eine Funktion $s \in S(\Delta, p, q)$ wird als *Spline* vom Grad p der Diffenzierbarkeitsklasse q zur Zerlegung Δ bezeichnet.

Die Spline-Interpolationsaufgabe lautet wie folgt:
Zu gegebener Funktion $f : [a, b] \to \mathbb{R}$ suchen wir ein $s \in S(\Delta, p, q)$ mit

$$s(x_k) = f(x_k) \quad \forall k = 0, \ldots, n.$$

Nachdem wir diese Aufgabe gelöst haben, wollen wir den Fehler

$$\|f - s\|_{C[a,b]}$$

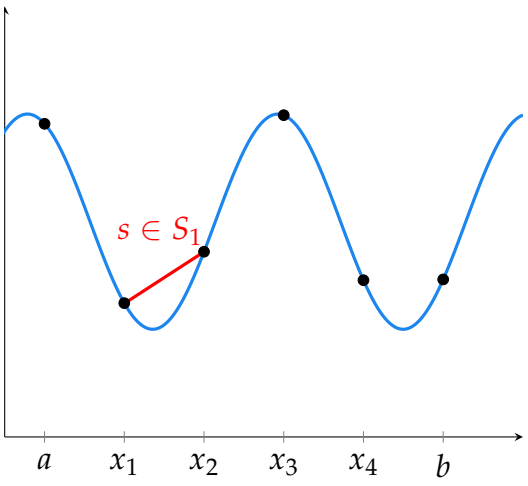
untersuchen.

Definition 7.15 (Stückweise lineare Splines)**.** Mit

$$S_1 := S(\Delta, 1, 0) = \left\{ s \in C[a,b] \mid s|_{[x_{k-1},x_k]} \in \Pi_1 \quad \forall k = 1, \dots, n \right\}$$

bezeichnen wir den Raum aller *stückweise linearer Splines*.

Beispiel 7.16.



Bemerkung 7.17. Ist $s \in S_1$, so gilt für jedes $k \in \{1, \dots, n\}$

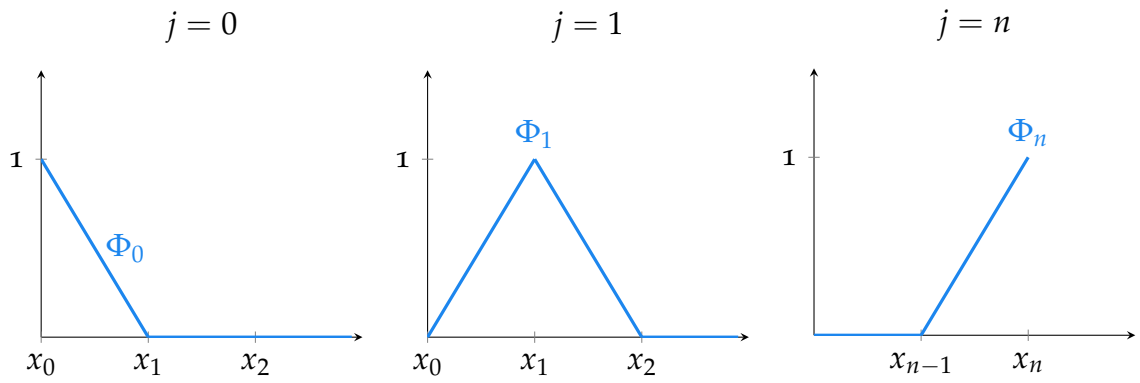
$$s|_{[x_{k-1},x_k]} \in \Pi_1,$$

also

$$\exists a_k, b_k \in \mathbb{R} : s(x) = a_k + b_k x \quad \forall x \in [x_{k-1}, x_k].$$

Definition 7.18 (Hutsche Basisfunktion)**.** Wir definieren die *Hutsche Basisfunktion* $\Phi_j \in S_1$ für $j = 0, \dots, n$ wie folgt:

$$\Phi_j(x) := \begin{cases} \frac{x - x_{j-1}}{x_j - x_{j-1}} & \text{falls } x \in [x_{j-1}, x_j] \text{ und } j > 0 \\ \frac{x_{j+1} - x}{x_{j+1} - x_j} & \text{falls } x \in [x_j, x_{j+1}] \text{ und } j < n \\ 0 & \text{sonst} \end{cases}.$$



Offenbar gilt für die Hutsche Basisfunktion

$$\Phi_j(x_i) = \delta_{ji} \quad \forall i, j = 0, \ldots, n.$$

Definition 7.19. Die Abbildung

$$I_1 : C[a,b] \to S_1, \quad (I_1 f)(x) := \sum_{j=0}^{n} f(x_j)\Phi_j(x)$$

heißt S_1-*Interpolationsoperator*.

Sei $f \in C[a,b]$. Dann gilt per Definition

$$I_1 f \in S_1$$

und

$$(I_1 f)(x_k) = \sum_{j=0}^{n} f(x_j) \underbrace{\Phi_j(x_k)}_{=\delta_{jk}} = f(x_k) \quad \forall k = 0, \ldots, n.$$

Somit löst $I_1 f \in S_1$ die Spline-Interpolationsaufgabe auf S_1.

Satz 7.20 (Fehlerabschätzung für den S_1-Interpolationsoperator). *Es sei $f \in C^2[a,b]$ und $h_{max} := \max_{1 \le k \le n} |x_{k-1} - x_k|$. Dann gilt*

$$\|f - I_1 f\|_{C[a,b]} \le \frac{1}{8} h_{max}^2 \left\| f^{(2)} \right\|_{C[a,b]},$$

wobei

$$\|\cdot\|_{C[a,b]} : C[a,b] \to \mathbb{R}, \quad \|v\|_{C[a,b]} = \max_{t \in [a,b]} |v(t)|.$$

Beweis. Nach Definition gilt

$$(I_1 f)(x_k) = f(x_k) \quad \forall k = 0, \ldots, n$$

und für jedes $k \in \{1, \dots, n\}$ ist $I_1 f|_{[x_{k-1}, x_k]} \in \Pi_1$. Sei nun $k \in \{1, \dots, n\}$ beliebig aber fest. Das Polynom $I_1 f|_{[x_{k-1}, x_k]} \in \Pi_1$ löst die Aufgabe:
Finde $p \in \Pi_1$ mit

$$\begin{aligned} p(x_{k-1}) &= f(x_{k-1}) \\ p(x_k) &= f(x_k). \end{aligned}$$

Nach unserem Satz über den Interpolationsfehler (Kapitel 7.1.3) gilt:
Zu jedem $\tilde{x} \in [x_{k-1}, x_k]$ gibt es ein $\xi \in (x_{k-1}, x_k)$ mit

$$f(\tilde{x}) - (I_1 f)(\tilde{x}) = \frac{w(\tilde{x})}{2!} f^{(2)}(\xi)$$

mit $w(\tilde{x}) = (\tilde{x} - x_{k-1})(\tilde{x} - x_k)$. Es folgt

$$\begin{aligned} |f(\tilde{x}) - (I_1 f)(\tilde{x})| &= \left| \frac{(\tilde{x} - x_{k-1})(\tilde{x} - x_k)}{2} \right| |f^{(2)}(\xi)| \\ &\overset{\tilde{x} \in [x_{k-1}, x_k]}{\leq} \frac{|x_k - x_{k-1}|^2}{8} |f^{(2)}(\xi)| \\ &\leq \frac{1}{8} h_{max}^2 |f^{(2)}(\xi)| \\ &\leq \frac{1}{8} h_{max}^2 \max_{x \in [a,b]} |f^{(2)}(x)| \\ &= \frac{1}{8} h_{max}^2 \left\| f^{(2)}(x) \right\|_{C[a,b]}. \end{aligned}$$

Daraus folgt

$$\max_{\tilde{x} \in [x_{k-1}, x_k]} |f(\tilde{x}) - (I_1 f)(\tilde{x})| \leq \frac{1}{8} h_{max}^2 \left\| f^{(2)}(x) \right\|_{C[a,b]}.$$

Da $k \in \{1, \dots, n\}$ beliebig aber fest gewählt wurde, folgt daraus

$$\max_{\tilde{x} \in [a,b]} |f(\tilde{x}) - (I_1 f)(\tilde{x})| \leq \frac{1}{8} h_{max}^2 \left\| f^{(2)}(x) \right\|_{C[a,b]}.$$

Mit anderen Worten:

$$\|f - I_1 f\|_{C[a,b]} \leq \frac{1}{8} h_{max}^2 \left\| f^{(2)}(x) \right\|_{C[a,b]}.$$

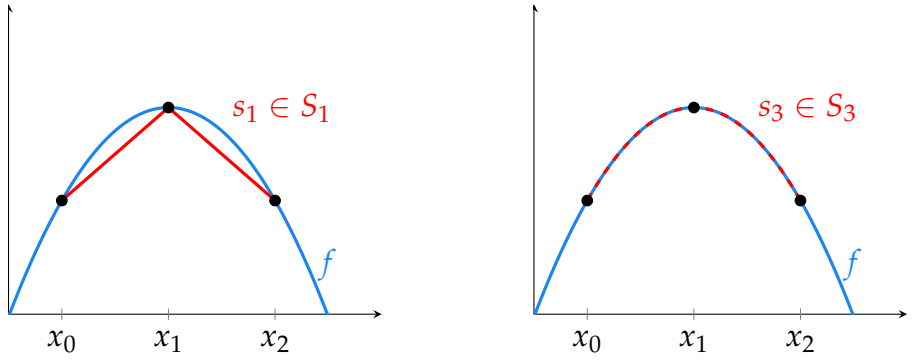
□

Definition 7.21 (Kubische Splines)**.** Mit

$$S_3 := S(\Delta, 3, 1) = \left\{ s \in C^1[a,b] \mid s|_{[x_{k-1}, x_k]} \in \Pi_3 \quad \forall k = 1, \dots, n \right\}$$

bezeichnen wir den Raum der *kubischen Hermite-Splines*.

Beispiel 7.22.



Definition 7.23 (Basisfunktion für S_3)**.** Wir definieren $\eta_j, \Psi_j \in S_3$, $j = 0, \dots, n$ wie folgt

$$\eta_j(x) := \begin{cases} \frac{(x-x_{j-1})^2(3x_j-x_{j-1}-2x)}{(x_j-x_{j-1})^3} & \text{falls } x \in [x_{j-1}, x_j] \text{ und } j > 0 \\ \frac{(x_{j+1}-x)^2(x_{j+1}-3x_j+2x)}{(x_{j+1}-x_j)^3} & \text{falls } x \in [x_j, x_{j+1}] \text{ und } j < n \cdot \\ 0 & \text{sonst} \end{cases}$$

$$\Psi_j(x) := \begin{cases} \frac{(x-x_{j-1})^2(x-x_j)}{(x_j-x_{j-1})^2} & \text{falls } x \in [x_{j-1}, x_j] \text{ und } j > 0 \\ \frac{(x_{j+1}-x)^2(x-x_j)}{(x_{j+1}-x_j)^2} & \text{falls } x \in [x_j, x_{j+1}] \text{ und } j < n \cdot \\ 0 & \text{sonst} \end{cases}$$

Die Basisfunktionen η_j und Ψ_j erfüllen

$$\begin{aligned} \eta_j(x_i) &= \delta_{ij}, \quad \eta_j'(x_i) = 0 \qquad \forall i, j = 0, \dots, n, \\ \Psi_j(x_i) &= 0, \quad \Psi_j'(x_i) = \delta_{ij} \quad \forall i, j = 0, \dots, n. \end{aligned}$$

Definition 7.24 (S_3-Interpolationsoperator)**.** Die Abbildung

$$I_3 : C^1[a,b] \to S_3, \quad (I_3 f)(x) := \sum_{j=0}^{n} f(x_j)\eta_j(x) + f'(x_j)\Psi_j(x)$$

heißt S_3-*Interpolationsoperator*.

Ist $f \in C^1[a,b]$, so erfüllt $I_3 f$

$$(I_3 f)(x_k) = \sum_{j=0}^{n} f(x_j) \underbrace{\eta_j(x_k)}_{=\delta_{jk}} + f'(x_j) \underbrace{\Psi_j(x_k)}_{=0} = f(x_k),$$

$$(I_3 f)'(x_k) = \sum_{j=0}^{n} f(x_j) \underbrace{\eta_j'(x_k)}_{=0} + f'(x_j) \underbrace{\Psi_j'(x_k)}_{=\delta_{jk}} = f'(x_k).$$

Insbesondere löst $I_3 f$ die Spline-Interpolationsaufgabe auf S_3. Nun wollen wir den Fehler

$$\|f - I_3 f\|_{C[a,b]}$$

untersuchen.

Lemma 7.25 (Hermite-Interpolationsfehler). *Es sei $n \in \mathbb{N}$ und $f \in C^{2n+2}[a,b]$. Ferner sei $p \in \Pi_{2n+1}$ eine Lösung der Hermite-Interpolationsaufgabe, d.h.*

$$\begin{aligned} p(x_k) &= f(x_k) \quad \forall k = 0, \dots, n, \\ p'(x_k) &= f'(x_k) \quad \forall k = 0, \dots, n. \end{aligned}$$

Dann gibt es zu jedem $\tilde{x} \in [a,b]$ ein $\xi \in (a,b)$, so dass

$$f(\tilde{x}) - p(\tilde{x}) = \frac{w^2(\tilde{x})}{(2n+2)!} f^{(2n+2)}(\xi)$$

mit $w(\tilde{x}) = (\tilde{x} - x_0) \cdot \ldots \cdot (\tilde{x} - x_n)$.

Beweis. Lemma von Rolle, siehe Kapitel 7.1.3. □

Satz 7.26 (Fehlerabschätzung für den S_3-Interpolationsoperator). *Es sei $f \in C^4[a,b]$ und $h_{max} := \max_{1 \le k \le n} |x_{k-1} - x_k|$. Dann gilt*

$$\|f - I_3 f\|_{C[a,b]} \le \frac{1}{384} h_{max}^4 \left\| f^{(4)} \right\|_{C[a,b]}.$$

Beweis. Nach Definition ist

$$\begin{aligned} (I_3 f)(x_k) &= f(x_k) \quad \forall k = 0, \dots, n, \\ (I_3 f)'(x_k) &= f'(x_k) \quad \forall k = 0, \dots, n \end{aligned} \tag{7.2}$$

und

$$I_3 f|_{[x_{k-1}, x_k]} \in \Pi_3 \quad \forall k = 0, \dots, n. \tag{7.3}$$

Sei nun $k \in \{1, \dots, n\}$ beliebig aber fest. Laut (7.2) und (7.3) löst $I_3 f|_{[x_{k-1}, x_k]} \in \Pi_{2n+1}$ (mit $n = 1$) die Hermite-Interpolationsaufgabe:
Finde $p \in \Pi_3$, so dass gilt

$$\begin{aligned} p(x_{k-1}) &= f(x_{k-1}), \quad p(x_k) &&= f(x_k) \\ p'(x_{k-1}) &= f'(x_{k-1}), \quad p'(x_k) &&= f'(x_k). \end{aligned}$$

Aus dem obigen Lemma (mit $n = 1, a = x_{k-1}, b = x_k$) folgt:
Zu jedem $\tilde{x} \in [x_{k-1}, x_k]$ gibt es ein $\xi \in (x_{k-1}, x_k)$, so dass

$$f(\tilde{x}) - p(\tilde{x}) = \frac{w^2(\tilde{x})}{4!} f^{(4)}(\xi).$$

Folglich gilt für jedes $\tilde{x} \in [x_{k-1}, x_k]$

$$\begin{aligned}
|f(\tilde{x}) - (I_3 f)(\tilde{x})| &= \frac{1}{4!}|\tilde{x} - x_{k-1}|^2 |\tilde{x} - x_k|^2 |f^{(4)}(\xi)| \\
&\leq \frac{1}{4!}|\tilde{x} - x_{k-1}|^2 |\tilde{x} - x_k|^2 \left\| f^{(4)}(x) \right\|_{C[a,b]} \\
&\overset{\tilde{x} \in [x_{k-1}, x_k]}{\leq} \frac{1}{4!}\frac{1}{4}|x_k - x_{k-1}|^2 \frac{1}{4}|x_k - x_k|^2 \left\| f^{(4)}(x) \right\|_{C[a,b]} \\
&= \frac{1}{384}|x_k - x_{k-1}|^4 \left\| f^{(4)}(x) \right\|_{C[a,b]} \\
&\leq \frac{1}{384} h_{max}^4 \left\| f^{(4)}(x) \right\|_{C[a,b]}.
\end{aligned}$$

Daraus folgt

$$\max_{\tilde{x} \in [x_{k-1}, x_k]} |f(\tilde{x}) - (I_3 f)(\tilde{x})| \leq \frac{1}{384} h_{max}^4 \left\| f^{(4)}(x) \right\|_{C[a,b]}.$$

Da $k \in \{1, \ldots, n\}$ beliebig gewählt wurde, folgt

$$\|f - I_3 f\|_{C[a,b]} \leq \frac{1}{384} h_{max}^4 \left\| f^{(4)}(x) \right\|_{C[a,b]}.$$

□

Fazit 7.27. Die Konvergenzrate beim S_3-Interpolationsoperator ist besser (2 Ordnunger mehr) als beim S_1- Interpolationsoperator. Der Nachteil ist, dass $f \in C^4[a, b]$ sein muss (bei I_1 nur $f \in C^2[a, b]$).

Kapitel 8

Numerische Integration

Wir untersuchen in diesem Kapitel numerische Verfahren zur Approximation von

$$\int_a^b f(x)\,\mathrm{d}x$$

für eine gegebene Funktion $f \in C[a,b]$.

8.1 Newton-Cotes-Formel

Idee:

(i) Wir betrachten eine äquidistante Zerlegung von $[a,b]$

$$x_k := a + kh, \quad k = 0,\dots,n, \quad n \in \mathbb{N}$$

mit

$$h := \frac{b-a}{n}.$$

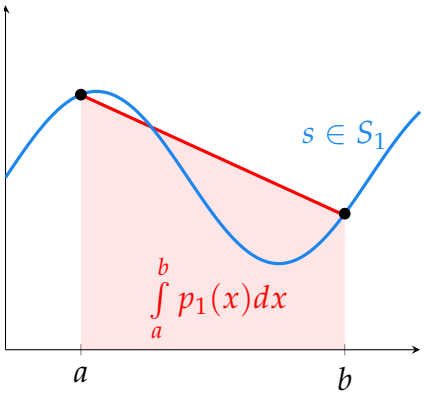
(ii) Wir lösen das Interpolationspolynom für die Punktepaare $(x_k, f(x_k))$, d.h. wir suchen ein Polynom $p \in \Pi_n$ mit

$$p(x_k) = f(x_k) \quad \forall k = 0,\dots,n.$$

(iii) Wir betrachten die Approximation

$$\int_a^b f(x)\,\mathrm{d}x \approx \int_a^b p(x)\,\mathrm{d}x.$$

Illustration:



Wir wissen bereits, dass die Interpolationsaufgabe in (ii) genau eine Lösung besitzt:

$$p(x) = \sum_{k=0}^{n} f(x_k) L_k(x),$$

wobei $L_k \in \Pi_n$ die Lagrange-Basispolynome bezeichnen. Folglich

$$\begin{aligned}
\int_a^b p(x)\,\mathrm{d}x &= \sum_{k=0}^{n} f(x_k) \int_a^b L_k(x)\,\mathrm{d}x \\
&= \sum_{k=0}^{n} f(x_k) \int_a^b \prod_{\substack{j=0 \\ j\neq k}}^{n} \frac{x - x_j}{x_k - x_j}\,\mathrm{d}x \\
&= \sum_{k=0}^{n} f(x_k) \int_a^b \prod_{\substack{j=0 \\ j\neq k}}^{n} \frac{x - (a+jh)}{(a+kh) - (a+jh)}\,\mathrm{d}x \\
&= \sum_{k=0}^{n} f(x_k) \int_a^b \prod_{\substack{j=0 \\ j\neq k}}^{n} \frac{\frac{x-a}{h} - j}{k - j}\,\mathrm{d}x.
\end{aligned}$$

Mit der Substitution

$$z = \frac{x-a}{h} \quad (\text{also } dz = \frac{1}{h}dx)$$

folgt

$$\begin{aligned}
\int_a^b p(x)\,\mathrm{d}x &= \sum_{k=0}^{n} f(x_k) h \int_0^n \prod_{\substack{j=0 \\ j\neq k}}^{n} \frac{z - j}{k - j}\,\mathrm{d}z \\
&= (b-a) \sum_{k=0}^{n} f(x_k) \frac{1}{n} \int_0^n \prod_{\substack{j=0 \\ j\neq k}}^{n} \frac{z - j}{k - j}\,\mathrm{d}z.
\end{aligned}$$

Definition 8.1 (Newton-Cotes-Quadraturformel)**.** Es sei $f \in C[a,b]$ und $n \in \mathbb{N}$. Die Näherungsformel

$$Q_n(f) := (b-a) \sum_{k=0}^{n} f(x_k)\sigma_k$$

mit den Gewichten

$$\sigma_k := \frac{1}{n} \int_0^n \prod_{\substack{j=0 \\ j\neq k}}^{n} \frac{z-j}{k-j} \, \mathrm{d}z, \quad k = 0, \ldots, n$$

heißt *Newton-Cotes-Quadraturformel*.

Beispiel 8.2.

(i) $n = 1$: Trapezregel

$$\sigma_0 = \frac{1}{1} \int_0^1 \frac{z-1}{0-1} \, \mathrm{d}z = -\frac{1}{2}(z-1)^2 \Big|_0^1 = \frac{1}{2},$$
$$\sigma_1 = \frac{1}{1} \int_0^1 \frac{z-0}{1-0} \, \mathrm{d}z = -\frac{1}{2}(z)^2 \Big|_0^1 = \frac{1}{2}.$$

Daraus folgt

$$\int_a^b f(x) \, \mathrm{d}x \approx Q_1(f) = (b-a) \sum_{k=0}^{1} f(x_k)\sigma_k = \frac{(b-a)}{2} \left(f(a) + f(b) \right),$$

denn $x_0 = a, x_1 = b$.

(ii) $n = 2$: Simpson-Regel

$$\sigma_0 = \frac{1}{2} \int_0^2 \frac{z-1}{0-1} \frac{z-2}{0-2} \, \mathrm{d}z = \frac{1}{4} \int_0^2 z^2 - 3z + 3 \, \mathrm{d}z = \frac{1}{6},$$
$$\sigma_1 = \frac{1}{2} \int_0^2 \frac{z-0}{1-0} \frac{z-2}{1-2} \, \mathrm{d}z = \frac{4}{6},$$
$$\sigma_2 = \frac{1}{6}.$$

(iii) $n = 3$: Newtonsche $\frac{3}{8}$-Regel

$$Q_3(f) = \frac{b-a}{8} \left(f(a) + 3f(\frac{2a+b}{3}) + 3f(\frac{a+2b}{3}) + f(b) \right).$$

Beachte, dass für die Newton-Cotes-Formel für $n \geq 8$ negative Gewichte σ_k auftreten. Deshalb ist die Formel nur für kleines $n \in \mathbb{N}$ gut geeignet. Um die Genauigkeit zu erhöhen, verwendet man die Newton-Cotes-Formel für kleines n ($n \in \{1,2,3\}$) auf Teilintervallen von $[a,b]$ und summiert auf. Dies führt auf die *summierte Newton-Cotes-Formel*.

n	σ_k				Name
1	$\frac{1}{2}$	$\frac{1}{2}$			Trapezregel
2	$\frac{1}{6}$	$\frac{4}{6}$	$\frac{1}{6}$		Simpson-Regel
3	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	Newtonsche $\frac{3}{8}$-Regel

Beispiel 8.3 (summierte Trapezregel)**.**

(i) Wir betrachten eine äquidistante Zerlegung von $[a, b]$

$$x_k = a + kh, \quad k = 0, \ldots, m, \quad k = \frac{b-a}{m}.$$

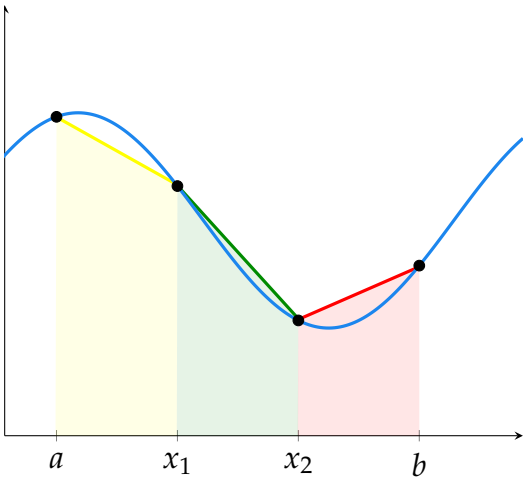
(ii) Das Integral von f auf jedem Teilintervall $[x_{k-1}, x_k]$ approximieren wir durch die Trapezregel

$$\int_{x_{k-1}}^{x_k} f(x)\,\mathrm{d}x \approx \frac{x_k - x_{k-1}}{2}\left(f(x_{k-1}) + f(x_k)\right).$$

(iii) Wir summieren die obigen Formeln auf

$$\begin{aligned}\int_a^b f(x)\,\mathrm{d}x &= \int_{x_0=a}^{x_1} f(x)\,\mathrm{d}x + \ldots + \int_{x_{m-1}}^{x_m=b} f(x)\,\mathrm{d}x \\ &\approx \frac{h}{2}\left(f(a) + f(x_1)\right) + \ldots + \frac{h}{2}\left(f(x_{m-1}) + f(b)\right) \\ &= h\left(\frac{1}{2}f(a) + \sum_{j=1}^{m-1} f(x_j) + \frac{1}{2}f(b)\right) \\ &=: T_m(F).\end{aligned}$$

Illustration:



8.2 Fehler von allgemeinen Quadraturformeln

Im Folgenden betrachten wir das Integral

$$I(f) = \int_0^1 f(x)\,\mathrm{d}x$$

und eine allgemeine Quadraturformel der Gestalt

$$Q(f) := \sum_{j=0}^{n} \omega_j f(x_j) \tag{8.1}$$

mit $x_j \in [0,1]$ und $\omega_j \in \mathbb{R}$ für alle $j = 0, \ldots, n$.

Definition 8.4. Eine Quadraturformel der Gestalt (8.1) hat die *Fehlerordnung* $m \in \mathbb{N}$ (den Genauigkeitsgrad m), falls gilt:

$$Q(p) = \int_0^1 p(x)\,\mathrm{d}x \quad \forall p \in \Pi_{m-1}$$
$$Q(\tilde{p}) \neq \int_0^1 \tilde{p}(x)\,\mathrm{d}x$$

für ein Polynom $\tilde{p} \in \Pi_m$.

Beispiel 8.5. Die Trapezregel ist eine Quadraturformel der Ordnung 2, denn

$$T(f) = \frac{1}{2}\left(f(0) + f(1)\right).$$

Also

$$T(p) = \int_0^1 p(x)\,\mathrm{d}x \quad \forall p \in \Pi_1 \quad \text{und} \quad T(x^2) = \frac{1}{2}(0^2 + 1^2) = \frac{1}{2} \neq \int_0^1 x^2\,\mathrm{d}x = \frac{1}{3}.$$

Satz 8.6. *Es sei Q eine Quadraturformel der Ordnung $m \in \mathbb{N}$. Dann gibt es eine Funktion $k : [0,1] \to \mathbb{R}$, so dass gilt*

$$\int_0^1 f(x)\,\mathrm{d}x - Q(f) = \int_0^1 k(x) f^{(m)}(x)\,\mathrm{d}x \quad \forall f \in C^m[0,1].$$

Die Funktion k heißt Peano-Kern der Quadraturformel.

Beweis. Es sei $f \in C^m[0,1]$. Der Satz von Taylor liefert

$$f(x) = \underbrace{\sum_{k=0}^{m-1} \frac{f^{(k)}(0)}{k!} x^k}_{=:p(x)} + \underbrace{\frac{1}{(m-1)!} \int_0^x f^{(m)}(t)(x-t)^{m-1}\,\mathrm{d}t}_{=:r(x)}.$$

Das heißt,

$$f(x) = p(x) + r(x).$$

Da $p \in \Pi_{m-1}$ ist und Q die Fehlerordnung m hat, gilt

$$Q(p) = \int_0^1 p(x)\,\mathrm{d}x.$$

Folglich gilt

$$\int_0^1 f(x)\,\mathrm{d}x - Q(f) = \int_0^1 p(x) + r(x)\,\mathrm{d}x - Q(p+r) = \int_0^1 r(x)\,\mathrm{d}x - Q(r) =: I.$$

Für I gilt

$$\begin{aligned} I &= \frac{1}{(m-1)!}\left(\int_0^1\int_0^x f^{(m)}(t)(x-t)^{m-1}\,\mathrm{d}t\,\mathrm{d}x - \sum_{j=0}^n \omega_j \int_0^{x_j} f^{(m)}(t)(x_j-t)^{m-1}\,\mathrm{d}t\right) \\ &= \frac{1}{(m-1)!}\left(\int_0^1\int_t^1 f^{(m)}(t)(x-t)^{m-1}\,\mathrm{d}t\,\mathrm{d}x - \sum_{j=0}^n \omega_j \int_0^1 f^{(m)}(t)(x_j-t)_+^{m-1}\,\mathrm{d}t\right), \end{aligned}$$

wobei $(c)_+ = \max\{c, 0\}$. Insgesamt gilt

$$\begin{aligned} I &= \frac{1}{(m-1)!}\left(\int_0^1 f^{(m)}(t)\frac{1}{m}(1-t)^{m-1}\,\mathrm{d}t - \sum_{j=0}^n \omega_j \int_0^1 f^{(m)}(t)(x_j-t)_+^{m-1}\,\mathrm{d}t\right) \\ &= \frac{1}{(m-1)!}\int_0^1 \underbrace{\left(\frac{1}{m}(1-t)^m - \sum_{j=0}^n \omega_j (x_j-t)_+^{m-1}\right)}_{=:k(t)} f^{(m)}(t)\,\mathrm{d}t. \end{aligned}$$

Es folgt also die Behauptung

$$\int_0^1 f(x)\,\mathrm{d}x - Q(t) = \int_0^1 k(x) f^{(m)}(x)\,\mathrm{d}x \quad \forall f \in C^m[0,1]$$

mit

$$k(x) = \frac{1}{(m-1)!}\left(\frac{1}{m}(1-t)^m - \sum_{j=0}^n \omega_j (x_j-t)_+^{m-1}\right).$$

□

Beispiel 8.7. Die Trapezregel

$$T(f) = \frac{1}{2}(f(0) + f(1))$$

ist eine Quadraturformel der Ordnung 2. Der Satz ist also anwendbar, und der Peano-Kern k ist gegeben durch

$$\begin{aligned} k_T(x) &= \frac{1}{1!}\left(\frac{1}{2}(1-x)^2 - \frac{1}{2}(-x)_+ - \frac{1}{2}(1-x)_+\right) \\ &= \frac{1}{2}(1-x)^2 - \frac{1}{2}(1-x) \quad \forall x \in [0,1] \\ &\left(= -\frac{1}{2}x(1-x)\right). \end{aligned}$$

Es gilt für jede Funktion $f \in C^2[0,1]$:

$$\int_0^1 f(x)\,\mathrm{d}x - T(f) = -\frac{1}{2}\int_0^1 x(1-x)f^{(2)}(x)\,\mathrm{d}x.$$