Definition 2.1: Maschinenzahlen / Gleitpunktzahlen

- Unter der zusätzlichen Normierungs–Bedingung $m_1 \neq 0$ (falls $x \neq 0$) ergibt sich eine eindeutige Darstellung der sogenannten ***maschinendarstellbaren Zahlen M*** zur Basis B:

$$M = \left\{ x \epsilon \mathbb{R} |\, x = \pm 0.m_1 m_2 m_3 ... m_n \cdot B^{\pm e_1 e_2 ... e_l} \right\} \cup \{0\}$$

  Dabei gilt m_i, $e_i \epsilon \{0,1,..., B\text{-}1\}$ für $i \neq 0$ und $B \in \mathbb{N},\ B > 1$)

- Der **Wert** einer solchen Zahl ist definiert als

$$\sum_{i=1}^{n} m_i B^{e-i}$$

  und ergibt gerade die (nicht normierte) Darstellung der Zahl im Dezimalsystem. Dabei ist e ebenfalls im Dezimalsystem zu nehmen, also $e = \sum_{i=1}^{l} e_i B^{l-i}$ und es gilt $e \in \mathbb{Z}$, d.h. e kann natürlich auch negativ sein. Weiter gibt es eine obere und untere Schranke: $e_{min} \leq e \leq e_{max}$.

- Man redet dann auch von einer ***n*-stelligen Gleitpunktzahl zur Basis *B*** (engl: floating point). Zahlen, die nicht in dieser Menge M liegen, müssen durch Rundung in eine maschinendarstellbare Zahl umgewandelt werden.

Definition 2.2: Absoluter / Relativer Fehler

- Hat man eine Näherung $\widetilde{x}$ zu einem exakten Wert x, so ist der Betrag der Differenz $|\, \widetilde{x} - x \,|$ der **absolute Fehler.**
- Falls $x \neq 0$, so ist $|\, \frac{\widetilde{x}-x}{x} \,|$bzw. $\frac{|\widetilde{x}-x|}{|x|}$der **relative Fehler** dieser Näherung.

  Bemerkung: In der Numerik ist der relative Fehler der wichtigere. Weshalb?

Definition 2.3: Maschinengenauigkeit

- Die Zahl $eps := 5 \cdot 10^{-n}$ heisst **Maschinengenauigkeit**. Bei allgemeiner Basis B gilt $eps := \frac{B}{2} \cdot B^{-n}$. Sie gibt den maximalen relativen Fehler, der durch Rundung entstehen kann.
- Alternative Definition: Die Maschinengenauigkeit ist die kleinste positive Maschinenzahl, für die auf dem Rechner $1 + eps \neq 1$ gilt.

$$|\, \frac{rd(x) - x}{x} \,| \leq 5 \cdot 10^{-n} \ (da\ x \geq 10^{e-1}).$$

- Näherung für den **absoluten Fehler bei Funktionsauswertungen**:

$$| f(\tilde{x}) - f(x) | \approx | f'(x) | \cdot | \tilde{x} - x |$$

- Näherung für den **relativen Fehler bei Funktionsauswertungen**:

$$\frac{| f(\widetilde{x}) - f(x) |}{| f(x) |} \approx \frac{| f'(x) | \cdot | x |}{| f(x) |} \cdot \frac{| \tilde{x} - x |}{| x |}$$

Definition 2.4: Konditionszahl

- Den Faktor

$$K := \frac{| f'(x) | \cdot | x |}{| f(x) |}$$

  nennt man **Konditionszahl**.

- Man unterscheidet **gut konditionierte Probleme**, d.h. die Konditionszahl ist klein, und **schlecht konditionierte Probleme** (ill posed problems) mit grosser Konditionszahl. Bei gut konditionierten Problemen wird der relative Fehler durch die Auswertung der Funktion nicht grösser.

Satz 3.1: Nullstellensatz von Bolzano

- Sei $f : [a, b] \to \mathbb{R}$ stetig mit $f(a) \leq 0 \leq f(b)$ oder $f(a) \geq 0 \geq f(b)$. Dann muss f in $[a, b]$ eine Nullstelle besitzen.

Satz 3.2: Bisektionsverfahren [1]

- Gegeben sei eine stetige Funktion $f : [a, b] \longrightarrow \mathbb{R}$ mit $f(a) \cdot f(b) < 0$. In jedem der über die Rekursion für $i = 0, 1, ...$erzeugten Intervalle

$$\begin{aligned} [a_0,\, b_0] &= [a,\, b]; \\ [a_{i+1},\, b_{i+1}] &= \begin{cases} \left[a_i,\, \frac{a_i+b_i}{2}\right] & \text{falls } f\left(\frac{a_i+b_i}{2}\right) \cdot f(a_i) \leq 0 \\ \left[\frac{a_i+b_i}{2},\, b_i\right] & \text{sonst} \end{cases} \end{aligned}$$

  befindet sich eine Nullstelle von f und es gilt

$$b_i - a_i = \frac{b-a}{2^i}, \text{ insbesondere also } \lim_{i \to \infty} (b_i - a_i) = 0$$

Definition 3.1: Fixpunktgleichung / Fixpunkt [1]

- Eine Gleichung der Form $F(x) = x$ heisst **Fixpunktgleichung**.
- Ihre Lösungen $\overline{x}$, für die $F(\overline{x}) = \overline{x}$ erfüllt ist, heissen **Fixpunkte** (da die Funktion F die Punkte $\overline{x}$ auf sich selbst abbildet.

Anstelle eines Nullstellenproblems kann man also ein dazu äquivalentes Fixpunktproblem betrachten. Dazu muss aber $f(x) = 0$ in die Fixpunktform $F(x) = x$ gebracht werden, wozu es viele Möglichkeiten gibt. Bei dieser Überführung muss unbedingt auf Äquivalenz geachtet werden, d.h. die Lösungsmenge darf nicht verändert werden.

Beispiel 3.2:

- Die Gleichung $p(x) = x^3 - x + 0.3$ soll in Fixpunktform gebracht werden.
  Lösung: Die einfachste Möglichkeit ist $p(x) = 0 \iff F(x) \equiv x^3 + 0.3 = x$
  Aber auch $F(x) \equiv \sqrt[3]{x - 0.3} = x$ ist möglich.
- Die Gleichung $x = \cos(x)$, die wir weiter oben graphisch gelöst haben, ist bereits in der Fixpunktform.

Definition 3.2: Fixpunktiteration [1]

- Gegeben sei $F : [a, b] \to \mathbb{R}$, mit $x_0 \in [a, b]$. Die rekursive Folge

$$x_{n+1} \equiv F(x_n), \qquad n = 0, 1, 2, ...$$

  heisst Fixpunktiteration von F zum Startwert x_0.

Satz 3.2 zur Fixpunktiteration [1]:

- Sei $F : [a, b] \longrightarrow \mathbb{R}$ mit stetiger Ableitung F' und $\overline{x} \in [a, b]$ ein Fixpunkt von F. Dann gilt für die Fixpunktiteration $x_{n+1} = F(x_n)$:
  - Ist $\mid F'(\overline{x}) \mid < 1$, so konvergiert x_n gegen $\bar{x}$, falls der Startwert x_0 nahe genug bei $\bar{x}$ liegt. Der Punkt $\bar{x}$ heisst dann **anziehender Fixpunkt.**
  - Ist $\mid F'(\overline{x}) \mid > 1$, so konvergiert x_n für keinen Startwert $x_0 \neq \bar{x}$. Der Punkt $\bar{x}$ heisst dann **abstossender Fixpunkt.**

Satz 3.3: Banachscher Fixpunktsatz [1]

- Sei $F : [a,b] \longrightarrow [a,b]$ (d.h. F bildet $[a,b]$ auf sich selber ab) und es existiere eine Konstante α mit $0 < \alpha < 1$ und

$$| F(x) - F(y) | \leq \alpha \, | x - y | \quad \text{für alle } x, y \in [a,b]$$

(d.h. F ist "Lipschitz-stetig" und "kontraktiv", α nennt man auch Lipschitz-Konstante). Dann gilt:

  - F hat genau einen Fixpunkt $\overline{x}$ in [a,b]
  - Die Fixpunktiteration $x_{n+1} = F(x_n)$ konvergiert gegen $\overline{x}$ für alle Startwerte $x_0 \in [a,b]$
  - Es gelten die Fehlerabschätzungen

$$\begin{aligned} | x_n - \overline{x} | &\leq \frac{\alpha^n}{1-\alpha} | x_1 - x_0 | \quad \text{a-priori Abschätzung} \\ | x_n - \overline{x} | &\leq \frac{\alpha}{1-\alpha} | x_n - x_{n-1} | \quad \text{a-posteriori Abschätzung} \end{aligned}$$

Bemerkungen:

- Aus $| F(x) - F(y) | \leq \alpha \, | x - y |$ für alle $x, y \in [a,b]$ folgt

$$\frac{| F(x) - F(y) |}{| x - y |} \leq \alpha,$$

wobei die linke Seite sämtliche möglichen Steigungen der Sekanten durch die beiden Punkte $(x, F(x))$ und $(y, F(y))$ für alle $x, y \in [a,b]$ darstellt. Aus diesem Grund kann man α als die grösstmögliche Steigung von $F(x)$ auf dem Intervall $[a,b]$ interpretieren, bzw.

$$\alpha = \max_{x_0 \in [a,b]} | F'(x_o) |$$

- Wählt man das Intervall $[a,b]$ sehr nahe um einen anziehenden Fixpunkt $\overline{x}$, so ist also $\alpha \approx | F'(\overline{x}) |$.
- In der Praxis gestaltet es sich meist schwierig, ein Intervall [a,b] zu finden, dass unter F auf sich selbst abgebildet wird. Hat man ein solches Intervall gefunden, dann sind die Fehlerabschätzungen aber recht nützlich. Wir werden diesen Satz nochmals im Zusammenhang mit der iterativen Lösung von linearen Gleichungssystemen in Kap. 4 aufgreifen.

Auflösen nach x_{n+1} liefert

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \qquad (n = 0, 1, 2, 3, ...).$$

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_0)} \qquad (n = 0, 1, 2, 3, ...).$$

Definition 3.3: Konvergenzordnung [1]

- Sei (x_n) eine gegen $\overline{x}$ konvergierende Folge. Dann hat das Verfahren die **Konvergenzordnung** q≥1 wenn es eine Konstante c > 0 gibt mit

$$| x_{n+1} - \overline{x} | \leq c \cdot | x_n - \overline{x} |^q$$

für alle n. Falls $q = 1$ verlangt man noch $c < 1$. Im Fall $q = 1$ spricht man von linearer, im Fall $q = 2$ von quadratischer Konvergenz.

Definition 4.1: Untere Dreiecksmatrix / Obere Dreiecksmatrix [6]

- Eine $n \times n$ Matrix $\boldsymbol{L} = (l_{ij})$ heisst **untere Dreiecksmatrix**, wenn $l_{ij} = 0$ für $j > i$ gilt; sie heisst **normierte untere Dreiecksmatrix**, wenn ausserdem $l_{ii} = 1$ für alle i gilt.
- Eine $n \times n$ Matrix $\boldsymbol{R} = (r_{ij})$ heisst **obere Dreiecksmatrix**, wenn $r_{ij} = 0$ für $i > j$ gilt; sie heisst **normierte obere Dreiecksmatrix**, wenn ausserdem $r_{ii} = 1$ für alle i gilt.

- Untere normierte Dreiecksmatrix:

$$\boldsymbol{L} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ l_{21} & 1 & 0 & \cdots & 0 \\ l_{31} & l_{32} & 1 & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & 0 \\ l_{n1} & l_{n2} & \cdots & l_{nn-1} & 1 \end{pmatrix}$$

- Obere Dreiecksmatrix:

$$\boldsymbol{R} = \begin{pmatrix} r_{11} & r_{12} & r_{13} & \cdots & r_{1n} \\ 0 & r_{22} & r_{23} & \cdots & r_{2n} \\ 0 & 0 & r_{33} & \cdots & r_{3n} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & r_{nn} \end{pmatrix}$$

Gauss-Algorithmus zur Transformation von $\boldsymbol{Ax} = \boldsymbol{b}$ auf ein oberes Dreiecksystem [1]:

- für $i = 1, ..., n-1$:

  erzeuge Nullen unterhalb des Diagonalelementes in der i-ten Spalte

  - Falls nötig und möglich, sorge durch Zeilenvertauschung für $a_{ii} \neq 0$:

    falls $a_{ii} \neq 0$: tue nichts

    falls $a_{ii} = 0$:
    - falls $a_{ji} = 0$ für alle $j = i+1, ..., n$:
      - A ist nicht regulär; stop;
    - wenn $a_{ji} \neq 0$ für ein $j = i+1, ..., n$:
      - sei $j \geq i+1$ der kleinste Index mit $a_{ji} \neq 0$
      - $z_i \longleftrightarrow z_j$
  - Eliminationsschritt:

    für $j = i+1, ..., n$ eliminiere das Element a_{ji} durch:

$$z_j := z_j - \frac{a_{ji}}{a_{ii}} \cdot z_i$$

Gauss-Algorithmus zur Transformation von $\boldsymbol{Ax} = \boldsymbol{b}$ mit Spaltenpivotisierung [1]:

- für $i = 1, ..., n-1$:

  erzeuge Nullen unterhalb des Diagonalelementes in der i-ten Spalte

  - Suche das betragsgrösste Element unterhalb der Diagonalen in der i-ten Spalte:
    Wähle k so, dass $|a_{ki}| = \max\{|a_{ji}| \mid j = i, ..., n\}$
    $\begin{cases} \text{falls } a_{ki} = 0: & A \text{ ist nicht regulär; stop;} \\ \text{falls } a_{ki} \neq 0: & z_k \longleftrightarrow z_i; \end{cases}$
  - Eliminationsschritt:
    für $j = i+1, ..., n$ eliminiere das Element a_{ji} durch:

$$z_j := z_j - \frac{a_{ji}}{a_{ii}} \cdot z_i$$

Satz 4.1: $\boldsymbol{LR}$-Zerlegung [1]
Zu jeder regulären $n \times n$ Matrix $\boldsymbol{A}$, für die der Gauss-Algorithmus ohne Zeilenvertauschung durchführbar ist, gibt es $n \times n$ Matrizen $\boldsymbol{L}$ und $\boldsymbol{R}$ mit den folgenden Eigenschaften:

- $\boldsymbol{L}$ ist eine normierte untere Dreiecksmatrix (also mit $l_{ii} = 1$ für $i = 1, ..., n$)
- $\boldsymbol{R}$ ist eine obere Dreiecksmatrix mit $r_{ii} \neq 0$ für $i = 1, ..., n$
- $\boldsymbol{A} = \boldsymbol{L} \cdot \boldsymbol{R}$ ist die $\boldsymbol{LR}$**-Zerlegung von** $\boldsymbol{A}$.

Aufwand: Die Berechnung der $\boldsymbol{LR}$-Zerlegung mit dem Gauss-Algorithmus benötigt $\frac{1}{3}(n^3 - n)$ Punktoperationen

4.5.2 Die Cholesky-Zerlegung

Im folgenden lernen wir ein weiteres Verfahren zur Dreieckszerlegung von Matrizen kennen, einen Spezialfall der $\boldsymbol{LR}$-Zerlegung. Dieses Verfahren ist nach seinem Entdecker André-Louis Cholesky (1875 -1918) benannt, einem französischen Mathematiker. Er entwickelte es für Anwendungen in der Geodäsie. [9] Die Cholesky-Zerlegung funktioniert nicht für allgemeine Matrizen sondern nur für *symmetrische, positiv definite* Matrizen. Falls anwendbar, ist es aber etwa um einen Faktor zwei effizienter als die allgemeine $\boldsymbol{LR}$-Zerlegung.

Definition 4.2: Symmetrische / positiv definite Matrizen [1]

- Eine Matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ heisst symmetrisch, falls $\boldsymbol{A}^T = \boldsymbol{A}$ gilt ($\boldsymbol{A}^T$ ist die transponierte Matrix).
- Eine Matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ heisst positiv definit, falls für alle $\boldsymbol{x} \in \mathbb{R}^n$, $x \neq 0$ gilt $\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} > 0$.

Satz 4.2: Cholesky Zerlegung [1]
Für jede positiv definite $n \times n$ Matrix $\boldsymbol{A}$ gibt es genau eine rechts-obere Dreiecksmatrix $\boldsymbol{R}$ mit $r_{ii} > 0$ für $i = 1, ..., n$ und $\boldsymbol{A} = \boldsymbol{R}^T \boldsymbol{R}$. Diese Zerlegung heisst **Cholesky-Zerlegung** von $\boldsymbol{A}$.

Die Berechnung der Cholesky-Zerlegung geschieht anhand des folgenden Algorithmus, der uns die Koeffizienten r_{ij} der oberen Dreiecksmatrix $\boldsymbol{R}$ berechnet und gleichzeitig überprüft, ob $\boldsymbol{A}$ positiv definit ist:

Cholesky-Algorithmus [1]:
Gegeben sei eine symmetrische $n \times n$ Matrix $\boldsymbol{A}$. Für $i = 1, ..., n$ berechne:

- $S = a_{ii} - \sum_{k=1}^{i-1} r_{ki}^2$ (für $i = 1$ ist also $S = a_{ii}$)
- falls $S \leq 0$, dann ist $\boldsymbol{A}$ nicht positiv definit $\rightarrow$stopp.
- falls $S > 0$:
  - $r_{ii} = \sqrt{S}$
  - für $j = i + 1, ..., n$: $r_{ij} = \frac{1}{r_{ii}} \left(a_{ij} - \sum_{k=1}^{i-1} r_{ki} r_{kj} \right)$

Definition 4.3: Vektornorm [1]
Eine Abbildung $\|.\| : \mathbb{R}^n \longrightarrow \mathbb{R}$ heisst Vektornorm, wenn die folgenden Bedingungen für alle $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, $\lambda \in \mathbb{R}$ erfüllt sind:

- $\|\boldsymbol{x}\| \geq 0$ und $\|\boldsymbol{x}\| = 0 \Longleftrightarrow x = 0$
- $\|\lambda \boldsymbol{x}\| = \mid \lambda \mid \|\boldsymbol{x}\|$
- $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$ "Dreiecksungleichung"

Definition 4.4: Vektornormen / Matrixnormen [1]

- Für Vektoren $\boldsymbol{x} = (x_1, x_2, ..., x_n)^{\mathrm{T}} \in \mathbb{R}^n$ gibt es die folgenden Vektornormen:

$$\begin{aligned} \text{1-Norm, Summennorm} &: \| \boldsymbol{x} \|_1 = \sum_{i=1}^{n} \mid x_i \mid \\ \text{2-Norm, euklidische Norm} &: \| \boldsymbol{x} \|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} \\ \infty\text{-Norm, Maximumnorm} &: \| \boldsymbol{x} \|_\infty = \max_{i=1,...,n} \mid x_i \mid \end{aligned}$$

- Für eine $n \times n$ Matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ sind mit den Vektornormen die folgenden Matrixnormen verbunden, welche die Eigenschaften der Definition 4.3 ebenfalls erfüllen:

$$\begin{aligned} \text{1-Norm, Spaltensummennorm} &: \| \boldsymbol{A} \|_1 = \max_{j=1,...,n} \sum_{i=1}^{n} \mid a_{ij} \mid \\ \text{2-Norm, Spektralnorm} &: \| \boldsymbol{A} \|_2 = \sqrt{\rho(\boldsymbol{A}^T \boldsymbol{A})} \\ \infty\text{-Norm, Zeilensummennorm} &: \| \boldsymbol{A} \|_\infty = \max_{i=1,...,n} \sum_{j=1}^{n} \mid a_{ij} \mid \end{aligned}$$

- Berechnen Sie die 1-, 2 -, und $\infty-$ Norm des Vektors $\begin{pmatrix} -1 \\ 2 \\ 3 \end{pmatrix}$ sowie die 1- und $\infty-$ Norm von $\begin{pmatrix} 1 & 2 & 3 \\ 3 & 4 & -2 \\ 7 & -3 & 5 \end{pmatrix}$.

  Lösung:

$$\left\| \begin{pmatrix} -1 \\ 2 \\ 3 \end{pmatrix} \right\|_1 = 1+2+3 = 6, \quad \left\| \begin{pmatrix} -1 \\ 2 \\ 3 \end{pmatrix} \right\|_2 = \sqrt{1+2^2+3^2} = \sqrt{14},$$

$$\left\| \begin{pmatrix} -1 \\ 2 \\ 3 \end{pmatrix} \right\|_\infty = \max\{1, 2, 3\} = 3$$

$$\left\| \begin{pmatrix} 1 & 2 & 3 \\ 3 & 4 & -2 \\ 7 & -3 & 5 \end{pmatrix} \right\|_1 = \max\{1+3+7, 2+4+3, 3+2+5\} = 11$$

$$\left\| \begin{pmatrix} 1 & 2 & 3 \\ 3 & 4 & -2 \\ 7 & -3 & 5 \end{pmatrix} \right\|_\infty = \max\{1+2+3, 3+4+2, 7+3+5\} = 15.$$

Satz 4.3: Abschätzung für fehlerbehaftete Vektoren [1]

- Sei $\| \cdot \|$eine Norm, $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ eine reguläre $n \times n$ Matrix und $\boldsymbol{x}$, $\widetilde{\boldsymbol{x}}$, $\boldsymbol{b}$, $\widetilde{\boldsymbol{b}} \in \mathbb{R}^n$ mit $\boldsymbol{Ax} = \boldsymbol{b}$ und $\boldsymbol{A}\widetilde{\boldsymbol{x}} = \widetilde{\boldsymbol{b}}$. Dann gilt für den absoluten und den relativen Fehler in $\boldsymbol{x}$:
  - $\| \boldsymbol{x} - \widetilde{\boldsymbol{x}} \| \leq \| \boldsymbol{A}^{-1} \| \cdot \| \boldsymbol{b} - \widetilde{\boldsymbol{b}} \|$
  - $\frac{\|\boldsymbol{x}-\widetilde{\boldsymbol{x}}\|}{\|\boldsymbol{x}\|} \leq \| \boldsymbol{A} \| \cdot \| \boldsymbol{A}^{-1} \| \cdot \frac{\|\boldsymbol{b}-\widetilde{\boldsymbol{b}}\|}{\|\boldsymbol{b}\|}$ falls $\| \boldsymbol{b} \| \neq 0$
- Die Zahl $\text{cond}(\boldsymbol{A}) = \| \boldsymbol{A} \| \cdot \| \boldsymbol{A}^{-1} \|$ nennt man Konditionszahl der Matrix $\boldsymbol{A}$ bzgl. der verwendeten Norm.

Satz 4.4: Abschätzung für fehlerbehaftete Matrix [1]
Sei $\| \cdot \|$eine Norm, $\boldsymbol{A}, \widetilde{\boldsymbol{A}} \in \mathbb{R}^{n \times n}$ reguläre $n \times n$ Matrizen und $\boldsymbol{x}$, $\widetilde{\boldsymbol{x}}$, $\boldsymbol{b}$, $\widetilde{\boldsymbol{b}} \in \mathbb{R}^n$ mit $\boldsymbol{Ax} = \boldsymbol{b}$ und $\boldsymbol{A}\widetilde{\boldsymbol{x}} = \widetilde{\boldsymbol{b}}$. Falls

$$\text{cond}(\boldsymbol{A}) \cdot \frac{\| \boldsymbol{A} - \widetilde{\boldsymbol{A}} \|}{\| \boldsymbol{A} \|} < 1$$

dann gilt:

$$\frac{\| \boldsymbol{x} - \widetilde{\boldsymbol{x}} \|}{\| \boldsymbol{x} \|} \leq \frac{\text{cond}(\boldsymbol{A})}{1 - \text{cond}(\boldsymbol{A}) \cdot \frac{\|\boldsymbol{A}-\widetilde{\boldsymbol{A}}\|}{\|\boldsymbol{A}\|}} \cdot \left(\frac{\| \boldsymbol{A} - \widetilde{\boldsymbol{A}} \|}{\| \boldsymbol{A} \|} + \frac{\| \boldsymbol{b} - \widetilde{\boldsymbol{b}} \|}{\| \boldsymbol{b} \|} \right)$$

Bemerkung:

- Für den Fall, dass $\boldsymbol{A}$ exakt gegeben ist, gilt $\frac{\|\boldsymbol{A}-\widetilde{\boldsymbol{A}}\|}{\|\boldsymbol{A}\|} = 0$ und der relative Fehler für $\boldsymbol{x}$ aus Satz 4.4 reduziert sich auf den relativen Fehler in Satz 4.3.

Beispiel 4.11 [1]:

- Nehmen Sie noch einmal das Beispiel 4.9 und untersuchen Sie die Fehlerfortpflanzung unter der zusätzlichen Annahme, dass die Matrix $\boldsymbol{A}$ um maximal 0.003 elementweise gestört ist.
- Lösung: Wir hatten bereits die folgenden Grössen berechnet

$$\| \boldsymbol{A} \|_\infty = 12.1,\ \mathrm{cond}(\boldsymbol{A}) = 732.05,\ \|\boldsymbol{b}\|_\infty = 1.5,\ \| \boldsymbol{b} - \widetilde{\boldsymbol{b}} \|_\infty \leq 0.1$$

Wenn nun jedes Element von $\boldsymbol{A}$ um maximal 0.003 gestört wird, summiert sich diese Störung in der $\infty-$Norm auf und wir erhalten $\|\boldsymbol{A} - \widetilde{\boldsymbol{A}}\|_\infty \leq 0.006$ und damit

$$\mathrm{cond}(\boldsymbol{A}) \cdot \frac{\| \boldsymbol{A} - \widetilde{\boldsymbol{A}} \|_\infty}{\| \boldsymbol{A} \|_\infty} \leq 0.363 < 1.$$

Wir können also die Abschätzung aus Satz 4.4 anwenden und erhalten

$$\frac{\| \boldsymbol{x} - \widetilde{\boldsymbol{x}} \|_\infty}{\| \boldsymbol{x} \|_\infty} \leq \frac{732.05}{1 - 0.363} \left(\frac{0.006}{12.1} + \frac{0.1}{1.5} \right) \leq 77.2$$

Definition 4.6: Jacobi- bzw. Gesamtschrittverfahren [1]

- Zu lösen sei $\boldsymbol{Ax} = \boldsymbol{b}$. Die Matrix $\boldsymbol{A} = (a_{ij})$ sei zerlegt in der Form

$$\boldsymbol{A} = \underbrace{\begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ a_{21} & 0 & 0 & \cdots & 0 \\ a_{31} & a_{32} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn-1} & 0 \end{pmatrix}}_{=: \boldsymbol{L}} + \underbrace{\begin{pmatrix} a_{11} & 0 & 0 & \cdots & 0 \\ 0 & a_{22} & 0 & \cdots & 0 \\ 0 & 0 & a_{33} & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & a_{nn} \end{pmatrix}}_{=: \boldsymbol{D}}$$

$$+ \underbrace{\begin{pmatrix} 0 & a_{12} & a_{13} & \cdots & a_{1n} \\ 0 & 0 & a_{23} & \cdots & a_{2n} \\ 0 & 0 & 0 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & a_{n-1,n} \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix}}_{=: \boldsymbol{R}}$$

Dann heisst die Fixpunktiteration

$$\begin{aligned} \boldsymbol{D}\boldsymbol{x}^{(k+1)} &= -(\boldsymbol{L} + \boldsymbol{R})\boldsymbol{x}^{(k)} + \boldsymbol{b} \text{ bzw.} \\ \boldsymbol{x}^{(k+1)} &= -\boldsymbol{D}^{-1}(\boldsymbol{L} + \boldsymbol{R})\boldsymbol{x}^{(k)} + \boldsymbol{D}^{-1}\boldsymbol{b} \end{aligned}$$

Gesamtschrittverfahren oder **Jacobi-Verfahren**.

Definition 4.7: Gauss-Seidel bzw. Einzelschrittverfahren [1]

- Zu lösen sei $\boldsymbol{Ax} = \boldsymbol{b}$. Die Matrix $\boldsymbol{A} = (a_{ij})$ sei zerlegt in der Form

$$\boldsymbol{A} = \underbrace{\begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ a_{21} & 0 & 0 & \cdots & 0 \\ a_{31} & a_{32} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn-1} & 0 \end{pmatrix}}_{=:\boldsymbol{L}} + \underbrace{\begin{pmatrix} a_{11} & 0 & 0 & \cdots & 0 \\ 0 & a_{22} & 0 & \cdots & 0 \\ 0 & 0 & a_{33} & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & a_{nn} \end{pmatrix}}_{=:\boldsymbol{D}}$$

$$+ \underbrace{\begin{pmatrix} 0 & a_{12} & a_{13} & \cdots & a_{1n} \\ 0 & 0 & a_{23} & \cdots & a_{2n} \\ 0 & 0 & 0 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & a_{n-1,n} \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix}}_{=:\boldsymbol{R}}$$

Dann heisst die Fixpunktiteration

$$\begin{aligned} (\boldsymbol{D}+\boldsymbol{L})\boldsymbol{x}^{(k+1)} &= -\boldsymbol{R}\boldsymbol{x}^{(k)} + \boldsymbol{b} \text{ bzw.} \\ \boldsymbol{x}^{(k+1)} &= -(\boldsymbol{D}+\boldsymbol{L})^{-1}\boldsymbol{R}\boldsymbol{x}^{(k)} + (\boldsymbol{D}+\boldsymbol{L})^{-1}\boldsymbol{b} \end{aligned}$$

Einzelschrittverfahren oder **Gauss-Seidel-Verfahren**.

Definition 4.8: anziehender / abstossender Fixpunkt [1]

- Gegeben sei eine Fixpunktiteration

$$\boldsymbol{x}^{(n+1)} = \boldsymbol{B}\boldsymbol{x}^{(n)} + \boldsymbol{b} =: F(\boldsymbol{x}^{(n)})$$

wobei $\boldsymbol{B}$ eine $n \times n$ Matrix ist und $\boldsymbol{b} \in \mathbb{R}^n$.Weiter sei $\| \cdot \|$eine der in Kap. 4.6.1 eingeführten Normen und $\overline{\boldsymbol{x}} \in \mathbb{R}^n$ erfülle $\overline{\boldsymbol{x}} = \boldsymbol{B}\overline{\boldsymbol{x}} + \boldsymbol{b} = F(\overline{\boldsymbol{x}})$. Dann heisst

  - $\overline{\boldsymbol{x}}$ anziehender Fixpunkt, falls $\| \boldsymbol{B} \| < 1$ gilt
  - $\overline{\boldsymbol{x}}$ abstossender Fixpunkt, falls $\| \boldsymbol{B} \| > 1$ gilt.

Satz 4.5: Abschätzungen [1]

- Gegeben sei wie in obiger Definition eine Fixpunktiteration

$$\boldsymbol{x}^{(n+1)} = \boldsymbol{B}\boldsymbol{x}^{(n)} + \boldsymbol{b} =: F(\boldsymbol{x}^{(n)})$$

und $\overline{\boldsymbol{x}} \in \mathbb{R}^n$ sei ein bezüglich der Norm $\| \cdot \|$ anziehender Fixpunkt. Dann konvergiert die Fixpunktiteration für alle Startvektoren $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$ gegen $\bar{\boldsymbol{x}}$ und es gelten die Abschätzungen

$$\begin{aligned} \| \boldsymbol{x}^{(n)} - \bar{\boldsymbol{x}} \| &\leq \frac{\| \boldsymbol{B} \|^n}{1 - \| \boldsymbol{B} \|} \| \boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)} \| \quad \text{a-priori Abschätzung} \\ \| \boldsymbol{x}^{(n)} - \bar{\boldsymbol{x}} \| &\leq \frac{\| \boldsymbol{B} \|}{1 - \| \boldsymbol{B} \|} \| \boldsymbol{x}^{(n)} - \boldsymbol{x}^{(n-1)} \| \quad \text{a-posteriori Abschätzung} \end{aligned}$$

Bemerkung: Der Vergleich mit den Definitionen für das Gesamt- und Einzelschrittverfahren liefert die Matrix $\boldsymbol{B}$:

- für das Gesamtschrittverfahren (Jacobi) ist

$$\boldsymbol{B} = -\boldsymbol{D}^{-1}(\boldsymbol{L}+\boldsymbol{R}),$$

- für das Einzelschrittverfahren (Gauss-Seidel) ist

$$\boldsymbol{B} = -(\boldsymbol{D}+\boldsymbol{L})^{-1}\boldsymbol{R}.$$

Definition 4.8: Diagonaldominanz [1]

- $\boldsymbol{A}$ ist eine **diagonaldominante Matrix**, falls eines der beiden folgenden Kriterien gilt:
  - für alle $i = 1, ..., n$: $\mid a_{ii} \mid > \sum_{j=1, j \neq i}^{n} \mid a_{i,j} \mid$ (Zeilensummenkriterium)
  - für alle $j = 1, ..., n$: $\mid a_{jj} \mid > \sum_{i=1, i \neq j}^{n} \mid a_{i,j} \mid$ (Spaltensummenkriterium)

Satz 4.6: Konvergenz [1]

- **Falls $\boldsymbol{A}$ diagonaldominant ist, konvergiert das Gesamtschrittverfahren (Jacobi) und auch das Einzelschrittverfahren (Gauss-Seidel) für $\boldsymbol{Ax = b}$.**

Bemerkungen:

- Die Bedingung $\parallel \boldsymbol{B} \parallel < 1$ für einen anziehenden Fixpunkt $\overline{\boldsymbol{x}}$ impliziert, dass $\boldsymbol{A}$ diagonaldominant ist.
- Diagonaldominanz ist nur ein hinreichendes Kriterium. Es gibt durchaus nicht diagonaldominante Matrizen, für die die Verfahren trotzdem konvergieren kann. Ein notwendiges und hinreichendes Kriterium ist, dass der Spektralradius $\rho(\boldsymbol{B}) < 1$.

Definition 5.1: Funktionen mit mehreren Variablen

- Unter einer Funktion mit n unabhängigen Variablen $x_1, ..., x_n$ und einer abhängigen Variablen y versteht man eine Vorschrift, die jedem geordneten Zahlentupel $(x_1, x_2, ..., x_n)$ aus einer Definitionsmenge $D \subset \mathbb{R}^n$ genau ein Element y aus einer Wertemenge $W \subset \mathbb{R}$ zuordnet. Symbolische Schreibweise:

$$\begin{array}{rcl} f: \ D \subset \mathbb{R}^n & \longrightarrow & W \subset \mathbb{R} \\ (x_1, x_2, ..., x_n) & \mapsto & y = f(x_1, x_2, ..., x_n) \end{array}$$

- Da das Ergebnis $y \in \mathbb{R}$ ein Skalar (eine Zahl) ist, redet man auch von einer **skalarwertigen** Funktion.

Bemerkungen:

- Die obige Definition lässt sich einfach erweitern auf beliebige **vektorwertige** Funktionen, die nicht einen Skalar, sondern einen Vektor als Wert zurückgeben:

$$\boldsymbol{f}: \mathbb{R}^n \longrightarrow \mathbb{R}^m,$$

mit

$$\boldsymbol{f}(x_1, ..., x_n) = \begin{pmatrix} y_1 = f_1(x_1, x_2, ..., x_n) \\ y_2 = f_2(x_1, x_2, ..., x_n) \\ \vdots \\ y_m = f_m(x_1, x_2, ..., x_n) \end{pmatrix},$$

wobei die m Komponenten $f_i : \mathbb{R}^n \longrightarrow \mathbb{R}$ $(i = 1, ..., m)$ von $\boldsymbol{f}$ wieder skalarwertige Funktionen sind, entsprechend Def. 5.1.

- Wie bei einem Vektor $\boldsymbol{x}$ stellen wir zur besseren Unterscheidbarkeit vektorwertige Funktionen $\boldsymbol{f}$ fett gedruckt dar, im Gegensatz zu einem Skalar x und einer skalarwertigen Funktion f.
- Wir werden uns bei der Lösung nichtlinearer Gleichungssysteme auf vektorwertige Funktionen $\boldsymbol{f} : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ konzentrieren.

Definition 5.2 [8]: Partielle Ableitungen 1. Ordnung
Unter den partiellen Ableitungen 1. Ordnung einer Funktion $z = f(x, y)$ and der Stelle (x, y) werden die folgenden Grenzwerte verstanden (falls sie vorhanden sind):

- Partielle Ableitung 1. Ordnung nach x:

$$\frac{\partial f}{\partial x}(x, y) = \lim_{\Delta x \to 0} \frac{f(x + \Delta x, y) - f(x, y)}{\Delta x}$$

- Partielle Ableitung 1. Ordnung nach y:

$$\frac{\partial f}{\partial y}(x, y) = \lim_{\Delta y \to 0} \frac{f(x, y + \Delta y) - f(x, y)}{\Delta y}$$

Definition 5.3: Jacobi-Matrix / Linearisierung / Tangentialebene

- Sei $\boldsymbol{f} : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ mit $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}) = \begin{pmatrix} y_1 = f_1(\boldsymbol{x}) \\ y_2 = f_2(\boldsymbol{x}) \\ \vdots \\ y_m = f_m(\boldsymbol{x}) \end{pmatrix}$ und $\boldsymbol{x} = (x_1, x_2, ..., x_n)^T \in \mathbb{R}^n$. Die **Jacobi-Matrix** enthält sämtliche partiellen Ableitung 1. Ordnung von $\boldsymbol{f}$ und ist definiert als

$$\boldsymbol{Df}(\boldsymbol{x}) := \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(x) & \frac{\partial f_1}{\partial x_2}(x) & \cdots & \frac{\partial f_1}{\partial x_n}(x) \\ \frac{\partial f_2}{\partial x_1}(x) & \frac{\partial f_2}{\partial x_2}(x) & \cdots & \frac{\partial f_2}{\partial x_n}(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(x) & \frac{\partial f_m}{\partial x_2}(x) & \cdots & \frac{\partial f_m}{\partial x_n}(x) \end{pmatrix}$$

- Die "verallgemeinerte Tangentengleichung"

$$\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x}^{(0)}) + \boldsymbol{Df}(\boldsymbol{x}^{(0)}) \cdot (\boldsymbol{x} - \boldsymbol{x}^{(0)})$$

beschreibt eine lineare Funktion und es gilt $\boldsymbol{f}(\boldsymbol{x}) \approx \boldsymbol{g}(\boldsymbol{x})$ in einer Umgebung eines gegebenen Vektors $\boldsymbol{x}^{(0)} = (x_1^{(0)}, x_2^{(0)}, ..., x_n^{(0)})^T \in \mathbb{R}^n$. Man spricht deshalb auch von der **Linearisierung** der Funktion $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$ in einer Umgebung von $\boldsymbol{x}^{(0)}$ (ein hochgestellter Index in Klammern $\boldsymbol{x}^{(k)}$ bezeichnet wie bisher einen Vektor aus $\mathbb{R}^n$ nach der $k-$ten Iteration).

- Für den speziellen Fall $f : \mathbb{R}^2 \longrightarrow \mathbb{R}$ mit $y = f(x_1, x_2)$ liefert die linearisierte Funktion

$$g(x_1, x_2) = f(x_1^{(0)}, x_2^{(0)}) + \frac{\partial f}{\partial x_1}(x_1^{(0)}, x_2^{(0)}) \cdot \begin{pmatrix} x_1 - x_1^{(0)} \end{pmatrix} + \frac{\partial f}{\partial x_2}(x_1^{(0)}, x_2^{(0)}) \cdot \begin{pmatrix} x_2 - x_2^{(0)} \end{pmatrix}$$

die Gleichung der **Tangentialebene**. Sie enthält sämtliche im Flächenpunkt $P = (x_1^{(0)}, x_2^{(0)}, f(x_1^{(0)}, x_2^{(0)}))$ an die Bildfläche von $y = f(x_1, x_2)$ angelegten Tangenten.

Newton-Verfahren für Systeme [1]:
Gesucht sind Nullstellen von $\boldsymbol{f} : \mathbb{R}^n \longrightarrow \mathbb{R}^n$. Sei $\boldsymbol{x}^{(0)}$ ein Startvektor in der Nähe einer Nullstelle. Das Newton-Verfahren zur näherungsweisen Bestimmung dieser Nullstelle lautet:

- für $n = 0, 1, ...$:
  - Berechne $\boldsymbol{\delta}^{(n)}$ als Lösung des linearen Gleichungssystems

$$\boldsymbol{Df}(\boldsymbol{x}^{(n)})\boldsymbol{\delta}^{(n)} = -\boldsymbol{f}(\boldsymbol{x}^{(n)})$$

  - Setze

$$\boldsymbol{x}^{(n+1)} := \boldsymbol{x}^{(n)} + \boldsymbol{\delta}^{(n)}$$

Vereinfachtes Newton-Verfahren für Systeme [1]:
Gesucht sind Nullstellen von $\boldsymbol{f} : \mathbb{R}^n \longrightarrow \mathbb{R}^n$. Sei $\boldsymbol{x}^{(0)}$ ein Startvektor in der Nähe einer Nullstelle. Das vereinfachte Newton-Verfahren zur näherungsweisen Bestimmung dieser Nullstelle lautet:

- für $n = 0, 1, ...$:
  - Berechne $\boldsymbol{\delta}^{(n)}$ als Lösung des linearen Gleichungssystems

$$\boldsymbol{Df}(\boldsymbol{x}^{(0)})\boldsymbol{\delta}^{(n)} = -\boldsymbol{f}(\boldsymbol{x}^{(n)})$$

  - Setze

$$\boldsymbol{x}^{(n+1)} := \boldsymbol{x}^{(n)} + \boldsymbol{\delta}^{(n)}$$

Gedämpftes Newton-Verfahren für Systeme [6]:
Gesucht sind Nullstellen von $\boldsymbol{f} : \mathbb{R}^n \longrightarrow \mathbb{R}^n$. Sei $\boldsymbol{x}^{(0)}$ ein Startvektor in der Nähe einer Nullstelle, $k_{max} \in \mathbb{N}$ sei vorgegeben. Das gedämpfte Newton-Verfahren zur näherungsweisen Bestimmung dieser Nullstelle lautet:

- für $n = 0, 1, ...$:
  - Berechne $\boldsymbol{\delta}^{(n)}$ als Lösung des linearen Gleichungssystems

$$\boldsymbol{Df}(\boldsymbol{x}^{(n)})\boldsymbol{\delta}^{(n)} = -\boldsymbol{f}(\boldsymbol{x}^{(n)})$$

  - Finde das minimale $k \in \{0, 1, ..., k_{max}\}$ mit

$$\| \boldsymbol{f}\left(\boldsymbol{x}^{(n)} + \frac{\boldsymbol{\delta}^{(n)}}{2^k}\right) \|_2 < \| \boldsymbol{f}\left(\boldsymbol{x}^{(n)}\right) \|_2$$

  - Falls kein minimales k gefunden werden kann, rechne mit $k = 0$ weiter
  - Setze

$$\boldsymbol{x}^{(n+1)} := \boldsymbol{x}^{(n)} + \frac{\boldsymbol{\delta}^{(n)}}{2^k}$$