Angewandte Mathematik für die Informatik

Sommersemester 2021, Teil 2

Louchka Popova-Zeugmann und Wolfgang Kössler

5. Juli 2021

Aufgabenstellung der Numerik

- Wie wirken sich Eingabe- und Rundungsfehler aus? gerundete, diskretisierte Daten, fehlerbehaftete Daten, Abbruchfehler
- Lösung von Aufgaben, die analytisch (z.B. mit Integralrechnung) nicht möglich ist oder nur mit relativ großem Aufwand. Bereitstellung von Methoden zur (näherungsweisen) Berechnung (mit dem Computer) hier: Approximation, Integration, Nichtlineare Gleichungssysteme
- Berechnung möglichst effizient

Literatur

Stoer oder Stoer/Bulirsch: Numerische Mathematik 1, Springer
Opfer, G.: Numerische Mathematik für Anfänger, Vieweg

Schaback, W., Wendland, H. Numerische Mathematik, Springer, 2005.
Hartmann, P. Mathematik für Informatiker, Vieweg, 2004.

Weitere Literatur in der Bibliothek unter SK900.

1 Fehleranalyse

1.1 Gleitkommazahlen

Um mit reellen Zahlen auf dem Computer zu arbeiten, gibt es die möglichen Darstellungen:

- symbolisch (z.B. in Computeralgebrasystemen wie Mathematica oder Maple)
- approximativ:
  - Festkommazahlen
  - Gleitkommazahlen

Wir werden uns in dieser Vorlesung ausschließlich mit Gleitkommazahlen beschäftigen.
Sei $b \in \mathbb{N}$, $b \geq 2$ eine Basis für die Zahlendarstellung (z.B. $b = 2$: Binärdarstelluung, $b = 10$: Dezimaldarstellung). Dann lässt sich jede reelle Zahl $x \in \mathbb{R}$ darstellen als

$$x \;=\; (-1)^v \cdot b^e \underbrace{\sum_{i=1}^{\infty} a_i b^{-i}}_{\in[0,1)}, \quad \text{wobei}$$

$v \in \{0,1\}$ (Vorzeichen-Bit)

$e \in \mathbb{Z}$ (Exponent)

$a_i \in \{0,1,...,b-1\}$ (Ziffern).

Beispiel 1: $b = 10$

$$\begin{aligned} -123.4 &= (-1)^1 \cdot 10^3 \cdot 0.1234 \\ &= (-1)^1 \cdot 10^3 (1 \cdot 10^{-1} + 2 \cdot 10^{-2} + 3 \cdot 10^{-3} + 4 \cdot 10^{-4}) \end{aligned}$$

Def. 1 (Menge der Gleitkommazahlen) $\mathbb{G}(b,l,E)$

Die Menge aller **Gleitkommazahlen** mit Basis $b \in \mathbb{N}$, $b \geq 2$, Mantissenlänge $l \in \mathbb{N}$, $l \geq 1$ und Exponent $e \in E \subseteq \mathbb{Z}$ ist:

$$\{x = (-1)^v \cdot b^e \sum_{i=1}^{l} a_i b^{-i} \mid v \in \{0,1\},\ e \in E,\ a_i \in \{0,...,b-1\}\ \forall i \leq l\}.$$

Eine Darstellung von $\mathbb{G}(b,l,E) \ni x = (-1)^v \cdot b^e \underbrace{\sum_{i=1}^{l} a_i b^{-i}}_{\in [\frac{1}{b},1)}$ heißt **normalisiert**, falls $a_1 \neq 0$.

Beispiel 2: $b = 10$ *nicht normalisiert-normalisiert*

$$0.0123 = 10^{-1} \cdot \underbrace{0.123}_{\in [\frac{1}{10},1)}$$

IEEE Gleitkommazahlen

		Mantissenlänge	Exponent	Vorzeichen	$\sum$
single	real*4	l=23 bit	e=8 bit	1 bit	32 bit
double	real*8	l=52 bit	e=11 bit	1 bit	64 bit

Die Darstellung $x = \pm m 2^{c-1022}$ spart das Vorzeichenbit für den Exponenten.

$$\begin{aligned} m &= 2^{-1} + a_2 2^{-2} + \cdots + a_{53} 2^{-53}, \quad a_i, c_i \in \{0,1\} \\ c &= c_0 2^0 + \cdots + c_{10} 2^{10} \in [1, 2046] \cap \mathbb{Z} \end{aligned}$$

$c = 0$ wird für $x = 0$, und $c = 2047$ für $x = NAN$ (not a number) verwendet.

Beispiel 3: $b = 2$, $x = \frac{3}{8}$

$a_1 = 0, a_2 = a_3 = 1 : x = 0 \cdot 2^{-1} + 1 \cdot 2^{-2} + 1 \cdot 2^{-3}$ nicht normalisiert

$a_1' = 1, a_2' = 1 :$ $x = 2^{-1}(1 \cdot 2^{-1} + 1 \cdot 2^{-2})$ normalisiert

$$(\frac{3}{8})_{10} = (\underbrace{0}_{V}\ \underbrace{01111111101}_{\text{Exponent}=-1}\ \underbrace{110\ldots0}_{\text{Mantisse}})_2$$

$c_0 = 1, c_1 = 0, c_2 = \cdots = c_9 = 1, c_{10} = 0, \quad c = 1021 = 1023 - 2$

Beispiel 4: $b = 2$, $x = 0.1 = \frac{1}{10}$

$$\begin{aligned} \frac{1}{10} &= \sum_{i=1}^{l} a_i \frac{1}{2^i} \quad \Leftrightarrow \quad 2^l = 10 \cdot \sum_{i=1}^{l} a_i 2^{l-i} \quad \text{keine Lösung für die } a_i \\ &\approx \underbrace{\underbrace{\frac{1}{16} + \frac{1}{32} + \frac{1}{256} + \frac{1}{512}}_{=0.0996094} + \frac{1}{2048}}_{=1.00098} = 2^{-3}(\frac{1}{2^1} + \frac{1}{2^2} + \frac{1}{2^5} + \frac{1}{2^6} + \frac{1}{2^8}) \end{aligned}$$

Zahlbereich (bis auf Vorzeichen): $|x| \in [2^{-1022}, (1 - 2^{-53})2^{1024}] \cup \{0\}$

- kleinste positive real*8 Gleitkommazahl: $a_1 = 1, a_i = 0 \quad \forall i = 2, \ldots, 53, c = 1$: $x = 2^{-1} \cdot 2^{1-1022} = 2^{-1022}$.
- größte real*8 Gleitkommazahl: $a_i = 1 \quad \forall i = 1, \ldots, 53,\ c = 2046$:

$$\begin{aligned} x &= \sum_{i=1}^{53} \frac{1}{2^i} \cdot 2^{2046-1022} = \Big(\frac{1 - \frac{1}{2^{54}}}{1 - \frac{1}{2}} - 1\Big)2^{1024} \\ &= (1 - 2^{-53})2^{1024}. \end{aligned}$$

Bem.: $x, y \in \mathbb{G}(b, l, E) \quad \not\Longrightarrow \quad x + y \in \mathbb{G}(b, l, E)$.

Beispiel 5: $b = 10,\ l = 3,\ E = \{0, ..., 3\},\ x = 500 = (-1)^0 \cdot 10^3 \cdot 0.500$.
Dann ist $x + x = 1000 = (-1)^0 \cdot 10^4 \cdot 0.100$. Aber $4 \notin E \Rightarrow$ Exponentenüberlauf!
(Aber auch ohne Exponentenüberlauf gilt dies nicht! siehe Übung)

Def. 2: Überlauf, Unterlauf
Ein <u>Überlauf</u> liegt vor wenn der Betrag der Gleitkommaoperation zu einem Wert führt, der größer als die größte Gleitkommazahl ist.
Ein <u>Unterlauf</u> liegt vor wenn der Betrag der Gleitkommaoperation zu einem Wert führt, der kleiner als die kleinste positive Gleitkommazahl ist.

Beispiel 6: Berechnung von $\sqrt{a^2 + b^2}$
naive Programmierung: es kann passieren, dass der Zahlbereich des Rechners überschritten oder unterschritten wird,
Zwischenergebnis größer als die maximale Gleitkommazahl oder
kleiner als die minimale positive Gleitkommazahl.
Sei $a \geq b > 0$. Dann berechnen wir besser

$$|a|\sqrt{1 + \big(\frac{b}{a}\big)^2}$$

1.2 Rundung

Wir fixieren jetzt b und l und betrachten im Folgenden nur $\mathbb{G} := \mathbb{G}(b, l, \mathbb{Z})$
(d.h. wir schließen einen Über-/Unterlauf des Exponenten aus).
Wir stellen eine beliebige relle Zahl $x \in \mathbb{R}$ durch die Gleitkommazahl $\mathrm{rd}(x) \in \mathbb{G}$ dar, die x am nächsten liegt, d.h.

$$|x - \mathrm{rd}(x)| \ \leq \ |x - g| \quad \forall g \in \mathbb{G}.$$

Aber $\mathrm{rd}(x)$ ist so i.A. noch nicht eindeutig bestimmt:
Sind $g_1, g_2 \in \mathbb{G}$ zwei „aufeinanderfolgende" Gleitkommazahlen (d.h. $[g_1, g_2] \cap \mathbb{G} = \{g_1, g_2\}$) und $x = (g_1 + g_2)/2$, so erfüllen g_1 <u>und</u> g_2 diese Bedingung. Definiere dann $\mathrm{rd}(x) := g_2$ (die größere von beiden).
Sind $g_1 < g_2 < g_3$ drei aufeinanderfolgende Gleitkommazahlen (d.h. $[g_1, g_3] \cap \mathbb{G} = \{g_1, g_2, g_3\}$), so wird genau das Intervall $I := [\frac{g_1+g_2}{2}, \frac{g_2+g_3}{2})$ durch rd auf g_2 abgebildet. Also

$$\mathrm{rd}(x) = g_2 \quad \Longrightarrow \quad x \in I.$$

Durch die Gleitkomma-Darstellung identifizieren wir somit alle Zahlen aus $I \subset \mathbb{R}$ mit $g_2 \in \mathbb{G}$. Die Menge $\mathbb{G}$ ist <u>nicht</u> abgeschlossen unter Addition (siehe Übung):

$$x, y \in \mathbb{G} \quad \not\Rightarrow \quad x + y \in \mathbb{G}.$$

Um Gleitkommazahlen zu addieren, müssen wir also $+ : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ approximieren durch eine Operation $\boxed{+} : \mathbb{G} \times \mathbb{G} \to \mathbb{G}$ (oder zumindest $\boxed{+} : \mathbb{R} \times \mathbb{R} \to \mathbb{G}$). Analoges gilt für andere Funktionen wie $-$, $\cdot$, $:$, sin, exp.

1.3 Gleitkommaoperationen

Gleitkommaoperationen

Die natürliche Approximation einer Funktion $f : X \to \mathbb{R}$ (X beliebige Menge) ist

$$\boxed{\mathrm{f}} : X \to \mathbb{G} \quad (\text{z.B. } x_1 \boxed{+} x_2 := \mathrm{rd}(x_1 + x_2))$$
$$\boxed{\mathrm{f}} := \mathrm{rd} \circ f \qquad (\text{d.h. } \boxed{\mathrm{f}}(x) := \mathrm{rd}(f(x))).$$

Vorsicht: Viele Gesetze wie Assoziativ- oder Distributivgesetze gelten für die Gleitkommaoperationen nicht mehr, z.B.

$$\Big(a \boxed{+} b\Big) \boxed{+} c \neq_{\text{i.A.}} a \boxed{+} \Big(b \boxed{+} c\Big).$$

(siehe Übung).

1.4 Absolute und relative Fehler

Erinnerung: Norm im $\mathbb{R}^n$

Die Abbildung $||.|| : \mathbb{R}^n \to [0, \infty)$ heißt eine Norm, falls für alle $\mathbf{x} \in R^n$ und für alle $\alpha \in \mathbb{R}$ gilt:

1. $||\mathbf{x}|| = 0$ genau dann, wenn $\mathbf{x} = \mathbf{0}$
2. $||\alpha \mathbf{x}|| = |\alpha| \cdot ||\mathbf{x}||$ und
3. $||\mathbf{x} + \mathbf{y}|| \leq ||\mathbf{x}|| + ||\mathbf{y}||$ (Dreiecksungleichung)

$$\begin{aligned}
||\mathbf{x}||_2 &= \sqrt{\sum_{i=1}^{n} x_i^2} && \text{Euklidische Norm} \\
||\mathbf{x}||_1 &= \sum_{i=1}^{n} |x_i| && \text{City Block Norm (Manhattan Norm)} \\
||\mathbf{x}||_\infty &= \max(|x_1, \ldots, |x_n|) && \text{Maximumnorm}
\end{aligned}$$

Die Norm von $\mathbf{x}$ ist der Abstand von $\mathbf{x}$ zum Nullpunkt.

Def. 3 (Absolute und relative Fehler)

Ist $\tilde{x}$ eine Approximation von x (für $x, \tilde{x} \in \mathbb{R}^n$ oder $x, \tilde{x} \in X$ mit $(X, ||\,.\,||)$ normierter Vektorraum), so heißt

$$||x - \tilde{x}|| \qquad \text{der } \textbf{absolute Fehler} \text{ von } \tilde{x} \text{ bzgl. } x,$$
$$\frac{||x - \tilde{x}||}{||x||} \qquad \text{der } \textbf{relative Fehler} \text{ von } \tilde{x} \text{ bzgl. } x \text{ (falls } x \neq 0).$$

Im Gegensatz zum relativen Fehler gibt der absolute Fehler keinen Aufschluss über die Anzahl der gültigen Stellen von $\tilde{x}$.

Beispiel 7:

1. $x = 1, \tilde{x} = 2. \Rightarrow \quad |x - \tilde{x}| = 1, \quad \frac{|x-\tilde{x}|}{|x|} = 1 = 10^0$ (Fehler in der 1. Stelle)
2. $x = 1000, \tilde{x} = 1001. \Rightarrow \quad |x - \tilde{x}| = 1, \quad \frac{|x-\tilde{x}|}{|x|} = 10^{-3}$ (Fehler in der 4. Stelle)

Satz 1

Für $x \in \mathbb{R}$, $x \neq 0$ gilt:

$$\frac{|x - \mathrm{rd}(x)|}{|x|} \leq \frac{b^{-l+1}}{2} =: \mathrm{eps}(b, l) =: \mathrm{eps} \quad (= \text{Maschinengenauigkeit}).$$

Beweis: OBdA. gelte $x > 0$. Wähle $g_1, g_2 \in \mathbb{G}$ aufeinanderfolgend mit $g_1 < x \leq g_2$ und

$$g_1 = b^e \sum_{i=1}^{l} a_i b^{-i}, \qquad a_1 \neq 0.$$

Dann ist $g_2 - g_1 = b^e \cdot b^{-l} = b^{e-l}$. Wegen $g_1 \geq b^e \cdot b^{-1}$ folgt

$$\frac{|x - \mathrm{rd}(x)|}{|x|} \leq \frac{\frac{g_2 - g_1}{2}}{g_1} \leq \frac{b^{e-l}}{2b^e \cdot b^{-1}} = \frac{b^{1-l}}{2}. \qquad \square$$

Bem.: Für IEEE-Gleitkommazahlen ist
eps $= 2^{-53} \approx 1,1 \cdot 10^{-16}$ für „double" (64 Bit-Gleitkommazahlen) und eps $= 2^{-24} \approx 6,0 \cdot 10^{-8}$ für „single" (32 Bit-Gleitkommazahlen).

Folgerung
Für alle $x \in \mathbb{R}$ existiert ein $\varepsilon \in \mathbb{R}$, $|\varepsilon| \leq$ eps mit

$$\mathrm{rd}(x) \;=\; (1+\varepsilon)x.$$

Beweis: Für $x \neq 0$ nimm $\varepsilon := -\frac{x - \mathrm{rd}(x)}{x}$. $\Rightarrow$ $\mathrm{rd}(x) = (1+\varepsilon)x$, $\Rightarrow_{\text{Satz 1}}$ $|\varepsilon| \leq$ eps. $\square$

1.5 Kondition von Abbildungen

Matrixnorm und Kondition

Def. 4 (Matrix-Norm)
Sei $\mathbf{A}$ eine $(m \times n)$ Matrix und $||.||$ eine beliebige Norm in $\mathbb{R}^n$. Dann heißt

$$||\mathbf{A}|| := \sup_{\mathbf{x} \neq \mathbf{0}} \frac{||\mathbf{A}\mathbf{x}||}{||\mathbf{x}||} = \sup_{||x||=1} ||\mathbf{A}\mathbf{x}||$$

(zugehörige) Matrixnorm.

z.z. 1. Supremum existiert, 2. $||.||$ ist eine Norm. (siehe Übung)
zweite Gleichung: $\frac{||\mathbf{A}\mathbf{x}||}{||\mathbf{x}||} = ||\mathbf{A}\frac{\mathbf{x}}{||\mathbf{x}||}|| = ||\mathbf{A}\mathbf{y}||$, wobei $y = \frac{\mathbf{x}}{||\mathbf{x}||}$, $||\mathbf{y}|| = 1$.
Die Menge $\{||\mathbf{x} = 1||\}$ ist kompakt und die lineare Abbildung $\mathbf{A}\mathbf{x}$ ist stetig, das Supremum existiert also.

Bem.: Es gibt andere Varianten der Matrixnormdefinition.
Matrixnormen, $||.||_k$ gehört zur entsprechenden Norm im $\mathbb{R}^n$

$$\begin{aligned}
||\mathbf{A}||_1 &= \max_j \sum_{i=1}^{n} |a_{ij}| && \text{Spaltensummennorm} \\
||\mathbf{A}||_\infty &= \max_i \sum_{j=1}^{n} |a_{ij}| && \text{Zeilensummennorm} \quad \text{(vgl. Satz 8)} \\
||\mathbf{A}||_2 &= \sqrt{\lambda_{max}(\mathbf{A}^T\mathbf{A})} && \text{Spektralnorm}, \quad \lambda_{max}: \text{ max. Eigenwert}
\end{aligned}$$

Wir verwenden hier nur die Spaltensummennorm, $k = 1$

$$\begin{aligned}
||\mathbf{A}||_1 &= \sup_{\mathbf{x} \neq \mathbf{0}} \frac{||\mathbf{A}\mathbf{x}||_1}{||\mathbf{x}||_1} = \sup_{\mathbf{x} \neq \mathbf{0}} \frac{1}{||\mathbf{x}||_1} \sum_{i=1}^{n} \Big| \sum_{j=1}^{m} a_{ij} x_j \Big| \\
&\leq \sup_{\mathbf{x} \neq \mathbf{0}} \frac{1}{||\mathbf{x}||_1} \sum_{i=1}^{n} \sum_{j=1}^{m} |a_{ij} x_j| \leq \sup_{\mathbf{x} \neq \mathbf{0}} \frac{1}{||\mathbf{x}||_1} \underbrace{\sum_{j=1}^{m} |x_j|}_{||\mathbf{x}||_1} \sum_{i=1}^{n} |a_{ij}| \\
&\leq \max_j \sum_{i=1}^{n} |a_{ij}|.
\end{aligned}$$

Sei $\mathbf{A} = (a_{ij})_{i=1,\dots,n,j=1,\dots,m}$ und $\mathbf{x} = (x_1, \dots, x_n)^T$.
Es gilt für die lineare Abbildung $\mathbf{Ax}$:

$$\mathbf{Ax} = \big(\sum_{j=1}^{m} a_{1j}x_j, \dots, \sum_{j=1}^{m} a_{mj}x_j\big)^T$$

$$\begin{aligned}
||\mathbf{Ax}||_1 &= \sum_{i=1}^{n} \big|\sum_{j=1}^{m} a_{ij}x_j\big| \leq \sum_{i=1}^{n}\sum_{j=1}^{m} |a_{ij}||x_j| \\
&= \underbrace{\sum_{j=1}^{m} |x_j|}_{||\mathbf{x}||_1} \sum_{i=1}^{n} |a_{ij}| \leq ||\mathbf{x}||_1 \max_j \sum_{i=1}^{n} |a_{ij}| \\
\frac{||\mathbf{Ax}||_1}{||\mathbf{x}||_1} &\leq \max_j \sum_{i=1}^{n} |a_{ij}|
\end{aligned}$$

Das Gleichheitszeichen gilt für $\mathbf{x} = (0, \dots, 0, 1, 0, \dots, 0)$, wobei die 1 an der Stelle j_0 steht, an der die letzte Summe maximal ist.
Wir wollen jetzt untersuchen, wie stark sich der absolute/relative Fehler unter Abbildungen ändert. Seien $X = \mathbb{R}^n$, $Y = \mathbb{R}^m$ mit fixierten Normen, $x \in X$, $f : X \to Y$, $x \neq 0$, $f(x) \neq 0$

Def. 5 (absolute und relative Kondition)

$$\begin{aligned}
\kappa_{\text{abs}} = \kappa_{\text{abs}}(x) &:= \overline{\lim_{\tilde{x}\to x}} \frac{||f(\tilde{x}) - f(x)||_Y}{||\tilde{x} - x||_X} = \overline{\lim_{\tilde{x}\to x}} \frac{\text{abs. Fehler von } f(\tilde{x})}{\text{abs. Fehler von } \tilde{x}} \\
\kappa_{\text{rel}} = \kappa_{\text{rel}}(x) &:= \overline{\lim_{\tilde{x}\to x}} \frac{\frac{||f(\tilde{x})-f(x)||_Y}{||f(x)||_Y}}{\frac{||\tilde{x}-x||_X}{||x||_X}} = \overline{\lim_{\tilde{x}\to x}} \frac{\text{rel. Fehler von } f(\tilde{x})}{\text{rel. Fehler von } \tilde{x}}
\end{aligned}$$

heißen absolute bzw. relative Kondition von f in x.
Eine Abbildung heißt in x **gut konditioniert**, falls ihre Kondition „klein" ist, andernfalls **schlecht konditioniert**.

Bem.: Diese Konditionen nennt man auch die **normweise Konditionen**. (Zur *komponentenweisen Kondition* siehe unten)

Bem.: Die Kondition hängt vom Argument x ab, wir werden jedoch meist das Argument weglassen. Auch den Index X bzw. Y werden wir im Folgenden weglassen.

Bem.: Zur Notation: Für $g : (X, ||\,.\,||) \to \mathbb{R}$ ist der limes superior in $x \in X$

$$\overline{\lim_{\tilde{x}\to x}} g(x) = \begin{cases} \text{Maximum/Supremum aller Grenzwerte } \lim g(x_n) \\ \text{konvergenter Teilfolgen } g(x_n) \text{ mit } x_n \to x\ , \\ \\ \infty \quad \text{falls keine konvergente Teilfolge } g(x_n) \text{ existiert.} \end{cases}$$

Kondition der linearen Abbildung

Beispiel 8: Kondition der linearen Abbildung $\mathbf{Ax}$

$$\begin{aligned}
\kappa_{\text{abs}} &= \overline{\lim_{\tilde{x}\to x}} \frac{||\mathbf{A}\tilde{\mathbf{x}} - \mathbf{Ax}||}{||\tilde{\mathbf{x}} - \mathbf{x}||} =_{(*)} \overline{\lim_{\tilde{x}\to x}} \frac{||\mathbf{A}||||\tilde{\mathbf{x}} - \mathbf{x}||}{||\tilde{\mathbf{x}} - \mathbf{x}||} = ||\mathbf{A}|| \\
\kappa_{\text{rel}} &= \frac{||\mathbf{x}||}{||\mathbf{Ax}||}||\mathbf{A}||. \qquad (*): \leq \text{ klar, } = \text{siehe Übung}
\end{aligned}$$

Beispiel 9: Kondition der Addition: $\mathbb{R}^2 \to \mathbb{R}$ $(x_1, x_2) \to x_1 + x_2 = \mathbf{Ax}$, *wobei* $\mathbf{A} = (1, 1), \mathbf{x} = (x_1, x_2)^T$

$$\begin{aligned}\kappa_{\text{abs}} &= ||(1,1)||_1 = 1 \quad \text{(Spaltensummennorm)}\\ \kappa_{\text{rel}} &= \frac{||\mathbf{x}||_1}{||\mathbf{Ax}||_1}||\mathbf{A}||_1 = \frac{|x_1|+|x_2|}{|x_1+x_2|}\end{aligned}$$

Haben x_1 und x_2 das gleiche Vorzeichen, so ist $\kappa_{\text{rel}} = 1$. Bei unterschiedlichen Vorzeichen aber kann κ_{rel} beliebig groß werden! In diesem Fall ist die Addition also schlecht konditioniert.

Beispiel 10: Seien $x_1 = 0.1234$, $\tilde{x}_1 = 0.1235$ eine Störung von x_1, $x_2 = -0.123$.
Außerdem sei $y := x_1 + x_2$ die Summe und $\tilde{y} := \tilde{x}_1 + x_2$ die gestörte Summe. Für die relativen Fehler gilt dann:

$$\begin{aligned}\frac{||\Delta\mathbf{x}||_1}{||\mathbf{x}||_1} = \frac{||(x_1,x_2)-(\tilde{x}_1,x_2)||_1}{||(x_1,x_2)||_1} &= \frac{0.0001}{0.2464} \approx 4.1\cdot 10^{-4},\\ \frac{||\Delta y||_1}{||y||_1} = \frac{|y-\tilde{y}|}{|y|} &= \frac{0.0001}{0.0004} = 2.5\cdot 10^{-1},\\ \kappa_{\text{rel}} = \frac{\text{relativer Fehler von } y}{\text{relativer Fehler von } \mathbf{x}} &= \frac{0.2464}{0.0004} = 616.\end{aligned}$$

Dieser Effekt des Verlustes an Genauigkeit heißt **Auslöschung**.

Faustregel: *Auslöschung durch Subtraktion annähernd gleicher Zahlen vermeiden!*

Beispiel 11: Subtraktion (etwa) gleich großer Zahlen $a - b$

$$a := Pi, \quad b := 100\frac{Pi}{100.}, \quad a - b = 0?$$

Mathematica gibt $a - b = -4.44089\cdot 10^{-16}$ ($\approx$ Maschinengenauigkeit) .

Schlussfolgerung: Gleitkommazahlen niemals auf Null testen.

Bem.: Bezüglich der normweisen Kondition kann auch die Multiplikation/Division schlecht konditioniert sein, bezüglich der komponentenweise Kondition ist sie aber immer gut konditioniert! (Übung)

Kondition einer differenzierbaren Abbildung

Sei $f : \mathbb{R}^n \to \mathbb{R}^m$:

$$f(x_1,\ldots,x_n) = (f_1(x_1,\ldots,x_n),\ldots,f_m(x_1,\ldots,x_n))^T$$

Def. 6 (partiell differenzierbare Abbildung)
f heißt partiell differenzierbar, falls alle $f_i(x_1,\ldots,x_n), i = 1,\ldots,m$ nach allen Komponenten x_j (partiell) differenzierbar ist.

Die Matrix $\mathbf{J}$ der partiellen Ableitungen, $f'(x) = \begin{pmatrix} \frac{\partial f_1(x)}{\partial x_1} & \cdots & \frac{\partial f_1(x)}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m(x)}{\partial x_1} & \cdots & \frac{\partial f_m(x)}{\partial x_n}\end{pmatrix} =: \mathbf{J}$ heißt Jacobi-Matrix. (Carl Gustav Jacob Jacobi, 1804-1851)

Satz 2 (Kondition einer differenzierbaren Abbildung)
Sei $f : X \to Y$ differenzierbar mit Jacobi-Matrix $\mathbf{J}$. Dann gilt

$$\kappa_{\text{abs}} = ||\mathbf{J}||, \qquad \kappa_{\text{rel}} = \frac{||x||}{||f(x)||}||\mathbf{J}||.$$

Beweis: Satz von Taylor für Funktionen mehrerer Variablen:

$$f(\mathbf{x}) - f(\mathbf{x}') = \mathbf{J}(\mathbf{x}-\mathbf{x}') + \rho(\mathbf{x},\mathbf{x}')||\mathbf{x}-\mathbf{x}'||, \quad \text{wobei } \lim_{\mathbf{x}'\to\mathbf{x}} \rho(\mathbf{x},\mathbf{x}') = \mathbf{0}$$

Normbildung, Dreiecksungleichung, Division durch $||\mathbf{x}-\mathbf{x}'||$ und Grenzwertbildung liefert $\kappa_{\text{abs}} \leq ||\mathbf{J}||$. Für $=$ nehmen wir eine Folge so dass $\frac{||\mathbf{J}(\mathbf{x}-\mathbf{x}')||}{||\mathbf{x}-\mathbf{x}'||} \to ||\mathbf{J}||$. □

Die Kondition von Abbildungen kann also leicht mit der Jacobi-Matrix bestimmt werden.

Bem.: Ist die Abbildung sogar zweimal stetig differenzierbar, so gilt nach dem Satz von Taylor für Funktionen von mehreren Variablen

$$f(\mathbf{x}) - f(\mathbf{x}') = \mathbf{J}(\mathbf{x}-\mathbf{x}') + \mathcal{O}(||\mathbf{x}-\mathbf{x}'||^2) = \underbrace{\left(\sum_{j=1}^{n} \frac{\partial f_i(\mathbf{x})}{\partial x_j}(x_j - x_j')\right)}_{i=1,\ldots,m} + \mathcal{O}(||\mathbf{x}-\mathbf{x}'||^2)$$

Satz 3: Für die komponentenweisen relativen Fehler gilt asymptotisch

$$\left|\frac{f_i(\mathbf{x}) - f_i(\mathbf{x}')}{f_i(\mathbf{x})}\right| \leq \sum_{j=1}^{n} \left|\frac{\partial f_i(\mathbf{x})}{\partial x_j}\right| \frac{|x_j - x_j'|}{|f_i(\mathbf{x})|} = \sum_{j=1}^{n} \underbrace{\left|\frac{\partial f_i(\mathbf{x})}{\partial x_j}\frac{x_j}{f_i(\mathbf{x})}\right|}_{\kappa_{ij}(x)=\kappa_{ij}} \left|\frac{\Delta x_j}{x_j}\right|$$

mit den **komponentenweisen Konditionen**

$$\kappa_{ij}(x) := \left|\frac{\partial f_i(\mathbf{x})}{\partial x_j}\right| \frac{|x_j|}{|f_i(\mathbf{x})|}$$

Die Kondition linearer Gleichungssysteme

Kondition der Abbildung $\mathbf{x} = \mathbf{A}^{-1}\mathbf{y}$*, vgl. auch Beispiel 8*
d.h. Lösen des linearen Gleichungssystems mit *fester* Matrix $\mathbf{A}$

$$\begin{aligned}
\kappa_{\text{abs}} &= ||\mathbf{A}^{-1}|| \\
\kappa_{\text{rel}} &= \frac{||\mathbf{y}||}{||\mathbf{A}^{-1}\mathbf{y}||}||\mathbf{A}^{-1}|| = \frac{||\mathbf{A}\mathbf{x}||}{||\mathbf{x}||}||\mathbf{A}^{-1}|| \leq \frac{||\mathbf{A}|| \cdot ||\mathbf{x}||}{||\mathbf{x}||}||\mathbf{A}^{-1}|| \\
&= ||\mathbf{A}|| \cdot ||\mathbf{A}^{-1}|| = \text{cond}(A)
\end{aligned}$$

Def. 7 (Kondition von Matrizen)
Sei $\mathbf{A}$ eine reguläre Matrix. Der Term

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

heißt Kondition der Matrix $\mathbf{A}$.

Wir betrachten jetzt die linearen Gleichungssysteme

$$\mathbf{A}\mathbf{x} = \mathbf{y} \qquad \text{und} \qquad \tilde{\mathbf{A}}\tilde{\mathbf{x}} = \tilde{\mathbf{y}} \quad \text{(gestörtes System)}$$

mit $\mathbf{A}, \tilde{\mathbf{A}} \in \mathbb{R}^{n\times n}$, $\mathbf{x}, \tilde{\mathbf{x}}, \mathbf{y}, \tilde{\mathbf{y}} \in \mathbb{R}^n$.
Setze $\Delta\mathbf{A} := \tilde{\mathbf{A}} - \mathbf{A}$, $\Delta\mathbf{x} := \tilde{\mathbf{x}} - \mathbf{x}$ und $\Delta\mathbf{y} := \tilde{\mathbf{y}} - \mathbf{y}$.
Wir wollen jetzt den absoluten und relativen Fehler ($||\Delta\mathbf{x}||$ bzw. $\frac{||\Delta\mathbf{x}||}{||\mathbf{x}||}$) der Lösung in Abhängigkeit von $\Delta\mathbf{A}$ und $\Delta\mathbf{y}$ ermitteln.

Satz 4 (zur Information)
Ist $\mathbf{A}$ invertierbar und $||\Delta\mathbf{A}|| \cdot ||\mathbf{A}^{-1}|| < 1$, so ist auch $\tilde{\mathbf{A}}$ invertierbar und es gilt

$$\frac{||\Delta\mathbf{x}||}{||\mathbf{x}||} \leq \frac{\text{cond}(\mathbf{A})}{1 - \text{cond}(\mathbf{A}) \cdot \frac{||\Delta\mathbf{A}||}{||\mathbf{A}||}} \left(\frac{||\Delta\mathbf{y}||}{||\mathbf{y}||} + \frac{||\Delta\mathbf{A}||}{||\mathbf{A}||}\right)$$

Bem.: Ist $\text{cond}(\mathbf{A}) \cdot ||\Delta\mathbf{A}|| \cdot ||\mathbf{A}||^{-1} \leq C_1 < 1$, so gilt

$$\frac{||\Delta\mathbf{x}||}{||\mathbf{x}||} \leq (1 - C_1)^{-1}\,\text{cond}(\mathbf{A})\left(\frac{||\Delta\mathbf{y}||}{||\mathbf{y}||} + \frac{||\Delta\mathbf{A}||}{||\mathbf{A}||}\right),$$

d.h. cond($\mathbf{A}$) kann als Verstärkungsfaktor der relativen Fehler interpretiert werden.
Faustregel: Für relative Fehler $\frac{||\Delta\mathbf{A}||}{||\mathbf{A}||} \approx \frac{||\Delta\mathbf{y}||}{||\mathbf{y}||} \approx 10^{-l}$ und cond($\mathbf{A}$) $\approx 10^k$ ist

$$\frac{||\Delta\mathbf{x}||}{||\mathbf{x}||} \approx 10^{k-l},$$

d.h. man verliert k Stellen Genauigkeit.

Lemma (benötigen wir zum Beweis von Satz 4)
Sei $\mathbf{B}$ (n,n) Matrix mit $||\mathbf{B}|| < 1$. Dann ist die Matrix $\mathbf{I}+\mathbf{B}$ regulär, und es gilt

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

Beweis: Es gilt für alle $\mathbf{x} \in \mathbb{R}^n, \mathbf{x} \neq \mathbf{0}$

$$||(\mathbf{I}+\mathbf{B})\mathbf{x}|| \geq ||\mathbf{x}|| - ||\mathbf{B}\mathbf{x}|| \geq \underbrace{(1-||\mathbf{B}||)}_{>0}||\mathbf{x}|| > 0$$

Die Gleichung $(\mathbf{I}+\mathbf{B})\mathbf{x} = \mathbf{0}$ besitzt also keine nichttriviale Lösung, also ist die Matrix $(\mathbf{I}+\mathbf{B})$ regulär.

$$\begin{aligned} 1 &= ||\mathbf{I}|| = ||(\mathbf{I}+\mathbf{B})(\mathbf{I}+\mathbf{B})^{-1}|| = ||(\mathbf{I}+\mathbf{B})^{-1} + \mathbf{B}(\mathbf{I}+\mathbf{B})^{-1}|| \\ &\geq ||(\mathbf{I}+\mathbf{B})^{-1}|| - ||\mathbf{B}||||(\mathbf{I}+\mathbf{B})^{-1}|| = ||(\mathbf{I}+\mathbf{B})^{-1}||(1-||\mathbf{B}||) > 0 \end{aligned}$$

woraus die Behauptung folgt. □

Beweis von Satz 4
Nach Voraussetzung ist $||\mathbf{A}^{-1}\Delta\mathbf{A}|| \leq ||\mathbf{A}^{-1}|| \cdot ||\Delta\mathbf{A}|| < 1$. Nach obigem Lemma ist auch $\mathbf{A}+\Delta\mathbf{A} = \mathbf{A}\underbrace{(\mathbf{I}+\mathbf{A}^{-1}\Delta\mathbf{A})}_{\text{regulär}}$ regulär.
Mit $\Delta\mathbf{x} = \mathbf{x} - \tilde{\mathbf{x}}$ folgt aus $\mathbf{A}\mathbf{x} = \mathbf{y}$ und $\tilde{\mathbf{A}}\tilde{\mathbf{x}} = \tilde{\mathbf{y}}$

$$\begin{aligned} (\mathbf{A}+\Delta\mathbf{A})\Delta\mathbf{x} &= (\mathbf{A}+\Delta\mathbf{A})(\mathbf{x}+\Delta\mathbf{x}-\mathbf{x}) \\ &= \underbrace{(\mathbf{A}+\Delta\mathbf{A})(\mathbf{x}+\Delta\mathbf{x})}_{\mathbf{y}+\Delta\mathbf{y}} - (\mathbf{A}+\Delta\mathbf{A})\mathbf{x} \\ &= \mathbf{y}+\Delta\mathbf{y} - (\mathbf{y}+\Delta\mathbf{A}\mathbf{x}) = \Delta\mathbf{y} - \Delta\mathbf{A}\mathbf{x} \\ \Delta\mathbf{x} &= (\mathbf{A}+\Delta\mathbf{A})^{-1}\big(\Delta\mathbf{y} - \Delta\mathbf{A}\mathbf{x}\big) \end{aligned}$$

$$\begin{aligned} ||\Delta\mathbf{x}|| &\leq ||(\mathbf{A}+\Delta\mathbf{A})^{-1}||\big(||\Delta\mathbf{y}|| + ||\Delta\mathbf{A}||||\mathbf{x}||\big) \\ &= ||(\mathbf{A}(\mathbf{I}+\mathbf{A}^{-1}\Delta\mathbf{A}))^{-1}||\big(||\Delta\mathbf{y}|| + ||\Delta\mathbf{A}||||\mathbf{x}||\big) \\ &= ||(\mathbf{I}+\mathbf{A}^{-1}\Delta\mathbf{A})^{-1}\mathbf{A}^{-1}||\big(||\Delta\mathbf{y}|| + ||\Delta\mathbf{A}||||\mathbf{x}||\big) \\ &\leq ||(\mathbf{I}+\mathbf{A}^{-1}\Delta\mathbf{A})^{-1}|| \cdot ||\mathbf{A}^{-1}||\big(||\Delta\mathbf{y}|| + ||\Delta\mathbf{A}||||\mathbf{x}||\big) \\ &\leq \frac{||\mathbf{A}^{-1}||}{1-||\mathbf{A}^{-1}|| \cdot ||\Delta\mathbf{A}||}\big(||\Delta\mathbf{y}|| + ||\Delta\mathbf{A}||||\mathbf{x}||\big) \qquad \text{(nach Lemma)} \\ &\leq \frac{|\mathbf{A}^{-1}|| \cdot ||\mathbf{A}|| \cdot ||\mathbf{x}||}{1-||\mathbf{A}^{-1}|| \cdot ||\Delta\mathbf{A}||}\left(\frac{||\Delta\mathbf{y}||}{||\mathbf{A}|| \cdot ||\mathbf{x}||} + \frac{||\Delta\mathbf{A}||}{||\mathbf{A}||}\right) \\ &\leq \frac{\text{cond(A)}}{1-\text{cond(A)}||\Delta\mathbf{A}|| \cdot ||\mathbf{A}||^{-1}}\left(\frac{||\Delta\mathbf{y}||}{||\mathbf{y}||} + \frac{||\Delta\mathbf{A}||}{||\mathbf{A}||}\right)||\mathbf{x}|| \end{aligned}$$

wegen $||\mathbf{y}|| = ||\mathbf{A}\mathbf{x}|| \leq ||\mathbf{A}||||\mathbf{x}||$ $\square$

Beispiel 12:

$$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 3 & 5 \end{pmatrix}, \qquad \mathbf{A}^{-1} = \begin{pmatrix} -5 & 2 \\ 3 & -1 \end{pmatrix}, \qquad \mathbf{y} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

Lösung von $\mathbf{A}\mathbf{x} = \mathbf{y}: \ \mathbf{x}^* = (-5, 3)'$.

$||\mathbf{A}||_1 = \max(1+3, 2+5) = 7, \qquad det(\mathbf{A}) = 1 \cdot 5 - 3 \cdot 2 = -1,$

$||\mathbf{A}^{-1}||_1 = \max(5+3, 2+1) = 8, \qquad det(\mathbf{A}^{-1}) = \frac{1}{det(\mathbf{A})} = -1,$

cond(A) $= ||\mathbf{A}||_1||\mathbf{A}^{-1}||_1 = 7 \cdot 8 = 56$

$$\tilde{\mathbf{A}} = \begin{pmatrix} 1 & 2 \\ 3.1 & 5 \end{pmatrix}, \qquad \tilde{\mathbf{A}}^{-1} = \begin{pmatrix} -4.1667 & 1.6667 \\ 2.5833 & -0.8333 \end{pmatrix}, \qquad \tilde{\mathbf{y}} = \begin{pmatrix} 1 \\ 0.1 \end{pmatrix}$$

Lösung von $\tilde{\mathbf{A}}\mathbf{x} = \tilde{\mathbf{y}}: \ \tilde{\mathbf{x}}^* = (-4, 2.5)'$.

Fortsetzung von Beispiel 12

$det(\tilde{\mathbf{A}}) = -1.2, \quad ||\Delta\mathbf{A}|| = 0.1, \quad ||\Delta\mathbf{A}||||\mathbf{A}^{-1}|| = 0.8 < 1$. Bedingung des Satzes erfüllt. Obere Schranke:
$\frac{56}{1-56\cdot\frac{0.1}{7}}\left(\frac{0.1}{1} + \frac{0.1}{7}\right) = \frac{56}{0.2} \cdot \frac{0.8}{7} = 32 > \frac{||\Delta\mathbf{x}||}{||\mathbf{x}||} = \frac{||(-5,3)-(-4,2.5)||}{||(-5,3)||} = \frac{1.5}{8}$

Beispiel 13:

$$\mathbf{A} = \begin{pmatrix} 1 & 0.99 \\ 0.99 & 0.98 \end{pmatrix} \qquad \mathbf{A}^{-1} = \begin{pmatrix} -9800 & 9900 \\ 9900 & -10000 \end{pmatrix}$$

$||\mathbf{A}||_1 = 1.99, \qquad det(\mathbf{A}) = 1 \cdot 0.98 - 0.99^2 = -0.0001, \qquad ||\mathbf{A}^{-1}||_1 = 19900, \qquad det(\mathbf{A}^{-1}) = \frac{1}{det(\mathbf{A})} = -10000,$
cond(A) $= ||\mathbf{A}||_1||\mathbf{A}^{-1}||_1 \approx 4 \cdot 10^4$

$$\tilde{\mathbf{A}} = \begin{pmatrix} 1 & 0.99 \\ 0.99 & 0.99 \end{pmatrix} \qquad \tilde{\mathbf{A}}^{-1} = \begin{pmatrix} 100 & -100 \\ -100 & \frac{10000}{99} \end{pmatrix}$$

$det(\tilde{\mathbf{A}}) = 1 \cdot 0.99 - 0.99^2 = 0.0099, \quad det(\tilde{\mathbf{A}}^{-1}) = \frac{1}{det(\tilde{\mathbf{A}})} \approx 100 \quad ||\Delta\mathbf{A}|| = 0.01, \quad ||\Delta\mathbf{A}||||\mathbf{A}^{-1}|| > 1$. Bedingung des Satzes nicht erfüllt (Übung).

Fortsetzung von Beispiel 13

Lösung von

$$\mathbf{A} = \begin{pmatrix} 1 & 0.99 \\ 0.99 & 0.98 \end{pmatrix} \mathbf{x} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}: \qquad \mathbf{x}^* = \mathbf{A}^{-1}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 100 \\ -100 \end{pmatrix}$$

Lösung von

$$\tilde{\mathbf{A}} = \begin{pmatrix} 1 & 0.99 \\ 0.99 & 0.99 \end{pmatrix} \mathbf{x} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}: \qquad \tilde{\mathbf{x}}^* = \tilde{\mathbf{A}}^{-1}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ \frac{100}{99} \end{pmatrix}$$

1.6 Einfluss des Algorithmus auf die Laufzeit

Beispiel 14: Berechnung von e^x für $x = x_0 = -15$ auf 10 Stellen genau

Der wahre Wert ist (auf 10 Stellen genau) $e^{x_0} = 3.059023205 \cdot 10^{-7}$. Potenzreihe

$$\sum_{i=0}^{50} \frac{(-15)^i}{i!} = 7.840047895 \cdot 10^{-7}$$

50 Summanden reichen nicht.

$$\sum_{i=0}^{70} \frac{(-15)^i}{i!} = 3.059023205 \cdot 10^{-7}$$

Berechnen erst $x = e^{15}$, und bilden dann $y = \frac{1}{x}$

$$\frac{1}{\sum_{i=0}^{45} \frac{15^i}{i!}} = 3.059023205 \cdot 10^{-7}$$

d.h. hier reichen 46 Summanden.

1.7 Einfluss des Algorithmus auf die Stabilität

Def. 8 (Stabilität)

Ein Algorithmus heißt numerisch stabil für ein gegebenes Problem, falls der relative Fehler $\leq$ relativen Kondition für das Problem ist.

Beispiel 15: Lösung der quadratischen Gleichung $x^2 - 2px + q = 0$: $f(p,q) = (x_1, x_2) = p \mp \sqrt{p^2 - q}$

Das Problem ist gut konditioniert, falls $q \ll p^2$: Die Jacobi-Matrix ist

$$\mathbf{J} = \begin{pmatrix} \frac{\partial x_1}{\partial p} & \frac{\partial x_1}{\partial q} \\ \frac{\partial x_2}{\partial p} & \frac{\partial x_2}{\partial q} \end{pmatrix} = \begin{pmatrix} 1 - \frac{p}{\sqrt{p^2-q}} & \frac{1}{2\sqrt{p^2-q}} \\ 1 + \frac{p}{\sqrt{p^2-q}} & \frac{-1}{2\sqrt{p^2-q}} \end{pmatrix}$$

Die absolute und relative Kondition

$$\begin{aligned} \kappa_{\text{abs}} &= ||\mathbf{J}|| = \max_j \sum_i |a_{ij}| \leq \sum_{i,j} |a_{ij}| \leq 2 + \frac{2|p|+1}{\sqrt{p^2-q}} \\ \kappa_{\text{rel}} &= \frac{||(p,q)||}{||f(p,q)||} ||\mathbf{J}|| = \frac{|p|+|q|}{|x_1|+|x_2|} ||\mathbf{J}|| \leq \frac{|p|+|q|}{2|p|} (2 + \frac{2|p|+1}{\sqrt{p^2-q}}) \end{aligned}$$

Fortsetzung von Beispiel 15:

Welchen Algorithmus wählen wir?

1. p-q Formel

$$x_{1,2} = p \mp \sqrt{p^2 - q} \qquad (1)$$

2. Vietascher Wurzelsatz ($x_1 x_2 = q$)

$$\begin{aligned} \text{Fall} \quad p &\leq 0 & \qquad \text{Fall} \quad p &> 0 \\ x_1 &= p - \sqrt{p^2-q} & x_2 &= p + \sqrt{p^2-q} \\ x_2 &= \frac{q}{x_1} & x_1 &= \frac{q}{x_2} \qquad (2) \end{aligned}$$

Zunächst berechnen wir $u = p^2, v = u - q, w = \sqrt{v}$ und zur Vermeidung von Auslöschung

$$\begin{aligned} x_1 &= p - w \quad \text{falls} \quad p < 0 \\ x_2 &= p + w \quad \text{falls} \quad p > 0 \end{aligned}$$

mit dem relativen Fehler (sei o.B.d.A. $p > 0$), vgl. Satz 3

$$\begin{aligned} \left|\frac{\Delta x_2}{x_2}\right| &\leq \left|\underbrace{\frac{\partial x_2}{\partial p}\frac{p}{x_2}}_{\kappa_{21}(x)}\right| \left|\frac{\Delta p}{p}\right| + \left|\underbrace{\frac{\partial x_2}{\partial w}\frac{w}{x_2}}_{\kappa_{22}(x)}\right| \left|\frac{\Delta w}{w}\right| \\ &= \left|1\right| \left|\frac{p}{x_2}\right| \left|\frac{\Delta p}{p}\right| + \left|1\right| \left|\frac{w}{x_2}\right| \left|\frac{\Delta w}{w}\right| \\ &= \left|\frac{p}{p+w}\right| \left|\frac{\Delta p}{p}\right| + \left|\frac{w}{p+w}\right| \left|\frac{\Delta w}{w}\right| \end{aligned}$$

Wenn $q \ll p^2$ so sind die komponentenweisen Konditionen etwa $\frac{1}{2}$.
Nach der p-q Formel können wir jetzt $x_1 = p - w$ (wenn $p > 0$) (1.) ausrechnen oder wir verwenden den Vietaschen Wurzelsatz und berechnen $x_1 = \frac{q}{x_2}$ (2.)

- $x_1 = p - w$. Bei $q \ll p^2$ ist $w \approx p$, d.h. wir haben Auslöschung. Rundungsfehler in p und w übertragen sich wie folgt (vgl. Satz 3):

$$\begin{aligned}\left|\frac{\Delta x_1}{x_1}\right| &\leq \underbrace{\left|\frac{\partial x_1}{\partial p}\frac{p}{x_1}\right|}_{\kappa_{11}(x)}\left|\frac{\Delta p}{p}\right| + \underbrace{\left|\frac{\partial x_1}{\partial w}\frac{w}{x_1}\right|}_{\kappa_{12}(x)}\left|\frac{\Delta w}{w}\right| = \left|\frac{p}{x_1}\right|\left|\frac{\Delta p}{p}\right| + \left|\frac{w}{x_1}\right|\left|\frac{\Delta w}{w}\right| \\ &= \underbrace{\left|\frac{p}{p-w}\right|}_{\gg 1}\underbrace{\left|\frac{\Delta p}{p}\right|}_{\leq \text{eps}} + \underbrace{\left|\frac{-w}{p-w}\right|}_{\gg 1}\underbrace{\left|\frac{\Delta w}{w}\right|}_{\approx \text{eps}}\end{aligned}$$

  Dieser Algorithmus ist also instabil.

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

$$\begin{aligned}\left|\frac{\Delta x_1}{x_1}\right| &\leq \left|\frac{\partial x_1}{\partial q}\frac{q}{x_1}\right|\left|\frac{\Delta q}{q}\right| + \left|\frac{\partial x_1}{\partial x_2}\frac{x_2}{x_1}\right|\left|\frac{\Delta x_2}{x_2}\right| \\ &= \left|\underbrace{\frac{q}{x_1 x_2}}_{=1}\right|\left|\frac{\Delta q}{q}\right| + \left|\underbrace{\frac{q}{x_2^2}\frac{x_2}{x_1}}_{=1}\right|\left|\frac{\Delta x_2}{x_2}\right|\end{aligned}$$

  Dieser Algorithmus ist also stabil.

Beispiel 16:, Fortsetzung von Beispiel 15 (Lösung der quadratischen Gleichung $x^2 - 2px + q = 0$ mit $q = 0.87$)

p	x_2	x_1 (1)	x_1 (2)	x_1 (2)- x_1 (1)
$9.728 \cdot 10^2$	194.556	0.00447173	0.00447173	$5.0307 \cdot 10^{-16}$
$2.55 \cdot 10^8$	$5.1 \cdot 10^8$	**0**	$1.70588 \cdot 10^{-9}$	$1.70588 \cdot 10^{-9}$

p-q-Formel liefert im zweiten Beispiel ein falsches Ergebnis.

1.8 Auswertung arithmetischer Ausdrücke

Beispiel 17:

$$y = f(x_1, x_2) = \underbrace{x_1^2 - x_2^2}_{A} = \underbrace{(x_1 - x_2)(x_1 + x_2)}_{B}$$

Relativer Problemfehler (nach Satz 3)

$$\begin{aligned}\left|\frac{\Delta y}{y}\right| &\leq \left|\frac{\partial f}{\partial x_1}\frac{x_1}{f}\right|\left|\frac{\Delta x_1}{x_1}\right| + \left|\frac{\partial f}{\partial x_2}\frac{x_2}{f}\right|\left|\frac{\Delta x_2}{x_2}\right| \\ &\leq \left(\left|\frac{2x_1 \cdot x_1}{x_1^2 - x_2^2}\right| + \left|\frac{-2x_2 \cdot x_2}{x_1^2 - x_2^2}\right|\right)\epsilon = 2\left(\left|\frac{x_1^2 + x_2^2}{x_1^2 - x_2^2}\right|\right)\epsilon\end{aligned}$$

wobei

$$\left|\frac{\Delta x_i}{x_i}\right| = \left|\frac{x_i(1+\epsilon_i) - x_i}{x_i}\right| = \epsilon_i \leq \epsilon$$

weiter mit Beispiel 17
Relative Kondition

$$\kappa_{rel} = \frac{||\mathbf{x}||}{||f(\mathbf{x})||}||\mathbf{J}|| = \frac{|x_1| + |x_2|}{|x_1^2 - x_2^2|}||(2x_1, 2x_2)|| = 2\frac{|x_1 + x_2|}{|x_1^2 - x_2^2|}\max(|x_1|, |x_2|)$$

schlechte Kondtionierung für $|x_1| \approx |x_2|$.

Betrachten wir Algorithmen A und B:

$$\begin{aligned} u_1 &:= x_1 \boxdot x_1 & v_1 &= x_1 \boxplus x_2 \\ u_2 &:= x_2 \boxdot x_2 & v_2 &= x_1 \boxminus x_2 \\ \tilde{y} &:= u_1 \boxminus u_2 & \tilde{z} &:= v_1 \boxdot v_2 \end{aligned}$$

Nach der Folgerung aus Satz 1 ist der Rundungsfehler bei Maschinenoperationen (z.B. $\circledast = \boxplus, \boxdot, \boxminus$) gleich: $rd(x_1 \circledast x_2) = (1+\varepsilon)(x_1 \circledast x_2)$, wobei $\varepsilon \leq \text{eps}$

Für Algorithmus (A) ergibt sich

$$\begin{aligned} u_1 &:= x_1^2(1+\varepsilon_1) \qquad u_2 := x_2^2(1+\varepsilon_2) \\ \tilde{y} &:= (x_1^2(1+\varepsilon_1) - x_2^2(1+\varepsilon_2))(1+\varepsilon_3) \\ &= \underbrace{x_1^2 - x_2^2}_{=y} + x_1^2\varepsilon_1 - x_2^2\varepsilon_2 + \underbrace{(x_1^2 - x_2^2)}_{=y}\varepsilon_3 + \mathcal{O}(\text{eps}^2) \\ \left|\frac{\Delta y}{y}\right| &\leq \text{eps}\frac{x_1^2 + x_2^2 + |x_1^2 - x_2^2|}{|x_1^2 - x_2^2|} \end{aligned}$$

d.h. der Rundungsfehlereinfluss kann groß sein, wenn $|x_1| \approx |x_2|$.

Für Algorithmus (B) ergibt sich

$$\begin{aligned} v_1 &:= (x_1 + x_2)(1+\varepsilon_1) \qquad v_2 := (x_1 - x_2)(1+\varepsilon_2) \\ \tilde{z} &:= (x_1 + x_2)(1+\varepsilon_1)(x_1 - x_2)(1+\varepsilon_2)(1+\varepsilon_3) \\ &= \underbrace{x_1^2 - x_2^2}_{=z} + \underbrace{(x_1^2 - x_2^2)}_{=z}(\varepsilon_1 + \varepsilon_2 + \varepsilon_3) + \mathcal{O}(\text{eps}^2) \\ \left|\frac{\Delta z}{z}\right| &\leq \left|\varepsilon_1 + \varepsilon_2 + \varepsilon_3\right| \leq 3 \cdot \text{eps} \end{aligned}$$

d.h. der Rundungsfehlereinfluss ist deutlich kleiner.

Auswertung von Polynomen $p(x) = \sum_{i=0}^{n} a_i x^i$
verlangt $\frac{n(n+1)}{2}$ Multiplikationen und n Additionen.
Schneller (nur noch n Multiplikationen) (Horner-Schema):

$$\begin{aligned} p(x) &= a_0 + x(a_1 + x(a_2 + x(a_3 + \cdots x(\underbrace{a_{n-1} + \underbrace{a_n}_{b_{n-1}} x}_{b_{n-2}} \underbrace{)\cdots)}_{n-1 \text{ mal}} \\ b_{n-1} &:= a_n \\ b_j &:= a_{j+1} + x \cdot b_{j+1} \quad j = n-2, \ldots, 0 \\ p(x) &:= a_0 + x \cdot b_0 \end{aligned}$$

verlangt nur noch n Multiplikationen und n Additionen

Beispiel 18: $p(x) = x^4 - 3x^3 - 8x^2 - 17x - 4$ *an der Stelle* $x = 2$

$$\begin{aligned} p(x) &= -4 + x(\underbrace{-17 + x(\underbrace{-8 + x(\underbrace{-3 + \underbrace{x}_{2}}_{-1})}_{-10})}_{-37}) \\ &= -4 + 2 \cdot (-37) = -78 \end{aligned}$$

Bem.: Die Auswertung nach dem Horner Schema ist auch numerisch stabiler.

Bem.: Es zeigt sich generell (Fehlerfortpflanzungsgesetz, s. z.B. Stoer), dass die zuletzt ausgeführte Operation den entscheidendsten Einfluss auf den Rundungsfehler hat, deshalb
Unvermeidbare numerisch instabile Operationen (z.B. Auslöschung) am Anfang durchführen!
Auswertung von

- irrationalen Konstanten z.B. durch schnell konvergierende Reihen, z.B. verwendet Mathematica für π eine hypergeometrische Reihe (Chudnovski-Algorithmus)
- trigonometrischen Funktionen durch Potenzreihenentwicklung
- rationalen Funktionen durch Kettenbruchbildung
- Logarithmus für $|x| < 1$ durch Potenzreihenentwicklung, und für $|x| > 1$ durch Logarithmusgesetze.
- teilweise sind mathematische Grundfunktionen schon in der Hardware implementiert.

2 Nichtlineare Gleichungssysteme

Wir betrachten das System von Gleichungen

$$f(\mathbf{x}) = \mathbf{0}, \qquad \mathbf{x} \in \mathbf{M} \subseteq \mathbb{R}^{\mathbf{n}}, \mathbf{f}(\mathbf{x}) \in \mathbb{R}^{\mathbf{n}}$$

d.h. wir untersuchen Nullstellen der Funktion f.
Lösungsmethode: Intervallschachtelung ($n = 1$) oder Iterationsverfahren

2.1 Intervallschachtelung

Problem: Für $f : [a, b] \to \mathbb{R}$, $a, b \in \mathbb{R}$ suche eine Nullstelle $x \in [a, b]$: $f(x) = 0$.

Satz 5 (Existenz einer Lösung)
Sei $f : [a, b] \to \mathbb{R}$ stetig und $f(a) \leq 0$ und $f(b) \geq 0$. Dann existiert ein $x \in [a, b]$ mit $f(x) = 0$.

Beweis: Folgt direkt aus dem Zwischenwertsatz für stetige Funktionen. □

Satz 6 (Konstruktion einer Lösung)
Sei $f : [a, b] \to \mathbb{R}$ stetig und $f(a) \leq 0$ und $f(b) \geq 0$. Setze $a_0 := a$, $b_0 := b$. Für $n \in \mathbb{N}$ definiere $x_n := \frac{1}{2}(a_n + b_n)$ und

$$\begin{cases} a_{n+1} := x_n, & b_{n+1} := b_n \quad \text{falls} \quad f(x_n) \leq 0 \\ a_{n+1} := a_n, & b_{n+1} := x_n \quad \text{falls} \quad f(x_n) > 0. \end{cases}$$

Dann gilt

1. $[a_{n+1}, b_{n+1}] \subset [a_n, b_n]$,
2. Die Folge (a_n) ist monoton wachsend und nach oben beschränkt, die Folge (b_n) monoton fallend und nach unten beschränkt,
3. $b_n - a_n = 2^{-n}(b - a)$,
4. $\lim a_n = \lim b_n = \tilde{x}$ und $f(\tilde{x}) = 0$.

Beweis: Folgt direkt aus Konvergenzsätzen. □

Beispiel 19: Berechnen die reelle Wurzel $3^{\frac{1}{3}}$, $f(x) = x^3 - 3$, $x^* \approx 1.44225$.

i	a_i	b_i	$f(a_i)$	$f(b_i)$	x_i	$f(x_i)$
0	1	3	$-2 < 0$	$24 > 0$	2	$5 > 0$
1	1	2		$5 > 0$	$\frac{3}{2}$	$\frac{3}{8} > 0$
2	1	$\frac{3}{2}$		$\frac{3}{8} > 0$	$\frac{5}{4}$	$-\frac{67}{64} < 0$
3	$\frac{5}{4}$	$\frac{3}{2}$	$-\frac{67}{64} < 0$		$\frac{11}{8}$	$-\frac{205}{512} < 0$
4	$\frac{11}{8}$	$\frac{3}{2}$	$-\frac{205}{512} < 0$		$\frac{23}{16} = 1.4375$	

Bem.:

- Die Intervallschachtelung ist numerisch sehr stabil, aber die Konvergenz ist sehr langsam! Im Allgemeinen konvergieren (a_n), (b_n), (x_n) linear gegen $\tilde{x}$. (Brauchen 10 Schritte, damit $b_n - a_n \leq 10^{-3}(b_0 - a_0)$ $\quad (2^{-10} = 1024 \approx 10^{-3})$).
- Funktioniert nur für Funktionen in $\mathbb{R}^1$ $(f : [a, b] \to \mathbb{R})$.
- Nur geringe Voraussetzungen an f nötig (Stetigkeit + Vorzeichen an Intervallgrenzen).
- Andere Wahl des Teilungspunktes x_n kann Konvergenz beschleunigen (z.B. „Regula falsi").

2.2 Iterationsverfahren

Wir betrachten die Iterationsvorschrift

$$\mathbf{x}_{n+1} = \Phi(\mathbf{x}_n), \qquad n = 0, 1, \ldots$$

wobei $\Phi(\mathbf{x})$ eine Iterationsfunktion $M \to M$ ist mit einem Fixpunkt $\boldsymbol{\xi}$, d.h. $\Phi(\boldsymbol{\xi}) = \boldsymbol{\xi}$. $\mathbf{x}_0$ ist ein geeignet zu wählender Startpunkt.
Fragen:

- Wie findet man ein geeignetes Verfahren?
- Unter welchen Bedingungen konvergiert die Folge der $\mathbf{x}_n$?
- Konvergenzgeschwindigkeit?

Verfahren:

- $\mathbf{x}_{n+1} = \mathbf{x}_n + f(\mathbf{x}_n)$ Einfache Iteration
- $\mathbf{x}_{n+1} = \mathbf{x}_n + C \cdot f(\mathbf{x}_n)$, Modifizierte Einfache Iteration
- $\mathbf{x}_{n+1} = \mathbf{x}_n + C(\mathbf{x}_n) f(\mathbf{x}_n)$, wobei $C(\mathbf{x})$ eine Funktionsmatrix ist Beispiel: Newton Verfahren (beruht auf Taylorreihenentwicklung)

2.3 Banach-Fixpunktsatz

Def. 9 (kontrahierende Abbildung)
Sei $(R, ||.||)$ normierter Raum. Eine eindeutige Abbildung T mit $Db(T), Wb(T) \subseteq R$ heißt kontrahierend, falls es eine reelle Zahl q, $0 \leq q < 1$ gibt, so dass

$$||T(x) - T(y)|| \leq q \cdot ||x - y|| \qquad \forall x, y \in R$$

Def. 10 (Fundamentalfolge)

Eine Folge $\{a_n\}$ heißt Fundamentalfolge, falls

$$\forall \varepsilon > 0 \quad \exists n_0 : \forall n, m \geq n_0 : \quad |a_n - a_m| < \varepsilon.$$

Def. 11 (Vollständiger normierter Raum)
Ein normierter Raum R heißt vollständig, falls jede Fundamentalfolge (=Cauchy-Folge) einen Grenzwert in R besitzt.

Beispiel 20:

- $(\mathbb{R}^n, ||.||)$ mit $||.||$ euklidische Norm oder Maximumnorm
- $(\mathbb{Q}^n, ||.||)$ ist nicht vollständig.

Satz 7 (Stefan Banach, 1892-1945, Fixpunktsatz)
Eine kontrahierende Abbildung T einer abgeschlossenen Teilmenge A eines vollständigen normierten Raumes in sich selbst besitzt genau einen Fixpunkt.

Beweis: Sei $x_1 \in A$ beliebig. Wir definieren eine Folge (x_n) wie folgt

$$x_{n+1} = T(x_n), \qquad n = 1, 2, \ldots$$

und zeigen als erstes, (x_n) ist Cauchy-Folge.
Zunächst ist (Beweis induktiv)

$$||x_n - x_{n+1}|| \leq q^{n-1}||x_1 - x_2|| \qquad (q < 1)$$

Weiter folgt $\forall n, m, n < m$:

$$\begin{aligned} ||x_n - x_m|| &\leq \sum_{i=n}^{m-1} ||x_i - x_{i+1}|| \leq \sum_{i=n}^{m-1} q^{i-1}||x_1 - x_2|| \\ &\leq q^{n-1}||x_1 - x_2|| \sum_{j=0}^{\infty} q^j = \frac{q^{n-1}}{1-q}||x_1 - x_2||. \end{aligned}$$

Die Folge auf der rechten Seite ist eine Nullfolge.
Sei $\varepsilon > 0$. Dann $\exists n_0 : \forall n \geq n_0$:

$$\frac{q^{n-1}}{1-q}||x_2 - x_1|| < \varepsilon$$

Damit $\forall n, m \geq n_0$

$$||x_n - x_m|| < \varepsilon$$

Wegen der Vollständigkeit von $(R, ||.||)$, A abgeschlossen und $A \subseteq R$ hat die Folge (x_n) einen Grenzwert in A, sei dieser x^*.
Wir zeigen, dass auch $\lim x_n = T(x^*)$. Dann folgt, da es (in R) nur einen Grenzwert geben kann: $x^* = T(x^*)$, d.h. x^* ist Fixpunkt.
Es gilt $\forall i \in \mathbb{N}$:

$$||x_{n+1} - T(x^*)|| = ||T(x_n) - T(x^*)|| \leq q \cdot ||x_n - x^*|| < ||x_n - x^*||$$

Da $(||x_n - x^*||)$ Nullfolge sind auch $(||x_{n+1} - T(x^*)||)$ und $(||x_n - T(x^*)||)$ Nullfolgen und obige Behauptung ist gezeigt.
Bleibt zu zeigen, x^* ist einziger Fixpunkt.
Angenommen, x^*, x^{**} sind Fixpunkte und $x^* \neq x^{**}$. Dann $||x^* - x^{**}|| \neq 0$ und

$$||x^* - x^{**}|| = ||T(x^*) - T(x^{**})|| \leq q \cdot ||x^* - x^{**}|| < ||x^* - x^{**}||$$

Widerspruch. □

Bem. Der Beweis ist konstruktiv. Das in dem Beweis angegebene Verfahren nennt man Verfahren der *sukzessiven Approximation.*

Beispiel 21: $n = 1,\ f(x) = \frac{1}{4} + \frac{1}{2}\sin(x),\ M = [0, \frac{\pi}{2}], \quad x^* \approx 0.481598$

$$f(0) = \frac{1}{4}, \quad f(\frac{\pi}{2}) = \frac{3}{4}, \quad f([0, \frac{\pi}{2}]) = [\frac{1}{4}, \frac{3}{4}] \subset M$$

$$f(\frac{\pi}{6}) = \frac{1}{2}, f(\frac{\pi}{3}) = \frac{\sqrt{3}}{4} + \frac{1}{4}$$

Nach dem Mittelwertsatz $\exists \xi \in [y, z]$ mit

$$|f(y) - f(z)| = |f'(\xi)||y - z| \leq \sup_x |f'(x)||y - z| \quad \text{(allgemein)}$$

$$\sup_{0 \leq x \leq \frac{\pi}{2}} |f'(x)| = \sup_{0 \leq x \leq \frac{\pi}{2}} \frac{1}{2}\cos(x) = \frac{1}{2} < 1$$

d.h. f ist kontrahierende Abildung und es existiert ein eindeutiger Fixpunkt x^*, d.h. $x^* = f(x^*)$

Beispiel 21 (Fortsetzung)
$n = 1,\ f(x) = \frac{1}{4} + \frac{1}{2}\sin(x),\ M = [0, \frac{\pi}{2}] \quad x^* \approx 0.481598$

x_0	$=$	0	x_6	$\approx$	0.47724
x_1	$=$	$\frac{1}{4}$	x_7	$\approx$	0.479665
x_2	$\approx$	0.3737	x_8	$\approx$	0.480741
x_3	$\approx$	0.432532	x_9	$\approx$	0.481218
x_4	$\approx$	0.459586	x_{10}	$\approx$	0.48143
x_5	$\approx$	0.471788	x_{11}	$\approx$	0.481523

2.4 Iterative Lösung Linearer Gleichungssysteme

Sei $\mathbf{A}$ (n, n)-Matrix mit vollem Rang. Dann besitzt das lineare Gleichungssystem

$$\mathbf{Ax} = \mathbf{b}$$

genau eine Lösung. Zur Iterativen Lösung betrachten wir die Fixpunktgleichung

$$\mathbf{x} = \mathbf{x} - \mathbf{C}(\mathbf{Ax} - \mathbf{b}) = \underbrace{(\mathbf{I} - \mathbf{CA})}_{\mathbf{B}}\mathbf{x} + \underbrace{\mathbf{Cb}}_{\mathbf{d}} = \mathbf{Bx} + \mathbf{d} =: \tilde{\mathbf{B}}\mathbf{x} \tag{1}$$

mit geeigneter regulärer Matrix $\mathbf{C}$.
Den Banach-Fixpunktsatz können wir anwenden, falls die Abbildung $\tilde{\mathbf{B}}$ kontrahierend:

$$||\tilde{\mathbf{B}}\mathbf{x} - \tilde{\mathbf{B}}\mathbf{y}|| = ||\mathbf{Bx} - \mathbf{By}|| \leq ||\mathbf{B}|| \cdot ||\mathbf{x} - \mathbf{y}||$$

d.h. wir brauchen $||\mathbf{B}|| < 1$ für eine geeignete Matrixnorm.

Jacobi-Verfahren

Satz 8 (Jacobi-Verfahren, Gesamtschritt-Verfahren)
Sei $\mathbf{A}$ regulär, $\mathbf{C} = (\text{diag}\mathbf{A})^{-1}$ und $||.||$ eine geeignete Matrixnorm. Dann besitzt die Gleichung

$$\mathbf{x} = \mathbf{Bx} + \mathbf{d}$$

für beliebiges $\mathbf{d}$ genau eine Lösung, und die Folge $(\mathbf{x}_m)$ mit

$$\mathbf{x}_{m+1} = \mathbf{B}\mathbf{x}_m + \mathbf{d} = (\mathbf{I} - \mathbf{C}\mathbf{A})\mathbf{x}_m + \mathbf{C}\mathbf{b}$$

konvergiert für beliebiges $\mathbf{x}_0$ gegen die Lösung falls die Matrix $\mathbf{A}$ diagonaldominant ist.

Def. 12 (Diagonaldominanz)
Eine Matrix $\mathbf{A} = (a_{ij})$ heißt diagonaldominant, falls

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}| \quad \forall i = 1, \ldots, n$$

Beweis von Satz 8:
Es genügt zu zeigen, $||\mathbf{B}|| < 1$ für eine geeignete Matrixnorm.
Es gilt für die Einträge der Matrix $\mathbf{B} = \mathbf{I} - \mathbf{C}\mathbf{A} = \mathbf{I} - (\mathrm{diag}(\mathbf{A})^{-1})\mathbf{A}$:

$$b_{ij} := \begin{cases} 0 & \text{falls} \quad j = i \\ -\frac{a_{ij}}{a_{ii}} & \text{sonst} \end{cases}$$

Für die i-te Komponente x_i von $\mathbf{x} = (\mathbf{I} - \mathbf{C}\mathbf{A})\mathbf{x} + \mathbf{C}\mathbf{b}$ folgt:

$$x_i = \sum_j b_{ij} x_j + \frac{b_i}{a_{ii}} = \frac{1}{a_{ii}}\big(b_i - \sum_{j \neq i} a_{ij} x_j\big)$$

Seien jetzt $\mathbf{x}_m = (x_{m,1} \ldots, x_{m,n})$ und $\mathbf{y}_m = (y_{m,1} \ldots, y_{m,n})$.

$$\begin{aligned} |x_{m+1,i} - y_{m+1,i}| &= \left| \frac{1}{a_{ii}}\big(b_i - \sum_{j \neq i} a_{ij} x_{m,j}\big) - \frac{1}{a_{ii}}\big(b_i - \sum_{j \neq i} a_{ij} y_{m,j}\big) \right| \\ &\leq \frac{1}{|a_{ii}|} \sum_{j \neq i} |a_{ij}||x_{m,j} - y_{m,j}| \\ &\leq \underbrace{\big(\frac{1}{|a_{ii}|} \sum_{j \neq i} |a_{ij}|\big)}_{<1} ||\mathbf{x}_m - \mathbf{y}_m||_\infty \end{aligned}$$

wegen der Diagonaldominanz der Matrix $\mathbf{A}$.

$$\begin{aligned} ||\mathbf{B}\mathbf{x}_m - \mathbf{B}\mathbf{y}_m||_\infty &= ||\mathbf{x}_{m+1} - \mathbf{y}_{m+1}||_\infty \\ &= \max_{1 \leq i \leq n} |x_{m+1,i} - y_{m+1,i}| \\ &\leq \underbrace{\max_{1 \leq i \leq n} \big(\frac{1}{|a_{ii}|} \sum_{j \neq i} |a_{ij}|\big)}_{||\mathbf{B}||_\infty < 1} ||\mathbf{x}_m - \mathbf{y}_m||_\infty \end{aligned}$$

□

$||\mathbf{B}||_\infty$ ist hier die Zeilensummennorm.

Beispiel 22: $\mathbf{A}\mathbf{x} = \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix} \mathbf{x} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, $\mathbf{A}$ *ist diagonaldominant.* $\quad \mathbf{x}^* = (\frac{2}{3}, \frac{2}{3})$

Sei $\mathbf{x}_o^T = (1, 0)$, $\mathbf{C} = \mathbf{I}$. $\quad$ Algorithmus: $\mathbf{x}_{n+1} = \begin{pmatrix} 0 & -\frac{1}{2} \\ -\frac{1}{2} & 0 \end{pmatrix} \mathbf{x}_n + \begin{pmatrix} 1 \\ 1 \end{pmatrix}$

$$
\begin{array}{rclrclrcl}
\mathbf{x}_1^T & = & (1, \frac{1}{2}), & \mathbf{x}_4^T & = & (\frac{11}{16}, \frac{5}{8}), & \mathbf{x}_7^T & = & (\frac{43}{64}, \frac{85}{128}) \\
\mathbf{x}_2^T & = & (\frac{3}{4}, \frac{1}{2}), & \mathbf{x}_5^T & = & (\frac{11}{16}, \frac{21}{32}), & \mathbf{x}_8^T & = & (\frac{171}{256}, \frac{85}{128}) \\
\mathbf{x}_3^T & = & (\frac{3}{4}, \frac{5}{8}), & \mathbf{x}_6^T & = & (\frac{43}{64}, \frac{21}{32}), & \mathbf{x}_9^T & = & (0.6680, 0.6660)
\end{array}
$$

Gauß-Seidel-Verfahren

Wir zerlegen jetzt die Matrix $\mathbf{A}$ in drei Teile, $\mathbf{A} = \mathbf{L} + \mathbf{D} + \mathbf{R}$, wobei $\mathbf{D} = \operatorname{diag} \mathbf{A}$ und $\mathbf{L}$ und $\mathbf{R}$ obere und untere Dreiecksmatrizen sind. Wir wählen jetzt $\mathbf{C} = (\mathbf{L}+\mathbf{D})^{-1}$. Damit wird die Fixpunktgleichung (1), d.h. $\mathbf{x} = \mathbf{x} - \mathbf{C}(\mathbf{A}\mathbf{x} - \mathbf{b})$ zu

$$
\begin{aligned}
\mathbf{x} &= (\mathbf{I} - (\mathbf{L}+\mathbf{D})^{-1}\,\mathbf{A})\,\mathbf{x} + (\mathbf{L}+\mathbf{D})^{-1}\,\mathbf{b} \\
&= (\mathbf{I} - (\mathbf{L}+\mathbf{D})^{-1}(\mathbf{L}+\mathbf{D}+\mathbf{R}))\,\mathbf{x} + (\mathbf{L}+\mathbf{D})^{-1}\,\mathbf{b} \\
&= -(\mathbf{L}+\mathbf{D})^{-1}\,\mathbf{R}\,\mathbf{x} + (\mathbf{L}+\mathbf{D})^{-1}\,\mathbf{b}
\end{aligned}
$$

Das Iterationsverfahren wird dann zu

$$
\begin{aligned}
\mathbf{x}_{m+1} &= -(\mathbf{L}+\mathbf{D})^{-1}\,\mathbf{R}\,\mathbf{x}_m + (\mathbf{L}+\mathbf{D})^{-1}\,\mathbf{b} \\
(\mathbf{L}+\mathbf{D})\mathbf{x}_{m+1} &= -\mathbf{R}\,\mathbf{x}_m + \mathbf{b} \qquad | - \mathbf{L}\mathbf{x}_{m+1} \quad | \cdot \mathbf{D}^{-1} \\
\mathbf{x}_{m+1} &= -\mathbf{D}^{-1}\,\mathbf{R}\,\mathbf{x}_m - \mathbf{D}^{-1}\,\mathbf{L}\,\mathbf{x}_{m+1} + \mathbf{D}^{-1}\,\mathbf{b}
\end{aligned}
$$

Bem.: Beim Jacobi-Verfahren haben wir $\mathbf{C} = (\operatorname{diag} \mathbf{A})^{-1}$ gewählt.

$$
\begin{aligned}
\mathbf{x}_{m+1} &= -\mathbf{D}^{-1}\,\mathbf{R}\,\mathbf{x}_m - \mathbf{D}^{-1}\,\mathbf{L}\,\mathbf{x}_{m+1} + \mathbf{D}^{-1}\,\mathbf{b} \\
x_{m+1,i} &= -\sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}} x_{m,j} - \sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} x_{m+1,j} + \frac{b_i}{a_{ii}}
\end{aligned}
$$

Zur Berechnung von $x_{m+1,i}$ werden also die ersten $i-1$ Komponenten von $\mathbf{x}_{m+1}$ mit verwendet.

Satz 9 (Gauß-Seidel-Verfahren, Einzelschritt-Verfahren)
Sei $\mathbf{A}$ regulär und diagonaldominant. Dann besitzt die Gleichung

$$\mathbf{x} = -(\mathbf{L}+\mathbf{D})^{-1}\,\mathbf{R}\mathbf{x} + \mathbf{d}$$

für beliebiges $\mathbf{d}$ genau eine Lösung, und die obige Folge $(\mathbf{x}_m)$
konvergiert für beliebiges $\mathbf{x}_0$ gegen die Lösung $\mathbf{x}^*$ und es gilt

$$||\mathbf{x}_m - \mathbf{x}^*|| \le \rho^m ||\mathbf{x}_1 - \mathbf{x}_0||, \quad \text{wobei } \rho < 1$$

Beweis von Satz 9:

Seien $\alpha_i = \sum_{j=1}^{i-1} |\frac{a_{ij}}{a_{ii}}|$ und $\beta_i = \sum_{j=i+1}^{n} |\frac{a_{ij}}{a_{ii}}|$. Dann erhalten wir

$$
\begin{aligned}
x_{m+1,i} - x_i^* &= -\sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}}(x_{m,j} - x_j^*) - \sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}}(x_{m+1,j} - x_j^*) \\
|x_{m+1,i} - x_i^*| &\le \sum_{j=i+1}^{n} |\frac{a_{ij}}{a_{ii}}||x_{m,j} - x_j^*| + \sum_{j=1}^{i-1} |\frac{a_{ij}}{a_{ii}}||x_{m+1,j} - x_j^*| \\
||\mathbf{x}_{m+1} - \mathbf{x}^*||_\infty &\le \underbrace{\sum_{j=i+1}^{n} |\frac{a_{ij}}{a_{ii}}|}_{\beta_{i(m)}} ||\mathbf{x}_m - \mathbf{x}^*||_\infty + \underbrace{\sum_{j=1}^{i-1} |\frac{a_{ij}}{a_{ii}}|}_{\alpha_{i(m)}} ||\mathbf{x}_{m+1} - \mathbf{x}^*||_\infty
\end{aligned}
$$

wobei $i = i(m)$ der Index ist, für den die rechte Summe maximal wird. Aus der letzten Gleichung folgt nacheinander $(i = i(m))$

$$\begin{aligned}
(1-\alpha_i)||\mathbf{x}_{m+1} - \mathbf{x}^*|| &\leq \beta_i ||\mathbf{x}_m - \mathbf{x}^*|| \\
||\mathbf{x}_{m+1} - \mathbf{x}^*|| &\leq \underbrace{\frac{\beta_i}{1-\alpha_i}}_{\leq \rho \text{ s.u.}} ||\mathbf{x}_m - \mathbf{x}^*|| \leq \rho ||\mathbf{x}_m - \mathbf{x}^*|| \\
||\mathbf{x}_m - \mathbf{x}^*|| &\leq \rho^m ||\mathbf{x}_0 - \mathbf{x}^*||
\end{aligned}$$

wobei

$$\rho := \max_i(\alpha_i + \beta_i) = \max_i\Big(\sum_{j\neq i}^{n} \big|\frac{a_{ij}}{a_{ii}}\big|\Big) < 1 \quad \text{wegen Diagonaldominanz}$$

Bleibt z.z.: $\frac{\beta_i}{1-\alpha_i} \leq \alpha_i + \beta_i$.
$0 \leq 1 - \alpha_i - \beta_i \quad \Rightarrow \quad 0 \leq \alpha_i(1-\alpha_i-\beta_i) \quad \Rightarrow \quad \beta_i \leq \alpha_i - \alpha_i^2 + \beta_i - \alpha_i\beta_i \quad \Rightarrow \quad \beta_i \leq (\alpha_i+\beta_i)(1-\alpha_i) \Rightarrow$ Beh. $\square$

Beispiel 23: $\mathbf{Ax} = \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix} \mathbf{x} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, $\mathbf{A}$ *ist diagonaldominant.* $\mathbf{x}^* = (\frac{2}{3}, \frac{2}{3})$, *Sei* $\mathbf{x}_o^T = (1,0)$, *vgl. Beispiel 22.*

$$\begin{aligned}
x_{m+1,i} &= -\sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}} x_{m,j} - \sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} x_{m+1,j} + \frac{b_i}{a_{ii}} \\
x_{m+1,1} &= -\frac{x_{m,2}}{2} + 1, \qquad x_{m+1,2} = -\frac{x_{m+1,1}}{2} + 1
\end{aligned}$$

$x_{1,1} = -\frac{1}{2} \cdot 0 + 1 = 1, \qquad x_{1,2} = -\frac{1}{2} \cdot 1 + 1 = \frac{1}{2}, \qquad \mathbf{x}_1^T = (1, \frac{1}{2})$
$x_{2,1} = -\frac{1}{2} \cdot \frac{1}{2} + 1 = \frac{3}{4}, \qquad x_{2,2} = -\frac{1}{2} \cdot \frac{3}{4} + 1 = \frac{5}{8}, \qquad \mathbf{x}_2^T = (\frac{3}{4}, \frac{5}{8})$
$x_{3,1} = -\frac{1}{2} \cdot \frac{5}{8} + 1 = \frac{11}{16}, \qquad x_{3,2} = -\frac{1}{2} \cdot \frac{11}{16} + 1 = \frac{21}{32}, \qquad \mathbf{x}_3^T = (\frac{11}{16}, \frac{21}{32})$
$x_{4,1} = -\frac{1}{2} \cdot \frac{21}{32} + 1 = \frac{43}{64}, \qquad x_{4,2} = -\frac{1}{2} \cdot \frac{43}{64} + 1 = \frac{85}{128}, \qquad \mathbf{x}_4^T = (\frac{43}{64}, \frac{85}{128})$

2.5 Einfache Iteration

Def. 13 (Jacobi-Matrix), Erinnerung
Sei $f : \mathbb{R}^k \to \mathbb{R}^k$ stetig partiell differenzierbar nach allen k Komponenten. Die Matrix

$$\mathbf{J} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_k} \\ . & \cdots & . \\ \frac{\partial f_k}{\partial x_1} & \cdots & \frac{\partial f_k}{\partial x_k} \end{pmatrix}$$

heißt Jacobi-Matrix (oder Funktionalmatrix).

Bem: die Stetigkeit der Ableitungen wird unten gebraucht um zu zeigen dass Suprema existieren.

Satz 10 (Einfache Iteration)
Sei die Menge $M \subseteq \mathbb{R}^k$ abgeschlossen, konvex, $f : M \to \mathbb{R}^k$, f besitze auf M stetig partielle Ableitungen. Weiter sei $f(\mathbf{x}) + \mathbf{x} \in M \quad \forall \mathbf{x} \in M$ und $\sup_{\mathbf{x}\in M} ||\mathbf{I} + \mathbf{J}(\mathbf{x})|| < 1$, wobei $\mathbf{I}$ die Einheitsmatrix und $\mathbf{J}(\mathbf{x})$ die Jacobi Matrix ist. Dann existiert genau ein $\mathbf{x}^* \in M$ mit $f(\mathbf{x}^*) = 0$ und für beliebige $\mathbf{x}_0 \in M$ konvergiert die Folge $(\mathbf{x}_n)$

$$\mathbf{x}_n := \mathbf{x}_{n-1} + f(\mathbf{x}_{n-1}), \quad n \in \mathbb{N} \tag{2}$$

gegen $\mathbf{x}^*$.

Beweis: (zu Satz 10). Sei $B\mathbf{x} = \mathbf{x} + f(\mathbf{x})$ für $\mathbf{x} \in M$. $B : M \to M$.
Wollen den Banach'schen Fixpunktsatz anwenden, und überprüfen, ob die Abbildung B die Kontraktivitätsbedingung erfüllt.

$$\begin{aligned} ||B\mathbf{x} - B\mathbf{y}|| &= ||\mathbf{x} - \mathbf{y} + f(\mathbf{x}) - f(\mathbf{y})|| \\ &= ||(\mathbf{I} + \mathbf{J}(\mathbf{x} + \underbrace{\vartheta}_{\in(0,1)}(\mathbf{y} - \mathbf{x})))(\mathbf{x} - \mathbf{y})|| \quad (Mittelwertsatz) \\ &\leq \underbrace{|| \sup_{\mathbf{x}\in M} |\mathbf{I} + \mathbf{J}(\mathbf{x}))|||}_{<\alpha<1} \, ||(\mathbf{x} - \mathbf{y})|| < \alpha||(\mathbf{x} - \mathbf{y})|| \end{aligned}$$

Nach dem Banachschen Fixpunktsatz besitzt die Abbildung B genau einen Fixpunkt $\mathbf{x}^*$: $B\mathbf{x}^* = \mathbf{x}^*$. Der Banachsche Fixpunktsatz liefert gleich das Verfahren (2) zum Finden des Fixpunktes. □

Einfache Iteration, Konvergenzgeschwindigkeit
Setzen wir in der Ungleichung $||B\mathbf{x} - B\mathbf{y}|| < \alpha||\mathbf{x} - \mathbf{y}||$:
$\mathbf{x} = \mathbf{x}_{n-1}$ und $\mathbf{y} = \mathbf{x}^*$ so erhalten wir

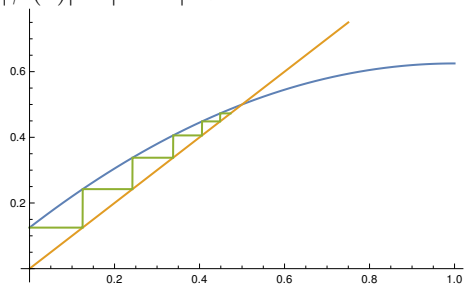
$$||\mathbf{x}_n - \mathbf{x}^*|| < \alpha||\mathbf{x}_{n-1} - \mathbf{x}^*|| < \alpha^n||\mathbf{x}_0 - \mathbf{x}^*||$$

Konvergenz ist linear.

Beispiel 24: $n = 1, M = [a, b], f(x) = 0$
Voraussetzung: $1 + f : M \to M$ und
$\sup |1 + f'(x)| < 1$, also $f' < 0$, d.h. f ist monoton fallend.
Verfahren: $x_n = x_{n-1} + f(x_{n-1}) =: \phi(x_{n-1})$, also $\sup |\phi'(x)| < 1$, d.h. ϕ darf nicht zu steilen Anstieg haben

Beispiel 25: $f(x) = -\frac{1}{2}x^2 + \frac{1}{8}$, $x \in (0, 1)$, $x^* = \frac{1}{2}$
Zunächst ist $\phi(x) := f(x) + x \in (0, 1)$, $|\phi'(x)| = |1 - x| < 1$

$$\begin{aligned} x_{n+1} &= x_n + \left(-\frac{1}{2}x_n^2 + \frac{1}{8}\right) \\ x_0 &= 0, \quad x_1 = \frac{1}{8}, \quad x_2 = \frac{31}{128}, \\ x_3 &= \frac{127}{256}, \quad \ldots, \quad x_8 \approx 0.49 \end{aligned}$$



Beispiel 26: $n = 2$, $M = [0, 1]^2$, $f(x_1, x_2) = \mathbf{0}$

$$\begin{aligned} f_1(x_1, x_2) &= x_1\left(\frac{1}{8}x_1 - 1\right) + \frac{1}{16}x_2 \\ f_2(x_1, x_2) &= \left(\frac{1}{16}x_1x_2 - 1\right)x_2 \end{aligned}$$

Man prüft leicht nach, dass $\mathbf{x} + f(\mathbf{x}) \in M \quad \forall \mathbf{x} \in M$.

$$\text{Jacobi-Matrix:} \qquad \mathbf{J}(x_1, x_2) = \begin{pmatrix} \frac{1}{4}x_1 - 1 & \frac{1}{16} \\ \frac{1}{16}x_2^2 & \frac{1}{8}x_1x_2 - 1 \end{pmatrix}$$

Damit

$$\mathbf{I} + \mathbf{J}(x_1, x_2) = \begin{pmatrix} \frac{1}{4}x_1 & \frac{1}{16} \\ \frac{1}{16}x_2^2 & \frac{1}{8}x_1x_2 \end{pmatrix}$$

$$\sup|\mathbf{I}+\mathbf{J}(x_1,x_2)| = \begin{pmatrix} \frac{1}{4} & \frac{1}{16} \\ \frac{1}{16} & \frac{1}{8} \end{pmatrix}$$

$$||\sup|\mathbf{I}+\mathbf{J}(x_1,x_2)|||<1$$

$$\begin{aligned}
\mathbf{x}_0 &= (1,1) \qquad \text{am weitesten von der Lösung } (0,0) \text{ entfernt} \\
\mathbf{x}_1 &= \mathbf{x}_0 + f(\mathbf{x}_0) = (\frac{3}{16},\frac{1}{16}) \\
\mathbf{x}_2 &= \mathbf{x}_1 + f(\mathbf{x}_1) \approx (0.0083, 0.000046) \\
\mathbf{x}_3 &\approx (0.0000114739, 1.08713 * 10^{-12})
\end{aligned}$$

Probleme bei der einfachen Iteration:

- geeignete Menge M, so dass $x + f(x) \in M$
- geeignete Norm, so dass $||B|| < 1$

2.6 Modifizierte Einfache Iteration

Hinweis: Der folgende Satz ist zur Information, vgl. Bemerkung im Anschluss an den Beweis.

Satz 11 (modifizierte Einfache Iteration)
Sei die Menge $M = \mathbb{R}^k$. $f : \mathbb{R}^k \to \mathbb{R}^k$, erfülle folgende Bedingung: $\exists L > 0, \gamma > 0$ so dass $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^k$

$$\begin{aligned}
||f(\mathbf{x}) - f(\mathbf{y})||_2 &\leq L||\mathbf{x}-\mathbf{y}||_2 \qquad \text{(Lipschitz-Stetigkeit)} \\
\big(f(\mathbf{x}) - f(\mathbf{y}), \mathbf{x}-\mathbf{y}\big) &\geq \gamma||\mathbf{x}-\mathbf{y}||_2^2 \qquad \text{(starke Monotonie)}
\end{aligned}$$

Dann existiert genau ein $\mathbf{x}^* \in \mathbb{R}^k$ mit $f(\mathbf{x}^*) = 0$ und für beliebige $\mathbf{x}_0 \in \mathbb{R}^k$ konvergiert die Folge $\{\mathbf{x}_n\}$

$$\mathbf{x}_n := \mathbf{x}_{n-1} + \beta f(\mathbf{x}_{n-1}), \quad n \in \mathbb{N}, \tag{3}$$

falls $\beta \in (-\frac{2\gamma}{L^2}, 0)$, gegen $\mathbf{x}^*$.

Beweis zu Satz 11
Sei $B\mathbf{x} := \mathbf{x} + \beta f(\mathbf{x})$ für $\mathbf{x} \in \mathbb{R}^k$ und $\beta \in (-\frac{2\gamma}{L^2}, 0)$.
Wollen wieder den Bachach'schen Fixpunktsatz anwenden, und überprüfen, ob die Kontraktivitätsbedingung erfüllt ist.

$$\begin{aligned}
||B\mathbf{x} - B\mathbf{y}||_2^2 &= ||\mathbf{x}-\mathbf{y}+\beta(f(\mathbf{x})-f(\mathbf{y}))||_2^2 \\
&= \langle \mathbf{x}-\mathbf{y}+\beta(f(\mathbf{x})-f(\mathbf{y}), \mathbf{x}-\mathbf{y}+\beta(f(\mathbf{x})-f(\mathbf{y})\rangle \\
&= ||\mathbf{x}-\mathbf{y}||_2^2 + 2\underbrace{\beta(f(\mathbf{x})-f(\mathbf{y}))(\mathbf{x}-\mathbf{y})}_{\leq \beta\gamma||\mathbf{x}-\mathbf{y}||_2^2 \quad \text{da} \quad \beta<0} \\
&\quad +\beta^2 \underbrace{||f(\mathbf{x})-f(\mathbf{y})||_2^2}_{\leq L^2||\mathbf{x}-\mathbf{y}||_2^2} \\
&\leq \underbrace{(1+2\beta\gamma+L^2\beta^2)}_{=:g(\beta)<1} ||\mathbf{x}-\mathbf{y}||_2^2
\end{aligned}$$

Kurvendiskussion $g(\beta)$:

$$g(-\frac{2\gamma}{L^2}) = g(0) = 1, \quad g'(\beta) = 2\gamma + 2\beta L^2 = 0, g''(\beta) = 2L^2 > 0$$

Minimalstelle an der Stelle $\beta = -\frac{\gamma}{L^2}$.

Konvergenzgeschwindigkeit
der modifizierten einfachen Iteration ist, wie bei der einfachen Iteration, linear

Bem.:
Wenn wir ein β finden so dass $||\mathbf{I} + \beta\mathbf{J}(\mathbf{x})|| < 1 \quad \forall \mathbf{x} \in M$ dann können wir uns die Überprüfung der anderen Voraussetzungen von Satz 11 sparen. Aber dann muss $\mathbf{x} + \beta f(\mathbf{x}) \in M$ sein.

Beispiel 27: Wir suchen eine positive Nullstelle $x \geq 0$ von

$$f(x) := \sqrt{x+1} + \frac{\sin x}{10} - 2.$$

Wähle $M = [0, 8]$. Wir haben für $x \in M$:

$$\frac{1}{15} = \frac{1}{6} - \frac{1}{10} \leq f'(x) = \frac{1}{2\sqrt{x+1}} + \frac{\cos x}{10} \leq \frac{1}{2} + \frac{1}{10} = \frac{3}{5}.$$

versuchen β so zu finden, dass $\sup|1 + \beta f'(x)| < 1$, d.h. $-2 < \beta f'(x) < 0 \quad \forall x \in M$. Wegen obiger Ungleichung gilt

$$-\frac{10}{3} < \beta < 0$$

Wir können somit $\beta := -3$ setzen und $\phi(x) := x - 3f(x) = 6 + x - 3\sqrt{x+1} - \frac{3}{10}\sin x$. Damit ist $|\phi'(x)| \leq 4/5 = 0.8$. Die Abbildung ist also kontraktiv.

Fortsetzung von Beispiel 27
Um zu zeigen, dass $\phi(M) \subset M$, betrachte zuerst $g(x) := x - 3\sqrt{x+1}$ und zeige, dass g sein Minimum auf M in $x = 5/4$ mit $g(5/4) = -13/4$ und sein Maximum in $x = 8$ mit $g(8) = -1$ annimmt. Somit gilt für $x \in M$:

$$0 \leq 6 - \frac{13}{4} - \frac{3}{10} \leq \phi(x) \leq 6 - 1 + \frac{3}{10} \leq 8.$$

Also konvergiert die Fixpunkiteration (x_n), $x_{n+1} = \phi(x_n)$ für beliebiges x_0 gegen die (eindeutig bestimmte) Nullstelle $x^* \approx 2.908$ von f auf M.

x_0	:=	8	x_4	=	3.30334	x_8	=	2.93364
x_1	=	4.70319	x_5	=	3.07067	x_9	=	2.92168
x_2	=	3.83877	x_6	=	2.99664	x_{10}	=	2.91527
x_3	=	3.43223	x_7	=	2.95593	x_{11}	=	2.91184

2.7 Newton Verfahren

Der Startwert x_0 bei der Iteration nach dem Newton-Verfahren muss nahe genug an der Nullstelle sein.
Betrachten zunächst den Fall der Dimension $n = 1$ und entwickeln $f(x)$ in der Umgebung von x_0 in eine Taylorreihe. Setzen dazu voraus: f zweimal (stetig) differenzierbar.

$$f(x) = f(x_o) + f'(x_0)(x - x_0) + \frac{f''(x_0 + \vartheta(x - x_0))}{2!}(x - x_0)^2$$

wobei $\vartheta \in (0, 1)$.
Idee: Betrachten statt $f(x) = 0$: $f(x_o) + f'(x_0)(x - x_0) = 0$ nur den linearen Anteil, was uns auf $x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$ führt.
Die Iteration

$$x_{m+1} = x_m - \frac{f(x_m)}{f'(x_m)}, \quad m = 1, \ldots,$$

beruht auf dieser Linearisierung.

Im Fall höherer Dimension $n \geq 1$ ist das Verfahren entsprechend

$$\mathbf{x}_{m+1} = \mathbf{x}_m - \mathbf{J}^{-1}(\mathbf{x}_m) f(\mathbf{x}_m), \quad m = 1, \ldots,$$

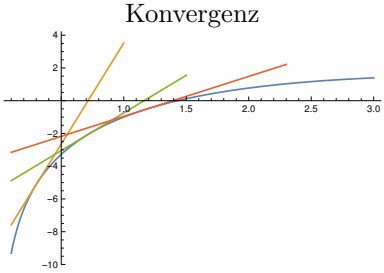
wobei $\mathbf{J}$ die Jacobi Matrix ist.
Im Vergleich zu den bisherigen Verfahren ($\mathbf{x}_{m+1} = \mathbf{x}_m + \mathbf{C} f(\mathbf{x}_m)$) mit *festem* $\mathbf{C}$ haben wir jetzt ein *variables* $\mathbf{C}(\mathbf{x}) = \mathbf{J}^{-1}(\mathbf{x})$.

Satz 12 (Newton-Raphson Verfahren)
Sei $f : M \to \mathbb{R}^n$, stetig differenzierbar, und $\exists \mathbf{x}^* \in int(M)$ mit $f(\mathbf{x}^*) = \mathbf{0}$ und $\mathbf{J}(\mathbf{x}^*)$ regulär. Dann $\exists \delta > 0$, so dass für $\mathbf{x}_0 \in U_\delta(\mathbf{x}^*) = \{\mathbf{x} | \mathbf{x} \in M, ||\mathbf{x} - \mathbf{x}^*|| < \delta\} \subset M$ die Folge $(\mathbf{x}_m)$ mit

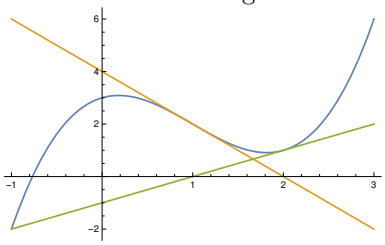
$$\mathbf{x}_{m+1} = \mathbf{x}_m - \mathbf{J}^{-1}(\mathbf{x}_m) f(\mathbf{x}_m), \quad m \in \mathbb{N}$$

gebildet werden kann und gegen $\mathbf{x}^*$ konvergiert.

Konvergenz



keine Konvergenz



Beweis: Wir führen den Beweis für den Fall $n = 1$. Es gilt:

$$\begin{aligned}
x_{m+1} - x^* &= x_m - f'^{-1}(x_m)(f(x_m) - \underbrace{f(x^*)}_{=0}) - x^* \\
&= x_m - x^* - f'^{-1}(x_m) f'(x_m + \vartheta(x^* - x_m))(x_m - x^*) \\
&= \underbrace{(1 - f'^{-1}(x_m) f'(x_m + \vartheta(x^* - x_m))}_{(*)}(x_m - x^*)
\end{aligned}$$

Konvergenz wenn $|(*)| < 1$. Sei $\varepsilon > 0$. Da f' stetig in x_m $\exists \delta > 0 : |f'(x_m + \vartheta(x^* - x_m)) - f'(x_m)| < \varepsilon$ falls $|\vartheta(x_m - x^*)| < \delta$.
Damit wird $|\frac{f'(x_m) - f'(x_m + \vartheta(x^* - x_m))}{f'(x_m)}| < \frac{\epsilon}{|f'(x_m)|} < 1$ und demzufolge $|(*)| < 1$ wenn ε hinreichend klein. □
Für den Fall $n \geq 1$ bekommen wir analog, wieder unter Anwendung des Mittelwertsatzes, diesmal für Funktionen von mehr Veränderlichen,

$$\begin{aligned}
\mathbf{x}_{m+1} - \mathbf{x}^* &= (\mathbf{I} - \mathbf{J}^{-1}(\mathbf{x}_m)\mathbf{J}(\mathbf{x}_m + \vartheta(\mathbf{x}^* - \mathbf{x}_m))(\mathbf{x}_m - \mathbf{x}^*) \\
&= \mathbf{J}^{-1}(\mathbf{x}_m)(\mathbf{J}(\mathbf{x}_m) - \mathbf{J}(\mathbf{x}_m + \vartheta(\mathbf{x}^* - \mathbf{x}_m))(\mathbf{x}_m - \mathbf{x}^*) \\
||\mathbf{x}_{m+1} - \mathbf{x}^*|| &\leq ||\mathbf{J}^{-1}(\mathbf{x}_m)||^{-1} ||(\mathbf{J}(\mathbf{x}_m) - \mathbf{J}(\mathbf{x}_m + \vartheta(\mathbf{x}^* - \mathbf{x}_m))|| \cdot \\
&\qquad \cdot ||\mathbf{x}_m - \mathbf{x}^*|| \\
&\leq ||\mathbf{J}^{-1}(\mathbf{x}_m)||^{-1} \cdot \varepsilon
\end{aligned}$$

für $||\mathbf{x}_m - (\mathbf{x}_m - \vartheta(\mathbf{x}^* - \mathbf{x}_m))|| \leq \delta$ wegen Stetigkeit von $\mathbf{J}$ (alle partiellen Ableitungen werden als stetig vorausgesetzt).

Beispiel 28: Nullstellen von $f(x) = 4\log(x) - x$
Die Nullstellen sind $x_1^* = 1.42961$ und $x_2^* = 8.61317$. Setzen den Startwert $x_o := 0.3$.

$$
\begin{aligned}
f'(x) &= \frac{4}{x} - 1 \\
x_1 &:= x_0 - \frac{f(x_0)}{f'(x_0)} = 0.3 - \frac{f(0.3)}{f'(0.3)} = 0.714802 \\
x_2 &:= x_1 - \frac{f(x_1)}{f'(x_1)} = 0.714802 - \frac{f(0.714802)}{f'(0.714802)} = 1.16254 \\
x_3 &:= x_2 - \frac{f(x_2)}{f'(x_2)} = 1.16254 - \frac{f(1.16254)}{f'(1.16254)} = 1.39202 \\
x_4 &:= x_3 - \frac{f(x_3)}{f'(x_3)} = 1.39202 - \frac{f(1.39202)}{f'(1.39202)} = 1.42885 \\
x_5 &:= x_4 - \frac{f(x_4)}{f'(x_4)} = 1.42885 - \frac{f(1.42885)}{f'(1.42885)} = 1.42961
\end{aligned}
$$

Beispiel 29: Nullstellen von $f(x) = x^3 - 3x^2 + x + 3$
Die einzige reelle Nullstelle ist $x^* = -0.769292$. Setzen wir den Startwert auf $x_0 := 1$, so

$$
\begin{aligned}
f'(x) &= 3x^2 - 6x + 1 \\
x_1 &:= x_0 - \frac{f(x_0)}{f'(x_0)} = 1 - \frac{f(1)}{f'(1)} = 1 - \frac{2}{-2} = 2 \\
x_2 &:= x_1 - \frac{f(x_1)}{f'(x_1)} = 2 - \frac{f(2)}{f'(2)} = 2 - \frac{1}{1} = 1 = x_0 \\
x_3 &:= x_1, \quad x_4 = x_2 \ldots
\end{aligned}
$$

Der Startwert ist offenbar zu weit entfernt von der Nullstelle.

Konvergenzgeschwindigkeit des Newton Verfahrens ($n = 1$)
Taylorreihenentwicklung ($f \in C_2[a,b], f'(x) \neq 0 \ \forall x \in [a,b]$)

$$
\underbrace{f(x^*)}_{=0} = f(x_m) + f'(x_m)(x^* - x_m) + \frac{f''(x_m + \vartheta(x^* - x_m))}{2!}(x^* - x_m)^2
$$

woraus folgt ($\theta := x_m + \vartheta(x^* - x_m)$)

$$
\begin{aligned}
-\frac{f(x_m)}{f'(x_m)} &= x^* - x_m + \frac{1}{2}\frac{f''(\theta)}{f'(x_m)}(x^* - x_m)^2 \\
|x_{m+1} - x^*| &= \left|x_m - x^* - \frac{f(x_m)}{f'(x_m)}\right| = \Big|\underbrace{\frac{1}{2}\frac{f''(\theta)}{f'(x_m)}}_{\leq K}\Big| |(x^* - x_m)|^2 \\
&\leq K|x^* - x_m|^2
\end{aligned}
$$

d.h. wir haben hier quadratische Konvergenz.

Vorteil: quadratische Konvergenz
Nachteil: Rechenaufwand

- in jedem Schritt ist die Ableitung oder sogar,
- im Fall $n > 1$, die Inverse der Jacobi-Matrix zu bestimmen.

Es gibt eine Reihe von Modifikationen des Newton-Verfahrens,
z.B. kann man statt jedesmal die Jacobi-Matrix (und deren Inverse) neu zu berechnen, $\mathbf{J}(x_0)$ verwenden.
ÜA: Führen Sie den Beweis für diese Modifikation von Satz 12 für den Fall $n = 1$ und untersuchen Sie die Konvergenzgeschwindigkeit!

Eine andere Möglichkeit ist, in jedem Schritt, das lineare Gleichungssystem

$$\mathbf{J}\Delta\mathbf{x}_m = -f(\mathbf{x}_m)$$

(numerisch) zu lösen und dann $\mathbf{x}_{m+1} = \mathbf{x}_m + \Delta\mathbf{x}_m$ zu setzen.

Beispiel 30: $f\begin{pmatrix} y \\ z \end{pmatrix} := \begin{pmatrix} 2y - z \\ 6y + z + 2z^2 \end{pmatrix}$, $\mathbf{J}\begin{pmatrix} y \\ z \end{pmatrix} = \begin{pmatrix} 2 & -1 \\ 6 & 1+4z \end{pmatrix}$

Für 2×2-Matrizen: $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \overset{\det A \neq 0}{\Longrightarrow} A^{-1} = \dfrac{1}{ad - bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$

$\overset{z \neq -1}{\Rightarrow} \left(\mathbf{J}\begin{pmatrix} y \\ z \end{pmatrix}\right)^{-1} = \dfrac{1}{8(z+1)}\begin{pmatrix} 1+4z & 1 \\ -6 & 2 \end{pmatrix}$

Wähle $\mathbf{x}_0 := \begin{pmatrix} y \\ z \end{pmatrix} = \begin{pmatrix} -2 \\ -3 \end{pmatrix}$.

$$\begin{aligned}
\mathbf{x}_1 &= x_0 - \mathbf{J}(\mathbf{x}_0)^{-1} f(\mathbf{x}_0) = \begin{pmatrix} -2 \\ -3 \end{pmatrix} + \frac{1}{16}\begin{pmatrix} -11 & 1 \\ -6 & 2 \end{pmatrix}\begin{pmatrix} -1 \\ 3 \end{pmatrix} \\
&= -\frac{1}{8}\begin{pmatrix} 9 \\ 18 \end{pmatrix} = \begin{pmatrix} -1.125 \\ -2.25 \end{pmatrix} \\
\mathbf{x}_2 &= \begin{pmatrix} -1.0125 \\ -2.025 \end{pmatrix}, \quad \mathbf{x}_3 \approx \begin{pmatrix} -1.000152 \\ -2.000309 \end{pmatrix}
\end{aligned}$$

(x_n) konvergiert gegen die Nullstelle $\begin{pmatrix} -1 \\ -2 \end{pmatrix}$.

2.8 Sekantenverfahren (Regula Falsi)

verwenden, im Gegensatz zum Newton-Verfahren, anstelle der Ableitung $f'(x_k)$ eine Approximation, den Differenzenquotienten

$$a_k := \frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}}$$

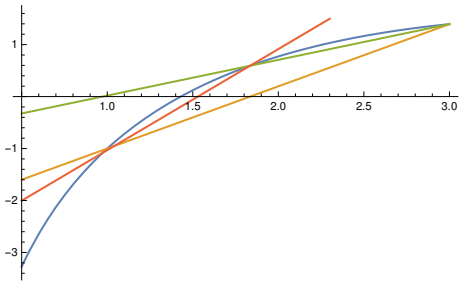
Diskretisiertes Newton-Verfahren

$$x_{k+1} = x_k - \frac{f(x_k)}{a_k}$$

Geometrische Interpretation: anstelle der **Tangente** wie beim Newton-Verfahren wird die **Sekante** durch die Punkte $(x_k, f(x_k))$ und $(x_{k-1}, f(x_{k-1}))$ genommen und der Schnittpunkt mit der x-Achse ist ist neue Wert x_{k+1}.
Konvergenz: $|x_{k+1} - x^*| \leq C|x_k - x^*||x_{k-1} - x^*|$ (siehe z.B. Stoer, Abschnitt 5.9)

Beispiel 31: Nullstellen von $f(x) = 4\log(x) - x$
Die Nullstellen sind $x_1^* = 1.42961$ und $x_2^* = 8.61317$. Setzen die Startwerte $x_o := 1, x_1 := 3$.

$$
\begin{aligned}
a_1 &:= \frac{f(x_1)-f(x_0)}{x_1-x_0} \approx 1.197 \\
x_2 &:= x_1 - \frac{f(x_1)}{a_1} \approx 1.835 \\
x_3 &:= x_2 - \frac{f(x_2)}{a_2} \approx 0.972 \\
x_4 &:= x_3 - \frac{f(x_3)}{a_3} \approx 1.53
\end{aligned}
$$



Reihenfolge der Sekanten: gelb, grün, rot.

2.9 Nullstellenbestimmung für Polynome

Sei $P(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots a_1 x + a_0$ Polynom n-ten Grades. Anwendung des Newton-Verfahrens

$$x_{m+1} = x_m - \frac{P(x_m)}{P'(x_m)}$$

verlangt die Entwicklung von $P(x)$ und $P'(x)$ an den Stellen x_m. Dazu setzen wir $\bar{x} := x_m$ und teilen $P(x)$ durch $(x - \bar{x})$:

$$
\begin{aligned}
P(x) &= a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 \\
&= \underbrace{(b_{n-1}x^{n-1} + b_{n-2}x^{n-2} + \cdots + b_0)}_{Q(x)}(x-\bar{x}) + b_{-1} \\
b_{n-1} &= a_n \\
b_j &= a_{j+1} + b_{j+1}\bar{x}, \quad j = n-2, \ldots, -1 \\
P(\bar{x}) &= b_{-1}
\end{aligned}
$$

wie man leicht durch Koeffizientenvergleich überprüft.
Dasselbe machen wir mit dem Polynom $Q(x)$:

$$
\begin{aligned}
Q(x) &= b_{n-1}x^{n-1} + b_{n-2}x^{n-2} + \cdots + b_1 x + b_0 \\
&= (c_{n-2}x^{n-2} + c_{n-3}x^{n-3} + \cdots + c_1 x + c_0)(x-\bar{x}) + c_{-1} \\
c_{n-2} &= b_{n-1} \\
c_{n-3} &= b_{n-2} + c_{n-2}\bar{x} \\
c_j &= b_{j+1} + c_{j+1}\bar{x}, \quad j = -1, \ldots, n-3 \\
Q(\bar{x}) &= c_{-1}
\end{aligned}
$$

$$
\begin{aligned}
P(x) &= Q(x)(x-\bar{x}) + b_{-1} \\
P'(x) &= Q'(x)(x-\bar{x}) + Q(x) \\
P'(\bar{x}) &= Q(\bar{x}) = c_{-1}.
\end{aligned}
$$

Def. 14 (Horner Schema)
Der Algorithmus zur Berechnung der Koeffizienten heißt Horner-Schema.

Bem.: In jedem Iterationsschritt haben wir also etwa je $2n$ Additionen und Multiplikationen.

Bem.: Natürlich konvergiert das Verfahren nur, wenn der Startwert hinreichend nahe der Nullstelle ist.

Beispiel 32: $P(x) = 4x^4 - 3x^3 + 2x^2 - x + 1$, *an der Stelle* $\bar{x} = 1$.

$$\begin{aligned}
b_3 &= a_4 = 4 \\
b_2 &= a_3 + b_3\bar{x} = -3 + 4 \cdot 1 = 1 \\
b_1 &= a_2 + b_2\bar{x} = 2 + 1 \cdot 1 = 3 \\
b_0 &= a_1 + b_1\bar{x} = -1 + 3 \cdot 1 = 2 \\
b_{-1} &= a_0 + b_0\bar{x} = 1 + 2 \cdot 1 = 3 = P(1) \\
Q(x) &= 4x^3 + x^2 + 3x + 2 \\
c_2 &= b_3 = 4 \\
c_1 &= b_2 + c_2\bar{x} = 1 + 4 \cdot 1 = 5 \\
c_0 &= b_1 + c_1\bar{x} = 3 + 5 \cdot 1 = 8 \\
c_{-1} &= b_0 + c_0\bar{x} = 2 + 8 \cdot 1 = 10 = Q(1) = P'(1)
\end{aligned}$$

$x_o = 1, \quad x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} = 1 - \frac{3}{10} = \frac{7}{10}$

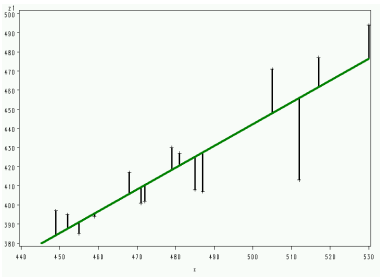
3 Ausgleichs- und Glättungsverfahren

3.1 Lineare Regression

Wir betrachten hier nur den Fall einer Einflussvariablen und einer Zielvariablen.

Einfache lineare Regression

$$Y_i = \theta_0 + \theta_1 X_i + \epsilon_i, \qquad \epsilon_i \sim (0, \sigma^2)$$



Die Summe der Quadrate der Länge der Streckenabschnitte soll minimal werden.

Minimumaufgabe:

$$\min_{\theta_0,\theta_1} \quad \sum_{i=1}^{n} (Y_i - \theta_0 - \theta_1 X_i)^2$$

Ableiten nach θ_0 und θ_1 und Nullsetzen:

$$\begin{aligned}
\sum_{i=1}^{n} (Y_i - (\theta_1 X_i + \theta_0)) \cdot X_i &= 0 \\
\sum_{i=1}^{n} (Y_i - (\theta_1 X_i + \theta_0)) \cdot 1 &= 0
\end{aligned}$$

$\Rightarrow$

$$\begin{aligned}
\sum_i X_i Y_i - \theta_1 \sum_i X_i^2 - \theta_0 \sum_i X_i &= 0 \\
\sum_i Y_i - \theta_1 \sum_i X_i - \theta_0 \cdot n &= 0
\end{aligned}$$

Die zweite Gleichung nach θ_0 auflösen:

$$\theta_0 = \frac{1}{n}\sum_i Y_i - \theta_1 \frac{1}{n}\sum_i X_i$$

und in die erste einsetzen:

Kleinste Quadrat-Schätzung

$$\sum_i X_i Y_i - \theta_1 \sum_i X_i^2 - \frac{1}{n}\sum_i Y_i \sum_i X_i + \theta_1 \frac{1}{n}\sum_i X_i \sum_i X_i = 0$$

$$\sum_i X_i Y_i - \frac{1}{n}\sum_i Y_i \sum_i X_i - \theta_1 \left(\sum_i X_i^2 - \frac{1}{n}\sum_i X_i \sum_i X_i\right) = 0$$

$\Rightarrow$

$$\begin{aligned}
\hat{\theta}_1 &= \frac{\sum_i X_i Y_i - \frac{1}{n}\sum_i X_i \sum_i Y_i}{\sum_i X_i^2 - \frac{1}{n}(\sum_i X_i)^2} = \frac{\frac{1}{n-1}\left(\sum_i X_i Y_i - \frac{1}{n}\sum_i X_i \sum_i Y_i\right)}{\frac{1}{n-1}\left(\sum_i X_i^2 - \frac{1}{n}(\sum_i X_i)^2\right)} \\
\hat{\theta}_0 &= \frac{1}{n}\left(\sum_i Y_i - \hat{\theta}_1 \sum_i X_i\right)
\end{aligned}$$

Dabei sind Zähler und Nenner bei θ_1 sogenannte empirische Kovarianz zwischen X und Y bzw. empirische Varianz von X. (Zur Definition von Varianz und Kovarianz siehe unten.)

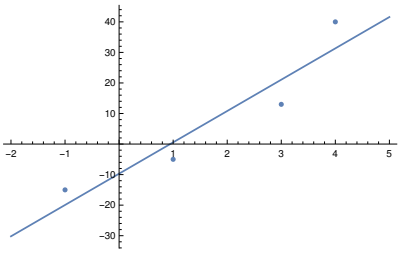
Beispiel 33: Regressionsgerade durch die Punkte $(-1,-15),(1,-5),(3,13),(4,40)$

$$\min_{\theta_0,\theta_1}\left((-15-(\theta_0-\theta_1))^2 + (-5-(\theta_0+\theta_1))^2 + (13-(\theta_0+3\theta_1))^2 + (40-(\theta_0+4\theta_1))^2\right)$$

Ableiten und Nullsetzen

$$\begin{aligned}
-66 + 8\theta_0 + 14\theta_1 &= 0 \\
-418 + 14\theta_0 + 54\theta_1 &= 0
\end{aligned}$$

$\theta_0 = -\frac{572}{59}, \theta_1 = \frac{605}{59}$



3.2 Nichtlineare Regression

Quasilineare Regression

z.B. Polynomregression $Y_i = a_0 + a_1 x_i + a_2 x_i^2 + a_3 x_i^3 + \epsilon_i$ wird auf lineare Regression zurückgeführt $x_{ij} := x_i^j$

Echt nichtlineare Regression, z.B. Wachstumskurven

$$\begin{aligned}
y &= \alpha + \frac{\gamma}{1+exp(-\beta(x-\mu))} && \text{logistische Funktion} \\
y &= \alpha + \gamma exp(-exp(-\beta(x-\mu))) && \text{Gompertzfunktion.} \\
y &= x\tan\beta - \frac{g}{2v_0^2\cos^2\beta}x^2 && \text{Wurfparabel}
\end{aligned}$$

zu schätzende Parameter: $\alpha, \beta, \gamma, \mu, v_0$ (g: Fallbeschleunigung)

Modell, f wird als bekannt angenommen

$$\begin{aligned}
Y &= f(x,\theta) + \epsilon \qquad \epsilon \sim (0,\sigma^2) \\
\mathbf{Y} &= \mathbf{F}(\mathbf{X},\boldsymbol{\theta}) + \boldsymbol{\epsilon}
\end{aligned}$$

$$L(\boldsymbol{\theta}) = \boldsymbol{\epsilon}'\boldsymbol{\epsilon} = \sum_i (Y_i - \mathbf{F}(\mathbf{X}_i,\boldsymbol{\theta}))^2 \longrightarrow min_\theta$$

Ableiten und Nullsetzen führt auf ein i.A. nichtlineares Gleichungssystem.
Dazu werden Iterationsverfahren verwendet (siehe oben).

3.3 Glättende Splines

Def. 15 (Spline)
Sei $M = [a, b] \subseteq \mathbb{R}$. Die Funktion

$$s(t) = \sum_{i=0}^{r-1} \theta_i t^i + \sum_{i=1}^{n} \vartheta_i (t - t_i)_+^{r-1}$$

heißt $(r-1)$-Spline mit Knoten in $t_1, \ldots, t_n \in M$, $a = t_1 \leq, \ldots, \leq t_n = b$.
Dabei sind $\theta_0, \ldots, \theta_{r-1}, \vartheta_1, \ldots, \vartheta_n$ irgendwelche reelle Koeffizienten.
Die Plus-Funktion ist definiert als:

$$x_+ = \begin{cases} x & \text{falls} \quad x > 0 \\ 0 & \text{sonst} \end{cases} = \max(x, 0)$$

Wir betrachten nur den Fall $r = 4$, d.h. kubische Splines.

Satz 13
Die Funktion $s(t)$ ist ein kubischer Spline gdw. $s(t)$ folgende Eigenschaften hat

- $s(t)$ ist stückweises Polynom vom Grad 3 auf $[t_i, t_{i+1})$.
- $s(t)$ ist zweimal stetig differenzierbar, d.h. $s(t) \in W_2(M)$.
- $s'''(t)$ ist Treppenfunktion mit Stufen in $t_1, \ldots, t_n$.

Beweis: Die Hinrichtung ist trivial. Die Rückrichtung ist schwieriger.

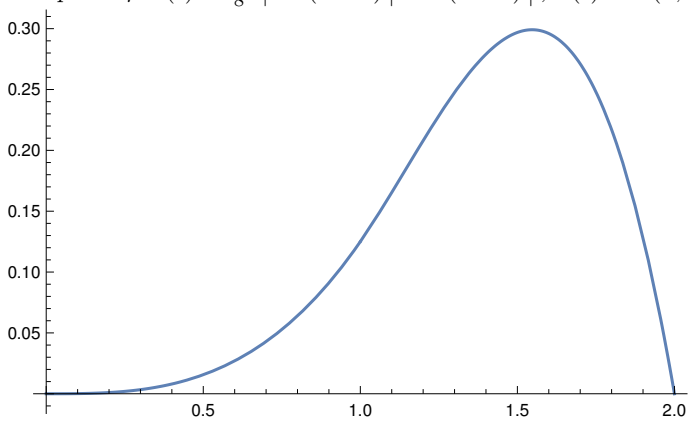
Def. 16 (natürlicher kubischer Spline)
Ein kubischer Spline heißt natürlicher kubischer Spline, falls es die sogenannten natürlichen Randbedingungen

$$s''(a) = s''(b) = s'''(a) = s'''(b) = 0.$$

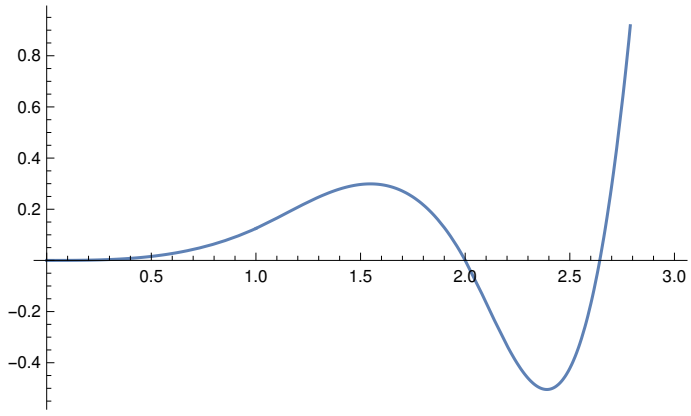
erfüllt.

Satz 14 (Basisfunktionen)
Die Menge $\{1, t, t^2, t^3, (t - t_i)_+^3, i = 1, \ldots, n\}$ bildet eine Basis für den Vektorraum $S(t_1, \ldots, t_n)$ der Splinefunktionen mit Knoten in $t_1, \ldots, t_n$.

Beweis: ÜA. Zeigen Sie: $S(t_1, \ldots, t_n)$ ein Vektorraum der Dimension $n + 4$. Die Funktionen $1, t, t^2, t^3, (t - t_i)_+^3, i = 1, \ldots, n$ sind linear unabhängig
Beispiel 34: $s(t) = \frac{1}{8} t_+^3 - (t-1)_+^3 + 8(t-2)_+^3$, $s(t) \in S(0, 1, 2)$



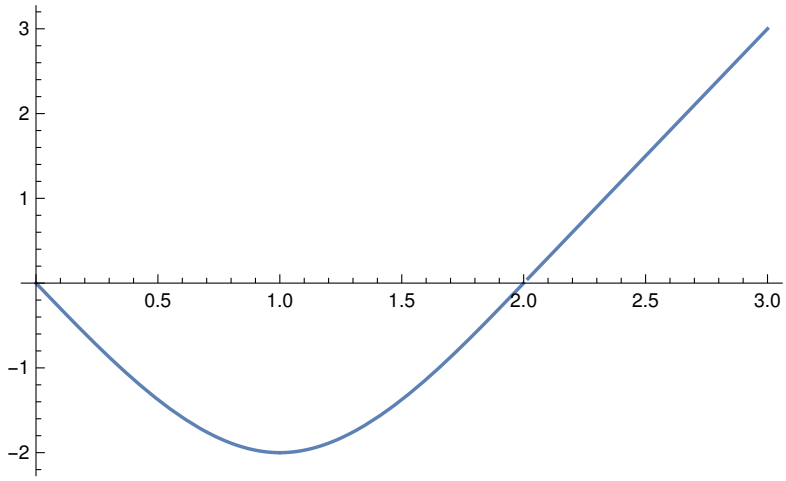
Beispiel 35: $s(t) = \frac{1}{8} t_+^3 - (t-1)_+^3 + 8(t-2)_+^3 + (t-3)_+^3$



$s(t) \in S(0,1,2,3)$

Beispiel 36: $s(t) = t_+^3 - 2(t-1)_+^3 + (t-2)_+^3 + (t-3)_+^3$

$$\begin{aligned}
s'(t) &= 3t_+^2 - 6(t-1)_+^2 + 3(t-2)_+^2 + 3(t-3)_+^2 \\
s''(t) &= 6t_+ - 12(t-1)_+ + 6(t-2)_+ + 6(t-3)_+ \\
s'''(t) &= 6 \cdot 1_{t>0} - 12 \cdot 1_{t>1} + 6 \cdot 1_{t>2} + 6 \cdot 1_{t>3} \\
s''(0) &= 0, \quad s''(3) = 6 \cdot 3 - 12 \cdot 2 + 6 = 0 \\
s'''(0) &= 0, \quad s'''(3) = 6 - 12 + 6 = 0 \quad \Rightarrow s(t) \in NS(0,1,2,3)
\end{aligned}$$



Satz 15 (Vektorraum der natürlichen Splinefunktionen)
Der Vektorraum $NS(t_1, \ldots, t_n)$ der natürlichen Splinefunktionen mit Knoten in $t_1, \ldots, t_n$ ist ein Teilraum von $S(t_1, \ldots, t_n)$ und hat die Dimension n.

Beweis: ÜA. Analog zu Satz 13 ist $NS(t_1, \ldots, t_n)$ ein Vektorraum. Wir haben 4 linear unabhängige Restriktionen an den Spline, $s''(a) = s''(b) = s'''(a) = s'''(b) = 0$, die die Dimension von $n+4$ auf n reduzieren. Insbesondere gilt $\theta_2 = \theta_3 = 0$, $\sum_{i=1}^{n-1} \vartheta_i = \sum_{i=1}^{n-1} \vartheta_i t_i = 0$.

Warum sind natürliche Splines so interessant?

Satz 16
Seien die Datenpunkte (t_i, y_i), $i = 1, \ldots, n$ gegeben, sei $f \in W_2$ (d.h. zweimal stetig differenzierbar) und sei $\lambda > 0$. Lösung der Minimumaufgabe

$$\min_f \left(\frac{1}{n} \sum_{i=1}^{n} (y_i - f(t_i))^2 + \lambda \int_a^b f''(t)^2 \, dt \right) \tag{4}$$

ist ein natürlicher glättender Spline mit Knoten in $t_1, \ldots, t_n$. Die Lösung ist, bei gegebenem λ, eindeutig.

Bem.: λ ist ein sogenannter Glättungsparameter, der "Nichtglattheit" bestraft.
$\lambda \to 0$: Lösung, die (4) minimiert ist ein interpolierender Spline.

Beweis von Satz 16: Sei $s(t)$ eine Funktion, die in (4) das Minumimum annimmt, und seien $\delta \neq 0$ und $f \in W_2$ beliebig. Dann gilt:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - (s(t_i) - \delta f(t_i)))^2 + \lambda \int_a^b (s''(t) - \delta f''(t))^2\, dt$$
$$\geq \frac{1}{n}\sum_{i=1}^{n}(y_i - s(t_i))^2 + \lambda \int_a^b s''(t)^2\, dt \quad (5)$$

Bezeichnen die linke Seite mit $\psi(s, f, \delta)$ Eine notwendige Bedingung für (5) ist

$$\frac{\partial}{\partial \delta}\psi(s, f, \delta)|_{\delta=0} = 0 \quad \forall f \in W_2, \quad \text{also}$$

$$-\frac{1}{n\lambda}\sum_{i=1}^{n} f(t_i)(y_i - s(t_i)) + \int_a^b f''(t)s''(t)\, dt = 0 \quad (6)$$

Da $f, s \in W_2$ sind Differentiation und Integration vertauschbar.

Beh.: Ein natürlicher kubischer Spline

$$s(t) = \sum_{j=1}^{n+4} \beta_j x_j(t)$$

mit $\beta_j = \theta_j (j = 1, \ldots, 4), \beta_j = \vartheta_{j-4} (j = 5, \ldots, n+4)$ und Spline-Basisfunktionen
$\mathbf{x} = (x_1(t), \ldots, x_{n+4}(t)) = (1, t, t^2, t^3, (t - t_i)_+^3, i = 1, \ldots, n)$ erfüllt Bedingung (6) für alle Funktionen $f \in W_2$.
(Anmerkung: $\mathbf{x}$ ist Basis von $S(t_1, \ldots, t_n)$, aber *keine* Basis von $NS(t_1, \ldots, t_n)$)
Setzen wir $s(t)$ in die linke Seite von (6) ein, erhalten wir

$$-\frac{1}{n\lambda}\sum_{i=1}^{n} f(t_i)(y_i - \sum_{j=1}^{n+4}\beta_j x_j(t)) + \int_a^b f''(t)\sum_{j=1}^{n+4}\beta_j x_j''(t)\, dt = 0 \quad (7)$$

Sei $t_0 = a, t_{n+1} = b$. Untersuchen erst das Integral

$$\int_{t_i}^{t_{i+1}} \underbrace{f''(t)}_{u'}\underbrace{x_j''(t)}_{v}\, dt = \underbrace{f'(t)}_{u}\underbrace{x''(t)}_{v}|_{t_i}^{t_{i+1}} - \int_{t_i}^{t_{i+1}} \underbrace{f'(t)}_{u}\underbrace{x_j'''(t)}_{v'=\delta_{ij}}\, dt$$
$$= \quad f'(t_{i+1})x_j''(t_{i+1}) - f'(t_i)x_j''(t_i) - \delta_{ij}(f(t_{i+1}) - f(t_i))$$

wobei

$$\delta_{ij} = x_j'''(t_i) \quad = \quad \begin{cases} 6 & \text{falls} \quad j = 4, \ldots, i+4 \\ 0 & \text{falls} \quad j = 1, 2, 3, i+5, \ldots, n+4 \end{cases}$$
$$= \quad \begin{cases} 6 & \text{falls} \quad i \geq j-4, j \geq 4 \\ 0 & \text{sonst} \end{cases}$$

Damit wird der Integralausdruck von (7) zu

$$\sum_{i=0}^{n}\sum_{j=1}^{n+4}\beta_j\big(f'(t_{i+1})x_j''(t_{i+1})-f'(t_i)x_j''(t_i)-\delta_{ij}(f(t_{i+1})-f(t_i))\big)$$

$$=\sum_{j=1}^{n+4}\beta_j\Bigg(\underbrace{\sum_{i=0}^{n}\big(f'(t_{i+1})x_j''(t_{i+1})-f'(t_i)x_j''(t_i)\big)}_{f'(b)x_j''(b)-f'(a)x_j''(a)}-\sum_{i=0}^{n}\delta_{ij}(f(t_{i+1})-f(t_i))\Bigg)$$

$$=f'(b)\underbrace{\sum_{j=1}^{n+4}\beta_j x_j''(b)}_{s''(b)=0}-f'(a)\underbrace{\sum_{j=1}^{n+4}\beta_j x_j''(a)}_{s''(a)=0}-\sum_{i=0}^{n}\sum_{j=1}^{n+4}\delta_{ij}\beta_j(f(t_{i+1})-f(t_i))$$

$$=-\sum_{j=4}^{n+4}\beta_j\underbrace{\sum_{i=0}^{n}(f(t_{i+1})-f(t_i))\delta_{ij}}_{6f(t_{j-4})}=-6\sum_{j=5}^{n+4}\beta_j f(t_{j-4})=-6\sum_{i=1}^{n}\beta_{i+4}f(t_i)$$

Nebenrechnung:

$$\begin{aligned}
&\sum_{i=0}^{n}(f(t_{i+1})-f(t_i))\delta_{ij}=\\
&-\delta_{0j}f(t_0)+\delta_{0j}f(t_1)-\delta_{1j}f(t_1)+\delta_{1j}f(t_2)-+\cdots\\
&-\delta_{nj}f(t_n)+\delta_{nj}f(t_{n+1})=\\
&-\delta_{0j}f(t_0)+\sum_{i=1}^{n}(\delta_{i-1,j}-\delta_{ij})f(t_i)+\delta_{nj}f(t_{n+1})=6f(t_{j-4})
\end{aligned}$$

$\delta_{0j}=\delta_{nj}=0$ da $x'''(a)=x'''(b)=0$. $\delta_{i-1,j}-\delta_{ij}=\begin{cases}6 & \text{falls} \quad i=j-4\\ 0 & \text{sonst}\end{cases}, i=1,\dots,n$

Die Gleichung (7) wird dann für alle $f\in W_2$ zu

$$-\frac{1}{n\lambda}\sum_{i=1}^{n}f(t_i)(y_i-\sum_{j=1}^{n+4}\beta_j x_j(t_i))=-6\sum_{i=1}^{n}\beta_{i+4}f(t_i)$$

$$\sum_{i=1}^{n}f(t_i)\Bigg(\frac{1}{n\lambda}(y_i-\sum_{j=1}^{n+4}\beta_j x_j(t_i))+6\beta_{i+4}\Bigg)=0$$

Die linke Seite wird zu Null wenn alle Klammerausdrücke Null werden, d.h. wenn $\boldsymbol{\beta}$ so gewählt werden kann, dass

$$y_i = 6n\lambda\beta_{i+4}+\sum_{j=1}^{n+4}\beta_j x_j(t_i)=\sum_{j=1}^{n+4}(x_j(t_i)+6n\lambda\eta_{ij})\beta_j \tag{8}$$

$$\mathbf{Y} = (\mathbf{X}+6n\lambda\mathbf{G})\boldsymbol{\beta} \tag{9}$$

wobei $\eta_{ij}=1$, falls $j=i+4$, 0 sonst, $\mathbf{Y}=(y_1,\dots,y_n)$, $\mathbf{X}=(x_j(t_i))_{i=1,\dots,n,j=1,\dots,n+4}$, $\mathbf{G}=(\eta_{ij})_{i=1,\dots,n,j=1,\dots,n+4}$
Bleibt zu zeigen, dass $\boldsymbol{\beta}$ so gewählt werden kann, dass (9) gilt.

Zusammen mit den natürlichen Spline-Nebenbedingungen $\theta_2 = 0, \theta_3 = 0$, $\sum_{i=1}^n \beta_i = 0$ und $\sum_{i=1}^n \beta_i t_i = 0$ wird (9) ein lineares Gleichungssystem mit je $n+4$ Gleichungen und Unbekannten.

$$\mathbf{Y}^* = \mathbf{B}\boldsymbol{\beta} \tag{10}$$

Das lineare Gleichungssystem (10) besitzt genau dann eine eindeutige Lösung wenn $rg(\mathbf{B}) = rg(\mathbf{B}, \mathbf{Y}^*) = n+4$
Betrachten dazu das homogene lineare Gleichungssystem $\mathbf{Bz} = \mathbf{0}$, d.h. in (10) setzen wir $\mathbf{Y}^* = \mathbf{0}$. (6) wird dann zu

$$\frac{1}{n\lambda}\sum_{i=1}^n f(t_i)s(t_i) + \int_a^b f''(t)s''(t)\,dt = 0 \quad \forall f \in W_2$$

insbesondere auch für $f(t) = s(t)$, d.h.

$$\frac{1}{n\lambda}\sum_{i=1}^n s^2(t_i) + \int_a^b (s''(t))^2\,dt = 0,$$

woraus $f(t_i) = s(t_i) = 0 \quad \forall i = 1, \ldots, n$ und $s''(t) = 0 \quad \forall t \in [a,b]$ folgt.
Daraus folgt $s(t)$ ist Polynom 2. Grades, das an allen Stellen t_i verschwindet. Wegen $n \geq 3$ folgt $s(t) = 0 \quad \forall t \in [a,b]$.
Da die $x_j(t)$ eine Basis von $S(t_1, \ldots, t_n)$ bilden, folgt aus $s(t) = \sum_{j+1}^{n+4} \beta_j x_j(t) = 0$ dass $\beta_j = 0 \quad \forall j$
d.h. Das Gleichungssystem $\mathbf{Bz} = \mathbf{0}$ hat nur die Nulllösung $\mathbf{z} = \mathbf{0}$ und der Rang von $\mathbf{B}$ ist voll und das Gleichungssystem (10) hat genau eine Lösung.

Bleibt noch zu zeigen, $s(t) = \sum_{j=1}^{n+4} \beta_j x_j(t)$ löst die Minimumaufgabe. Es gilt

$$\begin{aligned}
& \frac{1}{n}\sum_{i=1}^n (y_i - f(t_i))^2 + \lambda \int_a^b (f''(t))^2\,dt \\
= \; & \frac{1}{n}\sum_{i=1}^n (y_i - s(t_i))^2 + \frac{1}{n}\sum_{i=1}^n (s(t_i) - f(t_i))^2 \\
& + \underbrace{\frac{2}{n}\sum_{i=1}^n (f(t_i) - s(t_i))(s(t_i) - y_i)}_{:=A} \\
& + \lambda \int_a^b (s''(t) - f''(t))^2\,dt + \lambda \int_a^b (s''(t))^2\,dt \\
& + \underbrace{2\lambda \int_a^b s''(t)(f''(t) - s''(t))\,dt}_{:=B}
\end{aligned}$$

Es gilt $A + B = 0$, da $f - s \in W_2$ und (6) für alle Funktionen aus W_2.

$$\frac{1}{n}\sum_{i=1}^n (y_i - f(t_i))^2 + \lambda \int_a^b (f''(t))^2\,dt \geq \frac{1}{n}\sum_{i=1}^n (y_i - s(t_i))^2 + \lambda \int_a^b (s''(t))^2\,dt$$

Bleibt zu zeigen, s(t) ist einziges Minimum (ÜA).

Bem.:

- Bei gegebenen Stützstellen und Glättungsparameter kann ein glättender kubischer Spline z.B. durch Lösen eines linearen Gleichungssystems bestimmt werden. Die Lösung kann explizit angegeben werden, wenn man (10) von links mit $\mathbf{B}'$ multipliziert und dann die resultierende $(n+4, n+4)$ Matrix invertiert.
- Anstelle kubischer Polynomsplines kann man z.B. auch sogenannte B-Splines als Basisfunktionen verwenden, die populär geworden sind.

4 Interpolation und Numerische Integration

4.1 Interpolation durch Polynome

Gegeben sei eine Funktion f, die durch ein Polynom vom Grad n zu interpolieren ist. Seien $t_0, \ldots, t_n$ Stützstellen, und das Polynom soll durch die Punkte $(t_i, f(t_i)) =: (t_i, f_i), i = 0, \ldots, n$ gehen.
Definieren die Funktionen

$$\begin{aligned} \omega(t) &= \prod_{j=0}^{n}(t-t_j), \qquad \omega_k(t) = \prod_{j=0, j\neq k}^{n} \frac{t-t_j}{t_k-t_j} \\ L_n(t) &= \sum_{k=0}^{n} \omega_k(t) f_k \end{aligned}$$

$L_n(t)$ ist Interpolationspolynom, da

$$L_n(t_i) = f_i, \quad \text{wegen} \quad \omega_k(t_i) = \delta_{ik} = \begin{cases} 1 & \text{falls} \quad i = k \\ 0 & \text{sonst} \end{cases}$$

$$\begin{aligned} \omega(t) &= \prod_{j=0}^{n}(t-t_j), \qquad \omega_k(t) = \prod_{j=0, j\neq k}^{n} \frac{t-t_j}{t_k-t_j} \\ \omega'(t) &= \sum_{k=0}^{n} \prod_{j=0, j\neq k}^{n} (t-t_j) \\ \omega'(t_k) &= \prod_{j=0, j\neq k}^{n} (t_k-t_j) \quad \text{alle Summanden mit } j=k \text{ haben Faktor } 0 \\ \omega_k(t) &= \frac{\omega(t)}{(t-t_k)\omega'(t_k)} \\ L_n(t) &= \sum_{k=0}^{n} \omega_k(t) f_k = \sum_{k=0}^{n} \frac{\omega(t)}{(t-t_k)\omega'(t_k)} f_k \end{aligned}$$

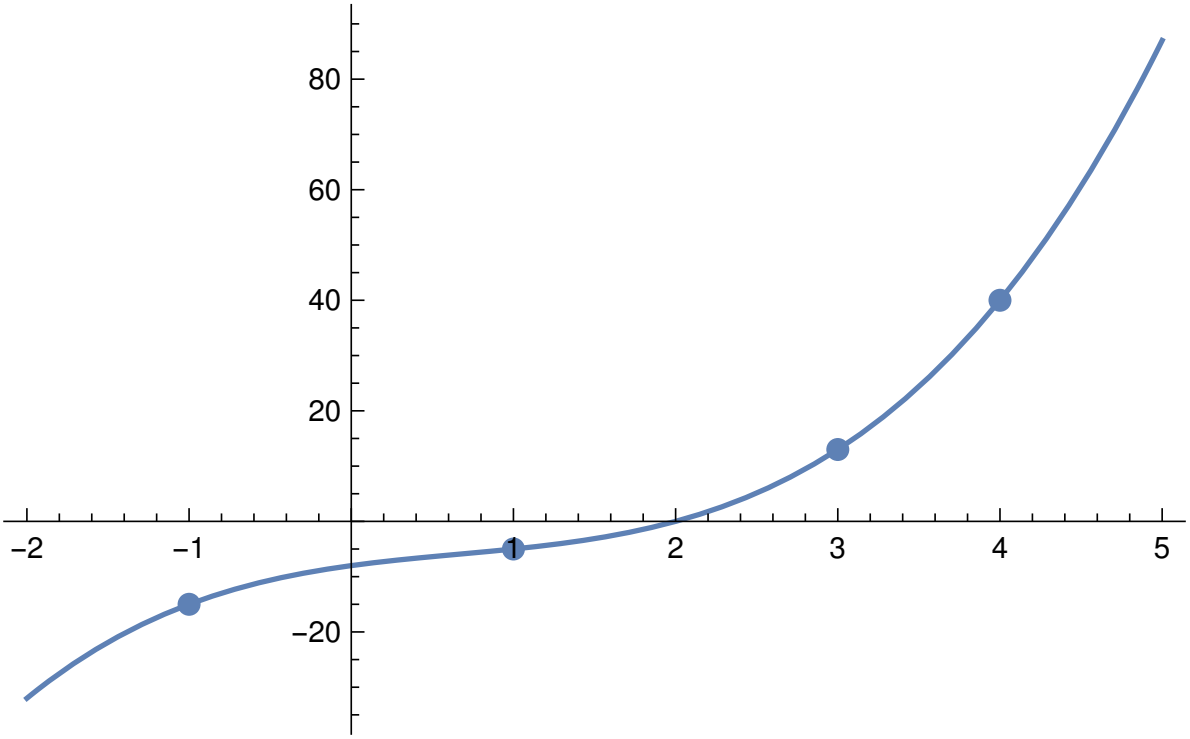
Interpolationsformel von Lagrange
Bem.: Es gibt noch andere Interpolationsformeln

Beispiel 37: Lagrange-Interpolationspolynom zu den Punkten $(-1,-15), (1,-5), (3,13), (4,40)$

Bestimmen die Funktionen

$$
\begin{aligned}
\omega_k(t) &= \prod_{j=0, j\neq k}^{n} \frac{t-t_j}{t_k-t_j} \\
\omega_0(t) &= \prod_{j=1}^{3} \frac{t-t_j}{t_0-t_j} = \frac{(t-1)}{(-1-1)}\frac{(t-3)}{(-1-3)}\frac{(t-4)}{(-1-4)} \\
&= -\frac{(t-1)(t-3)(t-4)}{40} \\
\omega_1(t) &= \prod_{j=0, j\neq 1}^{3} \frac{t-t_j}{t_1-t_j} = \frac{(t-(-1))}{(1-(-1))}\frac{(t-3)}{(1-3)}\frac{(t-4)}{(1-4)} \\
&= \frac{(t+1)(t-3)(t-4)}{12}
\end{aligned}
$$

$$
\begin{aligned}
\omega_2(t) &= \prod_{j=0, j\neq 2}^{3} \frac{t-t_j}{t_2-t_j} = \frac{(t-(-1))}{(3-(-1))}\frac{(t-1)}{(3-1)}\frac{(t-4)}{(3-4)} \\
&= -\frac{(t+1)(t-1)(t-4)}{8} \\
\omega_3(t) &= \prod_{j=0}^{2} \frac{t-t_j}{t_3-t_j} = \frac{(t-(-1))}{(4-(-1))}\frac{(t-1)}{(4-1)}\frac{(t-3)}{(4-3)} \\
&= \frac{(t+1)(t-1)(t-3)}{15} \\
L_n(t) &= \sum_{k=0}^{n} \omega_k(t) f_k = -15\omega_0(t) - 5\omega_1(t) + 13\omega_2(t) + 40\omega_3(t) \\
&= t^3 - 2t^2 + 4t - 8
\end{aligned}
$$



4.2 Numerische Integration

Problem: Berechnung von

$$\int_a^b f(t)\,dt$$

ist meistens anaytisch nicht möglich. Deshalb Approximation.

Vorgehen: Betrachten eine äquidistante Intervalleinteilung von $[a, b]$.

$$t_i = a + ih, \quad i = 0, \ldots, n, \qquad h = \frac{b-a}{n}, n > 0, n \in \mathbb{R}$$

Bestimmen das (Lagrange) Interpolationspolynom durch die Punkte $(t_i, f(t_i)) = (t_i, f_i), i = 0, \ldots, n$

$$L_n(t) = \sum_{k=0}^{n} \omega_k(t) f_k = \sum_{k=0}^{n} \frac{\omega(t)}{(t-t_k)\omega'(t_k)} f_k = \sum_{k=0}^{n} \prod_{j=0, j\neq k}^{n} \frac{t-t_j}{t_k-t_j} f_k$$

Approximieren

$$I = \int_a^b f(t)\,dt$$

durch

$$\begin{aligned}
\tilde{I}_n = \int_a^b L_n(t)\,dt &= \sum_{k=0}^{n} f_k \int_a^b \omega_k(t)\,dt \qquad t = a + hs, \quad dt = hds \\
&= h \sum_{k=0}^{n} f_k \int_0^n \omega_k(a+hs)\,ds \\
&= h \sum_{k=0}^{n} f_k \int_0^n \prod_{j=0, j\neq k}^{n} \frac{(a+hs)-(a+jh)}{(a+kh)-(a+jh)}\,ds \\
&= h \sum_{k=0}^{n} f_k \underbrace{\int_0^n \prod_{j=0, j\neq k}^{n} \frac{s-j}{k-j}\,ds}_{\alpha_k} = h \sum_{k=0}^{n} \alpha_k f_k
\end{aligned}$$

wobei α_k nur von n abhängt und nicht von f oder a, b.

Damit erhalten wir universelle Formeln (Newton-Cotes Formeln).

Für $n = 1$:

$$\begin{aligned}
\alpha_0 &= \int_0^1 \frac{s-1}{0-1}\,ds = -\frac{s^2}{2} + s\Big|_0^1 = \frac{1}{2} \\
\alpha_1 &= \int_0^1 \frac{s-0}{1-0}\,ds = \frac{s^2}{2}\Big|_0^1 = \frac{1}{2}
\end{aligned}$$

Für $n = 2$:

$$\begin{aligned}
\alpha_0 &= \int_0^2 \frac{s-1}{0-1}\frac{s-2}{0-2}\,ds = \frac{1}{2}\int_0^2 (s^2 - 3s + 2)\,ds = \frac{1}{3} \\
\alpha_1 &= \int_0^2 \frac{s-0}{1-0}\frac{s-2}{1-2}\,ds = \frac{1}{2}\int_0^2 (s^2 - 2s)\,ds = \frac{4}{3} \\
\alpha_2 &= \int_0^2 \frac{s-0}{2-0}\frac{s-1}{2-1}\,ds = \frac{1}{2}\int_0^2 (s^2 - s)\,ds = \frac{1}{3}
\end{aligned}$$

Damit erhalten wir die Näherungswerte ($h = \frac{b-a}{n}$)

$$\begin{aligned}
\tilde{I}_1 = \int_a^b L_1(t)\,dt &= \frac{b-a}{2}(f_0 + f_1) \qquad \text{Trapezregel} \\
\tilde{I}_2 = \int_a^b L_2(t)\,dt &= \frac{b-a}{6}(f_0 + 4f_1 + f_2) \qquad \text{Simpsonregel}
\end{aligned}$$

Zur Erinnerung:
$t_i = a + ih, \quad h = \frac{b-a}{n}, \quad f_i = f(t_i)$
bei der Trapezregel haben wir $n = 1, h = b - a$, also: $f_0 = f(a), f_1 = f(b)$
bei der Simpsonregel haben wir $n = 2, h = \frac{b-a}{n} = \frac{b-a}{2}$, also: $f_0 = f(a), f_1 = f\big(\frac{a+b}{2}\big)$, $f_2 = f(b)$

In der Praxis wird das Intervall $[a, b]$ in viele kleine Teilintervalle zerlegt und auf jedes dieser Intervalle eine der Regeln angewendet.

Fehlerabschätzung (Beweis siehe Stoer/Bulirsch)
Sei $m^{(i)} := \sup_{\xi\in[a,b]} \big|f^{(i)}(\xi)\big|$.
Trapezregel: $\big|I - \tilde{I}_1\big| = (b-a)^3 \frac{1}{12} m^{(2)}$
Simpsonregel: $\big|I - \tilde{I}_2\big| = (b-a)^5 \frac{1}{180} m^{(4)}$

Bem.: Wenn wir also das Intervall $[a, b]$ in N Teilintervalle zerlegen, und auf jedes dieser Intervalle eine der Regeln anwenden erhalten wir einen Fehler von $\mathcal{O}(\frac{1}{N^2})$ (Trapezregel) bzw. $\mathcal{O}(\frac{1}{N^4})$ (Simpsonregel) (unter der Voraussetzung, dass die Ableitungen existieren).

Beispiel 38: $f(x) = 1 + x, \quad x \in [0, 1], \qquad I = \int_0^1 f(x)\,dx$
Trapezregel $\quad f_0 = f(0), f_1 = f(1)$:

$$\tilde{I}_1 = \frac{1}{2}f_0 + \frac{1}{2}f_1 = \frac{1}{2}(1+0) + \frac{1}{2}(1+1) = \frac{3}{2}$$

Simpsonregel $\quad f_0 = f(0), f_1 = f(\frac{1}{2}), f_2 = f(1)$:

$$\tilde{I}_2 = \frac{1}{2}\big(\frac{1}{3}f_0 + \frac{4}{3}f_1 + \frac{1}{3}f_2\big) = \frac{1}{2}\big(\frac{1}{3}(1+0) + \frac{4}{3}(1+\frac{1}{2}) + \frac{1}{3}(1+1)\big) = \frac{3}{2}$$

Beide Regeln liefern sogar in einem Schritt das exakte Resultat

Beispiel 39: $f(x) = x^2, \quad x \in [0, 1], \qquad I = \int_0^1 f(x)\,dx$
Trapezregel $\quad f_0 = f(0), f_1 = f(1)$:

$$\tilde{I}_{1,1} = \frac{1}{2}f_0 + \frac{1}{2}f_1 = \frac{1}{2}\cdot 0 + \frac{1}{2}\cdot 1 = \frac{1}{2}$$

Simpsonregel $\quad f_0 = f(0), f_1 = f(\frac{1}{2}), f_2 = f(1)$:

$$\tilde{I}_{2,1} = \frac{1}{2}\big(\frac{1}{3}f_0 + \frac{4}{3}f_1 + \frac{1}{3}f_2\big) = \frac{1}{2}\big(\frac{1}{3}\cdot 0 + \frac{4}{3}\cdot(\frac{1}{2})^2) + \frac{1}{3}\cdot 1^2)\big) = \frac{1}{3}$$

Die Simpsonregel liefert das exakte Resultat, die Trapezregel nicht.

Fortsetzung von Beispiel 39: $f(x) = x^2, \quad x \in [0, 1], \qquad I = \int_0^1 f(x)\,dx$
Wenden die Trapezregel auf die beiden Teilintervalle $[0, \frac{1}{2}]$, $[\frac{1}{2}, 1]$ an.

$$\tilde{I}_{1,2} = \frac{1}{2}\big(\frac{1}{2}f(0) + \frac{1}{2}f(\frac{1}{2}) + \frac{1}{2}f(\frac{1}{2}) + \frac{1}{2}f(1))\big) = \frac{1}{2}\big(\frac{1}{8} + \frac{1}{8} + \frac{1}{2}\big) = \frac{3}{8}$$

Bei 4 Teilintervallen $[0, \frac{1}{4}], [\frac{1}{4}, \frac{1}{2}], [\frac{1}{2}, \frac{3}{4}], [\frac{3}{4}, 1]$:

$$\tilde{I}_{1,3} = \frac{1}{4}\big(\frac{1}{32} + (\frac{1}{32} + \frac{1}{8}) + (\frac{1}{8} + \frac{9}{32}) + (\frac{9}{32} + \frac{1}{2})\big) = \frac{11}{32}$$

Bei 8, 16 Teilintervallen: Übungsaufgabe.

Weitere Integrationsformeln

Manchmal bietet es sich an, nicht unbedingt äquidistante Stützstellen zu verwenden. Das ist insbesondere der Fall, wenn der Integrationsbereich unendlich ist. Wir wollen das Integral

$$\int_a^b \omega(t) f(t)\, dt$$

mit fester Gewichtsfunktion $\omega(t)$ approximieren ($a = -\infty, b = \infty$ ist erlaubt).
Voraussetzung an die Gewichtsfunktion:

- $\omega(x) \geq 0$ auf $[a, b]$
- Alle Momente $\int_a^b x^i \omega(x)\, dx < \infty \quad \forall i \in \mathbb{N}$

Als Stützstellen werden die Nullstellen von orthogonalen Polynomen genommen (bzgl. des Skalarprodukts in $L_{2,\omega}(a, b)$)

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

Solche Polynome sind z.B.

$[a, b]$	$\omega(x)$	
$[-1, 1]$	$\frac{1}{\sqrt{1-x^2}}$	Tschebyschev-Polynome
$[0, \infty)$	e^{-x}	Laguerre Polynome
$(-\infty, \infty)$	e^{-x^2}	Hermite-Polynome

Interessant ist, dass alle Nullstellen $x_1, \ldots, x_n$ dieser Polynome einfach und reell sind.
Sei n die Anzahl der gewählten Stützstellen und $p_k(x)$ gewählte Orthogonalpolynome vom Grad $k = 1, \ldots, n$. Dann werden die Gewichte ω_i berechnet als Lösung des linearen Gleichungssystems

$$\sum_{i=1}^n \omega_i p_k(x_i) = \begin{cases} (p_0, p_0) & \text{falls} \quad k = 0 \\ 0 & \text{sonst} \end{cases}$$

Das Integral wird dann approximiert durch

$$\int_a^b \omega(t) f(t)\, dt \approx \sum_{i=1}^n \omega_i f(x_i).$$

Bem.: Wie bei den Newton-Cotes Formeln sind die Gewichte ω_i unabhängig von f.

5 Grundbegriffe der Wahrscheinlichkeitsrechnung

Wahrscheinlichkeitsrechnung

Literatur

Greiner, M. und Tinhofer, G. (1996) Stochastik für Studienanfänger der Informatik, München
Steland, A. (2013). Basiswissen Statistik, Springer
Henze, N. (2004), Stochastik für Einsteiger, Wiesbaden
Dehling, H., Haupt, B. (2003). Einführung in die Wahrscheinlichkeitsrechnung, Springer
Büchter, A., Henn, H.-W. (2005). Elementare Stochastik, Springer

5.1 Zufällige Ereignisse

Def. 17 Ein <u>zufälliger Versuch</u> (Experiment)
ist ein Versuch mit ungewissem Ausgang.

Beispiel: Glücksspiele.

Wichtig bei solchen Experimenten ist:

- die Beschreibung des Experiments (Kartenspiele, Münzwurf),
- die Erfassung der Menge aller möglichen Ausgänge des Experiments.

Def. 18 (Ereignisse)

- Elementarereignis: möglicher Versuchsausgang, **Bez.**: ω, $\omega \in \Omega$.
- Ereignis: Menge von Elementarereignissen, $A \subseteq \Omega$
- sicheres Ereignis: Menge aller Elementarereignisse: Ω.
- unmögiches Ereignis: $\emptyset$.
- Komplementärereignis: $\overline{A} = \Omega \setminus A$

Ein Experiment kann diskret sein, d.h. endlich oder abzählbar viele Ausgänge besitzen, oder es kann überabzählbar viele Ausgänge haben.

Beispiel 40:

Würfeln (1 mal)
Elementarereignisse: $1, 2, 3, 4, 5, 6$

Werfen einer Münze, bis zum ersten Mal die Zahl fällt

$$\Omega = \{z, wz, wwz, wwz, wwwz, \ldots\}.$$

Lebensdauer einer Glühbirne

$$\Omega = [0, \infty[= \mathbb{R}^+.$$

Zufällige Funktionsverläufe

$$\Omega = L_2[a, b] \quad \text{Menge aller auf } [a, b] \text{ quadratisch integrierbaren Fkt}$$

Sei $\mathcal{E}$ eine Menge von Ereignissen, $\mathcal{E} \subseteq \mathcal{P}(\Omega)$.

Def. 19 ($\cup, \cap$, Komplement von Ereignissen)
Es seien $A_1 \in \mathcal{E}$ und $A_2 \in \mathcal{E}$ Ereignisse. Dann

- $A_3 := A_1 \cap A_2 = \{\omega \in \Omega \colon \omega \in A_1 \text{ und } \omega \in A_2\}$ das Ereignis, bei dem A_1 <u>und</u> A_2 eintreten;
- $A_4 := A_1 \cup A_2 = \{\omega \in \Omega \colon \omega \in A_1 \text{ oder } \omega \in A_2\}$ das Ereignis, bei dem A_1 <u>oder</u> A_2 eintreten;
- $\overline{A_1} = \Omega \setminus A_1 = \{\omega \in \Omega \colon \omega \notin A_1\}$ das zu A_1 komplementäre Ereignis.

Es gilt offenbar:

- $A \cup \overline{A} = \Omega$ (sicheres Ereignis),
- $A \cap \overline{A} = \emptyset$ (unmögliches Ereignis).

Satz 17 (Rechenregeln für Ereignisse)

(i) $A \cup B = B \cup A$ (Kommutativgesetz)

(ii) $(A \cup B) \cup C = A \cup (B \cup C)$ (Assoziativgesetz)

(iii) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

(iv) $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ (Distributivgesetze)

(v) (De'Morgansche Regeln)

$$\begin{aligned} \overline{(A \cup B)} &= \overline{A} \cap \overline{B} \\ \overline{(A \cap B)} &= \overline{A} \cup \overline{B} \end{aligned}$$

Def. 20 (Operationen mit abzählbar vielen Ereignissen)
Seien $A_1, A_2, \ldots,$ Ereignisse. Die <u>Vereinigung</u> $\bigcup_{i=1}^{\infty} A_i$ ist das Ereignis, das eintritt, wenn mindestens eines Ereignisse $A_1, A_2, \ldots$ eintritt. Der <u>Durchschnitt</u> $\bigcap_{i=1}^{\infty} A_i$ ist das Ereignis, das eintritt, wenn alle Ereignisse $A_1, A_2, A_3, \ldots$ eintreten.

Satz 18 (Verallgemeinerungen der Rechenregeln)

(iii) $A \cap (\bigcup_{i=1}^{\infty} A_i) = \bigcup_{i=1}^{\infty} (A \cap A_i)$

(iv) $A \cup (\bigcap_{i=1}^{\infty} A_i) = \bigcap_{i=1}^{\infty} (A \cup A_i)$

(v) $\overline{\bigcup_{i=1}^{\infty} A_i} = \bigcap_{i=1}^{\infty} \overline{A_i} \qquad \overline{\bigcap_{i=1}^{\infty} A_i} = \bigcup_{i=1}^{\infty} \overline{A_i}$

Def. 21 (Ereignisfeld)
$\mathcal{E} \subseteq \mathcal{P}(\Omega)$ heißt <u>Ereignisfeld</u> über Ω (σ–Algebra über Ω) falls folgendes gilt:

1. $\Omega \in \mathcal{E}$;
2. Gilt $A_i \in \mathcal{E}$ für $i \in \mathbf{N}$, dann folgt $\bigcap\limits_{i=1}^{\infty} A_i \in \mathcal{E}$;
3. $A \in \mathcal{E} \Longrightarrow \overline{A} \in \mathcal{E}$.

Folgerung, Beweis ÜA

1. Ist $A_i \in \mathcal{E} \quad \forall i \in \mathbf{N}$, so folgt: $\bigcup\limits_{i=1}^{\infty} A_i \in \mathcal{E}$.
2. Für das unmögliche Ereignis gilt: $\emptyset \in \mathcal{E}$.

5.2 Kolmogorov'sches Axiomensystem

Def. 22 (Wahrscheinlichkeit)
Sei $\mathcal{E}$ ein Ereignisfeld. Eine Abbildung $P\colon \mathcal{E} \longrightarrow \mathbb{R}$ heißt <u>Wahrscheinlichkeit</u>, falls sie die folgenden Eigenschaften hat:

1. Für alle $A \in \mathcal{E}$ gilt: $0 \leq P(A) \leq 1$;
2. $P(\Omega) = 1$;
3. Sind die Ereignisse $A_1, A_2, \ldots$ paarweise unvereinbar (d.h. $A_i \cap A_j = \emptyset$ für $i \neq j, i, j \in \mathbb{N}$), so gilt die sogenannte σ–Additivitätseigenschaft:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

Def. 23 (Wahrscheinlichkeitsraum)
Sei Ω die Menge der Elementarereignisse, $\mathcal{E}$ ein Ereignisfeld über Ω ($\mathcal{E} \subseteq \mathcal{P}(\Omega)$) und P genüge den Kolmogorov–Axiomen, dann heißt das Tripel $(\Omega, \mathcal{E}, P)$ <u>Wahrscheinlichkeitsraum</u>.

Def. 24 (klassische Definition der Wahrscheinlichkeit (Laplace))
$\Omega = \{\omega_1, \ldots, \omega_N\}, \quad \mathcal{E} = \mathcal{P}(\Omega). \;\; P(\omega) = P(\{\omega_i\}) = \frac{1}{N} \quad \forall i = 1, \ldots, N$. D.h. alle Elementarereignisse sind gleichwahrscheinlich.
Sei $A \in \mathcal{E}$.

$$P(A) = \frac{\#\{\omega, \omega \in A\}}{N} = \frac{\#\text{für } A \text{ günstigen Elementarereignisse}}{\#\text{möglichen Elementarereignisse}}$$

Def. 25 (Borel-Mengen, Wahrscheinlichkeitsraum $(\mathbb{R}, \mathcal{B}^1, P)$)
Es sei $\Omega = \mathbb{R}$ und $\mathcal{B}^1 := \mathcal{E}_{\mathcal{A}}$ die von der Menge der halboffenen Intervalle $\mathcal{A} = \{[a, b[\colon -\infty < a < b < \infty, a, b \in \mathbb{R}\} \subseteq \mathcal{P}(\Omega)$ erzeugte σ-Algebra. $A \in \mathcal{B}^1$ heißt Borel–Menge. $(\mathbb{R}, \mathcal{B}^1, P)$ ist dann ein Wahrscheinlichkeitsraum mit irgendeiner Wahrscheinlichkeit P.

Beispiel 41: $(\Omega, \mathcal{E}, Q)$ *mit* $\Omega = [0, \pi]$, $\mathcal{E} = \{A\colon A = B \cap [0, \pi], B \in \mathcal{B}^1\}$ *und* $Q\colon A \longrightarrow \mathbb{R}$ *mit* $Q(A) := \int\limits_A \frac{1}{2}\sin(x)dx$

$$Q(\Omega) = \int_0^{\pi} \frac{1}{2}\sin x\, dx = -\frac{1}{2}\cos x\Big|_0^{\pi} = -\frac{1}{2}(-1-1) = 1$$

$(\Omega, \mathcal{E}, Q)$ ist Wahrscheinlichkeitsraum.

Satz 19 (Folgerungen aus dem Kolmogorov-Axiomensystem)
Seien $(\Omega, \mathcal{E}, P)$ ein Wahrscheinlichkeitsraum und A, B Ereignisse.

1. $P(\overline{A}) = 1 - P(A)$.
2. $P(\emptyset) = 0$.
3. Sei $A \subseteq B$. Dann gilt:
   (a) $B \setminus A \in \mathcal{E}$;
   (b) $P(B \setminus A) = P(B) - P(A)$ (Subtraktivität);
   (c) $P(A) \leq P(B)$ (Monotonie der Wkt).
4. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, $\;P(A \cup B) \leq P(A) + P(B)$. Sind A und B unvereinbar, so gilt die Gleichheit.

Beweis: ÜA □

Satz 20 (Subadditivität von P)
Seien $A_1, A_2, \ldots$ Ereignisse. Dann gilt:

$$P(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} P(A_i)$$

Beweis:

$$\begin{aligned} B_1 &:= A_1 \\ B_2 &:= A_2 \setminus A_1 \\ B_3 &:= A_3 \setminus (A_1 \cup A_2) \\ &\ldots \\ B_i &:= A_i \setminus (\bigcup_{j<i} A_j) \ldots \end{aligned}$$

B_i paarw. disjunkt, $B_i \subseteq A_i$.

$$\begin{aligned} \bigcup_{i\geq 1} B_i &= \bigcup_{i\geq 1} A_i \Rightarrow \\ P(\bigcup_{i=1}^{\infty} A_i) &= P(\bigcup_{i=1}^{\infty} B_i) \\ &= \sum_{i=1}^{\infty} P(B_i) \quad \text{(3.Axiom)} \\ &\leq \sum_{i=1}^{\infty} P(A_i) \quad \text{(Monotonie)} \end{aligned}$$

□

5.3 Kombinatorik

Kombinatorik: Aufgabenstellung
Anzahl der verschiedenen Zusammenstellungen von Objekten. Je nach Art der zusätzlichen Forderungen, ist zu unterscheiden, welche Zusammenstellungen als gleich, und welche als verschieden angesehen werden.

- Permutation (ohne Wiederholung)
- Permutation mit Wiederholung
- Variation ohne Wiederholung
- Variation mit Wiederholung
- Kombination (ohne Wiederholung)
- Kombination mit Wiederholung

Permutation (ohne Wiederholung)
Jede eineindeutige Abbildung π der geordneten Menge $\{1, \ldots, n\}$ auf eine n-elementige Menge $M = \{s_1, \ldots, s_n\}$ heißt Permutation oder Permutation ohne Wiederholung,

$$\forall i \in \{1, \ldots, n\} : \pi(i) = s_i, s_i \in M, s_i \neq s_j (i \neq j)$$

Anzahl: $N = n!$

Beispiel 42: Wiewiel Möglichkeiten gibt es, die Eisenbahnwagen 32,33,34,35,36,37 hintereinander zu hängen?

$$N = 6!$$

Permutation mit Wiederholung
Sei $M = \{s_1, \ldots, s_k\}$, $k_i > 0 \ \forall i = 1, \ldots, k$ mit $\sum_{i=1}^{k} k_i = n$. Jedes geordnete n-Tupel von Elementen aus M, wobei jedes Element s_i genau k_i mal vorkommt, heißt Permutation mit Wiederholung.

$$\text{Anzahl:} \qquad N = \frac{n!}{k_1! \cdots k_k!}$$

Beispiel 43: Wieviele Möglichkeiten gibt es, 9 verschiedene Früchte auf 3 Kinder zu verteilen, so dass jedes Kind 3 Früchte bekommt

$$N = \frac{9!}{3!3!3!}$$

Variation ohne Wiederholung

Sei $M = \{s_1, \ldots, s_n\}$. Jedes geordnete k-Tupel, $k \leq n$ von verschiedenen Elementen aus M heißt Variation ohne Wiederholung. Die Anzahl der Variationen ohne Wiederholung ist

$$N = n(n-1)\cdots(n-k+1)$$

Aufteilung von k Elementen auf n Fächer.

Beispiel 44: Wieviele Möglichkeiten für die drei Erstplazierten in einem Schachturnier mit 10 Teilnehmern gibt es?

$$N = 10 \cdot 9 \cdot 8 = 720.$$

Variation mit Wiederholung

Auswahl von k Elementen aus einer Menge $M = \{s_1, \ldots, s_n\}$ mit Zurücklegen. Die Frage ist: Wieviel verschiedene Möglichkeiten gibt es, k Elemente aus dieser Menge zu entnehmen, wobei Elemente mehrfach entnommen werden können?

$$N = n^k.$$

Beispiel 45: Anzahl der 10stelligen Dualzahlen:

$$N = 2^{10}.$$

Kombinationen (ohne Wiederholung)

Jede k-elementige Teilmenge aus einer n-elementigen Menge M heißt Kombination (ohne Wiederholung) (von k aus n Elementen). Dabei sind Wiederholungen nicht erlaubt und die Reihenfolge der k Elemente wird nicht berücksichtigt.

$$N = \frac{n \cdot (n-1) \cdot \ldots \cdot (n-k+1)}{k!} = \binom{n}{k} = \frac{n!}{(n-k)!k!}.$$

Beispiel 46: Wieviele Möglichkeiten gibt es, bei 6 Mannschaften eine Spielpaarung auszuwählen?

$$N = \binom{6}{2} = \frac{6!}{2!(6-2)!} = 15$$

Kombination (mit Wiederholung)

Fasst man alle n^k Variationen mit Wiederholung (n Elemente, Ordnung k) zu Äquivalenzklassen zusammen, so dass sie aus aus den gleichen Elementen der gleichen Anzahl bestehen, so heißt jede solche Klasse Kombination mit Wiederholung.

$$N = \binom{n+k-1}{k}$$

Beispiel 47: Wie oft kann die 1 in Binärzahlen der Länge 4 vorkommen?

$n = 2, k = 4$: $\binom{n+k-1}{k} = \binom{2+4-1}{4} = 5$ Klassen: $\{1111\}, \{1110, 1101, 1011, 0111\}, \{1100, 1010, 1001, 0110, 0101, 0011\}$, $\{1000, 0100, 0010, 0001\}, \{0000\}$

Beispiel 48: Wieviele Möglichkeiten gibt es bei Würfen mit k nicht zu unterscheidenden Würfeln mit je n Flächen?

- Dazu nehmen wir für jede Fläche i eine Kugel mit der Nummer i, $i = 1, \ldots, n$

- Hinzu packen wir $k-1$ Kugeln mit den Nummern $n+1, \ldots, n+k-1$.
- Aus den jetzt insgesamt $n+k-1$ Kugeln ziehen wir k mal ohne Zurücklegen
- Ist keine der neuen Kugeln gezogen worden so bleibt die Ziehung
- Ist genau Kugel $n+i$ gezogen worden und alle anderen haben Nummern $\leq n$ so zählt die i größte Kugel doppelt.

Fortsetzung Beispiel 48

- Sind Kugeln $n+i$ und $n+j$ gezogen worden und alle anderen haben Nummern $\leq n$ so zählen die i größte und die j größte Kugel (unter den $k-2$ anderen) jeweils doppelt. $(i, j \leq k-2)$.

  Wenn $j = k-1$ so zählt die i-te Kugel dreifach. usw.

Damit haben wir das Problem auf eine Ziehung ohne Zurücklegen zurückgeführt, wobei k mal aus einer Urne mit $n+k-1$ Elementen gezogen wird.

6 Bedingte Wahrscheinlichkeit, Unabhängigkeit

6.1 Einführung

Beispiel 49: 2-maliges Würfeln
Menge der Elementarereignisse: $\Omega = \{11, 12, 13, \ldots, 56, 66\}$ $|\Omega| = 6^2 = 36 = N$ Wir definieren zwei Ereignisse:

A: Augensumme gerade, d.h. $A = \{11, 13, 15, 22, 24, 26, \ldots, 46, 66\}$

$$P(A) = \frac{n(A)}{N} = \frac{18}{36} = \frac{1}{2}.$$

B: es fällt mindestens eine 6: $B = \{16, 26, 36, 46, 56, 66, 65, 64, 63, 62, 61\}$

$$P(B) = \frac{n(B)}{N} = \frac{6^2 - 5^2}{36} = \frac{11}{36}.$$

Fortsetzung Beispiel 49: 2-maliges Würfeln
Angenommen, Ereignis B sei bereits eingetreten, d.h. $\Omega = B$, $|\Omega| = 11$ Wahrscheinlichkeit, daß unter dieser Bedingung das Ereignis A eintritt? Mit $A \cap B = \{26, 46, 66, 64, 62\}$ erhalten wir:

$$P(A, \text{falls } B \text{ bereits eingetreten ist}) = P(A/B) = \frac{5}{11}.$$

Def. 26 (Bedingte Wahrscheinlichkeit)
Es seien $A, B \in \mathcal{E}$ zwei zufällige Ereignisse und es gelte $P(B) > 0$. Dann heißt

$$P(A/B) = \frac{P(A \cap B)}{P(B)}.$$

<u>bedingte Wahrscheinlichkeit</u> von A unter der Bedingung B.

Def. 27 (Unabhängigkeit)
Zwei Ereignisse $A, B \in \mathcal{E}$ heißen <u>unabhängig</u>, wenn gilt:

$$P(A/B) = P(A).$$

Bem.: Für zwei unabhängige Ereignisse gilt also:

$$P(A \cap B) = P(A) \cdot P(B).$$

Fortsetzung Beispiel 49: 2-maliges Würfeln. Betrachten jetzt das Ereignis

C: es fällt mindestens eine 1: $C = \{11, 12, 13, 14, 15, 16, 61, 51, 41, 31, 21\}$

$$P(C) = \frac{n(C)}{N} = \frac{6^2 - 5^2}{36} = \frac{11}{36} \qquad \text{analog zu Ereignis} \quad B$$

Sind die beiden Ereignisse B und C voneinander unabhängig?

Fortsetzung Beispiel 49: 2-maliges Würfeln
Offenbar $P(B) = P(C) > 0$. Es sei mindestens eine 6 gewürfelt worden (Ereignis B also eingetreten). Wahrscheinlichkeit, dass dann auch eine 1 gewürfelt wurde

$$P(C|B) = \frac{P(C \cap B)}{P(B)} = \frac{\frac{2}{36}}{\frac{11}{36}} = \frac{2}{11} \neq \frac{11}{36} = P(C).$$

Folglich sind die Ereignisse B und C nicht unabhängig.

Satz 21
Es seien $A, B \in \mathcal{E}$ zwei Ereignisse, wobei $P(B) > 0$ gelte. Dann genügt die bedingte Wahrscheinlichkeit $P(.|B)$ den Kolmogorov–Axiomen. D.h. das Tripel $(\Omega, \mathcal{E}, P(.|B))$ ist ein Wahrscheinlichkeitsraum.
Beweis: ÜA □

Lemma
Es seien $A, B \in \mathcal{E}$ zwei unabhängige Ereignisse. Dann sind die Ereignisse A und $\overline{B}$ ebenfalls unabhängig. Gleiches gilt für die Ereignisse $\overline{A}$ und B sowie für $\overline{A}$ und $\overline{B}$.

Beweis: Wir zeigen die Aussage am Beispiel der Ereignisse A und $\overline{B}$. Sei $0 < P(B) < 1$.

$$\begin{aligned}
P(A|\overline{B}) &= \frac{P(A \cap \overline{B})}{P(\overline{B})} \\
&= \frac{P(A \setminus (A \cap B))}{1 - P(B)} \quad \text{(Satz 19.1))} \\
&= \frac{P(A) - P(A \cap B)}{1 - P(B)} \quad \text{(Satz 19.3b))} \\
&= \frac{P(A) - P(A)P(B)}{1 - P(B)} = \frac{P(A)(1 - P(B))}{1 - P(B)} = P(A)
\end{aligned}$$

□

6.2 Satz der Totalen Wahrscheinlichkeit

Def. 28 (Vollständigkeit einer Ereignisfolge)
Es sei $(\Omega, \mathcal{E}, P)$ ein Wahrscheinlichkeitsraum. Eine Folge von Ereignissen

$$\{A_n\}_{n=1}^{\infty} \; (A_n \in \mathcal{E}, \forall n \in \mathbb{N})$$

heißt <u>vollständig</u> falls folgende Bedingungen erfüllt sind:

1. $\bigcup\limits_{n=1}^{\infty} A_n = \Omega$;

2. $A_i \cap A_j = \emptyset$, für alle $i \neq j$.

Satz 22 (Satz der Totalen Wahrscheinlichkeit)
Es sei $A_1, A_2, \ldots$ eine vollständige Folge von Ereignissen. Weiterhin sei B ein beliebiges Ereignis und es gelte $P(A_i) \neq 0$ für alle i . Dann gilt:

$$P(B) = \sum_{i=1}^{\infty} P(B|A_i)P(A_i).$$

Beweis: Aus $B = B \cap (\bigcup_{i=1}^{\infty} A_i) = \bigcup_{i=1}^{\infty}(B \cap A_i)$ folgt (da die $(B \cap A_i)$ ebenfalls unvereinbar sind):

$$P(B) = P\left(\bigcup_{i=1}^{\infty}(B \cap A_i)\right) = \sum_{i=1}^{\infty} P(B \cap A_i) = \sum_{i=1}^{\infty} P(B|A_i)P(A_i)$$

□

Beispiel 50: Wir betrachten ein System (eine Markovsche Kette) mit 3 Zuständen und den Übergangswahrscheinlichkeiten

$$\begin{pmatrix} \frac{1}{6} & \frac{1}{3} & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & \frac{1}{6} \\ \frac{1}{2} & \frac{1}{6} & \frac{1}{3} \end{pmatrix}$$

Wie groß ist die Wahrscheinlichkeit, dass sich das System nach einem Schritt im Zustand 3 befindet?

Ereignisse:

- S_i: in Zustand i wird gestartet, $P(S_i) = p_i^0$, $p^0 := (\frac{1}{6}, \frac{1}{3}, \frac{1}{2})$
- E_j: in Zustand j kommen wir nach einem Schritt an

$$\begin{aligned} P(E_3) &= P(E_3|S_1)P(S_1) + P(E_3|S_2)P(S_2) + P(E_3|S_3)P(S_3) \\ &= \frac{1}{2} \cdot \frac{1}{6} + \frac{1}{6} \cdot \frac{1}{3} + \frac{1}{3} \cdot \frac{1}{2} = \frac{3+2+6}{36} = \frac{11}{36} \end{aligned}$$

6.3 Satz von Bayes

Gegeben: $P(A_i)$ und $P(A/A_i)$, $(i \in \mathbb{N})$. Gesucht: $P(A_i/A)$.

Satz 23 (Bayes)

$$P(A_i/A) = \frac{P(A_i) \cdot P(A/A_i)}{\sum_{j=1}^{\infty}(P(A/A_j) \cdot P(A_j))}$$

Beweis:

$$\begin{aligned} P(A_i/A) &= \frac{P(A_i \cap A)}{P(A)} && \text{Definition bedingte Wahrscheinlichkeit} \\ &= \frac{P(A_i) \cdot P(A/A_i)}{P(A)} && \text{analog, Bedingung vertauscht} \\ &= \frac{P(A_i) \cdot P(A/A_i)}{\sum_{j=1}^{\infty}(P(A/A_j) \cdot P(A_j))} && \text{Satz der Totalen Wkt.} \end{aligned}$$

Fortsetzung von Beispiel 50:

Angenommen, wir sind in Zustand 3 angekommen. Wie groß ist die Wahrscheinlichkeit, dass wir in Zustand 1 gestartet sind?

$$\begin{aligned} P(S_1|E_3) &= \frac{P(E_3|S_1)P(S_1)}{P(E_3|S_1)P(S_1) + P(E_3|S_2)P(S_2) + P(E_3|S_3)P(S_3)} \\ &= \frac{\frac{1}{2}\cdot\frac{1}{6}}{\frac{1}{2}\cdot\frac{1}{6} + \frac{1}{6}\cdot\frac{1}{3} + \frac{1}{3}\cdot\frac{1}{2}} = \frac{\frac{1}{12}}{\frac{11}{36}} = \frac{36}{11\cdot 12} = \frac{3}{11} \end{aligned}$$

7 Zufallsvariablen

7.1 Grundbegriffe

Def. 29 (Messbarkeit von Abbildungen)
Es seien $(\Omega_1, \mathcal{E}_1, P_1)$ und $(\Omega_2, \mathcal{E}_2, P_2)$ Wahrscheinlichkeitsräume. Eine Abbildung

$$X\colon \Omega_1 \longrightarrow \Omega_2$$

heißt <u>$\mathcal{E}_1$–$\mathcal{E}_2$–messbar</u>, falls für alle Ereignisse $A \in \mathcal{E}_2$ gilt:

$$X^{-1}(A) = \{\omega \in \Omega_1 \colon X(\omega) \in A\} \in \mathcal{E}_1.$$

Bem.: Wir werden nur den Fall $\mathcal{E}_2 = \mathcal{B}^1$ und $\Omega_2 \subseteq \mathbb{R}$ betrachten.

Def. 30 (Zufällige Variable, Zufallsgröße)
Es sei $(\Omega, \mathcal{E}, P)$ ein Wahrscheinlichkeitsraum. Eine $\mathcal{E}$–$\mathcal{B}^1$–meßbare Abbildung X von Ω in $\mathbb{R}$ heißt <u>(reellwertige) zufällige Variable</u> oder Zufallsgröße.

Bem.: $(\mathbb{R}, \mathcal{B}^1, P')$ bildet hier den zweiten Wahrscheinlichkeitsraum, wobei P' eine Abbildung von $\mathcal{B}^1$ in $\mathbb{R}$ ist, die den Kolmogorov–Axiomen genügt.

Beispiel 51: Augensumme beim zweimaligen Würfeln
$\Omega = \{(i,j), 1 \le i,j \le 6\}$: Paare von Augenzahlen $\mathcal{E} = \mathcal{P}(\Omega)$: Ereignisfeld $P(\omega) = P(i,j) = \frac{1}{36}$: Laplace-Wkt.

$$X: \quad \Omega \to \Omega'$$

$\Omega' = \{S : 2 \le S \le 12\}$ oder $\Omega' = \mathbb{R}$, S: Augensumme $\mathcal{E}' = \mathcal{P}(\Omega')$ oder $\mathcal{E}' = \mathcal{B}$: Ereignisfeld

$$P'(\omega') = P(S = s) = \frac{\#(i,j) : i + j = s}{36} = \frac{|X^{-1}(s)|}{36}$$

Bedingung z.B.: $X^{-1}(s) \in \mathcal{E}$ oder $X^{-1}(\{s_1, s_2\}) \in \mathcal{E}$

Def. 31 (Verteilungsfunktion von X)

$$F_X(x) := P(X \le x) = P_X\left((-\infty, x]\right)$$

Bem.: Der Einfachheit halber werden wir die Funktion F_X einfach nur mit F bezeichnen.
Bem.: Manchmal wird die Verteilungsfunktion auch durch

$$F_X(x) = P(X < x)$$

definiert.

7.2 Diskrete Zufallsvariablen

Wir beschreiben diskrete Zufallsvariablen X durch

$$X : \begin{pmatrix} x_1 & x_2 & x_3 & \cdots x_n & \cdots \\ p_1 & p_2 & p_3 & \cdots p_n & \cdots \end{pmatrix}$$

$$p_i = P(X = x_i) > 0, \quad i = 1, 2, 3, \ldots \quad \sum_{i=1}^{\infty} p_i = 1$$

Def. 32 (Wahrscheinlichkeitsfunktion, Zähldichte)
Die Funktion

$$f(x_i) = p_i$$

heißt <u>Wahrscheinlichkeitsfunktion</u>.

$x_i \in \mathbb{R}$: Werte, die die Zufallsgröße annehmen kann p_i: die entsrechenden Wahrscheinlichkeiten.

a) Zweimaliges Werfen einer Münze
$\Omega = \{ZZ, ZB, BZ, BB\}$ $X :=$ Anzahl von Blatt

$$X : \begin{pmatrix} 0 & 1 & 2 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{pmatrix}$$

b) Erfolge bei n Versuchen
Y: Anzahl der "Erfolge" bei n Versuchen, wobei jeder der n Versuche eine Erfolgswahrscheinlichkeit p hat.

$$P(Y = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{Binomialwahrscheinlichkeit} \tag{11}$$

Def. 33 (Binomialverteilung)
Eine Zufallsvariable Y mit Wahrscheinlichkeitsfunktion (11) heißt Binomial verteilt mit den Parametern n und p, bez. $Y \sim Bi(n, p)$ Im Fall $n = 1$, $Y \sim Bi(1, p)$ heißt Y auch Bernouilli verteilt.

Beispiel 52: Würfeln 3 mal. Wie groß ist die Wahrscheinlichkeit für mindestens 1 Sechs?
X: Anzahl der Sechsen.

$$\begin{aligned} P(X \geq 1) &= 1 - P(X \leq 0) = 1 - F_X(0) = 1 - \sum_{i=0}^{0} P(X = i) \\ &= 1 - \left(\frac{5}{6}\right)^3 \approx 1 - 0.579 = 0.421 \end{aligned}$$

Beispiel 53: Würfeln 6 mal. Wie groß ist die Wahrscheinlichkeit für mindestens 2 Sechsen?
Y: Anzahl der Sechsen beim sechsmaligen Wurf.

$$\begin{aligned} P(X \geq 2) &= 1 - P(X \leq 1) = 1 - F_X(1) = 1 - \sum_{i=0}^{1} P(X = i) \\ &= 1 - \left(\frac{5}{6}\right)^6 - \binom{6}{1}\left(\frac{1}{6}\right)^1\left(\frac{5}{6}\right)^5 \approx 0.26 \end{aligned}$$

Def. 34 (Poisson Verteilung)
Eine Zufallsvariable X,

$$X : \begin{pmatrix} 0 & 1 & 2 & 3 & \cdots \\ p_0 & p_1 & p_2 & p_3 & \cdots \end{pmatrix}$$

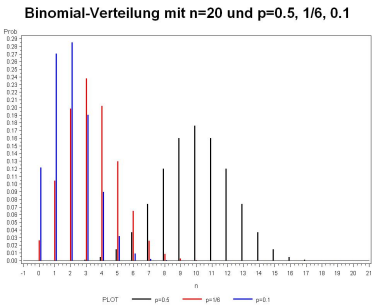
$$\text{mit} \qquad P(X=i) = p_i = \frac{\lambda^i}{i!} e^{-\lambda}, \quad \lambda > 0$$

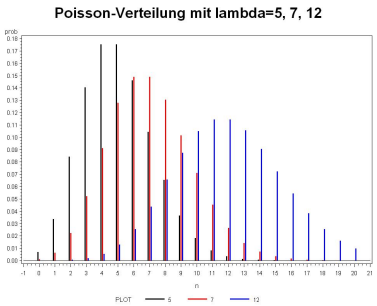
heißt Poisson-verteilt mit Parameter λ, bez. $X \sim Poi(\lambda)$.

Beispiel 54: X: Anzahl der Anrufe, die pro Zeiteinheit von einer Telefonzentrale vermittelt werden

$$\sum_{i=0}^{\infty} p_i = \underbrace{\sum_{i=0}^{\infty} \frac{\lambda^i}{i!}}_{e^\lambda} e^{-\lambda} = 1.$$

Binomial Poisson





Satz: Seien $X_n \sim Bi(n,p)$, $Y \sim Poi(\lambda)$
Für $\quad n \cdot p = \lambda$ gilt: $\quad P(X_n = k) \longrightarrow_{n\to\infty} P(Y = k)$.

Beweis: Stochastik-Vorlesung

Def. 35 (Geometrische Verteilung)
Eine Zufallsvariable X mit

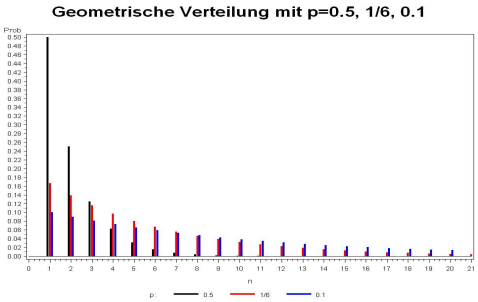
$$P(X=i) = p(1-p)^{i-1}, \quad i = 1, 2, \ldots$$

heißt <u>geometrisch</u> verteilt, bez. $X \sim Geo(p)$

X: Anzahl der Schritte bis zum ersten "Erfolg".

Beispiel 55: Würfeln solange bis eine Sechs kommt
$\Omega = \{6, i_1 6, i_2 i_3 6, ...\}, \qquad i_k \in 1, \ldots, 5 \quad X :=$ Anzahl der Würfe bis zur ersten Sechs.

$$X = \begin{pmatrix} 1 & 2 & 3 & 4 & \cdots & n & \cdots \\ \frac{1}{6} & (\frac{5}{6})\frac{1}{6} & (\frac{5}{6})^2\frac{1}{6} & (\frac{5}{6})^3\frac{1}{6} & \cdots & (\frac{5}{6})^{n-1}\frac{1}{6} & \cdots \end{pmatrix}$$



Beh.: Die in Def. 35 definierte geometrische Wahrscheinlichkeit ist eine Wahrscheinlichkeit.
Beweis: Übung.

Def. 36 (Hypergeometrische Verteilung)

Eine Zufallsvariable X mit

$$P(X = k) = \frac{\binom{n}{k} \cdot \binom{N-n}{m-k}}{\binom{N}{m}}$$

heißt hypergeometrisch verteilt, bez. $X \sim Hyper(N, n, m)$

Beispiel 56: Qualitätskontrolle
Gegeben sei eine Grundgesamtheit (z.B. eine Warenlieferung) mit N Stücken, von denen genau n schlecht seien. Wie groß ist die Wahrscheinlichkeit, dass in einer Stichprobe vom Umfang m höchstens k Stück schlecht sind?

X: zufällige Anzahl der schlechten Stücke in der Stichprobe.

7.3 Stetige Zufallsvariablen

Def. 37 (Dichtefunktion)

Eine Funktion $f\colon \mathbb{R} \longrightarrow \mathbb{R}$ heißt <u>Dichtefunktion</u>, falls sie die folgenden Eigenschaften hat:

1. Für alle $x \in \mathbb{R}$ gilt: $f(x) \geq 0$.
2. Es gilt: $\int\limits_{\mathbb{R}} f(x)\,dx = 1$.

Def. 38 (Stetige Zufallsvariable)
Eine zufällige Variable X heißt <u>stetig</u>, falls eine Dichtefunktion f_x existiert, so dass gilt:

$$P(X \leq x) = F(x) = \int\limits_{-\infty} f(t)\,dt.$$

Falls die Funktion f stetig ist, gilt: $F'(x) = f(x)$.

Stetige Zufallsvariablen

Bem.: Für die Wahrscheinlichkeit $P(X = x)$ gilt

$$P(X = x) = \int\limits_{x}^{x} f(t)\,dt = 0,$$

sogar wenn X den Wert x tatsächlich annehmen kann! D.h. z.B.

$$P(X \leq x) = P(X < x).$$

Außerdem gilt:

$$P(a \leq X \leq b) = \int\limits_{a}^{b} f(t)\,dt.$$

Def. 39 (Gleichverteilung)
Eine Zufallsvariable X auf dem Intervall (a, b) definiert mit der Dichtefunktion

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{falls } a \leq x < b \\ 0, & \text{sonst} \end{cases}.$$

heißt gleichverteilt auf dem Intervall (a, b), bez. $X \sim R(a, b)$.

Def. 40 (Exponentialverteilung) $X \sim Exp(\lambda)$
Eine Zufallsvariable X mit Verteilungsfunktion

$$F(x) = \begin{cases} 1 - e^{-\lambda \cdot x}, & \text{falls } x \geq 0 \\ 0, & \text{falls } x < 0 \end{cases} .$$

bzw. Dichtefunktion

$$f(x) = F'(x) = \begin{cases} \lambda \cdot e^{-\lambda \cdot x}, & \text{falls } x \geq 0 \\ 0, & \text{falls } x < 0 \end{cases} .$$

heißt exponentialverteilt, bez. $X \sim Exp(\lambda)$.

$\lim_{x \to -\infty} F(x) = 0, \qquad \lim_{x \to +\infty} F(x) = 1.$

Beispiel 57: Normalverteilung

$$X: \quad (\Omega, \mathcal{E}, P) \to (\mathbb{R}^1, \mathcal{B}^1, P_X)$$

sei der Messfehler bei Messung einer physikalischen Konstanten.
Der Wkt.raum $(\Omega, \mathcal{E}, P)$ ist ein Modell eines im Hintergrund wirkenden Zufallsmechanismus, der nicht näher beschrieben werden kann, Fehler im Messinstrument; zufällige äußere Einflüsse.
Er enthält alle nicht näher bestimmbaren zufälligen Effekte. Zur Beschreibung dient der Bildraum $(\mathbb{R}^1, \mathcal{B}^1, P_X)$.

Def. 41 (Normalverteilung)
Die Zufallsvariable X mit der Verteilungsfunktion

$$F(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt.$$

heißt normalverteilt mit den Parametern (μ, σ^2), bez.$X \sim \mathcal{N}(\mu, \sigma^2)$. Die zugehörige Dichtefunktion hat die Form:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad \sigma > 0.$$

offen: $f(x)$ ist Dichtefunktion. Siehe Stochastik-Vorlesung.

Bem: In den Wahrscheinlichkeiten können Parameter auftreten, die in der Regel unbekannt sind.

Die Parameter sind anhand der Beobachtungen (der Daten) zu bestimmen/zu <u>schätzen</u>! $\longrightarrow$ Aufgabe der Statistik

7.4 Der Erwartungswert

Der Erwartungswert

Beispiel 58: Eine Münze wird 3 mal geworfen.
Wie oft können wir erwarten, daß Blatt oben liegt? Wie oft wird im Mittel Blatt oben liegen?

$$X: \quad \begin{pmatrix} 0 & 1 & 2 & 3 \\ 1/8 & 3/8 & 3/8 & 1/8 \end{pmatrix}$$

Erwartungswert: $0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = \frac{12}{8} = 1.5$
D.h. bei 10maliger Durchführung des Experiments können wir im Mittel mit 15mal Blatt rechnen.

Sei X diskrete Zufallsvariable,

$$X : \begin{pmatrix} x_1 & ... & x_n & ... \\ p_1 & ... & p_n & ... \end{pmatrix}$$

Def. 42 (Erwartungswert, X diskret)
Die reelle Zahl

$$\mathbf{E}X = \sum_{i=1}^{\infty} p_i x_i$$

heißt Erwartungswert von X

Der Erwartungswert, Beispiele (1)

a) $X \sim Poisson(\lambda)$

$$X : \begin{pmatrix} 0 & 1 & 2 & 3 & ... \\ p_0 & p_1 & p_2 & p_3 & ... \end{pmatrix}$$

$$p_i = \frac{\lambda^i}{i!} e^{-\lambda}$$

$$\mathbf{E}X = \sum_{i=0}^{\infty} p_i i = \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} e^{-\lambda} \cdot i = \lambda \underbrace{\sum_{i=1}^{\infty} \frac{\lambda^{i-1}}{(i-1)!}}_{e^{\lambda}} e^{-\lambda} = \lambda.$$

z.B. mittlere Ankunftsrate.

Der Erwartungswert, Beispiele (2)

b) $X \sim Bi(n,p)$: $\mathbf{E}(X) = n \cdot p$ *(ÜA)*

$$X : \begin{pmatrix} 0 & 1 & 2 & ... & n \\ p_0 & p_1 & p_2 & ... & p_n \end{pmatrix}$$

$$p_i = \binom{n}{i} p^i (1-p)^{n-i}$$

c) $X \sim Geo(p)$ $\mathbf{E}(X) = \frac{1}{p}$ *(ÜA)*

$$X : \begin{pmatrix} 1 & 2 & 3 & ... & k & ... \\ p & pq & pq^2 & ... & pq^{k-1} & ... \end{pmatrix} \qquad q = 1 - p$$

Def. 43 (Erwartungswert, X stetig)
Sei X stetig mit Dichtefunktion $f(x)$. Die reele Zahl

$$\mathbf{E}X = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

heißt Erwartungswert von X.

Der Erwartungswert, Beispiele (4)

a) $X \sim \mathcal{N}(\mu, \sigma^2)$

$$\begin{aligned}
\mathbf{E}X &= \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi} \cdot \sigma} e^{-(\frac{x-\mu}{\sigma})^2/2} dx \\
&= \int_{-\infty}^{\infty} (\sigma t + \mu) \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}} dt \\
&= \mu + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \sigma \cdot t \cdot e^{\frac{-t^2}{2}} dt = \mu.
\end{aligned}$$

$\frac{x-\mu}{\sigma} = t, \quad dt = \frac{1}{\sigma} dx$

Der Erwartungswert, Beispiele (5)

b) $X \sim Exp(\lambda), \qquad \lambda > 0$

$$\mathbf{E}X = \int_0^{\infty} x \cdot \lambda \cdot e^{-\lambda \cdot x} \, dx = \frac{1}{\lambda} \qquad \text{(ÜA)}$$

c) $X \sim R(a, b)$*, gleichverteilt auf dem Intervall (a,b)*

$$\mathbf{E}X = \frac{1}{b-a} \int_a^b x \, dx = \frac{1}{b-a} \frac{x^2}{2} \bigg|_a^b = \frac{b^2 - a^2}{2(b-a)} = \frac{a+b}{2}.$$

Bemerkung: Die Erwartungswerte sind für stetige und diskrete Zufallsgrößen zweckmäßigerweise unterschiedlich definiert. Sie lässt sich jedoch (maßtheoretisch) vereinheitlichen.

Satz 24 (Eigenschaften des Erwartungswertes)
Seien X, X_1 und X_2 zufällige Variablen und $a, b, c \in \mathbb{R}$ beliebig. Dann gelten folgende Aussagen:

1. Wenn $P(X = c) = 1$, d.h. nimmt die zufällige Variable X genau einen festen Wert an, so folgt $\mathbf{E}X = \mathbf{E}c = c$.
2. Wenn $P(X \geq c) = 1$, so $\mathbf{E}X \geq c$.
3. $\mathbf{E}(c \cdot X) = c \cdot \mathbf{E}X$.
4. $\mathbf{E}(X + c) = \mathbf{E}X + \mathbf{E}c = \mathbf{E}X + c$.
5. $\mathbf{E}(a \cdot X_1 + b \cdot X_2) = a \cdot \mathbf{E}X_1 + b \cdot \mathbf{E}X_2$.
6. $\mathbf{E}g(X) = \begin{cases} \sum_{i=0}^{\infty} g(x_i) p_i & \text{falls } X \text{ diskret} \\ \int_{-\infty}^{\infty} g(x) f(x) \, dx & \text{falls } X \text{ stetig} \end{cases}$

Beweis: Wir beweisen stellvertretend Aussage 2.

- Es sei X eine diskrete Zufallsgröße,

$$X: \quad \begin{pmatrix} x_1 & x_2 & \dots & x_n & \dots \\ p_1 & p_2 & \dots & p_n & \dots \end{pmatrix}$$

  Nach Voraussetzung: $c \leq x_1 < x_2 < \ldots < x_n < \ldots$. Daraus folgt:

$$\mathbf{E}X = \sum_{i \in \mathbb{N}} x_i \cdot p_i \geq \sum_{i \in \mathbb{N}} c \cdot p_i = c \cdot \sum_{i \in \mathbb{N}} p_i = c.$$

□

Beweis: von Aussage 2 (Fortsetzung)

- Es sei X eine stetige zufällige Variable mit der Dichtefunktion f. Dann gilt:

$$P(X \geq c) = \int_c^{+\infty} f(x)\,dx = 1. \quad \Rightarrow$$

$$P(X < c) = \int_{-\infty}^{c} f(x)\,dx = 0. \quad \Rightarrow$$

$$\mathbf{E}X = \int_{-\infty}^{+\infty} x \cdot f(x)\,dx = \int_c^{+\infty} x \cdot f(x)\,dx \geq c \cdot \underbrace{\int_c^{+\infty} f(x)\,dx}_{=1} = c$$

□

7.5 Die Varianz

Es sei X eine zufällige Variable mit $\mathbf{E}X = \mu$.

Def. 44 (Varianz, bez. $\operatorname{Var} X$ oder σ_X^2)
Falls $\mathbf{E}(|X|^2) < \infty$, heißt der Erwartungswert $\mathbf{E}(X-\mu)^2$ <u>Varianz</u>

$$\mathbf{E}(X-\mu)^2 = \begin{cases} \int_{-\infty}^{+\infty} (x-\mu)^2 \cdot f(x)\,dx, & \text{falls } X \text{ stetig ist} \\ \sum_{i \in \mathbb{N}} (x_i - \mu)^2 \cdot p_i, & \text{falls } X \text{ diskret ist} \end{cases}$$

Def. 45 (Standardabweichung), σ, σ_X

$$\sigma = \sqrt{Var(X)}$$

Bem.: $\operatorname{Var}(X)$: mittlere quadratische Abweichung zwischen X und EX.

Satz 25 (Eigenschaften der Varianz)

1. Sei $c \in \mathbb{R}$. Wenn $P(X = c) = 1$, so $\operatorname{Var} X = 0$. Ist umgekehrt $\operatorname{Var} X = 0$, so existiert ein $c \in \mathbb{R}$, so daß gilt: $P(X = c) = 1$.
2. Für beliebige $c \in \mathbb{R}$ gilt: $\operatorname{Var}(X + c) = \operatorname{Var} X$.
3. Für beliebige $a \in \mathbb{R}$ gilt: $\operatorname{Var}(a \cdot X) = a^2 \cdot \operatorname{Var} X$.
4. Für zwei zufällige Variablen X_1 und X_2 gilt: $\operatorname{Var}(X_1 + X_2) = \operatorname{Var} X_1 + \operatorname{Var} X_2 + 2 \cdot \operatorname{cov}(X_1, X_2)$.

Wir zeigen nur 1. und 4., 2. und 3. ist ÜA.

Def. 46 (Kovarianz)

$$\operatorname{cov}(X_1, X_2) = \mathbf{E}\big((X_1 - \mathbf{E}(X_1))(X_2 - \mathbf{E}(X_2))\big)$$

heißt Kovarianz der Zufallsvariablen X_1 und X_2.

Beweis von Satz 25, (1) und (4)

Es seien X, X_1 und X_2 beliebige zufällige Variablen. $a, c \in \mathbb{R}$ seien ebenfalls beliebig gewählt. Die folgenden Aussagen folgen aus dem Satz über die Eigenschaften des Erwartungswertes.

1. Es gelte: $P(X = c) = 1$. Daraus folgt $\mathbf{E}X = c$.

$$\operatorname{Var} X = \mathbf{E}(X - \mathbf{E}X)^2 = \mathbf{E}(X - c)^2 = \mathbf{E}(c - c)^2 = 0$$

   Es sei nun $\operatorname{Var} X = 0 = \mathbf{E}(X - \mathbf{E}X)^2 = 0$. Allgemein gilt für $c \in \mathbb{R}$: $\mathbf{E}(X - c)^2 \geq 0$. Also, $P(X - EX = 0) = 1$. und $c := \mathbf{E}X$ leistet das Verlangte.

4.

$$\begin{aligned}
\operatorname{Var}(X_1 + X_2) &= \mathbf{E}(X_1 + X_2 - \mathbf{E}(X_1 + X_2))^2 \\
&= \mathbf{E}(X_1 + X_2 - \mathbf{E}X_1 - \mathbf{E}X_2)^2 \\
&= \mathbf{E}((X_1 - \mathbf{E}X_1) + (X_2 - \mathbf{E}X_2))^2 \\
&= \mathbf{E}((X_1 - \mathbf{E}X_1)^2 + (X_2 - \mathbf{E}X_2)^2 \\
&\quad +2 \cdot (X_1 - \mathbf{E}X_1) \cdot (X_2 - \mathbf{E}X_2)) \\
&= \mathbf{E}(X_1 - \mathbf{E}X_1)^2 + \mathbf{E}(X_2 - \mathbf{E}X_2)^2 \\
&\quad +2 \cdot \underbrace{\mathbf{E}((X_1 - \mathbf{E}X_1) \cdot (X_2 - \mathbf{E}X_2))}_{cov(X_1, X_2)}
\end{aligned}$$

a) Poisson-Verteilung, $X \sim Poi(\lambda)$

$$p_i = P(X = i) = \frac{\lambda^i}{i!} e^{-\lambda}, \quad i = 0, 1, 2, \ldots$$

$$\begin{aligned}
\operatorname{Var}(X) &= \mathbf{E}(X - \mathbf{E}X)^2 = \sum_{i=0}^{\infty} (i - \lambda)^2 p_i \\
&= \sum_{i=2}^{\infty} i \cdot (i-1) p_i + \sum_{i=0}^{\infty} i p_i - 2\lambda \sum_{i=0}^{\infty} i p_i + \lambda^2 \sum_{i=0}^{\infty} p_i \\
&= \lambda^2 \sum_{i=2}^{\infty} \frac{\lambda^{i-2}}{(i-2)!} e^{-\lambda} + \lambda - 2\lambda^2 + \lambda^2 = \lambda.
\end{aligned}$$

b) Binomialverteilung, $X \sim Bi(n, p)$.

$$\operatorname{Var}(X) = np(1 - p).$$

(ohne Beweis, ÜA)

c) Gleichverteilung auf (a, b), $X \sim R(a, b)$

$$f(x) = \begin{cases} \frac{1}{b-a} & x \in (a, b) \\ 0 & \text{sonst.} \end{cases} \qquad \begin{aligned} \mathbf{E}X &= \frac{a+b}{2} \\ \operatorname{Var}(X) &= \frac{(b-a)^2}{12} \end{aligned} \qquad \text{(ÜA)}$$

d) Exponentialverteilung

$$
\begin{aligned}
f(x) &= \begin{cases} \lambda e^{-\lambda \cdot x} & \text{falls } x \geq 0, \\ 0 & \text{sonst.} \end{cases} \\
\mathbf{E}X &= \frac{1}{\lambda}. \\
\mathbf{E}X^2 &= \int_0^\infty x^2 \lambda e^{-\lambda \cdot x}\, dx = \frac{2}{\lambda^2} \quad (\text{ÜA}). \\
\operatorname{Var}(X) &= \frac{1}{\lambda^2}.
\end{aligned}
$$

e) Normalverteilung

$$
\begin{aligned}
f(x) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \\
\mathbf{E}(X-\mu)^2 &= \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}\, dx \\
&= \sigma^2 \int_{-\infty}^{\infty} t^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt \\
&= \sigma^2 \int_{-\infty}^{\infty} (-t)(-t \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}})\, dt \\
&= \frac{\sigma^2}{\sqrt{2\pi}} \big(-t e^{-t^2/2} \big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} (-1) e^{-\frac{t^2}{2}}\, dt \big) \\
&= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{t^2}{2}}\, dt = \sigma^2.
\end{aligned}
$$

$$t = \tfrac{x-\mu}{\sigma}, \quad dt = \tfrac{1}{\sigma} dx$$

7.6 Korrelation und Unabhängigkeit

Def. 47 (Unabhängigkeit)
Zwei Zufallsvariablen X_1 und X_2 heißen <u>unabhängig</u>, falls für alle $x_1, x_2 \in \mathbf{R}$ gilt:

$$P(X_1 < x_1, X_2 < x_2) = P(X_1 < x_1) \cdot P(X_2 < x_2)$$

Def. 48 (Unkorreliertheit)
Zwei Zufallsvariablen X_1 und X_2 heißen unkorreliert falls $\operatorname{cov}(X_1, X_2) = 0$.

Satz 26
Zwei unabhängige Zufallsgrößen X_1 und X_2 sind unkorreliert.
Beweis: siehe Stochastik-Vorlesung □
Die Umkehrung der Aussage des Lemmas gilt im allgemeinen nicht, wie das folgende Beispiel zeigt:

Beispiel 59: $X_1 \sim \mathcal{N}(0,1), X_2 = X_1^2$
Offenbar, X_1 und X_2 sind abhängig. Wir berechnen die Kovarianz.

$$\mathbf{E}X_1 = 0, \quad \mathbf{E}X_2 = \mathbf{E}X_1^2 = 1, \quad \mathbf{E}(X_1 \cdot X_2) = \mathbf{E}X_1^3 = 0$$

$$\operatorname{cov}(X_1, X_2) = \mathbf{E}(X_1 \cdot X_2) - \mathbf{E}X_1 \cdot \mathbf{E}X_2 = 0 - 0 \cdot 1 = 0$$

Trotz der Abhängigkeit der beiden Zufallsgrößen X_1 und X_2 ist ihre Kovarianz gleich Null.

Folgerung

Falls zwei zufällige Variablen X_1 und X_2 unabhängig sind, gilt für die Varianz ihrer Summe:

$$\mathrm{Var}\,(X_1 + X_2) = \mathrm{Var}\,(X_1) + \mathrm{Var}\,(X_2).$$

Def. 49 (Korrelationskoeffizient)
Es seien X_1 und X_2 zwei zufällige Variablen, für die gilt: $0 < \sigma_{X_1}, \sigma_{X_2} < \infty$. Dann heißt der Quotient

$$\varrho(X_1, X_2) = \frac{\mathrm{cov}\,(X_1, X_2)}{\sigma_{X_1} \cdot \sigma_{X_2}}$$

<u>Korrelationskoeffizient</u> der Zufallsgrößen X_1 und X_2.

Satz 27
Es seien X_1 und X_2 zwei Zufallsgrößen mit $\sigma_{X_1}, \sigma_{X_2} > 0$. Dann gilt für den Korrelationskoeffizienten:

$$-1 \leq \varrho(X_1, X_2) \leq 1.$$

Beweis. Der Satz kann direkt aus der Cauchy-Schwarz'schen Ungleichung hergeleitet werden. Dazu betrachten Sie den Raum aller Zufallsvariablen mit endlicher Varianz und definieren das Skalarprodukt $(X, Y) := \mathbf{E}((X - \mathbf{E}(X))(Y - \mathbf{E}(Y))$ und die entsprechende Norm $||X||^2 = (X, X) = \mathrm{var}(X)$:

$$-1 \leq \frac{(X, Y)}{||X|| \cdot ||Y||} \leq 1$$

□

8 Ungleichungen und Grenzwertsätze

8.1 Markov-Ungleichung

Satz 28 (Ungleichung von Markov)
Sei X eine Zufallsgröße und $c > 0$. Dann gilt:

$$P(|X| \geq c) \leq \frac{\mathbf{E}|X|}{c}.$$

Beweis: Wir definieren eine Zufallsgröße Y wie folgt

$$Y: \quad \begin{pmatrix} 0 & c \\ P(|X| < c) & P(|X| \geq c) \end{pmatrix}$$

Offenbar gilt $0 \leq Y \leq |X|$ bzw. $P(|X| - Y \geq 0) = 1$ und

$$\begin{aligned} \mathbf{E}(|X| - Y) &\geq 0 \quad \text{bzw.} \quad \mathbf{E}|X| \geq EY \\ \mathbf{E}Y &= 0 \cdot P(|X| < c) + c \cdot P(|X| \geq c) \\ &= c \cdot P(|X| \geq c) \leq \mathbf{E}|X| \end{aligned}$$

Division durch c liefert die Behauptung. □

8.2 Tschebychev-Ungleichung

Satz (Ungleichung von Tschebychev)
Es sei $\varepsilon > 0$ und sei Y eine Zufallsgröße. Dann gilt:

$$P(|Y - \mathbf{E}Y| \geq \varepsilon) \leq \frac{\operatorname{Var} Y}{\varepsilon^2}.$$

Beweis: Wir verwenden die Markov-Ungleichung:

$$P(|X| \geq c) \leq \frac{\mathbf{E}|X|}{c}.$$

und setzen

$$X := (Y - \mathbf{E}Y)^{2i} \geq 0, \quad c := \varepsilon^{2i} \qquad (i \in \mathbb{N}).$$

Da $\varepsilon > 0$ gilt, ist die Voraussetzung der Markov- Ungleichung erfüllt. Wir erhalten:

$$P(|Y - \mathbf{E}Y| \geq \varepsilon) = P\left((Y - \mathbf{E}Y)^{2i} \geq \varepsilon^{2i}\right) \leq \frac{\mathbf{E}(Y - \mathbf{E}Y)^{2i}}{\varepsilon^{2i}}.$$

Für $i := 1$ ergibt sich:

$$P(|Y - \mathbf{E}Y| \geq \varepsilon) \leq \frac{\mathbf{E}(Y - \mathbf{E}Y)^2}{\varepsilon^2} = \frac{\operatorname{Var} Y}{\varepsilon^2}. \quad \square$$

Die Tschebyschev-Ungleichung kann nicht verschärft werden (ÜA)

8.3 Chernov-Ungleichung

Satz 29 (Chernov-Ungleichung)
Seien $X_1, \ldots, X_n$ Zufallsvariablen mit $X_i \sim Bi(p_i, 1)$, d.h. $P(X_i = 1) = \mathbf{E}(X_i) = p_i$. Dann gilt $\forall \delta \in (0, 1)$:

$$P\left(\frac{\overline{X}_n - \overline{p}}{\overline{p}} > \delta\right) < \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu}$$
$$P\left(-\frac{\overline{X}_n - \overline{p}}{\overline{p}} > \delta\right) < \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu}$$

wobei $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, $\overline{p}_n = \frac{1}{n}\sum_{i=1}^n p_i$ und $\mu = n\overline{p} = \sum_{i=1}^n p_i$.

Beweis: Aus der Markov-Ungleichung erhalten wir $\forall t > 0$

$$P(\sum_{i=1}^n X_i \geq \epsilon) = P(e^{t\sum_{i=1}^n X_i} \geq e^{t\epsilon}) \leq e^{-t\epsilon}\prod_{i=1}^n \mathbf{E}(e^{tX_i}).$$

Daraus folgt $\forall t > 0$

$$\begin{aligned} P\left(\frac{\overline{X}_n - \overline{p}}{\overline{p}} > \delta\right) &= P(\overline{X}_n > (1+\delta)\overline{p}) = P(\sum_{i=1}^n X_i > \underbrace{(1+\delta)\mu}_{=\epsilon}) \\ &\leq \frac{\prod_{i=1}^n \mathbf{E}(e^{tX_i})}{e^{t(1+\delta)\mu}} \leq \frac{\prod_{i=1}^n (p_i e^t + (1-p_i))}{e^{t(1+\delta)\mu}} \end{aligned}$$

Schreiben wir $p_i e^t + (1 - p_i) = \underbrace{p_i(e^t - 1)}_{=:x} + 1$ und bemerken, dass $1 + x < e^x \quad \forall x > 0$ so erhalten wir

$$\begin{aligned} P\left(\frac{\overline{X}_n - \overline{p}}{\overline{p}} > \delta\right) &< \frac{\prod_{i=1}^n exp(p_i(e^t - 1))}{e^{t(1+\delta)\mu}} \\ &= \frac{exp((e^t - 1)\sum_{i=1}^n p_i)}{e^{t(1+\delta)\mu}} = \frac{exp((e^t - 1)\mu)}{e^{t(1+\delta)\mu}}. \end{aligned}$$

Setzen wir nun $t = \log(1+\delta)$ ($t > 0$ für $\delta > 0$) so erhalten wir für den letzen Bruch

$$\frac{exp((e^t-1)\mu)}{e^{t(1+\delta)\mu}} = \left(\frac{e^{1+\delta-1}}{(1+\delta)^{1+\delta}}\right)^{\mu} = \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu}$$

Damit ist die erste Ungleichung bewiesen. □
Die zweite Ungleichung als ÜA.
Hinweis: Analog zur ersten Ungleichung. Beachte $e^{-x} > 1-x \quad \forall x \neq 0$ folgt $\forall t > 0$

Satz 30 (Chernov-Ungleichung, vereinfacht)
Seien $X_1, \ldots, X_n$ Zufallsvariablen mit $X_i \sim Bi(p_i, 1)$, d.h. $P(X_i = 1) = \mathbf{E}(X_i) = p_i$. Dann gilt $\forall \delta \in (0,1)$:

$$\begin{aligned} P(\frac{\overline{X}_n - \overline{p}}{\overline{p}} > \delta) &\leq e^{-\mu\frac{\delta^2}{2+\delta}} \leq e^{-\mu\frac{\delta^2}{3}} \\ P(-\frac{\overline{X}_n - \overline{p}}{\overline{p}} > \delta) &\leq e^{-\mu\frac{\delta^2}{2}} \end{aligned}$$

wobei $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, $\overline{p}_n = \frac{1}{n}\sum_{i=1}^n p_i$ und $\mu = \sum_{i=1}^n p_i$.

Beweis: Logarithmieren die rechte Seite der Chernov-Ungleichung und wenden die Ungleichungen $\log(1+x) \geq \frac{x}{1+x/2}$ (auf die erste Chernov-Ungleichung) und $\log(1-x) \geq -x - x^2/2$ (auf die zweite Chernov-Ungleichung) (Beweis unter Verwendung der Reihenentwicklung) an und erhalten ($0 < \delta < 1$)

$$\begin{aligned} \log\left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu} &= \mu(\delta - (1+\delta)\log(1+\delta)) \leq -\frac{\delta^2}{2+\delta}\mu \\ \log\left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu} &= \mu(-\delta - (1-\delta)\log(1-\delta)) \leq -\frac{\delta^2}{2}\mu \leq e^{-\mu\frac{\delta^2}{3}} \end{aligned}$$

□

8.4 Das Gesetz der Großen Zahlen

Der Erwartungswert einer zufälligen Variablen X ist in der Praxis meist nicht bekannt. Um ihn zu schätzen, sammelt man Beobachtungen $X_1, X_2, \ldots, X_n$, und bildet dann das arithmetische Mittel:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i =: \overline{X}_n$$

Beachten: die Beobachtungen $X_1, \ldots, X_n$ müssen unabhängig oder wenigstens unkorreliert sein.

Satz 31 (Schwaches Gesetz der Großen Zahlen)
Es seien $X_1, \ldots, X_n$ <u>unkorrelierte</u> zufällige Variablen mit $\mu := \mathbf{E}X_i$ und $\sigma^2 := \operatorname{Var} X_i < \infty$ (für alle $i = 1, \ldots, n$). Dann gilt für alle $\varepsilon > 0$:

$$\lim_{n\to\infty} P(|\overline{X}_n - \mu| > \varepsilon) = 0.$$

Beweis: Da die Zufallsgrößen $X_1, \ldots, X_n$ unkorreliert sind, gilt

$$\mathbf{E}\overline{X} = \mu, \qquad \operatorname{Var}\overline{X} = \frac{\sigma^2}{n}$$

Mittels der Tschebychev–Ungleichung erhalten wir:

$$P(|\overline{X}_n - \mu| > \varepsilon) = P(|\overline{X} - \mathbf{E}\overline{X}| \geq \varepsilon) \leq \frac{\operatorname{Var}\overline{X}}{\varepsilon^2} = \frac{\sigma^2}{n\cdot\varepsilon^2} \xrightarrow[n\to\infty]{} 0.$$

□

8.5 Der zentrale Grenzwertsatz

Satz 32 (Der Zentrale Grenzwertsatz)
Es seien $X_1, \ldots, X_n$ unabhängige, identisch verteilte Zufallsvariablen mit $\mu := \mathbf{E}X_i; \sigma^2 := \operatorname{Var} X_i$. Seien Zufallsgrößen Z_n, $\overline{Z}_n$ und Y_n definiert durch: $Z_n := \sum\limits_{i=1}^{n} X_i$ bzw. $\overline{Z}_n := \frac{Z_n}{n}$ und

$$Y_n = \sqrt{n} \cdot \frac{\overline{Z}_n - \mu}{\sigma} = \frac{Z_n - n\mu}{\sqrt{n}\sigma}$$

Dann gilt für alle reellen x:

$$\lim_{n\to\infty} P\left(\frac{Z_n - n\cdot\mu}{\sqrt{n}\cdot\sigma} \leq x\right) = \lim_{n\to\infty} P\left(Y_n \leq x\right) = \Phi(x)$$

Beweis: Als Hilfsmittel werden charakteristische Funktionen verwendet, siehe Stochastik-Vorlesung. □

Anwendungen:

- Simulation bei der Erzeugung einer normalverteilten Zufallsgröße aus Pseudozufallszahlen
- Approximation von Wahrscheinlichkeitsverteilungen (insbesondere von Schätz- und Teststatistiken)

Beispiel 60: Seien $X_1, \ldots, X_{12} \sim R(0,1)$

$$Z = \sum_{i=1}^{12} X_i, \qquad \mathbf{E}(Z) = 6, \quad \operatorname{var}(Z) = 1$$

$$\begin{aligned} P(5 < Z < 7) &= P(\frac{5-6}{1} < \underbrace{\frac{Z-6}{1}}_{\approx\sim\mathcal{N}(0,1)} < \frac{7-6}{1}) \\ &\approx \Phi(1) - \Phi(-1) \approx 0.84 - 0.16, \end{aligned}$$

wobei $\Phi(x)$ die Verteilungsfunktion der Standardnormalverteilung ist.

Der Grenzwertsatz von Moivre-Laplace

Satz 33 (Moivre–Laplace)
Es seien $X_i \sim Bi(1,p)$, unabhängig. Dann gilt für $Z_n = \sum_{i=1}^{n} X_i$ ($\sim Bi(n,p)$):

$$Z_n \to^D Z \sim \mathcal{N}\big(np, np(1-p)\big)$$

Beweis: Mit $\mathbf{E}Z_n = np$ und $\operatorname{Var} Z_n = np(1-p)$ folgt unter Anwendung des Zentralen Grenzwertsatzes

$$\begin{aligned} F_{Z_n}(y) = P(Z_n \leq y) &= P\left(\frac{Z_n - np}{\sqrt{np(1-p)}} \leq \frac{y - np}{\sqrt{np(1-p)}}\right) \\ &\to \Phi\left(\frac{y - np}{\sqrt{np(1-p)}}\right) \end{aligned}$$

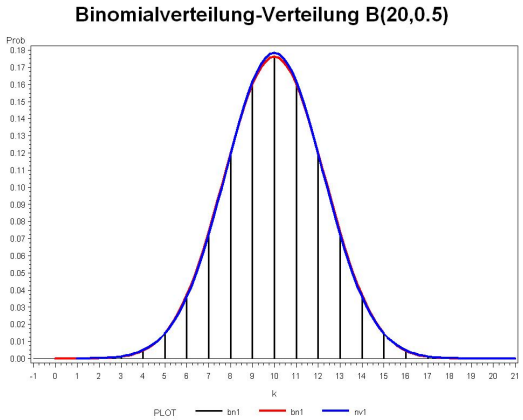
□

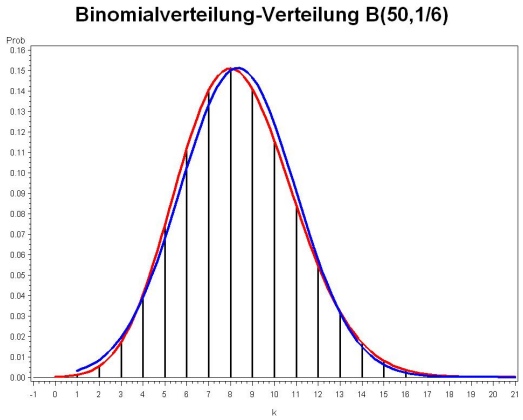
Beispiel 61: Es seien $X_i \sim Bi(1,p)$, $n = 100$ und $p = 0.1$. Gesucht werde die Wahrscheinlichkeit $P(Z_n < 15)$. Es gilt:

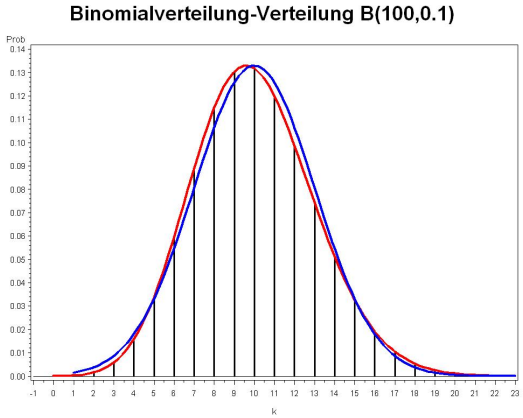
$$\begin{aligned} P(Z_n < 15) &= \sum_{x<15} P(Z_n = x) \\ &= \sum_{i=0}^{14} \binom{100}{i} 0.1^i (1-0.1)^{100-i} \qquad \text{rechenaufwändig} \\ &\approx \Phi\left(\tfrac{15-100\cdot 0.1}{\sqrt{100\cdot 0.1\cdot(1-0.1)}}\right) = \Phi\left(\tfrac{5}{\sqrt{9}}\right) = \Phi\left(\tfrac{5}{3}\right) \approx 0.952 \end{aligned}$$

Bem.: Faustregel: $n \cdot p \geq 10$ und $n \cdot (1-p) \geq 10$.

Satz von Moivre–Laplace: Illustration







Bedeutung des ZGWS beim Schätzen

Gesetz der Großen Zahlen: $\overline{X} \to \mu = \mathbf{E}(X)$.

Frage: Wie groß ist der Stichprobenumfang n zu wählen, um eine bestimmte Genauigkeit zu erreichen?

ε, δ vorgegeben, klein ($\varepsilon, \delta < 0.5$). $\qquad \sigma^2 := Var(X)$
n ist so zu wählen, dass

$$\begin{aligned} 1-\delta &\leq P(|\overline{X}-\mu| \leq \varepsilon) \\ &= P\Big(\underbrace{\sqrt{n}\frac{|\overline{X}-\mu|}{\sigma}}_{\sim\mathcal{N}(0,1)} \leq \sqrt{n}\frac{\varepsilon}{\sigma}\Big) \approx \Phi(\sqrt{n}\frac{\varepsilon}{\sigma}) - \Phi(-\sqrt{n}\frac{\varepsilon}{\sigma}) \\ n &\geq \left(\frac{\sigma\Phi^{-1}(1-\frac{\delta}{2})}{\varepsilon}\right)^2. \end{aligned}$$

Meist ist es anders herum, d.h. es sind n Beobachtungen gegeben, und Sie suchen ein Intervall das den wahren Parameter, hier μ mit vorgegebener Wahrscheinlichkeit $1-\alpha$, hier 0.95, überdeckt. (Das sogenannte Quantil u_β ist

wie folgt definiert: $\Phi(u_\beta) = \beta$.)

$$\begin{aligned}
0.95 &= 0.975 - 0.025 \\
&= \Phi(u_{0.975}) - \Phi(u_{0.025}) = \Phi(1.96) - \Phi(-1.96) \\
&\approx P(\frac{|\overline{X} - \mu|}{\sigma}\sqrt{n} \leq 1.96) \quad \text{nach dem ZGWS} \\
&= P(|\overline{X} - \mu| \leq \frac{\sigma \cdot 1.96}{\sqrt{n}}) = P(-\frac{\sigma \cdot 1.96}{\sqrt{n}} \leq \overline{X} - \mu \leq \frac{\sigma \cdot 1.96}{\sqrt{n}}) \\
&= P(\overline{X} - \frac{\sigma \cdot 1.96}{\sqrt{n}} \leq \mu \leq \overline{X} + \frac{\sigma \cdot 1.96}{\sqrt{n}})
\end{aligned}$$

Das Intervall $[\overline{X} - \frac{\sigma \cdot 1.96}{\sqrt{n}}, \overline{X} + \frac{\sigma \cdot 1.96}{\sqrt{n}}]$ ist ein 0.95-Vertauensintervall für den, hier unbekannten, Parameter μ.

Beispiel 62:
In der BRD gab es im Zeitraum 1970-1990 insgesamt 25 171 123 registrierte Lebendgeburten, davon waren 12 241 392 Mädchen.
Berechnen Sie die ein 95% Vertrauensintervall für die Wahrscheinlichkeit einer Mädchengeburt!

Das zufällige Ereignis einer Mädchengeburt wird dargestellt durch eine Bernoulli-verteilte Zufallsvariable, $X_i \sim Bi(1, p)$. Sei $n = 25171123$ und

$$S_n = \sum_{i=1}^{n} X_i \quad \text{die zufällige Anzahl der Mädchengeburten.}$$

Wir wissen, $\mathbf{E}S_n = n \cdot p$ und $\operatorname{Var} S_n = n \cdot p \cdot (1 - p)$.
Weiter sei $u_{0.975}$ das 0.975-Quantil von $\mathcal{N}(0, 1)$,

$$\Phi(u_{0.975}) = 0.975.$$

Nachsehen in der Tabelle oder Berechnung mit einem Standardprogramm liefert $u_{0.975} \approx 1.96$.
Aus dem Zentralen Grenzwertsatz folgt

$$P(\frac{|S_n - np|}{\sqrt{VarS_n}} \leq u_{0.975}) \approx 0.95.$$

Umstellen der Ungleichung liefert ein 0.95-Konfidenzintervall, hier: $[0.48613, 0.48652]$.

Beispiel 63:: Schätzung des Anteils der Todesfälle unter den registrierten Corona-Infizierten
Am 10.5.2020 gab es in Deutschland n=171324 registrierte Corona-Infizierte, von denen $k = 7545$ (unter anderem) an Corona starben.
Wie im letzten Beispiel erhalten wir ein 0.95-Konfidenzintervall für die (unbekannte) Sterblichkeit p. Das ist ein klein wenig mühselig, geht aber, da die Varianz, bei gegebenem p bekannt ist.
Wir können es uns auch einfacher machen, und die unbekannte Varianz schätzen in dem wir den Parameter p in der Varianzformel ($\operatorname{var}(S_n) = np(1 - p) = \sigma^2$) durch die Schätzung $\hat{p} = \frac{k}{n} = \overline{X}$ ersetzen (vgl. Formel vor dem letzten Beispiel).

Fortsetzung von Beipiel 63
In unserem Fall erhalten wir für die Todesfallrate p ein 0.95-Konfidenzintervall von $[0.043, 0.045]$. Ein 0.99-Konfidenzintervall erhalten Sie wenn Sie in den obigen Formeln das 0.975-Quantil $u_{0.975} = 1.96$ der Standardnormalverteilung ersetzen durch das 0.995-Quantil $u_{0.995} = 2.575$.
Bei einer inhaltlichen Bewertung ist (mindestens) zu beachten,

- Dunkelziffer, es handelt sich nur um die registrierten Fälle.

- Fallzahlen sind regional und zeitlich verschieden. Eine detailliertere Auswertung ist angebracht.

Auch andere Kennzahlen, wie z.B. die Reproduktionszahl R sind nur Schätzungen, und deren Güte kann deutlich schlechter sein.

9 Parameterschätzung

9.1 Eigenschaften von Schätzungen

Sei $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ eine Schätzung eines Parameters θ, die auf n Beobachtungen beruht.

- $\hat{\theta}_n \xrightarrow[n\to\infty]{} \theta$ "Konsistenz" (Minimalforderung)
- $E\hat{\theta}_n = \theta$ "Erwartungstreue" $E\hat{\theta}_n \xrightarrow[n\to\infty]{} \theta$ "Asymptotische Erwartungstreue"
- var $\hat{\theta}_n$ möglichst klein: "gute", "effiziente" Schätzung
- wenn var $\hat{\theta}_n$ den kleinstmöglichen Wert annimmt für alle e-treuen Schätzungen, $\hat{\theta}_n$: "optimale Schätzung"
- MSE = var $\hat{\theta}_n$ + bias^2 $\hat{\theta}_n$ = var $\hat{\theta}_n + (E\hat{\theta}_n - \theta)^2$ $\longrightarrow$ minimal oder möglichst klein.
- Eigenschaften sollten "möglichst" auch bei (kleinen) Abweichungen von der (Normal-)Verteilungsannahme gelten $\longrightarrow$ robuste Schätzung.

9.2 Schätzmethoden

Momentenmethode
Man drückt den zu schätzenden Parameter durch die Momente, z.B. $\mathbf{E}(X), \mathbf{E}(X^2)$, var$(X)$, aus. Dann werden die Momente durch die entsprechenden *empirischen* Momente, z.B. der Erwartungswert durch $\overline{X}$, ersetzt.

Maximum-Likelihood-Schätzung (ML-Schätzung)
Es wird der Schätzwert für den unbekannten Parameter ermittelt, der anhand der vorliegenden Daten, am meisten für diesen Paramter spricht (most likely).

Kleinste-Quadrat-Schätzung (KQS)
Sei θ der zu schätzende Parameter. Man geht aus von einem Modell, z.B.

$$Y_i = g(\theta, X_i) + \epsilon_i$$

Dannn versucht man die Summe der Fehlerquadrate

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - g(\theta, X_i))^2.$$

zu minimieren (Kleinste Quadrate).

9.3 Arithmetisches Mittel

$\overline{X}$ *(als Schätzung für den Erwartungswert)*

- Momentenschätzung
- Kleinste Quadrat Schätzung

- konsistent (Gesetz der großen Zahlen)
- erwartungstreu
- nicht robust (gegen Ausreißer)
- asymptotisch normalverteilt (Zentraler Grenzwertsatz)
- bei Normalverteilung:
  - Maximum-Likelihood-Schätzung
  - minimale Varianz, optimal
  - normalverteilt

10 Zusammenfassung

Fehlerrechnung

- Absolute und relative Fehler
- Kondition von Abbildungen

Nichtlineare Gleichungen

- Intervallschachtelung
- Banachscher Fixpunktsatz
- Iterative Lösung linearer Gleichungssysteme (Jacobi-Verfahren, Gauß-Seidel Verfahren)
- Einfache Iteration
- Newton Verfahren
- Regula Falsi
- Horner Schema

Ausgleichs- und Glättungsverfahren

- Lineare Regression
- Nichtlineare Regression
- Splines (nur Definition und Eigenschaften)

Interpolation und numerische Integration

- Lagrange Interpolation
- Trapezregel
- Simpsonregel

Grundlagen, Bedingte Wahrscheinlichkeit, Zufallsvariablen

- Wahrscheinlichkeitsbegriff
- Rechnen mit Wahrscheinlichkeiten
- Einfache kombinatorische Formeln

- Bedingte Wahrscheinlichkeiten, Unabhängigkeit
- Satz der Totalen Wahrscheinlichkeit
- Satz von Bayes
- Zufallsvariablen, Verteilungsfunktion
- Erwartungwert, Varianz, Rechnen mit Erwartungwert, Varianz

Wahrscheinlichkeitsmodelle

- Diskrete Gleichverteilung
- Binomialverteilung
- Poisson-Verteilung
- Geometrische Verteilung
- Gleichverteilung
- Exponentialverteilung
- Normalverteilung

Unabhängigkeit, Ungleichungen, Grenzwersätze

- Zweidimensionale Zufallsvariablen
- Unabhängigkeit und Korrelation
- Markov-Ungleichung, Tschebyschev-Ungleichung
- Gesetz der Großen Zahlen
- Zentraler Grenzwertsatz