Algorithmische Mathematik II

Vorlesungsskript SS 2010

Mario Bebendorf

Inhaltsverzeichnis

Vorwort

Dieses Skript fasst den Inhalt der von mir im Sommersemester 2010 an der Universität Bonn gehaltenen Vorlesung *Algorithmische Mathematik II* des zweiten Semesters im Bachelorstudiengang Mathematik zusammen und ist eine Überarbeitung des Skripts zu der von Andreas Eberle im Sommersemester 2009 gehaltenen gleichnamigen Vorlesungen. Mein Dank gebührt Herrn Maximilian Kirchner (`mkirchner@uni-bonn.de`), der dieses LaTeX-Dokument aus der Vorlesungsmitschrift erstellt hat. Korrekturvorschläge sind willkommen.

Bonn, 1. August 2010

Einleitung

Die algorithmische Mathematik vereint die algorithmischen Grundlagen aus verschiedenen Bereichen der Mathematik

- Diskrete Mathematik
- Numerische Mathematik
- Stochastik

Die Aufgabe der Algorithmischen Mathematik ist die Konstruktion und Analyse von Algorithmen zur Lösung mathematischer Probleme. Ursprung dieser Probleme können Aufgabenstellungen aus Technik, Naturwissenschaften, Wirtschaft und Sozialwissenschaften sein. Von erheblicher praktische Bedeutung ist deshalb die Umsetzung der Algorithmen in ein Computerprogramm.

Literaturangaben:

- U. Krengel: *Einführung in die Wahrscheinlichkeitstheorie und Statistik*, Vieweg, 2000
- P. Deuflhard und A. Hohmann: *Numerische Mathematik 1 u. 2*, de Gruyter Verlag
- M. Hanke-Bourgeois: *Grundlagen der numerischen Mathematik und des wissenschaftlichen Rechnens*, Teubner-Verlag
- J. Stoer: *Numerische Mathematik I*, Springer-Verlag
- A. Quarteroni, R. Sacco, F. Saleri: *Numerische Mathematik 1,2*, Springer-Verlag 2002

7 Diskrete Zufallsvariablen

Das Ziel dieses Kapitels ist die mathematische Modellierung von Zufallsprozessen. Wir wollen zunächst einige grundlegende Begriffe erklären.

7.1 Grundlegende Begriffe

Mit Ω bezeichnen wir im Folgenden die **Menge aller möglichen Fälle** eines Zufallsvorgangs. Hier werden wir uns ausschließlich mit abzählbaren Mengen $\Omega \neq \emptyset$ beschäftigen.

Beispiel 7.1.

(a) Beim Werfen eines Würfels ist $\Omega = \{1, 2, 3, 4, 5, 6\}$.

(b) Bei der Überprüfung von n verschiedenen Geräten auf Funktionstüchtigkeit ist $\Omega = \{0, 1\}^n$. Die Mächtigkeit dieser Menge ist $|\Omega| = 2^n$.

Jedem **Ereignis** ist eine Teilmenge $A \subset \Omega$ zugeordnet. Als **Elementarereignis** bezeichnet man jedes Element $\omega \in \Omega$.

Beispiel 7.2.

(a) Beim Würfeln ist dem Ereignis "Augenzahl ist gerade" die Menge $A = \{2, 4, 6\}$ zugeordnet.

(b) Beim Überprüfen von n Geräten ist dem Ereignis "es funktionieren mind. 3 Geräte" die Menge $A = \{\omega \in \{0, 1\}^n : \sum_{i=1}^n \omega_i \geq 3\}$ zugeordnet.

Wenn Ereignissen Mengen zugeordnet sind, läßt sich die Kombination von Ereignissen durch Mengenoperationen ausdrücken.

"A oder B tritt ein"	$A \cup B$
"mind. eines der Ereignisse A_i, $i \in I$, tritt ein"	$\bigcup_{i \in I} A_i$
"A und B tritt ein"	$A \cap B$
"jedes der Ereignisse A_i, $i \in I$, tritt ein"	$\bigcap_{i \in I} A_i$
"A tritt nicht ein" (sog. **Komplementärereignis**)	$A^c := \Omega \setminus A$.

Zwei Ereignisse A und B bezeichnet man als **unvereinbar**, falls $A \cap B = \emptyset$. Dem **sicheren Ereignis** entspricht $A = \Omega$, dem **unmöglichen Ereignis** entspricht die Menge $A = \emptyset$.

Beispiel 7.3. Beim Würfeln ist dem Ereignis

(a) "Augenzahl gerade oder kleiner als 4" die Menge $A = \{2, 4, 6\} \cup \{1, 2, 3\} = \{1, 2, 3, 4, 6\}$,

(b) "Augenzahl ist nicht gerade und größer als 5" die Menge $A = \{1, 3, 5\} \cap \{6\} = \emptyset$

zugeordnet.

Satz 7.4 (Rechenregeln für Mengen).

(i) *Kommutativgesetz* $A \cup B = B \cup A$, $A \cap B = B \cap A$,

(ii) *Assoziativgesetz* $(A \cup B) \cup C = A \cup (B \cup C)$, $(A \cap B) \cap C = A \cap (B \cap C)$,

(iii) *Distributivgesetz* $(A \cup B) \cap C = (A \cap B) \cup (B \cap C)$, $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$,

(iv) *De Morgansche Regeln* $(A \cup B)^c = A^c \cap B^c$, $(A \cap B)^c = A^c \cup B^c$.

Sei $\mathcal{A}$ die Menge der im Modell zugelassenen bzw. in Betracht gezogenen Ereignisse. Dann gilt $\mathcal{A} \subset \mathcal{P}(\Omega)$, wobei $\mathcal{P}(\Omega) := \{A : A \subset \Omega\}$ die **Potenzmenge**, d.h. die Menge aller Teilmengen von Ω bezeichnet. Die Menge $\mathcal{A}$ sollte unter obigen Mengenoperationen, d.h. abzählbaren Vereinigungen, Schnitten und Komplementdarstellung abgeschlossen sein.

Definition 7.5. *Eine Menge* $\mathcal{A} \subset \mathcal{P}(\Omega)$ *ist eine* ***σ-Algebra*** *oder* ***Ereignisalgebra****, falls*

(i) $\Omega \in \mathcal{A}$,

(ii) *für alle* $A \in \mathcal{A}$ *gilt* $A^c \in \mathcal{A}$,

(iii) *für* $A_i \in \mathcal{A}$ *gilt* $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$.

Bemerkung. Offenbar ist $\mathcal{P}(\Omega)$ eine σ-Algebra. Für jede σ-Algebra $\mathcal{A}$ gilt auch:

1. nach (i) und (ii) ist $\emptyset = \Omega^c \in \mathcal{A}$,
2. $A_i \in \mathcal{A}$, $i \in \mathbb{N} \overset{\text{(ii), (iii)}}{\Longrightarrow} \bigcap_{i=1}^{\infty} A_i = (\bigcup_{i=1}^{\infty} A_i^c)^c \in \mathcal{A}$,
3. $A, B \in \mathcal{A} \Longrightarrow A \setminus B = A \cap B^c \in \mathcal{A}$.

7.2 Wahrscheinlichkeitsverteilungen

Sei $\mathcal{A} \subset \mathcal{P}(\Omega)$ eine σ-Algebra. Wir wollen nun Ereignissen $A \in \mathcal{A}$ eine Wahrscheinlichkeit (engl. probability) $P(A)$ zuordnen.

Definition 7.6 (Kolmogorovsche Axiome). *Eine Abbildung* $P : \mathcal{A} \to \mathbb{R}$ *wird als* ***Wahrscheinlichkeitsverteilung*** *bezeichnet, falls*

(i) $P(A) \geq 0$ *für alle* $A \in \mathcal{A}$ *(Positivität),*

(ii) $P(\Omega) = 1$ *(Normierung),*

(iii) *für jede paarweise unvereinbare Folge* A_i, $i \in \mathbb{N}$, *(d.h.* $A_i \cap A_j = \emptyset$, $i \neq j$*) gilt*

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

Das Tripel $(\Omega, \mathcal{A}, P)$ *wird als* ***Wahrscheinlichkeitsraum*** *bezeichnet.*

Satz 7.7 (Rechenregeln). *Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum. Dann gilt*

(i) $P(\emptyset) = 0$,

(ii) *für $A, B \in \mathcal{A}$ gilt* $P(A \cup B) = P(A) + P(B) - P(A \cap B) \leq P(A) + P(B)$.
Insbesondere gilt für unvereinbare Ereignisse $P(A \cup B) = P(A) + P(B)$.

(iii) *für $A, B \in \mathcal{A}$, $A \subset B$, gilt* $P(B) = P(A) + P(B \setminus A)$.
Insbesondere gilt $P(B) \geq P(A)$, $P(A^c) = 1 - P(A)$ *und* $P(A) \leq 1$.

Beweis.

(i) Die Ereignisse $A_1 := \Omega$, $A_i = \emptyset$, $i > 1$, sind paarweise unvereinbar. Daher folgt

$$P(\Omega) = P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i) = P(\Omega) + \sum_{i=2}^{\infty} P(\emptyset)$$

und hieraus $P(\emptyset) = 0$.

(ii) Den zweiten Teil von (ii) erhält man aus (i) und

$$P(A \cup B) = P(A \cup B \cup \emptyset \cup \ldots) = P(A) + P(B) + P(\emptyset) + \ldots \overset{(i)}{=} P(A) + P(B).$$

(iii) Für $A \subset B$ ist $B = A \cup (B \setminus A)$. Weil diese Vereinigung disjunkt ist, folgt nach (ii)

$$P(B) = P(A) + P(B \setminus A) \geq P(A).$$

Insbesondere $1 = P(\Omega) = P(A) + P(A^c)$ und somit $P(A) \leq 1$.

(ii) Der erste Teil von (ii) ergibt sich aus (iii)

$$\begin{aligned} P(A \cup B) =& P(A) + P((A \cup B) \setminus A) = P(A) + P(B \cap A^c) \\ =& P(A) + P(B \cap (A^c \cup B^c)) = P(A) + P(B \setminus (A \cap B)) \\ \overset{(iii)}{=}& P(A) + P(B) - P(A \cap B). \end{aligned}$$

□

Nach (iii) gilt für drei Ereignisse $A, B, C \in \mathcal{A}$

$$\begin{aligned} P(A \cup B \cup C) &= P(A \cup B) + P(C) - P((A \cup B) \cap C) \\ &= P(A) + P(B) - P(A \cap B) + P(C) - [P(A \cap C) + P(B \cap C) - P(A \cap B \cap C)] \\ &= P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C). \end{aligned}$$

Die folgende Verallgemeinerung dieser Aussage auf n Ereignisse werden wir am Ende des Kapitels mit Hilfe des Erwartungswerts beweisen.

Korollar 7.8 (Einschluss-/Ausschlussprinzip). Für $n \in \mathbb{N}$ Ereignisse $A_1, \ldots, A_n$ gilt

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{k=1}^{n} (-1)^{k-1} \sum_{1 \le i_1 < \ldots < i_k \le n} P\left(\bigcap_{\ell=1}^{k} A_{i_\ell}\right).$$

7.3 Spezielle Wahrscheinlichkeitsverteilungen

Ist Ω endlich und sind alle Elementarereignisse gleich wahrscheinlich, d.h. gilt

$$P(\{\omega_i\}) = \frac{1}{|\Omega|}, \quad i = 1, \ldots, |\Omega|,$$

so spricht man von einem **Laplace-Modell**. Die entsprechende Wahrscheinlichkeitsverteilung bezeichnet man als **Gleichverteilung**. Aus den kolmogorovschen Axiomen erhält man für $A \in \mathcal{A} := P(\Omega)$

$$P(A) = \frac{|A|}{|\Omega|} = \frac{\text{“Anzahl günstiger Fälle”}}{\text{“Anzahl möglicher Fälle”}}.$$

Beispiel 7.9.

(i) Beim Werfen eines idealen Würfels sind alle Elementarereignisse in $\Omega = \{1, \ldots, 6\}$ gleich wahrscheinlich. Dem Ereignis “eine Primzahl wird gewürfelt” ist $A = \{2, 3, 5\}$ zugeordnet. Es ergibt sich $P(A) = \frac{|A|}{|\Omega|} = \frac{3}{6} = \frac{1}{2}$.

(ii) Beim n-fachen Werfen einer idealen Münze sind die Elementarereignisse Kopf, Zahl gleich wahrscheinlich. Es ist $\Omega = \{0, 1\}$ und $P(\{0\}) = P(\{1\}) = \frac{1}{2}$. Dem Ereignis “es wird abwechselnd Kopf bzw. Zahl geworfen” ist die Menge

$$A = \{(0, 1, 0, 1, \ldots), (1, 0, 1, 0, \ldots)\}$$

zugeordnet. Es gilt $P(A) = \frac{2}{2^n} = 2^{1-n}$.

Die Bestimmung der Mächtigkeit der Ereigniswege A (sog. Kombinatorik) kann je nach Problemstellung kompliziert sein. Im Folgenden geben wir die Mächtigkeit einiger typischer Ereignisse an. Dazu bedienen wir uns exemplarisch eines *Urnenmodells mit m unterscheidbaren Kugeln*. Ziel ist es, die Anzahl von Möglichkeiten beim Ziehen von $n \le m$ Kugeln zu bestimmen. Dabei ist zu berücksichtigen, ob eine entnommene Kugel vor Entnahme zurückgelegt wird und ob die Reihenfolge, in der die Kugeln gezogen werden, eine Rolle spielt.

1. Reihenfolge der Entnahme wird berücksichtigt.
   (a) Anzahl der Möglichkeiten mit Zurücklegen:

$$\underbrace{m \cdot m \cdot \ldots \cdot m}_{n\text{-mal}} = m^n.$$

   (b) Anzahl der Möglichkeiten ohne Zurücklegen:

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

2. Reihenfolge der Entnahme wird nicht berücksichtigt.
   (a) ohne Zurücklegen:

$$\binom{m}{n} = \frac{m!}{(m-n)! \cdot n!} \quad \text{Binomialkoeffizient “}m \text{ über } n\text{”}.$$

Um dies zu zeigen, bezeichne C_m^n die Anzahl der Möglichkeiten. Es gilt $C_1^0 = 1$, $C_m^m = 1$. Wir erhalten

$$\begin{aligned} C_{m+1}^n &= |\{(a_1, \ldots, a_n) : 1 \le a_1 < a_2 < \ldots < a_n \le m+1\}| \\ &= |\{(a_1, \ldots, a_n) : 1 \le a_1 < a_2 < \ldots < a_n \le m\}| \\ &\quad + |\{(a_1, \ldots, a_{n-1}, m+1) : 1 \le a_1 < a_2 < \ldots < a_{n-1} \le m\}| \\ &= C_m^n + C_m^{n-1}. \end{aligned}$$

Nach Induktionsvorraussetzung gilt $C_m^n = \binom{m}{n}$. Also folgt

$$\begin{aligned} C_{m+1}^n = C_m^n + C_m^{n-1} &= \binom{m}{n} + \binom{m}{n-1} \\ &= \frac{m!}{(m-n)!n!} + \frac{m!}{(m-n+1)!(n-1)!} \\ &= \frac{m!}{(m-n)!n!}(m+1-n+n) = \frac{(m+1)!}{(m+1-n)!n!} \\ &= \binom{m+1}{n} \end{aligned}$$

   (b) Anzahl der Möglichkeiten mit Zurücklegen:

$$\binom{m+n-1}{n}.$$

Dies folgt aus der Beobachtung, dass die Menge

$$\{(a_1, \ldots, a_n) : 1 \le a_1 < a_2 < \ldots < a_n \le m\}$$

durch die Bijektion $b_i = a_i + i - 1$ auf die Menge

$$\{(b_1, \ldots, b_n) : 1 \le b_1 < b_2 < \ldots < b_n \le m+n-1\}$$

abgebildet wird. Die Mächtigkeit der letzten Menge ist nach (a) $C_{m+n-1}^m = \binom{m+n-1}{m}$.

Beispiel 7.10 (Lotto 6 aus 49).
Beim Lotto werden 6 Kugeln aus einer Urne mit 49 Kugeln ohne Zurücklegen und ohne Beachtung der Reihenfolge gezogen. Das Ereignis $A_k \mathrel{\widehat{=}}$ “genau k Richtige werden getippt” hat die Mächtigkeit

$$|A_k| = \underbrace{\binom{6}{k}}_{\text{Richtige}} \cdot \underbrace{\binom{43}{6-k}}_{\text{Nieten}}$$

Die Anzahl möglicher Ereignisse Ω ist

$$|\Omega| = \binom{49}{6} = 13\,983\,816.$$

Gleichverteilung vorausgesetzt ergibt sich

$$P(A_k) = \frac{|A_k|}{|\Omega|} = \frac{\binom{6}{k}\binom{43}{6-k}}{\binom{49}{6}}.$$

Wir erhalten

$$\begin{aligned}
k=1: &\quad \tbinom{6}{1}=6 \quad \tbinom{43}{5}=962\,598 \quad P(A_1)=0.413\\
k=2: &\quad \tbinom{6}{2}=15 \quad \tbinom{43}{4}=123\,410 \quad P(A_2)=0.132\\
k=3: &\quad \tbinom{6}{3}=20 \quad \tbinom{43}{3}=12\,341 \quad P(A_3)=0.018\\
k=4: &\quad \tbinom{6}{4}=15 \quad \tbinom{43}{2}=903 \quad P(A_4)=9.7\cdot 10^{-4}\\
k=5: &\quad \tbinom{6}{5}=6 \quad \tbinom{43}{1}=43 \quad P(A_5)=1.8\cdot 10^{-5}\\
k=6: &\quad \tbinom{6}{6}=1 \quad \tbinom{43}{0}=1 \quad P(A_6)=7.2\cdot 10^{-8}
\end{aligned}$$

Beispiel 7.11 (Geburtstagsparadoxon).
Wir wollen die Wahrscheinlichkeit für das Ereignis A "mind. zwei von n Personen haben am gleichen Tag Geburtstag" bestimmen. Dabei setzen wir voraus, dass

(i) keiner am 29.2. Geburtstag hat (Schaltjahrproblem),

(ii) die übrigen 365 Tage als Geburtstag gleich wahrscheinlich sind.

Das Komplementärereignis A^c "alle Geburtstage sind verschieden" ist einfacher zu handhaben. Offenbar ist das Komplementärereignis isomorph zum Urnenmodell " Ziehen ohne Zurücklegen unter Berücksichtigung der Reihenfolge". Daher folgt wegen Gleichverteilung

$$P(A) = 1 - P(A^c) = 1 - \frac{\frac{365!}{(365-n)!}}{365^n}$$

Für $n = 23$ hat man $P(A) > 0.5$, während für $n = 57$ schon $P(A) > 0.99$ gilt.

Empirische Verteilung

Seien $x_1, \ldots, x_n \in \Omega$ Beobachtungsdaten, zum Beispiel n Schuhgrößen aller möglichen Schuhgrößen Ω. Sei

$$N(A) := |\{x_i \in A,\, i = 1, \ldots, n\}|$$

die Anzahl bzw. Häufigkeit der Werte x in A und

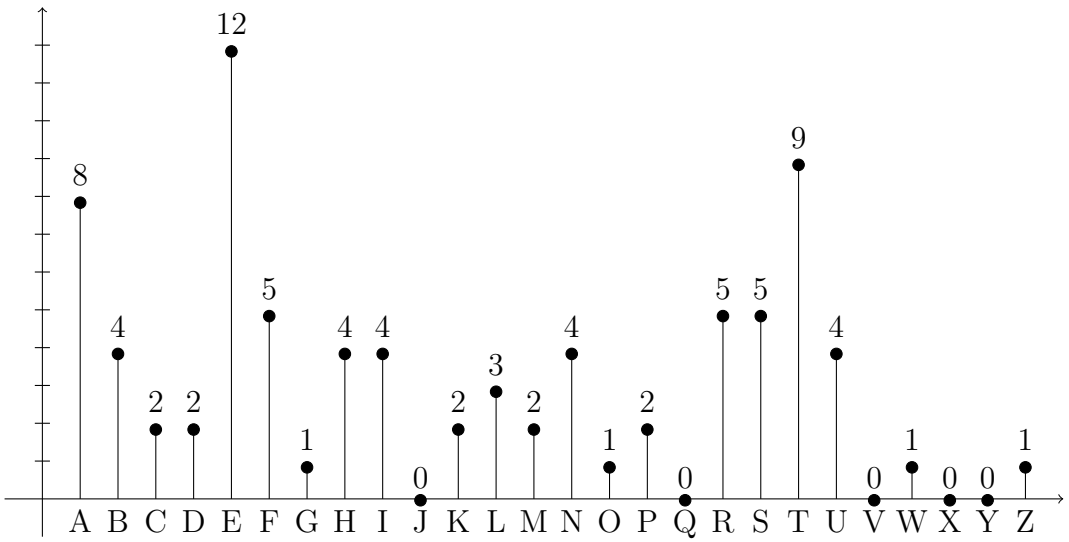
$$P(A) := \frac{N(A)}{n}$$

die relative Häufigkeit der Werte in A. Dann ist P eine Wahrscheinlichkeitsverteilung auf $(\Omega, \mathcal{P}(\Omega))$.

Beispiel 7.12. Laut Guinness-Buch ist das längste veröffentlichte Wort in deutscher Sprache

"Donaudampfschifffahrtselektrizitätenhauptbetriebswerkbauunterbeamtengesellschaft"

mit einer Länge von 81 Buchstaben. Dabei haben wir das "ä" als "ae" gezählt. Für die empirische Verteilung der 26 Buchstaben des Alphabets erhält man



7.4 Zufallsvariablen und ihre Verteilung

Mit Hilfe von Zufallsvariablen können weitere Verteilungen konstruiert werden.

Definition 7.13. *Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum. Eine **diskrete Zufallsvariable** ist eine Abbildung $X : \Omega \to S$ mit einer abzählbaren Menge S, so dass für alle $a \in S$ gilt*

$$X^{-1}(a) := \{\omega \in \Omega : X(\omega) = a\} \in \mathcal{A}.$$

Für das Urbild $X^{-1}(a)$ schreiben wir im Folgenden $\{X = a\}$.
*Die **Verteilung einer Zufallsvariablen** X ist die Wahrscheinlichkeitsverteilung*

$$\mu_X(E) := \sum_{a \in E} P(X = a), \quad E \subset S,$$

auf S. Dabei schreiben wir $P(X = a)$ für $P(\{X = a\})$.

Bemerkung.

(a) In den Übungsaufgaben zeigen wir, dass μ_X tatsächlich eine Wahrscheinlichkeitsverteilung ist.

(b) Für $E \subset S$ gilt

$$\{X \in E\} = \{\omega \in \Omega : X(\omega) \in E\} = \bigcup_{a \in E} \underbrace{\{X = a\}}_{\in \mathcal{A}} \in \mathcal{A}$$

und

$$\mu_X(E) = \sum_{a \in E} P(X = a) = P\left(\bigcup_{a \in E} \{X = a\}\right) = P(X \in E).$$

Die Wahrscheinlichkeitsverteilung μ_X gibt also an, mit welcher Wahrscheinlichkeit die Zufallsvariable X Werte in vorgegebenen Mengen annimmt.

Beispiel 7.14 (Zweimal würfeln). Sei $\Omega = \{(\omega_1, \omega_2) : \omega_i \in S\}$ mit $S := \{1, \dots, 6\}$. Sei P die Gleichverteilung.

(a) Sei $X_i : \Omega \to S$ mit $X_i(\omega) = \omega_i$, $i = 1, 2$. X_i ist eine diskrete Zufallsvariable. Für die Verteilung μ_{X_i} gilt

$$\mu_{X_i}(\{a\}) = P(X_i = a) = \frac{6}{36} = \frac{1}{6} \quad \text{für alle } a \in S.$$

Also ist μ_{X_i} gleichverteilt.

(b) Sei $Y : \Omega \to \{2, 3, \dots, 12\}$ mit $Y(\omega) = X_1(\omega) + X_2(\omega)$ die Summe der Augenzahlen. Dann gilt

$$P(Y = a) = \begin{cases} \frac{1}{36}, & \text{falls } a \in \{2, 12\}, \\ \frac{2}{36}, & \text{falls } a \in \{3, 11\}, \\ \frac{3}{36}, & \text{falls } a \in \{4, 10\}, \\ \dots \end{cases}$$

Also ist Y nicht mehr gleichverteilt.

Binomialverteilung

Wir erinnern uns an das Urnenmodell aus Abschnitt 7.3. Aus einer Urne mit Kugeln S sollen n Kugeln mit Zurücklegen unter Berücksichtigung der Reihenfolge gezogen werden. Dann ist

$$\Omega = S^n = \{\omega = (\omega_1, \dots, \omega_n),\ \omega_i \in S\}.$$

Wir nehmen an, dass alle kombinierten Stichproben gleichwahrscheinlich sind, d.h. P sei die Gleichverteilung auf Ω. Im Folgenden definieren wir zwei Zufallsvariablen.

1. i-ter Stichprobenwert

$$X_i(\omega) := \omega_i \quad \Longrightarrow \quad P(X_i = a) = \frac{|S|^{n-1}}{|\Omega|} = \frac{|S|^{n-1}}{|S|^n} = \frac{1}{|S|} \quad \text{für alle } a \in S.$$

Daher ist μ_{X_i} gleichverteilt auf S.

2. Sei $E \subset S$ eine Merkmalausprägung der Stichprobe, die wir im Folgenden als "Erfolg" bezeichnen (z.B. schwarze Kugel). Dann betrachten wir die Ereignisse $\{X_i \in E\}$ "Erfolg bei der i-ten Stichprobe". Es gilt $\mu_{X_i}(E) = P(X_i \in E) = \frac{|E|}{|S|}$.
Sei $p := \frac{|E|}{|S|}$ die Erfolgswahrscheinlichkeit und

$$N : \Omega \to \{0, 1, \dots, n\} \quad \text{mit} \quad N(\omega) := |\{1 \le i \le n : X_i(\omega) \in E\}|$$

die Anzahl der Einzelstichproben mit Merkmalausprägung E.

Lemma 7.15. *Für $k \in \{0, 1, \dots, n\}$ gilt*

$$P(N = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Beweis. Wir wollen die Mächtigkeit der Menge $\{N = k\}$ bestimmen. Sei $k \in \{0, 1, \ldots, n\}$. Es existieren $\binom{n}{k}$ Möglichkeiten, k Indizes aus $\{1, \ldots, n\}$ auszuwählen, für die ein Erfolg eintritt. Außerdem gibt es $|E|^k$ Möglichkeiten für jeden Erfolg und $|S \setminus E|^{n-k}$ Möglichkeiten für jeden Misserfolg. Daher gilt

$$P(N = k) = \frac{\binom{n}{k}|E|^k|S \setminus E|^{n-k}}{|S|^n} = \binom{n}{k}\left(\frac{|E|}{|S|}\right)^k\left(\frac{|S \setminus E|}{|S|}\right)^{n-k} = \binom{n}{k}p^k(1-p)^{n-k}.$$

□

Definition 7.16. *Sei $n \in \mathbb{N}$ und $p \in [0, 1]$. Die Wahrscheinlichkeitsverteilung*

$$P(\{k\}) = \binom{n}{k}p^k(1-p)^{n-k}$$

*auf $\Omega = \{0, 1, 2, \ldots, n\}$ heißt **Binomialverteilung** mit Parametern p und n*

Bemerkung. Wir weisen darauf hin, dass eine Wahrscheinlichkeitsverteilung bereits durch die Vorgabe der Wahrscheinlichkeiten der Elementarereignisse eindeutig definiert ist. Nach den Kolmogorovschen Axiomen gilt nämlich für $A \subset \Omega$

$$P(A) = \sum_{a \in A} P(\{a\}).$$

Beispiel 7.17. Ein idealer Würfel wird $n = 4$ mal geworfen. Mit welcher Wahrscheinlichkeit werden mindestens 2 Sechsen gewürfelt? Dazu sei E das Ereignis, dass eine Sechs gewürfelt wird. Dann ist $p = 1/6$ und N ist die Zufallsvariable, die die Anzahl der Sechsen bei $n = 4$ Würfen beschreibt. Also gilt

$$\begin{aligned} P(N \geq 2) &= 1 - P(N < 2) = 1 - P(N = 0) - P(N = 1) \\ &= 1 - \binom{4}{0}\left(\frac{1}{6}\right)^0\left(\frac{5}{6}\right)^4 - \binom{4}{1}\left(\frac{1}{6}\right)^1\left(\frac{5}{6}\right)^3 \\ &\approx 0.90355. \end{aligned}$$

Poissonverteilung

Beispiel 7.18 (Warteschlange). Um die Anzahl von Mitarbeitern in einem Callcenter zu planen, möchte ein Betreiber berechnen, mit welcher Wahrscheinlichkeit $P(N = k)$ k Anrufe in einer Stunde eingehen. Wir unterteilen die Stunde in n Intervalle $(\frac{i-1}{n}, \frac{i}{n}]$, $i = 1, \ldots, n$, und nehmen an, dass für große n die Wahrscheinlichkeit, im i-ten Intervall genau ein Anruf zu erhalten,

$$p := \frac{\lambda}{n}, \quad 0 < \lambda \in \mathbb{R},$$

ist. Nach dem Abschnitt zur Binomialverteilung gilt dann

$$P(N = k) \approx \binom{n}{k}p^k(1-p)^{n-k} =: p_{n,\frac{\lambda}{n}}(k).$$

Im folgenden Satz wird der Grenzwert $n \to \infty$ untersucht.

Satz 7.19 (Poissonapproximation der Binomialverteilung).
Sei $\lambda \in (0, \infty)$. Dann gilt

$$\lim_{n\to\infty} p_{n,\frac{\lambda}{n}}(k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \ldots.$$

Beweis. Es gilt

$$\begin{aligned} p_{n,\frac{\lambda}{n}}(k) &= \frac{n!}{(n-k)!k!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{\lambda^k}{k!} \underbrace{\frac{n \cdot (n-1) \cdot \ldots \cdot (n-k+1)}{n^k}}_{\to 1} \underbrace{\left(1 - \frac{\lambda}{n}\right)^n}_{e^{-\lambda}} \underbrace{\left(1 - \frac{\lambda}{n}\right)^{-k}}_{\to 1} \\ &\xrightarrow{n\to\infty} \frac{\lambda^k}{k!} e^{-\lambda} \end{aligned}$$

□

Definition 7.20. *Die Wahrscheinlichkeitsverteilung definiert durch*

$$P(\{k\}) = \frac{\lambda^k}{k!} e^{-\lambda}$$

*auf $\Omega = \{0, 1, 2, \ldots\}$ heißt **Poissonverteilung** mit Parameter λ.*

Wegen Satz 7.19 verwendet man die Poissonverteilung zur näherungsweisen Modellierung der Häufigkeit seltener Ereignisse (z.B. Tippfehler in einem Buch) und somit zur Approximation von Binomialverteilungen mit kleiner Erfolgswahrscheinlichkeit p.

Hypergeometrische Verteilung

Wir betrachten r rote und s schwarze Kugeln in einer Urne, von denen $n \leq min(r, s)$ Kugeln *ohne Zurücklegen* gezogen werden. Wir wollen die Wahrscheinlichkeit dafür bestimmen, dass k rote Kugeln gezogen werden. Sei N eine Zufallsvariable, die die Anzahl gezogener roter Kugeln beschreibt und $m = r + s$. Dann gilt wie in Beispiel 7.10 (Lotto 6 aus 49)

$$P(N = k) = \frac{\binom{r}{k}\binom{m-r}{n-k}}{\binom{m}{n}}, \quad k = 0, 1, \ldots, n.$$

Diese Wahrscheinlichkeitsverteilung wird als **hypergeometrische Verteilung** mit Paremetern m, r und n bezeichnet.

Bemerkung. Für $m, r \to \infty$ bei festem $p = \frac{r}{m}$ und festem n gilt

$$P(N = k) \to \binom{n}{k} p^k (1-p)^{n-k},$$

obwohl im Gegensatz zur Binomialverteilung die Kugeln nicht zurückgelegt werden. Bei großem m ist der Unterschied zwischen Ziehen mit und ohne Zurücklegen vernachlässigbar, weil nur selten dieselbe Kugel zweimal gezogen wird.

7.5 Zufallszahlengeneratoren

Ein (Pseudo-) Zufallszahlengenerator ist ein Algorithmus, der eine Folge von ganzen Zahlen mit Werten zwischen 0 und einem Maximalwert $m-1$ erzeugt. Dabei sind die erzeugten Werte durch eine vorgegebene Klasse statistischer Tests nicht von einer Folge von Stichproben unabhängiger, auf $\{0,1,\ldots,m-1\}$ gleichverteilter Zufallsgrößen unterscheidbar. Ein Zufallszahlengenerator erzeugt also nicht wirklich zufällige Zahlen, sie besitzen aber statistische Eigenschaften, die denen von echten Zufallszahlen in vielerlei (aber nicht in jeder) Hinsicht sehr ähnlich sind.

John von Neumann (1951): "*Anyone who considers arithmetical methods of producing random digits is, of course, in a state of sin.*"

Man unterscheidet zwischen nicht-deterministischen und deterministischen Generatoren. Nichtdeterministisch ist ein Generator dann, wenn er auch bei gleichen Ausgangsbedingungen unterschiedliche Werte liefert. Wir konzentrieren uns auf deterministische Generatoren. Konkret werden Pseudozufallszahlen über eine deterministische Rekurrenzrelation

$$x_{n+1} = f(x_{n-k+1},\ldots,x_n), \quad n = k, k+1, k+2, \ldots,$$

aus Startwerten $x_1,\ldots,x_k$ erzeugt. Wir betrachten folgende Beispiele:

Lineare Kongruenzgeneratoren (LCG)

Der folgende **lineare Kongruenzgenerator** wird in den Laufzeitbibliotheken vieler Programmiersprachen verwendet. Hier betrachtet man

$$x_{n+1} = (ax_n + b) \mod m, \quad n = 0,1,2,\ldots.$$

Dabei sind a, b und m Parameter. Im Fall $b = 0$ spricht man von einem **multiplikativen Kongruenzgenerator**. Pseudozufallszahlen in $[0,1)$ können durch Division mit m generiert werden.

Es existieren m Zustände. Daher muss nach spätestens m Schritten ein früherer Zustand wiederholt werden. Es wird also eine periodische Folge erzeugt, bei der die Periodenlänge wesentlich kleiner als m sein kann. Die maximale Periodenlänge m wird unter folgender Bedingung erreicht.

Satz 7.21 (Knuth). *Die Periodenlänge eines LCG ist genau dann m, wenn*

(i) b und m teilerfremd sind,

(ii) jeder Primfaktor von m teilt $a-1$,

(iii) ist 4 ein Teiler von m, so auch von $a-1$.

Beweis. D. Knuth, "The art of computer programming, vol. 2" □

Der multiplikative Generator muss somit eine Periodenlänge kleiner als m haben.

Beispiel 7.22.

ZX81 Generator	$m = 2^{16} + 1$,	$a = 75$,	$b = 0$
RANDU (IBM 360/361)	$m = 2^{31}$,	$a = 65539$,	$b = 0$
Marsaglia Generator	$m = 2^{32}$,	$a = 69069$,	$b = 1$
rand (Unix)	$m = 2^{31}$,	$a = 1103515245$,	$b = 12345$
rand48 (Unix)	$m = 2^{48}$,	$a = 25214903917$,	$b = 11$

Die durch LCG erzeugten Pseudozufallszahlen enthalten Abhängigkeiten. Dies wird durch den Satz von Marsaglia ausgedrückt.

Satz 7.23 (Marsaglia). *Bildet man aus der Folge x_n die k-Tupel $(x_0, \ldots, x_{k-1})$, $(x_1, \ldots, x_k)$, $(x_2, \ldots, x_{k+1})$, ..., so liegen dies im $\mathbb{R}^k$ auf maximal $\sqrt[k]{m \cdot k!}$ parallelen Hyperebenen.*

Beispiel 7.24 (Hyperebenen bei RANDU). Betrachte drei aufeinanderfolgende durch RANDU generierte Zahlen x_n, x_{n+1}, x_{n+2}

$$\begin{aligned} x_{n+2} &= 65539 x_{n+1} = (2^{16} + 3) x_{n+1} = (2^{16} + 3)^2 x_n = (2^{32} + 6 \cdot 2^{16} + 9) x_n \\ &= (6 \cdot 2^{16} + 9) x_n = (6 \cdot (2^{16} + 3) - 9) x_n = 6 x_{n+1} - 9 x_n. \end{aligned}$$

Dabei sind alle Ausdrücke modulo $m = 2^{31}$ zu verstehen. Wie man sieht, erfüllen die Punkte $p := (x_n, x_{n+1}, x_{n+2})^T$ die Ebenengleichung $p \cdot (9, -6, 1)^T = 0$. Als Folge dieser Abhängigkeit fallen die Punkte (x_n, x_{n+1}, x_{n+2}) auf 15 Hyperebenen im $\mathbb{R}^3$.

Bemerkung. RANDU wurde in den 70er Jahren oft verwendet. Viele Resultate aus dieser Zeit werden daher als "verdächtig" angesehen. Der 1972 vorgestellte Marsaglia-Generator zeigt keine solche Abhängigkeiten im $\mathbb{R}^3$, kann aber das prinzipielle Problem auch nicht beheben. Die aufwändigeren **inversen Kongruenzgeneratoren**

$$x_{n+1} = (a \hat{x}_n + b) \mod m, \quad n = 0, 1, 2, \ldots,$$

wobei $\hat{x} \cdot x = 1$, wenn $x \neq 0$, und $\hat{0} = 0$, haben nicht das Problem der Hyperebenenbildung.

Shift-Register-Generatoren

Die allgemeine Form von **Shift-Register-Generatoren** ist

$$x_{n+k} = \sum_{i=0}^{k-1} a_i x_{n+i} \mod 2,$$

wobei die $a_i \in \{0, 1\}$ sind. Die maximale Periodenlänge 2^k läßt sich mit nur zwei $a_i \neq 0$ erreichen, was einen schnellen Generator ermöglicht. Natürliche Zahlen lassen sich durch Aneinandersetzen der Bits $x_i \in \{0, 1\}$ konstruieren. In diesem Fall erhält man die einfache Form

$$x_n = x_{n-i} + x_{n-j} \mod 2.$$

Günstige Wahlen für (i, j) sind

$(35, 2)$	Tausworth, 1965,
$(23, 2)$	Canavos, 1968,
$(35, 3)$	Witlesey, 1968.

Kombinationen von Zufallszahlengeneratoren

Genertoren lassen sich leicht kombinieren, indem man die von mehreren Generatoren erzeugten Zahlen modulo m addiert. Der **KISS-Generator** (keep it simple and stupid) von Marsaglia kombiniert einen LCG mit zwei Shift-Register-Generatoren und besitzt die Periode 2^{95}.

Trotz aller Bemühungen werden nie wirkliche Zufallszahlen generiert. Daher hat man bei der eigentlichen Simulation nie die Gewissheit, dass das Ergebnis nicht verfälscht wird. Die Qualität von Pseudozufallszahlengeneratoren wird durch statistische Tests von Knuth und das DIEHARD-Paket von Marsaglia als Standard beurteilt.

7.6 Erwartungswert

In diesem Abschnitt stellen wir die Frage, welches Resultat ein Zufallsexperiment im Mittel liefert.

Beispiel 7.25. Beim Würfeln wird jede Augenzahl $i \in \{1, \dots, 6\}$ mit gleicher Wahrscheinlichkeit $\frac{1}{6}$ gewürfelt. Daher erhält man im Mittel die Augenzahl

$$\sum_{i=1}^{6} i \cdot \frac{1}{6} = \frac{1}{6} \cdot \frac{6 \cdot 7}{2} = 3.5.$$

Dies verallgemeinern wir für einen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ mit einer Zufallsvariable $X : \Omega \to S$ mit abzählbarem $S \subset \mathbb{R}$.

Definition 7.26. *Der **Erwartungswert** von X bzgl. P ist definiert als*

$$E(X) = \sum_{a \in S} a \cdot P(X = a),$$

falls die Summe wohldefiniert (d.h. unabhängig von der Summationsreihenfolge) ist.

Beispiel 7.27 (Erwartungswert der Poisson-Verteilung). Sei X Poisson-verteilt mit Parameter λ. Dann gilt

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

Daher kann λ als Erwartungswert oder als mittlere Häufigkeit des Experiments interpretiert werden.

Beispiel 7.28 (Erwartungswert der hypergeometrischen Verteilung). In den Übungen werden wir zeigen, dass für den Erwartungswert der hypergeometrischen Verteilung X mit den Parametern m, r und n gilt $E(X) = n \cdot \frac{r}{m}$.

Beispiel 7.29 (Erwartungswert der charakteristischen Funktion). Sei $A \subset \Omega$ ein Ereignis. Dann wird

$$\chi_A(\omega) := \begin{cases} 1, & \omega \in A, \\ 0, & \text{sonst}, \end{cases}$$

als **charakteristische Funktion** bzw. **Indikatorfunktion** von A bezeichnet. Es gilt

$$E(\chi_A) = 0 \cdot P(\chi_A = 0) + 1 \cdot P(\chi_A = 1) = P(A).$$

Sei nun S eine beliebige abzählbare Menge (nicht notwendigerweise eine Teilmenge von $\mathbb{R}$) und $g : S \to \mathbb{R}$ eine Funktion. Wir definieren die reellwertige Zufallsvariable $g(X) : \Omega \to \mathbb{R}$ durch $\omega \mapsto g(X(\omega))$.

Satz 7.30 (Transformationssatz). *Es gilt*

$$E(g(X)) = \sum_{a \in S} g(a) \cdot P(X = a),$$

falls die Summe wohldefiniert ist.

Beweis. Unter Verwendung der Additivität erhält man

$$\begin{aligned} E(g(x)) &= \sum_{b \in g(S)} b \cdot P(g(X) = b) \\ &= \sum_{b \in g(S)} b \cdot P\left(\bigcup_{a \in g^{-1}(b)} \{X = a\} \right) \\ &= \sum_{b \in g(S)} b \sum_{a \in g^{-1}(b)} P(X = a) \\ &= \sum_{b \in g(S)} \sum_{a \in g^{-1}(b)} g(a) \cdot P(X = a) \\ &= \sum_{a \in S} g(a) \cdot P(X = a). \end{aligned}$$

□

Bemerkung. Insbesondere gilt $E(|X|) = \sum_{a \in S} |a| \cdot P(X = a)$. Ist also $E(|X|)$ endlich, so konvergiert $E(X)$ absolut.

Satz 7.31 (Linearität des Erwartungswertes). *Seien $X : \Omega \to S_X \subset \mathbb{R}$ und $Y : \Omega \to S_Y \subset \mathbb{R}$ diskrete Zufallsvariablen auf $(\Omega, \mathcal{A}, P)$, für die $E(|X|)$ und $E(|Y|)$ endlich sind. Dann gilt*

$$E(\alpha X + \beta Y) = \alpha E(X) + \beta E(Y) \quad \text{für alle } \alpha, \beta \in \mathbb{R}.$$

Beweis. Wir definieren $g : S_X \times S_Y \to \mathbb{R}$ durch $g(x, y) = \alpha x + \beta y$. Dann ist $g(X, Y) =$

$\alpha X + \beta Y$ eine Zufallsvariable. Mit dem Transformationssatz folgt

$$\begin{aligned} E(\alpha X + \beta Y) &= E(g(X,Y)) \\ &= \sum_{a \in S_X} \sum_{b \in S_Y} g(a,b) P(X=a, Y=b) \qquad (7.1) \\ &= \sum_{a \in S_X} \sum_{b \in S_Y} (\alpha a + \beta b) P(X=a, Y=b) \\ &= \alpha \sum_{a \in S_X} a \sum_{b \in S_Y} P(X=a, Y=b) + \beta \sum_{b \in S_Y} b \sum_{a \in S_X} P(X=a, Y=b) \\ &= \alpha \sum_{a \in S_X} a P(X=a) + \beta \sum_{b \in S_Y} b P(Y=b) \\ &= \alpha E(X) + \beta E(Y). \end{aligned}$$

Hierbei konvergiert die Reihe (7.1) absolut, weil

$$\begin{aligned} \sum_{a \in S_X} \sum_{b \in S_Y} (\alpha a + \beta b) P(X=a, Y=b) &\leq |\alpha| \underbrace{\sum_{a \in S_X} |a| P(X=a)}_{E(|X|)} + |\beta| \underbrace{\sum_{b \in S_Y} |b| P(Y=b)}_{E(|X|)} \\ &\leq |\alpha| E(|X|) + |\beta| E(|Y|) \end{aligned}$$

nach Voraussetzung endlich ist. □

Korollar 7.32 (Monotonie des Erwartungswertes). Seien die Voraussetzungen von Satz 7.31 erfüllt und gelte $X(\omega) \leq Y(\omega)$ für alle $\omega \in \Omega$. Dann gilt $E(X) \leq E(Y)$.

Beweis. Wegen $(Y - X)(\omega) \geq 0$ für alle $\omega \in \Omega$ ist $E(Y - X) \geq 0$. Mit der Linearität des Erwartungswetes erhält man

$$E(Y) - E(X) = E(Y - X) \geq 0.$$

□

Wir kehren zum Beweis des Einschluss-/Ausschlussprinzips (Korollar 7.8)

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{k=1}^{n} (-1)^{k-1} \sum_{1 \leq i_1 < \ldots < i_k \leq n} P\left(\bigcap_{\ell=1}^{k} A_{i_\ell}\right)$$

zurück.

Beweis zu Korollar 7.8. Wir betrachten das Komplementärereignis. Es gilt unter Verwen-

dung der charakteristischen Funktion χ aus Beispiel 7.29

$$\begin{aligned}
P((A_1 \cup \ldots \cup A_n)^c) &= P(A_1^c \cap \ldots \cap A_n^c) = E(\chi_{A_1^c \cap \ldots \cap A_n^c}) \\
&= E\left(\prod_{i=1}^{n} \chi_{A_i^c}\right) = E\left(\prod_{i=1}^{n} (1 - \chi_{A_i})\right) \\
&= E\left(\sum_{k=0}^{n} (-1)^k \sum_{1 \le i_1 < \ldots < i_k \le n} \prod_{\ell=1}^{k} \chi_{A_{i_\ell}}\right) \\
&= \sum_{k=0}^{n} (-1)^k \sum_{1 \le i_1 < \ldots < i_k \le n} E\left(\prod_{\ell=1}^{k} \chi_{A_{i_\ell}}\right) \\
&= \sum_{k=0}^{n} (-1)^k \sum_{1 \le i_1 < \ldots < i_k \le n} E(\chi_{A_{i_1} \cap \ldots \cap A_{i_k}}) \\
&= \sum_{k=0}^{n} (-1)^k \sum_{1 \le i_1 < \ldots < i_k \le n} P(A_{i_1} \cap \ldots \cap A_{i_k}).
\end{aligned}$$

Die Behauptung folgt aus

$$P(A_1 \cup \ldots \cup A_n) = 1 - P((A_1 \cup \ldots \cup A_n)^c).$$

□

8 Bedingte Wahrscheinlichkeit und Unabhängigkeit

8.1 Bedingte Wahrscheinlichkeit

Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum. Das Konzept der bedingten Wahrscheinlichkeit berücksichtigt bei der Wahrscheinlichkeit eines Ereignisses $A \in \mathcal{A}$, ob ein anderes Ereignis $B \in \mathcal{A}$ eintritt. Anstelle aller möglichen Fälle $\omega \in \Omega$ werden nur die *relevanten* Fälle $\omega \in B$ berücksichtigt. Die günstigen Fälle sind daher $\omega \in A \cap B$.

Definition 8.1. *Seien $A, B \in \mathcal{A}$ mit $P(B) > 0$. Dann heißt*

$$P(A|B) := \frac{P(A \cap B)}{P(B)}$$

*die **bedingte Wahrscheinlichkeit von** A **gegeben** B.*

Bemerkung.

(a) $P(\,\cdot\,|B) : A \mapsto P(A|B)$ ist eine Wahrscheinlichkeitsverteilung auf $(\Omega, \mathcal{A})$.

(b) Ist P die Gleichverteilung auf einer endlichen Menge Ω, so gilt

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

Beispiel 8.2 (Zweimal würfeln). Mit einem idealen Würfel werden zwei Würfe ausgeführt. Es seien A, B die folgenden Ereignisse:

$$A \mathrel{\widehat{=}} \text{“beim ersten Wurf wird eine 1 gewürfelt”},$$
$$B \mathrel{\widehat{=}} \text{“die Augensumme beider Würfe ist 6”}.$$

Die Augensumme zweier Würfel ist in folgender Tabelle zusammengefasst:

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Man sieht, dass 5 der Ereignisse günstig für B sind und 6 Ereignisse für A. Also gilt $P(A) = \frac{1}{6}$ und $P(B) = \frac{5}{36}$. Nur im Fall, dass erst eine 1 und dann eine 5 gewürfelt wird, treten A und

B ein. Es gilt also $P(A \cap B) = \frac{1}{36}$. Die Anzahl der Ereignisse für A im Fall, dass B eintritt, ist 1 von 5 relevanten. Es gilt also

$$P(A|B) = \frac{1}{5} = \frac{P(A \cap B)}{P(B)}.$$

Satz 8.3 (Rechenregeln). *Es gilt*

(i) $P(A|A) = 1$,

(ii) $P(A^c|B) = 1 - P(A|B)$,

(iii) $P(A \cup B|C) = P(A|C) + P(B|C) - P(A \cap B|C)$.

Beweis.

(i) $P(A|A) = \frac{P(A \cap A)}{P(A)} = \frac{P(A)}{P(A)} = 1$.

(ii) Wegen $A \cup A^c = \Omega$ folgt $(A \cap B) \cup (A^c \cap B) = B$ und somit

$$1 = \frac{P(B)}{P(B)} = \frac{P((A \cap B) \cup (A^c \cap B))}{P(B)} = \frac{P(A \cap B)}{P(B)} + \frac{P(A^c \cap B)}{P(B)} = P(A|B) + P(A^c|B).$$

(iii)

$$\begin{aligned} P(A \cup B|C) &= \frac{P((A \cup B) \cap C)}{P(C)} = \frac{P((A \cap C) \cup (B \cap C))}{P(C)} \\ &= \frac{P(A \cap C) + P(B \cap C) - P(A \cap B \cap C)}{P(C)} \\ &= P(A|C) + P(B|C) - P(A \cap B|C). \end{aligned}$$

□

Aus Definition 8.1 erhält man

$$P(A \cap B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A), \tag{8.1}$$

falls $P(B) > 0$ bzw. $P(A) > 0$.

Beispiel 8.4. Nach einer Statistik besitzt das Ereignis

$$B \mathrel{\widehat{=}} \text{“Studierender schließt mit Note } 1, 2 \text{ oder } 3 \text{ ab”}$$

in der Mathematik die Wahrscheinlichkeit 80%. Von den Studierenden aus B genügen mit 25%iger Wahrscheinlichkeit dem Ereignis

$$A \mathrel{\widehat{=}} \text{“Studierender schließt mit Note 1 oder 2 ab”}.$$

Wie hoch ist die Wahrscheinlichkeit für das Ereignis A? Nach (8.1) gilt

$$P(A) = P(A \cap B) = P(A|B) \cdot P(B) = 0.25 \cdot 0.8 = 0.2.$$

Allgemeiner als (8.1) gilt

Satz 8.5 (Satz von der totalen Wahrscheinlichkeit). *Sei $B_1, B_2, \ldots$ eine Zerlegung von Ω, d.h. $\bigcup_{i=1}^{\infty} B_i = \Omega$, $B_i \cap B_j = \emptyset$, $i \neq j$. Dann gilt für $A \in \mathcal{A}$*

$$P(A) = \sum_{\substack{i=1 \\ P(B_i)>0}}^{\infty} P(A|B_i) \cdot P(B_i).$$

Beweis. Weil die B_i eine Zerlegung von Ω bilden, ist

$$A = A \cap \Omega = A \cap \bigcup_{i=1}^{\infty} B_i = \bigcup_{i=1}^{\infty} (A \cap B_i)$$

eine Zerlegung von A. Dann folgt wegen der paarweisen Unvereinbarkeit der $A \cap B_i$, $i = 1, 2, \ldots$

$$P(A) = \sum_{i=1}^{\infty} P(A \cap B_i) \overset{(8.1)}{=} \sum_{\substack{i=1 \\ P(B_i)>0}}^{\infty} P(A|B_i) \cdot P(B_i).$$

□

Beispiel 8.6. Urne A enthalte 3 rote und 4 schwarze Kugeln. Urne B enthalte 2 rote und 5 schwarze Kugeln. Wir legen eine Kugeln K_1 von Urne A in Urne B und ziehen eine Kugel K_2 aus Urne B. Mit welcher Wahrscheinlichkeit ist die gezogene Kugel K_2 rot?

$$\begin{aligned} P(K_2 \text{ rot}) &= P(K_2 \text{ rot}|K_1 \text{ rot}) \cdot P(K_1 \text{ rot}) + P(K_2 \text{ rot}|K_1 \text{ schwarz}) \cdot P(K_1 \text{ schwarz}) \\ &= \frac{3}{8} \cdot \frac{3}{7} + \frac{2}{8} \cdot \frac{4}{7} = \frac{17}{56}. \end{aligned}$$

Bayessche Regel

Die folgende Bayessche Regel stellt einen Zusammenhang zwischen der subjektiven Wahrscheinlichkeit, also der Wahrscheinlichkeit, die man einem Ereignis zubilligen würde, und dem Lernen aus Erfahrung her. Mit anderen Worten dient diese Regel zum Überprüfen von Hypothesen anhand neuer Indizien.

Beispiel 8.7. Wie groß ist die Wahrscheinlichkeit, dass man das Rentenalter erreicht? Jeder hat eine subjektive Wahrscheinlichkeit, sog. *a priori degree of belief*. Wenn zusätzliche Informationen (sog. *likelihood*) existieren, z.B. alle anderen Verwandten sind über 80 Jahre alt geworden, würde jeder seine subjektive Erwartungshaltung revidieren, sog. *a posteriori degree of belief*.

Technisch wird durch die Bayessche Regel $P(B_i|A)$ durch $P(A|B_i)$ ausgedrückt. Aus Satz 8.5 erhält man

Satz 8.8 (Bayessche Regel). *Sei $\{B_i\}$ eine Zerlegung von Ω. Für $A \in \mathcal{A}$ mit $P(A) > 0$ gilt*

$$P(B_j|A) = \frac{P(A|B_j) \cdot P(B_j)}{P(A)} = \frac{P(A|B_j) \cdot P(B_j)}{\sum_{\substack{i=1 \\ P(B_i)>0}}^{\infty} P(A|B_i) \cdot P(B_i)}$$

für alle $j = 1, 2, \ldots$ mit $P(B_j) > 0$.

Beweis. Es gilt

$$P(B_j|A) = \frac{P(A \cap B_j)}{P(A)} = \frac{P(A|B_j) \cdot P(B_j)}{\sum\limits_{\substack{i=1 \\ P(B_i)>0}}^{\infty} P(A|B_i) \cdot P(B_i)}.$$

□

Beispiel 8.9. In einer Bevölkerungsgruppe sei die Wahrscheinlichkeit für das Ereignis A, eine bestimmte Krankheit zu haben, $P(A) = 0.0002$. Um zu ermittelt, ob eine Personen diese Krankheit hat, wird ein Test verwendet, von dem der Hersteller garantiert, dass er die Krankheit zu 99% erkennt und nur zu 1% falsch anschlägt, obwohl keine Krankheit vorliegt. Mit B bezeichnen wir das Ereignis, dass der Test positiv ausfällt.
Mit welcher Wahrscheinlichkeit liegt eine Erkrankung vor, wenn der Test positiv ausfällt? Man beachte, dass die Informationen, wie der Test ausfällt, falls eine Erkrankung vorliegt, uns bereits bekannt ist. Es gilt $P(A) = 0.0002$, $P(B|A) = 0.99$, $P(B|A^c) = 0.01$ und somit

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B|A^c) \cdot P(A^c) + P(B|A) \cdot P(A)} = \frac{0.99 \cdot 0.0002}{0.01 \cdot 0.9998 + 0.99 \cdot 0.0002} \approx 0.019.$$

Obwohl der Test ihn als krank einschätzt, ist der Patient nur mit einer Wahrscheinlichkeit von etwa 2% tatsächlich krank. Der Grund für dieses überraschende Ergebnis ist, dass die Wahrscheinlichkeit, erkrankt zu sein, etwa um das Fünfzigfache geringer ist als die Wahrscheinlichkeit eines falschen Testergebnisses.

Auf der anderen Seite ist die Wahrscheinlichkeit, bei negativem Test tatsächlich gesund zu sein, aber ausreichend hoch:

$$P(A^c|B^c) = \frac{P(B^c|A^c) \cdot P(A^c)}{P(B^c|A^c) \cdot P(A^c) + P(B^c|A) \cdot P(A)} = \frac{0.99 \cdot 0.9998}{0.99 \cdot 0.9998 + 0.01 \cdot 0.0002} \approx 1.0.$$

8.2 Unabhängigkeit von Ereignissen

Definition 8.10. *Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum. Zwei Ereignisse $A, B \in \mathcal{A}$ heißen* ***unabhängig*** *bzgl. P, falls*

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

Eine Kollektion $A_i \in \mathcal{A}$, $i \in I$, von Ereignissen heißt **unabhängig** *bzgl. P, wenn für alle $n \in \mathbb{N}$ und alle paarweise verschiedenen $i_1, \ldots, i_n \in I$ gilt*

$$P(A_{i_1} \cap \ldots \cap A_{i_n}) = \prod_{k=1}^{n} P(A_{i_k}).$$

Bemerkung. Seien $A, B \in \mathcal{A}$ mit $P(B) > 0$. Dann sind A und B genau dann unabhängig, wenn $P(A|B) = P(A)$.

Beispiel 8.11. Sei $\Omega = \{1,2,3,4\}$ mit der Gleichverteilung P. Die Ereignisse $A = \{1,3\}$, $B = \{2,3\}$ und $C = \{1,2\}$ besitzen die Wahrscheinlichkeiten

$$P(A) = P(B) = P(C) = \frac{1}{2}.$$

Wegen

$$\begin{aligned} P(A \cap B) = P(\{3\}) &= \frac{1}{4} = P(A) \cdot P(B), \\ P(A \cap C) = P(\{1\}) &= \frac{1}{4} = P(A) \cdot P(C), \\ P(B \cap C) = P(\{2\}) &= \frac{1}{4} = P(B) \cdot P(C) \end{aligned}$$

sind die Paare $(A,B), (A,C)$ und (B,C) unabhängig. Allerdings ist die Kollektion (A,B,C) nicht unabhängig, denn es gilt

$$P(A \cap B \cap C) = P(\emptyset) = 0 \neq \frac{1}{8} = P(A) \cdot P(B) \cdot P(C).$$

Lemma 8.12. *Seien die Ereignisse $A_1, \ldots, A_n \in \mathcal{A}$ unabhängig und sei $B_j = A_j$ oder $B_j = A_j^c$, $j = 1, \ldots, n$. Dann sind die Ereignisse $B_1, \ldots, B_n$ unabhängig.*

Beweis. Sei o.B.d.A. $B_1 = A_1, \ldots, B_k = A_k$ und $B_{k+1} = A_{k+1}^c, \ldots, B_n = A_n^c$. Dann gilt wegen der Linearität des Erwartungswertes

$$\begin{aligned} P(B_1 \cap \ldots \cap B_n) &= P(A_1 \cap \ldots \cap A_k \cap A_{k+1}^c \cap \ldots \cap A_n^c) \\ &= E(\chi_{A_1} \cdot \ldots \cdot \chi_{A_k} \cdot (1 - \chi_{A_{k+1}}) \cdot \ldots \cdot (1 - \chi_{A_n})) \\ &= E\left(\chi_{A_1} \cdot \ldots \cdot \chi_{A_k} \cdot \sum_{J \subset \{k+1,\ldots,n\}} (-1)^{|J|} \prod_{j \in J} \chi_{A_j}\right) \\ &= \sum_{J \subset \{k+1,\ldots,n\}} (-1)^{|J|} E\left(\chi_{A_1} \cdot \ldots \cdot \chi_{A_k} \cdot \prod_{j \in J} \chi_{A_j}\right) \\ &= \sum_{J \subset \{k+1,\ldots,n\}} (-1)^{|J|} P\left(A_1 \cap A_2 \cap \ldots \cap A_k \cap \bigcap_{j \in J} A_j\right) \\ &= \sum_{J \subset \{k+1,\ldots,n\}} (-1)^{|J|} P(A_1) \cdot \ldots \cdot P(A_k) \cdot \prod_{j \in J} P(A_j) \\ &= P(A_1) \cdot \ldots \cdot P(A_k) \cdot (1 - P(A_{k+1})) \cdot \ldots \cdot (1 - P(A_n)) \\ &= P(B_1) \cdot \ldots \cdot P(B_n). \end{aligned}$$

□

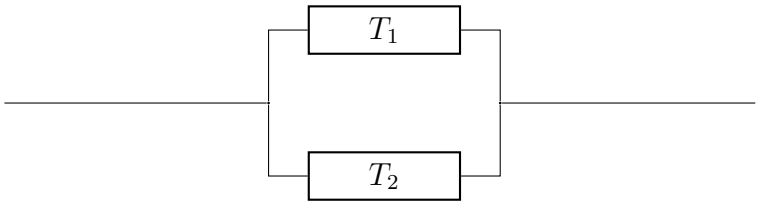
Beispiel 8.13 (Serien- und Parallelschaltung). Ein Gerät bestehe aus 2 Bauteilen T_1 und T_2, bei denen unabhängig voneinander Defekte auftreten können. Die unabhängigen Ereignisse A_1 mit $P(A_1) = p_1$ und A_2 mit $P(A_2) = p_2$ treten auf, wenn das jeweilige Bauteil funktioniert.

(1) Serienschaltung

Eine Serienschaltung funktioniert, falls sowohl T_1 als auch T_2 funktionieren, d.h.

$$P(A_1 \cap A_2) = P(A_1) \cdot P(A_2) = p_1 \cdot p_2.$$

(2) Parallelschaltung



Eine Parallelschaltung funktioniert, falls T_1 oder T_2 funktionieren, d.h.

$$\begin{aligned} P(A_1 \cup A_2) &= 1 - P(A_1^c \cap A_2^c) \\ &= 1 - P(A_1^c) \cdot P(A_2^c) \\ &= 1 - (1-p_1) \cdot (1-p_2). \end{aligned}$$

Verteilung für unabhängige Ereignisse

Beispiel 8.14. Ein Automat sei so eingerichtet, dass er sofort anhält, sobald ein defektes Teil produziert wird. Die Wahrscheinlichkeit dafür, dass ein Teil defekt ist, sei p. Die Defekte sind von produziertem Teil zu produziertem Teil unabhängig. Mit A_n bezeichnen wird das Ereignis, dass das n-te Teil defekt ist, und die Zufallsvariable

$$X(\omega) := \inf\{n \in \mathbb{N} : \omega \in A_{n+1}\}$$

beschreibe die Anzahl der produzierten einwandfreien Teile. Dann gilt

$$\begin{aligned} P(X=0) &= P(A_1) && = p, \\ P(X=1) &= P(A_1^c \cap A_2) = P(A_1^c) \cdot P(A_2) && = (1-p) \cdot p, \\ P(X=2) &= P(A_1^c \cap A_2^c \cap A_3) && = (1-p)^2 \cdot p, \\ &\vdots \\ P(X=n) &= P(A_1^c \cap \ldots \cap A_n^c \cap A_{n+1}) && = (1-p)^n \cdot p. \end{aligned}$$

Definition 8.15. *Sei $p \in [0,1]$. Die Wahrscheinlichkeitsverteilung P auf $\mathbb{N}$ definiert durch*

$$P(\{n\}) = p \cdot (1-p)^n$$

*heißt **geometrische Verteilung** zum Parameter p.*

Bemerkung.

(1) Für $p \neq 0$ gilt

$$\sum_{n=0}^{\infty} p \cdot (1-p)^n = p \cdot \frac{1}{1-(1-p)} = p \cdot \frac{1}{p} = 1.$$

(2) Das Für den Erwartungswert ergibt sich

$$\begin{aligned} E(X) &= \sum_{n=0}^{\infty} P(X=n) \cdot n = \sum_{n=0}^{\infty} p(1-p)^n \cdot n = p(1-p) \sum_{n=0}^{\infty} (1-p)^{n-1} \cdot n \\ &= p(1-p) \frac{1}{(1-(1-p))^2} = \frac{1-p}{p}. \end{aligned}$$

Dabei haben wir

$$\sum_{n=0}^{\infty} n x^{n-1} = \frac{d}{dx} \left(\sum_{n=0}^{\infty} x^n \right) = \frac{d}{dx} \frac{1}{1-x} = \frac{1}{(1-x)^2}$$

für $0 < x < 1$ verwendet.

Seien $A_1, \ldots, A_n \in \mathcal{A}$ unabhängige Ereignisse mit $P(A_i) = p \in [0,1]$. Wir haben in Beispiel 8.14 gesehen, dass die Zufallsvariable X auf die geometrische Verteilung führt. Sei

$$S_n(\omega) := |\{1 \leq i \leq n : \omega \in A_i\}| = \sum_{i=1}^{n} \chi_{A_i}(\omega)$$

die Anzahl der Ereignisse unter $A_1, \ldots, A_n$, die eintreten. Dann gilt

$$\begin{aligned} P(S_n = k) &= \sum_{\substack{I \subset \{1,\ldots,n\} \\ |I|=k}} P\left(\bigcap_{i \in I} A_i \cap \bigcap_{i \in I^c} A_i^c \right) \\ &= \sum_{\substack{I \subset \{1,\ldots,n\} \\ |I|=k}} \prod_{i \in I} P(A_i) \prod_{i \in I^c} P(A_i^c) \\ &= \sum_{\substack{I \subset \{1,\ldots,n\} \\ |I|=k}} p^k (1-p)^{n-k} \\ &= \binom{n}{k} p^k (1-p)^{n-k}. \end{aligned}$$

Für die Anzahl eintretender unabhängiger Ereignisse erhält man also die Binomialverteilung. Für den Erwartungswert von S_n gilt wegen der Linearität des Erwartungswertes und Beispiel 7.29

$$E(S_n) = \sum_{i=1}^{n} E(\chi_{A_i}) = \sum_{i=1}^{n} P(A_i) = p \cdot n.$$

8.3 Mehrstufige diskrete Modelle

Wir betrachten ein n-stufiges Zufallsexperiment. Sind die Mengen aller möglichen Fälle $\Omega_1, \ldots, \Omega_n$ der Teilexperimente abzählbar, dann ist

$$\Omega := \Omega_1 \times \ldots \times \Omega_n$$

die Menge der möglichen Fälle des Gesamtexperimentes. Für $\omega \in \Omega$ und $k \in \{1, \dots, n\}$ sei $X_k(\omega) = \omega_k$ der Ausgang des k-ten Teilexperiments. Angenommen, wir kennen

$$P(X_1 = a_1) =: p_1(a_1) \quad \text{für alle } a_1 \in \Omega_1 \tag{8.2}$$

sowie

$$P(X_k = a_k | X_1 = a_1, \dots, X_{k-1} = a_{k-1}) =: p_k(a_k | a_1, \dots, a_{k-1}) \tag{8.3}$$

die bedingte Wahrscheinlichkeit von X_k gegeben $X_1, \dots, X_{k-1}$ für $k = 2, \dots, n$ mit

$$P(X_1 = a_1, \dots, X_{k-1} = a_{k-1}) \neq 0.$$

Der folgende Satz gibt Auskunft über die Wahrscheinlichkeitsverteilung P auf $(\Omega, \mathcal{A})$.

Satz 8.16. *Seien durch $p_k(\cdot\,|a_1, \dots, a_{k-1})$, $k = 1, \dots, n$, mit $a_i \in \Omega_i$ Wahrscheinlichkeitsverteilungen auf Ω_k definiert. Dann existiert genau eine Wahrscheinlichkeitsverteilung P auf $(\Omega, \mathcal{A})$ mit (8.2), (8.3). Diese erfüllt*

$$P(X_1 = a_1, \dots, X_n = a_n) = p_1(a_1) \cdot p_2(a_2|a_1) p_3(a_3|a_1, a_2) \cdot \ldots \cdot p_n(a_n|a_1, \dots, a_{n-1}). \tag{8.4}$$

Beweis.

(a) Eindeutigkeit
Wir zeigen induktiv, dass für eine Wahrscheinlichkeitsverteilung P mit (8.2) und (8.3) die Eigenschaft (8.4) folgt. (8.2) liefert den Induktionsanfang. Sei die Behauptung für $k-1$ wahr. Dann folgt nach der Induktionsvoraussetzung und (8.3)

$$\begin{aligned} &P(X_1 = a_1, \dots, X_k = a_k) \\ &\quad = P(X_1 = a_1, \dots, X_{k-1} = a_{k-1}) \cdot P(X_1 = a_1, \dots, X_k = a_k | X_1 = a_1, \dots, X_{k-1} = a_{k-1}) \\ &\quad = p_1(a_1) \cdot \ldots \cdot p_{k-1}(a_{k-1}|a_1, \dots, a_{k-2}) \cdot p_k(a_k|a_1, \dots, a_{k-1}), \end{aligned}$$

falls $P(X_1 = a_1, \dots, X_{k-1} = a_{k-1}) \neq 0$. Andernfalls verschwinden beide Seiten und die Behauptung gilt trivialerweise.

(b) Existenz
P aus (8.4) ist eine Wahrscheinlichkeitsverteilung auf Ω, weil

$$\begin{aligned} &\sum_{a_1 \in \Omega_1} \dots \sum_{a_n \in \Omega_n} P(X_1 = a_1, \dots, X_n = a_n) \\ &\qquad = \sum_{a_1 \in \Omega_1} \dots \sum_{a_n \in \Omega_n} p_1(a_1) \cdot p_2(a_2|a_1) \cdot \ldots \cdot p_n(a_n|a_1, \dots, a_{n-1}) \\ &\qquad = \sum_{a_1 \in \Omega_1} p_1(a_1) \underbrace{\sum_{a_2 \in \Omega_2} p_2(a_2|a_1) \cdot \ldots \cdot \underbrace{\sum_{a_n \in \Omega_n} p_n(a_n|a_1, \dots, a_{n-1})}_{=1}}_{=1} \\ &\qquad = 1. \end{aligned}$$

Außerdem gilt

$$\begin{aligned} P(X_1 = a_1, \ldots, X_k = a_k) &= \sum_{a_{k+1} \in \Omega_{k+1}} \ldots \sum_{a_n \in \Omega_n} P(X_1 = a_1, \ldots, X_n = a_n) \\ &= \sum_{a_{k+1} \in \Omega_{k+1}} \ldots \sum_{a_n \in \Omega_n} p_1(a_1) \cdot \ldots \cdot p_n(a_n | a_1, \ldots, a_{n-1}) \\ &= p_1(a_1) \cdot p_2(a_2|a_1) \cdot \ldots \cdot p_k(a_k|a_1, \ldots, a_{k-1}). \end{aligned}$$

Hieraus folgen (8.2) und (8.3). □

Beispiel 8.17. Wie groß ist die Wahrscheinlichkeit, dass beim Skat jeder der drei Spieler genau einen der vier Buben erhält? Sei

$$\Omega = \{(\omega_1, \omega_2, \omega_3) : \omega_i \in \{0, 1, 2, 3, 4\}\}$$

mit $X_i(\omega) = \omega_i$ Anzahl der Buben von Spieler i. Es gilt entsprechend der hypergeometrischen Verteilung

$$\begin{aligned} p_1(a_1) &= \frac{\binom{4}{a_1}\binom{28}{10-a_1}}{\binom{32}{10}}, \\ p_2(a_2|a_1) &= \frac{\binom{4-a_1}{a_2}\binom{18+a_1}{10-a_2}}{\binom{22}{10}}, \\ p_3(a_3|a_1, a_2) &= \begin{cases} \frac{\binom{4-a_1-a_2}{a_3}\binom{8+a_1+a_2}{10-a_3}}{\binom{12}{10}} & \text{falls } 2 \leq a_1 + a_2 + a_3 \leq 4, \\ 0 & \text{sonst.} \end{cases} \end{aligned}$$

Hieraus folgt

$$P(X_1 = 1, X_2 = 1, X_3 = 1) = p_1(1) \cdot p_2(1|1) \cdot p_3(1|1, 1) \approx 5,56\%.$$

Im Folgenden werden zwei Klassen von mehrstufigen Modellen, Produktmodelle und Markov-Ketten betrachtet.

Produktmodelle

Angenommen, der Ausgang des i-ten Experiments hängt nicht von $a_1, \ldots, a_{i-1}$ ab. Dann sollte gelten

$$p_i(a_i|a_1, \ldots, a_{i-1}) = P_i(\{a_i\})$$

mit einer von $a_1, \ldots, a_{i-1}$ unabhängigen Wahrscheinlichkeitsverteilung P_i auf Ω_i.
Für die Wahrscheinlichkeitsverteilung P auf $\Omega = \Omega_1 \times \ldots \times \Omega_n$ gilt dann

$$P(X_1 = a_1, \ldots, X_n = a_n) = \prod_{i=1}^{n} P_i(\{a_i\}). \tag{8.5}$$

Definition 8.18. *Die Wahrscheinlichkeitsverteilung P in (8.5) auf $\Omega = \Omega_1 \times \ldots \times \Omega_n$ heißt* ***Produkt*** *von $P_1, \ldots, P_n$ und wird mit $P_1 \otimes \ldots \otimes P_n$ notiert.*

Bemerkung. Sind die Mengen $\Omega_i,\ i = 1, \ldots, n$ endlich und ist P_i die Gleichverteilung auf Ω_i, dann ist $P_1 \otimes \ldots \otimes P_n$ offenbar die Gleichverteilung auf $\Omega_1 \times \ldots \times \Omega_n$.

Beispiel 8.19. Wir betrachten $\Omega_1 = \ldots = \Omega_n = \{0, 1\}$ mit $P_i(\{1\}) = p$, $i = 1, \ldots, n$. Sei $k = \sum_{i=1}^n a_i$ die Anzahl der Einsen. Dann ist

$$P(X_1 = a_1, \ldots, X_n = a_n) = \prod_{i=1}^n P_i(\{a_i\}) = p^k(1-p)^{n-k}.$$

Diese Verteilung wird als n-dimensionale **Bernoulli-Verteilung** bezeichnet.

Allgemeiner als (8.5) gilt

Satz 8.20. *Im Produktmodell gilt für beliebige Ereignisse $A_i \subset \Omega_i,\ i = 1, \ldots, n$,*

$$P(X_1 \in A_1, \ldots, X_n \in A_n) = \prod_{i=1}^n P(X_i \in A_i) = \prod_{i=1}^n P_i(A_i).$$

Beweis. Es gilt

$$\begin{aligned} P(X_1 \in A_1, \ldots, X_n \in A_n) = P(A_1 \times \ldots \times A_n) &= \sum_{a \in A_i \times \ldots \times A_n} P(\{a\}) \\ &\overset{(8.5)}{=} \sum_{a_1 \in A_1} \cdots \sum_{a_n \in A_n} \prod_{i=1}^n P_i(\{a_i\}) \\ &= \prod_{i=1}^n \sum_{a_i \in A_i} P_i(\{a_i\}) = \prod_{i=1}^n P_i(A_i). \end{aligned}$$

Insbesondere gilt

$$\begin{aligned} P(X_i \in A_i) &= P(X_1 \in \Omega_1, \ldots, X_{i-1} \in \Omega_{i-1}, X_i \in A_i, X_{i+1} \in \Omega_{i+1}, \ldots, X_n \in \Omega_n) \\ &= \prod_{\substack{j=1 \\ j \neq i}}^n P_j(\Omega_j) \cdot P_i(A_i) = P_i(A_i). \end{aligned}$$

□

Markov-Ketten

Wir betrachten $\Omega = S^{n+1} = \{(\omega_1, \ldots, \omega_{n+1}),\ \omega_i \in S\}$ mit abzählbarem S. Während bei Produktexperimenten der Ausgang des nächsten Experiments weder vom aktuellen noch von den vorhergehenden abhängt, beeinflusst bei den sog. *Markov-Ketten* das aktuelle Experiment den Ausgang des nächsten ("kein Gedächtnis"), d.h.

$$p_{k+1}(a_{k+1} | a_1, \ldots, a_k) = p_{k+1}(a_k, a_{k+1}), \tag{8.6}$$

wobei $p_{k+1} : S \times S \to [0, 1]$ folgende Bedingungen erfüllt

(i) $p_{k+1}(x, y) \geq 0$ für alle $x, y \in S$,

(ii) $\sum_{y\in S} p_{k+1}(x,y) = 1$ für alle $x \in S$,

d.h. $p_{k+1}(x,\cdot)$ ist für alle $x \in S$ eine Wahrscheinlichkeitsverteilung auf S.

Definition 8.21. *Eine Matrix $p_{k+1}(x,y)$, $x,y \in S$, mit (i) und (ii) heißt* ***stochastische Matrix*** *auf S.*

Für das Mehrstufenmodell folgt nach Satz 8.16 aus (8.6)

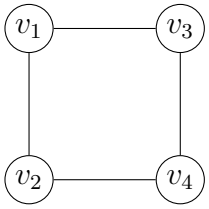
$$P(X_1 = a_1, \ldots, X_{n+1} = a_{n+1}) = \underbrace{p_1(a_1)}_{\text{Startverteilung}} \cdot \underbrace{p_2(a_1,a_2)\cdot \ldots \cdot p_n(a_{n-1},a_n)}_{\text{Übergangswahrscheinlichkeiten}} \tag{8.7}$$

für $a_1, \ldots, a_n \in S$. Die Folge der Zufallsvariablen $X_1, X_2, \ldots$ bezeichnet man als **Markov-Kette**. Sind die Übergangsmatrizen $p_{k+1}(x,y) = p(x,y)$ unabhängig von k, so nennt man die Markov-Kette **homogen**.

Beispiel 8.22.

(a) Für das Produktmodell gilt $p_{k+1}(x,y) = p_{k+1}(y)$ in (8.6).

(b) *Einfacher Random Walk*
Zum Zeitpunkt 1 befindet sich eine Person an der Ecke v_1 eines Häuserblocks.



Im den darauf folgenden Schritten geht die Person zu einer der beiden jeweils erreichbaren Ecken, je nachdem ob sie mit einer Münze Kopf oder Zahl wirft. Für jedes n sei X_n die Straßenecke zum Zeitpunkt n. Dann gilt

$$P(X_1 = v_1) = 1 \quad \text{und} \quad P(X_2 = v_2) = \frac{1}{2} = P(X_2 = v_3).$$

Die Übergangswahrscheinlichkeiten ergeben sich aus der Matrix

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

Satz 8.23. *Für alle $1 \le k < \ell \le n$ und $a_1, \ldots, a_\ell \in S$ mit $P(X_1 = a_1, \ldots, X_k = a_k) \neq 0$ gilt*

$$P(X_\ell = a_\ell | X_1 = a_1, \ldots, X_k = a_k) = P(X_\ell = a_\ell | X_k = a_k) = (p_{k+1}p_{k+2}\ldots p_\ell)(a_k, a_\ell),$$

wobei

$$(pq)(x,y) = \sum_{z\in S} p(x,z)q(z,y)$$

das Produkt der Matrizen p, q ist.

Beweis. Es gilt wegen (8.7)

$$\begin{aligned}
P(X_\ell = a_\ell | X_1 = a_1, \dots, X_k = a_k) &= \frac{P(X_1 = a_1, \dots, X_k = a_k, X_\ell = a_\ell)}{P(X_1 = a_1, \dots, X_k = a_k)} \\
&= \frac{\sum_{a_{k+1},\dots,a_{\ell-1} \in S} P(X_1 = a_1, \dots, X_\ell = a_\ell)}{P(X_1 = a_1, \dots, X_k = a_k)} \\
&\overset{(8.7)}{=} \frac{\sum_{a_{k+1},\dots,a_{\ell-1} \in S} p_1(a_1) \cdot p_2(a_1, a_2) \cdot \ldots \cdot p_\ell(a_{\ell-1}, a_\ell)}{p_1(a_1) \cdot p_2(a_1, a_2) \cdot \ldots \cdot p_k(a_{k-1}, a_k)} \\
&= \sum_{a_{k+1},\dots,a_{\ell-1} \in S} p_{k+1}(a_k, a_{k+1}) \cdot \ldots \cdot p_\ell(a_{\ell-1}, a_\ell) \\
&= (p_{k+1} \dots p_\ell)(a_k, a_\ell)
\end{aligned}$$

und

$$\begin{aligned}
P(X_\ell = a_\ell | X_k = a_k) &= \frac{P(X_k = a_k, X_\ell = a_\ell)}{P(X_k = a_k)} \\
&= \frac{\sum_{a_1,\dots,a_{k-1} \in S} \sum_{a_{k+1},\dots,a_{\ell-1} \in S} P(X_1 = a_1, \dots, X_\ell = a_\ell)}{\sum_{a_1,\dots,a_{k-1} \in S} P(X_1 = a_1, \dots, X_k = a_k)} \\
&\overset{(8.7)}{=} \frac{\sum_{a_1,\dots,a_{k-1} \in S} \sum_{a_{k+1},\dots,a_{\ell-1} \in S} p_1(a_1) \cdot p_2(a_1, a_2) \cdot \ldots \cdot p_\ell(a_{\ell-1}, a_\ell)}{\sum_{a_1,\dots,a_{k-1} \in S} p_1(a_1) \cdot p_2(a_1, a_2) \cdot \ldots \cdot p_k(a_{k-1}, a_k)} \\
&= \sum_{a_{k+1},\dots,a_{\ell-1} \in S} p_{k+1}(a_k, a_{k+1}) \cdot \ldots \cdot p_\ell(a_{\ell-1}, a_\ell) \\
&= (p_{k+1} \dots p_\ell)(a_k, a_\ell).
\end{aligned}$$

□

Bemerkung.

(a) Satz 8.23 zeigt die Markov-Eigenschaft. Die Weiterentwicklung hängt nur vom aktuellen Zustand a_k ab, aber nicht vom Verlauf $a_1, \dots, a_{k-1}$.

(b) Im Fall einer homogenen Markov-Kette hat man

$$P(X_\ell = a_\ell | X_k = a_k) = p^{\ell-k}(a_k, a_\ell).$$

Ist $S = \{S_1, \dots, S_m\}$ endlich, so definiere die Matrix $M \in \mathbb{R}^{m \times m}$ durch

$$M_{ij} = p(s_i, s_j), \quad i, j = 1, \dots, m.$$

Dann ist das Matrixprodukt $(pq)(x, y) = \sum_{z \in S} p(x, z) q(z, y)$ aus Satz 8.23 das übliche Matrixprodukt. Entsprechend gilt

$$P(X_\ell = a_\ell | X_k = a_k) = (M^{\ell-k})_{i_k i_\ell}, \tag{8.8}$$

wobei die Indizes i_k, i_ℓ durch $S_{i_k} = a_k$ und $S_{i_\ell} = a_\ell$ bestimmt sind.

(c) Das Produkt pq zweier stochastischer Matrizen p, q ist eine stochastische Matrix, weil

$$(pq)(x, y) = \sum_{z \in S} p(x, z) q(z, y) \geq 0$$

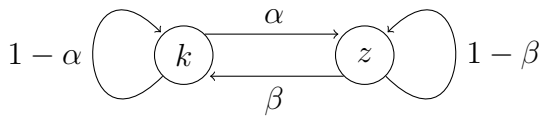
und

$$\sum_{y\in S}(pq)(x,y)=\sum_{y\in S}\sum_{z\in S}p(x,z)q(z,y)=\underbrace{\left(\sum_{z\in S}p(x,z)\right)}_{=1}\underbrace{\left(\sum_{y\in S}q(z,y)\right)}_{=1}=1.$$

Beispiel 8.24 (Abhängige Münzwürfe). Beim Werfen einer Münze sei die Wahrscheinlichkeit $(0<\alpha,\beta\leq 1)$

$1-\alpha$	für Kopf und	α	für Zahl, falls im letzten Wurf Kopf,
β	für Kopf und	$1-\beta$	für Zahl, falls im letzten Wurf Zahl

geworfen wurde.



Nach (8.8) ist

$$P(X_{n+1}=s_j|X_1=s_i)=(M^n)_{ij}$$

mit

$$M=\begin{bmatrix}1-\alpha & \alpha\\ \beta & 1-\beta\end{bmatrix}.$$

Weil M^{n-1} ebenfalls eine stochastische Matrix ist, gilt

$$\begin{aligned}(M^n)_{11}&=(M^{n-1})_{11}M_{11}+(M^{n-1})_{12}M_{21}\\&=(M^{n-1})_{11}(1-\alpha)+(1-(M^{n-1})_{11})\beta\\&=(1-\alpha-\beta)(M^{n-1})_{11}+\beta.\end{aligned}$$

Induktiv erhält man

$$(M^n)_{11}=\frac{\beta}{\alpha+\beta}+\frac{\alpha}{\alpha+\beta}(1-\alpha-\beta)^n$$

und analog

$$(M^n)_{22}=\frac{\alpha}{\alpha+\beta}+\frac{\beta}{\alpha+\beta}(1-\alpha-\beta)^n.$$

Hieraus folgt

$$M^n=\underbrace{\frac{1}{\alpha+\beta}\begin{pmatrix}\beta & \alpha\\ \beta & \alpha\end{pmatrix}}_{\text{gleiche Zeilen}}+\underbrace{\frac{(1-\alpha-\beta)^n}{\alpha+\beta}\begin{pmatrix}\alpha & -\alpha\\ -\beta & \beta\end{pmatrix}}_{\to 0 \text{ für } n\to\infty}.$$

Also gilt $P(X_{n+1}=a|X_1=K)\approx P(X_{n+1}=a|X_1=Z)$ für große n. Die Kette vergisst ihren Startwert exponentiell schnell.

Random Walk auf $\mathbb{Z}$

Definition 8.25. *Eine beliebige Kollektion $X_i : \Omega \to S_i,\ i \in I$, von diskreten Zufallsvariablen heißt* ***unabhängig****, falls die Ereignisse $\{X_i = a_i\}$, $i \in I$, für alle $a_i \in S_i$ unabhängig sind.*

Seien $X_1, X_2, X_3, \ldots$ unabhängige Zufallsvariablen auf $(\Omega, \mathcal{A}, P)$ mit

$$P(X_i = +1) = p, \quad P(X_i = -1) = 1 - p$$

mit $p \in (0, 1)$. Sei $a \in \mathbb{Z}$ gegeben. Wir betrachten die durch

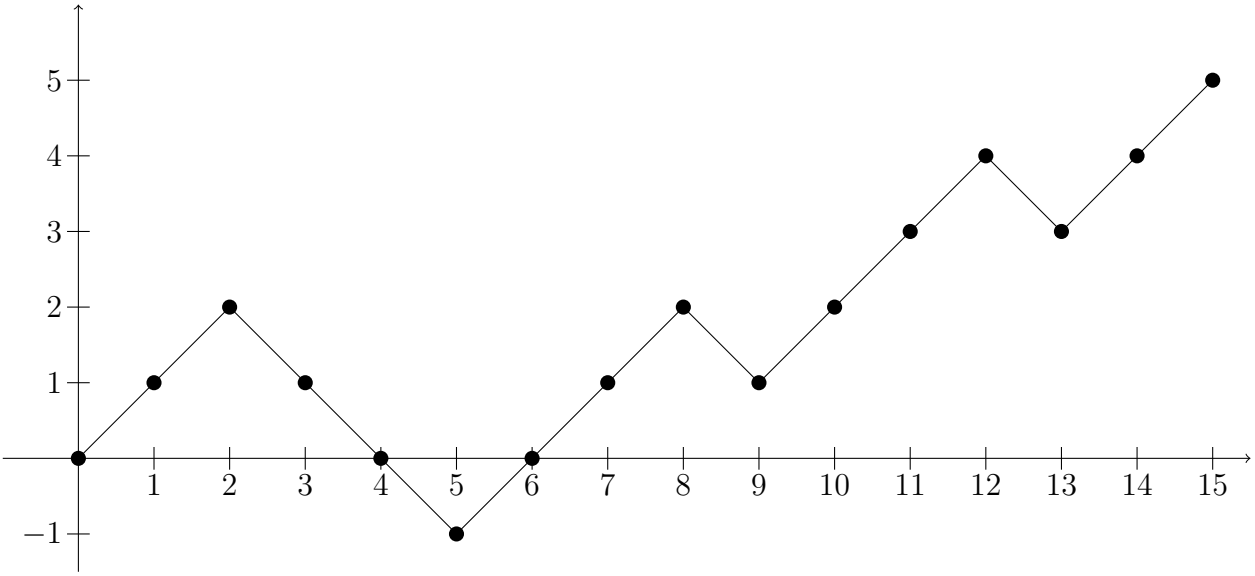
$$S_0 = a, \quad S_{n+1} = S_n + X_{n+1}, \quad n = 0, 1, 2, \ldots,$$

definierte zufällige Bewegung (engl. random walk) auf $\mathbb{Z}$.
Für die Position zum Zeitpunkt n gilt dann

$$S_n = a + X_1 + X_2 + \ldots + X_n.$$

Random Walks werden z.B. als einfache Modelle von Aktienkursen verwendet.

Beispiel 8.26.



Lemma 8.27 (Verteilung von S_n). *Für $k \in \mathbb{Z}$ gilt*

$$P(S_n = a + k) = \begin{cases} 0, & \text{falls } n + k \text{ ungerade oder } |k| > n, \\ \binom{n}{\frac{n+k}{2}} p^{\frac{n+k}{2}} (1-p)^{\frac{n-k}{2}}, & \text{sonst.} \end{cases}$$

Beweis. Der Fall $|k| > n$ kann nicht eintreten. Mit n_+ und n_- bezeichnen wir die Anzahl der Zufallsvariablen aus $X_1, \ldots, X_n$, die den Wert $+1$ bzw. -1 annehmen. Dann gilt $n_+ + n_- = n$ und $n_+ - n_- = k$. Dieses System hat genau dann die Lösung $n_+ = \frac{n+k}{2}$ und $n_- = \frac{n-k}{2}$, wenn $n + k$ und somit auch $n - k$ geradzahlig sind. □

Beispiel 8.28. Wir wollen die Wahrscheinlichkeit berechnen, mit der S_n zum Startwert a zurückkehrt. Zunächst sehen wir $P(S_{2n+1} = a) = 0$. Für geradzahlige Indizes verwenden wir die **Stirlingsche Formel**

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \quad \text{für } n \to \infty,$$

wobei zwei Folgen $\{a_n\}$ und $\{b_n\}$ **asymptotisch äquivalent** heißen ($a_n \sim b_n$), falls

$$\lim_{n\to\infty} \frac{a_n}{b_n} = 1.$$

Hiermit folgt

$$\begin{aligned} P(S_{2n} = a) &= \binom{2n}{n} p^n (1-p)^n = \frac{(2n)!}{(n!)^2} p^n (1-p)^n \\ &\sim \frac{\sqrt{4\pi n}}{2\pi n} \frac{\left(\frac{2n}{e}\right)^{2n}}{\left(\frac{n}{e}\right)^{2n}} p^n (1-p)^n \\ &= \frac{1}{\sqrt{\pi n}} (4p(1-p))^n \quad \text{für } n \to \infty. \end{aligned}$$

Im Fall $p \neq \frac{1}{2}$ ist $4p(1-p) < 1$ und $P(S_{2n} = a)$ konvergieren exponentiell gegen 0. Ist $p = \frac{1}{2}$, so konverigert $P(S_{2n} = a) \sim \frac{1}{\sqrt{\pi n}}$ nur langsam gegen 0.

9 Konvergenzsätze und Monte Carlo-Methoden

Sei μ eine Wahrscheinlichkeitsverteilung auf einer abzählbaren Menge S und $f : S \to \mathbb{R}$ eine Zufallsvariable. Nehmen wir an, die Mächtigkeit von S wäre so groß, dass es zu aufwändig ist, den Erwartungswert

$$E_\mu(f) = \sum_{a \in S} f(a)\mu(\{a\})$$

direkt auszurechnen. Um die Problematik zu umgehen, werden so genannte *Monte Carlo-Verfahren* verwendet. Dabei approximiert man $E_\mu(f)$ durch

$$\eta_n(\omega) := \frac{1}{n} \sum_{i=1}^{n} f(X_i(\omega))$$

mit einer großen Anzahl unabhängiger Stichproben $X_1, \ldots, X_n$ von μ. In diesem Kapitel soll der Approximationsfehler $\eta_n - E_\mu(f)$ und somit die Konvergenz $\eta_n \to E_\mu(f)$ für $n \to \infty$ untersucht werden. Nach Satz 7.30 und Satz 7.31 gilt wegen $P(X_i = a) = \mu(\{a\})$

$$E(\eta_n) = \frac{1}{n} \sum_{i=1}^{n} E(f(X_i)) = \frac{1}{n} \sum_{i=1}^{n} \sum_{a \in S} f(a)\mu(\{a\}) = E_\mu(f),$$

d.h. η_n ist ein **erwartungstreuer Schätzer** für $E_\mu(f)$. Für den mittleren quadratischen Fehler gilt daher

$$E(|\eta_n - E_\mu(f)|^2) = E(|\eta_n - E(\eta_n)|^2).$$

Den letzten Ausdruck werden wir im Folgenden untersuchen.

9.1 Varianz und Kovarianz

Sei wieder $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum und $X : \Omega \to S$ eine Zufallsvariable auf $(\Omega, \mathcal{A}, P)$, so dass $E(|X|)$ endlich ist.

Definition 9.1. *Die Größe*

$$\mathrm{Var}(X) := E([X - E(X)]^2) \quad \in \quad [0, \infty]$$

heißt ***Varianz*** *von X. Als* ***Standardabweichung*** *von X bezeichnet man*

$$\sigma(X) := \sqrt{\mathrm{Var}(X)}.$$

Die Varianz von X beschreibt die mittlere quadratische Abweichung der Zufallsvariablen X vom Erwartungswert $E(X)$. Wegen

$$\mathrm{Var}(X) = \sum_{a \in S} (a - E(X))^2 P(X = a)$$

gilt $\mathrm{Var}(X) = 0$ genau dann, wenn $P(X = E(X)) = 1$.

Satz 9.2 (Rechenregeln). *Es gilt*

(i) $\mathrm{Var}(\alpha X + \beta) = \alpha^2 \mathrm{Var}(X)$ *für alle* $\alpha, \beta \in \mathbb{R}$,

(ii) $\mathrm{Var}(X) = E(X^2) - (E(X))^2$.

Insbesondere ist $\mathrm{Var}(X) < \infty$ *genau dann, wenn* $E(X^2)$ *endlich ist.*

Beweis.

(i) Aus der Linearität des Erwartungswertes folgt

$$\begin{aligned}\mathrm{Var}(\alpha X + \beta) &= E([\alpha X + \beta - E(\alpha X + \beta)]^2) = E([\alpha X + \beta - \alpha E(X) - \beta]^2) \\ &= E([\alpha X - \alpha E(X)]^2) = \alpha^2 E([X - E(X)]^2) = \alpha^2 \mathrm{Var}(X).\end{aligned}$$

(ii)

$$\begin{aligned}\mathrm{Var}(X) &= E(X^2 - 2XE(X) + (E(X))^2) = E(X^2) - 2E(X)E(X) + (E(X))^2 \\ &= E(X^2) - (E(X))^2\end{aligned}$$

□

Beispiel 9.3 (Varianz der geometrischen Verteilung).
Sei X geometrisch verteilt mit Paramter $p \in (0, 1]$. In der Bemerkung zu Definition 8.15 haben wir gesehen, dass $E(X) = \frac{1-p}{p}$. Wegen

$$\begin{aligned}E(X(X+1)) &= \sum_{k=1}^{\infty} k(k+1)p(1-p)^k = p(1-p)\sum_{k=0}^{\infty}(k-1)k(1-p)^{k-2} \\ &= p(1-p)\frac{d^2}{dp^2}\sum_{k=0}^{\infty}(1-p)^k = p(1-p)\frac{d^2}{dp^2}\frac{1}{p} = \frac{2(1-p)}{p^2}\end{aligned}$$

und Satz 9.2 (ii) ist

$$\begin{aligned}\mathrm{Var}(X) &= E(X^2) - (E(X))^2 = E(X(X+1)) - E(X) - (E(X))^2 \\ &= \frac{2(1-p)}{p^2} - \frac{1-p}{p} - \frac{(1-p)^2}{p^2} = \frac{1-p}{p^2}.\end{aligned}$$

Beispiel 9.4 (Varianz der Poissonverteilung).
In Beispiel 7.27 haben wir den Erwartungswert der Poissonverteilung X berechnet. Wegen

$$E(X(X-1)) = \sum_{k=0}^{\infty} k(k-1)\frac{\lambda^k}{k!}e^{-\lambda} = \lambda^2 e^{-\lambda}\sum_{k=2}^{\infty}\frac{\lambda^{k-2}}{(k-2)!} = \lambda^2 e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} = \lambda^2$$

folgt mit $E(X) = \lambda$, dass $\mathrm{Var}(X) = E(X(X-1)) + E(X) - (E(X))^2 = \lambda$.

Beispiel 9.5 (Varianz der hypergeometrischen Verteilung).
Nach Beispiel 7.28 ist der Erwartungswert der hypergeometrischen Verteilung $n \cdot \frac{r}{m}$. Auf ähnliche Weise erhält man, dass $E(X(X-1)) = n(n-1)\frac{r(r-1)}{m(m-1)}$. Hieraus folgt

$$\begin{aligned}\mathrm{Var}(X) &= E(X^2) - (E(X))^2 = E(X(X-1)) + E(X) - (E(X))^2 \\ &= n\frac{r}{m}\left(1 - \frac{r}{m}\right)\frac{r-n}{r-1}.\end{aligned}$$

Im Folgenden betrachten wir Zufallsvariablen X mit endlichen $E(X^2)$, d.h. Elemente der Menge

$$\mathcal{L}^2(\Omega, \mathcal{A}, P) := \{X : \Omega \to \mathbb{R} : X \text{ ist diskrete Zufallsvariable mit } E(X^2) < \infty\}.$$

Lemma 9.6.

(i) Für $X, Y \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$ gilt $E(|XY|) \le E^{1/2}(X^2)\, E^{1/2}(Y^2) < \infty$.

(ii) Durch $(X, Y)_{\mathcal{L}^2} := E(XY)$ ist eine positiv semidefinite symmetrische Bilinearform (so genanntes Skalarprodukt) auf dem Vektorraum $\mathcal{L}^2(\Omega, \mathcal{A}, P)$ definiert.

Beweis.

(i) nach der Cauchy-Schwarz-Ungleichung

$$\sum_{i\in I} a_i b_i \le \left(\sum_{i\in I} |a_i|^2\right)^{1/2} \left(\sum_{i\in I} |b_i|^2\right)^{1/2}$$

mit abzählbarer Indexmenge I gilt

$$\begin{aligned}E(|XY|) &= \sum_{\substack{a\in X(\Omega)\\ b\in Y(\Omega)}} |ab| P(X=a, Y=b) \\ &= \sum_{\substack{a\in X(\Omega)\\ b\in Y(\Omega)}} |a| P^{1/2}(X=a, Y=b)\, |b| P^{1/2}(X=a, Y=b) \\ &\le \left(\sum_{\substack{a\in X(\Omega)\\ b\in Y(\Omega)}} a^2 P(X=a, Y=b)\right)^{1/2} \left(\sum_{\substack{a\in X(\Omega)\\ b\in Y(\Omega)}} b^2 P(X=a, Y=b)\right)^{1/2} \\ &= \left(\sum_{a\in X(\Omega)} a^2 P(X=a)\right)^{1/2} \left(\sum_{b\in Y(\Omega)} b^2 P(Y=b)\right)^{1/2} \\ &= E^{1/2}(X^2)\, E^{1/2}(Y^2).\end{aligned}$$

(ii) Seien $X, Y \in \mathcal{L}^2$ und $\alpha, \beta \in \mathbb{R}$. Dann ist $\alpha X + \beta Y$ eine diskrete Zufallsvariable, und es gilt wegen $2\alpha\beta XY \le \alpha^2 X^2 + \beta^2 Y^2$

$$\begin{aligned}E([\alpha X + \beta Y]^2) &= E(\alpha^2 X^2 + 2\alpha\beta XY + \beta^2 Y^2) \\ &= \alpha^2 E(X^2) + E(2\alpha\beta XY) + \beta^2 E(Y^2) \\ &\overset{\text{Satz 7.32}}{\le} \alpha^2 E(X^2) + \alpha^2 E(X^2) + \beta^2 E(Y^2) + \beta^2 E(Y^2) \\ &= 2\alpha^2 E(X^2) + 2\beta^2 E(Y^2) < \infty\end{aligned}$$

Daher ist $\alpha X + \beta Y \in \mathcal{L}^2$, und $\mathcal{L}^2$ ist ein linearer Raum.

Ferner ist $(X,Y)_{\mathcal{L}^2} = E(XY)$ bilinear und wegen

$$(X,X)_{\mathcal{L}^2} = E(X^2) \geq 0 \quad \text{für alle } X \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$$

positiv semidefinit. □

Bemerkung. Für $X \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$ folgt aus Lemma 9.6 (i)

$$E(|X|) \leq E^{1/2}(X^2)\, E^{1/2}(1^2) = E^{1/2}(X^2) < \infty.$$

Außerdem folgt die Cauchy-Schwarzsche Ungleichung auf $\mathcal{L}^2(\Omega, \mathcal{A}, P)$

$$E(XY) \leq E(|XY|) \leq E^{1/2}(X^2)\, E^{1/2}(Y^2) \quad \text{für alle } X, Y \in \mathcal{L}^2(\Omega, \mathcal{A}, P).$$

Definition 9.7. *Seien $X, Y \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$. Der Ausdruck*

$$\mathrm{Cov}(X,Y) := E([X - E(X)][Y - E(Y)]) = E(XY) - E(X)E(Y)$$

*wird als **Kovarianz** von X und Y bezeichnet. X und Y heißen **unkorreliert**, falls $\mathrm{Cov}(X,Y) = 0$, d.h. $E(XY) = E(X) \cdot E(Y)$. Gilt $\sigma(X), \sigma(Y) \neq 0$, so heißt*

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma(X)\sigma(Y)}$$

***Korrelationskoeffizient** von X und Y.*

Bemerkung. Die Abbildung $\mathrm{Cov} : \mathcal{L}^2 \times \mathcal{L}^2 \to \mathbb{R}$ ist eine symmetrische Bilinearform mit $\mathrm{Cov}(X,X) = \mathrm{Var}(X) \geq 0$ für alle $X \in \mathcal{L}^2$.

Der folgende Satz beschreibt den Zusammenhang von Unabhängigkeit und Unkorreliertheit.

Satz 9.8. *Seien $X : \Omega \to S$, $Y : \Omega \to T$ diskrete Zufallsvariablen auf $(\Omega, \mathcal{A}, P)$. X und Y sind genau dann unabhängig, falls $f(X)$ und $g(Y)$ für alle Funktionen $f : S \to \mathbb{R}$, $g : T \to \mathbb{R}$ mit $f(X), g(Y) \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$ unkorreliert sind.*

Beweis. Seien X und Y unabhängig. Dann gilt

$$\begin{aligned} E(f(X)g(Y)) &= \sum_{a \in S} \sum_{b \in T} f(a)g(b) P(X = a, Y = b) \\ &= \left(\sum_{a \in S} f(a) P(X = a) \right) \left(\sum_{b \in T} g(b) P(Y = b) \right) \\ &= E(f(X))\, E(g(Y)) \end{aligned}$$

und somit $\mathrm{Cov}(f(X), g(Y)) = 0$. Die Umkehrung der Aussage folgt aus

$$P(X = a, Y = b) = E(\chi_a(X)\chi_b(Y)) = E(\chi_a(X))\, E(\chi_b(Y)) = P(X = a)\, P(Y = b).$$

□

Beispiel 9.9. Sei $X = +1, 0, -1$ jeweils mit Wahrscheinlichkeit $\frac{1}{3}$. Dann sind X und $Y = X^2$ nicht unabhängig aber unkorreliert:

$$E(XY) = 0 = E(X)\,E(Y).$$

Dies steht nicht im Widerspruch zu Satz 9.8. Für $f(X) = X^2$ zeigt sich die Korreliertheit.

Satz 9.10 (Varianz von Summen). *Für $X_1, \ldots, X_n \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$ gilt*

$$\mathrm{Var}(X_1 + \ldots + X_n) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2 \sum_{\substack{i,j=1 \\ i<j}}^{n} \mathrm{Cov}(X_i, X_j).$$

Beweis. Wegen der Bilinearität der Kovarianz gilt

$$\begin{aligned}\mathrm{Var}(X_1 + \ldots + X_n) &= \mathrm{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\right) = \sum_{i,j=1}^{n} \mathrm{Cov}(X_i, X_j) \\ &= \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2 \sum_{\substack{i,j=1 \\ i<j}}^{n} \mathrm{Cov}(X_i, X_j).\end{aligned}$$

□

Beispiel 9.11. Wir wollen die Varianz der Binomialverteilung berechnen (siehe Bemerkung nach Definition 8.15). Sei

$$S_n = \sum_{i=1}^{n} X_i \quad \text{mit} \quad X_i = \begin{cases} 1 & \text{mit Wahrscheinlichkeit } p, \\ 0 & \text{mit Wahrscheinlichkeit } 1-p, \end{cases}$$

und unabhängigen Zufallsvariablen X_i. Nach Satz 9.8 und Satz 9.10 gilt

$$\mathrm{Var}(S_n) = \sum_{i=1}^{n} \mathrm{Var}(X_i) = n\,p(1-p),$$

weil

$$E(X_i^2) = E(X_i) = p$$

und

$$\mathrm{Var}(X_i) = E(X_i^2) - (E(X_i))^2 = p - p^2 = p(1-p).$$

Beispiel 9.12. Wir kehren zum Beginn des Kapitels zurück. Dort haben wir

$$\eta_n(\omega) = \frac{1}{n} \sum_{i=1}^{n} f(X_i(\omega))$$

als Approximation für den Erwartungswert $E_\mu(f)$ eingeführt. Sei $E_\mu(f^2) = \sum_{a \in S} f^2(a)\mu(\{a\})$ endlich. Nach Satz 9.8 gilt wegen der Unabhängigkeit der X_i, dass die $f(X_i)$ paarweise unkorreliert sind. Also folgt nach Satz 9.10 und wegen $P(X_i = a) = \mu(\{a\})$

$$\begin{aligned}\mathrm{Var}(\eta_n) &= \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}(f(X_i)) = \frac{1}{n^2} \sum_{i=1}^{n} (f(a) - E_\mu(f))^2 P(X_i = a) \\ &= \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}_\mu(f) = \frac{1}{n} \mathrm{Var}_\mu(f) < \infty.\end{aligned}$$

Für die durch das Skalarprodukt $(\cdot,\cdot)_{\mathcal{L}^2}$ (siehe Lemma 9.6) induzierte Norm $\|X\|_{\mathcal{L}^2} := \sqrt{(X,X)_{\mathcal{L}^2}}$ gilt somit

$$\|\eta_n - E_\mu(f)\|_{\mathcal{L}^2} = E^{1/2}(|\eta_n - E(\eta_n)|^2) = \sqrt{\mathrm{Var}(\eta_n)} = \frac{1}{\sqrt{n}}\sqrt{\mathrm{Var}_\mu(f)}.$$

Daher konvergiert η_n in dieser Norm gegen $E_\mu(f)$. Die Konvergenz ist im Vergleich zu deterministischen Verfahren (mehr dazu später) allerdings recht langsam.

Im folgenden Abschnitt wollen wir die stochastische Konvergenz untersuchen.

9.2 Schwaches Gesetz der großen Zahlen

Wie bisher sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum, $X_1, \ldots, X_n \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$ diskrete Zufallsvariablen und $S_n := X_1 + \ldots + X_n$. Ziel dieses Abschnittes ist es zu zeigen, dass für das arithmetische Mittel S_n/n der X_i in einem stochastischen Sinne gilt

$$\frac{S_n}{n} \approx \frac{E(S_n)}{n} \quad \text{für große } n,$$

d.h. der Zufall mittelt sich weg.

Definition 9.13. *Eine Folge von Zufallsvariablen $\{X_i\}$, $X_i : \Omega \to \mathbb{R}$, **konvergiert stochastisch** gegen $x \in \mathbb{R}$, falls für alle $\varepsilon > 0$*

$$P(|X_n - x| \geq \varepsilon) \xrightarrow{n\to\infty} 0.$$

Für den Beweis der Aussage von eben benötigen wir

Lemma 9.14 (Tschebyscheffsche Ungleichung). *Für alle $X \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$ und alle $\varepsilon > 0$ gilt*

$$P(|X - E(X)| \geq \varepsilon) \leq \frac{1}{\varepsilon^2}\mathrm{Var}(X).$$

Beweis. Für Elemente von $A := \{\omega \in \Omega : |X(\omega) - E(X)| \geq \varepsilon\}$ hat man offenbar $\frac{1}{\varepsilon^2}(X(\omega) - E(X))^2 \geq 1$. Also ist $\chi_A \leq \frac{1}{\varepsilon^2}(X - E(X))^2$ in Ω, und es folgt

$$P(A) = E(\chi_A) \leq E(\frac{1}{\varepsilon^2}[X - E(X)]^2) = \frac{1}{\varepsilon^2}E([X - E(X)]^2) = \frac{1}{\varepsilon^2}\mathrm{Var}(X).$$

□

Satz 9.15 (Schwaches Gesetz der großen Zahlen). *Seien $X_1, \ldots, X_n$ paarweise unkorrelierte Zufallsvariablen und $M_n := \max_{i=1,\ldots,n} \mathrm{Var}(X_i)$. Dann gilt für alle $\varepsilon > 0$*

$$P\left(\left|\frac{S_n}{n} - \frac{E(S_n)}{n}\right| \geq \varepsilon\right) \leq \frac{M_n}{\varepsilon^2 n}.$$

Ist die Folge $\{M_n/n\}$ eine Nullfolge und gilt $E(X_i) = S \in \mathbb{R}$ für alle i, so konvergiert $\{S_n/n\}$ stochastisch gegen S.

Beweis. Nach der Tschebyscheffschen Ungleichung und Satz 9.10 gilt

$$P\left(\left|\frac{S_n}{n} - \frac{E(S_n)}{n}\right| \geq \varepsilon\right) \leq \frac{1}{\varepsilon^2}\mathrm{Var}\left(\frac{S_n}{n}\right) = \frac{1}{n^2\varepsilon^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right)$$
$$= \frac{1}{n^2\varepsilon^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) \leq \frac{nM_n}{n^2\varepsilon^2} = \frac{M_n}{\varepsilon^2 n}.$$

□

Bemerkung. Die Aussage, dass $\{S_n/n\}$ stochastisch gegen S konvergiert, falls $E(X_i) = S$, ist im Allgemeinen falsch. Für $X_1 = \ldots = X_n$ mittelt sich der Zufall nicht weg, weil $S_n/n = X_1$.

Beispiel 9.16 (Monte Carlo-Verfahren für die Wahrscheinlichkeit). Sei $A \in \mathcal{A}$. Wir definieren die Zufallsvariablen

$$X_i = \begin{cases} 1, & A \text{ tritt im } i\text{-ten Versuch ein,} \\ 0, & A \text{ tritt im } i\text{-ten Versuch nicht ein.} \end{cases}$$

Dann gilt

$$E(X_i) = P(A), \quad \mathrm{Var}(X_i) = E(X_i^2) - (E(X_i))^2 = P(A)(1 - P(A)) \leq \frac{1}{4}.$$

Das schwache Gesetz der großen Zahlen zeigt die Konvergenz der **relativen Häufigkeit**

$$H_n(A) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

gegen die Wahrscheinlichkeit $P(A)$

$$P(|H_n(A) - P(A)| \geq \varepsilon) \leq \frac{1}{4n\varepsilon^2} \xrightarrow{n\to\infty} 0.$$

Mit Hilfe von Satz 9.15 kann auch die Anzahl der Stichproben für ein Konfidenzintervall bestimmt werden. Die Anzahl der Stichproben, die für einen relativen Fehler mit 95%iger Wahrscheinlichkeit unterhalb von 10% benötigt werden, ergibt sich aus

$$P(|H_n(A) - P(A)| \geq 0.1 \cdot P(A)) \leq \frac{P(A)(1 - P(A))}{n(0.1 \cdot P(A))^2} = \frac{100(1 - P(A))}{nP(A)} \leq 0.05,$$

falls

$$n \geq \frac{2000 \cdot (1 - P(A))}{P(A)}.$$

Für $P(A) = 10^{-5}$ ist daher $n \approx 2 \cdot 10^8$.

Beispiel 9.17. Wir wollen die Wahrscheinlichkeit $P(|\eta_n - E_\mu(f)| \geq \varepsilon)$ für den Monte Carlo-Schätzer aus Beispiel 9.12 abschätzen. Nach Satz 9.15 gilt

$$P(|\eta_n - E_\mu(f)| \geq \varepsilon) \leq \frac{1}{n\varepsilon^2}\mathrm{Var}_\mu(f).$$

Varianzreduktion durch Importance Sampling

In Beispiel 9.16 sind im Fall $P(A) = 10^{-5}$ der überwiegende Teil der rund $2 \cdot 10^8$ Summanden von H_n Null. Um dies zu verbessern, kann man ein alternatives Schätzverfahren einführen, das die "Wichtigkeit" der Stichproben berücksichtigt. Sei dazu ν eine weitere Wahrscheinlichkeitsverteilung auf S mit $\nu(\{a\}) > 0$ für alle $a \in S$. Dann lässt sich $E_\mu(f)$ auch als Erwartungswert bzgl. ν ausdrücken:

$$E_\mu(f) = \sum_{a \in S} f(a)\mu(\{a\}) = \sum_{a \in S} f(a)\rho(a)\nu(\{a\}) = E_\nu(f\rho),$$

wobei

$$\rho(a) := \frac{\mu(\{a\})}{\nu(\{a\})}.$$

Entsprechend erhalten wir einen alternativen Monte Carlo-Schätzer für $E_\mu(f)$

$$\tilde{\eta}_n = \frac{1}{n} \sum_{i=1}^{n} f(Y_i)\rho(Y_i)$$

mit unabhängigen Zufallsvariablen Y_i zur Verteilung ν. Dann ist auch $\tilde{\eta}_n$ erwartungstreu, weil

$$E_\nu(\tilde{\eta}_n) = E_\nu(f\rho) = E_\mu(f),$$

und für die Varianz von $\tilde{\eta}_n$ gilt

$$\mathrm{Var}(\tilde{\eta}_n) = \frac{1}{n}\mathrm{Var}_\nu(f\rho) = \frac{1}{n}\left(\sum_{a \in S} f^2(a)\rho^2(a)\nu\{a\} - (E_\mu(f))^2\right).$$

Durch geeignete Wahl von ν kann $\mathrm{Var}(\tilde{\eta}_n)$ gegenüber $\mathrm{Var}(\eta_n)$ deutlich reduziert werden. Als Faustregel gilt: $\nu(\{a\})$ sollte groß sein, wenn $|f(a)|$ groß ist.

9.3 Gleichgewichte von Markov-Ketten

Im Abschnitt 8.3 haben wir Markov-Ketten kennengelernt. In diesem Abschnitt wollen wir den Grenzwert homogener Markov-Ketten X_n untersuchen.

Lemma 9.18.

(i) Die Verteilung einer Markov-Kette mit Startverteilung ν und Übergangsmatrix p zum Zeitpunkt $n+1$ ist νp^n. Hierbei ist $(\nu p)(y) := \sum_{x \in S} \nu(x)p(x,y)$.

(ii) Gilt $\nu p = \nu$, so ist ν die Verteilung von X_n für alle $n \in \mathbb{N}$.

Beweis.

(i) Aus Satz 8.23 und der darauffolgenden Bemerkung erhält man

$$P(X_{n+1} = b | X_1 = a) = p^n(a,b)$$

für alle $n \in \mathbb{N}$ und $a, b \in S$ mit $P(X_1 = a) \neq 0$. Also folgt nach Satz 8.5

$$\begin{aligned} P(X_{n+1} = b) &= \sum_{\substack{a \in S \\ P(X_1 = a) \neq 0}} P(X_{n+1} = b | X_1 = a) \cdot P(X_1 = a) \\ &= \sum_{\substack{a \in S \\ \nu(a) \neq 0}} p^n(a, b)\nu(a) = (\nu p^n)(b). \end{aligned}$$

(ii) Aus $\nu p = \nu$ folgt $\nu p^n = \nu p^{n-1} = \ldots = \nu$ für alle $n \in \mathbb{N}$.

□

Definition 9.19. *Eine Wahrscheinlichkeitsverteilung μ auf S heißt* ***Gleichgewichtsverteilung*** *(oder* ***stationäre Verteilung****) der Übergangsmatrix p, falls $\mu p = \mu$, d.h. falls*

$$\sum_{x \in S} \mu(x)p(x, y) = \mu(y) \quad \text{für alle } y \in S.$$

Eine Wahrscheinlichkeitsverteilung μ auf S heißt ***reversibel*** *bzgl. p, falls*

$$\mu(x)p(x, y) = \mu(y)p(y, x) \quad \text{für alle } x, y \in S.$$

Bemerkung.

(a) Bei einer Startverteilung μ gilt

$$\mu(x)p(x, y) = P(X_1 = x, X_2 = y)$$

Interpretiert man $P(X_1 = x, X_2 = y)$ als Fluss von x nach y, so bedeutet anschaulich die

Reversibilität

$$\begin{aligned} \mu(x)p(x, y) &= \mu(y)p(y, x) \\ \text{Fluss von } x \text{ nach } y &= \text{Fluss von } y \text{ nach } x, \end{aligned}$$

Gleichgewichtsbedingung $\mu p = \mu$

$$\begin{aligned} \sum_{x \in S} \mu(x)p(x, y) &= \sum_{x \in S} \mu(y)p(y, x) \\ \text{Gesamter Fluss nach } y &= \text{Gesamter Fluss von } y. \end{aligned}$$

(b) Algebraisch bedeutet die Gleichgewichtsbedingung, dass μ ein linker Eigenvektor von p zum Eigenwert 1 ist.

Satz 9.20. *Ist μ reversibel bzgl. p, dann ist μ eine Gleichgewichtsverteilung von p.*

Beweis. Aus der Reversibilität folgt

$$\sum_{x \in S} \mu(x)p(x, y) = \sum_{x \in S} \mu(y)p(y, x) = \mu(y),$$

weil p eine stochastische Matrix ist. □

Beispiel 9.21. Wir betrachten nochmals Beispiel 8.24 mit $\alpha, \beta \in [0,1]$ und

$$M = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix}.$$

Dann ist die Gleichgewichtsbedingung $\mu p = \mu$, $\mu = \begin{bmatrix}\mu_1, \mu_2\end{bmatrix}^T$, äquivalent zu

$$\begin{aligned} \mu_1 &= \mu_1(1-\alpha) + \mu_2\beta, \\ \mu_2 &= \mu_1\alpha + (1-\beta)\mu_2 \end{aligned} \quad \Leftrightarrow \quad \begin{bmatrix}\mu_1 \\ \mu_2\end{bmatrix} = M^T \begin{bmatrix}\mu_1 \\ \mu_2\end{bmatrix}.$$

Da μ eine Wahrscheinlichkeitsverteilung ist, d.h. $\mu_1 + \mu_2 = 1$, sind beide Gleichungen äquivalent zu

$$\beta(1-\mu_1) = \alpha\mu_1,$$

welche äquivalent zur Reversibilität von μ ist. Im Fall $\alpha + \beta > 0$ ist $\mu = [\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta}]^T$ die eindeutige Gleichgewichtsverteilung. Ist $\alpha = \beta = 0$, so ist jede Wahrscheinlichkeitsverteilung μ eine Gleichgewichtsverteilung.

Konvergenz ins Gleichgewicht

In diesem Abschnitt zeigen wir die Konvergenz von νp^n gegen eine Gleichgewichtsverteilung. Sei $S = \{s_1, \ldots, s_m\}$ eine endliche Menge und

$$W(S) := \left\{ \mu = (\mu(s_1), \ldots, \mu(s_m)) : \mu(s_i) \geq 0, \sum_{i=1}^m \mu(s_i) = 1 \right\} \subset \mathbb{R}^m$$

die Menge aller Wahrscheinlichkeitsverteilungen auf S. Auf $W(S)$ führen wir die **Variationsdistanz** zweier Wahrscheinlichkeitsverteilungen $\mu, \nu \in W(S)$

$$d(\mu, \nu) = \frac{1}{2}\|\mu - \nu\|_1 = \frac{1}{2}\sum_{i=1}^m |\mu(s_i) - \nu(s_i)|$$

ein.

Bemerkung. Für alle $\mu, \nu \in W(S)$ gilt

$$d(\mu, \nu) \leq \frac{1}{2}\sum_{i=1}^m (\mu(s_i) + \nu(s_i)) = 1.$$

Wir betrachten im Folgenden eine stochastische Matrix p auf $S \times S$ mit Gleichgewicht μ. Die Verteilung einer Markov-Kette mit Startverteilung ν und Übergangsmatrix p zur Zeit n ist nach Lemma 9.18 νp^n. Für den folgenden Konvergenzbeweis von $\{\nu p^n\}$ ins Gleichgewicht nehmen wir zunächst die folgende Minorisierungsbedinung an. Es gibt $r \in \mathbb{N}$ und ein $\delta \in [0,1]$, so dass

$$p^r(x,y) \geq \delta\mu(y) \quad \text{für alle } x, y \in S. \tag{9.1}$$

Im Folgenden wird diese dann weiter untersucht werden.

Satz 9.22. *Gilt (9.1), dann konvergiert $\{\nu p^n\}$ für jede Startverteilung ν exponentiell gegen μ. Genauer gilt für alle $n \in \mathbb{N}$ und $\nu \in W(S)$:*

$$d(\nu p^n, \mu) \leq (1-\delta)^{\lfloor n/r \rfloor}.$$

Bemerkung. Insbesondere ist μ das eindeutige Gleichgewicht. Ist μ' nämlich eine andere Wahrscheinlichkeitsverteilung mit $\mu' p = \mu'$, dann folgt für $n \to \infty$

$$d(\mu', \mu) = d(\mu' p^n, \mu) \xrightarrow{n\to\infty} 0$$

und somit $d(\mu, \mu') = 0$, was $\mu = \mu'$ beweist.

Beweis von Satz 9.22. Mit δ, r aus (9.1) wird durch die Zerlegung

$$p^r(x, y) = \delta\mu(y) + (1 - \delta)q(x, y)$$

eine stochastische Matrix q definiert, weil aus (9.1) folgt, dass

$$(1 - \delta)q(x, y) = p^r(x, y) - \delta\mu(y) \geq 0,$$

und aus

$$\sum_{y\in S} p^r(x, y) = 1, \quad \sum_{y\in S} \mu(y) = 1$$

folgt, dass

$$\sum_{y\in S} q(x, y) = 1 \quad \text{für alle } x \in S.$$

Mit $\lambda := 1 - \delta$ gilt für alle $\nu \in W(S)$

$$\nu p^r = (1 - \lambda)\mu + \lambda\nu p. \tag{9.2}$$

Per vollständiger Induktion zeigen wir, dass

$$\nu p^{kr} = (1 - \lambda^k)\mu + \lambda^k \nu q^k \quad \text{für alle } k \geq 0,\ \nu \in W(S). \tag{9.3}$$

Für $k = 0$ ist diese Aussage trivial. Gilt (9.3) für ein $k \geq 0$, so erhalten wir durch Anwendung von (9.2) auf $\nu' p^r$ mit $\nu' := \nu q^k$

$$\begin{aligned}
\nu p^{(k+1)r} &= \nu p^{kr} p^r = [(1 - \lambda^k)\mu + \lambda^k \nu']p^r \\
&\overset{(9.2)}{=} (1 - \lambda^k)\mu p^r + (1 - \lambda)\lambda^k \mu + \lambda^{k+1}\nu' q \\
&= (1 - \lambda^k)\mu + (1 - \lambda)\lambda^k \mu + \lambda^{k+1}\nu' q \\
&= (1 - \lambda^{k+1})\mu + \lambda^{k+1}\nu q^{k+1}.
\end{aligned}$$

Für $n \in \mathbb{N}$, $n = kr + i$ mit $k \in \mathbb{N}$ und $0 \leq i < r$, folgt

$$\nu p^n = \nu p^{kr} p^i \overset{(9.3)}{=} (1 - \lambda^k)\mu p^i + \lambda^k \nu q^k p^i$$

und somit

$$\nu p^n - \mu = \lambda^k(\nu q^k p^i - \mu) \quad \text{für alle } \nu \in W(S).$$

Also gilt

$$d(\nu p^n, \mu) = \frac{1}{2}\|\nu p^n - \mu\|_1 = \lambda^k \underbrace{d(\nu q^k p^i, \mu)}_{\leq 1} \leq \lambda^k$$

nach der Bemerkung vor Satz 9.22. □

Wir kommen zur Minorisierungsbedingung (9.1) zurück. Welche Übergangsmatrizen p erfüllen diese?

Definition 9.23.

(i) Eine stochastische Matrix p heißt ***irreduzibel****, falls es für alle $x, y \in S$ ein $n \in \mathbb{N}$ gibt, so dass $p^n(x, y) > 0$.*

(ii) Die ***Periode*** *von $x \in S$ ist definiert als*

$$\text{Periode}(x) := \text{ggT}(R(x))$$

mit $R(x) := \{n \in \mathbb{N} : p^n(x, x) > 0\}$. p heißt ***aperiodisch****, falls $\text{Periode}(x) = 1$ für alle $x \in S$.*

Bemerkung. Eine stochastische Matrix p ist genau dann reduzibel, falls es eine Permutationsmatrix $P \in \mathbb{R}^{m\times m}$ und quadratische Matrizen A, C existieren, so dass für die Übergangsmatrix $M_{ij} = p(s_i, s_j)$, $i, j = 1, \ldots, m$, gilt

$$PMP^T = \begin{bmatrix} A & B \\ 0 & C \end{bmatrix}.$$

Wir benötigen das folgende Resultat aus der elementaren Zahlentheorie. Den Beweis geben wir der Vollständigkeit halber an.

Lemma 9.24. *Gegeben sei die Menge $A = \{a_1, a_2, \ldots\} \subset \mathbb{N}$ mit den Eigenschaften*

(i) $\text{ggT}(A) = 1$,

(ii) aus $a, b \in A$ folgt $a + b \in A$.

Dann existiert eine Zahl $N < \infty$, so dass $n \in A$ für alle $n \geq N$.

Beweis. (i) Wir zeigen zunächst, dass für beliebige Zahlen $a, b \in \mathbb{N}$ Zahlen $x, y \in \mathbb{N}$ existieren mit

$$ax - by = \text{ggT}(a, b).$$

Hierzu sei ohne Einschränkung der Allgemeinheit $\text{ggT}(A) = 1$, denn sonst betrachte einfach $a/\text{ggT}(a, b)$ und $b/\text{ggT}(a, b)$. Definieren wir die $b - 1$ Zahlen

$$\begin{aligned} z_1 &:= a \mod b, \\ z_2 &:= 2a \mod b, \\ &\vdots \\ z_{b-1} &:= (b-1)a \mod b, \end{aligned}$$

so gilt $0 \leq z_i < b$ für alle i. Weiter gilt $z_i \neq 0$ für alle i, denn sonst gäbe es ein $p \in \mathbb{N}$ mit $a = pb$, d.h. b teilt a im Widerspruch zu $\text{ggT}(a, b) = 1$. Ist hingegen $z_i \neq 1$ für alle i, so muss es $0 < k < \ell < b$ geben mit

$$ka \mod b = z_k = z_\ell = \ell a \mod b.$$

Dann ist aber $(\ell - k)a \mod b = 0$, d.h. b teilt $(\ell - k)a$, was wegen $\mathrm{ggT}(a, b) = 1$ auf den Widerspruch b teilt $0 \neq \ell - k < b$ führt. Folglich gibt es ein $0 < k < b$ mit $z_k = ka \mod b = 1$, oder anders ausgedrückt

$$ka - \ell b = 1.$$

(ii) Aussage (i) kann mittels vollständiger Induktion auf K Zahlen verallgemeinert werden. Denn gibt es $K - 1$ Zahlen $n_1, n_2, \ldots, n_{K-1} \in \mathbb{Z}$ mit

$$b = \sum_{k=1}^{K-1} n_k a_k = \mathrm{ggT}(a_1, a_2, \ldots, a_{K-1}),$$

so existieren nach Aussage (i) $x, y \in \mathbb{N}$ mit

$$bx - a_K y = \mathrm{ggT}(b, a_K) = \mathrm{ggT}(a_1, a_2, \ldots, a_K).$$

Setzen wir $\tilde{n}_k := xn_k \in \mathbb{Z}$ für alle $0 < k < K$ und $\tilde{n}_K := -y \in \mathbb{Z}$, so folgt

$$\sum_{k=1}^{K} \tilde{n}_k a_k = \mathrm{ggT}(a_1, a_2, \ldots, a_K).$$

(iii) Da jedes $a \in A$ eine endliche Primfaktorzerlegung besitzt, gibt es ein $K < \infty$ mit

$$\mathrm{ggT}(a_1, a_2, \ldots, a_K) = 1.$$

Gemäß Aussage (ii) existieren also Zahlen $n_1, n_2, \ldots, n_K \in \mathbb{Z}$ mit

$$\sum_{k=1}^{K} n_k a_k = 1.$$

Setzen wir

$$L := \max\{|n_1|, |n_2|, \ldots, |n_K|\}$$

und

$$N := La_1(a_1 + a_2 + \ldots + a_K),$$

dann gibt es zu jedem $n \geq N$ eine eindeutige Zerlegung

$$n = N + ka_1 + \ell$$

mit $k \geq 0$ und $0 \leq \ell < a_1$. Nun gilt aber

$$n = La_1(a_1 + a_2 + \ldots + a_K) + ka_1 + \ell \sum_{k=1}^{K} n_k a_k = \sum_{k=1}^{K} m_k a_k$$

mit nichtnegativen, ganzzahligen Koeffizienten

$$\begin{aligned} m_1 &= La_1 + k + n_1 \geq 0, \\ m_2 &= La_2 + n_2 \geq 0, \\ &\vdots \\ m_K &= La_K + n_K \geq 0, \end{aligned}$$

was wegen der Abgeschlossenheit von A bezüglich der Addition schließlich die Behauptung liefert.

□

Bemerkung. Für stochastische Matrizen p gilt mit $t, s \in \mathbb{N}$

$$p^{t+s}(x,y) \geq p^t(x,z) \cdot p^s(z,y) \quad \text{für alle } z \in S,$$

weil

$$p^{t+s}(x,y) = (p^t p^s)(x,y) = \sum_{a \in S} p^t(x,a) p^s(a,y) \geq p^t(x,z) p^s(z,y).$$

Lemma 9.25. *Sei p irreduzibel. Dann gilt*

(i) $\text{Periode}(x) = \text{Periode}(y)$ *für alle $x, y \in S$,*

(ii) es gibt ein $r > 0$ mit $p^r(x,y) > 0$ für alle $x, y \in S$, falls p zusätzlich aperiodisch ist.

Beweis.

(i) Weil p irreduzibel ist, existieren zu $x, y \in S$ Zahlen $t, s \in \mathbb{N}$ mit $p^s(x,y) > 0$ und $p^t(y,x) > 0$. Für $a := s + t$ folgt nach der letzten Bemerkung

$$p^a(x,x) \geq p^s(x,y) \cdot p^t(y,x) > 0.$$

Also ist $a \in R(x)$. Ferner gilt für $n \in R(y)$

$$p^{n+a}(x,x) \geq p^s(x,y) \cdot p^n(y,y) \cdot p^t(y,x) > 0,$$

was $n + a \in R(x)$ impliziert. $\text{Periode}(x)$ ist gemeinsamer Teiler von $R(x)$, also auch von a, $n + a$ und damit auch von n für alle $n \in R(y)$. Daher ist $\text{Periode}(x)$ ein gemeinsamer Teiler von $R(y)$, woraus $\text{Periode}(x) \leq \text{Periode}(y)$ folgt. Analog gilt $\text{Periode}(y) \leq \text{Periode}(x)$ und somit $\text{Periode}(x) = \text{Periode}(y)$.

(ii) $R(x)$ ist abgeschlossen unter Addition, weil für $s, t \in R(x)$ gilt

$$p^{s+t}(x,x) \geq p^s(x,x) \cdot p^t(x,x) > 0.$$

Da p aperiodisch ist, gilt $\text{ggT}(R(x)) = 1$ für alle $x \in S$. Nach Lemma 9.24 gibt es für alle x ein $r(x) \in \mathbb{N}$ mit $n \in R(x)$ für alle $n \geq r(x)$, d.h. $p^n(x,x) > 0$ für alle $n \geq r(x)$. Weil p irreduzibel ist, folgt, dass für alle $x, y \in S$ ein $r(x,y) \in \mathbb{N}$ existiert, so dass

$$p^{n+r(x,y)}(x,y) \geq p^n(x,x) \cdot p^{r(x,y)}(x,y) > 0 \quad \text{für alle } n \geq r(x).$$

Für $r \geq \max_{x,y \in S} r(x,y) + r(x)$ folgt dann $p^r(x,y) > 0$ für alle $x, y \in S$.

□

Satz 9.26 (Konvergenzsatz für endliche Markovketten). *Ist p irreduzibel und aperiodisch mit Gleichgewicht μ, so gilt*

$$\lim_{n \to \infty} d(\nu p^n, \mu) = 0 \quad \textit{für alle } \nu \in W(S).$$

Beweis. Weil p irreduzibel und aperiodisch ist, gibt es nach Lemma 9.25 (ii) ein $r \in \mathbb{N}$ mit

$$p^r(x,y) > 0 \quad \text{für alle } x, y \in S.$$

Sei $\delta = \min_{x,y \in S} p^r(x,y) > 0$. Dann gilt

$$p^r(x,y) \geq \delta \geq \delta \mu(y) \quad \text{für alle } x, y \in S.$$

Mit Satz 9.22 folgt die Behauptung. □

Die Markov-Ketten-Monte-Carlo-Methode

Wir wollen aus einer Menge S Stichproben mit Wahrscheinlichkeitsverteilung μ ziehen. Dabei sei μ nur bis auf eine Normierungskonstante bekannt, z.B. Gleichverteilung auf einer großen Menge S unbekannter Mächtigkeit. Bei der Markov-Ketten-Monte-Carlo-Methode (MCMC-Methode) konstruiert man eine Markov-Kette X_n, deren Gleichgewichtsverteilung μ ist. Nach Satz 9.26 kann man Stichproben von μ approximativ gewinnen, indem man X_n für großes n auswertet.

Wie konstruiert man eine Markov-Kette zu einem vorgegebenen Gleichgewicht? Wir stellen den Metropolis-Algorithmus vor. Sei eine Markov-Kette auf der endlichen Menge S mit irreduzibler Matrix q gegeben. Die Übergangswahrscheinlichkeiten $q(x,y)$ werden so umgewichtet, dass man eine Markov-Kette mit μ als Gleichgewichtsverteilung erhält.

Definition 9.27 (Metropolis-Kette). *Die Markov-Kette mit Übergangsmatrix*

$$p(x,y) = \begin{cases} \min\left(1, \frac{\mu(y)}{\mu(x)}\frac{q(y,x)}{q(x,y)}\right) q(x,y), & \text{falls } x \neq y, \\ 1 - \sum_{z \neq x} p(x,z), & \text{falls } x = y, \end{cases}$$

heißt ***Metropolis-Kette*** *mit Vorschlagsverteilung $q(x,y)$ und Gleichgewicht μ.*

Satz 9.28. *μ ist reversibel bzgl. p.*

Beweis. Der Fall $x = y$ ist trivial. Daher sei $x \neq y$. Dann gilt

$$\begin{aligned} \mu(x)p(x,y) &= \mu(x) \min\left(1, \frac{\mu(y)}{\mu(x)}\frac{q(y,x)}{q(x,y)}\right) q(x,y) = \min(\mu(x)q(x,y), \mu(y)q(y,x)) \\ &= \mu(y)p(y,x). \end{aligned}$$

□

Bemerkung.

(a) Man kann einen Übergang der p-Kette als zweistufiges Zufallsexperiment auffassen. Die Referenzkette q schlägt einen Übergang von x nach y mit Wahrscheinlichkeit $q(x,y)$ vor. Anschließend wirft man eine Münze mit Erfolgswahrscheinlichkeit

$$\alpha(x,y) = \min\left(1, \frac{\mu(y)}{\mu(x)}\frac{q(y,x)}{q(x,y)}\right).$$

Bei Erfolg wird der von der Referenzkette vorgeschlagene Übergang nach μ akzeptiert, ansonsten verharrt man im momentanen Zustand x.

(b) Ist q symmetrisch, d.h. $q(x,y) = q(y,x)$ für alle $x, y \in S$, dann ist

$$\alpha(x,y) = \min\left(1, \frac{\mu(y)}{\mu(x)}\right), \quad x \neq y.$$

Ein vorgeschlagener Übergang zu einem Zustand mit höherem Gewicht $\mu(y)$ wird also stets akzeptiert.

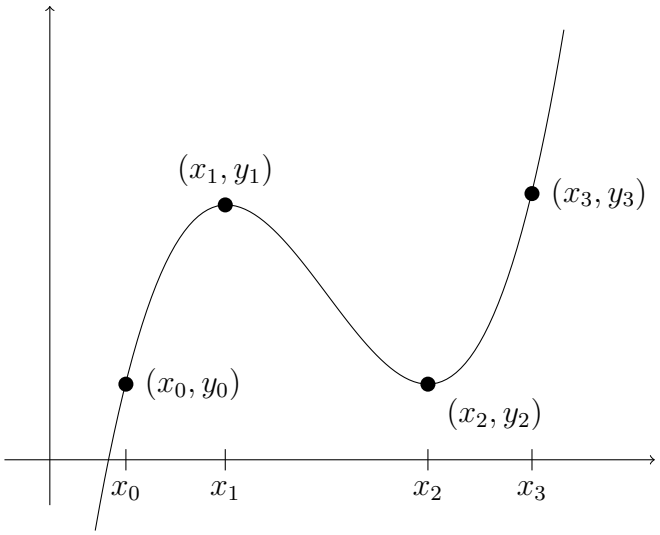
(c) Man beachte, dass zur Konstruktion von p nur der Quotient $\mu(y)/\mu(x)$ für $x \neq y$ mit $q(x,y) > 0$ bekannt sein muss. Die Aperiodizität und die Irreduzibilität von p ist von Fall zu Fall zu klären.

10 Interpolation

Bei der (lagrangeschen) Interpolation ist das Ziel, eine Funktion φ, die Interpolierende, in einem Funktionenraum Φ so zu finden, dass $\varphi \in \Phi$ an $n+1$ Stellen x_i, $i = 0, \ldots, n$, mit vorgegebenen Werten y_i, $i = 0, \ldots, n$, übereinstimmt, d.h.

$$\varphi(x_i) = y_i, \quad i = 0, \ldots, n. \tag{10.1}$$

Bemerkung. Der Interpolationsoperator $\mathcal{I}_n : \mathbb{K}^{n+1} \to \Phi$ definiert durch $y \mapsto \varphi$ ist ein linearer Operator. Dieser bildet diskrete Werte $y = [y_0, \ldots, y_n]^T$ auf Funktionen φ ab, auf die Methoden der Analysis angewendet werden können.



Die einfachste Wahl für Φ sind algebraische Polynome. Im Folgenden betrachten wir daher

$$\Phi = \Pi_n := \left\{ \sum_{j=0}^{n} a_j x^j, \ a_j \in \mathbb{K} \right\}$$

den **Raum der Polynome vom Grad höchstens n**.

Unter der Vorraussetzung, dass die Stützstellen x_i paarweise verschieden sind, kann alternativ zur **Monombasis** x^j, $j = 0, \ldots, n$, die so genannte *Lagrange-Basis* definiert werden.

Definition 10.1. *Die Polynome*

$$L_i(x) := \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j} \ \in \ \Pi_n, \quad i = 0, \ldots, n,$$

werden als ***Lagrange-Basispolynome*** *bezeichnet.*

Satz 10.2. *Es gilt $L_i(x_j) = \delta_{ij}$, $i, j = 0, \dots, n$, und (10.1) besitzt genau die Lösung*

$$p = \sum_{i=0}^{n} y_i L_i.$$

Insbesondere ist L_i, $i = 0, \dots, n$, eine Basis von Π_n.

Beweis. $L_i(x_j) = \delta_{ij}$ ist offensichtlich. Hiermit folgt sofort die Existenz aus

$$p(x_j) = \sum_{i=0}^{n} y_i \underbrace{L_i(x_j)}_{\delta_{ij}} = y_j, \quad j = 0, \dots, n.$$

Für die Eindeutigkeit von p betrachte zwei Lösungen p, q von (10.1). Dann folgt aus

$$(p - q)(x_i) = 0, \quad i = 0, \dots, n,$$

dass das Polynom $p - q \in \Pi_n$ genau $n + 1$ Nullstellen besitzt. Nach dem Fundamentalsatz der Algebra folgt daraus $p - q = 0$. Die Basiseigenschaft erhält man aus

$$p = \sum_{i=0}^{n} p(x_i) L_i \quad \text{für alle } p \in \Pi_n.$$

□

In der nächsten Bemerkung geben wir einen allgemeineren Beweis, der ohne den Fundamentalsatz der Algebra auskommt.

Bemerkung (Verallgemeinerung von Satz 10.2 auf beliebigen Funktionenräume). Für allgemeine Funktionenräume Φ mit Basis $\varphi_0, \dots, \varphi_n$ lassen sich Lagrange-Basisfunktionen L_i^Φ durch

$$L_i^\Phi(x) := \frac{\det M(i, x)}{\det M} \in \Phi, \quad i = 0, \dots, n,$$

definieren, falls die so genannte **Vandermonde-Matrix** $M \in \mathbb{K}^{(n+1)\times(n+1)}$ mit den Einträgen $M_{ij} = \varphi_j(x_i)$ regulär ist. Die Matrix $M(i, x) \in \mathbb{K}^{(n+1)\times(n+1)}$ entsteht aus

$$M = \begin{bmatrix} \varphi_0(x_0) & \cdots & \varphi_n(x_0) \\ \vdots & & \vdots \\ \varphi_0(x_n) & \cdots & \varphi_n(x_n) \end{bmatrix}$$

durch Ersetzen der i-ten Zeile mit dem Vektor $[\varphi_0(x), \dots, \varphi_n(x)]$. Dann gilt offenbar $L_i^\Phi(x_j) = \delta_{ij}$, $i, j = 0, \dots, n$, und durch

$$\varphi := \sum_{i=0}^{n} y_i L_i^\Phi \in \Phi$$

ist die eindeutige Lösung von (10.1) definiert. Die letzte Aussage erhält man aus der Basisdarstellung

$$\varphi = \sum_{i=0}^{n} \alpha_i \varphi_i$$

von $\varphi \in \Phi$. Dann ist das Problem (10.1) äquivalent zu dem linearen Gleichungssystem $M\alpha = y$ mit $\alpha = [\alpha_0, \ldots, \alpha_n]^T$ und $y = [y_0, \ldots, y_n]^T \in \mathbb{K}^{n+1}$. Dieses ist genau dann eindeutig lösbar, falls $\det M \neq 0$.

In den Übungsaufgaben zeigen wir, dass für $\Phi = \Pi_n$ die Lagrange-Basisfunktionen L_i^Φ mit den Lagrange-Basispolynomen aus Definition 10.1 übereinstimmen. Ferner werden wir sehen, dass im Fall $\Phi = \Pi_n$ gilt

$$\det M = \prod_{i=0}^{n} \prod_{j>i} (x_j - x_i).$$

Also gilt $\det M \neq 0$ genau dann, wenn die x_i paarweise verschieden sind.

Im folgenden Satz untersuchen wir die absolute Kondition $\kappa^\star(\mathcal{I}_n, y) = \|\mathcal{I}_n'(y)\|$ des Interpolationsproblems (10.1).

Satz 10.3. *Sei* $\det M \neq 0$. *Dann gilt für die absolute Kondition des Interpolationsproblems* $\mathcal{I}_n : \mathbb{K}^{n+1} \to \Phi$, $\mathcal{I}_n y = \varphi$, *definiert auf der kompakten Menge* D *bzgl. der Maximumnorm*

$$\max_{\substack{y \in \mathbb{K}^{n+1} \\ \|y\|_\infty = 1}} \kappa^\star(\mathcal{I}_n, y) = \Lambda_n$$

mit der so genannten ***Lebesgue-Konstanten***

$$\Lambda_n := \sup_{x \in D} \sum_{i=0}^{n} |L_i^\Phi(x)|.$$

Beweis. Weil $\mathcal{I}_n$ ein linearer Operator ist, gilt $\mathcal{I}_n'(y) = \mathcal{I}_n$. Wir müssen also zeigen, dass

$$\|\mathcal{I}_n\| = \max_{\|y\|_\infty = 1} \sup_{x \in D} |(\mathcal{I}_n y)(x)| = \Lambda_n.$$

Für alle $y \in \mathbb{K}^{n+1}$ gilt

$$|(\mathcal{I}_n y)(x)| = \left| \sum_{i=0}^{n} y_i L_i^\Phi(x) \right| \leq \sum_{i=0}^{n} |y_i| |L_i^\Phi(x)| \leq \|y\|_\infty \sum_{i=0}^{n} |L_i^\Phi(x)| \quad \text{für alle } x \in D$$

und somit $\|\mathcal{I}_n\| \leq \Lambda_n$. Für die umgekehrte Richtung sei $\hat{x} \in D$ so gewählt, dass

$$\sum_{i=0}^{n} |L_i^\Phi(\hat{x})| = \sup_{x \in D} \sum_{i=0}^{n} |L_i^\Phi(x)|$$

und $y \in \mathbb{K}^{n+1}$ sei der Vektor mit den Komponenten

$$y_i = \operatorname{sgn} L_i^\Phi(\hat{x}), \quad i = 0, \ldots, n.$$

Dann gilt $\|y\|_\infty = 1$ und

$$|(\mathcal{I}_n y)(\hat{x})| = \sum_{i=0}^{n} |L_i^\Phi(\hat{x})| = \sup_{x \in D} \sum_{i=0}^{n} |L_i^\Phi(x)|$$

und somit auch $\|\mathcal{I}_n\| \geq \Lambda_n$. □

In der folgenden Tabelle ist die Lebesgue-Konstante Λ_n für äquidistante Knoten $x_i = 2i/n - 1$ in Abhängigkeit von n angegeben. Offenbar wächst Λ_n für große n schnell über alle Grenzen. Der rechte Teil der Tabelle zeigt Λ_n für die so genannten **Tschebyscheff-Knoten**

$$x_j = \cos(\frac{2j+1}{2n+2}\pi), \quad j = 0, \dots, n$$

auf dem Intervall $D = [-1, 1]$.

n	Λ_n bei äquidistanten Knoten	Λ_n bei Tschebyscheff-Knoten
5	3.1	2.1
10	29.9	2.5
15	512.1	2.7
20	10986.5	2.9

Man kann nachweisen, dass im Fall der Tschebyscheff-Knoten $\Lambda_n \sim \log n$ gilt. Dies ist nachweislich das asymptotisch optimale Wachstumsverhalten.

Hermite-Interpolation

Sind zusätzlich zu den Funktionswerten y_i auch die Werte von Ableitungen an den Knoten vorgegeben, d.h. soll an (nicht notwendigerweise verschiedenenen) Knoten x_i, $i = 0, \dots, n$, gelten

$$p^{(d_i)}(x_i) = y_i \quad \text{für } i = 0, \dots, n, \tag{10.2}$$

mit $d_i := \max\{j : x_i = x_{i-j}\}$, so spricht man von **Hermite-Interpolation**, und $p \in \Pi_n$ wird als Hermite-Interpolierende bezeichnet. Dabei sind gleiche Knoten aufeinanderfolgend angeordnet, z.B.

x_i	x_0	$<$	x_1	$=$	x_2	$=$	x_3	$<$	x_4	$<$	x_5	$=$	x_6
d_i	0		0		1		2		0		0		1

Satz 10.4. *Es existiert genau ein $p \in \Pi_n$, das (10.2) erfüllt.*

Beweis. Das Problem (10.2) kann wieder als lineares Gleichungssystem in $n+1$ Unbekannten interpretiert werden. Es genügt zu zeigen, dass das homogene Problem

$$p^{(d_i)}(x_i) = 0, \quad i = 0, \dots, n,$$

nur die triviale Lösung besitzt. Dies sieht man, weil p (entspr. ihrer Vielfachheiten) insgesamt $n+1$ Nullstellen besitzt. Daher ist p das Nullpolynom. □

Bemerkung. Im Fall $d_i = 0$, $i = 0, \dots, n$, d.h. alle Knoten sind verschieden, erhält man Satz 10.2. Im Fall $x_0 = \cdots = x_n$, d.h. $d_i = i$, ist die Hermite-Interpolierende die abgebrochene **Taylor-Reihe**

$$\sum_{k=0}^{n} \frac{(x - x_0)^k}{k!} f^{(k)}(x_0)$$

von $f \in C^n$ um x_0, falls $y_i = f^{(d_i)}(x_0)$ gewählt wird.

10.1 Auswertung der Interpolierenden

Die Lagrange-Basispolyome sind gut für theoretische Zwecke, für die praktische Berechnung der Interpolierenden sind sie aber zu rechenaufwändig und instabil. Ist man an der Auswertung in nur einem Punkt x interessiert, so bietet sich folgende rekursive Berechnung an. Es bezeichne $p_{i,k} \in \Pi_k$ die eindeutig bestimmte Interpolierende zu den Daten (x_{i+j}, y_{i+j}), $j = 0, \dots, k$. Dann ist $p_{0,n}$ das gesuchte Polynom zu den Daten $(x_0, y_0), \dots, (x_n, y_n)$.

Lemma 10.5 (Aitken). *Es gilt:*

(i) $p_{i,0}(x) = y_i$, $i = 0, \dots, n$,

(ii)

$$p_{i,k}(x) = \frac{(x - x_i)p_{i+1,k-1}(x) - (x - x_{i+k})p_{i,k-1}(x)}{x_{i+k} - x_i}, \quad i = 0, \dots, n-k.$$

Beweis.

(i) ist klar per Definition.

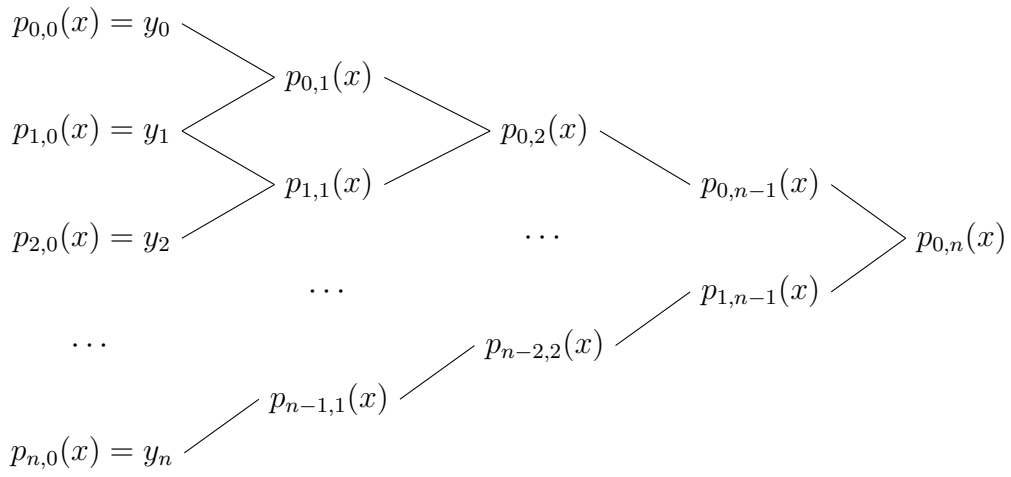
(ii) Sei $q(x)$ der Ausdruck auf der rechten Seite der Rekursionsformel. Dann ist $q \in \Pi_k$ und $q(x_{i+j}) = y_{i+j}$, $j = 1, \dots, k-1$, weil

$$p_{i+1,k-1}(x_{i+j}) = p_{i,k-1}(x_{i+j}) = y_{i+j}, \quad j = 1, \dots, k-1.$$

Ferner sieht man leicht, dass $q(x_i) = y_i$ und $q(x_{i+k}) = y_{i+k}$. Wegen der Eindeutigkeit der Interpolation folgt $p_{i,k} = q$.

□

Der gesuchte Wert $p(x) = p_{0,n}(x)$ der Interpolierenden ergibt sich wegen der Rekursionsformel aus Lemma 10.5 aus dem **Neville-Schema**



Beispiel 10.6. Man betrachte folgende Stützpunkte

i	0	1	2
x_i	1	4	16
y_i	1	2	4

Für $x = 2$ ergibt die Auswertung nach dem Neville-Schema

$$\begin{array}{ccccc} p_{0,0}(2) = y_0 = 1 & & & & \\ & \searrow & & & \\ & & p_{0,1}(2) = \frac{(2-1)\cdot 2-(2-4)\cdot 1}{4-1} = \frac{4}{3} & \searrow & \\ p_{1,0}(2) = y_1 = 2 & & & & p_{0,2}(2) = \frac{(2-1)\cdot\frac{5}{3}-(2-16)\cdot\frac{4}{3}}{16-1} = \frac{61}{45} \\ & \searrow & & \nearrow & \\ & & p_{1,1}(2) = \frac{(2-4)\cdot 4-(2-16)\cdot 2}{16-4} = \frac{5}{3} & & \\ p_{2,0}(2) = y_2 = 4 & \nearrow & & & \end{array}$$

Im nächsten Lemma zeigen wir eine Rekursionsfolem wie in Lemma 10.5 für die Hermite-Interpolation. Dazu sein p_J, $J \subset \mathbb{N}$, die Interpolierende zu den Stützstellen x_i, $i \in J$.

Lemma 10.7. *Unter der Vorraussetzung $x_i \neq x_j$ gilt für die Hermite-Interpolierende mit $i, j \in J$*

$$p_J(x) = \frac{(x_i - x)p_{J\setminus\{j\}}(x) - (x_j - x)p_{J\setminus\{i\}}(x)}{x_i - x_j}.$$

Beweis. Analog zum Beweis von Lemma 10.5 durch Überprüfen der Interpolationseigenschaften. □

Bemerkung. Eine Rekursionsformel für die Ableitungen von p_J kann durch Ableiten der der Formel in Lemma 10.7 gewonnen werden.

Soll ein Interpolationspolynom an $m \gg 1$ Stellen ausgewertet werden, so verwendet man anstelle des Aitken-Neville-Schemas, das $O(m \cdot n^2)$ Operationen benötigt, die folgende Newtonsche Interpolationsformel. Hierbei benötigt man $O(n^2 + m \cdot n)$ Operationen.

Newtonsche Interpolationsformel

Anders als beim Neville-Schema werden bei diesem Zugang zunächst die Koeffizienten des Polynoms in einer bestimmten Basis bestimmt und dann mittels des Horner-Schemas (siehe auch Abschnitt 1.1) ausgewertet. Als Basis des Polynomraums Π_n verwenden wir die so genannte *Newton-Basis*.

Definition 10.8. *Zu gegebenen $n+1$ Punkten $x_0, \ldots, x_n \in \mathbb{K}$ werden*

$$\omega_i(x) := \prod_{j=0}^{i-1}(x - x_j), \quad i = 0, \ldots, n,$$

*als **Newtonsche Basispolynome** bezeichnet. Dabei verwenden wir die Konvention, dass $\prod_{j=0}^{-1}(x - x_j) = 1$ ist.*

Sind die Koeffizienten $a_0, \ldots, a_n \in \mathbb{K}$ eines Polynoms $p \in \Pi_n$ in dieser Basis bestimmt, so kann für jedes $x \in \mathbb{K}$ der der Wert $p(x)$ mittels Horner-Schema

$$p(x) = a_0 + (x - x_0)(a_1 + (x - x_1)(a_2 + \ldots (a_{n-1} + (x - x_{n-1})a_n) \ldots))$$

ausgewertet werden, wobei die insgesamt $3n$ Operationen von rechts nach links auszuführen sind.

Im Folgenden behandeln wir die Berechnung der Koeffizienten $a_0, \ldots, a_n$. Diese können bei paarweise verschiedenen x_j, $j = 0, \ldots, n$, aus den Interpolationsbedingungen

$$\begin{aligned} y_0 &= p(x_0) = a_0, \\ y_1 &= p(x_1) = a_0 + a_1(x_1 - x_0), \\ y_2 &= p(x_2) = a_0 + a_1(x_2 - x_0) + a_2(x_2 - x_1)(x_2 - x_0), \end{aligned}$$

die zu

$$\begin{bmatrix} 1 & & & \\ \vdots & x_1 - x_0 & & \\ \vdots & x_2 - x_0 & (x_2 - x_1)(x_2 - x_0) & \\ \vdots & \vdots & \vdots & \ddots \\ 1 & x_n - x_0 & (x_n - x_1)(x_n - x_0) & \cdots \end{bmatrix} \begin{bmatrix} a_0 \\ \vdots \\ \vdots \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} y_0 \\ \vdots \\ \vdots \\ \vdots \\ y_n \end{bmatrix}$$

äquivalent sind, bestimmt werden. Da es sich hierbei um ein lineares Gleichungssystem mit unterer Dreiecksmatrix handelt, kann der Koeffizientenvektor $[a_0, \ldots, a_n]^T$ mittels Vorwärtseinsetzen (siehe Abschnitt 6.3) bestimmt werden. Wendet man dafür die spaltenweise Version (Algorithmus 6.30) an und berücksichtigt man die spezielle Struktur der Matrixeinträge des Gleichungssystems, so erhält man als Zwischenresultate die so genannten *dividierten Differenzen* (siehe Übungsaufgabe).

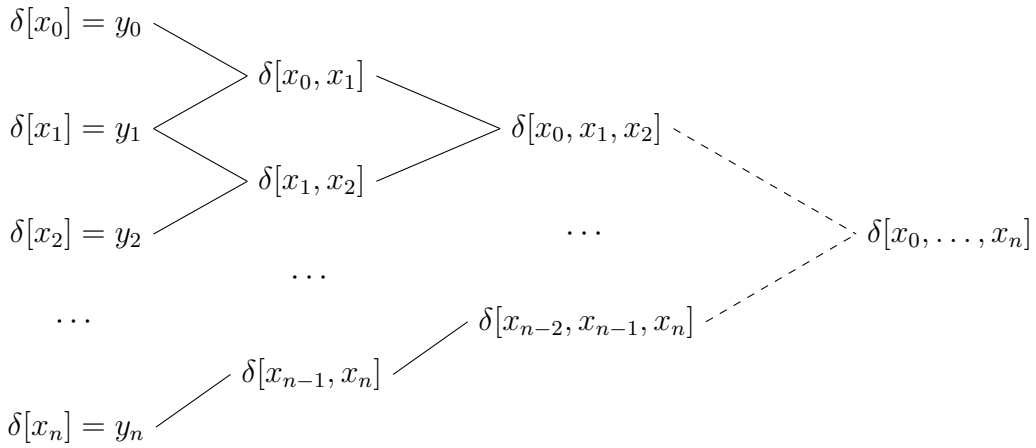
Definition 10.9. *Zu paarweise verschiedenen Punkten $x_0, \ldots, x_n$ und Werten $y_0, \ldots, y_n$ sind die* ***dividierten Differenzen*** *definiert als*

$$\delta[x_i] := y_i, \quad i = 0, \ldots, n,$$

und für $k = 1, \ldots, n$

$$\delta[x_i, x_{i+1}, \ldots, x_{i+k}] := \frac{\delta[x_{i+1}, x_{i+2}, \ldots, x_{i+k}] - \delta[x_i, \ldots, x_{i+k-1}]}{x_{i+k} - x_i}, \quad i = 0, \ldots, n - k.$$

Die rekursive Berechnungsformel kann durch folgendes Schema verdeutlicht werden.



Bemerkung.

(a) Die Berechnung aller dividierten Differenzen zu den Werten $(x_0, y_0), \ldots, (x_n, y_n)$ benötigt $O(n^2)$ Operationen.

(b) Sollen weitere Werte (x_{n+1}, y_{n+1}) aufgenommen werden, so können die zugehörigen dividierten Differenzen auf einfache Weise durch Anhängen einer weiteren Zeile berechnet werden.

Im Folgenden bestätigen wir, dass es sich bei $\delta[x_0, \ldots, x_k]$ tatsächlich um den gesuchten Koeffizienten a_k handelt. Ferner zeigen wir einige Eigenschaften der dividierten Differenzen auf.

Lemma 10.10. *Der führende Koeffizient in der Monombasis von $p_{i,k}$ (siehe Anfang von Abschnitt 10.1) stimmt mit $\delta[x_i, \ldots, x_{i+k}]$ überein.*

Beweis. Der Beweis erfolgt per Induktion über k. $k = 0$ ist klar nach Lemma 10.5 (i). Sei die Aussage für $k-1$ wahr. Nach dem Lemma von Aitken ist der führende Koeffizient von $p_{i,k}$ der führende Koeffizient von $\frac{p_{i+1,k-1} - p_{i,k-1}}{x_{i+k} - x_i}$. Nach Induktionsvorraussetzung ist dies

$$\frac{\delta[x_{i+1}, \ldots, x_{i+k}] - \delta[x_i, \ldots, x_{i+k-1}]}{x_{i+k} - x_i} = \delta[x_i, \ldots, x_{i+k}].$$

□

Wir haben die Bemerkung zu Satz 10.4 gesehen, dass die Hermite-Interpolation eine Verallgemeinerung der Lagrange-Interpolation ist. Um dividierte Differenenzen auch für die Hermite-Interpolation nutzen zu können, benötigt man eine Definition, die anders als Definition 10.9 auch gleiche Punkte zulässt. Lemma 10.10 zeigt eine alternative Definition über den führenden Koeffizienten des Hermite-Polynoms $p_{0,n}$ zu den Punkten $x_0, \ldots, x_n$ auf.

Definition 10.11. *Der führende Koeffizient des Hermite-Polynoms $p_{0,n}$ in der Monombasis zu den (nicht notwendig verschiedenen) Knoten $x_0, \ldots, x_n$ wird als* ***dividierte Differenz*** *$\delta[x_0, \ldots, x_n]$ bezeichnet.*

Wir wissen bereits, dass

$$\delta[x_0, \ldots, x_n] = \frac{f^{(n)}(x_0)}{n!}, \tag{10.3}$$

falls $x_0 = \ldots = x_n$ und $y_i = f^{(i)}(x_i)$, $i = 0, \ldots, n$, gilt. Ferner kann analog zum Beweis von Lemma 10.10 aus Lemma 10.7 die Rekursionsformel

$$\delta[x_0, \ldots, x_n] = \frac{\delta[x_0, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n] - \delta[x_0, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n]}{x_j - x_i} \tag{10.4}$$

hergeleitet werden, falls $x_i \neq x_j$. Mit (10.3) und (10.4) lassen sich die dividierten Differenzen aus den Funktionswerten und den Ableitungen einer Funktion f berechnen.

Satz 10.12. *Sei $p_{0,n}$ das Hermite-Interpolationspolynom. Dann ist*

$$p_{0,n} = \sum_{k=0}^{n} \delta[x_0, \ldots, x_k]\, \omega_k.$$

Ist $f \in C^{n+1}$ und $y_i = f^{(d_i)}(x_i)$, $i = 0, \ldots, n$, so gilt

$$f(x) = p_{0,n}(x) + \delta[x_0, \ldots, x_n, x]\, \omega_{n+1}(x).$$

Beweis. Wir zeigen die Behauptung per Induktion über n. Der Fall $n = 0$ ist klar. Sei also

$$p_{0,n-1} = \sum_{k=0}^{n-1} \delta[x_0, \ldots, x_k]\, \omega_k$$

das Interpolationspolynom zu den Punkten $x_0, \ldots, x_{n-1}$. Dann gilt nach Definition 10.11

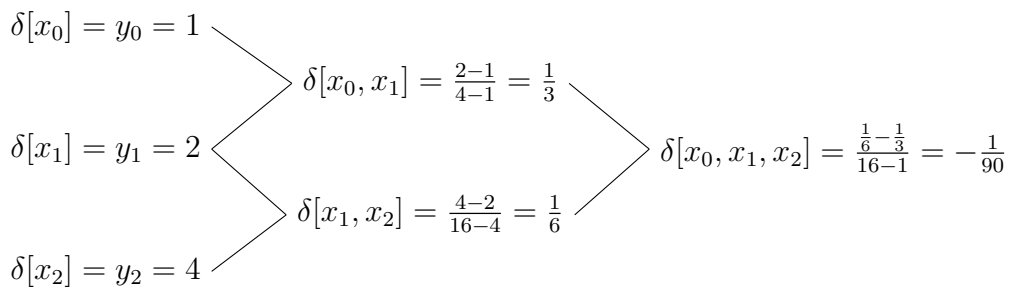
$$p_{0,n} = \delta[x_0, \ldots, x_n]x^n + b_{n-1}x^{n-1} + \ldots + b_0 = \delta[x_0, \ldots, x_n]\, \omega_n(x) + q(x)$$

mit einem $q \in \Pi_{n-1}$. Dann erfüllt $q = p_{0,n} - \delta[x_0, \ldots, x_n]\, \omega_n$ die Interpolationsbedingung für $x_0, \ldots, x_{n-1}$, woraus

$$q = p_{0,n-1} = \sum_{k=0}^{n-1} \delta[x_0, \ldots, x_{n-1}]\, \omega_k$$

folgt. Insbesondere folgt, dass $p_{0,n} + \delta[x_0, \ldots, x_n, x]\, \omega_{n+1}$ die Funktion f in den Punkten $x_0, \ldots, x_n$ und x interpoliert. □

Beispiel 10.13. Zu dem Interpolationsproblem aus Beispiel 10.6 soll das Newtonsche Interpolationspolynom mit Hilfe dividierter Differenzen bestimmt werden.



Damit ergibt sich

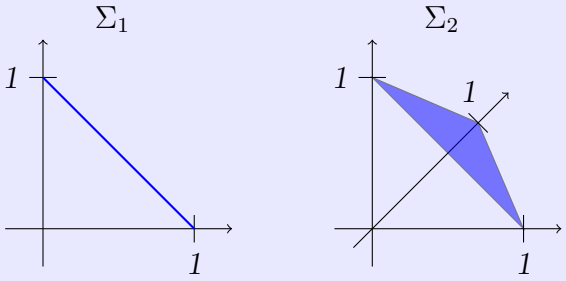
$$p_{0,3}(x) = 1 + \frac{1}{3}(x-1) - \frac{1}{90}(x-4)(x-1) = 1 + (x-1)\left(\frac{1}{3} - \frac{1}{90}(x-4)\right).$$

Satz 10.14 (Darstellungsformel für dividierte Differenzen). *Seien $x_0, \dots, x_n$ nicht notwendig verschiedene Punkte und $\delta[x_0, \dots, x_n]$ die dividierten Differenzen zu den Werten $y_i = f^{(d_i)}(x_i)$, $i = 0, \dots, n$, mit $f \in C^n$. Dann gilt*

$$\delta[x_0, \dots, x_n] = \int_{\Sigma_n} f^{(n)}\left(\sum_{i=0}^{n} s_i x_i\right) \mathrm{d}s$$

mit dem n-dimensionalen Simplex

$$\Sigma_n := \left\{ s = (s_0, \dots, s_n) \in \mathbb{R}^{n+1} : \sum_{i=0}^{n} s_i = 1,\ s_i \geq 0 \right\}.$$



Beweis. Wir zeigen die Aussage induktiv. Für $n = 0$ ist sie trivial. Sei die Behauptung für n wahr. Wenn alle Punkte x_i zusammenfallen, folgt die Behauptung aus (10.3). Wir dürfen annehmen, dass $x_0 \neq x_{n+1}$. Dann gilt

$$\begin{aligned}
&\int_{\sum_{i=0}^{n+1} s_i = 1} f^{(n+1)}\left(\sum_{i=0}^{n+1} s_i x_i\right) \mathrm{d}s \\
=&\int_{\sum_{i=1}^{n+1} s_i \leq 1} f^{(n+1)}\left(x_0 + \sum_{i=1}^{n+1} s_i (x_i - x_0)\right) \mathrm{d}s \\
=&\int_{\sum_{i=1}^{n} s_i \leq 1} \int_{s_{n+1}=0}^{1-\sum_{i=1}^{n} s_i} f^{(n+1)}\left(x_0 + \sum_{i=1}^{n} s_i (x_i - x_0) + s_{n+1}(x_{n+1} - x_0)\right) \mathrm{d}s \\
=&\frac{1}{x_{n+1} - x_0} \int_{\sum_{i=1}^{n} s_i \leq 1} \left[f^{(n)}\left(x_{n+1} + \sum_{i=1}^{n} s_i (x_i - x_{n+1})\right) - f^{(n)}\left(x_0 + \sum_{i=1}^{n} s_i (x_i - x_0)\right)\right] \mathrm{d}s \\
=&\frac{1}{x_{n+1} - x_0} \left(\delta[x_1, \dots, x_{n+1}] - \delta[x_0, \dots, x_n]\right) \overset{(10.4)}{=} \delta[x_0, \dots, x_{n+1}].
\end{aligned}$$

□

10.2 Interpolationsfehler

Sind die Daten y_i durch Auswertung einer Funktion f entstanden, d.h. ist $y_i = f^{(d_i)}(x_i)$, $i = 0, \ldots, n$, so kann der Interpolationsfehler

$$R_n(x) := f(x) - p_{0,n}(x)$$

betrachtet werden. Dabei ist $p_{0,n}$ das Hermite-Interpolationspolynom zu $x_0, \ldots, x_n$. Offenbar gilt $R_n(x_i) = 0$, $i = 0, \ldots, n$. Abhängig von n und der Glattheit von f ist $|R_n|$ auch zwischen den Punkten x_i klein.

Beispiel 10.15. Die Werte y_i in Beispiel 10.13 entstehen durch Auswertung der Funktion $f(x) = \sqrt{x}$ an den Stellen x_i, d.h. $y_i = \sqrt{x_i}$, $i = 0, \ldots, n$. An der Stelle $x = 2$ erhält man $p_{0,3}(2) = 1.3\overline{5}$. Der Wert von f an dieser Stelle beträgt $\sqrt{2} \approx 1.41$.

Wir schätzen den Approximationsfehler ab.

Satz 10.16. *Sei $f \in C^{n+1}[a, b]$. Dann gilt für den Approximationsfehler R_n der Hermite-Interpolierenden $p_{0,n} \in \Pi_n$ mit $x_i, x \in (a, b)$, $i = 0, \ldots, n$, dass*

$$R_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\,\omega_{n+1}(x)$$

für ein $\xi = \xi(x) \in (a, b)$.

Beweis. Nach Satz 10.12 und Satz 10.14 gilt

$$\begin{aligned} R_n(x) &= f(x) - p_{0,n}(x) = \delta[x_0, \ldots, x_n, x]\,\omega_{n+1}(x) \\ &= \omega_{n+1}(x) \int_{\Sigma_{n+1}} f^{(n+1)}\left(\sum_{i=0}^{n} s_i x_i + s_{n+1} x\right) \mathrm{d}s. \end{aligned}$$

Wegen $\mathrm{vol}(\Sigma_{n+1}) = \frac{1}{(n+1)!}$ und $x_i, x \in (a, b)$ gibt es nach dem Mittelwert der Integralrechnung ein $\xi \in (a, b)$ mit

$$\omega_{n+1}(x) \int_{\Sigma_{n+1}} f^{(n+1)}\left(\sum_{i=0}^{n} s_i x_i + s_{n+1} x\right) \mathrm{d}s = \frac{1}{(n+1)!} f^{(n+1)}(\xi)\,\omega_{n+1}(x).$$

□

Bemerkung. Im Fall $x_0 = \ldots = x_n$ ist

$$p_{0,n}(x) = \sum_{k=0}^{n} \frac{(x - x_0)^k}{k!} f^{(k)}(x_0)$$

die abgebrochene Taylor-Reihe und R_n ist das **Restglied der Taylorentwicklung**

$$f(x) - p_{0,n}(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x - x_0)^{n+1}.$$

Wie man aus Satz 10.16 sieht, hängt der Approximationsfehler R_n entscheidend von der Wahl der Punkte $x_0, \ldots, x_n$ in Form von

$$\omega_{n+1}(x) = \prod_{j=0}^{n}(x - x_j)$$

ab. Es stellt sich die Frage, ob bei wachsender Anzahl der Interpolationspunkte n der Interpolationsfehler R_n immer kleiner wird. Bei äquidistanten Punkten $x_i = a + i \cdot b_n$, $i = 0, \ldots, n$, $b_n = \frac{b-a}{n}$, lässt sich die Frage anhand des **Beispiels von Runge** verneinen:
Für die Funktion

$$f(x) = (1 + x^2)^{-1}$$

auf dem Intervall $[a, b] = [-5, 5]$ liegt zwar punktweise Konvergenz $\lim_{k\to\infty} |R_n(x)| = 0$ für $|x| \leq \tilde{x} \approx 3.63$ vor, für $|x| > \tilde{x}$ gilt jedoch $|R_n(x)| \xrightarrow{n\to\infty} \infty$ (siehe Übungsaufgabe). Die Ursache dafür ist die große Schwankung des Stützstellenpolynoms ω_{n+1} am Rande des Intervalls $[a, b]$.

Im nächsten Abschnitt zeigen wir, dass der Ausdruck

$$\max_{x\in[a,b]} |\omega_{n+1}(x)|$$

durch die schon bekannten Tschebyscheff-Punkte minimal wird.

10.3 Minimax-Eigenschaft der Tschebyscheff-Polynome

Wir haben bereits im Zusammenhang mit der Lebesgue-Konstanten gesehen, dass die Tschebyscheff-Knoten $x_j = \cos(\frac{2j+1}{2n+2}\pi)$, $j = 0, \ldots, n$, besonders günstige Eigenschaften besitzen. Bei diesen handelt es sich um die Nullstellen der in Beispiel 3.2 durch die Dreitermrekursion für $x \in \mathbb{R}$

$$T_k(x) := 2xT_{k-1}(x) - T_{k-2}(x), \quad T_0(x) = 1, \quad T_1(x) = x$$

eingeführten Tschebyscheff-Polynome T_n.

In diesem Abschnitt behandeln wir das folgende Minimax-Problem. Gesucht ist das Polynom $p \in \Pi_n$ mit führendem Koeffizienten 1 und minimaler Supremumsnorm, d.h.

$$\max_{x\in[a,b]} |p(x)| \to \min . \tag{10.5}$$

Wir dürfen annehmen, dass das Intervall $[a, b]$ mit $[-1, 1]$ übereinstimmt. Andernfalls betrachte die affine Transformation

$$y : [a, b] \to [-1, 1], \quad y \mapsto x(y) = 2\frac{y - a}{b - a} - 1,$$

und ihre Umkehrabbildung

$$y(x) = \frac{1 - x}{2}a + \frac{1 + x}{2}b.$$

Ist p eine Lösung von (10.5) auf $[-1, 1]$ mit führendem Koeffizienten 1, so ist $\hat{p}(y) := p(x(y))$ Lösung von (10.5) auf $[a, b]$ mit führendem Koeffizienten $2^n/(b - a)^n$.

Satz 10.17 (Eigenschaften der Tschebyscheff-Polynome).

(i) *die Koeffizienten von T_n sind ganzzahlig.*

(ii) *Der höchste Koeffizient von T_n, $n \geq 1$, ist $a_n = 2^{n-1}$.*

(iii) *T_n ist eine gerade Funktion, falls n gerade und ungerade, falls n ungerade ist.*

(iv) $T_n(1) = 1,\ T_n(-1) = (-1)^n$.

(v) $|T_n(x)| \leq 1$ *für* $x \in [-1, 1]$.

(vi) $|T_n(x)|$ *nimmt den Wert 1 an den sog. Tschebyscheff-Abszissen*

$$t_k := \cos\left(\frac{k}{n}\pi\right), \quad k = 0, \ldots, n,$$

an, d.h. $|T_n(x)| = 1 \iff x = t_k$ *für ein* $k = 0, \ldots, n$.

(vii) *Die Nullstellen von T_n sind*

$$x_k := \cos\left(\frac{2k-1}{2n}\pi\right), \quad k = 1, \ldots, n.$$

(viii) *Es gilt*

$$T_k(x) = \begin{cases} \cos(k \arccos(x)), & |x| \leq 1, \\ \cosh(k \arccos(x)), & x \geq 1, \\ (-1)^k \cosh(k \arccos(-x)), & x \leq -1. \end{cases}$$

(ix)

$$T_k(x) = \frac{1}{2}\left(\left(x + \sqrt{x^2-1}\right)^k + \left(x - \sqrt{x^2-1}\right)^k\right) \quad \text{für } x \in \mathbb{R}.$$

Beweis. (i)–(vii) überprüft man leicht. (viii) und (ix) beweist man, indem man nachweist, dass die Formeln der Dreitermrekursion (inkl. Startwerten) genügen. □

Satz 10.18. *Es bezeichne $\|f\|_\infty := \max_{x\in[-1,1]} |f(x)|$ die Supremumsnorm von f und a_n den führenden Koeffizienten von $p \in \Pi_n$. Dann gilt*

$$\|p\|_\infty \geq \frac{|a_n|}{2^{n-1}} \quad \text{für alle } p \in \Pi_n,\ a_n \neq 0.$$

Insbesondere sind die Tschebyscheff-Polynome minimal bzgl. der Supremumsnorm $\|\cdot\|_\infty$ unter den Polynomen vom Grad n mit führendem Koeffizienten $a_n = 2^{n-1}$.

Beweis. Angenommen, es existiert ein $p \in \Pi_n$ mit $a_n = 2^{n-1}$, $\|p\|_\infty < 1$. Dann ist $0 \neq$

$p - T_n \in \Pi_{n-1}$. An den Tschebyscheff-Abszissen $t_k = \cos(\frac{k}{n}\pi)$ gilt

$$\begin{aligned} T_n(t_{2k}) = \ \ 1, \quad p(t_{2k}) < \ \ 1 \quad &\Rightarrow \quad p(t_{2k}) - T_n(t_{2k}) < 0, \\ T_n(t_{2k+1}) = -1, \quad p(t_{2k+1}) > -1 \quad &\Rightarrow \quad p(t_{2k+1}) - T_n(t_{2k+1}) > 0. \end{aligned}$$

Also ist $p - T_n$ an den $n+1$ Punkten t_k abwechselnd positiv und negativ. Das Polynom $p - T_n$ besitz daher n verschiedene Nullstellen in $[-1, 1]$. Dies steht im Widerspruch zu $0 \neq p - T_n \in \Pi_{n-1}$.

Für ein beliebiges Polynom $p \in \Pi_n$ mit $a_n \neq 0$ folgt die Behauptung, weil $\tilde{p} := \frac{2^{n-1}}{a_n} p$ ein Polynom mit führendem Koeffizienten 2^{n-1} ist. □

Bemerkung. Wegen der Approximationsfehlerdarstellung in Satz 10.16 sind wir an Knoten $x_0, \ldots, x_n$ interessiert, für die

$$\|\omega_{n+1}\|_\infty = \| \prod_{i=0}^{n} (\cdot - x_i) \|_\infty$$

minimal wird. Anders gesagt suchen wir ein minimales Polynom ω_{n+1} mit führendem Koeffizienten 1. Nach Satz 10.18 ist dies $\omega_{n+1} = 2^{-n} T_{n+1}$ auf $[-1, 1]$, dessen Nullstellen gerade die Tschebyscheff-Knoten sind.

Für spätere Zwecke beweisen wir die folgende zweite Minimax-Eigenschaft der Tschebyscheff-Polynome.

Satz 10.19. *Sei $[a, b]$ ein beliebiges Intervall und $x_0 \notin [a, b]$. Dann ist das Polynom*

$$\hat{T}_n(x) := \frac{T_n(t)}{T_n(t_0)} \quad \text{mit } t(x) := 2\frac{x-a}{b-a} - 1, \ t_0 := t(x_0),$$

minimal bzgl. $\| \cdot \|_{\infty,[a,b]}$ unter den Polynomen $p \in \Pi_n$ mit $p(x_0) = 1$.

Beweis. Da alle Nullstellen von $T_n(t(x))$ in $[a, b]$ liegen, ist $c := T_n(t_0) \neq 0$ und $\hat{T}_n$ ist wohldefiniert. Ferner ist $\hat{T}_n(x_0) = 1$ und $|\hat{T}_n(x)| \leq 1/c$, $x \in [a, b]$.

Angenommen, es gebe ein Polynom $p \in \Pi_n$ mit $p(x_0) = 1$ und $|p(x)| < 1/c$ für alle $x \in [a, b]$. Dann ist x_0 eine Nullstelle von $\hat{T}_n - p$, d.h.

$$\hat{T}_n(x) - p(x) = q(x)(x - x_0)$$

für ein Polynom $0 \neq q \in \Pi_{n-1}$. Wie im Beweis von Satz 10.18 hat q an den Tschebyscheff-Abszissen $y_k := x(t_k)$ wechselndes Vorzeichen für $k = 0, \ldots, n$ und daher mindestens n verschiedene Nullstellen in $[a, b]$. Dies steht im Widerspruch zu $0 \neq q \in \Pi_{n-1}$. □

10.4 Grenzwertextrapolation

In diesem Abschnitt interessieren wir uns für die Berechnung des Grenzwertes

$$T^* := \lim_{h \to 0} T(h)$$

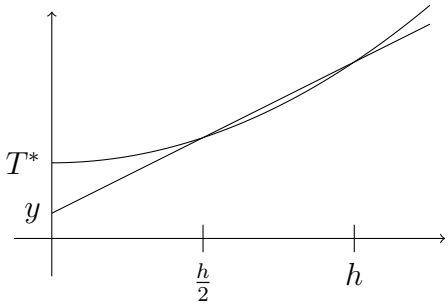
einer Funktion $T : (0, h_0] \to \mathbb{R}$, die nur für $h > 0$ ausgewertet werden kann. Eine typische Anwendung ist die Bestimmung der Ableitung einer Funktion f an einer Stelle x, d.h.

$$T(h) = \frac{f(x+h) - f(x)}{h}.$$

Um T^* zu approximieren, nutzt man die Existenz einer asymptotischen Entwicklung (z.B. Taylorentwicklung) von T in 0. Sei z.B.

$$T(h) = T^* + a_1 h + O(h^2). \tag{10.6}$$

Dann ist $T(\frac{h}{2}) = T^* + a_1 \frac{h}{2} + O(h^2)$, und es folgt $2T(\frac{h}{2}) - T(h) = T^* + O(h^2)$.



$y = 2T(\frac{h}{2}) - T(h)$ ist der Wert der Geraden durch $(\frac{h}{2}, T(\frac{h}{2})), (h, T(h))$ an der Stelle 0.

Gilt anstelle von (10.6), dass

$$T(h) = T^* + b_1 h^2 + O(h^4), \tag{10.7}$$

so ist $4T(\frac{h}{2}) - T(h) = 3T^* + O(h^4)$. In diesem Fall ist die Zahl $\frac{1}{3}(4T(\frac{h}{2}) - T(h))$ der Wert der interpolierenden Geraden für $((\frac{h}{2})^2, T(h)), (h^2, T(h))$ im Punkt 0, und wir haben

$$T^* - \frac{1}{3}(4T(\frac{h}{2}) - T(h)) = O(h^4).$$

Bemerkung. Für die Genauigkeitsaussage der jeweiligen Berechnungsformel ist nicht die Kenntniss der Koeffizienten a_1 bzw. b_1 in (10.6) und (10.7) erforderlich. Man muss nur wissen, dass (10.6) bzw. (10.7) gilt.

Für das folgende allgemeinere Extrapolationsverfahren machen wir die Annahme, dass $a_1, \ldots, a_n \in \mathbb{R}$ und $q \in \mathbb{N}$ existieren mit

$$T(h) = T^* + \sum_{i=1}^{n} a_i h^{qi} + O(h^{q(n+1)}).$$

Als Approximation an T^* verwenden wir $p(0)$, wobei $p \in \Pi_n$ das Interpolationspolynom zu $(h_0^q, T(h_0)), \ldots, (h_n^q, T(h_n))$ bei gegebenen Punkten $h_0, \ldots, h_n$ bezeichnet. $p(0)$ kann mittels Neville-Schema berechnet werden

$$\begin{aligned} T_{i,0} &= T(h_i), \quad i = 0, \ldots, n, \\ T_{i,k} &= T_{i,k-1} + \frac{T_{i+1,k-1} - T_{i,k-1}}{1 - \left(\frac{h_{i+k}}{h_i}\right)^q}, \quad i = 0, \ldots, n-k. \end{aligned}$$

Dann ist $T_{i,k}$ der Wert des Interpolationspolynoms für $(h_i^q, T(h_i)), \ldots, (h_{i+k}^q, T(h_{i+k}))$ an der Stelle 0.

Satz 10.20 (Konvergenz des Extrapolationsverfahrens). *Sei $T : (0, h_0] \to \mathbb{R}$ eine Abbildung und $h_0 > h_1 > \ldots > h_n > 0$. Angenommen, es exisitieren $q \in \mathbb{N}$, $a_1, \ldots, a_n \in \mathbb{R}$ und $c > 0$ mit*

$$T(h) = T^* + \sum_{i=1}^{n} a_i h^{qi} + a_{n+1}(h),$$

so dass $|a_{n+1}(h)| \leq c\, h^{q(n+1)}$. Gilt $h_{k+1} \leq \rho\, h_k$, $k = 0, \ldots, n-1$, für ein $\rho < 1$, dann gilt

$$|T_{i,k} - T^*| \leq \tilde{c} \prod_{j=0}^{k} h_{i+j}^q,$$

wobei $\tilde{c}$ nur von c, q und ρ abhängt. Insbesondere ist $|T_{0,n} - T^*| \leq \tilde{c}\, h_0^q \cdot h_1^q \cdot \ldots \cdot h_n^q$.

Beweis. Wir betrachten nur den Fall, dass $h_{k+1} = \rho h_k$, $k = 0, \ldots, n-1$. Setze $z_k = h_k^q$. Das Interpolationspolynom für $(z_0, T(h_0)), \ldots, (z_n, T(h_n))$ hat die Form

$$p(z) = \sum_{k=0}^{n} T(h_k) L_k(z), \quad L_k(z) := \prod_{\substack{i=0 \\ i \neq k}}^{n} \frac{z - z_i}{z_k - z_i}. \tag{10.8}$$

Wir benötigen einige Hilfsaussagen:

(1) Sei $p \in \Pi_n$, $p(x) = \sum_{k=0}^{n} c_k x^k$ mit positiven Nullstellen $x_1, \ldots, x_n$. Dann sind die c_k alternierend, d.h. $c_k c_{k+1} < 0$, $0 \leq k < n$.

Beweis. Einerseits gilt $p^{(k)}(0) = k c_k$ und andererseits wegen

$$p(x) = \alpha \prod_{k=1}^{n} (x - x_k)$$

für ein $\alpha \in \mathbb{R}$ nach der Produktregel $\operatorname{sgn}(p^{(k)}(0)) = (-1)^{n-k} \operatorname{sgn}(\alpha)$. □

(2) Es gilt

$$\sum_{k=0}^{n} z_k^i L_k(0) = \begin{cases} 1, & i = 0, \\ 0, & 1 \leq i \leq n. \end{cases}$$

Beweis. Für jedes $p \in \Pi_n$ gilt $p(z) = \sum_{k=0}^{n} p(z_k) L_k(z)$ und daher $z^i = \sum_{k=0}^{n} z_k^i L_k(z)$, $i = 0, \ldots, n$ und $z \in \mathbb{R}$. Die Behauptung folgt für $z = 0$. □

(3) Es gilt

$$\sum_{k=0}^{n} z_k^{n+1} |L_k(0)| \leq c' \prod_{k=0}^{n} z_k,$$

mit $c' := \prod_{i=1}^{\infty} \frac{1+\rho^{qi}}{1-\rho^{qi}}$.

Beweis. Betrachte $\tilde{p}(z) := \sum_{k=0}^{n} L_k(0) z^k$. Wegen $z_k/z_0 = \rho^{qk}$ folgt

$$\tilde{p}(\rho^{qi}) = \sum_{k=0}^{n} L_k(0)\rho^{qik} = z_0^{-i} \sum_{k=0}^{n} L_k(0) z_k^i.$$

Nach (ii) hat $\tilde{p}$ die Nullstellen ρ^{qi}, $1 \le i \le n$, und $\tilde{p}(1) = 1$. Daher ist nach (i) $L_k(0)L_{k+1}(0) < 0$, $0 \le k < n$, und wegen der Eindeutigkeit der Interpolation gilt $\tilde{p}(z) = \prod_{k=1}^{n} \frac{z-\rho^{qk}}{1-\rho^{qk}}$. Also folgt

$$\begin{aligned}
\sum_{k=0}^{n} z_k^{n+1} |L_k(0)| &= z_0^{n+1} \sum_{k=0}^{n} |L_k(0)| \rho^{qk(n+1)} = z_0^{n+1} |\tilde{p}(-\rho^{q(n+1)})| \\
&= z_0^{n+1} \prod_{k=1}^{n} \frac{\rho^{q(n+1)} + \rho^{qk}}{1-\rho^{qk}} \\
&= z_0^{n+1} \left(\prod_{k=1}^{n} \rho^{qk} \right) \left(\prod_{k=1}^{n} \frac{1+\rho^{q(n+1-k)}}{1-\rho^{qk}} \right) \\
&= \prod_{k=0}^{n} z_k \prod_{k=1}^{n} \frac{1+\rho^{qk}}{1-\rho^{qk}}.
\end{aligned}$$

Weil

$$\begin{aligned}
\prod_{k=1}^{n} \frac{1+\rho^{qk}}{1-\rho^{qk}} &= \exp\left(\ln \prod_{k=1}^{n} 1 + \frac{2\rho^{qk}}{1-\rho^{qk}} \right) = \exp \sum_{k=1}^{n} \ln\left(1 + \frac{2\rho^{qk}}{1-\rho^{qk}} \right) \\
&\le \exp\left(\sum_{k=1}^{n} 2 \frac{\rho^{qk}}{1-\rho^{qk}} \right)
\end{aligned}$$

und die Reihe im Exponenten für $n \to \infty$ konvergent ist, folgt die Behauptung. □

Mit diesen Hilfsaussagen setzen wir den Beweis von Satz 10.20 fort. Wegen (10.8) gilt

$$\begin{aligned}
T_{0,n} = p(0) &= \sum_{k=0}^{n} T(h_k) L_k(0) = \sum_{k=0}^{n} L_k(0) \left[T^* + \sum_{i=1}^{n} a_i h_k^{qi} + a_{n+1}(h_k) \right] \\
&= T^* \underbrace{\sum_{k=0}^{n} L_k(0)}_{=1} + \sum_{i=1}^{n} a_i \underbrace{\sum_{k=0}^{n} L_k(0) z_k^i}_{=0} + \sum_{k=1}^{n} L_k(0) a_{k+1}(h_k).
\end{aligned}$$

Also folgt

$$|T_{0,n} - T^*| = \left| \sum_{k=0}^{n} L_k(0) a_{k+1}(h_k) \right| \le c \sum_{k=0}^{n} |L_k(0)| z_k^{n+1} \le c\, c' \prod_{k=0}^{n} z_k, \quad c' = c'(\rho, q).$$

Damit ist der Satz für $T_{0,n}$ bewiesen. Für $T_{i,n}$ gilt er dann auch, weil wir $T_{i,n}$ als Endpunkt des Tableaus zu $(h_i, \ldots, h_{i+k})$ auffassen können. □

10.5 Trigonometrische Interpolation und die schnelle Fourier-Transformation

Definition 10.21. *Seien $c_0, \dots, c_{n-1} \in \mathbb{C}$ gegeben. Ist $c_{n-1} \neq 0$, so heißt $p : [0, 2\pi] \to \mathbb{C}$,*

$$p(x) := \sum_{j=0}^{n-1} c_j e^{\mathrm{i}jx},$$

***komplexes trigonometrisches Polynom** vom Grad $n-1$. Den Raum der komplexen trigonometrischen Polynome vom Grad höchstens $n-1$ bezeichnen wir mit $T_n^{\mathbb{C}}$. Mit $T_n^{\mathbb{R}}$ bezeichnen wir den Raum der **reellen trigonometrischen Polynome** der Form*

$$q(x) = \frac{a_0}{2} + \sum_{j=1}^{n} (a_j \cos(jx) + b_j \sin(jx)), \quad \text{falls } n = 2m+1 \text{ ungerade},$$

und

$$q(x) = \frac{a_0}{2} + \sum_{j=1}^{m-1} (a_j \cos(jx) + b_j \sin(jx)) + \frac{a_m}{2} \cos(mx), \quad \text{falls } n = 2m \text{ gerade}.$$

Hierbei sind $a_j, b_j \in \mathbb{R}$.

Entsprechend der Eigenschaften der Funktionen $e^{\mathrm{i}jx}, \cos(jx), \sin(jx)$ verwendet man trigonometrische Interpolation zur Analyse periodischer Funktionen $f(x) = f(x + 2\pi)$. Das Interpolationsproblem, finde $p \in T_n^{\mathbb{C}}$ mit

$$p(x_k) = y_k, \quad k = 0, \dots, n-1, \tag{10.9}$$

bei gegebenen (x_k, y_k), $k = 0, \dots, n-1$, mit $0 \leq x_0 < x_1 < \dots < x_{n-1} < 2\pi$ und $y_k \in \mathbb{C}$, kann durch die Transformation $x \mapsto e^{\mathrm{i}x} =: z$ kann dieses Problem in das algebraische Interpolationsproblem

$$\hat{p}(z_k) = y_k, \quad k = 0, \dots, n-1,$$

mit dem algebraischen Polynom $\hat{p}(z) := \sum_{j=0}^{n-1} c_j z^j \in \Pi_{n-1}$ und den paarweise verschiedenen Punkten $z_k := e^{\mathrm{i}x_k}$ äquivalent umgeformt werden. Aus den Aussagen für die algebraische Interpolation erhalten wir somit

Satz 10.22. *Seien die Punkte $0 \leq x_0 < x_1 < \dots < x_{n-1} < 2\pi$ gegeben. Dann gibt es genau ein Polynom $p \in T_n^{\mathbb{C}}$, das (10.9) löst.*

Bemerkung. Insbesondere ist die Bemerkung nach Satz 10.2 zur Definition der Lagrange-Funktionen anwendbar. Hieraus erhält man auch die eindeutige Lösbarkeit des reellen Interpolationsproblems: finde $q \in T_n^{\mathbb{R}}$ mit

$$q(x_k) = y_k, \quad k = 0, \dots, n-1, \tag{10.10}$$

mit $y_k \in \mathbb{R}$, indem man die lineare Unabhängigkeit der Basis $\{1, \cos(jx), \sin(jx)\}$ von $T_n^{\mathbb{R}}$ nachweist; siehe die Übungsaufgaben.

Im Fall äquidistanter Stützstellen $x_k = 2\pi k/n$, $0 \le k < n$, können die Koeffizienten des trigonometrischen Interpolationspolynoms explizit angegeben werden. Mit anderen Worten können wir in diesem Fall die Vandermonde-Matrix explizit invertieren. Wir zeigen zunächst die Orthogonalität der Basisfunktionen $\varphi_j(x) = e^{\mathrm{i}jx}$ bzgl. des Skalarprodukts

$$(f, g) := \frac{1}{n} \sum_{k=0}^{n-1} f(x_k)\overline{g(x_k)}.$$

Im Folgenden sei $\omega_n := e^{2\pi\mathrm{i}/n}$.

Lemma 10.23. *Für die n-te Einheitswurzel ω_n gilt*

$$\sum_{j=0}^{n-1} \omega_n^{jk} \omega_n^{-j\ell} = n\, \delta_{k\ell}.$$

Beweis. Der Fall $k = \ell$ ist klar. Sei $k \neq \ell$, so gilt

$$0 = \omega_n^{n(k-\ell)} - 1 = (\omega_n^{k-\ell} - 1) \sum_{j=0}^{n-1} \omega_n^{j(k-\ell)}$$

Weil $\omega_n^{k-\ell} \neq 1$, folgt hieraus $\sum_{j=0}^{n-1} \omega_n^{j(k-\ell)} = 0$. □

Satz 10.24. *Für äquidistante Stützstellen $x_k = 2\pi k/n$ ist die Lösung des komplexen trigonometrischen Interpolationspolynoms (10.9) gegeben durch*

$$p(x) = \sum_{j=0}^{n-1} c_j e^{\mathrm{i}jx}, \quad c_j := \frac{1}{n} \sum_{k=0}^{n-1} \omega_n^{-jk} y_k, \quad j = 0, \dots, n-1.$$

Beweis. Einsetzen der angegebenen c_j ergibt

$$p(x_k) = \sum_{j=0}^{n-1} \left(\frac{1}{n} \sum_{\ell=0}^{n-1} \omega_n^{-j\ell} y_\ell \right) \omega_n^{jk} = \sum_{\ell=0}^{n-1} y_\ell \left(\frac{1}{n} \sum_{j=0}^{n-1} \omega_n^{j(k-\ell)} \right) = y_k$$

nach Lemma 10.23. □

Die reelle Version von Satz 10.24 lautet:

Satz 10.25. *Seien $x_k = 2\pi k/n$, $k = 0, \dots, n-1$ äquidistante Stützstellen. Gilt $y_k \in \mathbb{R}$, $k = 0, \dots, n-1$, in (10.9), so gilt für das komplexe trigonometrische Interpolationspolynom $p \in T_n^{\mathbb{R}}$ mit den Koeffizienten*

$$a_j = 2 \operatorname{Re} c_j = c_j + c_{n-j}, \quad b_j = -2 \operatorname{Im} c_j = \mathrm{i}(c_j - c_{n-j}).$$

Insbesondere erhält man aus Satz 10.24 für die Lösung des reellen trigonometrischen Interpolationspolynoms (10.10)

$$a_j = \frac{2}{n}\sum_{k=0}^{n-1} y_k \cos(jx_k), \quad b_j = \frac{2}{n}\sum_{k=0}^{n-1} y_k \sin(jx_k).$$

Beweis. Wegen $e^{2\pi \mathrm{i}(n-j)/n} = e^{-2\pi \mathrm{i} j/n}$ gilt

$$\sum_{j=0}^{n-1} c_j e^{\mathrm{i}jx_k} = p(x_k) = y_k = \overline{y}_k = \sum_{j=0}^{n-1} \overline{c}_j e^{-\mathrm{i}jx_k} \overset{c_n := c_0}{=} \sum_{j=0}^{n-1} \overline{c}_{n-j} e^{\mathrm{i}jx_k},$$

woraus $c_j = \overline{c}_{n-j}$ folgt. Insbesondere sind c_0 und, falls $n = 2m$, auch c_m reell. Für ungerade $n = 2m+1$ erhält man

$$\begin{aligned} p(x_k) &= c_0 + \sum_{j=1}^{2m} c_j e^{\mathrm{i}jx_k} = c_0 + \sum_{j=1}^{m} c_j e^{\mathrm{i}jx_k} + \sum_{j=1}^{m} \overline{c}_j e^{-\mathrm{i}jx_k} \\ &= c_0 + \sum_{j=1}^{m} 2\,\mathrm{Re}\,(c_j e^{\mathrm{i}jx_k}) \\ &= c_0 + \sum_{j=1}^{m} \underbrace{2(\mathrm{Re}\, c_j)}_{a_j} \cos(jx_k) \underbrace{-2(\mathrm{Im}\, c_j)}_{b_j} \sin(jx_k). \end{aligned}$$

Aus der Eindeutigkeit der reellen trigonometrischen Interpolation folgt

$$a_j = 2\,\mathrm{Re}\, c_j = c_j + \overline{c}_j = c_j + c_{n-j} \quad \text{und} \quad b_j = -2\,\mathrm{Re}\, c_j = \mathrm{i}(c_j - \overline{c}_j) = \mathrm{i}(c_j - c_{n-j}).$$

Für $n = 2m$ folgt die Behauptung analog. □

Die Abbildung $y_k \mapsto c_j$ aus Satz 10.24 wird bis auf den Skalar $\frac{1}{n}$ als *diskrete Fourier-Transformation* bezeichnet.

Definition 10.26. *Die Abbildung $F_n : \mathbb{C}^n \to \mathbb{C}^n$ definiert durch $F_n f = g$,*

$$g_j := \sum_{k=0}^{n-1} \omega_n^{-jk} f_k, \quad 0 \le j < n,$$

heißt ***diskrete Fourier-Transformation*** *(DFT) oder* ***Fourier-Analyse*** *der Länge n.*

Bemerkung.

(a) Die Fourier-Transformation kann als Multiplikation der Matrix $F_n \in \mathbb{C}^{n\times n}$,

$$(F_n)_{k\ell} = \omega_n^{-k\ell},$$

mit einem Vektor aufgefasst werden. Dies benötigt bei naiver Vorgehensweise $O(n^2)$ Operationen. Die im Folgenden vorgestellte **schnelle Fourier-Transformation** (FFT) (engl. Fast Fourier Transformation) kommt mit $O(n \log_2 n)$ Operationen aus.

(b) Nach Lemma 10.23 gilt $F_n^H F_n = n\,I$. Daher ist $\frac{1}{\sqrt{n}} F_n$ unitär, also insbesondere invertierbar und offenbar symmetrisch. Die Umkehrabbildung

$$F_n^{-1} = \frac{1}{n} F_n^H = \frac{1}{n} \overline{F}_n$$

wird als **Fourier-Synthese** bezeichnet. Sie entspricht bis auf Skalare der Auswertung des trigonometrischen Polynoms $p \in T_n^{\mathbb{C}}$ an den Stellen $x_k = 2\pi k/n$, weil

$$y_k = p(x_k) = \sum_{j=0}^{n-1} c_j e^{2\pi \mathrm{i} jk/n} = \sum_{j=0}^{n-1} c_j \omega_n^{jk}.$$

Wegen $F_n^{-1} x = \frac{1}{n} \overline{F}_n x = \frac{1}{n} \overline{F_n \overline{x}}$ kann die Fourier-Synthese ebenfalls mit Hilfe der FFT berechnet werden.

Der schnellen Fourier-Transformation (Cooley & Tukey, 1965) liegt die Idee zu Grunde, die Multiplikation mit F_n auf zwei getrennte Multiplikationen (halber Länge) mit $F_{n/2}$ zurückzuführen. Wir beschränken uns auf den Fall $n = 2m$, $m \in \mathbb{N}$.

Lemma 10.27 (Danielson-Lanczos). *Sei $n = 2m$ und $\omega_n = e^{\pm 2\pi \mathrm{i}/n}$. Dann gilt für die Komponenten*

$$g_j = \sum_{k=0}^{n-1} \omega_n^{jk} f_k, \quad j = 0, \ldots, n-1,$$

dass für $j = 0, \ldots, m-1$

$$g_{2j} = \sum_{k=0}^{m-1} \omega_m^{jk} (f_k + f_{k+m}), \quad g_{2j+1} = \sum_{k=0}^{m-1} \omega_m^{jk} (f_k - f_{k+m}) \omega_n^k.$$

Die Berechnung der g_j kann auf zwei gleichartige Probleme halber Größe zurückgeführt werden.

Beweis. Für den geraden Fall folgt wegen $\omega_n^n = 1$ und $\omega_n^2 = \omega_m$, dass

$$g_{2j} = \sum_{k=0}^{n-1} \omega_n^{2jk} f_k = \sum_{k=0}^{m-1} \omega_n^{2jk} f_k + \omega_n^{2j(k+m)} f_{k+m} = \sum_{k=0}^{m-1} \omega_m^{jk} (f_k + f_{k+m}).$$

Für ungerade Indizes gilt wegen $\omega_n^m = -1$

$$g_{2j+1} = \sum_{k=0}^{n-1} \omega_n^{(2j+1)k} f_k = \sum_{k=0}^{m-1} \omega_n^{(2j+1)k} f_k + \omega_n^{(2j+1)(k+m)} f_{k+m} = \sum_{k=0}^{m-1} \omega_m^{jk} \omega_n^k (f_k - f_{k+m}).$$

□

Bemerkung. In Matrixschreibweise lässt sich Lemma 10.27 wie folgt ausdrücken. Sei

$$\Omega_m = \operatorname{diag}(\omega_n^0, \ldots, \omega_n^{m-1})$$

und $\Pi_n : \mathbb{C}^n \to \mathbb{C}^n$ die Permutationsmatrix mit

$$\Pi_n v = (v_0, v_2, v_4, \ldots, v_{n-2}, v_1, v_3, \ldots, v_{n-1})^T.$$

Dann gilt

$$\begin{aligned} \Pi_n F_n = \begin{bmatrix} F_m & 0 \\ 0 & F_m \end{bmatrix} \begin{bmatrix} I_m & I_m \\ \Omega_m & -\Omega_m \end{bmatrix} &\iff F_n \Pi_n^T = B_n \begin{bmatrix} F_m & 0 \\ 0 & F_m \end{bmatrix} \\ \iff F_n = B_n \begin{bmatrix} F_m & 0 \\ 0 & F_m \end{bmatrix} \Pi_n & \end{aligned} \tag{10.11}$$

mit den sog. Butterfly-Matrizen

$$B_n := \begin{bmatrix} I_m & \Omega_m \\ I_m & -\Omega_m \end{bmatrix}.$$

Die Bezeichnung Butterfly-Matrix wird klar, wenn man B_n auf einen Vektor anwendet

$$B_n \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_1 + \Omega_m x_2 \\ x_1 - \Omega_m x_2 \end{bmatrix}.$$

Aus der rekursiven Anwendung von (10.11) erhält man

Satz 10.28 (Cooley-Tukey Basis-2-Faktorisierung). *Sei $n = 2^p$, $p \in \mathbb{N}$. Dann ist*

$$F_n = A_p \cdot \ldots \cdot A_1 \cdot P_n$$

mit $P_2 = I_2$, $P_{2^{j+1}} = \begin{bmatrix} P_{2^j} & 0 \\ 0 & P_{2^j} \end{bmatrix} \cdot \Pi_{2^{j+1}}$, $j = 1, \ldots, p-1$ und

$$A_j = \operatorname{blockdiag}(B_{2^j}, \ldots, B_{2^j}) = \begin{bmatrix} B_{2^j} & & 0 \\ & \ddots & \\ 0 & & B_{2^j} \end{bmatrix}, \quad j = 1, \ldots, p.$$

Beispiel 10.29. Wir wollen die DFT des Vektors $f = [f_0, f_1, f_2, f_3]^T \in \mathbb{C}^4$ berechnen. Mit $\omega := \omega_4 = e^{-2\pi \mathrm{i}/4}$ folgt

$$\begin{bmatrix} g_0 \\ g_1 \\ g_2 \\ g_3 \end{bmatrix} = F_4 \begin{bmatrix} f_0 \\ f_1 \\ f_2 \\ f_3 \end{bmatrix} = B_4 \begin{bmatrix} F_2 & 0 \\ 0 & F_2 \end{bmatrix} \Pi_4 \begin{bmatrix} f_0 \\ f_1 \\ f_2 \\ f_3 \end{bmatrix} = B_4 \begin{bmatrix} F_2 & 0 \\ 0 & F_2 \end{bmatrix} \begin{bmatrix} f_0 \\ f_2 \\ f_1 \\ f_3 \end{bmatrix}.$$

Wegen $F_2 = B_2 \begin{bmatrix} F_1 & 0 \\ 0 & F_1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$ folgt

$$\begin{bmatrix} g_0 \\ g_1 \\ g_2 \\ g_3 \end{bmatrix} = B_4 \begin{bmatrix} f_0 + f_2 \\ f_0 - f_2 \\ f_1 + f_3 \\ f_1 - f_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & \omega \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -\omega \end{bmatrix} \begin{bmatrix} f_0 + f_2 \\ f_0 - f_2 \\ f_1 + f_3 \\ f_1 - f_3 \end{bmatrix} = \begin{bmatrix} f_0 + f_2 + f_1 + f_3 \\ f_0 - f_2 + \omega(f_1 - f_3) \\ f_0 + f_2 - (f_1 + f_3) \\ f_0 - f_2 - \omega(f_1 - f_3) \end{bmatrix}.$$

Der folgende Algorithmus wird mit $s = 1$ aufgerufen:

Algorithmus 10.30.

```
 Input:   unsigned n; complex f[n]; unsigned s;
 Output:  complex g[n];     // DFT von f

void FFT(n,f,g,s)
{
  if (n==1){
    g[0]=f[0];
  } else {
    // DFT von (f[0],f[2s],f[4s],...) -> (g[0],...,g[n/2-1])
    FFT(n/2,f,g,2s);
    // DFT von (f[s],f[3s],...) -> (g[n/2],...,g[n-1])
    FFT(n/2,f+s,g+n/2,2s);

    for (k=0; k<n/2; ++k){
      z:=g[k];
      w:=exp(-2*pi*i*k/n);
      x:=w*g[k+n/2];
      g[k]:=z+x;
      g[k+n/2]:=z-x;
    }
  }
}
```

Die Permutation beim Cooley-Tukey-Algorithmus

Wir versuchen, eine allgemeine Formel für die Wirkungsweise der Permutationen P_n zu finden. Dazu betrachten wir zunächst die Beispiele

$$\begin{aligned} P_8 x &= \begin{bmatrix} P_4 & \\ & P_4 \end{bmatrix} \Pi_8 x = \begin{bmatrix} \Pi_4 & \\ & \Pi_4 \end{bmatrix} \Pi_8 x = \begin{bmatrix} \Pi_4 & \\ & \Pi_4 \end{bmatrix} [x_0, x_2, x_4, x_6, x_1, x_3, x_5, x_7]^T \\ &= [x_0, x_4, x_2, x_6, x_1, x_5, x_3, x_7]^T \end{aligned}$$

und

$$P_{16} x = [x_0, x_8, x_4, x_{12}, x_2, x_{10}, x_6, x_{14}, x_1, x_9, x_5, x_{13}, x_3, x_{11}, x_7, x_{15}]^T.$$

Betrachtet man bei $P_{16}x$ das Element x_{12} an der Position 3 und x_7 an der Position 14, so stellt man für die Binärdarstellung der Indizes fest

$$\begin{array}{llll} \text{Index } 12 & = 1100 & \rightarrow & \text{Position } 3 \quad = 0011, \\ \text{Index } 7 & = 0111 & \rightarrow & \text{Position } 14 \quad = 1110. \end{array}$$

Die Binärdarstellung der Position entspricht also gerade der Spiegelung der Binärdarstellung des Index. Allgemein gilt

$$(P_n x)_k = x_{r(k)}, \quad k = 0, \ldots, n-1,$$

mit

$$r\left(\sum_{j=0}^{p-1} b_j 2^j\right) = \sum_{j=0}^{p-1} b_{p-1-j} 2^j \quad \text{für } b_j \in \{0,1\},\ j = 0, \ldots, p-1.$$

Bemerkung.

(a) Die Bitspiegelung des Eingabevektors kann einen signifikanten Beitrag zur Laufzeit haben, obwohl die asymptotische Komplexität $O(n)$ ist. Die Permutation des Eingabevektors wird daher nur bei "in-place" Implementierungen verwendet, d.h. im Vergleich zu Algorithmus 10.30 kommt man dabei ohne den Vektor g aus.

(b) Die Komplexität des Cooley-Tukey-Algorithmus ist $O(n \log n)$. Es bezeichnen a_p, m_p die Anzahl der Additionen bzw. Multiplikationen für $n = 2^p,\ p \in \mathbb{N}$. Dann gilt

$$\begin{aligned} a_{p+1} &= 2a_p + 2^{p+1}, \quad a_1 = 2, \\ m_{p+1} &= 2m_p + 2^p - 1, \quad m_1 = 0. \end{aligned}$$

Löst man die Rekursionsbeziehung auf, so erhält man (siehe auch das Master-Theorem Satz 4.13)

$$a_p = p \cdot 2^p = n \log_2 n \quad \text{und} \quad m_p = (p-2)2^{p-1} + 1 \leq \frac{1}{2} n \log_2 n.$$

(c) Für allgemeine n mit Primfaktorzerlegung $n = p_1 \cdot p_2 \cdot \ldots \cdot p_k$ ist eine Aufspaltung in k Stufen analog möglich (z.B. $n = 1000 = 2^3 5^3$).

Die reelle FFT

Wird F_n auf einen reellen Vektor $x \in \mathbb{R}^n$, $n = 2m$, angewendet, so kann $y := F_n x$ nach Satz 10.25 mit Hilfe der Sinus/Cosinus-Transformation berechnet werden kann. Man kann y aber auch mittels einer FFT halber Länge berechnen. Im Beweis zu Satz 10.25 haben wir bereits gesehen, dass $y_k = \overline{y}_{n-k}$, $k = 0, \ldots, n-1$ (dabei haben wir $y_n = y_0$ gesetzt). Es würde also genügen, die ersten $m+1$ Komponenten von y zu berechnen. Zunächst sehen wir, dass

$$y_m = \sum_{j=0}^{n-1} x_j \cos(\pi j) + \mathrm{i} \sum_{j=0}^{n-1} x_j \sin(\pi j) = \sum_{j=0}^{m-1} x_{2j} - x_{2j+1}.$$

Die anderen Komponenten $y_0, \ldots, y_{m-1}$ ergeben sich aus einer FFT halber Länge, wie das folgende Lemma besagt.

Lemma 10.31. *Definiere $u \in \mathbb{C}^m$ durch $u_j := x_{2j} + \mathrm{i} x_{2j+1}$, $j = 0, \ldots, m-1$, und $v := F_m u \in \mathbb{C}^m$. Dann gilt*

$$y_k = \frac{1}{2}(v_k + \overline{v}_{m-k}) + \frac{\omega_n^k}{2\mathrm{i}}(v_k - \overline{v}_{m-k}), \quad k = 0, \ldots, m-1.$$

Beweis. Wegen $\overline{\omega_m^{j(m-k)}} = \omega_m^{jk}$ gilt

$$\begin{aligned} v_k + \overline{v}_{m-k} &= \sum_{j=0}^{m-1} \omega_m^{jk}(x_{2j} + \mathrm{i} x_{2j+1}) + \overline{\omega_m^{j(m-k)}}(x_{2j} - \mathrm{i} x_{2j+1}) \\ &= 2 \sum_{j=0}^{m-1} \omega_m^{jk} x_{2j} = 2 \sum_{j=0}^{m-1} \omega_n^{2jk} x_{2j}. \end{aligned}$$

Entsprechend gilt

$$\frac{1}{2\mathrm{i}}(v_k - \overline{v}_{m-k}) = \sum_{j=0}^{m-1} \omega_n^{(2j+1)k} x_{2j+1}.$$

Insgesamt erhält man also (mit $v_m := v_0$)

$$y_k = \sum_{j=0}^{n-1} \omega_n^{jk} x_j = \frac{1}{2}(v_k + \overline{v}_{m-k}) + \frac{\omega_n^k}{2\mathrm{i}}(v_k - \overline{v}_{m-k}), \quad k = 0, \ldots, m-1.$$

□

Anwendung der FFT: Berechnung von Faltungsprodukten

Zu Vektoren $u = [u_0, \ldots, u_{n-1}]^T$ und $v = [v_0, \ldots, v_{n-1}]^T$, deren Komponenten bei Bedarf n-periodisch fortgesetzt werden, wird folgendes Produkt definiert:

$$z := u * v \iff z_k = \sum_{j=0}^{n-1} u_{k-j} v_j, \quad k = 0, \ldots, n-1.$$

Dieses **Faltungsprodukt** $*$ ist kommutativ und assoziativ. Die Faltung tritt häufig in der Bildverarbeitung und bei digitalen Filtern auf. Die Berechnung des Vektors z benötigt bei naiver Vorgehensweise wieder $O(n^2)$ Operationen. Das folgende Lemma zeigt aber einen Zusammenhang zur Fourier-Transformation auf.

Lemma 10.32. *Für $u, v \in \mathbb{C}^n$ und $z = u * v$ gilt*

$$(F_n z)_k = (F_n u)_k \cdot (F_n v)_k, \quad k = 0, \ldots, n-1.$$

Beweis. Wegen der Periodizität ist

$$\begin{aligned}
(F_n z)_k &= \sum_{j=0}^{n-1} z_j \omega_n^{jk} = \sum_{j=0}^{n-1} \sum_{\ell=0}^{n-1} u_{j-\ell} v_\ell \omega_n^{(j-\ell)k} \omega_n^{k\ell} \\
&= \sum_{\ell=0}^{n-1} v_\ell \omega_n^{k\ell} \sum_{j=-\ell}^{n-1-\ell} u_j \omega_n^{jk} = \left(\sum_{\ell=0}^{n-1} v_\ell \omega_n^{k\ell} \right) \left(\sum_{j=0}^{n-1} u_j \omega_n^{jk} \right) \\
&= (F_n u)_k \cdot (F_n v)_k.
\end{aligned}$$

□

Daher lässt sich die Berechnung einer Faltung auf drei Fourier-Transformationen (zwei Transformationen und eine Synthese) und eine komponentenweise Multiplikation zurückführen, was insgesamt mit $O(n \log n)$ Operationen durchführbar ist.

10.6 Splines

Wie in Abschnitt 10.2 anhand des Beispiels von Runge gesehen, eignen sich glatte Ansatzfunktionen wie Polynome nur bedingt zur Interpolation großer Datensätze. Als Ausweg bietet es sich an, Polynome niedrigeren Grades "aneinaderzusetzen".

Definition 10.33. *Sei $\triangle_n = \{x_0, \ldots, x_n\}$, $a = x_0 < x_1 < \ldots < x_n = b$, eine Zerlegung des Intervalls $[a, b]$. Eine Funktion $s : [a, b] \to \mathbb{R}$ heißt* ***Spline*** *vom Grad m zur Zerlegung $\triangle_n$, falls gilt*

(i) $s \in C^{m-1}[a, b]$,

(ii) $s|_{[x_j, x_{j+1}]} \in \Pi_m$, $0 \leq j < n$.

Den Raum solcher Funktionen bezeichnen wir mit $S_m(\triangle_n)$.

Bemerkung. Offenbar gilt $\Pi_m \subset S_m(\triangle_n)$.

Satz 10.34. *Es gilt* $\dim S_m(\triangle_n) = m + n$ *und* $\{p_0, \ldots, p_m, q_1, \ldots, q_{n-1}\}$ *ist eine Basis von $S_m(\triangle_n)$, wobei*

$$p_i(x) = (x - x_0)^i, \quad i = 0, \ldots, m,$$
$$q_j(x) = (x - x_j)_+^m := \begin{cases} (x - x_j)^m, & x \geq x_j, \\ 0, & x < x_j, \end{cases} \quad j = 1, \ldots, n-1.$$

Beweis. Offenbar gilt $p_i, q_j \in S_m(\triangle_n)$. Ist

$$s(x) = \sum_{i=0}^{m} a_i p_i(x) + \sum_{j=1}^{n-1} b_j q_j(x) = 0 \quad \text{für alle } x \in [a, b],$$

so folgt mit den linearen Funktionalen

$$G_k(f) := \frac{1}{m!} \left(f^{(m)}(x_k^+) - f^{(m)}(x_k^-) \right),$$

wobei $f^{(m)}(x^+)$ den rechtsseitigen Grenzwert von $f^{(m)}$ an der Stelle x bezeichnet, dass

$$0 = G_k(s) = \sum_{i=0}^{m} a_i \underbrace{G_k(p_i)}_{=0} + \sum_{j=1}^{n-1} b_j \underbrace{G_k(q_j)}_{=\delta_{jk}}, \quad k = 1, \ldots, n-1.$$

Also gilt $0 = \sum_{i=0}^{m} a_i p_i(x)$ für alle $x \in [a, b]$, woraus $a_i = 0$, $i = 0, \ldots, m$ folgt. Das System $\{p_0, \ldots, p_m, q_1, \ldots, q_{n-1}\}$ ist also linear unabhängig.

Wir müssen noch zeigen, dass jedes $s \in S_m(\triangle_n)$ durch $\{p_0, \ldots, p_n, q_1, \ldots, q_{m-1}\}$ darstellbar ist. Dazu betrachte

$$\tilde{s}(x) := \sum_{i=0}^{m} \frac{s^{(i)}(x_0)}{i!} p_i(x) + \sum_{j=1}^{n-1} G_j(s)\, q_j(x).$$

Dann ist $(\tilde{s}-s)|_{[x_j,x_{j+1}]} \in \Pi_m$, $0 \le j < n$. Ferner gilt

$$\tilde{s}^{(m)}(x_j^+) - \tilde{s}^{(m)}(x_j^-) = \sum_{i=1}^{n-1} G_i(s)\,(q_i^{(m)}(x_j^+) - q_i^{(m)}(x_j^-)).$$

Für $j < i$ und $j > i$ ist $q_i^{(m)}(x_j^+) - q_i^{(m)}(x_j^-) = 0$, weil dort q_i polynomial ist. Im Fall $i = j$ erhält man

$$q_i^{(m)}(x_i^+) - q_i^{(m)}(x_i^-) = q_i^{(m)}(x_i^+) = m!.$$

Daher folgt

$$\tilde{s}^{(m)}(x_j^+) - \tilde{s}^{(m)}(x_j^-) = m! \cdot G_j(s) = s^{(m)}(x_j^+) - s^{(m)}(x_j^-)$$

$$\iff (\tilde{s}-s)^{(m)}(x_j^+) = (\tilde{s}-s)^{(m)}(x_j^-)$$

und somit $\tilde{s} - s|_{[a,b]} \in \Pi_m$. Da $s^{(i)}(x_0) = \tilde{s}^{(i)}(x_0)$, $i = 0, \dots, m$, folgt $\tilde{s} - s = 0$. □

Will man das Interpolationsproblem

$$s(x_j) = y_j, \quad j = 0, \dots, n,$$

in $S_m(\triangle_n)$ lösen, so hat man $m + n$ freie Koeffizienten bei $n + 1$ Bedingungen. Der Spline wird also erst durch $m - 1$ zuätzliche Bedingungen eindeutig bestimmt. Ist $m = 2r + 1$, so können beispielsweise alternativ

(H) Hermite Bedingungen: $s^{(i)}(a) = y_{i,a}$, $s^{(i)}(b) = y_{i,b}$, $i = 1, \dots, r$, mit gegebenen Werten $y_{i,a}$, $y_{i,b}$,

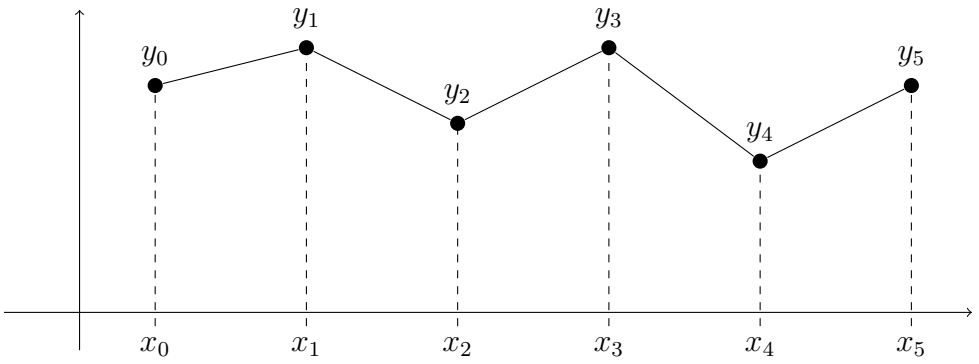
(N) natürliche Bedingungen $(r \le n)$**:** $s^{(i)}(a) = 0 = s^{(i)}(b)$, $i = r + 1, \dots, 2r$,

(P) periodische Bedingungen: $s^{(i)}(a) = s^{(i)}(b)$, $i = 1, \dots, 2r$,

gestellt werden.

Beispiel 10.35. Im Fall linearer Splines ($m = 1$) sind keine zusätzlichen Bedingungen erforderlich. Der Spline zu den Daten (x_j, y_j), $j = 0, \dots, n$, ist dann

$$s(x) = \frac{x_{j+1} - x}{x_{j+1} - x_j} y_j + \frac{x - x_j}{x_{j+1} - x_j} y_{j+1}, \quad x \in [x_j, x_{j+1}].$$



Satz 10.36. *Sei $m = 2r + 1$, $r \in \mathbb{N}$. Dann sind alle drei Interpolationsaufgaben (H), (N) und (P) eindeutig lösbar in $S_m(\triangle_n)$, und es gilt*

$$\int_a^b |s^{(r+1)}(x)|^2 \,\mathrm{d}x \le \int_a^b |g^{(r+1)}(x)|^2 \,\mathrm{d}x$$

für alle $g \in C^{r+1}[a,b]$, welches jeweils dieselbe (d.h. (H), (N), (P) mit denselben Daten) Interpolationsaufgabe wie $s \in S_m(\triangle_n)$ löst.

Bemerkung. Im Fall kubischer Splines ($m = 3$) hat man

$$\int_a^b |s''(x)|^2 \,\mathrm{d}x \le \int_a^b |g''(x)|^2 \,\mathrm{d}x.$$

Der letzte Ausdruck ist die sog. "linearisierte Biegungsenergie". Die Krümmung von g ist

$$\frac{g''(x)}{(1 + |g'(x)|^2)^{3/2}}.$$

"Linearisiert" bedeutet hier also $g'(x) \approx 0$. Kubische Splines werden daher als besonders "glatt" empfunden. Der Begriff "Spline" ist das englische Wort für eine lange dünne Latte im Schiffsbau, die an einzelnen Punkten fixiert und am Ende frei ist. Diese biegt sich annähernd wie ein kubischer Spline mit natürlichen Bedingungen (N).

Beweis.

(i) Wir zeigen: Sind $g \in C^{r+1}[a,b]$, $s \in S_m(\triangle_n)$ Funktionen, welche dasselbe Interpolationsproblem lösen, so gilt

$$0 \le \int_a^b |(g-s)^{(r+1)}(x)|^2 \,\mathrm{d}x = \int_a^b |g^{(r+1)}(x)|^2 \,\mathrm{d}x - \int_a^b |s^{(r+1)}(x)|^2 \,\mathrm{d}x. \tag{10.12}$$

Zunächst gilt

$$\begin{aligned}&\int_a^b |(g-s)^{(r+1)}(x)|^2 \,\mathrm{d}x \\ &\quad = \int_a^b |g^{(r+1)}(x)|^2 \,\mathrm{d}x - \int_a^b |s^{(r+1)}(x)|^2 \,\mathrm{d}x - 2\int_a^b (g-s)^{(r+1)}(x) s^{(r+1)}(x) \,\mathrm{d}x.\end{aligned}$$

Partielle Integration liefert

$$\begin{aligned}&\int_a^b (g-s)^{(r+1)}(x) s^{(r+1)} \,\mathrm{d}x \\ &= (g-s)^{(r)}(x) s^{(r+1)}(x)\Big|_a^b - \int_a^b (g-s)^{(r)}(x) s^{(r+2)}(x) \,\mathrm{d}x \\ &= \dots \\ &= \sum_{i=0}^{r-1} (-1)^i (g-s)^{(r-i)}(x) s^{(r+1+i)}(x)\Big|_a^b + (-1)^r \int_a^b (g-s)'(x) s^{(2r+1)}(x) \,\mathrm{d}x.\end{aligned}$$

Hierbei ist das letzte Integral nur stückweise definiert. Da $m = 2r + 1$, ist

$$s^{(2r+1)}\big|_{[x_j, x_{j+1}]} =: \alpha_j$$

konstant. Also folgt wegen $g(x_j) = s(x_j)$, $j = 0, \dots, n-1$, dass

$$\int_a^b (g-s)'(x) s^{(2r+1)}(x)\,\mathrm{d}x = \sum_{j=0}^{n-1} \alpha_j (g(x) - s(x))\Big|_{x_j}^{x_{j+1}} = 0.$$

Somit gilt

$$\int_a^b (g-s)^{(r+1)}(x) s^{(r+1)}(x)\,\mathrm{d}x = \sum_{i=0}^{r-1} (-1)^i (g-s)^{(r-i)}(x) s^{(r+1+i)}(x)\Big|_a^b = 0$$

bei jeder der drei Bedingungen (H), (N) und (P).

(ii) Eindeutigkeit der Interpolation: Seien $s, \tilde{s} \in S_m(\triangle_n)$ zwei Lösungen desselben Interpolationsproblems. Dann gilt nach (i)

$$\int_a^b |(s-\tilde{s})^{(r+1)}(x)|^2\,\mathrm{d}x = 0$$

und somit $(s-\tilde{s})^{(r+1)} = 0$, was $\rho := s - \tilde{s} \in \Pi_r$ beweist. Im Fall

(H) folgt $\rho = 0$ aus $\rho^{(i)}(a) = 0$, $i = 0, \dots, r$,

(N) folgt $\rho = 0$ aus $\rho(x_j) = 0$, $0 \le j \le n$, weil wir hier angenommen haben, dass $r \le n$,

(P) folgt, dass $\rho^{(r-1)}$ linear ist, weil $\rho \in \Pi_r$. Mit $\rho^{(r-1)}(a) = \rho^{(r-1)}(b)$ sieht man, dass $\rho^{(r-1)}$ konstant und somit $\rho \in \Pi_{r-1}$ ist. Induktiv erhält man, dass $\rho \in \Pi_1$. Aus $\rho(a) = \rho(b) = 0$ folgt $\rho = 0$.

(iii) Existenz der Interpolation: Dies folgt aus der Injektivität des Interpolationsproblems und der Tatsache, dass die Anzahl der Bedingungen mit der Dimension von $S_m(\triangle_n)$ übereinstimmt.

□

Bemerkung. Aus dem letzten Beweis sehen wir, dass

$$\int_a^b |s^{(r+1)}(x)|^2\,\mathrm{d}x \le \int_a^b |g^{(r+1)}(x)|^2\,\mathrm{d}x \tag{10.13}$$

sogar für jedes $g \in C^{r+1}[a,b]$ mit $g(x_i) = s(x_i)$, falls $s \in S_m(\triangle_n)$ die Bedingung (N) erfüllt. Das Minimum s von (10.13) erfüllt also automatisch (N). Dies erklärt die Bezeichnung "Natürliche Bedingungen".

Bell-Splines

Die in Satz 10.34 vorgestellte Basis ist nicht lokal (d.h. viele Basisfunktionen sind $\neq 0$ an jedem Punkt) und eignet sich auch wegen ihrer Konditionierung nur schlecht zur Berechnung eines interpolierenden Splines. Im Folgenden führen wir eine andere Basis, die so genannten *Bell-Splines*, mit deutlich besseren numerischen Eigenschaften ein.

Definition 10.37. *Sei $t_0 \leq \ldots \leq t_n$ eine beliebige Knotenfolge. Die* ***Bell-Splines*** *(B-Splines) B_{ik} der Ordnung $k = 0, \ldots, n-1$, $i = 0, \ldots, n-1-k$, sind rekursiv definiert durch*

$$B_{i0}(t) := \begin{cases} 1, & \text{falls } t_i \leq t < t_{i+1}, \\ 0, & \text{sonst}, \end{cases}$$

und

$$B_{ik} = \omega_{ik} B_{i,k-1} + (1 - \omega_{i+1,k}) B_{i+1,k-1}$$

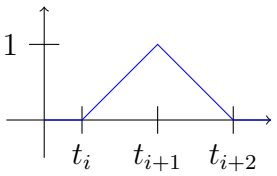
mit

$$\omega_{ik}(t) := \begin{cases} \frac{t-t_i}{t_{i+k}-t_i}, & \text{falls } t_i < t_{i+k}, \\ 0, & \text{sonst}. \end{cases}$$

Beispiel 10.38.

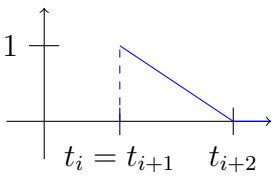
(a) Lineare B-Splines, sog. **Hütchenfunktionen**:
Falls $t_i < t_{i+1} < t_{i+2}$ gilt

$$\begin{aligned} B_{i1}(t) &= \frac{t-t_i}{t_{i+1}-t_i} B_{i0}(t) + \left(1 - \frac{t-t_{i+1}}{t_{i+2}-t_{i+1}}\right) B_{i+1,0}(t) \\ &= \begin{cases} \frac{t-t_i}{t_{i+1}-t_i}, & \text{falls } t \in [t_i, t_{i+1}], \\ \frac{t_{i+2}-t}{t_{i+2}-t_{i+1}}, & \text{falls } t \in [t_{i+1}, t_{i+2}], \\ 0, & \text{sonst}. \end{cases} \end{aligned}$$

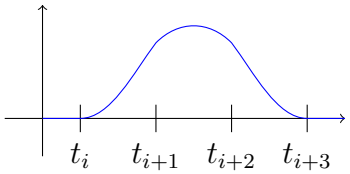


Im Fall $t_i = t_{i+1} < t_{i+2}$

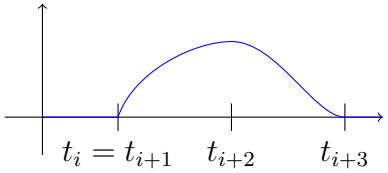
$$B_{i1}(t) = \left(1 - \frac{t-t_{i+1}}{t_{i+2}-t_{i+1}}\right) B_{i+1,0}(t) = \begin{cases} \frac{t_{i+2}-t}{t_{i+2}-t_{i+1}}, & \text{falls } t \in [t_{i+1}, t_{i+2}], \\ 0, & \text{sonst}. \end{cases}$$



(b) Quadratische B-Splines:
Falls $t_i < t_{i+1} < t_{i+2} < t_{i+3}$, ist B_{i2} an den Knoten stetig differenzierbar.



Falls $t_i = t_{i+1} < t_{t+2} < t_{i+3}$, ist B_{i2} in t_{i+2}, t_{i+3} stetig differenzierbar, aber in $t_i = t_{i+1}$ nur stetig.



Bemerkung. Man kann zeigen, dass B_{ik} an ℓ-fachen Knoten t_i, d.h. $t_{i-1} < t_i = \ldots = t_{t+\ell-1} < t_{i+\ell}$, $(k-\ell)$-fach stetig differenzierbar ist; siehe z.B. Deuflhard/Hohmann, *Numerische Mathematik 1.*

Lemma 10.39. *Es gilt*

(i) $B_{ik}|_{[t_j,t_{j+1}]} \in \Pi_k$ *für alle* $j = 0, \ldots, n-1$ *und* $i = 0, \ldots, n-1-k,\ k = 0, \ldots, n-1,$

(ii) für den Träger (engl. support) von B_{ik} *gilt*

$$\operatorname{supp}(B_{ik}) := \overline{\{t \in \mathbb{R} : B_{ik}(t) \neq 0\}} \subset [t_i, t_{i+k+1}],$$

d.h. B_{ik} *hat* ***lokalen Träger****,*

(iii) $B_{ik} \geq 0,\ i = 0, \ldots, n-1-k,$ *und* $\sum_{i=0}^{n-1-k} B_{ik}(t) = 1$ *für* $t \in [t_k, t_{n-k}]$ *und alle* $k = 0, \ldots, n-1$, *d.h.* $\{B_{ik},\ i = 0, \ldots, n-1-k\}$ *bildet eine* ***Zerlegung der Eins*** *auf* $[t_k, t_{n-k}]$.

Beweis. (i) und (ii) folgen direkt aus der Definition.
(iii): $B_{ik} \geq 0$ und $\sum_{i=0}^{n-1} B_{i0} = 1$ auf $[t_0, t_n]$ sind klar. Die Behauptung folgt per Induktion aus

$$\begin{aligned}
\sum_{i=0}^{n-1-k} B_{ik} &= \sum_{i=0}^{n-1-k} \omega_{ik} B_{i,k-1} + (1 - \omega_{i+1,k}) B_{i+1,k-1} \\
&= \sum_{i=0}^{n-1-k} B_{i+1,k-1} + \sum_{i=0}^{n-1-k} \omega_{ik} B_{i,k-1} - \omega_{i+1,k} B_{i+1,k-1} \\
&= \sum_{i=0}^{n-k} B_{i,k-1} - B_{0,k-1} + \omega_{0k} B_{0,k-1} - \omega_{n-k,k} B_{n-k,k-1} \\
&= \sum_{i=0}^{n-k} B_{i,k-1},
\end{aligned}$$

weil die letzten drei Summanden für $t \in [t_k, t_{n-k}]$, $k > 0$, verschwinden. □

Wir kehren nun zum Raum $S_m(\triangle_n)$ mit der Zerlegung $\triangle_n : a = x_0 < x_1 < \ldots < x_n = b$ zurück. Zur Konstruktion der B-Spline-Basis seien Knoten $t_0, \ldots, t_{n+2m}$ wie folgt definiert:

$$\begin{array}{ccccccccc}
 & & x_0 & < & x_1 & < \ldots < & x_n & & \\
 & & \| & & \| & & \| & & \\
t_0 \leq \ldots \leq t_{m-1} \leq t_m & < & t_{m+1} & < & \ldots & < & t_{m+n} & \leq t_{m+n+1} \leq \ldots \leq t_{n+2m}
\end{array} \tag{10.14}$$

Die Knoten $t_0, \dots, t_{m-1}$ und $t_{m+n+1}, \dots, t_{n+2m}$ können frei gewählt werden.

Im Folgenden zeigen wir, dass mit Hilfe der B-Splines eine Basis von $S_m(\triangle_n)$ konstruiert werden kann. Dazu benötigen wir die folgende **Marsden-Identität**.

Lemma 10.40. *Mit obigen Bezeichnungen gilt für alle $t \in [t_m, t_{m+n}]$ und $s \in \mathbb{R}$, dass*

$$(t-s)^m = \sum_{i=0}^{m+n-1} \varphi_{im}(s) B_{im}(t),$$

wobei

$$\varphi_{im}(s) := \prod_{j=1}^{m} (t_{i+j} - s).$$

Beweis. Der Fall $m = 0$ folgt aus Lemma 10.39 (iii). Sei die Aussage für alle $k < m$ bewiesen. Dann gilt wegen $B_{0,m-1}(t) = 0 = B_{0,m+n,m}(t)$ für $t \in [t_m, t_{m+n}]$, falls $t_i < t_{i+m}$, dass

$$\begin{aligned}
\sum_{i=0}^{m+n-1} \varphi_{im}(s) B_{im}(t) &= \sum_{i=1}^{m+n-1} \left[\frac{t-t_i}{t_{i+m}-t_i} \varphi_{im}(s) + \frac{t_{i+m}-t}{t_{i+m}-t_i} \varphi_{i-1,m}(s) \right] B_{i,m-1}(t) \\
&= \sum_{i=1}^{m+n-1} \prod_{j=1}^{m-1} (t_{i+j} - s) \underbrace{\left[\frac{t-t_i}{t_{i+m}-t_i} (t_{i+m} - s) + \frac{t_{i+m}-t}{t_{i+m}-t_i} (t_i - s) \right]}_{t-s} B_{i,m-1}(t) \\
&= (t-s) \sum_{i=1}^{m+n-1} \varphi_{i,m-1}(s) B_{i,m-1}(t) \\
&= (t-s)(t-s)^{m-1} = (t-s)^m.
\end{aligned}$$

Die vorletzte Identität erhält man durch die Anwendung der Induktionsvoraussetzung auf die Knoten $t_1, \dots, t_{n+2m}$ und das Intervall $[t_{m-1+1}, t_{m+n+1-1}] = [t_m, t_{m+n}]$. Der Fall $t_i = t_{i+m}$ hat wegen $B_{i,m-1} = 0$ keinen Beitrag zur Summe. □

Im folgenden Lemma zeigen wir die lineare Unabhängigkeit des Systems $\{B_{im},\ i = 0, \dots, m+n-1\}$. Es gilt sogar eine lokale Unabhängigkeit.

Lemma 10.41. *Die B-Splines B_{im}, $0 \le i < m+n$, sind lokal linear unabhängig, d.h. aus*

$$\sum_{i=0}^{m+n-1} \alpha_i B_{im}(t) = 0 \quad \text{für alle } t \in (c,d) \subset [a,b]$$

und $(c,d) \cap (t_j, t_{j+m+1}) \neq \emptyset$ für ein j folgt $\alpha_j = 0$.

Beweis. Wir dürfen annehmen, dass (c,d) keine Knoten enthält, sonst zerlegen wir (c,d) in Teilintervalle. Lemma 10.40 liefert für $\ell \le m$

$$t^\ell = (-1)^{m-\ell} \frac{\ell!}{m!} \left[(t-s)^m \right]^{(m-\ell)} \Big|_{s=0} = (-1)^{m-\ell} \frac{\ell!}{m!} \sum_{i=0}^{m+n-1} \varphi_{im}^{(m-\ell)}(0) B_{im}(t).$$

Daher lässt sich jedes $p \in \Pi_m$ auf (c,d) durch die B-Splines B_{im}, $0 \le i < m+n$, darstellen. Auf (c,d) sind aber nur $m+1 = \dim \Pi_m$ B-Splines von Null verschieden. Daher müssen diese B-Splines linear unabhängig sein. □

Satz 10.42. *Es bezeichnen B_{im}, $0 \le i < m+n$, die B-Splines zu den Knoten $t_0, \dots, t_{n+2m}$. Dann gilt*

$$S_m(\triangle_n) = \operatorname{span}\{B_{im},\ 0 \le i < m+n\}.$$

Ferner gilt

$$\sum_{i=0}^{m+n-1} B_{im}(x) = 1 \quad \text{für alle } x \in [a,b].$$

Beweis. Nach Lemma 10.39 (i) und der diesem Lemma vorangehende Bemerkung gilt $B_{im} \in S_m(\triangle_n)$, $0 \le i < m+n$. Nach Lemma 10.41 ist dieses Funktionensystem linear unabhängig und somit eine Basis von $S_m(\triangle_n)$. Der zweite Teil der Behauptung ist Lemma 10.39 (iii). □

Berechnung interpolierender kubischer Splines

Wir wollen die Lösung $s \in S_3(\triangle_n)$ des Interpolationsproblems $s(x_i) = y_i$, $i = 0, \dots, n$, mit einer der Bedingungen (H), (N) oder (P) unter Verwendung von kubischer B-Splines berechnen. Dazu seien die Knoten $t_0, \dots, t_{n+6}$ wie in (10.14) eingeführt. Nach Satz 10.42 existieren Koeffizienten α_i, $i = 0, \dots, n+2$, die sog. **de Boor-Punkte**, so dass

$$s(x) = \sum_{i=0}^{n+2} \alpha_i B_{i3}(x).$$

Diese Koeffizienten werden aus den $n+3$ Gleichungen

$$\sum_{i=0}^{n+2} \alpha_i B_{i3}(x_j) = y_j, \quad j = 0, \dots, n,$$

und im Fall

$$\text{(H)} \quad \sum_{i=0}^{n+2} \alpha_i B'_{i3}(a) = y_{1,a}, \quad \sum_{i=0}^{n+2} \alpha_i B'_{i3}(b) = y_{1,b},$$

$$\text{(N)} \quad \sum_{i=0}^{n+2} \alpha_i B''_{i3}(a) = 0, \quad \sum_{i=0}^{n+2} \alpha_i B''_{i3}(b) = 0,$$

$$\text{(P)} \quad \sum_{i=0}^{n+2} \alpha_i B'_{i3}(a) = \sum_{i=0}^{n+2} \alpha_i B'_{i3}(b), \quad \sum_{i=0}^{n+2} \alpha_i B''_{i3}(a) = \sum_{i=0}^{n+2} \alpha_i B''_{i3}(b)$$

bestimmt. Im Fall (H) ist also das lineare Gleichungssystem

$$\begin{bmatrix} B'_{03}(a) & \cdots & B'_{n+2,3}(a) \\ B_{03}(x_0) & \cdots & B_{n+2,3}(x_0) \\ \vdots & & \vdots \\ B_{03}(x_n) & \dots & B_{n+2,3}(x_n) \\ B'_{03}(b) & \dots & B'_{n+2,3}(b) \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \vdots \\ \vdots \\ \vdots \\ \alpha_{n+2} \end{bmatrix} = \begin{bmatrix} y_{1,a} \\ y_0 \\ \vdots \\ y_n \\ y_{1,b} \end{bmatrix}$$

zu lösen.

Bemerkung.

(a) Man beachte, dass wegen der lokalen Träger-Eigenschaft jede Zeile in obiger Koeffizientenmatrix höchstens $m = 3$ Einträge besitzt, die nicht verschwinden.

(b) Die im Gleichungssystem auftauchenden Ableitungen der B-Splines können mit Hilfe einer Aussage, die wir in den Übungsaufgaben zeigen werden, effizient berechnet werden.

(c) Die Auswertung von $s(x) = \sum_{i=0}^{n+2} \alpha_i B_{i3}(x)$ kann mit Hilfe des Algorithmus von de Boor effizient durchgeführt werden; vgl. auch hierzu die Übungsaufgaben.

Abschätzung des Interpolationsfehlers

Sei $\triangle_n : a = x_0 < \ldots < x_n = b$ eine Zerlegung und

$$h := \max_{j=0,\ldots,n-1} x_{j+1} - x_j$$

die **Gitterweite**. Wir schätzen den Interpolationsfehler für lineare und kubische Splines in der L^2-Norm

$$\|f\|^2_{L^2[a,b]} = \int_a^b |f(x)|^2 \,\mathrm{d}x$$

ab.

Satz 10.43. *Sei $f \in C^2[a,b]$ und $s \in S_1(\triangle_n)$ der eindeutig bestimmter lineare interpolierende Spline, d.h. $s(x_j) = f(x_j)$, $j = 0, \ldots, n$. Dann gilt*

$$\|f - s\|_{L^2[a,b]} \leq \frac{h^2}{2} \|f''\|_{L^2[a,b]}$$

und

$$\|(f - s)'\|_{L^2[a,b]} \leq \frac{h}{\sqrt{2}} \|f''\|_{L^2[a,b]}.$$

Beweis. Die Funktion $e := f - s$ besitzt die Nullstellen $x_0, \ldots, x_n$. Daher gilt nach der Cauchy-Schwarzschen Ungleichung die folgende **Poincaré-Ungleichung**

$$\begin{aligned}
\int_{x_j}^{x_{j+1}} |e(x)|^2 \,\mathrm{d}x &= \int_{x_j}^{x_{j+1}} \left| \int_{x_j}^{x} 1 \cdot e'(t) \,\mathrm{d}t \right|^2 \mathrm{d}x \\
&\leq \int_{x_j}^{x_{j+1}} \left(\int_{x_j}^{x} 1 \,\mathrm{d}t \right) \left(\int_{x_j}^{x} |e'(t)|^2 \,\mathrm{d}t \right) \mathrm{d}x \\
&\leq \int_{x_j}^{x_{j+1}} (x - x_j) \int_{x_j}^{x_{j+1}} |e'(t)|^2 \,\mathrm{d}t \,\mathrm{d}x \\
&\leq \frac{h^2}{2} \int_{x_j}^{x_{j+1}} |e'(t)|^2 \,\mathrm{d}t.
\end{aligned}$$

Durch Summation über j erhalten wir

$$\|f-s\|_{L^2[a,b]} \le \frac{h}{\sqrt{2}}\|(f-s)'\|_{L^2[a,b]}. \tag{10.15}$$

Partielle Integration liefert

$$\begin{aligned}\|(f-s)'\|^2_{L^2[a,b]} &= \sum_{j=0}^{n-1}(f-s)'(x)\underbrace{\overline{(f-s)(x)}}_{=0}\Big|_{x_j}^{x_{j+1}} - \int_{x_j}^{x_{j+1}}(f-s)''(x)\overline{(f-s)(x)}\,\mathrm{d}x\\ &= -\sum_{j=0}^{n-1}\int_{x_j}^{x_{j+1}} f''(x)\overline{(f-s)}(x)\,\mathrm{d}x.\end{aligned}$$

Nach der Cauchy-Schwarzschen Ungleichung und (10.15) folgt hieraus

$$\|(f-s)'\|^2_{L^2[a,b]} \le \|f-s\|_{L^2[a,b]}\|f''\|_{L^2[a,b]} \le \frac{h}{\sqrt{2}}\|(f-s)'\|_{L^2[a,b]}\|f''\|_{L^2[a,b]}.$$

Division durch $\|(f-s)'\|_{L^2[a,b]}$ und (10.15) liefern die Behauptung. □

Satz 10.44. *Sei $f \in C^4[a,b]$ und $s \in S_3(\triangle_n)$ bezeichne den eindeutig bestimmten kubischen interpolierenden Spline, d.h. $s(x_j) = f(x_j)$, $j = 0,\dots,n$, mit einer der Bedingungen (H), (N) oder (P). Dann gilt*

$$\|f-s\|_{L^2[a,b]} \le \frac{h^4}{4}\|f^{(4)}\|_{L^2[a,b]}.$$

Beweis. Es bezeichne $\mathcal{I}_1 : C[a,b] \to S_1(\triangle_n)$ den Interpolationsprojektor definiert durch $f \mapsto v$, wobei v den interpolierenden linearer Spline bezeichnet, und $\mathcal{I}_3 : C[a,b] \to S_3(\triangle_n)$ sei der entsprechende Projektor für kubische Splines. Wegen $\mathcal{I}_1(f-\mathcal{I}_3 f) = 0$ folgt nach Satz 10.43

$$\|f-\mathcal{I}_3 f\|_{L^2[a,b]} = \|f-\mathcal{I}_3 f-\mathcal{I}_1(f-\mathcal{I}_3 f)\|_{L^2[a,b]} \le \frac{h^2}{2}\|f''-(\mathcal{I}_3 f)''\|_{L^2[a,b]}. \tag{10.16}$$

Sei $s \in S_3(\triangle_n)$ mit $s'' = \mathcal{I}_1(f'')$ und s erfülle die geforderte Bedingung (H), (N) bzw. (P). Mit $e := f-s$ gilt wegen $\mathcal{I}_3 s = s$

$$\begin{aligned}\|f''-(\mathcal{I}_3 f)''\|^2_{L^2[a,b]} &= \|f''-s''-(\mathcal{I}_3 f)''+s''\| = \|f''-s''-[\mathcal{I}_3(f-s)]''\|^2_{L^2[a,b]}\\ &\le \|e''-(\mathcal{I}_3 e)''\|^2_{L^2[a,b]} + \|(\mathcal{I}_3 e)''\|^2_{L^2[a,b]} \overset{(10.12)}{=} \|e''\|^2_{L^2[a,b]}\\ &= \|f''-s''\|^2_{L^2[a,b]} = \|f''-\mathcal{I}_1(f'')\|^2_{L^2[a,b]}.\end{aligned}$$

Nochmalige Anwendung von Satz 10.43 ergibt

$$\|f''-(\mathcal{I}_3 f)''\|_{L^2[a,b]} \le \|f''-\mathcal{I}_1(f'')\|_{L^2[a,b]} \le \frac{h^2}{2}\|f^{(4)}\|_{L^2[a,b]}.$$

Die Behauptung folgt aus (10.16) □

11 Numerische Integration

Ziel dieses Kapitels ist die numerische Approximation von Integralen

$$I(f) := \int_a^b f(x)\,\mathrm{d}x,$$

die nicht in geschlossener Form durch Aufgabe einer Stammfunktionen integriert werden können.

Definition 11.1. *Eine Abbildung $Q_n : C[a,b] \to \mathbb{R}$ der Form*

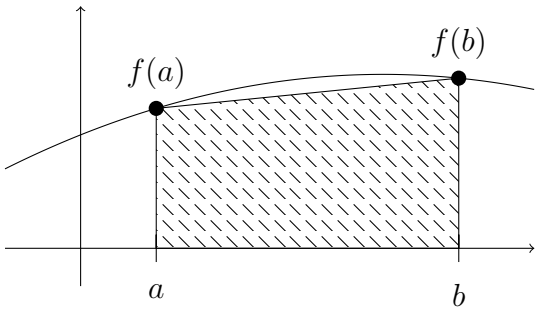
$$Q_n(f) = \sum_{j=0}^{n} w_j f(x_j)$$

mit Stützstellen $a \le x_0 < x_1 < \ldots < x_n \le b$ und Gewichten $w_0, \ldots, w_n \in \mathbb{R}$ heißt ***Quadraturformel****.*

Beispiel 11.2.

(a) **Mittelpunkt-Regel** $Q_0(f) = (b-a) f(\frac{a+b}{2})$

(b) **Trapez-Regel** $Q_1(f) = \frac{b-a}{2}(f(a) + f(b))$



(c) **Simpson-Regel** $Q_2(f) = \frac{b-a}{6}(f(a) + 4f(\frac{a+b}{2}) + f(b))$

Als Maß für die Qualität einer Quadraturformel führen wir den folgenden Begriff ein.

Definition 11.3. *Eine Quadraturformel Q_n hat* ***Exaktheitsgrad*** *k, falls*

$$Q_n(p) = I(p) \tag{11.1}$$

für alle $p \in \Pi_k$ gilt.

Bemerkung.

(a) Zur Bestimmung des Exaktheitsgrades genügt es, die Bedingung (11.1) für eine Basis von Π_k zu überprüfen, weil sowohl I als auch Q_n lineare Abbildungen sind.

(b) Weil $1 \in \Pi_0$, gilt

$$\sum_{j=0}^{n} w_j = \sum_{j=0}^{n} w_j \underbrace{f(x_j)}_{=1} = \int_a^b 1 \,\mathrm{d}x = b - a.$$

Beispiel 11.4. Die Mittelpunkt-Regel und Trapez-Regel haben Exaktheitsgrad $k = 1$, weil

$$Q_0(1) = Q_1(1) = b - a = I(1) \quad \text{und} \quad Q_0(x) = Q_1(x) = \frac{b-a}{2}(a+b) = \frac{1}{2}(b^2 - a^2) = I(x).$$

Die Simpson-Regel hat Exaktheitsgrad $k = 2$, weil

$$Q_2(1) = b - a = I(1), \quad Q_2(x) = \frac{b-a}{6}(a + 2(a+b) + b) = \frac{1}{2}(b^2 - a^2) = I(x)$$

und

$$Q_2(x^2) = \frac{b-a}{6}(a^2 + (a+b)^2 + b^2) = \frac{b-a}{3}(a^2 + ab + b^2) = \frac{1}{3}(b^3 - a^3) = I(x^2).$$

11.1 Newton-Côtes-Formeln

Mit Hilfe der Interpolation lässt sich eine Unterklasse der Quadraturformeln konstruieren. Dazu seien L_j, $j = 0, \ldots, n$, die Lagrange-Basispolynome (siehe Definition 10.1) zu den Stützstellen $x_0 < x_1 < \ldots < x_n$.

Definition 11.5. *Eine Quadraturformel Q_n zur Zerlegung $a \le x_0 < x_1 < \ldots < x_n \le b$ mit Gewichten*

$$w_j = \int_a^b L_j(x) \,\mathrm{d}x$$

heißt ***Interpolationsquadraturformel****. Im Fall äquidistanter Knoten*

$$x_j = a + \frac{b-a}{n} j \quad \text{oder} \quad x_j = a + \frac{b-a}{2(n+1)}(2j+1), \quad j = 0, \ldots, n,$$

spricht man von ***geschlossenen*** *bzw.* ***offenen Newton-Côtes-Formeln****.*

Satz 11.6. *Eine Quadraturformel Q_n ist genau dann eine Interpolationsquadraturformel, wenn sie Exaktheitsgrad $k = n$ hat. Insbesondere existiert genau eine Quadraturformel Q_n vom Exaktheitsgrad n.*

Beweis. Sei Q_n eine Interpolationsquadraturformel und $p \in \Pi_n$. Weil p sich selbst interpoliert, gilt

$$p(x) = \sum_{j=0}^{n} p(x_j) L_j(x).$$

Hieraus folgt

$$I(p) = \int_a^b p(x)\,\mathrm{d}x = \int_a^b \sum_{j=0}^n p(x_j)L_j(x)\,\mathrm{d}x = \sum_{j=0}^n p(x_j)\int_a^b L_j(x)\,\mathrm{d}x$$
$$= \sum_{j=0}^n w_j p(x_j) = Q_n(p).$$

Hat umgekehrt Q_n Exaktheitsgrad n, so gilt inbesondere für $L_j \in \Pi_n$, $j = 0, \dots, n$, dass

$$Q_n(L_j) = I(L_j), \quad j = 0, \dots, n.$$

Wegen

$$Q_n(L_j) = \sum_{i=0}^n w_i \underbrace{L_j(x_i)}_{=\delta_{ij}} = w_j$$

folgt $w_j = Q_n(L_j) = I(L_j) = \int_a^b L_j(x)\,\mathrm{d}x$. Daher ist Q_n eine Interpolationsquadraturformel. □

Bemerkung. Für Interpolationsquadraturformeln Q_n gilt

$$Q_n(f) = \sum_{j=0}^n w_j f(x_j) = \sum_{j=0}^n f(x_j)\int_a^b L_j(x)\,\mathrm{d}x = \int_a^b \sum_{j=0}^n f(x_j)L_j(x)\,\mathrm{d}x.$$

Das Polynom $p_{0,n} := \sum_{j=0}^n f(x_j)L_j \in \Pi_n$ ist das Interpolationspolynom zu $(x_j, f(x_j))$, $j = 0, \dots, n$. Die Approximation von $I(f)$ erfolgt bei Interpolationsquadraturformeln also durch Integration des Interpolationspolynoms zu f.

Beispiel 11.7. Wir wollen die Newton-Côtes-Formeln für $n = 1, 2$ berechnen.

(a) Im Fall $n = 1$ ist $x_0 = a$, $x_1 = b$. Dann folgt

$$w_0 = \int_a^b L_0(x)\,\mathrm{d}x = \int_a^b \frac{x-b}{a-b}\,\mathrm{d}x = \frac{1}{2}\frac{(x-b)^2}{a-b}\Big|_a^b = \frac{1}{2}(b-a)$$

und $w_1 = \frac{1}{2}(b-a)$ wegen der Bemerkung nach Definition 11.3. Wir erhalten also die Trapez-Regel.

(b) Ebenso leicht rechnet man nach, dass man für $n = 2$ die Simpson-Regel erhält.

(c) In folgender Tabelle sind die Gewichte der Newton-Côtes-Formeln bis $n = 6$ zusammengefasst.

n	$w_j/(b-a)$	Name
1	$\frac{1}{2}, \frac{1}{2}$	Trapez-Regel
2	$\frac{1}{6}, \frac{4}{6}, \frac{1}{6}$	Simpson-Regel
3	$\frac{1}{8}, \frac{3}{8}, \frac{3}{8}, \frac{1}{8}$	Newtons 3/8-Regel
4	$\frac{7}{90}, \frac{32}{90}, \frac{12}{90}, \frac{32}{90}, \frac{7}{90}$	Milne-Regel
5	$\frac{19}{288}, \frac{75}{288}, \frac{50}{288}, \frac{50}{288}, \frac{75}{288}, \frac{19}{288}$	–
6	$\frac{41}{840}, \frac{216}{840}, \frac{27}{840}, \frac{272}{840}, \frac{27}{840}, \frac{216}{840}, \frac{41}{840}$	Weddle-Regel

Unter folgenden Symmetrieannahmen haben Interpolationsquadraturformeln sogar noch einen etwas höheren Exaktheitsgrad.

Satz 11.8. *Sind die Stützstellen symmetrisch, d.h. gilt $x_j - a = b - x_{n-j}$, $j = 0, \dots, n$, so gilt für die Interpolationsquadraturformel Q_n*

(i) Q_n ist symmetrisch, d.h. $w_{n-j} = w_j$, $j = 0, \dots, n$.

(ii) Ist n gerade, so ist Q_n exakt auf Π_{n+1}.

Beweis.

(i) Sei $\tilde{Q}_n(f) := \sum_{j=0}^{n} w_{n-j} f(x_j)$ eine Quadraturformel. Wir betrachten die Basispolynome $p_i(x) := (x - \frac{a+b}{2})^i$, $i = 0, \dots, n$, von Π_n. Sei $\tilde{p}_i(x) := p_i(a + b - x)$, $i = 0, \dots, n$. Dann gilt

$$\begin{aligned}\tilde{Q}_n(p_i) &= \sum_{j=0}^{n} w_{n-j} p_i(x_j) = \sum_{j=0}^{n} w_{n-j} \tilde{p}_i(x_{n-j}) = Q_n(\tilde{p}_i) = I(\tilde{p}_i) \\ &= \int_a^b \tilde{p}_i(x)\,\mathrm{d}x = \int_a^b p_i(a+b-x)\,\mathrm{d}x = \int_a^b p_i(x)\,\mathrm{d}x = I(p_i).\end{aligned}$$

Also hat $\tilde{Q}_n$ Exaktheitsgrad $k = n$. Wegen der Eindeutigkeit (siehe Satz 11.6) folgt $\tilde{Q}_n = Q_n$ und somit $w_j = w_{n-j}$, $j = 0, \dots, n$.

(ii) Ist $n = 2m$, $m \in \mathbb{N}$, so gilt wegen $x_m = \frac{1}{2}(a+b)$ und $w_{m+j} = w_{n-m-j} = w_{m-j}$.

$$\begin{aligned}Q_n(p_{n+1}) &= \sum_{j=0}^{n} w_j \left(x_j - \frac{a+b}{2}\right)^{n+1} \\ &= \sum_{j=1}^{m} w_{m-j} \left(x_{m-j} - \frac{a+b}{2}\right)^{n+1} + \sum_{j=1}^{m} w_{m+j} \left(x_{m+j} - \frac{a+b}{2}\right)^{n+1} \\ &= \sum_{j=1}^{m} w_{m-j} \left[\underbrace{\left(x_{m-j} - \frac{a+b}{2}\right)^{n+1} + \left(\frac{a+b}{2} - x_{m-j}\right)^{n+1}}_{=0} \right] = 0.\end{aligned}$$

Andererseits gilt $\int_a^b p_{n+1}(x)\,\mathrm{d}x = 0 = Q_n(p_{n+1})$ weil p_{n+1} punktsymmetrisch bzgl. $\frac{1}{2}(a+b)$ ist. Da $\{p_0, \dots, p_{n+1}\}$ eine Basis von Π_{n+1} bildet, hat Q_n sogar Exaktheitsgrad $n+1$.

□

Bemerkung. Nach Satz 11.8 muss die Simpson-Regel nicht nur wie in Beispiel 11.4 nachgerechnet Exaktheitsgrad $k = 2$ sondern $k = 3$ haben. Dies rechnet man tatsächlich ebenso leicht nach.

Satz 11.9 (Fehlerabschätzung für Newton-Côtes-Formeln).

(i) Ist $f \in C^2[a,b]$, so gilt für die Trapez-Regel

$$|Q_1(f) - I(f)| \leq \frac{(b-a)^3}{12}\|f''\|_\infty.$$

(ii) Ist $f \in C^4[a,b]$, so gilt für die Simpson-Regel

$$|Q_2(f) - I(f)| \leq \frac{(b-a)^5}{2880}\|f^{(4)}\|_\infty.$$

Beweis. Wir verwenden die Fehlerformel aus Satz 10.16

$$|f(x) - p_{0,n}(x)| \leq \frac{\|f^{(n+1)}\|_\infty}{(n+1)!}|\omega_{n+1}(x)|.$$

(i) Sei $p \in \Pi_1$ das lineare Interpolationspolynom zu $(a, f(a))$, $(b, f(b))$. Dann gilt

$$\begin{aligned}|Q_1(f) - I(f)| &= \left|\int_a^b p(x) - f(x)\,\mathrm{d}x\right| \leq \int_a^b |p(x) - f(x)|\,\mathrm{d}x \\ &\leq \frac{\|f''\|_\infty}{2}\int_a^b (x-a)(x-b)\,\mathrm{d}x = \frac{(b-a)^3}{12}\|f''\|_\infty.\end{aligned}$$

(ii) Sei $p \in \Pi_3$ das Hermitesche Interpolationspolynom definiert durch $p(x_j) = f(x_j)$, $j = 0, 1, 2$, und $p'(x_1) = f'(x_1)$. Dann gilt, weil Q_2 Exaktheitsgrad 3 hat, dass

$$Q_2(f) = \sum_{j=1}^n w_j f(x_j) = \sum_{j=1}^n w_j p(x_j) = Q_2(p) = I(p)$$

und somit

$$\begin{aligned}|Q_2(f) - I(f)| &= \left|\int_a^b p(x) - f(x)\,\mathrm{d}x\right| \leq \int_a^b |p(x) - f(x)|\,\mathrm{d}x \\ &\leq \frac{\|f^{(4)}\|_\infty}{4!}\int_a^b (x-x_0)(x-x_1)^2(x-x_2)\,\mathrm{d}x = \frac{(b-a)^5}{2880}\|f^{(4)}\|_\infty.\end{aligned}$$

□

Interpolationsquadraturformeln werden nur für kleine n verwendet (vgl. das Beispiel von Runge). Um höhere Genauigkeiten zu erreichen, zerlegt man $[a,b]$ ähnlich wie bei den Splines das Integrationsgebiet und führt auf jedem Teilintervall eine Quadratur mit niedrigem Grad durch.

Zusammengesetzte Newton-Côtes-Formeln

Definition 11.10. *Sei das Intervall $[a,b]$ in Intervalle $[y_j, y_{j+1}]$ zerlegt, und Q_j sei eine Quadraturformel auf $[y_j, y_{j+1}]$, $j = 0, \ldots, m-1$. Dann wird*

$$\hat{Q}_m(f) := \sum_{j=0}^{m-1} Q_j(f)$$

als ***zusammengesetzte Quadraturformel*** *bezeichnet.*

Beispiel 11.11.

(a) **Zusammengesetze Trapez-Regel**
Sei $y_j = a + h \cdot j$, $h := \frac{b-a}{m}$, $j = 0, \ldots m$. Für die zusammengesetzte Trapez-Regel erhält man

$$T_h(f) = \sum_{j=0}^{m-1} \frac{y_{j+1} - y_j}{2}(f(y_j) + f(y_{j+1})) = \frac{h}{2}\left(f(a) + 2\sum_{j=1}^{m-1} f(y_j) + f(b)\right).$$

Als Fehlerabschätzung ergibt sich aus Satz 11.9 mit $f \in C^2[a,b]$

$$|T_h(f) - I(f)| \leq \sum_{j=0}^{m-1} \frac{h^3}{12}\|f''\|_\infty = \frac{b-a}{12}\|f''\|_\infty h^2.$$

(b) **Zusammengesetze Simpson-Regel**
Sei $x_j = a + h \cdot j$, $j = 0, \ldots, 2m$, $h := \frac{b-a}{2m}$ und $y_j = x_{2j}$, $j = 0, \ldots, m$. Für die zusammengesetze Simpson-Regel erhält man

$$S_h(f) := \sum_{j=0}^{m-1} \frac{y_{j+1} - y_j}{6}\left(f(y_j) + 4f(\underbrace{\frac{y_j + y_{j+1}}{2}}_{x_{2j+1}}) + f(y_{j+1})\right)$$
$$= \frac{h}{3}(f(a) + 4f(x_1) + 2f(x_2) + 4f(x_3) + \ldots + 4f(x_{2m-1}) + f(b)).$$

Wieder aus Satz 11.9 ergibt sich für $f \in C^4[a,b]$

$$|S_h(f) - I(f)| \leq \sum_{j=0}^{m-1} \frac{(2h)^5}{2880}\|f^{(4)}\|_\infty = \frac{b-a}{180}\|f^{(4)}\|_\infty h^4.$$

11.2 Das Romberg-Verfahren

Bei dem in diesem Abschnitt behandelten Romberg-Verfahren handelt es sich um die Anwendung der Grenzwertextrapolation (siehe Abschnitt 10.4) auf die zusammengesetzte Trapezregel für verschiedene Gitterweiten $h_0 > h_1 > \ldots > h_n > 0$. Dabei wird das Interpolationspolynom zu den Punkten (h_0^q, T_{h_0}), (h_1^q, T_{h_1}), $\ldots$, (h_n^q, T_{h_n}) an der Stelle 0 mit Hilfe des Neville-Schemas ausgewertet. Der "extrapolierte" Wert wird dann ein deutlich besserer Näherungswert als T_h sein. Genauere Auskunft über den Fehler gibt Satz 10.20.

Bevor wir die Vorraussetzung von Satz 10.20 durch die Euler-Maclaurinsche Summenformel (Satz 11.14) belegen können, benötigen wir die *Bernoullischen Zahlen*.

Definition 11.12. *Die durch* $B_0(t) := 1$ *und*

$$B_k'(t) := kB_{k-1}(t), \quad \int_0^1 B_k(t)\,\mathrm{d}t = 0, \quad k \geq 1,$$

eindeutig bestimmten Polynome heissen ***Bernoulli-Polynome****. Durch* $B_k := B_k(0)$ *sind die* ***Bernoullischen Zahlen*** *definiert.*

Beispielsweise ist $B_1(t) = t - \frac{1}{2}$ $B_2(t) = t^2 - t + \frac{1}{6}$.

Lemma 11.13. *Für die Bernoulli-Polynome gilt*

(i) $B_k(0) = B_k(1)$, $k \geq 2$,

(ii) $B_k(t) = (-1)^k B_k(1-t)$, $k \geq 0$,

(iii) $B_{2k+1}(0) = B_{2k+1}(\frac{1}{2}) = B_{2k+1}(1) = 0$, $k \geq 1$.

Beweis.

(i) $B_k(1) - B_k(0) = \int_0^1 B_k'(t)\,\mathrm{d}t = k\int_0^1 B_{k-1}(t)\,\mathrm{d}t = 0$ für $k \geq 2$.

(ii) Wir setzen $C_k(t) = (-1)^k B_k(1-t)$. Dann gilt

$$C_0(t) = 1, \quad C_k'(t) = (-1)^{k-1} k B_{k-1}(1-t) = kC_{k-1}(t)$$

und

$$\int_0^1 C_k(t)\,\mathrm{d}t = (-1)^k \int_0^1 B_k(1-t)\,\mathrm{d}t = 0, \quad k \geq 1.$$

Also genügt C_k denselben Rekrusionsformeln, und wir erhalten $C_k(t) = B_k(t)$.

(iii) Folgt aus (i) und (ii).

□

Der folgende Satz bestätigt die Existenz einer wie in Satz 10.20 vorausgesetzten asymptotischen Entwicklung

$$T_h(f) = I(f) + \sum_{i=1}^{n} a_i h^{qi} + a_{n+1}(h)$$

mit $q = 2$ und $|a_{n+1}(h)| \leq c\,h^{2(n+1)}$, wenn er auf $f \in C^{2(n+1)}[a,b]$ angewendet wird.

Satz 11.14 (Euler-Maclaurinsche Summenformel). *Sei* $f \in C^{2n}[a,b]$, $n \in \mathbb{N}$, *und* $h := \frac{b-a}{m}$, $m \in \mathbb{N}$. *Dann gilt für die zusammengesetze Trapez-Regel*

$$T_h(f) = I(f) + \sum_{k=1}^{n} \frac{h^{2k}}{(2k)!} B_{2k}(f^{(2k-1)}(b) - f^{(2k-1)}(a)) + O(h^{2n}).$$

Beweis. Sei $\varphi \in C^{2n}[0,1]$ eine beliebige Funktion. Dann gilt

$$\begin{aligned}
\int_0^1 \varphi(t)\,\mathrm{d}t &= \int_0^1 B_0(t)\varphi(t)\,\mathrm{d}t = B_1(t)\varphi(t)\big|_0^1 - \int_0^1 B_1(t)\varphi'(t)\,\mathrm{d}t \\
&= \frac{1}{2}(\varphi(0)+\varphi(1)) - \frac{1}{2}B_2(t)\varphi'(t)\big|_0^1 + \frac{1}{3}B_3(t)\varphi''(t)\big|_0^1 - \int_0^1 \frac{1}{3}B_3(t)\varphi'''(t)\,\mathrm{d}t \\
&= \frac{1}{2}(\varphi(0)+\varphi(1)) - \sum_{k=1}^{n} \frac{1}{(2k)!}B_{2k}\left(\varphi^{(2k-1)}(1) - \varphi^{(2k-1)}(0)\right) \\
&\quad + \int_0^1 \frac{1}{(2n)!}B_{2n}(t)\varphi^{(2n)}(t)\,\mathrm{d}t.
\end{aligned}$$

Wir setzen $\varphi_j(t) = h \cdot f(x_j + th)$, $0 \le j < n$. Dann gilt $\varphi_j(1) = \varphi_{j+1}(0) = h \cdot f(x_{j+1})$,

$$\int_0^1 \varphi_j(t)\,\mathrm{d}t = \int_{x_j}^{x_{j+1}} f(x)\,\mathrm{d}x, \quad \varphi_j^{(k)}(t) = h^{k+1} f^{(k)}(x_j + th)$$

und somit $\varphi_j^{(2k-1)}(1) = \varphi_{j+1}^{(2k-1)}(0)$, $0 \le j < n$. Also folgt

$$\begin{aligned}
\int_a^b f(x)\,\mathrm{d}x &= \sum_{j=0}^{m-1} \int_{x_j}^{x_{j+1}} f(x)\,\mathrm{d}x = \sum_{j=0}^{m-1} \int_0^1 \varphi_j(t)\,\mathrm{d}t \\
&= \sum_{j=0}^{m-1} \frac{1}{2}(\varphi_j(0)+\varphi_j(1)) - \sum_{j=0}^{m-1}\sum_{k=1}^{n} \frac{1}{(2k)!}B_{2k}\left(\varphi_j^{(2k-1)}(1) - \varphi_j^{(2k-1)}(0)\right) \\
&\quad + \sum_{j=0}^{m-1} \int_0^1 \frac{1}{(2n)!}B_{2n}(t)\varphi_j^{(2n)}(t)\,\mathrm{d}t \\
&= \sum_{j=0}^{m-1} \frac{h}{2}(f(x_j)+f(x_{j+1})) - \sum_{k=1}^{n} \frac{1}{(2k!)}B_{2k} \sum_{j=0}^{m-1} \varphi_j^{(2k-1)}(1) - \varphi_j^{(2k-1)}(0) \\
&\quad + \sum_{j=0}^{m-1} \int_0^1 \frac{1}{(2n)!}B_{2n}(t)\varphi_j^{(2n)}(t)\,\mathrm{d}t \\
&= \sum_{j=0}^{m-1} \frac{h}{2}(f(x_j)+f(x_{j+1})) - \sum_{k=1}^{n} \frac{h^{2k}}{(2k)!}B_{2k}\left(f^{(2k-1)}(b) - f^{(2k-1)}(a)\right) \\
&\quad + h^{2n+1} \sum_{j=0}^{m-1} \int_0^1 \frac{1}{(2n)!}B_{2n}(t) f^{(2n)}(x_j + th)\,\mathrm{d}t.
\end{aligned}$$

Wegen

$$\begin{aligned}
&\Big|h \sum_{j=0}^{m-1} \int_0^1 \frac{1}{(2n)!}B_{2n}(t) f^{(2n)}(x_j + th)\,\mathrm{d}t\Big| \\
&\qquad \le \frac{h}{(2n)!} \sum_{j=0}^{m-1} \left(\sup_{t\in[0,1]} |B_{2n}(t)|\right)\left(\sup_{t\in[0,1]} |f^{(2n)}(x_j+th)|\right) \\
&\qquad \le \frac{(b-a)}{(2n)!}\|B_{2n}\|_\infty \|f^{(2n)}\|_{\infty,[a,b]}
\end{aligned}$$

folgt die Behauptung. □

Bemerkung. Im Fall periodischer Funktionen f, d.h. falls $f^{(2k-1)}(a) = f^{(2k-1)}(b)$, $k = 1, \ldots, n$, liefert die zusammengesetzte Trapez-Regel nach dem letzten Satz bereits einen Fehler der Ordnung $O(h^{2n})$. Dies kann durch die Romberg-Quadratur nicht verbessert werden.

12 Iterative Lösungsverfahren

In vielen Anwendungen tritt das Problem auf, ein nicht-lineares Gleichungssystem lösen zu müssen. Für lineare Gleichungssysteme haben wir bereits in der Vorlesung *Algorithmische Mathematik I* direkte Lösungsverfahren kennengelernt. Diese eignen sich allerdings nicht für nicht-lineare Probleme. Hier werden üblicherweise *iterative* Verfahren angewendet, bei denen eine Folge von Approximationen $\{x_k\}_{k\in\mathbb{N}}$ durch

$$x_{k+1} = \Phi(x_k, \ldots, x_{k+1-m}), \quad k = m-1, m, m+1, m+2, \ldots,$$

mit gewählten Startwerten $x_0, \ldots, x_{m-1}$ für ein $m \in \mathbb{N}$ konstruiert wird. Die Funktion Φ wird als **Iterationsvorschrift** bezeichnet.

Die Konvergenzgeschwindigkeit einer Iteration gibt Auskunft über die Qualität eines Iterationsverfahrens.

Definition 12.1. *Eine Folge $\{x_k\}_{k\in\mathbb{N}} \subset \mathbb{K}^n$* ***konvergiert von (mind.) Ordnung $p \geq 1$*** *gegen $x \in \mathbb{K}^n$, falls ein $c > 0$ existiert mit*

$$\|x_{k+1} - x\| \leq c\|x_k - x\|^p, \quad k \geq m-1.$$

Im Fall $p = 1$ fordert man zusätzlich $c < 1$.

Bemerkung. Im Fall $p = 1$ spricht man von linearer, für $p = 2$ von quadratischer und für $p = 3$ von kubischer Konvergenz.

Bei nichtlinearen Problemen ist es zudem wichtig, zwischen lokaler und globaler Konvergenz zu unterscheiden.

Definition 12.2. *Ein Iterationsverfahren mit Iterierten $\{x_k\}_{k\in\mathbb{N}}$ heisst* ***lokal konvergent*** *gegen $x \in \mathbb{K}^n$, falls es eine Umgebung $U \subset \mathbb{K}^n$ und $x \in U$ gibt mit $x = \lim_{k\to\infty} x_k$ für alle Startwerte $x_0, \ldots, x_{m-1} \in U$. Ist $U = \mathbb{K}^n$, so heisst das Verfahren* ***global konvergent****.*

12.1 Der Banachsche Fixpunktsatz

Der folgende Banachsche Fixpunktsatz stellt die Grundlage für die Konvergenzanalyse aller hier untersuchten Iterationsverfahren dar. Wir benötigen einige Definitionen.

Definition 12.3. *Eine Abbildung Φ heisst* ***Selbstabbildung*** *von $M \subset \mathbb{K}^n$, falls $\Phi : M \to M$ gilt. Ist Φ Lipschitz-stetig, d.h. gilt*

$$\|\Phi(x) - \Phi(y)\| \leq L\,\|x - y\| \quad \text{für alle } x, y \in M,$$

mit Lipschitz-Konstante $0 \leq L < 1$, so heisst Φ ***kontrahierend****.*

Definition 12.4. *Sei $\Phi : M \to M$ eine Selbstabbildung von $M \subset \mathbb{K}^n$. Ein Punkt $x \in M$ heisst **Fixpunkt** von Φ, falls er der **Fixpunktgleichung** $x = \Phi(x)$ genügt.*

Satz 12.5. *Sei Φ eine kontrahierende Selbstabbildung der abgeschlossenen Menge $M \subset \mathbb{K}^n$ mit Lipschitz-Konstante $0 \le L < 1$. Dann existiert genau ein Fixpunkt $x \in M$ von Φ. Gegen diesen konvergiert die Folge $\{x_k\}_{k\in\mathbb{N}}$ definiert durch $x_{k+1} = \Phi(x_k)$ für alle $x_0 \in M$ linear, und es gilt für $k \in \mathbb{N}$*

(i)	$\|x_{k+1} - x\| \le L \|x_k - x\|$,	*"Monotonie"*
(ii)	$\|x_k - x\| \le \frac{L^k}{1-L} \|x_1 - x_0\|$,	*"a-priori Schranke"*
(iii)	$\|x_{k+1} - x\| \le \frac{L}{1-L} \|x_{k+1} - x_k\|$.	*"a-posteriori Schranke"*

Beweis. Aufgrund der Kontraktionseigenschaft von Φ gilt für $k \ge 1$

$$\|x_{k+1} - x_k\| = \|\Phi(x_k) - \Phi(x_{k-1})\| \le L \|x_k - x_{k-1}\| \le \ldots \le L^k \|x_1 - x_0\|. \tag{12.1}$$

Wir zeigen nun, dass $\{x_k\}_{k\in\mathbb{N}}$ eine Cauchy-Folge ist. Dazu seien $\varepsilon > 0$ und $m, n \in \mathbb{N}$ mit $m > n \ge N$, wo $N \in \mathbb{N}$ so gewählt ist, dass $L^N \|x_1 - x_0\| \le (1 - L)\varepsilon$. Aus (12.1) erhält man

$$\begin{aligned} \|x_m - x_n\| &\le \|x_m - x_{m-1}\| + \|x_{m-1} - x_{m-2}\| + \cdots + \|x_{n+1} - x_n\| \\ &\le (L^{m-1} + L^{m-2} + \cdots + L^n) \|x_1 - x_0\| \\ &\le \frac{L^n}{1-L} \|x_1 - x_0\| \le \varepsilon. \end{aligned} \tag{12.2}$$

Wegen der Vollständigkeit von M besitzt $\{x_k\}_{k\in\mathbb{N}}$ einen Grenzwert $x \in M$. Aus

$$\|x - \Phi(x)\| \le \|x - x_k\| + \|x_k - \Phi(x)\| \le \|x - x_k\| + L \|x_{k-1} - x\|$$

sieht man wegen des Verschwindens der rechten Seite für $k \to \infty$, dass $x = \Phi(x)$. Ist $\hat{x}$ ein weiterer Fixpunkt, dann hat man $\|\hat{x} - x\| = \|\Phi(\hat{x}) - \Phi(x)\| \le L \|\hat{x} - x\|$, woraus wegen $L < 1$ folgt, dass $\hat{x} = x$.

Die Monotonie (i) und Aussage (iii) gelten wegen

$$\begin{aligned} \|x_{k+1} - x\| &= \|\Phi(x_k) - \Phi(x)\| \le L \|x_k - x\| \\ &\le L \|x_k - x_{k+1} + x_{k+1} - x\| \le L \|x_{k+1} - x_k\| + L \|x_{k+1} - x\|. \end{aligned}$$

Die Abschätzung (ii) erhält man aus (12.2) im Grenzfall $m \to \infty$. □

Beispiel 12.6. Wir wollen die Lösung von $x^2 = 2$ mittels Fixpunktiteration bestimmen. Dazu benötigen wir nach dem letzten Satz eine Fixpunktgleichung $x = \Phi(x)$ und eine abgeschlossene Menge M mit $x \in M$, so dass $\Phi : M \to M$ eine Kontraktion ist. Sei $\Phi(x) = x/2 + 1/x$ und $M = [1, 2]$. Wegen

$$|\Phi(x) - \Phi(y)| = \left|\frac{1}{2}(x - y) - \frac{x - y}{xy}\right| = \left|\frac{1}{2} - \frac{1}{xy}\right| |x - y| \le \frac{1}{2} |x - y| \quad \text{für alle } x, y \in M$$

ist Φ eine Kontraktion und Selbstabbildung von M. Die Folge $x_{k+1} = x_k/2 + 1/x_k$ konvergiert für alle $x_0 \in M$ gegen den Fixpunkt $x = \sqrt{2}$.

Die Kontraktionseigenschaft ist oft schwierig zu überprüfen, weil sie meistens nicht global gegeben ist; beispielsweise ist Φ in Beispiel 12.6 keine Kontraktion auf $[\frac{1}{2}, 2]$. Das folgende Kriterium ist in der Regel leichter zu überprüfen.

Satz 12.7. *Seit $M \subset \mathbb{K}^n$ abgeschlossen und konvex und $\Phi : M \to M$ einmal stetig differenzierbar. Ferner sei $\|\cdot\|$ eine Vektornorm bzw. eine verträgliche Matrixnorm. Ist*

$$L := \sup_{y \in M} \|D\Phi(y)\| < 1,$$

so hat Φ genau einen Fixpunkt, gegen den die Folge $\{x_k\}_{k\in\mathbb{N}}$ definiert durch $x_{k+1} = \Phi(x_k)$ konvergiert.

Beweis. Nach dem Mittelwertsatz der Integralrechnung gilt für alle $y, z \in M$

$$\Phi(y) - \Phi(x) = \int_0^1 D\Phi((1-t)y + tz)(y - z)\,\mathrm{d}t.$$

Weil M konvex ist, gilt $(1-t)y + tz \in M$ für alle $t \in [0, 1]$. Also folgt

$$\|\Phi(y) - \Phi(z)\| \le \int_0^1 \|D\Phi((1-t)y + tz)\|\|y - z\|\,\mathrm{d}t \le L\|y - z\|,$$

und Φ ist eine Kontraktion auf der abgeschlossenen Menge M. □

Korollar 12.8. Sei $M \subset \mathbb{K}^n$ eine offene Menge und $\Phi : M \to \mathbb{K}^n$ stetig differenzierbar. Ferner sei $\|\cdot\|$ eine Vektornorm bzw. eine verträgliche Matrixnorm. Ist x ein Fixpunkt und gilt $\|D\Phi(x)\| < 1$, so konvergiert die Folge $\{x_k\}_{k\in\mathbb{N}}$ definiert durch $x_{k+1} = \Phi(x_k)$ lokal gegen x.

Beweis. Sei $\delta := 1 - \|D\Phi(x)\| > 0$. Wegen der Stetigkeit von $D\Phi$ existiert eine Umgebung $U := \{y \in \mathbb{K}^n : \|x - y\| \le \varepsilon\} \subset M$ von x mit $\|D\Phi(y)\| < 1 - \delta/2 =: L$ für alle $y \in U$. Für Satz 12.7 müssen wir noch zeigen, dass Φ eine Selbstabbildung der abgeschlossenen und konvexen Menge U ist. Dazu sei $y \in U$ beliebig gewählt. Dann gilt wegen $x = \Phi(x)$

$$\|x - \Phi(y)\| = \|\Phi(x) - \Phi(y)\| \le \int_0^1 \|D\Phi((1-t)x + ty)\|\|x - y\|\,\mathrm{d}t \le L\varepsilon \le \varepsilon$$

und somit $\Phi(y) \in U$. □

Wir beschäftigen uns in den beiden nächsten Abschnitten zunächst mit iterativen Lösungsverfahren für lineare Gleichungssysteme $Ax = b$. Während direkte Verfahren die Lösung in endlich vielen Schritten bestimmen, verbessern iterative Verfahren eine Anfangsnäherung sukzessive. Dabei wird im Gegensatz zu direkten Verfahren die Matrix aber nicht verändert, sondern geht nur durch Multiplikation mit Vektoren ein. Dies ist besonders dann von Vorteil, wenn A z.B. schwach besetzt ist, d.h. nur eine konstante Anzahl von Einträgen pro Zeile und Spalte nicht verschwinden. Im Gegensatz dazu besitzen die Faktoren der LR-Zerlegung signifikant mehr nicht-verschwindende Einträge pro Zeile/Spalte als A (sog. **fill-in**), was die Komplexität bedeutend verschlechtert.

12.2 Klassische Iterationsverfahren

In diesem Abschnitt betrachten wir Verfahren, die aus einer Zerlegung (sog. **reguläres Splitting**)

$$A = M + (A - M) \quad \text{von } A \in \mathbb{K}^{n\times n} \text{ regulär}$$

mit einer regulären Matrix $M \in \mathbb{C}^{n\times n}$ entstehen. Die Gleichung $Ax = b$ lässt sich damit auch als die Fixpunktgleichung

$$Mx = b - (A - M)x \iff x = (I - M^{-1}A)x + M^{-1}b$$

schreiben. Wenn die Lösung eines Gleichungssystems mit Koeffizientenmatrix M deutlich leichter fällt und die Matrix-Vektor-Multiplikation mit $A - M$ billig ist, kann es sinnvoll sein, die Folge $\{x_k\}_{k\in\mathbb{N}}$ definiert durch

$$x_{k+1} = Tx_k + c, \quad T := I - M^{-1}A,\ c = M^{-1}b, \tag{12.3}$$

mit Startvektor $x_0 \in \mathbb{K}^n$ zu berechnen. Der folgende Satz charakterisiert die Konvergenz des Interationsverfahrens (12.3).

Satz 12.9. *Sei $\{x_k\}_{k\in\mathbb{N}}$ die durch (12.3) definierte Folge. Dann gilt*

(i) $\{x_k\}$ konvergiert genau dann für jeden Startwert $x_0 \in \mathbb{K}^n$ gegen die Lösung von $Ax = b$, wenn $\rho(T) < 1$.

(ii) Ist $\|\cdot\|$ eine Norm auf $\mathbb{K}^n$ bzw. die zugeordnete Matrixnorm und ist $q := \|T\| < 1$, so konvergiert $\{x_k\}$ für jeden Startwert $x_0 \in \mathbb{K}^n$ gegen die Lösung von $Ax = b$. Ferner gilt

$$\|x_k - x\| \le \frac{q^k}{1-q}\|x_0 - x_1\| \quad \text{und} \quad \|x_{k+1} - x\| \le \frac{q}{1-q}\|x_{k+1} - x_k\|.$$

Beweis.

(i) Wegen $x = Tx + c$ gilt, dass $x_{k+1} - x = Tx_k + c - x = T(x_k - x)$. Sei vorausgesetzt, dass $\{x_k\}$ für jedes $x_0 \in \mathbb{K}^n$ konvergiert. Sei $\lambda \in \mathbb{C}$ der betragsmaximale Eigenwert von T und $v \in \mathbb{K}^n$ ein zugehöriger Eigenvektor. Betrachte den Startwert $x_0 = x + v$. Es gilt

$$x_k - x = T^k(x_0 - x) = T^k v = \lambda^k v.$$

Weil $\{x_k\}$ konvergiert, muss $\rho(T) = |\lambda| < 1$ gelten.
Ist umgekehrt $\rho(T) < 1$, so existiert $\varepsilon > 0$, so dass $p := \rho(T) + \varepsilon < 1$. Nach Satz 6.20 existiert eine zugeordnete Norm $\|\cdot\|_\varepsilon$ mit $\|T\|_\varepsilon \le \rho(T) + \varepsilon = p < 1$. Sei $x_0 \in \mathbb{K}^n$ beliebig. Dann ist

$$\|x_k - x\|_\varepsilon = \|T^k(x_0 - x)\|_\varepsilon \le \|T^k\|\|x_0 - x\| \le p^k\|x_0 - x\|.$$

Wegen $p < 1$ konvergiert $\{x_k\}$ gegen x.

(ii) Ist $\|T\| < 1$, so gilt nach Satz 6.19 $\rho(T) \leq \|T\| < 1$. Nach (i) konvergiert $\{x_k\}$ für jedes $x_0 \in \mathbb{K}^n$ gegen die Lösung von $Ax = b$. Ferner ist die Abbildung $\Phi(x) := Tx + c$ eine Kontraktion, weil

$$\|\Phi(x) - \Phi(y)\| = \|T(x-y)\| \leq \|T\|\|x-y\|.$$

Die Behauptung folgt aus dem Banachschen Fixpunktsatz.

□

Beispiel 12.10. Sei $A \in \mathbb{K}^{n\times n}$ regulär mit positiven Eigenwerten $\lambda_1 \leq \ldots \leq \lambda_n$ und $\alpha > 0$. Mit der Wahl $M = \alpha^{-1}I$ in (12.3) erhält man das so genannte **Richardson-Verfahren**

$$x_{k+1} = x_k + \alpha(b - Ax_k) = (I - \alpha A)x_k + \alpha b.$$

Weil die Iterationsmatrix $T = I - \alpha A$ die Eigenwerte $1 - \alpha\lambda_i$ besitzt, ist $\rho(T) < 1$ äquivalent mit

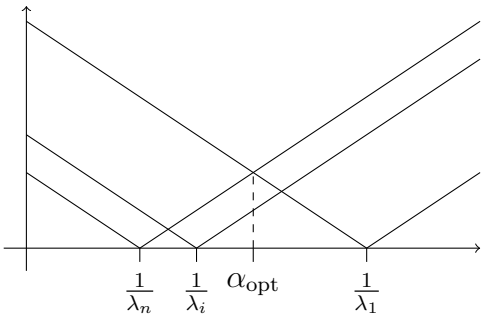
$$1 - \alpha\lambda_n > -1 \iff \alpha < \frac{2}{\lambda_n}.$$

Das Richardson-Verfahren konvergiert also für alle $\alpha \in (0, 2/\lambda_n)$. Wir gehen nun der Frage nach, für welches α der Spektralradius von T minimal ist. Wegen

$$\rho(T) = \max_{i=1,\ldots,n} |1 - \alpha\lambda_i| = \max\{|1 - \alpha\lambda_n|,\ |1 - \alpha\lambda_1|\}$$

ist α_{opt} die Schnittstelle der beiden Funktionen

$$f_1(\alpha) := |1 - \alpha\lambda_n| \quad \text{und} \quad f_2(\alpha) := |1 - \alpha\lambda_1|,$$



d.h.

$$-1 + \alpha_{\text{opt}}\lambda_n = 1 - \alpha_{\text{opt}}\lambda_1 \iff \alpha_{\text{opt}} = \frac{2}{\lambda_1 + \lambda_n}.$$

Der Spektralradius für diese Wahl ist

$$\rho(T) = \frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}.$$

Die Konvergenzgeschwindigkeit hängt also von der Verteilung der Eigenwerte von A ab. Diese Beobachtung werden wir auch für andere iterative Verfahren zur Lösung linearer Gleichungssysteme machen.

Im Folgenden stellen wir zwei weitere Verfahren vom Typ (12.3) vor. Dazu sei

$$A = A_L + A_D + A_R,$$

wobei $A_D = \operatorname{diag}(A)$, A_L und A_R die strikte (d.h. ohne Diagonale) untere bzw. obere Dreiecksmatrix von A bezeichnen. Wir nehmen an, dass A_D regulär ist. Dies kann immer durch Umsortieren der Zeilen-/Spaltenindizes erreicht werden. Beim **Jacobi-** oder **Gesamtschritt-Verfahren** wählt man $M = A_D$ in (12.3). Die Iterationsmatrix ist dann

$$T_J = I - A_D^{-1}A = -A_D^{-1}(A_L + A_R).$$

Komponentenweise bedeutet dies

$$x_i^{(k+1)} = \frac{1}{a_{ii}}\left(b_i - \sum_{j\neq i} a_{ij}x_j^{(k)}\right), \quad i = 1,\dots,n, \quad k = 0,1,2,\dots.$$

Beim **Einzelschritt-** oder **Gauß-Seidel-Verfahren** verwendet man im Vergleich zum Gesamtschrittverfahren alle bereits berechneten Komponenten von $x^{(k+1)}$, also

$$x_i^{(k+1)} = \frac{1}{a_{ii}}\left(b_i - \sum_{j<i} a_{ij}x_j^{(k+1)} - \sum_{j>i} a_{ij}x_j^{(k)}\right), \quad i = 1,\dots,n, \quad k = 0,1,2,\dots.$$

Dieses Verfahren entspricht der Wahl $M = A_D + A_L$ in (12.3). Die Iterationsmatrix ist folglich

$$T_{GS} = I - (A_D + A_L)^{-1}A = -(A_D + A_L)^{-1}A_R.$$

Der Name "Einzelschrittverfahren" rührt aus der Behandlung der Komponenten des Vektors $x^{(k+1)}$ her. Diese werden einzeln und nicht wie beim Gesamtschrittverfahren auf einmal berechnet.

Die Konvergenz von Gesamt- und Einzelschrittverfahren hängt von Eigenschaften der Matrix A ab. In der Literatur sind unterschiedliche hinreichende bekannt, die mehr oder weniger praktikabel sind. Wir konzentrieren uns auf eine einfache Bedingung.

Definition 12.11. *Eine Matrix $A \in \mathbb{K}^{n\times n}$ heißt **diagonaldominant**, falls*

$$|a_{ii}| > \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}| \quad \text{für alle } i = 1,\dots,n.$$

Satz 12.12. *Ist $A \in \mathbb{K}^{n\times n}$ diagonaldominant, dann konvergieren Gesamt- und Einzelschrittverfahren für jeden Startvektor $x_0 \in \mathbb{K}^n$ gegen die Lösung von $Ax = b$.*

Beweis. Nach Satz 12.9 müssen wir nur zeigen, dass $\|T_J\|_\infty <$ und $\|T_{GS}\|_\infty < 1$.

(i) $\|T_J\|_\infty = \|A_D^{-1}(A_L + A_R)\|_\infty = \max_{i=1,\dots,n} \sum_{j\neq i} \frac{|a_{ij}|}{|a_{ii}|} < 1.$

(ii) Sei $x \in \mathbb{K}^n$ mit $\|x\|_\infty = 1$ und $y = T_{GS}x$. Dann ist

$$y_i = -\frac{1}{a_{ii}} \left(\sum_{j<i} a_{ij} y_j + \sum_{j>i} a_{ij} x_j \right).$$

Wir zeigen induktiv, dass $|y_i| < 1$, $i = 1, \ldots, n$. Der Fall $i = 1$ ist klar, weil

$$|y_1| = \frac{1}{|a_{11}|} \left| \sum_{j>1} a_{1j} x_j \right| \leq \frac{1}{|a_{11}|} \sum_{j>1} |a_{1j}||x_j| \leq \sum_{j>1} \frac{|a_{1j}|}{|a_{11}|} < 1.$$

Angenommen, es gilt $|y_i| < 1,\ i = 1, \ldots, k-1$. Dann ist

$$|y_k| \leq \frac{1}{|a_{kk}|} \left(\sum_{j<k} |a_{kj}||y_j| + \sum_{j>k} |a_{kj}||x_j| \right) \leq \frac{1}{|a_{kk}|} \sum_{j \neq k} |a_{kj}| < 1.$$

Also gilt $\|y\|_\infty < 1$ und somit

$$\|T_{GS}\|_\infty = \max_{\|x\|_\infty = 1} \|T_{GS}x\|_\infty < 1.$$

□

Das Gauß-Seidel-Verfahren ist zwar aufwendiger, konvergiert aber für bestimmte Klassen von Matrizen schneller.

Relaxationsverfahren

Bei Relaxationsverfahren wird ausgehend von einem bekannten Verfahren ein Parameter $\omega > 0$ mit dem Ziel eingeführt, den Spektralradius zu verkleinern. Sei $z_i^{(k+1)}$ das Ergebnis einer Berechnungsvorschrift, die auf den bereits berechneten Komponenten von $x^{(k)}$ und $x^{(k+1)}$ basiert. Dann wird durch

$$x_i^{(k+1)} = (1 - \omega) x_i^{(k)} + \omega z_i^{(k+1)}$$

ein neues Iterationsverfahren, das sog. **Relaxationsverfahren** definiert. ω heißt Relaxationsparameter.

Bemerkung. Für $\omega < 1$ wird das Verfahren als unterrelaxiert, für $\omega > 1$ als überrelaxiert bezeichnet. Für $\omega = 1$ erhält man das ursprüngliche Verfahren.

Im Folgenden wollen wir ω so bestimmen, dass $\rho(T^{(\omega)})$ im Fall des relaxierten Jacobi- und des relaxierten Gauß-Seidel-Verfahrens möglichst klein ist.

Relaxiertes Jacobi-Verfahren

Das relaxierte Jacobi-Verfahren in Komponentenform lautet

$$x_i^{(k+1)} = (1 - \omega) x_i^{(k)} + \omega z_i^{(k+1)}, \quad z_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j \neq i} a_{ij} x_j^{(k)} \right)$$

also

$$x_i^{(k+1)} = x_i^{(k)} + \frac{\omega}{a_{ii}} \left(b_i - \sum_{j=1}^{n} a_{ij} x_j \right), \quad i = 1, \ldots, n, \quad k = 0, 1, 2, \ldots.$$

Dies entspricht der Wahl $M = \frac{1}{\omega} A_D$ im Splitting und somit $T_J^{(\omega)} = I - \omega A_D^{-1} A$.

Satz 12.13. *Die Matrix $A_D^{-1}A$ besitze nur reelle Eigenwerte $0 < \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$. Dann konvergiert das relaxierte Jacobi-Verfahren für alle $\omega \in (0, 2/\lambda_n)$. Der Spektralradius von $T_J^{(\omega)}$ wird minimal für $\omega_{\text{opt}} = \frac{2}{\lambda_1+\lambda_n}$, und es gilt*

$$\rho(T_J^{(\omega_{\text{opt}})}) = \frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}.$$

Beweis. $T_J^{(\omega)}$ besitzt die Eigenwerte $1-\omega\lambda_i$. Daher folgt die Aussage wie in Beispiel 12.10. □

Relaxiertes Gauß-Seidel-Verfahren

Das relaxierte Gauß-Seidel-Verfahren oder engl. **Successive Overrelaxation (SOR)** lautet in Komponentenform

$$x_i^{(k+1)} = (1-\omega)x_i^{(k)} + \omega z_i^{(k+1)}, \quad z_i^{(k+1)} = \frac{1}{a_{ii}}\left(b_i - \sum_{j<i} a_{ij}x_j^{(k+1)} - \sum_{j>i} a_{ij}x_j^{(k)}\right)$$

und somit $x_{k+1} = (1-\omega)x_k + \omega z_{k+1}$, $A_D z_{k+1} = b - A_L x_{k+1} - A_R x_k$. Dies ist äquivalent zu

$$\begin{aligned} & A_D x_{k+1} = (1-\omega)A_D x_k + \omega b - \omega A_L x_{k+1} - \omega A_R x_k \\ \iff & \left(\frac{1}{\omega}A_D + A_L\right) x_{k+1} = \left[\left(\frac{1}{\omega} - 1\right) A_D - A_R\right] x_k + b. \end{aligned}$$

Hier ist

$$M = \frac{1}{\omega}A_D + A_L, \quad T_{GS}^{(\omega)} = (A_D + \omega A_L)^{-1}[(1-\omega)A_D - \omega A_R].$$

Satz 12.14. *Es bezeichne $T_{GS}^{(\omega)}$ die Iterationsmatrix des SOR. Dann gilt $\rho(T_{GS}^{(\omega)}) \geq |\omega - 1|$. Daher kann das SOR nur für jeden Startwert konvergieren, falls $\omega \in (0, 2)$.*

Beweis. Wegen $T_{GS}^{(\omega)} = (A_D + \omega A_L)^{-1}[(1-\omega)AD - \omega A_R]$ gilt für jeden betragsmaximalen Eigenwert λ von $T_{GS}^{(\omega)}$ nach Lemma 6.13

$$\begin{aligned} |\lambda|^n \geq |\det T_{GS}^{(\omega)}| &= \frac{|\det(1-\omega)A_D - \omega A_R|}{|\det A_D + \omega A_L|} \\ &= \frac{|\det(1-\omega)A_D|}{|\det A_D|} = \frac{|1-\omega|^n |\det A_D|}{|\det A_D|} = |1-\omega|^n. \end{aligned}$$

□

Wir haben aber auch folgendes positives Resultat.

Satz 12.15. *Sei $A \in \mathbb{K}^{n\times n}$ positiv definit. Dann konvergiert das SOR-Verfahren für alle $\omega \in (0, 2)$; insbesondere konvergiert das Gauß-Seidel-Verfahren ($\omega = 1$).*

Beweis. Sei (λ, x) ein Eigenpaar von $T_{GS}^{(\omega)} = I - M^{-1}A$, $M = \omega^{-1}A_D + A_L$. Dann gilt

$$\lambda x = T_{GS}^{(\omega)} x = (I - M^{-1}A)x = x - M^{-1}Ax$$

und somit $Ax = (1-\lambda)Mx$. Weil A nicht singulär ist, ist $\lambda \neq 1$ und daher

$$\frac{1}{1-\lambda} = \frac{x^H M x}{x^H A x}.$$

Also folgt

$$2\,\mathrm{Re}\left(\frac{1}{1-\lambda}\right) = \frac{1}{1-\lambda} + \overline{\frac{1}{1-\lambda}} = \frac{x^H M x}{x^H A x} + \overline{\frac{x^H M x}{x^H A x}} = \frac{x^H (M + M^H) x}{x^H A x}.$$

Wegen $M = \omega^{-1}A_D + A_L$ ist $M + M^H = 2/\omega\, A_D + A_L + A_R = A + (2/\omega - 1)A_D$. Zusammengefasst gilt

$$2\,\mathrm{Re}\left(\frac{1}{1-\lambda}\right) = \frac{x^H \left(A + (2/\omega - 1)\, A_D\right) x}{x^H A x} = 1 + \underbrace{\left(\frac{2}{\omega} - 1\right)}_{>0} \underbrace{\frac{x^H A_D x}{x^H A x}}_{>0} > 1,$$

da mit A auch A_D positiv definit ist. Mit $\lambda = \alpha + \mathrm{i}\beta$ ist

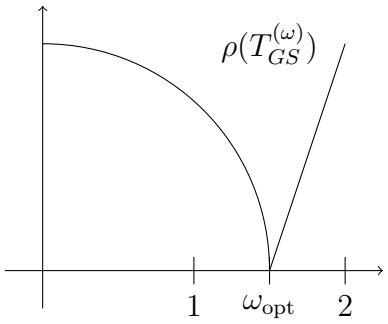
$$1 < 2\,\mathrm{Re}\left(\frac{1}{1-\lambda}\right) = 2\,\mathrm{Re}\left(\frac{1}{1-\alpha-\mathrm{i}\beta}\right) = 2\frac{1-\alpha}{(1-\alpha)^2 + \beta^2}$$

und somit

$$(1-\alpha)^2 + \beta^2 < 2 - 2\alpha \iff |\lambda|^2 = \alpha^2 + \beta^2 < 2 - 2\alpha - 1 + 2\alpha = 1.$$

Alle Eigenwerte von $T_{GS}^{(\omega)}$ sind also betragsmäßig kleiner als 1. Nach Satz 12.9 folgt die Behauptung. □

Bemerkung. Die Bestimmung des optimalen Relaxationsparameters ω_{opt} für das SOR-Verfahren fällt im Allgemeinen schwer. Beim SOR-Verfahren hat man aber folgendes qualitatives Verhalten.



$\rho(T_{GS}^{(\omega)})$ hat rechts von ω_{opt} die Steigung 1, die linksseitige Ableitung in ω_{opt} ist $-\infty$. Daher ist es besser, den optimalen Relaxationsparameter zu überschätzen als zu unterschätzen.

12.3 Gradientenverfahren

Im Folgenden betrachten wir positiv definite Matrizen. Bisher war diese Eigenschaft bzgl. des euklidischen Skalarproduktes zu verstehen. Dies verallgemeinern wir nun.

Definition 12.16. *Eine Matrix $A \in \mathbb{K}^{n\times n}$ heißt **positiv definit** bzgl. eines Skalarproduktes $(\cdot,\cdot)$ auf $\mathbb{K}^n$, falls A **selbstadjungiert** ist, d.h. es gilt $(Ax, y) = (x, Ay)$ für alle $x, y \in \mathbb{K}^n$, und falls $(x, Ax) > 0$ für alle $0 \neq x \in \mathbb{K}^n$.*

Man kann sich leicht davon überzeugen, dass für bzgl. $(\cdot,\cdot)$ positiv definite Matrizen A durch $(x, y)_A := (x, Ay)$ ein weiteres Skalarprodukt definiert ist. Die dadurch induzierte Norm $\|x\|_A := \sqrt{(x, Ax)}$ wird als **Energienorm** bezeichnet.

In diesesm Abschnitt werden wir einen anderen Zugang zur Lösung großdimensionierter Gleichungssysteme

$$Ax = b \tag{12.4}$$

mit positiv definiter Matrix $A \in \mathbb{K}^{n\times n}$ und gegebener rechten Seite $b \in \mathbb{K}^n$ kennenlernen. Dazu formulieren wir (12.4) als äquivalentes Minimierungsproblem der Funktion

$$f(y) = \frac{1}{2}(y, Ay) - \operatorname{Re}(y, b).$$

Lemma 12.17. *Die Lösung x von (12.4) ist das eindeutige Minimum von f, und für alle $y \in \mathbb{K}^n$ gilt*

$$f(y) - f(x) = \frac{1}{2}\|y - x\|_A^2.$$

Beweis. Wegen

$$f(y) = \frac{1}{2}(y, y)_A - \operatorname{Re}(y, x)_A = \frac{1}{2}(y - x, y - x)_A - \frac{1}{2}\|x\|_A^2 = \frac{1}{2}\|y - x\|_A^2 - \frac{1}{2}\|x\|_A^2$$

ist f minimal für $y = x$. □

Um das Minimum von f zu finden, verfolgen wir die Strategie, ausgehend von $x_k \in \mathbb{K}^n$ den nächsten Punkt $x_{k+1} \in \mathbb{K}^n$ durch Minimierung von f auf der Geraden durch x_k in einer gegebenen Richtung $p_k \in \mathbb{K}^n$ zu bestimmen (sog. **Liniensuche**), d.h.

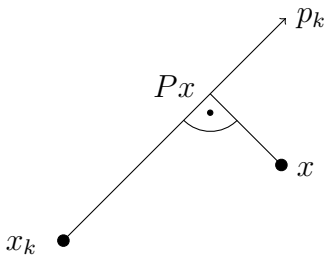
$$x_{k+1} = x_k + \alpha_k p_k, \tag{12.5}$$

wobei die Schrittweite α_k so gewählt ist, dass $f(x_{k+1}) = f(x) + \frac{1}{2}\|x_{k+1} - x\|_A^2$ minimal unter allen $x_k + \alpha p_k$, $\alpha \in \mathbb{K}$, ist.

Lemma 12.18. *Sei $U \subset \mathbb{K}^n$ ein Unterraum und $(\cdot,\cdot)$ ein Skalarprodukt mit induzierter Norm $\|\cdot\|$. Ist $P : X \to U$ eine Abbildung (sog. **Orthoprojektor**), so dass für alle $x \in \mathbb{K}^n$ gilt $(x - Px, u) = 0$ für alle $u \in U$, dann ist*

$$\|x - Px\| = \min_{u\in U} \|x - u\|.$$

Beweis. Wegen $\|x - u\|^2 = \|x - Px\|^2 + \|Px - u\|^2$ folgt $\|x - Px\| \leq \|x - u\|$ für alle $u \in U$. □



Wegen Lemma 12.18 wählen wir x_{k+1} als die bzgl. $(\cdot,\cdot)_A$ orthogonale Projektion von x auf die Gerade $\{x_k + \alpha p_k, \alpha \in \mathbb{K}\}$. Wir erhalten

$$x_{k+1} = x_k + (\frac{p_k}{\|p_k\|_A}, x - x_k)_A \frac{p_k}{\|p_k\|_A}$$

und somit

$$\alpha_k = \frac{(p_k, x - x_k)_A}{(p_k, p_k)_A} = \frac{(p_k, Ax - Ax_k)}{(p_k, Ap_k)} = \frac{(p_k, r_k)}{(p_k, Ap_k)}$$

mit dem **Residuum** $r_k = b - Ax_k$. Der Ausdruck $\|r_k\|$ ist ein Maß für den Fehler, und $\|r_k\| = 0$ impliziert $Ax_k = b$. Obige Wahl von α_k stellt sicher, dass $\{f(x_k)\}_{k\in\mathbb{N}}$ eine monoton fallende Folge ist. Denn es gilt mit (12.5)

$$\begin{aligned} f(x_k) - f(x_{k+1}) &= \frac{1}{2}(x_k, x_k)_A - \mathrm{Re}\,(x_k, x)_A - \frac{1}{2}(x_{k+1}, x_{k+1})_A + \mathrm{Re}\,(x_{k+1}, x)_A \\ &= \frac{1}{2}\frac{|(p_k, x - x_k)_A|^2}{\|p_k\|_A^2} \geq 0. \end{aligned}$$

Im Folgenden betrachten wir zwei Verfahren durch spezielle Wahl der Suchrichtung p_k.

Gradientenverfahren

Für dieses Verfahren nehmen wir an, dass $\mathbb{K} = \mathbb{R}$ und $(x, y) := x^H y$ das euklidische Skalarprodukt ist. Sei die Suchrichtung

$$p_k = -\nabla f(x_k) = -Ax_k + b = r_k$$

in Richtung des steilsten Abstieges gewählt. Dann gilt

$$\alpha_k = \frac{\|r_k\|^2}{(r_k, Ar_k)} \quad \text{und} \quad r_{k+1} = b - Ax_{k+1} = b - Ax_k - \alpha_k Ar_k = r_k - \alpha_k Ar_k.$$

Algorithmus 12.19 (Methode des steilsten Abstiegs).
`Input:` $A \in \mathbb{R}^{n\times n}$ positiv definit, $b, x_0 \in \mathbb{K}^n$ und Fehlertoleranz $\varepsilon > 0$.
`Output:` Folge $\{x_k\}_{k\in\mathbb{N}} \subset \mathbb{R}^n$

$r_0 = b - Ax_0$;
$k = 0$;
`do` {
  $\alpha_k = \frac{\|r_k\|^2}{(r_k, Ar_k)}$;
  $x_{k+1} = x_k + \alpha_k r_k$;
  $r_{k+1} = r_k - \alpha_k Ar_k$;
  $k = k + 1$;
} `while` $(\|r_k\| > \varepsilon)$;

Satz 12.20. *Ist $A \in \mathbb{R}^{n\times n}$ positiv definit, so konvergiert das Gradientenverfahren (Algorithmus 12.19) für jeden Startwert $x_0 \in \mathbb{R}^n$, d.h. es gilt*

$$\|x_{k+1} - x\|_A \leq \frac{\operatorname{cond}(A) - 1}{\operatorname{cond}(A) + 1}\|x_k - x\|_A,$$

wobei $\operatorname{cond}(A) = \|A\|_2\|A^{-1}\|_2$ die Konditionszahl von A bezeichnet.

Beweis. Sei $T_\alpha := I - \alpha A$, $\alpha \in \mathbb{R}$, die Iterationsmatrix des Richardson-Verfahrens aus Beispiel 12.10. Dann gilt

$$\|x_{k+1} - x\|_A \leq \|x_k + \alpha r_k - x\|_A = \|T_\alpha x_k + \alpha b - T_\alpha x - \alpha b\|_A = \|T_\alpha(x_k - x)\|_A.$$

Sei $v_1, \ldots, v_n \in \mathbb{R}^n$ eine Orthonormalbasis aus Eigenvektoren von A und $\lambda_1 \leq \ldots \leq \lambda_n$ die zugehörigen Eigenwerte. Für $y = \sum_{i=1}^n c_i v_i$ gilt

$$\begin{aligned}\|T_\alpha y\|_A^2 = (T_\alpha y, AT_\alpha y) &= \sum_{i,j=1}^n c_i(1-\alpha\lambda_i)c_j\lambda_j(1-\alpha\lambda_j)\underbrace{(v_i, v_j)}_{\delta_{ij}} \\ &= \sum_{i=1}^n c_i^2\lambda_i|1-\alpha\lambda_i|^2 \leq \underbrace{\max_{i=1,\ldots n}|1-\alpha\lambda_i|^2}_{\rho^2(T_\alpha)}\sum_{i=1}^n c_i^2\lambda_i = \rho^2(T_\alpha)\|y\|_A^2.\end{aligned}$$

Also folgt $\|x_{k+1} - x\|_A \leq \rho(T_\alpha)\|x_k - x\|_A$. Aus Beispiel 12.10 wissen wir bereits, dass für $\alpha = \alpha_{\text{opt}}$

$$\rho(T_{\alpha_{\text{opt}}}) = \frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1} = \frac{\frac{\lambda_n}{\lambda_1} - 1}{\frac{\lambda_n}{\lambda_1} + 1}$$

minimal ist. Die Behauptung folgt aus Satz 6.18, weil

$$\|A\|_2 = \rho(A) = \lambda_n, \quad \|A^{-1}\|_2 = \rho(A^{-1}) = \frac{1}{\lambda_1}.$$

□

Bemerkung. Wegen

$$\frac{\operatorname{cond}(A) - 1}{\operatorname{cond}(A) + 1} = 1 - \frac{2}{\operatorname{cond}(A) + 1} < 1$$

liegt zwar immer Konvergenz vor, die Konvergenzgeschwindigkeit kann bei großer Konditionszahl aber gering sein.

Verfahren der konjugierten Gradienten

Das folgende von Hestenes und Stiefel im Jahr 1952 vorgestellte konjugierte Gradienten-Verfahren (engl. Conjugate Gradients (CG) method) ist wohl das effizienteste bekannte Verfahren zur Lösung linearer Gleichungssysteme $Ax = b$ mit bzgl. eines Skalarproduktes $(\cdot, \cdot)$ positiv definiter Matrix A. Bei diesem Verfahren sind die Suchrichtungen p_k paarweise *konjugierte* Vektoren.

Definition 12.21. *Zwei Vektoren $x, y \in \mathbb{K}^n$ heißen* ***konjugiert bzgl. A und*** $(\cdot,\cdot)$*, falls* $(x,y)_A = (x, Ay) = 0$.

Bemerkung. Sind n Vektoren $v_i \neq 0$, $i = 1, \ldots, n$, paarweise A-konjugiert, so bilden sie eine Basis von $\mathbb{K}^n$. Dies sieht man, weil aus

$$\sum_{i=1}^{n} \beta_i v_i = 0$$

durch Multiplikation mit v_j folgt

$$\sum_{i=1}^{n} \beta_i (v_j, v_i)_A = \beta_j (v_j, v_j)_A$$

und hieraus $\beta_j = 0$, $j = 1, \ldots, n$.

Lemma 12.22. *Seien $p_0, \ldots, p_{n-1}$ paarweise A-konjugierte Vektoren. Dann liefert die durch (12.5) definierte Folge für jedes $x_0 \in \mathbb{K}^n$ nach (höchstens) n Schritten die Lösung $x = A^{-1}b$.*

Beweis. Wegen

$$r_n = r_{n-1} - \alpha_{n-1} A p_{n-1} = r_\ell - \sum_{i=\ell}^{n-1} \alpha_i A p_i$$

für $0 \leq \ell < n$ ergibt sich

$$(p_\ell, r_n) = (p_\ell, r_\ell) - \sum_{i=\ell}^{n-1} \alpha_i (p_\ell, A p_i) = (p_\ell, r_\ell) - \alpha_\ell (p_\ell, A p_\ell) = 0.$$

Weil $\{p_0, \ldots, p_{n-1}\}$ eine Basis von $\mathbb{K}^n$ bildet, ist somit $r_n = 0$. □

In der Regel ist ein A-konjugiertes System $\{p_0, \ldots, p_{n-1}\}$ von vornherein nicht vorhanden. Es kann aber schrittweise auf Basis des Residuums nach folgender Vorschrift gebildet werden:

$$p_0 = r_0, \quad p_{k+1} = r_{k+1} + \gamma_k p_k \quad \text{mit} \quad \gamma_k = -\frac{(p_k, A r_{k+1})}{(p_k, A p_k)}, \ k \geq 0.$$

Lemma 12.23. *Sei $r_j \neq 0$ für $j \leq k$. Dann gilt*

(i) $(p_j, r_k) = 0$ *für alle* $j < k$,

(ii) $(r_j, r_k) = 0$ *für alle* $j < k$,

(iii) die Vektoren $\{p_0, \ldots, p_k\}$ sind paarweise A-konjugiert.

Beweis. Wir bemerken zunächst, dass

$$(p_k, r_{k+1}) = (p_k, r_k - \alpha_k A p_k) = (p_k, r_k) - \alpha_k (p_k, A p_k) = 0 \tag{12.6}$$

nach Wahl von α_k. Wir zeigen die Behauptung per Induktion über k. Für $k = 1$ erhält man (i) und (ii) aus (12.6), (iii) folgt aus

$$(p_0, Ap_1) = (p_0, Ar_1) + \gamma_0 (p_0, Ap_0) = 0.$$

Die Behauptung sei wahr für ein k. Dann erhält man (i) für $j = k$ aus (12.6). Für $0 \leq j < k$ folgt (i) mit der Induktionsannahme aus

$$(p_j, r_{k+1}) = (p_j, r_k - \alpha_k A p_k) = \underbrace{(p_j, r_k)}_{=0} - \alpha_k \underbrace{(p_j, Ap_k)}_{=0} = 0.$$

Wegen $r_j = p_j - \gamma_{j-1} p_{j-1}$, $0 < j < k+1$, erhält man ferner (ii) aus (i). Die A-Konjugiertheit von p_k und p_{k+1} folgt wegen

$$(p_k, Ap_{k+1}) = (p_k, Ar_{k+1}) + \gamma_k (p_k, Ap_k) = 0.$$

Für $0 < j < k$ folgt mit der Induktionsannahme

$$(p_j, Ap_{k+1}) = (p_j, A(r_{k+1} + \gamma_k p_k)) = (p_j, Ar_{k+1}) + \gamma_k (p_j, Ap_k) = (p_j, Ar_{k+1})$$

und wegen (ii)

$$\overline{\alpha}_j (p_j, Ap_{k+1}) = \overline{\alpha}_j (p_j, Ar_{k+1}) = (r_j - r_{j+1}, r_{k+1}) = \underbrace{(r_j, r_{k+1})}_{=0} - \underbrace{(r_{j+1}, r_{k+1})}_{=0} = 0.$$

Dabei kann α_j nicht verschwinden, weil sonst $(r_j, p_j) = (r_{j+1}, p_j) = 0$ und somit

$$0 = (r_j, r_j + \gamma_{j-1} p_{j-1}) = (r_j, r_j) + \gamma_{j-1} \underbrace{(r_j, p_{j-1})}_{=0} = \|r_j\|^2$$

wäre. Dies widerspricht aber der Vorraussetzung. □

Im folgenden Lemma stellen wir eine Beziehung des CG-Verfahren zu dem sog. **Krylov-Raum**

$$\mathcal{K}_k(A, r_0) := \operatorname{span}\{r_0, Ar_0, \ldots, A^{k-1} r_0\}$$

her.

Lemma 12.24. *Sei $r_j \neq 0$ für $j < k$. Dann gilt*

$$\operatorname{span}\{x_1, \ldots, x_k\} = x_0 + \mathcal{K}_k(A, r_0) \tag{12.7}$$

und

$$\operatorname{span}\{p_0, \ldots, p_{k-1}\} = \operatorname{span}\{r_0, \ldots, r_{k-1}\} = \mathcal{K}_k(A, r_0). \tag{12.8}$$

Beweis. Wir zeigen den Beweis per Induktion über k. Für $k = 1$ sind (12.7) und (12.8) offenbar wahr. Aus $p_k = r_k + \gamma_{k-1} p_{k-1}$ erhält man

$$\operatorname{span}\{p_0, \ldots, p_k\} = \operatorname{span}\{r_0, \ldots, r_k\}.$$

Mit $r_k = r_{k-1} - \alpha_{k-1} A p_k$ sieht man

$$\operatorname{span}\{r_0, \ldots, r_k\} = \mathcal{K}_{k+1}(A, r_0).$$

(12.7) folgt nun aus $x_{k+1} = x_k + \alpha_k p_k$. □

Bemerkung.

(a) Man beachte, dass A nur durch die Anwendung auf Vektoren in Verfahren eingeht, die auf Krylov-Räumen basieren.

(b) Die Wahl der Parameter entsprechend Fletcher-Reeves

$$\alpha_k = \frac{(r_k, r_k)}{(p_k, A p_k)}, \quad \gamma_k = \frac{(r_{k+1}, r_{k+1})}{(r_k, r_k)}$$

liefert wegen der Orthogonalitätsbeziehung von Lemma 12.23 ein mathematisch äquivalentes Verfahren, das sich in der Praxis allerdings als stabiler und effizienter erwiesen hat.

Algorithmus 12.25.
`Input:` $A \in \mathbb{K}^{n\times n}$ positiv definit bzgl $(\cdot,\cdot)$, $b, x_0 \in \mathbb{K}^n$ und Fehlertoleranz $\varepsilon > 0$.
`Output:` Folge $\{x_k\}_{k\in\mathbb{N}}$

$p_0 := r_0 = b - A x_0$;
$k = 0$;
`do` {
  $\alpha_k = \frac{(r_k, r_k)}{(p_k, A p_k)}$;
  $x_{k+1} = x_k + \alpha_k p_k$;
  $r_{k+1} = r_k - \alpha_k A p_k$;
  $\gamma_k = \frac{(r_{k+1}, r_{k+1})}{(r_k, r_k)}$;
  $p_{k+1} = r_{k+1} + \gamma_k p_k$;
  $k = k + 1$;
} `while` $(\|r_k\| > \varepsilon)$;

Die Iterierten x_k des CG-Verfahrens erweisen sich als Bestapproximationen an x im Krylov-Raum $\mathcal{K}_k(A, r_0)$.

Lemma 12.26. *Es gilt*

$$\|x_k - x\|_A = \min_{y \in x_0 + \mathcal{K}_k(A, r_0)} \|y - x\|_A.$$

Insbesondere gilt wegen $\mathcal{K}_k(A, r_0) \subset \mathcal{K}_{k+1}(A, r_0)$, dass $\|x_{k+1} - x\|_A \leq \|x_k - x\|_A$. Das CG-Verfahren ist also monoton.

Beweis. Nach Lemma 12.24 wissen wir, dass $x_k \in x_0 + \mathcal{K}_k(A, r_0)$. Für $y \in x_0 + \mathcal{K}_k(A, r_0)$ setze $\delta := x_k - y \in \mathcal{K}_k(A, r_0)$. Dann gilt

$$\|y - x\|_A^2 = \|y - x_k + x_k - x\|_A^2 = \|y - x_k\|_A^2 + \|x_k - x\|_A^2 + 2\,\mathrm{Re}\,(\delta, \underbrace{A(x - x_k)}_{r_k}).$$

Nach Lemma 12.24 ist $\delta \in \mathcal{K}_k(A, r_0) = \mathrm{span}\,\{p_0, \ldots, p_{k-1}\}$, und Lemma 12.23 impliziert $(\delta, r_k) = 0$. Also wird das Minimum von

$$\|y - x\|_A^2 = \|\delta\|_A^2 + \|x_k - x\|_A^2$$

genau für $\delta = 0 \iff y = x_k$ angenommen. □

Weil normalerweise eine Genauigkeit $\varepsilon > 0$ der Iterierten x_k ausreichend ist, werden in Algorithmus 12.25 oft weniger als n Schritte ausgeführt. Um die Anzahl der Schritte, die für eine vorgegebene Genauigkeit benötigt werden, abzuschätzen, geben wir die folgende Fehlerabschätzung an.

Satz 12.27. *Ist $A \in \mathbb{K}^{n\times n}$ positiv definit bzgl. $(\cdot,\cdot)$, so konvergiert das CG-Verfahren und es gilt*

$$\|x_k - x\|_A \le 2\left(\frac{\sqrt{\mathrm{cond}(A)} - 1}{\sqrt{\mathrm{cond}(A)} + 1}\right)^k \|x_0 - x\|_A, \quad k = 1, \ldots, n-1.$$

Beweis. Nach Lemma 12.24 gilt $x_k = x_0 + p_k(A)r_0$ für ein $p_k \in \Pi_{k-1}$ und somit

$$x_k - x = x_0 + p_k(A)A(x - x_0) - x = q(A)(x_0 - x), \quad q(x) := 1 - x\,p_k(x).$$

Es gilt $q \in \tilde{\Pi}_k := \{p \in \Pi_k :\ p(0) = 1\}$. Ferner gilt nach Lemma 12.26

$$\|x_k - x\|_A = \min_{p\in\tilde{\Pi}_k} \|p(A)(x_0 - x)\|_A.$$

Sei $v_1, \ldots, v_n$ eine Orthonormalbasis aus Eigenvektoren zu den Eigenwerten $\lambda_1 \le \ldots \le \lambda_n$ von A und $x_0 - x = \sum_{i=1}^n \alpha_i v_i$. Dann folgt aus

$$p(A)(x_0 - x) = \sum_{i=1}^n \alpha_i p(A) v_i = \sum_{i=1}^n \alpha_i p(\lambda_i) v_i,$$

dass

$$\begin{aligned}
\|p(A)(x_0 - x)\|_A^2 = (p(A)(x_0 - x), p(A)(x_0 - x))_A &= \sum_{i,j=1}^n \overline{\alpha}_i \alpha_j \overline{p(\lambda_i)} p(\lambda_j)(v_i, v_j)_A \\
&= \sum_{i=1}^n |\alpha_i|^2 |p(\lambda_i)|^2 \|v_i\|_A^2 \le \max_{i=1,\ldots,n} |p(\lambda_i)|^2 \sum_{i=1}^n |\alpha_i|^2 \|v_i\|_A^2 \\
&= \max_{i=1,\ldots,n} |p(\lambda_i)|^2 \|x_0 - x\|_A^2
\end{aligned}$$

und somit

$$\|x_k - x\|_A \le \min_{p\in\tilde{\Pi}_k} \max_{i=1,\ldots,n} |p(\lambda_i)|\, \|x_0 - x\|.$$

Daher genügt es, ein Polynom in $\tilde{\Pi}_k$ zu finden, das die gewünschte Abschätzung liefert. Ist $\lambda_n = \lambda_1$, so existiert $p \in \tilde{\Pi}_k$ mit $p(\lambda_1) = 0$, was $\|x_k - x\|_A = 0$ zeigt. Im Fall $\lambda_n > \lambda_1$ verwenden wir nach Satz 10.19 das Tschebyscheff-Polynom

$$\hat{T}_k(x) = \frac{T_k(t(x))}{T_k(t_0)} \in \tilde{\Pi}_k, \quad t(x) = 2\frac{x - \lambda_1}{\lambda_n - \lambda_1} - 1, \quad t_0 = t(0) = \frac{\lambda_n + \lambda_1}{\lambda_n - \lambda_1} > 1.$$

Nach Satz 10.17 (ix) gilt

$$T_k(t_0) \geq \frac{1}{2}\left(t_0 + \sqrt{t_0^2 - 1}\right)^k = \frac{1}{2}\left(\frac{\sqrt{\frac{\lambda_n}{\lambda_1}} + 1}{\sqrt{\frac{\lambda_n}{\lambda_1}} - 1}\right)^k$$

und nach Satz 10.17 (v)

$$\max_{i=1,\ldots,n} |\hat{T}_k(\lambda_i)| = \frac{1}{T_k(t_0)} \leq 2\left(\frac{\sqrt{\text{cond}(A)} - 1}{\sqrt{\text{cond}(A)} + 1}\right)^k.$$

□

Bemerkung. Will man mit Hilfe des CG-Verfahrens die exakte Lösung berechnen, so müssen höchstens n Schritte durchgeführt werden, von denen jeder nur eine Multiplikation von A mit einem Vektor benötigt. Der Gesamtaufwand ist dann von der Ordnung n^3. Ist A schwachbesetzt, so genügen $O(n)$ Operationen pro Iterationsschritt. Ist ferner eine Genauigkeit $\varepsilon > 0$ der Approximation an die Lösung ausreichend, so gilt nach dem letzten Satz $\|x_k - x\|_A \leq 2\gamma^k \|x_0 - x\|_A$ mit einem $\gamma < 1$. Ist $2\gamma^k < \varepsilon \iff k \geq \log_\gamma(\varepsilon/2)$ und ist $x_0 = 0$, so hat x_k, $k \geq \log_\gamma(\varepsilon/2)$, mindestens Genaugkeit ε. In diesem Fall ist die Gesamtkomplexität von der Ordnung $n \log_\gamma \varepsilon$.

12.4 Newton-Verfahren zur Lösung nichtlinearer Gleichungen

Sei $D \subset \mathbb{K}^n$ offen und $f : D \to \mathbb{K}^n$ stetig differenzierbar. In diesem Abschnitt wollen wir die Nullstellen $x \in D$ der nichtlinearen Gleichung

$$f(x) = 0 \tag{12.9}$$

finden. Ist $x_0 \in D$ eine Approximation an die Nullstelle x, dann linearisieren wir (12.9), indem f durch die lineare Funktion

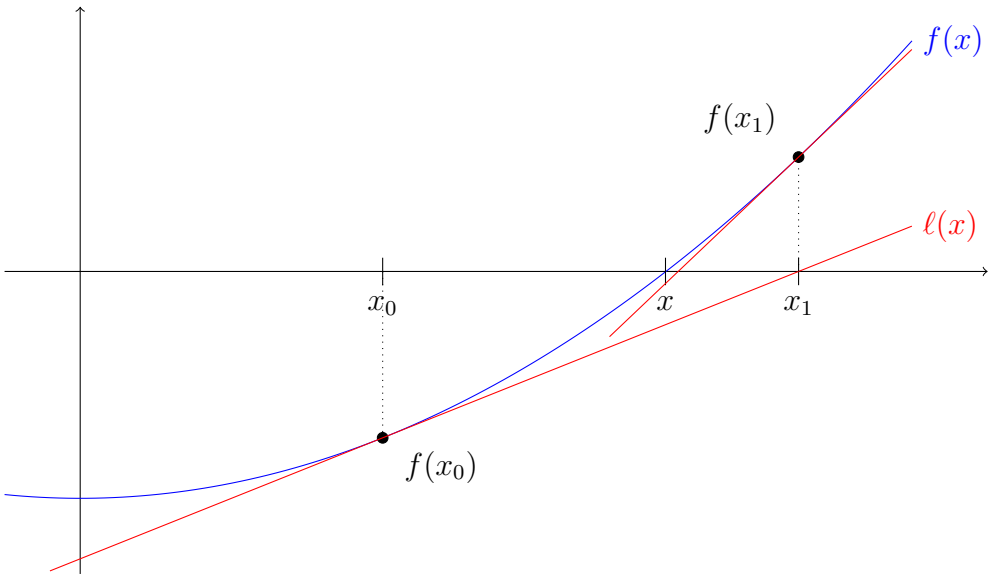
$$\ell(x) := f(x_0) + Df(x_0)(x - x_0)$$

approximiert wird. Existiert $(Df(x_0))^{-1} \in \mathbb{K}^{n\times n}$, so ist

$$x_1 := x_0 - (Df(x_0))^{-1} f(x_0)$$

die Nullstelle von $\ell(x) \approx f(x)$ und unter gewissen Umständen eine bessere Näherung an x als x_0. Setzt man diesen Prozess fort, so erhält man das **Newton-Verfahren**

$$x_{k+1} = x_k - (Df(x_k))^{-1} f(x_k), \quad k = 0, 1, 2, \ldots. \tag{12.10}$$



Im Fall $n = 1$ (eindimensionaler Fall) ist x_{k+1} der Schnittpunkt der Tangenten an f in x_k mit der x-Achse.

Im Folgenden Satz untersuchen wir die Konvergenz des Newton-Verfahrens. Dazu sein $\|\cdot\|$ eine Norm auf $\mathbb{K}^n$ bzw. eine verträgliche Matrizennorm auf $\mathbb{K}^{n\times n}$.

Satz 12.28. *Sei $D \subset \mathbb{K}^n$ offen und konvex, $f : D \to \mathbb{K}^n$ stetig differenzierbar und $x \in D$ eine Nullstelle von f. Es sei $Df(y)$ invertierbar für alle $y \in D$, und es existiere $L > 0$ mit*

$$\|(Df(z))^{-1}(Df(y) - Df(z))\| \le L\|y - z\| \quad \text{für alle } y, z \in D. \tag{12.11}$$

Ist

$$x_0 \in U := \{y \in \mathbb{K}^n : \|y - x\| \le \varepsilon\}, \quad 0 < \varepsilon \le \frac{2}{L},$$

und $U \subset D$, so gilt $x_k \in U$ und

$$\|x_{k+1} - x\| \le \frac{L}{2}\|x_k - x\|^2$$

für $k \ge 0$.

Beweis. Wir gehen zunächst davon aus, dass $x_k \in D$. Aus $f(x) = 0$ erhält man mit (12.10)

$$\begin{aligned} x_{k+1} - x &= x_k - (Df(x_k))^{-1} f(x_k) - x \\ &= x_k - x - (Df(x_k))^{-1}(f(x_k) - f(x)) \\ &= (Df(x_k))^{-1}(f(x) - f(x_k) - Df(x_k)(x - x_k)). \end{aligned}$$

Sei $\gamma(t) = (1 - t)x_k + tx \in D$, $t \in [0, 1]$. Dann gilt $\gamma'(t) = x - x_k$ und somit

$$f(x) - f(x_k) = f(\gamma(1)) - f(\gamma(0)) = \int_0^1 \frac{\mathrm{d}}{\mathrm{d}t}(f \circ \gamma)(t)\,\mathrm{d}t = \int_0^1 Df(\gamma(t))(x - x_k)\,\mathrm{d}t.$$

Wegen $Df(x_k)(x - x_k) = \int_0^1 Df(x_k)(x - x_k)\,\mathrm{d}t$ folgt mit (12.11)

$$\begin{aligned}
\|x_{k+1} - x\| &= \left\| \int_0^1 (Df(x_k))^{-1}\big(Df(\gamma(t)) - Df(x_k)\big)(x - x_k)\,\mathrm{d}t \right\| \\
&\le \int_0^1 \left\|(Df(x_k))^{-1}\big(Df(\gamma(t)) - Df(x_k)\big)\right\| \|x - x_k\|\,\mathrm{d}t \\
&\le L\|x - x_k\| \int_0^1 \|\gamma(t) - x_k\|\,\mathrm{d}t \\
&= L\|x - x_k\| \int_0^1 t\|x - x_k\|\,\mathrm{d}t \\
&= \frac{L}{2}\|x - x_k\|^2.
\end{aligned} \tag{12.12}$$

Die bisherigen Argumente setzen voraus, dass $x_k \in D$. Wir zeigen nun die Behauptung per Induktion über k. Für $k = 0$ ist nach Voraussetzung $x_0 \in U \subset D$. Daher folgt aus (12.12)

$$\|x_1 - x\| \le \frac{L}{2}\|x_0 - x\|^2.$$

Sei die Behauptung für ein k wahr. Dann gilt nach der Induktionsvorraussetzung, dass

$$\|x_{k+1} - x\| \le \frac{L}{2}\|x_k - x\|^2$$

und insbesondere $\|x_{k+1} - x\| \le \varepsilon^2 L/2 \le \varepsilon$, weil $\varepsilon \le 2/L$. Also ist $x_{k+1} \in U \subset D$ und (12.12) zeigt

$$\|x_{k+2} - x\| \le \frac{L}{2}\|x_{k+1} - x\|^2.$$

□

Bemerkung.

(a) Die Bedingung (12.11) ist invariant unter linearen Transformationen

$$f \mapsto Af, \quad A \in \mathbb{K}^{n\times n} \text{ regulär.}$$

(b) Ist $Df(x)$ invertierbar und Df Lipschitz-stetig in einer Umgebung von x, so existiert eine Umgebung U und eine Konstante L, so dass (12.11) erfüllt ist.

(c) In der Newton-Iteration (12.10)

$$x_{k+1} = x_k - \underbrace{(Df(x_k))^{-1} f(x_k)}_{\delta_k}, \quad k = 0, 1, 2, \ldots,$$

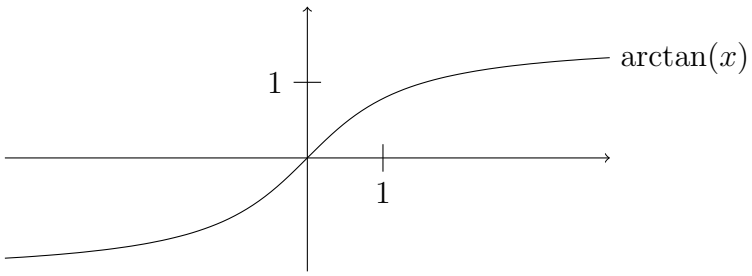
solte δ_k sollte nicht durch Anwendung der Inversen berechnet werden, sondern als Lösung eines linearen Gleichungssystems

$$Df(x_k)\delta_k = f(x_k).$$

Insbesondere erkennt man, dass das Newton-Verfahren ein nichtlineares Gleichungssystem in eine Folge linearer Systeme überführt.

Das folgende Beispiel zeigt, dass die Konvergenz des Newton-Verfahrens wie im letzten Satz bewiesen im Allgemeinen nur lokal ist.

Beispiel 12.29. Betrachte $f(x) = \arctan x$. Es gilt $f(0) = 0$ und $f'(x) = (1 + x^2)^{-1} \neq 0$.



Ferner ist (12.11) erfüllt, weil

$$(f'(y) - f'(z))/f'(z) = (1 + z^2)\left(\frac{1}{1 + y^2} - \frac{1}{1 + z^2}\right) = \frac{z + y}{1 + y^2}(z - x).$$

Ist x_0 so gewählt, dass

$$\arctan |x_0| \geq \frac{2|x_0|}{1 + |x_0|^2},$$

so divergiert das Newton-Verfahren

$$x_{k+1} = x_k - (1 + x_k^2) \arctan x_k, \quad k = 0, 1, 2, \ldots.$$

Index