Algorithmische Mathematik 2
(basierend auf Skript von Andreas Eberle)

Alexander Effland

18. Juni 2025

Inhaltsverzeichnis

Teil I.

Diskrete Stochastik

1. Diskrete Zufallsvariablen

Grundlegende Objekte im axiomatischen Aufbau der Wahrscheinlichkeitstheorie nach Kolmogorov sind die Menge Ω der in einem Modell in Betracht gezogenen *Fälle* ω, die Kollektion $\mathcal{A}$ der betrachteten *Ereignisse* A, sowie die *Wahrscheinlichkeitsverteilung* $\mathbb{P}$, die jedem Ereignis A eine Wahrscheinlichkeit $\mathbb{P}[A]$ zwischen 0 und 1 zuordnet. Dabei sind Ereignisse Teilmengen von Ω, und eine Wahrscheinlichkeitsverteilung ist eine Abbildung von $\mathcal{A}$ nach $[0,1]$. Zudem sind *Zufallsvariablen* X von zentralem Interesse, die jedem Fall ω einen Wert $X(\omega)$ zuweisen. Zur Illustration betrachten wir drei elementare Beispiele bevor wir die genannten Objekte formal definieren.

Beispiel (Würfeln und Münzwürfe).

a) Einmal Würfeln:
Die Menge der möglichen *Fälle* ist $\Omega = \{1,2,3,4,5,6\}$. Die Elemente $\omega \in \Omega$ bezeichnet man auch als *Elementarereignisse* und identifiziert sie mit den einelementigen Mengen $\{\omega\}$. Allgemeine *Ereignisse* werden durch Teilmengen von Ω beschrieben, zum Beispiel:

„Augenzahl ist 3“	$\{3\}$
„Augenzahl ist gerade“	$\{2,4,6\}$
„Augenzahl ist *nicht* gerade“	$\{1,3,5\} = \{2,4,6\}^C$
„Augenzahl ist größer als 3“	$\{4,5,6\}$
„Augenzahl ist gerade *und* größer als 3“	$\{4,6\} = \{2,4,6\} \cap \{4,5,6\}$
„Augenzahl gerade *oder* größer als 3“	$\{2,4,5,6\} = \{2,4,6\} \cup \{4,5,6\}$

Hierbei schreiben wir A^C für das Komplement $\Omega \setminus A$ der Menge A in der vorgegebenen Grundmenge Ω. für die Wahrscheinlichkeiten sollte im Falle eines „fairen“ Würfels gelten:

$$\mathbb{P}[\text{„3“}] = \frac{1}{6},$$
$$\mathbb{P}[\text{„Augenzahl gerade“}] = \frac{\text{Anzahl günstige Fälle}}{\text{Anzahl mögliche Fälle}} = \frac{|\{2,4,6\}|}{|\{1,2,3,4,5,6\}|} = \frac{3}{6} = \frac{1}{2},$$
$$\mathbb{P}[\text{„Augenzahl gerade oder größer als 3“}] = \frac{4}{6} = \frac{2}{3}.$$

Beispiele für *Zufallsvariablen* sind

$$X(\omega) = \omega, \qquad \text{„Augenzahl des Wurfs“, oder}$$
$$G(\omega) = \begin{cases} 1 & \text{falls } \omega \in \{1,2,3,4,5\}, \\ -5 & \text{falls } \omega = 6, \end{cases} \qquad \text{„Gewinn bei einem fairen Spiel“.}$$

In einem anderen (detaillierteren) Modell hätte man die Menge Ω auch anders wählen können, z.B. könnte Ω alle möglichen stabilen Anordnungen des Würfels auf dem Tisch beinhalten. Wir werden später sehen, dass die konkrete Wahl der Menge Ω oft gar nicht wesentlich ist - wichtig sind vielmehr die Wahrscheinlichkeiten, mit denen die relevanten Zufallsvariablen Werte in bestimmten Bereichen annehmen.

b) Endlich viele faire Münzwürfe:
Es ist naheliegend, als Menge der möglichen Fälle

$$\Omega = \{\omega = (x_1, \ldots, x_n) \mid x_i \in \{0,1\}\} = \{0,1\}^n$$

zu betrachten, wobei n die Anzahl der Münzwürfe ist, und 0 für „Kopf“ sowie 1 für „Zahl“ steht. Alle Ausgänge sind genau dann gleich wahrscheinlich, wenn $\mathbb{P}[\{\omega\}] = 2^{-n}$ für alle $\omega \in \Omega$ gilt. Dies wird im folgenden angenommen. Zufallsvariablen von Interesse sind beispielsweise das Ergebnis des i-ten Wurfs

$$X_i(\omega) \;=\; x_i,$$

oder die Häufigkeit

$$S_n(\omega) \;=\; \sum_{i=1}^{n} X_i(\omega)$$

von Zahl bei n Münzwürfen. Das Ereignis „i-ter Wurf ist Kopf“ wird durch die Menge

$$A_i \;=\; \{\omega \in \Omega \mid X_i(\omega) = 0\} \;=\; X_i^{-1}(0)$$

beschrieben. Diese Menge bezeichnen wir in intuitiver Kurznotation auch mit $\{X_i = 0\}$. Es gilt

$$\mathbb{P}[X_i = 0] \;:=\; \mathbb{P}[\{X_i = 0\}] \;=\; \mathbb{P}[A_i] \;=\; \frac{1}{2}.$$

Das Ereignis „genau k-mal Zahl“ wird entsprechend durch die Menge

$$A \;= \{\omega \in \Omega \mid S_n(\omega) = k\} \;=\; \{S_n = k\}$$

beschrieben und hat die Wahrscheinlichkeit

$$\mathbb{P}[S_n = k] \;=\; \binom{n}{k} 2^{-n}.$$

c) Unendlich viele Münzwürfe:
Hier kann man als Menge der möglichen Fälle den Raum

$$\Omega \;=\; \{\omega = (x_1, x_2, \ldots) \mid x_i \in \{0,1\}\} \;=\; \{0,1\}^{\mathbb{N}}$$

aller binären Folgen ansetzen. Diese Menge ist überabzählbar, da die durch die Dualdarstellung reeller Zahlen definierte Abbildung

$$(x_1, x_2, \ldots) \;\mapsto\; \sum_{i=1}^{\infty} x_i \cdot 2^{-i}$$

von Ω nach $[0,1]$ surjektiv ist. Dies hat zur Folge, dass es nicht möglich ist, *jeder* Teilmenge von Ω in konsistenter Weise eine Wahrscheinlichkeit zuzuordnen. Die formale Definition von Ereignissen und Wahrscheinlichkeiten ist daher in diesem Fall aufwändiger, und wird erst in der Vorlesung Einführung in die Wahrscheinlichkeitstheorie systematisch behandelt.

1.1. Ereignisse und ihre Wahrscheinlichkeit

Wir werden nun die Kolmogorovsche Definition eines Wahrscheinlichkeitsraums motivieren und formulieren, erste einfache Folgerungen daraus ableiten, und elementare Beispiele betrachten. Ein Wahrscheinlichkeitsraum besteht aus einer nichtleeren Menge Ω, die bis auf weiteres fest gewählt sei, einer Kollektion $\mathcal{A}$ von Teilmengen von Ω (den Ereignissen) und einer Abbildung $\mathbb{P} : \Omega \to [0,1]$, die bestimmte Axiome erfüllen.

Ereignisse als Mengen

Seien A, B, und A_i, $i \in I$, Ereignisse, d.h. Teilmengen von Ω. Hierbei ist I eine beliebige Indexmenge. Anschaulich stellen wir uns vor, dass ein Element $\omega \in \Omega$ zufällig ausgewählt wird, und das Ereignis A eintritt, falls ω in A enthalten ist. „Zufällig“ bedeutet dabei nicht unbedingt, dass alle Fälle gleich wahrscheinlich sind ! Wir werden manchmal auch die folgenden Notationen für die Menge A verwenden:

$$A = \{\omega \in \Omega \mid \omega \in A\} = \{\omega \in A\} = \{\text{ „}A \text{ tritt ein“ }\}.$$

Da Ereignisse durch Mengen beschrieben werden, können wir mengentheoretische Operationen benutzen, um mehrere Ereignisse zu kombinieren. Wir wollen uns überlegen, was Ereignisse wie A^C, $A \cup B$, $\bigcap_{i\in I} A_i$ usw. anschaulich bedeuten. Um dies herauszufinden, betrachtet man einen möglichen Fall ω und untersucht, wann dieser eintritt. Beispielsweise gilt

$$\omega \in A \cup B \qquad \Leftrightarrow \qquad \omega \in A \text{ oder } \omega \in B,$$

also in anschaulicher Sprechweise:

$$\text{„}A \cup B \text{ tritt ein“} \qquad \Leftrightarrow \qquad \text{„}A \text{ tritt ein oder } B \text{ tritt ein“}.$$

Entsprechend gilt

$$\omega \in \bigcup_{i\in I} A_i \qquad \Leftrightarrow \qquad \text{es gibt ein } i \in I \text{ mit } \omega \in A_i,$$

also

$$\text{„}\textstyle\bigcup_{i\in I} A_i \text{ tritt ein“} \qquad \Leftrightarrow \qquad \text{„mindestens eines der Ereignisse } A_i \text{ tritt ein“}.$$

Auf analoge Weise überlegen wir uns die Bedeutungen der folgenden Mengenoperationen:

$A \cap B$	„A und B treten ein“,
$\bigcap_{i\in I} A_i$	„jedes der A_i tritt ein“,
$A^C = \Omega \setminus A$	„A tritt nicht ein“,
$A = \emptyset$	„unmögliches Ereignis“ (tritt nie ein),
$A = \Omega$	„sicheres Ereignis“ (tritt immer ein),
$A = \{\omega\}$	„Elementarereignis“ (tritt nur im Fall ω ein).

Die Kollektion $\mathcal{A}$ aller im Modell zugelassenen bzw. in Betracht gezogenen Ereignisse besteht aus Teilmengen von Ω, d.h $\mathcal{A}$ ist eine Teilmenge der *Potenzmenge*

$$\mathcal{P}(\Omega) \;=\; \{A \mid A \subseteq \Omega\}$$

Die Kollektion $\mathcal{A}$ sollte unter den oben betrachteten Mengenoperationen (Vereinigungen, Durchschnitte, Komplementbildung) abgeschlossen sein. Genauer fordern wir die Abgeschlossenheit nur unter abzählbaren Vereinigungen und Durchschnitten, da $\mathcal{A}$ andernfalls immer gleich der Potenzmenge sein müsste sobald alle einelementigen Mengen enthalten sind. Eine effiziente Formulierung der Abgeschlossenheit unter abzählbaren Mengenoperationen führt auf die folgende Definition:

Definition 1.1. Eine Kollektion $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ von Teilmengen von Ω heißt **σ-Algebra**, falls gilt:

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

(ii) für alle $A \in \mathcal{A}$ gilt: $A^C \in \mathcal{A}$,

(iii) für $A_1, A_2, \ldots \in \mathcal{A}$ gilt: $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$.

Bemerkung. Aus der Definition folgt bereits, dass eine σ-Algebra $\mathcal{A}$ unter allen oben betrachteten endlichen und abzählbar unendlichen Mengenoperationen abgeschlossen ist, denn:

(a) Nach (i) und (ii) ist $\emptyset = \Omega^C \in \mathcal{A}$.

(b) Sind $A_1, A_2, \ldots \in \mathcal{A}$, dann folgt nach (ii) und (iii): $\bigcap_{i=1}^{\infty} A_i = (\bigcup_{i=1}^{\infty} A_i^C)^C \in \mathcal{A}$.

(c) Sind $A, B \in \mathcal{A}$, dann folgt nach (iii) und (a): $A \cup B = A \cup B \cup \emptyset \cup \emptyset \cup \ldots \in \mathcal{A}$.

(d) Entsprechend folgt $A \cap B \in \mathcal{A}$ aus (b) und (i).

Beispiele. a) Potenzmenge.
Die Potenzmenge $\mathcal{A} = \mathcal{P}(\Omega)$ ist stets eine σ-Algebra. In diskreten Modellen, in denen Ω abzählbar ist, werden wir diese σ-Algebra häufig verwenden. Bei nichtdiskreten Modellen kann man dagegen *nicht* jede Wahrscheinlichkeitsverteilung $\mathbb{P}$ auf einer σ-Algebra $\mathcal{A} \subset \mathcal{P}(\Omega)$ zu einer Wahrscheinlichkeitsverteilung auf $\mathcal{P}(\Omega)$ erweitern, siehe Beispiel c).

b) Partielle Information.
Wir betrachten das Modell für n Münzwürfe mit

$$\Omega = \{\omega = (x_1, \ldots, x_n) \mid x_i \in \{0,1\}\} = \{0,1\}^n.$$

Sei $k \leq n$. Dann ist die Kollektion $\mathcal{F}_k$ aller Mengen $A \subseteq \Omega$, die sich in der Form

$$A = \{(x_1, \ldots, x_n) \in \Omega \mid (x_1, \ldots, x_k) \in B\} = B \times \{0,1\}^{n-k}$$

mit $B \subseteq \{0,1\}^k$ darstellen lassen, eine σ-Algebra. Die Ereignisse in der σ-Algebra $\mathcal{F}_k$ sind genau diejenigen, von denen wir schon wissen ob sie eintreten oder nicht, wenn wir nur den Ausgang der ersten k Münzwürfe kennen. Die σ-Algebra $\mathcal{F}_k$ beschreibt also die *Information aus den ersten k Münzwürfen*.

c) Borelsche σ-Algebra. Man kann zeigen, dass es auf der Potenzmenge des reellen Intervalls $\Omega = [0,1]$ keine Wahrscheinlichkeitsverteilung $\mathbb{P}$ gibt, die jedem Teilintervall (a,b) die Länge als Wahrscheinlichkeit zuordnet. Andererseits gibt es eine kleinste σ-Algebra $\mathcal{B}$, die alle Teilintervalle enthält. Auf der σ-Algebra $\mathcal{B}$ existiert eine *kontinuierliche Gleichverteilung* mit der gerade beschriebenen Eigenschaft, siehe Analysis III. Sie enthält zwar alle offenen und alle abgeschlossenen Teilmengen von $[0,1]$, ist aber echt kleiner als die Potenzmenge $\mathcal{P}([0,1])$.

Wahrscheinlichkeitsverteilungen

Sei Ω eine nichtleere Menge und $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ eine σ-Algebra. Wir wollen nun die Abbildung $\mathbb{P}$ einführen, die jedem Ereigniss $A \in \mathcal{A}$ eine Wahrscheinlichkeit $\mathbb{P}[A]$ zuordnet. Welche Bedingungen (Axiome) sollten wir von $\mathbb{P}$ fordern ? Sind $A, B \in \mathcal{A}$ Ereignisse, dann ist $A \cup B$ ein Ereignis, welches genau dann eintritt, wenn A eintritt oder B eintritt. Angenommen, die beiden Ereignisse A und B *treten nicht gleichzeitig ein*, d.h. die Mengen A und B sind *disjunkt*. Dann sollte die Wahrscheinlichkeit von $A \cup B$ die Summe der Wahrscheinlichkeiten von A und B sein:

$$A \cap B = \emptyset \qquad \Rightarrow \qquad \mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B],$$

d.h. die Abbildung $\mathbb{P}$ ist *additiv*. Wir fordern etwas mehr, nämlich dass eine entsprechende Eigenschaft sogar für *abzählbar* unendliche Vereinigungen von disjunkten Mengen gilt. Dies wird sich als wichtig erweisen, um zu einer leistungsfähigen Theorie zu gelangen, die zum Beispiel Konvergenzaussagen für Folgen von Zufallsvariablen liefert.

Definition 1.2 (Axiome von Kolmogorov). Eine Abbildung $\mathbb{P} : \mathcal{A} \to [0, \infty]$, $A \mapsto \mathbb{P}[A]$, heißt **Wahrscheinlichkeitsverteilung** auf $(\Omega, \mathcal{A})$, falls gilt:

(i) $\mathbb{P}$ ist *„normiert“*, d.h.
$$\mathbb{P}[\Omega] = 1,$$

(ii) $\mathbb{P}$ ist *„σ-additiv“*, d.h. für Ereignisse $A_1, A_2, \ldots \in \mathcal{A}$ mit $A_i \cap A_j = \emptyset$ für $i \neq j$ gilt:
$$\mathbb{P}\Big[\bigcup_{i=1}^{\infty} A_i\Big] = \sum_{i=1}^{\infty} \mathbb{P}[A_i].$$

Ein **Wahrscheinlichkeitsraum** $(\Omega, \mathcal{A}, \mathbb{P})$ besteht aus einer nichtleeren Menge Ω, einer σ-Algebra $\mathcal{A} \subseteq \mathcal{P}(\Omega)$, und einer Wahrscheinlichkeitsverteilung $\mathbb{P}$ auf $(\Omega, \mathcal{A})$.

Bemerkung (Maße). Gilt nur Eigenschaft (ii) und $\mathbb{P}[\emptyset] = 0$, dann heißt $\mathbb{P}$ ein *Maß*. Eine Wahrscheinlichkeitsverteilung ist ein normiertes Maß, und wird daher auch äquivalent als **Wahrscheinlichkeitsmaß** bezeichnet. Maße spielen auch in der Analysis eine große Rolle, und werden in der Vorlesung Analysis III systematisch behandelt.

Man kann sich fragen, weshalb wir die Additivität nicht für beliebige Vereinigungen fordern. Würden wir dies tun, dann gäbe es nicht viele interessante Wahrscheinlichkeitsverteilungen auf kontinuierlichen Räumen. Beispielsweise sollte unter der Gleichverteilung auf dem Intervall $[0, 1]$ jede Menge, die nur aus einem Punkt besteht, die Wahrscheinlichkeit 0 haben, da sie in beliebig kleinen Intervallen enthalten ist. Würde Additivität für beliebige Vereinigungen gelten, dann müsste auch das ganze Intervall $[0, 1]$ Wahrscheinlichkeit 0 haben, da es die Vereinigung seiner einelementigen Teilmengen ist. Die Forderung der σ-Additivität liefert also einen angemessenen Kompromiss, der genügend viele interessante Modelle zulässt und es gleichzeitig ermöglicht, sehr weitreichende Aussagen herzuleiten.

Der folgende Satz zeigt, dass Wahrscheinlichkeitsverteilungen einige elementare Eigenschaften besitzen, die wir von der Anschauung her erwarten würden:

Satz 1.3 (Elementare Eigenschaften und erste Rechenregeln).
Ist $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum, dann gelten die folgenden Aussagen:

(i) $\mathbb{P}[\emptyset] = 0$,

(ii) für $A, B \in \mathcal{A}$ mit $A \cap B = \emptyset$ gilt
$$\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] \qquad \text{„endliche Additivität“}.$$

(iii) für $A, B \in \mathcal{A}$ mit $A \subseteq B$ gilt:
$$\mathbb{P}[B] = \mathbb{P}[A] + \mathbb{P}[B \backslash A].$$
Insbesondere folgt
$$\begin{aligned} \mathbb{P}[A] &\leq \mathbb{P}[B], && \text{„Monotonie“}, \\ \mathbb{P}[A^C] &= 1 - \mathbb{P}[A], && \text{„Gegenereignis“}, \\ \mathbb{P}[A] &\leq 1. \end{aligned}$$

(iv) für beliebige Ereignisse $A, B \in \mathcal{A}$ gilt

$$\mathbb{P}[A \cup B] \;=\; \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B] \;\leq\; \mathbb{P}[A] + \mathbb{P}[B].$$

Beweis. (i) Wegen der σ-Additivität von $\mathbb{P}$ gilt

$$1 \;=\; \mathbb{P}[\Omega] \;=\; \mathbb{P}[\Omega \cup \emptyset \cup \emptyset \cup \ldots] \;=\; \underbrace{\mathbb{P}[\Omega]}_{=1} + \underbrace{\mathbb{P}[\emptyset]}_{\geq 0} + \mathbb{P}[\emptyset] + \ldots,$$

und damit $\mathbb{P}[\emptyset] = 0$.

(ii) für disjunkte Ereignisse $A, B \in \mathcal{A}$ folgt aus der σ-Additivität und mit (i)

$$\begin{aligned}\mathbb{P}[A \cup B] \;&=\; \mathbb{P}[A \cup B \cup \emptyset \cup \emptyset \cup \ldots] \\ &=\; \mathbb{P}[A] + \mathbb{P}[B] + \mathbb{P}[\emptyset] + \mathbb{P}[\emptyset] + \ldots \\ &=\; \mathbb{P}[A] + \mathbb{P}[B].\end{aligned}$$

(iii) Gilt $A \subseteq B$, dann ist $B = A \cup (B \backslash A)$. Da diese Vereinigung disjunkt ist, folgt mit (ii)

$$\mathbb{P}[B] \;=\; \mathbb{P}[A] + \mathbb{P}[B \backslash A] \;\geq\; \mathbb{P}[A].$$

Insbesondere ist $1 = \mathbb{P}[\Omega] = \mathbb{P}[A] + \mathbb{P}[A^C]$, und somit $\mathbb{P}[A] \leq 1$.

(iv) für beliebige Ereignisse $A, B \in \mathcal{A}$ gilt nach (iii) gilt:

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

■

Aussage (iv) des Satzes lässt sich für endlich viele Ereignisse verallgemeinern. Beispielsweise folgt durch mehrfache Anwendung von (iv) für die Vereinigung von drei Ereignissen

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

Mit vollständiger Induktion ergibt sich eine Formel für die Wahrscheinlichkeit der Vereinigung einer beliebigen endlichen Anzahl von Ereignissen:

Korollar 1.4 (Einschluss-/Ausschlussprinzip). Für $n \in \mathbb{N}$ und Ereignisse $A_1, \ldots, A_n \in \mathcal{A}$ gilt:

$$\mathbb{P}[\underbrace{A_1 \cup A_2 \cup \ldots \cup A_n}_{\text{„eines der } A_i \text{ tritt ein“}}] \;=\; \sum_{k=1}^{n} (-1)^{k-1} \sum_{1 \leq i_1 < \ldots < i_k \leq n} \mathbb{P}[\; \underbrace{A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}}_{\text{„}A_{i_1}, A_{i_2}, \ldots \text{ und } A_{i_k} \text{ treten ein“}} \;].$$

Das Einschluss-/Ausschlussprinzip werden wir auf eine elegantere Weise am Ende dieses Kapitels beweisen (siehe Satz 1.18).

Diskrete Wahrscheinlichkeitsverteilungen

Ein ganz einfaches Beispiel für eine diskrete Wahrscheinlichkeitsverteilung ist das Grundmodell für einen Münzwurf oder ein allgemeineres 0-1-Experiment mit Erfolgswahrscheinlichkeit $p \in [0,1]$. Hier ist $\Omega = \{0,1\}$, $\mathcal{A} = \mathcal{P}(\Omega) = \{\emptyset, \{0\}, \{1\}, \Omega\}$, und $\mathbb{P}$ ist gegeben durch

$$\begin{aligned} \mathbb{P}[\{1\}] &= p, & \mathbb{P}[\emptyset] &= 0, \\ \mathbb{P}[\{0\}] &= 1-p, & \mathbb{P}[\Omega] &= 1. \end{aligned}$$

Die Verteilung $\mathbb{P}$ nennt man auch eine *(einstufige) Bernoulliverteilung* mit Parameter p.

Auf analoge Weise erhalten wir Wahrscheinlichkeitsverteilungen auf endlichen oder abzählbar unendlichen Mengen Ω. In diesem Fall können wir die Potenzmenge $\mathcal{P}[\Omega]$ als σ-Algebra verwenden, und Wahrscheinlichkeiten von beliebigen Ereignissen aus den Wahrscheinlichkeiten der Elementarereignisse berechnen.

Satz 1.5. (i) Sei $0 \le p(\omega) \le 1$, $\sum_{\omega\in\Omega} p(\omega) = 1$, eine Gewichtung der möglichen Fälle. Dann ist durch

$$\mathbb{P}[A] := \sum_{\omega\in A} p(\omega), \qquad (A \subseteq \Omega),$$

eine Wahrscheinlichkeitsverteilung auf $(\Omega, \mathcal{P}(\Omega))$ definiert.

(ii) Umgekehrt ist jede Wahrscheinlichkeitsverteilung $\mathbb{P}$ auf $(\Omega, \mathcal{P}(\Omega))$ von dieser Form mit

$$p(\omega) = \mathbb{P}[\{\omega\}] \qquad (\omega \in \Omega).$$

Definition 1.6. Die Funktion $p : \Omega \to [0,1]$ heißt **Massenfunktion** („probability mass function“) der diskreten Wahrscheinlichkeitsverteilung $\mathbb{P}$.

Für den Beweis des Satzes brauchen wir einige Vorbereitungen. Wir bemerken zunächst, dass für eine abzählbare Menge A die Summe der Gewichte $p(\omega)$ über $\omega \in A$ definiert ist durch

$$\sum_{\omega\in A} p(\omega) := \sum_{i=1}^{\infty} p(\omega_i), \tag{1.1}$$

wobei $\omega_1, \omega_2, \ldots$ eine beliebige Abzählung von A ist. Da die Gewichte nichtnegativ sind, existiert die Summe auf der rechten Seite (wobei der Wert $+\infty$ zugelassen ist). Der erste Teil des folgenden Lemmas zeigt, dass die Summe über $\omega \in A$ durch (1.1) wohldefiniert ist:

Lemma 1.7. *(i) Unabhängig von der gewählten Abzählung gilt*

$$\sum_{\omega\in A} p(\omega) = \sup_{\substack{F\subseteq A \\ |F|<\infty}} \sum_{\omega\in F} p(\omega). \tag{1.2}$$

Insbesondere hängt die Summe monoton *von A ab, d.h. für $A \subseteq B$ gilt*

$$\sum_{\omega\in A} p(\omega) \le \sum_{\omega\in B} p(\omega). \tag{1.3}$$

(ii) Ist $A = \bigcup_{i=1}^{\infty} A_i$ eine disjunkte Zerlegung, dann gilt:

$$\sum_{\omega \in A} p(\omega) \;=\; \sum_{i=1}^{\infty} \sum_{\omega \in A_i} p(\omega).$$

Beweis. (i) Sei $\omega_1, \omega_2, \ldots$ eine beliebige Abzählung von A. Aus $p(\omega_i) \geq 0$ für alle $i \in \mathbb{N}$ folgt, dass die Folge der Partialsummen $\sum_{i=1}^{n} p(\omega_i)$ monoton wachsend ist. Somit gilt

$$\sum_{i=1}^{\infty} p(\omega_i) = \sup_{n \in \mathbb{N}} \sum_{i=1}^{n} p(\omega_i).$$

Falls die Folge der Partialsummen von oben beschränkt ist, existiert dieses Supremum in $[0, \infty)$. Andernfalls divergiert die Folge der Partialsummen bestimmt gegen $+\infty$. Zu zeigen bleibt

$$\sup_{n \in \mathbb{N}} \sum_{i=1}^{n} p(\omega_i) \;=\; \sup_{\substack{F \subseteq A \\ |F| < \infty}} \sum_{\omega \in F} p(\omega).$$

Wir zeigen zunächst „≤“, und Anschließend „≥“:

„≤“: für alle $n \in \mathbb{N}$ gilt:

$$\sum_{i=1}^{n} p(\omega_i) \;\leq\; \sup_{\substack{F \subseteq A \\ |F| < \infty}} \sum_{\omega \in F} p(\omega),$$

da das Supremum auch über $F = \{\omega_1, \ldots, \omega_n\}$ gebildet wird. Damit folgt „≤“.

„≥“: Ist F eine endliche Teilmenge von A, dann gibt es ein $n \in \mathbb{N}$, so dass $F \subseteq \{\omega_1, \ldots, \omega_n\}$. Daher gilt

$$\sum_{\omega \in F} p(\omega) \;\leq\; \sum_{i=1}^{n} p(\omega_i) \;\leq\; \sup_{n \in \mathbb{N}} \sum_{i=1}^{n} p(\omega_i),$$

und es folgt „≥“.

(ii) Falls A endlich ist, dann gilt $A_i \neq \emptyset$ nur für endlich viele $i \in \mathbb{N}$ und alle A_i sind endlich. Die Behauptung folgt dann aus dem Kommutativ- und dem Assoziativgesetz. Wir nehmen nun an, dass A abzählbar unendlich ist. In diesem Fall können wir die Aussage aus der Aussage für endliche A unter Verwendung von (i) herleiten. Wir zeigen erneut „≤“ und „≥“ separat:

„≤“: Ist F eine endliche Teilmenge von A, so ist $F = \bigcup_{i=1}^{\infty} (F \cap A_i)$. Da diese Vereinigung wieder disjunkt ist, folgt mit σ-Additivität und Gleichung (1.3):

$$\sum_{\omega \in F} p(\omega) \;=\; \sum_{i=1}^{\infty} \sum_{\omega \in F \cap A_i} p(\omega) \;\leq\; \sum_{i=1}^{\infty} \sum_{\omega \in A_i} p(\omega).$$

Also folgt nach (i) auch:

$$\sum_{\omega \in A} p(\omega) \;=\; \sup_{\substack{F \subseteq A \\ |F| < \infty}} \sum_{\omega \in F} p(\omega) \;\leq\; \sum_{i=1}^{\infty} \sum_{\omega \in A_i} p(\omega).$$

„≥“: Seien $F_i \subseteq A_i$ endlich. Da die F_i wieder disjunkt sind, folgt mit σ-Additivität und Gleichung (1.3) für alle $n \in \mathbb{N}$:

$$\sum_{i=1}^{n} \sum_{\omega \in F_i} p(\omega) \;=\; \sum_{\omega \in \bigcup_{i=1}^{n} F_i} p(\omega) \;\le\; \sum_{\omega \in A} p(\omega).$$

Nach (i) folgt dann auch

$$\sum_{i=1}^{n} \sum_{\omega \in A_i} p(\omega) \;\le\; \sum_{\omega \in A} p(\omega),$$

und damit die Behauptung für $n \to \infty$. ■

Beweis (Beweis von Satz 1.5). (i) Nach Voraussetzung gilt $\mathbb{P}[A] \ge 0$ für alle $A \subseteq \Omega$ und $\mathbb{P}[\Omega] = \sum_{\omega \in \Omega} p(\omega) = 1$. Seien nun A_i $(i \in \mathbb{N})$ disjunkt. Dann folgt aus Lemma 1.7.(ii):

$$\mathbb{P}\Big[\bigcup_{i=1}^{\infty} A_i\Big] \;=\; \sum_{\omega \in \bigcup A_i} p(\omega) \;=\; \sum_{i=1}^{\infty} \sum_{\omega \in A_i} p(\omega) \;=\; \sum_{i=1}^{\infty} \mathbb{P}[A_i],$$

also die σ-Additivität von $\mathbb{P}$.

(ii) Umgekehrt folgt aus der σ-Additivität von $\mathbb{P}$ für $A \subseteq \Omega$ sofort

$$\mathbb{P}[A] \;=\; \mathbb{P}\Big[\underbrace{\bigcup_{\omega \in A} \{\omega\}}_{\text{disjunkt}}\Big] \;=\; \sum_{\omega \in A} \mathbb{P}[\{\omega\}].$$

Gleichverteilungen (Laplace-Modelle)

Ist Ω endlich, dann existiert auf $\mathcal{A} = \mathcal{P}(\Omega)$ eine eindeutige Wahrscheinlichkeitsverteilung $\mathbb{P}$ mit konstanter Massenfunktion

$$p(\omega) \;=\; \frac{1}{|\Omega|} \qquad \text{für alle } \omega \in \Omega.$$

Als Wahrscheinlichkeit eines Ereignisses $A \subseteq \Omega$ ergibt sich

$$\mathbb{P}[A] \;=\; \sum_{\omega \in A} \frac{1}{|\Omega|} \;=\; \frac{|A|}{|\Omega|} \;=\; \frac{\text{Anzahl „günstiger“ Fälle}}{\text{Anzahl aller Fälle}}. \tag{1.4}$$

Die Verteilung $\mathbb{P}$ heißt *Gleichverteilung* auf Ω und wird auch mit $\mathrm{Unif}(\Omega)$ bezeichnet. Laplace (1814) benutzte (1.4) als Definition von Wahrscheinlichkeiten. Dabei ist zu beachten, dass die Gleichverteilung nicht erhalten bleibt, wenn man zum Beispiel mehrere Fälle zu einem zusammenfasst. Der Laplacesche Ansatz setzt also voraus, dass man eine Zerlegung in gleich wahrscheinliche Fälle finden kann.

Beispiele. a) n FAIRE MÜNZWÜRFE:
Die Gleichverteilung $\mathrm{Unif}(\Omega)$ auf $\Omega = \{0,1\}^n$ hat die Massenfunktion

$$p(\omega) = \frac{1}{2^n}.$$

Die gleich wahrscheinlichen Fälle sind hier die 2^n möglichen Münzwurfsequenzen.

b) Zufällige Permutationen:
Sei $\Omega = \mathcal{S}_n$ die Menge aller Bijektionen $\omega : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}$. Der 1 können n verschiedene Zahlen zugeordnet geordnet werden, der 2 die verbleibenden $n-1$, usw. Somit gibt es insgesamt $n! = n \cdot (n-1) \cdot (n-2) \cdot \cdots \cdot 1$ dieser Permutationen. Bezüglich der Gleichverteilung auf $\mathcal{S}_n$ gilt also

$$\mathbb{P}[A] = \frac{|A|}{n!} \qquad \text{für alle } A \subseteq \mathcal{S}_n.$$

Anschauliche Beispiele für zufällige Permutationen sind die Anordnung eines gemischten Kartenspiels, oder das zufällige Vertauschen von n Hüten oder Schlüsseln. In letzterem Beispiel gilt:

$$\mathbb{P}[\text{„der } k\text{-te Schlüssel passt auf Schloss } i\text{“}] = \mathbb{P}[\{\omega \in \mathcal{S}_n \mid \omega(i) = k\}] = \frac{(n-1)!}{n!} = \frac{1}{n}.$$

Wie groß ist die Wahrscheinlichkeit, dass einer der Schlüssel sofort passt? Das Ereignis „Schlüssel i passt“ wird beschrieben durch die Menge

$$A_i \;=\; \{\omega \mid \omega(i) = i\} \;=\; \{\text{„}i \text{ ist Fixpunkt von } \omega\text{“ }\}.$$

Die Wahrscheinlichkeit für das Ereignis „ein Schlüssel passt“ lässt sich dann nach dem Einschluss-/Ausschlussprinzip (Satz 1.18) berechnen:

$$\begin{aligned}
&\mathbb{P}[\text{„es gibt mindestens einen Fixpunkt“}] \;=\; \mathbb{P}[A_1 \cup A_2 \cup \ldots \cup A_n] \\
&= \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \le i_1 < i_2 < \ldots < i_k \le n} \mathbb{P}[A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}] \\
&= \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \le i_1 < i_2 < \ldots < i_k \le n} \frac{(n-k)!}{n!} \\
&= \sum_{k=1}^{n} (-1)^{k+1} \binom{n}{k} \frac{(n-k)!}{n!} \;=\; -\sum_{k=1}^{n} \frac{(-1)^k}{k!}
\end{aligned}$$

Hierbei haben wir benutzt, dass es $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ k-elementige Teilmengen $\{i_1, \ldots, i_k\}$ von $\{1, \ldots, n\}$ gibt. für das Gegenereignis erhalten wir:

$$\begin{aligned}
\mathbb{P}[\text{„kein Schlüssel passt“}] \;&=\; 1 - \mathbb{P}[\text{„mindestens ein Fixpunkt“}] \\
&= 1 + \sum_{k=1}^{n} \frac{(-1)^k}{k!} \;=\; \sum_{k=0}^{n} \frac{(-1)^k}{k!}.
\end{aligned}$$

Die letzte Summe konvergiert für $n \to \infty$ gegen e^{-1}. Der Grenzwert existiert also und ist weder 0 noch 1! Somit hängt die Wahrscheinlichkeit, dass keiner der Schlüssel passt, für große n nur wenig von n ab.

Empirische Verteilungen

Sei $x_1, x_2, \ldots \in \Omega$ eine Liste von Beobachtungsdaten oder Merkmalsausprägungen, zum Beispiel das Alter aller Einwohner von Bonn. Für $k \in \mathbb{N}$ ist

$N_k[A] := |\{i \in \{1, \ldots, k\} \mid x_i \in A\}|$ die Häufigkeit der Werte in A unter $x_1, \ldots, x_k$, und
$\mathbb{P}_k[A] := N_k[A]/k$, die entsprechende relative Häufigkeit von Werten in A.

Für jedes feste k ist $\mathbb{P}_k$ eine Wahrscheinlichkeitsverteilung auf $(\Omega, \mathcal{P}(\Omega))$, deren Massenfunktion

$$p_k(\omega) = \frac{N_k[\{\omega\}]}{k}$$

durch die relativen Häufigkeit der möglichen Merkmalsausprägungen unter $x_1, \ldots, x_k$ gegeben ist. Die Wahrscheinlichkeitsverteilung $\mathbb{P}_k$ heißt *empirische Verteilung* der Werte $x_1, \ldots, x_k$. In der beschreibenden Statistik analysiert man empirische Verteilungen mithilfe verschiedener Kenngrößen.

Beispiele. a) Abzählung aller möglichen Fälle:
Sei $x_1, \ldots, x_k$ eine Abzählung der Elemente in Ω. Dann stimmt die empirische Verteilung $\mathbb{P}_k$ mit der Gleichverteilung auf Ω überein.

b) Empirische Verteilung von n Zufallszahlen aus $\{1, 2, 3, 4, 5, 6\}$:
Das **empirische Gesetz der großen Zahlen** besagt, dass sich die empirischen Verteilungen

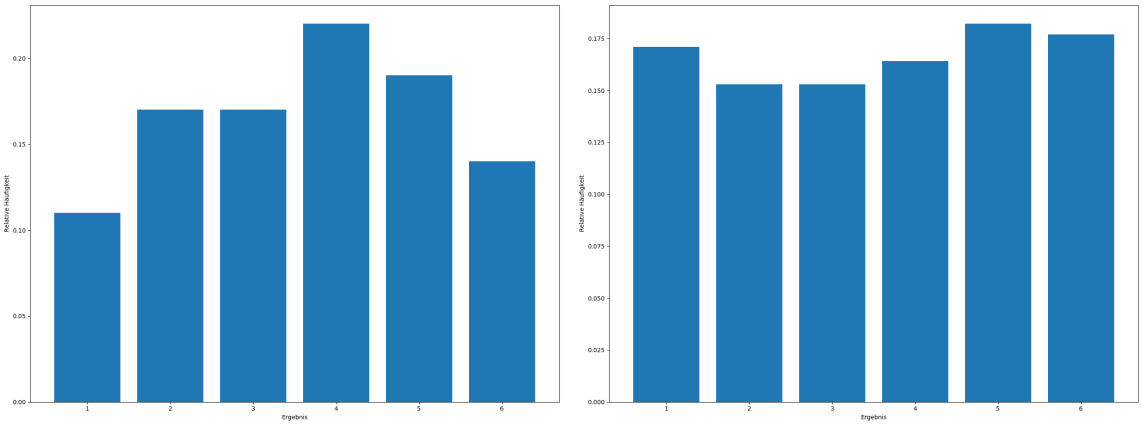


Abbildung 1.1.: Empirische Verteilung von 100 bzw. 1000 Würfen eines fairen Würfels.

für $k \to \infty$ der zugrundeliegenden Wahrscheinlichkeitsverteilung $\mathbb{P}$ (hier der Gleichverteilung auf $\{1, 2, \ldots, 6\}$) annähern:

$$\mathbb{P}_k[A] = \frac{|\{i \in \{1, \ldots, k\} \mid x_i \in A\}|}{k} \to \mathbb{P}[A] \qquad \text{für } k \to \infty.$$

Diese Aussage wird auch als frequentistische „Definition“ der Wahrscheinlichkeit von A in den empirischen Wissenschaften verwendet. Wir werden die Konvergenz der empirischen Verteilungen von unabhängigen, identisch verteilten Zufallsvariablen unten aus den Kolmogorovschen Axiomen herleiten.

c) Empirische Verteilung der Buchstaben „a“ bis „z“ in dem Wort „Eisenbahnschrankenwaerterhaeuschen“ und in einem englischen Wörterbuch:

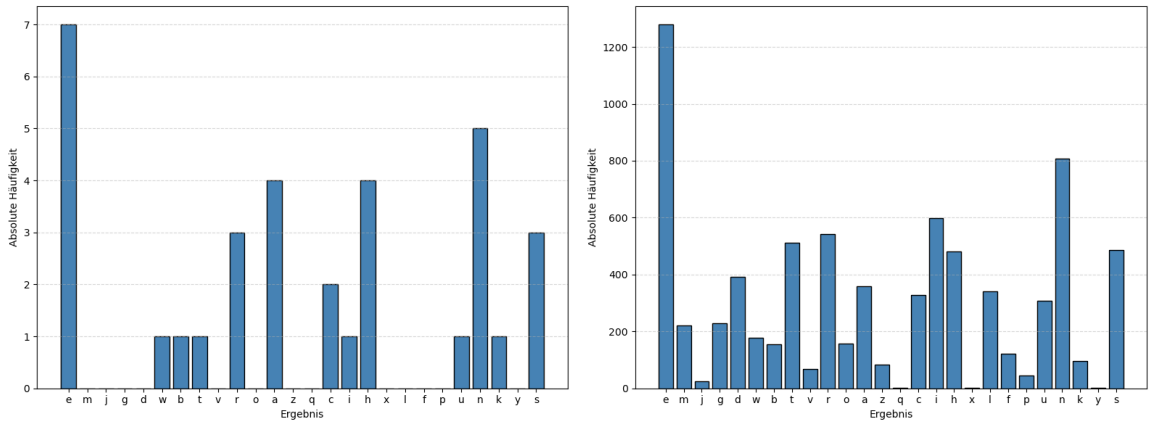
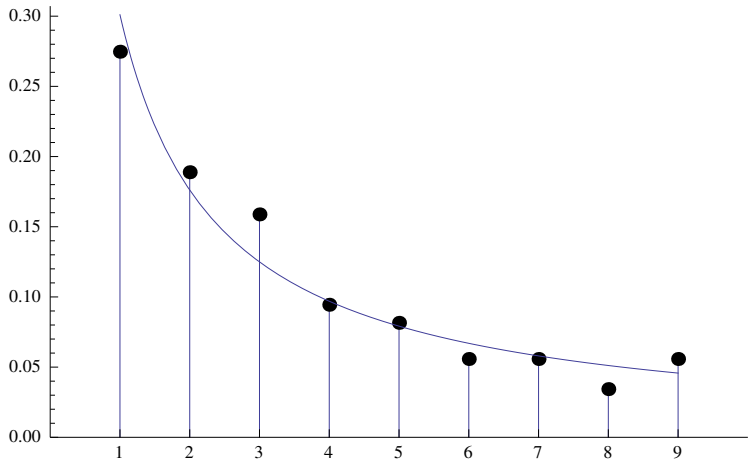


Abbildung 1.2.: Empirische Verteilung der Buchstaben in dem Wort „Eisenbahnschrankenwaerterhaeuschen“ bzw. in Faust 1.

d) Benfordsches Gesetz:
Das Benfordsche Gesetz beschreibt eine Gesetzmäßigkeit in der Verteilung der Anfangsziffern von Zahlen in empirischen Datensätzen. Es lässt sich etwa in Datensätzen über Einwohnerzahlen von Städten, Geldbeträge in der Buchhaltung, Naturkonstanten etc. beobachten. Ist d die erste Ziffer einer Dezimalzahl, so tritt sie nach dem Benfordschen Gesetz in empirischen Datensätzen näherungsweise mit folgenden relativen Häufigkeiten $p(d)$ auf:

$$p(d) \;=\; \log_{10}\left(1+\frac{1}{d}\right) \;=\; \log_{10}(d+1) - \log_{10} d.$$

In der Grafik unten (`Quelle`: „Wolfram Demonstrations Project“) werden die relativen Häufigkeiten der Anfangsziffern 1 bis 9 in den Anzahlen der Telefonanschlüsse in allen Ländern der Erde mit den nach dem Benfordschen Gesetz prognostizierten relativen Häufigkeiten verglichen.



1.2. Diskrete Zufallsvariablen und ihre Verteilung

Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein gegebener Wahrscheinlichkeitsraum. Meistens ist man nicht so sehr an den Elementen $\omega \in \Omega$ selbst interessiert, sondern an den Werten $X(\omega)$, die bestimmte von ω (also vom Zufall) abhängende Größen X annehmen. Entsprechende Abbildungen $\omega \to X(\omega)$ nennt man Zufallsvariablen, wenn die Ereignisse

$$\{X \in B\} \;=\; \{\omega \in \Omega : X(\omega) \in B\} \;=\; X^{-1}(B)$$

für hinreichend viele Teilmengen B des Wertebereichs von X in der zugrundeliegenden σ-Algebra $\mathcal{A}$ enthalten sind. Wir beschränken uns zunächst auf Zufallsvariablen mit abzählbarem Wertebereich.

Zufallsvariablen, Verteilung und Massenfunktion

Definition 1.8. (i) Eine **diskrete Zufallsvariable** ist eine Abbildung

$$X : \Omega \to S, \qquad S \text{ abzählbar},$$

so dass für alle $a \in S$ gilt:

$$X^{-1}(a) \;=\; \{\omega \in \Omega \mid X(\omega) = a\} \;\in\; \mathcal{A}. \tag{1.5}$$

für die Menge $X^{-1}(a)$ schreiben wir im folgenden kurz $\{X = a\}$.

(ii) Die **Verteilung** einer diskreten Zufallsvariable $X : \Omega \to S$ ist die Wahrscheinlichkeitsverteilung μ_X auf S mit Gewichten

$$p_X(a) \;=\; \mathbb{P}[\{X = a\}] \qquad (a \in S).$$

Statt $\mathbb{P}[\{X = a\}]$ schreiben wir auch kurz $\mathbb{P}[X = a]$.

Bemerkung. a) Man verifiziert leicht, dass p_X tatsächlich die Massenfunktion einer Wahrscheinlichkeitsverteilung μ_X auf S ist. In der Tat gilt $p_X(a) \geq 0$ für alle $a \in S$. Da die Ereignisse $\{X = a\}$ disjunkt sind, folgt zudem:

$$\sum_{a \in S} p_X(a) \;=\; \sum_{a \in S} \mathbb{P}[X = a] \;=\; \mathbb{P}\Big[\bigcup_{a \in S} \{X = a\}\Big] \;=\; \mathbb{P}[\Omega] \;=\; 1.$$

für eine beliebige Teilmenge $B \subseteq S$ des Wertebereichs von X ist $\{X \in B\}$ wieder ein Ereignis in der σ-Algebra $\mathcal{A}$, denn

$$\{X \in B\} \;=\; \underbrace{\{\omega \in \Omega : X(\omega) \in B\}}_{X^{-1}(B)} \;=\; \bigcup_{a \in B} \underbrace{\{X = a\}}_{\in \mathcal{A}} \;\in\; \mathcal{A}$$

nach der Definition einer σ-Algebra. Wegen der σ-Additivität von $\mathbb{P}$ gilt

$$\mathbb{P}[X \in B] \;=\; \sum_{a \in B} \mathbb{P}[X = a] \;=\; \sum_{a \in B} p_X(a) \;=\; \mu_X[B].$$

Die Verteilung μ_X gibt also an, mit welchen Wahrscheinlichkeiten die Zufallsvariable X Werte in bestimmten Teilmengen des Wertebereichs S annimmt.

b) Ist Ω selbst abzählbar und $\mathcal{A} = \mathcal{P}(\Omega)$, dann ist jede Abbildung $X : \Omega \to S$ eine Zufallsvariable.

c) Eine **reellwertige Zufallsvariable** ist eine Abbildung $X : \Omega \to \mathbb{R}$, so dass die Mengen $\{X \le c\} = X^{-1}((-\infty, c])$ für alle $c \in \mathbb{R}$ in der σ-Algebra $\mathcal{A}$ enthalten sind. Man überzeugt sich leicht, dass diese Definition mit der Definition oben konsistent ist, wenn der Wertebereich S eine abzählbare Teilmenge von $\mathbb{R}$ ist.

Wir beginnen mit einem elementaren Beispiel:

Beispiel (Zweimal Würfeln). Sei $\mathbb{P} = \mathrm{Unif}(\Omega)$ die Gleichverteilung auf der Menge

$$\Omega \;=\; \{\omega = (x_1, x_2) \;:\; x_i \in \{1, \dots, 6\}\}.$$

Die Augenzahl des i-ten Wurfs ($i = 1, 2$) wird durch $X_i(\omega) := x_i$ beschrieben. Die Abbildung

$$X_i : \Omega \to S := \{1, 2, 3, 4, 5, 6\}$$

ist eine diskrete Zufallsvariable. Die Verteilung μ_{X_i} hat die Massenfunktion

$$p_{X_i}(a) \;=\; \mathbb{P}[X_i = a] \;=\; \frac{6}{36} \;=\; \frac{1}{6} \qquad \text{für alle } a \in S,$$

d.h. μ_{X_i} ist die Gleichverteilung auf S.

Die Summe der Augenzahlen bei beiden Würfen wird durch die Zufallsvariable

$$Y(\omega) \;:=\; X_1(\omega) + X_2(\omega)$$

beschrieben. Die Gewichte der Verteilung von Y sind

$$p_Y(a) \;=\; \mathbb{P}[Y = a] \;=\; \begin{cases} \frac{1}{36} & \text{falls } a \in \{2, 12\}, \\ \frac{2}{36} & \text{falls } a \in \{3, 11\}, \\ \text{usw.} \end{cases}.$$

Die Zufallsvariable Y ist also nicht mehr gleichverteilt !

Das folgende Beispiel verallgemeinert die Situation aus dem letzten Beispiel:

Beispiel. Sei $\mathbb{P}$ die Gleichverteilung auf einer endlichen Menge $\Omega = \{\omega_1, \dots, \omega_n\}$ mit n Elementen, und sei $X : \Omega \to S$ eine beliebige Abbildung in eine Menge S. Setzen wir $x_i := X(\omega_i)$, dann ist X eine Zufallsvariable mit Massenfunktion

$$\mathbb{P}[X = a] \;=\; \frac{|\{\omega \in \Omega \;:\; X(\omega) = a\}|}{|\Omega|} \;=\; \frac{|\{1 \le i \le n \;:\; x_i = a\}|}{n}.$$

Die Verteilung μ_X von X unter der Gleichverteilung ist also die empirische Verteilung der Werte $x_1, \dots, x_n$.

Binomialverteilungen

Wir wollen nun zeigen, wie man von der Gleichverteilung zu anderen fundamentalen Verteilungen der Wahrscheinlichkeitstheorie gelangt. Dazu betrachten wir zunächst eine endliche Menge (Grundgesamtheit, Zustandsraum, Population) S. In Anwendungen können die Elemente von S alles mögliche beschreiben, zum Beispiel die Kugeln in einer Urne, die Einwohner von Bonn, oder die Fledermäuse

im Kottenforst. Wir wollen nun die zufällige Entnahme von n Einzelstichproben aus S mit Zurücklegen modellieren. Dazu setzen wir

$$\Omega \;=\; S^n \;=\; \{\omega = (x_1, \ldots, x_n) \;:\; x_i \in S\}.$$

Wir nehmen an, dass alle kombinierten Stichproben gleich wahrscheinlich sind, d.h. die zugrundeliegende Wahrscheinlichkeitsverteilung $\mathbb{P}$ sei die Gleichverteilung auf dem Produktraum Ω. Erste relevante Zufallsvariablen sind die Stichprobenwerte $X_i(\omega) = x_i$, $i = 1, \ldots, n$. Wie im ersten Beispiel oben gilt

$$\mathbb{P}[X_i = a] \;=\; \frac{|\{X_i = a\}|}{|\Omega|} \;=\; \frac{|S|^{n-1}}{|S|^n} \;=\; \frac{1}{|S|} \qquad \text{für alle } a \in S,$$

d.h. die Zufallsvariablen X_i sind gleichverteilt auf S. Sei nun $E \subseteq S$ eine Teilmenge des Zustandsraums, die für eine bestimmte Merkmalsausprägung der Stichprobe steht (zum Beispiel Ziehen einer roten Kugel oder Beobachtung einer bestimmten Fledermausart). Die Ereignisse $\{X_i \in E\}$, dass diese Merkmalsausprägung bei der i-ten Einzelstichprobe vorliegt, haben die Wahrscheinlichkeit

$$\mathbb{P}[X_i \in E] \;=\; \mu_{X_i}[E] \;=\; |E|/|S|.$$

Wir betrachten nun die Häufigkeit von E in der gesamten Stichprobe $(X_1, \ldots, X_n)$. Diese wird durch die Zufallsvariable $N : \Omega \to \{0, 1, 2, \ldots, n\}$,

$$N(\omega) \;:=\; |\{1 \le i \le n \;:\; X_i(\omega) \in E\}|$$

beschrieben. Ist $p = |E|/|S|$ die relative Häufigkeit des Merkmals E in der Population S, dann erhalten wir:

Lemma 1.9. *Für $k \in \{0, 1, \ldots, n\}$ gilt:*

$$\mathbb{P}[N = k] \;=\; \binom{n}{k} p^k \, (1-p)^{n-k}.$$

Beweis. Es gilt

$$|\{\omega \in \Omega \mid N(\omega) = k\}| = \binom{n}{k} |E|^k \, |S \backslash E|^{n-k}.$$

Hierbei gibt $\binom{n}{k}$ die Anzahl der Möglichkeiten an, k Indizes aus $\{1, \ldots, n\}$ auszuwählen (diejenigen, für die die Merkmalsausprägung E vorliegt), $|E|^k$ ist die Anzahl der Möglichkeiten für die nun festgelegten k Stichproben Werte aus E zu wählen, und $|S \backslash E|^{n-k}$ ist die Anzahl der Möglichkeiten für die verbleibenden $n - k$ Stichproben Werte aus $S \setminus E$ zu wählen. Da $\mathbb{P}$ die Gleichverteilung auf S^n ist, folgt

$$\mathbb{P}[N = k] \;=\; \frac{\binom{n}{k} |E|^k \, |S \backslash E|^{n-k}}{|S|^n} \;=\; \binom{n}{k} \left(\frac{|E|}{|S|}\right)^k \left(\frac{|S \backslash E|}{|S|}\right)^{n-k} \;=\; \binom{n}{k} p^k \, (1-p)^{n-k}. \qquad \blacksquare$$

Definition 1.10. Sei $n \in \mathbb{N}$ und $p \in [0, 1]$. Die Wahrscheinlichkeitsverteilung auf $\{0, 1, \ldots, n\}$ mit Massenfunktion

$$b_{n,p}(k) \;=\; \binom{n}{k} p^k \, (1-p)^{n-k}$$

heißt **Binomialverteilung mit Parametern n und p** (kurz: $\mathrm{Bin}(n, p)$).

Bemerkung. Dass $b_{n,p}$ die Massenfunktion einer Wahrscheinlichkeitsverteilung ist, kann man mit der allgemeinen binomischen Formel nachrechnen. Dies ist aber gar nicht notwendig, da sich diese Eigenschaft bereits aus Lemma 1.9 ergibt !

Wir haben gesehen, wie sich die Binomialverteilung aus der Gleichverteilung auf einer endlichen Produktmenge ableiten lässt. Binomialverteilungen treten aber noch allgemeiner auf, nämlich als Verteilung der Häufigkeit des Eintretens unabhängiger Ereignisse mit gleichen Wahrscheinlichkeiten. Ereignisse $E_1, \ldots, E_n$ heißen ***unabhängig***, falls

$$\mathbb{P}[E_{i_1} \cap E_{i_2} \cap \ldots \cap E_{i_k}] \;=\; \mathbb{P}[E_{i_1}] \cdot \mathbb{P}[E_{i_2}] \cdots \mathbb{P}[E_{i_k}]$$

für alle $k \leq n$ und $1 \leq i_1 < i_2 < \ldots < i_k \leq n$ gilt. Im Vorgriff darauf erwähnen wir schon die folgende wichtige Aussage:

Sind $E_1, \ldots, E_n$ unabhängige Ereignisse mit Wahrscheinlichkeit $\mathbb{P}[E_i] = p$, dann gilt

$$\mathbb{P}[\text{„genau } k \text{ der } E_i \text{ treten ein“}] \;=\; \binom{n}{k} p^k \, (1-p)^{n-k},$$

d.h. die Anzahl der Ereignisse, die eintreten, ist binomialverteilt.

Poissonverteilungen und Poissonscher Grenzwertsatz

Aus der Binomialverteilung lässt sich eine weitere Wahrscheinlichkeitsverteilung ableiten, die die Häufigkeit von seltenen Ereignissen beschreibt. Bevor wir den entsprechenden mathematischen Grenzwertsatz formulieren und beweisen, sehen wir, wie sich in diversen Anwendungssituationen aus einigen wenigen Grundannahmen dasselbe mathematische Modell ergibt, wenn man die Anzahl der Ereignisse, die in einem bestimmten Zeitintervall eintreten, beschreiben möchte.

Beispiel (Seltene Ereignisse in stetiger Zeit). Wir betrachten eine Folge von Ereignissen, die zu zufälligen Zeitpunkten eintreten. Dies können zum Beispiel eingehende Schadensfälle bei einer Versicherung, ankommende Anrufe in einer Telefonzentrale, oder radioaktive Zerfälle sein. Wir sind hier auf der Anwendungsebene - mit „Ereignissen“ meinen wir also im Moment keine mathematischen Objekte. Uns interessiert die Anzahl N der Ereignisse, die in einem festen Zeitintervall der Länge t eintreten. Der Einfachheit halber und ohne wesentliche Beschränkung der Allgemeinheit setzen wir $t = 1$. Wir treffen nun einige Grundannahmen, die näherungsweise erfüllt sein sollten. Diese Grundannahmen sind zunächst wieder auf der Anwendungsebene, und werden erst später durch Annahmen an das mathematische Modell präzisiert. Wir formulieren die Annahmen für die radioaktiven Zerfälle - entsprechende Annahmen gelten aber näherungsweise auch in vielen anderen Situationen.

Annahme 1: „Die Zerfälle passieren ‚unabhängig‘ voneinander zu ‚zufälligen‘ Zeitpunkten“.

Um die Verteilung der Anzahl der Zerfälle pro Zeiteinheit näherungsweise bestimmen zu können, unterteilen wir das Zeitintervall $(0, 1]$ in die n Teilintervalle $((k-1)/n, k/n]$, $k = 1, 2, \ldots, n$:

Annahme 2: „Wenn n sehr groß ist, dann passiert in einer Zeitspanne der Länge $\frac{1}{n}$ ‚fast immer‘ höchstens ein Zerfall“.

In einem stochastischen Modell repräsentiere E_i das Ereignis, dass im Zeitintervall $\left(\frac{i-1}{n}, \frac{i}{n}\right]$ mindestens ein radioaktiver Zerfall stattfindet. Die Wahrscheinlichkeit von E_i sei unabhängig von i und näherungsweise proportional zu $\frac{1}{n}$, also:

Annahme 3: „Es gilt $\mathbb{P}[E_i] \approx \lambda/n$ mit einer Konstanten $\lambda \in (0, \infty)$ (der Intensität bzw. Zerfallsrate)."

Wir gehen weiter davon aus, dass sich die erste Annahme dadurch präzisieren lässt, dass wir Unabhängigkeit der Ereignisse $E_1, \ldots, E_n$ fordern. Das ist nicht ganz offensichtlich, lässt sich aber in einem anspruchsvolleren mathematischen Modell, dass die Zeitpunkte aller Zerfälle beschreibt, rechtfertigen. Unter den Annahmen 1, 2 und 3 sollte für das Ereignis, dass genau k radioaktive Zerfälle im Zeitintervall $[0, 1]$ stattfinden, dann näherungsweise gelten, dass

$$\mathbb{P}[N = k] \approx \mathbb{P}[\text{„genau } k \text{ der } E_i \text{ treten ein"}] \approx b_{n,\frac{\lambda}{n}}(k),$$

wobei $b_{n,\frac{\lambda}{n}}(k)$ das Gewicht von k unter der Binomialverteilung mit Parametern n und $\frac{\lambda}{n}$ ist. Diese Näherung sollte zudem *„für große n immer genauer werden"*. Daher sollten wir die Anzahl der Zerfälle pro Zeiteinheit bei Intensität λ durch eine Zufallsvariable mit nichtnegativen ganzzahligen Werten beschreiben, deren Verteilung die Massenfunktion

$$p_\lambda(k) \;=\; \lim_{n\to\infty} b_{n,\frac{\lambda}{n}}(k)$$

hat. Der folgende Satz zeigt, dass p_λ in der Tat die Massenfunktion einer Wahrscheinlichkeitsverteilung ist, nämlich der Poissonverteilung mit Parameter λ.

Satz 1.11 (Poissonapproximation der Binomialverteilung). Sei $\lambda \in (0, \infty)$. Dann gilt:

$$\lim_{n\to\infty} b_{n,\frac{\lambda}{n}}(k) \;=\; \frac{\lambda^k}{k!}\, e^{-\lambda}, \qquad \text{für alle } k = 0, 1, 2, \ldots.$$

Beweis. Für $k \in \{0, 1, 2, \ldots\}$ und $n \to \infty$ gilt

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

■

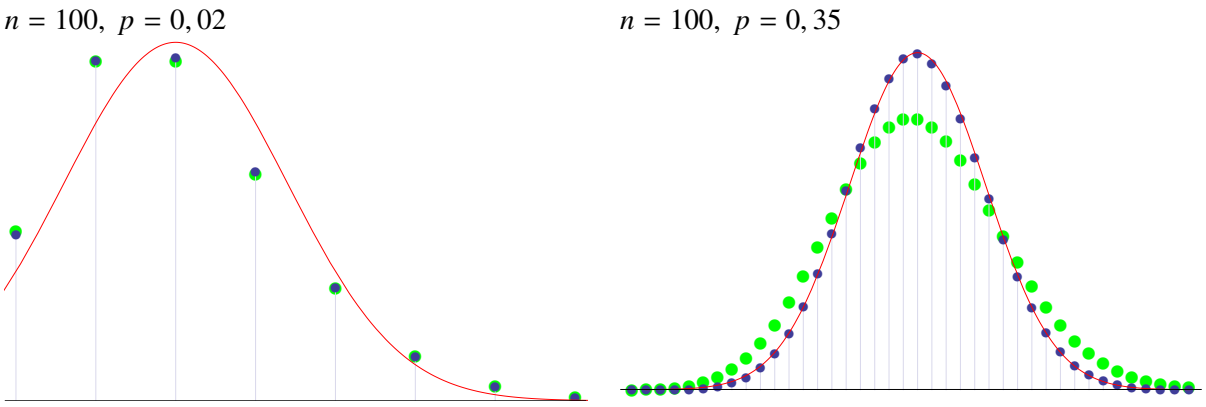
Definition 1.12. Die Wahrscheinlichkeitsverteilung auf $\{0, 1, 2, \ldots\}$ mit Massenfunktion

$$p_\lambda(k) \;=\; \frac{\lambda^k}{k!} e^{-\lambda}, \qquad k = 0, 1, 2, \ldots,$$

heißt **Poissonverteilung mit Parameter (Intensität)** λ.

Aufgrund des Satzes verwendet man die Poissonverteilung zur näherungsweisen Modellierung der *Häufigkeit seltener Ereignisse* (zum Beispiel Rechtschreibfehler in einer Zeitung, Programmfehler in einer Software, Lottogewinne, Unfälle oder Naturkatastrophen, Zusammenbrüche von Mobilfunknetzen, usw.), und damit zur „Approximation" von Binomialverteilungen mit kleinen Erfolgswahrscheinlichkeiten p.

Für häufigere Ereignisse (zum Beispiel wenn die Erfolgswahrscheinlichkeit p unabhängig von n ist) verwendet man hingegen besser eine Normalverteilung zur näherungsweisen Modellierung der (geeignet reskalierten) relativen Häufigkeit $\frac{k}{n}$ des Ereignisses für große n. Definition und Eigenschaften von Normalverteilungen werden wir später kennenlernen.
Die folgenden (mit MATHEMATICA erstellten) Graphiken zeigen die Poisson- und Normalapproximation (Poissonverteilung grün, reskalierte Dichte der Normalverteilung rot) der Binomialverteilung Bin(n,p) (blau) für unterschiedliche Parameterwerte:



Hypergeometrische Verteilungen

Abschließend zeigen wir, wie sich eine weitere Klasse von Wahrscheinlichkeitsverteilungen, die hypergeometrischen Verteilungen, aus Gleichverteilungen ableiten lässt. Diese Verteilungen treten bei der Entnahme von Stichproben ohne Zurücklegen aus einer Gesamtpopulation auf.

Beispiel (Stichproben ohne Zurücklegen). Wir betrachten eine Population S mit insgesamt m Objekten, z.B. die Kugeln in einer Urne, die Wähler in einem Bundesland, oder die Bäume in einem Waldstück. Unter den m Objekten seien r, die eine gewisse Eigenschaft/ Merkmalsausprägung besitzen (z.B. Wähler einer bestimmten Partei), und $m - r$, die diese Eigenschaft nicht besitzen. Wir wollen die Entnahme einer Zufallsstichprobe von n Objekten aus der Population beschreiben, wobei $n \leq \min(r, m - r)$ gelte. Dazu betrachten wir den Grundraum Ω, der aus allen Teilmengen (Stichproben) $\omega \subseteq S$ der Kardinalität n besteht. Die Menge Ω enthält $\binom{m}{n}$ Elemente. Gehen wir davon aus, dass alle Stichproben gleich wahrscheinlich sind, dann wählen wir als zugrundeliegende Wahrscheinlichkeitsverteilung in unserem Modell die Gleichverteilung

$$\mathbb{P} \;=\; \mathrm{Unif}(\Omega).$$

Sei nun $N(\omega)$ die Anzahl der Objekte in der Stichprobe ω, die die Merkmalsausprägung haben. Für die Wahrscheinlichkeit, dass genau k der n Objekte in der Stichprobe die Merkmalsausprägung haben, ergibt sich

$$\mathbb{P}[N = k] \;=\; \frac{|\{\omega \in \Omega \,:\, N(\omega) = k\}|}{|\Omega|} \;=\; \frac{\binom{r}{k}\binom{m-r}{n-k}}{\binom{m}{n}} \qquad (k = 0, 1, \ldots, n).$$

Definition 1.13. Die Wahrscheinlichkeitsverteilung auf $\{0, 1, 2, \ldots, n\}$ mit Massenfunktion

$$h_{m,r,n}(k) \;=\; \binom{r}{k}\binom{m-r}{n-k} \Big/ \binom{m}{n}$$

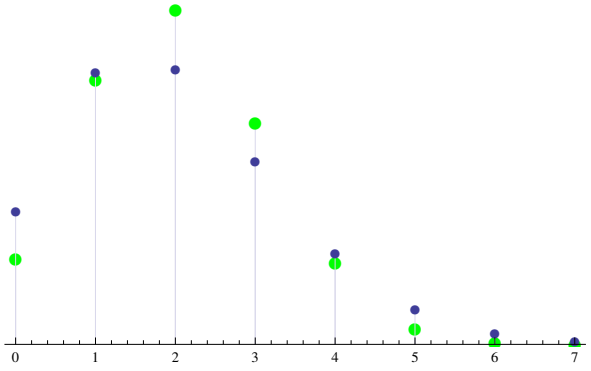
wird **hypergeometrische Verteilung mit Parametern** m**,** r **und** n genannt.

Ist die zugrundeliegende Population im Verhältnis zur Stichprobe groß, dann sollte sich kein wesentlicher Unterschied bei Ziehen mit und ohne Zurücklegen ergeben, da nur sehr selten dasselbe Objekt zweimal gezogen wird. Dies lässt sich mathematisch zum Beispiel folgendermaßen präzisieren: für ein festes $n \in \mathbb{N}$ und $m, r \to \infty$ mit $p = r/m$ fest gilt

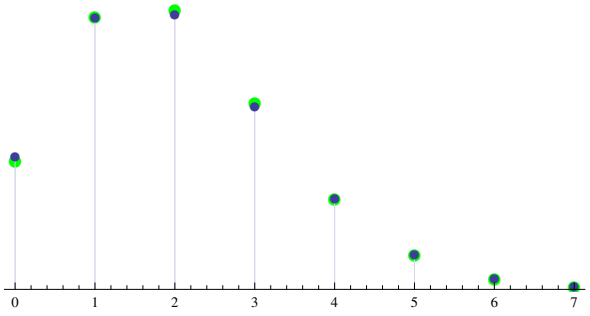
$$h_{m,r,n}(k) \longrightarrow \binom{n}{k} p^k (1-p)^{n-k},$$

d.h. die hypergeometrische Verteilung mit Parametern m, pm und n nähert sich der Binomialverteilung $\mathrm{Bin}(n, p)$ an. Der Beweis ist eine Übungsaufgabe. Die folgenden (mit MATHEMATICA erstellten) Graphiken zeigen die Gewichte der Binomialverteilung Bin(n,p) (blau) und der hypergeometrischen Verteilung Hyp(m,pm,n) (grün) für unterschiedliche Parameterwerte:

$n = 100,\ p = 0,02,\ m = 300$



$n = 100,\ p = 0,02,\ m = 3000$



1.3. Erwartungswert

Eine erste wichtige Kenngröße reellwertiger Zufallsvariablen ist ihr Erwartungswert. Wir betrachten eine Zufallsvariable $X : \Omega \to S$ auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$, deren Wertebereich S eine abzählbare Teilmenge der reellen Zahlen ist. In diesem Fall können wir den Erwartungswert (Mittelwert) von X bzgl. der zugrundeliegenden Wahrscheinlichkeitsverteilung $\mathbb{P}$ als gewichtetes Mittel der Werte von X definieren:

Definition 1.14. Der **Erwartungswert** von X bzgl. $\mathbb{P}$ ist gegeben durch

$$\mathbb{E}[X] := \sum_{a \in S} a \cdot \mathbb{P}[X = a] = \sum_{a \in S} a \cdot p_X(a),$$

sofern die Summe auf der rechten Seite wohldefiniert ist.

Nimmt die Zufallsvariable X nur nichtnegative Werte $X(\omega) \geq 0$ an, dann sind alle Summanden der Reihe nichtnegativ, und der Erwartungswert $\mathbb{E}[X]$ ist wohldefiniert in $[0, \infty]$. Weiterhin ist $\mathbb{E}[X]$ wohldefiniert und endlich, falls die Reihe absolut konvergiert. Allgemeiner können wir den Erwartungswert immer dann definieren, wenn

$$\sum_{a \in S, a < 0} |a| \cdot \mathbb{P}[X = a] < \infty \qquad \text{gilt.}$$

Der Erwartungswert $\mathbb{E}[X]$ wird häufig als *Prognosewert* für $X(\omega)$ verwendet, wenn keine weitere Information vorliegt.

Bemerkung. Nach der Definition *hängt der Erwartungswert nur von der Verteilung μ_X der Zufallsvariablen X ab* ! Wir bezeichnen $\mathbb{E}[X]$ daher auch als **Erwartungswert der Wahrscheinlichkeitsverteilung** μ_X auf $\mathbb{R}$.

Beispiel (Gleichverteilte Zufallsvariablen). Ist X gleichverteilt auf einer endlichen Teilmenge $S = \{a_1, \ldots, a_n\}$ von $\mathbb{R}$ mit $a_i \neq a_j$ für $i \neq j$, dann ist der Erwartungswert $\mathbb{E}[X]$ das arithmetische Mittel der Werte von X:

$$\mathbb{E}[X] \;=\; \frac{1}{n} \sum_{i=1}^{n} a_i.$$

Beispiel (Poissonverteilung). Für eine mit Parameter λ Poisson-verteilte Zufallsvariable N gilt

$$\mathbb{E}[N] = \sum_{k=0}^{\infty} k\, \mathbb{P}[N = k] = \sum_{k=0}^{\infty} k\, \frac{\lambda^k}{k!}\, e^{-\lambda} = \lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!}\, e^{-\lambda} = \lambda \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}\, e^{-\lambda} = \lambda.$$

Beschreibt N die Häufigkeit eines Ereignisses (pro Zeiteinheit), dann können wir den Parameter λ dementsprechend als *mittlere Häufigkeit* oder *Intensität* interpretieren.

Beispiel (Erwartungswerte von Indikatorfunktionen). Die Indikatorfunktion eines Ereignisses $A \in \mathcal{A}$ ist die durch

$$I_A(\omega) \;:=\; \begin{cases} 1 & \text{falls } \omega \in A, \\ 0 & \text{falls } \omega \in A^C, \end{cases}$$

definierte Zufallsvariable. Für den Erwartungswert gilt

$$\mathbb{E}[I_A] \;=\; 1 \cdot \mathbb{P}[X = 1] + 0 \cdot \mathbb{P}[X = 0] \;=\; \mathbb{P}[A].$$

Beträgt beispielsweise die Leistung in einem elementaren Versicherungskontrakt

$$Y(\omega) \;=\; \begin{cases} c & \text{falls } \omega \in A, \qquad \text{„Schadensfall“}, \\ 0 & \text{sonst}, \end{cases}$$

dann gilt $Y = c \cdot I_A$, und

$$\mathbb{E}[Y] \;=\; c \cdot \mathbb{P}[A].$$

Transformationssatz

Sei $X : \Omega \to S$ eine Zufallsvariable mit Werten in einer beliebigen abzählbaren Menge S (die nicht notwendig aus reellen Zahlen besteht). Dann können wir Erwartungswerte von Zufallsvariablen der Form

$$g(X)(\omega) \;:= g(X(\omega))$$

mit einer Funktion $g : S \to \mathbb{R}$ berechnen. Anstatt dabei über die Werte von $g(X)$ zu summieren, können wir den Erwartungswert auch direkt aus der Verteilung von X erhalten.

Satz 1.15 (Transformationssatz). Für jede reellwertige Funktion $g : S \to \mathbb{R}$ ist

$$g(X) = g \circ X : \Omega \to g(S) \subset \mathbb{R}$$

eine diskrete Zufallsvariable. Es gilt

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

falls die Summe wohldefiniert ist (also zum Beispiel falls g nichtnegativ ist, oder die Reihe absolut konvergiert).

Beweis. Wegen $\{g(X) = b\} = \bigcup_{a \in g^{-1}(b)} \{X = a\} \in \mathcal{A}$ für alle $b \in g(S)$ ist $g(X)$ wieder eine Zufallsvariable. Da die Vereinigung disjunkt ist, erhalten wir unter Verwendung der σ-Additivität:

$$\begin{aligned}\mathbb{E}[g(X)] &= \sum_{b \in g(S)} b \cdot \mathbb{P}[g(X) = b] = \sum_{b \in g(S)} b \cdot \sum_{a \in g^{-1}(b)} \mathbb{P}[X = a] \\ &= \sum_{b \in g(S)} \sum_{a : g(a) = b} g(a) \cdot \mathbb{P}[X = a] = \sum_{a \in S} g(a) \cdot \mathbb{P}[X = a].\end{aligned}$$

■

Beispiele. Sei $X : \Omega \to S \subset \mathbb{R}$ eine reellwertige Zufallsvariable mit abzählbarem Wertebereich S.

a) Für den Erwartungswert von $|X|$ ergibt sich

$$\mathbb{E}[|X|] = \sum_{a \in S} |a| \cdot \mathbb{P}[X = a].$$

Ist $\mathbb{E}[|X|]$ endlich, dann konvergiert $\mathbb{E}[X] = \sum a \cdot \mathbb{P}[X = a]$ absolut.

b) Die **Varianz** einer reellwertigen Zufallsvariable X mit $\mathbb{E}[|X|] < \infty$ ist definiert als mittlere quadratische Abweichung vom Erwartungswert, d.h.,

$$\mathrm{Var}[X] := \mathbb{E}\left[(X - \mathbb{E}[X])^2\right].$$

Kennen wir $\mathbb{E}[X]$, dann berechnet sich die Varianz als

$$\mathrm{Var}[X] = \sum_{a \in S} (a - \mathbb{E}[X])^2 \, \mathbb{P}[X = a] \in [0, \infty].$$

Ebenso wie der Erwartungswert hängt auch die Varianz nur von der Verteilung μ_X ab.

c) Ist Ω selbst abzählbar, dann können wir den Erwartungswert auch als *gewichtetes Mittel* über $\omega \in \Omega$ darstellen. In der Tat folgt für $X : \Omega \to \mathbb{R}$ durch Anwenden des Transformationssatzes:

$$\mathbb{E}[X] = \mathbb{E}[X \circ id_\Omega] = \sum_{\omega \in \Omega} X(\omega) \cdot \mathbb{P}[\{\omega\}],$$

wobei $id_\Omega(\omega) = \omega$ die identische Abbildung auf Ω bezeichnet. Ist $\mathbb{P}$ die Gleichverteilung auf Ω, so ist der Erwartungswert das *arithmetische Mittel*

$$\mathbb{E}[X] = \frac{1}{|\Omega|} \sum_{\omega \in \Omega} X(\omega).$$

Beispiel (Sankt-Petersburg-Paradoxon). Wir betrachten ein Glücksspiel mit fairen Münzwürfen $X_1, X_2, \ldots$, wobei sich der Gewinn in jeder Runde verdoppelt bis zum ersten Mal „Kopf“ fällt. Danach ist das Spiel beendet, und der Spieler erhält den Gewinn ausbezahlt. Wie hoch wäre eine faire Teilnahmegebühr für dieses Spiel?

Wir können den Gewinn beschreiben durch die Zufallsvariable

$$G(\omega) \;=\; 2^{T(\omega)}, \qquad \text{mit} \qquad T(\omega) \;=\; \min\{n \in \mathbb{N} \,:\, X_n(\omega) = 0\}.$$

Hierbei beschreibt T die Wartezeit auf den ersten „Kopf“. Als Erwartungswert des Gewinns erhalten wir nach dem Transformationssatz

$$\mathbb{E}[G] = \sum_{k=1}^{\infty} 2^k \,\mathbb{P}[T = k] = \sum_{k=1}^{\infty} 2^k \,\mathbb{P}[X_1 = \cdots = X_{k-1} = 1, X_k = 0] = \sum_{k=1}^{\infty} 2^k \, 2^{-k} = \infty.$$

Das Spiel sollte also auf den ersten Blick bei beliebig hoher Teilnahmegebühr attraktiv sein – dennoch wäre wohl kaum jemand bereit, einen sehr hohen Einsatz zu zahlen.

Eine angemessenere Beschreibung – vom Blickwinkel des Spielers aus betrachtet – erhält man, wenn man eine (üblicherweise als monoton wachsend und konkav vorausgesetzte) Nutzenfunktion $u(x)$ einführt, die den Nutzen beschreibt, den der Spieler vom Kapital x hat. Für kleine x könnte etwa $u(x) = x$ gelten, aber für große x wäre plausibler $u(x) < x$. Dann ist c ein fairer Einsatz aus Sicht des Spielers, wenn $u(c) = \mathbb{E}[u(G)]$ gilt.

Linearität und Monotonie des Erwartungswertes

Eine fundamentale Eigenschaft des Erwartungswerts ist, dass dieser linear von der Zufallsvariable abhängt. Dies kann häufig ausgenutzt werden, um Erwartungswerte zu berechnen, siehe dazu die Beispiele unten.

Satz 1.16 (Linearität des Erwartungswerts). Seien $X : \Omega \to S_X \subseteq \mathbb{R}$ und $Y : \Omega \to S_Y \subseteq \mathbb{R}$ diskrete reellwertige Zufallsvariablen auf $(\Omega, \mathcal{A}, \mathbb{P})$, für die $\mathbb{E}[|X|]$ und $\mathbb{E}[|Y|]$ endlich sind. Dann gilt:

$$\mathbb{E}[\lambda X + \mu Y] = \lambda \,\mathbb{E}[X] + \mu \,\mathbb{E}[Y] \qquad \text{für alle } \lambda, \mu \in \mathbb{R}.$$

Beweis. Wir betrachten die durch $g(x, y) = \lambda x + \mu y$ definierte Abbildung $g : S_X \times S_Y \to \mathbb{R}$. Nach dem Transformationssatz ist $g(X, Y) = \lambda X + \mu Y$ eine Zufallsvariable mit Werten in $\mathbb{R}$ und Erwartungswert

$$\begin{aligned}
\mathbb{E}[\lambda X + \mu Y] \;&=\; \mathbb{E}[g(X,Y)] \;=\; \sum_{(a,b) \in S_X \times S_Y} g(a,b)\,\mathbb{P}[(X,Y) = (a,b)] \qquad (1.6)\\
&= \sum_{a \in S_X} \sum_{b \in S_Y} (\lambda a + \mu b)\,\mathbb{P}[X = a, Y = b]\\
&= \lambda \sum_{a \in S_X} a \sum_{b \in S_Y} \mathbb{P}[X = a, Y = b] + \mu \sum_{b \in S_Y} b \sum_{a \in S_X} \mathbb{P}[X = a, Y = b]\\
&= \lambda \sum_{a \in S_X} a\,\mathbb{P}[X = a] + \mu \sum_{b \in S_Y} b\,\mathbb{P}[Y = b]\\
&= \lambda\,\mathbb{E}[X] + \mu\,\mathbb{E}[Y].
\end{aligned}$$

Hierbei haben wir benutzt, dass die Reihe in (1.6) absolut konvergiert, da nach einer analogen Rechnung

$$\begin{aligned}\sum_{a\in S_X}\sum_{b\in S_Y} |\lambda\, a + \mu\, b|\,\mathbb{P}[X=a, Y=b] \;&\le\; |\lambda| \sum_{a\in S_X} |a|\,\mathbb{P}[X=a] + |\mu| \sum_{b\in S_Y} |b|\,\mathbb{P}[Y=b] \\ &=\; |\lambda|\,\mathbb{E}[|X|] + |\mu|\,\mathbb{E}[|Y|]\end{aligned}$$

gilt. Die rechte Seite ist nach Voraussetzung endlich. ■

Beispiel (Varianz). Für die Varianz einer reellwertigen Zufallsvariable X mit $\mathbb{E}[|X|] < \infty$ gilt

$$\begin{aligned}\mathrm{Var}[X] \;&=\; \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \;=\; \mathbb{E}\left[X^2 - 2X\,\mathbb{E}[X] + \mathbb{E}[X]^2\right] \\ &=\; \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2\,.\end{aligned}$$

Aus der Linearität folgt auch, dass der Erwartungswert monoton von der Zufallsvariablen abhängt:

Korollar 1.17 (Monotonie des Erwartungswerts). Seien die Voraussetzungen von Satz 1.16 erfüllt. Ist $X(\omega) \le Y(\omega)$ für alle $\omega \in \Omega$, dann gilt

$$\mathbb{E}[X] \le \mathbb{E}[Y].$$

Beweis. Nach Voraussetzung gilt $(Y-X)(\omega) \ge 0$ für alle $\omega \in \Omega$, weshalb der Erwartungswert $\mathbb{E}[Y-X]$ nichtnegativ ist. Aufgrund der Linearität des Erwartungswerts folgt

$$0 \;\le\; \mathbb{E}[Y - X] \;=\; \mathbb{E}[Y] - \mathbb{E}[X].$$

Die folgenden Beispiele demonstrieren, wie die Linearität häufig ausgenutzt werden kann, um Erwartungswerte auf einfache Weise zu berechnen:

Beispiel (Unabhängige 0-1-Experimente, Erwartungswert der Binomialverteilung).
Seien $A_1, A_2, \ldots, A_n \in \mathcal{A}$ unabhängige Ereignisse mit Wahrscheinlichkeit p, und sei $X_i = I_{A_i}$ die Indikatorfunktion des Ereignisses A_i. Die Zufallsvariablen X_i sind *Bernoulli-verteilt mit Parameter* p, d.h. es gilt

$$X_i = \begin{cases} 1 & \text{mit Wahrscheinlichkeit } p, \\ 0 & \text{mit Wahrscheinlichkeit } 1-p. \end{cases}$$

Damit erhalten wir

$$\mathbb{E}[X_i] \;=\; \mathbb{E}[I_{A_i}] \;=\; \mathbb{P}[A_i] \;=\; p \qquad \forall\, i = 0, 1, \ldots n.$$

Die Anzahl

$$S_n = X_1 + X_2 + \cdots + X_n$$

der Ereignisse, die eintreten, ist binomialverteilt mit Parametern n und p, d.h.

$$\mathbb{P}[S_n = k] \;=\; \binom{n}{k} p^k\,(1-p)^{n-k}.$$

Den Erwartungswert kann man daher folgendermaßen berechnen:

$$\mathbb{E}[S_n] \;=\; \sum_{k=0}^{n} k \cdot \mathbb{P}[S_n = k] \;=\; \sum_{k=0}^{n} k \binom{n}{k} p^k\,(1-p)^{n-k} \;=\; \ldots \;=\; n\,p.$$

Einfacher benutzt man aber die Linearität des Erwartungswerts, und erhält direkt

$$\mathbb{E}[S_n] \;=\; \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] \;=\; \sum_{i=1}^{n} \mathbb{E}[X_i] \;=\; n\,p.$$

Dies gilt sogar wenn die Ereignisse $A_1, \ldots, A_n$ *nicht unabhängig* sind !

Beispiel (Abhängige 0-1-Experimente, Erwartungswert der hypergeometrischen Verteilung).

Wir betrachten eine Population aus m Objekten, darunter r, die eine gewisse Eigenschaft besitzen. Aus der Population wird eine Zufallsstichprobe aus n Objekten ohne Zurücklegen entnommen, wobei $n \leq \min(r, m-r)$ gelte. Sei A_i das Ereignis, dass das i-te Objekt in der Stichprobe die Eigenschaft besitzt, und sei $X_i = I_{A_i}$. Dann beschreibt die hypergeometrisch verteilte Zufallsvariable

$$S_n = X_1 + \cdots + X_n$$

die Anzahl der Objekte in der Stichprobe mit der Eigenschaft. Als Erwartungswert der Verteilung $\mathrm{Hyp}(m, r, n)$ erhalten wir daher analog zum letzten Beispiel:

$$\mathbb{E}[S_n] \;=\; \sum_{i=1}^{n} \mathbb{E}[X_i] \;=\; \sum_{i=1}^{n} \mathbb{P}[A_i] = n\,\frac{r}{m}.$$

Auch im nächsten Beispiel wird eine ähnliche Methode benutzt, um den Erwartungswert zu berechnen:

Beispiel (Inversionen von Zufallspermutationen und Sortieren durch Einfügen). Seien $\mathbb{P}$ die Gleichverteilung auf der Menge $\Omega = \mathcal{S}_n$ aller Bijektionen $\omega\colon \{1, \ldots, n\} \to \{1, \ldots, n\}$, und

$$N(\omega) = |\{(i,j) \,:\, i < j \text{ und } \omega(i) > \omega(j)\}|,$$

die Anzahl der Inversionen einer Permutation $\omega \in \mathcal{S}_n$. Dann gilt

$$N = \sum_{1 \leq i < j \leq n} I_{A_{i,j}}, \qquad \text{wobei} \qquad A_{i,j} \;=\; \{\omega \in \mathcal{S}_n : \omega(i) > \omega(j)\}$$

das Ereignis ist, dass eine Inversion von i und j auftritt. Damit erhalten wir

$$\mathbb{E}[N] = \sum_{i<j} \mathbb{E}[I_{A_{i,j}}] = \sum_{i<j} \mathbb{P}[\{\omega \in \mathcal{S}_n : \omega(i) > \omega(j)\}] = \sum_{i<j} \frac{1}{2} = \frac{1}{2}\binom{n}{2} = \frac{n\,(n-1)}{4}.$$

Anwendung: Beim Sortieren durch Einfügen („Insertion Sort“) werden die Werte einer Liste $\omega(1)$, $\omega(2)$, …, $\omega(n)$ der Reihe nach an der richtigen Stelle eingefügt. Dabei wird der Wert $\omega(i)$ für $i < j$ beim Einfügen von $\omega(j)$ genau dann verschoben, wenn $\omega(j) < \omega(i)$ gilt. Ist die Anfangsanordnung eine zufällige Permutation der korrekten Anordnung, dann ist die mittlere Anzahl der Verschiebungen, die der Algorithmus vornimmt, also gleich $n\,(n-1)/4$.

Beispiel (Sammelbilderproblem). Ein Sammelalbum enthält n Bilder. Wie viele (zufällig ausgewählte) Bilder muss man im Durchschnitt kaufen, bis das Sammelalbum voll ist? Wir nehmen der Einfachheit halber an, dass wir neue Bilder durch „Ziehen mit Zurücklegen” erhalten, und dass wir in einem Zug jedes der n Bilder mit derselben Wahrscheinlichkeit ziehen. Angenommen, wir haben schon k verschiedene Bilder gesammelt. Sei T_k eine Zufallsvariable, die die Anzahl der Züge beschreibt, die benötigt werden, um ein weiteres neues Bild zu ziehen. Ist A_i das Ereignis, dass wir beim i-ten Versuch ein neues Bild erhalten, dann gilt für $l \in \mathbb{N}$:

$$\mathbb{P}[T_k > l] \;=\; \mathbb{P}[A_1^C \cap A_2^C \cap \cdots \cap A_l^C] \;=\; \frac{k^l}{n^l} \;=\; \left(\frac{k}{n}\right)^l.$$

Nach einer Übungsaufgabe erhalten wir damit

$$\mathbb{E}[T_k] \;=\; \sum_{l=0}^{\infty} \mathbb{P}[T_k > l] \;=\; \sum_{l=0}^{\infty} \left(\frac{k}{n}\right)^l \;=\; \frac{1}{1-k/n} \;=\; \frac{n}{n-k}.$$

Die Gesamtzahl der benötigten Züge, um alle Bilder zu erhalten ist $T = T_0 + T_1 + \cdots + T_{n-1}$. Wegen der Linearität des Erwartungswerts ergibt sich als durchschnittliche Anzahl der Bilder, die wir kaufen müssen, um alle Bilder zu erhalten:

$$\mathbb{E}[T] \;=\; \sum_{k=0}^{n-1} \mathbb{E}[T_k] \;=\; \sum_{k=0}^{n-1} \frac{n}{n-k} \;=\; n \sum_{i=1}^{n} \frac{1}{i} \;\sim\; n \ln(n).$$

Hierbei haben wir benutzt, dass sich die Summe für $n \to \infty$ asymptotisch wie $\ln(n)$ verhält, wie man durch Vergleich mit dem Integral von $1/x$ sieht. Um ein Album mit 160 Sammelbildern zu füllen, muss man zum Beispiel im Durchschnitt 812 zufällig ausgewählte Bilder kaufen, falls jedes Bild mit derselben Wahrscheinlichkeit gezogen wird. Die Ordnung $n \ln(n)$ tritt auch als mittlerer Aufwand verschiedener Algorithmen auf, zum Teil aus ähnlichen Gründen.

Einschluss-/Ausschlussprinzip

Auch das schon oben erwähnte Einschluss-/Ausschlussprinzip lässt sich mithilfe von Indikatorfunktionen elegant beweisen. Dazu verwenden wir die elementaren Identitäten

$$I_{A\cap B} = I_A \cdot I_B \qquad \text{und} \qquad I_{A^C} = 1 - I_A.$$

Satz 1.18 (Einschluss-/Ausschlussprinzip). Für $n \in \mathbb{N}$ und Ereignisse $A_1, \ldots, A_n \in \mathcal{A}$ gilt:

$$\mathbb{P}\left[A_1 \cup A_2 \cup \ldots \cup A_n\right] = \sum_{k=1}^{n} (-1)^{k-1} \sum_{1 \le i_1 < \ldots < i_k \le n} \mathbb{P}\left[A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}\right].$$

Beweis. Wir betrachten zunächst das Gegenereignis, und drücken die Wahrscheinlichkeiten als Erwartungswerte von Indikatorfunktionen aus. Unter Ausnutzung der Linearität des Erwartungswerts erhalten wir:

$$\begin{aligned}
\mathbb{P}\left[(A_1 \cup \cdots \cup A_n)^C\right] \;&=\; \mathbb{P}\left[A_1^C \cap \cdots \cap A_n^C\right] \;=\; \mathbb{E}\big[I_{A_1^C \cap \cdots \cap A_n^C}\big] \\
&=\; \mathbb{E}\Big[\prod_{i=1}^{n} I_{A_i^C}\Big] \;=\; \mathbb{E}\Big[\prod_{i=1}^{n} (1 - I_{A_i})\Big] \\
&=\; \sum_{k=0}^{n} (-1)^k \sum_{1 \le i_1 < \ldots < i_k \le n} \mathbb{E}\big[I_{A_{i_1}} \cdot \cdots \cdot I_{A_{i_k}}\big] \\
&=\; \sum_{k=0}^{n} (-1)^k \sum_{1 \le i_1 < \ldots < i_k \le n} \mathbb{E}\big[I_{A_{i_1} \cap \cdots \cap A_{i_k}}\big] \\
&=\; \sum_{k=0}^{n} (-1)^k \sum_{1 \le i_1 < \ldots < i_k \le n} \mathbb{P}\big[A_{i_1} \cap \cdots A_{i_k}\big].
\end{aligned}$$

Damit folgt

$$\begin{aligned}\mathbb{P}\left[A_1 \cup \cdots \cup A_n\right] &= 1 - \mathbb{P}\left[(A_1 \cup \cdots \cup A_n)^C\right] \\ &= \sum_{k=1}^{n} (-1)^{k-1} \sum_{1 \le i_1 < \ldots < i_k \le n} \mathbb{P}\left[A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}\right].\end{aligned}$$

■

2. Bedingte Wahrscheinlichkeiten und Unabhängigkeit

Um den Zusammenhang zwischen mehreren Ereignissen oder Zufallsvariablen zu beschreiben sind bedingte Wahrscheinlichkeiten von zentraler Bedeutung. In diesem Kapitel werden bedingte Wahrscheinlichkeiten eingeführt, und mehrstufige Modelle mithilfe bedingter Wahrscheinlichkeiten konstruiert. Anschließend werden wir den Begriff der Unabhängigkeit von Ereignissen und Zufallsvariablen systematisch einführen, und erste wichtige Aussagen unter Unabhängigkeitsannahmen herleiten.

2.1. Bedingte Wahrscheinlichkeiten

Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein fester Wahrscheinlichkeitsraum, und seien $A, B \in \mathcal{A}$ Ereignisse. Angenommen, wir wissen bereits, dass das Ereignis B eintritt, und wir wollen die Wahrscheinlichkeit von A unter dieser Prämisse angeben. Dann sollten wir nur noch die Fälle $\omega \in B$ in Betracht ziehen, und für diese tritt das Ereignis ein, wenn ω in $A \cap B$ enthalten ist. Damit ist die folgende Definition naheliegend:

Definition 2.1. Sei $A, B \in \mathcal{A}$ mit $\mathbb{P}[B] \neq 0$. Dann heißt

$$\mathbb{P}[A|B] \;:=\; \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}$$

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

Eine weitere Motivation für die Definition liefern relative Häufigkeiten: Ist $\mathbb{P}$ eine empirische Verteilung, dann sind $\mathbb{P}[A \cap B]$ und $\mathbb{P}[B]$ die relativen Häufigkeiten von $A \cap B$ und B, und $\mathbb{P}[A|B]$ ist damit die relative Häufigkeit von $A \cap B$ unter Elementen aus B. Die Definition ist also auch konsistent mit einer frequentistischen Interpretation der Wahrscheinlichkeit als Grenzwert von relativen Häufigkeiten.

Bemerkung. a) Der Fall $\mathbb{P}[B] = 0$ muss ausgeschlossen werden, da sonst sowohl Zähler als auch Nenner in dem Bruch in der Definition gleich 0 sind. Bedingte Wahrscheinlichkeiten gegeben Nullmengen sind im Allgemeinen nicht wohldefiniert.

b) Ist $\mathbb{P}[B] \neq 0$, dann ist durch die Abbildung

$$\mathbb{P}[\,\bullet\,|B]\colon\; A \mapsto \mathbb{P}[A|B]$$

wieder eine Wahrscheinlichkeitsverteilung auf $(\Omega, \mathcal{A})$ gegeben, die ***bedingte Verteilung unter*** $\mathbb{P}$ ***gegeben*** B . Der Erwartungswert

$$\mathbb{E}[X|B] \;=\; \sum_{a \in S} a \cdot \mathbb{P}[X = a|B]$$

einer diskreten Zufallsvariable $X\colon \Omega \to S$ bzgl. der bedingten Verteilung heißt ***bedingte Erwartung von*** X ***gegeben*** B.

Beispiel (Gleichverteilung). Ist $\mathbb{P}$ die Gleichverteilung auf einer endlichen Menge Ω, dann gilt:

$$\mathbb{P}[A|B] \;=\; \frac{|A\cap B|/|\Omega|}{|B|/|\Omega|} \;=\; \frac{|A\cap B|}{|B|} \qquad \text{für alle } A, B \subseteq \Omega.$$

Erste Anwendungsbeispiele

Bei der mathematischen Modellierung von Anwendungsproblemen unter Verwendung bedingter Wahrscheinlichkeiten können leicht Fehler auftreten. An dieser Stelle sollte man also sehr sorgfältig argumentieren, und ggf. zur Kontrolle verschiedene Modellvarianten verwenden. Wir betrachten einige bekannte Beispiele.

Beispiel (Mädchen oder Junge). Wie groß ist die Wahrscheinlichkeit, dass in einer Familie mit zwei Kindern beide Kinder Mädchen sind, wenn mindestens eines der Kinder ein Mädchen ist ? Hier können wir als Wahrscheinlichkeitsraum

$$S \;=\; \{JJ, JM, MJ, MM\}$$

ansetzen. Wir nehmen vereinfachend an, daß alle Fälle gleich wahrscheinlich sind. Dann gilt:

$$\mathbb{P}[\text{„beide Mädchen“} \mid \text{„mindestens ein Mädchen“}] \;=\; \frac{|\{MM\}|}{|\{MM, JM, MJ\}|} \;=\; \frac{1}{3}.$$

Wir modifizieren die Fragestellung nun etwas. Angenommen, im Nachbarhaus ist heute eine neue Familie eingezogen. Alles, was wir wissen, ist, daß die Familie zwei Kinder hat. Nun sehen wir am Fenster ein Mädchen winken, und gehen davon aus, daß dies eines der beiden Kinder ist. Wie hoch ist nun die Wahrscheinlichkeit, daß beide Kinder Mädchen sind ? Die naheliegende Antwort $1/3$ ist in diesem Fall nicht richtig. Dadurch, daß eines der Kinder winkt, sind die Kinder für uns nicht mehr ununterscheidbar. Die Wahrscheinlichkeit, dass das zweite (nicht winkende) Kind ein Mädchen ist, beträgt dann $1/2$:

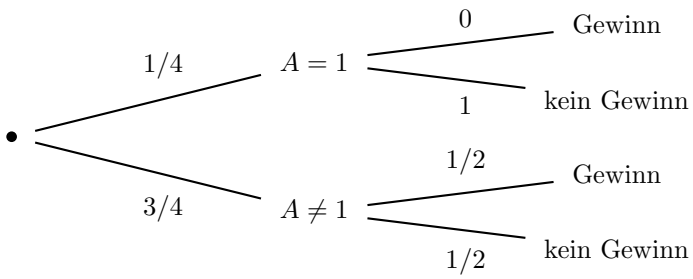
$$\mathbb{P}[\text{„beide Mädchen“} \mid \text{„das erste ist Mädchen“}] \;=\; \frac{|\{MM\}|}{|\{MM, MJ\}|} \;=\; \frac{1}{2}.$$

Haben wir noch Zweifel an der Richtigkeit dieser Aussage, könnten wir ein präziseres Modell aufstellen. Beispielsweise könnten wir das Geschlecht des älteren und des jüngeren Kindes durch Zufallsvariablen $X_1, X_2 : \Omega \to \{M, J\}$, und die Auswahl des winkenden Kindes durch eine weitere Zufallsvariable $K : \Omega \to \{1, 2\}$ beschreiben, wobei $K = 1, 2$ bedeutet, dass das ältere bzw. jüngere Kind winkt. Nehmen wir an, dass (X_1, X_2, K) gleichverteilt auf der Menge $\{M, J\}^2 \times \{1, 2\}$ ist, dann ergibt sich

$$\mathbb{P}[\text{„beide Mädchen“} \mid \text{„Mädchen winkt“}] \;=\; \frac{\mathbb{P}[X_1 = X_2 = M]}{\mathbb{P}[X_K = M]} \;=\; \frac{2/8}{4/8} \;=\; \frac{1}{2}.$$

Beispiel (Ziegenproblem). In einer leicht abgewandelten Version der Spielshow „Let's make a deal“ steht hinter einer von vier Türen ein Auto, und hinter den drei anderen Türen eine Ziege. Der Kandidat wählt zunächst eine der Türen aus (Tür 1). Anschließend öffnet der Moderator eine der verbleibenden Türen (Tür 2, 3 oder 4), wobei nie die Tür mit dem Auto geöffnet wird. Nun hat der Kandidat die Möglichkeit, die Tür nochmal zu wechseln, oder bei seiner ursprünglichen Wahl zu bleiben. Was ist die günstigere Strategie um das Auto zu gewinnen ?

Sie A die Nummer der Tür mit dem Auto. Bleibt der Kandidat bei seiner ursprünglichen Wahl, dann beträgt die Gewinnwahrscheinlichkeit offensichtlich $1/4$, da er bei zufälliger Position des Autos zu Beginn mit Wahrscheinlichkeit $1/4$ die richtige Tür gewählt hat. Die Situation beim Wechseln können wir uns durch das folgende Baumdiagramm klarmachen:



Steht das Auto hinter Tür 1, dann gewinnt der Spieler beim Wechseln nie. Steht das Auto dagegen hinter einer anderen Tür, dann öffnet der Moderator eine weitere Tür. Damit bleiben beim Wechseln nur noch zwei Türen zur Auswahl, und der Spieler gewinnt in diesem Fall mit Wahrscheinlichkeit $1/2$. Insgesamt beträgt die Gewinnwahrscheinlichkeit mit Wechseln also

$$p \;=\; \frac{1}{4}\cdot 0 + \frac{3}{4}\cdot\frac{1}{2} \;=\; \frac{3}{8},$$

d.h. Wechseln ist für den Kandidaten vorteilhaft.

Formal könnten wir die Situation durch Zufallsvariablen $A, M : \Omega \to \{1,2,3,4\}$ beschreiben, die die Nummern der Tür mit dem Auto und der vom Moderator geöffneten Tür angeben. Es ist dann naheliegend anzusetzen, dass A gleichverteilt ist, während M gegeben A bedingt gleichverteilt auf $\{2,3,4\} \setminus A$ ist, d.h.

$$\mathbb{P}[M = k | A = 1] = 1/3 \quad \text{für } k \neq 1, \qquad \mathbb{P}[M = k | A = 2] = \begin{cases} 1/2 & \text{für } k = 3,4, \\ 0 & \text{sonst,} \end{cases} \qquad \text{usw.}$$

Prüfen Sie selbst nach, dass sich in diesem Modell

$$\mathbb{P}[A = k | M \neq k] \;=\; 3/8 \qquad \text{für } k = 2,3,4$$

ergibt, d.h. bei Wechseln zu einer Tür $k \neq 1$, die der Moderator nicht geöffnet hat, beträgt die Gewinnwahrscheinlichkeit $3/8$.

Beispiel (Münzwürfe mit partieller Information). Bei 20 fairen Münzwürfen fällt 15-mal „Zahl“. Wie groß ist die Wahrscheinlichkeit, dass die ersten 5 Würfe „Zahl“ ergeben haben ? Sei $\mathbb{P}$ die Gleichverteilung auf

$$\Omega = \{0,1\}^{20} = \{\omega = (x_1, \dots x_{20}) : x_i \in \{0,1\}\},$$

und sei $X_i(\omega) = x_i$ der Ausgang des i-ten Wurfs. Dann gilt:

$$\mathbb{P}\left[X_1 = \ldots = X_5 = 1 \;\middle|\; \sum_{i=1}^{20} X_i = 15\right] = \frac{\mathbb{P}\left[X_1 = \ldots = X_5 = 1 \text{ und } \sum_{i=6}^{20} X_i = 10\right]}{\mathbb{P}\left[\sum_{i=1}^{20} X_i = 15\right]}$$
$$= \frac{\binom{15}{10}}{\binom{20}{15}} = \frac{15 \cdot 14 \cdot \dots \cdot 11}{20 \cdot 19 \cdot \dots \cdot 16} \approx \frac{1}{5}.$$

Dagegen ist $\mathbb{P}[X_1 = \ldots = X_5 = 1] = 1/32$.

Berechnung von Wahrscheinlichkeiten durch Fallunterscheidung

Wir zeigen nun wie man unbedingte Wahrscheinlichkeiten aus bedingten berechnet. Sei $\Omega = \bigcup H_i$ eine disjunkte Zerlegung von Ω in abzählbar viele Teilmengen H_i , $i \in I$. Die Mengen H_i beschreiben unterschiedliche Fälle (oder auch *Hypothesen* in statistischen Anwendungen).

Satz 2.2 (Formel von der totalen Wahrscheinlichkeit). Für alle $A \in \mathcal{A}$ gilt:

$$\mathbb{P}[A] = \sum_{\substack{i \in I \\ \mathbb{P}[H_i] \neq 0}} \mathbb{P}[A|H_i] \cdot \mathbb{P}[H_i] \tag{2.1}$$

Beweis. Es ist $A = A \cap (\bigcup_{i\in I} H_i) = \bigcup_{i\in I}(A \cap H_i)$ eine disjunkte Vereinigung, also folgt aus der σ-Additivität und wegen $\mathbb{P}[A \cap H_i] \leq \mathbb{P}[H_i]$:

$$\mathbb{P}[A] = \sum_{i\in I} \mathbb{P}[A \cap H_i] = \sum_{\substack{i\in I, \\ \mathbb{P}[H_i]\neq 0}} \mathbb{P}[A \cap H_i] = \sum_{\substack{i\in I, \\ \mathbb{P}[H_i]\neq 0}} \mathbb{P}[A|H_i] \cdot \mathbb{P}[H_i].$$

Beispiel (Zweistufiges Urnenmodell). Urne 1 enthalte 2 rote und 3 schwarze Kugeln, Urne 2 enthalte 3 rote und 4 schwarze Kugeln. Wir legen eine Kugel K_1 von Urne 1 in Urne 2 und ziehen eine Kugel K_2 aus Urne 2. Mit welcher Wahrscheinlichkeit ist K_2 rot?

Durch Bedingen auf die Farbe der ersten Kugel erhalten wir nach Satz 2.2:

$$\mathbb{P}[K_2 \text{ rot}] = \mathbb{P}[K_2 \text{ rot} \mid K_1 \text{ rot}] \cdot \mathbb{P}[K_1 \text{ rot}] + \mathbb{P}[K_2 \text{ rot} \mid K_1 \text{ schwarz}] \cdot \mathbb{P}[K_1 \text{ schwarz}]$$
$$= \frac{4}{8} \cdot \frac{2}{5} + \frac{3}{8} \cdot \frac{3}{5} = \frac{17}{40}.$$

Ein interessanter Effekt ist, dass bei Wechsel der zugrundeliegenden Wahrscheinlichkeitsverteilung die unbedingte Wahrscheinlichkeit eines Ereignisses A selbst dann abnehmen kann, wenn alle bedingten Wahrscheinlichkeiten in (2.1) zunehmen:

Beispiel (Simpson-Paradoxon). Die folgende (im wesentlichen auf Originaldaten basierende) Tabelle zeigt die Zahl der Bewerber und der aufgenommenen Studierenden an der Universität Berkeley in einem bestimmten Jahr:

Bewerbungen in Berkeley						
Statistik 1:	Männer		angenommen (A)	Frauen		angenommen (A)
Empirische Verteilung:	2083		996	1067		349
	$\mathbb{P}[A\|M]$	$\approx$	0,48	$\mathbb{P}[A\|F]$	$\approx$	0,33

Genauere Analyse durch Unterteilung in 4 Fachbereiche						
Statistik 2:	Männer	angenommen (A)		Frauen	angenommen (A)	
Bereich 1	825	511	62%	108	89	82%
Bereich 2	560	353	63%	25	17	68%
Bereich 3	325	110	34%	593	219	37%
Bereich 4	373	22	6%	341	24	7%

Sei $\mathbb{P}_F[A] = \mathbb{P}[A|F]$ die relative Häufigkeit der angenommenen Bewerber unter Frauen, und $\mathbb{P}_M[A] = \mathbb{P}[A|M]$ die entsprechende Annahmequote unter Männern. Hierbei steht $\mathbb{P}$ für die zugrundeliegende empirische Verteilung, und $\mathbb{P}_F$ sowie $\mathbb{P}_M$ sind dementsprechend die empirischen Verteilungen in den Unterpopulationen der weiblichen und männlichen Bewerber. Die vollständige Aufgliederung nach Fachbereichen ergibt folgende Zerlegung in Hypothesen:

$$\mathbb{P}_M[A] = \sum_{i=1}^{4} \mathbb{P}_M[A|H_i]\,\mathbb{P}_M[H_i], \qquad \mathbb{P}_F[A] = \sum_{i=1}^{4} \mathbb{P}_F[A|H_i]\,\mathbb{P}_F[H_i].$$

Im Beispiel ist $\mathbb{P}_F[A|H_i] > \mathbb{P}_M[A|H_i]$ für *alle* i, aber *dennoch* $\mathbb{P}_F[A] < \mathbb{P}_M[A]$. Obwohl die Annahmequoten unter männlichen Bewerbern insgesamt höher sind, schneiden also die Frauen in jedem der Fachbereiche besser ab.

Die Gesamtstatistik im Beispiel vermischt verschiedene Populationen und legt deshalb eventuell eine falsche Schlussfolgerung nahe. Bei statistischen Untersuchungen ist es daher wichtig, die Population zunächst in möglichst homogene Unterpopulationen aufzuspalten.

Das Simpson-Paradox tritt auch an vielen anderen Stellen auf. Beispielsweise kann bei der Steuerprogression der Steueranteil insgesamt steigen obwohl der Steuersatz in jeder Einkommensklasse sinkt, weil Personen in höhere Einkommensklassen aufsteigen.

Bayessche Regel

Eine direkte Konsequenz des Satzes von der totalen Wahrscheinlichkeit ist die Bayessche Regel. Wir betrachten erneut eine disjunkte Zerlegung von Ω in Teilmengen (Hypothesen) H_i.

Wie wahrscheinlich sind die Hypothesen H_i ? Ohne zusätzliche Information ist $\mathbb{P}[H_i]$ die Wahrscheinlichkeit von H_i. In der Bayesschen Statistik interpretiert man $\mathbb{P}[H_i]$ als unsere subjektive Einschätzung (aufgrund von vorhandenem oder nicht vorhandenem Vorwissen) über die vorliegende Situation („a priori degree of belief“).

Angenommen, wir wissen nun zusätzlich, dass ein Ereignis $A \in \mathcal{A}$ mit $\mathbb{P}[A] \neq 0$ eintritt, und wir kennen die bedingte Wahrscheinlichkeit („likelihood“) $\mathbb{P}[A|H_i]$ für das Eintreten von A unter der Hypothese H_i für jedes $i \in I$ mit $\mathbb{P}[H_i] \neq 0$. Wie sieht dann unsere neue Einschätzung der Wahrscheinlichkeiten der H_i („a posteriori degree of belief“) aus?

Korollar 2.3 (Bayessche Regel). Für $A \in \mathcal{A}$ mit $\mathbb{P}[A] \neq 0$ ist

$$\mathbb{P}[H_i|A] \;=\; \frac{\mathbb{P}[A|H_i]\cdot\mathbb{P}[H_i]}{\sum\limits_{\substack{k\in I\\ \mathbb{P}[H_k]\neq 0}} \mathbb{P}[A|H_k]\cdot\mathbb{P}[H_k]} \qquad \text{für alle } i \in I \text{ mit } \mathbb{P}[H_i] \neq 0,$$

d.h. es gilt die Proportionalität

$$\mathbb{P}[H_i|A] \;=\; c\cdot\mathbb{P}[H_i]\cdot\mathbb{P}[A|H_i],$$

wobei c eine von i unabhängige Konstante ist.

Beweis. Nach Satz 2.2 und der Definition der bedingten Wahrscheinlichkeit erhalten wir

$$\mathbb{P}[H_i|A] \;=\; \frac{\mathbb{P}[A\cap H_i]}{\mathbb{P}[A]} \;=\; \frac{\mathbb{P}[A|H_i]\cdot\mathbb{P}[H_i]}{\sum\limits_{\substack{k\in I\\ \mathbb{P}[H_k]\neq 0}} \mathbb{P}[A|H_k]\cdot\mathbb{P}[H_k]}.$$

Die Bayessche Regel besagt, dass die *A-posteriori-Wahrscheinlichkeiten* $\mathbb{P}[H_i|A]$ als Funktion von i proportional zum Produkt der *A-priori-Wahrscheinlichkeiten* $\mathbb{P}[H_i]$ und der *Likelihood-Funktion* $i \mapsto \mathbb{P}[A|H_i]$ sind. In dieser und ähnlichen Formen bildet sie das Fundament der Bayesschen Statistik.

Beispiel (Medizinische Tests). Von 10.000 Personen eines Alters habe einer die Krankheit K. Ein Test sei positiv (+) bei 96% der Kranken und bei 0,1% der Gesunden. Liegen keine weiteren Informationen vor (z.B. über Risikofaktoren), dann ergibt sich für die A-priori-und A-Posteriori-Wahrscheinlichkeiten für die Krankheit K vor und nach einem positiven Test:

$$\begin{array}{rrcl rrcl} \text{A priori:} & \mathbb{P}[K] & = & 0,0001, & \mathbb{P}[K^C] & = & 0,9999.\\ \text{Likelihood:} & \mathbb{P}[+|K] & = & 0,96, & \mathbb{P}[+|K^C] & = & 0,001. \end{array}$$

$$\begin{aligned} \text{A posteriori:} \qquad \mathbb{P}[K|+] \;&=\; \frac{\mathbb{P}[+|K]\cdot\mathbb{P}[K]}{\mathbb{P}[+|K]\cdot\mathbb{P}[K]+\mathbb{P}[+|K^C]\cdot\mathbb{P}[K^C]}\\ &=\; \frac{0,96\cdot 10^{-4}}{0,96\cdot 10^{-4}+10^{-3}\cdot 0,9999} \;\approx\; \frac{1}{11}. \end{aligned}$$

Daraus folgt insbesondere: $\mathbb{P}[K^C|+] \approx \frac{10}{11}$, d.h. ohne zusätzliche Informationen (z.B. durch einen weiteren Test) muss man in diesem Fall davon ausgehen, dass $\frac{10}{11}$ der positiv getesteten Personen in Wirklichkeit gesund sind!

2.2. Mehrstufige Modelle

Wir betrachten nun ein n-stufiges Zufallsexperiment. Der Ausgang des k-ten Teilexperiments ($k = 1,\dots,n$) werde durch eine Zufallsvariable $X_k : \Omega \to S_k$ auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ beschrieben, wobei wir wieder voraussetzen, dass der Wertebereich S_k abzählbar ist. Wir nehmen an, dass folgendes gegeben ist:

- Die *Verteilung bzw. Massenfunktion von* X_1:

$$\mathbb{P}[X_1 = x_1] \;=\; p_1(x_1) \qquad \text{für alle } x_1 \in S_1, \qquad \text{sowie} \tag{2.2}$$

- die *bedingten Verteilungen/Massenfunktionen von X_k gegeben $X_1, \dots, X_{k-1}$*:

$$\mathbb{P}[X_k = x_k \mid X_1 = x_1, \dots, X_{k-1} = x_{k-1}] = p_k(x_k \mid x_1, \dots, x_{k-1}) \tag{2.3}$$

für $k = 2, \dots n$ und alle $x_1 \in S_1, \dots, x_k \in S_k$ mit $\mathbb{P}[X_1 = x_1, \dots, X_{k-1} = x_{k-1}] \neq 0$.

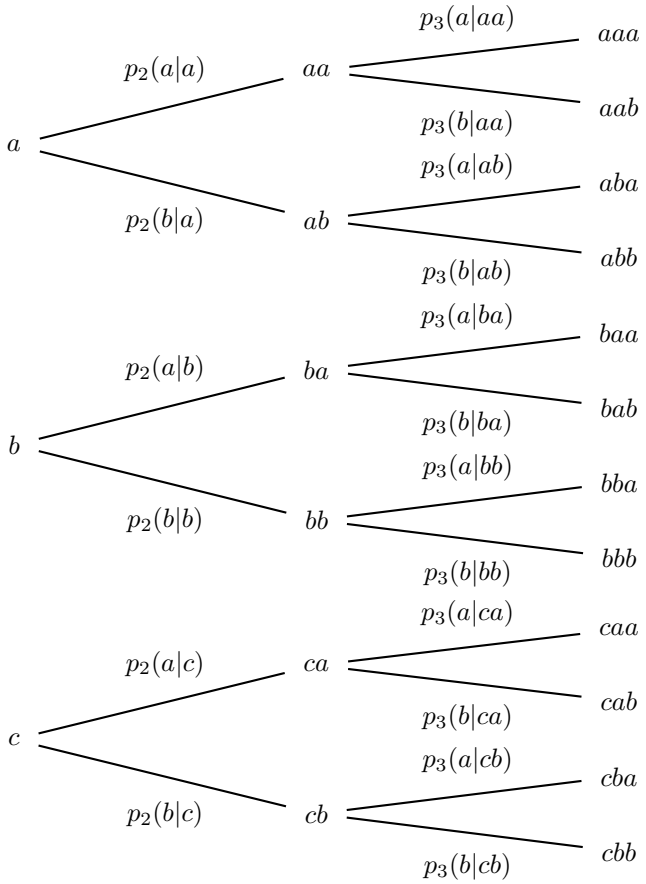


Abbildung 2.1.: Dreistufiges Modell mit $S_1 = \{a, b, c\}$ und $S_2 = S_3 = \{a, b\}$.

Zwei wichtige Spezialfälle sind

(i) *Produktmodelle*, in denen die bedingten Massenfunktionen $p_k(\bullet|x_1,\dots,x_{k-1})$ nicht von den vorherigen Werten $x_1,\dots,x_{k-1}$ abhängen, sowie

(ii) *Markovketten*, bei denen $p_k(\bullet|x_1,\dots,x_{k-1})$ nur vom letzten Zustand x_{k-1} abhängt.

Das kanonische Modell

Zufallsvariablen $X_1,\dots,X_n$, die (2.2) und (2.3) erfüllen, kann man zu gegebenen Massenfunktionen auf unterschiedlichen Wahrscheinlichkeitsräumen realisieren. Im „kanonischen Modell" realisiert man die Zufallsvariablen als Koordinatenabbildungen

$$X_k(\omega) \;=\; \omega_k, \qquad k = 1,\dots,n,$$

auf dem mit der σ-Algebra $\mathcal{A} = \mathcal{P}(\Omega)$ versehenen Produktraum

$$\Omega \;=\; S_1 \times \ldots \times S_n \;=\; \{(\omega_1,\dots,\omega_n) \,:\, \omega_i \in S_i\}.$$

Satz 2.4 (Kanonisches Mehrstufenmodell). Seien p_1 und $p_k(\,\bullet \mid x_1,\dots,x_{k-1})$ für jedes $k = 2,\dots,n$ und $x_1 \in S_1,\dots,x_{k-1} \in S_{k-1}$ Massenfunktionen von Wahrscheinlichkeitsverteilungen auf S_k. Dann existiert genau eine Wahrscheinlichkeitsverteilung $\mathbb{P}$ auf dem Produktraum $(\Omega,\mathcal{A})$ mit (2.2) und (2.3). Diese ist bestimmt durch die Massenfunktion

$$p(x_1,\dots,x_n) = p_1(x_1)\,p_2(x_2 \mid x_1)\,p_3(x_3 \mid x_1,x_2)\cdots p_n(x_n \mid x_1,\dots,x_{n-1}).$$

Beweis. Eindeutigkeit: Wir zeigen durch Induktion, dass für eine Verteilung $\mathbb{P}$ mit (2.2) und (2.3) und $k = 1,\dots n$ gilt:

$$\mathbb{P}[X_1 = x_1,\dots,X_k = x_k] \;=\; p_1(x_1)\cdot p_2(x_2 \mid x_1)\cdots p_k(x_k \mid x_1,\dots,x_{k-1}). \tag{2.4}$$

Nach (2.2) ist dies für $k = 1$ der Fall. Zudem folgt aus (2.4) für $k-1$ nach (2.3):

$$\begin{aligned}\mathbb{P}[X_1 = x_1,\dots,X_k = x_k] \;&= \mathbb{P}[X_1 = x_1,\dots,X_{k-1} = x_{k-1}]\\ &\quad\cdot \mathbb{P}[X_1 = x_1,\dots,X_k = x_k \mid X_1 = x_1,\dots,X_{k-1} = x_{k-1}]\\ &= p_1(x_1)\cdot p_2(x_2 \mid x_1)\cdots p_{k-1}(x_{k-1} \mid x_1,\dots,x_{k-2})\\ &\quad\cdot p_k(x_k \mid x_1,\dots,x_{k-1}),\end{aligned}$$

also die Behauptung (2.4) für k, falls $\mathbb{P}[X_1 = x_1,\dots,X_{k-1} = x_{k-1}] \neq 0$. Andernfalls verschwinden beide Seiten in (2.4) und die Behauptung ist trivialerweise erfüllt. Für $k = n$ erhalten wir die Massenfunktion von $\mathbb{P}$:

$$\mathbb{P}[X_1 = x_1,\dots,X_n = x_n] \;=\; p_1(x_1)\cdots p_n(x_n \mid x_1,\dots,x_{n-1}) \;=\; p(x_1,\dots,x_n).$$

Existenz: Die Funktion p ist Massenfunktion einer Wahrscheinlichkeitsverteilung $\mathbb{P}$ auf Ω, denn die Gewichte $p(x_1,\dots,x_n)$ sind nach Voraussetzung nichtnegativ mit

$$\begin{aligned}\sum_{x_1\in S_1}\cdots\sum_{x_n\in S_n} p(x_1,\dots,x_n) \;&=\; \sum_{x_1\in S_1} p_1(x_1)\sum_{x_2\in S_2} p_2(x_2 \mid x_1)\cdots \underbrace{\sum_{x_n\in S_n} p_n(x_n \mid x_1,\dots,x_n)}_{=n-1}\\ &=\; 1.\end{aligned}$$

Hierbei haben wir benutzt, dass die Funktionen $p_k(\bullet|x_1,\dots,x_{k-1})$ Massenfunktionen von Wahrscheinlichkeitsverteilungen auf S_k sind. Für die Wahrscheinlichkeitsverteilung $\mathbb{P}$ auf Ω gilt

$$\begin{aligned}\mathbb{P}[X_1 = x_1, \dots X_k = x_k] &= \sum_{x_{k+1}\in S_{k+1}} \dots \sum_{x_n \in S_n} p(x_1,\dots,x_n) \\ &= p_1(x_1)\, p_2(x_2 \mid x_1) \cdots p_k(x_k \mid x_1,\dots,x_{k-1})\end{aligned}$$

für $k = 1,\dots,n$. Hieraus folgt, dass $\mathbb{P}$ die Bedingungen (2.2) und (2.3) erfüllt. ■

Beispiel (Skat). Wie groß ist die Wahrscheinlichkeit, dass beim Skat jeder Spieler genau einen der vier Buben erhält ? Wir beschreiben die Anzahl der Buben der drei Spieler durch die Zufallsvariablen $X_i(\omega) = \omega_i$, $i = 1,2,3$, auf dem Produktraum

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

Da es insgesamt 32 Karten gibt, von denen jeder Spieler 10 erhält, sind die bedingten Verteilungen der Zufallvariablen X_1, X_2 und X_3 gegeben durch die hypergeometrischen Verteilungen

$$\begin{aligned}p_1(x_1) &= \binom{4}{x_1}\binom{28}{10-x_1}\Big/\binom{32}{10}, \\ p_2(x_2 \mid x_1) &= \binom{4-x_1}{x_2}\binom{18+x_1}{10-x_2}\Big/\binom{22}{10} \text{ falls } x_1+x_2 \le 4,\ 0 \text{ sonst, sowie} \\ p_3(x_3 \mid x_1,x_2) &= \binom{4-x_1-x_2}{x_3}\binom{8+x_1+x_2}{10-x_3}\Big/\binom{12}{10} \text{ falls } 2 \le x_1+x_2+x_3 \le 4,\ 0 \text{ sonst.}\end{aligned}$$

Damit erhalten wir für die gesuchte Wahrscheinlichkeit

$$p(1,1,1) = p_1(1)\, p_2(1 \mid 1)\, p_3(1 \mid 1,1) \approx 5,56\%.$$

Produktmodelle

Hängt der Ausgang des i-ten Teilexperiments nicht von $x_1,\dots,x_{i-1}$ ab, dann gilt

$$p_i(x_i \mid x_1,\dots,x_{i-1}) = p_i(x_i)$$

mit einer von $x_1,\dots,x_{i-1}$ unabhängigen Massenfunktion p_i einer Wahrscheinlichkeitsverteilung $\mathbb{P}_i$ auf S_i. Sind alle Teilexperimente voneinander unabhängig, dann hat die Wahrscheinlichkeitsverteilung $\mathbb{P}$ eines kanonischen n-stufigen Modells die Massenfunktion

$$p(x_1,\dots,x_n) = \prod_{i=1}^{n} p_i(x_i), \qquad x \in S_1 \times \cdot \times S_n. \tag{2.5}$$

Definition 2.5. Seien $\mathbb{P}_i$, $i = 1,\dots n$, Wahrscheinlichkeitsverteilungen auf abzählbaren Mengen S_i mit Massenfunktionen p_i. Die durch die Massenfunktion (2.5) bestimmte Wahrscheinlichkeitsverteilung $\mathbb{P} = \mathbb{P}_1 \otimes \dots \otimes \mathbb{P}_n$ auf $\Omega = S_1 \times \dots \times S_n$ heißt **Produkt** von $\mathbb{P}_1,\dots,\mathbb{P}_n$.

Beispiel (n-dimensionale Bernoulli-Verteilung). Wir betrachten n unabhängige 0-1-Experimente mit Erfolgswahrscheinlichkeit p, und setzen entsprechend

$$S_i = \{0,1\}, \qquad p_i(1) = p, \qquad p_i(0) = 1-p \qquad \text{für } i = 1,\dots,n.$$

Sei $k = \sum_{i=1}^n x_i$ die Anzahl der Einsen in einem n-Tupel $x \in \Omega = \{0,1\}^n$. Dann hat die Verteilung im Produktmodell die Massenfunktion

$$p(x_1, \ldots, x_n) \;=\; \prod_{i=1}^n p_i(x_i) \;=\; p^k\,(1-p)^{n-k},$$

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

Beispiel (Produkt von Gleichverteilungen). Sind die Mengen S_i, $i = 1, \ldots, n$, endlich, und ist $\mathbb{P}_i$ die Gleichverteilung auf S_i, dann ist $\mathbb{P}_1 \otimes \cdots \otimes \mathbb{P}_n$ die Gleichverteilung auf dem Produktraum $S_1 \times \ldots \times S_n$.

Die Multiplikativität gilt in Produktmodellen nicht nur für die Massenfunktionen, sondern allgemeiner für die Wahrscheinlichkeiten, dass in den Teilexperimenten bestimmte Ereignisse $A_1, \ldots, A_n$ eintreten:

Satz 2.6. Bezüglich des Produkts $\mathbb{P} = \mathbb{P}_1 \otimes \cdots \otimes \mathbb{P}_n$ gilt für beliebige Ereignisse $A_i \subseteq S_i, i = 1, \ldots, n$:

$$\begin{array}{rcl} \mathbb{P}[X_1 \in A_1, \ldots, X_n \in A_n] & = & \displaystyle\prod_{i=1}^n \mathbb{P}[X_i \in A_i] \\ || & & || \\ \mathbb{P}[A_1 \times \ldots \times A_n] & & \displaystyle\prod_{i=1}^n \mathbb{P}_i[A_i] \end{array} \tag{2.6}$$

Beweis. Wegen $(X_1, \ldots, X_n)(\omega) = (\omega_1, \ldots, \omega_n) = \omega$ ist $(X_1, \ldots, X_n)$ die identische Abbildung auf dem Produktraum, und es gilt

$$\begin{aligned} \mathbb{P}[X_1 \in A_1, \ldots, X_n \in A_n] \;&=\; \mathbb{P}[(X_1, \ldots, X_n) \in A_1 \times \cdots \times A_n] \;=\; \mathbb{P}[A_1 \times \cdots \times A_n] \\ &= \sum_{x \in A_1 \times \cdots \times A_n} p(x) \;=\; \sum_{x_1 \in A_1} \cdots \sum_{x_n \in A_n} \prod_{i=1}^n p_i(x_i) \\ &= \prod_{i=1}^n \sum_{x_i \in A_i} p_i(x_i) \;=\; \prod_{i=1}^n \mathbb{P}_i[A_i]. \end{aligned}$$

Insbesondere folgt

$$\mathbb{P}[X_i \in A_i] \;=\; \mathbb{P}[X_1 \in S_1, \ldots, X_{i-1} \in S_{i-1}, X_i \in A_i, X_{i+1} \in S_{i+1}, \ldots, X_n \in S_n] \;=\; \mathbb{P}_i[A_i],$$

für jedes $i \in \{1, \ldots n\}$, und damit die Behauptung. ■

Bemerkung (Unabhängigkeit). Satz 2.6 besagt, dass die Koordinatenabbildungen $X_i(\omega) = \omega_i$ im Produktmodell *unabhängige Zufallsvariablen* sind, siehe Abschnitt 2.3.

Markovketten

Zur Modellierung einer zufälligen zeitlichen Entwicklung mit abzählbarem Zustandsraum S betrachten wir den Stichprobenraum

$$\Omega \;=\; S^{n+1} \;=\; \{(x_0, x_1, \ldots, x_n) \;:\; x_i \in S\}.$$

Oft ist es naheliegend anzunehmen, dass die Weiterentwicklung des Systems nur vom gegenwärtigen Zustand, aber nicht vom vorherigen Verlauf abhängt („kein Gedächtnis“), d.h. es ist

$$p_k(x_k \mid x_0, \dots, x_{k-1}) \;=\; P_k(x_{k-1}, x_k), \tag{2.7}$$

wobei das „Bewegungsgesetz“ $P_k : S \times S \to [0,1]$ folgende Bedingungen erfüllt:

(i) $P_k(x,y) \geq 0$ für alle $x, y \in S$,

(ii) $\sum_{y \in S} P_k(x,y) = 1$ für alle $x \in S$.

Die Bedingungen (i) und (ii) besagen, dass $P_k(x, \bullet)$ für jedes $x \in S$ und $k \in \{1, \dots n\}$ die Massenfunktion einer Wahrscheinlichkeitsverteilung auf S ist. Diese Wahrscheinlichkeitsverteilung beschreibt die **Übergangswahrscheinlichkeiten** von einem Zustand x zum nächsten Zustand im k-ten Schritt. Die Übergangswahrscheinlichkeiten $P_k(x,y)$, $x, y \in S$, kann man in einer Matrix $P_k \in \mathbb{R}^{S \times S}$ zusammenfassen. Hat S unendlich viele Elemente, dann ist diese Matrix allerdings unendlich dimensional.

Definition 2.7. Eine Matrix $P_k = (P_k(x,y))_{x,y \in S} \in \mathbb{R}^{S \times S}$ mit (i) und (ii) heißt **stochastische Matrix** auf S.

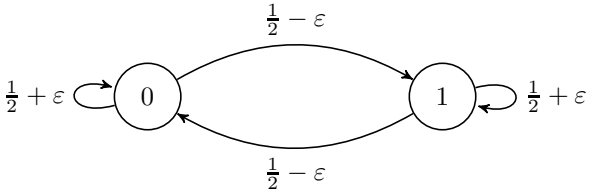
Sei $\nu : S \to [0,1]$ die Massenfunktion der Verteilung von X_0, also der **Startverteilung** der zufälligen Entwicklung. Als Massenfunktion des mehrstufigen Modells ergibt sich dann aus Gleichung (2.7):

$$p(x_0, x_1, \dots, x_n) \;=\; \nu(x_0)\, P_1(x_0, x_1)\, P_2(x_1, x_2) \cdots P_n(x_{n-1}, x_n) \qquad \text{für } x_0, \dots, x_n \in S,$$

Eine Folge $X_0, X_1, X_2, \dots, X_n$ von Zufallsvariablen, deren gemeinsame Verteilung durch das beschriebene mehrstufige Modell gegeben ist, nennt man eine **Markovkette** mit Übergangsmatrizen P_k, $k = 1, \dots, n$. Den Fall, in dem der Übergangsmechanismus $P_k(x,y) = P(x,y)$ unabhängig von k ist, bezeichnet man als ***zeitlich homogen***.

Beispiele. a) Produktmodell: Produktmodelle sind spezielle Markovketten mit Übergangswahrscheinlichkeiten $P_k(x,y) = p_k(y)$, die nicht von x abhängen.

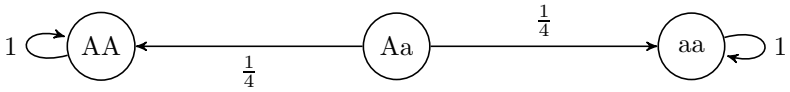
b) Abhängige Münzwürfe: Ein einfaches Modell für abhängige Münzwürfe ist eine Markovkette mit Zustandsraum $S = \{0,1\}$ und den folgenden Übergangswahrscheinlichkeiten:



Hierbei ist $\varepsilon \in \left[-\frac{1}{2}, \frac{1}{2}\right]$ ein Parameter, der die Abhängigkeit des nächsten Münzwurfs vom Ausgang des vorherigen Wurfs bestimmt. Die zeitunabhängige Übergangsmatrix ist

$$P \;=\; \begin{pmatrix} \frac{1}{2}+\varepsilon & \frac{1}{2}-\varepsilon \\ \frac{1}{2}-\varepsilon & \frac{1}{2}+\varepsilon \end{pmatrix}.$$

c) Selbstbefruchtung von Pflanzen: Die Selbstbefruchtung ist ein klassisches Verfahren zur Züchtung von Pflanzen vom Genotyp AA bzw. aa, wobei A und a zwei mögliche Allele des Pflanzen-Gens sind. Die Übergangswahrscheinlichkeiten zwischen den möglichen Genotypen AA, Aa und aa sind durch



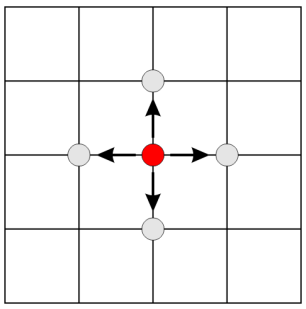
gegeben, und die Übergangsmatrix einer entsprechenden Markovkette ist

$$P = \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ 0 & 0 & 1 \end{pmatrix}.$$

d) Random Walks auf Graphen: Sei $S = V$ die Knotenmenge eines Graphen (V, E). Wir nehmen an, dass jeder Knoten $x \in V$ endlichen Grad $\deg(x)$ hat. Dann ist durch

$$P(x, y) = \begin{cases} \frac{1}{\deg(x)} & \text{falls } \{x, y\} \in E, \\ 0 & \text{sonst,} \end{cases}$$

die zeitunabhängige Übergangsmatrix eines Random Walks auf dem Graphen definiert. Beispielsweise ist der klassische Random Walk (Irrfahrt) auf $S = \mathbb{Z}^d$ die Markovkette, die sich in jedem Schritt zu einem zufällig (gleichverteilt) ausgewählten Nachbarpunkt des gegenwärtigen Zustands weiterbewegt:



Da in d Dimensionen jeder Gitterpunkt $2d$ Nachbarpunkte hat, sind die Übergangswahrscheinlichkeiten durch

$$P(x, y) = \begin{cases} \frac{1}{2d} & \text{falls } |x - y| = 1, \\ 0 & \text{sonst,} \end{cases}$$

gegeben. In Dimension $d = 1$ ist die Übergangsmatrix eine unendliche (mit $x \in \mathbb{Z}$ indizierte) Tridiagonalmatrix, die neben der Diagonale die Einträge $1/2$, und auf der Diagonalen die Einträge 0 hat.

Berechnung von Mehr-Schritt-Übergangswahrscheinlichkeiten

Wir berechnen nun die Übergangswahrscheinlichkeiten und Verteilungen einer Markovkette nach mehreren Schritten. Es stellt sich heraus, dass sich diese durch Matrizenmultiplikation der Übergangsmatrizen ergeben. Dazu interpretieren wir die Massenfunktion ν der Startverteilung als Zeilenvektor $(\nu(x))_{x \in S}$ in $\mathbb{R}^S$.

Satz 2.8 (Übergangswahrscheinlichkeiten und Verteilung nach mehreren Schritten).
Für alle $0 \le k < l \le n$ und $x_0, \dots, x_k, y \in S$ mit $\mathbb{P}[X_0 = x_0, \dots, X_k = x_k] \neq 0$ gilt

$$\begin{aligned} \mathbb{P}[X_l = y \mid X_0 = x_0, \dots, X_k = x_k] &= \mathbb{P}[X_l = y \mid X_k = x_k] \\ &= (P_{k+1}\, P_{k+2} \cdots P_l)(x_k, y), \qquad \text{und} \\ \mathbb{P}[X_l = y] &= (\nu P_1\, P_2 \cdots P_l)(y). \end{aligned}$$

Hierbei ist

$$(P\,\widetilde{P})(x,y) \;:=\; \sum_{z \in S} P(x,z)\,\widetilde{P}(z,y)$$

das Produkt zweier Übergangsmatrizen P und $\widetilde{P}$ an der Stelle (x, y), und

$$(\nu\widetilde{P})(y) \;=\; \sum_{x \in S} \nu(x)\widetilde{P}(x,y)$$

ist das Produkt des Zeilenvektors ν mit einer Übergangsmatrix $\widetilde{P}$, ausgewertet an der Stelle y.

Die Matrixprodukte in Satz 2.8 sind auch für abzählbar unendliche Zustandsräume S wohldefiniert, da die Komponenten der Übergangsmatrizen alle nicht-negativ sind.

Bemerkung. a) Markov-Eigenschaft: Der Satz zeigt, dass die Weiterentwicklung einer Markovkette auch für mehrere Schritte jeweils nur vom gegenwärtigen Zustand x_k abhängt, und nicht vom vorherigen Verlauf $x_0, x_1, \dots, x_{k-1}$.

b) n-Schritt-Übergangswahrscheinlichkeiten: Die Übergangswahrscheinlichkeiten für die ersten n Schritte sind nach dem Satz gegeben durch

$$\mathbb{P}[X_n = y \mid X_0 = x] \quad = \quad (P_1\, P_2 \cdots P_n)(x,y).$$

Im *zeitlich homogenen Fall* (d.h. $P_i \equiv P$ unabhängig von i) ist die n-Schritt-Übergangswahrscheinlichkeit von x nach y gleich $P^n(x,y)$.

c) Gleichgewichtsverteilungen: Weiterhin ist im *zeitlich homogenen Fall* $P_i \equiv P$ die Verteilung der Markovkette zur Zeit l gleich νP^l. Gilt $\nu = \nu P$, dann stimmt diese für jedes l mit der Startverteilung überein, d.h. die Wahrscheinlichkeitsverteilung ν ist ein *Gleichgewicht* der stochastischen Dynamik, die durch die Übergangsmatrix P beschrieben wird.

Beweis. Für $x_0, \dots, x_k, y$ wie im Satz vorausgesetzt gilt

$$\begin{aligned} \mathbb{P}[X_l = y \mid X_0 = x_0, \dots, X_k = x_k] &= \frac{\mathbb{P}[X_0 = x_0, \dots, X_k = x_k, X_l = y]}{\mathbb{P}[X_0 = x_0, \dots, X_k = x_k]} \\ &= \frac{\sum_{x_{k+1}, \dots, x_{l-1}} \nu(x_0)\, P_1(x_0,x_1) \cdots P_l(x_{l-1}, y)}{\nu(x_0)\, P_1(x_0,x_1) \cdots P_k(x_{k-1}, x_k)} \\ &= \sum_{x_{k+1}} \cdots \sum_{x_{l-1}} P_{k+1}(x_k, x_{k+1})\, P_{k+2}(x_{k+1}, x_{k+2}) \cdots P_l(x_{l-1}, y) \\ &= (P_{k+1}\, P_{k+2} \cdots P_l)(x_k, y). \end{aligned}$$

Entsprechend erhalten wir

$$\begin{aligned}\mathbb{P}[X_l = y \mid X_k = x_k] &= \frac{\mathbb{P}[X_k = x_k, X_l = y]}{\mathbb{P}[X_k = x_k]} \\ &= \frac{\sum_{x_1,\dots,x_{k-1}} \sum_{x_{k+1},\dots,x_{l-1}} \nu(x_0)\, P_1(x_0,x_1)\cdots P_l(x_{l-1},y)}{\sum_{x_1,\dots,x_{k-1}} \nu(x_0)\, P_1(x_0,x_1)\cdots P_k(x_{k-1},x_k)} \\ &= (P_{k+1}\, P_{k+2}\cdots P_l)(x_k,y).\end{aligned}$$

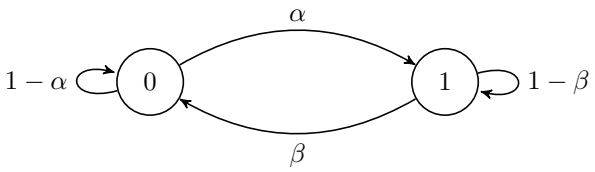
Für die unbedingten Wahrscheinlichkeiten ergibt sich

$$\begin{aligned}\mathbb{P}[X_l = y] &= \sum_{\substack{x\in S \\ \mathbb{P}[X_0=x]\neq 0}} \mathbb{P}[X_0 = x]\,\mathbb{P}[X_l = y \mid X_0 = x] \\ &= \sum_{\substack{x\in S \\ \nu(x)\neq 0}} \nu(x)\,(P_1\, P_2\cdots P_l)(x,y) = (\nu P_1\, P_2\cdots P_l)(y).\end{aligned}$$

■

Wir untersuchen abschließend den Spezialfall einer zeithomogenen Markovkette auf einem Zustandsraum mit zwei Elementen. Diesen können wir schon jetzt weitgehend vollständig analysieren:

Beispiel (Explizite Berechnung für Zustandsraum mit zwei Elementen). Wir betrachten eine allgemeine zeithomogene Markovkette mit Zustandsraum $S = \{0,1\}$. Die Übergangswahrscheinlichkeiten $P(x,y)$ sind durch



gegeben, wobei wir annehmen, dass $0 < \alpha, \beta \leq 1$ gilt. Die Wahrscheinlichkeitsverteilung μ mit Gewichten $\mu(0) = \frac{\beta}{\alpha+\beta}$ und $\mu(1) = \frac{\alpha}{\alpha+\beta}$ ist ein Gleichgewicht der Übergangsmatrix

$$P = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix},$$

denn für den Zeilenvektor $\mu = (\mu(0), \mu(1))$ gilt $\mu P = \mu$. Für $n \in \mathbb{N}$ erhalten wir durch Bedingen auf den Wert zur Zeit $n-1$:

$$\begin{aligned}P^n(0,0) &= P^{n-1}(0,0)\cdot P(0,0) + P^{n-1}(0,1)\cdot P(1,0) \\ &= P^{n-1}(0,0)\cdot(1-\alpha) + (1 - P^{n-1}(0,0))\cdot\beta \\ &= (1-\alpha-\beta)\cdot P^{n-1}(0,0) + \beta.\end{aligned}$$

Daraus folgt mit Induktion

$$\begin{aligned}P^n(0,0) &= \frac{\beta}{\alpha+\beta} + \frac{\alpha}{\alpha+\beta}\,(1-\alpha-\beta)^n, \qquad \text{und} \\ P^n(0,1) &= 1 - P^n(0,0) = \frac{\alpha}{\alpha+\beta} - \frac{\alpha}{\alpha+\beta}\,(1-\alpha-\beta)^n.\end{aligned}$$

Analoge Formeln erhält man für $P^n(1,0)$ und $P^n(1,1)$ durch Vertauschen von α und β. Für die n-Schritt-Übergangsmatrix ergibt sich also

$$P^n = \underbrace{\begin{pmatrix} \frac{\beta}{\alpha+\beta} & \frac{\alpha}{\alpha+\beta} \\ \frac{\beta}{\alpha+\beta} & \frac{\alpha}{\alpha+\beta} \end{pmatrix}}_{\text{Gleiche Zeilen}} + \underbrace{(1-\alpha-\beta)^n \begin{pmatrix} \frac{\alpha}{\alpha+\beta} & \frac{-\alpha}{\alpha+\beta} \\ \frac{-\beta}{\alpha+\beta} & \frac{\beta}{\alpha+\beta} \end{pmatrix}}_{\substack{\longrightarrow 0 \text{ exponentiell schnell,} \\ \text{falls } \alpha < 1 \text{ oder } \beta < 1}} .$$

Sind die Übergangswahrscheinlichkeiten α und β nicht beide gleich 1, dann gilt $P^n(0,\cdot) \approx P^n(1,\cdot) \approx \mu$ für große $n \in \mathbb{N}$. Die Kette „vergisst" also ihren Startwert X_0 exponentiell schnell („Exponentieller Gedächtnisverlust"), und die Verteilung von X_n nähert sich für $n \to \infty$ rasch der Gleichgewichtsverteilung μ an („Konvergenz ins Gleichgewicht") !

2.3. Unabhängigkeit

Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum. Hängen zwei Ereignisse $A, B \in \mathcal{A}$ nicht voneinander ab, dann sollte gelten:

$$\begin{aligned} \mathbb{P}[A|B] &= \mathbb{P}[A] \qquad \text{falls } \mathbb{P}[B] \neq 0, \qquad \text{sowie} \\ \mathbb{P}[B|A] &= \mathbb{P}[B] \qquad \text{falls } \mathbb{P}[A] \neq 0. \end{aligned}$$

Beide Aussagen sind äquivalent zu der Bedingung

$$\mathbb{P}[A \cap B] = \mathbb{P}[A] \cdot \mathbb{P}[B], \tag{2.8}$$

die im Fall $\mathbb{P}[A] = 0$ oder $\mathbb{P}[B] = 0$ automatisch erfüllt ist. Allgemeiner definieren wir für beliebige (endliche, abzählbare oder überabzählbare) Kollektionen von Ereignissen:

Definition 2.9. Eine Kollektion A_i, $i \in I$, von Ereignissen aus $\mathcal{A}$ heißt **unabhängig** (bzgl. $\mathbb{P}$), falls

$$\mathbb{P}[A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_n}] = \prod_{k=1}^{n} \mathbb{P}[A_{i_k}]$$

für alle $n \in \mathbb{N}$ und alle paarweise verschiedenen $i_1, \ldots, i_n \in I$ gilt.

Beispiele. a) Falls $\mathbb{P}[A] \in \{0,1\}$ gilt, dann ist A unabhängig von B für alle $B \in \mathcal{A}$. Deterministische Ereignisse sind also von allen anderen Ereignissen unabhängig.

b) Wir betrachten das kanonische Modell für zwei faire Münzwürfe, d.h. $\mathbb{P}$ ist die Gleichverteilung auf $\Omega = \{0,1\}^2$. Die drei Ereignisse

$$\begin{aligned} A_1 &= \{(1,0),(1,1)\} \quad \text{„erster Wurf Zahl"}, \\ A_2 &= \{(0,1),(1,1)\} \quad \text{„zweiter Wurf Zahl"}, \\ A_3 &= \{(0,0),(1,1)\} \quad \text{„beide Würfe gleich"}, \end{aligned}$$

sind *paarweise unabhängig*, denn es gilt:

$$\mathbb{P}[A_i \cap A_j] = \frac{1}{4} = \mathbb{P}[A_i] \cdot \mathbb{P}[A_j] \qquad \text{für alle } i \neq j.$$

Trotzdem ist die Kollektion A_1, A_2, A_3 aller drei Ereignisse *nicht unabhängig*, denn

$$\mathbb{P}[A_1 \cap A_2 \cap A_3] = \frac{1}{4} \neq \frac{1}{8} = \mathbb{P}[A_1] \cdot \mathbb{P}[A_2] \cdot \mathbb{P}[A_3].$$

Sind A und B unabhängige Ereignisse, so auch A und B^C, denn es gilt

$$\mathbb{P}[A \cap B^C] = \mathbb{P}[A] - \mathbb{P}[A \cap B] = \mathbb{P}[A] \cdot (1 - \mathbb{P}[B]) = \mathbb{P}[A] \cdot \mathbb{P}[B^C].$$

Allgemeiner folgt:

Lemma 2.10 (Stabilität von Unabhängigkeit unter Komplementbildung).
Sind die Ereignisse $A_1, \ldots, A_n \in \mathcal{A}$ unabhängig, und gilt $B_j = A_j$ oder $B_j = A_j^C$ für alle $j = 1, \ldots, n$, dann sind auch die Ereignisse $B_1, \ldots, B_n$ unabhängig.

Beweis. Da wir zum Nachweis der Unabhängigkeit beliebige Unterkollektionen von $\{B_1, \ldots B_n\}$ betrachten müssen, ist zu zeigen, dass

$$\mathbb{P}[C_1 \cap \ldots \cap C_n] \;=\; \mathbb{P}[C_1] \cdot \cdots \cdot \mathbb{P}[C_n]$$

gilt, falls die Ereignisse C_i jeweils gleich A_i, A_i^C oder Ω sind. Sei ohne Beschränkung der Allgemeinheit $C_i = A_i$ für $i \le k$, $C_i = A_i^C$ für $k < i \le l$, und $C_i = \Omega$ für $i > l$ mit $0 \le k \le l \le n$. Dann folgt unter Verwendung der Linearität des Erwartungswerts und der Unabhängigkeit von $A_1, \ldots, A_n$:

$$\begin{aligned}
\mathbb{P}[C_1 \cap \ldots \cap C_n] &= \mathbb{P}\big[A_1 \cap \ldots \cap A_k \cap A_{k+1}^C \cap \ldots \cap A_l^C\big] \\
&= \mathbb{E}\left[I_{A_1} \cdots I_{A_k} \cdot (1 - I_{A_{k+1}}) \cdots (1 - I_{A_l})\right] \\
&= \mathbb{E}\big[I_{A_1} \cdots I_{A_k} \cdot \sum_{J \subseteq \{k+1,\ldots,l\}} (-1)^{|J|} \prod_{j \in J} I_{A_j}\big] \\
&= \sum_{J \subseteq \{k+1,\ldots,l\}} (-1)^{|J|}\, \mathbb{P}\big[A_1 \cap \ldots \cap A_k \cap \bigcap_{j \in J} A_j\big] \\
&= \sum_{J \subseteq \{k+1,\ldots,l\}} (-1)^{|J|} \mathbb{P}[A_1] \cdots \mathbb{P}[A_k] \cdot \prod_{j \in J} \mathbb{P}[A_j] \\
&= \mathbb{P}[A_1] \cdots \mathbb{P}[A_k] \cdot (1 - \mathbb{P}[A_{k+1}]) \cdots (1 - \mathbb{P}[A_l]) \\
&= \mathbb{P}[C_1] \cdots \mathbb{P}[C_n].
\end{aligned}$$

■

Verteilungen für unabhängige Ereignisse

Seien $A_1, A_2, \ldots \in \mathcal{A}$ unabhängige Ereignisse (bzgl. $\mathbb{P}$) mit $\mathbb{P}[A_i] = p \in [0, 1]$. Diese beschreiben zum Beispiel unabhängige Wiederholungen eines Zufallsexperiments. Die Existenz von unendlich vielen unabhängigen Ereignissen auf einem geeigneten Wahrscheinlichkeitsraum setzen wir hier voraus – ein Beweis wird erst in der Vorlesung Einführung in die Wahrscheinlichkeitstheorie gegeben.

Geometrische Verteilung

Die *„Wartezeit“* auf das erste Eintreten eines der Ereignisse ist durch

$$T(\omega) \;=\; \inf\{n \in \mathbb{N} \;:\; \omega \in A_n\}$$

gegeben, wobei wir hier $\min \emptyset := \infty$ setzen. Mit Lemma 2.10 können wir die Verteilung der Zufallsvariable $T : \Omega \to \mathbb{N} \cup \{\infty\}$ berechnen. Für $n \in \mathbb{N}$ erhalten wir

$$\begin{aligned}
\mathbb{P}[T = n] &= \mathbb{P}[A_1^C \cap A_2^C \cap \ldots \cap A_{n-1}^C \cap A_n] \\
&= \mathbb{P}[A_n] \cdot \prod_{i=1}^{n-1} \mathbb{P}[A_i^C] \\
&= p \cdot (1 - p)^{n-1}.
\end{aligned}$$

Definition 2.11. Sei $p \in (0,1]$. Die Wahrscheinlichkeitsverteilung μ auf $\mathbb{N} \cup \{\infty\}$ mit Massenfunktion

$$\mu(n) = p \cdot (1-p)^{n-1} \qquad \text{für } n \in \mathbb{N}$$

heißt **geometrische Verteilung zum Parameter** p, und wird kurz mit $\mathrm{Geom}(p)$ bezeichnet.

Bemerkung. a) für $n \in \mathbb{N}$ gilt

$$\mathbb{P}[T > n] \;=\; \mathbb{P}\big[A_1^C \cap \ldots \cap A_n^C\big] \;=\; (1-p)^n.$$

Ist $p \neq 0$, dann folgt insbesondere $\mathbb{P}[T = \infty] = 0$, d.h. die geometrische Verteilung ist eine Wahrscheinlichkeitsverteilung auf den natürlichen Zahlen. für $p = 0$ gilt dagegen $\mathbb{P}[T = \infty] = 1$.

b) Wegen $T = \sum_{n=0}^{\infty} I_{\{T>n\}}$ ergibt sich als Erwartungswert der geometrischen Verteilung

$$\mathbb{E}[T] \;=\; \sum_{n=0}^{\infty} \mathbb{P}[T > n] \;=\; \frac{1}{1-(1-p)} \;=\; \frac{1}{p}.$$

Binomialverteilung

Die *Anzahl* der Ereignisse unter $A_1, \ldots, A_n$, die eintreten, ist durch die Zufallsvariable

$$S_n(\omega) \;=\; |\{1 \le i \le n \;:\; \omega \in A_i\}| \;=\; \sum_{i=1}^{n} I_{A_i}(\omega)$$

gegeben. Mithilfe von Lemma 2.10 können wir auch die Verteilung von S_n berechnen. Für $0 \le k \le n$ gilt

$$\begin{aligned}
\mathbb{P}[S_n = k] \;&=\; \sum_{\substack{I \subseteq \{1,\ldots,n\} \\ |I|=k}} \mathbb{P}\left[\bigcap_{i \in I} A_i \cap \bigcap_{i \in \{1,\ldots,n\}\setminus I} A_i^C\right] \;=\; \sum_{\substack{I \subseteq \{1,\ldots,n\} \\ |I|=k}} \prod_{i \in I} \mathbb{P}[A_i] \cdot \prod_{i \in I^C} \mathbb{P}[A_i^C] \\
&=\; \sum_{\substack{I \subseteq \{1,\ldots,n\} \\ |I|=k}} \prod_{i \in I} p \cdot \prod_{i \in I^C} (1-p) \;=\; \sum_{\substack{I \subseteq \{1,\ldots,n\} \\ |I|=k}} p^{|I|} \cdot (1-p)^{|I^C|} \\
&=\; \binom{n}{k} p^k \, (1-p)^{n-k},
\end{aligned}$$

d.h. S_n ist *binomialverteilt mit Parametern n und p*.

Exkurs zu gemeinsamen Verteilungen

Um den Zusammenhang zwischen mehreren Zufallsvariablen untersuchen, genügt es nicht, die Verteilungen der einzelnen Zufallsvariablen zu kennen. Stattdessen benötigen wir die *gemeinsame Verteilung* der Zufallvariablen. Diese ist folgendermaßen definiert: Sind $X_1 : \Omega \to S_1, \ldots, X_n : \Omega \to S_n$ diskrete Zufallsvariablen, dann ist auch $(X_1, \ldots, X_n)$ eine diskrete Zufallsvariable mit Werten im Produktraum $S_1 \times \cdots \times S_n$.

Definition 2.12. Die Verteilung $\mu_{X_1,\dots,X_n}$ des Zufallsvektors $(X_1, \dots, X_n)$ unter $\mathbb{P}$ heißt **gemeinsame Verteilung** der Zufallsvariablen $X_1, \dots, X_n$.

Die gemeinsame Verteilung ist eine Wahrscheinlichkeitsverteilung auf $S_1 \times \cdots \times S_n$ mit Massenfunktion

$$p_{X_1,\dots,X_n}(a_1, \dots, a_n) \;=\; \mathbb{P}[X_1 = a_1, \dots, X_n = a_n] \tag{2.9}$$

Sie enthält Informationen über den Zusammenhang zwischen den Zufallsgrößen X_i. Die Verteilungen der einzelnen Zufallsvariablen X_i nennt man dagegen **Randverteilungen**.

Beispiel (Zwei unabhängige Würfel). Beschreiben die Zufallsvariablen $X, Y : \Omega \to \{1, 2, 3, 4, 5, 6\}$ auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ die Augenzahlen beim Werfen zweier Würfel, die jede Augenzahl unabhängig voneinander mit Wahrscheinlichkeit $1/6$ annehmen, dann gilt

$$\mathbb{P}[X = a, Y = b] \;=\; \frac{1}{36} \qquad \text{für alle } a, b \in \{1, 2, 3, 4, 5, 6\}.$$

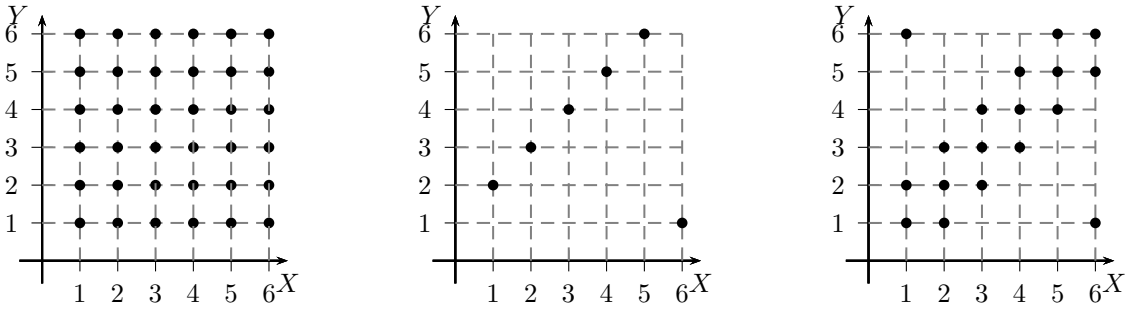
Die gemeinsame Verteilung von X und Y ist also die Gleichverteilung auf $\{1, 2, 3, 4, 5, 6\}^2$. Sei nun $M = \max(X, Y)$ die größere der beiden Augenzahlen. Dann erhalten wir für die gemeinsame Verteilung von M und X die folgenden Wahrscheinlichkeiten:

$\mathbb{P}[M = i, X = j]$	$j = 1$	2	3	4	5	6	$\mathbb{P}[M = i]$
$i = 1$	1/36	0	0	0	0	0	1/36
2	1/36	2/36	0	0	0	0	3/36
3	1/36	1/36	3/36	0	0	0	5/36
4	1/36	1/36	1/36	4/36	0	0	7/36
5	1/36	1/36	1/36	1/36	5/36	0	9/36
6	1/36	1/36	1/36	1/36	1/36	6/36	11/36
$\mathbb{P}[X = j]$	1/6	1/6	1/6	1/6	1/6	1/6	1

Die Massenfunktionen der Randverteilungen von X und M stehen in der letzten Zeile bzw. Spalte und ergeben sich durch Aufaddieren über die möglichen Werte der jeweils anderen Zufallsvariable.

Das folgende Beispiel zeigt, dass die gemeinsamen Verteilungen auch dann sehr unterschiedlich sein können wenn die Randverteilungen übereinstimmen.

Beispiel (Zwei abhängige Würfel). Seien $X, Y : \Omega \to \{1, 2, 3, 4, 5, 6\}$ gleichverteilte Zufallsvariablen. Für die Gewichte der gemeinsamen Verteilung von X und Y gibt es dann unter anderem die in Abbildung 2.2 gegebenen Möglichkeiten.



(a) X, Y unabhängig.
Die Gewichte der Punkte sind jeweils gleich $1/36$.
$\mu_{X,Y} = \mu_X \otimes \mu_Y$.

(b) $Y = (X + 1) \mod 6$.
Die Gewichte der Punkte sind jeweils gleich $1/6$.

(c) $Y = (X + Z) \mod 6$,
$Z \sim \mathrm{Unif}\{-1, 0, 1\}$.
Die Gewichte der Punkte sind jeweils gleich $1/18$.

Abbildung 2.2.: Gemeinsame Verteilungen mit identischen Randverteilungen.

Die gemeinsame Verteilung der Zufallsvariablen $X_1, \ldots, X_n$ benötigen wir, wenn wir Erwartungswerte von Funktionen berechnen wollen, die von mehreren dieser Zufallsvariablen abhängen. Ist $g : S_1 \times \cdots \times S_n \to [0, \infty)$ eine reellwertige Funktion, dann gilt nach dem Transformationssatz 1.15 nämlich

$$\mathbb{E}\left[g(X_1, \ldots, X_n)\right] = \sum_{(a_1, \ldots, a_n)} g(a_1, \ldots, a_n)\, \mathbb{P}\left[X_1 = a_1, \ldots, X_n = a_n\right].$$

Beispiel (Erwartungswerte für zwei Würfel). Sind X und Y die Augenzahlen zweier unabhängiger fairer Würfel, dann gilt

$$\mathbb{E}\left[X \cdot Y\right] = \frac{1}{36} \sum_{i,j=1}^{6} ij = \left(\frac{1}{6} \sum_{i=1}^{6} i\right)^2 = \left(\frac{7}{2}\right)^2 = \frac{49}{4},$$

$$\mathbb{E}\left[\max(X, Y)\right] = \frac{1}{36} \sum_{i,j=1}^{6} \max(i, j) = \frac{1}{36} \sum_{i=1}^{6} i + \frac{2}{36} \sum_{i=1}^{6} \sum_{j=1}^{i-1} i = \frac{161}{36}.$$

Unabhängigkeit von diskreten Zufallsvariablen

Wir erweitern den Begriff der Unabhängigkeit nun von Ereignissen auf Zufallsvariablen. Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum, und I eine beliebige Menge.

Definition 2.13. Eine Familie $X_i : \Omega \to S_i$ $(i \in I)$ von Zufallsvariablen auf $(\Omega, \mathcal{A}, \mathbb{P})$ mit abzählbaren Wertebereichen S_i heißt **unabhängig**, falls die Ereignisse $\{X_i \in A_i\}$ $(i \in I)$ für alle Teilmengen $A_i \subseteq S_i$ unabhängig sind.

Aus der Definition folgt unmittelbar, dass die Zufallsvariablen X_i $(i \in I)$ genau dann unabhängig sind, wenn jede endliche Teilkollektion unabhängig ist. Daher beschränken wir uns im folgenden auf den Fall $I = \{1, \ldots, n\}$ mit $n \in \mathbb{N}$.

Satz 2.14. Die folgenden Aussagen sind äquivalent:

(i) $X_1, \ldots, X_n$ sind unabhängig.

(ii) Die Ereignisse $\{X_1 = a_1\}, \ldots, \{X_n = a_n\}$ sind unabhängig für alle $a_i \in S_i$, $i = 1, \ldots, n$.

(iii) $p_{X_1,\ldots,X_n}(a_1, \ldots, a_n) = \prod_{i=1}^n p_{X_i}(a_i)$ für alle $a_i \in S_i$, $i = 1, \ldots, n$.

(iv) $\mu_{X_1,\ldots,X_n} = \bigotimes_{i=1}^n \mu_{X_i}$.

Beweis. (i)⇒(ii) folgt durch Wahl von $A_i = \{a_i\}$.

(iii)⇔(iv) gilt nach Definition des Produkts $\bigotimes_{i=1}^n \mu_{X_i}$ der Wahrscheinlichkeitsverteilungen μ_{X_i}.

(iv)⇒(i): Seien $A_i \subseteq S_i$ für $(i = 1, \ldots, n)$ und $1 \le i_1 < i_2 < \ldots < i_k \le n$. Um die Produkteigenschaft für die Ereignisse mit Indizes $i_1, \ldots, i_k$ zu zeigen, setzen wir $B_{i_j} := A_{i_j}$ für alle j und $B_i := S_i$ für $i \notin \{i_1, \ldots, i_k\}$. Mit (iv) folgt dann nach Satz 2.6:

$$\begin{aligned}
&\mathbb{P}[X_{i_1} \in A_{i_1}, \ldots, X_{i_k} \in A_{i_k}] = \mathbb{P}[X_1 \in B_1, \ldots, X_n \in B_n] \\
&\quad = \mathbb{P}[(X_1, \ldots, X_n) \in B_1 \times \ldots \times B_n] = \mu_{X_1,\ldots,X_n}[B_1 \times \ldots \times B_n] \\
&\quad = \prod_{i=1}^n \mu_{X_i}[B_i] = \prod_{i=1}^n \mathbb{P}[X_i \in B_i] = \prod_{j=1}^k \mathbb{P}[X_{i_j} \in A_{i_j}].
\end{aligned}$$

■

Als Konsequenz aus Satz 2.14 ergibt sich insbesondere:

Korollar 2.15. Sind $X_i : \Omega \to S_i$ $(i = 1, \ldots, n)$ diskrete Zufallsvariablen, und hat die Massenfunktion der gemeinsamen Verteilung eine Darstellung in Produktform

$$p_{X_1,\ldots,X_n}(a_1, \ldots, a_n) = c \cdot \prod_{i=1}^n g_i(a_i) \qquad \forall\, (a_1, \ldots, a_n) \in S_1 \times \ldots \times S_n \tag{2.10}$$

mit Funktionen $g_i : S_i \to [0, \infty)$ und einer Proportionalitätskonstanten $c \in \mathbb{R}$, dann sind $X_1, \ldots, X_n$ unabhängige Zufallsvariablen mit Massenfunktionen

$$p_{X_i}(a) = \frac{g_i(a)}{\sum\limits_{b \in S_i} g_i(b)}, \qquad a \in S_i. \tag{2.11}$$

Beweis. Durch Summieren über $a_1, a_2, \ldots, a_n$ in (2.10) folgt $\sum_{b \in S_i} g_i(b) < \infty$ für alle i. Daher ist die Funktion $\widetilde{g}_i(a) := g_i(a) / \sum_{b \in S_i} g_i(b)$ $(a \in S_i)$, die auf der rechten Seite von (2.11) steht, die Massenfunktion einer Wahrscheinlichkeitsverteilung μ_i auf S_i. Nach Voraussetzung gilt für $(a_1, \ldots, a_n) \in S_1 \times \ldots \times S_n$:

$$p_{X_1,\ldots,X_n}(a_1, \ldots, a_n) = \widetilde{c} \cdot \prod_{i=1}^n \widetilde{g}_i(a_i) \tag{2.12}$$

mit einer reellen Konstante $\widetilde{c}$. Da auf beiden Seiten von (2.12) bis auf den Faktor $\widetilde{c}$ die Massenfunktionen von Wahrscheinlichkeitsverteilungen stehen, gilt $\widetilde{c} = 1$. Also ist die gemeinsame Verteilung von $X_1, \ldots, X_n$ das Produkt der Verteilungen μ_i, und somit sind die Zufallsvariablen X_i unabhängig mit Verteilung μ_i, d.h. mit Massenfunktion $\widetilde{g}_i$. ■

Sei nun I eine beliebige Menge, und S_i sowie $\widetilde{S}_i$ $(i \in I)$ abzählbare Mengen. Sind X_i $(i \in I)$ unabhängige diskrete Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ mit Wertebereichen S_i, dann sind auch die Zufallsvariablen

$$Y_i(\omega) \;:=\; h_i(X_i(\omega)) \;=\; (h_i \circ X_i)(\omega) \qquad (i \in I)$$

für beliebige Funktionen $h_i : S_i \to \widetilde{S}_i$ wieder unabhängig. Dies folgt unmittelbar aus der Definition der Unabhängigkeit, denn für beliebige Teilmengen $A_i \subset \widetilde{S}_i$ gilt $\{Y_i \in A_i\} = \left\{X_i \in h_i^{-1}(A_i)\right\}$, und diese Ereignisse sind unabhängig. Das folgende Korollar liefert eine wichtige Verschärfung dieser Aussage.

Korollar 2.16 (Gruppierungslemma). Sei X_i $(i \in I)$ eine Kollektion unabhängiger Zufallsvariablen, und seien $I_1, I_2, \ldots, I_n$ disjunkte Teilmengen von I. Dann sind auch die Zufallsvariablen

$$X_{I_1} = (X_i)_{i \in I_1},\; X_{I_2} = (X_i)_{i \in I_2},\; \ldots,\; X_{I_n} = (X_i)_{i \in I_n}$$

wieder unabhängig. Zudem sind $h_1(X_{I_1}), \ldots, h_n(X_{I_n})$ für beliebige Funktionen $h_1, \ldots, h_n$ wieder unabhängige Zufallsvariablen.

Wir können also die unabhängigen Zufallsvariablen in disjunkte Blöcke einteilen, und beliebige Funktionen betrachten, die jeweils nur von den Zufallsvariablen in einem Block abhängen. Diese Funktionen sind dann wieder unabhängige Zufallsvariablen. Beispielsweise sind bei sechs unabhängigen Würfelwürfen $X_1, X_2, \ldots, X_6$ die Augensummen $X_1 + X_2$, $X_3 + X_4$ und $X_5 + X_6$ voneinander unabhängig, ebenso die maximalen Augenzahlen $\max(X_1, X_2, X_3)$ und $\max(X_4, X_5, X_6)$ bei den ersten und den letzten drei Würfen.

Beweis. Seien $a_i \in S_i$ $(i \in I)$, und sei $a_{I_k} = (a_i)_{i \in I_k}$ $(k = 1, 2, \ldots, n)$. Dann gilt

$$\begin{aligned}
\mathbb{P}\left[X_{I_1} = a_{I_1}, \ldots, X_{I_n} = a_{I_n}\right] &= \mathbb{P}\left[X_i = a_i \text{ für alle } i \in I_1 \cup \cdots \cup I_n\right] \\
&= \mathbb{P}\left[\bigcap_{i \in I_1 \cup \cdots \cup I_n} \{X_i = a_i\}\right] = \prod_{i \in I_1 \cup \cdots \cup I_n} \mathbb{P}\left[X_i = a_i\right] \\
&= \prod_{i \in I_1} \mathbb{P}\left[X_i = a_i\right] \cdot \cdots \cdot \prod_{i \in I_n} \mathbb{P}\left[X_i = a_i\right] \\
&= \mathbb{P}\left[X_{I_1} = a_{I_1}\right] \cdot \cdots \cdot \mathbb{P}\left[X_{I_n} = a_{I_n}\right].
\end{aligned}$$

Also ist die Massenfunktion der gemeinsamen Verteilung der Zufallsvariablen $X_{I_1}, \ldots, X_{I_n}$ das Produkt der einzelnen Massenfunktionen, d.h. die Zufallsvariablen sind unabhängig. Die Unabhängigkeit der Zufallsvariablen $h_1(X_{I_1}), \ldots, h_n(X_{I_n})$ folgt dann wie oben. ■

2.4. Summen von unabhängigen Zufallsvariablen

Wir berechnen nun Verteilungen und gemeinsame Verteilungen von Summen unabhängiger Zufallsvariablen.

Faltung von Wahrscheinlichkeitsverteilungen

Seien zunächst X und Y unabhängige diskrete Zufallsvariablen auf $(\Omega, \mathcal{A}, \mathbb{P})$ mit Werten im $\mathbb{R}^d$, Verteilungen μ bzw. ν und Massenfunktionen p bzw. q. Wir wollen die Verteilung von $X+Y$ bestimmen. Es gilt

$$\mathbb{P}[X+Y=z] \;=\; \sum_{x\in X(\Omega)} \underbrace{\mathbb{P}[X=x, Y=z-x]}_{=\mathbb{P}[X=x]\cdot\mathbb{P}[Y=z-x]} = \sum_{x\in X(\Omega)} p(x)q(z-x). \tag{2.13}$$

Die Verteilung von $X+Y$ ist also die Wahrscheinlichkeitsverteilung $\mu \star \nu$ mit Massenfunktion

$$(p\star q)(z) \;=\; \sum_{x\in X(\Omega)} p(x)q(z-x). \tag{2.14}$$

Diese Verteilung nennt man die **Faltung** der Wahrscheinlichkeitsverteilungen μ und ν.

Bemerkung (Eigenschaften der Faltung von Wahrscheinlichkeitsverteilungen). Die Faltung $\mu\star\nu$ zweier Wahrscheinlichkeitsverteilungen μ und ν auf $\mathbb{R}^d$ ist wieder eine Wahrscheinlichkeitsverteilung auf $\mathbb{R}^d$. Da die Addition von Zufallsvariablen kommutativ und assoziativ ist, hat die Faltung von Wahrscheinlichkeitsverteilungen dieselben Eigenschaften:

$$\begin{aligned} \mu\star\nu &= \nu\star\mu && (\text{da } X+Y=Y+X),\\ (\mu\star\nu)\star\eta &= \mu\star(\nu\star\eta) && (\text{da } (X+Y)+Z=X+(Y+Z)\,). \end{aligned}$$

Beispiele. (i) Sind X und Y unabhängig und $\mathrm{Bin}(n,p)$ bzw. $\mathrm{Bin}(m,p)$-verteilt, dann ist $X+Y$ eine $\mathrm{Bin}(n+m,p)$-verteilte Zufallsvariable. Zum Beweis bemerkt man, dass die gemeinsame Verteilung von X und Y mit der gemeinsamen Verteilung von $Z_1+\cdots+Z_n$ und $Z_{n+1}+\cdots+Z_{n+m}$ übereinstimmt, wobei die Zufallsvariablen Z_i $(1\le i\le n+m)$ unabhängig und Bernoulli(p)-verteilt sind. Also folgt:

$$\mu_{X+Y} \;=\; \mu_{Z_1+\cdots+Z_n+Z_{n+1}+\cdots+Z_{n+m}} \;=\; \mathrm{Bin}(n+m,p)\;.$$

Als Konsequenz erhalten wir (ohne zu rechnen):

$$\mathrm{Bin}(n,p)\star\mathrm{Bin}(m,p) \;=\; \mathrm{Bin}(n+m,p)\;,$$

d.h. die Binomialverteilungen bilden eine *Faltungshalbgruppe*. Explizit ergibt sich:

$$\sum_{k=0}^{l}\binom{n}{k}p^k(1-p)^{n-k}\binom{m}{l-k}p^{l-k}(1-p)^{m-(l-k)} \;=\; \binom{n+m}{l}p^l(1-p)^{n+m-l}\;,\quad \text{d.h.}$$

$$\sum_{k=0}^{l}\binom{n}{k}\binom{m}{l-k} \;=\; \binom{n+m}{l}. \tag{2.15}$$

Diese kombinatorische Formel ist auch als *Vandermonde-Identität* bekannt.

(ii) Sind X und Y unabhängig und Poisson-verteilt mit Parametern λ bzw. $\widetilde{\lambda}$, dann ist $X+Y$ Poisson-verteilt mit Parameter $\lambda+\widetilde{\lambda}$, denn nach der binomischen Formel gilt für $n\ge 0$:

$$\begin{aligned} (\mu_X\star\mu_Y)(n) &= \sum_{k=0}^{n}\mu_X(k)\cdot\mu_Y(n-k)\\ &= \sum_{k=0}^{n}\frac{\lambda^k}{k!}\,e^{-\lambda}\cdot\frac{\widetilde{\lambda}^{n-k}}{(n-k)!}\,e^{-\widetilde{\lambda}}\\ &= e^{-(\lambda+\widetilde{\lambda})}\cdot\sum_{k=0}^{n}\frac{\lambda^k}{k!}\frac{\widetilde{\lambda}^{n-k}}{(n-k)!}\\ &= e^{-(\lambda+\widetilde{\lambda})}\cdot\frac{(\lambda+\widetilde{\lambda})^n}{n!}\;. \end{aligned}$$

Also bilden auch die Poissonverteilungen eine Faltungshalbgruppe:

$$\text{Poisson}(\lambda) \star \text{Poisson}(\widetilde{\lambda}) = \text{Poisson}(\lambda + \widetilde{\lambda}).$$

Irrfahrten auf $\mathbb{Z}$

Seien $X_1, X_2, \ldots$ unabhängige und identisch verteilte („i.i.d." – independent and identically distributed) Zufallsvariablen auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ mit

$$\mathbb{P}[X_i = +1] = p, \qquad \mathbb{P}[X_i = -1] = 1 - p, \qquad p \in (0, 1).$$

Die Existenz von unendlich vielen unabhängigen identisch verteilten Zufallsvariablen auf einem geeigneten Wahrscheinlichkeitsraum (unendliches Produktmodell) wird in der Vorlesung Einführung in die Wahrscheinlichkeitstheorie gezeigt. Sei $a \in \mathbb{Z}$ ein fester Startwert. Wir betrachten die durch

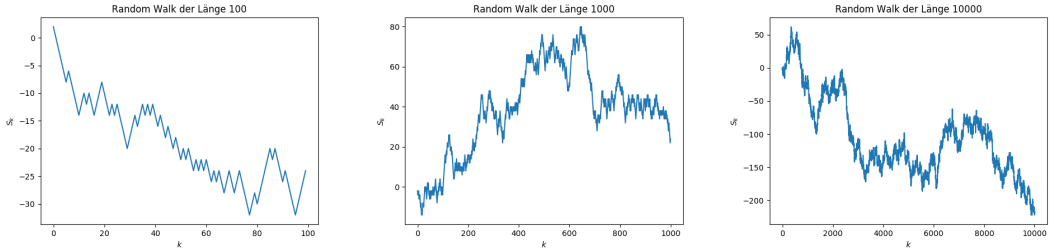
$$\begin{aligned} S_0 &= a, \\ S_{n+1} &= S_n + X_{n+1}, \end{aligned}$$

definierte zufällige Bewegung („Irrfahrt" oder „Random Walk") auf $\mathbb{Z}$. Als Position zur Zeit n ergibt sich

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

Irrfahrten werden unter anderem in vereinfachten Modellen für die Kapitalentwicklung beim Glücksspiel oder an der Börse (Aktienkurs), sowie die Brownsche Molekularbewegung (im Skalierungslimes Schrittweite $\to 0$) eingesetzt.

Beispiel (Symmetrische Irrfahrt, $p = 1/2$). Die folgenden Graphiken zeigen Simulationen der ersten 50, 500 bzw. 5000 Schritte eines Random Walks für $p = 1/2$.



(a) $n = 50$ (b) $n = 500$ (c) $n = 5000$

Wir wollen nun die Verteilung von verschiedenen, durch die Irrfahrt gegebenen, Zufallsvariablen berechnen. Die Verteilung von S_n selbst ist eine verzerrte Bimomialverteilung.

Lemma 2.17 (Verteilung von S_n). *Für $k \in \mathbb{Z}$ gilt*

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

Beweis. Es gilt

$$S_n = a + k \;\Leftrightarrow\; X_1 + \cdots + X_n = k \;\Leftrightarrow\; \begin{cases} X_i = 1 & \text{genau } \frac{n+k}{2} \text{ mal,} \\ X_i = -1 & \text{genau } \frac{n-k}{2} \text{ mal.} \end{cases}$$

■

Sei $\lambda \in \mathbb{Z}$. Weiter unten werden wir (im Fall $p = 1/2$) die Verteilung der Zufallsvariable

$$T_\lambda(\omega) \;:=\; \min\{n \in \mathbb{N} \;:\; S_n(\omega) = \lambda\}$$

bestimmen, wobei wir wieder $\min \emptyset := \infty$ setzen. Für $\lambda \neq a$ ist T_λ die erste *Trefferzeit von* λ, für $\lambda = a$ ist es hingegen die erste *Rückkehrzeit nach* a. Beschreibt die Irrfahrt beispielsweise die Kapitalentwicklung in einem Glücksspiel, dann kann man T_0 als Ruinzeitpunkt interpretieren. Da das Ereignis

$$\{T_\lambda \leq n\} \;=\; \bigcup_{i=1}^{n} \{S_i = \lambda\}$$

von den Positionen der Irrfahrt zu *mehreren* Zeiten abhängt, benötigen wir die *gemeinsame* Verteilung der entsprechenden Zufallsvariablen. Sei dazu

$$S_{0:n}(\omega) \;:=\; (S_0(\omega), S_1(\omega), \ldots, S_n(\omega))$$

der *Bewegungsverlauf bis zur Zeit* n . Dann ist $S_{0:n}$ eine Zufallsvariable, die Werte im Raum

$$\widehat{\Omega}_a^{(n)} \;:=\; \{(s_0, s_1, \ldots, s_n) \;:\; s_0 = a, s_i \in \mathbb{Z} \text{ mit } |s_i - s_{i-1}| = 1 \text{ für alle } i \in \{1, \ldots, n\}\}$$

aller möglichen Verläufe (Pfade) der Irrfahrt annimmt. Sei μ_a die Verteilung von $S_{0:n}$ unter $\mathbb{P}$.

Lemma 2.18. *Für* $(s_0, s_1, \ldots, s_n) \in \widehat{\Omega}_a^{(n)}$ *gilt*

$$\mu_a[\{(s_0, \ldots, s_n)\}] \;=\; p^{\frac{n+k}{2}} (1-p)^{\frac{n-k}{2}}, \qquad \textit{wobei } \; k = s_n - s_0. \tag{2.16}$$

Insbesondere ist μ_a *im Fall* $p = 1/2$ *die Gleichverteilung auf dem Pfadraum* $\widehat{\Omega}_a^{(n)} \subseteq \mathbb{Z}^{n+1}$.

Beweis. Für $s_0, \ldots, s_n \in \mathbb{Z}$ gilt

$$\begin{aligned} \mu_a[\{(s_0, \ldots, s_n)\}] &= \mathbb{P}[S_0 = s_0, \ldots, S_n = s_n] \\ &= \mathbb{P}[S_0 = s_0, X_1 = s_1 - s_0, \ldots, X_n = s_n - s_{n-1}]. \end{aligned}$$

Diese Wahrscheinlichkeit ist gleich 0, falls $s_0 \neq a$ oder $|s_i - s_{i-1}| \neq 1$ für ein $i \in \{1, \ldots, n\}$ gilt. Andernfalls, d.h. für $(s_0, \ldots, s_n) \in \widehat{\Omega}_a^{(n)}$, gilt (2.16), da für $s_n - s_0 = k$ genau $\frac{n+k}{2}$ der Inkremente $s_1 - s_0, \ldots, s_n - s_{n-1}$ gleich $+1$ und die übrigen gleich -1 sind. ■

Symmetrische Irrfahrt und Reflektionsprinzip

Ab jetzt betrachten wir nur noch die symmetrische Irrfahrt mit $p = \frac{1}{2}$. Lemma 2.18 ermöglicht es uns, Wahrscheinlichkeiten für die symmetrische Irrfahrt durch Abzählen zu berechnen. Dazu zeigen wir eine nützliche Invarianzeigenschaft bezüglich der Reflektion der Pfade beim ersten Erreichen eines Levels λ. Den Beweis des folgenden Satzes macht man sich am besten zunächst anhand von Abbildung 2.4 klar.

Satz 2.19 (Reflektionsprinzip). Seien $\lambda, b \in \mathbb{Z}$. Es gelte entweder ($a < \lambda$ und $b \leq \lambda$), oder ($a > \lambda$ und $b \geq \lambda$). Dann folgt

$$\mathbb{P}[T_\lambda \leq n, S_n = b] \;=\; \mathbb{P}[S_n = b^\star],$$

wobei $b^\star := \lambda + (\lambda - b) = 2\lambda - b$ die Spiegelung von b an λ ist.

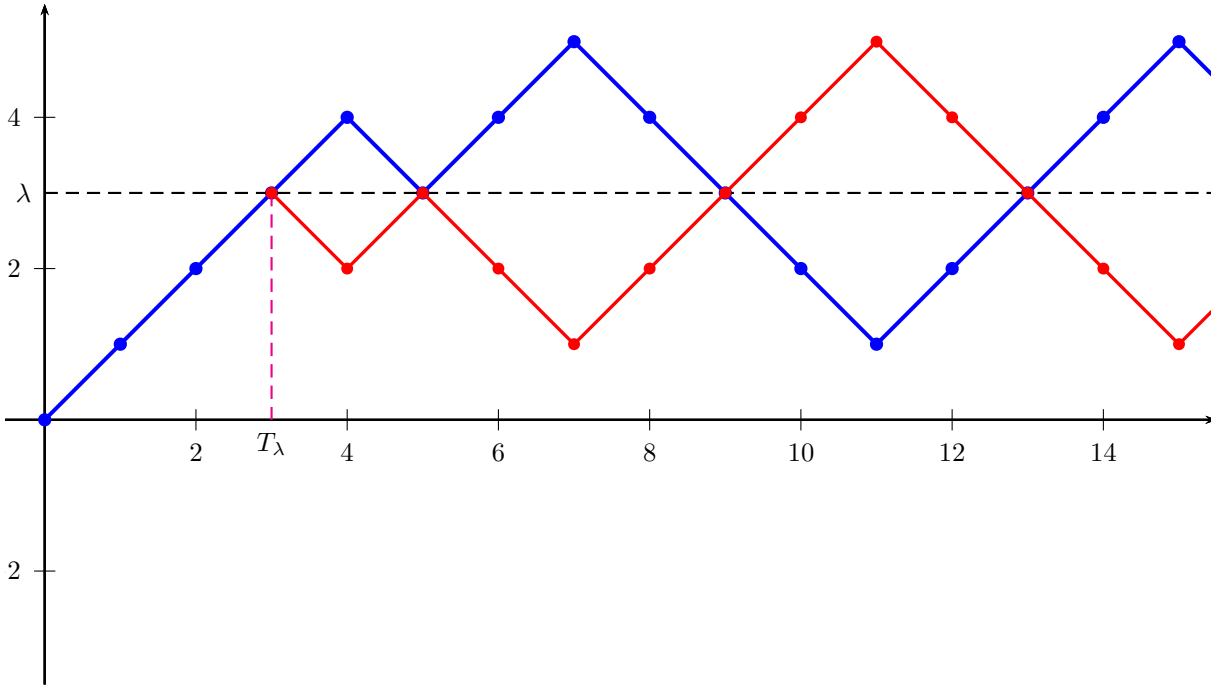


Abbildung 2.4.: Reflektionsprinzip

Beweis. Es gilt

$$\begin{aligned}\mathbb{P}[T_\lambda \le n, S_n = b] &= \mu_a[\overbrace{\{(s_0,\dots,s_n) \,:\, s_n = b, s_i = \lambda \text{ für ein } i \in \{1,\dots,n\}\}}^{=:A}],\\ \mathbb{P}[S_n = b^\star] &= \mu_a[\underbrace{\{(s_0,\dots,s_n) \,:\, s_n = b^\star\}}_{=:B}].\end{aligned}$$

Die in Abbildung 2.4 dargestellte Transformation (Reflektion des Pfades nach Treffen von λ) definiert eine Bijektion von A nach B. Also gilt $|A| = |B|$. Da μ_a die Gleichverteilung auf $\widehat{\Omega}_a^{(n)}$ ist, folgt

$$\mu_a[A] = \frac{|A|}{\left|\widehat{\Omega}_a^{(n)}\right|} = \frac{|B|}{\left|\widehat{\Omega}_a^{(n)}\right|} = \mu_a[B],$$

und damit die Behauptung. ■

Mithilfe des Reflektionsprinzips können wir nun die Verteilung der ersten Trefferzeiten explizit aus den uns schon bekannten Verteilungen der Zufallsvariablen S_n berechnen.

Korollar 2.20 (Verteilung der Trefferzeiten). Für $\lambda \in \mathbb{Z}$ und $n \in \mathbb{N}$ gilt:

(i)

$$\mathbb{P}[T_\lambda \le n] = \begin{cases}\mathbb{P}[S_n \ge \lambda] + \mathbb{P}[S_n > \lambda] & \text{falls } \lambda > a,\\ \mathbb{P}[S_n \le \lambda] + \mathbb{P}[S_n < \lambda] & \text{falls } \lambda < a.\end{cases}$$

(ii)

$$\mathbb{P}[T_\lambda = n] \;=\; \begin{cases} \frac{1}{2}\mathbb{P}[S_{n-1} = \lambda - 1] - \frac{1}{2}\mathbb{P}[S_{n-1} = \lambda + 1] & \text{falls } \lambda > a, \\ \frac{1}{2}\mathbb{P}[S_{n-1} = \lambda + 1] - \frac{1}{2}\mathbb{P}[S_{n-1} = \lambda - 1] & \text{falls } \lambda < a. \end{cases}$$

Beweis. Wir beweisen die Aussagen für $\lambda > a$, der andere Fall wird jeweils analog gezeigt.

(i) Ist $S_n \geq \lambda$, dann gilt stets $T_\lambda \leq n$. Daher folgt nach Satz 2.19:

$$\begin{aligned} \mathbb{P}[T_\lambda \leq n] &= \sum_{b \in \mathbb{Z}} \underbrace{\mathbb{P}[T_\lambda \leq n, S_n = b]}_{\substack{= \mathbb{P}[S_n = b] \text{ für } b \geq \lambda, \\ = \mathbb{P}[S_n = b^\star] \text{ für } b < \lambda.}} = \sum_{b \geq \lambda} \mathbb{P}[S_n = b] + \underbrace{\sum_{b < \lambda} \mathbb{P}[S_n = b^\star]}_{= \sum_{b > \lambda} \mathbb{P}[S_n = b]} \\ &= \mathbb{P}[S_n \geq \lambda] + \mathbb{P}[S_n > \lambda]. \end{aligned}$$

(ii) Aus (i) folgt

$$\begin{aligned} \mathbb{P}[T_\lambda = n] &= \mathbb{P}[T_\lambda \leq n] - \mathbb{P}[T_\lambda \leq n-1] \\ &= \underbrace{\mathbb{P}[S_n \geq \lambda] - \mathbb{P}[S_{n-1} \geq \lambda]}_{=:\mathbf{I}} + \underbrace{\mathbb{P}[S_n \geq \lambda + 1] - \mathbb{P}[S_{n-1} \geq \lambda + 1]}_{=:\mathbf{II}} \end{aligned}$$

Wegen

$$\mathbb{P}[A] - \mathbb{P}[B] = \mathbb{P}[A \backslash B] + \mathbb{P}[A \cap B] - \mathbb{P}[B \backslash A] - \mathbb{P}[B \cap A] = \mathbb{P}[A \backslash B] - \mathbb{P}[B \backslash A]$$

erhalten wir für den ersten Term:

$$\begin{aligned} \mathbf{I} &= \mathbb{P}[S_n \geq \lambda, S_{n-1} < \lambda] - \mathbb{P}[S_{n-1} \geq \lambda, S_n < \lambda] \\ &= \mathbb{P}[S_{n-1} = \lambda - 1, S_n = \lambda] - \mathbb{P}[S_{n-1} = \lambda, S_n = \lambda - 1] \\ &= \frac{1}{2}\mathbb{P}[S_{n-1} = \lambda - 1] - \frac{1}{2}\mathbb{P}[S_{n-1} = \lambda]. \end{aligned}$$

Mit einer analogen Berechnung für den zweiten Term erhalten wir insgesamt:

$$\begin{aligned} \mathbb{P}[T_\lambda = n] &= \mathbf{I} + \mathbf{II} \\ &= \frac{1}{2}\,(\mathbb{P}[S_{n-1} = \lambda - 1] - \mathbb{P}[S_{n-1} = \lambda] \\ &\quad + \mathbb{P}[S_{n-1} = (\lambda + 1) - 1] - \mathbb{P}[S_{n-1} = \lambda + 1]) \\ &= \frac{1}{2}\,(\mathbb{P}[S_{n-1} = \lambda - 1] - \mathbb{P}[S_{n-1} = \lambda + 1]). \end{aligned}$$

■

Aus der Verteilung der Trefferzeiten T_λ ergibt sich auch unmittelbar die Verteilung des Maximums

$$M_n \;:=\; \max(S_0, S_1, \ldots, S_n)$$

der Irrfahrt bis zur Zeit n.

Korollar 2.21 (Verteilung des Maximums). Für $\lambda > a$ gilt

$$\mathbb{P}[M_n \geq \lambda] \;=\; \mathbb{P}[T_\lambda \leq n] \;=\; \mathbb{P}[S_n \geq \lambda] + \mathbb{P}[S_n > \lambda].$$

3. Gesetze der großen Zahlen

In diesem Kapitel beweisen wir zwei ganz unterschiedliche Arten von Konvergenzaussagen für Folgen von Zufallsvariablen bzw. deren Verteilungen: zum einen Gesetze der großen Zahlen für relative Häufigkeiten von unabhängigen Ereignissen, und allgemeiner für Mittelwerte von schwach korrelierten Zufallsvariablen, zum anderen die Konvergenz ins Gleichgewicht der Verteilungen irreduzibler, aperiodischer Markovketten mit endlichem Zustandsraum. Beide Aussagen lassen sich auch zu einem Gesetz der großen Zahlen für Markovketten kombinieren.

3.1. Gesetz der großen Zahlen für unabhängige Ereignisse

Das empirische Gesetz der großen Zahlen (GGZ) besagt, dass sich die relative Häufigkeit für das Eintreten von gleich wahrscheinlichen unabhängigen Ereignissen $A_1, \dots, A_n$ für $n \to \infty$ der Erfolgswahrscheinlichkeit p annähert. Wir können diese Aussage nun mathematisch präzisieren, und aus den Kolmogorovschen Axiomen herleiten. Je nach Präzisierung des Konvergenzbegriffs unterscheidet man zwischen dem schwachen und dem starken Gesetz der großen Zahlen.

Bernstein-Ungleichung und schwaches Gesetz der großen Zahlen

Sei $A_1, A_2, \dots$ eine Folge unabhängiger Ereignisse auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ mit fester Wahrscheinlichkeit $\mathbb{P}[A_i] = p \in [0,1]$, und sei

$$S_n(\omega) \;=\; |\{1 \le i \le n \;:\; \omega \in A_i\}| \;=\; \sum_{i=1}^{n} I_{A_i}(\omega)$$

die Anzahl der Ereignisse unter $A_1, \dots, A_n$, die eintreten.

Satz 3.1 (Bernstein-Ungleichung, Schwaches GGZ für unabhängige Ereignisse).
für alle $\varepsilon > 0$ und $n \in \mathbb{N}$ gilt

$$\mathbb{P}\left[\frac{S_n}{n} \ge p + \varepsilon\right] \;\le\; e^{-2\varepsilon^2 n}, \qquad \text{und} \qquad \mathbb{P}\left[\frac{S_n}{n} \le p - \varepsilon\right] \;\le\; e^{-2\varepsilon^2 n}.$$

Insbesondere ist

$$\mathbb{P}\left[\left|\frac{S_n}{n} - p\right| \ge \varepsilon\right] \;\le\; 2\,e^{-2\varepsilon^2 n},$$

d.h. die Wahrscheinlichkeit für eine Abweichung der relativen Häufigkeit S_n/n von der Wahrscheinlichkeit p um mehr als ε fällt exponentiell schnell in n ab.

Bemerkung. a) Der Satz liefert eine nachträgliche Rechtfertigung der frequentistischen Interpretation der Wahrscheinlichkeit als asymptotische relative Häufigkeit.

b) Die Aussage kann man zum empirischen *Schätzen der Wahrscheinlichkeit p* verwenden: für große n gilt

$$p \approx \frac{S_n}{n} = \text{relative Häufigkeit des Ereignisses bei } n \text{ unabhängigen Stichproben.}$$

Simuliert man die Stichproben künstlich auf dem Computer, dann ergibt sich ein *Monte-Carlo-Verfahren* zur näherungsweisen Berechnung von p. Der Satz liefert eine recht präzise Fehlerabschätzung für den Schätz- bzw. Approximationsfehler.

c) Bemerkenswert ist, dass die Abschätzung aus der Bernstein-Ungleichung nicht nur asymptotisch für $n \to \infty$, sondern für jedes feste n gilt. Solche präzisen *nicht-asymptotischen Abschätzungen* sind für Anwendungen sehr wichtig, und oft nicht einfach herzuleiten.

Beweis. Der Beweis von Satz 3.1 besteht aus zwei Teilen: Wir leiten zunächst exponentielle Abschätzungen für die Wahrscheinlichkeiten her, welche von einem Parameter $\lambda \geq 0$ abhängen. Anschließend optimieren wir die erhaltene Abschätzung durch Wahl von λ.

Wir setzen $q := 1 - p$. Wegen $S_n \sim \mathrm{Bin}(n, p)$ gilt für $\lambda \geq 0$:

$$\begin{aligned}
\mathbb{P}[S_n \geq n\,(p+\varepsilon)] &= \sum_{k \geq np+n\varepsilon} \binom{n}{k} p^k\, q^{n-k} \\
&\leq \sum_{k \geq np+n\varepsilon} \binom{n}{k} e^{\lambda k}\, p^k\, q^{n-k}\, e^{-\lambda(np+n\varepsilon)} \\
&\leq \sum_{k=0}^{n} \binom{n}{k} \left(p\, e^{\lambda}\right)^k\, q^{n-k}\, e^{-\lambda np}\, e^{-\lambda n\varepsilon} \\
&= \left(p\, e^{\lambda} + q\right)^n\, e^{-\lambda np}\, e^{-\lambda n\varepsilon} \\
&= \left(p\, e^{\lambda q} + q\, e^{-\lambda p}\right)^n\, e^{-\lambda n\varepsilon}.
\end{aligned}$$

Wir werden unten zeigen, dass für alle $\lambda \geq 0$ die Abschätzung

$$p\, e^{\lambda q} + q\, e^{-\lambda p} \leq e^{\lambda^2/8} \tag{3.1}$$

gilt. Damit erhalten wir dann

$$\mathbb{P}[S_n \geq n\,(p+\varepsilon)] \leq e^{n\,(\frac{\lambda^2}{8} - \lambda\varepsilon)}.$$

Der Exponent auf der rechten Seite ist minimal für $\lambda = 4\varepsilon$. Mit dieser Wahl von λ folgt schließlich

$$\mathbb{P}[S_n \geq n\,(p+\varepsilon)] \leq e^{-2n\varepsilon^2}.$$

Die Abschätzung für $\mathbb{P}[S_n \leq n\,(p-\varepsilon)]$ zeigt man analog, und erhält so die Aussage des Satzes.

Nachzutragen bleibt nur noch der Beweis der Abschätzung (3.1). Sei dazu

$$f(\lambda) := \log\left(p\, e^{\lambda q} + q\, e^{-\lambda p}\right) = \log\left(e^{-\lambda p}\,(p\, e^{\lambda} + q)\right) = -\lambda\, p + \log\left(p\, e^{\lambda} + q\right).$$

Zu zeigen ist $f(\lambda) \le \lambda^2/8$ für alle $\lambda \ge 0$. Es gilt $f(0) = 0$,

$$f'(\lambda) \;=\; -p + \frac{p\,e^{\lambda}}{p\,e^{\lambda} + q} \;=\; -p + \frac{p}{p + q\,e^{-\lambda}}, \qquad f'(0) \;=\; 0,$$
$$f''(\lambda) \;=\; \frac{p\,q\,e^{-\lambda}}{(p + q\,e^{-\lambda})^2} \;\le\; \frac{1}{4}.$$

Hierbei haben wir im letzten Schritt die elementare Ungleichung

$$(a+b)^2 \;=\; a^2 + b^2 + 2\,a\,b \;\ge\; 4\,a\,b$$

benutzt. Damit folgt für $\lambda \ge 0$ wie behauptet

$$f(\lambda) \;=\; \int_0^{\lambda} f'(x)\,dx \;=\; \int_0^{\lambda}\int_0^{x} f''(y)\,dy\,dx \;\le\; \int_0^{\lambda} \frac{x}{4}\,dx \;\le\; \frac{\lambda^2}{8}. \qquad \blacksquare$$

Zur Illustration des Satzes simulieren wir den Verlauf von S_k und S_k/k für $k \le n$ und $p = 0.7$ mehrfach (30 mal, siehe Abbildung 3.1 und 3.2), und plotten die Massenfunktionen von S_n, Abbildung 3.3.

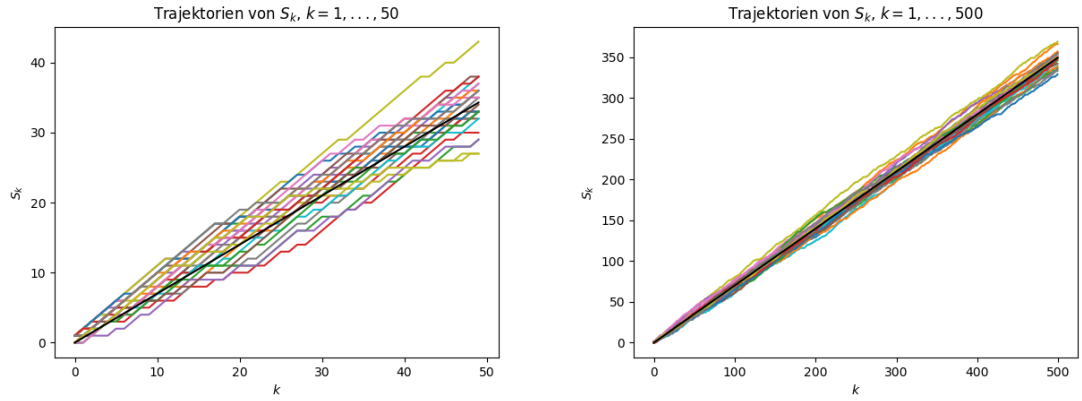


Abbildung 3.1.: Verlauf von S_k für $k \le 50$ bzw. $k \le 500$

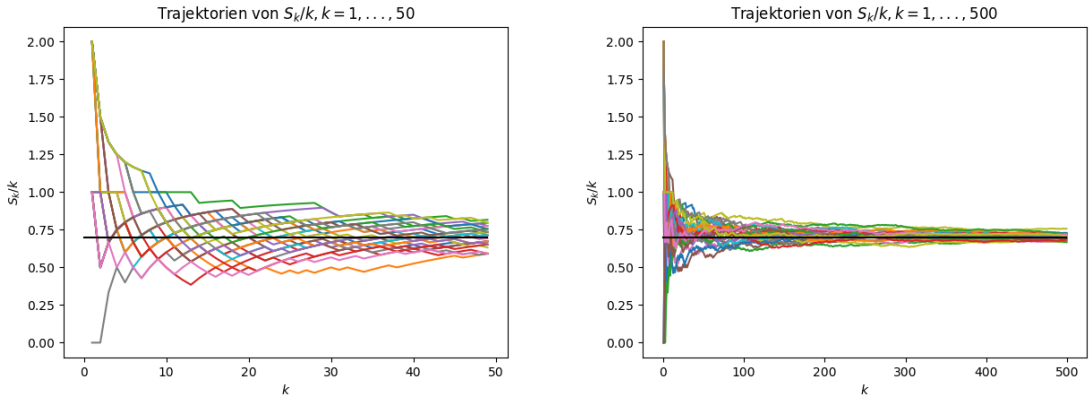


Abbildung 3.2.: Verlauf von S_k/k für $k \le 50$ bzw. $k \le 500$.

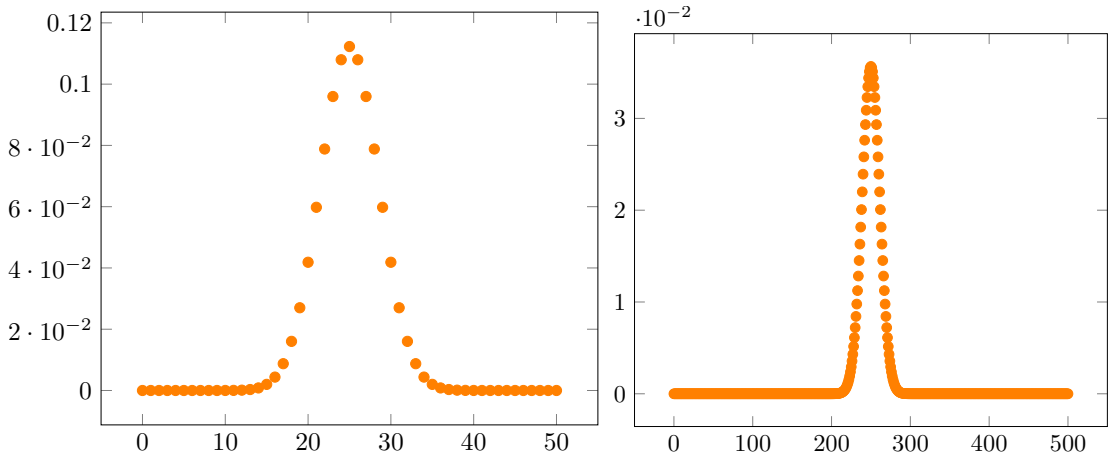


Abbildung 3.3.: Massenfunktion von S_{50} bzw. S_{500}.

Beispiel (Konfidenzintervalle bei Wahlumfragen). Sei p der Stimmenanteil einer Partei in der Gesamtheit aller Wähler. Um p zu schätzen befragen wir n Wähler. Der Einfachheit halber nehmen wir an, dass diese unabhängig voneinander und rein zufällig aus der Gesamtheit aller Wähler ausgewählt werden. Sei A_i das Ereignis, dass der i-te Wähler in unserer Zufallsstichprobe die Partei wählt. Dann sind die Ereignisse $A_1, \dots, A_n$ unabhängig mit Wahrscheinlichkeit p, und $\hat{p}_n := S_n/n$ ist der Stimmenanteil der Partei in unserer Stichprobe. Nach der Bernstein-Ungleichung gilt also

$$\mathbb{P}[|\hat{p}_n - p| \geq \varepsilon] \;\leq\; 2e^{-2\varepsilon^2 n}.$$

Ist zum Beispiel $\varepsilon = 2\%$ und $n = 5000$, dann ist die Wahrscheinlichkeit kleiner als $0,04$, d.h. für den gesuchten Stimmenanteil p gilt

$$\mathbb{P}\left[p \in (\hat{p}_n - \varepsilon, \hat{p}_n + \varepsilon)\right] \;\geq\; 0,96\,.$$

In diesem Fall nennt man das Intervall $(\hat{p}_n - \varepsilon, \hat{p}_n + \varepsilon)$ ein *Konfidenzintervall (Vertrauensintervall) zum Niveau* 96% für den gesuchten Wert p. In der Praxis ist es natürlich nicht möglich, die Stichprobe so zu wählen, dass jeder Wähler mit der gleichen Wahrscheinlichkeit befragt wird, und die Auswertung einer Befragung ist daher wesentlich komplizierter.

Starkes Gesetz der großen Zahlen für unabhängige Ereignisse

Wir zeigen nun, dass aus der Bernstein-Ungleichung auch ein *starkes Gesetz der großen Zahlen* für die relativen Häufigkeiten folgt. Dieses besagt, dass die Zufallsfolge S_n/n mit Wahrscheinlichkeit 1 für $n \to \infty$ gegen p konvergiert. Wir bemerken zunächst, dass $\{\lim S_n/n = p\}$ ein Ereignis in der σ-Algebra $\mathcal{A}$ ist, denn es gilt

$$\lim_{n\to\infty} \frac{S_n(\omega)}{n} = p \quad\Leftrightarrow\quad \forall\, k \in \mathbb{N}\ \exists\, n_0 \in \mathbb{N}\ \forall\, n \geq n_0 \;:\; \left|\frac{S_n(\omega)}{n} - p\right| \leq \frac{1}{k},$$

und damit

$$\left\{\lim_{n\to\infty} \frac{S_n}{n} = p\right\} \;=\; \bigcap_{k=1}^{\infty} \bigcup_{n_0=1}^{\infty} \bigcap_{n=n_0}^{\infty} \left\{\left|\frac{S_n}{n} - p\right| \leq \frac{1}{k}\right\} \;\in\; \mathcal{A}. \tag{3.2}$$

Korollar 3.2 (Starkes GGZ für unabhängige Ereignisse). Es gilt

$$\mathbb{P}\left[\lim_{n\to\infty} \frac{S_n}{n} = p\right] = 1.$$

Ein schwaches Gesetz der großen Zahlen für unabhängige Ereignisse wurde bereits 1689 von Jakob Bernoulli formuliert und bewiesen. Der erste Beweis eines starken Gesetzes der großen Zahlen wurde dagegen erst zu Beginn des 20. Jahrhunderts von Borel, Hausdorff und Cantelli gegeben.

Beweis. Wir zeigen mithilfe der Bernstein-Ungleichung, dass das Gegenereignis $\{S_n/n \not\to p\}$ Wahrscheinlichkeit Null hat. Nach (3.2) gilt

$$\left\{\lim \frac{S_n}{n} \neq p\right\} = \bigcup_{k=1}^{\infty} A_k \qquad \text{mit} \qquad A_k = \bigcap_{n_0=1}^{\infty} \bigcup_{n=n_0}^{\infty} \left\{\left|\frac{S_n}{n} - p\right| > \frac{1}{k}\right\}.$$

Es genügt also $\mathbb{P}[A_k] = 0$ für jedes $k \in \mathbb{N}$ zu zeigen. Sei dazu $k \in \mathbb{N}$ fest gewählt. Aus der Bernstein-Ungleichung folgt für $n_0 \in \mathbb{N}$:

$$\mathbb{P}[A_k] \leq \mathbb{P}\left[\bigcup_{n=n_0}^{\infty} \left\{\left|\frac{S_n}{n} - p\right| > \frac{1}{k}\right\}\right] \leq \sum_{n=n_0}^{\infty} 2e^{-2n/k^2}.$$

Wegen $\sum_{n=1}^{\infty} e^{-2n/k^2} < \infty$ konvergieren die Partialsummen auf der rechten Seite für $n_0 \to \infty$ gegen Null. Also folgt $\mathbb{P}[A_k] = 0$, und damit die Behauptung. ■

Im Beweis haben wir die folgende Aussage benutzt, die aus den Kolmogorovschen Axiomen folgt.

Lemma 3.3 (σ-Subadditivität). *Für beliebige Ereignisse $A_1, A_2, \ldots \in \mathcal{A}$ gilt*

$$\mathbb{P}\left[\bigcup_{n=1}^{\infty} A_n\right] \leq \sum_{n=1}^{\infty} \mathbb{P}[A_n].$$

Beweis. Die Mengen $B_n := A_n \setminus (A_{n-1} \cup \cdots \cup A_1)$ sind disjunkt mit $\bigcup_{n=1}^{\infty} B_n = \bigcup_{n=1}^{\infty} A_n$. Wegen $B_n \subseteq A_n$ erhalten wir $\mathbb{P}\left[\bigcup_{n=1}^{\infty} A_n\right] = \mathbb{P}\left[\bigcup_{n=1}^{\infty} B_n\right] = \sum_{n=1}^{\infty} \mathbb{P}[B_n] \leq \sum_{n=1}^{\infty} \mathbb{P}[A_n]$. ■

Beispiel (Irrfahrt auf $\mathbb{Z}$). Wir betrachten einen Random Walk

$$Z_n = X_1 + X_2 + X_3 + \ldots + X_n \quad (n \in \mathbb{N})$$

mit unabhängigen identisch verteilten Inkrementen $X_i, i \in \mathbb{N}$, mit

$$\mathbb{P}[X_i = 1] = p \quad \text{und} \quad \mathbb{P}[X_i = -1] = 1 - p, \quad p \in (0,1) \text{ fest.}$$

Die Ereignisse $A_i := \{X_i = 1\}$ sind unabhängig mit $\mathbb{P}[A_i] = p$ und es gilt:

$$X_i = I_{A_i} - I_{A_i^C} = 2I_{A_i} - 1,$$

also

$$Z_n = 2S_n - n, \quad \text{wobei} \quad S_n = \sum_{i=1}^{n} I_{A_i}.$$

Nach Korollar 3.2 folgt

$$\lim_{n\to\infty} \frac{Z_n}{n} = 2 \lim_{n\to\infty} \frac{S_n}{n} - 1 = 2p - 1 \qquad \mathbb{P}\text{-fast sicher (d.h. mit Wahrscheinlichkeit 1).}$$

Für $p \neq \frac{1}{2}$ wächst (bzw. fällt) Z_n mit Wahrscheinlichkeit 1 asymptotisch linear (siehe Abbildung 3.4), d.h. für $n \to \infty$ gilt

$$Z_n \sim (2p-1) \cdot n \quad \mathbb{P}\text{-fast sicher.}$$

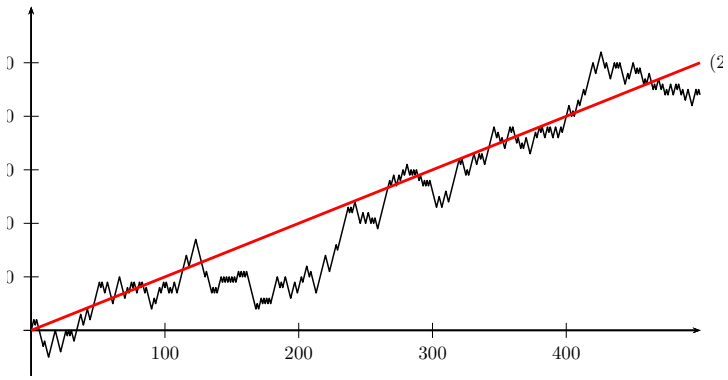


Abbildung 3.4.: Random Walk mit Drift: $p = 0.55, n = 500$

Für $p = \frac{1}{2}$ dagegen wächst der Random Walk sublinear, d.h. $\frac{Z_n}{n} \to 0$ $\mathbb{P}$-fast sicher. In diesem Fall liegt für hinreichend große n der Graph einer typischen Trajektorie $Z_n(\omega)$ in einem beliebig kleinen Sektor um die x-Achse (siehe Abbildung 3.5).

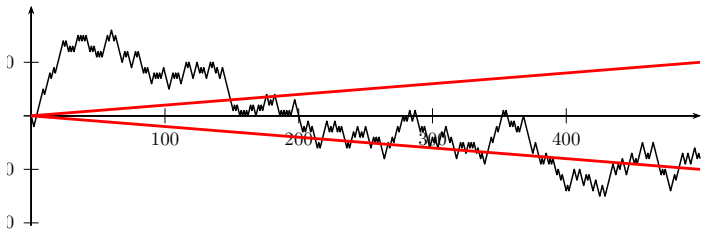


Abbildung 3.5.: Random Walk ohne Drift: $p = 0.5, n = 500$

3.2. Varianz und Kovarianz

Im nächsten Abschnitt werden wir ein Gesetz der großen Zahlen für schwach korrelierte Zufallsvariablen beweisen. Als Vorbereitung führen wir in diesem Abschnitt die Begriffe der Varianz und Standardabweichung, sowie Kovarianz und Korrelation reellwertiger Zufallsvariablen ein, und beweisen zwei wichtige Ungleichungen.

Varianz und Standardabweichung

Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $X : \Omega \to S \subseteq \mathbb{R}$ eine reellwertige Zufallsvariable auf $(\Omega, \mathcal{A}, \mathbb{P})$ mit abzählbarem Wertebereich S. Wir setzen voraus, dass $\mathbb{E}[|X|]$ endlich ist.

Definition 3.4. Die **Varianz** von X ist definiert als mittlere quadratische Abweichung vom Erwartungswert, d.h.

$$\mathrm{Var}[X] \;=\; \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \;\in\; [0, \infty].$$

Die Größe $\sigma[X] \;=\; \sqrt{\mathrm{Var}[X]}$ heißt **Standardabweichung** von X.

Die Varianz bzw. Standardabweichung kann als Kennzahl für die Größe der Fluktuationen (Streuung) der Zufallsvariablen X um den Erwartungswert $\mathbb{E}[X]$ und damit als Maß für das Risiko bei Prognose des Ausgangs $X(\omega)$ durch $\mathbb{E}[X]$ interpretiert werden.

Bemerkung (Eigenschaften der Varianz). a) Die Varianz einer Zufallsvariable hängt nur von ihrer Verteilung ab. Es gilt

$$\mathrm{Var}[X] \;=\; \sum_{a \in S} (a - m)^2 \, p_X(a),$$

wobei $m := \mathbb{E}[X] = \sum_{a \in S} a \, p_X(a)$ der Erwartungswert von X ist.

b) Aus der Linearität des Erwartungswerts folgt

$$\mathrm{Var}[X] \;=\; \mathbb{E}\left[X^2 - 2X \cdot \mathbb{E}[X] + \mathbb{E}[X]^2\right] \;=\; \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2.$$

Insbesondere ist die Varianz von X genau dann endlich, wenn $\mathbb{E}[X^2]$ endlich ist.

c) Entsprechend folgt aus der Linearität des Erwartungswerts

$$\mathrm{Var}[aX + b] \;=\; \mathrm{Var}[aX] \;=\; a^2 \, \mathrm{Var}[X] \qquad \text{für alle } a, b \in \mathbb{R}.$$

d) Die Varianz von X ist genau dann gleich 0, wenn X *deterministisch* ist, d.h. falls

$$\mathbb{P}[X = \mathbb{E}[X]] \;=\; 1.$$

Beispiele. a) Varianz von Bernoulli-Verteilungen: Sei $X = 1$ mit Wahrscheinlichkeit p, und $X = 0$ mit Wahrscheinlichkeit $1 - p$. Dann gilt $\mathbb{E}\left[X^2\right] = \mathbb{E}[X] = p$, und damit

$$\mathrm{Var}[X] \;=\; p - p^2 \;=\; p\,(1 - p).$$

b) Varianz von geometrischen Verteilungen: Sei T geometrisch verteilt mit Parameter $p \in (0, 1]$. Dann gilt $\mathbb{P}[T = k] = (1-p)^{k-1} p$ für alle $k \in \mathbb{N}$. Durch zweimaliges Differenzieren der Identität $\sum_{k=0}^{\infty} (1 - p)^k = 1/p$ erhalten wir

$$\mathbb{E}[T] \;=\; \sum_{k=1}^{\infty} k\,(1-p)^{k-1}\, p \;=\; -p\,\frac{d}{dp}\,\frac{1}{p} \;=\; \frac{1}{p}, \qquad \text{sowie}$$

$$\mathbb{E}[(T+1)\,T] \;=\; \sum_{k=1}^{\infty} (k+1)\,k\,(1-p)^{k-1}\, p \;=\; \sum_{k=2}^{\infty} k\,(k-1)\,(1-p)^{k-2}\, p \;=\; p\,\frac{d^2}{dp^2}\,\frac{1}{p} \;=\; \frac{2}{p^2}.$$

Damit ergibt sich $\mathbb{E}[T^2] = \frac{2}{p^2} - \frac{1}{p}$, und somit

$$\mathrm{Var}[T] \;=\; \mathbb{E}\left[T^2\right] - \mathbb{E}[T]^2 \;=\; \frac{1}{p^2} - \frac{1}{p} \;=\; \frac{1-p}{p^2}.$$

Im folgenden bezeichnen wir mit $\mathcal{L}^p(\Omega, \mathcal{A}, \mathbb{P})$ für $p \in [1, \infty)$ den Raum aller (diskreten) Zufallsvariablen $X : \Omega \to \mathbb{R}$ mit $\mathbb{E}[|X|^p] < \infty$. Dieser Raum ist ein Vektorraum. Ist der Wahrscheinlichkeitsraum fest vorgegeben, dann schreiben wir auch kurz $\mathcal{L}^p$ statt $\mathcal{L}^p(\Omega, \mathcal{A}, \mathbb{P})$. Die Zufallsvariablen aus $\mathcal{L}^1(\Omega, \mathcal{A}, \mathbb{P})$ haben einen endlichen Erwartungswert. Gilt $X \in \mathcal{L}^2(\Omega, \mathcal{A}, \mathbb{P})$, dann ist auch die Varianz von X endlich.

Die folgende wichtige Ungleichung spielt unter anderem im Beweis des Gesetzes der großen Zahlen im nächsten Abschnitt eine zentrale Rolle.

Satz 3.5 (Čebyšev-Ungleichung). Für $X \in \mathcal{L}^2(\Omega, \mathcal{A}, \mathbb{P})$ und $c > 0$ gilt:

$$\mathbb{P}[|X - \mathbb{E}[X]| \geq c] \;\leq\; \frac{1}{c^2}\,\mathrm{Var}[X].$$

Beweis. Es gilt

$$I_{\{|X-\mathbb{E}[X]|\geq c\}} \;\leq\; \frac{1}{c^2}(X - \mathbb{E}[X])^2,$$

denn der Term auf der rechten Seite ist nichtnegativ und ≥ 1 auf $\{|X - \mathbb{E}[X]| \geq c\}$. Durch Bilden des Erwartungswerts folgt

$$\mathbb{P}\left[|X - \mathbb{E}[X]| \geq c\right] \;=\; \mathbb{E}\left[I_{\{|X-\mathbb{E}[X]|\geq c\}}\right] \;\leq\; \mathbb{E}[\frac{1}{c^2}(X - \mathbb{E}[X])^2] \;=\; \frac{1}{c^2}\mathbb{E}\left[(X - \mathbb{E}[X])^2\right],$$

und damit die Behauptung. ∎

Kovarianz und Korrelation

Für Zufallsvariablen $X, Y \in \mathcal{L}^2$ können wir die Kovarianz und die Korrelation definieren.

Definition 3.6. Seien X und Y Zufallsvariablen in $\mathcal{L}^2(\Omega, \mathcal{A}, \mathbb{P})$.

(i) Die **Kovarianz** von X und Y ist definiert als

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

(ii) Gilt $\sigma[X]\,\sigma[Y] \neq 0$, so heißt

$$\varrho[X, Y] \;=\; \frac{\mathrm{Cov}[X, Y]}{\sigma[X]\,\sigma[Y]}$$

Korrelationskoeffizient von X und Y.

(iii) Die Zufallsvariablen X und Y heißen **unkorreliert**, falls $\mathrm{Cov}[X, Y] = 0$, d.h. falls

$$\mathbb{E}[XY] \;=\; \mathbb{E}[X] \cdot \mathbb{E}[Y].$$

Gilt $\mathrm{Cov}[X, Y] > 0$ bzw. < 0, dann heißen X und Y **positiv** bzw. **negativ korreliert**.

Satz 3.7 (Cauchy-Schwarz-Ungleichung für Kovarianz).

(i) Die Kovarianz ist eine symmetrische und bilineare Abbildung von $\mathcal{L}^2 \times \mathcal{L}^2$ nach $\mathbb{R}$ mit

$$\mathrm{Cov}[X, X] \;=\; \mathrm{Var}[X] \;\geq\; 0 \qquad \text{für alle } X \in \mathcal{L}^2.$$

(ii) Für $X, Y \in \mathcal{L}^2$ gilt die *Cauchy-Schwarz-Ungleichung*

$$|\mathrm{Cov}[X, Y]| \;\leq\; \sqrt{\mathrm{Var}[X]} \cdot \sqrt{\mathrm{Var}[Y]} \;=\; \sigma[X] \cdot \sigma[Y]. \tag{3.3}$$

Insbesondere gilt für den Korrelationskoeffizienten im Fall $\sigma[X] \cdot \sigma[Y] \neq 0$ stets

$$|\varrho[X, Y]| \quad\leq\quad 1. \tag{3.4}$$

(iii) Gleichheit gilt in den Ungleichungen (3.3) bzw. (3.4) genau dann, wenn Konstanten $a, b \in \mathbb{R}$ existieren, sodass

$$Y \;=\; aX + b \qquad \text{mit Wahrscheinlichkeit 1.}$$

In diesem Fall ist $\varrho[X, Y] = 1$ falls $a > 0$, und $\varrho[X, Y] = -1$ falls $a < 0$.

Beweis. Nach Definition gilt $\mathrm{Cov}[X, Y] = \mathrm{Cov}[Y, X]$ und $\mathrm{Cov}[X, X] = \mathrm{Var}[X]$. Außerdem folgt aus der Linearität des Erwartungswerts für $X, Y, Z \in \mathcal{L}^2$ und $a \in \mathbb{R}$:

$$\mathrm{Cov}[X, aY + Z] = \mathbb{E}\left[(X - \mathbb{E}[X])(aY + Z - \mathbb{E}[aY + Z])\right] = a\,\mathrm{Cov}[X, Y] + \mathrm{Cov}[X, Z].$$

Somit ist die Kovarianz linear in der zweiten Komponente und damit wegen der Symmetrie auch bilinear. Cov ist also eine nicht-negative definite symmetrische Bilinearform auf dem Vektorraum $\mathcal{L}^2$. Damit gilt insbesondere die Cauchy-Schwarz-Ungleichung, siehe die Vorlesung Lineare Algebra. Den letzten Teil der Aussage und auch die Cauchy-Schwarz-Ungleichung werden wir gleich nebenbei im Rahmen eines Exkurses zu linearen Prognosen beweisen. ■

Die Bilinearität und Symmetrie der Kovarianz können wir benutzen, um die Varianz von Summen von Zufallsvariablen zu berechnen. Zum Beispiel erhalten wir für $X, Y \in \mathcal{L}^2$:

$$\begin{aligned} \mathrm{Var}[X + Y] \;&=\; \mathrm{Cov}[X + Y, X + Y] \;=\; \mathrm{Cov}[X, X] + 2\,\mathrm{Cov}[X, Y] + \mathrm{Cov}[Y, Y] \\ &=\; \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\,\mathrm{Cov}[X, Y]. \end{aligned}$$

Der Kovarianzterm ist gleich 0 falls X und Y unkorreliert sind. Dies ist insbesondere für unabhängige Zufallsvariablen der Fall, denn für diese gilt

$$\mathrm{Cov}[X, Y] \;=\; E[X \cdot Y] - \mathbb{E}[X] \cdot \mathbb{E}[Y] \;=\; 0.$$

Allgemeiner gilt sogar:

Satz 3.8 (Zusammenhang von Unabhängigkeit und Unkorreliertheit). Seien $X : \Omega \to S$ und $Y : \Omega \to T$ diskrete Zufallsvariablen auf $(\Omega, \mathcal{A}, \mathbb{P})$. Dann sind äquivalent:

(i) X und Y sind unabhängig.

(ii) $f(X)$ und $g(Y)$ sind unkorreliert für beliebige Funktionen $f : S \to \mathbb{R}$ und $g : T \to \mathbb{R}$ mit $f(X), g(Y) \in \mathcal{L}^2$.

Bemerkung. Nach Satz 2.14 ist Bedingung (i) äquivalent zu

$$\mathbb{P}[X = a, Y = b] \;=\; \mathbb{P}[X = a]\,\mathbb{P}[Y = b] \qquad \text{für alle } a \in S \text{ und } b \in T.$$

Entsprechend ist Bedingung (ii) genau dann erfüllt, wenn

$$\mathbb{E}[f(X)\,g(Y)] = \mathbb{E}[f(X)]\,\mathbb{E}[g(Y)] \qquad \text{für alle } f : S \to \mathbb{R}, g : T \to \mathbb{R} \text{ mit } f(X), g(Y) \in \mathcal{L}^2 \text{ gilt.}$$

Beweis. (i)⇒(ii): Sind X und Y unabhängig, und $f(X), g(Y) \in \mathcal{L}^2$, dann folgt

$$\begin{aligned}\mathbb{E}[f(X)g(Y)] \;&=\; \sum_{a\in S}\sum_{b\in T} f(a)\,g(b)\,\mathbb{P}[X = a, Y = b] \\ &=\; \sum_{a\in S} f(a)\,\mathbb{P}[X = a] \sum_{b\in T} g(b)\,\mathbb{P}[Y = b] \;=\; \mathbb{E}[f(X)]\,\mathbb{E}[g(Y)].\end{aligned}$$

(ii)⇒(i): Durch Wahl von $f = I_{\{a\}}$ und $g = I_{\{b\}}$ folgt aus (ii) für $a \in S$ und $b \in T$:

$$\begin{aligned}\mathbb{P}[X = a, Y = b] \;&=\; \mathbb{E}[I_{\{a\}}(X)\,I_{\{b\}}(Y)] \\ &=\; \mathbb{E}[I_{\{a\}}(X)]\,\mathbb{E}[I_{\{b\}}(Y)] \;=\; \mathbb{P}[X = a]\,\mathbb{P}[Y = b].\end{aligned}$$

■

Das folgende einfache Beispiel zeigt, dass allein aus der Unkorreliertheit zweier Zufallsvariablen X und Y nicht deren Unabhängigkeit folgt.

Beispiel (Unkorreliertheit ohne Unabhängigkeit). Sei $X = +1$, 0, bzw. -1, jeweils mit Wahrscheinlichkeit $1/3$, und sei $Y = X^2$. Dann sind X und Y nicht unabhängig, aber unkorreliert, denn

$$\begin{aligned}\mathbb{P}[X = 0, Y = 0] \;&=\; 1/3 \;\neq\; 1/9 \;=\; \mathbb{P}[X = 0]\,\mathbb{P}[Y = 0], \\ \mathbb{E}[X\,Y] \;&=\; 0 \;=\; \mathbb{E}[X]\,\mathbb{E}[Y].\end{aligned}$$

Lineare Prognosen und Regressionsgeraden

Angenommen, wir wollen den Ausgang eines Zufallsexperiments vorhersagen, dass durch eine reellwertige Zufallsvariable $Y : \Omega \to \mathbb{R}$ beschrieben wird. Welches ist der *beste Prognosewert* b für $Y(\omega)$, wenn uns keine weiteren Informationen zur Verfügung stehen?

Die Antwort hängt offensichtlich davon ab, wie wir den Prognosefehler messen. Häufig verwendet man den mittleren quadratischen Fehler (*M*ean *S*quared *E*rror)

$$\text{MSE} \quad = \quad \mathbb{E}[(Y - b)^2]\,.$$

Satz 3.9 (Erwartungswert als bester Prognosewert im quadratischen Mittel). Ist Y eine Zufallsvariable in $\mathcal{L}^2(\Omega, \mathcal{A}, \mathbb{P})$, dann gilt für alle $b \in \mathbb{R}$:

$$\mathbb{E}[(Y-b)^2] \quad = \quad \mathrm{Var}[Y] + (b - \mathbb{E}[Y])^2 \quad \geq \quad \mathbb{E}[(Y - \mathbb{E}[Y])^2].$$

Der mittlere quadratische Fehler des Prognosewertes b ist also die Summe der Varianz von Y und des Quadrats des systematischen bzw. mittleren Prognosefehlers (engl. *Bias*) $b - \mathbb{E}[Y]$:

$$\mathrm{MSE} \quad = \quad \mathrm{Varianz} \; + \; \mathrm{Bias}^2.$$

Insbesondere ist der mittlere quadratische Fehler genau für $b = \mathbb{E}[Y]$ minimal.

Beweis. Für $b \in \mathbb{R}$ gilt wegen der Linearität des Erwartungswertes:

$$\mathbb{E}[(Y-b)^2] \quad = \quad \mathrm{Var}[Y-b] \; + \; \mathbb{E}[Y-b]^2 \; = \; \mathrm{Var}[Y] + (\mathbb{E}[Y] - b)^2. \qquad \blacksquare$$

Seien nun $X, Y \in \mathcal{L}^2(\Omega, \mathcal{A}, \mathbb{P})$ quadratintegrierbare Zufallsvariablen mit $\sigma[X] \neq 0$. Angenommen, wir kennen bereits den Wert $X(\omega)$ in einem Zufallsexperiment und suchen die beste *lineare* Vorhersage

$$\hat{Y}(\omega) \quad = \quad aX(\omega) + b, \qquad (a, b \in \mathbb{R}) \tag{3.5}$$

für $Y(\omega)$ im quadratischen Mittel. Zu minimieren ist jetzt der mittlere quadratischen Fehler

$$\mathrm{MSE} \quad := \quad \mathbb{E}[(\hat{Y} - Y)^2]$$

unter allen Zufallsvariablen $\hat{Y}$, die affine Funktionen von X sind. In diesem Fall erhalten wir

$$\mathrm{MSE} \quad = \quad \mathrm{Var}[Y - \hat{Y}] + \mathbb{E}[Y - \hat{Y}]^2 \; = \; \mathrm{Var}[Y - aX] + (\mathbb{E}[Y] - a\mathbb{E}[X] - b)^2.$$

Den zweiten Term können wir für gegebenes a minimieren, indem wir

$$b \quad = \quad \mathbb{E}[Y] - a\mathbb{E}[X]$$

wählen. Für den ersten Term ergibt sich

$$\begin{aligned} \mathrm{Var}[Y - aX] \quad &= \quad \mathrm{Cov}[Y - aX, Y - aX] \; = \; \mathrm{Var}[Y] - 2a\,\mathrm{Cov}[X,Y] + a^2\,\mathrm{Var}[X] \\ &= \quad \left(a \cdot \sigma[X] - \frac{\mathrm{Cov}[X,Y]}{\sigma[X]}\right)^2 + \mathrm{Var}[Y] - \frac{\mathrm{Cov}[X,Y]^2}{\mathrm{Var}[X]}. \end{aligned} \tag{3.6}$$

Dieser Ausdruck wird minimiert, wenn wir $a = \mathrm{Cov}[X,Y]/\sigma[X]^2$ wählen. Die bzgl. des mittleren quadratischen Fehlers optimale Prognose für Y gestützt auf X ist dann

$$\hat{Y}_{\mathrm{opt}} \quad = \quad aX + b \quad = \quad \mathbb{E}[Y] + a(X - \mathbb{E}[X]).$$

Damit haben wir gezeigt:

Satz 3.10 (Lineare Prognose/Regression von Y gestützt auf X). Der mittlere quadratische Fehler $\mathbb{E}[(\hat{Y} - Y)^2]$ ist minimal unter allen Zufallsvariablen der Form $\hat{Y} = aX + b$ mit $a, b \in \mathbb{R}$ für

$$\hat{Y}(\omega) \quad = \quad \mathbb{E}[Y] \; + \; \frac{\mathrm{Cov}[X,Y]}{\mathrm{Var}[X]} \cdot (X(\omega) - \mathbb{E}[X]).$$

Das Problem der linearen Prognose steht in engem Zusammenhang mit der Cauchy-Schwarz-Ungleichung für die Kovarianz. In der Tat ergibt sich diese Ungleichung unmittelbar aus Gleichung (3.6):

Beweis (Cauchy-Schwarz-Ungleichung, Satz 3.7 (ii) und (iii)). Im Fall $\sigma[X] = 0$ gilt $X = \mathbb{E}[X]$ mit Wahrscheinlichkeit 1, und die Ungleichung (3.3) ist trivialerweise erfüllt. Wir nehmen nun an, dass $\sigma[X] \neq 0$ gilt. Wählt man dann wie oben $a = \mathrm{Cov}[X,Y]/\sigma[X]^2$, so folgt aus (3.6) die Cauchy-Schwarz-Ungleichung

$$\mathrm{Var}[Y] - \frac{\mathrm{Cov}[X,Y]^2}{\mathrm{Var}[X]} \quad \geq \quad 0.$$

Die Ungleichung (3.4) folgt unmittelbar. Zudem erhalten wir nach (3.6) genau dann Gleichheit in (3.3) bzw. (3.4), wenn $\mathrm{Var}[Y - aX] = 0$ gilt, also wenn $Y - aX$ mit Wahrscheinlichkeit 1 konstant ist. In diesem Fall folgt $\mathrm{Cov}[X,Y] = \mathrm{Cov}[X, aX] = a\,\mathrm{Var}[X]$, also hat $\varrho[X,Y]$ dasselbe Vorzeichen wie a.■

Beispiel (Regressionsgerade, Methode der kleinsten Quadrate). Wenn die gemeinsame Verteilung von X und Y eine empirische Verteilung von Daten $(x_i, y_i) \in \mathbb{R}^2$, $i = 1, \ldots n$, ist, d.h. wenn

$$(X,Y) \; = \; (x_i, y_i) \qquad \text{mit Wahrscheinlichkeit } 1/n$$

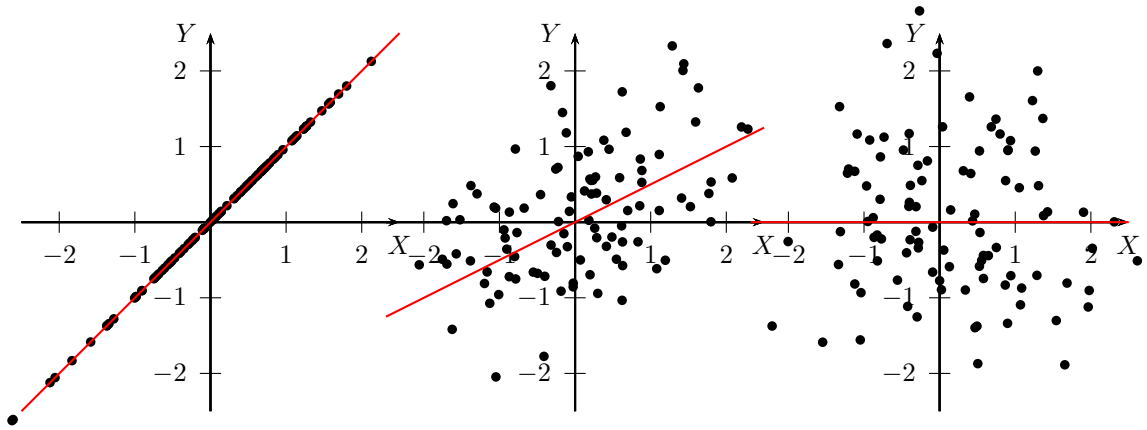
für $1 \leq i \leq n$ gilt, dann sind die Erwartungswerte und die Kovarianz gegeben durch

$$\begin{aligned} \mathbb{E}[X] \quad &= \quad \frac{1}{n}\sum_{i=1}^{n} x_i \;=:\; \overline{x}_n, \qquad \mathbb{E}[Y] \;=\; \overline{y}_n, \\ \mathrm{Cov}[X,Y] \quad &= \quad \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)(y_i - \overline{y}_n) \quad = \quad \frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i\right) - \overline{x}_n \overline{y}_n. \end{aligned}$$

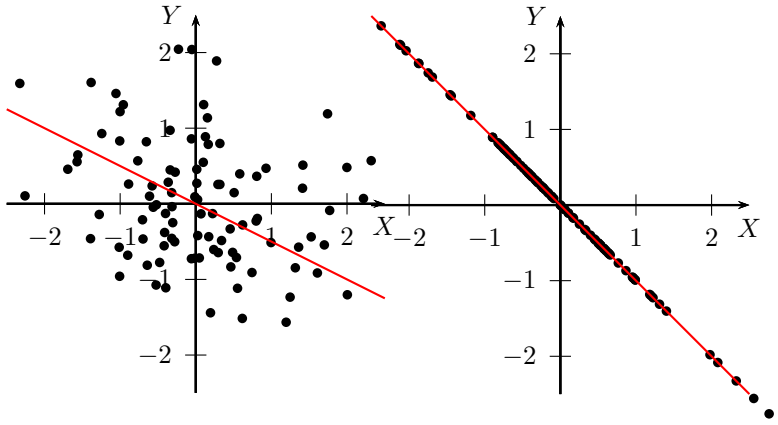
Der entsprechende *empirische Korrelationskoeffizient* der Daten (x_i, y_i), $1 \leq i \leq n$, ist

$$\varrho[X,Y] \quad = \quad \frac{\mathrm{Cov}[X,Y]}{\sigma[X]\sigma[Y]} \quad = \quad \frac{\sum\limits_{i=1}^{n}(x_i - \overline{x}_n)(y_i - \overline{y}_n)}{\left(\sum\limits_{i=1}^{n}(x_i - \overline{x}_n)^2\right)^{1/2}\left(\sum\limits_{i=1}^{n}(y_i - \overline{y}_n)^2\right)^{1/2}}$$

Diesen verwendet man als Schätzer für die Korrelation von Zufallsgrößen mit unbekannten Verteilungen. Die Grafiken in Abbildung 3.6 zeigen Datensätze mit verschiedenen Korrelationskoeffizienten ϱ.



(a) $\varrho = 1$ (b) $\varrho = \frac{1}{2}$ (c) $\varrho = 0$



(d) $\varrho = -\frac{1}{2}$ (e) $\varrho = -1$

Abbildung 3.6.: Korrelationskoeffizienten und Regressionsgeraden für verschiedene Datensätze

Als beste lineare Prognose von Y gestützt auf X im quadratischen Mittel erhalten wir die *Regressionsgerade* $y = ax + b$, die die Quadratsumme

$$\sum_{i=1}^{n} (ax_i + b - y_i)^2 \quad = \quad n \cdot \mathrm{MSE}$$

der Abweichungen minimiert. Hierbei gilt nach Satz 3.10:

$$a = \frac{\mathrm{Cov}[X,Y]}{\sigma[X]^2} = \frac{\sum (x_i - \overline{x}_n)(y_i - \overline{y}_n)}{\sum (x_i - \overline{x}_n)^2} \qquad \text{und} \qquad b = \mathbb{E}[Y] - a \cdot \mathbb{E}[X] = \overline{y}_n - a \cdot \overline{x}_n.$$

Die Regressionsgeraden sind in Abbildung 3.6 eingezeichnet.

3.3. Gesetz der großen Zahlen für schwach korrelierte Zufallsvariablen

Seien $X_1, X_2, \ldots : \Omega \to \mathbb{R}$ Zufallsvariablen, die auf einem gemeinsamen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ definiert sind (z.B. wiederholte Ausführungen desselben Zufallsexperiments), und sei

$$S_n(\omega) \quad = \quad X_1(\omega) + \cdots + X_n(\omega).$$

Wir betrachten die empirischen Mittelwerte

$$\frac{S_n(\omega)}{n} \quad = \quad \frac{X_1(\omega) + \ldots + X_n(\omega)}{n},$$

d.h. die arithmetischen Mittel der ersten n Beobachtungswerte $X_1(\omega), \ldots, X_n(\omega)$. Gesetze der großen Zahlen besagen, dass sich unter geeigneten Voraussetzungen die zufälligen „Fluktuationen" der X_i für große n wegmitteln, d.h. in einem noch zu präzisierenden Sinn gilt

$$\frac{S_n(\omega)}{n} \quad \approx \quad \mathbb{E}\left[\frac{S_n}{n}\right] \qquad \text{für große } n,$$

bzw.

$$\frac{S_n}{n} - \frac{\mathbb{E}[S_n]}{n} \quad \stackrel{n\to\infty}{\longrightarrow} \quad 0. \tag{3.7}$$

Ist insbesondere $\mathbb{E}[X_i] = m$ für alle i, dann sollten die empirischen Mittelwerte S_n/n gegen m konvergieren. Das folgende einfache Beispiel zeigt, dass wir ohne weitere Voraussetzungen an die Zufallsvariablen X_i kein Gesetz der großen Zahlen erwarten können.

Beispiel. Sind die Zufallsvariablen X_i alle gleich, d.h. $X_1 = X_2 = \ldots$, so gilt $\frac{S_n}{n} = X_1$ für alle n. Es gibt also kein Wegmitteln des Zufalls, somit kein Gesetz großer Zahlen.

Andererseits erwartet man ein Wegmitteln des Zufalls bei *unabhängigen* Wiederholungen desselben Zufallsexperiments. Wir werden nun zeigen, dass schon ein rasches Abklingen der Kovarianzen der Zufallsvariablen X_i genügt, um ein Gesetz der großen Zahlen zu erhalten. Dazu berechnen wir die Varianzen der Mittelwerte S_n/n, und schätzen Anschließend die Wahrscheinlichkeiten, dass die zentrierten Mittelwerte in (3.7) einen Wert größer als ε annehmen, durch die Varianzen ab.

Varianz von Summen

Die Varianz einer Summe von reellwertigen Zufallsvariablen können wir mithilfe der Kovarianzen berechnen:

Lemma 3.11. *Für Zufallsvariablen $X_1, \ldots, X_n \in \mathcal{L}^2$ gilt:*

$$\operatorname{Var}[X_1 + \cdots + X_n] \;=\; \sum_{i=1}^{n} \operatorname{Var}[X_i] \;+\; 2 \sum_{\substack{i,j=1 \\ i<j}}^{n} \operatorname{Cov}[X_i, X_j].$$

Falls $X_1, \ldots, X_n$ unkorreliert sind, folgt insbesondere:

$$\operatorname{Var}[X_1 + \cdots + X_n] \;=\; \sum_{i=1}^{n} \operatorname{Var}[X_i].$$

Beweis. Aufgrund der Bilinearität und Symmetrie der Kovarianz gilt

$$\begin{aligned} \operatorname{Var}[X_1 + \cdots + X_n] \;&=\; \operatorname{Cov}\Big[\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\Big] \;=\; \sum_{i,j=1}^{n} \operatorname{Cov}[X_i, X_j] \\ &=\; \sum_{i=1}^{n} \operatorname{Var}[X_i] \;+\; 2 \sum_{\substack{i,j=1 \\ i<j}}^{n} \operatorname{Cov}[X_i, X_j]. \end{aligned}$$

■

Beispiel (Varianz der Binomialverteilung). Eine mit Parametern n und p binomialverteilte Zufallsvariable ist gegeben durch $S_n = \sum_{i=1}^n X_i$ mit unabhängigen, Bernoulli(p)-verteilten Zufallsvariablen X_i, d.h.

$$X_i = \begin{cases} 1 & \text{mit Wahrscheinlichkeit } p, \\ 0 & \text{mit Wahrscheinlichkeit } 1-p. \end{cases}$$

Da unabhängige Zufallsvariablen auch unkorreliert sind, erhalten wir mit Lemma 3.11 für die Varianz der Binomialverteilung:

$$\operatorname{Var}[S_n] \;=\; \sum_{i=1}^n \operatorname{Var}[X_i] \;=\; n\,p\,(1-p).$$

Insbesondere ist die Standardabweichung einer Bin(n, p)-verteilten Zufallsvariable von der Ordnung $O(\sqrt{n})$.

Gesetz der großen Zahlen

Für den Beweis des Gesetzes der großen Zahlen nehmen wir an, dass $X_1, X_2, \ldots$ diskrete Zufallsvariablen aus $\mathcal{L}^2(\Omega, \mathcal{A}, \mathbb{P})$ sind, die die folgende Voraussetzung erfüllen:

Annahme (Schneller Abfall der Korrelationen): Es existiert eine Folge $c_n \in \mathbb{R}_+$ ($n \in \mathbb{Z}_{\geq 0}$) mit

$$\sum_{n=0}^{\infty} c_n \;<\; \infty \qquad \text{und} \qquad \operatorname{Cov}[X_i, X_j] \;\leq\; c_{|i-j|} \quad \text{für alle } i, j \in \mathbb{N}. \tag{3.8}$$

Die Annahme ist z.B. immer erfüllt, wenn die beiden folgenden Bedingungen erfüllt sind:

(i) Die Zufallsvariablen sind unkorreliert: $\operatorname{Cov}[X_i, X_j] \;=\; 0$ für alle $i \neq j$.

(ii) Die Varianzen sind beschränkt: $v \;:=\; \sup_{i \in \mathbb{N}} \operatorname{Var}[X_i] \;<\; \infty$.

In diesem Fall können wir in (3.8) $c_0 = v$ und $c_n = 0$ für $n \neq 0$ wählen. Insbesondere setzen wir keine Unabhängigkeit voraus, sondern nur Bedingungen an die Kovarianzen.

Satz 3.12 (Gesetz der großen Zahlen für schwach korrelierte Zufallsvariablen). Ist die Annahme erfüllt, dann gilt für alle $\varepsilon > 0$ und $n \in \mathbb{N}$:

$$\mathbb{P}\left[\left|\frac{S_n}{n} - \frac{\mathbb{E}[S_n]}{n}\right| \geq \varepsilon\right] \;\leq\; \frac{C}{\varepsilon^2 n} \qquad \text{mit} \qquad C \;:=\; c_0 + 2\sum_{n=1}^{\infty} c_n \;<\; \infty.$$

Ist insbesondere $\mathbb{E}[X_i] = m$ für alle $i \in \mathbb{N}$, dann ***konvergieren*** die Mittelwerte ***stochastisch*** gegen den Erwartungswert m, d.h.

$$\lim_{n\to\infty} \mathbb{P}\left[\left|\frac{S_n}{n} - m\right| \geq \varepsilon\right] = 0 \qquad \text{für jedes } \varepsilon > 0.$$

Der Beweis des Gesetzes der großen Zahlen ergibt sich unmittelbar aus Lemma 3.11 und Satz 3.5:

Beweis. Nach der Annahme und Lemma 3.11 gilt

$$\operatorname{Var}\left[\frac{S_n}{n}\right] \;=\; \frac{1}{n^2}\operatorname{Var}\left[\sum_{i=1}^{n} X_i\right] \;=\; \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\operatorname{Cov}[X_i, X_j] \;\leq\; \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} c_{|i-j|} \;\leq\; \frac{C}{n}.$$

Die Varianz der Mittelwerte fällt also mit Ordnung $O(1/n)$ ab. Mithilfe der Čebyšev-Ungleichung erhalten wir

$$\mathbb{P}\left[\left|\frac{S_n}{n} - \frac{\mathbb{E}[S_n]}{n}\right| \geq \varepsilon\right] \;\leq\; \frac{1}{\varepsilon^2}\operatorname{Var}\left[\frac{S_n}{n}\right] \;\leq\; \frac{C}{n\,\varepsilon^2}.$$

für alle $\varepsilon > 0$ und $n \in \mathbb{N}$. ■

Beispiel. Sind $X_1, X_2, \ldots$ unkorrelierte (also beispielsweise unabhängige) und identisch verteilte Zufallsvariablen aus $\mathcal{L}^2(\Omega, \mathcal{A}, \mathbb{P})$ mit $\mathbb{E}[X_i] = m$ und $\operatorname{Var}[X_i] = v$ für alle i, dann ist die Annahme mit $c_0 = v$ und $c_n = 0$ für $n \neq 0$ erfüllt, und wir erhalten die Abschätzung

$$\mathbb{P}\left[\left|\frac{S_n}{n} - m\right| \geq \varepsilon\right] \;\leq\; \frac{C}{\varepsilon^2\, n}$$

für den Abstand des Mittelwerts der Zufallsvariablen vom Erwartungswert.

Unter den Voraussetzungen aus Satz 3.12 gilt auch ein starkes Gesetz der großen Zahlen:

Satz 3.13 (Starkes Gesetz der großen Zahlen für schwach korrelierte Zufallsvariablen). Ist die Annahme oben erfüllt, und gilt $\mathbb{E}[X_i] = m$ für alle $i \in \mathbb{N}$, dann ***konvergieren*** die Mittelwerte ***fast sicher*** gegen den Erwartungswert m, d.h.

$$\mathbb{P}\left[\lim_{n\to\infty} \frac{S_n}{n} = m\right] = 1\,.$$

Der Beweis dieser Aussage wird in der Vorlesung Einführung in die Wahrscheinlichkeitstheorie gegeben.

Schätzen von Kenngrößen

Sei X eine reellwertige Zufallsvariable mit $\mathbb{E}[X^2] < \infty$. In vielen Anwendungen kennen wir die Verteilung μ von X nicht, oder wir können Erwartungswerte und Wahrscheinlichkeiten nicht explizit berechnen. In diesen Fällen können wir solche Kenngrößen aus unabhängigen Stichproben von X schätzen. Dies wird sowohl in der Statistik bei der Parameterschätzung, als auch in der stochastischen Simulation bei der Monte-Carlo Berechnung von Erwartungswerten verwendet. Im ersten Fall sind die Stichproben Beobachtungswerte, im zweiten Fall werden sie auf dem Computer simuliert.

Wir interpretieren die Stichproben als Realisierungen unabhängiger Zufallsvariablen $X_1, X_2, \ldots$, die auf einem gemeinsamen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ definiert sind.

(i) Schätzen des Erwartungswertes: Um den Erwartungswert $m = \mathbb{E}[X]$ zu schätzen, verwenden wir die *empirischen Mittelwerte*

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

Das empirische Mittel ist ein *erwartungstreuer Schätzer* für m, d.h. $\overline{X}_n$ ist eine Funktion von den Beobachtungswerten $X_1, \dots, X_n$ mit

$$\mathbb{E}[\overline{X}_n] \;=\; m\,.$$

Nach dem Gesetz der großen Zahlen ist $(\overline{X}_n)_{n\in\mathbb{N}}$ zudem eine *konsistente Folge von Schätzern* für m, d.h. es gilt

$$\overline{X}_n \quad\longrightarrow\quad m \qquad \mathbb{P}\text{-stochastisch bzw. } \mathbb{P}\text{-fast sicher.}$$

Zudem können wir basierend auf den Stichproben *Konfidenzintervalle* für m angeben. Sei dazu $\varepsilon > 0$. Dann gilt wie oben gezeigt

$$\mathbb{P}\left[m \notin (\overline{X}_n - \varepsilon, \overline{X}_n + \varepsilon)\right] \;=\; \mathbb{P}\left[|\overline{X}_n - m| \geq \varepsilon\right] \;\leq\; \frac{\mathrm{Var}[X]}{\varepsilon^2 n}\,.$$

Sind ε und n beispielsweise so gewählt, dass die rechte Seite kleiner oder gleich $0,05$ ist, dann folgt, dass das zufällige Intervall $I = (\overline{X}_n - \varepsilon, \overline{X}_n + \varepsilon)$ ein 95%-*Konfidenzintervall für m* ist, d.h. die Wahrscheinlichkeit, dass der tatsächliche Wert von m in diesem zufälligen Intervall liegt, beträgt mindestens $0,95$.

(ii) Schätzen der Varianz: Um die Varianz $v = \mathrm{Var}[X]$ zu schätzen, verwendet man meistens die *renormierte Stichprobenvarianz*

$$\widetilde{V}_n \quad:=\quad \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

Der Vorfaktor $\frac{1}{n-1}$ (statt $\frac{1}{n}$) gewährleistet, dass $\widetilde{V}_n$ ein *erwartungstreuer* Schätzer für v ist, denn aus

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \;&=\; \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2 \;-\; (\overline{X}_n - m)^2 \qquad (3.9)\\ \text{Stichprobenvarianz} \;&=\; \text{MSE} \;-\; \text{Stichprobenbias}^2 \end{aligned}$$

folgt

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2\right] \;=\; \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}[X_i] - \mathrm{Var}[\overline{X}_n] \;=\; \left(\frac{1}{n} - \frac{1}{n^2}\right)\sum_{i=1}^{n}\mathrm{Var}[X_i] \;=\; \frac{n-1}{n}v,$$

also $\mathbb{E}[\widetilde{V}_n] = v$. Um zu zeigen, dass $(\widetilde{V}_n)_{n\in\mathbb{N}}$ eine *konsistente* Folge von Schätzern für v ist, können wir erneut das Gesetz der großen Zahlen anwenden. Da die Zufallsvariablen $X_i - \overline{X}_n$, $1 \leq i \leq n$, selbst nicht unabhängig sind, verwenden wir dazu die Zerlegung (3.9). Aus dem starken Gesetz der großen Zahlen folgt dann

$$\frac{n-1}{n}\widetilde{V}_n \;=\; \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2 - (\overline{X}_n - m)^2 \quad\longrightarrow\quad v \qquad \mathbb{P}\text{-fast sicher,}$$

also auch $\widetilde{V}_n \to v$ $\mathbb{P}$-fast sicher.

(iii) Schätzen von allgemeinen Erwartungswerten: Allgemeiner können wir für jede Funktion f mit $\mathbb{E}[|f(X)|] < \infty$ den Erwartungswert $\theta = \mathbb{E}[f(X)]$ erwartungstreu durch die *empirischen Mittelwerte*

$$\widehat{\theta}_n \quad = \quad \frac{1}{n} \sum_{i=1}^{n} f(X_i)$$

schätzen. Da die Zufallsvariablen $f(X_i)$ wieder unabhängig und identisch verteilt sind mit Erwartungswert θ, gilt nach dem Gesetz der großen Zahlen:

$$\widehat{\theta}_n \quad \longrightarrow \quad \theta \qquad \mathbb{P}\text{-stochastisch und } \mathbb{P}\text{-fast sicher.} \tag{3.10}$$

(iv) Schätzen der Verteilung: Die gesamte Verteilung

$$\mu[B] \quad = \quad \mathbb{P}[X \in B]$$

können wir durch die *empirische Verteilung*

$$\widehat{\mu}_n[B] \quad = \quad \frac{1}{n} \sum_{i=1}^{n} I_B(X_i) \quad = \quad \frac{|\{i = 1, \ldots, n : X_i \in B\}|}{n}$$

der Zufallsstichprobe $X_1, \ldots, X_n$ schätzen. Diese ist eine „*zufällige Wahrscheinlichkeitsverteilung*“. Nach (iii) ist $\widehat{\mu}_n[B]$ ein erwartungstreuer Schätzer für $\mu[B]$, und es gilt

$$\widehat{\mu}_n[B] \quad \overset{n\to\infty}{\longrightarrow} \quad \mu[B] \qquad \mathbb{P}\text{-stochastisch und } \mathbb{P}\text{-fast sicher.} \tag{3.11}$$

Konfidenzintervalle für $\mu[B]$ kann man entweder wie oben mithilfe der Čebyšev-Ungleichung oder, wie im Beispiel in Abschnitt 3.1, über die Bernstein-Ungleichung herleiten.

3.4. Konvergenzsätze für Markov-Ketten

Sei S eine abzählbare Menge, ν eine Wahrscheinlichkeitsverteilung auf S, und $P = (P(x, y))_{x,y\in S}$ eine stochastische Matrix. Hier und im folgenden bezeichnen wir diskrete Wahrscheinlichkeitsverteilungen und die entsprechenden Massenfunktionen mit demselben Buchstaben, d.h. $\nu(x) := \nu[\{x\}]$. Wir interpretieren $\nu = (\nu(x))_{x\in S}$ auch als Zeilenvektor in $\mathbb{R}^S$.

In Abschnitt 2.2 haben wir das kanonische Modell für eine (zeithomogene) Markovkette mit Startverteilung ν und Übergangsmatrix P eingeführt. Allgemeiner definieren wir:

Definition 3.14. Eine Folge $X_0, X_1, \ldots : \Omega \to S$ von Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ heißt **zeitlich homogene Markov-Kette** mit Startverteilung ν und Übergangsmatrix P, falls die folgenden Bedingungen erfüllt sind:

(i) Für alle $x_0 \in S$ gilt $\mathbb{P}[X_0 = x_0] = \nu(x_0)$.

(ii) Für alle $n \in \mathbb{N}$ und $x_0, \ldots, x_{n+1} \in S$ mit $\mathbb{P}[X_0 = x_0, \ldots, X_n = x_n] \neq 0$ gilt

$$\mathbb{P}[X_{n+1} = x_{n+1} \mid X_0 = x_0, \ldots, X_n = x_n] \;=\; P(x_n, x_{n+1}).$$

Die Bedingungen (i) und (ii) sind äquivalent dazu, dass

$$\mathbb{P}[X_0 = x_0, \ldots, X_n = x_n] \;=\; \nu(x_0)\, P(x_0, x_1) \cdots P(x_{n-1}, x_n)$$

für alle $n \in \mathbb{Z}_{\geq 0}$ und $x_0, x_1, \ldots, x_n \in S$ gilt. Eine Folge $(X_k)_{k \in \mathbb{Z}_{\geq 0}}$ von Zufallsvariablen mit Werten in S ist also genau dann eine zeithomogene Markovkette mit Startverteilung ν und Übergangsmatrix P, wenn die gemeinsame Verteilung von $X_0, X_1, \ldots, X_n$ für jedes n mit der Verteilung im entsprechenden kanonischen Modell übereinstimmt.

Gleichgewichte und Detailed Balance

Satz 2.8 zeigt, dass die Verteilung einer zeithomogenen Markovkette zur Zeit n durch das Produkt νP^n des Zeilenvektors ν der Massenfunktion der Startverteilung mit dem n fachen Matrixprodukt der Übergangsmatrix P gegeben ist. Gilt $\nu\, P = \nu$, dann folgt $X_n \sim \nu$ für alle $n \in \mathbb{Z}_{\geq 0}$, d.h. die Markovkette mit Startverteilung ν ist „stationär“.

Definition 3.15. i) Eine Wahrscheinlichkeitsverteilung μ auf S heißt **Gleichgewichtsverteilung** (oder **invariante Verteilung**) der Übergangsmatrix P, falls $\mu\, P = \mu$ gilt, d.h. falls

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

ii) μ erfüllt die **Detailed Balance-Bedingung** bzgl. der Übergangsmatrix P, falls gilt:

$$\mu(x)\, P(x, y) = \mu(y)\, P(y, x) \qquad \text{für alle } x, y \in S \tag{3.12}$$

Satz 3.16. Erfüllt μ die Detailed Balance-Bedingung (3.12), dann ist μ eine Gleichgewichtsverteilung von P.

Beweis. Aus der Detailed Balance-Bedingung folgt

$$\sum_{x \in S} \mu(x)\, P(x, y) \;=\; \sum_{x \in S} \mu(y)\, P(y, x) \;=\; \mu(y) \qquad \text{für alle } y \in S.$$

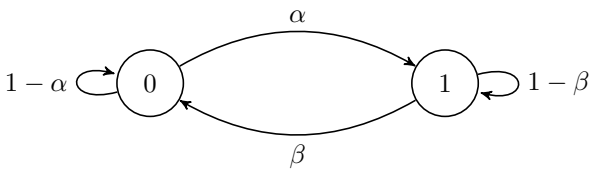
Bemerkung. Bei Startverteilung μ gilt:

$$\mu(x)\, P(x, y) = \mathbb{P}[X_0 = x, X_1 = y].$$

Wir können diese Größe als „Fluss der Wahrscheinlichkeitsmasse von x nach y“ interpretieren. Die Detailed Balance- und die Gleichgewichtsbedingung haben dann die folgenden anschaulichen Interpretationen:

Detailed Balance:	$\mu(x)\, P(x, y)$	=	$\mu(y)\, P(y, x)$
	„Fluss von x nach y“	=	„Fluss von y nach x“
Gleichgewicht:	$\sum_{x \in S} \mu(x)\, P(x, y)$	=	$\sum_{x \in S} \mu(y)\, P(y, x)$
	„Gesamter Fluss nach y“		„Gesamter Fluss von y“.

Beispiele. a) Markov-Kette auf $S = \{0, 1\}$:



Seien $\alpha, \beta \in [0, 1]$ und $P = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$. Dann ist die Gleichgewichtsbedingung $\mu\, p = \mu$ äquivalent zu den folgenden Gleichungen:

$$\mu(0) = \mu(0)\,(1-\alpha) + \mu(1)\,\beta,$$
$$\mu(1) = \mu(0)\,\alpha + \mu(1)\,(1-\beta).$$

Da μ eine Wahrscheinlichkeitsverteilung ist, sind beide Gleichungen äquivalent zu

$$\beta\,(1 - \mu(0)) = \alpha\,\mu(0).$$

Die letzte Gleichung ist äquivalent zur Detailed Balance-Bedingung (3.12). Auf einem Zustandsraum mit zwei Elementen erfüllt also jede Gleichgewichtsverteilung die Detailed Balance-Bedingung. Falls $\alpha + \beta > 0$ gilt, ist $\mu = \left(\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta}\right)$ das eindeutige Gleichgewicht. Falls $\alpha = \beta = 0$ gilt, ist jede Wahrscheinlichkeitsverteilung μ eine Gleichgewichtsverteilung.

b) Zyklischer Random Walk: Sei $S = \mathbb{Z}/n\mathbb{Z}$ ein diskreter Kreis, und

$$P(k + n\mathbb{Z}, k + 1 + n\mathbb{Z}) = p, \qquad P(k + n\mathbb{Z}, k - 1 + n\mathbb{Z}) = 1 - p.$$

Dann ist die Gleichverteilung $\mu(x) = \frac{1}{n}$ für jedes $p \in [0, 1]$ ein Gleichgewicht von P. Die Detailed Balance-Bedingung ist dagegen nur für $p = \frac{1}{2}$, d.h. im symmetrischen Fall, erfüllt.

c) Random Walks auf Graphen:
Sei (V, E) ein endlicher Graph, und $S = V$ die Menge der Knoten. Wir nehmen an, dass von jedem Knoten mindestens eine Kante ausgeht. Der klassische Random Walk auf dem Graphen hat die Übergangswahrscheinlichkeiten

$$P(x, y) \;=\; \begin{cases} \frac{1}{\deg(x)} & \text{falls } \{x, y\} \in E, \\ 0 & \text{sonst.} \end{cases}$$

Die Detailed Balance-Bedingung lautet in diesem Fall:

$$\frac{\mu(x)}{\deg(x)} \;=\; \frac{\mu(y)}{\deg(y)} \qquad \text{für alle } \{x, y\} \in E.$$

Sie ist erfüllt, falls

$$\mu(x) = \deg(x)/Z$$

gilt, wobei Z eine positive Konstante ist. Damit μ eine Wahrscheinlichkeitsverteilung ist, muss

$$Z \;=\; \sum_{x \in B} \deg(x) \;=\; 2\,|E|$$

gelten. Somit ergibt sich als Gleichgewichtsverteilung

$$\mu(x) \;=\; \frac{\deg(x)}{2|E|}.$$

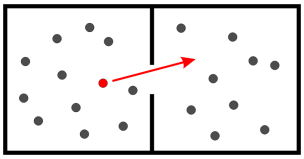
Alternativ können wir einen modifizierten Random Walk definieren, der die Gleichverteilung auf V als Gleichgewicht hat. Sei dazu $\triangle := \max_{x \in V} \deg(x)$ der maximale Grad, und

$$P(x,y) = \begin{cases} \frac{1}{\triangle} & \text{falls } \{x,y\} \in E, \\ 1 - \frac{\deg(x)}{\triangle} & \text{falls } x = y, \\ 0 & \text{sonst.} \end{cases}$$

Dann gilt $P(x,y) = P(y,x)$, und somit ist die Gleichverteilung auf V ein Gleichgewicht.

Ist der Graph regulär, also $\deg(x)$ konstant, dann stimmen die beiden Arten von Random Walks überein.

d) Urnenmodell von P. und T. Ehrenfest: Das Ehrenfestsche Urnenmodell ist ein einfaches Modell, dass den Austausch von Gasmolekülen zwischen zwei Behältern beschreibt, ohne die räumliche Struktur zu berücksichtigen. Im Modell ist eine feste Anzahl n von Kugeln (Molekülen) auf zwei Urnen (Behälter) verteilt. Typischerweise ist n sehr groß, z.B. $n = 10^{23}$. Zu jedem Zeitpunkt $t \in \mathbb{N}$ wechselt eine zufällig ausgewählte Kugel die Urne.



Wir können diesen Vorgang auf zwei ganz verschiedene Arten durch Markovketten beschreiben.

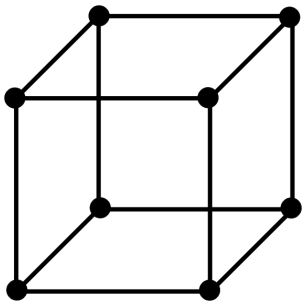
Mikroskopische Beschreibung: Ein detailliertes Modell ergibt sich, wenn wir für jede einzelne Kugel notieren, ob sich diese in der ersten Urne befindet. Der Zustandsraum ist dann

$$S = \{0,1\}^n = \{(\sigma_1, \ldots, \sigma_n) \ : \ \sigma_i \in \{0,1\} \ \forall\, i\},$$

wobei $\sigma_i = 1$ dafür steht, dass sich die i-te Kugel in der ersten Urne befindet. Man beachte, dass dieser Konfigurationsraum enorm viele Elemente enthält (z.B. $2^{10^{23}}$). Die Übergangswahrscheinlichkeiten sind durch

$$P(\sigma, \widetilde{\sigma}) \ = \ \begin{cases} \frac{1}{n} & \text{falls } \sum_{i=1}^n |\sigma_i - \widetilde{\sigma_i}| = 1, \\ 0 & \text{sonst,} \end{cases}$$

gegeben. Die resultierende Markov-Kette ist ein Random Walk auf dem (in der Regel sehr hochdimensionalen) diskreten HyperWürfel $\{0,1\}^n$, d.h. sie springt in jedem Schritt von einer Ecke des HyperWürfels zu einer zufällig ausgewählten benachbarten Ecke. Die Gleichverteilung auf dem Hyperwürfel ist ein Gleichgewicht.



Makroskopische Beschreibung: Wir betrachten nur die Anzahl der Kugeln in der ersten Urne. Der Zustandsraum ist dann

$$S = \{0,1,2,\ldots,n\},$$

und die Übergangswahrscheinlichkeiten sind durch

$$P(x,y) = \begin{cases} \frac{x}{n} & \text{falls } y = x-1, \\ \frac{n-x}{n} & \text{falls } y = x+1, \\ 0 & \text{sonst,} \end{cases}$$

gegeben, da in jedem Schritt mit Wahrscheinlichkeit x/n eine Kugel aus der ersten Urne gezogen wird, wenn sich x Kugeln dort befinden. Da sich im mikroskopischen Gleichgewicht jede Kugel mit Wahrscheinlichkeit $\frac{1}{2}$ in jeder der beiden Urnen befindet, können wir erwarten, dass die Binomialverteilung $\mu(x) = \binom{n}{x} 2^{-n}$ mit Parameter $p = \frac{1}{2}$ ein Gleichgewicht der makroskopischen Dynamik ist. Tatsächlich erfüllt die Binomialverteilung die Detailed Balance-Bedingung

$$\mu(x-1)\,P(x-1,x) \;=\; \mu(x)\,P(x,x-1) \qquad \text{für } x = 1,\dots,n,$$

denn es gilt

$$2^{-n}\,\frac{n!}{(x-1)!(n-(x-1))!}\,\frac{n-(x-1)}{n} \;=\; 2^{-n}\,\frac{n!}{x!(n-x)!}\,\frac{x}{n}.$$

Konvergenz ins Gleichgewicht

Wir wollen nun zeigen, dass sich unter geeigneten Voraussetzungen die Verteilung einer Markovkette zur Zeit n für $n \to \infty$ einer Gleichgewichtsverteilung annähert, die nicht von der Startverteilung abhängt. Um dies mathematisch zu präzisieren, benötigen wir einen Abstandsbegriff für Wahrscheinlichkeitsverteilungen. Sei

$$\mathrm{WV}(S) := \{\nu = (\nu(x))_{x\in S} : \nu(x) \geq 0\ \forall x, \quad \sum_{x\in S} \nu(x) = 1\}$$

die Menge aller (Massenfunktionen von) Wahrscheinlichkeitsverteilungen auf der abzählbaren Menge S. Ist S endlich mit m Elementen, dann ist $\mathrm{WV}(S)$ ein Simplex im $\mathbb{R}^m$. Wir führen nun einen Abstandsbegriff auf $\mathrm{WV}(S)$ ein:

Definition 3.17. Die **(totale) Variationsdistanz** zweier Wahrscheinlichkeitsverteilungen μ, ν auf S ist:

$$d_{TV}(\mu,\nu) \;:=\; \frac{1}{2}\|\mu-\nu\|_1 \;:=\; \frac{1}{2}\sum_{x\in S} |\mu(x)-\nu(x)|.$$

Man prüft leicht nach, dass d_{TV} tatsächlich eine Metrik auf $\mathrm{WV}(S)$ ist.

Bemerkung. a) Für alle $\mu, \nu \in \mathrm{WV}(S)$ gilt:

$$d_{TV}(\mu,\nu) \;\leq\; \frac{1}{2}\sum_{x\in S}(\mu(x)+\nu(x)) \;=\; 1.$$

b) Seien $\mu, \nu \in \mathrm{WV}(S)$ und $B := \{x \in S : \mu(x) \geq \nu(x)\}$. Dann gilt

$$d_{TV}(\mu,\nu) \;=\; \sum_{x\in B}(\mu(x)-\nu(x)) \;=\; \max_{A\subseteq S} |\mu(A)-\nu(A)|.$$

Diese Aussage zeigt, dass d_{TV} eine sehr natürliche Abstandsfunktion auf Wahrscheinlichkeitsverteilungen ist.

Wir betrachten nun eine stochastische Matrix $(P(x,y))_{x,y\in S}$ mit Gleichgewichtsverteilung μ. Die Verteilung einer Markov-Kette mit Startverteilung ν und Übergangsmatrix P zur Zeit n ist νP^n. Um Konvergenz ins Gleichgewicht zu zeigen, verwenden wir die folgende Annahme:

Minorisierungsbedingung: Es gibt ein $\delta \in (0,1)$ und ein $r \in \mathbb{N}$, so dass

$$P^r(x,y) \geq \delta \cdot \mu(y) \qquad \text{für alle } x, y \in S \text{ gilt.} \tag{3.13}$$

Satz 3.18 (Konvergenzsatz von W. Doeblin). Gilt die Minorisierungsbedingung (3.13), dann konvergiert νP^n für jede Startverteilung ν exponentiell schnell gegen μ. Genauer gilt für alle $n \in \mathbb{Z}_{\geq 0}$ und $\nu \in \mathrm{WV}(S)$:

$$d_{TV}(\nu P^n, \mu) \;\leq\; (1-\delta)^{\lfloor n/r \rfloor}.$$

Bemerkung. Insbesondere ist μ unter der Voraussetzung des Satzes das *eindeutige* Gleichgewicht von P, denn für eine beliebige Wahrscheinlichkeitsverteilung ν mit $\nu P = \nu$ gilt

$$d_{TV}(\nu, \mu) \;=\; d_{TV}(\nu P^n, \mu) \;\to\; 0 \qquad \text{für } n \to \infty,$$

und damit $\nu = \mu$.

Beweis. 1. Durch die Zerlegung

$$P^r(x,y) \;=\; \delta\,\mu(y) + (1-\delta)\,Q(x,y)$$

der r-Schritt-Übergangswahrscheinlichkeiten wird eine *stochastische* Matrix Q definiert, denn:

(i) Aus der Minorisierungsbedingung (3.13) folgt $Q(x,y) \geq 0$ für alle $x, y \in S$.

(ii) Aus $\sum_{y\in S} P^r(x,y) = 1$, $\sum_{y\in S} \mu(y) = 1$ folgt $\sum_{y\in S} Q(x,y) = 1$ für alle $x \in S$.

Wir setzen im folgenden $\lambda := 1-\delta$. Dann gilt für alle $\nu \in \mathrm{WV}(S)$:

$$\nu P^r = (1-\lambda)\,\mu + \lambda\,\nu Q. \tag{3.14}$$

2. Wir zeigen mit vollständiger Induktion:

$$\nu P^{kr} = (1-\lambda^k)\,\mu + \lambda^k\,\nu Q^k \qquad \text{für alle } k \geq 0, \quad \nu \in \mathrm{WV}(S). \tag{3.15}$$

für $k = 0$ ist die Aussage offensichtlich wahr. Gilt (3.15) für ein $k \geq 0$, dann erhalten wir durch Anwenden von Gleichung (3.14) auf $\widetilde{\nu} P^r$ mit $\widetilde{\nu} = \nu Q^k$:

$$\begin{aligned}
\nu P^{(k+1)r} &= \nu P^{kr} P^r \\
&= ((1-\lambda^k)\,\mu + \lambda^k \underbrace{\nu Q^k}_{=\widetilde{\nu}})\, P^r \\
&= (1-\lambda^k) \underbrace{\mu P^r}_{=\mu} + (1-\lambda)\,\lambda^k \mu + \lambda^{k+1}\,\nu Q^k Q \\
&= (1-\lambda^{k+1})\,\mu + \lambda^{k+1}\,\nu Q^{k+1}.
\end{aligned}$$

3. Sei $n \in \mathbb{Z}_{\geq 0}$. Dann gilt $n = k\,r + i$ mit $k \in \mathbb{Z}_{\geq 0}$ und $0 \leq i < r$. Damit folgt für $\nu \in \mathrm{WV}(S)$:

$$\nu P^n = \nu P^{kr} P^i = (1-\lambda^k) \underbrace{\mu P^i}_{=\mu} + \lambda^k \nu Q^k P^i, \qquad \text{also}$$

$$\nu P^n - \mu = \lambda^k (\nu Q^k P^i - \mu), \qquad \text{und damit}$$

$$d_{TV}(\nu P^n, \mu) = \frac{1}{2} \|\nu P^n - \mu\|_1 = \lambda^k d_{TV}(\nu Q^k P^i, \mu) \leq \lambda^k. \qquad \blacksquare$$

Auf abzählbar unendlichen Zustandsräumen ist die Minorisierungsbedingung eine relativ restriktive Annahme. Es gibt Erweiterungen des obigen Satzes, die unter deutlich schwächeren Voraussetzungen ähnliche Konvergenzaussagen liefern. Ist der Zustandsraum dagegen endlich, dann können wir den obigen Konvergenzsatz verwenden, um die Konvergenz ins Gleichgewicht unter minimalen Voraussetzungen zu beweisen. Dazu zeigen wir, dass die Minorisierungsbedingung immer erfüllt ist, wenn der Zustandsraum endlich, und die Übergangsmatrix *irreduzibel* ist und einen *aperiodischen Zustand* besitzt:

Definition 3.19. i) Eine stochastische Matrix P heißt **irreduzibel**, falls es für alle $x, y \in S$ ein $n \in \mathbb{N}$ gibt, so dass $P^n(x, y) > 0$ gilt.

ii) Ein Zustand $x \in S$ heißt **aperiodisch bzgl.** P, falls ein $n_0 \in \mathbb{N}$ existiert, so dass $P^n(x, x) > 0$ für alle $n \geq n_0$ gilt.

Bemerkung. a) Allgemeiner definiert man die **Periode** eines Zustands $x \in S$ als

$$\mathrm{Periode}(x) := \mathrm{ggT}\,\{n \in \mathbb{N} \mid P^n(x, x) > 0\}.$$

Man kann dann zeigen, dass x genau dann aperiodisch ist, wenn $\mathrm{Periode}(x) = 1$ gilt. Ein Beispiel für eine Übergangsmatrix mit Periode 2 ist die Matrix $P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ auf einem zweielementigen Zustandsraum. Die entsprechende Markovkette wechselt in jedem Schritt mit Wahrscheinlichkeit 1 den Zustand.

b) Ist P irreduzibel, dann folgt aus der Existenz eines aperiodischen Zustands bereits, dass alle Zustände aperiodisch sind.

Beispiel (Irreduzibilität von Random Walks auf Graphen). Die Übergangsmatrix eines Random Walks auf einem endlichen Graphen ist genau dann irreduzibel, wenn der Graph zusammenhängend ist.

Korollar 3.20 (Konvergenzsatz für endliche Markov-Ketten). Ist der Zustandsraum S endlich, die Übergangsmatrix P irreduzibel, und existiert ein aperiodischer Zustand $a \in S$, dann gilt:

$$\lim_{n\to\infty} d_{TV}(\nu P^n, \mu) = 0 \qquad \text{für alle } \nu \in \mathrm{WV}(S).$$

Beweis. Wir zeigen, dass zu jedem $x, y \in S$ eine natürliche Zahl $k(x, y)$ existiert, so dass

$$P^n(x, y) > 0 \qquad \text{für alle } n \geq k(x, y) \tag{3.16}$$

gilt. Da der Zustandsraum endlich ist, folgt hieraus, dass die Minorisierungsbedingung (3.13) mit

$$r = \max_{x,y \in S} k(x, y) < \infty \qquad \text{und} \qquad \delta = \min_{x,y \in S} P^r(x, y) > 0$$

erfüllt ist.

Zum Beweis der obigen Behauptung seien $x, y \in S$ fest gewählt. Wegen der Irreduzibilität von P existieren dann $i, j \in \mathbb{N}$ mit $P^i(x, a) > 0$ und $P^j(a, y) > 0$. Da a aperiodisch ist, existiert zudem ein $n_0 \in \mathbb{N}$ mit $P^n(a, a) > 0$ für alle $n \geq n_0$. Damit folgt

$$P^{i+n+j}(x, y) \;\geq\; P^i(x, a)\, P^n(a, a)\, P^j(a, y) > 0 \qquad \text{für alle } n \geq n_0,$$

und somit $P^n(x, y) > 0$ für alle $n \geq i + n_0 + j$. Also ist die Behauptung für x, y mit $k(x, y) = i + n_0 + j$ erfüllt. ■

Beispiel (Träger Random Walk auf endlichem Graphen). Ein Random Walk auf einem endlichen Graphen ist im Allgemeinen nicht aperiodisch; zum Beispiel hat der Random Walk auf $\mathbb{Z}/(n\mathbb{Z})$ Periode 2 falls n gerade ist. Um Aperiodizität zu gewährleisten genügt aber eine kleine Modifikation der Übergangsmatrix: Setzen wir

$$P(x, y) \;=\; \begin{cases} \varepsilon & \text{für } y = x, \\ \frac{1-\varepsilon}{\deg(x)} & \text{für } \{x, y\} \in E \text{ mit } x \neq y, \\ 0 & \text{sonst,} \end{cases}$$

mit einer festen Konstanten $\varepsilon > 0$, dann sind alle Zustände aperiodisch, und P hat weiterhin das Gleichgewicht $\mu(x) = \deg(x)/(2|E|)$. Die Markovkette mit Übergangsmatrix P ist ein „träger" Random Walk, der in jedem Schritt mit Wahrscheinlichkeit ε beim selben Zustand bleibt. Ist der Graph zusammenhängend, dann ist P irreduzibel. Es folgt, dass die Verteilung des trägen Random Walks zur Zeit n für eine beliebige Startverteilung gegen μ konvergiert.

Gesetz der großen Zahlen für stationäre Markovketten

Das Gesetz der großen Zahlen kann auch auf Mittelwerte von stationären Markovketten angewendet werden. Sei $(Y_n)_{n \in \mathbb{Z}_{\geq 0}}$ eine auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ definierte Markovkette mit abzählbarem Zustandsraum S und Übergangsmatrix $P = (P(x, y))_{x \in S}$. Wir nehmen an, dass die Markovkette im Gleichgewicht startet, d.h. die Verteilung μ von Y_0 ist ein Gleichgewicht von P. Dann gilt

$$Y_n \sim \mu \qquad \text{für alle } n \geq 0. \tag{3.17}$$

Wir betrachten nun die *Anzahl der Besuche*

$$S_n \;=\; \sum_{i=0}^{n-1} I_A(Y_i)$$

in einer Teilmenge A des Zustandsraums S während der ersten n Schritte der Markovkette. Erfüllt die Übergangsmatrix eine Minorisierungsbedingung, dann können wir zeigen, dass die Kovarianzen der Zufallsvariablen $X_i = I_A(Y_{i-1})$ rasch abklingen, und daher das Gesetz der großen Zahlen anwenden:

Korollar 3.21 (Gesetz der großen Zahlen für stationäre Markovketten). Ist die Minorisierungsbedingung (3.13) erfüllt, dann existiert eine Konstante $C \in (0, \infty)$, so dass

$$\mathbb{P}\left[\left|\frac{S_n}{n} - \mu[A]\right| \geq \varepsilon\right] \;\leq\; \frac{C}{\varepsilon^2\, n}$$

für alle $\varepsilon > 0$, $n \in \mathbb{N}$ und $A \subseteq S$ gilt.

Die Zufallsvariable S_n/n beschreibt die relative Häufigkeit von Besuchen in der Menge A während der ersten n Schritte der Markovkette. Das Korollar zeigt, dass sich diese relative Häufigkeit für $n \to \infty$ der Wahrscheinlichkeit $\mu[A]$ der Menge A bezüglich der Gleichgewichtsverteilung μ annähert. Dies kann zum näherungsweisen Berechnen der relativen Häufigkeiten für große n, oder aber umgekehrt zum Schätzen der Gleichgewichts-Wahrscheinlichkeiten durch relative Häufigkeiten verwendet werden.

Beweis (Beweis des Korollars). Seien $A \subseteq S$ und $i, n \in \mathbb{Z}_{\geq 0}$. Um die Annahme in Satz 3.12 zu verifizieren, schätzen wir die Kovarianzen der Zufallsvariablen $I_A(Y_i)$ und $I_A(Y_{i+n})$ ab. Nach (3.17) haben Y_i und Y_{i+n} beide die Verteilung μ. Zudem folgt aus der Markov-Eigenschaft, dass

$$\mathbb{P}[Y_i = a \text{ und } Y_{i+n} = b] \;=\; \mu(a) P^n(a, b) \qquad \text{für alle } a, b \in S$$

gilt. Damit erhalten wir

$$\begin{aligned}
\operatorname{Cov}\left[I_A(Y_i), I_A(Y_{i+n})\right] \;&=\; \mathbb{E}\left[I_A(Y_i)\, I_A(Y_{i+n})\right] - \mathbb{E}\left[I_A(Y_i)\right] \mathbb{E}\left[I_A(Y_{i+n})\right] \\
&= \sum_{a \in A} \sum_{b \in A} \mathbb{P}[Y_i = a, Y_{i+n} = b] - \sum_{a \in A} \mathbb{P}[Y_i = a] \sum_{b \in A} \mathbb{P}[Y_{i+n} = b] \\
&= \sum_{a \in A} \sum_{b \in A} \mu(a) P^n(a, b) - \sum_{a \in A} \mu(a) \sum_{b \in A} \mu(b) \\
&= \sum_{a \in A} \mu(a) \sum_{b \in A} \left(P^n(a, b) - \mu(b)\right) \\
&\leq\; 2 \sum_{a \in A} \mu(a)\; d_{TV}(P^n(a, \cdot), \mu) \\
&\leq\; 2 \sum_{a \in A} \mu(a)\, (1 - \delta)^{\lfloor n/r \rfloor} \;\leq\; 2\,(1 - \delta)^{\lfloor n/r \rfloor}.
\end{aligned}$$

Hierbei ist $P^n(a, \cdot)$ die Verteilung der Markovkette mit Start in a nach n Schritten. Die Abschätzung in der vorletzten Zeile gilt nach Definition der Variationsdistanz, und die zentrale Abschätzung in der letzten Zeile folgt nach Satz 3.18 aus der Minorisierungsbedingung (3.13).

Aus der Abschätzung sehen wir, dass die Zufallsvariablen $X_i := I_A(Y_{i-1})$ die Annahme in (3.8) mit $c_n = 2(1 - \delta)^{\lfloor n/r \rfloor}$ erfüllen. Wegen $\sum c_n < \infty$ können wir das Gesetz der großen Zahlen aus Satz 3.12 anwenden. Die Behauptung folgt dann wegen $S_n = \sum_{i=1}^n X_i$ und

$$\mathbb{E}[X_i] \;=\; \mathbb{P}[Y_{i-1} \in A] \;=\; \mu[A] \qquad \text{für alle } i \in \mathbb{N}. \qquad \blacksquare$$

4. Stochastische Simulation und Monte-Carlo-Verfahren

Simulationsverfahren für Stichproben von Wahrscheinlichkeitsverteilungen gehen in der Regel von der Existenz einer Folge von auf dem reellen Intervall $[0, 1]$ gleichverteilten, unabhängigen Zufallszahlen aus, die durch einen Zufallszahlengenerator erzeugt werden. In Wirklichkeit simulieren Zufallszahlengeneratoren natürlich nur auf $\{k\, m^{-1} : k = 0, 1, \ldots, m-1\}$ gleichverteilte Zufallszahlen, wobei m^{-1} die Darstellungsgenauigkeit des Computers ist. Außerdem ist eine Folge von vom Computer erzeugten Pseudozufallszahlen eigentlich gar nicht zufällig, sondern deterministisch. In Abschnitt 4.1 gehen wir kurz auf Verfahren und Probleme bei der Erzeugung von Pseudozufallszahlen mithilfe eines Zufallszahlengenerators ein. Im Abschnitt 4.2 betrachten wir dann verschiedene grundlegenden Verfahren, um Stichproben von allgemeineren Wahrscheinlichkeitsverteilungen aus Stichproben von unabhängigen gleichverteilten Zufallsvariablen zu erzeugen. Schließlich betrachten wir in Abschnitt 4.3 Monte-Carlo-Verfahren, die Gesetze der großen Zahlen verwenden, um Wahrscheinlichkeiten und Erwartungswerte mithilfe von simulierten Stichproben näherungsweise zu berechnen.

Um Simulationsverfahren zu analysieren, benötigen wir noch den Begriff einer auf dem Intervall $[0, 1]$ bzw., äquivalent dazu, auf dem offenen Intervall $(0, 1) \subseteq \mathbb{R}$ gleichverteilten reellwertigen Zufallsvariablen. Die Existenz solcher Zufallsvariablen auf einem geeigneten Wahrscheinlichkeitsraum wird in der Vorlesung Analysis III gezeigt, und hier zunächst vorausgesetzt.

Definition 4.1. Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum.

(i) Eine Abbildung $U : \Omega \to \mathbb{R}$ heißt **reellwertige Zufallsvariable**, falls die Menge $\{U \leq y\} = \{\omega \in \Omega : U(\omega) \leq y\}$ für jedes $y \in \mathbb{R}$ in der σ-Algebra $\mathcal{A}$ enthalten ist.

(ii) Eine reellwertige Zufallsvariable U heißt **gleichverteilt auf dem Intervall** $(0, 1)$, falls

$$\mathbb{P}[U \leq y] = y \qquad \text{für jedes } y \in (0, 1) \text{ gilt.}$$

Im folgenden schreiben wir kurz $U \sim \mathrm{Unif}(0, 1)$ falls U auf $(0, 1)$ gleichverteilt ist.

(iii) Sei I eine beliebige Indexmenge, und seien $U_i : \Omega \to \mathbb{R}$ $(i \in I)$ reellwertige Zufallsvariablen, und $X_i : \Omega \to S$ $(i \in I)$ diskrete Zufallsvariablen mit abzählbarem Zustandsraum S. Dann heißen die Zufallsvariablen U_i und X_i $(i \in I)$ **unabhängig**, falls die Ereignisse $\{U_i \leq y_i\}$ und $\{X_i = a_i\}$, $i \in I$, für alle $y_i \in \mathbb{R}$ und $a_i \in S$ unabhängig sind.

Die Definition ist ein Spezialfall der Definition von Zufallsvariablen mit allgemeinen Zustandsräumen und deren Unabhängigkeit, die in der Vorlesung Einführung in die Wahrscheinlichkeitstheorie gegeben werden.

4.1. Pseudozufallszahlen

Ein *(Pseudo-) Zufallszahlengenerator* ist ein Algorithmus, der eine deterministische Folge von ganzen Zahlen $x_1, x_2, x_3, \ldots$mit Werten zwischen 0 und einem Maximalwert $m-1$ erzeugt, welche durch eine vorgegebene Klasse statistischer Tests nicht von einer Folge von Stichproben unabhängiger, auf $\{0, 1, 2, \ldots, m-1\}$ gleichverteilter Zufallsgrößen unterscheidbar ist. Ein Zufallszahlengenerator erzeugt also nicht wirklich zufällige Zahlen. Die von „guten“ Zufallszahlengeneratoren erzeugten Zahlen haben aber statistische Eigenschaften, die denen von echten Zufallszahlen in vielerlei (aber nicht in jeder) Hinsicht sehr ähnlich sind.

Zufallszahlengeneratoren

Konkret werden Pseudozufallszahlen üblicherweise über eine deterministische Rekurrenzrelation vom Typ

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

aus *Saatwerten* $x_1, x_2, \ldots, x_k$ erzeugt. In vielen Fällen hängt die Funktion f nur von der letzten erzeugten Zufallszahl x_n ab. Beispiele von Pseudozufallszahlengeneratoren sind lineare Kongruenzgeneratoren und Shift-Register-Generatoren.

Lineare Kongruenzgeneratoren

Bei einem linearen Kongruenzgenerator (LCG) ist die Rekurrenzrelation vom Typ

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

Hierbei sind a, c und m geeignet zu wählende positive ganze Zahlen, zum Beispiel:

Generator	m	a	c
ZX81	$2^{16}+1$	75	0
RANDU, IBM 360/370	2^{31}	65539	0
Marsaglia	2^{32}	69069	1
Langlands	2^{48}	142412240584757	11

Ein erstes Problem, dass bei linearen Kongruenzgeneratoren auftreten kann, ist, dass die Folge von Pseudozufallszahlen periodisch mit einer Periode ist, die im Allgemeinen deutlich kleiner als die maximal mögliche Periode m sein kann:

> **Beispiel (LCG mit kleiner Periode).** Wählen wir $m = 63$, $a = 11$ und $c = 0$, dann hat die Folge der vom linearen Kongruenzgenerator erzeugten Pseudozufallszahlen die Periode 6, siehe Abbildung 4.1.

Dieses erste Problem lässt sich leicht mithilfe der folgenden Charakterisierung aller linearen Kongruenzgeneratoren mit der maximal möglichen Periode m umgehen:

Satz 4.2 (Knuth). Die Periode eines LCG ist gleich m genau dann, wenn

(i) c und m teilerfremd sind,

(ii) jeder Primfaktor von m ein Teiler von $a-1$ ist, und

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

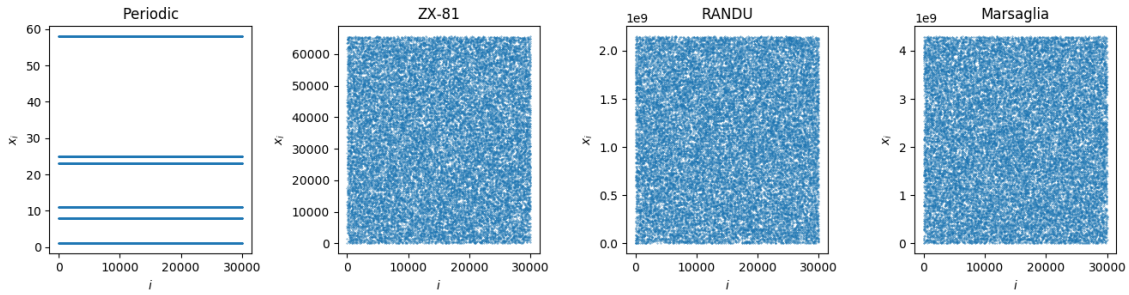


Abbildung 4.1.: Plots der Folgen $x_1, \ldots, x_{30000}$ für den LCG mit Periode 6 aus dem Beispiel, sowie für den ZX81-Generator, RANDU, und den Marsaglia-Generator.

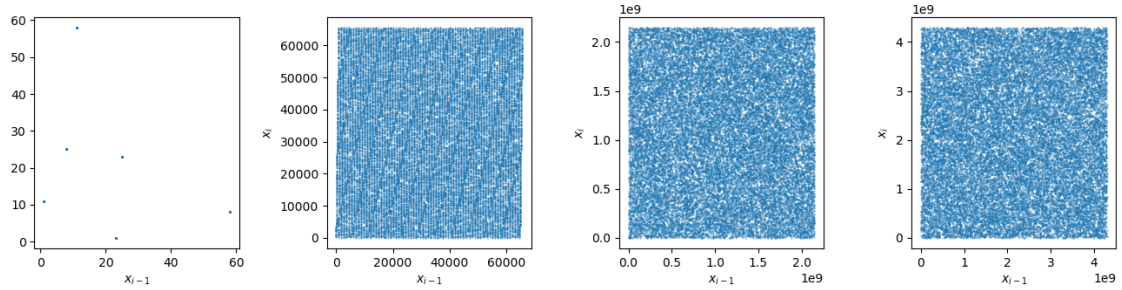


Abbildung 4.2.: Fassen wir Paare (x_i, x_{i+1}) von aufeinanderfolgenden Pseudozufallszahlen als Koordinaten eines zweidimensionalen Pseudozufallsvektors auf, und betrachten die empirische Verteilung dieser Vektoren, so ergibt sich beim ZX81-Generator keine besonders gute Approximation einer zweidimensionalen Gleichverteilung.

Auch wenn ein linearer Kongruenzgenerator die maximal mögliche Periode hat, können weitere Probleme durch versteckte Strukturen und Symmetrien auftreten. Bei einigen einfachen Generatoren werden diese Probleme schon sichtbar, wenn man die Pseudozufallszahlen benutzt, um zwei- oder dreidimensionale Pseudozufallsvektoren zu erzeugen. Dies ist in den Abbildungen 4.2 und 4.3 demonstriert. Der Marsaglia-Generator besteht alle drei Tests; da in Wirklichkeit aber auch dieser deterministische Werte liefert, kann man auch hier einen Test konstruieren, der die Pseudozufallszahlen von echten Zufallszahlen unterscheidet.

Shift-Register-Generatoren

Eine andere Rekurrenzrelation wird zur Erzeugung von Pseudozufallszahlen mit Shift-Register-Generatoren verwendet. Hier interpretiert man eine Zahl $x_n \in \{0,1,\ldots,2^k-1\}$ zunächst als Binärzahl bzw. als Vektor aus $\{0,1\}^k$, und wendet dann eine gegebene Matrix T darauf an, um x_{n+1} zu erhalten:

$$x_{n+1} = Tx_n, \qquad n = 0, 1, 2, \ldots.$$

Kombination von Zufallszahlengeneratoren

Generatoren von Pseudozufallszahlen lassen sich kombinieren, zum Beispiel indem man die von mehreren Zufallszahlengeneratoren erzeugten Folgen von Pseudozufallszahlen aus $\{0,1,\ldots,m-1\}$ modulo m addiert.

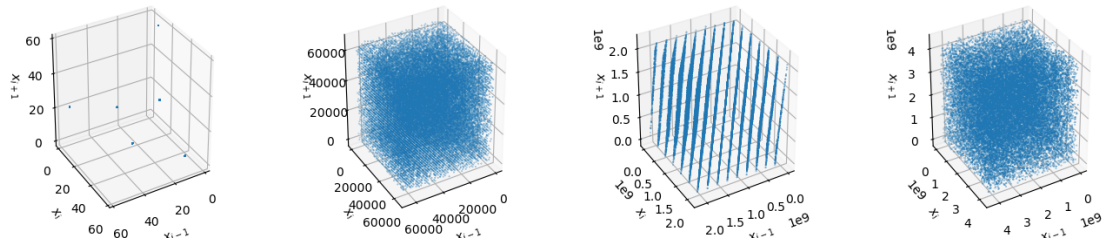


Abbildung 4.3.: Fassen wir analog jeweils drei aufeinanderfolgende Pseudozufallszahlen als dreidimensionale Vektoren auf, dann konzentrieren sich diese beim RANDU-LCG auf 15 Hyperebenen.

Physikalische Zufallszahlengeneratoren

Alternativ werden Zufallszahlen auch mithilfe von physikalischen und insbesondere quantenmechanischen Vorgängen erzeugt, z.B. durch radioaktive Zerfälle, thermisches Rauschen, Atmosphärenrauschen etc. Ein Nachteil ist, dass auf diese Weise nur eine begrenzte Anzahl unabhängiger Stichproben pro Zeiteinheit erzeugt werden kann. Zudem sind die erhaltenen Ergebnisse nicht reproduzierbar. Auch physikalische Zufallsgeneratoren können mit algorithmischen Pseudozufallszahlengeneratoren kombiniert werden.

Statistische Tests für Zufallszahlengeneratoren

Wie wir schon in den Abbildung 4.1, 4.2 und 4.3 gesehen haben, können Schwachstellen von Zufallszahlengeneratoren mithilfe statistischer Tests aufgezeigt werden. Wir wollen kurz auf die dabei zugrundeliegende Argumentation eingehen. Eine von einem Zufallszahlengenerator erzeugte Folge $x_1, x_2, x_3, \ldots$ soll eine Folge von Stichproben von *unabhängigen* Zufallsvariablen $X_1, X_2, X_3, \ldots$ simulieren, die auf der Menge $\{0, 1, \ldots, m-1\}$ *gleichverteilt* ist. Es stellt sich also die Frage, ob die erzeugte Zahlenfolge zu diesem mathematischen Modell passt. Um dies zu testen, leitet man aus den Modellannahmen Folgerungen her, und überprüft ob die erzeugte Zahlenfolge konsistent mit diesen Folgerungen ist.

> **Beispiel (Blocktest).** Sei d eine natürliche Zahl. Sind die Zufallsvariablen X_i unabhängig und gleichverteilt auf $\{0, 1, \ldots, m-1\}$, dann sind auch die Zufallsvektoren $(X_{(k-1)d+1}, X_{(k-1)d+2}, \ldots, X_{kd})$, $k \in \mathbb{N}$, wieder unabhängig und gleichverteilt auf dem Produktraum $\{0, 1, \ldots, m-1\}^d$. Genau dies haben wir in den Abbildungen 4.2 und 4.3 für $d = 2$ bzw. $d = 3$ graphisch getestet. In höheren Dimensionen versagt zwar der graphische Test, aber wir können weiterhin rechnerisch testen, ob sich zum Beispiel die relativen Häufigkeiten von Werten in einem bestimmten Bereich $A \subseteq \{0, 1, \ldots, m-1\}^d$ der simulierten Zufallsvektoren $(x_{(k-1)d+1}, x_{(k-1)d+2}, \ldots, x_{kd})$, $k = 1, \ldots, n$, für große n der Wahrscheinlichkeit von A unter der Gleichverteilung annähern.

Prinzipiell kann jede Folgerung aus den Modellannahmen zur Konzeption eines statistischen Tests verwendet werden. Beispielsweise haben wir in der Einleitung einen Test für 0-1-Zufallsfolgen betrachtet, der auf der Anzahl der Runs basiert. Da jeder Test nur einen bestimmten Aspekt berücksichtigen kann, ist auch für einen Zufallsgenerator, der viele der üblichen Tests besteht, noch nicht garantiert, dass er für eine konkrete Anwendung wirklich geeignet ist. Es kann daher sinnvoll sein, die Ergebnisse einer stochastischen Simulation mit verschiedenen Generatoren zu reproduzieren.

Simulation von Gleichverteilungen

Aus den von einem Pseudozufallszahlengenerator zunächst erzeugten Pseudozufallszahlen mit Werten in der endlichen Menge $\{0, 1, \ldots, m-1\}$ werden anschließend Pseudo-Stichproben von anderen Gleichverteilungen erzeugt.

Zufallszahlen aus $[0, 1)$

Ein Zufallszahlengenerator kann natürlich nicht wirklich reelle Pseudozufallszahlen erzeugen, die die Gleichverteilung auf dem Intervall $[0, 1)$ simulieren, denn dazu würden unendlich viele „zufällige“ Nachkommastellen benötigt. Stattdessen werden üblicherweise (pseudo-)zufällige Zahlen vom Typ

$$u_n = \frac{x_n}{m}, \qquad x_n \in \{0, 1, \ldots, m-1\},$$

erzeugt, wobei m vorgegeben ist (zum Beispiel Darstellungsgenauigkeit des Computers), und x_n eine Folge ganzzahliger Pseudozufallszahlen aus {0, 1, . . . , m - 1} ist.

Zufallspermutationen

Der folgende Algorithmus erzeugt eine (pseudo-)zufällige Permutation aus $\mathcal{S}_n$:

Algorithmus 1: Zufällige Permutation

Input: $n \in \mathbb{N}$
Output: Zufällige Permutation der Länge n

```
for i ← 1 to n do
    x_i ← i
for i ← 1 to n − 1 do
    k ← i + ZufälligeGanzzahl( 0,1,...,n-i );
    Vertausche(x_i, x_k);
return (x_i)_{i=1}^n;
```

Übung. Zeigen Sie, daß Algorithmus 1 tatsächlich eine Stichprobe einer gleichverteilten Zufallspermutation aus $\mathcal{S}_n$ simuliert. *Hinweis: Sei $\tau_{i,j}$ die Transposition von i und j. Zeigen Sie, daß die Abbildung*

$$X(\omega) = \tau_{n-1,\omega_{n-1}} \circ \cdots \circ \tau_{2,\omega_2} \circ \tau_{1,\omega_1}$$

eine Bijektion von $\Omega_n = \{1, 2, \ldots, n\} \times \{2, 3, \ldots, n\} \times \cdots \times \{n-1, n\}$ nach $\mathcal{S}_n$ ist.

4.2. Simulationsverfahren

Wir nehmen nun an, dass wir eine Folge $u_1, u_2, \ldots$ von Stichproben von auf $(0, 1)$ gleichverteilten, unabhängigen Zufallsvariablen $U_1, U_2, \ldots$ gegeben haben. Die in Abschnitt 4.1 beschriebenen Probleme beim Generieren solcher Stichproben werden wir im folgenden ignorieren. Stattdessen wollen wir uns nun überlegen, wie wir aus der Folge (u_n) Stichproben von einer vorgegebenen Wahrscheinlichkeitsverteilung μ auf einer abzählbaren Menge S erzeugen können. Dabei gehen wir in der Regel davon aus, dass wir die Gewichte $\mu(a) = \mu[\{a\}]$ zumindest bis auf eine Normierungskonstante kennen bzw. berechnen können.

Das direkte Simulationsverfahren

Sei $a_1, a_2, \ldots$ eine Abzählung der Elemente von S. Wir betrachten die durch

$$s_k := \sum_{i=1}^{k} \mu(a_i) = \mu[\{a_1, \ldots, a_k\}] \tag{4.1}$$

definierte *kumulative Verteilungsfunktion*. Wir gehen davon aus, dass wir die Werte $\mu(a_i)$ und damit auch s_i für jedes $i \in \mathbb{N}$ berechnen können. Für $n, i \in \mathbb{N}$ setzen wir

$$x_n := a_i \qquad \text{falls } s_{i-1} < u_n \leq s_i.$$

Dann ist x_n eine Stichprobe von der Zufallsvariable

$$X_n := \sum_i a_i \, I_{(s_{i-1}, s_i]}(U_n).$$

Lemma 4.3. *Sind U_n $(n \in \mathbb{N})$ unabhängige Zufallsvariablen mit Verteilung $U_n \sim \mathrm{Unif}(0,1)$, dann sind $X_n (n \in \mathbb{N})$ unabhängige Zufallsvariablen mit Verteilung $X_n \sim \mu$.*

Beweis. Für alle $i \in \mathbb{N}$ gilt

$$\mathbb{P}[X_n = a_i] = \mathbb{P}[s_{i-1} < U_n \leq s_i] = \mathbb{P}[U_n \leq s_i] - \mathbb{P}[U_n \leq s_{i-1}] = s_i - s_{i-1} = \mu(a_i).$$

Also hat X_n die Verteilung μ. Der Nachweis der Unabhängigkeit ist eine Übungsaufgabe. ■

Algorithmus 2: Direkte Simulation einer Stichprobe von einer diskreten Wahrscheinlichkeitsverteilung

Input: Gewichte $(\mu(a_i))_{i \in \mathbb{N}}$
Output: Zufallsstichprobe x von μ

```
1 i ← 1 ;
2 s ← μ(a_1) ;
3 u ← Stichprobe(Unif[0, 1]);
4 while u > s do
5 |   i ← i + 1;
6 |   s ← s + μ(a_i)
7 return x := a_i;
```

Bemerkung (Mittlere Laufzeit). Die mittlere Anzahl von Schritten des Algorithmus ist gleich $\sum i \, \mu(a_i)$.

Nach der Bemerkung ist das direkte Verfahren im Allgemeinen nur dann praktikabel, wenn die Gewichte $\mu(a_i)$ für große i rasch abfallen. In einigen einfachen Spezialfällen kann man jedoch eine explizite Formel zur Berechnung von x_n aus u_n angeben, für deren Auswertung die Schleife in Algorithmus 2 nicht durchlaufen werden muss:

Übung (Simulation von Stichproben einer geometrischen Verteilung). Geben Sie ein direktes Verfahren an, dass in einem Schritt aus einer Stichprobe von der Gleichverteilung auf dem Intervall $(0, 1)$ eine Stichprobe von der geometrischen Verteilung mit Parameter $p \in (0, 1)$ erzeugt.

Das Acceptance-Rejection-Verfahren

Da das direkte Verfahren oft nicht praktikabel ist, benötigen wir Alternativen. Eine häufig verwendete Methode besteht darin, zunächst unabhängige Stichproben von einer "einfacheren" Wahrscheinlichkeitsverteilung ν auf demselben Zustandsraum S zu generieren, und daraus mit einem Verwerfungsverfahren Stichproben von der Zielverteilung μ zu erzeugen. Dazu nehmen wir an, dass wir die Quotienten $\mu(x)/\nu(x)$ der Gewichte unter μ bzw. ν bis auf eine Proportionalitätskonstante kennen, d.h. für $x \in S$ gilt

$$\mu(x) \propto f(x)\nu(x) \tag{4.2}$$

mit einer explizit bekannten Funktion $f : S \to \mathbb{R}$. Beispielsweise können wir $f(x) = \mu(x)/\nu(x)$ setzen, wenn dieses Verhältnis explizit bekannt ist. Wir setzen zudem voraus, dass wir eine obere Schranke c für die Funktion f kennen, d.h.

$$\text{es gibt ein } c \in [1, \infty), \text{ so dass} \qquad f(x) \le c \qquad \text{für alle } x \in S. \tag{4.3}$$

Angenommen, wir können Folgen von Stichproben x_n, u_n $(n \in \mathbb{N})$ von unabhängigen Zufallsvariablen X_n, U_n mit Verteilung ν bzw. $\mathrm{Unif}(0,1)$ erzeugen. Dann können wir daraus Stichproben von der Zielverteilung μ generieren, indem wir die x_n als Vorschlagswerte betrachten, die mit einer Wahrscheinlichkeit proportional zu $f(x_n)$ akzeptiert, und ansonsten verworfen werden. Aufgrund der Annahme (4.3) können die *Akzeptanzwahrscheinlichkeiten* dabei gleich $f(x)/c$ gewählt werden.

Algorithmus 3: Acceptance-Rejection-Verfahren (AR)

Input: $f : S \to [0, \infty)$, $c \in [1, \infty)$ mit (4.3)
Output: Stichprobe x von Wahrscheinlichkeitsverteilung μ mit (4.2)

```
repeat
    x ← Stichprobe(ν) ;
    u ← Stichprobe(Unif(0, 1));
until u ≤ f(x)/c;
return x;
```

Wir wollen den Algorithmus nun analysieren. Seien dazu $X_n \sim \nu$ und $U_n \sim \mathrm{Unif}(0,1)$ $(n \in \mathbb{N})$ unabhängige Zufallsvariablen, die auf einem gemeinsamen Wahrscheinlichkeitsraum definiert sind. Die diskrete Zufallsvariable

$$T(\omega) \;=\; \min\{n \in \mathbb{N} \,:\, U_n(\omega) \le f(X_n(\omega))/c\}$$

beschreibt dann die Anzahl der Durchläufe der Schleife bis erstmals ein Vorschlag X_n akzeptiert wird, und

$$X_T(\omega) \;=\; X_{T(\omega)}(\omega)$$

ist der akzeptierte Wert, der schließlich ausgegeben wird.

Satz 4.4 (Laufzeit und Output des AR-Verfahrens).

(i) T ist *geometrisch verteilt* mit Parameter $p = \sum_{a \in S} \frac{f(a)\nu(a)}{c}$. Insbesondere ist T fast sicher endlich.

(ii) Die Zufallsvariable X_T hat die Verteilung μ.

Der Satz zeigt, dass der Algorithmus tatsächlich eine Stichprobe von der Verteilung μ liefert. Die mittlere Anzahl von Schritten, bis ein Vorschlag akzeptiert wird, beträgt $\mathbb{E}[T] = 1/p$. Ist $f = \mu/\nu$, dann ist $p = 1/c$, also die mittlere Laufzeit gleich c.

Beweis (von Theorem 4.4). (i) Sei $A_n := \{U_n \le f(X_n)/c\}$ das Ereignis, dass der n-te Vorschlag akzeptiert wird. Aus der Unabhängigkeit der Zufallsvariablen $X_1, U_1, X_2, U_2, \ldots$ folgt, daß auch die Ereignisse $A_1, A_2, \ldots$ unabhängig sind. Dies wird in der Vorlesung Einführung in die Wahrscheinlichkeitstheorie allgemein bewiesen, lässt sich im hier betrachteten Spezialfall aber auch direkt überprüfen. Zudem gilt wegen der Unabhängigkeit von X_n und U_n:

$$\begin{aligned}\mathbb{P}[A_n] &= \sum_{a\in S} \mathbb{P}\left[\{X_n = a\} \cap A_n\right] = \sum_{a\in S} \mathbb{P}\left[X_n = a, U_n \le f(a)/c\right] \\ &= \sum_{a\in S} \mathbb{P}\left[X_n = a\right] \cdot \mathbb{P}\left[U_n \le f(a)/c\right] = \sum_{a\in S} \nu(a) f(a)/c = p.\end{aligned}$$

Also ist $T(\omega) = \min\{n \in \mathbb{N} : \omega \in A_n\}$ geometrisch verteilt mit Parameter p.

(ii) Für $a \in S$ gilt

$$\begin{aligned}\mathbb{P}[X_T = a] &= \sum_{n=1}^{\infty} \mathbb{P}[\{X_T = a\} \cap \{T = n\}] \\ &= \sum_{n=1}^{\infty} \mathbb{P}[\{X_n = a\} \cap A_n \cap A_1^C \cap \ldots \cap A_{n-1}^C] \\ &= \sum_{n=1}^{\infty} \mathbb{P}[\{X_n = a, U_n \le f(a)/c\} \cap A_1^C \cap \ldots \cap A_{n-1}^C] \\ &= \sum_{n=1}^{\infty} \nu(a)\, \frac{f(a)}{c}\, (1-p)^{n-1} = \frac{f(a)\nu(a)}{pc}.\end{aligned}$$

Hierbei haben wir im letzten Schritt benutzt, dass die Ereignisse $\{X_n = a\}$, $\{U_n \le f(a)/c\}$, sowie $A_1^C, \ldots, A_{n-1}^C$ unabhängig sind. Da μ die einzige Wahrscheinlichkeitsverteilung ist, deren Massenfunktion proportional zu $f(a)\nu(a)$ ist, folgt $X_T \sim \mu$. ■

Übung (Unabhängigkeit). Sei S eine abzählbare Menge, $g : S \to \mathbb{R}$ eine Funktion, und seien $X_1, X_2, \ldots : \Omega \to S$ sowie $U_1, U_2, \ldots : \Omega \to \mathbb{R}$ unabhängige Zufallsvariablen auf $(\Omega, \mathcal{A}, \mathbb{P})$ mit Verteilungen $X_n \sim \mu$, $U_n \sim \mathrm{Unif}(0,1)$. Zeigen Sie, dass die Ereignisse

$$A_n := \{U_n \le g(X_n)\}, \qquad n \in \mathbb{N},$$

unabhängig sind. *(Hinweis: Zeigen Sie zunächst die Unabhängigkeit von A_1 und A_2).*

> **Beispiel (Simulation von bedingten Verteilungen).** Das Acceptance-Rejection-Verfahren kann prinzipiell verwendet werden, um Stichproben von einer bedingten Verteilung $\mu[A] = \nu[A|B]$ zu simulieren, wobei $B \subseteq S$ ein Ereignis mit $\nu[B] > 0$ ist. In diesem Fall gilt $\mu(x) = f(x)\nu(x)$ mit
>
> $$f(x) = I_B(x)/\nu[B] \le 1/\nu[B] \qquad \text{für alle } x \in S,$$
>
> so dass wir $c = 1/\nu[B]$ wählen können. Das AR-Verfahren erzeugt dann Stichproben von der Verteilung ν und akzeptiert diese mit Wahrscheinlichkeit $I_B(x)$, d.h., Stichproben in B werden stets akzeptiert. Da die mittlere Laufzeit gleich c ist, ist das Verfahren nur dann praktikabel, wenn die Wahrscheinlichkeit von B nicht zu klein ist.

Der Metropolis-Hastings-Algorithmus

Häufig sind direkte oder Acceptance-Rejection-Verfahren zur Simulation von Stichproben einer Wahrscheinlichkeitsverteilung μ nicht praktikabel. Eine Alternative ist die Simulation einer Markovkette (X_n) mit Gleichgewicht μ. Konvergiert die Markovkette ins Gleichgewicht, dann ist die Verteilung von X_n für hinreichend große n ungefähr gleich μ. Eine Stichprobe x_n von X_n ist daher auch eine Näherung einer Stichprobe von μ. Um eine Markovkette mit Gleichgewicht μ zu finden, benutzt man meistens die hinreichende Detailed-Balance-Bedingung (3.12). Die zwei wichtigsten Verfahren, die sich auf diese Weise ergeben, sind der *Metropolis-Hastings-Algorithmus* und der *Gibbs Sampler*.

Wir betrachten zunächst den Metropolis-Hastings-Algorithmus. Sei μ eine beliebige Wahrscheinlichkeitsverteilung auf S mit Gewichten $\mu(x) > 0$ für alle $x \in S$, und sei $Q = (Q(x,y))_{x,y\in S}$ eine stochastische Matrix, für die

$$Q(x,y) = 0 \qquad \Leftrightarrow \qquad Q(y,x) = 0 \tag{4.4}$$

für alle $x, y \in S$ gilt. Eine typische Wahl für Q ist beispielsweise die Übergangsmatrix eines Random Walks bezüglich einer geeigneten Graphenstruktur. Wie können wir die Matrix Q so modifizieren, daß die Detailed-Balance-Bedingung (3.12) bzgl. μ erfüllt ist ? Die Grundidee des Metropolis-Hastings-Algorithmus ist, Übergänge von x nach y mit den Wahrscheinlichkeiten $Q(x,y)$ vorzuschlagen, die Vorschläge aber nur mit einer geeignet gewählten *Akzeptanzwahrscheinlichkeit* $\alpha(x,y)$ zu akzeptieren. Wird ein Vorschlag nicht akzeptiert, dann bleibt die Markovkette an der Stelle x.

Algorithmus 4: Metropolis-Hastings-Algorithmus (MH)

Input: Stochastische Matrix Q, Wahrscheinlichkeitsverteilungen ν, μ
Output: Stichproben $x_0, x_1, \ldots$ von Markovkette mit Startverteilung ν und Gleichgewicht μ

```
1 n ← 0; x_0 ←Stichprobe(ν);
2 repeat
3 |   y_{n+1} ←Stichprobe(Q(x_n, •)) ;
4 |   u_{n+1} ←Stichprobe(Unif(0, 1));
5 |   if u_{n+1} ≤ α(x_n, y_{n+1}) then accept: x_{n+1} ← y_{n+1} else reject: x_{n+1} ← x_n;
6 |   n ← n + 1;
7 until Abbruchkriterium;
```

Die Übergangsmatrix der im Algorithmus simulierten Markovkette ist

$$P(x,y) \;:=\; \begin{cases} \alpha(x,y)\,Q(x,y) & \text{für } y \neq x, \\ 1 - \sum_{y\neq x} \alpha(x,y)\,Q(x,y) & \text{für } y = x. \end{cases} \tag{4.5}$$

Wir müssen noch spezifizieren, wie die Akzeptanzwahrscheinlichkeiten im Algorithmus gewählt werden, damit μ tatsächlich ein Gleichgewicht ist. Die Detailed-Balance-Bedingung lautet in diesem Fall

$$b(x,y) \;=\; b(y,x) \qquad \text{für alle } x, y \in S \text{ mit } x \neq y, \tag{4.6}$$

wobei wir

$$b(x,y) \;:=\; \mu(x)\,\alpha(x,y)\,Q(x,y) \tag{4.7}$$

setzen. Um sicherzustellen, dass die Markovkette nicht häufiger als unbedingt nötig an derselben Stelle stehen bleibt, wollen wir diese Bedingung mit möglichst großen Akzeptanzwahrscheinlichkeiten

$\alpha(x,y) \in [0,1]$, also mit möglichst großen Werten für $b(x,y)$ erfüllen. Wegen $\alpha(x,y) \le 1$ muss nach (4.7)

$$b(x,y) \;\le\; \min\,(\mu(x)Q(x,y),\, \mu(y)Q(y,x)) \tag{4.8}$$

gelten, falls die Symmetriebedingung (4.6) erfüllt ist. Damit ergibt sich als maximale Wahl von b mit(4.6) der Wert auf der rechten Seite von (4.8). Entsprechend erhalten wir die MH-Akzeptanzwahrscheinlichkeiten

$$\alpha(x,y) \;=\; \min\left(1,\, \frac{\mu(y)Q(y,x)}{\mu(x)Q(x,y)}\right) \quad \text{für alle } x,y \in S \text{ mit } Q(x,y) \neq 0. \tag{4.9}$$

Für $x,y \in S$ mit $Q(x,y) = 0$ können wir $\alpha(x,y)$ beliebig wählen, da in diesem Fall $P(x,y) = 0$ unabhängig von der Wahl von α gilt.

Beispiel (Metropolis-Algorithmus). In der ursprünglich von Metropolis, Rosenbluth, Rosenbluth, Teller und Teller betrachteten Version des Algorithmus ist die Vorschlagsmatrix symmetrisch, d.h. es gilt $Q(x,y) = Q(y,x)$ für alle $x,y \in S$. In diesem Fall vereinfacht sich die Formel für die Akzeptanzwahrscheinlichkeiten zu

$$\alpha(x,y) \;=\; \min(1, \mu(y)/\mu(x)). \tag{4.10}$$

Ist beispielsweise der Zustandsraum S ein regulärer Graph, dann liegt es nahe, als Vorschlagsmatrix die symmetrische Übergangsmatrix des Random Walks auf dem Graphen zu wählen.

Definition 4.5. Eine Markovkette (X_n) mit der durch (4.5) und (4.9) definierten Übergangsmatrix heißt **Metropolis-Hastings-Kette** mit Vorschlagsverteilung Q und Gleichgewicht μ.

Die Konvergenz ins Gleichgewicht einer Metropolis-Hastings-Kette folgt unter schwachen Voraussetzungen aus dem Konvergenzsatz für Markovketten:

Übung (Konvergenz ins Gleichgewicht für MH). Sei μ eine Wahrscheinlichkeitsverteilung auf einem endlichen Zustandsraum S mit Gewichten $\mu(x) > 0$, und sei $Q = (Q(x,y))_{x,y\in S}$ eine irreduzible und aperiodische stochastische Matrix auf S, die (4.4) erfüllt. Zeigen Sie, dass μ ein Gleichgewicht ist, und folgern Sie, dass die Verteilung von X_n für eine beliebige Startverteilung ν in Variationsdistanz gegen μ konvergiert.

Der Konvergenzsatz löst aber noch nicht die praktischen Probleme, denn die Konvergenz ins Gleichgewicht kann sehr langsam erfolgen! Wichtig sind daher Abschätzungen der Konvergenzgeschwindigkeit und explizite Fehlerschranken. Diese sind in der Regel stark problemabhängig, und in anwendungsrelevanten Fällen meist nicht leicht herzuleiten.

Übung (Independence Sampler). Sei μ eine Wahrscheinlichkeitsverteilung auf einer endlichen Menge S mit $\mu(x) > 0$ für alle $x \in S$. Der *Independence Sampler* ist ein spezieller MH-Algorithmus, bei dem die Vorschlagsverteilung $Q(x,\,\cdot\,)$ nicht vom Ausgangspunkt x abhängt, d.h.

$$Q(x,y) \;=\; \nu(y)$$

für eine feste Wahrscheinlichkeitsverteilung ν auf S mit $\nu(x) > 0$ für alle $x \in S$.

a) Geben Sie die Übergangsmatrix der entsprechenden Markovkette (X_n) an. Zeigen Sie, dass diese bzgl. des Gleichgewichts μ eine Minorisierungsbedingung mit Konstante $\delta = \min_{x\in S}(\nu(x)/\mu(x))$ erfüllt.

b) Leiten Sie eine Abschätzung für den Variationsabstand zwischen der Verteilung des Independence Samplers nach n Schritten und dem Gleichgewicht μ her.

c) Alternativ kann man in der obigen Situation eine Stichprobe von μ durch ein Acceptance-Rejection-Verfahren mit Vorschlagsverteilung ν erzeugen. Vergleichen Sie die beiden Verfahren.

Übung (Gibbs-Sampler). Sei μ eine Wahrscheinlichkeitsverteilung auf einem endlichen Produktraum $S = S_1 \times \cdots \times S_d$ mit strikt positiven Gewichten $\mu(x_1, \ldots, x_d)$, und sei

$$\mu_i(x_i \mid x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d) \;:=\; \frac{\mu(x_1, \ldots, x_d)}{\sum_{z \in S_i} \mu(x_1, \ldots, x_{i-1}, z, x_{i+1}, \ldots, x_d)}$$

die Massenfunktion der bedingten Verteilung der i-ten Komponente gegeben die Werte x_k $(k \neq i)$ der übrigen Komponenten. Zeigen Sie, dass durch Algorithmus 5 der Übergangsschritt einer Markovkette mit Gleichgewicht μ realisiert wird. *Hinweis: Schreiben Sie die Übergangsmatrix in der Form $P = P_d\, P_{d-1} \cdots P_1$ mit Übergangsmatrizen $P_1, \ldots, P_d$, die die Detailed Balance Bedingung bzgl. μ erfüllen.*

Algorithmus 5: Gibbs Sampler, Übergangsschritt

Input: $x = (x_1, \ldots, x_d) \in S$
Output: $y = (y_1, \ldots y_d) \in S$

```
y ← x;
for i = 1 to d do
    y_i ←Stichprobe(μ_i( • | y_1, ... y_{i-1}, y_{i+1}, ..., y_d)) ;
return y;
```

Simulated Annealing

Für viele Optimierungsprobleme, die in der Praxis auftreten, sind keine Lösungen in einer polynomiellen Anzahl von Schritten mit deterministischen Algorithmen bekannt. Zudem bleiben deterministische Optimierungsalgorithmen häufig in lokalen Minima stecken. Daher greift man in diesen Fällen auch auf heuristische stochastische Verfahren zurück. Angenommen, wir wollen das globale Minimum einer Funktion $U : S \to \mathbb{R}$ auf einem endlichen Zustandsraum S bestimmen. In typischen Anwendungen ist S beispielsweise ein hochdimensionaler Produktraum. Um die Gleichverteilung auf den globalen Minima von U anzunähern, betrachtet man die Wahrscheinlichkeitsverteilungen μ_β, $\beta \in [0, \infty)$, mit Gewichten

$$\mu_\beta(x) \;=\; \mathcal{Z}_\beta^{-1} \exp(-\beta U(x)), \qquad x \in S, \tag{4.11}$$

wobei $\mathcal{Z}_\beta$ eine Normierungskonstante ist. In der statistischen Physik ist μ_β die *Boltzmann-Gibbs-Verteilung* im thermodynamischen Gleichgewicht für die Energiefunktion U bei Temperatur $T = 1/\beta$. Für festes β können wir eine Markovkette mit Gleichgewicht μ_β mithilfe des Metropolis-Hastings-Algorithmus simulieren. Ist die Vorschlagsmatrix $Q(x, y)$ symmetrisch, dann sind die Akzeptanzwahrscheinlichkeiten nach (4.10) durch

$$\alpha_\beta(x, y) \;=\; \exp\left(-\beta(U(y) - U(x))^+\right) \tag{4.12}$$

gegeben. Wichtig ist, dass die rechte Seite nicht von $\mathcal{Z}_\beta$ abhängt, denn die Normierungskonstante ist meistens nicht explizit bekannt. Sei P_β die entsprechende Übergangsmatrix des MH-Algorithmus mit Gleichgewicht μ_β. Die Idee des Simulated Annealing Verfahrens („simuliertes Abkühlen“) besteht nun

darin, eine *zeitlich inhomogene* Markovkette (X_n) mit Übergangskernen $p_n = P_{\beta(n)}$ zu simulieren, wobei $\beta(n)$ eine Folge ist, die gegen unendlich konvergiert. Die Gleichgewichtsverteilung der Übergangskerne p_n nähert sich dann für $n \to \infty$ der Gleichverteilung auf der Menge $\mathcal{M}$ der globalen Minima von U an.

Mithilfe ähnlicher Abschätzungen wie im Beweis der Konvergenzsätze für Markovketten kann man zeigen, daß die Verteilung der inhomogenen Markovkette zur Zeit n gegen die Gleichverteilung auf $\mathcal{M}$ konvergiert, falls $\beta(n)$ nur sehr langsam (logarithmisch) gegen $+\infty$ geht. In praktischen Anwendungen wird das Verfahren aber in der Regel mit einem „schnelleren“ *cooling schedule* $\beta(n)$ verwendet. In diesem Fall findet die Markovkette (X_n) im allgemeinen kein globales Minimum von U, sondern kann, ähnlich wie deterministische Optimierungsverfahren, in lokalen Minima „steckenbleiben“. Das Auffinden eines globalen Minimums ist dann also nicht garantiert – trotzdem erhält man ein oft nützliches *heuristisches* Verfahren.

Übung (Konvergenz von Simulated Annealing).

a) Zeigen Sie, dass die Boltzmann-Gibbs-Verteilung μ_β in (4.11) für $\beta \to \infty$ in Variationsdistanz gegen die Gleichverteilung auf der Menge $\mathcal{M}$ der globalen Minima von U konvergiert.

b) Sei P_β die Übergangsmatrix des Metropolis-Hastings-Algorithmus mit Gleichgewicht μ_β und Vorschlagsverteilung $Q(x, \bullet) = \mathrm{Unif}(S)$. Zeigen Sie, dass P_β für jedes $\beta > 0$ eine Minorisierungsbedingung mit Konstante $\delta_\beta = \exp(-\beta(\max U - \min U))$ bezüglich der Gleichverteilung auf S erfüllt. Folgern Sie, dass es einen cooling schedule $\beta(n)$ gibt, für den die Verteilung von X_n in Variationsdistanz gegen $\mathrm{Unif}(\mathcal{M})$ konvergiert.

4.3. Monte-Carlo-Verfahren

Sei μ eine Wahrscheinlichkeitsverteilung mit Massenfunktion $\mu(x) = \mu[\{x\}]$ auf einer abzählbaren Menge S. Angenommen, wir wollen die Wahrscheinlichkeit

$$p := \mu[B] = \sum_{x \in S} I_B(x)\, \mu(x)$$

eines Ereignisses $B \subseteq S$ beziehungsweise, allgemeiner, den Erwartungswert

$$\theta := E_\mu[f] = \sum_{x \in S} f(x)\, \mu(x)$$

einer reellwertigen Zufallsvariable $f\colon S \to \mathbb{R}$ mit $E_\mu[f^2] < \infty$ (näherungsweise) berechnen, aber die Menge S ist zu groß, um die Summe direkt auszuführen. In einem solchen Fall können wir auf ein Monte-Carlo-Verfahren zurückgreifen. Hierbei simuliert man eine große Anzahl Stichproben $X_1(\omega), \dots, X_n(\omega)$ von unabhängigen Zufallsvariablen mit Verteilung μ (*klassisches Monte-Carlo-Verfahren*), beziehungsweise von einer konvergenten Markovkette mit Gleichgewicht μ (*Markov Chain Monte Carlo*). Nach dem Gesetz der großen Zahlen liefern dann die relativen Häufigkeiten

$$\widehat{p}_n(\omega) := \frac{1}{n} \sum_{i=1}^{n} I_B(X_i(\omega)).$$

bzw. die empirischen Mittelwerte

$$\widehat{\theta}_n(\omega) := \frac{1}{n} \sum_{i=1}^{n} f(X_i(\omega)).$$

Schätzwerte für p bzw. θ, die sich für $n \to \infty$ den gesuchten Werten annähern. Wir wollen nun verschiedene Abschätzungen für den Approximationsfehler $|\widehat{p}_n - p|$ bzw. $|\widehat{\theta}_n - \theta|$ vergleichen. Dazu nehmen wir an, dass die Zufallsvariablen X_i alle die Verteilung μ haben. Nach dem Transformationssatz (Satz 1.15) und der Linearität des Erwartungswerts gilt dann

$$\mathbb{E}[\widehat{\theta}_n] \;=\; \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[f(X_i)] \;=\; \frac{1}{n}\sum_{i=1}^{n} E_\mu[f] \;=\; E_\mu[f] \;=\; \theta,$$

d.h. $\widehat{\theta}_n$ ist ein *erwartungstreuer Schätzer*[1] für θ. Der *mittlere quadratische Fehler* („MSE“ = Mean Squared Error) des Schätzers ist daher durch die Varianz der Zufallsvariable $\widehat{\theta}_n$ gegeben:

$$\mathrm{MSE}\left[\widehat{\theta}_n\right] \;:=\; \mathbb{E}\left[\left|\widehat{\theta}_n - \theta\right|^2\right] \;=\; \mathrm{Var}\left[\widehat{\theta}_n\right].$$

Explizite Abschätzungen für den Approximationsfehler erhalten wir nun mit denselben Methoden wie beim Beweis von Gesetzen der großen Zahlen in Kapitel 3. Sind die Zufallsvariablen X_i beispielsweise unabhängig mit Verteilung μ, dann sind die Zufallsvariablen $f(X_i)$ unkorreliert. In diesem Fall ergibt sich ein mittlerer quadratische Fehler

$$\mathrm{MSE}\left[\widehat{\theta}_n\right] \;=\; \mathrm{Var}\left[\widehat{\theta}_n\right] \;=\; \frac{1}{n}\,\mathrm{Var}_\mu[f]$$

von der Ordnung $O(1/n)$. Der mittlere quadratische Fehler fällt also relativ langsam in n ab. Ein großer Vorteil ist jedoch, dass die Abschätzung völlig *problemunabhängig* ist. Aus diesem Grund sind Monte-Carlo-Verfahren sehr universell einsetzbar. In komplizierten Modellen sind sie oft die einzige praktikable Option um Erwartungswerte näherungsweise zu berechnen. Nach der Čebyšev-Ungleichung erhalten wir zudem für $\varepsilon > 0$ und $n \in \mathbb{N}$ die Fehlerabschätzung

$$\mathbb{P}\left[\left|\widehat{\theta}_n - \theta\right| \geq \varepsilon\right] \;\leq\; \frac{1}{\varepsilon^2}\mathbb{E}\left[\left|\widehat{\theta}_n - \theta\right|^2\right] \;=\; \frac{1}{n\,\varepsilon^2}\,\mathrm{Var}_\mu[f].$$

Insbesondere ist $\widehat{\theta}_n$ eine *konsistente Schätzfolge* für θ, d.h. für jedes $\varepsilon > 0$ gilt

$$\mathbb{P}\left[\left|\widehat{\theta}_n - \theta\right| \geq \varepsilon\right] \longrightarrow 0 \quad \text{für } n \to \infty.$$

Alternativ kann man statt der Čebyšev-Ungleichung auch exponentielle Abschätzungen verwenden, um den Schätzfehler zu kontrollieren. Dies demonstrieren wir im folgenden anhand der Monte-Carlo-Schätzung von Wahrscheinlichkeiten.

Bemerkung (Monte-Carlo-Schätzung von hochdimensionalen Integralen). Auch die Werte von mehrdimensionalen Integralen können mit Monte-Carlo-Verfahren näherungsweise berechnet werden. Dies ist besonders in hohen Dimensionen von Interesse, wo klassische numerische Verfahren in der Regel versagen. Soll beispielsweise der Wert des Integrals

$$\theta \;:=\; \int_{[0,1]^d} f(x)\,\mathrm{d}x \;:=\; \int_0^1 \cdots \int_0^1 f(x_1,\ldots,x_d)\,\mathrm{d}x_1 \ldots \mathrm{d}x_d.$$

[1] Als *Schätzer* bezeichnet man in der Statistik eine Funktion der gegebenen Daten (hier Stichproben von $X_1, \ldots, X_n$), die zum Schätzen eines unbekannten Parameters verwendet wird.

näherungsweise berechnet werden, dann können wir dazu Stichproben $u_1, u_2, \ldots, u_{dn}$ von unabhängigen Zufallsvariablen $U_i \sim \mathrm{Unif}(0,1)$ simulieren. Die d-dimensionalen Zufallsvektoren $X^{(i)} := (U_{di+1}, \ldots, U_{d(i+1)})$, $i = 1, \ldots, n$, sind dann unabhängig und gleichverteilt auf dem Produktraum $(0,1)^d$, siehe Einführung in die Wahrscheinlichkeitstheorie. Daher können wir den Wert θ des Integrals durch den Monte-Carlo-Schätzer

$$\hat{\theta}_n := \frac{1}{n}\sum_{i=1}^{n} f(x^{(i)}) = \frac{1}{n}\sum_{i=1}^{n} f(u_1, u_2, \ldots, u_{dn})$$

approximieren. Ist die Funktion f quadratintegrierbar, dann ergibt sich eine *dimensionsunabhängige* Abschätzung des mittleren quadratischen Fehlers, die nur von der Varianz von f bzgl. der Gleichverteilung auf dem Einheitswürfel $(0,1)^d$ abhängt. Da zum Erzeugen eines Stichprobenvektors $x^{(i)}$ d Zufallszahlen aus $(0,1)$ benötigt werden, beträgt der Aufwand $O(d)$, wenn ein vorgegebener mittlerer quadratischen Fehler für Funktionen mit Varianz kleiner gleich 1 unterschritten werden soll. Klassische numerische Integrationsverfahren haben dagegen in der Regel einen Aufwand, der exponentiell in der Dimension wächst.

Monte Carlo-Schätzung von Wahrscheinlichkeiten

Seien $X_1, X_2, \ldots$ auf $(\Omega, \mathcal{A}, \mathbb{P})$ unabhängige Zufallsvariablen mit Verteilung μ. Wir betrachten nun den klassischen Monte-Carlo-Schätzer $\hat{p}_n$ für die Wahrscheinlichkeit $p = \mu[B] = E_\mu[I_B]$ eines Ereignisses $B \subseteq S$. Obere Schranken für den Schätzfehler können sowohl mithilfe der Čebyšev-Ungleichung als auch über die Bernstein-Ungleichung hergeleitet werden. Wir wollen die entsprechenden Schranken nun vergleichen.

Fehlerkontrolle mittels Čebyšev. Mit der Čebyšev-Ungleichung ergibt sich

$$\mathbb{P}[|\widehat{p}_n - p| \geq \varepsilon] \leq \frac{1}{\varepsilon^2}\,\mathrm{Var}(\widehat{p}_n) = \frac{1}{n\varepsilon^2}\,\mathrm{Var}_\mu(I_B) = \frac{p\,(1-p)}{n\varepsilon^2} \leq \frac{1}{4n\varepsilon^2}.$$

Gilt beispielsweise $n \geq 5\varepsilon^{-2}$, dann erhalten wir

$$\mathbb{P}[p \notin (\widehat{p}_n - \varepsilon, \widehat{p}_n + \varepsilon)] \leq 5\%, \qquad \text{unabhängig von } p,$$

d.h. das zufällige Intervall $(\widehat{p}_n - \varepsilon, \widehat{p}_n + \varepsilon)$ ist ein *95%-Konfidenzintervall* für den gesuchten Wert p.

Fehlerkontrolle mittels Bernstein. Mithilfe der Bernstein-Ungleichung erhalten wir für $\delta > 0$:

$$\mathbb{P}\left[p \notin (\widehat{p}_n - \varepsilon, \widehat{p}_n + \varepsilon)\right] = \mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} I_B(X_i) - p\right| \geq \varepsilon\right] \leq 2e^{-2n\varepsilon^2} \leq \delta, \quad \text{falls } n \geq \frac{\log(2/\delta)}{2\varepsilon^2}.$$

Für kleine δ ist die erhaltene Bedingung an n wesentlich schwächer als eine entsprechende Bedingung, die man durch Anwenden der Čebyšev-Ungleichung erhält.

Für kleine Werte von p ist in der Regel nicht der absolute, sondern der *relative Schätzfehler* $(\widehat{p}_n - p)/p$ von Interesse. Für diesen ergibt sich die Abschätzung

$$\mathbb{P}\left[|\widehat{p}_n - p|/p \geq \varepsilon\right] = \mathbb{P}\left[|\widehat{p}_n - p| \geq \varepsilon p\right] \leq 2e^{-2n\varepsilon^2 p^2} \leq \delta \quad \text{für } n \geq \frac{\log(2/\delta)}{2\varepsilon^2 p^2}.$$

Die benötigte Anzahl von Stichproben für eine (ε, δ)-Approximation von p ist also polynomiell in den Parametern ε, $\log(1/\delta)$ und $1/p$. Mit einer etwas modifizierten Abschätzung kann man die Ordnung $\Omega(1/p^2)$ noch auf $\Omega(1/p)$ verbessern. Trotzdem ist eine direkte Anwendung des einfachen Monte-Carlo-Verfahren für sehr kleine Wahrscheinlichkeiten nicht effektiv.

Varianzreduktion durch Importance Sampling

Häufig ist es sinnvoll, das klassische Monte-Carlo-Verfahren zu modifizieren, indem man zu einer anderen Referenzverteilung übergeht. Beispielsweise wechselt man bei der Monte-Carlo-Berechnung von Wahrscheinlichkeiten seltener Ereignisse zu einer Wahrscheinlichkeitsverteilung, bezüglich der das relevante Ereignis nicht mehr selten ist. Sei also ν eine weitere Wahrscheinlichkeitsverteilung auf S mit Maßenfunktion $\nu(x) = \nu[\{x\}]$. Es gelte $\nu(x) > 0$ für alle $x \in S$. Dann können wir einen unbekannten Erwartungswert $\theta = E_\mu[f]$ auch als Erwartungswert bzgl. ν ausdrücken:

$$\theta \;=\; E_\mu[f] \;=\; \sum_{x\in S} f(x)\,\mu(x) \;=\; \sum_{x\in S} f(x)\,\frac{\mu(x)}{\nu(x)}\nu(x) \;=\; E_\nu[f\,\varrho],$$

wobei

$$\varrho(x) = \frac{\mu(x)}{\nu(x)}$$

der Quotient der beiden Massenfunktionen ist. Ein alternativer Schätzer für θ ist daher durch

$$\widetilde{\theta}_n \;=\; \frac{1}{n}\sum_{i=1}^{n} f(Y_i)\,\varrho(Y_i)$$

gegeben, wobei $Y_1, \dots, Y_n$ unabhängige Zufallsvariablen mit Verteilung ν sind. Auch $\widetilde{\theta}_n$ ist erwartungstreu, denn

$$E_\nu[\widetilde{\theta}_n] \;=\; E_\nu[f\,\varrho] \;=\; \theta.$$

Für die Varianz erhalten wir aufgrund der Unabhängigkeit

$$\mathrm{Var}_\nu[\widetilde{\theta}_n] \;=\; \frac{1}{n}\,\mathrm{Var}_\nu[f\,\varrho] \;=\; \frac{1}{n}\left(\sum_{x\in S} f(x)^2\varrho(x)^2\nu(x) - \theta^2\right).$$

Bei geeigneter Wahl der Referenzverteilung ν kann die Varianz von $\widetilde{\theta}_n$ deutlich kleiner sein als die des Schätzers $\widehat{\theta}_n$.

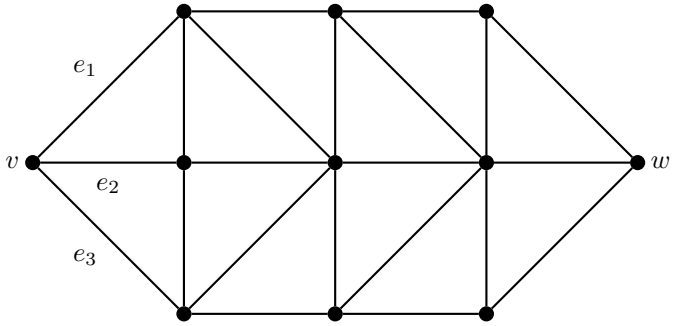
Übung (Varianzminimierung bei Importance Sampling). Zeigen Sie, dass für einen endlichen Zustandsraum S die eindeutige Lösung des Variationsproblems

$$\sum_{x\in S} f(x)^2\varrho(x)^2\nu(x) \;\overset{!}{=}\; \min \qquad \text{unter der Nebenbedingung} \qquad \sum_{x\in S}\nu(x) = 1$$

auf $\mathbb{R}^S$ durch die Massenfunktion der Wahrscheinlichkeitsverteilung ν mit Gewichten

$$\nu(x) \;\propto\; |f(x)|\,\mu(x) \tag{4.13}$$

gegeben ist. Die Referenzverteilung ν mit minimaler Varianz des Importance-Sampling-Schätzers $\widetilde{\theta}_n$ ist also durch (4.13) bestimmt. In Anwendungen ist es meistens nicht möglich, Stichproben von dieser optimalen Referenzverteilung zu erzeugen. Das obige Ergebnis motiviert aber die Faustregel, dass für eine „gute“ Referenzverteilung die Gewichte $\nu(x)$ groß sein sollten, wenn $|f(x)|$ groß ist - daher auch der Name „Importance Sampling“.



Beispiel (Zuverlässigkeit von Netzwerken). Wir beschreiben ein Netzwerk (z.B. Stromleitungen) durch einen endlichen Graphen (V, E). Dabei stehen die Kanten für Verbindungen, die unabhängig voneinander mit einer kleinen Wahrscheinlichkeit ε ausfallen. Seien nun $v, w \in E$ vorgegebene Knoten. Wir wollen die Wahrscheinlichkeit

$$p \;=\; \mathbb{P}[\text{„}v \text{ nicht verbunden mit } w \text{ durch intakte Kanten“}]$$

approximativ berechnen. Sei dazu

$$S = \{0,1\}^E = \{(x_e)_{e\in E} : x_e \in \{0,1\}\}$$

die Menge der Konfigurationen von intakten ($x_e = 0$) bzw. defekten ($x_e = 1$) Kanten, und sei μ die Wahrscheinlichkeitsverteilung auf S mit Massenfunktion

$$\mu(x) \;=\; \varepsilon^{k(x)}(1-\varepsilon)^{|E|-k(x)},$$

wobei $k(x) = \sum_{e\in E} x_e$ die Anzahl der defekten Kanten ist. Dann ist $p = \mu[B]$ die Wahrscheinlichkeit des Ereignisses

$$B \;=\; \{x \in S : v, w \text{ nicht verbunden durch Kanten } e \text{ mit } x_e = 0\}.$$

Der „klassische“ Monte Carlo-Schätzer

$$\widehat{p}_n \;=\; \frac{1}{n}\sum_{i=1}^{n} I_B(X_i), \quad X_i \text{ unabhängig mit Verteilung } \mu,$$

hat Varianz $p(1-p)/n$. Wir wollen den relativen Fehler $\sigma(\hat{p}_n)/p$ beschränken, wobei $\sigma(\hat{p}_n)$ die Standardabweichung bezeichnet. Fordern wir zum Beispiel

$$\sigma(\widehat{p}_n) \;=\; \sqrt{\frac{p(1-p)}{n}} \;\overset{!}{\leq}\; \frac{p}{10},$$

dann benötigen wir eine Stichprobenanzahl

$$n \geq \frac{100\,(1-p)}{p},$$

um diese Bedingung zu erfüllen. Für das in der Abbildung dargestellte (relativ kleine) Netzwerk mit Ausfallwahrscheinlichkeit $\varepsilon = 1\%$ können wir die Größenordnung von p folgendermaßen grob abschätzen:

$$10^{-6} = \mu[\text{„}e_1, e_2, e_3 \text{ versagen“}] \leq p \leq \mu(\text{„mindestens 3 Kanten versagen“}) = \binom{22}{3} \cdot 10^{-6} \approx 1,5 \cdot 10^{-3}.$$

Schon hier wird also eine sehr große Stichprobenanzahl benötigt, und für realistischere Netzwerke ist die Verwendung des klassischen Monte-Carlo-Schätzers nicht mehr praktikabel.

Um die benötigte Stichprobenanzahl zu reduzieren, wenden wir nun Importance Sampling an. Dazu wählen wir als Referenzverteilung die Wahrscheinlichkeitsverteilung ν auf S mit Gewichten

$$\nu(x) = t^{-k(x)}\,(1-t)^{|E|-k(x)}, \qquad k(x) = \sum_{e\in E} x_e,$$

die sich bei Ausfallwahrscheinlichkeit t ergibt. Im Netzwerk aus der Abbildung setzen wir beispielsweise $t := 3/22$, so dass unter ν im Schnitt 3 Kanten defekt sind. Der Ausfall der Verbindung ist dann bezüglich der Verteilung ν kein seltenes Ereignis mehr. Für den Importance-Sampling-Schätzer

$$\widetilde{p}_n \;=\; \frac{1}{n}\sum_{i=1}^{n} I_B(Y_i)\,\frac{\mu(Y_i)}{\nu(Y_i)}, \qquad Y_i \text{ unabhängig mit Verteilung } \nu,$$

erhalten wir

$$\sigma(\widetilde{p}_n)^2 \;=\; \mathrm{Var}(\widetilde{p}_n) \;=\; \frac{1}{n}\left(\sum_{x\in S} I_B(x)^2\frac{\mu(x)^2}{\nu(x)^2}\nu(x) - p^2\right).$$

Im Beispiel aus der Abbildung mit $\varepsilon = 0,01$ und $t = 3/22$ ergibt sich

$$\sigma(\widetilde{p}_n)^2 \;\le\; \frac{1}{n}\sum_{k=3}^{22}\binom{22}{k}\left(\frac{\varepsilon^2}{t}\right)^k\left(\frac{(1-\varepsilon)^2}{1-t}\right)^{22-k} \;\le\; 0,0053\,\frac{p}{n}.$$

Diese obere Schranke für die Varianz ist etwa um den Faktor 200 kleiner als die oben berechnete Varianz des einfachen Monte Carlo-Schätzers. Schon mit einem sehr einfachen Ansatz konnten wir also die Varianz, und damit die benötigte Stichprobenanzahl, deutlich reduzieren.

Der im Beispiel verwendete Ansatz, zu einer „kritischen“ Referenzverteilung überzugehen, bezüglich der die relevanten seltenen Ereignisse gerade eine nicht vernachlässigbare Wahrscheinlichkeit haben, ist typisch für den Einsatz von Importance Sampling auf praktische Problemstellungen. Die Hauptschwierigkeit ist dabei die geschickte Wahl der Referenzverteilung.

Markov Chain Monte Carlo

Häufig ist es nicht möglich oder zu aufwändig, unabhängige Stichproben von der Zielverteilung μ oder einer geeigneten Referenzverteilung zu simulieren. In diesem Fall kann man eine Markovkette (X_n) mit Gleichgewicht μ verwenden, um approximative Stichproben zu erhalten. Nach den Resultaten in Abschnitt 3.4 konvergiert die Verteilung der Verteilung von X_n für $n \to \infty$ unter geeigneten Voraussetzungen in Variationsdistanz gegen μ, sodass wir die Werte $X_n(\omega)$ der Markovkette für $n \ge b$, b hinreichend groß, als approximative Stichproben verwenden können. Diese Stichproben sind jedoch nicht mehr unabhängig, sondern korreliert. Wenn die Kovarianzen schnell abklingen, können wir trotzdem das Gesetz der großen Zahlen anwenden, um Wahrscheinlichkeiten $p = \mu[B]$ und, allgemeiner, Erwartungswerte $\theta = E_\mu[f]$ bezüglich der Gleichgewichtsverteilung durch empirische Mittelwerte der Form

$$\widehat{p}_{n,b} = \frac{1}{n}\sum_{k=b+1}^{b+n} I_B(X_k), \qquad \text{bzw.} \qquad \widehat{\theta}_{n,b} = \frac{1}{n}\sum_{k=b+1}^{b+n} f(X_k)$$

zu approximieren, siehe zum Beispiel Satz 3.12 und das anschließende Korollar.

Die Analyse des Schätzfehlers ist bei Markov Chain Monte Carlo Verfahren im Allgemeinen diffizil. Aus den Resultaten und Beweisen in den Abschnitten 3.4 und 3.3 lassen sich erste Fehlerabschätzungen herleiten. Für viele Anwendungsprobleme sind jedoch keine brauchbaren Fehlerabschätzungen

verfügbar, und man greift auf statistische Methoden zurück, um die Korrelationen abzuschätzen und zu testen, ob die Markovkette sich bereits dem Gleichgewicht angenähert hat. Da die statistischen Tests nicht immer zuverlässig sind, sind die Simulationsergebnisse dann mit entsprechender Vorsicht zu verwenden.

5. Iterationsverfahren

Beispiel. Gesucht: numerische Lösung $x^* \in \left(0, \frac{\pi}{2}\right)$ von $\tan(x) - 2x = 0$.

1. Erster Ansatz:

$$\tan(x) - 2x = 0 \qquad \Leftrightarrow \qquad x = \frac{1}{2}\tan(x) =: \phi(x) \qquad \text{„Fixpunktgleichung".}$$

Iterationsverfahren:

$$x^{(n+1)} = \phi\left(x^{(n)}\right).$$

Konvergiert das Iterationsverfahren?

- falls $x^{(0)} < x^*$ konvergiert $x^{(n)}$ gegen 0.
- falls $x^{(0)} > x^*$ ist $x^{(n)}$ monoton wachsend bis $x^{(n)} > \frac{\pi}{2}$, danach konvergiert $x^{(n)}$ gegen 0.

2. Zweiter Ansatz – alternative Fixpunktiteration:

$$\tan(x) - 2x = 0 \quad \Leftrightarrow \quad x = \arctan(2x) =: \widetilde{\phi}(x).$$

Die Fixpunktiteration

$$x^{(n+1)} = \widetilde{\phi}\left(x^{(n)}\right) = \arctan\left(2x^{(n)}\right)$$

konvergiert gegen x^* für alle Startwerte $x^{(0)} \in (0, \frac{\pi}{2})$.

3. Dritter Ansatz – Newton-Verfahren:
Gesucht ist die Nullstelle von

$$f(y) := \tan(y) - 2y = 0.$$

Idee: Lineare Approximation von f in x (Tangente)

$$f(y) \approx f(x) + (y - x)f'(x).$$

Die Nullstelle der Tangente ergibt

$$y := x - \frac{f(x)}{f'(x)} =: \phi(x).$$

Es gilt

$$f(x^*) = 0 \quad \Leftrightarrow \quad \phi(x^*) = x^*,$$

daher motiviert dies die *Fixpunktiteration*

$$x^{(n+1)} = \phi\left(x^{(n)}\right).$$

- Falls $x^{(0)} \in \left(x^*, \frac{\pi}{2}\right)$ oder $x^{(0)} \in \left(\phi^{-1}\left(\frac{\pi}{2}\right), x^*\right)$, konvergiert die Folge $(x^{(n)})$ sehr schnell gegen x^*.
- Falls $x^{(0)} \leq \phi^{-1}\left(\frac{\pi}{2}\right)$, konvergiert $(x^{(n)})$ nicht gegen x^*.

Das Beispiel zeigt, dass nicht jede Fixpunktiteration konvergiert. Die Konvergenz kann (abhängig vom Startwert) gegen verschiedene Fixpunkte erfolgen. Konvergenz und Konvergenzgeschwindigkeit hängen wesentlich von der Art der Fixpunktdarstellung einer Gleichung ab.

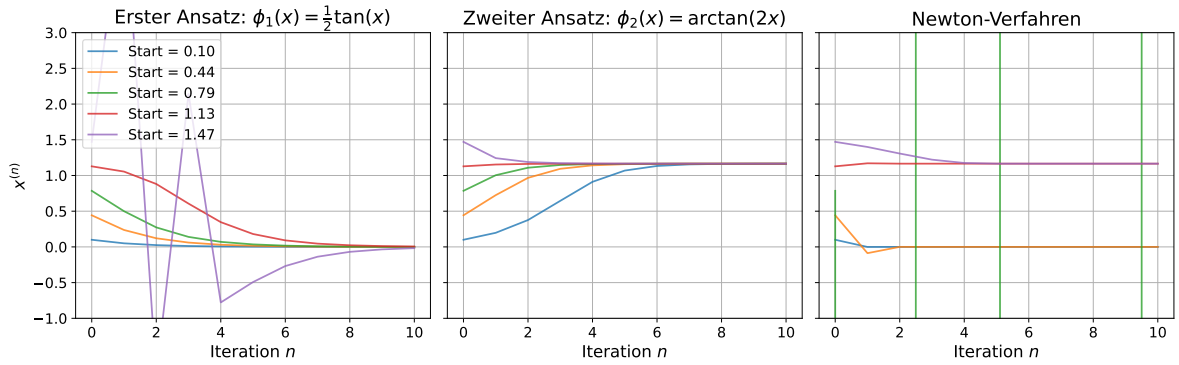


Abbildung 5.1.: Fixpunktiteration für die Gleichung $\tan(x) - 2x = 0$ mit drei verschiedenen Verfahren und fünf Startwerten.

5.1. Konvergenz von Fixpunktiterationen

Banachscher Fixpunktsatz

Allgemeiner Rahmen: (X, d) ein metrischer Raum, $\phi\colon X \to X$.
Fixpunktiteration:

$$x^{(n+1)} = \phi\left(x^{(n)}\right).$$

Satz 5.1 (Banachscher Fixpunktsatz). Ist (X, d) vollständig und ϕ eine *Kontraktion*, d.h., es existiert eine Konstante $L \in (0, 1)$ mit

$$d(\phi(x), \phi(y)) \leq L \cdot d(x, y) \qquad \text{für alle } x, y \in X, \tag{5.1}$$

dann gilt:

1. ϕ hat genau einen Fixpunkt $x^* \in X$.
2. Die Fixpunktiteration $x^{(n+1)} = \phi\left(x^{(n)}\right)$ konvergiert für jeden Startwert $x^{(0)} \in X$ gegen x^*, und für alle $n \in \mathbb{N}$ gilt:

$$d\left(x^{(n)}, x^*\right) \leq L\, d\left(x^{(n-1)}, x^*\right) \qquad \text{„Monotonie”} \tag{5.2}$$

$$d\left(x^{(n)}, x^*\right) \leq \frac{L^n}{1-L}\, d\left(x^{(1)}, x^{(0)}\right) \qquad \text{„A priori-Schranke”} \tag{5.3}$$

$$d\left(x^{(n)}, x^*\right) \leq \frac{L}{1-L}\, d\left(x^{(n)}, x^{(n-1)}\right) \qquad \text{„A posteriori-Schranke”} \tag{5.4}$$

Beweis. i) Aus (5.1) folgt durch Induktion:

$$d\left(x^{(n+1)}, x^{(n)}\right) \leq L^n\, d\left(x^{(1)}, x^{(0)}\right) \qquad \text{für alle } n \geq 0, \tag{5.5}$$

also nach Dreiecksungleichung

$$d\left(x^{(m)}, x^{(n)}\right) \leq \sum_{i=n}^{m-1} d\left(x^{(i+1)}, x^{(i)}\right) \leq \sum_{i=n}^{m-1} L^i \, d\left(x^{(1)}, x^{(0)}\right) \leq \frac{L^n}{1-L} \, d\left(x^{(1)}, x^{(0)}\right) \tag{5.6}$$

für alle $m \geq n \geq 0$. Wegen $L < 1$ ist $x^{(n)}$ eine Cauchy-Folge, also existiert wegen der vorausgesetzten Vollständigkeit

$$x^* := \lim_{n \to \infty} x^{(n)}.$$

Zudem gilt wegen der Stetigkeit von ϕ und Gleichung (5.5):

$$d\left(x^*, \phi\left(x^*\right)\right) = \lim_{n \to \infty} d\left(x^{(n)}, \phi\left(x^{(n)}\right)\right) = \lim_{n \to \infty} d\left(x^{(n)}, x^{(n+1)}\right) = 0,$$

d.h. x^* ist Fixpunkt. Ist y^* ein weiterer Fixpunkt, dann gilt:

$$d\left(x^*, y^*\right) = d\left(\phi\left(x^*\right), \phi\left(y^*\right)\right) \leq L \, d\left(x^*, y^*\right),$$

also $d\left(x^*, y^*\right) = 0$, d.h. $x^* = y^*$ und x^* ist eindeutiger Fixpunkt.

ii) 1. Es gilt:

$$d\left(x^{(n)}, x^*\right) = d\left(\phi\left(x^{(n-1)}\right), \phi\left(x^*\right)\right) \leq L \, d\left(x^{(n-1)}, x^*\right).$$

2.

$$d\left(x^{(n)}, x^*\right) = \lim_{m \to \infty} d\left(x^{(n)}, x^{(m)}\right) \leq \frac{L^n}{1-L} \, d\left(x^{(1)}, x^{(0)}\right).$$

3. Analog zu (5.6) ergibt sich:

$$d\left(x^{(m)}, x^{(n)}\right) \leq \frac{L}{1-L} \, d\left(x^{(n)}, x^{(n-1)}\right) \qquad \text{für alle } m \geq n \geq 0.$$

Die Behauptung folgt für $m \to \infty$. ■

Notation: $D\phi(x)$ bezeichnet die Jacobi-Matrix.

Satz 5.2. Sei $C \subseteq \mathbb{R}^n$ abgeschlossen und konvex und $\phi \colon C \to C$ einmal stetig differenzierbar. Gilt

$$L := \sup_{x \in C} \underbrace{\|D\phi(x)\|_2}_{l^2\text{-Matrixnorm}} = \sup_{x \in C} \sup_{v \in \mathbb{R}^n \setminus \{0\}} \frac{\|D\phi(x)(v)\|_2}{\|v\|_2} < 1,$$

dann hat ϕ genau einen Fixpunkt x^* und $x^{(n)}$ konvergiert gegen x^*.

Beweis. Im Spezialfall $\phi \equiv 0$ ist 0 der einzige Fixpunkt, da $\phi(x) = 0$ für alle $x \in C$ gilt. Die Fixpunktiteration konvergiert bereits nach dem ersten Schritt, da für beliebigen Startwert $x^{(0)} \in C$ stets

$$x^{(n)} = \phi(x^{(n-1)}) = 0 \quad \text{für alle } n \geq 1$$

gilt.

Für $x, y \in C$ gilt $[x, y] \subseteq C$, also mit $\gamma(t) := (1-t)\,x + t\,y$:

$$\begin{aligned}\|\phi(y) - \phi(x)\|_2 &= \|\phi(\gamma(1)) - \phi(\gamma(0))\|_2 \\ &= \left\| \int_0^1 \frac{d}{dt}\phi(\gamma(t))\,dt \right\|_2 \\ &\le \int_0^1 \|D\phi(\gamma(t))\,(\gamma'(t))\|_2\,dt \\ &\le \int_0^1 L\,\|\gamma'(t)\|_2 dt = L\,\|y - x\|_2,\end{aligned}$$

d.h. ϕ ist eine Kontraktion bzgl. $d(x, y) = \|y - x\|_2$. ■

Bemerkung (Lokale Konvergenz). Ist $D \subseteq \mathbb{R}^n$ offen, $\phi \in C^1(D, D)$ mit Fixpunkt x^* und gilt $\|D\phi\,(x^*)\,\|_2 < 1$, dann existiert ein $\varepsilon > 0$ mit:

$$\|D\phi(x)\|_2 \le L < 1 \qquad \text{für alle } x \in \overline{B_\varepsilon\,(x^*)}.$$

Nach dem Korollar konvergiert also $x^{(n)}$ gegen x^* für alle $x^{(0)}$ mit $\|x^{(0)} - x^*\|_2 < \varepsilon$ („lokale Konvergenz").

Beispiele. a) Gleichung von oben:

$$\tan(x) - 2x = 0.$$

1. $\phi(x) = \frac{1}{2}\tan(x)$. Dann ist

$$\phi'(x) = \frac{1}{2}\frac{1}{\cos^2(x)} = \begin{cases} < 1 & \text{für } 0 < x < \frac{\pi}{4}, \\ > 1 & \text{für } \frac{\pi}{4} < x < \frac{\pi}{2}. \end{cases}$$

   - für $t < \frac{\pi}{4}$ ist ϕ eine Kontraktion auf $[0, t]$ und $x^{(n)}$ konvergiert gegen 0.
   - ϕ ist keine Kontraktion auf $\left(\frac{\pi}{4}, \frac{\pi}{2}\right)$.

2. $\widetilde{\phi}(x) = \arctan(2x)$. Dann gilt:

$$\widetilde{\phi}'(x) = \frac{2}{1 + 4x^2} < 1 \qquad \Leftrightarrow \qquad x > \frac{1}{2}.$$

   - für alle $\varepsilon > 0$ ist $\widetilde{\phi}$ eine Kontraktion auf $[1/2 + \varepsilon, \infty)$ und $x^{(n)}$ konvergiert gegen $x^* \in [1/2, \infty)$.

b) Nichtlineares Gleichungssystem: Sei $c \in \mathbb{R}$ eine feste Konstante.

$$\begin{aligned} x_1 &= c\,\cos(x_1) - c\,\sin(x_2), \\ x_2 &= c\,\cos(x_1) - 2c\,\sin(x_2), \\ \phi(x) &= c\begin{pmatrix} \cos(x_1) - \sin(x_2) \\ \cos(x_1) - 2\sin(x_2) \end{pmatrix}, \\ D\phi(x) &= -c\begin{pmatrix} \sin(x_1) & \cos(x_2) \\ \sin(x_1) & 2\cos(x_2) \end{pmatrix}. \end{aligned}$$

Es gilt:

$$\|D\phi(x)\|_F = |c|\sqrt{2\,\sin^2(x_1) + 5\,\cos^2(x_2)} \le \sqrt{7}\,|c|.$$

Damit folgt:

$$\|D\phi(x)\|_2 = \sup_{v\neq 0} \frac{\|D\phi(x)(v)\|_2}{\|v\|_2} \leq \sup_{v\neq 0} \frac{\|D\phi(x)\|_F \cdot \|v\|_2}{\|v\|_2} \leq \|D\phi(x)\|_F \leq \sqrt{7}\,|c|.$$

Somit konvergiert die Fixpunktiteration für $|c| < 1/\sqrt{7}$ mit $L = \sqrt{7}\,|c|$.
Zum Beispiel ergibt sich für $c = 0.35$ mit Startwert $x^{(0)} = (0, 0)$ und $L = \sqrt{7}|c| \approx 0.925$:
Ziel: Für den Approximationsfehler soll gelten:

$$\varepsilon_n = \|x^{(n)} - x^*\|_2 \overset{!}{\leq} 5 \cdot 10^{-5} =: \delta.$$

Wir wissen, dass

$$x^{(1)} = \phi(x^{(0)}) = c \begin{pmatrix} \cos(0) - \sin(0) \\ \cos(0) - 2\sin(0) \end{pmatrix} = \begin{pmatrix} 0.35 \\ 0.35 \end{pmatrix}.$$

- A priori-Abschätzung:

$$\varepsilon_n \leq \frac{L^n}{1-L} \|x^{(1)} - x^{(0)}\|_2 \overset{!}{\leq} \delta.$$

  Die Ungleichung ist erfüllt, falls gilt:

$$n \geq \log_{L^{-1}} \frac{\|x^{(1)} - x^{(0)}\|_2}{(1-L)\,\delta} \approx \log_{1/0.925} \frac{0.495}{0.075 \cdot 5 \cdot 10^{-5}} \approx 121,9.$$

- A posteriori-Abschätzung:

$$\frac{L}{1-L} \|x^{(n)} - x^{(n-1)}\|_2 \overset{!}{\leq} \delta.$$

  Die Ungleichung ist schon für $n = 39$ erfüllt (Nachweis durch Simulation)!

c) Gleichgewichte von Markov-Ketten:
Sei S endlich, und $p(x, y)$, $(x, y \in S)$ eine stochastische Matrix. Ein Gleichgewicht μ von p ist ein Fixpunkt der Abbildung

$$\phi\colon \mathrm{WV}(S) \to \mathrm{WV}(S),$$
$$\phi(\mu) := \mu\, p.$$

Die Menge $\mathrm{WV}(S)$ ist ein Simplex im $\mathbb{R}^{|S|}$, also kompakt. Existiert ein $L \in (0, 1)$ mit

$$d_{TV}(\mu\, p, \nu\, p) \leq L \cdot d_{TV}(\mu, \nu) \qquad \text{für alle } \mu, \nu \in \mathrm{WV}(S), \tag{5.7}$$

dann existiert nach dem Banachschen Fixpunktsatz 5.1 genau ein Gleichgewicht μ von p, und es gilt:

$$d_{TV}(\mu\, p^n, \nu) \leq \frac{L^n}{1-L} \cdot d_{TV}(\mu, \nu) \leq \frac{L^n}{1-L} \qquad \text{für alle } \mu, \nu \in \mathrm{WV}(S).$$

In vielen Fällen gilt Gleichung (5.7) jedoch nur mit $L = 1$!

Konvergenzordnung

Definition 5.3 (Konvergenzordnung). Eine Nullfolge ε_k nicht-negativer reeller Zahlen konvergiert

(i) **linear**, falls es ein $L \in (0,1)$ und ein $n \in \mathbb{N}$ gibt, so dass gilt:

$$\varepsilon_{k+1} \leq L \cdot \varepsilon_k \qquad \text{für alle } k \geq n.$$

(ii) **superlinear**, falls es eine Nullfolge L_k gibt, so dass gilt:

$$\varepsilon_{k+1} \leq L_k \cdot \varepsilon_k \qquad \text{für alle } k \in \mathbb{N}.$$

(iii) **superlinear mit Ordnung** $p > 1$, falls es ein $C \in (0, \infty)$, $n \in \mathbb{N}$ gibt, so dass gilt:

$$\varepsilon_{k+1} \leq C \cdot \varepsilon_k^p \qquad \text{für alle } k \geq n.$$

Beispiele. a) $\varepsilon_k = e^{-\alpha k}$, $\alpha > 0$. Aufgrund von

$$\varepsilon_{k+1} = e^{-\alpha}\, \varepsilon_k$$

konvergiert ε_k linear.

b) $\varepsilon_k = e^{-k^2}$. Aufgrund von

$$\varepsilon_{k+1} = e^{-(k+1)^2} = e^{-2k-1}\, \varepsilon_k$$

konvergiert ε_k superlinear, aber nicht mit Ordnung $p > 1$.

c) $\varepsilon_k = e^{-p^k}$. Aufgrund von

$$\varepsilon_{k+1} = e^{-p \cdot p^k} = (\varepsilon_k)^p$$

konvergiert ε_k superlinear mit Ordnung p.

Der Banachsche Fixpunktsatz garantiert *lineare Konvergenz* (Konvergenzordnung im Sinne von (i) mit Konvergenzfaktor $L \in (0,1)$). Der folgende Satz gibt ein Kriterium für Konvergenz mit höherer Ordnung:

Satz 5.4 (Konvergenz mit höherer Ordnung). Sei $U \subseteq \mathbb{R}^d$ offen und $\phi \in C^p(U, \mathbb{R}^d)$ für ein $p \geq 2$. Ist x^* ein Fixpunkt von ϕ, und gilt

$$(D^k \phi)(x^*) \;=\; 0 \qquad \text{für alle } k = 1, \ldots, p-1,$$

dann existiert eine Konstante $\varepsilon > 0$, so dass die Iterationsfolge $x^{(n)}$ mit $x^{(n+1)} = \phi(x^{(n)})$ für jeden Startwert $x^{(0)} \in B_\varepsilon(x^*)$ (mindestens) mit Ordnung p gegen x^* konvergiert.

Beweis. Sei $\varepsilon > 0$ so, dass $D^p\phi$ auf $B_\varepsilon(x^*)$ beschränkt ist, also existiert $C_1 > 0$ mit $\|D^p\phi(x)\|_{\mathrm{op}} \leq C_1$ für alle $x \in B_\varepsilon(x^*)$. Die Norm $\|D^p\phi(x)\|_{\mathrm{op}}$ bezeichnet die Operatornorm der symmetrischen p-linearen Abbildung $D^p\phi(x)\colon (\mathbb{R}^d)^p \to \mathbb{R}^d$, definiert durch

$$\|D^p\phi(x)\|_{\mathrm{op}} := \sup_{\|h_1\| = \cdots = \|h_p\| = 1} \|D^p\phi(x)[h_1, \ldots, h_p]\|.$$

Dabei ist $\|\cdot\|$ jeweils die euklidische Norm auf $\mathbb{R}^d$. Für $x \in B_\varepsilon(x^*)$ entwickeln wir ϕ in eine Taylor-Reihe um x^*:

$$\phi(x) = \phi(x^*) + \sum_{|\alpha|=1}^{p-1} \frac{D^\alpha \phi(x^*)}{\alpha!}(x-x^*)^\alpha + \sum_{|\alpha|=p} \frac{D^\alpha \phi(\xi)}{\alpha!}(x-x^*)^\alpha$$

mit ξ auf der Verbindungslinie zwischen x und x^*. Wegen $\phi(x^*) = x^*$ und $D^k\phi(x^*) = 0$ für $k = 1, \ldots, p-1$ vereinfacht sich dies zu:

$$\phi(x) = x^* + \sum_{|\alpha|=p} \frac{D^\alpha \phi(\xi)}{\alpha!}(x-x^*)^\alpha.$$

Es gilt

$$\|\phi(x) - x^*\| \le C_2 \|x - x^*\|^p$$

für $C_2 := \frac{C_1}{p!}$ und wir erhalten:

$$\|x_{k+1} - x^*\| = \|\phi(x_k) - x^*\| \le C\|x_k - x^*\|^p,$$

was die Konvergenz der Ordnung p beweist. ■

Ist die Funktion ϕ vorgegeben, kann man in der Regel nicht erwarten, dass die Bedingung aus dem letzten Satz erfüllt ist. Andererseits kann man sich aber zu einem gegebenen Fixpunkt Iterationsabbildungen ϕ konstruieren, so dass die Konvergenz mit höherer Ordnung erfolgt. Ein wichtiges Beispiel liefert das Newton-Verfahren.

5.2. Das Newton-Verfahren

Linearisierung von Gleichungssystemen

Sei $D \subseteq \mathbb{R}^d$ offen, $f : D \to \mathbb{R}^d$ einmal stetig differenzierbar.
Gesucht ist $x^* \in D$ mit:

$$f(x^*) = 0.$$

Letztes Gleichungssystem hat d Gleichungen mit d Unbekannten und ist im Allgemeinen nicht linear.
Ansatz: Linearisierung

$$f(x) \approx \underbrace{f(x^{(0)}) + (Df)(x^{(0)})\,(x - x^{(0)})}_{:=l(x)} + o\left(\|x - x^{(0)}\|\right).$$

- löse $l(x) = 0$,
- Lösung $x^{(1)}$ liefert Näherung für Nullstelle von f,
- Iteration.

Es gilt:

$$l(x) = 0 \quad \Leftrightarrow \quad Df(x^{(0)})(x - x^{(0)}) = -f(x^{(0)}), \qquad \text{lineares Gleichungssystem.}$$

Falls $Df(x^{(0)})$ invertierbar ist, gilt ferner:

$$l(x) = 0 \quad \Leftrightarrow \quad x = x^{(0)} - Df(x^{(0)})^{-1} f(x^{(0)}).$$

Dies motiviert das Iterationsverfahren:

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

Eindimensionaler Fall:

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

$x^{(k+1)}$ ist in diesem Fall der Schnittpunkt der Tangente an f in $x^{(k)}$ mit der x-Achse. Falls f konvex und streng monoton wachsend ist, liegt die Tangente unter dem Graphen von f. In diesem Fall konvergiert die Newton-Iteration monoton für $x^{(0)} > x^*$.

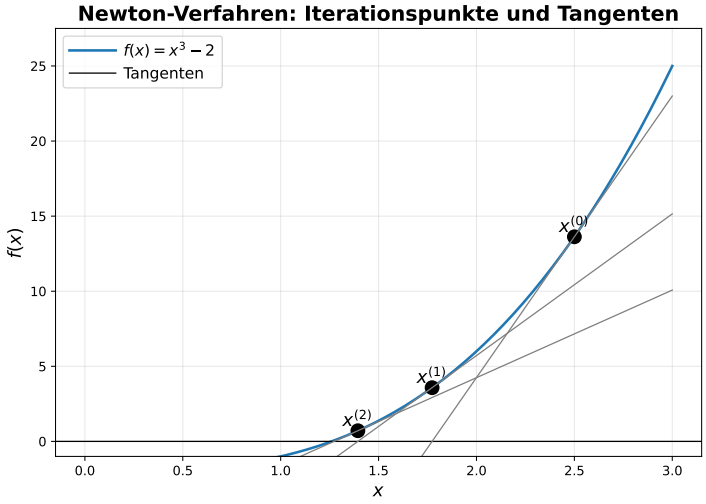
Beispiel (Iterationsverfahren zur Berechnung der n-ten Wurzel).

$$f(x) = x^n - a, \qquad n \in \mathbb{N}\backslash\{1\}, \qquad a \in \mathbb{R}^+.$$

Die Nullstellen von f entsprechen den n-ten Wurzeln von a.
Newton-Verfahren:

$$f'(x) = n\,x^{n-1}.$$

$$x^{(k+1)} = x^{(k)} - \frac{\left(x^{(k)}\right)^n - a}{n\left(x^{(k)}\right)^{n-1}} = \frac{n-1}{n}\,x^{(k)} + \frac{a}{n}\left(x^{(k)}\right)^{1-n}.$$



Algorithmus 6: Newton-Verfahren

Input: $f : D \to \mathbb{R}^d$, Anfangsnäherung $x^{(0)} \in D$, Fehlertoleranz $\varepsilon > 0$.
Output: Iterationsfolge $x^{(k)}$, $k = 0, 1, 2, \ldots$

```
k ← 0;
repeat
    Berechne(Lösung h ∈ ℝ^d des linearen Gleichungssystems Df(x^(k))h = −f(x^(k)));
    x^(k+1) ← x^(k) + h;
    if x^(k+1) ∉ D then
        error („overflow")
until ‖f(x^(k))‖ ≤ ε;
```

Im folgenden wollen wir die Konvergenz des Newton-Verfahrens untersuchen.

Lokal quadratische Konvergenz

Im eindimensionalen Fall zeigt Satz 5.4, dass das Newtonverfahren lokal quadratisch konvergiert. Ist beispielsweise $f \in C^2(\mathbb{R}, \mathbb{R})$ mit $f' > 0$ und $f(x^*) = 0$, dann gilt für die Iterationsabbildung $\phi(x) = x - f(x)/f'(x)$ des Newton-Verfahrens:

$$\phi'(x) \;=\; 1 - \frac{f'(x)}{f'(x)} + \frac{f(x)f''(x)}{f'(x)^2},$$

also $\phi'(x^*) = 0$. Nach Satz 5.4 folgt lokale Konvergenz mit Ordnung 2.

Der folgende Satz liefert eine quantitative Abschätzung im mehrdimensionalen Fall. Sei $\|\cdot\|_V$ eine Norm auf $\mathbb{R}^d$, und $\|\cdot\|_M$ eine verträgliche Matrixnorm.

Satz 5.5 (Lokal quadratische Konvergenz des Newton-Verfahrens). Sei $D \subseteq \mathbb{R}^d$ und $f : D \to \mathbb{R}^d$ einmal stetig differenzierbar, und sei $x^* \in D$ mit $f(x^*) = 0$. Existiert eine nichtleere Umgebung $U = \{x \in \mathbb{R}^d : \|x - x^*\|_V < \varepsilon\}$ von x^* mit $U \subseteq D$, und existiert ein $L \in (0, \infty)$, mit den Eigenschaften

$$\begin{aligned} &Df(x) \text{ ist invertierbar für alle } x \in U, \qquad \text{und} \\ &\|Df(x)^{-1}(Df(y) - Df(x))\|_M \le L\,\|x - y\|_V \qquad \text{für alle } x, y \in U, \end{aligned} \tag{5.8}$$

dann gilt

$$\|x^{(k+1)} - x^*\|_V \le \frac{L}{2}\,\|x^{(k)} - x^*\|_V^2$$

für alle $k \ge 0$ mit $x^{(k)} \in U$. Insbesondere konvergiert $\|x^{(k)} - x^*\|_V$ dann für alle $x^{(0)}$ mit $\|x^{(0)} - x^*\|_V < \min(2/L, \varepsilon)$ quadratisch gegen 0.

Bemerkung. a) Global gilt im allgemeinen keine Konvergenz (siehe Beispiel unten).

b) Bedingung (5.8) ist invariant unter linearen Transformationen

$$f(x) \mapsto \widetilde{f}(x) := A\,f(x), \qquad A \in \mathbb{R}^{d\times d} \text{ nicht-singulär.}$$

c) Falls $Df(x^*)$ invertierbar und Df Lipschitz-stetig in einer Umgebung von x^* ist, gibt es eine geeignete Umgebung U und Konstante L, so dass Bedingung (5.8) erfüllt ist.

Beweis. Für $k \ge 0$ mit $x^{(k)} \in U$ gilt:

$$\begin{aligned} x^{(k+1)} - x^* &= x^{(k)} - x^* - Df(x^{(k)})^{-1}(f(x^{(k)}) - f(x^*)) \\ &= Df(x^{(k)})^{-1}\left(f(x^*) - f(x^{(k)}) - Df(x^{(k)})(x^* - x^{(k)})\right). \end{aligned} \tag{5.9}$$

Sei nun

$$\gamma(t) := (1-t)x^{(k)} + tx^* \in U, \qquad 0 \le t \le 1,$$

die Verbindungsstrecke von $x^{(k)}$ nach x^*. Es gilt:

$$\gamma'(t) = x^* - x^{(k)}.$$

Damit folgt:

$$f(\gamma(1)) - f(\gamma(0)) = f(x^*) - f(x^{(k)}) = \int_0^1 \frac{d}{dt} f(\gamma(t))\, dt = \int_0^1 Df(\gamma(t))(x^* - x^{(k)})\, dt.$$

Außerdem gilt:

$$Df(x^{(k)})(x^* - x^{(k)}) = \int_0^1 Df(x^{(k)})(x^* - x^{(k)})\, dt.$$

Mit Bedingung (5.8) folgt:

$$\begin{aligned}
\|x^{(k+1)} - x^*\|_V &= \left\| \int_0^1 Df(x^{(k)})^{-1} \left[Df(\gamma(t)) - Df(x^{(k)}) \right] (x^* - x^{(k)})\, dt \right\|_V \\
&\leq \int_0^1 \|Df(x^{(k)})^{-1} \left[Df(\gamma(t)) - Df(x^{(k)}) \right] \|_M \, \|x^* - x^{(k)}\|_V \, dt \\
&\leq L \int_0^1 \|\gamma(t) - x^{(k)}\|_V \, dt \, \|x^* - x^{(k)}\|_V \\
&= L \int_0^1 t \, \|x^* - x^{(k)}\|_V \, dt \, \|x^* - x^{(k)}\|_V \\
&= \frac{L}{2} \|x^* - x^{(k)}\|_V^2.
\end{aligned}$$

Angenommen, $x^{(k)} \in U$, $\|x^{(k)} - x^*\|_V \leq \delta$ für ein $\delta \in (0, \frac{2}{L})$. Dann gilt:

$$\|x^{(k+1)} - x^*\|_V \leq \frac{L}{2} \|x^{(k)} - x^*\|_V^2 \leq \frac{\delta L}{2} \|x^{(k)} - x^*\|_V < \|x^{(k)} - x^*\|_V.$$

Somit ist $x^{(k+1)} \in U$ und $\|x^{(k+1)} - x^*\|_V \leq \delta$. Durch Induktion folgt die Behauptung für alle $k \geq 0$ falls $x^{(0)} \in U$ mit $\delta := \|x^{(0)} - x^*\|_V < \frac{2}{L}$. ■

Beispiel. $d = 1$, $f(x) = \arctan(x)$. $x^* = 0$ löst $f(x) = 0$.
Newton-Verfahren:

$$x^{(k+1)} = x^{(k)} - (1 + (x^{(k)})^2) \arctan x^{(k)}.$$

Ist

$$\arctan(|x^{(0)}|) > \frac{2\,|x^{(0)}|}{1 + |x^{(0)}|^2},$$

dann divergiert $x^{(k)}$ bestimmt gegen ∞. Das Newton-Verfahren konvergiert also nicht global.

Andere Iterationsverfahren im eindimensionalen Fall

Sei $f : [a, b] \to \mathbb{R}$ stetig. Gilt $f(a) \cdot f(b) \leq 0$, dann existiert nach dem Zwischenwertsatz ein $x^* \in [a, b]$ mit $f(x^*) = 0$. Das Newton-Verfahren konvergiert für $x^{(0)}$ nahe x^* quadratisch gegen x^*, konvergiert aber im allgemeinen nicht global. Außerdem benötigt es die Existenz und Berechnung der Ableitung f'. Daher werden häufig auch andere Verfahren mit einer kleineren Konvergenzordnung verwendet:

a) Bisektionsverfahren:
Sei $f \in C([a,b])$ mit $f(a) \cdot f(b) \leq 0$ (Vorzeichenwechsel im Intervall $[a,b]$).

Algorithmus 7: Bisektionsverfahren

Input: $f \in C([a,b])$ mit $f(a) \cdot f(b) \leq 0$
Output: Iterationsfolgen $a^{(k)}, b^{(k)},\ k = 0, 1, 2, \ldots$

```
a^(0) ← a;
b^(0) ← b;
for k ← 0, 1, ... do
    x^(k) ← (a^(k) + b^(k)) / 2;
    if f(a^(k)) · f(x^(k)) ≤ 0 then
        a^(k+1) ← a^(k);
        b^(k+1) ← x^(k);
    else
        a^(k+1) ← x^(k);
        b^(k+1) ← b^(k);
```

Das Bisektionsverfahren liefert Intervalle $[a^{(k)}, b^{(k)}]$ mit $|b^{(k)} - a^{(k)}| = 2^{-k}\,|b-a|$, deren Durchschnitt eine Nullstelle x^* von f enthält.
Fehlerabschätzung:

$$|x^{(k)} - x^*| \leq \frac{1}{2}|b^{(k)} - a^{(k)}| \leq 2^{-k-1}\,|b-a|.$$

Das Bisektionsverfahren konvergiert global, die Konvergenz ist im Allgemeinen nicht monoton und die Konvergenzgeschwindigkeit entspricht linearer Konvergenzordnung.

b) Sekantenverfahren:
Das Sekantenverfahren funktioniert wie das Newton-Verfahren, wobei man statt der Tangente die Sekante durch die letzten zwei Iterationswerte nimmt:

$$x^{(0)} = a, \qquad x^{(1)} = b,$$
$$x^{(k+1)} = x^{(k)} - \frac{x^{(k)} - x^{(k-1)}}{f(x^{(k)}) - f(x^{(k-1)})}\, f(x^{(k)}) \qquad \text{für alle } k \geq 1.$$

Der letzte Bruch ist der Kehrwert der Sekantensteigung. Das Sekantenverfahren benötigt keine Berechnung der Ableitung, konvergiert jedoch im allgemeinen nicht global. Wir zeigen die Konvergenzordnung des Sekantenverfahrens für lokale Konvergenz.

Satz 5.6 (Lokale Konvergenz des Sekantenverfahrens). Sei f zweimal stetig differenzierbar auf $[a,b]$ mit einer einfachen Nullstelle $x^* \in (a,b)$, also

$$f(x^*) = 0, \quad f'(x^*) \neq 0.$$

Dann konvergiert das Sekantenverfahren lokal gegen x^* mit der Konvergenzordnung

$$p = \frac{1+\sqrt{5}}{2} \approx 1.618.$$

Beweis. Da $f'(x^*) \neq 0$, ist f in einer Umgebung der Nullstelle von Null verschieden, also in einer Umgebung von x^* injektiv. Damit ist das Sekantenverfahren wohldefiniert.

Den Konvergenzbeweis unterteilen wir in drei Schritte:

1. Rekursionsgleichung für den Fehler:

Für den Fehler $e_k := x^* - x^{(k)}$ ergibt sich aus der Iterationsformel des Sekantenverfahrens

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

die Gleichung

$$e_{k+1} = x^* - x^{(k+1)} = e_k - \frac{e_k - e_{k-1}}{f(x^{(k)}) - f(x^{(k-1)})} f(x^{(k)}).$$

Multiplizieren mit $f(x^{(k)}) - f(x^{(k-1)})$ ergibt

$$e_{k+1}(f(x^{(k)}) - f(x^{(k-1)})) = e_{k-1} f(x^{(k)}) - e_k f(x^{(k-1)}).$$

Division durch $e_k e_{k-1}$ liefert:

$$\frac{e_{k+1}}{e_k e_{k-1}} = \frac{f(x^{(k-1)})}{f(x^{(k)}) - f(x^{(k-1)})} \cdot \frac{1}{e_k} - \frac{f(x^{(k)})}{f(x^{(k)}) - f(x^{(k-1)})} \cdot \frac{1}{e_{k-1}}.$$

Dies entspricht der Differenzenquotientenform

$$\frac{e_{k+1}}{e_k e_{k-1}} = \frac{g(x^{(k)}) - g(x^{(k-1)})}{f(x^{(k)}) - f(x^{(k-1)})}$$

mit der Hilfsfunktion

$$g(x) = -\frac{f(x)}{x - x^*}.$$

2. Anwendung des Mittelwertsatzes:

Die Ableitung von g ist gegeben durch

$$g'(x) = -\frac{f'(x)(x - x^*) - f(x)}{(x - x^*)^2}.$$

Da f zweimal stetig differenzierbar ist, lässt sich g stetig differenzierbar auf $x = x^*$ fortsetzen mit

$$g(x^*) = -f'(x^*), \quad g'(x^*) = -\frac{1}{2} f''(x^*).$$

Mit dem Mittelwertsatz folgt:

$$\frac{g(x^{(k)}) - g(x^{(k-1)})}{f(x^{(k)}) - f(x^{(k-1)})} = \frac{g'(\xi_k)}{f'(\xi_k)}$$

für ein ξ_k zwischen $x^{(k)}$ und $x^{(k-1)}$. Somit ergibt sich:

$$\frac{e_{k+1}}{e_k e_{k-1}} = \frac{g'(\xi_k)}{f'(\xi_k)}.$$

Entwickelt man $g'(\xi_k)$ mit Taylor um x^*, so ergibt sich:

$$\frac{e_{k+1}}{e_k e_{k-1}} = \frac{f(\xi_k) + f'(\xi_k)(x^* - \xi_k)}{(x^* - \xi_k)^2 f'(\xi_k)} = -\frac{1}{2} \cdot \frac{f(\zeta_k)}{f'(\xi_k)}$$

für ein geeignetes ζ_k zwischen x^* und ξ_k (nach Restglieddarstellung). Damit folgt:

$$\frac{e_{k+1}}{e_k e_{k-1}} = -\frac{1}{2} \cdot \frac{f(\zeta_k)}{f'(\xi_k)}.$$

In einer hinreichend kleinen Umgebung von x^* sind $f(\zeta_k)$ und $f'(\xi_k)$ beschränkt, sodass eine Konstante $C > 0$ existiert mit

$$|e_{k+1}| \leq C|e_k| \cdot |e_{k-1}|, \quad \text{für } k \geq 1.$$

3. Nachweis der Konvergenzordnung:

Setze

$$\varepsilon_k := \frac{|e_k|}{|e_{k-1}|^p}, \quad p := \frac{1+\sqrt{5}}{2}.$$

Dann gilt:

$$|e_{k+1}| = \alpha_k |e_k| \cdot |e_{k-1}| = \alpha_k \cdot \varepsilon_k |e_{k-1}|^p \cdot |e_{k-1}| = \alpha_k \varepsilon_k |e_{k-1}|^{p+1},$$

woraus folgt:

$$\varepsilon_{k+1} = \frac{|e_{k+1}|}{|e_k|^p} = \alpha_k \varepsilon_k |e_{k-1}|^{p+1} / |e_k|^p = \alpha_k \varepsilon_k \left(\frac{|e_{k-1}|}{|e_k|^{p/(p+1)}} \right)^{p+1}.$$

Mit der Definition $\gamma_k := \log \varepsilon_k$ folgt daraus eine lineare Rekursion:

$$\gamma_{k+1} = \log \alpha_k - \frac{1}{p} \gamma_k.$$

Diese lineare Rekursion lässt sich explizit lösen:

$$\gamma_{k+1} = \sum_{j=1}^{k} \left(-\frac{1}{p} \right)^{k-j} \log \alpha_j + \left(-\frac{1}{p} \right)^k \gamma_1.$$

Für hinreichend große k ist diese Summe beschränkt, da $\log \alpha_j$ beschränkt ist, also existiert ein $c > 0$ mit

$$|\gamma_k| \leq c \quad \Rightarrow \quad e^{-c} \leq \varepsilon_k \leq e^c.$$

Somit folgt aus

$$\varepsilon_k = \frac{|e_k|}{|e_{k-1}|^p}$$

die asymptotische Beziehung:

$$e^{-c} |e_{k-1}|^p \leq |e_k| \leq e^c |e_{k-1}|^p,$$

also Konvergenzordnung $p = \frac{1+\sqrt{5}}{2}$. ■

c) Regula Falsi-Verfahren:
Annahme:

$$f(a) \cdot f(b) \leq 0.$$

Das Regula Falsi-Verfahern funktioniert wie das Sekantenverfahren, aber anstelle von $x^{(k-1)}$ wird $x^{(l)}$ mit

$$l = \max\{i < k \mid f(x^{(i)}) \cdot f(x^{(k)}) < 0\}$$

verwendet. Das Regula Falsi-Verfahren hat lineare Konvergenzordnung. Es konvergiert global: die Nullstelle befindet sich zwischen $x^{(l)}$ und $x^{(k)}$.

d) Kombinierte Verfahren:
Zum Beispiel Bisektionsverfahren für Anfangsnäherung, anschließend Newton-Verfahren für schnelle genaue Approximation von x^*.

Modifikationen des Newton-Verfahrens im mehrdimensionalen Fall

Nachteile des Newton-Verfahrens sind die Notwendigkeit der Berechnung der Jacobi-Matrix $Df(x^{(k)})$, sowie die fehlende globale Konvergenz. In höheren Dimensionen ist die Berechnung von Df oft aufwändig, und in vielen Fällen nur näherungsweise möglich. Aus diesen Gründen verwendet man unter anderem die folgenden Varianten des Newton-Verfahrens:

a) Vereinfachtes Newton-Verfahren:
Hier wird die Ableitung $Df(x^{(k)})$ nicht in jedem Iterationsschritt, sondern nur gelegentlich berechnet. Dazwischen wird jeweils der zuletzt berechnete Wert $Df(x^{(l)})$, $l < k$, anstelle von $Df(x^{(k)})$ benutzt.

b) Approximation der Jacobi-Matrix:
Die Ableitung $Df(x^{(k)})$ wird in geeigneter Weise mithilfe von Differenzenquotienten approximiert. Ein Beispiel im eindimensionalen Fall ist das oben betrachtete Sekantenverfahren, bei dem $f'(x^{(k)})$ durch den Differenzenquotienten $\frac{f(x^{(k)})-f(x^{(k-1)})}{x^{(k)}-x^{(k-1)}}$ ersetzt wird.

c) Gedämpftes Newton-Verfahren:
Hier setzt man

$$x^{(k+1)} := x^{(k)} + \lambda^{(k)} p^{(k)}.$$

Dabei ist $p^{(k)}$ wie im Newton-Verfahren die Lösung des linearen Gleichungssystems

$$Df(x^{(k)})p^{(k)} = -f(x^{(k)}),$$

der Vektor $p^{(k)}$ wird aber mit einer Schrittweite $\lambda^{(k)} \in (0, 1]$ multipliziert. Die Schrittweite bestimmt man beispielsweise mit einem *Verwerfungsverfahren*: Man setzt zunächst $\lambda^{(k)} := 1$, und berechnet den entsprechenden Wert für $x^{(k+1)}$. Ist $\|f(x^{(k+1)})\|/\|f(x^{(k)})\|$ hinreichend klein, dann wird der erhaltene Wert für $x^{(k+1)}$ akzeptiert - andernfalls halbiert man die Schrittweite, und führt den letzten Schritt erneut durch. Eine Taylor-Entwicklung zeigt, dass eine Akzeptanzbedingung der Form

$$\|f(x^{(k+1)})\|/\|f(x^{(k)})\| < 1 - 2\mu\lambda^{(k)}$$

mit $\mu \in (0, 1)$ jeweils nach endlich vielen Schritten erfüllt ist, falls die Norm von einem Skalarprodukt induziert ist.

d) Homotopieverfahren:
Es werden ein zusätzlicher Parameter $t \in [0,1]$ eingeführt, und Funktionen $f_t(x)$ betrachtet, für die $f_1(x) = f(x)$ gilt, während das Gleichungssystem $f_0(x) = 0$ leicht lösbar ist. Man führt dann ein Newton-Verfahren für die Funktion f_t durch, wobei der Wert von t im Verlauf des Verfahrens schrittweise von 0 auf 1 erhöht wird.

5.3. Differenzengleichungen

Viele praktisch relevante Gleichungssysteme entstehen durch die Diskretisierung von Differentialgleichungen. Im einfachsten Fall verwendet man dazu ein Finite-Differenzen-Verfahren, bei dem Ableitungen durch Differenzenquotienten ersetzt werden.

Approximation von Ableitungen durch Differenzenquotienten

Die Taylor-Entwicklung 3. Ordnung einer Funktion $u \in C^4(\mathbb{R})$ liefert

$$u(x+h) = u(x) + hu'(x) + \frac{h^2}{2}u''(x) + \frac{h^3}{6}u'''(x) + O(h^4), \tag{5.10}$$

$$u(x-h) = u(x) - hu'(x) + \frac{h^2}{2}u''(x) - \frac{h^3}{6}u'''(x) + O(h^4). \tag{5.11}$$

Hieraus ergibt sich zunächst eine $O(h)$-Approximation der ersten Ableitung durch *Vorwärts- und Rückwärts-Differenzenquotienten*:

$$u'(x) = \frac{u(x+h)-u(x)}{h} + O(h) = \frac{u(x)-u(x-h)}{h} + O(h).$$

Subtrahiert man die Gleichungen (5.10) und (5.11), dann ergibt sich die Approximation

$$u'(x) = \frac{u(x+h)-u(x-h)}{2h} + O(h^2)$$

durch den *zentrierten Differenzenquotienten* mit *verbesserter Ordnung* $O(h^2)$.

Entsprechend erhalten wir durch Addition der Gleichungen (5.10) und (5.11) eine $O(h^2)$ Approximation der zweiten Ableitung:

$$\begin{aligned} u''(x) &= \frac{(u(x+h)-u(x)) - (u(x)-u(x-h))}{h^2} + O(h^2) \\ &= \frac{u(x+h) - 2u(x) + u(x-h)}{h^2} + O(h^2). \end{aligned} \tag{5.12}$$

Beispiel. Wir betrachten das lineare Randwertproblem

$$-u''(x) + \lambda(x)u(x) = g(x) \text{ für } x \in (0,1), \qquad u(0) = u(1) = 0,$$

wobei $\lambda, g : [0,1] \to \mathbb{R}$ vorgegebene stetige Funktionen sind. Gesucht ist eine Lösung $u \in C^2((0,1)) \cap C([0,1])$. Um die Lösung näherungsweise zu berechnen, wählen wir $n \in \mathbb{N}$, setzen $h := 1/(n+1)$, und betrachten die Funktionswerte an den Gitterpunkten $x_j = jh$, $j = 0, 1, \ldots, n+1$. Setzen wir $\lambda_j = \lambda(x_j)$ und $g_j = g(x_j)$, dann können wir die Werte $u(x_j)$ durch u_j approximieren, wobei $\vec{u} = (u_1, \ldots, u_n)$ das folgende lineare Gleichungssystem löst:

$$\frac{-u_{j-1} + 2u_j - u_{j+1}}{h^2} + \lambda_j u_j = g_j \qquad (j = 1, \ldots, n),$$

$$u_0 = u_{n+1} = 0,$$

Fehlerbetrachtung: Setzt man voraus, dass die Funktionswerte von u nur bis auf einen Rundungsfehler $\varepsilon > 0$ korrekt dargestellt sind, dann ist der Differenzenquotient zweiter Ordnung nur bis auf einen Rundungsfehler von der Ordnung εh^{-2} korrekt dargestellt. Zusammen mit dem Diskretisierungsfehler der Ordnung $O(h^2)$ ergibt sich ein Gesamtfehler, der sich durch $4\varepsilon h^{-2} + ch^2$ mit einer Konstante c abschätzen lässt. Der Fehler ist minimal für $h_{opt} = (4\varepsilon/c)^{1/4}$, und wächst für kleinere h rasch an ! Es macht also keinen Sinn, zu fein zu diskretisieren. Beispielsweise ist für $\varepsilon = 10^{-9}$ die in diesem Sinn optimale Maschenweite h_{opt} von der Größenordnung 10^{-2}.

LR-Zerlegung für Bandmatrizen

Die Matrix A in dem oben betrachteten linearen Gleichungssystem ist eine **Tridiagonalmatrix**, d.h. neben der Diagonale sind nur noch die Nachbarreihen mit von null verschiedenen Einträgen besetzt. Solche Matrizen treten allgemein bei der Diskretisierung von *gewöhnlichen* Differentialgleichungen zweiter Ordnung auf. Für Tridiagonalmatrizen gibt es eine LR-Zerlegung, bei der L und R eine untere bzw. obere Dreiecksmatrix ist, in der neben der Diagonale nur noch jeweils eine Reihe besetzt ist. Eine entsprechende Zerlegung lässt sich leicht explizit mit Aufwand $O(n)$ berechnen, siehe Vorlesung.

Allgemeiner kann man eine Zerlegung einer Bandmatrix, bei der oberhalb bzw. unterhalb der Diagonalen noch p bzw. q Nachbarreihen nichtverschwindende Einträge haben, mit Aufwand $O(pqn)$ berechnen.

Diskretisierung der Poissongleichung

Wir betrachten nun eine Finite-Differenzen-Diskretisierung eines mehrdimensionalen Randwertproblems. Sei $D \subset \mathbb{R}^d$ offen und beschränkt, und $g \in C(\overline{D})$. Gesucht ist eine Funktion $u \in C^2(D) \cap C(\overline{D})$, die das Dirichletproblem

$$\Delta u \;=\; g \;\text{ auf } D, \qquad u = 0 \;\text{ auf } \partial D,$$

löst. Im Fall $D = (0,1)^2$ diskretisieren wir D bzw. $\overline{D}$ durch das Gitter

$$\begin{aligned} D_h &= \{(ih, jh) : i, j = 1, \ldots, n\} \qquad \text{bzw.} \\ \overline{D_h} &= \{(ih, jh) : i, j = 0, \ldots, n+1\} \end{aligned}$$

mit $h = 1/(n+1)$, $n \in \mathbb{N}$. Approximiert man $\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}$ entsprechend zu (5.12) durch den Differenzenquotienten

$$\Delta_h u(x,y) \;=\; \frac{u(x+h,y) + u(x,y+h) - 4u(x,y) + u(x-h,y) + u(x,y-h)}{h^2},$$

dann ergibt sich die diskretisierte Gleichung

$$\begin{aligned} \Delta_h u(x,y) &= g(x,y) \qquad \text{für } (x,y) \in D_h, \\ u(x,y) &= 0 \qquad \text{für } (x,y) \in \overline{D_h} \setminus D_h. \end{aligned}$$

Ordnet man die Werte von u an den Gitterpunkten lexikographisch in dem Vektor

$$\vec{u} \;=\; (\, u(h,h), u(h,2h), \ldots, u(h,nh), u(2h,h), u(2h,2h), \ldots, u(nh,nh) \,)^T$$

an, dann ergibt sich für $\vec{u}$ ein lineares Gleichungssystem mit einer **Block-Tridiagonalmatrix** A der Dimension n^2, die auf den $n \times n$-Blöcken auf der Diagonalen Tridiagonalmatrizen stehen hat, und auf

den benachbarten $n \times n$-Blöcken negative Einheitsmatrizen, während die anderen $n \times n$-Blöcke nur Nullen enthalten. Die Matrix A ist **dünn besetzt**, da in jeder Zeile nur 5 von 0 verschiedene Einträge stehen - ihre **Bandbreite** $2n + 1$ wächst aber im Gegensatz zum eindimensionalen Fall linear in n, und damit in h^{-1}. Der Aufwand der LR-Zerlegung ist dementsprechend mit $O(n \cdot n \cdot n^2) = O(h^{-4})$ sehr hoch !

Für die Finite-Differenzen-Diskretisierung der Poissongleichung in Dimension $d > 2$ verschlechtert sich die Situation weiter: Hier ist die Dimension des Gleichungssystems von der Ordnung $O(h^{-d})$, die Bandbreite von der Ordnung $O(h^{1-d})$, und der Aufwand einer LR-Zerlegung damit von der Ordnung $O(h^{2-3d})$. Eine Zerlegung ist also für lineare Gleichungssysteme, die bei der Diskretisierung partieller Differentialgleichungen auftreten, im Allgemeinen nicht praktikabel. Aus diesem Grund werden wir im nächsten Abschnitt Iterationsverfahren zur Lösung großer linearer Gleichungssysteme betrachten. Im Gegensatz zur LR-Zerlegung liefern diese in der Regel nur eine Näherung der Lösung.

Eigenwerte, Matrixnormen und Kondition

Eine Matrixnorm $\|\cdot\|_M$ auf $\mathbb{R}^{d\times d}$ heißt **verträglich** mit einer Norm $\|\cdot\|_V$ auf $\mathbb{R}^d$, falls

$$\|Av\|_V \;\le\; \|A\|_M \cdot \|v\|_V \qquad \text{für alle } A \in \mathbb{R}^{d\times d} \text{ und } v \in \mathbb{R}^d.$$

Wichtigstes Beispiel einer verträglichen Matrixnorm ist die durch

$$\|A\|_{V,V} \;:=\; \sup_{v\neq 0} \frac{\|Av\|_V}{\|v\|_V} \;=\; \sup_{\|v\|_V=1} \|Av\|_V$$

definierte, von $\|\cdot\|_V$ induzierte **Operatornorm**. Im Folgenden schreiben wir kurz $\|A\|_V$ statt $\|A\|_{V,V}$. Wir verwenden vor allem die von den ℓ_p-Normen induzierten Matrixnormen

$$\|A\|_p \;:=\; \|A\|_{\ell^p,\ell^p}, \qquad p \in [1,\infty].$$

Es gilt

$$\|A\|_1 \;=\; \max_{j=1,\dots,d} \sum_{i=1}^{d} |a_{i,j}| \qquad \text{und} \qquad \|A\|_\infty \;=\; \max_{i=1,\dots,d} \sum_{j=1}^{d} |a_{i,j}|,$$

siehe Übung. Die ℓ^2-Operatornorm kann man durch die **Frobenius-Norm** abschätzen:

$$d^{-1/2}\,\|A\|_F \le \|A\|_2 \;\le\; \|A\|_F, \qquad \|A\|_F \;:=\; \left(\sum_{i,j=1}^{d} a_{i,j}^2\right)^{1/2}.$$

Diese Abschätzung ist jedoch nicht scharf. Um die ℓ^2-Operatornorm exakt zu berechnen, betrachten wir die Eigenwerte von A bzw. A^TA:

Definition 5.7. Der **Spektralradius** einer Matrix $T \in \mathbb{R}^{d\times d}$ ist

$$\varrho(T) := \max\{|\lambda| \,\big|\, \lambda \in \mathbb{C} \text{ Eigenwert von } T\}.$$

Lemma 5.8 (Spektralradius und Operatornormen).

a) Ist $\|\cdot\|_V$ eine Norm auf $\mathbb{R}^d$, und $\|\cdot\|_M$ eine verträgliche Matrixnorm auf $\mathbb{R}^{d\times d}$, dann gilt

$$\varrho(A) \;\le\; \|A\|_M \qquad \text{für alle } A \in \mathbb{R}^{d\times d}.$$

b) Für die ℓ^2-Operatornorm gilt

$$\|A\|_2 \;=\; \sqrt{\varrho(A^TA)}.$$

Ist A symmetrisch, dann gilt

$$\|A\|_2 = \varrho(A).$$

Beweis. a) Ist λ ein Eigenwert von A und v ein normierter Eigenvektor, dann folgt:

$$|\lambda| = \|\lambda\, v\|_V = \|A\, v\|_V \le \|A\|_M\, \|v\|_V = \|A\|_M.$$

b) Für $v \in \mathbb{R}^d$ gilt

$$\|Av\|_2^2 \;=\; \langle Av, Av\rangle \;=\; \langle v, A^TAv\rangle.$$

Hieraus folgt

$$\|A\|_2^2 \;=\; \sup_{\|v\|=1} \langle v, A^TAv\rangle \;=\; \varrho(A^TA).$$

Ist A symmetrisch, dann sind alle Eigenwerte reell, und die Eigenwerte von $A^TA = A^2$ sind die Quadrate der Eigenwerte von A. Also gilt $\varrho(A^TA) = \varrho(A)^2$. ■

Die ℓ^p-Matrixnormen können wir verwenden, um die Stabilität eines linearen Gleichungssystems $Ax = b$ zu untersuchen. Die ℓ^p-Kondition einer invertierbaren $d \times d$-Matrix A ist gegeben durch

$$\kappa_p(A) \;=\; \|A\|_p\, \|A^{-1}\|_p,$$

siehe Algorithmische Mathematik 1. Insbesondere folgt

$$\kappa_2(A) \;=\; \sqrt{\varrho(A^TA)\varrho((A^{-1})^TA^{-1})} \;=\; \sqrt{\frac{\lambda_{max}(A^TA)}{\lambda_{min}(A^TA)}},$$

wobei $\lambda_{max}(A^TA)$ bzw. $\lambda_{min}(A^TA)$ den größten bzw. kleinsten Eigenwert der symmetrischen, positiv definiten Matrix A^TA bezeichnet. Ist A symmetrisch, dann gilt entsprechend

$$\kappa_2(A) \;=\; \varrho(A)\varrho(A^{-1}) \;=\; \frac{\max\{|\lambda| : \lambda \text{ Eigenwert von } A\}}{\min\{|\lambda| : \lambda \text{ Eigenwert von } A\}}.$$

Beispiel (Newton-Verfahren). Hier ist das lineare Gleichungssystem

$$(Df)(x^{(k)})\left(x^{(k+1)} - x^{(k)}\right) \;=\; -f(x^{(k)})$$

zu lösen. Die ℓ^2-Kondition ist

$$\kappa_2 \;=\; \sqrt{\frac{\lambda_{max}((Df)^T(Df)(x^{(k)}))}{\lambda_{min}((Df)^T(Df)(x^{(k)}))}}.$$

Das Gleichungssystem ist dementsprechend schlecht konditioniert, wenn eine Richtung $v \in \mathbb{R}^d$ mit $\|v\| = 1$ und $(Df)(x^{(k)})v \approx 0$ existiert.

Beispiel (Diskretisierung der Poissongleichung auf $(0,1)^2$). Die Eigenwerte des diskretisierten Laplace-Operators Δ_h auf $D_h \subset (0,1)^2$ kann man explizit berechnen. Es gilt

$$-\Delta_h e_{k,l} \;=\; \lambda_{k,l} e_{k,l} \qquad \text{für } k,l = 1,\dots n$$

mit $e_{k,l}(x,y) = \sin(k\pi x)\sin(l\pi y)$ und

$$\lambda_{k,l} \;=\; \frac{4}{h^2}\left(\sin^2\left(\frac{k\pi h}{2}\right) + \sin^2\left(\frac{l\pi h}{2}\right)\right).$$

Damit folgt

$$\kappa_2(-\Delta_h) \;=\; \frac{\max_{k,l}\lambda_{k,l}}{\min_{k,l}\lambda_{k,l}} \;=\; \frac{\lambda_{n,n}}{\lambda_{1,1}} \;=\; \frac{\cos^2(\frac{\pi}{2}h)}{\sin^2(\frac{\pi}{2}h)} \;=\; O(h^{-2}).$$

Für kleine Maschenweiten h ist die Finite-Differenzen-Diskretisierung der Poissongleichung also schlecht konditioniert.

5.4. Iterationsverfahren für lineare Gleichungssysteme

Gegeben ist ein lineares Gleichungssystem

$$A\,x = b, \qquad \text{mit } b \in \mathbb{R}^d \text{ und invertierbarer Matrix } A \in \mathbb{R}^{d\times d}.$$

Gesucht ist die Lösung $x^* \in \mathbb{R}^d$.
Typisch in Anwendungen ist außerdem, dass d sehr groß ist und A „dünn besetzt" ist, d.h. dass nur wenige Einträge einer Zeile von A ungleich 0 sind.

Beispiel (Poissongleichung). Gesucht ist $u \in C(\overline{D}) \cap C^2(D)$ mit:

$$\begin{aligned} \triangle u(x) := \sum_{i=1}^{n} \frac{\partial^2 u}{\partial x_i}(x) &= g(x) \qquad \text{für alle } x \in D, \\ u(x) &= 0 \qquad \text{für alle } x \in \overline{D}\backslash D. \end{aligned}$$

Diskrete Version:

$$\begin{aligned} &D \subseteq \mathbb{Z}^n \text{ endlich}, \\ &\overline{D} := \{x \in \mathbb{Z}^n \mid \text{es gibt ein } y \in D : \|x-y\|_2 \le 1\}, \\ &g = (g_x)_{x\in D} \in \mathbb{R}^{|D|}. \end{aligned}$$

Gesucht ist in diesem Fall $u = (u_x)_{x\in\overline{D}}$ mit:

$$\begin{aligned} \sum_{i=1}^{n}\left[(u_{x+e_i} - u_x) - (u_x - u_{x-e_i})\right] &= g_x \qquad \text{für alle } x \in D, \\ u_x &= 0 \qquad \text{für alle } x \in \overline{D}\backslash D. \end{aligned}$$

Wir erhalten ein lineares Gleichungssystem

$$A\,u = b$$

mit hoher Dimension $d = |\overline{D}|$ und dünn besetzter Matrix:

$$A_{x,y} \neq 0 \qquad \text{nur falls} \qquad \|x-y\|_2 \le 1.$$

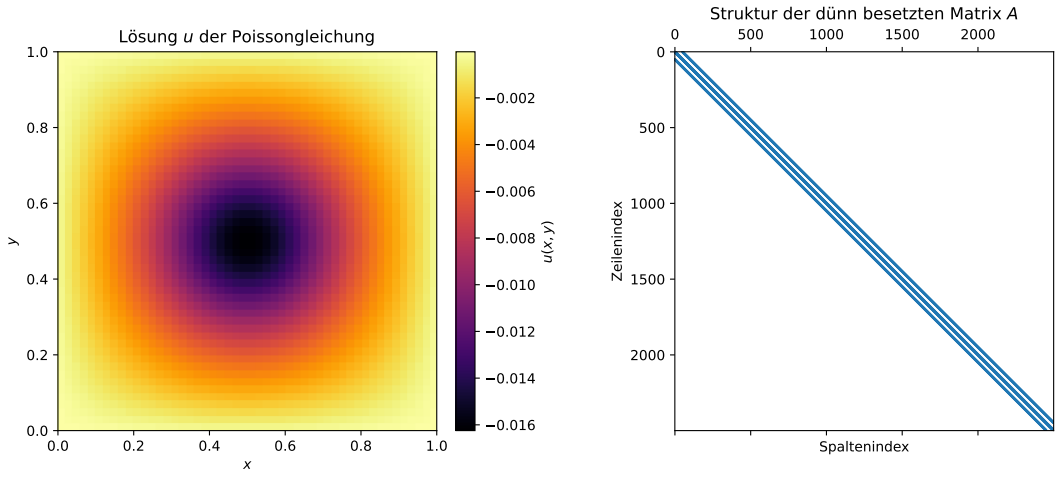


Abbildung 5.2.: Lösung der Poisson Gleichung auf einem Gitter $D \subset \mathbb{Z}^2$ mit Randbedingung $u = 0$ auf $\overline{D} \setminus D$ und Darstellung der Matrix A mit nichtnull Einträgen.

In Anwendungen wie dieser sind die klassischen Eliminationsverfahren meist zu aufwändig ($O(d^3)$), stattdessen verwendet man Fixpunktiterationen.

Allgemeines Iterationsverfahren:

$$A = P - N, \qquad P, N \in \mathbb{R}^{d \times d},$$

wobei P eine invertierbare Matrix, deren Inverses „leicht" berechenbar ist. P wird ***Präkonditionierer*** genannt. Es gilt:

$$\begin{aligned} A\,x = b \quad &\Leftrightarrow \quad P\,x = N\,x + b \quad \Leftrightarrow \quad x = T\,x + f =: \phi(x), \qquad \text{wobei} \\ T := P^{-1}\,N = P^{-1}\,(P - A) = I - P^{-1}\,A, \quad &f := P^{-1}\,b. \end{aligned}$$

Lineare Fixpunktiteration:

$$x^{(k+1)} = T\,x^{(k)} + f.$$

Residuendarstellung:

$$x^{(k+1)} = x^{(k)} - P^{-1}\,(A\,x^{(k)} - b).$$

Der Ausdruck $(A\,x^{(k)} - b)$ wird als ***Residuum von*** $x^{(k)}$ bezeichnet.

Bemerkung. Die Berechnung von $x^{(k+1)}$ erfordert Inversion von P!

Im folgenden wollen wir untersuchen, wann die Fixpunktiteration konvergiert.

Satz 5.9. 1. Die folgenden Aussagen sind äquivalent:

(i) Die Iterationsfolge $x^{(k)}$ konvergiert für jeden Startwert $x^{(0)} \in \mathbb{R}^d$ gegen die Lösung x^* von $A\,x = b$.

(ii) T^k konvergiert gegen 0 (d.h. $(T^k)_{i,j}$ konvergiert gegen 0 für alle i, j).

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

2. *Hinreichende Bedingung:*
Ist $\|\cdot\|_V$ eine Norm auf $\mathbb{R}^d$ und $\|\cdot\|_M$ eine verträgliche Matrixnorm mit $\|T\|_M < 1$, dann gilt:

$$\|x^{(k+1)} - x^*\|_V \leq \|T\|_M \, \|x^{(k)} - x^*\|_V \qquad \text{für alle } k \in \mathbb{N},$$

d.h. $x^{(k)}$ konvergiert bzgl. $\|\cdot\|_V$ monoton gegen x^*.

Beweis. i) • $(i) \Leftrightarrow (ii)$: Wegen

$$A\,x^* = b \qquad \Leftrightarrow \qquad x^* = \phi(x^*)$$

gilt

$$x^{(k+1)} - x^* = \phi(x^{(k)}) - \phi(x^*) = T(x^{(k)} - x^*),$$

für alle $k \in \mathbb{N}$, und somit:

$$x^{(k)} - x^* = T^k(x^{(0)} - x^*).$$

$x^{(k)}$ konvergiert also genau dann für alle $x_0 \in \mathbb{R}^d$ gegen x^*, wenn T^k gegen 0 konvergiert.

- $(ii) \Rightarrow (iii)$: Angenommen, $\varrho(T) \geq 1$. Dann gibt es einen Eigenwert $\lambda \in \mathbb{C}$ mit $|\lambda| \geq 1$ mit Eigenvektor $v \in \mathbb{C}^d \backslash \{0\}$ von T:

$$T\,v = \lambda\,v.$$

Daher konvergiert $T^k\,v = \lambda^k\,v$ nicht gegen 0. Somit konvergiert T^k nicht gegen 0 im Widerspruch zu (ii).

- $(iii) \Rightarrow (ii)$: Wir beweisen diese Richtung für den symmetrischen Fall ($T = T^T$). Da T symmetrisch ist, gibt es eine orthogonale Matrix $O \in \mathbb{R}^{d\times d}$, so dass gilt:

$$T = O\,D\,O^T, \qquad \text{wobei} \qquad D = \operatorname{diag}(\lambda_1, \ldots, \lambda_d) = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_d \end{pmatrix}$$

die Diagonalmatrix aus Eigenwerten von T ist. Wegen $O^T = O^{-1}$ gilt:

$$T^k = \left(O\,D\,O^{-1}\right)^k = O\,D^k\,O^T.$$

Aus $\varrho(T) < 1$ folgt nun für alle Eigenwerte $|\lambda_i| < 1$. Also konvergiert D^k und somit auch T^k gegen 0.
Den allgemeinen Fall zeigt man via Jordanscher Normalform, siehe z.B. „Hämmerlin/Hoffmann: *Numerische Mathematik*, Satz 8.1".

ii) Es gilt:

$$\|\phi(x) - \phi(y)\|_V = \|T\,x - T\,y\|_V \leq \|T\|_M \, \|x - y\|_V.$$

Wegen $\|T\|_M < 1$ ist ϕ eine Kontraktion bzgl. $\|\cdot\|_V$. Mit dem Banachschen Fixpunktsatz folgt die Behauptung. ■

Klassische Iterationsverfahren

Keine Präkonditionierung

$$P = I.$$

Dann lautet die Residuendarstellung:

$$x^{(k+1)} = x^{(k)} - (A\,x^{(k)} - b), \qquad \text{und} \qquad T = I - A.$$

Bezüglich Konvergenz gilt:

λ ist genau dann Eigenwert von A, wenn $1 - \lambda$ Eigenwert von T ist.

Also ist der Spektralradius

$$\varrho(T) = \max\{|1 - \lambda| \,\big|\, \lambda \text{ Eigenwert von } A\}.$$

Beispiel. Ist A symmetrisch und positiv definit („s.p.d."), dann sind die Eigenwerte von A reell und positiv. Also gilt

$$\varrho(T) = \max\{|1 - \lambda| \,\big|\, \lambda \text{ Eigenwert von } A\} < 1 \qquad \Leftrightarrow \qquad \lambda_{\max}(A) < 2.$$

Die Bedingung an die Eigenwerte ist bei diesem Verfahren zu restriktiv!

Jacobi-Verfahren / Gesamtschrittverfahren

Es gilt:

$$\begin{aligned} A\,x = b \qquad &\Leftrightarrow \qquad \sum_{j=1}^{d} a_{ij}\,x_j = b_i \qquad i = 1, \ldots, n \\ &\Leftrightarrow \qquad a_{ii}\,x_i = b_i - \sum_{j \neq i} a_{ij}\,x_j. \end{aligned}$$

Algorithmus 8: Jacobi-Verfahren

Input: Matrix $A \in \mathbb{R}^{d\times d}$ mit $a_{ii} \neq 0$ für alle $i \leq d$, $b \in \mathbb{R}^d$, Startwert $x^{(0)} \in \mathbb{R}^d$
Output: Iterationsfolge $x^{(k)} \in \mathbb{R}^d$

for $k \leftarrow 0, 1, \ldots$ **do**
  for $i \leftarrow 1$ **to** d **do**

$$x_i^{(k+1)} \leftarrow \frac{1}{a_{ii}} \left(b_i - \sum_{j \neq i} a_{ij}\,x_j^{(k)}\right) \tag{5.13}$$

Einordnung in allgemeinen Rahmen:
Die Matrix A wird in drei Matrizen L, D und R zerlegt:

$$A = -L + D - R, \qquad \text{mit}$$

$$L = -\begin{pmatrix} 0 & \cdots & \cdots & 0 \\ a_{21} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ a_{d1} & \cdots & a_{d,d-1} & 0 \end{pmatrix}, \quad D = \begin{pmatrix} a_{11} & \cdots & \cdots & 0 \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & a_{dd} \end{pmatrix}, \quad R = -\begin{pmatrix} 0 & a_{12} & \cdots & a_{1d} \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & a_{d-1,d} \\ 0 & \cdots & \cdots & 0 \end{pmatrix}.$$

Für das lineare Gleichungssystem gilt mit dieser Zerlegung:

$$A\,x = b \qquad \Leftrightarrow \qquad D\,x = (L+R)\,x + b \qquad \Leftrightarrow \qquad x = D^{-1}(L+R)\,x + D^{-1}\,b.$$

Als Iteration setzen wir:

$$x^{(k+1)} = D^{-1}\,(L+R)\,x^{(k)} + D^{-1}\,b.$$

Also gilt:

$$P = D \qquad \text{und} \qquad T = D^{-1}\,(L+R).$$

Gauß-Seidel-Verfahren / Einzelschrittverfahren

Idee: Verwende bereits berechnete Komponenten von $x^{(k+1)}$ sofort.

Algorithmus 9: Gauß-Seidel-Verfahren

Input: Matrix $A \in \mathbb{R}^{d\times d}$ mit $a_{ii} \neq 0$ für alle $i \leq d$, $b \in \mathbb{R}^d$, Startwert $x^{(0)} \in \mathbb{R}^d$
Output: Iterationsfolge $x^{(k)} \in \mathbb{R}^d$

```
1 for k ← 0, 1, ... do
2     for i ← 1 to d do
3         x_i^(k+1) ← 1/a_ii (b_i − Σ_{j<i} a_ij x_j^(k+1) − Σ_{j>i} a_ij x_j^(k))   (5.14)
```

In Matrixform:

$$D\,x^{(k+1)} = b + L\,x^{(k+1)} + R\,x^{(k)} \qquad \Leftrightarrow \qquad (D-L)\,x^{(k+1)} = R\,x^{(k)} + b.$$

Einordnung in allgemeinen Rahmen:

$$P = D - L, \qquad T = (D-L)^{-1}\,R.$$

Satz 5.10. Ist *A strikt diagonaldominant* . d.h.

$$|a_{ii}| > \sum_{j\neq i} |a_{ij}|, \qquad i = 1, \ldots, d,$$

dann konvergieren das Jacobi- und das Gauß-Seidel-Verfahren monoton bzgl. der Maximumsnorm $\|\cdot\|_\infty$ auf $\mathbb{R}^d$.

Beweis. Wir zeigen:

$$\|T\|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{n} |t_{ij}| < 1.$$

Da $\|\cdot\|_\infty$ die von der Maximumsnorm auf Vektoren induzierte Matrixnorm ist, folgt die Behauptung dann aus Satz 5.9.

1. Jacobi-Verfahren:

$$T = D^{-1}(L+R), \qquad t_{ij} = \begin{cases} -\frac{a_{ij}}{a_{ii}} & \text{für } j \neq i, \\ 0 & \text{für } j = i. \end{cases}$$

Damit folgt:

$$\|T\|_\infty \leq \max_{1\leq i\leq n} \frac{1}{|a_{ii}|} \sum_{j\neq i} |a_{ij}| =: c$$

Weil A strikt diagonaldominant ist, folgt $c < 1$.

2. Gauß-Seidel-Verfahren:

$$T = (D-L)^{-1} R.$$

Sei c wie oben definiert, wir behaupten:

$$\|T\|_\infty \leq c < 1.$$

Dazu ist zu zeigen:

$$\|Tx\|_\infty \leq c \cdot \|x\|_\infty \qquad \text{für alle } x \in \mathbb{R}^n.$$

Sei $y := Tx$. Nach Algorithmus 9 ist

$$y_i = \frac{1}{a_{ii}} \Big(-\sum_{j<i} a_{ij} y_j - \sum_{j>i} a_{ij} x_j \Big).$$

Wir zeigen durch vollständige Induktion nach i die Aussage:

$$|y_i| \leq c \cdot \|x\|_\infty.$$

Für $i = 1$ gilt:

$$|y_1| \leq \frac{1}{|a_{11}|} \sum_{j\neq 1} |a_{1j}| \cdot |x_j| \leq c \cdot \|x\|_\infty.$$

Mit der Induktionsvoraussetzung folgt:

$$\begin{aligned} |y_i| &\leq \frac{1}{|a_{ii}|} \Big(-\sum_{j=1}^{i-1} |a_{ij}|\,|y_j| + \sum_{j=i+1}^{d} |a_{ij}|\,|x_j| \Big) \\ &\leq \frac{1}{|a_{ii}|} \Big(\sum_{j=1}^{i-1} |a_{ij}|\, c\, \|x\|_\infty + \sum_{j=i+1}^{d} |a_{ij}|\, \|x\|_\infty \Big) \\ &\leq \frac{1}{|a_{ii}|} \sum_{j\neq i} |a_{ij}|\, \|x\|_\infty \leq c\, \|x\|_\infty \qquad \text{für alle } i = 1,\dots,d. \end{aligned}$$

■

Somit gilt:

$$\|Tx\|_\infty = \|y\|_\infty \leq c\, \|x\|_\infty.$$

Man kann nicht allgemein sagen, welches Verfahren „besser" ist. Für bestimmte Klassen von Matrizen kann man zeigen, dass das Gauß-Seidel-Verfahren schneller konvergiert. Wir werden nun zeigen, dass das Gauß-Seidel-Verfahren für symmetrische positiv definite Matrizen A stets konvergiert.
Sei $\langle v, w\rangle = \sum_{i=1}^{d} v_i\, w_i$ das euklidische Skalarprodukt im $\mathbb{R}^d$. Die von einer symmetrisch positiv definiten Matrix induzierte quadratische Form

$$(x, y)_A := \langle x, A\, y\rangle = \langle A\, x, y\rangle, \qquad (x, y \in \mathbb{R}^n)$$

nennt man auch ***Energieskalarprodukt***, die entsprechende Norm

$$\|x\|_A = \sqrt{(x, x)_A}$$

heißt auch ***Energienorm***.

Beispiel (Diskreter Laplace-Operator). Sei $D \subset \mathbb{R}^n$ offen und beschränkt, und $D_h = D \cap (h\mathbb{Z})^n$ die Diskretisierung von D durch ein Gitter der Maschenweite $h > 0$. Weiter sei $u = (u(x))_{x\in(h\mathbb{Z})^d}$ eine Funktion auf dem Gitter mit $u(x) = 0$ für alle $x \in D_h^C$ (Dirichlet-Randbedingungen). Ist $A = -\Delta_h$ mit der oben betrachteten Diskretisierung des Laplace-Operators, dann gilt

$$(A\, u)(x) \;=\; -(\triangle_h u)(x) \;=\; -h^{-2} \sum_{i=1}^{n} \big[(u(x + he_i) - u(x)) - (u(x) - u(x - he_i))\big].$$

Damit erhalten wir für die A-Energienorm von u:

$$\begin{aligned}
\|u\|_A^2 = \langle u, A\, u\rangle &= \sum_{x\in D_h} u(x)\,(A\, u)(x) = \sum_{x\in(h\mathbb{Z})^n} u(x)\,(A\, u)(x) \\
&= h^{-2} \sum_{i=1}^{n} \Big(\sum_x u(x)\,(u(x) - u(x - he_i)) - \sum_x u(x)\,(u(x + he_i) - u(x))\Big) \\
&= h^{-2} \sum_{i=1}^{n} \Big(\sum_x u(x + he_i)\,(u(x + he_i) - u(x)) - \sum_x u(x)\,(u(x + he_i) - u(x))\Big) \\
&= \sum_{x\in(h\mathbb{Z})^n} \sum_{i=1}^{n} \Big(\underbrace{\frac{u(x + he_i) - u(x)}{h}}_{\substack{\text{diskrete Richtungs-}\\ \text{ableitung } \partial_i u(x)}} \Big)^2 = \underbrace{\sum_{x\in(h\mathbb{Z})^n} \|\nabla_h u(x)\|^2}_{\text{Energie von } u},
\end{aligned}$$

wobei

$$\nabla_h u(x) \;=\; h^{-1}\big(u(x + he_1) - u(x), \ldots, u(x + he_n) - u(x)\big).$$

Es folgt, daß die lineare Abbildung $A = -\triangle_h$ auf

$$V \;=\; \{u = (u(x))_{x\in(h\mathbb{Z})^n} : u(x) = 0 \text{ für alle } x \in D_h^C\} \;\cong\; \mathbb{R}^{|D_h|}$$

symmetrisch und positiv definit ist, mit Energienorm

$$\|u\|_A^2 = \sum_{x\in(h\mathbb{Z})^n} \|\nabla_h u(x)\|^2.$$

Kontinuierliches Analogon:

$$-\int_D u\,\triangle u = \int_D \|\nabla u\|^2 \qquad \text{falls } u = 0 \text{ auf } \partial D \qquad \text{(siehe „Analysis III").}$$

Satz 5.11. Ist A symmetrisch und positiv definit, dann hat $A\,x = b$ eine eindeutige Lösung x^* und das Gauß-Seidel-Verfahren konvergiert monoton bzgl. $\|\cdot\|_A$ gegen x^*, d.h. es gibt ein $L < 1$ mit:

$$\|x^{(n+1)} - x^*\|_A \leq L\,\|x^{(n)} - x^*\|_A.$$

Bemerkung. Sei A symmetrisch positiv definit, dann gilt

$$a_{ii} = \langle e_i, A\,e_i\rangle > 0 \qquad \text{für alle } i.$$

Beweis. Zu zeigen ist:

$$\|T\|_A := \max_{\|x\|_A=1} \|T\,x\|_A < 1. \tag{5.15}$$

Die Behauptung folgt dann aus Satz 5.9

1. Es genügt zu zeigen:

   $$\|T\,x\|_A < 1 \qquad \text{für alle } x \in \mathbb{R}^n \text{ mit } \|x\|_A = 1, \tag{5.16}$$

   da $\{x \in \mathbb{R}^n \mid \|x\|_A = 1\}$ kompakt und die Abbildung $x \mapsto \|x\|_A$ stetig ist und somit in Gleichung (5.15) das Maximum angenommen wird.

2. Sei $T = I - P^{-1}\,A$, und P der Präkonditionierer, dann gilt:

   $$\begin{aligned}(T\,x, T\,x)_A &= \langle T\,x, A\,T\,x\rangle = \langle x, A\,T\,x\rangle - \langle P^{-1}\,A\,x, A\,T\,x\rangle \\ &= (x, T\,x)_A - \langle x, A^t\,(P^{-1})^t\,A\,T\,x\rangle \\ &= (x, T\,x)_A - (x, (P^{-1})^t\,A\,T\,x)_A \\ &= (x,x)_A - (x, Z\,x)_A \overset{!}{<} (x,x)_A,\end{aligned}$$

   wobei

   $$\begin{aligned}Z &= P^{-1}\,A + (P^t)^{-1}\,A - (P^t)^{-1}\,A\,P^{-1}\,A \\ &= (I + (P^t)^{-1}\,P - (P^t)^{-1}\,A)\,P^{-1}\,A \\ &= (P^t)^{-1}\,(P^t + P - A)\,P^{-1}\,A.\end{aligned}$$

3. Zu zeigen bleibt:

   $$(x, Z\,x)_A > 0 \qquad \text{für alle } x \in \mathbb{R}^n \backslash \{0\}.$$

   Es gilt:

   $$\begin{aligned}(x, Z\,x)_A &= \langle x, A\,(P^t)^{-1}\,(P^t + P - A)\,P^{-1}\,A\,x\rangle \\ &= \langle P^{-1}\,A\,x, (P^t + P - A)\,P^{-1}\,A\,x\rangle.\end{aligned}$$ ■

   Im Gauß-Seidel-Verfahren gilt:

   $$P = D - L, \qquad P^t = D - L^t = D - R,$$

da A symmetrisch ist. Es folgt:

$$P^t + P - A = (D - L) + (D - R) - (D - L - R) = D \qquad \text{ist positiv definit.}$$

Also folgt

$$(x, Z\,x)_A > 0 \qquad \text{für alle } x \in \mathbb{R}^n \backslash \{0\}.$$

Bemerkung. Im Beweis haben wir benutzt, daßA symmetrisch positiv definit und $P + P^t - A$ positiv definit ist!

Relaxationsverfahren

Allgemeines Iterationsverfahren:

$$x^{(k+1)} = T\,x^{(k)} + f, \qquad T = I - P^{-1}\,A, \qquad f = P^{-1}\,b.$$

Idee: erzwinge Konvergenz durch Dämpfungsparameter $\omega \in (0, 1]$:
Relaxationsverfahren:

$$x^{(k+1)} = \omega\,(T\,x^{(k)} + f) + (1 - \omega)\,x^{(k)}.$$

Bemerkung. a) x ist genau dann Fixpunkt, wenn $x = \omega\,(T\,x + f) + (1 - \omega)\,x$ gilt. Falls $\omega \neq 0$ ist dies äquivalent zu $x = T\,x + f$ bzw. $A\,x = b$.

b) Die Iterationsmatrix ist

$$T_\omega = \omega\,T + (1 - \omega)\,I = I - \omega\,P^{-1}\,A,$$

d.h. das Relaxationsverfahren ist ein Iterationsverfahren mit Präkonditionierer

$$P_\omega = \frac{1}{\omega}\,P.$$

Ziel: Wähle ω so, daß$\varrho(T_\omega)$ möglichst klein ist.

Relaxiertes Jacobi-Verfahren (Jacobi over-relaxation – „JOR")

Jacobi-Verfahren:

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

JOR-Verfahren:

$$\begin{aligned} x^{(k+1)} &= \omega\,D^{-1}\,(b + (L + R)\,x^{(k)}) + (1 - \omega)\,x^{(k)}, \qquad \text{d.h.} \\ x_i^{(k+1)} &= \frac{\omega}{a_{ii}}\big(b_i - \sum_{j \neq i} a_{ij}\,x_j^{(k)}\big) + (1 - \omega)\,x_i^{(k)}. \end{aligned}$$

Präkonditionierer:

$$P_\omega = \frac{1}{\omega}\,P.$$

Satz 5.12. Ist A eine symmetrisch positiv definite Matrix, dann konvergiert das JOR-Verfahren für alle $\omega \in (0, 2/\varrho(D^{-1}\,A))$.

Beweis.

$$T_\omega = I - \omega D^{-1} A,$$

also gilt:

$$\lambda \text{ ist genau dann Eigenwert von } D^{-1} A, \text{ wenn } 1 - \omega\lambda \text{ Eigenwert von } T_\omega \text{ ist.} \tag{5.17}$$

■

Ist A symmetrisch positiv definit, so auch $D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$, d.h. alle Eigenwerte von $D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$ sind reell und positiv. Da $D^{-1}A$ ähnlich zu $D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$ ist, hat auch $D^{-1}A$ nur reelle und positive Eigenwerte. Es folgt, dass das JOR-Verfahren genau dann konvergiert, wenn für den Spektralradius von T_ω gilt:

$$\varrho(T_\omega) = \max\{|1 - \omega\lambda| \,\big|\, \lambda \text{ Eigenwert von } D^{-1} A\} < 1.$$

Dies ist äquivalent dazu, dass $\omega\lambda < 2$ für alle Eigenwerte von $D^{-1} A$ gilt, und da $D^{-1} A$ nur reelle und positive Eigenwerte hat, ist dies äquivalent zu

$$\omega\, \varrho(D^{-1} A) < 2.$$

Sukzessives Relaxationsverfahren (Successive over-relaxation – „SOR")

Modifiziere JOR-Verfahren à la Gauß-Seidel-Verfahren:

$$x_i^{(k+1)} = \frac{\omega}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij}\, x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}\, x_j^{(k)} \right) + (1 - \omega)\, x_i^{(k)}.$$

In Matrixform:

$$x^{(k+1)} = \omega D^{-1} (b + L\, x^{(k+1)} + R\, x^{(k)}) + (1 - \omega)\, x^{(k)}.$$

Dies ist äquivalent zu:

$$\underbrace{\left(\frac{1}{\omega} D - L\right)}_{= P_\omega \text{ Präkonditionierer}} x^{(k+1)} = \underbrace{\left(R + \frac{1 - \omega}{\omega} D\right)}_{N_\omega} x^{(k)} + b.$$

Satz 5.13. Ist A symmetrisch positiv definit, dann konvergiert das SOR-Verfahren für alle $\omega \in (0, 2)$ monoton bzgl. $\|\cdot\|_A$.

Beweis. Verfahre wie bei Satz 5.11:

$$P_\omega = \frac{1}{\omega} D - L, \qquad P_\omega^t = \frac{1}{\omega} D - R, \qquad A = D - L - R.$$

Dann ist

$$P_\omega + P_\omega^t - A = \left(\frac{2}{\omega} - 1\right) D$$

positiv definit für alle $\omega < 2$. Damit folgt die Behauptung. ■

5.5. Abstiegsverfahren

Sei $A \in \mathbb{R}^{d\times d}$ symmetrisch und positiv definit, $b \in \mathbb{R}^d$. Wir betrachten wieder das lineare Gleichungssystem

$$A\,x = b. \tag{5.18}$$

Wie zuvor wollen wir die eindeutige Lösung $x^* = A^{-1}\,b$ numerisch berechnen. Dieses Problem lässt sich in ein *Minimierungsproblem* der folgenden Funktion ϕ umformulieren:

$$\phi(x) := \frac{1}{2}\langle x, A\,x\rangle - \langle x, b\rangle.$$

Lemma 5.14. *Die Lösung $x^* = A^{-1}\,b$ des linearen Gleichungssystems ist das eindeutige Minimum von ϕ, und für alle $x \in \mathbb{R}^d$ gilt:*

$$\phi(x) - \phi(x^*) = \frac{1}{2}\,\|x - x^*\|_A^2.$$

Beweis. Mit quadratischer Ergänzung sieht man, dass die Funktion

$$\phi(x) = \frac{1}{2}(x,x)_A - (x, A^{-1}b)_A = \frac{1}{2}\|x - x^*\|_A^2 - \frac{1}{2}\|x^*\|_A^2$$

minimal ist für $x = x^*$. Das Minimum ist dann

$$\phi(x^*) = -\frac{1}{2}\|x^*\|_A^2.$$

■

Allgemeines Iterationsverfahren zur Minimierung von $\phi : \mathbb{R}^d \to \mathbb{R}$ (Line Search):

$$x^{(k+1)} = x^{(k)} + \alpha_k\,p^{(k)}.$$

Hierbei bezeichnen

$$p^{(0)}, p^{(1)}, \ldots \in \mathbb{R}^d, \qquad \text{die „Suchrichtungen",}$$
$$\alpha_0, \alpha_1, \ldots \in \mathbb{R}, \qquad \text{die „Schrittweiten",}$$

Für die eben betrachtete Funktion ϕ wählt man die Schrittweiten α_k so, dass

$$\phi(x^{(k+1)}) = \phi(x^*) + \frac{1}{2}\|x^{(k+1)} - x^*\|_A^2$$

minimal unter allen $x^{(k)} + \alpha\,p^{(k)}$, $\alpha \in \mathbb{R}$, ist. Dies ist der Fall für

$$\begin{aligned} x^{(k+1)} &= A\text{-orthogonale Projektion von } x^* \text{ auf Gerade } x^{(k)} + \operatorname{span}(p^{(k)}) \\ &= x^{(k)} + \left(x^* - x^{(k)}, \frac{p^{(k)}}{\|p^{(k)}\|_A}\right)_A \frac{p^{(k)}}{\|p^{(k)}\|_A}, \end{aligned}$$

d.h.

$$\alpha_k^{\text{opt}} = \frac{(x^* - x^{(k)}, p^{(k)})_A}{(p^{(k)}, p^{(k)})_A} = \frac{\langle A\,x^* - A\,x^{(k)}, p^{(k)}\rangle}{\langle p^{(k)}, A\,p^{(k)}\rangle}.$$

Es gilt also:

$$\alpha_k^{\text{opt}} = \frac{\langle r^{(k)}, p^{(k)}\rangle}{\langle p^{(k)}, A\,p^{(k)}\rangle} \qquad \text{mit Residuum} \qquad r^{(k)} = b - A\,x^{(k)}.$$

Wir betrachten nun zwei spezielle Abstiegsverfahren.

Gradientenverfahren

Sei

$$p^{(k)} = -\nabla\phi(x^{(k)}), \qquad \text{die Richtung des steilsten Abstiegs.}$$

In unserem Fall gilt:

$$\begin{aligned}
\phi(x) &= \frac{1}{2}\langle x, A\,x\rangle - \langle x, b\rangle, && A \text{ symmetrisch,}\\
p^{(k)} &= -\nabla\phi(x^{(k)}) = b - A\,x^{(k)} = r^{(k)}, && \text{„Residuum von } x^{(k)}\text{“,}\\
\alpha_k &= \frac{\langle r^{(k)}, r^{(k)}\rangle}{\langle r^{(k)}, A\,r^{(k)}\rangle} = \frac{\|r^{(k)}\|^2}{\|r^{(k)}\|_A^2}.
\end{aligned}$$

Das Residuum lässt sich rekursiv berechnen:

$$\begin{aligned}
r^{(k+1)} &= b - A\,x^{(k+1)} = b - A\,x^{(k)} - \alpha_k\,A\,r^{(k)}\\
&= r^{(k)} - \alpha_k\,A\,r^{(k)}.
\end{aligned}$$

Damit erhalten wir folgenden Algorithmus:

Algorithmus 10: Steepest Descent

Input: $A \in \mathbb{R}^{d\times d}$ symmetrisch positiv definit, $b \in \mathbb{R}^d$, $x^{(0)} \in \mathbb{R}^d$
Output: Iterationsfolgen $x^{(k)} \in \mathbb{R}^d$, $r^{(k)} \in \mathbb{R}^d$

```
r^(0) ← b − A x^(0);
for k ← 0, 1, ... do
    α_k ← ⟨r^(k), r^(k)⟩ / ⟨r^(k), A r^(k)⟩ ;
    x^(k+1) ← x^(k) + α_k r^(k) ;
    r^(k+1) ← r^(k) − α_k A r^(k);
```

Bemerkung. Die Iterationsmatrix im k-ten Schritt des Gradientenverfahrens ist

$$T^{(k)} = I - \alpha_k\,A, \qquad \text{(keine Präkonditionierung!).}$$

Das Gradientenverfahren ist also ein ***nicht-stationäres*** (d.h. die Iterationsmatrix hängt von k ab), *nicht präkonditioniertes Verfahren mit Relaxationsparameter* α_k.

Satz 5.15. Ist A symmetrisch positiv definit, dann konvergiert das Gradientenverfahren für jeden Startwert $x^{(0)}$ monoton bzgl. der Energienorm $\|\cdot\|_A$, und es gilt:

$$\|x^{(k+1)} - x^*\|_A \le \frac{K_2(A) - 1}{K_2(A) + 1} \cdot \|x^{(k)} - x^*\|_A,$$

wobei $K_2(A) = \|A\|_2 \cdot \|A^{-1}\|_2$ die l^2-Kondition der Matrix A ist.

Wir zeigen zunächst:

Lemma 5.16. *Ist A symmetrisch positiv definit mit Eigenwerten $0 < \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_d$, dann gilt:*

$$\|A\|_2 = \max_{\|x\|=1} \|A\,x\| = \lambda_d = \varrho(A),$$

A^{-1} ist symmetrisch positiv definit mit Eigenwerten $\frac{1}{\lambda_1} \geq \frac{1}{\lambda_2} \geq \cdots \geq \frac{1}{\lambda_d} > 0$ und es gilt:

$$K_2(A) = \|A\|_2 \cdot \|A^{-1}\|_2 = \frac{\lambda_d}{\lambda_1}.$$

Beweis. Da A symmetrisch positiv definit ist, gibt es eine Orthonormalbasis von Eigenvektoren e_i von A. D.h. es gilt $\langle e_i, e_j \rangle = \delta_{ij}$ und jeder Vektor x lässt sich darstellen als $x = \sum_{i=1}^{d} c_i\, e_i$. Damit folgt:

$$A\,x = \sum_{i=1}^{d} c_i\,\lambda_i\,e_i, \qquad \|x\|^2 = \sum_{i=1}^{d} c_i^2, \qquad \|A\,x\|^2 = \sum_{i=1}^{d} c_i^2\,\lambda_i^2.$$

Weiterhin gilt

$$\|A\|_2 = \max_{x\neq 0} \frac{\|A\,x\|}{\|x\|} = \max_{i=1\ldots d} |\lambda_i| = \lambda_d.$$

Analog kann man

$$\|A^{-1}\|_2 = \max_{i=1\ldots d} \left|\frac{1}{\lambda_i}\right| = \frac{1}{\lambda_1}$$

zeigen. ■

Beweis (Beweis von Satz 5.15). Sei $T_\alpha := I - \alpha\,A$ die Iterationsmatrix des Verfahrens mit Relaxationsparameter α.

i) Vergleich mit T_α-Verfahren:
Wir behaupten, dass der Fehler des Gradientenverfahrens durch den Fehler des Verfahrens mit T_α für alle $\alpha \in \mathbb{R}$ beschränkt ist:

$$\|x^{(k+1)} - x^*\|_A \leq \|T_\alpha\,(x^{(k+1)} - x^*)\|_A.$$

Nach Definition ist $x^{(k+1)}$ die A-orthogonale Projektion von x^* auf die Gerade $x^{(k)} + \alpha\,r^{(k)}$, $\alpha \in \mathbb{R}$. Also gilt für alle $\alpha \in \mathbb{R}$:

$$\begin{aligned}\|x^{(k+1)} - x^*\|_A &\leq \|x^{(k)} + \alpha\,r^{(k)} - x^*\|_A \\ &= \|T_\alpha\,x^{(k)} + \alpha\,b - T_\alpha\,x^* - \alpha\,b\|_A = \|T_\alpha\,(x^{(k+1)} - x^*)\|_A.\end{aligned}$$

ii) Kontraktivität von T_α:
Seien $0 < \lambda < \ldots < \lambda_d$ die Eigenwerte von A. Wir behaupten:

$$\|T_\alpha\,y\|_A \leq \max_{i=1\ldots d} |1 - \alpha\,\lambda_i| \cdot \|y\|_A \qquad \text{für alle } y \in \mathbb{R}^d.$$

Seien e_i die Eigenvektoren zu λ_i und $\langle e_i, e_j \rangle = \delta_{ij}$. Dann gilt:

$$\begin{aligned}T_\alpha\,e_i &= (1 - \alpha\,\lambda_i)\,e_i, \qquad \text{und} \\ A\,T_\alpha\,e_i &= \lambda_i\,(1 - \alpha\,\lambda_i)\,e_i.\end{aligned}$$

Für $y = \sum_{i=1}^{n} c_i\, e_i$ folgt:

$$\begin{aligned}\|T_\alpha y\|_A^2 &= (T_\alpha\, y, A\, T_\alpha\, y)\\ &= \Big(\sum_{i=1}^{d} c_i\,(1-\alpha\,\lambda_i)\,e_i, \sum_{i=1}^{d} c_i\,\lambda_i\,(1-\alpha\,\lambda_i)\,e_i\Big) = \sum_{i=1}^{d} c_i^2\,\lambda_i\,(1-\alpha\,\lambda_i)^2.\end{aligned}$$

Mit

$$\|y\|_A^2 = (y, A\, y) = \sum_{i=1}^{d} c_i^2\,\lambda_i$$

folgt dann

$$\|T_\alpha y\|_A^2 \le \max_{i=1\ldots d}(1-\alpha\,\lambda_i)^2\,\|y\|_A^2.$$

iii) Optimieren der Abschätzung:
Nach i) und ii) gilt für alle $\alpha \ge 0$:

$$\|x^{(k+1)} - x^*\|_A \le \|T_\alpha\,(x^{(k)} - x^*)\|_A \le c(\alpha)\cdot\|x^{(k)} - x^*\|_A,$$

wobei

$$c(\alpha) := \max_{i=1\ldots d} |1-\alpha\,\lambda_i| = \max(1-\alpha\,\lambda_1, \alpha\,\lambda_d - 1)$$ ■

wegen $1-\alpha\,\lambda_1 \ge 1-\alpha\,\lambda_2 \ge \ldots \ge 1-\alpha\,\lambda_d$. Um eine möglichst gute Abschätzung zu erhalten, wählen wir a nun so, dass $c(\alpha)$ minimal wird. Dies ist genau dann der Fall, wenn:

$$\alpha\,\lambda_d - 1 = 1-\alpha\,\lambda_1, \qquad \text{d.h.} \quad \alpha = \frac{2}{\lambda_d + \lambda_1},$$

und für diese Wahl von α gilt:

$$c(\alpha) = 1-\alpha\,\lambda_1 = \frac{\lambda_d - \lambda_1}{\lambda_d + \lambda_1} = \frac{\frac{\lambda_d}{\lambda_1} - 1}{\frac{\lambda_d}{\lambda_1} + 1} = \frac{K_2(A) - 1}{K_2(A) + 1}.$$

Bemerkung. a) Wegen

$$\frac{K_2(A) - 1}{K_2(A) + 1} = 1 - \frac{2}{K_2(A) + 1}$$

kann die Konvergenzgeschwindigkeit des Gradientenverfahrens bei Matrizen mit großer Kondition relativ langsam sein.

b) Das Gradientenverfahren lässt sich auch zur Minimierung nicht-quadratischer Funktionen ϕ verwenden – aber es ist dann im Allgemeinen nicht klar, wie die Schrittweiten α_k geeignet zu wählen sind.

Verfahren der konjugierten Gradienten (CG)

Kann man im zweidimensionalen Fall erzwingen, dass x^* im nächsten Schritt getroffen wird (also das Verfahren abbricht)? Eine hinreichende Bedingung ist:

$$p^{(k+1)} \parallel x^* - x^{(k+1)}, \qquad \text{d.h. die } k+1\text{-te Suchrichtung ist parallel zu } x^* - x^{(k+1)}. \tag{5.19}$$

Nach Konstruktion ist $x^{(k+1)}$ die A-orthogonale Projektion von x^* auf $x^{(k)} + \operatorname{span}(p^{(k)})$, also gilt:

$$x^* - x^{(k+1)} \perp_A p^{(k)}, \qquad \text{d.h. die } k\text{-te Suchrichtung ist } A\text{-orthogonal zu } x^* - x^{(k+1)}.$$

Im zweidimensionalen Fall folgt aus $p^{(k+1)} \perp_A p^{(k)}$ die Gültigkeit von (5.19). Dies motiviert folgende *Modifikation des Gradientenverfahrens*:

$$p^{(k+1)} = r^{(k+1)} - \frac{(r^{(k+1)}, p^{(k)})_A}{(p^{(k)}, p^{(k)})_A} p^{(k)}.$$

Es gilt nun:

$$p^{(k+1)} \perp_A p^{(k)},$$

d.h. $p^{(k+1)}$ ist A-konjugiert zu $p^{(k)}$ („konjugierter Gradient"). $r^{(k+1)}$ bezeichnet das Residuum des Gradientenverfahren des vorigen Abschnitts.

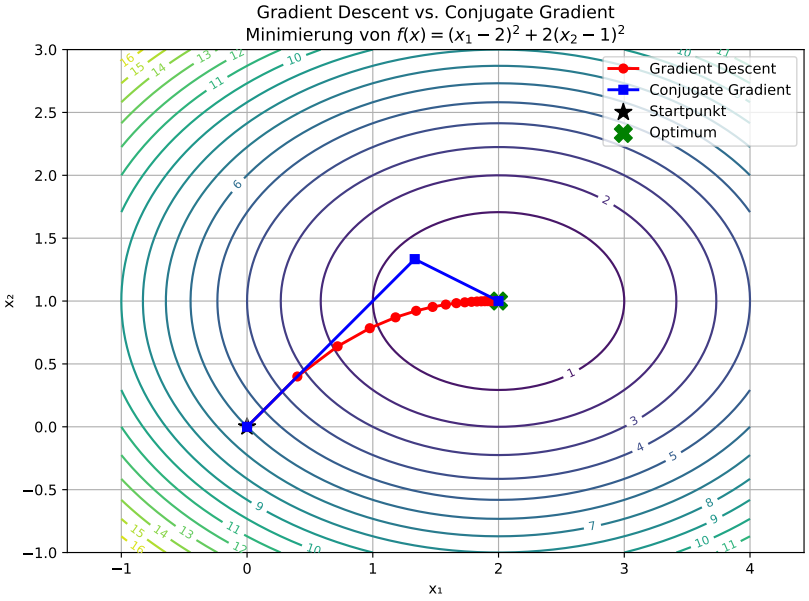


Abbildung 5.3.: Visualisierung von Gradientenverfahren und Verfahren der konjugierten Gradienten zum Minimum $(2, 1)$ der Funktion $f(x) = (x_1 - 2)^2 + 2(x_2 - 1)^2$.

Algorithmus 11: CG-Verfahren

Input: $A \in \mathbb{R}^{d\times d}$ symmetrisch positiv definit, $b \in \mathbb{R}^d$, $x^{(0)} \in \mathbb{R}^d$
Output: Iterationsfolgen $x^{(k)} \in \mathbb{R}^d$, $r^{(k)} \in \mathbb{R}^d$, $p^{(k)} \in \mathbb{R}^d$

$r^{(0)} \leftarrow b - A\,x^{(0)}$, $p^{(0)} \leftarrow r^{(0)}$;
for $k \leftarrow 0, 1, \ldots$ **do**
  $\alpha_k \leftarrow \frac{\langle r^{(k)}, p^{(k)}\rangle}{(p^{(k)}, p^{(k)})_A}$;
  $x^{(k+1)} \leftarrow x^{(k)} + \alpha_k\, p^{(k)}$;
  $r^{(k+1)} \leftarrow r^{(k)} - \alpha_k\, A\, p^{(k)}$;
  $\beta_k \leftarrow \frac{(r^{(k+1)}, p^{(k)})_A}{(p^{(k)}, p^{(k)})_A}$;
  $p^{(k+1)} \leftarrow r^{(k+1)} - \beta_k\, p^{(k)}$

Wir wollen nun zeigen, dass das CG-Verfahren im $\mathbb{R}^d$ nach höchstens d Schritten abbricht, und die exakte Lösung x^* liefert. Sei dazu

$$V_k := \operatorname{span}\{p^{(0)}, \ldots, p^{(k)}\}.$$

Durch Induktion folgt sofort

$$V_k = \operatorname{span}\{r^{(0)}, \ldots, r^{(k)}\} = \operatorname{span}\{r^{(0)}, A\,r^{(0)}, \ldots, A^k\, r^{(0)}\}.$$

V_k heißt auch ***Krylov-Raum***.

Satz 5.17. Für alle $k \geq 0$ gilt:

i) $x^{(k)} - x^* \perp_A V_k$,

ii) $r^{(k+1)} \perp_A V_k$, d.h. $r^{(0)}, r^{(1)}, \ldots, r^{(k+1)}$ sind orthogonal,

iii) $p^{(k+1)} \perp_A V_k$, d.h. $p^{(0)}, p^{(1)}, \ldots, p^{(k+1)}$ sind A-orthogonal.

Insbesondere folgt $r^{(k)} = 0$, also $x^{(k)} = x^*$, für alle $k \geq d$, d.h. das CG-Verfahren liefert nach maximal d Schritten die exakte Lösung.

Korollar 5.18 (Berechnung der Koeffizienten und Residuen).

$$\alpha_k = \frac{\langle r^{(k)}, r^{(k)}\rangle}{\langle p^{(k)}, A\,p^{(k)}\rangle}, \qquad \beta_k = -\frac{\langle r^{(k+1)}, r^{(k+1)}\rangle}{\langle r^{(k)}, r^{(k)}\rangle}.$$

Beweis. Der Beweis verbleibt als Übungsaufgabe. ■

Diese Art der Berechnung ist effizienter und numerisch stabiler als die Formel von oben, und sollte daher in Implementationen des CG-Verfahrens verwendet werden.

Beweis (Beweis von Satz 5.17). Die Aussagen werden durch vollständige Induktion nach k bewiesen. Der Induktionsanfang ist eine übungsaufgabe. Seien die drei Aussagen für k wahr.

i) Nach Konstruktion ist $x^{(k+1)}$ die A-orthogonale Projektion auf die Gerade $x^{(k)} + \operatorname{span}(p^{(k)})$, also gilt
$$(x^{(k+1)} - x^*, p^{(k)})_A = 0.$$
Zudem gilt nach ii):
$$x^{(k+1)} - x^* = x^{(k)} - x^* + \alpha\, p^{(k)}.$$
Der letzte Ausdruck ist nach Induktionsvoraussetzung A-orthogonal zu V_{k-1}, also folgt:
$$x^{(k+1)} - x^* \perp_A V_k.$$

ii) Für alle $z \in V^{(k)}$ gilt nach i):
$$0 = (x^{(k+1)} - x^*, z)_A = \langle A\, x^{(k+1)} - A\, x^*, z\rangle = \langle -r^{(k+1)}, z\rangle.$$

iii) Nach Konstruktion ist $p^{(k+1)}$ A-orthogonal zu $p^{(k)}$. Zudem gilt für $z \in V_{k-1}$ nach Induktionsvoraussetzung:
$$(p^{(k+1)}, z)_A = (r^{(k+1)}, z)_A - \beta_k\, (p^{(k)}, z)_A = \langle r^{(k+1)}, A\, z\rangle = 0.$$
Also gilt $p^{(k+1)} \perp_A V_{k-1}$, und damit $p^{(k+1)} \perp_A V_k$. ■

In der Praxis wird das CG-Verfahren häufig schon nach weniger als d Schritten beendet. Hier gilt folgende Fehlerabschätzung:

Satz 5.19. Ist A symmetrisch positiv definit, dann gilt
$$\|x^{(k)} - x^*\|_A \leq \frac{2\, c^k}{1 + c^{2k}}\, \|x^{(0)} - x^*\|_A,$$
wobei
$$c = \frac{\sqrt{K_2(A)} - 1}{\sqrt{K_2(A)} + 1} = 1 - \frac{2}{\sqrt{K_2(A)} + 1}.$$

Beweis. Das konjugierte Gradientenverfahren produziert nach k Schritten eine Näherung $x^{(k)}$. Wir definieren den Fehler nach k Iterationen als $e^{(0)} := x^* - x^{(0)}$ Da $e^{(k)} \perp_A V_k$ und $r^{(0)} := b - Ax^(0) = Ae^{(0)}$, lässt sich der Fehler als
$$e^{(k)} = p_k(A)(e^{(0)})$$
schreiben für ein Polynom p_k vom Grad maximal k mit $p_k(0) = 1$. Wir bezeichnen den Raum der Polynome mit Grad maximal k und $p_k(0) = 1$ mit Π_k^0. Da A symmetrisch positiv definit ist, existiert eine orthogonale Matrix Q ($Q^\top Q = QQ^\top = I$), sodass $A = Q\Lambda Q^\top$ mit Diagonalmatrix $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ und $0 < \lambda_{\min} = \lambda_d \leq \cdots \leq \lambda_1 = \lambda_{\max}$. Somit gilt $p(A) = Q^T p(\Lambda) Q$. Für ein Polynom $p \in \Pi_k^0$ erfüllt der Fehler in der A-Norm daher
$$\|e^{(k)}\|_A^2 = \|p_k(A)e^{(0)}\|_A^2 = \|Q p_k(\Lambda) Q^\top e^{(0)}\|_A^2.$$

Wir nutzen die Orthogonalität und die Substitution $y := Q^\top e^{(0)}$, sodass

$$\begin{aligned}\|Qp_k(\Lambda)Q^\top e^{(0)}\|_A^2 &= (e^{(0)})^\top Qp_k(\Lambda)Q^\top AQp_k(\Lambda)Q^\top e^{(0)} = y^\top p_k(\Lambda)\Lambda p_k(\Lambda)y \\ &\leq \max_{1\leq i\leq d} |p_k(\lambda_i)|^2 \sum_{i=1}^{d} y_i^2\lambda_i^2 = \max_{\lambda\in[\lambda_{\min},\lambda_{\max}]} |p_k(\lambda)|^2\|e^{(0)}\|_A^2.\end{aligned}$$

Damit gilt

$$\|e^{(k)}\|_A^2 \leq \min_{p_k\in\Pi_k^0} \max_{\lambda\in[\lambda_{\min},\lambda_{\max}]} |p_k(\lambda)|^2\|e^{(0)}\|_A^2.$$

Das optimale Polynom q_k zur Minimierung des Maximums $|q_k(\lambda)|$ über das Spektrum von A mit $q_k(0) = 1$ ist gegeben durch ein verschobenes und skaliertes Tschebyscheff-Polynom (siehe Quarteroni/Sacco/Saleri: „Numerische Mathematik 1). Für dieses gilt

$$\max_{\lambda\in[\lambda_{\min},\lambda_{\max}]} |q_k(\lambda)| = \frac{2c^k}{1+c^{2k}},$$

mit

$$c = \frac{\sqrt{\kappa_2(A)}-1}{\sqrt{\kappa_2(A)}+1}.$$

Daraus folgt unmittelbar

$$\|x^{(k)} - x^*\|_A = \|e^{(k)}\|_A \leq \frac{2c^k}{1+c^{2k}}\|e^{(0)}\|_A = \frac{2c^k}{1+c^{2k}}\|x^{(0)} - x^*\|_A.$$ ■

Bemerkung (Aufwand des CG-Verfahrens.). a) Maximaler Aufwand:

- pro Schritt: $O(d^2)$ (Multiplikation Matrix × Vektor),
- insgesamt: $O(d^3)$ (exakte Lösung nach d Schritten).

b) Matrix dünn besetzt (vgl. Programmieraufgabe):

- pro Schritt nur $O(d)$.
- insgesamt maximal $O(d^2)$.

c) Approximative Lösung: Schrittzahl $\ll d$. Daher ist weitere Reduktion des Aufwands möglich.

Acknowledgements

Besonderer Dank gilt Wassilij Gnedin, Felix Hoffmann, Andreas Haupt und Andreas Eberle für die Erstellung und Bereitstellung des Skripts, auf dem diese Vorlesung basiert. Darüber hinaus gilt Christoph Wiggers Dank für seine hilfreichen Hinweise zu Fehlern im Skript.

Index