Einführung in die Wahrscheinlichkeitstheorie

Prof. Dr. Andreas Eberle

28. September 2010

Inhaltsverzeichnis

Kapitel 1

Diskrete Zufallsvariablen

Unser Ziel in diesem Kapitel ist die mathematische Modellierung von **Zufallsvorgängen**. Einfache Beispiele für Zufallsvorgänge sind das Werfen eines Würfels oder Münzwürfe. Anhand dieser Beispiele wollen wir zunächst einige grundlegende Begriffe der Wahrscheinlichkeitstheorie veranschaulichen.
NOTATIONEN: $|A|$ bezeichnet die Anzahl der Elemente einer Menge A, A^C bezeichnet das Komplement der Menge A innerhalb einer bestimmten Menge B, die A enthält.

Beispiel (Werfen eines Würfels)**.**

- **Mögliche Fälle** sind $1, 2, 3, 4, 5, 6$. Mit $\Omega = \{1, 2, 3, 4, 5, 6\}$ wird die Menge aller möglichen Fälle bezeichnet. Ein **Elementarereignis** ist ein möglicher Fall, also ein Element $\omega \in \Omega$.

- **Ereignisse** sind die Objekte, denen man eine Wahrscheinlichkeit zuordnen kann, 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«	$\{2, 4, 5, 6\} = \{2, 4, 6\} \cup \{4, 5, 6\}$

  Jedes **Ereignis** kann durch eine **Teilmenge** A **von** Ω dargestellt werden!

- **Wahrscheinlichkeiten** werden mit P (für »probability«) bezeichnet. Zum Beispiel sollte für einen »fairen« Würfel gelten:

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

- **Zufallsvariablen** sind Abbildungen $X : \Omega \to S$, wobei S eine beliebige Menge ist, zum Beispiel:

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

Beispiel (Münzwürfe)**.** a) EIN MÜNZWURF:
Die Menge der möglichen Fälle ist $\Omega = \{0,1\}$, wobei 0 für »Kopf« und 1 für »Zahl« steht. Die Wahrscheinlichkeiten sind

$$P[\{1\}] = p \quad \text{und} \quad P[\{0\}] = 1 - p \quad \text{mit } 0 \le p \le 1.$$

Für $p = \frac{1}{2}$ ist der Münzwurf fair.

b) ENDLICH VIELE FAIRE MÜNZWÜRFE:
Die Menge der möglichen Fälle lautet

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

Alle Ausgänge sind genau dann gleich wahrscheinlich, wenn $P[\{\omega\}] = 2^{-n}$ für alle $\omega \in \Omega$ gilt. Dies wird im folgenden angenommen. Zufallsvariablen von Interesse sind beispielsweise:

- $X_i(\omega) := x_i$, das Ergebnis des i-ten Wurfs. Das Ereignis »i-ter Wurf ist Kopf« wird durch die Menge $A_i = \{\omega \in \Omega \mid X_i(\omega) = 0\} =: \{X_i = 0\}$ beschrieben, und hat die Wahrscheinlichkeit $P[A_i] = \frac{1}{2}$.
- $S_n(\omega) := \sum_{i=1}^n X_i(\omega)$, die Anzahl der Einsen in n Münzwürfen. Das Ereignis »genau k-mal Zahl« wird durch die Menge $A = \{\omega \in \Omega \mid S_n(\omega) = k\} =: \{S_n = k\}$ beschrieben und hat die Wahrscheinlichkeit $P[A] = \binom{n}{k} 2^{-n}$.

c) UNENDLICH VIELE MÜNZWÜRFE:
Die Menge der möglichen Fälle ist nun

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

Diese Menge ist überabzählbar, da die Abbildung

$$\Omega \to [0,1]$$
$$(x_1, x_2, \ldots) \mapsto 0.x_1x_2\ldots$$

surjektiv ist, (wobei das Einheitsintervall binär dargestellt wird). Die Definition von Ereignissen und Wahrscheinlichkeiten ist daher in diesem Fall aufwändiger. Wahrscheinlichkeitsverteilungen auf überabzählbaren Mengen werden systematisch in der Vorlesung »Einführung in die Wahrscheinlichkeitstheorie« betrachtet.

In dieser Vorlesung ist die Menge der möglichen Fälle Ω abzählbar. Solche Zufallsvorgänge werden **diskret** genannt.

1.1 Ereignisse und ihre Wahrscheinlichkeit

Ereignisse als Mengen

Sei Ω die Menge der möglichen Fälle und $A \subseteq \Omega$ ein Ereignis. Als Notationen für die Menge A werden wir auch verwenden:

$$A = \{\omega \in \Omega \mid \omega \in A\} = \{\omega \in A\} = \{\text{ »}A\text{ tritt ein« }\}.$$

Wir wollen nun **Kombinationen von Ereignissen** betrachten.
Seien A, B, A_i, $i \in I$, Ereignisse. Was bedeuten Ereignisse wie A^C, $A \cup B$, $\bigcap_{i\in I} A_i$ anschaulich? Um dies herauszufinden, betrachten wir einen möglichen Fall ω und untersuchen, wann dieser eintritt:

- $A \cup B$:

$$\begin{array}{lcl} \omega \in A \cup B & \Leftrightarrow & \omega \in A \text{ oder } \omega \in B, \\ \text{»}A \cup B \text{ tritt ein«} & \Leftrightarrow & \text{»}A \text{ tritt ein oder } B \text{ tritt ein«}. \end{array}$$

- $\bigcup_{i\in I} A_i$:

$$\begin{array}{lll} \omega \in \bigcup_{i\in I} A_i & \Leftrightarrow & \text{es gibt ein } i \in I \text{ mit } \omega \in A_i. \\ \text{»}\bigcup_{i\in I} A_i \text{ tritt ein«} & \Leftrightarrow & \text{»mindestens eines der Ereignisse } A_i \text{ tritt ein«.} \end{array}$$

- WEITERE BEISPIELE:

$$\begin{array}{lll} A \cap B & \Leftrightarrow & \text{»}A \text{ und } B \text{ treten ein«,} \\ \bigcap_{i\in I} A_i & \Leftrightarrow & \text{»jedes der } A_i \text{ tritt ein«,} \\ A^C = \Omega & \Leftrightarrow & \text{»}A \text{ tritt nicht ein«,} \\ A = \emptyset & \Leftrightarrow & \text{»unmögliches Ereignis« (tritt nie ein),} \\ A = \Omega & \Leftrightarrow & \text{»sicheres Ereignis« (tritt immer ein),} \\ A = \{\omega\} & \Leftrightarrow & \text{»Elementarereignis« (tritt nur im Fall } \omega \text{ ein).} \end{array}$$

Sei $\mathcal{A}$ die Kollektion aller im Modell zugelassenen bzw. in Betracht gezogenen Ereignisse.
$\mathcal{A}$ besteht aus Teilmengen von Ω, d.h.

$$\mathcal{A} \subseteq \mathcal{P}(\Omega), \qquad \text{wobei}$$
$$\mathcal{P}(\Omega) := \{A \mid A \subseteq \Omega\}$$

die Potenzmenge von Ω, d.h. die Menge aller Teilmengen von Ω bezeichnet. Die Kollektion $\mathcal{A}$ sollte unter den obigen Mengenoperationen, also abzählbaren Vereinigungen, Durchschnitten und Komplementbildung abgeschlossen sein. Wir fordern daher:

Axiom. *$\mathcal{A} \subseteq \mathcal{P}(\Omega)$ ist eine σ-**Algebra**, d.h.*

(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. Für σ-Algebren gilt auch:

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

b) Sind $A, B \in \mathcal{A}$, so gilt nach (iii) und a): $A \cup B = A \cup B \cup \emptyset \cup \emptyset \cup \ldots \in \mathcal{A}$.

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

Beispiel. Die Potenzmenge $\mathcal{A} = \mathcal{P}(\Omega)$ ist eine σ-Algebra.

Üblicherweise verwendet man $\mathcal{A} = \mathcal{P}(\Omega)$ bei **diskreten** Modellen, d.h. für abzählbare Ω. Bei nichtdiskreten Modellen kann man **nicht** jede Wahrscheinlichkeitsverteilung P auf einer σ-Algebra $\mathcal{A} \subset \mathcal{P}(\Omega)$ zu einer Wahrscheinlichkeitsverteilung auf $\mathcal{P}(\Omega)$ erweitern (siehe »Einführung in die Wahrscheinlichkeitstheorie«).

Wahrscheinlichkeitsverteilungen

Sei Ω eine nichtleere Menge und $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ eine σ-Algebra. Wir wollen nun Ereignissen $A \in \mathcal{A}$ eine Wahrscheinlichkeit $P[A]$ zuordnen. Für Ereignisse $A, B \in \mathcal{A}$ gilt:

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

Angenommen, A und B **treten nicht gleichzeitig ein**, d.h.

$$A \cap B = \emptyset, \qquad (A \text{ und } B \text{ sind »disjunkt«}).$$

Dann sollte »endliche Additivität« gelten:

$$P[A \cup B] = P[A] + P[B].$$

Axiom. *Eine Abbildung*

$$\begin{aligned} P \colon \mathcal{A} &\to [0, \infty] \\ A &\mapsto P[A] \end{aligned}$$

ist eine ***Wahrscheinlichkeitsverteilung*** *auf* $(\Omega, \mathcal{A})$*, wenn gilt:*

(i) 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:*

$$P\big[\bigcup_{i=1}^{\infty} A_i\big] = \sum_{i=1}^{\infty} P[A_i].$$

(ii) P *ist* ***»normiert«****, d.h.*

$$P[\Omega] = 1.$$

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

Satz 1.1 (Elementare Rechenregeln)**.** *Sei* $(\Omega, \mathcal{A}, P)$ *ein Wahrscheinlichkeitsraum.*

i) Es gilt $P[\emptyset] = 0$*,*

ii) Für $A, B \in \mathcal{A}$ *mit* $A \cap B = \emptyset$ *gilt* ***endliche Additivität****:*

$$P[A \cup B] = P[A] + P[B].$$

iii) Für $A, B \in \mathcal{A}$ mit $A \subseteq B$ gilt:

$$P[B] = P[A] + P[B \backslash A].$$

Insbesondere gilt:

$$\begin{aligned} P[A] &\leq P[B], && \textbf{\textit{»Monotonie«,}} \\ P[A^C] &= 1 - P[A], && \textbf{\textit{»Gegenereignis«,}} \\ P[A] &\leq 1. \end{aligned}$$

iv) Für $A, B \in \mathcal{A}$ gilt:

$$P[A \cup B] = P[A] + P[B] - P[A \cap B] \leq P[A] + P[B].$$

Beweis. i) Wegen der σ-Additivität von P gilt

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

und damit

$$P[\emptyset] = 0.$$

ii) Für disjunkte Ereignisse A, B folgt aus der σ-Additivität und mit i):

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

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

$$P[B] = P[A] + P[B \backslash A] \geq P[A].$$

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

iv) Nach iii) gilt:

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

□

Aussage iv) des Satzes lässt sich für endlich viele Ereignisse verallgemeinern. Nach iv) gilt für die Vereinigung von drei Ereignissen:

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

Mit vollständiger Induktion folgt:

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

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

Das Einschluss-/Ausschlussprinzip werden wir auf eine elegantere Weise am Ende dieses Kapitels beweisen (siehe Satz 1.9).

Diskrete Wahrscheinlichkeitsverteilungen

Als Beispiel für eine diskrete Wahrscheinlichkeitsverteilung haben wir den Münzwurf betrachtet:

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

ALLGEMEIN: Ist die Menge der möglichen Fälle Ω endlich oder abzählbar unendlich, dann setzen wir als zugehörige σ-Algebra $\mathcal{A} = \mathcal{P}[\Omega]$.

Satz 1.2. *i) Sei $0 \le p(\omega) \le 1$, $\sum_{\omega \in \Omega} p(\omega) = 1$ eine **Gewichtung der möglichen Fälle**. Dann ist durch*

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

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

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

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

*$p : \Omega \to [0,1]$ heißt **Massenfunktion** (»probability mass function«) der diskreten Wahrscheinlichkeitsverteilung P.*

Für den Beweis des Satzes brauchen wir einige Vorbereitungen.

Bemerkung (Vorbemerkung zu Summen mit positiven Summanden)**.** Sei A eine abzählbare Menge, $p(\omega) \geq 0$ für alle $\omega \in A$. Dann definieren wir

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

wobei $\omega_1, \omega_2, \ldots$ eine beliebige Abzählung von A ist.

Lemma 1.3. *i)* $\sum_{\omega \in A} p(\omega) \in [0, \infty]$ *und ist wohldefiniert (d.h. unabhängig von der Abzählung). Es gilt:*

$$\sum_{\omega \in A} p(\omega) = \sup_{\substack{F \subseteq A \\ |F| < \infty}} \sum_{\omega \in F} p(\omega). \tag{1.1.1}$$

Insbesondere gilt ***Monotonie:***

$$\sum_{\omega \in A} p(\omega) \leq \sum_{\omega \in B} p(\omega), \qquad (A \subseteq B). \tag{1.1.2}$$

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 Partialsummen $\sum_{i=1}^{n} p(\omega_i)$ monoton wachsend sind. Daraus folgt:

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

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

$$\sup_{n \in \mathbb{N}} \sum_{i=1}^{n} p(\omega_i) = \sup_{\substack{F \subseteq A \\ |F| < \infty}} \sum_{\omega \in F} p(\omega) \qquad \text{ist unabhängig von der Abzählung von A.}$$

»$\leq$«: 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 »$\leq$«.

»$\geq$«: Da $F \subseteq A$ endlich ist, gibt es ein $n \in \mathbb{N}$, so dass $F \subseteq \{\omega_1, \ldots, \omega_n\}$. Also gilt:

$$\sum_{\omega \in F} p(\omega) \leq \sum_{i=1}^{n} p(\omega_i) \leq \sum_{i=1}^{\infty} p(\omega_i).$$

Damit folgt »$\geq$«.

ii)
- Falls A endlich ist, 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.
- Sei andernfalls A abzählbar unendlich.

»$\leq$«: Da $F \subseteq A$ endlich, ist $F = \bigcup_i^{\infty} F \cap A_i$. Da diese Vereinigung disjunkt ist, folgt mit σ-Additivität und Gleichung (1.1.2):

$$P[F] = \sum_{i \in \mathbb{N}} P[F \cap A_i] \leq \sum_{i \in \mathbb{N}} P[A_i].$$

Mit (i)) gilt auch:

$$P[A] = \sup_{\substack{F \subseteq A \\ |F| < \infty}} P[F] \leq \sum_{i \in \mathbb{N}} P[A_i].$$

Damit folgt »$\leq$«.

»$\geq$«: Seien $F_i \subseteq A_i$ endlich. Da die F_i disjunkt sind, folgt mit σ-Additivität und Gleichung (1.1.2) für alle $n \in \mathbb{N}$:

$$\sum_{i=1}^{n} P[F_i] = P\left[\bigcup_{i=1}^{n} F_i\right] \leq P\left[\bigcup_{i=1}^{\infty} A_i\right] = P[A].$$

Mit (1.1.1) folgt

$$\sum_{i=1}^{n} P[A_i] \leq P[A],$$

und für $n \to \infty$ schließlich

$$\sum_{i=1}^{\infty} P[A_i] \leq P[A].$$

Damit folgt »$\geq$«.

□

Beweis von Satz 1.2. i) Es ist $P[\Omega] = \sum_{\omega \in \Omega} p(\omega) = 1$ nach Voraussetzung.

Seien A_i, $(i \in \mathbb{N})$ disjunkt und $A := \bigcup_{i=1}^{\infty} A_i$. Die σ-Additivität von P folgt aus Lemma 1.3.ii):

$$P\big[\bigcup_{i=1}^{\infty} A_i\big] = P[A] = \sum_{\omega \in A} p(\omega) = \sum_{i=1}^{\infty} \sum_{\omega \in A_i} p(\omega) = \sum_{i=1}^{\infty} P[A_i]$$

ii) Aus der σ-Additivität von P folgt:

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

□

Spezielle Wahrscheinlichkeitsverteilungen

Gleichverteilungen / Laplace-Modelle

Sei Ω endlich und nichtleer, $\mathcal{A} = \mathcal{P}(\Omega)$ und $p(\omega) = \frac{1}{|\Omega|}$ für alle $\omega \in \Omega$. Dann ist

$$P[A] = \frac{|A|}{|\Omega|} = \frac{\text{Anzahl »günstiger« Fälle}}{\text{Anzahl aller Fälle}}, \qquad (A \subseteq \Omega),$$

die Wahrscheinlichkeitsverteilung zu p und wird **Gleichverteilung** genannt.

Beispiele. a) n FAIRE MÜNZWÜRFE:

Sei $\Omega = \{0,1\}^n$ und P die Gleichverteilung. Dann ist

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

b) ZUFÄLLIGE PERMUTATIONEN:

Sei $\Omega = \mathcal{S}_n = \{\omega : \{1,2,\ldots,n\} \to \{1,2,\ldots,n\}$ bijektive Abbildungen $\}$ und P die Gleichverteilung. Dann ist

$$P[A] = \frac{|A|}{n!}.$$

Beispiele für zufällige Permutationen sind das Mischen eines Kartenspiels, Vertauschen von Hüten oder Umzug in die LWK, wobei n Schlüssel zufällig vertauscht werden. Es gilt:

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

Wie groß ist die Wahrscheinlichkeit, dass einer der Schlüssel sofort passt?

Das Ereignis »Schlüssel i passt« ist $A_i = \{\omega \mid \omega(i) = i\} = \{\text{»}i \text{ ist Fixpunkt« }\}$. Die Wahrscheinlichkeit für das Ereignis »ein Schlüssel passt« ist nach dem Einschluss-/Ausschlussprinzip (Satz 1.9):

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

wobei die innere Summe über alle k-elementigen Teilmengen läuft. Es folgt:

$$\begin{aligned} &= \sum_{k=1}^{n} (-1)^{k+1} \binom{n}{k} \frac{(n-k)!}{n!} \\ &= -\sum_{k=1}^{n} \frac{(-1)^k}{k!} \end{aligned}$$

Für das Gegenereignis erhalten wir:

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

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

Empirische Verteilungen

Seien $x_1, x_2, \ldots, x_n \in \Omega$ Beobachtungsdaten oder Merkmalsausprägungen, zum Beispiel das Alter aller Einwohner von Bonn. Sei

$$N[A] := |\{i \in \{1, \ldots, n\} \mid x_i \in A\}|, \qquad \text{die Anzahl bzw. Häufigkeit der Werte in } A, \text{ und}$$

$$P[A] := \frac{N[A]}{n}, \qquad \text{die relative Häufigkeit der Werte in } A.$$

Dann ist P eine Wahrscheinlichkeitsverteilung auf $(\Omega, \mathcal{P}(\Omega))$ mit Massenfunktion

$$p(\omega) = \frac{N[\{\omega\}]}{n}, \qquad \text{der relativen Häufigkeit der Merkmalsausprägungen.}$$

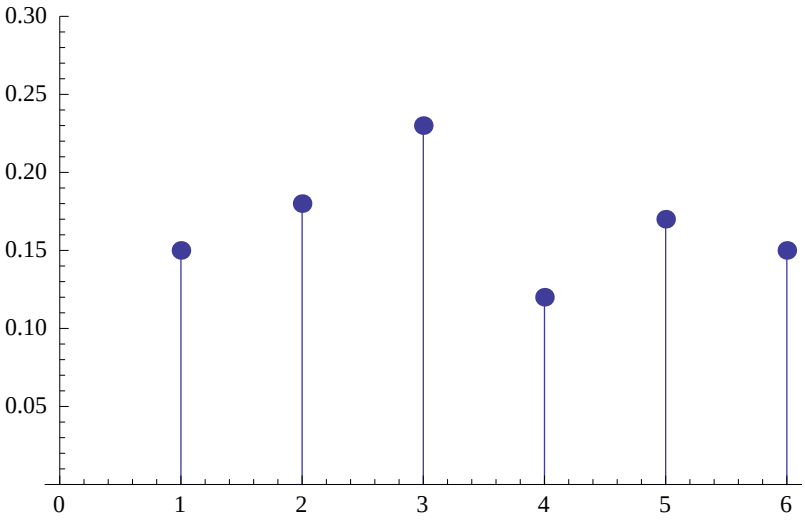
Beispiele. a) ABZÄHLUNG ALLER MÖGLICHEN FÄLLE:

Sei $x_1, \ldots, x_n$ eine Abzählung der Elemente in Ω. Dann stimmt die empirische Verteilung mit der Gleichverteilung überein.

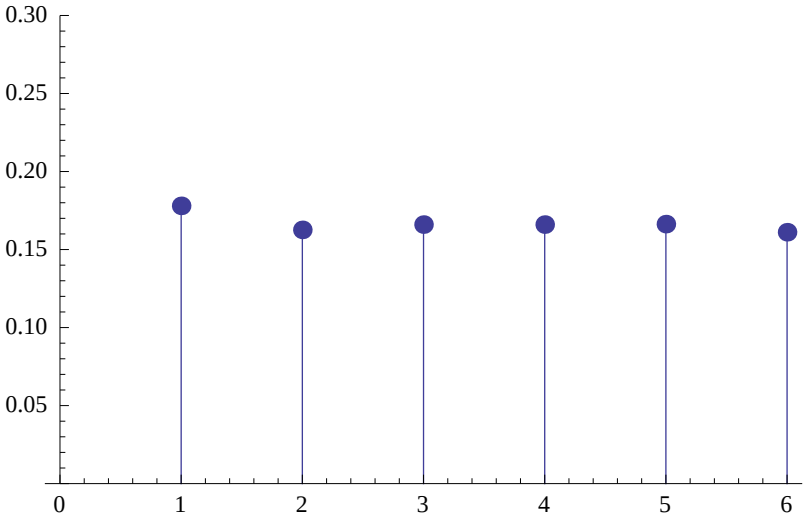
b) EMPIRISCHE VERTEILUNG VON n ZUFALLSZAHLEN AUS $\{1, 2, 3, 4, 5, 6\}$:

```
x=RandomChoice[{1,2,3,4,5,6},n];
ListPlot[BinCounts[x[[1 ;; n], {1, 7, 1}]/n,
 Filling -> Axis, PlotRange -> {0, 0.3},
 PlotStyle -> PointSize[Large]], {n, 1, 100, 1}
```

- $n = 100$:



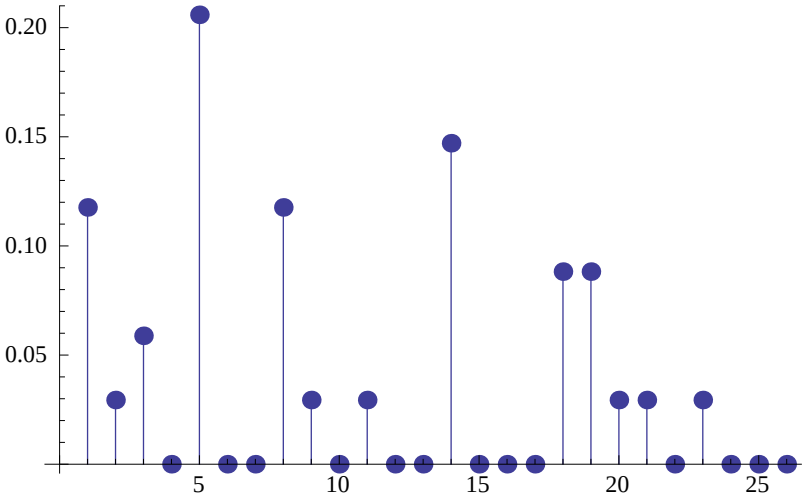
- $n = 10000$:



c) EMPIRISCHE VERTEILUNG DER BUCHSTABEN »A« BIS »Z«:

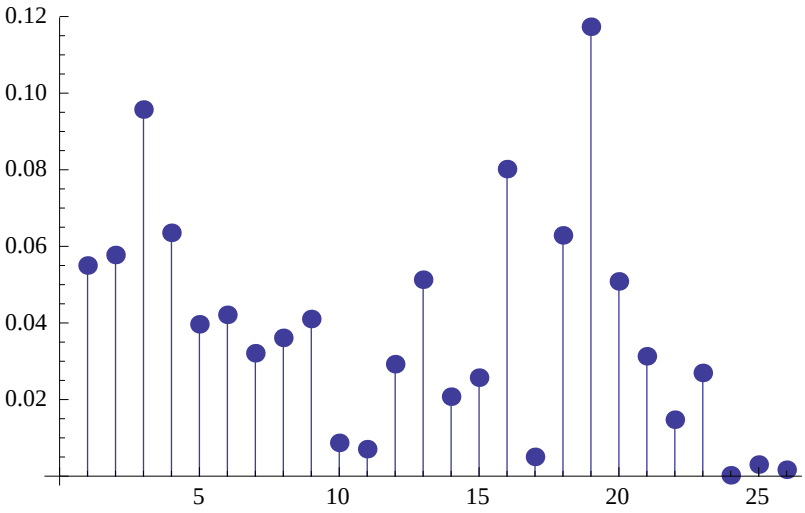
- in dem Wort »Eisenbahnschrankenwaerterhaeuschen«:

```
freq = StringCount["eisenbahnschrankenwaerterhaeuschen", #] & /@
   CharacterRange["a", "z"];
relfreq = freq/Total[freq];
ListPlot[relfreq, Filling -> Axis, PlotStyle -> PointSize[Large]]
```



- in einem englischen Wörterbuch:

```
freq = Length[DictionaryLookup[# ~~ ___]] & /@
   CharacterRange["a", "z"];
relfreq = freq/Total[freq];
ListPlot[relfreq, Filling -> Axis, PlotStyle -> PointSize[Large]]
```



d) BENFORDSCHES GESETZ:

Das Benfordsche Gesetz, auch Newcomb-Benford's Law (NBL) beschreibt eine Gesetzmäßigkeit in der Verteilung der Ziffernstrukturen von Zahlen in empirischen Datensätzen, zum Beispiel ihrer ersten Ziffern. Es lässt sich etwa in Datensätzen über Einwohnerzahlen von Städten, Geldbeträge in der Buchhaltung, Naturkonstanten etc. beobachten. Kurz gefasst besagt es:

»Je niedriger der zahlenmäßige Wert einer Ziffernsequenz bestimmter Länge an einer bestimmten Stelle einer Zahl ist, umso wahrscheinlicher ist ihr Auftreten. Für die Anfangsziffern in Zahlen des Zehnersystems gilt zum Beispiel: Zahlen mit der Anfangsziffer 1 treten etwa 6,5-mal so häufig auf wie solche mit der Anfangsziffer 9.«

1881 wurde diese Gesetzmäßigkeit von dem Mathematiker Simon Newcomb entdeckt und im „American Journal of Mathematics“ publiziert. Er soll bemerkt haben, dass in den benutzten Büchern mit Logarithmentafeln, die Seiten mit Tabellen mit Eins als erster Ziffer deutlich schmutziger waren als die anderen Seiten, weil sie offenbar öfter benutzt worden seien. Die Abhandlung Newcombs blieb unbeachtet und war schon in Vergessenheit geraten, als der Physiker Frank Benford (1883−1948) diese Gesetzmäßigkeit wiederentdeckte und darüber 1938 neu publizierte. Seither war diese Gesetzmäßigkeit nach ihm benannt, in neuerer Zeit wird aber durch die Bezeichnung »Newcomb-Benford's Law« (NBL) dem eigentlichen Urheber wieder Rechnung getragen. Bis vor wenigen Jahren war diese Gesetzmäßigkeit nicht einmal allen Statistikern bekannt. Erst seit der US-amerikanische Mathematiker Theodore Hill versucht hat, die Benford-Verteilung zur Lösung praktischer Probleme nutzbar zu machen, ist ihr Bekanntheitsgrad gewachsen.(`Quelle`: »Wikipedia«)

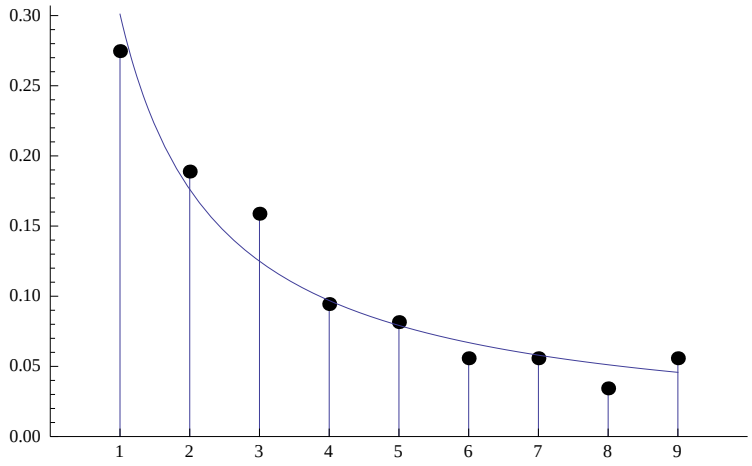
HÄUFIGKEITSVERTEILUNG DER ANFANGSZIFFERN VON ZAHLEN:

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:

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

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 Häu-

figkeiten verglichen.



1.2 Diskrete Zufallsvariablen und ihre Verteilung

Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum.

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

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

so dass für alle $a \in S$ gilt:

$$X^{-1}(a) := \{\omega \in \Omega \mid X(\omega) = a\} \in \mathcal{A}. \tag{1.2.1}$$

Für $X^{-1}(a)$ (das Urbild von a unter X) schreiben wir im folgenden $\{X = a\}$.

*ii) Die **Verteilung** von X ist die Wahrscheinlichkeitsverteilung μ_X auf S mit Gewichten*

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

Für $P[\{X = a\}]$ schreiben wir im folgenden $P[X = a]$.

Bemerkung. a) In der Tat ist p_X Massenfunktion einer Wahrscheinlichkeitsverteilung (siehe Satz 1.2):

i) Für alle $a \in S$ gilt: $\quad p_X(a) \geq 0$

ii) Da die Ereignisse $\{X = a\}$ disjunkt sind, folgt:

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

b) Für $B \subseteq S$ gilt:

$$\{X \in B\} := \underbrace{\{\omega \in \Omega | X(\omega) \in B\}}_{X^{-1}(B)} = \bigcup_{a \in B} \underbrace{\{X = a\}}_{\in \mathcal{A}} \in \mathcal{A}, \qquad \text{sowie}$$

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

Die Verteilung μ_X gibt also an, mit welchen Wahrscheinlichkeiten die Zufallsvariable X Werte in bestimmten Mengen annimmt.

Beispiele (Zweimal würfeln)**.** Sei $\Omega = \{\omega = (\omega_1, \omega_2) \mid \omega_i \in \{1, \dots, 6\}\}$ und sei P die Gleichverteilung.

a)

$$\text{Sei } X_i \colon \Omega \to S := \{1, 2, 3, 4, 5, 6\},$$
$$X(\omega) := \omega_i, \qquad \text{die Augenzahl des } i\text{-ten Wurfs.}$$

X_i ist eine diskrete Zufallsvariable mit Verteilung μ_{X_i}. Die Gewichte von μ_{X_i} sind

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

d.h. X_i ist gleichverteilt.

b)

$$\text{Sei } Y \colon \Omega \to \widetilde{S} := \{2, 3, \dots, 12\}$$
$$Y(\omega) := X_1(\omega) + X_2(\omega), \qquad \text{die Summe der Augenzahlen.}$$

Die Gewichte der Verteilung von Y sind

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

d.h. Y ist nicht mehr gleichverteilt!

Allgemeiner:

Beispiel. Sei $\Omega = \{\omega_1, \dots, \omega_n\}$ endlich, P die Gleichverteilung, $X \colon \Omega \to S$ eine Zufallsvariable und $x_i := X(\omega_i)$. Dann ist

$$P[X = a] = \frac{|\{\omega \in \Omega \mid X(\omega) = a\}|}{|\Omega|} = \frac{|\{1 \le i \le n \mid x_i = a\}|}{n},$$

also ist μ_x die empirische Verteilung von $x_1, \dots, x_n$.

Binomialverteilung

Beispiel (»Ziehen mit Zurücklegen«)**.** Wir betrachten eine endliche Grundgesamtheit (Population, Zustandsraum) S, zum Beispiel Kugeln in einer Urne, Vögel im Wald, Einwohner in NRW. Wir wollen nun die zufällige Entnahme von n Einzelstichproben mit Zurücklegen aus S beschreiben und setzen daher

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

Wir nehmen an, dass alle kombinierten Stichproben gleich wahrscheinlich sind, d.h. P sei die Gleichverteilung auf Ω.

RELEVANTE ZUFALLSVARIABLEN UND EREIGNISSE:

- i-ter Stichprobenwert:

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

  d.h. X_i ist gleichverteilt auf S.

  Sei $E \subseteq S$ eine bestimmte Merkmalsausprägung der Stichprobe, die wir im folgenden als »Erfolg« bezeichnen (zum Beispiel schwarze Kugel, Beobachtung einer Amsel). Dann können wir die Ereignisse

  $$\{X_i \in E\},\text{»Erfolg bei } i\text{-ter Stichprobe«},$$

  betrachten. Es gilt:

  $$P[X_i \in E] = \mu_{X_i}(E) = \frac{|E|}{|S|}.$$

  Wir setzen

  $$q := \frac{|E|}{|S|}, \qquad \text{»Erfolgswahrscheinlichkeit«}$$

- Häufigkeit von E / »Anzahl der Erfolge«:
  Sei nun

  $$N\colon \Omega \to \{0, 1, 2, \ldots, n\},$$
  $$N(\omega) := |\{1 \le i \le n \mid X_i(\omega) \in E\}|$$

  die Anzahl der Einzelstichproben mit Merkmalsausprägung E.

Lemma 1.4. *Für $k \in \{0, 1, \ldots, n\}$ gilt:*

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

Beweis. Es gilt

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

wobei

$$\begin{aligned}
\binom{n}{k} &= \text{Anzahl der Möglichkeiten } k \text{ Indizes aus } \{1, \ldots, n\} \text{ auszuwählen,} \\
&\quad \text{für die ein Erfolg eintritt,} \\
|E|^k &= \text{Anzahl der Möglichkeiten für jeden Erfolg,} \\
|S \backslash E|^{n-k} &= \text{Anzahl der Möglichkeiten für jeden Misserfolg.}
\end{aligned}$$

Also gilt:

$$\begin{aligned}
P[N = k] = \frac{\binom{n}{k} |E|^k \, |S \backslash E|^{n-k}}{|S|^n} &= \binom{n}{k} \left(\frac{|E|}{|S|}\right)^k \left(\frac{|S \backslash E|}{|S|}\right)^{n-k} \\
&= \binom{n}{k} p^k (1-p)^{n-k}.
\end{aligned}$$

□

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

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

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

Bemerkung. Dass $p_{n,p}$ eine Massenfunktion einer Wahrscheinlichkeitsverteilung ist, folgt aus Lemma 1.3!

Bemerkung. Ereignisse $E_1, \ldots, E_n$ heißen **unabhängig**, falls

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

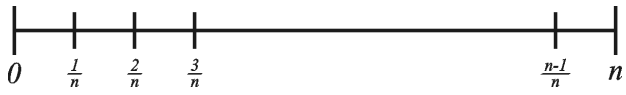
für alle $k \leq n$ und $1 \leq i_1 < i_2 < \ldots < i_k \leq n$ gilt.
Sind $E_1, \ldots, E_n$ unabhängig und $P[E_i] = p$, dann ist

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

d.h. die Anzahl der Ereignisse, die eintreten, ist binomialverteilt. Der Beweis folgt weiter unten.

Poissonverteilung

Beispiel (Warteschlange)**.** Angenommen, die Kunden in einer Warteschlange kommen **unabhängig voneinander** zu **zufälligen** (gleichverteilten) Zeitpunkten. Wie viele Kunden kommen in einer Zeitspanne der Länge t_0 an? Sei N die Anzahl dieser Kunden und $t_0 = 1$. Wir unterteilen das Intervall $[0, 1]$:



Wir machen die folgende Annahme (die natürlich in zu modellierenden Anwendungsproblemen zu überprüfen ist):

> »Wenn n sehr groß ist, dann kommt in einer Zeitspanne der Länge $\frac{1}{n}$ fast immer höchstens ein Kunde«.

E_i stehe für das Ereignis, dass ein Kunde im Zeitintervall $\left[\frac{i-1}{n}, \frac{i}{n}\right]$ ankommt ($1 \leq i \leq n$). Wir nehmen außerdem an, dass die Wahrscheinlichkeit unabhängig von i und näherungsweise proportional zu $\frac{1}{n}$ ist, also:

$$P[E_i] \approx \frac{\lambda}{n}\,, \qquad \lambda \in (0, \infty).$$

Für das Ereignis, dass genau k Kunden im Zeitintervall $[0, 1]$ ankommen, sollte dann gelten, dass

$$P[N = k] \approx P[\text{»genau } k \text{ der } E_i \text{ treten ein«}] \approx p_{n,\frac{\lambda}{n}}(k),$$

wobei $p_{n,\frac{\lambda}{n}}(k)$ das Gewicht von k unter der Binomialverteilung mit Parametern n und $\frac{\lambda}{n}$ ist. Diese Näherung sollte **»für große n immer genauer werden«**.

Satz 1.5 (Poissonapproximation der Binomialverteilung)**.** *Sei $\lambda \in (0, \infty)$. Dann gilt:*

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

Beweis. Es gilt:

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

□

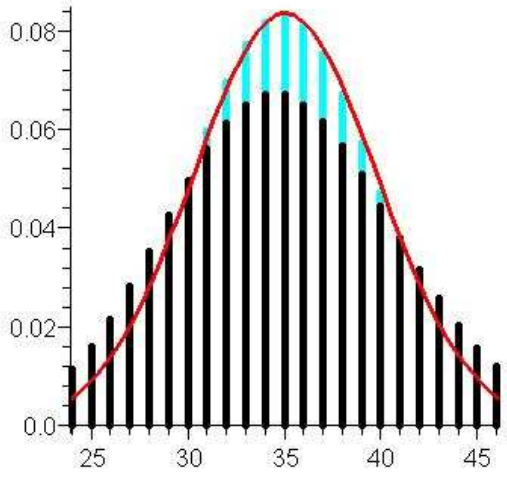
Definition. *Die Wahrscheinlichkeitsverteilung auf* $\{0, 1, 2, \ldots\}$ *mit Massenfunktion*

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

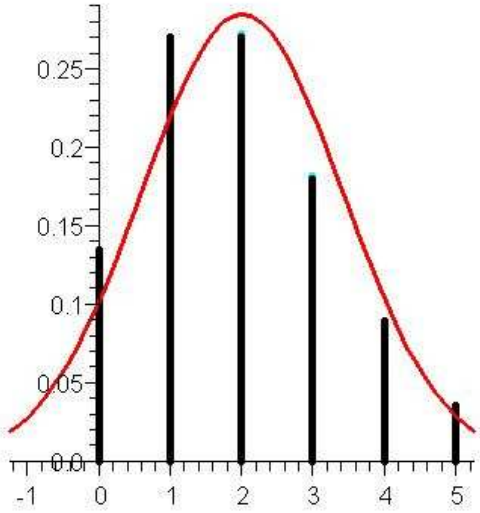
heißt ***Poissonverteilung mit Parameter*** λ.

Aufgrund des Satzes verwendet man die Poissonverteilung zur näherungsweisen Modellierung der Häufigkeit seltener Ereignisse (zum Beispiel Tippfehler in einem Buch, Schadensfälle bei Versicherung, Zusammenbrüche des T-Mobile-Netzes, ...) und damit zur »Approximation« von Binomialverteilungen mit kleiner Erfolgswahrscheinlichkeit p.
Für häufigere Ereignisse (zum Beispiel wenn 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 »Maple« erstellten) Graphiken zeigen die Poisson- und Normalapproximation (Poisson schwarz, Normalverteilung rot) der Binomialverteilung (blau) für unterschiedliche Parameterwerte:

$n = 100,\ p = 0,35$



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



Hypergeometrische Verteilung

Beispiel (Ziehen ohne Zurücklegen)**.** Wir betrachten m Kugeln in einer Urne (Wähler, Fische im See, ...), davon r rote und $m - r$ schwarze. Gezogen wird eine zufällige Stichprobe von n

Kugeln, $n \leq \min(r, m-r)$. Sind alle Stichproben gleich wahrscheinlich, dann ist ein geeignetes Modell gegeben durch:

$$\Omega = \mathcal{P}(\{1, \ldots, m\}) = \text{ alle Teilmengen von } \{1, \ldots, m\} \text{ der Kardinalität } n,$$
$$P = \text{ Gleichverteilung auf } \Omega.$$

Wir definieren eine Zufallsvariable $N \colon \Omega \to \{1, \ldots, m\}$ durch

$$N(\omega) := \text{Anzahl der roten Kugeln in } \omega.$$

Für das Ereignis, dass genau k rote Kugeln in der Stichprobe sind, gilt:

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

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

Bemerkung. Untersucht man die Asymptotik der hypergeometrischen Verteilung für $m \to \infty$, $r \to \infty$, $p = \frac{r}{m}$ fest und n fest, so gilt:

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

d.h. die hypergeometrische Verteilung nähert sich der Binomialverteilung an. Eine anschauliche Erklärung dafür ist:
Befinden sich sehr viele Kugeln in der Urne, dann ist der Unterschied zwischen Ziehen mit und ohne Zurücklegen vernachlässigbar, da nur sehr selten dieselbe Kugel zweimal gezogen wird.

1.3 Simulation von Gleichverteilungen

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.

Konkret werden die 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. Wir betrachten einige Beispiele:

Lineare Kongruenzgeneratoren (LCG)

Bei linearen Kongruenzgeneratoren ist die Rekurrenzrelation vom Typ

$$x_{n+1} = (ax_n + c) \bmod m, \qquad n = 0, 1, 2, \ldots.$$

Hierbei sind a, c und m geeignet zu wählende positive ganze Zahlen, zum Beispiel:

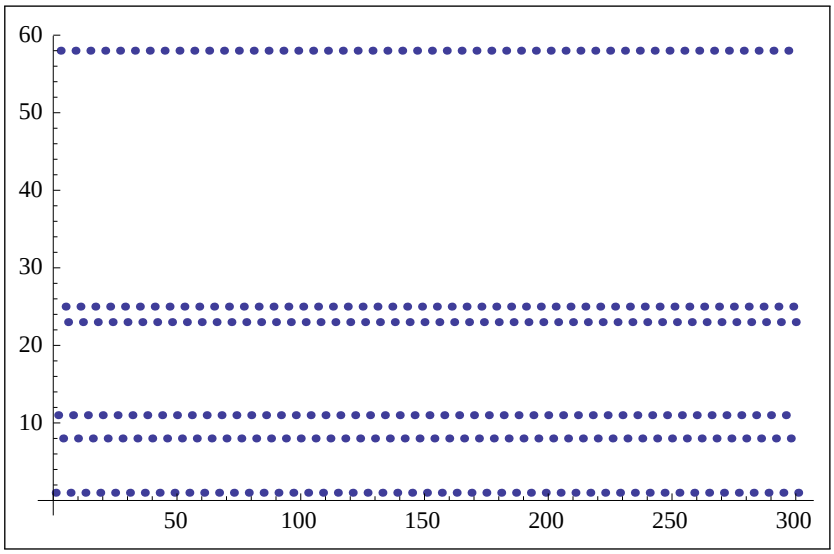
ZX81-Generator :	$m = 2^{16} + 1,$	$a = 75,$	$c = 0.$
RANDU, IBM 360/370 :	$m = 2^{31},$	$a = 65539,$	$c = 0.$
Marsaglia-Generator :	$m = 2^{32},$	$a = 69069,$	$c = 1.$
Langlands-Generator :	$m = 2^{48},$	$a = 142412240584757,$	$c = 11.$

Um einen ersten Eindruck zu erhalten, wie die Qualität der erzeugten Pseudozufallszahlen von a, c und m abhängt, implementieren wir die Generatoren mit »Mathematica«:

```
f[x_] := Mod[a x + c, m]
```

Beispiel. Wir beginnen zur Demonstration mit dem Beispiel eines ganz schlechten LCG:

```
a = 11; c = 0; m = 63;
pseudorandomdata = NestList[f, 1, 300];
ListPlot[pseudorandomdata]
```



Die Folge von Zufallszahlen ist in diesem Fall periodisch mit einer Periode, die viel kleiner ist als die maximal mögliche (63). Dies rechnet man auch leicht nach.

Periodizität mit Periode kleiner als m kann man leicht ausschließen. Es gilt nämlich:

Satz (Knuth)**.** *Die Periode eines LCG ist gleich m genau dann, wenn*

i) c *und* m *teilerfremd sind,*

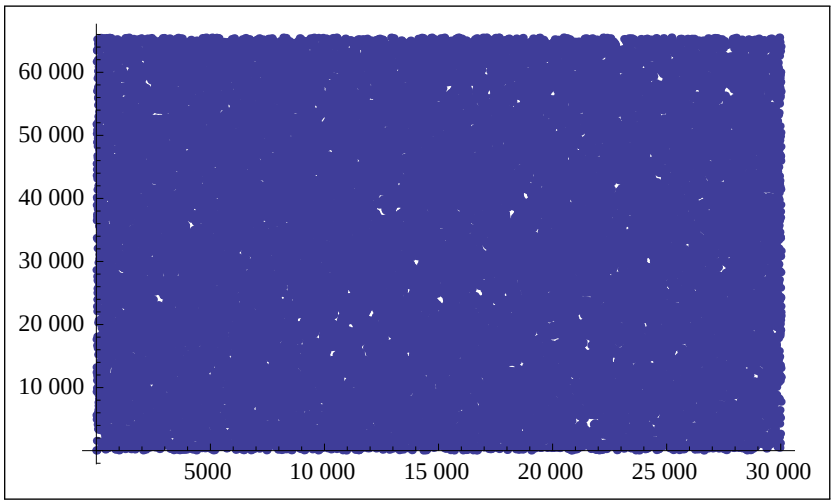
ii) jeder Primfaktor von m *ein Teiler von* $a - 1$ *ist, und*

iii) falls 4 *ein Teiler von* m *ist, dann auch von* $a - 1$.

Beweis. siehe `D. Knuth`: »The art of computer programming, Vol. 2.« □

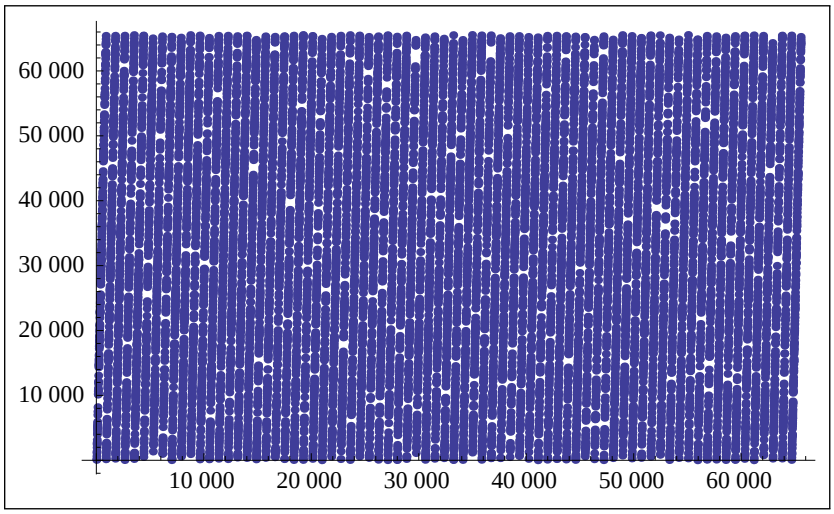
Beispiel (ZX 81-Generator)**.** Hier ergibt sich ein besseres Bild, solange wir nur die Verteilung der einzelnen Zufallszahlen betrachten:

```
a = 75; c = 0; m = 2^16 + 1;
pseudorandomdata = NestList[f, 1, 30000];
ListPlot[pseudorandomdata]
```



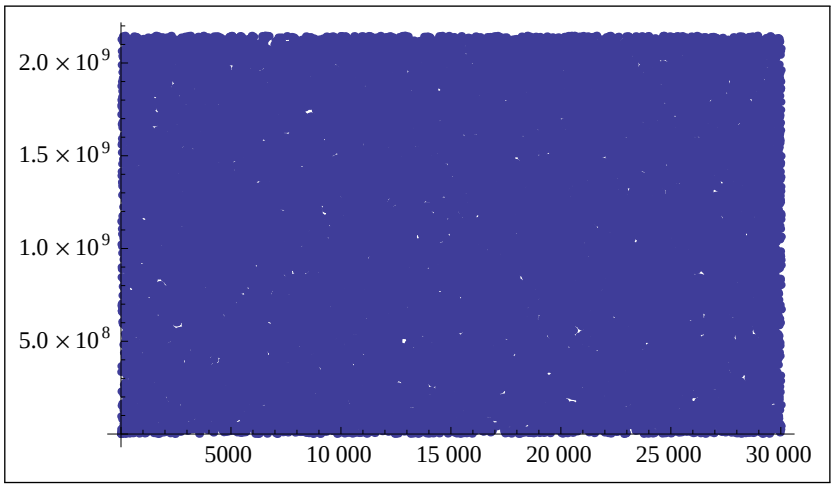
Fassen wir jedoch Paare (x_i, x_{i+1}) von aufeinanderfolgenden Pseudozufallszahlen als Koordinaten eines zweidimensionalen Pseudozufallsvektors auf, und betrachten die empirische Verteilung dieser Vektoren, so ergibt sich keine besonders gute Approximation einer zweidimensionalen Gleichverteilung:

```
blocks = Partition[pseudorandomdata, 2]; ListPlot[blocks]
```

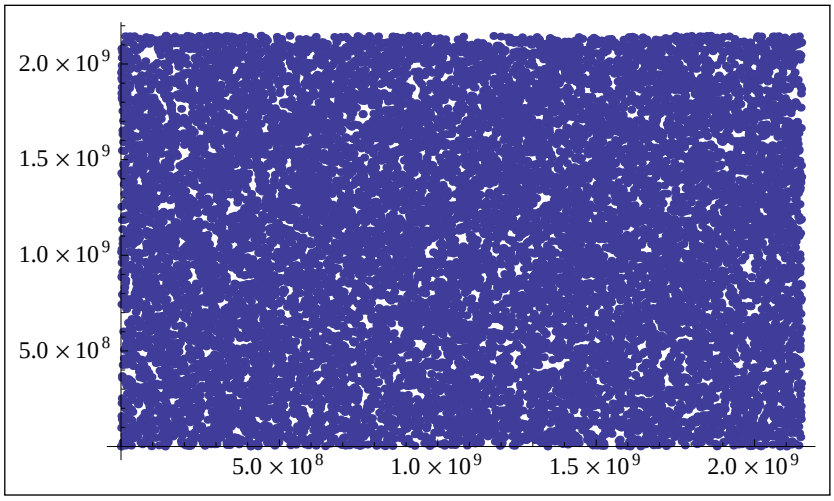


Beispiel (RANDU)**.** Hier scheinen sowohl die einzelnen Pseudozufallszahlen x_i als auch die Vektoren (x_i, x_{i+1}) näherungsweise gleichverteilt zu sein:

```
a = 65539; c = 0; m = 2^31;
pseudorandomdata = NestList[f, 1, 30000];
ListPlot[pseudorandomdata]
```

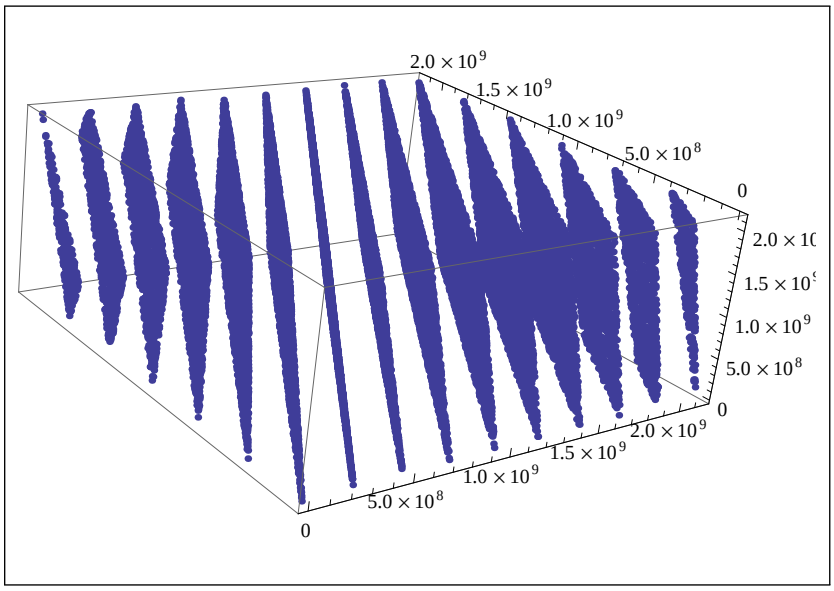



```
blocks = Partition[pseudorandomdata, 2]; ListPlot[blocks]
```



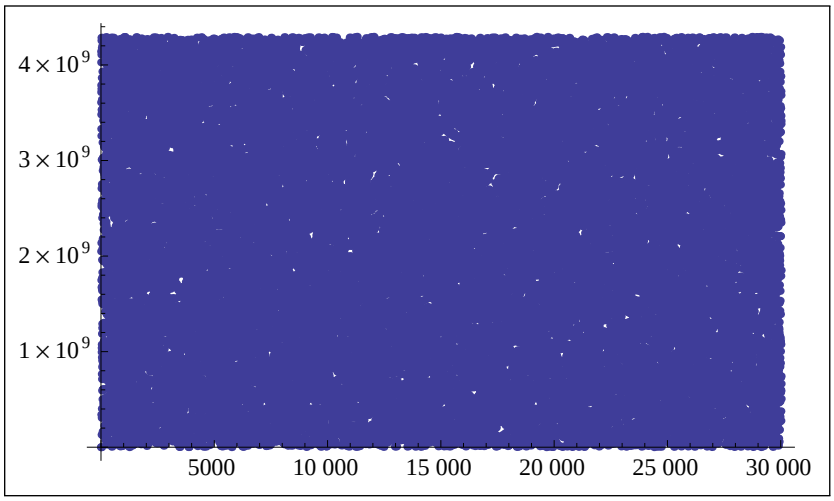
Fassen wir aber jeweils drei aufeinanderfolgende Pseudozufallszahlen als Koordinaten eines Vektors
(x_i, x_{i+1}, x_{i+2}) im $\mathbb{Z}^3$auf, dann ist die empirische Verteilung dieser Pseudozufallsvektoren keine Gleichverteilung mehr, sondern konzentriert sich auf nur 15 zweidimensionalen Hyperebenen:

```
blocks3 = Partition[pseudorandomdata, 3]; ListPointPlot3D[blocks3]
```

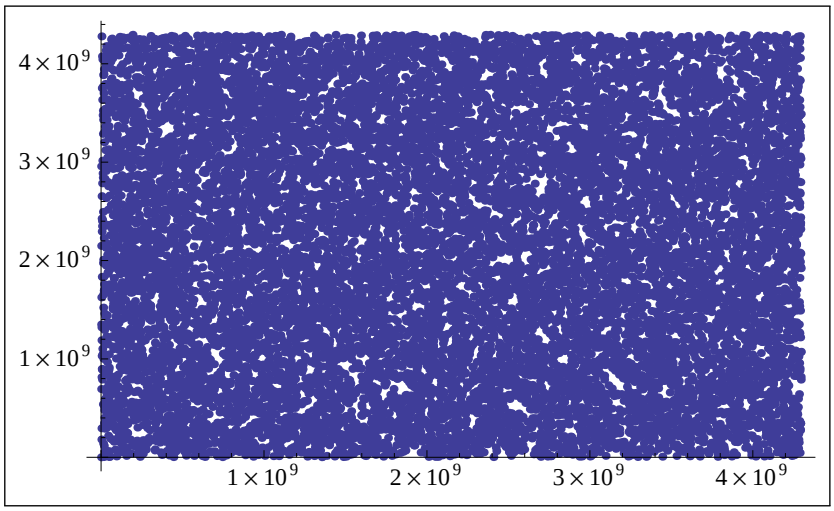


Beispiel (Marsaglia-Generator)**.** Der von Marsaglia 1972 vorgeschlagene LCG besteht dagegen alle obigen Tests (und einige weitere):

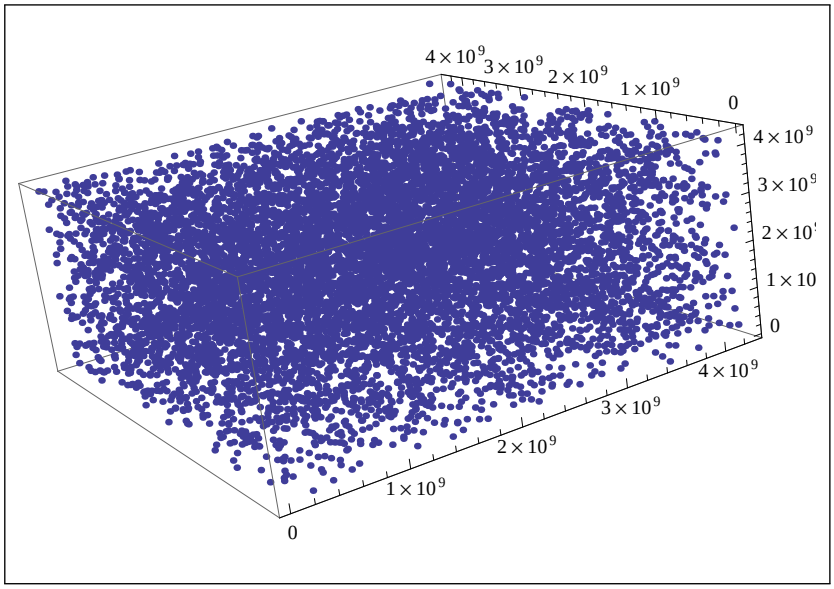
```
a = 60069; c = 1; m = 2^32;
pseudorandomdata = NestList[f, 1, 30000];
ListPlot[pseudorandomdata]
```




```
blocks = Partition[pseudorandomdata, 2];
ListPlot[blocks]
```




```
blocks3 = Partition[pseudorandomdata, 3];
ListPointPlot3D[blocks3]
```



Dies bedeutet natürlich nicht, daß die vom Marsaglia-Generator erzeugte Folge eine für **alle** Zwecke akzeptable Approximation einer Folge von unabhängigen Stichproben von der Gleichverteilung ist. Da die Folge in Wirklichkeit deterministisch ist, kann man einen Test konstruieren, der sie von einer echten Zufallsfolge unterscheidet.

Shift-Register-Generatoren

Bei Shift-Register-Generatoren 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

Zufallszahlengeneratoren lassen sich kombinieren, zum Beispiel indem man die von mehreren Zufallszahlengeneratoren erzeugten Folgen von Pseudozufallszahlen aus $\{0,1,\ldots,m-1\}$ modulo m addiert. Auf diese Weise erhält man sehr leistungsfähige Zufallszahlengeneratoren, zum Beispiel den Kiss-Generator von Marsaglia, der einen LCG und zwei Shift-Register-Generatoren kombiniert, Periode 2^{95} hat, und umfangreiche statistische Tests besteht.

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 Dezimalzahlen 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 $\{0, 1, \ldots, m - 1\}$ ist. In »Mathematica« kann man mit

```
RandomReal[spec, WorkingPrecision → k]
```

pseudozufällige Dezimalzahlen mit einer beliebigen vorgegebenen Anzahl k von Nachkommastellen erzeugen.

Zufallspermutationen

Der folgende Algorithmus erzeugt eine (pseudo-)zufällige Permutation aus S_n :

Algorithmus 1.6 (RPERM)**.**

```
rperm[n_] :=
 Module[{x = Range[n], k, a},   Beginn mit Liste 1,2,...,n
  Do[
   k = RandomInteger[{i, n}];
   a = x[[i]]; x[[i]] = x[[k]]; x[[k]] = a;    (Vertausche x[[i]] und x[[k]])
   , {i, n − 1}];   (Schleife, i läuft von 1 bis n − 1)
  x       (Ausgabe von x) ]
```

```
rperm[17]
{12, 5, 13, 8, 17, 9, 10, 6, 1, 7, 16, 15, 14, 4, 2, 3, 11}
```

ÜBUNG:
Sei $\Omega_n = \{1, 2, \ldots, n\} \times \{2, 3, \ldots, n\} \times \cdots \times \{n-1, n\}$.

a) 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 Ω_n nach S_n ist ($\tau_{i,j}$ bezeichnet die Transposition von i und j).

b) Folgern Sie, daß der Algorithmus oben tatsächlich eine Stichprobe einer gleichverteilten Zufallspermutation aus S_n simuliert.

1.4 Erwartungswert

Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum, $S \subseteq \mathbb{R}$ abzählbar und $X : \Omega \to S$ eine Zufallsvariable auf $(\Omega, \mathcal{A}, P)$.

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

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

sofern die Summe auf der rechten Seite wohldefiniert ist (d.h. unabhängig von der Abzählung von S).

Bemerkung. a) Falls $X(\omega) \geq 0$ für alle $\omega \in \Omega$ gilt, sind alle Summanden der Reihe nichtnegativ und der Erwartungswert $E[X] \in [0, \infty]$ wohldefiniert.

b) Falls die Reihe absolut konvergiert, d.h. falls $\sum_{a \in S} |a| \cdot P[X = a]$ endlich ist, ist der Erwartungswert $E[X] \in \mathbb{R}$ wohldefiniert.

$E[X]$ kann als der **Prognosewert** oder **(gewichteter) Mittelwert** für $X(\omega)$ interpretiert werden.

Beispiel (Indikatorfunktion eines Ereignisses $A \in \mathcal{A}$)**.** Sei

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

Dann ist der Erwartungswert

$$E[X] = 1 \cdot P[X = 1] + 0 \cdot P[X = 0] = P[A].$$

Ein Beispiel dafür ist ein elementarer Versicherungskontrakt mit Leistung

$$Y = \begin{cases} c & \text{falls } \omega \in A, \qquad \text{»Schadensfall«,} \\ 0 & \text{sonst.} \end{cases}$$

Dann gilt:

$$Y = c \cdot I_A \quad \text{und} \quad E[Y] = c \cdot P[A].$$

Beispiel (Poissonverteilung)**.** Sei X Poisson-verteilt mit Parameter λ. Dann ist der Erwartungswert

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

Wir können daher den Parameter λ als Erwartungswert oder die mittlere Häufigkeit des Ereignisses interpretieren.

Transformationssatz

Sei nun S eine beliebige abzählbare Menge, $g : S \to \mathbb{R}$ eine Funktion und $X : \Omega \to S$ eine Zufallsvariable. Wir definieren

$$g(X) \colon \Omega \to \mathbb{R},$$
$$\omega \mapsto g(X(\omega)).$$

$g(X)$ ist eine **reellwertige Zufallsvariable**.

Satz 1.7 (Transformationssatz)**.** *Es gilt*

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

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

Beweis. Es gilt mit Verwendung der σ-Additivität

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

□

Bemerkung. a) Insbesondere gilt:

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

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

b) Ist Ω abzählbar, dann folgt für $X : \Omega \to \mathbb{R}$:

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

wobei id_Ω die identische Abbildung auf Ω bezeichnet. Der Erwartungswert ist das **gewichtete Mittel**. Ist P die Gleichverteilung auf Ω, folgt weiter:

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

Der Erwartungswert ist in diesem Spezialfall das **arithmetische Mittel** .

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, dann ist das Spiel beendet.Wie hoch wäre eine faire Teilnahmegebühr für dieses Spiel?
Der Gewinn ist

$$\begin{aligned} &G(\omega) = 2^{T(\omega)}, && \text{mit} \\ &T(\omega) := \min\{n \in \mathbb{N} \mid X_n(\omega) = 1\}, && \text{der Wartezeit auf »Kopf«.} \end{aligned}$$

Für den erwarteten Gewinn ergibt sich

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

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 $u(c) = E[u(G)]$ gilt.

Linearität und Monotonie des Erwartungswertes

Satz 1.8 (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 $(\Omega, \mathcal{A}, P)$, für die $E[|X|]$ und $E[|Y|]$ endlich sind, dann gilt:*

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

Beweis. Wir definieren $g \colon S_X \times S_Y \to \mathbb{R}, \quad (x, y) \mapsto \lambda x + \mu y$. Dann ist $g(X, Y) = \lambda X + \mu Y$ eine Zufallsvariable mit Werten in $S_X \times S_Y$. Mit dem Transformationssatz folgt:

$$\begin{aligned}
E[\lambda X + \mu Y] &= E[g(X, Y)] \\
&= \sum_{a \in S_X} \sum_{b \in S_Y} g(a, b)\, P[X = a, Y = b] \qquad (1.4.1) \\
&= \sum_{a \in S_X} \sum_{b \in S_Y} (\lambda a + \mu b)\, P[X = a, Y = b] \\
&= \lambda \sum_{a \in S_X} a \sum_{b \in S_Y} P[X = a, Y = b] + \mu \sum_{b \in S_Y} b \sum_{a \in S_X} P[X = a, Y = b] \\
&= \lambda \sum_{a \in S_X} a\, P[X = a] + \mu \sum_{b \in S_Y} b\, P[Y = b] \\
&= \lambda E[X] + \mu E[Y].
\end{aligned}$$

Hierbei konvergiert die Reihe in (1.4.1) absolut, da

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

nach Voraussetzung endlich ist. □

Korollar (Monotonie des Erwartungswerts)**.** *Seien die Voraussetzungen von Satz 1.8 erfüllt. Sei zusätzlich $X(\omega) \leq Y(\omega)$ für alle $\omega \in \Omega$, dann gilt:*

$$E[X] \leq E[Y].$$

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

$$0 \leq E[Y - X] = E[Y] - E[X].$$

□

Beispiele (Unabhängige 0-1-Experimente)**.** Seien $A_1, A_2, \ldots, A_n \in \mathcal{A}$ unabhängige Ereignisse mit Wahrscheinlichkeit p, und sei

$$X_i = I_{A_i}, \qquad \text{die Indikatorfunktion des Ereignisses } A_i,\ i = 0, \ldots, n.$$

a) Die Zufallsvariablen X_i sind **Bernoulli-verteilt mit Parameter** p, d.h.

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

Also gilt:

$$E[X_i] = E[I_{A_i}] = P[A_i] = p,$$

analog zu Beispiel 1.4.

b) Die Anzahl

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

der Ereignisse, die eintreten, ist binomialverteilt mit Parametern n und p (siehe Übung), d.h.

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

Den Erwartungswert kann man daher wie folgt berechnen:

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

Einfacher benutzt man aber die Linearität des Erwartungswerts, und erhält

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

sogar **ohne Verwendung der Unabhängigkeit**!

Beispiel (Abhängige 0-1-Experimente)**.** Wir betrachten eine Population aus m Objekten, davon r rote, aus der eine Zufallsstichprobe aus n Objekten ohne Zurücklegen entnommen wird, $n \leq \min(r, m-r)$. Sei A_i das Ereignis, dass das i-te Objekt in der Stichprobe rot ist, und $X_i = I_{A_i}$. Die Anzahl

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

der roten Objekte in der Zufallsstichprobe ist dann hypergeometrisch verteilt mit Parametern m, r und n. Als Erwartungswert dieser Verteilung erhalten wir analog zum letzten Beispiel:

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

Beispiel (Inversionen von Zufallspermutationen)**.** Seien $\Omega = S_n$ die Menge aller Permutationen $\omega\colon \{1,\dots,n\} \to \{1,\dots,n\}$, P die Gleichverteilung auf Ω, und

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

die Anzahl der Inversionen einer Permutation $\omega \in \Omega$. Dann gilt

$$N = \sum_{1 \le i < j \le n} I_{A_{i,j}}, \qquad \text{wobei}$$
$$A_{i,j} = \{\omega \in S_n \mid \omega(i) > \omega(j)\}$$

das Ereignis ist, dass eine Inversion von i und j auftritt. Es folgt:

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

ANWENDUNG: Beim Sortieralgorithmus »Insertion Sort« wird der Wert $\omega(i)$ einer Liste $\{\omega(1), \omega(2), \dots, \omega(n)\}$ beim Einfügen von $\omega(j)$ genau dann verschoben, wenn $\omega(j) < \omega(i)$ gilt. Ist die Anfangsanordnung eine Zufallspermutation der korrekten Anordnung, dann ist die mittlere Anzahl der Verschiebungen, die der Algorithmus vornimmt, also gleich $\frac{n\,(n-1)}{4}$.

Satz 1.9 (Einschluss-/Ausschlussprinzip)**.** *Für $n \in \mathbb{N}$ und Ereignisse $A_1, \dots, A_n \in \mathcal{A}$ gilt:*

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

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

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

Es folgt:

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

□

Kapitel 2

Bedingte Wahrscheinlichkeiten und Unabhängigkeit

2.1 Bedingte Wahrscheinlichkeiten

Sei (Ω, $\mathcal{A}$, P) ein Wahrscheinlichkeitsraum und $A, B \in \mathcal{A}$ Ereignisse. Was ist die Wahrscheinlichkeit dafür, dass A eintritt, wenn wir schon wissen, dass B eintritt?

Relevante Fälle:	$\omega \in B$
Davon günstige Fälle:	$\omega \in A \cap B$

Definition. *Sei $P[B] \neq 0$. Dann heißt*

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

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

Bemerkung. a) $P[\,\bullet\,|B] : A \mapsto P[A|B]$ ist eine Wahrscheinlichkeitsverteilung auf $(\Omega, \mathcal{A})$, die **bedingte Verteilung gegeben** B . Der Erwartungswert

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

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

b) Ist P die Gleichverteilung auf einer endlichen Menge Ω, dann gilt:

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

Beispiele. a) Wir betrachten eine Familie mit 2 Kindern, und stellen die Frage nach dem Geschlecht der Kinder. Sei daher

$$\Omega = \{JJ, JM, MJ, MM\}.$$

Angenommen, alle Fälle wären gleich wahrscheinlich. Dann gilt:

$$P[\text{»beide Mädchen«} \mid \text{»eines Mädchen«}] = \frac{|\{MM\}|}{|\{MM, JM, MJ\}|} = \frac{1}{3},$$

$$P[\text{»beide Mädchen«} \mid \text{»das erste ist Mädchen«}] = \frac{|\{MM\}|}{|\{MM, MJ\}|} = \frac{1}{2}.$$

In Wirklichkeit sind die Kombinationen JJ und MM wahrscheinlicher.

b) Bei 20 fairen Münzwürfen fällt 15-mal »Zahl«. Wie groß ist die Wahrscheinlichkeit, dass die ersten 5 Würfe »Zahl« ergeben haben? Sei

$$\Omega = \{\omega = (x_1, \dots x_{20}) \mid x_i \in \{0, 1\}\}, \qquad \text{und}$$
$$X_i(\omega) = x_i, \qquad \text{der Ausgang des } i\text{-ten Wurfs.}$$

Es gilt:

$$P\big[X_1 = \ldots = X_5 = 1 \,\big|\, \sum_{i=1}^{20} X_i = 15\big] = \frac{P\big[X_1 = \ldots = X_5 = 1 \text{ und } \sum_{i=6}^{20} X_i = 10\big]}{P\big[\sum_{i=1}^{20} X_i = 15\big]}$$
$$= \frac{2^{-5} \cdot 2^{-15} \binom{15}{10}}{2^{-20} \binom{20}{15}} = \frac{15 \cdot 14 \cdot \dots \cdot 11}{20 \cdot 19 \cdot \dots \cdot 16} \approx \frac{1}{5}.$$

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

Berechnung von Wahrscheinlichkeiten durch Fallunterscheidung

Sei $\Omega = \bigcup H_i$ eine disjunkte Zerlegung von Ω in abzählbar viele Fälle (»Hypothesen«) H_i , $i \in I$.

Satz 2.1 (Formel von der totalen Wahrscheinlichkeit). *Für alle $A \in \mathcal{A}$ gilt:*

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

Beweis. Es ist $A = A \cap (\bigcup_{i\in I} H_i) = \bigcup_{i\in I}(A \cap H_i)$ eine disjunkte Vereinigung, also gilt nach σ-Additivität:

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

□

Beispiel. 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?

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

Beispiel (Simpson-Paradoxon)**.** Bewerbungen in Berkeley:

BEWERBUNGEN IN BERKELEY

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

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 $P_M[A] := P[A|M]$ die empirische Verteilung unter Männern und $P_F[A] := P[A|F]$ die empirische Verteilung unter Frauen, angenommen zu werden. Die Aufgliederung nach Fachbereichen ergibt folgende Zerlegung in Hypothesen:

$$P_M[A] = \sum_{i=1}^{4} P_M[A|H_i]\, P_M[H_i], \qquad P_F[A] = \sum_{i=1}^{4} P_F[A|H_i]\, P_F[H_i].$$

Im Beispiel ist $P_F[A|H_i] > P_M[A|H_i]$ für **alle** i, aber **dennoch** $P_F[A] < P_M[A]$. Die erste Statistik vermischt verschiedene Populationen und legt deshalb eventuell eine falsche Schlussfolgerung nahe.

Bayessche Regel

Wie wahrscheinlich sind die Hypothesen H_i? Ohne zusätzliche Information ist $P[H_i]$ die Wahrscheinlichkeit von H_i. In der Bayesschen Statistik interpretiert man $P[H_i]$ 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 $P[A] \neq 0$ eintritt, und wir kennen die bedingte Wahrscheinlichkeit (»likelihood«) $P[A|H_i]$ für das Eintreten von A unter der Hypothese H_i für jedes $i \in I$ mit $P[H_i] \neq 0$. Wie sieht dann unsere neue Einschätzung der Wahrscheinlichkeiten der H_i (»a posteriori degree of belief«) aus?

Korollar (Bayessche Regel)**.** *Für $A \in \mathcal{A}$ mit $P[A] \neq 0$ gilt:*

$$P[H_i|A] = \frac{P[A|H_i] \cdot P[H_i]}{\sum\limits_{\substack{j \in I \\ P[H_j] \neq 0}} P[A|H_j] \cdot P[H_j]} \qquad \textit{für alle } i \in I \textit{ mit } P[H_i] \neq 0\textit{, d.h.}$$

$$P[H_i|A] = c \cdot P[H_i] \cdot P[A|H_i],$$

wobei c eine von i unabhängige Konstante ist.

Beweis. Es gilt:

$$P[H_i|A] = \frac{P[A \cap H_i]}{P[A]} = \frac{P[A|H_i] \cdot P[H_i]}{\sum\limits_{\substack{j \in I \\ P[H_j] \neq 0}} P[A|H_j] \cdot P[H_j]}.$$

□

Beispiel. Von 10.000 Personen eines Alters habe einer die Krankheit K. Ein Test sei positiv $(+)$ bei 96% der Kranken und 0,1% der Gesunden.

A priori:	$P[K]$	=	$\frac{1}{10000}$.	$P[K^C]$	=	$\frac{9999}{10000}$.
Likelihood:	$P[+\|K]$	=	$0,96$.	$P[+\|K^C]$	=	$0,001$.

A posteriori:

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

Daraus folgt insbesondere: $P[K^C|+] \approx \frac{10}{11}$, d.h. ohne zusätzliche Informationen muss man davon ausgehen, dass $\frac{10}{11}$ der positiv getesteten Personen in Wirklichkeit gesund sind!

2.2 Mehrstufige diskrete Modelle

Wir betrachten ein n-stufiges Zufallsexperiment. Sind $\Omega_1, \ldots, \Omega_n$ abzählbare Stichprobenräume der Teilexperimente, dann können wir

$$\Omega = \Omega_1 \times \ldots \times \Omega_n = \{(\omega_1, \ldots, \omega_n) \mid \omega_i \in \Omega_i\}$$

als Stichprobenraum des Gesamtexperiments auffassen und setzen $\mathcal{A} = \mathcal{P}(\Omega)$. Für $\omega \in \Omega$ und $k = 1, \ldots, n$ sei

$$X_k(\omega) = \omega_k, \qquad \text{der Ausgang des } k\text{-ten Teilexperiments.}$$

Angenommen, wir kennen

$$P[X_1 = x_1] = p_1(x_1), \qquad \text{für alle } x_1 \in \Omega_1, \tag{2.2.1}$$

die Verteilung (Massenfunktion) von X_1, sowie

$$P[X_k \ = x_k \mid X_1 = x_1, \ldots, X_{k-1} = x_{k-1}] = p_k(x_k \mid x_1, \ldots, x_{k-1}), \tag{2.2.2}$$

die bedingte Verteilung von X_k gegeben $X_1, \ldots, X_{k-1}$ für $k = 2, \ldots n$, $x_i \in \Omega_i$ mit $P[X_1 = x_1, \ldots, X_{k-1} = x_{k-1}] \neq 0$.
Wie sieht die gesamte Wahrscheinlichkeitsverteilung P auf Ω aus?

Satz 2.2. *Seien p_1 und $p_k(\,\bullet \mid x_1, \ldots, x_{k-1})$ für jedes $k = 2, \ldots, n$ und $x_1 \in \Omega_1, \ldots, x_{k-1} \in \Omega_{k-1}$ die Massenfunktion einer Wahrscheinlichkeitsverteilung auf Ω_k. Dann existiert genau eine Wahrscheinlichkeitsverteilung P auf $(\Omega, \mathcal{A})$ mit* (2.2.1) *und* (2.2.2)*. Diese ist bestimmt durch die Massenfunktion*

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

Beweis.

- EINDEUTIGKEIT:
  Wir behaupten, dass für eine Verteilung P mit (2.2.1) und (2.2.2) gilt:

  $$P[X_1 = x_1, \ldots, X_k = x_k] = p_1(x_1) \cdot p_2(x_2 \mid x_1) \cdots p_k(x_k \mid x_1, \ldots, x_{k-1}), \qquad k = 1, \ldots, n.$$

Der Induktionsanfang folgt aus Bedingung (2.2.1). Sei die Induktionsbehauptung für $k-1$ wahr, dann folgt nach Induktionsannahme und (2.2.2):

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

falls $P[X_1 = x_1, \ldots, X_{k-1} = x_{k-1}] \neq 0$. Andernfalls verschwinden beide Seiten und die Behauptung folgt. Für $k = n$ erhalten wir als Massenfunktion von P:

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

- EXISTENZ:
  p ist Massenfunktion einer Wahrscheinlichkeitsverteilung P auf $\Omega_1 \times \cdots \times \Omega_n$, denn:

$$\begin{aligned} \sum_{x_1 \in \Omega_1} \cdots \sum_{x_n \in \Omega_n} p(x_1, \ldots, x_n) &= \sum_{x_1 \in \Omega_1} p_1(x_1) \sum_{x_2 \in \Omega_2} p_2(x_2 \mid x_1) \ldots \underbrace{\sum_{x_n \in \Omega_n} p_n(x_n \mid x_1, \ldots, x_n)}_{=1} \\ &= 1. \end{aligned}$$

  Für P gilt:

$$\begin{aligned} P[X_1 = x_1, \ldots X_k = x_k] &= \sum_{x_{k+1} \in \Omega_{k+1}} \cdots \sum_{x_n \in \Omega_n} p(x_1, \ldots, x_n) \\ &= p_1(x_1)\, p_2(x_2 \mid x_1) \cdots p_k(x_k \mid x_1, \ldots, x_{k-1}), \qquad k = 1, \ldots, n. \end{aligned}$$

  Damit folgen (2.2.1) und (2.2.2).

□

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

$$\begin{aligned} \Omega &= \{(\omega_1, \omega_2, \omega_3) \mid \omega_i \in \{0, 1, 2, 3, 4\}\}, \\ X_i(\omega) &= \omega_i = \text{Anzahl der Buben von Spieler } i. \end{aligned}$$

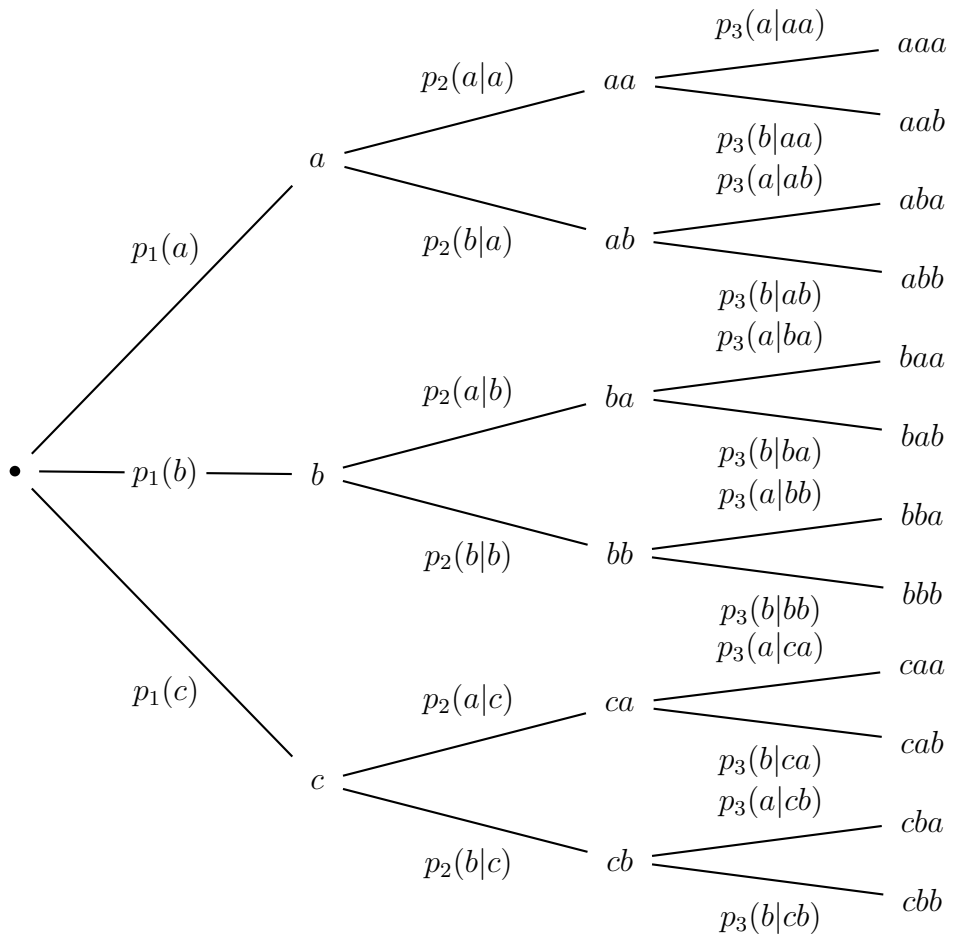


Abbildung 2.1: Baumdarstellung der Fallunterscheidungen

Es gilt:

$$p_1(x_1) = \frac{\binom{4}{x_1}\binom{28}{10-x_1}}{\binom{32}{10}}, \qquad \text{hypergeometrische Verteilung,}$$

$$p_2(x_2 \mid x_1) = \frac{\binom{4-x_1}{x_2}\binom{18+x_1}{10-x_2}}{\binom{22}{10}}$$

$$p_3(x_3 \mid x_1, x_2) = \begin{cases} \frac{\binom{4-x_1-x_2}{x_3}\binom{18+x_1+x_2}{10-x_3}}{\binom{12}{10}} & \text{falls } 2 \le x_1 + x_2 + x_3 \le 4, \\ 0 & \text{sonst.} \end{cases}$$

Damit folgt:

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

Im folgenden betrachten wir zwei fundamentale Klassen von mehrstufigen Modellen, Produktmodelle und Markov-Ketten.

Produktmodelle

Angenommen, der Ausgang des i-ten Experiments hängt nicht von $x_1, \ldots, x_{i-1}$ ab. Dann sollte gelten:

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

mit einer von $x_1, \ldots, x_{i-1}$ unabhängigen Massenfunktion p_i einer Wahrscheinlichkeitsverteilung P_i auf Ω_i. Die Wahrscheinlichkeitsverteilung P auf Ω hat dann die Massenfunktion

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p_i(x_i), \qquad x \in \Omega. \tag{2.2.3}$$

Definition. *Die Wahrscheinlichkeitsverteilung P auf $\Omega = \Omega_1 \times \ldots \times \Omega_n$ mit Massenfunktion* (2.2.3) *heißt **Produkt** von $P_1, \ldots, P_n$ und wird mit $P_1 \otimes \ldots \otimes P_n$ notiert.*

Beispiel (n-dimensionale Bernoulli-Verteilung)**.** Wir betrachten n unabhängige 0-1-Experimente mit Erfolgswahrscheinlichkeit p:

$$\Omega_1 = \ldots = \Omega_n = \{0, 1\}, \qquad p_i(1) = p, \qquad p_i(0) = 1 - p, \qquad i = 1, \ldots, n.$$

Sei $k = \sum_{i=1}^{n} x_i$ die Anzahl der Einsen. Dann ist

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

die ***n-dimensionale Bernoulli-Verteilung.***

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

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

Satz 2.3. *Im Produktmodell gilt für beliebige Ereignisse $A_i \subseteq \Omega_i, i = 1, \ldots, n$:*

$$\begin{array}{rcl} P[X_1 \in A_1, \ldots, X_n \in A_n] & = & \displaystyle\prod_{i=1}^{n} P[X_i \in A_i] \\ || & & || \\ P[A_1 \times \ldots \times A_n] & & \displaystyle\prod_{i=1}^{n} P_i[A_i] \end{array} \tag{2.2.4}$$

*(d.h. $X_1, \ldots, X_n$ sind **unabhängige** Zufallsvariablen, siehe nächsten Abschnitt).*

Beweis. Es gilt:

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

Insbesondere gilt:

$$P[X_i \in A_i] = P[X_1 \in \Omega, \ldots, X_{i-1} \in \Omega, X_i \in A_i, X_{i+1} \in \Omega, \ldots, X_n \in \Omega] = P_i[A_i].$$

□

Markov-Ketten

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

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

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 sollte gelten:

$$p_k(x_k \mid x_0, \ldots, x_{k-1}) = \underbrace{p_k(x_{k-1}, x_k)}_{\text{»Bewegungsgesetz«}}, \tag{2.2.5}$$

wobei $p_k : S \times S \to [0,1]$ 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 $x \in S$

d.h. $p_k(x, \bullet)$ ist für jedes $x \in S$ die Massenfunktion einer Wahrscheinlichkeitsverteilung auf S.

Definition. *Eine Matrix $p_k(x,y)$ $(x,y \in S)$ mit i) und ii) heißt **stochastische Matrix** (oder **stochastischer Kern**) auf S.*

Im **Mehrstufenmodell** folgt aus Gleichung (2.2.5):

$$p(x_0, x_1, \ldots, x_n) = \underbrace{p_0(x_0)}_{\text{»Startverteilung«}} \; p_1(x_0, x_1)\, p_2(x_1, x_2) \cdots p_n(x_{n-1}, x_n), \qquad \text{für } x_0, \ldots, x_n \in S.$$

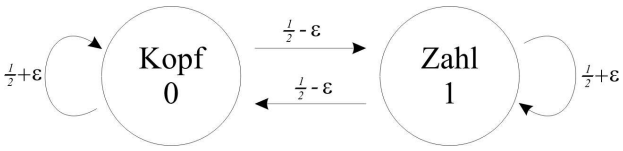
Den Fall, in dem der Übergangsmechanismus $p_k(x,y) = p(x,y)$ unabhängig von k ist, nennt man **zeitlich homogen**.

Beispiele. a) PRODUKTMODELL (siehe oben):

$$p_k(x, y) = p_k(y) \text{ ist unabhängig von } x.$$

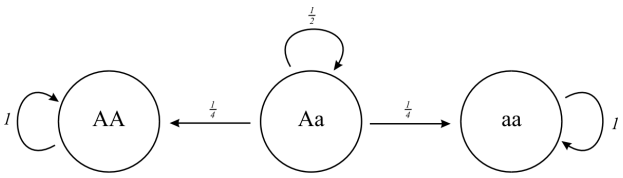
b) ABHÄNGIGE MÜNZWÜRFE:

$$S = \{0, 1\}, \qquad \varepsilon \in \big[-\frac{1}{2}, \frac{1}{2}\big].$$



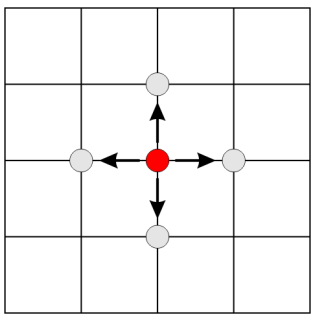
$$p = \begin{pmatrix} \frac{1}{2} + \varepsilon & \frac{1}{2} - \varepsilon \\ \frac{1}{2} - \varepsilon & \frac{1}{2} + \varepsilon \end{pmatrix}.$$

c) SELBSTBEFRUCHTUNG VON PFLANZEN:



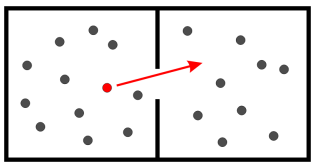
$$p = \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ 0 & 0 & 1 \end{pmatrix}$$

d) RANDOM WALK AUF $S = \mathbb{Z}^d$, ($d \in \mathbb{N}$):



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

e) URNENMODELL VON P. UND T. EHRENFEST (Austausch von Gasmolekülen in zwei Behältern):



Es seien N Kugeln auf zwei Urnen verteilt. Zu jedem Zeitpunkt $t \in \mathbb{N}$ wechselt eine zufällig ausgewählte Kugel die Urne.

MAKROSKOPISCHES MODELL:

$$S = \{0, 1, 2, \ldots, n\}.$$

$x \in S$ beschreibt die Anzahl Kugeln in der ersten Urne.

$$p(x,y) = \begin{cases} \frac{x}{n} & \text{falls } y = x-1, \\ \frac{n-x}{n} & \text{falls } y = x+1, \\ 0 & \text{sonst.} \end{cases}$$

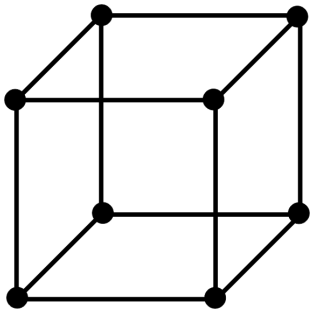
MIKROSKOPISCHES MODELL:

$$S = \{0,1\}^n = \{(\sigma_1, \ldots, \sigma_n) \mid \sigma_i \in \{0,1\}\}.$$

Es ist $\sigma_i = 1$ genau dann, wenn sich die i-te Kugel in Urne 1 befindet.

$$p(\sigma, \widetilde{\sigma}) = \begin{cases} \frac{1}{N} & \text{falls } \sum_{i=1}^n |\sigma_i - \widetilde{\sigma}_i| = 1, \\ 0 & \text{sonst.} \end{cases}$$

Die resultierende Markov-Kette ist ein Random Walk auf dem Hyperwürfel $\{0,1\}^n$, d.h. sie springt in jedem Schritt von einer Ecke des Hyperwürfels zu einer zufällig ausgewählten benachbarten Ecke.



Berechnung von Wahrscheinlichkeiten

Satz 2.4 (Markov-Eigenschaft)**.** *Für alle* $0 \le k < l \le n$ *und* $x_0, \dots, x_l \in S$ *mit* $P[X_0 = x_0, \dots, X_k = x_k] \neq 0$ *gilt:*

$$\begin{aligned} P[X_l = x_l \mid X_0 = x_0, \dots, X_k = x_k] &= P[X_l = x_l \mid X_k = x_k] \\ &= (p_{k+1}\, p_{k+2} \cdots p_l)(x_k, x_l), \end{aligned}$$

wobei

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

das Produkt der Matrizen p *und* q *ist.*

Bemerkung. a) MARKOV-EIGENSCHAFT:

Die Weiterentwicklung hängt jeweils nur vom gegenwärtigen Zustand x_k ab, und nicht vom vorherigen Verlauf $x_0, x_1, \dots, x_{k-1}$.

b) n-SCHRITT-ÜBERGANGSWAHRSCHEINLICHKEITEN:

$$\begin{aligned} P[X_n = y \mid X_0 = x] &= (p_1\, p_2 \cdots p_n)(x, y) \\ &= p^n(x, y) \qquad \text{falls } \textbf{zeitlich homogen}, \text{ d.h. } p_i \equiv p. \end{aligned}$$

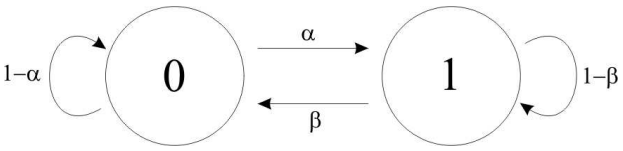
Beweis.

$$\begin{aligned} P[X_l = x_l \mid X_0 = x_0, \dots, X_k = x_k] &= \frac{P[X_0 = x_0, \dots, X_k = x_k, X_l = x_l]}{P[X_0 = x_0, \dots, X_k = x_k]} \\ &= \frac{\sum_{x_{k+1}, \dots, x_{l-1}} p_0(x_0)\, p_1(x_0, x_1) \cdots p_l(x_{l-1}, x_l)}{p_0(x_0)\, p_1(x_0, x_1) \cdots p_k(x_{k-1}, x_k)} \\ &= \sum_{x_{k+1}} \cdots \sum_{x_{l-1}} p_{k+1}(x_k, x_{k+1})\, p_{k+2}(x_{k+1}, x_{k+2}) \cdots p_l(x_{l-1}, x_l) \\ &= (p_{k+1}\, p_{k+2} \cdots p_l)(x_k, x_l). \\ P[X_l = x_l \mid X_k = x_k] &= \frac{P[X_k = x_k, X_l = x_l]}{P[X_k = x_k]} \\ &= \frac{\sum_{x_1, \dots, x_{k-1}} \sum_{x_{k+1}, \dots, x_{l-1}} p_0(x_0)\, p_1(x_0, x_1) \cdots p_l(x_{l-1}, x_l)}{\sum_{x_1, \dots, x_{k-1}} p_0(x_0)\, p_1(x_0, x_1) \cdots p_k(x_{k-1}, x_k)} \\ &= (p_{k+1}\, p_{k+2} \cdots p_l)(x_k, x_l). \end{aligned}$$

□

Beispiel.

$$S = \{0,1\}, \qquad 0 < \alpha, \beta \leq 1.$$



Für alle $n \in \mathbb{N}$ gilt:

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

Daraus folgt mit Induktion:

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

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

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

Insbesondere gilt $p^n(0,\cdot) \approx p^n(1,\cdot)$ für große $n \in \mathbb{N}$. Die Kette »vergisst« also ihren Startwert exponentiell schnell (»Exponentieller Gedächtnisverlust«)!.

2.3 Unabhängigkeit von Ereignissen

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

$$P[A|B] = P[A], \qquad \text{falls} P[B] \neq 0,$$

sowie

$$\underbrace{P[B|A]}_{\frac{P[B\cap A]}{P[A]}} = P[B], \qquad \text{falls} P[A] \neq 0,$$

also insgesamt

$$P[A \cap B] = P[A] \cdot P[B]. \tag{2.3.1}$$

Definition. *i) Zwei Ereignisse $A, B \in \mathcal{A}$ heißen **unabhängig** (bzgl. P), falls* (2.3.1) *gilt.*

*ii) Eine beliebige Kollektion A_i, $i \in I$, von Ereignissen heißt **unabhängig** (bzgl. P), falls*

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

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

Beispiele. a) Falls $P[A] \in \{0, 1\}$ gilt, ist A unabhängig von B für alle $B \in \mathcal{A}$.

b) Wir betrachten das Modell für ZWEI FAIRE MÜNZWÜRFE, also $\Omega = \{0, 1\}^2$ und P sei die Gleichverteilung. Die 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:

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

Allerdings ist die Kollektion A_1, A_2, A_3 **nicht unabhängig**, denn es gilt

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

Lemma 2.5. *Seien die Ereignisse $A_1, \ldots, A_n \in \mathcal{A}$ unabhängig, $B_j = A_j$ oder $B_j = A_j^C$ für alle $j = 1, \ldots, n$. Dann sind die Ereignisse $B_1, \ldots, B_n$ unabhängig.*

Beweis. Sei ohne Beschränkung der Allgemeinheit:

$$B_1 = A_1, \quad \ldots, \quad B_k = A_k, \quad B_{k+1} = A_{k+1}^C, \quad \ldots, \quad B_n = A_n^C$$

. Dann gilt unter Verwendung der Linearität des Erwartungswerts und der Unabhängigkeit von $A_1, \ldots, A_n$:

$$\begin{aligned} P[B_1 \cap \ldots \cap B_n] &= P\big[A_1 \cap \ldots \cap A_k \cap A_{k+1}^C \cap \ldots A_n^C\big] \\ &= E\big[I_{A_1} \cdots I_{A_k} \cdot (1 - I_{A_{k+1}}) \cdots (1 - I_{A_n})\big] \\ &= E\big[I_{A_1} \cdots I_{A_k} \cdot \sum_{J \subseteq \{k+1,\ldots,n\}} (-1)^{|J|} \prod_{j \in J} I_{A_j}\big] \\ &= \sum_{J \subseteq \{k+1,\ldots,n\}} (-1)^{|J|} P\big[A_1 \cap \ldots \cap A_k \cap \bigcap_{j \in J} A_j\big] \\ &= \sum_{J \subseteq \{k+1,\ldots,n\}} (-1)^{|J|} P[A_1] \cdots P[A_k] \cdot \prod_{j \in J} P[A_j] \\ &= P[A_1] \cdots P[A_k] \cdot (1 - P[A_{k+1}]) \ldots (1 - P[A_n]) = P[B_1] \cdots P[B_n]. \end{aligned}$$

□

Verteilungen für unabhängige Ereignisse

Seien $A_1, A_2, \ldots \in \mathcal{A}$ unabhängige Ereignisse (bzgl. P) mit $P[A_i] = p \in [0,1]$. 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

$$T(\omega) = \inf\{n \in \mathbb{N} \mid \omega \in A_n\}, \qquad \text{wobei } \min \emptyset := \infty.$$

Mit Lemma 2.5 folgt:

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

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

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

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

Bemerkung. a) Für $p \neq 0$ gilt:

$$\sum_{n=1}^{\infty} p \cdot (1-p)^{n-1} = 1,$$

d.h. die geometrische Verteilung ist eine Wahrscheinlichkeitsverteilung auf den natürlichen Zahlen, und

$$P[T = \infty] = 0.$$

b) Allgemein gilt:

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

c) Es gilt:

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

(siehe Übung).

Binomialverteilung

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

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

Es gilt:

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

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

Satz 2.6 (»Law of Averages«, Bernstein-Ungleichung)**.** *Für alle* $\varepsilon > 0$ *und* $n \in \mathbb{N}$ *gilt:*

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

Insbesondere gilt:

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

d.h. die Wahrscheinlichkeit für eine Abweichung des Mittelwerts $\frac{S_n}{n}$ *vom Erwartungswert* p *um mehr als* ε *fällt exponentiell in* n*.*

Bemerkung. a) Satz 2.6 ist eine erste Version des »Gesetzes der großen Zahlen«.

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

c) Anwendung auf Schätzen von p:

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

d) Anwendung auf näherungsweise Monte Carlo-Berechnung von p:
Simuliere n unabhängige Stichproben, $p \sim$ relative Häufigkeit.

Beweis. Sei $q := 1 - p$, $S_n \sim \mathrm{Bin}(n, p)$. Dann gilt für $\lambda > 0$:

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

Wir behaupten:

$$p\, e^{\lambda q} + q\, e^{-\lambda p} \leq e^{\frac{\lambda^2}{8}}.$$

Damit folgt:

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

Der Exponent ist minimal für $\lambda = 4\varepsilon$. Für diese Wahl von λ folgt schließlich

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

Beweis der Behauptung:

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

Zu zeigen ist nun

$$f(\lambda) \leq \frac{\lambda^2}{8} \qquad \text{für alle } \lambda \geq 0.$$

Es gilt:

$$\begin{aligned} &f(0) = 0, \\ &f'(\lambda) = -p + \frac{p\, e^{\lambda}}{p\, e^{\lambda} + q} = -p + \frac{p}{p + q\, e^{-\lambda}}, \qquad f'(0) = 0, \\ &f''(\lambda) = \frac{p\, q\, e^{-\lambda}}{(p + q\, e^{-\lambda})^2} \leq \frac{1}{4}. \end{aligned}$$

Die letzte Ungleichung folgt aus::

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

Damit folgt

$$\begin{aligned} f(\lambda) &= f(0) + \int_0^\lambda f'(x)\,dx \\ &= \int_0^\lambda \int_0^x f''(y)\,dy\,dx \leq \int_0^\lambda \frac{x}{4}\,dx \leq \frac{\lambda^2}{8} \qquad \text{für alle } \lambda \geq 0. \end{aligned}$$

□

Beispiel. Im letzten Satz wurde gezeigt:

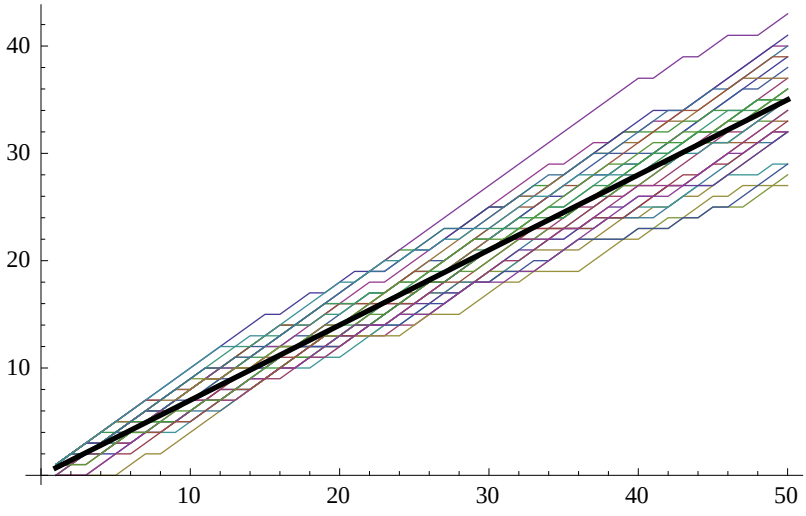
$$S_n = \sum_{i=1}^n I_{A_i}, \quad A_i \text{ unabhängig mit } P[A_i] = p \implies P\left[\left|\frac{S_n}{n} - p\right| \geq \varepsilon\right] \longrightarrow 0 \quad \text{für } n \to \infty.$$

Zur Demonstration simulieren wir den Verlauf von S_n und $\frac{S_n}{n}$ mehrfach (m-mal):

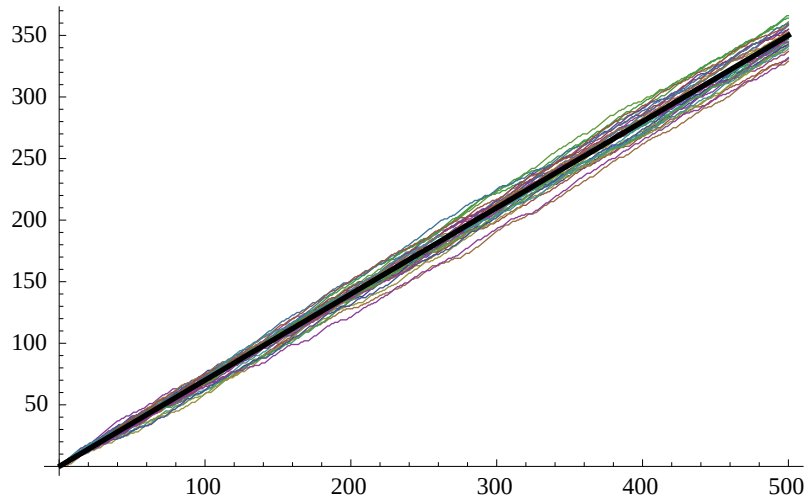
VERLAUF VON S_n

```
m = 30; nmax = 1000; p = 0.7;
(Wir erzeugen m × nmax Bernoulli-Stichproben mit Wahrscheinlichkeit p)
x = RandomChoice[{1 - p, p} -> {0, 1}, {nmax, m}]; s = Accumulate[x];
Das Feld s enthält m Verläufe von s_n = x_1 + ... + x_n, n = 1, ..., nmax
Manipulate[Show[
  ListLinePlot[Transpose[s[[1 ;; n]]]],
  ListLinePlot[p*Range[n], PlotStyle -> {Black, Thick}]]
  , {{n, 50}, 1, nmax, 1}]
(Vergleich der m Verläufe von s_n mit np)
```

- $n = 50$:



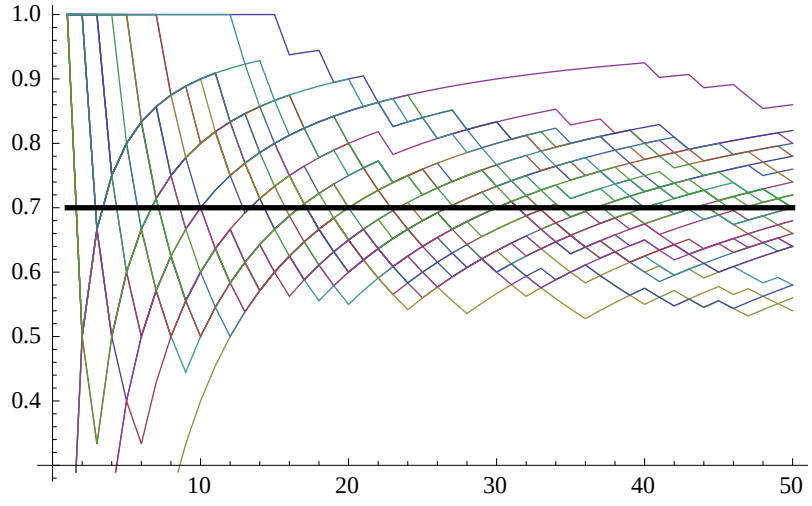
- $n = 500$:



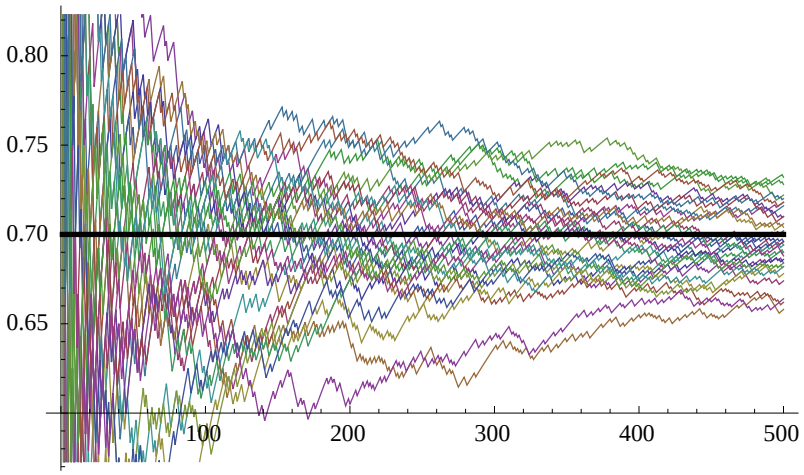
VERLAUF VON $\frac{S_n}{n}$

```
mean = s/Range[nmax];
(Das Feld mean enthält m Verläufe der Werte von s_n/n)
Manipulate[Show[
  ListLinePlot[Transpose[mean[[1 ;; n]]]],
  ListLinePlot[ConstantArray[p, n], PlotStyle -> {Black, Thick}] ], {{n,
    50}, 1, nmax, 1}]
```

- $n = 50$:



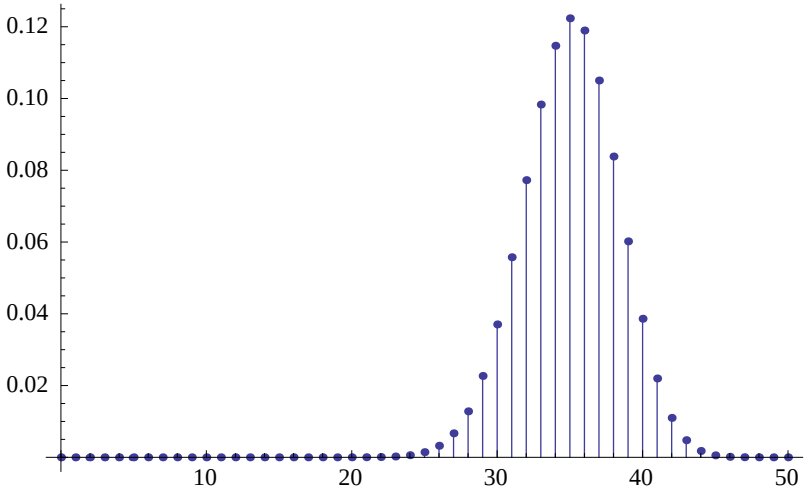
- $n = 500$:



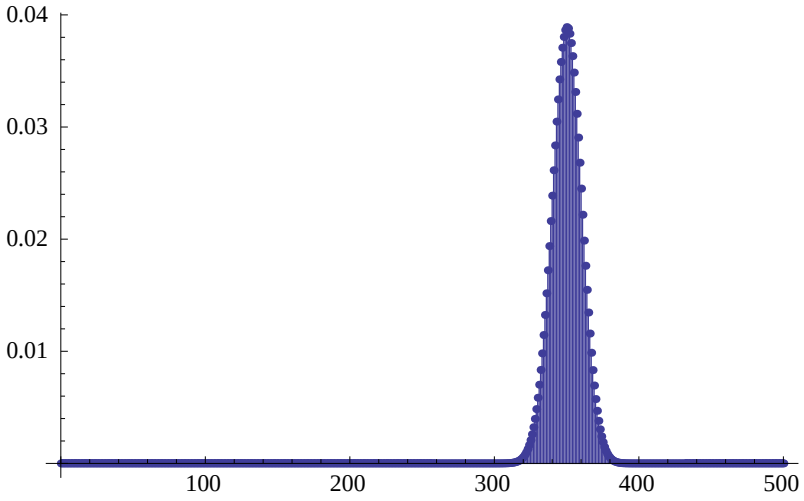
VERTEILUNG VON S_n

```
Manipulate[
 ListPlot[Table[{k, PDF[BinomialDistribution[n, p], k]}, {k, 0, n}],
  PlotRange -> All, Filling -> Axis]
 , {{n, 50}, 1, nmax, 1}]
```

- $n = 50$:



- $n = 500$:



2.4 Unabhängige Zufallsvariablen und Random Walk

Unabhängigkeit von diskreten Zufallsvariablen

Seien $X_i : \Omega \to S_i$, $i = 1, \ldots, n$, diskrete Zufallsvariablen auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$. Dann ist $(X_1, \ldots, X_n)$ eine Zufallsvariable mit Werten im Produktraum $S_1 \times \cdots \times S_n$.

Definition. *Die Verteilung $\mu_{X_1,\ldots,X_n}$ des Zufallsvektors $(X_1, \ldots, X_n)$ heißt **gemeinsame Verteilung** der Zufallsvariablen $X_1, \ldots, X_n$. Die Massenfunktion der gemeinsamen Verteilung lautet*

$$p_{X_1,\ldots,X_n}(a_1, \ldots, a_n) = P[X_1 = a_1, \ldots, X_n = a_n].$$

Definition. *Die diskreten Zufallsvariablen $X_1, \ldots, X_n$ heißen **unabhängig**, falls gilt:*

$$P[X_1 = a_1, \ldots, X_n = a_n] = \prod_{i=1}^{n} P[X_i = a_i] \qquad \text{für alle } a_i \in S_i, \qquad i = 1, \ldots, n.$$

Die gemeinsame Verteilung enthält Informationen über den Zusammenhang zwischen den Zufallsgrößen X_i.

Satz 2.7. *Die folgenden Aussagen sind äquivalent:*

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

(ii) $p_{X_1,\ldots,X_n}(a_1, \ldots, a_n) = \prod_{i=1}^{n} p_{X_i}(a_i)$.

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

(iv) Die Ereignisse $\{X_1 \in A_1\}, \ldots, \{X_n \in A_n\}$ *sind unabhängig für alle* $A_i \subseteq S_i$*,* $i = 1, \ldots, n$*.*

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

Beweis.

- (i) $\Leftrightarrow$ (ii) nach Definition von $p_{X_1,\ldots,X_n}$.
- (ii) $\Leftrightarrow$ (iii) nach Definition von $\bigotimes_{i=1}^n \mu_{X_i}$.
- (iii) $\Rightarrow$ (iv):
  Seien $1 \leq i_1 < i_2 < \ldots < i_m \leq n$ und $A_{i_k} \subseteq S_{i_k}$, $(k = 1, \ldots, m)$. Wir setzen $A_i := \Omega$ für $i \notin \{i_1, \ldots, i_m\}$. Mit (iii) folgt dann nach Satz 2.2:

$$\begin{aligned} P[X_{i_1} \in A_{i_1}, \ldots, X_{i_m} \in A_{i_m}] &= P[X_1 \in A_1, \ldots, X_n \in A_n] \\ &= P[(X_1, \ldots, X_n) \in A_1 \times \ldots \times A_n] \\ &= \mu_{X_1,\ldots,X_n}(A_1 \times \ldots \times A_n) \\ &= \prod_{i=1}^n \mu_{X_i}(A_i) = \prod_{i=1}^n P[X_i \in A_i] \\ &= \prod_{i=1}^m P[X_{i_k} \in A_{i_k}]. \end{aligned}$$

- (iv) $\Rightarrow$ (v) $\Rightarrow$ (i) ist klar.

□

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

Der Random Walk auf $\mathbb{Z}$

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

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

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:

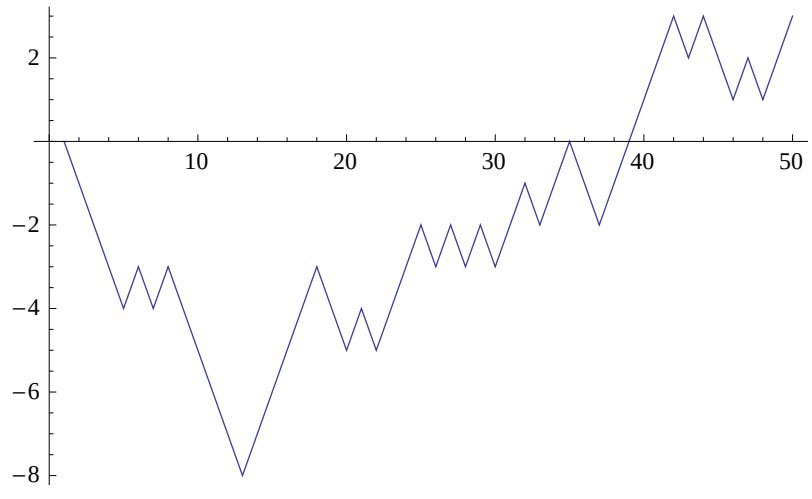
$$S_n = a + X_1 + X_2 + \cdots + X_n.$$

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

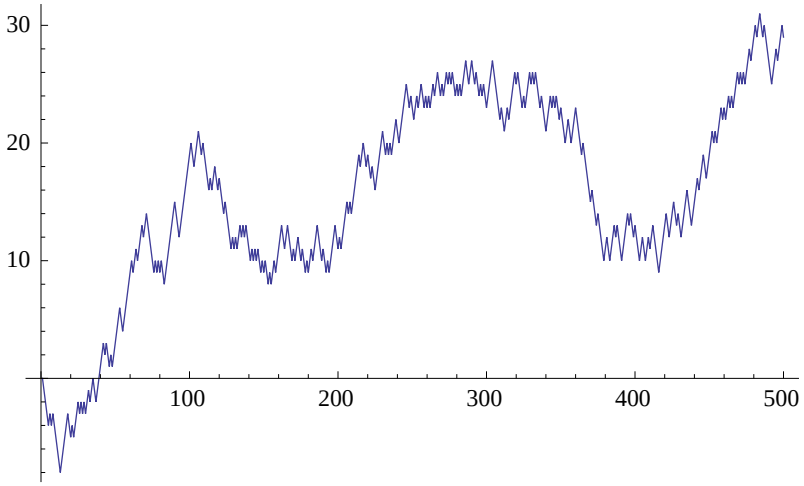
Beispiel (Symmetrischer Random Walk, $p = \frac{1}{2}$)**.**

```
zufall = RandomChoice[{-1, 1}, 10000];
randomwalk = FoldList[Plus, 0, zufall];
Manipulate[
 ListLinePlot[randomwalk[[1 ;; nmax]]], {nmax, 10, 10000, 10}]
```

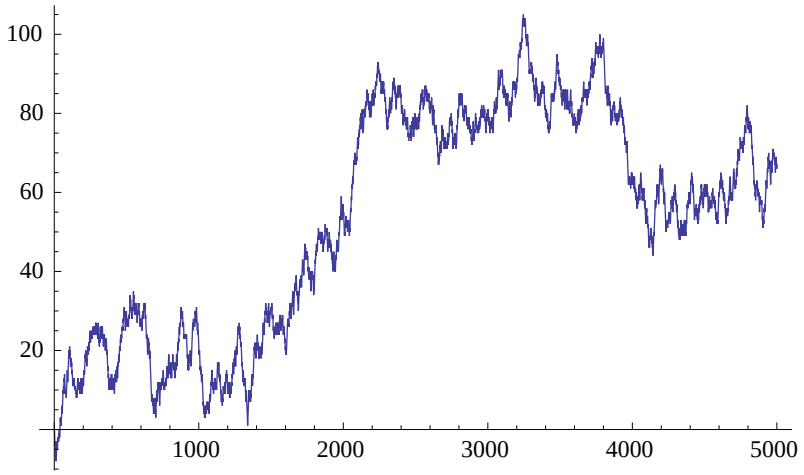
- $nmax = 50$:



- $nmax = 500$:



- $nmax = 5000$:



Lemma 2.8 (Verteilung von S_n)**.** *Für $k \in \mathbb{Z}$ gilt*

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

Beweis. Es gilt:

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

□

Beispiel (Rückkehrwahrscheinlichkeit zum Startpunkt)**.** Mithilfe der **Stirlingschen Formel**

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

folgt:

$$\begin{aligned} P[S_{2n+1} = a] &= 0, \\ P[S_{2n} = a] &= \binom{2n}{n} p^n (1-p)^n = \frac{(2n)!}{(n!)^2} p^n (1-p)^n \\ &\sim \frac{\sqrt{4\pi n}}{2\pi n} \frac{(\frac{2n}{e})^{2n}}{(\frac{n}{e})^{2n}} p^n (1-p)^n = \frac{1}{\sqrt{\pi n}} (4p(1-p))^n \qquad \text{für } n \to \infty, \end{aligned}$$

wobei zwei Folgen a_n und b_n **asymptotisch äquivalent** heißen $(a_n \sim b_n)$, falls $\lim_{n\to\infty} \frac{a_n}{b_n} = 1$ gilt.

- Falls $p \neq \frac{1}{2}$ gilt $4\,p\,(1-p) < 1$ und $P[S_{2n} = a]$ konvergiert exponentiell schnell gegen 0.
- Falls $p = \frac{1}{2}$ konvergiert $P[S_{2n} = a] \sim \frac{1}{\sqrt{\pi n}}$ nur langsam gegen 0.

Symmetrischer Random Walk

Ab jetzt betrachten wir den **symmetrischen** Random Walk, d.h. $p = \frac{1}{2}$.
Sei $\lambda \in \mathbb{Z}$. Wir wollen die Verteilung der Zufallsvariable

$$T_\lambda(\omega) := \inf\{n \in \mathbb{N} \mid S_n(\omega) = \lambda\}, \qquad (\min \emptyset := \infty),$$

bestimmen. Für $\lambda \neq a$ ist T_λ die erste **Trefferzeit von** λ, für $\lambda = a$ ist es die erste **Rückkehrzeit nach** a. Beschreibt der Random Walk beispielsweise die Kapitalentwicklung in einem Glücksspiel, dann kann man T_0 als Ruinzeitpunkt interpretieren.
Sei $n \in \mathbb{N}$. Wir wollen nun die Wahrscheinlichkeit

$$P\left[T_\lambda \leq n\right] = P\left[\bigcup_{i=1}^{n} \{S_i = \lambda\}\right]$$

berechnen. Da das Ereignis $\{T_\lambda \leq n\}$ von **mehreren** Positionen des Random Walks abhängt $(S_1, S_2, \ldots, S_n)$, benötigen wir die **gemeinsame** Verteilung dieser Zufallsvariablen. Sei also

$$S(\omega) := (S_0(\omega), S_1(\omega), \ldots, S_n(\omega))$$

der **Bewegungsverlauf bis zur Zeit** n . Dann ist S eine Zufallsvariable mit Werten im Raum

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

der möglichen Pfade des Random Walk. Sei μ_a die gemeinsame Verteilung von S unter P.

Lemma 2.9. *μ_a ist die **Gleichverteilung** auf dem Pfadraum $\widehat{\Omega}_a^{(n)}$.*

Beweis. Es gilt

$$\begin{aligned}\mu_a(\{(s_0,\ldots,s_n)\}) &= P[S_0 = s_0,\ldots,S_n = s_n]\\ &= P[S_0 = s_0, X_1 = s_1 - s_0,\ldots,X_n = s_n - s_{n-1}]\\ &= \begin{cases} 0 & \text{falls } s_0 \neq a \text{ oder } |s_i - s_{i-1}| \neq 1 \text{ für ein } i \in \{1,\ldots,n\},\\ & (\text{d.h. } (s_0,\ldots,s_n) \notin \widehat{\Omega}_a^{(n)}),\\ 2^{-n} & \text{sonst, d.h. falls } (s_0,\ldots,s_n) \in \widehat{\Omega}_a^{(n)}. \end{cases}\end{aligned}$$

□

Satz 2.10 (Reflektionsprinzip)**.** *Seien $\lambda, b \in \mathbb{Z}$. Es gelte entweder $(a < \lambda$ und $b \leq \lambda)$, oder $(a > \lambda$ und $b \geq \lambda)$. Dann gilt:*

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

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

Beweis. Es gilt:

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

Die im Bild 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:

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

□

Korollar (Verteilung der Trefferzeiten)**.** *Es gilt:*

i)

$$P[T_\lambda \leq n] = \begin{cases} P[S_n \geq \lambda] + P[S_n > \lambda], & \textit{falls } \lambda > a,\\ P[S_n \leq \lambda] + P[S_n < \lambda], & \textit{falls } \lambda < a. \end{cases}$$

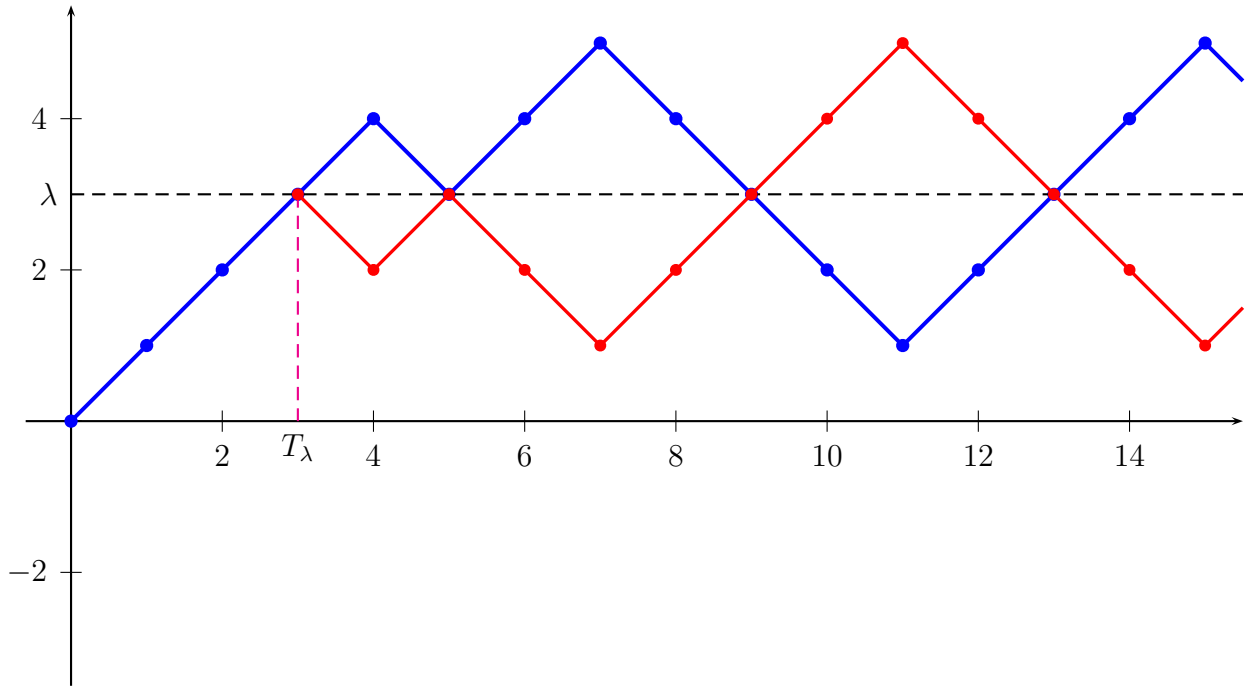


Abbildung 2.2: Spiegelung des Random Walks an $\lambda = 3$

ii)

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

$$= \begin{cases} \frac{\lambda - a}{n} \binom{n}{\frac{n+\lambda-a}{2}} 2^{-n} & \textit{falls } \lambda > a, \\ \frac{a - \lambda}{n} \binom{n}{\frac{n+\lambda-a}{2}} 2^{-n} & \textit{falls } \lambda < a. \end{cases}$$

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

i)

$$P[T_\lambda \leq n] = \sum_{b \in \mathbb{Z}} \underbrace{P[T_\lambda \leq n, S_n = b]}$$

$$= \begin{cases} P[S_n = b] & \text{falls } b \geq \lambda, \\ P[S_n = b^\star] & \text{falls } b < \lambda. \end{cases}$$

$$= \sum_{b \geq \lambda} P[S_n = b] + \underbrace{\sum_{b < \lambda} P[S_n = b^\star]}_{= \sum_{b > \lambda} P[S_n = b]} = P[S_n \geq \lambda] + P[S_n > \lambda].$$

ii)

$$P[T_\lambda = n] = P[T_\lambda \leq n] - P[T_\lambda \leq n-1]$$

Mit i) folgt

$$= \overbrace{P[\underbrace{S_n \geq \lambda}_{=:A}] - P[\underbrace{S_{n-1} \geq \lambda}_{=:B}]}^{=:\mathbf{I}} + \overbrace{P[S_n \geq \lambda+1] - P[S_{n-1} \geq \lambda+1]}^{=:\mathbf{II}}$$

Wegen

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

erhalten wir für den ersten Term:

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

Hierbei haben wir benutzt, dass

$$\begin{aligned}
&|\{(s_0, \ldots, s_n) \in \widehat{\Omega}_a^{(n)} \mid s_{n-1} = \lambda - 1\}| \\
&= |\{(s_0, \ldots, s_n) \mid s_{n-1} = \lambda - 1 \text{ und } s_n = \lambda\}| \\
&\quad + |\{(s_0, \ldots, s_n) \mid s_{n-1} = \lambda - 1 \text{ und } s_n = \lambda - 2\}| \\
&= 2 \cdot |\{(s_0, \ldots, s_n) | s_{n-1} = \lambda - 1, s_n = \lambda\}
\end{aligned}$$

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

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

□

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

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

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

2.5 Simulationsverfahren

Die Simulation von Stichproben verschiedener Wahrscheinlichkeitsverteilungen geht von auf $[0,1]$ gleichverteilten Pseudo-Zufallszahlen aus. In Wirklichkeit simuliert ein Zufallszahlengenerator natürlich nur auf $\{k\,m^{-1} \mid k = 0, 1, \ldots, m-1\}$ gleichverteilte Zufallszahlen, wobei m^{-1} die Darstellungsgenauigkeit des Computers ist. Dieser Aspekt wird im folgenden ignoriert. Um Simulationsverfahren zu analysieren, benötigen wir noch den Begriff einer auf $[0,1]$ gleichverteilten reellwertigen Zufallsvariablen. Die Existenz solcher Zufallsvariablen auf einem geeigneten Wahrscheinlichkeitsraum wird hier vorausgesetzt, und kann erst in der Vorlesung »Analysis III« bzw. in der »Einführung in die Wahrscheinlichkeitstheorie« gezeigt werden.
Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum, und $U\colon \Omega \to [0,1]$ eine Abbildung.

Definition. *i)* U *ist eine* ***reellwertige Zufallsvariable****, falls gilt:*

$$\{\omega \in \Omega \mid U(\omega) \le y\} \in \mathcal{A} \qquad \textit{für alle } y \in \mathbb{R}.$$

ii) Eine reellwertige Zufallsvariable $U\colon \Omega \to [0,1]$ *ist* ***gleichverteilt auf*** $[0,1]$*, falls*

$$P[U \le y] = y \qquad \textit{für alle } y \in [0,1].$$

Wir notieren dies im folgenden als $(U \sim \mathrm{Unif}[0,1])$.

iii) Reellwertige Zufallsvariablen $U_i\colon \Omega \to \mathbb{R}$*,* $i \in I$*, heißen* ***unabhängig****, falls die Ereignisse* $\{U_i \le y_i\}$*,* $i \in I$*, für alle* $y_i \in \mathbb{R}$ *unabhängig sind.*

Ein **Zufallszahlengenerator** simuliert Stichproben $u_1 = U_1(\omega)$, $u_2 = U_2(\omega)$, ... von auf $[0,1]$ gleichverteilten unabhängigen Zufallsvariablen. Wie erzeugt man daraus Stichproben von diskreten Verteilungen?

Das direkte Verfahren

Sei $S = \{a_1, a_2, \ldots\}$ endlich oder abzählbar unendlich, und μ eine Wahrscheinlichkeitsverteilung auf S mit Gewichten $p_i = p(a_i)$. Wir setzen

$$s_n := \sum_{i=1}^{n} p_i, \qquad n \in \mathbb{N}, \qquad \text{»kumulative Verteilungsfunktion«}.$$

Sei $U\colon \Omega \to [0,1)$ eine gleichverteilte Zufallsvariable. Wir setzen

$$X(\omega) := a_i, \qquad \text{falls } s_{i-1} < U(\omega) \le s_i, \qquad i \in \mathbb{N}.$$

Lemma 2.11. *Falls* $U \sim \mathrm{Unif}[0,1)$*, gilt* $X \sim \mu$*.*

Beweis. Für alle $i \in \mathbb{N}$ gilt:

$$P[X = a_i] = P[s_{i-1} < U \leq s_i] = P[U \leq s_i] - P[U \leq s_{i-1}] = s_i - s_{i-1} = p_i.$$

□

Algorithmus 2.12 (Direkte Simulation einer diskreten Verteilung)**.**

INPUT: Gewichte $p_1, p_2, \ldots,$

OUTPUT: Pseudozufallsstichprobe x von μ.

$n := 1$

$s := p_1$

erzeuge Zufallszahl $u \sim \mathrm{Unif}[0,1)$

while $u > s$ **do**

$n := n + 1$

$s := s + p_n$

end while

return $x := a_n$

Bemerkung. a) Die mittlere Anzahl von Schritten des Algorithmus ist

$$\sum_{n=1}^{\infty} n\, p_n = \text{ Erwartungswert von Wahrscheinlichkeitsverteilung } (p_n) \text{ auf } \mathbb{N}.$$

b) Für große Zustandsräume S ist das direkte Verfahren oft nicht praktikabel, siehe Übung.

Acceptance-Rejection-Verfahren

Sei S eine endliche oder abzählbare Menge, μ eine Wahrscheinlichkeitsverteilung auf S mit Massenfunktion $p(x)$, und ν eine Wahrscheinlichkeitsverteilung auf S mit Massenfunktion $q(x)$. Angenommen, wir können unabhängige Stichproben von ν erzeugen. Wie können wir daraus Stichproben von μ erhalten? IDEE: Erzeuge Stichprobe x von ν, akzeptiere diese mit Wahrscheinlichkeit proportional zu $\frac{p(x)}{q(x)}$, sonst verwerfe die Stichprobe und wiederhole.

ANNAHME:

$$\text{es gibt ein } c \in [1,\infty): \qquad p(x) \leq c\, q(x) \qquad \text{für alle } x \in S.$$

Aus der Annahme folgt:

$$\frac{p(x)}{c\, q(x)} \leq 1 \qquad \text{für alle } x \in S,$$

d.h. wir können $\frac{p(x)}{c\, q(x)}$ als **Akzeptanzwahrscheinlichkeit** wählen.

Algorithmus 2.13 (Acceptance-Rejection-Verfahren)**.**

INPUT: Gewichte $p(y)$, $q(y)$, c $\quad (y \in S)$,

OUTPUT: Stichprobe x von μ.

repeat

erzeuge Stichprobe $x \sim \nu$

erzeuge Stichprobe $u \sim \mathrm{Unif}[0,1]$

until $\frac{p(x)}{c\,q(x)} \geq u$ {akzeptiere mit Wahrscheinlichkeit $\frac{p(x)}{c\,q(x)}$}

return x

ANALYSE DES ALGORITHMUS

Für die verwendeten Zufallsvariablen gilt:

$$X_1, X_2, \ldots \sim \nu, \qquad (\text{Vorschläge}),$$
$$U_1, U_2, \ldots \sim \mathrm{Unif}[0,1].$$

Es gilt Unabhängigkeit, d.h.

$$P[X_1 = a_1, \ldots, X_n = a_n, U_1 \leq y_1, \ldots, U_n \leq q_n] = \prod_{i=1}^{n} P[X_i = a_i] \cdot \prod_{i=1}^{n} P[U_i \leq y_i]$$

für alle $n \in \mathbb{N}$, $a_i \in S$ und $y_i \in \mathbb{R}$.

Seien

$$\begin{aligned} T &= \min\left\{ n \in \mathbb{N} \;\middle|\; \tfrac{p(X_n)}{c\,q(X_n)} \geq U_n \right\} && \text{die »Akzeptanzzeit«, und} \\ X_T(\omega) &= X_{T(\omega)}(\omega) && \text{die ausgegebene Stichprobe.} \end{aligned}$$

des Acceptance-Rejection-Verfahrens. Wir erhalten:

Satz 2.14. *i) T ist **geometrisch verteilt** mit Parameter $1/c$,*

ii) $X_T \sim \mu$.

Bemerkung. Insbesondere ist die mittlere Anzahl von Schritten bis Akzeptanz:

$$E[T] = c.$$

Beweis. i) Sei

$$A_n := \left\{ \frac{p(X_n)}{c\,q(X_n)} \geq U_n \right\}.$$

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« bewiesen. Zudem gilt wegen der Unabhängigkeit von X_n und U_n:

$$\begin{aligned} P[A_n] &= \sum_{a \in S} P\left[\left\{U_n \leq \frac{p(a)}{c\,q(a)}\right\} \cap \{X_n = a\}\right] \\ &= \sum_{a \in S} P\left[\left\{U_n \leq \frac{p(a)}{c\,q(a)}\right\}\right] \cdot P[X_n = a] \\ &= \sum_{a \in S} \frac{p(a)}{c\,q(a)} \cdot q(a) = \frac{1}{c}. \end{aligned}$$

Also ist

$$T(\omega) = \min\{n \in \mathbb{N} \mid \omega \in A_n\}$$

geometrisch verteilt mit Parameter $1/c$.

ii)

$$\begin{aligned} P[X_T = a] &= \sum_{n=1}^{\infty} P[\{X_T = a\} \cap \{T = n\}] \\ &= \sum_{n=1}^{\infty} P[\{X_n = a\} \cap A_n \cap A_1^C \cap \ldots A_{n-1}^C] \\ &= \sum_{n=1}^{\infty} P[\{X_n = a\} \cap \left\{\frac{p(a)}{c\,q(a)} \geq U_n\right\} \cap A_1^C \cap \ldots A_{n-1}^C] \\ &= \sum_{n=1}^{\infty} q(a)\,\frac{p(a)}{c\,q(a)}\,\left(1 - \frac{1}{c}\right)^{n-1} \\ &= \frac{p(a)}{c} \sum_{n=1}^{\infty} \left(1 - \frac{1}{c}\right)^{n-1} \\ &= \frac{p(a)}{c}\,\frac{1}{1 - (1 - \frac{1}{c})} = p(a). \end{aligned}$$

□

Kapitel 3

Konvergenzsätze und Monte Carlo Verfahren

Sei μ eine Wahrscheinlichkeitsverteilung auf einer abzählbaren Menge S, und $f : S \to \mathbb{R}$ eine reellwertige Zufallsvariable. Angenommen, wir wollen den Erwartungswert

$$\theta := E_\mu[f] = \sum_{x \in S} f(x)\, \mu(x)$$

berechnen, aber die Menge S ist zu groß, um die Summe direkt auszuführen. In einem **Monte Carlo-Verfahren** simuliert man eine große Anzahl unabhängiger Stichproben $X_1(\omega), \ldots, X_n(\omega)$ von μ, und approximiert den Erwartungswert θ durch den **Monte Carlo-Schätzer**

$$\widehat{\theta}_n(\omega) := \frac{1}{n} \sum_{i=1}^{n} f(X_i(\omega)).$$

Wir wollen nun Methoden entwickeln, mit denen der Approximationsfehler $|\widehat{\theta}_n - \theta|$ abgeschätzt werden kann, und die Asymptotik des Approximationsfehlers für $n \to \infty$ untersuchen. Nach dem Transformationssatz (1.7) und der Linearität des Erwartungswerts (1.8) gilt:

$$E[\widehat{\theta}_n] = \frac{1}{n} \sum_{i=1}^{n} E[f(X_i)] = \frac{1}{n} \sum_{i=1}^{n} \sum_{x \in S} f(x)\, \mu(\{x\}) = E_\mu[f] = \theta,$$

d.h. $\widehat{\theta}_n$ ist ein **erwartungstreuer** Schätzer für θ. Der mittlere quadratische Fehler (»MSE« – mean squared error) des Schätzers ist daher:

$$\text{MSE} = E[|\widehat{\theta}_n - \theta|^2] = E[|\widehat{\theta}_n - E[\widehat{\theta}_n]|^2].$$

3.1 Varianz und Kovarianz

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

Definition.

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

*heißt **Varianz** von X und liegt in $[0, \infty]$.*

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

*heißt **Standardabweichung** von X.*

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

Bemerkung. (a) Die Varianz hängt nur von der Verteilung von X ab:

$$\mathrm{Var}(X) = \sum_{a \in S} (a - m)^2 \, p_X(a), \qquad \text{wobei} \qquad m = E[X] = \sum_{a \in S} a \, p_X(a).$$

(b) Es gilt

$$\mathrm{Var}(X) = 0 \qquad \text{genau dann, wenn} \qquad P[X = E[X]] = 1.$$

Bemerkung (Rechenregeln)**.** i)

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

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

ii)

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

Beweis. i) Nach der Linearität des Erwartungswerts gilt

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

ii) Wiederholte Anwendung der Linearität des Erwartungswerts liefert

$$\mathrm{Var}(aX + b) = E\left[(aX + b - E[aX + b])^2\right] = E\left[(aX - E[aX])^2\right] = a^2 \ \mathrm{Var}(X).$$

□

Beispiele. a) Sei $X = 1$ mit Wahrscheinlichkeit p und $X = 0$ mit Wahrscheinlichkeit $1 - p$. Dann ist der Erwartungswert von X:

$$E\left[X^2\right] = E[X] = p,$$

und die Varianz von X:

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

b) Sei T geometrisch verteilt ($T \sim Geom(p)$) mit Parameter $p \in (0,1]$. Der Erwartungswert von T beträgt:

$$E[T] = \sum_{k=1}^{\infty} k\,(1-p)^{k-1}\,p = -p\,\frac{p}{dp}\sum_{k=0}^{\infty}(1-p)^k = -p\,\frac{p}{dp}\frac{1}{p} = \frac{1}{p}.$$

Außerdem gilt:

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

Die Varianz von T ist somit:

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

Definition.

$$\mathcal{L}^2(\Omega, \mathcal{A}, P) \;:=\; \{X : \Omega \to \mathbb{R} \mid X \textit{ ist diskrete Zufallsvariable mit } E\left[X^2\right] < \infty\}$$

Lemma 3.1. *i) Für Zufallsvariablen $X, Y \in \mathcal{L}^2$ gilt:*

$$E[|XY|] \le \sqrt{E\left[X^2\right]}\,\sqrt{E\left[Y^2\right]} < \infty.$$

ii) $\mathcal{L}^2$ ist ein Vektorraum, und

$$(X,Y)_{\mathcal{L}^2} := E[X\,Y]$$

*ist eine **positiv semidefinite symmetrische Bilinearform (»Skalarprodukt«)** auf $\mathcal{L}^2$.*

Bemerkung. i) Insbesondere gilt die **Cauchy-Schwarz-Ungleichung**:

$$E[X\,Y]^2 \le E[|X\,Y|] \le E\left[X^2\right]\;E\left[Y^2\right] \qquad \text{für alle } X, Y \in \mathcal{L}^2.$$

ii) Für eine Zufallsvariable $X \in \mathcal{L}^2$ gilt

$$E[|X|] \leq \sqrt{E\left[X^2\right]}\,\sqrt{E\left[1^2\right]} < \infty.$$

Beweis. i) Nach der Cauchy-Schwarz-Ungleichung gilt:

$$\begin{aligned}
E[|X\,Y|] &= \sum_{\substack{a\in X(\Omega)\\ b\in Y(\Omega)}} |a\,b|\,P[X=a, Y=b] \\
&= \sum_{\substack{a\in X(\Omega)\\ b\in Y(\Omega)}} |a|\,\sqrt{P[X=a, Y=b]}\,|b|\,\sqrt{P[X=a, Y=b]} \\
&\leq \sqrt{\sum_{a,b} a^2\,P[X=a, Y=b]}\,\sqrt{\sum_{a,b} b^2\,P[X=a, Y=b]} \\
&= \sqrt{\sum_{a} a^2\,P[X=a]}\,\sqrt{\sum_{b} b^2\,P[Y=b]} \\
&= \sqrt{E\left[X^2\right]}\,\sqrt{E\left[Y^2\right]}.
\end{aligned}$$

ii) Seien $X, Y \in \mathcal{L}^2$, $a \in \mathbb{R}$. Dann ist $aX + Y$ eine diskrete Zufallsvariable, für die nach Monotonie und der Linearität des Erwartungswerts gilt:

$$E\left[(aX+Y)^2\right] = E\left[a^2X^2 + 2aX\,Y + Y^2\right] \leq 2a^2E\left[X^2\right] + 2E\left[Y^2\right] < \infty.$$

$(X,Y)_{\mathcal{L}^2} = E[X\,Y]$ ist bilinear, da $E[\,\bullet\,]$ linear und symmetrisch ist, und positiv semidefinit, aufgrund von:

$$(X,X)_{\mathcal{L}^2} = E\left[X^2\right] \geq 0 \qquad \text{für alle } X \in \mathcal{L}^2.$$

□

Definition. *Seien $X, Y \in \mathcal{L}^2$.*

i)

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

*heißt **Kovarianz** von X und 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) X *und* Y *heißen* ***unkorreliert****, falls* $\mathrm{Cov}(X,Y) = 0$*, d.h.*

$$E[X\,Y] \;=\; E[X] \cdot E[Y].$$

Bemerkung. $\mathrm{Cov} : \mathcal{L}^2 \times \mathcal{L}^2 \to \mathbb{R}$ ist eine symmetrische Bilinearform mit:

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

Satz 3.2 (Zusammenhang von Unabhängigkeit und Unkorreliertheit)**.** *Seien* $X : \Omega \to S$ *und* $Y : \Omega \to T$ *diskrete Zufallsvariablen auf* $(\Omega, \mathcal{A}, P)$*. Dann sind äquivalent:*

(i) X *und* Y *sind unabhängig, d.h.*

$$P[X \in A, Y \in B] = P[X \in A]\, P[Y \in B] \qquad \textit{für alle } A, B \in \mathcal{A}.$$

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

Beweis.

- (i)⇒ (ii): Seien X und Y unabhängig, dann gilt:

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

  Somit folgt:

$$\mathrm{Cov}(f(X), g(Y)) = 0.$$

- (ii)⇒ (i): Aus (ii) folgt für alle $a \in S$, $b \in T$:

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

□

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

$$E[X\,Y] = 0 = E[X]\, E[Y].$$

Satz 3.3 (Varianz von Summen)**.** *Für $X_1, \ldots, X_n \in \mathcal{L}^2$ gilt:*

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

Falls $X_1, \ldots, X_n$ unkorreliert sind, folgt insbesondere:

$$\mathrm{Var}(X_1 + \cdots + X_n) = \sum_{i=1}^{n} \mathrm{Var}(X_i).$$

Beweis. Nach Bilinearität der Kovarianz gilt:

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

□

Beispiel (Varianz der Binomialverteilung)**.** Sei

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

mit unabhängigen Zufallsvariablen X_i. Mit Satz 3.2 folgt:

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

Analog gilt für den Random Walk:

$$\sigma(S_n) = O(\sqrt{n}).$$

3.2 Das schwache Gesetz der großen Zahlen

Seien $X_1, X_2, \ldots : \Omega \to \mathbb{R}$ Zufallsvariablen, die auf einem gemeinsamen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ 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

$$\frac{S_n(\omega)}{n} \quad \approx \quad E\left[\frac{S_n}{n}\right] \qquad \text{für große } n,$$

bzw.

$$\frac{S_n}{n} - \frac{E[S_n]}{n} \quad \stackrel{n\to\infty}{\longrightarrow} \quad 0.$$

Ist insbesondere $E[X_i] = m$ 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 $\dfrac{S_n}{n} = X_1$ 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 sogar Unkorreliertheit und beschränkte Varianzen der Zufallsvariablen X_i genügen, um ein Gesetz der großen Zahlen zu erhalten. Dazu nehmen wir an, dass $X_1, X_2, \ldots$ diskrete Zufallsvariablen aus $\mathcal{L}^2(\Omega, \mathcal{A}, P)$ sind, die folgende Voraussetzungen erfüllen:

ANNAHMEN:

(i) Die Zufallsvariablen sind unkorreliert:

$$\mathrm{Cov}(X_i, X_j) = 0 \qquad \text{für alle } i \neq j.$$

(ii) Die Varianzen sind beschränkt:

$$v \;:=\; \sup_{i\in\mathbb{N}} \mathrm{Var}(X_i) \;<\; \infty.$$

Es wird **keine Unabhängigkeit vorausgesetzt**!

Satz 3.4 (Schwaches Gesetz der großen Zahlen)**.** *Unter den Voraussetzungen (i) und (ii) gilt für alle $\varepsilon > 0$:*

$$P\left[\left|\frac{S_n}{n} - \frac{E[S_n]}{n}\right| \geq \varepsilon\right] \leq \frac{v}{\varepsilon^2 n} \longrightarrow 0 \qquad \text{für } n \to \infty.$$

Gilt außerdem $E[X_i] = m$ für alle $i \in \mathbb{N}$, folgt $\frac{E[S_n]}{n} = m$ und $\frac{S_n}{n}$ ***konvergiert stochastisch*** *gegen m.*

Zum Beweis benötigen wir:

Lemma 3.5 (Čebyšev-Ungleichung)**.** *Für $X \in \mathcal{L}^2$ und $c > 0$ gilt:*

$$P[|X - E[X]| \geq c] \leq \frac{1}{c^2} \operatorname{Var}(X).$$

Beweis. Es gilt

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

$\frac{1}{c^2}(X - E[X])^2$ ist überall nichtnegativ und ≥ 1 auf $\{|X - E[X]| \geq c\}$. Durch Bilden des Erwartungswerts folgt:

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

□

Beweis von Satz 3.4. Nach der Čebyšev-Ungleichung und den Annahmen (i) und (ii) gilt für $\varepsilon > 0$:

$$P\left[\left|\frac{S_n}{n} - \frac{E[S_n]}{n}\right| \geq \varepsilon\right] \leq \frac{1}{\varepsilon^2} \operatorname{Var}\left(\frac{S_n}{n}\right) = \frac{1}{n^2\,\varepsilon^2} \operatorname{Var}(\sum_{i=1}^{n} X_i) = \frac{1}{n^2\,\varepsilon^2} \sum_{i=1}^{n} \operatorname{Var}(X_i) \leq \frac{v}{n\,\varepsilon^2}.$$

□

Bemerkung (Starkes Gesetz der großen Zahlen)**.**

$$\frac{S_n(\omega)}{n} \longrightarrow m \quad \text{mit Wahrscheinlichkeit } 1.$$

Dies wird in der Vorlesung »Einführung in die Wahrscheinlichkeitstheorie« bewiesen.

3.3 Monte Carlo-Verfahren

Sei S eine abzählbare Menge und μ eine Wahrscheinlichkeitsverteilung auf S. Wir bezeichnen im folgenden die Massenfunktion ebenfalls mit μ, d.h.

$$\mu(x) := \mu(\{x\}).$$

Sei $f : S \to \mathbb{R}$ eine reellwertige Funktion mit:

$$E_\mu[f^2] = \sum_{x \in S} f(x)^2 \, \mu(x) < \infty.$$

Wir wollen den Erwartungswert

$$\theta := E_\mu[f] = \sum_{x \in S} f(x) \, \mu(x)$$

näherungsweise berechnen bzw. schätzen. Dazu approximieren wir θ durch die **Monte Carlo-Schätzer**

$$\widehat{\theta}_n := \frac{1}{n} \sum_{i=1}^{n} f(X_i), \qquad n \in \mathbb{N},$$

wobei $X_1, X_2, \ldots$ unabhängige Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ mit Verteilung μ sind. Nach der Abschätzung aus dem Gesetz der großen Zahlen ergibt sich:

Korollar.

$$P[|\widehat{\theta}_n - \theta| \geq \varepsilon] \leq \frac{1}{n \, \varepsilon^2} \operatorname{Var}_\mu[f] \longrightarrow 0 \quad \textit{für } n \to \infty,$$

*d.h. $\widehat{\theta}_n$ ist eine **konsistente Schätzfolge** für θ.*

Beweis. Da die Zufallsvariablen X_i unabhängig sind, sind $f(X_i)$, $i \in \mathbb{N}$, unkorreliert. Zudem gilt

$$E[f(X_i)] = \sum_{x \in S} f(x) \, \mu(x) = E_\mu[f] = \theta, \qquad \text{und}$$
$$\operatorname{Var}[f(X_i)] = \sum_{x \in S} (f(x) - \theta)^2 \, \mu(x) = \operatorname{Var}_\mu[f] < \infty$$

nach Voraussetzung. Die Behauptung folgt nun aus Satz 3.4. □

Bemerkung. a) $\widehat{\theta}_n$ ist ein **erwartungstreuer Schätzer** für θ:

$$E[\widehat{\theta}_n] = \frac{1}{n} \sum_{i=1}^{n} E[f(X_i)] = E_\mu[f] = \theta.$$

b) Für den mittleren quadratischen Fehler des Schätzers ergibt sich nach a):

$$E\left[|\widehat{\theta}_n - \theta|^2\right] = \operatorname{Var}(\widehat{\theta}_n) = \frac{1}{n} \operatorname{Var}_\mu[f].$$

Insbesondere gilt:

$$\|\widehat{\theta}_n - \theta\|_{\mathcal{L}^2} = \sqrt{E[|\widehat{\theta}_n - \theta|^2]} = O(1/\sqrt{n}).$$

Beispiele. a) MONTE CARLO-SCHÄTZUNG VON $\theta = \int_{[0,1]^d} f(x)\,dx$:

Das mehrdimensionale Integral ist folgendermaßen definiert:

$$\int_{[0,1]^d} f(x)\,dx := \int_0^1 \dots \int_0^1 f(x_1,\dots,x_d)\,dx_1 \dots dx_d.$$

Der Wert von θ kann mit dem folgenden Algorithmus geschätzt werden.

erzeuge Pseudozufallszahlen $u_1, u_2, \dots, u_{nd} \in (0,1)$
$x^{(1)} := (u_1, \dots, u_d)$
$x^{(2)} := (u_{d+1}, \dots, u_{2d})$
$\dots$
$x^{(n)} := (u_{(n-1)d+1}, \dots, u_{nd})$
$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n f(x^{(i)})$ ist Schätzwert für θ.

b) MONTE CARLO-SCHÄTZUNG VON WAHRSCHEINLICHKEITEN:

Sei S abzählbar, $B \subseteq S$. Wir suchen:

$$p = \mu(B) = E_\mu[I_B]$$

Ein Monte Carlo-Schätzer ist

$$\widehat{p}_n = \frac{1}{n} \sum_{i=1}^n I_B(X_i), \qquad X_i \text{ unabhängig mit Verteilung } \mu.$$

FEHLERKONTROLLE:

- Mithilfe der Čebyšev-Ungleichung (Lemma 3.5) ergibt sich:

$$P[|\widehat{p}_n - p| \geq \varepsilon] \leq \frac{1}{\varepsilon^2}\ \mathrm{Var}(\widehat{p}_n) = \frac{1}{n\varepsilon^2}\ \mathrm{Var}_\mu(I_B) = \frac{p\,(1-p)}{n\varepsilon^2} \leq \frac{1}{4n\varepsilon^2}.$$

  Gilt beispielsweise $n \geq \frac{5}{\varepsilon^2}$, dann erhalten wir:

$$P[p \notin (\widehat{p}_n - \varepsilon, \widehat{p}_n + \varepsilon)] \leq 5\%, \qquad \text{unabhängig von } p,$$

  d.h. das zufällige Intervall $(\widehat{p}_n - \varepsilon, \widehat{p}_n + \varepsilon)$ ist ein **95%-Konfidenzintervall** für den gesuchten Wert p.

- Mithilfe der Bernstein-Ungleichung (Chernoff-Abschätzung) erhalten wir für $\delta > 0$ und $S_n := \sum_{i=1}^n I_B(X_i)$:

$$P[p \notin (\widehat{p}_n - \varepsilon, \widehat{p}_n + \varepsilon)] = P\big[\big|\frac{1}{n}S_n - p\big| \geq \varepsilon\big] \leq 2e^{-2n\varepsilon^2} \leq \delta, \qquad \text{falls } n \geq \frac{\log(2/\delta)}{2\varepsilon^2}.$$

Für kleine δ ist die erhaltene Bedingung an n wesentlich schwächer als eine entsprechende Bedingung, die man durch Anwenden der Čebyšev-Ungleichung erhält. Für den **relativen Schätzfehler** $(\widehat{p}_n - p)/p$ ergibt sich:

$$P[|\widehat{p}_n - p| \geq \varepsilon p] \leq 2e^{-2n\varepsilon^2 p^2} \leq \delta, \qquad \text{falls } n \geq \frac{\log(2/\delta)}{2\varepsilon^2 p^2}.$$

Die benötigte Anzahl von Stichproben für eine (ε, δ)-Approximation von p ist also polynomiell in ε, $\log(1/\delta)$ und $1/p$. Mit einer etwas modifizierten Abschätzung kann man statt der Ordnung $O(\frac{1}{p^2})$ sogar $O(\frac{1}{p})$ erhalten, siehe `Mitzenmacher und Upfal`: »Probability and Computing«.

Beispiel. Wie viele Stichproben sind nötig, damit der **relative Fehler** mit 95% Wahrscheinlichkeit unterhalb von 10% liegt? Mithilfe der Čebyšev-Ungleichung (Lemma 3.5) ergibt sich:

$$P[|\widehat{p}_n - p| \geq 0,1\, p] \leq \frac{p\,(1-p)}{n\,(0,1\, p)^2} = \frac{100\,(1-p)}{n\,p} \leq 0,05, \quad \text{falls } n \geq \frac{2000\,(1-p)}{p}.$$

So sind zum Beispiel für $p = 10^{-5}$ ungefähr $n \approx 2\,10^8$ Stichproben ausreichend. Dies ist nur eine obere Schranke, aber man kann zeigen, dass die tatsächlich benötigte Stichprobenzahl immer noch sehr groß ist. Für solch kleine Wahrscheinlichkeiten ist das einfache Monte Carlo-Verfahren ineffektiv, da die meisten Summanden von $\widehat{\theta}_n$ dann gleich 0 sind. Wir brauchen daher ein alternatives Schätzverfahren mit geringerer Varianz.

Varianzreduktion durch Importance Sampling

Sei 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 den gesuchten Wert θ 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 Monte Carlo-Schätzer für θ ist daher

$$\widetilde{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} f(Y_i)\,\varrho(Y_i),$$

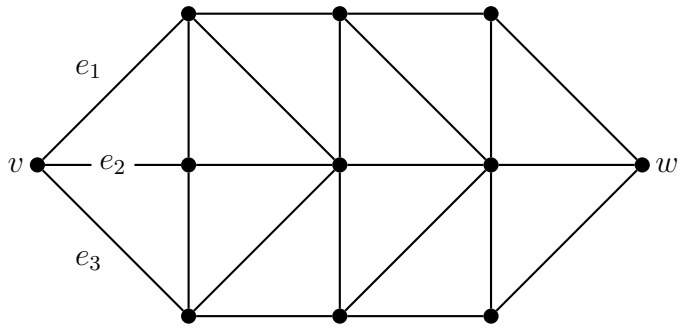


Abbildung 3.1: kleiner Beispielgraph für Perkolation

wobei die Y_i unabhängige Zufallsvariablen mit Verteilung ν sind. Auch $\widetilde{\theta}_n$ ist erwartungstreu:

$$E_\nu[\widetilde{\theta}_n] = E_\nu[f\,\varrho] = \theta.$$

Für die Varianz erhalten wir:

$$\mathrm{Var}_\nu(\widetilde{\theta}_n) = \frac{1}{n}\,\mathrm{Var}_\nu(f\,\varrho) = \frac{1}{n}\big(\sum_{x\in S} f(x)^2\varrho(x)^2\nu(x) - \theta^2\big).$$

Bei geeigneter Wahl von ν kann die Varianz von $\widetilde{\theta}_n$ deutlich kleiner sein als die des Schätzers $\widehat{\theta}_n$.
Faustregel für eine gute Wahl von ν : $\nu(x)$ sollte groß sein, wenn $|f(x)|$ groß ist.
»Importance Sampling«: Mehr Gewicht für die wichtigen x!

Beispiel (Zuverlässigkeit von Netzwerken; Perkolation)**.** Gegeben sei ein endlicher Graph (V, E), wobei V die Menge der Knoten und E die Menge der Kanten bezeichnet. Wir nehmen an, dass die Kanten unabhängig voneinander mit Wahrscheinlichkeit $\varepsilon \ll 1$ ausfallen. Seien $v, w \in E$ vorgegebene Knoten. Wir wollen die Wahrscheinlichkeit

$$p = P[\text{»}v \text{ nicht verbunden mit } w \text{ durch intakte Kanten«}]$$

approximativ berechnen. Sei

$$S = \{0,1\}^E = \{(x_e)_{e\in E} \mid x_e \in \{0,1\}\}$$

die Menge der Konfigurationen von intakten $(x_l = 0)$ bzw. defekten $(x_l = 1)$ Kanten und μ die Wahrscheinlichkeitsverteilung auf S mit Massenfunktion

$$\mu(x) = \varepsilon^{k(x)}(1-\varepsilon)^{|E|-k(x)}, \qquad k(x) = \sum_{e\in E} x_e.$$

Sei

$$A = \{x \in S \mid v, w \text{ nicht verbunden durch Kanten } e \text{ mit } x_e = 0\}.$$

Dann ist

$$p = \mu(A) = E_\mu[I_A].$$

Der »klassische Monte Carlo-Schätzer« für p ist

$$\widehat{p}_n \;=\; \frac{1}{n}\sum_{i=1}^{n} I_A(X_i), \quad X_i \text{ unabhängig mit Verteilung } \mu.$$

Fordern wir nun 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. Die Größenordnung von p für das in der obigen Graphik dargestellte Netzwerk mit $\varepsilon = 1\%$ lässt sich wie folgt abschätzen:

$$\begin{aligned} 10^{-6} = \mu(\text{»}e_1, e_2, e_3 \text{ versagen«}) \leq p \leq \mu(\text{»mindestens 3 Kanten versagen«}) \\ = \binom{22}{3} \cdot 10^{-6} \approx 1,5 \cdot 10^{-3}. \end{aligned}$$

Es sind also eventuell mehrere Millionen Stichproben nötig!
Um die benötigte Stichprobenanzahl zu reduzieren, wenden wir ein Importance Sampling-Verfahren an. Sei

$$\nu(x) = t^{-k(x)}\,(1-t)^{|E|-k(x)}, \qquad k(x) = \sum_{e \in E} x_e,$$

die Verteilung bei Ausfallwahrscheinlichkeit $t := \frac{3}{22}$. Da unter ν im Schnitt 3 Kanten defekt sind, ist der Ausfall der Verbindung bzgl. ν nicht mehr selten. Für den Schätzer

$$\widetilde{p}_n = \frac{1}{n}\sum_{i=1}^{n} I_A(Y_i)\frac{\mu(Y_i)}{\nu(Y_i)}, \qquad Y_i \text{ unabhängig mit Verteilung } \nu,$$

erhalten wir im Beispiel von oben:

$$\begin{aligned} \operatorname{Var}(\widetilde{p}_n) &= \frac{1}{n}\Big(\sum_{x \in S} I_A(x)^2 \frac{\mu(x)^2}{\nu(x)} - p^2\Big) \\ &\leq \frac{1}{n}\sum_{k=3}^{22} \binom{|E|}{k} \left(\frac{\varepsilon^2}{t}\right)^k \left(\frac{(1-\varepsilon)^2}{1-t}\right)^{|E|-k} \leq 0,0053\,\frac{p}{n}. \end{aligned}$$

Diese Abschätzung ist etwa um den Faktor 200 besser als die für den einfachen Monte Carlo-Schätzer erhaltene Abschätzung der Varianz.

3.4 Gleichgewichte von Markov-Ketten

Sei S eine abzählbare Menge, ν eine Wahrscheinlichkeitsverteilung auf S, und $p(x,y)$, $(x,y \in S)$, eine **stochastische Matrix** bzw. **Übergangsmatrix**, d.h. $p(x,y)$ erfüllt die folgenden Bedingungen:

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

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

Hier und im folgenden bezeichnen wir diskrete Wahrscheinlichkeitsverteilungen und die entsprechenden Massenfunktionen mit demselben Buchstaben, d.h. $\nu(x) := \nu(\{x\})$.

Definition. *Eine Folge $X_0, X_1, \ldots : \Omega \to S$ von Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ 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:

$$P[X_0 = x_0] = \nu(x_0)$$

(ii) Für alle $n \in \mathbb{N}$ und $x_0, \ldots, x_{n+1} \in S$ mit $P[X_0 = x_0, \ldots, X_n = x_n] \neq 0$ gilt:

$$P[X_{n+1} = x_{n+1} \mid X_0 = x_0, \ldots, X_n = x_n] = p(x_n, x_{n+1}).$$

Bemerkung. Die Bedingungen (i) und (ii) sind äquivalent zu:

$$P[X_0 = x_0, \ldots, X_n = x_n] = \nu(x_0)\, p(x_0, x_1) \cdots p(x_{n-1}, x_n) \qquad \text{für alle } n \in \mathbb{N}, x_i \in S.$$

Gleichgewichte und Stationarität

Für eine Wahrscheinlichkeitsverteilung μ mit Massenfunktion $\mu(x) = \mu(\{x\})$ und eine stochastische Matrix p auf S setzen wir

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

d.h. $\mu\, p$ ist der Zeilenvektor, den wir erhalten, wenn wir den Zeilenvektor $(\mu(x))_{x \in S}$ von links an die Matrix p multiplizieren.

Lemma 3.6. *i) Die Verteilung zur Zeit n einer Markov-Kette mit Startverteilung ν und Übergangsmatrix p ist $\nu\, p^n$.*

*ii) Gilt $\nu\, p = \nu$, dann folgt $X_n \sim \nu$ für alle $n \in \mathbb{N}$. **(»Stationarität«)***

Beweis. i) Wie im Beweis von Satz 2.4 erhalten wir

$$P[X_n = y \mid X_0 = x] = p^n(x, y)$$

für alle $n \in \mathbb{N}$ und $x, y \in S$ mit $P[X_0 = x] \neq 0$, und damit:

$$\begin{aligned} P[X_n = y] &= \sum_{\substack{x \in S \\ P[X_0 = x] \neq 0}} P[X_n = y \mid X_0 = x]\, P[X_0 = x] \\ &= \sum_{\substack{x \in S \\ \nu(x) \neq 0}} p^n(x, y)\, \nu(x) = (\nu\, p^n)(y). \end{aligned}$$

ii) Aus $\nu\, p = \nu$ folgt $\nu\, p^n = \nu$ für alle $n \in \mathbb{N}$.

□

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

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

*ii) μ erfüllt die **Detailed Balance-Bedingung** bzgl. der Übergangsmatrix p, falls gilt:*

$$\mu(x)\, p(x, y) = \mu(y)\, p(y, x) \qquad \textit{für alle } x, y \in S \tag{3.4.1}$$

Satz 3.7. *Erfüllt μ die Detailed Balance-Bedingung* (3.4.1)*, dann ist μ eine Gleichgewichtsverteilung von p.*

Beweis. Aus der Detailed Balance-Bedingung folgt:

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

□

Bemerkung. Bei Startverteilung μ gilt:

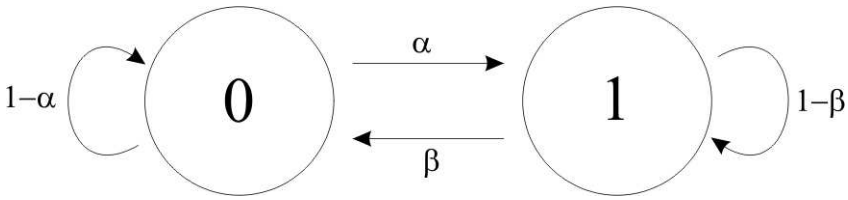
$$\mu(x)\, p(x, y) = P[X_0 = x, X_1 = y], \qquad \text{»Fluss von } x \text{ nach } y\text{«}.$$

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

$$\beta\,(1-\mu(0))=\alpha\,\mu(0).$$

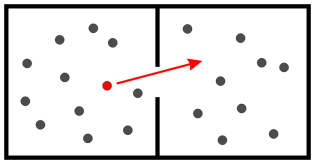
Die letzte Gleichung ist äquivalent zur Detailed Balance-Bedingung (3.4.1). Falls $\alpha+\beta>0$ gilt, ist $\mu=\left(\frac{\beta}{\alpha+\beta},\frac{\alpha}{\alpha+\beta}\right)$ die eindeutige Gleichgewichtsverteilung und erfüllt die Detailed Balance-Bedingung. Falls $\alpha=\beta=0$ gilt, ist jede Wahrscheinlichkeitsverteilung μ eine Gleichgewichtsverteilung mit Detailed Balance-Bedingung.

b) ZYKLISCHER RANDOM WALK: Sei $S=\mathbb{Z}/n\mathbb{Z}$ ein diskreter Kreis, und

$$p(k,k+1)=p,\qquad p(k,k-1)=1-p.$$

Die Gleichverteilung $\mu(x)=\frac{1}{n}$ ist ein Gleichgewicht. Die Detailed Balance-Bedingung ist dagegen nur für $p=\frac{1}{2}$, d.h. im symmetrischen Fall, erfüllt.

c) EHRENFEST-MODELL:



Sei $S=\{0,1,\dots,n\}$,

$$p(k,k-1)=\frac{k}{n},\qquad p(k,k+1)=\frac{n-k}{n}.$$

Man kann erwarten, dass sich im Gleichgewicht jede Kugel mit Wahrscheinlichkeit $\frac{1}{2}$ in jeder der beiden Urnen befindet. Tatsächlich erfüllt die Binomialverteilung $\mu(k) = \binom{n}{k} 2^{-n}$ mit Parameter $p = \frac{1}{2}$ die Detailed Balance-Bedingung:

$$\mu(k-1)\, p(k-1,k) = \mu(k)\, p(k,k-1) \qquad k = 1, \ldots, n$$

ist äquivalent zu

$$2^{-n} \frac{n!}{(k-1)!(n-(k-1))!} \frac{n-(k-1)}{n} = 2^{-n} \frac{n!}{k!(n-k)!} \frac{k}{n} \qquad k = 1, \ldots, n$$

d) RANDOM WALKS AUF GRAPHEN:
Sei (V, E) ein endlicher Graph, $S = V$ die Menge der Knoten.

- Sei
$$p(x,y) = \begin{cases} \frac{1}{\deg(x)} & \text{falls } \{x,y\} \in E, \\ 0 & \text{sonst.} \end{cases}$$
Die Detailed Balance-Bedingung lautet in diesem Fall:
$$\mu(x)\, p(x,y) = \mu(y)\, p(y,x).$$
Sie ist erfüllt, falls
$$\mu(x) = c\, \deg(x)$$
gilt, wobei c eine Konstante ist. Damit μ eine Wahrscheinlichkeitsverteilung ist, muss c so gewählt werden, dass gilt:
$$\sum_{x \in B} \deg(x) = 2\,|E|.$$
Somit ist die Gleichgewichtsverteilung:
$$\mu(x) = \frac{\deg(x)}{2|E|}.$$

- Sei $\triangle := \max_{x \in V} \deg(x)$,
$$p(x,y) = \begin{cases} \frac{1}{\triangle} & \text{falls } \{x,y\} \in E, \\ 1 - \frac{\deg(x)}{\triangle} & \text{sonst.} \end{cases}$$
Es gilt $p(x,y) = p(y,x)$ und somit ist die Gleichverteilung auf V die stationäre Verteilung.

Ist $\deg(x)$ konstant, dann stimmen die Random Walks in beiden Beispielen überein, und die Gleichverteilung ist ein Gleichgewicht.

Im nächsten Abschnitt zeigen wir:

Satz (Konvergenzsatz für Markov-Ketten)**.** *Ist S endlich, und p eine irreduzible und aperiodische stochastische Matrix mit Gleichgewicht μ, dann gilt für alle Wahrscheinlichkeitsverteilungen ν auf S:*

$$\lim_{n\to\infty} (\nu\, p^n)(x) = \mu(x) \qquad \textit{für alle } x \in S.$$

Aufgrund des Konvergenzsatzes können wir Stichproben von einer Wahrscheinlichkeitsverteilung μ näherungsweise erzeugen, indem wir eine Markov-Kette X_n mit Gleichgewicht μ simulieren, und für großes n auswerten. Wie findet man eine Markov-Kette mit einer vorgegebenen stationären Verteilung?

Metropolis-Algorithmus und Gibbs-Sampler

Die Metropolis-Kette

Sei $q(x, y)$ eine symmetrische stochastische Matrix, d.h. $q(x, y) = q(y, x)$ für alle $x, y \in S$. Dann erfüllt die Gleichverteilung die Detailed Balance-Bedingung (3.4.1). Sei nun μ eine beliebige Wahrscheinlichkeitsverteilung auf S mit $\mu(x) > 0$ für alle $x \in S$. Wie können wir die Übergangsmatrix q so modifizieren, dass die Detailed Balance-Bedingung bzgl. μ erfüllt ist?

Algorithmus 3.8 (Metropolis-Algorithmus (Update $x \to y$))**.** schlage Übergang $x \to y$ mit Wahrscheinlichkeit $q(x, y)$ vor
akzeptiere Übergang mit Wahrscheinlichkeit $\alpha(x, y) \in [0, 1]$
sonst verwerfe Vorschlag und bleibe bei x

ÜBERGANGSMATRIX:

$$p(x, y) := \begin{cases} \alpha(x, y)\, q(x, y) & \text{für } y \neq x, \\ 1 - \sum_{y\neq x} \alpha(x, y)\, q(x, y) & \text{für } y = x. \end{cases}$$

Die Detailed Balance-Bedingung lautet:

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

Sie ist äquivalent dazu, dass

$$b(x, y) := \mu(x)\, \alpha(x, y)$$

symmetrisch in x und y ist. Was ist die größtmögliche Wahl von $b(x,y)$?
Aus $\alpha(x,y) \leq 1$ folgen

$$b(x,y) \leq \mu(x),$$
$$b(x,y) = b(y,x) \leq \mu(y),$$

und somit

$$b(x,y) \leq \min(\mu(x), \mu(y)).$$

Der größtmögliche Wert $b(x,y) = \min(\mu(x), \mu(y))$ entspricht gerade

$$\alpha(x,y) = \min\big(1, \frac{\mu(y)}{\mu(x)}\big) = \begin{cases} 1 & \text{falls } \mu(y) \geq \mu(x), \\ \frac{\mu(y)}{\mu(x)} & \text{falls } \mu(x) \geq \mu(y). \end{cases}$$

Definition. *Die Markov-Kette mit Übergangsmatrix*

$$p(x,y) = \min\left(1, \frac{\mu(y)}{\mu(x)}\right) \cdot q(x,y) \qquad \textit{für } y \neq x$$

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

Satz 3.9. *μ erfüllt die Detailed Balance-Bedingung bzgl. p.*

Beweis. siehe oben. □

Der Gibbs-Sampler

Sei $S = S_1 \times \cdots \times S_d$ ein endlicher Produktraum, $\mu(x_1, \ldots, x_d)$ eine Wahrscheinlichkeitsverteilung auf S und

$$\mu_i(x_i \mid x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d) := \frac{\mu(x_1, \ldots, x_d)}{\sum_{z \in S_i} \mu(x_1, \ldots, x_{i-1}, z, x_{i+1}, \ldots, x_d)}$$

die bedingte Verteilung der i-ten Komponente gegeben die übrigen Komponenten.

Algorithmus 3.10 (Gibbs-Sampler (Update $x \to y$))**.** $y := x$
for $i := 1, \ldots d$ **do**
update $y_i \sim \mu_i(\,\bullet \mid y_1, \ldots y_{i-1}, y_{i+1}, \ldots, y_d)$
end for
return y

ÜBERGANGSMATRIX:

$$p = p_d\, p_{d-1} \cdots p_1,$$

wobei

$$p_i(x,y) = \begin{cases} \mu_i(y_i \mid y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_d) & \text{falls } y_k = x_k \text{ für alle } k \neq i, \\ 0 & \text{sonst.} \end{cases}$$

Satz 3.11. *i) μ erfüllt die Detailed Balance-Bedingung bzgl. p_i für alle $i = 1, \ldots, d$.*

ii) μ ist ein Gleichgewicht von p.

Beweis. i) Der Beweis der ersten Aussage ist eine Übungsaufgabe.

ii) Nach der ersten Aussage ist μ ein Gleichgewicht von p_i für alle i. Also gilt auch

$$\mu\, p = \mu\, p_d\, p_{d-1} \cdots p_1 = \mu.$$

□

Bemerkung. Zur Simulation von Y_n, $n \geq 0$, genügt es, die Massenfunktion $\mu(x)$ bis auf eine multiplikative Konstante zu kennen:
aus $\mu(x) = C\, f(X)$ folgt

$$\alpha(x,y) = \min\big(1, \frac{f(y)}{f(x)}\big) \qquad \text{unabhängig von } C.$$

Beispiel (Rucksackproblem)**.** Gegeben:

$$\begin{aligned} \omega_1, \ldots, \omega_d &\in \mathbb{R}, \quad \text{»Gewichte«}, \\ v_1, \ldots, v_d &\in \mathbb{R}, \quad \text{»Werte«}. \end{aligned}$$

Rucksack mit maximalem Gewicht $b > 0$, packe soviel Wert wie möglich ein.

$$\begin{aligned} S &= \{0,1\}^d, && \text{alle Konfigurationen}, \\ S_b &= \{(z_1, \ldots, z_d) \in S : \textstyle\sum_{i=1}^d z_i\, w_i \leq b\}, && \text{zulässige Konfigurationen}, \\ & z_i = 1 : i\text{-ter Gegenstand im Rucksack}. \end{aligned}$$

RUCKSACKPROBLEM:

$$\text{maximiere } V(z) = \textstyle\sum_{i=1}^d z_i\, v_i \text{ unter Nebenbedingung } z \in S_b.$$

Das Rucksackproblem ist **NP-vollständig**, insbesondere ist keine Lösung in $O(d^k)$ Schritten für ein $k \in \mathbb{N}$ bekannt.

STOCHASTISCHER ZUGANG: SIMULATED ANNEALING

Für $\beta > 0$ betrachten wir die Wahrscheinlichkeitsverteilung

$$\mu_\beta(z) = \begin{cases} \frac{1}{Z_\beta} e^{\beta V(z)} & \text{für } z \in S_b, \\ 0 & \text{für } z \in S \backslash S_b, \end{cases}$$

auf S, wobei $Z_\beta = \sum_{z \in S_b} e^{\beta V(z)}$ eine Konstante ist, die μ zu einer Wahrscheinlichkeitsverteilung normiert. Für $\beta = 0$ ist μ_β die Gleichverteilung auf S_b. Für $\beta \to \infty$ konvergiert μ_β gegen die Gleichverteilung auf der Menge der globalen Maxima von V, denn:

$$\mu_\beta(z) = \frac{e^{\beta V(z)}}{Z_\beta} = \frac{1}{\sum_{y \in S_b} e^{\beta (V(y) - V(z))}} \longrightarrow \begin{cases} 0 & \text{falls } V(z) \neq \max V, \\ \frac{1}{|\{y \mid V(y) = \max V\}|} & \text{falls } V(z) = \max V. \end{cases}$$

IDEE: Simuliere Stichprobe z von μ_β für β groß ($\beta \to \infty$). Dann ist $V(z)$ **wahrscheinlich** nahe dem Maximalwert.

METROPOLIS-ALGORITHMUS: Sei $x^+ := \max(x, 0)$ der Positivteil von x. Wir wählen als Vorschlagsmatrix die Übergangsmatrix

$$q(z, w) := \begin{cases} \frac{1}{d} & \text{falls } z_i \neq w_i \text{ für genau ein } i \in \{1, \ldots, d\}, \\ 0 & \text{sonst,} \end{cases}$$

des Random Walks auf $\{0, 1\}^d$. Für die Akzeptanzwahrscheinlichkeit ergibt sich

$$\alpha_\beta(z, w) = \min\left(1, \frac{\mu_\beta(w)}{\mu_\beta(z)}\right) = \begin{cases} e^{-\beta (V(z) - V(w))} & \text{für } z, w \in S_b, \\ 0 & \text{für } z \in S_b, w \notin S_b. \end{cases}$$

Der Vorschlag w wir also mit Wahrscheinlichkeit 1 akzeptiert, wenn $V(w) \geq V(z)$ gilt – andernfalls wird der Vorschlag nur mit Wahrscheinlichkeit $\exp -\beta (V(z) - V(w))$ akzeptiert.

Algorithmus 3.12 (Simulation einer Markov-Kette mit Gleichgewicht μ_β)**.** initialisiere $z^{(0)} \in S_b$

for $n = 1, 2, \ldots$ **do**

  $z^{(n)} := w := z^{(n-1)}$

  erzeuge $i \sim \mathrm{Unif}\{1, \ldots, d\}$

  $w_i := 1 - w_i$

  if $w \in S_b$ **then**

erzeuge $u \sim \mathrm{Unif}(0,1)$
if $u \leq \alpha_\beta(z,w)$ **then**
$z^{(n)} := w$
end if
end if
end for

Algorithmus 3.13 (Simulated Annealing)**.** Wie Algorithmus 3.12 aber mit $\beta = \beta(n) \to \infty$ für $n \to \infty$.

Bemerkung. a) PHYSIKALISCHE INTERPRETATIONEN:
μ_β ist die Verteilung im thermodynamischen Gleichgewicht für die Energiefunktion $H(z) = -V(z)$ bei der Temperatur $T = 1/\beta$. Der Grenzwert $\beta \to \infty$ entspricht $T \to 0$ (»simuliertes Abkühlen«).

b) Die beim Simulated Annealing-Verfahren simulierte zeitlich inhomogene Markov-Kette findet im allgemeinen nicht das globale Maximum von V, sondern kann in lokalen Maxima »steckenbleiben«. Man kann zeigen, dass die Verteilung der Markov-Kette zur Zeit n gegen die Gleichverteilung auf den Maximalstellen konvergiert, falls $\beta(n)$ nur sehr langsam (logarithmisch) gegen $+\infty$ geht. In praktischen Anwendungen wird der Algorithmus aber in der Regel mit einem schnelleren »Cooling schedule« $\beta(n)$ verwendet. Das Auffinden eines globalen Maximums ist dann nicht garantiert – trotzdem erhält man ein oft nützliches **heuristisches** Verfahren.

3.5 Konvergenz ins Gleichgewicht

Sei $S = \{x_1, \ldots, x_m\}$ eine endliche Menge, und

$$\mathrm{WV}(S) := \{\mu = (\mu(x_1), \ldots, \mu(x_m)) \mid \mu(x_i) \geq 0, \quad \sum_{i=1}^{m} \mu(x_i) = 1\} \subseteq \mathbb{R}^m$$

die Menge aller Wahrscheinlichkeitsverteilungen auf S. Geometrisch ist $\mathrm{WV}(S)$ ein Simplex im $\mathbb{R}^m$. Wir führen nun einen Abstandsbegriff auf $\mathrm{WV}(S)$ ein:

Definition. *Die **Variationsdistanz** zweier Wahrscheinlichkeitsverteilungen μ, ν auf S ist:*

$$d_{TV}(\mu,\nu) := \frac{1}{2}\|\mu - \nu\|_1 = \frac{1}{2}\sum_{x \in S} |\mu(x) - \nu(x)|.$$

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 μ, ν Wahrscheinlichkeitsverteilungen und $A := \{x \in S \mid \mu(x) \geq \nu(x)\}$. Dann gilt:

$$d_{TV}(\mu, \nu) = \sum_{x \in A} (\mu(x) - \nu(x)) = \max_{A \subseteq S} |\mu(A) - \nu(A)|.$$

Der Beweis dieser Aussage ist eine Übungsaufgabe.

Wir betrachten im folgenden eine stochastische Matrix $p(x, y)$, $(x, y \in S)$, mit Gleichgewicht μ. Die Verteilung einer Markov-Kette mit Startverteilung ν und Übergangsmatrix p zur Zeit n ist $\nu\, p^n$. 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 für alle $x, y \in S$ gilt:

$$p^r(x, y) \geq \delta \cdot \mu(y). \tag{3.5.1}$$

Satz 3.14. *Gilt die Minorisierungsbedingung* (3.5.1)*, dann konvergiert $\nu\, p^n$ für jede Startverteilung ν exponentiell schnell gegen μ. Genauer gilt für alle $n \in \mathbb{N}$ und $\nu \in \mathrm{WV}(S)$:*

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

Bemerkung. Insbesondere ist μ das **eindeutige** Gleichgewicht: Betrachte eine beliebige Wahrscheinlichkeitsverteilung ν mit $\nu\, p = \nu$. Dann folgt für $n \to \infty$:

$$d_{TV}(\nu, \mu) = d_{TV}(\nu\, p^n, \mu) \longrightarrow 0,$$

also $d_{TV}(\mu, \nu) = 0$, und somit $\mu = \nu$.

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 (3.5.1) 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 $\nu \in \mathrm{WV}(S)$:

$$\nu\, p^r = (1-\lambda)\,\mu + \lambda\,\nu\, q. \tag{3.5.2}$$

2. Wir wollen mit vollständiger Induktion zeigen:

$$\nu\, p^{kr} = (1-\lambda^k)\,\mu + \lambda^k\,\nu\, q^k \qquad \text{für alle } k \geq 0, \quad \nu \in \mathrm{WV}(S). \tag{3.5.3}$$

Für $k = 0$ ist die Aussage offensichtlich wahr. Gilt (3.5.3) für ein $k \geq 0$, dann erhalten wir durch Anwenden von Gleichung (3.5.2) 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. Für $n \in \mathbb{N}$, $n = k\,r + i$, $(k \in \mathbb{N}, 0 \leq i < r)$, folgt:

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

also

$$\nu\, p^n - \mu = \lambda^k\,(\nu\, q^k\, p^i - \mu) \qquad \text{für alle } \nu \in \mathrm{WV}(S),$$

und damit

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

nach der letzten Bemerkung.

□

Welche Übergangsmatrizen erfüllen die Minorisierungsbedingung?

Definition. *i) Die 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) Die **Periode** von $x \in S$ ist definiert als*

$$\mathrm{Periode}(x) := \mathrm{ggT}(\underbrace{\{n \in \mathbb{N} \mid p^n(x,x) > 0\}}_{=:R(x)}).$$

Lemma 3.15. *i) Falls p irreduzibel ist, gilt* $\mathrm{Periode}(x) = \mathrm{Periode}(y)$ *für alle* $x, y \in S$.

ii) Falls p irreduzibel und aperiodisch (d.h. $\mathrm{Periode}(x) = 1$ *für alle $x \in S$) ist, gibt es ein $r > 0$, so dass $p^r(x, y) > 0$ für alle $x, y \in S$ gilt.*

Beweis. Seien $x, y \in S$.

i) Sei p irreduzibel. Dann gibt es ein s und ein $t \in \mathbb{N}$, so dass gilt:

$$p^s(x, y) > 0 \qquad \text{und} \qquad p^t(y, x) > 0.$$

Für $a := s + t$ folgt:

- $p^a(x, x) \geq p^s(x, y)\, p^t(y, x) > 0$, also $a \in R(x)$.
- $p^{n+a}(x, x) \geq p^s(x, y)\, p^n(y, y)\, p^t(y, x) > 0$ für alle $n \in R(y)$, also $n + a \in R(x)$ für alle $n \in R(y)$.

$\mathrm{Periode}(x)$ ist ein gemeinsamer Teiler von $R(x)$, somit Teiler von a und $n + a$, also auch von n für alle $n \in R(y)$. Daher ist $\mathrm{Periode}(x)$ ein gemeinsamer Teiler von $R(y)$ und somit gilt:

$$\mathrm{Periode}(x) \leq \mathrm{Periode}(y).$$

»$\geq$« wird analog gezeigt. Es folgt:

$$\mathrm{Periode}(x) = \mathrm{Periode}(y).$$

ii) $R(x)$ ist abgeschlossen unter Addition, denn falls $s, t \in R(x)$ ist, gilt:

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

und somit $s + t \in R(x)$. Da p aperiodisch ist, folgt $\mathrm{ggT}(R(x)) = 1$ für alle $x \in S$. Nach einem Satz der Zahlentheorie gilt:
Da $R(x)$ additiv abgeschlossen, gibt es für alle x ein $r(x) \in \mathbb{N}$ mit $n \in R(x)$ für alle $n \geq r(x)$.
$n \in R(x)$ impliziert $p^n(x, x) > 0$. Da p irreduzibel ist, folgt, dass es für alle x, y ein $r(x, y) \in \mathbb{N}$ gibt, so dass gilt:

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

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

□

Satz 3.16 (Konvergenzsatz für **endliche** Markov-Ketten)**.** *Ist p irreduzibel und aperiodisch mit Gleichgewicht μ, dann gilt:*

$$\lim_{n\to\infty} d_{TV}(\nu p^n, \mu) = 0 \qquad \textit{für alle } \nu \in \mathrm{WV}(S).$$

Beweis. Da p irreduzibel und aperiodisch ist, gibt es ein $r \in \mathbb{N}$ mit:

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

Daher gibt es ein $r \in \mathbb{N}$ und ein $\delta > 0$, so dass gilt:

$$p^r(x,y) > \delta\, \mu(y) \qquad \text{für alle } x, y \in S,$$

(z.B. $\delta := \min_{x,y\in S} p^r(x,y)$). Mit Satz 3.14 folgt die Behauptung. □

Beispiel (Metropolis-Kette)**.** Sei S endlich, $\mu(x) > 0$ für alle $x \in S$, nicht konstant, und $q(x,y)$ irreduzibel. Dann ist $p(x,y)$ irreduzibel und aperiodisch. Somit folgt die Konvergenz ins Gleichgewicht nach Satz 3.16, allerdings evtl. sehr langsam!

ANWENDUNG: MARKOV-CHAIN-MONTE CARLO-VERFAHREN
Sei $\mu \in \mathrm{WV}(S)$, $f\colon S \to \mathbb{R}$.
GESUCHT:

$$\theta = E_\mu[f],$$

MARKOV-CHAIN-MONTE CARLO-SCHÄTZER:

$$\widehat{\theta}_{n,b} = \frac{1}{n} \sum_{k=b+1}^{b+n} f(X_k),$$

wobei $b \in \mathbb{N}$ eine feste Konstante (»burn-in-Zeit«) und $(X_k)_{k\in\mathbb{N}}$ irreduzible Markov-Ketten mit Gleichgewicht μ sind.

Satz (Ergodensatz / Gesetz der großen Zahlen für Markov-Ketten)**.** *: Für alle $b \in \mathbb{N}$ gilt:*

$$\lim_{n\to\infty} \widehat{\theta}_{n,b} = \theta \quad \textit{mit Wahrscheinlichkeit } 1,$$

Beweis. siehe Vorlesung »Stochastische Prozesse«. □

Die Analyse des Schätzfehler ist im Allgemeinen diffizil!

Kapitel 4

Stetige und Allgemeine Modelle

4.1 Unendliche Kombinationen von Ereignissen

Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum. Ist $(A_n)_{n\in\mathbb{N}}$ eine Folge von bzgl. P unabhängigen Ereignissen, $A_n \in \mathcal{A}$ mit fester Wahrscheinlichkeit

$$P[A_n] = p \in [0, 1]$$

und

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

die Anzahl der Ereignisse unter den ersten n, die eintreten, dann ist S_n binomialverteilt mit den Parametern n und p. Für die relative Häufigkeit $\frac{S_n}{n}$ der Ereignisse A_i gilt die Bernstein-Chernoff-Ungleichung

$$P\left[\left|\frac{S_n}{n} - p\right| \geq \varepsilon\right] \leq 2 \cdot e^{-2\varepsilon^2 n}, \tag{4.1.1}$$

d.h. die Verteilung von $\frac{S_n}{n}$ konzentriert sich für $n \to \infty$ sehr rasch in der Nähe von p, siehe Abschnitt 2.3. Insbesondere ergibt sich ein Spezialfall des schwachen Gesetzes der großen Zahlen: die Folge der Zufallsvariablen $\frac{S_n}{n}$ konvergiert P-stochastisch gegen p, d.h.

$$P\left[\left|\frac{S_n}{n} - p\right| \geq \varepsilon\right] \overset{n\to\infty}{\to} 0 \text{ für alle } \varepsilon > 0.$$

Definition. *(1). Eine P-**Nullmenge** ist ein Ereignis $A \in \mathcal{A}$ mit $P[A] = 0$.*

*(2). Ein Ereignis $A \in \mathcal{A}$ tritt P-**fast sicher** bzw. für P-**fast alle** $\omega \in \Omega$ ein, falls $P[A] = 1$ gilt, d.h. falls A^C eine P-Nullmenge ist.*

Wir wollen nun Methoden entwickeln, die es uns ermöglichen, zu zeigen, dass aus (4.1.1) sogar

$$\lim_{n\to\infty} \frac{S_n(\omega)}{n} = p \qquad \text{für } P\text{-fast alle } \omega \in \Omega \tag{4.1.2}$$

folgt. Das relevante Ereignis

$$L := \left\{ \omega \in \Omega : \lim_{n\to\infty} \frac{S_n(\omega)}{n} = p \right\}$$

lässt sich offensichtlich nicht durch endlich viele der A_i beschreiben.

Seien nun allgemein $A_1, A_2, \ldots \in \mathcal{A}$ beliebige Ereignisse. Uns interessieren zusammengesetzte Ereignisse wie z.B.

$\bigcup_{n=1}^{\infty} A_n$ („Eines der A_n tritt ein“)

$\bigcap_{n=1}^{\infty} A_n$ („Alle der A_n treten ein“)

$\bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n = \{\omega \in \Omega : \forall m \quad \exists n \geq m : \omega \in A_n\}$ („Unendlich viele der A_n treten ein“ oder „A_n tritt immer mal wieder ein“)

$\bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} A_n = \{\omega \in \Omega : \exists m \quad \forall n \geq m : \omega \in A_n\}$ („A_n tritt schließlich ein“)

Aufgrund der Eigenschaften einer σ-Algebra liegen alle diese Mengen wieder in $\mathcal{A}$. Das Ereignis L lässt sich wie folgt als abzählbare Kombination der A_i ausdrücken:

$$\begin{aligned}
\omega \in L \quad &\iff \quad \lim_{n\to\infty} \frac{S_n}{n} = p \\
&\iff \quad \forall \varepsilon \in \mathbb{Q}_+ : \left| \frac{S_n}{n} - p \right| \leq \varepsilon \text{ schließlich} \\
&\iff \quad \forall \varepsilon \in \mathbb{Q}_+ \quad \exists m \in \mathbb{N} \quad \forall n \geq m : \left| \frac{S_n}{n} - p \right| \leq \varepsilon
\end{aligned}$$

Somit gilt

$$\begin{aligned}
L &= \bigcap_{\varepsilon \in \mathbb{Q}_+} \left\{ \left| \frac{S_n}{n} - p \right| \leq \varepsilon \text{ schließlich} \right\} \\
&= \bigcap_{\varepsilon \in \mathbb{Q}_+} \bigcup_{m \in \mathbb{N}} \bigcap_{n \geq m} \left\{ \left| \frac{S_n}{n} - p \right| \leq \varepsilon \right\}.
\end{aligned}$$

Um Wahrscheinlichkeiten von solchen Ereignissen berechnen zu können, ist es wesentlich, dass eine Wahrscheinlichkeitsverteilung P nicht nur endlich additiv, sondern sogar σ-additiv ist. Der folgende Satz gibt eine alternative Charakterisierung der σ-Additivität:

Satz 4.1 (σ-Additivität und monotone Stetigkeit). *Sei $\mathcal{A}$ eine σ-Algebra und $P : \mathcal{A} \to [0, \infty]$ additiv, d.h.*

$$A \cap B = \emptyset \;\Rightarrow\; P[A \cup B] = P[A] + P[B].$$

(i) P ist σ-additiv genau dann, wenn:

$$A_1 \subseteq A_2 \subseteq \ldots \;\Rightarrow\; P\left[\bigcup_{n=1}^{\infty} A_n\right] = \lim_{n\to\infty} P[A_n]$$

(ii) Gilt $P[\Omega] = 1$, dann ist dies auch äquivalent zu:

$$A_1 \supseteq A_2 \supseteq \ldots \;\Rightarrow\; P\left[\bigcap_{n=1}^{\infty} A_n\right] = \lim_{n\to\infty} P[A_n]$$

Beweis. (i) Sei P σ-additiv und $A_1 \subseteq A_2 \subseteq \ldots$. Die Mengen $B_1 := A_1$, $B_2 := A_2 \backslash A_1$, $B_3 := A_3 \backslash A_2, \ldots$ sind disjunkt mit

$$\bigcup_{i=1}^{n} B_i = \bigcup_{i=1}^{n} A_i = A_n \quad \text{und} \quad \bigcup_{i=1}^{\infty} B_i = \bigcup_{i=1}^{\infty} A_i.$$

Also gilt:

$$\begin{aligned} P\left[\bigcup_{i=1}^{\infty} A_i\right] &= P\left[\bigcup_{i=1}^{\infty} B_i\right] \\ &\overset{\sigma-add.}{=} \sum_{i=1}^{\infty} P[B_i] \\ &= \lim_{n\to\infty} \sum_{i=1}^{n} P[B_i] \\ &= \lim_{n\to\infty} P\left[\bigcup_{i=1}^{n} B_i\right] \\ &= \lim_{n\to\infty} P[A_n]. \end{aligned}$$

Der Beweis der umgekehrten Implikation wird dem Leser als Übungsaufgabe überlassen.

(ii) Gilt $P[\Omega] = 1$, dann folgt

$$P\left[\bigcap_{i=1}^{\infty} A_i\right] \;=\; P\left[\left(\bigcup_{i=1}^{\infty} A_i^c\right)^c\right] \;=\; 1 - P\left[\bigcup_{i=1}^{\infty} A_i^c\right].$$

Die Behauptung folgt nun aus (i).

□

Ab jetzt setzen wir wieder voraus, dass P eine Wahrscheinlichkeitsverteilung ist. Eine weitere Folgerung aus der σ-Additivität ist:

Satz 4.2 (σ**-Subadditivität**). *Für beliebige Ereignisse $A_1, A_2, \ldots \in \mathcal{A}$ gilt:*

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

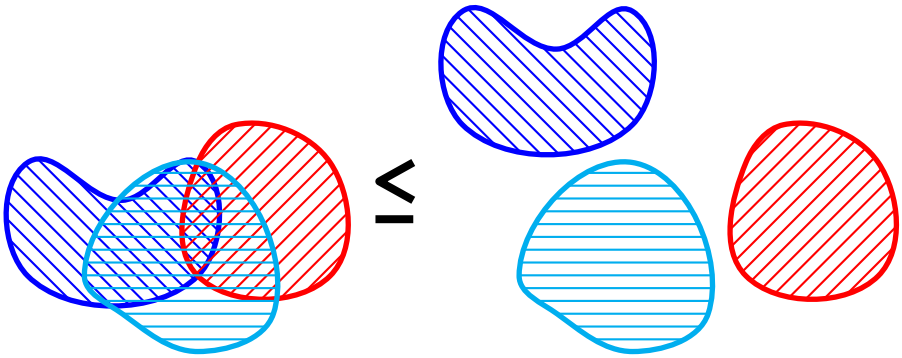


Abbildung 4.1: Darstellung von drei Mengen. Das Maß der Vereinigung von Mengen ist stets kleiner gleich als die Summe der Maße der einzelnen Mengen.

Beweis. Die Mengen

$$B_n = A_n \setminus (A_{n-1} \cup \cdots \cup A_1)$$

sind disjunkt mit $\bigcup_{n=1}^{\infty} B_n = \bigcup_{n=1}^{\infty} A_n$. Also gilt:

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

□

Bemerkung. Insbesondere ist eine Vereinigung von abzählbar vielen Nullmengen wieder eine Nullmenge.

Der folgende Satz spielt eine zentrale Rolle beim Beweis von Konvergenzaussagen für Zufallsvariablen:

Satz 4.3 (**1. Borel - Cantelli - Lemma**). *Für Ereignisse $A_1, A_2, \ldots \in \mathcal{A}$ mit*

$$\sum_{n=1}^{\infty} P[A_n] < \infty$$

gilt:

$$P[\text{„unendlich viele der } A_n \text{ treten ein“}] = P\left[\bigcap_m \bigcup_{n\geq m} A_n\right] = 0.$$

Beweis. Da die Folge $\bigcup_{n\geq m} A_n =: B_m$ von Ereignissen aus $\mathcal{A}$ monoton fallend ist, ergibt sich nach Satz 4.1 und 4.2:

$$\begin{aligned} P\left[\bigcap_m \bigcup_{n\geq m} A_n\right] &= P\left[\bigcap_m B_m\right] \\ &\overset{4.1}{=} \lim_{m\to\infty} P[B_m] \\ &= \lim_{m\to\infty} \underbrace{P\left[\bigcup_{n\geq m} A_n\right]}_{\overset{4.2}{\leq} \sum_{n=m}^{\infty} P[A_n]} \\ &\leq \liminf_{m\to\infty} \underbrace{\sum_{n=m}^{\infty} P[A_n]}_{\overset{m\to\infty}{\to} 0} = 0, \end{aligned}$$

da die Summe $\sum_{n=1}^{\infty} P[A_n]$ nach Voraussetzung konvergiert. □

Das erste Borel-Cantelli-Lemma besagt, dass mit Wahrscheinlichkeit 1 nur endlich viele der Ereignisse $A_n, n \in \mathbb{N}$ eintreten, falls $\sum P[A_n] < \infty$ gilt. Die Unabhängigkeit der Ereignisse ermöglicht die Umkehrung dieser Aussage. Es gilt sogar:

Satz 4.4 (2. Borel - Cantelli - Lemma). *Für unabhängige Ereignisse $A_1, A_2, \ldots \in \mathcal{A}$ mit*

$$\sum_{n=1}^{\infty} P[A_n] = \infty$$

gilt:

$$P[A_n \text{ unendlich oft}] = P\left[\bigcap_m \bigcup_{n\geq m} A_n\right] = 1$$

Bemerkung. Insbesondere ergibt sich ein ***0-1 Gesetz***:
Sind $A_1, A_2, \ldots \in \mathcal{A}$ unabhängige Ereignisse, dann beträgt die Wahrscheinlichkeit, dass unendlich viele der $A_n, n \in \mathbb{N}$, eintreten, entweder 0 oder 1 - je nachdem ob die Summe $\sum P[A_n]$ endlich oder unendlich ist.

Wir zeigen nun das zweite Borel-Cantelli-Lemma:

Beweis. Sind die Ereignisse $A_n, n \in \mathbb{N}$ unabhängig, so auch die Ereignisse A_n^C, siehe Lemma 2.5. Zu zeigen ist:

$$P[A_n \text{ nur endlich oft}] = P\left[\bigcup_m \bigcap_{n\geq m} A_n^C\right] = 0$$

Nach Satz 4.1 gilt:

$$P\left[\bigcup_m \bigcap_{n\geq m} A_n^C\right] = \lim_{m\to\infty} P\left[\bigcap_{n\geq m} A_n^C\right] \tag{4.1.3}$$

Wegen der Unabhängigkeit der Ereignisse A_n^C erhalten wir zudem

$$\begin{aligned} P\left[\bigcap_{n\geq m} A_n^C\right] &\overset{\text{mon. Stetigkeit}}{=} \lim_{k\to\infty} P\left[\bigcap_{n=m}^{k} A_n^C\right] \\ &\overset{\text{unabh.}}{=} \lim_{k\to\infty} \prod_{n=m}^{k} \underbrace{P[A_n^C]}_{=1-P[A_n]\leq \exp(-P[A_n])} \\ &\leq \liminf_{k\to\infty} \prod_{n=m}^{k} e^{-P[A_n]} \\ &= \liminf_{k\to\infty} e^{-\sum_{n=m}^{k} P[A_n]} = 0, \end{aligned} \tag{4.1.4}$$

da $\lim_{k\to\infty} \sum_{n=m}^{k} P[A_n] = \sum_{n=m}^{\infty} P[A_n] = \infty$ nach Voraussetzung.
Aus 4.1.3 und 4.1.4 folgt die Behauptung. □

Mithilfe des 1. Borel-Cantelli-Lemmas können wir nun eine erste Version eines starken Gesetzes großer Zahlen beweisen. Sei $p \in [0, 1]$.

Satz 4.5 (Starkes Gesetz großer Zahlen I, Borel 1909, Hausdorff 1914, Cantelli 1917). *Sind $A_1, A_2, \ldots \in \mathcal{A}$ unabhängige Ereignisse mit Wahrscheinlichkeit $P[A_n] = p$ für alle $n \in \mathbb{N}$, dann gilt für $S_n = \sum_{i=1}^{n} I_{A_i}$:*

$$\underbrace{\lim_{n\to\infty} \frac{S_n(\omega)}{n}}_{\substack{\text{asymptotische relative Häufigkeit des Ereignisses}}} = \underbrace{p}_{\text{W'keit}} \quad \text{für } P\text{-fast alle } \omega \in \Omega$$

Beweis. Sei

$$L := \left\{ \omega \in \Omega \,\middle|\, \frac{1}{n} S_n(\omega) \to p \text{ für } n \to \infty \right\}$$

Zu zeigen ist, dass $L^C \in \mathcal{A}$ mit $P[L^C] = 0$.
Wegen

$$\omega \in L^C \iff \frac{S_n(\omega)}{n} \not\to p \iff \exists \varepsilon \in \mathbb{Q}_+ : \left| \frac{S_n(\omega)}{n} - p \right| > \varepsilon \quad \text{unendlich oft}$$

gilt:

$$L^C = \bigcup_{\varepsilon \in \mathbb{Q}_+} \left\{ \left| \frac{S_n(\omega)}{n} - p \right| > \varepsilon \quad \text{unendlich oft} \right\} = \bigcup_{\varepsilon \in \mathbb{Q}_+} \bigcap_m \bigcup_{n \geq m} \left\{ \left| \frac{S_n(\omega)}{n} - p \right| > \varepsilon \right\} \in \mathcal{A}.$$

Zudem folgt aus der Bernstein-Chernoff-Abschätzung:

$$\sum_{n=1}^{\infty} P\left[\left| \frac{S_n}{n} - p \right| > \varepsilon \right] \leq \sum_{n=1}^{\infty} 2e^{-2n\varepsilon^2} < \infty$$

für alle $\varepsilon > 0$, also nach dem 1. Borel-Cantelli-Lemma:

$$P\left[\left| \frac{S_n}{n} - p \right| > \varepsilon \text{ unendlich oft} \right] = 0.$$

Also ist L^c eine Vereinigung von abzählbar vielen Nullmengen, und damit nach Satz 4.2 selbst eine Nullmenge. □

Das starke Gesetz großer Zahlen in obigem Sinn rechtfertigt nochmals im Nachhinein die empirische Interpretation der Wahrscheinlichkeit eines Ereignisses als asymptotische relative Häufigkeit bei unabhängigen Wiederholungen.

Beispiel (Random Walk/Irrfahrt). 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

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

Die Ereignisse $A_i := \{X_i = 1\}$ sind unabhängig mit $P[A_i] = p$ 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 Satz 4.5 folgt:

$$\lim_{n\to\infty} \frac{Z_n}{n} = 2 \lim_{n\to\infty} \frac{S_n}{n} - 1 = 2p - 1 \; P\text{-fast sicher.}$$

Für $p \neq \frac{1}{2}$ wächst (bzw. fällt) Z_n also mit Wahrscheinlichkeit 1 asymptotisch linear (siehe Abbildung 4.2):

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

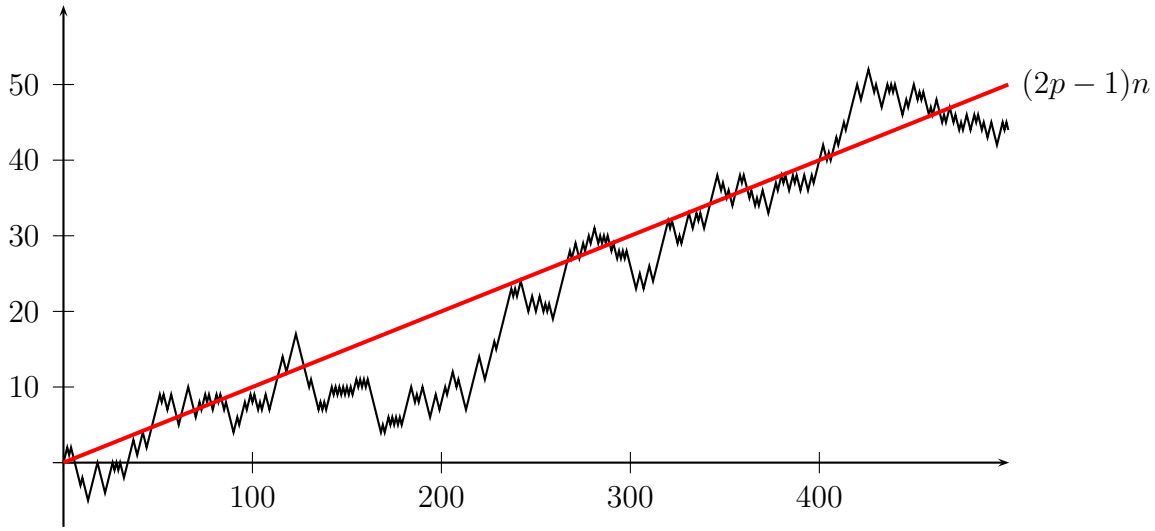


Abbildung 4.2: Random Walk mit Drift: $p = 0.55, n = 500$

Für $p = \frac{1}{2}$ dagegen wächst der Random Walk sublinear, d.h. $\frac{Z_n}{n} \to 0$ 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 4.3).

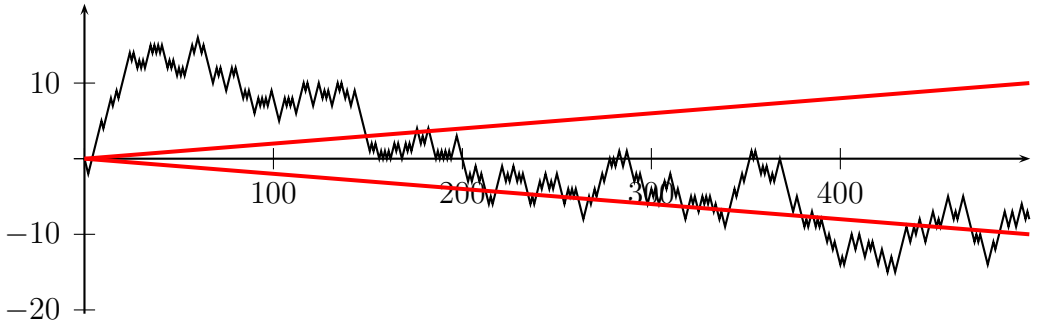


Abbildung 4.3: Random Walk ohne Drift: $p = 0.5, n = 500$

Eine viel präzisere Beschreibung der Asymptotik des Random Walk liefert der **Satz vom iterierten Logarithmus**:

$$\limsup_{n\to\infty} \frac{S_n(\omega)}{\sqrt{n \log\log n}} = +1 \; P\text{-fast sicher,}$$

$$\liminf_{n\to\infty} \frac{S_n(\omega)}{\sqrt{n \log\log n}} = -1 \; P\text{-fast sicher}$$

Mehr dazu: siehe Vorlesung „Stochastische Prozesse."

4.2 Allgemeine Wahrscheinlichkeitsräume

Bisher haben wir uns noch nicht mit der Frage befasst, ob überhaupt ein Wahrscheinlichkeitsraum existiert, auf dem unendlich viele unabhängige Ereignisse bzw. Zufallsvariablen realisiert werden können. Auch die Realisierung einer auf einem endlichen reellen Intervall gleichverteilten Zufallsvariable auf einem geeigneten Wahrscheinlichkeitsraum haben wir noch nicht gezeigt. Die Existenz solcher Räume wurde stillschweigend vorausgesetzt.

Tatsächlich ist es oft nicht notwendig, den zugrunde liegenden Wahrscheinlichkeitsraum explizit zu kennen - die Kenntnis der gemeinsamen Verteilungen aller relevanten Zufallsvariablen genügt, um Wahrscheinlichkeiten und Erwartungswerte zu berechnen. Dennoch ist es an dieser Stelle hilfreich, die grundlegenden Existenzfragen zu klären, und unsere Modelle auf ein sicheres Fundament zu stellen. Die dabei entwickelten Begriffsbildungen werden sich beim Umgang mit stetigen und allgemeinen Zufallsvariablen als unverzichtbar erweisen.

Beispiele von Wahrscheinlichkeitsräumen

Wir beginnen mit einer Auflistung von verschiedenen Wahrscheinlichkeitsräumen $(\Omega, \mathcal{A}, P)$, die wir gerne konstruieren würden:

Dirac-Maß

Sei Ω beliebig, $a \in \Omega$ fest, $\mathcal{A} = \mathcal{P}(\Omega), P = \delta_a$, wobei

$$\delta_a[A] := \begin{cases} 1 & \text{falls } a \in A \\ 0 & \text{sonst} \end{cases}$$

Dies ist eine deterministische Verteilung mit:

$$P[\{\omega = a\}] = 1$$

Diskrete Wahrscheinlichkeitsräume

Ist Ω eine abzählbare Menge und $p : \Omega \to [0,1]$ eine Gewichtsfunktion mit $\sum_{\omega\in\Omega} p(\omega) = 1$, dann haben wir bereits gezeigt, dass eine eindeutige Wahrscheinlichkeitsverteilung P auf der Potenzmenge $\mathcal{A} = \mathcal{P}(\Omega)$ existiert mit

$$P[A] = \sum_{a\in A} p(a) = \sum_{a\in\Omega} p(a)\delta_a[A] \quad \forall A \subseteq \Omega.$$

Jede diskrete Wahrscheinlichkeitsverteilung ist eine Konvexkombination von Diracmaßen:

$$P = \sum_{a\in\Omega} p(a)\delta_a$$

Endliche Produktmodelle

Auch die Konstruktion mehrstufiger diskreter Modelle ist auf diese Weise möglich: Ist beispielsweise

$$\Omega = \{(\omega_1, \ldots, \omega_n) : \omega_i \in \Omega_i\} = \Omega_1 \times \ldots \times \Omega_n$$

eine Produktmenge, und sind $p_1, \ldots, p_n$ Gewichtsfunktionen von Wahrscheinlichkeitsverteilungen $P_1, \ldots, P_n$ auf $\Omega_1, \ldots, \Omega_n$, dann ist

$$p(\omega) = \prod_{i=1}^{n} p_i(\omega_i)$$

die Gewichtsfunktion einer Wahrscheinlichkeitsverteilung $P = P_1 \otimes \ldots \otimes P_n$ auf Ω. Unter dieser Wahrscheinlichkeitsverteilung sind die Zufallsvariablen $X_i(\omega) = \omega_i$ unabhängig.

Unendliches Produktmodell (z.B. Münzwurffolge)

Es stellt sich die Frage, ob wir auch unendlich viele unabhängige Zufallsvariablen auf einem ähnlichen Produktraum realisieren können. Im einfachsten Fall möchten wir eine Folge unabhängiger fairer Münzwürfe (0-1-Experimente) auf dem Grundraum

$$\Omega = \{\omega = (\omega_1, \omega_2, \ldots) : \omega_i \in \{0,1\}\} = \{0,1\}^{\mathbb{N}}$$

modellieren. Ω ist überabzählbar, denn die Abbildung $X : (0,1) \to \Omega$, die einer reellen Zahl die Ziffernfolge ihrer Binärdarstellung zuordnet, ist injektiv. Genauer ist eine injektive Abbildung $X : (0,1) \to \Omega$ definiert durch

$$X(\omega) = (X_1(\omega),\ X_2(\omega),\ X_3(\omega),\ \ldots), \tag{4.2.1}$$

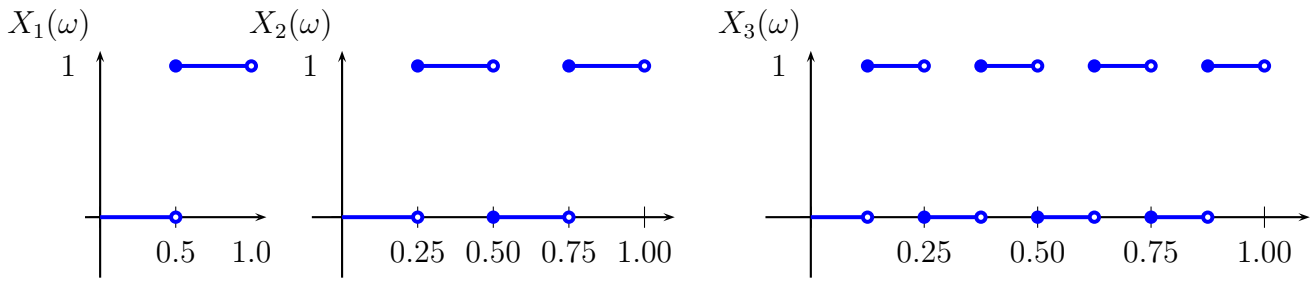


Abbildung 4.4: Darstellung der ersten drei $X_i(\omega)$.

wobei $X_n(\omega) = I_{D_n}(\omega), D_n = \bigcup_{i=1}^{2^{n-1}} [(2i-1)\cdot 2^{-n}, 2i\cdot 2^{-n})$.

Wir suchen eine Wahrscheinlichkeitsverteilung P auf Ω mit

$$P[\{\omega \in \Omega : \omega_1 = a_1, \omega_2 = a_2, \ldots, \omega_n = a_n\}] = 2^{-n} \tag{4.2.2}$$

Gibt es eine σ-Algebra $\mathcal{A}$, die alle diese Ereignisse enthält, und eine eindeutige Wahrscheinlichkeitsverteilung P auf $\mathcal{A}$ mit (4.2.2)?

Wir werden in Abschnitt 5.3 zeigen, dass dies der Fall ist; wobei aber

(1). $\mathcal{A} \neq \mathcal{P}(\Omega)$ und

(2). $P[\{\omega\}] = 0$ für alle $\omega \in \Omega$

gelten muss. Das entsprechende Produktmodell unterscheidet sich in dieser Hinsicht grundlegend von diskreten Modellen.

Kontinuierliche Gleichverteilung

Für die Gleichverteilung auf einem endlichen reellen Intervall $\Omega = [a, b], -\infty < a < b < \infty$, sollte gelten:

$$P[(c, d)] = P[[c, d]] = \frac{d-c}{b-a} \quad \forall a \leq c < d \leq b. \tag{4.2.3}$$

Gibt es eine σ-Algebra $\mathscr{B}$, die alle Teilintervalle von $[a, b]$ enthält, und eine Wahrscheinlichkeitsverteilung P auf $\mathscr{B}$ mit (4.2.3)?

Wieder ist die Antwort positiv, aber erneut gilt notwendigerweise $\mathscr{B} \neq \mathcal{P}(\Omega)$ und $P[\{\omega\}] = 0$ für alle $\omega \in \Omega$.

Tatsächlich sind die Probleme in den letzten beiden Abschnitten weitgehend äquivalent: die durch die Binärdarstellung (4.2.1) definierte Abbildung X ist eine Bijektion von $[0, 1)$ nach $\{0, 1\}^{\mathbb{N}} \setminus A$, wobei $A = \{\omega \in \Omega : \omega_n = 1 \text{ schließlich}\}$ eine abzählbare Teilmenge ist. Eine Gleichverteilung auf $[0, 1)$ wird durch X auf eine Münzwurffolge auf $\{0, 1\}^{\mathbb{N}}$ abgebildet, und umgekehrt.

Brownsche Bewegung

Simuliert man einen Random Walk, so ergibt sich in einem geeigneten Skalierungslimes mit Schrittweite $\to 0$ anscheinend eine irreguläre, aber stetige zufällige Bewegung in kontinuierlicher Zeit. Der entsprechende, 1923 von N. Wiener konstruierte stochastische Prozess heißt **Brown-**

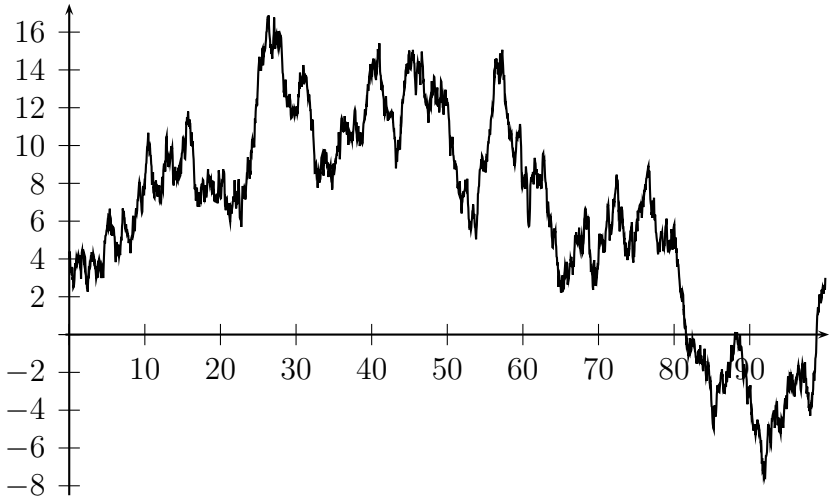


Abbildung 4.5: Graph einer Stichprobe der eindimensionalen Brownschen Bewegung

sche Bewegung, und kann durch eine Wahrscheinlichkeitsverteilung P (das Wienermaß) auf dem Raum

$$\Omega = C([0,1],\mathbb{R}) = \{\omega : [0,1] \to \mathbb{R} | \omega \text{ stetig}\}$$

beschrieben werden. Für diese, als Modell für Aktienkurse, zufällige Bewegungen, etc. in diversen Anwendungsbereichen fundamentale Wahrscheinlichkeitsverteilung gilt unter anderem:

$$P[\{\omega \in \Omega : \omega(t) \in [a,b)\}] = \frac{1}{\sqrt{2\pi t}} \int_a^b e^{-\frac{x^2}{2t}}\, \mathrm{d}x \qquad \text{für alle } t > 0,$$

siehe zum Beispiel die Vorlesung „Stochastische Prozesse“ im Sommersemester.

Um Wahrscheinlichkeitsverteilungen wie in den letzten beiden Beispielen zu konstruieren, benötigen wir zunächst geeignete σ-Algebren, die die relevanten Ereignisse bzw. Intervalle enthalten. Dazu verwenden wir die folgende Konstruktion:

Konstruktion von σ-Algebren

Sei Ω eine beliebige Menge, und $\mathscr{J} \subseteq \mathcal{P}(\Omega)$ eine Kollektion von Ereignissen, die auf jeden Fall in der zu konstruierenden σ-Algebra enthalten sein sollen (z.B. die Mengen aus den Beispielen zu unendlichen Produktmodellen und kontinuierlichen Gleichverteilungen auf Seite 111f).

Definition. *Die Kollektion*

$$\sigma(\mathscr{J}) := \bigcap_{\substack{\mathscr{F} \supseteq \mathscr{J} \\ \mathscr{F} \text{ } \sigma\text{-Algebra auf } \Omega}} \mathscr{F}$$

von Teilmengen von Ω heißt die von $\mathscr{J}$-erzeugte σ-Algebra.

Bemerkung. Wie man leicht nachprüft (Übung), ist $\sigma(\mathscr{J})$ tatsächlich eine σ-Algebra, und damit die kleinste σ-Algebra, die $\mathscr{J}$ enthält.

Beispiel (Borel'sche σ-Algebra auf $\mathbb{R}$). Sei $\Omega = \mathbb{R}$ und $\mathscr{J} = \{(s,t) | -\infty \leq s \leq t \leq \infty\}$ die Kollektion aller offenen Intervalle. Die von $\mathscr{J}$ erzeugte σ-Algebra

$$\mathcal{B}(\mathbb{R}) := \sigma(\mathscr{J})$$

heißt **Borel'sche σ-Algebra**. Man prüft leicht nach, dass $\mathcal{B}(\mathbb{R})$ auch alle abgeschlossenen und halboffenen Intervalle enthält. Die Borel'sche σ-Algebra wird auch erzeugt von der Kollektion aller abgeschlossenen bzw. aller kompakten Intervall. Ebenso gilt:

$$\mathcal{B}(\mathbb{R}) = \sigma(\{(-\infty, c] | c \in \mathbb{R}\})$$

Allgemeiner definieren wir:

Definition. *Sei Ω ein topologischer Raum (also z.B. ein metrischer Raum wie $\mathbb{R}^n, C([0,1],\mathbb{R})$ etc.), und sei τ die Kollektion aller offenen Teilmengen von Ω (die **Topologie**). Die von τ erzeugte σ-Algebra*

$$\mathcal{B}(\Omega) := \sigma(\tau)$$

*heißt **Borel'sche σ-Algebra auf** Ω.*

Wieder verifiziert man, dass $\mathcal{B}(\Omega)$ auch von den abgeschlossenen Teilmengen erzeugt wird. Im Fall $\Omega = \mathbb{R}$ ergibt sich die oben definierte, von den Intervallen erzeugte, σ-Algebra.

Bemerkung. Nicht jede Teilmenge von $\mathbb{R}$ ist in der Borelschen σ-Algebra $\mathcal{B}(\mathbb{R})$ enthalten - ein Beispiel wird in den Übungen gegeben.
Trotzdem enthält $\mathcal{B}(\mathbb{R})$ so gut wie alle Teilmengen von $\mathbb{R}$, die in Anwendungsproblemen auftreten; z.B. alle offenen und abgeschlossenen Teilmengen von $\mathbb{R}$, sowie alle Mengen, die durch Bildung von abzählbar vielen Vereinigungen, Durchschnitten und Komplementbildungen daraus entstehen.

Beispiel (Produkt σ-Algebra auf $\{0,1\}^{\mathbb{N}}$). Eine **Zylindermenge** auf dem Folgenraum

$$\Omega = \{0,1\}^{\mathbb{N}} = \{(\omega_1, \omega_2, \ldots) : \omega_i \in \{0,1\}\}$$

ist eine Teilmenge A von Ω von der Form

$$A = \{\omega \in \Omega : \omega_1 = a_1, \omega_2 = a_2, \ldots, \omega_n = a_n\}, \quad n \in \mathbb{N}, a_1, \ldots, a_n \in \{0,1\}.$$

In Beispiel 4.2 von oben betrachten wir die von der Kollektion $\mathscr{C}$ aller Zylindermengen erzeugte σ-Algebra $\mathcal{A} = \sigma(\mathscr{C})$ auf $\{0,1\}^{\mathbb{N}}$. $\mathcal{A}$ heißt **Produkt-σ-Algebra** auf Ω.

Allgemeiner sei I eine beliebige Menge, und $\Omega = \prod_{i \in I} \Omega_i$ eine Produktmenge (mit endlich, abzählbar, oder sogar überabzählbar vielen Faktoren $\Omega_i, i \in I$).

Definition. *Sind $\mathcal{A}_i, i \in I$ σ-Algebren auf Ω_i, dann heißt die von der Kollektion $\mathscr{C}$ aller Zylindermengen*

$$\{\omega = (\omega_i)_{i \in I} \in \Omega : \omega_{i_1} \in A_{i_1}, \omega_{i_2} \in A_{i_2}, \ldots, \omega_{i_n} \in A_{i_n}\},$$

$n \in \mathbb{N}, i_1, \ldots, i_n \in I, A_{i_1} \in \mathcal{A}_{i_1}, \ldots, A_{i_n} \in \mathcal{A}_{i_n}$, *erzeugte σ-Algebra*

$$\mathcal{A} = \bigotimes_{i \in I} \mathcal{A}_i := \sigma(\mathscr{C})$$

Produkt σ-Algebra auf Ω.

Man kann nachprüfen, dass die etwas anders definierte Produkt-σ-Algebra aus Beispiel 4.2 ein Spezialfall dieser allgemeinen Konstruktion ist.

Existenz und Eindeutigkeit von Wahrscheinlichkeitsverteilungen

Sei $(\Omega, \mathcal{A})$ ein **messbarer Raum**, d.h. Ω ist eine nichtleere Menge und $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ eine σ-Algebra. In der Regel sind die Wahrscheinlichkeiten $P[A]$ zunächst für Ereignisse A aus einer Teilmenge $\mathcal{J} \subseteq \mathcal{A}$ mit $\mathcal{A} = \sigma(\mathcal{J})$ gegeben, z.B. für Intervalle bei Wahrscheinlichkeitsverteilungen auf $\mathbb{R}$. Es stellt sich die Frage, ob hierdurch bereits die Wahrscheinlichkeiten aller Ereignisse in $\mathcal{A}$ eindeutig festgelegt sind, und ob sich P zu einer Wahrscheinlichkeitsverteilung auf $\mathcal{A}$ fortsetzen lässt. Diese Fragen beantworten die folgenden beiden fundamentalen Sätze.

Definition. *(1). Ein Mengensystem $\mathcal{J} \subseteq \mathcal{A}$ heißt **durchschnittsstabil**, falls*

$$A, B \in \mathcal{J} \quad \Rightarrow \quad A \cap B \in \mathcal{J}.$$

*(2). $\mathcal{J}$ heißt **Algebra**, falls*

(a) $\Omega \in \mathcal{J}$

(b) $A \in \mathcal{J} \quad \Rightarrow \quad A^C \in \mathcal{J}$

(c) $A, B \in \mathcal{J} \quad \Rightarrow \quad A \cup B \in \mathcal{J}$.

Eine Algebra ist stabil unter endlichen Mengenoperationen (Bilden von endlichen Vereinigungen, Durchschnitten und Komplementen). Insbesondere ist jede Algebra durchschnittsstabil.

Beispiel. (1). Die Kollektion aller offenen Intervalle ist eine durchschnittsstabile Teilmenge von $\mathcal{B}(\mathbb{R})$, aber keine Algebra. Dasselbe gilt für das Mengensystem $\mathcal{J} = \{(-\infty, c] | c \in \mathbb{R}\}$.

(2). Die Kollektion aller endlichen Vereinigungen von beliebigen Teilintervallen von $\mathbb{R}$ ist eine Algebra.

Satz 4.6 (Eindeutigkeitssatz). *Stimmen zwei Wahrscheinlichkeitsverteilungen P und $\widetilde{P}$ auf $(\Omega, \mathcal{A})$ überein auf einem **durchschnittsstabilen Mengensystem** $\mathcal{J} \subseteq \mathcal{A}$, so auch auf $\sigma(\mathcal{J})$.*

Den Satz werden wir am Ende dieses Abschnittes beweisen.

Beispiel. (1). Eine Wahrscheinlichkeitsverteilung P auf $\mathcal{B}(\mathbb{R})$ ist eindeutig festgelegt durch die Wahrscheinlichkeiten $P[(-\infty, c]],\ c \in \mathbb{R}$.

(2). Die Wahrscheinlichkeitsverteilung P im Modell der unendlich vielen Münzwürfe ist eindeutig festgelegt durch die Wahrscheinlichkeiten der Ausgänge der ersten n Würfe für alle $n \in \mathbb{N}$.

Nach dem Eindeutigkeitssatz 4.6 ist eine Wahrscheinlichkeitsverteilung durch die Wahrscheinlichkeiten der Ereignisse aus einem durchschnittsstabilen Erzeugendensystem festgelegt. Umgekehrt zeigt der folgende Satz, dass sich eine auf einem Erzeugendensystem $\mathcal{J}$ gegebene σ-additive Abbildung zu einem Maß auf der σ-Algebra fortsetzen lässt, falls $\mathcal{J}$ eine Algebra ist.

Satz 4.7 (Fortsetzungssatz von Carathéodory). *Ist $\mathcal{J}$ eine Algebra, und $P : \mathcal{J} \to [0, \infty]$ eine σ-additive Abbildung, dann besitzt P eine Fortsetzung zu einem Maß auf $\sigma(\mathcal{J})$.*

Den Beweis dieses klassischen Resultats findet man in vielen Maßtheorie-, Analysis- bzw. Wahrscheinlichkeitstheorie-Büchern (siehe z. B. Williams: „Probability with martingales“, Appendix A1). Wir verweisen hier auf die Analysisvorlesung, da für die weitere Entwicklung der Wahrscheinlichkeitstheorie in dieser Vorlesung der Existenzsatz zwar fundamental ist, das Beweisverfahren aber keine Rolle mehr spielen wird.

Bemerkung. Ist $P[\Omega] = 1$, bzw. allgemeiner $P[\Omega] < \infty$, dann ist die Maßfortsetzung nach Satz 4.6 eindeutig, denn eine Algebra ist durchschnittsstabil.

Als Konsequenz aus dem Fortsetzungssatz erhält man:

Korollar 4.8 (Existenz und Eindeutigkeit der kontinuierlichen Gleichverteilung). *Es existiert genau eine Wahrscheinlichkeitsverteilung $\mathcal{U}_{(0,1)}$ auf $\mathcal{B}((0,1))$ mit*

$$\mathcal{U}_{(0,1)}[(a,b)] = b - a \qquad \textit{für alle } 0 < a \leq b < 1. \tag{4.2.4}$$

Zum Beweis ist noch zu zeigen, dass die durch (4.2.4) definierte Abbildung $\mathcal{U}_{(0,1)}$ sich zu einer σ-additiven Abbildung auf die von den offenen Intervallen erzeugte Algebra $\mathcal{A}_0$ aller endlichen Vereinigungen von beliebigen (offenen, abgeschlossenen, halboffenen) Teilintervallen von $(0,1)$ fortsetzen lässt. Wie die Fortsetzung auf $\mathcal{A}_0$ aussieht, ist offensichtlich - der Beweis der σ-Additivität ist etwas aufwändiger. Wir verweisen dazu wieder auf die Analysisvorlesung, bzw. den Appendix A1 in Williams: „Probability with martingales."

Bemerkung. (1). Auf ähnliche Weise folgt die Existenz und Eindeutigkeit des durch

$$\lambda[(a_1,b_1) \times \ldots \times (a_d,b_d)] = \prod_{i=1}^{d}(b_i - a_i) \qquad \text{für alle } a_i, b_i \in \mathbb{R} \text{ mit } a_i \leq b_i$$

eindeutig festgelegten Lebesguemaßes λ auf $\mathcal{B}(\mathbb{R}^d)$, siehe Analysis III. Man beachte, dass wegen $\lambda[\mathbb{R}^d] = \infty$ eine Reihe von Aussagen, die wir für Wahrscheinlichkeitsverteilungen beweisen werden, nicht für das Lebesguemaß auf $\mathbb{R}^d$ gelten!

(2). Auch die Existenz der Wahrscheinlichkeitsverteilungen im Modell für unendlich viele faire Münzwürfe kann man mithilfe des Satzes von Carathéodory zeigen. Wir werden diese Wahrscheinlichkeitsverteilung stattdessen unmittelbar aus der Gleichverteilung $\mathcal{U}_{(0,1)}$ konstruieren.

Zum Abschluss dieses Abschnitts beweisen wir nun den Eindeutigkeitssatz. Dazu betrachten wir das Mengensystem

$$\mathcal{D} := \{A \in \mathcal{A} \mid P[A] = \widetilde{P}[A]\} \supseteq \mathcal{J}.$$

Zu zeigen ist: $\mathcal{D} \supseteq \sigma(\mathcal{J})$.

Dazu stellen wir fest, dass $\mathcal{D}$ folgende Eigenschaften hat:

(i) $\Omega \in \mathcal{D}$

(ii) $A \in \mathcal{D} \Rightarrow A^c \in \mathcal{D}$

(iii) $A_1, A_2, \ldots \in \mathcal{D}$ paarweise disjunkt $\Rightarrow \bigcup A_i \in \mathcal{D}$

Definition. *Ein Mengensystem $\mathcal{D} \subseteq \mathcal{P}(\Omega)$ mit (i) - (iii) heißt **Dynkinsystem**.*

Bemerkung. Für ein Dynkinsystem $\mathcal{D}$ gilt:

$$A, B \in \mathcal{D}\ ,\ A \subseteq B \ \Rightarrow\ B \backslash A = B \cap A^C = (\underbrace{B^C \cup A}_{disjunkt})^C \ \in \mathcal{D}$$

Lemma 4.9. *Jedes $\cap$ - stabile Dynkinsystem $\mathcal{D}$ ist eine σ - Algebra.*

Beweis. Für $A, B \in \mathcal{D}$ gilt:

$$A \cup B \ = \ A \underset{\substack{\uparrow \\ disjunkt}}{\cup} \underbrace{(\overbrace{B \backslash (A \cap B)}^{\in \mathcal{D}\ falls\ \cap - stabil})}_{\in \mathcal{D}\ nach\ Bem.} \ \in \mathcal{D}.$$

Hieraus folgt für $A_1, A_2, \ldots \in \mathcal{D}$ durch Induktion

$$B_n := \bigcup_{i=1}^{n} A_i \ \in \mathcal{D},$$

und damit

$$\bigcup_{i=1}^{\infty} A_i = \bigcup_{n=1}^{\infty} B_n = \underset{\substack{\uparrow \\ disjunkt}}{\bigcup_{n=1}^{\infty}} \underbrace{(B_n \backslash B_{n-1})}_{\in \mathcal{D}\ nach\ Bem.} \ \in \mathcal{D}.$$

□

Lemma 4.10. *Ist $\mathcal{J}$ ein $\cap$ - stabiles Mengensystem , so stimmt **das von $\mathcal{J}$ erzeugte Dynkinsystem***

$$\mathcal{D}(\mathcal{J}) \quad := \bigcap_{\substack{\mathcal{D}\ Dynkinsystem \\ \mathcal{D} \supseteq \mathcal{J}}} \mathcal{D}$$

mit der von $\mathcal{J}$ erzeugten σ - Algebra $\sigma(\mathcal{J})$ überein.

Aus Lemma (4.10) folgt der Eindeutigkeitssatz, denn $\{A \in \mathcal{A} \,|P[A] = \widetilde{P}[A]\}$ ist ein Dynkinsystem, das $\mathcal{J}$ enthält, und somit gilt nach dem Lemma

$$\{A \in \mathcal{A} \,|P[A] = \widetilde{P}[A]\} \supseteq \mathcal{D}(\mathcal{J}) = \sigma(\mathcal{J}),$$

falls $\mathcal{J}$ durchschnittsstabil ist.

Beweis. (von Lemma (4.10))
Jede σ - Algebra ist ein Dynkinsystem, also gilt $\mathcal{D}(\mathcal{J}) \subseteq \sigma(\mathcal{J})$.
Es bleibt zu zeigen, dass $\mathcal{D}(\mathcal{J})$ eine σ - Algebra ist (hieraus folgt dann $\mathcal{D}(\mathcal{J}) = \sigma(\mathcal{J})$). Nach dem ersten Lemma ist dies der Fall, wenn $\mathcal{D}(\mathcal{J})$ durchschnittsstabil ist. Dies zeigen wir nun in zwei Schritten:

Schritt 1: $B \in \mathcal{J}, A \in \mathcal{D}(\mathcal{J}) \Rightarrow A \cap B \in \mathcal{D}(\mathcal{J})$

Beweis: $\mathcal{D}_B :=$ { A $\in \mathcal{A} \mid A \cap B \in \mathcal{D}(\mathcal{J})\} \supseteq \mathcal{J}$ ist ein Dynkinsystem. Z.B. gilt

$$\begin{aligned} A \in \mathcal{D}_B &\Rightarrow A \cap B \in \mathcal{D}(\mathcal{J}) \\ &\Rightarrow A^C \cap B = \underset{\substack{\uparrow \\ \in \mathcal{D}(\mathcal{J})}}{B} \setminus (\underbrace{A \cap B}_{\in \mathcal{D}(\mathcal{J})}) \overset{Bem.}{\in} \mathcal{D}(\mathcal{J}) \\ &\Rightarrow A^C \in \mathcal{D}_B \text{ usw.} \end{aligned}$$

Also gilt $\mathcal{D}_B \supseteq \mathcal{D}(\mathcal{J})$, und damit $A \cap B \in \mathcal{D}(\mathcal{J})$ für alle $A \in \mathcal{D}(\mathcal{J})$.

Schritt 2: $A, B \in \mathcal{D}(\mathcal{J}) \Rightarrow A \cap B \in \mathcal{D}(\mathcal{J})$

Beweis: $\mathcal{D}_A :=$ { B $\in \mathcal{A} \mid A \cap B \in \mathcal{D}(\mathcal{J})\} \supseteq \mathcal{J}$ nach Schritt 1. Zudem ist $\mathcal{D}_A$ ein Dynkinsystem (Beweis analog zu Schritt 1), also gilt $\mathcal{D}_A \supseteq \mathcal{D}(\mathcal{J})$. □

4.3 Allgemeine Zufallsvariablen und ihre Verteilung

Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum. Wir wollen nun Zufallsvariablen $X : \Omega \to S$ mit Werten in einem allgemeinen messbaren Raum $(S, \mathcal{S})$ betrachten. Beispielsweise ist $S = \mathbb{R}$ oder $S = \mathbb{R}^d$ und $\mathcal{S}$ ist die Borelsche σ-Algebra. Oft interessieren uns die Wahrscheinlichkeiten von Ereignissen der Form

$$\{X \in B\} = \{\omega \in \Omega | X(\omega) \in B\} = X^{-1}(B),$$

„Der Wert der Zufallsgröße X liegt in B“

wobei $B \subseteq S$ eine Menge aus der σ-Algebra $\mathcal{S}$ auf dem Bildraum ist, also z.B. ein Intervall oder eine allgemeinere Borelmenge, falls $S = \mathbb{R}$ gilt.
Wir erweitern dementsprechend die zuvor eingeführten Konzepte einer Zufallsvariablen und ihrer Verteilung.

Allgemeine Zufallsvariablen

Definition. *Eine Abbildung $X : \Omega \to S$ heißt **messbar bzgl.** $\mathcal{A}/\mathcal{S}$, falls*

$$(M) \qquad X^{-1}(B) \in \mathcal{A} \qquad \textit{für alle } B \in \mathcal{S}.$$

*Eine **Zufallsvariable** ist eine auf einem Wahrscheinlichkeitsraum definierte messbare Abbildung.*

Bemerkung. (1). Ist Ω abzählbar und $\mathcal{A} = \mathcal{P}(\Omega)$, dann ist jede Abbildung $X : \Omega \to S$ eine Zufallsvariable.

(2). Ist S abzählbar und $\mathcal{S} = \mathcal{P}(S)$, dann ist X genau dann eine Zufallsvariable, falls

$$\{X = a\} = X^{-1}(\{a\}) \in \mathcal{A} \qquad \text{für alle } a \in S$$

gilt. Dies ist gerade die Definition einer diskreten Zufallsvariable von oben.

Stimmt die σ-Algebra $\mathcal{S}$ nicht mit der Potenzmenge $\mathcal{P}(S)$ überein, dann ist es meist schwierig, eine Bedingung (M) für **alle** Mengen $B \in \mathcal{S}$ explizit zu zeigen. Die folgenden Aussagen liefern handhabbare Kriterien, mit denen man in fast allen praktisch relevanten Fällen sehr leicht zeigen kann, dass die zugrunde liegenden Abbildungen messbar sind. Wir bemerken zunächst, dass es genügt die Bedingung (M) für alle Mengen aus einem Erzeugendensystem $\mathcal{J}$ der σ-Algebra $\mathcal{S}$ zu überprüfen:

Lemma 4.11. *Sei $\mathcal{J} \subseteq \mathcal{P}(S)$ mit $\mathcal{S} = \sigma(\mathcal{J})$. Dann gilt (M) bereits, falls*

$$X^{-1}(B) \in \mathcal{A} \qquad \textit{für alle } B \in \mathcal{J}.$$

Beweis. Das Mengensystem $\{B \in \mathcal{S} | X^{-1}(B) \in \mathcal{A}\}$ ist eine σ-Algebra, wie man leicht nachprüft. Diese enthält $\mathcal{J}$ nach Voraussetzung, also enthält sie auch die von $\mathcal{J}$ erzeugte σ-Algebra $\mathcal{S}$. □

Korollar (Reellwertige Zufallsvariablen). *Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum. Eine Abbildung $X : \Omega \to \mathbb{R}$ ist genau dann eine Zufallsvariable bzgl. der Borelschen σ-Algebra, wenn*

$$\{X \leq c\} = \{\omega \in \Omega \mid X(\omega) \leq c\} \in \mathcal{A} \quad \forall\, c \in \mathbb{R}, \qquad \textit{bzw. wenn}$$
$$\{X < c\} = \{\omega \in \Omega \mid X(\omega) < c\} \in \mathcal{A} \quad \forall\, c \in \mathbb{R}.$$

Beweis. Es gilt $\{X \leq c\} = X^{-1}((-\infty, c])$. Die Intervalle $(-\infty, c], c \in \mathbb{R}$, erzeugen $\mathcal{B}(\mathbb{R})$, also folgt die erste Aussage. Die zweite Aussage zeigt man analog. □

Beispiel (**Indikatorfunktionen**). Für eine Menge $A \subseteq \Omega$ gilt:

$$I_A \text{ ist Zufallsvariable } \Leftrightarrow \ A \in \mathcal{A},$$

denn

$$\{I_A \leq c\} = \begin{cases} \emptyset & \text{falls } c < 0 \\ \Omega & \text{falls } c \geq 1 \\ A^C & \text{falls } 0 \leq c < 1 \end{cases},$$

und A^C ist genau dann in $\mathcal{A}$ enthalten, wenn A in $\mathcal{A}$ enthalten ist.

Korollar (**Stetige Abbildungen sind messbar**). *Seien Ω und S topologische Räume, und $\mathcal{A}, \mathcal{S}$ die Borelschen σ-Algebren. Dann gilt:*

$$X : \Omega \to S \textit{ stetig } \Rightarrow \ X \textit{ messbar.}$$

Beweis. Sei $\mathcal{J}$ die Topologie von S, d.h. die Kollektion aller offenen Teilmengen von S. Nach Definition der Borelschen σ-Algebra gilt $\mathcal{S} = \sigma(\mathcal{J})$. Wegen

$$B \in \mathcal{J} \quad \Rightarrow B \text{ offen } \overset{X \text{ stetig}}{\Longrightarrow} \ X^{-1}(B) \text{ offen } \Longrightarrow \ X^{-1}(B) \in \mathcal{A}$$

folgt die Behauptung. □

Kompositionen von messbaren Abbildungen sind wieder messbar:

Lemma 4.12. *Sind $(\Omega_1, \mathcal{A}_1), (\Omega_2, \mathcal{A}_2)$ und $(\Omega_3, \mathcal{A}_3)$ messbare Räume, und ist $X_1 : \Omega_1 \to \Omega_2$ messbar bzgl. $\mathcal{A}_1/\mathcal{A}_2$ und $X_2 : \Omega_2 \to \Omega_3$ messbar bzgl. $\mathcal{A}_2/\mathcal{A}_3$, dann ist $X_2 \circ X_1$ messbar bzgl. $\mathcal{A}_1/\mathcal{A}_3$.*

$$\begin{array}{ccccc} \Omega_1 & \xrightarrow{X_1} & \Omega_2 & \xrightarrow{X_2} & \Omega_3 \\ \mathcal{A}_1 & & \mathcal{A}_2 & & \mathcal{A}_3 \end{array}$$

Beweis. Für $B \in \mathcal{A}_3$ gilt $(X_2 \circ X_1)^{-1}(B) = X_1^{-1}(\underbrace{X_2^{-1}(B)}_{\in \mathcal{A}_2}) \in \mathcal{A}_1$. □

Beispiel. (1). Ist $X : \Omega \to \mathbb{R}$ eine reellwertige Zufallsvariable und $f : \mathbb{R} \to \mathbb{R}$ eine messbare (z.B. stetige) Funktion, dann ist auch

$$f(X) := f \circ X : \Omega \to \mathbb{R}$$

wieder eine reellwertige Zufallsvariable. Beispielsweise sind $|X|, |X|^p, e^X$ usw. Zufallsvariablen.

(2). Sind $X, Y : \Omega \to \mathbb{R}$ reellwertige Zufallsvariablen, dann ist $(X, Y) : \omega \mapsto (X(\omega), Y(\omega))$ eine messbare Abbildung in den $\mathbb{R}^2$ mit Borelscher σ-Algebra.
Da die Abbildung $(x, y) \mapsto x + y$ stetig ist, ist $X + Y$ wieder eine reellwertige Zufallsvariable. Dies sieht man auch direkt wie folgt: Für $c \in \mathbb{R}$ gilt:

$$X + Y < c \iff \exists\, r, s \in \mathbb{Q} : r + s < c, X < r \text{ und } Y < s,$$

also

$$\{X + Y < c\} = \bigcup_{\substack{r,s \in \mathbb{Q} \\ r+s<c}} (\{X < r\} \cap \{Y < s\}) \in \mathcal{A}$$

Verteilungen von Zufallsvariablen

Um Zufallsexperimente zu analysieren, müssen wir wissen, mit welchen Wahrscheinlichkeiten die relevanten Zufallsvariablen Werte in bestimmten Bereichen annehmen. Dies wird durch die Verteilung beschrieben. Seien $(\Omega, \mathcal{A})$ und $(S, \mathcal{S})$ messbare Räume.

Satz 4.13 (Bild einer Wahrscheinlichkeitsverteilung unter einer ZV). *Ist P eine Wahrscheinlichkeitsverteilung auf $(\Omega, \mathcal{A})$, und $X : \Omega \to S$ messbar bzgl. $\mathcal{A}/\mathcal{S}$, dann ist durch*

$$\mu_X(B) := P[X \in B] = P[X^{-1}(B)] \qquad (B \in \mathcal{S})$$

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

Beweis. (1). $\mu_X(S) = P[X^{-1}(S)] = P[\Omega] = 1$

(2). Sind $B_n \in \mathcal{S}$, $n \in \mathbb{N}$, paarweise disjunkte Mengen, dann sind auch die Urbilder $X^{-1}(B_n)$, $n \in \mathbb{N}$, paarweise disjunkt. Also gilt wegen der σ-Additivität von P:

$$\mu_X\left[\bigcup_n B_n\right] = P\left[X^{-1}\left(\bigcup_n B_n\right)\right] = P\left[\bigcup_n X^{-1}(B_n)\right] = \sum_n P[X^{-1}(B_n)] = \sum_n \mu_X[B_n].$$

□

Definition. *Die Wahrscheinlichkeitsverteilung μ_X auf $(S, \mathcal{S})$ heißt **Bild von** P **unter** X oder **Verteilung (law) von** X **unter** P.*

Für μ_X werden häufig auch die folgenden Notationen verwendet:

$$\mu_X = P \circ X^{-1} = \mathcal{L}_X = P_X = X(P)$$

Charakterisierung der Verteilung

- Diskrete Zufallsvariablen:
  Die Verteilung μ_X einer diskreten Zufallsvariablen ist eindeutig durch die **Massenfunktion**

$$p_X(a) = P[X = a] = \mu_X[\{a\}], \qquad a \in S,$$

  festgelegt.

- Reelle Zufallsvariablen
  Die Verteilung μ_X einer reellwertigen Zufallsvariablen $X : \Omega \to \mathbb{R}$ ist eine Wahrscheinlichkeitsverteilung auf $\mathcal{B}(R)$. Sie ist eindeutig festgelegt durch die Wahrscheinlichkeiten

$$\mu_X[(-\infty, c]] = P[X \leq c], \qquad c \in \mathbb{R},$$

  da die Intervalle $(-\infty, c], c \in \mathbb{R}$, ein durchschnittsstabiles Erzeugendensystem der Borelschen σ-Algebra bilden.

Definition. *Die Funktion* $F_X : \mathbb{R} \to [0, 1]$*,*

$$F_X(c) := P[X \leq c] = \mu_X[(-\infty, c]]$$

heißt ***Verteilungsfunktion (distribution function)*** *der Zufallsvariable* $X : \Omega \to \mathbb{R}$ *bzw. der Wahrscheinlichkeitsverteilung* μ_X *auf* $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$*.*

Beispiel (Kontinuierliche Gleichverteilung)**.** Seien $a, b \in \mathbb{R}$ mit $a < b$. Eine Zufallsvariable $X : \Omega \to \mathbb{R}$ ist gleichverteilt auf dem Intervall (a, b), falls

$$F_X(c) = P[X \leq c] = \mathcal{U}_{(a,b)}[(a, c)] = \frac{c - a}{b - a} \qquad \text{für alle } c \in (a, b)$$

gilt. Eine auf $(0, 1)$ gleichverteilte Zufallsvariable ist zum Beispiel die Identität

$$U(\omega) = \omega$$

auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P) = ((0, 1), \mathcal{B}((0, 1)), \mathcal{U}_{(0,1)})$. Ist U gleichverteilt auf $(0, 1)$, dann ist die Zufallsvariable

$$X(\omega) = a + (b - a)U(\omega)$$

gleichverteilt auf (a, b).

Beispiel (Exponentialverteilung)**.** Angenommen, wir wollen die Wartezeit auf das erste Eintreten eines unvorhersehbaren Ereignisses (radioaktiver Zerfall, Erdbeben, ...) mithilfe einer Zufallsvariable $T : \Omega \to (0, \infty)$ beschreiben. Wir überlegen uns zunächst, welche Verteilung zur Modellierung einer solchen Situation angemessen sein könnte. Um die Wahrscheinlichkeit $P[T > t]$ zu approximieren, unterteilen wir das Intervall $(0, t]$ in eine große Anzahl $n \in \mathbb{N}$ von gleich großen Intervallen $(\frac{(k-1)t}{n}, \frac{kt}{n}], 1 \leq k \leq n$.

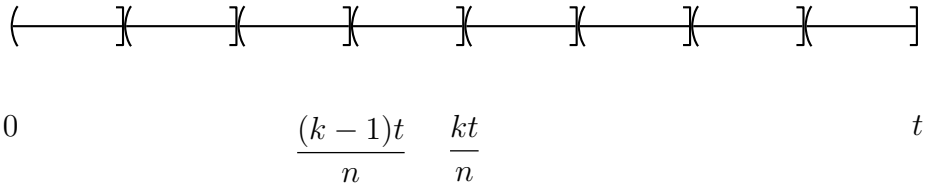


Abbildung 4.6: Unterteilung des Intervalls $(0, t]$ in n Teile.

Sei A_k das Ereignis, dass das unvorhersehbare Geschehen im Zeitraum $(\frac{(k-1)t}{n}, \frac{kt}{n}]$ eintritt. Ein nahe liegender Modellierungsansatz ist anzunehmen, dass die Ereignisse A_k unabhängig sind mit Wahrscheinlichkeit

$$P[A_k] \approx \lambda \frac{t}{n},$$

wobei $\lambda > 0$ die „Intensität“, d.h. die mittlere Häufigkeit des Geschehens pro Zeiteinheit, beschreibt, und die Approximation für $n \to \infty$ immer genauer wird. Damit erhalten wir:

$$P[T > t] = P[A_1^C \cap \ldots \cap A_n^C] \approx \left(1 - \frac{\lambda t}{n}\right)^n \quad \text{für großes } n.$$

Für $n \to \infty$ konvergiert die rechte Seite gegen $e^{-\lambda t}$.

Daher liegt folgende Definition nahe:

Definition. *Eine Zufallsvariable $T : \Omega \to [0, \infty)$ heißt **exponentialverteilt zum Parameter** λ>0, falls*

$$P[T > t] = e^{-\lambda t} \qquad \textit{für alle } t \geq 0 \textit{ gilt.}$$

*Die **Exponentialverteilung zum Parameter** λ ist dementsprechend die Wahrscheinlichkeitsverteilung $\mu = \mathrm{Exp}(\lambda)$ auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ mit*

$$\mu[(t, \infty)] = e^{-\lambda t} \qquad \textit{für alle } t \geq 0,$$

bzw. mit Verteilungsfunktion

$$F(t) = \mu[(-\infty, t]] = \begin{cases} 1 - e^{-\lambda t} & \text{für } t \geq 0 \\ 0 & \text{für } t < 0. \end{cases} \tag{4.3.1}$$

Nach dem Eindeutigkeitssatz ist die $\mathrm{Exp}(\lambda)$-Verteilung durch (4.3.1) eindeutig festgelegt.

Wir konstruieren nun explizit eine exponentialverteilte Zufallsvariable. Dazu bemerken wir, dass $T : \Omega \to \mathbb{R}$ genau dann exponentialverteilt mit Parameter λ ist, wenn

$$P[e^{-\lambda T} < u] = P\left[T > -\frac{1}{\lambda} \log u\right] = e^{\frac{\lambda}{\lambda} \log u} = u$$

für alle $u \in (0, 1)$ gilt, d.h. wenn $e^{-\lambda T}$ auf $(0, 1)$ gleichverteilt ist. Also können wir eine exponentialverteilte Zufallsvariable konstruieren, indem wir umgekehrt

$$T := -\frac{1}{\lambda} \log U \qquad U \sim \mathcal{U}_{(0,1)}$$

setzen. Insbesondere ergibt sich die folgende Methode zur Simulation einer exponentialverteilten Zufallsvariable:

Algorithmus 4.14 (Simulation einer exponentialverteilten Stichprobe).

Input: Intensität $\lambda > 0$

Output: Stichprobe x von $\mathrm{Exp}(\lambda)$

(1). Simuliere $u \sim \mathcal{U}_{(0,1)}$

(2). Setze $x := -\frac{1}{\lambda} \log u$

Wir werden in Abschnitt 4.5 zeigen, dass mit einem entsprechenden Verfahren beliebige reelle Zufallsvariablen konstruiert und simuliert werden können. Zum Abschluss dieses Abschnitts zeigen wir noch eine bemerkenswerte Eigenschaft exponentialverteilter Zufallsvariablen:

Satz 4.15 (Gedächtnislosigkeit der Exponentialverteilung). *Ist T exponentialverteilt, dann gilt für alle $s, t \geq 0$:*

$$P[T - s > t | T > s] = P[T > t].$$

Hierbei ist $T - s$ die verbleibende Wartezeit auf das erste Eintreten des Ereignisses. Also:

Auch wenn man schon sehr lange vergeblich gewartet hat,
liegt das nächste Ereignis nicht näher als am Anfang!

Beweis.

$$P[T-s>t|T>s] = \frac{P[T-s>t \text{ und } T>s]}{P[T>s]} = \frac{P[T>s+t]}{T>s} = \frac{e^{-\lambda(t+s)}}{e^{-\lambda s}} = e^{-\lambda t} = P[T>t].$$

□

4.4 Wahrscheinlichkeitsverteilungen auf $\mathbb{R}$

In diesem und im nächsten Abschnitt beschäftigen wir uns systematischer mit der Beschreibung, Konstruktion und Simulation reellwertiger Zufallsvariablen. Wir notieren dazu zunächst einige grundlegende Eigenschaften der Verteilungsfunktion

$$F(c) \quad = \quad P[X \leq c]$$

einer auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ definierten Zufallsvariable $X : \Omega \to \mathbb{R}$. Wir werden im nächsten Abschnitt sehen, dass umgekehrt jede Funktion mit den Eigenschaften (1)-(3) aus Satz 4.16 die Verteilungsfunktion einer reellen Zufallsvariable ist.

Eigenschaften der Verteilungsfunktion

Satz 4.16. *Für die Verteilungsfunktion $F : \mathbb{R} \to [0,1]$ einer reellwertigen Zufallsvariable X gilt:*

*(1). F ist **monoton wachsend**,*

(2). $\lim_{c \to -\infty} F(c) = 0$ und $\lim_{c \to \infty} F(c) = 1$,

*(3). F ist **rechtsstetig**, d.h. $F(c) = \lim_{y \searrow c} F(y)$ für alle $c \in \mathbb{R}$,*

(4). $F(c) = \lim_{y \nearrow c} F(y) + \mu_X[\{c\}]$.
Insbesondere ist F stetig bei c, falls $\mu_X[\{c\}] = 0$ gilt.

Beweis. Die Aussagen folgen unmittelbar aus der monotonen Stetigkeit und Normiertheit der zugrundeliegenden Wahrscheinlichkeitsverteilung P. Der Beweis der Eigenschaften (1)-(3) wird dem Leser als Übung überlassen. Zum Beweis von (4) bemerken wir, dass für $y < c$ gilt:

$$F(c) - F(y) = P[X \leq c] - P[X \leq y] = P[y < X \leq c].$$

Für eine monoton wachsende Folge $y_n \nearrow c$ erhalten wir daher aufgrund der monotonen Stetigkeit von P:

$$\begin{aligned} F(c) - \lim_{n \to \infty} F(y_n) &= \lim_{n \to \infty} P[y_n < X \leq c] = P\left[\bigcap_n \{y_n < X \leq c\}\right] \\ &= P[X = c] = \mu_X[\{c\}]. \end{aligned}$$

Da dies für alle Folgen $y_n \nearrow c$ gilt, folgt die Behauptung. □

Im Folgenden betrachten wir einige Beispiele von eindimensionalen Verteilungen und ihren Verteilungsfunktionen.

Diskrete Verteilungen

Die Verteilung μ einer reellen Zufallsvariable X heißt diskret, wenn $\mu[S] = 1$ für eine abzählbare Menge S gilt.

Beispiele. (1). BERNOULLI-VERTEILUNG MIT PARAMETER $p \in [0,1]$:

$$\mu[\{1\}] = p, \qquad \mu[\{0\}] = 1 - p.$$

Als Verteilungsfunktion ergibt sich

$$F(c) = \begin{cases} 0 & \text{für } c < 0 \\ 1-p & \text{für } c \in [0,1) \\ 1 & \text{für } c \geq 1. \end{cases}$$

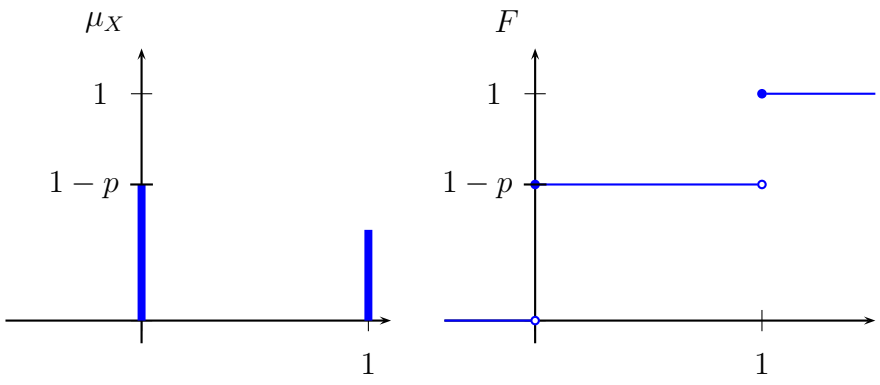


Abbildung 4.7: Massen- und Verteilungsfunktion einer $Ber(p)$-verteilten Zufallsvariablen.

(2). GEOMETRISCHE VERTEILUNG MIT PARAMETER $p \in [0,1]$:

$$\mu[\{k\}] = (1-p)^{k-1} \cdot p \quad \text{für } k \in \mathbb{N}.$$

Für eine geometrisch verteilte Zufallsvariable T gilt:

$$F(c) = P[T \leq c] = 1 - \underbrace{P[T > c]}_{=P[T > \lfloor c \rfloor]} = 1 - (1-p)^{\lfloor c \rfloor} \quad \text{für } c \geq 0,$$

wobei $\lfloor c \rfloor := \max\{n \in \mathbb{Z} \mid n \leq c\}$ der ganzzahlige Anteil von c ist.

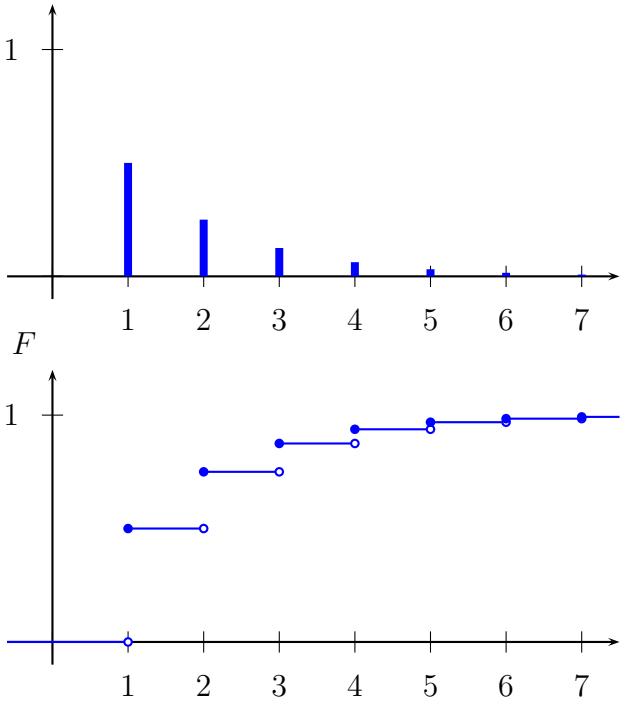


Abbildung 4.8: Massen- und Verteilungsfunktion einer $Geom(\frac{1}{2})$-verteilten Zufallsvariablen.

(3). BINOMIALVERTEILUNG MIT PARAMETERN n UND p:

$$\mu[\{k\}] = \binom{n}{k} p^k (1-p)^{n-k} \qquad \text{für} \quad k = 0, 1, \ldots, n$$

Somit ist die Verteilungsfunktion von $\mathrm{Bin}(n,p)$:

$$F(c) = \sum_{k=0}^{\lfloor c \rfloor} \binom{n}{k} p^k (1-p)^{n-k}$$

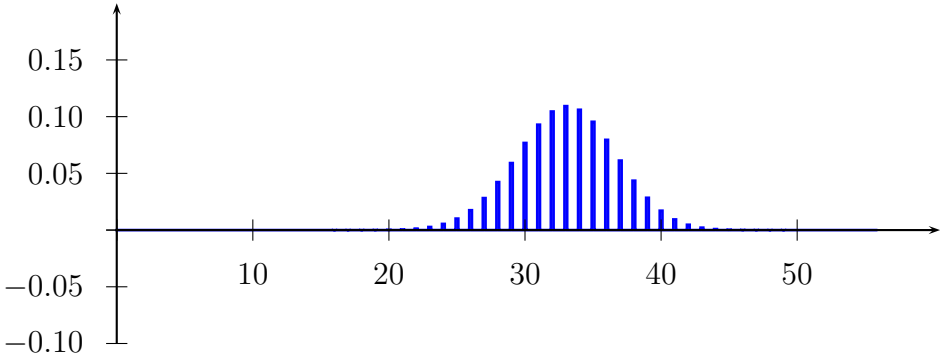


Abbildung 4.9: Massenfunktion einer $\mathrm{Bin}(55, 0.6)$-verteilten Zufallsvariable.

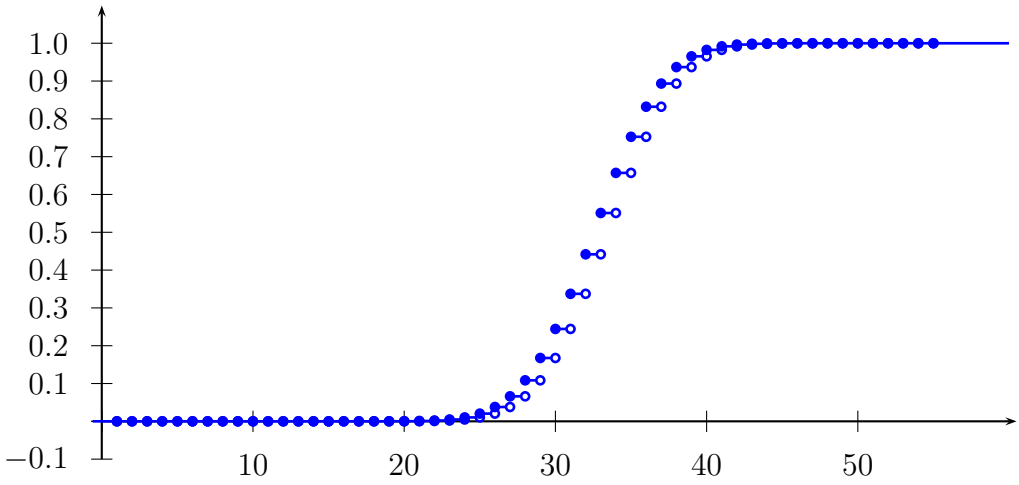


Abbildung 4.10: Verteilungsfunktion von $\mathrm{Bin}(55, 0.6)$

Allgemein sind die Unstetigkeitsstellen der Verteilungsfunktion F einer reellwertigen Zufallsvariable X nach Satz 4.16 (4) gerade die *Atome* der Verteilung, d.h. die $c \in \mathbb{R}$ mit $\mu_X[\{c\}] > 0$. Nimmt X nur endlich viele Werte in einem Intervall I an, dann ist F auf I stückweise konstant, und springt nur bei diesen Werten.

Stetige Verteilungen

Die Verteilung μ einer reellen Zufallsvariable X heißt **stetig**, bzw. **absolutstetig**, falls eine integrierbare Funktion $f : \mathbb{R} \to [0, \infty)$ existiert mit

$$F(c) = P[X \leq c] = \mu[(-\infty, c]] = \int_{-\infty}^{c} f(x)\,\mathrm{d}x \qquad \text{für alle } c \in \mathbb{R}. \tag{4.4.1}$$

Das Integral ist dabei im Allgemeinen als Lebesgueintegral zu interpretieren. Ist die Funktion f stetig, dann stimmt dieses mit dem Riemannintegral überein. Da μ eine Wahrscheinlichkeitsverteilung ist, folgt, dass f eine **Wahrscheinlichkeitsdichte** ist, d.h. $f \geq 0$ und

$$\int_{\mathbb{R}} f(x) = 1.$$

Definition. *Eine Lebesgue-integrierbare Funktion $f : \mathbb{R} \to [0, \infty)$ mit (4.4.1) heißt **Dichtefunktion** der Zufallsvariable X bzw. der Verteilung μ.*

Bemerkung. (1). Nach dem Hauptsatz der Differential- und Integralrechnung gilt

$$F'(x) \;=\; f(x) \tag{4.4.2}$$

für alle $x \in \mathbb{R}$, falls f stetig ist. Im Allgemeinen gilt (4.4.2) für λ-fast alle x, wobei λ das Lebesguemaß auf $\mathbb{R}$ ist.

(2). Aus (4.4.1) folgt aufgrund der Eigenschaften des Lebesgueintegrals (s. Kapitel 6 unten):

$$P[X \in B] = \mu_X[B] = \int_B f(x)\,dx, \tag{4.4.3}$$

für alle Mengen $B \in \mathcal{B}(\mathbb{R})$. Zum Beweis zeigt man, dass beide Seiten von (4.4.3) Wahrscheinlichkeitsverteilungen definieren, und wendet den Eindeutigkeitssatz an.

Beispiele. (1). GLEICHVERTEILUNG AUF (a, b) $(-\infty < a < b < \infty)$.

$$f(x) = \frac{1}{b-a}\, I_{(a,b)}(x), \qquad F(c) = \begin{cases} 0 & \text{für } c \le a \\ \frac{c-a}{b-a} & \text{für } a \le c \le b \\ 1 & \text{für } c \ge b \end{cases}.$$

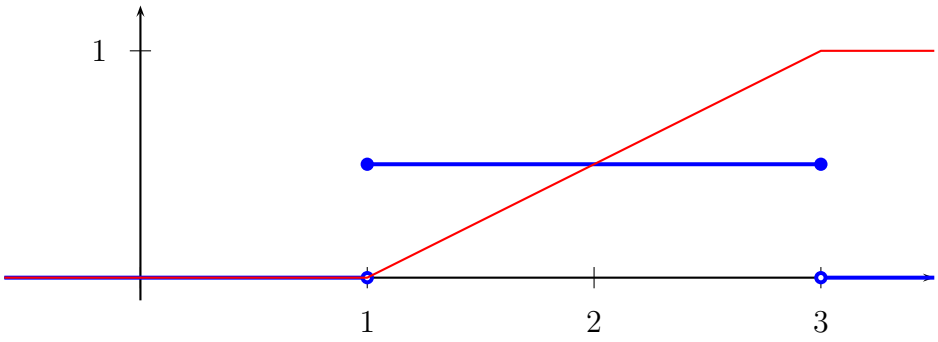


Abbildung 4.11: Dichte $f(x) = \mathbb{1}_{[1,3]}(x)$ einer uniform auf $[1, 3]$ verteilten Zufallsvariable (blau), und deren Verteilungsfunktion $F(c)$ (rot)

Affine Funktionen von gleichverteilten Zufallsvariablen sind wieder gleichverteilt.

(2). EXPONENTIALVERTEILUNG MIT PARAMETER $\lambda > 0$.

$$f(x) = \lambda e^{-\lambda x} I_{(0,\infty)}(x),$$

$$F(c) = \mu[(-\infty, c]] = (1 - e^{-\lambda c})^+ = \int_c^\infty f(x)dx.$$

Ist T eine exponentialverteilte Zufallsvariable zum Parameter λ, und $a > 0$, dann ist aT exponentialverteilt zum Parameter $\frac{\lambda}{a}$, denn

$$P[aT > c] = P[T > \frac{c}{a}] = e^{-\frac{\lambda}{a}c} \qquad \text{für alle } c \ge 0.$$

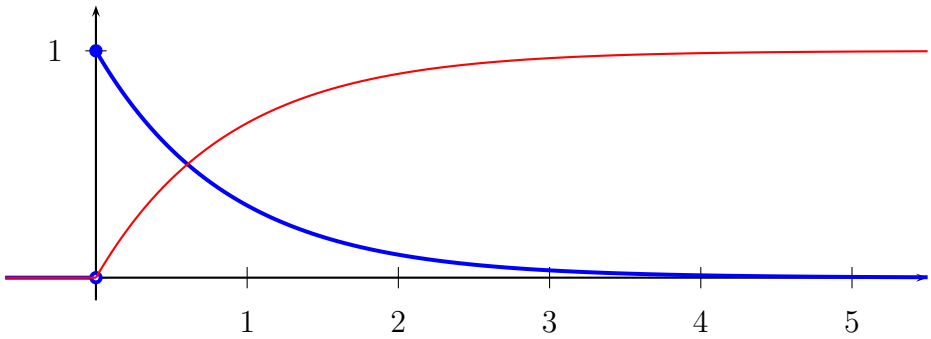


Abbildung 4.12: Dichte $f(x) = \mathbb{1}_{[0,\infty)}(x) \cdot e^{-x}$ einer zum Parameter 1 exponentialverteilten Zufallsvariable (blau) und deren Verteilungsfunktion $F(c)$ (rot)

(3). NORMALVERTEILUNGEN

Wegen $\int_{-\infty}^{\infty} e^{-z^2/2} dz = \sqrt{2\pi}$ ist die „Gaußsche Glockenkurve“

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \qquad z \in \mathbb{R},$$

eine Wahrscheinlichkeitsdichte. Eine stetige Zufallsvariable Z mit Dichtefunktion f heißt standardnormalverteilt. Die Verteilungsfunktion

$$\Phi(c) = \int_{-\infty}^{c} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

der Standardnormalverteilung ist i.A. nicht explizit berechenbar. Ist Z standardnormalverteilt, und

$$X(\omega) = \sigma Z(\omega) + m$$

mit $\sigma > 0, m \in \mathbb{R}$, dann ist X eine Zufallsvariable mit Verteilungsfunktion

$$F_X(c) = P[X \leq c] = P\left[Z \leq \frac{c-m}{\sigma}\right] = \Phi\left(\frac{c-m}{\sigma}\right).$$

Mithilfe der Substitution $z = \frac{x-m}{\sigma}$ erhalten wir:

$$F_X(c) = \int_{-\infty}^{\frac{c-m}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = \int_{-\infty}^{c} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2} dx$$

Definition. *Die Wahrscheinlichkeitsverteilung $N(m, \sigma^2)$ auf $\mathbb{R}$ mit Dichtefunktion*

$$f_{m,\sigma}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}$$

heißt ***Normalverteilung mit Mittel*** m ***und Varianz*** σ^2***.*** *Die Verteilung $N(0,1)$ heißt* ***Standardnormalverteilung****.*

Wir werden im nächsten Abschnitt sehen, dass die Binomialverteilung (also die Verteilung der Anzahl der Erfolge bei unabhängigen 0-1-Experimenten mit Erfolgswahrscheinlichkeit p) für große n näherungsweise durch eine Normalverteilung beschrieben werden kann. Entsprechendes gilt viel allgemeiner für die Verteilungen von Summen vieler kleiner unabhängiger Zufallsvariablen (*Zentraler Grenzwertsatz*, s.u.).

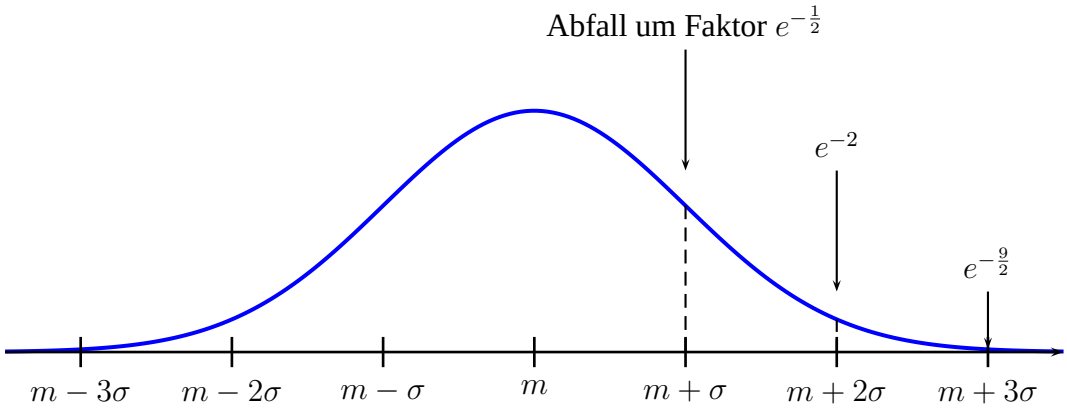


Abbildung 4.13: Dichte einer normalverteilten Zufallsvariable mit Mittelwert m und Varianz σ^2.

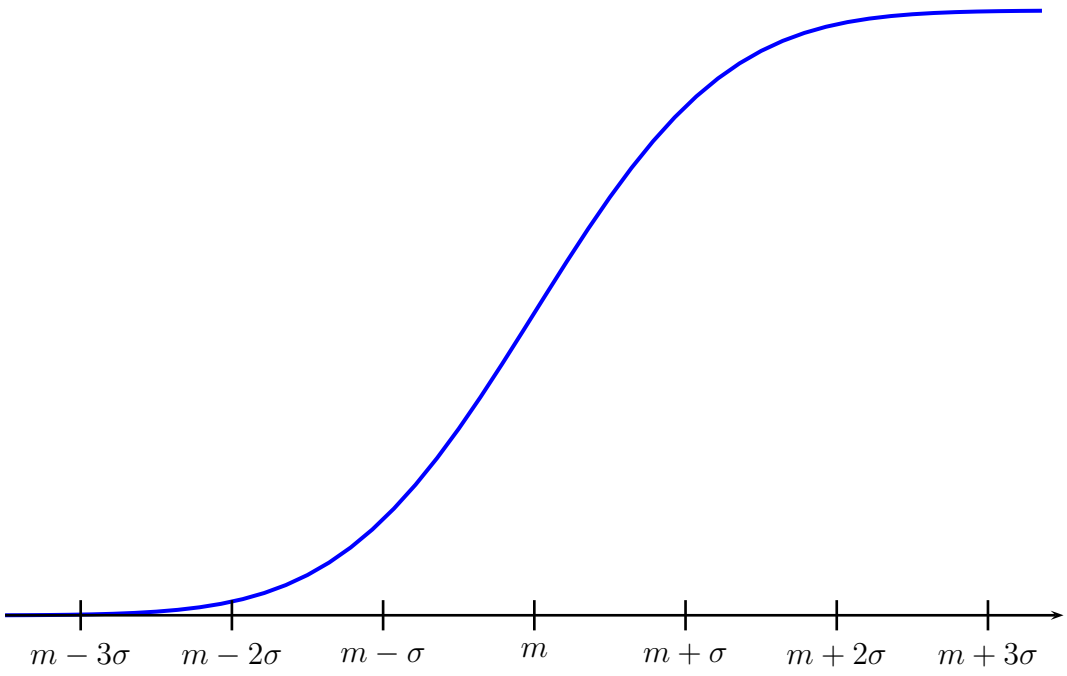


Abbildung 4.14: Verteilungsfunktion einer normalverteilten Zufallsvariable mit Mittelwert m und Varianz σ^2.

Die Dichte der Normalverteilung ist an der Stelle m maximal, und klingt außerhalb einer σ-Umgebung von m rasch ab. Beispielsweise gilt

$$f_{m,\sigma}(m \pm \sigma) = \frac{f_{m,\sigma}(m)}{\sqrt{e}}$$

$$f_{m,\sigma}(m \pm 2\sigma) = \frac{f_{m,\sigma}(m)}{e^2}$$

$$f_{m,\sigma}(m \pm 3\sigma) = \frac{f_{m,\sigma}(m)}{e^{9/2}}$$

Für die Wahrscheinlichkeit, dass eine normalverteilte Zufallsvariable Werte außerhalb der σ-, 2σ- und 3σ-Umgebungen annimmt, erhält man:

$$\begin{aligned} P[|X-m| > k\sigma] &= P\left[\left|\frac{X-m}{\sigma}\right| > k\right] \\ &= P[|Z| > k] = 2P[Z > k] = 2(1-\Phi(k)) \\ &= \begin{cases} 31.7\% & \text{für } k=1 \\ 4.6\% & \text{für } k=2 \\ 0.26\% & \text{für } k=3 \end{cases} \end{aligned}$$

Eine Abweichung der Größe σ vom Mittelwert m ist also für eine normalverteilte Zufallsvariable relativ typisch, eine Abweichung der Größe 3σ dagegen schon sehr selten.

Die folgenden expliziten Abschätzungen für die Wahrscheinlichkeiten großer Werte sind oft nützlich:

Lemma 4.17. *Für eine standardnormalverteilte Zufallsvariable Z gilt:*

$$(2\pi)^{-1/2} \cdot \left(\frac{1}{y} - \frac{1}{y^3}\right) \cdot e^{-y^2/2} \quad \leq \quad P[Z \geq y] \quad \leq \quad (2\pi)^{-1/2} \cdot \frac{1}{y} \cdot e^{-y^2/2} \qquad \forall y > 0$$

Beweis. Es gilt:

$$P[Z \geq y] = (2\pi)^{-1/2} \int_y^\infty e^{-z^2/2}\, dz$$

Um das Integral abzuschätzen, versuchen wir approximative Stammfunktionen zu finden. Zunächst gilt:

$$\frac{d}{dz}\left(-\frac{1}{z}e^{-z^2/2}\right) = \left(1 + \frac{1}{z^2}\right) \cdot e^{-z^2/2} \geq e^{-z^2/2} \qquad \forall z \geq 0,$$

also

$$\frac{1}{y}e^{-z^2/2} = \int_y^\infty \left(\frac{1}{y}e^{-z^2/2}\right) dz \geq \int_y^\infty e^{-z^2/2}\, dz,$$

woraus die obere Schranke für $P[Z \geq y]$ folgt.

Für die untere Schranke approximieren wir die Stammfunktion noch etwas genauer. Es gilt:

$$\frac{d}{dz}\left(\left(-\frac{1}{z}+\frac{1}{z^3}\right)e^{-z^2/2}\right) = \left(1+\frac{1}{z^2}-\frac{1}{z^2}-\frac{3}{z^4}\right)e^{-z^2/2} \leq e^{-z^2/2},$$

und damit

$$\left(\frac{1}{y}-\frac{1}{y^3}\right)e^{-y^2/2} \leq \int_y^\infty e^{-z^2/2}\, dz.$$

□

Für eine $N(m,\sigma^2)$-verteilte Zufallsvariable X mit $\sigma > 0$ ist $Z = \frac{X-m}{\sigma}$ standardnormalverteilt. Also erhalten wir für $y \geq m$:

$$P[X \geq y] \;=\; P\left[\frac{X-m}{\sigma} \;\geq\; \frac{y-m}{\sigma}\right] \leq \frac{1}{y-m}\cdot(2\pi\sigma)^{-1/2}\cdot e^{-\frac{(y-m)^2}{2\sigma^2}},$$

sowie eine entsprechende Abschätzung nach unten.

Transformation von absolutstetigen Zufallsvariablen

Wir haben in Beispielen bereits mehrfach die Verteilung von Funktionen von absolutstetigen Zufallsvariablen berechnet. Sei nun allgemein $I \subseteq \mathbb{R}$ ein offenes Intervall, und $X : \Omega \to I$ eine Zufallsvariable mit stetiger Verteilung.

Satz 4.18 (Eindimensionaler Dichtetransformationssatz). *Ist $\Phi : I \to J$ einmal stetig differenzierbar mit $\Phi'(x) \neq 0$ für alle $x \in I$, dann ist die Verteilung von $\Phi(X)$ absolutstetig mit Dichte*

$$f_{\Phi(X)}(y) \quad = \quad \begin{cases} f_X(\Phi^{-1}(y))\cdot|(\Phi^{-1})'(y)| & \text{für } y \in \Phi(I) \\ 0 & \text{sonst} \end{cases}. \tag{4.4.4}$$

Beweis. Nach der Voraussetzung gilt entweder $\Phi' > 0$ auf I oder $\Phi' < 0$ auf I. Wir betrachten nur den ersten Fall. Aus $\Phi' > 0$ folgt, dass Φ streng monoton wachsend ist, also eine Bijektion von I nach $\Phi(I)$. Daher erhalten wir

$$F_{\Phi(X)}(c) \quad = \quad P[\Phi(X) \leq c] \quad = \quad P[X \leq \Phi^{-1}(I)] \quad = \quad F_X(\Phi^{-1}(c))$$

für alle $c \in \Phi(I)$. Nach der Kettenregel ist dann $F_{\Phi(X)}$ für fast alle $c \in \Phi(I)$ differenzierbar, und es gilt

$$F'_{\Phi(X)}(c) \quad = \quad f_X(\Phi^{-1}(c)) \cdot (\Phi^{-1})'(c).$$

Die Behauptung folgt hieraus nach dem Hauptsatz der Differential- und Integralrechnung, da

$$P[\Phi(x) \notin \Phi(I)] = 0.$$

□

Beispiel (Geometrische Wahrscheinlichkeiten). Sei $\theta : \Omega \to [0, 2\pi)$ ein zufälliger, auf $[0, 2\pi)$ gleichverteilter, Winkel. Wir wollen die Verteilung von $\cos\theta$ berechnen. Da die Kosinusfunktion auf $[0, 2\pi)$ nicht streng monoton ist, ist (4.4.4) nicht direkt anwendbar. Wir können aber das Intervall $[0, 2\pi)$ in die Teile $[0, \pi)$ und $[\pi, 2\pi)$ zerlegen, und dann die Verteilung ähnlich wie im Beweis von Satz 4.18 berechnen. Wegen

$$\begin{aligned} P[\cos\theta > c] &= P[\cos\theta > c \quad \text{und} \quad \theta \in [0, \pi)] + P[\cos\theta > c \quad \text{und} \quad \theta \in [\pi, 2\pi)] \\ &= P[\theta \in [0, \arccos c)] + P[\theta \in [\pi - \arccos c, \pi)] \\ &= \frac{2}{2\pi} \cdot \arccos c \end{aligned}$$

erhalten wir, dass $\cos\theta$ eine sogenannte „Halbkreisverteilung“ mit Dichte

$$f_{\cos\theta}(x) \quad = \quad F'_{\cos\theta}(x) \quad = \quad \frac{1}{\pi} \cdot \frac{1}{\sqrt{1-x^2}}; \qquad x \in [-1, 1)$$

hat.

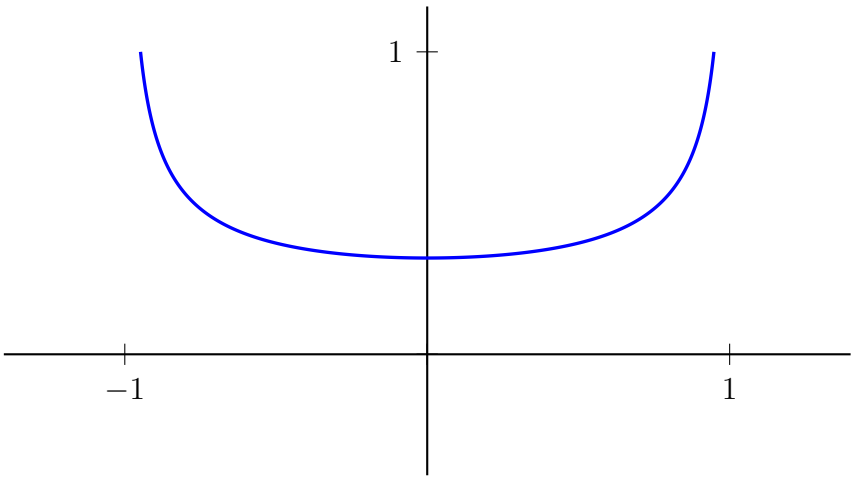


Abbildung 4.15: Abbildung der Dichtefunktion $f_{\cos\theta}$

Anstelle von (4.4.4) gilt in diesem Fall

$$f_{\cos\theta}(x) \quad = \quad f_X(\psi_1(x)) \cdot |\psi_1'(x)| + f_X(\psi_2(x)) \cdot |\psi_2'(x)|,$$

wobei $\psi_1(x) = \arccos x$ und $\psi_2(x) = 2\pi - \arccos x$ die Umkehrfunktionen auf den Teilintervallen sind. Entsprechende Formeln erhält man auch allgemein, wenn die Transformation nur stückweise bijektiv ist. Auf ähnliche Weise zeigt man für $a > 0$ (Übung):

$$f_{a\tan\theta}(x) \quad = \quad \frac{1}{\pi a} \cdot \frac{1}{1 + (x/a)^2}, \qquad x \in \mathbb{R}.$$

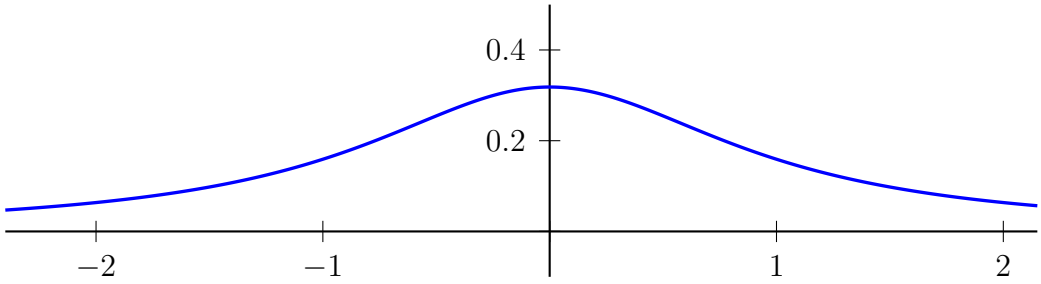
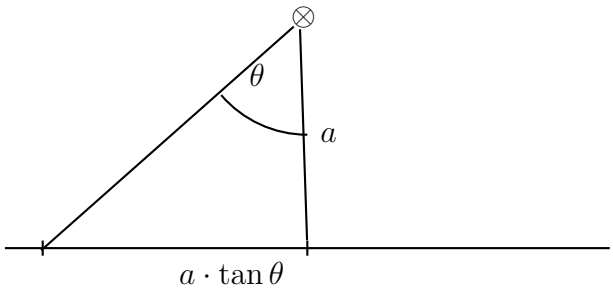


Abbildung 4.16: Abbildung der Dichtefunktion $f_{a\tan\theta}$

Die Verteilung mit dieser Dichte heißt **Cauchyverteilung** zum Parameter a. Sie beschreibt unter anderem die Intensitätsverteilung auf einer Geraden, die von einer in alle Richtungen gleichmäßig strahlenden Lichtquelle im Abstand a bestrahlt wird.



4.5 Quantile und Inversionsverfahren

Quantile sind Stellen, an denen die Verteilungsfunktion einen bestimmten Wert überschreitet. Mithilfe von Quantilen kann man daher verallgemeinerte Umkehrfunktionen der im Allgemeinen nicht bijektiven Verteilungsfunktion definieren. Diese Umkehrabbildungen werden wir nutzen, um reellwertige Zufallsvariablen mit einer gegebenen Verteilungsfunktion explizit zu konstruieren.

Quantile

In praktischen Anwendungen (z.B. Qualitätskontrolle) müssen häufig Werte berechnet werden, sodass ein vorgegebener Anteil der Gesamtmasse einer Wahrscheinlichkeitsverteilung auf $\mathbb{R}$ unterhalb dieses Wertes liegt. Sei $X : \Omega \to \mathbb{R}$ eine Zufallsvariable auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ mit Verteilungsfunktion F.

Definition. *Sei $u \in [0,1]$. Dann heißt $q \in \mathbb{R}$ ein u-**Quantil** der Verteilung von X, falls*

$$P[X < q] \leq u \qquad \textit{und} \qquad P[X > q] \leq 1 - u$$

*gilt. Ein $\frac{1}{2}$-Quantil heißt **Median**.*

Ist die Verteilungsfunktion nicht streng monoton wachsend, dann kann es mehrere u-Quantile zu einem Wert u geben.

Beispiel (Stichprobenquantile). Wir betrachten eine Stichprobe, die aus n reellwertigen Daten / Messwerten $x_1, \ldots, x_n$ mit $x_1 \leq x_2 \leq \ldots \leq x_n$ besteht. Die **empirische Verteilung der Stichprobe** ist die Wahrscheinlichkeitsverteilung

$$\mu = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$$

auf $(\mathbb{R}, \mathcal{P}(\mathbb{R}))$, d.h. für $B \subseteq \mathbb{R}$ ist

$$\mu[B] = \frac{1}{n} \left| \{x_i \in B, 1 \leq i \leq n\} \right|$$

die relative Häufigkeit des Bereichs B unter den Messwerten x_i. Die empirische Verteilung ergibt sich, wenn wir zufällig ein $i \in \{1, \ldots, n\}$ wählen, und den entsprechenden Messwert betrachten.

Die Quantile der empirischen Verteilung bezeichnet man als **Stichprobenquantile**. Für $u \in [0,1]$ sei

$$k_u := 1 + (n-1)u \in [1, n].$$

Ist k_u ganzzahlig, dann ist x_{k_u} das eindeutige u-Quantil der Stichprobe. Allgemein ist jedes $q \in [x_{\lfloor k_u \rfloor}, x_{\lceil k_u \rceil}]$ ein u-Quantil der Stichprobe, d.h. für $k_u \notin \mathbb{Z}$ gibt es mehrere u-Quantile.

Wir definieren nun zwei verallgemeinerte Inverse einer Verteilungsfunktion F, die ja im Allgemeinen nicht bijektiv ist. Für $u \in (0,1)$ sei

$$\underline{G}(u) := \inf\{x \in \mathbb{R} | F(x) \geq u\} = \sup\{x \in \mathbb{R} | F(x) < u\}$$

und

$$\overline{G}(u) := \inf\{x \in \mathbb{R} | F(x) > u\} = \sup\{x \in \mathbb{R} | F(x) \leq u\}.$$

Offensichtlich gilt $\underline{G}(u) \leq \overline{G}(u)$. Ist die Funktion F stetig und streng monoton wachsend, also eine Bijektion von $\mathbb{R}$ nach $(0,1)$, dann gilt $\underline{G}(u) = \overline{G}(u) = F^{-1}(u)$. Die Funktion $\underline{G}$ heißt daher auch die **linksstetige verallgemeinerte Inverse** von F. Der folgende Satz zeigt, dass $\underline{G}(u)$ das kleinste und $\overline{G}(u)$ das größte u-Quantil ist:

Satz 4.19. *Für $u \in (0,1)$ und $q \in \mathbb{R}$ sind die folgenden Aussagen äquivalent:*

(1). q ist ein u-Quantil.

(2). $F(q-) \leq u \leq F(q)$.

(3). $\underline{G}(u) \leq q \leq \overline{G}(u)$.

Hierbei ist $F(q-) := \lim_{y \nearrow q} F(y)$ der linksseitige Limes von F an der Stelle q.

Beweis. Nach Definition ist q genau dann ein u-Quantil, wenn

$$P[X < q] \leq u \leq 1 - P[X > q] = P[X \leq q]$$

gilt. Hieraus folgt die Äquivalenz von (1) und (2).

Um zu beweisen, dass (3) äquivalent zu diesen Bedingungen ist, müssen wir zeigen, dass $\underline{G}(u)$ das kleinste und $\overline{G}(u)$ das größte u-Quantil ist. Wir bemerken zunächst, dass $\underline{G}(u)$ ein u-Quantil ist, da

$$F(\underline{G}(u)-) = \lim_{x \nearrow \underline{G}(u)} \underbrace{F(x)}_{\substack{<u \\ \text{für } x<\underline{G}(u)}} \leq u,$$

und

$$F(\underline{G}(u)) = \lim_{x \searrow \underline{G}(u)} \underbrace{F(x)}_{\substack{\geq u \\ \text{für } x>\underline{G}(u)}} \geq u.$$

Andererseits gilt für $x < \underline{G}(u)$:

$$F(x) < u,$$

d.h. x ist kein u-Quantil. Somit ist $\underline{G}(u)$ das kleinste u-Quantil. Auf ähnliche Weise folgt, dass $\overline{G}(u)$ das größte u-Quantil ist (Übung!). □

Konstruktion und Simulation reellwertiger Zufallsvariablen

Wie erzeugt man ausgehend von auf $(0,1)$ gleichverteilten Zufallszahlen Stichproben von anderen Verteilungen μ auf $\mathbb{R}^1$?

Endlicher Fall: Gilt $\mu(S) = 1$ für eine endliche Teilmenge $S \subseteq \mathbb{R}$, dann können wir die Frage leicht beantworten: Sei $S = \{x_1, \ldots, x_n\} \subseteq \mathbb{R}$ mit $n \in \mathbb{N}$ und $x_1 < x_2 < \ldots < x_n$. Die Verteilungsfunktion einer Wahrscheinlichkeitsverteilung μ auf S ist

$$F(c) = \mu[(-\infty, c]] = \sum_{i:x_i \le c} \mu(\{x_i\}).$$

Ist U eine auf $(0,1)$ gleichverteilte Zufallsvariable, dann wird durch

$$X(\omega) = x_k \qquad \text{falls } F(x_{k-1}) < U(\omega) \le F(x_k), \quad x_0 := -\infty$$

eine Zufallsvariable mit Verteilung μ definiert, denn

$$P[X = x_k] = F(x_k) - F(x_{k-1}) = \mu[\{x_k\}].$$

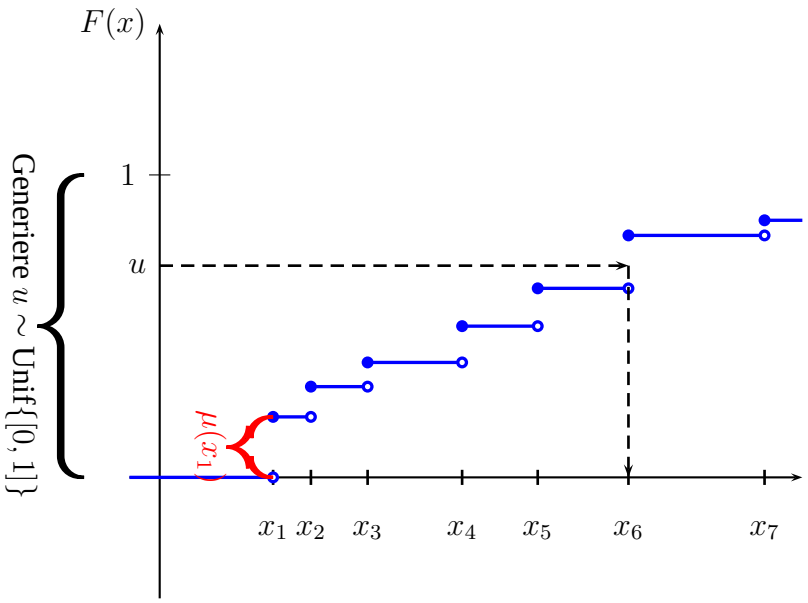


Abbildung 4.17: Wir generieren eine uniform auf $(0,1)$ verteilte Pseudozufallszahl u. Suche nun das minimale $k \in \mathbb{N}$, für das $\sum_{i=1}^{k} \mu(x_i) > u$. Dann ist $x = x_k$ eine Pseudozufallsstichprobe von der Verteilung μ.

Allgemeiner Fall: Wir wollen das Vorgehen nun verallgemeinern. Sei $F : \mathbb{R} \to [0,1]$ eine Funktion mit den Eigenschaften

(1). monoton wachsend: $\quad F(x) \leq F(y) \quad \forall\, x \leq y$

(2). rechtsstetig: $\quad \lim_{x \downarrow c} F(x) = F(c) \quad \forall\, c \in \mathbb{R}$

(3). normiert: $\quad \lim_{x \searrow -\infty} F(x) = 0 \quad , \quad \lim_{x \nearrow +\infty} F(x) = 1.$

Das folgende Resultat liefert eine explizite Konstruktion einer Zufallsvariable mit Verteilungsfunktion F:

Satz 4.20. *Ist $F : \mathbb{R} \to [0,1]$ eine Funktion mit (1)-(3), und*

$$\underline{G}(u) = \inf\{x \in \mathbb{R} | F(x) \geq c\}, \quad u \in (0,1),$$

die linksstetige verallgemeinerte Inverse, dann ist das Bild

$$\mu := \mathcal{U}_{(0,1)} \circ \underline{G}^{-1}$$

der Gleichverteilung auf $(0,1)$ unter $\underline{G}$ eine Wahrscheinlichkeitsverteilung auf $\mathbb{R}$ mit Verteilungsfunktion F.
Insbesondere gilt: Ist $U : \Omega \to (0,1)$ eine unter P gleichverteilte Zufallsvariable, dann hat die Zufallsvariable

$$X(\omega) := \underline{G}(U(\omega))$$

unter P die Verteilungsfunktion F.

Beweis. Da $\underline{G}(u)$ ein u-Quantil ist, gilt $F(\underline{G}(u)) \geq u$, also

$$\underline{G}(u) = \min\{x \in \mathbb{R} | F(x) \geq u\},$$

und somit für $c \in \mathbb{R}$:

$$\underline{G}(u) \leq c \iff F(x) \geq u \text{ für ein } x \leq c \iff F(c) \geq u.$$

Es folgt:

$$\begin{aligned} P[\underline{G}(U) \leq c] &= \mathcal{U}_{(0,1)}[\{u \in (0,1) | \underbrace{\underline{G}(u) \leq c}_{\iff F(c) \geq u}\}] \\ &= \mathcal{U}_{(0,1)}[(0, F(c))] = F(c). \end{aligned}$$

Also ist F die Verteilungsfunktion von $G(U)$ bzw. von μ. □

Bemerkung. (1). Ist F eine Bijektion von $\mathbb{R}$ nach $(0,1)$ (also stetig und streng monoton wachsend), dann ist $\underline{G} = F^{-1}$.

(2). Nimmt X nur endlich viele Werte $x_1 < x_2 < \ldots < x_n$ an, dann ist F stückweise konstant, und es gilt:

$$\underline{G}(u) = x_k \text{ für } F(x_{k-1}) < u \leq F(x_k), \quad x_0 := -\infty,$$

d.h. $\underline{G}$ ist genau die oben im endlichen Fall verwendete Transformation.

Das Resultat liefert einen

Existenzsatz: Zu jeder Funktion F mit (1)-(3) existiert eine reelle Zufallsvariable X bzw. eine Wahrscheinlichkeitsverteilung μ auf $\mathbb{R}$ mit Verteilungsfunktion F.

Zudem erhalten wir einen expliziten Algorithmus zur Simulation einer Stichprobe von μ:

Algorithmus 4.21 (Inversionsverfahren zur Simulation einer Stichprobe x von μ).

(1). Erzeuge (Pseudo)-Zufallszahl $u \in (0,1)$.

(2). Setze $x := \underline{G}(u)$.

Dieser Algorithmus funktioniert theoretisch immer. Er ist aber oft nicht praktikabel, da man $\underline{G}$ nicht immer berechnen kann, oder da das Anwenden der Transformation $\underline{G}$ (zunächst unwesentliche) Schwachstellen des verwendeten Zufallsgenerators verstärkt. Man greift daher oft selbst im eindimensionalen Fall auf andere Simulationsverfahren wie z.B. „Acceptance Rejection“ Methoden zurück.

Beispiel. (1). BERNOULLI(p)-VERTEILUNG AUF $\{0,1\}$. Hier gilt:

$$F = (1-p) \cdot I_{[}0,1) + 1 \cdot I_{[1,\infty)}$$

und $\underline{G} = \mathbb{1}_{(1-p,1)}$, siehe Abbildung 4.18.
Also ist die Zufallsvariable $\underline{G}(U) = I_{\{U<1-p\}}$ für $U \sim \mathcal{U}_{(0,1)}$ Bernoulli(p)-verteilt.

(2). GLEICHVERTEILUNG AUF (a,b):

$$F(c) = \frac{c-a}{b-a} \qquad \text{für } c \in [a,b],$$

$$\underline{G}(u) = a + (b-a)u,$$

siehe Abbildung 4.19. Also ist $a + (b-a)U$ für $U \sim \mathcal{U}_{(0,1)}$ gleichverteilt auf (a,b).

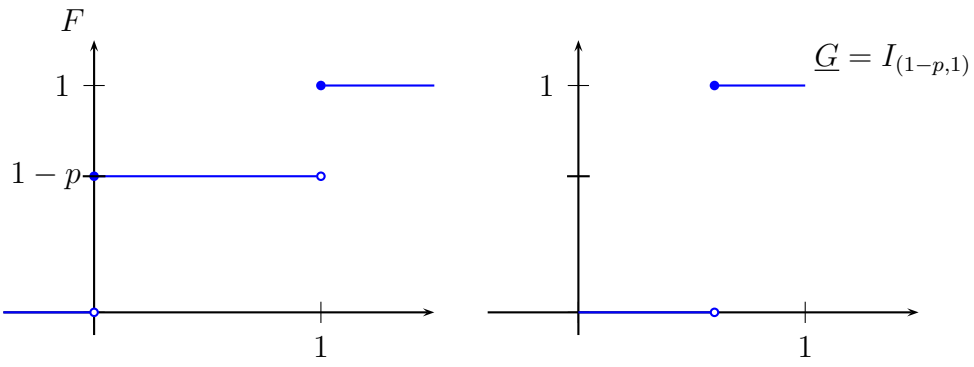


Abbildung 4.18: $\underline{G}(U) = I_{\{U>1-p\}}$ ist Bernoulli(p)-verteilt.

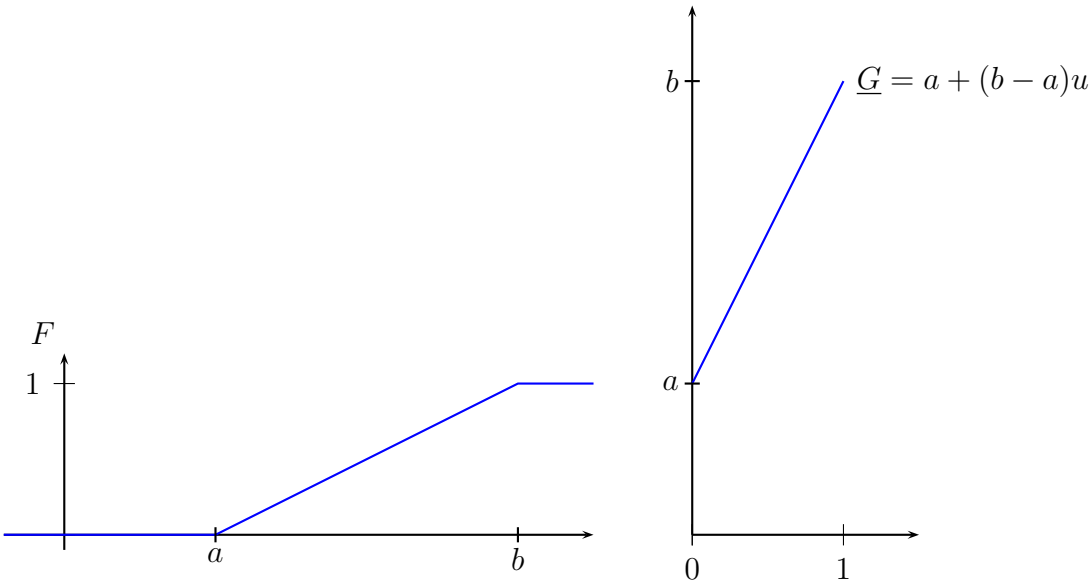


Abbildung 4.19: $\underline{G}(u) = a + (b-a)u$ ist (für $u \sim \text{unif}\{(0,1)\}$) uniform auf (a,b) verteilt.

(3). EXPONENTIALVERTEILUNG MIT PARAMETER $\lambda > 0$:

$$F(x) = 1 - e^{-\lambda x}, \quad G(u) = F^{-1}(u) = -\frac{1}{\lambda}\log(1-u).$$

Anwenden des Logarithmus transformiert also die gleichverteilte Zufallsvariable $1-u$ in eine exponentialverteilte Zufallsvariable.

4.6 Normalapproximation der Binomialverteilung

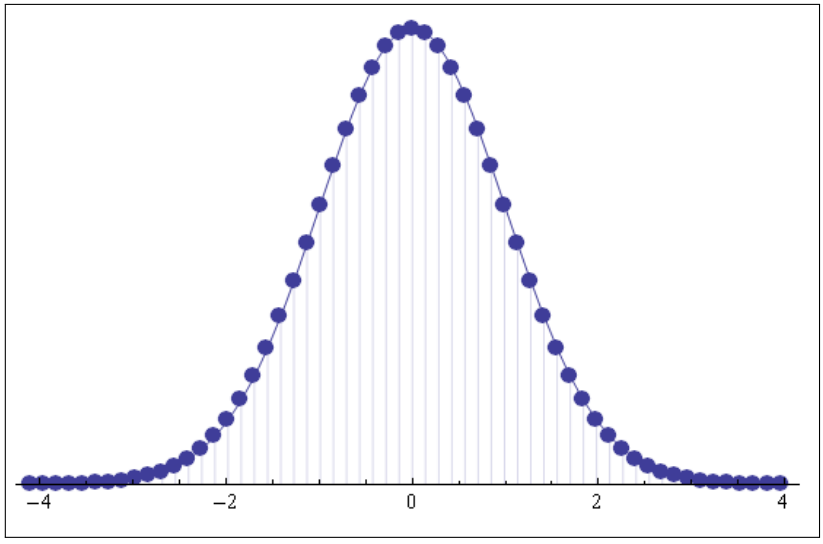
Die Binomialverteilung mit Parametern n und p beschreibt die Verteilung der Anzahl derjenigen unter n unabhängigen Ereignissen mit Wahrscheinlichkeit p, die in einem Zufallsexperiment eintreten. Viele Anwendungsprobleme führen daher auf die Berechnung von Wahrscheinlichkeiten bzgl. der Binomialverteilung. Für große n ist eine exakte Berechnung dieser Wahrscheinlichkeiten aber in der Regel nicht mehr möglich. Bei seltenen Ereignissen kann man die Poissonapproximation zur näherungsweisen Berechnung nutzen:

Konvergiert $n \to \infty$, und konvergiert gleichzeitig der Erwartungswert $n \cdot p_n$ gegen eine positive reelle Zahl $\lambda > 0$, dann nähern sich die Gewichte $b_{n,p_n}(k)$ der Binomialverteilung denen einer Poissonverteilung mit Parameter λ an:

$$b_{n,p_n}(k) = \binom{n}{k} p_n^k (1-p_n)^{n-k} \to \frac{\lambda^k}{k} e^{-\lambda} \qquad (k = 0, 1, 2, \ldots),$$

siehe Satz 1.5. Geht die Wahrscheinlichkeit p_n für $n \to \infty$ nicht gegen 0, sondern hat zum Beispiel einen festen Wert $p \in (0, 1)$, dann kann die Poissonapproximation nicht verwendet werden. Stattdessen scheinen sich die Gewichte der Binomialverteilung einer Gaußschen Glockenkurve anzunähern, wie z.B. die folgende mit Mathematica erstellte Grafik zeigt:

```
Manipulate[
 ListPlot[
  Table[{k, PDF[BinomialDistribution[n, Min[1, lambda /n]], k]}, {k, 0,
     IntegerPart[4 lambda]}],
  Filling -> Axis, PlotRange -> All,
  PlotMarkers -> {Automatic, Medium}, Axes -> {True, False}] , {{n, 10,
     "n"}, 3, 300,1},
  {{lambda, 5, "Erwartungswert: np=Lambda"}, 2, 20}]
```



Wir wollen diese Aussage nun mathematisch präzisieren und beweisen.

Der Satz von De Moivre - Laplace

Wir analysieren zunächst das asymptotische Verhalten von Binomialkoeffizienten mithilfe der Stirlingschen Formel.

Definition. *Zwei Folgen $a_n, b_n \in \mathbb{R}_+$, $n \in \mathbb{N}$, heißen **asymptotisch äquivalent** $(a_n \sim b_n)$, falls*

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

gilt.

Bemerkung.

(1). $a_n \sim b_n \quad \Longleftrightarrow \quad \exists \varepsilon_n \to 0 : a_n = b_n(1+\varepsilon_n) \quad \Longleftrightarrow \quad \log a_n - \log b_n \to 0$

(2). $a_n \sim b_n \quad \Longleftrightarrow \quad b_n \sim a_n \quad \Longleftrightarrow \quad \frac{1}{a_n} \sim \frac{1}{b_n}$

(3). $a_n \sim b_n, c_n \sim d_n \quad \Longrightarrow \quad a_n \cdot c_n \sim b_n \cdot d_n$

Satz 4.22 (Stirlingsche Formel).

$$n! \sim \sqrt{2\pi n} \cdot \left(\frac{n}{e}\right)^n$$

Zum Beweis nimmt man den Logarithmus, und schätzt die sich ergebende Summe mithilfe eines Integrals ab, siehe z.B. Forster: „Analysis I“.

Mithilfe der Stirlingschen Formel können wir die Gewichte

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

der Binomialverteilung für große n und k approximieren. Sei dazu S_n eine $\mathrm{Bin}(n,p)$-verteilte Zufallsvariable auf $(\Omega, \mathcal{A}, P)$. Für den Erwartungswert und die Standardabweichung von S_n gilt:

$$E[S_n] = np \quad \text{und} \quad \sigma(S_n) = \sqrt{Var[S_n]} = \sqrt{np(1-p)}.$$

Dies deutet darauf hin, dass sich die Masse der Binomialverteilung für große n überwiegend in einer Umgebung der Größenordnung $O(\sqrt{n})$ um np konzentriert.

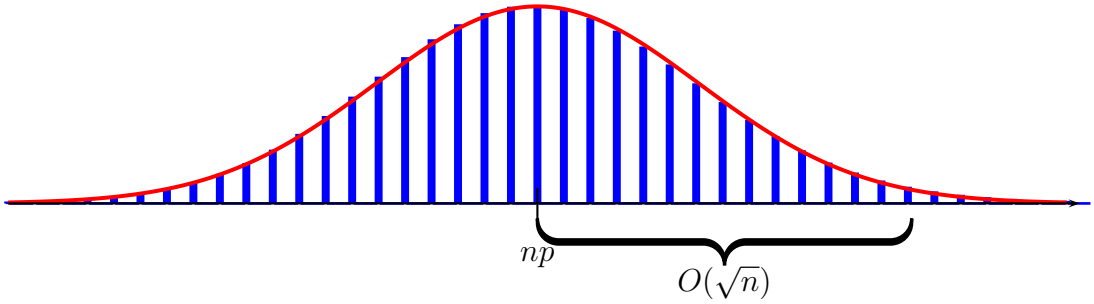


Abbildung 4.20: Die Gewichte der Binomialverteilung liegen für große n näherungsweise auf einer Glockenkurve mit Mittel np und Standardabweichung $\sqrt{np(1-p)}$.

Wir werden nun mithilfe der Stirlingschen Formel die Gewichte

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

der Binomialverteilung für große n und k in einer Umgebung der Größenordnung $O(\sqrt{n})$ von np ausgehend von der Stirlingschen Formel approximieren, und die vermutete asymptotische Darstellung präzisieren und beweisen.

Dazu führen wir noch folgende Notation ein: Wir schreiben

$$a_n(k) \approx b_n(k) \qquad (\text{„lokal gleichmäßig asymptotisch äquivalent“}),$$

falls

$$\sup_{k \in U_{n,r}} \left| \frac{a_n(k)}{b_n(k)} - 1 \right| \to 0 \quad \text{für alle } r \in \mathbb{R}_+ \text{ gilt,}$$

wobei

$$U_{n,r} \quad = \quad \{0 \le k \le n : |k - np| \le r \cdot \sqrt{n}\}.$$

Die Aussagen aus der Bemerkung oben gelten analog für diese Art der lokal gleichmäßigen asymptotischen Äquivalenz von $a_n(k)$ und $b_n(k)$.

Satz 4.23 (de Moivre 1733, Laplace 1819). *Sei $p \in (0,1)$ und $\sigma^2 = p(1-p)$. Dann gilt:*

(1). $$P[S_n = k] \quad = \quad b_{n,p}(k) \quad \approx \quad \frac{1}{\sqrt{2\pi n \sigma^2}} \exp\left(-\frac{1}{2\sigma^2} \left(\frac{k-np}{\sqrt{n}} \right)^2 \right) \quad =: \quad \widetilde{b}_{n,p}(k)$$

(2). $$P\left[a \le \frac{S_n - np}{\sqrt{n}} \le b \right] \quad \overset{n \nearrow \infty}{\longrightarrow} \quad \int_a^b \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}}_{\text{Gaußsche Glockenkurve}} \, dx$$

Beweis. (1). Wir beweisen die Aussage in zwei Schritten:

(a) Wir zeigen zunächst mithilfe der **Stirlingschen Formel**:

$$b_{n,p}(k) \quad \approx \quad \frac{1}{\sqrt{2\pi n \frac{k}{n}(1-\frac{k}{n})}} \cdot \left(\frac{p}{\frac{k}{n}}\right)^k \cdot \left(\frac{1-p}{1-\frac{k}{n}}\right)^{n-k} \quad =: \quad \bar{b}_{n,p}(k) \tag{4.6.1}$$

Es gilt

$$\lim_{n\to\infty} \frac{n!}{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n} \quad = \quad 1.$$

Für $k \in U_{n,r}$ gilt

$$k \quad \geq \quad np - A\cdot\sqrt{n} \qquad \stackrel{n\to\infty}{\longrightarrow} \quad \infty,$$

also folgt

$$\sup_{k\in U_{n,r}} \left| \frac{k!}{\sqrt{2\pi k}\left(\frac{k}{e}\right)^k} - 1 \right| \quad \longrightarrow \quad 0 \qquad \text{für } n\to\infty,$$

d.h.

$$k! \quad \approx \quad \sqrt{2\pi k}\left(\frac{k}{e}\right)^k.$$

Analog erhält man

$$(n-k)! \quad \approx \quad \sqrt{2\pi(n-k)}\left(\frac{n-k}{e}\right)^{n-k},$$

und damit

$$\begin{aligned} b_{n,p}(k) &= \frac{n!}{k!\cdot(n-k)!} p^k (1-p)^{n-k} \\ &\approx \frac{\sqrt{2\pi n}\cdot n^n \cdot p^k \cdot (1-p)^{n-k}}{2\pi\sqrt{k(n-k)}\cdot k^k \cdot (n-k)^{n-k}} \\ &= \sqrt{\frac{n}{2\pi k(n-k)}} \left(\frac{np}{k}\right)^k \left(\frac{n(1-p)}{n-k}\right)^{n-k} \\ &= \bar{b}_{n,p}(k). \end{aligned}$$

(b) Wir zeigen nun mithilfe einer **Taylorapproximation**:

$$\bar{b}_{n,p}(k) \quad \approx \quad \widetilde{b}_{n,p}(k) \tag{4.6.2}$$

Für $k \in U_{n,r}$ gilt

$$\left|\frac{k}{n} - p\right| \quad \leq \quad r\cdot n^{-\frac{1}{2}},$$

woraus folgt:

$$\sqrt{2\pi \cdot n \cdot \frac{k}{n} \cdot \left(1 - \frac{k}{n}\right)} \quad \approx \quad \sqrt{2\pi \cdot n \cdot p \cdot (1-p)} \quad = \quad \sqrt{2\pi \cdot n \cdot \sigma^2}. \tag{4.6.3}$$

Um die Asymptotik der übrigen Faktoren von $\overline{b}_{n,p}(k)$ zu erhalten, verwenden wir eine **Taylorapproximation für den Logarithmus**:

Wegen

$$x \log \frac{x}{p} \quad = \quad x - p + \frac{1}{2p}(x-p)^2 + O(|x-p|^3)$$

gilt:

$$\begin{aligned}
& \log \left(\left(\frac{p}{\frac{k}{n}} \right)^k \left(\frac{1-p}{1-\frac{k}{n}} \right)^{n-k} \right) \\
= \; & (-n) \left[\underbrace{\frac{k}{n} \log \left(\frac{\frac{k}{n}}{p} \right)}_{\overset{\text{Taylor}}{=} \frac{k}{n} - p + \frac{1}{2p}(\frac{k}{n}-p)^2 + O(|\frac{k}{n}-p|^3)} + \underbrace{\left(1 - \frac{k}{n}\right) \log \left(\frac{1-\frac{k}{n}}{1-p} \right)}_{= p - \frac{k}{n} + \frac{1}{2(1-p)}(p-\frac{k}{n})^2 + O(|\frac{k}{n}-p|^3)} \right] \\
= \; & \underbrace{\frac{1}{2p}(\frac{k}{n}-p)^2 + \frac{1}{2(1-p)}(p-\frac{k}{n})^2}_{= \frac{(p-\frac{k}{n})^2}{2} \underbrace{\left(\frac{1}{p} + \frac{1}{1-p} \right)}_{= \frac{1}{p(1-p)}}} + O(|\frac{k}{n}-p|^3) \\
= \; & \frac{1}{2p(1-p)}(p-\frac{k}{n})^2 + O(|\frac{k}{n}-p|^3)
\end{aligned}$$

Für $k \in U_{n,r}$ gilt:

$$\left| \frac{k}{n} - p \right|^3 \quad \leq \quad r^3 \cdot n^{-\frac{3}{2}}.$$

Also folgt:

$$\log \left(\left(\frac{p}{\frac{k}{n}} \right)^k \left(\frac{1-p}{1-\frac{k}{n}} \right)^{n-k} \right) \quad = \quad -\frac{1}{2\sigma^2} \left(\frac{k}{n} - p \right)^2 + R_{k,n},$$

wobei $|R_{k,n}| \leq \text{const.} \cdot r^3 n^{-\frac{1}{2}}$ für alle $k \in U_{n,r}$, d.h.

$$\left(\frac{p}{\frac{k}{n}} \right)^k \left(\frac{1-p}{1-\frac{k}{n}} \right)^{n-k} \quad \approx \quad \exp \left(-\frac{1}{2\sigma^2} \left(\frac{k}{n} - p \right)^2 \right). \tag{4.6.4}$$

Aussage (4.6.2) folgt dann aus (4.6.3) und (4.6.4).

(c) Aus (a) und (b) folgt nun Behauptung (1).

(2). Aufgrund von (1) erhalten wir für $a, b \in \mathbb{R}$ mit $a < b$:

$$\begin{aligned} P\left[a \le \frac{S_n - np}{\sqrt{n}} \le b\right] &= \sum_{\substack{k \in \{0,1,\ldots,n\} \\ a \le \frac{k-np}{\sqrt{n}} \le b}} \underbrace{P[S_n = k]}_{= b_{n,p}(k) \approx \widetilde{b}_{n,p}(k)} \\ &= \sum_{\substack{k \in \{0,1,\ldots,n\} \\ a \le \frac{k-np}{\sqrt{n}} \le b}} \widetilde{b}_{n,p}(k)(1 + \varepsilon_{n,p}(k)), \end{aligned}$$

wobei

$$\overline{\varepsilon}_{n,p} := \sup_{a \le \frac{k-np}{\sqrt{n}} \le b} |\varepsilon_{n,p}(k)| \longrightarrow 0 \qquad \text{für } n \to \infty. \tag{4.6.5}$$

Wir zeigen nun

$$\lim_{n\to\infty} \sum_{\substack{k \in \{0,1,\ldots,n\} \\ a \le \frac{k-np}{\sqrt{n}} \le b}} \widetilde{b}_{n,p}(k) = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{x^2}{2\sigma^2}\right) dx \tag{4.6.6}$$

Aus (4.6.5) und (4.6.6) folgt dann die Behauptung, da

$$\left| \sum_{\substack{k \in \{0,1,\ldots,n\} \\ a \le \frac{k-np}{\sqrt{n}} \le b}} \widetilde{b}_{n,p}(k) \cdot \varepsilon_{n,p}(k) \right| \le \underbrace{\overline{\varepsilon}_{n,p}}_{\to 0} \cdot \underbrace{\sum_{\substack{k \in \{0,1,\ldots,n\} \\ a \le \frac{k-np}{\sqrt{n}} \le b}} \widetilde{b}_{n,p}(k)}_{\to \int_a^b \ldots dx < \infty} \overset{n\to\infty}{\longrightarrow} 0$$

Zum Beweis von (4.6.6) sei

$$\Gamma_n := \left\{ \frac{k - np}{\sqrt{n}} \,\middle|\, k = 0, 1, \ldots, n \right\} \subseteq \mathbb{R}.$$

Dann ist Γ_n ein äquidistantes Gitter mit Maschenweite $\Delta = \frac{1}{\sqrt{n}}$, und es gilt

$$\sum_{\substack{k \in \{0,1,\ldots,n\} \\ a \le \frac{k-np}{\sqrt{n}} \le b}} \widetilde{b}_{n,p}(k) = \sum_{\substack{x \in \Gamma_n \\ a \le x \le b}} \frac{1}{\sqrt{2\pi\sigma^2}}^{-\frac{x^2}{2\sigma^2}} \Delta x.$$

Für $n \to \infty$ folgt (4.6.6), da die rechte Seite eine Riemannsummenapproximation des Integrals ist.

□

Der Satz von de Moivre/Laplace besagt, dass die Verteilungen der Zufallsvariablen $\frac{S_n - np}{\sqrt{n}}$ für $n \to \infty$ **schwach** gegen die Normalverteilung $N(0, \sigma^2)$ mit Varianz $\sigma^2 = p(1-p)$ **konvergieren**. Die allgemeine Definition der schwachen Konvergenz einer Folge von Wahrscheinlichkeitsverteilungen wird in Abschnitt 8.3 unten gegeben. Ist Z eine standardnormalverteilte Zufallsvariable, dann gilt:

$$\frac{S_n - np}{\sqrt{n}} \quad \xrightarrow{\mathcal{D}} \quad \sigma Z,$$

bzw.

$$\frac{S_n - E[S_n]}{\sigma(S_n)} \quad = \quad \frac{S_n - np}{\sigma\sqrt{n}} \quad \xrightarrow{\mathcal{D}} \quad Z, \tag{4.6.7}$$

wobei „$\xrightarrow{\mathcal{D}}$“ für schwache Konvergenz der Verteilungen der Zufallsvariablen steht (**Konvergenz in Verteilung**, s.u.).

Bemerkung. (1). Die Aussage (4.6.7) ist ein Spezialfall eines viel allgemeineren zentralen Grenzwertsatzes:

Sind $X_1, X_2, \ldots$ unabhängige, identisch verteilte Zufallsvariablen mit endlicher Varianz, und ist $S_n = X_1 + \ldots + X_n$, dann konvergieren die Verteilungen der standardisierten Summen

$$\frac{S_n - E[S_n]}{\sigma(S_n)}$$

schwach gegen eine Standardnormalverteilung, s.u.
Die Normalverteilung tritt also als universeller Skalierungslimes von Summen unabhängiger Zufallsvariablen auf.

(2). Heuristisch gilt für große n nach (4.6.7)

$$\text{„} \quad S_n \quad \overset{\mathcal{D}}{\approx} \quad np + \sqrt{np(1-p)} \cdot Z, \quad \text{“} \tag{4.6.8}$$

wobei „$\overset{\mathcal{D}}{\approx}$“ dafür steht, dass sich die Verteilungen der Zufallsvariablen einander in einem gewissen Sinn annähern. In diesem Sinne wäre für große n

$$\text{„}\mathrm{Bin}(n,p) \quad \overset{\mathcal{D}}{\approx} \quad N(np, np(1-p))\text{.“}$$

Entsprechende „Approximationen“ werden häufig in Anwendungen benutzt, sollten aber hinterfragt werden, da beim Übergang von (4.6.7) zu (4.6.8) mit dem divergierende Faktor $\sqrt{n}$ multipliziert wird. Die mathematische Präzisierung entsprechender heuristischer Argumentationen erfolgt üblicherweise über den Satz von de Moivre/Laplace.

Beispiel (Faire Münzwürfe). Seien $X_1, X_2, \ldots$ unabhängige Zufallsvariablen mit $P[X_i = 0] = P[X_i = 1] = \frac{1}{2}$ und sei $S_n = X_1 + \ldots + X_n$ (z.B. Häufigkeit von „Zahl“ bei n fairen Münzwürfen). In diesem Fall ist also $p = \frac{1}{2}$ und $\sigma = \sqrt{p(1-p)} = \frac{1}{2}$.

(1). **100 faire Münzwürfe:**

$$P[S_{100} > 60] \quad = \quad P[S_{100} - E[S_{100}] > 10] \quad = \quad P\left[\frac{S_{100} - E[S_{100}]}{\sigma(S_{100})} > \frac{10}{\sigma\sqrt{100}}\right]$$

Da $\frac{S_{100}-E[S_{100}]}{\sigma(S_{100})}$ nach (4.6.7) näherungsweise $N(0,1)$-verteilt ist, und $\frac{10}{\sigma\sqrt{100}} = 2$, folgt

$$P[S_{100} > 60] \quad \approx \quad P[Z > 2] \quad = \quad 1 - \Phi(2) \quad \approx \quad 0.0227 \quad = \quad 2.27\%.$$

(2). **16 faire Münzwürfe**:

$$\begin{aligned} P[S_{16} = 8] &= P[7.5 \leq S_{16} \leq 8.5] = P[|S_{16} - E[S_{16}]| \leq 0.5] \\ &= P\left[\left|\frac{S_{16} - E[S_{16}]}{\sigma(S_{16})}\right| \leq \frac{0.5}{\sigma\sqrt{16}}\right] \end{aligned}$$

Mit $\frac{0.5}{\sigma\sqrt{16}} = \frac{1}{4}$ folgt:

$$P[S_{16} = 8] \quad \approx \quad P[|Z| \leq 1.4] \quad = \quad 0.1974...$$

Der exakte Wert beträgt $P[S_{16} = 8] = 0.1964....$ Bei geschickter Anwendung ist die Normalapproximation oft schon für eine kleine Anzahl von Summanden relativ genau!

Approximative Konfidenzintervalle

Angenommen, wir wollen den Anteil p der Wähler einer Partei durch Befragung von n Wählern schätzen. Seien $X_1, \ldots, X_n$ unter P_p unabhängige und Bernoulli(p)-verteilte Zufallsvariablen, wobei $X_i = 1$ dafür steht, dass der i-te Wähler für die Partei A stimmen wird. Ein nahe liegender Schätzwert für p ist $\overline{X_n} := \frac{S_n}{n}$. Wie viele Stichproben braucht man, damit der tatsächliche Stimmenanteil mit 95% Wahrscheinlichkeit um höchstens $\varepsilon = 1\%$ von Schätzwert abweicht?

Definition. *Sei $\alpha \in (0,1)$. Das zufällige Intervall $[\overline{X_n} - \varepsilon, \overline{X_n} + \varepsilon]$ heißt Konfidenzintervall zum Konfidenzniveau $1 - \alpha$ (bzw. zum Irrtumsniveau α) für den unbekannten Parameter p, falls*

$$P_p[p \notin [\overline{X_n} - \varepsilon, \overline{X_n} + \varepsilon]] \quad \leq \quad \alpha$$

*für **alle** möglichen Parameterwerte $p \in [0,1]$ gilt.*

Im Prinzip lassen sich Konfidenzintervalle aus den Quantilen der zugrundeliegenden Verteilung gewinnen. In der Situation von oben gilt beispielsweise:

$$\begin{aligned} p \in [\overline{X_n} - \varepsilon, \overline{X_n} + \varepsilon] \quad &\iff \quad |\overline{X_n} - p| \leq \varepsilon \iff \overline{X_n} \in [p - \varepsilon, p + \varepsilon] \\ &\iff \quad S_n \in [n(p - \varepsilon), n(p + \varepsilon)] \end{aligned}$$

Diese Bedingung ist für $p \in [0, 1]$ mit Wahrscheinlichkeit $\geq 1 - \alpha$ erfüllt, falls z.B. $n(p - \varepsilon)$ oberhalb des $\frac{\alpha}{2}$-Quantils und $n(p + \varepsilon)$ unterhalb des $(1 - \frac{\alpha}{2})$-Quantils der Binomialverteilung $\mathrm{Bin}(n, p)$ liegt.

Praktikablere Methoden, um in unserem Modell Konfidenzintervalle zu bestimmen, sind zum Beispiel:

Abschätzung mithilfe der Čebyšev-Ungleichung:

$$P_p\left[\left|\frac{S_n}{n} - p\right| \geq \varepsilon\right] \quad \leq \quad \frac{1}{\varepsilon^2} \cdot Var\left(\frac{S_n}{n}\right) \quad = \quad \frac{p(1-p)}{n\varepsilon^2} \quad \leq \quad \frac{1}{4n\varepsilon^2} \quad \overset{!}{\leq} \quad \alpha \qquad \forall\, p \in [0, 1]$$

Dies ist erfüllt für $n \geq \frac{1}{4\varepsilon^2\alpha}$, also im Beispiel für $n \geq 50.000$.

Abschätzung über die exponentielle Ungleichung:

$$P_p\left[\left|\frac{S_n}{n} - p\right| \geq \varepsilon\right] \quad \leq \quad 2 \cdot e^{-2\varepsilon^2 n} \quad \leq \quad \alpha \qquad \forall\, p \in [0, 1],$$

ist erfüllt für $n \geq \frac{1}{2\varepsilon^2} \log(\frac{2}{\alpha})$, also im Beispiel für $n \geq 18445$.

Die exponentielle Abschätzung ist genauer - sie zeigt, dass bereits weniger als 20.000 Stichproben genügen. Können wir mit noch weniger Stichproben auskommen ? Dazu berechnen wir die Wahrscheinlichkeit, dass der Parameter im Intervall liegt, näherungsweise mithilfe des zentralen Grenzwertsatzes:

Approximative Berechnung mithilfe der Normalapproximation:

$$\begin{aligned} P_p\left[\left|\frac{S_n}{n} - p\right| \leq \varepsilon\right] \quad &= \quad P_p\left[\left|\frac{S_n - np}{\sqrt{np(1-p)}}\right| \leq \frac{n\varepsilon}{\sqrt{np(1-p)}}\right] \\ &\approx \quad N(0,1)\left(-\frac{\sqrt{n}\varepsilon}{\sqrt{p(1-p)}}\,,\, \frac{\sqrt{n}\varepsilon}{\sqrt{p(1-p)}}\right) \\ &= \quad 2\left(\Phi\left(\frac{\sqrt{n}\varepsilon}{\sqrt{p(1-p)}}\right) - \frac{1}{2}\right) \\ &\overset{p(1-p)\leq\frac{1}{4}}{\geq} \quad 2\Phi(2\sqrt{n}\varepsilon) - 1 \quad \geq \quad 1 - \alpha \qquad \forall\, p \in [0, 1], \end{aligned}$$

falls

$$n \quad \geq \quad \left(\frac{1}{2\varepsilon} \cdot \Phi^{-1} \left(1 - \frac{\alpha}{2} \right) \right)^2 .$$

Im Beispiel gilt

$$\Phi^{-1} \left(1 - \frac{\alpha}{2} \right) \quad \approx \quad 1.96$$

und die Bedingung ist für $n \geq 9604$ erfüllt. Also sollten bereits ca. 10.000 Stichproben ausreichen! *Exakte* (also ohne Verwendung einer Näherung hergeleitete) Konfidenzintervalle sind in vielen Fällen zu konservativ. In Anwendungen werden daher meistens *approximative* Konfidenzintervalle angegeben, die mithilfe einer Normalapproximation hergeleitet wurden. Dabei ist aber folgendes zu beachten:

Warnung: Mithilfe der Normalapproximation hergeleitete approximative Konfidenzintervalle erfüllen die Niveaubedingung im Allgemeinen nicht (bzw. nur näherungsweise). Da die Qualität der Normalapproximation für $p \to 0$ bzw. $p \to 1$ degeneriert, ist die Niveaubedingung im Allgemeinen selbst für $n \to \infty$ nicht erfüllt. Beispielsweise beträgt das Niveau von approximativen 99% Konfidenzintervallen asymptotisch tatsächlich nur 96.8%!

Kapitel 5

Unabhängigkeit und Produktmodelle

5.1 Unabhängigkeit in allgemeinen Modellen

Unabhängigkeit von Ereignissen

In Abschnitt 2.3 haben wir einen Unabhängigkeitsbegriff für Ereignisse eingeführt: Eine Kollektion $A_i, i \in I$, von Ereignissen aus derselben σ-Algebra $\mathcal{A}$ heißt **unabhängig** bzgl. einer Wahrscheinlichkeitsverteilung P, falls

$$P[A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_n}] \quad = \quad \prod_{k=1}^{n} P[A_{i_k}] \tag{5.1.1}$$

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

Beispiel. Ein Ereignis A ist genau dann unabhängig von sich selbst, wenn $P[A] = P[A \cap A] = P[A]^2$ gilt, also wenn die Wahrscheinlichkeit von A gleich 0 oder 1 ist. Solche Ereignisse nennt man auch deterministisch.

Wir wollen den obigen Unabhängigkeitsbegriff nun auf Ereignissysteme erweitern.

Definition. *Eine Kollektion $\mathcal{A}_i$ $(i \in I)$ von Mengensystemen $\mathcal{A}_i \subseteq \mathcal{A}$ heißt **unabhängig (bzgl. P)**, falls jede Kollektion A_i $(i \in I)$ von Ereignissen $A_i \in \mathcal{A}_i$ unabhängig ist, d.h.*

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

für alle $n \in \mathbb{N}, i_1, \ldots, i_n \in I$ paarweise verschieden, und $A_{i_k} \in \mathcal{A}_{i_k}$ $(1 \leq k \leq n)$.

Sind zum Beispiel A und B unabhängige Ereignisse, dann sind $\sigma(A) = \{\emptyset, A, A^C, \Omega\}$ und $\sigma(B) = \{\emptyset, B, B^C, \Omega\}$ unabhängige Mengensysteme. Allgemeiner:

Satz 5.1. *Seien $\mathcal{A}_i$ $(i \in I)$ unabhängige Mengensysteme. Jedes $\mathcal{A}_i$ sei durchschnittsstabil. Dann gilt:*

(1). Die σ-Algebren $\sigma(\mathcal{A}_i)$ $(i \in I)$ sind unabhängige Mengensysteme.

(2). Ist $I = \bigcup_{k \in K} I_k$ eine disjunkte Zerlegung von I, dann sind auch die σ-Algebren $\sigma(\bigcup_{i \in I_k} \mathcal{A}_i)$ $(k \in K)$ unabhängige Mengensysteme.

Beispiel. Sind $A_1, \ldots, A_n$ unabhängige Ereignisse, dann sind die Mengensysteme

$$\mathcal{A}_1 \quad = \quad \{A_1\}, \quad \ldots \quad , \mathcal{A}_n \quad = \quad \{A_n\}$$

unabhängig und durchschnittsstabil, also sind auch die σ-Algebren

$$\sigma(\mathcal{A}_i) \quad = \quad \{\emptyset, A_i, A_i^C, \Omega\} \qquad (i = 1, \ldots, n)$$

unabhängige Mengensysteme, d.h es gilt

$$P[B_1 \cap \ldots \cap B_n] \quad = \quad \prod_{i=1}^{n} P[B_i] \qquad \forall B_i \in \{\emptyset, A_i, A_i^C, \Omega\}.$$

Dies kann man auch direkt beweisen, siehe Lemma 2.5 oben.

Ein Beispiel zum zweiten Teil der Aussage von Satz 5.1 werden wir im Anschluss an den Beweis des Satzes betrachten.

Beweis. (1). Seien $i_1, \ldots, i_n \in I$ $(n \in \mathbb{N})$ paarweise verschieden. Wir müssen zeigen, dass

$$P[B_{i_1} \cap \ldots \cap B_{i_n}] \quad = \quad P[B_{i_1}] \cdot \ldots \cdot P[B_{i_n}] \tag{5.1.2}$$

für alle $B_{i_1} \in \sigma(\mathcal{A}_{i_1}), \ldots, B_{i_n} \in \sigma(\mathcal{A}_{i_n})$ gilt. Dazu verfahren wir schrittweise:

(a) Die Aussage (5.1.2) gilt nach Voraussetzung für $B_{i_1} \in \mathcal{A}_{i_1}, \ldots, B_{i_n} \in \mathcal{A}_{i_n}$.

(b) Für $B_{i_2} \in \mathcal{A}_{i_2}, \ldots, B_{i_n} \in \mathcal{A}_{i_n}$ betrachten wir das Mengensystem $\mathcal{D}$ aller $B_{i_1} \in \mathcal{A}$, für die (5.1.2) gilt. $\mathcal{D}$ ist ein Dynkinsystem, das $\mathcal{A}_{i_1}$ nach (a) enthält. Da $\mathcal{A}_{i_1}$ durchschnittsstabil ist, folgt

$$\mathcal{D} \quad \supseteq \quad \mathcal{D}(\mathcal{A}_{i_1}) \qquad = \qquad \sigma(\mathcal{A}_{i_1}).$$

Also gilt (5.1.2) für alle $B_{i_1} \in \sigma(\mathcal{A}_{i_1})$.

(c) Für $B_{i_1} \in \sigma(\mathcal{A}_{i_1})$ und $B_{i_3} \in \sigma(\mathcal{A}_{i_3}), \ldots, B_{i_n} \in \sigma(\mathcal{A}_{i_n})$ betrachten wir nun das Mengensystem aller $B_{i_2} \in \mathcal{A}$, für die (5.1.2) gilt. Wiederum ist $\mathcal{D}$ ein Dynkinsystem, das $\mathcal{A}_{i_2}$ nach (b) enthält. Wie im letzten Schritt folgt daher

$$\mathcal{D} \quad \supseteq \quad \mathcal{D}(\mathcal{A}_{i_2}) \quad = \quad \sigma(\mathcal{A}_{i_2}),$$

d.h. (5.1.2) ist für alle $B_{i_2} \in \sigma(\mathcal{A}_{i_2})$ erfüllt.
Anschließend verfahren wir auf entsprechende Weise weiter. Nach n-facher Anwendung eines analogen Arguments folgt die Behauptung.

(2). Für $k \in K$ gilt: $\sigma(\bigcup_{i \in I_k} \mathcal{A}_i) = \sigma(\mathcal{C}_k)$ mit

$$\mathcal{C}_k \quad := \quad \{B_{i_1} \cap \ldots \cap B_{i_n} | n \in \mathbb{N}, i_1, \ldots, i_n \in I_k \text{ paarw. verschieden }, B_{i_j} \in \mathcal{A}_{i_j}\}.$$

Die Mengensysteme $\mathcal{C}_k$, $k \in K$, sind durchschnittsstabil und unabhängig, da jede Kollektion von Ereignissen $B_i \in \mathcal{A}_i$, $i \in I$, nach Voraussetzung unabhängig ist. Also sind nach Teil (1) der Aussage auch die σ-Algebren $\sigma(\mathcal{C}_k)$, $k \in K$, unabhängig.

□

Beispiel (Affe tippt Shakespeare). Wir betrachten unabhängige 0-1-Experimente mit Erfolgswahrscheinlichkeit p. Sei $X_i(\omega) \in \{0, 1\}$ der Ausgang des i-ten Experiments. Für ein binäres Wort $(a_1, \ldots, a_n) \in \{0, 1\}^n$, $n \in \mathbb{N}$, gilt:

$$P[X_1 = a_1, \ldots, X_n = a_n] \quad = \quad P\left[\bigcap_{i=1}^{n} \{X_i = a_i\}\right] \quad \overset{\text{unabh.}}{=} \quad p^k \cdot (1-p)^{n-k},$$

wobei $k = a_1 + \ldots + a_n$ die Anzahl der Einsen in dem Wort ist. Wir zeigen nun:

Behauptung: $P[\text{Wort kommt unendlich oft in der Folge } X_1, X_2, \ldots \text{ vor}] = 1$, falls $p \notin \{0, 1\}$.
Zum Beweis bemerken wir, dass die Ereignisse

$$E_m \quad = \quad \{X_{mn+1} = a_1, X_{mn+2} = a_2, \ldots, X_{mn+n} = a_n\}, \qquad m \in \mathbb{N},$$
„Text steht im m-ten Block“

unabhängig sind. Nach Satz 5.1 sind nämlich die σ-Algebren

$$\sigma(\{\{X_{mn+1} = 1\}, \{X_{mn+2} = 1\}, \ldots, \{X_{mn+n} = 1\}\}), \qquad m \in \mathbb{N},$$

unabhängig, also auch die darin enthaltenen Ereignisse E_m. Für $p \neq 0$ gilt:

$$P[E_m] = p^k \cdot (1-p)^{n-k} > 0,$$

also

$$\sum_{m=1}^{\infty} P[E_m] \quad = \quad \infty.$$

Damit folgt nach Borel-Cantelli:

$$1 \quad = \quad P[E_m \text{ unendlich oft}] \quad \leq \quad P[\text{Wort kommt unendlich oft vor}].$$

□

Unabhängigkeit von Zufallsvariablen

Wir betrachten nun Zufallsvariablen mit Werten in einem messbaren Raum $(S, \mathcal{S})$.

Definition. *Seien $X, X_i : \Omega \to S$, $i \in I$, Abbildungen.*

(1). Das Mengensystem

$$\sigma(X) \quad := \quad \{X^{-1}(B) | B \in \mathcal{S}\} \quad \subseteq \quad \mathcal{P}(\Omega)$$

*heißt **die von** X **erzeugte** σ**-Algebra** auf Ω.*

(2). Allgemeiner heißt

$$\sigma(X_i | i \in I) \quad := \quad \sigma\left(\bigcup_{i \in I} \sigma(X_i)\right) \quad = \quad \sigma(\{X_i^{-1}(B) | B \in \mathcal{S}, i \in I\})$$

die von den Abbildungen X_i, $i \in I$, erzeugte σ-Algebra.

Bemerkung. (1). Man prüft leicht nach, dass $\sigma(X)$ tatsächlich eine σ-Algebra ist.

(2). Eine Abbildung $X : \Omega \to S$ ist messbar bzgl. $\mathcal{A}/\mathcal{S}$ genau dann, wenn $\sigma(X) \subseteq \mathcal{A}$ gilt. Somit ist $\sigma(X)$ die **kleinste** σ-Algebra auf Ω, bzgl. der X messbar ist.

(3). Entsprechend ist $\sigma(X_i, i \in I)$ die kleinste σ-Algebra auf Ω, bzgl. der alle Abbildungen X_i messbar sind.

Beispiel (Produkt-σ-Algebra). Sei $\Omega = \{0,1\}^{\mathbb{N}} = \{\omega = (x_1, x_2, \ldots) | x_i \in \{0,1\}\}$, oder ein allgemeiner Produktraum, und sei $X_i(\omega) = x_i$ die Projektion auf die i-te Komponente. Dann ist die Produkt-σ-Algebra $\mathcal{A}$ auf Ω gerade die von den Abbildungen X_i erzeugte σ-Algebra:

$$\begin{aligned} \mathcal{A} &= \sigma(\{X_1 = a_1, \ldots, X_n = a_n\} | n \in \mathbb{N}, a_1, \ldots, a_n \in \{0,1\}) \\ &= \sigma(\{X_i = 1\} | i \in \mathbb{N}) \\ &= \sigma(X_1, X_2, \ldots). \end{aligned}$$

Messbare Abbildungen auf $(\Omega, \mathcal{A})$ sind z.B.

$$S_n(\omega) \quad = \quad X_1(\omega) + \ldots + X_n(\omega),$$

$$\overline{L}(\omega) \quad = \quad \limsup_{n\to\infty} \frac{1}{n} S_n(\omega), \qquad \underline{L}(\omega) \quad = \quad \liminf_{n\to\infty} \frac{1}{n} S_n(\omega), \quad \text{etc.}$$

Wir können nun einen Unabhängigkeitsbegriff für allgemeine Zufallsvariablen einführen, der kompatibel mit dem oben definierten Unabhängigkeitsbegriff für Mengensysteme ist.

Definition. *Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum.*

*(1). Eine endliche Kollektion $X_1, \ldots, X_n : \Omega \to S$ von Zufallsvariablen auf $(\Omega, \mathcal{A}, P)$ heißt **unabhängig**, falls*

$$P[X_1 \in B_1, \ldots, X_n \in B_n] \quad = \quad \prod_{i=1}^{n} P[X_i \in B_i] \qquad \forall B_i \in \mathcal{S} \quad (1 \le i \le n). \quad (5.1.3)$$

*(2). Eine beliebige Kollektion $X_i, i \in I$, von Zufallsvariablen auf $(\Omega, \mathcal{A}, P)$ heißt **unabhängig**, falls jede endliche Teilkollektion $X_{i_1}, \ldots, X_{i_n}$ ($i_1, \ldots, i_n \in I$ paarweise verschieden) unabhängig ist.*

Bemerkung. (1). Die Definition ist **konsistent**: Jede endliche Teilkollektion einer unabhängigen endlichen Kollektion von Zufallsvariablen ist wieder unabhängig im Sinne von (5.1.3).

(2). Die Zufallsvariablen $X_i, i \in I$, sind genau dann unabhängig, wenn die σ-Algebren

$$\sigma(X_i) \quad = \quad \{\{X_i \in B\} | B \in \mathcal{B}(\mathcal{S})\}, \qquad i \in I,$$

unabhängige Mengensysteme sind.

Sei $(\widetilde{S}, \widetilde{\mathcal{S}})$ ein weiterer messbarer Raum. Eine sehr wichtige Konsequenz von Bemerkung (2) ist:

Satz 5.2 (Funktionen von unabhängigen Zufallsvariablen sind unabhängig). *Sind $X_i : \Omega \to S$, $i \in I$, unabhängige Zufallsvariablen auf $(\Omega, \mathcal{A}, P)$, und sind $h_i : S \to \widetilde{S}$ messbare Abbildungen, dann sind auch die Zufallsvariablen $Y_i := h_i(X_i)$, $i \in I$, unabhängig bzgl. P.*

Beweis.

$$\sigma(Y_i) \quad = \quad \{ \underbrace{Y_i^{-1}(B)}_{X_i^{-1}(h_i^{-1}(B))} \quad | \quad B \in \widetilde{\mathcal{S}} \}.$$

Da die σ-Algebren $\sigma(X_i)$, $i \in I$, unabhängig sind, sind auch $\sigma(Y_i)$, $i \in I$, unabhängige Mengensysteme. □

Aufgrund von Satz 5.1 kann man allgemeiner eine Kollektion X_i, $i \in I$, von unabhängigen Zufallsvariablen in disjunkte Gruppen X_i, $i \in I_k$, $I = \dot{\bigcup_k} I_k$, einteilen, und messbare Funktionen

$$Y_k \quad = \quad h_k(X_i | i \in I_k), \qquad k \in K$$

von den Zufallsvariablen der verschiedenen Gruppen betrachten. Auch die Y_k sind dann wieder unabhängige Zufallsvariablen.

Für unabhängige reellwertige Zufallsvariablen X_i $(i \in I)$ gilt insbesondere

$$P[X_{i-1} \leq c_1, \ldots, X_{i_n} \leq c_n] \quad = \quad \prod_{k=1}^{n} P[X_{i_k} \leq c_k] \tag{5.1.4}$$

für alle $n \in \mathbb{N}, i_1, \ldots, i_n \in I$ paarweise verschieden, und $c_i \in \mathbb{R}$.
Tatsächlich werden wir im nächsten Abschnitt zeigen, dass Bedingung (5.1.4) äquivalent zur Unabhängigkeit der X_i ist. Als erste Anwendung betrachten wir Extrema von unabhängigen exponentialverteilten Zufallsvariablen.

Beispiel (Maxima von exponentialverteilten Zufallsvariablen). Seien $T_1, T_2, \ldots$ unabhängige $\mathrm{Exp}(1)$-verteilte Zufallsvariablen. Wir wollen uns überlegen, wie sich die Extremwerte (Rekorde)

$$M_n \quad = \quad \max\{T_1, \ldots, T_n\}$$

asymptotisch für $n \to \infty$ verhalten. Dazu gehen wir in mehreren Schritten vor:

(1). Wir zeigen zunächst mithilfe des Borel-Cantelli-Lemmas:

$$\limsup_{n\to\infty} \frac{T_n}{\log n} = 1 \qquad P\text{-fast sicher.} \tag{5.1.5}$$

Zum Beweis berechnen wir für $c \in \mathbb{R}$:

$$\begin{aligned} P\left[\frac{T_n}{\log n} \geq c\right] &= P[T_n \geq c \cdot \log n] \\ &= e^{-c \log n} = n^{-c}. \end{aligned}$$

Für $c > 1$ gilt $\sum_{n=1}^{\infty} n^{-c} < \infty$. Nach dem 1. Borel-Cantelli-Lemma folgt daher

$$P\left[\limsup_{n\to\infty} \frac{T_n}{\log n} > c\right] \quad \leq \quad P\left[\frac{T_n}{\log n} \geq c \quad \text{unendlich oft}\right] \quad = \quad 0.$$

Für $c \searrow 1$ erhalten wir dann wegen der monotonen Stetigkeit von P:

$$P\left[\limsup_{n\to\infty} \frac{T_n}{\log n} > 1\right] \quad = \quad \lim_{c \searrow 1} P\left[\limsup_{n\to\infty} \frac{T_n}{\log n} > c\right] \quad = \quad 0. \tag{5.1.6}$$

Für $c < 1$ gilt $\sum_{n=1}^{\infty} n^{-c} = \infty$. Da die Ereignisse $\{T_n \geq c \log n\}, n \in \mathbb{N}$, unabhängig sind, folgt nach dem 2. Borel-Cantelli Lemma:

$$P\left[\limsup_{n\to\infty} \frac{T_n}{\log n} \geq c\right] \quad \geq \quad P\left[\frac{T_n}{\log n} \geq c \quad \text{unendlich oft}\right] \quad = \quad 1.$$

Für $c \nearrow 1$ erhalten wir mithilfe der monotonen Stetigkeit:

$$P\left[\limsup_{n\to\infty} \frac{T_n}{\log n} \geq 1\right] \quad = \quad \lim_{c\nearrow 1} P\left[\limsup_{n\to\infty} \frac{T_n}{\log n} \geq c\right] \quad = \quad 1 \qquad (5.1.7)$$

Aus (5.1.6) und (5.1.7) folgt die Behauptung (5.1.5).

(2). Als nächstes folgern wir:

$$M_n \sim \log n, \quad \text{d.h. } \lim_{n\to\infty} \frac{M_n}{\log n} = 1 \qquad P\text{-f.s.} \qquad (5.1.8)$$

Zum Beweis zeigen wir:

(a) $\limsup_{n\to\infty} \frac{M_n}{\log n} \leq 1 \quad P$-f.s., und

(b) $\liminf_{n\to\infty} \frac{M_n}{\log n} \geq 1 \quad P$-f.s.

Aussage (a) folgt aus (1), denn für $c \in \mathbb{R}$ gilt:

$$\begin{aligned}
& \limsup_{n\to\infty} \frac{M_n}{\log n} \quad > \quad c \\
\Rightarrow \quad & \max\{T_1, \ldots, T_n\} = M_n \quad > \quad c \cdot \log n \quad \text{unendlich oft} \\
\Rightarrow \quad & T_{k(n)} \quad > \quad c \cdot \log n \quad \text{für } k(n) \leq n \text{ für } \infty \text{ viele } n \\
\Rightarrow \quad & T_k \quad > \quad c \cdot \log k \quad \text{unendlich oft} \\
\Rightarrow \quad & \limsup \frac{T_k}{\log k} \geq c
\end{aligned}$$

Nach (1) hat das letztere Ereignis für $c > 1$ Wahrscheinlichkeit 0, also gilt wegen der monotonen Stetigkeit von P:

$$P\left[\limsup_{n\to\infty} \frac{M_n}{\log n} > 1\right] \quad = \quad \lim_{c\searrow 1} P\left[\limsup_{n\to\infty} \frac{M_n}{\log n} > c\right] \quad = \quad 0.$$

Zum Beweis von (b) genügt es wegen der monotonen Stetigkeit zu zeigen, dass für $c < 1$

$$P\left[\frac{M_n}{\log n} > c \quad \text{schließlich}\right] \quad = \quad P\left[\frac{M_n}{\log n} \leq c \quad \text{nur endlich oft}\right] \quad = \quad 1$$

gilt. Nach Borel-Cantelli I ist dies der Fall, wenn

$$\sum_{n\in\mathbb{N}} P\left[\frac{M_n}{\log n} \leq c\right] \quad < \quad \infty \qquad (5.1.9)$$

gilt. Für $c \in \mathbb{R}$ gilt aber wegen der Unabhängigkeit der T_i

$$\begin{aligned} P\left[\frac{M_n}{\log n} \le c\right] &= P\left[T_i \le c \cdot \log n \qquad \forall\, 1 \le i \le n\right] \\ &= P\left[T_1 \le c \cdot \log n\right]^n \quad = \quad (1 - e^{-c \log n})^n \\ &= (1 - n^{-c})^n \quad \le \quad e^{-n \cdot n^{-c}} \quad = \quad e^{-n^{1-c}}, \end{aligned}$$

und diese Folge ist für $c < 1$ summierbar. Also gilt (5.1.9) für alle $c < 1$, und damit (b).

(3). Abschließend untersuchen wir die Fluktuationen der Extremwerte M_n um $\log n$ noch genauer. Wir zeigen, dass die Zufallsvariable $M_n - \log n$ in Verteilung konvergiert:

$$P[M_n - \log n \le c] \overset{n\to\infty}{\longrightarrow} e^{-e^{-c}} \qquad \text{für alle } c \in \mathbb{R}. \tag{5.1.10}$$

Beweis. Wegen

$$\begin{aligned} P[M_n \le c] &= P[T_i \le c \quad \forall i = 1, \ldots, n] \\ &\overset{\text{i.i.d.}}{=} P[T_1 \le c]^n \\ &= (1 - e^{-c})^n \qquad \text{für alle } c \in \mathbb{R} \end{aligned}$$

folgt

$$P[M_n - \log n \le c] \quad = \quad P[M_n \le c + \log n] \quad = \quad (1 - \frac{1}{n} \cdot e^{-c})^n \quad \overset{n\to\infty}{\longrightarrow} e^{-e^{-c}}$$

□

Aussage (5.1.10) besagt, dass $M_n - \log n$ in Verteilung gegen eine Gumbel-verteilte Zufallsvariable X, d.h. eine Zufallsvariable mit Verteilungsfunktion $F_X(c) = e^{-e^{-c}}$ konvergiert. Für große n gilt also näherungsweise

$$M_n \quad \overset{\mathcal{D}}{\approx} \quad \log n + X, \qquad X \sim \text{Gumbel},$$

wobei $\log n$ die Asymptotik und X die Fluktuationen beschreibt.

Konfidenzintervalle für Quantile

Sei $(x_1, \ldots, x_n)$ eine n-elementige Stichprobe von einer unbekannten Wahrscheinlichkeitsverteilung μ auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Wir nehmen an, dass $x_1, \ldots, x_n$ Realisierungen von unabhängigen Zufallsvariablen mit stetiger Verteilung sind:

Annahme: $X_1, \ldots, X_n$ unabhängig unter P_μ mit stetiger Verteilung μ.

Wir wollen nun die Quantile (z.B. den Median) der zugrundeliegenden Verteilung auf der Basis der Stichprobe schätzen. Eine Funktion $T(X_1, \ldots, X_n), T : \mathbb{R}^n \to \mathbb{R}$ messbar, nennt man in diesem Zusammenhang auch ein *Statistik* der Stichprobe $(X_1, \ldots, X_n)$. Eine Statistik, deren Wert als Schätzwert für eine Kenngröße $q(\mu)$ der unbekannten Verteilung verwendet wird, nennt man auch einen *(Punkt-) Schätzer* für q. Nahe liegende Schätzer für die Quantile von μ sind die entsprechenden Stichprobenquantile. Unser Ziel ist es nun, *Konfidenzintervalle* für die Quantile anzugeben, d.h. von den Werten $X_1, \ldots, X_n$ abhängende Intervalle, in denen die Quantile *unabhängig von der tatsächlichen Verteilung* mit hoher Wahrscheinlichkeit enthalten sind. Seien dazu

$$X_{(1)} \quad \leq \quad X_{(2)} \quad \leq \quad \ldots \quad \leq \quad X_{(n)}$$

die der Größe nach geordneten Werte $X_1, \ldots, X_n$ – diese nennt man auch **Ordnungsstatistiken** der Stichprobe. Die Verteilung der Ordnungsstatistiken können wir explizit berechnen:

Satz 5.3 (Verteilung der Ordnungsstatistiken). *Ist μ eine absolutstetige Wahrscheinlichkeitsverteilung mit Verteilungsfunktion F, dann hat $X_{(k)}$ die Verteilungsfunktion*

$$\begin{aligned} F_{(k)}(c) &= \mathrm{Bin}(n, F(c))[\{k, k+1, \ldots, n\}] \\ &= \sum_{j=k}^{n} \binom{n}{j} F(c)^j \cdot (1 - F(c))^{n-j}. \end{aligned} \tag{5.1.11}$$

Beweis. Da die Ereignisse $\{X_i \leq c\}, 1 \leq i \leq n$, unabhängig sind mit Wahrscheinlichkeit $F(c)$, gilt

$$\begin{aligned} F_{(k)}(c) = P_\mu[X_{(k)} \leq c] &= P_\mu[X_i \leq c \quad \text{für mindestens } k \text{ verschiedene } i \in \{1, \ldots, n\}] \\ &= \mathrm{Bin}(n, F(c))[\{k, k+1, \ldots, n\}] \\ &= \sum_{j=k}^{n} \binom{n}{j} F(c)^j \cdot (1 - F(c))^{n-j}. \end{aligned}$$

□

Nach Satz 5.3 ist die Wahrscheinlichkeit, dass der Wert von $X_{(k)}$ unterhalb eines u-Quantils der zugrundeliegenden Verteilung μ liegt, für alle stetigen Verteilungen gleich! Damit folgt unmittelbar:

Korollar 5.4 (Ordnungsintervalle). *Sei $u \in (0, 1)$ und $0 \leq k < l \leq n$. Dann ist das zufällige Intervall $(X_{(k)}, X_{(l)})$ ein **Konfidenzintervall für das u-Quantil** der zugrundeliegenden Verteilung μ **zum Konfidenzniveau***

$$\beta \quad := \quad \mathrm{Bin}(n, k)[\{k, k+1, \ldots, l-1\}],$$

d.h. für jede absolutstetige Wahrscheinlichkeitsverteilung μ auf $\mathbb{R}$, und für jedes u-Quantil $q_u(\mu)$ gilt:

$$P_\mu[X_{(k)} < q_u(\mu < X_{(l)})] \quad \geq \quad \beta.$$

Beweis. Da die Verteilungen stetig sind, gilt $F_\mu(q_u(\mu)) = u$ für jedes u-Quantil, und damit nach Satz 5.3:

$$\begin{aligned} P_\mu[X_{(k)} < q_u(\mu) < X(l)] &= \mathrm{Bin}(n,u)[\{k, k+1, \ldots, n\}] - \mathrm{Bin}(n,u)[\{l, l+1, \ldots, n\}] \\ &= \mathrm{Bin}(n,u)[\{k, k+1, \ldots, l-1\}]. \end{aligned}$$

□

Für große n kann man die Quantile der Binomialverteilung näherungsweise mithilfe der Normalapproximation berechnen, und erhält daraus entsprechende Konfidenzintervalle für die Quantile von stetigen Verteilungen. Bemerkenswert ist, dass diese Konfidenzintervalle nicht nur für Verteilungen aus einer bestimmten Familie (z.B. der Familie der Normalverteilungen) gelten, sondern für *alle* stetigen Wahrscheinlichkeitsverteilungen auf $\mathbb{R}$ (*nichtparametrisches Modell*).

5.2 Gemeinsame Verteilungen und endliche Produktmodelle

Um Aussagen über den Zusammenhang mehrerer Zufallsvariablen $X_1, \ldots, X_n$ zu treffen, benötigen wir Kenntnisse über deren gemeinsame Verteilung, d.h. über die Verteilung des Zufallsvektors $X = (X_1, \ldots, X_n)$. Diese ist eine Wahrscheinlichkeitsverteilung auf dem Produkt der Wertebereiche der einzelnen Zufallsvariablen.

Wahrscheinlichkeitsverteilungen auf endlichen Produkträumen

Seien $(S_i, \mathcal{S}_i)$, $1 \leq i \leq n$, messbare Räume. Die Produkt-σ-Algebra $\mathcal{S}_1 \otimes \ldots \otimes \mathcal{S}_n$ auf $S_1 \times \ldots \times S_n$ wird von den endlichen Produkten von Mengen aus den σ-Algebren $\mathcal{S}_i$ erzeugt:

$$\mathcal{S}_1 \otimes \ldots \otimes \mathcal{S}_n \quad = \quad \sigma(\{B_1 \times \ldots \times B_n | B_i \in \mathcal{S}_i \quad \forall\, 1 \leq i \leq n\}).$$

Bezeichnen wir mit $\pi_i : S_1 \times \ldots \times S_n \to S_i$, $\pi_i(x_1, \ldots, x_n) := x_i$, die kanonische Projektion auf die i-te Komponente, so gilt

$$\mathcal{S}_1 \otimes \ldots \otimes \mathcal{S}_n \quad = \quad \sigma(\pi_1, \ldots, \pi_n).$$

Beispiel. Für die Borelsche σ-Algebra auf $\mathbb{R}^n$ gilt:

$$\mathcal{B}(\mathbb{R}^n) \quad = \quad \underbrace{\mathcal{B}(\mathbb{R}) \otimes \ldots \otimes \mathcal{B}(\mathbb{R})}_{n \text{ mal}} \quad = \quad \bigotimes_{i=1}^{n} \mathcal{B}(\mathbb{R}),$$

denn $\mathcal{B}(\mathbb{R}^n)$ wird zum Beispiel von den offenen Quadern, also Produkten von offenen Intervallen, erzeugt. Ein anderes Erzeugendensystem von $\mathcal{B}(\mathbb{R}^n)$ bilden die Produktmengen

$$(-\infty, c_1] \times (-\infty, c_2] \times \ldots \times (-\infty, c_n], \qquad c_1, \ldots, c_n \in \mathbb{R}. \tag{5.2.1}$$

Ist μ eine Wahrscheinlichkeitsverteilung auf $(S_1 \times \ldots \times S_n, \mathcal{S}_1 \otimes \ldots \otimes \mathcal{S}_n)$, dann heißen die Wahrscheinlichkeitsverteilungen

$$\mu_{\pi_i} \quad := \quad \mu \circ \pi_i^{-1}, \qquad 1 \leq i \leq n,$$

auf S_i **(eindimensionale) Randverteilungen (marginals)** von μ. Wir werden in Kapitel 9 allgemeine Wahrscheinlichkeitsverteilungen auf Produkträumen konstruieren und systematisch untersuchen. Im Moment beschränken wir uns meist auf eine spezielle Klasse von solchen Verteilungen: die endlichen Produktmodelle.

Definition (Endliches Produktmaß). *Seien $(S_i, \mathcal{S}_i, \mu_i)$ Wahrscheinlichkeitsräume, $1 \leq i \leq n$. Eine Wahrscheinlichkeitsverteilung μ auf $(S_1 \times \ldots \times S_n, \mathcal{S}_1 \otimes \ldots \otimes \mathcal{S}_n)$ heißt Produkt der μ_i, falls*

$$\mu[B_1 \times \ldots \times B_n] \quad = \quad \prod_{i=1}^{n} \mu_i[B_i] \qquad \forall\, B_i \in \mathcal{S}_i, 1 \leq i \leq n, \tag{5.2.2}$$

gilt.

Bemerkung. Das Produktmaß μ ist durch (5.2.2) *eindeutig* festgelegt, denn die Produktmengen bilden einen durchschnittsstabilen Erzeuger der σ-Algebra $\mathcal{S}_1 \otimes \ldots \otimes \mathcal{S}_n$. Die *Existenz* von Produktmaßen folgt aus dem Satz von Fubini, den wir in Abschnitt 9.1 beweisen. Für Wahrscheinlichkeitsverteilungen auf $\mathbb{R}$ zeigen wir die Existenz von Produktmaßen im nächsten Abschnitt.

Das nach der Bemerkung eindeutige Produktmaß der Wahrscheinlichkeitsverteilungen $\mu_1, \ldots, \mu_n$ bezeichnen wir mit $\mu_1 \otimes \ldots \otimes \mu_n$. Die eindimensionalen Randverteilungen eines Produktmaßes sind gerade die Faktoren μ_i.

Lemma 5.5. *Unter $\mu = \mu_1 \otimes \ldots \otimes \mu_n$ sind die Projektionen*

$$\pi_i : S_1 \times \ldots \times S_n \longrightarrow S_i, \quad \pi_i(x_1, \ldots, x_n) = x_i, \qquad 1 \leq i \leq n,$$

unabhängig mit Verteilung μ_i.

Beweis. Für $B_i \in \mathcal{S}_i, 1 \leq i \leq n$, gilt:

$$\begin{aligned}\mu[\pi_i \in B_i] &= \mu[S_1 \times \ldots \times S_{i-1} \times B_i \times S_{i+1} \times \ldots \times S_n] \\ &= \mu_i[B_i] \cdot \prod_{j \neq i} \underbrace{\mu_j[S_j]}_{=1} \quad = \quad \mu_i[B_i],\end{aligned}$$

und

$$\mu[\pi_1 \in B_1, \ldots, \pi_n \in B_n] \;=\; \mu[B_1 \times \ldots \times B_n] \;=\; \prod_{i=1}^n \mu_i[B_i] \;=\; \prod_{i=1}^n \mu_i[\pi_i \in B_i].$$

□

Sind die Mengen $S_1, \ldots, S_n$ abzählbar, dann gilt $\mu = \mu_1 \otimes \ldots \otimes \mu_n$ genau dann, wenn die Massenfunktion von μ das Produkt der einzelnen Massenfunktionen ist, d.h.

$$\mu(x_1, \ldots, x_n) \quad = \quad \prod_{i=1}^n \mu_i(x_i) \qquad \text{für alle } \; x_i \in S_i, 1 \leq i \leq n.$$

Im Fall $S_1 = \ldots = S_n = \mathbb{R}$ mit Borelscher σ-Algebra bilden die Mengen aus (5.2.1) einen durchschnittsstabilen Erzeuger der Produkt-σ-Algebra $\mathcal{B}(\mathbb{R}^n)$. Also ist $\mu = \mu_1 \otimes \ldots \otimes \mu_n$ genau dann, wenn

$$\mu[(-\infty, c_1] \times \ldots \times (-\infty, c_n]] \quad = \quad \prod_{i=1}^n \mu_i[(-\infty, c_i]] \qquad \text{für alle } \; c_1, \ldots, c_n \in \mathbb{R}$$

gilt. Die linke Seite ist die Verteilungsfunktion $F_\mu(c_1, \ldots, c_n)$ der multivariaten Verteilung μ, die rechte Seite das Produkt der Verteilungsfunktionen der μ_i.

Beispiel (Gleichverteilung auf n-dimensionalem Quader). Ist $\mu_i = \mathcal{U}_{(a_i,b_i)}$ die Gleichverteilung auf einem endlichen Intervall $(a_i, b_i), -\infty < a_i < b_i < \infty$, dann ist $\mu = \mu_1 \otimes \ldots \otimes \mu_n$ die Gleichverteilung auf dem Quader $S = \prod_{i=1}^n (a_i, b_i)$, denn für $c_1, \ldots, c_n \in S$ gilt:

$$\begin{aligned}\mu\left[\prod_{i=1}^n (-\infty, c_i]\right] &= \prod_{i=1}^n \mu - i[(-\infty, c_i]] \\ &= \prod_{i=1}^n \frac{c_i - a_i}{b_i - a_i} \\ &= \lambda^n \left[\prod_{i=1}^n (a_i, c_i]\right] \Big/ \lambda^n[S].\end{aligned}$$

Absolutstetigkeit von multivariaten Verteilungen

Absolutstetigkeit von endlichen Produktmodellen

Der Satz von Fubini, den wir in Abschnitt 9.1 in größerer Allgemeinheit beweisen werden, besagt unter anderem, dass das n-dimensionale Lebesgueintegral einer beliebigen Borel-messbaren nicht-negativen Funktion $f : \mathbb{R}^n \to \mathbb{R}$ existiert, und als Hintereinanderausführung von eindimensionalen Integralen nach den Koordinaten $x_1, \ldots, x_n$ berechnet werden kann:

$$\int_{\mathbb{R}^n} f(x)\, dx \quad = \quad \int \cdots \int f(x_1, \ldots, x_n)\, dx_n \cdots dx_1.$$

Hierbei können die eindimensionalen Integrationen in beliebiger Reihenfolge ausgeführt werden. Für den Beweis verweisen wir auf die Analysisvorlesung bzw. auf Abschnitt 9.1 unten.
In Analogie zum eindimensionalen Fall heißt eine Wahrscheinlichkeitsverteilung μ auf $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ *stetig* oder *absolutstetig*, falls eine $\mathcal{B}(\mathbb{R}^n)$-messbare *Dichtefunktion* $f : \mathbb{R}^n \to [0, \infty)$ existiert mit

$$\mu[B] \quad = \quad \int_B f(x)\, dx \quad := \quad \int I_B(x) f(x)\, dx$$

für jeden Quader, bzw. allgemeiner für jede Borelmenge $B \subseteq \mathbb{R}^n$. Endliche Produkte von eindimensionalen absolutstetigen Verteilungen sind wieder absolutstetig, und die Dichte ist das Produkt der einzelnen Dichten:

Lemma 5.6. *Sind $\mu_1, \ldots, \mu_n$ absolutstetige Wahrscheinlichkeitsverteilungen auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ mit Dichtefunktionen $f_1, \ldots, f_n$, dann ist das Produkt $\mu = \mu_1 \otimes \ldots \otimes \mu_n$ eine absolutstetige Wahrscheinlichkeitsverteilung auf $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ mit Dichtefunktion*

$$f(x_1, \ldots, x_n) \quad = \quad \prod_{i=1}^n f_i(x_i).$$

Beweis. Für jede Produktmenge $B = B_1 \times \ldots \times B_n$, $B_i \in \mathcal{B}(\mathbb{R})$, gilt nach dem Satz von Fubini:

$$\mu[B] \quad = \quad \prod_{i=1}^n \mu_i[B_i] \quad = \quad \prod_{i=1}^n \int_{B_i} f_i(x_i) dx_i \quad = \quad \int \cdots \int I_B(x_1, \ldots, x_n) \prod_{i=1}^n f_i(x_i) dx_1 \cdots dx_n.$$

□

Die Dichtefunktion der Gleichverteilung auf dem Quader $S = (a_1, b_1) \times \ldots \times (a_n, b_n)$ ist beispielsweise

$$f(x_1, \ldots, x_n) \quad = \quad \prod_{i=1}^n \frac{1}{b_i - a_i} I_{(a_i, b_i)}(x_i) = \frac{1}{\text{Volumen}[S]} I_S(x).$$

Ein anderes Produktmaß von fundamentaler Bedeutung für die Wahrscheinlichkeitstheorie ist die mehrdimensionale Standardnormalverteilung:

Beispiel (Standardnormalverteilung im $\mathbb{R}^n$). Die Wahrscheinlichkeitsverteilung

$$\mu \quad = \quad \bigotimes_{i=1}^{n} N(0,1)$$

auf $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ heißt n-**dimensionale Standardnormalverteilung**. Die mehrdimensionale Standardnormalverteilung ist absolutstetig mit Dichte

$$f(x_1,\ldots,x_n) \quad = \quad \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{x_i^2}{2}\right) = (2\pi)^{-n/2} e^{-\|x\|^2/2}, \qquad x \in \mathbb{R}^n.$$

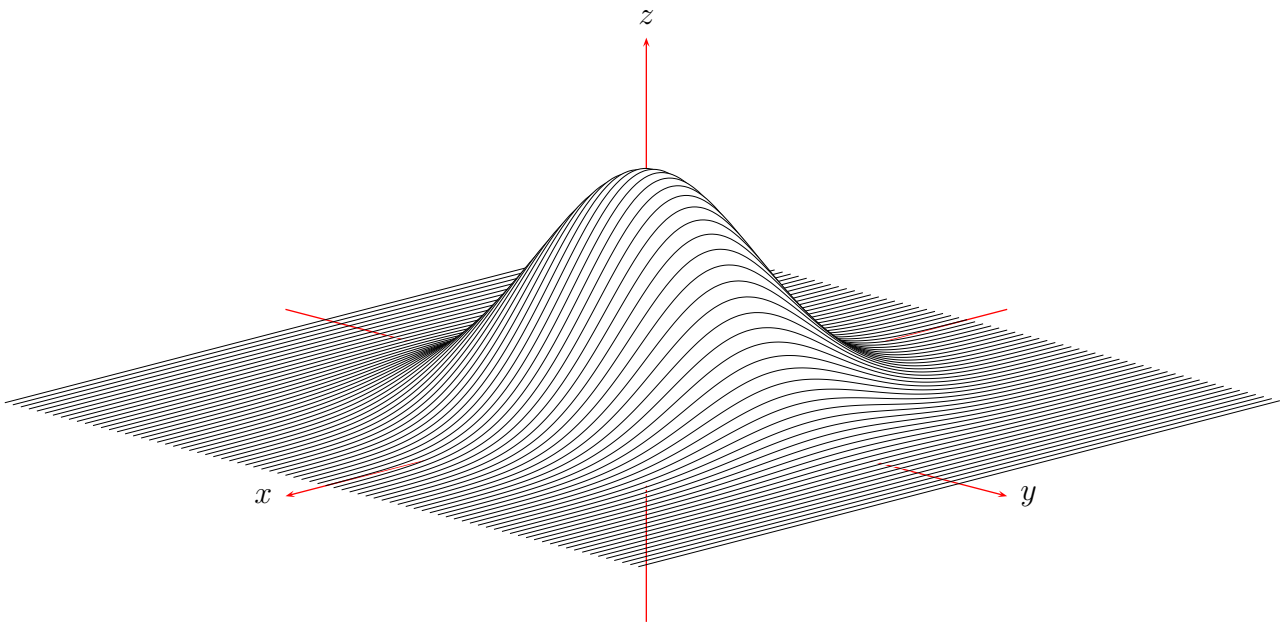


Abbildung 5.1: Dichte der Standardnormalverteilung in $\mathbb{R}^2$.

Gemeinsame Verteilungen

Sind $X_i : \Omega \to S_i, 1 \le i \le n$, beliebige Zufallsvariablen mit Werten in messbaren Räumen $(S_i, \mathcal{S}_i)$, welche auf einem gemeinsamen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ definiert sind, dann ist

$$(X_1,\ldots,X_n): \quad \Omega \quad \longrightarrow \quad S_1 \times \ldots \times S_n$$

eine Zufallsvariable mit Werten im Produktraum $(S_1 \times \ldots \times S_n, \mathcal{S}_1 \otimes \ldots \otimes \mathcal{S}_n)$, denn für $B_i \in \mathcal{S}_i, 1 \leq i \leq n$, gilt:

$$\{(X_1, \ldots, X_n) \in B_1 \times \ldots \times B_n\} \quad = \quad \bigcap_{i=1}^{n} \{X_i \in B\} \quad \in \mathcal{A}.$$

Wie zuvor im diskreten Fall (s. Abschnitt 2.4) definieren wir:

Definition. *Die Verteilung $\mu_{X_1,\ldots,X_n}$ des Zufallsvektors $(X_1, \ldots, X_n)$ auf $(S_1 \times \ldots \times S_n, \mathcal{S}_1 \otimes \ldots \otimes \mathcal{S}_n)$ heißt **gemeinsame Verteilung** der Zufallsvariablen $X_1, \ldots, X_n$.*

Der folgende Satz gilt analog zum diskreten Fall:

Satz 5.7. *Die folgenden Aussagen sind äquivalent:*

(1). Die Zufallsvariablen $X_1, \ldots, X_n$ sind unabhängig.

(2). Die gemeinsame Verteilung $\mu_{X_1,\ldots,X_n}$ ist ein Produktmaß.

(3). $\mu_{X_1,\ldots,X_n} \quad = \quad \mu_{X_1} \otimes \ldots \otimes \mu_{X_n}$.

Beweis. „ (1) $\Longrightarrow$ (3) “: folgt direkt aus der Definition der Unabhängigkeit und der gemeinsamen Verteilung: Sind $X_1, \ldots, X_n$ unabhängig, dann gilt

$$\begin{aligned}
\mu_{X_1,\ldots,X_n}[B_1 \times \ldots \times B_n] &= P[(X_1, \ldots, X_n) \in B_1 \times \ldots \times B_n] \\
&= P[X_i \in B_i, \quad \forall 1 \leq i \leq n] \\
&= \prod_{i=1}^{n} P[X_i \in B_i] \\
&= \prod_{i=1}^{n} \mu_{X_i}[B_i]
\end{aligned}$$

für alle $B_i \in \mathcal{S}_i$, $1 \leq i \leq n$.

„ (3) $\Longrightarrow$ (2) “: Die Implikation ist offensichtlich, und „ (2) $\Longrightarrow$ (1) “ folgt aus Lemma 5.5: Ist $\mu_{X_1,\ldots,X_n}$ ein Produktmaß, dann sind die kanonischen Projektionen $\pi_1, \ldots, \pi_n$ unabhängig unter $\mu_{X_1}, \ldots, \mu_{X_n}$. Also gilt für $B_i \in \mathcal{S}_i$:

$$\begin{aligned}
P[X_1 \in B_1, \ldots, X_n \in B_n] &= \mu_{X_1,\ldots,X_n}[B_1 \times \ldots \times B_n] \\
&= \mu_{X_1,\ldots,X_n}[\pi_1 \in B_1, \ldots, \pi_n \in B_n] \\
&= \prod_{i=1}^{n} \mu_{X_1,\ldots,X_n}[\pi_i \in B_i] \\
&= \prod_{i=1}^{n} P[\pi_i \in B_i]
\end{aligned}$$

□

Wir wenden die Aussage von Satz 5.7 nun speziell auf diskrete und reellwertige Zufallsvariablen an:

Diskrete Zufallsvariablen

Sind die Wertebereiche $S_1, \ldots, S_n$ der Zufallsvariablen $X_1, \ldots, X_n$ abzählbar, dann wird die gemeinsame Verteilung vollständig durch die gemeinsame Massenfunktion

$$p_{X_1,\ldots,X_n}(a_1,\ldots,a_n) \quad = \quad P[X_1 = a_1, \ldots, X_n = a_n], \qquad (a_1,\ldots,a_n) \in S_1 \times \ldots \times S_n$$

beschrieben. Die Zufallsvariablen $X_1, \ldots, X_n$ sind genau dann unabhängig, wenn die gemeinsame Massenfunktion das Produkt der einzelnen Massenfunktionen ist, s. Satz 2.7. Als Konsequenz aus Satz 5.7 ergibt sich zudem:

Korollar 5.8. *Sind $X_i : \Omega \to S_i$, $1 \leq i \leq n$, diskrete Zufallsvariablen, und hat die gemeinsame Massenfunktion eine Darstellung*

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

in Produktform mit einer Konstanten $c \in \mathbb{R}$, und Funktionen $g_i : S_i \to [0,\infty)$, dann sind $X_1, \ldots, X_n$ unabhängig mit Massenfunktion

$$p_{X_i}(a_i) \quad = \quad \frac{g_i(a_i)}{\sum\limits_{a \in S_i} g_i(a)}$$

Beweis. Die Werte

$$\widetilde{g}_i(a_i) \quad = \quad \frac{g_i(a_i)}{\sum\limits_{a \in S_i} g_i(a)}, \qquad a_i \in S_i,$$

sind die Gewichte eine Wahrscheinlichkeitsverteilung μ_i auf S_i. Nach Voraussetzung gilt

$$\begin{aligned} \mu_{X_1,\ldots,X_n}[\{a_1\} \times \ldots \times \{a_n\}] &= p_{X_1,\ldots,X_n}(a_1,\ldots,a_n) \\ &= \widetilde{c} \cdot \prod_{i=1}^{n} \widetilde{\mu}_{X_i}[\{a_i\}] \qquad \forall\, (a_1,\ldots,a_n) \in S_1 \times \ldots \times S_n \end{aligned} \tag{5.2.3}$$

mit einer reellen Konstante $\widetilde{c}$. Da auf beiden Seiten von (5.2.3) bis auf den Faktor $\widetilde{c}$ die Massenfunktionen von Wahrscheinlichkeitsverteilungen stehen, gilt $\widetilde{c} = 1$, und damit

$$\mu_{X_1,\ldots,X_n} \quad = \quad \bigotimes_{i=1}^{n} \mu_i.$$

Also sind die X_i unabhängig mit Verteilung μ_i, d.h. mit Massenfunktion $\widetilde{g}_i$. □

Beispiel (Zwei 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 beispielsweise folgende Möglichkeiten:

(1). X, Y unabhängig.

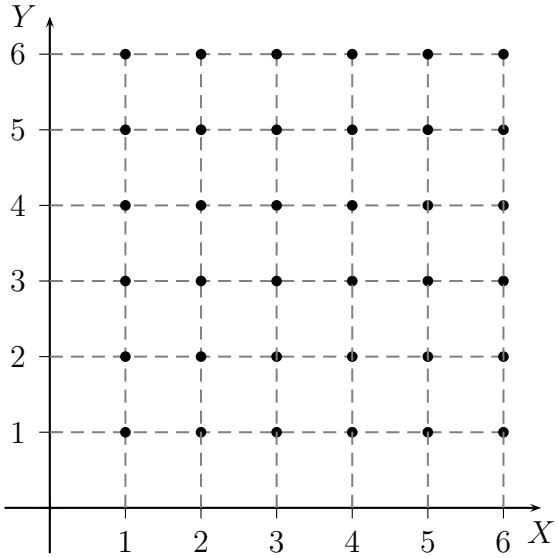


Abbildung 5.2: X, Y unabhängig; $\mu_{X,Y} = \mu_X \otimes \mu_Y$. Gewichte der Punkte sind jeweils $\frac{1}{36}$

(2). X, Y deterministisch korreliert, z.B. $Y = (X + 1) \mod 6$.

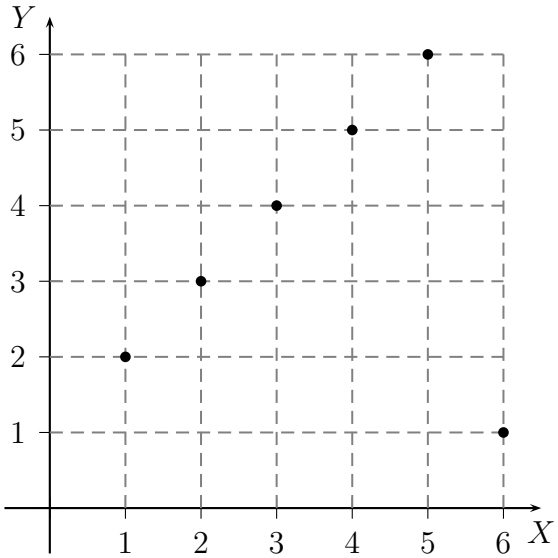


Abbildung 5.3: $Y = (X + 1) \mod 6$. Das Gewicht eines einzelnen Punktes ist $\frac{1}{6}$.

(3). $Y = (X + Z) \mod 6$, Z unabhängig von X, $Z = 0, \pm 1$ mit Wahrscheinlichkeit $\frac{1}{3}$.

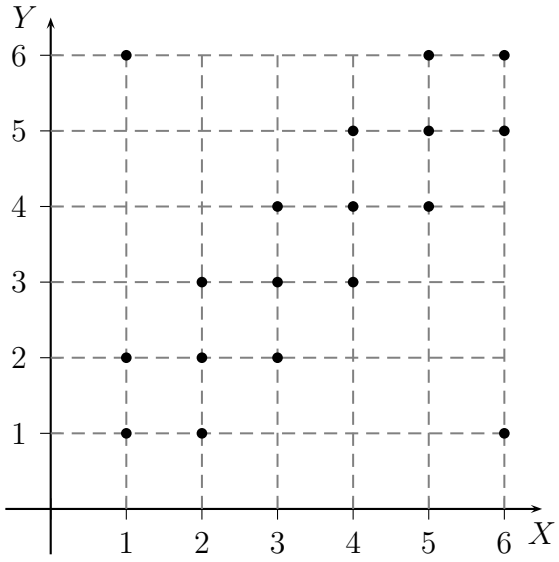


Abbildung 5.4: $Y = (X+Z) \mod 6$; $Z \sim \text{unif}\{-1,0,1\}$. Das Gewicht eines einzelnen Punktes ist $\frac{1}{18}$

Reelle Zufallsvariablen

Die gemeinsame Verteilung reellwertiger Zufallsvariablen $X_1, \ldots, X_n : \Omega \to \mathbb{R}$ auf der Produkt-σ-Algebra $\mathcal{B}(\mathbb{R}^n) = \bigotimes_{i=1}^{n} \mathcal{B}(\mathbb{R})$ ist vollständig durch die Werte

$$\begin{aligned} F_{X_1,\ldots,X_n}(c_1,\ldots,c_n) &:= \mu_{X_1,\ldots,X_n}[(-\infty,c_1] \times \ldots \times (-\infty,c_n]] \\ &= P[X_1 \le c_1, \ldots, X_n \le c_n], \quad (c_1,\ldots,c_n) \in \mathbb{R}^n, \end{aligned}$$

beschrieben. Die Funktion $F_{X_1,\ldots,X_n} : \mathbb{R}^n \to [0,1]$ heißt **gemeinsame Verteilungsfunktion**. Insbesondere sind $X_1, \ldots, X_n$ genau dann unabhängig, wenn

$$F_{X_1,\ldots,X_n}(c_1,\ldots,c_n) = \prod_{i=1}^{n} F_{X_i}(c_i) \qquad \forall (c_1,\ldots,c_n) \in \mathbb{R}^n$$

gilt. In Analogie zu Korollar 5.8 erhalten wir zudem:

Korollar 5.9. *Seien $X_1, \ldots, X_n : \Omega \to \mathbb{R}$ reellwertige Zufallsvariablen.*

(1). Sind $X_1, \ldots, X_n$ unabhängige Zufallsvariablen mit absolutstetigen Verteilungen mit Dichten $f_{X_1}, \ldots, f_{X_n}$, dann ist die gemeinsame Verteilung absolutstetig mit Dichte

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \prod_{i=1}^{n} f_{X_i}(x_i) \qquad \forall\, x \in \mathbb{R}^n.$$

(2). Umgekehrt gilt: Ist die gemeinsame Verteilung absolutstetig, und hat die Dichte eine Darstellung

$$f_{X_1,\dots,X_n}(x_1,\dots,x_n) \quad = \quad c \cdot \prod_{i=1}^{n} g_i(x_i) \qquad \forall\, x \in \mathbb{R}^n$$

in Produktform mit einer Konstante $c \in \mathbb{R}$ und integrierbaren Funktionen $g_i : \mathbb{R} \to [0,\infty)$, dann sind $X_1,\dots,X_n$ unabhängig, und die Verteilungen sind absolutstetig mit Dichten

$$f_{X_i}(x_i) \quad = \quad \frac{g_i(x_i)}{\int\limits_{\mathbb{R}} g_i(t)\,dt}.$$

Der Beweis verläuft ähnlich wie der von Korollar 5.8, und wird dem Leser zur Übung überlassen.

Beispiel (Zufällige Punkte in der Ebene). Seien X und Y unabhängige Zufallsvariablen, $N(0,\sigma^2)$-verteilte auf $(\Omega,\mathcal{A},P)$ mit $\sigma > 0$. Dann ist die gemeinsame Verteilung $\mu_{X,Y}$ absolutstetig mit Dichte

$$f_{X,Y}(x,y) \quad = \quad \frac{1}{2\pi\sigma^2} \cdot \exp(-\frac{x^2+y^2}{2\sigma^2}), \qquad (x,y) \in \mathbb{R}^2.$$

es gilt $(X,Y) \neq (0,0)$ P-fast sicher. Wir definieren den Radial- und Polaranteil

$$R \,:\, \Omega \to (0,\infty), \qquad \Phi \,:\, \Omega \to [0,2\pi)$$

durch

$$X \quad = \quad R \cdot \cos\Phi \qquad \text{und} \qquad Y \quad = \quad R \cdot \sin\Phi,$$

d.h. $R = \sqrt{X^2+Y^2}$ und $\Phi = \arg(X+iY)$ falls $(X,Y) \neq (0,0)$. Auf der Nullmenge $\{(X,Y) = (0,0)\}$ definieren wir (R,Φ) in beliebiger Weise, sodass sich messbare Funktionen ergeben. Wir berechnen nun die gemeinsame Verteilung von R und Φ:

$$\begin{aligned}
P[R \le r_0, \Phi \le \phi_0] &= P[(X,Y) \in \text{ „Kuchenstück“ mit Winkel } \phi_0 \text{ und Radius } r_0] \\
&= \int\limits_{\text{Kuchenstück}} \int f_{X,Y}(x,y)\,dx\,dy \\
&= \int_0^{r_0} \int_0^{\phi_0} f_{X,Y}(r\cos\phi, r\sin\phi) \underbrace{r}_{\substack{\text{Jacobideterminante} \\ \text{der Koordinatentrans. } f}} d\phi\,dr \\
&= \int_0^{r_0} \int_0^{\phi_0} \frac{r}{2\pi\sigma^2} e^{-r^2/(2\sigma^2)}\,d\phi\,dr.
\end{aligned}$$

Hierbei haben wir im 3. Schritt den Transformationssatz (Substitutionsregel) für mehrdimensionale Integrale verwendet - der Faktor r ist die Jacobideterminante der Koordinatentransformation (s. Analysis). Es folgt, dass die gemeinsame Verteilung $\mu_{R,\Phi}$ absolutstetig ist mit Dichte

$$f_{R,\Phi}(r,\phi) \quad = \quad \frac{1}{2\pi} \cdot \frac{r}{\sigma^2} \cdot e^{-r^2/(2\sigma^2)}.$$

Da die Dichte Produktform hat, sind R und Φ unabhängig. Die Randverteilung μ_Φ ist absolutstetig mit Dichte

$$f_\Phi(\phi) \quad = \quad \text{const.} \quad = \quad \frac{1}{2\pi} \qquad (0 \le \phi < 2\pi),$$

d.h. Φ ist gleichverteilt auf $[0, 2\pi)$. Somit ist μ_R absolutstetig mit Dichte

$$\phi_R(r) \quad = \quad \frac{r}{\sigma^2} \cdot e^{-r^2/(2\sigma^2)} \qquad (r > 0).$$

Die Berechnung können wir verwenden, um Stichproben von der Standardnormalverteilung zu simulieren:

Beispiel (Simulation von normalverteilten Zufallsvariablen). Die Verteilungsfunktion einer $N(0,1)$-verteilten Zufallsvariable X ist

$$F_X(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt\ .$$

Das Integral ist nicht explizit lösbar und die Inverse F_X^{-1} ist dementsprechend nur approximativ berechenbar. Daher ist die Simulation einer Standardnormalverteilung durch Inversion der Verteilungsfunktion relativ aufwendig. Ein einfacheres Simulationsverfahren ergibt sich, wenn wir eine zweidimensionale Standardnormalverteilung betrachten und auf Polarkoordinaten transformieren. Dann gilt für den Radialanteil:

$$F_R(x) \ = \ \int_0^x e^{-r^2/2} r\, dr \ = \ 1 - e^{-x^2/2}.$$

Das Integral ist also explizit berechenbar, und

$$F_R^{-1}(u) \ = \ \sqrt{-2\log(1-u)}\,, \quad u \in (0,1).$$

Der Winkelanteil Φ ist unabhängig von R und gleichverteilt auf $[0, 2\pi)$. Wir können Zufallsvariablen mit der entsprechenden gemeinsamen Verteilung erzeugen, indem wir

$$\begin{aligned} \Phi &:= 2\pi U_1\ , \\ R &:= \sqrt{-2\log(1-U_2)} \qquad \left(\text{bzw. } = \sqrt{-2\log U_2}\right), \end{aligned}$$

setzen, wobei U_1 und U_2 unabhängige, auf $(0,1)$ gleichverteilte Zufallsvariablen sind. Stichproben von U_1 und U_2 können durch Pseudozufallszahlen simuliert werden. die Zufallsvariablen

$$X := R\cos\Phi \qquad \text{und} \qquad Y := R\cdot\sin\Phi$$

sind dann unabhängig und $N(0,1)$-verteilt. Für $m\in\mathbb{R}$ und $\sigma>0$ sind $\sigma X+m$ und $\sigma Y+m$ unabhängige $N(m,\sigma^2)$-verteilte Zufallsvariable.

Wir erhalten also den folgenden Algorithmus zur Simulation von Stichproben einer Normalverteilung:

Algorithmus 5.10 (Box-Muller-Verfahren). Input: $m\in\mathbb{R}, \sigma>0$

Output: unabhängige Stichproben $\widetilde{x},\widetilde{y}$ von $N(m,\sigma^2)$.

1. Erzeuge unabhängige Zufallszahlen $u_1,u_2\sim\mathcal{U}_{(0,1)}$

2. $x:=\sqrt{-2\log u_1}\cos(2\pi u_2)$, $y:=\sqrt{-2\log u_1}\sin(2\pi u_2)$

3. $\widetilde{x}:=\sigma x+m$, $\widetilde{y}=\sigma y+m$

Beispiel (Ordnungsstatistiken). Für die gesamte Verteilung der Ordnungsstatistiken $X_{(1)}\le\ldots\le X_{(n)}$, unabhängiger, identisch verteilter, stetiger Zufallsvariablen $X_1,\ldots,X_n:\Omega\to\mathbb{R}$ gilt aus Symmetriegründen und wegen $P[X_i=X_j]=0$ für $i\neq j$:

$$\begin{aligned}
P[X_{(1)}\le c_1,\ldots,X_{(n)}\le c_n] &= \sum_{\pi\in S_n} P[X_{\pi(1)}\le c_1,\ldots,X_{\pi(n)}\le c_n, X_{\pi(1)}<\ldots<X_{\pi(n)}]\\
&= n!\,P[X_1\le c_1,\ldots,X_n\le c_n, X_1<X_2<\ldots<X_n]\\
&= n!\int_{-\infty}^{c_1}\cdots\int_{-\infty}^{c_n} I_{\{y_1<y_2<\ldots<y_n\}} f(y_1)\cdots f(y_n)\,dy_1\cdots dy_n.
\end{aligned}$$

Also ist die gemeinsame Verteilung von $X_{(1)},\ldots,X_{(n)}$ absolutstetig mit Dichte

$$f_{X_{(1)},\ldots,X_{(n)}}(y_1,\ldots,y_n) = n!\cdot I_{\{y_1<y_2<\ldots<y_n\}} f(y_1)\cdots f(y_n).$$

Durch Aufintegrieren erhält man daraus mithilfe des Satzes von Fubini und einer erneuten Symmetrieüberlegung die Dichten der Verteilungen der einzelnen Ordnungsstatistiken:

$$\begin{aligned}
P[X_{(k)}\le c] &= n!\int_{\mathbb{R}}\cdots\int_{\mathbb{R}} I_{\{y_1<y_2<\ldots<y_n\}} f(y_1)\cdots f(y_n)\cdot I_{Y_k\le c}\,dy_1\cdots dy_n\\
&= \int_{-\infty}^{c} f_{(k)}(y_k)\,dy_k
\end{aligned}$$

mit

$$f_{(k)}(y) \quad = \quad \frac{n!}{(k-1)!(n-k)!} F(y)^{k-1}(1-F(y))(n-k)f(y).$$

Dasselbe Resultat hätte man auch mit etwas Rechnen aus Satz 5.3 herleiten können.

Bemerkung (Beta-Verteilungen). Sind die Zufallsvariablen X_i auf $(0,1)$ gleichverteilt, dann hat $X_{(k)}$ die Dichte

$$f_{X_{(k)}}(u) \quad = \quad B(k, n-k+1)^{-1} \cdot u^{k-1} \cdot (1-u)^{n-k} \cdot I_{(0,1)}(u)$$

mit Normierungskonstante

$$B(a,b) \quad = \quad \int_0^1 u^{a-1}(1-u)^{b-1}\, du \qquad \left(= \frac{(a-1)!(b-1)!}{(a+b-1)!} \qquad \text{für } a, b \in \mathbb{N} \right).$$

Die entsprechende Verteilung heißt **Beta-Verteilung mit Parametern** $a, b > 0$, die Funktion B ist die *Euler'sche Beta-Funktion*.

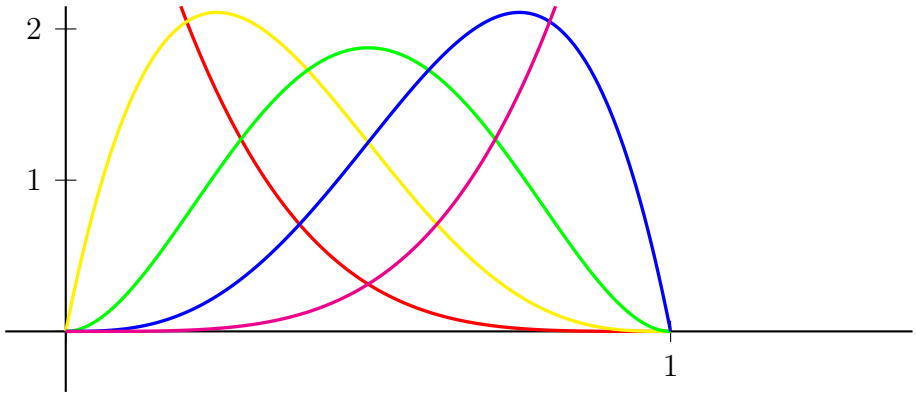


Abbildung 5.5: Abbildung der Dichtefunktionen der zugehörigen Verteilungen von $X_{(1)}, \ldots, X_{(5)}$ bei $n = 5$ in (rot, gelb, grün, blau, magenta).

5.3 Unendliche Produktmodelle

Konstruktion von unabhängigen Zufallsvariablen

Seien $\mu_1, \mu_2, \ldots$ vorgegebene Wahrscheinlichkeitsverteilungen auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Wir werden nun explizit unabhängige Zufallsvariablen $X_k, k \in \mathbb{N}$, mit Verteilungen μ_k konstruieren. Als Konsequenz ergibt sich die Existenz des unendlichen Produktmaßes $\bigotimes_{k=1}^{\infty} \mu_k$ als gemeinsame Verteilung der Zufallsvariablen X_k. Die Zufallsvariablen X_i können wir sogar auf den Raum $\Omega = (0,1)$ mit Gleichverteilung realisieren:

Satz 5.11. *Auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{B}((0,1)), \mathcal{U}_{(0,1)})$ existieren unabhängige Zufallsvariablen $X_k : \Omega \to \mathbb{R}, k \in \mathbb{N}$, mit Verteilungen*

$$P \circ X_k^{-1} \quad = \quad \mu_k \qquad \textit{für alle } 1 \leq i \leq n.$$

Beweis. Wir verfahren in drei Schritten:

(1). Wir konstruieren die Zufallsvariablen im Fall

$$\mu_k \quad = \quad \text{Bernoulli}\left(\frac{1}{2}\right) \quad = \quad \mathcal{U}_{(0,1)} \qquad \forall\, k \in \mathbb{N},$$

d.h. im fairen Münzwurfmodell. Dazu verwenden wir die schon in Abschnitt 4.2 eingeführte Transformation $X : (0,1) \to \{0,1\}^{\mathbb{N}}$, die einer reellen Zahl die Ziffernfolge ihrer Binärdarstellung zuordnet, d.h. wir setzen

$$X_k(\omega) \quad = \quad I_{D_k}(\omega), \qquad D_k \quad = \quad \bigcup_{i=1}^{2^{k-1}} [(2i-1)\cdot 2^{-k}, 2i \cdot 2^{-k}),$$

siehe Abbildung 4.4. Die Abbildungen $X_k : (0,1) \to \{0,1\}$ sind messbar, und es gilt

$$P[X_1 = a_1, \ldots, X_n = a_n] \quad = \quad 2^n \qquad \forall n \in \mathbb{N}, a_1, \ldots, a_n \in \{0,1\}, \tag{5.3.1}$$

da die Menge $\{\omega \in \Omega : X_1(\omega) = a_1, \ldots, X_n(\omega) = a_n\}$ gerade aus den Zahlen in $(0,1)$ besteht, deren Binärdarstellung mit den Ziffern $a_1, \ldots, a_n$ beginnt, und damit ein Intervall der Länge 2^{-n} ist. Nach (5.3.1) sind $X_1, \ldots, X_n$ für alle $X_k, k \in \mathbb{N}$, unabhängig mit Verteilung μ_k.

(2). Wir konstruieren die Zufallsvariablen im Fall

$$\mu_k \quad = \quad \mathcal{U}_{(0,1)} \qquad \forall\, k \in \mathbb{N}.$$

Dazu zerlegen wir die gerade konstruierte Folge $X_k(\omega) \in \{0,1\}, k \in \mathbb{N}$, in unendlich viele Teilfolgen, und konstruieren aus jeder Teilfolge wieder eine Zahl aus $[0,1]$ mit den entsprechenden Binärziffern. Genauer setzen wir in Binärdarstellung:

$$\begin{aligned} U_1 &:= 0.X_1X_3X_5X_7\cdots, \\ U_2 &:= 0.X_2X_6X_{10}X_{14}\cdots, \\ U_3 &:= 0.X_4X_{12}X_{20}X_{28}\cdots, \qquad \text{usw.}, \end{aligned}$$

also allgemein für $k \in \mathbb{N}$:

$$U_k(\omega) \quad := \quad \sum_{i=1}^{\infty} X_{k,i}(\omega) \cdot 2^{-i} \qquad \text{mit} \quad X_{k,i} := X_{(2i-1)\cdot 2^{k-1}}.$$

Da die Zufallsvariablen $X_{k,i}, i, k \in \mathbb{N}$, unabhängig sind, sind nach dem Zerlegungssatz auch die σ-Algebren

$$\mathcal{A}_k \quad = \quad \sigma(X_{k,i} | i \in \mathbb{N}), \qquad k \in \mathbb{N},$$

unabhängig, und damit auch die $\mathcal{A}_k$-messbaren Zufallsvariablen U_k, $k \in \mathbb{N}$. Zudem gilt für $n \in \mathbb{N}$ und

$$r = \sum_{i=1}^{n} a_i \cdot 2^{i-1} \quad \in \quad \{0, 1, \ldots, 2^n - 1\} :$$

$$P[U_k \in (r \cdot 2^{-n}, (r+1) \cdot 2^{-n})] \quad = \quad P[X_{k,1} = a_1, \ldots, X_{k,n} = a_n] \quad = \quad 2^{-n}.$$

Da die dyadischen Intervalle ein durchschnittsstabiles Erzeugendensystem der Borelschen σ-Algebra bilden, folgt, dass die Zufallsvariablen U_k auf $[0, 1]$ gleichverteilt sind.

(3). Im allgemeinen Fall konstruieren wir die Zufallsvariablen aus den gerade konstruierten unabhängigen gleichverteilten Zufallsvariablen $U_k, k \in \mathbb{N}$, mithilfe des Inversionsverfahrens aus Satz 4.19: Sind $\mu_k, k \in \mathbb{N}$, beliebige Wahrscheinlichkeitsverteilungen auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, und

$$\underline{G}_k(u) \quad = \quad \inf\{x \in \mathbb{R} : F_k(x) \geq u\}$$

die linksstetigen verallgemeinerten Inversen der Verteilungsfunktionen

$$F_k(c) \quad = \quad \mu_k[(-\infty, c]],$$

dann setzen wir

$$Y_k(\omega) \quad := \quad \underline{G}_k(U_k(\omega)), \qquad k \in \mathbb{N}, \omega \in \Omega.$$

Da die Zufallsvariablen $U_k, k \in \mathbb{N}$, unabhängig sind, sind nach Satz 5.2 auch die $Y_k, k \in \mathbb{N}$, wieder unabhängig. Zudem gilt nach Satz 4.19:

$$P \circ Y_k^{-1} \quad = \quad \mu_k \qquad \text{für alle } k \in \mathbb{N}.$$

□

Bemerkung. (1). Der Beweis von Satz 5.11 ist konstruktiv. Für numerische Anwendungen ist allerdings zumindest der erste Schritt des beschriebenen Konstruktionsverfahrens ungeeignet, da Defizite des verwendeten Zufallszahlengenerators und die Darstellungsungenauigkeit im Rechner durch die Transformation verstärkt werden.

(2). Mithilfe des Satzes kann man auch die Existenz einer Folge unabhängiger Zufallsvariablen $X_k, k \in \mathbb{N}$, mit Werten im $\mathbb{R}^d$, oder allgemeiner in vollständigen, separablen, metrischen

Räumen $S_k, k \in \mathbb{N}$, und vorgegebenen Verteilungen μ_k auf den Borelschen σ-Algebren $\mathcal{B}(S_k)$ zeigen. Sind beispielsweise $\phi_k : \mathbb{R} \to S_k$ Bijektionen, sodass ϕ_k und ϕ_k^{-1} messbar sind, und sind $\widetilde{X}_k : \Omega \to \mathbb{R}$ unabhängige reellwertige Zufallsvariablen mit Verteilungen $P[\widetilde{X}_k \in B] = \mu_K[\phi_k(B)]$, dann sind die transformierten Zufallsvariablen

$$X_k \quad = \quad \phi_k(\widetilde{X}_k) : \Omega \to S_k, \qquad \forall k \in \mathbb{N},$$

unabhängig mit Verteilungen μ_k.

Beispiel (Random Walks im $\mathbb{R}^d$). Sei μ eine Wahrscheinlichkeitsverteilung auf $(\mathbb{R}^d, \mathcal{B}(R^d))$, und seien $X_i, i \in \mathbb{N}$, unabhängige Zufallsvariablen mit identischer Verteilung $X_i \sim \mu$. Der durch

$$S_n \quad = \quad a + \sum_{i=1}^{n} X_i, \qquad n = 0, 1, 2, \ldots,$$

definierte stochastische Prozess heißt *Random Walk mit Startwert $a \in \mathbb{R}^d$ und Inkrementverteilung μ.*

Im Fall $d = 1$ können wir Stichproben von den Zufallsvariablen X_i, und damit vom Random Walk, beispielsweise mithilfe der Inversionsmethode, simulieren.

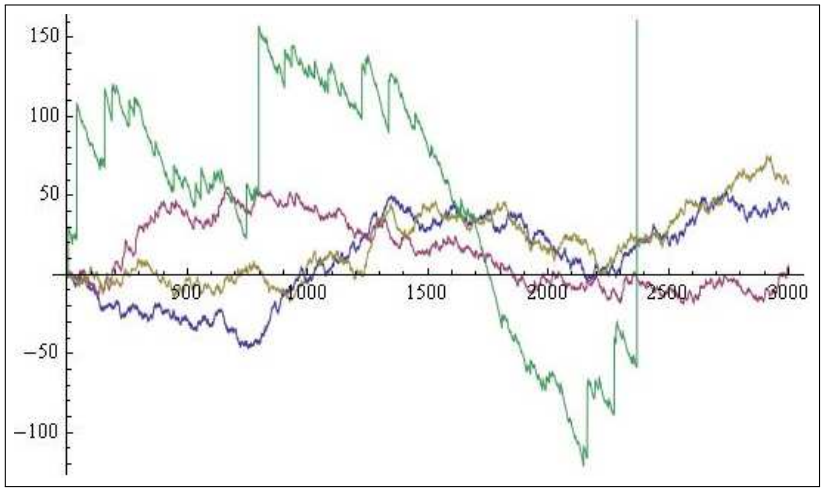


Abbildung 5.6: Grafiken von Trajektorien des Random Walks mit verschiedenen Inkrementverteilungen.

Abbildung 5.6 zeigt Grafiken von Trajektorien des Random Walks mit den Inkrementverteilungen

$$\mu \quad = \quad \frac{1}{2}(\delta_1 + \delta_{-1}) \qquad \text{(klassischer Random Walk (SSRW))},$$

$$\mu \quad = \quad N(0,1) \qquad \text{(diskrete Brownsche Bewegung)},$$

μ mit Dichte

$$f(x) \quad = \quad e^{-(x+1)} I_{(-1,\infty)(x)} \qquad \text{(zentrierte } \mathrm{Exp}(1)\text{-Verteilung)}$$

und μ mit Dichte

$$f(x) \quad = \quad 3{\cdot}2^{-5/2}{\cdot}(x{+}\frac{3}{2})^{-5/2}{\cdot}I_{(\frac{1}{2},\infty)}(x{+}\frac{3}{2}) \qquad \text{(zentrierte Pareto}(\alpha-1,\alpha)\text{-Verteilung mit } \alpha = \frac{3}{2}).$$

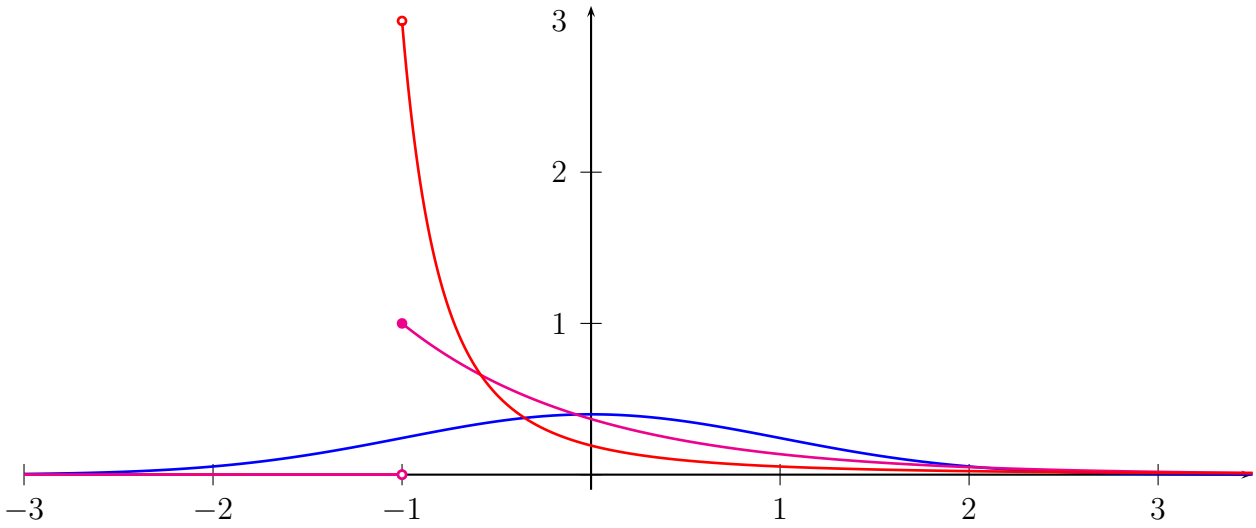


Abbildung 5.7: Dichten der drei stetigen Verteilungen aus Abbildung 5.6: $f_{N(0,1)}$ in Blau, $f_{\mathrm{Exp}(1)-1}$ in Magenta und $f_{\mathrm{Pareto}(\alpha-1,\alpha)}$ in Rot.

Im Gegensatz zu den anderen Verteilungen fällt die Dichte der Pareto-Verteilung für $x \to \infty$ nur sehr langsam ab („heavy tails“). Insbesondere hat die Verteilung unendliche Varianz. Die Trajektorien der Random Walks werden mit der folgenden Mathematica-Routine simuliert:

```
nmax = 10000; )
x = RandomChoice[{−1, 1}, nmax];
z = RandomReal[NormalDistribution[0, 1], nmax];
u = RandomReal[{0, 1}, nmax]; y = –Log[u] – 1;
$\alpha$ = 3/2; x0 = $\alpha$ – 1; p =
 RandomReal[ParetoDistribution[x0, $\alpha$], nmax];
 m = Mean[ParetoDistribution[x0, $\alpha$]]; q = p – m;

rwsimple = Accumulate[x]; rwexp = Accumulate[y];
rwnormal = Accumulate[z]; rwpareto = Accumulate[q];

ListLinePlot[{rwsimple[[1 ;; 3000]], rwexp[[1 ;; 3000]],
rwnormal[[1 ;; 3000]], rwpareto[[1 ;; 3000]]}]
```

Die Trajektorien des klassischen Random Walks, und der Random Walks mit exponential- und normalverteilten Inkrementen sehen in größeren Zeiträumen ähnlich aus. Die Trajektorien des Pareto-Random Walks (grün) verhalten sich dagegen anders, und werden auch in längeren Zeiträumen von einzelnen großen Sprüngen beeinflusst. Tatsächlich kann man zeigen, dass alle obigen Random Walks mit Ausnahme des Pareto-Random Walks in einem geeigneten Skalierungslimes mit Schrittweite gegen 0 in Verteilung gegen eine Brownsche Bewegung konvergieren (funktionaler zentraler Grenzwertsatz).

Unendliche Produktmaße

Als Konsequenz aus dem Satz können wir die Existenz von unendlichen Produktmaßen als gemeinsame Verteilung von unendlich vielen unabhängigen Zufallsvariablen zeigen. Dazu versehen wir den Folgenraum

$$\mathbb{R}^{\mathbb{N}} \quad = \quad \{(x_1, x_2, \ldots) | x_k \in \mathbb{R}, \forall\, k \in \mathbb{N}\}$$

mit der Produkt-σ-Algebra

$$\bigotimes_{k\in\mathbb{N}} \mathcal{B}(\mathbb{R}) \quad = \quad \sigma(\mathcal{C}) \quad = \quad \sigma(\pi_k | k \in \mathbb{N}),$$

die von der Kollektion $\mathcal{C}$ aller Zylindermengen

$$\{\pi_1 \in B_1, \ldots, \pi_n \in B_n\} \quad = \quad \{x = (x_k) \in \mathbb{R}^{\mathbb{N}} | x_1 \in B_1, \ldots, x_n \in B_n\},$$

$n \in \mathbb{N}, B_1, \ldots, B_n \in \mathcal{B}(\mathbb{R})$, von den Koordinatenabbildungen $\pi_k : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}, \pi_k(x) = x_k$.

Korollar 5.12 (Existenz von unendlichen Produktmaßen). *Zu beliebigen Wahrscheinlichkeitsverteilungen μ_k auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ existiert eine eindeutige Wahrscheinlichkeitsverteilung $\mu = \bigotimes_{k\in\mathbb{N}} \mu_k$ auf $(\mathbb{R}^{\mathbb{N}}, \bigotimes_{k\in\mathbb{N}} \mathcal{B}(\mathbb{R}))$ mit*

$$\mu[\pi_1 \in B_1, \ldots, \pi_n \in B_n] \quad = \quad \mu[B_1] \cdot \ldots \cdot \mu_n[B_n] \tag{5.3.2}$$

für alle $n \in \mathbb{N}$ und $B_1, \ldots, B_n \in \mathcal{B}(\mathbb{R})$.

Definition. *Die Wahrscheinlichkeitsverteilung μ mit (5.3.2) heißt **Produkt der Wahrscheinlichkeitsverteilungen** $\mu_k, k \in \mathbb{N}$.*

Beweis. Die Eindeutigkeit folgt, da die Zylindermengen aus $\mathcal{C}$ ein $\cap$-stabiles Erzeugendensystem der Produkt-σ-Algebra bilden.

Zum Beweis der Existenz: betrachten wir die Abbildung $X : \Omega \to \mathbb{R}^{\mathbb{N}}$ mit

$$X(\omega) \quad = \quad (X_1(\omega), X_2(\omega), \ldots),$$

wobei X_k unabhängige Zufallsvariablen mit Verteilung μ_k sind. X ist messbar bzgl. $\bigotimes_{k\in\mathbb{N}} \mathcal{B}(\mathbb{R})$, denn

$$X^{-1}[\{x \in \mathbb{R}^{\mathbb{N}} | (x_1, \ldots, x_n) \in B\}] \quad = \quad \{\omega \in \Omega | (X_1(\omega), \ldots, X_n(\omega)) \in B\} \quad \in \quad \mathcal{A}$$

für alle $n \in \mathbb{N}$ und $B \in \mathcal{B}(\mathbb{R}^n)$. Sei $\mu = P \circ X^{-1}$ die Verteilung von X auf $\mathbb{R}^{\mathbb{N}}$. Dann gilt

$$\begin{aligned} \mu[\pi_1 \in B_1, \ldots, \pi_m \in B_n] &= \mu[\{x \in \mathbb{R}^{\mathbb{N}} | x_1 \in B_1, \ldots, x_n \in B_n\}] \\ &= P[X_1 \in B_1, \ldots, X_n \in B_n] \\ &= \prod_{k=1}^{n} \mu_k[B_k] \end{aligned}$$

für alle $n \in \mathbb{N}$ und $B_1, \ldots, B_n \in \mathcal{B}(\mathbb{R})$. Also ist μ das gesuchte Produktmaß. □

Bemerkung. Auf analoge Weise folgt nach Bemerkung 2. von oben die Existenz des Produktmaßes $\bigotimes_{k\in\mathbb{N}} \mu_k$ von beliebigen Wahrscheinlichkeitsverteilungen $\mu_k, k \in \mathbb{N}$, auf vollständigen, separablen, messbaren Räumen S_k mit Borelschen σ-Algebren $\mathcal{S}_k$. Das Produktmaß sitzt auf dem Produktraum

$$\left(\bigtimes_{k\in\mathbb{N}} S_k, \bigotimes_{k\in\mathbb{N}} \mathcal{S}_k \right).$$

Der Satz von Carathéodory impliziert sogar die Existenz von beliebigen (auch überabzählbaren) Produkten von allgemeinen Wahrscheinlichkeitsräumen $(S_i, \mathcal{S}_i, \mu_i), i \in I$.

Sind $(S_i, \mathcal{S}_i, \mu_i)$ beliebige Wahrscheinlichkeitsräume, dann sind die Koordinatenabbildungen $\pi_k : \bigtimes_{i\in\mathbb{N}} S_i \to S_k$ unter dem Produktmaß $\bigotimes_{i\in I} \mu_i$ unabhängig und μ_k-verteilt. Man nennt den Produktraum

$$(\Omega, \mathcal{A}, P) \quad = \quad \left(\bigtimes S_i, \bigotimes \mathcal{S}_i, \bigotimes \mu_i \right)$$

daher auch das **kanonische Modell** für unabhängige μ_i-verteilte Zufallsvariablen.

5.4 Asymptotische Ereignisse

Sei X_i $(i \in I)$ eine unendliche Kollektion von Zufallsvariablen, die auf einem gemeinsamen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ definiert sind.

Definition. *Ein Ereignis $A \in \sigma(X_i \,|i \in I)$ heißt **asymptotisches Ereignis (tail event)**, falls*

$$A \;\in\; \sigma(X_i \mid i \in I\backslash I_0) \qquad \textit{für jede } \textbf{endliche} \textit{ Teilmenge } I_0 \subseteq\; I \textit{ gilt.}$$

Die Menge

$$\tau \quad = \quad \bigcap_{I_0 \subseteq I \text{ endlich}} \sigma(X_i \mid i \in I\backslash I_0)$$

*aller asymptotischen Ereignisse ist eine σ-Algebra. τ heißt **asymptotische σ-Algebra (tail field)**.*

Beispiel. (1). DYNAMISCH: Ist $X_n, n \in \mathbb{N}$ eine Folge von Zufallsvariablen (welche beispielsweise eine zufällige zeitliche Entwicklung beschreibt), dann gilt für ein Ereignis $A \in \sigma(X_n, n \in \mathbb{N})$:

$$A \text{ asymptotisch} \quad \Leftrightarrow \quad A \;\in\; \underbrace{\sigma(X_{n+1}, X_{n+2}, \;\ldots\,)}_{\text{Zukunft ab } n} \qquad \text{für alle } n.$$

Beispiele für asymptotische Ereignisse von reellwertigen Zufallsvariablen sind

$$\{X_n > 5n \quad \text{unendlich oft}\}, \qquad \left\{\limsup_{n\to\infty} X_n < c\right\}, \qquad \{\exists \lim_{n\to\infty} X_n\}, \qquad \left\{\exists \lim \frac{1}{n} S_n = m\right\},$$

wobei $S_n = X_1 + \ldots + X_n$. Die Ereignisse

$$\{\sup_{n\in\mathbb{N}} X_n = 3\} \qquad \text{und} \quad \{\lim S_n = 5\}$$

sind dagegen *nicht* asymptotisch.

(2). STATISCH: Eine Kollektion $X_i, i \in \mathbb{Z}^d$, von Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ heißt **stochastisches Feld** (random field). Beispielsweise basieren verschiedene grundlegende Modelle der statistischen Mechanik auf stochastischen Feldern $X_i : \Omega \to \{0, 1\}$, wobei $X_i = 1$ dafür steht, dass

- sich ein Teilchen am Gitterpunkt i befindet,
- ein Atom am Gitterpunkt i angeregt ist,
- der Gitterpunkt i durchlässig ist (Perkolationsmodell),
- etc.

Asymptotische Ereignisse beschreiben in diesem Fall „makroskopische“ Effekte.

Das 0-1-Gesetz von Kolmogorov

Satz 5.13 (0-1-Gesetz von Kolmogorov). *Sind X_i $(i \in I)$ unabhängige Zufallsvariablen auf $(\Omega, \mathcal{A}, P)$, dann gilt*

$$P[A] \in \{0,1\} \textbf{ für alle } A \in \tau.$$

„Asymptotische Ereignisse sind deterministisch."

Beweis. Der Übersichtlichkeit halber führen wir den Beweis im Fall $I = \mathbb{N}$ - der Beweis im allgemeinen Fall verläuft ähnlich. Es gilt: $X_1, X_2, ...$ unabhängige Zufallsvariablen
$\Longrightarrow \sigma(X_1), \sigma(X_2), ..., \sigma(X_n), \sigma(X_{n+1}), \sigma(X_{n+2}), ...$ unabhängige Mengensysteme
$\Longrightarrow \sigma(X_1, ..., X_n), \sigma(X_{n+1}, X_{n+2}, ...)$ sind unabhängig für alle $n \in \mathbb{N}$
$\Longrightarrow \sigma(X_1, ..., X_n)$ und τ sind unabhängig für alle $n \in \mathbb{N}$
$\Longrightarrow \tau$ unabhängig von $\sigma(X_1, X_2, ...) \supseteq \tau$
$\Longrightarrow$ Ereignisse $A \in \tau$ sind unabhängig von sich selbst
$\Longrightarrow P[A] \in \{0,1\} \qquad \forall\, A \in \tau$.
Hierbei gilt die zweite Implikation nach Satz 5.1 (2), und die vierte nach Satz 5.1 (1) □

Anwendungen auf Random Walks und Perkolationsmodelle

Beispiel (Rückkehr zum Startpunkt von Random Walks, Rekurrenz). Wir betrachten einen eindimensionalen klassischen Random Walk mit Startpunkt $a \in \mathbb{Z}$ und unabhängigen Inkrementen X_i mit Verteilung

$$P[X_i = 1] = p, \qquad P[X_i = -1] = 1 - p.$$

Für $n \in \mathbb{N}$ erhält man die Rückkehrwahrscheinlichkeiten

$$P[S_{2n+1} = a] = 0$$

$$P[S_{2n} = a] = \binom{2n}{n} \cdot p^n \cdot (1-p)^n = \frac{(2n)!}{(n!)^2} \cdot p^n \cdot (1-p)^n.$$

Wir betrachten nun die Asymptotik für $n \to \infty$ dieser Wahrscheinlichkeiten. Aus der **Stirlingschen Formel**

$$n! \sim \sqrt{2\pi n} \cdot \left(\frac{n}{e}\right)^n$$

folgt

$$P[S_{2n} = a] \sim \frac{\sqrt{4\pi n}}{2\pi n} \cdot \frac{(\frac{2n}{e})^{2n}}{(\frac{n}{e})^{2n}} \cdot p^n \cdot (1-p)^n = \frac{1}{\sqrt{\pi n}} (4p(1-p))^n \quad \text{für } n \to \infty.$$

Für $p \neq \frac{1}{2}$ fallen die Wahrscheinlichkeiten also exponentiell schnell ab. Insbesondere gilt dann

$$\sum_{m=0}^{\infty} P[S_m = a] \quad = \quad \sum_{n=0}^{\infty} P[S_{2n} = a] < \infty,$$

d.h. der asymmetrische Random Walk kehrt nach dem 1. Borel-Cantelli Lemma mit Wahrscheinlichkeit 1 nur endlich oft zum Startpunkt zurück (TRANSIENZ). Nach dem starken Gesetz großer Zahl gilt sogar

$$S_n \quad \sim \quad (2p-1)n \qquad P\text{-fast sicher.}$$

Für $p = \frac{1}{2}$ gilt dagegen $P[S_{2n} = a] \sim 1/\sqrt{\pi n}$, und damit

$$\sum_{m=0}^{\infty} P[S_m = a] \quad = \quad \sum_{n=0}^{\infty} P[S_{2n} = a] \quad = \quad \infty.$$

Dies legt nahe, dass der Startpunkt mit Wahrscheinlichkeit 1 unendlich oft besucht wird.

Ein Beweis dieser Aussage über das Borel-Cantelli-Lemma ist aber nicht direkt möglich, da die Ereignisse $\{S_{2n} = 0\}$ nicht unabhängig sind. Wir beweisen nun eine stärkere Aussage mithilfe des Kolmogorovschen 0-1-Gesetzes:

Satz 5.14 (Rekurrenz und unbeschränkte Oszillationen des symmetrischen Random Walks). *Für $p = \frac{1}{2}$ gilt*

$$P[\overline{\lim}\, S_n = +\infty \text{ und } \underline{\lim}\, S_n = -\infty] \quad = \quad 1.$$

*Insbesondere ist der eindimensionale Random Walk **rekurrent**, d.h.*

$$P[S_n = a \text{ unendlich oft}] = 1.$$

Tatsächlich wird nach dem Satz mit Wahrscheinlichkeit 1 sogar jeder Punkt $\lambda \in \mathbb{Z}$ unendlich oft getroffen.

Beweis. Für alle $k \in \mathbb{N}$ gilt:

$$P[S_{n+k} - S_n = k \text{ unendlich oft}] = 1,$$

denn nach dem Beispiel zu Satz 5.1 („Affe tippt Shakespeare“) gibt es P-fast sicher unendlich viele Blöcke der Länge k mit $X_{n+1} = X_{n+2} = ... = X_{n+k} = 1$. Es folgt

$$P[\overline{\lim}\, S_n - \underline{\lim}\, S_n = \infty] \quad \geq \quad P\left[\bigcap_k \bigcup_n \{S_{n+k} - S_n = k\}\right] \quad = \quad 1,$$

und damit

$$1 \quad = \quad P[\overline{\lim}\, S_n = +\infty \text{ oder } \underline{\lim}\, S_n = -\infty] \quad \leq \quad P[\overline{\lim}\, S_n = +\infty] + P[\underline{\lim}\, S_n = -\infty].$$

Also ist eine der beiden Wahrscheinlichkeiten auf der rechten Seite größer als $\frac{1}{2}$, und damit nach dem Kolmogorovschen 0-1-Gesetz gleich 1. Aus Symmetriegründen folgt

$$P[\underline{\lim}\, S_n = -\infty] \quad = \quad P[\overline{\lim}\, S_n = +\infty] \quad = \quad 1.$$

□

Das vorangehende Beispiel zeigt eine typische Anwendung des Kolmogorovschen 0-1-Gesetzes auf stochastische Prozesse. Um die Anwendbarkeit in räumlichen Modellen zu demonstrieren, betrachten wir ein einfaches Perkolationsmodell:

Beispiel (Perkolation im $\mathbb{Z}^d$). Sei $p \in (0,1)$ fest, und seien X_i $(i \in \mathbb{Z}^d)$ unabhängige Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ mit

$$P[X_i = 1] = p \quad , \qquad P[X_i = 0] = 1 - p\,.$$

Ein Gitterpunkt $i \in \mathbb{Z}^d$ heißt **durchlässig**, falls $X_i = 1$ gilt. Wir verbinden Gitterpunkte $i, j \in \mathbb{Z}^d$ mit $|i - j| = 1$ durch eine Kante. Sei A das Ereignis, dass bzgl. dieser Graphenstruktur eine unendliche Zusammenhangskomponente (Cluster) aus durchlässigen Gitterpunkten existiert (Eine Flüssigkeit könnte in diesem Fall durch ein makroskopisches Modellstück, das aus mikroskopischen Gitterpunkten aufgebaut ist, durchsickern - daher der Name „Perkolation"). A ist asymptotisch, also gilt nach dem Satz von Kolmogorov

$$P[A] \;\in\; \{0,1\}.$$

Hingegen ist es im Allgemeinen nicht trivial, zu entscheiden, welcher der beiden Fälle eintritt. Im Fall $d = 1$ zeigt man leicht (Übung):

$$P[A] = 0 \quad \text{für alle } p < 1.$$

Für $d = 2$ gilt:

$$P[A] = 1 \quad \Longleftrightarrow \quad p > \frac{1}{2}\,,$$

s. z.B. die Monografie *„Percolation"* von *Grimmett*. Für $d \geq 3$ ist nur bekannt, dass ein kritischer Parameter $p_c \in (0,1)$ existiert mit

$$P[A] = \begin{cases} 1 & \text{für } p > p_c. \\ 0 & \text{für } p < p_c. \end{cases}$$

Man kann obere und untere Schranken für p_c herleiten (z.B. gilt $\frac{1}{2d-1} \leq p_c \leq \frac{2}{3}$), aber der genaue Wert ist nicht bekannt. Man vermutet, dass $P[A] = 0$ für $p = p_c$ gilt, aber auch diese Aussage konnte bisher nur in Dimension $d \geq 19$ (sowie für $d = 2$) bewiesen werden, siehe das Buch von Grimmett.

Definition. *Eine Zufallsvariable $Y : \Omega \to [-\infty, \infty]$ heißt **asymptotisch**, wenn die bzgl. der asymptotischen σ-Algebra τ messbar ist.*

Das Perkolationsmodell ist ein Beispiel für ein sehr einfach formulierbares stochastisches Modell, das zu tiefgehenden mathematischen Problemstellungen führt. Es ist von großer Bedeutung, da ein enger Zusammenhang zu anderen Modellen der statistischen Mechanik und dabei auftretenden Phasenübergängen besteht. Einige elementare Aussagen über Perkolationsmodelle werden in den Wahrscheinlichkeitstheorie-Lehrbüchern von *Y. Sinai* und *A. Klenke* hergeleitet.

Korollar 5.15. *Sind X_i $(i \in I)$ unabhängige Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$, dann ist jede asymptotische Zufallsvariable $Y : \Omega \to [-\infty, \infty]$ P - fast sicher konstant, d.h.*

$$\exists\, c_0 \in [-\infty, \infty] : \quad P[Y = c_0] = 1 .$$

Beweis. Ist Y τ - messbar, dann sind die Ereignisse $\{Y \leq c\}, \quad c \in \mathbb{R}$, in τ enthalten. Aus dem Kolmogorovschen 0-1-Gesetz folgt:

$$F_Y(c) \quad = \quad P[Y \leq c] \quad \in \quad \{0,1\} \qquad \forall\ c \in \mathbb{R}.$$

Da die Verteilungsfunktion monoton wachsend ist, existiert ein $c_0 \in [-\infty, \infty]$ mit

$$P[Y \leq c] \quad = \quad \begin{cases} 0 & \text{für} \quad c < c_0 \\ 1 & \text{für} \quad c > c_0 \end{cases},$$

und damit $P[Y = c_0] \quad = \quad \lim\limits_{\varepsilon \downarrow 0}(F_Y(c_0) - F_Y(c_0 - \varepsilon)) \quad = \quad 1. = 1.$ □

Beispiele für asymptotische Zufallsvariablen im Fall $I = \mathbb{N}$ sind etwa

$$\underline{\lim_{n\to\infty}} X_n, \quad \overline{\lim_{n\to\infty}} X_n, \quad \underline{\lim_{n\to\infty}} \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \text{sowie} \quad \overline{\lim_{n\to\infty}} \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Insbesondere sind für unabhängige Zufallsvariablen $X_1, X_2, ... : \Omega \to \mathbb{R}$ sowohl

$$\underline{\lim}\, \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{als auch} \quad \overline{\lim}\, \frac{1}{n}\sum_{i=1}^{n} X_i \quad P \text{ - f.s. konstant.}$$

Hieraus ergibt sich die folgende *Dichotomie*: Sind $X_i, i \in \mathbb{N}$, unabhängige reellwertige Zufallsvariablen, dann gilt *entweder* ein Gesetz großer Zahlen, d.h.

$$\frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{konvergiert } P \text{ - f.s., und der Limes ist } P \text{ - f.s. konstant}$$

(falls der Limes inferior und Limes superior P-fast sicher übereinstimmen), *oder*

$$P\left[\frac{1}{n}\sum_{i=1}^{n} X_i \text{ konvergiert}\right] \;=\; 0.$$

Es ist bemerkenswert, dass für die Gültigkeit der Dichotomie keine Annahmen über die Verteilung der X_i benötigt werden. Insbesondere müssen die X_i nicht identisch verteilt sein!

Kapitel 6

Erwartungswert und Varianz

In diesem Kapitel definieren wir den Erwartungswert, die Varianz und die Kovarianz allgemeiner reellwertiger Zufallsvariablen, und beweisen grundlegende Eigenschaften und Abschätzungen. Da wir auch Grenzübergänge durchführen wollen, erweist es sich als günstig, die Werte $+\infty$ und $-\infty$ zuzulassen. Wir setzen daher $\overline{\mathbb{R}} = [-\infty, \infty]$. Der Raum $\overline{\mathbb{R}}$ ist ein topologischer Raum bzgl. des üblichen Konvergenzbegriffs. Die Borelsche σ-Algebra auf $\overline{\mathbb{R}}$ wird u.a. erzeugt von den Intervallen $[-\infty, c], c \in \mathbb{R}$. Die meisten Aussagen über reellwertige Zufallsvariablen aus den vorangegangenen Abschnitten übertragen sich unmittelbar auf Zufallsvariablen $X : \Omega \to \overline{\mathbb{R}}$, wenn wir die Verteilungsfunktion $F_X : \mathbb{R} \to [0, 1]$ definieren durch

$$F_X(c) \quad = \quad \mu_X[[-\infty, c]] \quad = \quad P[X \leq c].$$

6.1 Erwartungswert

Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum und $X : \Omega \to \overline{\mathbb{R}}$ eine Zufallsvariable. Wir wollen den Erwartungswert (Mittelwert, Prognosewert) von X bezüglich der Wahrscheinlichkeitsverteilung P in sinnvoller Weise definieren. Dazu gehen wir schrittweise vor:

Definition des Erwartungswerts

Elementare Zufallsvariablen

Nimmt X nur endlich viele Werte $c_1, ..., c_n \in \mathbb{R}$ an, dann soll gelten:

$$E[X] \quad = \quad \sum_{i=1}^{n} c_i \cdot P[X = c_i],$$

d.h. der Erwartungswert ist das Mittel der Werte c_i gewichtet mit den Wahrscheinlichkeiten der Ereignisse $A_i := \{X = c_i\}$.

Definition. *Eine Zufallsvariable von der Form*

$$X = \sum_{i=1}^{n} c_i I_{A_i} \quad (n \in \mathbb{N},\ c_i \in \mathbb{R},\ A_i \in \mathcal{A})$$

heißt ***elementar****. Ihr* ***Erwartungswert*** *bzgl.* P *ist*

$$E[X] \ := \ \sum_{i=1}^{n} c_i \cdot P[A_i].$$

Diese Definition ist ein Spezialfall der Definition des Erwartungswerts diskreter Zufallsvariablen aus Kapitel 1. Insbesondere ist der Erwartungswert $E[X]$ *wohldefiniert*, d.h. unabhängig von der gewählten Darstellung der Zufallsvariable X als Linearkombination von Indikatorfunktionen, und die Abbildung $X \mapsto E[X]$ ist *linear* und *monoton*:

$$E[aX + bY] \quad = \quad a \cdot E[X] + b \cdot E[Y] \qquad \text{für alle } a, b \in \mathbb{R},$$

$$X \le Y \quad \Longrightarrow \quad E[X] \le E[Y].$$

Die Definition des Erwartungswerts einer elementaren Zufallsvariable stimmt genau mit der des Lebesgueintegrals der Elementarfunktion X bzgl. des Maßes P überein:

$$E[X] \quad = \quad \int X \, dP \quad = \quad \int X(\omega) \, P(d\omega)$$

Für allgemeine Zufallsvariablen liegt es nahe, den Erwartungswert ebenfalls als Lebesgueintegral bzgl. des Maßes P zu definieren. Wir skizzieren hier die weiteren Schritte zur Konstruktion des Lebesgueintegrals bzw. des Erwartungswerts einer allgemeinen Zufallsvariable, siehe auch die Analysisvorlesung.

Nichtnegative Zufallsvariablen

Die Definition des Erwartungswerts einer nichtnegativen Zufallsvariable beruht auf der monotonen Approximation durch elementare Zufallsvariablen:

Lemma 6.1. *Sei* $X : \Omega \to [0, \infty]$ *eine nichtnegative Zufallsvariable auf* $(\Omega, \mathcal{A}, P)$*. Dann existiert eine monoton wachsende Folge elementarer Zufallsvariablen* $0 \le X_1 \le X_2 \le \ldots$ *mit*

$$X \quad = \quad \lim_{n \to \infty} X_n \quad = \quad \sup_{n \in \mathbb{N}} X_n.$$

Beweis. Für $n \in \mathbb{N}$ sei

$$X_n(\omega) \quad := \quad \begin{cases} (k-1)\cdot 2^{-n} & \text{falls } \ (k-1)\cdot 2^{-n} \le X(\omega) < k\cdot 2^{-n} \text{ für ein } k = 1, 2, \ldots, n\cdot 2^n \\ n & \text{falls } \ X(\omega) \ge n \end{cases}.$$

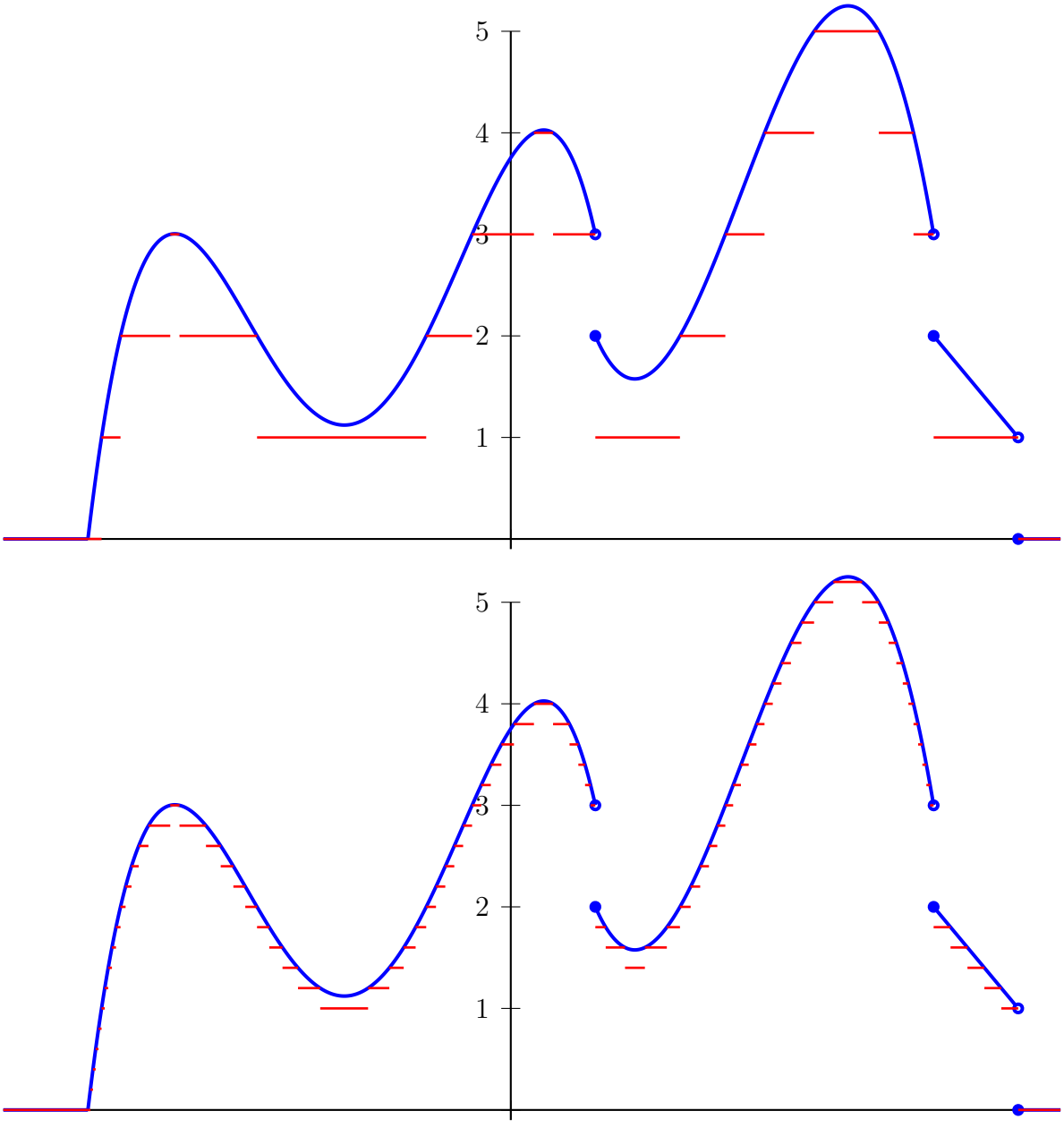


Abbildung 6.1: Approximation durch Elementarfunktionen. Hier ist die Annäherung in rot in zwei verschiedenen Feinheiten dargestellt.

Dann ist X_n eine elementare Zufallsvariable, denn es gilt

$$X_n \quad = \quad \sum_{k=0}^{n2^n-1} \frac{k}{2^n} I_{\left\{\frac{k}{2^n} \le X < \frac{k+1}{2^n}\right\}} + nI_{\{X \ge n\}}.$$

Die Folge $X_n(\omega)$ ist für jedes ω monoton wachsend, da die Unterteilung immer feiner wird, und

$$\sup_{n\in\mathbb{N}} X_n(\omega) \quad = \quad \lim_{n\to\infty} X_n(\omega) \quad = \quad X(\omega) \qquad \text{für alle } \omega \in \Omega.$$

□

Definition. *Sei $X : \Omega \to [0, \infty]$ eine nicht-negative Zufallsvariable.*
*Der **Erwartungswert** (bzw. das **Lebesgueintegral**) von X bzgl. P ist definiert als*

$$E[X] \;:=\; \lim_{n\to\infty} E[X_n] \;=\; \sup_{n\to\infty} E[X_n] \;\in\; [0,\infty], \tag{6.1.1}$$

wobei X_n eine monoton wachsende Folge von nichtnegativen elementaren Zufallsvariablen mit $X = \sup X_n$ ist.

Auch in diesem Fall ist der Erwartungswert wohldefiniert (in $[0, \infty]$):

Lemma 6.2. *Die Definition ist unabhängig von der Wahl einer monoton wachsenden Folge X_n von nichtnegativen Zufallsvariablen mit $X = \sup_{n\in\mathbb{N}} X_n$.*

Für den Beweis verweisen wir auf die Analysisvorlesung oder auf die Literatur, siehe z.B. Appendix 5 in WILLIAMS „Probability with martingales."

Bemerkung. Sind $X_n = I_{A_n}$ und $X = I_A$ Indikatorfunktionen, dann folgt (6.1.1) aus der monotonen Stetigkeit von P. In diesem Fall gilt nämlich:

$$X_n \nearrow X \qquad \Longleftrightarrow \qquad A_n \nearrow A \qquad (\text{d.h. } A_n \text{ monoton wachsend und } A = \bigcup A_n).$$

Aus der monotonen Stetigkeit von P folgt dann

$$E[X] \quad = \quad P[A] \quad = \quad \lim P[A_n] \quad = \quad \lim E[X_n].$$

Aus der Definition des Erwartungswerts folgt unmittelbar:

Lemma 6.3. *Für nichtnegative Zufallsvariablen X, Y mit $X \le Y$ gilt $E[X] \le E[Y]$.*

Beweis. Ist $X \le Y$, dann gilt auch $X_n \le Y_n$ für die approximierenden elementaren Zufallsvariablen aus Lemma 6.1, also

$$E[X] \quad = \quad \sup_{n\in\mathbb{N}} E[X_n] \quad \le \quad \sup_{n\in\mathbb{N}} E[Y_n] \quad = \quad E[Y].$$

□

Allgemeine Zufallsvariablen

Eine allgemeine Zufallsvariable $X : \Omega \to \overline{\mathbb{R}}$ können wir in ihren positiven und negativen Anteil zerlegen:

$$X \quad = \quad X^+ - X^- \qquad \text{mit} \qquad X^+ \quad := \quad \max(X, 0), \quad X^- \quad := \quad -\min(X, 0).$$

X^+ und X^- sind nichtnegative Zufallsvariablen. Ist mindestens einer der beiden Erwartungswerte $E[X^+]$ bzw. $E[X^-]$ endlich, dann können wir (ähnlich wie in Kapitel 1 für diskrete Zufallsvariablen) definieren:

Definition. *Der Erwartungswert einer Zufallsvariable $X : \Omega \to \overline{\mathbb{R}}$ mit $E[X^+] < \infty$ oder $E[X^-] < \infty$ ist*

$$E[X] \quad := \quad E[X^+] - E[X^-] \quad \in \quad [-\infty, \infty].$$

Notation: Der Erwartungswert $E[X]$ ist das Lebesgueintegral der messbaren Funktion $X : \Omega \to \overline{\mathbb{R}}$ bzgl. des Maßes P. Daher verwenden wir auch folgende Notation:

$$E[X] \quad = \quad \int X \, dP \quad = \quad \int X(\omega) \, P(d\omega).$$

Eigenschaften des Erwartungswerts

Nachdem wir den Erwartungswert einer allgemeinen Zufallsvariable $X : \Omega \to \overline{\mathbb{R}}$ definiert haben, fassen wir nun einige grundlegende Eigenschaften des Erwartungswerts zusammen. Dazu bezeichnen wir mit

$$\mathcal{L}^1 = \mathcal{L}^1(P) = \mathcal{L}^1(\Omega, \mathcal{A}, P) := \{X : \Omega \to \overline{\mathbb{R}} \text{ Zufallsvariable} \quad | \quad E[|X|] < \infty\}$$

die Menge aller bzgl. P integrierbaren Zufallsvariablen. Für Zufallsvariablen $X \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ ist nach Lemma 6.3 sowohl $E[X^+]$ als auch $E[X^-]$ endlich. Also ist der Erwartungswert $E[X]$ definiert und endlich.

Satz 6.4. *Für Zufallsvariablen $X, Y \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ und $a, b \in \mathbb{R}$ gilt:*

(1). $X \geq 0$ P-fast sicher $\Longrightarrow E[X] \geq 0$

(2). Die Zufallsvariable $aX + bY$ ist bzgl. P integrierbar, und

$$E[aX + bY] \quad = \quad a \cdot E[X] + b \cdot E[Y].$$

Insbesondere ist der Erwartungswert monoton:

(3). $X \leq Y$ *P-fast sicher* $\Longrightarrow E[X] \leq E[Y]$.

Zum Beweis der Eigenschaften (1) und (2) verweisen wir auf die Analysisvorlesung oder die Literatur. Eigenschaft (3) folgt unmittelbar aus (1) und (2).

Nach Aussage (2) des Satzes ist $\mathcal{L}^1(\Omega, \mathcal{A}, P)$ ein Vektorraum. Durch

$$X \sim Y \quad :\Longleftrightarrow \quad P[X = Y] = 1$$

wird eine Äquivalenzrelation auf diesem Raum definiert. Eine Konsequenz von Aussage (3) des Satzes ist, dass zwei äquivalente (also P-fast sicher identische= Zufallsvariablen denselben Erwartungswert haben:

$$X \sim Y \qquad \Longrightarrow \qquad E[X] = E[Y].$$

Daher ist der Erwartungswert einer Äquivalenzklasse von P-fast sicher gleichen Zufallsvariablen eindeutig definiert. In Zukunft verwenden wir häufig dieselbe Notation für die Äquivalenzklassen und Repräsentanten aus den Äquivalenzklassen. Satz 6.4 besagt, dass der Erwartungswert ein *positives lineares Funktional* auf dem Raum

$$L^1(\Omega, \mathcal{A}, P) \quad := \quad \mathcal{L}^1(\Omega, \mathcal{A}, P)/\sim$$

aller Äquivalenzklassen von integrierbaren Zufallsvariablen definiert. Aus dem Satz folgt zudem:

Korollar 6.5. *Durch*

$$\|X\|_{L^1(\Omega,\mathcal{A},P)} \quad = \quad E[|X|]$$

wird eine Norm auf $L^1(\Omega, \mathcal{A}, P)$ definiert. Insbesondere gilt für Zufallsvariablen $X : \Omega \to \overline{\mathbb{R}}$:

$$E[|X|] = 0 \qquad \Longrightarrow \qquad X = 0 \quad \textit{P-fast sicher.}$$

Beweis. Für eine Zufallsvariable $X : \Omega \to \overline{\mathbb{R}}$ mit $E[|X|] = 0$ und $\varepsilon > 0$ gilt wegen der Monotonie und Linearität des Erwartungswerts:

$$P[|X| \geq \varepsilon] \;=\; E[I_{\{|X|\geq\varepsilon\}}] \;\leq\; E\left[\frac{|X|}{\varepsilon}\right] \;=\; \frac{1}{\varepsilon}E[|X|] = 0.$$

Für $\varepsilon \searrow 0$ folgt

$$P[|X| > 0] = \lim_{\varepsilon \searrow 0} P[|X| \geq \varepsilon] = 0,$$

also $X = 0$ P-fast sicher.

Zudem folgt aus der Monotonie und Linearität des Erwartungswerts die Dreiecksungleichung:

$$E[|X + Y|] \quad \leq \quad E[|X| + |Y|] \quad = \quad E[|X|] + E[|Y|].$$

□

In der Analysis wird gezeigt, dass der Raum $L^1(\Omega, \mathcal{A}, P)$ bzgl. der im Korollar definierten Norm ein Banachraum ist.

Konvergenzsätze

Ein Vorteil des Lebesgueintegrals gegenüber anderen Integrationsbegriffen ist die Gültigkeit von sehr allgemeinen Konvergenzsätzen. Diese lassen sich zurückführen auf den folgenden fundamentalen Konvergenzsatz, der sich aus der oben skizzierten Konstruktion des Lebesgueintegrals ergibt:

Satz 6.6 (Satz von der monotonen Konvergenz, B. Levi). *Ist $X_n, n \in \mathbb{N}$, eine monoton wachsende Folge von Zufallsvariablen mit $E[X_1^-] < \infty$ (z.B. $X_1 \geq 0$), dann gilt:*

$$E[\sup_{n\in\mathbb{N}} X_n] \quad = \quad E[\lim_{n\to\infty} X_n] \quad = \quad \lim_{n\to\infty} E[X_n] \quad = \quad \sup_{n\in\mathbb{N}} E[X_n].$$

Der Beweis findet sich in zahlreichen Lehrbüchern der Integrations- oder Warscheinlichkeitstheorie, siehe z.B. WILLIAMS: PROBABILITY WITH MARTINGALES, APPENDIX 5.

Eine erste wichtige Konsequenz des Satzes von der monotonen Konvergenz ist:

Korollar 6.7. *Für nichtnegative Zufallsvariablen $X_i, i \in \mathbb{N}$, gilt:*

$$E\left[\sum_{i=1}^{\infty} X_i\right] \quad = \quad \sum_{i=1}^{\infty} E[X_i].$$

Beweis.

$$\begin{aligned} E\left[\sum_{i=1}^{\infty} X_i\right] &= E\left[\lim_{n\to\infty} \sum_{i=1}^{n} X_i\right] \\ &= \lim_{n\to\infty} E\left[\sum_{i=1}^{n} X_i\right] \qquad \text{(wegen monotoner Konvergenz)} \\ &= \lim_{n\to\infty} \sum_{i=1}^{n} E[X_i] \qquad \text{(wegen Linearität)} \\ &= \sum_{i=1}^{\infty} E[X_i]. \end{aligned}$$

□

Bemerkung (Abzählbare Wahrscheinlichkeitsräume, Summation als Spezialfall von Integration). Falls Ω abzählbar ist, können wir jede Zufallsvariable $X : \Omega \to \mathbb{R}$ auf die folgende Weise als abzählbare Linearkombination von Indikatorfunktionen darstellen:

$$X \quad = \quad \sum_{\omega\in\Omega} X(\omega) \cdot I_{\{\omega\}}.$$

Ist $X \geq 0$, dann gilt nach Korollar 6.7:

$$E[X] \quad = \quad \sum_{\omega \in \Omega} X(\omega) \cdot P[\{\omega\}].$$

Dieselbe Darstellung des Erwartungswerts gilt auch für allgemeine reellwertige Zufallsvariablen auf Ω, falls der Erwartungswert definiert ist, d.h. $E[X^+]$ oder $E[X^-]$ endlich ist.
Insbesondere sehen wir, dass *Summation ein Spezialfall von Integration* ist: Ist Ω abzählbar und $p(\omega) \geq 0$ für alle $\omega \in \Omega$, dann gilt

$$\sum_{\omega \in \Omega} X(\omega) \cdot p(\omega) \quad = \quad \int X \, dP,$$

wobei P das Maß mit Massenfunktion p ist. Beispielsweise gilt also

$$\sum_{\omega \in \Omega} X(\omega) \quad = \quad \int X \, d\nu,$$

wobei ν das durch $\nu[A] = |A|$, $A \subseteq \Omega$, definierte Zählmaß ist.
Konvergenzsätze wie der Satz von der monotonen Konvergenz lassen sich also auch auf Summen anwenden!

Beispiel. Ist P die Gleichverteilung auf einer endlichen Menge Ω, dann ist

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

das **arithmetische Mittel von** X.

Wir beweisen nun noch zwei wichtige Konvergenzsätze, die sich aus dem Satz von der monotonen Konvergenz ergeben:

Korollar 6.8 (Lemma von Fatou). *Seien $X_1, X_2, \cdots : \Omega \to \overline{\mathbb{R}}$ Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ und sei $Y \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ (z.B. $Y \equiv 0$).*

(1). Gilt $X_n \geq Y$ für alle $n \in \mathbb{N}$, dann folgt

$$E[\liminf X_n] \leq \liminf \; E[X_n].$$

(2). Gilt $X_n \leq Y$ für alle $n \in \mathbb{N}$, dann folgt

$$E[\limsup X_n] \geq \limsup \; E[X_n].$$

Beweis. Die Aussagen folgen aus dem Satz über monotone Konvergenz. Beispielsweise gilt:

$$\begin{aligned} E\left[\liminf\ X_n\right] &= E\left[\lim_{n\to\infty}\inf_{k\geq n} X_k\right] = \lim_{n\to\infty} E\left[\inf_{k\geq n} X_k\right] \\ &\leq \lim_{n\to\infty}\inf_{k\geq n} E[X_k] = \liminf_{n\to\infty} E[X_n], \end{aligned}$$

da die Folge der Infima monoton wachsend ist und durch die integrierbare Zufallsvariable Y nach unten beschränkt ist. Die zweite Aussage zeigt man analog. □

Korollar 6.9 (Satz von der majorisierten Konvergenz, Lebesgue). *Sei $X_n : \Omega \to \overline{\mathbb{R}}, n \in \mathbb{N}$, eine P-fast sicher konvergente Folge von Zufallsvariablen. Existiert eine Majorante $Y \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ mit $|X_n| \leq Y$ für alle $n \in \mathbb{N}$, dann gilt*

$$E[\lim X_n] \quad = \quad \lim\ E[X_n]. \tag{6.1.2}$$

Beweis. Nach dem Lemma von Fatou gilt:

$$E[\liminf X_n] \ \leq\ \liminf\ E[X_n] \ \leq\ \limsup\ E[X_n] \ \leq\ E[\limsup X_n],$$

da $X_n \geq -Y \in \mathcal{L}^1$ und $X_n \leq Y \in \mathcal{L}^1$ für alle $n \in \mathbb{N}$ gilt. Konvergiert X_n P-fast sicher, dann stimmen die linke und rechte Seite der obigen Ungleichungskette überein. □

Beispiel. Wir betrachten Setzen mit Verdoppeln auf »Null« für eine Folge von fairen Münzwürfen. Bei Anfangseinsatz 1 beträgt das Kapital des Spielers nach n Münzwürfen

$$X_n \quad = \quad 2^n \cdot I_{\{n<T\}}\ ,$$

wobei T die Wartezeit auf die erste »Eins« ist. Es folgt

$$E[X_n] \ =\ 2^n P[T > n] \ =\ 2^n \cdot 2^{-n} \ =\ 1 \quad \text{für alle } n \in \mathbb{N},$$

das Spiel ist also fair. Andererseits fällt aber P-fast sicher irgendwann eine »Eins«, d.h. es gilt:

$$\lim_{n\to\infty} X_n \ =\ 0 \quad P\text{-fast sicher.}$$

Die Aussage (6.1.2) des Satzes von Lebesgue ist in dieser Situation nicht erfüllt!

6.2 Berechnung von Erwartungswerten; Dichten

Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum. In diesem Abschnitt zeigen wir, wie man in verschiedenen Fällen den Erwartungswert einer Zufallsvariable $X : \Omega \to [0, \infty]$ aus der Verteilung von X berechnen kann.

Diskrete Zufallsvariablen

Falls X nur abzählbar viele Werte annimmt, können wir die Zufallsvariable X auf folgende Weise als abzählbare Linearkombination von Indikatorfunktionen darstellen:

$$X = \sum_{a \in X(\Omega)} a \cdot I_{\{X=a\}}.$$

Es folgt:

$$E[X] \;=\; \sum_{a \in X(\Omega)} E[a \cdot I_{\{X=a\}}] \;=\; \sum_{a \in X(\Omega)} a \cdot P[X=a].$$

Dieselbe Aussage gilt allgemeiner für diskrete reellwertige Zufallsvariablen X mit

$$E[X^+] < \infty \quad \text{oder} \quad E[X^-] < \infty.$$

Für Zufallsvariablen $X : \Omega \to S$, mit Werten in einer beliebigen abzählbaren Menge S, und eine Borel-messbare Funktion $h : S \to \overline{\mathbb{R}}$ erhalten wir entsprechend

$$E[h(X)] \;=\; \sum_{a \in X(\Omega)} h(a) \cdot P[X=a], \tag{6.2.1}$$

falls $E[h(X)]$ definiert ist, also z.B. falls $h \geq 0$ oder $h(X) \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ gilt.

Die allgemeine Definition des Erwartungswerts als Lebesgueintegral stimmt also für diskrete Zufallsvariablen mit der in Kapitel 1 gegebenen Definition überein.

Allgemeine Zufallsvariablen

Die Berechnungsmethode (6.2.1) für den Erwartungswert diskreter Zufallsvariablen lässt sich auf Zufallsvariablen mit beliebigen Verteilungen erweitern. Sei dazu $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum, $(S, \mathcal{S})$ ein messbarer Raum, $X : \Omega \to S$ eine Zufallsvariable, und $h : S \to [0, \infty]$ eine messbare Abbildung.

Satz 6.10 (Transformationssatz). *Unter den obigen Voraussetzungen gilt:*

$$E_P[h(X)] \quad = \quad \int h(X(\omega)) P(d\omega) \quad = \quad \int h(x)\, \mu(dx) \quad = E_\mu[h],$$

wobei $\mu = P \circ X^{-1}$ die Verteilung von X unter P ist, und E_P bzw. E_μ den Erwartungswert unter P bzw. μ bezeichnet.

Die Erwartungswerte hängen somit nur von der Verteilung von X ab!

Beweis. Der Beweis erfolgt in drei Schritten:

(1). Ist $h = I_B$ die Indikatorfunktion einer messbaren Menge $B \in \mathcal{S}$, dann gilt:

$$E[h(X)] \;=\; \int I_B(X(\omega))P(d\omega) \;=\; P[X^{-1}(B)] \;=\; \mu[B] \;=\; \int I_B \, d\mu,$$

da $I_B(X(\omega)) = I_{X^{-1}(B)}(\omega)$ gilt.

(2). Für Linearkombinationen $h = \sum_{i=1}^{n} a_i I_{B_i}$ von Indikatorfunktionen mit $n \in \mathbb{N}, a_i \in \mathbb{R}$, und $B_i \in \mathcal{S}$ gilt die Aussage auch, da das Lebesgueintegral linear vom Integranden abhängt.

(3). Für eine allgemeine messbare Funktion $h \geq 0$ existiert schließlich eine monoton wachsende Folge h_n von Elementarfunktionen mit $h_n(x) \nearrow h(x)$ für alle $x \in S$. Durch zweimalige Anwendung des Satzes von der monotonen Konvergenz erhalten wir erneut:

$$E[h(X)] = E[\lim h_n(X)] = \lim E[h_n(X)] = \lim \int h_n \, d\mu = \int h \, d\mu.$$

□

Das hier verwendete ***Beweisverfahren der »maßtheoretischen Induktion«*** wird noch sehr häufig auftreten: Wir zeigen eine Aussage

(1). für Indikatorfunktionen,

(2). für Elementarfunktionen,

(3). für nichtnegative messbare Funktionen,

(4). für allgemeine integrierbare Funktionen.

Mit maßtheoretischer Induktion zeigt man auch:

Übung: Jede $\sigma(X)$-messbare Zufallsvariable $Y : \Omega \to \overline{\mathbb{R}}$ ist vom Typ $Y = h(X)$ mit einer messbaren Funktion $h : S \to \mathbb{R}$.

Nach Satz 6.10 ist der Erwartungswert $E[T]$ einer reellwertigen Zufallsvariable $T : \Omega \to [0, \infty]$ eindeutig bestimmt durch die Verteilung $\mu_T = P \circ T^{-1}$:

$$E[T] \;=\; \int t \, \mu_T(dt) \, ,$$

also auch durch die Verteilungsfunktion

$$F_T(t) \;=\; P[T \leq t] \;=\; \mu_T[[0, t]], \qquad t \in \mathbb{R}.$$

Der folgende Satz zeigt, wie man den Erwartungswert konkret aus F_T berechnet:

Satz 6.11. *Für eine Zufallsvariable $T : \Omega \to [0, \infty]$ gilt*

$$E[T] \;=\; \int_0^\infty P[T > t]\, dt \;=\; \int_0^\infty (1 - F_T(t))\, dt.$$

Beweis. Wegen

$$T(\omega) = \int_0^{T(\omega)} dt = \int_0^\infty I_{\{T>t\}}(\omega) dt$$

erhalten wir

$$E[T] = E\left[\int_0^\infty I_{\{T>t\}} dt\right] = \int_0^\infty E\left[I_{\{T>t\}}\right] dt = \int_0^\infty P[T > t]\, dt.$$

Hierbei haben wir im Vorgriff auf Kapitel 9 den *Satz von Fubini* benutzt, der gewährleistet, dass man zwei Lebesgueintegrale (das Integral über t und den Erwartungswert) unter geeigneten Voraussetzungen (Produktmessbarkeit) vertauschen kann, siehe Satz 9.1. □

Bemerkung (Stieltjesintegral)**.** Das Lebesgue-Stieltjes-Integral $\int h\, dF$ einer messbaren Funktion $h : \mathbb{R} \to [0, \infty]$ bzgl. der Verteilungsfunktion F einer Wahrscheinlichkeitsverteilung μ auf $\mathbb{R}$ ist definiert als das Lebesgueintegral

$$\int h(t)\, dF(t) \quad := \quad \int h(t)\, \mu(dt).$$

Ist h stetig, dann lässt sich das Integral als Limes von Riemannsummen darstellen. Nach dem Transformationssatz gilt für eine Zufallsvariable $T : \Omega \to [0, \infty]$:

$$E[T] \quad = \quad \int t\, \mu_T(dt) \quad = \quad \int t\, dF_T(t).$$

Die Aussage von Satz 6.11 folgt hieraus formal durch partielle Integration.

Beispiel (Exponentialverteilung). Für eine exponentialverteilte Zufallsvariable T mit Parameter $\lambda > 0$ erhalten wir:

$$E[T] \quad = \quad \int_0^\infty P[T > t]\, dt \quad = \quad \int_0^\infty e^{-\lambda t}\, dt \quad = \quad \frac{1}{\lambda}.$$

Es gilt also

$$\text{Mittlere Wartezeit} \quad = \quad \frac{1}{\text{Mittlere relative Häufigkeit pro Zeiteinheit}}$$

.

Beispiel (Heavy tails). Sei $\alpha > 0$. Für eine Zufallsvariable $T : \Omega \to [0, \infty)$ mit

$$P[T > t] \quad \sim \quad t^{-\alpha} \qquad \text{für } t \to \infty$$

gilt

$$E[T] \quad = \quad \int_0^\infty P[T > t]\, dt \quad < \quad \infty$$

genau dann, wenn $\alpha > 1$. Allgemeiner ist das p**-te Moment**

$$E[T^p] \quad = \quad \int_0^\infty P[T^p > t]\, dt \quad = \quad \int_0^\infty \underbrace{P[T > t^{1/p}]}_{\sim t^{\alpha/p}}\, dt$$

nur für $p < \alpha$ endlich.

Zufallsvariablen mit Dichten

Die Verteilungen vieler Zufallsvariablen haben eine Dichte bzgl. des Lebesguemaßes, oder bzgl. eines anderen geeigneten Referenzmaßes. Wir wollen uns nun überlegen, wie man in diesem Fall den Erwartungswert berechnet.

Sei $(S, \mathcal{S})$ ein messbarer Raum und ν ein Maß auf $(S, \mathcal{S})$ (z.B. das Lebesguemaß oder eine Wahrscheinlichkeitsverteilung).

Definition. *Eine **Wahrscheinlichkeitsdichte** auf $(S, \mathcal{S}, \nu)$ ist eine messbare Funktion $\varrho : S \to [0, \infty]$ mit*

$$\int_S \varrho(x)\, \nu(dx) \;=\; 1.$$

Satz 6.12. *(1). Ist ϱ eine Wahrscheinlichkeitsdichte auf $(S, \mathcal{S}, \nu)$, dann wird durch*

$$\mu[B] \quad := \quad \int_B \varrho(x)\, \nu(dx) \quad = \quad \int I_B(x) \varrho(x)\, \nu(dx) \tag{6.2.2}$$

eine Wahrscheinlichkeitsverteilung μ auf $(S, \mathcal{S})$ definiert.

(2). Für eine messbare Funktion $h : S \to [0, \infty]$ gilt

$$\int h(x)\, \mu(dx) \quad = \quad \int h(x) \varrho(x)\, \nu(dx). \tag{6.2.3}$$

Insbesondere folgt nach dem Transformationssatz:

$$E[h(X)] \quad = \quad \int h(x) \varrho(x)\, \nu(dx)$$

für jede Zufallsvariable X mit Verteilung μ.

Beweis. Wir zeigen zunächst, dass μ eine Wahrscheinlichkeitsverteilung ist: Sind $B_1, B_2, ... \in \mathcal{S}$ disjunkt, so folgt

$$\begin{aligned}
\mu\left[\bigcup_{i=1}^{\infty} B_i\right] &= \int I_{\bigcup_{i=1}^{\infty} B_i}(x) \cdot \varrho(x)\, \nu(dx) \\
&= \lim_{n\to\infty} \int I_{\bigcup_{i=1}^{n} B_i}(x) \cdot \varrho(x)\, \nu(dx) \quad \text{(wegen } \varrho \geq 0 \text{ und monotoner Konvergenz)} \\
&= \lim_{n\to\infty} \sum_{i=1}^{n} \int_{B_i} \varrho(x)\, \nu(dx) \quad = \quad \lim_{n\to\infty} \sum_{i=1}^{n} \mu[B_i] \\
&= \sum_{i=1}^{\infty} \mu[B_i].
\end{aligned}$$

Zudem gilt:

$$\mu[S] \quad = \quad \int \varrho\, d\nu \quad = \quad 1.$$

Die Aussage (6.2.3) über den Erwartungswert beweisen wir durch maßtheoretische Induktion:

(1). Die Aussage folgt unmittelbar, wenn $h = I_B$ für $B \in \mathcal{S}$ gilt.

(2). Für Linearkombinationen $h = \sum_{i=1}^{n} c_i I_{B_i}$ folgt die Aussage aus der Linearität beider Seiten von (6.2.3) in h.

(3). Für allgemeine $h \geq 0$ existiert eine Teilfolge h_n aus Elementarfunktionen mit $h_n \nearrow h$. Mit monotoner Konvergenz folgt

$$\int h\, d\mu \quad = \quad \lim \int h_n\, d\mu \quad = \quad \lim \int h_n \varrho\, d\nu \quad = \quad \int h\varrho\, d\nu.$$

□

Bemerkung. Durch (6.2.2) wird die Dichte $\varrho(x)$ der Wahrscheinlichkeitsverteilung μ bzgl. des Maßes ν für ν-fast alle x eindeutig festgelegt: Existiert $\widetilde{\varrho} \in \mathcal{L}^1(S, \mathcal{S}, \nu)$ mit

$$\int_B \varrho\, d\nu \quad = \quad \mu[B] \quad = \quad \int_B \widetilde{\varrho}\, d\nu \qquad \text{für alle } B \in \mathcal{S},$$

dann folgt:

$$\begin{aligned}
\int_{\{\varrho > \widetilde{\varrho}\}} (\varrho - \widetilde{\varrho})\, d\nu \;&=\; \int_{\{\varrho < \widetilde{\varrho}\}} (\varrho - \widetilde{\varrho})\, d\nu \quad = \quad 0, \qquad \text{also} \\
\int (\varrho - \widetilde{\varrho})^+\, d\nu \;&=\; \int (\varrho - \widetilde{\varrho})^-\, d\nu \;=\; 0.
\end{aligned}$$

Somit erhalten wir:

$$(\varrho - \widetilde{\varrho})^+ = (\varrho - \widetilde{\varrho})^- = 0 \quad \nu\text{-fast überall,}$$

und damit $\varrho = \widetilde{\varrho} \quad \nu$-fast überall.

Notation: Die Aussage (6.2.3) rechtfertigt die folgende Notation für eine Wahrscheinlichkeitsverteilung μ mit Dichte ϱ bzgl. ν:

$$\mu(dx) \quad = \quad \varrho(x)\,\nu(dx) \qquad \text{bzw.} \qquad d\mu \quad = \quad \varrho\,d\nu \qquad \text{bzw.} \qquad \mu \quad = \quad \varrho \cdot \nu.$$

Für die nach der Bemerkung ν-fast überall eindeutig bestimmte Dichte von μ bzgl. ν verwenden wir dementsprechend auch die folgende Notation:

$$\varrho(x) \quad = \quad \frac{d\mu}{d\nu}(x).$$

Wichtige Spezialfälle:

(1). MASSENFUNKTION ALS DICHTE BZGL. DES ZÄHLMASSES.
Das Zählmaß auf einer abzählbaren Menge S ist das durch

$$\nu[B] \quad = \quad |B|, \qquad B \subseteq S,$$

definierte Maß auf S. Die Gewichtsfunktion $x \mapsto \mu[\{x\}]$ einer Wahrscheinlichkeitsverteilung μ auf S ist die Dichte von μ bzgl. des Zählmaßes ν. Insbesondere ist die Massenfunktion einer diskreten Zufallsvariable $X : \Omega \to S$ die Dichte der Verteilung von X bzgl. ν:

$$\mu_X[B] \quad = \quad P[X \in B] \quad = \quad \sum_{a \in B} p_X(a) \quad = \quad \int_B p_X(a)\nu(da), \qquad \text{für alle } B \subseteq S.$$

Die Berechnungsformel für den Erwartungswert diskreter Zufallsvariablen ergibt sich damit als Spezialfall von Satz 6.12:

$$E[h(X)] \quad \overset{6.12}{=} \quad \int h(a)p_X(a)\,\nu(da) \quad = \quad \sum_{a \in S} h(a)p_X(a) \qquad \text{für alle } h : S \to [0,\infty].$$

(2). DICHTEN BZGL. DES LEBESGUEMASSES
Eine Wahrscheinlichkeitsverteilung μ auf $\mathbb{R}^d$ mit Borelscher σ-Algebra hat genau dann eine Dichte ϱ bzgl. des Lebesguemaßes λ, wenn

$$\mu[(-\infty, c_1] \times \ldots \times (-\infty, c_d]] \quad = \quad \int_{-\infty}^{c_1} \cdots \int_{-\infty}^{c_d} \varrho(x_1, \ldots, x_d)\,\mathrm{d}x_d \cdots \mathrm{d}x_1$$

für alle $(c_1, \ldots, c_d) \in \mathbb{R}^d$ gilt. Insbesondere hat die Verteilung einer reellwertigen Zufallsvariable X genau dann die Dichte f_X bzgl. λ, wenn

$$F_X(c) \quad = \quad \mu_X[(-\infty, c]] \quad = \quad \int_{-\infty}^{c} f_X(x)\,\mathrm{d}x \qquad \text{für alle } c \in \mathbb{R}$$

gilt. Die Verteilungsfunktion ist in diesem Fall eine Stammfunktion der Dichte, und damit λ-fast überall differenzierbar mit Ableitung

$$F_X'(x) \quad = \quad f_X(x) \qquad \text{für fast alle } x \in \mathbb{R}.$$

Für den Erwartungswert ergibt sich:

$$E[h(X)] \quad = \quad \int_{\mathbb{R}} h(x) f_X(x) \, \mathrm{d}x$$

für alle messbaren Funktionen $h : \mathbb{R} \to \mathbb{R}$ mit $h \geq 0$ oder $h \in \mathcal{L}^1(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu)$.

Beispiel (Normalverteilungen). Die Dichte der Standardnormalverteilung bzgl. des Lebesguemaßes ist $\varrho(x) = (2\pi)^{-1/2} \cdot e^{-x^2/2}$. Damit ergibt sich für den Erwartungswert und die Varianz einer Zufallsvariable $Z \sim N(0,1)$:

$$\begin{aligned} E[Z] &= \int_{-\infty}^{\infty} x \cdot (2\pi)^{-1/2} \cdot e^{-x^2/2} \, dx \quad = \quad 0, \qquad \text{und} \\ \mathrm{Var}[Z] &= E[(Z - E[Z])^2] \quad = \quad E[Z^2] \\ &= \int_{-\infty}^{\infty} x^2 \cdot (2\pi)^{-1/2} \cdot e^{-x^2/2} \, dx \\ &= \int_{-\infty}^{\infty} 1 \cdot (2\pi)^{-1/2} \cdot e^{-x^2/2} \, dx \quad = \quad 1. \end{aligned}$$

Hierbei haben wir im letzten Schritt partielle Integration benutzt.
Ist X eine $N(m, \sigma^2)$-verteilte Zufallsvariable, dann ist $Z = \frac{X-m}{\sigma}$ standardnormalverteilt, und es gilt $X = m + \sigma Z$, also

$$E[X] \quad = \quad m + \sigma E[Z] \quad = \quad m,$$

und

$$\mathrm{Var}[X] \quad = \quad \mathrm{Var}[\sigma Z] \quad = \quad \sigma^2 \, \mathrm{Var}[Z] \quad = \quad \sigma^2.$$

Die Parameter m und σ geben also den Erwartungswert und die Standardabweichung der Normalverteilung an.

(3). RELATIVE DICHTEN: Seien μ und ν zwei Wahrscheinlichkeitsverteilungen auf einem messbaren Raum $(S, \mathcal{S})$ mit Dichten f bzw. g bezüglich eines Referenzmaßes λ (z.B. Zählmaß oder Lebesguemaß). Gilt $g > 0$ λ-fast überall, dann hat μ bzgl. ν die Dichte

$$\frac{d\mu}{d\nu} \quad = \quad \frac{f}{g} \quad = \quad \frac{d\mu/d\lambda}{d\nu/d\lambda},$$

denn nach Satz 6.12 gilt:

$$\begin{aligned}\mu[B] &= \int_B f\, d\lambda \quad = \quad \int_B \frac{f}{g} g\, d\lambda \\ &= \int_B \frac{f}{g}\, d\nu \qquad \text{für alle } B \in \mathcal{S}.\end{aligned}$$

In der Statistik treten relative Dichten als „Likelihoodquotienten“ auf, wobei $f(x)$ bzw. $g(x)$ die „Likelihood“ eines Beobachtungswertes x bzgl. verschiedener möglicher zugrundeliegender Wahrscheinlichkeitsverteilungen beschreibt, s. Abschnitt 9.1.

Existenz von Dichten

Wir geben abschließend ohne Beweis den Satz von Radon-Nikodym an. Dieser Satz besagt, dass eine Wahrscheinlichkeitsverteilung (oder allgemeiner ein σ-endliches Maß) μ genau dann eine Dichte bzgl. eines anderen (σ-endlichen) Maßes ν hat, wenn alle ν-Nullmengen auch μ-Nullmengen sind. Ein Maß μ auf einem messbaren Raum $(S, \mathcal{S})$ heißt **σ-endlich**, wenn eine Folge von messbaren Mengen $B_n \in \mathcal{S}$ mit $\mu[B_n] < \infty$ und $S = \bigcup_{n\in\mathbb{N}} B_n$ existiert.

Definition. *(1). Ein Maß μ auf $(S, \mathcal{S})$ heißt **absolutstetig** bzgl. eines anderen Maßes ν auf demselben messbaren Raum ($\mu \ll \nu$) falls für alle $B \in \mathcal{S}$ gilt:*

$$\nu[B] = 0 \qquad \Longrightarrow \qquad \mu[B] = 0$$

*(2). Die Maße μ und ν heißen **äquivalent** ($\mu \sim \nu$), falls $\mu \ll \nu$ und $\nu \ll \mu$.*

Beispiel. Ein Diracmaß $\delta_x, x \in \mathbb{R}$, ist nicht absolutstetig bzgl. das Lebesguemaßes λ auf $\mathbb{R}$, denn es gilt $\lambda[\{x\}] = 0$, aber $\delta_x[\{x\}] > 0$. Umgekehrt ist auch das Lebesguemaß nicht absolutstetig bzgl. des Diracmaßes.

Satz 6.13 (Radon-Nikodym). *Für σ-endliche Maße μ und ν gilt $\mu \ll \nu$ genau dann, wenn eine Dichte $\varrho \in \mathcal{L}^1(S, \mathcal{S}, \nu)$ existiert mit*

$$\mu[B] \quad = \quad \int_B \varrho\, d\nu \qquad \textit{für alle } B \in \mathcal{S}.$$

Die eine Richtung des Satzes zeigt man leicht: Hat μ eine Dichte bzgl. ν, und gilt $\nu[B] = 0$, so folgt

$$\mu[B] \quad = \quad \int_B \varrho\, d\nu \quad = \quad \int \varrho \cdot I_B\, d\nu \quad = \quad 0,$$

da $\varrho \cdot I_B = 0$ ν-fast überall. Der Beweis der Umkehrung ist nicht so einfach, und kann funktionalanalytisch erfolgen, siehe z.B. Klenke: „Wahrscheinlichkeitstheorie“. Einen stochastischen Beweis über Martingaltheorie werden wir in der Vorlesung „Stochastische Prozesse“ führen.

Beispiel (Absolutstetigkeit von diskreten Wahrscheinlichkeitsverteilungen). Sind μ und ν Wahrscheinlichkeitsverteilungen (oder σ-endliche Maße) auf einer abzählbaren Menge S, dann gilt $\mu \ll \nu$ genau dann, wenn $\mu(x) = 0$ für alle $x \in S$ mit $\nu(x) = 0$ gilt. In diesem Fall ist die Dichte von μ bzgl. ν durch

$$\frac{d\mu}{d\nu}(x) \quad = \quad \begin{cases} \dfrac{\mu(x)}{\nu(x)} & \text{falls } \nu(x) \neq 0 \\ \text{beliebig} & \text{sonst} \end{cases}$$

gegeben. Man beachte, dass die Dichte nur für ν-fast alle x, also für alle x mit $\nu(x) \neq 0$, eindeutig bestimmt ist.

6.3 Varianz, Kovarianz und lineare Regression

Varianz und Standardabweichung

Sei $X : \Omega \to \overline{\mathbb{R}}$ eine integrierbare Zufallsvariable auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$. Wie zuvor für diskrete Zufallsvariablen (s. Abschnitt 3.1) definieren wir auch im allgemeinen Fall die *Varianz* $\mathrm{Var}[X]$ und die Standardabweichung $\sigma[X]$ durch

$$\mathrm{Var}[X] \quad := \quad E[(X - E[X])^2], \qquad \sigma[X] \quad := \quad \sqrt{\mathrm{Var}[X]}.$$

Auch in diesem Fall folgen aus der Linearität des Erwartungswerts die Rechenregeln

$$\mathrm{Var}[X] \;=\; E[X^2] - E[X]^2, \qquad \text{und} \tag{6.3.1}$$

$$\mathrm{Var}[aX + b] \;=\; \mathrm{Var}[aX] \;=\; a^2 \cdot \mathrm{Var}[X] \qquad \text{für alle } a, b \in \mathbb{R}. \tag{6.3.2}$$

Insbesondere ist die Varianz genau dann endlich, wenn $E[X^2]$ endlich ist. Nach Korollar 6.5 gilt zudem genau dann $\mathrm{Var}[X] = 0$, wenn X P-f.s. konstant gleich $E[X]$ ist. Aufgrund des Transformationssatzes für den Erwartungswert können wir die Varianz auch allgemein aus der Verteilung $\mu_X = P \circ X^{-1}$ berechnen:

Korollar 6.14. *Die Varianz $\mathrm{Var}[X]$ hängt nur von der Verteilung $\mu_X = P \circ X^{-1}$ ab:*

$$\mathrm{Var}[X] \quad = \quad \int (x - \overline{x})^2 \, \mu_X(dx) \qquad \textit{mit} \qquad \overline{x} \;=\; E[X] \;=\; \int x \, \mu(dx).$$

Beweis. Nach Satz 6.12 gilt

$$\mathrm{Var}[X] \quad = \quad E[(X - E[X])^2] \quad = \quad \int (x - E[X])^2 \mu_X(dx)$$

mit $E[X] = \int x \mu_X(dx)$. □

Beispiel (Empirisches Mittel und empirische Varianz). Ist die zugrundeliegende Wahrscheinlichkeitsverteilung auf Ω eine empirische Verteilung

$$P \quad = \quad \frac{1}{n} \sum_{i=1}^{n} \delta_{\omega_i}$$

von n Elementen $\omega_1, \ldots, \omega_n$ aus einer Grundgesamtheit (z.B. alle Einwohner von Bonn, oder eine Stichprobe daraus), dann ist die Verteilung einer Abbildung $X : \Omega \to S$ (statistisches Merkmal, z.B. Alter der Einwohner von Bonn) gerade die empirische Verteilung der auftretenden Werte $x_i = X(\omega_i)$:

$$\mu_X \quad = \quad \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}.$$

Die Gewichte der empirischen Verteilung sind die relativen Häufigkeiten

$$\mu_X(a) \quad = \quad \frac{h(a)}{n}, \qquad h(a) \quad = \quad |\{1 \le i \le n : x_i = a\}|.$$

Für den Erwartungswert einer Funktion $g(X), g : S \to \mathbb{R}$, ergibt sich

$$E[g(X)] \quad = \quad \sum_{a \in \{x_1, \ldots, x_n\}} g(a) \cdot \frac{h(a)}{n} \quad = \quad \frac{1}{n} \sum_{i=1}^{n} g(x_i),$$

d.h. der Erwartungswert bzgl. der empirischen Verteilung ist das arithmetische Mittel der Werte $g(x_i)$.

Ist X reellwertig, so erhalten wir als Erwartungswert und Varianz das **empirische Mittel**

$$E[X] \quad = \quad \sum_{a \in \{x_1, \ldots, x_n\}} a \cdot \frac{h(a)}{n} \quad = \quad \frac{1}{n} \sum_{i=1}^{n} x_i \quad =: \quad \overline{x}_n,$$

und die **empirische Varianz**

$$\begin{aligned} \mathrm{Var}[X] \quad &= \quad E[(X - E[X])^2] \quad = \quad \sum_{a \in \{x_1, \ldots, x_n\}} (a - \overline{x}_n)^2 \cdot \frac{h(a)}{n} \\ &= \quad \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x}_n)^2 \quad = \overline{(x^2)}_n - (\overline{x}_n)^2 \quad =: \quad \sigma_n^2. \end{aligned}$$

Sind die x_i selbst unabhängige Stichproben von einer Wahrscheinlichkeitsverteilung μ, dann ist die empirische Verteilung $n^{-1} \sum_{i=1}^{n} \delta_{x_i}$ nach dem Gesetz der großen Zahlen eine Approximation von μ, siehe Abschnitt 7.2 unten. Daher verwendet man das Stichprobenmittel $\overline{x}_n$ und die Stichprobenvarianz σ_n^2 bzw. die renormierte Stichprobenvarianz

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x}_n)^2$$

in der Statistik, um den Erwartungswert und die Varianz einer zugrundeliegenden (unbekannten) Verteilung zu schätzen.

Beispiel (Exponentialverteilung). Für eine zum Parameter $\lambda > 0$ exponentialverteilte Zufallsvariable T gilt $E[T] = \frac{1}{\lambda}$. Mit partieller Integration folgt zudem:

$$\begin{aligned} E[T^2] &= \int_0^\infty t^2 f_T(t)\, dt &&= \int_0^\infty t^2 \lambda e^{-\lambda t}\, dt \\ &= \int_0^\infty 2te^{-\lambda t}\, dt &&= \frac{2}{\lambda} \int_0^\infty t f_T(t)\, dt \\ &= \frac{2}{\lambda} E[T] &&= \frac{2}{\lambda^2}, \end{aligned}$$

also

$$\sigma(T) = \sqrt{\mathrm{Var}[T]} = (E[T^2] - E[T]^2)^{1/2} = \frac{1}{\lambda}.$$

Die Standardabweichung ist also genauso groß wie der Erwartungswert!

Beispiel (Heavy Tails). Eine Zufallsvariable $X : \Omega \to \mathbb{R}$ mit Verteilungsdichte

$$f_X(x) \sim |x|^{-p} \qquad \text{für } |x| \to \infty$$

ist integrierbar für $p > 2$. Für $p \in (2, 3]$ gilt jedoch

$$\mathrm{Var}[X] = \int_{-\infty}^{\infty} (x - E[X])^2 f_X(x)\, dx = \infty.$$

Quadratintegrierbare Zufallsvariablen

Für einen gegebenen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ bezeichnen wir mit $\mathcal{L}^2(\Omega, \mathcal{A}, P)$ den Raum aller bezüglich P quadratintegrierbaren Zufallsvariablen:

$$\mathcal{L}^2(\Omega, \mathcal{A}, P) = \{X : \Omega \to \overline{\mathbb{R}} \text{ messbar } \mid \ E[X^2] < \infty\}.$$

Der Raum ist ein Unterraum des Vektorraums aller $\mathcal{A}/\mathcal{B}(\overline{\mathbb{R}})$ messbaren Abbildungen, denn für $X, Y \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$ und $a \in \mathbb{R}$ gilt:

$$E[(aX+Y)^2] \quad \leq \quad E[2(aX)^2 + 2Y^2] \quad = \quad 2a^2 E[X^2] + 2E[Y^2] \quad < \quad \infty.$$

Zudem gilt

$$\mathcal{L}^2(\Omega, \mathcal{A}, P) \quad \subseteq \quad \mathcal{L}^1(\Omega, \mathcal{A}, P),$$

denn aus $|X| \leq (X^2+1)/2$ folgt

$$E[|X|] \quad \leq \quad E\left[\frac{1}{2}(X^2+1)\right] \quad = \quad \frac{1}{2}(E[X^2]+1) \quad < \quad \infty$$

für alle $X \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$. Hierbei haben wir wesentlich benutzt, dass P ein endliches Maß ist - für unendliche Maße ist der Raum $\mathcal{L}^2$ *nicht* in $\mathcal{L}^1$ enthalten! Nach (6.3.1) ist umgekehrt eine Zufallsvariable aus $\mathcal{L}^1$ genau dann in $\mathcal{L}^2$ enthalten, wenn sie endliche Varianz hat.

Auf dem Vektorraum

$$L^2(\Omega, \mathcal{A}, P) \quad = \quad \mathcal{L}^2(\Omega, \mathcal{A}, P)/\sim$$

der Äquivalenzklassen von P-fast sicher gleichen quadratintegrierbaren Zufallsvariablen wird durch

$$(X,Y)_{L^2} \quad := \quad E[XY] \qquad \text{und} \qquad \|X\|_{L^2} \quad := \quad (X,X)_{L^2}^{1/2}$$

ein Skalarprodukt und eine Norm definiert. Hierbei ist der Erwartungswert $E[XY]$ wegen $|XY| \leq (X^2+Y^2)/2$ definiert. Insbesondere gilt die **Cauchy-Schwarz-Ungleichung**

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

In der Analysis wird gezeigt, dass $L^2(\Omega, \mathcal{A}, P)$ bzgl. des L^2-Skalarprodukts ein Hilbertraum, also vollständig bzgl. der L^2-Norm ist.

Beste Prognosen

Angenommen wir wollen den Ausgang eines Zufallsexperiments vorhersagen, dass durch eine reellwertige Zufallsvariable $X : \Omega \to \mathbb{R}$ beschrieben wird. Welches ist der beste Prognosewert a für $X(\omega)$, 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**quare **e**rror)

$$\text{MSE} \quad = \quad E[(X-a)^2]$$

bzw. die Wurzel (**r**oot **m**ean **s**quare **e**rror)

$$\text{RMSE} \quad = \quad \text{MSE}^{1/2} \quad = \quad \|X-a\|_{L^2(\Omega,\mathcal{A},P)}.$$

Satz 6.15 (Erwartungswert als bester L^2-Prognosewert). *Ist $X \in \mathcal{L}^2(\Omega,\mathcal{A},P)$, dann gilt für alle $a \in \mathbb{R}$:*

$$E[(X-a)^2] \quad = \quad \text{Var}[X] + (a-E[X])^2 \quad \geq \quad E[(X-E[X])^2]$$

Der mittlere quadratische Fehler des Prognosewertes a ist also die Summe der Varianz von X und des Quadrats des **Bias** (systematischer bzw. mittlerer Prognosefehler) $a - E[X]$:

$$\text{MSE} \quad = \quad \text{Varianz} + \text{Bias}^2.$$

Insbesondere ist der mittlere quadratische Fehler genau für $a = E[X]$ minimal.

Beweis. Für $a \in \mathbb{R}$ gilt wegen der Linearität des Erwartungswertes:

$$\begin{aligned} E[(X-a)^2] &= E[(X-E[X]+E[X]-a)^2] \\ &= E[(X-E[X])^2] + 2\underbrace{E[(X-E[X])\cdot(E[X]-a)]}_{=\underbrace{(E[X]-E[X])}_{=0}\cdot(E[X]-a)} + E[(E[X]-a)^2] \\ &= \text{Var}[X] + (E[X]-a)^2. \end{aligned}$$

□

Verwendet man eine andere Norm, um den Prognosefehler zu messen, dann ergeben sich im Allgemeinen andere beste Prognosewerte. Beispielsweise gilt:

Satz 6.16 (Median als bester L^1-Prognosewert). *Ist $X \in L^1(\Omega,\mathcal{A},P)$ und m ein Median der Verteilung von X, dann gilt für alle $a \in \mathbb{R}$:*

$$E[|X-a|] \quad \geq \quad E[|X-m|]$$

.

Beweis. Für $m \geq a$ folgt die Behauptung aus der Identität

$$|X-m| - |X-a| \quad \leq \quad (m-a)(I_{(-\infty,m)}(X) - I_{[m,\infty)}(X))$$

durch Bilden des Erwartungswertes. Der Beweis für $m \leq a$ verläuft analog. □

Insbesondere minimieren Stichprobenmittel und Stichprobenmedian einer Stichprobe $x_1,\ldots,x_n \in \mathbb{R}$ also die Summe der quadratischen bzw. absoluten Abweichungen $\sum(x_i-a)^2$ bzw. $\sum|x_i-a|$.

Kovarianz und Korrelation

Seien X und Y quadratintegrierbare reellwertige Zufallsvariablen, die auf einem gemeinsamen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ definiert sind. Wie schon für diskrete Zufallsvariablen definieren wir wieder die **Kovarianz**

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

und den **Korrelationskoeffizienten**

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

falls $\sigma[X] \cdot \sigma[Y] \neq 0$. Die Zufallsvariablen X und Y heißen unkorreliert, falls $\mathrm{Cov}[X,Y] = 0$ gilt, d.h. falls

$$E[XY] \quad = \quad E[X] \cdot E[Y].$$

Um die Kovarianz zu berechnen, benötigen wir die gemeinsame Verteilung der Zufallsvariablen X und Y. Aus dem Transformationssatz für den Erwartungswert folgt:

Korollar 6.17. *Die Kovarianz $\mathrm{Cov}[X,Y]$ hängt nur von der gemeinsamen Verteilung*

$$\mu_{X,Y} \quad = \quad P \circ (X,Y)^{-1}$$

der Zufallsvariablen X und Y ab:

$$\mathrm{Cov}[X,Y] \quad = \quad \int \left(x - \int z \, \mu_X(dz) \right) \left(y - \int z \, \mu_Y(dz) \right) \mu_{X,Y}(dx\, dy).$$

Beweis. Nach dem Transformationssatz gilt

$$\begin{aligned} \mathrm{Cov}[X,Y] &= E[(X - E[X])(Y - E[Y])] \\ &= \int \left(x - \int z \, \mu_X(dz) \right) \left(y - \int z \, \mu_Y(dz) \right) \mu_{X,Y}(dx\, dy). \end{aligned}$$

□

Aus der Linearität des Erwartungswertes folgt, dass die Abbildung $\mathrm{Cov} : \mathcal{L}^2 \times \mathcal{L}^2 \to \mathbb{R}$ symmetrisch und bilinear ist. Die Varianz $\mathrm{Var}[X] = \mathrm{Cov}[X,X]$ ist die zugehörige quadratische Form. Insbesondere gilt wie im diskreten Fall:

$$\mathrm{Var}\left[\sum_{i=1}^n X_i \right] \quad = \quad \sum_{i=1}^n \mathrm{Var}[X_i] + 2 \cdot \sum_{\substack{i,j=1 \\ i<j}}^n \mathrm{Cov}[X_i, X_j].$$

Sind die Zufallsvariablen $X_1, \ldots, X_n$ unkorreliert, dann folgt:

$$\operatorname{Var}[X_1 + \ldots + X_n] \quad = \quad \sum_{i=1}^{n} \operatorname{Var}[X_i].$$

Die folgende Aussage ist ein Spezialfall der Cauchy-Schwarz-Ungleichung. Wir geben trotzdem einen vollständigen Beweis, da dieser auch in Zusammenhang mit linearer Regression von Interesse ist.

Satz 6.18 (Cauchy-Schwarz)**.** *(1). Für $X, Y \in \mathcal{L}^2$ gilt:*

$$|\operatorname{Cov}[X,Y]| \quad \leq \quad \operatorname{Var}[X]^{1/2} \cdot \operatorname{Var}[Y]^{1/2} \quad = \quad \sigma[X] \cdot \sigma[Y]. \tag{6.3.3}$$

(2). Im Fall $\sigma[X] \cdot \sigma[Y] \neq 0$ gilt für den Korrelationskoeffizienten

$$|\varrho[X,Y]| \quad \leq \quad 1. \tag{6.3.4}$$

Gleichheit in (6.3.3) bzw. (6.3.4) gilt genau dann, wenn ein $a \neq 0$ und ein $b \in \mathbb{R}$ existieren, sodass $Y = aX + b$ P-fast sicher gilt. Hierbei ist $\varrho[X,Y] = 1$ im Falle $a > 0$ und $\varrho[X,Y] = -1$ für $a < 0$.

Beweis. Im Fall $\sigma[X] = 0$ gilt $X = E[X]$ P-fast sicher, und die Ungleichung (6.3.3) ist trivialerweise erfüllt. Wir nehmen nun an, dass $\sigma[X] \neq 0$ gilt.

(1). Für $a \in \mathbb{R}$ gilt:

$$\begin{aligned} 0 \quad &\leq \quad \operatorname{Var}[Y - aX] \quad = \quad \operatorname{Var}[Y] - 2a \operatorname{Cov}[X,Y] + a^2 \operatorname{Var}[X] \\ &= \quad \left(a \cdot \sigma[X] - \frac{\operatorname{Cov}[X,Y]}{\sigma[X]} \right)^2 - \frac{\operatorname{Cov}[X,Y]^2}{\operatorname{Var}[X]} + \operatorname{Var}[Y]. \end{aligned} \tag{6.3.5}$$

Da der erste Term für $a := \frac{\operatorname{Cov}[X,Y]}{\sigma[X]^2}$ verschwindet, folgt:

$$\operatorname{Var}[Y] - \frac{\operatorname{Cov}[X,Y]^2}{\operatorname{Var}[X]} \quad \geq \quad 0.$$

(2). Die Ungleichung $|\varrho[X,Y]| \leq 1$ folgt unmittelbar aus (6.3.3). Zudem gilt genau dann Gleichheit in (6.3.5) bzw. (6.3.3), wenn $\operatorname{Var}[Y - aX] = 0$ gilt, also $Y - aX$ P-fast sicher konstant ist. In diesem Fall folgt

$$\operatorname{Cov}[X,Y] \quad = \quad \operatorname{Cov}[X, aX] \quad = \quad a \operatorname{Var}[X],$$

also hat $\varrho[X,Y]$ dasselbe Vorzeichen wie a.

□

Beispiel (**Empirischer Korrelationskoeffizent**). Ist die zugrundeliegende Wahrscheinlichkeitsverteilung eine empirische Verteilung $P = \frac{1}{n}\sum_{i=1}^{n}\delta_{\omega_i}$, und sind $X, Y : \Omega \to \mathbb{R}$ reellwertige Abbildungen (statistische Merkmale), dann gilt

$$\mu_{X,Y} \quad = \quad \frac{1}{n}\sum_{i=1}^{n}\delta_{(x_i,y_i)} \qquad \text{mit} \quad x_i = X(\omega_i) \text{ und } y_i = Y(\omega_i).$$

Als Kovarianz ergibt sich

$$\mathrm{Cov}[X,Y] \quad = \quad \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)(y_i - \overline{y}_n) \quad = \quad \frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i\right) - \overline{x}_n\overline{y}_n.$$

Der entsprechende **empirische Korrelationskoeffizient** der Daten (x_i, y_i), $1 \le i \le n$, ist

$$\varrho[X,Y] \quad = \quad \frac{\mathrm{Cov}[X,Y]}{\sigma[X]\sigma[Y]} \quad = \quad \frac{\sum_{i=1}^{n}(x_i - \overline{x}_n)(y_i - \overline{y}_n)}{\left(\sum_{i=1}^{n}(x_i - \overline{x}_n)^2\right)^{1/2}\left(\sum_{i=1}^{n}(y_i - \overline{y}_n)^2\right)^{1/2}} \quad =: \quad r_n.$$

Den empirischen Korrelationskoeffizienten verwendet man als Schätzer für die Korrelation von Zufallsgrößen mit unbekannten Verteilungen.

Die Grafiken 6.3 und 6.3 zeigen Stichproben mit verschiedenen Korrelationskoeffizienten ϱ.

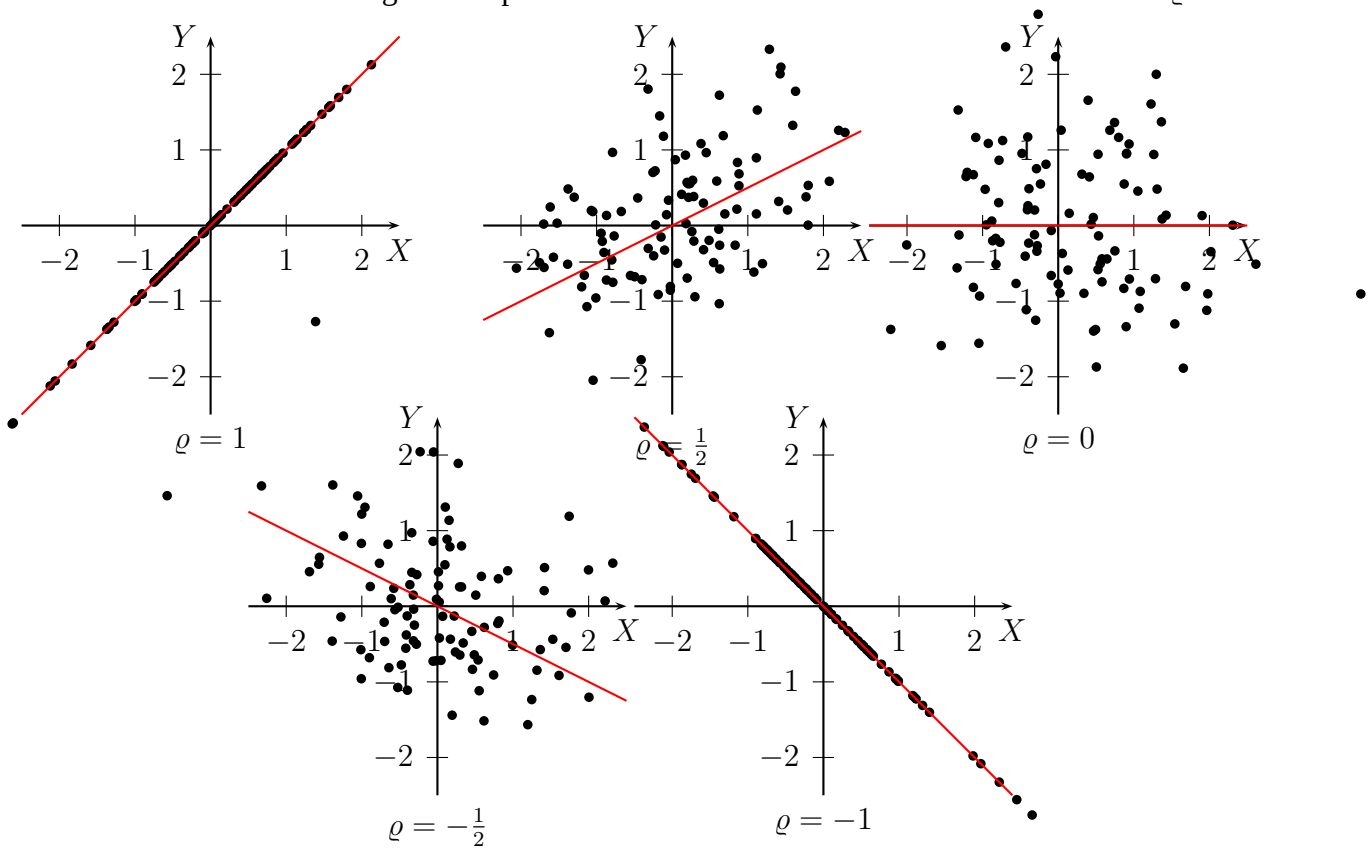


Abbildung 6.2: Stichprobe von 100 Punkten von korrelierten Standardnormalverteilungen

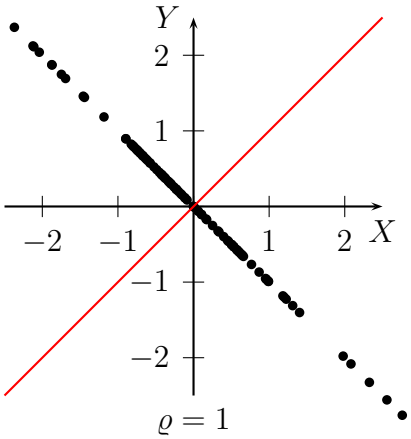


Abbildung 6.3: Stichprobe von 100 Punkten von korrelierten Standardnormalverteilungen

Anwendung auf lineare Prognose (Regression)

Seien $X, Y \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$ Zufallsvariablen mit $\sigma[X] \neq 0$. Angenommen, wir kennen den Wert $X(\omega)$ in einem Zufallsexperiment und suchen die beste *lineare* Vorhersage

$$\hat{Y}(\omega) \quad = \quad aX(\omega) + b, \qquad (a, b \in \mathbb{R}) \tag{6.3.6}$$

für $Y(\omega)$ im quadratischen Mittel, d.h. den Minimierer des mittleren quadratischen Fehlers,

$$\text{MSE} \quad := \quad E[(\hat{Y} - Y)^2],$$

unter alle Zufallsvariablen $\hat{Y}$, die affine Funktionen von X sind.

Korollar 6.19. *Der mittlere quadratische Fehler ist minimal unter allen Zufallsvariablen $\hat{Y} = aX + b$ $(a, b \in \mathbb{R})$ für*

$$\hat{Y}(\omega) \quad = \quad E[Y] + \frac{\text{Cov}[X, Y]}{\text{Var}[X]} \cdot (X(\omega) - E[X]).$$

Beweis. Es gilt

$$\begin{aligned} \text{MSE} &= \text{Var}[Y - \hat{Y}] + E[Y - \hat{Y}]^2 \\ &= \text{Var}[Y - aX] + (E[Y] - aE[X] - b)^2. \end{aligned}$$

Der zweite Term ist minimal für

$$b \quad = \quad E[Y] - aE[X],$$

und der erste Term für

$$a \quad = \quad \frac{\text{Cov}[X, Y]}{\sigma[X]^2},$$

siehe den Beweis der Cauchy-Schwarz-Ungleichung, Satz 6.18. Die bzgl. des mittleren quadratischen Fehlers optimale Prognose für Y gestützt auf X ist also

$$\hat{Y}_{\text{opt}} \quad = \quad aX + b \quad = \quad E[Y] + a(X - E[X]).$$

□

Beispiel (Regressionsgerade, Methode der kleinsten Quadrate). Im Beispiel der empirischen Verteilung von oben erhalten wir die Regressionsgerade $y = ax + b$, die die Quadratsumme

$$\sum_{i=1}^{n} (ax_i + b - y_i)^2 \quad = \quad n \cdot \text{MSE}$$

der Abweichungen minimiert. Es gilt

$$a \quad = \quad \frac{\text{Cov}[X,Y]}{\sigma[X]^2} \quad = \quad \frac{\sum_{i=1}^{n} (x_i - \overline{x}_n)(y_i - \overline{y}_n)}{\sum_{i=1}^{n} (x_i - \overline{x}_n)^2}$$

und

$$b \quad = \quad E[Y] - a \cdot E[X] \quad = \quad \overline{y}_n - a \cdot \overline{x}_n.$$

Die Regressionsgeraden sind in Grafik 6.3 eingezeichnet.

Beispiel (Zweidimensionale Normalverteilung). Die zweidimensionale Normalverteilung $N(m, C)$ ist die Verteilung im $\mathbb{R}^2$ mit Dichte

$$f_{m,C}(x) \quad = \quad \frac{1}{2\pi \cdot \sqrt{\det C}} \cdot \exp\left(-\frac{1}{2} (x - m) \cdot C^{-1} (x - m) \right), \qquad x \in \mathbb{R}^2.$$

Hierbei ist $m \in \mathbb{R}^2$ und $C = \begin{pmatrix} v_1 & c \\ c & v_2 \end{pmatrix}$ eine symmetrische positiv-definite Matrix mit Koeffizienten $c \in \mathbb{R}$ und $v_1, v_2 > 0$. Mit $\sigma_i := \sqrt{v_i}, i = 1, 2$, und $\varrho := \frac{c}{\sigma_1 \sigma_2}$ gilt:

$$\det C \quad = \quad v_1 v_2 - c^2 \quad = \quad \sigma_1^2 \sigma_2^2 \cdot (1 - \varrho^2), \qquad \text{und}$$

$$C^{-1} \quad = \quad \frac{1}{\det C} \begin{pmatrix} v_2 & -c \\ -c & v_1 \end{pmatrix} \quad = \quad \frac{1}{1 - \varrho^2} \cdot \begin{pmatrix} \frac{1}{\sigma_1^2} & -\frac{\varrho}{\sigma_1 \sigma_2} \\ -\frac{\varrho}{\sigma_1 \sigma_2} & \frac{1}{\sigma_2^2} \end{pmatrix},$$

also

$$f_{m,C}(x) \quad = \quad \frac{\exp\left(-\frac{1}{2(1-\varrho^2)} \left[\left(\frac{x_1 - m_1}{\sigma_1} \right)^2 - 2\varrho \frac{x_1 - m_1}{\sigma_1} \cdot \frac{x_2 - m_2}{\sigma_2} + \left(\frac{x_2 - m_2}{\sigma_2} \right)^2 \right] \right)}{2\pi \sigma_1 \sigma_2 \sqrt{1 - \varrho^2}}.$$

Die folgende Aussage zeigt, dass die Koeffizienten m_i, σ_i und ϱ tatsächlich der Mittelwert, die Standardabweichung und die Korrelation der Koordinaten x_1 und x_2 sind:

Behauptung:

(1). $f_{m,C}$ ist eine Wahrscheinlichkeitsdichte bzgl. des Lebesguemaßes im $\mathbb{R}^2$.

(2). Für reellwertige Zufallsvariablen X_1, X_2 mit gemeinsamer Verteilung $\mu_{X_1,X_2} = N(m, C)$ und $i = 1, 2$ gilt

$$E[X_i] \;=\; m_i, \qquad \mathrm{Var}[X_i] \;=\; v_i, \qquad \text{und} \qquad \mathrm{Cov}[X_1, X_2] \;=\; c, \tag{6.3.7}$$

d.h. m ist der Mittelwertvektor und $C = (\mathrm{Cov}[X_i, X_j])_{i,j}$ die Kovarianzmatrix der Normalverteilung $N(m, C)$.

Der Beweis der Behauptung wird der Leserin/dem Leser als Übung überlassen - wir zeigen nur exemplarisch die Berechnung der Kovarianz im Fall $m = 0$. Mit quadratischer Ergänzung können wir den Exponenten in der Dichte $f_{0,C}(x)$ schreiben als

$$-\frac{1}{2(1-\varrho^2)}\left(\frac{x_1}{\sigma_1} - \varrho\frac{x_2}{\sigma_2}\right)^2 - \frac{1}{2}\left(\frac{x_2}{\sigma_2}\right)^2.$$

Mit $\widetilde{m}(x_2) = \dfrac{x_2\varrho\sigma_1}{\sigma_2}$ erhalten wir dann nach dem Satz von Fubini:

$$\begin{aligned}
&\int_{\mathbb{R}^2} x_1 x_2 f_{0,C}(x)\, dx \\
&= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\varrho^2}} \int\int x_1 x_2 \exp[-\frac{1}{2(1-\varrho^2)\sigma_1^2}(x_1 - \widetilde{m}(x_2))^2]\, dx_1 \exp\left(-\frac{x_2^2}{2\sigma_2^2}\right)\, dx_2 \\
&= \frac{1}{\sqrt{2\pi\sigma_2^2}} \int \underbrace{x_2 \cdot \widetilde{m}(x_2)}_{x_2^2\varrho\sigma_1/\sigma_2} \cdot \exp\left(-\frac{x_2^2}{2\sigma_2^2}\right)\, dx_2 \quad = \quad \varrho\sigma_1\sigma_2 \quad = \quad c,
\end{aligned}$$

wobei wir im zweiten und dritten Schritt die Formeln für den Erwartungswert und die Varianz von eindimensionalen Normalverteilungen verwendet haben. Nach dem Transformationssatz ergibt sich:

$$E[X_1 X_2] \quad = \quad \int x_1 x_2\, \mu_{X_1,X_2}(dx) \quad = \quad c.$$

Da auf ähnliche Weise $E[X_1] = E[X_2] = 0$ folgt, ist c die Kovarianz von X_1 und X_2.

Bemerkung. Ist $X = (X_1, X_2)$ ein $N(m, C)$-verteilter Zufallsvektor, dann ist jede Linearkombination $Y = \alpha_1 X_1 + \alpha_2 X_2$, $\alpha \in \mathbb{R}^2$, normalverteilt mit Mittelwert $\alpha \cdot m$ und Varianz $\alpha \cdot C\alpha$. Auch dies kann man durch eine explizite Berechnung der Verteilungsfunktion aus der gemeinsamen Dichte von X_1 und X_2 zeigen. Wir werden multivariate Normalverteilungen systematischer in Abschnitt 9.3 untersuchen, und dort auch einen eleganteren Beweis der letzten Aussage mithilfe von charakteristischen Funktionen geben.

Beispiel (Autoregressiver Prozess). Seien X_0 und $Z_n, n \in \mathbb{N}$, unabhängige reellwertige Zufallsvariablen mit $Z_n \sim N(0,1)$ für alle n. Der durch das „stochastische Bewegungsgesetz“

$$X_n \quad = \quad \underbrace{\alpha X_{n-1}}_{\substack{\text{lineares} \\ \text{Bewegungsgesetz}}} \quad + \quad \underbrace{\varepsilon Z_n}_{\substack{\text{zufällige Störung,} \\ \text{Rauschen}}} \quad , \qquad n \in \mathbb{N}, \tag{6.3.8}$$

definierte stochastische Prozess $(X_n)_{n=0,1,2,\ldots}$ heißt **autoregressiver Prozess AR(1)** mit Parametern $\varepsilon, \alpha \in \mathbb{R}$. Autoregressive Prozesse werden zur Modellierung von Zeitreihen eingesetzt. Im allgemeineren autoregressiven Modell AR(p), $p \in \mathbb{N}$, mit Parametern $\varepsilon, \alpha_1, \ldots, \alpha_p \in \mathbb{R}$ lautet das Bewegungsgesetz

$$X_n \quad = \quad \sum_{i=1}^{p} \alpha_i X_{n-i} + \varepsilon Z_n, \qquad n \geq p.$$

Grafik 6.3 zeigt simulierte Trajektorien von AR(1)- und AR(2)-Prozessen:

Das folgende Lemma fasst einige grundlegende Eigenschaften des AR(1) Modells zusammen.

Lemma 6.20. *Für den AR(1)-Prozess mit Parametern ε, α und $m \in \mathbb{R}, \sigma > 0$ gilt:*

(1). $X_{n-1} \sim N(m, \sigma^2) \quad \Longrightarrow \quad X_n \sim N(\alpha m, \alpha^2\sigma^2 + \varepsilon^2)$.

(2). Für $|\alpha| < 1$ ist die Verteilung $\mu = N(0, \frac{\varepsilon^2}{1-\alpha^2})$ ein Gleichgewicht, d.h.

$$X_0 \sim \mu \quad \Longrightarrow \quad X_n \sim \mu \qquad \forall n \in \mathbb{N}.$$

Bei Startverteilung $P \circ X_0^{-1} = \mu$ gilt:

$$\mathrm{Cov}[X_n, X_{n-k}] \quad = \quad \alpha^k \cdot \frac{\varepsilon^2}{1-\alpha^2} \qquad \textit{für alle } 0 \leq k \leq n.$$

Exponentieller Abfall der Korrelationen

Beweis. Gilt $X_{n-1} \sim N(m, \sigma^2)$, dann ist (X_{n-1}, Z_n) bivariat normalverteilt, also ist auch die Linearkombination $X_n = aX_{n-1} + \varepsilon Z_n$ normalverteilt. Der Erwartungswert und die Varianz von X_n ergeben sich aus (6.3.7). Der Beweis der übrigen Aussagen wird dem Leser als Übungsaufgabe überlassen. □

Bemerkung. (1). Der AR(1)-Prozess ist eine *Markovkette* mit Übergangswahrscheinlichkeiten $p(x, \cdot) = N(\alpha x, \varepsilon^2)$, s. Abschnitt 9.1 unten.

(2). Ist die gemeinsame Verteilung der Startwerte $X_0, X_1, \ldots, X_{p-1}$ eine multivariate Normalverteilung, dann ist der AR(p)-Prozess ein *Gaussprozess*, d.h. die gemeinsame Verteilung von $X_0, X_1, \ldots, X_n$ ist für jedes $n \in \mathbb{N}$ eine multivariate Normalverteilung.

Unabhängigkeit und Unkorreliertheit

Wir zeigen abschließend, dass auch für allgemeine Zufallsvariablen X und Y aus Unabhängigkeit die Unkorreliertheit von beliebigen Funktionen $f(X)$ und $g(Y)$ folgt. Seien $X : \Omega \to S$ und $Y : \Omega \to T$ Zufallsvariablen mit Werten in messbaren Räumen $(S, \mathcal{S})$ und $(T, \mathcal{T})$.

Satz 6.21. *Es sind äquivalent:*

(1). Die Zufallsvariablen X und Y sind unabhängig, d.h.

$$P[X \in A, Y \in B] \quad = \quad P[X \in A] \cdot P[Y \in B] \qquad \textit{für alle } A \in \mathcal{S} \textit{ und } B \in \mathcal{T}$$

(2). Die Zufallsvariablen $f(X)$ und $g(Y)$ sind unkorreliert für alle messbaren Funktionen f, g mit $f, g \geq 0$ bzw. $f(X), g(Y) \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$, d.h.

$$E[f(X) \cdot g(Y)] \quad = \quad E[f(X)] \cdot E[g(Y)]. \tag{6.3.9}$$

Beweis. Offensichtlich folgt (1) aus (2) durch Wahl von $f = I_A$ und $g = I_B$. Die umgekehrte Implikation folgt durch maßtheoretische Induktion: Gilt (1), dann ist (6.3.9) für Indikatorfunktionen f und g erfüllt. Wegen der Linearität beider Seiten dieser Gleichung in f und g gilt (6.3.9) auch für beliebige Elementarfunktionen. Für messbare $f, g \geq 0$ betrachten wir Folgen von Elementarfunktionen f_n, g_n mit $f_n \nearrow f, g_n \nearrow g$. Die Aussage (6.3.9) folgt durch monotone Konvergenz. Allgemeine Funktionen zerlegen wir in ihren Positiv- und Negativanteil, und wenden die Aussage auf diese an. Also gilt $\mathrm{Cov}[f(X), g(Y)] = 0$ für alle messbaren f, g mit $f, g \geq 0$ bzw. $f(X), g(Y) \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$. □

Korollar 6.22. *Sind $X, Y \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ unabhängig, so gilt:*

$$X \cdot Y \quad \in \quad \mathcal{L}^1(\Omega, \mathcal{A}, P) \qquad \textit{und} \qquad E[XY] \quad = \quad E[X] \cdot E[Y].$$

Beweis. Nach Satz 6.21 gilt:

$$E[|XY|] \quad = \quad E[|X|] \cdot E[|Y|] \quad < \quad \infty.$$

Die Formel für $E[XY]$ folgt durch die Zerlegungen $X = X^+ - X^-$ und $Y = Y^+ - Y^-$. □

Kapitel 7

Gesetze der großen Zahlen

In diesem Kapitel beweisen wir verschiedene Gesetze der großen Zahlen, d.h. wir leiten Bedingungen her, unter denen die Mittelwerte $\frac{1}{n}\sum_{i=1}^{n} X_i$ einer Folge $(X_i)_{i\in\mathbb{N}}$ von reellwertigen Zufallsvariablen gegen ihren Erwartungswert konvergieren. Dabei unterscheiden wir verschiedene Arten der Konvergenz, die wir zunächst genauer untersuchen wollen.

7.1 Grundlegende Ungleichungen und Konvergenz von Zufallsvariablen

Konvergenzbegriffe für Zufallsvariablen

Seien $Y_n, n \in \mathbb{N},$ und Y reellwertige Zufallsvariablen, die auf einem gemeinsamen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ definiert sind. Wir betrachten die folgenden Konvergenzbegriffe für die Folge $(Y_n)_{n\in\mathbb{N}}$:

Definition. *(1). **Fast sichere Konvergenz:***

Die Folge $(Y_n)_{n\in\mathbb{N}}$ konvergiert P-fast sicher gegen Y, falls gilt:

$$P\left[\lim_{n\to\infty} Y_n = Y\right] \quad = \quad P[\{\omega \in \Omega | Y_n(\omega) \to Y(\omega)\}] \quad = \quad 1.$$

*(2). **Stochastische Konvergenz** (Convergence in probability):*

Die Folge $(Y_n)_{n\in\mathbb{N}}$ konvergiert P-stochastisch gegen Y (Notation $Y_n \xrightarrow{P} Y$), falls

$$\lim_{n\to\infty} P[|Y_n - Y| > \varepsilon] \quad = \quad 0 \qquad \textit{für alle } \varepsilon > 0 \textit{ gilt.}$$

(3). ***$\mathcal{L}^p$-Konvergenz*** *($1 \leq p < \infty$):*
Die Folge $(Y_n)_{n\in\mathbb{N}}$ konvergiert in $\mathcal{L}^p(\Omega, \mathcal{A}, P)$ gegen Y, falls

$$\lim_{n\to\infty} E[|Y_n - Y|^p] \quad = \quad 0.$$

Ein Gesetz der großen Zahlen bezüglich fast sicherer Konvergenz heißt **starkes Gesetz der großen Zahlen**, ein G.d.g.Z. bezüglich stochastischer Konvergenz heißt **schwaches Gesetz der großen Zahlen**. Wir wollen nun die Zusammenhänge zwischen den verschiedenen Konvergenzbegriffen untersuchen.

Satz 7.1. *(1). Fast sichere Konvergenz impliziert stochastische Konvergenz.*

(2). Die umgekehrte Implikation gilt im Allgemeinen nicht.

Beweis. (1). Konvergiert Y_n P-fast sicher gegen Y, dann gilt für $\varepsilon > 0$:

$$\begin{aligned}
1 &= P[|Y_n - Y| < \varepsilon \quad \text{schließlich}] \\
&= P\left[\bigcup_m \bigcap_{n\geq m} \{|Y_n - Y| < \varepsilon\}\right] \\
&= \lim_{m\to\infty} P\left[\bigcap_{n\geq m} \{|Y_n - Y| < \varepsilon\}\right] \\
&\leq \lim_{m\to\infty} \inf_{n\geq m} P[|Y_n - Y| < \varepsilon] \\
&= \liminf_{n\to\infty} P[|Y_n - Y| < \varepsilon].
\end{aligned}$$

Es folgt $\lim_{n\to\infty} P[|Y_n - Y| < \varepsilon] = 1$ für alle $\varepsilon > 0$, d.h. Y_n konvergiert auch P-stochastisch gegen Y.

(2). Sei andererseits P das Lebesguemaß auf $\Omega = (0, 1]$ mit Borelscher σ-Algebra. Wir betrachten die Zufallsvariablen

$$Y_1 = I_{(0,1]}, Y_2 = I_{(0,\frac{1}{2}]}, Y_3 = I_{(\frac{1}{2},1]}, Y_4 = I_{(0,\frac{1}{4}]}, Y_5 = I_{(\frac{1}{4},\frac{1}{2}]}, Y_6 = I_{(\frac{1}{2},\frac{3}{4}]}, Y_6 = I_{(\frac{3}{4},1]}, \ldots$$

Dann gilt

$$P[|Y_n| > \varepsilon] \quad = \quad P[Y_n = 1] \quad \to \quad 0 \qquad \text{für alle } \varepsilon > 0,$$

also konvergiert Y_n stochastisch gegen 0, obwohl

$$\limsup Y_n(\omega) \quad = \quad 1 \qquad \text{für alle } \omega \in \Omega \text{ gilt.}$$

□

Hier ist ein weiteres Beispiel, das den Unterschied zwischen stochastischer und fast sicherer Konvergenz zeigt:

Beispiel. Sind $T_1, T_2, \ldots$ unter P unabhängige $\mathrm{Exp}(1)$-verteilte Zufallsvariablen, dann konvergiert $T_n/\log n$ P-stochastisch gegen 0, denn

$$P\left[\left|\frac{T_n}{\log n}\right| \geq \varepsilon\right] \quad = \quad P[T_n \geq \varepsilon \cdot \log n] \quad = \quad n^{-\varepsilon} \quad \stackrel{n\to\infty}{\to} \quad 0$$

für alle $\varepsilon > 0$. Andererseits gilt nach (5.1.6) aber

$$\limsup_{n\to\infty} \frac{T_n}{\log n} \quad = \quad 1 \qquad P\text{-fast sicher,}$$

also konvergiert $T_n/\log n$ nicht P-fast sicher.

Obwohl die stochastische Konvergenz selbst nicht fast sichere Konvergenz impliziert, kann man aus einer Verschärfung von stochastischer Konvergenz die fast sichere Konvergenz schließen. Wir sagen, dass eine Folge $Y_n, n \in \mathbb{N}$, von Zufallsvariablen auf $(\Omega, \mathcal{A}, P)$ **schnell stochastisch** gegen Y **konvergiert**, falls

$$\sum_{n=1}^{\infty} P[|Y_n - Y| \geq \varepsilon] \quad < \quad \infty \qquad \text{für alle } \varepsilon > 0.$$

Lemma 7.2. *Aus schneller stochastischer Konvergenz folgt fast sichere Konvergenz.*

Beweis. Wir können o.B.d.A. $Y = 0$ annehmen. Konvergiert Y_n schnell stochastisch gegen 0, dann gilt:

$$P[\limsup |Y_n| \leq \varepsilon] \quad \geq \quad P[|Y_n| \geq \varepsilon \text{ nur endlich oft}] \quad = \quad 1.$$

Es folgt

$$P[\limsup |Y_n| \neq 0] \quad = \quad P\left[\bigcup_{\varepsilon\in\mathbb{Q}_+} \{\limsup |Y_n| > \varepsilon\}\right] \quad = \quad 0.$$

□

Ähnlich zeigt man:

Lemma 7.3. *Konvergiert Y_n P-stochastisch gegen Y, dann existiert eine Teilfolge Y_{n_k}, die P-fast sicher gegen Y konvergiert.*

Beweis. Wieder können wir o.B.d.A. $Y = 0$ annehmen. Konvergiert Y_n stochastisch gegen 0, dann existiert eine Teilfolge Y_{n_k} mit

$$P\left[|Y_{n_k}| \geq \frac{1}{k}\right] \quad \leq \quad \frac{1}{k^2}.$$

Nach dem Lemma von Borel-Cantelli folgt

$$P\left[|Y_{n_k}| \geq \frac{1}{k} \text{ nur endlich oft}\right] \quad = \quad 1,$$

also $Y_{n_k} \to 0$ P-fast sicher. □

Als nächstes beweisen wir eine Erweiterung der Čebyšev-Ungleichung, die wir an vielen Stellen verwenden werden. Insbesondere impliziert sie, dass stochastische Konvergenz schwächer ist als $\mathcal{L}^p$-Konvergenz.

Die Markov-Čebyšev-Ungleichung

Sei $X : \Omega \to \overline{\mathbb{R}}$ eine Zufallsvariable auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$. Wir verwenden die folgende Notation:

Notation: $E[X\,;\,A] := E[X \cdot I_A] = \int_A X dP$.

Satz 7.4 (Allgemeine Markov-Ungleichung). *Sei $h : [0,\infty] \to [0,\infty]$ monoton wachsend und Borel-messbar. Dann gilt*

$$P[|X| \geq c] \;\leq\; \frac{E[h(|X|)\,;\,|X| \geq c]}{h(c)} \;\leq\; \frac{E[h(|X|)]}{h(c)} \qquad \textit{für alle } c > 0 \textit{ mit } h(c) \neq 0.$$

Beweis. Da h nichtnegativ und monoton wachsend ist, gilt

$$h(|X|) \;\geq\; h(|X|) \cdot I_{\{|X| \geq c\}} \;\geq\; h(c) \cdot I_{\{|X| \geq c\}},$$

also auch

$$E[h(|X|)] \;\geq\; E[h(|X|)\,;\,|X| \geq c] \;\geq\; h(c) \cdot P[|X| \geq c].$$

□

Wichtige Spezialfälle:

(1). **Markov - Ungleichung:** Für $h(x) = x$ erhalten wir:

$$P[|X| \geq c] \;\leq\; \frac{E[|X|]}{c} \qquad \text{für alle } \; c > 0.$$

Insbesondere gilt für eine Zufallsvariable X mit $E[|X|] = 0$:

$$P[|X| \geq c] = 0 \qquad \text{für alle } \; c > 0,$$

also auch $P[|X| > 0] = 0$, d.h. $X = 0$ P-fast sicher.

(2). **Čebyšev - Ungleichung:** Für $h(x) = x^2$ und $X = Y - E[Y]$ mit $Y \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ erhalten wir:

$$P[|Y - E[Y]| \geq c] \leq \frac{E[(Y - E[Y])^2]}{c^2} = \frac{\mathrm{Var}[Y]}{c^2} \qquad \text{für alle } c > 0.$$

Diese Ungleichung haben wir bereits in Abschnitt 3.2 im Beweis des schwachen Gesetzes der großen Zahlen verwendet.

(3). **Exponentielle Abschätzung:** Für $h(x) = \exp(tx)$ mit $t > 0$ erhalten wir wegen $I_{\{X \geq c\}} \leq e^{-tc} e^{tX}$:

$$P[X \geq c] = E[I_{\{X \geq c\}}] \leq e^{-tc} \cdot E[e^{tX}].$$

Die Abbildung $t \mapsto E[e^{tX}]$ heißt **momentenerzeugende Funktion** der Zufallsvariablen X. Exponentielle Ungleichungen werden wir in Abschnitt 8.2 zur Kontrolle der Wahrscheinlichkeiten *großer Abweichungen* vom Gesetz der großen Zahlen verwenden.

Als erste Anwendung der allgemeinen Markovungleichung zeigen wir für reellwertige Zufallsvariablen X, X_n $(n \in \mathbb{N})$:

Korollar 7.5 ($\mathcal{L}^p$-Konvergenz impliziert stochastische Konvergenz). *Für $1 \leq p < \infty$ gilt:*

$$E[|X_n - X|^p] \to 0 \quad \Rightarrow \quad P[|X_n - X| > \varepsilon] \to 0 \quad \textit{für alle } \varepsilon > 0.$$

Beweis. Nach der Markovungleichung mit $h(x) = x^p$ gilt:

$$P[|X_n - X| \geq \varepsilon] \quad \leq \quad \frac{1}{\varepsilon^p} E[|X_n - X|^p].$$

□

Bemerkung. Aus stochastischer Konvergenz folgt im Allgemeinen nicht $\mathcal{L}^p$-Konvergenz (Übung). Es gilt aber: Konvergiert $X_n \to X$ stochastisch, und ist die Folge der Zufallsvariablen $|X_n|^p$ $(n \in \mathbb{N})$ **gleichmäßig integrierbar**, d.h.

$$\sup_{n \in \mathbb{N}} E[|X_n|^p \,;\, |X_n| \geq c] \to 0 \quad \text{für } c \to \infty,$$

dann konvergiert X_n gegen X in $\mathcal{L}^p$ *(Verallgemeinerter Satz von Lebesgue)*. Wir benötigen diese Aussage im Moment nicht, und werden sie daher erst in der Vorlesung »Stochastische Prozesse« beweisen.

Als nächstes wollen wir den Zusammenhang zwischen $\mathcal{L}^p$-Konvergenz für verschiedene Werte von $p \geq 1$ untersuchen. Dazu verwenden wir eine weitere fundamentale Ungleichung:

Die Jensensche Ungleichung

Ist $\ell(x) = ax + b$ eine affine Funktion auf $\mathbb{R}$, und $X \in \mathcal{L}^1$ eine integrierbare Zufallsvariable, dann folgt aus der Linearität des Lebesgueintegrals:

$$E[\ell(X)] \quad = \quad E[aX + b] \quad = \quad aE[X] + b \quad = \quad \ell(E[X]) \tag{7.1.1}$$

Da konvexe Funktionen Suprema einer Familie von affinen Funktionen (nämlich der Tangenten an den Funktionsgraphen der konvexen Funktion) sind, ergibt sich für konvexe Funktionen eine entsprechende *Ungleichung*:

Satz 7.6 (Jensensche Ungleichung). *Ist P eine Wahrscheinlichkeitsverteilung, $X \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ eine reellwertige Zufallsvariable, und $h : \mathbb{R} \to \mathbb{R}$ eine konvexe Abbildung, dann ist $E[h(X)^-] < \infty$, und es gilt*

$$h(E[X]) \ \leq \ E[h(X)].$$

Warnung: Diese Aussage gilt (wie auch (7.1.1)) nur für die Integration bzgl. eines Wahrscheinlichkeitsmaßes!
Bevor wir die Jensensche Ungleichung beweisen, erinnern wir kurz an die Definition und elementare Eigenschaften von konvexen Funktionen:

Bemerkung. Eine Funktion $h : \mathbb{R} \to \mathbb{R}$ ist genau dann konvex, wenn

$$h(\lambda x + (1-\lambda)y) \quad \leq \quad \lambda h(x) + (1-\lambda)h(y) \quad \text{für alle } \lambda \in [0,1] \text{ und } x, y \in \mathbb{R}$$

gilt, d.h. wenn alle Sekanten oberhalb des Funktionsgraphen liegen.

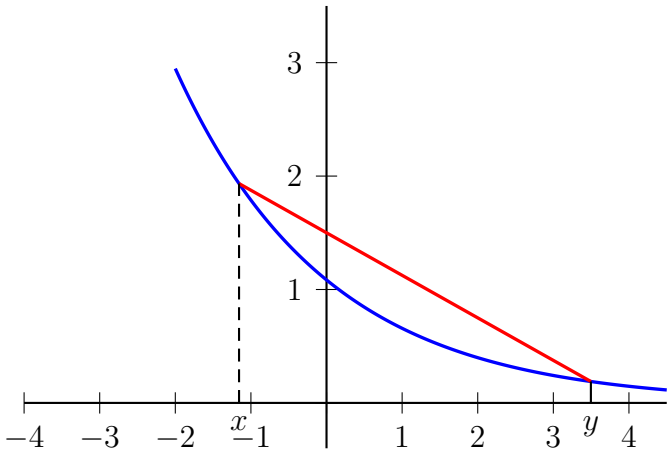


Abbildung 7.1: Sekante an konvexer Funktion

Hieraus folgt, dass jede konvexe Funktion stetig ist: Für $a < b < x < y < c < d$ gilt nämlich

$$\frac{h(b) - h(a)}{b - a} \quad \leq \quad \frac{h(y) - h(x)}{y - x} \quad \leq \quad \frac{h(d) - h(c)}{d - c}.$$

Also sind die Differenzenquotienten $\frac{h(y)-h(x)}{y-x}$ gleichmäßig beschränkt auf (b, c), und somit ist h gleichmäßig stetig auf (b, c). Da konvexe Funktionen stetig sind, sind sie auch messbar. Die Existenz des Erwartungswertes $E[h(X)]$ in $(-\infty, \infty]$ folgt dann aus $E[h(X)^-] < \infty$.

Wir beweisen nun die Jensensche Ungleichung:

Beweis. Ist h konvex, dann existiert zu jedem $x_0 \in \mathbb{R}$ eine affine Funktion ℓ (*Stützgerade*) mit $\ell(x_0) = h(x_0)$ und $\ell \leq h$, siehe die Analysis Vorlesung oder [A. KLENKE: „WAHRSCHEINLICHKEITSTHEORIE“, Abschnitt 7.2].

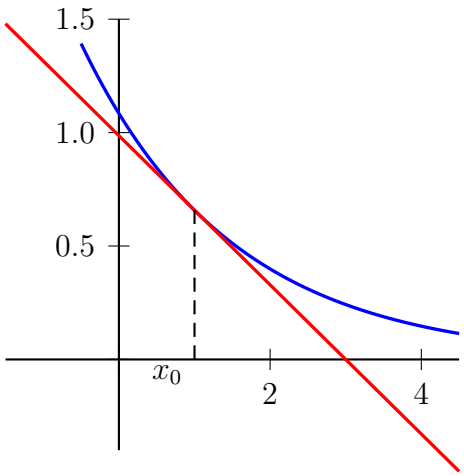


Abbildung 7.2: Darstellung von $\ell(x)$ und $h(x)$

Wählen wir $x_0 := E[X]$, dann folgt

$$h(E[X]) = \ell(E[X]) = E(\ell[X]) \leq E[h(X)].$$

Der Erwartungswert auf der rechten Seite ist definiert, da $h(X)$ durch die integrierbare Zufallsvariable $\ell(X)$ nach unten beschränkt ist. Insbesondere gilt $E[h(X)^-] \leq E[\ell(X)^-] < \infty$. □

Korollar 7.7 ($\mathcal{L}^q$-Konvergenz impliziert $\mathcal{L}^p$-Konvergenz). *Für $1 < p \leq q$ gilt:*

$$\|X\|_p \quad := \quad E[|X|^p]^{\frac{1}{p}} \quad \leq \quad \|X\|_q.$$

Insbesondere folgt $\mathcal{L}^p$-Konvergenz aus $\mathcal{L}^q$-Konvergenz.

Beweis. Nach der Jensenschen Ungleichung gilt

$$E[|X|^p]^{\frac{q}{p}} \quad \leq \quad E[|X|^q],$$

da die Funktion $h(x) = |x|^{q/p}$ für $q \geq p$ konvex ist. □

Nach dem Korollar gilt für $p \leq q$:

$$\mathcal{L}^p(\Omega, \mathcal{A}, P) \quad \supseteq \quad \mathcal{L}^q(\Omega, \mathcal{A}, P),$$

und

$$X_n \to X \text{ in } \mathcal{L}^q \quad \Rightarrow \quad X_n \to X \text{ in } \mathcal{L}^p.$$

Man beachte, dass diese Aussage nur für **endliche Maße** wahr ist, da im Beweis die Jensensche Ungleichung verwendet wird.
Mithilfe der Jensenschen Ungleichung beweist man auch die **Hölderungleichung**:

$$E[|XY|] \quad \leq \quad \|X\|_p \cdot \|Y\|_q \qquad \text{für } p, q \in [1, \infty] \text{ mit} \qquad \frac{1}{p} + \frac{1}{q} \quad = \quad 1.$$

7.2 Starke Gesetze der großen Zahlen

Wir werden nun Gesetze der großen Zahlen unter verschiedenen Voraussetzungen an die zugrundeliegenden Zufallsvariablen beweisen. Zunächst nehmen wir an, dass $X_1, X_2, \ldots \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$ quadratintegrierbare Zufallsvariablen sind, deren Varianzen gleichmäßig beschränkt sind, und deren Korrelationen hinreichend schnell abklingen:

Annahme: *„Schnelles Abklingen der positiven Korrelation“*

(A) Es existiert eine Folge $c_n \in \mathbb{R}_+$ $(n \in \mathbb{N})$ mit

$$\sum_{n=0}^{\infty} c_n < \infty$$

und

$$\operatorname{Cov}[X_i, X_j] \quad \leq \quad c_{|i-j|} \qquad \text{für alle } i, j \in \mathbb{N}. \tag{7.2.1}$$

Die Bedingung (A) ist insbesondere erfüllt, wenn die *Korrelationen exponentiell abfallen*, d.h. wenn

$$|\operatorname{Cov}[X_i, X_j]| \quad \leq \quad c \cdot \alpha^{|i-j|}$$

für ein $\alpha \in (0, 1)$ und $c \in \mathbb{R}^+$ gilt. Sind etwa die Zufallsvariablen X_i unkorreliert, und ist die Folge der *Varianzen beschränkt*, d.h. gilt

(A1) $\mathrm{Cov}[X_i, X_j] = 0$ für alle $i, j \in \mathbb{N}$, und

(A2) $v := \sup_i \mathrm{Var}[X_i] < \infty$,

dann ist die Annahme (A) mit $c_0 = v$ und $c_n = 0$ für $n > 0$ erfüllt. In diesem Fall haben wir bereits in Abschnitt 3.2 ein schwaches Gesetz der großen Zahlen bewiesen.

Wichtig: Es wird **keine Unabhängigkeit vorausgesetzt!**

Sei nun

$$S_n \quad = \quad X_1 + \ldots + X_n$$

die Summe der ersten n Zufallsvariablen.

Das schwache Gesetz der großen Zahlen

Den Beweis des schwachen Gesetzes der großen Zahlen aus Abschnitt 3.2 können wir auf den hier betrachteten allgemeinen Fall erweitern:

Satz 7.8 (Schwaches Gesetz der großen Zahlen, $\mathcal{L}^2$-Version)**.** *Unter der Voraussetzung (A) gilt für alle $n \in \mathbb{N}$ und $\varepsilon > 0$:*

$$E\left[\left(\frac{S_n}{n} - \frac{E[S_n]}{n}\right)^2\right] \leq \frac{v}{n}, \qquad \textit{und} \tag{7.2.2}$$

$$P\left[\left|\frac{S_n}{n} - \frac{E[S_n]}{n}\right| \geq \varepsilon\right] \leq \frac{v}{\varepsilon^2 n} \tag{7.2.3}$$

mit $v := c_0 + 2 \cdot \sum_{n=1}^{\infty} c_n < \infty$. Gilt insbesondere $E[X_i] = m$ für alle $i \in \mathbb{N}$, dann folgt

$$\frac{S_n}{n} \quad \to \quad m \qquad \textit{in } \mathcal{L}^2(\Omega, \mathcal{A}, P) \textit{ und } P\textit{-stochastisch.}$$

Beweis. Unter Verwendung der Voraussetzung an die Kovarianzen erhalten wir

$$\begin{aligned} E\left[\left(\frac{S_n}{n} - \frac{E[S_n]}{n}\right)^2\right] &= \mathrm{Var}\left[\frac{S_n}{n}\right] = \frac{1}{n^2}\mathrm{Var}[S_n] \\ &= \frac{1}{n^2}\sum_{i,j=1}^{n}\mathrm{Cov}[X_i, X_j] \leq \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} c_{|i-j|} \\ &\leq \frac{1}{n^2}\sum_{i=1}^{n}\sum_{k=-\infty}^{\infty} c_{|k|} = \frac{v}{n} \end{aligned}$$

Die zweite Behauptung folgt daraus durch Anwenden der Čebyšev-Ungleichung. □

Bemerkung. (1). Im Fall unkorrelierter Zufallsvariablen X_i (Annahmen (A1) und (A2)) ist die Aussage ein Spezialfall einer allgemeinen funktionalanalytischen Sachverhalts:

Das Mittel von beschränkten orthogonalen Vektoren im Hilbertraum
$L^2(\Omega, \mathcal{A}, P) = \mathcal{L}^2(\Omega, \mathcal{A}, P)/\sim$ *konvergiert gegen* 0.

Unkorreliertheit der X_i bedeutet gerade, dass die Zufallsvariablen

$$Y_i \quad := \quad X_i - E[X_i]$$

orthogonal in L^2 sind - beschränkte Varianzen der X_i ist gleichbedeutend mit der Beschränktheit der L^2 Normen der Y_i. Es gilt

$$S_n - E[S_n] \quad = \quad \sum_{i=1}^{n} Y_i,$$

also

$$\begin{aligned} E\left[\left(\frac{S_n}{n} - \frac{E[S_n]}{n}\right)^2\right] &= \left\|\frac{1}{n}\sum_{i=1}^{n} Y_i\right\|_{L^2}^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\langle Y_i, Y_j\rangle_{L^2} \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\|Y_i\|_{L^2}^2 \quad \leq \quad \frac{1}{n}\sup_i \|Y_i\|_{L^2}^2. \end{aligned}$$

(2). Die $\mathcal{L}^2$-Konvergenz und stochastische Konvergenz von $(S_n - E[S_n])/n$ gegen 0 gilt auch, falls die Korrelationen „langsam“ abklingen, d.h. falls (7.2.1) für eine nicht summierbare Nullfolge c_n erfüllt ist. In diesem Fall erhält man allerdings im Allgemeinen keine Abschätzung der Ordnung $O(\frac{1}{n})$ für den Fehler in (7.2.2) bzw. (7.2.3).

(3). Eine für große n deutlich bessere Abschätzung des Fehlers in (7.2.3) (mit exponentiellem Abfall in n) erhält man bei Unabhängigkeit und exponentieller Integrierbarkeit der X_i mithilfe der *exponentiellen Ungleichung*, siehe Satz 8.3 unten.

Das starke Gesetz für quadratintegrierbare Zufallsvariablen

Unter derselben Voraussetzung wie in Satz 7.8 gilt sogar P-fast sichere Konvergenz:

Satz 7.9 (Starkes Gesetz großer Zahlen, $\mathcal{L}^2$-Version). *Unter der Voraussetzung (A) konvergiert*

$$\frac{S_n(\omega)}{n} - \frac{E[S_n]}{n} \quad \longrightarrow \quad 0$$

für P-fast alle $\omega \in \Omega$. Insbesondere gilt

$$\frac{S_n}{n} \quad \longrightarrow \quad m \qquad P\text{-fast sicher,}$$

falls $E[X_i] = m$ für alle i.

Der Übersichtlichkeit halber führen wir den Beweis zunächst unter den stärkeren Voraussetzungen (A1) und (A2). Der allgemeine Fall ist eine Übungsaufgabe, die sich gut zum Wiederholen der Beweisschritte eignet:

Beweis unter den Annahmen (A1) und (A2). Wir können o.B.d.A. $E[X_i] = 0$ für alle i voraussetzen – andernfalls betrachten wir die zentrierten Zufallsvariablen $\widetilde{X_i} := X_i - E[X_i]$; diese sind wieder unkorreliert mit beschränkten Varianzen. Zu zeigen ist dann:

$$\frac{S_n}{n} \to 0 \qquad P\text{-fast sicher.}$$

Wir unterteilen den Beweis in mehrere Schritte:

(1). *Schnelle stochastische Konvergenz gegen 0 entlang der Teilfolge $n_k = k^2$:* Aus der Čebyšev-Ungleichung folgt:

$$P\left[\left|\frac{S_{k^2}}{k^2}\right| \geq \varepsilon\right] \leq \frac{1}{\varepsilon^2} \operatorname{Var}\left[\frac{S_{k^2}}{k^2}\right] \leq \frac{1}{\varepsilon^2 k^2} \sup_i \operatorname{Var}[X_i].$$

Da die Varianzen beschränkt sind, ist der gesamte Ausdruck durch die Summanden einer summierbaren Reihe beschränkt. Somit ergibt sich nach Borel-Cantelli:

$$\frac{S_{k^2}(\omega)}{k^2} \quad \to \quad 0$$

für alle ω außerhalb einer Nullmenge N_1.

(2). Wir untersuchen nun die Fluktuationen der Folge S_n zwischen den Werten der Teilfolge $n_k = k^2$. Sei

$$D_k \quad := \quad \max_{k^2 \leq l < (k+1)^2} |S_l - S_{k^2}|.$$

Wir zeigen *schnelle stochastische Konvergenz gegen 0 für D_k/k^2*. Für $\varepsilon > 0$ haben wir

$$\begin{aligned} P\left[\frac{D_k}{k^2} \geq \varepsilon\right] &= P\left[\bigcup_{k^2 \leq l < (k+1)^2} \{\, |S_l - S_{k^2}| > \varepsilon k^2 \,\}\right] \\ &\leq \sum_{l=k^2}^{k^2+2k} P[|S_l - S_{k^2}| > \varepsilon k^2] \leq \frac{\text{const.}}{k^2}, \end{aligned}$$

denn nach der Čebyšev-Ungleichung gilt für $k^2 \le l \le k^2 + 2k$:

$$\begin{aligned} P[|S_l - S_{k^2}| > \varepsilon k^2] &\le \frac{1}{\varepsilon^2 k^4} \operatorname{Var}[S_l - S_{k^2}] \le \frac{1}{\varepsilon^2 k^4} \operatorname{Var}\left[\sum_{i=k^2+1}^{l} X_i\right] \\ &\le \frac{l-k^2}{\varepsilon^2 k^4} \sup_i \operatorname{Var}[X_i] \le \text{const} \cdot \frac{k}{k^4}. \end{aligned}$$

Nach Lemma 7.2 folgt daher

$$\frac{D_k(\omega)}{k^2} \to 0$$

für alle ω außerhalb einer Nullmenge N_2.

(3). Zu gegebenem n wählen wir nun $k = k(n)$ mit $k^2 \le n < (k+1)^2$. Durch Kombination der ersten beiden Schritte erhalten wir:

$$\left|\frac{S_n(\omega)}{n}\right| \le \frac{|S_{k^2}(\omega)| + D_k(\omega)}{n} \le \left|\frac{S_{k^2}(\omega)}{k^2}\right| + \frac{D_k(\omega)}{k^2} \longrightarrow 0 \qquad \text{für } n \to \infty$$

für alle ω außerhalb der Nullmenge $N_1 \cup N_2$. Also konvergiert S_n/n P-fast sicher gegen 0.

□

Beispiel (Random Walk im $\mathbb{R}^d$). Sei $S_n = X_1 + ... + X_n$ ein Random Walk im $\mathbb{R}^d$ mit unabhängigen identisch verteilten Inkrementen X_i mit Verteilung μ. Gilt

$$E[\|X_i\|^2] = \int_{\mathbb{R}^d} \|x\|^2 \, \mu(dx) < \infty,$$

dann folgt nach dem schwachen Gesetz der großen Zahlen (angewandt auf die Komponenten $S_n^{(k)} = \sum_{i=1}^{n} X_i^{(k)}$ des Vektors S_n):

$$\frac{S_n(\omega)}{n} \longrightarrow m \qquad \text{für } P\text{-fast alle } \omega,$$

wobei $m = \int_{\mathbb{R}^d} x\mu(dx)$ der Schwerpunkt der Inkrementverteilung ist. Insbesondere gilt für $m \neq 0$:

$$S_n \sim m \cdot n \qquad \text{für } n \to \infty \quad P\text{-fast sicher,}$$

d.h. S_n wächst linear mit Geschwindigkeit m. Im Fall $m = 0$ gilt dagegen

$$\frac{S_n(\omega)}{n} \to 0 \quad P\text{-fast sicher,}$$

d.h. der Random Walk wächst sublinear. Eine viel präzisere Beschreibung der pfadweisen Asymptotik des Random Walk im Fall $m = 0$ liefert der *Satz vom iterierten Logarithmus*:

$$\begin{aligned} \limsup_{n\to\infty} \frac{S_n(\omega)}{\sqrt{n \log\log n}} &= +1 \qquad P\text{-fast sicher,} \\ \liminf_{n\to\infty} \frac{S_n(\omega)}{\sqrt{n \log\log n}} &= -1 \qquad P\text{-fast sicher,} \end{aligned}$$

siehe z.B. [BAUER: „WAHRSCHEINLICHKEITSTHEORIE“].

Beispiel (Wachstum in zufälligen Medien). Um ein zufälliges Populationswachstum zu beschreiben, definieren wir Zufallsvariablen X_n $(n \in \mathbb{N})$ durch

$$X_0 = 1, \quad X_n = Y_n \cdot X_{n-1},$$

d.h. $X_n = \prod_{i=1}^n Y_i$. Hierbei nehmen wir an, dass die Wachstumsraten Y_i unabhängige identisch verteilte Zufallsvariablen mit $Y_i > 0$ P-f.s. sind. Sei $m = E[Y_i]$.

(1). ASYMPTOTIK DER ERWARTUNGSWERTE: Da die Y_i unabhängig sind, gilt:

$$E[X_n] \;=\; \prod_{i=1}^n E[Y_i] \;=\; m^n.$$

Die mittlere Populationsgröße wächst also im *superkritischen Fall* $m > 1$ exponentiell und fällt im *subkritischen Fall* $m < 1$ exponentiell ab.

Konkretes Beispiel: In einem Glücksspiel setzt der Spieler in jeder Runde die Hälfte seines Kapitals. Mit Wahrscheinlichkeit $\frac{1}{2}$ erhält er das c-fache des Einsatzes zurück, und mit Wahrscheinlichkeit $\frac{1}{2}$ erhält er nichts zurück. Hier gilt:

$$Y_i \;=\; \begin{cases} \frac{1}{2}(1+c) & \text{mit } p = \frac{1}{2} \\ \frac{1}{2} & \text{mit } p = \frac{1}{2} \end{cases},$$

also

$$m \;=\; E[Y_i] = \frac{1}{4}(1+c) + \frac{1}{4} = \frac{2+c}{4}.$$

Das Spiel ist also „fair“ für $c = 2$ und „superfair“ für $c > 2$.

(2). ASYMPTOTIK VON $X_n(\omega)$: Wir nehmen nun an, dass $\log Y_1 \in \mathcal{L}^2$ gilt. Nach dem starken Gesetz der großen Zahlen folgt dann:

$$\frac{1}{n} \log X_n = \frac{1}{n} \sum_{i=1}^n \log Y_i \;\to\; E[\log Y_1] =: \alpha \qquad P\text{-f.s.}$$

Also existiert für $\varepsilon > 0$ ein $N(\omega)$ mit $N(\omega) < \infty$ P-fast sicher,

$$X_n(\omega) \;\leq\; e^{(\alpha+\varepsilon)n} \quad \text{und} \quad X_n(\omega) \;\geq\; e^{(\alpha-\varepsilon)n} \qquad \text{für alle } n \geq N(\omega).$$

Für $\alpha < 0$ fällt X_n also P-fast sicher exponentiell ab, während X_n für $\alpha > 0$ P-fast sicher exponentiell wächst.

(3). ZUSAMMENHANG VON α UND m: Nach der Jensenschen Ungleichung gilt:

$$\alpha = E[\log Y_1] \leq \log E[Y_1] = \log m.$$

Hierbei haben wir benutzt, dass der Logarithmus eine konkave, bzw. $-\log$ eine konvexe Funktion ist. Im subkritischen Fall $m < 1$ ist also auch α strikt negativ, d.h. X_n fällt auch P-f.s. exponentiell ab. Im superkritischen Fall $m > 1$ kann es aber passieren, dass *trotzdem* $\alpha < 0$ gilt, d.h. obwohl die Erwartungswerte exponentiell wachsen, fällt X_n P-fast sicher exponentiell! Im Beispiel

$$Y_i = \begin{cases} \frac{1}{2}(1+c) & \text{mit } p = \frac{1}{2} \\ \frac{1}{2} & \text{mit } p = \frac{1}{2} \end{cases}$$

von oben wachsen die Erwartungswerte exponentiell für $c > 2$, aber es gilt

$$\alpha = E[\log Y_i] = \frac{1}{2}\left(\log \frac{1+c}{2} + \log \frac{1}{2}\right) = \frac{1}{2} \log \frac{1+c}{4} \geq 0 \;\Leftrightarrow\; c \geq 3.$$

Für $c \in (2,3)$ ist das Spiel also superfair mit fast sicherem exponentiellem Bankrott!

Die Voraussetzungen des Satzes von Lebesgue sind in dieser Situation nicht erfüllt, denn es gilt:

$$E[X_n] \nearrow \infty, \quad \text{obwohl } X_n \to 0 \quad P\text{-fast sicher.}$$

Von $\mathcal{L}^2$ nach $\mathcal{L}^1$ mit Unabhängigkeit

Sind Zufallsvariablen $X, Y : \Omega \to S$ unabhängig, so sind $f(X)$ und $g(Y)$ für beliebige beschränkte oder nichtnegative Funktionen $f, g : S \to \mathbb{R}$ unkorreliert. Bisher konnten wir zeigen, dass das starke Gesetz der großen Zahlen für unkorrelierte (bzw. schwach korrelierte) Zufallsvariablen $X_n \in \mathcal{L}^2$ mit gleichmäßig beschränkten Varianzen gilt. Die Unabhängigkeit der X_n ermöglicht es, diese Aussage auf integrierbare Zufallsvariablen (d.h. $\mathcal{L}^1$ statt $\mathcal{L}^2$) zu erweitern:

Satz 7.10 (Kolmogorovs Gesetz der großen Zahlen). *Seien $X_1, X_2, ... \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ paarweise unabhängig und identisch verteilt mit $E[X_i] = m$. Dann gilt:*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i \;=\; m \qquad \textit{P-fast sicher.}$$

Kolmogorov hatte eine entsprechende Aussage unter der Annahme von Unabhängigkeit (statt paarweiser Unabhängigkeit) bewiesen. Der Beweis unter der schwächeren Voraussetzung stammt von Etemadi (1981).

Bemerkung (Dynamische Systeme, Ergodensatz). In einer dynamischen Interpretation bedeutet die Aussage

$$\frac{1}{n} \sum_{i=1}^{n} X_i(\omega) \longrightarrow m = \int x \, \mu_{X_i}(dx) \qquad P\text{-fast sicher,}$$

des starken Gesetzes der großen Zahlen, dass die „zeitlichen Mittelwerte“ der Zufallsvariablen X_i gegen den „räumlichen Mittelwert“ m konvergieren. Dies ist ein Spezialfall eines viel allgemeineren *Ergodensatzes*, der eine entsprechende Aussage für ergodische dynamische Systeme liefert, siehe z.B. BREIMAN: PROBABILITY oder DURRETT: PROBABILITY: THEORY AND EXAMPLES.

von Satz 7.10. Wir führen den Beweis in mehreren Schritten.

(1). *Reduktion auf nichtnegative Zufallsvariablen.*

Wir können o.B.d.A. $X_i \geq 0$ für alle $i \in \mathbb{N}$ voraussetzen. Andernfalls zerlegen wir $X_i = X_i^+ - X_i^-$. Die Zufallsvariablen $X_i^+, i \in \mathbb{N}$, bzw. $X_i^-, i \in \mathbb{N}$, sind jeweils Funktionen der X_i, und daher wieder paarweise unabhängig. Aus dem Gesetz der großen Zahlen für X_i^+ und X_i^- folgt das Gesetz der großen Zahlen für die Zufallsvariablen X_i.

(2). *Reduktion auf Gesetz der großen Zahlen für $Y_i := X_i \cdot I_{\{X_i \leq i\}}$.*

Nach dem Lemma von Borel-Cantelli gilt

$$P[Y_i \neq X_i \text{ unendlich oft}] = 0,$$

denn

$$\begin{aligned}
\sum_{i=1}^{\infty} P[Y_i \neq X_i] &= \sum_{i=1}^{\infty} P[X_i > i] \\
&= \sum_{i=1}^{\infty} P[X_1 > i] \qquad (X_i \text{ identisch verteilt}) \\
&\leq \int_0^{\infty} P[X_1 > x] \, dx \qquad (P[X_1 > x] \text{ monoton fallend}) \\
&= E[X_1] \quad < \quad \infty.
\end{aligned}$$

Also konvergiert $\frac{1}{n} \sum_{i=1}^{n} X_i$ P-fast sicher gegen m, falls dasselbe für $\frac{1}{n} \sum_{i=1}^{n} Y_i$ gilt.

Sei nun

$$S_n \;=\; \sum_{i=1}^{n} Y_i.$$

Die Zufallsvariablen Y_i sind wieder paarweise unabhängig, und es gilt $0 \leq Y_i \leq i$.

(3). *Konvergenz der Erwartungswerte.*

Da die Zufallsvariablen Y_i nicht mehr identisch verteilt sind, bestimmen wir zunächst den Grenzwert der Erwartungswerte der Mittelwerte S_n/n. Nach dem Satz von der monotonen Konvergenz gilt

$$E[Y_i] = E[X_i \,;\, X_i \leq i] = E[X_1 \cdot I_{\{X_1 \leq i\}}] \longrightarrow E[X_1] = m, \qquad \text{für } i \to \infty,$$

also auch

$$E\left[\frac{S_n}{n}\right] = \frac{1}{n}\sum_{i=1}^{n} E[Y_i] \longrightarrow m \qquad \text{für } n \to \infty.$$

(4). *P-fast sichere Konvergenz von $\frac{S_n}{n}$ entlang der Teilfolgen $k_n = \lfloor \alpha^n \rfloor, \alpha > 1$.*

Vorbemerkung: Es gilt

$$\sum_{n \geq m} \frac{1}{k_n^2} \;=\; \frac{1}{\lfloor \alpha^m \rfloor^2} + \frac{1}{\lfloor \alpha^{m+1} \rfloor^2} + \ldots \;\leq\; \frac{\text{const.}}{\lfloor \alpha^m \rfloor^2} \;=\; \frac{\text{const.}}{k_m^2}$$

mit einer von m unabhängigen Konstanten.

Behauptung:

$$\frac{S_{k_n}}{k_n} \;\longrightarrow\; \lim_{n\to\infty} E\left[\frac{S_{k_n}}{k_n}\right] = m \qquad P\text{-fast sicher.}$$

Beweis der Behauptung: Nach dem Lemma von Borel-Cantelli genügt es,

$$\sum_{n=1}^{\infty} P\left[\left|\frac{S_{k_n} - E[S_{k_n}]}{k_n}\right| \geq \varepsilon\right] \;<\; \infty$$

zu zeigen. Dies ist der Fall, wenn

$$\sum_{n=1}^{\infty} \mathrm{Var}\left[\frac{S_{k_n}}{k_n}\right] \;< \infty$$

gilt. Wegen

$$\mathrm{Var}[Y_i] \leq E[Y_i^2] = E[X_i^2 \,;\, X_i \leq i] = E[X_1^2 \,;\, X_1 \leq i]$$

erhalten wir mithilfe der Vorbemerkung

$$\begin{aligned}
\sum_{n=1}^{\infty} \mathrm{Var}\left[\frac{S_{k_n}}{k_n}\right] &= \sum_{n=1}^{\infty} \frac{1}{k_n^2} \cdot \sum_{i=1}^{k_n} \mathrm{Var}[Y_i] \\
&\leq \sum_{i=1}^{\infty} E[X_1^2\,;\, X_1 \leq i] \cdot \sum_{n:k_n \geq i} \frac{1}{k_n^2} \\
&\leq \text{const.} \cdot \sum_{i=1}^{\infty} E\left[X_1^2\,;\, X_1 \leq i\right] \cdot \frac{1}{i^2} \\
&\leq \text{const.} \cdot \sum_{i=1}^{\infty} \sum_{j=1}^{i} j^2 \cdot P[X_1 \in (j-1,j]] \cdot \frac{1}{i^2} \\
&= \text{const.} \cdot \sum_{j=1}^{\infty} j^2 \cdot P[X_1 \in (j-1,j]] \cdot \sum_{i=j}^{\infty} \frac{1}{i^2} \\
&\leq \text{const.} \cdot \sum_{j=1}^{\infty} j \cdot P[X_1 \in (j-1,j]] \\
&= \text{const.} \cdot E\left[\sum_{j=1}^{\infty} j \cdot I_{\{X_1 \in (j-1,j]\}}\right] \\
&\leq \text{const.} \cdot E[X_1 + 1] \;<\; \infty.
\end{aligned}$$

(5). *P-fast sichere Konvergenz von $\frac{S_n}{n}$.*

Für $l \in \mathbb{N}$ mit $k_n \leq l \leq k_{n+1}$ gilt wegen $Y_i \geq 0$:

$$S_{k_n} \leq S_l \leq S_{k_{n+1}}.$$

Es folgt

$$\frac{k_n}{k_{n+1}} \cdot \frac{S_{k_n}}{k_n} = \frac{S_{k_n}}{k_{n+1}} \leq \frac{S_l}{l} \leq \frac{S_{k_{n+1}}}{k_n} = \frac{k_{n+1}}{k_n} \cdot \frac{S_{k_{n+1}}}{k_{n+1}}.$$

Für $n \to \infty$ erhalten wir wegen $\frac{k_{n+1}}{k_n} \to \alpha$ und $\frac{S_{k_n}(\omega)}{k_n} \to m$:

$$\frac{m}{\alpha} \quad \leq \quad \liminf \frac{S_l(\omega)}{l} \quad \leq \quad \limsup \frac{S_l(\omega)}{l} \quad \leq \quad \alpha m$$

für alle ω außerhalb einer von α abhängenden Nullmenge N_α. Für ω außerhalb der Nullmenge $\bigcup_{\substack{\alpha > 1 \\ \alpha \in \mathbb{Q}}} N_\alpha$ folgt somit:

$$\lim_{l \to \infty} \frac{S_l(\omega)}{l} \;=\; m.$$

□

Korollar 7.11 (Gesetz der großen Zahlen ohne Integrierbarkeit). *Seien $X_1, X_2 ...$ paarweise unabhängige, identisch verteilte, nicht-negative Zufallsvariablen. Dann gilt:*

$$\lim_{n\to\infty} \frac{1}{n} \cdot \sum_{i=1}^{n} X_i(\omega) \;=\; E[X_1] \;\in [0,\infty] \qquad P\text{-fast sicher.}$$

Beweis. Nach Satz 7.10 gilt die Aussage im Fall $E[X_1] < \infty$. Für $E[X_1] = \infty$ erhalten wir für $k \in \mathbb{N}$:

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i \quad \geq \quad \liminf_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (X_i \wedge k) \quad = \quad E[X_1 \wedge k] \qquad P\text{-fast sicher.}$$

Für $k \to \infty$ folgt dann mit monotoner Konvergenz

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i \quad \geq \quad E[X_1] \quad = \quad \infty,$$

und damit die Behauptung. □

7.3 Empirische Verteilungen

Schätzen von Kenngrößen einer unbekannten Verteilung

Angenommen, wir haben eine Stichprobe aus reellen Beobachtungswerten $X_1, X_2, \ldots, X_n$ gegeben, und möchten die zugrundeliegende Wahrscheinlichkeitsverteilung μ auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ möglichst weitgehend rekonstruieren. Im einfachsten Modell interpretieren wir die Beobachtungswerte als Realisierungen unabhängiger Zufallsvariablen $X_1, X_2, \ldots$ mit Verteilung μ.

(1). SCHÄTZEN DES ERWARTUNGSWERTES: Sei $\int |x|\, \mu(dx) < \infty$. Um den Erwartungswert

$$m \quad = \quad \int x\, \mu(dx)$$

zu schätzen, verwenden wir das **empirische Mittel**

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

Das empirische Mittel ist ein *erwartungstreuer Schätzer* für m, d.h. $\overline{X}_n$ ist eine Funktion von den Beobachtungswerten $X_1, \ldots, X_n$ mit $E[\overline{X}_n] = m$. Obere Schranken für den Schätzfehler $P[|\overline{X}_n - m| > \varepsilon], \varepsilon > 0$, erhält man z.B. mithilfe der Čebyšev- oder der exponentiellen Markov-Ungleichung. Für $n \to \infty$ gilt nach dem Gesetz der großen Zahlen

$$\overline{X}_n \quad \longrightarrow \quad m \qquad P\text{-fast sicher,}$$

d.h. $\overline{X}_n$ ist eine *konsistente* Folge von Schätzern für m.

(2). SCHÄTZEN DER VARIANZ: Um die Varianz

$$v \quad = \quad \int (x-m)^2 \, \mu(dx)$$

der zugrundeliegenden Verteilung 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 unter anderem, dass $\widetilde{V}_n$ ein *erwartungstreuer* Schätzer für v ist, denn aus

$$\begin{array}{rcl} \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 & = & \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2 - (\overline{X}_n - m)^2 \\ \text{Stichprobenvarianz} & = & \text{MSE} \quad - \quad \text{Stichprobenbias}^2 \end{array} \tag{7.3.1}$$

folgt

$$E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2\right] \quad = \quad \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}[X_i] - \mathrm{Var}[\overline{X}_n] \quad = \quad \frac{n-1}{n}v,$$

also $E[\widetilde{V}_n] = v$.

Um zu zeigen, dass $\widetilde{V}_n$ 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 (7.3.1). Nach dem starken Gesetz der großen Zahlen für nichtnegative Zufallsvariablen erhalten wir

$$\frac{n-1}{n}\widetilde{V}_n \quad = \quad \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2 - (\overline{X}_n - m)^2 \quad \longrightarrow \quad v \qquad P\text{-fast sicher,}$$

also auch $\widetilde{V}_n \to v$ P-fast sicher.

(3). SCHÄTZEN VON INTEGRALEN: Allgemeiner können wir für jede Funktion $f \in \mathcal{L}^1(S, \mathcal{S}, \mu)$ das Integral

$$\theta \quad = \quad \int f \, d\mu$$

erwartungstreu durch die **empirischen Mittelwerte**

$$\widehat{\theta}_n \quad = \quad \frac{1}{n}\sum_{i=1}^{n} f(X_i)$$

schätzen. Dies haben wir schon in Kapitel 3 für Monte Carlo Verfahren verwendet. Da die Zufallsvariablen $f(X_i)$ wieder unabhängig und identisch verteilt sind mit Erwartungswert θ, gilt nach dem starken Gesetz der großen Zahlen:

$$\widehat{\theta}_n \quad \longrightarrow \quad \theta \qquad P\text{-fast sicher.} \tag{7.3.2}$$

(4). SCHÄTZEN DER VERTEILUNG: Die gesamte Verteilung μ können wir durch die **empirische Verteilung**

$$\widehat{\mu}_n(\omega) \quad = \quad \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i(\omega)}$$

der Zufallsstichprobe schätzen. $\widehat{\mu}_n$ ist eine *„zufällige Wahrscheinlichkeitsverteilung,“* d.h. eine Zufallsvariable mit Werten im Raum $WV(\mathbb{R})$ der Wahrscheinlichkeitsverteilungen auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Aus (7.3.2) ergibt sich die folgende Approximationseigenschaft der empirischen Verteilungen:

$$\int f \, d\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} f(X_i) \quad \stackrel{n\to\infty}{\longrightarrow} \quad \int f \, d\mu \tag{7.3.3}$$

P-fast sicher für alle $f \in \mathcal{L}^1(S, \mathcal{S}, \mu)$.

Konvergenz der empirischen Verteilungsfunktionen

Für die *empirischen Verteilungsfunktionen*

$$F_n(c) \quad = \quad \widehat{\mu}_n[(-\infty, c]] \quad = \quad \frac{1}{n}\,|\{1 \le i \le n \,:\, X_i \le c\}|$$

von unabhängigen, identisch verteilten, reellwertigen Zufallsvariablen $X_1, X_2, \ldots$ mit Verteilungsfunktion F ergibt sich wegen $F_n(c) = \int I_{(-\infty,c]}\, d\widehat{\mu}_n$:

$$\lim_{n\to\infty} F_n(c) \quad = \quad F(c) \qquad P\text{-fast sicher für alle } c \in \mathbb{R}. \tag{7.3.4}$$

Diese Aussage kann man noch etwas verschärfen:

Satz 7.12 (Glivenko-Cantelli). *Sind $X_1, X_2, \ldots$ unabhängig und identisch verteilt mit Verteilungsfunktion F, dann gilt für die empirischen Verteilungsfunktionen F_n:*

$$\sup_{c\in\mathbb{R}} |F_n(c) - F(c)| \quad \longrightarrow \quad 0 \qquad \textit{P-fast sicher.} \tag{7.3.5}$$

Beweis. Wir führen den Beweis unter der zusätzlichen Annahme, dass F stetig ist – für den allgemeinen Fall siehe z.B. *Klenke: Wahrscheinlichkeitstheorie*. Sie $\varepsilon > 0$ gegeben. Ist F stetig, dann existieren $k \in \mathbb{N}$ und Konstanten

$$-\infty = c_0 < c_1 < c_2 < \ldots < c_k = \infty \qquad \text{mit } F(c_i) - F(c_{i-1}) \leq \frac{\varepsilon}{2}$$

für alle $1 \leq i \leq k$. Da F_n nach 7.3.4 mit Wahrscheinlichkeit 1 punktweise gegen F konvergiert, existiert zudem ein $n_0 \in \mathbb{N}$ mit

$$\max_{0 \leq i \leq n} |F_n(c_i) - F(c_i)| < \frac{\varepsilon}{2} \qquad \text{für alle } n \geq n_0.$$

Wegen der Monotonie der Verteilungsfunktionen folgt dann

$$F_n(c) - F(c) \leq F_n(c_i) - F(c_{i-1}) \leq \frac{\varepsilon}{2} + F_n(c_i) - F(c_i) < \varepsilon,$$

und entsprechend

$$F(c) - F_n(c) \leq F(c_i) - F_n(c_{i-1}) \leq \frac{\varepsilon}{2} + F(c_i) - F_n(c_i) < \varepsilon,$$

für alle $n \geq n_0, c \in \mathbb{R}$, und $1 \leq i \leq k$ mit $c_{i-1} \leq c \leq c_i$. Also gilt auch

$$\sup_{c \in \mathbb{R}} |F_n(c) - F(c)| < \varepsilon \qquad \text{für alle } n \geq n_0.$$

□

Bemerkung (QQ-Plot). In parametrischen statistischen Modellen nimmt man von vornherein an, dass die beobachteten Daten Realisierungen von Zufallsvariablen sind, deren Verteilung aus einer bestimmten Familie von Wahrscheinlichkeitsverteilungen stammt, z.B. der Familie aller Normalverteilungen. Um zu entscheiden, ob eine solche Annahme für gegebene reellwertige Daten $x_1, \ldots, x_n$ gerechtfertigt ist, kann man die empirische Verteilungsfunktion mit der tatsächlichen Verteilungsfunktion vergleichen. Ein praktikables graphisches Verfahren ist der Quantil-Quantil-Plot, bei dem die Quantile der empirischen und der theoretischen Verteilung gegeneinander aufgetragen werden. Um auf Normalverteilung zu testen, plottet man beispielsweise die Punkte

$$\left(\Phi^{-1}\left(\frac{k - \frac{1}{2}}{n} \right), x_{(k)} \right), \qquad k = 1, 2, \ldots, n,$$

wobei Φ die Verteilungsfunktion der Standardnormalverteilung ist, und

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$$

die Ordnungsstatistiken von $x_1, \ldots, x_n$, also die $(k - \frac{1}{2})/n$-Quantile der empirischen Verteilung sind. Ist die zugrundeliegende Verteilung eine Normalverteilung mit Mittel m und Standardabweichung σ, dann liegen die Punkte für große n näherungsweise auf einer Geraden mit Steigung σ und Achsenabschnitt m, da für die Verteilungsfunktion und die Quantile der theoretischen Verteilung dann

$$F(c) \quad = \quad P[X \leq c] \quad = \quad P[\sigma Z + m \leq c] \quad = \quad P\left[Z \leq \frac{c-m}{\sigma}\right] \quad = \quad \Phi\left(\frac{c-m}{\sigma}\right),$$

bzw.

$$F^{-1}(u) \quad = \quad m + \sigma\Phi^{-1}(u)$$

gilt. Die folgende Grafik zeigt QQ-Plots bzgl. der Normalverteilung für verschiedene Datensätze.

Histogramme und Multinomialverteilung

Die empirische Verteilung $\widehat{\mu}_n(\omega) = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i(\omega)}$ von Zufallsvariablen $X_1, \ldots, X_n$ ist selbst eine Zufallsvariable mit Werten im Raum der Wahrscheinlichkeitsverteilungen. Wir wollen nun die Verteilung dieser Zufallsvariablen explizit berechnen, falls die X_i unabhängig und identisch verteilt mit endlichem Wertebereich S sind. Haben die Zufallsvariablen keinen endlichen Wertebereich, dann kann man die Aussagen trotzdem anwenden, indem man den Wertebereich in endlich viele Teilmengen (Klassen) zerlegt.

Das *Histogramm* von n Beobachtungswerten $x_1, \ldots, x_n$, die in einer endlichen Menge S liegen, ist der Vektor

$$\vec{h} = (h_a)_{a \in S}, \qquad h_a = |\{1 \leq i \leq n | x_i = a\}|,$$

der Häufigkeiten der möglichen Werte $a \in S$ unter $x_1, \ldots, x_n$. Graphisch stellt man ein Histogramm durch ein Balkendiagramm dar:

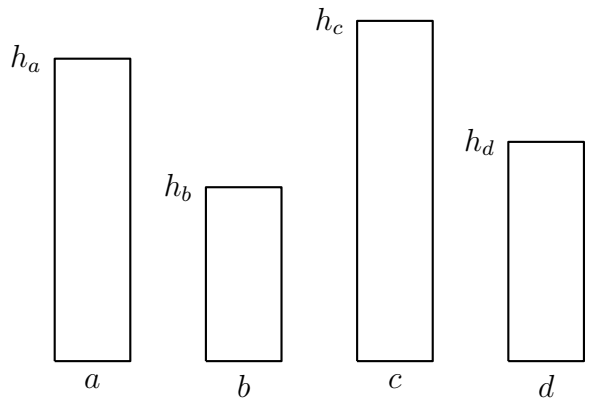


Abbildung 7.3: Histogramm der Klassen a, b, c und d mit den jeweiligen Häufigkeiten h_a, h_b, h_c und h_d

Der Raum Hist(n, S) aller möglichen Histogramme von n Beobachtungswerten ist eine Teilmenge von $\{0, 1, \ldots, n\}^S$:

$$\text{Hist}(n, S) \quad = \quad \{\vec{h} = (h_a)_{a \in S} | h_a \in \mathbb{Z}_+, \sum_{a \in S} h_a = n\} \quad \subseteq \quad \{0, 1, \ldots, n\}^S.$$

Sie nun μ eine Wahrscheinlichkeitsverteilung auf der endlichen Menge S. Wir wollen die Verteilung des Histogrammvektors bestimmen, wenn die Beobachtungswerte unabhängige Stichproben von der Verteilung μ sind. Wir betrachten also unabhängige Zufallsvariablen $X_1, \ldots, X_n$ auf einem Wahrscheinlichkeitsraum $(\Omega, \mathscr{A}, P)$ mit Verteilung μ und die Häufigkeiten

$$H_a(\omega) \quad := \quad |\{1 \leq i \leq n : X_i(\omega) = a\}|$$

der möglichen Werte $a \in S$. Die Zufallsvariable H_a ist $\text{Bin}(n, p)$-verteilt mit $p = \mu[\{a\}]$. Wir berechnen nun die **gemeinsame Verteilung** aller dieser Häufigkeiten, d.h. die Verteilung μ_H des Zufallsvektors

$$H \quad = \quad (H_a)_{a \in S} \quad : \quad \Omega \longrightarrow \text{Hist}(n, S)$$

mit Werten im Raum der Histogramme. Dazu verwenden wir die Unabhängigkeit der X_i. Mit $I = \{1, \ldots, n\}$ erhalten wir:

$$\begin{aligned}
\mu_H(\vec{k}) &= P[H_a = k_a \quad \forall a \in S] \\
&= P[X_i = a \text{ genau } k_a\text{-mal für alle } a \in S] \\
&= \sum_{\substack{I = \dot{\bigcup}_{a \in S} I_a \\ |I_a| = k_a}} P[X_i = a \quad \forall\, i \in I_a \quad \forall\, a \in S] \\
&= \sum_{\substack{I = \dot{\bigcup}_{a \in S} I_a \\ |I_a| = k_a}} \prod_{a \in S} \mu[\{a\}]^{k_a} \\
&= \binom{n}{\vec{k}} \prod_{a \in S} \mu[\{a\}]^{k_a}.
\end{aligned}$$

Hierbei laufen die Summen über alle disjunkten Zerlegungen von $I = \{0, 1, \ldots, n\}$ in Teilmengen $i_a, a \in S$, mit jeweils k_a Elementen, und der **Multinomialkoeffizient**

$$\binom{n}{\vec{k}} := \frac{n!}{\prod_{a \in S} k_a!}, \qquad k_a \in \{0, 1, \ldots, n\} \text{ mit } \sum_{a \in S} k_a = n,$$

gibt die Anzahl der Partitionen von n Elementen in Teilmengen von jeweils k_a Elementen an.

Definition. *Die Verteilung des Histogrammvektors H heißt **Multinomialverteilung für n Stichproben mit Ergebniswahrscheinlichkeiten** $\mu(a), a \in S$.*

Bemerkung. Im Fall $|S| = 2$ ist $H(\omega)$ eindeutig festgelegt durch $H_1(\omega)$, und die Zufallsvariable H_1 ist binomialverteilt mit Parametern n und $p = \mu[\{1\}]$. In diesem Sinn ergibt sich die Binomialverteilung als Spezialfall der Multinomialverteilung.

7.4 Entropie

Wir definieren nun die Entropie einer diskreten Wahrscheinlichkeitsverteilung. Mithilfe des Gesetzes der großen Zahlen können wir eine statistische Interpretation dieser Größe geben, aus der sich insbesondere der Quellenkodierungssatz von Shannon ergibt.

Definition und Eigenschaften

Wir bemerken zunächst, dass die auf $[0, \infty)$ definierte Funktion

$$u(x) := \begin{cases} x \log x & \text{für } x > 0 \\ 0 & \text{für } x = 0 \end{cases}$$

stetig und strikt konvex ist mit

$$u(x) \leq 0 \qquad \text{für alle } x \in [0, 1], \tag{7.4.1}$$

$$u(x) \geq x - 1 \qquad \text{für alle } x \geq 0, \tag{7.4.2}$$

und absolutem Minimum $u(1/e) = -1/e$.

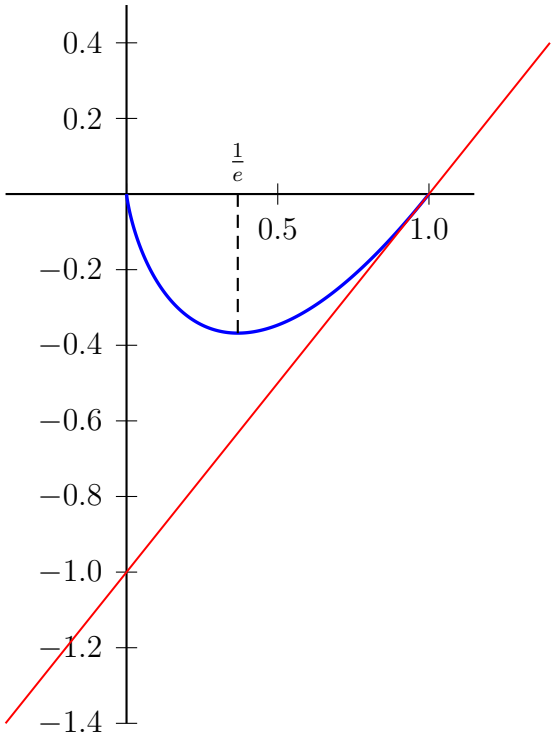


Abbildung 7.4: Graph der Funktion $u(x)$ (blau) und ihrer unteren Schranke $x - 1$ (rot)

Sei nun S eine abzählbare Menge, und $\mu = (\mu(x))_{x \in S}$ eine Wahrscheinlichkeitsverteilung auf S.

Definition. *Die Größe*

$$H(\mu) := -\sum_{\substack{x \in S \\ \mu(x) \neq 0}} \mu(x) \log \mu(x) = -\sum_{x \in S} u(\mu(x)) \in [0, \infty]$$

heißt ***Entropie*** *der Wahrscheinlichkeitsverteilung μ.*

Anschaulich können wir $-\log \mu(x)$ interpretieren als Maß für die »Überraschung« bzw. den »Informationsgewinn«, falls eine Stichprobe von der Verteilung μ den Wert x hat. Die »Überraschung« ist umso größer, je unwahrscheinlicher x ist. Die Entropie $H(\mu)$ ist dann die »mittlere Überraschung« bzw. der »mittlere Informationsgewinn« beim Ziehen einer Stichprobe von μ. Eine wichtige Eigenschaft der Entropie, die auch die Wahl des Logarithmus erklärt, ist:

Satz 7.13 (Faktorisierungseigenschaft). *Für beliebige diskrete Wahrscheinlichkeitsverteilungen μ und ν gilt:*

$$H(\mu \otimes \nu) \;=\; H(\mu) + H(\nu).$$

Der mittlere Informationszuwachs in einem aus zwei unabhängigen Experimenten zusammengesetzten Zufallsexperiment ist also die Summe der einzelnen mittleren Informationszuwächse.

Beweis. Nach Definition der Entropie gilt:

$$\begin{aligned} H(\mu \otimes \nu) &= \sum_{\substack{x,y \\ \mu(x)\nu(y)\neq 0}} \mu(x)\nu(y) \log(\mu(x)\nu(y)) \\ &= -\sum_{x:\mu(x)\neq 0} \mu(x) \log(\mu(x)) - \sum_{y:\nu(y)\neq 0} \nu(y) \log(\nu(y)) \\ &= H(\mu) + H(\nu). \end{aligned}$$

□

Wir bestimmen nun auf einer gegebenen abzählbaren Menge S die Wahrscheinlichkeitsverteilungen mit minimaler bzw. maximaler Entropie.

Extrema der Entropie:

(1). **Entropieminima:** Nach (7.4.1) ist die Entropie stets nicht-negativ, und es gilt:

$$H(\mu) = 0 \quad \Longleftrightarrow \quad \mu(x) \in \{0,1\} \quad \forall x \in S \quad \Longleftrightarrow \quad \mu \text{ ist ein Diracmaß.}$$

Die Diracmaße sind also die Entropieminima. Ist das Zufallsexperiment deterministisch, d.h. μ ein Diracmaß, dann tritt bei Ziehen einer Stichprobe von μ keine Überraschung bzw. kein Informationszuwachs auf.

(2). **Entropiemaximum:** Ist S endlich, dann gilt für alle Wahrscheinlichkeitsverteilungen μ auf S:

$$H(\mu) \;\leq\; -\log\left(\frac{1}{|S|}\right) \;=\; H(\mathcal{U}_S),$$

wobei $\mathcal{U}_S$ die Gleichverteilung auf S ist. Nach der Jensenschen Ungleichung gilt nämlich

$$\begin{aligned} -\sum_{x\in S} u(\mu(x)) &= -|S| \cdot \int u(\mu(x))\, \mathcal{U}_S(dx) \\ &\leq -|S| \cdot u\left(\int \mu(x)\, \mathcal{U}_S(dx)\right) \\ &= -|S| \cdot u\left(\frac{1}{|S|}\right) \quad = \quad -\log \frac{1}{|S|} \end{aligned}$$

mit Gleichheit genau dann, wenn μ die Gleichverteilung ist.

Die Gleichverteilung maximiert also die Entropie auf einem endlichen Zustandsraum. Anschaulich können wir die Gleichverteilung als eine »völlig zufällige« Verteilung auffassen – d.h. wir verwenden die Gleichverteilung als Modell, wenn wir keinen Grund haben, einen der Zustände zu bevorzugen. Die Entropie ist in diesem Sinne ein Maß für die *»Zufälligkeit«* (bzw. *»Unordnung«*) der Wahrscheinlichkeitsverteilung μ.

Auf einer abzählbar unendlichen Menge existiert keine Wahrscheinlichkeitsverteilung mit maximaler Entropie.

Beispiel (Entropie von Markovketten). Sei $p(x,y)$ $(x,y \in S)$ eine stochastische Matrix auf einer endlichen Menge S, die die Gleichverteilung $\mathcal{U}_S$ als Gleichgewicht hat, d.h. für alle $y \in S$ gilt:

$$\sum_{x\in S} p(x,y) \quad = \quad |S| \cdot \sum_{x\in S} \mathcal{U}_S(x)\, p(x,y) \quad = \quad |S| \cdot \mathcal{U}_S(y) \quad = \quad 1. \tag{7.4.3}$$

Beispielsweise ist p die Übergangsmatrix eines Random Walks auf dem diskreten Kreis $\mathbb{Z}_k = \mathbb{Z}/(k\mathbb{Z})$, der symmetrischen Gruppe S_n („Mischen eines Kartenspiels“), oder dem diskreten Hyperwürfel $\{0,1\}^n$ („Ehrenfestmodell“).

Der folgende Satz zeigt, dass die Entropie $H(\mu p^n)$ der Verteilung zur Zeit n einer Markovkette mit Startverteilung μ und Übergangsmatrix p monoton wächst:

Satz 7.14 (Zunahme der Entropie). *Ist p eine stochastische Matrix auf S mit (7.4.3), dann gilt:*

$$H(\mu p) \quad \geq \quad H(\mu)$$

für jede Wahrscheinlichkeitsverteilung μ auf S. Insbesondere ist $n \mapsto H(\mu p^n)$ monoton wachsend.

Beweis. Aus der Jensenschen Ungleichung folgt:

$$\begin{aligned} -H(\mu p) &= \sum_{y\in S} u\left(\sum_{x\in S}\mu(x)p(x,y)\right) \\ &\leq \sum_{y\in S}\sum_{x\in S} u(\mu(x))p(x,y) \\ &= \sum_{x\in S} u(\mu(x)) \quad = \quad -H(\mu). \end{aligned}$$

Hierbei haben wir im zweiten Schritt benutzt, dass die Funktion u konvex ist, und dass $x \mapsto p(x,y)$ nach (7.4.3) für jedes $y \in S$ die Gewichtsfunktion einer Wahrscheinlichkeitsverteilung ist. □

In der Interpretation der statistischen Physik geht die zeitliche Entwicklung auf makroskopischer Ebene (Thermodynamik) von einem geordneten hin zu einem ungeordneten Zustand maximaler Entropie (»thermodynamische Irreversibilität«). Trotzdem ist auf mikroskopischer Ebene die Dynamik rekurrent, d.h. jeder Zustand $x \in S$ wird von der Markovkette mit Wahrscheinlichkeit 1 unendlich oft besucht – dies dauert nur eventuell astronomisch lange. Die Einführung eines Markovmodells durch die österreichischen Physiker Tatjana und Paul Ehrenfest konnte eine entsprechende Kontroverse von Zermelo („Dynamik kehrt immer wieder zurück“) und Boltzmann („soll solange warten“) lösen.

Statistische Interpretation der Entropie

Sei μ eine Wahrscheinlichkeitsverteilung auf einer abzählbaren Menge S. Die Wahrscheinlichkeit einer Folge von Ausgängen $x_1, \ldots, x_n$ bei Entnehmen einer Stichprobe aus n unabhängigen Zufallsgrößen mit Verteilung μ beträgt

$$p_n(x_1, ..., x_n) = \prod_{i=1}^{n} \mu(x_i).$$

Der gemittelte Informationszuwachs durch Auswertung der Werte $x_1, \ldots, x_n$ ist also

$$-\frac{1}{n}\log p_n(x_1, ..., x_n).$$

Mithilfe des Gesetzes der großen Zahlen können wir die Asymptotik dieser Größen für $n \to \infty$ untersuchen:

Satz 7.15 (Shannon - Mc Millan). *Seien $X_1, X_2, \ldots : \Omega \to S$ unter P unabhängige Zufallsvariablen mit Verteilung μ. Dann gilt P-fast sicher*

$$-\frac{1}{n}\log p_n(X_1, \ldots, X_n) \quad \longrightarrow \quad H(\mu) \qquad \textit{für } n \to \infty.$$

Beweis. Mit Wahrscheinlichkeit 1 gilt $\mu(X_i) > 0$ für alle i, also nach Korollar 7.11:

$$-\frac{1}{n}\log p_n(X_1,\ldots,X_n) \quad = \quad -\frac{1}{n}\sum_{i=1}^{n}\log\mu(X_i) \quad \stackrel{n\to\infty}{\longrightarrow} \quad -\int \log\mu\, d\mu \quad = \quad H(\mu).$$

□

Bemerkung (Exponentielle Skala). Die Aussage des Satzes besagt, dass auf der „exponentiellen Skala“ fast sicher

$$p_n(X_1,\ldots,X_n) \quad \simeq \quad e^{-nH(\mu)}$$

gilt, d.h. beide Ausdrücke sind asymptotisch äquivalent bis auf subexponentielle (also z.B. polynomiell) wachsende Faktoren. Eine asymptotische Beschreibung von Wahrscheinlichkeiten auf der exponentiellen Skala ist Gegenstand der Theorie großer Abweichungen, siehe Abschnitt Satz 8.3 und Kapitel 11 unten.

Entropie und Kodierung

Wir betrachten nun eine Anwendung der Entropie auf die *möglichst effiziente Beschreibung/Kodierung einer Zufallsfolge.* Eine unbekannte Signalfolge mit Werten in einer endlichen Menge S (dem zugrundeliegenden „Alphabet“) beschreibt man im einfachsten A-Priori-Modell durch unabhängige Zufallsvariablen $X_1, X_2, ...$ mit Verteilung μ, wobei $\mu(x)$ die relative Häufigkeit des Buchstabens x in der verwendeten Sprache ist. Eine „perfekte“ Kodierung ordnet jedem Wort mit einer vorgegebenen Anzahl n von Buchstaben, also jedem Element des Produktraums S^n, eine Binärfolge zu. Will man alle Wörter mit n Buchstaben perfekt kodieren, werden $n \cdot \log|S|$ Bits benötigt. Wir betrachten stattdessen „effiziente“ Kodierungen, die nur den „meisten“ Wörtern mit n Buchstaben eindeutig eine Binärfolge zuordnen.

Definition. *Eine Folge von Mengen $B_n \subseteq S^n$ $(n \in \mathbb{N})$ heißt **wesentlich** bzgl. μ, falls*

$$P\left[(X_1,...,X_n) \in B_n\right] \quad = \quad \mu^n[B_n] \quad \to \quad 1 \qquad \textit{für } n \to \infty.$$

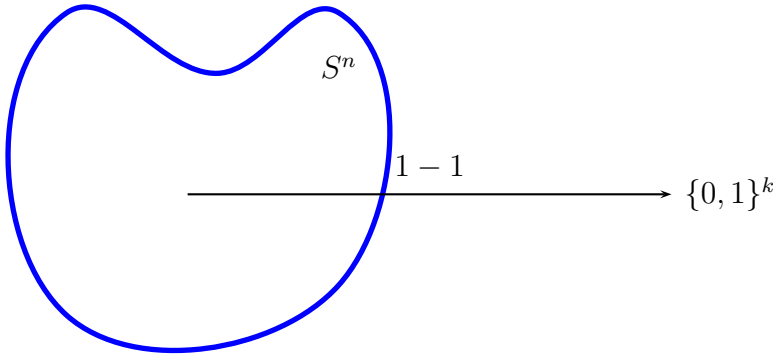


Abbildung 7.5: Perfekte Kodierung

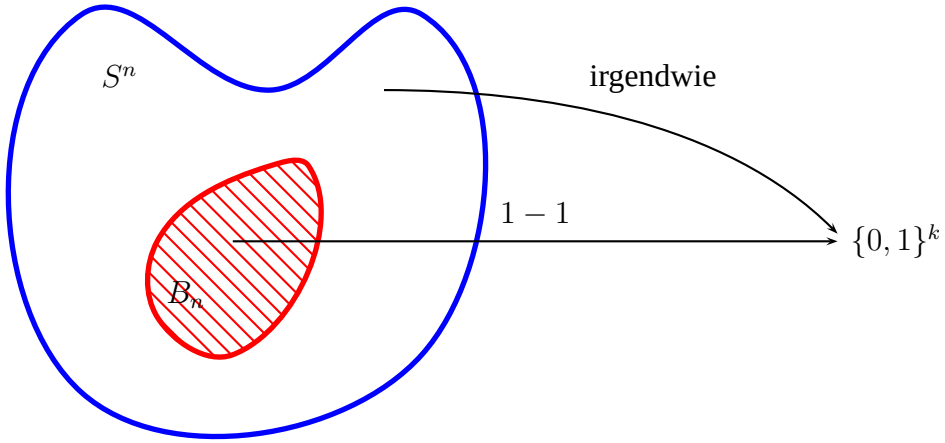


Abbildung 7.6: Effiziente Kodierung bzgl. einer Folge von wesentlichen Mengen B_n.

Korollar 7.16 (Maßkonzentrationssatz von McMillan). *Für jedes $\varepsilon > 0$ ist die Folge*

$$B_n := \left\{(x_1, ...x_n) \in S^n \;\middle|\; e^{-n(H(\mu)+\varepsilon)} \leq p_n(x_1, ..., x_n) \leq e^{-n(H(\mu)-\varepsilon)}\right\}, \qquad n \in \mathbb{N},$$

wesentlich bzgl. μ, und es gilt

$$|B_n| \leq e^{n(H(\mu)+\varepsilon)} \qquad \textit{für alle } n \in \mathbb{N}.$$

Beweis. Es gilt

$$B_n \quad = \quad \left\{(x_1, ...x_n) \in S^n \;\middle|\; H(\mu) - \varepsilon \leq -\frac{1}{n} \log p_n(x_1, ..., x_n) \leq H(\mu) + \varepsilon\right\}. \qquad (7.4.4)$$

Da aus der fast sicheren Konvergenz von $-\frac{1}{n}\log p_n(X_1, ..., X_n)$ gegen die Entropie $H(\mu)$ die stochastische Konvergenz folgt, ist die Folge B_n $(n \in \mathbb{N})$ nach Satz 7.15 wesentlich bzgl. μ. Zudem gilt wegen $p_n(x_1, ..., x_n) \geq e^{-n(H(\mu)+\varepsilon)}$ für $(x_1, \ldots, x_n) \in B_n$:

$$1 \;\geq\; P[(X_1, ..., X_n) \in B_n] \;=\; \sum_{x \in B_n} p_n(x_1, ..., x_n) \;\geq\; |B_n| \cdot e^{-n(H(\mu)+\varepsilon)},$$

also $|B_n| \leq e^{n(H(\mu)+\varepsilon)}$ □

Der Maßkonzentrationssatz zeigt, dass Folgen von wesentlichen Mengen existieren, die auf der exponentiellen Skala nicht viel schneller als $\exp(n \cdot H(\mu))$ wachsen.

Wie groß sind wesentliche Mengen mindestens? Für $p \in (0, 1)$ sei

$$K(n, p) \;=\; \inf\{|A_n| \,|A_n \subseteq S^n \text{ mit } P[(X_1, ..., X_n) \in A_n] \geq p\}$$

die mindestens benötigte Anzahl von Wörtern, um den Text $(X_1, ..., X_n)$ mit Wahrscheinlichkeit $\geq p$ korrekt zu erfassen. Dann ist $\log_2 K(n, p)$ die für eine korrekte binäre Kodierung von $(X_1, ..., X_n)$ mit Wahrscheinlichkeit $\geq p$ mindestens benötigte Anzahl von Bits.

Satz 7.17 (Quellenkodierungssatz von Shannon). *Für alle $p \in (0,1)$ gilt:*

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} \log K(n,p) &= H(\mu), \quad \textit{bzw.} \\ \lim_{n\to\infty} \frac{1}{n} \log_2 K(n,p) &= H_2(\mu) := -\sum_{x:\mu(x)\neq 0} \mu(x) \log_2 \mu(x). \end{aligned}$$

Insbesondere gilt: Ist A_n $(n \in \mathbb{N})$ wesentlich bzgl. μ, so ist

$$\liminf_{n\to\infty} \frac{1}{n} \log |A_n| \quad \geq \quad H(\mu).$$

Bemerkung. (1). Die Größe $\frac{1}{n} \log_2 K(n,p)$ kann als die für eine mit Wahrscheinlichkeit $\geq p$ korrekte Kodierung benötigte Zahl von Bits pro gesendetem Buchstaben interpretiert werden.

(2). Der Quellenkodierungssatz zeigt, dass es keine Folge von wesentlichen Mengen gibt, die auf der exponentiellen Skala deutlich langsamer wächst als die im Maßkonzentrationssatz konstruierten Folgen.

Beweis. Wir zeigen separat eine obere und eine untere Schranke für $\frac{1}{n} \log K(n,p)$:

Obere Schranke: $\limsup_{n\to\infty} \frac{1}{n} \log K(n,p) \leq H(\mu)$:

Zum Beweis sei $\varepsilon > 0$ gegeben. Nach Korollar 7.16 ist die Folge

$$B_n = \left\{ x \in S^n \,\middle|\, e^{-n(H(\mu)+\varepsilon)} \leq p_n(x_1,...,x_n) \leq e^{-n(H(\mu)-\varepsilon)} \right\}$$

wesentlich bzgl. μ, und $\frac{1}{n} \log |B_n| \leq H(\mu) + \varepsilon$. Wegen

$$\lim_{n\to\infty} P[(X_1,...,X_n) \in B_n] \quad = \quad 1 \quad > \quad p, \tag{7.4.5}$$

folgt

$$\limsup_{n\to\infty} \frac{1}{n} \log K(n,p) \quad \leq \quad \limsup_{n\to\infty} \frac{1}{n} \log |B_n| \quad \leq \quad H(\mu) + \varepsilon.$$

Die Behauptung ergibt sich für $\varepsilon \to 0$.

Untere Schranke: $\liminf_{n\to\infty} \frac{1}{n} \log K(n,p) \geq H(\mu)$:

Seien $A_n \subseteq S^n$ mit $P[(X_1,...,X_n) \in A_n] \geq p$. Dann gilt wegen (7.4.5) und (7.4.4) auch

$$p \quad \leq \quad \liminf_{n\to\infty} P[(X_1,...,X_n) \in A_n \cap B_n] \quad \leq \quad \liminf_{n\to\infty} \left(|A_n \cap B_n| \cdot e^{-n(H(\mu)-\varepsilon)} \right),$$

also für alle $\varepsilon > 0$

$$\liminf_{n\to\infty} \frac{1}{n} \log |A_n| \quad \geq \quad \liminf_{n\to\infty} \frac{1}{n} \log |A_n \cap B_n| \quad \geq \quad H(\mu) - \varepsilon.$$

Für $\varepsilon \to 0$ folgt

$$\liminf_{n\to\infty} \frac{1}{n} \log |A_n| \quad \geq \quad H(\mu).$$

□

Kapitel 8

Grenzwertsätze

Sind $X_i : \Omega \to \mathbb{R}, i \in \mathbb{N}$, unabhängige identisch verteilte (i.i.d.) Zufallsvariablen mit Erwartungswert m, dann konvergieren die Mittelwerte $\frac{S_n}{n}$ der Summen $S_n = \sum_{i=1}^{n} X_i$ nach dem Gesetz der großen Zahlen für $n \to \infty$ fast sicher gegen m. Wir wollen nun die Verteilung von S_n für große n genauer untersuchen. Dabei unterscheidet man zwei unterschiedliche Arten von Aussagen:

- *Zentrale Grenzwertsätze* beschreiben „typische“ Fluktuationen um den Grenzwert aus dem Gesetz der großen Zahlen, d.h. die asymptotische Form der Verteilung von S_n/n in Bereichen der Größenordnung $O(1/\sqrt{n})$ um den Erwartungswert m, siehe Abschnitt 8.4.

- Aussagen über *große Abweichungen* beschreiben asymptotisch die Wahrscheinlichkeiten der seltenen Abweichungen der Größenordnung $O(1)$ von S_n/n vom Erwartungswert m. Diese Wahrscheinlichkeiten fallen unter geeigneten Voraussetzungen exponentiell ab, siehe Abschnitt 8.2.

Mit dem Satz von de Moivre/Laplace bzw. der Bernsteinungleichung haben wir bereits entsprechende Aussagen kennengelernt, falls die X_i Bernoulli-verteilte Zufallsvariablen sind. In diesem Kapitel werden wir sehen, dass keine spezifische Form der Verteilung vorausgesetzt werden muss, sondern die Aussagen ganz allgemein unter geeigneten Integrierbarkeitsbedingungen gelten.

Ein wichtiges Hilfsmittel zum Beweis allgemeiner Grenzwertsätze sind momentenerzeugende und charakteristische Funktionen:

8.1 Charakteristische und Momentenerzeugende Funktionen

In diesem Abschnitt führen wir charakteristische und momentenerzeugende Funktionen von reellen Zufallsvariablen ein und beweisen einige grundlegende Aussagen über diese Funktionen. Insbesondere zeigen wir, dass sich die Verteilung einer reellen Zufallsvariable eindeutig aus ihrer charakteristischen Funktion rekonstruieren lässt.

Definition und Eigenschaften

Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum und $X : \Omega \to \mathbb{R}$ eine reellwertige Zufallsvariable mit Verteilung μ.

Definition. *(1). Die Funktion $M : \mathbb{R} \to (0, \infty]$,*

$$M(t) \quad := \quad E[e^{tX}] \quad = \quad \int_{\mathbb{R}} e^{tx}\, \mu(dx),$$

*heißt **momentenerzeugende Funktionen** der Zufallsvariable X bzw. der Verteilung μ.*

(2). Die Funktion $\phi : \mathbb{R} \to \mathbb{C}$,

$$\phi(t) \quad := \quad E[e^{itX}] \quad = \quad \int_{\mathbb{R}} e^{itx}\, \mu(dx),$$

*heißt **charakteristische Funktion** von X bzw. μ.*

Da die Funktionen $t \mapsto e^{tx}$ und $t \mapsto e^{itx}$ für $t \in \mathbb{R}$ nichtnegativ bzw. beschränkt sind, sind die Erwartungswerte definiert. Dabei wird der Erwartungswert einer komplexwertigen Zufallsvariable separat für Real- und Imaginärteil berechnet.

Rechenregeln Die folgenden Rechenregeln ergeben sich unmittelbar aus der Definition:

(1). Sind X und Y unabhängige reellwertige Zufallsvariablen auf $(\Omega, \mathcal{A}, P)$, dann gilt

$$M_{X+Y}(t) \quad = \quad M_X(t) \cdot M_Y(t) \qquad \text{und} \qquad \phi_{X+Y}(t) \quad = \quad \phi_X(t) \cdot \phi_Y(t)$$

für alle $t \in \mathbb{R}$.

(2). Für $a, b \in \mathbb{R}$ gilt

$$M_{aX+b}(t) \quad = \quad e^{bt} \cdot M_X(at) \qquad \text{und} \qquad \phi_{aX+b}(t) \quad = \quad e^{ibt} \cdot \phi_X(at)$$

für alle $t \in \mathbb{R}$.

(3). Für momentenerzeugende bzw. charakteristische Funktionen gilt stets

$$M(0) \quad = \quad \phi(0) \quad = \quad 1, \qquad \text{und}$$

$$\phi(-t) \quad = \quad \overline{\phi(t)} \qquad \text{für alle } t \in \mathbb{R}.$$

Die Funktion $\phi(-t) = \int e^{-itx}\,\mu(dx)$ ist die *Fouriertransformation* des Maßes μ. Ist μ absolutstetig bzgl. des Lebesguemaßes mit Dichte f, dann ist $\phi(-t)$ die Fouriertransformation der Funktion f:

$$\phi(t) \quad = \quad \int\limits_{\mathbb{R}} e^{-itx} f(x)\,dx \quad = \quad \widehat{f}(t).$$

Entsprechend ist

$$M(-t) \quad = \quad \int\limits_{\mathbb{R}} e^{-tx}\,\mu(dx) \qquad (t > 0)$$

die *Laplacetransformation* des Maßes μ bzw. der Dichte f.

Bemerkung (Zusammenhang von M und ϕ). (1). Gilt $M(s) < \infty$ für ein $s > 0$ (bzw. analog für ein $s < 0$), dann ist M auf dem Intervall $[0, s]$ (bzw. $[s, 0]$) endlich, denn nach der Jensenschen Ungleichung folgt:

$$M(t) \quad = \quad E[e^{tX}] \quad \leq \quad E[e^{sX}]^{t/s} \quad < \quad \infty \qquad \text{für alle } \; t \in [0, s] \text{ bzw. } t \in [s, 0].$$

(2). Gilt $M(t) < \infty$ auf $(-\delta, \delta)$ für ein $\delta > 0$, dann ist M analytisch fortsetzbar auf den Streifen $\{z \in \mathbb{C} \,:\, |Re(z)| < \delta\}$ in der komplexen Zahlenebene, und es gilt

$$\phi(t) \quad = \quad M(it) \qquad \text{für alle } t \in \mathbb{R}.$$

Die letzte Bemerkung ermöglicht manchmal eine vereinfachte Berechnung der charakteristischen Funktion.

Beispiel. (1). Für eine standardnormalverteilte Zufallsvariable Z gilt:

$$M_Z(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx - x^2/2} dx = e^{t^2/2}\, \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-t)^2/2} dx = e^{t^2/2} < \infty \quad \text{für alle } \; t \in \mathbb{R}.$$

Also ist die charakteristische Funktion gegeben durch

$$\phi_Z(t) \quad = \quad M_Z(it) \quad = \quad e^{-t^2/2} \qquad \text{für alle } \; t \in \mathbb{R}.$$

(2). Eine normalverteilte Zufallsvariable X mit Mittel m und Varianz σ^2 können wir darstellen als $X = \sigma Z + m$ mit $Z \sim N(0,1)$. Also gilt:

$$\begin{aligned} M_X(t) &= e^{mt}\, M_Z(\sigma t) = \exp\left(mt + \frac{\sigma^2 t^2}{2}\right), \\ \phi_X(t) &= \exp\left(imt - \frac{\sigma^2 t^2}{2}\right). \end{aligned}$$

Sind $X_1, ..., X_n$ unabhängige, $N(m, \sigma^2)$-verteilte Zufallsvariablen, dann erhalten wir:

$$\phi_{X_1+...+X_n}(t) = \prod_{i=1}^{n} \phi_{X_i}(t) = \exp\left(inmt - \frac{n\sigma^2 t^2}{2}\right).$$

Da die rechte Seite die charakteristische Funktion von $N(nm, n\sigma^2)$ ist, folgt nach dem Fourierinversionssatz (s.u., Satz 8.2):

$$X_1 + ... + X_n \;\sim\; N(nm, n\sigma^2)\,.$$

(3). Die Binomialverteilung mit Parametern n und p ist die Verteilung der Summe $\sum_{i=1}^{n} Y_i$ von unabhängigen $\mathrm{Bernoulli}(p)$-verteilten Zufallsvariablen $Y_1, ..., Y_n$. Also sind

$$\begin{aligned} \phi(t) &= \prod_{i=1}^{n} \phi_{Y_i}(t) = \left(1 - p + pe^{it}\right)^n, \quad \text{und} \\ M(t) &= \left(1 - p + pe^{t}\right)^n \end{aligned}$$

die charakteristische und momentenerzeugende Funktion von $\mathrm{Bin}(n, p)$.

(4). Die *Cauchyverteilung* ist die absolutstetige Wahrscheinlichkeitsverteilung auf $\mathbb{R}$ mit Dichte

$$f(x) \;=\; \frac{1}{\pi(1+x^2)} \qquad (x \in \mathbb{R}).$$

Für eine Cauchyverteilte Zufallsvariable X gilt $M_X(t) = \infty$ für alle $t \neq 0$ (und sogar $E[|X|^n] = \infty \quad \forall n \in \mathbb{N}$). Trotzdem existiert

$$\phi_X(t) = e^{-|t|} \qquad \text{für alle } t \in \mathbb{R}\,.$$

Die charakteristische Funktion ist allerdings bei 0 nicht differenzierbar.

Wir zeigen nun, dass sich die Momente $E[X^n]$ einer Zufallsvariable $X : \Omega \to \mathbb{R}$ unter geeigneten Voraussetzungen aus der momentenerzeugenden bzw. charakteristischen Funktion berechnen lassen. Die nötigen Voraussetzungen sind allerdings im Fall der momentenerzeugenden Funktion viel stärker:

Satz 8.1. *(1). Ist M endlich auf $(-\delta, \delta)$, $\delta > 0$, dann gilt*

$$E[e^{zX}] \quad = \quad \sum_{n=0}^{\infty} \frac{z^n}{n!} E[X^n] \qquad \textit{für alle } z \in \mathbb{C} \textit{ mit } |z| < \delta.$$

Insbesondere folgt

$$M(t) \quad = \quad \sum_{n=0}^{\infty} \frac{t^n}{n!}\, E[X^n] \qquad \textit{für alle } t \in (-\delta, \delta)\,,$$

und somit

$$M^{(n)}(0) = E[X^n] \qquad \textit{für alle } n \geq 0\,.$$

(2). Ist $E[|X|^n] < \infty$ für ein $n \in \mathbb{N}$, dann gilt $\phi \in C^n(\mathbb{R})$ und

$$\phi^{(n)}(t) \quad = \quad i^n \cdot E[X^n e^{itX}] \qquad \textit{für alle } t \in \mathbb{R}\,. \tag{8.1.1}$$

Beweis. (1). Aus der Voraussetzung und dem Satz von der monotonen Konvergenz folgt für $s \in (0, \delta)$:

$$\sum_{n=0}^{\infty} \frac{s^n}{n!} E[|X|^n] \quad = \quad E\left[e^{s|X|}\right] \quad \leq \quad E\left[e^{sX}\right] + E\left[e^{-sX}\right] \quad < \quad \infty\,.$$

Insbesondere existieren alle Momente $E[X^n], n \in \mathbb{N}$, sowie die exponentiellen Momente $E[e^{zX}]$ für $z \in \mathbb{C}$ mit $|\mathrm{Re}(z)| < \delta$. Nach dem Satz von Lebesgue erhalten wir für diese z zudem

$$\sum_{n=0}^{\infty} \frac{z^n}{n!}\, E[X^n] = \lim_{m\to\infty} E\left[\sum_{n=0}^{m} \frac{(zX)^n}{n!}\right] = E\left[\lim_{m\to\infty} \sum_{n=0}^{m} \frac{(zX)^n}{n!}\right] = E[e^{zX}]\,,$$

da $e^{s|X|}$ für $s \geq |z|$ eine Majorante der Partialsummen ist.

(2). Wir zeigen die Behauptung durch Induktion nach n. Für $n = 0$ gilt (8.1.1) nach Definition von $\phi(t)$. Ist $E[|X|^{n+1}] < \infty$, dann folgt nach Induktionsvoraussetzung und mit dem Satz von Lebesgue:

$$\begin{aligned} \frac{\phi^{(n)}(t+h) - \phi^{(n)}(t)}{h} \quad &= \quad \frac{1}{h} E\left[(iX)^n \left(e^{i(t+h)X} - e^{itX}\right)\right] \\ &= \quad E\left[(iX)^n \frac{1}{h} \int_t^{t+h} iX e^{isX}\, ds\right] \quad \to \quad E\left[(iX)^{n+1}\, e^{itX}\right] \end{aligned}$$

für $h \to 0$, also

$$\phi^{n+1}(t) \quad = \quad E[(iX)^{n+1} \cdot e^{itX}].$$

Die Stetigkeit der rechten Seite in t folgt ebenfalls aus dem Satz von Lebesgue und der Voraussetzung $E[|X|^{n+1}] < \infty$.

□

Beispiel. Für eine Zufallsvariable X mit Dichte $f_X(x) = const. \cdot e^{-|x|^{1/2}}$ gilt $E[|X|^n] < \infty$ für alle $n \in \mathbb{N}$. Also ist die charakteristische Funktion beliebig oft differenzierbar. Die momentenerzeugende Funktion $M(t) = E[e^{tX}]$ ist hingegen nur für $t = 0$ endlich.

Bemerkung (Satz von Bochner). Eine Funktion $\phi : \mathbb{R} \to \mathbb{C}$ ist genau dann eine charakteristische Funktion einer Wahrscheinlichkeitsverteilung auf $\mathbb{R}$, wenn gilt:

(1). $\phi(0) = 1$ und $|\phi(t)| \leq 1$ für alle $t \in \mathbb{R}$.

(2). ϕ ist gleichmäßig stetig.

(3). ϕ ist *nicht negativ definit*, d.h.

$$\sum_{i,j=1}^{n} \phi(t_i - t_j) z_i \overline{z_j} \geq 0 \quad \forall n \in \mathbb{N},\ t_1, ..., t_n \in \mathbb{R},\ z_1, ..., z_n \in \mathbb{C}.$$

Dass jede charakteristische Funktion einer Wahrscheinlichkeitsverteilung die Eigenschaften (1)-(3) hat, prüft man leicht nach (Übung). Der Beweis der umgekehrten Aussage findet sich z.B. in Vol. II des Lehrbuchs von Feller.

Inversion der Fouriertransformation

Die folgende zentrale Aussage zeigt, dass eine Wahrscheinlichkeitsverteilung *eindeutig* durch ihre charakteristische Funktion ϕ festgelegt ist, und liefert eine *explizite Formel* zur Rekonstruktion der Verteilung aus ϕ:

Satz 8.2 (Lévys Inversionsformel). *Sei ϕ die charakteristische Funktion einer Zufallsvariable X mit Verteilung μ. Dann gilt:*

(1).

$$\frac{1}{2}\,\mu[\{a\}] \;+\; \mu[(a,b)] \;+\frac{1}{2}\,\mu[\{b\}] \;=\; \frac{1}{2\pi}\,\lim_{T\to\infty} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it}\,\phi(t)\,dt \qquad \forall\, a < b\,.$$

(2). Gilt $\int_{-\infty}^{\infty} |\phi(t)|\,dt < \infty$, dann ist μ absolutstetig mit stetiger Dichte

$$f(x) \;=\; \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx}\,\phi(t)\,dt.$$

Bemerkung. (1). Die Verteilung μ ist durch (1) eindeutig festgelegt, denn für $c, d \in \mathbb{R}$ mit $c < d$ gilt:

$$\frac{1}{2}\,\mu[\{a\}] + \mu[(a,b)] + \frac{1}{2}\,\mu[\{b\}] \quad = \quad \frac{1}{2}\,\Big(\mu\Big[[a,b]\Big] + \mu\Big[(a,b)\Big]\Big) \quad \to \quad \mu[(c,d)]\,,$$

für $a \searrow c$ und $b \nearrow d$.

(2). Ist die Verteilung μ absolutstetig mit quadratintegrierbarer Dichte f, dann ist auch die entsprechende charakteristische Funktion

$$\phi(t) \quad = \quad \int_{-\infty}^{\infty} e^{itx} f(x)\, dx$$

quadratintegrierbar. Die Aussage (2) aus Satz 8.2 ist in diesem Fall die klassische *Fourierinversionsformel der Analysis*, siehe z.B. Forster „Analysis 3“.

Im Beweis der Inversionsformel verwenden wir den Satz von Fubini, der besagt, dass wir die Integrationsreihenfolge in Doppelintegralen vertauschen dürfen, wenn der Integrand produktintegrierbar ist. Für den Beweis des Satzes von Fubini verweisen wir auf die Analysisvorlesung oder Abschnitt 9.1.

von Satz 8.2. (1). Sei $T > 0$ und $a < b$. Nach dem Satz von Fubini können wir die Integrationsreihenfolge in dem folgendem Doppelintegral vertauschen, und erhalten:

$$\frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \underbrace{\phi(t)}_{= \int e^{itx}\, \mu(dx)}\, dt \quad = \quad \frac{1}{\pi} \int \underbrace{\int_{-T}^{T} \frac{e^{it(x-a)} - e^{it(x-b)}}{2it}\, dt}_{=:\, g(T,x)}\, \mu(dx) \quad (8.1.2)$$

Dabei haben wir benutzt, dass der Integrand produktintegrierbar ist, da aus der Lipschitz-Stetigkeit der Abbildung $y \mapsto e^{iy}$ mit Konstante $L = 1$ folgt, dass

$$\left| \frac{e^{it(x-a)} - e^{it(x-b)}}{it} \right| \leq \frac{|t \cdot (x-a) - t \cdot (x-b)|}{|t|} = |a-b| \qquad \text{gilt.}$$

Weiterhin erhalten wir, wegen $e^{it(x-a)} = \cos(t \cdot (x-a)) + i \sin(t \cdot (x-a))$, $\cos(x) = \cos(-x)$ und $\sin(x) = -\sin(-x)$:

$$\begin{aligned} g(T,x) &= \int_0^T \frac{\sin(t \cdot (x-a))}{t}\, dt - \int_0^T \frac{\sin(t \cdot (x-b))}{t}\, dt \\ &= \int_0^{T \cdot (x-a)} \frac{\sin u}{u}\, du - \int_0^{T \cdot (x-b)} \frac{\sin u}{u}\, du \\ &= S(T \cdot (x-a)) - S(T \cdot (x-b)) \end{aligned}$$

wobei

$$S(t) := \int_0^t \frac{\sin u}{u}\, du$$

der Integralsinus ist. Mithilfe des Residuensatzes (siehe Funktionentheorie) zeigt man:

$$\lim_{t \to \infty} S(t) \quad = \quad \frac{\pi}{2} \quad , \qquad \lim_{t \to -\infty} S(t) \quad = \quad -\frac{\pi}{2}\,.$$

Damit erhalten wir:

$$\lim_{T\to\infty} g(T,x) \;=\; \frac{\pi}{2}\operatorname{sgn}(x-a) \;-\; \frac{\pi}{2}\operatorname{sgn}(x-b) \;=\; \pi\cdot I_{(a,b)}(x) \;+\; \frac{\pi}{2}\cdot I_{\{a,b\}}(x)\;,$$

wobei wir $\operatorname{sgn}(0) := 0$ setzen. Da S beschränkt ist, ist auch $g(T,x)$ beschränkt in T und x. Nach dem Satz von Lebesgue folgt daher aus (8.1.2) für $T \to \infty$

$$\begin{aligned}\frac{1}{2\pi}\int_{-T}^{T}\frac{e^{-ita}-e^{-itb}}{it}\,\phi(t)\,dt \;&=\; \frac{1}{\pi}\int g(T,x)\,\mu(dx)\\ &\overset{T\to\infty}{\longrightarrow}\quad \mu[(a,b)] \;+\; \frac{1}{2}\,\mu[\{a,b\}]\;.\end{aligned}$$

(2). Ist ϕ integrierbar, dann ist die Funktion $(t,x)\mapsto e^{-itx}\,\phi(t)$ produktintegrierbar auf $[a,b]\times\mathbb{R}$ für alle $-\infty < a < b < \infty$. Also ist die Funktion

$$f(x) \;:=\; \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\,\phi(t)\,dt$$

integrierbar auf $[a,b]$, und es gilt nach dem Satz von Fubini und (1):

$$\int_a^b f(x)\,dx = \frac{1}{2\pi}\int_{-\infty}^{\infty}\phi(t)\underbrace{\int_a^b e^{-itx}\,dx}_{=\frac{e^{-ita}-e^{-itb}}{it}}dt \overset{(1)}{=} \frac{1}{2}\,\mu[\{a\}] + \mu[(a,b)] + \frac{1}{2}\,\mu[\{b\}]\;.$$

Insbesondere folgt

$$\int_{a+\varepsilon}^{b-\varepsilon} f(x)\,dx \;\le\; \mu\left[(a,b)\right] \;\le\; \int_a^b f(x)\,dx \quad \forall\,\varepsilon>0,$$

also für $\varepsilon \searrow 0$:

$$\mu[(a,b)] \;=\; \int_a^b f(x)\,dx\;.$$

□

8.2 Erste Anwendungen auf Grenzwertsätze

Charakteristische und momentenerzeugende Funktionen werden häufig beim Beweis von Grenzwertsätzen der Wahrscheinlichkeitstheorie vewendet. Wir skizzieren an dieser Stelle schon einmal die Anwendung charakteristischer Funktionen zum Beweis des zentralen Grenzwertsatzes und zeigen anschließend, wie obere Schranken für die Wahrscheinlichkeiten großer Abweichungen vom Gesetz der großen Zahlen mithilfe momentenerzeugender Funktionen hergeleitet werden können. Der detaillierte Beweis des zentralen Grenzwertsatzes wird dann nach weiteren Vorbereitungen in Abschnitt 8.3 ausgeführt. Die Analyse der Asymptotik der Wahrscheinlichkeiten großer Abweichungen auf der exponentiellen Skala werden wir in Kapitel 11 durch den Beweis einer unteren Schranke vervollständigen.

Zentraler Grenzwertsatz

Seien $X_1, X_2, ... \in \mathscr{L}^2(\Omega, \mathcal{A}, P)$ unabhängige und identisch verteilte Zufallsvariablen mit $E[X_i] = 0$ für alle i, und sei $S_n = X_1 + ... + X_n$. Nach dem Gesetz der großen Zahlen gilt:

$$\frac{S_n}{n} \quad \to \quad 0 \qquad P\text{-fast sicher.}$$

Wie sieht die Verteilung von S_n für große n aus?

Um eine asymptotische Darstellung zu erhalten, reskalieren wir zunächst so, dass die Varianz konstant ist. Es gilt

$$\mathrm{Var}[S_n] = n \cdot \mathrm{Var}[X_1],$$

also ist

$$\mathrm{Var}\left[\frac{S_n}{\sqrt{n}}\right] = \frac{1}{n} \cdot \mathrm{Var}[S_n] = \mathrm{Var}[X_1] =: \sigma^2$$

unabhängig von n.

Um die Asymptotik der Verteilungen der entsprechend standardisierten Summen $\frac{S_n}{\sqrt{n}}$ zu bestimmen, betrachten wir die charakteristischen Funktionen. Da die Summanden X_i unabhängig und identisch verteilt sind, erhalten wir

$$\phi_{\frac{S_n}{\sqrt{n}}}(t) = \phi_{S_n}\left(\frac{t}{\sqrt{n}}\right) \overset{X_i \text{ iid}}{=} \left[\phi_{X_1}\left(\frac{t}{\sqrt{n}}\right)\right]^n .$$

Wegen $X_1 \in \mathscr{L}^2(\Omega, \mathcal{A}, P)$ ist ϕ_{X_1} zweimal stetig differenzierbar, und die Taylorentwicklung bei $t = 0$ ist gegeben durch

$$\phi_{X_1}(t) \quad = \quad 1 + i \cdot E[X_1] \cdot t - \frac{1}{2}\, E[X_1^2] \cdot t^2 + o(t^2) \quad = \quad 1 - \frac{1}{2}\sigma^2 t^2 + o(t^2).$$

Damit folgt:

$$\begin{aligned} \phi_{\frac{S_n}{\sqrt{n}}}(t) \quad &= \quad \left(1 - \frac{\sigma^2 t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n \\ &\overset{n \nearrow \infty}{\longrightarrow} \quad \exp\left(-\frac{\sigma^2 t^2}{2}\right) \;=\; \phi_{N(0,\sigma^2)}(t) \qquad \forall\, t \in \mathbb{R}. \end{aligned}$$

Wir werden im nächsten Abschnitt zeigen, dass aus der Konvergenz der charakteristischen Funktionen unter geeigneten Voraussetzungen die schwache Konvergenz (Definition s.u.) der Verteilungen folgt. Somit ergibt sich:

Zentraler Grenzwertsatz: Die Verteilung der standardisierten Summen $\frac{S_n}{\sqrt{n}}$ konvergiert schwach gegen die Normalverteilung $N(0, \sigma^2)$.

Den detaillierten Beweis werden wir in Abschnitt 8.3 führen. Der zentrale Grenzwertsatz erklärt, warum die Normalverteilungen in der Stochastik von so großer Bedeutung sind:

Bemerkung (Universalität der Normalverteilung). Die *Limesverteilung im zentralen Grenzwertsatz ist unabhängig von der Verteilung von* X_1, vorausgesetzt, es gilt $X_1 \in \mathscr{L}^2(\Omega, \mathcal{A}, P)$.

Große Abweichungen vom Gesetz der großen Zahlen

Seien $X_1, X_2, ... \in \mathscr{L}^1(\Omega, \mathcal{A}, P)$ i.i.d. Zufallsvariablen mit Erwartungswert m und momentenerzeugender Funktion

$$M(t) = E[e^{tX_1}] \,,$$

und sei $S_n = X_1 + ... + X_n$.
Der folgende Satz verschärft die *nicht-asymptotische* obere Schranke für große Abweichungen vom Gesetz der großen Zahlen aus der Bernstein-Ungleichung (Satz 2.6), und verallgemeinert diese auf nicht Bernoulliverteilte Zufallsvariablen.

Satz 8.3 (Chernoff). *Für alle* $n \in \mathbb{N}$ *und* $a \in \mathbb{R}$ *gilt:*

$$\begin{aligned} P\left[\frac{S_n}{n} \geq a\right] &\leq e^{-nI(a)} \quad \textit{falls } a \geq m, \textit{ bzw.} \\ P\left[\frac{S_n}{n} \leq a\right] &\leq e^{-nI(a)} \quad \textit{falls } a \leq m, \end{aligned}$$

wobei die exponentielle Abfallrate $I(a)$ *gegeben ist durch*

$$I(a) = \sup_{t \in \mathbb{R}} (at - \log M(t)).$$

Beweis. Wir zeigen diese Aussage im Fall $a \geq m$ – der Beweis für $a \leq m$ verläuft analog. Der Beweis erfolgt in drei Schritten:

(1). *Zentrieren:* Wir können o.B.d.A. $m = 0$ annehmen. Andernfalls betrachten wir die zentrierten Zufallsvariablen $\widetilde{X}_i = X_i - E[X_i]$, die wieder unabhängig und identisch verteilt sind. Man überzeugt sich leicht, dass aus der Behauptung für $\widetilde{X_i}$ die Behauptung für X_i folgt (Übung).

(2). *Exponentielle Markovungleichung:* Für alle $t \geq 0$ gilt:

$$\begin{aligned} P\left[\frac{S_n}{n} \geq a\right] &= P[S_n \geq na] \leq e^{-tna} E[e^{tS_n}] \\ &\overset{X_i \text{ iid}}{=} e^{-tna}\, E[e^{tX_1}]^n = e^{-(at - \log M(t)) \cdot n}. \end{aligned}$$

(3). *Optimieren der Abschätzung:* Bilden wir das Infimum der für verschiedene $t \geq 0$ erhaltenen Abschätzungen, dann ergibt sich:

$$P\left[\frac{S_n}{n} \geq a\right] \quad \leq \quad \inf_{t\geq 0} e^{-(at-\log M(t))\cdot n} \quad = \quad e^{-\sup_{t\geq 0}(at-\log M(t))\cdot n}.$$

Es bleibt zu zeigen, dass

$$\sup_{t\geq 0}(at - \log M(t)) \quad = \quad \sup_{t\in\mathbb{R}}(at - \log M(t)) \quad = \quad I(a).$$

Dies ist in der Tat der Fall, denn für $t < 0$ und $a \geq 0$ gilt nach der Jensenschen Ungleichung und der Voraussetzung $m = 0$:

$$\begin{aligned} at - \log M(t) &\leq -\log E[e^{tX_1}] \leq -E[\log e^{tX_1}] \\ &= -tm = 0 = a\cdot 0 - \log M(0). \end{aligned}$$

□

Bemerkung (Kumulantenerzeugende Funktion, Legendretransformation). (1). Die Funktion $\Lambda(t) := \log M(t)$ heißt *logarithmische momentenerzeugende* oder *kumulantenerzeugende Funktion* von X_1. Diese Funktion hat u.a. folgende Eigenschaften:

(a) Λ ist konvex und *unterhalbstetig*, d.h. $\liminf_{s\to t} \Lambda(s) \geq \Lambda(t)$ für alle $t \in \mathbb{R}$.

(b) $\Lambda(0) = 0$.

(c) Gilt $M(t) < \infty$ auf $(-\delta, \delta)$ für ein $\delta > 0$, dann ist

$$\begin{aligned} \Lambda'(0) &= \frac{M'(0)}{M(0)} = m, \quad \text{und} \\ \Lambda''(0) &= \frac{M''(0)}{M(0)} - \frac{M'(0)^2}{M(0)^2} = E[X_1^2] - E[X_1]^2 = \mathrm{Var}[X_1]. \end{aligned}$$

Die höheren Ableitungen von Λ heißen *Kumulanten* von X_1.

(2). Die Ratenfunktion I ist die *Legendre-Transformation von* Λ:

$$I(a) = \sup_{t\in\mathbb{R}} f_a(t) \qquad \text{mit} \qquad f_a(t) = at - \Lambda(t),$$

d.h. $I(a)$ ist der negative Achsenabschnitt der (eindeutigen) Tangente an den Graphen von Λ mit Steigung a (wobei $I(a) = \infty$, falls keine solche Tangente existiert).

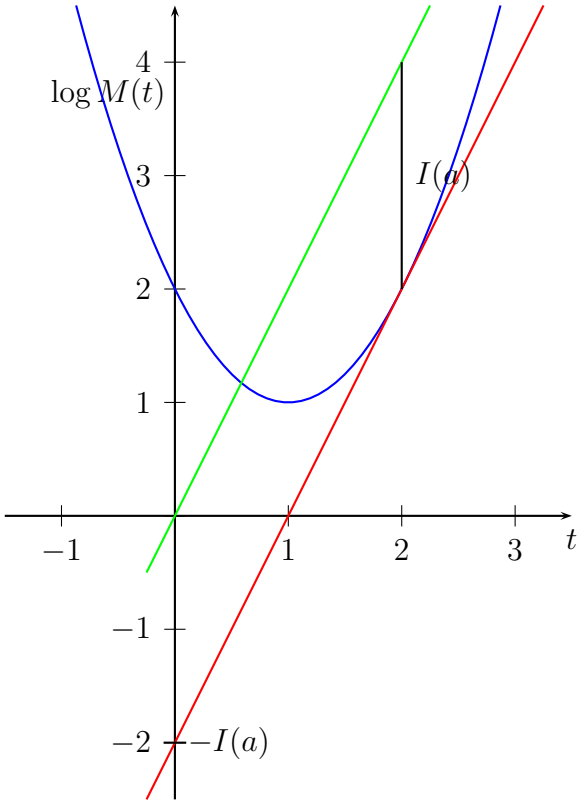


Abbildung 8.1: Geometrische Darstellung der Ratenfunktion $I(a)$ als negativer Achsenabschnitt der eindeutigen Tangente mit Steigung a (rot) an die Kumulantenerzeugende Funktion (blau)

Wichtige Eigenschaften der Ratenfunktion sind:

(a) I ist wieder konvex und unterhalbstetig.

(b) $I(a) \geq f_a(0) = 0 \qquad \forall a \in \mathbb{R}$.

(c) Gilt $M(t) < \infty$ auf $(-\delta, \delta)$ für ein $\delta > 0$, dann ist $f_a \in C^\infty(-\delta, \delta)$ mit $f_a(0) = 0$ und $f_a'(0) = a - m$. Also folgt:

$$I(a) \quad = \quad \sup f_a \quad > \quad 0 \qquad \forall\, a \neq m.$$

Unter der Voraussetzung der letzten Bemerkung (c) ist die exponentielle Abfallrate strikt positiv, d.h. es ergibt sich ein *exponentieller Abfall der Wahrscheinlichkeiten großer Abweichungen!* Sind die Zufallsvariablen X_i nicht exponentiell integrierbar, dann kann es auch passieren, dass $I(a) = 0$ für $a \neq m$. Die Wahrscheinlichkeiten großer Abweichungen fallen in diesem Fall langsamer als exponentiell ab, denn es gilt auch eine asymptotische untere Schranke mit derselben Ratenfunktion I, siehe Satz 12.7 unten.

Beispiel. Für konkrete Verteilungen der Zufallsvariablen X_i kann man die Kumulantenerzeugende Funktion Λ und die Ratenfunktion I häufig explizit berechnen:

(1). Für normalverteilte Zufallsvariablen $X_i \sim N(m, \sigma^2)$ gilt $I(a) = \frac{(a-m)^2}{2\sigma^2}$, also

$$P\left[\frac{S_n}{n} \geq a\right] \leq e^{-\frac{(a-m)^2 n}{2\sigma^2}} \qquad \text{für alle } a \geq m.$$

Die Ratenfunktion hat eine Nullstelle beim Erwartungswert m, da die Mittelwert S_n/n gegen diese konvergieren. Jenseits von m fallen die Wahrscheinlichkeiten exponentiell ab, und zwar mit einer Rate die quadratisch wächst.

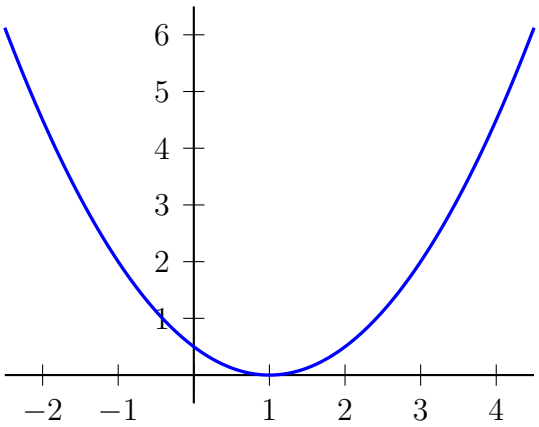


Abbildung 8.2: Legendre-Transformation der logarithmischen momentenerzeugenden Funktion einer $\mathcal{N}(1,1)$-verteilten Zufallsvariable

(2). Für $X_i \sim \mathrm{Exp}(\lambda)$ gilt

$$I(a) = \begin{cases} \lambda a - 1 - \log(\lambda a) & \text{für } a > 0 \\ \infty & \text{für } a \leq 0 \end{cases}.$$

In diesem Fall hat die Ratenfunktion eine Nullstelle beim Erwartungswert $1/\lambda$. Da nicht positive Werte mit Wahrscheinlichkeit 1 nicht auftreten, hat die Ratenfunktion auf dem Intervall $(-\infty, 0]$ den Wert $+\infty$.

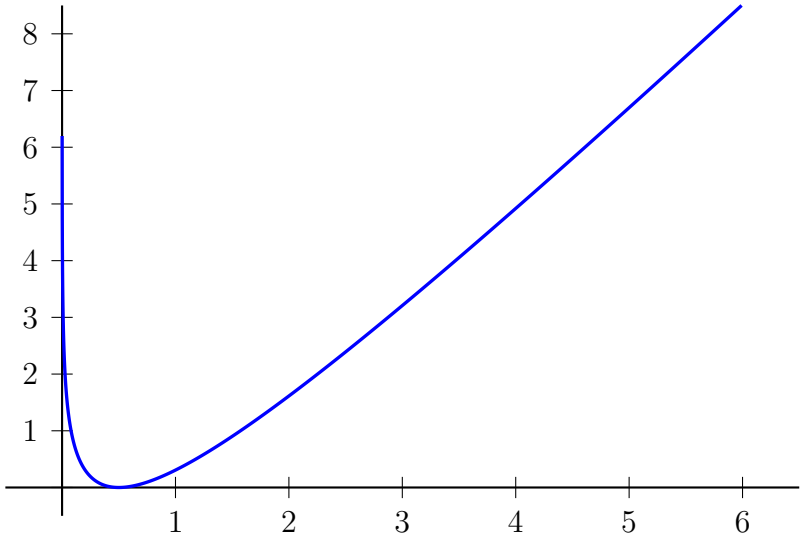


Abbildung 8.3: Legendre-Transformierte der logarithmischen momentenerzeugenden Funktion einer $\mathrm{Exp}(2)$-verteilten Zufallsvariable

(3). Für $X_i \sim \mathrm{Bernoulli}(p)$ erhält man

$$I(a) = a \log\left(\frac{a}{p}\right) + (1-a)\log\left(\frac{1-a}{1-p}\right) \qquad \text{für } a \in (0,1).$$

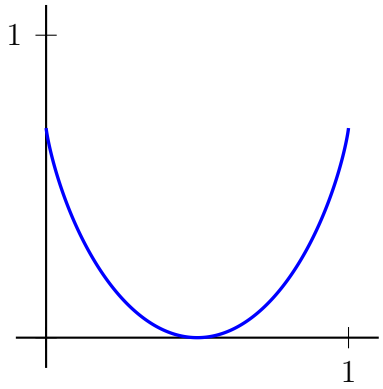


Abbildung 8.4: Legendre-Transformation der logarithmischen momentenerzeugenden Funktion einer Bernoulli$(1/2)$-verteilten Zufallsvariable

Wegen $I(a) \geq 2(a-p)^2$ verschärft die Abschätzung aus dem Satz von Chernoff in diesem Fall die in Satz 2.6 hergeleitete obere Schranke

$$P\left[\frac{S_n}{n} \geq a\right] \leq e^{-2(a-p)^2 n} \qquad \text{für } a \geq p.$$

Wir werden später sehen, dass $I(a)$ sich als relative Entropie der Bernoulli(a)-Verteilung bzgl. der Bernoulli (p)-Verteilung interpretieren lässt.

Beispiel (Ehrenfestmodell im Gleichgewicht). Es befinden sich $n = 10^{23}$ Moleküle in einem Gefäß. Jedes Molekül sei mit Wahrscheinlichkeit $\frac{1}{2}$ in der linken bzw. rechten Hälfte. Seien X_i ($1 \leq i \leq n$) Bernoulli$\left(\frac{1}{2}\right)$-verteilte unabhängige Zufallsvariablen, wobei $X_i = 1$ dafür steht, dass sich das i-te Molekül in der linken Hälfte befindet. Der Anteil S_n/n der Moleküle in dieser Hälfte konvergiert nach dem Gesetz der großen Zahlen fast sicher gegen $1/2$.

Wie groß ist $p := P\left[\frac{S_n}{n} \geq \frac{1}{2} + 10^{-10}\right]$?

Eine Abschätzung mit der Čebyšev-Ungleichung liefert:

$$p \quad \leq \quad 10^{20} \cdot \mathrm{Var}\left[\frac{S_n}{n}\right] \quad = \quad \frac{1}{4} \cdot 10^{-3} \quad = \quad \frac{1}{4000}.$$

Durch Anwenden der exponentiellen Abschätzung erhält man dagegen die viel präzisere Aussage

$$p \quad \leq \quad e^{-2n(10^{-10})^2} \quad = \quad e^{-2000}\,.$$

Eine Abweichung von der Größenordnung 10^{-10} vom Mittelwert ist also **praktisch unmöglich** ! Die makroskopische Größe S_n/n ist daher de facto deterministisch.

8.3 Verteilungskonvergenz

Sei S ein metrischer Raum mit Borelscher σ-Algebra $\mathcal{B}(S)$, zum Beispiel $S = \mathbb{R}$ oder $S = \mathbb{R}^d$. Wir wollen nun einen für den zentralen Grenzwertsatz angemessenen Konvergenzbegriff für die Verteilungen μ_n einer Folge Y_n von Zufallsvariablen mit Werten in S einführen. Naheliegend wäre es zu definieren, dass eine Folge μ_n von Wahrscheinlichkeitsverteilungen auf $(S, \mathcal{B}(S))$ gegen eine Wahrscheinlichkeitsverteilung μ konvergiert, wenn $\mu[A] = \lim \mu_n[A]$ für *jede* Menge $A \in \mathcal{B}(S)$ gilt. Ein solcher Konvergenzbegriff erweist sich jedoch sofort als zu restriktiv, z.B. würde eine Folge von diskreten Wahrscheinlichkeitsverteilungen in diesem Sinne niemals gegen eine Normalverteilung konvergieren. Einen angemesseneren Grenzwertbegriff erhält man durch Berücksichtigung der Topologie auf S:

Definition. *(1). **Schwache Konvergenz von Wahrscheinlichkeitsverteilungen**: Eine Folge $(\mu_n)_{n\in\mathbb{N}}$ von Wahrscheinlichkeitsverteilungen auf S (mit Borelscher σ-Algebra) **konvergiert schwach** gegen eine Wahrscheinlichkeitsverteilung μ auf S ($\mu_n \xrightarrow{w} \mu$), falls*

$$\int f\, d\mu_n \quad \longrightarrow \quad \int f\, d\mu \qquad \textit{für alle stetigen, beschränkten } f : S \to \mathbb{R} \textit{ gilt.}$$

(2). ***Konvergenz in Verteilung von Zufallsvariablen****: Eine Folge $(Y_n)_{n\in\mathbb{N}}$ von Zufallsvariablen mit Werten in S* ***konvergiert in Verteilung*** *gegen eine Zufallsvariable Y bzw. gegen die Verteilung von Y, falls*

$$\textit{Verteilung}(Y_n) \quad \xrightarrow{w} \quad \textit{Verteilung}(Y),$$

d.h. falls

$$E[f(Y_n)] \quad \longrightarrow \quad E[f(Y)] \qquad \textit{für alle } f \in C_b(S) \textit{ gilt.}$$

Konvergenz in Verteilung bezeichnet man auf Englisch als „convergence in distribution“ oder „convergence in law.“ Entsprechend verwendet man die Kurzschreibweisen $Y_n \xrightarrow{\mathcal{D}} Y$ oder $Y_n \xrightarrow{\mathcal{L}} Y$, falls Y_n in Verteilung gegen Y konvergiert.

Beachte: Die Zufallsvariablen $Y_n, n \in \mathbb{N},$ und Y können bei der Verteilungskonvergenz *auf verschiedenen Wahrscheinlichkeitsräumen* definiert sein!

Schwache Konvergenz von Wahrscheinlichkeitsverteilungen

Um den Begriff der schwachen Konvergenz besser zu erfassen, beginnen wir mit einigen Bemerkungen und Beispielen:

Bemerkung. (1). Die hier definierte Form der schwachen Konvergenz entspricht **nicht** der im funktionalanalytischen Sinn definierten schwachen Konvergenz auf dem Vektorraum aller beschränkten signierten Maße auf $(S, \mathcal{B}(S))$, sondern einer schwach$*$-Konvergenz auf diesem Raum, siehe z.B. ALT: LINEARE FUNKTIONALANALYSIS.

(2). Wir werden in Satz 8.5 zeigen, dass im Fall $S = \mathbb{R}$ die Folge μ_n genau dann schwach gegen μ konvergiert, wenn für die Verteilungsfunktionen

$$F_{\mu_n}(x) \quad \longrightarrow \quad F_\mu(x) \qquad \text{für alle Stetigkeitsstellen } x \text{ von } F,$$

d.h. für alle $x \in \mathbb{R}$ mit $\mu[\{x\}] = 0$, gilt.

Neben schwacher Konvergenz betrachtet man häufig u.a. auch die folgenden Konvergenzarten auf positiven bzw. beschränkten signierten Maßen:

- **Vage Konvergenz:** *μ_n konvergiert vage gegen μ,* falls

$$\int f \, d\mu_n \quad \longrightarrow \quad \int f \, d\mu$$

für alle stetigen Funktionen f mit kompaktem Träger gilt.

- **Konvergenz in Variationsdistanz:** *μ_n konvergiert μ in Variationsdistanz,* falls

$$\|\mu - \mu_n\|_{\text{TV}} \quad := \quad \frac{1}{2} \sup_{\substack{f:S\to\mathbb{R} \text{ messbar} \\ \text{mit } |f| \le 1}} \left| \int f \, d\mu - \int f \, d\mu_n \right| \quad \longrightarrow \quad 0.$$

  Die Variationsdistanz zweier Wahrscheinlichkeitsverteilungen lässt sich auch wie folgt darstellen:

$$\|\mu - \nu\|_{\text{TV}} \quad = \quad \sup_{A\in\mathcal{S}} |\mu[A] - \nu[A]|.$$

  Im diskreten Fall gilt

$$\|\mu - \nu\|_{\text{TV}} \quad = \quad \frac{1}{2} \sum_{x\in S} |\mu[\{x\}] - \nu[\{x\}]| \, .$$

  Diesen Abstandsbegriff haben wir bereits in Abschnitt 3.5 bei der Konvergenz ins Gleichgewicht von Markovketten verwendet.

Offensichtlich folgt aus der Konvergenz in Variationsdistanz die schwache Konvergenz, aus der wiederum die vage Konvergenz folgt:

$$\|\mu_n - \mu\|_{\text{TV}} \to 0 \quad \Longrightarrow \quad \mu_n \xrightarrow{w} \mu \quad \Longrightarrow \quad \mu_n \to \mu \quad \text{vage}.$$

Die folgenden Beispiele verdeutlichen die unterschiedlichen Konvergenzbegriffe:

Beispiel. (1). ***Diracmaße:*** Für $x, x_n \in S$ $(n \in \mathbb{N})$ mit $x_n \to x$ gilt $\delta_{x_n} \xrightarrow{w} \delta_x$.

Beweis:

$$\int f \, d\delta_{x_n} \quad = \quad f(x_n) \quad \to \quad f(x) \quad = \quad \int f \, d\delta_x \qquad \text{für alle } f \in C_b(\mathbb{R}).$$

Alternativer Beweis im Fall $S = \mathbb{R}$:

$$F_{\delta_{x_n}}(c) \quad = \quad I_{[x_n,\infty)}(c) \quad \overset{n\to\infty}{\to} \quad I_{[x,\infty)}(c) \quad = \quad F_{\delta_x}(c) \qquad \text{für alle } c \neq x,$$

d.h. für alle Stetigkeitsstellen von F_{δ_x}.

In diesem Beispiel gilt i.A. keine Konvergenz in Variationsnorm, denn $\|\delta_{x_n} - \delta_x\|_{\text{TV}} = 1$ für $x_n \neq x$.

(2). ***Degeneration/Diracfolge:*** Auf $S = \mathbb{R}^1$ konvergiert die Folge $\mu_n := N(0, \frac{1}{n})$ von Normalverteilungen mit degenerierender Varianz schwach gegen das Diracmaß δ_0, denn mit dem Satz von Lebesgue folgt für $f \in C_b(\mathbb{R})$

$$\begin{aligned}
\int f\, d\mu_n &= \int f(x)\, \frac{1}{\sqrt{2\pi/n}}\, e^{-\frac{x^2}{2/n}}\, dx \\
&\overset{y=\sqrt{n}x}{=} \int f\left(\frac{y}{\sqrt{n}}\right) \frac{1}{\sqrt{2\pi}}\, e^{-\frac{y^2}{2}}\, dy \\
&\overset{Lebesgue}{\longrightarrow} f(0) \cdot \underbrace{\int \frac{1}{\sqrt{2\pi}}\, e^{-\frac{y^2}{2}}\, dy}_{=1} \\
&= \int f\, d\delta_0.
\end{aligned}$$

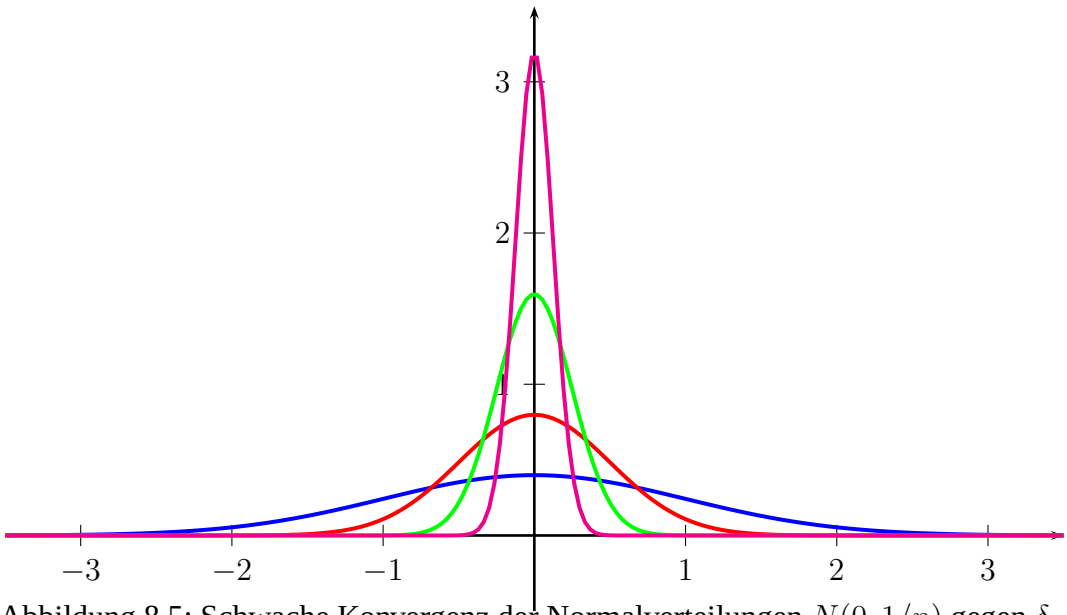


Abbildung 8.5: Schwache Konvergenz der Normalverteilungen $N(0, 1/n)$ gegen δ_0.

(3). ***Schwache vs. vage Konvergenz:*** Die Folge $\mu_n = N(0, n)$ konvergiert vage gegen das Nullmaß μ mit $\mu[A] = 0$ für alle A. In der Tat gilt für $f \in C(\mathbb{R})$ mit $f(x) = 0$ für $x \notin [-K, K]$:

$$\left|\int f\, d\mu_n\right| = \left|\int_{-K}^{K} f(x) \cdot \frac{1}{\sqrt{2\pi n}} e^{-x^2/2n} dx\right| \leq \frac{2K}{\sqrt{2\pi n}} \cdot \sup|f| \overset{n\to\infty}{\longrightarrow} 0.$$

Es gilt aber keine schwache Konvergenz, da

$$\int 1\, d\mu_n = \mu_n[\mathbb{R}] = 1 \not\to 0.$$

Die Masse wandert in diesem Fall ins Unendliche ab.

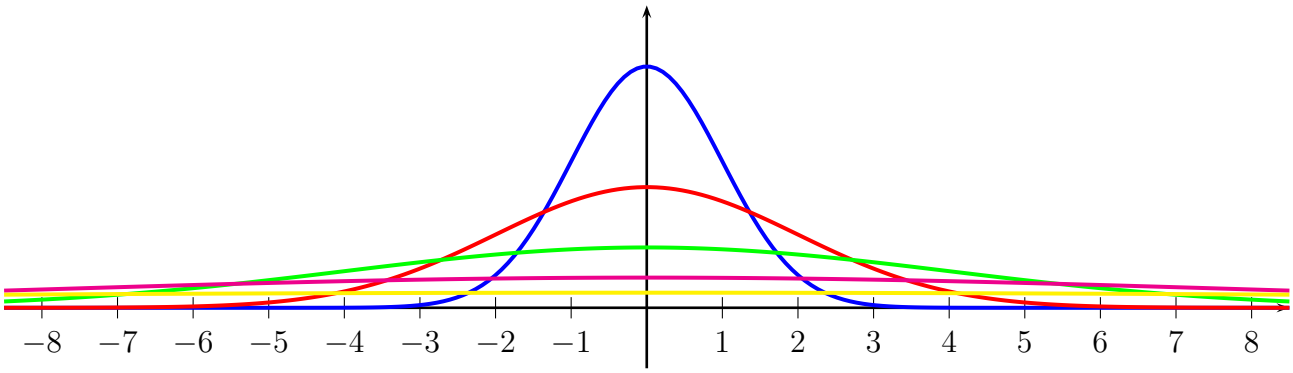


Abbildung 8.6: Konvergenz der Dichten der Normalverteilungen $N(0,n)$ gegen die Nullfunktion.

(4). ***Wartezeiten:*** Die Wartezeit T_p auf den ersten Erfolg bei unabhängigen Ereignissen mit Erfolgswahrscheinlichkeit $p \in (0,1)$ ist geometrisch verteilt:

$$P[T_p > k] \quad = \quad (1-p)^k \qquad \text{für alle } k \in \mathbb{N}.$$

Sei nun eine Intensität $\lambda > 0$ gegeben. Um kontinuierliche Wartezeiten zu approximieren, betrachten wir unabhängige Ereignisse, die zu den Zeitpunkten i/n, $n \in \mathbb{N}$, mit Wahrscheinlichkeit λ/n stattfinden. Dann ist $\frac{1}{n}T_{\lambda/n}$ die Wartezeit bis zum ersten Eintreten eines Ereignisses. Für $n \to \infty$ gilt:

$$P\left[\frac{1}{n}\, T_{\frac{\lambda}{n}} > x\right] \;=\; P\left[T_{\frac{\lambda}{n}} > nx\right] \;=\; \left(1-\frac{\lambda}{n}\right)^{\lfloor nx \rfloor} \quad \overset{n\nearrow\infty}{\longrightarrow} \quad e^{-\lambda x} \qquad \forall x \geq 0.$$

Also konvergiert die Verteilung von $\frac{1}{n}T_{\lambda/n}$ schwach gegen die Exponentialverteilung mit Parameter λ. Konvergenz in Variationsdistanz gilt nicht, da die approximierenden Verteilungen diskret, und die Grenzverteilungen stetig sind.

(5). ***Diskrete Approximation von Wahrscheinlichkeitsverteilungen:*** Allgemeiner können wir eine gegebene Wahrscheinlichkeitsverteilung auf verschiedene Arten durch diskrete Wahrscheinlichkeitsverteilungen, also Konvexkombinationen von Diracmaßen approximieren:

(a) **Klassische numerische Approximation:** Sei μ eine absolutstetige Wahrscheinlichkeitsverteilung auf $[0,1]$ mit Dichtefunktion proportional zu $g(x)$, und sei

$$\mu_n \quad := \quad \sum_{i=1}^{n} w_n^{(i)} \delta_{\frac{i}{n}},$$

mit

$$w_n^{(i)} \quad = \quad \frac{g(\frac{i}{n})}{\sum_{j=1}^{n} g(\frac{j}{n})}.$$

Dann konvergiert μ_n schwach gegen μ, denn

$$\int f \, d\mu_n \quad = \quad \sum_{i=1}^{n} w_n^{(i)} \, f\left(\frac{i}{n}\right) \quad = \quad \frac{\frac{1}{n} \sum_{i=1}^{n} f\left(\frac{i}{n}\right) \, g\left(\frac{i}{n}\right)}{\frac{1}{n} \sum_{i=1}^{n} g\left(\frac{i}{n}\right)}$$

$$\stackrel{n \nearrow \infty}{\longrightarrow} \quad \frac{\int_0^1 fg \, dx}{\int_0^1 g \, dx} \quad = \quad \int f \, d\mu \qquad \forall \, f \in C([0,1]).$$

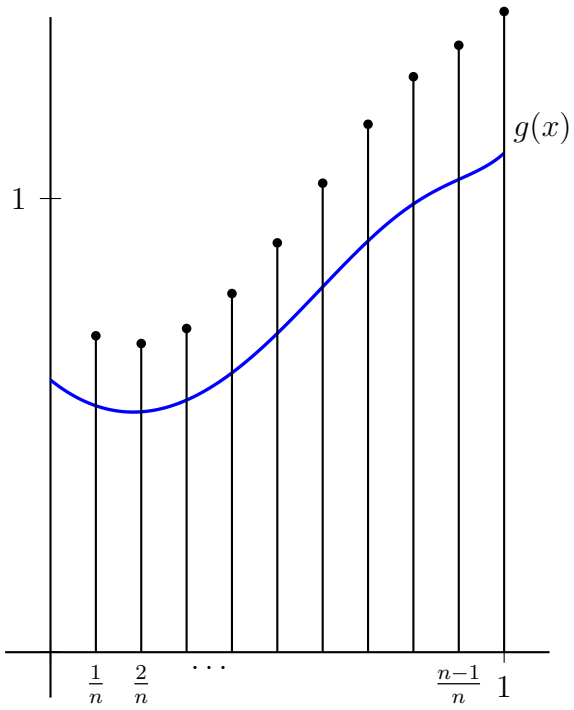


Abbildung 8.7: Stützstellen und Gewichte einer deterministischen Approximation von μ.

Die Stützstellen i/n und die Gewichte $w_n^{(i)}$ können natürlich auch auf andere Art gewählt werden, z.B. kann die hier verwendete naive Approximation des Integrals durch eine andere deterministische Quadraturformel ersetzt werden.

(b) **Monte-Carlo-Approximation**: Sei $(S, \mathcal{S}, \mu)$ ein beliebiger Wahrscheinlichkeitsraum. Sind $X_1, X_2, ... : \Omega \to S$ unabhängige Zufallsvariablen auf $(\Omega, \mathcal{A}, P)$ mit Verteilung μ, dann konvergieren die **empirischen Verteilungen**

$$\widehat{\mu}_n(\omega, \bullet) \quad := \quad \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i(\omega)}$$

P-f.s. schwach gegen μ, denn für $f \in C_b(S)$ gilt nach dem starken Gesetz großer Zahlen für P-fast alle ω

$$\int f \, d\widehat{\mu}_n(\omega, \bullet) \quad = \quad \frac{1}{n} \sum_{i=1}^{n} \underbrace{f(X_i(\omega))}_{\substack{\text{iid,} \\ \text{beschränkt}}} \quad \overset{GdgZ}{\longrightarrow} \quad E[f(X_1)] \quad = \quad \int f \, d\mu.$$

Konvergenz der Verteilungen von Zufallsvariablen

Im Gegensatz zu anderen Konvergenzbegriffen für eine Folge $(Y_n)_{n\in\mathbb{N}}$ von Zufallsvariablen bezieht sich die Verteilungskonvergenz nur auf die Verteilungen der Y_n. Insbesondere können die Zufallsvariablen Y_n und der Grenzwert Y alle auf unterschiedlichen Wahrscheinlichkeitsräumen definiert sein. Wir untersuchen nun den Zusammenhang der schwachen Konvergenz der Verteilungen mit anderen Konvergenzarten in dem Fall, dass Y_n $(n \in \mathbb{N})$ und Y *reellwertige* Zufallsvariablen sind, die auf einem *gemeinsamen Wahrscheinlichkeitsraum* $(\Omega, \mathcal{A}, P)$ definiert sind.

Satz 8.4. *Konvergiert Y_n P-fast sicher oder P-stochastisch gegen Y, dann konvergiert Y_n auch in Verteilung gegen Y.*

Beweis. Sei $f \in C_b(\mathbb{R})$. Konvergiert Y_n fast sicher gegen Y, dann konvergiert auch $f(Y_n)$ fast sicher gegen $f(Y)$. Nach dem Satz von Lebesgue folgt

$$E[f(Y_n)] \quad \longrightarrow \quad E[f(Y)].$$

Konvergiert Y_n nur stochastisch gegen Y, dann hat jede Teilfolge $(Y_{n_k})_{k\in\mathbb{N}}$ von $(Y_n)_{n\in\mathbb{N}}$ eine fast sicher gegen Y konvergente Teilfolge $(Y_{n_{k_l}})_{l\in\mathbb{N}}$. Wie zuvor folgt

$$E[f(Y_{n_{k_l}})] \quad \longrightarrow \quad E[f(Y)].$$

Also hat jede Teilfolge der Folge $(E[f(Y_n)])_{n\in\mathbb{N}}$ der Erwartungswerte eine gegen $E[f(Y)]$ konvergente Teilfolge, d.h. es gilt erneut

$$E[f(Y_n)] \quad \longrightarrow \quad E[f(Y)].$$

□

Wir beweisen nun eine partielle Umkehrung der Aussage aus Satz 8.4:

Satz 8.5 (Skorokhod - Darstellung und Charakterisierung der schwachen Konvergenz).

Seien μ_n, μ Wahrscheinlichkeitsverteilungen auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ mit Verteilungsfunktionen F_n bzw. F. Dann sind äquivalent:

(1). Die Folge $(\mu_n)_{n\in\mathbb{N}}$ konvergiert schwach gegen μ.

(2). $F_n(c) \to F(c)$ für alle Stetigkeitsstellen c von F.

(3). Es existieren Zufallsvariablen G_n, G auf

$$(\Omega, \mathcal{A}, P) = ((0,1), \mathcal{B}((0,1)), \mathcal{U}_{(0,1)})$$

mit Verteilungen μ_n bzw. μ, sodass $G_n \to G$ P-fast sicher.

Beweis. „(3) $\Rightarrow$ (1)“ folgt aus Satz 8.4.

„(1) $\Rightarrow$ (2)“: Für $c \in \mathbb{R}$ gilt:

$$F_n(c) = \int I_{(-\infty,c]}\, d\mu_n \quad \text{und} \quad F(c) = \int I_{(-\infty,c]}\, d\mu. \tag{8.3.1}$$

Sei $\varepsilon > 0$. Wir definieren stetige Approximationen der Indikatorfunktion $I_{(-\infty,c]}$ durch

$$f_\varepsilon(x) = \begin{cases} 1 & \text{für } x \le c-\varepsilon \\ 0 & \text{für } x \ge c \\ \frac{c-x}{\varepsilon} & \text{für } x \in [(c-\varepsilon, c) \end{cases}, \text{ und } \quad g_\varepsilon(x) = \begin{cases} 1 & \text{für } x \le c \\ 0 & \text{für } x \ge c+\varepsilon \\ \frac{c+\varepsilon-x}{\varepsilon} & \text{für } x \in (c, c+\varepsilon) \end{cases}.$$

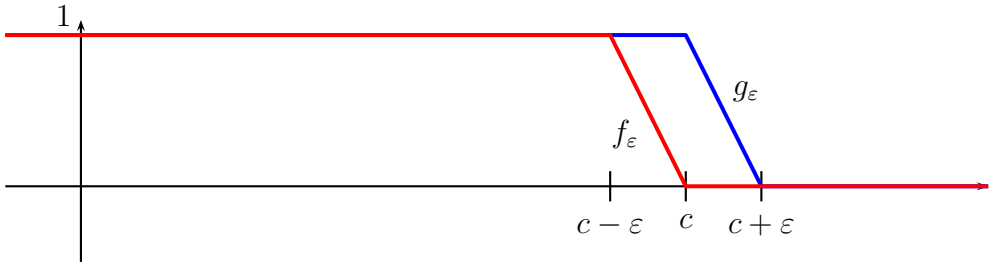


Abbildung 8.8: Stetige Approximationen von $I_{(-\infty,c]}$.

Es gilt

$$I_{(-\infty,c-\varepsilon]} \le f_\varepsilon \le I_{(-\infty,c]} \le g_\varepsilon \le I_{(-\infty,c+\varepsilon]}. \tag{8.3.2}$$

Konvergiert μ_n schwach gegen μ, dann folgt nach (8.3.1) und (8.3.2):

$$\liminf F_n(c) \ge \liminf \int f_\varepsilon\, d\mu_n = \int f_\varepsilon\, d\mu \ge F(c-\varepsilon), \quad \text{und}$$
$$\limsup F_n(c) \le \limsup \int g_\varepsilon\, d\mu_n = \int g_\varepsilon\, d\mu \le F(c+\varepsilon).$$

Für $\varepsilon \searrow 0$ erhalten wir

$$\limsup F_n(c) \le F(c) = \lim_{\varepsilon \searrow 0} F(c+\varepsilon),$$

und

$$\liminf F_n(c) \quad \geq \quad F(c) \quad = \quad \lim_{\varepsilon \searrow 0} F(c-\varepsilon),$$

falls F bei c stetig ist.

„(2) $\Rightarrow$ (3)“: Für $u \in (0,1)$ betrachten wir die minimalen und maximalen u-Quantile

$$\underline{G}(u) := \inf\{x \in \mathbb{R} \mid F(x) \geq u\}, \qquad \text{und} \qquad \overline{G}(u) := \inf\{x \in \mathbb{R} \mid F(x) > u\}$$

der Verteilung μ, siehe Abschnitt 4.4. Entsprechend seien $\underline{G}_n$ und $\overline{G}_n$ die minimalen und maximalen u-Quantile der Verteilung μ_n. Analog zum Beweis von Satz 4.20 zeigt man, dass $\underline{G}$ und $\overline{G}$ bzw. $\underline{G}_n$ und $\overline{G}_n$ unter der Gleichverteilung $P = \mathcal{U}_{(0,1)}$ Zufallsvariablen mit Verteilung μ bzw. μ_n sind. Wir zeigen nun, dass aus (2) folgt:

Behauptung: $\underline{G}_n \to \underline{G}$ P-fast sicher und $\overline{G}_n \to \overline{G}$ P-fast sicher.

Damit ist dann die Implikation „(2) $\Rightarrow$ (3)“ bewiesen. Den Beweis der Behauptung führen wir in mehreren Schritten durch:

(a) Offensichtlich gilt $\underline{G} \leq \overline{G}$, und $\underline{G}_n \leq \overline{G}_n$ für alle $n \in \mathbb{N}$.

(b) $\underline{G} = \overline{G}$ und $\underline{G}_n = \overline{G}_n$ P-fast sicher, denn:

$$P[\underline{G} \neq \overline{G}] \quad = \quad P[\underline{G} < \overline{G}] \quad = \quad P\left[\bigcup_{c \in \mathbb{Q}} \{\underline{G} \leq c < \overline{G}\}\right]$$

$$\leq \quad \sum_{c \in \mathbb{Q}} P[\{\underline{G} \leq c\} \setminus \{\overline{G} \leq c\}] \quad = \quad \sum_{c \in \mathbb{Q}} (\underbrace{P[\{\underline{G} \leq c\}]}_{=F(c)} - \underbrace{P[\{\overline{G} \leq c\}]}_{=F(c)}) \quad = \quad 0\,.$$

(c) Wir zeigen nun:

$$\limsup \overline{G}_n(u) \quad \leq \quad \overline{G}(u), \qquad \text{und} \qquad \liminf\ \underline{G}_n(u) \quad \geq \quad \underline{G}(u). \tag{8.3.3}$$

Zum Beweis der ersten aussage genügt es zu zeigen, dass

$$\limsup \overline{G}_n(u) \quad \leq \quad c \qquad \text{für alle } c > \overline{G}(u) \qquad \text{mit } \mu[\{c\}] = 0 \tag{8.3.4}$$

gilt, denn es existieren höchstens abzählbar viele c mit $\mu[\{c\}] \neq 0$. Für $c > \overline{G}(u)$ mit $\mu[\{c\}] = 0$ gilt aber nach Definition von $\overline{G}$ und nach (2):

$$u \quad < \quad F(c) \quad = \quad \lim_{n \to \infty} F_n(c),$$

also existiert ein $n_0 \in \mathbb{N}$ mit

$$F_n(c) \quad > \quad u \qquad \text{für alle } n > n_0. \tag{8.3.5}$$

Aus (8.3.5) folgt

$$\overline{G}_n(u) \le c \qquad \text{für } n \ge n_0,$$

und somit

$$\limsup \overline{G}_n(u) \quad \le \quad c.$$

Damit haben wir die erste Aussage in (8.3.3) bewiesen. Die zweite Aussage zeigt man auf ähnliche Weise.

(d) Aus (a)-(c) folgt P-fast sicher:

$$\limsup \underline{G}_n \overset{(a)}{\le} \limsup \overline{G}_n \overset{(c)}{\le} \overline{G} \overset{(b)}{=} \underline{G} \overset{(3)}{\le} \liminf \overline{G}_n \overset{(a)}{\le} \liminf \overline{G}_n,$$

also

$$\lim \underline{G}_n \;=\; \underline{G} \quad \text{und} \quad \lim \overline{G}_n \;=\; \overline{G}.$$

□

Ein wesentlicher Schritt, um den oben skizzierten Beweis des Zentralen Grenzwertsatzes zu vervollständigen, ist es, zu zeigen, dass die Verteilungen der standardisierten Summen von unabhängigen, identisch verteilten, quadratintegrierbaren Zufallsvariablen eine schwach konvergente Teilfolge haben:

Existenz schwach konvergenter Teilfolgen

Eine Folge von Wahrscheinlichkeitsverteilungen auf einer *endlichen* Menge $S = \{x_1, \ldots, x_d\}$ können wir als beschränkte Folge in $\mathbb{R}^d$ auffassen. Daher existiert stets eine konvergente Teilfolge – der Grenzwert ist wieder eine Wahrscheinlichkeitsverteilung auf S. Für unendliche Mengen S gilt eine entsprechende Aussage im Allgemeinen nicht. Wir beweisen nun ein Kriterium für die Existenz schwach konvergenter Teilfolgen für Folgen von Wahrscheinlichkeitsverteilungen auf $\mathbb{R}^1$. Dazu setzen wir voraus, dass die Masse nicht ins unendliche abwandert:

Definition. *Eine Folge $\mu_n \in WV(\mathbb{R})$ heißt **straff** (engl. tight), falls zu jedem $\varepsilon > 0$ ein $c \in (0, \infty)$ existiert mit*

$$\mu_n([-c, c]) \;\ge\; 1 - \varepsilon \qquad \textit{für alle } \; n \in \mathbb{N}.$$

Eine straffe Folge von Wahrscheinlichkeitsverteilungen ist also gleichmäßig auf Kompakta konzentriert. Die Masse kann daher für $n \to \infty$ nicht ins Unendliche abwandern.

Beispiel. Die Folge $\mu_n = N(m_n, \sigma_n^2), m_n \in \mathbb{R},\ \sigma_n > 0$, ist genau dann straff, wenn die Folgen m_n und σ_n der Mittelwerte und Standardabweichungen beschränkt sind.

Satz 8.6 (Helly-Bray). *Jede straffe Folge $\mu_n \in WV(\mathbb{R})$ hat eine schwach konvergente Teilfolge.*

Bemerkung. (1). Das Kriterium lässt sich deutlich verallgemeinern: Eine entsprechende Aussage gilt für Folgen von Wahrscheinlichkeitsverteilungen auf beliebigen vollständigen separablen metrischen Räumen (Satz von Prohorov, siehe z.B. *Billingsley: Convergence of probability measures*). Die endlichen Intervalle $[-c, c]$ in der Definition von Straffheit ersetzt man in diesem Fall durch kompakte Mengen.

(2). Der Raum $WV(\overline{\mathbb{R}})$ aller Wahrscheinlichkeitsverteilungen auf $[-\infty, \infty]$ ist sogar **kompakt** bezüglich der schwachen Topologie, d.h. **jede** Folge $\mu_n \in WV(\overline{\mathbb{R}})$ hat eine schwach konvergente Teilfolge. Der Beweis verläuft analog zu dem von Satz 8.6. Es folgt, dass jede Folge $\mu_n \in WV(\mathbb{R})$ eine vag konvergente Teilfolge hat. Der Limes ist jedoch i.A. kein Wahrscheinlichkeitsmaß auf $\mathbb{R}$, da die Masse ins unendliche abwandern kann. Allgemeiner gilt: Ist S kompakt, dann ist $WV(S)$ kompakt bzgl. der schwachen Konvergenz.

Wir beweisen nun den Satz von Helly-Bray:

Beweis. Sei μ_n $(n \in \mathbb{N})$ eine straffe Folge von Wahrscheinlichkeitsverteilungen auf $\mathbb{R}$. Um die Existenz einer schwach konvergenten Teilfolge zu zeigen, betrachten wir die Folge der Verteilungsfunktionen F_n. Wir zeigen die Aussage in mehreren Schritten:

(1). *Es existiert eine Teilfolge $(F_{n_k})_{k\in\mathbb{N}}$, sodass $F_{n_k}(x)$ für alle $x \in \mathbb{Q}$ konvergiert:*

Zum Beweis verwenden wir ein Diagonalverfahren: Sei $x_1, x_2, ...$ eine Abzählung von $\mathbb{Q}$. Wegen $0 \le F_n \le 1$ existiert eine Teilfolge $(F_{n_k^{(1)}})_{k\in\mathbb{N}}$, für die $F_{n_k^{(1)}}(x_1)$ konvergiert. Ebenso existiert eine Teilfolge $(F_{n_k^{(2)}})_{k\in\mathbb{N}}$ von $(F_{n_k^{(1)}})_{k\in\mathbb{N}}$, für die $F_{n_k^{(2)}}(x_2)$ konvergiert, usw. Die Diagonalfolge $F_{n_k}(x) := F_{n_k^{(k)}}(x)$ konvergiert dann für alle $x \in \mathbb{Q}$.
Für $x \in \mathbb{Q}$ setzen wir $\overline{F}(x) := \lim_{k\to\infty} F_{n_k}(x)$. Nach (1) existiert der Grenzwert, außerdem ist die Funktion $\overline{F} : \mathbb{Q} \to [0, 1]$Der Limes existiert nach 1. für $x \in \mathbb{Q}$ und die Funktion $\overline{F} : \mathbb{Q} \to [0, 1]$ monoton wachsend, da die Funktionen F_{n_k} monoton wachsend sind.

(2). *Stetige Fortsetzung von $\overline{F}$ auf $[0, 1]$:* Für $x \in \mathbb{R}$ setzen wir

$$F(x) := \inf\{\overline{F}(y) \mid y \in \mathbb{Q}, y > x\}.$$

Die folgenden Eigenschaften der Funktion F prüft man leicht nach:

(a) Die Funktion F ist rechtsstetig, monoton wachsend, und es gilt $0 \le F \le 1$.

(b) $F_{n_k}(x) \to F(x)$ für alle $x \in \mathbb{R}$, an denen F stetig ist.

(3). Aus (a)folgt, dass durch

$$\mu[(a,b]] := F(b) - F(a), \qquad -\infty < a \leq b < \infty,$$

ein positives Maß auf $\mathbb{R}$ definiert wird mit

$$\mu[\mathbb{R}] = \lim_{c\to\infty} \mu[(-c,c]] \ \in [0,1].$$

Wir zeigen nun, dass μ eine *Wahrscheinlichkeitsverteilung* auf $\mathbb{R}$ ist, falls die Folge $(\mu_n)_{n\in\mathbb{N}}$ *straff* ist. Es gilt nämlich:

$$\mu[(-c,c]] \ = \ F(c) - F(-c) \ = \ \lim_{k\to\infty} (F_{n_k}(c) - F_{n_k}(-c)) \ = \ \lim_{k\to\infty} \mu_{n_k}[(-c,c]] \quad (8.3.6)$$

für fast alle c. Aus der Straffheit von $(\mu_n)_{n\in\mathbb{N}}$ folgt, dass zu jedem $\varepsilon > 0$ ein $c(\varepsilon) \in \mathbb{R}$ existiert mit

$$\mu_{n_k}[(-c,c]] \geq 1 - \varepsilon \qquad \text{für alle } k.$$

Aus (8.3.6) folgt dann $\mu[(-c,c]] \geq 1-\varepsilon$, falls c groß genug ist, und damit für $\varepsilon \searrow 0$:

$$\mu[\mathbb{R}] \geq 1, \qquad \text{also} \quad \mu(\mathbb{R}) = 1.$$

(4). Aus (b) folgt nun nach Satz 8.5, dass die Folge $(\mu_{n_k})_{k\in\mathbb{N}}$ schwach gegen μ konvergiert.

□

Schwache Konvergenz über charakteristische Funktionen

Unter Verwendung der Existenz schwach konvergenter Teilfolgen einer straffen Folge von Wahrscheinlichkeitsverteilungen zeigen wir nun, dass eine Folge von Wahrscheinlichkeitsverteilungen auf $\mathbb{R}$ genau dann schwach konvergiert, wenn die charakteristischen Funktionen gegen eine Grenzfunktion konvergieren, die bei 0 stetig ist:

Satz 8.7 (Stetigkeitssatz, Konvergenzsatz von Lévy). *Seien $(\mu_n)_{n\in\mathbb{N}}$ Wahrscheinlichkeitsverteilungen auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ mit charakteristischen Funktionen*

$$\phi_n(t) \quad = \quad \int e^{itx}\, \mu_n(dx).$$

Dann gilt:

(1). Konvergiert μ_n schwach gegen eine Wahrscheinlichkeitsverteilung μ, dann konvergieren auch die charakteristischen Funktionen:

$$\phi_n(t) \quad \to \quad \phi(t) \quad := \quad \int e^{itx}\, \mu(dx) \qquad \textit{für alle } \ t \in \mathbb{R}.$$

(2). Konvergiert umgekehrt $\phi_n(t)$ für alle $t \in \mathbb{R}$ gegen einen Limes $\phi(t)$, und ist ϕ stetig bei $t = 0$, dann ist ϕ die charakteristische Funktion einer Wahrscheinlichkeitsverteilung μ, und μ_n konvergiert schwach gegen μ.

Bemerkung. (1). Die Stetigkeit von ϕ bei 0 ist wesentlich. Zum Beispiel ist die Folge $\mu_n = N(0, n)$ nicht schwach konvergent, aber die charakteristischen Funktionen konvergieren punktweise:

$$\phi_n(t) = e^{-\frac{t^2}{2n}} \overset{n\uparrow\infty}{\to} \begin{cases} 0 & \text{falls } t \neq 0 \\ 1 & \text{falls } t = 0 \end{cases}.$$

(2). Eine Aussage wie im Satz gilt auch für Wahrscheinlichkeitsverteilungen auf $\mathbb{R}^d$. Hier definiert man die charakteristische Funktion $\phi : \mathbb{R}^d \to \mathbb{C}$ durch

$$\phi(t) \quad = \quad \int_{\mathbb{R}^d} e^{it\cdot x}\, \mu(dx), \qquad t \in \mathbb{R}^d.$$

Beweis. Der erste Teil der Aussage folgt unmittelbar aus $e^{itx} \quad = \quad \cos(tx) + i\sin(tx)$, denn Kosinus und Sinus sind beschränkte stetige Funktionen.
Der Beweis des zweiten Teils der Aussage erfolgt nun in mehreren Schritten. Wir nehmen an, dass die charakteristischen Funktionen $\phi_n(t)$ punktweise gegen eine bei 0 stetige Grenzfunktion $\phi(t)$ konvergieren.

(1). *Relative Kompaktheit: Jede Teilfolge von $(\mu_n)_{n\in\mathbb{N}}$ hat eine schwach konvergente Teilfolge.*

Dies ist der zentrale Schritt im Beweis. Nach dem Satz von Helly-Bray genügt es zu zeigen, dass μ_n $(n \in \mathbb{N})$ unter den Voraussetzungen straff ist. Dazu schätzen wir die Wahrscheinlichkeiten $\mu_n[|x| \geq c]$ mithilfe der charakteristischen Funktionen ab. Da die Funktion $f(u) = 1 - \frac{\sin u}{u}$ für $u \neq 0$ strikt positiv ist mit $\lim_{|u|\to\infty} f(u) = 1$, existiert eine Konstante $a > 0$ mit $f(u) \geq a$ für alle $|u| \geq 1$. Damit erhalten wir für $\varepsilon > 0$:

$$\begin{aligned} & \mu_n\left[|x| \geq \frac{1}{\varepsilon}\right] \\ = \quad & \mu_n\left[\{x \in \mathbb{R} \mid |\varepsilon x| \geq 1\}\right] \quad \leq \quad \frac{1}{a}\int \underbrace{\left(1 - \frac{\sin \varepsilon x}{\varepsilon x}\right)}_{=\frac{1}{2\varepsilon}\int_{-\varepsilon}^{\varepsilon}(1-\cos(xt))dt} \mu_n(dx) \end{aligned} \tag{8.3.7}$$

$$\overset{Fubini}{=} \quad \frac{1}{2a\varepsilon}\int_{-\varepsilon}^{\varepsilon}(1 - \mathrm{Re}(\phi_n(t)))dt \overset{\substack{n\nearrow\infty \\ \longrightarrow \\ Lebesgue}}{} \frac{1}{2a\varepsilon}\cdot\int_{-\varepsilon}^{\varepsilon}(1 - \mathrm{Re}(\phi(t)))dt.$$

Sei nun $\delta > 0$ vorgegeben. Ist ε hinreichend klein, dann gilt wegen der vorausgesetzten Stetigkeit von ϕ bei 0:

$$|1 - \mathrm{Re}(\phi(t))| \quad = \quad |\mathrm{Re}(\phi(0) - \phi(t))| \quad \leq \quad \frac{\delta a}{2} \qquad \text{für alle } t \in [-\varepsilon, \varepsilon].$$

Also können wir die rechte Seite von (8.3.7) durch $\delta/2$ abschätzen, und somit existiert ein $n_0 \in \mathbb{N}$ mit

$$\mu_n \left[|x| \geq \frac{1}{\varepsilon} \right] \quad \leq \quad \delta \qquad \text{für alle } n \geq n_0. \tag{8.3.8}$$

Diese Aussage gilt natürlich auch, falls wir ε noch kleiner wählen. Zudem gilt (8.3.8) auch für alle $n < n_0$, falls ε klein genug ist. Also ist μ_n $(n \in \mathbb{N})$ straff.

(2). *Der Grenzwert **jeder** schwach konvergenten Teilfolge von $(\mu_n)_{n\in\mathbb{N}}$ hat die charakteristische Funktion ϕ.*

Zum Beweis sei $(\mu_{n_k})_{k\in\mathbb{N}}$ eine Teilfolge von $(\mu_n)_{n\in\mathbb{N}}$ und μ eine Wahrscheinlichkeitsverteilung mit $\mu_{n_k} \xrightarrow{w} \mu$. Dann gilt nach dem ersten Teil der Aussage des Satzes:

$$\phi_\mu(t) \quad = \quad \lim_{k\to\infty} \phi_{n_k}(t) \quad = \quad \phi(t) \qquad \text{für alle } \ t \in \mathbb{R}.$$

(3). *Schwache Konvergenz von $(\phi_n)_{n\in\mathbb{N}}$.*

Nach dem Inversionssatz existiert höchstens eine Wahrscheinlichkeitsverteilung μ mit charakteristischer Funktion ϕ. Also konvergieren nach (2) alle schwach konvergenten Teilfolgen von $(\mu_n)_{n\in\mathbb{N}}$ gegen denselben Limes μ. Hieraus folgt aber, zusammen mit (1), dass $(\mu_n)_{n\in\mathbb{N}}$ schwach gegen μ konvergiert, denn für $f \in \mathcal{C}_b(S)$ hat jede Teilfolge von $\int f \, d\mu_n$ eine gegen $\int f \, d\mu$ konvergente Teilfolge, und somit gilt $\int f \, d\mu_n \to \int f \, d\mu$.

□

8.4 Der Zentrale Grenzwertsatz

Wir können nun den in Abschnitt 8.2 skizzierten Beweis des Zentralen Grenzwertsatzes (engl. *Central Limit Theorem*) vervollständigen. Wir zeigen zunächst, dass ein zentraler Grenzwertsatz für Summen beliebiger unabhängiger, identisch verteilter Zufallsvariablen mit endlicher Varianz gilt. Diese Aussage wurde zuerst 1900 von Lyapunov bewiesen, der damit den Satz von de Moivre/Laplace (1733) deutlich verallgemeinern konnte. Am Ende dieses Abschnitts beweisen wir eine noch allgemeinere Version des Zentralen Grenzwertsatzes, die auf Lindeberg und Feller zurückgeht.

Zentraler Grenzwertsatz für Summen von i.i.d. Zufallsvariablen

Satz 8.8 (Zentraler Grenzwertsatz – 1. Version). *Seien $X_1, X_2, ... \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$ unabhängige, identisch verteilte Zufallsvariablen mit Varianz σ^2 und sei*

$$S_n \;=\; X_1 + ... + X_n \;.$$

Dann konvergieren die Verteilungen der standardisierten Summen

$$\hat{S}_n \;=\; \frac{S_n - E[S_n]}{\sqrt{n}} \;=\; \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - E[X_i])$$

schwach gegen $N(0, \sigma^2)$.

Bemerkung. (1). Alternativ kann man die standardisierten Summen auf Varianz 1 normieren, und erhält

$$\frac{S_n - E[S_n]}{\sigma \cdot \sqrt{n}} \quad \xrightarrow{\mathcal{D}} \quad Z,$$

wobei Z eine standardnormalverteilte Zufallsvariable ist.

(2). Die Voraussetzung $X_i \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$ ist wesentlich. Bei unendlicher Varianz der X_i können sich andere Grenzverteilungen für die geeignet renormierten Summen $\frac{S_n - a_n}{b_n} \quad (a_n \in \mathbb{R}, b_n > 0)$ ergeben. Als Grenzverteilungen können i.A. die sogenannten stabilen Verteilungen auftreten, siehe dazu z.B. Satz 8.12 unten.

(3). Im Fall $\sigma^2 = 0$ gilt die Aussage auch. Hierbei interpretieren wir das Diracmaß δ_m als degenerierte Normalverteilung $N(m, 0)$.

Wir beweisen nun den Zentralen Grenzwertsatz in der oben stehenden Form:

Beweis. O.B.d.A. sei $E[X_i] = 0$, ansonsten betrachten wir die zentrierten Zufallsvariablen $\widetilde{X}_i := X_i - E[X_i]$. Nach dem Konvergenzsatz von Lévy genügt es zu zeigen, dass die charakteristischen Funktionen der standardisierten Summen $\widehat{S}_n$ punktweise gegen die charakteristische Funktion der Normalverteilung $N(0, \sigma^2)$ konvergieren, d.h.

$$\phi_{\hat{S}_n}(t) \;\to\; \phi_{N(0,\sigma^2)}(t) \;=\; e^{-\frac{\sigma^2 t^2}{2}} \qquad \forall\, t \in \mathbb{R}. \tag{8.4.1}$$

Da die Zufallsvariablen X_i unabhängig, identisch verteilt und zentriert sind, gilt für $t \in \mathbb{R}$:

$$\phi_{\hat{S}_n}(t) \quad \overset{E[S_n]=0}{=} \quad \phi_{S_n}\left(\frac{t}{\sqrt{n}}\right) \quad \overset{X_i \text{ iid}}{=} \quad \left(\phi_{X_1}\left(\frac{t}{\sqrt{n}}\right)\right)^n.$$

Aus $X_1 \in \mathcal{L}^2$ folgt $\phi_{X_1} \in C^2(\mathbb{R})$, und

$$\phi_{X_1}(t) \quad = \quad E[e^{itX_1}] \quad = \quad 1 + itE[X_1] + \frac{(it)^2}{2}E[X_1^2] + o(t^2) \quad = \quad 1 - \frac{t^2\sigma^2}{2} + o(t^2),$$

wobei o für eine Funktion $o : \mathbb{R}^+ \to \mathbb{C}$ mit $\lim_{\varepsilon \downarrow 0} \frac{|o(\varepsilon)|}{\varepsilon} = 0$ steht. Damit erhalten wir:

$$\phi_{\widehat{S}_n}(t) \quad = \quad \left(1 - \frac{t^2\sigma^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n.$$

Wir vermuten, dass dieser Ausdruck für $n \to \infty$ gegen $e^{-\frac{t^2\sigma^2}{2}}$ strebt. Dies kann man beweisen, indem man den Logarithmus nimmt, und die Taylorapproximation $\log(1+w) = w+o(|w|)$ verwendet. Da die charakteristische Funktion komplexwertig ist, muss dazu allerdings der Hauptzweig der komplexen Logarithmusfunktion verwendet werden.
Wir zeigen stattdessen die Konvergenz ohne Verwendung von Aussagen aus der Funktionentheorie: Für komplexe Zahlen $z_i, w_i \in \mathbb{C}$ mit $|z_i|, |w_i| \leq 1$ gilt nach der Dreiecksungleichung

$$\begin{aligned}\left|\prod_{i=1}^n z_i - \prod_{i=1}^n w_i\right| &= |(z_1 - w_1)z_2z_3\cdots z_n + w_1(z_2 - w_2)z_3z_4\cdots z_n + \ldots + w_1\cdots w_{n-1}(z_n - w_n)| \\ &\leq \sum_{i=1}^n |z_i - w_i|.\end{aligned}$$

Damit erhalten wir:

$$\begin{aligned}\left|\phi_{\widehat{S}_n}(t) - \exp\left(-\frac{t^2\sigma^2}{2}\right)\right| &= \left|\left(1 - \frac{t^2\sigma^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n - \exp\left(-\frac{t^2\sigma^2}{2n}\right)^n\right| \\ &\leq n \cdot \left|1 - \frac{t^2\sigma^2}{2n} + o\left(\frac{t^2}{n}\right) - \exp\left(-\frac{t^2\sigma^2}{2n}\right)\right|.\end{aligned}$$

Da die rechte Seite für $n \to \infty$ gegen 0 konvergiert, folgt (8.4.1) und damit die Behauptung. □

Beispiel. (1). Sind $X_1, X_2, \ldots$ unabhängig mit $P[X_i = 1] = p$ und $P[X_i = 0] = 1 - p$, dann ist $S_n = \sum_{i=1}^n X_i$ binomialverteilt mit Parametern n und p. Die Aussage des Zentralen Grenzwertsatzes folgt in diesem Fall aus dem Satz von de Moivre/Laplace.

(2). Sind die Zufallsvariablen X_i unabhängig und Poissonverteilt mit Parameter $\lambda > 0$, dann ist $S_n = \sum_{i=1}^n X_i$ Poissonverteilt mit Parameter $n\lambda$. Der Zentrale Grenzwertsatz liefert in diesem Fall eine Normalapproximation für Poissonverteilungen mit großer Intensität (Übung).

(3). Sind $X_1, X_2, \ldots$ unabhängige, $N(m, \sigma^2)$-verteilte Zufallsvariablen, dann gilt

$$\widehat{S}_n \quad = \quad \frac{X_1 + X_2 + \ldots + X_n - nm}{\sqrt{n}} \quad \sim \quad N(0, \sigma^2)$$

für alle $n \in \mathbb{N}$ (und nicht nur asymptotisch!).

Warum tritt die Normalverteilung im Limes auf? Wie schon im letzten Beispiel bemerkt, gilt

$$X_i \sim N(0,\sigma^2) \text{ unabhängig} \quad \Rightarrow \quad \frac{X_1+\ldots+X_n}{\sqrt{n}} \sim N(0,\sigma^2).$$

Die zentrierten Normalverteilungen sind also „invariant“ unter der *Reskalierungstransformation* aus dem zentralen Grenzwertsatz. Man kann sich leicht plausibel machen, dass eine Grenzverteilung der standardisierten Summen unabhängiger quadratintegrierbarer Zufallsvariablen eine entsprechende Invarianzeigenschaft haben muss. Tatsächlich sind die zentrierten Normalverteilungen die einzigen nichtdegenerierten Wahrscheinlichkeitsverteilungen mit dieser Invarianz. Aus dem Zentralen Grenzwertsatz folgt sogar:

Korollar 8.9. *Sei μ eine Wahrscheinlichkeitsverteilung auf $\mathbb{R}$ mit $\int x^2\mu(dx) < \infty$. Gilt*

$$X, Y \sim \mu \textit{ unabhängig} \quad \Rightarrow \quad \frac{X+Y}{\sqrt{2}} \sim \mu, \tag{8.4.2}$$

dann ist μ eine zentrierte Normalverteilung.

Bemerkung. Die Aussage gilt auch ohne die Voraussetzung $\int x^2\mu(dx) < \infty$; der Beweis ist aber aufwändiger, siehe z.B. BREIMAN: PROBABILITY.

Beweis. Seien $X_1, X_2, \ldots$ unabhängige Zufallsvariablen mit Verteilung μ. Aus der Voraussetzung (8.4.2) folgt $E[X_i] = \int x\,\mu(dx) = 0$ für alle $i \in \mathbb{N}$, und durch Induktion:

$$\frac{(X_1+\ldots+X_n)}{\sqrt{n}} \quad \sim \quad \mu \qquad \text{für } n = 2^k, k \in \mathbb{N}.$$

Wegen $\int x^2\mu(dx) < \infty$ sind die X_i quadratintegrierbar. Durch Anwenden des zentralen Grenzwertsatzes auf die standardisierten Summen folgt, dass μ eine zentrierte Normalverteilung ist. □

Normalapproximationen

Die Normalverteilungsasymptotik der standardisierten Summen wird häufig verwendet, um Wahrscheinlichkeiten näherungsweise zu berechnen. Wir betrachten zunächst ein typisches Beispiel:

Beispiel (Versicherungsgesellschaft mit n Verträgen). Eine Versicherungsgesellschaft habe mit n Kunden Verträge abgeschlossen. Beim Eintreten des Schadenfalls für Vertrag i muss die Leistung $X_i \geq 0$ gezahlt werden. Wir nehmen an, dass gilt:

$$X_i \in \mathcal{L}^2 \text{ i.i.d.} \quad \text{mit} \quad E[X_i] = m,\ \mathrm{Var}[X_i] = \sigma^2.$$

Die Prämie pro Vertrag betrage $\Pi = m + \lambda\sigma^2$, wobei m die erwartete Leistung ist und $\lambda\sigma^2$ mit $\lambda > 0$ einem Risikozuschlag entspricht. Die Einnahmen nach einer Zeitperiode betragen dann $n \cdot \Pi$, die Ausgaben $S_n = X_1 + ... + X_n$. Wir wollen die Wahrscheinlichkeit des Ruinereignisses

$$S_n > k + n\Pi,$$

berechnen, wobei k das Anfangskapital bezeichnet. Hierbei nehmen wir implizit an, dass nicht verzinst wird, und die Abrechnung nur am Schluß einer Zeitperiode erfolgt. Wenn die standardisierten Schadenssummen mithilfe einer ZGS-Näherung approximiert werden, ergibt sich:

$$\begin{aligned} P[\text{Ruin}] &= P[S_n > k + n\Pi] = P[S_n - E[S_n] > k + n\lambda\sigma^2] \\ &= P\left[\frac{S_n - E[S_n]}{\sigma\sqrt{n}} > \frac{k}{\sigma\sqrt{n}} + \lambda\sigma\sqrt{n}\right] \\ &\approx P\left[Z > \frac{k}{\sigma\sqrt{n}} + \lambda\sigma\sqrt{n}\right], \end{aligned}$$

wobei Z eine standardnormalverteilte Zufallsvariable ist. Der Ausdruck auf der rechten Seite geht für $n \to \infty$ gegen 0. Eine große Anzahl von Verträgen sollte also eine kleine Ruinwahrscheinlichkeit implizieren. Für $n = 2000$, $\sigma = 60$ und $\lambda = 0,05\%$ ergibt sich beispielsweise:

$$\begin{aligned} k = 0 &: P[\text{Ruin}] \approx 9\%, \\ k = 1500 &: P[\text{Ruin}] \approx 3\%. \end{aligned}$$

Nach einer solchen Überschlagsrechnung sollte man das verwendete Modell und die Approximationsschritte einer kritischen Analyse unterziehen. In unserem Fall stellen sich unmittelbar mehrere Fragen:

(1). Wir haben die ZGS-Näherung verwendet, obwohl die auftretenden Schranken für die standardisierten Summen von n abhängen. Ist das in diesem Fall zulässig?

(2). Ist die Quadratintegrierbarkeit der X_i eine sinnvolle Modellannahme, und was ergibt sich andernfalls?

(3). In einem realistischen Modell kann man nicht davon ausgehen, dass die X_i identisch verteilt sind. Gilt trotzdem ein Zentraler Grenzwertsatz?

(4). Ist die Unabhängigkeitsannahme gerechtfertigt?

Wir werden nun auf die ersten drei Fragen näher eingehen. Das folgende Beispiel zeigt, dass man in der Tat vorsichtig sein sollte, wenn man von n abhängige Quantile von standardisierten Summen durch entsprechende Quantile von Normalverteilungen ersetzt:

Beispiel (Eine zu naive ZGS-Approximation). Seien $X_i, i \in \mathbb{N}$, unabhängige, identisch verteilte Zufallsvariablen mit $E[X_i] = 0$ und $\mathrm{Var}[X_i] = 1$, und sei $a > 0$. Mit einer ZGS-Approximation erhalten wir für große n:

$$\begin{aligned} P\left[\frac{1}{n}\sum_{i=1}^{n} X_i \geq a\right] &= P\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i \geq a\sqrt{n}\right] \\ &\approx \frac{1}{\sqrt{2\pi}}\int_{a\sqrt{n}}^{\infty} e^{-\frac{x^2}{2}}\,dx \\ &= e^{-\frac{na^2}{2}}\cdot\frac{1}{\sqrt{2\pi}}\int_0^{\infty} e^{-a\sqrt{n}y-\frac{y^2}{2}}\,dy \qquad \left(x = a\sqrt{n}+y\right) \\ &= e^{-\frac{na^2}{2}}\cdot\frac{1}{\sqrt{2\pi n}}\int_0^{\infty} e^{-az-\frac{z^2}{2n}}\,dz \qquad \left(z = \sqrt{n}y\right) \\ &\sim \frac{1}{\sqrt{2\pi a^2 n}}\cdot\exp\left(-\frac{na^2}{2}\right) \end{aligned}$$

Dies ist aber **nicht** die korrekte Asymptotik für $n \to \infty$. Auf der exponentiellen Skala gilt nämlich

$$P\left[\frac{1}{n}\sum_{i=1}^{n} X_i \geq a\right] \sim \exp\left(-nI(a)\right),$$

wobei $I(a)$ die Ratenfunktion aus dem Satz von Chernoff ist. Diese ist im Allgemeinen von $na^2/2$ verschieden. Die ZGS-Approximation ist hier nicht anwendbar, da $a\sqrt{n}$ von n abhängt!

Dass die Näherung aus dem Beispiel oben trotzdem recht gut funktioniert, wenn die Zufallsvariablen X_i dritte Momente haben, garantiert die folgende *Abschätzung der Konvergenzgeschwindigkeit im Zentralen Grenzwertsatz:*

Satz 8.10 (Berry-Esséen). *Seien $X_i \in \mathcal{L}^3$ i.i.d. Zufallsvariablen, $Z \sim N(0,1)$, und seien*

$$\begin{aligned} F_n(x) &:= P\left[\frac{S_n - E[S_n]}{\sigma\sqrt{n}} \leq x\right], \\ \Phi(x) &:= P[Z \leq x]. \end{aligned}$$

Dann gilt folgende Abschätzung:

$$\sup_{x\in\mathbb{R}} |F_n(x) - \Phi(x)| \leq \frac{3\cdot E[|X_1 - E[X_1]|^3]}{\sigma^3\sqrt{n}}.$$

Den Beweis dieser Aussage findet man etwa im Buch PROBABILITY THEORY von R. Durrett (4.10).

Für die Normalapproximation der Binomialverteilung $Bin(n,p)$ ergibt sich beispielsweise

$$\frac{3 \cdot E[|X_1 - E[X_1]|^3]}{\sigma^3 \sqrt{n}} = \frac{3 \cdot ((1-p)^2 + p^2)}{\sqrt{np(1-p)}}.$$

Für $p \to 0$ oder $p \to 1$ divergiert die rechte Seite. Wir erhalten also möglicherweise einen hohen Approximationsfehler für p nahe 0 oder 1. In diesen Fällen empfiehlt sich in der Tat die Verwendung der Poisson-Approximation anstelle des zentralen Grenzwertsatzes.

Heavy Tails, Konvergenz gegen α-stabile Verteilungen

Als nächstes betrachten wir ein Beispiel, welches zeigt, dass die Voraussetzung der Quadratintegrierbarkeit der Zufallsvariablen essentiell für den zentralen Grenzwertsatz ist:

Seien $\alpha \in (1,2), r \in (0,\infty)$, und seien $X_1, X_2, \ldots : \Omega \to \mathbb{R}$ unabhängige identisch verteilte absolutstetige Zufallsvariablen, deren Dichtefunktion

$$f_{X_i}(x) = |x|^{-\alpha-1} \qquad \text{für alle } |x| \geq r$$

erfüllt. Da die Dichte für $|x| \to \infty$ nur langsam abfällt, sind die Zufallsvariablen nicht quadratintegrierbar; sie sind aber integrierbar. Daher ergibt sich ein anderes asymptotisches Verhalten der charakteristischen Funktionen für $t \to 0$:

Lemma 8.11. *Für $t \to 0$ gilt*

$$\phi_{X_i}(t) = 1 + imt - c|t|^\alpha + O(t^2)$$

mit $m = E[X_i]$ und $c = \int_{\mathbb{R}} (1 - \cos u)|u|^{-\alpha-1}\, du \in (0,\infty)$.

Beweis. Sei $t \neq 0$. Wegen $e^{iu} - 1 - iu = O(u^2)$ und $\cos u - 1 = O(u^2)$ erhalten wir

$$\begin{aligned}
\phi_{X_i}(t) - 1 - imt &= \int_{-\infty}^{\infty} (e^{itx} - 1 - itx) f(x)\, dx \\
&= \int_{-\infty}^{\infty} (e^{iu} - 1 - iu) f\left(\frac{u}{t}\right) \frac{1}{|t|}\, du \\
&= \frac{1}{|t|} \int_{-tr}^{tr} (e^{iu} - 1 - iu) f\left(\frac{u}{t}\right)\, du + |t^\alpha| \int_{[-tr,tr]^C} (\cos u - 1)|u|^{-\alpha-1}\, du \\
&= -c|t|^\alpha + O(t^2).
\end{aligned}$$

□

Für die zentrierten Summen $S_n = \sum_{i=1}^{n}(X_i - m)$ folgt nach dem Lemma:

$$\phi_{S_n}(t) \quad = \quad (1 - c|t|^\alpha + O(t^2))^n.$$

Um Konvergenz der charakteristischen Funktionen zu erhalten, müssen wir X_n nun mit $n^{-1/\alpha}$ statt $n^{-1/2}$ reskalieren:

$$\begin{aligned} \phi_{n^{-1/\alpha}S_n}(t) &= \phi_{S_n}(n^{-1/\alpha}t) \quad = \quad (1 - c|t|^\alpha n^{-1} + O(n^{-2/\alpha}))^n \\ &\to \exp(-c|t|^\alpha) \qquad \text{für } n \to \infty. \end{aligned}$$

Nach dem Konvergenzsatz von Lévy folgt:

Satz 8.12. *Für $n \to \infty$ gilt*

$$n^{-1/\alpha}S_n \quad \xrightarrow{\mathcal{D}} \quad \mu_{c,\alpha},$$

wobei $\mu_{c,\alpha}$ die Wahrscheinlichkeitsverteilung mit charakteristischer Funktion

$$\phi_{c,\alpha}(t) \quad = \quad \exp(-c|t|^\alpha)$$

ist.

Definition. *Seien $\alpha \in (0, 2]$ und $m \in \mathbb{R}$. Die Wahrscheinlichkeitsverteilungen mit charakteristischer Funktion*

$$\phi(t) \quad = \quad \exp(imt - c|t|^\alpha),$$

*$c \in (0, \infty)$, heißen **symmetrische α-stabile Verteilungen** mit Mittelwert m.*

Die Dichten der α-stabilen Verteilungen sind für $\alpha \neq 1, 2$ nicht explizit berechenbar, fallen aber für $|x| \to \infty$ wie $|x|^{-\alpha-1}$ ab. Für $\alpha = 1$ erhält man die Cauchyverteilungen, für $\alpha = 2$ die Normalverteilungen. Satz 8.12 ist ein Spezialfall eines allgemeineren Grenzwertsatzes für Summen von Zufallsvariablen mit polynomiellen Tails, siehe z.B. BREIMAN, THEOREM 9.34.

Der Satz von Lindeberg-Feller

Wir wollen nun die Annahme fallen lassen, dass die Summanden X_i identisch verteilt sind, und zeigen, dass trotzdem ein zentraler Grenzwertsatz gilt. Sei

$$\widehat{S}_n \quad = \quad Y_{n,1} + Y_{n,2} + ... + Y_{n,n} \qquad \text{mit } Y_{n,i} \in \mathcal{L}^2(\Omega, \mathcal{A}, P).$$

Die Zufallsvariablen $Y_{n,i}$ können etwa kleine Störungen oder Messfehler beschreiben. Setzen wir

$$Y_{n,i} = \frac{X_i - E[X_i]}{\sqrt{n}} \quad \text{mit } X_i \in \mathcal{L}^2 \text{ unabhängig}, \tag{8.4.3}$$

so erhalten wir das Setup von oben.

Satz 8.13 (ZGS von Lindeberg-Feller). *Sei $\sigma \in (0, \infty)$. Es gelte:*

(i) $Y_{n,i}$ $(1 \le i \le n)$ *sind unabhängig für jedes* $n \in \mathbb{N}$ *mit* $E[Y_{n,i}] = 0$,

(ii) $\mathrm{Var}[\widehat{S}_n] = \sum_{i=1}^n \mathrm{Var}[Y_{n,i}] \overset{n\uparrow\infty}{\longrightarrow} \sigma^2$,

(iii) $\gamma_{n,\varepsilon} := \sum_{i=1}^n E[Y_{n,i}^2;\ |Y_{n,i}| > \varepsilon] \overset{n\uparrow\infty}{\longrightarrow} 0 \quad \forall\, \varepsilon > 0.$

Dann konvergiert die Verteilung von $\widehat{S}_n$ schwach gegen $N(0, \sigma^2)$.

Der Satz zeigt, dass die Summe vieler kleiner unabhängiger Störungen unter geeigneten Voraussetzungen ungefähr normalverteilt ist. Dies rechtfertigt bis zu einem gewissen Grad, dass Zufallsgrößen mit unbekannter Verteilung, die durch Überlagerung vieler kleiner Effekte entstehen, häufig durch normalverteilte Zufallsvariablen modelliert werden.

Bemerkung. (1). Der Zentrale Grenzwertsatz von oben ist ein Spezialfall des Satzes von Lindeberg-Feller: Sind $X_i \in \mathcal{L}^2$ i.i.d. Zufallsvariablen mit $E[X_i] = m$ und $\mathrm{Var}[X_i] = \sigma^2$, und definieren wir $Y_{n,i}$ wie in (8.4.3), dann gilt:

$$\mathrm{Var}[\widehat{S}_n] = \frac{1}{n}\sum_{i=1}^n \mathrm{Var}[X_i] = \mathrm{Var}[X_1] = \sigma^2, \qquad \text{für alle } n \in \mathbb{N},$$

und, für $\varepsilon > 0$

$$\begin{aligned}\gamma_{n,\varepsilon} &= \sum_{i=1}^n E\left[Y_{n,i}^2; |Y_{n,i}| > \varepsilon\right] = \frac{1}{n}\sum_{i=1}^n E\left[|X_i - m|^2; |X_i - m| > \varepsilon\sqrt{n}\right] \\ &= E\left[|X_1 - m|^2; |X_1 - m| > \varepsilon\sqrt{n}\right] \to 0 \qquad \text{für } n \to \infty,\end{aligned}$$

da X_1 quadratintegrierbar ist.

(2). Die Bedingung (iii) ist insbesondere erfüllt, wenn die *Lyapunovbedingung*

$$\sum_{i=1}^n E[|Y_{n,i}|^p] \overset{n\to\infty}{\longrightarrow} 0 \quad \text{für ein } p > 2 \text{ gilt,}$$

denn für $\varepsilon > 0$ ist $E[Y_{n,i}^2; |Y_{n,i}| \ge \varepsilon] \le E[|Y_{n,i}|^p]/\varepsilon^{p-2}$.

Wir beweisen nun den Satz von Lindeberg-Feller: Der Beweis basiert wieder auf einer Analyse der Asymptotik der charakteristischen Funktionen. Dazu zeigen wir zunächst einige asymptotische Abschätzungen:

Beweis. (a) ***Vorüberlegungen:*** Sei $t \in \mathbb{R}$ fest.

(I) *Taylorapproximation für* $\phi_{n,i}(t) := E[e^{itY_{n,i}}]$*:*

Aus den verschiedenen Abschätzungen des Taylorrestglieds erhält man

$$e^{ix} = 1 + ix - \frac{x^2}{2} + R(x) \quad \text{mit} \quad |R(x)| \leq \min\left(\frac{|x|^3}{6}\,,\, x^2\right). \tag{8.4.4}$$

Damit ergibt sich

$$\phi_{n,i}(t) \;=\; 1 + itE[Y_{n,i}] - \frac{t^2}{2}\,E[Y_{n,i}^2] + E[R(tY_{n,i})] \;=\; 1 - \frac{t^2\sigma_{n,i}^2}{2} + R_{n,i},$$

wobei für $R_{n,i} := E[R(tY_{n,i})]$ die Abschätzung

$$|R_{n,i}| \;\leq\; E\left[\min\left(\frac{|tY_{n,i}|^3}{6}, t^2Y_{n,i}^2\right)\right] \tag{8.4.5}$$

gilt.

(II) Wir zeigen $\sum_{i=1}^n |R_{n,i}| \to 0$ *für* $n \to \infty$:

Für $\varepsilon > 0$ gilt nach (8.4.5):

$$|R_{n,i}| \leq \frac{1}{6} \cdot E\left[|tY_{n,i}|^3; |Y_{n,i}| \leq \varepsilon\right] + E[|tY_{n,i}|^2; |Y_{n,i}| > \varepsilon].$$

Mit $E\left[|tY_{n,i}|^3; |Y_{n,i}| \leq \varepsilon\right] \;\leq\; |t|^3\varepsilon \cdot \sigma_{n,i}^2$ erhalten wir

$$\sum_{i=1}^n |R_{n,i}| \;\leq\; \frac{|t|^3\varepsilon}{6}\sum_{i=1}^n \sigma_{n,i}^2 + t^2\gamma_{n,\varepsilon},$$

und somit nach Voraussetzung (ii) und (iii)

$$\limsup_{n\to\infty} \sum_{i=1}^n |R_{n,i}| \;\leq\; \frac{\sigma^2|t|^3}{6}\varepsilon\,.$$

Die Behauptung folgt für $\varepsilon \to 0$.

(III) Wir zeigen $\sup_{1\leq i\leq n} \sigma_{n,i}^2 \to 0$ *für* $n \to \infty$:

Für $\varepsilon > 0$ und $1 \leq i \leq n$ gilt

$$\sigma_{n,i}^2 \;=\; E[Y_{n,i}^2; |Y_{n,i}| \leq \varepsilon] + E[Y_{n,i}^2; |Y_{n,i}| > \varepsilon] \;\leq\; \varepsilon^2 + \gamma_{n,\varepsilon}.$$

Wegen $\gamma_{n,\varepsilon} \to 0$ für $n \to \infty$ ergibt sich

$$\limsup_{n\to\infty} \sup_{1\leq i\leq n} \sigma_{n,i}^2 \;\leq\; \varepsilon^2.$$

Die Behauptung folgt wieder für $\varepsilon \to 0$.

(b) ***Hauptteil des Beweises:*** Zu zeigen ist

$$\phi_{\hat{S}_n}(t) = \prod_{i=1}^{n} \phi_{n,i}(t) \overset{n\to\infty}{\longrightarrow} \exp\left(-\frac{t^2\sigma^2}{2}\right), \tag{8.4.6}$$

die Aussage folgt dann aus dem Konvergenzsatz von Lévy.

Wir zeigen:

$$\left|\prod_{i=1}^{n} \phi_{n,i}(t) - \prod_{i=1}^{n}\left(1 - \frac{t^2\sigma_{n,i}^2}{2}\right)\right| \overset{n\to\infty}{\longrightarrow} 0, \qquad \text{und} \tag{8.4.7}$$

$$\prod_{i=1}^{n}\left(1 - \frac{t^2\sigma_{n,i}^2}{2}\right) \overset{n\to\infty}{\longrightarrow} e^{-\frac{t^2\sigma^2}{2}}. \tag{8.4.8}$$

Daraus folgt (8.4.6), und damit die Behauptung.

Beweis von (8.4.7): Wie oben gezeigt, gilt für $z_i, w_i \in \mathbb{C}$ mit $|z_i|, |w_i| \le 1$:

$$\left|\prod_{i=1}^{n} z_i - \prod_{i=1}^{n} w_i\right| \le \sum_{i=1}^{n} |z_i - w_i|.$$

Zudem gilt $|\phi_{n,i}(t)| \le 1$, und nach der 3. Vorüberlegung existiert ein $n_0 \in \mathbb{N}$ mit

$$1 - \frac{t^2\sigma_{n,i}^2}{2} \in (0,1) \qquad \text{für alle } n \ge n_0 \text{ und } 1 \le i \le n. \tag{8.4.9}$$

Damit erhalten wir für $n \ge n_0$:

$$\left|\prod_{i=1}^{n} \phi_{n,i}(t) - \prod_{i=1}^{n}\left(1 - \frac{t^2\sigma_{n,i}^2}{2}\right)\right| \le \sum_{i=1}^{n}\left|\phi_{n,i}(t) - \left(1 - \frac{t^2\sigma_{n,i}^2}{2}\right)\right| = \sum_{i=1}^{n} |R_{n,i}|$$

Die rechte Seite konvergiert nach der 2. Vorüberlegung gegen 0.

Beweis von (8.4.8): Wegen (8.4.9) erhalten wir

$$\begin{aligned}\log\left(\prod_{i=1}^{n}\left(1 - \frac{t^2\sigma_{n,i}^2}{2}\right)\right) &= \sum_{i=1}^{n}\log\left(1 - \frac{t^2\sigma_{n,i}^2}{2}\right) \\ &= -\sum_{i=1}^{n}\frac{t^2\sigma_{n,i}^2}{2} + \sum_{i=1}^{n}\widetilde{R}_{n,i},\end{aligned}$$

wobei $|\widetilde{R}_{n,i}| \le C \cdot \left(t^2\sigma_{n,i}^2\right)^2$ mit $\widetilde{C} \in (0,\infty)$. Die rechte Seite konvergiert nach Voraussetzung (ii) für $n \to \infty$ gegen $-\frac{t^2\sigma^2}{2}$, denn

$$\sum_{i=1}^{n}|\widetilde{R}_{n,i}| \le Ct^4 \cdot \sum_{i=1}^{n}\sigma_{n,i}^4 \le Ct^4 \cdot \sum_{i=1}^{n}\sigma_{n,i}^2 \cdot \sup_{1\le i\le n}\sigma_{n,i}^2 \to 0$$

nach der 3. Vorüberlegung.

□

Bemerkung (Zentrale Grenzwertsätze für Summen abhängiger Zufallsvariablen). In allen Fällen haben wir bisher angenommen, dass die Zufallsvariablen X_i unabhängig sind. Tatsächlich hat man zentrale Grenzwertsätze auch für viele große Modellklassen mit Abhängigkeit gezeigt, beispielsweise für Martingale, additive Funktionale von Markovketten, Skalierungslimiten von Teilchensystemen, unterschiedliche Folgen von Parameterschätzern in der Statistik, usw. Wir werden darauf in weiterführenden Vorlesungen zurückkommen.

8.5 Vom Random Walk zur Brownschen Bewegung

Kapitel 9

Multivariate Verteilungen und statistische Anwendungen

9.1 Mehrstufige Modelle

Seien $(S_i, \mathcal{S}_i), 1 \leq i \leq n$, messbare Räume. Wir wollen allgemeine Wahrscheinlichkeitsverteilungen auf dem Produktraum $S_1 \times ... \times S_n$ konstruieren und effektiv beschreiben. In Analogie zu diskreten, mehrstufigen Modellen versuchen wir diese in der Form

$$P(dx_1...dx_n) = \mu(dx_1)p(x_1, dx_2)p((x_1, x_2), dx_3) \cdots p((x_1, ..., x_{n-1}), dx_n)$$

darzustellen.

Stochastische Kerne und der Satz von Fubini

Wir betrachten zunächst den Fall $n = 2$, der allgemeine Fall ergibt sich dann durch Iteration der Konstruktion. Seien also $(S, \mathcal{S})$ und $(T, \mathscr{T})$ messbare Räume, und sei

$$\Omega := S \times T \quad \text{und} \quad \mathcal{A} := \mathcal{S} \otimes \mathscr{T} \quad \text{die Produkt-}\sigma\text{-Algebra.}$$

Unser Ziel ist die Konstruktion einer Wahrscheinlichkeitsverteilung auf $(\Omega, \mathcal{A})$ vom Typ

$$P(dxdy) = \mu(dx)p(x, dy).$$

Definition. *Eine Abbildung*

$$p : S \times \mathscr{T} \longrightarrow [0, 1], \qquad (x, C) \mapsto p(x, C),$$

heißt ***stochastischer Kern*** *(oder* ***Übergangswahrscheinlichkeit****), wenn gilt:*

(i) $p(x, \bullet)$ *ist für jedes* $x \in S$ *eine Wahrscheinlichkeitsverteilung auf* $(T, \mathscr{T})$*,*

(ii) $p(\bullet, C)$ *ist für jedes* $C \in \mathscr{T}$ *eine messbare Funktion auf* $(S, \mathcal{S})$*.*

Bemerkung (Diskreter Spezialfall). Sind S und T abzählbar mit $\mathcal{S} = \mathscr{P}(S), \mathscr{T} = \mathscr{P}(T)$, dann ist p eindeutig festgelegt durch die Matrix mit Komponenten

$$p(x, y) \ := \ p(x, \{y\}) \qquad (x \in S \ , \ y \in T).$$

Da p ein stochastischer Kern ist, ist $p(x, y)$ $(x \in S, y \in T)$ eine *stochastische Matrix*.

Der folgende Satz zeigt im allgemeinen Fall die Existenz eines zweistufigen Modells mit μ als Verteilung der ersten Komponente, und $p(x, \bullet)$ als bedingte Verteilung der zweiten Komponente gegeben den Wert x der ersten Komponente. Der Satz zeigt zudem, dass Erwartungswerte im mehrstufigen Modell durch Hintereinanderausführen von Integralen berechnet werden können.

Satz 9.1 (Fubini). *Sei* $\mu(dx)$ *eine Wahrscheinlichkeitsverteilung auf* $(S, \mathcal{S})$ *und* $p(x, dy)$ *ein stochastischer Kern von* $(S, \mathcal{S})$ *nach* $(T, \mathscr{T})$*. Dann existiert eine eindeutige Wahrscheinlichkeitsverteilung* $\mu \otimes p$ *auf* $(\Omega, \mathcal{A})$ *mit*

$$(\mu \otimes p)[B \times C] = \int_B \mu(dx) \, p(x, C) \qquad \textit{für alle } B \in \mathcal{S}, \ C \in \mathscr{T}. \tag{9.1.1}$$

Für diese Wahrscheinlichkeitsverteilung gilt:

$$\int f \, d(\mu \otimes p) = \int \left(\int f(x, y) \, p(x, dy) \right) \mu(dx) \qquad \textit{für alle } \mathcal{A}\textit{-messbaren } f : \Omega \to \mathbb{R}_+. \tag{9.1.2}$$

Beweis. (1). *Eindeutigkeit:* Das Mengensystem $\{B \times C \mid B \in \mathcal{S}, \ C \in \mathscr{T}\}$ ist ein durchschnittsstabiler Erzeuger der Produkt-σ-Algebra $\mathcal{A}$. Also ist die Wahrscheinlichkeitsverteilung $\mu \otimes \nu$ durch (9.1.1) eindeutig festgelegt.

(2). *Existenz:* Wir wollen die Wahrscheinlichkeitsverteilung $\mu \otimes \nu$ über (9.1.2) mit $f = I_A, \ A \in \mathcal{A}$, definieren. Dazu müssen wir überprüfen, ob die rechte Seite in diesem Fall definiert ist (d.h. ob die Integranden messbar sind), und ob

$$(\mu \otimes p)[A] \ := \ \int \left(\int I_A(x, y) \, p(x, dy) \right) \mu(dx)$$

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

Für Produktmengen $A = B \times C$ $(B \in \mathcal{S}, C \in \mathscr{T})$ ist die Funktion $x \mapsto \int I_A(x, y) p(x, dy)$ nach Definition des stochastischen Kerns messbar. Da die Mengen $A \in \mathcal{A}$, für die diese Funktion messbar ist, ein Dynkinsystem bilden, folgt die Messbarkeit für alle $A \in \mathcal{A}$.

$\mu \otimes p$ ist eine Wahrscheinlichkeitsverteilung, denn einerseits folgt

$$(\mu \otimes p)[\Omega] = (\mu \otimes p)[S \times T] = \int \left(\int I_S(x) I_T(y) p(x, dy) \right) \mu(dx) = \mu[S] = 1$$

aus $\int_T p(x, dy) = p(x, T) = 1$; andererseits gilt für disjunkte Mengen A_i $(i \in \mathbb{N})$

$$I_{\bigcup A_i} = \sum I_{A_i},$$

woraus unter zweimaliger Anwendung des Satzes von der monotonen Konvergenz folgt:

$$\begin{aligned} (\mu \otimes p) \left[\bigcup_i A_i \right] &= \int \left(\int \sum_i I_{A_i}(x, y)\, p(x, dy) \right) \mu(dx) \\ &= \sum_i \int \left(\int I_{A_i}(x, y)\, p(x, dy) \right) \mu(dx) \\ &= \sum_i (\mu \otimes p)[A_i]. \end{aligned}$$

Durch maßtheoretische Induktion zeigt man nun, dass die Wahrscheinlichkeitsverteilung $\mu \otimes p$ auch (9.1.2) erfüllt.

□

Als nächstes wollen wir die **Randverteilungen** des gerade konstruierten zweistufigen Modells berechnen. Sei also $P := \mu \otimes p$, und seien

$$\begin{array}{rccc} X: & S \times T & \to & S \\ & (x, y) & \mapsto & x \end{array} \quad , \quad \begin{array}{rccc} Y: & S \times T & \to & T \\ & (x, y) & \mapsto & y \end{array},$$

die Projektionen auf die 1. bzw. 2. Komponente. Wegen $p(x, T) = 1$ gilt:

$$P[X \in B] = P[B \times T] = \int_B \mu(dx)\, p(x, T) = \mu[B] \qquad \forall\, B \in \mathcal{S},$$

also ist die Verteilung $P \circ X^{-1}$ der ersten Komponente gleich μ. Für die Verteilung der zweiten Komponente erhalten wir

$$P[Y \in C] = P[S \times C] = \int_S \mu(dx)\, p(x, C) \qquad \forall\, C \in \mathscr{T}.$$

Definition. *Die durch*

$$(\mu p)[C] \quad := \quad \int \mu(dx)\, p(x, C), \qquad C\}in\mathscr{T},$$

*definierte Wahrscheinlichkeitsverteilung auf $(T, \mathscr{T})$ heißt **Mischung** der Wahrscheinlichkeitsverteilungen $p(x, \bullet)$ bezüglich μ.*

Wie gerade gezeigt, ist $\mu p = P \circ Y^{-1}$ die Verteilung der zweiten Komponente im zweistufigen Modell.

Bemerkung. Sind S und T abzählbar, dann sind $\mu \otimes p$ und μp die schon in Abschnitt 2.3 betrachteten Wahrscheinlichkeitsverteilungen mit Gewichten

$$\begin{aligned}(\mu \otimes p)(x,y) &= \mu(x)\, p(x,y),\\ (\mu p)(y) &= \sum_{x\in S} \mu(x)\, p(x,y).\end{aligned}$$

Die Massenfunktionen von $\mu \otimes p$ und μp sind also das Tensor- bzw. Matrixprodukt des Zeilenvektors μ und der stochastischen Matrix p.

Wichtige Spezialfälle

Produktmaße: Ist $p(x,\bullet) \equiv \nu$ eine feste (von x unabhängige) Wahrscheinlichkeitsverteilung auf $(T,\mathscr{T})$, dann ist $\mu \otimes p$ das Produkt $\mu \otimes \nu$ der Wahrscheinlichkeitsverteilungen μ und ν. Der Satz von Fubini liefert also die Existenz des Produktmaßes, und die schon mehrfach verwendete Berechnungsformel

$$\int f\, d(\mu\otimes\nu) = \int_S \left(\int_T f(x,y)\, \nu(dy)\right) \mu(dx) \tag{9.1.3}$$

für die Integrale nicht-negativer oder integrierbarer messbarer Funktionen bzgl. des Produktmaßes. Die Integrationsreihenfolge kann man in diesem Fall vertauschen, denn wegen

$$(\mu\otimes\nu)[B\times C] = \mu[B]\nu[C] \qquad \text{für alle } B\in S,\ C\in\mathscr{T} \tag{9.1.4}$$

gilt $(\nu\otimes\mu)\circ R^{-1} = \mu\otimes\nu$, wobei $R(x,y)=(y,x)$, und damit nach dem Transformationssatz:

$$\begin{aligned}\int \left(\int f(x,y)\, \mu(dx)\right) \nu(dy) &\overset{Fub.}{=} \int f\circ R\ d(\nu\otimes\mu)\\ &= \int f\, d(\mu\otimes\nu)\\ &\overset{Fub.}{=} \int \left(\int f(x,y)\, \nu(dy)\right) \mu(dx).\end{aligned}$$

Durch wiederholte Anwendung dieses Arguments erhalten wir zudem:

Korollar 9.2. *Seien $(S_i,\mathcal{S}_i,\mu_i)$ Wahrscheinlichkeitsräume $(1\le i\le n)$. Dann existiert eine eindeutige Wahrscheinlichkeitsverteilung $\mu_1\otimes\ldots\otimes\mu_n$ auf $(S_1\times\ldots\times S_n, \mathcal{S}_1\otimes\ldots\otimes\mathcal{S}_n)$ mit:*

$$(\mu_1\otimes\ldots\otimes\mu_n)\,[B_1\times\ldots\times B_n] = \prod_{i=1}^n \mu_i[B_i] \qquad \textit{für alle } B_i\in\mathcal{S}_i \quad (1\le i\le n).$$

Für alle produktmessbaren Funktionen $f : S_1 \times ... \times S_n \to [0, \infty)$ gilt:

$$\int f \, d(\mu_1 \otimes ... \otimes \mu_n) = \int ... \left(\int f(x_1, ..., x_n) \mu_n(dx_n) \right) ... \mu_1(dx_1),$$

wobei die Integration auch in beliebiger anderer Reihenfolge ausgeführt werden kann.

Beweis. Die Existenz folgt durch wiederholte Anwendung des Satzes von Fubini, die Eindeutigkeit aus dem Eindeutigkeitssatz. Dass die Integrationsreihenfolge vertauscht werden kann, zeigt man ähnlich wie im oben betrachteten Fall $n = 2$. □

Deterministische Kopplung: Gilt $p(x, \bullet) = \delta_{f(x)}$ für eine messbare Funktion $f : S \to T$, dann folgt $(\mu \otimes p)[\{(x, y) \mid y = f(x)\}] = 1$. Die zweite Komponente ist also durch die erste Komponente mit Wahrscheinlichkeit 1 eindeutig festgelegt. Die Verteilung der zweiten Komponente ist in diesem Fall das Bild von μ unter f:

$$\mu p = \mu \circ f^{-1}.$$

Übergangskerne von Markovschen Ketten: Gilt $S = T$, dann können wir $p(x, dy)$ als Übergangswahrscheinlichkeit (Bewegungsgesetz) einer Markovkette auf $(S, \mathcal{S})$ auffassen. In Analogie zum diskreten Fall definieren wir:

Definition. *Eine Wahrscheinlichkeitsverteilung μ auf $(S, \mathcal{S})$ heißt **Gleichgewicht (stationäre oder auch invariante Verteilung)** von p, falls $\mu p = \mu$ gilt, d.h. falls*

$$\int \mu(dx) p(x, B) = \mu[B] \qquad \textit{für alle } B \in \mathcal{S}.$$

Beispiel (Autoregressiver Prozess). Der AR(1)-Prozess mit Parametern $\varepsilon, \alpha \in \mathbb{R}$ ist eine Markovkette mit Übergangskern $p(x, \bullet) = N(\alpha x, \varepsilon^2)$. Die Normalverteilung $N\left(0, \frac{\varepsilon^2}{1-\alpha^2}\right)$ ist für $\alpha \in (0, 1)$ ein Gleichgewicht. Für $\alpha \geq 1$ existiert kein Gleichgewicht.

Bedingte Dichten und Bayessche Formel

Wir betrachten nun Situationen mit nichttrivialer Abhängigkeit zwischen den Komponenten im kontinuierlichen Fall. Seien $X : \Omega \to \mathbb{R}^n$ und $Y : \Omega \to \mathbb{R}^m$ Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$, deren gemeinsame Verteilung absolutstetig ist mit Dichte $f_{X,Y}$, d.h.

$$P[x \in B, Y \in C] \quad = \quad \int_B \int_C f_{X,Y}(x, y) \, dy \, dx \qquad \text{für alle } B \in \mathcal{B}(\mathbb{R}^n), C \in \mathcal{B}(\mathbb{R}^m).$$

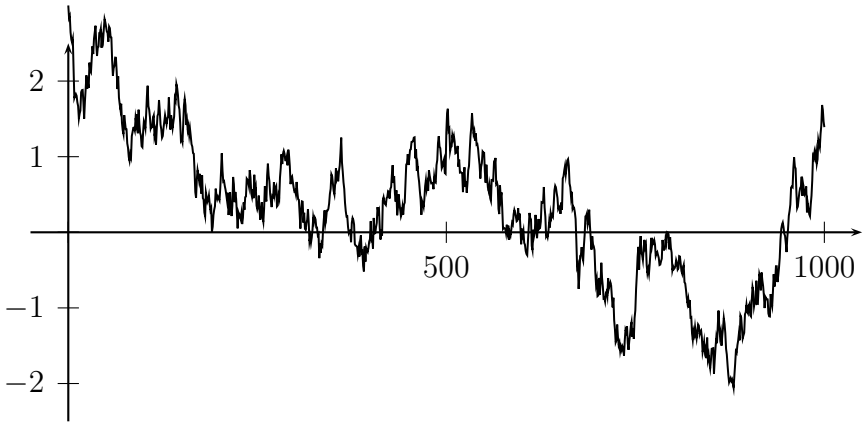


Abbildung 9.1: Simulation einer Trajektorie eines AR(1)-Prozesses mit Parametern $\alpha = 0.8$ und $\varepsilon^2 = 1.5$.

Nach dem Satz von Fubini sind dann auch die Verteilungen von X und Y absolutstetig mit dichten

$$f_X(x) \quad = \quad \int\limits_{\mathbb{R}^m} f_{X,Y}(x,y)\, dy$$

und

$$f_Y(x) \quad = \quad \int\limits_{\mathbb{R}^n} f_{X,Y}(x,y)\, dx.$$

Obwohl bedingte Wahrscheinlichkeiten gegeben $Y = y$ nicht im herkömmlichen Sinn definiert werden können, da das Ereignis $\{Y = y\}$ eine Nullmenge ist, können wir die bedingte Dichte und die bedingte Verteilung von X gegeben Y in diesem Fall sinnvoll definieren. Anschaulich beträgt die Wahrscheinlichkeit, dass der Wert X in einem infinitesimal kleinen Volumenelement dx liegt, gegeben, dass der Wert von Y in einem entsprechenden infinitesimalen Volumenelement dy liegt:

$$\begin{aligned} P[X \in dx | Y \in dy] \quad &= \quad \frac{P[X \in dx, Y \in dy]}{P[Y \in dy]} \quad = \quad \frac{f_{X,Y}(x,y)\, dx\, dy}{f_Y(y)\, dy} \\ &= \quad \frac{f_{X,Y}(x,y)}{f_Y(y)} dx \end{aligned}$$

Diese heuristische Überlegung motiviert die folgende Definition:

Definition. *Die Funktion $f_{X|Y} : \mathbb{R}^n \times \mathbb{R}^m \to [0,\infty]$ mit*

$$f_{X|Y} \quad = \quad \begin{cases} \dfrac{f_{X,Y}(x,y)}{f_Y(y)} & \textit{falls } f_Y(y) \neq 0 \\ f_X(x) & \textit{falls } f_Y(y) = 0 \end{cases}$$

heißt ***bedingte Dichte von*** X ***gegeben*** Y*, und die von* y *abhängende Wahrscheinlichkeitsverteilung*

$$\mu_{X|Y}(y, B) \quad := \quad \int\limits_B f_{X|Y}(x, y)\, dx, \qquad \textit{für } B \in \mathcal{B}(\mathbb{R}^n),$$

heißt ***bedingte Verteilung von*** X ***gegeben*** Y.

Bemerkung. (1). Für festes y ist die bedingte Dichte eine Wahrscheinlichkeitsdichte auf $\mathbb{R}^n$. Da $f_{X|Y}$ produktmessbar ist, ist die bedingte Verteilung $\mu_{X|Y}$ nach dem Satz von Fubini ein *stochastischer Kern* von $\mathbb{R}^m$ nach $\mathbb{R}^n$.

(2). Auf der Nullmenge $\{y \in \mathbb{R}^m | f_Y(y) = 0\}$ sind $f_{X|Y}(x|y)$ und $\mu_{X|Y}(y, dx)$ nicht eindeutig festgelegt - die oben getroffene Definition über die unbedingte Dichte ist relativ willkürlich.

Aus der Definition der bedingten Dichte ergibt sich unmittelbar eine Variante der Bayesschen Formel für absolutstetige Zufallsvariablen:

Satz 9.3 (Bayessche Formel). *Für* $(x, y) \in \mathbb{R}^n \times \mathbb{R}^m$ *mit* $f_X(x) > 0$ *und* $f_Y(y) > 0$ *gilt*

$$f_{X|Y}(x|y) \quad = \quad \frac{f_X(x) f_{Y|X}(y|x)}{\int\limits_{\mathbb{R}^n} f_X(x) f_{Y|X}(y|x)\, dx}.$$

Beweis. Aus der Definition folgt

$$f_{X|Y}(x|y) \quad = \quad \frac{f_{X,Y}(x, y)}{f_Y(y)} \quad = \quad \frac{f_{X,Y}(x, y)}{\int\limits_{\mathbb{R}^n} f_{X,Y}(x, y)\, dx},$$

und damit die Behauptung. □

In Modellen der Bayesschen Statistik interpretiert man $f_X(x)$ als Dichte der *a priori* angenommenen Verteilung eines unbekannten Parameters X, und $f_{Y|X}(y|x)$ als Maß für die Plausibilität („Likelihood“) des Parameterwertes x, wenn der Wert y der Zufallsgröße Y beobachtet wird. Die Bayessche Formel besagt dann, dass die Verteilung von X, von der man *a posteriori* (d.h. nach der Beobachtung von y) ausgeht, die Dichte

$$\begin{aligned} f_{X|Y}(x|y) \quad &= \quad \text{const.}(y) \cdot f_X(x) \cdot f_{Y|X}(y|x) \\ \text{A posteriori Dichte} \quad &\propto \quad \text{A priori Dichte} \ \times \ \text{Likelihood} \end{aligned}$$

hat. Trotz der einfachen Form der Bayesschen Formel ist es im Allgemeinen nicht trivial, Stichproben von der A-posteriori-Verteilung zu simulieren, und Erwartungswerte numerisch zu berechnen. Problematisch ist u.A., dass die Berechnung der Normierungskonstanten die Auswertung eines (häufig hochdimensionalen) Integrals erfordert. Ein wichtiges Verfahren zur Simulation von Stichproben in diesem Zusammenhang ist der Gibbs-Sampler.

Sind X und Y gemeinsam normalverteilt, dann kann man die wichtigsten Erwartungswerte bzgl. der A-posteriori-Verteilung im Prinzip exakt berechnen. Wir demonstrieren dies nun in einem grundlegenden Beispiel eines zweistufigen Modells. Ähnliche Modelle treten in zahlreichen Anwendungen auf.

Beispiel (Signalverarbeitung). Sei $S = T = \mathbb{R}^1$, also

$$S \times T = \mathbb{R}^2 = \{(x, y) \mid x, y \in \mathbb{R}\}.$$

Wir interpretieren die erste Komponente x als Größe eines nicht direkt beobachtbaren Signals, und die zweite Komponente y als verrauschte Beobachtung von x. In einem einfachen Bayesschen Modell nimmt man z.B. a priori an, dass Signal und Beobachtung normalverteilt sind:

$$\begin{aligned} \text{Signal} \quad x &\sim N(0, v)\,, \quad v > 0, \\ \text{Beobachtung} \quad y &\sim N(x, \varepsilon)\,, \quad \varepsilon > 0. \end{aligned}$$

Die Verteilung der ersten Komponente und der Übergangskern zur zweiten Komponente sind dann:

$$\begin{aligned} \mu(dx) &= f_X(x)\ \lambda(dx) \\ p(x, dy) &= f_{Y|X}(y|x)\ \lambda(dy) \end{aligned}$$

mit den Dichten

$$\begin{aligned} f_X(x) &:= \frac{1}{\sqrt{2\pi v}}\, e^{-\frac{x^2}{2v}} && \text{(Dichte der Verteilung der ersten Komponente } X\text{)}, \\ f_{Y|X}(y|x) &:= \frac{1}{\sqrt{2\pi\varepsilon}}\, e^{-\frac{(y-x)^2}{2\varepsilon}} && \text{(bedingte Dichte der zweiten Komponente } Y \text{ gegeben } X = x\text{)}. \end{aligned}$$

Die gemeinsame Verteilung von Signal und Beobachtungswert ist

$$\begin{aligned} (\mu \otimes p)(dxdy) &= \mu(dx)\ p(x, dy) \\ &= \frac{1}{2\pi\sqrt{v\varepsilon}}\ \exp\left(-\frac{(\varepsilon + v)x^2 - 2vxy + vy^2}{2v\varepsilon}\right)\ \lambda(dx)\lambda(dy) \\ &= \frac{1}{2\pi\sqrt{\det C}}\ \exp\left(-\frac{1}{2}\begin{pmatrix} x \\ y \end{pmatrix} \cdot C^{-1}\begin{pmatrix} x \\ y \end{pmatrix}\right)\ \lambda^2(dx\ dy). \end{aligned}$$

D.h. $\mu \otimes p$ ist eine zweidimensionale Normalverteilung mit Kovarianzmatrix

$$C = \begin{pmatrix} v & v \\ v & v + \varepsilon \end{pmatrix} \quad .$$

Mit anderen Worten: Die gemeinsame Verteilung von X und Y ist absolutstetig bzgl. des zweidimensionalen Lebesguemaßes mit Dichte

$$f_{X,Y}(x,y) \quad = \quad f_X(x) f_{Y|X}(y|x) \quad = \quad \frac{1}{2\pi\sqrt{\det C}} \exp\left(-\frac{1}{2}\begin{pmatrix}x\\y\end{pmatrix}^\top \cdot C^{-1}\begin{pmatrix}x\\y\end{pmatrix}\right).$$

Als Dichte der Verteilung μp von Y ergibt sich:

$$f_Y(y) = \int f_{X,Y}(x,y)\, dx.$$

Nach der Bayesschen Formel erhalten wir für die A-posteriori dichte des Signals gegeben die Beobachtung y:

$$\begin{aligned} f_{X|Y}(x|y) &:= \frac{f_{X,Y}(x,y)}{f_Y(y)} \\ &= \frac{f_X(x) f_{Y|X}(y|x)}{\int f_X(x) f_{Y|X}(y|x)\, \lambda(dx)} \\ &= const(y) \cdot \exp\left(-\frac{\varepsilon+v}{2v\varepsilon}(x - \frac{v}{v+\varepsilon}y)^2\right). \end{aligned} \tag{9.1.5}$$

Die bedingte Verteilung des Signals gegeben die Beobachtung ist also $N(\hat{x}, u)$, wobei

$$\begin{aligned} \hat{x} &= \frac{v}{v+\varepsilon}\, y && \text{der Prognosewert ist, und} \\ u &= \frac{v\varepsilon}{v+\varepsilon} = \left(\frac{1}{v}+\frac{1}{\varepsilon}\right)^{-1} && \text{die Varianz der Prognose.} \end{aligned}$$

In einem Bayesschen Modell würden wir also nach der Beobachtung mit einer Standardabweichung $\sigma = \sqrt{u}$ prognostizieren, dass der Signalwert gleich $\hat{x}$ ist.

Ähnliche Modellierungsansätze werden auch in viel allgemeinerem Kontext verwendet. Beispielsweise wird in stochastischen Filterproblemen das Signal durch eine Markovkette (oder einen zeitstetigen Markovprozess) beschrieben, und die Folge der Beobachtungen durch einen von der Markovkette angetriebenen stochastischen Prozess. Sind alle gemeinsamen Verteilungen Gaußsch, dann kann man auch hier die a posteriori Verteilung im Prinzip exakt berechnen – andernfalls muss man auf numerische Näherungsmethoden (z.B. Partikelfilter) zurückgreifen.

9.2 Summen unabhängiger Zufallsvariablen, Faltung

Seien X und Y unabhängige reellwertige Zufallsvariablen auf $(\Omega, \mathcal{A}, P)$ mit Verteilungen μ bzw. ν. Wir wollen die Verteilung von $X+Y$ bestimmen. Für diskrete Zufallsvariablen ergibt sich:

$$P[X+Y=z] \quad = \quad \sum_{x \in X(\Omega)} \underbrace{P[X=x, Y=z-x]}_{=P[X=x]\cdot P[Y=z-x]} = \sum_{x\in X(\Omega)} \mu(x)\nu(z-x) \tag{9.2.1}$$

Die Wahrscheinlichkeitsverteilung mit Massenfunktion

$$(\mu \star \nu)(z) \quad = \quad \sum_{x \in X(\Omega)} \mu(x)\nu(z-x)$$

heißt Faltung von μ und ν. Eine entsprechende Aussage erhält man auch im allgemeinen Fall:

Verteilungen von Summen unabhängiger Zufallsvariablen

Satz 9.4. *Seien X und Y unabhängige reellwertige Zufallsvariablen mit Verteilungen μ bzw. ν. Dann ist die Verteilung von $X+Y$ die durch*

$$(\mu \star \nu)[B] \quad := \quad \int \mu(dx)\, \nu[B-x]\,, \qquad B \in \mathcal{B}(\mathbb{R}),$$

*definierte **Faltung** der Wahrscheinlichkeitsverteilungen μ und ν.*

Beweis. Sei $\widetilde{B} := \{(x,y) \mid x+y \in B\}$. Da X und Y unabhängig sind, erhalten wir mit dem Satz von Fubini

$$\begin{aligned} P[X+Y \in B] \quad &= \quad P[(X,Y) \in \widetilde{B}] \quad = \quad (\mu \otimes \nu)[\widetilde{B}] \\ &\overset{Fubini}{=} \quad \int \mu(dx) \int \nu(dy) \underbrace{I_B(x+y)}_{=I_{B-x}(y)} \quad = \quad \int \mu(dx)\, \nu[B-x]. \end{aligned}$$

□

Bemerkung. Die Faltung $\mu \star \nu$ zweier Wahrscheinlichkeitsverteilungen μ und ν auf $\mathbb{R}^1$ ist wieder eine Wahrscheinlichkeitsverteilung auf $\mathbb{R}^1$. Da die Addition von Zufallsvariablen kommutativ und assoziativ ist, hat die Faltung von Wahrscheinlichkeitsverteilungen nach Satz 9.4 dieselben Eigenschaften:

$$\mu \star \nu \quad = \quad \nu \star \mu \qquad (\text{da } X+Y = Y+X) \tag{9.2.2}$$

$$(\mu \star \nu) \star \eta \quad = \quad \mu \star (\nu \star \eta) \qquad (\text{da } (X+Y)+Z = X+(Y+Z)\,). \tag{9.2.3}$$

Im diskreten Fall ist $\mu \star \nu$ nach (9.2.2) die Wahrscheinlichkeitsverteilung mit Gewichten

$$(\mu \star \nu)(z) \quad = \quad \sum_x \mu(x)\nu(z-x).$$

Eine entsprechende Berechnungsformel ergibt sich auch für absolutstetige Wahrscheinlichkeitsverteilungen:

Lemma 9.5. *Ist ν absolutstetig mit Dichte g, dann ist auch $\mu \star \nu$ absolutstetig mit Dichte*

$$\varrho(z) \;=\; \int \mu(dx)\; g(z-x).$$

Ist zusätzlich auch μ absolutstetig mit Dichte f, dann gilt

$$\varrho(z) \;=\; \int_{\mathbb{R}} f(x)\; g(z-x)\; dx \;=:\; (f \star g)(z)$$

Beweis. Wegen der Translationsinvarianz des Lebesguemaßes gilt

$$(\mu \star \nu)[B] = \int \mu(dx)\nu[B-x] = \int \mu(dx) \underbrace{\int_{B-x} g(y)dy}_{=\int_B g(z-x)dz} \overset{Fub.}{=} \int_B \left(\int \mu(dx) g(z-x) \right) dz \; .$$

Also ist $\mu \star \nu$ absolutstetig mit Dichte ϱ. Die zweite Behauptung folgt unmittelbar. □

Beispiel. (1). 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 $Z_1 + ... + Z_n$ und $Z_{n+1} + ... + Z_{n+m}$ übereinstimmt, wobei die Zufallsvariablen Z_i $(1 \le i \le n+m)$ unabhängig und $\mathrm{Bernoulli}(p)$-verteilt sind. Also folgt:

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

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:

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

d.h.

$$\sum_{k=0}^{l} \binom{n}{k} \binom{m}{l-k} = \binom{n+m}{l}. \tag{9.2.4}$$

Die kombinatorische Formel (9.2.4) ist auch als *Vandermonde-Identität* bekannt.

(2). 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 $n \geq 0$:

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

Also bilden auch die Poissonverteilungen eine Faltungshalbgruppe:

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

(3). Sind X und Y unabhängig und normalverteilt mit Parametern (m,σ^2) bzw. $(\widetilde{m},\widetilde{\sigma}^2)$, dann ist $X+Y$ normalverteilt mit Parametern $(m+\widetilde{m},\sigma^2+\widetilde{\sigma}^2)$, siehe **??**. Dies verifiziert man leicht mithilfe der charakteristischen Funktionen. Die Normalverteilungen bilden also eine zweiparametrige Faltungshalbgruppe.

Wartezeiten, Gamma-Verteilung

Seien $T_1, T_2, ...$ sukzessive Wartezeiten auf das Eintreten eines unvorhersehbaren Ereignisses. In einem einfachen Modell nehmen wir an, dass die T_i $(i \in \mathbb{N})$ unabhängige exponentialverteilte Zufallsvariablen sind, d.h. die Verteilungen der T_i sind absolutstetig mit Dichte

$$f(t) \ = \ \lambda \cdot e^{-\lambda t} \cdot I_{(0,\infty)}(t)\ .$$

Die Verteilung der Gesamtwartezeit

$$S_n = T_1 + ... + T_n$$

bis zum n-ten Ereignis ist dann

$$\mu_{S_n} = \mu_{T_1} \star \mu_{T_2} \star ... \star \mu_{T_n}\ .$$

Insbesondere ist die Verteilung von S_2 absolutstetig mit Dichte

$$(f \star f)(s) = \int_{\mathbb{R}} \underbrace{f(x)}_{\substack{=0 \\ \text{für } x<0}} \underbrace{f(s-x)}_{\substack{=0 \\ \text{für } x>s}} = \int_0^s \lambda^2 e^{-\lambda x} e^{-\lambda(s-x)} dx = \lambda^2 e^{-\lambda s} \int_0^s dx = \lambda^2 s e^{-\lambda s}$$

für $s \geq 0$, bzw. $(f \star f)(s) = 0$ für $s < 0$. Durch Induktion ergibt sich allgemein:

Lemma 9.6. *Die Verteilung von S_n ist absolutstetig mit Dichte*

$$f_{\lambda,n}(s) = \frac{\lambda^n}{\Gamma(n)} \cdot s^{n-1} \cdot e^{-\lambda s} \cdot I_{(0,\infty)}(s) \ ,$$

wobei

$$\Gamma(n) \ := \ \int_0^\infty t^{n-1}\, e^{-t}\, dx \ \stackrel{n\in\mathbb{N}}{=} \ (n-1)! \quad .$$

Definition. *Die Wahrscheinlichkeitsverteilung auf $\mathbb{R}_+$ mit Dichte $f_{\lambda,n}$ heißt **Gammaverteilung** mit Parametern $\lambda, n \in (0,\infty)$.*

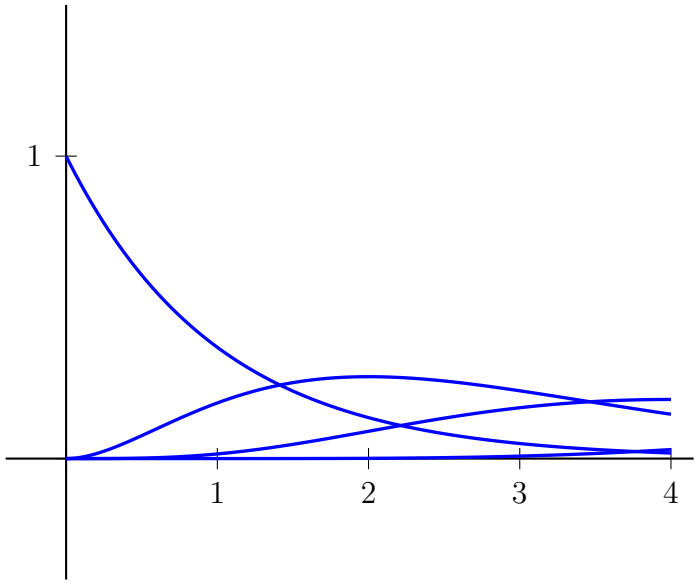


Abbildung 9.2: Dichtefunktionen der Gammaverteilung $\Gamma_{1,n}$ für verschiedene n.

Die Gammaverteilung ist auch für nicht-ganzzahlige n definiert, Γ ist dann die Eulersche Gammafunktion. Für $n = 1$ ergibt sich die Exponentialverteilung als Spezialfall der Gammaverteilung. Allgemein gilt:

$$\Gamma(\lambda, r) \ \star \ \Gamma(\lambda, s) \quad = \quad \Gamma(\lambda, r+s) \ ,$$

d.h. die Gammaverteilungen mit festem Parameter λ bilden eine Faltungshalbgruppe.
Durch Anwenden des zentralen Grenzwertsatzes auf die Zufallsvariable S_n erhalten wir:

Korollar 9.7 (Normalapproximation der Gammaverteilungen). *Sei $\lambda > 0$. Dann gilt für $\Gamma(\lambda, n)$ verteilte Zufallsvariablen S_n:*

$$n^{-1/2} \cdot \left(S_n - n\lambda^{-1}\right) \quad \stackrel{\mathcal{D}}{\to} \quad N(0, \lambda^{-2}) \qquad \textit{für } n \to \infty.$$

Bemerkung (Poissonprozess). Die Anzahl der bis zur Zeit $t \geq 0$ eingetretenen Ereignisse im obigen Modell ist

$$N_t \;=\; \max\{n \geq 0 \mid S_n \leq t\}\;.$$

Die Zufallsvariablen N_t sind Poissonverteilt mit Parameter $\lambda \cdot t$ (Übung). Die Kollektion N_t $(t \geq 0)$ der Zufallsvariablen heißt ***Poissonprozess mit Intensität*** λ. Der Poissonprozess ist ein monoton wachsender stochastischer Prozess mit ganzzahligen Werten. Er ist selbst eine zeitstetige Markovkette und ist von grundlegender Bedeutung für die Konstruktion allgemeiner Markovketten in kontinuierlicher Zeit. Wir werden den Poissonprozess in der Vorlesung „Stochastische Prozesse“ genauer betrachten.

9.3 Transformationen, Gaußmodelle und Parameterschätzung

Der Dichtetransformationssatz

Allgemein gibt es zwei ganz verschiedene Arten, eine Wahrscheinlichkeitsverteilung $\mu(dx)$ zu transformieren:

(1). Koordinatentransformation: $\quad y = \phi(x), \quad \mu(dx) \to \mu \circ \phi^{-1}(dy)$

(2). Maßwechsel durch Dichte: $\quad \mu(dx) \to \varrho(x)\mu(dx)$.

In bestimmten regulären Fällen lassen sich beide Transformationen in Beziehung setzen: Ein Koordinatenwechsel hat denselben Effekt wie eine absolutstetige Maßtransformation mit einer geeigneten Dichte ϱ. Wir demonstrieren dies hier im Fall absolutstetiger Verteilungen im $\mathbb{R}^d$. Die entsprechende Koordinatentransformationsformel verwenden wir dann, um multivariate Normalverteilungen, und verschiedene für die Statistik zentrale Verteilungen zu untersuchen.
Seien $S, T \subseteq \mathbb{R}^n$ offen, und sei $X : \Omega \to S$ eine Zufallsvariable auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ mit absolutstetiger Verteilung μ_X mit Dichte f_X.

Satz 9.8 (Mehrdimensionaler Dichtetransformationssatz). *Ist $\phi : S \to T$ ein Diffeomorphismus (C^1) mit $\det D\phi(x) \neq 0$ für alle $x \in S$, dann ist die Verteilung von $\phi(X)$ absolutstetig mit Dichte*

$$f_{\phi(X)}(y) = f_X(\phi^{-1}(y)) \cdot |\det D\phi^{-1}(y)|,$$

wobei $\det D\phi^{-1}(y) = \det(\frac{\partial x_i}{\partial y_j})$ die Jacobideterminante der Koordinatentransformation ist.

Beweis. Die Behauptung folgt aus dem Transformationssatz der multivariaten Analysis:

$$\begin{aligned} P[\phi(X) \in B] &= P[X \in \phi^{-1}(B)] \\ &= \int_{\phi^{-1}(B)} f_X(x)\, dx \overset{Subst.}{=} \int_B f_X(\phi^{-1}(y)) \cdot |\det D\phi^{-1}(y)|\, dy\ . \end{aligned}$$

□

Beispiel (Sukzessive Wartezeiten). Seien T und $\widetilde{T}$ unabhängige, zum Parameter $\lambda > 0$ exponentialverteilte Zufallsvariablen (z.B. sukzessive Wartezeiten), und sei $S = T + \widetilde{T}$. Nach dem Dichtetransformationssatz gilt dann

$$\begin{aligned} f_{T,S}(t,s) &= f_{T,\widetilde{T}}(t, s-t) \cdot |\det \frac{\partial(t, s-t)}{\partial(t,s)}| \\ &\propto e^{-\lambda t} \cdot I_{(0,\infty)}(t) \cdot e^{-\lambda(s-t)} \cdot I_{(0,\infty)}(s-t) \\ &= e^{-\lambda s} \cdot I_{(0,s)}(t). \end{aligned}$$

Somit ist die bedingte Dichte $f_{S|T}(s|t)$ für festes $t > 0$ proportional zu $e^{-\lambda s} \cdot I_{(t,\infty)}(s)$. Dies ist auch anschaulich sofort plausibel, da s eine um die unabhängige Zufallsvariable T verschobene exponentialverteilte Zufallsvariable ist.
Interessanter ist die Berechnung der bedingten Dichte von T gegeben S: Für festes $s > 0$ ist $f_{T|S}(t|s)$ proportional zu $I_{(0,s)}(t)$, d.h.

$$f_{T|S}(t|s) = \frac{1}{s} \cdot I_{(0,s)}(t).$$

Gegeben die Summe S der beiden Wartezeiten ist die erste Wartezeit T also gleichverteilt auf $[0, S]$!

Wir betrachten nun verschiedene weiterreichende Anwendungen des Dichtetransformationssatzes.

Multivariate Normalverteilungen und multivariater ZGS

Sei $Z = (Z_1, Z_2, ..., Z_d)$ mit unabhängigen, $N(0,1)$-verteilten Zufallsvariablen Z_i. Die Verteilung des Zufallsvektors Z ist dann absolutstetig bzgl. des Lebesguemaßes im $\mathbb{R}^d$ mit Dichte

$$f_Z(x) = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x_i^2}{2}} = (2\pi)^{-\frac{d}{2}} e^{-\frac{|x|^2}{2}} \quad \textbf{(}d\textbf{-dimensionale Standardnormalverteilung)}.$$

Sei nun $m \in \mathbb{R}^d$ und $\sigma \in \mathbb{R}^{d\times d}$ eine $d \times d$-Matrix. Wir betrachten den Zufallsvektor

$$Y = \sigma Z + m\ .$$

Wir zeigen zunächst, dass Y Erwartungswert m und Kovarianzmatrix $C = \sigma\sigma^T$ hat, und berechnen die charakteristische Funktion: **Erwartungswert:** $E[Y_i] = \sum_{k=1}^d \sigma_{ik} E[Z_k] + m_i = m_i$.

Kovarianz: $Cov(Y_i, Y_j) = Cov(\sum_k \sigma_{ik} Z_k + m_i, \sum_l \sigma_{jl} Z_l + m_j)$
$= \sum_{k,l} \sigma_{ik}\sigma_{jl} \cdot Cov(Z_k, Z_l) = \sum_k \sigma_{ik}\sigma_{jk} = C_{ij}.$

Charakteristische Funktion: Für einen Vektor $p \in \mathbb{R}^d$ gilt

$$\begin{aligned} \varphi_Y(p) &:= E\left[e^{ip\cdot Y}\right] = E\left[e^{i(\sigma^T p)\cdot Z}\right] e^{ip\cdot m} = e^{-\frac{1}{2}|\sigma^T p|^2 + ip\cdot m} \\ &= e^{-\frac{1}{2} p\cdot Cp + ip\cdot m}. \end{aligned} \tag{9.3.1}$$

Ist σ *regulär*, dann können wir die Dichte der Verteilung von Y sofort mithilfe des Transformationssatzes explizit berechnen:

$$\begin{aligned} f_Y(y) &= f_X(\sigma^{-1}(y-m)) \cdot |\det \sigma^{-1}| \\ &= \frac{1}{\sqrt{(2\pi)^d |\det C|}} \exp\left(-\frac{1}{2}(y-m)C^{-1}(y-m)\right). \end{aligned}$$

Auch im $\mathbb{R}^d$ ist eine Wahrscheinlichkeitsverteilung durch ihre charakteristische Funktion eindeutig festgelegt, s. z.B. Bauer: Wahrscheinlichkeitstheorie. Allgemein (also auch für nicht reguläre σ) können wir die Verteilung von Y auch über die Fourierinversionsformel berechnen.

Definition. *Sei $m \in \mathbb{R}^d$ und $C \in \mathbb{R}^{d\times d}$ eine symmetrische, nicht-negativ definite Matrix. Die Verteilung $N(m, C)$ im $\mathbb{R}^d$ mit charakteristischer Funktion $\phi_Y = \exp(-\frac{1}{2} p\dot{C}p + ipm)$ heißt d-**dimensionale Normalverteilung** mit Mittel m und Kovarianzmatrix C.*

Bemerkung/Übung. Mithilfe von charakteristischen Funktionen beweist man die folgenden Transformationsformeln und Charakterisierungen für multivariate Normalverteilungen:

(1). Für $a \in \mathbb{R}^k$ und $A \in \mathbb{R}^{k\times d}$ gilt

$$X \sim N(m, C) \Rightarrow AX + a \sim N(Am + a, ACA^T).$$

(2). Folgende Aussagen sind äquivalent:

- $X \sim N(0, C)$ ist multivariat normalverteilt mit Kovarianzmatrix C.
- $p \cdot X \sim N(0, p \cdot Cp) \qquad \forall\, p \in \mathbb{R}^d$.

Auch im $\mathbb{R}^d$ gilt ein zentraler Grenzwertsatz :

Satz 9.9 (Multivariater zentraler Grenzwertsatz). *Seien $X_1, X_2, ... : \Omega \to \mathbb{R}^d$ unabhängige, identisch verteilte, quadratintegrierbare Zufallsvektoren auf $(\Omega, \mathcal{A}, P)$, und sei $S_n = X_1 + \ldots + X_n$. Dann gilt*

$$\frac{S_n - E[S_n]}{\sqrt{n}} \xrightarrow{\mathcal{D}} Z \sim N(0, C),$$

wobei $C_{jk} = Cov(X_{1,j}, X_{1,k})$ die Kovarianzmatrix der Zufallsvektoren X_i ist.

Der Beweis basiert auf folgender Charakterisierung der schwachen Konvergenz von Zufallsvektoren:

Lemma 9.10 (Cramér-Wold Device). *Für Zufallsvariablen $Y, Y_1, Y_2, ... : \Omega \to \mathbb{R}^d$ gilt:*

$$Y_n \xrightarrow{\mathcal{D}} Y \quad \Leftrightarrow \quad p \cdot Y_n \xrightarrow{\mathcal{D}} p \cdot Y \qquad \forall\, p \in \mathbb{R}^d.$$

Beweisskizze. Die Richtung „⇒“ ist klar, da $Y \mapsto p \cdot Y$ stetig ist. Umgekehrt gilt:

$$p \cdot Y_n \xrightarrow{\mathcal{D}} p \cdot Y \;\Rightarrow\; E[\exp(ip \cdot Y_n)] \;\to\; E[\exp(ip \cdot Y)] \quad \forall\, p \in \mathbb{R}^d.$$

Mit einem ähnlichen Beweis wie im $\mathbb{R}^1$ folgt dann aus der Konvergenz der charakteristischen Funktionen die schwache Konvergenz $Y_n \xrightarrow{\mathcal{D}} Y$. Um die relative Kompaktheit zu zeigen (Satz von Helly-Bray), verwendet man dabei im $\mathbb{R}^d$ die multivariaten Verteilungsfunktionen

$$F_n(x_1, ..., x_d) := P[Y_{n,1} \le x_1, ..., Y_{n,d} \le x_d], \qquad (x_1, \ldots, x_d) \in \mathbb{R}^d.$$

Wir beweisen nun den zentralen Grenzwertsatz:

Beweis. Für $p \in \mathbb{R}^d$ gilt nach dem eindimensionalen zentralen Grenzwertsatz:

$$\begin{aligned} p \cdot \left(\frac{S_n - E[S_n]}{\sqrt{n}} \right) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (p \cdot X_i - E[p \cdot X_i]) \\ &\xrightarrow{\mathcal{D}} N\left(0, \mathrm{Var}[p \cdot X_1]\right) = N(0, p \cdot Cp), \end{aligned}$$

da

$$\mathrm{Var}[p \cdot X_1] = \mathrm{Cov}\left[\sum_k p_k X_{1,k} \,,\, \sum_l p_k X_{1,l} \right] = \sum_{k,l} p_k p_l C_{kl} = p \cdot Cp.$$

Ist Y ein $N(0, C)$-verteilter Zufallsvektor, dann ist $N(0, p \cdot Cp)$ die Verteilung von $p \cdot Y$. Mithilfe der Cramér-Wold Device folgt also

$$(S_n - E[S_n]) / \sqrt{n} \quad \xrightarrow{\mathcal{D}} \quad Y.$$

□

Beispiel (Vom Random Walk zur Brownschen Bewegung). Sei $S_n = X_1 + ... + X_n$, wobei die X_i unabhängige Zufallsvariablen mit

$$E[X_i] = 0 \text{ und } \operatorname{Var}[X_i] = 1$$

sind. Beispielsweise ist S_n ein klassischer Random Walk. Um einen stochastischen Prozess in kontinuierlicher Zeit zu erhalten, interpolieren wir $n \mapsto S_n$ linear. Anschließend reskalieren wir in Raum und Zeit, und setzen

$$\widetilde{S}_t^{(n)} := \frac{1}{\sqrt{n}} S_{nt}, \qquad t \in \mathbb{R}_+.$$

GRAPHIK SKALIERTER RANDOM WALK
Aus dem Zentralen Grenzwertsatz folgt:

$$\widetilde{S}_t^{(n)} = \sqrt{t}\,\frac{1}{\sqrt{nt}}\,S_{nt} \xrightarrow{\mathcal{D}} \sim N(0,t) \qquad \text{für jedes feste } t \in \mathbb{R}_+,$$

d.h. die eindimensionalen Randverteilungen der Prozesse $\widetilde{S}^{(n)} = (\widetilde{S}_t^{(n)})_{t\geq 0}$ konvergieren. Allgemeiner zeigt man mithilfe des multivariaten zentralen Grenzwertsatzes, dass auch endlich dimensionale Randverteilungen schwach konvergieren:

$$\left(\widetilde{S}_{t_1}^{(n)}, \widetilde{S}_{t_2}^{(n)}, ..., \widetilde{S}_{t_k}^{(n)}\right) \xrightarrow{\mathcal{D}} (B_{t_1}, ..., B_{t_k}), \qquad \text{für alle } 0 \leq t_1 < t_2 < \ldots < t_k, k \in \mathbb{N},$$

wobei $(B_{t_1}, ..., B_{t_k})$ multivariat normalverteilt ist mit

$$E[B_{t_j}] = 0 \quad \text{und} \quad \operatorname{Cov}[B_{t_j}\,,\, B_{t_k}] = \min(t_j, t_k).$$

Eine noch allgemeinere Aussage erhält man mithilfe eines **funktionalen zentralen Grenzwertsatzes** (Invarianzprinzip von Donsker, ZGS auf dem Banachraum $C([0,1],\mathbb{R})$): Der gesamte stochastische Prozess $(\widetilde{S}_t^{(n)})_{0\leq t\leq 1}$ konvergiert in Verteilung gegen eine **Brownsche Bewegung** $(B_t)_{0\leq t\leq 1}$. Mehr dazu in den weiterführenden Vorlesungen »Stochastische Prozesse« und »Grundzüge der stochastischen Analysis«.

Wir betrachten noch eine weitere Anwendung des Dichtetransformationssatzes auf Normalverteilungen.

Beispiel (χ^2-Verteilungen). Wir berechnen nun die Verteilung vom Quadrat des Abstandes vom Ursprung eines standardnormalverteilten Zufallsvektors im $\mathbb{R}^d$:

$$Z \;=\; (Z_1, ..., Z_d) \;\sim\; N(0, I_d), \quad \|Z\|^2 \;=\; \sum_{i=1}^{d} Z_i^2.$$

Wegen $f_{|Z_i|}(x) = \frac{2}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}} \cdot I_{(0,\infty)}(x)$ folgt durch Anwenden des Dichtetransformationssatzes mit $Y = \phi(x) := x^2$:

$$f_{Z_i^2}(y) = \sqrt{\frac{2}{\pi}}\, e^{-\frac{y}{2}} \cdot I_{(0,\infty)}(y) \cdot \frac{1}{2\sqrt{y}}\,,$$

d.h. Z_i^2 ist $\Gamma(\frac{1}{2}, \frac{1}{2})$-verteilt. Da die Zufallsvariablen $Z_i^2,\ 1 \le i \le d$, unabhängig sind, folgt:

$$||Z||^2 = \sum_{i=1}^{d} Z_i^2 \ \sim\ \Gamma\left(\frac{1}{2}, \frac{d}{2}\right)\,.$$

Definition. *Die Gamma-Verteilung mit Parametern $\frac{1}{2}$ und $\frac{d}{2}$ heißt auch **Chiquadrat-Verteilung** $\chi^2(d)$ **mit** d **Freiheitsgraden**.*

Parameterschätzung im Gaußmodell

Angenommen, wir beobachten reellwertige Messwerte (Stichproben, Daten), die von einer unbekannten Wahrscheinlichkeitsverteilung μ auf $\mathbb{R}$ stammen. Ziel der Statistik ist es, Rückschlüsse auf die zugrundeliegende Verteilung aus den Daten zu erhalten. Im einfachsten Modell (Gaußmodell) nimmt man an, dass die Daten unabhängige Stichproben von einer Normalverteilung mit unbekanntem Mittelwert und/oder Varianz sind:

$$\mu = N(m, v), \quad m, v \text{ unbekannt.}$$

Eine partielle Rechtfertigung für die Normalverteilungsannahme liefert der zentrale Grenzwertsatz. Letztendlich muss man aber in jedem Fall überprüfen, ob eine solche Annahme gerechtfertigt ist.Ein erstes Ziel ist es nun, den Wert von m auf der Basis von n unabhängigen Stichproben $X_1(\omega) = x_1, \ldots, X_n(\omega) = x_n$ zu schätzen, und zu quantifizieren.

Problemstellung: Schätzung des Erwartungswerts

- Schätze m auf der Basis von n unabhängigen Stichproben $X_1(\omega), ..., X_n(\omega)$ von μ.
- Herleitung von Konfidenzintervallen.

Im mathematischen Modell interpretieren wir die Beobachtungswerte als Realisierungen von unabhängigen Zufallsvariablen $X_1, \ldots, X_n$. Da wir die tatsächliche Verteilung nicht kennen, untersuchen wir alle in Betracht gezogenen Verteilungen simultan:

$$X_1, \ldots, X_n \sim N(m, v) \quad \text{unabhängig unter } P_{m,v}. \tag{9.3.2}$$

Ein naheliegender Schätzer für m ist der *empirische Mittelwert*

$$\overline{X}_n(\omega) := \frac{X_1(\omega) + ... + X_n(\omega)}{n} .$$

Wir haben oben bereits gezeigt, dass dieser Schätzer *erwartungstreu (unbiassed)* und *konsistent* ist, d.h. für alle m, v gilt:

$$E_{m,v}[\overline{X}_n] = m$$

und

$$\overline{X}_n \to m \quad P_{m,v}\text{-stochastisch für } n \to \infty.$$

Wie wir den Schätzfehler quantifizieren hängt davon ab, ob wir die Varianz kennen.

Schätzung von m bei bekannter Varianz v.
Um den Schätzfehler zu kontrollieren, berechnen wir die Verteilung von $\overline{X}_n$:

$$\begin{aligned} X_i \sim N(m,v) \text{ unabh.} \quad &\Rightarrow \quad X_1 + ... + X_n \sim N(nm, nv) \\ &\Rightarrow \quad \overline{X}_n \sim N(m, \frac{v}{n}) \\ &\Rightarrow \quad \frac{\overline{X}_n - m}{\sqrt{v/n}} \sim N(0,1) \end{aligned}$$

Bezeichnet Φ die Verteilungsfunktion der Standardnormalverteilung, dann erhalten wir

$$P_{m,v}\left[|\overline{X}_n - m| < q\sqrt{\frac{v}{n}}\right] = N(0,1)(-q,q) = 2\left(\Phi(q) - \frac{1}{2}\right) \quad \text{für alle } m \in \mathbb{R}.$$

Satz 9.11. *Im Gaußmodell (9.3.2) mit bekannter Varianz v ist das zufällige Intervall*

$$\left(\overline{X}_n - \Phi^{-1}(\alpha)\sqrt{\frac{v}{n}}\,,\ \overline{X}_n + \Phi^{-1}(\alpha)\sqrt{\frac{v}{n}}\right)$$

*ein $(2\alpha - 1) \cdot 100\%$ **Konfidenzintervall** für m, d.h.*

$$P_{m,v}[m \in \textit{Intervall}] \geq 2\alpha - 1 \qquad \textit{für alle } m \in \mathbb{R}.$$

Man beachte, dass die Länge des Konfidenzintervalls in diesem Fall nicht von den beobachteten Stichproben abhängt!

Schätzung von m bei unbekannter Varianz v. In Anwendungen ist meistens die Varianz unbekannt. In diesem Fall können wir das Intervall oben nicht verwenden, da es von der unbekannten Varianz v abhängt. Stattdessen schätzen wir m und v simultan, und konstruieren ein

Konfidenzintervall für m mithilfe beider Schätzwerte. Erwartungstreue Schätzer für m und v sind

$$\overline{X}_n \;=\; \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{und} \quad V_n = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \overline{X}_n)^2 \; .$$

Um ein Konfidenzintervall für m zu erhalten, bestimmen wir mithilfe des Transformationssatzes die gemeinsame Verteilung von $\overline{X}_n$ und V_n:

Lemma 9.12. *$\overline{X}_n$ und V_n sind unabhängig unter $P_{m,v}$ mit Verteilung*

$$\overline{X}_n \;\sim\; N\left(m, \frac{v}{n}\right) \quad , \quad \frac{n-1}{v} V_n \sim \chi^2(n-1) \; .$$

Beweis. Wir führen eine lineare Koordinatentransformation $Y = OX$ durch, wobei O eine orthogonale $n \times n$-Matrix vom Typ

$$O = \begin{pmatrix} \frac{1}{\sqrt{n}} \cdots \frac{1}{\sqrt{n}} \\ \text{beliebig} \end{pmatrix}$$

ist. Eine solche Matrix erhalten wir durch Ergänzen des normierten Vektors $(\frac{1}{\sqrt{n}}, ..., \frac{1}{\sqrt{n}})$ zu einer Orthonormalbasis des $\mathbb{R}^n$. In den neuen Koordinaten gilt:

$$\begin{aligned} \overline{X}_n \qquad &= \qquad \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{\sqrt{n}} Y_1, \quad \text{und} \\ (n-1)V_n \qquad &= \qquad \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}_n^2 = ||X||_{\mathbb{R}^n}^2 - n\overline{X}_n^2 \\ &\overset{O \text{ orthogonal}}{=} \quad ||Y||_{\mathbb{R}^n}^2 - Y_1^2 = \sum_{i=2}^{n} Y_i^2 \; . \end{aligned}$$

Da die Zufallsvariablen X_i $(1 \le i \le n)$ unabhängig und $N(m, v)$-verteilt sind, ist der Zufallsvektor $X = (X_1, ..., X_n)$ multivariat normalverteilt mit Mittel $(m, \ldots, m)$ und Kovarianzmatrix $v \cdot I_n$. Nach dem Transformationssatz folgt

$$Y \sim N\left(O \begin{pmatrix} m \\ \vdots \\ m \end{pmatrix}, v \cdot O I_n O^T\right) = N\left(\begin{pmatrix} m\sqrt{n} \\ 0 \\ \vdots \\ 0 \end{pmatrix}, v \cdot I_n\right) .$$

Also sind $Y_1, ..., Y_n$ unabhängige Zufallsvariablen mit Verteilungen

$$Y_1 \sim N(m\sqrt{n}, v) \quad , \quad Y_i \sim N(0, v) \quad \text{für } i \ge 2.$$

Es folgt, dass

$$\overline{X}_n = \frac{Y_1}{\sqrt{n}} \quad \text{und} \quad \frac{n-1}{v} V_n = \sum_{i=2}^{n} \left(\frac{Y_i}{\sqrt{v}}\right)^2$$

unabhängige Zufallsvariablen mit Verteilungen $N(m, \frac{v}{n})$ bzw. $\chi^2(n-1)$ sind. □

Bei bekannter Varianz v hatten wir Konfidenzintervalle für m vom Typ $\overline{X}_n \pm q \cdot \sqrt{\frac{v}{n}}$ erhalten, wobei q ein geeignetes Quantil der Standardnormalverteilung ist. Daher liegt es nahe, zu versuchen, bei unbekannter Varianz Konfidenzintervalle vom Typ $\overline{X}_n \pm q \cdot \sqrt{\frac{V_n}{n}}$ herzuleiten. Es gilt:

$$P_{m,v}\left[|\overline{X}_n - m| \geq q\sqrt{\frac{V_n}{n}}\right] = P_{m,v}[|T_{n-1}| \geq q] \qquad \text{mit}$$

$$T_{n-1} := \frac{\sqrt{n} \cdot (\overline{X}_n - m)}{\sqrt{V_n}} \, .$$

Die Zufallsvariable T_{n-1} heißt **Studentsche t-Statistik mit $n-1$ Freiheitsgraden**.[1] Unsere Überlegungen zeigen, dass wir aus Quantilen der Studentschen t-Statistik Konfidenzintervalle für das Gaußmodell herleiten können. Wir müssen nur noch die Verteilung von T_n berechnen:

Satz 9.13 (Student[2])**.** *Die Verteilung von T_n ist absolutstetig mit Dichte*

$$f_{T_n}(t) \quad = \quad B\left(\frac{1}{2}, \frac{n}{2}\right)^{-1} \cdot n^{-1/2} \cdot \left(1 + \frac{t^2}{2}\right)^{-n/2} \qquad (t \in \mathbb{R}).$$

»Studentsche t-Verteilung mit n Freiheitsgraden«*. Hierbei ist*

$$B\left(\frac{1}{2}, \frac{n}{2}\right) = \frac{1}{\sqrt{n}} \int_{-\infty}^{\infty} (1 + s^2)^{-\frac{n}{2}} ds$$

*die **Eulersche Beta-Funktion**, die als Normierungsfaktor auftritt.*
Insbesondere ist das zufällige Intervall

$$\overline{X}_n \pm q \cdot \sqrt{\frac{V_n}{n}}$$

ein $100 \cdot (1-2\alpha)\%$ Konfidenzintervall für m, falls

$$q = F^{-1}_{T_{n-1}}(1-\alpha)$$

ein $(1-\alpha)$-Quantil der t-Verteilung mit $n-1$ Freiheitsgraden ist.

Beweis. Direkt oder mithilfe des Transformationssatzes zeigt man: Sind Z und Y unabhängige Zufallsvariablen mit Verteilungen $N(0,1)$ bzw. $\chi^2(n-1)$, dann ist $Z/\sqrt{\frac{1}{n-1}Y}$ absolutstetig mit dichte $f_{T_{n-1}}$.

[1] In der Statistik bezeichnet man eine messbare Funktion der Beobachtungsdaten als Statistik - ein (Punkt-) Schätzer ist eine Statistik, die zum Schätzen eines unbekannten Parameters verwendet wird, ein Konfidenzintervall nennt man auch Intervallschätzer.

[2] Synonym von W. S. Gosset, der als Angestellter der Guiness-Brauerei nicht publizieren durfte.

Der Satz folgt dann nach Lemma 9.12 mit

$$Z := \frac{\overline{X}_n - m}{\sqrt{v/n}} \quad \text{und} \quad Y := \frac{n-1}{v} V_n \,.$$

□

Bemerkung (Nichtparametrische und Verteilungsunabhängige Konfidenzintervalle). In Anwendungen ist es oft unklar, ob eine Normalverteilungsannahme an die Beobachtungswerte gerechtfertigt ist. Zudem können einzelne größere Ausreißer in den Daten (z.B. aufgrund von Messfehlern) das Stichprobenmittel relativ stark beeinflussen. Der Stichprobenmedian ist dagegen in den meisten Fällen ein deutlich stabilerer Schätzwert für den Median der zugrundeliegenden Verteilung, und die in Abschnitt 5.1 hergeleiteten, auf Ordnungsstatistiken basierenden, Konfidenzintervalle für den Median und andere Quantile werden ebenfalls in der Regel weniger stark durch Ausreißer beeinflusst. Zudem gelten diese Konfidenzintervalle simultan für alle stetigen Verteilungen. Ist man sich daher nicht sicher, ob eine Normalverteilungsannahme aufgrund der Daten gerechtfertigt ist, empfiehlt es sich, auf die stabileren Ordnungsintervalle zurückzugreifen.

Beispiel. (NOCH EINZUFÜGEN)

Hypothesentests

In Anwendungen werden statistische Aussagen häufig nicht über Konfidenzintervalle, sondern als Hypothesentest formuliert. Mathematisch passiert dabei nichts wirklich Neues – es handelt sich nur um eine durch praktische Erwägungen motivierte Umformulierung derselben Resultate: Angenommen, wir haben n unabhängige reellwertige Stichproben $X_1, ..., X_n$ von einer unbekannten Verteilung vorliegen und wir gehen davon aus, daß die zugrundeliegende Verteilung aus einer Familie μ_θ $(\theta \in \Theta)$ von Wahrscheinlichkeitsverteilungen kommt, z.B. der Familie aller Normalverteilungen $\mu_{m,v}, \theta = (m, v) \in \mathbb{R} \times \mathbb{R}_+$. Die gemeinsame Verteilung von $X_1, \ldots, X_n$ ist dann das Produktmaß $\mu_\theta^n = \bigotimes_{i=1}^n \mu_\theta$. Sei nun Θ_0 eine Teilmenge des Parameterbereichs. Wir wollen entscheiden zwischen der

$$\textit{Nullhypothese } H_0\text{:} \qquad \text{»}\theta \in \Theta_0\text{«}$$

und der

$$\textit{Alternative } H_1\text{:} \qquad \text{»}\theta \notin \Theta_0\text{«}$$

Ein **Hypothesentest** für ein solches Problem ist bestimmt durch eine messbare Teilmenge $C \subseteq \mathbb{R}^n$ (den **Verwerfungsbereich**) mit zugehöriger Entscheidungsregel:

$$\text{Akzeptiere } H_0 \iff (X_1, ..., X_n) \notin C.$$

Beispiel (t-Test). Seien $X_1, X_2, \ldots, X_n$ unabhängige Stichproben von einer Normalverteilung mit unbekanntem Parameter $(m, v) \in \Theta = \mathbb{R} \times \mathbb{R}^+$. Wir wollen testen, ob der Mittelwert der Verteilung einen bestimmten Wert m_0 hat:

$$\textit{Nullhypothese } H_0\text{:} \qquad »m = m_0« \,, \qquad \Theta_0 = \{m_0\} \times \mathbb{R}^+ \,.$$

Ein solches Problem tritt z.B. in der Qualitätskontrolle auf, wenn man überprüfen möchte, ob ein Sollwert m_0 angenommen wird. Eine andere Anwendung ist der Vergleich zweier Verfahren, wobei X_i die Differenz der mit beiden Verfahren erhaltenen Messwerte ist. Die Nullhypothese mit $m_0 = 0$ besagt hier, daß kein signifikanter Unterschied zwischen den Verfahren besteht.

Im *t–Test* für obiges Testproblem wird die Nullhypothese akzeptiert, falls der Betrag der *Student-schen t-Statistik* unterhalb einer angemessen zu wählenden Konstanten c liegt, bzw. verworfen, falls

$$|T_{n-1}| \;=\; \left| \frac{\sqrt{n} \cdot (\overline{X}_n - m_0)}{\sqrt{V_n}} \right| \;>\; c$$

gilt.

Seien nun allgemein $X_1, X_2, \ldots$ unter P_θ unabhängige Zufallsvariablen mit Verteilung μ_θ. Bei einem Hypothesentest können zwei Arten von Fehlern auftreten:

Fehler 1. Art: *H_0 wird verworfen, obwohl wahr.* Die Wahrscheinlichkeit dafür beträgt:

$$P_\theta[(X_1, ..., X_n) \in C] = \mu_\theta^n[C] \,, \qquad \theta \in \Theta_0.$$

Fehler 2. Art: *H_0 wird akzeptiert, obwohl falsch.* Die Wahrscheinlichkeit beträgt:

$$P_\theta[(X_1, ..., X_n) \notin C] = \mu_\theta^n[C^C] \,, \qquad \theta \in \Theta \setminus \Theta_0.$$

Obwohl das allgemeine Testproblem im Prinzip symmetrisch in H_0 und H_1 ist, interpretiert man beide Fehler i.a. unterschiedlich. Die Nullhypothese beschreibt in der Regel den Normalfall, die Alternative eine Abweichung oder einen zu beobachtenden Effekt. Da ein Test Kritiker überzeugen soll, sollte die Wahrscheinlichkeit für den Fehler 1. Art (Effekt prognostiziert, obgleich nicht vorhanden) unterhalb einer vorgegebenen (kleinen) Schranke α liegen. Die Wahrscheinlichkeit

$$\mu_\theta^n[C] \,, \qquad \theta \in \Theta \setminus \Theta_0 \,,$$

daß kein Fehler 2. Art auftritt, sollte unter dieser Voraussetzung möglichst groß sein.

Definition. *Die Funktion*

$$G(\theta) \;=\; P_\theta[(X_1, ..., X_n) \in C] \;=\; \mu_\theta^n[C]$$

heißt ***Gütefunktion*** *des Tests. Der Test hat* ***Niveau*** α*, falls*

$$G(\theta) \;\leq\; \alpha \quad \text{für alle } \theta \in \Theta_0$$

gilt. Die Funktion $G(\theta)$ mit $\theta \in \Theta_1$ heißt ***Macht*** *des Tests.*

Aus Satz 9.13 und der Symmetrie der Studentschen t-Verteilung folgt unmittelbar:

Korollar 9.14. *Der Studentsche t-Test hat Niveau α falls c ein $(1 - \frac{\alpha}{2})$-Quantil der Studentschen t-Verteilung mit $n - 1$ Freiheitsgraden ist.*

Allgemeiner gilt:

Satz 9.15 (Korrespondenz Konfidenzintervalle $\leftrightarrow$ Hypothesentests). *Für einen reellwertigen Parameter $\gamma = c(\theta)$, ein Irrtumsniveau $\alpha \in (0, 1)$, und messbare Abbildungen (Statistiken) $\hat{\gamma}, \varepsilon : \mathbb{R}^n \to \mathbb{R}$ sind äquivalent:*

(i) Das Intervall

$$[\hat{\gamma}(X_1, \ldots, X_n) - \varepsilon(X_1, \ldots, X_n)\,,\; \hat{\gamma}(X_1, \ldots, X_n) + \varepsilon(X_1, \ldots, X_n)]$$

ist ein $(1 - \alpha) \cdot 100\,\%$ Konfidenzintervall für γ.

(ii) Für jedes $\gamma_0 \in \mathbb{R}$ ist

$$C \;=\; \{(x_1, ..., x_n) \,:\, |\hat{\gamma}(x_1, \ldots, x_n) - \gamma_0| > \varepsilon(x_1, \ldots, x_n)\}$$

der Verwerfungsbereich eines Test der Nullhypothese $\gamma = \gamma_0$ zum Niveau α.

Beweis. Das Intervall ist genau dann ein Konfidenzintervall für γ zum Irrtumsniveau α, wenn

$$P_\theta\;[|\hat{\gamma}(X_1, \ldots, X_n) - c(\theta)| > \varepsilon(X_1, ..., X_n)] \;\leq\; \alpha \qquad \forall\, \theta \in \Theta$$

gilt, also wenn der entsprechende Test der Nullhypothesen $c(\theta) = \gamma_0$ für jedes γ_0 Niveau α hat. □

Kapitel 10

Bedingte Erwartungen

Zur Analyse von stochastischen Modellen mit Abhängigkeiten verwendet man in der Regel bedingte Wahrscheinlichkeiten und Erwartungswerte gegeben die Werte von Zufallsvariablen. Beispielsweise beschreibt man einen stochastischen Prozess X_n , $n \in \mathbb{N}$, durch die bedingten Verteilungen des nächsten Zustands X_{n+1} gegeben den Verlauf $X_{0:n} = (X_0, X_1, \ldots, X_n)$ bis zur Zeit n.

10.1 Bedingen auf diskrete Zufallsvariablen

Wir betrachten zunächst das Bedingen auf den Ausgang einer diskreten Zufallsvariable $Y : \Omega \to S$, S abzählbar. In diesem Fall können wir die *bedingte Wahrscheinlichkeitsverteilung*

$$P[A \mid Y = z] \quad = \quad \frac{P[A \cap \{Y = z\}]}{P[Y = z]}, \qquad A \in \mathcal{A},$$

und die *bedingten Erwartungswerte*

$$E[X \mid Y = z] \quad = \quad \frac{E[X; Y = z]}{P[Y = z]}, \qquad X \in \mathcal{L}^1(\Omega, \mathcal{A}, P),$$

für alle $z \in S$ mit $P[Y = z] > 0$ auf elementare Weise wie in Abschnitt 2.1 definieren. Für $z \in S$ mit $P[Y = z] = 0$ sind die bedingten Wahrscheinlichkeiten nicht definiert.

Bedingte Erwartungen als Zufallsvariablen

Es wird sich als praktisch erweisen, die bedingten Wahrscheinlichkeiten und Erwartungswerte nicht als Funktion des Ausgangs z, sondern als Funktion der Zufallsvariable Y zu interpretieren. Die bedingten Wahrscheinlichkeiten und Erwartungswerte sind dann selbst Zufallsvariablen:

Definition. *Sei $X : \Omega \to \mathbb{R}$ eine Zufallsvariable mit $E[X^-] < \infty$, und $Y : \Omega \to S$ eine diskrete Zufallsvariable. Die durch*

$$E[X \mid Y] \quad := \quad g(Y) \quad = \quad \sum_{z \in S} g(z) \cdot I_{\{Y=z\}}$$

mit

$$g(z) \quad := \quad \begin{cases} E[X \mid Y = z] & \text{falls } P[Y = z] > 0 \\ \text{beliebig} & \text{falls } P[Y = z] = 0 \end{cases}$$

*P-fast sicher eindeutig definierte Zufallsvariable $E[X \mid Y]$ heißt **(Version der) bedingte(n) Erwartung von X gegeben Y**. Für ein Ereignis $A \in \mathcal{A}$ heißt die Zufallsvariable*

$$P[A \mid Y] \quad := \quad E[I_A \mid Y]$$

(Version der) bedingte(n) Wahrscheinlichkeit von A gegeben Y.

Die bedingte Erwartung $E[X \mid Y]$ und die bedingte Wahrscheinlichkeit $P[A \mid Y]$ sind also Zufallsvariablen mit den Werten $E[X \mid Y = z]$ bzw. $P[A \mid Y = z]$ auf den Mengen $\{Y = z\}, z \in S$ mit $P[Y = z] > 0$. Auf jeder der Nullmengen $\{Y = z\}, z \in S$ mit $P[Y = z] = 0$, wird der bedingten Erwartung ein willkürlicher konstanter Wert zugewiesen, d.h. die Definition ist nur P-fast überall eindeutig. Wir fassen zunächst einige elementare Eigenschaften der so definierten bedingten Erwartung zusammen:

Lemma 10.1 (Eigenschaften der bedingten Erwartung).

(1). Die Abbildung $X \mapsto E[X \mid Y]$ ist P-fast sicher linear und monoton.

(2). Sind X und Y unabhängig, dann gilt $E[X \mid Y] = E[X]$ P-fast sicher.

(3). Herausziehen, was bekannt ist:
Für alle $f : S \to \mathbb{R}$ mit $f(Y) \cdot X \geq 0$ bzw. $f(Y) \cdot X \in \mathcal{L}^1$ gilt

$$E[f(Y) \cdot X \mid Y] \; = \; f(Y) \cdot E[X \mid Y] \qquad P\text{-fast sicher}.$$

Insbesondere gilt

$$E[f(Y) \mid Y] = f(Y) \qquad P\text{-fast sicher}.$$

Beweis. (2). Sind X und Y unabhängig, dann gilt

$$E[X \mid Y = z] \quad = \quad \frac{E[X \cdot I_{\{Y=z\}}]}{P[Y = z]} \quad = \quad E[X]$$

für alle $z \in S$ mit $P[Y = z] > 0$, also $E[X \mid Y] = E[X]$ P-fast sicher. Die ebenso elementaren Beweise von (1) und (3) werden dem Leser als Übung überlassen.

□

Anschaulich können wir die zweite Aussage folgendermaßen interpretieren: Sind X und Y unabhängig, dann liefert die Kenntnis des Wertes $Y(\omega)$ keine zusätzlichen Informationen über $X(\omega)$. Daher ist die beste L^2-Prognose für $X(\omega)$ wie im unbedingten Fall durch den Erwartungswert $E[X]$ gegeben.

Formel von der totalen Wahrscheinlichkeit

Die aus Satz 2.1 bekannte Formel von der totalen Wahrscheinlichkeit können wir mithilfe der obigen Definition in kompakter Weise schreiben.

Satz 10.2 (Formel von der totalen Wahrscheinlichkeit). *Sei $Y : \Omega \to S$ eine diskrete Zufallsvariable mit Verteilung $\mu(z) = P[Y = z]$. Für alle messbaren $X : \Omega \to \mathbb{R}_+$ gilt:*

$$E[X] = \sum_{z:\ \mu(z)\neq 0} E[X \mid Y = z]\, \mu(z) = E[E[X \mid Y]]$$

Insbesondere gilt

$$P[A] = E[P[A \mid Y]] \qquad \textit{für alle } A \in \mathcal{A}.$$

Beweis. Wegen $\Omega = \dot{\bigcup_{z\in S}} \{Y = z\}$ gilt nach dem Transformationssatz

$$\begin{aligned} E[X] &= \sum_{z\in S} E[X; Y = z] \quad = \quad \sum_{z:\ \mu(z)\neq 0} E[X; Y = z] \\ &= \sum_{z:\ \mu(z)\neq 0} E[X \mid Y = z] \cdot \mu(z) \quad = \quad \sum_{z:\ \mu(z)\neq 0} g(z) \cdot \mu(z) \\ &= E[g(Y)], \end{aligned}$$

wobei $g : S \to \mathbb{R}$ eine beliebige Funktion mit $g(z) = E[X \mid Y = z]$ für alle $z \in S$ mit $\mu(z) \neq 0$ ist. Die Aussage folgt wegen $g(Y) = E[X \mid Y]$ P-fast sicher. □

Bemerkung. Für $X \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ folgt aus der Monotonie der bedingten Erwartung

$$|E[X \mid Y]| \quad \leq \quad E[|X| \,\big|\, Y]$$

und damit die Ungleichung

$$E[|E[X \mid Y]|] \quad \leq \quad E\left[E[|X| \,\big|\, Y]\right] \quad = \quad E[|X|].$$

Die Abbildung $X \mapsto E[X \mid Y]$ ist also eine Kontraktion auf $L^1(\Omega, \mathcal{A}, P)$. Die Aussage von Satz 10.2 gilt entsprechend auch für $X \in \mathcal{L}^1$.

Bedingte Varianz

Sei nun $X : \Omega \to \overline{\mathbb{R}}$ eine bzgl. P integrierbare Zufallsvariable

Definition.

$$\mathrm{Var}[X \mid Y] \quad := \quad E\left[(X - E[X \mid Y])^2 \mid Y\right]$$

heißt ***bedingte Varianz*** *von X gegeben Y.*

Ist X quadratintegrierbar, dann gelten die folgenden Aussagen:

Lemma 10.3. *Für $X \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$ gilt:*

(1). ***L^2-Kontraktivität:*** $E\left[\left|E[X \mid Y]\right|^2\right] \leq E[X^2]$.

(2). $\mathrm{Var}[X \mid Y] = E[X^2 \mid Y] - E[X \mid Y]^2$ *P-fast sicher.*
Insbesondere folgt für $z \in S$ mit $\mu(z) \neq 0$:

$$\mathrm{Var}[X \mid Y] \quad = \quad \mathrm{Var}[X \mid Y = z] \qquad \text{auf } \{Y = z\}. \tag{10.1.1}$$

Beweis. (1). folgt aus Satz 10.2, da für alle $z \in S$ mit $P[Y = z] \neq 0$ auf $\{Y = z\}$ gilt:

$$|E[X \mid Y]|^2 \quad = \quad |E[X \mid Y = z]|^2 \quad \leq \quad E[X^2 \mid Y = z] \quad = \quad E[X^2 \mid Y].$$

(2). Nach Lemma 10.1, (1) und (3), ergibt sich dann ähnlich wie für die unbedingte Varianz:

$$\begin{aligned} \mathrm{Var}[X \mid Y] &= E[X^2 \mid Y] - 2 \cdot E[X \cdot E[X \mid Y] \mid Y] + E[E[X \mid Y]^2 \mid Y] \\ &= E[X^2 \mid Y] - E[X \mid Y]^2 \qquad P\text{-fast sicher.} \end{aligned}$$

□

Die folgende Zerlegungsformel kann häufig verwendet werden, um Varianzen zu berechnen oder abzuschätzen:

Satz 10.4 (Formel von der bedingten Varianz). *Für eine Zufallsvariable $X \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$ gilt:*

$$\begin{aligned} \mathrm{Var}[X] &= E[\mathrm{Var}[X \mid Y]] + \mathrm{Var}[E[X \mid Y]] \\ &= \sum_{z:\mu(z)\neq 0} \mathrm{Var}[X \mid Y = z] \cdot \mu(z) + \sum_{z:\mu(z)\neq 0} (E[X \mid Y = z] - E[X])^2 \cdot \mu(z). \end{aligned}$$

Beweis. Es gilt

$$\begin{aligned}\mathrm{Var}[X] &= E[X^2] - E[X]^2 = E[E[X^2 \mid Y]] - E[E[X \mid Y]]^2 \\ &= E[E[X^2 \mid Y]] - E[E[X \mid Y]^2] + E[E[X \mid Y]^2] - E[E[X \mid Y]]^2 \\ &= E[\mathrm{Var}[X \mid Y]] + \mathrm{Var}[E[X \mid Y]].\end{aligned}$$

Der zweite Teil der Behauptung folgt nun aus (10.1.1) und der entsprechenden Eigenschaft für die bedingte Erwartung. □

Anwendung auf zufällige Summen

Als erste Anwendung betrachten wir eine Summe

$$S_N(\omega) \quad := \quad \sum_{i=1}^{N(\omega)} X_i(\omega)$$

von unabhängigen, identisch verteilten Zufallsvariablen $X_i \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ mit zufälliger Anzahl N von Summanden. Hierbei sei $N : \Omega \to \{0, 1, 2, \ldots\}$ eine von den X_i unabhängige Zufallsvariable. Seien $m = E[X_1]$ und $\sigma^2 = \mathrm{Var}[X_1]$. Wir berechnen nun die verschiedenen Kenngrößen der Verteilung von S_N.

Berechnung des Erwartungswertes: Da S_k und N unabhängig sind, erhalten wir

$$E[S_N \mid N = k] \quad = \quad E[S_k \mid N = k] \quad = \quad E[S_k] \quad = \quad k \cdot m \qquad \text{für alle } k \in \mathbb{N},$$

also $E[S_N \mid N] = N \cdot m$, und damit nach Satz 10.2:

$$E[S_N] \quad = \quad E[E[S_N \mid N]] \quad = \quad E[N] \cdot m.$$

Berechnung der Varianz: Erneut folgt wegen der Unabhängigkeit von S_k und N:

$$\mathrm{Var}[S_N \mid N = k] \quad = \quad \mathrm{Var}[S_k \mid N = k] \quad = \quad \mathrm{Var}[S_k] \quad = \quad k \cdot \sigma^2,$$

also $\mathrm{Var}[S_N \mid N] = N \cdot \sigma^2$, und damit nach Satz 10.4:

$$\mathrm{Var}[S_N] \quad = \quad E[\mathrm{Var}[S_N \mid N]] + \mathrm{Var}[E[S_N \mid N]] \quad = \quad E[N] \cdot \sigma^2 + \mathrm{Var}[N] \cdot m^2.$$

Berechnung der momentenerzeugenden Funktion: Für $t \in \mathbb{R}$ gilt

$$\begin{aligned}M_{S_N}(t) &= E\left[e^{tS_N}\right] = E\left[E[e^{tS_N} \mid N]\right] = E\left[\prod_{i=1}^{N} E[e^{tX_i}]\right] \\ &= E\left[E[e^{tX_1}]^N\right] = E\left[M_{X_1}(t)^N\right] = M_N\left(\log M_{X_1}(t)\right).\end{aligned}$$

Mithilfe von M_{S_N} kann man die Momente der zufälligen Summe S_N berechnen:

$$E[S_N^m] \quad = \quad M_{S_N}^{(m)}(0) \qquad \text{für alle } m \in \mathbb{N}.$$

Im Prinzip erhält man die Verteilung von S_N durch Laplace-Inversion, was aber nicht immer praktikabel ist. Nehmen die Zufallsvariablen X_i nur nichtnegative ganzzahlige Werte an, kann man statt der momentenerzeugenden Funktion die erzeugende Funktion verwenden, und daraus die Verteilung berechnen. Wir gehen darauf im folgenden Abschnitt ein.

Charakterisierende Eigenschaften der bedingten Erwartung

Zum Abschluss dieses Abschnitts beweisen´wir eine alternative Charakterisierung der bedingten Erwartung gegeben eine diskrete Zufallsvariable $Y : \Omega \to S$, S abzählbar. Diese Charakterisierung werden wir in Abschnitt 10.3 verwenden, um bedingte Erwartungen für allgemeine Bedingungen zu definieren. Sei $X : \Omega \to \mathbb{R}_+$ eine nichtnegative (bzw. integrierbare) Zufallsvariable auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$.

Satz 10.5. *Eine reellwertige Zufallsvariable $\overline{X} \geq 0$ (bzw. $\overline{X} \in \mathcal{L}^1$) auf $(\Omega, \mathcal{A}, P)$ ist genau dann eine Version der bedingten Erwartung $E[X \mid Y]$, wenn gilt:*

(I) $\overline{X} = g(Y)$ *für eine Funktion $g : S \to \mathbb{R}$, und*

(II) $E\left[\overline{X} \cdot f(Y)\right] \;=\; E[X \cdot f(Y)]$ *für alle nichtnegativen bzw. beschränkten Funktionen $f : S \to \mathbb{R}$.*

Beweis. Ist $\overline{X}$ eine Version von $E[X \mid Y]$, dann gilt (I). Außerdem folgt nach Lemma 10.1 (3) und der Formel von der totalen Wahrscheinlichkeit:

$$E\left[\overline{X} \cdot f(Y)\right] \quad = \quad E\left[E\left[X \,\middle|\, Y\right] \cdot f(Y)\right] \quad = \quad E\left[E\left[X \cdot f(Y) \,\middle|\, Y\right]\right] \quad = \quad E[X \cdot f(Y)]$$

für jede nichtnegative bzw. beschränkte Funktion $f : S \to \mathbb{R}$.

Umgekehrt folgt aus (I), dass $\overline{X} = g(z)$ auf $\{Y = z\}$ gilt. Ist außerdem (II) erfüllt, dann folgt weiter

$$\begin{aligned} g(z) \;&=\; E\left[\overline{X} \mid Y = z\right] \;=\; \frac{E\left[\overline{X} \cdot I_{\{z\}}(Y)\right]}{P[Y = z]} \\ &=\; \frac{E\left[X \cdot I_{\{z\}}(Y)\right]}{P[Y = z]} \;=\; E[X \mid Y = z] \end{aligned}$$

für alle $z \in S$ mit $P[Y = z] > 0$, d.h. $\overline{X} = g(Y)$ ist eine Version der bedingten Erwartung $E[X \mid Y]$. □

In einigen Fällen können die charakterisierenden Eigenschaften direkt überprüft werden, um bedingte Erwartungen zu identifizieren:

Beispiel (Summen austauschbarer Zufallsvariablen). Seien $X_1, X_2, \ldots, X_n \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ integrierbare Zufallsvariablen, deren gemeinsame Verteilung invariant unter Koordinatenpermutationen ist, d.h. $(X_{\pi(1)}, X_{\pi(2)}, \ldots, X_{\pi(n)}) \sim (X_1, X_2, \ldots, X_n)$ für alle $\pi \in \mathcal{S}_n$. Zufallsvariablen mit dieser Eigenschaft heißen **austauschbar** – beispielsweise sind unabhängige identisch verteilte Zufallsvariablen austauschbar. Wir zeigen:

$$E[X_i \mid S_n] \quad = \quad \frac{1}{n} S_n \qquad P\text{-fast sicher für alle } i = 1, \ldots, n,$$

wobei $S_n = X_1 + \ldots + X_n$. Zum Beweis überprüfen wir, dass $\overline{X}_i := \frac{1}{n} S_n$ die Bedingungen (I) und (II) aus Satz 10.5 für $Y = S_n$ erfüllt. (I) ist offensichtlich. Zudem gilt wegen der Austauschbarkeit für jede beschränkte messbare Funktion $f : \mathbb{R} \to \mathbb{R}$:

$$E[X_i \cdot f(S_n)] \quad = \quad E[X_j \cdot f(S_n)] \qquad \text{für alle } i, j = 1, \ldots, n,$$

also

$$E\left[\frac{1}{n} S_n \cdot f(S_n)\right] \quad = \quad \frac{1}{n} \sum_{j=1}^{n} E[X_j \cdot f(S_n)] \quad = \quad E[X_i \cdot f(S_n)]$$

für alle $i = 1, \ldots, n$, d.h. (II) ist auch erfüllt.

10.2 Erzeugende Funktionen, Verzweigungsprozesse, und Erneuerungen

Wir wollen die Methoden aus dem letzten Abschnitt nun verwenden, um Verzweigungs- und Erneuerungsprozesse zu untersuchen. Ein wichtiges Hilfsmittel sind in beiden Fällen erzeugende Funktionen:

Erzeugende Funktionen von ganzzahligen Zufallsvariablen

Sei $X : \Omega \to \{0, 1, 2, \ldots\}$ eine auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ definierte Zufallsvariable mit nichtnegativen *ganzzahligen* Werten.

Definition. *Die durch*

$$G(s) \quad = \quad E[s^X] \quad = \quad \sum_{k=0}^{\infty} P[X = k] s^k, \qquad s \in [-1, 1],$$

definierte Funktion heißt ***erzeugende Funktion*** *der Zufallsvariable X bzw. der Folge $\mu(k) = P[X = k]$ der Gewichte von X.*

Durch Vergleich mit der geometrischen Reihe sieht man, dass der Konvergenzradius der Potenzreihe stets größer oder gleich 1 ist. Also ist die erzeugende Funktion analytisch auf $(-1, 1)$, und es gilt

$$P[X = k] \quad = \quad \frac{G^{(k)}(0)}{k!} \qquad \text{für alle } k = 0, 1, 2, \ldots.$$

Kennen wir also die erzeugende Funktion explizit, dann können wir die Gewichte der Verteilung berechnen.
Durch zweimaliges Ableiten zeigt man zudem, dass G monoton und konvex auf $[0, 1]$ ist. Für $s \in (0, 1]$ gilt nach Definition $G(s) = M(\log s)$. Daher lassen sich aus der erzeugenden Funktion die Momente von X berechnen – beispielsweise gilt $E[X] = G'(1-)$ (linksseitige Ableitung von $G(s)$ bei $s = 1$), falls der Erwartungswert endlich ist.

Für die erzeugende Funktion einer Summe $X + Y$ von unabhängigen, nichtnegativen, ganzzahligen Zufallsvariablen X und Y gilt offensichtlich

$$G_{X+Y}(s) \quad = \quad G_X(s) \cdot G_Y(s) \qquad \text{für alle } s \in [-1, 1].$$

Somit ist die erzeugende Funktion der Faltung

$$(\mu * \nu)(k) \quad = \sum_{i=0}^{k} \mu(i)\nu(k-i) \qquad (k = 0, 1, 2, \ldots)$$

zweier Wahrscheinlichkeitsverteilungen μ und ν auf $\mathbb{N} \cup \{0\}$ das Produkt der einzelnen erzeugenden Funktionen.

Erzeugende Funktionen können in verschiedenen Situationen für explizite Berechnungen verwendet werden. Wir demonstrieren dies hier in einigen grundlegenden Beispielen. Viele weitere entsprechende Anwendungen finden sich in den Wahrscheinlichkeitstheorie-Lehrbüchern von Feller und Grimmett/Stirzacker.

Erzeugende Funktionen zufälliger Summen

Sind $N, X_1, X_2, \ldots : \Omega \to \{0, 1, 2, \ldots\}$ unabhängige Zufallsvariablen, dann erhalten wir für die Summe $S_N = \sum_{i=1}^{N} X_i$:

$$G_{S_N}(s) \quad = \quad E[s^{S_N}] \quad = \quad E[E[s^{S_N} \mid N]] \quad = \quad E[G(s)^N] \quad = \quad G_N(G(s)), \quad (10.2.1)$$

wobei G die erzeugende Funktion der Summanden X_i ist. Für die Verteilung von S_N ergibt sich

$$P[S_N = k] \quad = \quad \frac{1}{k!}(G_N \circ G)^{(k)}(0) \qquad \text{für alle } k \geq 0.$$

Beispiel (Ausdünnungseigenschaft von Poissonverteilungen). Ein Huhn lege eine mit Parameter $\lambda > 0$ Poissonverteilte Anzahl N von Eiern, von denen aus jedem unabhängig voneinander und von N mit Wahrscheinlichkeit p ein Küken schlüpfe. Die erzeugende Funktion der Poissonverteilung ist

$$G_N(s) \quad = \quad E[s^N] \quad = \quad \sum_{k=0}^{\infty} s^k \cdot \frac{\lambda^k}{k!} e^{-\lambda} \quad = \quad e^{\lambda(s-1)}.$$

Die Anzahl der geschlüpften Küken ist $S_N = \sum_{i=1}^{N} X_i$, wobei die X_i untereinander und von N unabhängige, Bernoulli(p)-verteilte Zufallsvariablen sind. Wir erhalten also

$$G_{S_N}(s) \quad = \quad G_N(G_{X_1}(s)) \quad = \quad G_N(1 - p + p \cdot s) \quad = \quad e^{p\lambda \cdot (s-1)},$$

d.h. die Zahl der geschlüpften Küken ist wieder Poissonverteilt mit Parameter $p \cdot \lambda$. Eine ausgedünnte Poissonverteilung ist also wieder eine Poissonverteilung!

Galton-Watson-Verzweigungsprozesse

Wir betrachten das folgende Modell für ein zufälliges Populationswachstum: Alle Individuen der Population erzeugen unabhängig voneinander eine zufällige Anzahl von Nachkommen in der nächsten Generation mit Verteilung ν. Hierbei sei ν eine feste Wahrscheinlichkeitsverteilung auf $\{0, 1, 2, \ldots\}$ mit $\nu[\{2, 3, \ldots\}] \neq 0$. Dieses Modell wurde 1889 von Galton und Watson eingeführt, um die Aussterbewahrscheinlichkeit englischer Adelstitel zu untersuchen. Ähnlich wie beim Random Walk handelt es sich um ein fundamentales stochastisches Modell mit unzähligen Erweiterungen und Anwendungen, z.B. auf das Wachstum von Zellpopulationen, die Ausbreitung von Epidemien, die Zunahme der Neutronenzahl in einem Reaktor, oder auch die näherungsweise Berechnung von genetischen Abständen oder der Anzahl von Zuständen in einem großen zufälligen Graphen (z.B. dem Internet), die man in einer bestimmten Anzahl von Schritten erreichen kann. Die Nachkommensstruktur eines einzelnen Individuums bestimmt einen zufälligen verwurzelten Baum, s. Grafik 10.1. Dementsprechend spielen Verzweigungsprozesse auch eine zentrale Rolle bei der Analyse diverser stochastischer Modelle auf Bäumen, s. z.B. [Peres: Probablity on trees].

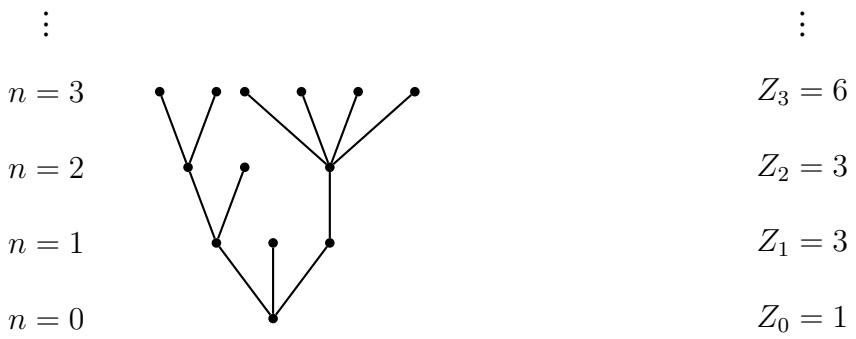


Abbildung 10.1: Beispiel für eine Realisierung eines Galton-Watson-Prozesses.

Wir beschreiben die Nachkommenszahlen der einzelnen Individuen in der $(n-1)$-ten Generation eines Verzweigungsprozesses durch unabhängige Zufallsvariablen

$$N_i^n : \Omega \to \{0, 1, 2, \ldots\}, \qquad i, n = 1, 2, \ldots,$$

mit Verteilung ν. Für die Gesamtzahl der Individuen in der n-ten Generation erhalten wir die folgende rekursive Darstellung:

$$Z_N \quad = \quad \sum_{i=1}^{Z_{n-1}} N_i^n \qquad \text{für alle } n \geq 1.$$

Ohne wesentliche Einschränkungen nehmen wir $Z_0 = 1$ an. Enthält die Anfangspopulation stattdessen mehrere Individuen, dann erzeugen diese voneinander unabhängige, identisch verteilte Unterpopulationen. Da Z_{n-1} nur von den Zufallsvariablen N_i^k für $k \leq n-1$ abhängt, sind Z_{n-1} und N_i^n ($i \in \mathbb{N}$) unabhängige Zufallsvariablen. Durch Bedingen auf Z_{n-1} erhalten wir für die mittleren Populationsgrößen die Rekursion

$$E[Z_n] \quad = \quad E[Z_{n-1}] \cdot m,$$

wobei $m := \sum_{i=1}^{\infty} i \cdot \nu(i)$ die mittlere Nachkommenszahl eines Individuums ist. Wir unterscheiden die folgenden Fälle:

$m > 1$:	Exponentielles Wachstum der Erwartungswerte	(superkritisch)
$m = 1$:	Erwartungswerte konstant	(kritisch)
$m < 1$:	Exponentieller Abfall der Erwartungswerte	(subkritisch)

Wir wollen nun untersuchen, mit welcher Wahrscheinlichkeit die Population in den einzelnen Fällen ausstirbt. Nach (10.2.1) gilt für die erzeugenden Funktionen der Populationsgrößen die Rekursionsformel

$$G_{Z_n}(s) = E\left[s^{\sum_{i=1}^{Z_{n-1}} N_i^n}\right] = G_{Z_{n-1}}(G(s)),$$

wobei G die erzeugende Funktion der Verteilung ν der Anzahl N_i^n der Kinder eines Individuums ist. Per Induktion folgt wegen $G_{Z_1}(s) = G(s)$:

$$G_{Z_n}(s) \quad = \quad \underbrace{G(G(\ldots G(s)\ldots))}_{n-\text{mal}} \quad = \quad G^n(s) \qquad \text{für alle } n \in \mathbb{N} \text{ und } s \in [0,1].$$

Für die Wahrscheinlichkeiten π_n, dass der Prozess zur Zeit n ausgestorben ist, erhalten wir die Rekursionsformel

$$\pi_n \; = \; P[Z_n = 0] \quad = \quad G_{Z_n}(0) \quad = \quad G^n(0) \quad = \quad G(\pi_{n-1}). \tag{10.2.2}$$

Sei nun π die Wahrscheinlichkeit, dass die Population in endlicher Zeit ausstirbt. Da die Ereignisse $\{Z_n = 0\}$ monoton wachsend sind, gilt

$$\pi \quad = \quad P\left[\bigcup_n \{Z_n = 0\}\right] \quad = \quad \lim_{n\to\infty} \pi_n.$$

Da G auf $[0,1]$ stetig ist, folgt aus (10.2.2)

$$\pi \quad = \quad G(\pi),$$

d.h. die Aussterbewahrscheinlichkeit π ist ein Fixpunkt der erzeugenden Funktion. Wie oben bemerkt, ist die erzeugende Funktion $G : [0,1] \to [0,1]$ strikt konvex mit $G(1) = 1$ und $G'(1-) = E[N_i^n] = m$. Hieraus folgt, dass 1 im Fall $m \leq 1$ der einzige Fixpunkt von G in $[0,1]$ ist, während im superkritischen Fall $m > 1$ ein weiterer Fixpunkt $\pi^* \in [0,1)$ existiert, siehe auch Grafik 10.2. Aus den Skizzen erkennt man zudem, dass die Aussterbewahrscheinlichkeit $\pi = \lim \pi_n$ der kleinste Fixpunkt von G auf $[0,1]$ ist. Also stirbt der Prozess im subkritischen bzw. kritischen Fall mit Wahrscheinlichkeit 1 aus, während er im superkritischen Fall mit einer strikt positiven Wahrscheinlichkeit überlebt.

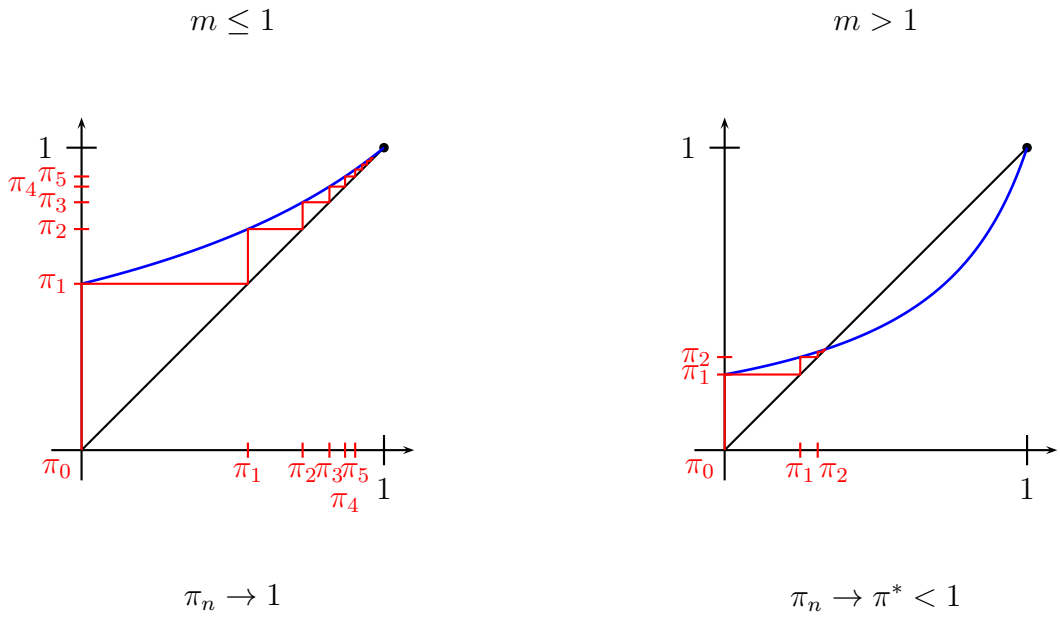


Abbildung 10.2: Erzeugendenfunktionen von Galton-Watson-Prozessen mit unterschiedlichen Verteilungen für die Anzahl der Nachkommen. In Rot: Fixpunktiteration

Beispiel (Geometrische Nachkommensverteilung). Ist die Verteilung

$$\nu(k) \quad = \quad p^k(1-p) \qquad (k = 0, 1, 2 \ldots)$$

der Anzahl der Nachkommen eine geometrische Verteilung mit Parameter $p \in (0, 1)$, dann ergibt sich

$$G(s) \quad = \quad \sum_{k=0}^{\infty} s^k p^k (1-p) \quad = \quad \frac{1-p}{1-ps} \qquad \text{für alle } s \in [0, 1].$$

Fixpunkte dieser Funktion sind 1 und $\frac{1-p}{p}$. Für $1 - p \geq p$ (subkritischer Fall) ist 1 der einzige Fixpunkt in $[0, 1]$, also stirbt die Population P-fast sicher aus. Im superkritischen Fall $1 - p < p$ beträgt die Aussterbewahrscheinlichkeit π dagegen nur $\frac{1-p}{p}$.

Rekurrente Ereignisse und Erneuerungsgleichung

Als weitere Anwendung von erzeugenden Funktionen betrachten wir eine Folge von unvorhersehbaren Ereignissen, die zu diskreten Zeitpunkten $n \in \mathbb{N}$ eintreten. Die Ereignisse bezeichnen wir auch als „Erneuerungen“ (engl. renewals), und denken dabei z.B. an den wiederholten Ausfall und Austausch eines Verschleißteils in einer Maschine, oder das wiederholte Abarbeiten einer

Warteschlange. Wir beschreiben den Zeitpunkt, an dem die k-te Erneuerung stattfindet, durch eine Zufallsvariable

$$S_k \quad = \quad T_1 + T_2 + \ldots + T_k.$$

T_1 ist also der Zeitpunkt der ersten Erneuerung, und für $k \geq 2$ ist T_k der zeitliche Abstand der $(k-1)$-ten und der k-ten Erneuerung. In einem einfachen Modell nehmen wir an, dass $T_1, T_2, \ldots : \Omega \to \mathbb{N}$ unabhängige Zufallsvariablen sind, und, dass $T_2, T_3, \ldots$ identisch verteilt sind (aber nicht T_1!). Wir wollen nun die Wahrscheinlichkeiten p_n der Ereignisse

$$A_n = \{\exists k \in \mathbb{N} : \ S_k = n\} \qquad \text{„Erneuerung zur Zeit } n\text{“}$$

aus den Verteilungen der Wartezeiten berechnen. Bedingen auf den Wert von T_1 liefert für $n \geq m$:

$$\begin{aligned} P[A_n \mid T_1 = m] &= P[\exists k \in \mathbb{N} : \ T_1 + \ldots + T_k = n \mid T_1 = m] \\ &= P[\exists k \in \mathbb{N} : \ T_2 + \ldots + T_k = n - m \mid T_1 = m] \\ &= P[\exists k \in \mathbb{N} : \ T_2 + \ldots + T_k = n - m], \end{aligned}$$

und damit

$$P[A_n \mid T_1 = m] \quad = \quad P[A_{n-m+1} \mid T_1 = 1] \quad = \quad P[A_{n-m+1} \mid A_1].$$

Nach der Formel von der totalen Wahrscheinlichkeit erhalten wir für $n \in \mathbb{N}$:

$$p_n \quad = \quad \sum_{m=1}^{n} q_{n-m} \cdot P[T_1 = m] \tag{10.2.3}$$

mit $q_n := P[A_{n+1} \mid A_1]$. Um die bedingten Wahrscheinlichkeiten q_n zu berechnen, bedingen wir zusätzlich auf T_2. Da $T_2, T_3, \ldots$ unabhängig und identisch verteilt sind, gilt für $n \geq m$:

$$\begin{aligned} P[A_{n+1} \mid A_1 \cap \{T_2 = m\}] &= P[\exists k \in \mathbb{N} : \ T_1 + \ldots + T_k = n + 1 \mid T_1 = 1, T_2 = m] \\ &= P[\exists k \geq 2 : \ T_3 + \ldots + T_k = n - m \mid T_1 = 1, T_2 = m] \\ &= P[\exists k \geq 2 : \ T_3 + \ldots + T_k = n - m] \\ &= P[\exists k \geq 2 : \ T_2 + \ldots + T_{k-1} = n - m] \\ &= P[A_{n-m+1} \mid A_1] \quad = \quad q_{n-m}. \end{aligned}$$

Wegen

$$q_n \quad = \quad P[A_{n+1} \mid A_1] \quad = \quad \sum_{m=1}^{n} P[A_{n+1} \mid A_1 \cap \{T_2 = m\}] \cdot P[T_2 = m]$$

erhalten wir

$$q_n \quad = \quad \sum_{m=1}^{n} q_{n-m} \cdot P[T_2 = m] \qquad \text{für alle } n \geq 1. \tag{10.2.4}$$

Die Gleichungen (10.2.3) und (10.2.4) heißen *Erneuerungsgleichungen*. Auf den rechten Seiten dieser Gleichungen stehen (wegen $T_1, T_2 \geq 1$) die Faltungen der Folge $q_n, n \in \mathbb{N}$, mit der Folge der Gewichte der Wartezeiten T_1 bzw. T_2. Daher ist es zweckmäßig, zu den erzeugenden Funktionen

$$G_p(s) \quad = \quad \sum_{n=1}^{\infty} p_n s^n$$

und

$$G_q(s) \quad = \quad \sum_{n=0}^{\infty} q_n s^n$$

überzugehen. Für $|s| < 1$ erhalten wir aus (10.2.3)

$$G_p(s) \quad = \quad G_q(s) \cdot G_{T_1}(s).$$

Aus (10.2.4) ergibt sich, da die rechte Seite für $n = 0$ verschwindet:

$$G_q(s) - 1 = \sum_{n=1}^{\infty} q_n s^n \quad = \quad G_q(s) \cdot G_{T_2}(s).$$

Es folgt $G_q(s) = (1 - G_{T_2}(s))^{-1}$, und damit

$$G_p(s) \quad = \quad \frac{G_{T_1}(s)}{1 - G_{T_2}(s)}. \tag{10.2.5}$$

(10.2.5) liefert den gesuchten Zusammenhang zwischen der Verteilung der Wartezeiten, und den Wahrscheinlichkeiten p_n, dass zur Zeit n eine Erneuerung stattfindet.

Sei nun die Verteilung der Lebensdauern $T_2, T_3, \ldots$ vorgegeben. Dann können wir untersuchen, welche Verteilung die Anfangswartezeit T_1 haben muss, damit die Wahrscheinlichkeiten p_n nicht von n abhängen (*Stationarität*). Für $\alpha \in [0, 1]$ gilt $p_n = \alpha$ für alle $n \in \mathbb{N}$ genau dann, wenn

$$G_p(s) \quad = \quad \sum_{n=1}^{\infty} p_n s^n \quad = \quad \frac{\alpha}{1 - s} \qquad \text{für alle } s \in (-1, 1),$$

d.h. wenn

$$G_{T_1}(s) \quad = \quad \alpha \cdot \frac{1 - G_{T_2}(s)}{1 - s} \qquad \text{für alle } s \in (-1, 1). \tag{10.2.6}$$

Da G_{T_1} und G_{T_2} erzeugende Funktionen von Wahrscheinlichkeitsverteilungen sind, muss dann gelten:

$$\begin{aligned} 1 \;&=\; G_{T_1}(1) \quad = \quad \lim_{s \uparrow 1} G_{T_1}(s) \\ &=\; \alpha \lim_{s \uparrow 1} \frac{G_{T_2}(s) - 1}{s - 1} \quad = \quad \alpha G'_{T_2}(1-) \\ &=\; \alpha \cdot E[T_2]. \end{aligned}$$

Also muss T_2 endlichen Erwartungswert haben, und

$$\alpha \quad = \quad 1/E[T_2] \tag{10.2.7}$$

gelten. Dies ist auch anschaulich plausibel: Im stationären Fall ist die Erneuerungswahrscheinlichkeit zu einem festen Zeitpunkt der Kehrwert des mittleren zeitlichen Abstandes zwischen zwei Erneuerungen. Gilt (10.2.7), dann ergibt sich aus (10.2.6) durch Koeffizientenvergleich:

$$P[T_1 = n] \quad = \quad \alpha \cdot \left(1 - \sum_{k=1}^{n} P[T_2 = k]\right) \quad = \quad \frac{P[T_2 > n]}{E[T_2]}. \tag{10.2.8}$$

Die Folge p_n der Erneuerungswahrscheinlichkeiten ist also genau dann konstant, wenn die Verteilung von T_1 durch (10.2.8) gegeben ist („stationärer Erneuerungsprozess“). Weiter kann man ausgehend von (10.2.6) zeigen, dass für *beliebige* Verteilungen der ersten Erneuerungszeit die Wahrscheinlichkeiten p_n für $n \to \infty$ gegen $1/E[T_2]$ konvergieren („asymptotische Stationarität“), falls der Erwartungswert endlich ist und keine *Periodizität* auftritt, d.h.

$$\mathrm{ggT}(\{n | P[T_2 = n] > 0\}) \quad = \quad 1.$$

Den Beweis dieses **Erneuerungssatzes** über erzeugende Funktionen findet man im Klassiker von W.Feller (An Introduction to Probability Theory and its Applications, Vol. 1).

10.3 Bedingen auf allgemeine Zufallsvariablen

Ist Y eine reellwertige Zufallsvariable auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ mit stetiger Verteilungsfunktion, dann gilt $P[Y = z] = 0$ für alle $z \in \mathbb{R}$. Bedingte Wahrscheinlichkeiten gegeben $Y = z$ können daher nicht wie für diskrete Zufallsvariablen definiert werden. Alternativ könnte man versuchen, bedingte Wahrscheinlichkeiten gegeben Y als Grenzwert zu definieren:

$$P[A \mid Y = z] \quad = \quad \lim_{h \searrow 0} P[A \mid z - h \leq Y \leq z + h]. \tag{10.3.1}$$

Dies ist in bestimmten Fällen möglich, aber im allgemeinen ist die Existenz des Grenzwertes nicht gewährleistet.

Stattdessen definiert man bedingte Erwartungen gegeben allgemeine Zufallsvariablen Y mithilfe der Charakterisierung aus Satz 10.5. Bedingte Wahrscheinlichkeiten gegeben Y erhält man als Spezialfall bedingter Erwartungen:

$$P[A \mid Y] \quad := \quad E[I_A \mid Y]. \tag{10.3.2}$$

Bedingte Wahrscheinlichkeiten wie in (10.3.1) sind im Allgemeinen nicht im herkömmlichen Sinn definiert. Es ist allerdings ausgehend von (10.3.1) allgemein möglich, für ein festes Ereignis A die Abbildung $z \mapsto P[A \mid Y = z]$ bis auf Modifikation auf Nullmengen bzgl. der Verteilung von Y zu definieren.

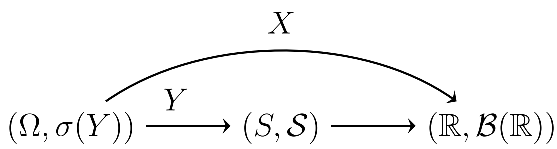
Das Faktorisierungslemma

Wir beweisen zunächst eine wichtige maßtheoretische Aussage. Diese wird es uns unter Anderem ermöglichen, die charakterisierenden Eigenschaften bedingter Erwartungen aus Satz 10.5 noch etwas eleganter zu formulieren:

Satz 10.6 (Faktorisierungslemma). *Sei $(S, \mathcal{S})$ ein messbarer Raum und $Y : \Omega \to S$ eine Abbildung. Eine Abbildung $X : \Omega \to \mathbb{R}$ ist genau dann $\sigma(Y)$-messbar, wenn*

$$X = f(Y) = f \circ Y$$

für eine $\mathcal{S}$-messbare Funktion $f : S \to \mathbb{R}$ gilt.



Beweis. (1). Ist $X = f \circ Y$ für eine messbare Funktion f, dann gilt

$$X^{-1}(B) \quad = \quad Y^{-1}(f^{-1}(B)) \in \sigma(Y) \qquad \text{für alle } B \in \mathcal{B}(\mathbb{R}),$$

da $f^{-1}(B) \in \mathcal{S}$. Daher ist X $\sigma(Y)$-messbar.

(2). Für die umgekehrte Richtung müssen wir zeigen, dass aus der $\sigma(Y)$-Messbarkeit von X folgt, dass X eine messbare Funktion von Y ist. Dazu gehen wir schrittweise vor („maßtheoretische Induktion“):

(a) Ist $X = I_A$ eine Indikatorfunktion mit $A \in \sigma(Y)$, dann gilt $A = Y^{-1}(B)$ mit $B \in \mathcal{S}$, und damit

$$X(\omega) \quad = \quad I_{Y^{-1}(B)}(\omega) \quad = \quad I_B(Y(\omega)) \qquad \text{für alle } \omega \in \Omega.$$

(b) Für $X = \sum_{i=1}^n c_i I_{A_i}$ mit $A_i \in \sigma(Y)$ und $c_i \in \mathbb{R}$ gilt entsprechend

$$X = \sum_{i=1}^n c_i I_{B_i}(Y),$$

wobei B_i Mengen aus S mit $A_i = Y^{-1}(B_i)$ sind.

(c) Für eine beliebige nichtnegative, $\sigma(Y)$-messbare Abbildung $X : \Omega \to \mathbb{R}$ existiert eine Folge (X_n) von $\sigma(Y)$-messbaren Elementarfunktionen mit $X_n \nearrow X$. Nach (b) gilt $X_n = f_n(Y)$ mit $\mathcal{S}$-messbaren Funktionen f_n. Damit folgt:

$$X \quad = \quad \sup X_n \quad = \quad \sup f_n(Y) \quad = \quad f(Y),$$

wobei $f = \sup f_n$ wieder $\mathcal{S}$-messbar ist.

(d) Für eine allgemeine $\sigma(Y)$-messbare Abbildung $X : \Omega \to \mathbb{R}$ sind sowohl X^+ als auch X^- messbare Funktionen von Y, also auch X selbst.

□

Mithilfe des Faktorisierungslemmas können wir die *charakterisierenden Eigenschaften* (I) und (II) bedingter Erwartungen gegeben eine diskrete Zufallsvariable Y aus Satz 10.5 wie folgt umformulieren:

$\overline{X}$ ist genau dann eine Version von $E[X \mid Y]$, wenn gilt:

(i) $\overline{X}$ ist $\sigma(Y)$-messbar,

(ii) $E[\overline{X}\,;\, A] \,=\, E[X\,;\, A]$ *für alle* $A \in \sigma(Y)$.

Die Äquivalenz von (I) und (i) folgt aus dem Faktorisierungslemma, und die Äquivalenz von (II) und (ii) ergibt sich durch maßtheoretische Induktion, denn (ii) besagt gerade, dass

$$E[\overline{X} \cdot I_B(Y)] \quad = \quad E[X \cdot I_B(Y)] \qquad \text{für alle } B \in \mathcal{S} \text{ gilt.}$$

Definition allgemeiner bedingter Erwartungen

Eine bemerkenswerte Konsequenz der Charakterisierung bedingter Erwartungen durch die Bedingungen (i) und (ii) ist, dass die *bedingte Erwartung $E[X \mid Y]$ von der Zufallsvariablen Y nur über die von Y erzeugte σ-Algebra $\sigma(Y)$ abhängt!* Sind zwei Zufallsvariablen Y und Z Funktionen voneinander, dann ist $\sigma(Y) = \sigma(Z)$, und damit stimmen auch die bedingten Erwartungen $E[X \mid Y]$ und $E[X \mid Z]$ überein (mit Wahrscheinlichkeit 1). Daher liegt es nahe, gleich von der bedingten Erwartung gegeben eine σ-Algebra zu sprechen. Die σ-Algebra (z.B. $\sigma(Y)$ oder $\sigma(Y_1, \ldots, Y_n)$) beschreibt dann die zur Verfügung stehende „Information“, auf die bedingt wird.

Die Charakterisierung bedingter Erwartungen durch (i) und (ii) können wir sofort auf den Fall allgemeiner bedingter Erwartungen gegeben eine σ-Algebra oder gegeben beliebige Zufallsvariablen übertragen. Sei dazu $X : \Omega \to \overline{\mathbb{R}}$ eine nichtnegative (oder integrierbare) Zufallsvariable auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$.

Definition (Bedingte Erwartung, allgemein). *(1). Sei $\mathcal{F} \subseteq \mathcal{A}$ eine σ-Algebra. Eine nichtnegative (bzw. integrierbare) Zufallsvariable $\overline{X} : \Omega \to \overline{\mathbb{R}}$ heißt **Version der bedingten Erwartung** $E[X \mid \mathcal{F}]$, falls gilt:*

(a) $\overline{X}$ ist $\mathcal{F}$-messbar, und

(b) $E[\overline{X}\ ;\ A] = E[X\ ;\ A]$ für alle $A \in \mathcal{F}$.

(2). Für beliebige Zufallsvariablen $Y, Y_1, Y_2, \ldots, Y_n$ auf $(\Omega, \mathcal{A}, P)$ definieren wir

$$\begin{aligned} E[X \mid Y] &:= E[X \mid \sigma(Y)], \\ E[X \mid Y_1, \ldots Y_n] &:= E[X \mid (Y_1, \ldots, Y_n)] \quad = \quad E[X \mid \sigma(Y_1, \ldots, Y_n)]. \end{aligned}$$

(3). Für ein Ereignis $A \in \mathcal{A}$ definieren wir

$$P[A \mid \mathcal{F}] \quad := \quad E[I_A \mid \mathcal{F}], \quad \text{und entsprechend} \quad P[A \mid Y] \quad = \quad E[A \mid Y].$$

Bemerkung. Durch maßtheoretische Induktion zeigt man, dass Bedingung (b) äquivalent ist zu:

(b') $E[\overline{X} \cdot Z] = E[X \cdot Z]$ für alle nichtnegativen (bzw. beschränkten) $\mathcal{F}$-messbaren $Z : \Omega \to \mathbb{R}$.

Satz 10.7 (Existenz und Eindeutigkeit der bedingten Erwartung). *Sei $X \geq 0$ oder $X \in \mathcal{L}^1$, und $\mathcal{F} \subseteq \mathcal{A}$ eine σ-Algebra. Dann gilt*

(1). Es existiert eine Version der bedingten Erwartung $E[X \mid \mathcal{F}]$.

(2). Zwei Versionen stimmen P-fast sicher überein.

Beweis. Die Existenz kann man unmittelbar aus dem Satz von Radon-Nikodym folgern, s. z.B. [A.Klenke, Wahrscheinlichkeitstheorie]. Wir geben stattdessen am Ende von Abschnitt 10.4 einen Existenzbeweis, der mit elementaren Methoden auskommt.
Zum Beweis der Eindeutigkeit seien $\overline{X}$ und $\widetilde{X}$ zwei Versionen der bedingten Erwartung $E[X \,|\, \mathcal{F}]$. Dann sind $\overline{X}$ und $\widetilde{X}$ beide $\mathcal{F}$-messbar, und

$$E[\overline{X}\ ;\ A] \quad = \quad E[\widetilde{X}\ ;\ A] \qquad \text{für alle } A \in \mathcal{F}.$$

Hieraus folgt $\overline{X} = \widetilde{X}$ P-fast sicher. □

Bemerkung (Probleme mit Ausnahmemengen). Man beachte, dass die bedingte Erwartung $E[X \mid \mathcal{F}]$ und damit auch die bedingte Wahrscheinlichkeit $P[A \mid \mathcal{F}]$ nur für jede *feste* Zufallsvariable X bzw. jedes *feste* Ereignis A bis auf Modifikation auf Nullmengen eindeutig definiert sind. Ein weiteres Problem ist, dass wir allgemein zwar bedingte Erwartungen gegeben eine Zufallsvariable Y definieren können, aber nicht solche gegeben das Ereignis $Y = z$ für festes z. In vielen Fällen kann man die beschriebenen Probleme durch Auswahl einer „regulären Version der bedingten Verteilung gegeben Y“ umgehen. Wir kommen darauf in Korollar 10.9 zurück.

Bemerkung ($\mathbf{E[X \mid Y = z]}$). Obwohl $E[X \mid Y = z]$ für ein festes z im Allgemeinen nicht definiert ist, kann man die Funktion $z \mapsto E[X \mid Y = z]$ bis auf Modifikation auf Nullmengen bzgl. der Verteilung von Y sinnvoll definieren: Ist $Y : \Omega \to S$ eine Zufallsvariable mit Werten in einem messbaren Raum $(S, \mathcal{S})$, dann ist jede Version der bedingten Erwartung $E[X \,|\, Y]$ nach Definition $\sigma(Y)$-messbar. Also gilt nach dem Faktorisierungslemma:

$$E[X \mid Y] \quad = \quad g(Y) \qquad \text{für eine messbare Funktion } g : S \to \mathbb{R}. \tag{10.3.3}$$

Da die Versionen der bedingten Erwartung bis auf Modifikation auf P-Nullmengen eindeutig festgelegt sind, ist die Funktion g bis auf Modifikation auf μ_Y-Nullmengen eindeutig festgelegt. In Anlehnung an den diskreten Fall setzt man manchmal:

$$E[X \mid Y = z] \quad := \quad g(z). \tag{10.3.4}$$

Genauer definieren wir für eine nichtnegative Zufallsvariable X:

Definition. *Eine messbare Funktion $g : S \to \mathbb{R}^+$ heißt **Version der bedingten Erwartung** $z \mapsto E[X \mid Y = z]$ **von** X **gegeben** $Y = z$, wenn gilt:*

$$E[X\,;\, Y \in B] \quad = \quad \int\limits_B g(z)\mu_Y(dz) \qquad \textit{für alle } B \in \mathcal{S}. \tag{10.3.5}$$

Die charakterisierende Bedingung (10.3.5) ist nichts anderes als eine allgemeine Variante der *Formel von der totalen Wahrscheinlichkeit*. Mithilfe des Transformationssatzes sieht man, dass g genau dann (10.2.3) erfüllt, wenn $g(Y)$ eine Version von $E[X \mid Y]$ ist.

WARNUNG: Bei der Definition ist zu beachten, dass $E[X \mid Y = z]$ für ein festes z im Allgemeinen nicht definiert ist, sondern nur die Funktion $z \mapsto E[X \mid Y = z]$ modulo Modifikation auf μ_Y-Nullmengen! Das formale Rechnen mit bedingten Erwartungen wir in (10.3.4) ist daher eine häufige Fehlerquelle.

Trotz dieser Gefahren ist die Notation $E[X \mid Y = z]$ oft nützlich, um Argumentationen transparenter zu machen, oder um anschauliche Überlegungen in mathematische Formeln zu übersetzen. Wir werden sie daher auch hier gelegentlich verwenden.

Diskreter und absolutstetiger Fall

In einigen Fällen kann man die Definition direkt anwenden, um bedingte Erwartungswerte zu berechnen. Wir betrachten zunächst noch einmal den Spezialfall eine *diskreten Bedingung:*
Gilt $\mathcal{F} = \sigma(\{H_i | i \in \mathbb{N}\})$ für eine disjunkte Zerlegung $\Omega = \dot{\bigcup_{i\in\mathbb{N}}} H_i$ in abzählbar viele messbare Teilmengen („Hypothesen“) $H_i \in \mathcal{A}$, dann sind $\mathcal{F}$-messbare Zufallsvariablen konstant auf jeder der Mengen H_i. Aus der Definition der bedingten Erwartung folgt dann

$$E[X \mid \mathcal{F}] \quad = \quad E[X \mid H_i] \qquad \text{auf } H_i$$

für alle $i \in \mathbb{N}$ mit $P[H_i] > 0$.

Beispiel (Unbedingte Erwartungen). Die bedingte Erwartung einer Zufallsvariable X gegeben die triviale σ-Algebra $\{\emptyset, \Omega\}$ ist der Erwartungswert von X.

Beispiel (Bedingen auf eine Partition). Ist $P = \mathcal{U}_{[0,1)}$ die Gleichverteilung auf $[0, 1)$, und $\mathcal{F} = \sigma(\{[t_{i-1}, t_i) | i = 1, \ldots, n\})$ die von einer Partition $0 = t_0 < t_1 < t_2 < \ldots < t_n = 1$ erzeugte σ-Algebra, denn ist die bedingte Erwartung $E[g \,|\, \mathcal{F}]$ einer integrierbaren Funktion $g : [0, 1) \to \mathbb{R}$ die durch

$$E[g \mid \mathcal{F}] \quad = \quad \frac{1}{t_i - t_{i-1}} \int_{t_{i-1}}^{t_i} g(u)\, du \qquad \text{auf } [t_{i-1}, t_i)$$

definierte Funktion.

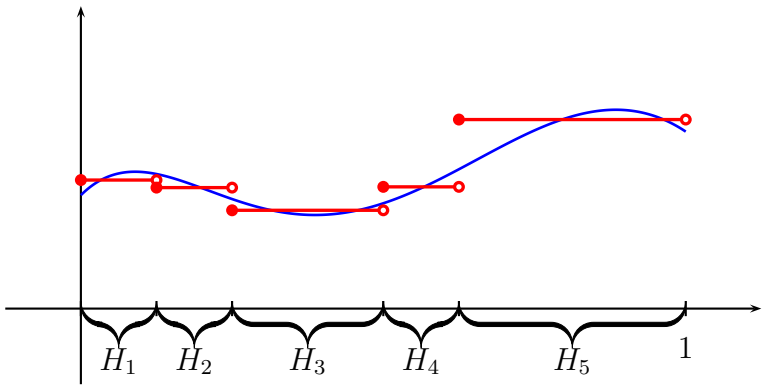


Abbildung 10.3: Die Funktion $g(\omega)$ ist hier blau dargestellt und $E[g|\mathcal{F}]$ in rot.

Ist die gemeinsame Verteilung aller relevanten Zufallsvariablen absolutstetig, dann kann man bedingte Erwartungen mithilfe von bedingten Dichten berechnen:

Satz 10.8 (Berechnung bedingter Erwartungen im absolutstetigen Fall). *Seien $X : \Omega \to \mathbb{R}^n$ und $Y : \Omega \to \mathbb{R}^m$ Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$, deren gemeinsame Verteilung $\mu_{X,Y}$ absolutstetig ist, und sei $h : \mathbb{R}^n \times \mathbb{R}^m \to [0, \infty]$ messbar. Dann ist*

$$E[h(X,Y) \mid Y](\omega) \quad = \quad \int\limits_{\mathbb{R}^n} h(x, Y(\omega)) f_{X|Y}(x|Y(\omega))\, dx \tag{10.3.6}$$

eine Version der bedingten Erwartung von $h(X,Y)$ gegeben Y.

Beweis. Nach dem Satz von Fubini ist die rechte Seite von (10.3.6) eine messbare Funktion von $Y(\omega)$, und es gilt

$$\begin{aligned} E\left[g(Y) \cdot \int h(x,Y) f_{X|Y}(x|Y)\, dx\right] &= \int\int g(y) h(x,y) f_{X|Y}(x|y) f_Y(y)\, dx\, dy \\ &= E[g(Y)h(X,Y)] \end{aligned}$$

für jede messbare Funktion $g : \mathbb{R}^m \to [0, \infty]$. □

Mit der Notation aus (10.3.4) lautet die Aussage des Satzes:

$$E[h(X,Y)|Y = z] \quad = \quad \int\limits_{\mathbb{R}^n} h(x,z) f_{X|Y}(x|z)\, dx \qquad \text{für } \mu_Y\text{-fast alle } z \in S.$$

Um die bedingte Erwartung zu berechnen, müssen wir also den uns bekannten Wert von Y einsetzen, und die Funktion bzgl. der bedingten Dichte $f_{X|Y}$ nach x integrieren.

Beispiel (Bedingen auf eine Koordinate). Ist $P = \mathcal{U}_\Omega$ die Gleichverteilung auf einer beschränkten, messbaren Menge $\Omega \subseteq \mathbb{R}^2$, und ist

$$Y : \Omega \to \mathbb{R}, \quad Y(x,y) = y,$$

die Projektion auf die zweite Komponente, dann gilt

$$E[h|Y](x,y) \quad = \quad \frac{1}{\lambda(\Omega_y)} \int\limits_{\Omega_y} h(x,y)\, dx \qquad P\text{-fast sicher} \tag{10.3.7}$$

für jede integrierbare Funktion $h : \Omega \to \mathbb{R}$. Hierbei ist $\Omega_y = \{x \in \mathbb{R} | (x,y) \in \mathbb{R}\}$ der y-Schnitt von Ω. Bedingen auf Y entspricht hier also dem normierten „Herausintegrieren“ der komplementären Koordinate x.

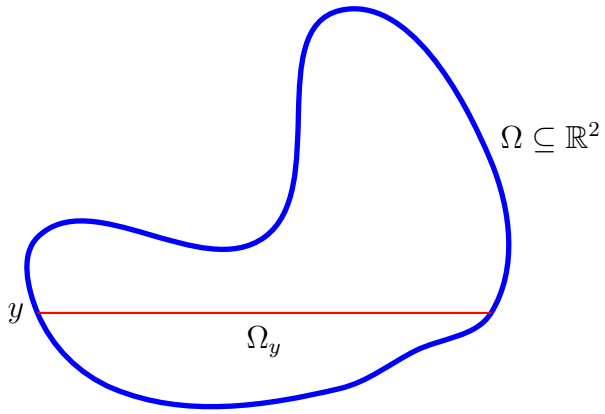


Abbildung 10.4: In Rot: Der y-Schnitt der Menge Ω.

Reguläre bedingte Verteilungen

Beim Bedingen auf diskrete Zufallsvariablen konnten wir bedingte Wahrscheinlichkeitsverteilungen auf elementare Weise definieren. Für allgemeine Zufallsvariablen sind die bedingten Wahrscheinlichkeiten

$$P[X \in B \,|\, Y] \quad = \quad E[I_B(X) \,|\, Y]$$

für jede feste messbare Menge B nur bis auf Modifikation auf P-Nullmengen eindeutig definiert. Dies ist ein Nachteil, da die Ausnahmemenge von B abhängen kann, und im Allgemeinen überabzählbar viele messbare Mengen existieren. Die bedingte Verteilung von X gegeben Y ist daher zunächst nicht definiert. Im absolutstetigen Fall können wir das Problem umgehen, indem wir die über die bedingte Dichte gegebene Version

$$\mu_{X|Y}(y, dx) \quad := \quad f_{X|Y}(x|y)dx$$

der bedingten Verteilung verwenden. Aus Satz 10.8 folgt unmittelbar, dass wir bedingte Wahrscheinlichkeiten gegeben Y aus $\mu_{X|Y}$ berechnen können:

Korollar 10.9. *Ist die gemeinsame Verteilung der Zufallsvariablen $X : \Omega \to \mathbb{R}^n$ und $Y : \Omega \to \mathbb{R}^m$ absolutstetig, dann ist $\mu_{X|Y}$ eine **reguläre Version der bedingten Verteilung von X gegeben** Y, d.h.*

(1). $\mu_{X|Y}$ ist ein stochastischer Kern von $\mathbb{R}^m$ nach $\mathbb{R}^n$.

(2). Für jedes $B \in \mathcal{B}(\mathbb{R}^n)$ ist

$$P[X \in B \,|\, Y] \quad = \quad \mu_{X|Y}(Y, B)$$

eine Version der bedingten Wahrscheinlichkeit von $\{X \in B\}$ gegeben Y.

Bemerkung (Existenz von regulären Versionen bedingter Verteilungen). Die Existenz von regulären Versionen von bedingten Verteilungen gegeben eine Zufallsvariable Y kann man allgemein beweisen, wenn Y Werte in einem vollständigen, separablen, metrischen Raum (kurz: polnischen Raum) annimmt, siehe z.B. [Breiman, Ch. 4.3.]. Eine explizite Berechnung über bedingte Dichten ist natürlich im Allgemeinen nicht möglich.

Wenn wir uns auf eine bestimmte reguläre Version $\mu_{X|Y}$ festlegen, dann können wir die bedingten Wahrscheinlichkeiten $P[X \in B|Y = z]$ durch

$$P[X \in B \mid Y = z] \;=\; \mu_{X|Y}(z, B)$$

für *alle* $z \in S$ definieren. Die Festlegung auf eine bestimmte reguläre Version der bedingten Verteilung ist im Allgemeinen willkürlich. Manchmal gibt es aber eine kanonische Version, die sich auszeichnet. Dies ist zum Beispiel der Fall, wenn die Dichte der gemeinsamen Verteilung von X und Y eine stetige Version hat.

Beispiel (Bivariate Normalverteilung). Ist (X, Y) bivariat normalverteilt mit Mittel $(0, 0)$ und Kovarianzmatrix $\begin{pmatrix} 1 & \varrho \\ \varrho & 1 \end{pmatrix}$, $\varrho \in (-1, 1)$, dann gilt

$$f_{X,Y}(x, y) \quad = \quad \frac{1}{2\pi\sqrt{1-\varrho^2}} \cdot \exp\left(-\frac{x^2 - 2\varrho xy + y^2}{2(1-\varrho^2)}\right).$$

Für ein festes $x \in \mathbb{R}$ folgt

$$f_{Y|X}(y|x) \quad \propto \quad f_{X,Y}(x, y) \quad \propto \quad \exp\left(-\frac{(y-\varrho x)^2}{2(1-\varrho^2)}\right)$$

als Funktion von y. Also ist

$$\mu_{Y|X}(x, \bullet) \quad = \quad N(\varrho x, 1-\varrho^2)$$

eine kanonische reguläre Version der bedingten Verteilung von Y gegeben X.

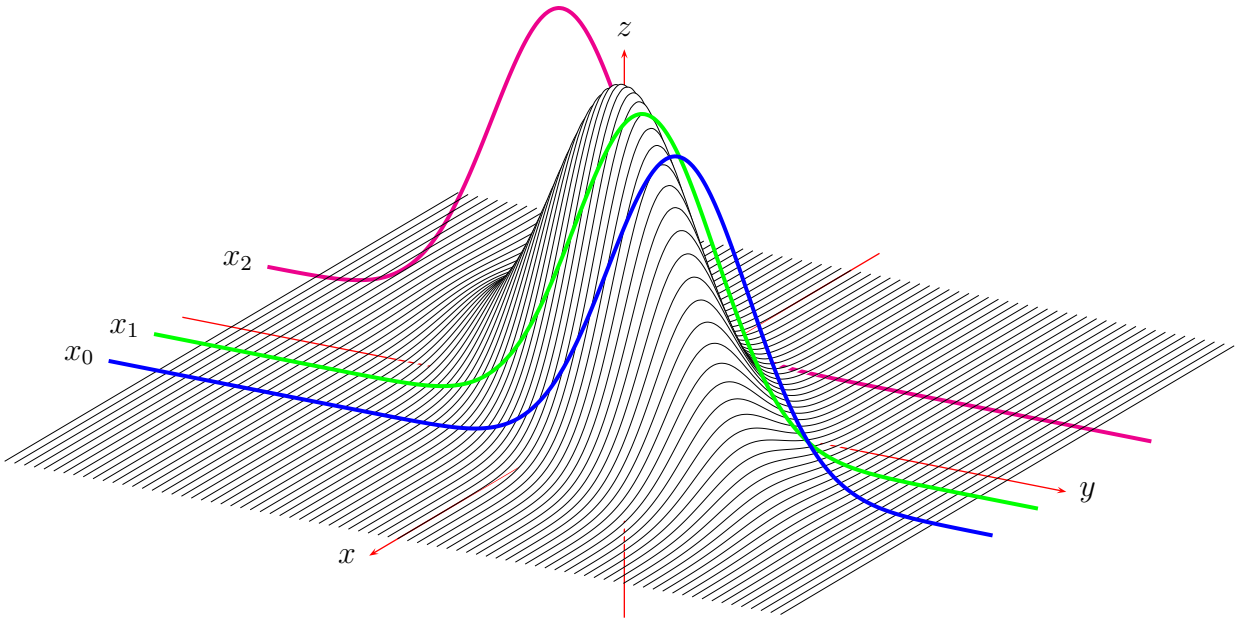


Abbildung 10.5: Die Dichte $f_{X,Y}(x,y)$ und in Blau, Grün und Magenta $f_{Y|X}(y|x_i)$ für $i \in \{0,1,2\}$. Man beachte, dass $f_{Y|X}(y|x_i) \propto f_{X,Y}(x_i,y)$ als Funktion von y.

Beispiel (Grenzen naiven Bedingens). Sei (X,Y) gleichverteilt auf dem Viertelkreis

$$S \quad = \quad \{(x,y) \in \mathbb{R}^2 | x > 0, y > 0, x^2 + y^2 < 1\}.$$

Wir versuchen auf zwei Arten eine „bedingte Verteilung von X gegeben $X = Y$“ zu berechnen. Dazu betrachten wir die Zufallsvariablen $V = Y - X$ und $W = Y/X$. Wegen $f_{X,Y} \propto I_S$ erhalten wir mithilfe des Dichtetransformationssatzes für fast jedes v:

$$\begin{aligned} f_{X|V}(x|v) \quad &\propto \quad f_{X,V}(x,v) \quad = \quad f_{X,Y}(x,v+x) \cdot \left| \det \frac{\partial(x,v+x)}{\partial(x,v)} \right| \\ &\propto \quad I_S(x,v+x), \end{aligned}$$

wobei „$\propto$“ für „proportional als Funktion von x“ steht. Wählen wir die normierte rechte Seite als kanonische Version der bedingten Dichte, so ergibt sich

$$f_{X|V}(x|0) \quad \propto \quad I_S(x,x) \quad = \quad I_{(0,1/\sqrt{2})}(x).$$

Gegeben $Y - X = 0$ ist X also gleichverteilt auf $(0, 1/\sqrt{2})$.

Andererseits erhalten wir für fast jedes w:

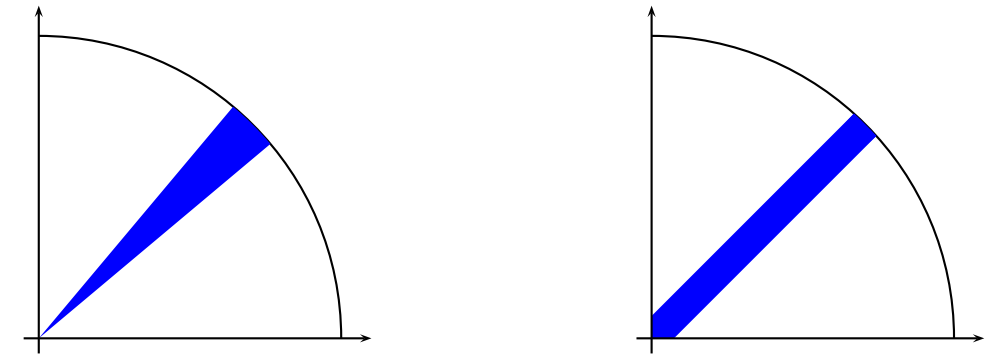
$$\begin{aligned} f_{X|W}(x|w) \quad &\propto \quad f_{X,W}(x,w) \quad = \quad f_{X,W}(x,wx) \cdot \left| \det \frac{\partial(x,wx)}{\partial(x,w)} \right| \\ &\propto \quad I_S(x,wx) \cdot x. \end{aligned}$$

Wählen wir wieder die rechte Seite als kanonische Version, so ergibt sich

$$f_{X|W}(x|1) \quad \propto \quad x \cdot I_S(x,x) \quad = \quad x \cdot I_{(0,1/\sqrt{2})}(x).$$

Die bedingte Verteilung von X gegeben $Y/X = 1$ unterscheidet sich also von der bedingten Verteilung von X gegeben $Y - X = 0$. Bedingte Wahrscheinlichkeiten gegeben $X = Y$ sind daher nicht wohldefiniert!

Eine anschauliche Erklärung für das Phänomen ist, dass wir in den beiden Fällen oben auf unterschiedliche infinitesimale Umgebungen der Diagonale $\{(x,y) \in S | x = y\}$ bedingen, wie die folgende Grafik veranschaulicht:



$\left\{(x,y) \in S \,:\, \left|\frac{x}{y} - 1\right| < \delta\right\}$ $\qquad$ $\{(x,y) \in S \,:\, |y - x| < \delta\}$

Abbildung 10.6: Zwei verschiedene Arten die Diagonale zu approximieren.

10.4 Rechnen mit bedingten Erwartungen; Poissonprozess

In vielen Fällen tritt eine Kombination bedingter Erwartungen bezüglich verschiedener Zufallsvariablen und/oder σ-Algebren auf. Die bedingten Erwartungswerte können dann meist nicht unmittelbar berechnet werden, lassen sich aber mithilfe grundlegender Eigenschaften und Rechenregeln schrittweise umformen und ggf. vereinfachen. Wir leiten nun aus der Definition einige fundamentale Eigenschaften bedingter Erwartungen her, die wir in diesem Zusammenhang häufig verwenden werden.

Als eine erste Anwendung untersuchen wir zeitliche und räumliche Poissonprozesse. Zeitliche Poissonprozesse sind die einfachsten Beispiele von zeitstetigen stochastischen Prozessen mit stationären unabhängigen Inkrementen, bzw. von zeitstetigen Markovketten. Räumliche Poissonprozesse (Poissonsche Punktprozesse) sind grundlegende Modelle für zufällige Punktmengen. Beide Arten von Prozessen spielen in etlichen Anwendungsbereichen eine wichtige Rolle

(z.B. Warteschlangen, Versicherungsmathematik, Materialwissenschaften, stochastische Geometrie etc.), und bilden die Basis für die Konstruktion vieler komplexerer stochastischer Modelle.

Eigenschaften der bedingten Erwartung

Wir leiten zunächst aus der Definition einige fundamentale Eigenschaften der bedingten Erwartung her, die wir häufig bei der Berechnung bedingter Erwartungswerte verwenden werden:

Satz 10.10. *Seien X, Y und X_n $(n \in \mathbb{N})$ nichtnegative oder integrierbare Zufallsvariablen auf $(\Omega, \mathcal{A}, P)$, und seien $\mathcal{F}, \mathcal{G} \subseteq \mathcal{A}$ σ-Algebren.*
Es gelten folgende Aussagen:

(1). Linearität: $E[\lambda X + \mu Y \mid \mathcal{F}] = \lambda\, E[X \mid \mathcal{F}] + \mu\, E[Y \mid \mathcal{F}]$ P-fast sicher für alle $\lambda, \mu \in \mathbb{R}$.

(2). Monotonie: Aus $X \geq 0$ P-fast sicher folgt $E[X \mid \mathcal{F}] \geq 0$ P-fast sicher.

(3). Aus $X = Y$ P-fast sicher folgt $E[X \mid \mathcal{F}] = E[Y \mid \mathcal{F}]$ P-fast sicher.

(4). Monotone Konvergenz: Ist (X_n) monoton wachsend mit $X_1 \geq 0$, dann gilt

$$E[\sup X_n \mid \mathcal{F}] \quad = \quad \sup E[X_n \mid \mathcal{F}] \qquad P\text{-fast sicher.}$$

(5). Projektivität / Tower Property: Ist $\mathcal{G} \subseteq \mathcal{F}$, dann gilt

$$E[E[X \mid \mathcal{F}] \mid \mathcal{G}] \quad = \quad E[X \mid \mathcal{G}] \qquad P\text{-fast sicher.}$$

Insbesondere:

$$E[E[X \mid Y, Z] \mid Y] \quad = \quad E[X|Y] \qquad P\text{-fast sicher.}$$

(6). Herausziehen, was bekannt ist: Sei Y $\mathcal{F}$-messbar mit $Y \cdot X \in \mathcal{L}^1$ bzw. ≥ 0. Dann gilt

$$E[Y \cdot X \mid \mathcal{F}] \quad = \quad Y \cdot E[X \mid \mathcal{F}] \qquad P\text{-fast sicher.}$$

(7). Unabhängigkeit: Ist X unabhängig von $\mathcal{F}$, dann gilt $E[X \mid \mathcal{F}] = E[X]$ P-fast sicher.

(8). Seien $(S, \mathcal{S})$ und $(T, \mathcal{T})$ messbare Räume. Ist $Y : \Omega \to S$ $\mathcal{F}$-messbar, und $X : \Omega \to T$ unabhängig von $\mathcal{F}$, und $f : S \times T \to [0, \infty)$ eine produktmessbare Abbildung, dann gilt

$$E[f(X, Y) \mid \mathcal{F}](\omega) \quad = \quad E[f(X, Y(\omega))] \qquad \text{für } P\text{-fast alle } \omega.$$

Beweis. (1). Aus der Linearität des Erwartungswertes folgt, dass $\lambda E[X \mid \mathcal{F}] + \mu E[Y \mid \mathcal{F}]$ eine Version der bedingten Erwartung $E[\lambda X + \mu Y \mid \mathcal{F}]$ ist.

(2). Sei $\overline{X}$ eine Version von $E[X \mid \mathcal{F}]$. Aus $X \geq 0$ P-fast sicher folgt wegen $\{\overline{X} < 0\} \in \mathcal{F}$:

$$E[\overline{X}\,;\, \overline{X} < 0] \quad = \quad E[X\,;\, \overline{X} < 0] \quad \geq \quad 0,$$

und damit $\overline{X} \geq 0$ P-fast sicher.

(3). Dies folgt unmittelbar aus (1) und (2).

(4). Ist $X_n \geq 0$ und monoton wachsend, dann ist $\sup E[X_n \mid \mathcal{F}]$ eine nichtnegative $\mathcal{F}$-messbare Zufallsvariable (mit Werten in $[0, \infty]$), und nach dem „klassischen " Satz von der monotonen Konvergenz (siehe Satz 6.6) gilt:

$$E[\sup E[X_n \mid \mathcal{F}] \cdot Z] \;=\; \sup E[E[X_n \mid \mathcal{F}] \cdot Z] \;=\; \sup E[X_n \cdot Z] \;=\; E[\sup X_n \cdot Z]$$

für jede nichtnegative $\mathcal{F}$-messbare Zufallsvariable Z. Also ist $\sup E[X_n \mid \mathcal{F}]$ eine Version der bedingten Erwartung von $\sup X_n$ gegeben $\mathcal{F}$.

(5). Wir zeigen, dass jede Version von $E[X \mid \mathcal{G}]$ auch eine Version von $E[E[X \mid \mathcal{F}] \mid \mathcal{G}]$ ist, also die Eigenschaften (i) und (ii) aus der Definition der bedingten Erwartung erfüllt:

(i) $E[X \mid \mathcal{G}]$ ist nach Definition $\mathcal{G}$-messbar.

(ii) Für $A \in \mathcal{G}$ gilt auch $A \in \mathcal{F}$, und somit $E[E[X\,|\mathcal{G}]\,; A] = E[X\,; A] = E[E[X\,|\mathcal{F}]\,; A]$.

(6) und (7). Auf ähnliche Weise verifiziert man, dass die Zufallsvariablen, die auf der rechten Seite der Gleichungen in (6) und (7) stehen, die definierenden Eigenschaften der bedingten Erwartungen auf der linken Seite erfüllen (Übung).

(8). Dies folgt aus (6) und (7) in drei Schritten:

(a) Gilt $f(x, y) = g(x) \cdot h(y)$ mit messbaren Funktionen $g, h \geq 0$, dann folgt nach (6) und (7) P-fast sicher:

$$\begin{aligned} E[f(X,Y) \mid \mathcal{F}] \;&=\; E[g(X) \cdot h(Y) \mid \mathcal{F}] \quad = \quad h(Y) \cdot E[g(X)|\mathcal{F}] \\ &=\; h(Y) \cdot E[g(X)], \end{aligned}$$

und somit

$$E[f(X,Y) \mid \mathcal{F}](\omega) \;=\; E[g(X) \cdot h(Y(\omega))] \;=\; E[f(X, Y(\omega))] \qquad \text{für } P\text{-fast alle } \omega.$$

(b) Um die Behauptung für Indikatorfunktionen $f(x,y) = I_B(x,y)$ von produktmessbaren Mengen B zu zeigen, betrachten wir das Mengensystem

$$\mathcal{D} \quad = \quad \{B \in \mathcal{S} \otimes \mathcal{T} \mid \text{Behauptung gilt für } f = I_B\}.$$

$\mathcal{D}$ ist ein Dynkinsystem, das nach (a) alle Produkte $B = B_1 \times B_2$ mit $B_1 \in \mathcal{S}$ und $B_2 \in \mathcal{T}$ enthält. Also gilt auch

$$\mathcal{D} \quad \supseteq \quad \sigma(\{B_1 \times B_2 \mid B_1 \in \mathcal{S}, B_2 \in \mathcal{T}\}) \quad = \quad \mathcal{S} \otimes \mathcal{T}.$$

(c) Für beliebige produktmessbare Funktionen $f : S \times T \to \mathbb{R}_+$ folgt die Behauptung nun durch maßtheoretische Induktion.

□

Bemerkung (Konvergenzsätze für bedingte Erwartungen). Aus dem Satz von der monotonen Konvergenz (Eigenschaft (4)) folgen auch Versionen des Lemmas von Fatou und des Satzes von der dominierten Konvergenz für bedingte Erwartungen. Der Beweis verläuft ähnlich wie im unbedingten Fall (Übung).

Die letzte Eigenschaft aus Satz 10.10 ist oft sehr nützlich. Für unabhängige Zufallsvariablen X und Y ergibt sich insbesondere

$$E[f(X,Y) \mid Y](\omega) \quad = \quad E[f(X, Y(\omega))] \qquad \text{für } P\text{-fast alle } \omega, \tag{10.4.1}$$

d.h.

$$E[f(X,Y) \mid Y = z] \quad = \quad E[f(X,z)] \qquad \text{für } \mu_Y\text{-fast alle } z. \tag{10.4.2}$$

Die Unabhängigkeit von X und Y ist wesentlich für (10.4.1) bzw. (10.4.2):

Beispiel. Ist $Y = X$, dann gilt offensichtlich

$$\begin{aligned} E[X \cdot Y \mid Y = z] &= E[Y^2 \mid Y = z] \quad = \quad z^2 \qquad \text{für } \mu_Y\text{-fast alle } z, \qquad \text{aber} \\ E[X \cdot z] &= z \cdot E[X] \quad = \quad z \cdot E[Y]. \end{aligned}$$

Das Anwenden der Formeln (10.4.1) und (10.4.2) ohne dass Unabhängigkeit vorliegt ist ein sehr häufiger Fehler beim Rechnen mit bedingten Erwartungen!

Beispiel (Summen von Wartezeiten). Für eine exponential-verteilte Zufallsvariable gilt

$$P[T > t + h | T > t] \quad = \quad P[T > h] \qquad \text{für alle } t \geq 0 \text{ und } h \in \mathbb{R}.$$

Durch Bedingen können wir diese Aussage deutlich verallgemeinern:

Lemma 10.11 (Erweiterte Gedächtnislosigkeit). *Sind T und R unabhängige nichtnegative Zufallsvariablen, und ist T exponentialverteilt, dann gilt*

$$P[T+R>t+h \mid T>t] \quad = \quad P[T+R>h] \qquad \textit{für alle } t\geq 0 \textit{ und } h\in\mathbb{R}.$$

Beweis. Durch Bedingen auf R erhalten wir nach (10.4.2) für $t \geq 0$:

$$\begin{aligned} P[T+R>t+h \text{ und } T>t \mid R=r] &\overset{(*)}{=} P[T+r>t+h \text{ und } T>t] \\ &= P[T>t+h-r \mid T>t]\cdot P[T>t] \\ &= P[T>h-r]\cdot P[T>t] \end{aligned}$$

für fast alle $r>0$, also

$$\begin{aligned} P[T+R>t+h \text{ und } T>t] &= \int P[T+R>t+h, T>t \mid R=r]\,\mu_R(dr) \\ &= \int P[T>h-r]\,\mu_R(dr)\cdot P[T>t] \\ &\overset{(**)}{=} P[T+R>h]\cdot P[T>t]. \end{aligned}$$

Hierbei haben wir in $(*)$ und $(**)$ wesentlich benutzt, dass T und R unabhängig sind. □

Das Lemma zeigt, dass für Summen von unabhängigen Wartezeiten eine Gedächtnislosigkeitseigenschaft gilt, sofern der erste Summand exponentialverteilt ist. Diese Tatsache ist von grundlegender Bedeutung um nachzuweisen, dass die zukünftige Weiterentwicklung von zeitstetigen Markovketten nicht vom Verlauf in der Vergangenheit, sondern nur vom gegenwärtigen Zustand abhängt. Wir betrachten zunächst exemplarisch den einfachsten Fall einer solchen zeitstetigen Markovkette - den Poissonprozess.

Poissonprozesse

Ein Poissonprozess mit Intensität $\lambda>0$ ist ein zeitstetiger stochastischer Prozess, d.h. eine Kollektion $N_t, t\in[0,\infty)$, von Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega,\mathcal{A},P)$, mit nichtnegativen ganzzahligen Werten. Der Prozess wartet jeweils eine $\mathrm{Exp}(\lambda)$-verteilte Zeit ab, und springt dann um eine Einheit nach oben. Naheliegende Anwendungen sind z.B. die Modellierung einer Warteschlange, oder der Anzahl der bei einer Versicherung auflaufenden Schadensfälle.

Um einen Poissonprozess zu konstruieren, wählen wir unabhängige exponentialverteilte Zufallsvariablen $T_1, T_2, \ldots \geq 0$ mit festem Parameter $\lambda>0$ auf einem Wahrscheinlichkeitsraum $(\Omega,\mathcal{A},P)$, und setzen

$$S_n = T_1+T_2+\ldots+T_n, \qquad n\in\mathbb{N}, \qquad \text{und}$$

$$N_t = \#\{n \in \mathbb{N} \mid S_n \leq t\}, \qquad t \in [0, \infty).$$

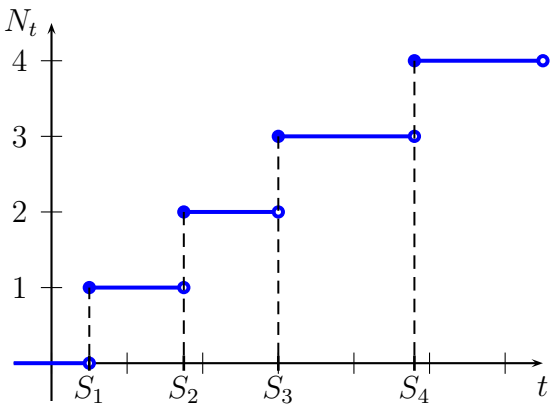


Abbildung 10.7: Darstellung von $N_t(\omega)$.

Dann ist $t \mapsto N_t(\omega)$ für alle ω monoton wachsend mit ganzzahligen Werten und $N_0(\omega) = 0$. Die Wartezeit S_n bis zum n-ten Sprung ist $\Gamma(\lambda, n)$-verteilt, s. Lemma 9.5. Durch Bedingen können wir die Verteilungen des Prozesses $(N_t)_{t\geq 0}$ auf elegante Weise berechnen. Beispielsweise folgt aus der erweiterten Gedächtnislosigkeit (Lemma 10.11) für $t, h \geq 0$ unmittelbar

$$\begin{aligned} P[N_{t+h} < k \mid N_t = 0] &= P[S_k > t + h \mid S_1 > t] \\ &= P[T_1 + T_2 + \ldots + T_k > t + h \mid T_1 > t] \\ &= P[T_1 + T_2 + \ldots + T_k > h] \\ &= P[N_h < k] \quad \text{für alle } k \in \mathbb{N}, \end{aligned}$$

d.h. die bedingte Verteilung von N_{t+h} gegeben $N_t = 0$ stimmt mit der Verteilung von N_h überein. Allgemeiner erhalten wir:

Satz 10.12. *Für $t, h \geq 0$ gilt:*

(1). $N_t \sim \text{Poisson}(\lambda t)$

(2). Stationarität: $N_{t+h} - N_t \sim N_h$

(3). Unabhängige Inkremente: $N_{t+h} - N_t \perp\!\!\!\perp \sigma(N_s \mid 0 \leq s \leq t)$.

Beweis. (1). *Verteilung von N_t:* Da $S_k = T_1 + \ldots + T_k$ unabhängig von T_{k+1} und $\Gamma(\lambda, k)$-verteilt ist, erhalten wir für $k \in \mathbb{N}$ nach (10.4.2):

$$\begin{aligned} P[N_t = k] &= P[S_k \le t < S_{k+1}] \\ &= \int P[S_k \le t < S_k + T_{k+1} \mid S_k = u] : \mu_{S_k}(du) \\ &= \int I_{(0,t]}(u) \cdot P[t < u + T_{k+1}] \, \mu_{S_k}(du) \\ &= \int_0^t e^{-\lambda(t-u)} \cdot \frac{1}{(k-1)!} \lambda^k u^{k-1} e^{-\lambda u} \, du \\ &= \frac{(\lambda t)^k}{k!} e^{-\lambda t}. \end{aligned}$$

Also ist N_t Poisson-verteilt zum Parameter λt.

(2). *Gemeinsame Verteilung von N_t und N_{t+h}:* Seien $k, l \ge 0$. Wegen $S_k = T_1 + \ldots + T_k$ und $S_{k+l} = S_k + T_{k+1} + \ldots + T_{k+l}$ erhalten wir nach (10.4.1) aufgrund der Unabhängigkeit der T_i:

$$\begin{aligned} &P[N_{t+h} < k + l, N_t = k \mid T_1, \ldots, T_k](\omega) \\ &= P[S_{k+l} > t + h, S_k \le t < S_{k+1} \mid T_1, \ldots, T_k](\omega) \\ &= P[S_k(\omega) + T_{k+1} + \ldots + T_{k+l} > t + h, S_k(\omega) \le t < S_k(\omega) + T_{k+1}] \quad (10.4.3) \\ &= P[T_{k+1} + \ldots + T_{k+l} > h] \cdot P[T_{k+1} > t - S_k(\omega)] \cdot I_{\{S_k \le t\}}(\omega) \\ &= P[N_h < l] \cdot P[N_t = k \mid T_1, \ldots, T_k](\omega) \end{aligned}$$

für P-fast alle ω. Hierbei haben wir im vorletzten Schritt Lemma 10.11 verwendet. Aus (a) folgt:

$$\begin{aligned} P[N_{t+h} - N_t < l, N_t = k] &= E[P[N_{t+h} < k + l, N_t = k \mid T_1, \ldots, T_k]] \\ &= P[N_h < l] \cdot P[N_t = k], \qquad (10.4.4) \end{aligned}$$

d.h.

$$P[N_{t+h} - N_t < l \mid N_t = k] \quad = \quad P[N_h < l] \qquad \text{für alle } k, l \ge 0.$$

Also ist das Inkrement $N_{t+h} - N_t$ unabhängig von N_t mit Verteilung

$$P \circ (N_{t-h} - N_t)^{-1} \quad = \quad P \circ N_h^{-1} \quad = \quad \text{Poisson}(\lambda h).$$

(3). *Unabhängigkeit von $N_{t+h} - N_t$ und $\sigma(N_s \mid 0 \le s \le t)$:* Wir bemerken zunächst, dass für jedes Ereignis $A \in \sigma(N_s \mid 0 \le s \le t)$ und $k \ge 0$ ein Ereignis $A_k \in \sigma(T_1, \ldots, T_k)$ existiert mit

$$A \cap \{N_t = k\} \quad = \quad A_k \cap \{N_t = k\}. \tag{10.4.5}$$

Zum Beweis kann man sich auf Ereignisse der Form $A = \{N_s = l\}$ mit $s \in [0,t]$ und $l \ge 0$ beschränken, da diese die σ-Algebra $\sigma(N_s \mid 0 \le s \le t)$ erzeugen. Für solche Ereignisse A gilt in der Tat

$$A \cap \{N_t = k\} = \{N_s = l, N_t = k\} = \{S_l \le s < S_{l+1}, S_k \le t < S_{k+1}\} = A_k \cap \{N_t = k\}$$

wobei

$$A_k \quad := \quad \begin{cases} \emptyset & \text{falls } l > k, \\ \{S_l \le s\} & \text{falls } l = k, \\ \{S_l \le s < S_{l+1}\} & \text{falls } l < k, \end{cases}$$

ein Ereignis ist, dass nur von $T_1, \ldots, T_k$ abhängt.
Nach (10.4.5) erhalten wir für $A \in \sigma(N_s \mid 0 \le s \le t)$ und $k, l \ge 0$ analog zu (10.4.4):

$$\begin{aligned} & P[\{N_{t+h} - N_t < l\} \cap A \cap \{N_t = k\}] \\ &= E[P[N_{t+h} - N_t < l, N_t = k \mid T_1, \ldots, T_k]; A_k] \\ &= P[N_h < l] \cdot P[A_k \cap \{N_t = k\}] \\ &= P[N_{t+h} - N_t < l] \cdot P[A \cap \{N_t = k\}]. \end{aligned}$$

Durch Summieren über k folgt die Unabhängigkeit von $N_{t+h} - N_t$ und A.

□

Aus Satz 10.12 folgt, dass für jede Partition $t_0 < t_1 < \ldots < t_k$ die Inkremente $N_{t_1} - N_{t_0}, N_{t_2} - N_{t_1}, \ldots, N_{t_k} - N_{t_{k-1}}$ unabhängige Zufallsvariablen mit Verteilung

$$N_t - N_s \quad \sim \quad \text{Poisson}(\lambda \cdot (t-s)), \qquad 0 \le s \le t, \tag{10.4.6}$$

sind. Insbesondere sind die Inkremente **stationär**, d.h. die Verteilung von $N_t - N_s$ hängt nur von $t - s$ ab.

Definition. *(1). Ein stochastischer Prozess $(N_t)_{t \ge 0}$ auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ heißt **Lévy-Prozess**, falls*

(a) die Inkremente $N_t - N_s, 0 \le s \le t$, stationär sind, und

(b) Inkremente über disjunkten Intervallen unabhängig sind.

(2). Ein Lévy-Prozess heißt ***Poissonprozess mit Intensität*** $\lambda > 0$*, falls (10.4.6) gilt.*

Weitere wichtige Beispiele von Lévy-Prozessen sind Brownsche Bewegungen und α-stabile Prozesse. Eine *Brownsche Bewegung* $(B_t)_{t\geq 0}$ ist ein Lévy-Prozess mit normalverteilten Inkrementen $B_t - B_s \sim N(0, t-s), 0 \leq s \leq t$, dessen Pfade $t \mapsto B_t(\omega)$ für P-fast alle ω stetig sind.

Prozesse in diskreter Zeit mit unabhängigen stationären Inkrementen sind Random Walks. Lévy-Prozesse kann man aus Random Walks durch Grenzübergänge mit unterschiedlichen Skalierungen erhalten (Poissonapproximation, zentraler Grenzwertsatz, Grenzwertsatz für Inkremente mit heavy tails etc.). Den Poissonprozess erhält man beispielsweise als Grenzwert für $k \to \infty$ der reskalierten Random Walks $N_t^{(k)} = S_{\lfloor kt \rfloor}^{(k)}$,

$$S_n^{(k)} \quad = \quad \sum_{i=1}^{n} X_i^{(k)}, \qquad X_i^{(k)} \text{ unabhängig, } \sim \text{Bernoulli}(\lambda/k).$$

Die Simulation in Abbildung 5.6 deutet an, wie andere Lévyprozesse als Skalierungslimiten von Random Walks auftreten.

Ein weiteres Beispiel für Lévy-Prozesse sind zusammengesetzte (compound) Poissonprozesse:

Beispiel (Compound Poisson-Prozess). Sei μ eine Wahrscheinlichkeitsverteilung auf $\mathbb{R}^d$ und $\lambda > 0$. Dann heißt der stochastische Prozess

$$S_t \quad = \quad \sum_{i=1}^{N_t} X_i, \qquad t \geq 0,$$

mit unabhängigen Zufallsvariablen X_i mit Verteilung μ und einem von den X_i unabhängigen Poissonprozess $(N_t)_{t\geq 0}$ mit Intensität λ, **Compound-Poisson-Prozess mit Sprungverteilung** μ **und Intensität** λ. Der Compound-Poisson-Prozess ist eine zeitstetige Version des Random Walks mit Inkrementen X_i. Er wartet jeweils eine $\text{Exp}(\lambda)$-verteilte Zeit ab, und macht dann einen Sprung gemäß der Verteilung μ. Entsprechende Prozesse werden u. A. in der Versicherungsmathematik zur Modellierung der akkumulierten Schadenshöhe bis zur Zeit t verwendet. Die Verteilung S_t für ein festes $t \geq 0$ kann man mit den oben eingeführten Methoden für zufällige Summen berechnen. Zudem kann man beweisen, dass $(S_t)_{t\geq 0}$ in der Tat ein Prozess mit stationären unabhängigen Inkrementen ist.

Poissonscher Punktprozess

Die Sprungzeitpunkte eines Poissonprozesses in einem endlichen Zeitintervall $(s, t]$ kann man auch auf andere Weise konstruieren: Ist Z eine Poisson-verteilte Zufallsvariable mit Parame-

ter $\lambda \cdot (t-s)$, und sind $U_1, U_2, \ldots$ unabhängig voneinander und von Z, und gleichverteilt auf $(s,t]$, dann sind $U_1, \ldots, U_Z$ die Sprungzeiten eines Poissonprozesses mit Parameter λ (s. Korollar 10.14). Allgemeiner sei nun ν ein endliches Maß auf einem messbaren Raum $(S, \mathcal{S})$. Wir wollen eine zufällige „Punktwolke“ in S mit Intensität ν konstruieren. Dazu wählen wir unabhängige Zufallsvariablen $X_1, X_2, \ldots : \Omega \to S$ mit Verteilung $\mu = \frac{\nu}{\nu(S)}$, und setzen für $A \subseteq S$:

$$N(A) \quad = \quad \sum_{i=1}^{Z} \delta_{X_i}[A] \quad = \quad \#\{1 \le i \le Z | X_i \in A\}, \tag{10.4.7}$$

wobei Z (Gesamtzahl der Punkte) unabhängig von den X_i und Poisson-verteilt mit Parameter $\nu(S)$ ist. Die Abbildung $A \mapsto N(A)$ ist die Häufigkeitsverteilung der Punkte $X_1, \ldots, X_Z$, und damit ein zufälliges Maß. Hat das Intensitätsmaß ν keine Atome (d.h. gilt $\nu[\{x\}] = 0$ für alle $x \in S$), dann sind die Punkte X_i mit Wahrscheinlichkeit 1 alle verschieden, und wir können N P-fast sicher mit der zufälligen Punktmenge $\{X_1, X_2, \ldots, X_Z\} \subseteq S$ identifizieren.

Satz 10.13 (Konstruktion von Poissonschen Punktprozessen). *Das durch (10.4.7) definierte zufällige Maß N ist ein **Poissonscher Punktprozess mit Intensitätsmaß** ν, d.h. für beliebige $k \in \mathbb{N}$ und disjunkte Teilmenge $A_1, \ldots, A_k \subseteq S$, sind die Zufallsvariablen $N(A_1), \ldots, N(A_k)$ unabhängig mit Verteilung*

$$N(A_i) \quad \sim \quad \text{Poisson}(\nu(A_i)).$$

Zum Beweis benötigen wir die erzeugende Funktion der gemeinsamen Verteilung mehrerer Zufallsvariablen:

Definition (Erzeugende Funktion und gemeinsame Verteilung). *Seien $N_1, \ldots, N_k : \Omega \to \{0, 1, 2, \ldots\}$ nichtnegative ganzzahlige Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$, und sei*

$$\nu(n_1, \ldots, n_k) \quad = \quad P[N_1 = n_1, \ldots, N_k = n_k].$$

*Die **erzeugende Funktion** des Zufallsvektors $(N_1, \ldots, N_n)$ bzw. der Wahrscheinlichkeitsverteilung ν auf $\{0, 1, 2, \ldots\}^k$ ist die durch*

$$G(s_1, \ldots, s_k) \quad = \quad E[s_1^{N_1} s_2^{N_2} \cdot \ldots \cdot s_k^{N_k}] \quad = \quad \sum_{n_1, \ldots, n_k = 0}^{\infty} \nu(n_1, \ldots, n_k) \cdot s_1^{n_1} s_2^{n_2} \cdot \ldots \cdot s_k^{n_k}$$

definierte Funktion $G : [0,1]^k \to [0,1]$.

Die gemeinsame Verteilung ν ist ähnlich wie im eindimensionalen Fall eindeutig durch die erzeugende Funktion festgelegt, denn für $n_1, \ldots, n_k \in \{0, 1, 2, \ldots\}$ gilt:

$$\nu(n_1, \ldots, n_k) \quad = \quad \frac{1}{n_1! \cdot \ldots \cdot n_k!} \cdot \frac{\partial^{n_1 + n_2 + \ldots + n_k}}{\partial_{s_1}^{n_1} \cdot \ldots \cdot \partial_{s_k}^{n_k}}(0, \ldots, 0).$$

Beweis. O.B.d.A. können wir $S = \bigcup_{i=1}^{k} A_i$ annehmen. Wir berechnen für diesen Fall die erzeugende Funktion der gemeinsamen Verteilung von $N(A_1), \ldots, N(A_k)$. Für $s_1, \ldots, s_k \in [0,1)$ gilt

$$\prod_{j=1}^{k} s_j^{N(A_j)} \quad = \quad \prod_{i=1}^{Z} \prod_{j=1}^{k} s_j^{I_{A_j}(X_i)},$$

also wegen der Unabhängigkeit von Z und den X_i:

$$E\left[\prod_{j=1}^{k} s_j^{N(A_j)} \,\middle|\, Z\right] \quad = \quad \prod_{i=1}^{Z} E\left[\prod_{j=1}^{k} s_j^{I_{A_j}(X_i)}\right] \quad = \quad \left(\sum_{j=1}^{k} s_j \cdot \mu[A_j]\right)^{Z}.$$

Hierbei haben wir im letzten Schritt verwendet, dass das Produkt über j gleich s_j ist, falls X_i in der Menge A_j liegt. Da Z Poisson-verteilt ist mit Parameter $\nu(S)$, erhalten wir

$$\begin{aligned} E\left[\prod_{j=1}^{k} s_j^{N(A_j)}\right] &= G_Z\left(\sum_{j=1}^{k} s_j \cdot \mu[A_j]\right) \\ &= \exp\left(\nu(S) \cdot \left(\sum_{j=1}^{k} s_j \cdot \mu[A_j] - 1\right)\right) \\ &= \prod_{j=1}^{k} \exp\left(\nu(A_j) \cdot (s_j - 1)\right), \end{aligned}$$

d.h. die erzeugende Funktion von $(N(A_1), \ldots, N(A_k))$ ist das Produkt der erzeugenden Funktionen von Poissonverteilungen mit Parametern $\nu(A_j)$. Hieraus folgt, dass die gemeinsame Verteilung der Zufallsvariablen $N(A_1), \ldots, N(A_k)$ das Produkt dieser Poissonverteilungen ist. □

Poissonsche Punktprozesse bezeichnet man auch synonym als *räumliche Poissonprozesse, Poissonsche Zufallsmaße,* oder *Poissonsche Felder*. Sie spielen eine wichtige Rolle bei der Modellierung zufälliger räumlicher Strukturen, z.B. in der stochastischen Geometrie. Satz 10.13 liefert uns einen einfachen Algorithmus zur Simulation Poissonscher Punktprozesse. Graphik **??** wurde mit diesem Algorithmus erzeugt. Als eindimensionalen Spezialfall von Satz 10.13 erhalten wir eine alternative Konstruktion von zeitlichen Poissonprozessen:

Korollar 10.14. *Seien $\lambda, a \in (0,\infty)$. Sind $Z, U_1, U_2, \ldots$ unabhängige Zufallsvariablen mit Verteilungen $Z \sim \mathrm{Poisson}(\lambda \cdot a)$ und $U_1, U_2, \ldots \sim \mathrm{Unif}_{(0,a)}$, dann ist*

$$N_t \quad := \quad \sum_{i=1}^{Z} I_{[0,t]}(U_i), \qquad 0 \le t \le a,$$

ein Poissonprozess mit Intensität λ.

Beweis. Es gilt $N_t = \overline{N}([0,t])$, wobei $\overline{N}$ der wie in (10.4.7) definierte Poissonsche Punktprozess auf $S = [0,a]$ mit homogenem Intensitätsmaß $\lambda \cdot dt$ ist. Nach Satz 10.13 folgt, dass für jede Partition $0 \leq t_0 < t_1 < \ldots < t_k \leq a$ die Inkremente

$$N_{t_j} - N_{t_{j-1}} \quad = \quad \overline{N}((t_{j-1}, t_j]), \qquad 1 \leq j \leq k,$$

unabhängig und Poisson$(\lambda \cdot (t_j - t_{j-1}))$-verteilt sind. □

Poissonsche Punktprozesse lassen sich durch verschiedene Transformationen wieder in Poissonsche Punktprozesse überführen. Bildet man beispielsweise die Punkte $X_i, 1 \leq i \leq Z$, eines Poissonschen Punktprozesses N mit Intensitätsmaß ν mit einer (messbaren) Abbildung ϕ ab, dann erhält man einen Poissonschen Punktprozess

$$\widetilde{N}(A) \quad := \quad \sum_{i=1}^{Z} \delta_{\phi(X_i)}[A] \quad = \quad \sum_{i=1}^{Z} I_A(\phi(X_i)) \quad = \quad \sum_{i=1}^{Z} I_{\phi^{-1}(A)}(X_i)$$

mit Intensitätsmaß $\widetilde{\nu} = \nu \circ \phi^{-1}$. Zudem gilt eine Ausdünnungseigenschaft:
Seien $Z, X_1, X_2, \ldots, U_1, U_2, \ldots$ unabhängige Zufallsvariablen mit Verteilungen

$$Z \sim \text{Poisson}(\nu(S)), \quad X_i \sim \frac{\nu}{\nu(S)}, \quad U_i \sim \text{Unif}_{(0,1)},$$

und sei $\alpha : S \to [0,1]$ eine messbare Funktion (Akzeptanzwahrscheinlichkeit). Wir konstruieren einen ausgedünnten Punktprozess N_α, indem wir einen Punkt X_i nur mit Wahrscheinlichkeit $\alpha(X_i)$ berücksichtigen:

$$N_\alpha \quad := \quad \sum_{i=1}^{Z} I_{\{U_i \leq \alpha(X_i)\}} \delta_{X_i}.$$

Satz 10.15 (Färbungssatz, Ausdünnungseigenschaft). *N_α ist ein Poissonscher Punktprozess mit Intensitätsmaß $\alpha(x)\nu(dx)$.*

Der Beweis wird dem Leser als Übung überlassen. Bemerkenswert ist unter Anderem, dass die beschriebene Konstruktion eine Kopplung von Poissonprozessen mit verschiedenen Intensitätsmaßen, d.h. eine simultane Konstruktion dieser Prozesse auf einem gemeinsamen Wahrscheinlichkeitsraum ermöglicht.

10.5 Bedingte Erwartung als beste L^2-Approximation

In diesem Abschnitt zeigen wir, dass sich die bedingte Erwartung einer quadratintegrierbaren Zufallsvariable X gegeben eine σ-Algebra $\mathcal{F}$ charakterisieren lässt als beste Approximation von

X im Unterraum der $\mathcal{F}$-messbaren quadratintegrierbaren Zufallsvariablen, bzw. als orthogonale Projektion von X auf diesen Unterraum. Neben naheliegenden Anwendungen auf nichtlineare Prognosen liefert uns dies auch einen einfachen Existenzbeweis für die bedingte Erwartung.

Jensensche Ungleichung

Die Jensensche Ungleichung gilt auch für bedingte Erwartungen.
Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum, $X \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ eine integrierbare Zufallsvariable und $\mathcal{F} \subseteq \mathcal{A}$ eine σ-Algebra.

Satz 10.16 (Jensen). *Ist $u : \mathbb{R} \to \mathbb{R}$ eine konvexe Funktion mit $u(X) \in \mathcal{L}^1$ oder $u \geq 0$, dann gilt*

$$E[u(X) \mid \mathcal{F}] \quad \geq \quad u(E[X \mid \mathcal{F}]) \qquad P\text{-fast sicher.}$$

Beweis. Jede konvexe Funktion u lässt sich als Supremum von abzählbar vielen affinen Funktionen darstellen, d.h. es gibt $a_n, b_n \in \mathbb{R}$ mit

$$u(x) \quad = \quad \sup_{n \in \mathbb{N}} (a_n x + b_n) \qquad \text{für alle } x \in \mathbb{R}.$$

Zum Beweis betrachtet man die Stützgeraden an allen Stellen einer abzählbaren dichten Teilmenge von $\mathbb{R}$, siehe z.B. [Williams: Probability with martingales, 6.6]. Wegen der Monotonie und Linearität der bedingten Erwartung folgt

$$E[u(X) \mid \mathcal{F}] \quad \geq \quad E[a_n X + b_n \mid \mathcal{F}] \quad = \quad a_n \cdot E[X \mid \mathcal{F}] + b_n$$

P-fast sicher für alle $n \in \mathbb{N}$, also auch

$$E[u(X) \mid \mathcal{F}] \quad \geq \quad \sup_{n \in \mathbb{N}} (a_n \cdot E[X \mid \mathcal{F}] + b_n) \qquad P\text{-fast sicher.}$$

□

Korollar 10.17 (L^p-Kontraktivität). *Die Abbildung $X \mapsto E[X \mid \mathcal{F}]$ ist eine Kontraktion auf $\mathcal{L}^p(\Omega, \mathcal{A}, P)$ für alle $p \geq 1$, d.h.*

$$E\left[|E[X \mid \mathcal{F}]|^p\right] \quad \leq \quad E[|X|^p] \qquad \textit{für alle } X \in \mathcal{L}^1(\Omega, \mathcal{A}, P).$$

Beweis. Nach der Jensenschen Ungleichung gilt:

$$|E[X \mid \mathcal{F}]|^p \quad \leq \quad E[|X|^p \mid \mathcal{F}] \qquad P\text{-fast sicher.}$$

Die Behauptung folgt durch Bilden des Erwartungswertes. □

Im Beweis des Korollars haben wir insbesondere gezeigt, dass für eine Zufallsvariable $X \in \mathcal{L}^p$ auch die bedingte Erwartung $E[X \mid \mathcal{F}]$ in $\mathcal{L}^p$ enthalten ist. Wir beschränken uns nun auf den Fall $p = 2$.

Bedingte Erwartung als beste L^2-Prognose

Der Raum $L^2(\Omega, \mathcal{A}, P) = \mathcal{L}^2(\Omega, \mathcal{A}, P)/\sim$ der Äquivalenzklassen von quadratintegrierbaren Zufallsvariablen ist ein Hilbertraum mit Skalarprodukt $(X, Y)_{L^2} = E[XY]$. Ist $\mathcal{F} \subseteq \mathcal{A}$ eine Unter-σ-Algebra, dann ist $L^2(\Omega, \mathcal{F}, P)$ ein **abgeschlossener Unterraum** von $L^2(\Omega, \mathcal{A}, P)$, denn Grenzwerte von $\mathcal{F}$-messbaren Zufallsvariablen sind wieder $\mathcal{F}$-messbar. Nach der Jensenschen Ungleichung ist für $X \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$ jede Version der bedingten Erwartung $E[X \mid \mathcal{F}]$ im Unterraum $\mathcal{L}^2(\Omega, \mathcal{F}, P)$ der $\mathcal{F}$-messbaren quadratintegrierbaren Zufallsvariablen enthalten. Außerdem respektiert die bedingte Erwartung Äquivalenzklassen, s. Satz 10.7. Die Zuordnung $X \mapsto E[X \mid \mathcal{F}]$ definiert also eine lineare Abbildung vom Hilbertraum $L^2(\Omega, \mathcal{A}, P)$ der Äquivalenzklassen auf den Unterraum $L^2(\Omega, \mathcal{F}, P)$.

Satz 10.18. *Für $Y \in \mathcal{L}^2(\Omega, \mathcal{F}, P)$ sind äquivalent:*

(1). Y ist eine Version der bedingten Erwartung $E[X \mid \mathcal{F}]$.

*(2). Y ist eine „**beste Approximation**“ von X im Unterraum $\mathcal{L}^2(\Omega, \mathcal{F}, P)$, d.h.*

$$E[(X-Y)^2] \quad \leq \quad E[(X-Z)^2] \qquad \text{für alle } Z \in \mathcal{L}^2(\Omega, \mathcal{F}, P).$$

*(3). Y ist eine Version der **orthogonalen Projektion** von X auf den Unterraum $L^2(\Omega, \mathcal{F}, P) \subseteq L^2(\Omega, \mathcal{A}, P)$, d.h.*

$$E[(X-Y)\cdot Z] \quad = \quad 0 \qquad \text{für alle } Z \in \mathcal{L}^2(\Omega, \mathcal{F}, P).$$

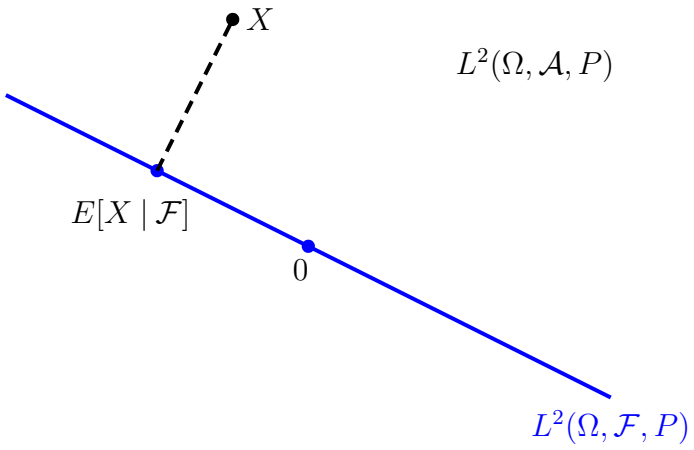


Abbildung 10.8: Darstellung von $X \mapsto E[X \mid \mathcal{F}]$ als orthogonale Projektion auf den Unterraum $L^2(\Omega, \mathcal{F}, P)$.

Beweis. **(1)** $\iff$ **(3):** Für $Y \in \mathcal{L}^2(\Omega, \mathcal{F}, P)$ gilt:

$$\begin{aligned} & Y \text{ ist eine Version von } E[X \mid \mathcal{F}] \\ \iff \quad & E[Y \cdot I_A] \;=\; E[X \cdot I_A] \quad \text{für alle } A \in \mathcal{F} \\ \iff \quad & E[Y \cdot Z] \;=\; E[X \cdot Z] \quad \text{für alle } Z \in \mathcal{L}^2(\Omega, \mathcal{F}, P) \\ \iff \quad & E[(X - Y) \cdot Z] \;=\; 0 \quad \text{für alle } Z \in \mathcal{L}^2(\Omega, \mathcal{F}, P) \end{aligned}$$

Hierbei zeigt man die zweite Äquivalenz mit den üblichen Fortsetzungsverfahren (maßtheoretische Induktion).

(3) $\Rightarrow$ **(2):** Sei Y eine Version der orthogonalen Projektion von X auf $L^2(\Omega, \mathcal{F}, P)$. Dann gilt für alle $Z \in \mathcal{L}^2(\Omega, \mathcal{F}, P)$:

$$\begin{aligned} E[(X - Z)^2] &= E[((X - Y) + (Y - Z))^2] \\ &= E[(X - Y)^2] + E[(Y - Z)^2] + 2E[(X - Y) \underbrace{(Y - Z)}_{\in \mathcal{L}^2(\Omega, \mathcal{F}, P)}] \\ &\geq E[(X - Y)^2] \end{aligned}$$

Hierbei haben wir im letzten Schritt verwendet, dass $Y - Z$ im Unterraum $\mathcal{L}^2(\Omega, \mathcal{F}, P)$ enthalten, also orthogonal zu $X - Y$ ist.

(2) $\Rightarrow$ **(3):** Ist umgekehrt Y eine beste Approximation von X in $\mathcal{L}^2(\Omega, \mathcal{F}, P)$ und $Z \in \mathcal{L}^2(\Omega, \mathcal{F}, P)$, dann gilt

$$\begin{aligned} E[(X - Y)^2] &\leq E[(X - Y + tZ)^2] \\ &= E[(X - Y)^2] + 2tE[(X - Y)Z] + t^2 E[Z^2] \end{aligned}$$

für alle $t \in \mathbb{R}$, also $E[(X - Y) \cdot Z] = 0$.

□

Die Äquivalenz von (2) und (3) ist eine bekannte funktionalanalytische Aussage: die beste Approximation eines Vektors in einem abgeschlossenen Unterraum eines Hilbertraums ist die orthogonale Projektion des Vektors auf diesen Unterraum. Die dahinterstehende geometrische Intuition verdeutlicht man sich leicht anhand von Abbildung 10.8.

Satz 10.18 rechtfertigt die Verwendung der bedingten Erwartung als Prognoseverfahren. Beispielsweise ist $E[X \mid Y]$ nach dem Faktorisierungslemma die beste L^2-Prognose für X unter allen Funktionen vom Typ $g(Y), g : \mathbb{R} \to \mathbb{R}$ messbar.

Beispiel (Nichtlineare Prognose). Seien $S, T : \Omega \to \mathbb{R}_+$ unabhängige Zufallsvariablen, die zum Beispiel die Ausfallzeiten zweier Komponenten eines Systems beschreiben. S sei exponentialverteilt mit Parameter $\lambda > 0$ - die Verteilung von T ist beliebig. Angenommen, wir können nur den Ausfall der einen Komponente (mit Ausfallzeit T) beobachten, und wir möchten den Wert der ersten Ausfallzeit

$$X \quad = \quad \min(T, S)$$

aufgrund des beobachteten Wertes $T(\omega)$ prognostizieren. Nach Satz 10.18 ist der beste Prognosewert für X bzgl. des mittleren quadratischen Fehlers durch

$$\hat{X}(\omega) \quad = \quad E[X \mid T](\omega)$$

gegeben. Explizit erhalten wir wegen der Unabhängigkeit von T und S:

$$\begin{aligned} E[X \mid T](\omega) &= E[\min(T(\omega), S)] \\ &= \int_0^\infty \min(T(\omega), s)\lambda e^{-\lambda s}\, ds \\ &= \int_0^{T(\omega)} s\lambda e^{-\lambda s}\, ds + \int_{T(\omega)}^\infty T(\omega)\lambda e^{-\lambda s}\, ds \\ &= \frac{1}{\lambda}(1 - e^{-\lambda T(\omega)}) \qquad \text{für } P\text{-fast alle } \omega. \end{aligned}$$

Die beste Prognose im quadratischen Mittel hängt also in diesem Fall *nichtlinear* von T ab. Sie unterscheidet sich damit von der *besten linearen Prognose* (Regressionsgerade), die wie in Abschnitt 6.3 gezeigt durch

$$\hat{X}_{\text{lin}} \;=\; aT + b \qquad \text{mit} \quad a \;=\; \frac{\text{Cov}[X,T]}{\text{Var}[T]}, \quad b \;=\; E[X] - aE[T]$$

gegeben ist. Dass sich $\hat{X}$ und $\hat{X}_{\text{lin}}$ unterscheiden ist die Regel. Eine wichtige Ausnahme ergibt sich, wenn die gemeinsame Verteilung von X und T eine Gaußverteilung ist - in diesem Fall ist die beste L^2 Prognose $E[X \mid T]$ stets eine affine Funktion von T.

Existenz der bedingten Erwartung

Durch die Charakterisierung der bedingten Erwartung als beste L^2-Approximation ergibt sich die Existenz der bedingten Erwartung einer quadratintegrierbaren Zufallsvariable unmittelbar aus der Existenz der Bestapproximation eines Vektors in einem abgeschlossenen Unterraum eines Hilbertraums. Durch monotone Approximation folgt hieraus die Existenz der bedingten Erwartung auch für beliebige nichtnegative bzw. integrierbare Zufallsvariablen:

Satz 10.19. *Für jede Zufallsvariable $X \geq 0$ bzw. $X \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ und jede σ-Algebra $\mathcal{F} \subseteq \mathcal{A}$ existiert eine Version der bedingten Erwartung $E[X \mid \mathcal{F}]$.*

Beweis. (1). Wir betrachten zunächst den Fall $X \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$. Wie eben bemerkt, ist der Raum $L^2(\Omega, \mathcal{F}, P)$ ein abgeschlossener Unterraum des Hilbertraums $L^2(\Omega, \mathcal{A}, P)$. Sei $d = \inf\{\|Z - X\|_{L^2} \mid Z \in \mathcal{L}^2(\Omega, \mathcal{F}, P)\}$ der Abstand von X zu diesem Unterraum. Um zu zeigen, dass eine beste Approximation von X in $L^2(\Omega, \mathcal{F}, P)$ existiert, wählen wir eine Folge (X_n) aus diesem Unterraum mit $\|X_n - X\|_{L^2} \to d$. Mithilfe der Parallelogramm-Identität folgt für $n, m \in \mathbb{N}$:

$$\begin{aligned}
\|X_n - X_m\|_{L^2}^2 &= \|(X_n - X) - (X_m - X)\|_{L^2}^2 \\
&= 2 \cdot \|X_n - X\|_{L^2}^2 + 2 \cdot \|X_m - X\|_{L^2}^2 - \|(X_n - X) + (X_m - X)\|_{L^2}^2 \\
&= 2 \cdot \underbrace{\|X_n - X\|_{L^2}^2}_{\to d^2} + 2 \cdot \underbrace{\|X_m - X\|_{L^2}^2}_{\to d^2} - 4 \underbrace{\left\| \frac{X_n + X_m}{2} - X \right\|_{L^2}^2}_{\leq d^2},
\end{aligned}$$

und damit

$$\limsup_{n,m \to \infty} \|X_n - X_m\|_{L^2}^2 \quad \leq \quad 0.$$

Also ist die Minimalfolge (X_n) eine CauchyLfolge in dem vollständigen Raum $L^2(\Omega, \mathcal{F}, P)$, d.h. es existiert ein $Y \in \mathcal{L}^2(\Omega, \mathcal{F}, P)$ mit

$$\|X_n - Y\|_{L^2} \quad \to \quad 0.$$

Für Y gilt

$$\|Y - X\|_{L^2} \quad = \quad \| \lim_{n \to \infty} X_n - X\|_{L^2} \quad \leq \quad \liminf_{n \to \infty} \|X_n - X\|_{L^2} \quad \leq \quad d,$$

d.h. Y ist die gesuchte Bestapproximation, und damit eine Version der bedingten Erwartung $E[X \mid \mathcal{F}]$.

(2). Für eine beliebige nichtnegative Zufallsvariable X auf $(\Omega, \mathcal{A}, P)$ existiert eine monoton wachsende Folge (X_n) nichtnegativer quadratintegrierbarer Zufallsvariablen mit $X = \sup X_n$. Man verifiziert leicht, dass $\sup_n E[X_n \mid \mathcal{F}]$ eine Version von $E[X \mid \mathcal{F}]$ ist.

(3). Entsprechend verifiziert man, dass für allgemeine $X \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ durch $E[X \mid \mathcal{F}] = E[X^+ \mid \mathcal{F}] - E[X^- \mid \mathcal{F}]$ eine Version der bedingten Erwartung gegeben ist.

□

Kapitel 11

Markovketten

In diesem Kapitel werden wir Markovketten genauer untersuchen. Ein wichtiges Hilfsmittel dabei ist der Zusammenhang von Markovketten und Differenzengleichungen.

11.1 Grundlagen

Sei $(S, \mathcal{S})$ ein messbarer Raum. Eine Folge $X_0, X_1, \ldots$ von auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ definierten Zufallsvariablen $X_n : \Omega \to S$ heißt **(zeitdiskreter) stochastischer Prozess mit Zustandsraum** S. Den Index „n“ interpretieren wir entsprechend als „Zeit.“ Für $m \leq n$ setzen wir:

$$X_{m:n} \quad := \quad (X_m, X_{m+1}, \ldots, X_n).$$

Seien nun $p_n(x, dy), n = 1, 2, 3, \ldots,$ stochastische Kerne auf $(S, \mathcal{S})$. Wir verwenden die Notation

$$(p_n f)(x) \quad := \quad \int p_n(x, dy) f(y)$$

für den Erwartungswert einer messbaren Funktion $f : S \to \mathbb{R}$ bzgl. der Wahrscheinlichkeitsverteilung $p_n(x, \bullet)$. Insbesondere gilt

$$(p_n I_A)(x) \quad = \quad p_n(x, A) \qquad \text{für alle } A \in \mathcal{S}.$$

Definition. *Ein stochastischer Prozess (X_n) mit Zustandsraum S heißt **Markovkette mit Übergangswahrscheinlichkeiten** $p_n(x, dy)$, falls gilt:*

$$P[X_{n+1} \in A \mid X_{0:n}] \;=\; p_{n+1}(X_n, A) \qquad P\text{-f.s.} \quad \textit{für alle } A \in S \textit{ und } n \geq 0, \tag{11.1.1}$$

bzw. dazu äquivalent

$$E[f(X_{n+1}) \mid X_{0:n}] \;=\; (p_{n+1} f)(X_n) \qquad P\text{-f.s.} \quad \textit{für alle } \mathcal{S}\textit{-messbaren } f : S \to \mathbb{R}_+ \textit{ und } n \geq 0. \tag{11.1.2}$$

*Die Markovkette heißt **zeitlich homogen**, falls p_n nicht von n abhängt. Die Verteilung von X_0 heißt **Startverteilung** der Markovkette. Gilt $P \circ X_0^{-1} = \delta_x$, dann sagen wir, die **Markovkette startet in** x.*

Die Äquivalenz von (11.1.1) und (11.1.2) ergibt sich durch maßtheoretische Induktion. Die definierende Eigenschaft (11.1.1) besagt, dass bedingt auf X_n der nächste Zustand X_{n+1} unabhängig von $X_0, \dots, X_{n-1}$ mit Verteilung $p_{n+1}(X_n, \bullet)$ ist. Eine Markovkette „vergisst“ also den vorherigen Verlauf bis zur Zeit $n-1$, und startet in jedem Schritt neu im gegenwärtigen Zustand X_n.

Bemerkung. Allgemeiner heißt ein stochastischer Prozess (X_n) Markovkette, falls

$$P[X_{n+1} \in A \mid X_{0:n}] \;=\; P[X_{n+1} \in A \mid X_n] \qquad P\text{-f.s. für alle } A \in \mathcal{S} \text{ und } n \geq 0 \qquad (11.1.3)$$

gilt. Die Existenz eines Übergangskerns folgt aus (11.1.3) unter Regularitätsvoraussetzungen an $(S, \mathcal{S})$, z.B. falls S ein polnischer (d.h. vollständiger separabler metrischer) Raum ist mit Borelscher σ-Algebra $\mathcal{S} = \mathcal{B}(S)$.

Beispiel (Diskreter Zustandsraum). Ist S abzählbar, dann können wir einen stochastischen Kern p_n auf S mit der stochastischen Matrix $p_n(x,y) = p_n(x, \{y\})$ identifizieren. Ein stochastischer Prozess (X_n) ist genau dann eine Markovkette mit Übergangsmatrizen $p_n(x,y)$, wenn

$$P[X_{n+1} = x_{n+1} \mid X_{0:n} = x_{0:n}] \quad = \quad p_{n+1}(x_n, x_{n+1})$$

für alle $x_0, \dots, x_{n+1} \in S$ mit $P[X_{0:n} = x_{0:n}] \neq 0$ gilt.

Zufällige dynamische Systeme als Markovketten, Beispiele

Markovketten erhält man insbesondere als zufällige Störungen dynamischer Systeme.

Sei $(T, \mathcal{T})$ ein messbarer Raum. Wir betrachten einen stochastischen Prozess (X_n) mit Zustandsraum S, der rekursiv durch

$$X_{n+1} \quad = \quad \Phi_{n+1}(X_n, W_{n+1}), \qquad n = 0, 1, 2, \dots,$$

definiert ist, wobei $X_0 : \Omega \to S$ und $W_1, W_2, \dots; \Omega \to T$ unabhängige Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$, und $\Phi : S \times T \to S, n \in \mathbb{N}$, messbare Abbildungen sind. Die Abbildungen Φ_n beschreiben das Bewegungsgesetz des dynamischen Systems, und die Zufallsvariablen W_n die zufälligen Einflussfaktoren (Rauschen, noise).

Satz 11.1. *(1). (X_n) ist eine Markovkette mit Übergangswahrscheinlichkeiten*

$$p_n(x, A) \quad = \quad P[\Phi_n(x, W_n) \in A], \qquad x \in S, A \in \mathcal{S}.$$

(2). Hängen die Abbildungen Φ_n nicht von n ab, und sind die Zufallsvariablen W_n identisch verteilt, dann ist die Markovkette (X_n) zeitlich homogen.

Beweis. (1). Für $n \geq 0$ ist $X_{0:n}$ eine Funktion von $X_0, W_1, W_2, \ldots, W_n$. Also ist W_{n+1} unabhängig von $X_{0:n}$, und für $A \in \mathcal{S}$ folgt

$$\begin{aligned} P[X_{n+1} \in A \mid X_{0:n}](\omega) &= P[\Phi_{n+1}(X_n, W_{n+1}) \in A \mid X_{0:n}](\omega) \\ &= P[\Phi_{n+1}(X_n(\omega), W_{n+1}) \in A] \quad = \quad p_{n+1}(X_n(\omega), A) \end{aligned}$$

für P-fast alle $\omega \in \Omega$.

(2). Hängen Φ_n und die Verteilung von W_n nicht von n ab, dann hängt auch p_n nicht von n ab, d.h. die Markovkette ist zeitlich homogen.

□

Beispiel. (1). *Random Walks auf $\mathbb{Z}^d$ bzw. $\mathbb{R}^d$:* Sind die Zufallsvariablen W_n unabhängig und identisch verteilt mit Werten in $\mathbb{Z}^d$ oder $\mathbb{R}^d$, dann wird durch

$$X_{n+1} \quad = \quad X_n + W_{n+1}, \quad X_0 = x,$$

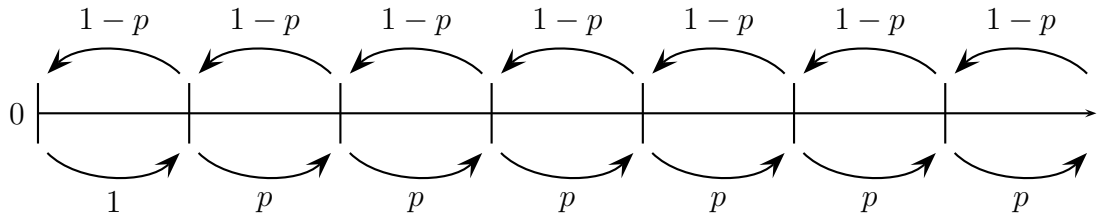
ein d-dimensionaler Random Walk definiert. $(X_n)_n$ ist eine zeitlich homogene Markovkette mit Start in x und Übergangskern $p(x, \bullet) = \mu \circ \tau_x^{-1}$, wobei μ die Verteilung von W_n und $\tau_x(y) = y + x$ die Translation um x ist.

(2). *Random Walk auf $\{0, 1, 2, \ldots\}$ mit Reflexion bzw. Absorption bei 0*: Durch

$$X_{n+1} \quad = \quad \begin{cases} X_n + W_{n+1} & \text{falls } X_n > 0 \\ 1 \text{ bzw. } 0 & \text{falls } X_n = 0 \end{cases}$$

mit unabhängigen, identisch verteilten Zufallsvariablen W_n mit $P[W_n = 1] = p$ und $P[W_n = -1] = 1 - p, p \in [0, 1]$, wird ein Random Walk auf $\{0, 1, 2, \ldots\}$ definiert, der bei 0 reflektiert bzw. absorbiert wird. (X_n) ist eine zeitlich homogene Markovkette mit Übergangswahrscheinlichkeiten wie in Graphik 11.1 dargestellt.

Random Walk mit Reflexion bei 0.



Random Walk mit Absorption bei 0.

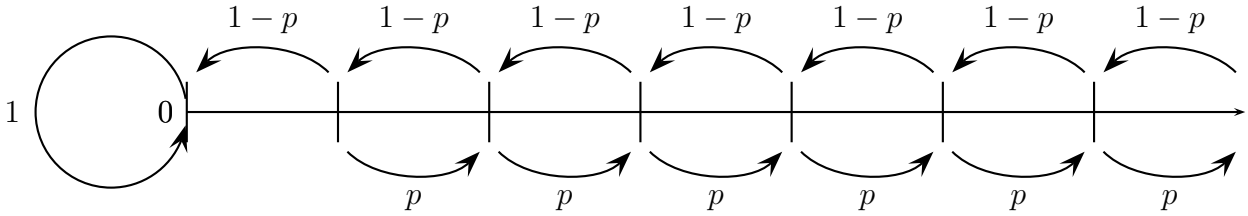


Abbildung 11.1: Darstellung der Übergangswahrscheinlichkeiten von Random Walks auf $\{0, 1, 2, \ldots\}$ mit Reflexion bzw. Absorption in 0.

(3). *Warteschlange mit einem Server:* In einer einfachen Warteschlange wird pro Zeiteinheit ein Kunde bedient, während A_n neue Kunden ankommen. Die Anzahlen A_n der Ankünfte in einer Bedienzeit sind unabhängige Zufallsvariablen mit Werten in $\{0, 1, 2, \ldots\}$. Die Zahl X_n der wartenden Kunden ist dann eine Markovkette mit Übergangsmechanismus

$$X_{n+1} \quad = \quad (X_n - 1 + A_{n+1})_+.$$

(4). *Autoregressive Prozesse:* Ein $AR(p)$-Prozess mit Parametern $\varepsilon, \alpha_1, \ldots, \alpha_p \in \mathbb{R}$ ist durch die Rekursionsformel

$$X_n = \sum_{i=1}^{p} \alpha_i X_{n-i} + \varepsilon \cdot W_n, \qquad n \geq p,$$

mit unabhängigen, standardnormalverteilten Zufallsvariablen W_n gegeben. Für $p = 1$ ergibt sich eine zeithomogene Markovkette mit Übergangskern

$$p(x, \cdot) \quad = \quad N(\alpha_1 x, \varepsilon^2).$$

Für $p \geq 2$ und $\alpha_p, \varepsilon \neq 0$ ist der $AR(p)$-Prozess dagegen keine Markovkette, da der nächste Zustand nicht nur vom gegenwärtigen Zustand, sondern auch vom vorherigen Verlauf abhängt. Wir können jedoch eine Markovkette erhalten, indem wir statt X_n die aus den letzten p Zuständen gebildeten Vektoren

$$\overline{X}_n \quad = \quad (X_n, X_{n-1}, \ldots, X_{n-p+1}), \qquad n = p-1, p, p+1, \ldots,$$

betrachten. $(\overline{X}_n)$ ist eine zeithomogene Markovkette mit Zustandsraum S^p, denn für $n \geq p$ gilt

$$\overline{X}_n \quad = \quad \begin{pmatrix} \alpha_1 & \alpha_2 & \alpha_3 & \cdots & \alpha_p \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & \cdots & 0 & 1 & 0 \end{pmatrix} \overline{X}_{n-1} + \varepsilon \cdot \begin{pmatrix} W_n \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

(5). *Galton-Watson-Verzweigungsprozesse:* Der Galton-Watson-Prozess ist eine zeithomogene Markovkette auf $S = \{0, 1, 2, \ldots\}$, denn für $n \geq 0$ gilt

$$Z_n = \sum_{i=1}^{Z_{n-1}} N_i^n$$

mit unabhängigen, identisch verteilten Zufallsvariablen N_i^n $(i, n \in \mathbb{N})$. Als Übergangskern ergibt sich

$$p(k, \bullet) \quad = \quad P \circ \left(\sum_{i=1}^{k} N_i^n \right)^{-1} \quad = \quad \nu^{*k},$$

wobei ν^{*k} die k-fache Faltung der Nachkommensverteilung $\nu = P \circ (N_i^n)^{-1}$ ist.

(6). *Wrightsches Evolutionsmodell:* In diesem Modell besteht die Population zu jedem Zeitpunkt n auf seiner festen Anzahl m von Individuen, von denen jedes genau eines der Merkmale aus einer endlichen Menge T besitzt. Die Merkmale werden gemäß folgendem Mechanismus von einer Generation zur nächsten vererbt:

Algorithmus 11.2 (Multinomiales Resampling).

for $i := 1, \ldots, m$ **do**
  erzeuge $w \sim \mathrm{Unif}\{1, \ldots, m\}$
  $x_{n+1}^{(i)} := x_n^{(w)}$
end for

Jedes Individuum der Nachkommensgeneration sucht sich also zufällig und unabhängig voneinander einen Vorfahren in der Elterngeneration, und nimmt dessen Merkmalsausprägungen an. Durch den Algorithmus wird eine Markovkette (X_n) mit Zustandsraum T^m und Übergangskern

$$p(x, \bullet) \quad = \quad \bigotimes_{i=1}^{m} \hat{\mu}(x)$$

definiert, wobei $\hat{\mu}(x) = \frac{1}{m} \sum_{i=1}^{m} \delta_{x^{(i)}}$ die empirische Verteilung der Merkmalsausprägungen $x = (x^{(1)}, \ldots, x^{(m)})$ in der vorherigen Population ist.
Anstatt die Merkmalsausprägungen $X_n^{(i)}$ aller Individuen einer Generation zu betrachten („mikroskopische Beschreibung“), genügt es die Häufigkeiten

$$H_n(a) \quad = \quad \left|\left\{i \in \{1, \ldots, m\} \,:\, X_n^{(i)} = a\right\}\right|, \qquad a \in T,$$

aller möglichen Merkmalsausprägungen a zu notieren („makroskopische Beschreibung“). Die Histogrammvektoren $H_n = (H_n(a))_{a \in T}$ bilden eine zeithomogene Markovkette mit Werten im Raum $\mathrm{Hist}(m, T)$ der Histogramme von m Beobachtungswerten aus T. Der Übergangskern ist durch

$$p(h, \bullet) \quad = \quad \mathrm{Mult}\left(h \,\Big/ \sum_{a \in S} h(a)\right), \qquad h \in \mathrm{Hist}(m, T),$$

gegeben, d.h. der Histogrammvektor im nächsten Schritt ist multinomialverteilt mit Ergebniswahrscheinlichkeiten der Merkmalsausprägungen $a \in T$ proportional zu den Häufigkeiten $h(a)$ im letzten Schritt. Dies erklärt auch die Bezeichnung „Multinomiales Resampling.“ Multinomiale Resamplingschritte werden u.a. in genetischen Algorithmen und sequentiellen Monte-Carlo Verfahren eingesetzt.

Aus der Darstellung von Markovketten als zufällige dynamische Systeme ergibt sich unmittelbar ein explizites Konstruktionsverfahren für Markovketten mit Zustandsraum $\mathbb{R}$:
Seien μ eine Wahrscheinlichkeitsverteilung und $p_n, n \in \mathbb{N}$, stochastische Kerne auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Wir betrachten die linksstetigen Inversen

$$\begin{aligned} \underline{G}_0(u) &= \inf\{c \in \mathbb{R} \,:\, F_0(c) \geq u\} \qquad \text{und} \\ \underline{G}_n(x, u) &= \inf\{c \in \mathbb{R} \,:\, F_n(x, c) \geq u\} \end{aligned}$$

der Verteilungsfunktionen $F_0(c) = \mu[(-\infty, c)]$ und $F_n(x, c) = p_n(x\,,\,(-\infty, c])$ der Wahrscheinlichkeitsverteilungen μ und $p_n(x, \bullet)$. Aus Satz 11.1 und Satz 4.20 folgt unmittelbar:

Korollar 11.3 (Existenzsatz und Konstruktionsverfahren für Markovketten). *Sei $U_0, U_1, U_2, \ldots$ eine Folge von unabhängigen, auf $(0,1)$ gleichverteilten Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$. Dann ist der durch*

$$X_0 \;=\; \underline{G}_0(U_0), \qquad X_{n+1} \;=\; \underline{G}_{n+1}(X_n, U_{n+1})$$

definierte stochastische Prozess eine Markovkette mit Startverteilung μ und Übergangskernen p_n.

Bemerkung. Auch auf anderen Zustandsräumen kann man Markovketten oft auf ähnliche Weise explizit konstruieren, siehe z.B. die Übung für den diskreten Fall. Die Konstruktion liefert unmittelbar einen Algorithmus zur Simulation der Markovkette:

Algorithmus 11.4 (Simulation einer reellwertigen Markovkette).

erzeuge $U_0 \sim \mathrm{Unif}(0,1)$; $y_0 := \underline{G}_0(u_0)$
  for $n := 1, 2, \ldots$ **do**
    erzeuge $u_n \sim \mathrm{Unif}(0,1)$; $y_n := \underline{G}_n(y_{n-1}, u_n)$
  end for

Endlichdimensionale Randverteilung eine Markovkette

Wir wollen nun Verteilungen von Markovketten berechnen. Sei (X_n) ein auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ definierter stochastischer Prozess mit Zustandsraum $(S, \mathcal{S})$.

Satz 11.5. *Es sind äquivalent:*

(1). (X_n) ist eine Markovkette mit Übergangswahrscheinlichkeiten p_n und Startverteilung μ.

(2). Für jedes $n \geq 0$ hat $(X_0, X_1, \ldots, X_n)$ die Verteilung

$$\mu(dx_0)p_1(x_0, dx_1)p_2(x_1, dx_2) \cdot \ldots \cdot p_n(x_{n-1}, dx_n),$$

d.h. für alle messbaren Funktionen $f : S^{n+1} \to \mathbb{R}_+$ gilt

$$E[f(X_0, \ldots, X_n)] \;=\; \int \mu(dx_0) \int p_1(x_0, dx_1) \cdots \int p_n(x_{n-1}, dx_n) f(x_0, \ldots, x_n). \tag{11.1.4}$$

Beweis. **„(1) ⇒ (2)“ :** Ist (X_n) eine Markovkette mit Startverteilung μ und Übergangskernen p_n, dann gilt für $n \in \mathbb{N}$ und $B_0, \ldots, B_n \in \mathcal{S}$:

$$\begin{aligned} P[X_{0:n} \in B_0 \times \ldots \times B_n] \;&=\; E\left[P\left[X_n \in B_n \mid X_{0:n-1}\right] \,;\, X_{0:n-1} \in B_0 \times \ldots \times B_{n-1}\right] \\ &=\; \int\limits_{B_0 \times \ldots \times B_{n-1}} p(x_{n-1}, B_n)\, \mu_{X_{0:n-1}}(dx_{0:n-1}). \end{aligned}$$

Durch Induktion nach n folgt

$$P[X_{0:n} \in B_0 \times \ldots \times B_n] = \int_{B_n} \cdots \int_{B_1} \int_{B_0} \mu(dx_0) p_1(x_0, dx_1) \cdot \ldots \cdot p_n(x_{n-1}, dx_n)$$

für alle $n \geq 0$ und $B_i \in \mathcal{S}$. Also gilt $X_{0:n} \sim \mu \otimes p_1 \otimes \ldots \otimes p_n$, und damit (11.1.4).

„(2) $\Rightarrow$ (1)“ : Gilt (11.1.4), dann hat X_0 die Verteilung μ, und $(p_{n+1}f)(X_n)$ ist für alle messbaren Funktionen $f : S \to [0, \infty)$ eine Version der bedingten Erwartung $E[f(X_{n+1}) \mid X_{0:n}]$. Zum Beweis überprüfen wir die definierenden Eigenschaften der bedingten Erwartung: $(p_{n+1}f)(X_n)$ ist eine Funktion von $X_{0:n}$, und es gilt

$$\begin{aligned}
& E[f(X_{n+1}) \cdot g(X_{0:n})] \\
= & \int \mu(dx_0) \int p_1(x_0, dx_1) \cdot \ldots \cdot \int p_n(x_{n-1}, dx_n) g(x_{0:n}) \int p_{n+1}(x_n, dx_{n+1}) f(x_{n+1}) \\
= & \int \mu(dx_0) \int p_1(x_0, dx_1) \cdot \ldots \cdot \int p_n(x_{n-1}, dx_n) g(x_{0:n}) (p_{n+1}f)(x_n) \\
= & E[(p_{n+1}f)(X_n) \cdot g(X_{0:n})]
\end{aligned}$$

für alle messbaren Funktionen $g : S^{n+1} \to [0, \infty)$.

□

Seien μ eine Wahrscheinlichkeitsverteilung, p, q, r stochastische Kerne, und f eine messbare nicht-negative Funktion auf $(S, \mathcal{S})$. Wir bezeichnen mit

$$(\mu p)(dy) \quad = \quad \int \mu(dx) p(x, dy)$$

die Verteilung der 2. Komponente unter dem Maß $\mu \otimes p$, und mit

$$(pq)(x, dz) \quad = \quad \int p(x, dy) q(y, dz)$$

den stochastischen Kern, der durch Hintereinanderausführen von p und q entsteht. Aus dem Satz von Fubini ergeben sich die folgenden *Rechenregeln für stochastische Kerne*:

$$\int f \, d(\mu p) = \int \int \mu(dx) p(x, dy) f(y) = \int (pf) \, d\mu \tag{11.1.5}$$

$$p(qf) = (pq)f \tag{11.1.6}$$

$$(\mu p)q = \mu(pq) \tag{11.1.7}$$

$$p(qr) = (pq)r \tag{11.1.8}$$

Als Verteilung der Markovkette zur Zeit n erhalten wir dementsprechend

$$P \circ X_n^{-1} \quad = \quad \mu p_1 p_2 \cdot \ldots \cdot p_n, \tag{11.1.9}$$

wobei das Produkt wegen (11.1.7) und (11.1.8) nicht von der Klammerung abhängt.
Ist der Zustandsraum S abzählbar, dann gelten die folgenden Identifikationen:

μ	$\leftrightarrow$	$(\mu(x)\|x \in S)$	Zeilenvektor
f	$\leftrightarrow$	$(f(x)\|x \in S)$	Spaltenvektor
p	$\leftrightarrow$	$(p(x,y)\|x,y \in S)$	stochastische Matrix
$(\mu p)(y)$	$=$	$\sum_x \mu(x)p(x,y)$	Multiplikation mit Zeilenvektor von links
$(pf)(x)$	$=$	$\sum_y p(x,y)f(y)$	Multiplikation mit Spaltenvektor von rechts
$(pq)(x,z)$	$=$	$\sum_y p(x,y)q(y,z)$	Matrizenprodukt.

Beispiel (Zeithomogene Markovkette mit endlichem Zustandsraum). Wir betrachten einen endlichen Zustandsraum S mit k Elementen, und eine stochastische Matrix p, die nicht von n abhängt. Die Verteilung zur Zeit n einer zeithomogenen Markovkette mit Startverteilung μ und Übergangsmatrix p ist dann

$$P \circ X_n^{-1} \quad = \quad \mu p^n.$$

Um die Verteilung und deren Asymptotik zu berechnen, können wir die Spektraldarstellung der Übergangsmatrix verwenden. Seien $\lambda_1, \ldots, \lambda_k \in \mathbb{C}$ die Eigenwerte von p, d.h. die Nullstellen des charakteristischen Polynoms $\chi(\lambda) = \det(p - \lambda I)$. Da p eine stochastische Matrix ist, gilt Folgendes:

(1). $|\lambda_j| \leq 1$ für alle j,
(dies folgt wegen $\|pf\|_\infty = \max_x \left|\sum_y p(x,y)f(y)\right| \leq \|f\|_\infty$ für alle f).

(2). $\lambda_1 = 1$ ist Eigenwert mit Rechtseigenvektor $f_1 = (1, \ldots, 1)^T$.

(3). Nichtreelle Eigenwerte treten in Paaren $\lambda, \overline{\lambda}$ auf.

Wir nehmen nun der Einfachheit halber an, dass alle Eigenwerte einfach sind, d.h. $\lambda_i \neq \lambda_j$ für $i \neq j$. In diesem Fall existieren Rechts- und Linkseigenvektoren f_j, ν_j $(1 \leq j \leq k)$ mit

$$pf_j = \lambda_j f_j, \qquad \nu_j p = \lambda_j \nu_j, \qquad \text{und} \qquad \langle \nu_i, f_j \rangle = \sum_{x \in S} \nu_i(x) f_j(x) = \delta_{ij}.$$

Mithilfe der aus den Rechts- und Linkseigenvektoren gebildeten Matrizen

$$U = (f_1, \ldots, f_k), \qquad V = \begin{pmatrix} \nu_1 \\ \nu_2 \\ \vdots \\ \nu_k \end{pmatrix}, \qquad V \cdot U = I,$$

erhalten wir die Spektraldarstellung

$$p \quad = \quad \sum_{j=1}^{k} \lambda_j f_j \otimes \nu_j \quad = \quad U \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_k \end{pmatrix} V,$$

für die Übergangsmatrix p, und damit auch für p^n:

$$p^n \quad = \quad \sum_{j=1}^{k} \lambda_j^n f_j \otimes \nu_j \quad = \quad U \begin{pmatrix} \lambda_1^n & 0 & \cdots & 0 \\ 0 & \lambda_2^n & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_k^n \end{pmatrix} V.$$

Für die Verteilung der Markovkette zur Zeit n ergibt sich

$$P \circ X_n^{-1} \quad = \quad \sum_{j=1}^{n} \lambda_j^n \langle \mu, f_j \rangle \nu_j. \tag{11.1.10}$$

Insbesondere folgt:

Satz 11.6 (Exponentielle Konvergenz ins Gleichgewicht). *Sind die Eigenwerte einer stochastischen Matrix $p \in \mathbb{R}^{k \times k}$ einfach, und gilt $|\lambda_j| < 1$ für alle $j \neq 1$, dann existiert eine Gleichgewichtsverteilung ν von p, und für jede Startverteilung μ gilt*

$$\mu p^n \quad = \quad \nu + O\left(\max_{j \neq 1} |\lambda_j|^n \right) \qquad \textit{für } n \to \infty.$$

Beweis. Nach (11.1.10) gilt

$$\mu p^n \quad = \quad \langle \mu, f_1 \rangle \nu_1 + \sum_{j=2}^{k} \lambda_j^n \langle \mu, f_j \rangle \nu_j \qquad \text{für alle } n \geq 0.$$

Aus $\langle \mu, f_1 \rangle = \langle \mu, (1, \ldots, 1)^T \rangle = \sum \mu(x) = 1$ folgt

$$\mu p^n \quad = \quad \nu_1 + O\left(\max_{j \neq 1} |\lambda_j|^n \right)$$

Insbesondere ist $\nu_1 = \lim \mu p^n$ eine Wahrscheinlichkeitsverteilung mit $\nu_1 p = \nu_1$, also ein Gleichgewicht von p. Ist umgekehrt μ ein beliebiges Gleichgewicht von p, dann gilt $\mu p^n = \mu$ für alle $n \geq 0$, und damit

$$\mu \quad = \quad \lim_{n \to \infty} \mu p^n \quad = \quad \nu_1.$$

□

Bemerkung. (1). Sind die Eigenwerte nicht einfach, dann folgt eine ähnliche Aussage über die Jordansche Normalformdarstellung der Übergangsmatrix p. Als Konvergenzgeschwindigkeit ergibt sich in diesem Fall $O(n^{m-1} \max_{i \neq 1} |\lambda_i|^n)$, wobei m die größte Multiplizität des betragsmäßig zweitgrößten Eigenwertes ist (Satz von Perron-Frobenius).

(2). Entscheidend für die exponentielle Konvergenzrate ist die Lücke zwischen dem Eigenwert 1 und dem Rest des Spektrums. Eine entsprechende Aussage kann man auch auf allgemeinen Zustandsräumen mithilfe des Spektralsatzes für selbstadjungierte Operatoren zeigen, falls die Gleichgewichtsverteilung die Detailed Balance Bedingung erfüllt.

Beispiel. (1). Die Übergangsmatrix der Markovkette aus Abbildung 11.2 ist

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

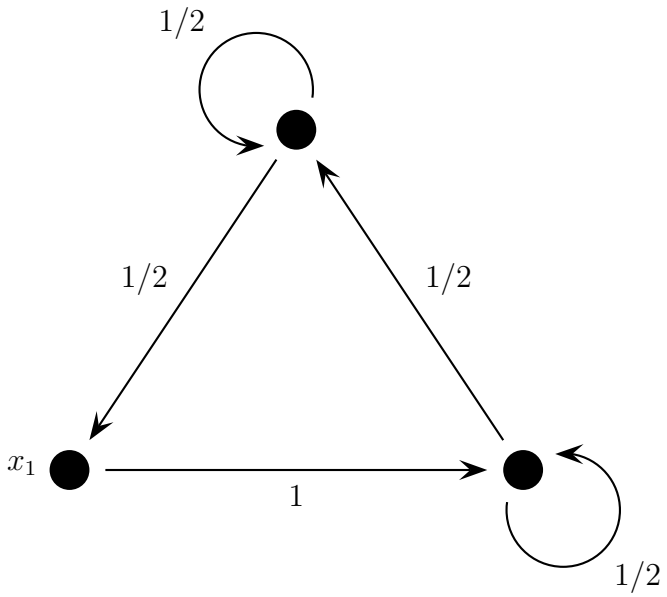


Abbildung 11.2: Markovkette mit zugehöriger Übergangsmatrix p.

Eigenwerte sind $\lambda_1 = 1, \lambda_2 = i/2$ und $\lambda_3 = -i/2$. Es folgt:

$$p^n \quad = \quad A + B \cdot \left(\frac{i}{2}\right)^n + C \cdot \left(-\frac{i}{2}\right)^n$$

mit Matrizen $A, B, C \in \mathbb{C}^{3\times 3}$. Wegen $p^0(x_1, x_1) = 1$ und $p^1(x_1, x_1) = p^2(x_1, x_1) = 0$ folgt

$$p^n(x_1, x_1) \quad = \quad \frac{1}{5} + \left(\frac{1}{2}\right)^n \cdot \left(\frac{4}{5}\cos\frac{n\pi}{2} - \frac{2}{5}\sin\frac{n\pi}{2}\right) \qquad \text{für alle } n \geq 0.$$

Der Wert $1/5$ ist die erste Komponente des Gleichgewichtsvektors $\nu_1 = (1/5, 2/5, 2/5)$. Für $n \to \infty$ konvergieren die Übergangswahrscheinlichkeiten mit Rate $O(2^{-n})$ gegen ν_1.

(2). Die Übergangsmatrix einer deterministischen Rotation auf dem diskreten Kreis $\mathbb{Z}/k\mathbb{Z}, k \in \mathbb{N}$ ist

$$p = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ 1 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix}.$$

Das charakteristische Polynom ist $\chi(\lambda) = (-1)^k \cdot (\lambda^k - 1)$, und die Eigenwerte von p sind dementsprechend die k-ten Einheitswurzeln $\lambda_j = \exp(2\pi i \cdot (j-1)/k), j = 1, \ldots, k$. Da alle Eigenwerte Betrag 1 haben, gilt keine exponentielle Konvergenz ins Gleichgewicht. Tatsächlich ist die Markovkette mit Übergangsmatrix p periodisch: $X_{n+mk} = X_n$ P-fast sicher für alle $n, m \geq 0$.

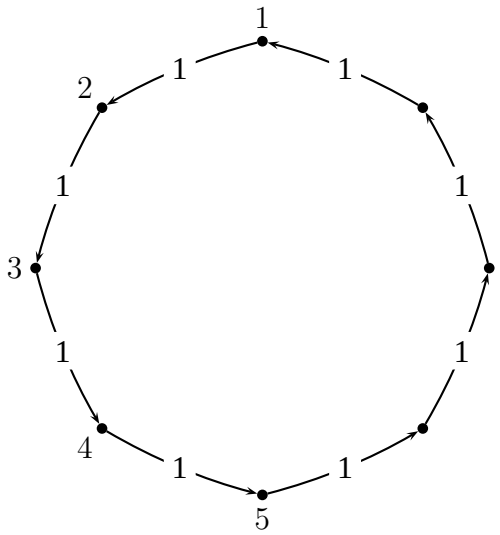


Abbildung 11.3: Darstellung eines gerichteten Graphen einer Markovkette auf $\mathbb{Z}/m\mathbb{Z}$.

Verteilung auf dem Pfadraum; kanonisches Modell

In Satz 11.5 haben wir die endlich-dimensionalen Verteilungen $P \circ (X_0, X_1, \ldots, X_n)^{-1}$ einer Markovkette $(X_n)_{n\geq 0}$ berechnet. Viele relevante Ereignisse hängen aber von unendlich vielen

der Zufallsvariablen X_n ab. Die gemeinsame Verteilung aller dieser Zufallsvariablen ist eine Wahrscheinlichkeitsverteilung auf dem unendlichen Produktraum

$$\hat{S} \quad := \quad S^{\{0,1,2,\ldots\}} \quad = \quad \{x = (x_0, x_1, x_2, \ldots) \mid x_i \in S\}$$

aller diskreten Pfade (Folgen) mit Werten in S. Wir versehen die Menge $\hat{S}$ wie üblich mit der von den Koordinatenabbildungen

$$\pi_k : \quad \hat{S} \to S, \qquad \pi_k(x) = x_k,$$

erzeugten Produkt-σ-Algebra

$$\mathcal{F} \quad = \quad \sigma(\pi_k | k \geq 0) \quad = \quad \bigotimes_{k \geq 0} \mathcal{S}.$$

Einen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ definierten stochastischen Prozess $(X_n)_{n \geq 0}$ können wir auch als Abbildung

$$X = (X_n) : \Omega \to \hat{S}$$

auffassen. Die Abbildung X ist eine $\hat{S}$-wertige Zufallsvariable, also messbar bzgl. der σ-Algebren $\mathcal{A}/\mathcal{F}$, denn $\mathcal{F}$ wird von den Koordinatenabbildungen π_k erzeugt, und $\pi_k(X) = X_k$ ist für alle $k \geq 0$ messbar. Wir können daher die Verteilung

$$\mu_X[A] = P[(X_n) \in A], \qquad A \in \mathcal{F},$$

des stochastischen Prozesses (X_n) auf dem Pfadraum $(\hat{S}, \mathcal{F})$ betrachten.

Wir beschränken uns nun wieder auf Markovketten. Seien $p_1, p_2, \ldots$ stochastische Kerne, und μ eine Wahrscheinlichkeitsverteilung auf $(S, \mathcal{S})$.

Satz 11.7 (Existenz und Eindeutigkeit in Verteilung von Markovketten). *(1). Es existiert genau eine Wahrscheinlichkeitsverteilung P_μ auf dem unendlichen Produktraum $(\hat{S}, \mathcal{F})$, bzgl. der die Folge $(\pi_n)_{n \geq 0}$ der Koordinatenabbildungen eine Markovkette mit Startverteilung $\mu(dx)$ und Übergangskern $p_n(x, dy)$ ist.*

(2). Ist $(X_n)_{n \geq 0}$ auf $(\Omega, \mathcal{A}, P)$ eine beliebige Markovkette mit Startverteilung μ und Übergangswahrscheinlichkeiten p_n, dann gilt

$$P[(X_n) \in A] \quad = \quad P_\mu[A] \qquad \textit{für alle } A \in \mathcal{F},$$

d.h. P_μ ist die Verteilung von (X_n) auf $(\hat{S}, \mathcal{F})$.

Bemerkung (Unendliches mehrstufiges Modell). Die Verteilung P_μ der Markovkette entspricht einem mehrstufigen Modell auf dem unendlichen Produktraum $\hat{S} = S^{\{0,1,2,\ldots\}}$:

$$P_\mu(dx) \quad = \quad \mu(dx_0)p_1(x_0, dx_1)p_2(x_1, dx_2) \cdot \ldots .$$

Beweis. Nach Satz 11.5 ist ein stochastischer Prozess (X_n) genau dann eine Markovkette zu μ und p_n, wenn $(X_0, \ldots, X_n)$ für jedes $n \geq 0$ die Verteilung

$$\mu_{0:n}(dx_{0:n}) := \mu(dx_0)p_1(x_0, dx_1) \cdot \ldots \cdot p_n(x_{n-1}, dx_n)$$

hat. Zu zeigen ist, dass zu der Familie $\mu_{0:n}, n \geq 0$, von Wahrscheinlichkeitsverteilungen auf den endlichdimensionalen Produkträumen $S^{\{0,1,\ldots,n\}}$ eine eindeutige Wahrscheinlichkeitsverteilung P_μ auf den unendlichen Produktraum $\hat{S}$ existiert, bzgl. der die ersten $n+1$ Koordinaten $x_0, \ldots, x_n$ für jedes n die Verteilung $\mu_{0:n}$ haben. Die Folge $\pi_n(x) = x_n$ der Koordinatenabbildungen ist dann unter P_μ eine Markovkette mit den vorgegebenen Übergangswahrscheinlichkeiten.

Existenz: Die Wahrscheinlichkeitsverteilungen $\mu_{0:n}$ auf den endlichdimensionalen Produkträumen $S^{\{0,1,\ldots,n\}}$ sind *konsistent*, d.h. für $m \leq n$ stimmt die Verteilung der ersten $m+1$ Koordinaten unter $\mu_{0:n}$ mit $\mu_{0:m}$ überein. Aus dem Fortsetzungssatz von Carathéodory folgt nun allgemein, dass zu einer Familie von konsistenten endlichdimensionalen Verteilungen eine Wahrscheinlichkeitsverteilung auf dem unendlichen Produktraum mit den entsprechenden Randverteilungen existiert (*Fortsetzungssatz von Kolmogorov*). Wir verzichten hier auf den Beweis dieser maßtheoretischen Aussage, der sich in vielen Lehrbüchern zur Wahrscheinlichkeitstheorie findet, s. z.B. [Bauer], [Klenke], oder den Anhang in [Durrett: Probability - Theory and Examples].

Eindeutigkeit: Ein stochastischer Prozess (X_n) auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ ist genau dann eine Markovkette mit Parametern μ und p_n, wenn

$$\int_{B_0} \mu(dx_0) \int_{B_1} p_1(x_0, dx_1) \cdots \int_{B_n} p_n(x_{n-1}, dx_n) = P[X_0 \in B_0, \ldots, X_n \in B_n] = P[X \in A] \tag{11.1.11}$$

für jede Zylindermenge der Form

$$A = B_0 \times B_1 \times \ldots \times B_n \times S \times S \times \ldots \quad = \quad \{\pi_0 \in B_0, \ldots, \pi_n \in B_n\},$$

mit $n \in \mathbb{N}$ und $B_0, \ldots, B_n \in \mathcal{S}$ gilt. Da die Zylindermengen ein durchschnittsstabiles Erzeugendensystem der Produkt-σ-Algebra $\mathcal{F}$ bilden, ist die Verteilung von X auf $(\hat{S}, \mathcal{F})$ durch (11.1.11) eindeutig festgelegt.

Ist $X_n = \pi_n$ der Koordinatenprozess auf dem Produktraum $(\Omega, \mathcal{A}) = (\hat{S}, \mathcal{F})$, dann stimmt die Verteilung von X mit dem zugrundeliegenden Wahrscheinlichkeitsmaß P überein, d.h. P ist durch (11.1.11) eindeutig festgelegt.

□

Bemerkung (Konstruktive Existenzbeweise). Im Fall $S = \mathbb{R}$ erhalten wir die Wahrscheinlichkeitsverteilung P_μ auch direkt als Verteilung der im letzten Abschnitt explizit konstruierten Markovkette (X_n) mit Startverteilung μ und Übergangswahrscheinlichkeiten p_n. Auch auf allgemeineren Zustandsräumen kann man die Existenz von P_μ auf ähnliche Weise aus der Existenz einer Folge von auf $(0, 1)$ gleichverteilten, unabhängigen Zufallsvariablen herleiten (z.B. durch eine messbare Transformation des Zustandsraums nach $\mathbb{R}$).

Nach Satz 11.7 können wir eine Markovkette mit beliebigen Übergangswahrscheinlichkeiten durch die Koordinatenabbildungen auf dem unendlichen Produktraum $\hat{S} = S^{\{0,1,2,\dots\}}$ realisieren.

Definition. *Der durch die Koordinatenabbildungen $\pi_n(x) = x_n$ gegebene stochastische Prozess auf dem Wahrscheinlichkeitsraum $(\hat{S}, \mathcal{F}, P_\mu)$ heißt **kanonisches Modell** der Markovkette mit Startverteilung μ und Übergangswahrscheinlichkeiten p_n.*

Allgemein kann man jeden stochastischen Prozess im kanonischen Modell realisieren, indem man zur Verteilung des Prozesses auf dem Pfadraum übergeht.

11.2 Markoveigenschaft und Differenzengleichungen

In diesem Abschnitt werden wir die wichtige Verbindung von Markovketten und Differenzengleichungen betrachten. Dazu beweisen wir zunächst eine weitergehende Form der definierenden Eigenschaft einer Markovkette.

Sei $(X_n)_{n\geq 0}$ auf $(\Omega, \mathcal{A}, P)$ eine Markovkette mit Startverteilung μ und Übergangskernen p_n. Ist $(S, \mathcal{S})$ der Zustandsraum, dann hat (X_n) nach Satz 11.7 die Verteilung

$$P_\mu(dx) \quad = \quad \mu(dx_0)p_1(x_0, dx_1)p_2(x_1, dx_2) \cdot \dots$$

auf dem unendlichen Produktraum $\hat{S} = S^{\{0,1,2,\dots\}}$. Wir bezeichnen im Folgenden die Verteilung P_{δ_x} der Markovkette bei Startwert x kurz mit P_x. Entsprechend sei $P_x^{(n)}$ die Verteilung der Markovkette mit Start in x und Übergangskernen $p_{n+1}, p_{n+2}, \dots$.

Die Markoveigenschaft

In Erweiterung der definierenden Eigenschaft einer Markovkette können wir sogar die bedingte Verteilung der um n Schritte verschobenen Kette gegeben den Verlauf bis zur Zeit n identifizieren:

Satz 11.8 (Markoveigenschaft). *Für alle $n \geq 0$ und alle $\mathcal{F}$-messbaren Funktionen $F : \hat{S} \to [0, \infty)$ gilt:*

$$E[F(X_n, X_{n+1}, \ldots) \mid X_{0:n}] \quad = \quad E^{(n)}_{X_n}[F] \qquad P\text{-fast sicher}. \tag{11.2.1}$$

Bemerkung. (1). Für *zeitlich homogene Markovketten* gilt $P^{(n)}_x = P_x$ für alle n.

(2). Für diskrete Zustandsräume ergibt sich, dass $(X_n, X_{n+1}, \ldots)$ unter der bedingten Verteilung gegebene $X_{0:n} = x_{0:n}$ für jedes $n \geq 0$ und $x_{0:n} \in S^{n+1}$ mit $P[X_{0:n} = x_{0:n}] \neq 0$ eine Markovkette mit Start in x_n und Übergangskernen $p_{n+1}, p_{n+2}, \ldots$ ist.

Beweis. Der Beweis erfolgt in mehreren Schritten:

(1). Wir nehmen zunächst an, dass die Funktion F nur von endlich vielen Variablen abhängt, d.h.

$$F(x_0, x_1, \ldots) \; = \; f(x_{0:k}) \quad \text{für ein } k \geq 0 \text{ und eine messbare Funktion } f : S^{k+1} \to \mathbb{R}_+. \tag{11.2.2}$$

In diesem Fall können wir direkt verifizieren, dass $E^{(n)}_{X_n}[F]$ eine Version der bedingten Erwartung in (11.2.1) ist:

(a) Es gilt $E^{(n)}_{X_n}[F] = g(X_n)$ mit

$$g(z) \quad = \quad E^{(n)}_z[F] \quad = \quad \int p_1(z, dx_1) \int p_2(x_1, dx_2) \cdots \int p_k(x_{k-1}, dx_k) f(x_{0:k}).$$

Da $f : S^{k+1} \to \mathbb{R}_+$ produktmessbar ist, ist $g : S \to \mathbb{R}_+$ messbar.

(b) Für $n \geq 0$ und eine messbare Funktion $h : S^{n+1} \to \mathbb{R}_+$ gilt

$$\begin{aligned} & E[F(X_n, X_{n+1}, \ldots) h(X_{0:n})] \; = \; E[f(X_{n:n+k}) h(X_{(0:n)})] \\ = \; & \int \mu(dx_0) \int p_1(x_0, dx_1) \cdots \int p_n(x_{n-1}, dx_n) h(x_{0:n}) \times \\ & \times \underbrace{\int p_{n+1}(x_n, dx_{n+1}) \cdots \int p_{n+k}(x_{n+k-1}, dx_{n+k}) f(x_{n:n+k})}_{E^{(n)}_{X_n}[F]} \\ = \; & E\left[E^{(n)}_{X_n}[F] \cdot h(X_{0:n}) \right]. \end{aligned}$$

(2). Nach (1) gilt (11.2.1) für Indikatorfunktionen $F = I_A$ von Zylindermengen der Form $A = \{x \in \hat{S} : x_0 \in B_0, \ldots, x_n \in B_n\}$ mit $n \in \mathbb{N}$ und $B_0, \ldots, B_n \in \mathcal{S}$. Wir zeigen nun, dass die Aussage dann auch für Indikatorfunktionen von beliebigen Mengen A aus der Produkt-σ-Algebra $\mathcal{F}$ gilt. Dazu bemerken wir, dass das System $\mathcal{D}$ aller Mengen $A \in \mathcal{F}$, für die (11.2.1) mit $F = I_A$ gilt, ein Dynkinsystem ist. Sind beispielsweise $A_1, A_2, \ldots \in \mathcal{D}$ disjunkt, dann ist auch $\bigcup_k A_k$ in $\mathcal{D}$ enthalten, denn

$$\begin{aligned} E[I_{\bigcup A_k}(X_n, X_{n+1}, \ldots) \mid X_{0:n}] &= \sum_k E[I_{A_k}(X_n, X_{n+1}, \ldots) \mid X_{0:n}] \\ &= \sum_k E_{X_n}^{(n)}[I_{A_k}] \quad = \quad E_{X_n}^{(n)}[I_{\bigcup A_k}] \qquad P\text{-fast sicher.} \end{aligned}$$

Da die Zylindermengen ein durchschnittsstabiles Erzeugendensystem der Produkt-σ-Algebra bilden, folgt $\mathcal{D} = \mathcal{F}$, d.h. (11.2.1) gilt für alle $F = I_A$ mit $A \in \mathcal{F}$.

(3). Die Aussage (11.2.1) für beliebige nicht-negative $\mathcal{F}$-messbare Funktionen F folgt nun wie üblich durch maßtheoretische Induktion.

□

Bemerkung (Markoveigenschaft im kanonischen Modell). Im kanonischen Modell können wir die Markoveigenschaft noch etwas kompakter formulieren. Sei $\theta : \hat{S} \to \hat{S}$ die durch

$$\theta(x_0, x_1, \ldots) \quad = \quad (x_1, x_2 \ldots)$$

definierte Shiftabbildung auf dem Pfadraum $\hat{S}$, und seien $X_n : \hat{S} \to S$,

$$X_n(x_0, x_1, \ldots) \quad = \quad x_n,$$

die Koordinatenabbildungen. Dann gilt:

$$E_\mu[F \circ \theta^n \mid X_{0:n}] \quad = \quad E_{X_n}^{(n)}[F] \qquad P\text{-fast sicher} \tag{11.2.3}$$

für alle Wahrscheinlichkeitsverteilungen μ auf $(S, \mathcal{S})$ und alle messbaren Funktionen $F : \hat{S} \to \mathbb{R}_+$.

Das folgende Korollar liefert eine weitere äquivalente Formulierung der Markoveigenschaft.

Korollar 11.9 (Markoveigenschaft, 2. Version). *Ist (X_n) unter P eine Markovkette mit Parametern μ und p_n, dann ist $(X_n, X_{n+1}, \ldots)$* ***bedingt unabhängig von*** *$(X_0, \ldots, X_n)$* ***gegeben*** *X_n mit bedingter Verteilung $P_{X_n}^{(n)}$, d.h.*

$$\begin{aligned} &E[F(X_n, X_{n+1}, \ldots) g(X_0, \ldots, X_n) \mid X_n] \\ &\quad = \quad E_{X_n}^{(n)}[F] \cdot E[g(X_0, \ldots, X_n) \mid X_n] \\ &\quad = \quad E[F(X_n, X_{n+1}, \ldots) \mid X_n] \cdot E[g(X_0, \ldots, X_n) \mid X_n] \qquad P\textit{-fast sicher} \end{aligned}$$

für alle messbaren $F : \hat{S} \to [0, \infty)$ *und* $g : S^{n+1} \to [0, \infty)$.

Beweis. Wegen der Projektivität der bedingten Erwartung gilt nach Satz 11.8:

$$\begin{aligned} E[F(X_{n:\infty})g(X_{0:n}) \mid X_n] &= E[E[F(X_{n:\infty})g(X_{0:n}) \mid X_{0:n}] \mid X_n] \\ &= E\left[E^{(n)}_{X_n}[F]g(X_{0:n}) \,\middle|\, X_n\right] \quad = \quad E^{(n)}_{X_n}[F] \cdot E[g(X_{0:n}) \mid X_n]. \end{aligned}$$

□

Das Korollar besagt anschaulich, dass, gegeben den gegenwärtigen Zustand X_n, die zukünftige Entwicklung einer Markovkette bedingt unabhängig von der vorherigen Entwicklung ist:

„Die Zukunft ist bedingt unabhängig von der Vergangenheit gegeben die Gegenwart."

Beispiel (**Das klassische Ruinproblem**)**.** Wir wollen nun den Zusammenhang von Markovketten und Differenzengleichungen zunächst in einem einfachen Beispiel betrachten. In jeder Runde eines Glücksspiels trete einer der folgenden Fälle ein:

- Mit Wahrscheinlichkeit $p \in (0, 1)$ gewinnt der Spieler 1 Euro dazu.
- Mit Wahrscheinlichkeit $q = 1 - p$ verliert der Spieler 1 Euro.

Die Entwicklung des Kapitals X_n des Spielers kann dann durch einen Random Walk auf $\mathbb{Z}$ mit Übergangswahrscheinlichkeiten $p(x, x + 1) = p, p(x, x - 1) = q$ beschrieben werden. Sei $x \in \mathbb{Z}$ das Startkapital, und seien $a, b \in \mathbb{Z}$ mit $a \leq x \leq b$. Wir können den Random Walk ohne Beschränkung der Allgemeinheit im kanonischen Modell betrachten, d.h. P_x ist die Verteilung bei Startwert x auf dem Produktraum $\Omega = \mathbb{Z}^{\{0,1,2,\dots\}}$ und $X_n(\omega) = \omega_n$ ist die n-te Koordinatenabbildung.
Das Glücksspiel soll folgende mögliche Ausgänge haben:

- Im Fall $X_n \leq a$ ist der Spieler bankrott.
- Im Fall $X_n \geq b$ ist der Gegenspieler (bzw. die Spielbank) bankrott.

Die Zeit, zu der eines dieser beiden Ereignisse zum ersten Mal eintritt, wird durch die Zufallsvariable

$$T(\omega) \quad := \quad \min\{n \geq 0 \mid X_n(\omega) \leq a \quad \text{oder} \quad X_n(\omega) \geq b\}$$

beschrieben, wobei wir $\min \emptyset = \infty$ setzen. Wegen $\limsup |X_n| = +\infty$ gilt $T < \infty$ P_X-fast sicher für alle x. Also ist der Austrittspunkt

$$X_T(\omega) \quad := \quad X_{T(\omega)}(\omega)$$

des Random Walks (X_n) aus dem Intervall (a, b) P_x-fast sicher definiert, und mit Wahrscheinlichkeit 1 gilt $X_T = a$ (Spieler bankrott) oder $X_T = b$ (Spielbank bankrott). Wegen

$$X_T \quad = \quad \sum_{n=0}^{\infty} X_n \cdot I_{\{T=n\}}$$

ist auch X_T eine Zufallsvariable. Uns interessiert die *Ruinwahrscheinlichkeit*

$$h(x) \quad := \quad P_x[X_T = a]$$

des Spielers bei Startkapital x. Um diese zu berechnen, bedingen wir auf den ersten Schritt des Random Walks („first step analysis“). Sei dazu

$$\widetilde{X}_n(\omega) \quad := \quad X_{n+1}(\omega) \quad = \quad X_n(\theta(\omega))$$

der um einen Schritt verschobene Prozess, und sei

$$\widetilde{T} \quad = \quad \min\{n \geq 0 \mid \widetilde{X}_n \leq a \quad \text{oder} \quad \widetilde{X}_n \geq b\}.$$

Für $a < x < b$ gilt $T \geq 1$, also

$$X_T(\omega) \quad = \quad \widetilde{X}_{\widetilde{T}}(\omega) \quad = \quad X_T(\theta(\omega)) \qquad \text{für alle } \omega \in \Omega.$$

Daher folgt mit der Markoveigenschaft:

$$\begin{aligned} h(x) \quad &= \quad P_x[X_T = a] \quad = \quad P_x[X_T \circ \theta = a] \\ &= \quad P_x[X_T \circ \theta = a \mid X_1 = x+1] \cdot P_x[X_1 = x+1] + \\ &\quad +P_x[X_T \circ \theta = a \mid X_1 = x-1] \cdot P_x[X_1 = x-1] \\ &\overset{(11.2.1)}{=} \quad P_{x+1}[X_T = a] \cdot p + P_{x-1}[X_T = a] \cdot q \\ &= \quad p \cdot h(x+1) + q \cdot h(x-1). \end{aligned}$$

Die Funktion h hat also die *gewichtete Mittelwerteigenschaft*

$$h(x) \quad = \quad p \cdot h(x+1) + q \cdot h(x-1), \qquad \text{für alle } a < x < b.$$

Diese Eigenschaft ist äquivalent zu den Differenzengleichungen

$$0 \quad = \quad p \cdot (h(x+1) - h(x)) - q \cdot (h(x) - h(x-1)) \qquad \text{bzw.} \tag{11.2.4}$$

$$0 \quad = \quad q(\underbrace{(h(x+1) - h(x)) - (h(x) - h(x-1))}_{\text{diskrete 2. Ableitung}}) + (p-q) \underbrace{(h(x+1) - h(x))}_{\text{diskrete 1. Ableitung}} \tag{11.2.5}$$

Die gesuchte Ruinwahrscheinlichkeit $h(x)$ löst (11.2.3) bzw. (11.2.4) bzw. (11.2.5) mit den Randbedingungen

$$h(a) \;=\; P_a[X_T = a] \;=\; 1, \qquad h(b) \;=\; P_b[X_T = a] \;=\; 0.$$

Die Lösung der Differenzengleichung können wir leicht berechnen. Dazu verfahren wir ähnlich wir bei linearen gewöhnlichen Differentialgleichungen. Nach (11.2.4) gilt für die erste Differenz $v(x) := h(x+1) - h(x)$:

$$v(x) \quad = \quad \frac{q}{p} \cdot v(x-1) \qquad \text{für alle } a < x < b,$$

d.h. $v(x) = c \cdot (q/p)^x$ für ein $c \in \mathbb{R}$. Wir unterscheiden folgende Fälle:

(1). *Faire Münzwürfe* ($p = q = \frac{1}{2}$): In diesem Fall ist

$$h(x) \quad = \quad cx + d \qquad \text{mit } c, d \in \mathbb{R}$$

die allgemeine Lösung von (11.2.4) bzw. (11.2.5). Aus den Randbedingungen folgt:

$$h(x) \quad = \quad \frac{b-x}{b-a} \qquad (a \le x \le b).$$

(2). $p \neq \frac{1}{2}$: In diesem Fall erhalten wir

$$h(x) \quad = \quad c \cdot \left(\frac{q}{p}\right)^x + d \qquad \text{mit } c, d \in \mathbb{R}$$

als allgemeine Lösung. Aus den Randbedingungen folgt:

$$h(x) \quad = \quad \frac{\left(\frac{q}{p}\right)^b - \left(\frac{q}{p}\right)^x}{\left(\frac{q}{p}\right)^b - \left(\frac{q}{p}\right)^a} \quad = \quad \frac{1 - \left(\frac{p}{q}\right)^{b-x}}{1 - \left(\frac{p}{q}\right)^{b-a}}.$$

Wir haben damit die Ruinwahrscheinlichkeit in allen Fällen berechnet. Ist die Erfolgswahrscheinlichkeit p kleiner als $1/2$, dann gilt $\frac{p}{q} < 1$ und somit $h(x) \ge 1 - (p/q)^{b-x}$. Der letzte Ausdruck hängt nicht von dem Betrag a ab, bei dem der Spieler ruiniert ist. Beispielsweise gilt bei Roulette mit Höchsteinsatz 1 stets:

$$h(x) \ge 1 - \left(\frac{18}{19}\right)^{b-x}.$$

Bei genügend kleinem Höchsteinsatz geht also mit an Sicherheit grenzender Wahrscheinlichkeit der Spieler zuerst bankrott - selbst wenn das Kapital, das er mobilisieren kann, über dem der Bank liegt!

Differenzengleichungen für Markovketten

Die beim Ruinproblem verwendete Methode, die Berechnung von Wahrscheinlichkeiten und Erwartungswerten von Markovketten durch Konditionieren auf den ersten Schritt auf eine Differenzengleichung zurückzuführen, ist viel allgemeiner anwendbar. Wir betrachten im Folgenden eine beliebige zeithomogene Markovkette (X_n) mit Zustandsraum $(S, \mathcal{S})$ und Übergangskern $p(x, dy)$ im kanonischen Modell. Sei $D \in \mathcal{S}$ eine messbare Teilmenge des Zustandsraums, und sei

$$T(\omega) \quad := \quad \min\{n \geq 0 \,:\, X_n(\omega) \in D^C\}$$

die **erste Trefferzeit** von $D^C = S \setminus D$, d.h. die **erste Austrittszeit** der Markovkette aus dem Gebiet D. Hierbei setzen wir wieder $\min \emptyset = \infty$. Wir wollen Erwartungswerte von Typ

$$u(x) \quad = \quad E_x\left[\sum_{n=0}^{T-1} c(X_n)\right] + E_x\left[f(X_T)\,;\, T < \infty\right] \tag{11.2.6}$$

berechnen, wobei $c : D \to \mathbb{R}$ und $f : D^C \to \mathbb{R}$ gegebene nichtnegative, messbare Funktionen sind. Interpretieren wir beispielsweise $c(x)$ als Kosten, wenn die Markovkette den Punkt x durchläuft, und $f(x)$ als Zusatzkosten, wenn die Markovkette im Punkt x aus der Menge D austritt, dann gibt $u(x)$ die mittleren Gesamtkosten an, die beim Start in x bis zum Austritt aus der Menge D anfallen. Man beachte, dass sich eine Reihe wichtiger Wahrscheinlichkeiten und Erwartungswerte von Markovketten in der Form (11.2.6) darstellen lassen.

Beispiel. (1). $c \equiv 0, f \equiv 1$*: Austrittswahrscheinlichkeit aus D bzw. Trefferwahrscheinlichkeit von D^C:*

$$u(x) \quad = \quad P_x[T < \infty].$$

(2). $c \equiv 0, f = I_B$*: Verteilung des Austrittspunktes X_T:*

$$u(x) \quad = \quad P_x[X_T \in B\,;\, T < \infty].$$

(3). $c \equiv 1, f \equiv 0$*: Mittlere Austrittszeit aus D:*

$$u(x) \quad = \quad E_x[T].$$

(4). $c = I_B, f \equiv 0$*: Mittlere Anzahl der Besuche in B vor Austritt aus D:*

$$u(x) \quad = \quad E_x\left[\sum_{n=0}^{T-1} I_B(X_n)\right] \quad = \quad \sum_{n=0}^{\infty} P_x[X_n \in B\,,\, n < T].$$

Satz 11.10 (Poissongleichung). *u ist die minimale nichtnegative Lösung des Randwertproblems*

$$\begin{aligned} u(x) - (pu)(x) &= c(x) \qquad \text{für } x \in D, \\ u(x) &= f(x) \qquad \text{für } x \in D^C. \end{aligned} \tag{11.2.7}$$

Beweis. (1). Wir zeigen zunächst durch Bedingen auf den ersten Schritt, dass u das Randwertproblem löst. Dazu betrachten wir – wie oben – die verschobene Markovkette $\widetilde{X}_n = X_{n+1}$ und die entsprechende Austrittszeit $\widetilde{T} = \min\{n \geq 0 \,:\, \widetilde{X}_n \in D^C\}$. Für $x \in D$ gilt P_x-fast sicher $T \geq 1$, also

$$X_T = \widetilde{X}_{\widetilde{T}} \qquad \text{und} \qquad \sum_{n=0}^{T-1} c(X_n) = c(X_0) + \sum_{n=0}^{\widetilde{T}-1} c(\widetilde{X}_n).$$

Damit erhalten wir unter Verwendung der Markoveigenschaft:

$$\begin{aligned} & E_x\left[\sum_{n=0}^{T-1} c(X_n) + f(X_T)\cdot I_{\{T<\infty\}} \,\middle|\, X_1\right] \\ = \; & E_x\left[c(x) + \sum_{n=0}^{\widetilde{T}-1} c(\widetilde{X}_n) + f(\widetilde{X}_{\widetilde{T}})\cdot I_{\{\widetilde{T}<\infty\}} \,\middle|\, X_1\right] \\ = \; & c(x) + E_{X_1}\left[\sum_{n=0}^{T-1} c(X_n) + f(X_T)\cdot I_{\{T<\infty\}}\right] \\ = \; & c(x) + u(X_1) \qquad P\text{-fast sicher,} \end{aligned}$$

wobei wir $f(X_T)\cdot I_{\{T<\infty\}} := 0$ auf $\{T = \infty\}$ setzen. Durch Bilden des Erwartungswertes bzgl. P_x ergibt sich:

$$u(x) = c(x) + E_x[u(X_1)] = c(x) + (pu)(x) \qquad \text{für alle } x \in D.$$

Für $x \in D^C$ gilt $T = 0$ P_x-fast sicher, und damit

$$u(x) = E_x[f(X_0)] = f(x) \qquad \text{für alle } x \in D^C.$$

Also löst u das Randwertproblem (11.2.7).

(2). Sei nun $v \geq 0$ eine beliebige Lösung des Randwertproblems. Wir wollen zeigen, dass $v \geq u$ gilt. Dazu betrachten wir für $m \in \mathbb{N}$ die Funktion

$$u_m(x) := E_x\left[\sum_{n=0}^{(T\wedge m)-1} c(X_n) + f(X_T)\cdot I_{\{T\leq m\}}\right], \qquad x \in S.$$

Nach dem Satz über monotone Konvergenz gilt $u(x) = \sup_{m \geq 1} u_m(x)$. Durch Konditionieren auf den ersten Schritt erhalten wir ähnlich wie oben:

$$\begin{aligned} u_{m+1}(x) &= c(x) + (p\,u_m)(x) \qquad \text{für } x \in D, \qquad \text{und} \\ u_{m+1}(x) &= f(x) \qquad \text{für } x \in D^C. \end{aligned} \tag{11.2.8}$$

Wir zeigen nun durch Induktion nach m:

$$v \geq u_m \qquad \text{für alle } m \geq 0. \tag{11.2.9}$$

Für $m = 0$ ist (11.2.9) erfüllt, denn nach Voraussetzung gilt

$$v(x) \geq 0 = u_0(x) \quad \text{für alle } x \in D, \text{ und} \quad v(x) = f(x) = u_0(x) \quad \text{für alle } x \in D^C.$$

Gilt (11.2.9) für ein $m \geq 0$, dann folgt zudem

$$\begin{aligned} v &= pv + c \geq pu_m + c \overset{(11.2.8)}{=} u_{m+1} \qquad \text{auf } D, \text{ und} \\ v &= f = u_{m+1} \qquad \text{auf } D^C, \end{aligned}$$

d.h. (11.2.9) gilt auch für $m + 1$. Also ist (11.2.9) für alle $m \geq 0$ erfüllt. Damit folgt aber auch

$$v \geq \sup u_m = u,$$

d.h. u ist tatsächlich die *minimale* nichtnegative Lösung von (11.2.7).

□

Wir wollen uns nun das erhaltene Randwertproblem genauer ansehen. In kompakter Notation können wir (11.2.7) schreiben als

$$\begin{aligned} -\mathscr{L}u &= c \qquad \text{auf } D, \\ u &= f \qquad \text{auf } D^C \end{aligned} \tag{11.2.10}$$

mit

$$(\mathscr{L}u)(x) := (pu)(x) - u(x) = \int p(x, dy)(u(y) - u(x)).$$

Der lineare Operator $\mathscr{L} = p - I$ heißt **Generator** der Markovkette. Auf diskreten Zustandsräumen ist $\mathscr{L}$ ein Differenzenoperator:

$$(\mathscr{L}u)(x) = \sum_{y \in S} p(x, y)(u(y) - u(x)).$$

Beispiel (Random Walk auf $\mathbb{Z}^d$, Poissongleichung und Dirichletproblem). Für den klassischen d-dimensionalen Random Walk gilt

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

Damit ergibt sich

$$\begin{aligned} (\mathscr{L}u)(x) &= \frac{1}{2d}\sum_{i=1}^{d}(u(x+e_i)-u(x)+u(x-e_i)-u(x)) \\ &= \frac{1}{2d}\sum_{i=1}^{d}\left((u(x+e_i)-u(x))-(u(x)-u(x-e_i))\right). \end{aligned}$$

Also ist

$$\mathscr{L} \quad = \quad \frac{1}{2d}\Delta_{\mathbb{Z}^d}$$

der diskrete Laplace-Operator multipliziert mit der Übergangswahrscheinlichkeit. (11.2.10) ist also ein Randwertproblem für die *diskrete Poissongleichung*

$$(\Delta_{\mathbb{Z}^d}u)(x) \quad = \quad -2dc(x).$$

Beispielsweise ist die mittlere Austrittszeit $u(x)$ des Random Walks mit Start in x aus einer Menge D durch die minimale nichtnegative Lösung des Randwertproblems

$$\begin{aligned} \Delta_{\mathbb{Z}^d}u &= -2d \qquad \text{auf } D, \\ u &= 0 \qquad \text{auf } D^C, \end{aligned}$$

gegeben. Wollen wir die Verteilung des Austrittspunktes X_T berechnen (wie z.B. beim Ruinproblem), dann müssen wir $c \equiv 0$ setzen. In diesem Fall ist (11.2.10) ein *diskretes Dirichletproblem*: Gesucht ist eine Funktion $u : \mathbb{Z}^d \to \mathbb{R}$ mit

$$\begin{aligned} \Delta_{\mathbb{Z}^d}u &= 0 \qquad \text{auf } D, \\ u &= f \qquad \text{auf } D^C. \end{aligned}$$

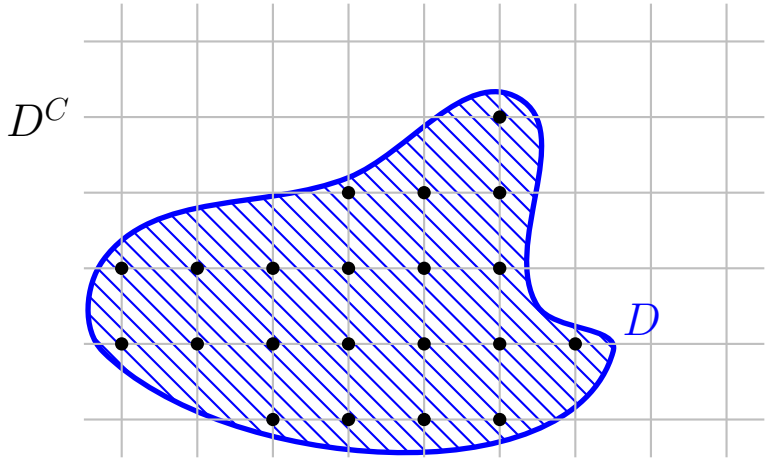


Abbildung 11.4: Diskretes Dirichletproblem auf einer Menge $D \subset \mathbb{Z}^2$.

Dirichletproblem und Austrittsverteilung

Allgemein nennen wir Funktionen $h : S \to \mathbb{R}$ mit $\mathscr{L}h = 0$ harmonisch.

Definition. *Eine nach unten beschränkte, messbare Funktion $h : S \to \mathbb{R}$ heißt **harmonisch auf der Menge D bzgl. des stochastischen Kerns** p, falls*

$$(\mathscr{L}h)(x) \quad = \quad (ph)(x) - h(x) \quad = \quad 0 \qquad \textit{für alle } x \in D$$

*gilt, d.h. falls h die **verallgemeinerte Mittelwerteigenschaft***

$$\int p(x, dy)h(y) \quad = \quad h(x) \qquad \textit{für alle } x \in D \tag{11.2.11}$$

besitzt.

Als Spezialfall von Satz 11.10 erhalten wir:

Korollar 11.11 (Stochastische Lösung des Dirichletproblems). *Die Funktion*

$$u(x) \; = \; E_x[f(X_T)\,;\, T < \infty]$$

ist die minimale nichtnegative Lösung des Dirichletproblems

$$u \textit{ harmonisch auf } D, \qquad u \; = \; f \quad \textit{auf } D^C. \tag{11.2.12}$$

Bemerkung (Lokalität). Ist S abzählbar, dann sind für die Lösung des Dirichletproblems nur die Werte von f auf dem äußeren Rand

$$\partial_{\text{ext}} D \quad = \quad \{y \in D^C \mid p(x, y) > 0 \quad \text{für ein } x \in D\}$$

relevant. In der Tat gilt für $u : S \to \mathbb{R}$ und $x \in D$:

$$(pu)(x) \quad = \quad \sum_{y \in S} p(x,y)u(y) \quad = \quad \sum_{y \in D} p(x,y)u(y) + \sum_{y \in \partial D} p(x,y)u(y),$$

d.h. $(\mathscr{L}u)(x)$ hängt nicht von den Werten von u auf $D^C \setminus \partial D$ ab.

Bemerkung (Eindeutigkeit des Dirichletproblems). (1). Im Allgemeinen können mehrere Lösungen des Dirichletproblems (11.2.12) existieren. Ist beispielsweise p der Übergangskern eines klassischen Random Walks auf $\{0, 1, 2, \ldots\}$ und $D = \{1, 2, \ldots\}$, dann sind die Funktionen $h_a(x) = ax, a \in \mathbb{R}$, alle harmonisch mit Randwerten $h_a(0) = 0$. Ebenso ist die Lösung nicht eindeutig, falls ein $z \in S$ mit $P_z[T = \infty] \neq 0$ existiert, denn in diesem Fall ist $h(x) = P_x[T = \infty]$ eine nichttriviale harmonische Funktion mit Nullrandwerten.

(2). Ist die Funktion f beschränkt, und ist die Austrittszeit T für alle $x \in S$ P_x-fast sicher endlich, dann ist u die eindeutige beschränkte Lösung von (11.2.12). Dies kann man z.B. mit dem Stoppsatz für Martingale beweisen.

Satz 11.10 und Korollar 11.11 sind erste Aspekte weitreichender Beziehungen zwischen Wahrscheinlichkeitstheorie und Analysis (Potentialtheorie) mit fundamentalen Konsequenzen auch für andere Gebiete der Mathematik wie z.B. Diskrete Mathematik, Differentialgeometrie, Numerik und mathematische Physik. Wir erwähnen hier einige wichtige Gesichtspunkte und Konsequenzen des gefundenen Zusammenhangs. Dazu setzen wir $T < \infty$ P_x-fast sicher für alle $x \in S$ voraus. Unter dieser Annahme ist

$$u(x) \quad = \quad E_x[f(X_T)] \tag{11.2.13}$$

für eine nichtnegative bzw. beschränkte Funktion f auf D^C die minimale nichtnegative, bzw. die eindeutige beschränkte Lösung des Dirichletproblems.

Monte-Carlo- Methode zur Berechnung harmonischer Funktionen: Nach dem Gesetz der großen Zahlen gilt

$$u(x) \quad \approx \quad \frac{1}{k} \sum_{i=1}^{k} f(X^{(i)}_{T^{(i)}}) \qquad \text{für große } k,$$

wobei $X^{(1)}, X^{(2)}, \ldots$ unabhängige Markovketten mit Start in x und Übergangskern p sind, und $T^{(i)}$ die Austrittszeit von $X^{(i)}$ aus der Menge D bezeichnet. Die Simulation von Markovketten kann daher in sehr allgemeinem Rahmen zur näherungsweisen Berechnung harmonischer Funktionen verwendet werden.

Stochastische Darstellung der Lösung des Dirichletproblems als Pfadintegral: Nach (11.2.13) können wir die harmonische Funktion u schreiben als Integral

$$u(x) \quad = \int\limits_{S^{\{0,1,2,\ldots\}}} f(X_T(\omega))\, P(d\omega)$$

über den Raum aller diskreten Pfade auf S. Ähnliche Pfadintegraldarstellungen spielen in der Quantenphysik eine wichtige Rolle, siehe z.B. die Lecture Notes von R. Feynman.

Integralformel für harmonische Funktionen: Sei $\mu_x := P_x \circ X_T^{-1}$ die Austrittsverteilung der Markovkette mit Start in x. Dann gilt:

$$u(x) \quad = \int\limits_{D^C} f(y)\, \mu_x(dy).$$

Die Austrittsverteilung μ_x ist also das **harmonische Maß** der Potentialtheorie, das eine Berechnung harmonischer Funktionen aus den Randwerten ermöglicht.

Beispiele harmonischer Funktionen

Diskrete Zustandsräume

Ist S abzählbar, dann ist

$$h_y(x) \quad := \quad P_x[T < \infty \text{ und } X_T = y]$$

für jedes $y \in D^C$ eine nichtnegative, beschränkte, harmonische Funktion auf D mit Randwerten

$$h_y(x) \quad = \quad I_{\{y\}}(x) \qquad \text{für alle } x \in D^C.$$

Eine Lösung u des Dirichletproblems zu beliebigen Randwerten $f : D^C \to \mathbb{R}_+$ erhält man als Linearkombination der Funktionen h_y: Gilt $P_x[T = \infty] = 0$ für alle $x \in S$, dann gibt es genau eine beschränkte Lösung des Dirichlet-Problems. Damit folgt, dass die Funktionen $h_y, y \in D^C$, eine Basis des Vektorraums aller beschränkten, harmonischen Funktionen bilden. Wir erhalten also einen Zusammenhang zwischen beschränkten harmonischen Funktionen und den möglichen Austrittspunkten $y \in D^C$ der Markovkette.

Beispiel. (1). *Ruinproblem:* Für den Random Walk auf $\{a, a+1, \ldots, b\} \subset \mathbb{Z}$ mit Übergangskern $p(x, x+1) = p, p(x, x-1) = q = 1-p$, gilt

$$h_a(x) \quad = \quad P_x[X_T = a] \quad = \quad \frac{h(b) - h(x)}{h(b) - h(a)}$$

und

$$h_b(x) \quad = \quad P_x[X_T = b] \quad = \quad \frac{h(x) - h(a)}{h(b) - h(a)}$$

mit

$$h(x) \quad := \quad \begin{cases} x & \text{für } p = q \\ (q/p)^x & \text{für } p \neq q \end{cases}.$$

Die Funktionen h_a und h_b bilden eine Basis des Vektorraums $\{c \cdot h + d \mid c, d \in \mathbb{R}\}$ aller harmonischer Funktionen.

(2). *Eine transiente Markovkette auf $\mathbb{Z}$:* Sei $p \in (\frac{1}{2}, 1)$ und $q = 1 - p$. Wir betrachten die Markovkette (X_n, P_x) auf $\mathbb{Z}$ mit den folgenden Übergangswahrscheinlichkeiten:

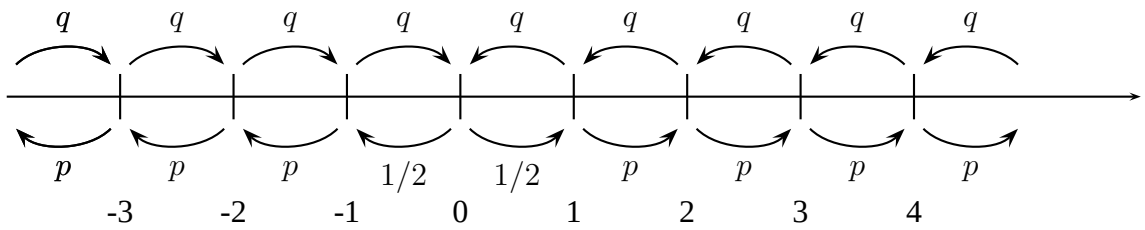


Abbildung 11.5: Übergangswahrscheinlichkeiten der transienten Markovkette (X_n, P_x)

Für $x > 0$ gilt

$$p(x, x+1) \quad = \quad p \quad > \quad q \quad = \quad p(x, x-1),$$

für $x < 0$ dagegen umgekehrt

$$p(x, x+1) \quad = \quad q \quad < \quad p \quad = \quad p(x, x-1).$$

Daher haben die Ereignisse $\{\lim X_n = \infty\}$ und $\{\lim X_n = -\infty\}$ beide positive Wahrscheinlichkeit. Die Funktion

$$h_+(x) \quad := \quad \begin{cases} 1 - \frac{1}{2}\left(\frac{q}{p}\right)^x & \text{für } x > 0 \\ \frac{1}{2}\left(\frac{q}{p}\right)^{-x} & \text{für } x \leq 0 \end{cases}$$

ist harmonisch mit Randbedingungen

$$\lim_{x \to \infty} h_+(x) \quad = \quad 1$$

und

$$\lim_{x \to -\infty} h_+(x) \quad = \quad 0.$$

Entsprechend ist $h_-(x) = h_+(-x)$ harmonisch mit

$$\lim_{x\to-\infty} h_-(x) \quad = \quad 1$$

und

$$\lim_{x\to\infty} h_-(x) \quad = \quad 0,$$

und jede harmonische Funktion ist eine Linearkombination von h_+ und h_-. Durch Bedingen auf den ersten Schritt der Markovkette zeigt man

$$h_+(x) \quad = \quad P_x[\lim X_n = \infty]$$

und

$$h_-(x) \quad = \quad P_x[\lim X_n = -\infty].$$

Die harmonischen Funktionen h_+ und h_- beschreiben in diesem Fall die möglichen Asymptotiken der Markovkette.

Rotationssymmetrischer Fall

Wir betrachten eine Markovkette auf $S = \mathbb{R}^d$, deren Übergangsverteilungen $p(x, dy)$ für jedes x rotationssymmetrisch mit Zentrum x sind.

Beispielsweise sei $X_n = x + \sum_{i=1}^{n} Y_i$ ein Random Walk, dessen Inkremente Y_i unabhängig mit identischer rotationssymmetrischer Verteilung sind. Dann ist jede Funktion $u \in \mathcal{C}^2(\mathbb{R}^d)$ mit

$$\Delta u \quad = \quad \sum_{i=1}^{d} \frac{\partial^2 u}{\partial x_i^2} \quad = \quad 0$$

(also jede harmonische Funktion des Laplaceoperators) auch eine harmonische Funktion des Übergangskerns p, falls u für alle $x \in \mathbb{R}^d$ bzgl. $p(x, dy)$ integrierbar ist. Aus der Greenschen Formel folgt nämlich die Mittelwerteigenschaft

$$u(x) \quad = \quad \text{Mittelwert von } u \text{ auf } \partial B_r(x)$$

für alle Sphären $\partial B_r(x) = \{y \in \mathbb{R}^d \,:\, |y - x| = r\}, r > 0$, siehe z.B. [Forster, Analysis III]. Da $p(x, dy)$ rotationssymmetrisch ist, erhalten wir durch Integration über den Radius:

$$u(x) \quad = \quad \int p(x, dy)u(y),$$

d.h. u ist in der Tat harmonisch bzgl. p.

Mittlere Aufenthaltszeiten und Greenfunktion

Die mittlere Aufenthaltszeit

$$u(x) \quad = \quad E_x\left[\sum_{n=0}^{T-1} I_B(X_n)\right] \quad = \quad \sum_{n=0}^{\infty} P_x[X_n \in B\ ,\ n < T],$$

einer Markovkette mit Übergangskern p in einer Menge $B \in \mathcal{S}$ vor Austritt aus D löst das Randwertproblem

$$\begin{aligned} u - pu &= I_A \qquad \text{auf } D \\ u &= 0 \qquad \text{auf } D^C. \end{aligned}$$

Wir betrachten nun den diskreten Fall: Sei S abzählbar, $D \subset S$, und sei

$$B_y^D \quad := \quad \sum_{n=0}^{T-1} I_{\{y\}}(X_n), \qquad y \in S,$$

die Anzahl der Besuche der Markovkette in y vor Austritt aus D. Für die mittlere Anzahl der Besuche in y bei Start in x gilt

$$E_x[B_y^D] \quad = \quad E_x\left[\sum_{n=0}^{\infty} I_{\{X_n \in B\,,\, n<T\}}\right] \quad = \quad \sum_{n=0}^{\infty} p_n^D(x,y),$$

wobei

$$p_n^D(x,y) \quad = \quad P_x[X_n = y\ ,\ n < T]$$

die n-Schritt-Übergangswahrscheinlichkeit der Markovkette mit Absorption bei Austritt aus D bezeichnet.

Definition. *Die durch*

$$G^D(x,y) \quad := \quad \sum_{n=0}^{\infty} p_n^D(x,y)$$

*definierte Funktion $G^D : S \times S \to [0,\infty]$ heißt **Greensche Funktion** der Markovkette im Gebiet D.*

Korollar 11.12. *(1). $G^D(\bullet, y)$ ist die minimale Lösung des Randwertproblems*

$$\begin{aligned} (I-p)G^D(\bullet,y) &= I_{\{y\}} \qquad \textit{auf } D, \\ G^D(\bullet,y) &= 0 \qquad \textit{auf } D^C. \end{aligned}$$

(2). Für alle Funktionen $f : S \to [0, \infty]$ gilt

$$E_x\left[\sum_{n=0}^{T-1} f(X_n)\right] \quad = \quad (G^D f)(y).$$

Beweis. Die erste Aussage folgt unmittelbar aus Satz 11.10. Für eine Funktion $f \geq 0$ gilt:

$$E_x\left[\sum_{n=0}^{T-1} f(X_n)\right] \;=\; E_x\left[\sum_{n=0}^{T-1}\sum_{y\in S} f(y)\cdot I_{\{y\}}(X_n)\right] \;=\; \sum_{y\in S} G^D(x,y)f(y) \;=\; (G^D f)(y).$$

□

Beispiel (Random Walk auf $\mathbb{Z}^d$). Die Greensche Funktion des klassischen Random Walks auf $\mathbb{Z}^d$ ist die minimale nichtnegative Lösung des Randwertproblems

$$\begin{aligned} \Delta_{\mathbb{Z}^d} G^D(\bullet, y) &= -2dI_{\{y\}} \qquad \text{auf } D, \\ G^D(\bullet, y) &= 0 \qquad \text{auf } D^C. \end{aligned}$$

Sie ist damit ein diskretes Analogon zur Greenschen Funktion der Analysis, die als Fundamentallösung der Poissongleichung definiert ist. Beispielsweise erhält man für den klassischen eindimensionalen Random Walk als Greensche Funktion eines Intervalls $D = \{a, a+, \ldots, b\} \subset \mathbb{Z}$:

$$G^D(x,y) \quad = \quad \begin{cases} 2\dfrac{(b-y)(x-a)}{b-a} & \text{für } a \leq x < y \\ 2\dfrac{(y-a)(b-x)}{b-a} & \text{für } y \leq x \leq b \end{cases}.$$

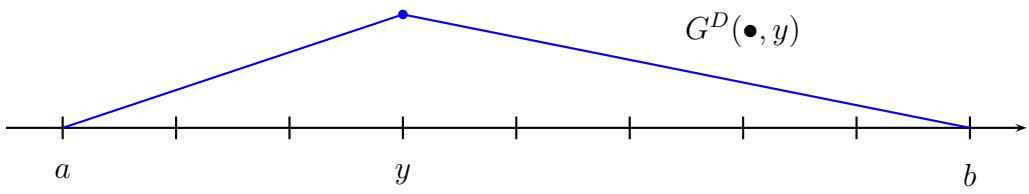


Abbildung 11.6: Darstellung des Graphen der Funktion $G^D(\bullet, y)$.

11.3 Rekurrenz und Transienz

Sei $p(x, y)$ $(x, y \in S)$ eine stochastische Matrix auf einer abzählbaren Menge S. Wir betrachten eine zeithomogene Markovkette (X_n, P_x) mit Übergangsmatrix p im kanonischen Modell, d.h.

$$\Omega \;=\; S^{\{0,1,2,\ldots\}}, \quad X_n(\omega) \;=\; \omega_n, \quad \mathcal{A} \;=\; \sigma(X_n \,|\, n \geq 0),$$

und P_x ist die Verteilung der Markovkette bei Start in x. Für $y \in S$ sei

$$B_y(\omega) \quad = \quad \sum_{n=0}^{\infty} I_{\{y\}}(X_n(\omega))$$

die *Anzahl der Besuche* (*Aufenthaltszeit*) der Markovkette im Punkt y. Wir wollen untersuchen, ob die Markovkette immer wieder zu ihrem Startpunkt zurückkehrt.

Definition. *Ein Punkt $x \in S$ heißt **transient**, falls $P_x[B_x = \infty] = 0$ gilt, und **rekurrent**, falls $P_x[B_x = \infty] = 1$.*

Sei nun

$$G(x,y) \quad = \quad E_x[B_y] \quad = \quad \sum_{n=0}^{\infty} p^n(x,y)$$

die mittlere Anzahl der Besuche der Markovkette im Punkt y bei Start in x. Offensichtlich ist x transient, wenn

$$G(x,x) \quad = \quad E_x[B_x] \quad < \quad \infty$$

gilt. Wir werden in Korollar 11.15 zeigen, dass umgekehrt x rekurrent ist, wenn $G(x,x) = \infty$ gilt. Insbesondere ergibt sich ein 0-1-Gesetz: Jeder Punkt ist entweder transient oder rekurrent. Allgemeiner werden wir sehen, dass bei irreduziblen Markovketten sogar entweder alle Punkte transient oder alle Punkte rekurrent sind – wir nennen die Markovkette in diesem Fall *transient* bzw. *rekurrent*.

Intuitiv können wir diese Dichotomie folgendermaßen erklären: Jedes Mal, wenn die Markovkette zum Startpunkt x zurückkehrt, startet sie aufgrund der Markoveigenschaft wieder neu in diesem Punkt – unabhängig vom vorherigen Verlauf. Kehrt die Kette also mit Wahrscheinlichkeit 1 wieder zum Startpunkt zurück, dann kehrt sie auch mit Wahrscheinlichkeit 1 immer wieder, also unendlich oft nach x zurück. Ist die Markovkette zudem irreduzibel, dann erreicht sie jeden festen Punkt y auf jeder Exkursion mit einer konstanten strikt positiven Wahrscheinlichkeit – trifft also insgesamt den Punkt y mit Wahrscheinlichkeit 1 unendlich oft.

Kehrt die Kette dagegen mit einer strikt positiven Wahrscheinlichkeit $\varepsilon > 0$ nicht zum Startpunkt x zurück, dann wird sie auch bei jedem weiteren Erreichen von x mit derselben Wahrscheinlichkeit ε nicht wieder zurückkehren – unabhängig vom vorherigen Verlauf. Also wird sie mit Wahrscheinlichkeit 1 schließlich nicht mehr nach x zurückkehren – sie durchläuft also jeden Punkt nur endlich oft.

Um diese *Dichotomie von Rekurrenz und Transienz* rigoros zu beweisen, benötigen wir eine Markoveigenschaft für die *zufälligen* (!) Rückkehrzeiten zum Startpunkt. Bevor wir eine entsprechende „starke Markoveigenschaft“ beweisen, betrachten wir schon mal eine Anwendung auf mehrdimensionale Random Walks.

Beispiel (Rekurrenz und Transienz von Random Walks in $\mathbb{Z}^d$). Sei (X_n, P_x) der klassische Random Walk auf $\mathbb{Z}^d$ mit Übergangswahrscheinlichkeiten $p(x,y) = \frac{1}{2d}$ falls $|x-y| = 1$, $p(x,y) = 0$ sonst. Wir untersuchen Rekurrenz und Transienz in Abhängigkeit von der Dimension d:

$d = 1$**:** Im eindimensionalen Fall erhalten wir für die Rückkehrwahrscheinlichkeiten zum Ausgangspunkt x mithilfe der Stirling-Approximation:

$$\begin{aligned} p^{2n}(x,x) &= \binom{2n}{n} \cdot 2^{-2n} = \frac{(2n)!}{(n!)^2} 2^{-2n} \\ &\sim \frac{\sqrt{4\pi n}}{2\pi n} \frac{(2n)^{2n}}{n^{2n}} \cdot 2^{-2n} = \frac{1}{\sqrt{\pi n}}. \end{aligned}$$

Also gilt $G(x,x) = \sum_{n=0}^{\infty} p^n(x,x) = \infty$, d.h. jeder Punkt $x \in \mathbb{Z}$ ist *rekurrent*.

$d = 2$**:** Beim klassischen Random Walk $X_n = (X_n^{(1)}, X_n^{(2)})$ auf $\mathbb{Z}^2$ sind die Komponenten $X_n^{(1)}$ und $X_n^{(2)}$ nicht unabhängig.

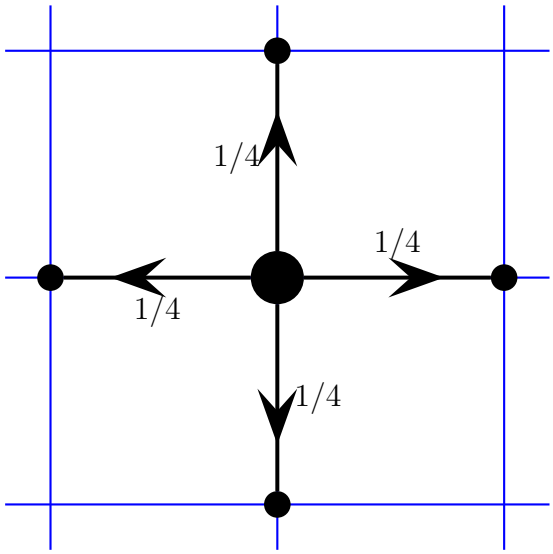


Abbildung 11.7: Übergangswahrscheinlichkeiten des klassischen Random Walks.

Durch eine $45°$ Drehung des Koordinatensystems, können wir den Prozess aber in einen zweidimensionalen Random Walk

$$Y_n = (X_n^{(1)} + X_n^{(2)}, X_n^{(1)} - X_n^{(2)})$$

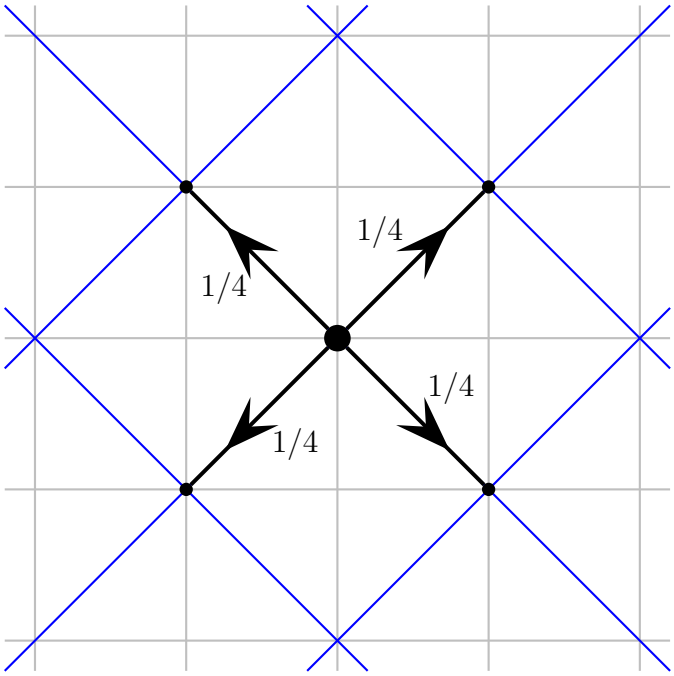


Abbildung 11.8: Übergangswahrscheinlichkeiten des um $45°$ gedrehten Random Walks.

überführen, dessen Komponenten $Y_n^{(1)}$ und $Y_n^{(2)}$ unabhängige eindimensionale Random Walks sind. Offensichtlich gilt:

$$X_n \text{ rekurrent} \iff Y_n \text{ transient.}$$

Die Übergangswahrscheinlichkeiten für Y_n sind

$$\begin{aligned} p^{2n}(x,x) &= P_x[Y_{2n}^{(1)} = x_1, Y_{2n}^{(2)} = x_2] = P_{x_1}[Y_{2n}^{(1)} = x_1] \cdot P_{x_2}[Y_{2n}^{(2)} = x_2] \\ &= \left(\binom{2n}{n} \cdot 2^{-2n} \right)^2 \sim \frac{1}{\pi n}. \end{aligned}$$

Also gilt erneut $G(x,x) = \infty$, d.h. jedes $x \in \mathbb{Z}^2$ ist *rekurrent*.

$d = 3$**:** Betrachten wir einen dreidimensionalen Random Walk

$$Y_n = (X_n^{(1)}, X_n^{(2)}, X_n^{(3)}),$$

dessen Komponenten $X_n^{(i)}$ unabhängige klassische Random Walks auf $\mathbb{Z}^1$ sind, dann gilt entsprechend

$$p^{2n}(x,x) = \left(\binom{2n}{n} \cdot 2^{-2n} \right)^3 \sim \frac{1}{(\pi n)^{3/2}},$$

und damit

$$G(x,x) = \sum_{n=0}^{\infty} p^{2n}(x,x) < \infty$$

Der Prozess ist also *transient*. Auch der klassische Random Walk auf $\mathbb{Z}^3$ ist transient – der Beweis erfordert allerdings etwas mehr Kombinatorik, da sich der Prozess in Dimension 3 nicht durch eine Drehung in einen Prozess mit unabhängigen Komponenten überführen lässt. Die Details werden in einer Übungsaufgabe ausgeführt. Analog folgt Transienz in höheren Dimensionen. Zwischen Dimension 2 und 3 gibt es also einen Übergang von rekurrentem zu transientem Verhalten. Anschaulich steht in Dimension $d > 2$ soviel Raum zur Verfügung, dass der Random Walk der Startpunkt schließlich nicht mehr trifft.

Starke Markoveigenschaft

Wir beweisen nun die angekündigte Erweiterung der Markoveigenschaft auf zufällige Zeiten (Stoppzeiten). Die Information, die über einen stochastischen Prozess $(X_n)_{n\geq 0}$ bis zur Zeit n vorliegt, wird beschrieben durch die σ-Algebra

$$\mathcal{F}_n \quad = \quad \sigma(X_0, X_1, \ldots, X_n).$$

Sei $T : \Omega \to \{0, 1, 2, \ldots\} \cup \{\infty\}$ eine nichtnegative ganzzahlige Zufallsvariable. T heißt eine **Stoppzeit** (bzgl. der σ-Algebren $\mathcal{F}_n$), falls

$$\{T = n\} \in \mathcal{F}_n \qquad \text{für alle } n \geq 0 \text{ gilt.}$$

Nach dem Faktorisierungslemma ist T genau dann eine Stoppzeit bzgl. $(\mathcal{F}_n)_n$, wenn $I_{\{T=n\}}$ für jedes n eine Funktion von $X_0, \ldots, X_n$ ist. Anschaulich bedeutet dies, dass aufgrund der Information, die bis zur Zeit n vorliegt, entscheidbar ist, ob T den Wert n annimmt.

Beispiel (Trefferzeiten). (1). Die **erste Treffer- bzw. Rückkehrzeit**

$$T_B \quad = \quad \min\{n \geq 1 \mid X_n \in B\} \qquad (\min \emptyset := \infty)$$

einer messbaren Teilmenge B des Zustandsraumes S ist eine Stoppzeit, denn es gilt

$$\{T_B = n\} \quad = \quad \{X_1 \in B^C, \ldots, X_{n-1} \in B^C, X_n \in B\} \quad \in \quad \mathcal{F}_n \qquad \text{für alle } n \geq 0.$$

Hat man beispielsweise beschlossen, eine Aktie zu verkaufen, sobald ihr Kurs X_n den Wert λ überschreitet, dann ist der Verkaufszeitpunkt gleich $T_{(\lambda,\infty)}$, also eine Stoppzeit.

(2). Die **letzte Besuchszeit**

$$L_B \quad := \quad \sup\{n \geq 0 \mid X_n \in B\} \qquad (\sup \emptyset := 0)$$

ist dagegen in der Regel keine Stoppzeit (Übung). Um zu entscheiden, ob $L_B = n$ gilt, benötigt man nämlich Informationen über die zukünftige Entwicklung des Prozesses.

Die Information, die bis zu einer Stoppzeit vorliegt, wird beschrieben durch die σ-Algebra

$$\mathcal{F}_T \quad = \quad \{A \in \mathcal{A} \mid A \cap \{T = n\} \in \mathcal{F}_n \quad \text{für alle } n \geq 0\},$$

der *„bis zur Zeit T beobachtbaren" Ereignisse*. Durch maßtheoretische Induktion zeigt man, dass eine Abbildung $Y : \Omega \to \mathbb{R}$ genau dann bzgl. $\mathcal{F}_T$ messbar ist, wenn $Y \cdot I_{\{T=n\}}$ für jedes $n \geq 0$ $\mathcal{F}_n$-messbar, also eine Funktion von $X_0, \ldots, X_n$ ist. Insbesondere ist die Position X_T des Prozesses (X_n) zur Stoppzeit T eine $\mathcal{F}_T$-messbare Zufallsvariable, denn für $n \geq 0$ ist

$$X_T \cdot I_{\{T=n\}} \quad = \quad X_n \cdot I_{\{T=n\}} \qquad \mathcal{F}_n\text{-messbar.}$$

Wir setzen nun wieder voraus, dass (X_n, P_x) eine zeithomogene Markovkette im kanonischen Modell ist.

Satz 11.13 (Starke Markoveigenschaft). *Ist $T : \Omega \to \{0, 1, 2, \ldots\} \cup \{\infty\}$ eine Stoppzeit bzgl. der σ-Algebren $\mathcal{F}_n = \sigma(X_0, X_1, \ldots, X_n)$, dann gilt*

$$E_\nu[F(X_T, X_{T+1}, \ldots) \mid \mathcal{F}_T] \quad = \quad E_{X_T}[F(X_0, X_1, \ldots)] \qquad P_\nu\text{-fast sicher auf } \{T < \infty\}$$

für alle Wahrscheinlichkeitsverteilungen ν auf $(S, \mathcal{S})$ und alle messbaren Funktionen $F : S^{\{0,1,2,\ldots\}} \to \mathbb{R}_+$, wobei $F(X_T, X_{T+1}, \ldots)$ auf $\{T = \infty\}$ willkürlich definiert ist.

Beweis. Sei $\theta(x_0, x_1, \ldots) = (x_1, x_2, \ldots)$ der Shiftoperator auf $S^{\{0,1,2,\ldots\}}$. Wir müssen zeigen, dass

$$E_\nu[F \circ \theta^T \mid \mathcal{F}_T] \cdot I_{\{T<\infty\}} \quad = \quad E_{X_T}[F] \cdot I_{\{T<\infty\}} \qquad P_\nu\text{-fast sicher} \tag{11.3.1}$$

gilt, wobei wir die rechte Seite für $T = \infty$ gleich 0 setzen. Für $A \in \mathcal{F}_T$ und $n \geq 0$ gilt $A \cap \{T = n\} \in \mathcal{F}_n$, also nach der Markoveigenschaft:

$$\begin{aligned} E_\nu[F \circ \theta^T \,;\, A \cap \{T = n\}] &= E_\nu[F \circ \theta^n \,;\, A \cap \{T = n\}] \\ &= E_\nu\left[E_{X_n}[F] \,;\, A \cap \{T = n\}\right] \\ &= E_\nu\left[E_{X_T}[F] \,;\, A \cap \{T = n\}\right] \end{aligned}$$

Durch Summieren über n erhalten wir:

$$E_\nu[F \circ \theta^T \,;\, A \cap \{T < \infty\}] \quad = \quad E_\nu\left[E_{X_T}[F] \,;\, A \cap \{T < \infty\}\right].$$

Also stimmen die Integrale beider Seiten von (11.3.1) über eine beliebige Menge $A \in \mathcal{F}_T$ überein. Da beide Seiten in (11.3.1) $\mathcal{F}_T$-messbar sind, folgt, dass diese P_ν-fast sicher übereinstimmen. □

Anschaulich startet eine zeithomogene Markovkette also auch zu einer Stoppzeit T neu im Zustand X_T, d.h. der weitere Verlauf ist unabhängig vom vorherigen Verlauf gegeben den gegenwärtigen Zustand X_T.

Rekurrenz und Transienz von einzelnen Zuständen

Mithilfe der starken Markoveigenschaft können wir die Verteilung der Aufenthaltszeit B_y der Markovkette in einem Punkt $y \in S$ aus den Trefferwahrscheinlichkeiten

$$f_{x,y} \quad := \quad P_x[T_y < \infty]$$

berechnen. Hierbei bezeichnen wir mit

$$T_y \quad = \quad \min\{n \geq 1 \,:\, X_n = y\}$$

die erste Trefferzeit des Zustandes y, bzw. die erste Rückkehrzeit nach y, falls die Markovkette in y startet.

Satz 11.14. *Für alle $x, y \in S$ gilt*

$$P_x[B_y \geq n] \quad = \quad \begin{cases} f_{x,y} \cdot f_{y,y}^{n-1} & \textit{falls } x \neq y \\ f_{y,y}^{n-1} & \textit{falls } x = y \end{cases}.$$

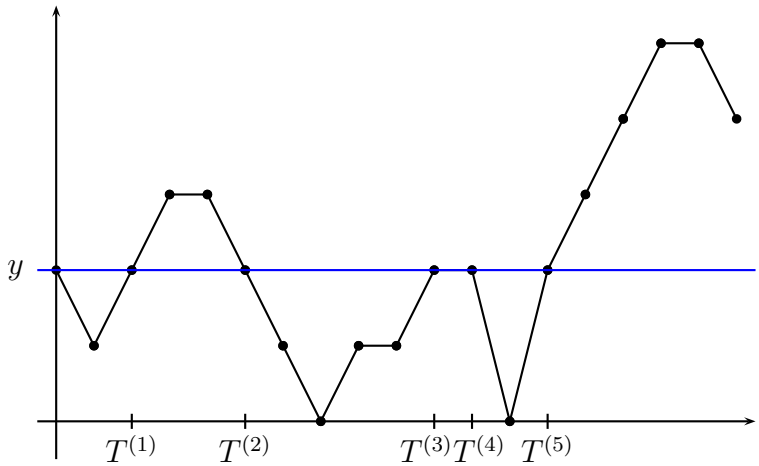
Insbesondere ist jedes $y \in S$ entweder rekurrent oder transient, und es gilt:

$$\begin{aligned} y \textit{ rekurrent} \quad &\Longleftrightarrow \quad f_{y,y} = 1, \\ y \textit{ transient} \quad &\Longleftrightarrow \quad f_{y,y} < 1. \end{aligned}$$

Beweis. Sei $T^{(0)} := 0$, und sei

$$T^{(n)} \quad := \quad T^{(n-1)} + T_y \circ \theta^{T^{(n-1)}}$$

die n-te Besuchszeit (bei Start außerhalb von y) bzw. Rückkehrzeit (bei Start in y) des Zustands y.



Es folgt

$$\begin{aligned} P_x[T^{(n)} < \infty] &= P_x[T^{(n-1)} < \infty \text{ und } T_y \circ \theta^{T^{(n-1)}} < \infty] \\ &= E_x[P_x[T_y \circ \theta^{T^{(n-1)}} < \infty \mid \mathcal{F}_{T^{(n-1)}}] \,;\, T^{(n-1)} < \infty] \\ &= E_x[P_{X_{T^{(n-1)}}}[T_y < \infty] \,;\, T^{(n-1)} < \infty] \\ &= P_y[T_y < \infty] \cdot P_x[T^{(n-1)} < \infty], \end{aligned}$$

also durch Induktion nach n:

$$P_x[T^{(n)} < \infty] \quad = \quad P_x[T_y < \infty] \cdot P_y[T_y < \infty]^{n-1} \quad = \quad f_{x,y} \cdot f_{y,y}^{n-1}.$$

Die Aussage folgt wegen

$$P_x[B_y \geq n] \quad = \quad \begin{cases} P_x[T^{(n)} < \infty] & \text{falls } x \neq y, \\ P_x[T^{(n-1)} < \infty] & \text{falls } x = y. \end{cases}$$

□

Bemerkung. Die Wahrscheinlichkeiten $f_{x,y}$ (und damit die Verteilungen der Aufenthaltszeiten) kann man im Prinzip durch Lösen eines Dirichletproblems berechnen: Nach Korollar 11.11 bilden die Trefferwahrscheinlichkeiten

$$h(x) \quad = \quad P_x[\widetilde{T}_y < \infty], \quad \widetilde{T}_y := \min\{n \geq 0 \mid X_n = y\},$$

die minimale nichtnegative Lösung von

$$ph \quad = \quad h \qquad \text{auf } S \setminus \{y\}; \qquad h(y) = 1.$$

Wegen $\widetilde{T}_y = T_y$ P_x-fast sicher für alle $x \neq y$ folgt $f_{x,y} = h(x)$ für $x \neq y$, und

$$f_{y,y} \quad = \quad \sum_{x \in S} p(y,x) f_{x,y} \quad = \quad (ph)(y).$$

Beispiel (Kartenhaus, Maschinenerneuerung). Wir betrachten eine Markovkette mit Zustandsraum $S = \{0, 1, 2, \ldots\}$ und Übergangswahrscheinlichkeiten

$$p(x, x+1) \quad = \quad 1 - \varrho_x, \qquad p(x, 0) \quad = \quad \varrho_x, \qquad \varrho_x \in (0, 1).$$

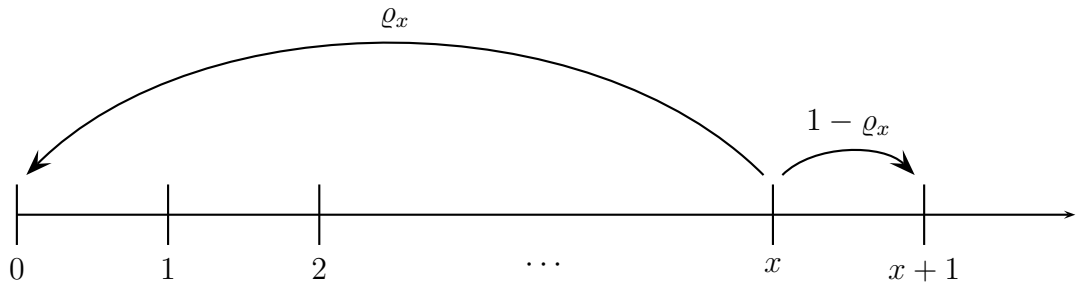


Abbildung 11.9: Übergangswahrscheinlichkeiten der durch p gegebenen Markovkette.

Hier gilt

$$P_0[T_0 > n] \quad = \quad \prod_{x=0}^{n-1} (1 - \varrho_x),$$

also:

$$0 \text{ rekurrent} \iff P_0[T_0 = \infty] = \prod_{x=0}^{\infty} (1 - \varrho_x) = 0 \iff \sum_{x=0}^{\infty} \varrho_x = \infty$$

Aus Satz 11.14 folgt unmittelbar die schon oben erwähnte Charakterisierung rekurrenter Zustände über die Greensche Funktion:

Korollar 11.15 (Rekurrenzkriterium). *Für alle $x \in S$ gilt*

$$G(x, x) \quad = \quad \frac{1}{1 - f_{x,x}} \quad = \quad \frac{1}{P_x[T_x = \infty]} \qquad \textit{falls } P_x[T_x = \infty] > 0,$$

bzw. $G(x, x) = \infty$ falls $P_x[T_x = \infty] = 0$. Insbesondere ist x genau dann rekurrent, wenn $G(x, x)$ unendlich ist.

Beweis. Für $x \in S$ gilt nach Satz 11.14:

$$\begin{aligned} G(x, x) \quad &= \quad E_x[B_x] \quad = \quad \sum_{n=1}^{\infty} P_x[B_x \geq n] \\ &= \quad \sum_{n=1}^{\infty} f_{x,x}^{n-1} \quad = \quad \sum_{n=0}^{\infty} f_{x,x}^{n}. \end{aligned}$$

□

Leider ist das Kriterium zwar für die Theorie wichtig, aber praktisch nur selten einsetzbar. Leichter verifizierbare hinreichende Bedingungen für Rekurrenz und Transienz basieren auf stochastischen Lyapunovfunktionen und dem Martingalkonvergenzsatz, s. [Stochastische Analysis].

Kommunikationsklassen und globale Rekurrenz

Wir wollen nun untersuchen, wie die Rekurrenz verschiedener Zustände $x, y \in S$ miteinander zusammenhängt.

Definition. *Der Zustand y heißt **erreichbar** von x für die Markovkette (X_n, P_x), falls*

$$P_x[T_y < \infty] > 0$$

gilt.

Bemerkung. (1). Ein Zustand y ist genau dann erreichbar von x, wenn ein $n \in \mathbb{N}$ mit $p^n(x,y) > 0$ existiert. Insbesondere gilt für $y \neq x$:

$$y \text{ ist erreichbar von } x \quad \Longleftrightarrow \quad G(x,y) > 0.$$

(2). Ist y erreichbar von x und z erreichbar von y, dann ist z erreichbar von x.

(3). Ist die Übergangsmatrix irreduzibel, dann ist jeder Zustand von jedem anderen Zustand aus erreichbar.

Wir wollen zeigen, dass mit einem Zustand $x \in S$ auch jeder von x aus erreichbare Zustand y rekurrent ist. Dazu bemerken wir zunächst:

Lemma 11.16. *Für $x, y \in S$ mit $y \neq x$ gilt*

$$G(x,y) \quad = \quad P_x[T_y < \infty] \cdot G(y,y).$$

Beweis. Für $y \neq x$ gilt P_x-fast sicher $X_0 \neq y$, also

$$B_y \quad = \quad B_y \circ \theta^{T_y} \qquad \text{auf } \{T_y < \infty\}.$$

Mit der starken Markoveigenschaft folgt

$$\begin{aligned} E_x[B_y] &= E_x[B_y \,;\, T_y < \infty] \quad = \quad E_x[B_y \circ \theta^{T_y} \,;\, T_y < \infty] \\ &= E_y[B_y] \cdot P_x[T_y < \infty]. \end{aligned}$$

□

Satz 11.17. *Ist x rekurrent, und y von x aus erreichbar, dann ist auch x von y aus erreichbar, y ist rekurrent, und es gilt*

$$B_y = \infty \quad P_x\text{-fast sicher} \qquad \text{und} \qquad B_x = \infty \quad P_y\text{-fast sicher.}$$

Insbesondere gilt also

$$G(x,y) \quad = \quad G(y,x) \quad = \quad G(y,y) \quad = \quad \infty.$$

Beweis. (1). *y ist rekurrent:* Da y von x aus erreichbar ist, existiert $m \geq 0$ mit $p^m(x,y) > 0$. Nach dem Lemma folgt:

$$\begin{aligned} G(y,y) \;&\geq\; G(x,y) \quad \geq \quad \sum_{n=0}^{\infty} p^{n+m}(x,y) \\ &\geq \sum_{n=0}^{\infty} p^n(x,x) p^m(x,y) \\ &= \underbrace{G(x,x)}_{=\infty} \cdot \underbrace{p^m(x,y)}_{>0} \quad = \quad \infty. \end{aligned}$$

(2). *Wir zeigen $P_y[B_x = \infty] = 1$:* Da y von x aus erreichbar und x rekurrent ist, gilt nach der starken Markoveigenschaft

$$\begin{aligned} 0 \quad < \quad P_x[T_y < \infty] \;&\overset{x \text{ rek.}}{=}\; P_x[T_y < \infty, T_x \circ \theta_{T_y} < \infty] \\ &\overset{\text{SME}}{=}\; P_x[T_y < \infty] \cdot P_y[T_x < \infty], \end{aligned}$$

also $f_{y,x} = P_y[T_x < \infty] = 1$. Da x rekurrent ist, gilt zudem $f_{x,x} = 1$, also nach Satz 11.14

$$P_y[B_x = \infty] \quad = \quad \lim_{n\to\infty} (f_{y,x} \cdot f_{x,x}^{n-1}) \quad = \quad 1.$$

Insbesondere ist x von y aus erreichbar.

(3). Analog erhalten wir $P_x[B_y = \infty] = 1$ durch Vertauschen der Rolle von x und y.

□

Der Satz zeigt, dass für eine Markovkette mit irreduzibler Übergangsmatrix und einem rekurrenten Zustand alle Zustände rekurrent sind, und jeder Zustand bei beliebiger Startverteilung mit Wahrscheinlichkeit 1 unendlich oft durchlaufen wird:

Korollar 11.18 (Dichotomie von Rekurrenz und Transienz). *Für eine zeithomogene Markovkette mit irreduzibler Übergangsmatrix gilt entweder*

(1). Alle $x \in S$ sind rekurrent, und $P_x[B_y = \infty] = 1$ für alle $x, y \in S$, oder

(2). Alle $x \in S$ sind transient, und $E_x[B_y] < \infty$ für alle $x, y \in S$.

Ist S endlich, dann kann nur der erste Fall eintreten.

Beweis. Existiert ein rekurrenter Zustand, dann sind nach Satz 11.17 alle Zustände rekurrent, und $P_x[B_y = \infty] = 1$ für alle $x, y \in S$. Andernfalls sind nach Satz 11.14 alle $x \in S$ transient, und nach Korollar 11.15 gilt $G(x, x) < \infty$. Nach Lemma 11.16 folgt dann $E_x[B_y] < \infty$ für alle $x, y \in S$. Ist S endlich, dann kann der zweite Fall wegen

$$\sum_{y \in S} E_x[B_y] \quad = \quad E_x\left[\sum_{y \in S} B_y\right] \quad = \quad \infty$$

nicht eintreten. □

Was können wir aussagen, wenn die Übergangsmatrix nicht irreduzibel ist?

Allgemein ist die Relation

$$x \rightsquigarrow y \qquad \text{„}y \text{ ist von } x \text{ aus erreichbar“}$$

eine Äquivalenzrelation auf der Menge S_{rek} der rekurrenten Zustände in S. Die zugehörigen Äquivalenzklassen $S_i, i \in I$, heißen **Rekurrenzklassen**. Wir erhalten also eine disjunkte Zerlegung

$$S = S_{\text{trans}} \dot{\cup} \dot{\bigcup_{i \in I}} S_i$$

des Zustandsraums in die Menge S_{trans} der transienten Zustände, und die verschiedenen Rekurrenzklassen

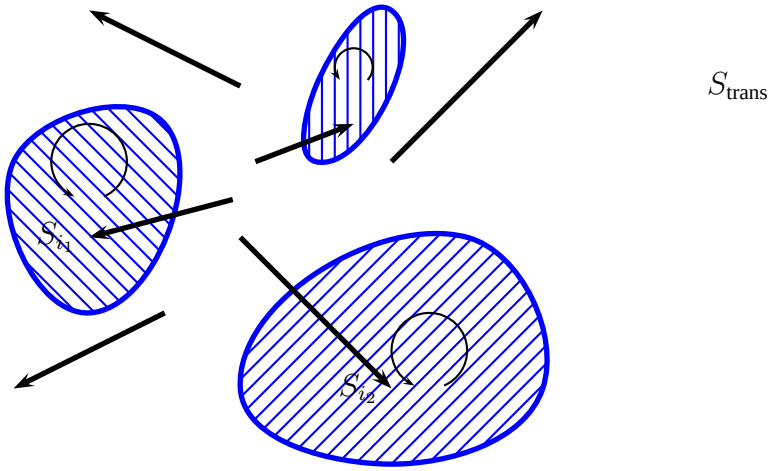


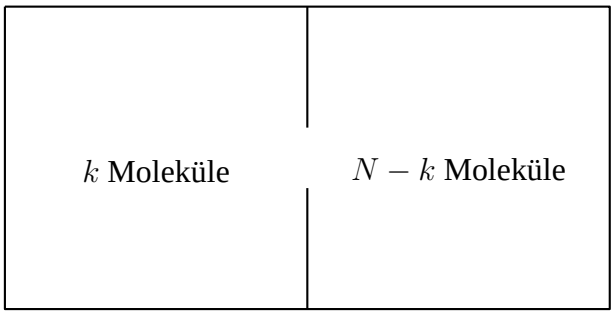
Abbildung 11.10: Zerlegung der Menge S in die transienten Zustände und die einzelnen Rekurrenzklassen

Gelangt die Markovkette in eine Rekurrenzklasse, dann bleibt sie dort mit Wahrscheinlichkeit 1 und durchläuft alle Zustände der Rekurrenzklasse unendlich oft. Startet die Markovkette in einem transienten Zustand, dann läuft sie entweder in eine Rekurrenzklasse, oder sie verbleibt im transienten Bereich, verlässt aber jede endliche Teilmenge von S_{trans} schließlich mit Wahrscheinlichkeit 1.

Beispiel. (1). *Ehrenfestmodell:* Die Markovkette aus dem Ehrenfestmodell (s. Abschnitt 2.2) ist rekurrent, da der Zustandsraum $S = \{0, 1, \ldots, N\}$ endlich, und die Übergangsmatrix

$$\begin{aligned} p(k, k-1) &= k/N \\ p(k, k+1) &= (N-k)/N \end{aligned}$$

irreduzibel ist.



Jeder Zustand wird also unendlich oft durchlaufen, was der thermodynamischen Irreversibilität zunächst zu widersprechen scheint (Einwand von Zermelo, vgl. die Bemerkung

unter Satz 7.17). Tatsächlich kann man zeigen, dass die mittlere Zeit $E_0[T_{N/2}]$ für den Übergang vom geordneten Zustand $k = 0$ in den ungeordneten Zustand $k = N/2$ von der Größenordnung $N \log N$ ist, die mittlere Zeit $E_{N/2}[T_0]$ für den umgekehrten Übergang dagegen von der Größenordnung $\frac{1}{2N} 2^{2N}$. Da N zum Beispiel gleich 10^{23} ist, ist die Rekurrenz jenseits des ungeordneten Zustandes de facto nicht beobachtbar – im makroskopischen Skalierungslimes $N \to \infty$ ergibt sich bei geeigneter Zeitreskalierung eine irreversible Dynamik.

(2). *Kartenhaus/Maschinenerneuerung:* Im Fall $\sum_{x=0}^{\infty} \varrho_x = \infty$ sind alle Zustände der Markovkette aus dem Beispiel von oben rekurrent, da 0 rekurrent und die Übergangsmatrix irreduzibel ist. Andernfalls sind alle Zustände transient.

(3). *Galton-Watson-Prozess:* Für den Galton-Watson-Verzweigungsprozess mit Nachkommensverteilung ν ist 0 ein *absorbierender* Zustand, d.h. kein anderer Zustand ist von 0 aus erreichbar. Insbesondere ist $\{0\}$ eine Rekurrenzklasse. Gilt $\nu(0) \neq 0$, dann ist umgekehrt 0 von jedem Zustand $x \in \mathbb{N}$ aus erreichbar, also sind alle $x \neq 0$ transient. Es folgt dann:

$$P_x[Z_n = 0 \text{ schließlich oder } Z_n \to \infty] \quad = \quad 1 \qquad \text{für alle } x \geq 0.$$

11.4 Stationäre stochastische Prozesse

In vielen Fällen nähert sich die Verteilung eines zeitlich verschobenen stochastischen Prozesses $(Y_n, Y_{n+1}, \ldots)$ mit Zustandsraum $(S, \mathcal{S})$ für $n \to \infty$ einer Grenzverteilung P auf dem Produktraum $\Omega = S^{\{0,1,2,\ldots\}}$ mit Produkt-σ-Algebra $\mathcal{A}$ an („asymptotische Stationarität“). Die Grenzverteilung P sollte dann selbst invariant unter Verschiebungen sein, d.h. für den Koordinatenprozess $X_n(\omega) = \omega_n$ sollte gelten:

$$(X_n, X_{n+1}, \ldots) \sim (X_0, X_1, \ldots) \qquad \text{unter } P \text{ für alle } n \geq 0. \tag{11.4.1}$$

Wir wollen stochastische Prozesse mit der Eigenschaft (11.4.1) nun genauer untersuchen.

Stationarität und Reversibilität

Definition. *(1). Eine Wahrscheinlichkeitsverteilung P auf $(\Omega, \mathcal{A})$ bzw. ein stochastischer Prozess $((X_n), P)$ heißt **stationär**, falls (11.4.1) gilt.*

*(2). Der Prozess $((X_n), P)$ heißt **reversibel**, falls die endlichdimensionalen Verteilungen invariant unter Zeitumkehr sind, d.h. falls*

$$(X_0, X_1, \ldots, X_n) \quad \sim \quad (X_n, X_{n-1}, \ldots, X_0) \qquad \textit{unter } P \textit{ für alle } n \geq 0. \qquad (11.4.2)$$

Bemerkung. Eine Wahrscheinlichkeitsverteilung P auf $(\Omega, \mathcal{A})$ ist genau dann stationär, wenn die Shiftabbildung $\theta : \Omega \to \Omega$ eine *maßerhaltende* Abbildung auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ ist, d.h. wenn $P = P \circ \theta^{-1}$ gilt.

Beispiel. (1). *IID Folgen:* Eine Folge $(X_n)_{n\geq 0}$ unabhängiger, identisch verteilter Zufallsvariablen ist ein stationärer und reversibler stochastischer Prozess.

(2). *Gaußprozesse:* Ein Gaußprozess ist ein reellwertiger stochastischer Prozess (X_n), dessen Randverteilungen Normalverteilungen sind. Beispielsweise ist ein $AR(p)$-Prozess ein Gaußprozess, wenn die Startwerte normalverteilt sind. Gaußprozesse sind eindeutig durch die Mittelwerte $E[X_n]$ und die Kovarianzen $\mathrm{Cov}[X_n, X_m]$ festgelegt. Stationarität gilt genau dann, wenn $E[X_n] = \alpha$ nicht von n abhängt, und $\mathrm{Cov}[X_n, X_m] = c_{n-m}$ nur von der Differenz $n - m$ abhängt.

(3). *Deterministische Rotationen:* Ist X_0 gleichverteilt auf dem Einheitskreis S^1, und $X_{n+1} = e^{i\phi} \cdot X_n$ mit $\phi \in [0, 2\pi)$, dann ist $(X_n)_{n\geq 0}$ stets ein stationärer Prozess. Reversibilität gilt für $\phi \neq 0$ nicht.

Satz 11.19. *Ein reversibler Prozess ist stationär.*

Beweis. Aus der Reversibilität folgt durch Zeitumkehr auf $\{0, 1, \ldots, n+1\}$ und $\{0, 1, \ldots, n\}$:

$$P \circ (X_1, X_2, \ldots, X_{n+1})^{-1} \quad = \quad P \circ (X_n, X_{n-1}, \ldots, X_0)^{-1} \quad = \quad P \circ (X_0, \ldots, X_n)^{-1}$$

für alle $n \geq 0$. Also gilt

$$P[(X_1, X_2, \ldots) \in A] \quad = \quad P[(X_0, X_1, \ldots) \in A]$$

für alle Zylindermengen $A \in \mathcal{A}$, und damit für alle $A \in \mathcal{A}$. □

Stationarität bzw. Reversibilität zeithomogener Markovketten ist durch die Startverteilung und den Übergangskern charakterisierbar:

Satz 11.20 (Stationarität und Reversibilität von Markovketten). *Für eine zeithomogene Markovkette (X_n, P_μ) im kanonischen Modell gilt:*

(1). $P_\mu \circ (X_n, X_{n+1}, \ldots)^{-1} = P_{\mu p^n}$ *für alle* $n \geq 0$.

(2). P_μ *ist genau dann stationär, wenn* μ *ein Gleichgewicht des Übergangskerns* p *ist.*

(3). P_μ *ist genau dann reversibel, wenn* μ *die Detailed-Balance-Bedingung*

$$\mu(dx)p(x, dy) \quad = \quad \mu(dy)p(y, dx) \tag{11.4.3}$$

erfüllt, d.h. wenn die Wahrscheinlichkeitsverteilung $\mu \otimes p$ *auf* $S \times S$ *invariant unter der Abbildung* $(x, y) \mapsto (y, x)$ *ist.*

Beweis. (1). Für $A \in \mathcal{A}$ und $n \geq 0$ gilt nach der Markoveigenschaft

$$\begin{aligned} P_\mu[(X_n, X_{n+1}, \ldots) \in A] &= E_\mu[I_A \circ \theta^n] \\ &= E_\mu[P_{X_n}[A]] \\ &= \int P_x[A](\mu p^n)(dx) \\ &= P_{\mu p^n}[A]. \end{aligned}$$

(2). folgt unmittelbar aus (1).

(3). Aus der Reversibilität von (X_n, P_μ) folgt, dass

$$\mu \otimes p \quad = \quad P_\mu \circ (X_0, X_1)^{-1}$$

invariant unter Koordinatentausch ist.
Umgekehrt folgt aus der Detailed-Balance-Bedingung durch Induktion

$$\begin{aligned} \mu(dx_0)p(x_0, dx_1) \cdots p(x_{n-1}, dx_n) &= \mu(dx_1)p(x_1, dx_2) \cdots p(x_{n-1}, dx_n)p(x_1, dx_0) \\ &= \ldots \quad = \quad \mu(dx_n)p(x_n, dx_{n-1}) \cdot \ldots \cdot p(x_1, dx_0) \end{aligned}$$

für alle $n \geq 0$; also

$$P_\mu \circ (X_0, \ldots, X_n)^{-1} \quad = \quad P_\mu \circ (X_n, \ldots, X_0)^{-1}.$$

□

Rekurrenz von stationären Prozessen

Stationäre stochastische Prozesse haben starke Rekurrenzeigenschaften. Die folgende Aussage zeigt unter Anderem, dass die mittlere Rückkehrzeit in eine Menge B endlichen Erwartungswert hat, wenn der Prozess mit positiver Wahrscheinlichkeit in B startet:

Satz 11.21 (Wiederkehrsatz von Kac). *Sei (X_n, P) ein stationärer stochastischer Prozess mit Zustandsraum $(S, \mathcal{S})$, und sei*

$$T_B \quad = \quad \min\{n \geq 1 : X_n \in B\}$$

die erste Eintritts- bzw. Rückkehrzeit in eine Menge $B \in \mathcal{S}$. Dann gilt

$$E[T_B\,;\, X_0 \in B] \quad = \quad P[T_B < \infty], \tag{11.4.4}$$

also mit anderen Worten

$$E[T_B \mid X_0 \in B] \quad = \quad \frac{P[T_B < \infty]}{\mu[B]} \qquad \textit{falls } \mu[B] > 0, \qquad \textit{und} \tag{11.4.5}$$

$$P[T_B < \infty] \quad = \quad 0 \qquad \textit{falls } \mu[B] = 0, \tag{11.4.6}$$

wobei $\mu = P \circ X_0^{-1}$ die Startverteilung des Prozesses ist.

Bemerkung. (1). Nach (11.4.5) ist die mittlere Rückkehrzeit in die Menge B der Kehrwert des Quotienten $\frac{P[X_0 \in B]}{P[T_B < \infty]}$, also des Anteils von $\{X_0 \in B\}$ an allen Pfaden, die B treffen.

(2). Allgemeiner gilt für jede messbare Teilmenge $A \in \mathcal{A}$ des Pfadraumes:

$$E[\tau_A\,;\, A] \quad = \quad P[\tau_A < \infty],$$

wobei $\tau_A = \min\{n \geq 1 : (X_n, X_{n+1}, \ldots) \in A\}$ die erste Zeit ist, zu der der verschobene Pfad in A liegt.

Beweis. Für $n \in \mathbb{N}$ gilt wegen der Stationarität des Prozesses:

$$\begin{aligned} E[\min(T_B, n)\,;\, X_0 \in B] &= \sum_{k=0}^{n-1} P[T_B > k \text{ und } X_0 \in B] \\ &= \sum_{k=0}^{n-1} P[X_0 \in B, X_1 \notin B, \ldots, X_k \notin B] \\ &= \sum_{k=0}^{n-1} P[X_{n-k} \in B, X_{n-k+1} \notin B, \ldots, X_n \notin B] \\ &= P[T_B \leq n]. \end{aligned}$$

Hierbei haben wir verwendet, dass $T_B \leq n$ genau dann gilt, wenn zu einer der Zeiten $n - k, k = 0, 1, \ldots, n-1$, ein letzter Besuch in B vor der Zeit n stattfindet. Die Aussage folgt für $n \to \infty$. □

Nach dem Wiederkehrsatz von Kac kehrt der Prozess (X_n) auf der Menge $\{X_0 \in B\}$ P-fast sicher nach B zurück. Durch Anwenden dieser Aussage auf die Teilfolgen $(X_{nk})_{n\geq 0}, k \in \mathbb{N}$, die alle wieder stationäre Prozesse unter P sind, erhalten wir sogar:

Korollar 11.22. *Jeder stationäre Prozess (X_n, P) ist rekurrent in folgendem Sinne: Für alle $B \in \mathcal{S}$ gilt $X_n \in B$ unendlich oft P-fast sicher auf $\{X_0 \in B\}$.*

Bemerkung (Wiederkehrsatz von Poincaré). Allgemeiner gilt für $A \in \mathcal{A}$:

$$(X_n(\omega), X_{n+1}(\omega), \ldots) \in A \qquad \text{unendlich oft für } P\text{-fast alle } \omega \in A.$$

Anwendung auf Markovketten

Wir betrachten nun eine zeithomogene Markovkette (X_n, P_x) mit abzählbarem Zustandsraum S im kanonischen Modell.

Definition. *Ein Zustand $x \in S$ heißt **positiv rekurrent**, falls die mittlere Rückkehrzeit $E_x[T_x]$ endlich ist.*

Aus dem Wiederkehrsatz von Kac folgt unmittelbar:

Korollar 11.23 (Gleichgewichte und mittlere Rückkehrzeiten). *(1). Ist μ ein Gleichgewicht der Markovkette, dann gilt*

$$\mu(x) \cdot E_x[T_x] \quad = \quad P_\mu[T_x < \infty] \qquad \textit{für alle } x \in S.$$

Insbesondere sind alle Zustände x mit $\mu(x) > 0$ positiv rekurrent.

(2). Ist zudem die Übergangsmatrix irreduzibel, dann sind sogar alle $x \in S$ positiv rekurrent mit

$$\mu(x) \quad = \quad \frac{1}{E_x[T_x]}. \tag{11.4.7}$$

Insbesondere ist das Gleichgewicht in diesem Fall eindeutig.

Beweis. (1). Da die Markovkette mit Startverteilung μ stationär ist, gilt nach dem Satz von Kac:

$$\mu(x) \cdot E_x[T_x] \quad = \quad E_\mu[T_x \,;\, X_0 = x] \quad = \quad P_\mu[T_x < \infty] \qquad \text{für alle } x \in S.$$

(2). Bei Irreduzibilität folgt globale Rekurrenz, also $P_y[T_x < \infty] = 1$ für alle $x, y \in S$, und damit $\mu(x) \cdot E_x[T_x] = 1$ für alle x.

□

Beispiel (Eindimensionale Markovkette, Birth-Death-Process). Wir betrachten eine zeithomogene Markovkette auf $S = \{0, 1, 2, \ldots\}$ mit Übergangswahrscheinlichkeiten

$$p(x, x+1) \;=\; p_x, \quad p(x, x-1) \;=\; q_x, \quad p(x, x) \;=\; r_x,$$

$p_x, q_x, r_x > 0$ mit $p_x + q_x + r_x = 1, q_0 = 0$, und $p_x, q_x > 0$ für alle $x \geq 1$.

Offensichtlich gilt Irreduzibilität. Das Gleichungssystem für eine Gleichgewichtsverteilung μ lautet

$$\begin{aligned} \mu(0) \cdot r_0 + \mu(1) \cdot q_1 &= \mu(0), \\ \mu(x-1) \cdot p_{x-1} + \mu(x) \cdot r_x + \mu(x+1) \cdot q_{x+1} &= \mu(x) \qquad \text{für } x \in \mathbb{N}. \end{aligned}$$

Da die Lösung sich rekursiv aus $\mu(0)$ berechnen lässt, ist der Lösungsvektorraum des linearen Gleichungssystems eindimensional. Aus der hinreichenden Detailed-Balance-Bedingung

$$\mu(x-1) \cdot p_{x-1} \quad = \quad \mu(x) \cdot q_x \qquad \text{für alle } x \in \mathbb{N} \tag{11.4.8}$$

erhalten wir daher in diesem Fall bereits die allgemeine Lösung

$$\mu(x) \quad = \quad \mu(0) \cdot \frac{p_0 \cdot p_1 \cdot \ldots \cdot p_{x-1}}{q_1 \cdot q_2 \cdot \ldots \cdot q_x}. \tag{11.4.9}$$

Sei

$$Z = \sum_{x=0}^{\infty} \frac{p_0 \cdot p_1 \cdot \ldots \cdot p_{x-1}}{q_1 \cdot q_2 \cdot \ldots \cdot q_x}.$$

Gilt $Z < \infty$, dann ist durch (11.4.9) mit $\mu(0) = 1/Z$ das eindeutige Gleichgewicht der Markovkette gegeben, und für die mittleren Rückkehrzeiten folgt

$$E_x[T_x] \quad = \quad 1/\mu(x) \qquad \text{für alle } x \geq 0.$$

Die Bedingung $Z < \infty$ bedeutet, dass die Wachstumswahrscheinlichkeiten $p(x-1, x)$ nicht zu groß im Vergleich zu den Abfallwahrscheinlichkeiten $p(x, x-1)$ sind. Gilt dagegen $Z = \infty$,

dann existiert keine Gleichgewichtsverteilung. Wir werden in 11.25 sehen, dass in diesem Fall auch keiner der Zustände $x \in S$ positiv rekurrent ist. Durch Lösen des Dirichletproblems kann man zudem zeigen, dass die Markovkette genau dann rekurrent ist, wenn

$$\sum_{x=0}^{\infty} \frac{q_1 q_2 \cdots q_x}{p_1 p_2 \cdots p_x} \quad = \quad \infty$$

gilt (s. Übung).

11.5 Ergodizität

In diesem Abschnitt werden wir ein Gesetz der großen Zahlen für positiv rekurrente Markovketten beweisen. Dabei verwenden wir, dass die Verläufe der Markovkette während verschiedener Exkursionen von einem Punkt aus unabhängig voneinander und identisch verteilt sind. Langzeitmittelwerte verhalten sich daher asymptotisch wie der Erwartungswert des zeitlichen Mittelwerts über eine Exkursion. Als Vorbereitung überlegen wir uns, dass der Anteil der mittleren Exkursionszeit, den die Markovkette in bestimmten Bereichen verbringt, eine Gleichgewichtsverteilung definiert.
Wie zuvor sei (X_n, P_x) eine zeithomogene Markovkette mit abzählbarem Zustandsraum S und Übergangsmatrix $p(x, y)$ im kanonischen Modell. Ferner sei

$$T_x \quad = \quad \min\{n \geq 1 \,:\, X_n = x\}$$

die erste Treffer- bzw. Rückkehrzeit zum Punkt x.

Positive Rekurrenz und Gleichgewichte

Für einen Zustand $x \in S$ sei

$$\mu_x[B] \quad := \quad E_x\left[\sum_{n=0}^{T_x - 1} I_B(X_n)\right] \quad = \quad \sum_{n=0}^{\infty} P[X_n \in B\,;\, n < T_x] \tag{11.5.1}$$

die *mittlere Anzahl der Besuche in einer Menge $B \subset S$ während einer Exkursion von x*. Ein positives Maß ν auf S heißt **invariant** bzgl. der Übergangsmatrix p, falls

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

gilt. Ein Gleichgewicht ist also eine invariante Wahrscheinlichkeitsverteilung.

Satz 11.24. *(1). Ist $x \in S$ ein rekurrenter Zustand der Markovkette, dann ist μ_x ein invariantes Maß mit Gesamtmasse $\mu_x[S] = E_x[T_x]$.*

(2). Ist x positiv rekurrent, dann ist das normierte Maß

$$\overline{\mu}_x[B] \quad = \quad \frac{\mu_x[B]}{E_x[T_x]} \qquad \left(= \frac{\textit{mittlere Aufenthaltszeit in } B}{\textit{mittlere Exkursionsdauer}} \right)$$

ein Gleichgewicht der Markovkette.

Bei positiver Rekurrenz existiert also stets ein Gleichgewicht. Umgekehrt haben wir in Korollar 11.22 bereits gezeigt, dass Gleichgewichtsverteilungen nur positiv rekurrenten Zuständen eine strikt positive Gesamtmasse zuordnen. Ist die Markovkette zudem irreduzibel, dann ist die Gleichgewichtsverteilung nach Korollar 11.22 eindeutig, d.h. die Verteilung $\overline{\mu}_x$ hängt nicht vom Startpunkt x ab.

Beweis. (1). Ist x rekurrent, dann gilt P_x-fast sicher $T_x < \infty$, und damit $X_{T_x} = x = X_0$. Für $B \subseteq S$ folgt

$$\sum_{n=0}^{T_x-1} I_B(X_n) \quad = \quad \sum_{n=0}^{T_x-1} I_B(X_{n+1}).$$

Mit der Markoveigenschaft erhalten wir damit

$$\begin{aligned}
\mu_x[B] &= E_x\left[\sum_{n=0}^{T_x-1} I_B(X_{n+1})\right] \\
&= \sum_{n=0}^{\infty} P_x[X_{n+1} \in B\,;\, n < T_x] \\
&\overset{\text{ME}}{=} \sum_{n=0}^{\infty} E_x\left[P_{X_n}[X_1 \in B\,];\, n < T_x\right] \\
&= \sum_{z\in S}\sum_{n=0}^{\infty} P_x[X_n = z\,;\, n < T_x] \cdot p(z,B) \\
&= \sum_{z\in S} \mu_x[\{z\}] \cdot p(z,B) \quad = \quad (\mu_x p)[B],
\end{aligned}$$

d.h. μ_x ist ein invariantes Maß. Die Gesamtmasse ist

$$\mu_x[S] \quad = \quad E_x\left[\sum_{n=0}^{T_x-1} I_S(X_n)\right] \quad = \quad E_x[T_x].$$

(2). Ist x positiv rekurrent, dann hat μ_x endliche Gesamtmasse, also erhält man durch Normieren ein Gleichgewicht.

□

Ein Gesetz der großen Zahlen für Markovketten

Wir können nun das Hauptresultat dieses Abschnitts formulieren. Für $n \in \mathbb{N}$ und $y \in S$ sei

$$B_y(n) \quad := \quad \sum_{i=0}^{n-1} I_{\{X_i=y\}}$$

die Anzahl der Besuche der Markovkette im Zustand y vor der Zeit n.

Satz 11.25 (Ergodensatz für Markovketten, 1. Version). *Sei (X_n, P) eine irreduzible homogene Markovkette mit abzählbarem Zustandsraum S.*

(1). Ist die Markovkette rekurrent, dann gilt

$$\lim_{n\to\infty} \frac{1}{B_y(n)} \sum_{i=0}^{n-1} f(X_i) \quad = \quad E_y\left[\sum_{i=0}^{T_y-1} f(X_i)\right] \quad = \quad \int f \, d\mu_y$$

P-fast sicher für jede Funktion $f : S \to \mathbb{R}_+$ und alle $y \in S$. Hierbei ist μ_y das durch (11.5.1) definierte invariante Maß.

(2). Existiert eine Gleichgewichtsverteilung $\overline{\mu}$, dann folgt $\overline{\mu}_y = \overline{\mu}$ für alle $y \in S$ und

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} f(X_i) \quad = \quad \int f \, d\overline{\mu} \qquad P\text{-fast sicher.}$$

Die letzte Aussage ist ein Gesetz der großen Zahlen für irreduzible, positiv rekurrente Markovketten, und eine erste Version eines Ergodensatzes für Markovketten: Die „zeitlichen“ Mittelwerte $\frac{1}{n}\sum_{i=0}^{n-1} f(X_i)$ konvergieren fast sicher gegen den „räumlichen“ Mittelwert $\int f \, d\overline{\mu}$ der Funktion f bzgl. der Gleichgewichtsverteilung. Insbesondere ergibt sich

$$\overline{\mu}(x) \quad = \quad \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} I_{\{x\}}(X_i) \qquad P\text{-fast sicher für alle } x \in S,$$

d.h. die Gewichte der Gleichgewichtsverteilung sind die asymptotischen relativen Häufigkeiten der Zustände $x \in S$. Dieser Zusammenhang kann in beide Richtungen verwendet werden:

(1). *Berechnung der asymptotischen relativen Häufigkeiten* durch Lösen des linearen Gleichungssystems $\overline{\mu} = \overline{\mu}p$.

(2). *Schätzen der Gleichgewichtsverteilung:*

$$\overline{\mu} \quad \approx \quad \frac{1}{n} \sum_{i=0}^{n-1} \delta_{X_i} \qquad \text{für große } n.$$

Beweis von Satz 11.25. Da die Zufallsvariablen X_i nicht unabhängig sind, können wir nicht wie im Beweis des klassischen GdgZ verfahren. Stattdessen nutzen wir aus, dass die Markovkette jedes mal, wenn sie den Punkt x trifft, neu startet – unabhängig vom vorherigen Verlauf. Durch Zerlegen der Summe in Teilsummen über diese verschiedenen Zykel erhalten wir eine Summe von unabhängigen Zufallsvariablen, auf die sich das klassische GdgZ anwenden lässt:

(1). Wir betrachten die Markovkette o.B.d.A. im kanonischen Modell. Sei $T^{(k)}$ die k-te Besuchszeit bzw. Rückkehrzeit zu einem festen Zustand $y \in S$, d.h. $T^{(0)} = 0$, und

$$T^{(k+1)} \quad = \quad T^{(k)} + T_y \circ \theta^{T^{(k)}} \qquad \text{für alle } k \geq 0.$$

Da die Kette irreduzibel und rekurrent ist, gilt $T^{(k)} < \infty$ P-fast sicher für alle k, und damit

$$\sum_{i=1}^{T^{(l)}} f(X_i) \quad = \quad \sum_{k=0}^{l-1} Y_k \qquad \text{mit} \quad Y_k := \sum_{i=T^{(k)}+1}^{T^{(k+1)}} f(X_i).$$

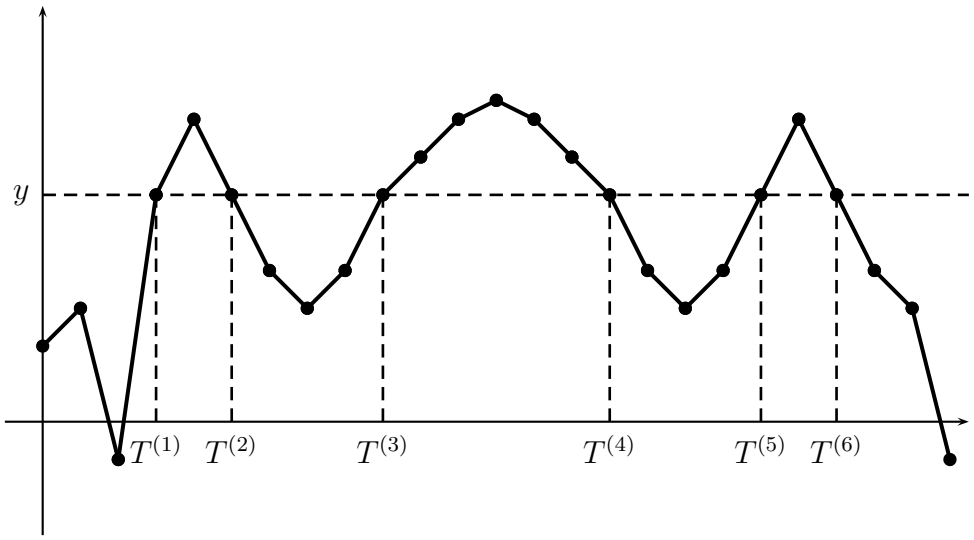


Abbildung 11.11: Regenerative Zyklen.

Wir zeigen nun, dass aufgrund der starken Markoveigenschaft die Zufallsvariablen Y_k $(k \geq 1)$ unter P unabhängig und identisch verteilt sind. Es gilt nämlich

$$Y_k \quad = \quad \sum_{i=T^{(k)}+1}^{T^{(k)}+T_y \circ \theta^{T^{(k)}}} f(X_i) \quad = \quad \sum_{j=1}^{T_y} f(X_j \circ \theta^{T^{(k)}}) \quad = \quad Y_0 \circ \theta^{T^{(k)}},$$

also

$$P[Y_k \in B \mid \mathcal{F}_{T^{(k)}}] \quad \overset{\text{SME}}{=} \quad P_y[Y_0 \in B] \qquad \text{für alle } B \subset S,$$

d.h. Y_k ist unabhängig von $\mathcal{F}_{T^{(k)}}$ mit Verteilung $P_y \circ Y_0^{-1}$. Da die Zufallsvariablen $Y_0, \ldots, Y_{k-1}$ $\mathcal{F}_{T^{(k)}}$-messbar sind, folgt die Unabhängigkeit der $Y_k, k \geq 0$, unter P. Zudem erhalten wir für $k \geq 1$:

$$E[Y_k] \quad = \quad E_y[Y_0] \quad = \quad E_y\left[\sum_{i=1}^{T_y} f(X_i)\right] \quad = \quad \int f \, d\mu_y.$$

Nach dem *Gesetz der großen Zahlen* folgt dann:

$$\lim_{l\to\infty} \frac{1}{l} \sum_{i=1}^{T^{(l)}} f(X_i) \quad = \quad \lim_{l\to\infty} \frac{1}{l} \sum_{k=1}^{l-1} Y_k \quad = \quad \int f \, d\mu_y \qquad P\text{-fast sicher.} \qquad (11.5.2)$$

Ist die Anzahl $B_y(n)$ der Besuche in y vor der Zeit n gleich l, dann gilt

$$T^{(l-1)} \quad < \quad n \quad \leq \quad T^{(l+1)},$$

also

$$\frac{1}{l} \sum_{i=1}^{T^{(l-1)}} f(X_i) \quad \leq \quad \frac{1}{B_y(n)} \sum_{i=1}^{n} f(X_i) \quad \leq \quad \frac{1}{l} \sum_{i=1}^{T^{(l+1)}} f(X_i). \qquad (11.5.3)$$

Für $n \to \infty$ konvergiert auch $l = B_y(n)$ gegen unendlich, da die Markovkette rekurrent ist. Da die linke und rechte Seite von (11.5.3) nach (11.5.2) für $l \to \infty$ gegen $\int f \, d\mu_y$ konvergieren, folgt

$$\lim_{n\to\infty} \frac{1}{B_y(n)} \sum_{i=0}^{n-1} f(X_i) \quad = \quad \lim_{n\to\infty} \frac{1}{B_y(n)} \sum_{i=1}^{n} f(X_i) \quad = \quad \int f \, d\mu_y \qquad P\text{-fast sicher.}$$

(2). Anwenden von Aussage (1) mit der konstanten Funktion $f \equiv 1$ liefert

$$\frac{n}{B_y(n)} \quad \overset{n\to\infty}{\longrightarrow} \quad \mu_y[S] \qquad P\text{-fast sicher.}$$

Da eine invariante Verteilung existiert, ist die Kette positiv rekurrent, d.h. $\mu_y[S] < \infty$. Daher folgt für $f \geq 0$:

$$\frac{1}{n} \sum_{i=0}^{n-1} f(X_i) \quad = \quad \frac{B_y(n)}{n} \cdot \frac{1}{B_y(n)} \sum_{i=0}^{n-1} f(X_i) \quad \overset{n\to\infty}{\longrightarrow} \quad \frac{\int f \, d\mu_y}{\mu_y[S]} \quad = \quad \int f \, d\overline{\mu}_y$$

P-fast sicher für $n \to \infty$. Da die Markovkette nach Voraussetzung irreduzibel ist, ist die Gleichgewichtsverteilung nach Korollar 11.22 eindeutig. Also gilt $\overline{\mu}_y = \overline{\mu}$ für alle $y \in S$.

□

Beispiel (Kartenhaus / Maschinenerneuerung). Wir betrachten die Markovkette aus dem Beispiel von oben.

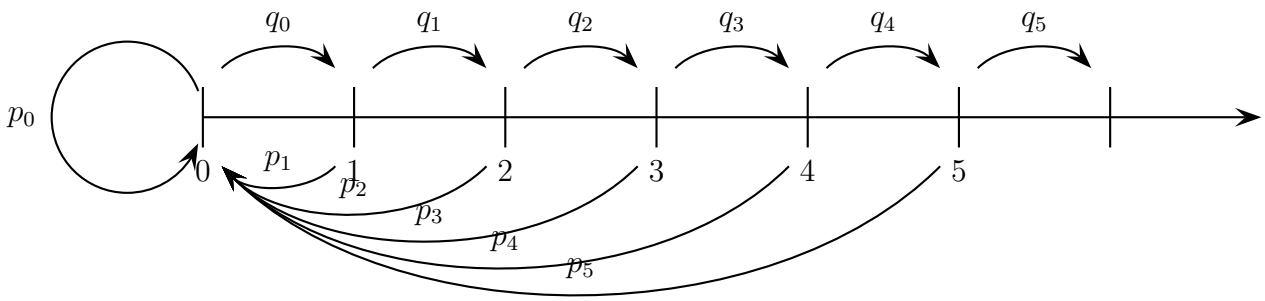


Abbildung 11.12: Übergangswahrscheinlichkeiten der durch p gegeben Markovkette.

Für Übergangswahrscheinlichkeiten $p(i,0) = p_i \in (0,1)$ und $p(i,i+1) = q_i = 1 - p_i$ erhalten wir

$$\begin{aligned} P_0[T_0 > n] &= q_0 \cdot q_1 \cdot \ldots \cdot q_{n-1}, \qquad \text{und damit} \\ E_0[T_0] &= \sum_{n=0}^{\infty} \prod_{i=0}^{n-1} q_i. \end{aligned}$$

Gilt $E_0[T_0] < \infty$, dann ist die Kette irreduzibel und positiv rekurrent. Für die asymptotische relative Häufigkeit des Zusammenfallens des Kartenhauses folgt dann nach Satz 11.25 und Korollar 11.22:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} I_{\{0\}}(X_i) = \overline{\mu}(0) = \frac{1}{E_0[T_0]} \qquad P_x\text{-fast sicher für alle } x \in S.$$

Beispiel (Markov Chain Monte Carlo Verfahren (MCMC)). Sei μ eine Wahrscheinlichkeitsverteilung auf eine abzählbaren Menge S, deren Gewichte wir bis auf eine Normierungskonstante kennen bzw. berechnen können. Um Erwartungswerte von Funktionen $f : S \to \mathbb{R}_+$ bzgl μ approximativ zu berechnen, können wir dann wie in Kapitel 3 beschrieben eine irreduzible Übergangsmatrix p mit Gleichgewicht μ bestimmen, und eine Markovkette (X_n, P) mit dieser Übergangsmatrix simulieren. Nach Satz 11.26 liefern die empirischen Mittelwerte

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} f(X_i)$$

dann eine konsistente Folge von Schätzern für den gesuchten Erwartungswert

$$\theta = \int f \, d\mu.$$

Für praktische Anwendungen ist es wichtig, den Schätzfehler zu quantifizieren. Eine erste Aussage in diese Richtung liefert ein zentraler Grenzwertsatz für Markovketten, siehe z.B. [T. Komorowski, C. Landim, S. Olla: Fluctuations in Markov Processes].

Allgemeinere Ergodensätze

Die Aussage von Satz 11.25 lässt sich wesentlich allgemeiner formulieren. Wir notieren zunächst eine elementare, aber wichtige Erweiterung:

Satz 11.26 (Ergodensatz für Markovketten, 2. Version). *Ist (X_n, P) eine irreduzible homogene Markovkette, und $\overline{\mu}$ ein Gleichgewicht des Übergangskerns p, dann gilt*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} f(X_i, X_{i+1}, \ldots, X_{i+r}) \quad = \quad \int \cdots \int f(x_0, x_1, \ldots, x_r) \overline{\mu}(dx_0) p(x_0, dx_1) \cdots p(x_{r-1}, dx_r)$$

P-fast sicher für alle $r \geq 0$ und $f : S^{r+1} \to \mathbb{R}_+$.

Wir geben nur die Beweisidee an, und überlassen die Ausführung der Details dem Leser als Übung:

Beweis-Skizze. Der Prozess

$$\widetilde{X}_i \quad := \quad (X_i, X_{i+1}, \ldots, X_{i+r})$$

ist eine Markovkette mit Zustandsraum

$$\widetilde{S} \quad = \quad \{(x_0, \ldots, x_r) \in S^{r+1} \mid p(x_i, x_{i+1}) > 0 \quad \forall 0 \leq i < r\},$$

Übergangsmatrix

$$\widetilde{p}((x_0, \ldots, x_r), (y_0, \ldots, y_r)) \quad = \quad \delta_{x_1}(y_0)\delta_{x_2}(y_1) \cdots \delta_{x_r}(y_{r-1}) p(x_r, y_r),$$

und Gleichgewichtsverteilung

$$\widetilde{\mu}(x_0, \ldots, x_r) \quad = \quad \overline{\mu}(x_0) \cdot p(x_0, x_1) \cdot \ldots \cdot p(x_{r-1}, x_r).$$

Ist (X_n) irreduzibel, so auch $(\widetilde{X}_n)$. Die Behauptung folgt daher aus Satz 11.25. □

Eine wichtige Anwendung von Satz 11.26 ist das *Schätzen der Übergangsmatrix* einer Markovkette: Für $x, y \in S$ gilt

$$p(x, y) \quad = \quad \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} I_{\{X_i = x, X_{i+1} = y\}} \qquad P\text{-fast sicher,}$$

d.h. die Übergangswahrscheinlichkeiten sind die asymptotischen relativen Häufigkeiten der Übergänge.

Beispiel (Neues im I.I.D. Fall)**.** Auch im i.i.d. Fall liefert Satz 11.26 eine neue Aussage: Ist $X_0, X_1, \ldots$ eine Folge unabhängiger, identisch verteilter Zufallsvariablen („Buchstaben“) mit Werten in einer endlichen oder abzählbaren Menge S („Alphabet“), dann ergibt sich für die asymptotische relative Häufigkeit eines Wortes $(a_0, a_1, \ldots, a_k) \in S^{k+1}$:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} I_{\{X_i=a_0, X_{i+1}=a_1, \ldots, X_{i+k}=a_k\}} \quad = \quad \prod_{j=0}^{k} \mu(a_j) \qquad P\text{-fast sicher},$$

wobei $\mu(a) = P[X_i = a]$ die Wahrscheinlichkeit des Buchstabens a ist.

Mit abstrakteren Argumenten kann man Ergodensätze im allgemeinen Rahmen dynamischer Systeme beweisen. Zum Abschluss dieses Abschnittes geben wir kurz ein entsprechendes zentrales Resultat ohne Beweis wieder. Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum und $\theta : \Omega \to \Omega$ eine *maßerhaltende Abbildung*, d.h. $P \circ \theta^{-1} = P$. Den Raum $(\Omega, \mathcal{A}, P)$ zusammen mit der maßerhaltenden Abbildung θ nennt man auch ein *dynamisches System*. Beispielsweise ist der Shiftoperator θ auf dem Pfadraum maßerhaltend bzgl. der Verteilung P eines stationären stochastischen Prozesses. Die *σ-Algebra $\mathcal{J}$ der θ-invarianten Ereignisse* ist definiert als

$$\mathcal{J} \quad = \quad \{A \in \mathcal{A} \mid \theta^{-1}(A) = A\}.$$

Beispielsweise sind die Zufallsvariablen $\liminf \frac{1}{n} \sum_{i=0}^{n-1} F \circ \theta^i$ und $\limsup \frac{1}{n} \sum_{i=0}^{n-1} F \circ \theta^i$ für jede $\mathcal{A}$-messbare Abbildung $F : \Omega \to \mathbb{R}$ messbar bzgl. $\mathcal{J}$. Allgemein sind alle θ-invarianten Ereignisse asymptotisch. Das Maß P heißt **ergodisch**, falls $P[A] \in \{0, 1\}$ für alle $A \in \mathcal{J}$ gilt. In dieser allgemeinen Situation kann man zeigen:

Satz 11.27 (Birkhoffs individueller Ergodensatz). *Für jede Funktion $F \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ gilt*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} F(\theta^i(\omega)) \quad = \quad E[F \mid \mathcal{J}](\omega) \qquad \textit{für } P\textit{-fast alle } \omega \in \Omega.$$

Ist P ergodisch, dann folgt

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} F \circ \theta^i \quad = \quad E[F] \qquad P\textit{-fast sicher.}$$

Der Beweis findet sich z.B. in den Wahrscheinlichkeitstheorie-Büchern von Breiman oder Durrett. Die $\mathcal{L}^2$-Konvergenz lässt sich mit wesentlich einfacheren funktionalanalytischen Methoden zeigen (Ergodensatz von Neumann, siehe z.B. [Varadhan: Probability Theory])

11.6 Zeitstetige Markovprozesse

Für viele Anwendungsprobleme sind Modelle, die auf Markovprozessen in kontinuierlicher Zeit basieren, natürlicher. Ändert der Prozess nur an abzählbar vielen zufälligen Zeitpunkten seinen Zustand, dann nennt man ihn eine zeitstetige Markovkette. Ein Markovprozess mit stetigen Pfaden heißt dagegen Diffusionsprozess.
Klassische Anwendungsbereiche zeitstetiger Markovketten sind die Modellierung von Warteschlangen und chemischen Reaktionen. Wir zeigen hier, wie man zeitstetige aus zeitdiskreten Markovketten konstruiert und beschreibt. Viele der Aussagen aus den letzten Abschnitten haben Entsprechungen im zeitstetigen Fall – wir verweisen dazu auf das einführende Lehrbuch [J. Norris: Markov Chains].
Der wichtigste Diffusionsprozess ist die Brownsche Bewegung, die sich ausgehend vom zentralen Grenzwertsatz als universeller zeitstetiger Skalierungslimes von Random Walks mit quadratintegrierbaren Inkrementen ergibt. In der stochastischen Analysis konstruiert man andere Diffusionsprozesse über stochastische Differentialgleichungen aus der Brownschen Bewegung – mit zahlreichen Anwendungen z.B. in der Finanzmathematik, Physik und mathematischen Biologie, aber auch mit weitreichenden Konsequenzen für viele Bereiche der Mathematik.

Übergangskerne und Markovprozesse

Seien $p_{s,t}(x, dy), 0 \leq s \leq t < \infty$, stochastische Kerne auf einem messbaren Raum $(S, \mathcal{S})$.

Definition (Markovprozess).

*(1). Ein auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ definierter zeitstetiger stochastischer Prozess $X_t : \Omega \to S, t \in [0, \infty)$, heißt **Markovprozess mit Übergangswahrscheinlichkeiten** $p_{s,t}(x, dy)$, falls*

$$P[X_{t+h} \in B \mid \mathcal{F}_t^X] \quad = \quad p_{t,t+h}(X_t, B) \qquad P\text{-fast sicher}$$

für alle $B \in \mathcal{S}$ und $t, h \geq 0$ gilt, wobei

$$\mathcal{F}_t^X \quad = \quad \sigma(X_s \mid 0 \leq s \leq t)$$

die vom Prozess erzeugten σ-Algebren sind.

*(2). Der Markovprozess heißt **zeitlich homogen**, falls die Übergangswahrscheinlichkeiten $p_{t,t+h}$ nur von h abhängen:*

$$p_{t,t+h}(x, dy) \quad = \quad p_h(x, dy) \qquad \text{für alle } t, h \geq 0.$$

Bemerkung. Einen Markovprozess mit stetigen Pfaden $t \mapsto X_t(\omega)$ nennt man einen **Diffusionsprozess**. Eine **zeitstetige Markovkette** ist ein Markovprozess, dessen Pfade stückweise konstant sind, und nur an abzählbar vielen (von ω abhängigen) Zeitpunkten springen. Allgemeine Markovprozesse können sowohl stetige als auch Sprunganteile haben – es ist auch möglich, das sich die Sprünge häufen.

Die zeithomogenen reellwertigen Markovprozesse mit räumlich homogenen Übergangswahrscheinlichkeiten sind gerade die Lévy-Prozesse:

Satz 11.28 (Lévy-Prozesse als Markovprozesse). *Ein $\mathbb{R}^d$-wertiger stochastischer Prozess (X_t, P) mit stationären unabhängigen Inkrementen $X_{t_0}, X_{t_1} - X_{t_0}, \ldots, X_{t_n} - X_{t_{n-1}} \quad (0 \le t_0 < t_1 < \ldots < t_n)$ ist ein zeithomogener Markovprozess mit translationsinvarianten Übergangswahrscheinlichkeiten*

$$p_h(x, B) \quad = \quad P[X_{t+h} - X_t \in B - x], \qquad t, h \ge 0, \quad B \in \mathcal{B}(\mathbb{R}^d).$$

Beweis. Für jede Partition $0 = t_0 < t_1 < \ldots < t_n = t$ eines Intervalls $[0, t]$ sind die Inkremente $X_{t+h} - X_t$ für $h \ge 0$ unabhängig von $\sigma(X_{t_0}, X_{t_1} - X_{t_0}, \ldots, X_{t_n} - X_{t_{n-1}})$. Wegen

$$X_{t_k} \quad = \quad X_{t_0} + \sum_{i=1}^{k} (X_{t_i} - X_{t_{i-1}})$$

erzeugen die Zufallsvariablen $X_{t_0}, X_{t_1}, \ldots, X_{t_n}$ dieselbe σ-Algebra. Also ist $X_{t+h} - X_t$ auch unabhängig von der σ-Algebra

$$\mathcal{F}_t^X \quad = \quad \sigma(X_s \mid 0 \le s \le t) \quad = \quad \sigma\left(\bigcup_{0=t_0<t_1<\ldots<t_n} \sigma(X_{t_0}, X_{t_1}, \ldots, X_{t_n}) \right).$$

Damit folgt

$$\begin{aligned} P[X_{t+h} \in B \mid \mathcal{F}_t^X](\omega) &= P[X_t + (X_{t+h} - X_t) \in B \mid \mathcal{F}_t^X](\omega) \\ &= P[X_{t+h} - X_t \in B - X_t(\omega)] \quad = \quad p_h(X_t(\omega), B) \end{aligned}$$

für P-fast alle ω. □

Beispiel. (1). *Poissonprozess:* Ein Poissonprozess mit Parameter $\lambda > 0$ ist eine zeitstetige Markovkette mit Zustandsraum $S = \{0, 1, 2, \ldots\}$ und Übergangswahrscheinlichkeiten

$$p_t(x, y) \;=\; e^{-\lambda t} \frac{(\lambda t)^{y-x}}{(y-x)!} \quad \text{für } y \ge x, \qquad \text{bzw. } p_t(x, y) \;=\; 0 \quad \text{für } y < x.$$

(2). *Brownsche Bewegung:* Eine d-dimensionale Brownsche Bewegung ist ein zeitlich homogener Diffusionsprozess mit Zustandsraum $S = \mathbb{R}^d$ und absolutstetigen Übergangswahrscheinlichkeiten $p_t(x, dy)$ mit Dichte

$$p_t(x,y) \quad = \quad (2\pi t)^{-d/2} \cdot \exp\left(-\frac{\|x-y\|^2}{2t}\right), \qquad t > 0, \quad x, y \in \mathbb{R}^d.$$

Damit die Übergangswahrscheinlichkeiten eines Markovprozesses für verschiedene Zeitintervalle konsistent sind, muss

$$p_{s,u} \quad = \quad p_{s,t}p_{t,u} \qquad \text{für alle } 0 \le s \le t \le u, \tag{11.6.1}$$

bzw., im zeithomogenen Fall,

$$p_{s+t} \quad = \quad p_s p_t \quad = \quad p_t p_s \qquad \text{für alle } s, t \ge 0 \tag{11.6.2}$$

gelten. (11.6.1) und (11.6.2) werden auch als **Chapman-Kolmogorov-Gleichungen** bezeichnet. Im zeithomogenen Fall besagt die Chapman-Kolmogorov-Gleichung (11.6.2), dass die Übergangskerne $p_t, t \ge 0$, eine **Halbgruppe** bilden.

Ist $(X_t)_{t\in[0,\infty)}$ bzgl. P ein zeitstetiger Markovprozess und (t_n) eine aufsteigende Folge in $\mathbb{R}_+$, dann ist (X_{t_n}) eine zeitdiskrete Markovkette mit Übergangskernen p_{t_{n-1},t_n}. Insbesondere erhalten wir mit Satz 11.5:

Korollar 11.29 (Endlichdimensionale Randverteilungen). *Für jedes $n \ge 0$ und $0 = t_0 < t_1 < \ldots < t_n$ hat $(X_{t_0}, X_{t_1}, \ldots, X_{t_n})$ die Verteilung*

$$\mu(dx_0)p_{t_0,t_1}(x_0, dx_1)p_{t_1,t_2}(x_1, dx_2)\cdots p_{t_{n-1},t_n}(x_{n-1}, dx_n),$$

wobei $\mu = P \circ X_0^{-1}$ die Startverteilung des Markovprozesses ist.

Beispiel (Brownsche Bewegung). Für eine d-dimensionale Brownsche Bewegung (B_t) mit Start in x_0 sind die Verteilungen von $(B_{t_1}, \ldots, B_{t_n})$ für $0 = t_0 < t_1 < \ldots < t_n$ absolutstetig mit Dichten

$$f_{B_{t_1},\ldots,B_{t_n}}(x_1,\ldots,x_n) = \prod_{i=1}^n p_{t_i-t_{i-1}}(x_{i-1},x_i) = \prod_{i=1}^n (2\pi(t_i-t_{i-1}))^{-d/2} \exp\left(-\frac{\|x_i-x_{i-1}\|^2}{2(t_i-t_{i-1})}\right).$$

Insbesondere ist eine Brownsche Bewegung ein *Gaußprozess*, d.h. alle endlichdimensionalen Randverteilungen sind multivariate Normalverteilungen.

Bemerkung (Eindeutigkeit in Verteilung, Modifikationen). Nach dem Korollar ist die Verteilung eines Markovprozesses $((X_t)_{t\geq 0}, P)$ auf dem Produktraum $S^{[0,\infty)}$ mit Produkt-σ-Algebra eindeutig durch die Startverteilung und die Übergangswahrscheinlichkeiten festgelegt. Da es überabzählbar viele Zeitpunkte $t \in \mathbb{R}_+$ gibt, ist die Situation allerdings etwas subtiler als im zeitdiskreten Fall. Beispielsweise ist das Ereignis, dass die Pfade $t \mapsto X_t(\omega)$ des Prozesses stetig bzw. rechtsstetig sind, *nicht messbar* bzgl. der Produkt-σ-Algebra. Tatsächlich kann man zu einem Markovprozess (X_t) mit (rechts-)stetigen Pfaden in der Regel einen modifizierten Prozess $(\widetilde{X}_t)$ mit

$$P[\widetilde{X}_t = X_t] = 1 \qquad \text{für jedes } t \in \mathbb{R}_+$$

finden, der keine (rechts-)stetigen Pfade hat. Der Prozess $(\widetilde{X}_t)$ hat dann dieselben endlichdimensionalen Randverteilungen wie (X_t), und ist daher ein Markovprozess mit derselben Startverteilung und denselben Übergangswahrscheinlichkeiten!

Zeitstetige Markovketten

Wir wollen nun (umgekehrt wie oben) aus einer zeitdiskreten Markovkette einen zeitstetigen Markovprozess konstruieren, der dieselben Zustände durchläuft, aber zu zufälligen kontinuierlichen Zeitpunkten von einem Zustand zum nächsten springt. Dazu betrachten wir der Übersichtlichkeit halber nur den Fall eines abzählbaren Zustandsraumes S. Einen zeitstetigen Markovprozess auf S charakterisieren wir dann durch die infinitesimalen Übergangsraten

$$\mathscr{L}_t(x,y) = \lim_{h\searrow 0} \frac{p_{t,t+h}(x,y) - \delta(x,y)}{h}, \qquad t \geq 0. \tag{11.6.3}$$

Wir beschränken uns im Folgenden auf den zeithomogenen Fall. Hier hängen die Übergangswahrscheinlichkeiten nicht von t ab, und es gilt

$$\mathscr{L}_t(x,y) = \mathscr{L}(x,y) = \lim_{h\searrow 0} \frac{p_h(x,y) - \delta(x,y)}{h} \qquad \text{für alle } t \geq 0. \tag{11.6.4}$$

Wegen

$$\begin{aligned} p_h(x,y) &= h \cdot \mathscr{L}(x,y) + o(h) \qquad \text{für } x \neq y, \text{ und} \\ p_h(x,x) &= 1 + h \cdot \mathscr{L}(x,x) + o(h), \end{aligned}$$

ist $\mathscr{L}(x,y)$ für $x \neq y$ die Sprungrate von x nach y, und $\mathscr{L}(x,x)$ ist die negative Wegsprungrate von x. Erfüllen die Übergangswahrscheinlichkeiten eines zeithomogenen Markovprozesses auf S die Bedingung (11.6.4) bzgl. eines zu spezifizierenden Konvergenzbegriffes, dann heißt die

Matrix $\mathscr{L}(x,y)$ $(x,y \in S)$ **infinitesimaler Generator** des Markovprozesses. Da $p_h(x,\bullet)$ für alle $h \geq 0$ eine Wahrscheinlichkeitsverteilung ist, sollte in diesem Fall gelten:

$$\mathscr{L}(x,x) \quad = \quad -\sum_{y \in S} \mathscr{L}(x,y) \qquad \text{für alle } x \in S. \tag{11.6.5}$$

Sei nun $\mathscr{L}(x,y)$ $(x,y \in S)$ eine gegebene Matrix mit $\mathscr{L}(x,y) \geq 0$ für alle $x,y \in S$ und (11.6.5). Wir setzen zudem voraus, dass die Wegsprungraten $\mathscr{L}(x,x)$ beschränkt sind:

Annahme: Es existiert $\lambda > 0$ mit

$$\sum_{y \in S} \mathscr{L}(x,y) \quad = \quad -\mathscr{L}(x,x) \quad \leq \quad \lambda \qquad \text{für alle } x \in S. \tag{11.6.6}$$

Um einen Markovprozess mit Sprungraten $\mathscr{L}(x,y)$ zu konstruieren, betrachten wir unabhängige, $\mathrm{Exp}(\lambda)$-verteilte Zufallsvariablen $T_1, T_2, \ldots$ auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$, die die zeitlichen Abstände zwischen möglichen Sprüngen des Prozesses beschreiben. Sei ferner $(Y_n)_{n=0,1,2,\ldots}$ eine von $\sigma(T_1, T_2, \ldots)$ unabhängige Markovkette auf $(\Omega, \mathcal{A}, P)$ mit Übergangswahrscheinlichkeiten

$$\begin{aligned} \pi(x,y) &= \frac{1}{\lambda}\mathscr{L}(x,y) \qquad \text{für } y \neq x, \\ \pi(x,x) &= 1 - \sum_{y \neq x} \pi(x,y). \end{aligned}$$

Die Kette (Y_n) beschreibt die Zustände, die der zu konstruierende zeitstetige Sprungprozess durchläuft. Mit

$$N_t \quad = \quad \#\{n \in \mathbb{N} | T_1 + T_2 + \ldots + T_n \leq t\}$$

erhalten wir:

Satz 11.30 (Konstruktion von zeitstetigen Markovketten). *Der Prozess $X_t := Y_{N_t}$ ist ein zeitstetiger Markovprozess mit Zustandsraum S, Übergangswahrscheinlichkeiten*

$$p_t(x,y) \quad = \quad e^{-\lambda t} \sum_{n=0}^{\infty} \frac{(\lambda t)^n}{n!} \pi^n(x,y), \qquad x,y \in S, \tag{11.6.7}$$

und Generator $\mathscr{L}(x,y)$. Genauer gilt

$$\lim_{h \searrow 0} \sup_{x \in S} \sum_{y \in S} \left| \frac{p_h(x,y) - \delta(x,y)}{h} - \mathscr{L}(x,y) \right| \quad = \quad 0. \tag{11.6.8}$$

Bemerkung. (1). *Matrixexponentialfunktion:* Die Übergangswahrscheinlichkeiten haben die Form

$$p_t \quad = \quad e^{-\lambda t} e^{\lambda t \pi} \quad = \quad e^{\lambda t(\pi - I)}, \tag{11.6.9}$$

wobei

$$(e^q)(x,y) \quad = \quad \sum_{n=0}^{\infty} \frac{1}{n!} q^n(x,y)$$

die Matrixexponentialfunktion ist. Hierbei ist $e^{\lambda t \pi}$ auch im abzählbar unendlichen Fall definiert, da die Matrizen $(\lambda t \pi)^n$ für alle $n \geq 0$ nichtnegativ sind. Die Reihe $e^{\lambda t(\pi - I)}$ konvergiert bzgl. der multiplikativen Matrixnorm

$$\|q\| \quad := \quad \sup_{x \in S} \sum_{y \in S} |q(x,y)|, \tag{11.6.10}$$

da $\|\lambda^n t^n (\pi - I)^n\| \leq (2\lambda t)^n$ für alle $n \geq 0$ gilt, und die Identität (11.6.9) folgt wegen $e^{\lambda t(\pi - I)} = e^{-\lambda t I} e^{\lambda t \pi} = e^{-\lambda t} e^{\lambda t \pi}$.

(2). *Konvergenzbegriff:* Die Aussage (11.6.8) besagt, dass

$$\lim_{h \searrow 0} \frac{p_h - I}{h} \quad = \quad \mathscr{L}$$

bzgl. der durch (11.6.10) definierten Matrixnorm gilt. Die Voraussetzung (11.6.6) gewährleistet gerade, dass die Norm von $\mathscr{L}$ endlich ist. In anderer Form ausgedrückt bedeutet (11.6.8), dass die signierten Maße $\frac{1}{h}(p_h(x,\bullet) - \delta(x,\bullet))$ für $h \searrow 0$ gleichmäßig in Variationsnorm (ℓ^1-Norm) gegen $\mathscr{L}(x,\bullet)$ konvergieren. Eine entsprechende Aussage gilt auch (mit analogem Beweis), wenn der Zustandsraum S nicht abzählbar ist.

Beweis. Seien $t, h \geq 0$ und $y \in S$. Um die Markoveigenschaft

$$P[X_{t+h} \,=\, y \,|\, \mathcal{F}_t^X] \quad = \quad p_h(X_t, Y) \qquad P\text{-fast sicher} \tag{11.6.11}$$

zu zeigen, verfahren wir ähnlich wie für Poisson-Prozesse in Satz 10.12. Seien zunächst $k, l \in \{0,1,2,\ldots\}$ fest, und sei

$$\mathcal{G}_k \quad = \quad \sigma(T_1, \ldots, T_k, Y_0, Y_1, \ldots, Y_k)$$

die σ-Algebra, die den Verlauf des Prozesses bis zum k-ten Sprung beschreibt. Da die Wartezeiten T_i $(i \in \mathbb{N})$ und die Markovkette (Y_n) unabhängig voneinander sind, und N_t messbar bzgl. $\sigma(T_i \mid i \in \mathbb{N})$ ist, erhalten wir nach (10.4.3):

$$\begin{aligned} & P[N_t = k, N_{t+h} = k+l, Y_{k+l} = y \mid \mathcal{G}_k] \\ = \quad & P[N_t = k, N_{t+h} = k+l \mid T_1, \ldots, T_k] \cdot P[Y_{k+l} = y \mid Y_0, Y_1, \ldots, Y_k] \\ \overset{(10.4.3)}{=} \quad & P[N_t = k \mid T_1, \ldots, T_k] \cdot P[N_h = l] \cdot \pi^l(Y_k, y) \\ = \quad & P[N_t = k \mid \mathcal{G}_k] \cdot e^{-\lambda h} \frac{(\lambda h)^l}{l!} \pi^l(Y_k, y) \qquad P\text{-fast sicher.} \end{aligned}$$

Durch Summieren über l folgt:

$$P[N_t = k, X_{t+h} = y \mid \mathcal{G}_k] \;=\; P[N_t = k \mid \mathcal{G}_k] \cdot e^{-\lambda h} \sum_{l=0}^{\infty} \frac{(\lambda h)^l}{l!} \pi^l(Y_k, y) \quad P\text{-f.s.} \qquad (11.6.12)$$

Sei nun $A \in \mathcal{F}_t^X$. Ähnlich wie im Beweis von Satz 10.12 (3) zeigt man, dass dann ein Ereignis $A_k \in \mathcal{G}_k$ existiert mit

$$A \cap \{N_t = k\} \quad = \quad A_k \cap \{N_t = k\},$$

d.h. für $N_t = k$ hängt der Verlauf von X_s für $0 \le s \le t$ nur von den Zufallsvariablen $T_1, \ldots, T_k$ und $Y_0, \ldots, Y_k$ ab. Nach (11.6.12) folgt dann

$$\begin{aligned} & P[\{N_t = k\} \cap \{X_{t+h} = y\} \cap A] \\ = \quad & E[P[N_t = k, X_{t+h} = y \mid \mathcal{G}_k] \,;\, A_k] \\ = \quad & E[P[N_t = k \mid \mathcal{G}_k] \cdot p_h(Y_k, y) \,;\, A_k] \\ = \quad & E[p_h(X_t, y) \,;\, A \cap \{N_t = k\}], \end{aligned}$$

wobei p_h wie in (11.6.7) definiert ist. Hierbei haben wir im letzten Schritt benutzt, dass $X_t = Y_k$ auf $\{N_t = k\}$ gilt. Durch Summieren über k erhalten wir schließlich

$$P[\{X_{t+h} = y\} \cap A] \quad = \quad E[p_h(X_t, y) \,;\, A] \qquad \text{für alle } A \in \mathcal{F}_t^X,$$

und damit (11.6.11).

Um den Generator zu identifizieren, bemerken wir, dass für $y \neq x$ aus (11.6.7) wegen $\mathscr{L}(x, y) = \lambda \pi(x, y)$ folgt:

$$p_h(x, y) - h\mathscr{L}(x, y) \quad = \quad (e^{-\lambda h} - 1)\lambda h \pi(x, y) + e^{\lambda h} \sum_{n=2}^{\infty} \frac{(\lambda h)^n}{n!} \pi^n(x, y).$$

Wegen $\sum_{y \in S} \pi^n(x,y) = 1$ für alle $n \geq 0$ erhalten wir dann die Abschätzung

$$\sup_{x \in S} \sum_{y \neq x} |p_h(x,y) - h\mathscr{L}(x,y)| \quad = \quad O(h^2).$$

Die Aussage (11.6.8) folgt hieraus, da

$$p_h(x,x) - \delta(x,x) - h\mathscr{L}(x,x) \quad = \quad -\sum_{y \neq x} (p_h(x,y) - h\mathscr{L}(x,y))$$

für alle $x \in S$ gilt. □

Vorwärts- und Rückwärtsgleichungen für Markovketten

Wir leiten nun Gleichungen für die Zeitentwicklung der Übergangswahrscheinlichkeiten von Markovketten her.

Zeitdiskreter Fall. Für die n-Schritt Übergangswahrscheinlichkeiten einer zeitdiskreten Markovkette mit Übergangskern π gilt

$$\pi^{n+1} - \pi^n \quad = \quad (\pi - I)\pi^n \quad = \quad \pi^n(\pi - I) \qquad \text{für alle } n \geq 0. \tag{11.6.13}$$

Hierbei ist $\pi - I$ der Generator der Markovkette.

Zeitstetiger Fall. Im zeitstetigen Fall erhalten wir als infinitesimale Versionen von (11.6.13) Differentialgleichungen für die Zeitentwicklung der Übergangswahrscheinlichkeiten. Aus (11.6.8) und der Chapman-Kolmogorov-Gleichung (11.6.1) folgt:

Satz 11.31 (Kolmogorovsche Vorwärts- und Rückwärtsgleichung). *Für die Übergangsmatrizen $p_t(x,y)$ des in Satz 11.30 konstruierten Markovprozesses gilt*

$$\lim_{h \to 0} \frac{p_{t+h} - p_t}{h} \quad = \quad p_t\mathscr{L} \quad = \quad \mathscr{L}p_t \qquad \textit{für alle } t \geq 0$$

mit Konvergenz bzgl. der in 11.6.10 definierten Matrixnorm $\| \bullet \|$. Insbesondere erfüllen die Übergangswahrscheinlichkeiten die ***Kolmogorovsche Vorwärtsgleichung (Mastergleichung)***

$$\frac{d}{dt} p_t(x,y) \quad = \quad \sum_{z \in S} p_t(x,z)\mathscr{L}(z,y), \qquad t \geq 0, \tag{11.6.14}$$

sowie die ***Kolmogorovsche Rückwärtsgleichung***

$$\frac{d}{dt} p_t(x,y) \quad = \quad \sum_{z \in S} \mathscr{L}(x,z)p_t(z,y), \qquad t \geq 0. \tag{11.6.15}$$

Beweis. Nach (11.6.8) gilt $\lim_{h \searrow 0} h^{-1}(p_h - I) = \mathscr{L}$ bzgl. der Matrixnorm $\| \bullet \|$. Da die Norm multiplikativ mit $\|p_t\| \leq 1$ ist, folgt für $t, h > 0$ nach der Chapman-Kolmogorov-Gleichung

$$\begin{aligned} \left\| \frac{p_{t+h} - p_t}{h} - p_t \mathscr{L} \right\| &= \left\| p_t \left(\frac{p_h - I}{h} - \mathscr{L} \right) \right\| \leq \|p_t\| \cdot \left\| \frac{p_h - I}{h} - \mathscr{L} \right\| \\ &\leq \left\| \frac{p_h - I}{h} - \mathscr{L} \right\| \overset{h \searrow 0}{\longrightarrow} 0. \end{aligned}$$

Entsprechend konvergiert auch

$$\begin{aligned} \left\| \frac{p_{t-h} - p_t}{-h} - p_t \mathscr{L} \right\| &= \left\| p_{t-h} \left(\frac{p_h - I}{h} - p_h \mathscr{L} \right) \right\| \leq \|p_{t-h}\| \cdot \left\| \frac{p_h - I}{h} - p_h \mathscr{L} \right\| \\ &\leq \left\| \frac{p_h - I}{h} - \mathscr{L} \right\| + \|I - p_h\| \cdot \|\mathscr{L}\| \end{aligned}$$

für $h \searrow 0$ gegen 0. Damit haben wir die Vorwärtsgleichung

$$\lim_{h \to 0} h^{-1}(p_{t+h} - p_t) = p_t \mathscr{L}$$

für $t > 0$ gezeigt. Der Beweis der Rückwärtsgleichung verläuft ähnlich. □

Anschaulich können wir die *Vorwärtsgleichung* folgendermaßen interpretieren: Sei $x \in S$ ein fester Zustand. Dann beschreibt die Funktion

$$u(t, y) = p_t(x, y) = P[X_t = y \mid X_0 = x], \qquad t \geq 0, \quad y \in S,$$

die Zeitentwicklung der Aufenthaltswahrscheinlichkeiten der Markovkette in Zuständen $y \in S$. Die Vorwärtsgleichung besagt, dass u das Anfangswertproblem

$$\begin{aligned} \frac{\partial u}{\partial t}(t, y) &= \sum_{z \in S} u(t, z) \mathscr{L}(z, y) \qquad \text{für } t \geq 0, \\ u(0, y) &= \delta_x(y) \end{aligned}$$

löst. Die Wahrscheinlichkeitsmasse im Punkt y ändert sich also dadurch, dass Übergänge von anderen Zuständen z nach y mit den Raten $\mathscr{L}(z, y)$, bzw. Übergänge von y in andere Zustände mit der negativen Rate $\mathscr{L}(y, y)$ stattfinden. Bei der Analyse chemischer Reaktionen spielt die Vorwärtsgleichung eine wichtige Rolle – sie wird in den Naturwissenschaften meist als Mastergleichung bezeichnet.

Für die *Rückwärtsgleichung* ergibt sich eine ähnliche, aber andere Interpretation: Seien hier $y \in S$ und $t \geq 0$ fest, und

$$v(s, x) = p_{t-s}(x, y) = P[X_t = y \mid X_s = x], \qquad s \in [0, t], \quad x \in S.$$

Die Funktion v beschreibt die Abhängigkeit der Aufenthaltswahrscheinlichkeiten von dem zurückliegenden Startzeitpunkt und Anfangszustand des Markovprozesses. Die Rückwärtsgleichung besagt dann, dass v das „Endwertproblem“

$$\begin{aligned} \frac{\partial v}{\partial s}(s,x) &= \sum_{z\in S} \mathscr{L}(x,z)v(s,z), \qquad s \in [0,t], \\ v(t,x) &= \delta_y(x) \end{aligned}$$

löst.

Allgemeiner ergeben sich aus der Vorwärtsgleichung Zeitentwicklungsgleichungen für die Verteilungen μp_t des Markovprozesses mit beliebiger Startverteilung μ, und aus der Rückwärtsgleichung Zeitentwicklungsgleichungen für die Erwartungswerte $E[f(X_t) \,|\, X_s = x]$ von Funktionen $f : S \to \mathbb{R}$. Die Rückwärtsgleichung liefert auch eine infinitesimale Charakterisierung von Gleichgewichtsverteilungen zeitstetiger Markovketten:

Korollar 11.32 (Gleichgewichte zeitstetiger Markovketten). *Ist die Annahme (11.6.6) erfüllt, dann sind für eine Wahrscheinlichkeitsverteilung μ auf S die folgenden Aussagen äquivalent:*

(1). μ ist ein Gleichgewicht der Übergangshalbgruppe $(p_t)_{t\geq 0}$ aus (11.6.7), d.h.

$$\mu p_t = \mu \qquad \text{für alle } t \geq 0.$$

(2). $\mu\mathscr{L} = 0$, d.h.

$$\sum_{x\in S} \mu(x)\mathscr{L}(x,y) = 0 \qquad \text{für alle } y \in S.$$

Hierbei gewährleistet die Annahme (11.6.6) unter anderem, dass $\mu\mathscr{L}$ auch im abzählbar unendlichen Fall definiert ist.

Beweis. Anschaulich folgt aus der Rückwärtsgleichung

$$\frac{d}{dt}\mu p_t = \mu\mathscr{L}p_t \qquad \text{für } t \geq 0, \qquad \mu p_0 = \mu, \tag{11.6.16}$$

und damit die Aussage. Um dies auch im abzählbar unendlichen Fall zu rechtfertigen, verwenden wir die Variationsnorm (ℓ^1-Norm) $\|\nu\|_{TV} = \sum\limits_{x\in S} |\nu(x)|$ von signierten Maßen. Für eine Matrix $q(x,y)$ $(x,y \in S)$ und eine Wahrscheinlichkeitsverteilung μ gilt:

$$\|\mu q\|_{TV} \leq \|\mu\|_{TV} \cdot \|q\| = \|q\|.$$

Nach Satz 11.30 erhalten wir

$$\lim_{h\searrow 0} \left\| \frac{\mu p_{t+h} - \mu p_t}{h} - \mu\mathscr{L}p_t \right\|_{TV} \leq \|\mu\|_{TV} \cdot \lim_{h\searrow 0} \left\| \frac{p_{t+h} - p_t}{h} - \mathscr{L}p_t \right\| = 0,$$

und somit (11.6.16), wobei die Ableitung als Grenzwert der Differenzenquotienten in Variationsnorm definiert ist. □

Aufbauend auf den obigen Resultaten kann man nun, ähnlich wie im zeitdiskreten Fall, die Rekurrenz und Transienz von zeitstetigen Markovketten untersuchen, mittlere Rückkehrzeiten und Trefferwahrscheinlichkeiten berechnen, und einen Ergodensatz beweisen. Unter der Annahme (11.6.6) können Rekurrenz und Transienz vollständig auf den zeitdiskreten Fall zurückgeführt werden, da der zeitstetige Prozess $X_t = Y_{N_t}$ dieselben Zustände durchläuft wie die zeitdiskrete Markovkette $(Y_n)_{n=0,1,2,\ldots}$. Für die Herleitung von Differenzengleichungen für mittlere Rückkehrzeiten, Trefferwahrscheinlichkeiten usw., sowie den Beweis des Gesetzes der großen Zahlen im zeitstetigen Fall verweisen wir aus das Buch 'Markov Chains' von J. R. Norris. Wir sehen uns hier noch ein Beispiel an, das einen wichtigen Anwendungsbereich zeitstetiger Markovketten kurz anreißt:

Beispiel (M/M/1-Warteschlangenmodell). Im einfachsten Modell einer Warteschlange gibt es nur einen Server. Die Aufträge kommen jeweils nach unabhängigen, mit einem Parameter $\lambda > 0$ exponentialverteilten Wartezeiten beim Server an, und die Abstände zwischen den Bearbeitungszeiten zweier Aufträge sind ebenfalls unabhängig, und mit einem Parameter ν exponentialverteilt. Die beiden „M"s in M/M/1 stehen für gedächtnislose (engl. memoryless) Ankunfts- und Bearbeitungszeiten, die „1" für die Anzahl der Server.
Unter diesen (sehr restriktiven) Annahmen wird die Warteschlange durch eine zeitstetige Markovkette mit Zustandsraum $S = \{0, 1, 2, \ldots\}$ und Übergangsraten

$$\mathscr{L}(x, x+1) \quad = \quad \lambda, \qquad \mathscr{L}(x, x-1) \quad = \quad \nu,$$

beschrieben, d.h. durch einen zeitstetigen Birth-Death-Process.

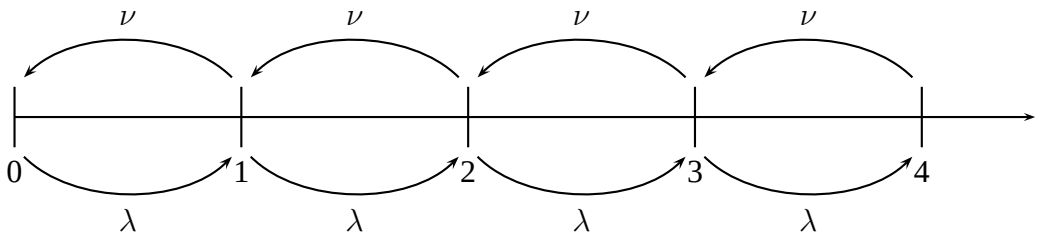


Abbildung 11.13: Übergangsraten einer M/M/1-Warteschlange.

Die Wegsprungraten $-\mathscr{L}(x,x)$ sind durch $\lambda+\nu$ beschränkt, und die Sprungkette (Y_n) hat die Übergangswahrscheinlichkeiten

$$\pi(x,x+1) \quad = \quad \frac{\lambda}{\lambda+\nu}, \qquad \pi(x,x-1) \quad = \quad \frac{\nu}{\lambda+\nu} \quad \text{für } x>0,$$

und $\pi(0,0)=\dfrac{\nu}{\lambda+\nu}$. Insbesondere sind die Sprungkette, und damit auch die zeitstetige Markovkette, genau dann rekurrent, wenn $\lambda \le \nu$ gilt. Die Gleichgewichtsbedingung $\mu\mathscr{L}=0$ für den zeitstetigen Prozess lautet

$$\begin{aligned} -\mu(0)\cdot\lambda+\mu(1)\cdot\nu &= 0, \\ \mu(x-1)\cdot\lambda-\mu(x)(\lambda+\nu)+\mu(x+1)\cdot\nu &= 0 \qquad \text{für } x\in\mathbb{N}. \end{aligned}$$

Für $\lambda \ge \nu$ existiert keine Gleichgewichsverteilung, für $\lambda<\nu$ ist die geometrische Verteilung

$$\mu(x) \quad = \quad \left(1-\frac{\lambda}{\nu}\right)\cdot\left(\frac{\lambda}{\nu}\right)^x, \qquad x=0,1,2,\ldots,$$

das eindeutige Gleichgewicht. Aus dem Ergodensatz folgt dann beispielsweise, dass die mittlere Länge $\frac{1}{t}\int_0^t X_s\,ds$ der Warteschlange sich asymptotisch wie der Erwartungswert $\dfrac{\lambda}{\nu-\lambda}$ der Gleichgewichtsverteilung verhält.

Vorwärts- und Rückwärtsgleichung für die Brownsche Bewegung

Für allgemeine Markovprozesse ist die Herleitung von Vorwärts- und Rückwärtsgleichungen technisch häufig deutlich aufwändiger, da der infinitesimale Generator $\mathscr{L}$ im Allgemeinen ein unbeschränkter linearer Operator ist. Dies ist bereits bei zeitstetigen Markovketten der Fall, wenn die Wegsprungraten nicht beschränkt sind. Für die Brownsche Bewegung erhalten wir die Kolmogorovschen Gleichungen unmittelbar aus der expliziten Form der Übergangsdichten

$$p_t(x,y) \quad = \quad (2\pi t)^{-d/2}\cdot\exp\left(-\frac{\|x-y\|^2}{2t}\right).$$

Als infinitesimaler Generator ergibt sich der Laplaceoperator:

Satz 11.33 (Brownsche Bewegung und Wärmeleitungsgleichung). *Die Übergangsdichten $p_t(x,y)$ der Brownschen Bewegung bilden die Fundamentallösung der Wärmeleitungsgleichung, d.h. es gilt*

$$\frac{\partial}{\partial t}p_t(x,y) \quad = \quad \frac{1}{2}\Delta_x p_t(x,y) \quad = \quad \frac{1}{2}\Delta_y p_t(x,y) \tag{11.6.17}$$

mit Anfangsbedingung

$$\lim_{t\searrow 0}\int p_t(x,y)f(y)\,dy \quad = \quad f(x) \qquad \textit{für alle } f\in C_b(\mathbb{R}^d) \textit{ und } x\in\mathbb{R}^d, \tag{11.6.18}$$

bzw.

$$\lim_{t \searrow 0} \int g(x) p_t(x,y)\, dy \quad = \quad g(y) \qquad \textit{für alle } g \in C_b(\mathbb{R}^d) \textit{ und } y \in \mathbb{R}^d. \tag{11.6.19}$$

Hierbei ist $\Delta_x = \sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2}$ *der Laplace-Operator in der* x*-Variable.*

Beweis. Die Gleichung (11.6.17) verifiziert man durch Nachrechnen. Für $x \in \mathbb{R}^d$ ist $p_t(x,y)dy$ eine Normalverteilung mit Mittelwertvektor x und Kovarianzmatrix $t \cdot I_d$. Hieraus folgt (11.6.18), da diese Wahrscheinlichkeitsverteilung für $t \searrow 0$ analog zu Beispiel 2 in Abschnitt 8.3 schwach gegen das Diracmaß δ_x konvergiert. Die Identität (11.6.19) folgt aus (11.6.18) wegen $p_t(x,y) = p_t(y,x)$. □

Die Gleichung

$$\frac{\partial}{\partial t} p_t(x,y) \quad = \quad \frac{1}{2} \Delta_y p_t(x,y)$$

ist die *Vorwärtsgleichung*, und die Gleichung

$$\frac{\partial}{\partial t} p_t(x,y) \quad = \quad \frac{1}{2} \Delta_x p_t(x,y)$$

die *Rückwärtsgleichung* der Brownschen Bewegung. Anschaulich können wir die Vorwärtsgleichung auch folgendermaßen interpretieren: Für jedes Gebiet $D \subset \mathbb{R}^d$ mit glattem Rand gilt:

$$\begin{aligned} \frac{\partial}{\partial t} p_t(x,D) &= \int\limits_D \frac{\partial}{\partial t} p_t(x,y) dy \\ &= \frac{1}{2} \int\limits_D \Delta_y p_t(x,y) dy \\ &= \frac{1}{2} \int\limits_{\partial D} n(y) \cdot \nabla_y p_t(x,y) \nu(dy), \end{aligned}$$

wobei n der äußere Normalenvektor und ν das Oberflächenmaß auf ∂D ist.

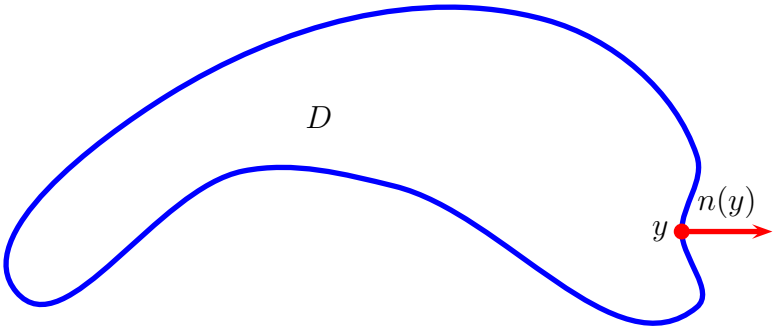


Abbildung 11.14: Äußerer Normalenvektor der Menge D im Punkt y.

Also beschreibt $\frac{1}{2} n \cdot \nabla_y p_t(x,y)$ den Nettozufluss von Wahrscheinlichkeitsmasse pro Flächeneinheit durch ein infinitesimales Flächenstück mit Ausrichtung n am Punkt y.
Für Funktionen $f \in C_b^2(\mathbb{R}^d)$ ergeben sich aus (11.6.17) die Zeitentwicklungsgleichungen

$$\frac{\partial}{\partial t} p_t f \quad = \quad \frac{1}{2} \Delta p_t f \quad = \quad \frac{1}{2} p_t \Delta f$$

für die Erwartungswerte

$$(p_t f)(x) \quad = \quad \int_{\mathbb{R}^d} p_t(x,y) f(y)\, dy \quad = \quad E_x[f(B_t)].$$

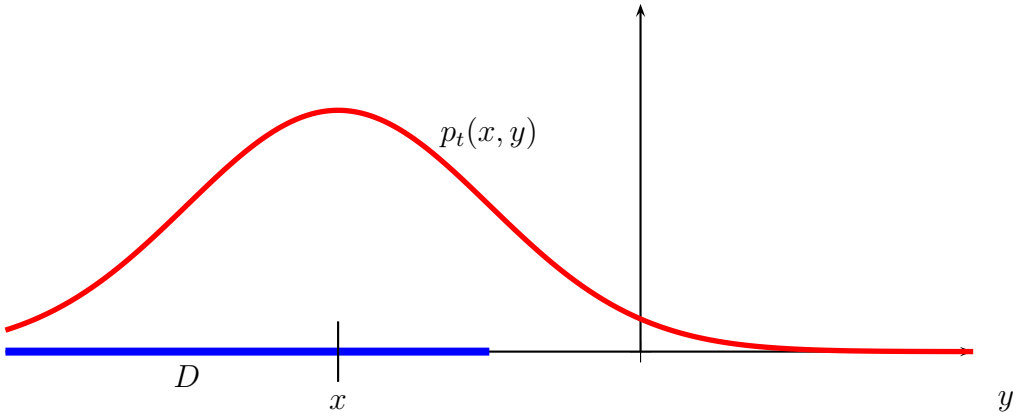


Abbildung 11.15: Nettoabfluss $\left| \frac{1}{2} \frac{\partial}{\partial y} p_t(x,y) \right|$.

Kapitel 12

Importance Sampling und große Abweichungen

Um Wahrscheinlichkeiten seltener Ereignisse zu untersuchen, geht man häufig zu einer neuen absolutstetigen Wahrscheinlichkeitsverteilung über, bzgl. der das relevante Ereignis nicht mehr selten ist. Der Maßwechsel geschieht dabei typischerweise mit einer exponentiellen Dichte. Auf diese Weise erhält man unter Anderem asymptotische Aussagen über die Wahrscheinlichkeiten großer Abweichungen. Eine zentrale Rolle spielt dabei der Begriff der relativen Entropie, die die statistische Unterscheidbarkeit zweier Wahrscheinlichkeitsverteilungen misst. Anwendungen liegen in der Asymptotik von Likelihood basierten Schätz und Testverfahren, und der asymptotischen Effizienz von Importance Sampling Schätzern.

12.1 Relative Dichten und Importance Sampling

Oft ist es günstig, Wahrscheinlichkeitsverteilungen mit einer relativen Dichte bzgl. leichter handhabbarer Verteilungen darzustellen. Die relative Dichte ist dabei häufig nur bis auf eine multiplikative Konstante explizit bekannt. Wir stellen hier zunächst einige Grundlagen über relative Dichten zusammen, und betrachten dann Monte-Carlo Verfahren in diesem Kontext.

Relative Dichten

Seien μ und ν Wahrscheinlichkeitsverteilungen auf einem messbaren Raum $(S, \mathcal{S})$. Das Maß μ heißt **absolutstetig** bzgl. ν ($\mu \ll \nu$), falls jede ν-Nullmenge auch eine μ-Nullmenge ist. Der

Satz von Radon-Nikodym besagt, dass μ genau dann absolutstetig bzgl. ν ist, wenn eine relative Dichte $d\mu/d\nu \in \mathcal{L}^1(S, \mathcal{S}, \nu)$ existiert mit $\mu[B] = \int\limits_B \frac{d\mu}{d\nu}(x)\nu(dx)$ für alle $B \in \mathcal{S}$, bzw.

$$\int f \, d\mu \quad = \quad \int f \cdot \frac{d\mu}{d\nu} \, d\nu \qquad \text{für alle messbaren } f : S \to \mathbb{R}^+. \tag{12.1.1}$$

Die relative Dichte ist ν-fast sicher eindeutig festgelegt. Ein stochastischer Beweis des Satzes von Radon-Nikodym basierend auf dem Martingal-Konvergenzsatz findet sich z.B. in [Williams: Prob. with martingales]. Die folgenden elementaren Aussagen ergeben sich unmittelbar aus (12.1.1):

Satz 12.1. *(1). Ist μ absolutstetig bzgl. ν mit ν-fast überall strikt positiver relativer Dichte, dann ist auch ν absolutstetig bzgl. μ und*

$$\frac{d\nu}{d\mu}(x) \quad = \quad \left(\frac{d\mu}{d\nu}(x)\right)^{-1} \qquad \textit{für } \mu\textit{-fast alle } x \in S.$$

(2). Sind μ und ν beide absolutstetig bzgl. eines Referenzmaßes λ mit Dichten f und g, und gilt $g > 0$ λ-fast überall, dann ist μ absolutstetig bzgl. ν mit relativer Dichte

$$\frac{d\mu}{d\nu}(x) \quad = \quad \frac{f(x)}{g(x)} \qquad \textit{für } \nu\textit{-fast alle } x \in S.$$

(3). Sind $\mu_1, \ldots, \mu_n$ und $\nu_1, \ldots, \nu_n$ Wahrscheinlichkeitsverteilungen auf messbaren Räumen $(S_1, \mathcal{S}_1), \ldots, (S_n, \mathcal{S}_n)$ mit $\mu_i \ll \nu_i$ für alle $1 \le i \le n$, dann ist auch $\mu_1 \otimes \mu_2 \otimes \ldots \otimes \mu_n$ absolutstetig bzgl. $\nu_1 \otimes \nu_2 \otimes \ldots \otimes \nu_n$ mit relativer Dichte

$$\frac{d(\mu_1 \otimes \ldots \mu_n)}{d(\nu_1 \otimes \ldots \otimes \nu_n)}(x_1, \ldots, x_n) \quad = \quad \prod_{i=1}^n \frac{d\mu_i}{d\nu_i}(x_i).$$

Die letzte Aussage gilt nicht für unendliche Produkte.

Beispiel (Singularität von unendlichen Produktmaßen). Sind μ und ν zwei unterschiedliche Wahrscheinlichkeitsverteilungen auf einem messbaren Raum $(S, \mathcal{S})$, dann ist das unendliche Produkt $\mu^\infty := \bigotimes\limits_{i \in \mathbb{N}} \mu$ **nicht** absolutstetig bzgl. $\nu^\infty := \bigotimes\limits_{i \in \mathbb{N}} \nu$. In der Tat gilt nämlich nach dem Gesetz der großen Zahlen:

$$\begin{aligned}
\mu^\infty \left[\left\{ (x_1, x_2, \ldots) \in S^\infty \, : \, \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n I_B(x_i) \, = \, \mu[B] \right\} \right] \quad &= \quad 1 \\
\nu^\infty \left[\left\{ (x_1, x_2, \ldots) \in S^\infty \, : \, \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n I_B(x_i) \, = \, \nu[B] \right\} \right] \quad &= \quad 1
\end{aligned}$$

für alle $B \in \mathcal{S}$. Ist $\mu \neq \nu$, dann existiert eine Menge $B \in \mathcal{S}$ mit $\mu[B] \neq \nu[B]$. Also sind die Wahrscheinlichkeitsverteilungen μ^∞ und ν^∞ in diesem Fall singulär.

In Satz 12.10 werden wir sehen, dass die relativen Dichten $d\mu^n/d\nu^n$ der endlichen Produktmaße für $\mu \neq \nu$ und $n \to \infty$ exponentiell schnell anwachsen.

Sind μ und ν Wahrscheinlichkeitsverteilungen auf einem messbaren Raum $(S, \mathcal{S})$ mit *beschränkter relativer Dichte,* dann können wir ein Acceptance-Rejection Verfahren verwenden, um Stichproben von der Verteilung μ aus Stichproben der Verteilung ν zu erzeugen.
In vielen praktischen Anwendungen ist die Dichte nur bis auf eine Normierungskonstante explizit bekannt. Wir nehmen daher an, dass eine beschränkte Funktion $\varrho : S \to \mathbb{R}_+$ und Konstanten $Z, C \in (0, \infty)$ gegeben sind mit

$$\frac{d\mu}{d\nu}(x) \quad = \quad \frac{1}{Z} \cdot \varrho(x), \qquad \varrho(x) \quad \leq \quad C \quad \text{für alle } x \in S. \tag{12.1.2}$$

Dies ist beispielsweise der Fall, wenn μ und ν absolutstetige Verteilungen auf $\mathbb{R}^d$ mit Dichten proportional zu $f(x)$ bzw. $g(x)$ sind, und

$$f(x) \quad \leq \quad C \cdot g(x) \qquad \text{für alle } x \in \mathbb{R}^d$$

gilt. In diesem Fall können wir $\varrho = f/g$ wählen. Die Konstante C sollte explizit bekannt sein – die Normierungskonstante $Z = \int \varrho \, d\nu$ kennt man dagegen meistens nicht. Gilt (12.1.2), dann können wir μ folgendermaßen als bedingte Verteilung darstellen:

Lemma 12.2. *Sei X eine Zufallsvariable mit Verteilung ν, und sei U eine unabhängige, auf $(0,1)$ gleichverteilte Zufallsvariable. Dann gilt:*

$$\mu[B] \quad = \quad P\left[X \in B \,\middle|\, U \leq \frac{\varrho(X)}{C}\right] \qquad \textit{für alle } B \in \mathcal{S}.$$

Beweis. Die gemeinsame Verteilung von X und U ist $\nu \otimes \mathcal{U}_{(0,1)}$. Also gilt nach dem Satz von Fubini:

$$\begin{aligned} P\left[X \in B, U \leq \frac{\varrho(X)}{C}\right] &= \int_B \int_{(0, \frac{\varrho(x)}{C})} \lambda(du)\nu(dx) \\ &= \frac{1}{C} \cdot \int_B \varrho(x)\nu(dx) \\ &= \frac{Z}{C} \cdot \mu[B], \end{aligned}$$

und insbesondere

$$P\left[U \leq \frac{\varrho(X)}{C}\right] \quad = \quad \frac{Z}{C} \cdot \mu[S] \quad = \quad \frac{Z}{C}.$$

Die bedingte Wahrscheinlichkeit ist der Quotient der beiden Ausdrücke. □

Das Lemma motiviert das folgende Verwerfungsverfahren zur Simulation von Stichproben von der Wahrscheinlichkeitsverteilung μ:

Algorithmus 12.3 (Acceptance-Rejection-Verfahren). **repeat**
erzeuge unabhängige Stichproben $x \sim \nu$ und $u \sim \mathcal{U}_{(0,1)}$
until $u \leq \frac{\varrho(x)}{C}$
return x

Der folgende Satz zeigt, dass der Algorithmus im Mittel nach C/Z Schritten eine Stichprobe von μ liefert:

Satz 12.4. *Seien $X_1, X_2, \ldots : \Omega \to S$ und $U_1, U_2, \ldots : \Omega \to (0,1)$ unter P unabhängige Zufallsvariablen mit Verteilungen ν bzw. $\mathcal{U}_{(0,1)}$. Dann ist die erste Akzeptanzzeit*

$$T(\omega) \quad := \quad \min\left\{ k \in \mathbb{N} \;\middle|\; U_k(\omega) \leq \frac{\varrho(X_k(\omega))}{C} \right\}$$

geometrisch verteilt mit Parameter Z/C, und die (fast überall definierte) Zufallsvariable

$$Y(\omega) \quad := \quad X_{T(\omega)}(\omega)$$

hat die Verteilung μ.

Beweis. Da die Ereignisse $E_k := \{U_k \leq \frac{\varrho(X_k)}{C}\}$ unabhängig sind, ist die Zufallsvariable $T(\omega) = \min\{k \in \mathbb{N} \mid \omega \in E_k\}$ geometrisch verteilt mit Parameter

$$p \quad = \quad P[E_k] \quad = \quad P\left[U_k \leq \frac{\varrho(X_k)}{C}\right] \quad \overset{\text{Lemma 12.2}}{=} \quad \frac{Z}{C}.$$

Weiterhin folgt nach Lemma 12.2:

$$\begin{aligned} P[Y \in B] &= \sum_{k=1}^{\infty} P[X_T \in B, T = k] \quad = \quad \sum_{k=1}^{\infty} P[\{X_k \in B\} \cap E_1^C \cap \ldots \cap E_{k-1}^C \cap E_k] \\ &= \sum_{k=1}^{\infty} P[\{X_k \in B\} \cap E_k] \prod_{i=1}^{k-1} P[E_i^C] \quad = \quad \sum_{k=1}^{\infty} P[X_k \in B \mid E_k] \cdot p \cdot (1-p)^{k-1} \\ &= \mu[B]. \end{aligned}$$

□

Bemerkung. (1). Im Algorithmus kommt nur das Verhältnis $\varrho(x)/C$ vor, und die Konstante C kann frei gewählt werden, solange $C \geq \sup \varrho$ gilt. Um das Acceptance-Rejection-Verfahren einzusetzen, benötigen wir daher lediglich eine obere Schranke für die *unnormierte* Dichte ϱ.

(2). Die mittlere Anzahl von Versuchen bis zur Akzeptanz beträgt $E[T] = C/Z$. Der Algorithmus ist also umso effizienter, je kleiner C gewählt wird.

Die letzte Bemerkung zeigt auch eine Schwäche des AR-Verfahrens: Damit die Methode praktikabel ist, muss die relative Dichte *gleichmäßig* durch eine Konstante beschränkt sein, die nicht zu groß ist. Dies ist besonders in hohen Dimensionen häufig nicht der Fall. Ist man nur an Schätzern von Erwartungswerten, und nicht an der Simulation einzelner Stichproben interessiert, dann bietet es sich an, Importance Sampling anstelle eines AR-Verfahrens zu verwenden. In diesem Fall wird zumindest keine gleichmäßige Schranke für die relative Dichte benötigt, s.u. Alternative Verfahren, um Stichproben zu generieren sind Markov Chain Monte Carlo (MCMC) Methoden.

Beispiel (**Abgeschnittene Normalverteilungen**). Für $a > 0$ sei

$$\mu \quad := \quad N(0,1)[\ \bullet\ |\ (a,\infty)\ ]$$

die auf Werte größer als a konditionierte Standardnormalverteilung. Die Dichte ist proportional zu

$$f(x) \quad = \quad e^{-x^2/2} I_{(a,\infty)}(x).$$

Eine naive Methode zur Simulation einer Stichprobe von μ ist, solange Stichproben von $N(0,1)$ zu erzeugen, bis ein Wert größer als a auftritt. Für große a ist dieses Verfahren jedoch extrem ineffizient, da die Akzeptanzwahrscheinlichkeit $N(0,1)[(a,\infty)]$ sehr klein ist. Besser geht man wie folgt vor: Für $x > a$ gilt

$$f(x) \quad = \quad e^{-(a+(x-a))^2/2} \quad = \quad e^{-a^2/2} \cdot e^{-a(x-a)-(x-a)^2/2}.$$

Wir schätzen diese Dichte durch die Dichte

$$g_\lambda(x) \quad = \quad \lambda \cdot e^{-\lambda(x-a)} \cdot I_{(a,\infty)}(x)$$

einer verschobenen Exponentialverteilung mit Parameter $\lambda \geq a$ ab.

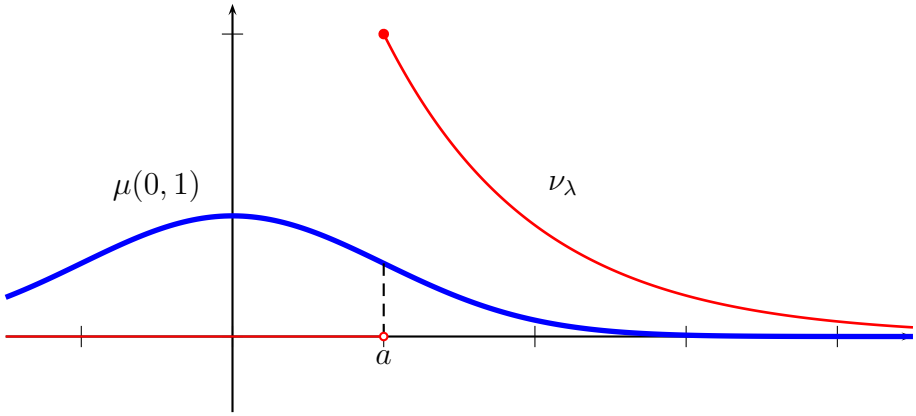


Abbildung 12.1: Dichten der Normalverteilung und der Sampling-Verteilung für die abgeschnittene Normalverteilung (hier die Dichte der Exponentialverteilung).

Maximieren von $\varrho_\lambda = f/g_\lambda$ liefert:

$$C(\lambda) \quad := \quad \sup_{x \geq a} \varrho_\lambda(x) \quad = \quad \frac{1}{\lambda} \exp((\lambda - a)^2/2).$$

Diese Funktion ist für $\lambda = (a + \sqrt{a^2+1})/2$ minimal. Damit bietet sich der folgende Algorithmus zum Simulieren einer Stichprobe von μ an:

Setze $\lambda := (a + \sqrt{a^2+1})/2$
repeat
  erzeuge unabhängige Stichproben u_1, u_2 von $\mathcal{U}_{(0,1)}$
  setze $x := a - \frac{1}{\lambda} \log u_1$ (simuliert Stichprobe von ν_λ)
until $u_2 \leq \frac{\varrho_\lambda(x)}{C(\lambda)} \qquad \left(= \frac{f(x)}{g_\lambda(x) \cdot C(\lambda)}\right)$
return x

Seltene Ereignisse und Importance Sampling

Sei μ eine Wahrscheinlichkeitsverteilung auf einem messbaren Raum $(S, \mathcal{S})$. Angenommen, wir wollen die Wahrscheinlichkeit

$$\theta \quad = \quad \mu[A] \quad = \quad \int I_A \, d\mu$$

eines Ereignisses $A \in \mathcal{S}$ mit einem Monte-Carlo-Verfahren näherungsweise berechnen. Der klassische Monte-Carlo-Schätzer

$$\hat{\theta}_k \quad = \quad \frac{1}{k} \sum_{i=1}^{k} I_A(Y_i), \qquad Y_i \text{ unabhängig mit Verteilung } \mu,$$

ist erwartungstreu mit Varianz

$$\mathrm{Var}[\hat{\theta}_k] \quad = \quad \frac{1}{k} \mathrm{Var}_\mu[I_A] \quad = \quad \frac{\theta \cdot (1-\theta)}{k}$$

und relativem Fehler

$$E[|\hat{\theta}_k - \theta|^2]^{1/2}/\theta \quad = \quad \sigma(\hat{\theta}_k)/\theta \quad = \left(\frac{1-\theta}{k \cdot \theta} \right)^{1/2}.$$

Für seltene Ereignisse ist der relative Fehler hoch, und das Schätzverfahren ineffizient. Ist ν eine andere Wahrscheinlichkeitsverteilung mit $\mu \ll \nu$, dann können wir alternativ den Importance Sampling Schätzer

$$\widetilde{\theta}_k \quad = \quad \frac{1}{k} \sum_{i=1}^{k} I_A(X_i) \frac{d\mu}{d\nu}(X_i), \qquad X_i \text{ unabhängig mit Verteilung } \nu,$$

verwenden. Auch $\widetilde{\theta}_k$ ist erwartungstreu, denn

$$\theta \quad = \quad \mu[A] \quad = \quad \int I_A \frac{d\mu}{d\nu} \, d\nu \quad = \quad E[\widetilde{\theta}_k].$$

Zudem gilt

$$\mathrm{Var}[\widetilde{\theta}_k] \quad = \quad \frac{1}{k} \mathrm{Var}_\nu \left[I_A \cdot \frac{d\mu}{d\nu} \right]. \tag{12.1.3}$$

Es stellt sich die Frage, wie wir eine Wahrscheinlichkeitsverteilung ν finden, von der wir Stichproben simulieren können, und für die die Varianz in (12.1.3) möglichst klein ist. Wir betrachten zunächst ein Beispiel:

Beispiel (Berechnung Gaußscher Wahrscheinlichkeiten). Sei C eine strikt positiv definite symmetrische $d \times d$-Matrix, und sei $\mu = N(0, C)$ die multivariate Normalverteilung im $\mathbb{R}^d$ mit Dichte

$$f(x) \quad = \quad \frac{1}{\sqrt{(2\pi)^d (\det C)}} \exp\left(-\frac{1}{2} x \cdot C^{-1} x \right).$$

Angenommen, wir wollen die Wahrscheinlichkeit $\theta = \mu[A]$ einer offenen Menge $A \subseteq \mathbb{R}^d$ mit einem Monte-Carlo Verfahren berechnen. Ist der Nullpunkt in der Menge A enthalten, dann ist A ein „typisches“ Ereignis bzgl. μ, und wir können in der Regel den klassischen Monte-Carlo

Schätzer $\hat{\theta}_k$ verwenden.

Hier interessiert uns der Fall $0 \notin A$. In diesem Fall ist die Wahrscheinlichkeit θ evtl. sehr klein – wir wenden daher ein Importance Sampling Verfahren an. Um eine geeignete Referenzverteilung ν zu erhalten, wählen wir einen Punkt x^* aus dem Abschluss $\overline{A}$ mit

$$f(x^*) \quad = \quad \sup_{x \in A} f(x), \qquad \text{d.h.} \qquad x^* \cdot C^{-1} x^* \; = \; \inf_{x \in A} x \cdot C^{-1} x, \tag{12.1.4}$$

und setzen

$$\nu \quad := \quad N(x^*, C).$$

Wir verschieben die Verteilung also so, dass sie in der Umgebung des „wahrscheinlichsten“ Punktes $x^* \in \overline{A}$ bzgl. μ, d.h. des Punktes mit maximaler Dichte, konzentriert ist. Die Verteilung ν ist absolutstetig mit Dichte

$$g(x) \quad = \quad \frac{1}{\sqrt{(2\pi)^d \det C}} \exp\left(-\frac{1}{2}(x - x^*) \cdot C^{-1}(x - x^*)\right).$$

Damit erhalten wir

$$\frac{d\mu}{d\nu}(x) \quad = \quad \frac{f(x)}{g(x)} \quad = \quad \exp\left(-x^* \cdot C^{-1} x + \frac{1}{2} x^* \cdot C^{-1} x^*\right).$$

Ist die Menge A konvex, dann gilt

$$x^* \cdot C^{-1}(x - x^*) \quad \geq \quad 0 \qquad \text{für alle } x \in A,$$

da x^* der Minimierer der quadratischen Form $x \mapsto x \cdot C^{-1} x$ in $\overline{A}$ ist. Damit erhalten wir

$$\sup_{x \in A} \frac{d\mu}{d\nu}(x) \quad = \quad \exp\left(-\frac{1}{2} x^* \cdot C^{-1} x^*\right),$$

und somit nach (12.1.3)

$$\begin{aligned} \operatorname{Var}[\widetilde{\theta}_k] \; &\leq \; \frac{1}{k} \int_A \left(\frac{d\mu}{d\nu}\right)^2 d\nu \quad = \quad \frac{1}{k} \int_A \frac{d\mu}{d\nu} \, d\mu \\ &\leq \; \frac{\theta}{k} \cdot \exp\left(-\frac{1}{2} x^* \cdot C^{-1} x^*\right). \end{aligned}$$

Offensichtlich ist dieser Wert in vielen Fällen deutlich kleiner als die Varianz $\theta(1-\theta)/k$ des klassischen Monte-Carlo Schätzers.

Wir wollen nun Importance Sampling Schätzer systematischer untersuchen. Sei allgemein

$$\theta \quad = \quad \int \phi \, d\mu \qquad \text{mit } \phi \in \mathcal{L}^1(\mu),$$

und sei ν eine zu μ absolutstetige Verteilung mit relativer Dichte

$$w \quad = \quad \frac{d\nu}{d\mu} \quad > \quad 0 \qquad \mu\text{-fast überall.}$$

Dann ist auch μ absolutstetig bzgl. ν mit relativer Dichte $1/w$, und es gilt $\phi/w \in \mathcal{L}^1(\nu)$. Wegen

$$\theta \quad = \quad \int \phi \, d\mu \quad = \quad \int \frac{\phi}{w} \, d\nu$$

ist der Importance Sampling Schätzer

$$\widetilde{\theta}_k \quad = \quad \frac{1}{k} \sum_{i=1}^{k} \phi(X_i)/w(X_i), \qquad X_i \text{ unabhängig mit Verteilung } \nu,$$

erwartungstreu, und nach dem Gesetz der großen Zahlen konsistent, d.h. $\widetilde{\theta}_k \to \theta$ P-fast sicher für $k \to \infty$. Für den mittleren quadratischen Fehler ergibt sich:

Satz 12.5 (MSE von Importance Sampling). *(1). Es gilt* $E[|\widetilde{\theta}_k - \theta|^2] = \sigma_\nu^2/k$ *mit*

$$\sigma_\nu^2 \quad = \quad \mathrm{Var}_\nu \left[\frac{\phi}{w} \right] \quad = \quad \left(\int \frac{\phi^2(x)}{w(x)} \, \mu(dx) \right) - \theta^2.$$

(2). Der mittlere quadratische Fehler ist minimal, falls w *proportional zu* $|\phi|$ *ist.*

Beweis. (1). Die Aussage folgt, da $\widetilde{\theta}_k$ erwartungstreu ist mit

$$k \cdot \mathrm{Var}[\widetilde{\theta}_k] \quad = \quad \mathrm{Var}_\nu[\phi/w] \quad = \quad \int \left(\frac{\phi(x)}{w(x)} - \theta \right)^2 w(x) \mu(dx).$$

(2). Aus der Cauchy-Schwarz Ungleichung ergibt sich

$$\left(\int |\phi| \, d\mu \right)^2 \quad = \quad \left(\int \frac{|\phi|}{\sqrt{w}} \sqrt{w} \, d\mu \right)^2 \leq \int \frac{\phi^2}{w} \, d\mu \cdot \int w \, d\mu \; = \; \sigma_\nu^2 + \theta^2$$

Dies liefert eine untere Schranke für den mittleren quadratischen Fehler. Zudem gilt Gleichheit in der Cauchy-Schwarz Ungleichung genau dann, wenn $\sqrt{w}$ proportional zu $|\phi|/\sqrt{w}$ ist, also, wenn $w \propto |\phi|$ ist.

□

Das Optimalitätsresultat aus Satz 12.5 ist eher von theoretischer als von praktischer Bedeutung, wie das folgende Beispiel zeigt:

Beispiel (Seltene Ereignisse). Ist $\theta = \mu[A]$ für eine Menge $A \in \mathcal{S}$, also $\phi = I_A$, dann ist

$$w \quad = \quad \frac{I_A}{\mu[A]}, \qquad \text{d.h. } \nu = \mu[\bullet \mid A],$$

die Importance Sampling Verteilung mit minimalem quadratischen Fehler. Die Simulation von Stichproben von der bedingten Verteilung ist jedoch für Ereignisse A mit kleiner Wahrscheinlichkeit oft nicht praktikabel. Das AR-Verfahren ist in diesem Fall ineffizient, da die mittlere Akzeptanzzeit mindestens $1/\mu[A]$ beträgt.

Zumindest liefert Satz 12.5 eine gewisse Rechtfertigung für die Faustregel, dass man bei der Auswahl einer IS Verteilung ν darauf achten sollte, die relative Dichte w von ν bzgl. μ dort groß zu wählen, wo auch der Integrand ϕ betragsmäßig große Werte annimmt.

Da die optimale Importance Sampling Verteilung gewöhnlich nicht realisierbar ist, betrachtet man stattdessen üblicherweise nur Verteilungen aus einer ein- oder mehrparametrigen Familie $(\nu_t)_{t\in\Theta}$ von Wahrscheinlichkeitsverteilungen, und versucht σ_ν^2 innerhalb dieser Familie zu minimieren. Am wichtigsten sind dabei die im nächsten Abschnitt betrachteten exponentiellen Familien, da diese eine Minimierungseigenschaft bzgl. der relativen Entropie besitzen, s. Satz 12.13 unten.

Bemerkung. (1). *Asymptotische Normalität:* Ist $\phi/w \in \mathcal{L}^2(\nu)$, dann folgt aus dem zentralen Grenzwertsatz die asymptotische Normalität des Importance Sampling Schätzers:

$$\sqrt{k}(\widetilde{\theta}_k - \theta) \quad \xrightarrow{\mathcal{D}} \quad N(0, \sigma_\nu^2) \qquad \text{für } k \to \infty.$$

Für praktische Anwendungen ist der nicht-asymptotische mittlere quadratische Fehler allerdings wichtiger.

(2). *Importance Sampling mit unnormierten Dichten:* In Anwendungen ist die relative Dichte oft nur bis auf eine Normierungskonstante bekannt, d.h. es gilt

$$\frac{d\mu}{d\nu}(x) \quad = \quad \frac{1}{w(x)} \quad \propto \quad \varrho(x)$$

mit einer explizit bekannten Funktion $\varrho(x)$, aber einem unbekannten Proportionalitätsfaktor. In diesem Fall können wir die Darstellung

$$\theta \quad = \quad \int \phi \, d\mu \quad = \quad \int \phi \frac{d\mu}{d\nu} d\nu \quad = \quad \frac{\int \phi\varrho \, d\nu}{\int \varrho \, d\nu}$$

nutzen, und θ durch den Schätzer

$$\overline{\theta}_k \quad = \quad \frac{\frac{1}{k}\sum_{i=1}^{k}\phi(X_i)\varrho(X_i)}{\frac{1}{k}\sum_{i=1}^{k}\varrho(X_i)} \quad = \quad \frac{\sum_{i=1}^{k}\phi(X_i)\varrho(X_i)}{\sum_{i=1}^{k}\varrho(X_i)}, \qquad X_i \text{ i.i.d.} \sim \nu, \tag{12.1.5}$$

approximieren. Nach dem Gesetz der großen Zahlen ist $\overline{\theta}_k$ konsistent, d.h. $\overline{\theta}_k \to \theta$ fast sicher für $k \to \infty$. Ein zentraler Grenzwertsatz gilt ebenfalls. Allerdings ist $\overline{\theta}_k$ i. A. nicht erwartungstreu, und der nicht-asymptotische mittlere quadratische Fehler ist nicht so leicht zu kontrollieren, da der Nenner in (12.1.5) degenerieren kann.

(3). *Schätzen der Varianz:* Ein weiteres zentrales Problem in Anwendungen ist, dass die Varianz σ_ν^2 in der Regel nicht bekannt ist, und häufig auch keine guten Abschätzungen für σ_ν^2 vorliegen. Daher behilft man sich in der Praxis oft damit, die Varianz empirisch zu schätzen, z.B. durch

$$s_k^2 \quad = \quad \frac{1}{k-1}\sum_{i=1}^{k}\left(\frac{\phi(X_i)}{w(X_i)} - \widetilde{\theta}_k\right)^2.$$

Die empirische Schätzung kann jedoch irreführend sein, wie das folgende warnende Beispiel zeigt:

Beispiel. Sei μ das Lebesguemaß auf $\mathbb{R}$,

$$\phi(x) \quad = \quad (2\pi)^{-1/2}\exp(-|x-m|^2/2)$$

die Dichte der Normalverteilung mit Varianz 1 und Mittelwert $m \in \mathbb{R}$, und ν die Standardnormalverteilung. In diesem Fall gilt

$$\theta \quad = \quad \int \phi \, d\mu \quad = \quad 1.$$

Obwohl μ keine Wahrscheinlichkeitsverteilung ist, können wir wie oben Importance Sampling mit Referenzverteilung ν durchführen. Es gilt dann $w(x) = (2\pi)^{-1/2}\exp(-x^2/2)$ und

$$\sigma_\nu^2 \;=\; \int \frac{\phi(x)^2}{w(x)}\,dx - 1 \;=\; (2\pi)^{-1/2}\int e^{m^2-(x-2m)^2/2}\,dx - 1 \;=\; e^{m^2} - 1.$$

Hieraus folgt, dass schon für $m = 5$ mindestens $k \geq 6.5 \cdot 10^{15}$ Stichproben benötigt werden, damit für den mittleren quadratischen Fehler

$$E[|\widetilde{\theta}_k - \theta|^2]^{1/2} \quad = \quad \sigma_\nu \cdot k^{-1/2} \quad < \quad \frac{1}{3}$$

gilt. Empirisches Schätzen der Varianz in Simulationsläufen liefert ganz andere Ergebnisse. Beispielsweise erhielten wir für $k = 10^6$ Stichproben in einem typischen Simulationslauf

$$s_k^2 = 6816, \qquad \text{d.h. } s_k \cdot k^{-1/2} \approx 0.08 < 1/3.$$

Die empirische Varianzschätzung suggeriert hier also die falsche Aussage, dass der Schätzwert bereits für $k = 10^6$ genau genug ist!

Die Ursache ist in diesem Fall, dass praktisch keine Stichproben im relevanten Bereich $x \approx m$ landen. Daher ist nicht nur der Schätzwert für θ, sondern auch die empirische Varianz sehr klein. Tatsächlich rechnet man in dem Beispiel leicht nach, dass

$$\operatorname{Var}[s_k^2] \quad = \quad \frac{1}{k-1} \operatorname{Var}_\nu[\phi/w] \quad \leq \quad \frac{1}{k-1}(e^{6m^2} - 1)$$

gilt – der Schätzer s_k^2 für die Varianz ist also völlig unbrauchbar. Das Problem ist, dass in vielen Anwendungen ähnliche Effekte auftreten können, aber nicht so leicht zu erkennen sind.

12.2 Exponentielle Familien und große Abweichungen

In diesem Abschnitt wollen wir uns überlegen, wie die Wahrscheinlichkeiten großer Abweichungen vom Gesetz der großen Zahlen sowohl asymptotisch als auch numerisch berechnet werden können. In beiden Fällen hilft uns dasselbe Prinzip weiter: Ein Maßwechsel zu einer Verteilung aus einer exponentiellen Familie.

Exponentielle Familien

Sei μ ein positives Maß auf $(S, \mathcal{S})$, $U : S \to \mathbb{R}^d$ eine messbare Funktion, und

$$Z(t) \quad = \quad \int e^{t \cdot U} \, d\mu, \qquad t \in \mathbb{R}^d,$$

die momentenerzeugende Funktion von U mit Definitionsbereich

$$\Theta \quad = \quad \{t \in \mathbb{R}^d \mid Z(t) < \infty\}.$$

Für $t \in \Theta$ sei

$$\Lambda(t) \quad = \quad \log Z(t)$$

die kumulantenerzeugende Funktion.

Definition. *Die Familie der Wahrscheinlichkeitsverteilungen*

$$\nu_t(dx) \quad := \quad \frac{1}{Z(t)} e^{t \cdot U(x)} \mu(dx) \quad = \quad e^{t \cdot U(x) - \Lambda(t)} \mu(dx), \qquad t \in \Theta,$$

heißt ***exponentielle Familie zu*** μ ***und*** U.

Bemerkung (Boltzmannverteilung). In der statistischen Physik treten exponentielle Familien als Gleichgewichtsverteilungen auf. Beispielsweise ist die Verteilung im thermodynamischen Gleichgewicht in einem abgeschlossenen System bei inverser Temperatur $\beta = 1/T$ gleich ν_β, wobei μ die Gleichverteilung bzw. das Lebesguemaß auf dem Zustandsraum und $U(x) = -H(x)$ die negative Energie des Zustandes x ist. Die Normierungskonstante $Z(\beta)$ heißt in der statistischen Physik *Partitionsfunktion*.

Wir betrachten nun einige elementare Beispiele von exponentiellen Familien:

Beispiel. (1). **Exponential und Gammaverteilungen.** Ist μ die Exponentialverteilung mit Parameter $\lambda > 0$, und $U(x) = -x$, dann ist $M(t)$ für $t > -\lambda$ endlich, und es gilt

$$\nu_t \quad = \quad \mathrm{Exp}(\lambda + t) \qquad \text{für alle } t > -\lambda.$$

Die exponentielle Familie besteht also aus allen Exponentialverteilungen.
Ist $\mu = \Gamma(\alpha, \lambda)$ eine Gammaverteilung, dann gilt entsprechend $\nu_t = \Gamma(\alpha, \lambda + t)$.

(2). **Bernoulli-, Binomial- und Poissonverteilungen** Ist μ die Bernoulliverteilung mit Parameter p und $U(k) = k$, dann gilt $\nu_t(1) = p_t$ mit

$$p_t \quad = \quad \frac{e^t p}{e^t p + 1 - p} \quad = \quad \frac{p}{p + (1-p)e^{-t}},$$

d.h. ν_t ist die Bernoulliverteilung mit Parameter p_t. Entsprechend gilt für $U(k) = k$:

$$\begin{aligned} \mu &= \mathrm{Bin}(n, p) \quad \Rightarrow \quad \nu_t = \mathrm{Bin}(n, p_t), \qquad \text{und} \\ \mu &= \mathrm{Poisson}(\lambda) \quad \Rightarrow \quad \nu_t = \mathrm{Poisson}(\lambda e^t). \end{aligned}$$

Die exponentielle Familie besteht also jeweils aus allen Bernoulliverteilungen, Binomialverteilungen mit festem n, bzw. Poissonverteilungen.

(3). **Normalverteilungen.** Ist $\mu = N(m, C)$ eine d-dimensionale Normalverteilung, und $U(x) = x$, dann gilt $\nu_t = N(m + Ct, C)$ für $t \in \mathbb{R}^d$. Im nichtdegenerierten Fall enthält die exponentielle Familie also alle Normalverteilungen mit fester Kovarianzmatrix C. Für $d = 1, \mu = N(m, \sigma^2)$, und

$$U(x) \quad = \quad -\frac{(x-m)^2}{2}$$

erhält man

$$\nu_t \quad = \quad N\left(m, \left(\frac{1}{\sigma^2} + \frac{1}{t}\right)^{-1}\right) \qquad \text{für } t > 0,$$

d.h. die exponentielle Familie besteht aus Normalverteilungen mit festem Mittelwert m. Entsprechend kann man die Familie der eindimensionalen Normalverteilungen als zweiparametrige exponentielle Familie bzgl. einer Referenz-Normalverteilung interpretieren.

Wir beschränken uns nun auf den Fall $d = 1$. Sei $(\nu_t)_{t\in\Theta}$ eine einparametrige exponentielle Familie zu μ und U, und sei $\overset{\circ}{\Theta} = \Theta \setminus \partial\Theta$ der offene Kern des Definitionsbereichs.

Lemma 12.6 (Eigenschaften exponentieller Familien).

(1). Es gilt $Z \in C^\infty(\overset{\circ}{\Theta})$. Für $t \in \overset{\circ}{\Theta}$ existieren die Erwartungswerte und Varianzen

$$m(t) \quad = \quad \int U \, d\nu_t \qquad \textit{bzw.} \qquad v(t) \quad = \quad \mathrm{Var}_{\nu_t}[U],$$

und es gilt

$$m(t) \quad = \quad \Lambda'(t) \qquad \textit{und} \qquad v(t) \quad = \quad \Lambda''(t).$$

(2). Die Funktion m ist auf $\overset{\circ}{\Theta}$ beliebig oft differenzierbar und monoton wachsend. Ist U nicht ν-fast überall konstant, dann ist m sogar strikt monoton wachsend. Im Fall $\Theta = \mathbb{R}$ gilt zudem

$$\lim_{t\to\infty} m(t) \quad = \quad \operatorname{esssup} U \quad = \quad \inf\{a \in \mathbb{R} \,:\, \mu[U > a] = 0\}, \quad \textit{und} \tag{12.2.1}$$

$$\lim_{t\to-\infty} m(t) \quad = \quad \operatorname{essinf} U \quad = \quad \sup\{a \in \mathbb{R} \,:\, \mu[U < a] = 0\}, \tag{12.2.2}$$

d.h. $m : \mathbb{R} \to (\operatorname{essinf} U, \operatorname{esssup} U)$ ist bijektiv.

Beweis. (1). Sei $t \in \overset{\circ}{\Theta}$. Wir betrachten die momentenerzeugende Funktion

$$M(s) \quad = \quad \int e^{sU} \, d\nu_t$$

der Verteilung ν_t. Wegen $t \in \overset{\circ}{\Theta}$ gilt

$$M(s) \quad = \quad \int \frac{1}{Z(t)} e^{(s+t)U} \, d\mu \quad = \quad Z(s+t)/Z(t) \quad < \quad \infty \tag{12.2.3}$$

für alle s in einer Umgebung $(-\varepsilon, \varepsilon)$ der 0, also $M \in C^\infty(-\varepsilon, \varepsilon)$. Wegen (12.2.3) folgt $Z \in C^\infty(t-\varepsilon, t+\varepsilon)$,

$$\begin{aligned} \int U \, d\nu_t \quad &= \quad M'(0) \quad = \quad \frac{Z'(t)}{Z(t)} \quad = \quad \Lambda'(t), \qquad \text{und} \\ \mathrm{Var}_{\nu_t}[U] \quad &= \quad (\log M)''(0) \quad = \quad \Lambda''(t). \end{aligned}$$

(2). Aus (1) folgt $m = \Lambda' \in C^\infty(\overset{\circ}{\Theta})$ und $m' = v$. Also ist m monoton wachsend, und strikt monoton wachsend, falls $\mathrm{Var}_{\nu_t}[U] > 0$. Für $a \in (\operatorname{essinf} U, \operatorname{esssup} U)$ folgt mit monotoner Konvergenz

$$\frac{\nu_t[U \le a]}{\nu_t[U > a]} \quad = \quad \frac{\int e^{tU} \cdot I_{\{U \le a\}}\, d\mu}{\int e^{tU} \cdot I_{\{U > a\}}\, d\mu} \quad = \quad \frac{\int e^{t(U-a)} \cdot I_{\{U \le a\}}\, d\mu}{\int e^{t(U-a)} \cdot I_{\{U > a\}}\, d\mu} \quad \to \quad 0$$

für $t \to \infty$, also $\lim_{t\to\infty} \nu_t[U > a] = 1$. Hieraus folgt

$$\liminf_{t\to\infty} m(t) \quad \ge \quad a \cdot \liminf_{t\to\infty} \nu_t[U > a] \quad = \quad a \qquad \text{für alle } a < \operatorname{esssup} U,$$

also (12.2.1). Die Aussage (12.2.2) zeigt man analog.

□

Beispiel (Isingmodell). Das Isingmodell wurde 1925 in der Dissertation von Ernst Ising mit der Absicht eingeführt, Phasenübergänge von ferromagnetischen Materialien in einem vereinfachten mathematischen Modell nachzuweisen. Heute spielt das Isingmodell eine wichtige Rolle als einfach zu formulierendes, aber schwer zu analysierendes grundlegendes mathematisches Modell, das auch in unterschiedlichen Anwendungsbereichen wie z.B. der Bildverarbeitung eingesetzt wird.

Sei $S = \{-1, 1\}^V$, wobei V die Knotenmenge eines endlichen Graphen (V, E) ist, z.B.

$$V \quad = \quad \{-k, -k+1, \ldots, k-1, k\}^d \subseteq \mathbb{Z}^d, \quad d, k \in \mathbb{N}.$$

Ein Element $\sigma = (\sigma_i | i \in V)$ aus S interpretieren wir physikalisch als Konfiguration von Spins $\sigma_i \in \{-1, 1\}$ an dem Knoten $i \in V$, wobei $\sigma_i = +1$ für einen Spin in Aufwärtsrichtung und $\sigma_i = -1$ für einen Spin in Abwärtsrichtung steht. Da benachbarte Spins sich vorzugsweise gleich ausrichten, ist die Energie einer Konfiguration σ durch

$$H(\sigma) \quad = \quad \sum_{(i,j)\in E} |\sigma_i - \sigma_j|^2 + h \cdot \sum_{i \in V} \sigma_i$$

gegeben, wobei die erste Summe über alle Kanten des Graphen läuft, und der zweite Term die Wechselwirkung mit einem äußeren Magnetfeld mit Stärke $h \in \mathbb{R}$ beschreibt. Als Gleichgewichtsverteilung bei inverser Temperatur $\beta = 1/T$ ergibt sich die Verteilung $\mu_{\beta,h}$ auf S mit Gewichten

$$\mu_{\beta,h}(\sigma) \quad \propto \quad \exp(-\beta \sum_{(i,j)\in E} |\sigma_i - \sigma_j|^2 - \beta h \cdot \sum_{i \in V} \sigma_i).$$

Die folgende Grafik zeigt Stichproben von der Verteilung $\mu_{\beta,h}$ auf einem ?×? Gitter V für verschiedene Werte von β und h. Für $\beta = 0$ (d.h. bei unendlicher Temperatur) ergibt sich eine

Gleichverteilung. Für $\beta \to \infty$ (Temperatur $\to 0$) konzentriert sich die Verteilung dagegen auf den energieminimierenden Konfigurationen. Dieses sind für $h = 0$ die beiden konstanten Konfigurationen $\sigma_i \equiv +1$ und $\sigma_i \equiv -1$, für $h \neq 0$ hat dagegen nur eine dieser Konfigurationen minimale Energie.

Der Satz von Cramér

Sei μ eine Wahrscheinlichkeitsverteilung auf einem messbaren Raum $(S, \mathcal{S})$, $U : S \to \mathbb{R}$ eine messbare Funktion, und sei $(X_i)_{i\in\mathbb{N}}$ eine Folge unabhängiger Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ mit Verteilung μ. Wir setzen voraus:

Annahmen:

(1). Alle exponentiellen Momente der Zufallsvariablen $U(X_i)$ existieren, d.h.

$$\Lambda(t) \quad = \quad \log \int e^{tU} d\mu \quad < \quad \infty \qquad \text{für alle } t \in \mathbb{R}.$$

(2). U ist nicht μ-fast sicher konstant.

Sei $a \in \mathbb{R}$ fest. Wir möchten nun die Asymptotik der Wahrscheinlichkeiten

$$\theta_n \quad = \quad P[S_n \geq an], \qquad S_n = \sum_{i=1}^{n} U(X_i),$$

für $n \to \infty$ genauer untersuchen. Nach dem Gesetz der großen Zahlen gilt:

$$S_n/n \quad \longrightarrow \quad m \, = \int U \, d\mu \qquad P\text{-fast sicher.}$$

Für $a > m$ ist das Ereignis $\{S_n \geq an\}$ also eine große Abweichung vom typischen Verhalten. Der Satz von Chernoff liefert eine obere Schranke der Wahrscheinlichkeiten θ_n. Um die Asymptotik genauer zu verstehen, führen wir eine Maßtransformation durch. Es gilt

$$\theta_n \quad = \quad \mu^n[A_n] \quad \text{mit } A_n = \left\{ x \in S^n \, : \, \sum_{i=1}^{n} U(x_i) \geq an \right\}. \tag{12.2.4}$$

Wir wollen zu einer Verteilung übergehen, bzgl. der das Ereignis A_n nicht mehr selten, sondern typisch ist. Dazu betrachten wir die Produktmaße $\nu_t^n, t \in \mathbb{R}$, wobei ν_t absolutstetig bzgl. μ ist mit Dichte

$$\frac{d\nu_t}{d\mu}(x) \quad = \quad \exp\left(tU(x) - \Lambda(t)\right).$$

Die relative Dichte von ν_t^n bzgl. μ^n ist dann

$$w_t^n(x_1,\ldots,x_n) \quad = \quad \prod_{i=1}^n \frac{d\nu_t}{d\mu}(x_i) \quad = \quad \exp\left(t\sum_{i=1}^n U(x_i) - n\Lambda(t)\right). \tag{12.2.5}$$

Man beachte, dass $(\nu_t^n)_{t\in\mathbb{R}}$ wieder eine exponentielle Familie ist. Es gilt

$$w_t^n(X_1,\ldots,X_n) \quad = \quad \exp(tS_n - n\Lambda(t)).$$

Bemerkung. Der stochastische Prozess $M_n = \exp(tS_n - n\Lambda(t)), n = 0,1,2,\ldots$, ist ein exponentielles Martingal. Exponentielle Martingale spielen in der stochastischen Analysis eine wichtige Rolle, s. [Introduction to Stochastical Analysis].

Wir wollen uns nun überlegen, wie wirden Parameter t in angemessener Weise wählen können. Wenn wir t zu klein wählen, dann hat das Ereignis A_n für große n nur eine geringe Wahrscheinlichkeit bzgl. ν_t^n. Wählen wir umgekehrt t sehr groß, dann liegt die Wahrscheinlichkeit $\nu_t^n[A_n]$ für große n nahe bei 1. In beiden Fällen sind Abschätzungen für $\nu_t^n[A_n]$ daher nur bedingt aussagekräftig. Um eine präzisere Aussage zu erhalten, sollten wir t so groß wählen, dass das Ereignis A_n „gerade typisch wird." Der Erwartungswert

$$m(t) \quad = \quad \int U\, d\nu_t, \qquad t\in\mathbb{R},$$

ist nach Lemma 12.6 strikt monoton wachsend. Wählen wir t^* mit

$$m(t^*) \quad = \quad a,$$

dann gilt nach dem Gesetz der großen Zahlen

$$\lim_{n\to\infty} \nu_{t^*}^n\left[\left\{x\in S^n \,:\, \frac{1}{n}\sum_{i=1}^n U(x_i)\in(a-\varepsilon,a+\varepsilon)\right\}\right] \quad = \quad 1 \qquad \text{für alle } \varepsilon>0,$$

und nach dem zentralen Grenzwertsatz

$$\lim_{n\to\infty} \nu_{t^*}^n\left[\left\{x\in \mathbb{R}^n \,:\, \frac{1}{n}\sum_{i=1}^n U(x_i)\geq a\right\}\right] \quad = \quad \frac{1}{2},$$

d.h. t^* ist gerade der gesuchte „Schwellenwert."

Die Umsetzung unserer Überlegungen führt zu einer ersten Aussage über die Asymptotik der Wahrscheinlichkeiten großer Abweichungen vom Gesetz der großen Zahlen auf der exponentiellen Skala:

Satz 12.7 (Cramér). *Unter den Annahmen von oben gilt*

$$\lim_{n\to\infty} \frac{1}{n} \log P\left[\frac{S_n}{n} \geq a\right] \quad = \quad -I(a) \qquad \textit{für alle } a \in (m, \operatorname{esssup} U),$$

wobei die Ratenfunktion

$$I(a) \quad = \quad \sup_{t\in\mathbb{R}} (ta - \Lambda(t))$$

die Legendretransformation von Λ ist.

Bemerkung. Der Satz von Cramér besagt, dass sich die Wahrscheinlichkeiten $\theta_n = P[S_n/n \geq a]$ asymptotisch wie $\exp(-n \cdot I(a))$ verhalten, wenn man subexponentiell wachsende Faktoren vernachlässigt. Er besagt *nicht*, dass die Folgen (θ_n) und $(\exp(-n \cdot I(a)))$ asymptotisch äquivalent sind!

Beweis. Der Beweis setzt sich zusammen aus einer nicht-asymptotischen Abschätzung der Wahrscheinlichkeiten

$$\theta_n \quad = \quad P[S_n \geq an] \quad = \quad \mu^n[A_n], \qquad A_n \;=\; \{x \in S^n \;:\; \sum_{i=1}^n U(x_i) \geq an\},$$

nach oben, und einer asymptotischen Abschätzung der Wahrscheinlichkeit nach unten.

(1). *Obere Schranke.* Die nicht-asymptotische obere Schranke

$$\frac{1}{n} \log \theta_n \leq -I(a) \qquad \text{für alle } n \in \mathbb{N}$$

liefert der Satz von Chernoff (Satz 8.3). Zur Illustration schreiben wir das Hauptargument aus dem Beweis von oben noch einmal so auf, dass der Zusammenhang mit einer Maßtransformation verdeutlicht wird: Für $t > 0$ gilt nach (12.2.5):

$$\begin{aligned} \theta_n \quad = \quad \mu^n[A_n] \quad &= \quad \int_{A_n} \frac{1}{w_t^n}\, d\nu_t^n \\ &= \quad \int_{A_n} \exp\left(-t \sum_{i=1}^n U(x_i) + \Lambda(t)n\right)\, d\nu_t^n \\ &\leq \quad e^{-(ta-\Lambda(t))n} \cdot \nu_t^n[A_n] \\ &\leq \quad e^{-(ta-\Lambda(t))n}. \end{aligned}$$

Hieraus folgt die Behauptung wie im Beweis von Satz 8.3 durch Optimieren der Abschätzung in t.

(2). *Untere Schranke.* Wir zeigen nun die asymptotische untere Schranke

$$\liminf_{n\to\infty} \frac{1}{n} \log \mu^n[A_n] \quad \geq \quad -I(a). \tag{12.2.6}$$

Zusammen mit der oberen Schranke folgt dann

$$\lim_{n\to\infty} \frac{1}{n} \log \mu^n[A_n] \quad = \quad -I(a),$$

d.h. die obere Schranke ist asymptotisch „scharf“. Zum Beweis von (12.2.6) gehen wir zu der Verteilung $\nu_{t^*}^n$ zum Schwellenwert $t^* = m^{-1}(a)$ über. Nach Lemma 12.6 ist $m : \mathbb{R} \to (\operatorname{essinf} U, \operatorname{esssup} U)$ bijektiv, also existiert $m^{-1}(a) > 0$ für $a \in (m, \operatorname{esssup} U)$. Für $\varepsilon > 0$ sei

$$A_{n,\varepsilon} \quad = \quad \left\{ x \in S^n \,:\, a \leq \frac{1}{n} \sum_{i=1}^n U(x_i) \leq a + \varepsilon \right\}.$$

Ähnlich wie bei der oberen Schranke erhalten wir

$$\begin{aligned} \mu^n[A_n] \quad \geq \quad \mu^n[A_{n,\varepsilon}] \quad &= \quad \int\limits_{A_{n,\varepsilon}} \exp\left(-t^* \sum_{i=1}^n U(x_i) + \Lambda(t) n \right) \, d\nu_{t^*}^n \\ &\geq \quad e^{-(t^*(a+\varepsilon) - \Lambda(t^*))n} \nu_{t^*}^n[A_{n,\varepsilon}] \\ &\geq \quad e^{-I(a)\cdot n} e^{-t^*\varepsilon n} \cdot \nu_{t^*}^n[A_{n,\varepsilon}] \end{aligned} \tag{12.2.7}$$

Wegen $\int U \, d\nu_{t^*} = m(t^*) \,=\, a$ gilt nach dem zentralen Grenzwertsatz:

$$\begin{aligned} \nu_{t^*}^n[A_{n,\varepsilon}] \quad &= \quad \nu_{t^*}^n \left[0 \leq \frac{1}{\sqrt{n}} \sum_{i=1}^n (U(x_i) - a) \leq \varepsilon \sqrt{n} \right] \\ &\overset{n\to\infty}{\longrightarrow} \quad N(0, \operatorname{Var}[U])[[0,\infty)] \quad = \quad \frac{1}{2}, \end{aligned} \tag{12.2.8}$$

d.h. die große Abweichung ist typisch unter $\nu_{t^*}^n$.
Für die Wahrscheinlichkeiten bzgl. μ^n ergibt sich dann nach (12.2.7):

$$\liminf \frac{1}{n} \log \mu^n[A_n] \geq -I(a) - t^*\varepsilon.$$

Die Behauptung folgt für $\varepsilon \searrow 0$.

□

Bemerkung. Ähnliche Aussagen über die Asymptotik von Wahrscheinlichkeiten großer Abweichungen wurden auch in vielen Modellen mit Abhängigkeit bewiesen. Sie spielen unter anderem in der mathematischen statistischen Mechanik eine wichtige Rolle.

Asymptotische Effizienz von IS Schätzern

Der beschriebene Maßwechsel ermöglicht nicht nur die asymptotische Berechnung der Wahrscheinlichkeiten

$$\theta_n \quad = \quad P[S_n \geq an] \quad = \quad \mu^n[A_n].$$

Wir können den Maßwechsel auch praktisch verwenden, um die Wahrscheinlichkeiten θ_n numerisch mithilfe von Importance Sampling zu berechnen. Wählen wir ν_t^n als Referenzmaß, dann erhalten wir nach (12.2.5) die Importance Sampling Schätzer

$$\begin{aligned}\widetilde{\theta}_n^{(k)} &= \frac{1}{k}\sum_{j=1}^{k}\left(\frac{I_{A_n}}{w_t^n}\right)(X_1^{(j)},\ldots,X_n^{(j)}) \\ &= \frac{1}{k}\sum_{j=1}^{k} I_{\{S_n^{(j)}\geq an\}}\cdot\exp\left(-tS_n^{(j)}+\Lambda(t)\cdot n\right)\end{aligned}$$

mit unabhängigen Zufallsvariablen $X_i^{(j)}$ mit Verteilung ν_t und $S_n^{(j)} = \sum_{i=1}^{n} X_i^{(j)}$. Wir können vermuten, dass auch diese Schätzer für große n nur für t nahe t^* effizient sind, da ansonsten das Ereignis A_n eine Wahrscheinlichkeit nahe 0 oder 1 bzgl. ν_t^n hat, und daher die überwiegende Mehrheit der Stichproben $S_n^{(j)}$ außerhalb bzw. in A_n liegt. Diese Vermutung lässt sich bestätigen. Auf ähnliche Weise wie beim Beweis des Satzes von Cramér erhalten wir:

Lemma 12.8. *Der Schätzer $\widetilde{\theta}_n^{(k)}$ ist für jedes $t \in \mathbb{R}$ und $k, n \in \mathbb{N}$ erwartungstreu. Für die Varianz gelten folgende Abschätzungen:*

$$\operatorname{Var}[\widetilde{\theta}_n^{(k)}] \quad \leq \quad \frac{1}{k}e^{-2n\cdot(at-\Lambda(t))}, \tag{12.2.9}$$

$$\liminf_{n\to\infty}\frac{\log\operatorname{Var}[\widetilde{\theta}_n^{(k)}]}{\log\theta_n^2} \quad \geq \quad \frac{at-\Lambda(t)}{I(a)}. \tag{12.2.10}$$

Bemerkung. Die zweite Aussage sieht auf den ersten Blick wie eine untere Schranke aus. Tatsächlich handelt es sich aber um eine Abschätzung der Varianz nach oben, da der Nenner $\log\theta_n^2$ negativ ist.

Beweis. Für die Varianz erhalten wir ähnlich wie beim Beweis der oberen Schranke im Satz von Cramér:

$$\begin{aligned}k\cdot\operatorname{Var}[\widetilde{\theta}_n^{(k)}] &= \operatorname{Var}_{\nu_n}[I_{A_n}/w_t^n] \quad \leq \quad \int_{A_n}(w_t^n)^{-2}\,d\nu_t^n \\ &= \int_{A_n}\exp\left(-2t\sum_{i=1}^{n}U(x_i)+2\Lambda(t)\cdot n\right)\,\nu_t^n(dx) \\ &\leq \exp(-2\cdot(ta-\Lambda(t))n),\end{aligned}$$

wobei $w_t^n = d\nu_t^n/d\mu^n$ die relative Dichte ist. In Kombination mit der unteren Schranke aus dem Satz von Cramér folgt

$$\begin{aligned}\liminf_{n\to\infty} \frac{\log \mathrm{Var}[\widetilde{\theta}_n^{(k)}]}{\log \theta_n^2} &= \frac{1}{2} \liminf_{n\to\infty} \frac{-n^{-1} \log \mathrm{Var}[\widetilde{\theta}_n^{(k)}]}{-n^{-1} \log \theta_n} \\ &\geq \frac{1}{2} \cdot \frac{-\limsup n^{-1} \log \mathrm{Var}[\widetilde{\theta}_n^{(k)}]}{-\liminf n^{-1} \log \theta_n} \\ &\geq \frac{at - \Lambda(t)}{I(a)}.\end{aligned}$$

Hierbei haben wir die Vorzeichen eingefügt, da $\log \theta_n$ negativ ist. □

Aus dem Lemma ergibt sich:

Satz 12.9 (Logarithmische Effizienz). *Gilt $t = t^*$, dann ist die Folge $(\widetilde{\theta}_n^{(k)})_{n\in\mathbb{N}}$ von Schätzern für die Wahrscheinlichkeiten θ_n für jedes $k \in \mathbb{N}$ logarithmisch effizient, d.h. für jedes $\varepsilon > 0$ gilt*

$$\limsup_{n\to\infty} \frac{E[|\widetilde{\theta}_n^{(k)} - \theta_n|^2]}{\theta_n^{2-\varepsilon}} < \infty.$$

Beweis. Die Funktion $f(t) = ta - \Lambda(t)$ hat ein globales Maximum bei $t^* = m^{-1}(a)$, denn es gilt

$$\begin{aligned} f'(t) &= a - \Lambda'(t) = a - m(t) = 0 \qquad \text{für } t = t^*, \text{ und} \\ f''(t) &= -\Lambda''(t) = -\mathrm{Var}_{\nu_t}[U] < 0 \qquad \text{für alle } t \in \mathbb{R}. \end{aligned}$$

Also gilt

$$I(a) = \sup_{t\in\mathbb{R}}(ta - \Lambda(t)) = t^*a - \Lambda(t^*).$$

Für $t = t^*$ folgt dann aus Lemma 12.8

$$\limsup_{k\to\infty} \frac{\log \mathrm{Var}[\widetilde{\theta}_k^{(n)}]}{-\log \theta_k^2} \leq -1,$$

d.h. zu jedem $\varepsilon > 0$ existiert ein $n_0 \in \mathbb{N}$ mit

$$\log \mathrm{Var}[\widetilde{\theta}_n^{(k)}] \leq -(-1+\varepsilon) \log \theta_n^2,$$

bzw.

$$\mathrm{Var}[\widetilde{\theta}_n^{(k)}] \leq \theta_n^{2+2\varepsilon} \qquad \text{für alle } n \geq n_0.$$

□

Umgekehrt kann man zeigen, dass bei anderer Wahl von t keine logarithmische Effizienz vorliegt. Dies rechtfertigt die zunächst anschaulich motivierte Wahl von t als Schwellenwert $t^* = m^{-1}(a)$.

12.3 Relative Entropie und statistische Unterscheidbarkeit

In diesem Abschnitt werden wir den Wechsel des zugrundeliegenden Wahrscheinlichkeitsmaßes systematischer untersuchen. Dabei spielt der Begriff der relativen Entropie eine zentrale Rolle.

Relative Entropie

Seien μ und ν Wahrscheinlichkeitsverteilungen auf $S = \mathbb{R}^d$ oder einem diskreten Raums mit Dichten (bzw. Massenfunktionen) $f, g > 0$. Die relative Dichte w von ν bzgl. μ ist

$$w(x) \quad := \quad \frac{d\nu}{d\mu}(x) \quad = \quad \frac{g(x)}{f(x)} \qquad \text{für } \mu\text{-fast alle } x \in S.$$

Die Dichte bzw. Massenfunktion

$$L_n(\mu; x_1, \ldots, x_n) \quad = \quad \prod_{i=1}^{n} f(x_i)$$

der Verteilung n unabhängiger Stichproben $X_1, \ldots, X_n$ von μ bezeichnet man auch als **Likelihood** der Verteilung μ bzgl. der Daten $(x_1, \ldots, x_n)$.

Wie kann man anhand von unabhängigen Stichproben erkennen, welche der beiden Verteilungen μ und ν in einem Zufallsexperiment vorliegt? Dazu betrachten wir den **Likelihoodquotienten**

$$w_n(x_1, \ldots, x_n) \quad := \quad \frac{L_n(\nu; x_1, \ldots, x_n)}{L_n(\mu; x_1, \ldots, x_n)} \quad = \quad \frac{\prod_{i=1}^{n} g(x_i)}{\prod_{i=1}^{n} f(x_i)} \quad = \quad \prod_{i=1}^{n} w(x_i).$$

Definition. *Die durch*

$$\begin{aligned} H(\nu \mid \mu) &= \int \log w \, d\nu = \int w \log w \, d\mu \quad \textit{falls } \nu \ll \mu \textit{ mit Dichte } w, \\ H(\nu \mid \mu) &= \infty \qquad \textit{sonst,} \end{aligned}$$

*definierte Größe $H(\nu \mid \mu) \in [0, \infty]$ heißt **relative Entropie** (oder **Kullback-Leibler Information**) von ν bzgl. μ.*

Um eine anschauliche Interpretation der relativen Entropie zu geben, bemerken wie, dass

$$H(\nu \mid \mu) \quad = \quad \int \log \frac{g}{f} \, d\nu \quad = \quad \int (-\log f(x) - (-\log g(x))) \, \nu(dx)$$

gilt. Wir können $-\log f(x)$ und $-\log g(x)$ als Maß für die Überraschung (den Informationsgewinn) bei Eintreten von x interpretieren, falls μ bzw. ν das zugrundeliegende Modell ist. Wenn

wir also μ als Modell annehmen, aber tatsächlich ν die zugrundeliegende Verteilung ist, dann erhöht sich die Überraschung (der Informationszuwachs) bei Ziehen einer Stichprobe im Vergleich zum korrekten Modell im Mittel um $H(\nu \mid \mu)$.

Satz 12.10 (Shannon-Mac Millan). *Seien $X_1, X_2, \ldots : \Omega \to S$ unabhängige Zufallsvariablen unter P_μ bzw. P_ν mit Verteilung μ bzw. ν. Dann gilt für $n \to \infty$:*

(1).
$$\frac{1}{n} \log w_n(X_1, \ldots, X_n) \quad \longrightarrow \quad H(\nu \mid \mu) \qquad P_\nu\text{-fast sicher.}$$

(2).
$$\frac{1}{n} \log w_n(X_1, \ldots, X_n) \quad \longrightarrow \quad -H(\mu \mid \nu) \qquad P_\mu\text{-fast sicher.}$$

Beweis. (1). Für $n \to \infty$ gilt nach dem Gesetz der großen Zahlen

$$\frac{1}{n} \log w_n(X_1, \ldots, X_n) \quad = \quad \frac{1}{n} \sum_{i=1}^{n} \log w(X_i) \quad \longrightarrow \quad \int \log w \, d\nu \qquad P_\nu\text{-fast sicher.}$$

Das Gesetz der großen Zahlen ist anwendbar, da

$$\int (\log w)^- \, d\nu \quad = \quad \int (w \log w)^- \, d\mu \quad \leq \quad \frac{1}{e} \quad < \quad \infty.$$

(2). Da μ absolutstetig bzgl. ν mit Dichte $1/w$ ist, gilt entsprechend

$$\begin{aligned} \frac{1}{n} \log w_n(X_1, \ldots, X_n) \quad &= \quad -\frac{1}{n} \sum_{i=1}^{n} \log \frac{1}{w(X_i)} \\ &\overset{\text{GdgZ}}{\longrightarrow} \quad -\int \log \frac{1}{w} \, d\mu \quad = \quad -H(\mu \mid \nu) \qquad P_\mu\text{-fast sicher.} \end{aligned}$$

□

Der Satz zeigt, dass sich die Produktdichte (der Likelihoodquotient) asymptotisch auf der exponentiellen Skala (d.h. unter Vernachlässigung subexponentiell wachsender Faktoren) folgendermaßen verhält:

$$w_n(X_1, \ldots, X_n) \quad \simeq \quad \begin{cases} e^{nH(\nu \mid \mu)} & P_\nu\text{-fast sicher} \\ e^{-nH(\mu \mid \nu)} & P_\mu\text{-fast sicher} \end{cases}.$$

Das folgende Lemma fasst einige elementare Eigenschaften der relativen Entropie zusammen:

Lemma 12.11 (Eigenschaften der relativen Entropie).

(1). Es gilt $H(\nu \mid \mu) \geq 0$ mit Gleichheit genau dann, wenn $\nu = \mu$.

(2). Ist μ die Gleichverteilung auf einer endlichen Menge S, dann gilt

$$H(\nu \mid \mu) \quad = \quad \log|S| - H(\nu). \tag{12.3.1}$$

(3). $H(\nu_1 \otimes \ldots \otimes \nu_n \mid \mu_1 \otimes \ldots \otimes \mu_n) \quad = \quad \sum_{i=1}^{n} H(\nu_i \mid \mu_i)$.

Beweis. (1). Aus der Jensenschen Ungleichung folgt

$$H(\nu \mid \mu) \quad = \quad \int w \log w \, d\mu \quad \geq \quad \int w \, d\mu \cdot \log \int w \, d\mu \quad = \quad 0.$$

Gleichheit gilt genau dann, wenn w μ-fast sicher konstant, also $\nu = \mu$ ist.

(2). In diesem Fall gilt $w(x) = \nu(x) \cdot |S|$, also

$$H(\nu \mid \mu) \quad = \quad \sum_{x \in S} \nu(x) \log(\nu(x) \cdot |S|) \quad = \quad \log|S| - H(\nu).$$

(3). Übung.

□

Nach (12.3.1) liefern Aussagen über die relative Entropie als Spezialfall entsprechende Aussagen für die Entropie.

Beispiel. (1). *Bernoulliverteilungen*: Für die Bernoulliverteilungen μ_p mit $\mu_p(1) = p$ und $\mu_p(0) = 1 - p$ gilt:

$$H(\mu_a \mid \mu_p) \quad = \quad a \log\left(\frac{a}{p}\right) + (1-a) \log\left(\frac{1-a}{1-p}\right) \qquad \text{für alle } a, p \in (0,1).$$

(2). *Normalverteilungen*: Für $m, \widetilde{m} \in \mathbb{R}$ und $v, \widetilde{v} > 0$ gilt:

$$\begin{aligned} H(N(\widetilde{m}, \widetilde{v}) \mid N(m,v)) &= \frac{1}{2}\left(\log\left(\frac{v}{\widetilde{v}}\right) + \frac{\widetilde{v}}{v} - 1 + \frac{(\widetilde{m}-m)^2}{v}\right), \qquad \text{also insbesondere} \\ H(N(\widetilde{m}, v) \mid N(m,v)) &= \frac{(\widetilde{m}-m)^2}{2v}. \end{aligned}$$

Die relative Entropie ist ein im Allgemeinen *nichtsymmetrischer Abstandsbegriff* für Wahrscheinlichkeitsverteilungen. Ihre statistische Interpretation werden wir im nächsten Abschnitt noch weiter präzisieren. Zuvor bemerken wir, dass die relative Entropie Aussagen über die Größe wesentlicher Mengen bei Wechsel der zugrundeliegenden Wahrscheinlichkeitsverteilung ermöglicht:

Maßwechsel und untere Schranken für große Abweichungen

Seien $X_1, X_2, \ldots$ unter P_μ bzw. P_ν unabhängige Zufallsvariablen mit Verteilung μ bzw. ν. Wie in Abschnitt 7.4 nennen wir eine Folge B_n von messbaren Teilmengen der Produkträume S^n wesentlich bzgl. ν, falls

$$P_\nu[(X_1, \ldots, X_n) \in B_n] \quad = \quad \nu^n[B_n] \quad \longrightarrow \quad 1 \qquad \text{für } n \to \infty.$$

Die folgende Aussage verallgemeinert den Maßkonzentrationssatz von MacMillan und den Quellenkodierungssatz von Shannon aus Abschnitt 7.4.

Korollar 12.12. *(1). Für jedes $\varepsilon > 0$ ist die Folge*

$$B_{n,\varepsilon} := \{(x_1, \ldots, x_n) \mid e^{n(H(\nu \mid \mu)-\varepsilon)} \leq w_n(x_1, \ldots, x_n) \leq e^{n(H(\nu \mid \mu)+\varepsilon)}\} \subseteq S^n$$

wesentlich bzgl. ν, und

$$\mu^n[B_{n,\varepsilon}] \quad \leq \quad e^{-n(H(\nu \mid \mu)-\varepsilon)} \qquad \textit{für alle } n \in \mathbb{N}. \tag{12.3.2}$$

(2). Für beliebige messbare Mengen $A_n \subseteq S^n$ mit

$$\liminf \nu^n[A_n] > 0 \tag{12.3.3}$$

gilt

$$\liminf \frac{1}{n} \log \mu^n[A_n] \geq -H(\nu \mid \mu). \tag{12.3.4}$$

Bemerkung. Der Maßkonzentrationssatz von MacMillan und der Quellenkodierungssatz von Shannon ergeben sich als Spezialfall von (1) bzw. (2), wenn S endlich und ν die Gleichverteilung ist.

Wir beweisen nun das Korollar.

Beweis. (1). Die Mengen $B_{n,\varepsilon}, n \in \mathbb{N}$, sind wesentlich bzgl. ν nach Satz 12.10. Zudem gilt:

$$1 \geq \nu^n[B_{n,\varepsilon}] \quad = \quad \int\limits_{B_{n,\varepsilon}} w_n \, d\mu^n \quad \geq \quad \mu^n[B_{n,\varepsilon}] \cdot e^{n(H(\nu \mid \mu)-\varepsilon)}.$$

(2). beweist man analog zum Quellenkodierungssatz (Satz 7.17): Aus

$$\mu^n[A_n] \quad = \quad \int\limits_{A_n} \frac{1}{w_n} \, d\nu_n \quad \geq \quad e^{-n(H(\nu \mid \mu)+\varepsilon)} \nu^n[A_n \cap B_{n,\varepsilon}]$$

folgt

$$\begin{aligned}\liminf \frac{1}{n} \log \mu^n[A_n] &\geq -(H(\nu \mid \mu) + \varepsilon) + \liminf \frac{1}{n} \log \nu^n[A_n \cap B_{n,\varepsilon}] \\ &= -(H(\nu \mid \mu) + \varepsilon),\end{aligned}$$

da $\liminf \nu^n[A_n \cap B_{n,\varepsilon}] = \liminf \nu^n[A_n] > 0$ nach (1) gilt. Die Behauptung folgt für $\varepsilon \to 0$.

□

Die zweite Aussage der Korollars können wir als eine allgemeine untere Schranke für große Abweichungen interpretieren: Ist $A_n \subseteq S^n$ eine Folge von Ereignissen, deren Wahrscheinlichkeit bzgl. μ^n gegen 0 geht, dann liefert uns (12.3.4) für jede Wahrscheinlichkeitsverteilung ν mit (12.3.3) eine asymptotische Schranke für die Wahrscheinlichkeiten

$$P_\mu[(X_1, \ldots, X_n) \in A_n] \quad = \quad \mu^n[A_n]$$

auf der exponentiellen Skala.

Als erste Anwendung betrachten wir nochmal die Situation aus dem Satz von Cramér: Sei $U : S \to \mathbb{R}$ eine messbare Funktion mit $\int e^{tU} \, d\mu < \infty$ für alle $t \in \mathbb{R}$, und sei

$$a \quad > \quad m \quad = \quad \int U \, d\mu.$$

Um aus (12.3.4) eine bestmögliche asymptotische untere Schranke für die Wahrscheinlichkeiten $\mu^n[A_n]$ der großen Abweichungen

$$A_n \quad = \quad \left\{(x_1, \ldots, x_n) \in S^n \, : \, \frac{1}{n} \sum_{i=1}^n U(x_i) \geq a\right\}$$

zu erhalten, müssen wir eine Wahrscheinlichkeitsverteilung ν finden, die die relative Entropie $H(\nu \mid \mu)$ unter allen Wahrscheinlichkeitsverteilungen ν mit (12.3.3) minimiert. Die Bedingung (12.3.3) ist aber genau dann erfüllt, wenn $\int U \, d\nu \geq a$ gilt, denn aus dem Gesetz der großen Zahlen und dem zentralen Grenzwertsatz folgt:

$$\lim_{n\to\infty} \nu^n \left[\frac{1}{n} \sum_{i=1}^n U(x_i) \, \geq \, a\right] \quad = \quad \begin{cases} 1 & \text{für } a < \int U \, d\nu \\ 1/2 & \text{für } a = \int U \, d\nu \\ 0 & \text{für } a > \int U \, d\nu \end{cases}. \tag{12.3.5}$$

Das sich ergebende Variationsproblem

$$\begin{aligned} H(\nu \mid \mu) \quad &= \quad \int w \log w \, d\mu \quad \stackrel{!}{=} \quad \min \\ \text{unter der Nebenbedingung } \int U \, d\nu \quad &= \quad \int Uw \, d\mu \quad \geq \quad a \end{aligned}$$

kann man formal durch Variationsrechnung lösen. Als eindeutige Lösung erhält man gerade die Verteilung ν_{t^*} aus der exponentiellen Familie

$$\nu_t(dx) \quad = \quad \frac{1}{Z(t)} \exp\left(tU(x)\right) \mu(dx), \qquad Z(t) \;=\; \int e^{tU}\, d\mu,$$

zum eindeutigen Schwellenwert t^* mit $\int U\, d\nu_{t^*} \;=\; a$:

Satz 12.13 (Variationsprinzip für die relative Entropie). *Sei $t \geq 0$ und $m(t) = \int U\, d\nu_t$. Dann minimiert das Maß ν_t die relative Entropie bzgl. μ unter allen Wahrscheinlichkeitsverteilungen ν mit $\int U\, d\nu \geq m(t)$:*

$$\begin{aligned} H(\nu_t \mid \mu) &= t \cdot m(t) - \log Z(t) \\ &= \min\{H(\nu \mid \mu) \,:\, \nu \text{ Wahrscheinlichkeitsmaß mit } \int U\, d\nu \geq m(t)\} \end{aligned} \tag{12.3.6}$$

Beweis. Sei ν eine Wahrscheinlichkeitsverteilung mit $H(\nu \mid \mu) < \infty$ und $\int U\, d\nu \geq m(t)$. Dann gilt $\nu \ll \mu$ und

$$\begin{aligned} H(\nu \mid \mu) &= \int \log \frac{d\nu}{d\mu} d\nu \quad = \quad \int \log \frac{d\nu}{d\nu_t} d\nu + \int \log \frac{d\nu_t}{d\mu} d\nu \\ &= H(\nu \mid \nu_t) + \left(t \int U\, d\nu - \log Z(t) \right) \\ &\geq tm(t) - \log Z(t). \end{aligned}$$

Für $\nu = \nu_t$ ergibt sich Gleichheit. □

Wir beweisen nun die untere Schranke aus dem Satz von Cramér zur Illustration noch einmal mithilfe von Korollar 12.12:
Für $\nu = \nu_{t^*}$ gilt $\int U\, d\nu = m(t^*) = a$, also nach 12.3.5 $\lim \nu^n[A_n] = \frac{1}{2}$. Damit erhalten wir nach Korollar 12.12(2) und (12.3.6) die untere Schranke

$$\liminf \frac{1}{n} \log \mu^n[A_n] \quad \geq \quad -H(\nu \mid \mu) \quad = \quad t^* \cdot m(t^*) - \log Z(t^*) \quad \geq \quad -I(a),$$

wobei I die Ratenfunktion aus Satz 12.7 ist.

Das beschriebene Vorgehen ergibt nicht nur die untere Schranke. Es demonstriert auch, dass der Maßwechsel über die exponentielle Familie sinnvoll ist, da er asymptotisch die bestmöglichen Abschätzungen liefert.

Große Abweichungen für empirische Verteilungen

Mithilfe von Korollar 12.12 können wir noch eine stärkere Form der unteren Schranke für große Abweichungen vom Gesetz der großen Zahlen herleiten. Seien dazu

$$L_n(\omega) \quad = \quad \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i(\omega)}, \qquad n \in \mathbb{N},$$

die empirischen Verteilungen einer Folge $(X_i)_{i\in\mathbb{N}}$ unabhängiger Zufallsvariablen mit Verteilung μ bzgl. P_μ. Aus dem Gesetz der großen Zahlen folgt die fast sichere schwache Konvergenz der empirischen Verteilungen

$$L_n(\omega) \quad \xrightarrow{\omega} \quad \mu \qquad \text{für } P_\mu\text{-fast alle } \omega. \tag{12.3.7}$$

Insbesondere konvergiert die Wahrscheinlichkeit $P_\mu[L_n \notin \mathcal{U}]$ für jede Umgebung $\mathcal{U}$ der Wahrscheinlichkeitsverteilung μ bzgl. der Topologie der schwachen Konvergenz gegen 0. Die Konvergenzgeschwindigkeit auf der exponentiellen Skala lässt sich durch ein Prinzip der großen Abweichungen auf dem Raum WV(S) der Wahrscheinlichkeitsverteilungen auf $(S, \mathcal{S})$ mit der Topologie der schwachen Konvergenz beschreiben:

Satz 12.14 (Sanov). *Die empirischen Verteilungen $L_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$ erfüllen das folgende Prinzip der großen Abweichungen:*

(1). Obere Schranke: Für jede abgeschlossene Menge $\mathcal{A} \subseteq$WV(S) gilt:

$$\limsup_{n\to\infty} \frac{1}{n} \log P_\mu[L_n \in \mathcal{A}] \quad \leq \quad -\inf_{\nu\in\mathcal{A}} H(\nu \mid \mu).$$

(2). Untere Schranke: Für jede offene Menge $\mathcal{O} \subseteq$WV(S) gilt:

$$\liminf_{n\to\infty} \frac{1}{n} \log P_\mu[L_n \in \mathcal{O}] \quad \geq \quad -\inf_{\nu\in\mathcal{O}} H(\nu \mid \mu).$$

Beweis. (2). Zum Beweis der unteren Schranke wechseln wir wieder das zugrundeliegende Maß, und wenden Korollar 12.12 an. Sei $\mathcal{O} \subseteq$ WV(S) offen und $\nu \in \mathcal{O}$. Nach (12.3.7) ist dann die Folge

$$A_n \quad = \quad \{(x_1, \ldots, x_n) \in S^n \mid \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i} \in \mathcal{O}\}$$

wesentlich bzgl. ν, denn

$$\nu^n[A_n] \quad = \quad P_\nu[L_n \in \mathcal{O}] \quad \longrightarrow \quad 1$$

für $n \to \infty$. Daher folgt nach Korollar 12.12(2):

$$\liminf_{n\to\infty} \frac{1}{n} \log P_\mu[L_n \in \mathcal{O}] \quad = \quad \liminf_{n\to\infty} \frac{1}{n} \log \mu^n[A_n] \quad \geq \quad -H(\nu \mid \mu).$$

Die Behauptung ergibt sich, da dies für alle $\nu \in \mathcal{O}$ gilt.

(1). Die obere Schranke beweisen wir hier nur für endliche Zustandsräume S, s. z.B. [Dembo und Zeitouni: Large Deviations] für den Beweis im allgemeinen Fall. Ist S endlich, und ν eine bzgl. μ absolutstetige Wahrscheinlichkeitsverteilung mit Dichte $w = d\nu/d\mu$, dann gilt für alle $(x_1, \ldots, x_n) \in S^n$ mit empirischer Verteilung $\frac{1}{n}\sum_{i=1}^n \delta_{x_i} = \nu$:

$$\begin{aligned} \frac{d\nu^n}{d\mu^n}(x_1, \ldots, x_n) &= \prod_{i=1}^n \frac{d\nu}{d\mu}(x_i) \quad = \quad \exp\left(\sum_{i=1}^n \log\left(\frac{d\nu}{d\mu}(x_i)\right)\right) \\ &= \exp\left(n \int \log\left(\frac{d\nu}{d\mu}\right) d\nu\right) \quad = \quad \exp(n \cdot H(\nu \mid \mu)). \end{aligned}$$

Damit folgt

$$\begin{aligned} P_\mu[L_n = \nu] &= \mu^n\left[\left\{(x_1, \ldots, x_n) \;\middle|\; \frac{1}{n}\sum_{i=1}^n \delta_{x_i} = \nu\right\}\right] \\ &= e^{-nH(\nu\mid\mu)} \cdot \nu^n\left[\left\{(x_1, \ldots, x_n) \;\middle|\; \frac{1}{n}\sum_{i=1}^n \delta_{x_i} = \nu\right\}\right] \qquad (12.3.8) \\ &\leq e^{-nH(\nu\mid\mu)}. \end{aligned}$$

Jeder empirischen Verteilung von n Elementen $x_1, \ldots, x_n \in S$ entspricht ein Histogramm $\vec{h} = (h_a)_{a\in S} \in \{0, 1, \ldots, n\}^S$. Für die Anzahl der möglichen empirischen Verteilungen gilt daher

$$\left|\left\{\frac{1}{n}\sum_{i=1}^n \delta_{x_i} \;\middle|\; (x_1, \ldots, x_n) \in S^n\right\}\right| \quad \leq \quad (n+1)^{|S|}.$$

Nach (12.3.8) erhalten wir nun für eine beliebige Menge $\mathcal{A} \subseteq \mathrm{WV}(S)$ die (nicht-asymptotische) Abschätzung

$$P_\mu[L_n \in \mathcal{A}] \quad = \quad \sum_{\nu\in\mathcal{A}} P_\mu[L_n = \nu] \quad \leq \quad (n+1)^{|S|} \cdot e^{-n \inf_{\nu\in\mathcal{A}} H(\nu\mid\mu)},$$

aus der die asymptotische obere Schranke wegen $|S| < \infty$ folgt.

□

Bemerkung. Wie der Beweis schon andeutet, gilt auch die obere Schranke in diesem Fall nur noch asymptotisch und modulo subexponentiell wachsender Faktoren. Der Übergang von endlichen zu allgemeinen Zustandsräumen ist bei der oberen Schranke nicht trivial, s. [Dembo/Zeitouni].

Den Satz von Sanov bezeichnet man gelegentlich auch als ein „Prinzip der großen Abweichungen auf Level II“, d.h. für die empirischen Verteilungen. Wir bemerken abschließend, dass sich eine Version des Satzes von Cramér, d.h. ein „Prinzip der großen Abweichungen auf Level I“ als Spezialfall ergibt:

Für $U : S \to \mathbb{R}$ und eine offene Menge $B \subseteq \mathbb{R}$ gilt nach dem Satz von Sanov:

$$\liminf_{n\to\infty} \frac{1}{n} \log P_\mu \left[\frac{1}{n} \sum_{i=1}^{n} U(X_i) \in B \right] \quad = \quad \liminf_{n\to\infty} \frac{1}{n} \log P_\mu[L_n \in \mathcal{O}] \quad \geq \quad - \inf_{\nu\in\mathcal{O}} H(\nu \mid \mu)$$

mit $\mathcal{O} = \{\nu \in \text{WV}(S) \mid \int U \, d\nu \in B\}$. Entsprechend ergibt sich eine analoge obere Schranke, falls B abgeschlossen ist.

12.4 Likelihood

Praktisch unterscheidet man Wahrscheinlichkeitsverteilungen in der Schätz- und Testtheorie durch Likelihood-basierte statistische Verfahren. Der Zusammenhang von relativer Entropie und statistischer Unterscheidbarkeit kann genutzt werden, um die Qualität dieser Verfahren asymptotisch zu beurteilen.

Konsistenz von Maximum-Likelihood-Schätzern

Sei $(\mu_\theta)_{\theta\in\Theta}$ eine Familie von Wahrscheinlichkeitsverteilungen auf $S = \mathbb{R}^d$ (oder einem diskreten Raum) mit Dichten (bzw. Massenfunktionen) f_θ wobei θ ein unbekannter Parameter ist. Ferner sei

$$L_n(\theta; x_1, ..., x_n) \; = \; \prod_{i=1}^{n} f_\theta(x_i), \qquad \theta \in \Theta,$$

die Likelihoodfunktion zu n unabhängigen Stichproben $x_1, ..., x_n$ von μ_θ. Ein wichtiges Ad-hoc-Verfahren zur Konstruktion eines Schätzers für θ ist das

Maximum-Likelihood-Prinzip: Wähle $\hat{\theta}(x_1, ..., x_n)$ als den Parameterwert θ, für den die Likelihood der beobachteten Werte $x_1, \ldots, x_n$ maximal ist.

Definition. *(1). Eine Zufallsvariable vom Typ $\hat{\theta}(X_1, \ldots, X_n)$, $\hat{\theta} : S^n \to \Theta$ messbar, heißt* **Statistik** *der Daten $X_1, \ldots, X_n$.*

*(2). Die Statistik heißt **Maximum-Likelihood-Schätzer (MLE)** für den Parameter θ, falls*

$$L_n(\hat{\theta}(x_1, ..., x_n);\, x_1, ..., x_n) \quad = \quad \max_{\theta\in\Theta} L_n(\theta;\, x_1, ..., x_n) \qquad \textit{für alle } x_1, \ldots, x_n \in S \textit{ gilt.}$$

Um einen Maximum-Likelihood-Schätzer zu berechnen, ist es oft günstig, die **log-Likelihood**

$$\theta \mapsto \log L_n(\theta;\, x_1, ..., x_n) = \sum_{i=1}^{n} \log f_\theta(x_i) \qquad \text{zu maximieren.}$$

Beispiel. (1). **Gaußmodell:** $\Theta = \{(m, v) \mid m \in \mathbb{R}, v > 0\},\ \mu_{m,v} = N(m, v)$.

$$L_n(m, v; X_1, ..., X_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi v}}\, e^{-\frac{(X_i - m)^2}{2v}}$$

ist maximal für $\hat{m}(X) = \overline{X}_n$, $\hat{v}(X) = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X}_n)^2$. Dieser Maximum-Likelihood-Schätzer ist **nicht** erwartungstreu, da die Stichprobenvarianz mit dem Faktor $\frac{1}{n}$ statt $\frac{1}{n-1}$ gebildet wird.

(2). **Doppelexponentialverteilung:** $\Theta = \mathbb{R}$, $f_\theta(X_i) = \frac{1}{2}\, e^{-|X_i - \theta|}$.

$$\log L_n(\theta; X_1, ..., X_n) = -n \log 2 - \sum_{i=1}^{n} |X_i - \theta|$$

ist maximal, falls $\hat{\theta}$ ein Median von $X_1, ..., X_n$ ist.

(3). **Zufallszahlen** aus $[0, \theta]$, $\theta > 0$ unbekannt.

$$\begin{aligned} f_\theta(X_i) &= \frac{1}{\theta} I_{[0,\theta]}(X_i), \\ L_n(\theta; X_1, ..., X_n) &= \frac{1}{\theta^n} I_{[0,\theta]}(\max_{1 \le i \le n} X_i). \end{aligned}$$

Der Maximum-Likelihood-Schätzer ist $\hat{\theta}(X_1, ..., X_n) = \max_{1 \le i \le n} X_i$. Dieser Schätzer ist sicher nicht optimal, da mit Wahrscheinlichkeit 1 $\theta > \hat{\theta}(X_1, \ldots, X_n)$ gilt !

Wie das letzte Beispiel zeigt, sind Maximum-Likelihood-Schätzer für ein festes n nicht immer optimal. Unter bestimmten Voraussetzungen haben sie aber gute asymptotische Eigenschaften für $n \to \infty$. Sei etwa μ_θ $(\theta \in \Theta)$ eine einparametrige (d.h. $\Theta \subseteq \mathbb{R}$) Familie von Wahrscheinlichkeitsverteilungen mit Dichten bzw. Massenfunktionen f_θ. Es gelte:

Annahme (Unimodalität): Für alle $n \in \mathbb{N}$ und $x \in S^n$ existiert ein $\hat{\theta}_n(x_1, ..., x_n)$, sodass

$$\theta \mapsto L_n(\theta; x_1, ..., x_n) \begin{cases} \text{ist monoton wachsend für } \theta \le \hat{\theta}_n(x_1, ..., x_n). \\ \text{ist monoton fallend für } \theta \ge \hat{\theta}_n(x_1, ..., x_n). \end{cases}$$

Bemerkung. (1). Die Annahme ist z.B. erfüllt, falls $\theta \mapsto \log f_\theta(x)$ für jedes x konkav ist - denn dann ist auch $\log L_n(\theta, x_1, ..., x_n) = \sum_{i=1}^{n} \log f_\theta(x_i)$ konkav in θ.

(2). $\hat{\theta}_n(X_1, ..., X_n)$ ist im unimodalen Fall eindeutiger Maximum-Likelihood-Schätzer für θ.

Satz 12.15. *Es gelte die Annahme, sowie $\mu_\theta \neq \mu_{\widetilde{\theta}}$ für $\theta \neq \widetilde{\theta}$. Dann ist $\hat{\theta}_n(X_1, \ldots, X_n)$ $(n \in \mathbb{N})$ eine **konsistente** Folge von Schätzern für θ, d.h. für jedes $\varepsilon > 0$ gilt:*

$$P_\theta[|\hat{\theta}_n(X_1, ..., X_n) - \theta| < \varepsilon] \to 1 \quad \textit{für } n \to \infty.$$

Beweis. Wegen der Unimodalität gilt $\hat{\theta}_n(x_1, ..., x_n) \in (\theta - \varepsilon, \theta + \varepsilon)$ falls

$$L_n(\theta; x_1, ..., x_n) \; > \; L_n(\theta \pm \varepsilon; x_1, ..., x_n).$$

Also:

$$P_\theta[|\hat{\theta}_n(X_1, ..., X_n) - \theta| < \varepsilon] \geq P_\theta \left[\frac{L_n(\theta; X_1, ..., X_n)}{L_n(\theta \pm \varepsilon; X_1, ..., X_n)} > 1 \right].$$

Die rechte Seite konvergiert aber für $n \to \infty$ nach Satz 12.10 für jedes θ gegen 1. □

Bemerkung (Asymptotische Normalität von Maximum-Likelihood-Schätzern). Unter geeigneten Regularitätsvoraussetzungen an die Dichten f_θ gilt für die Maximum-Likelihood-Schätzer neben der Konsistenz (also dem Gesetz der großen Zahlen) auch ein zentraler Grenzwertsatz:

Satz (Fisher, Wilkes, Wold). Unter geeigneten Voraussetzungen gilt:

$$\sqrt{n}(\hat{\theta}_n(X_1, ..., X_n) - \theta) \; \xrightarrow{\mathcal{D}} \; N\left(0, \frac{1}{I(\theta)}\right),$$

wobei

$$I(\theta) \; = \; \int \left| \frac{\partial}{\partial \theta} \; \log f_\theta(x) \right|^2 \mu_\theta(dx) \; = \; \lim_{\varepsilon \to 0} \; \frac{2}{\varepsilon^2} \; H(\mu_{\theta+\varepsilon} | \; \mu_\theta)$$

die **Fisher-Information** des statistischen Modells ist.

Da man andererseits unter geeigneten Regularitätsbedingungen zeigen kann, daß die Varianz eines erwartungstreuen Schätzers für θ basierend auf n unabhängigen Stichproben stets größer als $\frac{1}{nI(\theta)}$ ist (*Informationsungleichung von Cramér-Rao*), folgt, daß Maximum-Likelihood-Schätzer in gewisser Hinsicht asymptotisch optimal sind.

Asymptotische Macht von Likelihoodquotiententests

Angenommen, wir haben n unabhängige Stichproben $X_1, ..., X_n$ von einer unbekannten Verteilung vorliegen und wir gehen davon aus, daß die zugrundeliegende Verteilung aus einer Familie μ_θ $(\theta \in \Theta)$ von Wahrscheinlichkeitsverteilungen kommt. Sei Θ_0 eine Teilmenge des Parameterbereichs. Wir wollen entscheiden zwischen der

$$\textit{Nullhypothese } H_0\text{:} \qquad \text{»}\theta \in \Theta_0\text{«}$$

und der

$$\textit{Alternative } H_1\text{:} \qquad \text{»}\theta \notin \Theta_0\text{«}$$

Ein **Hypothesentest** für ein solches Problem ist bestimmt durch eine messbare Teilmenge $C \subseteq S^n$ (den **Verwerfungsbereich**) mit zugehöriger Entscheidungsregel:

$$\text{akzeptiere } H_0 \iff (X_1, ..., X_n) \notin C.$$

Beispiel (t-Test). Seien $X_1, X_2, \ldots, X_n$ unabhängige Stichproben von einer Normalverteilung mit unbekanntem Parameter $(m, v) \in \Theta = \mathbb{R} \times \mathbb{R}^+$. Wir wollen testen, ob der Mittelwert der Verteilung einen bestimmten Wert m_0 hat:

$$\textit{Nullhypothese } H_0\text{:} \qquad \text{»}m = m_0\text{«}\,, \qquad \Theta_0 = \{m_0\} \times \mathbb{R}^+\,.$$

Ein solches Problem tritt z.B.in der Qualitätskontrolle auf, wenn man überprüfen möchte, ob ein Sollwert m_0 angenommen wird. Eine andere Anwendung ist der Vergleich zweier Verfahren, wobei X_i die Differenz der mit beiden Verfahren erhaltenen Messwerte ist. Die Nullhypothese mit $m_0 = 0$ besagt hier, daß kein signifikanter Unterschied zwischen den Verfahren besteht.

Im *t–Test* für obiges Testproblem wird die Nullhypothese akzeptiert, falls der Betrag der *Studentschen t-Statistik* unterhalb einer angemessen zu wählenden Konstanten c liegt, bzw. verworfen, falls

$$|T_{n-1}| \;=\; \left|\frac{\sqrt{n}\cdot(\overline{X}_n - m_0)}{\sqrt{V_n}}\right| \;>\; c$$

gilt.

Seien nun allgemein $X_1, X_2, \ldots$ unter P_θ unabhängige Zufallsvariablen mit Verteilung μ_θ. Bei einem Hypothesentest können zwei Arten von Fehlern auftreten:

Fehler 1. Art: H_0 wird verworfen, obwohl wahr. Wahrscheinlichkeit:

$$P_\theta[(X_1, ..., X_n) \in C] = \nu_\theta^n(C)\,, \qquad \theta \in \Theta_0.$$

Fehler 2. Art: H_0 wird akzeptiert, obwohl falsch. Wahrscheinlichkeit:

$$P_\theta[(X_1, ..., X_n) \notin C] = \mu_\theta^n(C^C) \,, \qquad \theta \in \Theta \setminus \Theta_0.$$

Obwohl das allgemeine Testproblem im Prinzip symmetrisch in H_0 und H_1 ist, interpretiert man beide Fehler i.a. unterschiedlich. Die Nullhypothese beschreibt in der Regel den Normalfall, die Alternative eine Abweichung oder einen zu beobachtenden Effekt. Da ein Test Kritiker überzeugen soll, sollte die Wahrscheinlichkeit für den Fehler 1. Art (Effekt prognostiziert, obgleich nicht vorhanden) unterhalb einer vorgegebenen (kleinen) Schranke α liegen. Die Wahrscheinlichkeit

$$\mu_\theta^n(C) \,, \qquad \theta \in \Theta \setminus \Theta_0 \,,$$

daß kein Fehler 2. Art auftritt, sollte unter dieser Voraussetzung möglichst groß sein.

Definition. *Die Funktion*

$$G(\theta) \;=\; P_\theta[(X_1, ..., X_n) \in C] \;=\; \mu_\theta^n(C)$$

heißt ***Gütefunktion*** *des Tests. Der Test hat* ***Niveau*** α*, falls*

$$G(\theta) \;\leq\; \alpha \quad \textit{für alle } \theta \in \Theta_0$$

gilt. Die Funktion $G(\theta)$ mit $\theta \in \Theta_1$ heißt ***Macht*** *des Tests.*

Beispiel. Der Studentsche t-Test hat Niveau α falls c ein $(1 - \frac{\alpha}{2})$-Quantil der Studentschen t-Verteilung mit $n - 1$ Freiheitsgraden ist.

Ein Ziel bei der Konstruktion eines Testverfahrens sollte es sein, die Machtfunktion bei vorgegebenem Niveau zu maximieren. Dies ist im Allgemeinen nicht simultan für alle Parameter $\theta \in \Theta \setminus \Theta_0$ möglich. Eine Ausnahme bildet der Fall einer einfachen Hypothese und Alternative, in dem ein optimaler Test existiert:

a) Einfache Hypothese und Alternative

Angenommen, wir wissen, daß die Stichproben von einer der beiden Verteilungen $\mu_0 := \nu$ und $\mu_1 := \mu$ stammen und wir wollen entscheiden zwischen der

$$\textit{Nullhypothese } H_0\text{:} \qquad \text{»}X_i \sim \nu\text{«}$$

und der

$$\textit{Alternative } H_1\text{:} \qquad \text{»}X_i \sim \mu\text{«}.$$

Ein solches Problem tritt in Anwendungen zwar selten auf, bildet aber einen ersten Schritt zum Verständnis allgemeinerer Testprobleme. Sei

$$\varrho_n(x_1, \ldots, x_n) \;=\; \frac{L_n(\mu; x_1, ..., x_n)}{L_n(\nu; x_1, ..., x_n)} \;=\; \prod_{i=1}^{n} \frac{f(x_i)}{g(x_i)}$$

der Quotient der Likelihoods der Stichproben $x_1, \ldots, x_n$ im Produktmodell. Hierbei sind f und g die Dichte bzw. Massenfunktion der Verteilungen μ und ν.

Definition. *Ein Test mit Entscheidungsregel*

$$\textit{Akzeptiere } H_0 \iff \varrho_n(X_1, ..., X_n) \leq c,$$

$c \in (0, \infty)$*, heißt* ***Likelihoodquotiententest****.*

Der Verwerfungsbereich eines Likelihoodquotiententests ist also $C = \{\varrho_n > c\}$, die Wahrscheinlichkeit für den Fehler 1. Art beträgt

$$\alpha := \nu^n(\varrho_n > c).$$

Satz 12.16 (Neyman-Pearson-Lemma). *Der Likelihoodquotiententest mit Parameter c ist der* ***beste Test zum Niveau*** *α, d.h. jeder Test mit*

$$\textit{Wahrscheinlichkeit (Fehler 1.Art)} \leq \alpha$$

hat eine kleinere Macht (d.h. eine höhere Wahrscheinlichkeit für den Fehler 2. Art).

Beweis. Sei $A \subseteq S^n$ der Verwerfungsbereich eines Tests mit $\nu^n(A) \leq \alpha$, und sei

$$\chi = I_C - I_A = I_{A^C} - I_{C^C}.$$

Zu zeigen ist:

$$0 \leq \mu^n(A^C) - \mu^n(C^C) = \int \chi \, d\mu^n.$$

Offensichtlich gilt $\chi \geq 0$ auf $C = \{\varrho_n > c\}$ und $\chi \leq 0$ auf $C^C = \{\varrho_n \leq c\}$, also $\chi \cdot (\varrho_n - c) \geq 0$. Durch Integration erhalten wir:

$$0 \leq \int \chi \cdot (\varrho_n - c) \, d\nu^n = \int \chi \, d\mu^n - c \cdot \int \chi \, d\nu^n \leq \int \chi \, d\mu^n,$$

da $\int \chi \, d\nu^n = \nu^n(C) - \nu^n(A) \geq 0$. □

Wie gut ist der Likelihoodquotiententest (also der beste Test zur Unterscheidung von ν und μ) asymptotisch für große n? Wir betrachten ein festes Niveau $\alpha \in (0,1)$, und wählen $c_n \in (0,\infty)$ $(n \in \mathbb{N})$ mit

$$\nu^n(\varrho_n > c_n) \leq \alpha \leq \nu^n(\varrho_n \geq c_n) \tag{12.4.1}$$

Satz 12.17 (Asymptotische Macht des Likelihoodquotiententests). *Es gilt:*

(i)
$$\frac{1}{n} \log c_n \longrightarrow -H(\nu|\mu) \qquad \textit{für } n \to \infty.$$

(ii)
$$\frac{1}{n} \log \mu^n(\varrho_n \leq c_n) \longrightarrow -H(\nu|\mu) \qquad \textit{für } n \to \infty,$$

d.h. die Wahrscheinlichkeit für den Fehler 2. Art fällt exponentiell mit Rate $H(\nu|\mu)$.

Beweis. (i) Sei $\varepsilon > 0$. Für große n gilt nach dem Satz von Shannon-McMillan:

$$\nu^n\left(\varrho_n > e^{-n(H(\nu|\mu)+\varepsilon)}\right) > \alpha \overset{12.4.1}{\geq} \nu^n(\varrho_n > c_n).$$

Es folgt $e^{-n(H(\nu|\mu)+\varepsilon)} < c_n$. Analog zeigt man $e^{-n(H(\nu|\mu)-\varepsilon)} > c_n$. Die Behauptung folgt dann für $\varepsilon \to 0$.

(ii) a) *Untere Schranke:* Wegen

$$\nu^n(\varrho_n \leq c_n) \geq 1 - \alpha > 0 \qquad \forall\, n \in \mathbb{N}$$

folgt nach Korollar 12.12:

$$\underline{\lim} \frac{1}{n} \log \mu^n(\varrho_n \leq c_n) \geq -H(\nu|\mu).$$

Obere Schranke: Wegen

$$\mu^n(\varrho_n \leq c_n) = \int_{\varrho_n \leq c_n} \varrho_n \, d\nu^n \leq c_n$$

folgt nach (i)

$$\overline{\lim} \frac{1}{n} \log \mu^n(\varrho_n \leq c_n) \leq \overline{\lim} \frac{1}{n} \log c_n = -H(\nu|\mu).$$

□

Der Satz demonstriert erneut, daß die relative Entropie ein gutes Maß für die Unterscheidbarkeit zweier Wahrscheinlichkeitsverteilungen ist.

b) Zusammengesetzte Hypothesen und/oder Alternativen

Wenn Θ_0 und/oder Θ_1 aus mehr als einem Element bestehen, kann man den **verallgemeinerten Likelihoodquotienten**

$$\bar{\varrho}_n(x_1, \ldots, x_n) \;=\; \frac{\sup_{\theta\in\Theta_1} L_n(\theta; x_1, ..., x_n)}{\sup_{\theta\in\Theta_0} L_n(\theta; x_1, ..., x_n)} \;=\; \frac{\text{max. Lik. von } x\text{, falls } H_1 \text{ wahr}}{\text{max. Lik. von } x\text{, falls } H_0 \text{ wahr}}$$

betrachten. Der entsprechende Likelihoodquotiententest ist ähnlich wie der Maximum-Likelihood-Schätzer ein häufig verwendetes ad hoc Verfahren. Im Gegensatz zum Fall einer einfachen Hypothese und Alternative ist der verallgemeinerte Likelihoodquotiententest allerdings nicht immer optimal.

Beispiel. Im Beispiel von oben ist der t-Test der Likelihoodquotiententest. Mit einem Neyman-Pearson-Argument kann man zeigen, daß er im Gaußschen Produktmodell der beste unverfälschte Test zu einem vorgegebenen Niveau α ist, d.h. der mächtigste Test mit

$$G(\theta) \leq \alpha \quad \forall\, \theta \in \Theta_0 \quad \text{und} \quad G(\theta) \geq \alpha \quad \forall\, \theta \in \Theta_1.$$

Auch in nicht-Gaußschen Modellen wird häufig der t-Test eingesetzt – eine partielle Rechtfertigung dafür liefert der zentrale Grenzwertsatz.

12.5 Bayessche Modelle und MCMC Verfahren

Stichwortverzeichnis