WAHRSCHEINLICHKEITSTHEORIE I und II

Wolfgang König

Vorlesungsskript

TU Berlin und WIAS Berlin

SS 2005 und WS 2005/06
überarbeitet im WS 2008/09, kleine Korrekturen im März 2012

Version vom März 2012

Inhaltsverzeichnis

Vorwort

Dies ist das Vorlesungsskript zu zwei vierstündigen einführenden Vorlesungen über Wahrscheinlichkeitstheorie, gehalten an der Universität Leipzig 2005/06 und 2008/09 sowie an der Technischen Universität Berlin 2012/13. Es werden die grundlegenden Begriffe motiviert und entwickelt, wie Wahrscheinlichkeitsräume, Zufallsgrößen, Erwartungswert und Varianz, bedingte Wahrscheinlichkeiten und Erwartungswerte, Unabhängigkeit, Konvergenzbegriffe, stochastische Prozesse und vieles mehr. Ein Steilkurs stellt die benötigten maßtheoretischen Konzepte und Hilfsmittel zusammen. Ferner werden viele der wichtigen Verteilungen und ihre Einsatzgebiete und Beziehungen unter einander vorgestellt. Auch fundamentale stochastische Prozesse wie die einfache Irrfahrt, der Poisson-Prozess, Markovketten, stationäre Prozesse und die Brown'sche Bewegung werden eingeführt und untersucht.

In den ersten Kapiteln beschränken wir uns auf diskrete Wahrscheinlichkeitsräume und lernen die wichtigen diskreten Verteilungen kennen. Sodann behandeln wir Wahrscheinlichkeiten mit Dichten und wichtige Beispiele. Anschließend bringen wir den Begriff des allgemeinen Wahrscheinlichkeitsraums unter Verwendung von Maßtheorie. Dazu tragen wir zunächst die grundlegenden Begriffe und Konzepte der Maßtheorie zusammen, um dann die allgemeine Wahrscheinlichkeitstheorie zu präsentieren und auszubauen, womit ungefähr der Stoff der ersten der beiden Vorlesungen endet.

In der zweiten Vorlesung behandeln wir Unabhängigkeit, bedingte Erwartungswerte, diverse Konvergenzbegriffe und ihre Beziehungen untereinander (einschließlich des Zentralen Grenzwertsatzes und des Satzes von Prohorov), Markovketten, allgemeine stochastische Prozesse in diskreter Zeit, stationäre Prozesse und den Ergodensatz sowie die Konstruktion der Brown'schen Bewegung mit Hilfe des Satzes von Donsker.

Das vorliegende Skript wurde natürlich aus mehreren verschiedenen Quellen gespeist, die man nicht mehr alle im Einzelnen zurück verfolgen kann. Die Lehrbücher, die den meisten Einfluss auf die Gestaltung des ersten Teils dieses Skriptes hatten, sind die viel benutzten klassischen Texte [Kr02] von U. Krengel und [Ge02] von H.-O. Georgii, ferner das Lehrbuch [Kl06] von A. Klenke. Es sind auch Hinweise meiner Kollegen N. Gantert und A. Klenke eingeflossen, denen ich hiermit dafür danke. In der Überarbeitung vom Wintersemester 2008/09 korrigierte ich ein paar Fehler, verbesserte stellenweise die Präsentation und fügte manche Erläuterungen und Übungsaufgaben hinzu. Im März 2012 trug ich etliche kleine Korrekturen ein, die mir dankenswerter Weise Albert Haase und Jonka Hähnel übermittelten. Frau Leili Riazy steuerte netterweise die Illustration zu Beispiel 5.2.6 bei.

Berlin, im März 2012

Kapitel 1

Diskrete Wahrscheinlichkeitsräume

In diesem Abschnitt führen wir die grundlegenden Begriffe der diskreten Wahrscheinlichkeitstheorie ein, das heißt der Theorie der Wahrscheinlichkeiten auf höchstens abzählbar unendlich großen Mengen. Wir geben eine Auswahl an Beispielen und sammeln Eigenschaften dieses Konzeptes.

1.1 Grundbegriffe

Wir beginnen mit einem einführenden Beispiel.

Beispiel 1.1.1. Wir würfeln mit zwei fairen[1] Würfeln. Wie groß ist die Wahrscheinlichkeit, dass die Augensumme nicht kleiner als zehn ist?

Es bieten sich (mindestens) zwei Möglichkeiten der Problemlösung an. Wir können zum Beispiel ansetzen, dass jedes Element aus $\Omega = \{1, 2, \dots, 6\}^2$ (das ist die Menge aller Zahlenpaare mit Koeffizienten zwischen 1 und 6) die selbe Wahrscheinlichkeit besitzt (also $1/36$), wobei wir ein solches Paar identifizieren als die Ergebnisse der beiden Würfel. (Wir behandeln also die beiden Würfel als unterscheidbar, obwohl davon sicher die gesuchte Wahrscheinlichkeit nicht abhängen wird.) Dann zählen wir die günstigen Elementarereignisse, also diejenigen Paare, die das gesuchte Ereignis realisieren. Wir kommen auf die sechs Paare $(4, 6), (5, 5), (5, 6), (6, 4), (6, 5)$ und $(6, 6)$. Also antworten wir, dass die gesuchte Wahrscheinlichkeit gleich $6/36$ sei, also $1/6$.

Eine zweite Möglichkeit ist, die Menge $\Omega = \{2, 3, 4, \dots, 12\}$ von möglichen Augensummen zu betrachten. Allerdings müssen wir beachten, dass diese elf Elementarereignisse nicht gleich wahrscheinlich sind. Zum Beispiel hat die 2 die Wahrscheinlichkeit $1/36$, und die 3 hat die Wahrscheinlichkeit $1/18$. Nun müssen wir also die Wahrscheinlichkeiten der 10, 11 und 12 ermitteln und sie addieren, um die gesuchte Wahrscheinlichkeit zu erhalten. Die 10 hat die Wahrscheinlichkeit $1/12$, denn es gibt drei Möglichkeiten, die Summe 10 zu würfeln. Die 11 und die 12 haben die Wahrscheinlichkeiten $2/36$ bzw. $1/36$. Eine Addition ergibt, dass die gesuchte Wahrscheinlichkeit gleich $1/6$ beträgt, also das selbe Ergebnis wie oben. $\diamond$

Dieses Beispiel macht mehrere Dinge deutlich:

1. Es empfiehlt sich, eine Grundmenge von Elementarereignissen zu definieren, deren Wahrscheinlichkeiten einzeln bestimmt werden. Letzteres ist besonders einfach, wenn alle die

[1] Mit einem ‘fairen’ Würfel meinen wir einen idealisierten, der jedes der sechs Ergebnisse mit der selben Wahrscheinlichkeit $\frac{1}{6}$ ausgibt.

gleiche Wahrscheinlichkeit besitzen.

2. Das Ereignis, nach dem gefragt ist, identifiziert man mit einer Teilmenge der Grundmenge. Ihre Wahrscheinlichkeit ist die Summe ihrer Einzelwahrscheinlichkeiten.

3. Es gibt im Allgemeinen mehrere Möglichkeiten, eine Grundmenge zu wählen.

Das zu Grunde liegende Konzept fassen wir wie folgt zusammen. Mit $\mathcal{P}(\Omega)$ bezeichnen wir die *Potenzmenge* einer Menge Ω, also die Menge aller Teilmengen von Ω.

Definition 1.1.2 (diskreter Wahrscheinlichkeitsraum). *Ein* diskreter Wahrscheinlichkeitsraum *ist ein Tupel (Ω, p), bestehend aus einer endlichen oder höchstens abzählbar unendlichen Menge Ω und einer Abbildung $p\colon \Omega \to [0,1]$ mit der Eigenschaft*

$$\sum_{\omega\in\Omega} p(\omega) = 1.$$

Wir nennen Ω den Ereignisraum*, seine Elemente die* Elementarereignisse*, seine Teilmengen die* Ereignisse *und die $p(\omega)$ die* Einzelwahrscheinlichkeiten*.*

Die Abbildung $\mathbb{P}\colon \mathcal{P}(\Omega) \to [0,1]$, definiert durch

$$\mathbb{P}(A) = \sum_{\omega\in A} p(\omega) \qquad \textit{für alle } A \subset \Omega,$$

heißt das von den Einzelwahrscheinlichkeiten induzierte Wahrscheinlichkeitsmaß*. Man spricht auch von einer* Verteilung $\mathbb{P}$ *auf Ω.*

Da alle Einzelwahrscheinlichkeiten nicht negativ sind, spielt in dem Ausdruck $\sum_{\omega\in\Omega} p(\omega)$ die Reihenfolge der Summanden keine Rolle. Genau genommen handelt es sich um den Grenzwert $\lim_{n\to\infty} \sum_{i=1}^n p(\omega_i)$, wobei $\omega_1, \omega_2, \dots$ eine Abzählung von Ω ist.

Der ersten Lösung im Beispiel 1.1.1 lag also der Wahrscheinlichkeitsraum (Ω, p) mit $\Omega = \{1, 2, \dots, 6\}^2$ und $p(\omega) = \frac{1}{36}$ für jedes $\omega \in \Omega$ zu Grunde.

Der Begriff des Wahrscheinlichkeitsraums in Definition 1.1.2 ist ein Spezialfall eines allgemeinen und fundamentalen Konzepts, das erst in Kapitel 7 behandelt werden wird. Die fundamentalen Eigenschaften des Wahrscheinlichkeitsmaßes $\mathbb{P}$ in Definition 1.1.2 sind die folgenden.

Bemerkung 1.1.3 (Kolmogorov'sche Axiome). Sei (Ω, p) ein diskreter Wahrscheinlichkeitsraum. Das Wahrscheinlichkeitsmaß $\mathbb{P}$ hat die beiden Eigenschaften

(i) $\mathbb{P}(\Omega) = 1$ (Normierung),

(ii) für alle Folgen $(A_i)_{i\in\mathbb{N}}$ von paarweise disjunkten[2] Ereignissen gilt

$$\mathbb{P}\Big(\bigcup_{i\in\mathbb{N}} A_i\Big) = \sum_{i\in\mathbb{N}} \mathbb{P}(A_i) \qquad \text{(abzählbare Additivität).}$$

Die Gültigkeit dieser beiden Eigenschaften ist evident. ◇

[2] Wir nennen Ereignisse A_i mit $i \in I$ *paarweise disjunkt*, wenn $A_i \cap A_j = \emptyset$ für alle $i, j \in I$ mit $i \neq j$.

Diese sogenannten Kolmogorovschen Axiome sind für uns also Folgerungen aus Definition 1.1.2. In allgemeinerem Rahmen (siehe Kapitel 7) fordert man sie üblicherweise als Axiome, muss aber zunächst definieren, was Ereignisse sein sollen. Es ist leicht zu sehen, dass jede endliche oder höchstens abzählbare Menge Ω, auf deren Potenzmenge eine Abbildung $\mathbb{P}$ definiert ist, die den Kolmogorovschen Axiomen genügt, durch die Definition

$$p(\omega) = \mathbb{P}(\{\omega\}) \qquad \text{für } \omega \in \Omega$$

zu einem diskreten Wahrscheinlichkeitsraum (Ω, p) gemacht wird.

Wir sprechen also von Teilmengen von Ω als von Ereignissen. Wir listen die wichtigsten gängigen Sprechweisen für Ereignisse und ihre Entsprechung in der mathematischen Mengenschreibweise auf:

Sprache der Ereignisse	**Mengenschreibweise**
A, B und C sind Ereignisse	$A, B, C \subset \Omega$
A und B	$A \cap B$
A oder B	$A \cup B$
nicht A	$A^{\mathrm{c}} = \Omega \setminus A$
A und B schließen sich aus	$A \cap B = \emptyset$
A impliziert B	$A \subset B$

Wahrscheinlichkeiten genügen einigen einfachen Regeln:

Lemma 1.1.4. *Sei (Ω, p) ein diskreter Wahrscheinlichkeitsraum, dann hat das zugehörige Wahrscheinlichkeitsmaß $\mathbb{P}$ die folgenden Eigenschaften:*

(a) $\mathbb{P}(\emptyset) = 0$,

(b) $A \subset B \Longrightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$ *für alle Ereignisse A und B,*

(c) $\mathbb{P}(A \cup B) + \mathbb{P}(A \cap B) = \mathbb{P}(A) + \mathbb{P}(B)$ *für alle Ereignisse A und B,*

(d) $\mathbb{P}(\bigcup_{i\in\mathbb{N}} A_i) \leq \sum_{i\in\mathbb{N}} \mathbb{P}(A_i)$ *für alle Folgen $(A_i)_{i\in\mathbb{N}}$ von Ereignissen,*

(e) Falls für eine Folge $(A_i)_{i\in\mathbb{N}}$ von Ereignissen und ein Ereignis A gilt:[3] *$A_i \downarrow A$ oder $A_i \uparrow A$, so gilt* $\lim_{i\to\infty} \mathbb{P}(A_i) = \mathbb{P}(A)$.

Beweis. Die Beweise von (a) - (d) sind Übungsaufgaben. Der Beweis von (e) geht wie folgt. Falls $A_i \uparrow A$, so ist A gleich der disjunkten Vereinigung der Mengen $A_i \setminus A_{i-1}$ mit $i \in \mathbb{N}$ (wobei wir $A_0 = \emptyset$ setzen), und daher gilt nach Bemerkung 1.1.3:

$$\mathbb{P}(A) = \mathbb{P}\Big(\bigcup_{i=1}^{\infty}(A_i \setminus A_{i-1})\Big) = \sum_{i=1}^{\infty} \mathbb{P}(A_i \setminus A_{i-1}) = \lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{P}(A_i \setminus A_{i-1}) = \lim_{n\to\infty} \mathbb{P}(A_n).$$

Der Beweis der Aussage unter der Voraussetzung $A_i \downarrow A$ ist eine Übungsaufgabe. □

Beispiel 1.1.5. Wir betrachten ein Kartenspiel mit $2n$ Karten, darunter zwei Jokern. Wir bilden zwei gleich große Stapel. Wir groß ist die Wahrscheinlichkeit, dass sich beide Joker im selben Stapel befinden?

[3] Wir schreiben $A_i \downarrow A$ für $i \to \infty$, falls $A_{i+1} \subset A_i$ für jedes i und $\bigcap_{i\in\mathbb{N}} A_i = A$. Analog ist $A_i \uparrow A$ definiert.

Wir benutzen den Wahrscheinlichkeitsraum (Ω, p) mit

$$\Omega = \big\{(i,j) \in \{1,2,\dots,2n\}^2 \colon i \neq j\big\} \qquad \text{und} \qquad p((i,j)) = \frac{1}{|\Omega|} = \frac{1}{2n(2n-1)}.$$

Wir interpretieren i und j als die Plätze der beiden Joker im Kartenspiel. Die Plätze 1 bis n gehören zum ersten Stapel, die anderen zum zweiten. Also betrachten wir das Ereignis

$$A = \big\{(i,j) \in \Omega \colon i,j \leq n\big\} \cup \big\{(i,j) \in \Omega \colon i,j \geq n+1\big\}.$$

Man sieht leicht, dass A genau $2n(n-1)$ Elemente besitzt. Also ist $\mathbb{P}(A) = \frac{n-1}{2n-1}$. $\diamond$

1.2 Urnenmodelle

Wie aus Beispiel 1.1.1 klar wurde, ist eine der einfachsten Verteilungen durch die *Gleichverteilung* (auch *Laplace-Verteilung*) auf einer endlichen Menge Ω gegeben. Hier gilt $p(\omega) = 1/|\Omega|$ für jedes $\omega \in \Omega$, wobei $|\Omega|$ die *Kardinalität* von Ω bezeichnet, also die Anzahl der Elemente in Ω. Im Folgenden geben wir eine Liste von wichtigen Beispielen von Laplace-Räumen, die von Urnenmodellen her rühren. Dabei werden einige sehr wichtige Formeln der elementaren Kombinatorik motiviert und hergeleitet.

Beispiel 1.2.1 (Urnenmodelle)**.** Es sei eine Urne gegeben, in der N Kugeln mit den Aufschriften $1, 2, \dots, N$ liegen. Wir ziehen nun n Kugeln aus der Urne. Es stellt sich die Frage, wieviele unterschiedliche Ergebnisse wir dabei erhalten können. Ein Ergebnis ist hierbei ein Tupel $(k_1, \dots, k_n)$ in $\{1, \dots, N\}^n$, wobei k_i dafür stehen soll, dass in der i-ten Ziehung die Kugel mit der Nummer k_i gezogen wird.

Die Anwort auf diese Frage hängt stark davon ab, ob wir die einzelnen Ziehungen mit oder ohne zwischenzeitlichem Zurücklegen durchführen und ob wir Ergebnisse als unterschiedlich ansehen wollen, wenn sie sich nur in der Reihenfolge unterscheiden. Mit $\mathcal{M}$ bezeichnen wir die Menge $\{1, \dots, N\}$.

(I) mit Zurücklegen, mit Reihenfolge. Wir legen also nach jeder Ziehung die gezogene Kugel jeweils wieder in die Urne zurück, und wir betrachten die Ziehung unterschiedlicher Kugeln in unterschiedlicher Reihenfolge als unterschiedlich. Als Modell bietet sich die Menge

$$\Omega_I = \mathcal{M}^n = \{(k_1, \dots, k_n) \colon k_1, \dots, k_n \in \mathcal{M}\},$$

die Menge aller n-Tupel mit Koeffizienten aus $\mathcal{M}$, an. Es ist klar, dass Ω_I genau N^n Elemente besitzt.

(II) ohne Zurücklegen, mit Reihenfolge. Hier legen wir zwischendurch keine gezogenen Kugeln zurück, insbesondere müssen wir voraus setzen, dass $n \leq N$. Ein geeignetes Modell ist

$$\Omega_{II} = \{(k_1, \dots, k_n) \in \mathcal{M}^n \colon k_1, \dots, k_n \in \mathcal{M} \text{ paarweise verschieden}\}.$$

Die Anzahl der Elemente von Ω_{II} ist leicht aus der folgenden Argumentation als

$$|\Omega_{II}| = N(N-1)(N-2) \cdot \dots \cdot (N-n+1) = \frac{N!}{(N-n)!}$$

zu ermitteln:[4] Bei der ersten Ziehung haben wir N Kugeln zur Auswahl, bei der zweiten nur noch $N-1$ und so weiter. Jede dieser insgesamt n Ziehungen führt zu einem anderen n-Tupel.

(III) ohne Zurücklegen, ohne Reihenfolge. Hier legen wir keine gezogenen Kugeln zurück, und uns interessiert zum Schluss nicht, in welcher Reihenfolge wir die n Kugeln gezogen haben. Also empfiehlt sich die Grundmenge

$$\Omega_{III} = \{A \subset \mathcal{M} \colon |A| = n\} = \{\{k_1, \dots, k_n\} \subset \mathcal{M} \colon k_1, \dots, k_n \text{ paarweise verschieden}\},$$

die Menge der n-elementigen Teilmengen von $\mathcal{M}$. Die Kardinalität von Ω_{III} kann man ermitteln, indem man sich überlegt, dass die Menge Ω_{II} alle Teilmengen aus Ω_{III} genau $n!$ Mal auflistet, nämlich als Tupel in sämtlichen möglichen Reihenfolgen. Also gilt

$$|\Omega_{III}| = \frac{|\Omega_{II}|}{n!} = \frac{N!}{(N-n)!n!} = \binom{N}{n}.$$

Den Ausdruck $\binom{N}{n}$ liest man 'N über n', er wird *Binomialkoeffizient* genannt. Eine seiner wichtigsten Bedeutungen ist, dass er die Anzahl der n-elementigen Teilmengen einer N-elementigen Menge angibt.

(IV) mit Zurücklegen, ohne Reihenfolge. Wir legen also die jeweils gezogene Kugel nach jeder Ziehung zurück, und am Ende ordnen wir die Aufschriften der gezogenen Kugeln und zählen, wie oft wir welche gezogen haben. In einer gewissen Abwandlung bedeutet also k_i nicht mehr die Aufschrift der Kugel, die wir als i-te gezogen haben, sondern die i-t kleinste Aufschrift, die wir gezogen haben. Dies legt die Grundmenge

$$\Omega_{IV} = \{(k_1, \dots, k_n) \in \mathcal{M}^n \colon k_1 \leq k_2 \leq \dots \leq k_n\},$$

nahe, die Menge der n-Tupel in nichtabsteigender Reihenfolge. Zur Ermittlung der Kardinalität verwenden wir einen kleinen Trick. Wir betrachten die Abbildung $(k_1, \dots, k_n) \mapsto (k'_1, \dots, k'_n)$ mit $k'_i = k_i + i - 1$. Diese Abbildung ist bijektiv von Ω_{IV} in die Menge

$$\Omega = \big\{(k'_1, \dots, k'_n) \in (\mathcal{M}')^n \colon k'_1 < k'_2 < \dots < k'_n\big\},$$

wobei $\mathcal{M}' = \{1, \dots, N+n-1\}$. Offensichtlich besitzt diese Menge Ω die selbe Kardinalität wie die oben behandelte Menge Ω'_{III}, wobei Ω'_{III} allerdings nicht mit $\mathcal{M} = \{1, \dots, N\}$ gebildet wird, sondern mit $\mathcal{M}'$. Nun können wir die obige Formel für die Kardinalität von Ω_{III} benutzen, nachdem wir N durch $N+n-1$ ersetzt haben, und wir erhalten

$$|\Omega_{IV}| = |\Omega| = |\Omega'_{III}| = \binom{N+n-1}{n}.$$

◇

Es stellt sich heraus, dass eine Fülle von weiteren Modellen auf die obigen Urnenmodelle zurückgeführt und mit Hilfe der Formeln in Beispiel 1.2.1 behandelt werden können.

Beispiel 1.2.2. Wir würfeln mit vier Würfeln. Wie groß ist die Wahrscheinlichkeit dafür, dass wir vier verschiedene Augenzahlen erhalten? (Übungsaufgabe) ◇

[4] Mit $n! = n \cdot (n-1) \cdot (n-2) \cdot \dots \cdot 2 \cdot 1$ (lies 'n Fakultät') bezeichnet man das Produkt der ersten n natürlichen Zahlen. Man sieht leicht, dass es genau $n!$ verschiedene Permutationen von n unterscheidbaren Objekten gibt.

Beispiel 1.2.3. Wie groß ist beim Zahlenlotto '6 aus 49' die Wahrscheinlichkeit für genau sechs bzw. für genau vier Richtige? (Übungsaufgabe) ◇

Beispiel 1.2.4 (Geburtstagszwillinge)**.** Wie groß ist die Wahrscheinlichkeit p, dass unter $n \in \mathbb{N}$ Personen keine zwei Personen am selben Tag Geburtstag haben? (Natürlich setzen wir voraus, dass das Jahr 365 Tage hat, dass $n \leq 365$ gilt und dass alle Geburtstage unabhängig von einander die gleiche Wahrscheinlichkeit haben.)

Die Menge aller Geburtstagskombinationen der n Personen ist Ω_I mit $N = 365$ aus Beispiel 1.2.1, und die Menge von Kombinationen, die das gesuchte Ereignis realisieren, ist die Menge Ω_{II}. Also ist die Anwort, dass die gesuchte Wahrscheinlichkeit gleich

$$p = \frac{|\Omega_{II}|}{|\Omega_I|} = 1\Big(1 - \frac{1}{N}\Big)\Big(1 - \frac{2}{N}\Big)\dots\Big(1 - \frac{n-1}{N}\Big) = \exp\Big\{\sum_{i=1}^{n-1} \log\Big(1 - \frac{i}{N}\Big)\Big\}$$

ist. Bei $N = 365$ und $n = 25$ ist dieser Wert etwa gleich $0,432$. Für allgemeine n und N können wir eine Approximation dieses unhandlichen Ausdrucks vornehmen, wenn n sehr klein im Verhältnis zu N ist. In disem Fall benutzen wir die Näherung $\log(1 + x) \approx x$ für kleine $|x|$ und erhalten

$$p \approx \exp\Big\{-\sum_{i=1}^{n-1} \frac{i}{N}\Big\} = \exp\Big\{-\frac{n(n-1)}{2N}\Big\},$$

was sich viel leichter berechnen lässt. ◇

Beispiel 1.2.5. Wie viele Möglichkeiten gibt es, n ununterscheidbare Murmeln auf N Zellen zu verteilen?

Diese Frage führt auf die Frage der Kardinalität von Ω_{IV} in Beispiel 1.2.1, und wir benutzen die Gelegenheit, eine alternative Lösung anzubieten. Wir denken uns die n Murmeln in eine Reihe gelegt. Sie in N Zellen einzuteilen, ist äquivalent dazu, $N - 1$ Trennwände zwischen die n Murmeln zu setzen, denn dadurch werden ja die N Zellen definiert. (Links von der ersten Trennwand ist der Inhalt der ersten Zelle, zwischen der $(i-1)$-ten und i-ten Trennwand ist der Inhalt der i-ten Zelle, und rechts von der $(N-1)$-ten Trennwand ist der Inhalt der letzten Zelle.) Dadurch erhalten wir eine Reihe von $N + n - 1$ Objekten: n Murmeln und $N - 1$ Trennwände. Jede der $\binom{N+n-1}{n}$ möglichen Anordnungen (hier benutzen wir die Formel für Ω_{II}) korrespondiert mit genau einer Möglichkeit, die Murmeln in N Zellen einzuteilen. Also ist diese Anzahl gleich $\binom{N+n-1}{n}$. ◇

1.3 Weitere Beispiele von Verteilungen

Bei Ziehungen aus einer Urne, in der zwei verschiedene Typen von Kugeln liegen, bietet sich das folgende Modell an.

Beispiel 1.3.1 (Hypergeometrische Verteilung)**.** Aus einer Urne mit S schwarzen und W weißen Kugeln ziehen wir n Kugeln ohne Zurücklegen. Wir setzen dabei voraus, dass $S, W, n \in \mathbb{N}_0$ mit $n \leq S + W$. Wir nehmen wie immer an, dass die Kugeln in der Urne gut gemischt sind. Uns interessiert die Wahrscheinlichkeit, dass dabei genau s schwarze und $w = n - s$ weiße Kugeln gezogen wurden, wobei $s, w \in \mathbb{N}_0$ mit $s \leq S$ und $w \leq W$. Das heißt, dass s die Bedingungen

$\max\{0, n-W\} \leq s \leq \min\{S, n\}$ erfüllen muss. Die gesuchte Wahrscheinlichkeit ist gleich

$$\mathrm{Hyp}_{n,S,W}(s) = \frac{\binom{S}{s}\binom{W}{n-s}}{\binom{S+W}{n}}, \qquad s \in \big\{\max\{0, n-W\}, \dots, \min\{S, n\}\big\}.$$

Dies ergibt sich aus einer mehrfachen Anwendung der Formel für $|\Omega_{III}|$ in Beispiel 1.2.1: Der Nenner ist die Anzahl der Möglichkeiten, n Kugeln aus $S+W$ auszuwählen, und im Zähler steht die Anzahl der Möglichkeiten, s Kugeln aus S auszuwählen und $n-s$ aus W. Wir können die Menge $\Omega = \{\max\{0, n-W\}, \dots, \min\{S, n\}\}$ zusammen mit $\mathrm{Hyp}_{n,S,W}$ als einen Wahrscheinlichkeitsraum auffassen. Der analytische Nachweis, dass $\mathrm{Hyp}_{n,S,W}$ wirklich auf Ω normiert ist, ist etwas unangenehm; wir begnügen uns mit der probabilistischen Argumentation, dass wir die Summe der Wahrscheinlichkeiten aller möglichen Ereignisse erhalten, indem wir $\mathrm{Hyp}_{n,S,W}(s)$ über $s \in \Omega$ summieren. Man nennt $\mathrm{Hyp}_{n,S,W}$ die *hypergeometrische Verteilung* auf Ω mit Parametern n, S und W. ◇

Beispiel 1.3.2. Wie groß ist die Wahrscheinlichkeit, dass nach dem Geben beim Skatspiel Spieler A genau drei Asse besitzt? (Wir gehen davon aus, dass 32 Karten an drei Spieler, die je zehn Karten erhalten, und den Skat verteilt werden, und dass genau vier Asse unter den 32 Karten sind. Natürlich nehmen wir wieder an, dass das Spiel gut gemischt ist.)

Um diese Frage zu beantworten, benutzen wir die hypergeometrische Verteilung mit $S = 4$, $W = 28$ und $n = 10$ und $s = 3$, denn Spieler A erhält zehn von 32 Karten, und es gibt vier Asse und 28 andere Karten im Spiel. Die gesuchte Wahrscheinlichkeit ist also

$$\mathrm{Hyp}_{10,4,28}(3) = \frac{\binom{4}{3}\binom{28}{7}}{\binom{32}{10}} = \frac{66}{899}.$$

◇

Eine der wichtigsten Verteilungen, die uns in dieser Vorlesung immer wieder begegnen wird, taucht auf bei der Frage nach Erfolgszahlen bei Serien von Glücksspielen.

Beispiel 1.3.3 (Bernoulliverteilung)**.** Wir spielen ein Glücksspiel, bei dem es mit Wahrscheinlichkeit $p \in [0,1]$ einen Erfolg gibt und sonst keinen. Insgesamt spielen wir es n Mal, und alle n Spiele sind unabhängig voneinander. (Zur Präzisierung dieses Begriffs siehe Abschnitt 2.2, insbesondere Beispiel 2.2.3.) Man sagt, wir führen ein *Bernoulli-Experiment* der Länge n mit Erfolgswahrscheinlichkeit p durch. Dieses Experiment können wir mit dem Wahrscheinlichkeitsraum (Ω, q) gut beschreiben, wobei $\Omega = \{0,1\}^n$ (die Menge aller Vektoren der Länge n mit Koeffizienten 0 oder 1, wobei '1' für 'Erfolg' steht und '0' für 'Misserfolg'), und die Einzelwahrscheinlichkeiten sind gegeben durch

$$q(\omega) = p^{\sum_{i=1}^n \omega_i}(1-p)^{n-\sum_{i=1}^n \omega_i}, \qquad \omega = (\omega_1, \dots, \omega_n) \in \Omega.$$

(Man beachte, dass $\sum_{i=1}^n \omega_i$ gleich der Anzahl der Einsen im Vektor ω ist und $n - \sum_{i=1}^n \omega_i$ die Anzahl der Nullen.) Den Nachweis, dass q tatsächlich eine Wahrscheinlichkeitsverteilung auf Ω induziert (d. h., dass die $q(\omega)$ sich zu Eins aufsummieren), führt man am besten über eine vollständige Induktion. ◇

Beispiel 1.3.4 (Binomialverteilung)**.** Wir führen ein Bernoulli-Experiment der Länge n mit Erfolgswahrscheinlichkeit $p \in [0,1]$ durch wie in Beispiel 1.3.3. Wir groß ist die Wahrscheinlichkeit, genau k Erfolge zu haben, wobei $k \in \{0, \dots, n\}$? Formaler gefragt, wie groß ist die Wahrscheinlichkeit des Ereignisses $A_k = \{\omega \in \Omega \colon \sum_{i=1}^n \omega_i = k\}$?

Die Lösung anzugeben fällt uns nicht schwer. Es gibt $\binom{n}{k}$ Anzahlen von Spielverläufen, in denen es genau k Erfolge gibt, und die k Erfolge geschehen mit Wahrscheinlichkeit p^k, und die $n-k$ Misserfolge mit Wahrscheinlichkeit $(1-p)^{n-k}$. Also ist die gesuchte Wahrscheinlichkeit gegeben durch die Formel

$$\mathrm{Bi}_{n,p}(k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k \in \{0, \dots, n\}.$$

Die Verteilung auf $\Omega = \{0, \dots, n\}$ mit den Einzelwahrscheinlichkeiten $\mathrm{Bi}_{n,p}$ heißt die *Binomialverteilung* mit Parametern $p \in [0,1]$ und $n \in \mathbb{N}$. Wir erinnern an den Binomischen Lehrsatz

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}, \qquad x, y \in \mathbb{R}, n \in \mathbb{N}, \tag{1.3.1}$$

der eine analytische Rechtfertigung dafür liefert, dass $\mathrm{Bi}_{n,p}$ tatsächlich eine Wahrscheinlichkeitsverteilung induziert. Ohne Probleme leitet man die Beziehungen

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n, \qquad \sum_{k=0}^{n} (-1)^k \binom{n}{k} = 0, \qquad \sum_{k=0}^{n} k \binom{n}{k} = n 2^{n-1}, \qquad \binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1} \tag{1.3.2}$$

her (Übungsaufgabe). ◇

In den folgenden beiden Beispielen führen wir die wohl wichtigsten Verteilungen auf einem abzählbar *unendlichen* Raum ein.

Beispiel 1.3.5 (Geometrische Verteilung)**.** Wir stellen uns vor, dass wir das Glücksspiel aus Beispiel 1.3.4 so lange spielen, bis wir zum ersten Male einen Erfolg verbuchen können. Mit welcher Wahrscheinlichkeit passiert dies beim k-ten Spiel (wobei $k \in \mathbb{N}$)?

Auch diese Frage beantworten wir leicht. Das gesuchte Ereignis ist dadurch charakterisiert, dass wir genau $k-1$ Mal hinter einander keinen Erfolg haben und im k-ten Spiel endlich einen. Also ist die gesuchte Wahrscheinlichkeit gleich

$$\mathrm{Geo}_p(k) = p(1-p)^{k-1}, \qquad k \in \mathbb{N}. \tag{1.3.3}$$

Die Verteilung auf $\Omega = \mathbb{N}$ mit den Einzelwahrscheinlichkeiten Geo_p heißt *geometrische Verteilung* mit Parameter $p \in (0,1)$. Sie ist also die Verteilung der Wartezeit auf den ersten Erfolg in einer Serie von Glücksspielen, die jeweils mit Erfolgswahrscheinlichkeit p ablaufen. Manche Autoren definieren die geometrische Verteilung auf der Menge $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$ mittels der Formel $p(1-p)^k$ für $k \in \mathbb{N}_0$. Wir werden diese Verteilung mit $\widetilde{\mathrm{Geo}}_p$ bezeichnen.

Die Interpretation als Wartezeit auf den ersten Erfolg in einer möglicherweise unendlich langen Serie von Glücksspielen scheint den überabzählbaren Wahrscheinlichkeitsraum $\Omega = \{0,1\}^{\mathbb{N}}$ zu erfordern, die Menge aller unendlich langen 0-1-Folgen. Das ist allerdings tatsächlich nicht nötig, da die Betrachtung des Ereignisses, dass das k-te Spiel den ersten Erfolg bringt, nur den endlichen Wahrscheinlichkeitsraum $\Omega = \{0,1\}^k$ erfordert. ◇

Beispiel 1.3.6 (Poisson-Verteilung)**.** Mit einem Parameter $\alpha \in (0, \infty)$ betrachten wir die Verteilung auf $\Omega = \mathbb{N}_0 = \{0, 1, 2, \dots\}$, die gegeben ist durch

$$\mathrm{Po}_\alpha(k) = e^{-\alpha} \frac{\alpha^k}{k!}, \qquad k \in \mathbb{N}_0. \tag{1.3.4}$$

Die Verteilung Po_α heißt die *Poisson-Verteilung* zum Parameter α. Eine ihrer wichtigsten Bedeutungen und Anwendungen kommt von der folgenden Überlegung her. Wir denken uns, dass im Zeitintervall $(0,t]$ eine zufällige Anzahl von Zeitpunkten zufällig verteilt sind. Um dies zu präzisieren, müssen wir ein paar zusätzliche, plausible Annahmen machen. Wir teilen das Intervall in n gleiche Stücke der Länge t/n ein und nehmen an, dass jedes dieser Teilintervalle für großes n höchstens einen dieser zufälligen Punkte enthält. Außerdem nehmen wir an, dass die Entscheidung, ob die n Teilintervalle jeweils einen solchen Zeitpunkt enthalten, *unabhängig* fällt (ein Begriff, den wir in Abschnitt 2.2 präzisieren werden), und zwar mit Wahrscheinlichkeit p_n. Mit anderen Worten, wir nehmen an, dass wir es mit einem Bernoulli-Experiment der Länge n mit Erfolgswahrscheinlichkeit p_n zu tun haben. Den Parameter $p = p_n$ legen wir so fest, dass für großes n die erwartete Gesamtzahl der zufälligen Punkte im Intervall $(0,t]$ (dies ist die Zahl np_n, was in Abschnitt 3.3 präzisiert werden wird) ungefähr ein fester Wert $\alpha \in (0,\infty)$ ist. Der folgende Satz sagt, dass die Verteilung der Anzahl der zufälligen Punkte in $(0,t]$ dann etwa durch die Poisson-Verteilung gegeben ist:

Satz 1.3.7 (Poissonapproximation der Binomialverteilung; Poisson'scher Grenzwertsatz)**.** *Falls* $\lim_{n\to\infty} np_n = \alpha$, *so gilt für jedes* $k \in \mathbb{N}_0$

$$\lim_{n\to\infty} \mathrm{Bi}_{n,p_n}(k) = \mathrm{Po}_\alpha(k).$$

Beweis. Wir benutzen die (gebräuchliche) Notation $a_n \sim b_n$, wenn $\lim_{n\to\infty} a_n/b_n = 1$. Also sieht man, dass

$$\binom{n}{k} = \frac{n^k}{k!}\frac{n(n-1)(n-2)\cdot\dots\cdot(n-k+1)}{n^k} \sim \frac{n^k}{k!}, \tag{1.3.5}$$

da wir den zweiten Bruch als Produkt von k Brüchen schreiben können, die jeweils gegen Eins konvergieren. Nun benutzen wir die Tatsache $\lim_{t\to 0}(1+t)^{1/t} = e$, um zu schlussfolgern, dass

$$\begin{aligned}\mathrm{Bi}_{n,p_n}(k) &= \binom{n}{k} p_n^k (1-p_n)^n (1-p_n)^{-k} \sim \frac{n^k}{k!} p_n^k \big[(1-p_n)^{1/p_n}\big]^{np_n} \sim \frac{(np_n)^k}{k!} e^{-np_n}\\ &\to \frac{\alpha^k}{k!} e^{-\alpha} = \mathrm{Po}_\alpha(k).\end{aligned}$$

□◇

Beispiel 1.3.8 (Verallgemeinerte hypergeometrische Verteilung und Multinomialverteilung)**.** In einem Teich gibt es N Fische von k unterschiedlichen Sorten, und zwar genau N_i Stück von Sorte $i \in \{1,\dots,k\}$, wobei $N = N_1 + \dots + N_k$. Wir fischen $n \in \{0,\dots,N\}$ Fische zufällig aus dem Teich heraus. Mit welcher Wahrscheinlichkeit fangen wir dabei genau $n_i \in \mathbb{N}_0$ Fische von der Sorte i für jedes $i \in \{1,\dots,k\}$? Wir setzen natürlich voraus, dass $n = n_1 + \dots + n_k$ und $n_i \in \{0,\dots,N_i\}$ für jedes i.

Die Antwort ist offensichtlich (bei 'guter Mischung' der Fische im Teich) gegeben durch die Formel

$$\mathrm{Hyp}_{n;N_1,\dots,N_k}(n_1,\dots,n_k) = \frac{\prod_{i=1}^k \binom{N_i}{n_i}}{\binom{N}{n}}.$$

Die Verteilung $\mathrm{Hyp}_{n;N_1,\dots,N_k}$ auf der Menge

$$\Omega_n = \big\{(n_1,\dots,n_k) \in \mathbb{N}_0^k \colon n_1 + \dots + n_k = n\big\}$$

heißt die *verallgemeinerte hypergeometrische Verteilung* mit Parametern n und $N_1, \dots, N_k$. Man kann Ω_n auffassen als die Menge aller Zerlegungen einer Menge von n ununterscheidbaren Objekten in k Teilmengen, siehe auch Beispiel 1.2.5.

Falls man annehmen kann, dass sich von jeder Sorte sehr viele Fische im Teich befinden, dann kann man die verallgemeinerte hypergeometrische Verteilung auch approximieren mit einer leichter handhabbaren Verteilung. Wir nehmen dabei an, dass $N_1, \dots, N_k$ so groß sind, dass die obige Wahrscheinlichkeit im Wesentlichen nur noch von den relativen Anteilen N_i/N abhängt. Dann kann man in etwa davon ausgehen, bei jedem Fang eines einzelnen Fisches die Sorte i mit Wahrscheinlichkeit $p_i = N_i/N$ zu erwischen, unabhängig davon, wieviele und welche Fische man schon vorher gefischt hat. Dies wird im folgenden Satz präzisiert.

Satz 1.3.9 (Multinomialapproximation)**.** *Für jedes $n \in \mathbb{N}$ und alle $n_1, \dots, n_k \in \mathbb{N}$ mit $n_1 + \cdots + n_k = n$ und für alle $p_1, \dots, p_k \in [0,1]$ mit $p_1 + \cdots + p_k = 1$ gilt*

$$\lim_{\substack{N, N_1, \dots, N_k \to \infty \\ N_i/N \to p_i \forall i}} \mathrm{Hyp}_{n; N_1, \dots, N_k}(n_1, \dots, n_k) = \mathrm{Mul}_{n; p_1, \dots, p_k}(n_1, \dots, n_k) = \binom{n}{n_1, \dots, n_k} \prod_{i=1}^k p_i^{n_i}.$$

Der *Multinomialkoeffizient*

$$\binom{n}{n_1, \dots, n_k} = \frac{n!}{n_1!\, n_2! \dots n_k!}, \qquad \text{für } n_1 + \cdots + n_k = n,$$

gibt offensichtlich die Anzahl der Möglichkeiten an, eine Menge von n unterscheidbaren Objekten in k Teilmengen mit $n_1, n_2, \dots, n_k$ Elementen zu zerlegen. Die in Satz 1.3.9 eingeführte Verteilung $\mathrm{Mul}_{n; p_1, \dots, p_k}$ auf Ω_n heißt die *Multinomialverteilung* mit Parametern $p_1, \dots, p_k$. Sie beschreibt die Wahrscheinlichkeit, genau n_i Mal das Ergebnis i zu erhalten für jedes $i \in \{1, \dots, k\}$, wenn n Mal hintereinander ein Zufallsexperiment mit den möglichen Ausgängen $1, \dots, k$, die jeweils die Wahrscheinlichkeit $p_1, \dots, p_k$ haben, durchgeführt wird. Daher ist die Multinomialverteilung eine Verallgemeinerung der Binomialverteilung auf mehr als zwei mögliche Ausgänge des Experiments. Der *Multinomialsatz*

$$(x_1 + \cdots + x_k)^n = \sum_{\substack{n_1, \dots, n_k \in \mathbb{N}_0 \\ n_1 + \cdots + n_k = n}} \binom{n}{n_1, \dots, n_k} \prod_{i=1}^k x_i^{n_i} \qquad \text{für alle } x_1, \dots, x_k \in \mathbb{R},$$

liefert die Begründung, dass $\mathrm{Mul}_{n; p_1, \dots, p_k}$ tatsächlich eine Wahrscheinlichkeitsverteilung auf Ω_n induziert.

Beweis von Satz 1.3.9. Man sieht leicht wie in (1.3.5), dass im Grenzwert $N_i \to \infty$ gilt

$$\binom{N_i}{n_i} \sim \frac{N_i^{n_i}}{n_i!},$$

wobei wir (wie allgemein üblich) $a \sim b$ schreiben, wenn der Quotient von a und b gegen 1 strebt. Also erhalten wir für den Grenzwert $N, N_1, \dots, N_k \to \infty$ mit $N_i/N \to p_i$ für alle i:

$$\mathrm{Hyp}_{n; N_1, \dots, N_k}(n_1, \dots, n_k) \sim \frac{\prod_{i=1}^k \frac{N_i^{n_i}}{n_i!}}{\frac{N^n}{n!}} = \binom{n}{n_1, \dots, n_k} \prod_{i=1}^k \Big(\frac{N_i}{N}\Big)^{n_i} \sim \binom{n}{n_1, \dots, n_k} \prod_{i=1}^k p_i^{n_i}.$$

□◇

Kapitel 2

Bedingte Wahrscheinlichkeit und Unabhängigkeit

In vielen Situationen liegt schon eine gewisse Information vor, wenn man die Wahrscheinlichkeit eines Ereignisses bestimmen möchte. Bei Abschluss einer Lebensversicherung kennt man schon das aktuelle Alter des Antragstellers, beim Skatspiel kennt man schon die eigenen Karten, bei Meinungsumfragen hat man vor der eigentlichen Befragung schon die Bevölkerungsgruppe des oder der Befragten festgestellt. Das heißt, dass man über das Eintreten oder Nichteintreten eines Ereignisses B schon informiert ist, wenn man die Wahrscheinlichkeit eines Ereignisses A bestimmen will. In diesem Abschnitt wollen wir mathematisch präzisieren, was eine bedingte Wahrscheinlichkeit von A gegeben B ist, und was es heißen soll, dass die Ereignisse A und B unabhängig sind.

Dem gesamten Kapitel legen wir einen diskreten Wahrscheinlichkeitsraum (Ω, p) zu Grunde.

2.1 Bedingte Wahrscheinlichkeiten

Wir beginnen wieder mit einem einführenden Beispiel.

Beispiel 2.1.1 (Umfrage)**.** In einer Meinungsumfrage soll der Anzahl der Raucher an der Bevölkerung fest gestellt werden, das heißt die Wahrscheinlichkeit des Ereignisses A, dass eine zufällig heraus gegriffene Person Raucher ist. Hierbei möchte man allerdings mehrere Bevölkerungsgruppen unterscheiden, zum Beispiel die Gruppe der 21- bis 30jährigen Frauen, was das Ereignis B darstellen soll. Man möchte also die bedingte Wahrscheinlichkeit von A unter der Voraussetzung, dass B eintritt, feststellen. Dazu wird man normalerweise die Anzahl der rauchenden 21- bis 30jährigen Frauen durch die Anzahl der 21- bis 30jährigen Frauen teilen, d.h., den Anteil der Raucherinnen unter den 21- bis 30jährigen Frauen bestimmen. Dies führt auf die plausible Formel

$$\mathbb{P}(\text{Raucher(in)} \mid \text{21- bis 30jährige Frau}) = \frac{|\{\text{21- bis 30jährige Raucherinnen}\}|}{|\{\text{21- bis 30jährige Frauen}\}|} = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

$\diamond$

Dieses Beispiel nehmen wir nun als Anlass zu einer allgemeinen Definition. Wir müssen allerdings beachten, dass die obige Formel Gebrauch davon macht, dass wir es mit einer Gleichverteilung zu tun haben.

Definition 2.1.2 (bedingte Wahrscheinlichkeit)**.** *Es seien A und B zwei Ereignisse mit $\mathbb{P}(B) > 0$. Mit*

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

bezeichnen wir die bedingte Wahrscheinlichkeit von A gegeben B.

Beispiel 2.1.3. Es sei p_k die Wahrscheinlichkeit, dass man während des k-ten Lebensjahrs stirbt. (Hier machen wir natürlich wie üblich eine ganze Reihe an stark vereinfachenden Annahmen, z. B. dass diese Wahrscheinlichkeit nicht abhängt vom Geschlecht, dem Geburtsjahr, dem Wohnort, dem Beruf etc.) Dann ist $s_k = p_k + p_{k+1} + p_{k+2} + \dots$ die Wahrscheinlichkeit, dass man das Alter k erreicht. Wenn man für einen Zeitgenossen die Wahrscheinlichkeit, im k-ten Lebensjahr zu sterben, ermitteln möchte, sollte man beachten, dass dieser gerade lebt und das Alter l hat (mit $l \leq k$). Also sollte man tatsächlich die *bedingte* Wahrscheinlichkeit gegeben, dass er gerade l Jahre alt ist, angeben, und dies ist der Quotient p_k/s_l. ◇

Wir sammeln ein paar einfache, aber wichtige Eigenschaften der bedingten Wahrscheinlichkeit.

Lemma 2.1.4 (totale Wahrscheinlichkeit, Bayessche Formel)**.** *Es sei B ein Ereignis mit $\mathbb{P}(B) > 0$.*

(i) *$\mathbb{P}(\cdot \mid B)$ erfüllt die Kolmogorovschen Axiome (siehe Bemerkung 1.1.3), d. h. es gilt $\mathbb{P}(\Omega \mid B) = 1$, und für alle Folgen $(A_i)_{i\in\mathbb{N}}$ von paarweise disjunkten Ereignissen gilt $\mathbb{P}(\bigcup_{i\in\mathbb{N}} A_i \mid B) = \sum_{i\in\mathbb{N}} \mathbb{P}(A_i \mid B)$.*

(ii) *Es gilt die* Formel von der totalen Wahrscheinlichkeit*: Für jedes Ereignis A und jede Folge $(B_i)_{i\in\mathbb{N}}$ von paarweise disjunkten Ereignissen mit $B = \bigcup_{i\in\mathbb{N}} B_i$ und $\mathbb{P}(B_i) > 0$ für alle $i \in \mathbb{N}$ gilt*

$$\mathbb{P}(A \cap B) = \sum_{i\in\mathbb{N}} \mathbb{P}(B_i)\mathbb{P}(A \mid B_i).$$

(iii) *Es gilt die* Formel von Bayes*: Für jedes Ereignis A mit $\mathbb{P}(A) > 0$ und jede Folge $(B_i)_{i\in\mathbb{N}}$ von paarweise disjunkten Ereignissen mit $\Omega = \bigcup_{i\in\mathbb{N}} B_i$ und $\mathbb{P}(B_i) > 0$ für alle $i \in \mathbb{N}$ gilt*

$$\mathbb{P}(B_i \mid A) = \frac{\mathbb{P}(B_i)\mathbb{P}(A \mid B_i)}{\sum_{j\in\mathbb{N}} \mathbb{P}(B_j)\mathbb{P}(A \mid B_j)}, \qquad i \in \mathbb{N}.$$

Beweis. einfache Übungsaufgabe. □

Die folgende Anwendung der Bayesschen Formel zeigt, dass man nicht zu alarmiert sein muss, wenn ein nicht ganz hundertprozentiger Test auf eine seltene Krankheit anspricht.

Beispiel 2.1.5 (Test auf eine seltene Krankheit)**.** Eine seltene Krankheit liegt bei ungefähr 0,5 Prozent der Bevölkerung vor. Es gibt einen (recht guten) Test auf diese Krankheit, der bei 99 Prozent der Kranken anspricht, aber auch bei 2 Prozent der Gesunden. Mit welcher Wahrscheinlichkeit ist eine getestete Person tatsächlich krank, wenn der Test bei ihr angesprochen hat?

Wir teilen also die Gesamtmenge Ω aller getesteten Personen ein in die Ereignisse B_1 der kranken und B_2 der gesunden getesteten Personen. Es sei A das Ereignis, dass die getestete Person auf den Test anspricht. Gesucht ist die bedingte Wahrscheinlichkeit $\mathbb{P}(B_1 \mid A)$. Dem Aufgabentext entnimmt man die Angaben

$$\mathbb{P}(B_1) = 0,005, \qquad \mathbb{P}(A \mid B_1) = 0,99, \qquad \mathbb{P}(A \mid B_2) = 0,02.$$

Insbesondere ist natürlich $\mathbb{P}(B_2) = 0,995$. Nun können wir die Bayessche Formel benutzen, um die Aufgabe zu lösen:

$$\mathbb{P}(B_1 \mid A) = \frac{\mathbb{P}(B_1)\mathbb{P}(A \mid B_1)}{\mathbb{P}(B_1)\mathbb{P}(A \mid B_1) + \mathbb{P}(B_2)\mathbb{P}(A \mid B_2)} = \frac{0,005 \cdot 0,99}{0,005 \cdot 0,99 + 0,995 \cdot 0,02} = \frac{495}{2485} \approx 0,2.$$

Also ist trotz positiven Testergebnisses die Wahrscheinlichkeit, diese Krankheit zu haben, nicht alarmierend hoch. Man sollte unter Beobachtung bleiben. $\diamond$

In manchen Fällen erweist sich die folgend vorgestellte Formel als nützlich.

Lemma 2.1.6 (Multiplikationsformel). *Für jedes $n \in \mathbb{N}$ und alle Ereignisse $A_1, \dots, A_n \subset \Omega$ mit $\mathbb{P}(A_1 \cap \dots \cap A_{n-1}) \neq 0$ gilt*

$$\mathbb{P}(A_1 \cap \dots \cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2 \mid A_1)\mathbb{P}(A_3 \mid A_1 \cap A_2) \dots \mathbb{P}(A_n \mid A_1 \cap \dots \cap A_{n-1}).$$

Beweis. Klar. $\square$

Beispiel 2.1.7. Mit welcher Wahrscheinlichkeit besitzt beim Skat jeder der drei Spieler nach dem Geben genau ein Ass?

Wir verteilen also zufällig 32 Karten, darunter vier Asse, an drei Spieler, die je zehn Karten erhalten, und den Skat. Es sei A_i das Ereignis, dass der i-te Spieler genau ein Ass erhält. Mit Hilfe der Multiplikationsformel aus Lemma 2.1.6 errechnen wir:

$$\begin{aligned}\mathbb{P}(A_1 \cap A_2 \cap A_3) &= \mathbb{P}(A_1)\mathbb{P}(A_2 \mid A_1)\mathbb{P}(A_3 \mid A_1 \cap A_2) \\ &= \frac{\binom{4}{1}\binom{28}{9}}{\binom{32}{10}} \times \frac{\binom{3}{1}\binom{19}{9}}{\binom{22}{10}} \times \frac{\binom{2}{1}\binom{10}{9}}{\binom{12}{10}} = 10^3 \frac{2 \cdot 4!}{32 \cdot 31 \cdot 30 \cdot 29} \approx 0,0556.\end{aligned}$$

$\diamond$

2.2 Unabhängigkeit von Ereignissen

In diesem Abschnitt präzisieren wir, was es heißt, dass zwei oder mehr Ereignisse unabhängig sind. Intuitiv bedeutet Unabhängigkeit der Ereignisse A und B, dass das Eintreffen oder Nichteintreffen von A nicht beeinflusst wird vom Eintreffen oder Nichteintreffen von B. Ein elementares Beispiel, in dem man den Unterschied zwischen Abhängigkeit und Unabhängigkeit gut sehen kann, ist das Ziehen zweier Kugeln aus einer Urne ohne Zurücklegen im Vergleich zum Ziehen mit Zurücklegen: Das Ziehen der ersten Kugel sollte das Ergebnis der zweiten Ziehung beeinflussen, wenn man die erste nicht wieder zurück legt, aber nicht, wenn man sie zurück legt und damit den Zustand, der vor der ersten Ziehung vorlag, wieder her stellt.

Wie formalisiert man aber Unabhängigkeit zweier Ereignisse A und B? Eine gute Idee ist es, zu fordern, dass die Wahrscheinlichkeit von A überein stimmen muss mit der bedingten

Wahrscheinlichkeit von A gegeben B (zumindest, falls $\mathbb{P}(B) > 0$), also $\mathbb{P}(A) = \mathbb{P}(A \mid B)$. Dies ist auch tatsächlich eine plausible Formel, doch sollte man in der Definition für Symmetrie sorgen. Ferner brauchen wir einen tragfähigen Begriff der Unabhängigkeit mehrerer Ereignisse:

Definition 2.2.1 (Unabhängigkeit). *(i) Zwei Ereignisse A und B heißen* unabhängig, *falls gilt:* $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

(ii) Eine Familie $(A_i)_{i\in I}$ (wobei I eine beliebige Indexmenge ist) heißt unabhängig, *falls für jede endliche Teilmenge J von I die folgende* Produktformel *gilt:*

$$\mathbb{P}\Big(\bigcap_{i\in J} A_i\Big) = \prod_{i\in J} \mathbb{P}(A_i). \tag{2.2.1}$$

Beispiel 2.2.2. Beim n-fachen Wurf mit einem fairen Würfel (also $\Omega = \{1,\dots,6\}^n$ mit der Gleichverteilung) sind die n Ereignisse 'i-ter Wurf zeigt a_i' für $i = 1,\dots,n$ bei beliebigen $a_1,\dots,a_n \in \{1,\dots,6\}$ unabhängig, denn ihre jeweilige Wahrscheinlichkeit ist $\frac{1}{6}$, und die Wahrscheinlichkeit eines beliebigen Schnittes dieser Ereignisse ist $\frac{1}{6}$ hoch die Anzahl der geschnittenen Mengen. ◇

Beispiel 2.2.3 (Bernoulli-Experiment). Das in Beispiel 1.3.3 eingeführte Bernoulli-Experiment besteht tatsächlich aus n unabhängigen Experimenten, die jeweils Erfolgswahrscheinlichkeit p haben. Mit anderen Worten, die Ereignisse $A_i = \{\omega \in \Omega\colon \omega_i = 1\}$ ('i-tes Spiel hat Erfolg') sind unabhängig für $i = 1,\dots,n$. Der Nachweis der Unabhängigkeit ist eine Übungsaufgabe. ◇

Beispiel 2.2.4. Beim zweimaligen Ziehen je einer Kugel aus einer gut gemischten Urne mit s schwarzen und w weißen Kugeln sind die Ereignisse 'erste Kugel ist weiß' und 'zweite Kugel ist weiß' unabhängig, wenn nach dem ersten Ziehen die gezogene Kugel zurück gelegt wurde, aber nicht unabhängig, wenn dies nicht geschah. (Übungsaufgabe: Man beweise dies.) Dies unterstreicht, dass Unabhängigkeit nicht nur eine Eigenschaft der Ereignisse ist, sondern auch entscheidend vom Wahrscheinlichkeitsmaß abhängt. ◇

Bemerkung 2.2.5. (a) Es ist klar, dass jede Teilfamilie einer Familie von unabhängigen Ereignissen unabhängig ist.

(b) Die Forderung, dass die Produktformel in (2.2.1) für jede endliche Teilauswahl J gelten soll (und nicht nur für $J = I$ oder nur für alle zweielementigen Teilauswahlen) ist sehr wesentlich. Der (weitaus weniger wichtige) Begriff der *paarweisen Unabhängigkeit* fordert die Produktformel nur für die zweielementigen Teilmengen J von I. Im folgenden Beispiel haben wir drei paarweise unabhängige Ereignisse, die nicht unabhängig sind: In $\Omega = \{1,2,3,4\}$ mit der Gleichverteilung betrachten wir die Ereignisse $A_1 = \{1,2\}$, $A_2 = \{2,3\}$ und $A_3 = \{3,1\}$. (Übungsaufgabe: Verifiziere, dass A_1, A_2 und A_3 paarweise unabhängig, aber nicht unabhängig sind.)

(c) Unabhängigkeit ist keine Eigenschaft von *Mengen* von Ereignissen, sondern von *Tupeln* von Ereignissen, die allerdings nicht von der Reihenfolge abhängt. Dies ist wichtig, wenn z. B. eines der Ereignisse in dem betrachteten Tupel mehrmals auftritt. Falls das Paar (A, A) unabhängig ist, so folgt $\mathbb{P}(A) = \mathbb{P}(A \cap A) = \mathbb{P}(A)\mathbb{P}(A) = \mathbb{P}(A)^2$, also $\mathbb{P}(A) \in \{0,1\}$.

(d) Unabhängigkeit kann sogar trotz Kausalität vorliegen, wie das folgende Beispiel zeigt. Beim Würfeln mit zwei fairen Würfeln (also $\Omega = \{1,\dots,6\}^2$ mit der Gleichverteilung) betrachten wir die Ereignisse $A = \{\text{Augensumme ist } 7\} = \{(i,j) \in \Omega\colon i+j=7\}$ und $B = \{\text{erster Würfel zeigt } 6\} = \{(i,j) \in \Omega\colon i = 6\}$. Dann gilt $\mathbb{P}(A \cap B) = \mathbb{P}(\{(6,1)\}) = \frac{1}{36} = \frac{1}{6} \times \frac{1}{6} = \mathbb{P}(A)\mathbb{P}(B)$. Also sind A und B unabhängig, obwohl offensichtlich eine Kausalität zwischen diesen beiden Ereignissen vorliegt.

$\diamond$

Ein hilfreiches Kriterium für Unabhängigkeit von Ereignissen ist das Folgende. Es sagt, dass die Unabhängigkeit von n Ereignissen gleichbedeutend ist mit der Gültigkeit der Produktformel in (2.2.1) für $J = \{1,\dots,n\}$, aber man muss für jedes der Ereignisse auch das Komplement zulassen können. Zur bequemsten Formulierung benötigen wir die Notation $A^1 = A$ für Ereignisse A, während A^{c} wie üblich das Komplement von A ist.

Lemma 2.2.6. *Ereignisse $A_1,\dots,A_n$ sind genau dann unabhängig, wenn für alle $k_1,\dots,k_n \in \{1,\mathrm{c}\}$ gilt:*

$$\mathbb{P}\Big(\bigcap_{i=1}^n A_i^{k_i}\Big) = \prod_{i=1}^n \mathbb{P}(A_i^{k_i}). \tag{2.2.2}$$

Beweis.

'$\Longrightarrow$': Wir setzen die Unabhängigkeit der $A_1,\dots,A_n$ voraus und zeigen (2.2.2) mit einer Induktion nach n. Für $n=1$ ist nichts zu zeigen. Wir nehmen nun den Schluss von n auf $n+1$ vor und setzen voraus, dass $A_1,\dots,A_{n+1}$ unabhängig sind. Nun zeigen wir (2.2.2) für $n+1$ statt n mit einer weiteren Induktion über die Anzahl m der 'c' in $k_1,\dots,k_{n+1}$.

Für $m=0$ folgt die Behauptung aus der Definition der Unabhängigkeit der $A_1,\dots,A_{n+1}$. Nun machen wir den Induktionsschluss von $m \in \{0,\dots,n\}$ auf $m+1$: Es seien $m+1$ Komplementzeichen in $k_1,\dots,k_{n+1}$. Durch eine geeignete Permutation der Ereignisse können wir annehmen, dass $k_{n+1} = \mathrm{c}$ ist. Dann haben wir

$$\mathbb{P}\Big(\bigcap_{i=1}^{n+1} A_i^{k_i}\Big) = \mathbb{P}\Big(\bigcap_{i=1}^{n} A_i^{k_i} \cap A_{n+1}^{\mathrm{c}}\Big) = \mathbb{P}\Big(\bigcap_{i=1}^{n} A_i^{k_i}\Big) - \mathbb{P}\Big(\bigcap_{i=1}^{n} A_i^{k_i} \cap A_{n+1}\Big).$$

Der erste Summand ist nach Induktionsvoraussetzung an n gleich $\prod_{i=1}^n \mathbb{P}(A_i^{k_i})$, und der zweite nach Induktionsvoraussetzung an m gleich $[\prod_{i=1}^n \mathbb{P}(A_i^{k_i})]\mathbb{P}(A_{n+1})$. Nun setzt man diese zwei Gleichungen ein, klammert geeignet aus und erhält die beabsichtigte Gleichung:

$$\mathbb{P}\Big(\bigcap_{i=1}^{n+1} A_i^{k_i}\Big) = \prod_{i=1}^{n+1} \mathbb{P}(A_i^{k_i}).$$

'$\Longleftarrow$': Wir setzen also die Gültigkeit von (2.2.2) voraus und zeigen nun die Unabhängigkeit von $A_1,\dots,A_n$. Sei $\{i_1,\dots,i_k\} \subset \{1,\dots,n\}$, und sei $\{j_1,\dots,j_m\}$ das Komplement von $\{i_1,\dots,i_k\}$ in $\{1,\dots,n\}$. Dann lässt sich $\bigcap_{l=1}^k A_{i_l}$ wie folgt als disjunkte Vereinigung schreiben:

$$\bigcap_{l=1}^k A_{i_l} = \bigcup_{k_1,\dots,k_m \in \{1,\mathrm{c}\}} \bigcap_{l=1}^k A_{i_l} \cap \bigcap_{s=1}^m A_{j_s}^{k_s}.$$

Die Wahrscheinlichkeit von der Menge auf der rechten Seite ist nach unserer Voraussetzung gleich

$$\sum_{k_1,\dots,k_m \in \{1,\mathrm{c}\}} \prod_{l=1}^k \mathbb{P}(A_{i_l}) \times \prod_{s=1}^m \mathbb{P}(A_{j_s}^{k_s}) = \prod_{l=1}^k \mathbb{P}(A_{i_l}),$$

wobei wir auch die Additivität bei paarweiser Disjunktheit benutzten. □

Da Lemma 2.2.6 symmetrisch ist in den Ereignissen und ihren Komplementen, ist die folgende Folgerung klar.

Korollar 2.2.7. *Ereignisse $A_1, \dots, A_n$ sind genau dann unabhängig, wenn ihre Komplemente $A_1^c, \dots, A_n^c$ unabhängig sind.*

Eine sehr hübsche Anwendung betrifft die Riemannsche Zetafunktion und die Euler'sche Produktformel:

Beispiel 2.2.8 (Euler'sche Primzahlformel)**.** Die *Riemann'sche Zetafunktion* ist definiert durch

$$\zeta(s) = \sum_{k=1}^{\infty} k^{-s}, \qquad s > 1.$$

Die *Euler'sche Primzahlformel* oder *Produktdarstellung* dieser Funktion lautet

$$\zeta(s) = \prod_{p \text{ prim}} \left(1 - p^{-s}\right)^{-1}.$$

Diese Formel wollen wir nun mit probabilistischen Mitteln beweisen. Wir skizzieren hier nur den Weg, die Ausformulierung der Details ist eine Übungsaufgabe.

Wir erhalten einen Wahrscheinlichkeitsraum $(\mathbb{N}, q)$, indem wir $q(k) = k^{-s}\zeta(s)^{-1}$ für $k \in \mathbb{N}$ setzen. Man zeigt, dass die Mengen $p\mathbb{N} = \{pk \colon k \in \mathbb{N}\}$ (die Mengen der durch p teilbaren Zahlen) unabhängig sind, wenn p über alle Primzahlen genommen wird. Aufgrund von Folgerung 2.2.7 sind also auch die Komplemente $(p\mathbb{N})^c$ unabhängig. Deren Schnitt ist gleich der Menge $\{1\}$, deren Wahrscheinlichkeit gleich $\zeta(s)^{-1}$ ist. Die Unabhängigkeit liefert, dass der (unendliche) Schnitt der Mengen $(p\mathbb{N})^c$ mit p eine Primzahl die Wahrscheinlichkeit hat, die durch das Produkt der Wahrscheinlichkeiten dieser Mengen gegeben ist. Diese identifiziert man mit $(1-p^{-s})$, und daraus folgt die Euler'sche Produktdarstellung der Riemann'schen Zetafunktion. ◇

2.3 Produkträume

Der Begriff der Unabhängigkeit ist eng verknüpft mit Produkträumen von mehreren Wahrscheinlichkeitsräumen. Man denke an n nacheinander und unabhängig von einander ausgeführten Zufallsexperimenten, die jeweils durch einen Wahrscheinlichkeitsraum beschrieben werden. Die gesamte Versuchsreihe dieser n Experimente wird in natürlicher Weise durch den Produktraum beschrieben.

Definition 2.3.1. *Es seien $(\Omega_1, p_1), \dots, (\Omega_n, p_n)$ diskrete Wahrscheinlichkeitsräume. Auf der Produktmenge*

$$\Omega = \Omega_1 \times \cdots \times \Omega_n = \big\{(\omega_1, \dots, \omega_n) \colon \omega_1 \in \Omega_1, \dots, \omega_n \in \Omega_n\big\}$$

definieren wir $p \colon \Omega \to [0,1]$ durch

$$p\big((\omega_1, \dots, \omega_n)\big) = \prod_{i=1}^n p_i(\omega_i).$$

Dann ist (Ω, p) ein diskreter Wahrscheinlichkeitsraum, der der Produktwahrscheinlichkeitsraum *der Räume $(\Omega_1, p_1), \dots, (\Omega_n, p_n)$ genannt wird. Er wird auch mit $\otimes_{i=1}^n (\Omega_i, p_i)$ bezeichnet. Falls die Räume $(\Omega_1, p_1), \dots, (\Omega_n, p_n)$ identisch sind, schreiben wir auch $(\Omega, p) = (\Omega_1, p_1)^{\otimes n}$.*

Das Beispiel 2.2.2 wird im folgenden Satz verallgemeinert.

Satz 2.3.2. *Seien $(\Omega_1, p_1), \dots, (\Omega_n, p_n)$ diskrete Wahrscheinlichkeitsräume, und seien $A_1 \subset \Omega_1, \dots, A_n \subset \Omega_n$ Ereignisse in den jeweiligen Räumen. Dann sind die Ereignisse $A_1^{(1)}, \dots, A_n^{(n)}$ unabhängig im Produktraum (Ω, p) der Räume $(\Omega_1, p_1), \dots, (\Omega_n, p_n)$, wobei*

$$A_i^{(i)} = \{(\omega_1, \dots, \omega_n) \in \Omega \colon \omega_i \in A_i\}.$$

Die Ereignisse $A_i^{(i)}$ sind nichts weiter als eine Art Einbettung des Ereignisses 'A_i tritt ein' in den Produktraum an die i-te Stelle. Korrekter ausgedrückt, wenn $\pi_i \colon \Omega \to \Omega_i$ die Projektion $\pi_i((\omega_1, \dots, \omega_n)) = \omega_i$ auf den i-ten Faktor bezeichnet, dann ist

$$A_i^{(i)} = \{\omega \in \Omega \colon \pi_i(\omega) \in A_i\} = \pi_i^{-1}(A_i)$$

das Urbild von A_i unter π_i.

Beweis. Mit $\mathbb{P}$ bezeichnen wir das von p auf dem Produktraum Ω induzierte Wahrscheinlichkeitsmaß.

Sei $J \subset \{1, \dots, n\}$, dann ist die Produktformel in (2.2.1) für die Ereignisse $A_i^{(i)}$ mit $i \in J$ zu zeigen. Die Wahrscheinlichkeit des Schnittes dieser Mengen drücken wir durch Summation über alle Einzelereignisse aus:

$$\begin{aligned}\mathbb{P}\Big(\bigcap_{i \in J} A_i^{(i)}\Big) &= \mathbb{P}\big(\{\omega \colon \omega_i \in A_i \text{ für alle } i \in J\}\big)\\ &= \sum_{\omega \colon \omega_i \in A_i \forall i \in J} p(\omega) = \sum_{\omega_1, \dots, \omega_n \colon \omega_i \in A_i \forall i \in J} p_1(\omega_1) \dots p_n(\omega_n).\end{aligned}$$

Den letzten Ausdruck können wir mit der Notation $B_i = A_i$, falls $i \in J$ und $B_i = \Omega_i$, falls $i \notin J$, zusammenfassen als

$$\sum_{\omega_1 \in B_1} p_1(\omega_1) \cdots \sum_{\omega_n \in B_n} p_n(\omega_n) = \prod_{i=1}^n \sum_{\omega_i \in B_i} p_i(\omega_i) = \prod_{i=1}^n \mathbb{P}_i(B_i) = \prod_{i \in J} \mathbb{P}_i(A_i),$$

da ja $\mathbb{P}_i(B_i) = \mathbb{P}_i(\Omega_i) = 1$ für $i \notin J$. (Natürlich bezeichnet $\mathbb{P}_i$ das zu (Ω_i, p_i) gehörige Wahrscheinlichkeitsmaß.) Nun beachte man, dass gilt:

$$A_i^{(i)} = \Omega_1 \times \cdots \times \Omega_{i-1} \times A_i \times \Omega_{i+1} \times \cdots \times \Omega_n,$$

woraus folgt, dass $\mathbb{P}_i(A_i) = \mathbb{P}(A_i^{(i)})$. Dies beendet den Beweis. $\square$

Kapitel 3

Zufallsgrößen, Erwartungswerte und Varianzen

Wir wollen nicht nur die Wahrscheinlichkeiten von Ereignissen bestimmen, sondern auch zufällige Größen behandeln. In diesem Abschnitt präzisieren wir, was eine Zufallsgröße ist, und was wir unter ihrem Erwartungswert und ihrer Varianz verstehen wollen. Außerdem erläutern wir, was Unabhängigkeit von Zufallsgrößen ist.

Wir legen wieder dem gesamten Kapitel einen diskreten Wahrscheinlichkeitsraum (Ω, p) zu Grunde.

3.1 Zufallsgrößen

Definition 3.1.1 (Zufallsgröße)**.** *Jede Abbildung $X\colon \Omega \to \mathbb{R}$ heißt eine (reellwertige)* Zufallsgröße *oder* Zufallsvariable*.*

Beispiel 3.1.2. Die Augensumme bei zwei Würfen mit einem fairen Würfel ist die auf $\Omega = \{1, \dots, 6\}^2$ mit der Gleichverteilung definierte Zufallsvariable $X\colon \Omega \to \mathbb{R}$ mit $X(i,j) = i + j$. $\diamond$

Beispiel 3.1.3. Die Anzahl der Erfolge im Bernoulli-Experiment (also $\Omega = \{0,1\}^n$ mit $q(\omega) = p^{\sum_{i=1}^n \omega_i}(1-p)^{n-\sum_{i=1}^n \omega_i}$, wobei $p \in [0,1]$ ein Parameter ist; siehe Beispiel 1.3.3) ist die Zufallsvariable $X\colon \Omega \to \mathbb{R}$, gegeben durch $X(\omega) = \sum_{i=1}^n \omega_i$. $\diamond$

Im Umgang mit Zufallsgrößen haben sich einige sehr handliche Konventionen eingebürgert. Wir schreiben $X(\Omega)$ für die (höchstens abzählbare) Menge $\{X(\omega) \in \mathbb{R}\colon \omega \in \Omega\}$, das Bild von Ω unter X. Für eine Menge $A \subset \mathbb{R}$ ist die Menge $X^{-1}(A) = \{\omega \in \Omega\colon X(\omega) \in A\}$ (das Urbild von A unter X) das Ereignis 'X nimmt einen Wert in A an' oder 'X liegt in A'. Wir benutzen die Kurzschreibweisen

$$\begin{aligned}
\{X \in A\} &= X^{-1}(A) = \{\omega \in \Omega\colon X(\omega) \in A\},\\
\{X = z\} &= X^{-1}(\{z\}) = \{\omega \in \Omega\colon X(\omega) = z\},\\
\{X \leq z\} &= X^{-1}((-\infty, z]) = \{\omega \in \Omega\colon X(\omega) \leq z\}
\end{aligned}$$

und so weiter. Statt $\mathbb{P}(\{X \in A\})$ schreiben wir $\mathbb{P}(X \in A)$ etc. Mit $\mathbb{P}(X \in A, Y \in B)$ meinen wir immer $\mathbb{P}(\{X \in A\} \cap \{Y \in B\})$, d. h., das Komma steht immer für 'und' bzw. für den mengentheoretischen Schnitt.

Definition 3.1.4 (Verteilung einer Zufallsgröße)**.** *Sei X eine Zufallsgröße, dann ist das Paar $(X(\Omega), p_X)$ definiert durch $p_X(x) = \mathbb{P}(X = x)$, ein diskreter Wahrscheinlichkeitsraum. Das induzierte Wahrscheinlichkeitsmaß $\mathbb{P} \circ X^{-1}$, definiert durch $\mathbb{P} \circ X^{-1}(A) = \sum_{x \in A} p_X(x)$, erfüllt $\mathbb{P} \circ X^{-1}(A) = \mathbb{P}(X \in A)$ für alle $A \subset X(\Omega)$ und wird die* Verteilung von X *genannt.*

Bemerkung 3.1.5. Wenn man X^{-1} als Urbildoperator von der Potenzmenge von $X(\Omega)$ in die Potenzmenge von Ω auffasst, dann ist $\mathbb{P} \circ X^{-1}$ die Hintereinanderausführung der beiden Abbildungen $X^{-1}\colon \mathcal{P}(X(\Omega)) \to \mathcal{P}(\Omega)$ und $\mathbb{P}\colon \mathcal{P}(\Omega) \to [0, 1]$.

Wir können $\mathbb{P} \circ X^{-1}(A) = \mathbb{P}(X \in A)$ für *jede* Teilmenge von $\mathbb{R}$ betrachten und meinen damit $\mathbb{P} \circ X^{-1}(A \cap X(\Omega))$. $\diamond$

Je nachdem, ob das Wahrscheinlichkeitsmaß $\mathbb{P} \circ X^{-1}$ die Binomial-, geometrische oder hypergeometrische Verteilung ist, nennen wir die Zufallsgröße X binomialverteilt, geometrisch verteilt oder hypergeometrisch verteilt, analog für jede andere Verteilung. Insbesondere ist also die Anzahl der Erfolge in einem Bernoulli-Experiment binomialverteilt, und die Wartezeit auf den ersten Erfolg ist geometrisch verteilt. Letztere Verteilung besitzt eine interessante charakteristische Eigenschaft, die (bei oberflächlicher Betrachtung) der Intuition widerspricht: Wenn man eine große Anzahl von erfolglosen Spielen beobachtet hat, erwartet man vielleicht, dass die Erfolgswahrscheinlichkeit im nächsten Spiel größer sein sollte, denn im Durchschnitt sollten sich ja über lange Spielserien die Erfolge und Misserfolge im Verhältnis ihrer Wahrscheinlichkeiten ausgleichen.[1] Der nächste Satz zeigt, dass dies ein Trugschluss ist.

Lemma 3.1.6 (Gedächtnislosigkeit der geometrischen Verteilung)**.** *Sei X eine geometrisch auf $\mathbb{N}$ verteilte Zufallsgröße, und sei $n \in \mathbb{N}$. Dann ist für jedes $k \in \mathbb{N}$*

$$\mathbb{P}(X = k) = \mathbb{P}(X = n - 1 + k \mid X \geq n).$$

Insbesondere ist also, wenn man $n-1$ erfolglosen Spielen beigewohnt hat, die Wahrscheinlichkeit für den nächsten Erfolg nach genau k weiteren Spielen unabhängig von n, also unabhängig vom bisherigen Spielverlauf.

Beweis. Wir rechnen die rechte Seite explizit aus:

$$\begin{aligned}\mathbb{P}(X = n - 1 + k \mid X \geq n) &= \frac{\mathbb{P}(X = n - 1 + k, X \geq n)}{\mathbb{P}(X \geq n)} = \frac{\mathbb{P}(X = n - 1 + k)}{\sum_{l=n}^{\infty} \mathbb{P}(X = l)} \\ &= \frac{p(1-p)^{n+k-2}}{\sum_{l=n}^{\infty} p(1-p)^{l-1}} = \frac{(1-p)^{k-1}}{\sum_{l=1}^{\infty} (1-p)^{l-1}} = p(1-p)^{k-1} \\ &= \mathbb{P}(X = k).\end{aligned}$$

$\square$

[1] Dieser Gedanke wird in Satz 5.1.4 unter dem Namen 'Gesetz der Großen Zahlen' präzisiert werden.

3.2 Unabhängigkeit von Zufallsgrößen

Wir beginnen mit einer recht allgemeinen Definition der Unabhängigkeit von diskreten Zufallsgrößen.

Definition 3.2.1. *Sei $(X_i)_{i\in I}$ eine Familie von Zufallsgrößen $X_i\colon \Omega \to \mathbb{R}$, wobei I eine beliebige Indexmenge ist. Wir sagen, die Familie $(X_i)_{i\in I}$ ist* unabhängig *(oder auch, die X_i mit $i \in I$ seien* unabhängig*), wenn für jede Familie $(A_i)_{i\in I}$ von reellen Mengen $A_i \subset \mathbb{R}$ die Familie der Ereignisse $(\{X_i \in A_i\})_{i\in I}$ unabhängig ist.*

Tatsächlich lässt das Konzept von diskreten Wahrscheinlichkeitsräumen, das wir in diesem Skript behandeln, es nicht zu, mehr als endlich viele unabhängige Zufallsgrößen zu konstruieren. Die Existenz einer unendlich großen Familie von unabhängigen Zufallsgrößen erfordert die Theorie von Wahrscheinlichkeiten auf überabzählbar großen Mengen (die wir später behandeln werden), denn alleine wenn man schon abzählbar unendlich viele unabhängige Zufallsgrößen definieren möchte, die jeweils nur zwei verschiedene Werte annnehmen, sagen wir, 0 und 1, muss man das auf einer Menge tun, die die Menge $\{0,1\}^{\mathbb{N}}$ (die Menge aller unendlichen Folgen mit Koeffizienten in $\{0,1\}$) enthält, aber diese Menge ist überabzählbar. Im Folgenden werden wir also ausschließlich *endlich viele* unabhängige Zufallsgrößen betrachten.

Wir bringen zunächst eine Charakterisierung der Unabhängigkeit von endlich vielen Zufallsgrößen. Ausführlicher könnte man die Unabhängigkeit von Zufallsgrößen $X_1,\dots,X_n$ formulieren, indem man die Gültigkeit die Produktformel (2.2.1) für alle endlichen Teilmengen J von $\{1,\dots,n\}$ und alle Ereignisse der Form $\{X_i \in A_i\}$ fordert. Das folgende Lemma sagt, dass man sich zurück ziehen kann auf einelementige Mengen A_i und auf die Wahl $J = \{1,\dots,n\}$. Der Beweis zeigt, dass man nämlich alle anderen Ereignisse aus dieser speziellen Wahl kombinieren kann durch Vereinigungsbildung.

Lemma 3.2.2. *Zufallsgrößen $X_1,\dots,X_n$ sind genau dann unabhängig, wenn für alle $x_1 \in X_1(\Omega),\dots,x_n \in X_n(\Omega)$ gilt:*

$$\mathbb{P}(X_1 = x_1,\dots,X_n = x_n) = \prod_{i=1}^{n} \mathbb{P}(X_i = x_i).$$

Beweis.

'$\Longrightarrow$': Dies folgt aus einer Anwendung der Definition 2.2.1 auf die Teilmenge $J = \{1,\dots,n\}$ und die Ereignisse $\{X_i = x_i\}$.

'$\Longleftarrow$' Seien $A_1,\dots,A_n \subset \mathbb{R}$, und sei J eine beliebige nichtleere Teilmenge der Indexmenge $\{1,\dots,n\}$. Wir zeigen die Gültigkeit der Produktformel (2.2.1) für die Ereignisse $\{X_i \in A_i\}$ mit $i \in J$, wobei wir voraus setzen dürfen, dass $A_i \subset X_i(\Omega)$. Hierzu schreiben wir den Schnitt dieser Ereignisse als eine disjunkte Vereinigung von Schnitten von Ereignissen der Form $\{X_i = x_i\}$ mit $i \in \{1,\dots,n\}$. Auf diese Ereignisse können wir die Produktformel laut Voraussetzung anwenden.

Die Einführung der Notation $B_i = A_i$, falls $i \in J$, und $B_i = X_i(\Omega)$ sonst, vereinfacht die

Notation. Also gilt

$$
\begin{aligned}
\mathbb{P}\Big(\bigcap_{i\in J}\{X_i \in A_i\}\Big) &= \mathbb{P}\Big(\bigcap_{i=1}^n\{X_i \in B_i\}\Big) = \mathbb{P}\Big(\bigcup_{x_1\in B_1,\dots,x_n\in B_n}\bigcap_{i=1}^n\{X_i = x_i\}\Big)\\
&= \sum_{x_1\in B_1,\dots,x_n\in B_n}\mathbb{P}\Big(\bigcap_{i=1}^n\{X_i = x_i\}\Big) = \sum_{x_1\in B_1,\dots,x_n\in B_n}\prod_{i=1}^n\mathbb{P}(X_i = x_i)\\
&= \prod_{i=1}^n\sum_{x_i\in B_i}\mathbb{P}(X_i = x_i) = \prod_{i=1}^n\mathbb{P}(X_i \in B_i) = \prod_{i\in J}\mathbb{P}(X_i \in A_i).
\end{aligned}
$$

□

Beispiel 3.2.3 (Indikatorvariable)**.** Für ein beliebiges Ereignis A bezeichnen wir mit $\mathbb{1}_A$ die durch

$$
\mathbb{1}_A(\omega) = \begin{cases} 1, & \text{falls } \omega \in A,\\ 0, & \text{falls } \omega \notin A,\end{cases}
$$

definierte *Indikatorvariable* auf A. Es folgt leicht aus einer Kombination von Lemma 2.2.6 und Lemma 3.2.2, dass Ereignisse $A_1, \dots, A_n$ genau dann unabhängig sind, wenn die Indikatorvariablen $\mathbb{1}_{A_1}, \dots, \mathbb{1}_{A_n}$ unabhängig sind. ◇

Beispiel 3.2.4 (Bernoulli-Zufallsgrößen)**.** In einem oft auftretenden Spezialfall nehmen die unabhängigen Zufallsgrößen $X_1, \dots, X_n$ die Werte 1 und 0 jeweils mit Wahrscheinlichkeit p bzw. $1-p$ an. Man sagt, $X_1, \dots, X_n$ sind *Bernoulli-Zufallsgrößen*. Man kann diese Größen auf dem in Beispiel 1.3.3 (siehe auch Beispiele 2.2.3 und 3.1.3) eingeführten Wahrscheinlichkeitsraum (Ω, q) formal definieren, indem man $X_i(\omega) = \omega_i$ setzt, d. h., X_i ist 1, wenn das i-te Spiel einen Erfolg hatte. Also kann man Bernoulli-Zufallsgrößen auch auffassen als Indikatorvariable auf unabhängigen Ereignissen, die jeweils mit Wahrscheinlichkeit p eintreten. ◇

Bernoulli-Zufallsgrößen benutzt man oft, um gewisse Modelle der statistischen Physik zu definieren:

Beispiel 3.2.5 (Perkolation)**.** Ein Modell für die (zufällige) Durchlässigkeit eines porösen Materials erhält man auf folgende Weise. Wir sagen, zwei Punkte i und j im Gitter $\mathbb{Z}^d$ sind *benachbart*, und wir schreiben $i \sim j$, falls sich i und j nur in genau einer Komponenten unterscheiden, und zwar um 1 oder -1. Wir betrachten eine endliche Box $\Lambda \subset \mathbb{Z}^d$, die den Ursprung enthält. Jede Kante zwischen Nachbarn in Λ habe den Wert 'offen' mit Wahrscheinlichkeit $p \in (0,1)$ und den Wert 'geschlossen' sonst. Die Offenheit der Kanten sei unabhängig. Also haben wir eine Kollektion von Bernoulli-Zufallsgrößen $(X_{\{i,j\}})_{i,j\in\Lambda, i\sim j}$ mit Werten in einer zweielementigen Menge.

Wir denken uns eine Wasserquelle im Ursprung, und das Wasser kann frei entlang offener Kanten laufen, aber nicht entlang der anderen Kanten. Eine typische Frage, die man sich nun stellt, ist: 'Mit welcher Wahrscheinlichkeit gibt es an der Oberfläche von Λ feuchte Punkte?'. Besonders interessant ist diese Frage im Grenzübergang $\Lambda \uparrow \mathbb{Z}^d$, d. h., die Frage nach der Existenz eines offenen Weges nach Unendlich. (Man sagt in diesem Fall, dass die Flüssigkeit durchsickert, d. h. sie *perkoliert*.) Eine andere natürliche Frage ist die nach der erwarteten Anzahl (siehe den nächsten Abschnitt) von durchfeuchteten Zellen in dem Material, und im Fall $\Lambda \uparrow \mathbb{Z}^d$ die Frage, ob der feuchte Teil erwartungsgemäß endlich ist oder nicht.

Es ist klar, dass die Antworten auf diese Fragen nur von p abhängen, und man erwartet einen sogenannten *Phasenübergang* zwischen kleinen und großen Werten von p, d. h. einen kritischen Wert von p, an dem die Antworten drastisch umschlagen. $\diamond$

Ein weiteres Standardmodell der statistischen Physik wird mit Hilfe von abhängigen Zufallsgrößen konstruiert:

Beispiel 3.2.6 (Ising-Modell)**.** Ein Modell für die Magnetisierung eines Stücks Eisen in einem Magnetfeld wird folgendermaßen definiert. Wieder sei $\Lambda \subset \mathbb{Z}^d$ eine endliche Box, und wir definieren die Verteilung von $\{-1,1\}$-wertigen Zufallsgrößen X_i mit $i \in \Lambda$ wie folgt. (Wir interpretieren X_i als die Ladung im Punkt i.) Für jede Konfiguration $(x_i)_{i\in\Lambda} \in \{-1,1\}^\Lambda$ sei

$$\mathbb{P}\big((X_i)_{i\in\Lambda} = (x_i)_{i\in\Lambda}\big) = \frac{1}{Z_{\beta,h,\Lambda}} \exp\Big\{\beta \sum_{i,j\in\Lambda, i\sim j} x_i x_j + h \sum_{i\in\Lambda} x_i\Big\},$$

wobei $\beta, h > 0$ Parameter seien und $Z_{\beta,h,\Lambda}$ eine geeignete Normierungskonstante. Das Maß $\mathbb{P}$ favorisiert Konfigurationen $(x_i)_{i\in\Lambda}$ mit vielen Übergängen zwischen gleichen Ladungen und mit vielen positiven Ladungen. Die Parameter β bzw. h regeln die Stärke dieser Effekte; insbesondere ist h die Stärke des äußeren Magnetfeldes. $\diamond$

Im folgenden Beispiel wird eine Ahnung davon gegeben, dass man bei geometrisch verteilten Zufallsgrößen sehr viele Wahrscheinlichkeiten explizit ausrechen kann.

Beispiel 3.2.7 (unabhängige geometrische Zufallsgrößen)**.** Es seien X und Y zwei unabhängige, zum Parameter $p \in [0,1]$ geometrisch auf $\mathbb{N}_0$ verteilte Zufallsgrößen, d. h., wir haben $\mathbb{P}(X = k) = \mathbb{P}(Y = k) = (1-p)^k p$ für jedes $k \in \mathbb{N}_0$, und für alle $k, n \in \mathbb{N}_0$ gilt $\mathbb{P}(X = k, Y = n) = \mathbb{P}(X = k)\mathbb{P}(Y = n)$. Wir wollen die Wahrscheinlichkeiten gewisser mit X und Y beschriebener Ereignisse errechnen.

Als Beispiel berechnen wir $\mathbb{P}(X = Y)$, also die Wahrscheinlichkeit, dass X und Y den selben Wert annehmen. Da das Ereignis $\{X = Y\}$ die disjunkte Vereinigung der Ereignisse $\{X = k, Y = k\}$ mit $k \in \mathbb{N}_0$ ist, erhalten wir

$$\begin{aligned}\mathbb{P}(X = Y) &= \sum_{k\in\mathbb{N}_0} \mathbb{P}(X = k, Y = k) = \sum_{k\in\mathbb{N}_0} \mathbb{P}(X = k)\mathbb{P}(Y = k)\\ &= p^2 \sum_{k\in\mathbb{N}_0} (1-p)^k (1-p)^k = p^2 \sum_{k\in\mathbb{N}_0} [(1-p)^2]^k = p^2 \frac{1}{1-(1-p)^2}\\ &= \frac{p}{2-p}.\end{aligned}$$

Als Übungsaufgabe berechne man die Werte der Wahrscheinlichkeiten $\mathbb{P}(X \leq Y)$, $\mathbb{P}(X < Y)$ und $\mathbb{P}(X = Y = Z)$, wobei X, Y und Z drei geometrisch verteilte Zufallsgrößen sind. $\diamond$

Wie bei Unabhängigkeit von Ereignissen gibt es auch einen engen Zusammenhang zwischen Unabhängigkeit von Zufallsgrößen und Produkträumen. Im folgenden Lemma charakterisieren wir die Unabhängigkeit von Zufallsgrößen mit Hilfe der gemeinsamen Verteilung der Größen. Zunächst erklären wir, was man unter der gemeinsamen Verteilung von mehreren Zufallsgrößen versteht.

Definition 3.2.8 (Gemeinsame Verteilung, Randverteilung). *Es seien $X_1, \dots, X_n$ Zufallsgrößen. Dann verstehen wir unter der* gemeinsamen Verteilung *der $X_1, \dots, X_n$ die* Verteilung des Zufallsvektors $X = (X_1, \dots, X_n)$. *Dies ist (analog zu Definition 3.1.4) das Wahrscheinlichkeitsmaß $\mathbb{P} \circ X^{-1}$, das durch p_X induziert wird, und p_X wird auf der Bildmenge $X(\Omega) = \{(X_1(\omega), \dots, X_n(\omega)) : \omega \in \Omega\}$ definiert durch*

$$p_X(x_1, \dots, x_n) = \mathbb{P}\big(X = (x_1, \dots, x_n)\big) = \mathbb{P}(X_1 = x_1, \dots, X_n = x_n).$$

Die Verteilung der einzelnen Zufallsgrößen X_i erhält man, indem man die i-te Randverteilung *oder* Marginalverteilung *von p_X bildet, die gegeben ist durch*

$$\mathbb{P}(X_i = x_i) = p_{X_i}(x_i) = \sum_{x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n} p_X(x_1, \dots, x_n), \qquad x_i \in X_i(\Omega).$$

Beispiel 3.2.6 enthielt ein Beispiel für eine gemeinsame Verteilung von nicht unabhängigen Zufallsgrößen. Nun folgt der angekündigte Zusammenhang zwischen Unabhängigkeit und Produktverteilungen.

Lemma 3.2.9 (Zufallsvektoren mit unabhängigen Komponenten). *Es seien $X_1, \dots, X_n$ Zufallsgrößen. Dann sind $X_1, \dots, X_n$ genau dann unabhängig, wenn die Verteilung von X gleich der Produktverteilung der Verteilungen der $X_1, \dots, X_n$ ist.*

Beweis. Diese Aussage ist nur eine Umformulierung von Lemma 3.2.2, wie man sich leicht klar macht. □

In der Situation von Lemma 3.2.9 sagt man, der Zufallsvektor X habe unabhängige Komponenten $X_1, \dots, X_n$.

Die folgende Bemerkung ist analog zu Lemma 2.3.2.

Bemerkung 3.2.10. Falls die Zufallsgrößen $X_1, \dots, X_n$ auf eventuell verschiedenen Wahrscheinlichkeitsräumen $(\Omega_1, p_1), \dots, (\Omega_n, p_n)$ definiert sind, können wir sie auf dem Produktraum (Ω, p) der $(\Omega_1, p_1), \dots, (\Omega_n, p_n)$ (siehe Definition 2.3.1) in einer solchen Weise definieren, dass sie unabhängig sind. Dazu betrachten wir die Zufallsgrößen $X_1^{(1)}, \dots, X_n^{(n)}$, die durch $X_i^{(i)}((\omega_1, \dots, \omega_n)) = X_i(\omega_i)$ definiert sind. Mit Hilfe der kanonischen Projektion $\pi_i \colon \Omega \to \Omega_i$ können wir auch $X_i^{(i)} = X_i \circ \pi_i$ schreiben. Dann ist die Verteilung von $X_i^{(i)}$ (auf Ω) gleich der von X_i (auf Ω_i), wie man sich leicht klar macht. Ferner sind die Variablen $X_1^{(1)}, \dots, X_n^{(n)}$ unabhängig, denn für alle $x_1, \dots, x_n$ gilt

$$\begin{aligned}
\mathbb{P}(X_1^{(1)} = x_1, \dots, X_n^{(n)} = x_n) &= \sum_{\omega\in\Omega\colon X_i^{(i)}(\omega)=x_i\forall i} p(\omega_1, \dots, \omega_n) \\
&= \sum_{\omega_1\in\Omega_1\colon X_1(\omega_1)=x_1} \cdots \sum_{\omega_n\in\Omega_n\colon X_n(\omega_n)=x_n} \prod_{i=1}^n p_i(\omega_i) \\
&= \prod_{i=1}^n \sum_{\omega_i\in\Omega_i\colon X_i(\omega_i)=x_1} p_i(\omega_i) \\
&= \prod_{i=1}^n \mathbb{P}_i(X_i = x_i) = \prod_{i=1}^n \mathbb{P}(X_i^{(i)} = x_i).
\end{aligned}$$

Hierbei ist p in Definition 2.3.1 eingeführt worden, und $\mathbb{P}$ und $\mathbb{P}_i$ sind die zu p bzw. zu p_i gehörigen Wahrscheinlichkeitsmaße auf Ω bzw. auf Ω_i. $\diamond$

Aus Familien unabhängiger Zufallsgrößen kann man andere Familien unabhängiger Zufallsgrößen gewinnen, indem man sie auf disjunkte Weise irgendwie mit einander kombiniert:

Lemma 3.2.11 (Kombinationen unabhängiger Zufallsgrößen)**.** *Es sei $(X_i)_{i\in I}$ eine Familie unabhängiger Zufallsgrößen, wobei I eine beliebige Indexmenge sei. Ferner seien $I_1, I_2, \dots$ endliche, paarweise disjunkte Teilmengen von I, und für jedes $j \in \mathbb{N}$ sei Y_j eine Zufallsgröße, die aus den Zufallsvariablen X_i mit $i \in I_j$ gebildet sei, also $Y_j = f_j((X_i)_{i\in I_j})$ für eine geeignete Funktion f_j. Dann sind die Zufallsvariablen Y_j mit $j \in \mathbb{N}$ unabhängig.*

Beweis. (Wegen der notationellen Unhandlichkeit formulieren wir diesen Beweis nicht voll aus.)

Es seien $k \in \mathbb{N}$ und $y_1 \in Y_1(\Omega), \dots, y_k \in Y_k(\Omega)$, dann müssen wir die Gültigkeit der Produktformel $\mathbb{P}(Y_1 = y_1, \dots, Y_k = y_k) = \prod_{j=1}^k \mathbb{P}(Y_j = y_j)$ zeigen. Um dies zu tun, zerlegen wir das Ereignis $\{Y_j = y_j\}$ zunächst in eine disjunkte Vereinigung von Ereignissen, die mit Hilfe der X_i mit $i \in I_j$ gebildet werden:

$$\{Y_j = y_j\} = \bigcup_{(x_i)_{i\in I_j}\colon f_j\big((x_i)_{i\in I_j}\big)=y_j} \big\{(X_i)_{i\in I_j} = (x_i)_{i\in I_j}\big\}. \tag{3.2.1}$$

Nun erhält man einen Ausdruck für das Produkt $\prod_{j=1}^k \mathbb{P}(Y_j = y_j)$, indem man in (3.2.1) zu den Wahrscheinlichkeiten übergeht, die Summenformel für disjunkte Vereinigungen benutzt, über $j = 1, \dots, k$ multipliziert und die Wahrscheinlichkeiten der Ereignisse auf der rechten Seite von (3.2.1) mit Hilfe der Unabhängigkeit der X_i in Produkte von einzelnen Wahrscheinlichkeiten zerlegt.

Auf der anderen Seite erhält man analog zu (3.2.1) eine Darstellung des Ereignisses $\{Y_1 = y_1, \dots, Y_k = y_k\}$ als disjunkte Vereinigung von Ereignissen, die mit Hilfe von $(X_i)_{i\in I_j}$ für $j \in \{1, \dots, k\}$ gebildet werden. Die Wahrscheinlichkeit dieser Vereinigung errechnet man analog und stellt durch Vergleich mit dem Resultat der obigen Rechnung fest, dass sie identisch ist mit dem Produkt der Wahrscheinlichkeiten der Ereignisse $\{Y_j = y_j\}$. $\square$

Das folgende wichtige Korollar zu Lemma 3.2.11 erhält man durch die Wahl von einelementigen Teilindexmengen:

Korollar 3.2.12. *Falls die Zufallsvariablen X_i mit $i \in \mathbb{N}$ unabhängig sind, so auch die Zufallsvariablen $f_i(X_i)$ mit $i \in \mathbb{N}$, wobei f_i beliebige Funktionen sind.*

3.3 Erwartungswerte

Wir führen einen zentralen Begriff der Wahrscheinlichkeitstheorie ein: den Begriff des 'erwarteten Wertes' einer Zufallsgröße. Man erhält diesen Wert, indem man über alle Werte, die die Zufallsgröße annehmen kann, mittelt mit den Gewichten, die durch die Einzelwahrscheinlichkeiten gegeben sind.

Definition 3.3.1 (Erwartungswert). *Wir sagen, eine Zufallsgröße* $X\colon \Omega \to \mathbb{R}$ besitzt einen endlichen Erwartungswert*, falls die Reihe* $\sum_{\omega\in\Omega} p(\omega)|X(\omega)|$ *konvergiert. In diesem Fall schreiben wir auch* $X \in \mathcal{L}^1(\mathbb{P})$ *(oder auch einfach* $X \in \mathcal{L}^1$*) und definieren den* Erwartungswert *von* X *als die Zahl*

$$\mathbb{E}(X) = \sum_{\omega\in\Omega} p(\omega)X(\omega).$$

In Erweiterung von Definition 3.3.1 können wir auch für jede *nicht negative* Zufallsgröße X den Erwartungswert von X als $\mathbb{E}(X) = \sum_{\omega\in\Omega} p(\omega)X(\omega)$ definieren, auch wenn die Reihe divergiert. In letzterem Fall setzen wir $\mathbb{E}(X) = \infty$. In der selben Weise können wir auch Zufallsgrößen behandeln, die nach unten beschränkt sind oder nach oben beschränkt.

Die Voraussetzung der *absoluten* Konvergenz der Reihe $\sum_{\omega\in\Omega} p(\omega)X(\omega)$ sichert, dass der Wert dieser Reihe nicht von der gewählten Aufzählung von Ω abhängt. Außerdem impliziert sie etliche Rechenregeln, die man vom Begriff des Erwartungswertes erwartet:

Lemma 3.3.2 (Eigenschaften des Erwartungswertes). *(a) Eine Zufallsgröße* X *liegt genau dann in* $\mathcal{L}^1$*, wenn die Reihe* $\sum_{x\in X(\Omega)} |x|\mathbb{P}(X = x)$ *konvergiert. In diesem Fall gilt* $\mathbb{E}(X) = \sum_{x\in X(\Omega)} x\mathbb{P}(X = x)$*.*

(b) Falls $X, Y \in \mathcal{L}^1$ *mit* $X \le Y$*, so gilt* $\mathbb{E}(X) \le \mathbb{E}(Y)$*.* (Monotonie des Erwartungswertes)

(c) Falls $X, Y \in \mathcal{L}^1$ *und* $c \in \mathbb{R}$*, so ist* $X + cY \in \mathcal{L}^1$*, und es gilt* $\mathbb{E}(X + cY) = \mathbb{E}(X) + c\mathbb{E}(Y)$*.* (Linearität des Erwartungswertes)

(d) Falls $X, Y \in \mathcal{L}^1$ *unabhängig sind, so ist auch* $XY \in \mathcal{L}^1$*, und es gilt* $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$*.* (Produktregel bei Unabhängigkeit)

Beweis.

(a) Die erste Aussage wird gezeigt durch die Rechnung

$$\begin{aligned}\sum_{x\in X(\Omega)} |x|\mathbb{P}(X = x) &= \sum_{x\in X(\Omega)} |x| \sum_{\omega\colon X(\omega)=x} p(\omega) = \sum_{x\in X(\Omega)} \sum_{\omega\colon X(\omega)=x} |X(\omega)|p(\omega)\\ &= \sum_{\omega\in\Omega} |X(\omega)|p(\omega),\end{aligned}$$

und die zweite Aussage folgt aus einer Wiederholung dieser Rechnung ohne Betragstriche.

(b) folgt mit Hilfe von (a) aus den Regeln für absolut konvergente Reihen.

(c) folgt mit Hilfe von Definition 3.3.1 aus den Rechenregeln für absolut konvergente Reihen.

(*d*) Wir spalten auf nach allen Werten von X:

$$\begin{aligned}\sum_z |z|\mathbb{P}(XY=z) &= \sum_{z\neq 0}\sum_x |z|\mathbb{P}(XY=z, X=x) = \sum_{x,z\neq 0} |z|\mathbb{P}(X=x, Y=z/x)\\ &= \sum_{x,y\neq 0} |x|\,|y|\mathbb{P}(X=x, Y=y) = \sum_{x,y\neq 0} |x|\,|y|\mathbb{P}(X=x)\mathbb{P}(Y=y)\\ &= \sum_x |x|\mathbb{P}(X=x)\sum_y |y|\mathbb{P}(Y=y),\end{aligned}$$

wobei wir im vorletzten Schritt die Unabhängigkeit von X und Y benutzten. Dies zeigt, dass $XY \in \mathcal{L}^1$. Die behauptete Gleichung folgt durch eine Wiederholung der obigen Rechnung ohne Betragstriche. □

Für die Behandlung des Erwartungswertes von zusammengesetzten Zufallsgrößen ist das folgende Lemma nützlich. Wir erinnern an die Definition 3.2.8.

Lemma 3.3.3. *Es seien $X_1, \dots, X_n$ Zufallsgrößen, und es sei $g\colon X_1(\Omega)\times\cdots\times X_n(\Omega)\to\mathbb{R}$ eine Abbildung. Dann existiert der Erwartungswert der Zufallsgröße $Y = g(X_1,\dots,X_n) = g\circ(X_1,\dots,X_n)$ genau dann, wenn die Reihe*

$$\sum_{x_1\in X_1(\Omega)}\cdots\sum_{x_n\in X_n(\Omega)} g(x_1,\dots,x_n)\mathbb{P}(X_1=x_1,\dots,X_n=x_n)$$

absolut konvergiert, und der Wert der Reihe ist dann gleich $\mathbb{E}(Y)$.

Beweis. Wir machen $\Omega' = X_1(\Omega)\times\cdots\times X_n(\Omega)$ zu einem Wahrscheinlichkeitsraum (Ω', p'), indem wir $p'(x_1,\dots,x_n) = \mathbb{P}(X_1=x_1,\dots,X_n=x_n)$ setzen. Dann ist die Verteilung von g auf Ω' identisch mit der von Y auf Ω. Also folgt die Aussage aus Lemma 3.3.2(a). □

Nun bestimmen wir die Erwartungswerte einiger wichtiger Verteilungen.

Beispiel 3.3.4 (Erwartungswert der Binomialverteilung)**.** Sei X eine binomialverteilte Zufallsgröße, also $\mathbb{P}(X=k) = \mathrm{Bi}_{n,p}(k) = \binom{n}{k}p^k(1-p)^{n-k}$ für $k\in\{0,\dots,n\}$, wobei $n\in\mathbb{N}_0$ und $p\in[0,1]$. Es ist durchaus möglich, den Erwartungswert von X mit Hilfe von Lemma 3.3.2(a) zu berechnen (Übungsaufgabe: Man führe dies durch), doch wir schlagen einen weniger rechnenaufwändigen Weg ein, indem wir Lemma 3.3.2(c) benutzen. Wir erinnern uns (siehe Beispiele 3.1.3 und 3.2.4), dass wir X auf dem Raum $\Omega=\{0,1\}^n$ mit $q(\omega) = p^{\sum_{i=1}^n \omega_i}(1-p)^{n-\sum_{i=1}^n\omega_i}$ definieren können, indem wir $X(\omega)=\sum_{i=1}^n\omega_i$ setzen. Mit anderen Worten, wir haben $X=\sum_{i=1}^n X_i$ mit den in Beispiel 3.2.4 definieren Bernoulli-Größen. Jedes X_i nimmt die Werte 1 und 0 mit Wahrscheinlichkeit p bzw. $1-p$ an, also sieht man leicht, dass $\mathbb{E}(X_i) = 1\cdot p + 0\cdot(1-p) = p$ ist. Mit Hilfe der Linearität des Erwartungswertes erhält man also leicht, dass $\mathbb{E}(X)=\sum_{i=1}^n\mathbb{E}(X_i) = np$. ◇

Beispiel 3.3.5 (Erwartungswert der geometrischen Verteilung)**.** Es sei X eine zum Parameter $p\in(0,1)$ auf $\mathbb{N}$ geometrisch verteilte Zufallsgröße, also $\mathbb{P}(X=k)=\mathrm{Geo}_p(k) = p(1-p)^{k-1}$ für $k\in\mathbb{N}$. Zur Berechnung des Erwartungswertes von X (und für viele andere Berechnungen für geometrisch verteilte Zufallsgrößen) ist der folgende Trick sehr hilfreich. Eine gliedweise Differenziation der Identität $\sum_{k=1}^\infty x^k = \frac{x}{1-x}$ nach x für $|x|<1$ liefert die Gleichung $\sum_{k=1}^\infty kx^{k-1} = \frac{1}{(1-x)^2}$. Eine Anwendung auf $x=1-p$ liefert dann mit Hilfe von Lemma 3.3.2(a):

$$\mathbb{E}(X) = \sum_{k=1}^\infty k\mathbb{P}(X=k) = p\sum_{k=1}^\infty k(1-p)^{k-1} = p\frac{1}{p^2} = \frac{1}{p}.$$

Eine auf $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$ geometrisch verteilte Zufallsgröße hat die Verteilung von $X - 1$, also den Erwartungswert $\frac{1}{p} - 1$. ◇

Beispiel 3.3.6 (Erwartungswert der Poisson-Verteilung). Sei X eine zum Parameter $\alpha > 0$ Poisson-verteilte Zufallsgröße, also $\mathbb{P}(X = k) = \frac{\alpha^k}{k!} e^{-\alpha}$ für $k \in \mathbb{N}_0$. Der Erwartungswert von X kann leicht mit Lemma 3.3.2(a) berechnet werden:

$$\mathbb{E}(X) = \sum_{k \in \mathbb{N}_0} k\mathbb{P}(X = k) = e^{-\alpha} \sum_{k=1}^{\infty} \frac{\alpha^k}{(k-1)!} = \alpha e^{-\alpha} \sum_{k=0}^{\infty} \frac{\alpha^k}{k!} = \alpha,$$

wobei wir im vorletzten Schritt eine Laufindexverschiebung durchführten. ◇

Beispiel 3.3.7. Wir erinnern an die in Beispiel 2.2.8 eingeführte Verteilung auf $\mathbb{N}$, die durch $q(k) = k^{-s}\zeta(s)^{-1}$ gegeben ist, wobei $s > 1$. Da die Reihe $\sum_{k \in \mathbb{N}} kq(k)$ genau dann konvergiert, wenn $s > 2$, existiert der Erwartungswert dieser Verteilung also nur im Fall $s > 2$. Ihr Wert ist dann $\zeta(s-1)/\zeta(s)$. ◇

Die folgende Formel ist manchmal hilfreich, wenn eine $\mathbb{N}_0$-wertige Zufallsgröße X nur durch Angabe ihrer ‘Schwänze’ $\mathbb{P}(X > k)$ gegeben ist. (Letzteres ist manchmal einfacher, weil man bei dieser Festlegung der Verteilung von X nicht auf Normierung achten muss, sondern nur auf Monotonie.)

Lemma 3.3.8. *Der Erwartungswert einer $\mathbb{N}_0$-wertigen Zufallsgröße X (egal, ob er endlich ist oder nicht) ist gegeben durch*

$$\mathbb{E}(X) = \sum_{k=0}^{\infty} \mathbb{P}(X > k).$$

Beweis. Übungsaufgabe. □

3.4 Varianzen

Die Kenngröße ‘Erwartungswert’ gibt natürlich bei weitem keine erschöpfende Information über die Zufallsgröße. Eine nützliche zusätzliche Information ist zum Beispiel die Antwort auf die Frage, wie stark die Zufallsgröße im Durchschnitt von ihrem Erwartungswert abweicht. Eine Kenngröße, die dies angibt, ist die Varianz.

Definition 3.4.1. *Es sei X eine Zufallsgröße mit existierendem Erwartungswert $\mathbb{E}(X)$ sowie mit $\mathbb{E}(X^2) < \infty$. Die* Varianz *von X ist der Ausdruck*

$$\mathbb{V}(X) = \sum_x (x - \mathbb{E}(X))^2 \mathbb{P}(X = x) \in [0, \infty),$$

und die Standardabweichung *von X als $S(X) = \sqrt{\mathbb{V}(X)}$.*

Man sieht leicht mit Hilfe von Lemma 3.3.2(a) ein, dass $\mathbb{V}(X)$ der Erwartungswert der Zufallsgröße $\omega \mapsto (X(\omega) - \mathbb{E}(X))^2$ ist, also

$$\mathbb{V}(X) = \mathbb{E}\big((X - \mathbb{E}(X))^2\big).$$

Beispiel 3.4.2 (Varianz einer Gleichverteilung)**.** Wenn eine Zufallsgröße X auf einer n-elementigen Menge $\{x_1, \dots, x_n\}$ gleichverteilt ist, so sind

$$\mathbb{E}(X) = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{und} \qquad \mathbb{V}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mathbb{E}(X))^2.$$

Also ist $\mathbb{E}(X)$ der Mittelwert der $x_1, \dots, x_n$, und $\mathbb{V}(X)$ ist die mittlere quadratische Abweichung davon. $\diamond$

Beispiel 3.4.3 (Varianz der Bernoulli-Verteilung)**.** Eine Bernoulli-verteilte Zufallsgröße X nimmt die Werte 1 und 0 mit Wahrscheinlichkeit $p \in [0,1]$ bzw. $(1-p)$ an. Also ist $\mathbb{E}(X) = p$, und die Varianz berechnet sich nach Definition 3.4.1 zu $\mathbb{V}(X) = (0-\mathbb{E}(X))^2\mathbb{P}(X = 0) + (1-\mathbb{E}(X))^2\mathbb{P}(X = 1) = p^2(1-p) + (1-p)^2 p = p(1-p)$. $\diamond$

Die Varianz einer binomialverteilten Zufallsgröße kann man durchaus unter Verwendung der Definition 3.4.1 direkt berechnen, aber in Beispiel 3.5.4 werden wir einen eleganten Weg präsentieren.

Einige einfache Eigenschaften der Varianz sind im folgenden Lemma aufgelistet.

Lemma 3.4.4. *Es seien $X, Y \in \mathcal{L}^1$.*

(a) Die Varianz von X existiert genau dann, wenn $\mathbb{E}(X^2) < \infty$. In diesem Fall schreiben wir $X \in \mathcal{L}^2$, und es gilt die Formel

$$\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2.$$

(b) Seien $a, b \in \mathbb{R}$. Falls die Varianz von X existiert, dann auch die von $a + bX$, und es gilt $\mathbb{V}(a + bX) = b^2\mathbb{V}(X)$.

(c) Falls X und Y unabhängig sind mit existierenden Varianzen, dann existiert auch die Varianz von $X + Y$, und es gilt $\mathbb{V}(X + Y) = \mathbb{V}(X) + \mathbb{V}(Y)$. (Satz von Bienaymé)

(d) Falls $\mathbb{V}(X)$ existiert und $\mathbb{V}(X) = 0$ ist, so existiert ein $x \in \mathbb{R}$ mit $\mathbb{P}(X = x) = 1$.

Beweis.

(a): Wegen

$$(x - \mathbb{E}(X))^2\mathbb{P}(X = x) = x^2\mathbb{P}(X = x) - 2x\mathbb{E}(X)\mathbb{P}(X = x) + \mathbb{E}(X)^2\mathbb{P}(X = x)$$

ist die erste Behauptung klar, denn die absolute Konvergenz von $\sum_x x\mathbb{P}(X = x)$ ist vorausgesetzt und die von $\sum_x \mathbb{P}(X = x)$ ist klar. Die behauptete Gleichung erechnet man leicht, indem man die obige Beziehung über x summiert und zusammenfasst.

(b): Dies errechnet sich leicht mit Hilfe von (a) und der Linearität des Erwartungswertes.

(c): Wir benutzen (a) für $X + Y$ und multiplizieren aus und erhalten

$$\mathbb{V}(X + Y) = \mathbb{E}(X^2) + 2\mathbb{E}(XY) + \mathbb{E}(Y^2) - \mathbb{E}(X)^2 - 2\mathbb{E}(X)\mathbb{E}(Y) - \mathbb{E}(Y)^2.$$

Nun sehen wir mit Hilfe von Lemma 3.3.2(d), dass die rechte Seite gleich $\mathbb{E}(X^2) - \mathbb{E}(X)^2 + \mathbb{E}(Y^2) - \mathbb{E}(Y)^2$ ist, nach (a) also gleich $\mathbb{V}(X) + \mathbb{V}(Y)$.

(d): Dies sieht man leicht aus der Definition 3.4.1: Falls $\mathbb{V}(X) = 0$, so muss für jedes $x \in \mathbb{R}$ entweder $x = \mathbb{E}(X)$ oder $\mathbb{P}(X = x) = 0$ sein.

$\square$

Beispiel 3.4.5 (Varianz der Poisson-Verteilung)**.** Die Varianz einer zum Parameter $\alpha > 0$ Poisson-verteilten Zufallsgröße X (siehe Beispiel 3.3.6) ist $\mathbb{V}(X) = \alpha$. (Übungsaufgabe) ◇

Beispiel 3.4.6 (Varianz der geometrischen Verteilung)**.** Als Übungsaufgabe berechne man die Varianz einer geometrisch verteilten Zufallsgröße. (Man verwende den in Beispiel 3.3.5 erläuterten Trick ein zweites Mal.) ◇

Die Varianz bzw. der Erwartungswert besitzt die folgende Minimaleigenschaft:

Lemma 3.4.7 (Minimale quadratische Abweichung)**.** *Für jede Zufallsgröße $X \in \mathcal{L}^2$ gilt die Abschätzung*

$$\mathbb{E}\big((X-a)^2\big) \geq \mathbb{V}(X), \qquad a \in \mathbb{R},$$

mit Gleichheit genau dann, wenn $a = \mathbb{E}(X)$.

Beweis. Mit Hilfe der Linearität des Erwartungswerts errechnet man leicht, dass $\mathbb{E}((X-a)^2) = \mathbb{V}(X) + (a - \mathbb{E}(X))^2$ für jedes $a \in \mathbb{R}$. Also ist die Aussage evident. □

3.5 Kovarianzen

In diesem Abschnitt stellen wir eine Kenngröße vor, die über die Abhängigkeiten zweier gegebener Zufallsgrößen eine gewisse Aussage macht.

Definition 3.5.1. *Es seien X und Y zwei Zufallsgrößen mit existierenden Varianzen. Die* Kovarianz *von X und Y ist die Zahl $\operatorname{cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$. Wir nennen X und Y* unkorreliert*, falls $\operatorname{cov}(X,Y) = 0$.*

Die Kovarianz ist wohldefiniert, denn der Erwartungswert von XY existiert auf Grund der Abschätzung $2|XY| \leq X^2 + Y^2$ und Lemma 3.4.4(a).

Einige evidente Eigenschaften der Kovarianz werden nun gesammelt:

Lemma 3.5.2. *(a) Für je zwei Zufallsgrößen X und Y mit existierenden Varianzen gilt*

$$\operatorname{cov}(X,Y) = \mathbb{E}\big((X-\mathbb{E}(X))(Y-\mathbb{E}(Y))\big).$$

Insbesondere gelten $\operatorname{cov}(X,X) = \mathbb{V}(X)$ und $\operatorname{cov}(X,Y) = \operatorname{cov}(Y,X)$, und $\operatorname{cov}(\cdot,\cdot)$ ist linear in jedem der beiden Argumente.

(b) Für je n Zufallsgrößen $X_1, \dots, X_n$ gilt

$$\mathbb{V}\Big(\sum_{i=1}^n X_i\Big) = \sum_{i=1}^n \mathbb{V}(X_i) + \sum_{\substack{i,j=1\\ i\neq j}}^n \operatorname{cov}(X_i, X_j).$$

(c) Falls X und Y unabhängige Zufallsgrößen mit existierenden Varianzen sind, so sind X und Y unkorreliert.

Beweis. Übungsaufgabe. □

Aus der ersten Beziehung in Lemma 3.5.2(a) folgt auch, dass sich die Kovarianz von X und Y nicht ändert, wenn zu X oder Y Konstanten addiert werden.

Die Aussage in Lemma 3.5.2(c) kann nicht ohne Weiteres umgekehrt werden, wie das folgende Beispiel zeigt. Wir betrachten auf $\Omega = \{-1, 0, 1\}$ mit der Gleichverteilung die Zufallsgröße X, gegeben durch $X(\omega) = \omega$. Als Übungsaufgabe vergewissere man sich, dass die beiden Zufallsgrößen X und $|X|$ zwar unkorreliert, aber nicht unabhängig sind.

Eine direkte Folgerung aus Lemma 3.5.2(b) ist die folgende Aussage.

Korollar 3.5.3 (Satz von Bienaymé). *Für paarweise unkorrelierte Zufallsgrößen $X_1, \dots, X_n$ mit existierenden Varianzen gilt $\mathbb{V}(X_1 + \cdots + X_n) = \mathbb{V}(X_1) + \cdots + \mathbb{V}(X_n)$.*

Beispiel 3.5.4 (Varianz der Binomialverteilung). Im Kontext von Beispiel 3.3.4 wissen wir schon aus Beispiel 3.4.3, dass $\mathbb{V}(X_i) = p(1-p)$. Da die Zufallsgrößen $X_1, \dots, X_n$ unabhängig sind, erlaubt der Satz von Bienaymé, die Varianz von X wie folgt zu berechen: $\mathbb{V}(X) = \mathbb{V}(X_1 + \cdots + X_n) = \mathbb{V}(X_1) + \cdots + \mathbb{V}(X_n) = np(1-p)$. ◇

Die folgende Minimaleigenschaft der Kovarianz ist manchmal hilfreich, wenn man eine (eventuell schwer zugängliche) Zufallsvariable Y mit Hilfe einer linearen Funktion einer (leichter zu erhaltenden) Zufallsvariablen X annähern möchte.

Lemma 3.5.5 (Beste lineare Vorhersage). *Es seien $X, Y \in \mathcal{L}^2$ mit $\mathbb{V}(X) = 1$. Dann wird die quadratische Abweichung*

$$\mathbb{E}\big((Y - a - bX)^2\big)$$

zwischen Y und der linearen Funktion $a + bX$ von X minimiert genau für $b = \operatorname{cov}(X, Y)$ und $a = \mathbb{E}(Y - bX)$. Falls insbesondere X und Y unkorreliert sind, so hängt die Lösung $b = 0$ und $a = \mathbb{E}(Y)$ nicht von X ab.

Beweis. Eine Ausnutzung der Linearität des Erwartungswertes zeigt, dass der betrachtete Ausdruck ein Polynom zweiter Ordnung in a und b ist, das im Unendlichen gegen Unendlich geht und daher sein Minimum an der Nullstelle des Gradienten annimmt. Diese Nullstelle ist durch die behaupteten Gleichungen bestimmt. □

Eine der wichtigsten Ungleichungen ist die folgende.

Satz 3.5.6 (Cauchy-Schwarz-Ungleichung). *Für je zwei Zufallsgrößen X und Y gilt*

$$|\mathbb{E}(XY)| \leq \sqrt{\mathbb{E}(X^2)}\,\sqrt{\mathbb{E}(Y^2)}.$$

Gleichheit gilt genau dann, wenn es reelle Zahlen a, b gibt, die nicht beide Null sind, sodass $\mathbb{P}(aX + bY = 0) = 1$, d. h. wenn X und Y konstante Vielfache von einander sind.

Beweis. Wir setzen $\alpha = \mathbb{E}(Y^2)$ und $\beta = \mathbb{E}(XY)$. Wir dürfen annehmen, dass $\alpha > 0$, denn sonst wäre $\mathbb{P}(Y = 0) = 1$, also auch $\mathbb{E}(XY) = 0$, und die behauptete Ungleichung stimmt trivialerweise.

Nun errechnen wir durch Ausmultiplizieren und mit Hilfe der Linearität des Erwartungswertes:

$$\begin{aligned} 0 \le \mathbb{E}((\alpha X - \beta Y)^2) &= \alpha^2 \mathbb{E}(X^2) - 2\alpha\beta\, \mathbb{E}(XY) + \beta^2 \mathbb{E}(Y^2) \\ &= \alpha\big(\mathbb{E}(X^2)\mathbb{E}(Y^2) - \mathbb{E}(XY)^2\big). \end{aligned}$$

Da $\alpha > 0$, folgt die behauptete Ungleichung.

Falls Gleichheit gilt, so ist der Erwartungswert von $(\alpha X - \beta Y)^2$ gleich Null, also folgt $\mathbb{P}(\alpha X - \beta Y = 0) = 1$. Falls $\alpha > 0$, so können wir $\alpha = a$ und $\beta = b$ wählen. Falls $\alpha = 0$, so können wir $a = 0$ und $b = 1$ wählen. □

Aus einer Anwendung der Cauchy-Schwarz-Ungleichung auf die Zufallsgrößen $X - \mathbb{E}(X)$ und $Y - \mathbb{E}(Y)$ folgt insbesondere die Ungleichung

$$-1 \le \frac{\operatorname{cov}(X,Y)}{S(X)S(Y)} \le 1$$

für alle Zufallsgrößen X und Y, deren Varianzen existieren, wobei wir daran erinnern, dass $S(X) = \sqrt{\mathbb{V}(X)}$ die Standardabweichung von X ist.

Die Bedingung $\operatorname{cov}(X,Y) > 0$ bedeutet, dass eine Tendenz vorliegt, nach der das Ereignis $\{X > \mathbb{E}(X)\}$ öfter mit dem Ereignis $\{Y > \mathbb{E}(Y)\}$ zusammen auftritt als mit $\{Y \le \mathbb{E}(Y)\}$. Man sagt, X und Y seien *positiv korreliert*. Das impliziert allerdings noch lange nicht, dass X eine (Mit-)Ursache für Y ist oder umgekehrt, auch wenn dieser Fehlschluss oft und gerne gemacht wird.

3.6 Faltungen

Wenn X und Y zwei unabhängige Zufallsgrößen sind, was ist dann die Verteilung der Summe $X + Y$? Wir geben eine Antwort durch Summation über alle Werte, die X annehmen kann. Die Ausnutzung der Unabhängigkeit führt uns auf natürliche Weise zum Begriff der Faltung. In diesem Abschnitt werden wir nur $\mathbb{Z}$-wertige Zufallsgrößen betrachten. Daher ist die Menge $\mathbb{Z}$ der natürliche Indexbereich der von uns betrachteten Folgen.

Definition 3.6.1. *Die* Faltung *zweier absolut summierbarer Folgen $a = (a_x)_{x\in\mathbb{Z}}$ und $b = (b_y)_{y\in\mathbb{Z}}$ ist die Folge $c = (c_z)_{z\in\mathbb{Z}}$, die gegeben ist durch $c_z = \sum_{x\in\mathbb{Z}} a_x b_{z-x}$. Wir schreiben $c = a \star b$.*

Man sieht leicht, dass $a \star b = b \star a$ und dass $a \star b$ eine absolut summierbare Folge ist, wenn a und b dies sind.

Satz 3.6.2 (Faltungssatz)**.** *Seien X und Y zwei unabhängige $\mathbb{Z}$-wertige Zufallsgrößen mit Verteilungen p_X und p_Y, d. h. $p_X(x) = \mathbb{P}(X = x)$ und $p_Y(y) = \mathbb{P}(Y = y)$ für alle $x, y \in \mathbb{Z}$. Dann ist die Verteilung von $X + Y$ gleich der Faltung $p_X \star p_Y$, d. h. $\mathbb{P}(X + Y = z) = (p_X \star p_Y)(z)$ für alle $z \in \mathbb{Z}$.*

Beweis. Wir summieren über alle Werte, die X annehmen kann, und erhalten für alle $z \in \mathbb{Z}$:

$$\begin{aligned}\mathbb{P}(X+Y=z) &= \sum_{x\in\mathbb{Z}} \mathbb{P}(X=x, Y=z-x) = \sum_{x\in\mathbb{Z}} \mathbb{P}(X=x)\mathbb{P}(Y=z-x)\\ &= \sum_{x\in\mathbb{Z}} p_X(x)p_Y(z-x) = (p_X \star p_Y)(z).\end{aligned}$$

□

Beispiel 3.6.3 (Binomialverteilung)**.** Die Summe zweier unabhängiger binomialverteiler Zufallsgrößen zu den Parametern n_1 und p bzw. n_2 und p ist eine zu den Parametern $n_1 + n_2$ und p binomialverteilte Zufallsgröße. Dies sieht man am einfachsten ein, indem man die beiden Zufallsgrößen jeweils als Summe von n_1 bzw. n_2 unabhängigen Bernoulli-Zufallsgrößen darstellt. Man sagt, die Familie der zum Parameter n binomialverteilten Zufallsgrößen (mit festem zweiten Parameter p) bildet eine *Faltungshalbgruppe*.

Als eine Anwendung von Satz 3.6.2 beweisen wir die eingangs erwähnte Aussage noch einmal, indem wir zeigen, dass $\mathrm{Bi}_{n_1,p} \star \mathrm{Bi}_{n_2,p} = \mathrm{Bi}_{n_1+n_2,p}$ gilt, wobei wir die Folge $\mathrm{Bi}_{n,p}$ (die ja nur auf $\{0,\dots,n\}$ definiert ist) trivial mit Null zu einer Folge mit Indexmenge $\mathbb{Z}$ fortsetzen.

Es sei also $k \in \{0,\dots,n_1+n_2\}$, dann gilt

$$\begin{aligned}\mathrm{Bi}_{n_1,p} \star \mathrm{Bi}_{n_2,p}(k) &= \sum_{l\in\mathbb{Z}} \mathrm{Bi}_{n_1,p}(l)\mathrm{Bi}_{n_2,p}(k-l)\\ &= \sum_{l=\max\{0,k-n_2\}}^{\min\{n_1,k\}} \binom{n_1}{l} p^l(1-p)^{n_1-l}\binom{n_2}{k-l}p^{k-l}(1-p)^{n_2-k+l}\\ &= \mathrm{Bi}_{n_1+n_2,p}(k) \sum_{l=\max\{0,k-n_2\}}^{\min\{n_1,k\}} \frac{\binom{n_1}{l}\binom{n_2}{k-l}}{\binom{n_1+n_2}{k}},\end{aligned}$$

wie eine direkte Rechnung ergibt. Nun beachte man, dass der Quotient hinter dem Summenzeichen die hypergeometrische Wahrscheinlichkeit mit Parametern n_1, n_2 und k ist, ausgewertet in l. Da der Summationsbereich genau derjenige ist, auf der diese Verteilung definiert ist, ist also die Summe gleich Eins. Dies beweist die Behauptung $\mathrm{Bi}_{n_1,p} \star \mathrm{Bi}_{n_2,p} = \mathrm{Bi}_{n_1+n_2,p}$ auf $\{0,\dots,n_1+n_2\}$, und man sieht leicht, dass sie trivialerweise auch in $\mathbb{Z} \setminus \{0,\dots,n_1+n_2\}$ erfüllt ist. ◇

Beispiel 3.6.4 (Negative Binomialverteilung)**.** Es seien $X_1,\dots,X_n$ unabhängige, zum Parameter $p \in (0,1)$ auf $\mathbb{N}_0$ geometrisch verteilte Zufallsgrößen (siehe Beispiel 1.3.5), also $\mathbb{P}(X_i = k) = p(1-p)^k$ für $k \in \mathbb{N}_0$. Wir setzen $X = X_1 + \dots + X_n$, also hat X die Verteilung $\widetilde{\mathrm{Geo}}_p^{\star n}$, womit wir die n-fache Faltung der geometrischen Verteilung auf $\mathbb{N}_0$ bezeichnen.

Wir behaupten, dass die Verteilung von X identifiziert wird als

$$\mathbb{P}(X=k) = \binom{n+k-1}{k} p^n(1-p)^k = \binom{-n}{k} p^n(p-1)^k = \mathrm{Neg}_{n,p}(k), \qquad k \in \mathbb{N}_0, \qquad (3.6.1)$$

wobei

$$\binom{-n}{k} = \frac{(-n)(-n-1)(-n-2)\dots(-n-k+1)}{k!}, \qquad n \in (0,\infty),$$

der allgemeine Binomialkoeffizient ist. Die Verteilung $\mathrm{Neg}_{n,p}$ ist unter dem Namen *negative Binomialverteilung* zu den Parametern p und n bekannt. Sie kann ohne Probleme auch für beliebiges

$n \in (0, \infty)$ definiert werden, besitzt aber die Interpretation als Summe von geometrisch verteilten Zufallsgrößen nur für natürliche Zahlen n. Insbesondere ist $\mathrm{Neg}_{1,p} = \widetilde{\mathrm{Geo}}_p$.

Wir bieten nun zwei Wege an, die Faltungsformel

$$\widetilde{\mathrm{Geo}}_p^{\star n} = \mathrm{Neg}_{n,p}$$

für $n \in \mathbb{N}$ zu zeigen, ein dritter Weg (der sogar für alle $n \in (0, \infty)$ funktioniert), folgt in Beispiel 3.7.11. Der erste Weg macht Gebrauch von der oben erwähnten Interpretation von $X_i + 1$ als die Wartezeit nach dem $(i-1)$-ten bis zum i-ten Erfolg in einem unendlich langen Bernoulli-Experiment. Dann ist also $X + n$ die Wartezeit auf den n-ten Erfolg, gerechnet ab dem Beginn der Serie. Also ist das Ereignis $\{X = k\} = \{X + n = k + n\}$ das Ereignis, dass unter den ersten $k + n - 1$ Spielen genau $n - 1$ Erfolge und k Misserfolge sind und dass das $(n + k)$-te Spiel erfolgreich ist. Jede einzelne dieser Serien der Länge $n + k$ hat die Wahrscheinlichkeit $p^n(1-p)^k$, und es gibt genau $\binom{n-1+k}{k}$ solche Serien. Dies zeigt, dass die Verteilung von X tatsächlich durch die Formel in (3.6.1) gegeben ist.

Der zweite Weg benutzt den Faltungssatz und eine kombinatorische Überlegung. Man errechnet leicht, dass für $n_1, n_2 \in \mathbb{N}$ gilt:

$$\begin{aligned}
\mathrm{Neg}_{n_1,p} \star \mathrm{Neg}_{n_2,p}(k) &= \sum_{l \in \mathbb{Z}} \mathrm{Neg}_{n_1,p}(l)\mathrm{Neg}_{n_2,p}(k-l) \\
&= \sum_{l=0}^{k} \binom{n_1 - 1 + l}{l} p^{n_1}(1-p)^l \binom{n_2 - 1 + k - l}{k - l} p^{n_2}(1-p)^{k-l} \\
&= \mathrm{Neg}_{n_1+n_2,p}(k) \sum_{l=0}^{k} \frac{\binom{n_1-1+l}{l}\binom{n_2-1+k-l}{k-l}}{\binom{n_1+n_2-1+k}{k}}.
\end{aligned}$$

Dass die Summe über l den Wert Eins hat, sieht man folgendermaßen ein. Die Zahl $\binom{n_1-1+l}{l}$ ist die Anzahl der Möglichkeiten, n_1 Einsen und l Nullen hinter einander in eine Reihe zu legen, so dass die Reihe mit einer Eins endet. Analog ist $\binom{n_2-1+k-l}{k-l}$ die Anzahl der Möglichkeiten, n_2 Einsen und $k - l$ Nullen in eine Reihe zu legen, so dass sie mit einer Eins endet. Wenn man je eine solche Reihe hintereinander legt für irgendein $l \in \{0, \dots, k\}$, so erhält man eine Reihe mit $n_1 + n_2$ Einsen und k Nullen, die mit einer Eins endet. Anders herum kann man jede solche Reihe eindeutig aufspalten in zwei Reihen mit n_1 bzw. n_2 Einsen und l bzw. $k - l$ Nullen für ein geeignetes $l \in \{0, \dots, k\}$, so dass diese beiden Teilreihen jeweils mit einer Eins enden. Dies zeigt auf kombinatorische Weise, dass $\sum_{l=0}^{k} \binom{n_1-1+l}{l}\binom{n_2-1+k-l}{k-l} = \binom{n_1+n_2-1+k}{k}$ für $n_1, n_2 \in \mathbb{N}_0$. $\diamond$

Beispiel 3.6.5 (Poisson-Verteilung)**.** Eine weitere Faltungshalbgruppe ist die der Poisson-Verteilungen: Die Summe zweier unabhängiger zum Parameter $\alpha > 0$ bzw. $\beta > 0$ Poisson-verteilter Zufallsgrößen ist eine zum Parameter $\alpha + \beta$ Poisson-verteilte Zufallsgröße. Den Beweis führt man wiederum mit Hilfe von Satz 3.6.2 (Übungsaufgabe). Ein eleganterer Beweis folgt in Beispiel 3.7.12. $\diamond$

3.7 Erzeugende Funktionen

In diesem Abschnitt betrachten wir ausschließlich Verteilungen auf $\mathbb{N}_0 = \{0, 1, 2, \dots, \}$ und kombinieren ihre Betrachtung mit der gewisser Potenzreihen. Auf diesem Wege setzen wir einige bekannte Ergebnisse aus der Analysis für die Beschreibung von Zufallsgrößen ein.

Definition 3.7.1 (erzeugende Funktion)**.** *Die* erzeugende Funktion *einer Wahrscheinlichkeitsverteilung $(p_k)_{k\in\mathbb{N}_0}$ auf $\mathbb{N}_0$ ist die Funktion φ, die gegeben ist durch*

$$\varphi(s) = \sum_{k\in\mathbb{N}_0} p_k s^k, \qquad |s| < 1.$$

Bemerkung 3.7.2. Da die Reihe der p_k absolut konvergiert, hat die zugehörige erzeugende Funktion φ mindestens den Konvergenzradius Eins, d. h., sie konvergiert mindestens im Innern des (komplexen) Einheitskreises. Insbesondere kann man die Potenzreihe beliebig oft im Intervall $(-1, 1)$ gliedweise differenzieren. Auf Grund des Satzes von Taylor kann man aus der erzeugenden Funktion mit Hilfe der Formel

$$p_k = \frac{\varphi^{(k)}(0)}{k!}, \qquad k \in \mathbb{N}_0,$$

eindeutig die Koeffizienten p_k erhalten, wobei $\varphi^{(k)}$ die k-te Ableitung bedeutet. Es besteht also ein eineindeutiger Zusammenhang zwischen der Verteilung und ihrer erzeugenden Funktion. ◇

Beispiel 3.7.3 (Binomialverteilung)**.** Für $n \in \mathbb{N}$ und $p \in [0, 1]$ ist die erzeugende Funktion der Binomialverteilung zu den Parametern n und p (siehe Beispiel 1.3.4) gegeben durch

$$\varphi(s) = \sum_{k=0}^{n} \mathrm{Bi}_{n,p}(k) s^k = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} s^k = (1 - p + sp)^n,$$

wobei wir den binomischen Lehrsatz benutzten. ◇

Beispiel 3.7.4 (Negative Binomialverteilung)**.** Die erzeugende Funktion einer negativ zu den Parametern $n \in (0, \infty)$ und $p \in [0, 1]$ binomialverteilten Zufallsgröße X (siehe Beispiel 3.6.4) errechnet man als

$$\varphi(s) = p^n \sum_{k\in\mathbb{N}_0} \binom{-n}{k} (p-1)^k s^k = \Big(\frac{p}{1+sp-s}\Big)^n \sum_{k\in\mathbb{N}_0} \mathrm{Neg}_{n,1+sp-s}(k) = \Big(\frac{p}{1+sp-s}\Big)^n,$$

zunächst nur für $s \in \mathbb{R}$ mit $1 + sp - s \in (0, 1]$. Doch da beide Seiten der Gleichung analytisch im Einheitskreis sind, gilt sie auch dort. ◇

Beispiel 3.7.5 (Poisson-Verteilung)**.** Die erzeugende Funktion der Poisson-Verteilung zum Parameter $\alpha \in (0, \infty)$ (siehe Beispiel 1.3.6) ist gegeben durch

$$\varphi(s) = \sum_{k\in\mathbb{N}_0} \mathrm{Po}_\alpha(k) s^k = \sum_{k\in\mathbb{N}_0} e^{-\alpha} \frac{\alpha^k}{k!} s^k = e^{-\alpha(1-s)}.$$

◇

Da die erzeugende Funktion die Verteilung eindeutig fest legt, ist es klar, dass auch der Erwartungswert und die Varianz der Verteilung mit Hilfe der erzeugenden Funktion ausgedrückt werden können. Mit $\mathbb{E}(P) = \sum_{k\in\mathbb{N}_0} kp_k$ und $\mathbb{V}(P) = \sum_{k\in\mathbb{N}_0}(k - \mathbb{E}(P))^2 p_k$ bezeichnen wir den Erwartungswert und die Varianz einer Verteilung $P = (p_k)_{k\in\mathbb{N}_0}$ auf $\mathbb{N}_0$.

Satz 3.7.6. *Es sei $P = (p_k)_{k\in\mathbb{N}_0}$ eine Verteilung auf $\mathbb{N}_0$ mit erzeugender Funktion φ.*

(i) $\mathbb{E}(P)$ existiert genau dann, wenn $\varphi'(1-) = \lim_{s\uparrow 1} \varphi'(s)$ existiert, und dann gilt $\mathbb{E}(P) = \varphi'(1-)$.

(ii) $\mathbb{V}(P)$ existiert genau dann, wenn $\varphi''(1-) = \lim_{s\uparrow 1} \varphi''(s)$ existiert, und dann gilt $\mathbb{V}(P) = \varphi''(1-) - \mathbb{E}(P)^2 + \mathbb{E}(P)$.

Beweis.

(i) Für $|s| < 1$ existiert $\varphi'(s)$, und für $s \uparrow 1$ gilt

$$\varphi'(s) = \sum_{k\in\mathbb{N}_0} p_k k s^{k-1} \uparrow \sum_{k\in\mathbb{N}_0} p_k k = \mathbb{E}(P).$$

(ii) Für $|s| < 1$ existiert $\varphi''(s)$, und für $s \uparrow 1$ gilt

$$\varphi''(s) = \sum_{k\in\mathbb{N}_0} p_k k(k-1) s^{k-2} \uparrow \sum_{k\in\mathbb{N}_0} p_k (k^2 - k).$$

Es ist leicht zu sehen, dass die Reihe $\sum_{k\in\mathbb{N}_0} p_k(k^2-k)$ genau dann konvergiert, wenn $\sum_{k\in\mathbb{N}_0} p_k k^2$ konvergiert, d. h., wenn $\mathbb{V}(P)$ existiert. In diesem Fall errechnet man leicht, dass

$$\varphi''(1-) - \mathbb{E}(P)^2 + \mathbb{E}(P) = \sum_{k\in\mathbb{N}_0} p_k k^2 - \mathbb{E}(P)^2 = \mathbb{V}(P).$$

□

Beispiel 3.7.7. Indem man die in den Beispielen 3.7.3 bzw. 3.7.5 errechneten erzeugenden Funktionen der Binomial- bzw. Poisson-Verteilung bei Eins ein bzw. zwei Mal differenziert, erhält man als eine Übungsaufgabe, dass sie die Erwartungswerte np bzw. α und die Varianzen $np(1-p)$ bzw. α besitzen. (Natürlich stehen diese Ergebnisse in Übereinstimmung mit Beispielen 3.3.4 und 3.5.4 bzw. 3.3.6 und 3.4.5.) ◇

Den Zusammenhang zwischen Verteilungen auf $\mathbb{N}_0$ und den zugehörigen erzeugenden Funktionen kann man auch ausnützen bei der Behandlung von Summen unabhängiger Zufallsgrößen. Wir bezeichnen nun die erzeugende Funktion einer Verteilung $P = (p_k)_{k\in\mathbb{N}_0}$ auf $\mathbb{N}_0$ mit φ_P. Die erzeugende Funktion φ_X einer $\mathbb{N}_0$-wertigen Zufallsvariablen X ist definiert als die erzeugende Funktion der Verteilung von X, also

$$\varphi_X(s) = \varphi_{\mathbb{P}\circ X^{-1}}(s) = \sum_{k\in\mathbb{N}_0} \mathbb{P}(X = k) s^k = \mathbb{E}(s^X), \qquad |s| < 1.$$

Wir erinnern daran (siehe Abschnitt 3.6), dass die Faltung $P_1 \star P_2$ zweier Verteilungen P_1 und P_2 auf $\mathbb{N}_0$ die Verteilung der Summe zweier unabhängiger Zufallsgrößen mit Verteilung P_1 bzw. P_2 ist. Es stellt sich heraus, dass die erzeugende Funktion dieser Summe gleich dem punktweisen Produkt der beiden erzeugenden Funktionen ist:

Satz 3.7.8 (Faltung und Unabhängigkeit). *(i) Für je zwei unabhängige $\mathbb{N}_0$-wertige Zufallsgrößen X_1 und X_2 gilt*

$$\varphi_{X_1+X_2}(s) = \varphi_{X_1}(s)\varphi_{X_2}(s), \qquad |s| < 1.$$

(ii) Für je zwei Verteilungen P_1 und P_2 auf $\mathbb{N}_0$ gilt

$$\varphi_{P_1 \star P_2}(s) = \varphi_{P_1}(s)\varphi_{P_2}(s), \qquad |s| < 1.$$

Beweis. (i) Nach Korollar 3.2.12 sind auch die beiden Zufallsvariablen s^{X_1} und s^{X_2} unabhängig. Nach dem Produktsatz für Erwartungswerte bei Unabhängigkeit (siehe Satz 3.3.2(d)) haben wir

$$\varphi_{X_1+X_2}(s) = \mathbb{E}(s^{X_1+X_2}) = \mathbb{E}(s^{X_1}s^{X_2}) = \mathbb{E}(s^{X_1})\mathbb{E}(s^{X_2}) = \varphi_{X_1}(s)\varphi_{X_2}(s).$$

Die Aussage in (ii) ist der Spezialfall von (i) für identische Zufallsgrößen X_1 und X_2. □

Bemerkung 3.7.9. Den Satz 3.7.8 kann man statt auf probabilistischem Wege auch mit Hilfe aus der Analysis beweisen, wenn man benutzt, dass das punktweise Produkt $\varphi_{X_1}(s)\varphi_{X_2}(s)$ wieder eine Potenzreihe ist, deren Koeffizientenfolge die Faltung der Koeffizientenfolgen von $\varphi_{X_1}(s)$ und $\varphi_{X_2}(s)$ sind, also

$$\begin{aligned}\varphi_{X_1}(s)\varphi_{X_2}(s) &= \Big(\sum_{k=0}^\infty p_{X_1}(k)s^k\Big)\Big(\sum_{k=0}^\infty p_{X_2}(k)s^k\Big) = \sum_{k=0}^\infty (p_{X_1} \star p_{X_2})(k)s^k \\ &= \sum_{k=0}^\infty p_{X_1+X_2}(k)s^k = \varphi_{X_1+X_2}(s),\end{aligned}$$

wobei im vorletzten Schritt der Faltungssatz 3.6.2 benutzt wurde. ◇

Die bemerkenswerte Aussage von Satz 3.7.8 gibt uns ein weiteres Mittel in die Hand, Verteilungen von Summen unabhängiger Zufallsgrößen zu identifizieren:

Beispiel 3.7.10 (Binomialverteilung). Die erzeugende Funktion der Binomialverteilung mit Parametern n und p (siehe Beispiel 3.7.3) ist offensichtlich das n-fache Produkt der erzeugenden Funktion einer Binomialverteilung mit Parametern 1 und p. Dies reflektiert die bekannte Tatsache (siehe Beispiel 3.6.3) dass eine binomialverteilte Zufallsgröße eine Summe unabhängiger Bernoulli-Zufallsgrößen ist, oder auch die Summe unabhängiger binomialverteilter Zufallsgrößen mit gewissen Parametern. ◇

Beispiel 3.7.11 (Negative Binomialverteilung). Wie man im Beispiel 3.7.4 gesehen hat, ist auch die erzeugende Funktion der Negativen Binomialverteilung mit Parametern $n \in (0,\infty)$ und $p \in [0,1]$ die n-te Potenz derselben Verteilung mit Parametern 1 und p. Daher ist die Summe unabhängiger negativ zu den Parametern $n_1 \in (0,\infty)$ und p bzw. $n_2 \in (0,\infty)$ und p binomialverteilter Zufallsgrößen negativ binomialverteilt zu den Parametern n_1+n_2 und p. Also gilt insbesondere die Faltungsformel

$$\mathrm{Neg}_{n_1,p} \star \mathrm{Neg}_{n_2,p} = \mathrm{Neg}_{n_1+n_2,p},$$

die wir in Beispiel 3.6.4 nur für natürliche Zahlen n_1 und n_2 bewiesen. ◇

Beispiel 3.7.12 (Poisson-Verteilung)**.** Ein sehr eleganter und kurzer Beweis für die Tatsache (siehe Beispiel 3.6.5), dass die Summe zweier unabhängiger Poisson-verteilter Zufallsgrößen X und Y mit Parametern α und β Poisson-verteilt ist mit Parameter $\alpha + \beta$, ist nun mit Satz 3.7.8 und Beispiel 3.7.5 möglich: Die erzeugende Funktion der Summe ist gegeben durch

$$\varphi_{X+Y}(s) = \varphi_X(s)\varphi_Y(s) = e^{\alpha(s-1)}e^{\beta(s-1)} = e^{(\alpha+\beta)(s-1)}.$$

Also wird die erzeugende Funktion von $X + Y$ identifiziert mit der einer Poisson-Verteilung mit Parameter $\alpha + \beta$. Wegen der Eindeutigkeit der erzeugenden Funktion ist $X + Y$ also Poisson-verteilt mit diesem Parameter. $\diamond$

Kapitel 4

Wahrscheinlichkeit mit Dichten

In diesem Kapitel erweitern bzw. übertragen wir die bisher behandelte Theorie auf Wahrscheinlichkeitsmaße bzw. Zufallsgrößen, die mit Hilfe von Integralen über Dichten beschrieben werden. Alle in diesem Kapitel auftretenden Integrale werden als Riemann-Integrale aufgefasst, sodass Vorkenntnisse in Analysis I ausreichen und eine Kenntnis des Lebesgue-Integrals nicht benötigt wird. Wir werden auch darauf verzichten, Wahrscheinlichkeitsräume anzugeben, denn dies wird nicht nötig sein für eine Behandlung der Theorie auf dem Niveau, auf dem wir sie betreiben. Eine mathematische Fundierung wird später nachgeholt werden.

Man beachte zum Beispiel die Unmöglichkeit, die diskrete Wahrscheinlichkeitstheorie und die Theorie der Wahrscheinlichkeit mit Dichten unter das Dach eines übergeordneten Konzepts zu stellen; sie werden scheinbar verbindungslos neben einander stehen bleiben, mit allerdings nicht zu übersehenden Verwandtschaften. Tatsächlich werden wir nicht einmal im Stande sein, eine diskrete Zufallsgröße und eine über eine Dichte definierte Zufallsgröße miteinander zu addieren. Dieser beklagenswerte Zustand wird allerdings später aufgehoben werden, wenn wir das Riemann-Integral durch das Lebesgue-Integral ersetzen werden (das ja kompatibel ist mit abzählbaren Operationen, im Gegensatz zum Riemann-Integral) und Maßtheorie benutzen werden.

4.1 Grundbegriffe

Definition 4.1.1 (Dichte, Verteilungsfunktion). *(a) Eine Abbildung $f\colon \mathbb{R} \to [0,\infty)$, so dass $\int_{\mathbb{R}} f(x)\,\mathrm{d}x$ existiert und den Wert Eins hat, heißt eine* Wahrscheinlichkeitsdichte *oder kurz eine* Dichte.

(b) Eine Abbildung $F\colon \mathbb{R} \to [0,1]$ heißt eine Verteilungsfunktion*, falls gelten:*

(i) F ist monoton steigend,

(ii) $\lim_{t\to\infty} F(t) = 1$ und $\lim_{t\to-\infty} F(t) = 0$,

(iii) F ist rechtsseitig stetig (d. h. $\lim_{s\downarrow t} F(s) = F(t)$ für jedes $t \in \mathbb{R}$).

Bemerkung 4.1.2. (a) Falls f eine Dichte ist, so definiert $F(t) = \int_{-\infty}^{t} f(x)\,\mathrm{d}x$ eine Verteilungsfunktion, und zwar sogar eine stetige. Man nennt dann f eine Dichte von F. Nicht jede stetige Verteilungsfunktion besitzt eine Dichte.

(b) Falls eine Dichte f in endlich vielen Punkten abgeändert wird, erhält man eine neue Dichte $\tilde{f}$. Für jedes Intervall I gilt dann $\int_I f(x)\,\mathrm{d}x = \int_I \tilde{f}(x)\,\mathrm{d}x$.

(c) Falls eine Dichte f einer Verteilungsfunktion F stetig in einem Punkte a ist, so gilt $F'(a) = f(a)$ nach dem Hauptsatz der Differential- und Integralrechnung.

$\diamond$

Definition 4.1.3 (Verteilungsfunktion und Dichte einer Zufallsgröße). *Für eine reellwertige Zufallsgröße X heißt die Abbildung $F_X\colon \mathbb{R} \to [0,1]$, definiert durch $F_X(t) = \mathbb{P}(X \leq t)$, die* Verteilungsfunktion von X. *Wir sagen, X* hat eine Dichte, *wenn ihre Verteilungsfunktion F_X eine hat.*

Bemerkung 4.1.4. (a) Falls X eine diskrete Zufallsgröße ist, so ist F_X die rechtsstetige Treppenfunktion, die in den Punkten x mit $\mathbb{P}(X = x) > 0$ einen Sprung der Größe $\mathbb{P}(X = x)$ macht. Insbesondere hat X keine Dichte.

(b) Wenn eine Zufallsgröße X eine Dichte f hat, dann gilt

$$\mathbb{P}\big(X \in A\big) = \mathbb{P} \circ X^{-1}(A) = \int_A f(x)\,\mathrm{d}x, \tag{4.1.1}$$

für alle Mengen $A \subset \mathbb{R}$, für die die Abbildung $f\mathbb{1}_A$ Riemann-integrierbar ist, also mindestens für alle endlichen Vereinigungen A von Intervallen. Insbesondere ist $\mathbb{P}(X = x) = 0$ für alle $x \in \mathbb{R}$, denn

$$0 \leq \mathbb{P}(X = x) \leq \mathbb{P}\Big(x \leq X \leq x + \frac{1}{n}\Big) = \int_x^{x+\frac{1}{n}} f(y)\,\mathrm{d}y \longrightarrow 0 \qquad \text{für } n \to \infty.$$

Da X einzelne Werte mit Wahrscheinlichkeit Null annimmt, gilt insbesondere $\mathbb{P}(X \in [a,b]) = \mathbb{P}(X \in (a,b]) = \mathbb{P}(X \in (a,b))$ etc. Wir werden allerdings keinen Wahrscheinlichkeitsraum angeben, auf dem X definiert wäre.

(c) Wir sagen, ein Wahrscheinlichkeitsmaß $\mathbb{P}$ auf $\Omega = \mathbb{R}$ besitzt eine Dichte f, wenn die identische Zufallsgröße $X(\omega) = \omega$ eine hat. In diesem Fall gilt also $\mathbb{P}(A) = \int_A f(x)\,\mathrm{d}x$ für alle Mengen $A \subset \mathbb{R}$, für die $f\mathbb{1}_A$ Riemann-integrierbar ist, mindestens aber für alle endlichen Vereinigungen von Intervallen.

$\diamond$

4.2 Übertragung der bisherigen Ergebnisse

Die meisten allgemeinen Aussagen der voran gegangenen Kapitel über Wahrscheinlichkeiten, Erwartungswerte, Varianzen, Kovarianzen und Unabhängigkeit gelten analog auch für Wahrscheinlichkeitsmaße bzw. Zufallsgrößen mit Riemann-integrierbaren Dichten, und die Beweise laufen analog. In den meisten Fällen genügt es, die auftretenden Summen über $x \in X(\Omega)$ durch die entsprechenden Integrale $\int_{\mathbb{R}} \dots \,\mathrm{d}x$ und den Term $\mathbb{P}(X = x)$ durch die Dichte $f(x)$ zu ersetzen und die Sprechweise anzupassen. Ausdrücke, die explizit den Wahrscheinlichkeitsraum Ω involvieren, müssen wir außen vor lassen.

Da das Riemann-Integral allerdings (anders als das Lebesgue-Integral) keine abzählbaren Additivitätseigenschaften aufweist[1], können also diejenigen Eigenschaften, die auf der zweiten Aussage in den Kolmogorovschen Axiomen in Bemerkung 1.1.3 beruhen, nicht gefolgert werden und müssen durch die entsprechende *endliche* Additivität ersetzt werden.

Im Einzelnen gilt Folgendes.

Grundbegriffe

Die Rechenregeln für Wahrscheinlichkeiten in Lemma 1.1.4(a)-(c) gelten wörtlich auch für Wahrscheinlichkeitsmaße $\mathbb{P}$ mit Dichten, aber die Aussagen in (d) und (e) müssen auf endliche Familien von Ereignissen eingeschränkt werden. Die bedingte Wahrscheinlichkeit wird wie in Definition 2.1.2 definiert, und ihre Eigenschaften in Lemma 2.1.4 (allerdings nur für *endliche* Familien von Ereignissen) und die Multiplikationsformel in Lemma 2.1.6 gelten ebenso für Wahrscheinlichkeiten mit Dichten.

Gemeinsame Verteilungen und Randdichten

Definition 4.2.1 (gemeinsame Verteilungsfunktion und Dichte)**.** *Die* gemeinsame Verteilungsfunktion *von Zufallsgrößen $X_1,\dots,X_n$, also die Verteilungsfunktion des Zufallsvektors $X = (X_1,\dots,X_n)$, ist definiert durch*

$$\begin{aligned} F_X(t_1,\dots,t_n) &= \mathbb{P}\big(X_1 \le t_1,\dots,X_n \le t_n\big) \\ &= \mathbb{P}\circ X^{-1}\big((-\infty,t_1]\times\cdots\times(-\infty,t_n]\big), \qquad t_1,\dots,t_n \in \mathbb{R}. \end{aligned}$$

Wir sagen, $X_1,\dots,X_n$ haben eine gemeinsame Dichte $f\colon \mathbb{R}^n \to [0,\infty)$ *(oder der Zufallsvektor $X = (X_1,\dots,X_n)$ habe eine Dichte f), falls gilt*

$$\mathbb{P}(X\in A) = \mathbb{P}\circ X^{-1}(A) = \int_A f(x_1,\dots,x_n)\,\mathrm{d}x_1\dots\mathrm{d}x_n$$

für alle $A \subset \mathbb{R}^n$, für die die Abbildung $f1\!\!1_A$ Riemann-integrierbar ist.

Bemerkung 4.2.2. (a) Wenn $X_1,\dots,X_n$ die Verteilungsfunktion F_X hat, so hat jedes X_i die Verteilungsfunktion

$$\begin{aligned} F_{X_i}(t_i) &= \mathbb{P}(X_1<\infty,\dots,X_{i-1}<\infty, X_i \le t_i, X_{i+1}<\infty,\dots,X_n<\infty) \\ &= \lim_{t_1,\dots,t_{i-1},t_{i+1},\dots,t_n\to\infty} F_X(t_1,\dots,t_n). \end{aligned}$$

(b) Wenn $X_1,\dots,X_n$ eine gemeinsame Dichte f haben, dann gilt insbesondere

$$\begin{aligned} F_X(t_1,\dots,t_n) &= \mathbb{P}\circ X^{-1}\Big(\prod_{i=1}^n(-\infty,t_i]\Big) = \int_{\prod_{i=1}^n(-\infty,t_i]} f(x_1,\dots,x_n)\,\mathrm{d}x_1\dots\mathrm{d}x_n \\ &= \int_{-\infty}^{t_n}\cdots\int_{-\infty}^{t_1} f(x_1,\dots,x_n)\,\mathrm{d}x_1\dots\mathrm{d}x_n, \qquad t_1,\dots,t_n\in\mathbb{R}. \end{aligned}$$

[1]Zum Beispiel ist für jedes $x \in \mathbb{Q}\cap[0,1]$ die Abbildung $1\!\!1_{\{x\}}$ Riemann-integrierbar, nicht aber ihre abzählbare Summe $1\!\!1_{\mathbb{Q}\cap[0,1]}$.)

(Der *Satz von Fubini* garantiert, dass der Wert dieses n-fachen Integrals nicht von der Reihenfolge der Integration abhängt.) Insbesondere besitzen auch die einzelnen Zufallsgrößen $X_1, \dots, X_n$ jeweils eine Dichte, und zwar erhält man eine Dichte von X_i, indem man f über alle Werte, die die anderen Zufallsgrößen annehmen können, integriert:

$$\begin{aligned}\mathbb{P}(X_i \le t_i) &= \int_{\mathbb{R}^{i-1}\times(-\infty,t_i]\times\mathbb{R}^{n-i}} f(x_1,\dots,x_n)\,\mathrm{d}x_1\dots\mathrm{d}x_n \\ &= \int_{(-\infty,t_i]} \Big(\int_{\mathbb{R}^{n-1}} f(x_1,\dots,x_n)\,\mathrm{d}x_1\dots\mathrm{d}x_{i-1}\mathrm{d}x_{i+1}\dots\mathrm{d}x_n\Big)\mathrm{d}x_i.\end{aligned}$$

Der Ausdruck in den Klammern als Funktion von x_i nennt man die i-te *Randdichte* von f; diese Funktion ist eine (eindimensionale) Dichte von X_i. In analoger Weise kann man für jeden Teilvektor von $X = (X_1, \dots, X_n)$ eine Dichte erhalten, indem man über alle Indices, die nicht in dem Vektor verwendet werden, ausintegriert.

◇

Unabhängigkeit

Für die Definition der Unabhängigkeit von Zufallsgrößen adaptieren wir die Aussage von Lemma 3.2.2:

Definition 4.2.3 (Unabhängigkeit von Zufallsgrößen)**.** *Es seien $X_1, \dots, X_n$ beliebige Zufallsgrößen. Wir nennen $X_1, \dots, X_n$* unabhängig*, falls für alle $t_1, \dots, t_n \in \mathbb{R}$ gilt:*

$$\mathbb{P}(X_1 \le t_1, \dots, X_n \le t_n) = \prod_{i=1}^n \mathbb{P}(X_i \le t_i).$$

Wie in Lemma 3.2.9 sind die Zufallsgrößen $X_1, \dots, X_n$ mit Dichten genau dann unabhängig, wenn die Verteilung des Vektors $X = (X_1, \dots, X_n)$ gleich dem Produkt der Verteilungen der $X_1, \dots, X_n$ ist. Unabhängigkeit schlägt sich auch in der Produktstruktur der Dichten nieder:

Lemma 4.2.4. *Es seien Zufallsgrößen $X_1, \dots, X_n$ mit Dichten $f_1, \dots, f_n \colon \mathbb{R} \to [0, \infty)$ gegeben (wir setzen nicht voraus, dass eine gemeinsame Dichte existiert). Dann sind $X_1, \dots, X_n$ genau dann unabhängig, wenn eine gemeinsame Dichte gegeben ist durch die Abbildung*

$$(x_1, \dots, x_n) \mapsto \prod_{i=1}^n f_i(x_i), \qquad x_1, \dots, x_n \in \mathbb{R}.$$

Beweis. Übungsaufgabe. □

Erwartungswerte, Varianzen und Kovarianzen

Erwartungswerte, Varianzen und Kovarianzen von Zufallsgrößen mit Dichten werden analog zu den jeweils entsprechenden Begriffen für diskrete Zufallsgrößen definiert:

Definition 4.2.5 (Erwartungswert,Varianz). *Es sei X eine Zufallsgröße mit Dichte f.*

(a) Der Erwartungswert *von X existiert genau dann (und wir schreiben dann $X \in \mathcal{L}^1$), wenn das Integral $\int_{\mathbb{R}} |x| f(x)\,\mathrm{d}x$ konvergiert, und der Erwartungswert ist dann gegeben als*

$$\mathbb{E}(X) = \int_{\mathbb{R}} x f(x)\,\mathrm{d}x.$$

(b) Wenn $X \in \mathcal{L}^1$, so heißt

$$\mathbb{V}(X) = \int_{\mathbb{R}} (x - \mathbb{E}(X))^2 f(x)\,\mathrm{d}x = \mathbb{E}\big((X - \mathbb{E}(X))^2\big)$$

die Varianz *von X und $S(X) = \sqrt{\mathbb{V}(X)}$ die* Standardabweichung *von X.*

Die Eigenschaften des Erwartungswertes in Lemma 3.3.2 und die Formel für Erwartungswerte von zusammengesetzen Zufallsgrößen in Lemma 3.3.3 gelten wörtlich bzw. analog; in Lemma 3.3.2(a) und Lemma 3.3.3 müssen die Summen durch Integrale ersetzt werden. Die Eigenschaften von Varianzen und Kovarianzen in Lemmas 3.4.4 und 3.5.2 sowie die Cauchy-Schwarz-Ungleichung in Satz 3.5.6 gelten wörtlich auch für Zufallsvariable mit Dichten.

Faltung

Den Begriff der *Faltung* in Definition 3.6.1 überträgt man wie folgt auf Integrale: Die Faltung zweier integrierbarer Funktionen $f, g\colon \mathbb{R} \to \mathbb{R}$ ist definiert als die Funktion $f \star g\colon \mathbb{R} \to \mathbb{R}$, gegeben durch

$$f \star g(y) = \int_{\mathbb{R}} f(x) g(y - x)\,\mathrm{d}x, \qquad y \in \mathbb{R}.$$

Man kann leicht zeigen, dass $f \star g = g \star f$, und dass $f \star g$ absolut integrierbar ist, falls f und g es sind. Es gilt das Analogon zum Faltungssatz 3.6.2:

Satz 4.2.6 (Faltungssatz). *Für je zwei unabhängige Zufallsgrößen X und Y mit Dichten f bzw. g hat die Zufallsgröße $X + Y$ die Dichte $f \star g$.*

Beweis. Nach Lemma 4.2.4 hat (X, Y) die Dichte $(x, y) \mapsto f(x)g(y)$. Es sei $z \in \mathbb{R}$ und $A_z = \{(x, y) \in \mathbb{R}^2 \colon x + y \leq z\}$. Dann ist

$$\begin{aligned}
\mathbb{P}(X + Y \leq z) = \mathbb{P}\big((X, Y) \in A_z\big) &= \int_{A_z} f(x)g(y)\,\mathrm{d}x\mathrm{d}y = \int_{-\infty}^{\infty} \mathrm{d}x\, f(x) \int_{-\infty}^{z-x} \mathrm{d}y\, g(y) \\
&= \int_{-\infty}^{\infty} \mathrm{d}x\, f(x) \int_{-\infty}^{z} \mathrm{d}y\, g(y - x) = \int_{-\infty}^{z} \mathrm{d}y \Big(\int_{\mathbb{R}} \mathrm{d}x\, g(y - x) f(x)\Big) \\
&= \int_{-\infty}^{z} \mathrm{d}y\, f \star g(y).
\end{aligned}$$

Also ist $f \star g$ eine Dichte von $X + Y$. $\square$

4.3 Beispiele

Beispiel 4.3.1 (Gleichförmige Verteilung). Seien $a, b \in \mathbb{R}$ mit $a < b$, dann ist durch

$$f(t) = \frac{1}{b-a} \mathbb{1}_{[a,b]}(t), \qquad \text{für } t \in \mathbb{R},$$

eine Dichte auf $\mathbb{R}$ gegeben, die Dichte der *gleichförmigen Verteilung auf* $[a, b]$. Die zugehörige Verteilungsfunktion F hat ein lineares Stück vom Wert Null bei a zum Wert Eins bei b. Eine Zufallsgröße X mit Dichte f besitzt den Erwartungswert $\mathbb{E}(X) = \frac{a+b}{2}$ und die Varianz $\mathbb{V}(X) = \frac{1}{12}(b-a)^2$, wie man leicht ausrechnet. Analog werden gleichförmige Verteilungen auf beliebigen Teilmengen des $\mathbb{R}^d$ definiert, deren Indikatorfunktion Riemann-integrierbar ist. $\diamond$

Beispiel 4.3.2 (Exponentialverteilung). Wir definieren mit einem Parameter $\alpha \in (0, \infty)$ die Dichte

$$f(t) = \alpha e^{-\alpha t} \mathbb{1}_{[0,\infty)}(t), \qquad \text{für } t \in \mathbb{R},$$

die die Dichte der *Exponentialverteilung* zum Parameter α genannt wird. Die zugehörige Verteilungsfunktion ist gegeben durch $F(t) = (1 - e^{-\alpha t})\mathbb{1}_{[0,\infty)}(t)$, und den Erwartungswert errechnet man mit Hilfe einer partiellen Integration als

$$\mathbb{E}(X) = \int_{\mathbb{R}} t f(t)\, \mathrm{d}t = \int_0^\infty t\alpha e^{-\alpha t}\, \mathrm{d}t = -te^{-\alpha t}\Big|_0^\infty + \int_0^\infty e^{-\alpha t}\, \mathrm{d}t = \frac{1}{\alpha}.$$

Mit einer weiteren partiellen Integration errechnet man leicht, dass die Varianz gegeben ist als $\mathbb{V}(X) = 1/\alpha^2$ (Übungsaufgabe).

Die Exponentialverteilung ist das kontinuierliche Gegenstück zur geometrischen Verteilung. Insbesondere besitzt sie ebenfalls die Eigenschaft der Gedächtnislosigkeit (siehe Lemma 3.1.6):

Lemma 4.3.3 (Gedächtnislosigkeit der Exponentialverteilung). *Sei X eine exponentiell verteilte Zufallsvariable. Dann gilt für jede $s, t > 0$*

$$\mathbb{P}(X > s + t \mid X > s) = \mathbb{P}(X > t).$$

Beweis. Übungsaufgabe. $\square$

Die Interpretation ist analog zu der von Lemma 3.1.6: Wenn man auf das Eintreten einer exponentiell verteilten Zufallszeit wartet und sie bis zum Zeitpunkt s noch nicht eingetreten ist, so ist die Wahrscheinlichkeit, dass sie nach weiteren t Zeiteinheiten eintritt, die gleiche, als wenn man das Nichteintreten in den letzten s Zeiteinheiten nicht kennen würde. Weitere interessante Eigenschaften dieser Wartezeitverteilung treten bei der Behandlung des Poisson-Prozesses in Abschnitt 4.4 auf.

Das Maximum unabhängiger gleichförmig verteiler Zufallsgrößen steht in einer interessanten Beziehung zur Exponentialverteilung (Beweis als Übungsaufgabe):

Lemma 4.3.4. *Es seien $X_1, \dots, X_n$ unabhängige, auf dem Intervall $[0, \alpha]$ gleichförmig verteilte Zufallsgrößen. Dann ist eine Dichte der Zufallsgröße $M_n = \max\{X_1, \dots, X_n\}$ gegeben durch $x \mapsto \mathbb{1}_{[0,\alpha]}(x) n\alpha^{-n} x^{n-1}$, und ihr Erwartungswert ist $\mathbb{E}(M_n) = \frac{n}{n+1}\alpha$. Für $n \to \infty$ konvergiert die Verteilungsfunktion der Zufallsgröße $Y_n = n(\alpha - M_n)$ gegen die Verteilungsfunktion der Exponentialverteilung mit Parameter $\frac{1}{\alpha}$.*

$\diamond$

Beispiel 4.3.5 (Gamma-Verteilung). Mit zwei Parametern $\alpha > 0$ und $r > 0$ definieren wir die Dichte $\gamma_{\alpha,r} \colon (0,\infty) \to [0,\infty)$ durch

$$\gamma_{\alpha,r}(t) = \frac{\alpha^r}{\Gamma(r)} t^{r-1} e^{-\alpha t} \mathbb{1}_{[0,\infty)}(t),$$

wobei $\Gamma \colon (0,\infty) \to (0,\infty)$ die bekannte Gamma-Funktion[2] ist:

$$\Gamma(r) = \int_0^\infty y^{r-1} e^{-y} \, \mathrm{d}y, \qquad r > 0.$$

Mit Hilfe der Substitution $\alpha t = y$ sieht man leicht, dass $\gamma_{\alpha,r}$ eine Dichte ist, die Dichte der *Gamma-Verteilung* mit Parametern α und r. Der Spezialfall $r = 1$ ist die Dichte der Exponentialverteilung, siehe Beispiel 4.3.2. Die Gamma-Verteilung ist das kontinuierliche Analogon zur negativen Binomial-Verteilung (siehe Beispiel 3.6.4). Sie besitzt ein paar bemerkenswerte Beziehungen zur Exponential- und zur Poisson-Verteilung. Insbesondere stellt sich heraus, dass die Faltungsgleichung $\gamma_{\alpha,r_1} \star \gamma_{\alpha,r_2} = \gamma_{\alpha,r_1+r_2}$ gilt, d. h., dass die Familie der Gamma-Verteilungen eine Faltungshalbgruppe bildet (siehe auch Abschnitt 3.6):

Lemma 4.3.6 (Gamma-, Exponential- und Poisson-Verteilung). *Es sei $\alpha > 0$.*

(i) Die Summe zweier unabhängiger Gamma-verteilter Zufallsgrößen mit Parametern α und r_1 bzw. α und r_2 ist Gamma-verteilt mit Parametern α und $r_1 + r_2$. Insbesondere ist für $k \in \mathbb{N}$ die Gamma-Verteilung mit Parameter α und k identisch mit der Verteilung der Summe von k unabhängigen, zum Parameter α exponentiell verteilten Zufallsgrößen.

(ii) Für ein $t > 0$ sei $N_{\alpha t}$ eine zum Parameter αt Poisson-verteilte Zufallsgröße und $X_{\alpha,k}$ eine zum Parameter α und $k \in \mathbb{N}$ Gamma-verteilte Zufallsgröße. Dann gilt $\mathbb{P}(N_{\alpha t} \geq k) = \mathbb{P}(X_{\alpha,k} \leq t)$.

Beweis. (i) Es genügt, die Gleichung $\gamma_{\alpha,r_1} \star \gamma_{\alpha,r_2} = \gamma_{\alpha,r_1+r_2}$ zu beweisen: Für $s > 0$ gilt

$$\begin{aligned}
\gamma_{\alpha,r_1} \star \gamma_{\alpha,r_2}(s) &= \int_0^s \gamma_{\alpha,r_1}(t) \gamma_{\alpha,r_2}(s-t) \, \mathrm{d}t \\
&= \frac{\alpha^{r_1}}{\Gamma(r_1)} \frac{\alpha^{r_2}}{\Gamma(r_2)} e^{-\alpha s} \int_0^s t^{r_1-1}(s-t)^{r_2-1} \, \mathrm{d}t \\
&= \gamma_{\alpha,r_1+r_2}(s) \frac{\Gamma(r_1+r_2)}{\Gamma(r_1)\Gamma(r_2)} \int_0^1 u^{r_1-1}(1-u)^{r_2-1} \, \mathrm{d}u,
\end{aligned}$$

wobei wir die Substitution $t/s = u$ benutzten. Das Integral auf der rechten Seite ist bekannt als das Eulersche Beta-Integral mit Parametern r_1 und r_2, und es ist bekannt, dass sein Wert gleich dem Kehrwert des Bruches davor ist. Also ist die gesamte rechte Seite gleich $\gamma_{\alpha,r_1+r_2}(s)$, und die Aussage ist bewiesen.

[2]Die Gamma-Funktion ist die einzige logarithmisch konvexe Funktion $f \colon (0,\infty) \to (0,\infty)$ mit $f(1) = 1$, die die Funktionalgleichung $\Gamma(r+1) = r\Gamma(r)$ für jedes $r > 0$ erfüllt. Sie interpoliert also auf $\mathbb{N}$ die Fakultät, d. h. $f(k) = (k-1)!$ für $k \in \mathbb{N}$.

(ii)

$$\begin{aligned}
\mathbb{P}(N_{\alpha t} \geq k) &= 1 - \mathbb{P}(N_{\alpha t} \leq k-1) = 1 - e^{-\alpha t} \sum_{n=0}^{k-1} \frac{(\alpha t)^n}{n!} \\
&= \int_0^t \frac{\alpha^k}{(k-1)!} x^{k-1} e^{-\alpha x} \, \mathrm{d}x = \int_0^t \gamma_{\alpha,k}(x) \, \mathrm{d}x = \mathbb{P}(X_{\alpha,k} \leq t),
\end{aligned}$$

wobei der dritte Schritt mit Hilfe einer Differenziation nach t eingesehen wird. $\square$ $\diamond$

Beispiel 4.3.7 (Normal- oder Gaußverteilung)**.** Mit zwei Parametern $\mu \in \mathbb{R}$ und $\sigma \in (0, \infty)$ definieren wir $\varphi_{\mu,\sigma} \colon \mathbb{R} \to (0, \infty)$ durch

$$\varphi_{\mu,\sigma}(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(t-\mu)^2}{2\sigma^2}\Big), \qquad \text{für } t \in \mathbb{R}.$$

Wir benutzen nun einen kleinen Trick, um zu zeigen, dass $\varphi_{\mu,\sigma}$ tatsächlich eine Wahrscheinlichkeitsdichte ist, d. h. dass $\int_{\mathbb{R}} \varphi_{\mu,\sigma}(t) \, \mathrm{d}t = 1$. Zunächst bemerken wir, dass es genügt, dies nur für $\mu = 0$ und $\sigma = 1$ zu tun, denn das Integral erstreckt sich über die gesamte reelle Achse und kann um μ verschoben werden, und eine Substitution $t = \tilde{t}\sigma$ führt die Frage auf den Fall $\sigma = 1$ zurück. Wir werden zeigen, dass

$$\Big(\int_{\mathbb{R}} e^{-t^2/2} \, \mathrm{d}t\Big)^2 = 2\pi,$$

woraus die Behauptung folgt. Wir schreiben das Quadrat der Integrale als ein zweidimensionales Integral über $x = (x_1, x_2) \in \mathbb{R}^2$ (man beachte, dass $x_1^2 + x_2^2 = \|x\|_2^2$) und gehen über zu Polarkoordinaten:

$$\Big(\int_{\mathbb{R}} e^{-t^2/2} \, \mathrm{d}t\Big)^2 = \int_{\mathbb{R}} e^{-x_1^2/2} \, \mathrm{d}x_1 \int_{\mathbb{R}} e^{-x_2^2/2} \, \mathrm{d}x_2 = \int_{\mathbb{R}^2} e^{-\|x\|_2^2/2} \, \mathrm{d}x = \int_0^{2\pi} \int_0^\infty r e^{-r^2/2} \, \mathrm{d}r \mathrm{d}t.$$

Der Lohn dieses Ansatzes ist, dass wir eine explizite Stammfunktion für den Integranden $r \mapsto re^{-r^2/2}$ haben, und zwar $r \mapsto -e^{-r^2/2}$. Daher ist das innere Integral offensichtlich gleich Eins, und der gesamte Ausdruck gleich 2π, wie behauptet. Die Funktion $\varphi_{\mu,\sigma}$ ist also tatsächlich eine Wahrscheinlichkeitsdichte, und zwar die Dichte der *Normal-* oder *Gaußverteilung*. Auf den 10-DM-Scheinen, die bis Ende 2001 im Umlauf waren, war der Graf von $\varphi_{\mu,\sigma}$ abgebildet, man nennt ihn die *Gaußsche Glockenkurve*.

Für die zugehörige Verteilungsfunktion

$$\Phi_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^x \exp\Big(-\frac{(t-\mu)^2}{2\sigma^2}\Big) \, \mathrm{d}t$$

gibt es keinen geschlossenen Ausdruck, aber Tabellen für viele ihrer Werte. Die Rolle der Parameter μ und σ wird klar, wenn man den Erwartungswert und die Varianz ausrechnet. Es sei X eine Zufallsgröße mit Dichte $\varphi_{\mu,\sigma}$, dann gilt:

$$\begin{aligned}
\mathbb{E}(X) &= \int_{\mathbb{R}} t\varphi_{\mu,\sigma}(t) \, \mathrm{d}t = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} (t+\mu) \exp\Big(-\frac{t^2}{2\sigma^2}\Big) \, \mathrm{d}t \\
&= \mu + \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} t \exp\Big(-\frac{t^2}{2\sigma^2}\Big) \, \mathrm{d}t \\
&= \mu,
\end{aligned}$$

denn die Funktion $t \mapsto te^{-t^2/(2\sigma^2)}$ ist antisymmetrisch auf $\mathbb{R}$, und da das Integral existiert (wie man leicht etwa mit Vergleichskriterien sieht), ist sein Wert gleich Null.

Außerdem errechnet man die Varianz mit Hilfe einer Substitution und einer partiellen Integration zu

$$\begin{aligned}\mathbb{V}(X) &= \int_{\mathbb{R}} (t - \mathbb{E}(X))^2 \varphi_{\mu,\sigma}(t)\,\mathrm{d}t = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} (t-\mu)^2 \exp\Big(-\frac{(t-\mu)^2}{2\sigma^2}\Big)\,\mathrm{d}t \\ &= \frac{\sigma^2}{\sqrt{2\pi}} \int_{\mathbb{R}} s^2 e^{-s^2/2}\,\mathrm{d}s = \frac{\sigma^2}{\sqrt{2\pi}} \Big[-se^{-s^2/2}\Big|_{-\infty}^{\infty} + \int_{\mathbb{R}} e^{-s^2/2}\,\mathrm{d}s\Big] \\ &= \sigma^2.\end{aligned}$$

Also ist μ der Erwartungswert und σ^2 die Varianz der Normalverteilung mit Parametern μ und σ. Man bezeichnet diese Verteilung auch oft mit $\mathcal{N}(\mu, \sigma^2)$. Im Fall $\mu = 0$ und $\sigma^2 = 1$ sprechen wir von der *Standardnormalverteilung* $\mathcal{N}(0, 1)$.

Die Normalverteilung besitzt mehrere spezielle Eigenschaften und tritt in universeller Weise auf als Grenzverteilung im sehr wichtigen Zentralen Grenzwertsatz, siehe Satz 5.2.2. Wir wollen ihre Faltungseigenschaft beweisen: Die Summe zweier unabhängiger normalverteilter Zufallsgrößen ist wiederum normalverteilt, und ihre ersten Parameter addieren sich und die Quadrate der zweiten ebenfalls:

Lemma 4.3.8 (Faltungseigenschaft der Normalverteilung). *Für alle $\mu_1, \mu_2 \in \mathbb{R}$ und alle $\sigma_1, \sigma_2 \in (0, \infty)$ gilt*

$$\varphi_{\mu_1,\sigma_1} \star \varphi_{\mu_2,\sigma_2} = \varphi_{\mu_1+\mu_2,\sigma}, \qquad \textit{wobei } \sigma^2 = \sigma_1^2 + \sigma_2^2.$$

Beweis. Wir dürfen $\mu_1 = \mu_2 = 0$ annehmen. Sei $t \in \mathbb{R}$. Offensichtlich gilt

$$\frac{\varphi_{0,\sigma_1} \star \varphi_{0,\sigma_2}(t)}{\varphi_{0,\sigma}(t)} = \frac{\sigma}{\sigma_1\sigma_2} \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \exp\Big\{\frac{t^2}{2\sigma^2} - \frac{s^2}{2\sigma_1^2} - \frac{(t-s)^2}{2\sigma_2^2}\Big\}\,\mathrm{d}s. \tag{4.3.1}$$

Eine langweilige, aber unkomplizierte Rechnung identifiziert den Term im Exponenten:

$$\frac{t^2}{2\sigma^2} - \frac{s^2}{2\sigma_1^2} - \frac{(t-s)^2}{2\sigma_2^2} = -\frac{1}{2}\Big(s\frac{\sigma}{\sigma_1\sigma_2} - \frac{t}{\sigma}\frac{\sigma_1}{\sigma_2}\Big)^2.$$

Nun benutzt man dies im Integral in (4.3.1) und substituiert den Term zwischen den Klammern im Integral. Also sieht man, dass die rechte Seite gleich Eins ist. □ ◇

Beispiel 4.3.9 (Mehrdimensionale Normalverteilung). Es seien $X_1, \dots, X_n$ unabhängige, standardnormalverteilte Zufallsgrößen. Dann hat der Vektor $X = (X_1, \dots, X_n)^{\mathrm{T}}$ die Dichte

$$f(x) = f(x_1, \dots, x_n) = \frac{1}{(2\pi)^{n/2}} \exp\Big\{-\frac{1}{2}\sum_{i=1}^n x_i^2\Big\}, \qquad x = (x_1, \dots, x_n) \in \mathbb{R}^n.$$

Wir nennen den Vektor X *n-dimensional standardnormalverteilt*. Wir gehen im Folgenden davon aus, dass X ein *Spaltenvektor* ist; mit X^{T} bezeichnen wir den Zeilenvektor $(X_1, \dots, X_n)$.

Sei nun A eine reguläre $n \times n$-Matrix und $\mu \in \mathbb{R}^n$ ein Vektor, sowie $\theta(x) = Ax + \mu$ für $x \in \mathbb{R}^n$. Wir sagen, dass der Vektor $Y = \theta(X) = (Y_1, \dots, Y_n)^{\mathrm{T}}$ eine *(allgemeine) Normalverteilung* besitzt. Dann besitzt Y die Dichte

$$g(y) = g(y_1, \dots, y_n) = \frac{1}{|\det(C)|^{1/2}(2\pi)^{n/2}} \exp\Big\{-\frac{1}{2}(y-\mu)^{\mathrm{T}} C^{-1}(y-\mu)\Big\}, \qquad y \in \mathbb{R}^n,$$

wobei $C = AA^{\mathrm{T}}$, und A^{T} ist die Transponierte der Matrix A. Dass die Dichte von $Y = AX + \mu$ diese Form haben muss, sieht man ein, indem man die lineare Substitution $y = Ax + \mu$, also $x = A^{-1}(y - \mu)$, durchführt und beachtet, dass gilt:

$$(y-\mu)^{\mathrm{T}} C^{-1}(y-\mu) = (y-\mu)^{\mathrm{T}}(A^{\mathrm{T}})^{-1}A^{-1}(y-\mu) = x^{\mathrm{T}}x = \sum_{i=1}^n x_i^2.$$

Aus der Theorie der mehrdimensionalen Integration ist bekannt, dass die Integrationsvariablen der Regel $\mathrm{d}y = |\det A|\,\mathrm{d}x = \sqrt{|\det C|}\,\mathrm{d}x$ gehorchen.

Wir definieren Erwartungswerte von Zufallsvektoren komponentenweise, also ist $\mathbb{E}(Y) = (\mathbb{E}(Y_1), \dots, \mathbb{E}(Y_n))^{\mathrm{T}}$, und es ergibt sich, dass $\mathbb{E}(Y) = A\mathbb{E}(X) + \mu = \mu$ ist. Ferner ist die *Kovarianzmatrix* $\mathrm{cov}(Y,Y) = (\mathrm{cov}(Y_i, Y_j))_{i,j=1,\dots,n}$ gegeben durch

$$\mathrm{cov}(Y_i, Y_j) = \mathbb{E}\big((Y_i - \mathbb{E}(Y_i))(Y_j - \mathbb{E}(Y_j))\big), \qquad i,j = 1, \dots, n.$$

Sie erfüllt

$$\begin{aligned}\mathrm{cov}(Y,Y) &= \mathbb{E}\big((Y - \mathbb{E}(Y))(Y - \mathbb{E}(Y))^{\mathrm{T}}\big) = \mathbb{E}\big((AX)(AX)^{\mathrm{T}}\big)\\ &= \mathbb{E}(AXX^{\mathrm{T}}A^{\mathrm{T}}) = A\mathbb{E}(XX^{\mathrm{T}})A^{\mathrm{T}} = AA^{\mathrm{T}}\\ &= C.\end{aligned}$$

Also ist (wie im eindimensionalen Fall) die Verteilung des normalverteilten Vektors Y festgelegt durch den Erwartungswert und die Kovarianzmatrix. Man nennt Y die (n-dimensionale) Normalverteilung mit Kovarianzmatrix C und Erwartungswertvektor μ und schreibt auch oft $Y \sim \mathcal{N}(\mu, C)$. ◇

Beispiel 4.3.10 (Cauchy-Verteilung)**.** Die Dichte der *Cauchy-Verteilung* mit Parameter $c \in (0, \infty)$ ist gegeben durch

$$f(t) = \frac{c}{\pi}\,\frac{1}{t^2 + c^2}, \qquad \text{für } t \in \mathbb{R}.$$

Die Verteilungsfunktion ist gegeben durch $F(t) = \frac{1}{2} + \frac{1}{\pi}\arctan(\frac{t}{c})$. Der Erwartungswert der Cauchy-Verteilung existiert nicht, da die Funktion $t \mapsto \frac{|t|}{t^2+1}$ nicht über $\mathbb{R}$ integrierbar ist. ◇

4.4 Der Poisson-Prozess

In diesem Abschnitt diskutieren wir ein wichtiges mathematisches Modell für das Eintreten einer Folge von zufälligen Zeitpunkten. Dieses Modell wurde schon in Beispiel 1.3.6 angedeutet, doch für eine befriedigende Behandlung ist eine Kenntnis der Exponential- und der Gamma-Verteilungen notwendig (siehe Beispiele 4.3.2 und 4.3.5). Wir werden zunächst den Poisson-Prozess axiomatisch einführen und den Zusammenhang mit Poisson-verteilten Zufallsvariablen diskutieren. Danach geben wir eine Charakterisierung in Termen von exponentiellen Zufallsvariablen. Die Existenz des Poisson-Prozesses können wir hier noch nicht beweisen, und wir werden auch nicht alle Beweise ausformulieren.

Gegeben seien zufällige Zeitpunkte auf der positiven Zeitachse $(0,\infty)$. Dies können Zeitpunkte von Abfahrten von Bussen von einer Haltestelle sein oder von eingehenden Telefonanrufen oder vieles Andere mehr. Für jedes endliche halboffene Zeitintervall $I = (a, b]$ mit $0 \le a < b < \infty$ (für ein solches Intervall werden wir im Folgenden nur kurz 'Intervall' sagen) sei N_I die Anzahl derjenigen zufälligen Zeitpunkte, die in das Intervall I fallen. Für die Kollektion der $\mathbb{N}_0$-wertigen Zufallsgrößen N_I machen wir folgende Annahmen:

(P1) Die Verteilung von N_I hängt nur von der Länge des Intervalls I ab.

(P2) Wenn $I_1, \dots, I_k$ paarweise disjunkte Intervalle sind, dann sind $N_{I_1}, \dots, N_{I_k}$ unabhängig.

(P3) Für jedes Intervall I existiert $\mathbb{E}(N_I)$.

(P4) Es gibt ein Intervall I mit $\mathbb{P}(N_I > 0) > 0$.

(P5) Es gilt $\lim_{\varepsilon \downarrow 0} \varepsilon^{-1} \mathbb{P}(N_{(0,\varepsilon]} \ge 2) = 0$.

Eine Kollektion von Zufallsgrößen N_I, die die Eigenschaften (P1)-(P5) erfüllen, nennen wir einen *Poisson'schen Punktprozess* oder kurz einen *Poisson-Prozess*. Oft nennt man auch den Prozess $(N_{(0,t]})_{t \in [0,\infty)}$ einen Poisson-Prozess. (P1) und (P2) sind starke Strukturannahmen, die die mathematische Behandlung vereinfachen bzw. ermöglichen sollen. (P3) und (P4) schließen unerwünschte pathologische Fälle aus, und (P5) verhindert, dass sich die Zeitpunkte zu stark häufen können. Wir werden nicht beweisen, dass ein Poisson-Prozess existiert, sondern wir werden dies annehmen und diesen Prozess genauer untersuchen. Es stellt sich heraus, dass alle Zählvariablen N_I Poisson-verteilt sind, eine Tatsache, die den Namen erklärt:

Lemma 4.4.1. *Wenn (P1)-(P5) erfüllt sind, so existiert ein $\alpha > 0$, sodass für alle $t, s > 0$ die Zufallsvariable $N_{(t,t+s]}$ Poisson-verteilt ist mit Parameter αs. Insbesondere ist $\mathbb{E}(N_I) = \alpha |I|$ für jedes Intervall I.*

Beweis. Zunächst identifizieren wir den Erwartungswert jedes N_I, der ja nach (P3) endlich ist. Die Funktion $\alpha(t) = \mathbb{E}(N_{(0,t]})$ erfüllt $\alpha(0) = 0$ sowie

$$\alpha(t+s) = \mathbb{E}(N_{(0,t]} + N_{(t,t+s]}) = \mathbb{E}(N_{(0,t]}) + \mathbb{E}(N_{(0,s]}) = \alpha(t) + \alpha(s),$$

wobei wir im zweiten Schritt (P1) benutzten. Mit Hilfe von ein wenig Maßtheorie sieht man, dass $\lim_{t \downarrow 0} \alpha(t) = 0$, d. h. $\alpha(\cdot)$ ist stetig in 0. Eine beliebte Übungsaufgabe aus der Analysis zeigt, dass es ein $\alpha \ge 0$ gibt mit $\alpha(t) = \alpha t$ für jedes $t \ge 0$. Wegen (P4) muss $\alpha > 0$ sein.

Nun beweisen wir die erste Aussage des Lemmas. Wegen (P1) reicht es, $N_{(0,s]}$ zu betrachten. Wir zerlegen $(0, s]$ in die Intervalle $I_j^{(k)} = (\frac{s}{k}(j-1), \frac{s}{k} j]$ mit $j \in \{1, \dots, k\}$ und betrachten die Zufallsvariable $X_j^{(k)} = N_{I_j^{(k)}}$. Dann gilt offensichtlich $N_{(0,s]} = \sum_{j=1}^k X_j^{(k)}$. Ferner approximieren wir $X_j^{(k)}$ mit der Indikatorvariable $\overline{X}_j^{(k)} = \mathbb{1}\{X_j^{(k)} > 0\}$, die registriert, ob das j-te Intervall $I_j^{(k)}$ mindestens einen der zufälligen Punkte erhält oder nicht. Wegen (P2) in Kombination mit Korollar 3.2.12 sind die Variablen $\overline{X}_1^{(k)}, \dots, \overline{X}_k^{(k)}$ unabhängig, und wegen (P1) sind sie identisch verteilt. Mit anderen Worten, sie sind Bernoulli-Variable mit Parameter $p_k = \mathbb{P}(N_{(0,s/k]} > 0)$.

Nun definieren wir $\overline{N}_{(0,s]}^{(k)} = \sum_{j=1}^k \overline{X}_j^{(k)}$, dann ist $\overline{N}_{(0,s]}^{(k)}$ eine binomialverteilte Zufallsvariable mit Parametern k und p_k. Es gilt offensichtlich $\overline{N}_{(0,s]}^{(k)} \le N_{(0,s]}$.

Wir benutzen nun (P5), um zu zeigen, dass für große k die Variablen $\overline{N}^{(k)}_{(0,s]}$ und $N_{(0,s]}$ sehr nahe bei einander liegen, d. h. dass gilt:

$$\lim_{k\to\infty} \mathbb{P}(\overline{N}^{(k)}_{(0,s]} = m) = \mathbb{P}(N_{(0,s]} = m), \qquad m \in \mathbb{N}_0. \tag{4.4.1}$$

Dies sieht man ein, indem man abschätzt:

$$\mathbb{P}(\overline{N}^{(k)}_{(0,s]} \neq N_{(0,s]}) = \mathbb{P}(\overline{N}^{(k)}_{(0,s]} < N_{(0,s]}) \leq \sum_{j=1}^{k} \mathbb{P}(X^{(k)}_j \geq 2) = k\mathbb{P}(N_{(0,s/k]} \geq 2),$$

und dies konvergiert gegen Null für $k \to \infty$ wegen (P5). Da $|\mathbb{P}(\overline{N}^{(k)}_{(0,s]} = m) - \mathbb{P}(N_{(0,s]} = m)| \leq 2\mathbb{P}(\overline{N}^{(k)}_{(0,s]} \neq N_{(0,s]})$, folgt (4.4.1).

Mit Hilfe von (4.4.1) zeigt man nun, dass $\lim_{k\to\infty} kp_k = \alpha s$ ist: Wir haben

$$\lim_{k\to\infty} kp_k = \lim_{k\to\infty} \mathbb{E}(\overline{N}^{(k)}_{(0,s]}) = \lim_{k\to\infty} \sum_{l=1}^{\infty} \mathbb{P}(\overline{N}^{(k)}_{(0,s]} \geq l) = \sum_{l=1}^{\infty} \mathbb{P}(N_{(0,s]} \geq l) = \mathbb{E}(N_{(0,s]}) = \alpha s,$$

wobei wir im zweiten und im vorletzten Schritt Lemma 3.3.8 benutzten und im dritten eine allgemeine Tatsache über Reihen (wie im Beweis von (4.4.1) zeigt man nämlich, dass $\lim_{k\to\infty} \mathbb{P}(\overline{N}^{(k)}_{(0,s]} \geq l) = \mathbb{P}(N_{(0,s]} \geq l)$ für jedes $l \in \mathbb{N}_0$ gilt). Also kann man den Poisson'schen Grenzwertsatz (Satz 1.3.7) anwenden und erhält

$$\lim_{k\to\infty} \mathbb{P}(\overline{N}^{(k)}_{(0,s]} = m) = \lim_{k\to\infty} \mathrm{Bi}_{k,p_k}(m) = \mathrm{Po}_{\alpha s}(m), \qquad m \in \mathbb{N}_0.$$

D. h., die Zufallsvariable $\overline{N}^{(k)}_{(0,s]}$ ist asymptotisch Poisson-verteilt mit Parameter αs. Da diese Grenzverteilung mit der Verteilung von $N_{(0,s]}$ übereinstimmt, ist der Beweis beendet. □

Wir nähern uns nun einem anderen Zugang zum Poisson-Prozess. Ausgangspunkt der Überlegung ist folgende Kombination der beiden Beobachtungen von Lemma 4.3.6: Die Wahrscheinlichkeit, dass im Intervall $(0,t]$ mindestens k Punkte liegen (dies ist gleich $\mathbb{P}(N_{(0,t]} \geq k)$, und $N_{(0,t]}$ ist nach Lemma 4.4.1 Poisson-verteilt mit Parameter αt), ist für jedes $k \in \mathbb{N}_0$ gegeben durch $\mathbb{P}(\sum_{i=1}^k \tau_i \leq t)$, wobei $\tau_1, \tau_2, \dots$ eine Folge unabhängiger, zum Parameter α exponentiell verteilten Zufallsgrößen ist (denn $\sum_{i=1}^k \tau_i$ ist Gamma-verteilt mit Parametern α und k). Dies legt den Schluss nahe, dass die Zufallszeiten τ_i eine Interpretation als die Wartezeiten zwischen dem Eintreffen des $(i-1)$-ten und i-ten zufälligen Zeitpunktes zulassen, und genau das ist der Fall:

Satz 4.4.2. *Es sei $(\tau_i)_{i\in\mathbb{N}}$ eine Folge unabhängiger, zum Parameter α exponentiell verteilten Zufallsgrößen. Wir definieren $T_k = \sum_{i=1}^k \tau_i$ für $k \in \mathbb{N}$. Für jedes endliche halboffene Intervall I definieren wir $N_I = |\{k \in \mathbb{N} \colon T_k \in I\}|$ als die Zahl der T_k, die in I fallen. Dann erfüllt die Kollektion der Zufallsgrößen N_I die Bedingungen (P1)-(P5).*

Beweisskizze. Wir werden nicht die volle Aussage beweisen; der vollständige Beweis des Satzes ist komplizierter, aber analog. Wir zeigen nur die Aussage: Für jedes $0 < s < t$ sind $N_{(0,s]}$ und $N_{(s,t]}$ unabhängige, zu den Parametern αs bzw. $\alpha(t-s)$ Poisson-verteilte Zufallsgrößen, d. h.

$$\mathbb{P}(N_{(0,s]} = k, N_{(s,t]} = l) = \mathrm{Po}_{\alpha s}(k)\mathrm{Po}_{\alpha(t-s)}(l) = e^{-\alpha t}\alpha^{k+l}\frac{s^k}{k!}\frac{(t-s)^l}{l!}, \qquad k,l \in \mathbb{N}_0. \tag{4.4.2}$$

Zunächst sieht man, dass $\{N_{(0,s]} = k, N_{(s,t]} = l\} = \{T_k \leq s < T_{k+1}, T_{k+l} \leq t < T_{k+l+1}\}$. Das betrachtete Ereignis lässt sich also mit Hilfe der τ_i ausdrücken als das Ereignis, dass der Zufallsvektor $(\tau_1, \dots, \tau_{k+l+1})$ in der Menge

$$A = \big\{x \in [0,\infty)^{k+l+1} \colon S_k(x) \leq s < S_{k+1}(x), S_{k+l}(x) \leq t < S_{k+l+1}(x)\big\}$$

liegt (wobei wir $S_n(x) = x_1 + \dots + x_n$ setzten):

$$\{N_{(0,s]} = k, N_{(s,t]} = l\} = \big\{(\tau_1, \dots, \tau_{k+l+1}) \in A\big\}.$$

Nach Lemma 4.2.4 ist eine Dichte des Zufallsvektors $(\tau_1, \dots, \tau_{k+l+1})$ gegeben durch

$$x = (x_1, \dots, x_{k+l+1}) \mapsto 1\!\!1_{[0,\infty)^{k+l+1}}(x)\alpha^{k+l+1} e^{-\alpha S_{k+l+1}(x)}.$$

Wir zeigen nun die Gleichung (4.4.2) für $l \geq 1$ (der Fall $l = 0$ ist analog). Es gilt

$$\begin{aligned}
\mathbb{P}(N_{(0,s]} = k, N_{(s,t]} = l) = \mathbb{P}\big((\tau_1, \dots, \tau_{k+l+1}) \in A\big) &= \int_A \mathrm{d}x\, \alpha^{k+l+1} e^{-\alpha S_{k+l+1}(x)} \\
&= \alpha^{k+l} \int_0^\infty \cdots \int_0^\infty \mathrm{d}x_1 \dots \mathrm{d}x_{k+l+1}\, \alpha e^{-\alpha S_{k+l+1}(x)} \\
&\qquad \times 1\!\!1\{S_k(x) \leq s < S_{k+1}(x), S_{k+l}(x) \leq t < S_{k+l+1}(x)\}.
\end{aligned}$$

Wir integrieren nun schrittweise von innen nach außen. Zuerst halten wir $x_1, \dots, x_{k+l}$ fest und substituieren $z = S_{k+l+1}(x)$:

$$\int_0^\infty \mathrm{d}x_{k+l+1}\, \alpha e^{-\alpha S_{k+l+1}(x)} 1\!\!1\{t < S_{k+l+1}(x)\} = \int_t^\infty \mathrm{d}z\, \alpha e^{-\alpha z} = e^{-\alpha t}.$$

Nun halten wir $x_1, \dots, x_k$ fest und substituieren $y_1 = S_{k+1}(x) - s, y_2 = x_{k+2}, \dots, y_l = x_{k+l}$:

$$\begin{aligned}
&\int_0^\infty \cdots \int_0^\infty \mathrm{d}x_{k+1} \dots \mathrm{d}x_{k+l}\, 1\!\!1\{s < S_{k+1}(x), S_{k+l}(x) \leq t\} \\
&\qquad = \int_0^\infty \cdots \int_0^\infty \mathrm{d}y_1 \dots \mathrm{d}y_l\, 1\!\!1\{y_1 + \dots + y_l \leq t - s\} = \frac{(t-s)^l}{l!},
\end{aligned}$$

wobei wir eine Induktion über l benutzten. Die restlichen k Integrale behandeln wir genauso und erhalten

$$\int_0^\infty \dots \int_0^\infty \mathrm{d}x_1 \dots \mathrm{d}x_k\, 1\!\!1\{S_k(x) \leq s\} = \frac{s^k}{k!}.$$

Wenn man dies alles zusammensetzt, ergibt sich die Behauptung. □

Die Konstruktion des Poisson-Prozesses, die durch Satz 4.4.2 gegeben wird, hängt für uns in der Luft, da wir noch nicht unendlich viele unabhängige Zufallsgrößen mathematisch korrekt konstruieren können. Aber wir haben den Poisson-Prozess sehr befriedigend charakterisiert: Wir können ihn uns vorstellen als der Prozess von Zeitpunkten, zwischen denen unabhängige exponentielle Wartezeiten mit Parameter α vergehen. Mehr noch, die folgenden drei Charakterisierungen sind zueinander äquivalent:

1. Die $\mathbb{N}_0$-wertigen Zufallsgrößen N_I erfüllen (P1)–(P5).
2. Die Zeitdifferenzen τ_i zwischen dem $(i-1)$-ten und i-ten der zufälligen Zeitpunkte sind unabhängige, zu einem gemeinsamen Parameter α exponentiell verteilte Zufallsgrößen.

3. Mit einem geeigneten Parameter α ist für jedes Intervall I die Zufallsgröße N_I zum Parameter $\alpha|I|$ Poisson-verteilt, und bei paarweiser Disjunktheit der Intervalle $I_1, \dots, I_k$ sind $N_{I_1}, \dots, N_{I_k}$ unabhängig.

Die Äquivalenz dieser drei Aussagen zu beweisen, ist eine (anspruchsvolle) Übungsaufgabe, wobei wesentliche Teile in Lemma 4.4.1 und Satz 4.4.2 enthalten sind. Im Folgenden werden wir die Äquivalenz benutzen.

Zum Schluss betrachten wir noch ein kleines Kuriosum, das *Wartezeitparadox*. Wenn man einen Poisson-Prozess als Modell für Zeitpunkte, an denen Busse von einer Haltestelle abfahren, verwendet, stellt man sich vielleicht folgende Frage: Wenn ich zum (festen) Zeitpunkt $t > 0$ an der Haltestelle ankomme und nicht weiß, wann der letzte Bus abgefahren ist, was ist die Verteilung der Wartezeit auf den nächsten Bus? Ist die erwartete Wartezeit vielleicht kürzer als $1/\alpha$? Und wie steht es um die Zeit, die seit der Abfahrt des letzten Busses bis jetzt (also den Zeitpunkt t) vergangen ist? Hat die Summe dieser zwei Zeiten den Erwartungswert $1/\alpha$? Die Antwort ist ein wenig überraschend und folgt im nächsten Lemma. Wir nehmen an, dass $N_t = N_{(0,t]}$, wobei $(N_I)_I$ ein wie im Satz 4.4.2 konstruierter Poisson-Prozess ist. Ferner definieren wir für $t > 0$

$$W_t = -t + \min\{T_k \colon k \in \mathbb{N}, T_k > t\} \qquad V_t = t - \max\{T_k \colon k \in \mathbb{N}, T_k \leq t\},$$

wobei wir $\max \emptyset = 0$ setzen. In Worten: W_t ist die Wartezeit ab t bis zur Abfahrt des nächsten Busses, und V_t ist die Zeitdifferenz zwischen der letzten Abfahrt vor dem Zeitpunkt t und t.

Lemma 4.4.3 (Wartezeitparadox). *Für jedes $t > 0$ ist W_t zum Parameter α exponentialverteilt, und V_t hat die Verteilung von $\min\{W_t, t\}$. Insbesondere gelten $\mathbb{E}(W_t) = 1/\alpha$ und $\mathbb{E}(V_t) = \frac{1}{\alpha}(1 - e^{-\alpha t})$. Die Zufallszeiten W_t und V_t sind unabhängig.*

Beweis. Das Ereignis $\{W_t > s\}$ ist identisch mit dem Ereignis, dass zwischen den Zeitpunkten t und $t + s$ keines der T_k eintrifft. Also gilt $\mathbb{P}(W_t > s) = \mathbb{P}(N_{(t,t+s]} = 0) = \mathrm{Po}_{\alpha s}(0) = e^{-\alpha s}$. Dies zeigt die erste Aussage.

Das Ereignis $\{V_t > s\}$ ist für $s < t$ identisch mit dem Ereignis, dass zwischen den Zeitpunkten $t-s$ und t keines der T_k eintrifft. Analog zu dem Obigen erhält man, dass $\mathbb{P}(V_t > s) = \mathbb{P}(N_{(t-s,t]} = 0) = e^{-\alpha s}$ für $s < t$. Für $s \geq t$ ist $\{V_t > s\} = \emptyset$, also $\mathbb{P}(V_t > s) = 0$. Setzt man diese zwei Teilaussagen zusammen, erhält man die zweite Aussage des Lemmas.

Die Berechnung des Erwartungswertes von $\min\{W_t, t\}$ ist eine Übungsaufgabe. Die Unabhängigkeit der Zufallszeiten W_t und V_t ist eine einfache Konsequenz von Satz 4.4.2, denn W_t und V_t hängen nur von der Zahl der T_k in (t, ∞) bzw. in $(0, t]$ ab, und diese Anzahlen sind nach (P2) unabhängig (siehe Korollar 3.2.12). □

Die Wartezeit zwischen der Abfahrt des letzten Busses vor dem Zeitpunkt t und des nächsten nach t hat also *nicht* die Verteilung irgendeiner der Wartezeiten τ_i. Tatsächlich ist ihr Erwartungswert größer und tendiert für große t gegen das Zweifache des Erwartungswertes von τ_i. Das Phänomen, dass dieses beobachtete der zufälligen Zeitintervalle größer ist als jedes der anderen, liegt nicht an der exponentiellen Verteilung, sondern an der einfachen Überlegung, dass größere solche Intervalle natürlich auch mit einer höheren Wahrscheinlichkeit 'erwischt' werden als kleinere.

Kapitel 5

Grenzwertsätze

In diesem Kapitel behandeln wir die zwei wichtigsten Grenzwertsätze für Wahrscheinlichkeitsverteilungen: das *Gesetz der Großen Zahlen* und den *Zentralen Grenzwertsatz*, zumindest für wichtige Spezialfälle. Beide Sätze machen asymptotische Aussagen über sehr oft wiederholte unabhängige identisch verteilte Zufallsexperimente: Das Gesetz der Großen Zahlen formuliert, dass der Durchschnitt der dabei auftretenden Ergebnisse sich in gewissem Sinne dem erwarteten Wert eines dieser Experimente annähert, und der Zentrale Grenzwertsatz macht eine feinere Aussage über die Fluktuationen um diesen asymptotischen Wert.

Unter einer Zufallsgröße verstehen wir in diesem Kapitel eine diskrete im Sinne der Definition 3.1.1 oder eine stetige im Sinne von Abschnitt 4.1.

5.1 Das Gesetz der Großen Zahlen

Wir beginnen mit zwei einfachen, aber wichtigen allgemeinen Ungleichungen für die Wahrscheinlichkeit von Abweichungen einer Zufallsgröße von Null bzw. von ihrem Erwartungswert. Wir erinnern daran, dass der Erwartungswert einer nicht negativen Zufallsgröße immer definiert ist, aber eventuell gleich ∞ ist.

Satz 5.1.1 (Markov-Ungleichung)**.** *Es sei X eine Zufallsgröße und $\varphi\colon (0,\infty) \to (0,\infty)$ eine monoton wachsende Funktion. Dann gilt für jedes $\varepsilon > 0$*

$$\mathbb{P}(|X| \geq \varepsilon) \leq \frac{\mathbb{E}(\varphi \circ |X|)}{\varphi(\varepsilon)}.$$

Beweis. Auf der Menge $\{|X| \geq \varepsilon\} = \{\omega \in \Omega\colon |X(\omega)| \geq \varepsilon\}$ gilt wegen der Monotonie von φ, dass $\varphi(\varepsilon) \leq \varphi(|X(\omega)|)$. Also gilt die Abschätzung

$$\mathbb{1}_{\{|X| \geq \varepsilon\}} \leq \frac{\varphi \circ |X|}{\varphi(\varepsilon)}.$$

Nun bildet man auf beiden Seiten den Erwartungswert und erhält die behauptete Ungleichung. $\square$

Indem man die Markov-Ungleichung auf die Zufallsvariable $X - \mathbb{E}(X)$ statt X und $\varphi(x) = x^2$ anwendet, erhält man die sehr nützliche folgende Ungleichung:

Korollar 5.1.2 (Tschebyscheff-Ungleichung). *Für jede Zufallsgröße $X \in \mathcal{L}^2$ und jedes $\varepsilon > 0$ gilt*

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq \varepsilon) \leq \frac{\mathbb{V}(X)}{\varepsilon^2}.$$

Die Tschebyscheff-Ungleichung kann im Allgemeinen nicht verbessert werden, wie das folgende Beispiel zeigt. Man betrachte eine Zufallsvariable X, die die Werte ε, 0 und $-\varepsilon$ annimmt mit den Wahrscheinlichkeiten $(2\varepsilon^2)^{-1}$, $1 - \varepsilon^{-2}$ und $(2\varepsilon^2)^{-1}$. Dann sind $\mathbb{E}(X) = 0$ und $\mathbb{V}(X) = 1$, und es gilt Gleichheit in der Tschebyscheffschen Ungleichung.

Der Wert der Tschebyscheff-Ungleichung liegt in ihrer einfachen Handhabbarkeit und universellen Anwendbarkeit, sie gibt einen recht guten Eindruck von der Größenordnung der betrachteten Wahrscheinlichkeit. Sie ist immerhin gut genug, um einen kurzen Beweis des zentralen Ergebnisses dieses Abschnittes zu ermöglichen (siehe Satz 5.1.4).

Wir kommen nun zum Gesetz der Großen Zahlen. Wie so oft betrachten wir eine oftmalige Wiederholung eines zufälligen Experiments, bei dem jeweils ein zufälliges Ergebnis erzielt wird, das man mit einer Zufallsgröße angeben kann. Sei also $X_i \in \mathbb{R}$ das Ergebnis der i-ten Ausführung. Wir nehmen an, dass jede Zufallsgröße X_i den gleichen Erwartungswert $E = \mathbb{E}(X_1) = \mathbb{E}(X_2) = \dots$ besitzt. Die Intuition sagt, dass die Folge der Mittelwerte $\frac{1}{n} S_n = \frac{1}{n}(X_1 + \dots + X_n)$ sich für große n der Zahl E annähern sollte. Doch in welchen Sinn sollte das passieren? Eine Aussage wie '$\lim_{n\to\infty} \frac{1}{n} S_n = E$' können wir noch nicht exakt behandeln, denn dazu müssten alle (unendlich vielen) Zufallsgrößen $X_1, X_2, \dots$ gleichzeitig definiert sein.[1] Wir werden die Annäherung von $\frac{1}{n} S_n$ an E in der Form formulieren, dass die Wahrscheinlichkeit, dass $\frac{1}{n} S_n$ von E einen gewissen positiven Abstand hat, mit $n \to \infty$ gegen Null geht. Dies gibt Anlass zu einer Definition:

Definition 5.1.3 (Konvergenz in Wahrscheinlichkeit). *Wir sagen, eine Folge $(Y_n)_{n\in\mathbb{N}}$ von Zufallsgrößen* konvergiert in Wahrscheinlichkeit *oder* konvergiert stochastisch *gegen eine Zufallsgröße Y, falls für jedes $\varepsilon > 0$ gilt:*

$$\lim_{n\to\infty} \mathbb{P}(|Y_n - Y| > \varepsilon) = 0.$$

In diesem Fall schreiben wir auch $Y_n \xrightarrow{\mathbb{P}} Y$.

Es ist klar, dass $Y_n \xrightarrow{\mathbb{P}} Y$ genau dann gilt, wenn $Y_n - Y \xrightarrow{\mathbb{P}} 0$. Man beachte, dass die Konvergenz in Wahrscheinlichkeit nicht von den Zufallsgrößen abhängt, sondern nur von ihrer Verteilung. Insbesondere müssen für diesen Konvergenzbegriff nicht unendlich viele Zufallsgrößen auf einem Wahrscheinlichkeitsraum definiert werden, sondern jeweils nur eine einzige, nämlich $Y_n - Y$, dies allerdings für jedes $n \in \mathbb{N}$.

Der Begriff der Konvergenz in Wahrscheinlichkeit ist tatsächlich ein sehr geeigneter für die oben gestellte Frage nach einer sinnvollen Formulierung der Annäherung von $\frac{1}{n} S_n$ an E. Wir

[1] Tatsächlich kann man diese Aussage (unter Zuhilfenahme von Maßtheorie) präzisieren und unter geeigneten Voraussetzungen beweisen; siehe später. Eine solche Aussage nennt man das *Starke Gesetz der Großen Zahlen.*

erinnern daran (siehe Abschnitt 3.5), dass zwei Zufallsvariable X und Y unkorreliert heißen, wenn $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.

Satz 5.1.4 (Schwaches Gesetz der Großen Zahlen). *Für jedes $n \in \mathbb{N}$ seien paarweise unkorrelierte Zufallsgrößen $X_1, \dots, X_n$ gegeben, die alle den gleichen Erwartungswert $E \in \mathbb{R}$ und die gleiche Varianz $V < \infty$ besitzen. Sei $\frac{1}{n}S_n = \frac{1}{n}(X_1 + \dots + X_n)$ der Mittelwert. Dann gilt für jedes $\varepsilon > 0$:*

$$\lim_{n\to\infty} \mathbb{P}(|\tfrac{1}{n}S_n - E| > \varepsilon) = 0,$$

d. h., $\frac{1}{n}S_n$ konvergiert in Wahrscheinlichkeit gegen E.

Beweis. Auf Grund der Linearität des Erwartungswerts ist $\mathbb{E}(\frac{1}{n}S_n) = E$, und auf Grund der paarweisen Unkorreliertheit ist

$$\mathbb{V}(\tfrac{1}{n}S_n) = \tfrac{1}{n^2}\mathbb{V}(X_1 + \dots + X_n) = \tfrac{1}{n^2}n\mathbb{V}(X_1) = \tfrac{1}{n}V; \tag{5.1.1}$$

siehe den Satz von Bienaymé, Korollar 3.5.3. Also liefert eine Anwendung der Tschebyscheff-Ungleichung:

$$\mathbb{P}(|\tfrac{1}{n}S_n - E| > \varepsilon) = \mathbb{P}(|\tfrac{1}{n}S_n - \mathbb{E}(\tfrac{1}{n}S_n)| > \varepsilon) \leq \frac{\mathbb{V}(\frac{1}{n}S_n)}{\varepsilon^2} = \frac{1}{n}\frac{V}{\varepsilon^2},$$

und dies konvergiert gegen Null für $n \to \infty$. □

Insbesondere gilt das Gesetz der Großen Zahlen also auch für unabhängige identisch verteilte Zufallsgrößen mit existierenden Varianzen. Natürlich gibt es weit stärkere Versionen des Gesetzes der Großen Zahlen in der Literatur, aber darum kümmern wir uns hier nicht. Der Mittelwert $\frac{1}{n}S_n$ der n Ausführungen der Experimente liegt also in der Nähe des Erwartungswertes einer einzelnen Ausführung in dem Sinne, dass Abweichungen einer gegebenen positiven Größe eine verschwindende Wahrscheinlichkeit haben. Man beachte aber, dass im Allgemeinen *nicht* $\lim_{n\to\infty} \mathbb{P}(\frac{1}{n}S_n \neq E) = 0$ gilt.

Unter gewissen zusätzlichen Integrierbarkeitsannahmen kann man sogar erhalten, dass die Geschwindigkeit der Konvergenz im Gesetz der Großen Zahlen sogar *exponentiell* ist:

Lemma 5.1.5 (Große Abweichungen). *Für jedes $n \in \mathbb{N}$ seien $X_1, \dots, X_n$ unabhängige, identisch verteilte Zufallsvariablen. Es existiere ein $\alpha > 0$, so dass $\mathbb{E}(e^{\alpha|X_1|}) < \infty$. Wir definieren wieder $E = \mathbb{E}(X_1)$ und $S_n = X_1 + \dots + X_n$. Dann existiert für jedes $\varepsilon > 0$ ein $C > 0$ (das von ε und von α abhängt), so dass*

$$\mathbb{P}(|\tfrac{1}{n}S_n - E| > \varepsilon) \leq e^{-Cn}, \qquad n \in \mathbb{N}.$$

Beweis. Wir dürfen voraus setzen, dass $E = \mathbb{E}(X_1) = 0$. Wir werden nur zeigen, dass $\mathbb{P}(\frac{1}{n}S_n > \varepsilon) \leq e^{-Cn}$ für alle $n \in \mathbb{N}$ und ein geeignetes C. Der Beweis der anderen Hälfte der Aussage, $\mathbb{P}(\frac{1}{n}S_n < -\varepsilon) \leq e^{-Cn}$, läuft analog, und dann folgt die Aussage des Lemmas mit einem eventuell anderen Wert von $C > 0$.

Wir fixieren ein $\beta \in (0, \alpha/2)$ und benutzen die Markov-Ungleichung (siehe Satz 5.1.1) für die Abbildung $\varphi(x) = e^{\beta x}$ wie folgt:

$$\mathbb{P}(\tfrac{1}{n}S_n > \varepsilon) = \mathbb{P}(X_1 + \dots + X_n > \varepsilon n) = \mathbb{P}\big(e^{\beta(X_1+\dots+X_n)} > e^{\beta\varepsilon n}\big) \leq \mathbb{E}\big(e^{\beta(X_1+\dots+X_n)}\big)e^{-\beta\varepsilon n}.$$

Den auftretenden Erwartungswert kann man mit Hilfe der Unabhängigkeit von $e^{\beta X_1}, \dots, e^{\beta X_n}$ (siehe Korollar 3.2.12) und Lemma 3.3.2(d) berechnen zu

$$\mathbb{E}\big(e^{\beta(X_1+\dots+X_n)}\big) = \mathbb{E}(e^{\beta X_1})^n.$$

Auf Grund unserer Integrierbarkeitsvoraussetzung und wegen $\beta < \alpha$ ist diese obere Schranke endlich. Wenn wir sie oben einsetzen, erhalten wir die Abschätzung

$$\mathbb{P}(\tfrac{1}{n} S_n > \varepsilon) \le \exp\Big\{-n\big[\beta\varepsilon - \log \mathbb{E}(e^{\beta X_1})\big]\Big\}. \tag{5.1.2}$$

Nun zeigen wir, dass für genügend kleines $\beta > 0$ der Term in $[\dots]$ in (5.1.2) positiv ist. Wir schreiben nun X statt X_1. Für jedes $\beta \in (0, \frac{\alpha}{2})$ ist $e^{\beta X} = 1 + \beta X + \beta^2 Y_\beta$ mit

$$Y_\beta = X^2 \sum_{k=2}^{\infty} \frac{(\beta X)^{k-2}}{k!}.$$

Um zu sehen, dass $\sup_{\beta\in(0,\alpha/2)} |Y_\beta|$ integrierbar ist, schätzen wir ab mit Hilfe eines $R > 0$, das so groß (nur abhängig von α) ist, dass $x^2 \le e^{\frac{\alpha}{2}|x|}$ für jedes $x \in \mathbb{R}$ mit $|x| > R$ gilt:

$$Y_\beta \le X^2 e^{\beta|X|} \le X^2 e^{\frac{\alpha}{2}|X|} \le \mathbb{1}_{\{|X|\le R\}} R^2 e^{\frac{\alpha}{2}R} + \mathbb{1}_{\{|X|>R\}} e^{\alpha|X|} \le R^2 e^{\frac{\alpha}{2}R} + e^{\alpha|X|}.$$

Also ist $\sup_{\beta\in(0,\alpha/2)} |Y_\beta|$ integrierbar, und wir erhalten

$$\mathbb{E}(e^{\beta X}) = 1 + \beta\mathbb{E}(X) + \beta^2 \mathbb{E}(Y_\beta) = 1 + \mathcal{O}(\beta^2)$$

für $\beta \downarrow 0$. Nun ist leicht zu sehen, dass für genügend kleines $\beta > 0$ der Term in $[\dots]$ in (5.1.2) positiv ist, und dies beendet den Beweis. □

5.2 Der Zentrale Grenzwertsatz

Wir gehen zurück zu der im vorigen Abschnitt beschriebenen Situation von oftmaligen Ausführungen eines Zufallsexperiments und fragen uns: Wenn also $\frac{1}{n} S_n - E$ gegen Null geht, mit welcher Rate passiert denn das?[2] Ist diese Größe typischerweise von der Ordnung n^{-1} oder $1/\log n$ oder e^{-n} oder von welcher sonst? Eine grobe Antwort wird gegeben durch eine etwas trickreichere Anwendung der Tschebyscheffschen Ungleichung, als dies im Beweis des Gesetzes der Großen Zahlen geschehen ist: Sie liefert für jedes $\alpha > 0$ die Abschätzung

$$\mathbb{P}\big(n^\alpha |\tfrac{1}{n} S_n - E| > \varepsilon\big) \le n^{2\alpha-1} \frac{V}{\varepsilon^2},$$

und dies konvergiert für jedes $\alpha < \frac{1}{2}$ gegen Null. Dies legt die Vermutung nahe, dass $\frac{1}{n} S_n - E$ von der Größenordnung $n^{-1/2}$ sein sollte und dass die 'aufgeblähte' Zufallsvariable $\sqrt{n}(\frac{1}{n} S_n - E)$ gegen etwas Nichttriviales konvergieren könnte. Dies stellt sich auch als korrekt heraus, wie wir im folgenden Satz 5.2.2 sehen werden. Tatsächlich kommt sogar in einer Vielzahl von Fällen immer die selbe Grenzverteilung heraus, und zwar die Normalverteilung (siehe Beispiel 4.3.7). Der Sinn, in dem die Konvergenz stattfindet, ist der folgende.

[2] Man beachte, dass diese Frage sehr verschieden ist von der Frage, mit welcher Rate die Wahrscheinlichkeit $\mathbb{P}(|\frac{1}{n} S_n - E| \ge \varepsilon)$ gegen Null geht.

Definition 5.2.1 (Schwache Konvergenz)**.** *Es seien Zufallsgrößen X und $X_1, X_2, \dots$ gegeben mit Verteilungsfunktionen F bzw. $F_1, F_2, \dots$. Wir sagen, dass X_n* in Verteilung *oder* schwach gegen *X* konvergiert*, geschrieben $X_n \stackrel{\mathrm{w}}{\to} X$, falls für jedes $t \in \mathbb{R}$, in dem F stetig ist, gilt*

$$\lim_{n\to\infty} F_n(t) = F(t).$$

Die Notation $X_n \stackrel{\mathrm{w}}{\to} X$ lehnt sich an dem englischen Ausdruck 'weak convergence' für schwache Konvergenz an. Natürlich konvergiert X_n genau dann schwach gegen X, wenn $X_n - X$ schwach gegen Null konvergiert. Wie beim Begriff der Konvergenz in Wahrscheinlichkeit im vorigen Abschnitt braucht man streng genommen keine Zufallsgrößen, um die Konvergenz zu formulieren, sondern nur deren Verteilungen, und zwar nur die von $X_n - X$ für jedes $n \in \mathbb{N}$. Man kann leicht zeigen, dass $X_n \stackrel{\mathrm{w}}{\to} X$ genau dann gilt, wenn für alle $a < b$, so dass die Verteilungsfunktion von X in a und in b stetig ist, gilt

$$\lim_{n\to\infty} \mathbb{P}(X_n \in [a,b]) = \mathbb{P}(X \in [a,b]).$$

Wir formulieren nun das zentrale Ergebnis dieses Abschnitts, das wir hier allerdings nur in einem Spezialfall beweisen werden, siehe Satz 5.2.5.

Satz 5.2.2 (Zentraler Grenzwertsatz)**.** *Für jedes $n \in \mathbb{N}$ seien unabhängige und identisch verteilte Zufallsgrößen $X_1, \dots, X_n$ gegeben, die alle den gleichen Erwartungswert E und die gleiche Varianz V besitzen. Sei $\frac{1}{n} S_n = \frac{1}{n}(X_1 + \dots, X_n)$ der Mittelwert. Dann gilt für jedes $t \in \mathbb{R}$*

$$\lim_{n\to\infty} \mathbb{P}\Big(\sqrt{\frac{n}{V}}\Big(\frac{1}{n} S_n - E\Big) \le t\Big) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^t e^{-\frac{x^2}{2}}\,\mathrm{d}x. \tag{5.2.1}$$

Mit anderen Worten, $\sqrt{\frac{n}{V}}(\frac{1}{n} S_n - E)$ konvergiert schwach gegen eine standardnormalverteilte Zufallsgröße:

$$\sqrt{\frac{n}{V}}\Big(\frac{1}{n} S_n - E\Big) \stackrel{\mathrm{w}}{\to} \mathcal{N}(0,1).$$

Beweis. Siehe Abschnitt 8.3. □

Man beachte, dass

$$\sqrt{\frac{n}{V}}\Big(\frac{1}{n} S_n - E\Big) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{X_i - \mathbb{E}(X_i)}{\sqrt{\mathbb{V}(X_i)}},$$

gleich $1/\sqrt{n}$ Mal eine Summe von n unabhängigen, identisch verteilten standardisierten Zufallsgrößen ist, wobei eine Zufallsgröße mit Erwartungswert Null und Varianz Eins *standardisiert* genannt wird. Der Zentrale Grenzwertsatz sagt also in Kürze:

$$X_1, \dots, X_n \text{ u.i.v. und standardisiert} \qquad \Longrightarrow \qquad \frac{1}{\sqrt{n}}(X_1 + \dots + X_n) \stackrel{\mathrm{w}}{\to} \mathcal{N}(0,1).$$

Bemerkung 5.2.3 (ZGWS $\Longrightarrow$ GGZ)**.** Die Aussage des Zentralen Grenzwertsatzes (ZGWS) ist stärker als die des Schwachen Gesetzes der Großen Zahlen (GGZ). Dies ist überhaupt nicht verwunderlich, wenn man sich klar macht, dass im ZGWS die Konvergenz von $S_n/\sqrt{n}$ behauptet wird (in welchem Sinn auch immer) und im GGZ die Konvergenz von S_n/n gegen Null (wir haben uns o.b.d.A. auf standardisierte Größen beschränkt, also $E = 0$ und $V = 1$ vorausgesetzt). Hier ist ein Beweis: Sei also die Gültigkeit der Aussage des Satzes 5.2.2 vorausgesetzt, und wir wollen zeigen, dass daraus das Schwache Gesetz der Großen Zahlen folgt. Sei also ein $\varepsilon > 0$ gegeben. Für jedes $C > 0$ und alle genügend großen n gilt $\varepsilon > C/\sqrt{n}$, also ist das Ereignis $\{|\frac{1}{n}S_n| \geq \varepsilon\}$ enthalten in $\{|\frac{1}{n}S_n| \geq C/\sqrt{n}\}$. Daher haben wir

$$\limsup_{n\to\infty} \mathbb{P}\Big(\Big|\frac{1}{n}S_n\Big| > \varepsilon\Big) \leq \lim_{n\to\infty} \mathbb{P}\Big(\Big|\frac{1}{n}S_n\Big| > \frac{C}{\sqrt{n}}\Big) = 2(1 - \Phi(C)),$$

und dies konvergiert für $C \to \infty$ gegen Null. Also gilt die Aussage des Schwachen Gesetzes der Großen Zahlen. ◇

Bemerkung 5.2.4 (Der Zentrale Grenzwertsatz als Fixpunktsatz)**.** Es ist bemerkenswert, dass die Grenzverteilung nicht abhängt von den Details der Verteilung der X_i und immer gleich der Normalverteilung $\mathcal{N}(0,1)$ ist. Ein wenig Plausibilität dafür kommt von der speziellen Eigenschaft der Normalverteilung in Lemma 4.3.8 her: Falls alle X_i exakt $\mathcal{N}(0,1)$-verteilt sind, so ist auch $\sqrt{\frac{n}{V}}(\frac{1}{n}S_n - E) = n^{-1/2}\sum_{i=1}^n X_i$ exakt $\mathcal{N}(0,1)$-verteilt. Dies heißt, dass die Normalverteilung ein Fixpunkt ist unter der Abbildung, die eine Wahrscheinlichkeitsverteilung $\mathbb{P}$ abbildet auf die Verteilung von $2^{-1/2}$ Mal die Summe zweier unabhängiger Zufallsgrößen mit Verteilung $\mathbb{P}$. In Termen der zugehörigen Dichten ist dies die Abbildung $\varphi \mapsto \varphi \star \varphi(\,\cdot\, 2^{-1/2})2^{-1/2}$. Man kann zumindest die Teilfolge $2^{-n/2}\sum_{i=1}^{2^n} X_i$ (wobei $X_1, X_2, \dots$ standardisierte unabhängige identisch verteilte Zufallsgrößen im $\mathcal{L}^2$ sind) auffassen als die Iterationsfolge unter der oben beschriebenen Abbildung. Da die Standardnormalverteilung ein Fixpunkt dieser Abbildung ist, ist es zumindest nicht unplausibel, dass ein 'Fixpunktsatz' wie der Zentrale Grenzwertsatz gelten könnte. ◇

Hier kommt also der angekündigte Spezialfall von Satz 5.2.2 für Bernoulli-Zufallsgrößen:

Satz 5.2.5 (Satz von de Moivre-Laplace)**.** *Es seien $X_1, \dots, X_n$ Bernoulli-Zufallsgrößen mit Parameter $p \in (0,1)$, d. h. $S_n = \sum_{i=1}^n X_i$ ist binomialverteilt mit Parametern n und p. Dann gilt für alle $a, b \in \mathbb{R}$ mit $a < b$*

$$\lim_{n\to\infty} \mathbb{P}\Big(a \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq b\Big) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2}}\,\mathrm{d}x.$$

Beweis. Wir schreiben q statt $1 - p$.

Wir schreiben zunächst die betrachtete Wahrscheinlichkeit als eine Summe von Einzelwahrscheinlichkeiten und skalieren diese Summe zu einem Integral, wobei wir die Randeffekte bei a und bei b pauschal mit $o(1)$ für $n \to \infty$ beschreiben:

$$\begin{aligned}
\mathbb{P}\Big(a \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq b\Big) &= \sum_{k \approx np + a\sqrt{npq}}^{\approx np + b\sqrt{npq}} \mathrm{Bi}_{n,p}(k) + o(1) \\
&= \int_{np + a\sqrt{npq}}^{np + b\sqrt{npq}} \mathrm{Bi}_{n,p}(\lfloor t \rfloor)\,\mathrm{d}t + o(1) \\
&= \int_a^b \sqrt{npq}\,\mathrm{Bi}_{n,p}\big(\lfloor np + x\sqrt{npq} \rfloor\big)\,\mathrm{d}x + o(1).
\end{aligned}$$

Der Rest des Beweises besteht darin zu zeigen, dass der Integrand gleichmäßig in $x \in [a,b]$ gegen $\varphi_{0,1}(x) = (2\pi)^{-1/2} e^{-x^2/2}$ konvergiert. Wir schreiben zunächst die Definition von $\mathrm{Bi}_{n,p}(k)$ aus und setzen Stirlings Formel asymptotisch für $n \to \infty$ ein:

$$n! \sim \sqrt{2\pi n}\Big(\frac{n}{e}\Big)^n, \qquad n \to \infty, \tag{5.2.2}$$

wobei wir wie immer $\sim$ für asymptotische Äquivalenz benutzen. Man beachte, dass diese und alle folgenden Approximationen gleichmäßig in $x \in [a,b]$ gelten.

$$\begin{aligned} \sqrt{npq}\,\mathrm{Bi}_{n,p}\big(\lfloor np + x\sqrt{npq}\rfloor\big) &\sim \sqrt{npq}\frac{\big(\frac{n}{e}\big)^n\sqrt{2\pi n}}{\big(\frac{np+x\sqrt{npq}}{ep}\big)^{np+x\sqrt{npq}}\sqrt{2\pi np}\big(\frac{nq-x\sqrt{npq}}{eq}\big)^{nq-x\sqrt{npq}}\sqrt{2\pi nq}} \\ &= \frac{1}{\sqrt{2\pi}}\Big(1+\frac{x}{\sqrt{n}}\sqrt{\frac{q}{p}}\Big)^{-np-x\sqrt{npq}}\Big(1-\frac{x}{\sqrt{n}}\sqrt{\frac{p}{q}}\Big)^{-nq+x\sqrt{npq}}. \end{aligned} \tag{5.2.3}$$

Nun schreiben wir die letzten beiden Terme mit Hilfe von $\exp\{\dots\}$ und benutzen die Asymptotik $(1+\frac{c_1}{\sqrt{n}})^{c_2\sqrt{n}} \to e^{c_1 c_2}$. Also ist die rechte Seite von (5.2.3) asymptotisch gleich

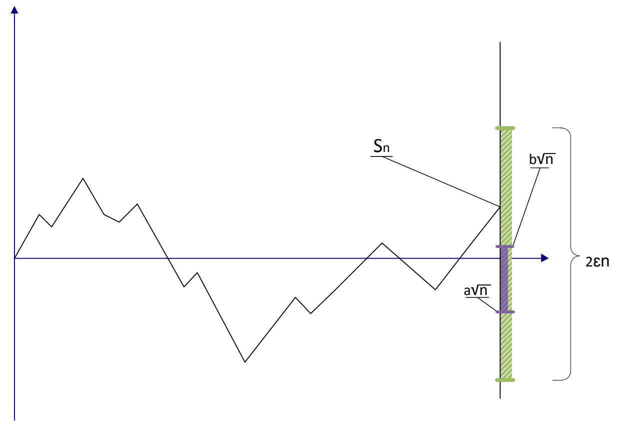
$$\frac{1}{\sqrt{2\pi}}\exp\Big\{-np\log\Big(1+\frac{x}{\sqrt{n}}\sqrt{\frac{q}{p}}\Big) - nq\log\Big(1-\frac{x}{\sqrt{n}}\sqrt{\frac{p}{q}}\Big)\Big\} e^{-x^2\sqrt{q/p}\sqrt{pq}} e^{-x^2\sqrt{p/q}\sqrt{pq}}. \tag{5.2.4}$$

Nun benutzen wir eine Taylor-Approximation für den Logarithmus: $\log(1+h) = h - h^2/2 + \mathcal{O}(h^3)$ für $h \to 0$. Also ist der Term in (5.2.4) asymptotisch äquivalent zu

$$\frac{1}{\sqrt{2\pi}}\exp\Big\{-np\frac{x}{\sqrt{n}}\sqrt{\frac{q}{p}} + np\frac{x^2}{2n}\frac{q}{p} + nq\frac{x}{\sqrt{n}}\sqrt{\frac{p}{q}} + nq\frac{x^2}{2n}\frac{p}{q} - x^2 q - x^2 p\Big\}. \tag{5.2.5}$$

Elementares Zusammenfassen zeigt, dass dies identisch ist mit $\varphi_{0,1}(x)$. $\square$

Beispiel 5.2.6 (Irrfahrt)**.** Sowohl beim Schwachen Gesetz der Großen Zahlen als auch beim Zentralen Genzwertsatz fragt man sich, mit welcher Wahrscheinlichkeit eine standardisierte Irrfahrt auf $\mathbb{Z}$ (zum Beispiel die einfache Irrfahrt, die mit gleicher Wahrscheinlichkeit jeweils zum nächsten Nachbarn springt) zum Zeitpunkt n eine gewisse "Zielscheibe" trifft. Diese ist beim GGZ für große n allerdings viel größer als beim ZGWS, nämlich $[-\varepsilon n, \varepsilon n]$ im Vergleich zu $[a\sqrt{n}, b\sqrt{n}]$. Man beachte, dass erstere um Null zentriert ist, aber die zweitere nicht unbedingt. Man sieht auch, dass εn durch jede Skala $\gg \sqrt{n}$ ersetzt werden kann, und das GGZ mit dieser Zielscheibe gilt immer noch.



$\diamond$

Beispiel 5.2.7 (Normalapproximation)**.** Eine typische Anwendung des Satzes von de Moivre-Laplace ist die Folgende. Wenn man einen fairen Würfel 1200 Mal wirft, wie groß ist dann die Wahrscheinlichkeit, dass dabei zwischen 190 und 200 Sechsen fallen? Wir nehmen natürlich an, dass die 1200 Würfe unabhängig voneinander geschehen.

Wir können die Zufallsgrößen X_i interpretieren als 1, falls der i-te Wurf eine Sechs hervor bringt und als 0 andernfalls. Dann ist die Anzahl der gewürfelten Sechsen gleich $S_{1200} = X_1 + \cdots + X_{1200}$, und wir sind in der Situation des Satzes von de Moivre-Laplace mit $p = \frac{1}{6}$ und $n = 1200$. Insbesondere sind $np = 200$ und $\sqrt{np(1-p)} = \sqrt{\frac{500}{3}} \approx 13$. Die gesuchte Wahrscheinlichkeit müssen wir wie folgt umschreiben:

$$
\begin{aligned}
\mathbb{P}(190 \leq S_{1200} \leq 200) &\approx \mathbb{P}\Big(\frac{190-200}{13} \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq 0\Big)\\
&= \mathbb{P}\Big(-\frac{10}{13} \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq 0\Big) \approx \Phi(0) - \Phi(-\tfrac{10}{13}).
\end{aligned}
$$

Es ist klar, dass $\Phi(0) = \frac{1}{2}$, und in einer Tabelle finden wir den Wert $\Phi(-\frac{10}{13}) = 1 - \Phi(\frac{10}{13}) \approx 1 - \Phi(0,77) \approx 1 - 0,7794 = 0,2306$. Also ist die gesuchte Wahrscheinlichkeit ungefähr gleich $0,2794$. ◇

Beispiel 5.2.8 (Wahlprognose)**.** Bei einer Wahl erhält Kandidat A einen unbekannten Anteil $p \in (0,1)$ der Stimmen. Um den Wert von p zu ermitteln, werten wir die ersten n Wahlzettel aus (bzw. befragen wir n zufällig gewählte Wähler). Wir groß sollte n sein, damit die Wahrscheinlichkeit eines Irrtums von mehr als einem Prozent nicht größer als $0,05$ ist?

Wenn wir n Zettel auswerten bzw. n Personen befragen, dann bekommen wir S_n Stimmen für den Kandidaten A, und S_n ist binomialverteilt mit Parametern n und p. Die Wahrscheinlichkeit des Ereignisses $\{|\frac{1}{n}S_n - p| > 0,01\}$ soll unter $0,05$ liegen, also

$$
\begin{aligned}
0,95 &\approx \mathbb{P}\Big(-\frac{0,01n}{\sqrt{np(1-p)}} \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq \frac{0,01n}{\sqrt{np(1-p)}}\Big)\\
&\approx \Phi\Big(0,01\sqrt{\frac{n}{p(1-p)}}\Big) - \Phi\Big(-0,01\sqrt{\frac{n}{p(1-p)}}\Big)\\
&= 2\Phi\Big(0,01\sqrt{\frac{n}{p(1-p)}}\Big) - 1.
\end{aligned}
$$

Also wollen wir ein (möglichst kleines) n bestimmen mit $\Phi\Big(0,01\sqrt{\frac{n}{p(1-p)}}\Big) \approx 0,975$. Einer Tabelle entnehmen wir, dass $\Phi(1,96) \approx 0,975$. Also sollte $n \approx p(1-p)10000 \cdot 1,96^2$ ausreichen. Wenn man keine Informationen über p besitzt, dann kann man (indem man den maximalen Wert $p = \frac{1}{2}$ einsetzt) davon ausgehen, dass $n \approx \frac{1}{4}10000 \cdot 1,96^2 \approx 9600$ ausreicht. ◇

Kapitel 6

Maßtheorie

In diesem Kapitel geben wir eine Einführung in die Maß- und Integrationstheorie unter besonderer Berücksichtigung der Wahrscheinlichkeitstheorie. Wir erläutern die fundamentalen Begriffe und Konzepte wie σ-Algebren, Dynkin-Systeme, Maße, Messbarkeit, Lebesgue-Maß, Verteilungsfunktionen, Integration, L^p-Räume und Produkträume, und stellen die wichtigsten Hilfsmittel bereit wie den Satz von Carathéodory, den monotonen und den beschränkten Konvergenzsatz und das Lemma von Fatou.

6.1 σ-Algebren und Maße

Es sei Ω eine nichtleere, eventuell sehr große Menge. Wir möchten auf Ω Wahrscheinlichkeiten einführen, um die Teilmengen von Ω zu messen. In anderen Worten, wir möchten eine Abbildung $\mathbb{P}$ auf der Menge der Teilmengen von Ω mit Werten in $[0,1]$ fest legen, so dass wir mit $\mathbb{P}(A)$ die Wahrscheinlichkeit einer Teilmenge A von Ω angeben können. Zu den Grundanforderungen an $\mathbb{P}$ gehören sicher die *Kolmogoroffschen Axiome*, die besagen, dass $\mathbb{P}(\Omega) = 1$ und $\mathbb{P}(\bigcup_{n\in\mathbb{N}} A_n) = \sum_{n\in\mathbb{N}} \mathbb{P}(A_n)$ für jede Folge von paarweise disjunkten Teilmengen $A_1, A_2, \dots$ gelten sollen. Insbesondere sollte man erwarten dürfen, dass $\mathbb{P}(A^{\mathrm{c}}) = 1 - \mathbb{P}(A)$ gilt.

Dies sind auch die Grundanforderungen, die man an ein Maß $\mathbb{P}$ stellt. Allerdings definiert man ein Maß fast nie auf der *gesamten* Menge aller Teilmengen von Ω (der sogenannten *Potenzmenge* $\mathcal{P}(\Omega)$ von Ω), sondern auf Teilmengensystemen, die kompatibel sind mit den oben erwähnten Rechengesetzen:

Definition 6.1.1 ($\boldsymbol{\sigma}$-Algebra)**.** *Ein Mengensystem $\mathcal{F}$ auf Ω (also eine Teilmenge $\mathcal{F}$ von $\mathcal{P}(\Omega)$) heißt eine* σ-Algebra*, falls gelten:*

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

(ii) falls $A \in \mathcal{F}$, so gilt auch $A^{\mathrm{c}} \in \mathcal{F}$,

(iii) falls $A_1, A_2, A_3, \dots \in \mathcal{F}$, so gilt auch $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{F}$.

Die Elemente von $\mathcal{F}$ heißen messbare Mengen *und das Paar $(\Omega, \mathcal{F})$ ein* messbarer Raum.

Also sind σ-Algebren geeignete Definitionsbereiche für Maße, siehe Definition 6.1.7 weiter unten. Die folgenden Bemerkungen sind offensichtlich oder Übungsaufgaben.

Bemerkung 6.1.2. (a) Die Potenzmenge und $\{\emptyset, \Omega\}$ sind zwei σ-Algebren, und zwar die beiden trivialen. Alle anderen σ-Algebren liegen zwischen diesen beiden.

(b) Jede σ-Algebra $\mathcal{F}$ ist nicht nur gegen Komplementbildung und abzählbare Vereinigung abgeschlossen, sondern auch gegen abzählbare Schnittbildung, d. h. auch $\bigcap_{n\in\mathbb{N}} A_n$ liegt in $\mathcal{F}$, wenn $A_1, A_2, A_3, \dots \in \mathcal{F}$, denn $\bigcap_{n\in\mathbb{N}} A_n = (\bigcup_{n\in\mathbb{N}} A_n^{\mathrm{c}})^{\mathrm{c}}$.

(c) Beliebige Schnitte von σ-Algebren sind ebenfalls σ-Algebren, aber im Allgemeinen sind Vereinigungen von σ-Algebren selber keine.

(d) Für jede Menge Ω ist das System der Mengen $A \subset \Omega$, so dass A oder A^{c} abzählbar ist, eine σ-Algebra.

(e) Wenn Ω und Ω' zwei Mengen sind und $f\colon \Omega \to \Omega'$ eine Abbildung, dann ist für jede σ-Algebra $\mathcal{F}'$ auf Ω' das System aller Urbilder $f^{-1}(A') = \{\omega \in \Omega\colon f(\omega) \in A'\}$ mit $A' \in \mathcal{F}'$ eine σ-Algebra auf Ω.

$\diamond$

Im Allgemeinen gibt es sehr viele σ-Algebren auf Ω, und viele σ-Algebren sind nicht nur selber gigantisch groß, sondern auch kaum explizit zu beschreiben. Daher ist es oft nützlich, Erzeugendensysteme gut zu kennen:

Lemma 6.1.3 (erzeugte $\boldsymbol{\sigma}$-Algebra)**.** *Zu jedem Mengensystem $\mathcal{C}$ auf Ω gibt es genau eine kleinste σ-Algebra $\sigma(\mathcal{C})$, die $\mathcal{C}$ enthält, d. h. eine σ-Algebra, die $\mathcal{C}$ enthält und die in jeder σ-Algebra enthalten ist, die $\mathcal{C}$ enthält. Wir nennen $\sigma(\mathcal{C})$ die* von $\mathcal{C}$ erzeugte σ-Algebra.

Beweis. Wir können $\sigma(\mathcal{C})$ angeben als[1]

$$\sigma(\mathcal{C}) = \bigcap\{\mathcal{A}\colon \mathcal{A} \text{ ist eine } \sigma\text{-Algebra mit } \mathcal{C} \subset \mathcal{A}\}.$$

Diese Definition ist sinnvoll (also ein nichtleerer Schnitt), denn mindestens $\mathcal{P}(\Omega)$ ist in dem Mengensystem enthalten, über das der Schnitt gebildet wird. Man verifiziert leicht, dass dieses Mengensystem die beiden Eigenschaften besitzt. $\square$

Es ist klar, dass immer $\mathcal{C} \subset \sigma(\mathcal{C})$ gilt und dass Gleichheit genau dann gilt, wenn $\mathcal{C}$ eine σ-Algebra ist. Ferner gilt $\sigma(\mathcal{C}_1) \subset \sigma(\mathcal{C}_2)$, falls $\mathcal{C}_1 \subset \mathcal{C}_2$.

Bemerkung 6.1.4. Wenn für jedes i aus einer beliebigen Indexmenge I das System $\mathcal{F}_i$ eine σ-Algebra ist, so ist ja (wie schon in Bemerkung 6.1.2(c) erwähnt) die Vereinigung $\bigcup_{i\in I} \mathcal{F}_i$ im Allgemeinen keine σ-Algebra. Die kleinste σ-Algebra, die diese Vereinigung enthält, wird mit

$$\bigvee_{i\in I} \mathcal{F}_i = \sigma\Big(\bigcup_{i\in I} \mathcal{F}_i\Big)$$

bezeichnet. Im Fall $I = \{1, \dots, n\}$ schreibt man auch $\mathcal{F}_1 \vee \dots \vee \mathcal{F}_n$. $\diamond$

Die wichtigste σ-Algebra auf $\mathbb{R}$ bzw. $\mathbb{R}^n$ ist die folgende.

[1] Wir schreiben $\bigcap\{\mathcal{A}_i : i \in I\}$ an Stelle von $\bigcap_{i\in I} \mathcal{A}_i = \{M \subset \Omega\colon M \in \mathcal{A}_i \,\forall i \in I\}$.

Beispiel 6.1.5 (Borel-$\boldsymbol{\sigma}$-Algebra)**.** (a) Auf $\mathbb{R}$ betrachten wir die von den offenen Mengen erzeugte σ-Algebra, also $\mathcal{B} = \sigma(\mathcal{O})$, wobei $\mathcal{O}$ die Menge aller offenen Teilmengen von $\mathbb{R}$ ist. Die σ-Algebra $\mathcal{B}$ heißt die *Borel-σ-Algebra* auf $\mathbb{R}$. Es gilt auch

$$\begin{aligned}\mathcal{B} &= \sigma(\{(a,b)\colon -\infty \le a < b \le \infty\} \\ &= \sigma(\{(-\infty,t]\colon t \in \mathbb{R}\}) \\ &= \sigma(\{(-\infty,t]\colon t \in \mathbb{Q}\}),\end{aligned}$$

d. h. $\mathcal{B}$ wird auch von dem System der offenen Intervalle, dem der links unbeschränkten und rechts beschränkten, abgeschlossenen Intervalle (mit rationalem Randpunkt) erzeugt (Übungsaufgaben). Es gibt kein direktes Kriterium, an Hand dessen man von einer gegebenen Teilmenge von $\mathbb{R}$ erkennen kann, ob sie zu $\mathcal{B}$ gehört oder nicht.[2]

(b) Auf $\mathbb{R}^n$ betrachten wir die von den offenen Mengen erzeugte σ-Algebra $\mathcal{B}_n$, die die *Borel-σ-Algebra* auf $\mathbb{R}^n$ genannt wird. Auch $\mathcal{B}_n$ wird von einer Reihe anderer Mengensysteme erzeugt, wie etwa vom System der offenen Kuben $\prod_{i=1}^n (a_i, b_i)$ mit $-\infty \le a_i < b_i \le \infty$ für $i = 1, \dots, n$, dem der abgeschlossenen Mengen oder auch von dem der kompakten Mengen (Übungsaufgabe).

◇

Bemerkung 6.1.6 (Spur-$\boldsymbol{\sigma}$-Algebra)**.** Wenn $\mathcal{F}$ eine σ-Algebra über Ω ist und $\widetilde{\Omega} \subset \Omega$ eine beliebige Teilmenge von Ω, so sieht man leicht, dass

$$\widetilde{\mathcal{F}} = \{A \cap \widetilde{\Omega}\colon A \in \mathcal{F}\}$$

eine σ-Algebra über $\widetilde{\Omega}$ ist. Wir nennen $\widetilde{\mathcal{F}}$ die *Spur-σ-Algebra* von $\mathcal{F}$ über $\widetilde{\Omega}$. Falls $\widetilde{\Omega}$ selber messbar ist, so besteht $\widetilde{\mathcal{F}}$ aus allen messbaren Teilmengen von $\widetilde{\Omega}$. Beispielsweise werden wir die Spur-σ-Algebra der Borel-σ-Algebra auf einer Teilmenge I von $\mathbb{R}^n$ mit $\mathcal{B}_I$ bezeichnen. ◇

Der zweite zentrale Begriff der Maßtheorie ist das Maß:

Definition 6.1.7 (Maß, Maßraum, Wahrscheinlichkeitsraum)**.** *Es sei $(\Omega, \mathcal{F})$ ein messbarer Raum. Es sei $\mu\colon \mathcal{F} \to [0, \infty]$ eine Abbildung.*

(a) μ heißt ein Maß*, falls gelten:*
  (i) $\mu(\emptyset) = 0$,
  *(ii) μ ist σ-*additiv*, d. h. für alle Folgen $(A_n)_{n\in\mathbb{N}}$ von paarweise disjunkten Mengen $A_n \in \mathcal{F}$ gilt $\mu(\bigcup_{n\in\mathbb{N}} A_n) = \sum_{n\in\mathbb{N}} \mu(A_n)$.*

  In diesem Fall heißt das Tripel $(\Omega, \mathcal{F}, \mu)$ ein Maßraum*.*

*(b) μ heißt σ-*endlich*, falls eine aufsteigende Folge $(\Omega_n)_{n\in\mathbb{N}}$ von messbaren Teilmengen von Ω existiert mit $\Omega = \bigcup_{n\in\mathbb{N}} \Omega_n$ und $\mu(\Omega_n) < \infty$ für jedes $n \in \mathbb{N}$.*

(c) μ heißt endlich*, falls $\mu(\Omega) < \infty$.*

(d) μ heißt ein Wahrscheinlichkeitsmaß*, falls μ ein Maß ist mit $\mu(\Omega) = 1$. In diesem Fall heißt das Tripel $(\Omega, \mathcal{F}, \mu)$ ein* Wahrscheinlichkeitsraum*, und wir schreiben meist $\mathbb{P}$ statt μ.*

[2]Dies ist etwa bei dem Begriff der Offenheit einer Menge anders: Falls zu jedem Punkt der Menge noch eine ganze Umgebung dazu gehört, hat die Menge diese Eigenschaft.

Für den Wahrscheinlichkeitstheoretiker wird also das Tripel $(\Omega, \mathcal{F}, \mathbb{P})$, der Wahrscheinlichkeitsraum, das Hauptobjekt der Untersuchung sein. Ein beliebiges endliches Maß $\mu \neq 0$ kann man leicht zu einem Wahrscheinlichkeitsmaß $\mathbb{P} = \mu/\mu(\Omega)$ normieren. Allerdings werden auch nichtendliche Maße von großer Bedeutung sein, wie zum Beispiel das natürlichste Maß auf der Borel-σ-Algebra:

Beispiel 6.1.8 (Lebesgue-Maß)**.** Auf dem messbaren Raum $(\mathbb{R}^n, \mathcal{B}_n)$ (siehe Beispiel 6.1.5) wollen wir ein Maß λ_n betrachten, das jedem Kubus $Q = \prod_{i=1}^n [a_i, b_i]$ seinen elementargeometrischen Inhalt $\lambda_n(Q) = \prod_{i=1}^n (b_i - a_i)$ zuordnet. Ein solches Maß nennen wir das *Lebesgue-Maß* auf $(\mathbb{R}^n, \mathcal{B}_n)$. Allerdings müssen wir die Frage nach der Existenz und Eindeutigkeit noch ein wenig zurück stellen, siehe Korollar 6.3.3 und Beispiel 6.10.7. ◇

Die Existenz und Eindeutigkeit von Maßen mit gewissen gewünschten Eigenschaften behandeln wir in Abschnitt 6.2. Wir sammeln zunächst ein paar fundamentale Eigenschaften von Maßen:

Lemma 6.1.9 (Eigenschaften von Maßen)**.** *Es sei μ ein Maß auf einer σ-Algebra $\mathcal{F}$. Dann hat μ die folgenden Eigenschaften.*

(a) μ ist stetig von unten und von oben, *d. h.*

(i) für jede Folge $(A_n)_{n\in\mathbb{N}}$ in $\mathcal{F}$ mit $A_n \uparrow A$[3] für ein $A \in \mathcal{F}$ gilt $\lim_{n\to\infty} \mu(A_n) = \mu(A)$,

(ii) für jede Folge $(A_n)_{n\in\mathbb{N}}$ in $\mathcal{F}$ mit $\mu(A_n) < \infty$ für ein $n \in \mathbb{N}$ und $A_n \downarrow A$ für ein $A \in \mathcal{F}$ gilt $\lim_{n\to\infty} \mu(A_n) = \mu(A)$.

(b) μ ist σ-subadditiv, *d. h. für jede Folge $(A_n)_{n\in\mathbb{N}}$ in $\mathcal{F}$ gilt $\mu(\bigcup_{n\in\mathbb{N}} A_n) \leq \sum_{n\in\mathbb{N}} \mu(A_n)$.*

(c) μ ist monoton *oder* isoton, *d. h. für je zwei Mengen $A, B \in \mathcal{F}$ mit $A \subset B$ gilt $\mu(A) \leq \mu(B)$.*

Beweis. (a) Wir behandeln nur den Fall $A_n \uparrow A$. Mit $A_0 = \emptyset$ definieren wir $B_n = A_n \setminus A_{n-1}$, dann sind die Mengen $B_1, B_2, B_3, \dots$ paarweise disjunkt, und ihre Vereinigung ist A. Also gilt

$$\mu(A) = \mu\Big(\bigcup_{n\in\mathbb{N}} B_n\Big) = \sum_{n\in\mathbb{N}} \mu(B_n) = \lim_{N\to\infty} \sum_{n=1}^N \mu(B_n) = \lim_{N\to\infty} \mu\Big(\bigcup_{n=1}^N B_n\Big) = \lim_{N\to\infty} \mu(A_N).$$

(b) und (c) Übungsaufgabe. □

Die diskrete Wahrscheinlichkeitstheorie ist ein Spezialfall des allgemeinen Konzepts von Definition 6.1.7:

Beispiel 6.1.10 (diskrete Maße, Dirac-Maß)**.** Es sei Ω beliebig und $\widetilde{\Omega} \subset \Omega$ eine höchstens abzählbare Teilmenge. Mit einer Abbildung $p\colon \widetilde{\Omega} \to (0, \infty]$ definieren wir

$$\mu(A) = \sum_{\omega \in A \cap \widetilde{\Omega}} p(\omega), \qquad A \subset \Omega.$$

Dann ist μ ein Maß auf der Potenzmenge von Ω, das auf der diskreten Teilmenge $\widetilde{\Omega}$ konzentriert[4] ist. Es gilt dann $\mu(\{\omega\}) = p(\omega) > 0$ für alle $\omega \in \widetilde{\Omega}$, und man nennt ω ein *Atom* von μ. Das

[3] Wir schreiben $A_n \uparrow A$, falls $A_n \subset A_{n+1}$ für jedes $n \in \mathbb{N}$ und $A = \bigcup_{n\in\mathbb{N}} A_n$, und wir schreiben $A_n \downarrow A$, falls $A_{n+1} \subset A_n$ für jedes $n \in \mathbb{N}$ und $A = \bigcap_{n\in\mathbb{N}} A_n$.

[4] Wir sagen, ein Maß ist *konzentriert* auf einer messbaren Menge $\widetilde{\Omega} \subset \Omega$, falls $\mu(\Omega \setminus \widetilde{\Omega}) = 0$ gilt.

Maß μ (oder auch seine Einschränkung auf die Potenzmenge von $\widetilde{\Omega}$) heißt ein *diskretes Maß*. Falls $p(\omega) < \infty$ für alle $\omega \in \widetilde{\Omega}$, so ist μ σ-endlich. Falls $\sum_{\omega \in \widetilde{\Omega}} p(\omega) = 1$, so ist μ ein diskretes Wahrscheinlichkeitsmaß. Falls $p(\omega) = 1$ für jedes $\omega \in \widetilde{\Omega}$, nennen wir μ das *Zählmaß auf* $\widetilde{\Omega}$; dann zählt $\mu(A) = \#(A \cap \widetilde{\Omega})$ die Anzahl der Punkte aus $\widetilde{\Omega}$ in A.

Falls $\widetilde{\Omega}$ einelementig ist, also $\widetilde{\Omega} = \{\omega_0\}$ für ein $\omega_0 \in \Omega$, so heißt μ das *Dirac-Maß* in ω_0, und wir schreiben $\mu = \delta_{\omega_0}$. Es gilt also

$$\delta_{\omega_0}(A) = \begin{cases} 1, & \text{falls } \omega_0 \in A, \\ 0 & \text{sonst.} \end{cases} \tag{6.1.1}$$

Ein beliebiges diskretes Maß μ wie oben kann man dann auch als eine Linearkombination von Dirac-Maßen auffassen, denn es gilt $\mu = \sum_{\omega \in \widetilde{\Omega}} p(\omega)\delta_\omega$.

Jede aus der diskreten Wahrscheinlichkeitstheorie bekannte Verteilung lässt sich auffassen als ein diskretes Maß auf $(\mathbb{R}, \mathcal{P}(\mathbb{R}))$. Die wichtigsten Beispiele sind die folgenden:

geometrische Verteilung	$\sum_{n \in \mathbb{N}_0} p(1-p)^n \delta_n,$	Parameter $p \in (0,1)$,
Poisson-Verteilung	$\sum_{n \in \mathbb{N}_0} \frac{\alpha^n}{n!} e^{-\alpha} \delta_n,$	Parameter $\alpha > 0$,
Binomial-Verteilung	$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \delta_k,$	Parameter $n \in \mathbb{N}, p \in (0,1)$.

◇

Warum konstruieren wir nicht alle Maße, die uns interessieren, gleich auf der Potenzmenge von Ω, sondern machen uns die Mühe, auf σ-Algebren einzuschränken, die ja ein wenig widerspenstig sind? Ein berühmtes Beispiel von Vitali (1905) gibt eine überzeugende Antwort:

Beispiel 6.1.11 (Potenzmenge ist zu groß)**.** Es gibt kein Maß auf der Potenzmenge von $\mathbb{R}$, das jedem Intervall gerade seine Länge zuordnet, d. h. man kann das Lebesguemaß aus Beispiel 6.1.8 nicht auf $\mathcal{P}(\mathbb{R})$ erweitern.

Nehmen wir an, dass $\lambda\colon \mathcal{P}(\mathbb{R}) \to [0,\infty]$ ein solches Maß sei, und betrachten wir die Äquivalenzrelation $\sim$ auf $\mathbb{R}$, die definiert wird durch $x \sim y \iff x - y \in \mathbb{Q}$. Unter Benutzung des Auswahlaxioms kann man eine Menge $A \subset [0,1]$ konstruieren, die mit jeder Äquivalenzklasse genau ein Element gemeinsam hat. Daher ist $\mathbb{R}$ gleich der abzählbaren disjunkten Vereinigung $\mathbb{R} = \bigcup_{q \in \mathbb{Q}} (q + A)$, wobei $q + A$ die um q verschobene Menge A ist. Wegen der abzählbaren Additivität ist also

$$\infty = \lambda(\mathbb{R}) = \sum_{q \in \mathbb{Q}} \lambda(q + A) = \sum_{q \in \mathbb{Q}} \lambda(A),$$

wobei wir die Translationsinvarianz von λ benutzt haben. Nun zeigen wir aber, dass $\lambda(A) = 0$ gilt, womit ein Widerspruch hergestellt ist. Dies geht so. Für die Menge

$$C = \bigcup_{q \in \mathbb{Q} \cap [0,1]} (q + A) \subset [0,2]$$

gilt

$$2 = \lambda([0,2]) \geq \lambda(C) = \sum_{q \in \mathbb{Q} \cap [0,1]} \lambda(q + A) = \sum_{q \in \mathbb{Q} \cap [0,1]} \lambda(A).$$

Daraus folgt $\lambda(A) = 0$, und der Widerspruch ist komplett. ◇

6.2 Konstruktion von Maßen

Wie wir schon in Beispiel 6.1.8 gesehen haben, ist es eine wichtige Aufgabe, die Existenz und Eindeutigkeit von Maßen zu sichern, die auf gewissen Teilsystemen bestimmte Eigenschaften aufweisen. Das Hauptmittel ist hierbei der *Satz von Carathéodory* (siehe Satz 6.2.7), der besagt, dass Mengenfunktionen, die auf geeigneten Teilsystemen gewisse Struktureigenschaften haben, auf die davon erzeugte σ-Algebra erweitert werden können. Betrachten wir zunächst geeignete Teilmengensysteme:

Definition 6.2.1 (durchschnittstabil, Dynkin-System). *Es sei $\mathcal{C} \subset \mathcal{P}(\Omega)$ ein nichtleeres Mengensystem.*

(a) $\mathcal{C}$ heißt durchschnittstabil, *falls für je zwei Mengen $A, B \in \mathcal{C}$ auch $A \cap B \in \mathcal{C}$ gilt.*

(b) $\mathcal{C}$ heißt ein Dynkin-System, *falls gelten:*
   (i) $\Omega \in \mathcal{C}$,
   (ii) falls $A \in \mathcal{C}$, so gilt auch $A^{\mathrm{c}} \in \mathcal{C}$,
   (iii) falls $A_1, A_2, A_3, \dots \in \mathcal{C}$ paarweise disjunkt sind, so gilt auch $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{C}$.

Ein Dynkin-System unterscheidet sich also nur darin von einer σ-Algebra, dass die Abgeschlossenheit gegen abzählbare Vereinigung nur für Folgen *paarweise disjunkter* Teilmengen gefordert wird. Insbesondere ist jede σ-Algebra ein Dynkin-System, aber nicht umgekehrt.[5] Dieser kleine Unterschied macht es in vielen Fällen viel einfacher, die Eigenschaft eines Dynkin-Systems zu verifizieren als die einer σ-Algebra. Der Zusammenhang ist allerdings sehr eng:

Lemma 6.2.2. *Jedes durchschnittstabile Dynkin-System ist eine σ-Algebra.*

Beweis. Es sei $\mathcal{C}$ ein durchschnittstabiles Dynkin-System. Seien $A_1, A_2, A_3, \dots \in \mathcal{C}$, dann müssen wir zeigen, dass auch $A = \bigcup_{n\in\mathbb{N}} A_n$ in $\mathcal{C}$ liegt.

Mit $A_0 = \emptyset$ betrachten wir die Mengen $B_n = A_n \setminus (A_1 \cup \dots \cup A_{n-1})$, die ja offensichtlich paarweise disjunkt sind. Offensichtlich ist die Vereinigung der B_n gleich der Vereinigung der A_n. Nun zeigen wir mit einer Vollständigen Induktion nach n, dass B_n und $A_1 \cup \dots \cup A_n$ zu $\mathcal{C}$ gehören, was also den Beweis beendet.

Für $n = 1$ ist nichts zu zeigen. Für $n \geq 2$ gilt

$$B_n = A_n \cap \big(A_1 \cup \dots \cup A_{n-1}\big)^{\mathrm{c}}.$$

Nach Induktionsvoraussetzung ist $A_1 \cup \dots \cup A_{n-1}$ in $\mathcal{C}$, also auch sein Komplement, also auch der Schnitt mit A_n, das heißt B_n liegt in $\mathcal{C}$. Ferner ist

$$A_1 \cup \dots \cup A_n = (A_1 \cup \dots \cup A_{n-1}) \cup B_n$$

eine disjunkte Vereinigung von zwei Mengen, die in $\mathcal{C}$ liegen, also liegt auch $A_1 \cup \dots \cup A_n$ in $\mathcal{C}$. Damit ist die Induktion beendet und damit auch der Beweis des Lemmas. □

Analog zu Lemma 6.1.3 erzeugt jedes Mengensystem $\mathcal{C}$ auf Ω auch ein Dynkin-System, das wir mit $d(\mathcal{C})$ bezeichnen wollen, also

$$d(\mathcal{C}) = \bigcap \{\mathcal{A} \subset \mathcal{P}(\Omega) \colon \mathcal{A} \text{ ist ein Dynkin-System mit } \mathcal{C} \subset \mathcal{A}\}. \tag{6.2.1}$$

Ein weiterer wichtiger Zusammenhang ist der folgende.

[5] Falls Ω eine (endliche) gerade Anzahl von Elementen > 2 enthält, ist das System der Teilmengen mit gerader Elementanzahl ein Dynkin-System, aber keine σ-Algebra.

Satz 6.2.3. *Für jedes durchschnittstabile Mengensystem $\mathcal{C}$ ist $d(\mathcal{C}) = \sigma(\mathcal{C})$.*

Beweis. Da jede σ-Algebra ein Dynkin-System ist, ist klar, dass $d(\mathcal{C}) \subset \sigma(\mathcal{C})$. Wenn wir zeigen, dass $d(\mathcal{C})$ eine σ-Algebra ist, folgt also Gleichheit. Nach Lemma 6.2.2 reicht es zu zeigen, dass $d(\mathcal{C})$ durchschnittstabil ist. Dazu betrachten wir das Mengensystem

$$\mathcal{A} = \{A \subset \Omega \colon A \cap C \in d(\mathcal{C}) \text{ für alle } C \in \mathcal{C}\}.$$

Da $\mathcal{C}$ durchschnittstabil ist, gilt $\mathcal{C} \subset \mathcal{A}$. Nun zeigen wir, dass $\mathcal{A}$ ein Dynkin-System ist, indem wir die drei definierenden Eigenschaften nachweisen. Die erste ist klar. Für jedes $A \in \mathcal{A}$ ist für jedes $C \in \mathcal{C}$ auch die Menge $A^c \cap C = (C^c \cup (A \cap C))^c$ in $d(\mathcal{C})$, wie man leicht sieht, also liegt auch A^c in $\mathcal{A}$. Und wenn $A_1, A_2, \dots \in \mathcal{A}$ paarweise disjunkt sind, so liegt für jedes $C \in \mathcal{C}$ auch die Menge $(\bigcup_{n\in\mathbb{N}} A_n) \cap C = \bigcup_{n\in\mathbb{N}}(A_n \cap C)$ in $d(\mathcal{C})$, also $\bigcup_{n\in\mathbb{N}} A_n$ in $\mathcal{A}$. Also ist $\mathcal{A}$ ein Dynkin-System. Insbesondere gilt $d(\mathcal{C}) \subset \mathcal{A}$.

Nun betrachten wir das System

$$\widetilde{\mathcal{A}} = \{A \subset \Omega \colon A \cap A' \in d(\mathcal{C}) \text{ für alle } A' \in d(\mathcal{C})\}.$$

Wegen $d(\mathcal{C}) \subset \mathcal{A}$ gilt $\mathcal{C} \subset \widetilde{\mathcal{A}}$. Wie bei $\mathcal{A}$ zeigt man, dass $\widetilde{\mathcal{A}}$ ebenfalls ein Dynkin-System ist. Also folgt auch $d(\mathcal{C}) \subset \widetilde{\mathcal{A}}$. Dies ist aber äquivalent dazu, dass $d(\mathcal{C})$ durchschnittstabil ist. □

Unser Ziel in diesem Abschnitt ist, Mengenfunktionen auf gewissen Mengensystemen zu Maßen auf σ-Algebren zu erweitern. Dazu brauchen wir zunächst geeignete Mengensysteme, die einfacher zu behandeln sind als σ-Algebren:

Definition 6.2.4 (Algebra)**.** *Ein nichtleeres Mengensystem $\mathcal{A} \subset \mathcal{P}(\Omega)$ heißt eine* Algebra*, falls für alle $A, B \in \mathcal{A}$ auch A^c und $A \cap B$ und $A \cup B$ in $\mathcal{A}$ liegen.*

Einer Algebra fehlt also nur die Abgeschlossenheit gegen *abzählbare* Vereinigungsbildung zu einer σ-Algebra.[6] Die Bedeutung des Begriffes der Algebra liegt darin, dass Algebren (im Gegensatz zu σ-Algebren) oft sehr explizit angegeben werden können. Zum Beispiel ist das System aller endlichen Vereinigungen von Intervallen eine Algebra über $\mathbb{R}$.

Analog zur Algebra gibt es auch eine nur *endlich* additive Version von Maßen:

Definition 6.2.5 (Inhalt, Prämaß)**.** *(a) Ein* Inhalt *μ auf einer Algebra $\mathcal{A}$ ist eine Abbildung $\mu \colon \mathcal{A} \to [0, \infty]$ mit den Eigenschaften $\mu(\emptyset) = 0$ und $\mu(A \cup B) = \mu(A) + \mu(B)$ für alle disjunkten Mengen $A, B \in \mathcal{A}$.*

(b) Ein σ-additiver Inhalt heißt ein Prämaß*. Genauer: Ein Inhalt μ auf einer Algebra $\mathcal{A}$ heißt ein* Prämaß*, wenn für jede Folge $(A_n)_{n\in\mathbb{N}}$ paarweise disjunkter Mengen $A_n \in \mathcal{A}$ mit $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{A}$ gilt: $\mu(\bigcup_{n\in\mathbb{N}} A_n) = \sum_{n\in\mathbb{N}} \mu(A_n)$.*

Einem Inhalt fehlt also nur die *abzählbare* Additivität zu einem Maß, vorausgesetzt, dass sein Definitionsbereich eine σ-Algebra ist. Wenn der Definitionsbereich eines Prämaßes nicht nur eine Algebra, sondern sogar eine σ-Algebra ist, so ist es schon ein Maß. Es ist klar, dass jeder Inhalt μ *monoton* ist (d. h. $\mu(A) \leq \mu(B)$ für alle $A, B \in \mathcal{A}$ mit $A \subset B$), *endlich additiv* (d. h. $\mu(\bigcup_{i=1}^n A_i) = \sum_{i=1}^n \mu(A_i)$ für paarweise disjunkte $A_1, \dots, A_n \in \mathcal{A}$) und *endlich subadditiv* (d. h. $\mu(\bigcup_{i=1}^n A_i) \leq \sum_{i=1}^n \mu(A_i)$ für beliebige $A_1, \dots, A_n \in \mathcal{A}$).

Die σ-Additivität ist sogar äquivalent zur Stetigkeit in $\emptyset$:

[6] Zum Beispiel ist das System der Teilmengen A von Ω, so dass A oder A^c endlich sind, eine Algebra, aber für unendliches Ω keine σ-Algebra.

Lemma 6.2.6. *Es sei μ ein endlicher Inhalt auf einer Algebra $\mathcal{A}$. Dann sind äquivalent:*

(i) μ ist σ-additiv (d. h. ein Prämaß),

(ii) μ ist stetig in der leeren Menge, d. h. für alle Folgen von Mengen $A_1, A_2, \dots \in \mathcal{A}$ mit $A_n \downarrow \emptyset$ gilt $\lim_{n\to\infty} \mu(A_n) = 0$.

Beweis. (i)$\Longrightarrow$(ii): Sei $A_n \downarrow \emptyset$. Wir definieren $B_n = A_n \setminus A_{n+1}$ für $n \in \mathbb{N}$. Dann sind die Mengen B_n paarweise disjunkt, und es gilt $A_n = \bigcup_{m=n}^\infty B_m$ für alle $n \in \mathbb{N}$. Also ist $\mu(A_n) = \sum_{m=n}^\infty \mu(B_m)$. Da diese Summe konvergiert, gilt $\lim_{n\to\infty} \mu(A_n) = 0$.

(ii)$\Longrightarrow$(i): Es sei $(B_n)_{n\in\mathbb{N}}$ eine Folge von paarweise disjunkten Mengen in $\mathcal{A}$, so dass $B = \bigcup_{n\in\mathbb{N}} B_n$ in $\mathcal{A}$ liegt. Für $n \in \mathbb{N}$ betrachte $A_{n+1} = \bigcup_{m=n+1}^\infty B_m = B \setminus (B_1 \cup \dots \cup B_n) \in \mathcal{A}$. Dann sind die Mengen $B_1, \dots, B_n, A_{n+1}$ paarweise disjunkt und in $\mathcal{A}$. Wegen der endlichen Additivität gilt

$$\mu(B) = \sum_{m=1}^n \mu(B_m) + \mu(A_{n+1}).$$

Wegen $A_{n+1} \downarrow \emptyset$ gilt $\lim_{n\to\infty} \mu(A_{n+1}) = 0$. Also folgt $\mu(B) = \sum_{m=1}^\infty \mu(B_m)$, d. h. μ ist σ-additiv. □

Nun kommen wir endlich zum Hauptergebnis dieses Abschnittes, das wir allerdings nicht in voller Stärke beweisen werden:

Satz 6.2.7 (Carathéodory)**.** *Es sei μ_0 ein σ-endliches Prämaß auf einer Algebra $\mathcal{A}$. Dann gibt es genau ein Maß μ auf $\sigma(\mathcal{A})$, das μ_0 erweitert, d. h. das auf $\mathcal{A}$ mit μ_0 übereinstimmt.*

Der Existenzteil dieses Satzes wird z. B. in [Ba91, I.5] bewiesen. Wir zeigen hier nur die Eindeutigkeit, die direkt aus dem folgenden Resultat folgt, das wir noch oft benutzen werden.

Satz 6.2.8 (Eindeutigkeitssatz)**.** *Es sei $\mathcal{F}$ eine σ-Algebra auf Ω und $\mathcal{C}$ ein durchschnittsstabiles Erzeugendensystem von $\mathcal{F}$. Falls zwei auf $\mathcal{F}$ definierte Maße μ, ν auf $\mathcal{C}$ mit einander übereinstimmen und eine Folge $(\Omega_n)_{n\in\mathbb{N}}$ in $\mathcal{C}$ existiert mit $\Omega_n \uparrow \Omega$ und $\mu(\Omega_n) = \nu(\Omega_n) < \infty$ für jedes $n \in \mathbb{N}$, so gilt $\mu = \nu$ auf $\mathcal{F}$.*

Beweis. Zunächst behandeln wir den Spezialfall, wo $\mu(\Omega) = \nu(\Omega) < \infty$. Wir betrachten das Mengensystem

$$\widetilde{\mathcal{F}} = \{A \in \mathcal{F} \colon \mu(A) = \nu(A)\}$$

und brauchen nur zu zeigen, dass $\mathcal{F} \subset \widetilde{\mathcal{F}}$. Hierzu reicht es zu zeigen, dass $\widetilde{\mathcal{F}}$ ein Dynkin-System ist, denn dann gilt nach Satz 6.2.3 schon $\mathcal{F} = \sigma(\mathcal{C}) = d(\mathcal{C}) \subset \widetilde{\mathcal{F}}$, und $\widetilde{\mathcal{F}}$ enthält offensichtlich $\mathcal{C}$. Dass $\widetilde{\mathcal{F}}$ ein Dynkin-System ist, rechnet man leicht an Hand der Definition nach: Offensichtlich ist $\Omega \in \widetilde{\mathcal{F}}$, und für jedes $D \in \widetilde{\mathcal{F}}$ gilt $\mu(D^c) = \mu(\Omega) - \mu(D) = \nu(\Omega) - \nu(D) = \nu(D^c)$, also liegt auch D^c in $\widetilde{\mathcal{F}}$. Und für jede Folge $(D_n)_{n\in\mathbb{N}}$ von paarweise disjunkten Mengen in $\widetilde{\mathcal{F}}$ gilt

$$\mu\Big(\bigcup_{n\in\mathbb{N}} D_n\Big) = \sum_{n\in\mathbb{N}} \mu(D_n) = \sum_{n\in\mathbb{N}} \nu(D_n) = \nu\Big(\bigcup_{n\in\mathbb{N}} D_n\Big),$$

also liegt $\bigcup_{n\in\mathbb{N}} D_n$ auch in $\widetilde{\mathcal{F}}$. Damit ist der Spezialfall $\mu(\Omega) = \nu(\Omega) < \infty$ bewiesen.

Im allgemeinen Fall betrachten wir für $n \in \mathbb{N}$ die endlichen Maße μ_n und ν_n, die definiert sind durch $\mu_n(A) = \mu(A \cap \Omega_n)$ bzw. $\nu_n(A) = \nu(A \cap \Omega_n)$ für alle $A \in \mathcal{F}$. Auf μ_n und ν_n können wir die oben bewiesene Aussage anwenden und erhalten, dass $\mu_n = \nu_n$ für jedes $n \in \mathbb{N}$. Durch Grenzübergang $n \to \infty$ erhält man $\mu(A) = \nu(A)$ für jedes $A \in \mathcal{F}$, also stimmen μ und ν überein. $\square$

Korollar 6.2.9. *Falls zwei endliche Maße μ, ν mit $\mu(\Omega) = \nu(\Omega)$ auf einem durchschnittstabilen Erzeugendensystem mit einander überein stimmen, so sind sie schon gleich.*

Ein oft benutzter Begriff im Zusammenhang mit Maßen ist der folgende.

Definition 6.2.10 (Nullmengen, Vollständigkeit)**.** *Es sei $(\Omega, \mathcal{F}, \mu)$ ein Maßraum. Eine Menge $A \subset \Omega$ (auch wenn sie nicht in $\mathcal{F}$ liegt) heißt eine* μ-Nullmenge*, falls eine messbare Menge $F \in \mathcal{F}$ existiert mit $\mu(F) = 0$ und $A \subset F$. Der Maßraum $(\Omega, \mathcal{F}, \mu)$ heißt* vollständig*, falls $\mathcal{F}$ alle μ-Nullmengen enthält.*

Als eine einfache Übungsaufgabe zeigt man leicht:

Lemma 6.2.11. *Abzählbare Vereinigungen von Nullmengen sind Nullmengen.*

Bemerkung 6.2.12 (Vervollständigung)**.** Jeder Maßraum $(\Omega, \mathcal{F}, \mu)$ lässt sich sehr leicht vervollständigen. Das Mengensystem[7]

$$\mathcal{G} = \{A \subset \Omega \colon \text{Es gibt ein } B \in \mathcal{F}, \text{ so dass } A \triangle B \text{ eine } \mu\text{-Nullmenge ist}\}$$

kann man leicht als eine σ-Algebra erkennen. Ferner lässt sich μ leicht auf $\mathcal{G}$ erweitern: Für $A \in \mathcal{G}$ existiert ja ein $B \in \mathcal{F}$, so dass $A \triangle B$ eine μ-Nullmenge ist. Nun setzt man $\overline{\mu}(A) = \mu(B)$. Als Übungsaufgabe verifiziert man leicht, dass $\overline{\mu}$ wohldefiniert ist (d. h. dass diese Definition nicht von der Wahl von B abhängt) und dass $(\Omega, \mathcal{G}, \overline{\mu})$ ein vollständiger Maßraum ist. Man nennt ihn die *Vervollständigung* von $(\Omega, \mathcal{F}, \mu)$. Es hat manchmal Vorteile, mit einer Vervollständigung zu arbeiten, aber man darf nicht vergessen, dass die σ-Algebra $\mathcal{G}$ vom Maß μ abhängt. $\diamond$

6.3 Verteilungsfunktionen

Wir erinnern daran (siehe Beispiel 6.1.5), dass die Borel-σ-Algebra $\mathcal{B}$ auf $\mathbb{R}$ insbesondere von den Mengen $(-\infty, t]$ mit $t \in \mathbb{R}$ erzeugt wird. Wir zeigen in diesem Abschnitt, dass Maße μ auf $(\mathbb{R}, \mathcal{B})$ eindeutig fest gelegt werden durch die Abbildung $t \mapsto \mu((-\infty, t])$. Im Fall eines Wahrscheinlichkeitsmaßes μ nennt man diese Abbildung die *Verteilungsfunktion* von μ. Im Folgenden finden wir hinreichende und notwendige Eigenschaften von Funktionen $F \colon \mathbb{R} \to \mathbb{R}$, so dass es ein Maß μ gibt mit $F(t) = \mu((-\infty, t])$ für alle $t \in \mathbb{R}$. Zunächst eine einfache Beobachtung.

Lemma 6.3.1. *Für jedes Maß μ auf $(\mathbb{R}, \mathcal{B})$ mit $\mu((s, t]) < \infty$ für alle $s < t$ ist die Abbildung $t \mapsto \mu((-\infty, t])$ monoton steigend und rechtsseitig stetig.*

Beweis. Die Monotonie ist klar, und die rechtsseitige Stetigkeit folgt aus der Darstellung $(-\infty, t] = \bigcap_{n \in \mathbb{N}} (-\infty, t + \frac{1}{n}]$ sowie der Stetigkeit des Maßes μ von oben, siehe Lemma 6.1.9(a). $\square$

Tatsächlich erweisen sich diese beiden Eigenschaften schon als entscheidend:

[7] Mit $A \triangle B = (A \setminus B) \cup (B \setminus A)$ bezeichnen wir die *symmetrische Differenz* von A und B.

Satz 6.3.2. *Zu jeder monoton steigenden, rechtsseitig stetigen Abbildung $F\colon \mathbb{R} \to \mathbb{R}$ gibt es genau ein Maß μ auf $(\mathbb{R}, \mathcal{B})$ mit $F(t) - F(s) = \mu((s,t])$ für alle $s, t \in \mathbb{R}$ mit $s < t$.*

Beweis. Es sei $\mathcal{J}$ die Menge aller Intervalle $(s,t]$ mit $-\infty \le s \le t < \infty$ und aller Intervalle (s,∞) mit $-\infty \le s \le \infty$. Ferner sei $a(\mathcal{J})$ die von $\mathcal{J}$ erzeugte Algebra, das heißt die kleinste Algebra, die $\mathcal{J}$ enthält. Man kann leicht zeigen, dass $a(\mathcal{J})$ aus allen endlichen disjunkten Vereinigungen der Intervalle in $\mathcal{J}$ besteht. Wir definieren

$$\begin{aligned} \mu((s,t]) &= F(t) - F(s), & -\infty \le s \le t < \infty, \\ \mu((s,\infty)) &= F(\infty) - F(s), & -\infty \le s < \infty, \end{aligned}$$

wobei wir $F(\infty) = \lim_{t\to\infty} F(t) \in (-\infty, \infty]$ und $F(-\infty) = \lim_{s\to-\infty} F(s) \in [-\infty, \infty)$ setzen. Natürlich können wir μ sofort leicht auf $a(\mathcal{J})$ erweitern: Eine disjunkte Vereinigung von Intervallen dieser Form erhält als μ-Wert natürlich die Summe der μ-Werte der Intervalle. Es ist evident, dass μ ein Inhalt auf $a(\mathcal{J})$ ist. Offensichtlich ist μ auch σ-endlich, denn $\mathbb{R} = \bigcup_{n\in\mathbb{N}}(-n,n]$, und $\mu((-n,n]) = F(n) - F(-n) < \infty$.

Nach dem Satz von Carathéodory genügt es zu zeigen, dass μ ein Prämaß auf $a(\mathcal{J})$ ist. Diese Eigenschaft zeigen wir nur für den Fall $-\infty < F(-\infty) < F(\infty) < \infty$, der allgemeine Fall ist eine Übungsaufgabe. Nach Lemma 6.2.6 müssen wir zeigen, dass $\mu(A_n) \to 0$ für jede Folge $(A_n)_{n\in\mathbb{N}}$ in $a(\mathcal{J})$ mit $A_n \downarrow \emptyset$ gilt. Dies beweisen wir indirekt: Sei $(A_n)_{n\in\mathbb{N}}$ eine fallende Folge in $a(\mathcal{J})$ mit $a = \lim_{n\to\infty} \mu(A_n) > 0$, dann werden wir zeigen, dass $\bigcap_{n\in\mathbb{N}} A_n \ne \emptyset$ gilt.

(i) Zunächst zeigen wir, dass für jedes Intervall $I \in \mathcal{J}$ und jedes $\varepsilon > 0$ ein Intervall $I' \in \mathcal{J}$ und ein kompaktes Intervall L existieren mit

$$I' \subset L \subset I \qquad \text{und} \qquad \mu(I') \ge \mu(I) - \varepsilon. \tag{6.3.1}$$

Im Fall $I = (s,t]$ mit $s, t \in \mathbb{R}$ wählen wir $I' = (s', t]$ und $L = [(s+s')/2, t]$, wobei $s' \in (s,t)$ so gewählt wurde, dass $F(s') \le F(s) + \varepsilon$ gilt. Dann ist offensichtlich (6.3.1) erfüllt. Im Fall $I = (-\infty, t]$ wählen wir ein $s' \in (-\infty, t]$ mit $F(s') \le F(-\infty) + \varepsilon$ und setzen $I' = (s', t]$ und $L = [s'-1, t]$. Die Fälle $I = (s, \infty)$ und $I = \mathbb{R}$ gehen analog.

(ii) Jedes $A_n \in a(\mathcal{J})$ ist eine endliche Vereinigung von Intervallen in $\mathcal{J}$. Aus (i) folgt daher, dass wir $B_n \in a(\mathcal{J})$ und eine kompakte Menge $K_n \subset \mathbb{R}$ finden können mit $B_n \subset K_n \subset A_n$ und $\mu(B_n) \ge \mu(A_n) - a2^{-n-1}$.

(iii) Es gilt für jedes $n \in \mathbb{N}$

$$A_n \setminus (B_1 \cap \cdots \cap B_n) = (A_1 \cap \cdots \cap A_n) \setminus (B_1 \cap \cdots \cap B_n) \subset \bigcup_{i=1}^n (A_i \setminus B_i)$$

und daher auch

$$\begin{aligned} \mu(B_1 \cap \cdots \cap B_n) &\ge \mu(A_n) - \mu\Big(\bigcup_{i=1}^n (A_i \setminus B_i)\Big) \\ &\ge \mu(A_n) - \sum_{i=1}^n \mu(A_i \setminus B_i) \ge a - \sum_{i=1}^n a2^{-i-1} \ge a/2. \end{aligned}$$

Daher ist $B_1 \cap \cdots \cap B_n \neq \emptyset$ für jedes $n \in \mathbb{N}$. Wegen $B_n \subset K_n$ ist auch die Menge $\overline{K}_n = K_1 \cap \cdots \cap K_n$ nicht leer für jedes $n \in \mathbb{N}$. Da $(\overline{K}_n)_{n \in \mathbb{N}}$ eine absteigende Folge nichtleerer kompakter Mengen ist, ist ihr Schnitt $\bigcap_{n \in \mathbb{N}} \overline{K}_n$ nicht leer. Wegen $K_n \subset A_n$ ist auch $\bigcap_{n \in \mathbb{N}} A_n$ nicht leer, und das war zu zeigen. □

Nachzutragen ist noch, dass μ im Falle $F(\infty) - F(-\infty) = 1$ ein *Wahrscheinlichkeits*maß ist. In diesem Fall kann man ohne Einschränkung voraus setzen, dass $F(-\infty) = 0$ gilt, und man nennt F die *Verteilungsfunktion* von μ. Wir haben dann $\mu((-\infty, t]) = F(t)$ für jedes $t \in \mathbb{R}$. Ferner haben wir endlich einen Existenz- und Eindeutigkeitsbeweis für das Lebesgue-Maß, siehe Beispiel 6.1.8:

Korollar 6.3.3 (Lebesgue-Maß). *Es gibt genau ein Maß λ auf $(\mathbb{R}, \mathcal{B})$ mit $\lambda((s, t]) = t - s$ für alle $s < t$. Dieses Maß λ heißt das* Lebesgue-Maß *auf* $\mathbb{R}$.

Beweis. Wir wenden Satz 6.3.2 auf $F(t) = t$ an. □

Korollar 6.3.4 (stetige Dichten). *Es sei $f\colon \mathbb{R} \to [0, \infty)$ eine stetige nichtnegative Funktion, so dass das Riemann-Integral $\int_a^b f(x)\,\mathrm{d}x$ endlich ist für alle $a, b \in \mathbb{R}$ mit $a < b$. Dann existiert genau ein Maß μ_f auf $(\mathbb{R}, \mathcal{B})$ mit $\mu_f((a, b]) = \int_a^b f(x)\,\mathrm{d}x$ für alle $a < b$. Wir sagen, μ_f habe die* Dichte *f.*

Beweis. Wir wenden Satz 6.3.2 auf $F(t) = \int_{-\infty}^t f(x)\,\mathrm{d}x$ an. □

Das Beispiel aus Korollar 6.3.4 kann man stark verallgemeinern; siehe Abschnitt 6.6.

6.4 Messbare Abbildungen

Bei Abbildungen zwischen messbaren Räumen muss die Kompatibilität mit den jeweiligen σ-Algebren beachtet werden.

Definition 6.4.1 (messbare Abbildung). *Es seien $(\Omega, \mathcal{F})$ und $(\Omega', \mathcal{F}')$ zwei messbare Räume. Eine Abbildung $f\colon \Omega \to \Omega'$ heißt* $\mathcal{F}$-$\mathcal{F}'$-messbar *oder auch nur kurz* messbar*, falls für jedes $A \in \mathcal{F}'$ das Urbild $f^{-1}(A) = \{\omega \in \Omega\colon f(\omega) \in A\}$ in $\mathcal{F}$ liegt.*

Eine Abbildung $f\colon \Omega \to \Omega'$ heißt also messbar, falls die Urbild-σ-Algebra

$$f^{-1}(\mathcal{F}') = \{f^{-1}(A)\colon A \in \mathcal{F}'\}$$

eine Teilmenge von $\mathcal{F}$ ist. Im Fall $(\Omega', \mathcal{F}') = (\mathbb{R}, \mathcal{B})$ sprechen wir auch einfach von $\mathcal{F}$-Messbarkeit oder von Borel-Messbarkeit.

Bemerkung 6.4.2. Wenn über Ω a priori keine σ-Algebra existiert, so kann man mit Hilfe eines messbaren Raums $(\Omega', \mathcal{F}')$ und einer Abbildung $f\colon \Omega \to \Omega'$ über Ω eine σ-Algebra erklären, indem man $\mathcal{F} = \sigma(f) = f^{-1}(\mathcal{F}')$ setzt. Diese σ-Algebra ist die kleinste σ-Algebra über Ω, so dass f bezüglich ihr messbar ist. Man nennt $\sigma(f)$ auch die *durch f auf Ω erzeugte* σ-Algebra. ◇

Für das Urbild $f^{-1}(A)$ werden wir manchmal auch abkürzend $\{f \in A\}$ schreiben. Darüber hinaus werden wir auch folgende Konventionen benutzen:

$$\begin{aligned}\{f < t\} &= \{\omega \in \Omega\colon f(\omega) < t\},\\ \{f = g\} &= \{\omega \in \Omega\colon f(\omega) = g(\omega)\},\\ \{f = 0\} &= \{\omega \in \Omega\colon f(\omega) = 0\}\end{aligned}$$

und so weiter.

Es ist im Folgenden bequem, den Wertebereich $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\} \cup \{-\infty\}$ zuzulassen. Eine Funktion $f\colon \Omega \to \overline{\mathbb{R}}$ belegen wir auch mit der unsinnigen, aber üblichen Bezeichnung *numerisch*. Auf $\overline{\mathbb{R}}$ betrachten wir die σ-Algebra $\overline{\mathcal{B}}$, die von $\mathcal{B}$, $\{\infty\}$ und $\{-\infty\}$ erzeugt wird. Eine $\mathcal{F}$-$\overline{\mathcal{B}}$-messbare Funktion ist dann also eine messbare numerische Funktion.

Es ist schlicht unmöglich, alle Urbildmengen $f^{-1}(A)$ auf Zugehörigkeit zu $\mathcal{F}$ zu prüfen. Glücklicherweise kann man sich stark einschränken:

Lemma 6.4.3. *Eine Abbildung $f\colon \Omega \to \Omega'$ ist genau dann $\mathcal{F}$-$\mathcal{F}'$-messbar, wenn es ein Erzeugendensystem $\mathcal{C}$ von $\mathcal{F}'$ gibt mit $f^{-1}(\mathcal{C}) \subset \mathcal{F}$.*

Beweis. Man weist leicht als Übungsaufgabe nach, dass das System

$$\mathcal{A}_f = \{A \subset \Omega'\colon f^{-1}(A) \in \mathcal{F}\}$$

eine σ-Algebra ist. Da nach Voraussetzung $\mathcal{C} \subset \mathcal{A}_f$ gilt, gilt auch $\mathcal{F}' = \sigma(\mathcal{C}) \subset \mathcal{A}_f$. Dies aber bedeutet, dass $f^{-1}(\mathcal{F}') \subset \mathcal{F}$ gilt. □

Insbesondere ist eine Abbildung $f\colon \Omega \to \mathbb{R}$ schon Borel-messbar, wenn für alle $t \in \mathbb{R}$ gilt: $\{\omega \in \Omega\colon f(\omega) \le t\} = f^{-1}((-\infty, t]) \in \mathcal{F}$, denn die Intervalle $(-\infty, t]$ bilden ein Erzeugendensystem von $\mathcal{B}$.

Messbarkeit wird unter nahezu allen nur denkbaren sinnvollen Operationen erhalten. Wir sammeln diese Ergebnisse in den folgenden Lemmata.

Lemma 6.4.4. *Hintereinanderausführungen von messbaren Funktionen sind messbar. Ausführlich: Falls $(\Omega_i, \mathcal{F}_i)$ mit $i = 1, 2, 3$ drei messbare Räume sind und $f_1\colon \Omega_1 \to \Omega_2$ sowie $f_2\colon \Omega_2 \to \Omega_3$ messbar sind, so ist auch $f_2 \circ f_1\colon \Omega_1 \to \Omega_3$ messbar.*

Beweis. Dies folgt leicht aus $(f_2 \circ f_1)^{-1}(A) = f_1^{-1}(f_2^{-1}(A)) \in \mathcal{F}_1$ für $A \in \mathcal{F}_3$. □

Lemma 6.4.5. *Es sei $(\Omega, \mathcal{F})$ ein messbarer Raum und $(f_n)_{n\in\mathbb{N}}$ eine Folge messbarer numerischer Funktionen auf Ω. Dann sind $\underline{f} = \liminf_{n\to\infty} f_n$ und $\overline{f} = \limsup_{n\to\infty} f_n$ ebenfalls messbare numerische Funktionen. Falls alle f_n reellwertig sind und $\underline{f}$ ebenfalls, so ist $\underline{f}$ Borel-messbar, analog für $\overline{f}$. Insbesondere sind (punktweise) Grenzwerte von messbaren Funktionen messbar.*

Beweis. Wir zeigen die Aussage nur für $\underline{f}$. Es genügt zu zeigen, dass $\{\omega \in \Omega\colon \underline{f}(\omega) \le t\} \in \mathcal{F}$ für jedes $t \in \mathbb{R}$. Es ist $\underline{f} = \lim_{n\to\infty} \inf_{m\ge n} f_m$, wobei dieser Grenzwert monoton steigend ist. Also können wir schreiben:

$$\begin{aligned}\{\omega\colon \underline{f}(\omega) \le t\} &= \bigcap_{n\in\mathbb{N}} \Big\{\omega\colon \inf_{m\ge n} f_m(\omega) \le t\Big\} = \bigcap_{n\in\mathbb{N}} \bigcap_{k\in\mathbb{N}} \Big\{\omega\colon \inf_{m\ge n} f_m(\omega) < t + \frac{1}{k}\Big\}\\ &= \bigcap_{n\in\mathbb{N}} \bigcap_{k\in\mathbb{N}} \bigcup_{m\in\mathbb{N}\colon m\ge n} \Big\{\omega\colon f_m(\omega) < t + \frac{1}{k}\Big\} \in \mathcal{F}. \qquad \square\end{aligned}$$

Messbarkeit erhält sich auch unter allen algebraischen Verknüpfungen:

Lemma 6.4.6. *(a) Jede konstante Funktion ist messbar.*

(b) Sind f und g messbare Funktionen, so sind auch $f+g$ und fg messbar.

(c) Sind f und g messbare Funktionen und $g(\omega) \neq 0$ für jedes ω, so ist f/g messbar.

(d) Falls A eine messbare Menge ist, so ist die Indikatorfunktion $\mathbb{1}_A$ messbar.

Hierbei ist die *Indikatorfunktion* $\mathbb{1}_A \colon \Omega \to \mathbb{R}$ auf einer Menge $A \subset \Omega$ definiert durch

$$\mathbb{1}_A(\omega) = \begin{cases} 1, & \text{falls } \omega \in A, \\ 0 & \text{sonst.} \end{cases}$$

Beweis von Lemma 6.4.6. Die Aussagen (a) und (d) sind klar. Offensichtlich ist die Menge

$$\{f < g\} = \bigcup_{q \in \mathbb{Q}} \{f < q\} \cap \{q < g\}$$

messbar, wenn f und g messbar sind, und leicht sieht man auch die Messbarkeit der Mengen $\{f \leq g\}$, $\{f = g\}$ und $\{f \neq g\}$ ein. Auch sieht man leicht die Messbarkeit von $\alpha + \tau g$ für alle $\alpha, \tau \in \mathbb{R}$. Die Messbarkeit von $f+g$ folgt dann aus der Messbarkeit der Menge $\{f + g \leq t\} = \{f \leq t - g\}$ für jedes $t \in \mathbb{R}$. Dass fg messbar ist, sieht man an der Darstellung $fg = \frac{1}{4}(f+g)^2 - \frac{1}{4}(f-g)^2$, denn auch die Messbarkeit von Quadraten messbarer Funktionen ist leicht zu sehen. Der Beweis von (c) ist eine Übungsaufgabe. □

Bemerkung 6.4.7 (Borel-$\boldsymbol{\sigma}$-Algebren)**.** Die Borel-σ-Algebra kann natürlicherweise auf jedem *topologischen* Raum Ω definiert werden als die von dem System der offenen Mengen erzeugte σ-Algebra. Dies erzwingt auf natürliche Weise eine Kompatibilität mit der Stetigkeit: Jede stetige Abbildung zwischen messbaren Räumen, die mit der entsprechenden Borel-σ-Algebra versehen sind, ist messbar. Dies folgt direkt aus der Tatsache, dass die Urbilder offener Mengen unter stetigen Abbildungen offen sind, in Kombination mit Lemma 6.4.3. ◇

Lemma 6.4.8. *Ist f eine messbare Funktion, so sind $f^+ = \max\{f, 0\}$, $f^- = \max\{-f, 0\}$ und $|f|$ messbar.*

Beweis. Man kombiniere Lemma 6.4.4, Lemma 6.4.6 und Bemerkung 6.4.7. □

Definition 6.4.9 (einfache Funktion)**.** *Es sei $(\Omega, \mathcal{F})$ ein messbarer Raum. Eine* einfache Funktion *ist eine endliche Linearkombination von Indikatorfunktionen auf messbaren Mengen, d. h. eine Funktion der Form $\sum_{i=1}^n a_i \mathbb{1}_{A_i}$ mit $n \in \mathbb{N}$, $a_1, \dots, a_n \in \mathbb{R}$ und $A_1, \dots, A_n \in \mathcal{F}$.*

Die Menge der einfachen Funktionen ist offensichtlich abgeschlossen gegen (punktweise) Addition, Multiplikation, Maximum- und Minimumbildung. Mit ihnen kann man alle nichtnegativen messbaren numerischen Funktionen approximieren durch monotone Grenzwertbildung:

Lemma 6.4.10. *Für jede nichtnegative messbare numerische Funktion f gibt es eine monoton wachsende Folge von einfachen Funktionen $f_1, f_2, f_3, \dots$ mit $f_n \uparrow f$ für $n \to \infty$.*

Beweis. Wähle

$$f_n = \sum_{k=1}^{n2^n} (k-1)2^{-n} \mathbb{1}_{\{(k-1)2^{-n} \leq f < k2^{-n}\}} + n\mathbb{1}_{\{f \geq n\}}.$$

□

Mit diesem Ergebnis kann man die Menge aller messbaren numerischen Funktionen wie folgt charakterisieren:

Satz 6.4.11. *Es sei $(\Omega, \mathcal{F})$ ein messbarer Raum und Γ eine Menge von nichtnegativen messbaren numerischen Funktionen mit den folgenden Eigenschaften:*

(i) Falls $f, g \in \Gamma$ und $a, b \in (0, \infty)$, so gilt $af + bg \in \Gamma$.

(ii) Falls $f_1, f_2, f_3, \dots \in \Gamma$ mit $f_n \uparrow f$ für $n \to \infty$, so gilt $f \in \Gamma$.

(iii) Γ enthält alle Indikatorfunktionen $\mathbb{1}_A$ mit $A \in \mathcal{F}$.

Dann ist Γ die Menge aller nichtnegativen messbaren numerischen Funktionen.

Beweis. Aus (i) und (iii) folgt, dass Γ alle nichtnegativen einfachen Funktionen enthält. Mit (ii) und Lemma 6.4.10 folgt die Aussage. □

Satz 6.4.11 wird routinemäßig für diverse Beweise über Eigenschaften messbarer Funktionen oder verwandter Objekte eingesetzt werden. Wenn gezeigt werden soll, dass alle nichtnegativen messbaren numerischen Funktionen eine gewisse Eigenschaft besitzen, so reicht es also zu zeigen, dass alle nichtnegativen einfachen Funktionen diese Eigenschaft besitzen und dass diese Eigenschaft sich erhält unter Bildung von Linearkombinationen und monotonen Grenzwerten.

6.5 Integration

In diesem Abschnitt sei ein σ-endlicher Maßraum $(\Omega, \mathcal{F}, \mu)$ vorgegeben. Wir wollen das Integral $\int_\Omega f \, \mathrm{d}\mu$ von messbaren Funktionen $f \colon \Omega \to \mathbb{R}$ bezüglich μ einführen. Dies machen wir in drei Schritten:

(i) für einfache Funktionen durch eine offensichtliche Formel,

(ii) für nichtnegative Funktionen durch monotonen Grenzübergang,

(iii) für beliebige Funktionen durch Zerlegung in Positiv- und Negativteil.

Dieses Vorgehen ist fundamental in der Integrationstheorie, und es wird sich später an vielen Stellen zeigen, dass es auch ein effektives Beweisgerüst liefert für eine Vielzahl von Aussagen über Integrale messbarer Funktionen (siehe auch die Bemerkung nach Satz 6.4.11).

Definition 6.5.1 (Definition des Integrals)**.**

(i) Das Integral einer nichtnegativen einfachen Funktion $f = \sum_{i=1}^n a_i \mathbb{1}_{A_i}$ *mit* $A_1, \dots, A_n \in \mathcal{F}$ *und* $a_1, \dots, a_n > 0$ *wird definiert als*

$$\int f \,\mathrm{d}\mu = \sum_{i=1}^n a_i \mu(A_i).$$

(ii) Das Integral einer nichtnegativen messbaren numerischen Funktion $f \colon \Omega \to [0, \infty]$ *wird definiert als*

$$\int f \,\mathrm{d}\mu = \lim_{n\to\infty} \int f_n \,\mathrm{d}\mu \in [0, \infty],$$

wobei $(f_n)_{n\in\mathbb{N}}$ *eine Folge nichtnegativer einfacher Funktionen sei mit* $f_n \uparrow f$*.*

(iii) Sei $f \colon \Omega \to \overline{\mathbb{R}}$ *messbar und numerisch, so dass* $\int f^+ \,\mathrm{d}\mu < \infty$ *und* $\int f^- \,\mathrm{d}\mu < \infty$*. In diesem Fall nennen wir* f integrierbar bezüglich μ*, und das Integral von* f *wird definiert als*

$$\int f \,\mathrm{d}\mu = \int f^+ \,\mathrm{d}\mu - \int f^- \,\mathrm{d}\mu$$

Wir schreiben $\mathcal{L}^1(\Omega, \mathcal{F}, \mu)$ *(oder auch kurz* $\mathcal{L}^1(\mu)$ *oder einfach nur* $\mathcal{L}^1$*) für die Menge aller bezüglich* μ *integrierbaren Funktionen.*

Es sind etliche Bemerkungen notwendig:

Bemerkung 6.5.2. (a) Um zu zeigen, dass Definition 6.5.1(i) sinnvoll ist, muss man zeigen, dass der Ausdruck $\sum_{i=1}^n a_i\mu(A_i)$ nicht von der gewählten Darstellung $f = \sum_{i=1}^n a_i \mathbb{1}_{A_i}$ abhängt. Dies kann man leicht tun, nachdem man sich überlegt hat, dass man ohne Einschränkung davon ausgehen kann, dass die Mengen $A_1, \dots, A_n$ paarweise disjunkt sind; siehe auch [Ba91, Lemma 10.2].

(b) Eine ähnliche, aber etwas schwierigere Aufgabe hat man bei Definition 6.5.1(ii): Man muss zeigen, dass der Ausdruck $\lim_{n\to\infty} \int f_n \,\mathrm{d}\mu$ nicht von der gewählten Folge $(f_n)_{n\in\mathbb{N}}$ abhängt. (Die Existenz einer solchen Folge wird ja durch Lemma 6.4.10 garantiert.) Man zeige dies als Übungsaufgabe oder konsultiere [Ba91, Korollar 11.2]. Wir weisen darauf hin, dass dieses hier von uns unter den Tisch gekehrte technische Detail die eigentliche Schwierigkeit ist beim Beweis der folgenden Resultate in Lemma 6.5.4 und den Sätzen 6.8.1 und 6.8.6. Man könnte auch das Integral einer nichtnegativen messbaren Funktion definieren als *Supremum* aller Integrale messbarer nichtnegativer einfacher Funktionen, die nicht größer sind als die betrachtete Funktion, und verschiebt die oben genannte Schwierigkeit.

(c) Statt $\int f \,\mathrm{d}\mu$ schreibt man auch oft ausführlicher $\int_\Omega f(\omega)\,\mu(\mathrm{d}\omega)$ oder $\int_\Omega f(\omega)\,\mathrm{d}\mu(\omega)$.

(d) Für nichtnegative numerische messbare Funktionen f ist $\int f \,\mathrm{d}\mu$ also immer definiert, aber es kann gleich ∞ sein. Eine beliebige numerische messbare Funktion f nennt man manchmal auch noch dann integrierbar (allerdings nur in unbestimmtem Sinn), wenn $\int f^+ \,\mathrm{d}\mu = \infty$ und $\int f^- \,\mathrm{d}\mu < \infty$ (dann mit Wert $\int f \,\mathrm{d}\mu = \infty$) oder wenn $\int f^+ \,\mathrm{d}\mu < \infty$ und $\int f^- \,\mathrm{d}\mu = \infty$ (dann mit Wert $\int f \,\mathrm{d}\mu = -\infty$).

(e) Falls $f \geq 0$ und $\mu(\{\omega\colon f(\omega) = \infty\}) > 0$, so gilt $\int f \,\mathrm{d}\mu = \infty$, wie man leicht als eine Übungsaufgabe zeigt.

(f) Für eine messbare Menge $A \in \mathcal{F}$ setzt man $\int_A f \,\mathrm{d}\mu = \int (f \mathbb{1}_A) \,\mathrm{d}\mu$.

(g) Falls $g\colon \Omega \to [0, \infty]$ nichtnegativ und integrierbar ist und f eine numerische messbare Funktion mit $|f| \leq g$, so ist auch f integrierbar.

(h) Die Abbildung $f \mapsto \int f \,\mathrm{d}\mu$ ist linear und monoton (manchmal auch isoton genannt), d. h. für $a, b \in \mathbb{R}$ und integrierbare Funktionen f und g ist auch $af + bg$ integrierbar mit $\int (af + bg) \,\mathrm{d}\mu = a \int f \,\mathrm{d}\mu + b \int g \,\mathrm{d}\mu$, und falls $f \leq g$, ist $\int f \,\mathrm{d}\mu \leq \int g \,\mathrm{d}\mu$. Insbesondere ist $\mathcal{L}^1(\Omega, \mathcal{F}, \mu)$ ein Vektorraum. Die Linearität sieht man leicht für einfache Funktionen direkt ein, durch Grenzübergang überträgt sie sich leicht auf nichtnegative Funktionen, und der allgemeine Fall folgt wiederum leicht durch Zerlegung in Positiv- und Negativteil unter Benutzung der Dreiecksungleichung und Bemerkung (g). Dies ist unser erstes Beispiel für ein Beweisverfahren, das strukturell dem Prinzip der Definition 6.5.1 folgt und formal mit Hilfe von Satz 6.4.11 durchgeführt wird. Die Monotonie sieht man durch Anwendung der Linearität auf die nichtnegative Funktion $g - f$.

(i) Falls $\mu = \mathbb{P}$ ein Wahrscheinlichkeitsmaß ist, so schreiben wir auch $\mathbb{E}(f)$ (oder $\mathbb{E}_{\mathbb{P}}(f)$, falls $\mathbb{P}$ betont werden soll) statt $\int f \,\mathrm{d}\mathbb{P}$ und nennen $\mathbb{E}(f)$ den *Erwartungswert* von f.

◇

Bemerkung 6.5.3 (fast überall, fast sicher)**.** Eine wichtige, oft benutzte Begriffsbildung im Zusammenhang mit Eigenschaften in Bezug auf Elemente in Ω ist die folgende. Wir sagen, eine Eigenschaft gilt *μ-fast überall* oder kurz *μ-f. ü.*, falls die Menge derjenigen ω, für die sie nicht gilt, eine μ-Nullmenge ist; siehe Definition 6.2.10. Falls also z. B. zwei messbare Funktionen f und g sich nur auf einer μ-Nullmenge von einander unterscheiden, d. h. wenn $\mu(\{\omega\colon f(\omega) \neq g(\omega)\}) = 0$, so schreiben wir $f = g$ μ-f. ü. In dem Fall, dass $\mu = \mathbb{P}$ ein Wahrscheinlichkeitsmaß ist, sagen wir *$\mathbb{P}$-fast sicher* und schreiben $\mathbb{P}$-f. s. ◇

Lemma 6.5.4. *Es seien $f, g \in \mathcal{L}^1$. Dann gelten:*

(a) Falls $f = g$ μ-f. ü., so gilt $\int f \,\mathrm{d}\mu = \int g \,\mathrm{d}\mu$.

(b) Falls $f \geq g$ und $\int f \,\mathrm{d}\mu = \int g \,\mathrm{d}\mu$, so gilt $f = g$ μ-f. ü.

Beweis. (a) Durch Zerlegung in Positiv- und Negativteil können wir annehmen, dass f und g nichtnegativ sind. Es sei $(f_n)_{n\in\mathbb{N}}$ eine Folge von einfachen Funktionen mit $f_n \uparrow f$. Für die (messbare) Menge $D = \{\omega \in \Omega\colon f(\omega) = g(\omega)\}$ gelten $\mu(D^{\mathrm{c}}) = 0$ und $\mathbb{1}_D f_n \uparrow \mathbb{1}_D f = \mathbb{1}_D g$. Man beachte, dass für jede einfache Funktion h gilt: $\int_D h \,\mathrm{d}\mu = \int h \,\mathrm{d}\mu$. Insbesondere ist $\int_D f_n \,\mathrm{d}\mu = \int f_n \,\mathrm{d}\mu$ für jedes $n \in \mathbb{N}$, und der Grenzübergang $n \to \infty$ lässt $\int_D f \,\mathrm{d}\mu = \int f \,\mathrm{d}\mu$ folgen. Analog gilt $\int_D g \,\mathrm{d}\mu = \int g \,\mathrm{d}\mu$. Wegen $\mathbb{1}_D f = \mathbb{1}_D g$ folgt die Aussage.

(b) Nach Voraussetzung gelten $f - g \geq 0$ und $\int (f - g) \,\mathrm{d}\mu = 0$. Es genügt also, den Fall $g = 0$ zu betrachten. Für $n \in \mathbb{N}$ sei $A_n = \{f \geq \frac{1}{n}\}$. Dann gilt $f \geq \frac{1}{n} \mathbb{1}_{A_n}$. Aus $\int f \,\mathrm{d}\mu = 0$ folgt somit $\int \frac{1}{n} \mathbb{1}_{A_n} \,\mathrm{d}\mu = \frac{1}{n} \mu(A_n) = 0$, also $\mu(A_n) = 0$. Da $\{f > 0\} = \bigcup_{n\in\mathbb{N}} A_n$ eine abzählbare Vereinigung von Nullmengen ist, ist $\{f > 0\}$ nach Lemma 6.2.11 selber eine. □

Bemerkung 6.5.5 (Riemann-Integral)**.** Die Integration bezüglich des Lebesgue-Maßes λ_n auf $(\mathbb{R}^n, \mathcal{B}_n)$ wird Lebesgue-Integration genannt. Der größte und wichtigste Unterschied zum Riemann-Integral ist, dass die Unterteilungen beim Riemann-Integral im Definitionsbereich gemacht werden und beim Lebesgue-Integral im Wertebereich. Das Riemann-Integral kommt nicht von einem

Maß her, denn dieser Integralbegriff ist nicht kompatibel mit abzählbaren Operationen.[8] Trotzdem gibt es enge Zusammenhänge zwischen dem Riemann- und dem Lebesgue-Integral. Wir zitieren die wichtigsten Tatsachen (Beweis siehe etwa [Ba91, Satz 16.4 und Korollar 16.5]):

(a) Jede auf einem kompakten Intervall messbare, Riemann-integrierbare Funktion ist auch Lebesgue-integrierbar, und die beiden Integrale stimmen überein.

(b) Jede auf einem unbeschränkten Intervall I messbare nichtnegative Funktion, die über jedes kompakte Teilintervall Riemann-integrierbar ist, ist genau dann Lebesgue-integrierbar über I, wenn ihr uneigentliches Riemann-Integral über I existiert, und die beiden Integrale stimmen dann überein.

Dass man in (b) nicht auf die Nichtnegativität verzichten kann, sieht man an dem Beispiel $(0, \infty) \ni x \mapsto \frac{1}{x} \sin x$. $\diamond$

6.6 Bildmaße und Dichten

Mit Hilfe von messbaren Abbildungen kann man aus gegebenen Maßen weitere Maße konstruieren. Wir stellen die beiden wichtigsten Prinzipien hier vor.

Bildmaße

Es sei $(\Omega, \mathcal{F}, \mu)$ ein Maßraum und $(\Omega', \mathcal{F}')$ ein weiterer messbarer Raum. Ferner sei $f\colon \Omega \to \Omega'$ eine messbare Funktion. Wir definieren $\mu \circ f^{-1}\colon \mathcal{F}' \to [0, \infty]$ durch

$$\mu \circ f^{-1}(A) = \mu\big(f^{-1}(A)\big), \qquad A \in \mathcal{F}', \tag{6.6.1}$$

wobei $f^{-1}(A)$ wie immer das Urbild von A unter f bezeichnet.

Satz 6.6.1. *(a) $\mu \circ f^{-1}$ ist ein Maß. Falls μ ein Wahrscheinlichkeitsmaß ist, so auch $\mu \circ f^{-1}$.*

(b) Für jede nichtnegative messbare numerische Funktion $\phi\colon \Omega' \to [0, \infty]$ gilt

$$\int_{\Omega'} \phi \, \mathrm{d}(\mu \circ f^{-1}) = \int_{\Omega} (\phi \circ f) \, \mathrm{d}\mu. \tag{6.6.2}$$

(c) Für jede messbare Funktion $\phi\colon \Omega' \to \mathbb{R}$ gilt

$$\phi \in \mathcal{L}^1(\Omega', \mathcal{F}', \mu \circ f^{-1}) \qquad \Longleftrightarrow \qquad \phi \circ f \in \mathcal{L}^1(\Omega, \mathcal{F}, \mu),$$

und in diesem Fall gilt (6.6.2) *ebenfalls.*

Beweis. (a) folgt aus den allgemein gültigen Rechenregeln für die Abbildung $f^{-1}\colon \mathcal{F}' \to \mathcal{F}$. Die Aussage (b) für Indikatorfunktionen $\phi = \mathbb{1}_A$ mit $A \in \mathcal{F}'$ folgt direkt aus der Definition (6.6.1), und die allgemeine Aussage (b) folgt mit Satz 6.4.11 auf die übliche Weise. (c) ergibt sich direkt aus einer Anwendung von (b) auf $|\phi|$, und die Formel (6.6.2) für allgemeine integrierbare Funktionen zeigt man leicht durch Zerlegung in Positiv- und Negativteil. $\square$

[8] Zum Beispiel ist $\mathbb{1}_{\{\omega\}}$ für jedes $\omega \in [0, 1]$ Riemann-integrierbar, aber nicht die abzählbare Summe $\mathbb{1}_{\mathbb{Q} \cap [0,1]} = \sum_{\omega \in \mathbb{Q} \cap [0,1]} \mathbb{1}_{\{\omega\}}$.

Definition 6.6.2 (Bildmaß). *Das Maß* $\mu \circ f^{-1}$ *heißt das* durch f induzierte Maß *oder das* Bildmaß *von* μ *unter* f. *Es sind auch die Schreibweisen* μf^{-1} *und* $f(\mu)$ *gebräuchlich.*

Falls das Bildmaß $\mu \circ f^{-1}$ σ-endlich ist, so auch μ, aber die Umkehrung gilt im Allgemeinen nicht (Übungsaufgabe).

Dichten

Nun kommt die nach Korollar 6.3.4 angekündigte starke Verallgemeinerung des Begriffs der Dichte. Wir erinnern uns, dass $\mathcal{B}$ die Borel-σ-Algebra auf $\mathbb{R}$ bezeichnet.

Definition 6.6.3 (Dichten). *Es seien* μ *und* ν *zwei Maße auf einem messbaren Raum* $(\Omega, \mathcal{F})$. *Eine* $\mathcal{F}$-$\mathcal{B}$-*messbare Abbildung* $\varphi\colon \Omega \to [0, \infty)$ *heißt eine* Dichte *von* ν *bezüglich* μ, *wenn* $\nu(A) = \int_A \varphi \, \mathrm{d}\mu$ *für alle* $A \in \mathcal{F}$ *gilt. Wir schreiben dann auch* $\varphi = \frac{\mathrm{d}\nu}{\mathrm{d}\mu}$ *und* $\mathrm{d}\nu = \varphi \, \mathrm{d}\mu$.

Wenn φ eine Dichte von ν bezüglich μ ist, dann ist jede messbare Abbildung $\phi\colon \Omega \to \mathbb{R}$ genau dann ν-integrierbar, wenn $\phi\varphi$ integrierbar bezüglich μ ist.

Lemma 6.6.4. *Es seien* μ *und* ν *zwei Maße auf einem messbaren Raum* $(\Omega, \mathcal{F})$. *Falls eine Dichte von* ν *bezüglich* μ *existiert, so ist sie eindeutig bis auf Gleichheit* μ-*f. ü.*

Beweis. Seien φ_1 und φ_2 zwei Dichten von ν bezüglich μ. Wir betrachten die messbare Menge $A = \{\varphi_1 > \varphi_2\} = \{\omega \in \Omega \colon \varphi_1(\omega) > \varphi_2(\omega)\}$. Dann gilt

$$\int \mathbb{1}_A(\varphi_1 - \varphi_2) \, \mathrm{d}\mu = \int_A \varphi_1 \, \mathrm{d}\mu - \int_A \varphi_2 \, \mathrm{d}\mu = \nu(A) - \nu(A) = 0.$$

Wegen $\mathbb{1}_A(\varphi_1 - \varphi_2) \geq 0$ folgt mit Lemma 6.5.4(b), dass $\mathbb{1}_A(\varphi_1 - \varphi_2) = 0$ μ-f. ü., also $\varphi_1 \leq \varphi_2$ μ-f. ü. Die komplementäre Ungleichung $\varphi_1 \geq \varphi_2$ μ-f. ü. folgt analog. □

Bemerkung 6.6.5 (Absolutstetigkeit). Wenn ν eine Dichte φ bezüglich μ hat, so sind alle μ-Nullmengen auch ν-Nullmengen. Den letzteren Sachverhalt drücken wir auch aus mit der Notation $\nu \ll \mu$ und sagen dass ν *absolutstetig* bezüglich μ sei. Der berühmte *Satz von Radon-Nikodym* (den wir nicht behandeln werden) dreht die Aussage um: Falls $\nu \ll \mu$, so existiert eine Dichte von ν bezüglich μ. ◇

Diskrete Maße (siehe Beispiel 6.1.10) haben keine Dichten bezüglich des Lebesgue-Maßes. Ein Maß kann durchaus aus einem diskreten und einem absolutstetigen Anteil bestehen. Es folgen die wichtigsten Beispiele mit Dichten:

Beispiel 6.6.6 (Lebesgue-Dichten). Wichtige Beispiele von Wahrscheinlichkeitsmaßen auf $\mathbb{R}$ mit Lebesgue-Dichten sind die folgenden:

Normalverteilung $\mathcal{N}(a, \sigma^2)$	$\varphi_{a,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big\{-\frac{(x-a)^2}{2\sigma^2}\Big\}$	Parameter $a \in \mathbb{R}, \sigma > 0$,
Exponentialverteilung	$\varphi_\alpha(x) = \alpha e^{-\alpha x} \mathbb{1}_{[0,\infty)}(x)$,	Parameter $\alpha > 0$,
Cauchy-Verteilung	$\varphi_c(x) = \frac{c}{\pi(c^2 + x^2)}$,	Parameter $c > 0$.

◇

Beispiel 6.6.7 (mehrdimensionale Normalverteilung). Die *Standardnormalverteilung* $\mathcal{N}$ auf dem messbaren Raum $(\mathbb{R}^n, \mathcal{B}_n)$ ist definiert als das Maß, das die Dichte

$$\varphi(x) = (2\pi)^{-n/2} \exp\Big\{-\frac{1}{2}\sum_{i=1}^n x_i^2\Big\}, \qquad x = (x_1, \dots, x_n) \in \mathbb{R}^n,$$

besitzt. Jedes Bildmaß $\mathcal{N} \circ f^{-1}$ mit einer affinen Abbildung $f\colon \mathbb{R}^n \to \mathbb{R}^n$, d. h. $f(x) = Ax + b$ mit einer $n \times n$-Matrix A und einem Vektor $b \in \mathbb{R}^n$, heißt eine *Normalverteilung*. Dieses Maß $\mathcal{N} \circ f^{-1}$ besitzt genau dann eine Dichte, wenn A invertierbar ist, und zwar ist dann eine Dichte gegeben durch

$$\varphi_{b,\Sigma}(x) = \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} \exp\Big\{-\frac{1}{2}(x-b)^{\mathrm{T}}\Sigma^{-1}(x-b)\Big\}, \qquad x \in \mathbb{R}^n,$$

wobei $\Sigma = AA^{\mathrm{T}}$. Falls nämlich A nicht invertierbar ist, ist das Bild $f(\mathbb{R}^n)$ in einer affinen Hyperebene enthalten, besitzt also das Maß Null unter λ_n, aber es gilt $\mathcal{N} \circ f^{-1}(f(\mathbb{R}^n)) = 1$. Also hat $\mathcal{N} \circ f^{-1}$ keine Lebesgue-Dichte. Wenn andererseits A invertierbar ist, so ist auch $f\colon \mathbb{R}^n \to \mathbb{R}^n$ bijektiv, und der Transformationssatz für lineare Transformationen unter dem Lebesgue-Maß (der aus der Analysis bekannt sein sollte) impliziert die Aussage.

Daher schreibt man oft auch $\mathcal{N}(b, \Sigma)$ an Stelle von $\mathcal{N} \circ f^{-1}$, und dann ist $\mathcal{N} = \mathcal{N}(0, I)$, wobei I die $n \times n$-Einheitsmatrix ist. ◇

6.7 Die L^p-Räume

Es sei $(\Omega, \mathcal{F}, \mu)$ ein beliebiger Maßraum. In diesem Abschnitt führen wir die zur Potenz $p \geq 1$ integrierbaren Funktionen ein und leiten ein paar wichtige Ungleichungen her.

Für eine messbare numerische Funktion $f\colon \Omega \to \overline{\mathbb{R}}$ und $p \in [1, \infty]$ definieren wir

$$\|f\|_p = \begin{cases} \big(\int_\Omega |f|^p \,\mathrm{d}\mu\big)^{\frac{1}{p}}, & \text{falls } p < \infty, \\ \sup\{a \geq 0\colon \mu(|f| \geq a) > 0\}, & \text{falls } p = \infty, \end{cases}$$

und

$$\mathcal{L}^p(\Omega, \mathcal{F}, \mu) = \mathcal{L}^p(\mu) = \{f\colon f \text{ ist messbar und numerisch mit } \|f\|_p < \infty\}.$$

Dies ist offensichtlich eine Erweiterung der Definition von $\mathcal{L}^1$. Man überlegt sich als Übungsaufgabe, dass $\|f\|_\infty = \min\{c \geq 0\colon |f| \leq c\ \mu\text{-f.ü.}\}$, d. h. $\|f\|_\infty = c$ gilt genau dann, wenn es eine Nullmenge N gibt mit $\sup_{N^c} |f| = c$, d. h. $\|f\|_\infty$ ist das Supremum von $|f|$ außerhalb einer μ-Nullmenge.

Eines unserer Ziele ist es, $(\mathcal{L}^p, \|\cdot\|_p)$ als einen normierten Raum aufzufassen. Immerhin ist $\|\cdot\|_p$ positiv homogen (d. h. $\|\alpha f\|_p = |\alpha|\|f\|_p$ für $\alpha \in \mathbb{R}$ und $f \in \mathcal{L}^p(\mu)$), und in Satz 6.7.8 werden wir die Gültigkeit der Dreiecksungleichung zeigen. Allerdings ist $\|\cdot\|_p$ nicht definit, denn aus Lemma 6.5.4 folgt: $\|f\|_p = 0 \iff f = 0$ μ-f. ü. Dies bringt uns auf die Idee, zu dem folgenden Quotientenraum über zu gehen: Mit

$$\mathcal{N} = \{f \in \mathcal{L}^p(\mu)\colon f = 0\ \mu\text{-fast überall}\}$$

definieren wir

$$L^p(\Omega,\mathcal{F},\mu) = L^p(\mu) = \mathcal{L}^p(\Omega,\mathcal{F},\mu)/\mathcal{N},$$

also ist $L^p(\mu)$ die Menge aller Äquivalenzklassen von $\mathcal{L}^p(\mu)$ bezüglich der Äquivalenzrelation 'Gleichheit μ-fast überall'. Für eine Klasse $[f] \in L^p(\mu)$ setzen wir $\|[f]\|_p = \|f\|_p$, und wiederum Lemma 6.5.4 sagt uns, dass diese Definition sinnvoll ist, d. h. nicht vom gewählten Vertreter f abhängt.

Satz 6.7.1. *Für jedes $p \in [1,\infty]$ ist $(L^p(\mu), \|\cdot\|_p)$ ein Banachraum. Im Spezialfall $p = 2$ ist $(L^2(\mu), \|\cdot\|_2)$ ein Hilbertraum mit dem durch $\langle [f],[g]\rangle = \int fg\,\mathrm{d}\mu$ definierten Skalarprodukt.*

Den Beweis von Satz 6.7.1 werden wir nicht ausformulieren, insbesondere die Vollständigkeit werden wir hier nicht beweisen (siehe dazu etwa [Ba91, Paragraf 15]). Die Dreiecksungleichung für $\|\cdot\|_p$ wird in Satz 6.7.8 bewiesen, und daraus folgen auch die Vektorraumeigenschaften. Dass $\langle\cdot,\cdot\rangle$ ein wohldefiniertes Skalarprodukt ist, das die Norm $\|\cdot\|_2$ induziert, zeigt man auf elementare Weise.

Wir diskutieren nun die Jensen'sche Ungleichung, eine der wichtigsten Ungleichungen der Wahrscheinlichkeitstheorie, und ein paar Anwendungen. Wir erinnern (siehe Bemerkung 6.5.2), dass wir $\mathbb{E}(X)$ statt $\int X\,\mathrm{d}\mathbb{P}$ schreiben, wenn $\mathbb{P}$ ein Wahrscheinlichkeitsmaß ist.

Satz 6.7.2 (Jensen'sche Ungleichung)**.** *Es sei $(\Omega,\mathcal{F},\mathbb{P})$ ein Wahrscheinlichkeitsraum, und es sei $X\colon \Omega \to \mathbb{R}$ integrierbar. Ferner sei $I \subset \mathbb{R}$ ein Intervall und $\varphi\colon I \to \mathbb{R}$ eine konvexe Funktion. Es gelte $\mathbb{P}(X \notin I) = 0$. Dann gilt*

$$\mathbb{E}(\varphi\circ X) \geq \varphi\big(\mathbb{E}(X)\big).$$

Beweis. In dem Fall, dass $\mathbb{E}(X)$ im Innern von I liegt, gibt es eine Gerade unterhalb des Grafen von φ, die durch den Punkt $(\mathbb{E}(X), \varphi(\mathbb{E}(X)))$ geht und eine Steigung α hat:

$$\varphi(y) \geq \alpha(y - \mathbb{E}(X)) + \varphi(\mathbb{E}(X)), \qquad y \in I.$$

Nun ersetzen wir y durch $X(\omega)$ und integrieren über ω, um zu erhalten:

$$\mathbb{E}(\varphi\circ X) = \int \varphi(X(\omega))\,\mathbb{P}(\mathrm{d}\omega) \geq \alpha\int (X(\omega) - \mathbb{E}(X))\,\mathbb{P}(\mathrm{d}\omega) + \int \varphi(\mathbb{E}(X))\,\mathbb{P}(\mathrm{d}\omega) = \varphi\big(\mathbb{E}(X)\big).$$

Im anderen Fall, wo $\mathbb{E}(X)$ ein Randpunkt von I ist, gilt $X = \mathbb{E}(X)$ $\mathbb{P}$-fast sicher, und die Behauptung gilt trivialerweise. □

Daraus sieht man leicht als Übungsaufgabe, dass für endliche Maßräume bei wachsendem p die $\mathcal{L}^p$- und die L^p-Räume absteigend in einander enthalten sind:

Korollar 6.7.3. *Es sei $(\Omega,\mathcal{F},\mu)$ ein endlicher Maßraum. Dann gelten für $1 \leq p \leq p' \leq \infty$ die Inklusionen $\mathcal{L}^{p'}(\mu) \subset \mathcal{L}^p(\mu)$ und $L^{p'}(\mu) \subset L^p(\mu)$. Falls $\mu(\Omega) = 1$, so ist die Abbildung $[1,\infty] \to [0,\infty]$, $p \mapsto \|f\|_p$ für jede messbare numerische Funktion f monoton nichtfallend.*

Bemerkung 6.7.4 (Entropien)**.** Die *relative Entropie* eines Wahrscheinlichkeitsmaßes $\mathbb{P}$ bezüglich eines Wahrscheinlichkeitsmaßes $\widetilde{\mathbb{P}}$ definieren wir durch

$$H\big(\mathbb{P} \mid \widetilde{\mathbb{P}}\big) = \int \log\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\widetilde{\mathbb{P}}}\,\mathrm{d}\mathbb{P},$$

falls die Dichte $\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\widetilde{\mathbb{P}}}$ existiert, und $H(\mathbb{P} \mid \widetilde{\mathbb{P}}) = \infty$ sonst. Als eine einfache Übungsaufgabe wendet man die Jensen'sche Ungleichung auf die konvexe Abbildung $x \mapsto x \log x$ an, um zu zeigen, dass die Entropie nicht negativ ist. Außerdem ist leicht zu sehen, dass $H(\cdot \mid \widetilde{\mathbb{P}})$ eine konvexe Abbildung ist. Entropien spielen wichtige Rollen in der Theorie der Großen Abweichungen, der Informationstheorie und beim Umgang mit empirischen Maßen. ◇

Eine weitere wichtige, berühmte Ungleichung ist die folgende, die wir nun als eine Anwendung der Jensenschen Ungleichung beweisen. Wie zu Beginn sei $(\Omega, \mathcal{F}, \mu)$ ein beliebiger Maßraum.

Satz 6.7.5 (Hölder'sche Ungleichung). *Es seien $p, q \in [1, \infty]$ mit $\frac{1}{p} + \frac{1}{q} = 1$. Ferner seien $f \in \mathcal{L}^p(\mu)$ und $g \in \mathcal{L}^q(\mu)$. Dann gelten $fg \in \mathcal{L}^1(\mu)$ und*

$$\|fg\|_1 \leq \|f\|_p \cdot \|g\|_q.$$

Beweis. Der Fall $p = \infty$ ist sehr leicht; wir behandeln nur den Fall $1 < p < \infty$. Wir dürfen auch voraussetzen, dass $f \neq 0$ und $g \neq 0$. Wir betrachten die Funktionen $\widetilde{f} = (|f|/\|f\|_p)^p$ und $\widetilde{g} = (|g|/\|g\|_q)^q$. Dann gilt $\int \widetilde{f}\,\mathrm{d}\mu = \int \widetilde{g}\,\mathrm{d}\mu = 1$, und $\mathbb{P} = \widetilde{f}\,\mathrm{d}\mu$ ist ein Wahrscheinlichkeitsmaß. Ferner sieht man, dass

$$\int \frac{|fg|}{\|f\|_p \|g\|_q}\,\mathrm{d}\mu = \int \widetilde{f}^{1/p} \widetilde{g}^{1/q}\,\mathrm{d}\mu = \int \Big(\frac{\widetilde{g}}{\widetilde{f}}\Big)^{1/q} \widetilde{f}\,\mathrm{d}\mu = \mathbb{E}\Big[\Big(\frac{\widetilde{g}}{\widetilde{f}}\Big)^{1/q}\Big],$$

wobei wir $\mathbb{E}$ für den Erwartungswert bezüglich $\mathbb{P}$ geschrieben haben. (Man beachte auch, dass der Quotient $\mathbb{P}$-fast sicher wohldefiniert ist.) Nun benutzen wir die Konvexität der Abbildung $x \mapsto x^q$ und erhalten mit Hilfe der Jensen'schen Ungleichung:

$$\int \frac{|fg|}{\|f\|_p \|g\|_q}\,\mathrm{d}\mu = \mathbb{E}\Big[\Big(\frac{\widetilde{g}}{\widetilde{f}}\Big)^{1/q}\Big] \leq \mathbb{E}\Big[\frac{\widetilde{g}}{\widetilde{f}}\Big]^{1/q} = \Big(\int \widetilde{g}\,\mathrm{d}\mu\Big)^{1/q} = 1,$$

also die Hölder'sche Ungleichung. (Genau genommen, dürfen wir die Jensen'sche Ungleichung nur anwenden, wenn wir wissen, dass der Erwartungswert von $(\widetilde{g}/\widetilde{f})^{1/q}$ endlich ist, aber indem man den obigen Beweis zunächst nur für nichtnegative einfache Funktionen durchführt und dann zu monotonen Grenzwerten übergeht, erhält man den allgemeinen Fall.) □

Bemerkung 6.7.6 (Momenten erzeugende Funktion I). Es sei μ ein Maß auf $(\mathbb{R}, \mathcal{B})$. Für $t \in \mathbb{R}$ definieren wir $M(t) = \int \mathrm{e}^{tx}\,\mu(\mathrm{d}x) \in [0, \infty]$. Man nennt M die *Momenten erzeugende Funktion* von μ (dieser Name wird in Bemerkung 6.8.10 erhellt.) Es sei $I = \{t \in \mathbb{R} \colon M(t) < \infty\}$ nicht leer. Man sieht leicht, dass I ein Intervall ist. Eine Anwendung der Hölder'schen Ungleichung zeigt, dass die Abbildung $\log M$ konvex auf I ist. ◇

Der symmetrische Spezialfall der Hölder-Ungleichung ist ebenfalls berühmt:

Korollar 6.7.7 (Cauchy-Schwarz-Ungleichung). *Für je zwei Funktionen $f, g \in \mathcal{L}^2(\mu)$ gelten $fg \in \mathcal{L}^1(\mu)$ und $\|fg\|_1 \leq \|f\|_2 \cdot \|g\|_2$.*

Als eine Anwendung der Hölder'schen Ungleichung zeigen wir nun die Dreiecksungleichung für $\|\cdot\|_p$:

Satz 6.7.8 (Minkowski-Ungleichung). *Es sei $p \in [1, \infty]$. Dann gilt für alle $f, g \in \mathcal{L}^p(\mu)$:*

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p.$$

Beweis. Die Fälle $p = \infty$ und $p = 1$ sind leicht bzw. schon bekannt, so dass wir nur den Fall $1 < p < \infty$ behandeln. Die Dreiecksungleichung impliziert, dass $\|f + g\|_p \leq \||f| + |g|\|_p$, sodass wir annehmen dürfen, dass $f \geq 0$ und $g \geq 0$. Wegen $(f + g)^p \leq [2 \max\{f, g\}]^p \leq 2^p(f^p + g^p)$ ist schon klar, dass $\|f + g\|_p < \infty$. Wir schreiben

$$\int (f + g)^p \, \mathrm{d}\mu = \int f(f + g)^{p-1} \, \mathrm{d}\mu + \int g(f + g)^{p-1} \, \mathrm{d}\mu$$

und wenden die Hölder'sche Ungleichung (mit einem $q > 1$, so dass $\frac{1}{p} + \frac{1}{q} = 1$) auf die beiden Integrale an, um zu erhalten:

$$\int (f + g)^p \, \mathrm{d}\mu \leq \|f\|_p \|(f + g)^{p-1}\|_q + \|g\|_p \|(f + g)^{p-1}\|_q = \big(\|f\|_p + \|g\|_p\big) \Big(\int (f + g)^p \, \mathrm{d}\mu\Big)^{1/q}.$$

Elementare Umformungen lassen die Aussage folgen. □

6.8 Die fundamentalen Konvergenzsätze

In diesem Abschnitt sei $(\Omega, \mathcal{F}, \mu)$ wieder ein gegebener Maßraum. Man steht oft vor dem Problem, ein Integral mit einem Grenzwert zu vertauschen, also eine Regel wie $\int \lim_{n\to\infty} f_n \, \mathrm{d}\mu = \lim_{n\to\infty} \int f_n \, \mathrm{d}\mu$ anzuwenden. Wir geben die zwei berühmten hinreichenden Kriterien, unter denen man dies tun darf. Man beachte, dass die Voraussetzungen extrem schwach sind im Vergleich etwa zum Riemann-Integralbegriff.

Satz 6.8.1 (Monotoner Konvergenzsatz). *Sei $(f_n)_{n\in\mathbb{N}}$ eine Folge nichtnegativer, messbarer numerischer Funktionen mit $f_n \uparrow f$ μ-f. ü. für $n \to \infty$. Dann gilt*

$$\lim_{n\to\infty} \int f_n \, \mathrm{d}\mu = \int f \, \mathrm{d}\mu.$$

Man beachte, dass wir nicht vorausgesetzt haben, dass $\int f \, \mathrm{d}\mu$ endlich ist.

Beweis. Durch eine Modifikation auf Nullmengen (die nach Lemma 6.5.4(a) nichts an den Werten der Integrale ändert) können wir annehmen, dass $f_n(\omega) \uparrow f(\omega)$ sogar für *alle* $\omega \in \Omega$ gilt. Wegen der Monotonie des Integrals gilt natürlich $\int f_n \, \mathrm{d}\mu \leq \int f \, \mathrm{d}\mu$ für jedes $n \in \mathbb{N}$. Wegen Monotonie der Folge $(f_n)_{n\in\mathbb{N}}$ existiert der Grenzwert $\lim_{n\to\infty} \int f_n \, \mathrm{d}\mu$ in $[0, \infty]$. Wir müssen also nur noch zeigen, dass er nicht kleiner als $\int f \, \mathrm{d}\mu$ ist.

Wir wählen für jedes $n \in \mathbb{N}$ eine Folge $(f_{n,m})_{m\in\mathbb{N}}$ nichtnegativer einfacher Funktionen mit $f_{n,m} \uparrow f_n$ für $m \to \infty$. Wir betrachten die einfache Funktion

$$g_{n,m} = \max\{f_{1,m}, \dots, f_{n,m}\}.$$

Es gilt auch $g_{n,m} \uparrow f_n$ für $m \to \infty$ für jedes $n \in \mathbb{N}$, denn die Monotonie in m ist klar, und es gilt $f_n \geq g_{n,m} \geq f_{n,m}$. Außerdem ist offensichtlich $g_{n,m}$ monoton in n, d. h. $g_{n+1,m} \geq g_{n,m}$ für alle n, m. Daraus folgt sehr einfach, dass $g_{n,n} \uparrow f$ für $n \to \infty$. Damit ergibt sich

$$\int f \,\mathrm{d}\mu = \lim_{n\to\infty} \int g_{n,n} \,\mathrm{d}\mu \leq \lim_{n\to\infty} \int f_n \,\mathrm{d}\mu,$$

denn $g_{n,n} \leq f_n$. □

Eine Variante des Monotonen Konvergenzsatzes ist der folgende Satz (Beweis als Übungsaufgabe):

Korollar 6.8.2 (Satz von Beppo-Lévy). *Sei $(f_n)_{n\in\mathbb{N}}$ eine Folge integrierbarer Funktionen mit $f_n \uparrow f$ μ-f. ü. für $n \to \infty$ und $\sup_{n\in\mathbb{N}} \int f_n \,\mathrm{d}\mu < \infty$. Dann ist auch f integrierbar, und es gilt*

$$\lim_{n\to\infty} \int f_n \,\mathrm{d}\mu = \int f \,\mathrm{d}\mu.$$

Weitere Folgerungen aus dem Satz von der monotonen Konvergenz:

Korollar 6.8.3. *Für beliebige nichtnegative messbare Funktionen $f_1, f_2, f_3, \ldots$ gilt*

$$\int \sum_{n\in\mathbb{N}} f_n \,\mathrm{d}\mu = \sum_{n\in\mathbb{N}} \int f_n \,\mathrm{d}\mu.$$

Insbesondere gilt für paarweise disjunkte messbare Mengen $A_1, A_2, A_3, \ldots$ und jede nichtnegative messbare Funktion f:

$$\int_{\bigcup_{n\in\mathbb{N}} A_n} f \,\mathrm{d}\mu = \sum_{n\in\mathbb{N}} \int_{A_n} f \,\mathrm{d}\mu.$$

Satz 6.8.4 (Lemma von Fatou). *Es sei $(f_n)_{n\in\mathbb{N}}$ eine beliebige Folge nichtnegativer messbarer numerischer Funktionen. Dann gilt*

$$\int \liminf_{n\to\infty} f_n \,\mathrm{d}\mu \leq \liminf_{n\to\infty} \int f_n \,\mathrm{d}\mu.$$

Beweis. Nach Definition des Limes Inferior konvergiert $\inf_{m\colon m\geq n} f_m$ monoton steigend gegen $\liminf_{n\to\infty} f_n$. Nach dem Monotonen Konvergenzsatz gilt

$$\int \liminf_{n\to\infty} f_n \,\mathrm{d}\mu = \lim_{n\to\infty} \int \inf_{m\colon m\geq n} f_m \,\mathrm{d}\mu \leq \lim_{n\to\infty} \inf_{m\colon m\geq n} \int f_m \,\mathrm{d}\mu = \liminf_{n\to\infty} \int f_n \,\mathrm{d}\mu.$$

□

Bemerkung 6.8.5. Es ist leicht, Beispiele zu geben, in denen die Ungleichung im Lemma von Fatou strikt ist. Etwa für $(\Omega, \mathcal{F}, \mu) = (\mathbb{R}, \mathcal{B}, \lambda)$ kann man 'Zackenfunktionen' wählen, deren Graf die lineare Interpolation der Punkte $(0,0)$, $(\frac{1}{n}, n)$ und $(\frac{2}{n}, 0)$ ist. ◇

Der wohl nützlichste und stärkste Satz über die Vertauschung von Grenzwert und Integral ist der folgende.

Satz 6.8.6 (Satz von Lebesgue, Satz von der majorisierten Konvergenz). *Es sei $(f_n)_{n\in\mathbb{N}}$ eine Folge von μ-integrierbaren Funktionen. Es existiere eine nichtnegative Funktion $g \in \mathcal{L}^1(\mu)$ mit $|f_n(\omega)| \leq g(\omega)$ für alle $\omega \in \Omega$ und alle $n \in \mathbb{N}$. Falls $f(\omega) = \lim_{n\to\infty} f_n(\omega)$ für μ-fast alle $\omega \in \Omega$ existiert, so gilt*

$$\lim_{n\to\infty} \int f_n \,\mathrm{d}\mu = \int f \,\mathrm{d}\mu.$$

Beweis. Durch Änderung auf einer Nullmenge können wir erreichen, dass $f_n(\omega) \to f(\omega)$ für *alle* $\omega \in \Omega$ gilt. Es gilt $f_n + g \geq 0$, und das Lemma von Fatou impliziert

$$\int f \,\mathrm{d}\mu = \int (f+g) \,\mathrm{d}\mu - \int g \,\mathrm{d}\mu \leq \liminf_{n\to\infty} \int (f_n + g) \,\mathrm{d}\mu - \int g \,\mathrm{d}\mu \leq \liminf_{n\to\infty} \int f_n \,\mathrm{d}\mu.$$

Wenn wir den selben Trick auf $g - f_n \geq 0$ anwenden, erhalten wir die komplementäre Ungleichung $\int f \,\mathrm{d}\mu \geq \limsup_{n\to\infty} \int f_n \,\mathrm{d}\mu$. □

Der Spezialfall eines diskreten Maßraums enthält aus der Analysis bekannte Sätze über Doppelfolgen:

Bemerkung 6.8.7 (Zahlenfolgen). Es ist eine der Stärken der Maßtheorie, dass viele bekannte Sätze über reelle Zahlenfolgen einen Spezialfall darstellen. Wendet man zum Beispiel den Satz von Lebesgue auf das Zählmaß auf $\mathbb{N}$ an, also $(\Omega, \mathcal{F}, \mu) = (\mathbb{N}, \mathcal{P}(\mathbb{N}), \mu)$ mit dem Zählmaß $\mu = \sum_{n\in\mathbb{N}} \delta_n$ (siehe Beispiel 6.1.10), dann erhält man das folgende Resultat. Zunächst muss man sich damit vertraut machen, dass $\mathcal{L}^1(\mathbb{N}, \mathcal{P}(\mathbb{N}), \mu)$ nichts anderes ist als der bekannte Folgenraum $\ell^1(\mathbb{N})$. Der Satz von Lebesgue besagt dann, dass für jede Doppelfolge $(a_{i,n})_{i,n\in\mathbb{N}}$, die die Bedingungen

$$a_i = \lim_{n\to\infty} a_{i,n} \text{ existiert für alle } i \in \mathbb{N},$$
$$|a_{i,n}| \leq b_i \,\forall i, n \in \mathbb{N} \text{ mit } \sum_{i\in\mathbb{N}} b_i < \infty,$$

erfüllt, die Summation mit dem Grenzwert vertauschbar ist, d. h. $\lim_{n\to\infty} \sum_{i\in\mathbb{N}} a_{i,n} = \sum_{i\in\mathbb{N}} a_i$ gilt. ◇

Wir ziehen noch zwei Folgerungen aus dem Satz von Lebesgue, deren Beweise Übungsaufgaben sind:

Korollar 6.8.8 (Stetigkeitslemma). *Es sei $f \colon \Omega \times \mathbb{R} \to \mathbb{R}$ eine Funktion und $x_0 \in \mathbb{R}$, so dass gelten:*

(a) Für jedes $x \in \mathbb{R}$ ist $f(\cdot, x) \in \mathcal{L}^1(\mu)$,

(b) für fast alle $\omega \in \Omega$ ist $f(\omega, \cdot)$ stetig in x_0,

(c) es gibt eine Umgebung $\mathcal{U}$ von x_0, so dass die Abbildung $\sup_{x\in\mathcal{U}} |f(\cdot, x)|$ in $\mathcal{L}^1(\mu)$ liegt.

Dann ist die Abbildung $\int f(\omega, \cdot) \,\mu(\mathrm{d}\omega)$ stetig in x_0.

Korollar 6.8.9 (Differenziationslemma). *Es sei $I \subset \mathbb{R}$ ein nichttriviales Intervall und $f\colon \Omega \times I \to \mathbb{R}$ eine Abbildung, so dass gelten:*

(a) Für jedes $x \in I$ ist $f(\cdot, x) \in \mathcal{L}^1(\mu)$,

(b) für fast alle $\omega \in \Omega$ ist $f(\omega, \cdot)$ differenzierbar (wir bezeichnen die Ableitung mit f'),

(c) die Abbildung $\sup_{x\in I} |f'(\cdot, x)|$ liegt im $\mathcal{L}^1(\mu)$.

Dann liegt $f'(\cdot, x)$ für jedes $x \in I$ in $\mathcal{L}^1(\mu)$, und die Funktion $F(x) = \int f(\omega, x)\,\mu(\mathrm{d}\omega)$ ist differenzierbar mit $F'(x) = \int f'(\omega, x)\,\mu(\mathrm{d}\omega)$.

Bemerkung 6.8.10 (Momenten erzeugende Funktion II). Wir betrachten wieder die Momenten erzeugende Funktion $M(t) = \int e^{tx}\,\mu(\mathrm{d}x)$ eines Maßes μ auf $(\mathbb{R}, \mathcal{B})$ wie in Bemerkung 6.7.6. Man sieht leicht, dass der Definitionsbereich $I = \{t \in \mathbb{R} : M(t) < \infty\}$ ein Intervall ist. Als eine Anwendung des Differenziationslemmas sieht man, dass M im Innern von I beliebig oft differenzierbar ist mit $M^{(k)}(t) = \int x^k e^{tx}\,\mu(\mathrm{d}x)$ für jedes $k \in \mathbb{N}_0$ und $t \in I^\circ$. Wenn insbesondere $0 \in I^\circ$, so existieren alle *Momente* $\int x^k\,\mu(\mathrm{d}x)$, $k \in \mathbb{N}_0$, von μ. Mit Hilfe der zweiten Ableitung erhält man als Übungsaufgabe insbesondere einen zweiten Beweis für die Konvexität der Abbildung $\log M$, siehe Bemerkung 6.7.6. ◇

6.9 Produkte messbarer Räume

In der Wahrscheinlichkeitstheorie möchte man auch gerne ganze *Tupel* oder gar unendliche *Folgen* von zufälligen Experimenten (sogenannte stochastische Prozesse) behandeln und in die allgemeine Maßtheorie einbetten. Das natürlichste mathematische Konstrukt hierfür ist das *Produkt* der einzelnen Mengen, die die einzelnen Experimente beschreiben. In diesem Abschnitt führen wir die natürlichste σ-Algebra auf diesem Produkt ein, und zwar die kleinste, die es zulässt, dass die einzelnen Experimente in dem Produktraum genauso gut beschrieben werden können wie für sich alleine.

Es sei I eine beliebige Indexmenge, und für jedes $i \in I$ sei $(E_i, \mathcal{E}_i)$ ein messbarer Raum. Zur besseren Anschauung stelle man sich $I = \mathbb{N}$ vor, aber das Folgende ist nicht beschränkt auf abzählbare Indexmengen. Auf der Produktmenge $\prod_{i\in I} E_i$ wollen wir eine natürliche σ-Algebra definieren. Dazu betrachten wir die Mengen der Form

$$\mathcal{A}^{(j)}(A_j) = \Big\{(x_i)_{i\in I} \in \prod_{i\in I} E_i : x_j \in A_j\Big\} \qquad \text{mit } A_j \in \mathcal{E}_j, j \in I. \tag{6.9.1}$$

In Worten: $\mathcal{A}^{(j)}(A_j)$ ist die Menge derjenigen Elemente im Produktraum, deren j-te Komponente eine Bedingung erfüllt, nämlich dass sie in A_j liegt. Diese Mengen wählen wir als Erzeuger der gesuchten σ-Algebra:

Definition 6.9.1 (Produkt-$\boldsymbol{\sigma}$-Algebra). *Die σ-Algebra auf dem Produkt $\prod_{i\in I} E_i$, die von den Mengen $\mathcal{A}^{(j)}(A_j)$ mit $A_j \in \mathcal{E}_j$ und $j \in I$ erzeugt wird, heißt die* Produkt-σ-Algebra *auf $\prod_{i\in I} E_i$ und wird mit $\bigotimes_{i\in I} \mathcal{E}_i$ bezeichnet. Den messbaren Raum $(\prod_{i\in I} E_i, \bigotimes_{i\in I} \mathcal{E}_i)$ nennt man den* Produktraum *der $(E_i, \mathcal{E}_i)$. Falls alle $(E_i, \mathcal{E}_i)$ gleich sind, also $(E_i, \mathcal{E}_i) = (E, \mathcal{E})$, so schreibt man auch $(E^I, \mathcal{E}^{\otimes I})$ für den Produktraum. Im Fall $I = \{1, \dots, n\}$ schreiben wir auch $\mathcal{E}^{\otimes n}$ statt $\mathcal{E}^{\otimes I}$.*

Bemerkung 6.9.2. Die Produkt-σ-Algebra ist die kleinste σ-Algebra auf dem Produkt, so dass die Projektionen

$$\pi_j\colon \prod_{i\in I} E_i \to E_j, \qquad \pi_j\big((x_i)_{i\in I}\big) = x_j,$$

messbar sind. Für die Mengen in (6.9.1) gilt $\mathcal{A}^{(j)}(A_j) = \pi_j^{-1}(A_j)$. Also haben wir auch auch $\bigotimes_{i\in I}\mathcal{E}_i = \bigvee_{i\in I}\pi_i^{-1}(\mathcal{E}_i) = \bigvee_{i\in I}\sigma(\pi_i)$; siehe Bemerkung 6.4.2. $\diamond$

Bemerkung 6.9.3 (durchschnittstabile Erzeuger des Produktraums)**.** Die Mengen der Form (6.9.1) bilden kein *durchschnittstabiles* Erzeugendensystem. Insbesondere ist der Eindeutigkeitssatz (Satz 6.2.8) nicht auf dieses System anwendbar. Daher benutzt man oft lieber das Erzeugendensystem, das aus den *endlichen Durchschnitten* von Mengen der Form (6.9.1) besteht. Genauer: Für eine endliche Teilmenge J von I und Mengen $A_j \in \mathcal{E}_j$ für $j \in J$ setzen wir

$$\mathcal{A}^{(J)}((A_j)_{j\in J}) = \bigcap_{j\in J}\mathcal{A}^{(j)}(A_j) = \{(x_i)_{i\in I}\colon x_j \in A_j \text{ für alle } j \in J\}. \qquad (6.9.2)$$

Das Mengensystem $\mathcal{D}$ aller solcher $\mathcal{A}^{(J)}((A_j)_{j\in J})$ ist offensichtlich ein durchschnittstabiles Erzeugendensystem der Produkt-σ-Algebra auf $\prod_{i\in I} E_i$. Die Mengen $\mathcal{A}^{(J)}((A_j)_{j\in J})$ nennt man auch *endlichdimensionale Zylinder*. Auf $\mathcal{D}$ kann man Satz 6.2.8 problemlos anwenden. Im Spezialfall von Folgenräumen (also $I = \mathbb{N}_0$) und $J = \{0, 1, \dots, n\}$ hat $\mathcal{A}^{(J)}$ die Gestalt

$$\mathcal{A}^{(J)}((A_j)_{j\in J}) = A_0 \times A_1 \times \cdots \times A_n \times \prod_{i=n+1}^{\infty} E_i.$$

$\diamond$

Beispiel 6.9.4. Als eine Übungsaufgabe mache man sich klar, dass die Borel-σ-Algebra $\mathcal{B}_n$ auf $\mathbb{R}^n$ gleich dem n-fachen Produkt der Borel-σ-Algebra $\mathcal{B}$ auf $\mathbb{R}$ ist, also $\mathcal{B}_n = \mathcal{B}^{\otimes n}$ (siehe Beispiel 6.1.5). $\diamond$

Beispiel 6.9.5 (Unendliche Folgenräume)**.** Für die mathematische Modellierung eines unendlich oft wiederholten Versuchs benötigt man eine geeignete σ-Algebra auf dem unendlichen Folgenraum $E^{\mathbb{N}}$, wobei $(E, \mathcal{E})$ ein gegebener messbarer Raum ist, der einen einzelnen Versuch beschreibt. Eines der einfachsten und meist bemühten Beispiele ist der Münzwurf, wo man $E = \{K, Z\}$ setzt und die Potenzmenge $\mathcal{E} = \mathcal{P}(\{K, Z\})$ wählt. Die Produkt-σ-Algebra $(\mathcal{P}(\{K, Z\}))^{\otimes\mathbb{N}}$ besitzt das durchschnittstabile Erzeugendensystem, das aus den Zylindermengen$\{(x_n)_{n\in\mathbb{N}}\colon x_1 \in A_1, \dots, x_m \in A_m\}$ mit $m \in \mathbb{N}$ und $A_1, \dots, A_m \in \mathcal{P}(\{K, Z\})$ besteht. Diese Mengen legen Bedingungen fest, die die ersten m Komponenten der betrachteten Folge betreffen. Ein Maß auf dem Produktraum $(\{K, Z\}^{\mathbb{N}}, (\mathcal{P}(\{K, Z\}))^{\otimes\mathbb{N}})$ ist eindeutig festgelegt, wenn es auf allen diesen Zylindermengen festgelegt ist.

Man beachte, dass z. B. die Mengen 'Fast alle Würfe zeigen Kopf' oder 'Unendliche viele Würfe sind Kopf' in der Produkt-σ-Algebra enthalten sind, denn es gilt z. B.

$$\big\{(x_n)_{n\in\mathbb{N}} \in \{K, Z\}^{\mathbb{N}}\colon x_n = K \text{ für fast alle } n \in \mathbb{N}\big\} = \bigcup_{M\in\mathbb{N}}\ \bigcap_{N\in\mathbb{N}\colon N\geq M}\big\{(x_n)_{n\in\mathbb{N}}\colon x_N = K\big\},$$

und die Menge auf der rechten Seite ist messbar als abzählbare Vereinigung eines abzählbaren Schnitts des Urbilds von $\{K\}$ unter der Projektion π_N.

Die Produkt-σ-Algebra $(\mathcal{P}(\{K,Z\}))^{\otimes\mathbb{N}}$ ist tatsächlich echt kleiner als die Menge $\mathcal{P}(\{K,Z\}^{\mathbb{N}})$ *aller* Teilmengen des Folgenraums. Dies ist nicht leicht einzusehen, man muss das Beispiel von Vitali (siehe Beispiel 6.1.11) geeignet modifizieren. Man kann zeigen, dass die Produkt-σ-Algebra von der Mächtigkeit des Kontinuums ist, d. h. dass es eine bijektive Abbildung $(\mathcal{P}(\{K,Z\}))^{\otimes\mathbb{N}} \to \mathbb{R}$ gibt, während die Potenzmenge $\mathcal{P}(\{K,Z\}^{\mathbb{N}})$ von der Mächtigkeit von $\mathcal{P}(\mathbb{R})$ ist. $\diamond$

Beispiel 6.9.6 (Räume reeller Zahlenfolgen)**.** Die Produkt-σ-Algebra $\mathcal{B}^{\otimes\mathbb{N}}$ auf der Menge aller Zahlenfolgen $\mathbb{R}^{\mathbb{N}}$ enthält unter Anderem auch die Menge aller beschränkten Folgen, die Menge aller Nullfolgen und die Menge aller konvergenten Folgen (Übungsaufgaben). Die Messbarkeit der Menge aller beschränkten Folgen sieht man so:

$$\big\{(x_n)_{n\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}} \colon \exists M \in \mathbb{N}\,\forall N \in \mathbb{N} \colon |x_N| \leq M\big\} = \bigcup_{M\in\mathbb{N}} \bigcap_{N\in\mathbb{N}} \pi_N^{-1}([-M,M]),$$

und die Menge auf der rechten Seite ist eine abzählbare Vereinigung eines abzählbares Schnittes von Urbildern messbarer Mengen unter Projektionen. $\diamond$

6.10 Produktmaße

Auf der Produkt-σ-Algebra möchten wir nun auch Maße definieren, die in geeignetem Sinn als das Produkt von Maßen auf den einzelnen Faktoren aufgefasst werden können. Dies tun wir in diesem Abschnitt für nur endliche viele Faktoren, genauer gesagt, zunächst nur für zwei.[9] Es seien also zwei messbare Räume $(\Omega_1,\mathcal{F}_1)$ und $(\Omega_2,\mathcal{F}_2)$ gegeben. Wir erinnern uns, dass die Produkt-σ-Algebra $\mathcal{F}_1 \otimes \mathcal{F}_2$ auf $\Omega_1 \times \Omega_2$ durch das durchschnittstabile Mengensystem

$$\mathcal{C} = \{A_1 \times A_2 \colon A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2\} \tag{6.10.1}$$

erzeugt wird. Für Mengen $A \subset \Omega_1 \times \Omega_2$ nennen wir die Mengen

$$\begin{aligned} A^{(2)}_{\omega_2} &= \{\omega_1 \in \Omega_1 \colon (\omega_1,\omega_2) \in A\}, \qquad && \omega_2 \in \Omega_2, \\ A^{(1)}_{\omega_1} &= \{\omega_2 \in \Omega_1 \colon (\omega_1,\omega_2) \in A\}, \qquad && \omega_1 \in \Omega_1, \end{aligned}$$

die *Schnitte* von A. Allgemeiner nennen wir für eine Abbildung $f\colon \Omega_1 \times \Omega_2 \to \mathbb{R}$ die beiden Abbildungen $f^{(2)}_{\omega_2}(\cdot) = f(\cdot,\omega_2)$ und $f^{(1)}_{\omega_1}(\cdot) = f(\omega_1,\cdot)$ die *Schnitte* von f. Man sieht leicht, dass $1\!\!1_{A^{(2)}_{\omega_2}} = (1\!\!1_A)^{(2)}_{\omega_2}$ und eine analoge Formel für $A^{(1)}_{\omega_1}$ gelten.

Satz 6.10.1. *Ist $f\colon \Omega_1 \times \Omega_2 \to \mathbb{R}$ eine messbare Funktion, so sind alle Schnitte messbar bezüglich der jeweiligen σ-Algebra, d. h. für jedes $\omega_2 \in \Omega_2$ ist $f^{(2)}_{\omega_2}$ $\mathcal{F}_1$-messbar, und für jedes $\omega_1 \in \Omega_1$ ist $f^{(1)}_{\omega_1}$ $\mathcal{F}_2$-messbar.*

Beweis. Wir zeigen nur die erste Aussage, und zwar zunächst nur für alle Funktionen der Form $f = 1\!\!1_A$ mit $A \in \mathcal{F}_1 \otimes \mathcal{F}_2$. (Der Rest des Beweises folgt dann wie gewohnt etwa mit Hilfe von Satz 6.4.11.) Betrachten wir das Mengensystem

$$\mathcal{G} = \{A \in \mathcal{F}_1 \otimes \mathcal{F}_2 \colon A^{(2)}_{\omega_2} \in \mathcal{F}_1 \text{ für alle } \omega_2 \in \Omega_2\}.$$

[9]Siehe Abschnitt 10 für abzählbar unendlich viele Faktoren. Dort ist das Hauptmittel der Satz von Carathéodory, hier konstruieren wir auf elementare Weise.

Wir möchten gerne zeigen, dass $\mathcal{G} = \mathcal{F}_1 \otimes \mathcal{F}_2$, denn das beendet den Beweis. Zunächst sehen wir, dass $\mathcal{C} \subset \mathcal{G}$ (siehe (6.10.1)), denn

$$(A_1 \times A_2)^{(2)}_{\omega_2} = \begin{cases} A_1, & \text{falls } \omega_2 \in A_2, \\ \emptyset & \text{sonst.} \end{cases} \tag{6.10.2}$$

Wegen $(A^{\mathrm{c}})^{(2)}_{\omega_2} = (A^{(2)}_{\omega_2})^{\mathrm{c}}$ für alle $A \subset \Omega_1 \times \Omega_2$ und $(\bigcup_{n\in\mathbb{N}} B_n)^{(2)}_{\omega_2} = \bigcup_{n\in\mathbb{N}} (B_n)^{(2)}_{\omega_2}$ für alle Folgen von Mengen $B_1, B_2, B_3, \dots \subset \Omega_1 \times \Omega_2$ ist $\mathcal{G}$ eine σ-Algebra. Also gilt $\mathcal{G} = \mathcal{F}_1 \otimes \mathcal{F}_2$. □

Nun können wir Produktmaße auf $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2)$ definieren. Dazu legen wir einen Kandidaten zunächst auf dem durchschnittstabilen Erzeugendensystem $\mathcal{C}$ aus (6.10.1) mit einer offensichtlichen Formel fest und erweitern ihn dann auf die Produkt-σ-Algebra:

Satz 6.10.2 (Produktmaße). *Seien $(\Omega_1, \mathcal{F}_1, \mu_1)$ und $(\Omega_2, \mathcal{F}_2, \mu_2)$ zwei σ-endliche Maßräume. Dann gibt es genau ein Maß μ auf dem Produktraum $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2)$, so dass $\mu(A_1 \times A_2) = \mu_1(A_1)\mu_2(A_2)$ für alle $A_1 \in \mathcal{F}_1$ und $A_2 \in \mathcal{F}_2$ gilt. Das Maß μ heißt das* Produktmaß *von μ_1 und μ_2 und wird mit $\mu = \mu_1 \otimes \mu_2$ bezeichnet.*

Bemerkung 6.10.3 (Produkte unendlich vieler Maße)**.** Das Produktmaß $\mu = \otimes_{i\in I}\mu_i$ von *un*endlich vielen Maßen μ_i mit $i \in I$ wird analog definiert (wenn es existiert) durch die Forderung

$$\mu\Big(\mathcal{A}^{(J)}((A_j)_{j\in J})\Big) = \prod_{i\in J} \mu_i(A_i), \qquad J \subset I \text{ endlich}, A_i \in \mathcal{E}_i \text{ für } i \in J;$$

siehe (6.9.2). Die Existenz von μ werden wir in Abschnitt 10 mit Hilfe des Satzes 6.2.7 von Carathéodory zumindest im abzählbaren Fall unter gewissen weiteren Einschränkungen beweisen; siehe den Satz 10.1.14 von Ionescu Tulcea. Hier geben wir hier einen konstruktiven expliziten Beweis für endliche Produkte. ◇

Beweis von Satz 6.10.2. Die Eindeutigkeit ergibt sich sofort aus dem Eindeutigkeitssatz 6.2.8. Die Existenz zeigen wir nun konstruktiv mit Hilfe von Satz 6.10.1: Sei $A \in \mathcal{F}_1 \otimes \mathcal{F}_2$. Nach Satz 6.10.1 ist $A^{(2)}_{\omega_2} \in \mathcal{F}_1$ für alle $\omega_2 \in \Omega_2$, also dürfen wir $\mu_1(A^{(2)}_{\omega_2})$ bilden. Nun zeigen wir, dass die Abbildung $\omega_2 \mapsto \mu_1(A^{(2)}_{\omega_2})$ eine $\mathcal{F}_2$-messbare numerische Funktion ist. Da μ_1 σ-endlich ist, gibt es eine Folge $(B_n)_{n\in\mathbb{N}}$ von messbaren Teilmengen von Ω_1 mit $B_n \uparrow \Omega_1$ und $\mu_1(B_n) < \infty$ für jedes $n \in \mathbb{N}$. Wegen $\mu_1(A^{(2)}_{\omega_2}) = \lim_{n\to\infty} \mu_1(A^{(2)}_{\omega_2} \cap B_n)$ genügt es daher zu zeigen, dass für jedes $n \in \mathbb{N}$ die Abbildung $\omega_2 \mapsto \mu_1(A^{(2)}_{\omega_2} \cap B_n)$ eine $\mathcal{F}_2$-messbare Funktion ist. Man kann elementar zeigen, dass das Mengensystem

$$\{A \in \mathcal{F}_1 \otimes \mathcal{F}_2 \colon \omega_2 \mapsto \mu_1(A^{(2)}_{\omega_2} \cap B_n) \text{ ist } \mathcal{F}_2\text{-messbar}\}$$

ein Dynkin-System ist, das $\mathcal{C}$ enthält. Also ist es gleich $\mathcal{F}_1 \otimes \mathcal{F}_2$. Insgesamt haben wir also die Messbarkeit der Abbildung $\omega_2 \mapsto \mu_1(A^{(2)}_{\omega_2})$ gezeigt.

Nun können wir definieren:

$$\mu(A) = \int_{\Omega_2} \mu_1(A^{(2)}_{\omega_2})\, \mu_2(\mathrm{d}\omega_2), \qquad A \in \mathcal{F}_1 \otimes \mathcal{F}_2.$$

Eine Anwendung von Korollar 6.8.3 zeigt, dass μ ein Maß ist. Und mit Hilfe von (6.10.2) sieht man leicht, dass μ auf $\mathcal{C}$ die gewünschte Gestalt besitzt. □

Um problemlos bezüglich $\mu_1 \otimes \mu_2$ integrieren zu können, brauchen wir noch zwei Vertauschungssätze. Es seien $(\Omega_1, \mathcal{F}_1, \mu_1)$ und $(\Omega_2, \mathcal{F}_2, \mu_2)$ zwei σ-endliche Maßräume.

Satz 6.10.4 (Satz von Tonelli). *Es sei $f\colon \Omega_1 \times \Omega_2 \to [0, \infty]$ eine messbare numerische Funktion. Dann sind die Abbildungen $\omega_1 \mapsto \int f^{(1)}_{\omega_1}\,\mathrm{d}\mu_2$ und $\omega_2 \mapsto \int f^{(2)}_{\omega_2}\,\mathrm{d}\mu_1$ messbar, und es gilt*

$$\int f\,\mathrm{d}(\mu_1 \otimes \mu_2) = \int_{\Omega_1} \Big(\int_{\Omega_2} f^{(1)}_{\omega_1}(\omega_2)\,\mu_2(\mathrm{d}\omega_2)\Big)\mu_1(\mathrm{d}\omega_1) = \int_{\Omega_2} \Big(\int_{\Omega_1} f^{(2)}_{\omega_2}(\omega_1)\,\mu_1(\mathrm{d}\omega_1)\Big)\mu_2(\mathrm{d}\omega_2).$$

Beweis. Diese Aussage ist nach Konstruktion richtig für alle Funktionen der Form $f = \mathbb{1}_A$ mit $A \in \mathcal{F}_1 \otimes \mathcal{F}_2$. Die allgemeine Aussage folgt wie gewohnt mit Hilfe von Satz 6.4.11. □

Mit Hilfe der üblichen Zerlegung in Positiv- und Negativteil erweitert man den Satz von Tonelli leicht auf alle integrierbare Funktionen $\Omega_1 \times \Omega_2 \to \mathbb{R}$ und erhält den berühmten Satz von Fubini, der vereinfacht sich so ausdrücken lässt: Falls $f \in \mathcal{L}^1(\mu_1 \otimes \mu_2)$, so kann man die Integrationsreihenfolge vertauschen:

$$\int f\,\mathrm{d}(\mu_1 \otimes \mu_2) = \int \Big(\int f(\omega_1, \omega_2)\,\mu_1(\mathrm{d}\omega_1)\Big)\mu_2(\mathrm{d}\omega_2) = \int \Big(\int f(\omega_1, \omega_2)\,\mu_2(\mathrm{d}\omega_2)\Big)\mu_1(\mathrm{d}\omega_1).$$

Satz 6.10.5 (Satz von Fubini). *Sei $f \in \mathcal{L}^1(\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2, \mu_1 \otimes \mu_2)$. Dann liegt für μ_1-fast alle $\omega_1 \in \Omega_1$ die Funktion $f^{(1)}_{\omega_1}$ in $\mathcal{L}^1(\Omega_2, \mathcal{F}_2, \mu_2)$, und für μ_2-fast alle $\omega_2 \in \Omega_2$ liegt die Funktion $f^{(2)}_{\omega_2}$ in $\mathcal{L}^1(\Omega_1, \mathcal{F}_1, \mu_1)$. Setzt man*

$$f_1(\omega_1) = \begin{cases} \int_{\Omega_2} f^{(1)}_{\omega_1}(\omega_2)\,\mu_2(\mathrm{d}\omega_2), & \textit{falls } f^{(1)}_{\omega_1} \in \mathcal{L}^1, \\ 0 & \textit{sonst,} \end{cases}$$

und $f_2(\omega_2)$ analog, so gilt

$$\int f\,\mathrm{d}(\mu_1 \otimes \mu_2) = \int f_1(\omega_1)\,\mu_1(\mathrm{d}\omega_1) = \int f_2(\omega_2)\,\mu_2(\mathrm{d}\omega_2).$$

Bemerkung 6.10.6 (endlich viele Faktoren). Die obigen Betrachtungen können problemlos auf endlich viele Faktoren ausgedehnt werden, wobei man sich klar machen muss, dass die Produkte beliebig umgeklammert werden dürfen. Wenn also $(\Omega_i, \mathcal{F}_i, \mu_i)$ für jedes $i \in \{1, \dots, n\}$ ein σ-endlicher Maßraum ist, so definiert man rekursiv den Produktraum $\bigotimes_{i=1}^n (\Omega_i, \mathcal{F}_i, \mu_i) = (\prod_{i=1}^n \Omega_i, \bigotimes_{i=1}^n \mathcal{F}_i, \bigotimes_{i=1}^n \mu_i)$ und meint damit das Produkt $\prod_{i=1}^n \Omega_i = (\dots((\Omega_1 \times \Omega_2) \times \Omega_3) \times \dots \times \Omega_n)$ und analoge Formeln für die Produkte der σ-Algebren und die der Maße. Es ist dann klar, dass $\prod_{i=1}^n \Omega_i = (\Omega_1 \times \dots \times \Omega_k) \times (\Omega_{k+1} \times \dots \times \Omega_n)$ für jedes $k \in \{1, \dots, n\}$ gilt und analoge Formeln für die Produkte der σ-Algebren und die der Maße. Falls alle Faktoren $(\Omega_i, \mathcal{F}_i, \mu_i) = (\Omega, \mathcal{F}, \mu)$ übereinstimmen, schreiben wir natürlich Ω^n sowie $\mathcal{F}^{\otimes n}$ und $\mu^{\otimes n}$ oder auch Ω^I sowie $\mathcal{F}^{\otimes I}$ und $\mu^{\otimes I}$, falls $I = \{1, \dots, n\}$. ◇

Beispiel 6.10.7 (***n***-dimensionales Lebesgue-Maß)**.** Im wichtigen Spezialfall, wo (siehe Beispiel 6.1.5) $(\Omega_i, \mathcal{F}_i, \mu_i) = (\mathbb{R}, \mathcal{B}, \lambda)$ für $i \in \{1, \dots, n\}$, identifizieren wir das n-fache Produkt mit $(\mathbb{R}^n, \mathcal{B}_n, \lambda^{\otimes n})$ (siehe auch Beispiel 6.9.4) und führen das Lebesgue-Maß auf dem $\mathbb{R}^n$ ein als das n-fache Produkt des eindimensionalen Lebesgue-Maßes. Es ist klar, dass $\lambda^{\otimes n}$ übereinstimmt mit den eindeutig bestimmten Maß λ_n auf $\mathcal{B}_n$, das jedem Quader $Q = \prod_{i=1}^n [a_i, b_i]$ seinen elementargeometrischen Inhalt $\lambda_n(Q) = \prod_{i=1}^n (b_i - a_i)$ zuweist. ◇

Beispiel 6.10.8 (Bernoulli-Maß)**.** Für eine endliche Menge I heißt das Maß $(p\delta_1 + (1-p)\delta_0)^{\otimes I}$ auf $\{0, 1\}^I$ mit der Potenzmenge (siehe Beispiel 6.1.10) das *Bernoulli-Maß* zum Parameter $p \in (0, 1)$. Es beschreibt eine endliche Folge von unabhängigen Versuchen, die jeweils mit Wahrscheinlichkeit p Erfolg haben und sonst nicht.

Wenn zum Beispiel I die Menge aller Kanten zwischen benachbarten Punkten in dem Gitter $\mathbb{Z}^d \cap [-n, n]^d$ bezeichnet, dann kann $(p\delta_1 + (1-p)\delta_0)^{\otimes I}$ als ein stochastisches Modell für einen porösen Stein dienen, den man sich als in regelmäßige Kammern unterteilt vorstellt, die jeweils mit ihrer Nachbarkammer verbunden sind durch eine Wand, die mit Wahrscheinlichkeit p durchlässig ist und sonst nicht. Dies ist der Ausgangspunkt des Forschungszweiges der *Perkolation*, die untersucht, ob solche auf zufällige Weise poröse Steine eine Flüssigkeit durchsickern lassen oder nicht. Am interessantesten wird diese Frage natürlich, wenn wir statt eine endliche Box $\mathbb{Z}^d \cap [-n, n]^d$ den gesamten Raum $\mathbb{Z}^d$ betrachten, aber *un*endliche Produktmaße werden wir erst später im Zusammenhang mit dem Satz von Ionescu Tulcea behandeln. ◇

Kapitel 7

Die allgemeine Wahrscheinlichkeitstheorie

In diesem Kapitel wenden wir uns der allgemeinen Wahrscheinlichkeitstheorie auf Basis der soeben vorgestellten Maßtheorie zu. Wir werden Grundlagen vorstellen, die unverzichtbar sind für die tiefere Beschäftigung mit der Wahrscheinlichkeitstheorie. Zu Beginn dieses Kapitels stellen wir den Jargon der Theorie vor und sammeln die wichtigsten Werkzeuge, dann behandeln wir die Unabhängigkeit von Ereignissen und von Zufallsgrößen, und zum Schluss führen wir bedingte Erwartungswerte in großer Allgemeinheit ein.

7.1 Grundlegende Begriffe

Es sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum, also ein Maßraum mit $\mathbb{P}(\Omega) = 1$. In diesem Spezialfall eines Maßraums haben sich etliche spezielle Sprechweisen eingebürgert. So nennt man $\mathbb{P}$ auch eine *Verteilung*, die Elemente der σ-Algebra $\mathcal{F}$ werden meist *Ereignisse* genannt und $\mathbb{P}(A)$ die *Wahrscheinlichkeit* des Ereignisses $A \in \mathcal{F}$. Das Komplement A^{c} steht für 'A tritt nicht ein', und die Vereinigung $A \cup B$ für 'A oder B trifft ein', der Schnitt $A \cap B$ für 'A und B treffen ein'.

Für einen messbaren Raum $(E, \mathcal{E})$ werden messbare Abbildungen $\Omega \to E$ *E-wertige Zufallsgrößen* genannt, im Falle $(E, \mathcal{E}) = (\mathbb{R}, \mathcal{B})$ einfach nur *Zufallsgrößen* und im Falle $(E, \mathcal{E}) = (\mathbb{R}^n, \mathcal{B}_n)$ *Zufallsvektoren*. Wir werden Zufallsgrößen nun meist mit X und Y usw. bezeichnen statt mit f und g usw. Auf Grund von Bemerkung 6.9.4 sind Abbildungen $X_1, \dots, X_n \colon \Omega \to \mathbb{R}$ genau dann Zufallsgrößen, wenn die vektorwertige Abbildung $X = (X_1, \dots, X_n) \colon \Omega \to \mathbb{R}^n$ ein Zufallsvektor ist. Die von einer Zufallsgröße $X \colon \Omega \to E$ erzeugte σ-Algebra wird wie zuvor mit

$$X^{-1}(\mathcal{E}) = \{X^{-1}(A) \colon A \in \mathcal{E}\} = \sigma(X)$$

bezeichnet. Falls für jedes i aus einer Indexmenge I eine Zufallsgröße $X_i \colon \Omega \to E_i$ in einen messbaren Raum $(E_i, \mathcal{E}_i)$ definiert ist, so heißt

$$\bigvee_{i \in I} X_i^{-1}(\mathcal{E}_i) = \sigma\Big(\bigcup_{i \in I} X_i^{-1}(\mathcal{E}_i)\Big)$$

die von den X_i auf Ω erzeugte σ-Algebra. Dies ist die kleinste Teil-σ-Algebra von $\mathcal{F}$, so dass jedes X_i messbar ist.

Die *Verteilung* einer Zufallsgröße $X\colon \Omega \to E$ ist per definitionem das Bildmaß $\mathbb{P} \circ X^{-1}$ auf $(E, \mathcal{E})$ (siehe Abschnitt 6.6). Es ordnet jedem $A \in \mathcal{E}$ die Wahrscheinlichkeit $\mathbb{P} \circ X^{-1}(A) = \mathbb{P}(X \in A)$ des Ereignisses $\{X \in A\}$ zu. Man schreibt auch manchmal $\mathcal{L}(X)$ statt $\mathbb{P} \circ X^{-1}$, wenn man den Raum $(\Omega, \mathcal{F}, \mathbb{P})$ nicht betonen oder gar nicht erst erwähnen möchte. Man beachte, dass zwei $(E, \mathcal{E})$-wertige Zufallsgrößen X und Y nicht auf dem selben Wahrscheinlichkeitsraum definiert sein müssen, um die selbe Verteilung zu haben. Für $\mathbb{R}$-wertige Zufallsgrößen X nennt man die Abbildung $t \mapsto \mathbb{P}(X \leq t)$ die *Verteilungsfunktion* von X. Sie ist monoton steigend und rechtsstetig mit Grenzwerten 0 und 1 bei $-\infty$ bzw. ∞.

Für eine integrierbare Zufallsgröße $X\colon \Omega \to \mathbb{R}$ schreibt man $\mathbb{E}(X)$ statt $\int_\Omega X(\omega)\,\mathbb{P}(\mathrm{d}\omega)$ und nennt $\mathbb{E}(X)$ den *Erwartungswert* von X. Der Erwartungswert von X existiert also genau dann, wenn $\mathbb{E}(|X|) < \infty$. Für eine $(E, \mathcal{E})$-wertige Zufallsgröße X und eine messbare Funktion $f\colon E \to \mathbb{R}$ liegt $f(X)$ (dies schreiben wir meist statt $f \circ X$) genau dann im $\mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$, wenn f im $\mathcal{L}^1(E, \mathcal{E}, \mathbb{P} \circ X^{-1})$ liegt, und dann gilt $\mathbb{E}(f(X)) = \int_E f \,\mathrm{d}\mathbb{P} \circ X^{-1}$. Der Erwartungswert ist eine monotone und lineare Abbildung, d. h. für Zufallsgrößen X und $Y \in \mathcal{L}^1(\mathbb{P})$ gilt $X \leq Y \Longrightarrow \mathbb{E}(X) \leq \mathbb{E}(Y)$, und für $a, b \in \mathbb{R}$ ist auch $aX + bY \in \mathcal{L}^1(\mathbb{P})$, und $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$.

Die beiden Spezialfälle, dass (1) X diskret ist und (2) X eine Dichte besitzt, führen auf folgende Formeln. Falls die reellwertige Zufallsgröße X nur die Werte $x_1, x_2, x_3, \dots \in \mathbb{R}$ mit positiver Wahrscheinlichkeit annimmt und auf $\{x_n \colon n \in \mathbb{N}\}$ konzentriert ist[1], so ist also $\mathbb{P} \circ X^{-1} = \sum_{n \in \mathbb{N}} \mathbb{P}(X = x_n)\delta_{x_n}$, und wir haben für jede Abbildung $f\colon \mathbb{R} \to \mathbb{R}$:

$$\mathbb{E}(f(X)) = \sum_{n \in \mathbb{N}} f(x_n)\,\mathbb{P}(X = x_n),$$

falls diese Reihe absolut konvergiert. Falls X eine (Lebesgue-)Dichte $\varphi\colon \mathbb{R} \to [0, \infty)$ besitzt, also wenn $\mathbb{P}(X \in A) = \int_A \varphi(x)\,\lambda(\mathrm{d}x)$ für alle messbaren Mengen $A \in \mathcal{B}$ gilt, so gilt für jede messbare Abbildung $f\colon \mathbb{R} \to \mathbb{R}$:

$$\mathbb{E}(f(X)) = \int_{\mathbb{R}} f(x)\varphi(x)\,\lambda(\mathrm{d}x),$$

falls das Integral absolut konvergiert.

Eine wichtige Maßzahl einer Zufallsgröße $X \in \mathcal{L}^1(\mathbb{P})$, die ihre mittlere quadratische Abweichung von ihrem Erwartungswert angibt, ist die *Varianz* , die definiert wird als

$$\mathbb{V}(X) = \mathbb{E}\big((X - \mathbb{E}(X))^2\big) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 \in [0, \infty].$$

(Die zweite Gleichung nennt man die *Steiner'sche Formel.*) Die Varianz von $X \in \mathcal{L}^1(\mathbb{P})$ ist also immer definiert, könnte aber gleich ∞ sein. Sie ist genau dann endlich, wenn $X \in \mathcal{L}^2(\mathbb{P})$, und sie ist genau dann Null, wenn $X = \mathbb{E}(X)$ $\mathbb{P}$-fast sicher.

Wir erinnern in Kurzform an die wichtigsten Ungleichungen, spezialisiert auf einen Wahrscheinlichkeitsraum:

Jensen: φ konvex $\Longrightarrow$ $\mathbb{E}(\varphi(X)) \geq \varphi(\mathbb{E}(X))$,

siehe Satz 6.7.2,

Hölder: $\dfrac{1}{p} + \dfrac{1}{q} = 1$ $\Longrightarrow$ $\mathbb{E}(XY) \leq \mathbb{E}(X^p)^{1/p}\mathbb{E}(Y^q)^{1/q}$,

[1] Wir sagen, dass eine Zufallsgröße X auf einer messbaren Menge A *konzentriert* ist, falls $\mathbb{P}(X \in A) = 1$.

siehe Satz 6.7.5,

Cauchy-Schwarz: $\mathbb{E}(XY) \leq \mathbb{E}(X^2)^{1/2}\mathbb{E}(Y^2)^{1/2}$,

siehe Korollar 3.5.6,

Markov: $\varphi\colon (0,\infty) \to (0,\infty)$ wachsend, $\varepsilon > 0 \quad \Longrightarrow \quad \mathbb{P}(|X| > \varepsilon) \leq \dfrac{\mathbb{E}(\varphi(|X|))}{\varphi(\varepsilon)}$,

siehe Satz 5.1.1,

Tschebyscheff: $\varepsilon > 0 \quad \Longrightarrow \quad \mathbb{P}(|X - \mathbb{E}(X)| \geq \varepsilon) \leq \dfrac{\mathbb{V}(X)}{\varepsilon^2}$,

siehe Korollar 5.1.2.

Mit Hilfe der Tschebyscheff-Ungleichung kann man leicht einen der wichtigsten Grenzwertsätze der Wahrscheinlichkeitstheorie herleiten:

Satz 7.1.1 (Schwaches Gesetz der Großen Zahlen)**.** *$X_1, \dots, X_n \in \mathcal{L}^1$ paarweise unkorreliert, $\mathbb{E}(X_i) = E$, $\mathbb{V}(X_i) = V < \infty$, dann gilt*

$$\lim_{n\to\infty} \mathbb{P}(|\tfrac{1}{n}(X_1 + \dots + X_n) - E| > \varepsilon) = 0, \qquad \varepsilon > 0.$$

Den Konvergenztyp in Satz 7.1.1 nennt man die *Konvergenz in Wahrscheinlichkeit*, siehe auch Abschnitt 8.1.

Ferner erinnern wir kurz an die wichtigen Vertauschungssätze für Grenzwerte und Integrale, die wir hier auf einen Wahrscheinlichkeitsraum einschränken:

Monotoner Konvergenzsatz: $0 \leq X_n \uparrow X$ f.s. $\quad \Longrightarrow \quad \lim\limits_{n\to\infty} \mathbb{E}(X_n) = \mathbb{E}(X)$,

siehe Satz 6.8.1,

Satz von Lebesgue: $|X_n| \leq Y \in \mathcal{L}^1, X_n \to X$ f.s. $\quad \Longrightarrow \quad \lim\limits_{n\to\infty} \mathbb{E}(X_n) = \mathbb{E}(X)$,

siehe Satz 6.8.6,

Lemma von Fatou: $0 \leq X_n \quad \Longrightarrow \quad \liminf\limits_{n\to\infty} \mathbb{E}(X_n) \geq \mathbb{E}\Big(\liminf\limits_{n\to\infty} X_n\Big)$,

siehe Satz 6.8.4.

Den Erwartungswert eines Zufallsvektors $X = (X_1, \dots, X_n)$ definiert man komponentenweise, also

$$\mathbb{E}\big((X_1, \dots, X_n)\big) = \big(\mathbb{E}(X_1), \dots, \mathbb{E}(X_n)\big).$$

Definition 7.1.2 (Kovarianz). *(a) Die* Kovarianz *zweier Zufallsgrößen X und Y mit $XY \in \mathcal{L}^1(\mathbb{P})$ wird definiert durch*

$$\operatorname{cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \mathbb{E}\big((X - \mathbb{E}(X))(Y - \mathbb{E}(Y)\big).$$

(b) Die Kovarianzmatrix *eines Zufallsvektors $X = (X_1, \dots, X_n)$ mit $X_i \in \mathcal{L}^1(\mathbb{P})$ und $X_iX_j \in \mathcal{L}^1(\mathbb{P})$ für alle $i, j \in \{1, \dots, n\}$ ist definiert als die Matrix $\Sigma(X) = (\sigma_{i,j}(X))_{i,j\in\{1,\dots,n\}}$ mit $\sigma_{i,j}(X) = \operatorname{cov}(X_i, X_j)$.*

(c) Zwei Zufallsgrößen X und Y mit $XY \in \mathcal{L}^1(\mathbb{P})$ heißen unkorreliert*, wenn $\operatorname{cov}(X, Y) = 0$.*

Falls der Zufallsvektor X als Spalte geschrieben wird, so ist $\Sigma(X) = \mathbb{E}\big((X - \mathbb{E}(X))(X - \mathbb{E}(X))^{\mathrm{T}}\big)$, wobei y^{T} der transponierte Vektor ist. Als Übungsaufgabe sieht man leicht ein, dass $\Sigma(X)$ symmetrisch und positiv semidefinit ist. Die Kovarianz zweier eindimensionaler Zufallsgrößen X und Y ist definiert, sobald $X, Y \in \mathcal{L}^2(\mathbb{P})$, wie leicht aus der Cauchy-Schwarz-Ungleichung folgt. Wenn die Koeffizienten $X_1, \dots, X_n$ von X paarweise unkorreliert sind, ist $\Sigma(X)$ eine Diagonalmatrix, und man errechnet leicht, dass dann gilt: $\mathbb{V}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathbb{V}(X_i)$. Die letzte Aussage nennt man auch den *Satz von Bienaymé*.

Bemerkung 7.1.3 (Normalverteilung). Die mehrdimensionale Normalverteilung $\mathcal{N}(b, \Sigma)$ aus Beispiel 6.6.7 besitzt den Erwartungswertvektor a und die Kovarianzmatrix Σ, wie man mit Hilfe von quadratischen Ergänzungen im Exponenten der Dichte errechnet. ◇

7.2 Unabhängigkeit

In diesem Abschnitt geben wir eine präzise Definition der Unabhängigkeit von Ereignissen, Ereignisfamilien und Zufallsgrößen und listen die wichtigsten Beziehungen auf. Außerdem behandeln wir terminale Ereignisse und das Kolmogorovsche 0-1-Gesetz. Im Folgenden sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum.

Definition 7.2.1 (unabhängig). *(a) Endlich viele Teilmengen $\mathcal{E}_1, \dots, \mathcal{E}_n$ von $\mathcal{F}$ heißen* unabhängig*, falls für jede Auswahl von paarweise verschiedenen $i_1, \dots, i_k \in \{1, \dots, n\}$ und jede Wahl von $A_{i_l} \in \mathcal{E}_{i_l}$ für $l = 1, \dots, k$ gilt:*

$$\mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k}) = \prod_{l=1}^{k} \mathbb{P}(A_{i_l}). \tag{7.2.1}$$

(b) Sei I eine beliebige Indexmenge und $\mathcal{E}_i$ für $i \in I$ eine Teilmenge von $\mathcal{F}$. Dann heißen die $\mathcal{E}_i$ mit $i \in I$ unabhängig*, wenn je endlich viele unabhängig sind.*

(c) Eine beliebige Familie $(A_i)_{i\in I}$ von Ereignissen $A_i \in \mathcal{F}$ heißt unabhängig*, wenn die Mengensysteme $\{A_i, \Omega\}$ mit $i \in I$ unabhängig sind.*

Unabhängigkeit von Ereignissen oder von Mengensystemen hängt also nicht von der Reihenfolge der Ereignisse bzw. der Mengensysteme in der Indexmenge ab, aber durchaus von der Anzahl des Auftretens jedes einzelnen Ereignisses bzw. Mengensystems in der Familie. Wenn zum Beispiel die Familie $(A_i)_{i\in\{1,2\}}$ unabhängig ist mit $A_1 = A_2 = A$, so heißt das insbesondere:

$\mathbb{P}(A) = \mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1)\mathbb{P}(A_2) = \mathbb{P}(A)^2$, also $\mathbb{P}(A) \in \{0,1\}$. In diesem Fall ist also A von sich selber unabhängig. Dieses Beispiel scheint spitzfindig, seine Idee ist aber grundlegend bei diversen 0-1-Gesetzen (wie etwa auch im Beweis von Satz 7.2.8); dies sind Aussagen der Form, dass gewisse Ereignisse nur die Wahrscheinlichkeit 0 oder 1 haben können.

Wir klären nun den engen Zusammenhang zwischen Unabhängigkeit und Produktmaßräumen für abzählbar viele Faktoren. Wir erinnern an Produkte messbarer Räume und Produktmaße, siehe Definition 6.9.1 und Bemerkung 6.10.3.

Lemma 7.2.2 (Produkträume und Unabhängigkeit)**.** *Für $i \in \mathbb{N}$ sei $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$ ein Wahrscheinlichkeitsraum, und es sei $(\Omega, \mathcal{F}, \mathbb{P}) = (\prod_{i\in\mathbb{N}} \Omega_i, \otimes_{i\in\mathbb{N}}\mathcal{F}_i, \otimes_{i\in\mathbb{N}}\mathbb{P}_i)$ der Produktraum. Dann ist für beliebige Wahl von messbaren Mengen $A_i \in \mathcal{F}_i$ die Familie der Mengen $\pi_i^{-1}(A_i)$ unabhängig im Produktraum.*

Beweis. Übungsaufgabe. □

Im Folgenden ist $(\Omega, \mathcal{F}, \mathbb{P})$ ein gegebener Wahrscheinlichkeitsraum und I eine beliebige nicht leere Indexmenge. Teilsysteme von unabhängigen Systemen sind unabhängig:

Lemma 7.2.3. *(a) Falls $\mathcal{E}_i$ für $i \in I$ unabhängig sind und $\mathcal{D}_i \subset \mathcal{E}_i$ für $i \in I$ gilt, so sind die $\mathcal{D}_i$ ebenfalls unabhängig.*

(b) Falls $\mathcal{D}$ und $\mathcal{E}_i$ für jedes $i \in I$ unabhängig sind, so sind auch $\mathcal{D}$ und $\bigcup_{i\in I} \mathcal{E}_i$ unabhängig.

Beweis. einfache Übungsaufgabe. □

Wir diskutieren nun ein paar Möglichkeiten, Unabhängigkeiten von Mengensystemen auf größere Systeme hoch zu ziehen.

Satz 7.2.4. *Wenn für $i \in I$ die Systeme $\mathcal{D}_i \subset \mathcal{F}$ unabhängig sind und durchschnittstabil, so sind die σ-Algebren $\sigma(\mathcal{D}_i)$ unabhängig.*

Beweis. Es genügt, für jede endliche Teilmenge J von I (sagen wir, $J = \{1, \dots, n\}$) zu zeigen, dass $\mathbb{P}(\bigcap_{i=1}^n A_i) = \prod_{i=1}^n \mathbb{P}(A_i)$ für jede Wahl von $A_i \in \sigma(\mathcal{D}_i)$ für $i \in \{1, \dots, n\}$ gilt. Für $k = 0, \dots, n$ sei L_k die folgende Aussage:

$$\mathbb{P}\Big(\bigcap_{i=1}^n A_i\Big) = \prod_{i=1}^n \mathbb{P}(A_i) \qquad \forall A_i \in \sigma(\mathcal{D}_i) \text{ für } i \leq k, \forall A_i \in \mathcal{D}_i \cup \{\Omega\} \text{ für } i > k.$$

Die Aussage L_0 gilt wegen der Unabhängigkeit der $\mathcal{D}_i$. Es sei $0 \leq k < n$, und es gelte L_k. Wir betrachten das Mengensystem

$$\mathcal{A}_{k+1} = \Big\{A_{k+1} \in \sigma(\mathcal{D}_{k+1}) \colon \mathbb{P}\Big(\bigcap_{i=1}^n A_i\Big) = \prod_{i=1}^n \mathbb{P}(A_i)$$
$$\forall A_i \in \sigma(\mathcal{D}_i) \text{ für } i \leq k, \forall A_i \in \mathcal{D}_i \cup \{\Omega\} \text{ für } i > k+1\Big\}.$$

Da L_k gilt, ist $\mathcal{D}_{k+1}$ in $\mathcal{A}_{k+1}$ enthalten. Wir zeigen, dass $\mathcal{A}_{k+1}$ ein Dynkin-System ist, denn dann gilt $d(\mathcal{D}_{k+1}) = \sigma(\mathcal{D}_{k+1})$ nach Satz 6.2.3 und daher $\mathcal{A}_{k+1} = \sigma(\mathcal{D}_{k+1})$.

Nach Definition ist $\Omega \in \mathcal{A}_{k+1}$. Wenn $D \in \mathcal{A}_{k+1}$, so gilt für alle $A_i \in \sigma(\mathcal{D}_i)$ für $i \leq k$, und für alle $A_i \in \mathcal{D}_i$ für $i > k+1$:

$$\begin{aligned}\mathbb{P}\Big(\bigcap_{j=1}^{k} A_j \cap D^{\mathrm{c}} \cap \bigcap_{j=k+2}^{n} A_j\Big) &= \mathbb{P}\Big(\bigcap_{j=1}^{k} A_j \cap \bigcap_{j=k+2}^{n} A_j\Big) - \mathbb{P}\Big(\bigcap_{j=1}^{k} A_j \cap D \cap \bigcap_{j=k+2}^{n} A_j\Big)\\ &= \prod_{j\colon j\neq k+1} \mathbb{P}(A_j) - \mathbb{P}(D) \prod_{j\colon j\neq k+1} \mathbb{P}(A_j)\\ &= \mathbb{P}(D^{\mathrm{c}}) \prod_{j\colon j\neq k+1} \mathbb{P}(A_j),\end{aligned}$$

wobei im zweiten Schritt die Induktionsvoraussetzung und die Tatsache $D \in \mathcal{A}_{k+1}$ ausgenutzt wurden. Also gilt auch $D^{\mathrm{c}} \in \mathcal{A}_{k+1}$. Analog zeigt man, dass $\mathcal{A}_{k+1}$ abgeschlossen ist gegen abzählbare paarweise disjunkte Vereinigungsbildung; man muss in der obigen Rechnung nur D^{c} durch eine abzählbare paarweise disjunkte Vereinigung ersetzen und analog verfahren. D. h. $\mathcal{A}_{k+1}$ ist ein Dynkin-System, und dies beendet den Beweis. □

Bemerkung 7.2.5. Falls die Ereignisse A_i mit $i \in I$ unabhängig sind, so sind also auch die σ-Algebren $\sigma(\{A_i, \Omega\}) = \{\emptyset, A_i, A_i^{\mathrm{c}}, \Omega\}$ unabhängig. Insbesondere sind also auch die Komplemente A_i^{c} unabhängig. ◇

Korollar 7.2.6. *Es seien $\mathcal{D}_i \subset \mathcal{F}$ für $i \in I$ unabhängig und durchschnittsstabil. Ferner sei $(I_k)_{k\in K}$ eine Familie von paarweise disjunkten Teilmengen von I. Dann sind die $\sigma(\bigcup_{j\in I_k} \mathcal{D}_j)$ für $k \in K$ unabhängig.*

Beweis. Für $k \in K$ sei $\widehat{\mathcal{D}}_k$ die Familie der endlichen Durchschnitte von Elementen aus $\bigcup_{j\in I_k} \mathcal{D}_j$. Das Mengensystem $\widehat{\mathcal{D}}_k$ ist offenbar durchschnittsstabil, und da die $\mathcal{D}_j$ durchschnittsstabil sind, hat jedes Element aus $\widehat{\mathcal{D}}_k$ die Gestalt $A_{j_1} \cap \dots \cap A_{j_n}$ mit $n \in \mathbb{N}$, $A_j \in \mathcal{D}_j$ und paarweise verschiedenen $j_1, \dots, j_n \in I_k$. Daraus folgt, dass die $\widehat{\mathcal{D}}_k$ für $k \in K$ unabhängig sind. Da $\mathcal{D}_j \subset \widehat{\mathcal{D}}_k$ für alle $j \in I_k$, gilt $\sigma(\bigcup_{j\in I_k} \mathcal{D}_j) = \sigma(\widehat{\mathcal{D}}_k)$. Daher folgt die Aussage aus Satz 7.2.4. □

Eine interessante Kollektion von Ereignissen ist die folgende:

Definition 7.2.7 (terminale Ereignisse). *Es sei $(\mathcal{F}_n)_{n\in\mathbb{N}}$ eine Folge von Teil-σ-Algebren von $\mathcal{F}$, und wir setzen $\widehat{\mathcal{F}}_n = \bigvee_{k=n}^{\infty} \mathcal{F}_k$ für $n \in \mathbb{N}$. Dann heißt die σ-Algebra*

$$\mathcal{T}_\infty = \bigcap_{n\in\mathbb{N}} \widehat{\mathcal{F}}_n = \bigcap_{n\in\mathbb{N}} \bigvee_{k=n}^{\infty} \mathcal{F}_k$$

die σ-Algebra der terminalen Ereignisse *oder die* terminale σ-Algebra *der Folge $(\mathcal{F}_n)_{n\in\mathbb{N}}$.*

In $\mathcal{T}_\infty$ sind diejenigen Ereignisse enthalten, die für jedes $n \in \mathbb{N}$ nur von den σ-Algebren $\mathcal{F}_n, \mathcal{F}_{n+1}, \dots$ abhängen. Typische Anwendungen betreffen Fälle, in denen $\mathcal{F}_n$ die 'Gegenwart' zum Zeitpunkt n beschreibt, also beschreibt $\widehat{\mathcal{F}}_n$ die 'Zukunft' ab dem Zeitpunkt n. In diesem Fall enthält $\mathcal{T}_\infty$ alle Ereignisse der *infinitesimalen Zukunft*, d. h. alle Ereignisse, deren Eintreten auch dann noch beurteilt werden kann, wenn man die Beobachtung beliebig spät beginnt, vorausgesetzt, man betrachtet danach *alle* späteren Zeitpunkte. Wenn etwa $(X_n)_{n\in\mathbb{N}}$ eine Folge von Zufallsgrößen ist und X_n eine Größe zum Zeitpunkt n beschreibt, dann liegen etwa die Ereignisse 'die Folge $(X_n)_{n\in\mathbb{N}}$ besucht den Punkt i unendlich oft' und 'die Folge $(X_n)_{n\in\mathbb{N}}$ bleibt ab einem Zeitpunkt nur noch in A' (für einen gegebenen Punkt i bzw. Menge A) in $\mathcal{T}_\infty$, wobei wir $\mathcal{F}_n = \sigma(X_n)$ gesetzt haben.

Der folgende, verblüffende Satz (mit verblüffendem Beweis) sagt, dass bei Unabhängigkeit der $\mathcal{F}_n$ die terminale σ-Algebra in gewissem Sinn trivial ist:

Satz 7.2.8 (Kolmogorov'sches 0-1-Gesetz). *Es sei $(\mathcal{F}_n)_{n\in\mathbb{N}}$ eine Folge von unabhängigen Teil-σ-Algebren von $\mathcal{F}$ mit terminaler σ-Algebra $\mathcal{T}_\infty$. Dann gilt $\mathbb{P}(A) \in \{0,1\}$ für jedes $A \in \mathcal{T}_\infty$.*

Beweis. Es sei $n \in \mathbb{N}$ fest. Nach Korollar 7.2.6 sind $\widehat{\mathcal{F}}_{n+1}$ und $\bigvee_{k=1}^n \mathcal{F}_k$ unabhängig, also sind nach Lemma 7.2.3(a) auch $\mathcal{T}_\infty$ und $\bigvee_{k=1}^n \mathcal{F}_k$ unabhängig. Daher sind nach Lemma 7.2.3(b) auch $\mathcal{T}_\infty$ und $\bigcup_{n\in\mathbb{N}} \bigvee_{k=1}^n \mathcal{F}_k$ unabhängig. Die letztere Menge ist als eine Vereinigung aufsteigender σ-Algebren durchschnittstabil, also sind nach Satz 7.2.4 auch $\mathcal{T}_\infty$ und die von ihr erzeugte σ-Algebra unabhängig, und dies ist die σ-Algebra $\bigvee_{n\in\mathbb{N}} \mathcal{F}_n$. Diese enthält $\widehat{\mathcal{F}}_n$ für jedes $n \in \mathbb{N}$, also enthält sie auch deren Schnitt $\mathcal{T}_\infty$. Nach Lemma 7.2.3 ist also $\mathcal{T}_\infty$ von sich selber unabhängig. Daraus folgt für jedes $A \in \mathcal{T}_\infty$, dass $\mathbb{P}(A) = \mathbb{P}(A \cap A) = \mathbb{P}(A)\mathbb{P}(A) = \mathbb{P}(A)^2$, also muss $\mathbb{P}(A) \in \{0,1\}$ sein. □

Man nennt eine σ-Algebra $\mathcal{T}$ mit $\mathbb{P}(A) \in \{0,1\}$ für jedes $A \in \mathcal{T}$ auch $\mathbb{P}$-*trivial*. Eine $\mathbb{P}$-triviale σ-Algebra ist natürlich $\{\emptyset, \Omega\}$, aber terminale σ-Algebren sind meist sehr viel größer.

Beispiel 7.2.9 (Perkolation). Es sei $\mathbb{K}$ die Menge aller Kanten zwischen je zwei Nachbarn auf dem Gitter $\mathbb{Z}^d$. Mit einem Parameter $p \in (0,1)$ sei $(X_i)_{i\in\mathbb{K}}$ eine Kollektion unabhängiger Zufallsgrößen, die jeweils mit Wahrscheinlichkeit p den Wert 'offen' und sonst den Wert 'undurchlässig' annehmen. Dies ist ein Modell für einen porösen Stein, dessen Kammern durch Wände verbunden sind, die zufällig durchlässig sind oder nicht. In einem Punkt denken wir uns eine Wasserquelle. Das Wasser kann nur entlang durchlässiger Kanten fließen. Als Übungsaufgabe zeige man, dass das Ereignis 'Es gibt einen Punkt, von dem aus das Wasser einen unendlich langen Weg fließen kann' nur die Wahrscheinlichkeiten 0 oder 1 haben kann. Dieses Ereignis ist gleich bedeutend mit der Existenz eines unendlich großen Clusters, d. h. einer unendlich großen Teilmenge von $\mathbb{Z}^d$, deren Punkte alle mit durchlässigen Kanten mit einander verbunden sind. (In dem Fall, dass dieses Ereignis eintritt, sagt man, dass *Perkolation eintritt.*) ◇

Lemma 7.2.10. *Es sei $\mathcal{T} \subset \mathcal{F}$ eine $\mathbb{P}$-triviale σ-Algebra. Ferner sei X eine $(\overline{\mathbb{R}}, \overline{\mathcal{B}})$-wertige, $\mathcal{T}$-messbare Zufallsgröße. Dann existiert ein $c \in \overline{\mathbb{R}}$ mit $\mathbb{P}(X = c) = 1$.*

Beweis. Es sei F die Verteilungsfunktion von X. Da $\{X \le t\} = X^{-1}([-\infty, t]) \in \mathcal{T}$ für jedes $t \in \mathbb{R}$, gilt $F(t) = \mathbb{P}(X \le t) \in \{0,1\}$ für jedes $t \in \mathbb{R}$. Da F nicht fällt, sind drei Fälle möglich.

1. Falls $F(t) = 0$ für alle $t \in \mathbb{R}$, so ist $\mathbb{P}(X > n) = 1$ für jedes $n \in \mathbb{N}$, also $\mathbb{P}(X = \infty) = 1$.

2. Falls $F(t) = 1$ für alle $t \in \mathbb{R}$, so ist $\mathbb{P}(X \le -n) = 1$ für jedes $n \in \mathbb{N}$, also $\mathbb{P}(X = -\infty) = 1$.

3. Falls F an einer Stelle $t_0 \in \mathbb{R}$ von 0 auf 1 springt, ist $\mathbb{P}(X \in (t_0 - \frac{1}{n}, t_0 + \frac{1}{n}]) = F(t_0 + \frac{1}{n}) - F(t_0 - \frac{1}{n}) = 1$ für jedes $n \in \mathbb{N}$. Durch den Grenzübergang $n \to \infty$ erhält man, dass $\mathbb{P}(X = t_0) = 1$.

□

Definition 7.2.11 (Unabhängigkeit von Zufallsgrößen)**.** *Für jedes $i \in I$ sei $X_i \colon \Omega \to E_i$ eine Zufallsgröße mit Werten in einem messbaren Raum $(E_i, \mathcal{E}_i)$. Die Zufallsgrößen X_i mit $i \in I$ heißen* unabhängig, *wenn die σ-Algebren $X_i^{-1}(\mathcal{E}_i)$ mit $i \in I$ unabhängig sind.*

Bemerkung 7.2.12 (Konstruktion unabhängiger Zufallsgrößen)**.** Man kann aus Zufallsgrößen, die a priori nichts mit einander zu tun haben, eine Familie *unabhängiger* Zufallsgrößen mit den selben Verteilungen konstruieren, indem man alle diese Größen auf dem Produktraum definiert. Wenn für jedes $i \in I$ eine Zufallsgröße $X_i \colon \Omega_i \to E_i$ auf einem Wahrscheinlichkeitsraum $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$ definiert ist und Werte in einem messbaren Raum $(E_i, \mathcal{E}_i)$ hat, so kann man auf dem Produktraum $(\Omega, \mathcal{F}, \mathbb{P})$ (siehe Bemerkung 6.10.3) Zufallsgrößen $\widetilde{X}_i \colon \Omega \to E_i$ durch $\widetilde{X}_i(\omega) = X_i(\omega_i)$ für $\omega = (\omega_i)_{i\in I} \in \Omega$ definieren. Dann sind die Größen $\widetilde{X}_i$ mit $i \in I$ unabhängig, und die Verteilungen von $\widetilde{X}_i$ und X_i stimmen überein. Dies sieht man leicht ein, wenn man ausnutzt, dass $\widetilde{X}_i = X_i \circ \pi_i$ gilt, wobei $\pi_i \colon \Omega \to \Omega_i$ die kanonische Projektion ist. Dann ist nämlich

$$\mathbb{P} \circ \widetilde{X}_i^{-1} = \mathbb{P} \circ \pi_i^{-1} \circ X_i^{-1} = \mathbb{P}_i \circ X_i^{-1},$$

denn $\mathbb{P}_i$ ist das Bildmaß des Produktmaßes $\mathbb{P}$ unter der Projektion π_i. ◇

Wir sammeln ein paar Eigenschaften unabhängiger Zufallsgrößen, die alle leicht einzusehen sind:

Lemma 7.2.13. *Für jedes $i \in I$ sei $X_i \colon \Omega \to E_i$ eine Zufallsgröße mit Werten in einem messbaren Raum $(E_i, \mathcal{E}_i)$.*

(a) Falls $\mathcal{D}_i \subset \mathcal{F}$ für $i \in I$ unabhängig sind und X_i messbar bezüglich $\mathcal{D}_i$, so sind die X_i unabhängig.

(b) Falls die X_i unabhängig sind und $\varphi_i \colon E_i \to E_i'$ Abbildungen, die $\mathcal{E}_i$-$\mathcal{E}_i'$-messbar sind (wobei $\mathcal{E}_i'$ jeweils eine σ-Algebra auf E_i' sei), so sind die Abbildungen $\varphi_i \circ X_i$ ebenfalls unabhängig.

(c) Die X_i sind genau dann unabhängig, wenn für jedes $n \in \mathbb{N}$ und alle paarweise verschiedenen $i_1, \dots, i_n \in I$ und alle $A_1 \in \mathcal{E}_{i_1}, \dots, A_n \in \mathcal{E}_{i_n}$ gilt:

$$\mathbb{P}(X_{i_1} \in A_1, \dots, X_{i_n} \in A_n) = \prod_{j=1}^{n} \mathbb{P}(X_{i_j} \in A_j).$$

Eine der angenehmsten Eigenschaften unabhängiger Zufallsgrößen ist die folgende Rechenregel für den Erwartungswert von Produkten.

Satz 7.2.14. *Es seien X und Y zwei unabhängige reelle Zufallsgrößen. Falls X und Y nichtnegativ sind oder falls $X, Y \in \mathcal{L}^1(\mathbb{P})$, so gilt $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.*

Beweis. Zunächst seien X und Y als nichtnegativ vorausgesetzt. Die beiden σ-Algebren $\mathcal{F}_1 = X^{-1}(\mathcal{B})$ und $\mathcal{F}_2 = Y^{-1}(\mathcal{B})$ sind unabhängig. Für $A \in \mathcal{F}_1$ erfüllt die Menge aller nichtnegativen $\mathcal{F}_2$-messbaren Zufallsgrößen Y' mit $\mathbb{E}(\mathbb{1}_A Y') = \mathbb{P}(A)\mathbb{E}(Y')$ die Eigenschaften (i) - (iii) von Satz 6.4.11. Demzufolge gilt die Gleichung $\mathbb{E}(\mathbb{1}_A Y') = \mathbb{P}(A)\mathbb{E}(Y')$ für alle nichtnegativen $\mathcal{F}_2$-messbaren Zufallsgrößen, also insbesondere für Y selbst. Die Menge der nichtnegativen $\mathcal{F}_1$-messbaren Zufallsgrößen X' mit $\mathbb{E}(X'Y) = \mathbb{E}(X')\mathbb{E}(Y)$ erfüllt ebenfalls die Bedingungen von

Satz 6.4.11. Also ergibt das gleiche Argument wie oben, dass die letzte Gleichung für $X' = X$ gilt.

Nun sei $X, Y \in \mathcal{L}^1(\mathbb{P})$ vorausgesetzt. Da X und Y unabhängig sind, sind dies auch $|X|$ und $|Y|$. Also folgt aus dem Bisherigen, dass $\mathbb{E}(|XY|) = \mathbb{E}(|X|)\mathbb{E}(|Y|)$ gilt. Die Gleichung $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ folgt daraus durch eine Zerlegung in Positiv- und Negativteil. □

Als Anwendung von Satz 7.2.14 zeigen wir, dass unter stärkeren Integrierbarkeitsannahmen die Konvergenzgeschwindigkeit im Schwachen Gesetz der Großen Zahlen (Satz 7.1.1) bei Unabhängigkeit sogar exponentiell ist und nicht nur linear.

Lemma 7.2.15 (Große Abweichungen). *Sei $(X_n)_{n\in\mathbb{N}}$ eine Folge von unabhängigen, identisch verteilten Zufallsgrößen. Setze $S_n = X_1 + \cdots + X_n$. Dann gilt für jedes $c > 0$:*

$$\frac{1}{n} \log \mathbb{P}(S_n > cn) \le - \sup_{\alpha\in(0,\infty)} \big[c\alpha - \log \mathbb{E}(e^{\alpha X_1})\big].$$

Man erhält exponentielle Konvergenz in Satz 7.1.1, wenn man Lemma 7.2.15 anwendet auf $c = \mathbb{E}(X_1) + \varepsilon$ und ein zweites Mal auf $\tilde{X}_i = -X_i$ und $c = \mathbb{E}(\tilde{X}_1) + \varepsilon$, und wenn man voraus setzt, dass $\mathbb{E}(e^{\alpha X_1}) < \infty$ für jedes $\alpha \in \mathbb{R}$.

Beweis von Lemma 7.2.15. Wir wenden die Markov-Ungleichung von Satz 5.1.1 wie folgt an: Für jedes $\alpha > 0$ gilt

$$\mathbb{P}(S_n > cn) = \mathbb{P}(e^{\alpha S_n} > e^{\alpha cn}) \le e^{-\alpha cn}\mathbb{E}(e^{\alpha S_n}).$$

Nun schreiben wir $e^{\alpha S_n} = \prod_{i=1}^n e^{\alpha X_i}$ und benutzen Lemma 7.2.13(b) und Satz 7.2.14, um den Erwartungswert wie folgt zu berechnen:

$$\mathbb{E}(e^{\alpha S_n}) = \mathbb{E}\Big[\prod_{i=1}^n e^{\alpha X_i}\Big] = \prod_{i=1}^n \mathbb{E}(e^{\alpha X_i}) = \mathbb{E}(e^{\alpha X_1})^n.$$

Dies setzen wir oben ein, nehmen den Logarithmus, teilen durch n und optimieren über α. □

Wir beleuchten kurz die Verteilung der Summe zweier unabhängiger Zufallsvektoren:

Definition 7.2.16 (Faltung). *Die* Faltung *zweier Wahrscheinlichkeitsmaße $\mathbb{P}_1$ und $\mathbb{P}_2$ auf $(\mathbb{R}^d, \mathcal{B}_d)$ ist das Bildmaß*

$$\mathbb{P}_1 * \mathbb{P}_2 = (\mathbb{P}_1 \otimes \mathbb{P}_2) \circ S^{-1},$$

wobei $S\colon \mathbb{R}^{2d} \to \mathbb{R}^d$ definiert ist als $S(x_1, x_2) = x_1 + x_2$.

Als Übungsaufgabe zeigt man, dass die Faltung zweier Maße mit (Lebesgue-)Dichten f_1 und f_2 die Dichte $f_1 * f_2$ besitzt, wobei $f_1 * f_2$ die übliche Faltung zweier Funktionen f_1 und f_2 bezeichnet. Ferner sieht man ebenfalls leicht, dass die Faltung zweier Wahrscheinlichkeitsmaße $\sum_{n\in\mathbb{Z}^d} p_n\delta_n$ und $\sum_{n\in\mathbb{Z}^d} q_n\delta_n$ auf $\mathbb{Z}^d$ gleich dem Wahrscheinlichkeitsmaß $\sum_{n\in\mathbb{Z}^d} r_n\delta_n$ ist, wobei $r = (r_n)_{n\in\mathbb{Z}^d}$ die Faltung der Folgen $(p_n)_{n\in\mathbb{Z}^d}$ und $(q_n)_{n\in\mathbb{Z}^d}$ ist. Mit Hilfe dieser Tatsachen errechnet man als eine Übungsaufgabe die folgenden Faltungseigenschaften der Normal- und der Poisson-Verteilungen:

Lemma 7.2.17. *(i) Die Summe zweier unabhängiger, $\mathcal{N}(a, \sigma^2)$- bzw. $\mathcal{N}(b, \tau^2)$- verteilter Zufallsgrößen ist $\mathcal{N}(a + b, \sigma^2 + \tau^2)$-verteilt.*

(ii) Die Summe zweier unabhängiger Poisson-verteilter Zufallsgrößen mit Parametern $\alpha > 0$ bzw. $\beta > 0$ ist Poisson-verteilt mit Parameter $\alpha + \beta$.

Ein anderer Beweis von Lemma 7.2.17 kann mit Hilfe von charakteristischen Funktionen geführt werden, siehe Abschnitt 8.3.

7.3 Bedingte Erwartungswerte

Wir wollen das Konzept der bedingten Wahrscheinlichkeiten, das wir in Abschnitt 2.1 in der diskreten Situation einführten, stark erweitern. Insbesondere möchten wir auf viele Ereignisse (statt nur auf eines) bedingen, und wir möchten erklären, in welchem Sinne man auf eine Nullmenge bedingen darf. Sammeln wir zunächst ein paar Tatsachen, deren Rechtfertigung uns keinerlei Probleme bereitet.

Sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $A \in \mathcal{F}$, dann kann man ohne Probleme für jedes $B \in \mathcal{F}$ die bedingte Wahrscheinlichkeit von B gegeben A einführen als

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)}, \qquad \text{falls } \mathbb{P}(A) > 0.$$

Wir erinnern an die Interpretation von $\mathbb{P}(B \mid A)$ als die Wahrscheinlichkeit von B unter der Voraussetzung, dass A eingetreten ist. Im Fall $\mathbb{P}(A) = 0$ kann man $\mathbb{P}(\cdot \mid A)$ als irgendein Wahrscheinlichkeitsmaß definieren, wenn man möchte oder muss, aber das ist willkürlich. Man sieht leicht, dass mit dieser Erweiterung der Definition 2.1.2 die Formeln von der totalen Wahrscheinlichkeit und von Bayes (siehe Lemma 2.1.4) gelten. Außerdem ist auch leicht zu sehen, dass $\mathbb{P}(\cdot \mid A)$ ein Wahrscheinlichkeitsmaß ist. Also kann man auch problemlos den *bedingten Erwartungswert* einer Zufallsgröße $X \in \mathcal{L}^1(\mathbb{P})$ definieren, indem man setzt:

$$\mathbb{E}[X \mid A] = \frac{\mathbb{E}[X \, \mathbb{1}_A]}{\mathbb{P}(A)} = \int_\Omega X(\omega) \, \mathbb{P}(\mathrm{d}\omega \mid A). \tag{7.3.1}$$

Natürlich haben wir dann auch die Formel $\mathbb{E}[\mathbb{1}_B \mid A] = \mathbb{P}(B \mid A)$.

Nun möchten wir gerne auf viele Ereignisse bedingen anstatt nur auf ein A. Mit anderen Worten, wir wollen voraussetzen, dass das Eintreten oder Nichteintreten einer gewissen Vielzahl von Ereignissen bekannt sei. Nun allerdings kann die bedingte Wahrscheinlichkeit nicht mehr eine einzige Zahl sein, denn sie wird davon abhängen müssen, welche der bedingenden Ereignisse eingetreten sind und welche nicht. Es fällt uns leicht, den folgenden Spezialfall zu behandeln: Falls $(A_i)_{i \in I}$ eine (höchstens abzählbare) Familie disjunkter messbarer Mengen ist, deren Vereinigung Ω ist, dann definieren wir $\mathcal{A} = \sigma(A_i \colon i \in I)$ und für $X \in \mathcal{L}^1(\mathbb{P})$

$$\mathbb{E}[X \mid \mathcal{A}](\omega) = \mathbb{E}[X \mid A_i] \qquad \Longleftrightarrow \qquad \omega \in A_i, \tag{7.3.2}$$

also $\mathbb{E}[X \mid \mathcal{A}] = \sum_{i \in I} \mathbb{1}_{A_i} \mathbb{E}[X \mid A_i]$. Diese Abbildung hat die folgenden zwei wichtigen Eigenschaften.

Lemma 7.3.1. *(i)* $\mathbb{E}[X \mid \mathcal{A}]$ *ist* $\mathcal{A}$*-messbar,*

(ii) $\mathbb{E}[X \mid \mathcal{A}] \in \mathcal{L}^1(\mathbb{P})$*, und für jedes* $A \in \mathcal{A}$ *gilt* $\int_A \mathbb{E}[X \mid \mathcal{A}] \, \mathrm{d}\mathbb{P} = \int_A X \, \mathrm{d}\mathbb{P}$.

Beweis. (i) $\mathbb{E}[X \mid \mathcal{A}]$ ist die Hintereinanderschaltung der beiden messbaren Abbildungen $\Omega \ni \omega \mapsto i(\omega) \in I$, wobei $i(\omega)$ definiert ist durch $\omega \in A_{i(\omega)}$, und $I \ni i \mapsto \mathbb{E}[X \mid A_i]$.

(ii) Für $A \in \mathcal{A}$ gibt es eine Teil-Indexmenge $J \subset I$ mit $A = \cup_{i\in J} A_i$. O.B.d.A. haben wir $\mathbb{P}(A_i) > 0$ für jedes $i \in J$. Dann folgt

$$\int_A \mathbb{E}[X \mid \mathcal{A}] \, \mathrm{d}\mathbb{P} = \sum_{i\in J} \mathbb{P}(A_i)\, \mathbb{E}[X \mid A_i] = \sum_{i\in J} \mathbb{E}[X \mathbb{1}_{A_i}] = \int_A X \, \mathrm{d}\mathbb{P}.$$

□

Dies gibt Anlass zu einer Definition. Im Folgenden sei immer $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum, $\mathcal{A} \subset \mathcal{F}$ eine Teil-σ-Algebra und $X \in \mathcal{L}^1(\mathbb{P})$ eine Zufallsgröße.

Definition 7.3.2 (Bedingte Erwartung). *Eine Zufallsgröße Y heißt eine* bedingte Erwartung *von X bezüglich $\mathcal{A}$, und wir schreiben $Y = \mathbb{E}[X \mid \mathcal{A}]$, falls gelten:*

(i) Y ist $\mathcal{A}$-messbar,

(ii) Für jedes $A \in \mathcal{A}$ gilt $\mathbb{E}[X \mathbb{1}_A] = \mathbb{E}[Y \mathbb{1}_A]$.

Falls $B \in \mathcal{F}$, so heißt $\mathbb{P}(B \mid \mathcal{A}) = \mathbb{E}[\mathbb{1}_B \mid \mathcal{A}]$ die bedingte Wahrscheinlichkeit *von B gegeben $\mathcal{A}$.*

Wenn man (ii) für $A = \Omega$ ausnutzt, erhält man insbesondere, dass $\mathbb{E}[\mathbb{E}[X \mid \mathcal{A}]] = \mathbb{E}[X]$, d. h. die bedingte Erwartung von X hat den selben Erwartungswert wie X.

Bedingte Erwartungen und bedingte Wahrscheinlichkeiten werden also nur bis auf fast sichere Gleichheit definiert. Auch wenn wir nicht immer darauf hinweisen werden, sind also alle Gleichungen und Aussagen, in denen sie auftreten, nur fast sicher zu verstehen.

Bemerkung 7.3.3. Auch wenn die Schreibweise sehr suggestiv ist, ist es überhaupt nicht klar (und es gilt auch nicht in jeder Situation), dass die Abbildung $B \mapsto \mathbb{P}(B \mid \mathcal{A})$ ein Wahrscheinlichkeitsmaß ist. Genauer gesagt, würde man gerne zumindest für fast alle $\omega \in \Omega$ haben wollen, dass $\mathbb{P}(\cdot \mid \mathcal{A})(\omega)$ ein Wahrscheinlichkeitsmaß ist, aber auch hierfür muss man zusätzliche Voraussetzungen machen und diese Aussage beweisen. Grob gesagt, ist das Problem, dass gewisse Ausnahmemengen *a priori* von B abhängen, und man muss zeigen, dass es eine Version von $B \mapsto \mathbb{P}(B \mid \mathcal{A})(\omega)$ gibt, so dass diese Ausnahmemengen *nicht* von B abhängen. ◇

Bemerkung 7.3.4. Wir können mit den bisherigen Methoden zum Beispiel nicht die folgende, intuitiv lösbar scheinende, Aufgabe behandeln: Gegeben seien eine gleichförmig auf $[0, 1]$ verteilte Zufallsgröße X und weitere Zufallsgrößen $Y_1, \dots, Y_n$, die bei Kenntnis des Ereignisses $\{X = x\}$ unabhängig und Bernoulli-verteilt mit Parameter x sind. Können wir der bedingten Wahrscheinlichkeit $\mathbb{P}(Y_1 = \cdots = Y_n = 1 \mid X = x)$ einen Sinn geben? Intuitiv sollte das gleich x^n sein, aber wie erklärt man die Bedingung auf die Nullmenge $\{X = x\}$? Wir brauchen also einen Begriff des Erwartungswertes bedingt auf Nullmengen, der unserer Intuition entspricht. ◇

Beim sogleich folgenden Beweis der Existenz der bedingten Erwartung werden wir einen der wichtigsten Sätze der Maßtheorie benötigen, der auch schon in Bemerkung 6.6.5 erwähnt wurde. Wir erinnern daran, dass wir $\nu \ll \mu$ schreiben, wenn das Maß ν absolutstetig bezüglich dem Maß μ ist, d. h. wenn jede μ-Nullmenge auch eine ν-Nullmenge ist.

Satz 7.3.5 (Satz von Radon-Nikodym)**.** *Seien μ und ν zwei σ-endliche Maße auf einem messbaren Raum $(\Omega, \mathcal{F})$. Dann gilt*

$$\nu \text{ hat eine Dichte bezüglich } \mu \qquad \Longleftrightarrow \qquad \nu \ll \mu.$$

In diesem Fall ist die Dichte $\mathcal{F}$-messbar, μ-fast überall endlich und μ-fast überall eindeutig. Die Dichte $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$ heißt die Radon-Nikodym-Dichte *von ν nach μ.*

Beweise finden sich in jedem guten Maßtheoriebuch.

Satz 7.3.6. *Die bedingte Erwartung $\mathbb{E}[X \mid \mathcal{A}]$ existiert und ist fast sicher eindeutig.*

Beweis. Seien Y und Y' bedingte Erwartungen von X bezüglich $\mathcal{A}$. Für die Menge $A = \{Y > Y'\} \in \mathcal{A}$ gilt dann $\mathbb{E}[(Y - Y')1\!\!1_A] = \mathbb{E}[Y 1\!\!1_A] - \mathbb{E}[Y' 1\!\!1_A] = 0$. Da $(Y - Y')1\!\!1_A \geq 0$, folgt $Y \leq Y'$ fast sicher. Analog folgt die umgekehrte Ungleichung, also die fast sichere Eindeutigkeit.

Wir zerlegen $X = X^+ - X^-$ in den Positiv- und den Negativteil $X^+ = X \vee 0$ bzw. $X^- = X^+ - X$. Durch $Q^+(A) = \mathbb{E}[X^+ 1\!\!1_A]$ bzw. $Q^-(A) = \mathbb{E}[X^- 1\!\!1_A]$ werden zwei endliche Maße auf $(\Omega, \mathcal{A})$ definiert. Beide Maße sind offensichtlich absolutstetig bezüglich $\mathbb{P}$, also liefert der Satz von Radon-Nikodym die Existenz von $\mathcal{A}$-messbaren Dichten Y^+ bzw. Y^-, so dass $Q^+(A) = \int_A Y^+ \,\mathrm{d}\mathbb{P} = \mathbb{E}[Y^+ 1\!\!1_A]$ und $Q^-(A) = \int_A Y^- \,\mathrm{d}\mathbb{P} = \mathbb{E}[Y^- 1\!\!1_A]$ für jedes $A \in \mathcal{A}$ gelten. Dann ist $Y = Y^+ - Y^-$ eine bedingte Erwartung von X gegeben $\mathcal{A}$, denn Y ist $\mathcal{A}$-messbar, und für jedes $A \in \mathcal{A}$ gilt $\mathbb{E}[Y 1\!\!1_A] = \mathbb{E}[Y^+ 1\!\!1_A] - \mathbb{E}[Y^- 1\!\!1_A] = \mathbb{E}[X^+ 1\!\!1_A] - \mathbb{E}[X^- 1\!\!1_A] = \mathbb{E}[X 1\!\!1_A]$. □

Nun können wir auch die bedingte Erwartung bezüglich einer anderen Zufallsgröße definieren:

Definition 7.3.7. *Für jede Zufallsgröße Y heißt $\mathbb{E}[X \mid Y] = \mathbb{E}[X \mid \sigma(Y)]$ die bedingte Erwartung von X bezüglich Y.*

Die wichtigsten Eigenschaften der bedingten Erwartung werden nun aufgelistet:

Satz 7.3.8 (Eigenschaften der bedingten Erwartung)**.** *Seien $\mathcal{A}_1 \subset \mathcal{A} \subset \mathcal{F}$ zwei Teil-σ-Algebren, und sei $Y \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ und $\lambda \in \mathbb{R}$, dann gelten:*

(i) *(Linearität)* $\mathbb{E}[\lambda X + Y \mid \mathcal{A}] = \lambda \mathbb{E}[X \mid \mathcal{A}] + \mathbb{E}[Y \mid \mathcal{A}]$.

(ii) *(Monotonie) Falls $X \geq Y$ fast sicher, so $\mathbb{E}[X \mid \mathcal{A}] \geq \mathbb{E}[Y \mid \mathcal{A}]$ fast sicher.*

(iii) *Falls $\mathbb{E}[|XY|] < \infty$ und Y messbar bezüglich $\mathcal{A}$, so gelten*

$$\mathbb{E}[XY \mid \mathcal{A}] = Y \mathbb{E}[X \mid \mathcal{A}] \qquad \text{und insbesondere} \qquad \mathbb{E}[Y \mid Y] = Y.$$

(iv) $\mathbb{E}[\mathbb{E}[X \mid \mathcal{A}] \mid \mathcal{A}_1] = \mathbb{E}[\mathbb{E}[X \mid \mathcal{A}_1] \mid \mathcal{A}] = \mathbb{E}[X \mid \mathcal{A}_1]$.

(v) $|\mathbb{E}[X \mid \mathcal{A}]| \leq \mathbb{E}[|X| \mid \mathcal{A}]$.

(vi) *Falls $\sigma(X)$ und $\mathcal{A}$ unabhängig sind, so ist $\mathbb{E}[X \mid \mathcal{A}] = \mathbb{E}[X]$.*

(vii) *Falls $\mathbb{P}(A) \in \{0,1\}$ für jedes $A \in \mathcal{A}$, so gilt $\mathbb{E}[X \mid \mathcal{A}] = \mathbb{E}[X]$.*

(viii) *(Majorisierte Konvergenz) Falls $Y \geq 0$ und $(X_n)_{n\in\mathbb{N}}$ eine Folge von Zufallsgrößen ist mit $|X_n| \leq Y$ für jedes $n \in \mathbb{N}$ sowie $\lim_{n\to\infty} X_n = X$ fast sicher, so gilt $\lim_{n\to\infty} \mathbb{E}[X_n \mid \mathcal{A}] = \mathbb{E}[X \mid \mathcal{A}]$ fast sicher und in $\mathcal{L}^1(\mathbb{P})$.*

Beweis. (i), (ii): Übungsaufgaben.

(iii): Es ist klar, dass $Y\mathbb{E}[X \mid \mathcal{A}]$ die erste Bedingung in der Definition 7.3.2 der bedingten Erwartung von XY bezüglich $\mathcal{A}$ erfüllt. Wir wenden uns der zweiten zu. Es reicht, den Fall $X \geq 0$ und $Y \geq 0$ zu betrachten, denn der allgemeine Fall wird mit Hilfe der Linearität und (i) durch eine Zerlegung in Positiv- und Negativteil erledigt. Die Folge der $Y_n = 2^{-n}\lfloor 2^n Y\rfloor$ konvergiert steigend gegen Y. (Mit $\lfloor x\rfloor$ bezeichnen wir die größte ganze Zahl $\leq x$.) Das gilt auch für $Y_n\mathbb{E}[X \mid \mathcal{A}]$ gegen $Y\mathbb{E}[X \mid \mathcal{A}]$, denn diese Zufallsgrößen sind nach (ii) nichtnegativ. Nach dem Satz von der monotonen Konvergenz ist dann auch für jedes $A \in \mathcal{A}$

$$\lim_{n\to\infty} \mathbb{E}[\mathbb{1}_A Y_n \mathbb{E}[X \mid \mathcal{A}]] = \mathbb{E}[\mathbb{1}_A Y \mathbb{E}[X \mid \mathcal{A}]].$$

Andererseits ist wegen der Messbarkeit von Y bezüglich $\mathcal{A}$ (man beachte, dass $\{Y_n = k2^{-n}\} = \{Y \in [k2^{-n}, (k+1)2^{-n})\} \in \mathcal{A}$ für jedes $n, k \in \mathbb{N}_0$)

$$\begin{aligned}
\mathbb{E}[\mathbb{1}_A Y_n \mathbb{E}[X \mid \mathcal{A}]] &= \sum_{k=1}^{\infty} \mathbb{E}\big[\mathbb{1}_A \mathbb{1}_{\{Y_n = k2^{-n}\}} k2^{-n} \mathbb{E}[X \mid \mathcal{A}]\big] \\
&= \sum_{k=1}^{\infty} \mathbb{E}\big[\mathbb{1}_A \mathbb{1}_{\{Y_n = k2^{-n}\}} k2^{-n} X\big] = \mathbb{E}[\mathbb{1}_A Y_n X],
\end{aligned}$$

und dies konvergiert gegen $\mathbb{E}[\mathbb{1}_A Y X]$. Also gilt $\mathbb{E}[\mathbb{1}_A Y \mathbb{E}[X \mid \mathcal{A}]] = \mathbb{E}[\mathbb{1}_A Y X]$ für jedes $A \in \mathcal{A}$. Also erfüllt $Y\mathbb{E}[X \mid \mathcal{A}]$ auch die zweite Bedingung in der Definition der bedingten Erwartung von XY bezüglich $\mathcal{A}$.

(iv) Übungsaufgabe.

(v) Folgt aus (i) und (ii) mit Hilfe einer Zerlegung in Positiv- und Negativteil.

(vi) Für $A \in \mathcal{A}$ sind X und $\mathbb{1}_A$ unabhängig, also gilt $\mathbb{E}[X\mathbb{1}_A] = \mathbb{E}[X]\mathbb{E}[\mathbb{1}_A]$.

(vii) $\mathcal{A}$ ist von jeder σ-Algebra unabhängig, also auch von $\sigma(X)$. Daher folgt die Aussage aus (vi).

(viii) Die Zufallsgrößen $Z_n = \sup_{k\geq n} |X_k - X|$ erfüllen $0 \leq Z_n \leq 2Y$ und $\lim_{n\to\infty} Z_n = 0$ fast sicher. Insbesondere gilt $\lim_{n\to\infty} \mathbb{E}[Z_n] = 0$ nach dem Satz 6.8.6 von Lebesgue. Mit Hilfe von (i) und (v) sieht man, dass gilt

$$\big|\mathbb{E}[X_n \mid \mathcal{A}] - \mathbb{E}[X \mid \mathcal{A}]\big| \leq \mathbb{E}[Z_n \mid \mathcal{A}]. \tag{7.3.3}$$

Also ist der Erwartungswert der linken Seite nicht größer als $\mathbb{E}[\mathbb{E}[Z_n \mid \mathcal{A}]] = \mathbb{E}[Z_n]$, der ja gegen Null konvergiert. Dies zeigt die behauptete L^1-Konvergenz von $\mathbb{E}[X_n \mid \mathcal{A}]$ gegen $\mathbb{E}[X \mid \mathcal{A}]$.

Nach dem Lemma 6.8.4 von Fatou gilt wegen der Monotonie von $n \mapsto Z_n$

$$\mathbb{E}\big[\limsup_{n\to\infty} \mathbb{E}[Z_n \mid \mathcal{A}]\big] = \mathbb{E}\big[\liminf_{n\to\infty} \mathbb{E}[Z_n \mid \mathcal{A}]\big] \leq \lim_{n\to\infty} \mathbb{E}[Z_n] = 0.$$

Daher ist auch $\limsup_{n\to\infty} \mathbb{E}[Z_n \mid \mathcal{A}] = 0$ fast sicher, also auch $\lim_{n\to\infty} \mathbb{E}[Z_n \mid \mathcal{A}] = 0$ fast sicher. Nun folgt die fast sichere Konvergenz von $\mathbb{E}[X_n \mid \mathcal{A}]$ gegen $\mathbb{E}[X \mid \mathcal{A}]$ aus (7.3.3). □

Bemerkung 7.3.9 (Interpretation)**.** Wie im diskreten Zusammenhang in Abschnitt 2.1 ist $\mathbb{E}[X \mid \mathcal{A}]$ der erwartete Wert von X unter der Voraussetzung, dass uns sämtliche Information aus der σ-Algebra $\mathcal{A}$ zur Verfügung steht. Über diese Ereignisse wird nicht mehr gemittelt, sie werden als deterministisch angesehen. Im Fall $\sigma(X) \subset \mathcal{A}$ kennen wir also X schon vollständig und haben $\mathbb{E}[X \mid \mathcal{A}] = X$, siehe (iii). Im anderen Extrem, wenn X und $\mathcal{A}$ unabhängig sind, gibt uns die Kenntnis von $\mathcal{A}$ keine relevante Information über X, und wir haben $\mathbb{E}[X \mid \mathcal{A}] = \mathbb{E}[X]$, siehe (vii). ◇

Bemerkung 7.3.10. Man zeige durch ein Gegenbeispiel, dass durchaus $\mathbb{E}[\mathbb{E}[X \mid \mathcal{A}] \mid \mathcal{A}_1] \neq \mathbb{E}[\mathbb{E}[X \mid \mathcal{A}_1] \mid \mathcal{A}]$ gelten kann, wenn nämlich die σ-Algebren $\mathcal{A}_1$ und $\mathcal{A}$ nicht ineinander enthalten sind. ◇

Bei quadratisch integrierbaren Zufallsgrößen können wir die bedingte Erwartung sogar als eine Orthogonalprojektion auffassen. Wir erinnern, dass wir nicht an allen Stellen den Zusatz 'fast sicher' einbauen; insbesondere machen wir keinen Unterschied zwischen dem Raum $\mathcal{L}^2(\Omega, \mathcal{A}, \mathbb{P})$ und dem Hilbertraum $L^2(\Omega, \mathcal{A}, \mathbb{P})$, siehe Abschnitt 6.7.

Lemma 7.3.11 (Bedingte Erwartung als Orthogonalprojektion)**.** *Sei $X \in \mathcal{L}^2(\mathbb{P})$, und sei $\mathcal{A} \subset \mathcal{F}$ eine Teil-σ-Algebra. Dann ist $\mathbb{E}[X \mid \mathcal{A}]$ die Orthogonalprojektion von X auf $\mathcal{L}^2(\Omega, \mathcal{A}, \mathbb{P})$. Mit anderen Worten, es gilt für jede $\mathcal{A}$-messbare Zufallsgröße $Y \in \mathcal{L}^2(\mathbb{P})$:*

$$\mathbb{E}[(X - Y)^2] \geq \mathbb{E}\big[(X - \mathbb{E}[X \mid \mathcal{A}])^2\big]$$

mit Gleichheit genau dann, wenn $Y = \mathbb{E}[X \mid \mathcal{A}]$.

Beweis. Übungsaufgabe. □

Beispiel 7.3.12. Seien $X_1, \dots, X_N$ unabhängige zentrierte Zufallsgrößen. Wir betrachten die Irrfahrt $S_n = X_1 + \dots + X_n$ und die σ-Algebren $\mathcal{A}_n = \sigma(X_1, \dots, X_n)$ für $n = 1, \dots, N$. Dann gilt für $n \geq m$

$$\mathbb{E}[S_n \mid \mathcal{A}_m] = \mathbb{E}[X_1 \mid \mathcal{A}_m] + \dots + \mathbb{E}[X_n \mid \mathcal{A}_m] = X_1 + \dots + X_m + \mathbb{E}[X_{m+1}] + \dots + \mathbb{E}[X_n] = S_m.$$

Damit haben wir gezeigt, dass die Folge $(S_n)_{n=1,\dots,N}$ ein sogenanntes $(\mathcal{A}_n)_{n=1,\dots,N}$-*Martingal* ist. (Die Theorie der Martingale ist ein wichtiger Teilbereich der Theorie der Stochastischen Prozesse.) Nach Satz 7.3.8(iv) ist wegen $\sigma(S_m) \subset \mathcal{A}_m$ auch

$$\mathbb{E}[S_n \mid S_m] = \mathbb{E}[\mathbb{E}[S_n \mid \mathcal{A}_m] \mid S_m] = \mathbb{E}[S_m \mid S_m] = S_m.$$

Die natürliche Erweiterung der Formel von Bayes aus Lemma 2.1.4 ist die folgende:

Lemma 7.3.13 (Bayes'sche Formel)**.** *Für alle $A \in \mathcal{A}$ und $B \in \mathcal{F}$ mit $\mathbb{P}(B) > 0$ gilt*

$$\mathbb{P}(A \mid B) = \frac{\int_A \mathbb{P}(B \mid \mathcal{A}) \,\mathrm{d}\mathbb{P}}{\int \mathbb{P}(B \mid \mathcal{A}) \,\mathrm{d}\mathbb{P}} = \frac{\mathbb{E}\big[\mathbb{1}_A \mathbb{P}(B \mid \mathcal{A})\big]}{\mathbb{E}\big[\mathbb{P}(B \mid \mathcal{A})\big]}.$$

Beweis. Dies folgt leicht aus der zweiten definierenden Eigenschaft in Definition 7.3.2. □

Der Spezialfall, dass $\mathcal{A}$ von einer höchstens abzählbaren Familie paarweise disjunkter Mengen mit positiven Wahrscheinlichkeiten erzeugt wird, ist gerade Lemma 2.1.4.

Bedingte Wahrscheinlichkeiten haben etliche Eigenschaften, die analog sind zu jenen, die Wahrscheinlichkeitsmaße haben, und wir werden gleich ein paar wichtige davon behandeln. Man behalte aber im Hinterkopf, dass bedingte Wahrscheinlichkeiten Zufallsgrößen sind und dass daher die Beweise der jeweiligen Eigenschaft aus der Definition 7.3.2 heraus geführt werden müssen.

Satz 7.3.14 (Bedingte Markov-Ungleichung). *Für jede monoton steigende messbare Abbildung $\varphi\colon (0,\infty) \to (0,\infty)$ gilt*

$$\mathbb{P}(|X| \geq \varepsilon \mid \mathcal{A}) \leq \frac{\mathbb{E}[\varphi(|X|) \mid \mathcal{A}]}{\varphi(\varepsilon)}, \qquad \varepsilon > 0.$$

Beweis. Wie im Beweis der gewöhnlichen Markov-Ungleichung in Satz 5.1.1 nutzt man aus, dass fast sicher gilt: $\mathbb{1}_{\{|X|\geq\varepsilon\}} \leq \varphi(|X|)/\varphi(\varepsilon)$. Der Rest des Beweises folgt aus der Monotonie und Linearität der bedingten Erwartung. □

Satz 7.3.15 (Bedingte Jensen-Ungleichung). *Sei $I \subset \mathbb{R}$ ein Intervall und $\varphi\colon I \to \mathbb{R}$ konvex, und sei $X \in \mathcal{L}^1(\mathbb{P})$ eine I-wertige Zufallsgröße. Dann gilt*

$$\varphi\big(\mathbb{E}[X \mid \mathcal{A}]\big) \leq \mathbb{E}[\varphi(X) \mid \mathcal{A}] \leq \infty.$$

Beweis. Dies ist eine Anpassung des Beweises von Satz 6.7.2. Wir sagen, dass eine Aussage A auf einem Ereignis B fast sicher gelte, wenn $\mathbb{P}(B \setminus A) = 0$.

Auf dem Ereignis $\{\mathbb{E}[X \mid \mathcal{A}]$ ist ein Randpunkt von $I\}$ ist X fast sicher konstant gleich diesem Randpunkt. Dies sieht man wie folgt ein. Sei o. B. d. A. 0 der linke Randpunkt von I, also gilt $X \geq 0$ fast sicher. Daher ist auch $\mathbb{E}[X \mid \mathcal{A}] \geq 0$ fast sicher. Nun haben wir

$$\mathbb{E}[X] = \mathbb{E}\big[\mathbb{E}[X \mid \mathcal{A}]\big] = \mathbb{E}\big[\mathbb{E}[X \mid \mathcal{A}]\mathbb{1}_{\{\mathbb{E}[X|\mathcal{A}]>0\}}\big] = \mathbb{E}\big[X\mathbb{1}_{\{\mathbb{E}[X|\mathcal{A}]>0\}}\big],$$

da das Ereignis $\{\mathbb{E}[X \mid \mathcal{A}] > 0\}$ in $\mathcal{A}$ liegt. Durch Differenzbildung der obigen beiden Seiten folgt $\mathbb{E}[X\mathbb{1}_{\{\mathbb{E}[X|\mathcal{A}]=0\}}] = 0$, also auch $\mathbb{P}(X > 0, \mathbb{E}[X \mid \mathcal{A}] = 0) = 0$. Dies heißt, dass $X = 0$ fast sicher auf dem Ereignis $\{\mathbb{E}[X \mid \mathcal{A}] = 0\}$ gilt. Auf diesem Ereignis ist $\mathbb{E}[\varphi(X) \mid \mathcal{A}]$ fast sicher gleich der Konstanten $\varphi(0)$, denn die eventuelle Abänderung der Funktion $\omega \mapsto \mathbb{E}[\varphi(X) \mid \mathcal{A}](\omega)$ auf einer Nullmenge auf $\{\mathbb{E}[X \mid \mathcal{A}] = 0\}$ ändert nichts an der Eigenschaft der bedingten Erwartung von $\varphi(X)$ gegeben $\mathcal{A}$, wie man leicht nachrechnet. Daher ist die behauptete Ungleichung trivial auf dem Ereignis $\{\mathbb{E}[X \mid \mathcal{A}]$ ist ein Randpunkt von $I\}$.

Wir betrachten das Ereignis $B = \{\mathbb{E}[X \mid \mathcal{A}]$ ist ein innerer Punkt von $I\}$. Für $y \in I^\circ$ sei $D^+\varphi(y) = \max\{t \in \mathbb{R}\colon \varphi(x) - \varphi(y) \geq (x-y)t \,\forall x \in I\}$ die maximale Tangentensteigung von φ in y, d. h. die Linksableitung von φ in y. Also gilt $\varphi(x) \geq (x-y)D^+\varphi(y) + \varphi(y)$ für jedes $x \in I$. Wegen Konvexität ist $D^+\varphi$ steigend, also messbar, und daher ist $D^+\varphi(\mathbb{E}[X \mid \mathcal{A}])$ eine $\mathcal{A}$-messbare Zufallsgröße. Es folgt fast sicher auf B:

$$\mathbb{E}[\varphi(X) \mid \mathcal{A}] \geq \mathbb{E}\Big[\big(X - \mathbb{E}[X \mid \mathcal{A}]\big)D^+\varphi(\mathbb{E}[X \mid \mathcal{A}]) + \varphi\big(\mathbb{E}[X \mid \mathcal{A}]\big) \,\Big|\, \mathcal{A}\Big] = \varphi\big(\mathbb{E}[X \mid \mathcal{A}]\big)$$

fast sicher auf B. □

Korollar 7.3.16. *Sei $p \in [1,\infty]$, dann ist die Abbildung $\mathcal{L}^p(\Omega,\mathcal{F},\mathbb{P}) \to \mathcal{L}^p(\Omega,\mathcal{A},\mathbb{P})$, definiert durch $X \mapsto \mathbb{E}[X \mid \mathcal{A}]$, eine Kontraktion (d. h. $\|\mathbb{E}[X \mid \mathcal{A}]\|_p \le \|X\|_p$) und insbesondere stetig. Insbesondere gilt für Folgen $(X_n)_{n\in\mathbb{N}}$ im $\mathcal{L}^p(\Omega,\mathcal{F},\mathbb{P})$ mit $\lim_{n\to\infty}\|X_n - X\|_p = 0$ auch $\lim_{n\to\infty}\|\mathbb{E}[X_n \mid \mathcal{A}] - \mathbb{E}[X \mid \mathcal{A}]\|_p = 0$.*

Beweis. Für $p \in [1,\infty)$ benutze man die bedingte Jensensche Ungleichung für $\varphi(x) = |x|^p$. Für $p = \infty$ benutze man $|\mathbb{E}[X \mid \mathcal{A}]| \le \mathbb{E}[|X| \mid \mathcal{A}] \le \mathbb{E}[\|X\|_\infty \mid \mathcal{A}] = \|X\|_\infty$. □

Bemerkung 7.3.17 (Verkleinerung der Varianz)**.** Als eine Übungsaufgabe folgere man, dass für jede quadratisch integrierbare Zufallsgröße X die Varianz von $\mathbb{E}[X \mid \mathcal{A}]$ nicht größer ist als die von X. ◇

Ebenfalls kann man den folgenden wichtigen Satz als Folgerung aus der bedingten Jensenschen Ungleichung beweisen:

Satz 7.3.18 (Bedingte Cauchy-Schwarz-Ungleichung)**.** *Für alle $X, Y \in \mathcal{L}^2(\mathbb{P})$ gilt*

$$\mathbb{E}[XY \mid \mathcal{A}]^2 \le \mathbb{E}[X^2 \mid \mathcal{A}]\,\mathbb{E}[Y^2 \mid \mathcal{A}].$$

Beweis. Übungsaufgabe. □

Im Folgenden wollen wir uns endlich der Frage nähern, die in Bemerkungen 7.3.3 und 7.3.4 aufgeworfen wurde. Ein wichtiges technisches Hilfsmittel wird dabei die folgende Tatsache über die Messbarkeit von Hintereinanderausführungen sein. Wir erinnern uns daran, dass $\overline{\mathcal{B}}$ die von $\mathcal{B}$ und $\{\infty,-\infty\}$ auf $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty,-\infty\}$ erzeugte σ-Algebra ist.

Lemma 7.3.19 (Faktorisierungslemma)**.** *Seien $(\Omega,\mathcal{F})$ und $(\Omega',\mathcal{F}')$ zwei Messräume und $f\colon \Omega \to \Omega'$ messbar. Dann ist eine Abbildung $g\colon \Omega \to \overline{\mathbb{R}}$ genau dann messbar bezüglich $\sigma(f)$ und $\overline{\mathcal{B}}$, wenn es eine messbare Abbildung $\varphi\colon \Omega' \to \overline{\mathbb{R}}$ gibt mit $g = \varphi \circ f$. Wenn f surjektiv ist, ist φ eindeutig festgelegt, und wir schreiben dann $\varphi = g \circ f^{-1}$.*

Beweis. Wegen Lemma 6.4.4 ist nur die Richtung '$\Longrightarrow$' zu zeigen. Mit Hilfe einer Zerlegung in Positiv- und Negativteil sieht man, dass es reicht, die Aussage für nichtnegative g zu zeigen. Sei also $g\colon \Omega \to [0,\infty]$ messbar bezüglich $\sigma(f)$ und $\overline{\mathcal{B}}$, dann gibt es $A_1, A_2, \dots \in \sigma(f)$ und $\alpha_1, \alpha_2, \dots \in [0,\infty)$ mit $g = \sum_{n=1}^\infty \alpha_n 1\!\!1_{A_n}$. Dies sieht man ein, indem man die einfache Funktion $f_n = (2^{-n}\lfloor 2^n f\rfloor) \wedge n$ aus dem Beweis von Lemma 6.4.10 betrachtet (man erinnere sich, dass $f_n \uparrow f$ für $n \to \infty$) und beachtet, dass

$$f_n - f_{n-1} = \sum_{i=1}^{2^n} i2^{-n} 1\!\!1_{\{f_n - f_{n-1} = i2^{-n}\}}.$$

Wegen $A_n \in \sigma(f)$ gibt es ein $B_n \in \mathcal{F}'$ mit $f^{-1}(B_n) = A_n$, also mit $1\!\!1_{A_n} = 1\!\!1_{B_n} \circ f$. Nun leistet $\varphi = \sum_{n=1}^\infty \alpha_n 1\!\!1_{B_n}$ das Gewünschte. Die Eindeutigkeit bei Surjektivität sieht man leicht ein. □

Wir weisen noch darauf hin, dass im Falle der Surjektivität die Schreibweise $\varphi = g \circ f^{-1}$ nicht etwa heißen soll, dass eine Inverse von f existiert, sondern dass gemeint ist: $\varphi(\omega') = g(\omega)$ mit irgendeinem $\omega \in \Omega$, das $f(\omega) = \omega'$ erfüllt.

Nun sei X eine E-wertige Zufallsvariable, wobei $(E, \mathcal{E})$ ein beliebiger messbarer Raum ist. Unser Ziel ist, für jedes $x \in E$ ein Wahrscheinlichkeitsmaß $\mathbb{P}(\cdot \mid X = x)$ anzugeben, sodass für jedes $A \in \mathcal{F}$ auf dem Ereignis $\{X = x\}$ gilt: $\mathbb{P}(A \mid X) = \mathbb{P}(A \mid X = x)$. Die Schwierigkeit besteht darin, dass $\{X = x\}$ im Allgemeinen eine Nullmenge ist. Wir werden die Aufgabe im Allgemeinen auch nicht für jedes einzelne x lösen, sondern nur (in einem gewissen Sinne) für fast alle x.

Definition 7.3.20. *Sei $Y \in \mathcal{L}^1(\mathbb{P})$ und X eine E-wertige Zufallsvariable, wobei $(E, \mathcal{E})$ ein beliebiger messbarer Raum ist. Dann ist $Z = \mathbb{E}[Y \mid X]$ messbar bezüglich $\sigma(X)$ und $\mathcal{B}$. Nach Lemma 7.3.19 existiert eine $\mathcal{E}-\mathcal{B}$-messbare Abbildung $\varphi\colon E \to \mathbb{R}$ mit $Z = \varphi(X)$. Diese Abbildung φ heißt die* bedingte Erwartung *von Y gegeben $X = x$, kurz $\varphi(x) = \mathbb{E}[Y \mid X = x]$.*

Wir setzen weiterhin $\mathbb{P}(A \mid X = x) = \mathbb{E}[\mathbb{1}_A \mid X = x]$ für $A \in \mathcal{F}$.

Falls X surjektiv ist, so ist die bedingte Erwartung von Y gegeben $X = x$ eindeutig festgelegt, und wir können $\varphi = Z \circ X^{-1}$ schreiben.

Wir sind nun schon soweit, dass wir $\mathbb{P}(Y \in B \mid X = x) = \mathbb{E}[1_{\{Y \in B\}} \mid X = x)$ erklären können für jede einzelne messbare Menge $B \in \mathcal{E}'$, wenn Y eine Zufallsgröße mit Werten in einem beliebigen messbaren Raum $(E', \mathcal{E}')$ ist. Nun entsteht die natürliche Frage, ob die Kollektion der Zufallsvariablen $(\mathbb{P}(Y \in B \mid X = x))_{B \in \mathcal{E}'}$ aufgefasst werden kann als ein (durch x indiziertes, also gewissermaßen zufälliges) Wahrscheinlichkeitsmaß. Dieses Maß könnte man dann die bedingte Verteilung von Y gegeben $\{X = x\}$ nennen. Diese Frage liegt tief, denn diese Zufallsvariablen $\mathbb{P}(Y \in B \mid X = x)$ sind jeweils nur über fast sichere Eigenschaften definiert worden, und die jeweilige Ausnahmemenge hängt von B ab. Da wir von überabzählbar vielen B reden, ist es zunächst unklar, ob man sich auf *eine einzige* Ausnahmemenge mit Maß Null zurückziehen kann, die nicht von B abhängt. Um diese Situation genau zu fassen, benötigen wir den folgenden Begriff.

Definition 7.3.21 (Markovkern)**.** *Es seien $(E_1, \mathcal{E}_1)$ und $(E_2, \mathcal{E}_2)$ zwei messbare Räume. Ein* Markovkern *K von $(E_1, \mathcal{E}_1)$ nach $(E_2, \mathcal{E}_2)$ ist eine Abbildung $K\colon E_1 \times \mathcal{E}_2 \to [0,1]$ mit den folgenden Eigenschaften:*

(i) Für jedes $x \in E_1$ ist $K(x, \cdot)$ ein Wahrscheinlichkeitsmaß auf $(E_2, \mathcal{E}_2)$.

(ii) Für jedes $A \in \mathcal{E}_2$ ist $K(\cdot, A)$ $\mathcal{E}_1$-messbar.

Wir sagen auch kurz, K sei ein Kern von E_1 nach E_2.

Beispiel 7.3.22. (a) Mit einem Wahrscheinlichkeitsmaß ν auf $(E_2, \mathcal{E}_2)$ wird ein Markovkern K durch $K(x, A) = \nu(A)$ definiert.

(b) Mit einer messbaren Abbildung $f\colon E_1 \to E_2$ wird ein Markovkern K durch $K(x, A) = \mathbb{1}_A(f(x))$ definiert.

(c) Es sei P eine stochastische Matrix auf einer abzählbaren Menge I (d. h. $P = (p_{i,j})_{i,j \in I}$ mit $p_{i,j} \geq 0$ und $\sum_{j \in I} p_{i,j} = 1$ für jedes $i \in I$), und es sei $(E_1, \mathcal{E}_1) = (E_2, \mathcal{E}_2) = (I, \mathcal{P}(I))$, dann wird durch $K(i, A) = \sum_{j \in A} p_{i,j}$ ein Markovkern definiert.

◇

Bemerkung 7.3.23. Es reicht in Definition 7.3.21 die Eigenschaft (ii) nur für Mengen A aus einem durchschnittstabilen Erzeugendensystem von $\mathcal{E}_2$ zu fordern, denn die Menge

$$\{A \in \mathcal{E}_2 \colon K(\cdot, A) \text{ ist } \mathcal{E}_1\text{-messbar}\}$$

ist ein Dynkin-System, wie man leicht sieht. Siehe Satz 6.2.3.

Betreffend die Frage, die wir vor Definition 7.3.21 stellten, wünschen wir uns also einen Markovkern $K = K_{Y,\mathcal{A}}$ mit $K(x,B) = \mathbb{P}(Y \in B \mid X = x)$ für (fast) alle $x \in E$ und alle $B \in \mathcal{E}'$, wobei das 'fast' nicht von B abhängen soll, genau wie in Definition 7.3.21. Wir formalisieren dies nun.

Definition 7.3.24 (Reguläre Version der bedingten Verteilung)**.** *Sei Y eine Zufallsvariable mit Werten in einem messbaren Raum $(E',\mathcal{E}')$, und sei $\mathcal{A} \subset \mathcal{F}$ eine Teil-σ-Algebra. Ein Markovkern $K_{Y,\mathcal{A}}$ von $(\Omega,\mathcal{A})$ nach $(E',\mathcal{E}')$ heißt eine* reguläre Version der bedingten Verteilung *von Y gegeben $\mathcal{A}$, falls $K_{Y,\mathcal{A}}(\omega,B) = \mathbb{P}(Y \in B \mid \mathcal{A})(\omega)$ für $\mathbb{P}$-fast alle $\omega \in \Omega$ und für alle $B \in \mathcal{E}'$ gilt. (Man beachte, dass das 'fast' nicht von B abhängt!)*

Sei speziell $\mathcal{A} = \sigma(X)$ für eine Zufallsvariable X mit Werten in einem messbaren Raum $(E,\mathcal{E})$. Dann heißt der Markovkern $K_{Y,X}$ definiert durch $K_{Y,X}(x,B) = \mathbb{P}(Y \in B \mid X = x) = K_{Y,\sigma(X)}(X^{-1}(x),B)$ (die Funktion φ aus Lemma 7.3.19 mit beliebiger Festsetzung für $x \notin X(\Omega)$, wobei $f = X$ und $g = K_{Y,\sigma(X)}(\,\cdot\,,B)$) eine reguläre Version der bedingten Verteilung *von Y gegeben X.*

Falls er existiert, ist also der Markovkern $K_{Y,X}$ genau das Objekt, nach dem wir vor Definition 7.3.21 fragten. Nun beweisen wir seine Existenz (sogar die Existenz eines Kernes $K_{Y,\mathcal{A}}$), allerdings unter der Einschränkung, dass Y seine Werte in einem polnischen Raum E' annimmt, d. h. in einem separablen[2] vollständigen metrischen Raum. (Die σ-Algebra, die wir auf so einem Raum betrachten, ist natürlich die Borel-σ-Algebra $\mathcal{E}'$, die von dem System der offenen Mengen erzeugt wird.) Diese Einschränkung ist allerdings recht harmlos, denn die meisten Wertebereiche von Interesse sind polnisch, wie etwa alle separablen Banachräume oder die Menge aller stetigen Abbildungen $K \to \mathbb{R}$ auf einem Kompaktum $K \subset \mathbb{R}^d$, zusammen mit der Supremumsnorm. Es ist anschaulich relativ klar, dass wir eine Annahme machen müssen, die es uns erlaubt, die σ-Algebra $\mathcal{F}$ mit nur abzählbar vielen messbaren Mengen in geeigneter Weise zu approximieren (wie es gerade in polnischen Räumen möglich ist), denn wir haben ja oben gesehen, dass es darauf ankommt, überabzählbar viele Nullmengen zu einer Nullmenge zu vereinigen.

Satz 7.3.25. *Sei die Situation von Definition 7.3.24 gegeben. Falls $(E',\mathcal{E}')$ ein polnischer Raum ist (d. h. $\mathcal{E}'$ die Borel-σ-Algebra), so existiert eine reguläre Version $K_{Y,\mathcal{A}}$ der bedingten Verteilungen $\mathbb{P}(Y \in \cdot \mid \mathcal{A})$.*

Beweis. Wir geben hier nur den Beweis für $(E',\mathcal{E}') = (\mathbb{R},\mathcal{B})$. Im allgemeinen Fall konstruiert man einen Maßraum-Isomorphismus von $(\mathbb{R},\mathcal{B})$ in den polnischen Raum, siehe [Du02, Theorem 13.1.1].

Für $r \in \mathbb{Q}$ sei $F(r,\cdot)$ eine Version der bedingten Wahrscheinlichkeit $\mathbb{P}(Y \in (-\infty,r] \mid \mathcal{A})$. Da die Abbildung $r \mapsto \mathbb{1}_{\{Y \in (-\infty,r]\}}$ steigend ist, gibt es nach Satz 7.3.8(ii) (Monotonie der bedingten Erwartung) für jedes $r \leq s$ eine Nullmenge $A_{r,s} \in \mathcal{A}$ mit $F(r,\omega) \leq F(s,\omega)$ für jedes $\omega \in \Omega \setminus A_{r,s}$. Nach Satz 7.3.8(viii) (Satz von der majorisierten Konvergenz) gibt es eine Nullmenge $C \in \mathcal{A}$ und für jedes $r \in \mathbb{Q}$ eine Nullmenge $B_r \in \mathcal{A}$, sodass

$$\lim_{n\to\infty} F\Big(r + \frac{1}{n},\omega\Big) = F(r,\omega), \qquad \omega \in \Omega \setminus B_r,$$

[2]Ein topologischer Raum heißt *separabel*, wenn er eine abzählbare dichte Teilmenge enthält.

und

$$\lim_{n\to\infty} F(-n,\omega) = 0 \qquad \text{und} \qquad \lim_{n\to\infty} F(n,\omega) = 1, \qquad \omega \in \Omega \setminus C.$$

Sei N die Vereinigung aller Nullmengen $A_{r,s}$, B_r und C mit $s, r \in \mathbb{Q}$ und $r \leq s$, dann ist ja N auch eine Nullmenge. Für $\omega \in \Omega \setminus N$ ist dann

$$F(z,\omega) = \inf\{F(r,\omega) \colon r \in \mathbb{Q}, r \geq z\}, \qquad z \in \mathbb{R},$$

wohldefiniert. Für $\omega \in N$ setze $F(z,\omega) = F_0(z)$, wobei F_0 eine beliebige Verteilungsfunktion ist. Nach Konstruktion ist $F(\cdot,\omega)$ für jedes $\omega \in \Omega$ eine Verteilungsfunktion. Dies ist für $\omega \in N$ klar, und für $\omega \notin N$ ist zunächst klar, dass $F(\cdot,\omega)$ steigend ist auf $\mathbb{R}$ mit Grenzwerten 0 und 1 bei $-\infty$ bzw. ∞ (denn $\omega \notin C$). Die Rechtsstetigkeit in einem beliebigen $z \in \mathbb{R}$ sieht man wie folgt. Wegen Monotonie ist $F(z,\omega) \leq \lim_{r\downarrow z} F(r,\omega)$ für jedes $z \in \mathbb{R}$. Falls $z \in \mathbb{Q}$, so folgt die Rechtsstetigkeit von $F(\cdot,\omega)$ in z daraus, dass $\omega \notin B_r$, zusammen mit der Monotonie von $F(\cdot,\omega)$. Sei $z \in \mathbb{R} \setminus \mathbb{Q}$, und sei $(\varepsilon_n)_{n\in\mathbb{N}}$ eine Nullfolge in $(0,\infty)$, und sei $\delta > 0$. Auf Grund der Definition von $F(z,\omega)$ gibt es ein $r \in \mathbb{Q}$ mit $r > z$ und $F(z,\omega) \geq F(r,\omega) - \delta$. Es gilt $z + \varepsilon_n < r$ für alle genügend großen n, also auch $F(r,\omega) \geq F(z+\varepsilon_n,\omega)$, und daraus folgt $F(z,\omega) \geq \liminf_{n\to\infty} F(z+\varepsilon_n,\omega) - \delta$. Nun kann man $\delta \downarrow 0$ gehen lassen und erhält die Rechtsstetigkeit von $F(\cdot,\omega)$ in z.

Nach Satz 6.3.2 gibt es für jedes $\omega \in \Omega$ ein Wahrscheinlichkeitsmaß $K(\omega,\cdot)$ auf $(\Omega,\mathcal{F})$, dessen Verteilungsfunktion $F(\cdot,\omega)$ ist. Nun müssen wir noch zeigen, dass K ein Markovkern ist (hierfür ist nur die Messbarkeit von $K(\cdot,B)$ für jedes $B \in \mathcal{B}$ zu zeigen) und dass K eine Version der bedingten Verteilungen ist.

Für $r \in \mathbb{Q}$ und $B = (-\infty, r]$ ist dann

$$K(\omega,B) = \mathbb{P}(Y \in B \mid \mathcal{A})(\omega)\, \mathbb{1}_{N^c}(\omega) + F_0(r)\mathbb{1}_N(\omega). \tag{7.3.4}$$

Insbesondere ist $\omega \mapsto K(\omega,B)$ messbar. Da $\{(-\infty,r] \colon r \in \mathbb{Q}\}$ ein durchschnittstabiler Erzeuger von $\mathcal{B}$ ist, gilt die Messbarkeit sogar für jedes $B \in \mathcal{B}$ (siehe Bemerkung 7.3.23). Also ist K ein Markovkern. Dieser Kern ist auch eine Version der bedingten Verteilungen: Für $A \in \mathcal{A}$, $r \in \mathbb{Q}$ und $B = (-\infty,r]$ ist nach (7.3.4)

$$\int_A K(\omega,B)\,\mathbb{P}(\mathrm{d}\omega) = \int_A \mathbb{P}(Y \in B \mid \mathcal{A})\,\mathrm{d}\mathbb{P} = \mathbb{P}(A \cap \{Y \in B\}).$$

Als Funktion von B sind beide Seiten endliche Maße, die auf dem System $\{(-\infty,r] \colon r \in \mathbb{Q}\}$ übereinstimmen. Nach dem Eindeutigkeitssatz 6.2.8 gilt die Gleichung sogar für jedes $B \in \mathcal{B}$. Also ist $K(\cdot,B) = \mathbb{P}(Y \in B \mid \mathcal{A})$ $\mathbb{P}$-fast sicher. Daraus folgt, dass $K = K_{Y,\mathcal{A}}$. □

Satz 7.3.25 besagt, dass man, wenn Y Werte in einem polnischen Raum $(E',\mathcal{E}')$ annimmt, die Abbildungen $\omega \mapsto \mathbb{P}(Y \in B \mid \mathcal{A})(\omega)$ mit $B \in \mathcal{E}'$ jeweils auf Nullmengen so abändern kann, dass sie übereinstimmen mit einer messbaren Abbildung $\omega \mapsto K(\omega,B)$, die in B für jedes ω ein Wahrscheinlichkeitsmaß auf $(E',\mathcal{E}')$ ist. Dies beantwortet also die in Bemerkung 7.3.3 aufgeworfene Frage. Um zu prüfen, ob ein gegebener Markovkern K diese Eigenschaft hat, muss man also nur prüfen, ob für jedes B die Abbildung $\omega \mapsto K(\omega,B)$ ein bedingter Erwartungswert von Y gegeben $\mathcal{A}$ ist.

Beispiel 7.3.26. Sind X und Y unabhängige reelle Zufallsgrößen, so ist für $\mathbb{P} \circ X^{-1}$-fast alle $x \in \mathbb{R}$

$$\mathbb{P}(X + Y \in \cdot \mid X = x) = \delta_x * (\mathbb{P} \circ Y^{-1}).$$

Mit anderen Worten: Gegeben $X = x$, ist die bedingte Verteilung von $X+Y$ gleich der von $x+Y$. Diesen Sachverhalt vermuteten wir schon lange, und wir konnten ihn auch in dem Fall beweisen, wo $\{X = x\}$ positive Wahrscheinlichkeit hat, aber die korrekte Formulierung im allgemeinen Fall erfordert den Apparat, den wir soeben einführten. $\diamond$

Beispiel 7.3.27. Die motivierende Frage aus Bemerkung 7.3.4 können wir nun auch korrekt behandeln. Sei also X auf $[0,1]$ uniform verteilt, und gegeben $\{X = x\}$ seien $Y_1,\dots,Y_n$ unabhängige Bernoulli-verteilte Zufallsgrößen mit Parameter x. Die Zufallsvariable $Y = (Y_1,\dots,Y_n)$ hat Werte in dem polnischen Raum $\mathbb{R}^n$. Also existieren die regulären bedingten Verteilungen

$$K_{Y,X}(x,\cdot) = \mathbb{P}(Y \in \cdot \mid X = x), \qquad x \in [0,1].$$

Die Verteilung $\mathbb{P}(Y \in \cdot \mid X = x)$ ist natürlich identisch mit dem n-fachen Produkt des Bernoulli-Maßes mit Parameter x. $\diamond$

Beispiel 7.3.28. Seien Z_1 und Z_2 zwei unabhängige normalverteilte reelle Zufallsgrößen mit Erwartungswerten $\mu_1,\mu_2 \in \mathbb{R}$ und Varianzen $\sigma_1^2,\sigma_2^2 \in (0,\infty)$, siehe Beispiel 4.3.7 und Lemma 4.3.8. Dann kann man als eine Übungsaufgabe zeigen, dass die reguläre bedingte Verteilung $\mathbb{P}(Z_1 \in \cdot \mid Z_1 + Z_2 = x)$ für $x \in \mathbb{R}$ existiert und für fast alle x gleich der Normalverteilung mit Erwartungswert $\mu_1 + \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2}(x - \mu_1 - \mu_2)$ und Varianz $\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}$ ist. $\diamond$

Beispiel 7.3.29. Seien X und Y zwei reelle Zufallsvariablen mit gemeinsamer (Lebesgue-)Dichte $f\colon \mathbb{R}^2 \to [0,\infty)$. Wir betrachten die *Randdichte* bezüglich X, die gegeben ist durch

$$f_X(x) = \int_{\mathbb{R}} f(x,y)\,\lambda(\mathrm{d}y), \qquad x \in \mathbb{R}.$$

Dann ist f_X positiv $\mathbb{P}\circ X^{-1}$-fast sicher. Wir behaupten, dass die reguläre Version der bedingten Verteilung von Y gegeben X die Dichte

$$\frac{\mathbb{P}(Y \in \mathrm{d}y \mid X = x)}{\mathrm{d}y} = f_{Y|X}(x,y) = \frac{f(x,y)}{f_X(x)} \qquad \text{für } \mathbb{P}\circ X^{-1}\text{-fast alle } x \in \mathbb{R},$$

besitzt. Um dies zu zeigen, benutzt man zunächst den Satz 6.10.5 von Fubini, um zu sehen, dass die Abbildung $x \mapsto \int_B f_{Y|X}(x,y)\,\lambda(\mathrm{d}y)$ für jedes $B \in \mathcal{B}$ messbar ist. Ferner überzeugt man sich davon, dass für alle $A, B \in \mathcal{B}$ gilt:

$$\begin{aligned}
\int_A \mathbb{P}(X \in \mathrm{d}x) \int_B f_{Y|X}(x,y)\,\lambda(\mathrm{d}y) &= \int_A \mathbb{P}(X \in \mathrm{d}x)\frac{1}{f_X(x)} \int_B f(x,y)\,\lambda(\mathrm{d}y)\\
&= \int_A \lambda(\mathrm{d}x) \int_B f(x,y)\,\lambda(\mathrm{d}y) = \int_{A\times B} f(x,y)\,\lambda(\mathrm{d}(x,y))\\
&= \mathbb{P}(X \in A, Y \in B).
\end{aligned}$$

$\diamond$

Um sinnvoll mit regulären bedingten Verteilungen umgehen zu können, muss man natürlich noch wissen, wie man bezüglich ihrer integriert:

Satz 7.3.30. *Sei Y eine Zufallsvariable mit Werten in einem polnischen Raum $(E',\mathcal{E}')$. Sei $\mathcal{A} \subset \mathcal{F}$ eine Teil-σ-Algebra und $K_{Y,\mathcal{A}}$ eine reguläre Version der bedingten Verteilungen $\mathbb{P}(Y \in \cdot \mid \mathcal{A})$. Sei ferner $f\colon E' \to \mathbb{R}$ messbar mit $\mathbb{E}[|f(Y)|] < \infty$. Dann gilt*

$$\mathbb{E}[f(Y) \mid \mathcal{A}](\omega) = \int f(y)\,K_{Y,\mathcal{A}}(\omega,\mathrm{d}y) \qquad \textit{für } \mathbb{P}\textit{-fast alle } \omega \in \Omega. \tag{7.3.5}$$

Beweisskizze. Es reicht, $f \geq 0$ voraus zu setzen. Mit Hilfe einer Approximation durch einfache Funktionen (dies geht ähnlich wie im Beweis des Satzes 6.10.2 über Produktmaße) zeigt man, dass die rechte Seite von (7.3.5) als Funktion von ω messbar bezüglich $\mathcal{A}$ ist. Dies zeigt, dass die erste definierende Eigenschaft der rechten Seite als bedingte Erwartung von $f(Y)$ gegeben $\mathcal{A}$ erfüllt ist. Im Folgenden zeigen wir die zweite.

Wie wir im Beweis des Faktorisierungslemmas 7.3.19 gesehen haben, gibt es messbare Mengen $A_1, A_2, \dots \in \mathcal{E}'$ und $\alpha_1, \alpha_2, \dots \in [0, \infty)$, sodass die Funktion $f_n = \sum_{i=1}^n \alpha_i \mathbb{1}_{A_i}$ punktweise monoton gegen f konvergiert. Für jedes $n \in \mathbb{N}$ und jedes $B \in \mathcal{A}$ errechnet man

$$\begin{aligned}
\mathbb{E}[f_n(Y)\mathbb{1}_B] &= \sum_{i=1}^n \alpha_i \mathbb{P}(\{Y \in A_i\} \cap B) = \sum_{i=1}^n \alpha_i \int_B \mathbb{P}(\{Y \in A_i\} \mid \mathcal{A})(\omega)\, \mathbb{P}(\mathrm{d}\omega) \\
&= \sum_{i=1}^n \alpha_i \int_B K_{Y,\mathcal{A}}(\omega, A_i)\, \mathbb{P}(\mathrm{d}\omega) = \int_B \sum_{i=1}^n \alpha_i K_{Y,\mathcal{A}}(\omega, A_i)\, \mathbb{P}(\mathrm{d}\omega) \\
&= \int_B \Big(\int f_n(y)\, K_{Y,\mathcal{A}}(\omega, \mathrm{d}y) \Big)\, \mathbb{P}(\mathrm{d}\omega).
\end{aligned}$$

Nach dem Satz von der monotonen Konvergenz konvergiert für fast jedes ω das innere Integral auf der rechten Seite gegen $\int f(y)\, K_{Y,\mathcal{A}}(\omega, \mathrm{d}y)$. Erneute Anwendung dieses Satzes liefert

$$\mathbb{E}[f(Y)\mathbb{1}_B] = \lim_{n\to\infty} \mathbb{E}[f_n(Y)\mathbb{1}_B] = \int_B \int f(y)\, K_{Y,\mathcal{A}}(\omega, \mathrm{d}y)\, \mathbb{P}(\mathrm{d}\omega).$$

□

Man beachte, dass Satz 7.3.30 einen großen Fortschritt gegenüber der Situation nach der Einführung bedingter Erwartungen darstellt, denn *a priori* wussten wir zunächst nur, dass $\mathbb{E}[f(Y) \mid \mathcal{A}]$ eine fast sicher definierte Zufallsgröße ist, und Ausnahmemengen hingen von f ab. Nun können wir auch leicht die folgende wichtige Ungleichung folgern.

Korollar 7.3.31 (Bedingte Hölder'sche Ungleichung)**.** *Seien $p, q \in (1, \infty)$ mit $\frac{1}{p} + \frac{1}{q} = 1$, und seien $X \in \mathcal{L}^p(\mathbb{P})$ und $Y \in \mathcal{L}^q(\mathbb{P})$. Sei $\mathcal{A} \subset \mathcal{F}$ eine Teil-σ-Algebra. Dann gilt*

$$\mathbb{E}[|XY| \mid \mathcal{A}] \leq \mathbb{E}[|X|^p \mid \mathcal{A}]^{\frac{1}{p}}\, \mathbb{E}[|Y|^q \mid \mathcal{A}]^{\frac{1}{q}}.$$

Beweis. Leichte Übungsaufgabe. □

Kapitel 8

Konvergenzbegriffe

In diesem Kapitel behandeln wir die wichtigsten wahrscheinlichkeitstheoretischen Konzepte von Konvergenz und ihre Beziehungen untereinander. Grenzwerte von Zufallsgrößen (bzw. derer Verteilungen) werden in Abschnitt 8.1 betrachtet, die schwache Konvergenz von Wahrscheinlichkeitsmaßen in Abschnitt 8.2, und eines der wichtigsten Hilfsmittel, die charakteristischen Funktionen, in Abschnitt 8.3, wo wir auch den Zentralen Grenzwertsatz beweisen.

8.1 Konvergenz von Zufallsgrößen

Im Folgenden sei $(X_n)_{n\in\mathbb{N}}$ eine Folge von Zufallsgrößen $X_n\colon \Omega \to \mathbb{R}$, die alle auf dem selben Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, \mathbb{P})$ definiert sind. In der Wahrscheinlichkeitstheorie sind die folgenden drei Konvergenzbegriffe besonders wichtig:

Definition 8.1.1 (Konvergenzbegriffe). *(a) Die Folge* $(X_n)_{n\in\mathbb{N}}$ konvergiert fast sicher *gegen eine Zufallsgröße X, falls*

$$\mathbb{P}\big(\{\omega \in \Omega\colon \lim_{n\to\infty} X_n(\omega) = X(\omega)\}\big) = 1.$$

Wir schreiben dann $X_n \to X$ $\mathbb{P}$-fast sicher. (Die Menge $\{\lim_{n\to\infty} X_n = X\}$ ist messbar, wie aus Lemma 6.4.5 leicht folgt.)

(b) Es sei $p > 0$, und es gelte $X_n \in \mathcal{L}^p(\mathbb{P})$ für jedes n. Dann konvergiert $(X_n)_{n\in\mathbb{N}}$ im p-ten Mittel *gegen eine Zufallsgröße $X \in \mathcal{L}^p(\mathbb{P})$, falls*

$$\lim_{n\to\infty} \mathbb{E}(|X_n - X|^p) = \lim_{n\to\infty} \|X_n - X\|_p^p = 0.$$

(c) Die Folge $(X_n)_{n\in\mathbb{N}}$ konvergiert in Wahrscheinlichkeit *oder* stochastisch *gegen eine Zufallsgröße X, falls*

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0, \qquad \textit{für jedes } \varepsilon > 0.$$

Man beachte, dass für die fast sichere Konvergenz alle Zufallsgrößen $X_1, X_2, X_3, \dots$ 'gleichzeitig' auf einem Wahrscheinlichkeitsraum definiert sein müssen, während die anderen beiden

Begriffe nur von der Verteilung von $X_n - X$ für alle $n \in \mathbb{N}$ abhängen. Man sieht als eine leichte Übungsaufgabe, dass alle drei Konvergenzarten den Grenzwert bis auf fast sichere Gleichheit festlegen.

Die Beziehungen zwischen diesen Konvergenzarten sind die folgenden.

Lemma 8.1.2. *(a) Fast sichere Konvergenz impliziert Konvergenz in Wahrscheinlichkeit.*

(b) Konvergenz im p-ten Mittel impliziert Konvergenz in Wahrscheinlichkeit.

Beweis. (a) Es sei $Y_n = \mathbb{1}_{\{|X_n - X| > \varepsilon\}}$ für ein $\varepsilon > 0$. Wenn $X_n \to X$ fast sicher, so gilt $Y_n \to 0$ fast sicher. Wegen $|Y_n| \leq 1$ folgt aus dem Satz 6.8.6 von Lebesgue, dass $\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{E}(Y_n) \to 0$.

(b) folgt direkt aus der Markov-Ungleichung (siehe Satz 5.1.1) mit $\varphi(x) = x^p$. □

Die anderen denkbaren Implikationen sind nicht richtig, wie die folgenden Beispiele belegen:

Beispiel 8.1.3. Es sei $(\Omega, \mathcal{F}, \mathbb{P}) = ([0,1], \mathcal{B}_{[0,1]}, \lambda)$.

(a) Es sei $p > 0$. Betrachte $X_n = n^{1/p} \mathbb{1}_{[0,1/n]}$ für $n \in \mathbb{N}$. Dann gilt $X_n \to 0$ fast sicher und in Wahrscheinlichkeit, aber $\mathbb{E}(|X_n|^p) = 1$ für alle $n \in \mathbb{N}$, also konvergiert $(X_n)_{n \in \mathbb{N}}$ nicht im p-ten Mittel gegen $X = 0$.

(b) Ist $n = 2^m + k$ für $m \in \mathbb{N}_0$ und $0 \leq k < 2^m$, so setzt man $X_n = \mathbb{1}_{[k2^{-m},(k+1)2^{-m}]}$. Offenbar konvergiert $(X_n(\omega))_{n \in \mathbb{N}}$ für kein $\omega \in \Omega$. Andererseits gelten $\mathbb{P}(|X_n| > \varepsilon) \leq 2^{-m}$ für jedes $\varepsilon > 0$ und $\mathbb{E}(|X_n|^p) = 2^{-m}$ für jedes $p > 0$. Also konvergiert $(X_n)_{n \in \mathbb{N}}$ gegen $X = 0$ in Wahrscheinlichkeit und im p-ten Mittel. ◇

Unter einer Majorantenbedingung impliziert die fast sichere Konvergenz die Konvergenz im p-ten Mittel:

Lemma 8.1.4. *Sei $(X_n)_{n \in \mathbb{N}}$ eine Folge von Zufallsgrößen, die fast sicher gegen eine Zufallsgröße X konvergiert. Falls eine Zufallsgröße $Y \in \mathcal{L}^p(\mathbb{P})$ für ein $p > 0$ existiert, so dass $|X_n| \leq Y$ fast sicher gilt, so gilt $X_n \to X$ im p-ten Mittel.*

Beweis. Es gelten $|X_n - X|^p \leq (|X_n| + |X|)^p \leq (2Y)^p \leq 2^p Y^p \in \mathcal{L}^1(\mathbb{P})$ und $|X_n - X|^p \to 0$ fast sicher. Also folgt aus dem Satz 6.1.8 von Lebesgue, dass $\mathbb{E}(|X_n - X|^p) \to 0$. □

Eines der wichtigsten Hilfsmittel bei fast sicheren Konvergenzaussagen ist der folgende Satz 8.1.5. Zunächst bemerken wir, dass für eine Folge $(A_n)_{n \in \mathbb{N}}$ von Ereignissen $A_n \in \mathcal{F}$ die Mengen

$$\limsup_{n \to \infty} A_n = \bigcap_{n \in \mathbb{N}} \bigcup_{m \in \mathbb{N}\colon m \geq n} A_m \qquad \text{und} \qquad \liminf_{n \to \infty} A_n = \bigcup_{n \in \mathbb{N}} \bigcap_{m \in \mathbb{N}\colon m \geq n} A_m \tag{8.1.1}$$

in $\mathcal{F}$ liegen. Offensichtlich fällt die Folge $(\bigcup_{m \in \mathbb{N}\colon m \geq n} A_m)_{n \in \mathbb{N}}$ gegen $\limsup_{n \to \infty} A_n$, und die Folge $(\bigcap_{m \in \mathbb{N}\colon m \geq n} A_m)_{n \in \mathbb{N}}$ steigt gegen $\liminf_{n \to \infty} A_n$. Man sieht leicht, dass

$$\begin{aligned} \limsup_{n \to \infty} A_n &= \{\omega \in \Omega\colon \omega \in A_n \text{ für unendlich viele } n \in \mathbb{N}\}, \\ \liminf_{n \to \infty} A_n &= \{\omega \in \Omega\colon \omega \in A_n \text{ für fast alle } n \in \mathbb{N}\}. \end{aligned} \tag{8.1.2}$$

Satz 8.1.5 (1. Borel-Cantelli-Lemma). *Es sei $(A_n)_{n\in\mathbb{N}}$ eine Folge von Ereignissen mit $\sum_{n\in\mathbb{N}} \mathbb{P}(A_n) < \infty$. Dann gilt $\mathbb{P}(\limsup_{n\to\infty} A_n) = 0$.*

Beweis. Sei $B_m = \bigcup_{k\in\mathbb{N}\colon k\geq m} A_k$. Wegen $B_m \downarrow \limsup_{n\to\infty} A_n$ und Satz 6.1.9(a) und (b) gilt

$$\mathbb{P}\Big(\limsup_{n\to\infty} A_n\Big) = \lim_{m\to\infty} \mathbb{P}(B_m) \leq \lim_{m\to\infty} \sum_{k=m}^{\infty} \mathbb{P}(A_k) = 0.$$

□

Das 1. Borel-Cantelli-Lemma besitzt eine partielle Umkehrung, siehe Satz 8.1.10. Es folgen ein paar Anwendungen des 1. Borel-Cantelli-Lemmas, zunächst auf das fast sichere asymptotische Verhalten von identisch verteilten Zufallsgrößen:

Lemma 8.1.6. *Es sei $(X_n)_{n\in\mathbb{N}}$ eine Folge identisch verteilter integrierbarer Zufallsgrößen. Dann gilt $\lim_{n\to\infty} X_n/n = 0$ $\mathbb{P}$-fast sicher.*

Beweis. Es gilt

$$\Big\{\lim_{n\to\infty} \frac{X_n}{n} = 0\Big\}^{\mathrm{c}} = \bigcup_{j=1}^{\infty} \limsup_{n\to\infty} \{j\,|X_n| > n\},$$

denn eine Folge $(y_n)_{n\in\mathbb{N}}$ konvergiert genau dann nicht gegen Null, wenn es eine positive Zahl $1/j$ mit $j \in \mathbb{N}$ gibt, so dass $|y_n| > 1/j$ für unendlich viele n gilt. Nach dem 1. Borel-Cantelli-Lemma reicht es nun zu zeigen, dass für jedes j die Wahrscheinlichkeiten der Mengen $\{j\,|X_n| > n\}$ über $n \in \mathbb{N}$ summierbar sind. Man sieht leicht, dass dies gilt:

$$\sum_{n\in\mathbb{N}} \mathbb{P}(j\,|X_n| > n) = \sum_{n\in\mathbb{N}} \mathbb{P}(j\,|X_1| > n) = \int \sum_{n\in\mathbb{N}} 1\!\!1_{\{j\,|X_1|>n\}}\,\mathrm{d}\mathbb{P} \leq j\,\mathbb{E}(|X_1|) < \infty.$$

□

Mit Hilfe von Satz 8.1.5 können wir hinreichende Kriterien für fast sichere Konvergenz entlang einer Teilfolge angeben:

Satz 8.1.7. *Sei $(X_n)_{n\in\mathbb{N}}$ eine Folge von Zufallsgrößen, die in Wahrscheinlichkeit gegen eine Zufallsgröße X konvergiert. Dann existiert eine Teilfolge $(n_k)_{k\in\mathbb{N}}$, so dass $X_{n_k} \to X$ für $k \to \infty$ fast sicher.*

Beweis. Zu jedem $k \in \mathbb{N}$ existiert nach Voraussetzung ein $n_k \in \mathbb{N}$ mit $\mathbb{P}(|X_{n_k} - X| > 1/k) \leq 1/k^2$. Wir können annehmen, dass $(n_k)_{k\in\mathbb{N}}$ strikt wachsend ist. Da die Reihe $\sum_{k\in\mathbb{N}} 1/k^2$ konvergiert, folgt $\mathbb{P}(\limsup_{k\to\infty}\{|X_{n_k} - X| > 1/k\}) = 0$ nach Satz 8.1.5. Für $\omega \notin \limsup_{k\to\infty}\{|X_{n_k} - X| > 1/k\}$ gilt $|X_{n_k} - X| \leq 1/k$ für alle genügend großen k, also gilt $\lim_{k\to\infty} X_{n_k}(\omega) = X(\omega)$ für alle ω, die nicht in der Nullmenge $\limsup_{k\to\infty}\{|X_{n_k} - X| > 1/k\}$ liegen. □

Alle drei Konvergenzbegriffe in Definition 8.1.1 sind vollständig. Dies ist klar für die fast sichere Konvergenz, denn $\mathbb{R}$ ist vollständig. Die Vollständigkeit der Konvergenz im p-ten Mittel beweisen wir nicht (wir haben sie in Satz 6.7.1 für $p \geq 1$ erwähnt), siehe etwa [Ba91, Paragraf 15]. Für die Konvergenz in Wahrscheinlichkeit beweisen wir die Vollständigkeit nun mit Hilfe von Satz 8.1.5:

Satz 8.1.8. *Sei $(X_n)_{n\in\mathbb{N}}$ eine Folge von Zufallsgrößen mit*

$$\lim_{m,n\to\infty} \mathbb{P}(|X_n - X_m| > \varepsilon) = 0 \qquad \text{für jedes } \varepsilon > 0.$$

Dann existiert eine Zufallsgröße X mit $X_n \to X$ in Wahrscheinlichkeit.

Beweis. Ähnlich wie im Beweis von Satz 8.1.7 wählt man eine Teilfolge $(n_k)_{k\in\mathbb{N}}$ mit

$$\mathbb{P}(|X_{n_k} - X_{n_{k+1}}| \geq 1/k^2) \leq 1/k^2, \qquad k \in \mathbb{N}.$$

Wie oben folgt aus dem 1. Borel-Cantelli-Lemma

$$\mathbb{P}\Big(\limsup_{k\to\infty}\{|X_{n_k} - X_{n_{k+1}}| \geq 1/k^2\}\Big) = 0.$$

Für $\omega \notin \limsup_{k\to\infty}\{|X_{n_k} - X_{n_{k+1}}| \geq 1/k^2\}$ ist $(X_{n_k}(\omega))_{k\in\mathbb{N}}$ eine Cauchy-Folge in $\mathbb{R}$. Wegen dessen Vollständigkeit konvergiert diese Folge gegen ein $X(\omega)$ für diese ω, d. h. $X_{n_k} \to X$ fast sicher für $k \to \infty$, nach Lemma 8.1.2 auch in Wahrscheinlichkeit. Für jedes $\varepsilon > 0$ gilt

$$\mathbb{P}(|X_m - X| > \varepsilon) \leq \mathbb{P}\Big(|X_m - X_{n_k}| > \frac{\varepsilon}{2}\Big) + \mathbb{P}\Big(|X_{n_k} - X| > \frac{\varepsilon}{2}\Big), \qquad m, k \in \mathbb{N}.$$

Wählt man k als die kleinste Zahl mit $n_k \geq m$, so folgt $\lim_{m\to\infty} \mathbb{P}(|X_m - X| > \varepsilon) = 0$. □

Als eine weitere Anwendung des 1. Borel-Cantelli-Lemmas (Satz 8.1.5) beweisen wir eine nicht optimale Version des Gesetzes der Großen Zahlen (siehe Korollar 10.2.22 für die übliche Version, in der auf die Voraussetzung $X_i \in \mathcal{L}^4$ verzichtet wird):

Satz 8.1.9 (Starkes Gesetz der Großen Zahlen)**.** *Es sei $(X_n)_{n\in\mathbb{N}}$ eine Folge unabhängiger identisch verteilter Zufallsgrößen $X_n \in \mathcal{L}^4$. Dann gilt $\lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^n X_j = \mathbb{E}(X_1)$ fast sicher.*

Beweis. Man sieht leicht, dass wir voraussetzen dürfen, dass $\mathbb{E}(X_1) = 0$ gilt. Wir kürzen $S_n = \sum_{j=1}^n X_j$ ab und betrachten das Ereignis $A_n = \{|\frac{1}{n}S_n| \geq n^{-1/8}\}$. Mit der Markov-Ungleichung (Satz 5.1.1) schätzen wir ab:

$$\mathbb{P}(A_n) \leq n^{1/2}\frac{1}{n^4}\mathbb{E}(S_n^4). \tag{8.1.3}$$

Der Erwartungswert berechnet sich zu

$$\begin{aligned}\mathbb{E}(S_n^4) &= \mathbb{E}\Big(\Big(\sum_{j=1}^n X_j\Big)^4\Big) = \mathbb{E}\Big(\sum_{j_1,j_2,j_3,j_4=1}^n X_{j_1}X_{j_2}X_{j_3}X_{j_4}\Big)\\ &= \sum_{j_1,j_2,j_3,j_4=1}^n \mathbb{E}\big(X_{j_1}X_{j_2}X_{j_3}X_{j_4}\big) = \sum_{j=1}^n \mathbb{E}(X_j^4) + \sum_{\substack{i,j=1\\ i\neq j}}^n \mathbb{E}(X_i^2X_j^2).\end{aligned}$$

Um die letzte Gleichung einzusehen, beachte man, dass alle Terme verschwinden, in denen ein Index verschieden von den anderen ist, denn mit Hilfe der Unabhängigkeit kann man den entsprechenden Erwartungswert isolieren, und er ist Null nach unserer Voraussetzung. Also haben wir $\mathbb{E}(S_n^4) = n\mathbb{E}(X_1^4) + n(n-1)\mathbb{E}(X_1^2)^2 \leq \mathcal{O}(n^2)$ für $n \to \infty$. Setzt man dies in (8.1.3) ein, erhält man sofort, dass $\sum_{n\in\mathbb{N}} \mathbb{P}(A_n) < \infty$. Nach dem 1. Borel-Cantelli-Lemma ist $\mathbb{P}(\limsup_{n\to\infty} A_n) = 0$.

Für jedes $\omega \in (\limsup_{n\to\infty} A_n)^{\mathrm{c}}$ gilt für alle genügend großen n, dass $|\frac{1}{n}S_n(\omega)| < n^{-1/8}$, also konvergiert $\frac{1}{n}S_n(\omega)$ gegen Null. $\square$

Nun kommt die angekündigte partielle Umkehrung des 1. Borel-Cantelli-Lemmas:

Satz 8.1.10 (2. Borel-Cantelli-Lemma). *Es seien $A_1, A_2, A_3, \dots$ unabhängige Ereignisse mit $\sum_{n\in\mathbb{N}} \mathbb{P}(A_n) = \infty$. Dann gilt $\mathbb{P}(\limsup_{n\to\infty} A_n) = 1$.*

Beweis. Für jedes $n \in \mathbb{N}$ gilt

$$\mathbb{P}\Big(\bigcup_{m=n}^{\infty} A_m\Big) = 1 - \mathbb{P}\Big(\bigcap_{m=n}^{\infty} A_m^{\mathrm{c}}\Big) = 1 - \lim_{k\to\infty} \mathbb{P}\Big(\bigcap_{m=n}^{k} A_m^{\mathrm{c}}\Big)$$
$$= 1 - \prod_{m=n}^{\infty} \mathbb{P}(A_m^{\mathrm{c}}) = 1 - \prod_{m=n}^{\infty} (1 - \mathbb{P}(A_m)).$$

Da die Reihe der $\mathbb{P}(A_m)$ divergiert, ist das Produkt $\prod_{m=n}^{\infty}(1-\mathbb{P}(A_m))$ gleich Null, denn für jedes $k \geq n$ gilt $\prod_{m=n}^{k}(1 - \mathbb{P}(A_m)) = \exp\{\sum_{m=n}^{k} \log(1 - \mathbb{P}(A_m))\} \leq \exp\{-\sum_{m=n}^{k} \mathbb{P}(A_m)\}$. Daher folgt $\mathbb{P}\Big(\bigcup_{m=n}^{\infty} A_m\Big) = 1$ für jedes $n \in \mathbb{N}$. Also folgt die Aussage des Satzes, denn die Mengen $\bigcup_{m=n}^{\infty} A_m$ fallen für $n \to \infty$ gegen $\limsup_{n\to\infty} A_n$. $\square$

Auf die Unabhängigkeit kann natürlich nicht verzichtet werden, wie man an der trivialen Wahl $A_n = A$ für alle $n \in \mathbb{N}$ mit $0 < \mathbb{P}(A) < 1$ sieht.

Wenn man unendlich viele unabhängige Versuche unternimmt, ein Ereignis mit positiver Wahrscheinlichkeit zu erreichen, wird man unendlich oft Erfolg haben:

Beispiel 8.1.11. Wenn ein Affe auf einer Schreibmaschine unendlich oft auf unabhängige Weise jede Taste mit gleicher Wahrscheinlichkeit bedient, dann ist die Wahrscheinlichkeit, dass er dabei unendlich oft den gesamten Bibeltext schreibt, gleich Eins.

Denn sei I die Menge aller Tasten, und sei μ die Gleichverteilung auf I, dann hat das Ergebnis der Bemühungen des Affen die Verteilung $\mathbb{P} = \mu^{\otimes\mathbb{N}}$ auf dem Folgenraum $I^{\mathbb{N}}$. Wenn der Bibeltext die Länge b hat, so betrachten wir das Ereignis A_n, dass der unendlich lange Text von der n-ten bis zur $(n + b - 1)$-ten Stelle genau den Bibeltext enthält. Die Ereignisse A_{bn} mit $n \in \mathbb{N}$ sind unabhängig, und ihre Wahrscheinlichkeit ist jeweils $|I|^{-b}$. Also divergiert die Reihe der $\mathbb{P}(A_n)$, und das 2. Borel-Cantelli-Lemma impliziert die Aussage. $\diamond$

Eine weitere Anwendung des 2. Borel-Cantelli-Lemmas zeigt, dass es unmöglich ist, im Starken Gesetz der Großen Zahlen auf die Integrierbarkeit der Summanden zu verzichten:

Lemma 8.1.12. *Es sei $(X_n)_{n\in\mathbb{N}}$ eine Folge unabhängiger identisch verteilter Zufallsgrößen, so dass $Y_n = \frac{1}{n}\sum_{j=1}^{n} X_j$ fast sicher gegen eine Zufallsvariable Y konvergiere. Dann ist jedes X_j integrierbar, und es gilt $Y = \mathbb{E}(X_1)$ fast sicher.*

Beweis. Da ja $Y_n \to Y$ fast sicher, konvergiert $\frac{1}{n}X_n = Y_n - \frac{n-1}{n}Y_{n-1}$ fast sicher gegen Null. Für das Ereignis $A_n = \{|X_n| \geq n\}$ gilt also $\mathbb{P}(\limsup_{n\to\infty} A_n) = 0$, denn das Ereignis A_n kann nicht mit positiver Wahrscheinlichkeit für unendlich viele n eintreten. Wegen der Unabhängigkeit der X_n ist auch die Folge der Ereignisse A_n unabhängig. Nach dem 2. Borel-Cantelli-Lemma gilt

$\sum_{n\in\mathbb{N}} \mathbb{P}(A_n) < \infty$. Wir schätzen ab:

$$\begin{aligned}\mathbb{E}(|X_1|) &= \int_{(0,\infty)} \mathbb{P}(|X_1| > t)\,\mathrm{d}t = \sum_{n=1}^\infty \int_{n-1}^n \mathbb{P}(|X_1| > t)\,\mathrm{d}t \leq \sum_{n=1}^\infty \mathbb{P}(|X_1| \geq n-1)\\ &= 1 + \sum_{n=1}^\infty \mathbb{P}(|X_1| \geq n) = 1 + \sum_{n=1}^\infty \mathbb{P}(A_n) < \infty.\end{aligned}$$

Also ist jedes X_j integrierbar. Nach dem Starken Gesetz der Großen Zahlen (hier müssen wir Korollar 10.2.22 zitieren) ist Y fast sicher konstant gleich $\mathbb{E}(X_1)$. □

Mit gemeinsamer Hilfe der beiden Borel-Cantelli-Lemmata kann man die exakte limsup-Asymptotik von Folgen unabhängiger Zufallsgrößen ermitteln. Wir beschränken uns auf einen Spezialfall, den wir als eine Übungsaufgabe dem Leser überlassen:

Lemma 8.1.13. *Für jede Folge unabhängiger, zum Parameter Eins exponentiell verteilter Zufallsgrößen $X_1, X_2, \dots$ gilt $\limsup_{n\to\infty} X_n/\log n = 1$ fast sicher.*

Eine andere Übungsaufgabe ist der Beweis der Transienz der unsymmetrischen eindimensionalen Irrfahrt:

Beispiel 8.1.14. Sei $p \in (0,1) \setminus \{\frac{1}{2}\}$, und seien $X_1, X_2, \dots$ unabhängige Zufallsgrößen, die die Werte 1 und -1 mit Wahrscheinlichkeit p bzw. $1-p$ annehmen. Dann nennt man die Folge der $S_n = X_1 + \dots + X_n$ (wobei $S_0 = 0$) eine unsymmetrische eindimensionale Irrfahrt. Diese Irrfahrt ist *transient* in dem Sinn, dass sie den Ursprung mit Wahrscheinlichkeit Eins nur endlich oft besucht. ◇

Wir streifen einen Integrierbarkeitsbegriff, der enge Zusammenhänge mit $\mathcal{L}^1$-Konvergenz aufweist:

Definition 8.1.15 (gleichgradig integrierbar)**.** *Eine Familie $(X_i)_{i\in I}$ von Zufallsgrößen (mit beliebiger Indexmenge I) heißt* gleichgradig integrierbar *oder* uniform integrabel*, falls*

$$\lim_{R\to\infty} \sup_{i\in I} \mathbb{E}[|X_i| \mathbb{1}_{\{|X_i|>R\}}] = 0.$$

Natürlich ist jede endliche Familie von integrierbaren Zufallsgrößen gleichgradig integrierbar, und gleichgradige Integrierbarkeit überträgt sich leicht auf Summen, Differenzen und Teilfamilien. Es wird sich gleich herausstellen (und das ist wohl die wichtigste Eigenschaft der gleichgradigen Integrierbarkeit), dass die gleichgradige Integrierbarkeit es erlaubt, von Konvergenz in Wahrscheinlichkeit auf $\mathcal{L}^1$-Konvergenz zu schließen, also die Implikation von Satz 8.1.2(b) für $p = 1$ zu einer Äquivalenz zu machen (wir erinnern an Beispiel 8.1.3). Wir bringen zunächst ein notwendiges und hinreichendes Kriterium für gleichgradige Integrierbarkeit:

Lemma 8.1.16. *Eine Familie $(X_i)_{i\in I}$ von Zufallsgrößen ist genau dann gleichgradig integrierbar, wenn $\sup_{i\in I} \mathbb{E}[|X_i|] < \infty$ gilt und es zu jedem $\varepsilon > 0$ ein $\delta > 0$ gibt mit $\mathbb{E}[|X_i| \mathbb{1}_A] < \varepsilon$ für jedes $i \in I$ und jede messbare Menge A mit $\mathbb{P}(A) < \delta$.*

Beweis. '$\Longrightarrow$': Sei $(X_i)_{i\in I}$ gleichgradig integrierbar. Dann ist für jedes $R > 0$ natürlich $\mathbb{E}[|X_i|] \leq \mathbb{E}[|X_i| \mathbb{1}_{|X_i|>R}] + R$, und der erste Summand verschwindet gleichmäßig in $i \in I$ für $R \to \infty$, also ist die erste Bedingung im Lemma erfüllt. Ferner gilt für jede messbare Menge A und für jedes $R > 0$ und $i \in I$:

$$\mathbb{E}[|X_i|\mathbb{1}_A] \leq \mathbb{E}[|X_i|\mathbb{1}_{|X_i|>R}] + \mathbb{E}[R\mathbb{1}_A] \leq \mathbb{E}[|X_i|\mathbb{1}_{|X_i|>R}] + R\mathbb{P}(A).$$

Sei nun $\varepsilon > 0$ gegeben, dann wählen wir R so groß, dass $\mathbb{E}[|X_i|\mathbb{1}_{|X_i|>R}] < \varepsilon/2$ für alle $i \in I$, und danach wählen wir $\delta = \varepsilon/(2R)$. Dann gilt nämlich für jedes A mit $\mathbb{P}(A) < \delta$ die gewünschte Ungleichung $\sup_{i\in I} \mathbb{E}[|X_i|\mathbb{1}_A] < \varepsilon$, wovon man sich leicht überzeugt.

'$\Longleftarrow$': Sei $\varepsilon > 0$, und sei $\delta > 0$ gewählt nach Voraussetzung. Dann gibt es ein $R > 0$, so dass $\mathbb{P}(|X_i| > R) < \delta$ für jedes $i \in I$, denn nach der Markov-Ungleichung gilt $\mathbb{P}(|X_i| > R) \leq \frac{1}{R}\mathbb{E}[|X_i|] \leq \frac{1}{R}\sup_{j\in I} \mathbb{E}[|X_j|]$. Also gilt für jedes $i \in I$ und die Menge $A_i = \{|X_i| > R\}$, dass $\mathbb{E}[|X_j|\mathbb{1}_{A_i}] < \varepsilon$ für jedes $j \in I$, also auch für $j = i$. Dies zeigt die gleichgradige Integrierbarkeit der Familie $(X_i)_{i\in I}$. □

Man zeigt als Übungsaufgabe, dass für jede integrierbare Zufallsgröße X und jede Familie Φ von Teil-σ-Algebren von $\mathcal{F}$ die Familie der bedingten Erwartungswerte $\{\mathbb{E}[X \mid \mathcal{G}] \colon \mathcal{G} \in \Phi\}$ gleichgradig integrierbar ist.

Der Beweis des folgenden hinreichenden Kriteriums ist ebenfalls eine Übungsaufgabe:

Korollar 8.1.17. *Eine Familie $(X_i)_{i\in I}$ von Zufallsgrößen ist gleichgradig integrierbar, wenn eine Funktion $H\colon [0,\infty) \to [0,\infty)$ existiert mit $\lim_{x\to\infty} H(x)/x = \infty$ und $\sup_{i\in I} \mathbb{E}[H(|X_i|)] < \infty$.*

Insbesondere ist jede Familie von Zufallsgrößen, die beschränkt in $\mathcal{L}^p(\mathbb{P})$ für ein $p > 1$ ist, gleichgradig integrierbar.

Nun kommt der angekündigte Zusammenhang zwischen gleichgradiger Integrierbarkeit und $\mathcal{L}^1$-Konvergenz:

Lemma 8.1.18. *Eine Folge $(X_n)_{n\in\mathbb{N}}$ von integrierbaren Zufallsgrößen konvergiert genau dann in $\mathcal{L}^1$, wenn sie in Wahrscheinlichkeit konvergiert und gleichgradig integrierbar ist.*

Beweis. '$\Longrightarrow$': Sei $(X_n)_{n\in\mathbb{N}}$ in $\mathcal{L}^1$ gegen eine Zufallsgröße X konvergent. Nach Lemma 8.1.2(b) liegt auch Konvergenz in Wahrscheinlichkeit vor. Nun zeigen wir, dass $(X_n)_{n\in\mathbb{N}}$ gleichgradig integrierbar ist. Leicht ist zu sehen, dass $K = \sup_{n\in\mathbb{N}} \mathbb{E}[|X_n|]$ endlich ist. Sei $\varepsilon > 0$, and sei $N_\varepsilon \in \mathbb{N}$ gewählt mit $\mathbb{E}[|X_n - X|] \leq \varepsilon/2$ für alle $n \geq N_\varepsilon$. Da X integrierbar ist, ist $\{X\}$ insbesondere gleichgradig integrierbar. Nach Lemma 8.1.16 existiert ein $\delta > 0$ mit $\mathbb{E}[|X|\mathbb{1}_A] < \varepsilon/2$ für alle messbaren Mengen A mit $\mathbb{P}(A) < \delta$. Nun wählen wir R so groß, dass $R > 2K/\delta$. Dann können wir die Eigenschaft von δ anwenden auf die Menge $A = \{|X_n| > R\}$, denn für alle $n \geq N_\varepsilon$ folgt mit Hilfe der Markov-Ungleichung, dass $\mathbb{P}(|X_n| > R) \leq \frac{1}{R}\mathbb{E}[|X_n|] \leq \delta$. Also können wir für diese n und R abschätzen:

$$\mathbb{E}[|X_n|\mathbb{1}_{\{|X_n|>R\}}] \leq \mathbb{E}[|X - X_n|\mathbb{1}_{\{|X_n|>R\}}] + \mathbb{E}[|X|\mathbb{1}_{\{|X_n|>R\}}] \leq \varepsilon/2 + \varepsilon/2 = \varepsilon.$$

Nach eventueller Vergrößerung von R gilt die Ungleichung $\mathbb{E}[|X_n|\mathbb{1}_{\{|X_n|>R\}}] \leq \varepsilon$ auch noch für $n = 1, \dots, N_\varepsilon$. Dies zeigt die gleichmäßige Integrierbarkeit.

'$\Longleftarrow$': Es gelte nun $X_n \to X$ in Wahrscheinlichkeit für eine Zufallsgröße X (die *a priori* nicht integrierbar sein muss), und $(X_n)_{n\in\mathbb{N}}$ sei gleichgradig integrierbar. Wir nehmen an, dass $(X_n)_{n\in\mathbb{N}}$ nicht in $\mathcal{L}^1$ gegen X konvergiert. Dann gibt es ein $\varepsilon > 0$ und eine Teilfolge $(X_{n_k})_{k\in\mathbb{N}}$ mit $\mathbb{E}[|X - X_{n_k}|] > 2\varepsilon$ für jedes $k \in \mathbb{N}$.

Nach Satz 8.1.7 können wir annehmen, dass $(X_{n_k})_{k\in\mathbb{N}}$ fast überall gegen X konvergiert. Nach dem Lemma von Fatou haben wir $\mathbb{E}[|X|] \leq \liminf_{k\to\infty} \mathbb{E}[|X_{n_k}|] < \infty$, also ist X integrierbar. Auch die Folge $(X - X_{n_k})_{k\in\mathbb{N}}$ ist gleichgradig integrierbar, also existiert ein $R > 0$ mit $\mathbb{E}[|X - X_{n_k}|\mathbb{1}_{\{|X-X_{n_k}|>R\}}] < \varepsilon/2$ für alle $k \in \mathbb{N}$. Die Folge $(|X - X_{n_k}|\mathbb{1}_{\{|X-X_{n_k}|\leq R\}})_{k\in\mathbb{N}}$ hingegen konvergiert fast sicher gegen Null und ist beschränkt durch R, also konvergiert ihr Erwartungswert nach dem Satz von Lebesgue gegen Null. Zusammen ergibt dies für alle genügend großen k, dass $\mathbb{E}[|X - X_{n_k}|] \leq \varepsilon$ im Widerspruch zur Annahme. □

8.2 Schwache Konvergenz und der Satz von Prohorov

Einer der natürlichsten und meistbenutzten Konvergenzbegriffe für Maße ist der der schwachen Konvergenz. In diesem Abschnitt führen wir ihn ein und charakterisieren ihn in verschiedener Weise. Insbesondere geben wir das nützliche Kompaktheitskriterium von Prohorov.

Zunächst stellen wir fest, was es heißen soll, dass Maße gegen einander konvergieren. Dabei ist es üblich, Maße auf einem metrischen Raum zu betrachten. Auf einem solchen betrachten wir wie immer die Borel-σ-Algebra, also die von dem System der offenen Mengen erzeugte σ-Algebra, siehe Bemerkung 6.4.7. Der meist benutzte Konvergenzbegriff ist der folgende.

Definition 8.2.1 (schwache Konvergenz). *Es sei $(E, \mathtt{d})$ ein metrischer Raum, und seien $\mathbb{P}, \mathbb{P}_1, \mathbb{P}_2, \mathbb{P}_3, \dots$ Wahrscheinlichkeitsmaße auf E. Wir sagen, dass $\mathbb{P}_n$ schwach für $n \to \infty$ gegen $\mathbb{P}$ konvergiert, und schreiben $\mathbb{P}_n \Longrightarrow \mathbb{P}$ oder $\mathbb{P} = \text{w-lim}_{n\to\infty} \mathbb{P}_n$, wenn für jede beschränkte stetige Funktion $f\colon E \to \mathbb{R}$ gilt: $\lim_{n\to\infty} \int f \,\mathrm{d}\mathbb{P}_n = \int f \,\mathrm{d}\mathbb{P}$.*

Anstelle von $\Longrightarrow$ benutzt man auch $\stackrel{\mathrm{w}}{\to}$, wobei das 'w' an *weak convergence* erinnert. Im Folgenden lassen wir den Term 'schwach' auch manchmal weg, denn wir werden nur diesen Konvergenzbegriff für Maße benutzen. Im Folgenden bezeichnen wir die Menge der Wahrscheinlichkeitsmaße auf E mit $\mathcal{M}_1(E)$ und die Menge der beschränkten stetigen Funktionen $E \to \mathbb{R}$ mit $\mathcal{C}_{\mathrm{b}}(E)$.

Bemerkung 8.2.2. (i) In topologischer Sprechweise ist die Topologie der schwachen Konvergenz, d. h. die *schwache Topologie*, die gröbste Topologie auf $\mathcal{M}_1(E)$, sodass die Funktionale $\mathbb{P} \mapsto \int_E f \,\mathrm{d}\mathbb{P}$ stetig sind für jedes beschränkte stetige $f\colon E \to \mathbb{R}$.

(ii) Falls E separabel ist, kann die schwache Topologie metrisiert werden mit der sogenannten *Prohorov-Metrik*

$$\varrho(\mathbb{P}_1, \mathbb{P}_2) = \inf\{\varepsilon > 0\colon \mathbb{P}_1(A) \leq \mathbb{P}_2(A^\varepsilon) + \varepsilon \text{ und } \mathbb{P}_2(A) \leq \mathbb{P}_1(A^\varepsilon) + \varepsilon \\ \text{für jedes messbare } A \subset E\},$$

wobei $A^\varepsilon = \{x \in E\colon \mathtt{d}(x, A) < \varepsilon\}$ die offene ε-Umgebung von A ist; siehe etwa [Bi68].

(iii) Man kann diesen Konvergenzbegriff leicht auf die Menge aller endlichen signierten Maße ausdehnen (was wir hier nicht machen wollen). Dann ist die schwache Topologie das, was man in der Funktionalanalysis die *Schwach-*-Topologie* nennt, denn die Menge aller endlichen signierten Maße ist der Dualraum der Menge der beschränkten stetigen Funktionen.

(iv) Im Gegensatz zur schwachen Topologie, die ja durch die Integrale gegen alle beschränkten stetigen Funktionen erzeugt wird, nennt man manchmal diejenige Topologie, die durch Integrale gegen alle beschränkten messbaren Funktionen erzeugt wird, die *starke Topologie*. $\mathbb{P}_n$ konvergiert dann also stark gegen $\mathbb{P}$, wenn $\lim_{n\to\infty} \mathbb{P}_n(A) = \mathbb{P}(A)$ für jede messbare Menge A gilt. Allerdings ist dieser Begriff zu stark für die meisten Anwendungen; eine starke Konvergenz liegt nur in ganz wenigen interessanten Fällen vor.

◇

Die folgenden Beispiele verifiziert man als Übungsaufgaben.

Beispiel 8.2.3. (i) Eine Folge $(x_n)_{n\in\mathbb{N}}$ in E konvergiert genau dann gegen ein $x \in E$, wenn die Folge der Diracmaße δ_{x_n} gegen δ_x konvergiert.

(ii) Es sei $E = \mathbb{R}$, dann konvergiert $\frac{1}{n}\sum_{i=1}^n \delta_{i/n}$ gegen die Einschränkung des Lebesgue-Maßes auf $[0, 1]$.

(iii) Die Familie $(\delta_n)_{n\in\mathbb{N}}$ konvergiert nicht schwach.

(iv) Die Familie der gleichförmigen Verteilungen auf $[-n, n]$ mit $n \in \mathbb{N}$ konvergiert nicht schwach.

◇

Wie bei jedem vernünftigen Konvergenzbegriff ist der schwache Grenzwert eindeutig:

Lemma 8.2.4 (Eindeutigkeit des schwachen Grenzwerts)**.** *Es sei $(E, \mathtt{d})$ ein metrischer Raum, und eine Folge $(\mathbb{P}_n)_{n\in\mathbb{N}}$ von Wahrscheinlichkeitsmaßen auf E schwach konvergent sowohl gegen $\mathbb{P}$ als auch gegen $\widetilde{\mathbb{P}}$. Dann gilt $\mathbb{P} = \widetilde{\mathbb{P}}$.*

Beweis. Wir haben also $\int_E f\,\mathrm{d}\mathbb{P} = \int_E f\,\mathrm{d}\widetilde{\mathbb{P}}$ für jede beschränkte stetige Funktion $f : E \to \mathbb{R}$. Wir zeigen, dass daraus folgt, dass $\mathbb{P}(A) = \widetilde{\mathbb{P}}(A)$ für jede abgeschlossene Teilmenge A von E gilt. Dies beendet den Beweis, da die Menge der abgeschlossenen Mengen ein durchschnittstabiles Erzeugendensystem der Borel-σ-Algebra ist, siehe den Eindeutigkeitssatz, bzw. Korollar 6.2.9.

Sei also $A \subset E$ abgeschlossen, und sei $\varepsilon > 0$. Mit $\mathtt{d}(x, A) = \inf\{\mathtt{d}(x, a) : a \in A\}$ bezeichnen wir den Abstand von x zur Menge A. Betrachte $\varphi(r) = (r \vee 0) \wedge 1$ für $r \in \mathbb{R}$, dann ist φ stetig und beschränkt. Daher ist auch die Funktion $f_\varepsilon : E \to [0, 1]$, definiert durch $f_\varepsilon(x) = 1 - \varphi(\frac{1}{\varepsilon}\mathtt{d}(x, A))$ stetig und beschränkt. Insbesondere gilt $\int_E f_\varepsilon\,\mathrm{d}\mathbb{P} = \int_E f_\varepsilon\,\mathrm{d}\widetilde{\mathbb{P}}$ für jedes $\varepsilon > 0$. Die Funktion f_ε ist konstant gleich 1 auf A und gleich Null außerhalb der ε-Umgebung von A. Ferner gilt $f_\varepsilon \downarrow \mathbb{1}_A$ für $\varepsilon \downarrow 0$, wofür wir die Abgeschlossenheit von A benutzen. Nach dem monotonen Konvergenzsatz folgt

$$\mathbb{P}(A) = \int_E \mathbb{1}_A\,\mathrm{d}\mathbb{P} = \lim_{\varepsilon\downarrow 0}\int_E f_\varepsilon\,\mathrm{d}\mathbb{P} = \lim_{\varepsilon\downarrow 0}\int_E f_\varepsilon\,\mathrm{d}\widetilde{\mathbb{P}} = \int_E \mathbb{1}_A\,\mathrm{d}\widetilde{\mathbb{P}} = \widetilde{\mathbb{P}}(A).$$

□

Wir haben schwache Konvergenz mit Hilfe von Testintegralen gegen beschränkte stetige Funktionen definiert. Aber man möchte natürlich auch gerne wissen, ob dann auch die Wahrscheinlichkeiten von Ereignissen konvergieren. Dies ist allerdings bei weitem nicht für alle Ereignisse richtig:

Bemerkung 8.2.5. Die Folge $(\delta_{1/n})_{n\in\mathbb{N}}$ der Punktmaße in $1/n$ konvergiert offensichtlich schwach gegen das Punktmaß δ_0. Für die abgeschlossene Menge $(-\infty, 0]$ gilt allerdings

$$\lim_{n\to\infty} \delta_{1/n}((-\infty,0]) = 0 < 1 = \delta_0((-\infty,0]),$$

was man damit kommentieren kann, dass am Rande Masse 'eingewandert' ist. Analog 'wandert' aus offenen Mengen Masse 'aus', wie man an dem folgenden Beispiel sieht: $\lim_{n\to\infty} \delta_{1/n}((0,\infty)) = 1 > 0 = \delta_0((0,\infty))$. Durch diese Eigenschaft werden wir gleich die schwachen Konvergenz charakterisieren. $\diamond$

Es folgen Charakterisierungen der schwachen Konvergenz. Wir erinnern daran, dass eine Funktion $f\colon E \to \mathbb{R}$ *Lipschitz-stetig* heißt, wenn ein $L > 0$ existiert mit $|f(x) - f(y)| \leq L\mathtt{d}(x,y)$ für alle $x, y \in E$. Mit $\mathcal{U}_f = \{x \in E\colon f \text{ ist nicht stetig in } x\}$ bezeichnen wir die Menge aller Unstetigkeitsstellen von f.[1] Mit $\mathbb{E}_n$ bezeichnen wir den Erwartungswert bezüglich des Wahrscheinlichkeitsmaßes $\mathbb{P}_n$.

Satz 8.2.6 (Portmanteau-Theorem)**.** *Es sei $(E, \mathtt{d})$ ein metrischer Raum, und es seien $\mathbb{P}$, $\mathbb{P}_1$, $\mathbb{P}_2$, $\mathbb{P}_3, \dots$ Wahrscheinlichkeitsmaße auf E. Dann sind folgende Aussagen äquivalent.*

(i) $\mathbb{P}_n \Longrightarrow \mathbb{P}$.

(ii) Für jede beschränkte Lipschitz-stetige Funktion $f\colon E \to \mathbb{R}$ gilt $\lim_{n\to\infty} \mathbb{E}_n(f) = \mathbb{E}(f)$.

(iii) Für jede beschränkte messbare Funktion $f\colon E \to \mathbb{R}$ mit $\mathbb{P}(\mathcal{U}_f) = 0$ gilt $\lim_{n\to\infty} \mathbb{E}_n(f) = \mathbb{E}(f)$.

(iv) Für jede abgeschlossene Menge $F \subset E$ gilt $\limsup_{n\to\infty} \mathbb{P}_n(F) \leq \mathbb{P}(F)$.

(v) Für jede offene Menge $G \subset E$ gilt $\liminf_{n\to\infty} \mathbb{P}_n(G) \geq \mathbb{P}(G)$.

(vi) Für jede messbare Menge $A \subset E$ mit $\mathbb{P}(\partial A) = 0$ gilt $\lim_{n\to\infty} \mathbb{P}_n(A) = \mathbb{P}(A)$.

Beweis. Die trivialen bzw. leicht einzusehenden Richtungen sind (iv) $\Longleftrightarrow$ (v) $\Longrightarrow$ (vi) und (iii) $\Longrightarrow$ (i) $\Longrightarrow$ (ii). Also reicht es, nur noch (ii) $\Longrightarrow$ (iv) und (vi) $\Longrightarrow$ (iii) zu zeigen.

'(ii) $\Longrightarrow$ (iv)': Sei $F \subset E$ abgeschlossen, und sei $\varepsilon > 0$. Wie im Beweis von Lemma 8.2.4 approximieren wir $\mathbb{1}_F$ mit $f_\varepsilon\colon E \to [0,1]$, definiert durch $f_\varepsilon(x) = 1 - \varphi(\frac{1}{\varepsilon}\mathtt{d}(x,F))$, wobei $\varphi\colon \mathbb{R} \to [0,1]$ durch $\varphi(r) = (r \vee 0) \wedge 1$ definiert ist. Tatsächlich ist $\mathtt{d}(\cdot, F)$ sogar Lipschitz-stetig mit Lipschitz-Konstante 1, wie man aus der Dreiecksungleichung folgert. Also ist die Abbildung f_ε ebenfalls Lipschitz-stetig. Sie ist auf F konstant gleich 1 und hat den Wert 0 für alle $x \in E$ mit $\mathtt{d}(x,F) \geq \varepsilon$. Wegen $\mathbb{1}_F \leq f_\varepsilon$ gilt $\mathbb{P}_n(F) = \mathbb{E}_n(\mathbb{1}_F) \leq \mathbb{E}_n(f_\varepsilon)$, also $\limsup_{n\to\infty} \mathbb{P}_n(F) \leq \limsup_{n\to\infty} \mathbb{E}_n(f_\varepsilon) = \mathbb{E}(f_\varepsilon)$ für jedes $\varepsilon > 0$. Da $f_\varepsilon \downarrow \mathbb{1}_F$ für $\varepsilon \downarrow 0$, folgt aus dem monotonen Konvergenzsatz die Behauptung.

'(vi) $\Longrightarrow$ (iii)': Sei $f\colon E \to \mathbb{R}$ beschränkt und messbar mit $\mathbb{P}(\mathcal{U}_f) = 0$. Zunächst überlegt man sich als eine (topologische) Übungsaufgabe, dass gilt:

$$\partial f^{-1}(B) \subset f^{-1}(\partial B) \cup \mathcal{U}_f, \qquad B \subset \mathbb{R}. \tag{8.2.1}$$

(Mit anderen Worten: Wenn man f anwendet auf einen Randpunkt des Urbilds von B und f ist stetig in diesem Punkt, dann erhält man einen Randpunkt von B.)

[1] Man kann sich als eine Übungsaufgabe klar machen, dass $\mathcal{U}_f$ messbar ist, wenn f messbar ist.

Setze $C_n = \{y \in \mathbb{R} \colon \mathbb{P}(f^{-1}(\{y\})) > \frac{1}{n}\}$. Dann ist C_n eine endliche Menge, und die Vereinigung $C = \bigcup_{n\in\mathbb{N}} C_n = \{y \in \mathbb{R} \colon \mathbb{P}(f^{-1}(\{y\})) > 0\}$ ist abzählbar.

Sei $\varepsilon > 0$. Dann gibt es ein $N \in \mathbb{N}$ und $y_0 \le -\|f\|_\infty < y_1 < \cdots < y_{N-1} < \|f\|_\infty < y_N$, so dass $y_i \in \mathbb{R} \setminus C$ und $|y_{i+1} - y_i| < \varepsilon$ für alle i. Sei $E_i = f^{-1}([y_{i-1}, y_i))$ für $i = 1, \dots, N$. Dann ist E die disjunkte Vereinigung der $E_1, \dots, E_N$, und wegen (8.2.1) haben wir

$$\mathbb{P}(\partial E_i) \le \mathbb{P}(f^{-1}(\{y_{i-1}\})) + \mathbb{P}(f^{-1}(\{y_i\})) + \mathbb{P}(\mathcal{U}_f) = 0, \qquad i = 1, \dots, N.$$

Daher gilt

$$\begin{aligned}\limsup_{n\to\infty} \mathbb{E}_n(f) &\le \limsup_{n\to\infty} \sum_{i=1}^N y_i \mathbb{P}_n(E_i) \le \varepsilon + \limsup_{n\to\infty} \sum_{i=1}^N y_{i-1} \mathbb{P}_n(E_i)\\ &\le \varepsilon + \sum_{i=1}^N y_{i-1} \mathbb{P}(E_i) \le \varepsilon + \mathbb{E}(f).\end{aligned}$$

Wir lassen $\varepsilon \downarrow 0$ und erhalten $\limsup_{n\to\infty} \mathbb{E}_n(f) \le \mathbb{E}(f)$. Durch Übergang zu $-f$ statt f erhalten wir die komplementäre Ungleichung, die den Beweis beendet. □

Bemerkung 8.2.7. Als eine Übungsaufgabe gebe man einen direkten Beweis der Implikation (vi) $\Longrightarrow$ (v). (Hinweis: Man betrachte $G_\delta = \{x \in G \colon \mathrm{d}(x, G^{\mathrm{c}}) > \delta\}$ für eine geeignete Folge von $\delta \downarrow 0$.) ◇

Mit Hilfe der schwachen Konvergenz erhalten wir (hinzu zu den Begriffen in Definition 8.1.1) einen weiteren Konvergenzbegriff für Zufallsgrößen:

Bemerkung 8.2.8 (Verteilungskonvergenz)**.** Wir sagen, dass eine Folge $(X_n)_{n\in\mathbb{N}}$ von E-wertigen Zufallsgrößen schwach gegen eine Zufallsgröße X konvergiert, wenn die jeweiligen Verteilungen dies tun, und dann schreiben wir auch $X_n \Longrightarrow X$. Hierzu sagt man auch, dass die Folge der Zufallsgrößen *in Verteilung* konvergiert. Daher nennt man diesen Begriff auch oft die *Verteilungskonvergenz*. Es gilt also $X_n \Longrightarrow X$ genau dann, wenn $\lim_{n\to\infty} \mathbb{E}(f(X_n)) = \mathbb{E}(f(X))$ für jede beschränkte stetige Funktion f gilt.[2] Man benutzt auch die Notation $X_n \xrightarrow{\mathcal{D}} X$ oder $X_n \xrightarrow{\mathcal{L}} X$, wobei '$\mathcal{D}$' und '$\mathcal{L}$' an *distribution* bzw. an *law* erinnern. ◇

Wir möchten die Beziehung der Verteilungskonvergenz zur Konvergenz in Wahrscheinlichkeit klären. Letzteren Begriff möchten wir daher auf beliebige E-wertige Zufallsgrößen ausdehnen. Es scheint klar, wie man ihn definieren muss: Eine Folge von E-wertigen Zufallsgrößen $X_1, X_2, \dots$ konvergiert in Wahrscheinlichkeit gegen eine Zufallsgröße X, wenn für jedes $\varepsilon > 0$ gilt: $\lim_{n\to\infty} \mathbb{P}(\mathtt{d}(X_n, X) > \varepsilon) = 0$. Obwohl das ganz harmlos klingt, gibt es einen Haken an dieser Definition: Das Ereignis $\{\mathtt{d}(X_n, X) > \varepsilon\}$ muss gar nicht messbar sein! Zwar ist klar, dass (X_n, X) eine $E \times E$-wertige Zufallsgröße ist und damit messbar bezüglich dem Produkt $\mathcal{B}_E \otimes \mathcal{B}_E$ der Borel-σ-Algebra auf E. Ferner ist leicht zu zeigen, dass $\mathtt{d} \colon E \times E \to [0, \infty)$ stetig ist. Aber aus Letzterem folgt nur, dass $\mathtt{d}$ messbar ist bezüglich der Borel-σ-Algebra $\mathcal{B}_{E\times E}$ auf $E \times E$. Falls also $\mathcal{B}_{E\times E} = \mathcal{B}_E \otimes \mathcal{B}_E$ gilt, so ist die Abbildung $\mathtt{d}(X_n, X)$ eine Zufallsgröße, also messbar. Dies gilt jedoch im Allgemeinen nur, wenn E separabel ist:

Lemma 8.2.9. *Wenn E separabel ist, so gilt $\mathcal{B}_{E\times E} = \mathcal{B}_E \otimes \mathcal{B}_E$.*

[2] Warnung: Hier ist $\mathbb{P}$ kein Maß auf E, sondern auf irgendeinem abstrakten messbaren Raum. Es handelt sich dann um die schwache Konvergenz der Bildmaße $\mathbb{P} \circ X_n^{-1}$ auf E.

Beweisskizze. Wir benötigen keine Separabilität, um zu zeigen, dass $\mathcal{B}_E \otimes \mathcal{B}_E \subset \mathcal{B}_{E\times E}$ gilt, denn für je zwei offene Mengen $A, B \subset E$ ist $A \times B$ offen in $E \times E$, und $\mathcal{B}_E \otimes \mathcal{B}_E$ wird ja per Definition von diesen Kreuzprodukten erzeugt.

Wenn E separabel ist, so gibt es eine abzählbare Basis $(U_i)_{i\in\mathbb{N}}$ der Topologie auf E, d. h. jede offene Menge in E ist eine Vereinigung von geeigneten dieser U_i. Per Definition der Produkt-Topologie ist dann $(U_i \times U_j)_{i,j\in\mathbb{N}}$ eine Basis der Topologie auf $E \times E$. Also ist jede offene Menge in $E \times E$ eine Vereinigung von geeigneten dieser $U_i \times U_j$ und liegt damit auch in $\mathcal{B}_E \otimes \mathcal{B}_E$. □

Wenn wir voraussetzen, dass E separabel ist, können wir also problemlos von Konvergenz in Wahrscheinlichkeit E-wertiger Zufallsgrößen sprechen. Klären wir also den Zusammenhang mit Verteilungskonvergenz:

Bemerkung 8.2.10 (Konvergenz in Verteilung und in Wahrscheinlichkeit)**.** Es sei E ein separabler metrischer Raum.

(i) Wenn eine Folge von E-wertigen Zufallsgrößen X_n in Wahrscheinlichkeit gegen eine Zufallsgröße X konvergiert, so gilt auch $X_n \Longrightarrow X$. Dies zeigen wir mit Hilfe des Kriteriums (ii) im Portmanteau-Theorem: Für jede beschränkte Lipschitz-stetige Funktion $f\colon E \to \mathbb{R}$ mit Lipschitz-Konstante L und für jedes $\varepsilon > 0$ gilt

$$\begin{aligned}\big|\mathbb{E}[f(X_n)] - \mathbb{E}[f(X)]\big| &\le \mathbb{E}\big[|f(X_n) - f(X)|\mathbb{1}_{\mathtt{d}(X_n,X)\le\varepsilon}\big] + 2\|f\|_\infty \mathbb{P}(\mathtt{d}(X_n,X) > \varepsilon)\\ &\le L\varepsilon + 2\|f\|_\infty \mathbb{P}(\mathtt{d}(X_n,X) > \varepsilon),\end{aligned}$$

und dies konvergiert gegen $L\varepsilon$ für $n \to \infty$. Da die linke Seite nicht von ε abhängt, folgt $\lim_{n\to\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)]$, also die schwache Konvergenz von X_n gegen X.

(ii) Die Umkehrung der Aussage in (i) gilt, falls die Limes-Zufallsgröße X fast sicher konstant ist (Übungsaufgabe), aber nicht im Allgemeinen: Wenn zum Beispiel die Zufallsgrößen $X, X_1, X_2, \dots$ unabhängig und identisch verteilt sind mit einer nichttrivialen Verteilung, so konvergiert $(X_n)_{n\in\mathbb{N}}$ trivialer Weise in Verteilung gegen X, aber nicht in Wahrscheinlichkeit.

◇

Verteilungskonvergenz überträgt sich auf Bilder unter fast überall stetigen Funktionen. Das folgende Ergebnis wird auch *Continuous mapping theorem* genannt.

Lemma 8.2.11. *Es seien $(E_1, \mathtt{d}_1)$ und $(E_2, \mathtt{d}_2)$ zwei metrische Räume und $h\colon E_1 \to E_2$ eine Borel-messbare Funktion, so dass die Unstetigkeitsmenge $\mathcal{U}_h$ messbar sei.*

(a) Seien $\mathbb{P}, \mathbb{P}_1, \mathbb{P}_2, \mathbb{P}_3, \dots$ Wahrscheinlichkeitsmaße auf E_1 mit $\mathbb{P}(\mathcal{U}_h) = 0$ und $\mathbb{P}_n \Longrightarrow \mathbb{P}$. Dann gilt $\mathbb{P}_n \circ h^{-1} \Longrightarrow \mathbb{P} \circ h^{-1}$.

(b) Sei $E_1 = E_2 = \mathbb{R}$, und seien $X, X_1, X_2, X_3, \dots$ Zufallsgrößen mit $\mathbb{P}(X \in \mathcal{U}_h) = 0$ und $X_n \Longrightarrow X$. Dann gilt auch $h(X_n) \Longrightarrow h(X)$.

Beweis. Der Beweis von (a) ist eine elementare Übungsaufgabe mit Hilfe des Portmanteau-Theorems, und (b) ist nur eine Umformulierung. □

Für Verteilungsfunktionen gibt es natürlich auch einen Konvergenzbegriff, der gut mit der schwachen Konvergenz der zugehörigen Maße zusammenpasst:

Definition 8.2.12 (schwache Konvergenz von Verteilungsfunktionen). *Seien F bzw. $F_1, F_2, \ldots$ Verteilungsfunktionen von Wahrscheinlichkeitsmaßen auf $\mathbb{R}$. Wir sagen, $(F_n)_{n\in\mathbb{N}}$* konvergiert schwach *gegen F, und wir schreiben $F_n \Longrightarrow F$ oder $F = \text{w-}\lim_{n\to\infty} F_n$, falls $F(x) = \lim_{n\to\infty} F_n(x)$ für alle $x \in \mathbb{R}$ gilt, in denen F stetig ist.*

Bemerkung 8.2.13. Verteilungsfunktionen von Wahrscheinlichkeitsmaßen können natürlich auch gegen Funktionen konvergieren, die keine solchen mehr sind; es kann Masse verschwinden. Wenn zum Beispiel G die Verteilungsfunktion eines Wahrscheinlichkeitsmaßes ist und $F_n(x) = G(x+n)$ ihre Verschiebung um n, so konvergiert zwar F_n in allen Punkten, aber dennoch nicht im Sinne der Definition 8.2.12, denn der Limes ist konstant. Das Gleiche gilt für die Verschiebungen $F_n(x) = G(x-n)$.

Falls $F, F_1, F_2, \ldots$ Verteilungsfunktionen von endlichen Maßen mit Gesamtmasse ≤ 1 sind (sogenannten *Subwahrscheinlichkeitsmaßen*), dann sagen wir, dass $(F_n)_{n\in\mathbb{N}}$ schwach gegen F konvergiert, wenn $F(x) = \lim_{n\to\infty} F_n(x)$ für alle $x \in \mathbb{R}$ gilt, in denen F stetig ist, *und zusätzlich* noch $F(\infty) - F(-\infty) \ge \limsup_{n\to\infty}(F_n(\infty) - F_n(-\infty))$ gilt. (Hierbei haben wir natürlich $F(\infty) = \lim_{x\to\infty} F(x)$ usw. gesetzt.) $\diamond$

Lemma 8.2.14. *Es seien $\mathbb{P}$ und $(\mathbb{P}_n)_{n\in\mathbb{N}}$ eine Folge von Wahrscheinlichkeitsmaßen auf $\mathbb{R}$ mit zugehörigen Verteilungsfunktionen F und $(F_n)_{n\in\mathbb{N}}$. Dann konvergiert $\mathbb{P}_n$ genau dann schwach gegen $\mathbb{P}$, wenn F_n schwach gegen F konvergiert.*

Beweis.'$\Longrightarrow$': Sei F in $x \in \mathbb{R}$ stetig. Dann ist $\mathbb{P}(\partial(-\infty, x]) = \mathbb{P}(\{x\}) = 0$. Also folgt aus dem Portmanteau-Theorem 8.2.6(vi), dass $\lim_{n\to\infty} F_n(x) = \lim_{n\to\infty} \mathbb{P}_n((-\infty, x]) = \mathbb{P}((-\infty, x]) = F(x)$.

'$\Longleftarrow$': Nach dem Portmanteau-Theorem 8.2.6(ii) reicht es zu zeigen, dass für jede beschränkte Lipschitz-stetige Funktion $f\colon \mathbb{R} \to \mathbb{R}$ gilt: $\lim_{n\to\infty} \int_{\mathbb{R}} f \,\mathrm{d}\mathbb{P}_n = \int_{\mathbb{R}} f \,\mathrm{d}\mathbb{P}$. Ohne Beschränkung der Allgemeinheit wählen wir eine Lipschitz-stetige Funktion $f\colon \mathbb{R} \to [0,1]$ mit Lipschitz-Konstante 1. Sei $\varepsilon > 0$. Wir wählen $N \in \mathbb{N}$ und $N+1$ Stetigkeitspunkte $y_0 < y_1 < \cdots < y_N$ von F, so dass $F(y_0) < \varepsilon$, $F(y_N) > 1 - \varepsilon$ und $y_i - y_{i-1} < \varepsilon$ für jedes $i = 1, \ldots, N$. Wir teilen den Integrationsbereich $\mathbb{R}$ auf in $(-\infty, y_0]$, die Intervalle zwischen den y_i und $[y_N, \infty)$. Auf den Intervallen $[y_{i-1}, y_i]$ schätzen wir f ab gegen $\sup_{[y_{i-1}, y_i]} f$ und dies mit Hilfe der Lipschitz-Konstante Eins gegen $f(y_i) + \varepsilon$. Wenn wir noch beachten, dass $f \le 1$ gilt, erhalten wir

$$\int_{\mathbb{R}} f \,\mathrm{d}\mathbb{P}_n \le F_n(y_0) + \sum_{i=1}^{N} (f(y_i) + \varepsilon)(F_n(y_i) - F_n(y_{i-1})) + 1 - F_n(y_N).$$

Nach Voraussetzung gilt $\lim_{n\to\infty} F_n(y_i) = F(y_i)$ für jedes $i = 0, \ldots, N$, also folgt

$$\limsup_{n\to\infty} \int_{\mathbb{R}} f \,\mathrm{d}\mathbb{P}_n \le 3\varepsilon + \sum_{i=1}^{N} f(y_i)(F(y_i) - F(y_{i-1})) \le 4\varepsilon + \int_{\mathbb{R}} f \,\mathrm{d}\mathbb{P}.$$

Also haben wir $\limsup_{n\to\infty} \int_{\mathbb{R}} f \,\mathrm{d}\mathbb{P}_n \le \int_{\mathbb{R}} f \,\mathrm{d}\mathbb{P}$. Indem wir f durch $1 - f$ ersetzen, folgt die komplementäre Ungleichung, und dies beendet den Beweis. $\square$

Schwache Konvergenz von Verteilungsfunktionen wird nicht nur nützlich sein beim Beweis des Satzes von Prohorov, sondern gibt uns ein weiteres nützliches Kriterium in die Hand, schwache Konvergenz zu zeigen:

Beispiel 8.2.15. Ein berühmtes Beispiel einer schwachen Konvergenz ist der *Satz von Moivre-Laplace* (siehe Satz 5.2.5), der Folgendes besagt. Wenn $(X_n)_{n\in\mathbb{N}}$ eine Folge unabhängiger, zum Parameter $p \in (0,1)$ Bernoulli-verteilter Zufallsgrößen ist (also $\mathbb{P}(X_n = 1) = 1 - \mathbb{P}(X_n = 0) = p$ und insbesondere $\mathbb{E}(X_n) = p$ und $\mathbb{V}(X_n) = p(1-p)$), dann gilt für alle $t \in \mathbb{R}$:

$$\lim_{n\to\infty} \mathbb{P}\Big(\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{X_i - \mathbb{E}(X_i)}{\sqrt{\mathbb{V}(X_i)}} \leq t\Big) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^t e^{-x^2/2}\,\mathrm{d}x = \mathcal{N}((-\infty, t]).$$

Mit anderen Worten, die Verteilungsfunktion der standardisierten Summe $n^{-1/2}\sum_{i=1}^n \widetilde{X}_i$ mit $\widetilde{X}_i = (X_i - E(X_i))/\sqrt{\mathbb{V}(X_i)}$ konvergiert punktweise gegen die Verteilungsfunktion der Standard-Normalverteilung $\mathcal{N}$. Dies ist aber schon die schwache Konvergenz dieser Verteilungsfunktionen. Aus Satz 8.2.14 folgt die schwache Konvergenz $n^{-1/2}\sum_{i=1}^n \widetilde{X}_i \Longrightarrow \mathcal{N}$. Eines unserer Hauptergebnisse wird eine Version der obigen Aussage für *beliebige* Verteilungen an Stelle der Bernoulli-Verteilung sein, und dies ist der Zentrale Grenzwertsatz. $\diamond$

Nun wenden wir uns der Frage zu, unter welchen Umständen eine gegebene Folge von Wahrscheinlichkeitsmaßen schwach konvergiert, genauer gesagt, unter welchen Umständen sie mindestens eine schwach konvergente Teilfolge besitzt.[3] An dem simplen Beispiel $(\delta_n)_{n\in\mathbb{N}}$ der Punktmaße in $n \in \mathbb{N}$ sieht man, dass mindestens eine Bedingung gestellt werden muss, die verhindert, dass die Masse nach Unendlich abwandert. Der folgende Begriff tut genau dies.

Definition 8.2.16 (Straffheit). *Eine beliebige Familie $(\mathbb{P}_i)_{i\in I}$ von Wahrscheinlichkeitsmaßen auf einem metrischen Raum $(E, \mathtt{d})$ heißt* straff, *falls zu jedem $\varepsilon > 0$ eine kompakte Menge $K \subset E$ existiert mit $\mathbb{P}_i(K^c) < \varepsilon$ für jedes $i \in I$.*

Natürlich überträgt sich Straffheit auf Teilmengen und auf endliche Vereinigungen, aber es ist *a priori* nicht klar, ob endliche Mengen straff sind (außer in polnischen Räumen; siehe Lemma 8.2.18).

Bemerkung 8.2.17. (i) Ist E kompakt, so ist die Menge aller Wahrscheinlichkeitsmaße auf E straff.

(ii) Falls eine Familie $(X_i)_{i\in I}$ von Zufallsgrößen $\mathcal{L}^1$-beschränkt ist (d. h. $\sup_{i\in I} \mathbb{E}[|X_i|] < \infty$ gilt), so ist die Menge ihrer Verteilungen straff, wie man als eine Übungsaufgabe zeigt.

(iii) Die Familie $(\delta_n)_{n\in\mathbb{N}}$ ist nicht straff.

(iv) Die Familie der gleichförmigen Verteilungen auf $[-n, n]$ mit $n \in \mathbb{N}$ ist nicht straff.

(v) Sei E ein metrischer Raum und $A \subset E$. Dann ist $\{\delta_a : a \in A\}$ genau dann straff, wenn A relativkompakt ist (d. h. $\overline{A}$ kompakt). $\diamond$

Lemma 8.2.18. *Falls E polnisch ist, so ist jede einelementige Familie von Wahrscheinlichkeitsmaßen auf E straff, also auch jede endliche.*

Beweis. Sei $\mathbb{P}$ ein Wahrscheinlichkeitsmaß auf E. Sei $\varepsilon > 0$. Da E separabel ist, gibt es zu jedem $n \in \mathbb{N}$ Elemente $x_1^{(n)}, x_2^{(n)}, \dots \in E$, so dass $E = \bigcup_{i=1}^\infty B_{1/n}(x_i^{(n)})$, wobei $B_{1/n}$ die offene $1/n$-Kugel

[3] Man sagt in diesem Falle, sie sei *schwach relativ folgenkompakt*.

ist. Wähle $N_n \in \mathbb{N}$ so, dass $\mathbb{P}(E \setminus \bigcup_{i=1}^{N_n} B_{1/n}(x_i^{(n)})) < \varepsilon 2^{-n}$. Die Menge $A = \bigcap_{n\in\mathbb{N}} \bigcup_{i=1}^{N_n} B_{1/n}(x_i^{(n)})$ ist nach Konstruktion total beschränkt[4]. Wegen Vollständigkeit von E ist $\overline{A}$ kompakt. Außerdem erfüllt sie $\mathbb{P}(\overline{A}^{\mathrm{c}}) \le \mathbb{P}(E \setminus A) < \sum_{n\in\mathbb{N}} \varepsilon 2^{-n} = \varepsilon$. □

Beispiel 8.2.19. Eine Familie $\{\mathcal{N}_{\mu,\sigma^2} : (\mu, \sigma^2) \in L\}$ von Normalverteilungen auf $\mathbb{R}$ ist genau dann straff, wenn die Indexmenge $L \subset \mathbb{R} \times (0, \infty)$ beschränkt ist. (Übungsaufgabe.) ◇

Beispiel 8.2.20. Sei $(X_i)_{i\in I}$ eine Familie von nichtnegativen Zufallsgrößen mit $\mathbb{E}[X_i] = 1$ für jedes $i \in I$. Für $i \in I$ betrachten wir die Wahrscheinlichkeitsmaße $\mathbb{P}_i$ auf $\mathbb{R}$, die durch $\mathbb{P}_i(A) = \int_A x\, \mathbb{P} \circ X_i^{-1}(\mathrm{d}x)$ für $A \in \mathcal{B}$ gegeben ist. (Man nennt $\mathbb{P}_i$ auch die *größenverzerrte Verteilung* zu X_i.) Als Übungsaufgabe zeige man, dass $(\mathbb{P}_i)_{i\in I}$ genau dann straff ist, wenn $(X_i)_{i\in I}$ gleichgradig integrierbar ist. ◇

Nun kommen wir zum Hauptergebnis zu der Frage, wann man aus einer Familie von Wahrscheinlichkeitsmaßen eine schwach konvergente Teilfolge extrahieren kann.

Satz 8.2.21 (Satz von Prohorov)**.** *Sei $(E, \mathtt{d})$ ein metrischer Raum und $M \subset \mathcal{M}_1(E)$.*

(i) Falls M straff ist, so auch schwach relativ folgenkompakt.

(ii) Sei nun E zusätzlich polnisch. Dann ist M straff, wenn M schwach relativ folgenkompakt ist.

Korollar 8.2.22. *Wenn E ein kompakter metrischer Raum ist, ist die Menge aller Wahrscheinlichkeitsmaße auf E schwach folgenkompakt.*

Satz 8.2.21(ii) gilt nicht in allgemeinen metrischen Räumen, denn zum Beispiel sind einelementige Familien immer schwach folgenkompakt, aber nur unter Zusatzvoraussetzungen auch straff, siehe etwa Lemma 8.2.18. Die weitaus nützlichere Aussage des Satzes von Prohorov (und die deutlich schwieriger zu beweisende) ist die in (i), denn man steht immer wieder vor der Ausgabe, aus einer Folge von Wahrscheinlichkeitsmaßen eine konvergente Teilfolge zu extrahieren. Siehe etwa Beispiel 8.2.24. Eine wichtige Rolle wird die Aussage (i) auch in unserer Konstruktion der Brown'schen Bewegung in Kapitel 11 spielen. Den Beweis von Satz 8.2.21(i) werden wir nur für $E = \mathbb{R}$ durchführen. Für den allgemeinen Beweis siehe etwa [Bi68].

Beweis von Satz 8.2.21(ii). Sei $M \subset \mathcal{M}_1(E)$ schwach relativ folgenkompakt. Wir wählen eine abzählbare dichte Teilmenge $\{x_1, x_2, \dots\}$ von E. Für $n \in \mathbb{N}$ sei $A_{n,N} = \bigcup_{i=1}^{N} B_{1/n}(x_i)$, dann gilt $A_{n,N} \uparrow E$ für $N \to \infty$. Setze

$$\delta = \sup_{n\in\mathbb{N}} \inf_{N\in\mathbb{N}} \sup_{\mathbb{P}\in M} \mathbb{P}(A_{n,N}^{\mathrm{c}}).$$

Dann gibt es ein $n \in \mathbb{N}$, so dass für jedes $N \in \mathbb{N}$ ein $\mathbb{P}_N \in M$ existiert mit $\mathbb{P}_N(A_{n,N}^{\mathrm{c}}) \ge \delta/2$. Nach Voraussetzung besitzt $(\mathbb{P}_N)_{N\in\mathbb{N}}$ eine schwach konvergente Teilfolge $(\mathbb{P}_{N_k})_{k\in\mathbb{N}}$ mit einem Grenzwert $\mathbb{P}$. Nach dem Portmanteau-Theorem gilt für jedes $N \in \mathbb{N}$

$$\mathbb{P}(A_{n,N}^{\mathrm{c}}) \ge \limsup_{k\to\infty} \mathbb{P}_{N_k}(A_{n,N}^{\mathrm{c}}) \ge \liminf_{k\to\infty} \mathbb{P}_{N_k}(A_{n,N_k}^{\mathrm{c}}) \ge \delta/2.$$

[4] Eine Menge A heißt *total beschränkt* oder *präkompakt*, falls sie zu jedem $\varepsilon > 0$ mit endlich vielen Kugeln vom Radius ε überdeckt werden kann. In vollständigen metrischen Räumen fallen die Begriffe 'präkompakt' und 'relativkompakt' zusammen.

Andererseits gilt $A^{\mathrm{c}}_{n,N} \downarrow \emptyset$ für $N \to \infty$, also $\mathbb{P}(A^{\mathrm{c}}_{n,N}) \to 0$. Dies zeigt, dass $\delta = 0$.

Sei nun $\varepsilon > 0$ beliebig. Nach dem Obigen können wir zu jedem $n \in \mathbb{N}$ ein $N'_n \in \mathbb{N}$ wählen, so dass $\mathbb{P}(A^{\mathrm{c}}_{n,N'_n}) < \varepsilon 2^{-n}$ für jedes $\mathbb{P} \in M$ gilt. Die Menge $A = \bigcap_{n\in\mathbb{N}} A_{n,N'_n}$ ist nach Konstruktion total beschränkt, also relativ kompakt. Ferner gilt für jedes $\mathbb{P} \in M$

$$\mathbb{P}((\overline{A})^{\mathrm{c}}) \leq \mathbb{P}(A^{\mathrm{c}}) \leq \sum_{n\in\mathbb{N}} \mathbb{P}(A^{\mathrm{c}}_{n,N'_n}) \leq \varepsilon.$$

Also ist M straff. □

Nun wollen wir Satz 8.2.21(i) beweisen, zumindest im Fall $E = \mathbb{R}$. Im Laufe des Beweises muss man sich einen Kandidaten für einen schwachen Grenzwert entlang einer geeigneten Teilfolge verschaffen. Im Fall $E = \mathbb{R}$ ist dies nach Lemma 8.2.14 das selbe Problem, wie einen schwachen Limespunkt einer Folge von Verteilungsfunktionen zu konstruieren. Diesen Punkt des Beweises wird uns der Auswahlsatz von *Helly* leisten.

Satz 8.2.23 (Helly'scher Auswahlsatz)**.** *Sei $(F_n)_{n\in\mathbb{N}}$ eine Folge von Verteilungsfunktionen $\mathbb{R} \to [0,1]$. Dann gibt es eine Teilfolge $(n_k)_{k\in\mathbb{N}}$ und eine rechtsstetige steigende Funktion $F\colon \mathbb{R} \to [0,1]$, sodass F_{n_k} schwach für $k \to \infty$ gegen F konvergiert.*

Beweis. Der Beweis beruht auf einem gewöhnlichen Diagonalfolgenargument. Sei $\mathbb{Q} = \{q_n \colon n \in \mathbb{N}\}$ eine Abzählung von $\mathbb{Q}$. Nach dem Satz von Bolzano-Weierstraß besitzt die beschränkte Folge $(F_n(q_1))_{n\in\mathbb{N}}$ eine konvergente Teilfolge $(F_{n_k^{(1)}}(q_1))_{k\in\mathbb{N}}$. Aus dem gleichen Grund finden wir eine Teilfolge $(n_k^{(2)})_{k\in\mathbb{N}}$ von $(n_k^{(1)})_{k\in\mathbb{N}}$, so dass auch $(F_{n_k^{(2)}}(q_2))_{k\in\mathbb{N}}$ konvergiert. Iterativ erhalten wir für jedes $i \in \mathbb{N}$ eine Teilfolge $(n_k^{(i+1)})_{k\in\mathbb{N}}$ von $(n_k^{(i)})_{k\in\mathbb{N}}$, sodass $(F_{n_k^{(i+1)}}(q_{i+1}))_{k\in\mathbb{N}}$ konvergiert. Nun betrachten wir die Teilfolge $n_k = n_k^{(k)}$ und wissen, dass $(F_{n_k}(q))_{k\in\mathbb{N}}$ für jedes $q \in \mathbb{Q}$ gegen ein $F(q) \in [0,1]$ konvergiert. Wir setzen F von $\mathbb{Q}$ auf $\mathbb{R}$ fort, indem wir setzen

$$F(x) = \inf\{F(q) \colon q \in \mathbb{Q} \text{ mit } q > x\}, \qquad x \in \mathbb{R} \setminus \mathbb{Q}.$$

Natürlich ist $F\colon \mathbb{R} \to [0,1]$ rechtsstetig und steigend. Nun müssen wir noch zeigen, dass $(F_{n_k})_{k\in\mathbb{N}}$ schwach gegen F konvergiert.

Sei F stetig in $x \in \mathbb{R}$, und sei $\varepsilon > 0$. Dann gibt es $q^-, q^+ \in \mathbb{Q}$ mit $q^- < x < q^+$ und $F(q^-) \geq F(x) - \varepsilon$ und $F(q^+) \leq F(x) + \varepsilon$. Nach Konstruktion ist dann

$$\limsup_{k\to\infty} F_{n_k}(x) \leq \lim_{k\to\infty} F_{n_k}(q^+) = F(q^+) \leq F(x) + \varepsilon.$$

Also ist $\limsup_{k\to\infty} F_{n_k}(x) \leq F(x)$. Analog erhält man auch, dass $\liminf_{k\to\infty} F_{n_k}(x) \geq F(x)$. Also konvergiert $(F_{n_k})_{k\in\mathbb{N}}$ schwach gegen F. □

Man beachte, dass die Limesfunktion F in Hellys Satz nicht die Verteilungsfunktion eines Wahrscheinlichkeitsmaßes sein muss. Wir erinnern an das Verschiebungsbeispiel $F_n(x) = G(x-n)$ von Bemerkung 8.2.13.

Beweis von Satz 8.2.21(i) im Fall $E = \mathbb{R}$. Sei $M \subset \mathcal{M}_1(\mathbb{R})$ eine straffe Familie von Wahrscheinlichkeitsmaßen, und sei $(\mathbb{P}_n)_{n\in\mathbb{N}}$ eine Folge in M mit zugehörigen Verteilungsfunktionen $(F_n)_{n\in\mathbb{N}}$. Nach dem Satz 8.2.23 von Helly konvergiert eine Teilfolge $(F_{n_k})_{k\in\mathbb{N}}$ von $(F_n)_{n\in\mathbb{N}}$ schwach gegen eine rechtsstetige steigende Funktion $F\colon \mathbb{R} \to [0,1]$.

Wir müssen noch zeigen, dass F die Verteilungsfunktion eines Wahrscheinlichkeitsmaßes ist, denn dann folgt die Aussage aus Lemma 8.2.14. Wegen der Straffheit von M existiert zu jedem $\varepsilon > 0$ ein $R \in (0, \infty)$ mit $F_n(\infty) - F_n(x) < \varepsilon$ und $F_n(-x) - F_n(-\infty) < \varepsilon$ für alle $n \in \mathbb{N}$ und alle $x > R$. Dies nutzen wir für ein $x > R$ aus, sodass sowohl x als auch $-x$ eine Stetigkeitsstelle für F sind (dies ist möglich, da F nur höchstens abzählbar viele Unstetigkeitsstellen, also Sprungstellen, haben kann). Dann folgt

$$\begin{aligned} 1 &= \limsup_{k\to\infty} \big(F_{n_k}(\infty) - F_{n_k}(-\infty)\big) \leq 2\varepsilon + \limsup_{k\to\infty} \big(F_{n_k}(x) - F_{n_k}(-x)\big) \\ &= 2\varepsilon + F(x) - F(-x) \leq 2\varepsilon + F(\infty) - F(-\infty). \end{aligned}$$

Also ist $F(\infty) - F(-\infty) = 1$ und der Beweis beendet. □

Eine typische Anwendung des Satzes von Prohorov ist die folgende.

Beispiel 8.2.24 (Variationsproblem)**.** Die Aussage (i) im Satz von Prohorov ist nützlich, wenn man zeigen möchte, dass das Variationsproblem

$$V = \inf_{\mathbb{P}\in\mathcal{M}_1(\mathbb{R}):\ \int x\,\mathbb{P}(\mathrm{d}x)=1} \int x^2\,\mathbb{P}(\mathrm{d}x)$$

einen Minimierer besitzt, wobei das Infimum über alle Wahrscheinlichkeitsmaße $\mathbb{P}$ mit den Eigenschaften $\int |x|\,\mathbb{P}(\mathrm{d}x) < \infty$ und $\int x\,\mathbb{P}(\mathrm{d}x) = 1$ geht.

Dies zeigen wir wie folgt. Sei $(\mathbb{P}_n)_{n\in\mathbb{N}}$ eine Folge von approximativen Minimierern, also eine Folge in $\mathcal{M}_1(\mathbb{R})$ mit $\int x\,\mathbb{P}_n(\mathrm{d}x) = 1$ für jedes n und $\lim_{n\to\infty} \int x^2\,\mathbb{P}_n(\mathrm{d}x) = V$. Insbesondere gibt es ein $C > 0$ mit $\int x^2\,\mathbb{P}_n(\mathrm{d}x) \leq C$ für jedes $n \in \mathbb{N}$. Die Folge $(\mathbb{P}_n)_{n\in\mathbb{N}}$ ist straff, denn für jedes $R > 0$ und jedes $n \in \mathbb{N}$ gilt nach der Markov-Ungleichung

$$\mathbb{P}_n([-R, R]^{\mathrm{c}}) = \int_{\mathbb{R}} \mathbb{1}_{\{|x|>R\}}\,\mathbb{P}_n(\mathrm{d}x) \leq \frac{1}{R^2}\int x^2\,\mathbb{P}_n(\mathrm{d}x) \leq \frac{C}{R^2}.$$

Nach dem Satz von Prohorov gibt es eine Teilfolge $(n_k)_{k\in\mathbb{N}}$, so dass $\mathbb{P}_{n_k} \Longrightarrow \mathbb{P}$ für ein $\mathbb{P} \in \mathcal{M}_1(\mathbb{R})$. Nun zeigen wir, dass dieses $\mathbb{P}$ ein Minimierer des Variationsproblems ist. Dazu muss gezeigt werden, dass (i) $\mathbb{P}$ optimal ist, also $\int x^2\,\mathbb{P}(\mathrm{d}x) = V$ erfüllt, und (ii) $\mathbb{P}$ zulässig ist, also $\int x\,\mathbb{P}(\mathrm{d}x) = 1$ erfüllt.

Der erste Punkt ergibt sich daraus, dass für jedes $R > 0$ wegen der schwachen Konvergenz gilt:

$$\int x^2 \wedge R\,\mathbb{P}(\mathrm{d}x) = \lim_{n\to\infty} \int x^2 \wedge R\,\mathbb{P}_n(\mathrm{d}x) \leq \lim_{n\to\infty} \int x^2\,\mathbb{P}_n(\mathrm{d}x) \leq V,$$

wobei wir benutzt haben, dass die Abbildung $x \mapsto x^2 \wedge R$ beschränkt und stetig ist. Mit $R \to \infty$ ergibt sich (i).

Den zweiten Punkt zeigt man in zwei Schritten, indem man zunächst bemerkt, dass für jedes $R > 0$ gilt

$$\liminf_{n\to\infty} \int |x|\,\mathbb{P}_n(\mathrm{d}x) \geq \liminf_{n\to\infty} \int |x| \wedge R\,\mathbb{P}_n(\mathrm{d}x) = \int |x| \wedge R\,\mathbb{P}(\mathrm{d}x),$$

also $\liminf_{n\to\infty} \int |x|\,\mathbb{P}_n(\mathrm{d}x) \geq \int |x|\,\mathbb{P}(\mathrm{d}x)$. Um zu sehen, dass auch die komplementäre Aussage, $\limsup_{n\to\infty} \int |x|\,\mathbb{P}_n(\mathrm{d}x) \leq \int |x|\,\mathbb{P}(\mathrm{d}x)$, gilt, muss man nur zeigen, dass man beim '$\geq$' in der obigen Rechnung asymptotisch nichts verloren hat. Dies ergibt sich aus der Abschätzung

$$\sup_{n\in\mathbb{N}} \int \big||x| - |x| \wedge R\big|\,\mathbb{P}_n(\mathrm{d}x) \leq 2\sup_{n\in\mathbb{N}} \int |x|\mathbb{1}_{|x|>R}\,\mathbb{P}_n(\mathrm{d}x) \leq \sup_{n\in\mathbb{N}} \frac{2}{R}\int x^2\,\mathbb{P}_n(\mathrm{d}x) \leq \frac{2C}{R}.$$

Also haben wir gezeigt, dass $\lim_{n\to\infty} \int |x|\, \mathbb{P}_n(\mathrm{d}x) = \int |x|\, \mathbb{P}(\mathrm{d}x)$ gilt. Auf ähnliche Weise (etwa durch Zerlegung des Integrationsbereichs in $(-\infty, 0]$ und $[0, \infty)$) zeigt man, dass wir auch haben: $1 = \lim_{n\to\infty} \int x\, \mathbb{P}_n(\mathrm{d}x) = \int x\, \mathbb{P}(\mathrm{d}x)$, d. h. dass (ii) gilt. ◇

Mit Hilfe des Satzes von Prohorov kann man die schwache Konvergenz auf eine sehr nützliche Weise charakterisieren. Hinzu zur Straffheit muss man nur noch prüfen, ob die Integrale gegen eine genügend große Klasse von stetigen beschränkten Funktionen konvergieren. Was 'genügend groß' heißen soll, definieren wir zunächst:

Definition 8.2.25 (trennende Familie)**.** *Wir nennen eine Teilmenge $\mathcal{C}$ der Menge der beschränkten stetigen Funktionen $E \to \mathbb{R}$* trennend*, falls die folgende Implikation für je zwei Wahrscheinlichkeitsmaße $\mathbb{P}_1$ und $\mathbb{P}_2$ gilt:*

$$\int_E f\, \mathrm{d}\mathbb{P}_1 = \int_E f\, \mathrm{d}\mathbb{P}_2 \text{ für alle } f \in \mathcal{C} \qquad \Longrightarrow \qquad \mathbb{P}_1 = \mathbb{P}_2.$$

Ein Wahrscheinlichkeitsmaß ist also durch die Integrale gegen alle Mitglieder einer trennenden Famile schon eindeutig festgelegt. Zum Beispiel ist die Menge aller Lipschitz-stetigen Funktionen $f\colon E \to [0,1]$ trennend, wie man sich als eine Übungsaufgabe überlegt.

Satz 8.2.26. *Sei $(E, \mathtt{d})$ ein polnischer Raum und $(\mathbb{P}_n)_{n\in\mathbb{N}}$ eine Folge von Wahrscheinlichkeitsmaßen auf E. Dann konvergiert $(\mathbb{P}_n)_{n\in\mathbb{N}}$ genau dann schwach gegen ein Wahrscheinlichkeitsmaß $\mathbb{P}$ auf E, wenn $(\mathbb{P}_n)_{n\in\mathbb{N}}$ straff ist und es eine trennende Familie $\mathcal{C}$ gibt, so dass für jedes $f \in \mathcal{C}$ gilt: $\lim_{n\to\infty} \int_E f\, \mathrm{d}\mathbb{P}_n = \int_E f\, \mathrm{d}\mathbb{P}$.*

Beweis. Die Richtung '$\Longrightarrow$' ist klar, da Straffheit aus Konvergenz folgt und da $\mathcal{C}$ eine Menge von beschränkten stetigen Funktionen ist. Die Richtung '$\Longleftarrow$' folgt daraus, dass ja auf Grund des Satzes von Prohorov jede Teilfolge von $(\mathbb{P}_n)_{n\in\mathbb{N}}$ ihrerseits eine Teilfolge hat, die schwach gegen ein Wahrscheinlichkeitsmaß konvergiert, das durch die Integrale gegen alle Mitglieder von $\mathcal{C}$ schon eindeutig festgelegt wird, unabhängig von der Teilfolge. Also stimmen alle Häufungspunkte von $(\mathbb{P}_n)_{n\in\mathbb{N}}$ miteinander überein, d. h., die Folge konvergiert. □

8.3 Charakteristische Funktionen und der Zentrale Grenzwertsatz

In diesem Abschnitt geben wir unter Anderem einen Beweis des berühmten Zentralen Grenzwertsatzes, den wir schon in Satz 5.2.2 formulierten. Dieser besagt ja, dass die Verteilung von $n^{-\frac{1}{2}} \sum_{i=1}^n X_i$ gegen die Standardnormalverteilung konvergiert, wenn $(X_n)_{n\in\mathbb{N}}$ eine Folge unabhängiger zentrierter Zufallsgrößen mit Varianz Eins ist. Einen Spezialfall, den berühmten Satz 5.2.5 von Moivre-Laplace, bewiesen wir schon, und der Beweis beruhte auf expliziter Rechnung. Hier geben wir den Beweis im allgemeinen Fall.

Unser Hauptbeweismittel werden die charakteristischen Funktionen sein, denen wir uns zunächst eine Weile lang widmen werden. Die charakteristische Funktion einer Wahrscheinlichkeitsverteilung spiegelt viele wichtige Eigenschaften der Verteilung auf analytische Weise wider, so dass sich etliche neue Möglichkeiten der Untersuchung der Verteilung eröffnen.

Wir werden auch *komplex*wertige Funktionen integrieren, was völlig natürlich durch Zerlegung in Real- und Imaginärteil geschieht, also $\mathbb{E}(X + \mathrm{i}Y) = \mathbb{E}(X) + \mathrm{i}\mathbb{E}(Y)$ für integrierbare (reellwertige) Zufallsgrößen X und Y. Mit $\langle\cdot,\cdot\rangle$ bezeichnen wir das Standardskalarprodukt auf dem $\mathbb{R}^d$, und i ist die imaginäre Einheit.

Definition 8.3.1 (charakteristische Funktion, Fourier-Transformierte)**.** *Die* charakteristische Funktion *oder* Fourier-Transformierte *eines Wahrscheinlichkeitsmaßes $\mathbb{P}$ auf $(\mathbb{R}^d, \mathcal{B}_d)$ ist die Abbildung $\widehat{\mathbb{P}}\colon \mathbb{R}^d \to \mathbb{C}$, definiert durch*

$$\widehat{\mathbb{P}}(x) = \int \mathrm{e}^{\mathrm{i}\langle x,y\rangle}\,\mathbb{P}(\mathrm{d}y) = \int \cos(\langle x,y\rangle)\,\mathbb{P}(\mathrm{d}y) + \mathrm{i}\int \sin(\langle x,y\rangle)\,\mathbb{P}(\mathrm{d}y), \qquad x \in \mathbb{R}^d.$$

Die charakteristische Funktion einer $\mathbb{R}^d$-wertigen Zufallsvariablen X ist die charakteristische Funktion der Verteilung von X; statt $\widehat{\mathbb{P}\circ X^{-1}}$ schreiben wir φ_X, also $\varphi_X(x) = \mathbb{E}(\mathrm{e}^{\mathrm{i}\langle x,X\rangle})$.

Beispiel 8.3.2. (a) Die charakteristische Funktion der d-dimensionalen Standardnormalverteilung $\mathcal{N} = \mathcal{N}(0, I)$ (siehe Beispiel 6.6.7) ist $\widehat{\mathcal{N}}(x) = \mathrm{e}^{-\frac{1}{2}\|x\|^2}$, wie man mit Hilfe einer quadratischen Ergänzung im Exponenten und dem Cauchy'schen Integralsatz zeigt. Wir skizzieren den Beweis für den Spezialfall $d = 1$. Zunächst haben wir

$$\widehat{\mathcal{N}}(x) = \frac{\mathrm{e}^{-\frac{1}{2}x^2}}{\sqrt{2\pi}} \int_{\mathbb{R}} \mathrm{e}^{-\frac{1}{2}(y-\mathrm{i}x)^2}\,\mathrm{d}y = \frac{\mathrm{e}^{-\frac{1}{2}x^2}}{\sqrt{2\pi}} \lim_{R\to\infty} \int_{-R-\mathrm{i}x}^{R-\mathrm{i}x} \mathrm{e}^{-\frac{1}{2}z^2}\,\mathrm{d}z,$$

wobei der Integrationsweg die gerichtete Strecke von $-R-\mathrm{i}x$ bis $R-\mathrm{i}x$ ist. Nun ergänzen wir diesen Weg durch die drei gerichteten Strecken von $R-\mathrm{i}x$ nach R, von R nach $-R$ und von $-R$ nach $-R-\mathrm{i}x$ und erhalten eine geschlossene Kurve ohne Doppelpunkte. Nach dem Cauchy'schen Integralsatz ist das Kurvenintegral über diese geschlossene Kurve gleich Null, denn $z \mapsto \mathrm{e}^{-\frac{1}{2}z^2}$ ist analytisch in der ganzen komplexen Ebene. Außerdem ist leicht zu sehen, dass die Integralteilstücke über die beiden Strecken von $R-\mathrm{i}x$ nach R und von $-R$ nach $-R-\mathrm{i}x$ im Grenzwert $R \to \infty$ verschwinden. Also folgt

$$\widehat{\mathcal{N}}(x) = \frac{\mathrm{e}^{-\frac{1}{2}x^2}}{\sqrt{2\pi}} \lim_{R\to\infty} \int_{-R}^{R} \mathrm{e}^{-\frac{1}{2}z^2}\,\mathrm{d}z = \mathrm{e}^{-\frac{1}{2}x^2}.$$

Für eine allgemeine positiv definite symmetrische $d \times d$-Matrix Σ und $b \in \mathbb{R}^d$ ist die charakteristische Funktion von $\mathcal{N}(b, \Sigma)$ gegeben durch

$$\widehat{\mathcal{N}(b,\Sigma)}(x) = \exp\Big\{\mathrm{i}\langle x,b\rangle - \frac{1}{2}\langle x, \Sigma x\rangle\Big\}, \qquad x \in \mathbb{R}^d.$$

(b) Die charakteristische Funktion der Cauchy-Verteilung $\mathbb{P}$ auf $\mathbb{R}$ zum Parameter $c > 0$ (siehe Beispiel 6.6.6) ist gegeben durch $\widehat{\mathbb{P}}(x) = \mathrm{e}^{-c|x|}$, wie man mit Hilfe des Residuensatzes zeigt. Den Beweis skizzieren wir für $c = 1$ und $x > 0$. Zunächst sehen wir, dass gilt:

$$\widehat{\mathbb{P}}(x) = \frac{1}{\pi} \lim_{R\to\infty} \int_{-R}^{R} \frac{\mathrm{e}^{\mathrm{i}xy}}{1+y^2}\,\mathrm{d}y.$$

Nun ergänzen wir die gerichtete Strecke von $-R$ nach R durch den Halbkreis (gegen den Uhrzeigersinn durchlaufen) um Null mit Radius R von R über $\mathrm{i}R$ nach $-R$. Die entstehende geschlossene Kurve nennen wir γ_R. Man muss sich überlegen, dass das Integral über den Halbkreis im Grenzwert für $R \to \infty$ verschwindet (hier benötigen wir, dass x positiv ist). Also erhalten wir

$$\widehat{\mathbb{P}}(x) = 2\mathrm{i} \lim_{R\to\infty} \frac{1}{2\pi\mathrm{i}} \int_{\gamma_R} \frac{\mathrm{e}^{\mathrm{i}xz}/(z+\mathrm{i})}{z-\mathrm{i}}\,\mathrm{d}z.$$

Die Kurve γ_R besitzt die Windungszahl Eins um die Singularität des Nenners, i. Nach dem Residuensatz ist für jedes $R > 1$ der Term hinter '$\lim_{R\to\infty}$' gleich dem Wert des Zählers an der Stelle dieser Singularität, also gleich $\mathrm{e}^{\mathrm{i}x\mathrm{i}}(\mathrm{i}+\mathrm{i}) = \mathrm{e}^{-|x|}/(2\mathrm{i})$. Daraus folgt die Behauptung.

(c) Die gleichförmige Verteilung auf dem Intervall $[0, 1]$ hat die charakteristische Funktion $t \mapsto \frac{\mathrm{e}^{\mathrm{i}t}-1}{\mathrm{i}t}$.

(d) Die Exponentialverteilung zum Parameter $\lambda > 0$ (Dichte $x \mapsto \lambda\mathrm{e}^{-\lambda x}\mathbb{1}_{[0,\infty)}(x)$) hat die charakteristische Funktion $t \mapsto \frac{\lambda}{\lambda-\mathrm{i}t}$ (Übungsaufgabe).

(e) Die Binomial-Verteilung $\sum_{k=0}^n \binom{n}{k} p^k(1-p)^{n-k}\delta_k$ zu den Parametern $n \in \mathbb{N}$ und $p \in (0, 1)$ hat die charakteristische Funktion $t \mapsto (1 + p(\mathrm{e}^{\mathrm{i}t} - 1))^n$.

(f) Die Poisson-Verteilung $\mathrm{e}^{-\alpha} \sum_{k\in\mathbb{N}_0} \frac{\alpha^k}{k!}\delta_k$ zum Parameter $\alpha > 0$ hat die charakteristische Funktion $t \mapsto \mathrm{e}^{\alpha(\mathrm{e}^{\mathrm{i}t}-1)}$.

(g) Die Verteilung auf $[-a, a]$ mit der Dichte $x \mapsto \frac{1}{a}(1 - |x|/a)^+$ hat die charakteristische Funktion $t \mapsto \frac{2}{a^2t^2}(1 - \cos(at))$.

$\diamond$

Die große Bedeutung der charakteristischen Funktionen φ_X von Zufallsgrößen X gründet sich auf folgenden Tatsachen:

(i) Die Abbildung $X \mapsto \varphi_X$ hat schöne Eigenschaften unter Transformationen und unabhängigen Summen,

(ii) φ_X bestimmt die Verteilung von X eindeutig,

(iii) punktweise Konvergenz der charakteristischen Funktionen ist äquivalent zur schwachen Konvergenz der zugehörigen Verteilungen.

Dies formulieren wir nun im Einzelnen, zunächst (i).

Lemma 8.3.3. *Für jede $\mathbb{R}^d$-wertige Zufallsgröße X gelten die folgenden Aussagen.*

(a) Für jedes $x \in \mathbb{R}^d$ ist $|\varphi_X(x)| \leq 1$, und es gilt $\varphi_X(0) = 1$.

(b) Für alle $a \in \mathbb{R}$ und $b \in \mathbb{R}^d$ ist $\varphi_{aX+b}(x) = \varphi_X(ax)\mathrm{e}^{\mathrm{i}\langle x,b\rangle}$.

(c) φ_X ist genau dann reellwertig, wenn $\mathbb{P} \circ X^{-1} = \mathbb{P} \circ (-X)^{-1}$.

(d) Wenn X und Y zwei unabhängige Zufallsgrößen sind, so gilt $\varphi_{X+Y} = \varphi_X \cdot \varphi_Y$.

Beweis. einfache Übungsaufgaben. Beim Beweis von (d) benutze man Lemma 7.2.13(b) (genauer gesagt, eine Erweiterung auf komplexwertige Zufallsgrößen). □

Wir beleuchten noch den Zusammenhang mit höheren Momenten:

Lemma 8.3.4. *Sei X eine reelle Zufallsvariable mit charakteristischer Funktion φ.*

(a) Falls $\mathbb{E}[|X|^n] < \infty$, so ist φ n Mal stetig differenzierbar, und die Ableitungen sind gegeben durch $\varphi^{(k)}(t) = \mathbb{E}[(\mathrm{i}X)^k \mathrm{e}^{\mathrm{i}tX}]$ für $k = 0, \dots, n$.

(b) Falls $\mathbb{E}[X^2] < \infty$, so ist

$$\varphi(t) = 1 + \mathrm{i}t\mathbb{E}[X] - \frac{1}{2}t^2\mathbb{E}[X^2] + o(t^2), \qquad t \to 0.$$

(c) Sei $h \in \mathbb{R}$ mit $\lim_{n\to\infty} \frac{1}{n!}|h|^n\mathbb{E}[|X|^n] = 0$, so gilt

$$\varphi(t+h) = \sum_{k=0}^{\infty} \frac{(\mathrm{i}h)^k}{k!}\mathbb{E}\big[\mathrm{e}^{\mathrm{i}tX}\, X^k\big], \qquad t \in \mathbb{R}.$$

Insbesondere gilt dies unter der Voraussetzung, dass $\mathbb{E}[\mathrm{e}^{|hX|}] < \infty$.

Beweis. Übungsaufgaben. Im Beweis von (a) betrachte man z. B. die Zufallsgröße $Y_k(t,h,x) = k!h^{-k}\mathrm{e}^{\mathrm{i}tx}\big(\mathrm{e}^{\mathrm{i}hx} - \sum_{l=0}^{k-1}\frac{(\mathrm{i}hx)^l}{l!}\big)$, benutze eine Taylor-Entwicklung der Exponentialfunktion und geeignete Abschätzungen und gehe ähnlich vor wie im Beweis des Differenziationslemmas 6.8.9. (b) folgt aus der Taylorformel unter Verwendung von (a). Der Beweis von (c) ist elementar. □

Wir werden in Kürze zeigen, dass die Verteilung einer Zufallsvariable eindeutig durch seine charakteristische Funktion gegeben ist, also durch die Angabe der Integrale gegen die Funktionen $x \mapsto \mathrm{e}^{\mathrm{i}xy}$ mit $y \in \mathbb{R}$. Zuvor aber betrachten wir die Frage, ob sie vielleicht schon durch die Angabe aller Momente festgelegt ist. Diese Frage ist von einiger Bedeutung bei Zufallsgrößen, deren Verteilung nur schwer angegeben werden kann, aber deren Momente zugänglich sind.

Korollar 8.3.5 (Momentenproblem)**.** *Sei X eine reelle Zufallsvariable mit*

$$\alpha = \limsup_{n\to\infty} \frac{1}{n}\mathbb{E}[|X|^n]^{1/n} < \infty.$$

Dann ist die charakteristische Funktion φ von X analytisch, und die Verteilung von X ist eindeutig durch die Momente $\mathbb{E}[X^n]$, $n \in \mathbb{N}$, festgelegt. Speziell gilt dies, wenn $\mathbb{E}[\mathrm{e}^{t|X|}] < \infty$ für ein $t > 0$.

Beweis. Die Stirling'sche Formel besagt, dass $n! \sim n^n\mathrm{e}^{-n}\sqrt{2\pi n}$ für $n \to \infty$. Für alle $|h| < 1/(3\alpha)$ gilt daher

$$\limsup_{n\to\infty} \frac{1}{n!}|h|^n\mathbb{E}[|X|^n] = \limsup_{n\to\infty} \frac{1}{\sqrt{2\pi n}}\Big(\mathbb{E}[|X|^n]^{1/n}\,|h|\,\frac{\mathrm{e}}{n}\Big)^n \le \limsup_{n\to\infty}\Big(\frac{\mathrm{e}}{3}\Big)^n = 0.$$

Nach Lemma 8.3.4(c) ist φ um jeden Punkt $t \in \mathbb{R}$ in eine Potenzreihe entwickelbar mit Konvergenzradius $\geq 1/(3\alpha)$, insbesondere also analytisch. Damit ist sie insbesondere festgelegt durch die Koeffizienten der Reihe um $t = 0$, nach Lemma 8.3.4(a) also durch die Momente von X. Die Zusatzaussage zeigt man als eine elementare Übungsaufgabe. □

Die Momente legen also die Verteilung der Zufallsgröße X fest, wenn sie nicht schneller wachsen als $\mathbb{E}[|X|^n] \leq (Cn)^n$, $n \in \mathbb{N}$, für ein $C > 0$.

Beispiel 8.3.6. (i) Eine normalverteilte Zufallsgröße $X \sim \mathcal{N}(\mu, \sigma^2)$ hat die Momenten erzeugende Funktion $\mathbb{E}[\mathrm{e}^{tX}] = \mathrm{e}^{\mu t + t^2\sigma^2/2} < \infty$. Also ist X die einzige Zufallsgröße, die die selben Momente wie $\mathcal{N}(\mu, \sigma^2)$ besitzt. Die charakteristische Funktion $\varphi(t) = \mathrm{e}^{\mathrm{i}\mu t - t^2\sigma^2/2}$ ist in der Tat analytisch.

(ii) Eine zum Parameter $\lambda > 0$ exponential-verteilte Zufallsgröße X hat die Momenten erzeugende Funktion $\mathbb{E}[\mathrm{e}^{tX}] = \frac{\lambda}{\lambda - t}$, und dies ist endlich für $t < \lambda$. Also ist X durch seine Momente eindeutig bestimmt. Die charakteristische Funktion $\varphi(t) = \lambda/(\lambda - \mathrm{i}t)$ ist analytisch. Die Singularität bei $t = -\mathrm{i}\lambda$ korrespondiert mit dem Konvergenzradius λ und der Tatsache, dass die t-ten exponentiellen Momente für $t \geq \lambda$ nicht endlich sind.

(iii) Wenn Y eine Standard-Normalverteilte Zufallsgröße ist, so besitzt $X = \mathrm{e}^Y$ per Definition die sogenannte *Log-Normalverteilung*. Da nY die Verteilung einer Summe von n^2 unabhängigen Kopien von Y hat, kann man die n-ten Momente von X leicht zu $\mathbb{E}[X^n] = \mathrm{e}^{n^2/2}$ errechnen. Hier ist also $\alpha = \infty$, und Korollar 8.3.5 kann nicht angewendet werden. Tatsächlich kann man auch eine andere Verteilung konstruieren, die die selben Momente wie X besitzt.

$\diamond$

Die eindeutige Beziehung zwischen einer Zufallsgröße und ihren Momenten kann man sich auch für die schwache Konvergenz nutzbar machen:

Satz 8.3.7 (Fréchet-Shohat). *Sei X eine Zufallsgröße, so dass $(\frac{1}{k}\mathbb{E}[|X|^k]^{1/k})_{k\in\mathbb{N}}$ beschränkt ist, und sei $(X_n)_{n\in\mathbb{N}}$ eine Folge von Zufallsgrößen, so dass für jedes $k \in \mathbb{N}$ gilt: $\lim_{n\to\infty} \mathbb{E}[X_n^k] = \mathbb{E}[X^k]$. Dann konvergiert X_n schwach gegen X.*

Beweis. Übungsaufgabe. Man zeige die Straffheit von $(X_n)_{n\in\mathbb{N}}$ und identifiziere die Momente aller möglichen Häufungspunkte. Auch das Continuous mapping theorem (Satz 8.2.11) und Lemma 8.1.18 sind hilfreich. $\square$

Die schwache Konvergenz der Zufallsgrößen X_n kann auch aus der Konvergenz der Momenten erzeugenden Funktionen $M_n(t) = \mathbb{E}[\mathrm{e}^{tX_n}]$ gefolgert werden:

Korollar 8.3.8. *Seien $X, X_1, X_2, \dots$ Zufallsgrößen mit Momenten erzeugenden Funktionen $M, M_1, M_2, \dots$. Es existiere ein $t_0 > 0$ mit $M(t) < \infty$ und $M_n(t) < \infty$ für jedes $n \in \mathbb{N}$ und $t \in [-t_0, t_0]$. Es gelte $\lim_{n\to\infty} M_n(t) = M(t)$ für jedes $t \in [-t_0, t_0]$. Dann konvergiert X_n schwach gegen X.*

Beweis. Übungsaufgabe. Man zeige Konvergenz der Momente und benutze den Satz 8.3.7 von Fréchet-Shohat. $\square$

Beispiel 8.3.9. Für $n \in \mathbb{N}$ sei N_n eine zum Parameter n Poisson-verteilte Zufallsgröße, und wir setzen $X_n = (N_n - n)/\sqrt{n}$. Als eine Übungsaufgabe zeigt man mit Hilfe von Korollar 8.3.8, dass X_n gegen die Standardnormalverteilung konvergiert. Daraus folgert man einerseits, dass $\lim_{n\to\infty} \mathrm{e}^{-n} \sum_{k=0}^n \frac{n^k}{k!} = \frac{1}{2}$ gilt und andererseits (durch Betrachtung des Negativteils von X_n), dass die Stirling-Asymptotik $n! \sim (\frac{n}{\mathrm{e}})^n \sqrt{2\pi n}$ gilt. $\diamond$

Nun kommen wir zu (ii), also dem eindeutigen Bezug zwischen einer Verteilung und ihrer charakteristischen Funktion. Die folgende Aussage kan man auch äquivalent damit formulieren, dass die Vereinigung der Familien der Funktionen $\cos(\langle x, \cdot\rangle)$ und $\sin(\langle x, \cdot\rangle)$ mit $x \in \mathbb{R}^d$ trennend ist.

Satz 8.3.10. *Es seien $\mathbb{P}_1$ und $\mathbb{P}_2$ zwei Wahrscheinlichkeitsmaße auf $(\mathbb{R}^d, \mathcal{B}_d)$ mit $\widehat{\mathbb{P}}_1 = \widehat{\mathbb{P}}_2$. Dann gilt $\mathbb{P}_1 = \mathbb{P}_2$.*

Beweis. Nach dem Eindeutigkeitssatz 6.2.8 reicht es, für jede kompakte Menge $K \subset \mathbb{R}^d$ zu zeigen, dass $\mathbb{P}_1(K) = \mathbb{P}_2(K)$; siehe auch Beispiel 6.1.5. Sei also $K \subset \mathbb{R}^d$ kompakt. Mit $\operatorname{dist}(K, x) = \inf\{\mathtt{d}(y, x) \colon y \in K\}$ bezeichnen wir den Abstand von $x \in \mathbb{R}^d$ zu K. Für $m \in \mathbb{N}$ sei $f_m \colon \mathbb{R}^d \to [0, 1]$ definiert durch

$$f_m(x) = \begin{cases} 1, & \text{falls } x \in K, \\ 0, & \text{falls } \operatorname{dist}(K, x) \geq \frac{1}{m}, \\ 1 - m \operatorname{dist}(K, x) & \text{sonst.} \end{cases}$$

(Dies ist im Wesentlichen die Funktion f_ε aus dem Beweis des Lemmas 8.2.4.) Dann ist f_m stetig mit Werten in $[0, 1]$ und kompaktem Träger, und es gilt $f_m \downarrow \mathbb{1}_K$ für $m \to \infty$. Auf Grund des monotonen Konvergenzsatzes (Satz 6.8.1) wären wir also fertig, wenn wir gezeigt hätten, dass $\int f_m \, \mathrm{d}\mathbb{P}_1 = \int f_m \, \mathrm{d}\mathbb{P}_2$ für jedes $m \in \mathbb{N}$ gilt. Dies zeigen wir nun. Sei also $m \in \mathbb{N}$ fest, und wir schreiben f statt f_m.

Sei $\varepsilon > 0$, dann können wir ein $N > 0$ so groß wählen, dass der Kubus $B_N = [-N, N]^d$ die Menge $\{x \in \mathbb{R}^d \colon f(x) \neq 0\}$ enthält und so dass $\mathbb{P}_1(B_N^{\mathrm{c}}) \leq \varepsilon$ und $\mathbb{P}_2(B_N^{\mathrm{c}}) \leq \varepsilon$ gelten. Nach dem Weierstraß'schen Approximationssatz gibt es eine Funktion $g \colon \mathbb{R}^d \to \mathbb{C}$ der Form $g(x) = \sum_{j=1}^n c_j \exp\{\mathrm{i}\langle \frac{\pi}{N} t_j, x\rangle\}$ mit $n \in \mathbb{N}$, $c_1, \dots, c_n \in \mathbb{C}$ und $t_1, \dots, t_n \in \mathbb{Z}^d$, die f bis auf ε auf B_N approximiert, d. h. $\sup\{|f(x) - g(x)| \colon x \in B_N\} \leq \varepsilon$. Es folgt, dass $\sup_{x \in \mathbb{R}^d} |g(x)| \leq 1 + \varepsilon$. Wir betrachten die Abschätzung

$$\Big|\int f \, \mathrm{d}\mathbb{P}_1 - \int f \, \mathrm{d}\mathbb{P}_2\Big| \leq \Big|\int f \, \mathrm{d}\mathbb{P}_1 - \int g \, \mathrm{d}\mathbb{P}_1\Big| + \Big|\int g \, \mathrm{d}\mathbb{P}_1 - \int g \, \mathrm{d}\mathbb{P}_2\Big| + \Big|\int g \, \mathrm{d}\mathbb{P}_2 - \int f \, \mathrm{d}\mathbb{P}_2\Big|.$$

Der zweite Summand auf der rechten Seite ist nach Voraussetzung $\widehat{\mathbb{P}}_1 = \widehat{\mathbb{P}}_2$ gleich Null. Der erste Summand kann wegen $|g(x)| \leq 1 + \varepsilon$ folgendermaßen abgeschätzt werden:

$$\begin{aligned} \Big|\int f \, \mathrm{d}\mathbb{P}_1 - \int g \, \mathrm{d}\mathbb{P}_1\Big| &\leq \Big|\int_{B_N} f \, \mathrm{d}\mathbb{P}_1 - \int_{B_N} g \, \mathrm{d}\mathbb{P}_1\Big| + \int_{B_N^{\mathrm{c}}} |f| \, \mathrm{d}\mathbb{P}_1 + \int_{B_N^{\mathrm{c}}} |g| \, \mathrm{d}\mathbb{P}_1 \\ &\leq \int_{B_N} |f - g| \, \mathrm{d}\mathbb{P}_1 + (1 + \varepsilon)\mathbb{P}_1(B_N^{\mathrm{c}}) \\ &\leq \varepsilon \mathbb{P}_1(B_N) + (1 + \varepsilon)\mathbb{P}_1(B_N^{\mathrm{c}}) \leq \varepsilon(2 + \varepsilon). \end{aligned}$$

Der dritte Summand wird analog abgeschätzt. Da $\varepsilon > 0$ beliebig war, folgt die Aussage $\int f \, \mathrm{d}\mathbb{P}_1 = \int f \, \mathrm{d}\mathbb{P}_2$, und der Beweis ist beendet. □

Für die folgende Anwendung erinnern wir an die mehrdimensionale Normalverteilung von Beispiel 6.6.7 sowie an Bemerkung 7.1.3.

Korollar 8.3.11. *Sei $\mu = \mathcal{N}(a, \Sigma)$ die Normalverteilung auf $\mathbb{R}^n$ mit Erwartungswertvektor $a \in \mathbb{R}^n$ und Kovarianzmatrix $\Sigma \in \mathbb{R}^{n \times n}$, und sei $\phi \colon \mathbb{R}^n \to \mathbb{R}^m$ eine affine Abbildung, also $\phi(x) = Ax + b$ mit einer $(m \times n)$-Matrix A und $b \in \mathbb{R}^m$. Dann ist das Bildmaß $\mu \circ \phi^{-1}$ die Normalverteilung auf $\mathbb{R}^m$ mit Erwartungswertvektor $Aa + b$ und Kovarianzmatrix $A\Sigma A^{\mathrm{T}}$; kurz: $\mathcal{N}(a, \Sigma) \circ \phi^{-1} = \mathcal{N}(Aa + b, A\Sigma A^{\mathrm{T}})$.*

Beweis. Wir identifizieren die charakteristische Funktion:

$$\begin{aligned}\widehat{\mu\circ\phi^{-1}}(t) &= \int \mathrm{e}^{\mathrm{i}\langle t,x\rangle}\,\mu\circ\phi^{-1}(\mathrm{d}x) = \int \mathrm{e}^{\mathrm{i}\langle t,Ax+b\rangle}\,\mu(\mathrm{d}x)\\ &= \mathrm{e}^{\mathrm{i}\langle t,b\rangle}\int \mathrm{e}^{\mathrm{i}\langle A^{\mathrm{T}}t,x\rangle}\,\mu(\mathrm{d}x) = \mathrm{e}^{\mathrm{i}\langle t,b\rangle}\exp\Big\{\mathrm{i}\langle A^{\mathrm{T}}t,a\rangle - \frac12\langle A^{\mathrm{T}}t,\Sigma A^{\mathrm{T}}t\rangle\Big\}\\ &= \exp\Big\{\mathrm{i}\langle t, Aa+b\rangle - \frac12\langle t, A\Sigma A^{\mathrm{T}}t\rangle\Big\}.\end{aligned}$$

Nun folgt die Aussage aus Satz 8.3.10 und Beispiel 8.3.2. □

Ferner erhalten wir aus Satz 8.3.10 eine Charakterisierung der mehrdimensionalen Normalverteilung in Termen ihrer Projektionen:

Lemma 8.3.12. *Ein $\mathbb{R}^d$-wertiger Zufallsvektor X mit positiv definiter Kovarianzmatrix Σ ist genau dann normalverteilt, wenn für jedes $\lambda\in\mathbb{R}^d$ die Zufallsgröße $\langle\lambda, X\rangle$ normalverteilt ist, und zwar mit Erwartungswert $\langle\lambda,\mathbb{E}[X]\rangle$ und Varianz $\langle\lambda,\Sigma\lambda\rangle$.*

Beweis. Übungsaufgabe. □

Als Letztes formulieren wir (iii) aus, also den Bezug zwischen schwacher Konvergenz der Verteilungen und punktweiser Konvergenz der zugehörigen charakteristischen Funktionen:

Satz 8.3.13 (Stetigkeitssatz)**.** *Es seien $\mathbb{P},\mathbb{P}_1,\mathbb{P}_2,\mathbb{P}_3\dots$ Wahrscheinlichkeitsmaße mit charakteristischen Funktionen $\varphi,\varphi_1,\varphi_2,\varphi_3,\dots$. Dann gelten:*

(a) Falls $\mathbb{P}_n\Longrightarrow\mathbb{P}$ für $n\to\infty$, so konvergiert φ_n gegen φ lokal gleichmäßig.

(b) Falls φ_n punktweise gegen eine Funktion $f\colon\mathbb{R}^d\to\mathbb{C}$ konvergiert, die in 0 stetig ist, so existiert ein Wahrscheinlichkeitsmaß Q, dessen charakteristische Funktion f ist, und $\mathbb{P}_n$ konvergiert schwach gegen Q.

Der Beweis des Stetigkeitssatzes 8.3.13 benötigt Vorbereitungen. Bevor wir ihn bringen, geben wir ein Beispiel dafür, dass die charakteristischen Funktionen konvergieren können, ohne dass die zugehörigen Verteilungen konvergieren:

Beispiel 8.3.14. Sei $\mathbb{P}_n$ die Gleichverteilung auf dem Intervall $[-n,n]$, dann errechnet sich die charakteristische Funktion φ_n zu

$$\varphi_n(t) = \frac{1}{2n}\int_{-n}^{n}\mathrm{e}^{\mathrm{i}tx}\,\mathrm{d}x = \frac{1}{2nt}\big[\sin(tn)-\sin(-tn)\big] = \frac{\sin(tn)}{tn},\qquad t\in\mathbb{R}\setminus\{0\}.$$

Also konvergiert φ_n punktweise gegen die Funktion, die in 0 gleich 1 ist und Null sonst. Da diese Grenzfunktion in Null unstetig ist, sagt der Stetigkeitssatz nichts über die Konvergenz von $\mathbb{P}_n$ aus. Allerdings sieht man leicht, dass $(\mathbb{P}_n)_{n\in\mathbb{N}}$ nicht schwach konvergiert, denn für jedes beschränkte Intervall I gilt $\lim_{n\to\infty}\mathbb{P}_n(I)=0$. ◇

Ohne Mühe erhalten wir aus dem Stetigkeitssatz 8.3.13 auch den Poisson'schen Grenzwertsatz (siehe Satz 1.3.7):

Beispiel 8.3.15 (Poisson'scher Grenzwertsatz)**.** Es sei $\alpha>0$, und für jedes $n\in\mathbb{N}$ seien $X_1,\dots,X_n$ unabhängige, zum Parameter $\frac{\alpha}{n}$ Bernoulli-verteilte Zufallsgrößen. Also ist $S_n=$

$X_1 + \cdots + X_n$ zu den Parametern n und $\frac{\alpha}{n}$ Binomial-verteilt. Die charakteristische Funktion $\varphi_n(t) = (1+\frac{\alpha}{n}(\mathrm{e}^{\mathrm{i}t}-1))^n$ von S_n kennen wir aus Beispiel 8.3.2(e). Offensichtlich konvergiert $\varphi_n(t)$ gegen $e^{\alpha(\mathrm{e}^{\mathrm{i}t}-1)}$, und dies ist nach Beispiel 8.3.2(f) die charakteristische Funktion der Poisson-Verteilung zum Parameter α. Nach dem Stetigkeitssatz konvergiert also S_n schwach gegen diese Verteilung. $\diamond$

Nun kommt der erste Schritt auf dem Wege zum Beweis des Stetigkeitssatzes:

Lemma 8.3.16. *Sei φ die charakteristische Funktion eines Wahrscheinlichkeitsmaßes $\mathbb{P}$ auf $\mathbb{R}^d$, dann gilt*

$$|\varphi(t) - \varphi(s)|^2 \leq 2\big(1 - \Re(\varphi(t-s))\big) \qquad t, s \in \mathbb{R}^d.$$

Beweis. Mit Hilfe der Cauchy-Schwarz-Ungleichung errechnet man

$$\begin{aligned}
|\varphi(t) - \varphi(s)|^2 &= \Big| \int_{\mathbb{R}^d} \big(\mathrm{e}^{\mathrm{i}\langle t,x\rangle} - \mathrm{e}^{\mathrm{i}\langle s,x\rangle}\big) \,\mathbb{P}(\mathrm{d}x)\Big|^2 = \Big| \int_{\mathbb{R}^d} \big(\mathrm{e}^{\mathrm{i}\langle t-s,x\rangle} - 1\big)\mathrm{e}^{\mathrm{i}\langle s,x\rangle} \,\mathbb{P}(\mathrm{d}x)\Big|^2 \\
&\leq \int_{\mathbb{R}^d} \big|\mathrm{e}^{\mathrm{i}\langle t-s,x\rangle} - 1\big|^2 \,\mathbb{P}(\mathrm{d}x) \int_{\mathbb{R}^d} \big|\mathrm{e}^{\mathrm{i}\langle s,x\rangle}\big|^2 \,\mathbb{P}(\mathrm{d}x) \\
&= \int_{\mathbb{R}^d} \big(\mathrm{e}^{\mathrm{i}\langle t-s,x\rangle} - 1\big)\big(\mathrm{e}^{\mathrm{i}\langle s-t,x\rangle} - 1\big) \,\mathbb{P}(\mathrm{d}x) \\
&= 2\big(1 - \Re(\varphi(t-s))\big).
\end{aligned}$$

$\square$

Nun folgt der zweite Schritt. Wir nennen eine Familie $\{f_i : i \in I\}$ von reellen Abbildungen auf einem metrischen Raum $(E, \mathtt{d})$ *gleichgradig gleichmäßig stetig*, falls zu jedem $\varepsilon > 0$ ein $\delta > 0$ existiert, so dass für alle $i \in I$ und alle $s, t \in E$ mit $\mathtt{d}(s,t) < \delta$ gilt: $|f_i(t) - f_i(s)| < \varepsilon$. Zunächst benötigen wir eine elementare analytische Aussage:

Lemma 8.3.17. *Seien $(E, \mathtt{d})$ ein metrischer Raum und $f, f_1, f_2, \dots : E \to \mathbb{R}$ Funktionen, sodass $\lim_{n\to\infty} f_n(t) = f(t)$ für jedes $t \in E$. Falls $(f_n)_{n\in\mathbb{N}}$ gleichgradig gleichmäßig stetig ist, so konvergiert f_n sogar gleichmäßig auf jedem Kompaktum gegen f, und f ist gleichmäßig stetig.*

Beweis. Übungsaufgabe. $\square$

Lemma 8.3.18. *Sei $M \subset \mathcal{M}_1(\mathbb{R}^d)$ straff, dann ist die Menge $\{\widehat{\mathbb{P}} : \mathbb{P} \in M\}$ der zugehörigen charakteristischen Funktionen gleichgradig gleichmäßig stetig. Insbesondere ist jede charakteristische Funktion gleichmäßig stetig.*

Beweis. Sei $\varepsilon > 0$, dann ist zu zeigen, dass ein $\delta > 0$ existiert, so dass für alle $t, s \in \mathbb{R}$ mit $|t - s| \leq \delta$ und für jedes $\mathbb{P} \in M$ gilt: $|\widehat{\mathbb{P}}(t) - \widehat{\mathbb{P}}(s)| \leq \varepsilon$.

Da M straff ist, gibt es ein $N \in (0, \infty)$ mit $\mathbb{P}([-N, N]^d) > 1 - \varepsilon^2/6$ für alle $\mathbb{P} \in M$. Weiterhin existiert ein $\delta > 0$, sodass für alle $x \in [-N, N]^d$ und alle $u \in \mathbb{R}^d$ mit $|u| < \delta$ gilt: $|1 - \mathrm{e}^{\mathrm{i}\langle u,x\rangle}| < \varepsilon^2/6$. Daraus folgt für jedes $\mathbb{P} \in M$:

$$1 - \Re(\widehat{\mathbb{P}}(u)) \leq \int_{\mathbb{R}^d} |1 - \mathrm{e}^{\mathrm{i}\langle u,x\rangle}| \,\mathbb{P}(\mathrm{d}x) \leq \frac{\varepsilon^2}{3} + \int_{[-N,N]^d} |1 - \mathrm{e}^{\mathrm{i}\langle u,x\rangle}| \,\mathbb{P}(\mathrm{d}x) \leq \frac{\varepsilon^2}{3} + \frac{\varepsilon^2}{6} = \frac{\varepsilon^2}{2}.$$

Wenn wir dies für $u = t - s$ benutzen und Lemma 8.3.16 aufrufen, folgt die Aussage. $\square$

Nun kommt der Beweis des Stetigkeitssatzes:

Beweis des Satzes 8.3.13. *(a):* Es gilt $\lim_{n\to\infty}\varphi_n(t)=\varphi(t)$ für jedes $t\in\mathbb{R}^d$, denn die Abbildung $\mathbb{R}^d\ni x\mapsto \mathrm{e}^{\mathrm{i}\langle t,x\rangle}$ ist stetig und beschränkt. Nach Lemma 8.3.18 ist $(\varphi_n)_{n\in\mathbb{N}}$ gleichgradig gleichmäßig stetig. Daraus folgt mit Hilfe von Lemma 8.3.17, dass $(\varphi_n)_{n\in\mathbb{N}}$ sogar lokal gleichmäßig gegen φ konvergiert.

(b): Wir wissen aus Satz 8.3.10, dass die Vereinigung der Familien der Abbildungen $\cos(\langle x,\cdot\rangle)$ mit $x\in\mathbb{R}^d$ und der Abbildungen $\sin(\langle x,\cdot\rangle)$ mit $x\in\mathbb{R}^d$ trennend ist. Nach Satz 8.2.26 reicht es also, die Straffheit der Folge $(\mathbb{P}_n)_{n\in\mathbb{N}}$ zu zeigen. Dies tun wir nur für $d=1$, der Fall allgemeiner Dimension geht genauso.

Wegen $\varphi_n(0)=1$ für jedes $n\in\mathbb{N}$ gilt $f(0)=1$. Wir betrachten die Abbildung $h\colon\mathbb{R}\to[0,\infty)$, die durch $h(x)=1-\frac{1}{x}\sin x$ für $x\neq 0$ und $h(0)=0$ definiert ist, dann ist h stetig differenzierbar. Man sieht leicht, dass $\alpha=\inf_{|x|\geq 1}h(x)=1-\sin 1>0$ ist. Für jedes $K>0$ errechnen wir

$$\begin{aligned}\mathbb{P}_n\big([-K,K]^{\mathrm{c}}\big)&\leq\frac{1}{\alpha}\int_{[-K,K]^{\mathrm{c}}}h(x/K)\,\mathbb{P}_n(\mathrm{d}x)\leq\frac{1}{\alpha}\int_{\mathbb{R}}h(x/K)\,\mathbb{P}_n(\mathrm{d}x)\\&=\frac{K}{\alpha}\int_{\mathbb{R}}\Big(\int_0^{1/K}(1-\cos(xt))\,\mathrm{d}t\Big)\,\mathbb{P}_n(\mathrm{d}x)\\&=\frac{K}{\alpha}\int_0^{1/K}\int_{\mathbb{R}}(1-\cos(xt))\,\mathbb{P}_n(\mathrm{d}x)\mathrm{d}t\\&=\frac{K}{\alpha}\int_0^{1/K}\big(1-\Re(\varphi_n(t))\big)\,\mathrm{d}t,\end{aligned}$$

wobei wir die Markov'sche Ungleichung und den Satz von Fubini benutzten. Mit Hilfe des Satzes von der Majorisierten Konvergenz bekommen wir nun

$$\begin{aligned}\limsup_{n\to\infty}\mathbb{P}_n\big([-K,K]^{\mathrm{c}}\big)&\leq\limsup_{n\to\infty}\frac{1}{\alpha}\int_0^1\big(1-\Re(\varphi_n(t/K))\big)\,\mathrm{d}t\\&=\frac{1}{\alpha}\int_0^1\Big(\lim_{n\to\infty}\big(1-\Re(\varphi_n(t/K))\big)\Big)\,\mathrm{d}t=\frac{1}{\alpha}\int_0^1\big(1-\Re(f(t/K))\big)\,\mathrm{d}t.\end{aligned}$$

Da f stetig ist mit $f(0)=1$, konvergiert das letzte Integral gegen 0 für $K\to\infty$. Also ist $(\mathbb{P}_n)_{n\in\mathbb{N}}$ straff. $\square$

Nun können wir das Hauptergebnis dieses Abschnitts beweisen:

Satz 8.3.19 (Zentraler Grenzwertsatz). *Sei $(X_i)_{i\in\mathbb{N}}$ eine Folge unabhängiger, identisch verteilter Zufallsgrößen mit Erwartungswert Null und Varianz Eins. Wir setzen $S_n=X_1+\dots+X_n$. Dann konvergiert die Verteilung von $\frac{1}{\sqrt{n}}S_n$ schwach gegen die Standardnormalverteilung $\mathcal{N}$. Insbesondere gilt*

$$\lim_{n\to\infty}\mathbb{P}\big(a\sqrt{n}\leq S_n\leq b\sqrt{n}\big)=\frac{1}{\sqrt{2\pi}}\int_a^b\mathrm{e}^{-x^2/2}\,\mathrm{d}x,\qquad-\infty\leq a<b\leq\infty.$$

Beweis. Es sei φ die charakteristische Funktion von X_1. Dann ist nach Lemma 8.3.4(b)

$$\varphi(t)=1-\frac{1}{2}t^2+o(t^2),\qquad t\to 0.$$

Sei φ_n die charakteristische Funktion von $\frac{1}{\sqrt{n}}S_n$. Nach Lemma 8.3.3 gilt für jedes $t \in \mathbb{R}$ im Limes $n \to \infty$:

$$\varphi_n(t) = \varphi\Big(\frac{t}{\sqrt{n}}\Big)^n = \Big(1 - \frac{t^2(1+o(1))}{2n}\Big)^n \to \mathrm{e}^{-t^2/2} = \widehat{\mathcal{N}}(t).$$

Also haben wir gesehen, dass die charakteristische Funktion φ_n von $\frac{1}{\sqrt{n}}S_n$ gegen die charakteristische Funktion $t \mapsto \mathrm{e}^{-t^2/2}$ der Standardnormalverteilung (siehe Beispiel 8.3.2) konvergiert. Nach dem Stetigkeitssatz 8.3.13 ist der Beweis der schwachen Konvergenz der Verteilung von $\frac{1}{\sqrt{n}}S_n$ gegen $\mathcal{N}$ beendet. Die Zusatzaussage folgt aus Lemma 8.2.14, weil $\mathcal{N}$ eine Dichte hat. □

Anwendungen des Zentralen Grenzwertsatzes (genauer: seines Spezialfalles für Bernoulli-Größen, des Satzes von Moivre-Laplace) wurden in Abschnitt 5.2 diskutiert. Man beachte, dass die Aussage des Zentralen Grenzwertsatzes trivial ist für standardnormalverteilte Zufallsgrößen $X_1, X_2, \dots$, denn wir wissen aus Lemma 7.2.17(i), dass dann $\frac{1}{\sqrt{n}}(X_1 + \dots + X_n)$ ebenfalls exakt standardnormalverteilt ist für jedes $n \in \mathbb{N}$. Dies ist allerdings eine sehr spezielle Eigenschaft der Normalverteilung.

Die Voraussetzung der Unabhängigkeit und identischen Verteilung im Zentralen Grenzwertsatz können natürlich wesentlich abgeschwächt werden. Hinreichende Bedingungen, die oft in der Literatur erwähnt werden, sind die sogenannten *Lindeberg-Feller-Bedingungen*, die wir allerdings nicht behandeln werden.

Um einen mehrdimensionalen Zentralen Grenzwertsatz zu erhalten, ist die folgende Aussage hilfreich:

Lemma 8.3.20 (Satz von Cramér-Wold)**.** *Seien $X, X_1, X_2, \dots$ Zufallsvektoren mit Werten im $\mathbb{R}^d$. Dann gilt $X_n \Longrightarrow X$ genau dann, wenn für jedes $\lambda \in \mathbb{R}^d$ gilt: $\langle \lambda, X_n \rangle \Longrightarrow \langle \lambda, X \rangle$.*

Beweis. Übungsaufgabe. Die Richtung '$\Longrightarrow$' folgt aus Lemma 8.2.11(b). Beim Beweis der Richtung '$\Longleftarrow$' zeigt man zunächst die Straffheit und betrachtet dann die charakteristischen Funktionen, oder man benutzt gleich den Stetigkeitssatz. □

Mit Hilfe von Lemmas 8.3.20 und 8.3.12 erhält man aus dem Zentralen Grenzwertsatz 8.3.19 als eine Übungsaufgabe die folgende mehrdimensionale Variante:

Satz 8.3.21 (Zentraler Grenzwertsatz im $\mathbb{R}^d$)**.** *Sei $(X_i)_{i \in \mathbb{N}}$ eine Folge unabhängiger, identisch verteilter $\mathbb{R}^d$-wertiger Zufallsvektoren mit Erwartungswertvektor Null und positiv definiter Kovarianzmatrix $\Sigma \in \mathbb{R}^{d \times d}$. Dann konvergiert die Folge der*

$$\frac{1}{\sqrt{n}}(X_1 + \dots + X_n), \qquad n \in \mathbb{N},$$

in Verteilung gegen die Normalverteilung $\mathcal{N}(0, \Sigma)$.

Wir ziehen noch ein paar weitere Folgerungen aus dem Stetigkeitssatz. Man fragt sich vielleicht, ob man einer Funktion ansehen kann, ob sie die charakteristische Funktion einer Wahrscheinlichkeitsverteilung ist. Hier ist ein hinreichendes Kriterium:

Satz 8.3.22 (Satz von Polya)**.** *Sei $f : \mathbb{R} \to \mathbb{R}$ stetig mit $f(0) = 1$ und $f(-x) = f(x)$ für jedes $x \in \mathbb{R}$. Ferner sei f auf $[0, \infty)$ konvex und fallend. Dann ist f die charakteristische Funktion einer Wahrscheinlichkeitsverteilung.*

Beweisskizze. Wir können f auf $[0,\infty)$ durch konvexe fallende Polygonzüge approximieren, indem wir etwa $f_n(k/n) = f(k/n) - f(n)$ setzen für $k = 0, 1, \dots, n^2$ und auf $[n,\infty)$ konstant stetig mit Null fortsetzen. Für $x < 0$ setzen wir $f_n(x) = f_n(-x)$. Offensichtlich konvergiert dann f_n punktweise gegen f, denn $\lim_{n\to\infty} f(n) = 0$. Man kann durch eine explizite Konstruktion (die auf Beispiel 8.3.2(g) basiert) zeigen, dass f_n die charakteristische Funktion einer Wahrscheinlichkeitsverteilung ist. Dann ist der Beweis mit Hilfe des Stetigkeitssatzes beendet. $\square$

Beispiel 8.3.23 (stabile Verteilungen)**.** Für jedes $\alpha \in (0,2]$ und $a > 0$ ist $\varphi_{a,\alpha}(t) = \mathrm{e}^{-a|t|^\alpha}$ die charakteristische Funktion eines symmetrischen Wahrscheinlichkeitsmaßes $\mu_{a,\alpha}$ auf $\mathbb{R}$. Dies folgt für $\alpha \leq 1$ aus dem Satz von Polya. Der Spezialfall $\alpha = 2$ ist der der Normalverteilung.

Diese Verteilungen haben die folgende spezielle Eigenschaft: Sie sind *α-stabil*, was bedeutet, dass für jedes n und für je n unabhängige Zufallsgrößen $X_1, \dots, X_n$ mit Verteilung $\mu_{a,\alpha}$ gilt, dass die Summe $X_1 + \cdots + X_n$ die Verteilung von $n^{1/\alpha} X_1$ hat. Dies folgt direkt aus der Funktionalgleichung $\varphi_{a,\alpha}(t)^n = \varphi_{a,\alpha}(n^{1/\alpha} t)$, zusammen mit der Eindeutigkeit der charakteristischen Funktion einer Verteilung und Satz 8.3.3(d).

Die Verteilung $\mu_{a,\alpha}$ tritt unter bestimmten Umständen als universeller Verteilungsgrenzwert von $n^{-1/\alpha}(X_1 + \cdots + X_n)$ auf, wenn $X_1, X_2, \dots$ unabhängige, identisch verteilte Zufallsgrößen sind mit $\alpha = \sup\{\gamma > 0 \colon \mathbb{E}[|X_1|^\gamma < \infty\}$. $\diamond$

Kapitel 9

Markovketten

In diesem Kapitel behandeln wir einen der wichtigsten stochastischen Prozesse, die Markovketten auf einem diskreten Raum in diskreter Zeit. Man stelle sich ein Teilchen vor, das sich durch eine höchstens abzählbare Menge I zufällig bewegt und zu den Zeitpunkten $1, 2, 3, \dots$ jeweils zu einem (eventuell anderen) Punkt springt. Die besondere Eigenschaft der Sprungentscheidungen, die die Markoveigenschaft ausmacht, ist die Tatsache, dass diese Entscheidung nur von der aktuellen Position des Teilchens abhängt, aber nicht von dem Verlauf des ganzen bisherigen Pfades, den das Teilchen zurückgelegt hat.

Markovketten spielen eine wichtige Rolle bei der Modellierung vieler zeitlich sich entwickelnder Prozesse: bei Mischungsvorgängen, bei Sortieralgorithmen und anderen stochastischen Algorithmen, bei der Modellierung physikalischer Prozesse oder von Finanzmärkten und Vielem mehr.

Im gesamten Kapitel sei I eine nichtleere, endliche oder höchstens abzählbar unendliche Menge.

9.1 Definition und einfache Eigenschaften

Wir beginnen mit der Definition der Markoveigenschaft: Selbst wenn der gesamte bisherige Verlauf bekannt ist, hat nur die aktuelle Position einen Einfluss auf die Sprungentscheidung.

Definition 9.1.1 (Markoveigenschaft)**.** *Wir sagen, eine (endliche oder unendliche) Folge $X_0, X_1, X_2, \dots$ von I-wertigen Zufallsgrößen* besitzt die Markoveigenschaft*, wenn für jedes $n \in \mathbb{N}_0$ und alle $i_0, i_1, \dots, i_{n+1} \in I$ gilt:*

$$\mathbb{P}(X_{n+1} = i_{n+1} \mid X_n = i_n, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) = \mathbb{P}(X_{n+1} = i_{n+1} \mid X_n = i_n), \qquad (9.1.1)$$

sofern alle auftretenden bedingten Wahrscheinlichkeiten wohldefiniert sind.

Das legt nahe, dass eine Markovkette im Wesentlichen durch die bedingten Wahrscheinlichkeiten $\mathbb{P}(X_{n+1} = i_{n+1} \mid X_n = i_n)$ festgelegt wird. Ihre Kollektion ist also ein ganz wesentliches Objekt:

Definition 9.1.2 (stochastische Matrix)**.** *Eine Matrix $P = (p_{i,j})_{i,j \in I}$ heißt* stochastisch*, falls $p_{i,j} \in [0, 1]$ für alle $i, j \in I$ gilt und $\sum_{j \in I} p_{i,j} = 1$ für jedes $i \in I$ gilt.*

Wir werden im Folgenden immer still schweigend davon ausgehen, dass die Koeffizienten einer stochastischen Matrix P mit $p_{i,j}$ bezeichnet sind.

Die Sprungwahrscheinlichkeiten einer Folge von Zufallsgrößen, die die Markoveigenschaft besitzt, sind also durch stochastische Matrizen gegeben, die *a priori* noch von dem Zeitpunkt des Sprunges abhängen dürfen. Wir werden im Folgenden nur solche Folgen betrachten, deren Sprungwahrscheinlichkeiten nicht von diesem Zeitpunkt abhängen:

Definition 9.1.3 (Markovkette). *Sei P eine stochastische Matrix. Eine (endliche oder unendliche) Folge $X_0, X_1, X_2, \dots$ von I-wertigen Zufallsgrößen heißt eine (zeitlich homogene)* Markovkette mit Übergangsmatrix *P, falls für alle $n \in \mathbb{N}_0$ und alle $i_0, i_1, \dots, i_{n+1} \in I$ mit $\mathbb{P}(X_n = i_n, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) > 0$ gilt:*

$$\mathbb{P}(X_{n+1} = i_{n+1} \mid X_n = i_n, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) = p_{i_n, i_{n+1}}. \tag{9.1.2}$$

Die Einträge $p_{i,j}$ von P heißen die Übergangswahrscheinlichkeiten*, und die* Startverteilung *ν der Kette ist definiert durch $\nu(i) = \mathbb{P}(X_0 = i)$ für $i \in I$.*

Eine Startverteilung ist also eine Wahrscheinlichkeitsverteilung, oder kurz eine Verteilung, auf I, genau wie jede Zeile einer stochastischen Matrix. Wir schreiben auch oft $\mathbb{P}_\nu$ an Stelle von $\mathbb{P}$, um die Startverteilung zu betonen. Im Fall, dass ν in einem $i \in I$ konzentriert ist (also $\nu(i) = 1$), schreiben wir $\mathbb{P}_i$. Die Elemente von I nennt man auch oft *Zustände*, die Menge I selber den *Zustandsraum.*

Beispiel 9.1.4 (Irrfahrten). Es seien I eine kommutative additive Gruppe (etwa $\mathbb{Z}^d$ oder $\mathbb{R}^d$) und $(Y_n)_{n\in\mathbb{N}}$ eine Folge unabhängiger und identisch verteilter I-wertiger Zufallsgrößen. Wir betrachten die durch $X_0 = 0$ und $X_n = \sum_{k=1}^n Y_k$ definierte Partialsummenfolge $(X_n)_{n\in\mathbb{N}_0}$. Als Übungsaufgabe beweist man, dass $(X_n)_{n\in\mathbb{N}_0}$ eine Markovkette ist, deren Übergangsmatrix die Koeffizienten $p_{i,j} = \mathbb{P}(Y_1 = j-i)$ besitzt, die also nur von der Differenz der Indizes abhängt. Man nennt $(X_n)_{n\in\mathbb{N}_0}$ eine *Irrfahrt* auf I (wobei allerdings der englische Begriff *random walk* sicherlich sinnvoller ist). $\diamond$

Es folgen Charakterisierungen von Markovketten. Notationell ist es angenehm, für $s < t$ den Pfad $(X_s, X_{s+1}, \dots, X_t)$ mit $X_{[s,t]}$ abzukürzen, ebenso schreiben wir für (nicht zufällige) Vektoren $i_{[s,t]}$ statt $(i_s, i_{s+1}, \dots, i_t) \in I^{t-s+1}$.

Satz 9.1.5. *Es seien $(X_n)_{n\in\mathbb{N}_0}$ eine Folge von I-wertigen Zufallsgrößen, ν eine Verteilung auf I und P eine stochastische Matrix.*

(a) $(X_n)_{n\in\mathbb{N}_0}$ ist genau dann eine Markovkette mit Übergangsmatrix P und Startverteilung ν, wenn für alle $n \in \mathbb{N}_0$ und alle $i_0, i_1, \dots, i_n \in I$ gilt

$$\mathbb{P}(X_0 = i_0, X_1 = i_1, \dots, X_n = i_n) = \nu(i_0) p_{i_0,i_1} p_{i_1,i_2} \cdots p_{i_{n-1},i_n}. \tag{9.1.3}$$

(b) Falls $(X_n)_{n\in\mathbb{N}_0}$ eine Markovkette ist, so gilt für alle $n < m$, $i_n \in I$ und alle $A \subset I^n$ mit $\mathbb{P}(X_{[0,n-1]} \in A, X_n = i_n) > 0$ und für alle $B \subset I^{m-n}$:

$$\mathbb{P}(X_{[n+1,m]} \in B \mid X_{[0,n-1]} \in A, X_n = i_n) = \mathbb{P}(X_{[n+1,m]} \in B \mid X_n = i_n). \tag{9.1.4}$$

Beweis. (a) Der Beweis von (9.1.3) wird leicht mit einer Vollständigen Induktion über n geführt, und eine Folge $(X_n)_{n\in\mathbb{N}_0}$, die (9.1.3) für alle $n \in \mathbb{N}_0$ und alle $i_0, i_1, \dots, i_n \in I$ erfüllt, wird mit Hilfe der Definition der bedingten Wahrscheinlichkeit leicht als eine Markovkette identifiziert.

(b) Mit Hilfe der Definition der bedingten Wahrscheinlichkeit und Teil (a) errechnet man

$$\begin{aligned}\mathbb{P}(X_{[n+1,m]} \in B \mid X_{[0,n-1]} \in A, X_n = i_n) &= \frac{\mathbb{P}(X_{[n+1,m]} \in B, X_{[0,n-1]} \in A, X_n = i_n)}{\mathbb{P}(X_{[0,n-1]} \in A, X_n = i_n)}\\ &= \frac{\sum_{i_{[n+1,m]}\in B}\sum_{i_{[0,n-1]}\in A}\nu(i_0)p_{i_0,i_1}\cdots p_{i_{m-1},i_m}}{\sum_{i_{[0,n-1]}\in A}\nu(i_0)p_{i_0,i_1}\cdots p_{i_{n-1},i_n}}\\ &= \sum_{i_{[n+1,m]}\in B} p_{i_n,i_{n+1}}p_{i_{n+1},i_{n+2}}\cdots p_{i_{m-1},i_m}.\end{aligned}$$

Der Ausdruck auf der rechten Seite hängt nicht von A ab. Wir können insbesondere $A = I^n$ setzen und erhalten

$$\mathbb{P}(X_{[n+1,m]} \in B \mid X_n = i_n) = \sum_{i_{[n+1,m]}\in B} p_{i_n,i_{n+1}}p_{i_{n+1},i_{n+2}}\cdots p_{i_{m-1},i_m}.$$

Wenn wir dies in der obigen Rechnung wieder einsetzen, ist der Beweis von (9.1.4) beendet. □

Die Aussage in (b) kann man auch einprägsam wie folgt formulieren:

Korollar 9.1.6 (Unabhängigkeit von Zukunft und Vergangenheit bei gegebener Gegenwart). *Falls $(X_n)_{n\in\mathbb{N}_0}$ eine Markovkette ist, so gilt für alle $n < m$, $i_n \in I$ mit $\mathbb{P}(X_n = i_n) > 0$ und alle $A \subset I^n$ und $B \subset I^{m-n}$:*

$$\mathbb{P}(X_{[0,n-1]} \in A, X_{[n+1,m]} \in B \mid X_n = i_n) = \mathbb{P}(X_{[0,n-1]} \in A \mid X_n = i_n)\mathbb{P}(X_{[n+1,m]} \in B \mid X_n = i_n). \tag{9.1.5}$$

Beweis. Im Fall $\mathbb{P}(X_{[0,n-1]} \in A, X_n = i_n) > 0$ ergibt sich (9.1.5) direkt aus (9.1.4) nach Multiplikation mit $\mathbb{P}(X_{[0,n-1]} \in A \mid X_n = i_n)$. Ansonsten steht Null auf beiden Seiten von (9.1.5), und die Aussage gilt trivialerweise. □

Man überlege sich an einem Beispiel, dass die Aussage von Korollar 9.1.6 im Allgemeinen falsch wird, wenn man das Ereignis $\{X_n = i_n\}$ ersetzt durch $\{X_n \in C\}$ für beliebige Teilmengen C von I.

Wir schneiden kurz die Frage der Existenz und Konstruktion von Markovketten an. Eine abschließende Diskussion wird erst im Rahmen der allgemeinen Wahrscheinlichkeitstheorie möglich sein.

Bemerkung 9.1.7 (Konstruktion von Markovketten). Eine endlich lange Markovkette $(X_0, X_1, \dots, X_n)$ im Sinne von Definition 9.1.3 kann leicht auf einem diskreten Wahrscheinlichkeitsraum konstruiert werden. Wir setzen $\Omega_n = I^{n+1}$ und definieren die Einzelwahrscheinlichkeiten als $p((i_0, \dots, i_n)) = \nu(i_0)p_{i_0,i_1}\dots p_{i_{n-1},i_n}$, wobei ν eine Verteilung auf I ist und P eine stochastische Matrix. Dann bilden die Projektionsabbildungen $X_i\colon I^{n+1} \to I$ eine Markovkette mit Startverteilung ν und Übergangsmatrix P.

Man beachte, dass dieser Wahrscheinlichkeitsraum von der Länge der Kette abhängt und dass man genauer $\mathbb{P}_n$ statt $\mathbb{P}$ für das induzierte Wahrscheinlichkeitsmaß schreiben müsste. Allerdings überzeugt man sich leicht davon, dass die ersten $n+1$ Zufallsgrößen einer Markovkette der Länge $m > n$ selber eine Markovkette der Länge n bilden.[1] Bei der Untersuchung einer Markovkette in endlicher Zeit reicht es also für viele Zwecke aus, eine genügend späte Zeit m zu fixieren und den für dieses m konstruierten Wahrscheinlichkeitsraum zu Grunde zu legen.

Eine andere Frage ist allerdings, wie *unendlich lange* Markovketten $(X_n)_{n\in\mathbb{N}_0}$ konstruiert werden. Dies erfordert maßtheoretische Vorbereitungen und wird erst später erschöpfend (positiv) beantwortet werden, siehe den Satz von Ionescu Tulcea. Wir werden daher im Folgenden aus Bequemlichkeit immer von einer unendlich langen Markovkette ausgehen und den benutzten Wahrscheinlichkeitsraum nicht mit einer endlichen Länge indizieren. ◇

Der Zusammenhang zwischen dem Markovschen Mechanismus und der Matrixmultiplikation ist sehr eng, wie wir uns kurz klar machen wollen.

Bemerkung 9.1.8 (Potenzen stochastischer Matrizen)**.** Stochastische Matrizen P und Q kann man ohne Probleme im Sinne der Matrixmultiplikation mit einander multiplizieren, und das Produkt ist ebenfalls eine stochastische Matrix. Die Koeffizienten der n-ten Potenz P^n von P bezeichnen wir mit $P^n = (p_{i,j}^{(n)})_{i,j\in I}$. Es ist $P^0 = (\delta_{i,j})_{i,j\in I}$ die Einheitsmatrix, wobei $\delta_{i,j}$ das Kroneckersymbol bezeichnet. Wenn man die Gleichung $P^nP^m = P^{n+m}$ ausschreibt, erhält man die sogenannten *Chapman-Kolmogorov-Gleichungen*:

$$p_{i,j}^{(n+m)} = \sum_{k\in I} p_{i,k}^{(n)} p_{k,j}^{(m)}, \qquad i,j \in I. \tag{9.1.6}$$

Insbesondere haben wir

$$p_{i,j}^{(n+m)} \geq p_{i,k}^{(n)} p_{k,j}^{(m)}, \qquad i,j,k \in I, n,m \in \mathbb{N}_0. \tag{9.1.7}$$

Auf Grund des folgenden Lemmas 9.1.9 nennt man die Koeffizienten $p_{i,j}^{(n)}$ von P^n auch die *n-stufigen Übergangswahrscheinlichkeiten*. ◇

Wir stellen uns eine Verteilung ν als *Zeilenvektoren* vor (wie auch z. B. die Zeilen einer stochastischen Matrix), so dass das Matrixprodukt νP wohldefiniert ist, also $(\nu P)_j = \sum_{i\in I} \nu(i) p_{i,j}$.

Lemma 9.1.9. *Es sei $(X_n)_{n\in\mathbb{N}_0}$ eine Markovkette mit Startverteilung ν und Übergangsmatrix P. Dann gilt $\mathbb{P}_\nu(X_n = j) = (\nu P^n)_j$ für alle $n \in \mathbb{N}$ und alle $j \in I$. Insbesondere gilt $\mathbb{P}_i(X_n = j) = p_{i,j}^{(n)}$ für alle $i,j \in I$.*

Beweis. Wir summieren die Gleichung (9.1.3) (mit $j = i_n$) über alle $i_0, \dots, i_{n-1} \in I$ und beachten die Regeln der Matrixmultiplikation. □

Die Verteilung einer Markovkette zum Zeitpunkt n ist also nichts Anderes als die n-te Potenz der Übergangsmatrix, multipliziert von links mit der Startverteilung. Wir halten noch fest, wie sich dies auf zeitliche Verschiebung auswirkt:

Korollar 9.1.10. *Es sei $(X_n)_{n\in\mathbb{N}_0}$ eine Markovkette mit Übergangsmatrix P. Dann gilt für alle $n,m \in \mathbb{N}_0$ und alle $i,j \in I$ mit $\mathbb{P}(X_m = i) > 0$:*

$$\mathbb{P}(X_{m+n} = j \mid X_m = i) = p_{i,j}^{(n)}.$$

[1] Diese Eigenschaft nennt man die *Konsistenz* der Kette, eine Eigenschaft, der wir uns in Abschnitt 10 viel allgemeiner widmen werden.

9.2 Beispiele

Ein sehr allgemeines Prinzip, nach dem man Markovketten konstruieren kann (siehe Übungsaufgabe), ist das folgende. Mit einer Folge von unabhängigen identisch verteilten Zufallsgrößen $Y_1, Y_2, Y_3, \dots$ mit Werten in einer beliebigen abzählbaren Menge J und mit einer Funktion $f\colon I \times J \to I$ setzt man rekursiv $X_{n+1} = f(X_n, Y_{n+1})$ für $n \in \mathbb{N}_0$. Die Kette wird also rekursiv fortgesetzt, indem man den aktuellen Wert der Kette mit einem festgelegten Mechanismus einem unabhängigen Zufall unterwirft.

Beispiel 9.2.1 (unabhängige identisch verteilte Folgen)**.** Wenn $(X_n)_{n\in\mathbb{N}_0}$ eine Folge unabhängiger und identisch verteilter I-wertiger Zufallsgrößen ist, so ist sie auch eine Markovkette. Die Übergangsmatrix ist gegeben durch $p_{i,j} = q_j$ für alle $i, j \in I$, wobei q die Verteilung von X_0 ist. Natürlich ist q auch die Startverteilung. Andersherum ist eine Markovkette, deren Übergangsmatrix identische Zeilen besitzt (d. h. deren $p_{i,j}$ nicht von i abhängen), eine Folge unabhängiger und identisch verteilter Zufallsgrößen. ◇

Beispiel 9.2.2 (Irrfahrten auf Gruppen)**.** Es sei $I = G$ eine abzählbare Gruppe, die wir multiplikativ schreiben wollen und nicht unbedingt als kommutativ voraus setzen wollen. Ferner sei μ eine beliebige Verteilung auf G. Wir definieren $p_{g,h} = \mu(g^{-1}h)$ für alle $g, h \in G$. Wegen der Gruppeneigenschaft ist für jedes g die Abbildung $h \mapsto g^{-1}h$ bijektiv auf G, und es gilt

$$\sum_{h\in G} p_{g,h} = \sum_{h\in G} \mu(g^{-1}h) = \sum_{h'\in G} \mu(h') = 1,$$

also ist $P = (p_{g,h})_{g,h\in G}$ eine stochastische Matrix. Die zugehörige Markovkette heißt die *μ-Irrfahrt* auf G. Dieses Beispiel ist die multiplikative Variante der Irrfahrt von Beispiel 9.1.4 (bis auf die dort voraus gesetzte Kommutativität). Mit Hilfe einer Folge von unabhängigen, nach μ verteilten Zufallsgrößen $Y_1, Y_2, Y_3, \dots$ erhält man eine μ-Irrfahrt, indem man rekursiv setzt: $X_0 = 1$ und $X_{n+1} = X_n Y_n$, denn dann gilt für alle $n \in \mathbb{N}_0$ und alle $g, h \in G$ mit $\mathbb{P}(X_n = g) > 0$:

$$\mathbb{P}(X_{n+1} = h \mid X_n = g) = \mathbb{P}(gY_n = h) = \mu(g^{-1}h) = p_{g,h}.$$

Die Wahl von $I = G$ als die Menge der Permutationen einer endlichen Menge führt zum Beispiel auf ein Modell für die Mischung eines Kartenstapels; man beachte, dass diese Gruppe nicht kommutativ ist. ◇

Beispiel 9.2.3 (eindimensionale Irrfahrt)**.** Der folgende Spezialfall von Beispiel 9.1.4 wird die *eindimensionale Irrfahrt* genannt. Setze $I = \mathbb{Z}$, und Y_n nehme die Werte 1 und -1 mit Wahrscheinlichkeiten p und $q = 1 - p$ an, wobei $p \in [0, 1]$ ein Parameter sei. Dann beschreibt die Markovkette $(X_n)_{n\in\mathbb{N}_0}$ den Weg eines Teilchens durch die diskrete Achse mit unabhängigen Sprüngen, wobei es zu jedem Zeitpunkt mit Wahrscheinlichkeit p um eine Einheit nach rechts springt und sonst nach links. Die Übergangsmatrix besitzt die Einträge p auf der rechten Nebendiagonalen und $1 - p$ auf der linken, ansonsten besteht sie aus Nullen. Im Fall $p = \frac{1}{2}$ wird die Irrfahrt *symmetrisch* genannt. ◇

Beispiel 9.2.4 (Irrfahrt auf $\mathbb{Z}^d$)**.** Die d-dimensionale Variante von Beispiel 9.2.3 (die ebenfalls ein Spezialfall von Beispiel 9.1.4 ist) ist auf $I = \mathbb{Z}^d$ gegeben, indem die Übergangsmatrix durch $p_{i,j} = \frac{1}{2d}$ für $|i - j| = 1$ fest gelegt wird. (Hier ist $|\cdot|$ die ℓ^1-Norm auf $\mathbb{Z}^d$.) Die zugehörige Markovkette beschreibt einen Nächstnachbarschaftspfad durch $\mathbb{Z}^d$, wobei jeder Nachbar mit gleicher Wahrscheinlichkeit ausgewählt wird, unabhängig von allen anderen Sprungentscheidungen. ◇

Beispiel 9.2.5 (Irrfahrten auf $\{0,\dots,N\}$)**.** Ähnlich wie in Beispiel 9.2.3 soll ein Sprung innerhalb von $I=\{0,\dots,N\}$ mit Wahrscheinlichkeit p zum rechten Nachbarn und mit Wahrscheinlichkeit $1-p$ zum linken ausgeführt werden. Für die Sprungentscheidungen an den Rändern 0 und N müssen wir allerdings gesonderte Vereinbarungen treffen, und es gibt dafür mehrere Möglichkeiten. Ein Randpunkt, sagen wir 0, heißt *absorbierend*, falls $p_{0,0}=1$ gilt, falls also das springende Teilchen nie mehr von der 0 sich entfernen kann. Im Fall $p_{0,1}=1$, wenn also das Teilchen sofort wieder unweigerlich zurück springen muss, heißt der Randpunkt 0 *reflektierend*. ◇

Beispiel 9.2.6 (Polyas Urnenschema)**.** In einer Urne liegen gewisse (endliche) Anzahlen roter und schwarzer Kugeln. Zu jedem Zeitpunkt wird eine Kugel zufällig gezogen und zusammen mit einer neuen Kugel der selben Farbe in die Urne zurück gelegt. Dann bildet das Paar der Anzahlen der roten und der schwarzen Kugeln zu den Zeitpunkten $0,1,2,\dots$ eine Markovkette auf $\mathbb{N}_0^2$. Die Übergangswahrscheinlichkeiten sind gegeben durch $p_{(r,s),(r+1,s)}=\frac{r}{r+s}$ und $p_{(r,s),(r,s+1)}=\frac{s}{r+s}$; alle anderen sind Null. ◇

Beispiel 9.2.7 (Ehrenfests Urnenmodell)**.** Insgesamt N Kugeln liegen in zwei Urnen. Zu jedem Zeitpunkt $1,2,\dots$ wählen wir eine der Kugeln mit gleicher Wahrscheinlichkeit und lassen sie die Urne wechseln. Dann ist die Anzahl der Kugeln in der linken Urne zum Zeitpunkt n eine Markovkette auf $I=\{0,\dots,N\}$ im Zeitparameter n. Die Übergangsmatrix P ist gegeben durch $p_{k,k-1}=\frac{k}{N}$ und $p_{k,k+1}=1-\frac{k}{N}$, und alle anderen Übergangswahrscheinlichkeiten sind Null. Siehe Beispiel 9.7.4 für interessante Eigenschaften dieses Modells. ◇

Beispiel 9.2.8 (Bernoulli-Laplace-Diffusionsmodell)**.** In zwei Behältern A und B befinden sich insgesamt w weiße und s schwarze Kugeln, wobei s Kugeln in A liegen, und es sei $w\le s$. Zu den diskreten Zeitpunkten $n=1,2,3,\dots$ wird jeweils in A und in B eine Kugel zufällig ausgewählt und in den jeweils anderen Behälter gelegt. Dann ist die Anzahl der weißen Kugeln in A eine Markovkette auf $\{0,1,\dots,w\}$ im Zeitparameter n. Siehe Beispiel 9.5.9 für interessante Eigenschaften dieses Modells. ◇

Beispiel 9.2.9 (Irrfahrten-Maxima)**.** Falls $(S_n)_{n\in\mathbb{N}_0}$ eine Irrfahrt auf $\mathbb{Z}$ wie in Beispiel 9.1.4 ist, dann bildet die Folge der Maxima $M_n=\max\{S_0,S_1,\dots,S_n\}$ im Allgemeinen keine Markovkette, aber die Folge $(M_n,S_n)_{n\in\mathbb{N}_0}$ der Paare (Übungsaufgabe). ◇

9.3 Klasseneigenschaften, Rekurrenz, Transienz

In diesem Abschnitt sei P eine stochastische Matrix auf I. Wir wollen die Frage untersuchen, ob die zugehörige Markovkette gegebene Punkte in I mit Sicherheit besucht oder nicht. Es wird sich in diesem Abschnitt heraus stellen, dass der Zustandsraum zerlegt werden kann in rekurrente und transiente Klassen. Eine in einer rekurrenten Klasse gestartete Markovkette verlässt die Klasse nie und besucht jeden Punkt dieser Klasse mit Sicherheit.

Wir müssen zunächst die Frage untersuchen, ob ein gegebener Punkt überhaupt mit positiver Wahrscheinlichkeit jemals erreicht werden kann. Diese Unterscheidung induziert auf natürliche Weise eine Einteilung von I in Klassen.

Definition 9.3.1 (erreichbar)**.** *Ein Punkt $j\in I$ heißt von einem Punkt $i\in I$ aus* erreichbar*, falls ein $n\in\mathbb{N}_0$ existiert mit $p_{i,j}^{(n)}>0$. Wir schreiben dann $i\rightsquigarrow j$. Falls $i\rightsquigarrow j$ und $j\rightsquigarrow i$, so schreiben wir $i\leftrightsquigarrow j$.*

Bemerkung 9.3.2 ($\leftrightsquigarrow$ als Äquivalenzrelation)**.** Die Relation $\rightsquigarrow$ auf I ist reflexiv und transitiv, denn wegen $p_{i,i}^{(0)} = 1$ gilt $i \rightsquigarrow i$ für jedes $i \in I$, und falls $i \rightsquigarrow j$ und $j \rightsquigarrow k$ gelten, so folgt aus (9.1.7) leicht, dass auch $i \rightsquigarrow k$ gilt. Also ist $\leftrightsquigarrow$ eine Äquivalenzrelation auf I und teilt I folglich in Klassen ein. Für zwei Klassen $A, B \subset I$ schreiben wir $A \rightsquigarrow B$, falls es $i \in A$ und $j \in B$ gibt mit $i \rightsquigarrow j$, und wir sagen dann, dass B von A aus erreichbar ist. (Offensichtlich ist diese Sprechweise wohldefiniert, d. h. die Definition unabhängig von den gewählten Repräsentanten.) $\diamond$

Definition 9.3.3 (abgeschlossen, irreduzibel)**.** *(a) Eine Teilmenge J von I heißt* abgeschlossen*, wenn $J \not\rightsquigarrow I \setminus J$ gilt, d. h. wenn keine zwei Elemente $j \in J$ und $i \in I \setminus J$ existieren mit $j \rightsquigarrow i$.*

(b) P heißt irreduzibel*, falls I aus einer einzigen Klasse besteht, d. h. wenn je zwei Elemente aus I äquivalent sind.*

Man sieht leicht, dass die Einschränkung $P_J = (p_{i,j})_{i,j \in J}$ von P auf eine abgeschlossene Klasse J von I ihrerseits eine stochastische Matrix auf J ist. Falls P irreduzibel ist, so existieren keine abgeschlossenen echten Teilmengen von I.

Beispiel 9.3.4. (a) Die symmetrische Irrfahrt auf $\mathbb{Z}^d$ ist irreduzibel.

(b) In Polyas Urnenschema (siehe Beispiel 9.2.6) sind keine zwei Elemente aus $I = \mathbb{N}_0^2$ äquivalent. Für jede $r_0, s_0 \in \mathbb{N}$ ist die Menge $\{(r, s) \in I : r \geq r_0, s \geq s_0\}$ abgeschlossen.

(c) Die Irrfahrt auf $\{0, \dots, N\}$(siehe Beispiel 9.2.5) mit absorbierenden Rändern besitzt die Äquivalenzklassen $\{0\}$, $\{1, \dots, N-1\}$ und $\{N\}$. Die Mengen $\{0\}$ und $\{N\}$ sind abgeschlossen, und es gelten $\{1, \dots, N-1\} \rightsquigarrow \{0\}$ und $\{1, \dots, N-1\} \rightsquigarrow \{N\}$.

$\diamond$

Es sei $(X_n)_{n \in \mathbb{N}_0}$ die zur stochastischen Matrix gehörige Markovkette. Wir führen nun die *Ersteintrittszeit* von $i \in I$ ein:

$$T_i = \inf\{n \in \mathbb{N} : X_n = i\}. \tag{9.3.1}$$

Falls die Kette den Punkt i gar nicht besucht, dann setzen wir $T_i = \infty$. Man beachte, dass die Definition von T_i die gesamte unendlich lange Kette $(X_n)_{n \in \mathbb{N}_0}$ erfordert, deren Konstruktion wir noch nicht durchgeführt haben (siehe Bemerkung 9.1.7). Allerdings hängt das Ereignis $\{T_i = n\}$ nur von der Zeit bis n ab, kann also mit unseren Mitteln korrekt behandelt werden. Wir definieren

$$f_{i,j}^{(n)} = \mathbb{P}_i(T_j = n) = \mathbb{P}_i(X_1 \neq j, X_2 \neq j, \dots, X_{n-1} \neq j, X_n = j), \qquad i, j \in I, n \in \mathbb{N}. \tag{9.3.2}$$

In Worten: $f_{i,j}^{(n)}$ ist die Wahrscheinlichkeit, dass eine in i gestartete Kette zum Zeitpunkt n den Punkt j zum ersten Mal trifft. Die Summe

$$f_{i,j} = \sum_{n=1}^{\infty} f_{i,j}^{(n)} \tag{9.3.3}$$

liegt in $[0, 1]$, da die Ereignisse $\{T_j = 1\}, \{T_j = 2\}, \dots$ disjunkt sind. Man kann (und sollte) die unendliche Summe $f_{i,j}$ als $\mathbb{P}_i(T_j < \infty)$ interpretieren, d. h. als die Wahrscheinlichkeit, dass eine in i gestartete Kette jemals den Punkt j besucht.

Satz 9.3.5 (Erneuerungsgleichung). *Es gilt*

$$p_{i,i}^{(n)} = \sum_{k=1}^{n} f_{i,i}^{(k)} p_{i,i}^{(n-k)}, \qquad n \in \mathbb{N}, i \in I. \tag{9.3.4}$$

Beweis. Gemäß Lemma 9.1.9 gilt $p_{i,i}^{(n)} = \mathbb{P}_i(X_n = i)$. Wir spalten das Ereignis $\{X_n = i\}$ nach dem ersten Zeitpunkt, an dem die Kette i erreicht, auf und erhalten

$$p_{i,i}^{(n)} = \sum_{k=1}^{n} \mathbb{P}_i(T_i = k, X_n = i) = \sum_{k=1}^{n} \mathbb{P}_i(X_n = i \mid X_1 \neq i, \dots, X_{k-1} \neq i, X_k = i) f_{i,i}^{(k)}.$$

Nun wenden wir (9.1.4) und Korollar 9.1.10 an und erhalten

$$p_{i,i}^{(n)} = \sum_{k=1}^{n} \mathbb{P}_i(X_n = i \mid X_k = i) f_{i,i}^{(k)} = \sum_{k=1}^{n} f_{i,i}^{(k)} p_{i,i}^{(n-k)}.$$

□

Nun kommen wir zu weiteren zentralen Begriffen der Theorie der Markovketten.

Definition 9.3.6 (Rekurrenz, Transienz). *Ein Zustand $i \in I$ heißt* rekurrent*, falls $f_{i,i} = 1$, und* transient *sonst.*

Interpretationsgemäß heißt also i rekurrent, falls die in i gestartete Kette mit Sicherheit wieder irgendwann einmal zu i zurück kehrt. Wir können diese Eigenschaft in Termen der Potenzen von P charakterisieren:

Satz 9.3.7. *Ein Zustand $i \in I$ ist genau dann rekurrent, wenn $\sum_{n\in\mathbb{N}_0} p_{i,i}^{(n)} = \infty$ gilt.*

Beweis. Für jedes $s \in (0,1)$ erhalten wir aus der Erneuerungsgleichung (siehe Satz 9.3.5)

$$\begin{aligned}\sum_{n\in\mathbb{N}_0} p_{i,i}^{(n)} s^n &= 1 + \sum_{n\in\mathbb{N}} s^n p_{i,i}^{(n)} = 1 + \sum_{n\in\mathbb{N}} s^n \sum_{k=1}^{n} f_{i,i}^{(k)} p_{i,i}^{(n-k)} = 1 + \sum_{k\in\mathbb{N}} f_{i,i}^{(k)} s^k \sum_{n=k}^{\infty} p_{i,i}^{(n-k)} s^{n-k}\\ &= 1 + \sum_{k\in\mathbb{N}} f_{i,i}^{(k)} s^k \sum_{n\in\mathbb{N}_0} p_{i,i}^{(n)} s^n.\end{aligned}$$

Also haben wir für die beiden Funktionen

$$\pi(s) = \sum_{n\in\mathbb{N}_0} p_{i,i}^{(n)} s^n \qquad \text{und} \qquad \phi(s) = \sum_{k\in\mathbb{N}} f_{i,i}^{(k)} s^k$$

die Beziehung $\pi(s) = 1 + \pi(s)\phi(s)$ hergeleitet. Im Fall $1 = f_{i,i} = \phi(1)$ machen wir den Grenzübergang $s \uparrow 1$ und erhalten $\pi(1) = 1 + \pi(1)$, was natürlich impliziert, dass $\sum_{n\in\mathbb{N}_0} p_{i,i}^{(n)} = \pi(1) = \infty$. Falls $f_{i,i} < 1$, so formen wir zunächst um zu $\pi(s) = \frac{1}{1-\phi(s)}$ und lassen dann $s \uparrow 1$, was auf $\sum_{n\in\mathbb{N}_0} p_{i,i}^{(n)} = \frac{1}{1-f_{i,i}} < \infty$ führt und den Beweis beendet. □

Die in Satz 9.3.7 auftretende Reihe kann interpretiert werden:

$$\sum_{n\in\mathbb{N}_0} p_{i,i}^{(n)} = \sum_{n\in\mathbb{N}_0} \mathbb{E}_i(\mathbb{1}_{\{X_n=i\}}) = \mathbb{E}_i\Big(\sum_{n\in\mathbb{N}_0} \mathbb{1}_{\{X_n=i\}}\Big),$$

also ist sie gleich der erwarteten Anzahl der Besuche in i für die unendlich lange in i gestartete Markovkette. Der Zustand i ist also nach Satz 9.3.7 genau dann transient, wenn diese Kette ihren Startpunkt erwartungsgemäß nur endlich oft besucht.

Rekurrenz und Transienz sind Klasseneigenschaften:

Satz 9.3.8. *Es seien $i, j \in I$ mit $i \leftrightsquigarrow j$. Dann ist i genau dann rekurrent, wenn j es ist.*

Beweis. Übungsaufgabe. $\square$

Also können und werden wir in Zukunft auch von rekurrenten bzw. transienten Klassen sprechen und bei Irreduzibilität von rekurrenten bzw. transienten Markovketten. Das Kriterium aus Satz 9.3.7 kann man benutzen, um beispielsweise die Rekurrenz der d-dimensionalen Irrfahrt aus Beispiel 9.2.4 zu entscheiden:

Satz 9.3.9. *Die d-dimensionale symmetrische Irrfahrt auf $\mathbb{Z}^d$ ist rekurrent für $d \in \{1, 2\}$ und transient für $d \geq 3$.*

Beweisskizze. Wegen der Verschiebungsinvarianz genügt es, die Divergenz bzw. Konvergenz der Reihe $\sum_{n \in \mathbb{N}} p_{0,0}^{(n)}$ zu untersuchen. Wir sind in der glücklichen Lage, den Wert von $p_{0,0}^{(n)}$ ausrechnen zu können. Natürlich ist immer $p_{0,0}^{(2n-1)} = 0$ für alle $n \in \mathbb{N}$.

In $d = 1$ ist leicht zu sehen, dass $p_{0,0}^{(2n)} = \binom{2n}{n} 2^{-2n}$ für alle $n \in \mathbb{N}$, und Stirlings Formel zeigt, dass $p_{0,0}^{(2n)} \sim (\pi n)^{-1/2}$ für $n \to \infty$. Also ist $\sum_{n \in \mathbb{N}} p_{0,0}^{(n)} = \infty$ für $d = 1$.

In $d = 2$ schreiben wir $X_n = (X_n^{(1)}, X_n^{(2)})$ für die beiden Komponenten und nutzen die Beobachtung aus, dass die beiden Folgen $(X_n^{(1)} - X_n^{(2)})_{n \in \mathbb{N}_0}$ und $(X_n^{(1)} + X_n^{(2)})_{n \in \mathbb{N}_0}$ zwei unabhängige eindimensionale symmetrische Irrfahrten auf $\mathbb{Z}$ sind. Folglich ist $p_{0,0}^{(2n)} = \left(\binom{2n}{n} 2^{-2n}\right)^2$, und wie oben sieht man, dass dies sich wie $\frac{1}{\pi n}$ verhält, also ebenfalls nicht summierbar ist.

Den Fall $d \geq 3$ kann man auf den Grenzfall $d = 3$ zurück führen. Im Fall $d = 3$ stellt man den Wert von $p_{0,0}^{(2n)}$ mit Hilfe einer Doppelsumme dar (nach einer Aufspaltung, wieviele Schritte jeweils in die drei Dimensionsrichtungen geschehen) und schätzt diese mit einiger Arbeit geeignet ab, bis man sieht, dass $p_{0,0}^{(2n)}$ von der Ordnung $n^{-3/2}$ ist. $\square$

Im Folgenden zeigen wir insbesondere, dass die Markovkette, wenn sie in einer rekurrenten Klasse gestartet wird, *jeden* Zustand in dieser Klasse mit Sicherheit besucht, siehe Lemma 9.3.13. Zunächst ziehen wir eine Folgerung aus der Erneuerungsgleichung.

Lemma 9.3.10. *Für alle $i, j \in I$ gilt*

$$\sum_{n \in \mathbb{N}_0} p_{i,j}^{(n)} = f_{i,j} \sum_{n \in \mathbb{N}_0} p_{j,j}^{(n)}.$$

Beweis. Genau wie im Beweis der Erneuerungsgleichung (Satz 9.3.5) leitet man die Beziehung

$$p_{i,j}^{(n)} = \sum_{k=1}^{n} f_{i,j}^{(k)} p_{j,j}^{(n-k)}$$

für alle $n \in \mathbb{N}$ her. Nun summiert man diese Beziehung über $n \in \mathbb{N}$, vertauscht auf der rechten Seite die beiden Summationen und verschiebt die eine davon (ähnlich wie im Beweis von Satz 9.3.7). $\square$

Korollar 9.3.11. *Für alle $i, j \in I$ gilt:*

$$j \text{ transient} \quad \Longrightarrow \quad \sum_{n\in\mathbb{N}_0} p_{i,j}^{(n)} < \infty.$$

Aus dem folgenden Ergebnis folgt insbesondere, dass Klassen, die nicht abgeschlossen sind, zwangsläufig transient sind. (Abgeschlossene Klassen können hingegen sowohl rekurrent als auch transient sein.)

Lemma 9.3.12. *Es seien $i, j \in I$ mit $i \rightsquigarrow j$. Falls i rekurrent ist, so gilt auch $j \rightsquigarrow i$, und j ist dann ebenfalls rekurrent.*

Beweis. Wir dürfen $i \neq j$ annehmen. Wir führen einen Widerspruchsbeweis und nehmen an, dass i nicht von j aus erreichbar ist, d. h. dass $p_{j,i}^{(n)} = 0$ für alle $n \in \mathbb{N}_0$ ist.

Sei $N \in \mathbb{N}$ die kleinste Zahl n mit $p_{i,j}^{(n)} > 0$. Für alle $n \in \mathbb{N}$ gilt dann $\mathbb{P}_i(X_N = j, X_n = i) = 0$, denn für $n > N$ gilt ja $\mathbb{P}_i(X_N = j, X_n = i) = p_{i,j}^{(N)} p_{j,i}^{(n-N)} = 0$, und für $n < N$ gilt $\mathbb{P}_i(X_N = j, X_n = i) = p_{i,i}^{(n)} p_{i,j}^{(N-n)} = 0$, denn N ist ja das *kleinste* n mit $p_{i,j}^{(n)} > 0$. Daher haben wir

$$\mathbb{P}_i(T_i \leq M, X_N = j) = \sum_{n=1}^{M} \mathbb{P}_i(T_i = n, X_N = j) \leq \sum_{n=1}^{M} \mathbb{P}_i(X_n = i, X_N = j) = 0.$$

Also folgt

$$\sum_{n=1}^{M} f_{i,i}^{(n)} = \mathbb{P}_i(T_i \leq M) = \mathbb{P}_i(T_i \leq M, X_N \neq j) \leq \mathbb{P}_i(X_N \neq j) = 1 - \mathbb{P}_i(X_N = j) = 1 - p_{i,j}^{(N)}.$$

Nun lassen wir $M \uparrow \infty$ und beachten, dass die rechte Seite der Abschätzung nicht von M abhängt. Mit der Rekurrenz von i folgt der Widerspruch $1 = f_{i,i} = \sum_{n\in\mathbb{N}} f_{i,i}^{(n)} \leq 1 - p_{i,j}^{(N)} < 1$. $\square$

Insbesondere sind rekurrente Klassen abgeschlossen. (Transiente Klassen müssen nicht abgeschlossen sein.) Insbesondere ist die Einschränkung von P auf eine rekurrente Klasse (siehe die Bemerkung nach Definition 9.3.3) wieder eine stochastische Matrix, die dann natürlich auch irreduzibel ist. Daher lassen sich also die einzelnen rekurrenten Klassen getrennt diskutieren.

Lemma 9.3.13. *Wenn i und j in der selben rekurrenten Klasse liegen, so gilt $f_{i,j} = f_{j,i} = 1$.*

Beweis. Wir dürfen $i \neq j$ annehmen. Sei N das kleinste $n \in \mathbb{N}$ mit $p_{j,i}^{(n)} > 0$. Für $M > N$ gilt

$$\begin{aligned}
\mathbb{P}_j(T_j \leq M, X_N = i) &= \sum_{n=1}^{M} \mathbb{P}_j(T_j = n, X_N = i) \\
&= \sum_{n=1}^{N-1} f_{j,j}^{(n)} p_{j,i}^{(N-n)} + \sum_{n=N+1}^{M} \mathbb{P}_j(T_j \geq N, X_N = i) f_{i,j}^{(n-N)}.
\end{aligned}$$

Nach Definition ist aber $p_{j,i}^{(N-n)} = 0$ für jedes $n \in \{1, \dots, N-1\}$, also ist die erste Summe auf der rechten Seite gleich Null. In der zweiten schätzen wir ab: $\mathbb{P}_j(T_j \geq N, X_N = i) \leq p_{j,i}^{(N)}$, also erhalten wir aus der obigen Rechnung

$$\mathbb{P}_j(T_j \leq M, X_N = i) \leq \sum_{n=N+1}^{M} p_{j,i}^{(N)} f_{i,j}^{(n-N)} \leq p_{j,i}^{(N)} f_{i,j}.$$

Nun lassen wir $M \uparrow \infty$ und beachten, dass $\lim_{M\to\infty} \mathbb{P}_j(T_j \le M) = \lim_{M\to\infty} \sum_{k=1}^M f_{j,j}^{(k)} = f_{j,j} = 1$, also folgt

$$p_{j,i}^{(N)} = \mathbb{P}_j(X_N = i) = \lim_{M\to\infty} \mathbb{P}_j(T_j \le M, X_N = i) \le p_{j,i}^{(N)} f_{i,j}.$$

Wegen $f_{i,j} \le 1$ und $p_{j,i}^{(N)} > 0$ ergibt sich $f_{i,j} = 1$, und dies beendet den Beweis. □

Ein sehr handliches und simples hinreichendes Kriterium für Rekurrenz ist das Folgende.

Satz 9.3.14. *Wenn I endlich ist, so ist jede irreduzible Kette rekurrent.*

Beweis. Wegen $\sum_{j\in I} p_{i,j}^{(n)} = 1$ für jedes $n \in \mathbb{N}$ und jedes $i \in I$ folgt, dass zu jedem i ein j existiert mit $\sum_{n\in\mathbb{N}} p_{i,j}^{(n)} = \infty$, denn es gilt ja $\sum_{j\in I}(\sum_{n\in\mathbb{N}} p_{i,j}^{(n)}) = \infty$, und irgendeiner der endlich vielen Summanden $\sum_{n\in\mathbb{N}} p_{i,j}^{(n)}$ muss gleich ∞ sein. Aus Lemma 9.3.10 folgt $\sum_{n\in\mathbb{N}} p_{j,j}^{(n)} = \infty$, und es folgt die Rekurrenz von j. □

9.4 Stoppzeiten und die starke Markov-Eigenschaft

Stoppzeiten sind zufällige Zeiten, die nicht in die Zukunft blicken können. Im Zusammenhang mit Markovketten sind Stoppzeiten die Eintrittszeitpunkte gewisser Ereignisse, deren Eintreten durch die bisherige 'Geschichte' der Kette beurteilt werden kann.

Im Folgenden sei $(X_n)_{n\in\mathbb{N}_0}$ eine Markovkette auf I mit Übergangsmatrix P. In diesem Abschnitt empfiehlt es sich anzunehmen, dass alle Zufallsgrößen X_n mit $n \in \mathbb{N}$ auf einem gemeinsamen Wahrscheinlichkeitsraum Ω definiert sind, und das werden wir tun.

Definition 9.4.1 (Filtrierung, Stoppzeit). *(a) Für jedes $n \in \mathbb{N}_0$ sei $\mathcal{F}_n$ die Menge der Ereignisse von der Form $\{\omega \in \Omega : X_{[0,n]}(\omega) \in B\}$ mit $B \subset I^{n+1}$, d. h. die Menge der Ereignisse, die mit Hilfe der Zufallsgrößen $X_0, \dots, X_n$ beschrieben werden können. Die Familie $\mathcal{F} = (\mathcal{F}_n)_{n\in\mathbb{N}_0}$ nennt man die zu $(X_n)_{n\in\mathbb{N}_0}$ gehörige* Filtrierung.

(b) Eine Abbildung $T : \Omega \to \mathbb{N}_0 \cup \{\infty\}$ heißt eine Stoppzeit, *wenn für jedes $n \in \mathbb{N}_0$ das Ereignis $\{T = n\}$ in $\mathcal{F}_n$ liegt.*

Eine Filtrierung ist immer aufsteigend, denn es gilt ja $\mathcal{F}_n \subset \mathcal{F}_{n+1}$ für jedes n. Das Ereignis $\{T = \infty\}$ kann man interpretieren als das Ereignis, dass die Stoppzeit T nie eintritt. Ein Beispiel von Stoppzeiten sind die Ersteintrittszeiten $T_i = \inf\{n \in \mathbb{N} : X_n = i\}$ in einen Punkt i für eine Markovkette $(X_n)_{n\in\mathbb{N}_0}$. Etwas allgemeiner definiert man die Ersteintrittszeit in eine nichtleere Menge $A \subset I$ als

$$T_A = \inf\{n \in \mathbb{N} : X_n \in A\}.$$

Falls man den Zeitpunkt 0 einbeziehen möchte, benutzt man die Stoppzeit

$$S_A = \inf\{n \in \mathbb{N}_0 : X_n \in A\}.$$

Falls $i \in A$, so gilt $\mathbb{P}_i(S_A = 0) = 1$. Ferner gilt für $i \notin A$ die Beziehung $\mathbb{P}_i(T_A = S_A) = 1$.

Wir beginnen mit einer technischen Vorbereitung.

Lemma 9.4.2. *Für alle $i \in I$, $n \in \mathbb{N}_0$ und alle nichtleeren $A \subset I$ gilt*

$$\mathbb{P}_i(T_A \le n+1) = \sum_{j\in I} p_{i,j}\mathbb{P}_j(S_A \le n).$$

Beweis. Übungsaufgabe. □

Sei $A \subset I$ nicht leer. Die Funktion $h_A\colon I \to [0,1]$, definiert durch

$$h_A(i) = \mathbb{P}_i(S_A < \infty),$$

spielt in der Theorie der Markovketten eine wichtige Rolle. Insbesondere ist sie die minimale Lösung eines gewissen Systems von linearen Gleichungen:

Satz 9.4.3. *Sei $A \subset I$ nicht leer. Dann ist h_A die kleinste Funktion $I \to [0,\infty)$, die das Gleichungssystem*

$$h_A(i) = \sum_{j\in I} p_{i,j} h_A(j), \qquad i \in I \setminus A, \tag{9.4.1}$$

löst mit Randbedingung $h_A(i) = 1$ für alle $i \in A$.

Beweis. Dass h_A die Randbedingung $h_A(i) = 1$ für alle $i \in A$ erfüllt, ist klar. Sei nun $i \in I \setminus A$, und wir wollen die Gleichung $h_A(i) = \sum_{j\in I} p_{i,j}h_A(j)$ beweisen. Zunächst benutzen wir die Markoveigenschaft zum Zeitpunkt 1, um zu sehen, dass für alle $N \in \mathbb{N}$ gilt:

$$\begin{aligned}
\mathbb{P}_i(1 \le S_A \le N) &= \sum_{j\in I} \mathbb{P}_i(X_1 = j, 1 \le S_A \le N)\\
&= \sum_{j\in I} \mathbb{P}_i(X_1 = j)\mathbb{P}_i(1 \le S_A \le N \mid X_1 = j)\\
&= \sum_{j\in I} p_{i,j}\mathbb{P}_j(0 \le S_A \le N-1).
\end{aligned}$$

Nun lassen wir auf beiden Seiten $N \to \infty$ gehen und beachten, dass $h_A(i) = \mathbb{P}_i(S_A < \infty) = \lim_{N\to\infty} \mathbb{P}_i(1 \le S_A \le N)$. Da die Summanden der rechten Seite, $\mathbb{P}_j(0 \le S_A \le N-1)$, monoton gegen $h_A(j)$ konvergieren, dürfen wir die Summation mit dem Grenzwert vertauschen, und es folgt die behauptete Gleichung.

Als Letztes zeigen wir die Minimalität von h_A. Sei also $g\colon I \to [0,\infty)$ eine Lösung von $g(i) = \sum_{j\in I} p_{i,j}g(j)$ für $i \in I \setminus A$ mit Randbedingung $g(i) = 1$ für $i \in A$. Wir zeigen mit Induktion nach $n \in \mathbb{N}_0$, dass

$$g(i) \ge \mathbb{P}_i(S_A \le n) \qquad \text{für alle } i \in I. \tag{9.4.2}$$

(Daraus folgt durch Grenzübergang $n \to \infty$, dass $g(i) \ge \lim_{n\to\infty} \mathbb{P}_i(S_A \le n) = \mathbb{P}_i(S_A < \infty) = h_A(i)$, was zu zeigen war.)

Für $n = 0$ ist (9.4.2) offensichtlich, ebenso für alle $i \in A$. Sei (9.4.2) für ein n vorausgesetzt, und sei $i \in I \setminus A$. Dann ist

$$g(i) = \sum_{j\in I} p_{i,j}g(j) \ge \sum_{j\in I} p_{i,j}\mathbb{P}_j(S_A \le n) = \mathbb{P}_i(T_A \le n+1)$$

nach Lemma 9.4.2. Da $i \notin A$ ist, gilt aber $\mathbb{P}_i(T_A \leq n+1) = \mathbb{P}_i(S_A \leq n+1)$. Dies zeigt (9.4.2) für $n+1$ und beendet den Beweis. □

Das Gleichungssystem in (9.4.1) kann man auch als $\mathbb{E}_i[h_A(X_1)] = h_A(i)$ für jedes $i \in I \setminus A$ formulieren. Man sagt, dass h_A in $I \setminus A$ *harmonisch* sei mit Randbedingung Eins (in A). Im trivialen Fall $A = \emptyset$ heißt das, dass $h_\emptyset$ ein Rechtseigenvektor der Matrix P zum Eigenwert Eins ist (wenn er nicht Null wäre wie $h_\emptyset$). Ein solcher Eigenvektor ist natürlich immer der Vektor, der aus Einsen besteht, denn P ist ja stochastisch.

Wir kommen nun zur *starken Markov-Eigenschaft*. Diese besagt im Wesentlichen, dass die gewöhnliche Markov-Eigenschaft, die in Satz 9.1.5 formuliert wurde, auch an Stoppzeiten gilt, d. h. dass die ab einer Stoppzeit T betrachtete Restkette $(X_T, X_{T+1}, \dots)$, gegeben das Ereignis $\{X_T = i\}$ für ein $i \in I$, wieder eine Markovkette mit der selben Übergangsmatrix und Start in i ist und dass sie ferner unabhängig von der Kette vor dem Zeitpunkt T ist, also unabhängig von $(X_0, X_1, \dots, X_T)$. Die Menge aller Ereignisse, die der Kette $(X_0, X_1, \dots, X_T)$ zustoßen können, ist ein wichtiges System von Ereignissen:

Definition 9.4.4 (Prä-$\boldsymbol{T}$-Ereignis)**.** *Es sei $T\colon \Omega \to \mathbb{N}_0 \cup \{\infty\}$ eine Stoppzeit. Wir nennen ein Ereignis $A \subset \Omega$ ein* Prä-T-Ereignis, *falls das Ereignis $A \cap \{T = n\}$ für jedes $n \in \mathbb{N}_0$ in $\mathcal{F}_n$ liegt. Die Menge aller Prä-T-Ereignisse bezeichnen wir mit $\mathcal{F}_T$.*

Wir definieren die I-wertige Zufallsgröße X_T als $X_{T(\omega)}(\omega)$, allerdings nur auf der Menge $\{\omega \in \Omega\colon T(\omega) < \infty\}$. Falls $T(\omega) = \infty$, so ist also X_T nicht definiert. Daraus folgt, dass das Ereignis $\{X_T = i\}$ insbesondere impliziert, dass $T < \infty$.

Nun folgt eine exakte Formulierung der starken Markov-Eigenschaft.

Satz 9.4.5 (starke Markov-Eigenschaft)**.** *Es sei $(X_n)_{n\in\mathbb{N}_0}$ eine Markovkette, T eine Stoppzeit, $A \in \mathcal{F}_T$ und $B \subset I^m$ für ein $m \in \mathbb{N}$. Falls $\mathbb{P}(X_T = i, A) > 0$, so gilt*

$$\mathbb{P}(X_{[T+1,T+m]} \in B \mid X_T = i, A) = \mathbb{P}_i(X_{[1,m]} \in B). \tag{9.4.3}$$

Beweis. Wir multiplizieren mit $\mathbb{P}(X_T = i, A)$ und brauchen nur die Gleichung

$$\mathbb{P}(X_{[T+1,T+m]} \in B, X_T = i, A) = \mathbb{P}_i(X_{[1,m]} \in B)\mathbb{P}(X_T = i, A)$$

zu zeigen. Dies tun wir durch Aufspaltung nach allen (endlichen) Werten, die T annehmen kann: Die linke Seite ist gleich

$$\sum_{n\in\mathbb{N}_0} \mathbb{P}(X_{[n+1,n+m]} \in B, X_n = i, A, T = n).$$

Da $A \in \mathcal{F}_T$, liegt das Ereignis $A \cap \{T = n\}$ in $\mathcal{F}_n$. Also können wir die gewöhnliche Markov-Eigenschaft anwenden (siehe etwa Korollar 9.1.6) und erhalten

$$\mathbb{P}(X_{[n+1,n+m]} \in B, X_n = i, A, T = n) = \mathbb{P}(X_n = i, A, T = n)\mathbb{P}_i(X_{[1,m]} \in B).$$

Aufsummation über $n \in \mathbb{N}_0$ beendet den Beweis. □

Damit können wir die Intuition hinter Satz 9.3.7 (siehe die Bemerkung danach) verschärfen: Die Anzahl der Besuche zu einem gegebenen *rekurrenten* Punkt hat nicht nur unendlichen Erwartungswert, sondern ist sogar konstant gleich ∞:

Korollar 9.4.6. *Es sei $i \in I$ mit $p_i = \mathbb{P}_i(T_i = \infty) < 1$. Es sei $V_i = \sum_{n\in\mathbb{N}_0} \mathbb{1}_{\{X_n=i\}}$ die Anzahl der Besuche in i. Dann gilt $\mathbb{P}_i(V_i > k) = (1-p_i)^k$ für jedes $k \in \mathbb{N}$, d. h., V_i ist unter $\mathbb{P}_i$ geometrisch verteilt, falls i transient ist, und konstant gleich ∞, falls i rekurrent ist.*

Beweis. Wir definieren die sukzessiven Zeitpunkte der Besuche in i durch $T_i^{(0)} = 0$ und $T_i^{(k+1)} = \inf\{n > T_i^{(k)} : X_n = i\}$ für $k \in \mathbb{N}_0$. Das Ereignis $\{V_i > k\}$ ist also gleich dem Ereignis $\{T_i^{(k)} < \infty\}$, falls die Kette in i startet. Ferner gilt offensichtlich $\mathbb{P}_i(X_{T_i^{(k)}} = i \mid T_i^{(k)} < \infty) = 1$. Also gilt für alle $k, l \in \mathbb{N}$:

$$\begin{aligned}
\mathbb{P}_i(V_i > k + l \mid V_i > l) &= \mathbb{P}_i(T_i^{(k+l)} < \infty \mid T_i^{(l)} < \infty)\\
&= \lim_{n\to\infty} \mathbb{P}_i(T_i^{(k+l)} \leq T_i^{(l)} + n \mid X_{T_i^{(l)}} = i, T_i^{(l)} < \infty)\\
&= \lim_{n\to\infty} \mathbb{P}_i(T_i^{(k)} \leq n) = \mathbb{P}_i(T_i^{(k)} < \infty)\\
&= \mathbb{P}_i(V_i > k).
\end{aligned}$$

Im dritten Schritt benutzten wir die starke Markov-Eigenschaft (9.4.3) mit $A = \{T_i^{(l)} < \infty\}$ und $B = \{T_i^{(k)} \leq n\}$.

Also ist V_i auf $\mathbb{N}_0$ geometrisch verteilt, und zwar mit Parameter $\mathbb{P}_i(V_i = 1) = 1 - \mathbb{P}_i(V_i > 1) = 1 - \mathbb{P}_i(T_i < \infty) = \mathbb{P}_i(T_i = \infty) = p_i$. Wie man leicht sieht, gilt dies auch im Fall $p_i = 0$, wobei dann $V_i = \infty$ mit $\mathbb{P}_i$-Wahrscheinlichkeit Eins. $\square$

Beispiel 9.4.7. Es sei $(X_n)_{n\in\mathbb{N}_0}$ die Irrfahrt auf $\mathbb{Z}$ wie in Beispiel 9.2.3. Wir wollen berechnen, wie lange die Irrfahrt im Durchschnitt benötigt, um einen gegebenen Punkt zu erreichen. Genauer: Wir wollen $\mathbb{P}_0(T_1 < \infty)$ und $\mathbb{E}_0(T_1)$ berechnen.

Zunächst sieht man mit Hilfe der starken Markov-Eigenschaft und der räumlichen Verschiebungsinvarianz der Irrfahrt ein, dass $\mathbb{P}_0(T_2 < \infty) = \mathbb{P}_0(T_1 < \infty)\mathbb{P}_1(T_2 < \infty) = \mathbb{P}_0(T_1 < \infty)^2$ gilt (der formale Beweis ist eine Übungsaufgabe). Ferner benutzt man die (gewöhnliche) Markov-Eigenschaft zum Zeitpunkt Eins, um die Gleichung $\mathbb{P}_0(T_1 < \infty) = p + (1-p)\mathbb{P}_{-1}(T_1 < \infty)$ aufzustellen. Wieder mit Hilfe der Verschiebungsinvarianz und der ersten Beobachtung erhält man die quadratische Gleichung $\mathbb{P}_0(T_1 < \infty) = p + (1-p)\mathbb{P}_0(T_1 < \infty)^2$. Diese hat die beiden Lösungen

$$\mathbb{P}_0(T_1 < \infty) = \frac{1 \pm \sqrt{1 - 4p(1-p)}}{2(1-p)} = \frac{1 \pm |2p-1|}{2(1-p)} = \begin{cases} 1 \\ \text{oder} \\ \frac{p}{1-p}. \end{cases}$$

Im Fall $p \geq \frac{1}{2}$ ergibt sich sofort, dass $\mathbb{P}_0(T_1 < \infty) = 1$, denn $\frac{p}{1-p} > 1$ für $p > \frac{1}{2}$. Im Fall $p < \frac{1}{2}$ gilt $\mathbb{P}_0(T_1 < \infty) = \frac{p}{1-p} < 1$, aber unsere Mittel reichen derzeit noch nicht aus, dies zu beweisen.[2] Für $p \geq \frac{1}{2}$ erreicht die Irrfahrt also den Punkt 1 von 0 aus mit Sicherheit, und man zeigt mit der starken Markov-Eigenschaft leicht, dass dann auch jedes andere $i \in \mathbb{N}$ mit Sicherheit erreicht wird.

Mit der (gewöhnlichen) Markov-Eigenschaft leitet man im Fall $p \geq \frac{1}{2}$ auch leicht die Beziehung $\mathbb{E}_0(T_1) = p + (1-p)[1 + \mathbb{E}_{-1}(T_1)]$ her, und die starke impliziert, dass $\mathbb{E}_{-1}(T_1) = \mathbb{E}_0(T_2) = 2\mathbb{E}_0(T_1)$. Also gilt $\mathbb{E}_0(T_1) = \frac{1}{1-2(1-p)}$, was im symmetrischen Fall gleich ∞ ist. $\diamond$

[2]Dies folgt etwa aus dem Starken Gesetz der Großen Zahlen, zusammen mit dem Hinweis auf $\mathbb{E}_0(S_1) < 0$, d. h. es gilt insbesondere $\lim_{n\to\infty} S_n = -\infty$ mit $\mathbb{P}_0$-Wahrscheinlichkeit Eins.

9.5 Gleichgewichtsverteilungen

Im Allgemeinen hängt die Verteilung einer Markovkette zu einem gegebenen Zeitpunkt auf sehr komplizierte Weise ab von der Startverteilung (und natürlich von dem Zeitpunkt). Oft gibt es aber für eine gegebene Übergangsmatrix spezielle Verteilungen, so dass die mit dieser Verteilung gestartete Kette zu jedem Zeitpunkt dieselbe Verteilung besitzt, nämlich diese Startverteilung. Man sagt, die Kette sei dann *im Gleichgewicht* bzw. sie sei *stationär*.

Wie bisher sei I eine höchstens abzählbare Menge und P eine stochastische Matrix auf I. Wir wollen jede Abbildung $\nu\colon I \to [0,\infty)$ ein *Maß* nennen und erinnern uns, dass ein Maß ν eine *Verteilung* heißt, falls $\sum_{i\in I} \nu(i) = 1$ gilt. Außerdem erinnern wir uns, dass νP das Matrixprodukt $(\nu P)(j) = \sum_{i\in I} \nu(i)p_{i,j}$ bezeichnet.

Definition 9.5.1 (invariantes Maß, Gleichgewichtsverteilung)**.** *Ein Maß ν heisst* invariant *für P, falls $\nu P = \nu$ gilt, d. h. falls ν ein (nichtnegativer) Links-Eigenvektor von P zum Eigenwert* 1 *ist. Falls ν eine Verteilung und invariant ist, nennt man ν auch eine* Gleichgewichtsverteilung *für P.*

Bemerkung 9.5.2. (a) Wenn I endlich ist, kann jedes invariante Maß zu einer Gleichgewichtsverteilung normiert werden.

(b) Die Frage, ob P ein invariantes Maß besitzt, ist *a priori* nicht leicht zu beantworten. Man muss im Prinzip das Gleichungssystem $\nu P = \nu$ lösen und prüfen, ob es eine nichtnegative Lösung ν gibt.

(c) Offensichtlich ist ein für P invariantes Maß auch für jede Potenz von P invariant. Falls P irreduzibel ist und $\nu \neq 0$ ein invariantes Maß ist, so ist $\nu(i) > 0$ für jedes $i \in I$. Denn wegen $\nu \neq 0$ existiert ein $i_0 \in I$ mit $\nu(i_0) > 0$, und dann folgt für jedes $i \in I$ (mit einem $n \in \mathbb{N}$, so dass $p_{i_0,i}^{(n)} > 0$): $\nu(i) \geq \nu(i_0)p_{i_0,i}^{(n)} > 0$.

(d) Wenn ν eine Gleichgewichtsverteilung ist, dann gilt für jedes $n \in \mathbb{N}_0$ und jedes $j \in I$:

$$\mathbb{P}_\nu(X_n = j) = \sum_{i\in I} \nu(i)p_{i,j}^{(n)} = \nu(j),$$

d. h. die mit Verteilung ν gestartete Kette hat zu jedem Zeitpunkt die selbe Verteilung ν.

$\diamond$

Sehr oft kann das invariante Maß nicht sinnvoll explizit angegeben werden, und es gibt wenige Beispiele, in denen das invariante Maß eine der bekannten Verteilungen ist. Für die symmetrische Irrfahrt auf $\mathbb{Z}$ (siehe Beispiel 9.2.3) ist die Gleichverteilung auf $\mathbb{Z}$ invariant, aber offensichtlich kann man sie nicht zu einer Gleichgewichtsverteilung normieren. Für die Irrfahrt auf $\{0,\dots,N\}$ (siehe Beispiel 9.2.5) mit absorbierenden Rändern 0 und N sind die beiden Verteilungen invariant, die auf $\{0\}$ bzw. auf $\{N\}$ konzentriert sind. Für weitere Beispiele siehe Lemma 9.5.8 und Beispiele 9.5.9 und 9.7.4.

Für den Rest dieses Abschnittes setzen wir voraus, dass P irreduzibel ist, also dass I aus einer einzigen Äquivalenzklasse besteht. Zunächst zeigen wir, dass im Falle der Rekurrenz immer mindestens ein invariantes Maß existiert, und zwar, für beliebiges $k \in I$, das Maß γ_k, gegeben als

$$\gamma_k(i) = \mathbb{E}_k\Big(\sum_{n=1}^{T_k} \mathbb{1}_{\{X_n=i\}}\Big).$$

In Worten: $\gamma_k(i)$ ist die erwartete Zahl der Besuche in i für eine in k gestartete Markovkette, die betrachtet wird bis zur ersten Rückkehr zum Startpunkt k.

Satz 9.5.3. *Sei P irreduzibel und rekurrent, und sei $k \in I$. Dann gelten:*

(a) γ_k ist ein invariantes Maß.

(b) Für jedes $i \in I$ gilt $0 < \gamma_k(i) < \infty$.

(c) γ_k ist das einzige invariante Maß mit Wert 1 in k.

Beweis. (a) Wir schreiben zunächst $\gamma_k(i)$ wie folgt:

$$\begin{aligned}\gamma_k(i) &= \mathbb{E}_k\Big(\sum_{n\in\mathbb{N}} \mathbb{1}_{\{X_n=i,n\leq T_k\}}\Big) = \sum_{n\in\mathbb{N}} \mathbb{P}_k(X_n = i, n \leq T_k)\\ &= \sum_{n\in\mathbb{N}}\sum_{j\in I} \mathbb{P}_k(X_n = i, X_{n-1} = j, n \leq T_k).\end{aligned}$$

Man beachte, dass das Ereignis $\{n \leq T_k\} = \{T_k \leq n-1\}^{\mathrm{c}}$ in $\mathcal{F}_{n-1}$ liegt. Also können wir die Markov-Eigenschaft zum Zeitpunkt $n-1$ anwenden:

$$\begin{aligned}\mathbb{P}_k(X_n = i, X_{n-1} = j, n \leq T_k) &= \mathbb{P}_k(X_{n-1} = j, n \leq T_k)\mathbb{P}_j(X_1 = i)\\ &= \mathbb{P}_k(X_{n-1} = j, n-1 \leq T_k - 1)p_{j,i}.\end{aligned}$$

Dies setzen wir oben ein und erhalten nach einer Indexverschiebung

$$\begin{aligned}\gamma_k(i) &= \sum_{j\in I} p_{j,i} \sum_{n\in\mathbb{N}_0} \mathbb{P}_k(X_n = j, n \leq T_k - 1) = \sum_{j\in I} p_{j,i}\mathbb{E}_k\Big(\sum_{n=0}^{T_k-1} \mathbb{1}_{\{X_n=j\}}\Big)\\ &= \sum_{j\in I} p_{j,i}\mathbb{E}_k\Big(\sum_{n=1}^{T_k} \mathbb{1}_{\{X_n=j\}}\Big) = \sum_{j\in I} \gamma_k(j)p_{j,i}.\end{aligned}$$

(b) Auf Grund von (a) ist γ_k auch invariant für P^n für jedes $n \in \mathbb{N}$, also folgt insbesondere $1 = \gamma_k(k) \geq \gamma_k(j)p_{j,k}^{(n)}$ für jedes $j \in I$. Wegen der Irreduzibilität gibt es für jedes j ein n mit $p_{j,k}^{(n)} > 0$, und somit folgt $\gamma_k(j) < \infty$ für jedes j. Andererseits folgt auch $\gamma_k(j) \geq \gamma_k(k)p_{k,j}^{(n)} = p_{k,j}^{(n)}$, was positiv ist für geeignetes n.

(c) Es sei λ ein invariantes Maß mit $\lambda(k) = 1$. Dann haben wir $\lambda(j) = \sum_{i\in I\setminus\{k\}} \lambda(i)p_{i,j} + p_{k,j}$ für jedes j. Nun ersetzen wir auf der rechten Seite $\lambda(i)$ ebenfalls mit Hilfe der Invarianz und erhalten

$$\begin{aligned}\lambda(j) &= \sum_{i\in I\setminus\{k\}} \Big(\sum_{i_1\in I\setminus\{k\}} \lambda(i_1)p_{i_1,i} + p_{k,i}\Big)p_{i,j} + p_{k,j}\\ &= \sum_{i,i_1\in I\setminus\{k\}} \lambda(i_1)p_{i_1,i}p_{i,j} + \sum_{i\in I\setminus\{k\}} p_{k,i}p_{i,j} + p_{k,j}\\ &= \sum_{i,i_1\in I\setminus\{k\}} \lambda(i_1)p_{i_1,i}p_{i,j} + \mathbb{P}_k(T_k \geq 2, X_2 = j) + \mathbb{P}_k(T_k \geq 1, X_1 = j).\end{aligned}$$

In dieser Weise fahren wir iterativ fort und erhalten für jedes $n \in \mathbb{N}$

$$\begin{aligned}\lambda(j) &= \sum_{i_0,i_1,\dots,i_n\in I\setminus\{k\}} \lambda(i_n)\Big(\prod_{r=1}^{n} p_{i_r,i_{r-1}}\Big)p_{i_0,j} + \sum_{r=1}^{n+1}\mathbb{P}_k(T_k\geq r, X_r=j)\\ &\geq \sum_{r=1}^{n+1}\mathbb{P}_k(T_k\geq r, X_r=j) = \mathbb{E}_k\Big(\sum_{r=1}^{\min\{T_k,n+1\}} \mathbb{1}_{\{X_r=j\}}\Big).\end{aligned}$$

Nun lassen wir $n\to\infty$ und erhalten auf der rechten Seite $\gamma_k(j)$ als Grenzwert. Also haben wir gezeigt, dass $\lambda(j)\geq\gamma_k(j)$ für alle $j\in I$. Daher ist $\lambda-\gamma_k$ ebenfalls ein invariantes Maß. Da dieses Maß allerdings eine Nullstelle besitzt (nämlich in k), ist es nach Bemerkung 9.5.2(c) identisch gleich Null. Dies beendet den Beweis. □

Zusammen mit Satz 9.3.14 wissen wir nun, dass jede *endliche* irreduzible stochastische Matrix immer eine Gleichgewichtsverteilung besitzt.

Der folgende Satz 9.5.5 klärt die Frage nach der Existenz von Gleichgewichtsverteilungen und ist einer der wichtigsten Sätze über Markovketten. Die Antwort ist eng verknüpft mit der *erwarteten Rückkehrzeit* $\mu_i=\mathbb{E}_i(T_i)$ zu einem Startpunkt $i\in I$. Falls $\mathbb{P}_i(T_i=\infty)>0$, so ist natürlich $\mu_i=\infty$. Ansonsten kann man schreiben

$$\mu_i=\mathbb{E}_i(T_i)=\sum_{n\in\mathbb{N}} n\mathbb{P}_i(T_i=n)=\sum_{n\in\mathbb{N}} nf_{i,i}^{(n)}\in[0,\infty].$$

Falls $\mu_i<\infty$ (d. h. falls die Irrfahrt nach endlicher erwarteter Zeit zu ihrem Startpunkt i zurück kehrt), dann ist natürlich $\mathbb{P}_i(T_i<\infty)=1$, die Irrfahrt also rekurrent. Mit Hilfe von μ_i können wir also rekurrente Zustände feiner unterteilen:

Definition 9.5.4 (positive und Nullrekurrenz)**.** *Ein Zustand $i\in I$ heißt* positiv rekurrent*, falls $\mu_i<\infty$, und* nullrekurrent*, falls i rekurrent, aber nicht positiv rekurrent ist.*

Der folgende Satz zeigt insbesondere, dass die Existenz von Gleichgewichtsverteilungen äquivalent zur positiven Rekurrenz ist:

Satz 9.5.5. *Sei P irreduzibel, dann sind die folgenden drei Aussagen äquivalent.*

(i) Es existiert eine Gleichgewichtsverteilung.

(ii) Es gibt einen positiv rekurrenten Zustand in I.

(iii) Alle Zustände in I sind positiv rekurrent.

Sind diese Bedingungen erfüllt, so ist die Gleichgewichtsverteilung π eindeutig bestimmt und durch $\pi(i)=\frac{1}{\mu_i}$ gegeben.

Beweis. Die Richtung $(iii)\Longrightarrow(ii)$ ist trivial. Wir zeigen nun $(ii)\Longrightarrow(i)$. Sei $k\in I$ mit $\mu_k<\infty$, dann ist wegen Irreduzibilität die Kette rekurrent. Nach Satz 9.5.3 ist γ_k ein invariantes Maß. Man sieht, dass

$$\sum_{j\in I}\gamma_k(j)=\sum_{j\in I}\mathbb{E}_k\Big(\sum_{n=1}^{T_k}\mathbb{1}_{\{X_n=j\}}\Big)=\mathbb{E}_k\Big(\sum_{n=1}^{T_k}\sum_{j\in I}\mathbb{1}_{\{X_n=j\}}\Big)=\mathbb{E}_k(T_k)=\mu_k<\infty. \qquad (9.5.1)$$

Also kann man γ_k zu einer Gleichgewichtsverteilung normieren, d. h. (i) folgt.

Nun beweisen wir $(i) \Longrightarrow (iii)$. Sei π eine Gleichgewichtsverteilung, und sei $k \in I$. Dann ist $\gamma = \pi/\pi(k)$ ein invariantes Maß mit $\gamma(k) = 1$. Nach Satz 9.5.3 gilt $\gamma = \gamma_k$. Wie in (9.5.1) sieht man, dass

$$\mu_k = \sum_{j \in I} \gamma_k(j) = \sum_{j \in I} \gamma(j) = \frac{1}{\pi(k)} \sum_{j \in I} \pi(j) = \frac{1}{\pi(k)} < \infty.$$

Also ist (iii) gezeigt.

Die Zusatzaussage ergibt sich aus den obigen Beweisen. □

Also ergibt sich eine Trichotomie für irreduzible Markovketten:

Bemerkung 9.5.6 (Klassifikation)**.** Jede irreduzible Markovkette gehört immer zu genau einem der folgenden drei Fälle:

- Die Kette ist *transient*, für alle $i, j \in I$ gilt $\mathbb{P}_i(T_j < \infty) < 1$, und es gibt keine invariante Verteilung.
- Die Kette ist *nullrekurrent*, für alle $i, j \in I$ gilt $\mathbb{P}_i(T_j < \infty) = 1$ und $\mathbb{E}_i(T_j) = \infty$, und es gibt zwar ein invariantes Maß, aber keine invariante Verteilung.
- Die Kette ist *positiv rekurrent*, für alle $i, j \in I$ gilt $\mathbb{E}_i(T_j) < \infty$, und es gibt eine invariante Verteilung.

◇

Bemerkung 9.5.7. (i) Auf Grund von Satz 9.3.14 kann bei endlichem Zustandsraum I nur der positiv rekurrente Fall auftreten.

(ii) Die eindimensionale einfache Irrfahrt (siehe Beispiel 9.2.3) ist nullrekurrent. Denn sie ist rekurrent (siehe Satz 9.3.9) und besitzt das konstante Maß als ein invariantes Maß, welches sich nicht normieren lässt.

(iii) Die dreidimensionale einfache Irrfahrt (siehe Beispiel 9.2.4) ist transient und besitzt ein invariantes Maß (das konstante Maß), aber keine invariante Verteilung. ◇

In den meisten Fällen, selbst bei endlichem I, kann man die Gleichgewichtsverteilung nicht sinnvoll explizit angeben, die Formel wäre zu kompliziert. Ein schönes Gegenbeispiel ist die μ-Irrfahrt auf einer Gruppe, siehe Beispiel 9.2.2:

Lemma 9.5.8. *Es seien G eine endliche Gruppe und μ eine Verteilung auf G. Dann ist die Gleichverteilung auf G eine Gleichgewichtsverteilung der μ-Irrfahrt. Falls die Irrfahrt irreduzibel ist, so ist sie auch die einzige Gleichgewichtsverteilung.*

Beweis. Für jedes $g \in G$ ist die Abbildung $h \mapsto g^{-1}h$ bijektiv auf G, und es gilt

$$\sum_{g \in G} p_{g,h} = \sum_{g \in G} \mu(g^{-1}h) = \sum_{g' \in G} \mu(g') = 1.$$

□

Beispiel 9.5.9 (Bernoulli-Laplace-Diffusionsmodell)**.** Noch ein Beispiel, in dem die invariante Verteilung explizit angegeben werden kann, ist das Diffusionsmodell aus Beispiel 9.2.8. Hier ist die hypergeometrische Verteilung $\nu(i) = \binom{s}{i}\binom{w}{w-i}/\binom{s+w}{w}$ invariant auf $\{0, 1, \dots, w\}$. ◇

9.6 Konvergenz gegen die Gleichgewichtsverteilung

Das Hauptergebnis dieses Abschnittes (und eine der wichtigsten Eigenschaften von Markovketten) ist die Tatsache, dass jede irreduzible aperiodische und positiv rekurrente Markovkette bei divergierender Laufzeit gegen ihre Gleichgewichtsverteilung konvergiert. Hierbei nennen wir eine Markovkette bzw. ihre Übergangsmatrix P *aperiodisch*, wenn für jedes $i \in I$ die sogenannte *Periode von i*,

$$d_i = \mathrm{ggT}\{n \in \mathbb{N} \colon p_{i,i}^{(n)} > 0\},$$

gleich Eins ist. Wir wollen uns im Folgenden auf aperiodische Markovketten beschränken, d. h. wir werden voraussetzen, dass $d_i = 1$ für jedes $i \in I$ gilt. Unser Hauptergebnis ist:

Satz 9.6.1. *Es sei P irreduzibel, aperiodisch und positiv rekurrent mit Gleichgewichtsverteilung π. Dann gilt für alle $i, j \in I$:*

$$\lim_{n\to\infty} p_{i,j}^{(n)} = \pi(j).$$

Als Übungsaufgabe zeigt man leicht, dass auch für jede Startverteilung ν und jedes $j \in I$ gilt: $\lim_{n\to\infty} \mathbb{P}_\nu(X_n = j) = \pi(j)$. Falls die Markovkette nicht aperiodisch ist, also wenn sie etwa die Periode $d > 1$ hat[3], dann zeigt man leicht, dass die Matrix P^d auf gewissen Teilmengen aperiodisch ist, und kann Satz 9.6.1 auf geeignete Teilmatrizen von P^d anwenden.

Als Vorbereitung des Beweises von Satz 9.6.1 charakterisieren wir die Aperiodizität:

Lemma 9.6.2. *P ist genau dann irreduzibel und aperiodisch, wenn für jede $i, j \in I$ ein $n_0 \in \mathbb{N}$ existiert, so dass $p_{i,j}^{(n)} > 0$ für alle $n \geq n_0$ gilt.*

Beweis. Dieser Beweis ist länglich und benutzt ausschließlich zahlentheoretische Argumente und wird daher weggelassen. Siehe etwa [Se81]. □

Beweis von Satz 9.6.1. Wie immer sei $(X_n)_{n\in\mathbb{N}_0}$ eine Markovkette mit Übergangsmatrix P und Start in i unter $\mathbb{P}_i$. Wir wollen also zeigen, dass $\mathbb{P}_i(X_n = j)$ für $n \to \infty$ gegen $\pi(j)$ konvergiert.

Wir benutzen ein sogenanntes *Kopplungsargument*, das die Untersuchung der betrachteten Markovkette $(X_n)_{n\in\mathbb{N}_0}$ koppelt mit einer weiteren, unabhängigen Markovkette $(Y_n)_{n\in\mathbb{N}_0}$ mit der selben Übergangsmatrix, wobei allerdings $(Y_n)_{n\in\mathbb{N}_0}$ mit der Gleichgewichtsverteilung π gestartet wird. Daraus folgt, dass die Verteilung von Y_n zu jedem festen Zeitpunkt n exakt gleich π ist, siehe Bemerkung 9.5.2(d). Unsere Hauptaufgabe besteht nun darin zu beweisen, dass die erste Kette irgendwann einmal die zweite trifft, d. h. dass sich beide Ketten jemals im selben Zustand befinden. Mit anderen Worten, wir werden zeigen, dass die *Treffzeit*

$$T = \inf\{n \in \mathbb{N}_0 \colon X_n = Y_n\}$$

mit Sicherheit endlich ist. Dann definieren wir die gekoppelte Kette $(Z_n)_{n\in\mathbb{N}_0}$ durch

$$Z_n = \begin{cases} X_n & \text{für } n \leq T, \\ Y_n & \text{für } n > T, \end{cases}$$

[3]Man kann leicht zeigen, dass die Periode eine Klasseneigenschaft ist, d. h. dass die Abbildung $i \mapsto d_i$ konstant auf den Klassen ist.

und zeigen, dass $(Z_n)_{n\in\mathbb{N}_0}$ ebenfalls eine Markovkette mit Übergangsmatrix P ist. Danach ist der Beweis schnell beendet.

Wir kommen zu den Details. Als Erstes sieht man, dass die Paar-Markovkette $(X_n, Y_n)_{n\in\mathbb{N}_0}$ die Übergangsmatrix $\widehat{P}$ auf $I\times I$ mit $\widehat{p}_{(i,j),(k,l)} = p_{i,k}p_{j,l}$ für alle $i,j,k,l\in I$ besitzt. Für alle $n\in\mathbb{N}$ gilt auch $\widehat{p}^{(n)}_{(i,j),(k,l)} = p^{(n)}_{i,k}p^{(n)}_{j,l}$, wovon man sich leicht überzeugt. Die Aperiodizität von $\widehat{P}$ zeigt man leicht als Übungsaufgabe mit Hilfe von Lemma 9.6.2 und (9.1.7). Da P die Gleichgewichtsverteilung π besitzt, hat $\widehat{P}$ die Gleichgewichtsverteilung $\widehat{\pi}$, die gegeben ist durch $\widehat{\pi}(i,j) = \pi(i)\pi(j)$ für $i,j\in I$. Also ist $\widehat{P}$ insbesondere positiv rekurrent (siehe Satz 9.5.5).

Ab jetzt sei $X_0 = i$ fest. Die Markovkette $(X_n, Y_n)_{n\in\mathbb{N}_0}$ auf $I\times I$ besitzt die Startverteilung $\widehat{\nu}(k,l) = \delta_i(k)\pi(l)$, wobei $\delta_i(k) = \delta_{i,k}$ das Kroneckersymbol ist.

Nun argumentieren wir, dass die Treffzeit T mit Wahrscheinlichkeit Eins endlich ist. Für einen beliebigen Zustand $b\in I$ sei $T_{(b,b)}$ der erste Zeitpunkt, an dem die Kette $(X_n, Y_n)_{n\in\mathbb{N}_0}$ den Punkt (b,b) trifft. Offensichtlich ist $T\le T_{(b,b)}$. Wegen der Rekurrenz von $\widehat{P}$ ist $T_{(b,b)}$ endlich, also auch T.

Nun beweisen wir, dass die oben definierte Folge $(Z_n)_{n\in\mathbb{N}_0}$ eine Markovkette mit Übergangsmatrix P ist. Dazu reicht es nach Lemma 9.1.5(a) zu zeigen, dass für alle $n\in\mathbb{N}$ und alle $i_0,\dots,i_n\in I$ gilt:

$$\mathbb{P}_{\widehat{\nu}}(Z_{[0,n]} = i_{[0,n]}) = \delta_i(i_0)\prod_{r=1}^n p_{i_{r-1},i_r}. \tag{9.6.1}$$

Dies zeigen wir durch Aufspaltung nach allen Werten, die T annehmen kann:

$$\begin{aligned}\mathbb{P}_{\widehat{\nu}}(Z_{[0,n]} = i_{[0,n]}) &= \sum_{r=0}^n \mathbb{P}_{\widehat{\nu}}(Z_{[0,n]} = i_{[0,n]}, T=r) + \mathbb{P}_{\widehat{\nu}}(Z_{[0,n]} = i_{[0,n]}, T>n)\\ &= \sum_{r=0}^n \mathbb{P}_{\widehat{\nu}}(X_{[0,r]} = i_{[0,r]}, Y_{[r+1,n]} = i_{[r+1,n]}, Y_0\neq i_0,\dots,Y_{r-1}\neq i_{r-1}, Y_r = i_r)\\ &\qquad + \mathbb{P}_{\widehat{\nu}}(X_{[0,n]} = i_{[0,n]}, Y_0\neq i_0,\dots,Y_n\neq i_n).\end{aligned}$$

Nun nutzt man die Unabhängigkeit der Ketten $(X_n)_{n\in\mathbb{N}_0}$ und $(Y_n)_{n\in\mathbb{N}_0}$ sowie die Markov-Eigenschaft der Kette $(Y_n)_{n\in\mathbb{N}_0}$ zum Zeitpunkt r bzw. n aus, um zu erhalten:

$$\begin{aligned}&\mathbb{P}_{\widehat{\nu}}(X_{[0,r]} = i_{[0,r]}, Y_{[r+1,n]} = i_{[r+1,n]}, Y_0\neq i_0,\dots,Y_{r-1}\neq i_{r-1}, Y_r = i_r)\\ &\quad = \mathbb{P}_i(X_{[0,r]} = i_{[0,r]})\mathbb{P}_\pi(Y_{[r+1,n]} = i_{[r+1,n]} \mid Y_0\neq i_0,\dots,Y_{r-1}\neq i_{r-1}, Y_r = i_r)\\ &\qquad\qquad \times \mathbb{P}_\pi(Y_0\neq i_0,\dots,Y_{r-1}\neq i_{r-1}, Y_r = i_r)\\ &\quad = \delta_i(i_0)\Big(\prod_{j=1}^n p_{i_{j-1},i_j}\Big)\mathbb{P}_\pi(Y_0\neq i_0,\dots,Y_{r-1}\neq i_{r-1}, Y_r = i_r)\end{aligned}$$

und

$$\mathbb{P}_{\widehat{\nu}}(X_{[0,n]} = i_{[0,n]}, Y_0\neq i_0,\dots,Y_n\neq i_n) = \delta_i(i_0)\Big(\prod_{j=1}^n p_{i_{j-1},i_j}\Big)\mathbb{P}_\pi(Y_0\neq i_0,\dots,Y_n\neq i_n).$$

Nun kombiniere man all diese Gleichungen und beachte, dass

$$\sum_{r=0}^n \mathbb{P}_\pi(Y_0\neq i_0,\dots,Y_{r-1}\neq i_{r-1}, Y_r = i_r) + \mathbb{P}_\pi(Y_0\neq i_0,\dots,Y_n\neq i_n) = 1,$$

um bei (9.6.1) anzukommen. Also ist $(Z_n)_{n\in\mathbb{N}_0}$ eine Markovkette mit Übergangsmatrix P.

Nun ist der Beweis von Satz 9.6.1 leicht zu beenden. Wir schreiben $p_{i,j}^{(n)}$ mit Hilfe von $(Z_n)_{n\in\mathbb{N}_0}$ als

$$p_{i,j}^{(n)} = \mathbb{P}_{\widehat{\nu}}(Z_n = j) = \mathbb{P}_{\widehat{\nu}}(Z_n = j, T \le n) + \mathbb{P}_{\widehat{\nu}}(Z_n = j, T > n)$$

und analog $\pi(j)$ als

$$\pi(j) = \mathbb{P}_\pi(Y_n = j) = \mathbb{P}_{\widehat{\nu}}(Z_n = j, T \le n) + \mathbb{P}_{\widehat{\nu}}(Y_n = j, T > n).$$

Somit folgt

$$\left|p_{i,j}^{(n)} - \pi(j)\right| \le 2\mathbb{P}_{\widehat{\nu}}(T > n),$$

und dies konvergiert für $n \to \infty$ gegen Null, da ja $\mathbb{P}_{\widehat{\nu}}(T < \infty) = 1$. □

9.7 Reversible Markovketten

Einen wichtigen Spezialfall von Gleichgewichtsverteilungen stellen Verteilungen dar, die die sogenannte *detailed balance condition* erfüllen:

Definition 9.7.1 (reversibel)**.** *Ein Maß ν auf I heißt* reversibel *bezüglich der stochastischen Matrix P, falls für alle $i,j \in I$ gilt: $\nu(i)p_{i,j} = \nu(j)p_{j,i}$. In diesem Fall nennen wir auch P reversibel.*

Bemerkung 9.7.2. (a) Es ist offensichtlich, dass jedes reversible Maß invariant ist. Im positiv rekurrenten Fall ist also die Gleichgewichtsverteilung der einzige Kandidat für eine reversible Verteilung. Allerdings sind bei weitem nicht alle Gleichgewichtsverteilungen reversibel. Die Reversibilität ist eine sehr spezielle Eigenschaft.

(b) Der Begriff der Reversibilität kommt von der folgenden Eigenschaft her: Wenn π eine reversible Verteilung ist, so ist die Verteilung der Markovkette unter $\mathbb{P}_\pi$ invariant unter Zeitumkehr, d. h. für alle $n \in \mathbb{N}$ und alle $i_0, i_1, \dots, i_n \in I$ gilt

$$\mathbb{P}_\pi(X_0 = i_0, \dots, X_n = i_n) = \mathbb{P}_\pi(X_0 = i_n, \dots, X_n = i_0).$$

Diese Eigenschaft rechnet man leicht als Übungsaufgabe nach.

◇

Es folgen zwei Beispiele, in denen die reversible Gleichgewichtsverteilung existiert und explizit angegeben werden kann.

Beispiel 9.7.3. Sei G eine endliche Gruppe und μ eine Verteilung auf G mit $\mu(g) = \mu(g^{-1})$ für jedes $g \in G$. Dann ist die Gleichverteilung auf G reversibel für die μ-Irrfahrt auf G (siehe Beispiele 9.2.2 und 9.5.8). Dies sieht man leicht, indem man nachrechnet, dass die Übergangsmatrix P symmetrisch ist. ◇

Beispiel 9.7.4 (Ehrenfests Urnenmodell)**.** Wir betrachten das Ehrenfest-Modell von Beispiel 9.2.7. Als Übungsaufgabe rechnet man nach, dass die Binomialverteilung $\pi(k) = \binom{N}{k}2^{-N}$ reversibel ist.

Eine andere Betrachtungsweise macht es möglich, die reversible Verteilung durch Deduktion zu finden. Wir erweitern das Modell, indem wir den Aufenthaltsort jeder einzelnen Kugel registrieren: '0' für die linke Urne, '1' für die rechte. Dann ist der Vektor der N Aufenthaltsorte der N Kugeln ein Element des neuen Zustandsraumes $\{0,1\}^N$. Wir können diesen Zustandsraum als eine Gruppe auffassen, wobei die Verknüpfung die Addition modulo 2 ist. Die Markovkette der N Labels gehorcht der Regel, dass zu den Zeitpunkten $1, 2, \dots$ ein zufällig gewähltes der N Indizes geflippt wird. Dies ist eine μ-Irrfahrt im Sinne von Beispiel 9.2.2 mit $\mu(x) = \frac{1}{N}$, falls $x = (x_1, \dots, x_N)$ genau eine von Null verschiedene Komponente hat. Nach Beispiel 9.7.3 ist die Gleichverteilung reversibel für diese Irrfahrt. Projiziert man diese Verteilung wieder zurück auf die Anzahl der Kugeln in der linken Urne, erhält man gerade die Binomialverteilung.

Da die Kehrwerte der Verteilung $\pi(k)$ gerade die erwarteten Rückkehrzeiten zu k sind, erhält man die interessante Folgerung, dass das System im Mittel viel später in seinen Anfangszustand zurückkehrt, wenn es in einem extremen Zustand startet (d. h. alle Kugeln in einer Urne), als wenn es mit etwa ausgeglichenen Kugelbeständen beginnt. $\diamond$

Beispiel 9.7.5 (MCMC). In vielen Anwendungen des Konvergenzsatzes 9.6.1 dreht man den Spieß folgendermaßen um: Auf einer endlichen Menge I hat man eine gewisse Verteilung π gegeben und möchte eine Stichprobe mit dieser Verteilung ziehen, d. h. man möchte eine Zufallsvariable mit der Verteilung π generieren. Diese Aufgabe ist zwar endlich, aber in den meisten Fällen höchst nichttrivial, weil I oft gigantisch groß ist. Ein beliebter Ansatz ist die sogenannte *Markov Chain Monte Carlo* Methode: Man wählt eine Markovkette, deren invariante Verteilung π ist, und man lässt diese Kette mit einem beliebigen Startwert 'genügend lange' laufen, bis man auf Grund von Satz 9.6.1 meint, dass die Kette 'nahe genug' an π heran gekommen ist. Dabei wählt man meist eine Übergangsmatrix mit vielen Nullen, so dass π sogar reversibel ist, etwa die Matrix $P = (p_{i,j})_{i,j \in I}$ mit

$$p_{i,j} = \begin{cases} q_{i,j} \min\Big\{1, \frac{\pi(j) q_{j,i}}{\pi(i) q_{i,j}}\Big\}, & \text{falls } i \neq j, \\ 1 - \sum_{k \in I \setminus \{i\}} p_{i,k}, & \text{falls } i = j, \end{cases}$$

wobei $Q = (q_{i,j})_{i,j \in I}$ eine stochastische 'Referenzmatrix' auf I ist (mit vielen Nullen). Dieser Ansatz wird der *Metropolis-Algorithmus* genannt. Man beachte, dass nur die Quotienten $\pi(j)/\pi(i)$ bekannt sein müssen, die bei geeigneter Wahl von Q meist nicht extrem klein oder extrem groß sind und deren Berechnung dann auch nicht aufwendig ist.

Der Vorteil einer solchen MCMC-Methode ist, dass sie einfach zu definieren und zu implementieren ist. Der Nachteil ist, dass man oft keine guten Kriterien hat, wann man die Iterationen abbrechen sollte. In jedem Fall werden Abschätzungen für die Konvergenzgeschwindigkeit benötigt, und es ist oft nicht leicht, effektive Abschätzungen zu erhalten.

Viele Beispiele für zu simulierende Verteilungen kommen aus der statistischen Physik, wo auf Gitterkonfigurationsräumen der Form $I = \{0,1\}^{\Lambda_n}$, wobei $\Lambda_n = \mathbb{Z}^d \cap [-n,n]^d$ eine große Box ist, gewisse Verteilungen der Form $\pi(i) = \frac{1}{Z_{n,\beta}} e^{-\beta H_n(i)}$ gegeben sind, wobei $Z_{n,\beta}$ die Normierung ist, $\beta > 0$ ein Stärkenparameter und $H_n \colon I \to \mathbb{R}$ eine *Hamiltonsche Abbildung*. Dies ist ein Maß, das Konfigurationen i mit kleinem Wert von $H_n(i)$ bevorzugt, und dieser Effekt ist desto stärker, je größer β ist. Ein Beispiel ist die Anzahl der 0-1-Übergänge $H_n(i) = \sum_{x,y \in \Lambda_n \colon |x-y|=1} |i_x - i_y|$, das ein Stück Eisen modelliert (Ladung in Nachbarpunkten soll möglichst oft gleich gerichtet sein), oder die Anzahl der Nachbarpaare, in denen ein Molekül sitzt, $H_n = \#\{(x,y) \in \Lambda_n^2 \colon |x-y| = 1, i_x = i_y = 1\}$. $\diamond$

Kapitel 10

Stochastische Prozesse

In diesem Abschnitt wenden wir uns Folgen von Zufallsgrößen und ihrer Verteilung zu. In Abschnitt 10.1 diskutieren wir die Frage, ob die ganze Folge als eine einzige Zufallsgröße aufgefasst werden kann, und wodurch die Verteilung dieser Zufallsgröße festgelegt werden kann. Insbesondere werden wir Folgen von unabhängigen und identisch verteilten Zufallsgrößen konstruieren und unendlich lange Markovketten. Das Hauptmittel dabei wird der *Satz von Ionescu Tulcea* sein; siehe Satz 10.1.14. In Abschnitt 10.2 stellen wir einen weiteren wichtigen Typ stochastischer Prozesse vor: jene, deren Verteilung sich nicht bei zeitlicher Verschiebung ändert. Das Hauptergebnis dieses Abschnittes ist der Birkhoff'sche Ergodensatz, eine kraftvolle Verallgemeinerung des Starken Gesetzes der Großen Zahlen.

10.1 Konstruktion stochastischer Prozesse

Im gesamten Abschnitt seien $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $(E, \mathcal{E})$ ein messbarer Raum.

Definition 10.1.1 (stochastischer Prozess). *Jede Folge $\mathbb{X} = (X_n)_{n\in\mathbb{N}_0}$ von $(E,\mathcal{E})$-wertigen Zufallsgrößen $X_n\colon \Omega \to E$ heißt ein E-*wertiger stochastischer Prozess*. Im Fall $(E,\mathcal{E}) = (\mathbb{R},\mathcal{B})$ spricht man einfach von einem* stochastischen Prozess.

Wir können also einen stochastischen Prozess $\mathbb{X} = (X_n)_{n\in\mathbb{N}_0}$ als eine $E^{\mathbb{N}_0}$-wertige Abbildung auffassen, deren Projektionen $X_n = \pi_n \circ \mathbb{X}$ Zufallsgrößen sind. Die sogenannten *eindimensionalen Projektionen* $\pi_n\colon E^{\mathbb{N}_0} \to E$ (siehe Bemerkung 6.9.2) sind gegeben durch

$$\pi_n\big((x_k)_{k\in\mathbb{N}_0}\big) = x_n.$$

Machen wir uns kurz klar, dass dann $\mathbb{X}$ eine $(E^{\mathbb{N}_0}, \mathcal{E}^{\otimes\mathbb{N}_0})$-wertige Zufallsgröße ist. Dazu erinnern wir zunächst daran, dass wir in Abschnitt 6.9 auf dem Folgenraum $E^{\mathbb{N}_0}$ die Produkt-σ-Algebra $\mathcal{E}^{\otimes\mathbb{N}_0}$ eingeführt haben. Nach Definition 6.9.1 ist $\mathcal{E}^{\otimes\mathbb{N}_0}$ die kleinste σ-Algebra auf $E^{\mathbb{N}_0}$, so dass alle eindimensionalen Projektionen π_n messbar sind.

Lemma 10.1.2. *Es seien $X_1, X_2, X_3, \ldots\colon \Omega \to E$ Abbildungen. Dann ist $\mathbb{X} = (X_n)_{n\in\mathbb{N}_0}$ genau dann ein stochastischer Prozess, wenn $\mathbb{X}$ eine $(E^{\mathbb{N}_0}, \mathcal{E}^{\otimes\mathbb{N}_0})$-wertige Zufallsgröße ist.*

Beweis. Falls $\mathbb{X}\colon \Omega \to E^{\mathbb{N}_0}$ messbar ist, so ist es auch jedes $X_n = \pi_n \circ \mathbb{X}$, denn π_n ist messbar. Wenn umgekehrt jedes $X_n\colon \Omega \to E$ messbar ist, so ist wegen $\mathbb{X}^{-1}(\pi_n^{-1}(A)) = X_n^{-1}(A) \in \mathcal{F}$ für jedes $A \in \mathcal{E}$ und alle $n \in \mathbb{N}_0$ das Urbild-Mengensystem $\mathbb{X}^{-1}(\bigcup_{n\in\mathbb{N}_0} \pi_n^{-1}(\mathcal{E}))$ in $\mathcal{F}$ enthalten. Da das Mengensystem $\bigcup_{n\in\mathbb{N}_0} \pi_n^{-1}(\mathcal{E})$ per definitionem die σ-Algebra $\mathcal{E}^{\otimes\mathbb{N}_0}$ erzeugt, folgt die Messbarkeit von $\mathbb{X}$ aus Lemma 6.4.3. $\square$

Die Verteilung von $\mathbb{X}$ ist also einfach seine Verteilung als eine $(E^{\mathbb{N}_0}, \mathcal{E}^{\otimes\mathbb{N}_0})$-wertige Zufallsgröße, also das Wahrscheinlichkeitsmaß $\mathbb{P}\circ\mathbb{X}^{-1}$ auf $(E^{\mathbb{N}_0}, \mathcal{E}^{\otimes\mathbb{N}_0})$. Die uns im Folgenden interessierenden Fragen werden nur von dieser Verteilung abhängen, also ist es keine Einschränkung, wenn wir annehmen, dass der zu Grunde liegende Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, \mathbb{P})$ von der Form $(E^{\mathbb{N}_0}, \mathcal{E}^{\otimes\mathbb{N}_0}, \mathbb{P})$ ist und dass die Komponenten X_n des Prozesses die Koordinatenprojektionen π_n sind: Falls nämlich $\mathbb{X}$ ein stochastischer Prozess auf einem abstrakten Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, \mathbb{P})$ ist, so bilden die Projektionen $\pi_n\colon E^{\mathbb{N}_0} \to E$ einen auf $(E^{\mathbb{N}_0}, \mathcal{E}^{\otimes\mathbb{N}_0}, \mathbb{P}\circ\mathbb{X}^{-1})$ definierten Prozess, der die selbe Verteilung wie $\mathbb{X}$ besitzt.

Die Beschreibung der Verteilung eines stochastischen Prozesses ist nicht ganz einfach und wird meist über die Verteilungen der *endlichdimensionalen Projektionen* $\varrho_n\colon E^{\mathbb{N}_0} \to E^{n+1}$ gemacht, die definiert sind durch

$$\varrho_n\big((x_k)_{k\in\mathbb{N}_0}\big) = (x_0, \dots, x_n).$$

Mit Hilfe von Bemerkung 6.9.3 sieht man leicht, dass $\mathcal{E}^{\otimes\mathbb{N}_0}$ auch die kleinste σ-Algebra auf $E^{\mathbb{N}_0}$ ist, so dass alle endlichdimensionalen Projektionen ϱ_n messbar sind.

Definition 10.1.3 (endlichdimensionale Verteilungen)**.** *Mit den* endlichdimensionalen Verteilungen *eines Wahrscheinlichkeitsmaßes Q auf $(E^{\mathbb{N}_0}, \mathcal{E}^{\otimes\mathbb{N}_0})$ bezeichnen wir die Bildmaße $Q\circ\varrho_n^{-1}$ auf $(E^{n+1}, \mathcal{E}^{\otimes(n+1)})$.*

Es ist meist leicht, für jedes $n \in \mathbb{N}_0$ die Verteilung eines Zufallsvektors $(X_0^{(n)}, \dots, X_n^{(n)})$ auf E^{n+1} anzugeben. Dann erhebt sich die natürliche Frage, ob wir diese Zufallsgrößen als den Beginn genau eines stochastischen Prozesses $(X_0, X_1, X_2, \dots)$ auffassen können, d. h. ob es einen Prozess $\mathbb{X}$ so gibt, dass die Verteilungen der Vektoren $(X_0^{(n)}, \dots, X_n^{(n)})$ und $(X_0, \dots, X_n)$ für jedes $n \in \mathbb{N}_0$ überein stimmen. Mit anderen Worten, wir stellen die beiden Fragen:

1. Legen die endlichdimensionalen Verteilungen einer Verteilung auf $(E^{\mathbb{N}_0}, \mathcal{E}^{\otimes\mathbb{N}_0})$ diese Verteilung eindeutig fest?

2. Gibt es zu einer vorgegebenen Folge von Verteilungen Q_n auf $(E^{n+1}, \mathcal{E}^{\otimes(n+1)})$ eine Verteilung Q auf $(E^{\mathbb{N}_0}, \mathcal{E}^{\otimes\mathbb{N}_0})$ mit $Q\circ\varrho_n^{-1} = Q_n$ für alle $n \in \mathbb{N}_0$?

Die Anwort auf die erste Frage ist leicht zu geben und lautet immer *Ja*:

Lemma 10.1.4. *Die Folge der endlichdimensionalen Verteilungen $Q\circ\varrho_n^{-1}$ bestimmt das Maß Q auf $(E^{\mathbb{N}_0}, \mathcal{E}^{\otimes\mathbb{N}_0})$ eindeutig.*

Beweis. Durch die Verteilungen $Q\circ\varrho_n^{-1}$ wird Q auf dem durchschnittstabilen Erzeugendensystem $\bigcup_{n\in\mathbb{N}_0} \varrho_n^{-1}(\mathcal{E}^{\otimes(n+1)})$ von $\mathcal{E}^{\otimes\mathbb{N}_0}$ (siehe Bemerkung 6.9.3) eindeutig festgelegt. Nach dem Eindeutigkeitssatz 6.2.8 ist Q auch auf $\mathcal{E}^{\otimes\mathbb{N}_0}$ eindeutig festgelegt. $\square$

Beispiel 10.1.5. Die *ein*dimensionalen Verteilungen legen Q im Allgemeinen nicht eindeutig fest: Man betrachte eine Folge $(X_n)_{n\in\mathbb{N}_0}$ unabhängiger identisch verteilter Zufallsgrößen und die Folge $(X'_n)_{n\in\mathbb{N}_0}$, die durch $X'_n = X_0$ für alle $n \in \mathbb{N}_0$ definiert wird. $\diamond$

Die Antwort auf die zweite Frage ist viel schwieriger und lautet *Ja* unter einigen einschränkenden Bedingungen. Zunächst sieht man relativ leicht ein, dass eine gewisse Verträglichkeitsbedingung erfüllt sein muss, damit überhaupt eine Chance besteht, ein Wahrscheinlichkeitsmaß Q zu finden mit $Q \circ \varrho_n^{-1} = Q_n$ für alle n. Bezeichnet nämlich

$$\varphi_n \colon E^{n+1} \to E^n, \qquad \varphi_n(x_0, \dots, x_n) = (x_0, \dots, x_{n-1}),$$

die Projektion auf die ersten n Koordinaten, so gilt $\varrho_{n-1} = \varphi_n \circ \varrho_n$, also auch $\varrho_{n-1}^{-1} = \varrho_n^{-1} \circ \varphi_n^{-1}$. Wenn also $Q \circ \varrho_n^{-1} = Q_n$ für alle n, so gilt

$$Q_{n-1} = Q \circ \varrho_{n-1}^{-1} = Q \circ \varrho_n^{-1} \circ \varphi_n^{-1} = Q_n \circ \varphi_n^{-1}. \tag{10.1.1}$$

Definition 10.1.6 (verträglich, konsistent)**.** *Eine Folge von Maßen Q_n auf $(E^{n+1}, \mathcal{E}^{\otimes(n+1)})$ heißt* verträglich *oder* konsistent*, falls für jedes $n \in \mathbb{N}$ gilt: $Q_{n-1} = Q_n \circ \varphi_n^{-1}$.*

Eine Folge $(X_0^{(n)}, \dots, X_n^{(n)})_{n\in\mathbb{N}_0}$ von Zufallsvektoren heißt konsistent, wenn die Folge der Verteilungen von $(X_0^{(n)}, \dots, X_n^{(n)})$ konsistent ist, d. h. wenn gilt:

$$(X_0^{(n)}, \dots, X_{n-1}^{(n)}) \stackrel{\mathcal{D}}{=} (X_0^{(n-1)}, \dots, X_{n-1}^{(n-1)}), \qquad n \in \mathbb{N},$$

wobei wir an die Schreibweise $\stackrel{\mathcal{D}}{=}$ erinnern, die Gleichheit in Verteilung bezeichnet. Die Konsistenz der Folge ist intuitiv gleichbedeutend damit, dass die Kette sich sukzessive zeitlich fortschreitend entwickelt, ohne in die Zukunft zu sehen. Man spricht auch manchmal – etwas schlampig – von der Konsistenz der Folge $(X_n)_n$, muss aber immer im Gedächtnis behalten, dass die Konsistenz der Folge der endlich dimensionalen Verteilungen gemeint ist.

Beispiel 10.1.7. (i) Folgen von unabhängigen Zufallsgrößen und Markovketten im Sinne der Definition 9.1.3 sind konsistent, siehe auch Bemerkung 9.1.7. Genauer ausgedrückt, gilt Folgendes: Wenn die Verteilung des Zufallsvektors $(X_0^{(n)}, \dots, X_n^{(n)})$ auf einer diskreten Menge I^{n+1} gegeben ist durch

$$\mathbb{P}\big((X_0^{(n)}, \dots, X_n^{(n)}) = (i_0, \dots, i_n)\big) = \nu(i_0) p_{i_0,i_1} \cdot \dots \cdot p_{i_{n-1},i_n}, \qquad i_0, \dots, i_n \in I,$$

wobei ν eine Verteilung auf I ist und $P = (p_{i,j})_{i,j\in I}$ eine stochastische Matrix, so rechnet man leicht nach, dass $(X_0^{(n)}, \dots, X_{n-1}^{(n)})$ und $(X_0^{(n-1)}, \dots, X_{n-1}^{(n-1)})$ die selbe Verteilung besitzen. Das gilt auch noch für zeitlich inhomogene Markovketten, wo die Übergangsmatrix vom Zeitpunkt abhängen darf.

(ii) Ein weiteres Beispiel für eine konsistente Folge ist ein Prozess $(X_n)_{n\in\mathbb{N}_0}$, der dem Baugesetz $X_{n+1} = f_n(X_0, \dots, X_n)$ für eine Folge von geeigneten zufälligen Abbildungen $f_n \colon E^{n+1} \to E$ gehorcht, wobei der eingespeiste Zufall unabhängig ist von $(X_n)_n$.

(iii) Keine Konsistenz liegt vor, wenn $(X_0^{(n)}, \dots, X_n^{(n)})$ für jedes $n \in \mathbb{N}_0$ gleichförmig verteilt ist auf der Menge aller n-schrittigen Nächstnachbarschaftspfade im $\mathbb{Z}^d$ (mit $d \geq 2$), die im Ursprung starten und keinen Punkt zweimal besuchen (Übungsaufgabe). Dies ist ein Modell für eine Polymerkette und wird der *self-avoiding walk* oder *selbstvermeidende Irrfahrt* genannt.

(iv) Pfadmaße $\mathbb{P}_n$ von der Form

$$\frac{\mathrm{d}\mathbb{P}_n}{\mathrm{d}\mathbb{P}}(x_0, \dots, x_n) = \frac{1}{Z_{n,\beta}} \mathrm{e}^{-\beta H_n(x_0,\dots,x_n)},$$

wobei $H_n\colon (\mathbb{R}^d)^{n+1} \to \mathbb{R}$ eine Energiefunktion, $\beta > 0$ ein Stärkenparameter und $Z_{n,\beta}$ die Normierungskonstante sind und $\mathbb{P}$ die Verteilung einer Irrfahrt im $\mathbb{R}^d$, sind im Allgemeinen nicht konsistent. Dies kann man als eine Übungsaufgabe am Beispiel des *self-repellent walk* nachrechnen, wo $\mathbb{P}$ die Verteilung der einfachen Irrfahrt auf dem $\mathbb{Z}^d$ ist (siehe Beispiel 9.2.4) und $H_n(x_0, \dots, x_n) = \sum_{0 \leq i < j \leq n} \mathbb{1}_{\{x_i = x_j\}}$ die Anzahl der Selbstüberschneidungen des Pfades ist.

(v) Eine Ausnahme von der Faustregel in (iv) liegt vor, wenn die Dichte, mit der $\mathbb{P}_n$ aus $\mathbb{P}$ konstruiert wird, die folgende besondere Eigenschaft besitzt. Sei $(X_n)_{n\in\mathbb{N}}$ eine Folge von Zufallsgrößen, und sei $(M_n)_{n\in\mathbb{N}_0}$ ein nichtnegatives Martingal, d. h., für jedes $n \in \mathbb{N}$ sei $M_n \geq 0$ integrierbar und messbar bezüglich $\sigma(X_1, \dots, X_n)$ mit $\mathbb{E}[M_n \mid X_1, \dots, X_{n-1}] = M_{n-1}$, sowie $M_0 = 1$. Dann ist die durch $\mathrm{d}\mathbb{P}_n = M_n \,\mathrm{d}\mathbb{P}$ gebildete Folge von Wahrscheinlichkeitsmaßen $\mathbb{P}_n$ konsistent (Übungsaufgabe).

Bei der Frage nach der Existenz eines Maßes Q mit $Q \circ \varrho_n^{-1} = Q_n$ für alle n müssen wir also mindestens die Verträglichkeit der Folge $(Q_n)_{n\in\mathbb{N}_0}$ voraussetzen. Allerdings lautet auch in diesem Fall die Antwort nicht immer 'Ja'. Wir werden aber im Satz 10.1.14 ein hinreichendes Kriterium angeben, das die Existenz von Q in praktisch allen für die Wahrscheinlichkeitstheorie interessanten Fälle sichert. In diesem Satz wird vorausgesetzt, dass die Maße Q_n auf eine bestimmte Weise mit Hilfe von Markovkernen gegeben sind, die wir ja in Definition 7.3.21 eingeführt haben. Machen wir uns zunächst klar, wie diese Konstruktion funktioniert und dass diese Voraussetzung zumindest für polnische Räume E keine Einschränkung darstellt.

Lemma 10.1.8. *(i) Falls für jedes $n \in \mathbb{N}_0$ ein Markovkern K_n von $(E^{n+1}, \mathcal{E}^{\otimes(n+1)})$ nach $(E, \mathcal{E})$ existiert, so dass gilt:*

$$Q_{n+1}\big(\mathrm{d}(x_0, \dots, x_{n+1})\big) = Q_n\big(\mathrm{d}(x_0, \dots, x_n)\big) K_n\big((x_0, \dots, x_n), \mathrm{d}x_{n+1}\big), \tag{10.1.2}$$

so ist die Familie $(Q_n)_{n\in\mathbb{N}_0}$ konsistent.

(ii) Falls $(E, \mathcal{E})$ ein polnischer Raum mit seiner Borel-σ-Algebra ist und $(Q_n)_{n\in\mathbb{N}_0}$ eine konsistente Familie von Wahrscheinlichkeitsmaßen Q_n auf $(E^{n+1}, \mathcal{E}^{\otimes(n+1)})$, so gibt es zu jedem $n \in \mathbb{N}_0$ ein Markovkern K_n von $(E^{n+1}, \mathcal{E}^{\otimes(n+1)})$ nach $(E, \mathcal{E})$, so dass (10.1.2) gilt.

Wir verzichten auf einen strengen Beweis von Lemma 10.1.8 und begnügen uns mit ein paar Erläuterungen. Die Verteilung Q_{n+1} von $(X_0, \dots, X_{n+1})$ wird also in (10.1.2) rekursiv konstruiert, indem man mit Q_n den Vektor $(X_0, \dots, X_n)$ 'auswürfelt' und dann mit dem Wahrscheinlichkeitsmaß $K_n((x_0, \dots, x_n), \cdot)$ die Zufallsgröße X_{n+1}, und die Verteilung des resultierenden

Vektors $(X_0, \dots, X_{n+1})$ ist Q_{n+1}. Man sollte $K_n((x_0, \dots, x_n), A)$ als die bedingte Wahrscheinlichkeit interpretieren, dass X_{n+1} in A liegt, gegeben $\{X_0 = x_0, \dots, X_n = x_n\}$. Mit anderen Worten, der Kern K_n ist eine Version der bedingten Verteilung von X_{n+1} gegeben $(X_0, \dots, X_n)$; siehe Abschnitt 7.3. Die Konsistenz der Q_n ist leicht aus (10.1.2) herzuleiten, sie folgt auch aus (10.1.3) weiter unten. In dem Spezialfall, dass $K_n((x_0, \dots, x_n), \mathrm{d}x_{n+1}) = K(x_n, \mathrm{d}x_{n+1})$ gar nicht von $x_0, \dots, x_{n-1}$ oder n abhängt, haben wir also eine homogene Markovkette vor uns, und wenn $K(x_n, \mathrm{d}x_{n+1}) = \mu(\mathrm{d}x_{n+1})$ nicht einmal von x_n abhängt, so sind die X_n unabhängig.

Unter der Verträglichkeitsbedingung $Q_n = Q_{n+1} \circ \varphi_{n+1}^{-1}$ ist die Verteilung von $X_{n+1}^{(n+1)}$ tatsächlich die bedingte Verteilung (unter Q_{n+1}) gegeben $(X_0^{(n+1)}, \dots, X_n^{(n+1)})$, was ja die selbe Verteilung wie $(X_0^{(n)}, \dots, X_n^{(n)})$ besitzt. Aus dem Satz 7.3.25 über die Existenz einer regulären Version einer bedingten Verteilung (man beachte, dass auch $(E^{n+1}, \mathcal{E}^{\otimes(n+1)})$ polnisch ist) folgt dann, dass eine reguläre Version K_n wie oben existiert, dass also Q_{n+1} wie in (10.1.2) mit Hilfe von Q_n und K_n gegeben ist. Es wird also ausreichend vorauszusetzen, dass die Q_n auf diese Weise gegeben sind, und dies werden wir in unserer Version des Satzes von Ionescu Tulcea tun.

Im Folgenden stellen wir die nötigen theoretischen Hilfsmittel bereit. Zunächst verallgemeinern wir den Begriff des Produktmaßes:

Definition 10.1.9. *Es seien $(E_1, \mathcal{E}_1)$ und $(E_2, \mathcal{E}_2)$ zwei messbare Räume und K ein Kern von E_1 nach E_2. Ferner sei μ ein Wahrscheinlichkeitsmaß auf $(E_1, \mathcal{E}_1)$. Dann ist das Wahrscheinlichkeitsmaß $\mu \otimes K$ auf $(E_1 \times E_2, \mathcal{E}_1 \otimes \mathcal{E}_2)$ definiert durch*

$$(\mu \otimes K)(A) = \int_{E_1} K(\omega_1, A^{(1)}_{\omega_1})\, \mu(\mathrm{d}\omega_1), \qquad A \in \mathcal{E}_1 \otimes \mathcal{E}_2,$$

wobei $A^{(1)}_{\omega_1} = \{\omega_2 \in E_2 \colon (\omega_1, \omega_2) \in A\}$ der ω_1-Schnitt von A ist.

Bemerkung 10.1.10. Man muss sich Einiges überlegen, um einzusehen, dass Definition 10.1.9 sinnvoll ist:

- Nach Satz 6.10.1 ist $A^{(1)}_{\omega_1} \in \mathcal{E}_2$, also $K(\omega_1, A^{(1)}_{\omega_1})$ wohldefiniert.
- Dass $\omega_1 \mapsto K(\omega_1, A^{(1)}_{\omega_1})$ messbar ist, sieht man ein, indem man nachprüft, dass

  $$\mathcal{D} = \{A \in \mathcal{E}_1 \otimes \mathcal{E}_2 \colon \omega_1 \mapsto K(\omega_1, A^{(1)}_{\omega_1}) \text{ ist messbar}\}$$

  ein Dynkin-System ist (einfache Übungsaufgabe) und die Mengen der Form $A_1 \times A_2$ mit $A_i \in \mathcal{E}_i$ für $i \in \{1, 2\}$ enthält (dies folgt aus $K(\omega_1, (A_1 \times A_2)^{(1)}_{\omega_1}) = \mathbb{1}_{A_1}(\omega_1) K(\omega_1, A_2)$). Dann folgt aus Satz 6.2.3, dass $\mathcal{D} = \mathcal{E}_1 \otimes \mathcal{E}_2$, d. h. $\mu \otimes K(A)$ ist wohldefiniert für jedes $A \in \mathcal{E}_1 \otimes \mathcal{E}_2$.
- Mit Hilfe des Satzes 6.8.6 von Lebesgue zeigt man, dass $\mu \otimes K$ auch σ-additiv, also ein Wahrscheinlichkeitsmaß ist.

$\diamond$

Also können wir die Gleichung (10.1.2) auch kurz mit $Q_{n+1} = Q_n \otimes K_n$ ausdrücken.

Beispiel 10.1.11. (a) Für den Kern K aus Beispiel 7.3.22(a) (also $K(x, A) = \nu(A)$ für ein Maß ν) ist $\mu \otimes K$ gleich dem Produktmaß $\mu \otimes \nu$.

(c) Für den Kern K aus Beispiel 7.3.22(c) (also $K(i,A)=\sum_{j\in A}p_{i,j}$ mit einer stochastischen Matrix $P=(p_{i,j})_{i,j\in i})$ ist $\mu\otimes K$ gleich der Verteilung des Paars (X_0,X_1), wenn $(X_n)_{n\in\mathbb{N}_0}$ die Markovkette mit Übergangsmatrix P und Startverteilung μ auf I ist.

$\diamond$

Bemerkung 10.1.12 (Randverteilungen)**.** Wir betrachten kurz die Randverteilungen des Maßes $\mu\otimes K$ auf E_1 bzw. E_2. Zunächst zur Begriffsbildung: Wenn $\pi_i\colon E_1\times E_2\to E_i$ die i-te Projektion bezeichnet, so ist $\mathfrak{m}\circ\pi_i^{-1}$ die i-te Randverteilung eines Maßes $\mathfrak{m}$ auf $E_1\times E_2$. Für $i=1$ ist

$$(\mu\otimes K)\circ\pi_1^{-1}=\mu, \tag{10.1.3}$$

wie man sich leicht klar macht, und die zweite Randverteilung wird mit μK bezeichnet, also

$$\mu K(A)=(\mu\otimes K)(E_1\times A)=\int\mu(\mathrm{d}x)\,K(x,A),\qquad A\in\mathcal{E}_2. \tag{10.1.4}$$

$\diamond$

Wir benötigen noch Verallgemeinerungen der Sätze 6.10.4 von Tonelli und 6.10.5 von Fubini:

Satz 10.1.13. *Es seien $(E_1,\mathcal{E}_1)$ und $(E_2,\mathcal{E}_2)$ zwei messbare Räume, μ ein Wahrscheinlichkeitsmaß auf $(E_1,\mathcal{E}_1)$, K ein Markovkern von E_1 nach E_2 und $f\colon E_1\times E_2\to\mathbb{R}$ eine messbare Abbildung.*

(a) Falls f nichtnegativ ist, so gilt

$$\int f\,\mathrm{d}\big(\mu\otimes K\big)=\int\Big(\int f(x,y)\,K(x,\mathrm{d}y)\Big)\,\mu(\mathrm{d}x). \tag{10.1.5}$$

(b) Falls $f\in\mathcal{L}^1(\mu\otimes K)$, so ist für μ-fast alle $x\in E_1$ die Abbildung $f(x,\cdot)$ integrierbar bezüglich $K(x,\cdot)$. Ferner ist $\int f(\cdot,y)\,K(\cdot,\mathrm{d}y)$ bezüglich μ integrierbar, und es gilt (10.1.5)*.*

Beweis. Der Beweis folgt dem üblichen Schema mit Hilfe von Satz 6.4.11. $\square$

Wir formulieren nun endlich das Hauptergebnis dieses Abschnittes.

Satz 10.1.14 (Satz von Ionescu Tulcea)**.** *Es sei $(E,\mathcal{E})$ ein messbarer Raum und μ ein Wahrscheinlichkeitsmaß auf $(E,\mathcal{E})$. Für jedes $n\in\mathbb{N}_0$ sei K_n ein Markovkern von $(E^{n+1},\mathcal{E}^{\otimes(n+1)})$ nach $(E,\mathcal{E})$. Wir betrachten die induktiv durch*

$$Q_0=\mu\qquad\textit{und}\qquad Q_n=Q_{n-1}\otimes K_{n-1},\qquad n\in\mathbb{N},$$

definierten Wahrscheinlichkeitsmaße Q_n auf $(E^{n+1},\mathcal{E}^{\otimes(n+1)})$. Dann existiert genau ein Wahrscheinlichkeitsmaß Q auf $(E^{\mathbb{N}_0},\mathcal{E}^{\otimes\mathbb{N}_0})$ mit $Q\circ\varrho_n^{-1}=Q_n$ für jedes $n\in\mathbb{N}_0$.

Man beachte, dass $(Q_n)_{n\in\mathbb{N}_0}$ verträglich ist nach (10.1.3), angewendet auf $E_1=E^n$ und $\mu=Q_{n-1}$, denn dann ist $\pi_1=\varphi_n$.

Im Beweis des Satzes 10.1.14 benötigen wir Verallgemeinerungen der Kerne K_n, und zwar Kerne $K_{n,m}$ von E^{n+1} nach E^m für beliebiges $m\in\mathbb{N}$, die die m-stufigen Übergänge des Prozesses beschreiben. (Die Interpretation dieser Kerne ist, dass $K_{n,m}((x_0,\dots,x_n),A)$ gleich der bedingten

Wahrscheinlichkeit von $\{(X_{n+1},\dots,X_{n+m}) \in A\}$ gegeben $\{(X_0,\dots,X_n) = (x_0,\dots,x_n)\}$ ist, wenn $(X_k)_{k\in\mathbb{N}_0}$ der Prozess unter Q ist.) Wir definieren induktiv $K_{n,1} = K_n$ und

$$K_{n,m+1}(x,A) = \int_{E^m} K_{n,m}(x,\mathrm{d}y) \int_E K_{n+m}((x,y),\mathrm{d}z) \mathbb{1}_A(y,z), \qquad x \in E^{n+1}, A \in \mathcal{E}^{\otimes(m+1)}.$$

Dann haben wir nämlich:

Lemma 10.1.15. *Für alle $n \in \mathbb{N}_0$ und $m \in \mathbb{N}$ gilt $Q_{n+m} = Q_n \otimes K_{n,m}$.*

Beweis. Wir führen eine Induktion nach m. Der Fall $m = 1$ ist klar nach Definition. Es sei $m \geq 2$. Für $A \in \mathcal{E}^{\otimes(n+1)}$, $B \in \mathcal{E}^{\otimes(m-1)}$ und $C \in \mathcal{E}$ benutzen wir nach einander die Definition von Q_{n+m}, die Induktionsvoraussetzung, die Definition von $K_{n,m-1}$ und die Definition von $Q_n \otimes K_{n,m}$, um zu erhalten:

$$\begin{aligned}
Q_{n+m}(A \times B \times C) &= \int_{A\times B} Q_{n+m-1}(\mathrm{d}(x,y))\, K_{n+m-1}((x,y),C) \\
&= \int_A Q_n(\mathrm{d}x) \int_B K_{n,m-1}(x,\mathrm{d}y)\, K_{n+m-1}((x,y),C) \\
&= \int_A Q_n(\mathrm{d}x)\, K_{n,m}(x, B\times C) \\
&= \big(Q_n \otimes K_{n,m}\big)(A\times B\times C).
\end{aligned}$$

Da die Mengen der Form $A \times B \times C$ mit $A \in \mathcal{E}^{\otimes(n+1)}$, $B \in \mathcal{E}^{\otimes(m-1)}$ und $C \in \mathcal{E}$ ein durchschnittstabiles Erzeugendensystem von $\mathcal{E}^{\otimes(n+m+1)}$ bilden, folgt die Aussage aus dem Eindeutigkeitssatz 6.2.8. □

Beweis des Satzes 10.1.14. Wir werden den Satz 6.2.7 von Carathéodory anwenden, müssen aber zuvor die Maße Q_n (die ja jeweils auf verschiedenen Räumen definiert sind) auf den Raum $\Omega = E^{\mathbb{N}_0}$ 'liften'. Wir betten also die Definitionsbereiche $(E^{n+1}, \mathcal{E}^{\otimes(n+1)})$ von Q_n in Ω ein, indem wir E^{n+1} durch $E^{n+1} \times E^{\mathbb{N}_0}$ ersetzen und $\mathcal{E}^{\otimes(n+1)}$ durch die Urbild-σ-Algebra

$$\mathcal{F}_n = \varrho_n^{-1}(\mathcal{E}^{\otimes(n+1)})$$

auf Ω. Man beachte, dass $(\mathcal{F}_n)_{n\in\mathbb{N}_0}$ eine aufsteigende Folge von σ-Algebren ist, deren Vereinigung

$$\mathcal{A} = \bigcup_{n\in\mathbb{N}_0} \mathcal{F}_n$$

eine Algebra ist, die die Produkt-σ-Algebra $\mathcal{E}^{\otimes\mathbb{N}_0}$ erzeugt. Es ist klar, dass $\varrho_n^{-1}\colon \mathcal{E}^{\otimes(n+1)} \to \mathcal{F}_n$ bijektiv ist, und daher das Wahrscheinlichkeitsmaß

$$\overline{Q}_n = Q_n \circ \varrho_n,$$

auf $(\Omega, \mathcal{F}_n)$ eineindeutig dem Wahrscheinlichkeitsmaß Q_n auf $(E^{n+1}, \mathcal{E}^{\otimes(n+1)})$ zugeordnet ist. Die Verträglichkeit der Folge $(Q_n)_{n\in\mathbb{N}_0}$ ist gleichbedeutend damit, dass für alle $n \leq m$ die Maße $\overline{Q}_m$ und $\overline{Q}_n$ auf $\mathcal{F}_m$ übereinstimmen. Daher bestimmen die $\overline{Q}_n$ eindeutig eine Abbildung $\overline{Q}\colon \mathcal{A} \to [0,1]$. Es ist klar, dass $\overline{Q}$ ein Inhalt auf der Algebra $\mathcal{A}$ mit $\overline{Q}(\Omega) = 1$ ist.

Nach dem Satz 6.2.7 von Carathéodory ist also nur noch zu zeigen, dass $\overline{Q}$ σ-additiv ist, denn dann lässt sich $\overline{Q}$ eindeutig zu einem Maß Q auf $\sigma(\mathcal{A}) = \mathcal{E}^{\otimes\mathbb{N}_0}$ erweitern, und dann gilt $Q \circ \varrho_n^{-1} = Q_n$ für jedes $n \in \mathbb{N}_0$, und das beendet den Beweis.

Nach Lemma 6.2.6 reicht es zu zeigen, dass $\overline{Q}$ stetig in $\emptyset$ ist. Es sei also $(B_n)_{n\in\mathbb{N}_0}$ eine fallende Folge von Elementen aus $\mathcal{A}$ mit $\lim_{n\to\infty}\overline{Q}(B_n)>0$, dann müssen wir zeigen, dass der Schnitt $\bigcap_{n\in\mathbb{N}_0} B_n$ nicht leer ist. Als Übungsaufgabe überlege man sich, dass man voraussetzen kann, dass $B_n\in\mathcal{F}_n$ für jedes $n\in\mathbb{N}_0$ gilt, und dies wollen wir nun voraussetzen.

Für $n\in\mathbb{N}_0$ sei $A_n\in\mathcal{E}^{\otimes(n+1)}$ die eindeutig bestimmte Menge mit $B_n=\varrho_n^{-1}(A_n)$. Dann gilt $B_n=\varrho_{n+1}^{-1}(A_n\times E)$ und daher $A_{n+1}\subset A_n\times E$. Wir konstruieren nun eine Folge $x=(x_n)_{n\in\mathbb{N}_0}$ mit $(x_0,\dots,x_n)\in A_n$ für jedes $n\in\mathbb{N}_0$. Wenn wir dies getan haben, sind wir fertig, denn es gilt dann $x\in B_n$ für jedes $n\in\mathbb{N}_0$, also ist der Schnitt $\bigcap_{n\in\mathbb{N}_0} B_n$ nicht leer.

Für $m\in\mathbb{N}$ definieren wir Funktionen $f_{n,m}\colon E^{n+1}\to[0,1]$ durch $f_{n,m}(y)=K_{n,m}(y,A_{m-1})$, wobei wir $f_{n,1}$ als $\mathbb{1}_{A_n}$ interpretieren. Man beachte, dass $(f_{n,m})_{m\in\mathbb{N}}$ für jedes $n\in\mathbb{N}$ eine fallende Folge ist, denn für jedes $y\in E^{n+1}$ gilt wegen $A_m\subset A_{m-1}\times E$:

$$\begin{aligned}
f_{n,m+1}(y)&=K_{n,m+1}(y,A_m)\le K_{n,m+1}(y,A_{m-1}\times E)\\
&=\int_{E^m}K_{n,m}(y,\mathrm{d}x)\int_E K_{n+m}((y,x),\mathrm{d}z)\mathbb{1}_{A_{m-1}\times E}(x,z)\\
&=\int_{E^m}K_{n,m}(y,\mathrm{d}x)\,\mathbb{1}_{A_{m-1}}(x)K_{n+m}((y,x),E)\\
&=K_{n,m}(y,A_{m-1})=f_{n,m}(y).
\end{aligned}$$

Wir konstruieren nun rekursiv eine Folge $x=(x_n)_{n\in\mathbb{N}_0}$ mit

$$\lim_{m\to\infty} f_{n,m}((x_0,\dots,x_n))>0 \tag{10.1.6}$$

für jedes $n\in\mathbb{N}_0$. Wegen der Monotonie in m gilt dann $\mathbb{1}_{A_n}((x_0,\dots,x_n))=f_{n,0}((x_0,\dots,x_n))>0$, also $(x_0,\dots,x_n)\in A_n$, was den Beweis beendet.

Wir beginnen mit $n=0$. Es gilt

$$\int f_{0,m}(y)\,\mu(\mathrm{d}y)=\int K_{0,m}(y,A_m)\,Q_0(\mathrm{d}y)=Q_m(A_m)=\overline{Q}_m(B_m)=\overline{Q}(B_m).$$

Wegen $\lim_{m\to\infty}\overline{Q}(B_m)>0$ ist also $(f_{0,m})_{m\in\mathbb{N}_0}$ eine monoton fallende Folge, deren punktweiser Grenzwert nicht identisch Null ist. Also existiert ein $x_0\in E$, so dass (10.1.6) gilt für $n=0$.

Seien nun $x_0,\dots,x_n$ konstruiert, so dass (10.1.6) gilt. Dann gilt für $m\in\mathbb{N}$:

$$\begin{aligned}
&\int_E f_{n+1,m}((x_0,\dots,x_n,y))\,K_n((x_0,\dots,x_n),\mathrm{d}y)\\
&\qquad=\int_E K_{n+1,m}((x_0,\dots,x_n,y),A_m)\,K_n((x_0,\dots,x_n),\mathrm{d}y)\\
&\qquad=K_{n,m}((x_0,\dots,x_n),A_m)\\
&\qquad=f_{n,m}((x_0,\dots,x_n)).
\end{aligned}$$

Somit ist die Folge der Abbildungen $E\ni y\mapsto f_{n+1,m}((x_0,\dots,x_n,y))$ eine in m fallende Folge, deren Grenzwert nicht identisch Null ist. Wegen (10.1.6) für n gibt es also ein $x_{n+1}\in E$, so dass (10.1.6) auch für $n+1$ gilt. □

Beispiel 10.1.16. (a) Der Satz 10.1.14 von Ionescu Tulcea garantiert die Existenz von Folgen $(X_n)_{n\in\mathbb{N}_0}$ unabhängiger Zufallsgrößen mit beliebig vorgegebenen Verteilungen der einzelnen

X_n. Wenn nämlich μ_n für jedes $n \in \mathbb{N}_0$ ein Wahrscheinlichkeitsmaß auf einem messbaren Raum $(E, \mathcal{E})$ ist, so ist das Maß $Q_n = \mu_0 \otimes \cdots \otimes \mu_n$ von der Form $Q_n = Q_{n-1} \otimes K_{n-1}$ mit $K_n(x, A) = \mu_{n+1}(A)$ (siehe Beispiel 7.3.22(a)). Also gibt es ein Maß Q auf $(E^{\mathbb{N}_0}, \mathcal{E}^{\otimes \mathbb{N}_0})$, so dass die n-te Projektion $X_n : E^{\mathbb{N}_0} \to E$ die Verteilung μ_n hat.

(b) Auch die Existenz von Markovketten im Sinne von Kapitel 9 wird durch den Satz von Ionescu Tulcea abgedeckt. (Dadurch wird endlich die in Bemerkung 9.1.7 aufgeworfene Frage positiv beantwortet.) Es sei P eine stochastische Matrix auf einer abzählbaren Menge I, und es sei μ ein Wahrscheinlichkeitsmaß auf $(I, \mathcal{P}(I))$. Wir betrachten die durch $K_n((i_0, \dots, i_n), i_{n+1}) = p_{i_n, i_{n+1}}$ definierten Kerne K_n. Durch $Q_0 = \mu$ und $Q_n = Q_{n-1} \otimes K_{n-1}$ wird eine Folge $(Q_n)_{n \in \mathbb{N}_0}$ von Wahrscheinlichkeitsmaßen festgelegt. Man sieht leicht, dass Q_n die Verteilung einer endlichen Markovkette $(X_0, \dots, X_n)$ mit Startverteilung μ und Übergangsmatrix P ist. Nach dem Satz von Ionescu Tulcea existiert genau ein Wahrscheinlichkeitsmaß Q auf $(I^{\mathbb{N}_0}, \mathcal{P}(I)^{\otimes \mathbb{N}_0})$, dessen endlichdimensionale Verteilungen die Q_n sind. Q ist die Verteilung einer unendlich langen Markovkette $(X_n)_{n \in \mathbb{N}_0}$ mit Startverteilung μ und Übergangsmatrix P. Natürlich gilt dies auch für zeitlich inhomogene Ketten und für andere Verallgemeinerungen.

(c) Es seien I und J abzählbare Mengen, und es sei $(Y_n)_{n \in \mathbb{N}_0}$ eine Folge von unabhängigen identisch verteilten J-wertigen Zufallsgrößen. Es seien I-wertige Zufallsgrößen $X_0, X_1, X_2, \dots$ gegeben, die dem Baugesetz $X_{n+1} = f_n(X_0, \dots, X_n, Y_n)$ für eine Funktion $f_n : I^{n+1} \times J \to I$ gehorchen. In anderen Worten: X_{n+1} ist eine zufällige Funktion des gesamten 'Vorgeschichte' $X_0, \dots, X_n$. Wieder zeigt der Satz von Ionescu Tulcea, dass genau ein Wahrscheinlichkeitsmaß Q auf $(I^{\mathbb{N}_0}, \mathcal{P}(I)^{\otimes \mathbb{N}_0})$ existiert, dessen endlichdimensionale Verteilungen die Verteilungen der $(X_0, \dots, X_n)$ sind. Prozesse mit dieser Abhängigkeitsstruktur sind es, die man bei diesem Satz im Auge hatte.

$\diamond$

10.2 Stationäre Prozesse und der Ergodensatz

In diesem Abschnitt behandeln wir stochastische Prozesse, deren Verteilung durch zeitliche Verschiebung nicht geändert wird, so genannte *stationäre Prozesse*. Wir beleuchten den Zusammenhang mit Maß erhaltenden Transformationen, und als Hauptergebnis beweisen wir den Birkhoff'schen Ergodensatz, der das Starke Gesetz der Großen Zahlen als einen Spezialfall enthält. Es sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum.

Definition 10.2.1 (stationärer Prozess). *Ein stochastischer Prozess $\mathbb{X} = (X_n)_{n \in \mathbb{N}_0}$ heißt* stationär, *falls $\mathbb{X}$ die selbe Verteilung wie der zeitlich verschobene Prozess $(X_{n+1})_{n \in \mathbb{N}_0}$ hat.*

Wenn $\theta : \mathbb{R}^{\mathbb{N}_0} \to \mathbb{R}^{\mathbb{N}_0}$ der Shift-Operator ist, also $\theta((x_n)_{n \in \mathbb{N}_0}) = (x_{n+1})_{n \in \mathbb{N}_0}$ für alle Folgen $(x_n)_{n \in \mathbb{N}_0} \in \mathbb{R}^{\mathbb{N}_0}$, dann ist $\mathbb{X}$ stationär, falls $\mathbb{X}$ und $\theta \circ \mathbb{X}$ die gleiche Verteilung besitzen. Dies kann man auch schreiben als $\mathbb{P} \circ \mathbb{X}^{-1} = \mathbb{P} \circ (\theta \circ \mathbb{X})^{-1} = (\mathbb{P} \circ \mathbb{X}^{-1}) \circ \theta^{-1}$, also ist die Verteilung $\mathbb{P} \circ \mathbb{X}^{-1}$ gleich seinem Bildmaß unter θ. Dies nehmen wir zum Anlass zu einer Definition, die zunächst allgemeiner scheint:

Definition 10.2.2 (Maß erhaltende Transformation)**.** *Eine messbare Abbildung $T\colon \Omega \to \Omega$ heißt eine* Maß erhaltende Transformation, *wenn $\mathbb{P} \circ T^{-1} = \mathbb{P}$ gilt. Wir sagen dann auch, dass* $\mathbb{P}$ invariant unter T *ist oder kurz* T-invariant*.*

Nach der obigen Bemerkung ist ein stochastischer Prozess $\mathbb{X}$ also genau dann stationär, wenn der Shift θ eine Maß erhaltende Transformation auf dem Wahrscheinlichkeitsraum $(\mathbb{R}^{\mathbb{N}_0}, \mathcal{B}^{\otimes \mathbb{N}_0}, \mathbb{P} \circ \mathbb{X}^{-1})$ ist. In diesem Falle ist $\mathbb{P} \circ \mathbb{X}^{-1}$ natürlich auch die Verteilung der Folge der Projektionen $\pi_n\colon \mathbb{R}^{\mathbb{N}_0} \to \mathbb{R}$, denn $(\pi_n)_{n\in\mathbb{N}_0}$ ist die identische Abbildung auf $\mathbb{R}^{\mathbb{N}_0}$. Offensichtlich ist dann $\pi_{n+1} = \pi_n \circ \theta$ für jedes $n \in \mathbb{N}_0$.

Umgekehrt kann man aus einer beliebigen, auf einem abstrakten Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, \mathbb{P})$ definierten Maß erhaltenden Transformation einen stationären Prozess erhalten:

Lemma 10.2.3. *Es sei Y eine Zufallsgröße und T eine Maß erhaltende Transformation auf $(\Omega, \mathcal{F}, \mathbb{P})$. Definiere $X_0 = Y$ und $X_{n+1} = X_n \circ T$ für $n \in \mathbb{N}_0$. Dann ist der Prozess $\mathbb{X} = (X_n)_{n\in\mathbb{N}_0}$ stationär.*

Beweis. Für alle $n \in \mathbb{N}_0$ und $t_0, \dots, t_n \in \mathbb{R}$ gilt

$$\begin{aligned} \mathbb{P}(X_0 \le t_0, \dots, X_n \le t_n) &= (\mathbb{P} \circ T^{-1})(X_0 \le t_0, \dots, X_n \le t_n) \\ &= \mathbb{P}(X_0 \circ T \le t_0, \dots, X_n \circ T \le t_n) \\ &= \mathbb{P}(X_1 \le t_0, \dots, X_{n+1} \le t_n). \end{aligned}$$

Dies zeigt, dass die endlich dimensionalen Verteilungen von $\mathbb{X}$ und $\theta \circ \mathbb{X}$ übereinstimmen, nach Lemma 10.1.4 also auch die von $\mathbb{X}$ und $\theta \circ \mathbb{X}$. □

Also entsprechen sich die beiden Konzepte 'stationäre Prozesse' und 'Maß erhaltende Transformationen' eineindeutig: Zu jedem stationären Prozess $\mathbb{X}$ existiert eine $\mathbb{P} \circ \mathbb{X}^{-1}$-invariante Transformation T (nämlich der Shift θ auf dem Folgenraum $(\mathbb{R}^{\mathbb{N}_0}, \mathcal{B}^{\otimes \mathbb{N}_0}, \mathbb{P} \circ \mathbb{X}^{-1})$), und dann gilt $\pi_{n+1} = \pi_n \circ \theta$ für die Projektionen, und zu jeder $\mathbb{P}$-invarianten Transformation T und jeder Zufallsgröße Y existiert ein stationärer Prozess $\mathbb{X}$ mit $X_{n+1} = X_n \circ T$ für $n \in \mathbb{N}_0$. Die Tupel

$$\big(\Omega, \mathcal{F}, \mathbb{P}, T, (X_n)_{n\in\mathbb{N}_0}\big) \qquad \text{und} \qquad \big(\mathbb{R}^{\mathbb{N}_0}, \mathcal{B}^{\otimes \mathbb{N}_0}, \mathbb{P} \circ \mathbb{X}^{-1}, \theta, (\pi_n)_{n\in\mathbb{N}_0}\big)$$

entsprechen einander eineindeutig. Daher ist es nur eine Sache des Geschmacks, ob man mit einem abstrakten Raum $(\Omega, \mathcal{F}, \mathbb{P}, T)$ arbeitet oder mit dem Folgenraum $(\mathbb{R}^{\mathbb{N}_0}, \mathcal{B}^{\otimes \mathbb{N}_0}, \mathbb{P} \circ \mathbb{X}^{-1}, \theta)$. Wir werden den abstrakten Rahmen bevorzugen, aber die wichtigsten Ergebnisse auch im Zusammenhang mit stationären Prozessen formulieren.

Bemerkung 10.2.4. Eine messbare Abbildung $T\colon \Omega \to \Omega$ ist genau dann eine Maß erhaltende Transformation, wenn $\mathbb{P} \circ T^{-1}(A) = \mathbb{P}(A)$ für alle A aus einem durchschnittstabilen Erzeugendensystem von $\mathcal{F}$ gilt. Dies folgt direkt aus dem Eindeutigkeitssatz 6.2.8. ◇

Beispiel 10.2.5. (a) Eine Folge von unabhängigen identisch verteilten Zufallsgrößen ist ein stationärer Prozess. Anders ausgedrückt: Für jedes Wahrscheinlichkeitsmaß μ auf $(\mathbb{R}, \mathcal{B})$ ist das unendliche Produktmaß $\mu^{\otimes \mathbb{N}_0}$ shift-invariant.

(b) Wir fassen die Menge $\mathbb{D} = \{z \in \mathbb{C} \colon |z| = 1\}$ als einen Kreis im $\mathbb{R}^2$ auf und betrachten die gleichförmige Verteilung (das normierte Lebesgue-Maß) $\lambda_{\mathbb{D}}$ auf $\mathbb{D}$ mit der Spur-Borel-σ-Algebra $\mathcal{B}_{\mathbb{D}}$ auf $\mathbb{D}$. Für $c \in \mathbb{D}$ sei der Drehoperator $T_c\colon \mathbb{D} \to \mathbb{D}$ definiert durch $T_c(\omega) = c\omega$.

Ist $A \subset \mathbb{D}$ ein Intervall, so gilt offensichtlich $\lambda_{\mathbb{D}} \circ T_c^{-1}(A) = \lambda_{\mathbb{D}}(A)$. Aus Bemerkung 10.2.4 folgt, dass T_c Maß erhaltend ist.

(c) Auf $([0,1), \mathcal{B}_{[0,1)}, \lambda)$ betrachten wir die Transformation T, die definiert wird durch

$$T(\omega) = \begin{cases} 2\omega, & \text{falls } \omega \in [0,1/2), \\ 2\omega - 1, & \text{falls } \omega \in [1/2,1). \end{cases}$$

Für $0 \leq a < b < 1$ ist $T^{-1}([a,b)) = [\frac{a}{2}, \frac{b}{2}) \cup [\frac{a+1}{2}, \frac{b+1}{2})$, also folgt $\lambda \circ T^{-1}([a,b)) = b - a = \lambda([a,b))$. Aus Bemerkung 10.2.4 folgt, dass T Maß erhaltend ist.

(d) Jede Markovkette, die in ihrer invarianten Verteilung gestartet wird (vorausgesetzt, eine solche existiert), ist ein stationärer Prozess, siehe auch Bemerkung 9.5.2(d), wo die eindimensionalen Verteilungen betrachtet werden. Falls π eine Gleichgewichtsverteilung einer stochastischen Matrix P auf einer abzählbaren Menge I ist, so gilt nämlich für die zugehörige Markovkette $\mathbb{X} = (X_n)_{n\in\mathbb{N}_0}$, dass für jedes $n \in \mathbb{N}$ und alle $i_0, \dots, i_n \in I$:

$$\mathbb{P}_\pi(X_{[0,n]} = i_{[0,n]}) = \mathbb{P}_\pi(X_{[1,n+1]} = i_{[0,n]}),$$

wie man leicht mit Hilfe der Invarianz von π unter P nachweist. Also ist der Shift-Operator θ auch auf $(I^{\mathbb{N}_0}, \mathcal{P}(I)^{\otimes\mathbb{N}_0}, \mathbb{P} \circ \mathbb{X}^{-1})$ eine Maß erhaltende Transformation.

$\diamond$

Definition 10.2.6 (invariante Menge, ergodische Abbildung). *(a) Es sei $T\colon \Omega \to \Omega$ eine Abbildung. Eine Teilmenge A von Ω heißt* invariant unter T *oder* T-invariant, *falls $T^{-1}(A) = A$ gilt.*

(b) Eine Maß erhaltende Abbildung T auf $(\Omega, \mathcal{F}, \mathbb{P})$ heißt ergodisch, *falls für jede T-invariante messbare Menge A gilt: $\mathbb{P}(A) \in \{0,1\}$. Wir nennen dann T auch $\mathbb{P}$-ergodisch und $\mathbb{P}$ auch T-ergodisch.*

Bemerkung 10.2.7. Sei $T\colon \Omega \to \Omega$ messbar. Die Menge der T-invarianten messbaren Mengen,

$$\mathcal{J}_T = \{A \in \mathcal{F}\colon T^{-1}(A) = A\}, \tag{10.2.1}$$

ist eine Teil-σ-Algebra von $\mathcal{F}$. Falls T ergodisch ist, so sagt man auch, dass $\mathcal{J}_T$ $\mathbb{P}$-*trivial* ist. Nach Lemma 7.2.10 ist in diesem Fall jede $\mathcal{J}_T$-messbare Abbildung fast sicher konstant. $\diamond$

Wenn der Shift-Operator θ auf dem Folgenraum $\mathbb{R}^{\mathbb{N}_0}$ ergodisch ist, nennt man den zugehörigen Prozess ebenfalls ergodisch. Ein wichtiges Beispiel sind Folgen von unabhängigen identisch verteilten Zufallsgrößen:

Lemma 10.2.8. *Es sei $\mathbb{X} = (X_n)_{n\in\mathbb{N}_0}$ eine Folge von unabhängigen identisch verteilten Zufallsgrößen auf $(\Omega, \mathcal{F}, \mathbb{P})$. Dann ist der Shift-Operator θ auf $(\mathbb{R}^{\mathbb{N}_0}, \mathcal{B}^{\otimes\mathbb{N}_0}, \mathbb{P} \circ \mathbb{X}^{-1})$ ergodisch.*

Beweis. Sei $A \in \mathcal{B}^{\otimes\mathbb{N}_0}$ shift-invariant. Dann gilt für jedes $n \in \mathbb{N}$ auch $A = (\theta^n)^{-1}(A) = \{(x_k)_{k\in\mathbb{N}_0}\colon (x_{n+k})_{k\in\mathbb{N}_0} \in A\}$. Also gilt

$$\mathbb{X}^{-1}(A) = \{\omega \in \Omega\colon (X_{n+k}(\omega))_{k\in\mathbb{N}_0} \in A\} \in \widehat{\mathcal{F}}_n = \bigvee_{k=n}^{\infty} X_k^{-1}(\mathcal{B}).$$

Da dies für jedes $n \in \mathbb{N}$ gilt, liegt $\mathbb{X}^{-1}(A)$ in der terminalen σ-Algebra $\mathcal{T}_\infty$ (siehe Definition 7.2.7), die ja nach dem Kolmogorov'schen 0-1-Gesetz (siehe Satz 7.2.8; hier geht die Unabhängigkeit ein) $\mathbb{P}$-trivial ist. Also folgt $\mathbb{P} \circ \mathbb{X}^{-1}(A) = \mathbb{P}(\mathbb{X}^{-1}(A)) \in \{0, 1\}$. □

Dieser Beweis zeigt insbesondere, dass $\mathcal{J}_\theta \subset \mathcal{T}_\infty$, d. h. jedes shift-invariante Ereignis ist ein Ereignis der infinitesimalen Zukunft. Allerdings ist im Allgemeinen $\mathcal{J}_\theta \neq \mathcal{T}_\infty$, und es gibt auch ergodische Maße, bezüglich denen $\mathcal{T}_\infty$ *nicht* trivial ist.[1]

Wir streifen kurz den Zusammenhang mit zwei wichtigen Mischungsbegriffen:

Definition 10.2.9 (mischend)**.** *Eine Maß erhaltende Transformation T auf $(\Omega, \mathcal{F}, \mathbb{P})$ heißt* mischend*, wenn für alle $A, B \in \mathcal{F}$ gilt:*

$$\lim_{n\to\infty} \mathbb{P}(A \cap T^{-n}(B)) = \mathbb{P}(A)\mathbb{P}(B), \tag{10.2.2}$$

und sie heißt schwach mischend*, wenn für alle $A, B \in \mathcal{F}$ gilt:*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \mathbb{P}(A \cap T^{-k}(B)) = \mathbb{P}(A)\mathbb{P}(B). \tag{10.2.3}$$

Im Fall des Shift-Operators $T = \theta$ auf dem Folgenraum kann man $A \cap T^{-n}(B)$ für großes n interpretieren als den Schnitt zweier zeitlich weit von einander entfernt liegenden Ereignisse, die also asymptotisch unabhängig werden, wenn T mischend ist. Die Begriffsbildung 'mischend' wird klar aus dem folgenden Beispiel.

Beispiel 10.2.10. Zum Zeitpunkt Null befindet sich in einem Gefäß, das den Raum $\Omega \subset \mathbb{R}^3$ einnehme, Tinte im Teil $A \subset \Omega$ des Gefäßes und Wasser im verbleibenden Teil $\Omega \setminus A$. Auf Ω betrachten wir das normierte Lebesgue-Maß $\mathbb{P}$. Die Abbildung $T\colon \Omega \to \Omega$ beschreibe einen Schritt eines Mischungsvorgangs, etwa ein einmaliges Umrühren. Diese Abbildung erhalte das Maß $\mathbb{P}$, also soll der Mischungsvorgang die Flüssigkeit insbesondere nirgends komprimieren oder auseinanderziehen. Für eine beliebige Menge $B \subset \Omega$ ist dann

$$A \cap T^{-n}(B) = \{x \in A\colon T^n(x) \in B\}$$

die Menge aller Tintenpartikel, die nach n-maligem Umrühren in der Menge B sind. Daher ist

$$\frac{\mathbb{P}(A \cap T^{-n}(B))}{\mathbb{P}(B)}$$

der relative Anteil der Tintenpartikel in B nach n-maligem Umrühren. Wenn T mischend ist, dann konvergiert dies gegen $\mathbb{P}(A)$, den Anteil der Tintenpartikel im Gefäß. ◇

Als Übungsaufgabe zeigt man leicht, dass unabhängige identisch verteilte Prozesse (wenn auf dem Folgenraum also ein Produktmaß betrachtet wird) mischend sind. Dieses Resultat ist aber auch ein Spezialfall von Satz 10.2.15 weiter unten. Eine andere Übungsaufgabe zeigt, dass Ergodizität äquivalent zur schwachen Mischungseigenschaft ist. Der Zusammenhang zur Mischungseigenschaft wird wie folgt geklärt:

[1]Ein Beispiel für diesen beiden Aussagen liefert $\mathbb{P} = \frac{1}{2}(\delta_{\omega^{(0)}} + \delta_{\omega^{(1)}})$ mit $\omega^{(0)} = (0, 1, 0, 1, \dots,)$ und $\omega^{(1)} = (1, 0, 1, 0, \dots) \in \{0, 1\}^{\mathbb{N}_0}$ (Übungsaufgabe).

Lemma 10.2.11. *Jede mischende Maß erhaltende Transformation T ist ergodisch.*

Beweis. Sei $A \in \mathcal{F}$ eine T-invariante Menge, dann gilt

$$\mathbb{P}(A) = \mathbb{P}(A \cap A) = \mathbb{P}(A \cap T^{-n}(A)) \to \mathbb{P}(A)\mathbb{P}(A) \qquad \text{für } n \to \infty.$$

Daraus folgt sofort $\mathbb{P}(A) = \mathbb{P}(A)^2$, also $\mathbb{P}(A) \in \{0, 1\}$. □

Als Übungsaufgabe zeigt man, dass der Drehoperator T_c aus Beispiel 10.2.5(b) für kein $c \in \mathbb{D}$ mischend ist. Auf Grund des folgenden Ergebnisses ist also Ergodizität nicht äquivalent zur Mischungseigenschaft:

Satz 10.2.12. *Der Drehoperator T_c aus Beispiel 10.2.5(b) ist genau dann ergodisch, wenn c keine Einheitswurzel ist.*

Als Vorbereitung benötigen wir das folgende maßtheoretische Resultat (Beweis als Übungsaufgabe oder siehe [Ba91, Satz 5.7]):

Lemma 10.2.13. *Es sei $\mathcal{A}$ eine Algebra mit $\mathcal{F} = \sigma(\mathcal{A})$. Dann existiert zu jedem $\varepsilon > 0$ und zu jedem $A \in \mathcal{F}$ ein $B \in \mathcal{A}$ mit $\mathbb{P}(A \triangle B) < \varepsilon$.*

Beweis von Satz 10.2.12. Falls c eine Einheitswurzel ist, also $c^n = 1$ für ein $n \in \mathbb{N}$, so ist T_c^n die identische Abbildung. Somit ist für jedes $A \in \mathcal{B}_{\mathbb{D}}$ die Menge $A \cup T_c^{-1}(A) \cup \cdots \cup T_c^{-n+1}(A)$ invariant unter T_c. Wir wählen ein $A \in \mathcal{B}_{\mathbb{D}}$ mit $0 < \lambda_{\mathbb{D}}(A) < \frac{1}{n}$ und erhalten

$$0 < \lambda_{\mathbb{D}}\big(A \cup T_c^{-1}(A) \cup \cdots \cup T_c^{-n+1}(A)\big) \leq \sum_{k=0}^{n-1} \lambda_{\mathbb{D}}(T_c^{-k}(A)) = n\lambda_{\mathbb{D}}(A) < 1,$$

also ist T_c nicht ergodisch.

Sei nun c keine Einheitswurzel. Zunächst machen wir uns klar, dass die Menge $\{c^n : n \in \mathbb{N}_0\}$ dicht in $\mathbb{D}$ liegt: Diese Folge hat zunächst (mindestens) einen Häufungspunkt ω_0 in $\mathbb{D}$. Sei $\varepsilon > 0$, und seien $m > n$ mit $|c^n - \omega_0| < \varepsilon$ und $|c^m - \omega_0| < \varepsilon$. Dann ist $0 < |c^{m-n} - 1| < 2\varepsilon$. Daraus folgt, dass zu jedem $\omega \in \mathbb{D}$ ein $k \in \mathbb{N}$ existiert mit $|\omega - c^{k(m-n)}| < 2\varepsilon$. Da $\varepsilon > 0$ beliebig ist, folgt die Dichtheit der Menge $\{c^n : n \in \mathbb{N}_0\}$.

Wir betrachten das Mengensystem $\mathcal{A}$, das aus allen endlichen Vereinigungen paarweise disjunkter Intervalle in $\mathbb{D}$ besteht. Dann ist $\mathcal{A}$ eine Algebra. Sei $A \in \mathcal{B}_{\mathbb{D}}$ invariant unter T_c mit $\lambda_{\mathbb{D}}(A) > 0$, und sei $0 < \varepsilon < \frac{1}{2}$. Nach Lemma 10.2.13 existiert ein $B \in \mathcal{A}$, also ein B von der Form $B = \bigcup_{i=1}^n I_i$ mit $n \in \mathbb{N}$ und disjunkten Intervallen $I_1, \ldots, I_n \subset \mathbb{D}$, so dass $\lambda_{\mathbb{D}}(A \triangle B) < \varepsilon\lambda_{\mathbb{D}}(A)$. Wir können annehmen, dass $\lambda_{\mathbb{D}}(I_i) < \varepsilon$ für jedes $i \in \{1, \ldots, n\}$.

Wir haben

$$\lambda_{\mathbb{D}}(A \triangle B) \leq \varepsilon\lambda_{\mathbb{D}}(A) \leq 2\varepsilon(1 - \varepsilon)\lambda_{\mathbb{D}}(A) \leq 2\varepsilon\big(\lambda_{\mathbb{D}}(A) - \lambda_{\mathbb{D}}(A \triangle B)\big) \leq 2\varepsilon\lambda_{\mathbb{D}}(B),$$

und daraus folgt

$$\sum_{i=1}^n \lambda_{\mathbb{D}}(A \cap I_i) = \lambda_{\mathbb{D}}(A \cap B) \geq \lambda_{\mathbb{D}}(B) - \lambda(A \triangle B) \geq (1 - 2\varepsilon)\lambda_{\mathbb{D}}(B) = (1 - 2\varepsilon) \sum_{i=1}^n \lambda_{\mathbb{D}}(I_i).$$

Mindestens eines der $I_1, \ldots, I_n$ – kurz mit I bezeichnet – erfüllt also die Ungleichung $\lambda_{\mathbb{D}}(A \cap I) \geq (1 - 2\varepsilon)\lambda_{\mathbb{D}}(I)$. Wir schreiben im Folgenden T statt T_c. Auf Grund der Dichtheit der

Menge $\{c^n\colon n \in \mathbb{N}_0\}$ existieren $k \in \mathbb{N}$ und $n_1, \dots, n_k \in \mathbb{N}$ der Art, dass die Intervalle $T^{-n_1}(I)$, $T^{-n_2}(I), \dots, T^{-n_k}(I)$ paarweise disjunkt sind und $\mathbb{D}$ bis auf eine Menge von kleinerem Maß als 2ε ausfüllen. Wegen der T-Invarianz von A und $\lambda_{\mathbb{D}}$ gilt für $j \in \{1, \dots, k\}$:

$$\begin{aligned}\lambda_{\mathbb{D}}(A \cap T^{-n_j}(I)) &= \lambda_{\mathbb{D}}(T^{-n_j}(A) \cap T^{-n_j}(I)) = \lambda_{\mathbb{D}} \circ T^{-n_j}(A \cap I)\\ &= \lambda_{\mathbb{D}}(A \cap I) \geq (1-2\varepsilon)\lambda_{\mathbb{D}}(I) = (1-2\varepsilon)\lambda_{\mathbb{D}}(T^{-n_j}(I)),\end{aligned}$$

und dies führt auf

$$\lambda_{\mathbb{D}}(A) \geq \sum_{j=1}^{k} \lambda_D(A \cap T^{-n_j}(I)) \geq (1-2\varepsilon) \sum_{j=1}^{k} \lambda_{\mathbb{D}}(T^{-n_j}(I)) \geq (1-2\varepsilon)^2.$$

Daraus folgt $\lambda_{\mathbb{D}}(A) = 1$. □

Hier ist eine Hilfestellung für den Nachweis der Eigenschaft des Mischens:

Lemma 10.2.14. *Sei $\mathcal{A}$ eine Algebra mit $\mathcal{F} = \sigma(\mathcal{A})$. Falls* (10.2.2) *für alle $A, B \in \mathcal{A}$ gilt, so ist T mischend.*

Der Beweis ist eine elementare Übungsaufgabe, man benutze Lemma 10.2.13.

Für das folgende Beispiel erinnern wir an Beispiel 10.2.5(d).

Satz 10.2.15. *Jede irreduzible aperiodische positiv rekurrente stationäre Markovkette ist mischend bezüglich des Shift-Operators.*

Beweis. Es sei $\mathbb{X} = (X_n)_{n\in\mathbb{N}_0}$ die betrachtete Markovkette, die wir also als eine Zufallsgröße mit Werten in dem messbaren Raum $(I^{\mathbb{N}_0}, \mathcal{P}(I)^{\otimes\mathbb{N}_0})$ auffassen. Wir betrachten die σ-Algebren $\mathcal{F}_n = \sigma(X_0, \dots, X_n)$ und die Algebra $\mathcal{A} = \bigcup_{n\in\mathbb{N}_0} \mathcal{F}_n$. Es ist leicht zu sehen, dass $\sigma(\mathcal{A}) = \mathcal{F}$. Nach Lemma 10.2.14 brauchen wir nur für alle $A, B \in \mathcal{A}$ die Gleichung (10.2.2) beweisen. Jedes $A \in \mathcal{A}$ ist von der Form $A = (\mathbb{X}^{(m)})^{-1}(C)$ mit $m \in \mathbb{N}_0$ und $C \subset I^{m+1}$, wobei $\mathbb{X}^{(n)} = (X_0, \dots, X_n)$. Also reicht es zu zeigen, dass für alle $m, k \in \mathbb{N}_0$ gilt:

$$\lim_{n\to\infty} \mathbb{P}(\mathbb{X}^{(m)} \in C, (\theta^n\mathbb{X})^{(k)} \in D) = \mathbb{P}(\mathbb{X}^{(m)} \in C)\mathbb{P}(\mathbb{X}^{(k)} \in D), \qquad C \subset I^{m+1}, D \subset I^{k+1}. \quad (10.2.4)$$

Zunächst machen wir uns klar, dass es hierbei reicht, (10.2.4) nur für endliche Mengen C und D zu zeigen. Für beliebiges $\varepsilon > 0$ und $D \subset I^{k+1}$ gibt es eine endliche Menge $D_0 \subset I^{k+1}$ mit $|\mathbb{P}(\mathbb{X}^{(k)} \in D) - \mathbb{P}(\mathbb{X}^{(k)} \in D_0)| < \varepsilon$. Also folgt für beliebiges $C \subset I^{m+1}$ mit Hilfe der Shift-Invarianz:

$$\begin{aligned}\big|\mathbb{P}(\mathbb{X}^{(m)} \in C, &(\theta^n\mathbb{X})^{(k)} \in D) - \mathbb{P}(\mathbb{X}^{(m)} \in C, (\theta^n\mathbb{X})^{(k)}) \in D_0)\big|\\ &\leq \big|\mathbb{P}((\theta^n\mathbb{X})^{(k)}) \in D) - \mathbb{P}((\theta^n\mathbb{X})^{(k)} \in D_0)\big|\\ &= |\mathbb{P}(\mathbb{X}^{(k)} \in D) - \mathbb{P}(\mathbb{X}^{(k)} \in D_0)|\\ &< \varepsilon.\end{aligned}$$

Ein analoges Argument gibt es für C. Also reicht es, (10.2.4) nur für endliche Mengen C und D zu zeigen.

Wir bezeichnen wie immer die Koeffizienten der zu Grunde liegenden stochastischen Matrix mit $p_{i,j}$. Mit ν bezeichnen wir die stationäre Startverteilung. Für endliche Mengen $C \subset I^{m+1}$ und $D \subset I^{k+1}$ und für $n > m$ gilt

$$\begin{aligned}&\mathbb{P}(\mathbb{X}^{(m)} \in C, (\theta^n\mathbb{X})^{(k)} \in D)\\ &\qquad = \sum_{(i_0,\dots,i_m)\in C} \sum_{(j_0,\dots,j_k)\in D} \nu(i_0)\Big(\prod_{s=1}^{m} p_{i_{s-1},i_s}\Big) p_{i_m,j_0}^{(n-m)} \Big(\prod_{t=1}^{k} p_{j_{t-1},j_t}\Big).\end{aligned}$$

Nach Satz 9.6.1 konvergiert der Term $p^{(n-m)}_{i_m,j_0}$ für $n \to \infty$ gegen $\nu(j_0)$, also (wegen der Endlichkeit der Summen) der gesamte Ausdruck gegen

$$\sum_{(i_0,\dots,i_m)\in C} \nu(i_0)\Big(\prod_{s=1}^{m} p_{i_{s-1},i_s}\Big) \sum_{(j_0,\dots,j_k)\in D} \nu(j_0)\Big(\prod_{t=1}^{k} p_{j_{t-1},j_t}\Big) = \mathbb{P}(\mathbb{X}^{(m)} \in C)\mathbb{P}(\mathbb{X}^{(k)} \in D).$$

□

Nun kommen wir zum Hauptergebnis dieses Abschnitts:

Satz 10.2.16 (Birkhoffs Ergodensatz). *Es sei T eine Maß erhaltende Transformation auf $(\Omega, \mathcal{F}, \mathbb{P})$ und $X \in \mathcal{L}^1(\mathbb{P})$. Dann konvergiert $S_n = \frac{1}{n}\sum_{i=0}^{n-1} X \circ T^i$ fast sicher für $n \to \infty$ gegen eine $\mathcal{J}_T$-messbare Zufallsgröße Y, für die $\mathbb{E}(X) = \mathbb{E}(Y)$ gilt.*

Beweis. Wir dürfen voraussetzen, dass X nichtnegativ ist, denn anderen Falls bemühen wir die Zerlegung $X = X^+ - X^-$. Sei also $X \geq 0$.

Wir betrachten die beiden Zufallsgrößen $\overline{X} = \limsup_{n\to\infty} S_n$ und $\underline{X} = \liminf_{n\to\infty} S_n$. Man sieht leicht, dass $\overline{X} \circ T = \overline{X}$ und $\underline{X} = \underline{X} \circ T$ gelten, also sind $\overline{X}$ und $\underline{X}$ messbar bezüglich $\mathcal{J}_T$. Der Beweis des Satzes wird durch den Beweis der Aussage

$$\int \overline{X}\,\mathrm{d}\mathbb{P} \leq \int X\,\mathrm{d}\mathbb{P} \leq \int \underline{X}\,\mathrm{d}\mathbb{P} \tag{10.2.5}$$

beendet werden, denn dann folgt wegen $\underline{X} \leq \overline{X}$ sogar Gleichheit $\mathbb{P}$-fast sicher nach Lemma 6.5.4(b), d. h. S_n ist $\mathbb{P}$-fast sicher sogar konvergent. Natürlich reicht es, nur die erste Ungleichung in (10.2.5) zu beweisen, die andere geht analog.

Die Grundidee des Beweises der ersten Ungleichung in (10.2.5) ist die folgende: Für gegebenes $\varepsilon > 0$ werden wir geeignete zufällige Zeitpunkte $0 = n_0 < n_1 < n_2 < \dots$ betrachten, an denen der durchschnittliche gemittelte Zuwachs dem Limes Superior bis auf ε nahe kommt, d. h.

$$X(T^{n_k(\omega)}(\omega)) + \dots + X(T^{n_{k+1}(\omega)-1}(\omega)) \geq [n_{k+1}(\omega) - n_k(\omega)]\overline{X}(\omega) - \varepsilon.$$

Dann summieren wir über alle k, so dass $n_{k+1} \leq n$, integrieren über $\omega \in \Omega$, teilen durch n, lassen $n \to \infty$ und zum Schluss $\varepsilon \downarrow 0$, und dies beendet den Beweis.

Dabei haben wir ein paar technische Probleme: Erstens kann $\overline{X}$ unbeschränkt sein oder sogar eventuell den Wert ∞ annehmen. Daher schneiden wir $\overline{X}$ ab: Für $M > 0$ sei $\overline{X}_M = \min\{\overline{X}, M\}$. Natürlich hat auch $\overline{X}_M$ die Eigenschaft $\overline{X}_M = \overline{X}_M \circ T$. Nun können wir einen wichtigen Baustein zur Definition der Folge der n_j einführen: Sei $\varepsilon > 0$, dann setzen wir

$$n(\omega) = \min\{k \in \mathbb{N} \colon S_k(\omega) \geq \overline{X}_M(\omega) - \varepsilon\}.$$

Die Abbildung $n \colon \Omega \to \mathbb{N}$ ist $\mathcal{F}$-$\mathcal{P}(\mathbb{N})$-messbar, denn für jedes $k \in \mathbb{N}$ gilt

$$\{\omega \in \Omega \colon n(\omega) = k\} = \bigcap_{j=1}^{k-1} \{S_j < \overline{X}_M - \varepsilon\} \cap \{S_k \geq \overline{X}_M - \varepsilon\}.$$

Die zweite technische Schwierigkeit besteht darin, dass die Abbildung $n \colon \Omega \to \mathbb{N}$ eventuell nicht beschränkt ist. Aber da $\lim_{N\to\infty} \mathbb{P}(\{\omega \colon n(\omega) > N\}) = 0$ ist, kann man ein $N_\varepsilon \in \mathbb{N}$ finden, so dass

$$\mathbb{P}(A) \leq \varepsilon, \qquad \text{wobei } A = \{\omega \colon n(\omega) > N_\varepsilon\}.$$

Nun stutzen wir auch $n(\omega)$ und $X(\omega)$ zurecht: Es sei

$$\widetilde{X}(\omega)=\begin{cases}X(\omega) & \text{für } \omega\in\Omega\setminus A,\\ M & \text{für } \omega\in A,\end{cases}\qquad\text{und}\qquad \widetilde{n}(\omega)=\begin{cases}n(\omega) & \text{für } \omega\in\Omega\setminus A,\\ 1 & \text{für } \omega\in A.\end{cases}$$

Falls $X(\omega)>M$, so ist $n(\omega)=1$ (wegen $S_1(\omega)=X(\omega)$), das heißt, es gilt $\omega\in A$. Also haben wir $X(\omega)\leq\widetilde{X}(\omega)$ für alle $\omega\in\Omega$. Für $\omega\in A$ gilt wegen $\widetilde{n}(\omega)=1$ und $\widetilde{X}(\omega)=M$ die Abschätzung

$$\frac{1}{\widetilde{n}(\omega)}\sum_{j=0}^{\widetilde{n}(\omega)-1}\widetilde{X}(T^j(\omega))\geq\overline{X}_M(\omega)-\varepsilon. \tag{10.2.6}$$

Diese Ungleichung gilt auch für $\omega\notin A$, denn dann ist $\widetilde{n}(\omega)=n(\omega)$, und wegen $X\leq\widetilde{X}$ ist die linke Seite von (10.2.6) nicht kleiner als $S_{n(\omega)}(\omega)$, also auch nicht kleiner als $\overline{X}_M(\omega)-\varepsilon$ nach Definition von $n(\omega)$. Wegen $\mathbb{P}(A)\leq\varepsilon$ gilt

$$\int\widetilde{X}\,\mathrm{d}\mathbb{P}=\int_{A^c}\widetilde{X}\,\mathrm{d}\mathbb{P}+\int_A\widetilde{X}\,\mathrm{d}\mathbb{P}\leq\int X\,\mathrm{d}\mathbb{P}+M\varepsilon. \tag{10.2.7}$$

Nun definieren wir rekursiv die Folge der n_j, die wir am Beginn des Beweises erwähnten:

$$n_0(\omega)=0\qquad\text{und}\qquad n_k(\omega)=n_{k-1}(\omega)+\widetilde{n}(T^{n_{k-1}(\omega)}(\omega)),\quad k\in\mathbb{N}.$$

Für $m\in\mathbb{N}$ sei $K_m(\omega)=\max\{k\in\mathbb{N}\colon n_k(\omega)\leq m\}$. Wegen $\widetilde{n}(\omega)\leq N_\varepsilon$ gilt $m-n_{K_m(\omega)}(\omega)\leq N_\varepsilon$. Es folgt

$$\begin{aligned}\sum_{j=0}^{m-1}\widetilde{X}(T^j(\omega))&\geq\sum_{j=0}^{n_{K_m(\omega)}(\omega)-1}\widetilde{X}(T^j(\omega))\\ &=\sum_{j=0}^{n_1(\omega)-1}\widetilde{X}(T^j(\omega))+\sum_{j=n_1(\omega)}^{n_2(\omega)-1}\widetilde{X}(T^j(\omega))+\dots+\sum_{j=n_{K_m(\omega)-1}(\omega)}^{n_{K_m(\omega)}(\omega)-1}\widetilde{X}(T^j(\omega)).\end{aligned}$$

Nun wenden wir die Ungleichung (10.2.6) nach einander an auf ω, $T^{n_1(\omega)}(\omega)$, $T^{n_2(\omega)}(\omega)$, $\dots$, $T^{n_{K_m(\omega)-1}(\omega)}(\omega)$ und erhalten aus der vorherigen Ungleichung

$$\begin{aligned}\sum_{j=0}^{m-1}\widetilde{X}(T^j(\omega))&\geq n_1(\omega)\big[\overline{X}_M(\omega)-\varepsilon\big]+(n_2(\omega)-n_1(\omega))\big[\overline{X}_M(T^{n_1(\omega)}(\omega))-\varepsilon\big]\\ &\quad+\dots+(n_{K_m(\omega)}(\omega)-n_{K_m(\omega)-1}(\omega))\big[\overline{X}_M(T^{n_{K_m(\omega)-1}(\omega)}(\omega))-\varepsilon\big]\\ &=n_{K_m(\omega)}(\omega)\big[\overline{X}_M(\omega)-\varepsilon\big]\\ &\geq m\overline{X}_M(\omega)+(n_{K_m(\omega)}-m)\overline{X}_M(\omega)-m\varepsilon\\ &\geq m\overline{X}_M(\omega)-N_\varepsilon M-m\varepsilon.\end{aligned}$$

wobei wir im zweiten Schritt ausnutzten, dass $\overline{X}_M=\overline{X}_M\circ T$.

Nun dividieren wir durch m, integrieren über $\omega\in\Omega$ bezüglich $\mathbb{P}$ und beachten, dass T Maß erhaltend ist, um zu erhalten:

$$\int\widetilde{X}\,\mathrm{d}\mathbb{P}=\int\frac{1}{m}\sum_{j=0}^{m-1}\widetilde{X}\circ T^j\,\mathrm{d}\mathbb{P}\geq\int\overline{X}_M\,\mathrm{d}\mathbb{P}-\frac{N_\varepsilon M}{m}-\varepsilon.$$

Wegen (10.2.7) haben wir

$$\int X \,\mathrm{d}\mathbb{P} \geq \int \overline{X}_M \,\mathrm{d}\mathbb{P} - \frac{N_\varepsilon M}{m} - \varepsilon - M\varepsilon.$$

Nun können wir $m \to \infty$ und $\varepsilon \to 0$ gehen lassen und erhalten $\int X \,\mathrm{d}\mathbb{P} \geq \int \overline{X}_M \,\mathrm{d}\mathbb{P}$ für jedes $M > 0$. Mit dem monotonen Konvergenzsatz (siehe Satz 6.8.1) erhalten wir die erste Ungleichung in (10.2.5). Damit ist der Beweis beendet. □

Bemerkung 10.2.17. (a) Die Konvergenz in Satz 10.2.16 gilt auch im $\mathcal{L}^1$-Sinn (Beweis als Übungsaufgabe oder in [Bi65]). Die $\mathcal{L}^1$-Konvergenz kann allerdings im Allgemeinen nicht mit Hilfe des majorisierten Konvergenzsatzes (siehe Satz 6.8.6) bewiesen werden.

(b) Falls T ergodisch ist, so ist die Zufallsgröße Y in Birkhoffs Ergodensatz nach Bemerkung 10.2.7 fast sicher konstant.

Im folgenden Lemma charakterisieren wir die Zufallsgröße Y aus dem Birkhoff'schen Ergodensatz als den bedingten Erwartungswert von X gegeben $\mathcal{J}_T$, siehe Abschnitt 7.3.

Lemma 10.2.18 (Charakterisierung des Grenzwerts)**.** *In der Situation von Satz 10.2.16 gilt* $Y = \mathbb{E}(X \mid \mathcal{J}_T)$.

Beweis. Dass Y messbar bezüglich $\mathcal{J}_T$ ist, haben wir schon in Satz 10.2.16 erwähnt. Sei nun $A \in \mathcal{J}_T$. Wir wollen zeigen, dass $\int_A X \,\mathrm{d}\mathbb{P} = \int_A Y \,\mathrm{d}\mathbb{P}$ gilt. Wir wenden Satz 10.2.16 auf $\mathbb{1}_A X$ an und erhalten, dass eine $\mathcal{J}_T$-messbare Zufallsgröße Y_A existiert mit

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} (\mathbb{1}_A X) \circ T^j = Y_A \qquad \text{fast sicher,}$$

und es gilt $\int Y_A \,\mathrm{d}\mathbb{P} = \int \mathbb{1}_A X \,\mathrm{d}\mathbb{P} = \int_A X \,\mathrm{d}\mathbb{P}$. Da A invariant unter T ist, haben wir andererseits aber auch

$$\frac{1}{n} \sum_{j=0}^{n-1} (\mathbb{1}_A X) \circ T^j = \mathbb{1}_A \frac{1}{n} \sum_{j=0}^{n-1} X \circ T^j, \qquad n \in \mathbb{N}.$$

Daher ist $Y_A = Y \mathbb{1}_A$ fast sicher. Daraus folgt $\int_A X \,\mathrm{d}\mathbb{P} = \int_A Y \,\mathrm{d}\mathbb{P}$, also besitzt Y auch die zweite definierende Eigenschaft von $\mathbb{E}(X \mid \mathcal{J}_T)$. □

Wir formulieren den Birkhoff'schen Ergodensatz im Konzept stationärer Prozesse; siehe die Diskussion nach Lemma 10.2.3. Wir erinnern daran, dass $\mathcal{J}_\theta$ die σ-Algebra der unter dem Shift-Operator θ invarianten messbaren Teilmengen des Folgenraums $\mathbb{R}^{\mathbb{N}_0}$ ist, siehe Bemerkung 10.2.7.

Satz 10.2.19 (Ergodensatz für stationäre Prozesse)**.** *Es sei* $\mathbb{X} = (X_n)_{n \in \mathbb{N}_0}$ *ein stationärer Prozess, so dass* X_n *integrierbar ist für jedes* n. *Dann gilt*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} X_j = \mathbb{E}[X_0 \mid \mathcal{J}_\theta] \qquad \textit{fast sicher.}$$

Beweis. Der Shift-Operator $\theta\colon \mathbb{R}^{\mathbb{N}_0} \to \mathbb{R}^{\mathbb{N}_0}$ ist Maß erhaltend für $\mathbb{P} \circ \mathbb{X}^{-1}$, und die Projektion $\pi_0\colon \mathbb{R}^{\mathbb{N}_0} \to \mathbb{R}$ auf den nullten Faktor liegt im $\mathcal{L}^1(\mathbb{P} \circ \mathbb{X}^{-1})$. Nach Satz 10.2.16, kombiniert mit

Lemma 10.2.18, konvergiert $\frac{1}{n}\sum_{j=0}^{n-1} X_j = \frac{1}{n}\sum_{j=0}^{n-1} \pi_0 \circ \theta^j$ fast sicher gegen den bedingten Erwartungswert von π_0 unter $\mathbb{P}\circ \mathbb{X}^{-1}$ gegeben $\mathcal{J}_\theta$. Da aber $(\pi_n)_{n\in\mathbb{N}_0}$ auf dem Raum $\mathbb{R}^{\mathbb{N}_0}$ die identische Abbildung ist, ist dies auch gleich $\mathbb{E}[X_0 \mid \mathcal{J}_\theta]$. □

Beispiel 10.2.20. Ein Spezialfall von Satz 10.2.19 ist der folgende: Für jede irreduzible, aperiodische, positiv rekurrente Markovkette $(X_n)_{n\in\mathbb{N}_0}$ auf einer abzählbaren Menge I mit invarianter Verteilung π und für jede π-integrierbare Abbildung $f\colon I \to \mathbb{R}$ gilt fast sicher

$$\lim_{n\to\infty} \frac{1}{n}\sum_{j=0}^{n-1} f(X_j) = \sum_{i\in I} f(i)\,\pi(i).$$

Diese Aussage folgt aus Satz 10.2.19 in dem Fall, dass die Markovkette mit π gestartet wird, also stationär ist, denn man sieht leicht mit Hilfe von Lemma 10.2.15, dass auch $(f(X_n))_{n\in\mathbb{N}_0}$ ergodisch ist. Aber sie gilt auch für jede beliebige Startverteilung, wie man mit Hilfe von Satz 9.6.1 als Übungsaufgabe zeigt.

Wendet man dies auf $f = \mathbb{1}_{\{i\}}$ mit einem $i \in I$ an, so erhält man, dass die sogenannten *Lokalzeiten* $\ell_n(i) = \sum_{j=0}^{n-1} \mathbb{1}_{\{X_j=i\}}$ erfüllen: $\lim_{n\to\infty} \frac{1}{n}\ell_n(i) = \pi(i)$ fast sicher, d. h. der mittlere Anteil der Treffer in i konvergiert fast sicher gegen die invariante Verteilung in i. ◇

Beispiel 10.2.21 (Range einer Irrfahrt)**.** Sei $(X_n)_{n\in\mathbb{N}_0}$ ein stationärer Prozess im $\mathbb{R}^d$, und sei durch $S_n = \sum_{i=1}^n X_i$ die Irrfahrt mit Schritten X_i definiert. Der *Range* der Irrfahrt nach n Schritten ist die Anzahl der besuchten Punkte, also $R_n = \#\{S_1,\dots,S_n\}$. Dann zeigt man als eine Übungsaufgabe, dass fast sicher gilt: $\lim_{n\to\infty} R_n/n = \mathbb{P}(A \mid J_\theta)$, wobei $A = \{S_n \neq 0$ für jedes $n \in \mathbb{N}_0\}$ das Fluchtereignis ist. ◇

Der Spezialfall von Satz 10.2.19 von unabhängigen identisch verteilten Zufallsgrößen ist berühmt:

Korollar 10.2.22 (Starkes Gesetz der Großen Zahlen)**.** *Es sei $(X_n)_{n\in\mathbb{N}_0}$ eine Folge von unabhängigen identisch verteilten integrierbaren Zufallsgrößen. Dann gilt*

$$\mathbb{P}\Big(\lim_{n\to\infty} \frac{1}{n}\sum_{j=0}^{n-1} X_j = \mathbb{E}(X_1)\Big) = 1.$$

Beweis. Nach Beispiel 10.2.5(a) ist $\mathbb{X} = (X_n)_{n\in\mathbb{N}_0}$ insbesondere ein stationärer Prozess, also konvergiert nach Satz 10.2.19 $\frac{1}{n}\sum_{i=0}^{n-1} X_i$ fast sicher gegen $\mathbb{E}[X_1 \mid \mathcal{J}_\theta]$. Da der Shift-Operator nach Lemma 10.2.8 ergodisch ist, ist dies nach Bemerkung 10.2.7 fast sicher konstant, also gleich $\mathbb{E}(X_1)$. □

Kapitel 11

Irrfahrten und die Brownsche Bewegung

Der wichtigste stochastische Prozess und eines der beziehungsreichsten und bedeutendsten Objekte der Wahrscheinlichkeitstheorie ist die *Brown'sche Bewegung*, ein stetiger Prozess mit Gauß'schen endlich-dimensionalen Verteilungen. Die überragende Bedeutung dieses Prozesses gründet sich unter Anderem darauf, dass er – ähnlich wie die Normalverteilung – als natürlicher universeller Grenzwert einer Vielzahl von stochastischen Prozessen auftaucht. In diesem Kapitel führen wir die Brown'sche Bewegung ein als Grenzverteilung einer geeignet reskalierten Irrfahrt auf $\mathbb{R}$. Zunächst betrachten wir in Abschnitt 11.1 eine spezielle Irrfahrt, die *einfache Irrfahrt*, ein wenig genauer und ermitteln die Verteilungen von ein paar Funktionalen ihres Pfades. Dann konstruieren wir in Abschnitt 11.2 die Brown'sche Bewegung mit Hilfe des Zentralen Grenzwertsatzes, des Straffheitskriteriums von Prohorov und des Kompaktheitskriteriums von Arzelà-Ascoli.

11.1 Die einfache Irrfahrt

In diesem Abschnitt betrachten wir ein grundlegendes stochastisches Modell ein wenig genauer: die *einfache Irrfahrt*, die wir als eine Markovkette schon in Beispiel 9.2.3 kennenlernten. Hier wollen wir mit Hilfe elementarer Kombinatorik ein paar explizite Rechnungen durchführen, um die Verteilung dieses Prozesses näher zu charakterisieren.

Sei eine Folge $X_1, X_2, \dots$ von unabhängigen symmetrisch verteilten $\{-1,1\}$-wertigen Zufallsgrößen gegeben, dann definiert $S_n = X_1 + X_2 + \dots + X_n$ die Irrfahrt $(S_n)_{n\in\mathbb{N}_0}$. Der Pfad $(S_0, S_1, \dots, S_n)$ ist also uniform verteilt auf der Menge

$$\Omega_n = \{(s_0, \dots, s_n) \in \mathbb{Z}^{n+1} \colon s_0 = 0 \text{ und } |s_i - s_{i-1}| = 1 \text{ für alle } i = 1, \dots, n\}.$$

Das zugehörige Wahrscheinlichkeitsmaß bezeichen wir mit $\mathbb{P}_n$.

Uns interessieren unter Anderem die Verteilungen der folgenden Funktionale des Pfades $(S_0, S_1, \dots, S_n)$, und zwar sowohl für festes n als auch asymptotisch für $n \to \infty$:

- der Endpunkt S_n,
- das Maximum von $S_0, S_1, \dots, S_n$,
- die Zeitdauer, die der Pfad in $\mathbb{N}$ verbringt,

- der Zeitpunkt der ersten Rückkehr zur Null.

Das wichtigste analytische Hilfsmittel bei den asymptotischen Aussagen wird die *Stirling'sche Formel* sein, die besagt:

$$n! \sim \sqrt{2\pi n}\Big(\frac{n}{e}\Big)^n, \qquad n \to \infty, \tag{11.1.1}$$

wobei wir wie immer $\sim$ für asymptotische Äquivalenz benutzen.

Wenden wir uns zunächst der Verteilung von S_n zu.

Lemma 11.1.1. *Für alle $n \in \mathbb{N}$ und $i \in \mathbb{Z}$ gilt*

$$\mathbb{P}_n(S_n = i) = \begin{cases} 0 & \textit{falls } n+i \textit{ ungerade oder } |i| > n, \\ 2^{-n}\binom{n}{(n+i)/2} & \textit{sonst.} \end{cases}$$

Beweis. Es ist klar, dass das Ereignis $\{S_n = i\}$ nicht eintreten kann, wenn $|i| > n$.

Da das Teilchen zu jedem Zeitpunkt um eine Einheit springen muss, kann es zu einem geraden Zeitpunkt nicht in einem ungeraden Raumpunkt sein und umgekehrt. Falls $n+i$ gerade ist, dann muss das Teilchen, um zum Zeitpunkt n in i zu sein, genau $(n+i)/2$ Mal aufwärts springen und genau $(n-i)/2$ Mal abwärts. Es gibt genau $\binom{n}{(n+i)/2}$ Pfade, die dies tun. □

Man beachte, dass die Wahrscheinlichkeiten $\mathbb{P}_{2n}(S_{2n} = 2i)$ in i steigend sind für $i \leq 0$ und fallend für $i \geq 0$. Insbesondere ist

$$u_{2n} = \mathbb{P}_{2n}(S_{2n} = 0)$$

der größte unter ihnen. Die Asymptotik von u_{2n} ermittelt man als Übungsaufgabe mit Hilfe von Lemma 11.1.1 und der Stirlingschen Formel in (11.1.1):

Korollar 11.1.2. *Es gilt*

$$u_{2n} = 2^{-2n}\binom{2n}{n} \sim \frac{1}{\sqrt{\pi n}}, \qquad n \to \infty.$$

Nun wenden wir uns der Verteilung des Minimums des Pfades zu, also der Zufallsgröße

$$M_n = \min\{S_0, \dots, S_n\}.$$

Eines der wichtigsten Hilfsmittel hierfür ist das *Spiegelungsprinzip*, das durch geschicktes Spiegeln eines Teils des Pfades einen Vergleich zwischen gewissen Pfadklassen herstellt. Im Folgenden bestimmen wir die Wahrscheinlichkeit der Menge der Pfade, die den Punkt $j \in -\mathbb{N}_0$ erreichen und nach insgesamt n Schritten in $i \geq j$ enden. Zur Veranschaulichung des folgenden Prinzips ist eine Skizze hilfreich.

Lemma 11.1.3 (Spiegelungsprinzip). *Für alle $n \in \mathbb{N}$ und alle $i, j \in \mathbb{Z}$ mit $j \leq 0$ und $i \geq j$ gilt $\mathbb{P}_n(M_n \leq j, S_n = i) = \mathbb{P}_n(S_n = i - 2j)$.*

Beweis. Wir brauchen nur den Fall zu betrachten, dass $n+i$ gerade ist, sonst sind beide betrachteten Ereignisse leer.

Für einen Pfad s in $\{M_n \leq j, S_n = i\}$ betrachten wir das kleinste $k \in \{1, \dots, n\}$ mit $s_k = j$, also den ersten Zeitpunkt, an dem das Teilchen den Wert j erreicht. Nun spiegeln wir das

Pfadstück $(s_k, \dots, s_n)$ an $s_k = j$ und erhalten einen Pfad $\tilde{s} = (\tilde{s}_0, \dots, \tilde{s}_n) \in \Omega_n$ mit $\tilde{s}_n = 2j - i$. Dieser Pfad liegt also in dem Ereignis $\{S_n = 2j - i\}$.

Nun überlegt man sich leicht, dass die oben durchgeführte Abbildung (Spiegeln ab dem ersten Erreichenszeitpunkt des Niveaus j) sogar eine bijektive Abbildung zwischen den Ereignissen $\{M_n \leq j, S_n = i\}$ und $\{S_n = 2j - i\}$ ist. Die Umkehrabbildung erhält man, indem man einen Pfad aus $\{S_n = 2j - i\}$ ab dem ersten Zeitpunkt, an dem er das Niveau j erreicht, spiegelt. (Er muss dieses Niveau spätestens zum Zeitpunkt n erreichen, da $2j - i \leq j \leq 0$.) Also enthalten die Mengen $\{M_n \leq j, S_n = i\}$ und $\{S_n = 2j - i\}$ die selbe Anzahl von Pfaden. Daraus folgt die Behauptung, denn $\mathbb{P}_n(M_n \leq j, S_n = i) = \mathbb{P}_n(S_n = 2j - i) = \mathbb{P}_n(S_n = i - 2j)$. □

Mit Hilfe des Spiegelungsprinzips können wir nun die gemeinsame Verteilung des Endpunkts S_n und des Minimums M_n bestimmen:

Satz 11.1.4 (Verteilung des Minimums des Pfades)**.** *Für alle $n \in \mathbb{N}$ und alle $i, j \in \mathbb{Z}$ mit $j \leq 0$ und $i \geq j$ gelten:*

$$\begin{aligned} \mathbb{P}_n(M_n = j, S_n = i) &= \mathbb{P}_n(S_n = i - 2j) - \mathbb{P}_n(S_n = i - 2j + 2), \\ \mathbb{P}_n(M_n = j) &= \mathbb{P}_n(S_n \in \{j, j-1\}). \end{aligned}$$

Beweis. Übungsaufgabe.

□

Aus Symmetriegründen erhält man aus Satz 11.1.4 natürlich auch die gemeinsame Verteilung des Endpunkts und des Maximums des Pfades.

Nun betrachten wir die Ereignisse, dass das Teilchen erst nach $2n$ Schritten zum Ursprung zurück kehrt bzw. nicht mehr bzw. nie den negativen Bereich betritt:

$$\begin{aligned} A_{2n} &= \big\{S_1 \neq 0, \dots, S_{2n-1} \neq 0, S_{2n} = 0\big\}, \\ B_{2n} &= \big\{S_i \neq 0 \text{ für alle } i \in \{1, \dots, 2n\}\big\}, \\ C_{2n} &= \big\{S_i \geq 0 \text{ für alle } i \in \{1, \dots, 2n\}\big\}. \end{aligned}$$

Wir erinnern daran, dass u_{2n} die Wahrscheinlichkeit dafür ist, dass das Teilchen zum Zeitpunkt $2n$ sich in Null befindet. Offensichtlich ist also $\mathbb{P}_{2n}(A_{2n}) \leq u_{2n}$. In den beiden Ereignissen B_{2n} und A_{2n} kann sich das Teilchen im Zeitintervall $\{1, \dots, 2n - 1\}$ entweder in $\mathbb{N}$ oder in $-\mathbb{N}$ aufhalten.

Lemma 11.1.5. *Für jedes $n \in \mathbb{N}$ gelten die Beziehungen*

$$\mathbb{P}_{2n}(A_{2n}) = \frac{1}{2n} u_{2n-2} = u_{2n-2} - u_{2n}, \tag{11.1.2}$$

$$\mathbb{P}_{2n}(B_{2n}) = u_{2n}, \tag{11.1.3}$$

$$\mathbb{P}_{2n}(C_{2n}) = u_{2n}. \tag{11.1.4}$$

Beweis. (11.1.2): Wir zählen die Pfade in A_{2n}, die im positiven Bereich verlaufen, und multiplizieren deren Anzahl mit 2. Ein solcher Pfad ist zu den Zeitpunkten 1 und $2n - 1$ in 1. Die Zahl der $(2n - 2)$-schrittigen in 1 startenden und endenden Pfade, die nie die Null betreten, ist gleich der Zahl der $(2n - 2)$-schrittigen in 1 startenden und endenden Pfade minus die Zahl solcher

Pfade, die zwischendurch die Null betreten. Diese beiden Anzahlen sind gleich der Anzahl der Pfade in dem Ereignis $\{S_{2n-2}=0\}$ bzw. in $\{M_{2n-2}\leq -1, S_{2n-2}=0\}$. Also haben wir

$$\begin{aligned}\mathbb{P}_{2n}(A_{2n}) &= 2\mathbb{P}_{2n}(S_1>0,\dots,S_{2n-1}>0,S_{2n}=0)\\ &= 2^{-2n+1}\big(|\{S_{2n-2}=0\}|-|\{M_{2n-2}\leq -1,S_{2n-2}=0\}|\big).\end{aligned}$$

Nach dem Spiegelungsprinzip ist der letzte Term gleich $|\{S_{2n-2}=2\}|$. Mit Hilfe von Lemma 11.1.1 können wir nun ausrechnen:

$$\mathbb{P}_{2n}(A_{2n}) = 2^{-2(n-1)}\frac{1}{2}\left(\binom{2n-2}{n-1}-\binom{2n-2}{n}\right) = \frac{1}{2n}u_{2n-2}.$$

Dies beweist die erste Gleichung in (11.1.2); die zweite rechnet man leicht nach.

(11.1.3): Das Gegenereignis von B_{2n} ist das Ereignis, dass das Teilchen zu einem der Zeitpunkte $2j$ mit $j\in\{1,\dots,n\}$ zum ersten Mal zurück zur Null kehrt, also die disjunkte Vereinigung der Ereignisse $\{S_1\neq 0,\dots,S_{2j-1}\neq 0,S_{2j}=0\}$. Man überlegt sich leicht, dass dieses Ereignis die Wahrscheinlichkeit $\mathbb{P}_{2j}(A_{2j})$ hat. Wenn wir (11.1.2) für j an Stelle von n anwenden, erhalten wir also, dass

$$\mathbb{P}_{2n}(B_{2n}) = 1-\sum_{j=1}^n \mathbb{P}_{2j}(A_{2j}) = 1-\sum_{j=1}^n (u_{2(j-1)}-u_{2j}) = u_{2n}.$$

(11.1.4): Übungsaufgabe. □

Bemerkung 11.1.6 (Rekurrenz und Nullrekurrenz)**.** Aus (11.1.3) in Kombination mit der Tatsache $\lim_{n\to\infty}u_{2n}=0$ (siehe Korollar 11.1.2) erhalten wir einen alternativen Beweis der Rekurrenz der einfachen Irrfahrt, siehe Satz 9.3.9. Wir betrachten den ersten Zeitpunkt einer Rückkehr zum Startpunkt:

$$T=\inf\{k\in\mathbb{N}\colon S_k=0\}\in\mathbb{N}_0\cup\{\infty\}.$$

Dann sehen wir, dass

$$\mathbb{P}(T<\infty)=\lim_{n\to\infty}\mathbb{P}(T\leq 2n)=\lim_{n\to\infty}\mathbb{P}(B_{2n}^{\mathrm{c}})=\lim_{n\to\infty}(1-u_{2n})=1,$$

also die Rekurrenz.

Eine weitere Folgerung aus Lemma 11.1.5 betrifft die erwartete Rückkehrzeit $\mathbb{E}(T)$ zum Ursprung, denn wegen $\{T=2n\}=A_{2n}$ liefert die erste Gleichung in (11.1.2), dass diese Reihe divergiert:

$$\mathbb{E}(T)=\sum_{n=1}^\infty 2n\mathbb{P}(A_{2n})=\sum_{n=1}^\infty 2n\frac{1}{2n}u_{2n-2}=\sum_{n=1}^\infty u_{2n-2},$$

und wegen Korollar 11.1.2 divergiert diese Reihe. Dies zeigt noch einmal die Nullrekurrenz, siehe Definition 9.5.4 und Beispiel 9.5.7. ◇

Nun bearbeiten wir die Frage, mit welcher Wahrscheinlichkeit das Teilchen sich eine gegebene Anzahl von Zeitpunkten im positiven Bereich aufhält. Wir betrachten also die Zufallsgröße, die die Zeit angibt, die der zufällige $2n$-schrittige Pfad im positiven Bereich verbringt:

$$Z_{2n}=2|\{i\in\{1,\dots,n\}\colon S_{2i-1}>0\}|.$$

Die Zufallsgröße Z_{2n} nimmt nur Werte in $\{0, 2, 4, \dots, 2n\}$ an. Das Ereignis $\{Z_{2n} = 0\}$ ist das Ereignis $\{S_i \leq 0$ für alle $i \in \{1, \dots, 2n\}\}$, und das Ereignis $\{Z_{2n} = 2n\}$ ist identisch mit $\{S_i \geq 0$ für alle $i \in \{1, \dots, 2n\}\}$, also mit C_{2n}. Nach (11.1.4) ist also $\mathbb{P}_{2n}(Z_{2n} = 0) = \mathbb{P}_{2n}(Z_{2n} = 2n\} = u_{2n}$. Außerdem ist die Abbildung $j \mapsto \mathbb{P}_{2n}(Z_{2n} = 2j)$ symmetrisch in dem Sinne, dass $\mathbb{P}_{2n}(Z_{2n} = 2j) = \mathbb{P}_{2n}(Z_{2n} = 2(n-j))$. Wir bestimmen nun die Verteilung von Z_{2n}:

Lemma 11.1.7. *Für jedes $n \in \mathbb{N}$ und alle $j \in \{0, \dots, n\}$ gilt $\mathbb{P}_{2n}(Z_{2n} = 2j) = u_{2j} u_{2(n-j)}$.*

Beweis. Wir führen eine Induktion nach n. Der Fall $n = 1$ ist klar.

Wir nehmen nun an, die Aussage trifft zu für alle $k \leq n-1$, und wir beweisen sie für n. Oben hatten wir schon darauf hin gewiesen, dass die Aussage für $j = 0$ und für $j = n$ zutrifft, also behandeln wir nur noch den Fall $1 \leq j \leq n-1$. Im Ereignis $\{Z_{2n} = 2j\}$ muss das Teilchen zu einem der Zeitpunkte $2, 4, 6, \dots, 2n-2$ in Null sein, und wir spalten auf nach dem ersten solchen Zeitpunkt und danach, ob der Pfad zuvor im positiven oder im negativen Bereich verläuft. Diesen Zeitpunkt nennen wir $2l$. Im ersten Fall (d. h., wenn der Pfad bis $2l$ in $\mathbb{N}$ verläuft) muss $l \leq j$ gelten, und nach dem Zeitpunkt $2l$ bleibt er genau $2(j-l)$ Zeitpunkte im positiven Bereich. Im zweiten Fall muss $l \leq n-j$ gelten, denn nach dem Zeitpunkt $2l$ muss der Pfad ja noch genau $2j$ Zeitpunkte im Positiven bleiben. Diese Überlegungen führen zu der Formel

$$\begin{aligned}\mathbb{P}_{2n}(Z_{2n} = 2j) = &\sum_{l=1}^{j} \mathbb{P}_{2l}(S_1 > 0, \dots, S_{2l-1} > 0, S_{2l} = 0) \mathbb{P}_{2(n-l)}(Z_{2(n-l)} = 2(j-l)) \\ &+ \sum_{l=1}^{n-j} \mathbb{P}_{2l}(S_1 < 0, \dots, S_{2l-1} < 0, S_{2l} = 0) \mathbb{P}_{2(n-l)}(Z_{2(n-l)} = 2j).\end{aligned}$$

Nun können wir die Induktionsvoraussetzung einsetzen und erhalten

$$\begin{aligned}\mathbb{P}_{2n}(Z_{2n} = 2j) = &\frac{1}{2} u_{2(n-j)} \sum_{l=1}^{j} \mathbb{P}_{2l}(S_1 \neq 0, \dots, S_{2l-1} \neq 0, S_{2l} = 0) u_{2(j-l)} \\ &+ \frac{1}{2} u_{2j} \sum_{l=1}^{n-j} \mathbb{P}_{2l}(S_1 \neq 0, \dots, S_{2l-1} \neq 0, S_{2l} = 0) u_{2(n-j-l)}.\end{aligned}$$

Nun beachte man, dass die erste Summe die Wahrscheinlichkeit der disjunkten Vereinigung über $l \in \{1, \dots, j\}$ der Ereignisse ist, dass der Pfad zum Zeitpunkt $2l$ zum ersten Mal in Null ist und dann zum Zeitpunkt $2j$ ebenfalls. Also ist die erste Summe gleich der Wahrscheinlichkeit des Ereignisses $\{S_{2j} = 0\}$, d. h. gleich u_{2j}. Analog ist die zweite Summe gleich $u_{2(n-j)}$. Damit ist der Beweis beendet. □

Aus Lemma 11.1.7 und Korollar 11.1.2 erhält man mit ein wenig Rechnung, dass

$$\frac{\mathbb{P}_{2n}(Z_{2n} = 2j)}{\mathbb{P}_{2n}(Z_{2n} = 2(j+1))} = 1 + \frac{n-1-2j}{(n-j)(2j+1)},$$

und der Quotient auf der rechten Seite ist positiv für $j \in [0, \frac{1}{2}(n-1)]$ und negativ für $j \in [\frac{1}{2}(n-1), n]$. Die Wahrscheinlichkeiten $\mathbb{P}_{2n}(Z_{2n} = 2j)$ fallen also in der linken Hälfte des Definitionsbereiches $\{0, \dots, n\}$ und steigen in der rechten. Sie sind also für $j = 0$ und $j = n$ am größten. Wenn also zwei gleich starke Tennisspieler eine Serie von Matches gegeneinander spielen, so ist es viel wahrscheinlicher, dass einer von ihnen die gesamte Zeit über führt, als dass die Dauer der Führung sich ausgleicht. Dies sieht man auch an den Asymptoten

$$\mathbb{P}_{2n}(Z_{2n} = 0) = \mathbb{P}_{2n}(Z_{2n} = 2n) \sim \frac{1}{\sqrt{\pi n}} \qquad \text{und} \qquad \mathbb{P}_{2n}(Z_{2n} = 2\lfloor n/2 \rfloor) \sim \frac{2}{\pi n},$$

das heißt, die Wahrscheinlichkeit für einen Ausgleich der Führungsdauern geht sogar doppelt so schnell gegen Null wie die Wahrscheinlichkeit für ständige Führung eines der Spieler.

Mit Hilfe der Asymptotik in Korollar 11.1.2 erhalten wir sogar einen sehr schönen Grenzwertsatz:

Satz 11.1.8 (Arcussinus-Gesetz). *Für alle* $0 < a < b < 1$ *gilt*

$$\lim_{n\to\infty} \mathbb{P}_{2n}\Big(a \le \frac{Z_{2n}}{2n} \le b\Big) = \frac{1}{\pi}\int_a^b \frac{1}{\sqrt{x(1-x)}}\,\mathrm{d}x = \frac{2}{\pi}\big(\arcsin\sqrt{b} - \arcsin\sqrt{a}\big).$$

Mit anderen Worten, $\frac{1}{2n}Z_{2n}$ *konvergiert in Verteilung gegen die* Arcussinus-Verteilung.

Beweisskizze. Es gilt für $n \to \infty$:

$$\begin{aligned}
\mathbb{P}_{2n}\Big(a \le \frac{Z_{2n}}{2n} \le b\Big) &\sim \sum_{j\approx an}^{\approx bn} \mathbb{P}_{2n}(Z_{2n} = 2j) = \sum_{j\approx an}^{\approx bn} u_{2j}u_{2(n-j)} \sim \sum_{j\approx an}^{\approx bn} \frac{1}{\sqrt{\pi j \pi (n-j)}} \\
&= \frac{1}{\pi}\frac{1}{n}\sum_{j\approx an}^{\approx bn} \frac{1}{\sqrt{\frac{j}{n}(1-\frac{j}{n})}},
\end{aligned}$$

wobei wir Randeffekte bei $j \approx an, bn$ vernachlässigten. Der letzte Ausdruck ist offensichtlich eine Riemannsumme für das Integral $\frac{1}{\pi}\int_a^b (x(1-x))^{-1/2}\,\mathrm{d}x$ für äquidistante Unterteilung in Intervalle der Länge $\frac{1}{n}$. Also konvergiert dieser Ausdruck gegen das Integral. □

Man nennt eine Zufallsgröße Z *Arcussinus-verteilt*, wenn für alle $0 < a < b < 1$ gilt:

$$\mathbb{P}(a \le Z \le b) = \frac{1}{\pi}\int_a^b \frac{1}{\sqrt{x(1-x)}}\,\mathrm{d}x = \frac{2}{\pi}\big(\arcsin\sqrt{b} - \arcsin\sqrt{a}\big).$$

11.2 Konstruktion der Brown'schen Bewegung

Wenn man den Graphen einer Irfahrt $(S_0, \dots, S_n)$ auf $\mathbb{Z}$ zu einem Polygonzug interpoliert, erhält man eine zufällige stetige Funktion $[0, n] \to \mathbb{R}$. Nach dem Zentralen Grenzwertsatz (Satz 8.3.19) ist zumindest der Endpunkt dieser Funktion für $n \to \infty$ von der Größenordnung $\sqrt{n}$. Also kann man auf die Idee kommen, diese Funktion zeitlich mit n und räumlich mit $\sqrt{n}$ zu reskalieren und nach dem Verhalten dieser Funktion $[0, 1] \to \mathbb{R}$ zu fragen. Wir betrachten also die stückweise lineare, zufällige Funktion

$$B_t^{(n)} = \frac{1}{\sqrt{n}}\Big(S_{\lfloor tn \rfloor} + \big(tn - \lfloor tn \rfloor\big)X_{\lfloor tn \rfloor + 1}\Big), \qquad t \in [0, 1], \tag{11.2.1}$$

wobei die X_i die Schritte der Irrfahrt sind, also $S_n = X_1 + \dots + X_n$. Es sei $\mathcal{C}_0 = \mathcal{C}_0([0, 1])$ die Menge der stetigen Funktionen $[0, 1] \to \mathbb{R}$, die in Null starten. Daher liegt $B^{(n)} = (B_t^{(n)})_{t\in[0,1]}$

in $\mathcal{C}_0$. Wir fassen $\mathcal{C}_0$, zusammen mit der Supremumsmetrik, als einen polnischen Raum auf. Wie man sich als eine Übungsaufgabe überlegt, gilt dann für die zugehörige Borel-σ-Algebra $\mathcal{B}_{\mathcal{C}_0}$:

$$\mathcal{B}_{\mathcal{C}_0} = \bigvee_{t\in[0,1]} \pi_t^{-1}(\mathcal{B}_\mathbb{R}) = \sigma(\pi_t : t \in [0,1]), \tag{11.2.2}$$

wobei π_t die eindimensionalen Projektionen $f \mapsto f(t)$ sind. Offensichtlich ist $B_t^{(n)}$ für jedes $t \in [0,1]$ eine messbare Zufallsgröße. Wegen (11.2.2) ist $B^{(n)}$ messbar bezüglich $\mathcal{B}_{\mathcal{C}_0}$, also eine $\mathcal{C}_0$-wertige Zufallsgröße.

Wie oben angedeutet, konvergiert der Endpunkt $B_1^{(n)}$ dieser Funktion in Verteilung gegen die Standardnormalverteilung $\mathcal{N}$. Hierfür benötigt man nur, dass die Schritte $X_1, X_2, \dots$ unabhängige, identisch verteilte standardisierte Zufallsgrößen sind, also $\mathbb{E}[X_i] = 0$ und $\mathbb{E}[X_i^2] = 1$ erfüllen, und dies wollen wir in diesem Abschnitt voraussetzen.

Aus dem Zentralen Grenzwertsatz kann man allerdings noch viel mehr herausholen als nur die Konvergenz von $B_1^{(n)} = S_n/\sqrt{n}$ gegen $\mathcal{N}$:

Lemma 11.2.1. *Sei $S_n = X_1 + \dots + X_n$ eine Irrfahrt auf $\mathbb{R}$ mit unabhängigen, identisch verteilten standardisierten Schritten $X_1, X_2, \dots$. Seien $m \in \mathbb{N}$ und $0 < t_1 < t_2 < \dots < t_m \leq 1$, dann konvergiert der Vektor $(B_{t_1}^{(n)}, B_{t_2}^{(n)}, \dots, B_{t_m}^{(n)})$ in Verteilung gegen die m-dimensionale Normalverteilung mit Erwartungswertvektor Null und Kovarianzmatrix $(\min\{t_i, t_j\})_{i,j=1,\dots,m}$. In anderen Worten, die gemeinsame Verteilung der Zufallsgrößen*

$$B_{t_1}^{(n)} - B_0^{(n)}, B_{t_2}^{(n)} - B_{t_1}^{(n)}, \dots, B_{t_m}^{(n)} - B_{t_{m-1}}^{(n)} \tag{11.2.3}$$

konvergiert gegen m unabhängige Normalverteilungen mit Erwartungswerten Null und Varianzen t_1, $t_2 - t_1, \dots, t_m - t_{m-1}$.

Beweisskizze. Um die technischen Details gering zu halten, zeigen wir die letzte der beiden Aussagen nur für den Fall, dass die t_i ersetzt werden durch $t_i^{(n)} \in \frac{1}{n}\mathbb{N}_0$, die gegen t_i konvergieren (mit $t_0^{(n)} = 0$). Dann nämlich sind die Zufallsgrößen in (11.2.3) exakt gleich den Zufallsgrößen

$$\frac{1}{\sqrt{n}} \sum_{k=nt_{i-1}^{(n)}+1}^{nt_i^{(n)}} X_k, \qquad i = 1, \dots, m;$$

man beachte, dass die Summationsgrenzen nun natürliche Zahlen sind. Diese Zufallsgrößen sind offensichtlich unabhängig, und ihre Limesverteilung ist jeweils $\mathcal{N}(0, t_i - t_{i-1})$, wie aus dem Zentralen Grenzwertsatz folgt. Dies zeigt also die zweite Aussage.

Die erste erhält man zum Beispiel durch folgendes Argument. Die gemeinsame Grenzverteilung der Größen in (11.2.3) ist eine m-dimensionale Normalverteilung, und den Vektor $(B_{t_1}^{(n)}, B_{t_2}^{(n)}, \dots, B_{t_m}^{(n)})$ erhält man durch eine gewisse explizite bijektive lineare Abbildung aus dem Vektor der Größen in (11.2.3). Also konvergiert er ebenfalls gegen eine m-dimensionale Normalverteilung, die das Bildmaß von

$$\mathcal{N}(0, \mathrm{Diag}((t_i - t_{i-1})_{i=1}^m) = \bigotimes_{i=1}^m \mathcal{N}(0, t_i - t_{i-1})$$

unter dieser Abbildung ist, die man ja explizit ausrechnen kann, was wir hier nicht machen wollen.

Statt dessen benutzen wir, dass zwei zentrierte Normalverteilungen schon dann übereinstimmen, wenn ihre Kovarianzmatrizen übereinstimmen, und berechnen einfach nur die Kovarianzen. Wir bezeichnen den Grenzzufallsvektor mit $(B_{t_1}, B_{t_2}, \dots, B_{t_m})$, dann gilt für alle $i, j \in \{1, \dots, m\}$ mit $i < j$:

$$\begin{aligned}\operatorname{cov}(B_{t_i}, B_{t_j}) &= \mathbb{E}\big[B_{t_i} B_{t_j}\big] = \mathbb{E}\big[B_{t_i} B_{t_{j-1}}\big] + \mathbb{E}\big[B_{t_i}(B_{t_j} - B_{t_{j-1}})\big] \\ &= \mathbb{E}\big[B_{t_i} B_{t_{j-1}}\big] + \sum_{l=1}^{i} \mathbb{E}\big[(B_{t_l} - B_{t_{l-1}})(B_{t_j} - B_{t_{j-1}})\big] \\ &= \mathbb{E}\big[B_{t_i} B_{t_{j-1}}\big],\end{aligned}$$

denn wir schrieben B_{t_i} als eine Teleskopsumme der $B_{t_l} - B_{t_{l-1}}$ und nutzten aus, dass diese von $B_{t_j} - B_{t_{j-1}}$ unabhängig sind und zentriert. Falls auch noch $i < j - 1$ gilt, können wir die selbe Rechnung anwenden und erhalten, dass $\operatorname{cov}(B_{t_i}, B_{t_j}) = \mathbb{E}[B_{t_i} B_{t_{j-2}}]$ gilt. Iteration ergibt, dass $\operatorname{cov}(B_{t_i}, B_{t_j}) = \mathbb{E}[B_{t_i} B_{t_i}] = \mathbb{V}(B_{t_i})$ gilt. Um diese Varianz zu bestimmen, schreiben wir wieder B_{t_i} als eine Teleskopsumme der $B_{t_l} - B_{t_{l-1}}$ und sehen leicht, dass $\mathbb{V}(B_{t_i}) = t_i$ gilt, denn die Varianzen addieren sich. Dies zeigt, dass die Kovarianzmatrix von $(B_{t_1}, B_{t_2}, \dots, B_{t_m})$ gerade die angegebene ist. □

Also wissen wir schon, dass die endlich-dimensionalen Verteilungen des Prozesses $B^{(n)}$ konvergieren, und zwar gegen einen gewissen *Gauß'schen Prozess*:

Definition 11.2.2 (Brown'sche Bewegung)**.** *Eine* Brown'sche Bewegung *ist ein reellwertiger stochastischer Prozess $B = (B_t)_{t\in[0,1]}$ in stetiger Zeit mit den folgenden drei charakteristischen Eigenschaften:*

(i) $B_0 = 0$ fast sicher,

(ii) Für jedes $m \in \mathbb{N}$ und jede $0 \le t_0 < t_1 < \dots < t_m \le 1$ ist der Vektor $(B(t_m) - B(t_{m-1}), B(t_{m-1}) - B(t_{m-2}), \dots, B(t_1) - B(t_0))$ normalverteilt mit Erwartungswertvektor 0 und Kovarianzmatrix $\operatorname{Diag}(t_m - t_{m-1}, t_{m-1} - t_{m-2}, \dots, t_1 - t_0)$,

(iii) Die Pfade $t \mapsto B_t$ sind fast sicher stetig.

Das berühmte *Wiener-Maß* ist die Verteilung einer Brown'schen Bewegung:

Definition 11.2.3 (Wiener-Maß)**.** *Sei B eine Brown'sche Bewegung, also eine Zufallsgröße mit Werten in $\mathcal{C}_0([0,1])$, definiert auf einem geeigneten Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, \mathbb{P})$. Dann heißt das Bildmaß $\mathbb{P} \circ B^{-1}$ auf $\mathcal{C}_0([0,1])$ das* Wiener-Maß.

Meist definiert man eine Brown'sche Bewegung und das Wiener-Maß auf dem Zeitintervall $[0, \infty)$ an Stelle von $[0, 1]$, aber das soll uns hier nicht kümmern. Die Brown'sche Bewegung und das Wiener-Maß sind die bedeutungsvollsten und beziehungsreichsten Objekte der Wahrscheinlichkeitstheorie, und eine enorme Literatur über sie hat sich in den letzten Jahrzehnten angesammelt. Mehr über die Historie, Bedeutung, Konstruktionen und Eigenschaften der Brown'schen Bewegung findet man in der Vorlesung *Stochastische Prozesse.*

Zurück zu unserem zufälligen Polygonzug. Falls $(B^{(n)})_{n\in\mathbb{N}}$ in Verteilung konvergiert, so konvergieren alle endlich-dimensionalen Verteilungen ebenfalls, denn für jedes $m \in \mathbb{N}_0$ und jede

$0 \leq t_0 < t_1 < \cdots < t_m \leq 1$ ist die Abbildung $f \mapsto (f(t_0), \dots, f(t_m))$ stetig von $\mathcal{C}_0([0,1])$ nach $\mathbb{R}^{m+1}$; siehe das *Continuous mapping theorem* in Lemma 8.2.11. Der einzige mögliche Häufungspunkt (im Sinne der Verteilungskonvergenz) von $(B^{(n)})_{n\in\mathbb{N}}$ ist also eine Brown'sche Bewegung. Die Familie der Abbildungen $f \mapsto (f(t_0), \dots, f(t_m))$ mit $m \in \mathbb{N}_0$ und $0 \leq t_0 < t_1 < \cdots < t_m \leq 1$ ist sogar eine trennende Familie (siehe Definition 8.2.25), denn nach einer zeitlich stetigen Variante von Lemma 10.1.4 (die wir hier nicht näher ausformulieren oder beweisen) bestimmen die endlich-dimensionalen Verteilungen den Prozess eindeutig. Nach Satz 8.2.26 konvergiert die Folge $(B^{(n)})_{n\in\mathbb{N}}$ genau dann gegen eine Brown'sche Bewegung, wenn sie straff ist. Mit anderen Worten, die Straffheit der Folge $(B^{(n)})_{n\in\mathbb{N}}$ sichert die Existenz einer Brown'schen Bewegung. Dies ist der Inhalt des folgenden Satzes.

Satz 11.2.4 (Donsker'sches Invarianzprinzip)**.** *Sei $S_n = X_1 + \cdots + X_n$ eine Irrfahrt auf $\mathbb{R}$ mit unabhängigen, identisch verteilten standardisierten Schritten $X_1, X_2, \dots$. Dann ist die Folge $(B^{(n)})_{n\in\mathbb{N}}$ von $\mathcal{C}_0([0,1])$-wertigen Zufallsvariablen straff. Insbesondere konvergiert $(B^{(n)})_{n\in\mathbb{N}}$ in Verteilung gegen eine Brown'sche Bewegung.*

Man nennt den Satz 11.2.4 ein *Invarianzprinzip* oder einen *funktionalen Grenzwertsatz*. Der erste Begriff betont, dass der Grenzprozess invariant ist unter der zugrunde liegenden Schrittverteilung der Irrfahrt (solange sie standardisiert ist), und der zweite weist darauf hin, dass die Aussage eine Ausweitung des Zentralen Grenzwertsatzes auf zufällige Funktionen ist. Satz 11.2.4 liefert eine Konstruktion der Brown'schen Bewegung sowie eine Approximationsaussage mit reskalierten Irrfahrten. Letztere werden wir in Lemmas 11.2.7 und 11.2.8 einsetzen, um die Verteilung gewisser Funktionale des Brown'schen Pfades zu ermitteln; es werden Ergebnisse aus Abschnitt 11.1 hilfreich sein.

Im Laufe des Beweises von Satz 11.2.4 muss die Kompaktheit gewisser Teilmengen von $\mathcal{C}_0([0,1])$ gezeigt werden. Das geeignete Kriterium liefert der folgende berühmte Satz aus der Funktionalanalysis. Wir führen die folgende Maßgröße für die Stetigkeit einer Funktion ein:

$$w_\delta(f) = \sup\big\{|f(s) - f(t)| \colon s,t \in [0,1], |s-t| \leq \delta\big\}, \qquad f \colon [0,1] \to \mathbb{R}, \delta > 0. \tag{11.2.4}$$

Eine Funktion $f \colon [0,1] \to \mathbb{R}$ ist also genau dann gleichmäßig stetig, wenn $\lim_{\delta\downarrow 0} w_\delta(f) = 0$ ist.

Satz 11.2.5 (Arzelà-Ascoli)**.** *Eine Teilmenge K von $\mathcal{C}([0,1])$ hat genau dann einen kompakten Abschluss, wenn gelten:*

$$(i) \quad \sup_{f\in K} |f(0)| < \infty \qquad \textit{und} \qquad (ii) \quad \lim_{\delta\downarrow 0} \sup_{f\in K} w_\delta(f) = 0.$$

Beweise finden sich in vielen Lehrbüchern der Funktionalanalysis. Wenn wir nun die Straffheit der Folge $(B^{(n)})_{n\in\mathbb{N}}$ zeigen wollen, so ist nur die Eigenschaft (ii), die *gleichmäßige gleichgradige Stetigkeit* von K, zu beachten, denn die $B^{(n)}$ starten alle in Null. Die Aufgabe lautet also, zu jedem $\varepsilon > 0$ eine kompakte Menge $K \subset \mathcal{C}_0([0,1])$ zu finden, so dass $\mathbb{P}(B^{(n)} \in K^{\mathrm{c}}) < \varepsilon$ für jedes $n \in \mathbb{N}$ gilt.

Wir geben nun einen Beweis von Satz 11.2.4. Zunächst passen wir die Aussage des Satzes von Arzelà-Ascoli auf die Straffheit von $(B^{(n)})_{n\in\mathbb{N}}$ an:

Lemma 11.2.6. *Sei $(\mu_n)_{n\in\mathbb{N}}$ eine Folge von Wahrscheinlichkeitsmaßen auf $\mathcal{C}_0([0,1])$. Falls gilt:*

$$\lim_{\delta\downarrow 0}\limsup_{n\to\infty}\sup_{t\in[0,1-\delta]}\frac{1}{\delta}\,\mu_n\Big(\Big\{f\in\mathcal{C}_0\colon \sup_{s\in[t,t+\delta]}|f(t)-f(s)|\geq\eta\Big\}\Big)=0,\qquad \eta>0, \tag{11.2.5}$$

so ist $(\mu_n)_{n\in\mathbb{N}}$ straff.

Beweis. Wir machen uns zunächst klar, dass es genügt, für jedes $\varepsilon,\eta>0$ ein $\delta>0$ und ein $n_0\in\mathbb{N}$ zu finden mit

$$\mu_n(\{f\colon w_\delta(f)>\eta\})\leq\varepsilon,\qquad n\geq n_0. \tag{11.2.6}$$

Dies wenden wir nämlich für gegebenes $\varepsilon>0$ und jedes $k\in\mathbb{N}$ an auf ε ersetzt durch $\varepsilon 2^{-k-1}$ und η ersetzt durch $1/k$ und erhalten ein $\delta_k>0$ und $n_0(k)\in\mathbb{N}$ mit $\mu_n(\{f\colon w_{\delta_k}(f)>1/k\})\leq\varepsilon 2^{-k-1}$ für jedes $n\geq n_0(k)$. Da $\lim_{\delta\downarrow 0}w_\delta(f)=0$ für jedes f gilt, kann man nach eventueller Verkleinerung von δ_k davon ausgehen, dass die Ungleichung $\mu_n(\{f\colon w_{\delta_k}(f)>1/k\})\leq\varepsilon 2^{-k-1}$ sogar für jedes $n\in\mathbb{N}$ gilt. Nach dem Satz 11.2.5 von Arzelà-Ascoli hat die Menge

$$K=K_\varepsilon=\bigcap_{k\in\mathbb{N}}\Big\{f\colon w_{\delta_k}(f)\leq\frac{1}{k}\Big\}$$

kompakten Abschluss, und es gilt

$$\mu_n(\overline{K}^{\mathrm{c}})\leq\mu_n(K^{\mathrm{c}})\leq\sum_{k\in\mathbb{N}}\mu_n\Big(\Big\{f\colon w_{\delta_k}(f)>\frac{1}{k}\Big\}\Big)\leq\sum_{k\in\mathbb{N}}\varepsilon 2^{-k-1}\leq\varepsilon$$

für alle $n\in\mathbb{N}$. Also folgt die Straffheit von $(\mu_n)_{n\in\mathbb{N}}$.

Seien also $\varepsilon,\eta>0$ vorgegeben. Auf Grund der Voraussetzung in (11.2.5) haben wir ein $\delta_0>0$ und ein $n_0\in\mathbb{N}$ mit

$$\frac{1}{\delta_0}\,\mu_n\Big(\Big\{f\in\mathcal{C}_0\colon \sup_{s\in[t,t+\delta_0]}|f(t)-f(s)|\geq\frac{\eta}{2}\Big\}\Big)\leq\frac{\varepsilon}{3},\qquad n\geq n_0, t\in[0,1-\delta_0].$$

Sei $m\in\mathbb{N}$ minimal mit $1/m<\delta_0$. Wir setzen $\delta=\frac{1}{2m}$. Mit diesem δ und dem soeben gewählten n_0 zeigen wir nun, dass (11.2.6) erfüllt ist.

Für jedes $f\in\mathcal{C}_0$ mit $w_\delta(f)\geq\eta$ gibt es $t,s\in[0,1]$ mit $t<s$ und $|f(t)-f(s)|\geq\eta$ und $|t-s|<\delta$. Weiterhin existiert ein $k\in\mathbb{N}_0$ mit $k\leq 2m-2$ und $\frac{k}{2m}\leq t<s\leq\frac{k}{2m}+2\delta=\frac{k}{2m}+\frac{1}{m}$. Dann ist $|f(t)-f(\frac{k}{2m})|\geq\frac{\eta}{2}$ oder $|f(s)-f(\frac{k}{2m})|\geq\frac{\eta}{2}$. Aus diesen Überlegungen folgt die Inklusion

$$\{f\colon w_\delta(f)\geq\eta\}\subset\bigcup_{k=0}^{2m-2}\Big\{f\colon \sup_{s\in[\frac{k}{2m},\frac{k}{2m}+\frac{1}{m}]}\Big|f(s)-f\Big(\frac{k}{2m}\Big)\Big|\geq\frac{\eta}{2}\Big\}.$$

Also folgt für jedes $n\geq n_0$:

$$\mu_n(\{f\colon w_\delta(f)\geq\eta\})\leq(2m-1)\delta_0\frac{\varepsilon}{3}\leq(2+\delta_0)\frac{\varepsilon}{3}\leq\varepsilon.$$

Dies zeigt, dass (11.2.6) erfüllt ist, und beendet den Beweis. □

Beweis von Satz 11.2.4. Wir müssen also nur das Kriterium in (11.2.5) für die Verteilungen $\mu_n=\mathbb{P}\circ(B^{(n)})^{-1}$ der Polygonzüge $B^{(n)}$ verifizieren. Seien $\eta,\delta>0$ und $t\in[0,1-\delta]$. Für unsere Wahl von μ_n ist natürlich

$$\mu_n\Big(\Big\{f\in\mathcal{C}_0\colon \sup_{s\in[t,t+\delta]}|f(t)-f(s)|\geq\eta\Big\}\Big)=\mathbb{P}\Big(\sup_{s\in[t,t+\delta]}|B^{(n)}(t)-B^{(n)}(s)|\geq\eta\Big).$$

Wir möchten das Supremum in Termen der Irrfahrt ausdrücken und zeigen nun, dass gilt:

$$\mathbb{P}\Big(\sup_{s\in[t,t+\delta]}|B^{(n)}(t)-B^{(n)}(s)|\geq\eta\Big)\leq\mathbb{P}\Big(\frac{1}{\sqrt{n}}\max_{i=1}^{\lfloor 3n\delta\rfloor}|S_i|\geq\frac{\eta}{2}\Big) \qquad (11.2.7)$$

Da $B^{(n)}$ ein Polygonzug ist, ist für $t,s\in\frac{1}{n}\mathbb{N}_0$ und $t<s$ einfach

$$\sup_{s\in[t,t+\delta]}|B^{(n)}(t)-B^{(n)}(s)|=\frac{1}{\sqrt{n}}\max_{i=0}^{(s-t)n}|S_{\lfloor nt\rfloor+i}-S_{\lfloor nt\rfloor}|.$$

Im Allgemeinen gibt es $j,k\in\{0,\dots,n\}$ mit $k<j$ und $\frac{k}{n}\leq t<\frac{k+1}{n}$ sowie $\frac{j-1}{n}<t+\delta\leq\frac{j}{n}$. Dann gilt für jedes $s\in[t,t+\delta]$:

$$\begin{aligned}|B^{(n)}(t)-B^{(n)}(s)|&\leq\Big|B^{(n)}(t)-B^{(n)}\Big(\frac{k}{n}\Big)\Big|+\max_{i=1}^{j-k}\Big|B^{(n)}\Big(\frac{k+i}{n}\Big)-B^{(n)}\Big(\frac{k}{n}\Big)\Big|\\&\leq 2\max_{i=1}^{j-k}\Big|B^{(n)}\Big(\frac{k+i}{n}\Big)-B^{(n)}\Big(\frac{k}{n}\Big)\Big|.\end{aligned}$$

Also folgt

$$\sup_{s\in[t,t+\delta]}|B^{(n)}(t)-B^{(n)}(s)|\leq 2\max_{i=1}^{j-k}\Big|B^{(n)}\Big(\frac{k+i}{n}\Big)-B^{(n)}\Big(\frac{k}{n}\Big)\Big|=\frac{2}{\sqrt{n}}\max_{i=1}^{j-k}|S_{k+i}-S_k|.$$

Wegen $j-k-2\leq n\delta$ gilt $j-k\leq 3n\delta$ für alle genügend großen n, also können wir die rechte Seite gegen $2\max_{i=1}^{\lfloor 3n\delta\rfloor}|S_{k+i}-S_k|/\sqrt{n}$ abschätzen. Die Verteilung dieser Zufallsgröße hängt aber gar nicht von k ab. Also ist (11.2.7) bewiesen.

In (11.2.7) substituieren wir nun $m=\lfloor 3n\delta\rfloor$, also müssen wir zum Beweis von (11.2.5) nur noch zeigen:

$$\lim_{\delta\downarrow 0}\frac{1}{\delta}\limsup_{m\to\infty}\mathbb{P}\Big(\max_{i=1}^{m}|S_i|\geq\frac{\eta}{\sqrt{\delta}}\sqrt{m}\Big)=0,\qquad \eta>0. \qquad (11.2.8)$$

Den Beweis von (11.2.8) werden wir auf eine Anwendung des Zentralen Grenzwertsatzes zurück führen, indem wir zeigen:

$$\mathbb{P}\Big(\max_{i=1}^{m}|S_i|\geq\lambda\sqrt{m}\Big)\leq 2\mathbb{P}(|S_m|\geq(\lambda-\sqrt{2})\sqrt{m}),\qquad m\in\mathbb{N},\lambda>0. \qquad (11.2.9)$$

Tatsächlich folgt aus einer Anwendung von (11.2.9) und Satz 8.3.19 (es sei Φ die Verteilungsfunktion der Standardnormalverteilung):

$$\frac{1}{\delta}\limsup_{m\to\infty}\mathbb{P}\Big(\max_{i=1}^{m}|S_i|\geq\frac{\eta}{\sqrt{\delta}}\sqrt{m}\Big)\leq\frac{2}{\delta}\limsup_{m\to\infty}\mathbb{P}\Big(\frac{|S_m|}{\sqrt{m}}\geq\frac{\eta}{\sqrt{\delta}}-\sqrt{2}\Big)=\frac{4}{\delta}\Phi\Big(\sqrt{2}-\frac{\eta}{\sqrt{\delta}}\Big).$$

Die rechte Seite verschwindet für $\delta\downarrow 0$, wie man leicht sieht. Also folgt (11.2.8) aus (11.2.9).

Nun beweisen wir (11.2.9).[1] Wir dürfen $\lambda>\sqrt{2}$ voraussetzen. Die Ereignisse

$$A_j=\bigcap_{i=1}^{j-1}\{|S_i|<\lambda\sqrt{m}\}\cap\{|S_j|\geq\lambda\sqrt{m}\},\qquad j=1,\dots,m,$$

[1] Man beachte, dass im Spezialfall der einfachen Irrfahrt (11.2.9) sehr leicht aus Satz 11.1.4 folgt.

sind paarweise disjunkt mit $A := \{\max_{i=1}^m |S_i| \geq \lambda\sqrt{m}\} = \bigcup_{j=1}^m A_j$. Es gilt

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}\big(A \cap \{|S_m| \geq (\lambda - \sqrt{2})\sqrt{m}\}\big) + \mathbb{P}\big(A \cap \{|S_m| < (\lambda - \sqrt{2})\sqrt{m}\}\big)\\ &\leq \mathbb{P}(|S_m| \geq (\lambda - \sqrt{2})\sqrt{m}) + \sum_{j=1}^{m-1} \mathbb{P}\big(A_j \cap \{|S_m| < (\lambda - \sqrt{2})\sqrt{m}\}\big),\end{aligned}$$

da $A_m \cap \{|S_m| < (\lambda - \sqrt{2})\sqrt{m}\} = \emptyset$. Für $j = 1, \dots, m-1$ gilt

$$A_j \cap \{|S_m| < (\lambda - \sqrt{2})\sqrt{m}\} \subset A_j \cap \{|S_m - S_j| > \sqrt{2m}\}.$$

Da die Ereignisse A_j und $\{|S_m - S_j| > \sqrt{2m}\}$ unabhängig sind, folgt aus dem Obigen

$$\mathbb{P}(A) \leq \mathbb{P}(|S_m| \geq (\lambda - \sqrt{2})\sqrt{m}) + \sum_{j=1}^{m-1} \mathbb{P}(A_j)\mathbb{P}(|S_m - S_j| > \sqrt{2m}).$$

Wegen

$$\mathbb{P}(|S_m - S_j| > \sqrt{2m}) \leq \frac{1}{2m}\mathbb{E}\Big[\Big(\sum_{k=j+1}^m X_k\Big)^2\Big] = \frac{1}{2m}\sum_{k=j+1}^m \mathbb{E}[X_k^2] \leq \frac{1}{2}$$

kann die rechte Seite weiter abgeschätzt werden durch

$$\mathbb{P}(A) \leq \mathbb{P}(|S_m| \geq (\lambda - \sqrt{2})\sqrt{m}) + \frac{1}{2}\sum_{j=1}^{m-1} \mathbb{P}(A_j) \leq \mathbb{P}(|S_m| \geq (\lambda - \sqrt{2})\sqrt{m}) + \frac{1}{2}\mathbb{P}(A).$$

Daraus folgt (11.2.9), was den Beweis des Satzes 11.2.4 von Donsker beendet. □

Wir wollen den Satz von Donsker anwenden, um exemplarisch die Verteilung zweier Funktionale der Brown'schen Bewegung zu identifizieren. Dabei erhalten wir auch die asymptotische Verteilung der entsprechenden Funktionale der zugrundeliegenden Irrfahrten. Wie immer sei Φ die Verteilungsfunktion der Standardnormalverteilung.

Lemma 11.2.7 (Verteilung des Maximums). *Sei $S_n = X_1 + \dots + X_n$ eine Irrfahrt auf $\mathbb{R}$ mit unabhängigen, identisch verteilten standardisierten Schritten $X_1, X_2, \dots$. Dann gilt für jedes $x \in \mathbb{R}$*

$$\mathbb{P}\Big(\max_{t\in[0,1]} B_t \leq x\Big) = \lim_{n\to\infty} \mathbb{P}\Big(\frac{1}{\sqrt{n}}\max_{i=1}^n S_i \leq x\Big) = \max\{2\Phi(x) - 1, 0\}. \tag{11.2.10}$$

Beweis. Natürlich können wir $x \geq 0$ voraussetzen. Wir betrachten $h\colon \mathcal{C}_0 \to \mathbb{R}$, definiert durch $h(f) = \max_{t\in[0,1]} f(t)$. Offensichtlich ist h stetig, und wir haben $h(B) = \max_{t\in[0,1]} B_t$ sowie $h(B^{(n)}) = \max_{i=0}^n S_i/\sqrt{n}$. Also ist die erste Gleichung in (11.2.10) eine direkte Folge aus dem Donsker'schen Invarianzprinzip, zusamen mit dem Continuous mapping theorem, Satz 8.2.11.

Um die zweite zu zeigen, benutzen wir, dass wir ohne Einschränkung die Schrittverteilung der Irrfahrt wählen dürfen, denn der Grenzwert hängt nicht davon ab. Also wählen wir die einfache Irrfahrt, wo X_i die Werte 1 und -1 jeweils mit Wahrscheinlichkeit $\frac{1}{2}$ annimmt. Hier kommt uns Satz 11.1.4 zu Pass, nach dem gilt (man beachte, dass wegen Symmetrie $h(B^{(n)})$ in Verteilung gleich $-M_n/\sqrt{n}$ ist):

$$\begin{aligned}\mathbb{P}(h(B^{(n)}) \geq x) &= \sum_{j\geq x\sqrt{n}} \mathbb{P}(M_n = -j) = \sum_{j\leq -x\sqrt{n}} \mathbb{P}_n(S_n \in \{j, j-1\})\\ &= 2\mathbb{P}(S_n > x\sqrt{n}) + \mathbb{P}(S_n = \lfloor -x\sqrt{n}\rfloor).\end{aligned}$$

Der letzte Term auf der rechten Seite ist (wenn er nicht sowieso schon Null ist) höchstens von der Ordnung $1/\sqrt{n}$, und der erste konvergiert nach dem Zentralen Grenzwertsatz gegen $2(1-\Phi(x))$. Daraus folgt die zweite Gleichung in (11.2.10) für die einfache Irrfahrt. □

Die zweite Anwendung betrifft das Lebesguemaß der Menge der Zeitpunkte, die die Brown'sche Bewegung im Positiven verbringt:

Lemma 11.2.8 (Arcussinus-Gesetz für die Brownsche Bewegung)**.** *Sei $S_n = X_1 + \cdots + X_n$ eine Irrfahrt auf $\mathbb{R}$ mit unabhängigen, identisch verteilten standardisierten Schritten $X_1, X_2, \dots$. Dann gilt für jedes $x \in [0,1]$*

$$\mathbb{P}\Big(\big|\{t \in [0,1]\colon B_t \geq 0\}\big| \leq x\Big) = \lim_{n\to\infty} \mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{S_i \geq 0\}} \leq x\Big) = \frac{2}{\pi}\arcsin\sqrt{x}. \qquad (11.2.11)$$

Beweis. Wir betrachten die Abbildung $h\colon \mathcal{C}_0 \to [0,1]$, definiert durch $h(f) = |\{t \in [0,1]\colon f(t) \geq 0\}|$. Leider ist h nicht stetig, aber immerhin ist seine Unstetigkeitsmenge $\mathcal{U}_h$ eine Nullmenge, wie wir gleich sehen werden.

Aber zunächst zeigen wir die Messbarkeit von h. Die Abbildung $\psi\colon \mathcal{C}_0 \times [0,1] \to \mathbb{R}$, definiert durch $\psi(f,t) = f(t)$, ist stetig, also $\mathcal{B}_{\mathcal{C}_0\times[0,1]}$-$\mathcal{B}$-messbar. Wie im Beweis von Lemma 8.2.9 beweist man, dass $\mathcal{B}_{\mathcal{C}_0\times[0,1]} = \mathcal{B}_{\mathcal{C}_0} \otimes \mathcal{B}_{[0,1]}$ gilt. Sei $A = \{(f,t)\colon f(t) \geq 0\} = \psi^{-1}([0,\infty)) \in \mathcal{B}_{\mathcal{C}_0} \otimes \mathcal{B}_{[0,1]}$. Wegen $h(f) = |\{t \in [0,1]\colon (f,t) \in A\}|$ ist h messbar bezüglich $\mathcal{B}_{\mathcal{C}_0}$.

Nun zeigen wir, dass $\mathcal{U}_h$ in $\{f\colon |\{t \in [0,1]\colon f(t) = 0\}| > 0\}$ enthalten ist. Sei $f \in \mathcal{C}_0$ mit $|\{t \in [0,1]\colon f(t) = 0\}| = 0$, und sei $(f_n)_{n\in\mathbb{N}}$ eine Folge in $\mathcal{C}_0$ mit $f_n \to f$ im Sinne der Supremumsmetrik. Dann gilt $\mathbb{1}_{[0,\infty)}(f_n(t)) \to \mathbb{1}_{[0,\infty)}(f(t))$ für fast alle $t \in [0,1]$. Nach dem Satz von Lebesgue über majorisierte Konvergenz folgt

$$\lim_{n\to\infty} h(f_n) = \lim_{n\to\infty} \int_{[0,1]} \mathbb{1}_{[0,\infty)}(f_n(t))\,\mathrm{d}t = \int_{[0,1]} \mathbb{1}_{[0,\infty)}(f(t))\,\mathrm{d}t = h(f).$$

Also gilt $\{f\colon |\{t \in [0,1]\colon f(t) = 0\}| > 0\}^{\mathrm{c}} \subset \mathcal{U}_h^{\mathrm{c}}$.

Nun zeigen wir, dass $\{f\colon |\{t \in [0,1]\colon f(t) = 0\}| > 0\}$ eine Nullmenge bezüglich des Wiener-Maßes $\mu = \mathbb{P} \circ B^{-1}$ ist. (Die Messbarkeit von $|\{t \in [0,1]\colon f(t) = 0\}|$ zeigt man wie oben.) Mit Hilfe des Satzes von Fubini sieht man

$$\begin{aligned}
&\int_{\mathcal{C}_0} |\{t \in [0,1]\colon f(t) = 0\}|\,\mu(\mathrm{d}f)\\
&\qquad = \int_{\mathcal{C}_0}\int_{[0,1]} \mathbb{1}_{f(t)=0}\,\mathrm{d}t\,\mu(\mathrm{d}f) = \int_{[0,1]}\int_{\mathcal{C}_0} \mathbb{1}_{f(t)=0}\,\mu(\mathrm{d}f)\,\mathrm{d}t\\
&\qquad = \int_{[0,1]} \mu(\{f\colon f(t) = 0\})\,\mathrm{d}t = \int_{[0,1]} \mathbb{P}(B_t = 0)\,\mathrm{d}t.
\end{aligned}$$

Die rechte Seite ist in der Tat Null, denn B_t hat ja eine Dichte für jedes $t > 0$. Also haben wir gezeigt, dass die Zufallsgröße $f \mapsto |\{t \in [0,1]\colon f(t) = 0\}|$ den Erwartungswert Null unter μ besitzt. Da sie nichtnegativ ist, ist sie also gleich Null fast sicher, d. h., $\mu(\{f\colon |\{t \in [0,1]\colon f(t) = 0\}| > 0\}) = 0$.

Das heißt, dass wir gezeigt haben, dass h μ-fast überall stetig ist. Also kann das Continuous mapping theorem auf h angewendet werden. Nach Lemma 8.2.11(b) konvergiert $h(B^{(n)})$ also in

Verteilung gegen $h(B) = |\{t \in [0,1] \colon B_t \geq 0\}|$. Man sieht leicht, dass

$$\frac{1}{n}\sum_{i=1}^{n} 1\!\!1_{\{S_i \geq 0\}} = \int_0^1 1\!\!1_{\{S_{\lceil tn \rceil} \geq 0\}}\,\mathrm{d}t = h\big(n^{-\frac{1}{2}} S_{\lceil \cdot\, n \rceil}\big),$$

und ein klein wenig Arbeit (Übungsaufgabe) ist nötig um einzusehen, dass die Differenz zwischen diesem und $h(B^{(n)})$ in Wahrscheinlichkeit gegen Null konvergiert. Also haben wir die erste Gleichung in (11.2.11). Für die einfache Irrfahrt folgt die zweite Gleichung aus Satz 11.1.8. Da die Grenzverteilung nicht von der Schrittverteilung abhängt, ist das Lemma bewiesen. $\square$

Bibliographie

[Ba91] H. Bauer, *Maß- und Integrationstheorie,* 2. Auflage, Walter de Gruyter, Berlin–New York, 1991.

[Ba02] H. Bauer, *Wahrscheinlichkeitstheorie*, 5. Auflage, Walter de Gruyter, Berlin–New York, 2002.

[Bi65] P. Billingsley, *Ergodic Theory and Information*, Wiley, New York 1965.

[Bi68] P. Billingsley, *Convergence of probability measures*, Wiley, New York 1968.

[Bi86] P. Billingsley, *Probability and Measure*, Wiley, New York, 1986.

[Ch78] K. L. Chung, *Elementare Wahrscheinlichkeitstheorie und stochastische Prozesse*, Springer, Berlin, 1978.

[Ch67] K. L. Chung, *Markov Chains with Stationary Transition Probabilities*, Springer, Berlin 1967.

[Co80] D.L. Cohn, *Measure Theory,* Birkhäuser, Basel 1980.

[Du02] R.M. Dudley, *Real Analysis and Probability*, Vol. 74 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge 2002, Revised preprint of the 1989 original.

[Fe68] W. Feller, *An Introduction to Probability Theory and its Applications*, Vol. I, 3rd ed., Wiley, New York, 1968.

[Ge02] H.-O. Georgii, *Stochastik. Einführung in die Wahrscheinlichkeitstheorie und Statistik*, Walter de Gruyter, 2002.

[H02] O. Häggström, *Finite Markov Chains and Algorithmic Applications*, Cambridge University Press, 2002.

[Kl06] A. Klenke, *Wahrscheinlichkeitstheorie*, Springer, Berlin, 2006.

[Kr02] U. Krengel, *Einführung in die Wahrscheinlichkeitstheorie und Statistik*, 6. Auflage, Vieweg, 2002.

[No97] J. Norris, *Markov Chains*, Cambridge University Press, 1997.

[Sch98] K. Schürger, *Wahrscheinlichkeitstheorie*, Oldenbourg, München, 1998.

[Se81] E. Seneta, *Non-Negative Matrices and Markov Chains*, Springer, New York, 1981.

[St94] D. Stirzaker, *Elementary Probability*, Cambridge University Press, Cambridge, 1994.

Index