Anton Bovier

Einführung in die Wahrscheinlichkeitstheorie

Vorlesung Winter 2012/13, Bonn

15. Februar 2013

universitätbonn iam

Inhaltsverzeichnis

Kapitel 1
Wahrscheinlichkeit

Il est remarquable qu'une science, qui a commencé par la considération des jeux, ce soit élévée aux plus importants objects des connaissances humaines[a].
Pierre Simon de Laplace, Théorie Analytique des Probabilités

[a] Es ist bemerkenswert, dass eine Wissenschaft, die mit der Betrachtung von Glücksspielen begonnen hat, sich zu einem der wichtigsten Gegenstände der menschlichen Erkenntnis erhoben hat.

In dieser Vorlesung werden wir ein Gebiet der Mathematik behandeln, dass sich von anderen dadurch hervorhebt, dass viele seiner Begriffe weitgehend Eingang in die Umgangssprache gefunden haben, ja, dass Fragen behandelt werden, die viele Menschen im täglichen Leben betreffen und von denen fast jedermann gewisse, ob falsche oder richtige, Vorstellungen hat.

Der zentrale Begriff, der uns hier beschäftigt, ist der des *Zufalls*. Was Zufall ist, oder ob es so etwas überhaupt gibt, ist eine tiefe philosphische Frage, der wir uns hier nur in wenigen Punkten annähern können; sie ist auch nicht der zentrale Gegenstand der Vorlesung. Grob gesprochen reden wir von "Zufall", wenn es sich um den Eintritt von *Ereignissen* handelt, die wir nicht oder nicht im Detail vorhersehen können. Typischerweise sind für ein solches Ereignis mehrere Varianten möglich, und wir reden von der Wahrscheinlichkeit des einen oder anderen Ausgangs. Ein beliebtes Beispiel ist etwa die Frage, ob es morgen regnet. In vielen Fällen ist dies möglich, aber nicht sicher. Der Wetterbericht macht darüber zwar Vorhersagen, aber auch diese treffen nur "mit einer gewissen Wahrscheinlichkeit ein". Wir können die Frage auch noch weiter spezifizieren, etwa danach wieviel Regen morgen fallen wird, und werden noch weniger sichere Vorhersagen bekommen. Gleiches gilt für sehr viele Vorkommnisse des täglichen Lebens. Der Begriff des Zufalls und der Wahrscheinlichkeit wird gebraucht, um solche Unsicherheiten qualitativ und quantitativ genauer zu beschreiben.

Unsicherheit tritt in vielen Situationen auf und wird sehr unterschiedlich wahrgenommen. Vielfach betrachten wir sie als Ärgernis und suchen eigentlich nach einer deterministischen Gesetzmässigkeit, die genauere Vorhersagen erlaubt. Dies betrifft insbesondere viele Bereiche von Naturwissenschaft und Technik, wo uns der Zufall vielfach nur in der Form von "Fehlern" und Un-

genauigkeiten begegnet, und wir bestrebt sind seine Effekte möglichst zu eleminiern oder doch zu minimieren.

In anderen Fällen ist der Zufall wesentlicher Motor des Geschehens und seine Existenz ist sogar gewollt und wird gezielt ausgenutzt. Am ausgeprägtesten ist dies sicher im *Glückspiel*, und in vieler Hinsicht ist hier die Wahrscheinlichkeitstheorie genuin zuhause and kann in ihrer reinsten Form beobachtet werden. Wie das Zitat von Laplace am Anfang dieses Kapitels belegt, sind die grundlegenden Prinzipien der Wahrscheinlichkeitstheorie zunächst in diesem Kontext entwickelt worden. In diesem Zusammenhang steht auch der Erfolg der Wahrscheinlichkeit unter dem Namen Finanzmathematik. Interessanterweise sind viele der mathematischen Prinzipien die hier entwickelt wurden, von der genauen Interpretation von Zufall gar nicht abhängig.

Literaturhinweise: Es gibt eine grosse Zahl von Lehrbüchern zur Wahrscheinlichkeitstheorie. Für die Vorlesung beziehe ich mich vielfach auf das Buch von Hans-Otto Georgii [6]. Ein Klassiker ist das zweibändige Werk von Feller [3, 4]. Persönlich gefällt mir auch das Buch von Chow und Teicher [2], dass allerdings in vielen Teilen schon eher das Niveau der Wahrscheinlichkeitstheorie 2 Vorlesung hat. Ein neueres Buch auf ähnlichem Niveau ist die Wahrscheinlichkeitstheorie von Achim Klenke [9]. Eine sehr elementare schöne Einführung ist ein neues Buch von Kersting und Wakolbinger [8].

1.1 Zufallsexperimente und Glückspiele

Die meisten klassischen Glückspiele beruhen auf einer Vorrichtung, die es erlaubt in unvorhersahbarer Weise wiederholbar eines aus einer Reihe möglicher Ausgänge eines Experiments zu produzieren. Typische Beispiele sind:

- **Münzwurf.** Eine Münze mit zwei unterschiedlich bedruckten Seiten ("Kopf" und "Zahl") wird in die Luft geworfen. Sie kommt schließlich auf dem Boden zu liegen und zeigt nun mit einer ihrer Seiten nach oben. Diese zwei möglichen Ausgänge stellen die zwei Ereignisse "Kopf" oder "Zahl" dar. Wir gehen davon aus, dass es uns nicht möglich ist den Ausgang vorherzusehen, wir betrachten diesen als völlig zufällig [dies mag eine Idealisierung sein, da ein sehr geschickter Münzwerfer den Ausgang des Experiments beeinflussen kann. Wir wollen hiervon aber absehen]. Wichtig ist hier, dass wir einen solchen Wurf beliebig oft wiederholen können, ohne irgendeine zusätzliche Information über den Ausgang des nächsten Wurfes zu bekommen.
- **Roulette.** Hier wird eine Kugel auf eine sich drehende Scheibe geworfen, die 37 numerierte identische Vertiefungen enthält, in einer von denen die Kugel am Ende des Experiments liegenbleibt. Auch hier wird eines der 37 möglichen Ereignisse in unvorhersehbarer Weise realisiert.

- **Würfeln.** Ähnlich wie der Münzwurf, es sind hier aber 6 Ereignisse möglich.
- **Lotto.** Aus einem Behälter, der 49 numerierte Kugeln enthält, werden 6 davon mit einem komplizierten Mechanismus herausgefischt. Aufgrund der Durchmischung am Anfang ist das Ergebnis nicht vorhersehbar. Die möglichen Ereignisse sind "sechs Zahlen aus den 49 ersten natürlichen Zahlen", zum Beispiel $3, 8, 19, 23, 25, 45$. Die Zahl der möglichen Ausgänge ist recht gross, nämlich $49!/43!/6! = \binom{49}{6} = 1\,398\,316$.
- **Zufallszahlengeneratoren.** Zufallszahlengeneratoren sind numerische Algorithmen, mit denen ein Computer Zahlenreihen (etwa aus $\{0, 1\}$) produziert, die möglichst zufällig sein sollen. In Wirklichkeit sind diese Reihen allerdings völlig deterministisch, können aber sehr irregulär von einem Anfangswert ("seed") abhängen. Die Erzeugung von Zufallszahlen ist ein wichtiges Problem, dem wir uns aber zunächst nicht weiter widmen wollen.

Wir wollen die Durchführung eines solchen "Experiments" in Zukunft als *Zufallsexperiment* bezeichnen. Jedem Zufallsexperiment kommt eine Menge möglicher Ausgänge zu. Diese Menge bezeichnen wir meißt mit Ω; sie wird den Namen *Wahrscheinlichkeitsraum* erhalten.

Ein Glücksspiel besteht nun darin, auf den Ausgang eines (oder mehrerer) Zufallsexperiments zu wetten. Der Übersichtlichkeit halber wollen wir uns auf das Roulettespiel konzentrieren. Hier gibt es "Spieler" sowie eine "Bank". Jeder Spieler hat die Möglichkeit einen von ihm gewählten Geldbetrag, g, darauf zu wetten, dass die nächste Ausführung des Zufallsexperiments "Ball-auf-Scheibe-werfen" damit endet, dass die Kugel in einer bestimmten Untermenge, $A \subset \Omega = \{0, \ldots, 36\}$, liegen bleibt. Wir wollen den Ausgang des Experimentes mit X bezeichnen. Als mögliche Untermengen sind eine Reihe Optionen auf dem Tisch vorgegeben, unter anderem aber auch jede beliebige Zahl von 0 bis 36. Die Wette besteht darin, dass die Bank den Einsatz, g, des Spielers einstreicht und verspricht, wenn das vom Spieler vorhergesagte Ereigniss, also $X \in A$, eintritt, ein festgelegtes Vielfaches des Einsatzes, gn_A, an den Spieler auszuzahlen (beachte, dass der Gewinn natürlich nur $(n_A - 1)g$ ist). Die Zahlen n_A sind von der Bank von Anfang an festgesetzt.

Die Bank wettet also mit $n_A : 1$ gegen das Eintreten des Ereignisses $X \in A$, der Spieler setzt $1 : n_A$ dafür. Diese Verhältnisse ("odds") geben in gewisser objektiver (jedenfalls aus Sicht der rational handelnden Bank) eine Einschätzung der Gewinnchancen wieder. Letzlich sind sie in gewisser Weise "objektive", weil in Geld umsetzbare, Bewertungen der Wahrscheinlichkeiten dieser Ereignisse.

Die Frage, wie solche Bewertungen gewählt werden sollen, ist die grundlegende Frage des Anwenders an den Mathematiker und steht am historischen Ursprung der Wahrscheinlichkeitstheorie. Wir wollen uns daher diesem Problem von verschiedenen Seiten zuwenden.

1.2 Allgemeine Eigenschaften von Bewertungen.

Im Fall des Roulette Spiels wird man sich leicht davon überzeugen lassen, dass die Bewertugen n_A umgekehrt proportional zu der Grösse der Menge A sein sollten (bereits bei einem elektronischen Roulette, dessen Programm man nicht kennt, wird man wesentlich skeptischer sein). Wir wollen aber vorerst von solchen speziellen Annahmen absehen und Eigenschaften herleiten, die unter allen Umständen gelten müssen, damit die Bank nicht unversehens ruiniert werden kann. Wir betrachen dazu einen viel allgemeineren Fall als das Roulette Spiel. Dazu sei Ω zunächst nicht weiter spezifiziert. Den Spielern sei eine Menge, $\mathfrak{A}$, von Teilmengen von Ω gegeben auf die sie beliebige Geldbeträge setzen dürfen. Über die Menge $\mathfrak{A}$ sei folgendes angenommen:

- Wenn $A, B \in \mathfrak{A}$, dann ist auch $A \cup B \in \mathfrak{A}$.
- Wenn $A \in \mathfrak{A}$, dann ist auch $A^c \equiv \Omega \backslash A \in \mathfrak{A}$.
- Der Form halber nehmen wir an, dass $\Omega \in \mathfrak{A}$ und somit auch $\emptyset \in \mathfrak{A}$.

Der erste Punkt ist unvermeidbar wenn A und B disjunkt sind, andernfalls ist diese Konvention eher vom mathematischen Standpunkt aus notwendig. Die zweite Bedingung erlaubt es dem Spieler "mit" der Bank zu spielen, was einer gewissen Fairness entspricht.

Die Bank möchte nun alle Mengen $A \in \mathfrak{A}$ bewerten. Dabei muss sie zunächst folgendes Prinzip beachten:
Keine risikofreien Gewinne: Es darf für die Spieler nicht möglich sein Einsätze zu tätigen, die ihnen mit Sicherheit, d.h. unabhängig vom Ausgang des Zufallsexperiments, einen Gewinn versprechen. Wir nennen eine solche Bewertung *zulässig*.

Lemma 1.1. *Jede zulässige Bewertung muss die Eigenschaft*

$$n_A^{-1} + n_{A^c}^{-1} \geq 1 \tag{1.2.1}$$

erfüllen.

Beweis. Ein Spieler könnte die Strategie verfolgen Beträge g_A und g_{A^c} auf die Mengen A und A^c so zu setzen, dass die erzielte Auszahlung, $g_A n_A \mathbb{1}_A + g_{A^c} n_{A^c} \mathbb{1}_{A^c}$, unabhängig von Ausgang des Experiments wird. ($\mathbb{1}_A$ bezeichnet hier die *Indikatorfunktion* des Ereignisses "die Kugel fällt in die Menge A" und nimmt den Wert 1 an, falls das Ereignis eintritt, und den Wert 0, falls das Ereignis nicht eintritt). Dazu muss lediglich

$$g_A n_A = g_{A^c} n_{A^c}$$

gelten, also $g_{A^c} = g_A n_A / n_{A^c}$. Es muss sichergestellt sein, dass in diesem Fall die Auszahlung, $g_A n_A$, den Einsatz, $g_A + g_{A^c}$, nicht übersteigt, also

$$g_A n_A \leq g_A + g_{A^c} = g_A \left(1 + n_A / n_{A^c}\right),$$

also

$$1 \leq n_A^{-1} + n_{A^c}^{-1},$$

wie behauptet. □

Insbesondere muss natürlich auch $n_\Omega \leq 1$ gelten, falls $\Omega \in \mathfrak{A}$.

In der Tat wählt die Bank, etwa im Roulette, Bewertungen so, dass die Ungleichung in (1.2.1) streng ist. Dies ist der Grund, warum Spielbanken meißt viel Geld verdienen. Im Gegensatz zu dieser Praxis stehen

Faire Bewertungen: Eine zulässige Bewertung heißt fair (oder maximal), wenn für jede Menge $A \in \mathfrak{A}$ gilt, dass

$$n_A^{-1} + n_{A^c}^{-1} = 1 \tag{1.2.2}$$

Die Bezeichnung "fair" begründet sich daher, dass hiermit dem Spieler, der auf A^c setzt, die gleiche Chance eingeräumt wird wie der Bank, wenn der Spieler auf A setzt. Die Bezeichnung "maximal" begründet sich daher, dass die Bank nicht systematisch unterboten werden kann, d.h. es ist nicht möglich eine Bewertung, n', zu finden mit der Eigenschaft, dass für alle $A \in \mathfrak{A}$, $n_A \leq n'_A$, ohne dass $n_A = n'_A$, für alle $A \in \mathfrak{A}$.

Satz 1.2. *Eine maximale zulässige Bewertung hat die Eigenschaft, dass, für alle $A, B \in \mathfrak{A}$,*

$$n_{A\cup B}^{-1} = n_A^{-1} + n_B^{-1} - n_{A\cap B}^{-1} \tag{1.2.3}$$

Insbesondere gilt, wenn $A \cap B = \emptyset$,

$$n_A^{-1} + n_B^{-1} = n_{A\cup B}^{-1} \tag{1.2.4}$$

Beweis. Wir zeigen zunächst (1.2.4). Wegen der Fairness der Bewertung ist schon einmal $n_{A\cup B}^{-1} = 1 - n_{(A\cup B)^c}^{-1}$, und der Spieler kann auf $A \cup B$ und $(A \cup B)^c$ so setzen, dass er sicher seinen Einsatz zurückerhält. Nun könnte er versuchen den Einsatz auf $A \cup B$ dadurch zu reproduzieren, dass er getrennt auf A und B die Beträge g_A, g_B setzt, so dass $n_A g_A = n_B g_B$ ist, d.h. es werden $g_A n_A$ ausgezahlt, wenn immer $X \in A \cup B$. Ferner soll dies der Auszahlung entsprechen, die der Spieler im umgekehrten Fall erhält, nämlich $n_{(A\cup B)^c} g_{(A\cup B)^c}$. Es folgt, dass $g_B = g_A \frac{n_A}{n_B}$ und $g_{(A\cup B)^c} = g_A \frac{n_A}{n_{(A\cup B)^c}}$. Damit ist der gesamte Einsatz

$$g_A + g_B + g_{(A\cup B)^c} = g_A \left(1 + \frac{n_A}{n_B} + \frac{n_A}{n_{(A\cup B)^c}}\right).$$

Die sichere Auszahlung, $n_A g_A$, darf diesen Betrag nicht überschreiten, was bedeutet, dass

$$n_A \leq 1 + \frac{n_A}{n_B} + \frac{n_A}{n_{(A\cup B)^c}} = 1 + n_A + \frac{n_A}{n_B} - \frac{n_A}{n_{A\cup B}}, \tag{1.2.5}$$

oder,

$$\frac{1}{n_{A\cup B}} \leq \frac{1}{n_A} + \frac{1}{n_B}. \tag{1.2.6}$$

Um zu zeigen, dass auch die umgekehrte Ungleichung gelten muss, müssen wir zeigen, dass es andernfalls möglich ist, statt auf $(A\cup B)^c$, auf A^c und B^c zu setzen um einen Einsatz auf $A\cup B$ abzusichern, und damit einen sicheren Gewinn zu machen. Die nötigen Einsätze sind dabei: $g_{A^c}, g_{B^c} = g_{A^c}\frac{n_{A^c}}{n_{B^c}}$, und $g_{A\cup B} = g_{A^c}\frac{n_{A^c}}{n_{A\cup B}}$. Es sei dem Leser überlassen, nachzuprüfen, dass dies einen sicheren Gewinn abwirft, ausser wenn

$$\frac{1}{n_{A\cup B}} \geq \frac{1}{n_A} + \frac{1}{n_B}. \tag{1.2.7}$$

Damit ist (1.2.4) gezeigt.

Falls A und B nicht-leeren Durchschnitt haben, können wir $A\cup B$ in die drei disjunktem Mengen $A\backslash B$, $B\backslash A$, und $A\cap B$ zerlegen, und das vorherige Resultat ausnutzen um (1.2.3) zu erhalten. □

Wir wollen noch schnell den Umkehrschluss machen und nachprüfen, dass die Eigenschaften von Theorem 1.2 ausreichend sind, so dass kein risikofreier Einsatz mit Gewinnoption existiert. Dazu betrachten wir einen allgemeinen Einsatz mit Wetten g_A auf alle Mengen $A\in\mathfrak{A}$. Wir nehmen der Einfachheit halber an, dass Ω eine endliche Menge ist, und dass alle einpunktigen Mengen, $x\in\Omega$, in $\mathfrak{A}$ enthalten sind. Der Gewinn bzw. Verlust im Fall des Ausgangs $X = x\in\Omega$ ist dann

$$r(x) = \sum_{A\in\mathfrak{A}} g_A n_A \mathbb{1}_{x\in A} - \sum_{A\in\mathfrak{A}} g_A$$

Nun ist $\sum_{x\in\Omega} n_x^{-1} = 1$, und daher

$$\begin{aligned}\sum_{x\in\Omega} n_x^{-1} r(x) &= \sum_{x\in\Omega} n_x^{-1} \sum_{A\in\mathfrak{A}} g_A n_A \mathbb{1}_{x\in A} - \sum_{A\in\mathfrak{A}} g_A \\ &= \sum_{A\in\mathfrak{A}} g_A \left(\sum_{x\in A} n_x^{-1} n_A - 1\right) = 0,\end{aligned} \tag{1.2.8}$$

weil nach (1.2.4)

$$\sum_{x\in A} n_x^{-1} n_A = 1.$$

Falls also in der Summe über $x\in\Omega$ einer der Terme $n_x^{-1}r(x) > 0$, so muss mindestens ein anderer Term $n_y^{-1}r(y) < 0$ sein. Unser Resultat zeigt, dass aus dem einfachen Prinzip, dass keine "sicheren" Gewinne in einer Spielbank möglich sein dürfen, erhebliche Einschränkungen an maximal mögliche Bewertung der verschiedenen Wetten hergeleitet werden können. Natürlich sind weiterhin noch viele Freiheiten vorhanden, und die Bank ist gut beraten, die genaue Auswahl sorgsam zu treffen. Auf diese Frage kommen wir gleich ausführlicher zu sprechen.

1.3 Faire Bewertungen und Wahrscheinlichkeitsmaße.

Wir wollen nun konzeptuell den Begriff der Wahrscheinlichkeit mit dem einer fairen Bewertung verbinden. Es scheint nämlich naheliegend, die Aussage "morgen regnet es mit 90-prozentiger Wahrscheinlichkeit" mit dem Angebot "ich wette zehn zu 1 darauf, dass es morgen regnen wird" gleichzusetzen. Wie sonst soll nämlich eine solche Aussage einen Nutzen haben? Im Roulettespiel heißt dass: Die Aussage,"die Kugel fällt in die Menge A mit Wahrscheinlichkeit $\mathbb{P}(A)$" bedeutet, dass die Bank dem Spieler das $n_A = 1/\mathbb{P}(A)$-fache seines Einsatzes, g_A, auszahlt, wenn dieses Ereignis eintritt. (Dass Banken unfaire Bewertungen anwenden wollen wir in diesem Zusammenhang nicht berücksichtigen). Natürlich sind diese so definierten Wahrscheinlichkeiten im Prinzip *subjektiv*: a priori könnte die Bank jede zulässige Bewertung anwenden.

Die oben diskutierten Eigenschaften von fairen Bewertungen legen nun eine sehr allgemeine *axiomatische* Definition von *Wahrscheinlichkeitsmaßes* nahe.

Zunächst wird der Begriff der *möglichen Wetten* zum Begriff der σ-Algebra erweitert.

Definition 1.3. Sei Ω eine Menge und sei $\mathfrak{A}$ eine Menge von Teilmengen ("Mengensystem"). von Ω. Man stattet $\mathfrak{A}$ mit den Operationen $\cup$ ("Vereinigung") und definiert als *Komplement*, A^c, die kleinste Menge in Ω, so dass $A \cup A^c = \Omega$. Falls $\mathfrak{A}$ die leere Menge $\emptyset$ enthält, und mit $A, B \in \mathfrak{A}$ auch $A \cup B \in \mathfrak{A}$ und $A^c \in \mathfrak{A}$, so heisst $\mathfrak{A}$ eine (Mengen)-*Algebra*.

Aus Vereinigung und Komplementbildung kann man auch den Durchschnitt von Mengen kostruieren als $A \cup B = (A^c \cup B^c)^2$. Somit ist eine Mengenalgebra auch unter dem Durchschnitt abgeschlossen. Klarerweise entspricht $\cup$ der Addition und $\cap$ der Multiplikation. Die Menge $\emptyset$ ist das neutrale Element der Addition und Ω das neutrale Element der Multiplikation.

Anmerkung. Im Sinne der Aussagenlogik entsprechen die Mengenoperationen der **Negation**, dem logischen **oder** und dem logischen **und**. Oft werden in der Wahrscheinlichkeitstheorie die Mengen A mit der Aussage "ein Zufallsexperiment hat einen Ausgang in der Menge A" identifiziert, und die Mengenoperationen daher mit den logischen Operationen bezeichnet.

Mengenalgebren scheinen zunächst der richtige Spielplatz für die Wahrscheinlichkeitstheorie. Für den Fall endlicher Mengen Ω ist das auch so. Wir werden aber sehen, dass wir im Allgemeinen um interessante Dinge machen zu können, noch eine zusätzliche Forderung stellen müssen.

Definition 1.4. Sei Ω eine beliebige Menge, und sei $\mathfrak{A}$ eine Menge von Teilmengen (ein *"Mengensystem"*) von Ω mit der Eigenschaft, dass

(i) $\Omega \in \mathfrak{A}$ und $\emptyset \in \mathfrak{A}$,

(ii) Falls $A \in \mathfrak{A}$, dann ist auch $A^c \equiv \Omega \backslash A \in \mathfrak{A}$.
(iii) Falls $A_n \in \mathfrak{A}$, für alle $n \in \mathbb{N}$, dann ist auch $\cup_{n\in\mathbb{N}} A_n \in \mathfrak{A}$.

Dann heißt $\mathfrak{A}$ eine σ-Algebra, und das Paar $(\Omega, \mathfrak{A})$ heißt ein *Messraum*.

Die neue Forderung (iii) wird es uns erlauben, Wahrscheinlichkeitsaussagen über Grenzwerte zu machen. Dies bringt gegenüber der elementaren kombinatorischen Wahrscheinlickeit ganz neue und interessante Fragestellungen.

Definition 1.5. Sei $(\Omega, \mathfrak{A})$ ein Messraum, und sei $\mathbb{P} : \mathfrak{A} \to \mathbb{R}_+$ eine Abbildung von $\mathfrak{A}$ in die positiven reellen Zahlen, mit folgenden Eigenschaften:

(i) $\mathbb{P}(\Omega) = 1$.
(ii)$\mathbb{P}(\emptyset) = 0$.
(iii) Falls die Mengen $A_i \in \mathfrak{A}$, $i \in \mathbb{N}$, disjunkt sind, dann gilt

$$\mathbb{P}\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} \mathbb{P}(A_i). \tag{1.3.1}$$

Dann heißt $\mathbb{P}$ ein *Wahrscheinlichkeitsmaß* auf dem Messraum $(\Omega, \mathfrak{A})$, und das Tripel $(\Omega, \mathfrak{A}, \mathbb{P})$ wird ein *Wahrscheinlichkeitsraum* genannt.

Anmerkung. In der Wahrscheinlichkeitstheorie benutzen wir stets die Konvention $0 \times \infty = 0$, bzw. $\sum_{n=k}^{\infty} 0 = 0$. Zeige, dass damit aus Punkt (iii) notwendig $\mathbb{P}(\emptyset) = 0$ folgt, und dass andererseits diese Konvention nötig ist, damit (1.3.1) mit $A_i = \emptyset$ für alle i gelten kann.

Anmerkung. Die Punkte (i) und (ii) aus Definition 1.4 sowie (i) und (ii) aus der Definition 1.5 sind evident aus den obigen Überlegungen. Die Punkte (iii) wären nur für endliche Vereinigungen zwingend, die Forderung dass die σ-Algebra auch unendliche Vereinigungen enthält ist aber mathematisch bequem, um im Fall unendlicher Mengen Ω nicht an unendlichen Iterationen zu scheitern. Ebenso ist Punkt (iii) in Defintion 1.5 in diesem Fall praktisch. Eigenschaft (iii) nennt man σ-Additivität. Die in der Definition 1.5 aufgestellten Bedingungen heißen *Kolmogorov's Axiome*. Sie bilden die Grundlage der abstrakten Theorie der Wahrscheinlichkeitsmaße.

Terminologie. Man verwendet gemeinhin die Bezeichnungen *Wahrscheinlichkeitsmaß, Wahrscheinlichkeitsverteilung* oder auch einfach *Verteilung* synonym. Die ebenfalls synonyme Bezeichnung *Wahrscheinlichkeitsgesetz* ist im Deutschen eher veraltet, wird aber sowohl im Englischen *"probability law", "law"*, wie auch im Französischen *"loi de probabilités", "loi"*, noch gängig gebraucht.

Für unseren späteren Gebrauch definieren wir gleich noch einige Verallgemeinerungen des Maßkonzepts.

Definition 1.6. Eine Abbildung $\mu : \Omega \to [0, +\infty]$, die alle Eigenschaften der Definition 1.5 erfüllt ausser $\mu(\Omega) = 1$ heißt ein Maß auf $(\Omega, \mathfrak{F})$. Falls $\mu(\Omega) < \infty$ heißt es ein endliches Maß. Ein Maß heißt σ-endlich, falls eine aufsteigende Folge, $\Omega_n \in \mathfrak{F}$, existiert, so dass $\Omega = \cup_{n=0}^{\infty} \Omega_n$, und $\mu(\Omega_n) < \infty$ für jedes n.

1.4 Die Gleichverteilung.

Im einfachsten Fall, wenn Ω eine endliche Menge ist (das ist in unseren Beispielen vom Roulette , wie überhaupt in den meisten Glückspielen, der Fall), gibt es eine privilegierte Wahrscheinlichkeitsverteilung, die Gleichverteilung, wo jedes Element, i, von Ω dieselbe Wahrscheinlichkeit, $\mathbb{P}(i) = 1/|\Omega|$, zugeordnet bekommt. Im Roulette oder beim Würfeln entspricht es der anscheinenden Symmetrie des physikalischen Experiments, dass dem Spiel zugrunde liegt, dass jeder elementare Spielausgang gleich wahrscheinlich erscheint, und es a priori keinen Grund gibt, etwa die Zahl 2 anders zu bewerten als die 36. Im allgemeinen Sprachgebrauch werden die Begriffe "zufällig" und "gleichverteilt" oft synonym gebraucht.

Tatsächlich ist die Gleichverteilung die priviligierte Verteilung, die vom sogenannten "Baysianischen" Standpunkt zu verwenden ist, wenn wir keinerlei Information über den Ausgang eines Zufallsexperiments vorliegen haben. Im Fall des Roulettespiels gehen wir ja auch davon aus, dass das Gerät so konstruiert ist, dass die faire Bewertung gerade der Gleichverteilung auf $\{0, \ldots, 36\}$ entspricht,

In der *kombinatorischen Wahrscheinlichkeitstheorie* geht es dann darum, auf der Basis einer solchen angenommenen Gleichverteilung, Wahrscheinlichkeiten komplizierterer Mengen auszurechnen; also etwa die Wahrscheinlichkeit zu berechnen, dass, wenn k Münzen mit gleichverteiltem Ausgang 0 oder 1 geworfen werden, die Summe der Ergebnisse gerade m ist. Klarerweise ist ja in diesem Fall für jede Menge A, $\mathbb{P}(A) = |A|/|\Omega|$, und alles was wir tun müssen ist die Grösse uns interessierender Mengen zu berechnen. Dies kann allerdings schwierig genug sein.

1.5 Wahrscheinlichkeit und Frequenz

Wir haben bisher das Konzept eines Wahrscheinlichkeitsmaßes mit einem Wettangebot identifiziert. Im Prinzip besteht damit noch überhaupt kein Zusammenhang zwischen einem solchen Maß und dem betrachteten Zufallsexperiment. Vielmehr ist es als eine subjektive Bewertung der Ereignisse durch die Spielbank zu betrachten. In den vorhergehenden Abschnitten haben wir nur gesehen, welche Restriktionen solche Bewertungen erfüllen müssen um

überhaupt akzeptabel zu sein, ganz unabhängig vom Ausgang des Zufallsexperiments.

Es stellt sich im Weiteren die Frage, wie irgend jemand, etwa eine Spielbank, zur Wahl einer konkreten Bewertung, also der Wahl einer Wahrscheinlichkeitsverteilung kommt. Dabei will eine Spielbank ja klarerweise Geld zu verdienen. Unter Annahme einer fairen Bewertung ist dies freilich nicht mit Sicherheit möglich; die Bank wird also versuchen die Aufgabe zu lösen, unter allen Bewertungen diejenige zu finden, bei der ihr auf lange Sicht der geringste Verlust droht, unabhängig davon, wie die Spieler agieren (und dann etwa weniger auszuzahlen). Es muss also die Bewertung in irgendeiner Form mit dem Ausgang der Zufallsexperimente in Bezug gesetzt werden. Dies ist die Aufgabe der *Statistik*.

Wir gehen dabei zunächst von der Premisse wiederholbarer Spiele aus. Wir nehmen an, dass die Bank ihre Bewertung ein für alle mal festlegt. Weiter nehmen wir (der Einfachheit halber) an, dass ein Spieler eine (beliebig) grosse Anzahl von Spielen zu spielen bereit ist, und dabei stets gleiche Einsätze macht[1].

Wir definieren nun die *Frequenzen* der Ausgänge der Roulettespiele,

$$f_k(A) \equiv \frac{1}{k} \sum_{i=1}^{k} \mathbb{1}_{X_i \in A}, \tag{1.5.1}$$

für $A \in \mathfrak{A}$, wo X_i der Ausgang des i-ten Spiels ist.

Notation: Wir schreiben ohne Unterscheidung

$$\mathbb{1}_{X \in A} = \mathbb{1}_A(X) = \begin{cases} 1, \text{wenn } X \in A, \\ 0, \text{ wenn } X \notin A. \end{cases}$$

Wir bemerken zunächst:

Lemma 1.7. *Die Abbildung $f_k : \mathfrak{A} \to \mathbb{R}_+$ ist ein Wahrscheinlichkeitsmaß.*

Beweis. Der Beweis ist eine Übungsaufgabe. □

Die Wahrscheinlichkeitsverteilung f_k heißt auch die *empirische Verteilung*, das heißt, die tatsächlich beobachtete Verteilung der Ausgänge.

Lemma 1.8. *Falls die (faire) Bewertung der Bank, n, die Gleichung $n_A = 1/f_k(A)$ für jedes $A \in \mathfrak{A}$ erfüllt, dann gilt für jeden möglichen Einsatz g_A, dass die Summe aller Auszahlungen der Bank in den k betrachteten Spielen genau der Summe der Einsätze des Spielers entspricht.*

Für jede andere faire Bewertung gibt es eine mögliche Einsatzstrategie des Spielers, die diesem einen positiven Gewinn sichert.

[1] Diese Annahme ist nicht notwendig, vereinfacht aber die Diskussion an dieser Stelle. Wir behandeln den allgemeinen Fall später.

Beweis. Falls $n_A = 1/f_k(A)$, so beträgt die Auszahlung der Bank

$$\sum_{i=1}^{k} \sum_{A} g_A n_A \mathbb{1}_{X_i \in A} = \sum_{A} g_A n_A k f_k(A) = k \sum_{A} g_A$$

was genau der Einsatz des Spielers ist.

Falls dagegen für irgendein $A \in \mathfrak{A}$ gilt, dass $n_A \neq 1/f_k(A)$, dann muss entweder $n_A > 1/f_k(A)$ gelten oder aber $n_{A^c} > 1/f_k(A^c)$. Wir können (modulo Umbenennung) annehmen, dass der erste Fall vorliegt. Dann setzen wir einen Betrag $g_A = 1$ auf A und nichts auf alle anderen Mengen.

Der Einsatz in k Spielen ist dann k, die Auszahlung der Bank aber

$$\sum_{i=1}^{k} n_A \mathbb{1}_{X_i \in A} = k n_A f_k(A) > k.$$

□

Nun kann die Bank n_A nicht so wählen wie im obigen Lemma, da die Bewertung ja vorab erfolgen muss und sich nicht am Ausgang der Spiele orientieren kann. Genausowenig kann der Spieler einen Einsatz in Abhängigkeit von f_k tätigen. Eine sinnvolle Bewertung ergibt sich, falls die oben eingeführten Frequenzen *konvergieren*.

Lemma 1.9. *Es sei angenommen, dass die Frequenzen $f_k(A)$ für alle $A \in \mathfrak{A}$ konvergieren, d.h.*

$$\lim_{k\to\infty} f_k(A) \equiv f(A)$$

existiert. Dann ist $f : \mathfrak{A} \to \mathbb{R}_+$ ein Wahrscheinlichkeitsmaß, und die Bewertung $n_A = 1/f(A)$ optimal im Sinne, dass sie die einzige Bewertung ist, so dass, für jede Einsatzstrategie g_A,

$$\lim_{k\to\infty} \frac{1}{k} \sum_{i=1}^{k} \sum_{A} g_A \left(n_A \mathbb{1}_{X_i \in A} - 1\right) = 0 \tag{1.5.2}$$

während es für jede andere Bewertung eine Strategie g_A gibt, so dass

$$\lim_{k\to\infty} \frac{1}{k} \sum_{i=1}^{k} \sum_{A} g_A \left(n_A \mathbb{1}_{X_i \in A} - 1\right) > 0 \tag{1.5.3}$$

Beweis. Übung! □

Die Idee ist hier natürlich, dass man eine grosse Anzahl, sagen wir k, Experimente durchführt und sich mit f_k eine gute Approximation des Limes f verschafft, bevor man den Spielbetrieb aufnimmt. f_k heißt in der Statistik ein *Schätzer* für die tatsächlichen Wahrscheinlichkeiten f.

Anmerkung. Mathematisch steht obiges Lemma auf sehr wackeligen Beinen. f_k ist ja eine Funktion der Ausgänge der Spiele 1 bis k, also von $X_1, \ldots, X_k$. Wir könnten das Lemma mit Sinn erfüllen, wenn wir etwa fordern, dass der limes für alle mögliche Spielausgänge existiert und unabhängig von diesem ist. Man kann sich aber leicht davon überzeugen, dass dies praktisch nie der Fall sein wird (man betrachte etwa den trivialen Fall $X_1 = X_2 = X_3 = \cdots = \omega$). Wir werden später sehen, dass es sinnvolle Konvergenzbegriffe für Folgen zufälliger Grössen gibt (insbes. die *fast sichere Konvergenz*), die es erlauben, sinnvolle und mathematisch rigorose Versionen dieses Lemmas zu formulieren.

Die obigen Beobachtungen bilden die Grundlage der *frequentistischen* Betrachtung von Wahrscheinlichkeiten. Ihr liegt immer die Annahme zugrunde, ein zufälliges Experiment könne beliebig oft wiederholt werden. Wenn dann die so gemessenen Frequenzen konvergieren, stellen sie ein Maß für die Wahrscheinlichkeitsverteilung des Ausgangs dar, was nach obigem Lemma offenbar sinnvoll ist. Viele Statistiker lassen nur diese Interpretation von Wahrscheinlichkeit gelten, womit aber nur in speziellen Situationen tatsächlich von Wahrscheinlichkeit gesprochen werden kann. Das Glückspiel ist offenbar ein Beispiel dafür.

Die frequentistische Interpretation erlaubt der Bank ihre Bewertung an Erfahrungswerte anzupassen. So wird sich beim Roulette heraustellen, dass nach vielen Spielen, jede Zahl mit einer Frequenz nahe 1/37 herauskommt. Dabei mag es auch Roulettetische geben, bei denen andere Werte beobachtet werden. Den Spielern ist diese Information in der Regel nicht zugänglich. Sie vertrauen darauf, dass dies dennoch so ist. Natürlich kann die Bank hier manipuliert haben. Eigentlich hat sie daran aber kein Interesse, da ihre Bewertung ja für diese Frequenzen optimiert ist. Gäbe es Abweichungen, und ein Spieler würde abweichende Frequenzen beobachten, könnte er seinen Einsatz dem anpassen, und so einen Vorteil erlangen.

1.6 Wahrscheinlichkeit und Information

Die frequentistische Interpretation von Wahrscheinlichkeit ist in vielen Fällen, in denen dennoch gerne von "Wahrscheinlichkeit" geredet wird, nicht sinnvoll, da es keine Wiederholung des Experiments unter gleichen Bedingungen geben kann oder wird. Das betrifft etwa die Aussage des Wetterberichts "die Wahrscheinlichkeit, dass es morgen regnet ist 30%". Am nächsten Tag wird es entweder regnen oder nicht regnen, und die Interpretation, dass es in 30 Prozent der Fälle morgen regnet, ist sinnlos. Allenfalls kann man sagen, dass Wettervorhersagen im allgemeinen mit einer gewissen Wahrscheinlichkeit richtig sind, was hier aber nicht gemeint ist.

Dasselbe Problem tritt bei manchen Formen des Glückspiels ein, insbesondere etwa bei Pferdewetten. Da auch hier kein Rennen wie ein anderes ist, stellt sich für die Bank hier die Frage nach der Bewertung der Ergebnisses

anders als im Roulette. Tatsächlich wird hier die Bank auch keine festen "a priori" Bewertungen verwenden, sondern diese werden von Rennen zu Rennen festgesetzt, und zwar *nachdem* die Spieler ihre Wetteinsätze getätigt haben. Dies erlaubt der Bank eine faire Bewertung zu finden, die wiederum für sie völlig risikofrei ist (und mittels eines Abschlags an eine faire Bewertung, sogar risikofrei Geld zu verdienen). Betrachten wir dies im einfachsten Fall, in dem jeweils nur auf den Sieg eines Pferdes der Betrag g_i gesetzt werden kann. Dann stellt $\mathbb{P}(i) \equiv g_i / \sum_{j \in \Omega} g_j$ eine Wahrscheinlichkeitsverteilung auf Ω dar, die die Erwartungen der Spieler über den Ausgang des Rennens wiederspiegelt. Wenn die Bank nun die Auszahlungen so wählt, dass beim Sieg von i eine Quote $n_i = 1/\mathbb{P}(i)$ auf den Einsatz g_i gezahlt wird, so zahlt sie unabhängig vom Ausgang des Rennens gerade den gesamten Einsatz wieder aus.

1.7 Wahrscheinlichkeit und Versicherung.

Bisher hatten wir Wahrscheinlichkeit stark in einem "spielerischen" Kontext gesehen. Oft sind wir aber unvorhersehbaren Ereignissen ausgesetzt und wollen unser handeln an Wahrscheinlichkeitsbewertungen solcher Ereignisse ausrichten. Dabei handelt es sich in aller Regel, zumindest aus der Sicht der Betroffenen, nicht um reproduzierbare Ereignisse. Machen wir das an einem einfachen Beispiel klar.

Ein Landwirt wird im Falle einer längeren Dürreperiode eine Verlust von $X = 10000\$$ hinnehmen müssen. Er möchte naturgemäss das Risiko, dem er ausgesetzt ist, bewerten. Dazu würde er gerne Aussagen über die Wahrscheinlichkeit des Ereignisses "Dürre" heranziehen. Angenommen, er bekommt eine Einschätzung dieser Wahrscheinlichkeit als $p = 0.001$. Wenn diese Aussage mit einem Wettangebot gekoppelt ist, kann er nun folgendes machen: Er setzt einen Betrag Y auf das Ereignis "Dürre" derart, dass er im Fall des Eintritts aus der Wette gerade seinen Verlust $X = 10000\$$ ausgleicht. Dazu muss er nur $10\$$ einsetzen (da $(1/p) * Y = 1000 * 10 = 10000 = X$). Er wird nun in jedem Fall, d.h. egal ob die Dürre kommt oder nicht jeweils nur seinen Einsatz von $10\$ = p * X$ verlieren. Das Dürrerisiko ist damit mit $10\$$ vernünftig bewertet. Für den Landwirt ist nunmehr gleich, was mit der Wahrscheinlichkeit p gemeint ist: worauf es ankommt, ist ein damit gekoppeltes Wettangebot, dass ihm erlaubt seinen Schaden unabhängig von der Dürre zu machen, d.h. sein Risiko gegen eine "Versicherungsprämie" zu eliminieren.

Fragt sich, warum der Versicherer ein solches Wettangebot machen kann. Idealerweise könnte der Versicherer sein Angebot auf einer frequentistischen Wahrscheinlichkeitsinterpretation aufbauen: er macht sehr viele solcher Geschäfte die vergleichbar sind und er mag Erfahrungen über die Häufigkeit solcher Ereignisse haben.

Anmerkung. Damit für den Versicherten die Risikobewertung durch eine Versicherung Sinn macht, muss von der Solvenz des Versicherers ausgegangen werden. Bei enorm grossen Schadenssummen ist letztere durchaus zweifelhaft. Aus pragmatischer Sicht ist dann eine wahrscheinlichkeitsbasierte Risikobewertung fragwürdig. Weiter ist die Sinnhaftigkeit einer solchen Risikobewertung auch nur dann gegeben, wenn eine entsprechender Versicherungswette auch abgeschlossen wird. Die gegenwärtige Finanzkrise ist teilweise auch dadurch begründet, dass Risikobewertung auf der Basis von Wahrscheinlichkeiten vorgenommen wurden, die entsprechenden risikobegrenzenden Wetten aber nie abgeschlossen wurden.

Kapitel 2
Elemente der Maßtheorie

On voit, par cet Essaie, que la théorie des probabilités n'est, au fond, que le bon sens reduit au calcul; elle fait apprécier avec exactitude ce que les esprits justes sentent par une sorte d'instinct, sans qu'ils puissent souvent s'en rendre compte[a]. Pierre Simon de Laplace, Théorie Analytique des Probabilités

[a] Man sieht durch diese Abhandlung, dass die Wahrscheinlichkeitstheorie im Grunde nur gesunder Menschverstand reduziert auf Berechnung ist; sie lässt mit Genauigkeit das erkennen, was verständige Geister durch eine Art Instinkt erfühlen, oft ohne dass sie dafür Rechenschaft ablegen könnten.

Wir haben im ersten Kapitel gesehen, dass unter einer vernünftig erscheinenden Definition des Wahrscheinlichkeitsbegriffes, in natürlicher Weise der Begriff eines Wahrscheinlichkeitsmaßes in der Form der Definition 1.5 auftaucht. Diese nunmehr *axiomatisch* definierten Objekte können nun mathematisch untersucht werden. In diesem Kapitel wollen wir einige der wichtigsten Eigenschaften von und Sätze über Wahrscheinlichkeitsmaße zusammentragen. Eine intensivere Behandlung wird in der Analysis III gegeben, die sehr zu empfehlen ist.

2.1 Wahrscheinlichkeitsmaße auf endlichen Mengen

Wenn auch die Theorie der W-Maße auf endlichen Mengen fast trivial ist, ist es nützlich, sich mit einigen Konzepten in diesem einfachen Zusammenhang vertraut zu machen.

Es sei also nun Ω eine endliche Menge, die wir ohne Beschränkung der Allgemeinheit als $\Omega = \{1, \ldots, N\}$ wählen können. Betrachten wir zunächst den einfachsten Fall, in dem die σ-Algebra von Ω jedes Element von Ω enthält. Dann ist die σ-Algebra von Ω die Menge aller Teilmengen von Ω, die sog. Potenzmenge von Ω, $\mathcal{P}(\Omega)$ (warum?). Ein Wahrscheinlichkeitsmaß, $\mathbb{P}$, auf Ω, ist dann ebenfalls durch die Angabe der Werte $\mathbb{P}(i)$, $i \in \Omega$, eindeutig festgelegt.

Lemma 2.1. *Sei $\Omega = \{1, \ldots, N\}$. Sei $\mathbb{P}$ ein W-Maß auf $(\Omega, \mathcal{P}(\Omega))$. Dann gilt:*

- $\mathbb{P}$ *ist durch die Angabe der Werte* $\mathbb{P}(i), i \in \Omega$*, eindeutig festgelegt, und es gilt* $\sum_{i\in\Omega} \mathbb{P}(i) = 1$*.*
- *Jede Sammlung positiver Zahlen* $p_i \geq 0$*,* $i \in \Omega$*, so dass* $\sum_{i\in\Omega} p_i = 1$ *definiert ein Wahrscheinlichkeitsmaß* $\mathbb{P}$ *auf* Ω *mit* $\mathbb{P}(i) = p_i$*.*

Beweis. Übung!! □

In obigen einfachen Kontext würden wir sagen, dass die σ-Algebra durch die Menge der ein-punktigen Mengen, $\{1\}, \{2\}, \ldots, \{N\}$, *erzeugt* wird. Darüber hinaus ist diese Untermenge der σ-Algebra *maßbestimmend*, d.h. die Werte des Maßes auf diesen Mengen legen das Maß fest.
Übung: Finde im obigen Fall eine andere erzeugende und maßbestimmende Menge von Teilmengen der σ-Algebra.

Es ist instruktiv, sich klarzumachen, dass nach unserem bisherigen Verständnis die Wahl der Potenzmenge als σ-Algebra über Ω durchaus nicht zwingend ist. So könnten wir zum Beispiel die Mengen (es sei N gerade) $\{1, 2\}, \{3, 4\}, \ldots, \{N-1, N\}$ als Basis einer σ-Algebra wählen. Es ist leicht zu sehen, dass die hiervon erzeugte σ-Algebra kleiner ist als die vorherige. Insbesondere sind die Elemente der zuvor betrachteten Basis, die ein-punktigen Mengen, hier nicht enthalten. Demnach ordnet ein Wahrscheinlichkeitsmaß, dass bezüglich dieser σ-Algebra definiert ist, diesen Einpunktmengen auch keine Werte zu.

Üblicherweise geht man bei der Beschreibung einer σ-Algebra so vor, dass man eine gewisse Menge von Teilmengen, die man in der σ-Algebra haben möchte vorgibt, und diese dann zu einer σ-Algebra ergänzt, indem man alle gemäß der Definition nötigen Mengen dazufügt.

Definition 2.2. Sei $\mathcal{E}$ eine Menge von Teilmengen von Ω. Die kleinste σ-Algebra, die $\mathcal{E}$ enthält, heisst die von $\mathcal{E}$ *erzeugte* σ-Algebra. Wir bezeichnen diese oft mit $\sigma(\mathcal{E})$. Für eine gegebene σ-Algebra, $\mathfrak{A}$, heisst eine Menge von Mengen, $\mathcal{E}$, *Erzeuger* (oder *Generator*) von $\mathfrak{A}$, wenn $\sigma(\mathcal{E}) = \mathfrak{A}$.

Wenn Ω endlich ist, ist es recht einfach, sowohl alle σ-Algebren (die dann auch einfach Algebren sind) zu beschreiben, sowie alle Wahrscheinlichkeitsmaße auf $(\Omega, \mathfrak{A})$ anzugeben. Der Grund ist folgendes einfaches Lemma.

Lemma 2.3. *Sei* $(\Omega, \mathfrak{A})$ *ein Messraum und* Ω *endlich. Dann enthält* $\mathfrak{A}$ *eine eindeutige minimale Partition,* $\Pi = (\pi_1, \ldots, \pi_n)$*, von* Ω *mit folgenden Eigenschaften:*

(i) $\bigcup_{i=1}^n \pi_i = \Omega$*;*
(ii) Für alle $B \in \mathfrak{A}$ *und alle* $k = 1, \ldots, n$*, gilt* $B \cap \pi_k \in \{\emptyset, \pi_k\}$*. Insbesondere gilt für alle* $i \neq j$*, dass* $\pi_i \cap \pi_j = \emptyset$*.*

Beweis. (Erst mal als Übung!) □

Proposition 2.4. *Sei* Ω *eine endliche Menge und* $(\Omega, \mathfrak{A}, \mathbb{P})$ *ein Wahrscheinlichkeitsraum. Dann ist das Maß* $\mathbb{P}$ *eindeutig durch die Werte* $p_i = \mathbb{P}(\pi_i)$*,* $i = 1, \ldots, n$*, festgelegt. Umgekehrt gibt es für jede Sammlung von Werten* $p_i \geq 0$*,* $i = 1, \ldots, n$*, mit* $\sum_{i=1}^n p_i = 1$ *ein Wahrscheinlichkeitsmaß auf* $(\Omega, \mathfrak{A})$*, so dass* $p_i = \mathbb{P}(\pi_i)$*.*

Beweis. Übung! □

2.1.1 Messbare Funktionen

Ein wesentliches Bestreben der Maßtheorie ist es, Funktionen gegen Maße zu integrieren. Im diskreten Fall scheint das weitgehend trivial, wir wollen aber doch einige allgemeine Ideen in diesem Fall entwickeln. Betrachten wir zunächst den Fall in dem die σ-Algebra die Potenzmenge ist. Sei dann $f : \Omega \to \mathbb{R}$ eine beliebige Funktion. Es ist klar dass wir mit dem Integral von f gegen $\mathbb{P}$ den Ausdruck

$$\int_\Omega f \, \mathrm{d}\mathbb{P} \equiv \sum_{i \in \Omega} f(i)\mathbb{P}(i) \tag{2.1.1}$$

meinen. Dies setzt aber die Existenz der Werte $\mathbb{P}(i)$ voraus. Hätten wir die kleinere σ-Algebra aus dem vorherigen Beispiel gewählt, könnten wir so offenbar nicht vorgehen.

Es lohnt sich also, nochmals über die Bedeutung des Integrals einer Funktion nachzudenken. Dazu empfiehlt sich die frequentistische Interpretation von $\mathbb{P}$. Sei z.B. $f(i)$ die Auszahlung, die beim Eintritt des Ereignisses $X = i$ anfällt. Wir sind dann an der "auf lange Sicht" erwarteten Rate der Auszahlung interessiert. Nun wird in unserem Fall f endlich viele Werte annehmen. Uns interessiert, wie häufig diese Werte vorkommen. Dies führt zu folgender Definition des "Integrals" einer solchen Funktion.

Definition 2.5. Sei $(\Omega, \mathfrak{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum, und sei $f : \Omega \to \mathbb{R}$ eine Funktion, die nur k Werte, $w_1, \ldots, w_k$, annimmt. Dann ist

$$\int_\Omega f \, \mathrm{d}\mathbb{P} = \sum_{\ell=1}^{k} w_\ell \mathbb{P}\left(\{i \in \Omega : f(i) = w_\ell\}\right), \tag{2.1.2}$$

genau dann wenn für alle ℓ

$$\{i \in \Omega : f(i) = w_\ell\} \in \mathfrak{F}.$$

Wir sehen also: der Ausdruck (2.1.2) kann genau dann berechnet werden, wenn alle Mengen $\{i \in \Omega : f(i) = w_\ell\}$ in der σ-Algebra bezüglich derer unser Wahrscheinlichkeitsmaß definiert ist enthalten sind!! Dies ist offenbar eine Eigenschaft einer Funktion bezüglich einer σ-Algebra. Wir wollen diese (vorläufig) wie folgt formalisieren.

Definition 2.6. Sei $(\Omega, \mathfrak{F})$ ein Messraum, und $f : \Omega \to \mathbb{R}$ eine reell-wertige Funktion. Dann heisst f *messbar* bezüglich $\mathfrak{F}$ (oder $\mathfrak{F}$-messbar), genau dann, wenn, für jedes $w \in \mathbb{R}$,

$$\{i \in \Omega : f(i) \leq w\} \in \mathfrak{F}. \tag{2.1.3}$$

Eine reell-wertige messbare Funktion auf $(\Omega, \mathfrak{F})$ heisst eine *Zufallsvariable* auf $(\Omega, \mathfrak{F})$.

Die Definition des Integrals mittels der Formel (2.1.2) hat den formalen Nachteil, dass sie die Kenntnis der Werte, w_i, die f annimmt voraussetzt. Dies wird bei der Verallgemeinerung auf allgemeine Messräume hinderlich sein. Wir können aber leicht eine Formel angeben, die mit (2.1.2) übereinstimmt, formal aber keine implizite Information über f voraussetzt.

Lemma 2.7. *Sei $(\Omega, \mathfrak{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum, und sei $f : \Omega \to \mathbb{R}$ eine messbare Funktion bezüglich $\mathfrak{F}$, die nur endlich viele Werte annimmt. Dann ist das Integral von f bezüglich $\mathbb{P}$ gegeben durch*

$$\int_\Omega f \, \mathrm{d}\mathbb{P} \equiv \lim_{\epsilon \downarrow 0} \sum_{k=-\infty}^{+\infty} k\epsilon \, \mathbb{P}\left(\{i \in \Omega : k\epsilon \leq f(i) < (k+1)\epsilon\}\right) \tag{2.1.4}$$

Beweis. Der Beweis ist recht einfach. Wenn $w_1, \ldots, w_k$ die Werte sind, die f annimmt, dann ist $\delta = \min_{i \neq j} |w_i - w_j| > 0$. Dann gilt zunächst, dass, für alle $0 < \epsilon < \delta/2$, jedes Intervall $(k\epsilon, (k+1)\epsilon]$ höchstens einen der Werte w_i enthalten kann. Für solche ϵ sei k_l so, dass $w_l \in (k_l\epsilon, (k_l+1)\epsilon]$. Dann ist

$$\begin{aligned}
\sum_{l=1}^{k} w_l \mathbb{P}\left(\{i \in \Omega : f(i) = w_l\}\right) &= \sum_{l=1}^{k} w_l \mathbb{P}\left(\{i \in \Omega : f(i) \in (k_l\epsilon, (k_l+1)\epsilon]\}\right) \\
&\geq \sum_{l=1}^{k} \epsilon k_l \mathbb{P}\left(\{i \in \Omega : f(i) \in (k_l\epsilon, (k_l+1)\epsilon]\}\right) \\
&= \sum_{k=-\infty}^{\infty} \epsilon k \mathbb{P}\left(\{i \in \Omega : f(i) \in (k\epsilon, (k+1)\epsilon]\}\right)
\end{aligned}$$

sowie auch

$$\begin{aligned}
\sum_{l=1}^{k} w_l \mathbb{P}\left(\{i \in \Omega : f(i) = w_l\}\right) &\leq \sum_{l=1}^{k} \epsilon(k_l+1) \mathbb{P}\left(\{i \in \Omega : f(i) \in (k_l\epsilon, (k_l+1)\epsilon]\}\right) \\
&= \sum_{k=-\infty}^{\infty} \epsilon k \mathbb{P}\left(\{i \in \Omega : f(i) \in (k\epsilon, (k+1)\epsilon]\}\right) \\
&\quad + \epsilon \sum_{k=-\infty}^{\infty} \mathbb{P}\left(\{i \in \Omega : f(i) \in (k\epsilon, (k+1)\epsilon]\right) \\
&= \sum_{k=-\infty}^{\infty} \epsilon k \mathbb{P}\left(\{i \in \Omega : f(i) \in [k\epsilon, (k+1)\epsilon)\}\right) + \epsilon
\end{aligned}$$

da die letzte Summe gerade dass Maß von Ω, also 1 ist. Da diese Ungleichungen für jedes $\epsilon < \delta/2$ gelten, folgt, dass

$$\begin{aligned}
&\limsup_{\epsilon\downarrow 0} \sum_{k=-\infty}^{\infty} \epsilon k \mathbb{P}\left(\{i \in \Omega : f(i) \in (k\epsilon, (k+1)\epsilon]\}\right) \\
&\leq \sum_{l=1}^{k} w_l \mathbb{P}\left(\{i \in \Omega : f(i) = w_l\}\right) \qquad (2.1.5) \\
&\leq \liminf_{\epsilon\downarrow 0} \sum_{k=-\infty}^{\infty} \epsilon k \mathbb{P}\left(\{i \in \Omega : f(i) \in (k\epsilon, (k+1)\epsilon]\right).
\end{aligned}$$

Dies beweist das Lemma und die Existenz des Limes in (2.1.4). □

Wir werden später sehen (siehe Section 2.2.4), dass wir mit der obigen Definition schon sehr nahe am allgemeinen Fall sind. Die einzige verbleibende Frage wird die der Konvergenz der Summen über k sein.

Das Integral einer messbaren Funktion, f, d.h. einer Zufallsvariablen, wir in der Regel auch als die *Erwartung* von f oder der *Erwartungswert*, oder *Mittelwert* von f, bezeichnet. Wir schreiben

$$\int_\Omega f \, \mathrm{d}\mathbb{P} \equiv \mathbb{E}_{\mathbb{P}} f \equiv \mathbb{E} f. \qquad (2.1.6)$$

Manchmal spricht man auch vom *mathematischen Erwartung* oder dem *mathematischen Mittel* von f. Dies wird getan um den Unterschied zum sogennanten *empirischen Mittel* zu betonen, der das arithmetische Mittel der Funktion f über n Wiederholungen eines Experiments darstellt,

$$\mathbb{E}_n^{\mathrm{emp}} f \equiv n^{-1} \sum_{k=1}^{n} f(X_k).$$

Der Zusammenhang zwischen mathematischem und empirischen Mittel ist eine der grundlegenden Fragen der Wahrscheinlichkeitstheorie.

2.1.2 Erwartungswerte und Risiko.

Wir wollen in Anknüpfung an unsere Diskussion aus Section 1.6 noch eine andere Interpretation des Erwartungswertes geben. Wir interpretieren wir die Werte $w_1, \ldots, w_k$ als die Verluste, die eine Person erleiden könnte. Die Angabe der Verlustwahrscheinlichkeiten, $\mathbb{P}(\{i \in \Omega : f(i) = w_\ell\})$, interpretieren wir als Wettangebote. Wir können nun Wetten so abschliessen, dass unser Verlust in jedem Fall gerade durch den Wettgewinn ausgeglichen wird, wir also unabhängig vom Zufall nur gerade unseren Wetteinsatz verlieren. Da-

zu müssen wir die Beträge $w_\ell \mathbb{P}(\{i \in \Omega : f(i) = w_\ell\})$ auf das Eintreten der Ereignisse $\{f = w_\ell\}$ setzen. Unser gesamter Einsatz, also unsere Versicherungsprämie, ist dann

$$\sum_\ell w_\ell \mathbb{P}(\{i \in \Omega : f(i) = w_\ell\}) = \int_\Omega f \, \mathrm{d}\mathbb{P}. \tag{2.1.7}$$

Damit haben wir dem Erwartungswert, zunächts im Fall positiver Zufallsvariablen eine eindeutige Interpretation als die Prämie gegeben, die wir aufbringen müssen, um uns vor jedem Risiko abzusichern.

Im Fall, dass wir neben Verlusten auch Gewinne erwarten, können wir die obige Formel problemlos übertragen, wenn wir davon ausgehen, dass wir im Fall negativer w_k unsererseits als Bank auftreten.

Diese Interpretation des Begriffs der Erwartung findet sich schon vor 200 Jahren bei Laplace [10]. Er schreibt: "La probabilité des événements sert à déterminer l'espérance ou la crainte des personnes interesées à leur existence. Le mot *espérance* a diverses acceptions: il exprime généralement l'avantage de celui qui attend un bien quelqu'onque, dans des suppositions qui ne sont que probables. Cet avantage, dansl la théorie des hasards, est le produit de la somme espérée par la probabilité de d'obtenir : c'est la somme partielle qui doit revenir lorsqu'on ne veut pas courir les risques de l'événement, en supposant que la repartition se fasse proportionellement aux probabilités. Cette repartition est la seule équitable, lorsqu'on fait sbstraction de toutes circonstances étrangères, parce qu'un égal degré de probabilité donne un droit égal sur la somme espérée. Nous nommerons cet avantage *espérance mathématique*[1]".

2.1.3 Erwartungswerte und Verteilungsfunktionen.

Wir wollen nun eine weitere nützliche Interpretation des Integralbegriffes untersuchen. Hierzu wollen wir den Ausdruck (2.1.2) in der Form

$$\int_\Omega f \, \mathrm{d}\mathbb{P} = \int_{\mathbb{R}} x \, \mathrm{d}P_f$$

[1] Die Wahrscheinlichkeit von Ereignissen dient zur Bestimmung der Erwartung oder der Furcht von Personen, die an ihrer Existenz interessiert sind. Das Wort. *Erwartung* hat verschiedene Bedeutungen: es drückt im allgemeinen den Vorteil desjenigen aus, der irgendeinen Vorteil erwartet, und zwar unter Annahmen, die nur wahrscheinlich sind. Dieser Vorteil ist in der Theorie der Zufälle das Produkt der erwarteten Summe und der Wahrscheinlichkeit sie zu erhalten: es ist die Teilsumme die man erhalten muss, wenn man das Risiko des Ereignisses nicht eingehen will, unter der Annahme, dass die Verteilung proportional zu den Wahrscheinlichkeiten erfolgt. Diese Verteilung ist die einzig gerechte, sofern man von allen fremden Umständen abstrahiert, da ein gleicher Grad von Wahrscheinlichkeit einen gleichen Anspruch an die erwartete Summe gibt. Wir nennen dieses Vorteil die *mathematische Erwartung*.

uminterpretieren, wobei nun P_f ein Maß auf den reellen Zahlen ist, dass jedem halb-offenen Intervall, $(x, y]$, die Maße

$$P_f\left((x,y]\right) \equiv \mathbb{P}\left(\{\omega \in \Omega : x < f(\omega) \leq y\}\right)$$

zuteilt. Es ist leicht zu sehen, dass diese Definition konsistent ist, wenn wir die Definition des Integrals (die wir bislang nur für endliche Mengen Ω begründet haben) formal auf den Fall $\Omega = \mathbb{R}$ ausdehnen, mit einer σ-Algebra, die die Menge aller halboffenen Intervalle enthält. Die Wahrscheinlichkeitsverteilung P_f ist die Verteilung der Werte von f in den reellen Zahlen, mithin die Verteilung der (reellen) Zufallsvariablen f (die wir hinfort häufig gerne mit X bezeichnen werden). Wir nennen P_f auch das Bild des Maßes $\mathbb{P}$ unter der Abbildung f. Eine besonders interessante Grösse ist dann die sogenannte *Verteilungsfunktion*, $F : \mathbb{R} \to [0, 1]$, die durch

$$F(x) = \mathbb{P}\left(\{\omega \in \Omega : f(\omega) \leq x\}\right) = P_f\left((-\infty, x]\right) \tag{2.1.8}$$

definiert ist. Beachte dass eine Verteilungsfunktion von dem Maß $\mathbb{P}$ und der Zufallsvariablen f abhängt, aber eindeutig durch die Verteilung P_f auf $\mathbb{R}$ bestimmt wird.

In unserem Fall eines endlichen Zustandsraumes ist die Verteilungsfunktion jeder Zufallsvariablen eine Stufenfunktion mit endlich vielen Sprüngen. Diese Sprünge liegen an den Punkten w_i, welche die Zufallsvariable f annimmt. Die Funktion F springt and der Stelle w_i um den Betrag $P_f(w_i) \equiv \mathbb{P}\left(\{\omega \in \Omega : f(\omega) = w_i\}\right)$, d.h.

$$F(w_i) = \lim_{x \uparrow w_i} F(x) P_f(w_i).$$

insbesondere ist F *wachsend* und *rechtsstetig*.

2.2 Wahrscheinlichkeitsmaße auf $\mathbb{R}$.

Wir sehen aus der obigen Diskussion, dass die Behandlung von Wahrscheinlichkeitsmaßen ausschließlich auf endlichen Mengen unbequem ist. Zumindest sollten wir in der Lage sein, Wahrscheinlichkeitsmaße auf den reellen Zahlen, $\mathbb{R}$, zu behandeln. Wie sich zeigen wird, ist dann der allgemeine Fall im wesentlichen sehr ähnlich.

2.2.1 Die Borel'sche σ-Algebra.

Grundsätzlich können wir genau wie im endlichen Fall vorgehen, und zunächst eine σ-Algebra konstruieren. Dazu brauchen wir erst mal eine Klasse von

Mengen, die darin enthalten sein sollen. Obwohl es hier natürlich viele Wahlmöglichkeiten gibt, wollen wir uns auf den kanonischen und wichstigsten Fall beschränken, der zu der sogenannten *Borel'schen* σ-Algebra, $\mathfrak{B} \equiv \mathfrak{B}(\mathbb{R})$, führt. Dazu fordern wir, dass $\mathfrak{B}$ die leere Menge und alle offenen Intervalle in $\mathbb{R}$ enthalten soll. Nach Definition einer σ-Algebra enthält $\mathfrak{B}$ dann alle Mengen, die durch abzählbare Vereinigung und Bildung von Komplementen, sowie die Grenzwertbildung von solchen Operationen erhalten werden können. Die Borel'sche σ-Algebra ist nun genau diejenige σ-Algebra, die eben auch gerade nur diese Mengen enthält, d.h. sie ist *die kleinste σ-Algebra, die alle offenen Intervalle* enthält.

Die in $\mathfrak{B}$ enthaltenen Teilmengen der reellen Zahlen heissen *Borel-Mengen.*

Die Borel-Mengen stellen eine äußerst reiche Klasse von Mengen dar. Insbesondere sind die folgenden Mengen allesamt Borel'sch:

(i) alle offenen Mengen;
(ii) alle abgeschlossenen Mengen.

Dies ist aber bei Weitem nicht alles. Eine "explizite" Angabe aller Borel-Mengen ist nicht möglich.

Anmerkung. Die Borel'sche σ Algebra ist strikt kleiner als die Potenzmenge von $\mathbb{R}$, d.h. es gibt Untermengen von $\mathbb{R}$, die nicht in $\mathfrak{B}$ enthalten sind. Solche Mengen sind in der Regel durch implizite Beschreibungen definiert. Die Borel'sche σ-Algebra ist für unsere Zwecke reich genug. Insbesondere kann auf ihr in sinnvoller Weise ein uniformes Maß, dass Lebesgue-Maß, definiert werden.

Beispiel einer nicht-Borel'schen Menge.

Wir definieren zunächst eine Äquivalenzrelation $\sim$ auf den reellen Zahlen in $[0,1]$ wie folgt: $x \sim y$ genau dann, wenn sie sich um eine rationale Zahl unterscheiden, also $x - y \in \mathbb{Q}$. Damit wird $[0,1]$ (und als Folge auch $\mathbb{R}$) in Äquivalenzklassen zerlegt. Wähle nun aus jeder Äquivalenzklasse ein Element aus (dies ist möglich unter Berufung auf das Auswahlaxiom) und bilde die Vereinigungsmenge, A, dieser ausgewählten Elemente. Dann gilt offenbar dass die reellen Zahlen die disjunkte Vereinigung der Mengen $A + q$, mit $q \in \mathbb{Q}$ sind (hier ist $A + q = \cup_{y \in A}\{y + q\}$. Die Menge A ist nicht Borel'sch. Das interessante an ihr ist, dass es unmöglich ist, ihr in konsistenter Weise eine Masse unter der Gleichverteilung μ zuzuordnen. Es muss dann nämlich gelten, dass $\mu(A) = \mu(A + q)$ für alle $q \in \mathbb{R}$; wenn nun aber $\mu(A) > 0$, dann gilt für jedes Intervall $I = [a, b]$

$$\sum_{q \in \mathbb{Q} \cap I} \mu(A + q) = \infty,$$

obwohl sicher

$$\cup_{q\in\mathbb{Q}\cap I}\{A+q\} \subset I' = [a, b+1]$$

und somit

$$\sum_{q\in\mathbb{Q}\cap I} \mu(A+q) = \mu\Big(\bigcup_{q\in\mathbb{Q}\cap I}\{A+q\}\Big) \leq \mu(I') < \infty$$

gelten muss. Also bliebe nur die Option $\mu(A) = 0$; dann aber wäre

$$\mu(\mathbb{R}) = \sum_{q\in\mathbb{Q}} \mu(A+q) = 0,$$

was offenbar auch nicht in unserem Sinn sein kann. Daher ist es besser, den Versuch dieser Menge eine Maße zu geben, zu unterlassen.

Wir sehen dass das Problem darin liegt, dass wir $\mathbb{R}$ (oder jedes Intervall in $\mathbb{R}$) in abzählbar viele gleichgroße Teile zerlegen wollen. Das Summierbarkeitsaxiom steht dieser Möglichkeit im Wege. Die Tatsache, dass die Menge A nicht Borel'sch zeigt man indirekt dadurch, dass das Lebesgue-Maß (das wir später konstruiren werden), jeder Borel-Menge eine Masse zuordnet.

Die Borel'sche σ-Algebra enthält ansonsten alle "vernünftigen" Mengen. Insbesondere enthält sie alle Punkte, $x \in \mathbb{R}$, alle kompakten Intervalle, alle halb-offenen Intervalle, sowie alle Halbachsen. Auch gibt es viele andere Charakterisierungen. Insbesondere die folgende ist für uns interessant (wegen Theorem 2.16).

Lemma 2.8. *Die Borel'sche σ-Algebra über $\mathbb{R}$ ist die kleinste σ-Algebra, die alle Mengen der Form*

$$\{y \in \mathbb{R} : y \leq x\}$$

enthält.

Beweis. Übung!! □

2.2.2 Maßbestimmende Mengen und Satz von Carathéodory.

Für unsere Zwecke ist das wichtigste Problem der Maßtheorie das folgende: Wie können wir in minimaler Weise ein Maß charakterisieren? Im Fall endlicher Mengen war das einfach; schlimmstenfalls hätten wir die Werte auf allen (endlich vielen!) Elementen der σ-Algebra angegeben, aber wie sich herausstellt genügt wegen der Additivität bereits die Kenntnis der Werte auf einer viel kleineren Menge, etwa auf allen Elementen von Ω. Im Fall des $\mathbb{R}$ ist das Problem dringlicher: die gesamte Borel σ-Algebra ist viel zu gross und unhandlich, als das wir die Maße aller ihrer Mengen angeben wollten. Wir

machen es also wie die Mathematiker es immer machen: Wir arbeiten einmal, und zeigen eine kleinere Menge von Mengen auf, die ausreicht, dass Maß auf allen Borel Mengen zu bestimmen. Das wird etwa die Menge der im vorherigen Lemma angegebenen Mengen sein. Diese einmalige Anstrengung wird uns später dann das Leben enorm erleichtern.

Wir werden dazu jetzt etwas abstrakter. Das macht die Dinge erstens einfacher, und zweitens arbeiten wir schon für später vor.

Als erstes definieren wir den Begriff von *durchschnitts-stabilen* Mengensystemen und *Dynkin-Systemen*.

Definition 2.9. Sei Ω eine Menge und $\mathfrak{A}$ eine Algebra von Teilmengen. Sei Ω eine Menge, und $\mathfrak{C}$ eine nicht-leere Teilmenge der Potenzmenge von Ω. Wir nennen $\mathfrak{C}$ ein Mengensystem.

(i) $\mathfrak{C}$ heisst *durchschnittsstabil*, falls für jedes $A, B \in \mathfrak{C}$ auch $A \cap B \in \mathfrak{C}$ gilt.
(ii) $\mathfrak{C}$ heisst ein Dynkin-System, genau dann wenn

a) $\Omega \in \mathfrak{C}$.
b) wenn $A, B \in \mathfrak{C}$ und $A \subset B$, dann ist auch $B \setminus A \in \mathfrak{C}$;
c) falls $A_1, A_2, \cdots \in \mathfrak{C}$ paarweise disjunkt sind, dann gilt $\cup_{n\in\mathbb{N}} A_n \in \mathfrak{C}$.

Dynkin-Systeme können viel kleiner sein als σ-Algebren. Andererseits fehlt Dynkin-Systemen zur σ-Algebra nur die Durchschnittsstabilität.

Lemma 2.10. *Jede σ-Algebra ist ein Dynkin-System. Jedes durchschnittstabile Dynkin-System ist eine σ-Algebra.*

Beweis. Da σ-Algebren sogar allgemeine Vereinigungen enthalten, sind sie inbesondere auch Dynkin-Systeme. Zu beweisen ist die zweite Aussage des Lemmas. Sei $\mathcal{D}$ ein Dynkin-System für das gilt, dass aus $A, B \in \mathcal{D}$ auch $A \cap B \in \mathcal{D}$. Wir wollen zeigen, dass dann $\mathcal{D}$ eine σ-Algebra ist. Dazu zeigen wir zunächst, dass $\mathcal{D}$ unter endlichen Vereinigungen abgeschlossen ist. Wenn $A, B \in \mathcal{D}$ sind, so sind dies auch A^c, B^c (da $A^c = \Omega \setminus A$). Dann ist auch $A^c \cap B^c \in \mathcal{D}$, weil $\mathcal{D}$ durchschnittsstabil ist. Dann ist aber auch $A \cup B = (A^c \cap B^c)^c \in \mathcal{D}$.

Nachdem wir wissen, dass endliche Vereinigungen in $\mathcal{D}$ liegen, können wir nun jede abzählbare Vereinigung, $\bigcup_{n\in\mathbb{N}} A_n$, in eine abzählbare Vereinigung, $\bigcup_{n\in\mathbb{N}} B_n$, paarweise disjunkter Mengen, $B_n \equiv A_n \setminus \bigcup_{k<n} A_k$, verwandeln, die dann wegen der Dynkin-Eigenschaft in $\mathcal{D}$ enthalten ist. Damit ist $\mathcal{D}$ eine σ-Algebra. □

Ferner gilt der Satz von Dynkin:

Satz 2.11. *Wenn $\mathfrak{C}$ ein durchschnittstabiles Mengensystem ist, dann ist das kleinste Dynkin-System, das $\mathfrak{C}$ enthält, gerade die von $\mathfrak{C}$ erzeugte σ-Algebra.*

Beweis. Hier gehen wir etwas indirekt vor. Für $A \in \mathfrak{C}$ betrachten wir die Menge $D_A \equiv \{B \in \mathcal{D}(\mathfrak{C}) : A \cap B \in \mathcal{D}(\mathfrak{C})\} \subset \mathcal{D}(\mathfrak{C})$. D_A ist ein Dynkin-System,

weil: (1) $A \cap \Omega = A \in \mathfrak{C} \subset \mathcal{D}(\mathfrak{C})$; (2) Wenn $B_1 \subset B_2$, und $A \cap B_i \in \mathcal{D}(\mathfrak{C})$, dann ist $(B_2 \setminus B_1) \cap A = (B_2 \cap A) \setminus (B_1 \cap A)$ und letztere Menge ist in $\mathcal{D}(\mathfrak{C})$ weil $\mathcal{D}(\mathfrak{C})$ ein Dynkin-System ist; (3) wenn $B_n, n \in \mathbb{N}$ paarweise disjunkt sind, und $B_n \cap A \in \mathcal{D}(\mathfrak{C})$, dann ist $(\cup_{n\in\mathbb{N}} B_n) \cap A = \cup_{n\in\mathbb{N}}(B_n \cap A)$; letzteres ist eine Vereinigung paarweise disjunkter Mengen $(B_n \cap A)$ aus $\mathcal{D}(\mathfrak{C})$, also auch in $\mathcal{D}(\mathfrak{C})$ (weil $\mathcal{D}(\mathfrak{C})$).

Damit ist für jedes $A \in \mathfrak{C}$ das Mengensystem D_A ein Dynkin-System; offenbar ist $\mathfrak{C} \subset D_A$, also ist $\mathcal{D}(\mathfrak{C}) \subset D_A$.

Damit sind wir noch nicht am Ziel: wir haben erst gezeigt, dass alle Durchschnitte von Mengen des Dynkin-Systems mit jeder Menge des Erzeugers in $\mathcal{D}(\mathfrak{C})$ liegen. Wir können nunmehr aber dieselbe Idee nochmals anwenden, d.h. wir definieren wieder D_A, diesmal aber für alle Mengen $A \in \mathcal{D}(\mathfrak{C})$. Nach dem vorher gezeigten Resultat sind nun alle Mengen des Erzeugers in jeder dieser Mengen enthalten. Andererseits sieht man mit exakt denselben Argumenten wie zuvor, dass D_A wiederum ein Dynkin System ist, und damit $\mathcal{D}(\mathfrak{C}) \subset D_A$, für alle $A \in \mathcal{D}(\mathfrak{C})$. Somit ist per Konstruktion $\mathcal{D}(\mathfrak{C})$ durschnittsstabil und daher nach Lemma 2.10 $\mathcal{D}(\mathfrak{C})$ eine σ-Algebra, die $\mathfrak{C}$ enthält, und $\sigma(\mathfrak{C}) \subset \mathcal{D}(\mathfrak{C})$.

Da ausserdem $\sigma(\mathfrak{C})$ ein Dynkin-System ist, dass $\mathfrak{C}$ enthält, gilt auch, wegen der Minimalität, dass $\mathcal{D}(\mathfrak{C}) \subset \sigma(\mathfrak{C})$, mithin $\mathcal{D}(\mathfrak{C}) = \sigma(\mathfrak{C})$. □

Der Unterschied einer Algebra zur σ-Algebra ist, dass keine abzählbaren Vereinigungen in $\mathfrak{A}$ enthalten sein müssen. Daher ist die durch ein Mengensystem erzeugte Algebra (die kleinste Algebra, die dieses Mengensystem enthält) viel kleiner, als die davon erzeugte σ-Algebra. (Insbesondere ist die σ-Algebra auch dann, wenn der Erzeuger abzählbar unendlich ist, überabzählbar, während die erzeugte Algebra nur abzählbar wäre).

Auf einer Algebra definiert man nun etwas, was schon fast ein Maß ist:

Definition 2.12. (i) Eine Abbildung $\mu : \mathfrak{A} \to \mathbb{R}_+$, heisst ein *Inhalt*, wenn $\mu(\emptyset) = 0$ und für alle disjunkten Mengen $A, B \in \mathfrak{A}$, $\mu(A \cup B) = \mu(A) + \mu(B)$.

(ii) Ein Inhalt heisst ein *Prämaß*, wenn für Folgen disjunkter Mengen $A_1, A_2 \cdots \in \mathfrak{A}$ für die $\cup_{n\in\mathbb{N}} A_n \in \mathfrak{A}$,

$$\mu\Big(\bigcup_{n\in\mathbb{N}} A_n\Big) = \sum_{n\in\mathbb{N}} \mu(A_n) \tag{2.2.1}$$

gilt.

Anmerkung. Falls $\mathfrak{A}$ eine σ-Algebra ist, und μ ein Prämaß, dann ist μ ein Maß. Wenn darüber hinaus $\mu(\Omega) = 1$, dann ist μ ein Wahrscheinlichkeitsmaß. Die Eigenschaft (ii) heisst σ-Additivität.

Die σ-Additivität ist in der Regel nicht sonderlich schwer nachzuprüfen. Das folgende Lemma macht dies transparent, und erklärt zum Teil warum wir die abzählbare Additivität für Maße fordern.

Lemma 2.13. *Sei μ ein endlicher Inhalt auf einer Algebra $\mathfrak{A}$. Dann sind die folgenden Aussagen äquivalent:*

(i)μ ist ein Prämaß.
(ii) Für alle monotone Folgen von Mengen $A_1, A_2, \cdots \in \mathfrak{A}$, so dass $A_n \downarrow \emptyset$, gilt $\lim_{n\to\infty} \mu(A_n) = 0$.

Beweis. Wir zeigen zunächst, dass (i) (ii) impliziert. Dazu sei $B_n \equiv A_n \backslash A_{n+1}$. Die Mengen B_n sind disjunkt, und $A_n = \cup_{m=n}^{\infty} B_m$, für jedes n. Also ist nach (i) $\sum_{m=n}^{\infty} \mu(B_m) = \mu(A_n)$. Die Konvergenz der Summe impliziert dass $\mu(A_n)$ nach Null konvergiert.

Wir zeigen nun die Gegenrichtung. Es sei nun B_n eine Folge disjunkter Mengen in $\mathfrak{A}$ so dass $B \equiv \cup_{n\in\mathbb{N}} B_n \in \mathcal{A}$. Setze nun $A_{n+1} \equiv \cup_{m=n+1} B_m = B \setminus \cup_{i=1}^{n} B_i \in \mathcal{A}$. Wegen der endlichen Additivität des Inhalts gilt

$$\mu(B) = \sum_{i=1}^{n} \mu(B_i) + \mu(A_{n+1}).$$

Da aber nach (ii) $\mu(A_{n+1}) \downarrow 0$, (denn $A_n \downarrow \emptyset$), so folgt dass $\mu(B) = \sum_{i=1}^{\infty} \mu(B_i)$, und der Beweis ist erbracht. □

Satz 2.14. *Sei $\mathfrak{F}$ eine σ-Algebra über Ω, und $\mathfrak{C}$ ein durchschnittstabiles Mengensystem das $\mathfrak{F}$ erzeugt. Falls zwei Wahrscheinlichkeitsmaße μ und ν auf $\mathfrak{C}$ übereinstimmen, dann gilt $\mu = \nu$ auf $\mathfrak{F}$.*

Beweis. Wir definieren

$$\widetilde{\mathfrak{F}} \equiv \{A \in \mathfrak{F} : \mu(A) = \nu(A)\}.$$

Wir wollen zeigen, dass $\mathfrak{F} = \widetilde{\mathfrak{F}}$. Dazu genügt es zu zeigen, dass $\widetilde{\mathfrak{F}}$ ein Dynkin-System ist. Denn da $\mathfrak{C}$ durchschnittsstabil ist, ist das kleinste Dynkin-System, dass $\mathfrak{C}$ enthält ja auch gerade die von $\mathfrak{C}$ erzeugte σ-Algebra, also $\mathfrak{F}$. Da aber nach Voraussetzung $\widetilde{\mathfrak{F}}$ gerade $\mathfrak{C}$ enthält, wären wir fertig. Prüfen wir also ob $\widetilde{\mathfrak{F}}$ ein Dynkin-System ist. Zunächst testen wir, ob relative Komplemente enthalten sind. Es ist aber, wenn $\mathcal{A}, B \in \widetilde{\mathfrak{F}}$, $A \subset B$,

$$\mu(B \setminus A) = \mu(B) - \mu(A) = \nu(B) - \nu(A) = \nu(B \setminus A),$$

also $B \setminus A \in \widetilde{\mathfrak{F}}$. Für paarweise disjunkte Mengen $D_n \in \widetilde{\mathfrak{F}}$ 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 ist auch $\cup_{n\in\mathbb{N}} D_n \in \widetilde{\mathfrak{F}}$. Damit ist die Behauptung bewiesen. □

Anmerkung. Die Aussage des Satzes gilt für allgemeine Maße, wenn zusätzlich angenommen wird, dass $\mathfrak{C}$ eine Folge von Mengen Ω_n mit den Eigenschaften

$\mu(\Omega_n) < \infty$ und $\Omega_n \uparrow \Omega$. Dies ist der Fall, wenn μ σ-endlich ist. Der Beweis in diesem Fall besteht darin, zu beobachten, dass die Maße μ_n und ν_n, definiert durch $\mu_n(A) \equiv \mu(A \cap \Omega_n)$, bzw. $\nu_n(A) \equiv \nu(A \cap \Omega_n)$ identisch sind, und andererseits $\mu_n \to \mu$, resp. $\nu_n \to \nu$ gilt.

Ein Mengensystem, dass die Voraussetzung des Satzes erfüllt nennt man maßbestimmend.

Zu unserem Glück fehlt nun nur noch die Beobachtung, dass aus Prämaßen Maße werden. Dies besagt der folgende wichtige Satz.

Satz 2.15 (Satz von Carathéodory). *Sei μ_0 ein (σ-)endliches Prämaß auf einer Algebra $\mathfrak{A}$. Dann gibt es genau ein Maß, μ, auf der von $\mathfrak{A}$ erzeugten σ-Algebra, das mit μ_0 auf $\mathfrak{A}$ übereinstimmt. μ heisst die Erweiterung von μ_0 auf $\sigma(\mathfrak{A})$.*

Anmerkung. Ich habe den Satz in voller Allgemeinheit für σ-endliche Maße angegeben; für den Zweck der Vorlesung können wir uns auf den Fall beschränken, wo μ_0 ein endliches Prämaß ist.

Anmerkung. Die Eindeutigkeit folgt aus dem vorhergehenden Satz sofort. Der Existenzbeweis würde hier zu weit führen. Er wird in der Vorlesung Maßtheorie erbracht. Interessanterweise zeigt dieser auch, dass die Borel'sch σ-Algebra im wesentlichen die grösstmögliche σ-Algebra ist auf der sich Maße konstruieren lassen, die die abzählbare Additivitätseigenschaft besitzen.

2.2.3 Verteilungsfunktionen.

Die für uns zunächste wichtigste Anwendung des Satzes von Carathéodory ist die Beobachtung, dass ein Wahrscheinlichkeitsmaß auf $\mathbb{R}$ durch seine Verteilungsfunktion eindeutig charakterisiert ist.

Satz 2.16. *Zu jeder monoton wachsenden, rechtsstetigen Funktion $F : \mathbb{R} \to \mathbb{R}$ gibt es genau ein Maß, μ, auf $(\mathbb{R}, \mathfrak{B})$, so dass $\mu((s,t]) = F(t) - F(s)$ ist, für alle $s < t \in \mathbb{R}$.*

Beweis. Wir nehmen ein Mengensystem $\mathfrak{C}$ dass aus allen Intervallen der Form $(s,t]$ besteht, mit $-\infty \leq s < t < \infty$, sowie zusätzlich allen Intervallen $(s, +\infty)$. Es sei $a(\mathfrak{C})$ die von diesen Intervallen erzeugte Algebra. Offenbar sind dies gerade alle endlichen Vereinigungen von halb-offenen Intervallen. Wir können nun für jedes solche Intervall den Wert von μ festsetzen als

$$\mu((s,t]) \equiv F(t) - F(s),$$

bzw.

$$\mu((s,\infty) = \lim_{t\to\infty} F(t) - F(s) \equiv F(\infty) - F(s).$$

Wir sehen auch, dass durch endliche Additivität diese Funktion auf die ganze Algebra erweitert werden kann, die Massen von disjunkten Vereinigungen sind gerade die Summe der Massen. Wichtig ist dabei die Konsistenz, nämlich, dass

$$\mu((s,t]) + \mu((t,r]) = \mu((s,r]),$$

wie man leicht nachprüft. Damit können wir μ auf $a(\mathfrak{C})$ erweitern und erhalten einen Inhalt. Um den Satz von Carathéodory anwenden zu können, bleibt nur noch übrig zu zeigen, dass μ ein Prämaß ist. Dann liefert dieser Existenz und Eindeutigkeit des Maßes μ auf der Borel σ-Algebra.

Dazu benutzen wir Lemma 2.13 und zeigen, dass für jede Folge $A_n \downarrow \emptyset$ in $a(\mathfrak{C})$, $\mu(A_n) \downarrow 0$. Dies wieder werden wir dadurch beweisen, dass aus der Annahme $\lim_{n\to\infty} \mu(A_n) > 0$ folgt, dass $\cap_{n\in\mathbb{N}} A_n \neq \emptyset$.

Es sei dafür A_n eine absteigende Folge von Teilmengen von $a(\mathfrak{C})$ mit $\lim_{n\to\infty} \mu(A_n) = a > 0$; ohne Beschränkung der Allgemeinheit können wir A_n aus $\mathfrak{C}$ wählen. Wir wollen nun zeigen, dass in jedem A_n noch eine nichtleere kompakte Menge $\bar{K}_n$ steckt, derart, dass die Folge $\bar{K}_n$ absteigend ist. Der unendliche Durchschnitt dieser Mengen kann aber nicht leer sein, andererseits ist er in $\cap_{n\in\mathbb{N}} A_n$ enthalten, weshalb auch letztere nicht leer sein kann.

Wie konstruieren wir nun diese Mengen? Wir zeigen zunächst, dass für jedes Intervall $I \in \mathfrak{C}$ mit $\mu(I) > 0$ und jedes $\epsilon > 0$ eine kompakte Menge L und ein Intervall $I' \in \mathfrak{C}$ existieren, so dass

$$I' \subset L \subset I, \quad \text{und} \quad \mu(I') \geq \mu(I) - \epsilon.$$

Sei nämlich $I = (s,t]$, so wähle man $I' = (s',t]$ mit $s' \in (s,t)$ derart, dass $F(s') \leq F(s) + \epsilon$ (dies ist stets möglich, da F rechtsstetig ist). Dann wählen wir zum Beispiel $L = [(s+s')/2, t]$, wenn $s \in \mathbb{R}$. Wenn $s = -\infty$ ist, wählt man stattdessen $L = [s'-1, t]$.

Wir konstruieren mit diesem Verfahren nun für jede Folge $A_n \in a(\mathfrak{C})$ mit $A_n \downarrow \emptyset$ Mengen B_n, K_n so dass

$$B_n \subset K_n \subset A_n, \quad \text{und} \quad \mu(B_n) \geq \mu(A_n) - a2^{-n-1}.$$

Nun ist leicht zu sehen, dass

$$\mu(B_1 \cap \cdots \cap B_n) \geq \mu(A_n) - \mu(\cup_{i=1}^n A_i \setminus B_i)$$

und da nach Konstruktion $\mu(A_i \setminus B_i) \leq \mu(A_i) - \mu(B_i) \leq a2^{-n-1}$ ist, folgt

$$\mu(B_1 \cap \cdots \cap B_n) \geq \mu(A_n) - \sum_{i=1}^n a2^{-n-1} \geq a - a/2 = a/2$$

Also ist $B_1 \cap \cdots \cap B_n$ für jedes n nicht leer und ist in der kompakten Menge $K_1 \cap \cdots \cap K_n \equiv \bar{K}_n$ enthalten. Letztere ist die gesuchte absteigende Folge

nichtleerer kompakter Mengen, die in $A_1 \cap \cdots \cap A_n$ enthalten ist. Damit kann $\cap_{i\in\mathbb{N}} A_i$ nicht leer sein. □

Anmerkung. Wir benutzen hier ein Resultat der Topologie: Falls K_n, $n \in \mathbb{N}$ kompakte Mengen sind so dass der Durchschnitt jeder endlichen Teilmenge dieser Mengen nicht leer ist, so ist $\cap_{n\in\mathbb{N}} K_n \neq \emptyset$. Der Beweis ist einfach: Falls die Aussage nicht wahr ist, so ist es etwa für jedes $x \in K_m$ (für gegebenes m) $x \in \cup_{n\in\mathbb{N}} K_n^c$. Da die Mengen K_n^c offen sind, so bilden Sie eine offenen Überdeckung von K_m. Da K_m kompakt ist, so besitzt nach Definition jede offenen Überdeckung eine endliche Teilüberdeckung, also $K_m \subset \cup_{i=1}^{\ell} K_{n_i}^c$. Es folgt dann aber, dass $K_m \cup_{i=1}^{\ell} K_{n_i} = \emptyset$, was einen Widerspruch darstellt.

Korollar 2.17. *Es existiert ein Maß auf $(\mathbb{R}, \mathfrak{B})$, das jedem Intervall gerade seine Länge zuordnet. Dieses Maß heisst das Lebesgue-Maß* [2].

Beweis. Wähle $F(t) = t$ im vorhergehenden Satz! □

Falls $F(\infty) - F(-\infty) = 1$, so ist das resultierende Maß ein Wahrscheinlichkeitsmaß, P. Indem wir noch $F(-\infty) = 0$ festlegen, ist F gerade die Verteilungsfunktion von P,

$$F(t) = P\left((-\infty, t]\right)$$

Definition 2.18. Wenn $(\Omega, \mathfrak{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum ist und $X : \Omega \to \mathbb{R}$ eine Zufallsvariable, so heisst die Funktion

$$F_X(x) \equiv \mathbb{P}(X \leq x), \tag{2.2.2}$$

die *Verteilungsfunktion* der Zufallsvariablen X.

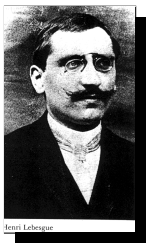
Klarerweise ist F_X gerade die Verteilungsfunktion des Wahrscheinlikcheitsmaßes P_X, nämlich $F_X(x) = P_X((-\infty, x])$.

Wir fassen als Korollar zusammen:

Korollar 2.19. *Jedes Wahrscheinlichkeitsmaß P auf $(\mathbb{R}, \mathfrak{B})$ ist eindeutig durch seine Verteilungsfunktion $F(t) = P\left((-\infty, t]\right)$ bestimmt. Umgekehrt ist jede rechtstetige, wachsende Funktion $F : \mathbb{R} \to [0, 1]$ mit $F(-\infty) = 0$ und $F(+\infty) = 1$ Verteilungsfunktionen eines Wahrscheinlichkeitsmaßes auf $\mathbb{R}$.*

2.2.4 Integration

[2] Benannt nach dem französischen Mathematiker Henri Léon Lebesgue (28.06.1875–26.07.1941).



Nachdem wir nun Maße auf $\mathbb{R}$ definiert haben, wollen wir uns erneut der Frage der Integration von Funktionen zuwenden. Zunächst liegt es nahe, unsere Definition der Messbarkeit im Lichte der Diskussion von Maßen auf $\mathbb{R}$ neu zu interpretieren.

Definition 2.20. Sei $(\Omega, \mathfrak{F})$ ein Messraum, und $f : \Omega \to \mathbb{R}$ eine reell-wertige Funktion. Dann heisst f eine *messbare Funktion* von $(\Omega, \mathfrak{F})$ nach $(\mathbb{R}, \mathfrak{B})$, genau dann, wenn für alle $B \in \mathfrak{B}$,

$$f^{-1}(B) \equiv \{\omega \in \Omega : f(\omega) \in B\} \in \mathfrak{F}.$$

Diese Definition stimmt mit unserer früheren Definition 2.6 der messbaren Funktionen überein, lässt sich aber leicht auf Funktionen zwischen beliebigen Messräumen übertragen:

Definition 2.21. Seien $(\Omega, \mathfrak{F})$ und $(\widetilde{\Omega}, \widetilde{\mathfrak{F}})$ Messräume, und $f : \Omega \to \widetilde{\Omega}$ eine Funktion. Dann heisst f eine messbare Funktion von $(\Omega, \mathfrak{F})$ nach $(\widetilde{\Omega}, \widetilde{\mathfrak{F}})$, genau dann, wenn für alle $B \in \widetilde{\mathfrak{F}}$,

$$f^{-1}(B) \equiv \{\omega \in \Omega : f(\omega) \in B\} \in \mathfrak{F}.$$

Eine nützliche Beobachtung, die insbesondere die Nachprüfung der Messbarkeit von Funktionen erleichtert, ist die folgende:

Lemma 2.22. *Sei $\mathfrak{F}$ eine σ-Algebra, und sei $f : \Omega \to \widetilde{\Omega}$. Sei $\mathfrak{A}$ die Menge aller Mengen der Form*

$$\mathfrak{A} \equiv \{A \subset \widetilde{\Omega} : f^{-1}(A) \in \mathfrak{F}\}.$$

Dann ist $\mathfrak{A}$ eine σ-Algebra.

Beweis. Zunächst ist klar, dass $f^{-1}(\widetilde{\Omega}) = \Omega$, so dass $\widetilde{\Omega} \in \mathfrak{A}$. Auch ist $f^{-1}(\emptyset) = \emptyset \in \mathfrak{F}$, so dass auch $\emptyset \in \mathfrak{A}$. Sei $A \in \mathfrak{A}$; dann ist

$$f^{-1}(A^c) \equiv \{\omega \in \Omega : f(\omega) \notin A\} = \{\omega : f(\omega) \in \mathfrak{A}\}^c,$$

also das Komplement einer Menge in $\mathfrak{F}$, mithin selbst in $\mathfrak{F}$. Somit ist auch $A^c \in \mathfrak{A}$. Seien schließlich A_i, $i \in \mathbb{N}$ in $\mathfrak{A}$. Dann ist

$$f^{-1}(\cup_i A_i) \equiv \{\omega \in \Omega : f(\omega) \in \cup_i A_i\} = \cup_i \{\omega \in \Omega : f(\omega) \in A_i\} \in \mathfrak{F},$$

und so $\cup_i A_i \in \mathfrak{A}$. Mithin ist $\mathfrak{A}$ eine σ-Algebra. □

Korollar 2.23. *Falls $\mathfrak{C}$ ein Mengensystem ist, das $\widetilde{\mathfrak{F}}$ erzeugt, dann ist f messbar, wenn für alle $C \in \mathfrak{C}$, $f^{-1}(C) \in \mathfrak{F}$.*

Beweis. Der Beweis ist denkbar einfach. Einerseits ist die Menge $\mathfrak{A} \equiv \{A : f^{-1}(A) \in \mathfrak{F}\}$ nach dem vorigen Lemma eine σ-Algebra, andererseits enthält sie einen Erzeuger, $\mathfrak{C}$ der σ-Algebra. Dann enthält sie mindestens die erzeugte σ-Algebra, mithin $\widetilde{\mathfrak{F}}$. □

Korollar 2.24. *Sei $f : \mathbb{R} \to \mathbb{R}$ eine stetige Funktion. Dann ist f messbar als Funktion von $(\mathbb{R}, \mathfrak{B}(\mathbb{R})) \to (\mathbb{R}, \mathfrak{B}(\mathbb{R}))$.*

Beweis. Wir müssen nur (z.B.) zeigen, dass die Urbilder von offenen Intervallen unter f Borelmengen sind. Nun ist aber das Urbild von offenen Mengen unter stetigen Abbildungen offen, und alle offenen Mengen sind Borel-Mengen. □

Im Fall das f eine beschränkte messbare Funktion und $\mathbb{P}$ ein Wahrscheinlichkeitsmaß auf $(\Omega, \mathfrak{F})$ ist, lässt sich die Definition des Integrals, die wir in (2.1.4) gegeben haben ohne weiteres wieder anwenden, wenn Ω nicht endlich, sondern etwa $\Omega = \mathbb{R}$ ist. Allerdings müssen wir aufgrund der in (2.1.4) auftauchenden unendlichen Summe etwas vorsichtiger sein und insbesondere die Existenz der verschiedenen Limiten abklären. Dem wollen wir uns nun zuwenden.

Eine zweckmässige Vorgehensweise (aber nicht die einzige) ist es, zunächst das Integral für sogenannte *einfache Funktionen* zu erkären.

Definition 2.25. Eine Funktion $g : \Omega \to \mathbb{R}$ heisst *einfach*, wenn sie nur endlich viele Werte annimmt, d.h. wenn es Zahlen $w_1, \ldots, w_k$ und Mengen $A_i \in \mathfrak{F}$ mit $\cup_{i=1}^k A_i = \Omega$, so dass $A_i = \{\omega \in \Omega : g(\omega) = w_i\}$. g kann dann geschrieben werden als

$$g(\omega) = \sum_{i=1}^{k} w_i \mathbb{1}_{A_i}(\omega).$$

Wir bezeichnen den Raum aller einfachen messbaren Funktion mit $\mathcal{E}$, und den Raum aller positiven einfachen messbaren Funktionen mit $\mathcal{E}_+$.

Es ist elementar zu sehen, dass jede einfache Funktion messbar ist. Für einfache Funktionen ist das Integral nun wie früher erklärt. (Im folgenden schreiben wir P für ein Maß, das nicht notwendig ein Wahrscheinlichkeitsmaß sein muss. Wer möchte, kann sich aber auf diesen Fall beschränken).

Definition 2.26. Sei $(\Omega, \mathfrak{F}, P)$ ein Maßraum und $g = \sum_{i=1}^k w_i \mathbb{1}_{A_i}$. Dann ist

$$\int_\Omega g \, \mathrm{d}P = \sum_{i=1}^{k} w_i P(A_i) \tag{2.2.3}$$

Diese Definition ist die einzig sinnvolle, wenn wir fordern, dass dass das Integral einer Indikatorfunktion einer Menge gerade das Maß dieser Menge ist, und dass das Integral eine lineare Abbildung sein soll.

Sei nun f eine positive, messbare Funktion. Die Grundidee ist, dass wir f durch einfache Funktionen annähern. Daher definieren wir

Definition 2.27. Sei f positiv und messbar. Dann ist

$$\int_\Omega f \, \mathrm{d}P \equiv \sup_{g \le f, g \in \mathcal{E}_+} \int_\Omega g \, \mathrm{d}P \tag{2.2.4}$$

Beachte, dass der Wert des Integrals in $[0, +\infty]$ liegt.

Schließlich zerlegt man eine allgemeine Funktion in ihren positiven und negativen Teil durch

$$f(\omega) = \mathbb{1}_{f(\omega)\geq 0} f(\omega) + \mathbb{1}_{f(\omega)<0} f(\omega) \equiv f_+(\omega) - f_-(\omega)$$

und definiert:

Definition 2.28. Sei f eine messbare Funktion und sei entweder $\int_\Omega f_+ \,\mathrm{d}P < \infty$ oder $\int_\Omega f_- \,\mathrm{d}P < \infty$. Dann ist das Integral von f bezüglich P gegeben durch

$$\int_\Omega f \,\mathrm{d}P \equiv \int_\Omega f_+(\omega) - \int_\Omega f_-(\omega) \,\mathrm{d}P. \tag{2.2.5}$$

Eine messbare Funktion heißt *integrierbar* (oder *absolut integrierbar*) bezüglich P, wenn $\int_\Omega f_+ \,\mathrm{d}P < \infty$ und $\int_\Omega f_- \,\mathrm{d}P < \infty$, oder, equivalent, $\int_\Omega |f| \,\mathrm{d}P < \infty$.

Man bezeichnet den Raum der gegen P integrierbaren Funktionen mit $L^1(\Omega, \mathfrak{F}, P)$ oder einfacher $L^1(\Omega, P)$.

Man benutzt die folgenden Notationen ohne Unterschied:

$$\int_\Omega f \,\mathrm{d}P = \int_\Omega f(\omega) \,\mathrm{d}P(\omega) = \int_\Omega f(\omega) P(\mathrm{d}\omega),$$

wobei wir die Angabe des Integrationsgebietes der Bequemlichkeit halber auch oft weglassen.

Der Satz von der monotonen Konvergenz stellt eine der wichtigsten Eigenschaften des Integrals fest.

Satz 2.29 (Monotone Konvergenz). *Sei $(\Omega, \mathfrak{F}, P)$ ein Maßraum und f eine nicht-negative reellwertige messbare Funktion. Sei $f_1 \leq f_2 \leq \cdots \leq f$ eine monoton wachsende Folge von nicht-negativen messbaren Funktionen, die punktweise gegen f streben, d.h., für jedes $\omega \in \Omega$ gilt $\lim_{n\to\infty} f_n(\omega) = f(\omega)$. Dann gilt*

$$\int_\Omega f \,\mathrm{d}P = \lim_{n\to\infty} \int_\Omega f_n \,\mathrm{d}P \tag{2.2.6}$$

Beweis. Es ist klar, dass

$$\int_\Omega f_n \,\mathrm{d}P \leq \int_\Omega f \,\mathrm{d}P, \tag{2.2.7}$$

und damit auch $\lim_{n\to\infty} \int_\Omega f_n \,\mathrm{d}P \leq \int_\Omega f \,\mathrm{d}P$. Wir müssen nur die umgekehrte Ungleichung beweisen. Für beliebiges $h = \sum_{i=1}^k h_i \mathbb{1}_{A_i} \in \mathcal{E}_+$ mit $h \leq f$ und $a < 1$ wollen wir zunächst zeigen, dass

$$\lim_{n\uparrow\infty} \int_\Omega f_n \,\mathrm{d}P \geq a \int_\Omega h \,\mathrm{d}P.$$

Sei E_n die messbare Menge $E_n \equiv \{\omega \in \Omega : ah(\omega) \leq f_n(\omega)\}$. Da $a < 1$ und $f_n \uparrow f$, muss die Folge E_n wachsend sein und $\Omega = \cup_n E_n$. Wir setzen

$$h_n(\omega) = ah(\omega)\mathbb{1}_{E_n}(\omega).$$

Dann ist $h_n \leq f_n$. Also ist

$$\int_\Omega f_n \,\mathrm{d}P \equiv \sup_{g \leq f_n, g \in \mathcal{E}_+} \int_\Omega g \,\mathrm{d}P \geq \int_\Omega h_n \,\mathrm{d}P = a \sum_{i=1}^{k} h_i P(A_i \cap E_n).$$

Da nun aber $E_n \uparrow \Omega$, gilt auch $A_i \cap E_n \uparrow A_i$, wenn $n \to \infty$ und somit auch $P(A_i \cap E_n) \uparrow P(A_i)$. Also ist

$$\lim_{n\to\infty} \int_\Omega f_n \,\mathrm{d}P \geq a \sum_{i=1}^{k} h_i P(A_i) = a \int_\Omega h \,\mathrm{d}P.$$

Da letzteres für jedes $a < 1$ und $h \in \mathcal{E}_+, h \leq f$ gilt, ist auch

$$\lim_{n\to\infty} \int_\Omega f_n \,\mathrm{d}P \geq \sup_{a<1} \sup_{h\in\mathcal{E}_+, h\leq f} a \int_\Omega h \,\mathrm{d}P = \sup_{a<1} a \int_\Omega f \,\mathrm{d}P = \int_\Omega f \,\mathrm{d}P. \tag{2.2.8}$$

Hieraus folgt mit (2.2.7) die Behauptung sofort. □

Der Satz von der monotonen Konvergenz erlaubt uns nun eine "explizite" Konstruktion (im Geiste von (2.1.4)) anzugeben.

Lemma 2.30. *Sei f eine nicht-negative messbare Funktion. Dann ist*

$$\int_\Omega f \,\mathrm{d}P \equiv \lim_{n\to\infty} \left[\sum_{k=0}^{n2^n-1} 2^{-n} k P\left(\omega : 2^{-n}k \leq f(\omega) < 2^{-n}(k+1)\right) + nP\left(\omega : f(\omega) \geq n\right) \right] \tag{2.2.9}$$

Beweis. Wir bemerken, dass auf der rechten Seite der Gleichung der Limes der Integrale der messbaren positiven, einfachen Funktionen

$$f_n \equiv \sum_{k=0}^{n2^n-1} 2^{-n} k \mathbb{1}_{\{\omega : 2^{-n}k \leq f(\omega) < 2^{-n}(k+1)\}} + n\mathbb{1}_{\{\omega : f(\omega) \geq n\}}$$

steht. Diese sind offenbar monoton wachsend und streben gegen f. Damit folgt das Lemma aus dem Satz von der monotonen Konvergenz. □

Anmerkung. Lemma 2.30 impliziert insbesondere, dass für zwei positive messbare Fuktionen f, g, $\int (f+g) dP = \int f dP + \int g dP$ gilt, d.h. die Integraloperation ist linear, was natürlich notwendig ist, damit der Integralbegriff

sinnvoll ist. Man könnte die Defininition 2.2.4 des Integrals auch auf nichtmessbare Funktionen ausdehenen. Dann ginge allerdings, wie man sich leicht an einfachen Beispielen klar machen kann, diese Eigenschaft verloren. Daher sind in der Tat nur messbare Funktionen sinnvolle Integranden.

Anmerkung. Falls P das Lebesguemaß und $\Omega = \mathbb{R}$, so heisst das so definierte Integral *Lebesgue Integral*. Im Fall $\Omega = \mathbb{R}$ heisst das Integral *Lebesgue-Stieltjes Integral*. Das Lebesgue Integral verallgemeinert das *Riemann Integral* insofern, als sehr viel mehr Funktionen im Lebesgue'schen Sinn integrierbar sind als im Riemann'schen. Andererseits gilt, dass jede Riemann integrierbare Funktion auch Lebesgue integrierbar ist, und dass in diesem Fall beide Integrale übereinstimmen. Dasselbe gilt auch für die Stieltjes-Varianten.

Die zwei folgenden Eigenschaften des Integrals werden immer wieder benötigt und sollen daher hier bewiesen werden. Der erste ist das *Lemma von Fatou:*

Lemma 2.31 (Lemma von Fatou). *Sei f_n eine Folge positiver messbarer Funktionen. Dann gilt*

$$\int_\Omega \liminf_n f_n \, \mathrm{d}P \leq \liminf_n \int_\Omega f_n \, \mathrm{d}P. \tag{2.2.10}$$

Beweis. Es ist

$$\liminf_n f_n(\omega) = \lim_{k\to\infty} \left(\inf_{n\geq k} f_n(\omega) \right)$$

wobei das Infimum in der Klammer eine monoton wachsende Funktionenfolge ist ist. Daher liefert der Satz von der monotonen Konvergenz, dass

$$\int_\Omega \liminf_n f_n(\omega) \, \mathrm{d}P(\omega) = \lim_{k\to\infty} \int_\Omega \left(\inf_{n\geq k} f_n(\omega) \right) \, \mathrm{d}P(\omega). \tag{2.2.11}$$

Andererseits ist für jedes $p \geq k$, und jedes $\omega \in \Omega$

$$\inf_{n\geq k} f_n(\omega) \leq f_p(\omega).$$

Deswegen ist

$$\int_\Omega \left(\inf_{n\geq k} f_n(\omega) \right) \, \mathrm{d}P(\omega) \leq \inf_{p\geq k} \int_\Omega f_p(\omega) \, \mathrm{d}P(\omega).$$

Daher erhalten wir aber, dass

$$\lim_{k\to\infty} \int_\Omega \left(\inf_{n\geq k} f_n(\omega) \right) \, \mathrm{d}P(\omega) \leq \lim_{k\to\infty} \inf_{p\geq k} \int_\Omega f_p(\omega) \, \mathrm{d}P(\omega) = \liminf_p \int_\Omega f_p(\omega) \, \mathrm{d}P(\omega), \tag{2.2.12}$$

was zu zeigen war. □

Der zweite zentrale Satz ist Lebesgue's Satz von der dominierten Konvergenz.

Wir sagen dass eine Folge von Funktionen f_n *P-fast überall* gegen eine Funktion f konvergiert, wenn

$$P\left(\{\omega : \lim_{n\to\infty} f_n(\omega) \neq f(\omega)\}\right) = 0.$$

Satz 2.32 (Dominierte Konvergenz). *Sei $(\Omega, \mathfrak{F}, P)$ ein Maßraum, f_n eine Folge von absolut gegen P integrierbaren Funktionen, f eine messbare Funktion und es gelte*

$$\lim_n f_n(\omega) = f(\omega) \quad P\text{-fast überall.} \tag{2.2.13}$$

Sei ferner $g \geq 0$ eine positive Funktion so dass $\int g \,\mathrm{d}P < \infty$ und es gelte

$$|f_n(\omega)| \leq g(\omega) \quad P\text{-fast überall.} \tag{2.2.14}$$

Dann ist f absolut integrierbar bezüglich P und

$$\lim_{n\to\infty} \int_\Omega f_n \,\mathrm{d}P = \int_\Omega f \,\mathrm{d}P. \tag{2.2.15}$$

Beweis. Wir nehmen zunächst an, dass die Annahmen, die fast überall gelten sollen sogar überall zutreffen.

Dann ist zunächst einmal $|f| \leq g$, und somit die absolute Integrierbarkeit von f eine direkte Folge der Integrierbarkeit von g. Da ferner $|f_n - f| \leq 2g$, und $|f_n - f| \to 0$, folgt mit Fatou's Lemma, dass

$$\liminf_n \int_\Omega (2g - |f_n - f|) \,\mathrm{d}P \geq \int_\Omega \liminf_n (2g - |f_n - f|) \,\mathrm{d}P = 2\int_\Omega g \,\mathrm{d}P. \tag{2.2.16}$$

Wegen der Linearität des Integrals ist das aber äquivalent zu

$$2\int_\Omega g \,\mathrm{d}P - \limsup_n \int_\Omega |f_n - f| \,\mathrm{d}P \geq 2\int_\Omega g \,\mathrm{d}P, \tag{2.2.17}$$

und daher

$$\limsup_n \int_\Omega |f_n - f| \,\mathrm{d}P = 0.$$

Dann folgt das Resultat wegen

$$\left|\int_\Omega f \,\mathrm{d}P - \int_\Omega f_n \,\mathrm{d}P\right| \leq \int_\Omega |f_n - f| \,\mathrm{d}P.$$

Um den allgemeinen Fall mit den nur fast sicheren Annahmen zu behandeln, setzen wir

$$A = \{\omega : f_n(\omega) \to f(\omega) \text{ und } |f_n(\omega)| \leq g(\omega) \text{ für alle } n\}.$$

Dann ist $P(A^c) = 0$. Aus dem vorherigen folgt, dass für die Funktionen $\tilde{f}_n \equiv f_n \mathbb{1}_A, \tilde{f} \equiv f\mathbb{1}_A$, die Aussage des Satzes gilt, während andererseits

$$\int_\Omega f_n \mathbb{1}_{A^c} \mathrm{d}P = \int_\Omega f \mathbb{1}_{A^c} \mathrm{d}P = 0.$$

Damit ist der Satz bewiesen. □

Ein einfaches Beispiel für eine Funktionenfolge, die die Voraussetzunges des Satzen von Lebesgue nicht erfüllt, ist

$$f_n(x) = \mathbb{1}_{[n.n+1]}(x).$$

Offensichtlich gilt für jedes $x \in \mathbb{R}$, $\lim_{n\uparrow\infty} f_n(x) = 0$. Die kleinste Majorante, die wir für f_n finden können ist $\mathbb{1}_{\mathbb{R}_+}$. Sei nun P das Lebesguemaß. Dann ist das Integral dieser Majorange unendlich. In der Tat gilt aber auch, dass

$$\int_{\mathbb{R}} f_n(x)dx = 1, \quad \text{für alle } n,$$

und somit $1 = \lim_{n\uparrow\infty} \int f_n dx \neq \int \lim_n f_n dx = 0$.

2.2.5 Abbildungen von Maßen

Wir kommen an dieser Stelle nochmals auf die bereits im diskreten angesprochene Frage der Verteilung einer Zufallsvariablen zurück. Diese Frage stellt sich jetzt so. Wir haben zwei Messräume, $(\Omega, \mathfrak{F})$ und $(\widetilde{\Omega}, \widetilde{\mathfrak{F}})$, eine W-Maß, $\mathbb{P}$, auf $(\Omega, \mathfrak{F})$ und eine messbare Abbildung $f : (\Omega, \mathfrak{F}) \to (\widetilde{\Omega}, \widetilde{\mathfrak{F}})$. Dann können wir auf $(\widetilde{\Omega}, \widetilde{\mathfrak{F}})$ ein neues Maß, P_f definieren durch die Forderung, dass für alle $A \in \widetilde{\mathfrak{F}}$,

$$P_f(A) \equiv \mathbb{P}(\{\omega \in \Omega : f(\omega) \in A\}) = \mathbb{P}\left(f^{-1}(A)\right). \tag{2.2.18}$$

Aufgrund der Messbarkeit von f ist dieses Maß offenbar wohldefiniert. Wir schreiben häufig

$$P_f \equiv \mathbb{P} \circ f^{-1}, \tag{2.2.19}$$

und nennen P_f das von f auf $(\widetilde{\Omega}, \widetilde{\mathfrak{F}})$ *induzierte* Maß oder das Bildmaß von $\mathbb{P}$ unter f.

Wenn insbesondere $(\widetilde{\Omega}, \widetilde{\mathfrak{F}}) = (\mathbb{R}, \mathfrak{B}(\mathbb{R}))$ ist, nennen wir P_f auch die Verteilung der Zufallsvariablen f.

Lemma 2.33. *Sei $(\Omega, \mathfrak{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $f : \Omega \to \mathbb{R}$ eine reellwertige integrierbare Zufallsvariable. Dann gilt*

$$\int_\Omega f(\omega)\,\mathrm{d}\mathbb{P}(\omega) = \int_{\mathbb{R}} x\,\mathrm{d}P_f(x). \tag{2.2.20}$$

Weiter ist, wenn $g : \mathbb{R} \to \mathbb{R}$ eine reellwertige messbare Funktion ist und $g \circ f$[3] integrierbar ist, dass

$$\int_\Omega g \circ f(\omega)\,\mathrm{d}\mathbb{P}(\omega) = \int_{\mathbb{R}} g(x)\,\mathrm{d}P_f(x) \tag{2.2.21}$$

Beweis. Es genügt Eq. (2.2.21) zu zeigen, da (2.2.20) ein Spezialfall mit $g(x) = x$ ist. Wir nehmen zunächst $g(x) = \mathbb{1}_B(x)$, mit $B \in \mathfrak{B}(\mathbb{R})$. Dann ist

$$\begin{aligned}\int_\Omega (\mathbb{1}_B \circ f)(\omega)\mathrm{d}\mathbb{P}(\omega) &= \int_\Omega \mathbb{1}_B(f(\omega))\mathrm{d}\mathbb{P}(\omega) \\ &= \mathbb{P}(\{\omega \in \Omega : f(\omega) \in B\}) = P_f(B) = \int_{\mathbb{R}} \mathbb{1}_B(x)\mathrm{d}P_f(x),\end{aligned} \tag{2.2.22}$$

d.h. (2.2.21) gilt für diesen Fall. Wenn g eine einfache Funktion ist, so folgt (2.2.21) aus (2.2.22) und der Linearität des Integrals. Als nächstes sei g positiv. Dann wählen wir eine Folge $g_n \uparrow g$ von positiven einfachen Funktionen, die punktweise gegen g konvergiert. Dann gilt auch, dass die Funktionen $h_n \equiv g_n \circ f : \Omega \to \mathbb{R}$ einfache Funktionen sind, die monoton gegen $h \equiv g \circ f$ konvergieren. Es gilt dann nach dem Satz von der monotonen Konvergenz, dass

$$\int_\Omega g\circ f(\omega)\mathrm{d}\mathbb{P}(\omega) = \lim_{n\uparrow\infty} g_n\circ f(\omega)\mathrm{d}\mathbb{P}(\omega) = \lim_{n\uparrow\infty}\int_{\mathbb{R}} g_n(x)\mathrm{d}P_f(x) = \int_{\mathbb{R}} g(x)\mathrm{d}P_f(x). \tag{2.2.23}$$

Schliesslich zerlegt man eine allgemeine messbare Fuktion g in ihren positiven und negativen Teil und benutzt das schon bewiesene für beide Teile. □

Insoweit wir uns nur für die Zufallsvariable f interessieren, können wir durch diese Abbildung unser Problem auf den Wahrscheinlichkeitsraum $(\mathbb{R}, \mathfrak{B}(\mathbb{R}), P_f)$ zurückführen auf dem unsere Zufallsvariable gerade die identische Abbildung ist. Für praktische Zwecke ist daher eine Zufallsvariable insbesondere durch ihre Verteilung charakterisiert.

Anmerkung. Wir haben oft folgendes Bild vor Augen: Wir beginnen mit einem Wahrscheinlichkeitsraum $(\Omega, \mathfrak{F}, \mathbb{P})$, den wir oft einen *abstrakten Wahrscheinlichkeitsraum* nennen. Auf diesem definieren wir dann Zufallsvariablen, die wir durch ihre Verteilungen charakterisieren (während wir nie weder das Maß $\mathbb{P}$ noch die Zufallsvariablen als Abbildungen explizit angeben.

[3] $\circ$ steht für Verkettung, also $g \circ f(\omega) \equiv g(f(\omega))$.

2.2.6 Beispiele von Wahrscheinlichkeitsmaßen.

Das einfachste Wahrscheinlichkeitsmaß aus $\mathbb{R}$ ist das sogenannte *Dirac-Maß* an einem Punkt $t \in \mathbb{R}$, δ_t. Es ist definiert durch

$$\delta_t(A) \equiv \mathbb{1}_A(t),$$

für jede Borel-Menge $A \in \mathfrak{B}$.

Das Dirac-Maß δ_t ist die Verteilung einer Zufallsvariablen, die stets den Wert t annimmt. Eine solche Zufallsvariable nennt man "deterministisch".

2.2.6.1 Diskrete Wahrscheinlichkeitsmaße.

Aus Dirac-Maßen kann man nicht-triviale Zufallsmaße durch die Bildung von konvexen Linearkombiationen bilden. Dazu benutzen wir den allgemein gültigen einfachen Satz:

Lemma 2.34. *Seien $\nu_1, \nu_2, \ldots$ Wahrscheinlichkeitsmaße auf einem Messraum $(\Omega, \mathfrak{F})$, und $p_i \geq 0$ für alle $i \in \mathbb{N}$ positive reelle Zahlen mit $\sum_{i\in\mathbb{N}} p_i = 1$, dann ist*

$$\mu \equiv \sum_{i\in\mathbb{N}} p_i \nu_i$$

ebenfalls ein Wahrscheinlichkeitsmaß auf $(\Omega, \mathfrak{F})$.

Beweis. Übung! □

Einige besonders wichtige diskrete Verteilungen sind:

Bernoulli Verteilung $\mathrm{Ber}(p)$.

$$\mathbb{P} = p\,\delta_1 + (1-p)\,\delta_0.$$

Diese Verteilung kommt von einem Münzwurf, in dem mit Wahrscheinlichkeit p Kopf (und mit Wahrscheinlichkeit $(1-p)$ Zahl erscheint). Die Zufallsvariable f, definiert durch $f(\text{Kopf}) = 1, f(\text{Zahl}) = 0$ hat dann die Verteilung $\mathbb{P}$.

Binomialverteilung $\mathrm{Bin}(n, p)$.

Eine besonders wichtige Verteilung ist die Binomialverteilung. Wir betrachten n Münzen aus dem vorherigen Beispiel, die mit Wahrscheinlichkeit p Kopf $(= 0)$ zeigen und die gleichzeitig geworfen werden. Der Zustandsraum dieses Experiments ist $\Omega = \{0, 1\}^n$. Wir definieren nun eine Funktion f auf Ω, durch

$$f(\omega) = \sum_{i=1}^{n} \mathbb{1}_{\{0\}}(\omega_i),$$

wo $\omega = (\omega_1, \ldots, \omega_n)$. Offenbar nimmt f Werte in $\{0, \ldots, n\}$ an. Wir überlegen uns leicht, dass

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

Daraus sehen wir, dass die Verteilung von f gegeben ist durch

$$P_{n,p} = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \delta_k.$$

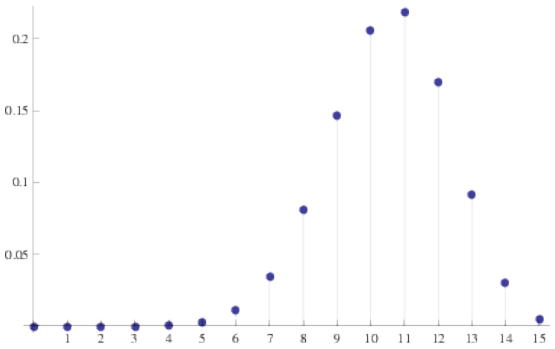


Abb. 2.1 Wahrscheinlichkeiten für $\mathrm{Bin}(n = 15, p = 0.7)$.

Poissonverteilung $\mathrm{Poi}(\rho)$.

Eine weitere wichtige Verteilung is die Poissonverteilung, eingeführt von Simón-Denis Poisson (1781–1840). Sie ist gegeben durch

$$P_\rho = \sum_{n=0}^{\infty} \frac{\rho^n}{n!} e^{-\rho} \delta_n.$$

wobei $\rho > 0$ ein Parameter ist. Die Poissonverteilung hängt mit der Binomialverteilung durch einen Grenzübergang zusammen. So können wir leicht sehen dass, wenn $p = \rho/n$ gewählt wird, die Koeffizienten $P_{n,\rho/n}(k)$ der Binomialverteilung gegen $P_\rho(k)$ (für festes k) konvergieren (im $n \to \infty$ Limes):

$$P_{n,\rho/n}(k) = \frac{n!}{k!(n-k)!} \frac{\rho^k}{n^k} (1-\rho/n)^{n-k} \to \frac{\rho^k}{k!} e^{-\rho},$$

denn

$$\frac{n!}{n^k(n-k)!} \to 1$$

und

$$(1-\rho/n)^n \to e^{-\rho}$$

und $(1-\rho/n)^{-k} \to 1$.

Wir werden in Kürze sehen, dass solche Grenzwertbildungen von zentralem Interesse in der W-Theorie sind und diese Problematik dementsprechend gründlich behandeln.

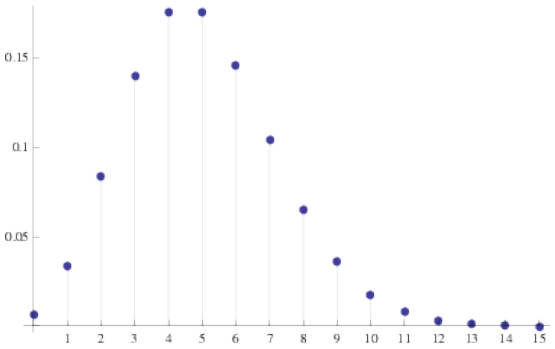


Abb. 2.2 Wahrscheinlichkeiten für Poi($\rho = 5$).

Geometrische Verteilung Geo(q).

Dies ist wieder eine Verteilung auf den positiven ganzen Zahlen mit

$$P_q(k) = q^k(1-q), \quad k \geq 0.$$

Sie hat eine wichtige Interpretation im Kontext des unendlich oft wiederholten Münzwurfs mit Parameter q: Wenn N die Nummer des Münzwurfs bezeichnet, bei dem erstmalig "Zahl" $(= 0)$ erscheint, dann ist

$$\mathbb{P}(\{N = k\}) = q^{k-1}(1-q) = P_q(k-1).$$

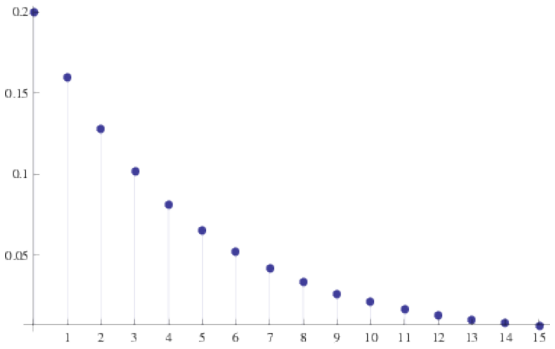


Abb. 2.3 Wahrscheinlichkeiten für Geo($q = 0.2$).

2.2.7 Absolut stetige Maße. Wahrscheinlichkeitsdichten.

Ein besonderer Fall von Wahrscheinlichkeitsmaßen auf $\mathbb{R}$ liegt in dem Fall vor, dass die Verteilungsfunktion, F, 'differenzierbar' ist. Genauer:

Definition 2.35. Sei F Verteilungsfunktion eines Maßes auf $(\mathbb{R}, \mathfrak{B}(R)$. Dann heisst F absolut stetig (bezüglich des Lebesgue Maßes), falls es eine positive, messbare Funktion $\rho : \mathbb{R} \to [0, \infty)$ gibt, so dass für alle $s < t \in \mathbb{R}$,

$$P((s,t]) = F(t) - F(s) = \int_s^t \rho(x)\, \mathrm{d}\lambda(x) \tag{2.2.24}$$

gilt, wobei λ das Lebesgue-Maß[4] ist. Wir nennen in diesem Fall die Funktion ρ die *Wahrscheinlichkeitsdichte* des Wahrscheinlichkeitsmaßes P.

Jede positive messbare Funktion ρ mit der Eigenschaft, dass $\int_0^\infty \rho(x) d\lambda(x) = 1$ bestimmt ein Wahrscheinlichkeitsmaß auf $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$.

Beachte, dass eine Wahrscheinlichkeitsdichte nicht eindeutig bestimmt ist. Wenn ρ, ρ' Wahrscheinlichkeitdichten sind und ausser auf einer Menge vom Lebesgue Maß Null $\rho(x) = \rho'(x)$, dann bestimmen ρ und ρ' das selbe Wahrscheinlichkeitsmaß.

Es gilt ferner, dass wenn F absolut stetig ist, dann ist F fast überall differenzierbar und für jede Dichte ρ von F gilt, dass für Lebesgue-fast alle x, $\rho(x) = F'(x)$. (Der Beweis dieser Aussage findet sich in fast jedem Lehrbuch der Maßtheorie, z.B. Satz 31.3 in [1]).

Warnung: In der nicht-mathematischen Literatur werden die Begriffe Verteilungsfunktion und Wahrscheinlichkeitsdichte häufig durcheinander geworfen. Vor allem in der englischsprachigen Literatur, wo diese *probability distribution function* und *probability density (function)* heissen, ist die Gefahr der

[4] Oft schreiben wir auch einfach dx für das Integral bezl. des Lebesgue Maßes.

Verwechslung gross. In der physikalischen Literatur wird häufig die Fiktion aufrechterhalten, alle Wahrscheinlichkeitsvereilungen besässen Dichten. Dazu wird insbesondere der Begriff der Dirac'schen Delta-Funktion eingeführt, der die Gleichung $\delta_x(y) = \delta(x-y)\,\mathrm{d}y$ zu schreiben erlaubt. Man muss sich aber klar sein, dass es viele Maße gibt, die werder eine Dichte haben, noch als abzählbare Summen von Dirac-Maßen geschrieben werden können.

Eine Vielzahl in der Praxis verwendeter Wahrscheinlichkeitsmaße ist absolut stetig. Dies liegt, wenigstens zum Teil, daran, dass diese einfacher zu handhaben sind wenn es um konkrete Berechnungen geht. Wichtige Beispiele sind etwa:

Gleichverteilung $\mathcal{U}_I$.

Für ein Intervall $I \subset \mathbb{R}$ ist die Gleichverteilung auf I definiert als

$$\mathrm{d}P_I(x) = |I|^{-1}\mathbb{1}_I(x)\,\mathrm{d}x$$

wo $\mathrm{d}x$ für das Lebesgue-Maß steht. Die Funktion $|I|^{-1}\mathbb{1}_I(x)$ ist die Wahrscheinlichkeitsdichte.

Gaußverteilung $\mathcal{N}(m,\sigma^2)$.

Die mit Abstand wichtigste Verteilung hat die Dichte

$$\phi_{m,\sigma^2}(x) \equiv \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-m)^2}{2\sigma^2}\right)$$

wobei $m \in \mathbb{R}$ Mittelwert, $\sigma > 0$ Standardabweichung und σ^2 Varianz heisst. Parameter sind auf die wir noch zu sprechen kommen. Aus vielen guten Gründen ist die Gaußverteilung die erste Wahl, wenn es um die Verteilung von Abweichungen um ein typisches Verhalten geht. Der Grund hierfür wird sich bei der Diskussion des zentralen Grenzwertsatzes offenbaren.

Interessanterweise wurde die Gauß'verteilung von dem in England lebenden Franzosen Abraham de Moivre (26.05.1667–27.11.1754) 1733 als Approximation der Binomialverteilung eingeführt. Gauß benutzte sie erst 1794 (publiziert 1809) in der Fehlerrechnung (Methode der kleinsten Quadrate).

Exponentialverteilung $\mathrm{Exp}(a)$.

Hier ist die Dichte

$$\rho(x) = ae^{-ax}\mathbb{1}_{[0,\infty)}(x)$$

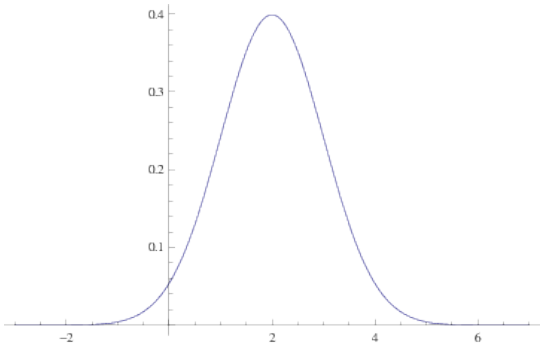


Abb. 2.4 Dichte der Gaussverteilung für $m = 2$ und $\sigma = 1$.

Die Exponentialverteilung tritt insbesondere als Verteilung von Wartezeiten gerne auf. Ihr Characteristikum ist die "Gedächtnislosigkeit". $a > 0$ is ein Parameter.

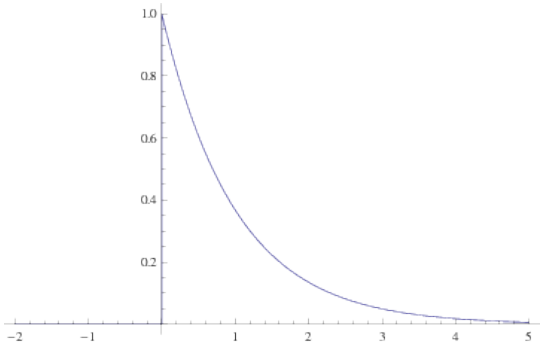


Abb. 2.5 Dichte der Exponentialverteilung mit $a = 1$.

Cauchy-Verteilung Cauchy(a).

Diese hat die Dichte

$$\rho(x) = \frac{a}{\pi} \frac{1}{a^2 + x^2}$$

Diese Verteilung zeichnet sich dadurch aus, dass die Funktion x nicht gegen sie integrierbar ist, d.h. dass kein Mittelwert existiert.

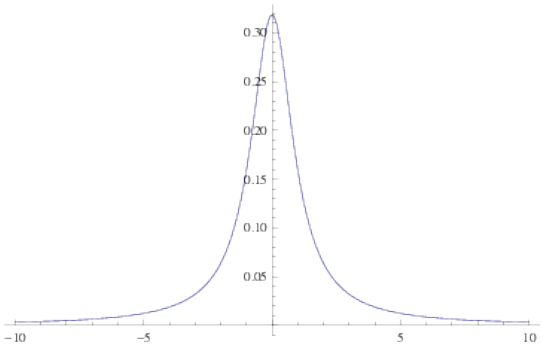


Abb. 2.6 Dichte der Cauchyverteilung mit $a = 1$.

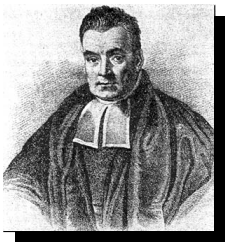
Kapitel 3
Bedingte Wahrscheinlichkeiten, Unabhängigkeit, Produktmaße

Si l'on considère les méthodes analytiques auxquelles cette théorie a donné naissance, la vérité des principes qui lui servent des base, la logique fine et délicate qu'exige leur emploi dans la solution des problemes, les établissements d'utilité publique qui s'appuient sur elle, et l'extension qu'elle a reçue et qu'elle peut recevoir encore par son application aux questions les plus importantes de la Philosopie naturelle et des Sciences morales; si l'on observe ensuite que, dans les choses mêmes qui ne peuvent être soumises au calcul, elle donne les aperçus les plus sûrs qui puissent nous guider dans nos jugements, et qu'elle apprend à se garantir des illusions qui souvent nous égarent, on verra qu'il n'est point de science plus digne des nos méditations et qu'il soit plus utile de faire entrer dans le système de l'instruction publique[a].
Pierre Simon de Laplace, Theorie Analytique des Probabilités

[a] Bedenkt man die analytischen Methoden, die diese Theorie hervorgebracht hat, die Wahrheit der ihr zugrundeliegenden Prinzipien, die feine und delikate Logik, die ihr Gebrauch bei der Lösung von Problemen erfordert, die gemeinnützigen Einrichtungen, die auf ihr beruhen, sowie die Erweiterungen, die sie erfahren hat und durch ihre Anwendung auf die wichtigsten Fragen der Naturphilosopie und der Geisteswissenschaften noch erfahren kann; wenn man weiter beobachtet, dass selbst in den Dingen, die sich der Berechnbarkeit entziehen, sie die gesichertesten Erkenntnissen liefert, die unser Urteil lenken können, und dass sie lehrt, sich vor Illusionen, die uns häufig in die Irre führen, zu bewahren, so sieht man, dass es keine Wissenschaft gibt, die unserer Meditationen würdiger wäre, und die in das öffentliche Bildungssystem aufzunehmen nützlicher wäre.

Bisher haben wir Wahrscheinlichkeitstheorie weitgehend wie einen Teil der Analysis behandelt. In diesem Kapitel kommen wir nun zu zentralen Konzepten, die mathematisch die Eigenständigkeit der Wahrscheinlichkeitstheorie begründen.

3.1 Bedingte Wahrscheinlichkeiten



Wir betrachten nunmehr einen beliebigen Wahrscheinlichkeitsraum $(\Omega, \mathfrak{F}, \mathbb{P})$. Es seien $A, B \in \mathfrak{F}$ zwei Ereignisse. Die Wahrscheinlichkeit von $A \cap B$, d.h. das gleichzeitige Eintreten beider Ereignisse ist $\mathbb{P}(A \cap B) \leq \min(\mathbb{P}(A), \mathbb{P}(B))$. Was uns nun interessiert ist, wie Information über das Ereignis B unsere Annahmen über das Ereignis A beeinflussen. Dazu definieren wir die *bedingte Wahrscheinlichkeit:*

Definition 3.1. Sei $(\Omega, \mathfrak{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und seien $A, B \in \mathfrak{F}$. Sei $\mathbb{P}(B) > 0$. Dann heisst

$$\mathbb{P}(A|B) \equiv \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \tag{3.1.1}$$

die *bedingte Wahrscheinlichkeit* von A gegeben B.

Diese Definition der bedingten Wahrscheinlichkeit ist einleuchtend und kompatibel mit der frequentistischen Interpretation von Wahrscheinlichkeiten: Wenn $\mathbb{P}$ eine empirische Verteilung ist, dann stellt $\mathbb{P}(A|B)$ offenbar die Frequenz des Eintretens von A unter all den Experimenten mit Ausgang in B dar.

Die bedingte Wahrscheinlichkeit hat zwei wichtige Eigenschaften:

Satz 3.2. *Sei $B \in \mathfrak{F}$ mit $\mathbb{P}(B) > 0$.*

(i) Die bedingte Wahrscheinlichkeit, $\mathbb{P}(\cdot|B)$ definiert ein Wahrscheinlichkeitsmaß auf dem Raum $(B, \mathfrak{F} \cap B)$, wo

$$\mathfrak{F} \cap B \equiv \{A \cap B, A \in \mathfrak{F}\} \tag{3.1.2}$$

(ii) Sei $B_n \in \mathfrak{F}$, $n \in \mathbb{N}$, eine paarweise disjunkte Folge von Mengen, so dass (a) $\cup_{n\in\mathbb{N}} B_n = \Omega$, (b) $\mathbb{P}(B_n) > 0$, für alle n. Dann gilt, dass, für alle $A \in \mathfrak{F}$,

$$\sum_{n\in\mathbb{N}} \mathbb{P}(A|B_n)\mathbb{P}(B_n) = \mathbb{P}(A) \tag{3.1.3}$$

Beweis. Bevor wir mit dem Beweis von (i) beginnen, müssen wir zeigen, dass $\mathfrak{F} \cap B$ eine σ-Algebra über B ist. Dies lässt sich aber sofort durch Nachprüfen der Axiome bestätigen. Als nächstes prüfen wir, ob $\mathbb{P}(\cdot|B)$ ein Wahrscheinlichkeitsmaß ist. Offenbar gilt $\mathbb{P}(B|B) = 1$ und $\mathbb{P}(\emptyset|B) = 0$. Weiterhin gilt, dass

$$\begin{aligned}\mathbb{P}(B\backslash A|B) &= \frac{\mathbb{P}(B\backslash A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B\backslash A)}{\mathbb{P}(B)} \\ &= \frac{\mathbb{P}(B) - \mathbb{P}(A \cap B)}{\mathbb{P}(B)} = 1 - \mathbb{P}(A|B).\end{aligned}$$

Sei schliesslich A_n eine Folge paarweise disjunkter Teilmengen von B. Dann gilt

$$\mathbb{P}\Big(\bigcup_n A_n|B\Big) = \frac{\mathbb{P}(\cup_n A_n \cap B)}{\mathbb{P}(B)} = \sum_n \frac{\mathbb{P}(A_n \cap B)}{\mathbb{P}(B)} = \sum_n \mathbb{P}(A_n|B),$$

und somit gilt (i).

Wegen (ii) schreiben wir

$$\begin{aligned}\sum_{n\in\mathbb{N}} \mathbb{P}(A|B_n)\mathbb{P}(B_n) &= \sum_{n\in\mathbb{N}} \mathbb{P}(A \cap B_n) \\ &= \mathbb{P}(A \cap \cup_n B_n) = \mathbb{P}(A \cap \Omega) = \mathbb{P}(A).\end{aligned}$$

□

Definition 3.3. Zwei Ereignisse $A, B \in \mathfrak{F}$, mit $\mathbb{P}(B) > 0$ und $\mathbb{P}(A) > 0$, heissen *unabhängig*, genau dann wenn

$$\mathbb{P}(A|B) = \mathbb{P}(A), \tag{3.1.4}$$

beziehungsweise (was das gleiche ist), wenn

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B). \tag{3.1.5}$$

Allgemeiner heissen n Ereignisse, $A_1, \ldots, A_n$ unabhängig, genau dann, wenn für alle $m \leq n$, und $1 \leq i_1 < i_2 < \ldots < i_m \leq n$ gilt

$$\mathbb{P}\Big(\bigcap_{k=1}^{m} A_{i_k}\Big) = \prod_{k=1}^{m} \mathbb{P}(A_{i_k}) \tag{3.1.6}$$

Anmerkung. Falls $\mathbb{P}(A) = 0$ und $\mathbb{P}(B) > 0$, so gilt stets $\mathbb{P}(A|B) = 0$.

Ein triviales Korollar aus der Definition der bedingten Wahrscheinlichkeit ist die berühmte *Bayes'sche Formel:*

Satz 3.4. *Seien $A, B \in \mathfrak{F}$ und $\mathbb{P}(A) > 0$, $\mathbb{P}(B) > 0$. Dann gilt*

$$\mathbb{P}(B|A) = \mathbb{P}(A|B)\frac{\mathbb{P}(B)}{\mathbb{P}(A)} \tag{3.1.7}$$

Beweis. Der Beweis ist trivial. □

Die Formel ist in der Statistik von grosser Bedeutung. Thomas Bayes (1702 - 1761) (siehe das Bild am Kapitelanfang) hat diesen Satz in seinem Werk *"Essay towards solving a problem in the doctrine of chances"* in einem speziellen Fall hergeleitet. Da Bayes von Beruf Priester war, ist sein Interesse an Wahrscheinlichkeiten wohl rein akademischer Natur gewesen. Ein Beispiel soll

zeigen, dass man aus ihr durchaus nicht völlig intuitive Ergebnisse gewinnen kann.

Beispiel. Ein Test auf Vogelgrippe liefert mit Wahrscheinlichkeit von 99% ein korrektes Ergebnis. Ein bekanntes Pnharmaunternehmen empfiehlt, sich sofort testen zu lassen, und bei positivem Resultat sofort Oseltamivirphosphate prophylaktisch einzunehmen. Für wen ist das sinnvoll?

Wir nehmen dazu an, dass der tatsächliche Durchseuchungsgrad x beträgt. Wir bezeichnen das Ereignis "krank" mit A und das Ereignis "Test richtig" mit B. Dann ist das Ereignis $C =$"positiv auf Vogelgrippe getestet" gegeben durch

$$C = (A \cap B) \cup (A^c \cap B^c)$$

Offenbar gilt

$$\mathbb{P}(A \cap B) = x \times 0.99$$

und

$$\mathbb{P}(A^c \cap B^c) = (1 - x) \times 0.01$$

Insbesondere ist $\mathbb{P}(C) \geq 1\%$, unabhängig vom tatsächlichen Wert von x.

Angenommen nun, eine Versuchsperson sei positiv getestet worden. Wie wahrscheinlich ist es, dass sie auch krank ist? Dazu müssen wir $\mathbb{P}(A|C)$ berechnen. Nach der Formel von Bayes ist dann

$$\begin{aligned} \mathbb{P}(A|C) = \mathbb{P}(C|A)\frac{\mathbb{P}(A)}{\mathbb{P}(C)} = \frac{\mathbb{P}(C \cap A)}{\mathbb{P}(C)} = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(C)} \\ = \frac{x \times 0.99}{x \times 0.99 + (1 - x) \times 0.01}. \end{aligned} \tag{3.1.8}$$

Wenn $x \ll 1$ ist, dann ist im wesentlichen $\mathbb{P}(A|C) = 100\mathbb{P}(A) \ll 1$, d.h. der Test hat eigentlich keine neue Information gebracht, bzw. fast alle positiv getesteten erweisen sich im Nachhinein als gesund....

3.2 Unabhängige Zufallsvariablen

Wir betrachten wieder einen Wahrscheinlichkeitsraum $(\Omega, \mathfrak{F}, \mathbb{P})$. Wir wollen nun den Begriff der von einer Zufallsvariablen erzeugten σ-Algebra einführen.

Definition 3.5. Sei $(\Omega, \mathfrak{F})$ ein Messraum, und $f : \Omega \to \mathbb{R}$ eine messbare Funktion. Sei $\sigma(f)$ die kleinste Unter-σ-Algebra von $\mathfrak{F}$ mit der Eigenschaft dass f bezüglich $\sigma(f)$ messbar ist. Wir sagen $\sigma(f)$ sei die von f erzeugte σ-Algebra.

Die σ-Algebra $\sigma(f)$ kann wie folgt konstruiert werden: Es sei $f^{-1}(\mathfrak{B})$ die Menge aller Urbilder von Elementen der Borel'schen σ-Algebra. Dann ist $\sigma(f)$

die kleinste σ-Algebra, die $f^{-1}(\mathfrak{B})$ enthält. Andererseits sieht man leicht, dass $f^{-1}(\mathfrak{B})$ selbst eine σ-Algebra ist. Daher ist $\sigma(f) = f^{-1}(\mathfrak{B})$.

Definition 3.6. Sei $(\Omega, \mathfrak{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum, und seien X_1, X_2 Zufallsvariablen. X_1 und X_2 heissen *unabhängig*, wenn folgendes gilt: Für jedes Paar von Ereignissen $A \in \sigma(X_1), B \in \sigma(X_2)$ mit $\mathbb{P}(A) > 0, \mathbb{P}(B) > 0$ ist

$$\mathbb{P}(A|B) = \mathbb{P}(A). \tag{3.2.1}$$

Wir sagen in diesem Fall auch: X_1 ist unabhängig von der σ-Algebra $\sigma(X_2)$.

Anmerkung. Da $\sigma(X) = X^{-1}(\mathfrak{B})$, folgt sofort, dass zwei Zufallsvariablen X_1, X_2, genau dann unabhängig sind, wenn für alle Mengen $B_1, B_2 \in \mathfrak{B}$,

$$\mathbb{P}(\{X_1 \in B_1\} \cap \{X_2 \in B_2\}) = \mathbb{P}(\{X_1 \in B_1\})\mathbb{P}(\{X_2 \in B_2\}). \tag{3.2.2}$$

Das folgende Lemma gibt eine alternative Definition der Unabhängigkeit.

Lemma 3.7. *Sei $(\Omega, \mathfrak{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum, und seien X_1, X_2 unabhängige Zufallsvariablen. Seien g_1, g_2 messbare Funktionen von $(\mathbb{R}, \mathfrak{B})$ nach $(\mathbb{R}, \mathfrak{B})$. Es seien ferner $\int_\Omega |g_i(X_i)| \, \mathrm{d}\mathbb{P} < \infty$. Dann gilt*

$$\int_\Omega g_1(X_1) g_2(X_2) \, \mathrm{d}\mathbb{P} = \int_\Omega g_1(X_1) \, \mathrm{d}\mathbb{P} \int_\Omega g_2(X_2) \, \mathrm{d}\mathbb{P} \tag{3.2.3}$$

Beweis. Wir bemerken zunächts, dass unter den Annahmen das Satzes $g_i(X_i)$ messbare Abbildungen von $(\Omega, \sigma(X_i))$ nach $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$ sind. Denn offenbar ist $(g_i(X_i))^{-1}(\mathfrak{B}) = X_i^{-1} \circ g_i^{-1}(\mathfrak{B}) \subset X_i^{-1}(\mathfrak{B}) = \sigma(X_i)$. Wir zeigen als erstes, dass (3.2.3) gilt wenn g_i Indikator-Funktionen sind. Denn für $A_i \in \mathfrak{B}(\mathbb{R})$, $i = 1, 2$, ist

$$\int_\Omega \mathbb{1}_{A_i}(X_i) \, \mathrm{d}\mathbb{P} = \mathbb{P}(X_i \in A_i), \tag{3.2.4}$$

und

$$\begin{aligned} \int_\Omega \mathbb{1}_{A_1}(X_1) \mathbb{1}_{A_2}(X_2) \, \mathrm{d}\mathbb{P} &= \mathbb{P}(\{X_1 \in A_1\} \cap \{X_2 \in A_2\}) \\ &= \mathbb{P}(X_1 \in A_1)\mathbb{P}(X_2 \in A_2) \end{aligned} \tag{3.2.5}$$

was sofort (3.2.3) für diesen Fall liefert.

Als nächstes folgt dann, unter Benutzung der Linearität des Integrals, dass (3.2.3) für alle positiven einfachen Funktionen gilt.

Der entscheidende Schritt ist jetzt, dass der Satz von der monotonen Konvergenz erlaubt, hieraus die Gültigkeit für positive messbare Funktionen zu zeigen. Dazu seien $h_n^{(i)}$, $i = 1, 2$, zwei monoton (in n) wachsende Folgen einfacher Funktionen die punktweise gegen die positiven messbaren Funktionen g_i konvergieren. Somit ist

$$\int_\Omega g_i(X_i)\,\mathrm{d}\mathbb{P} = \lim_{n\to\infty}\int_\Omega h_n^{(i)}(X_i)\,\mathrm{d}\mathbb{P}. \tag{3.2.6}$$

Da auch $h_n^{(1)}(X_1)h_n^{(2)}(X_1)$ eine wachsende Folge positiver einfacher Funktionen ist, die gegen $g_1(X_1)g_2(X_2)$ konvergiert, ist auch

$$\int_\Omega g_1(X_1)g_2(X_2)\,\mathrm{d}\mathbb{P} = \lim_{n\to\infty}\int_\Omega h_n^{(1)}(X_1)h_n^{(2)}(X_2)\,\mathrm{d}\mathbb{P}. \tag{3.2.7}$$

Andererseits ist wegend er Gültigkeit von (3.2.3) für einfache Funktionen,

$$\begin{aligned}(3.2.7) &= \lim_{n\to\infty}\int_\Omega h_n^{(1)}(X_1)h_n^{(2)}(X_2)\,\mathrm{d}\mathbb{P}\\ &= \lim_{n\to\infty}\int_\Omega h_n^{(1)}(X_1)\,\mathrm{d}\mathbb{P}\int_\Omega h_n^{(2)}(X_2)\,\mathrm{d}\mathbb{P}\\ &= \lim_{n\to\infty}\int_\Omega h_n^{(1)}(X_1)\,\mathrm{d}\mathbb{P}\lim_{n\to\infty}\int_\Omega h_n^{(2)}(X_2)\,\mathrm{d}\mathbb{P}.\end{aligned} \tag{3.2.8}$$

Hieraus folgt (3.2.3) sofort.

Zum Schluss zeigt man noch mittels der Zerlegung in positive und negative Teile, dass (3.2.3) auch für allgemeine integrierbare Funktionen gilt. □

Übung. Beweisen Sie den Umkehrschluss zu Lemma 3.7, d.h., wenn (3.2.3) gilt für alle Wahl von g_1, g_2, dann sind X_1 und X_2 unabhängig.

Eine Eigenschaft, die der aus dem Lemma ähnlich sieht, aber deutlich schwächer ist, ist die sogenannte *Unkorreliertheit* von Zufallsvariablen.

Definition 3.8. Sei $(\Omega, \mathfrak{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum, und seien X_1, X_2 Zufallsvariablen. X_1 und X_2 heissen *unkorreliert*, genau dann wenn gilt

$$\int_\Omega X_1X_2\,\mathrm{d}\mathbb{P} = \int_\Omega X_1\,\mathrm{d}\mathbb{P}\int_\Omega X_2\,\mathrm{d}\mathbb{P}. \tag{3.2.9}$$

Offensichtlich ist die Unkorreliertheit viel leichter nachzuprüfen als die Unabhängigkeit. Häufig wird erstere darum auch als erstes Indiz für die Unabhängigkeit benutzt. Allerdings muss man sich klarmachen, dass dieses Indiz keinesfalls schlüssig ist. So seien X, Y zwei unabhängige, gleichverteilte Zufallsvariablen, und $Z_+ \equiv X+Y$, $Z_- \equiv X-Y$. Dann sind Z_+, Z_- unkorreliert. Im allgemeinen sind sie aber nicht unabhängig. Dazu betrachten wir den Fall der Bernoulli Verteilung mit Parameter $p = 1/2$. Dann ist

$$\mathbb{P}(Z_- = 0|Z_+ = 2) = 1 \quad \text{aber} \quad \mathbb{P}(Z_- = 0|Z_+ = 1) = 0,$$

was sofort die Unabhängigkeit falsifiziert.

Anmerkung. Wir werden später sehen, dass es genau eine Verteilungsklasse gibt, in der Unkorreliertheit zur Unabhängigkeit äquivalent ist, nämlich die Gaußverteilungen.

3.3 Produkträume

Unabhängige Zufallsvariablen können wir explizit konstruieren. Dazu betrachten wir zwei Wahrscheinlichkeitsräume, $(\Omega_1, \mathfrak{F}_1, \mathbb{P}_1)$ und $(\Omega_2, \mathfrak{F}_2, \mathbb{P}_2)$ und messbare Funktionen $f_1 : \Omega_1 \to \mathbb{R}$, $f_2 : \Omega_2 \to \mathbb{R}$. Die Idee ist, einen Wahrscheinlichkeitsraum über dem Produktraum $\Omega_1 \times \Omega_2$ zu konstruieren, bezüglich dessen f_1 und f_2 unabhängige Zufallsvariablen sind. Dazu führen wir zunächst die entsprechende σ-Algebra ein.

Definition 3.9. Die *Produkt-σ-Algebra*, $\mathfrak{F}_1 \otimes \mathfrak{F}_2$, ist die kleinste σ-Algebra, die alle Mengen der Form $C = A \times B$ mit $A \in \mathfrak{F}_1, B \in \mathfrak{F}_2$ enthält.

Wir nennen Mengen der Form $A \times B$ gelegentlich Rechtecke, obwohl das etwas irreführend ist. Man beachte, dass die Menge aller Rechtecke ein durchschnittsstabiler Erzeuger der Produkt-σ-Algebra ist, da $(A_1 \times B_1) \cap (A_2 \times B_2) = (A_1 \cap A_2) \times (B_1 \cap B_2)$.

Der nächste Schritt ist die Konstruktion eines W-Maßes auf $(\Omega_1 \times \Omega_2, \mathfrak{F}_1 \otimes \mathfrak{F}_2)$ für das die Unter-σ-Algebren $\mathfrak{F}_1 \times \Omega_2$ und $\Omega_1 \times \mathfrak{F}_2$ unabhängig sind.

Sei $C \in \mathfrak{F}_1 \otimes \mathfrak{F}_2$. Für jedes $x \in \Omega_1$ und jedes $y \in \Omega_2$ führen wir die Mengen

$$C_x \equiv \{y \in \Omega_2 : (x, y) \in C\} \tag{3.3.1}$$

und

$$C^y \equiv \{x \in \Omega_1 : (x, y) \in C\} \tag{3.3.2}$$

ein. Entsprechend definieren wir auch für jede messbare Funktion f auf $\Omega_1 \times \Omega_2$ für jedes $x \in \Omega_1$ die Funktion $f_x(y) \equiv f(x, y)$ und für jedes $y \in \Omega_2$ die Funktion $f^y(x) \equiv f(x, y)$. Dann gilt folgendes:

Lemma 3.10. *Mit den Definitionen von oben gilt:*

(i) Für jedes $C \in \mathfrak{F}_1 \otimes \mathfrak{F}_2$ und $x \in \Omega_1, y \in \Omega_2$ ist $C_x \in \mathfrak{F}_2$ und $C^y \in \mathfrak{F}_1$.
(ii) Für jede messbare Funktion, $f : \Omega_1 \times \Omega_2 \to \mathbb{R}$, und $x \in \Omega_1, y \in \Omega_2$ ist f_x messbar bezüglich $\mathfrak{F}_2$ und f^y messbar bezüglich $\mathfrak{F}_1$.

Beweis. Wir setzen für $x \in \Omega_1$ (für $y \in \Omega_2$ ist das Beweis analog),

$$\mathfrak{C}_x \equiv \{C \in \mathfrak{F}_1 \otimes \mathfrak{F}_2 : C_x \in \mathfrak{F}_2\}.$$

Dann enthält $\mathfrak{C}_x$ sicher die einfachen Mengen $C = A \times B$ mit $A \in \mathfrak{F}_1$ und $B \in \mathfrak{F}_2$. Denn entweder ist dann $x \in A$ und $C_x = B$, oder $x \notin A$ und $C_x = \emptyset$. Beidesmal ist $C_x \in \mathfrak{F}_2$. Nun kann man andererseits leicht nachweisen, dass $\mathfrak{C}_x$ eine σ-Algebra ist. Da dies aber den Erzeuger von $\mathfrak{F}_1 \otimes \mathfrak{F}_2$ enthält, andererseits per Konstruktion nicht grösser als $\mathfrak{F}_1 \otimes \mathfrak{F}_2$ ist, muss $\mathfrak{C}_x = \mathfrak{F}_1 \otimes \mathfrak{F}_2$ gelten.

Weiter ist für jede messbare Menge $D \subset \mathbb{R}$,

$$\begin{aligned} f_x^{-1}(D) &= \{y \in \Omega_2 : f_x(y) \in D\} = \{y \in \Omega_2 : f(x, y) \in D\} \\ &= \{y \in \Omega_2 : (x, y) \in f^{-1}(D)\} = (f^{-1}(D))_x, \end{aligned} \tag{3.3.3}$$

die aber nach (i) in $\mathfrak{F}_2$ liegt. Damit ist das Lemma bewiesen. □

Satz 3.11. *Seien $\mathbb{P}_1, \mathbb{P}_2$ Wahrscheinlichkeitsmaße auf $(\Omega_1, \mathfrak{F}_1)$, bzw. $(\Omega_2, \mathfrak{F}_2)$.*

(i) Dann existiert ein einziges Wahrscheinlichkeitsmaß, $\mathbb{P} \equiv \mathbb{P}_1 \otimes \mathbb{P}_2$, genannt das Produktmaß*, auf der Produkt-σ-Algebra, $\mathfrak{F}_1 \otimes \mathfrak{F}_2$, mit der Eigenschaft, dass für alle $A \in \mathfrak{F}_1$ und $B \in \mathfrak{F}_2$*

$$\mathbb{P}_1 \otimes \mathbb{P}_2(A \times B) = \mathbb{P}_1(A)\mathbb{P}_2(B). \tag{3.3.4}$$

(ii) Wenn $C \in \mathfrak{F}_1 \otimes \mathfrak{F}_2$, so gilt dass

$$\mathbb{P}_1 \otimes \mathbb{P}_2(C) = \int_{\Omega_1} \mathbb{P}_2(C_x)\mathbb{P}_1(\mathrm{d}x) = \int_{\Omega_2} \mathbb{P}_1(C^y)\mathbb{P}_2(\mathrm{d}y). \tag{3.3.5}$$

Beweis. Die Tatsache, dass es nur ein Wahrscheinlichkeitsmass geben kann, dass (3.3.4) erfüllt folgt aus der Tatsache, dass die Mengen der Rechtecke $A \times B$ ein durchschnittstabiles Mengensystem bilden und $\mathfrak{F}_1 \otimes \mathfrak{F}_2$ erzeugen.

Um die Existenz und die zweite Aussage zu beweisen, setzen wir zunächst für $C \in \mathfrak{F}_1 \otimes \mathfrak{F}_2$

$$\mathbb{P}(C) \equiv \int_{\Omega_1} \mathbb{P}_2(C_x)\mathbb{P}_1(\mathrm{d}x). \tag{3.3.6}$$

Dies ist wohldefiniert, wenn $\mathbb{P}_2(C_x)$ messbar bzgl. $\mathfrak{F}_1$ ist. In der Tat ist zunächst $\mathbb{P}_2(C_x)$ wohldefiniert, da $C_x \in \mathfrak{F}_2$ wegen Lemma 3.10. Setzen wir nun

$$\mathfrak{G} \equiv \{C \in \mathfrak{F}_1 \otimes \mathfrak{F}_2 : \mathbb{P}_2(C_x) \text{ ist } \mathfrak{F}_1\text{-messbar}\}. \tag{3.3.7}$$

Für einfache Mengen $C = A \times B$ gilt, dass $\mathbb{P}_2(C_x) = \mathbb{1}_A(x)\mathbb{P}_2(B)$, was offenbar eine $\mathfrak{F}_1$-messbare Funktion ist. Daher sind alle solchen Mengen in $\mathfrak{G}$ enthalten. Wir zeigen noch, dass $\mathfrak{G}$ ein Dynkin-System ist. Wir wissen schon, dass $\Omega_1 \times \Omega_2 \in \mathfrak{G}$. Ferner sieht man aus der Definition, dass $(C^c)_x = (C_x)^c$, und so $\mathbb{P}_2((C^c)_x) = 1 - \mathbb{P}_2(C_x)$, so dass mit C auch $C^c \in \mathfrak{G}$. Weiter ist, wenn $C_i \in \mathfrak{G}$ eine abzählbare Familie disjunkter Mengen sind,

$$\left(\cup_i C_i\right)_x = \cup_i (C_i)_x,$$

wobei auch die $(C_i)_x$ paarweise disjunkt sind. Mithin ist wegen der σ-Additivität

$$\mathbb{P}_2\left(\left(\cup_i C_i\right)_x\right) = \sum_i \mathbb{P}_2\left((C_i)_x\right),$$

was als abzählbare Summe messbarer Funktionen ebenfalls messbar ist. Damit ist $\left(\cup_i C_i\right)_x \in \mathfrak{G}$, und $\mathfrak{G}$ ist ein Dynkin-System dass den durchschnittstabilen Erzeuger von $\mathfrak{F}_1 \otimes \mathfrak{F}_2$ enthält. Also ist $\mathfrak{G} = \mathfrak{F}_1 \otimes \mathfrak{F}_2$. Damit aber sind alle Funktionen $\mathbb{P}_2(C_x)$ messbar bezüglich $\mathfrak{F}_1$, und $\mathbb{P}(C)$ ist durch (3.3.6) wohldefiniert. Wir sehen auch, dass, wenn $C = A \times B$ ist,

$$\mathbb{P}(A \times B) = \mathbb{P}_2(B) \int_{\Omega_1} \mathbb{1}_A(x)\mathbb{P}_1(\mathrm{d}x) = \mathbb{P}_2(B)\mathbb{P}_1(A).$$

Es bleibt zu zeigen, dass $\mathbb{P}$ ein Wahrscheinlichkeitsmass ist. Wir haben aber schon gesehen, dass für disjunkte Familien C_i, $i \in \mathbb{N}$,

$$\begin{aligned}\mathbb{P}(\cup_i C_i) &= \int_{\Omega_1} \mathbb{P}_2((\cup_i C_i)_x)\mathbb{P}_1(\mathrm{d}x) \\ &= \sum_i \int_{\Omega_1} \mathbb{P}_2((C_i)_x)\mathbb{P}_1(\mathrm{d}x) = \sum_i \mathbb{P}(C_i),\end{aligned}$$

d.h. $\mathbb{P}$ ist σ-additiv. Da auch $\mathbb{P}(\Omega_1 \times \Omega_2) = 1$ gilt, ist $\mathbb{P}$ ein W-Maß auf unserem Produktraum, dass der Bedingung (i) des Satzes genügt. Damit ist die Existenz gezeigt. Die alternative Formel in der rechten Seite von (3.3.5) beweist man in völlig gleicher Weise, und die Gleichheit beider Ausdrücke folgt aus der schon bewiesenen Eindeutigkeit. □

Der Punkt ist nun, dass, wenn f_i Zufallsvariablen auf $(\Omega_i, \mathfrak{F}_i)$, $i = 1, 2$, sind, dann sind f_1 und f_2 unabhängige Zufallsvariablen auf dem Wahrscheinlichkeitsraum $(\Omega_1 \times \Omega_2, \mathfrak{F}_1 \otimes \mathfrak{F}_2, \mathbb{P}_1 \otimes \mathbb{P}_2)$ sind. Dies ist die kanonische Konstruktion von unabhängigen Zufallsvariablen.

Es ist offensichtlich, dass durch Iteration die obige Konstruktion auf beliebige endliche Produkte von Wahrscheinlichkeitsmaßen ausgedehnt werden kann.

Beispiel. Wir betrachten das Werfen von n Münzen. Der Zustandsraum jeder Münze ist $\Omega_i = \{0, 1\}$. Dann ist der Zustandsraum der n Würfe $\Omega_1 \times \cdots \times \Omega_n = \{0, 1\}^n$. Jede einzelne Münze hat eine Bernoulliverteillung mit Parameter p. Die Zufallsvariablen $X_1, \ldots, X_n$, wo $X_i(\omega_1, \ldots, \omega_n) = \omega_i$ sind dann unter dem n-fachen Produktmaß unabhängig und gleichverteilt.

Beispiel. Sei $\Omega = \mathbb{R}$, dann ist der $\mathbb{R}^n$ ein Produktraum mit $\mathfrak{B}^n$ der Produkt-Borel-σ-Algebra. Das Gauß'sche Maß mit Dichte

$$\frac{1}{(2\pi)^{n/2} \prod_{i=1}^n \sigma_i} \exp\left(-\sum_{i=1}^n \frac{x_i^2}{2\sigma_i^2}\right)$$

auf $\mathbb{R}^n$ is dann ein Produktmaß. Die Koordinaten des Vektors $X = (x_1, \ldots, x_n)$ sind dann unabhängige Zufallsvariablen.

Unabhängige Zufallsvariablen sind ein wesentlicher Baustein der Wahrscheinlichkeitstheorie. Vielfach wird im alltäglichen Sprachgebrauch der Begriff Unabhängigkeit mit dem der Zufälligkeit gleichgesetzt. So geht man stillschweigend davon aus, dass die sukzessiven Ausgänge eines Roulettspiels unabhängig sind, und wird dies als den zufälligen Charakter des Spiels betrachten.

Beispiel. (Gewinnen mit bedingter Wahrscheinlichkeit). Ein schönes Beispiel, das zeigt wie man Nutzen aus der Kenntnis des Konzepts der bedingten Wahrscheinlichkeit und Produktmaß ziehen kann, ist folgendes Spiel. Alice schreibt zwei Zahlen, auf je einen Zettel. Dann wirft sie eine faire Münze und zeigt Bob je nach Ausgang des Wurfs entweder den einen oder den anderen

Zettel. Nennen wir die gezeigte Zahl im folgenden y und die versteckte Zahl x. Die Aufgabe von Bob besteht darin, zu erraten, ob $x > y$ oder ob $x < y$. Alice bietet Bob eine Wette mit Quote $1 : 2$ an. Soll Bob die Wette annehmen?

Die Antwort auf die Frage ist ja, und zwar weil Bob in der Lage ist, die richtige Antwort mit einer Wahrscheinlichkeit vom mehr als $1/2$ zu geben. Dazu muss er sich nur eine geschickte Strategie ausdenken!

Eine solche Strategie sieht so aus: Bob zieht gemäß einer Gaußverteilung $\mathcal{N}(0, 100)$ eine Zufallszahl, Z. Nun vergleicht er x mit Z: Wenn $Z \geq y$, so rät er $y < x$, wenn $Z < y$ rät er $x < y$.

Um zu sehen, warum das funktioniert, wollen wir das ganze etwas formalisieren. Gegeben sind zwei Zahlen, $x_0 < x_1$. Ferner gibt es eine Bernoulli Zufallsvariable, B, mit Parameter $1/2$, definiert auf einem W-Raum $(\Omega_1, \mathfrak{F}_1, \mathbb{P}_1)$. Die Bob zugängliche Information ist nur die Zufallsvariable $Y = x_B$. Ziel des Spiels ist es, B zu schätzen, denn wenn Bob B kennt, kann es sagen, ob Y gleich x_0 oder x_1 ist, mithin ob es die grössere oder die kleinere Zahl war. Das bedeutet, dass Bob eine neue Zufallsvariable konstruieren will, die von Y abhängt und B voraussagen lässt. Dazu führt der Spieler einen neuen Wahrscheinlichkeitsraum $(\Omega_2, \mathfrak{F}_2, \mathbb{P}_2)$ ein, auf dem er eine Gauß'sche Zufallsvariable, Z konstruiert. Nun betrachten wir den Produktraum, $(\Omega_1 \times \Omega_2, \mathfrak{F}_1 \otimes \mathfrak{F}_2, \mathbb{P} \equiv \mathbb{P}_1 \otimes \mathbb{P}_2)$. Auf diesem sind die Zufallsvariablen B und Z unabhängig. Bob's Strategie ist es, auf diesem Produktraum eine neue Zufallsvariable, A, zu konstruieren, deren Wert nur von (den dem Spieler bekannten Werten von) Z und Y abhängt ist, die aber mit B positiv korreliert in dem Sinne, dass

$$\mathbb{P}(A = B) > 1/2.$$

Die Wahl von A ist

$$A \equiv \mathbb{1}_{Z<Y}$$

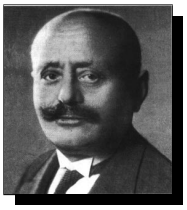
Wir sehen, dass, da Y ja von B abhängt, A und B nicht unabhängig sind. In der Tat ist

Nun können wir benutzen, dass, wenn $B = 1$, $Y = x_1$, und wenn $B = 0$, $Y = x_0$. Also folgt

$$\begin{aligned}
\mathbb{P}(A = B) &= \mathbb{P}(\{Z < Y\} \cap \{B = 1\}) + \mathbb{P}(\{Z \geq Y\} \cap \{B = 0\}) \\
&= \frac{1}{2}\mathbb{P}(\{Z < x_B\}|\{B = 1\}) + \frac{1}{2}\mathbb{P}(\{Z \geq x_B\}|\{B = 0\} \\
&= \frac{1}{2}\left(\mathbb{P}_2(Z < x_1) + \mathbb{P}_2(Z \geq x_0)\right) \\
&= \frac{1}{2} + \frac{1}{2}\mathbb{P}_2(x_0 \leq Z < x_1) > \frac{1}{2}.
\end{aligned}$$

Das wollten wir aber nur zeigen.

3.4 Der Satz von Fubini



Eines der wichtigsten Hilfmittel zur Berechnung komplizierter Integrale auf Produkträumen ist die Vertauschung der Integrationsreihenfolge. Bedingungen die solche Operationen erlauben sind durch ein nach Guido Fubini (19.01.1879–6.06.1943) benanntes Theorem gegeben.

Der erste Schritt ist ein entsprechender Satz für *positive* Funtionen. Hier braucht es erstaunlicherweise gar keine Voraussetzungen.

Satz 3.12 (Fubini-Tonnelli). *Seien $(\Omega_1, \mathfrak{F}_1, \mathbb{P}_1)$ und $(\Omega_2, \mathfrak{F}_2, \mathbb{P}_2)$ zwei Wahrscheinlichkeitsräume, und sei f eine reellwertige, positive, messbare Funktion auf $(\Omega_1 \times \Omega_2, \mathfrak{F}_1 \otimes \mathfrak{F}_2)$. Dann sind die Funktionen*

$$h(x) \equiv \int_{\Omega_2} f(x,y)\mathbb{P}_2(\mathrm{d}y) \quad \textit{und} \quad g(y) \equiv \int_{\Omega_1} f(x,y)\mathbb{P}_1(\mathrm{d}x)$$

messbare bezüglich $\mathfrak{F}_1$ bzw. $\mathfrak{F}_2$, und es gilt

$$\int_{\Omega_1\times\Omega_2} f\,\mathrm{d}(\mathbb{P}_1 \otimes \mathbb{P}_2) = \int_{\Omega_1} h\,\mathrm{d}\mathbb{P}_1 = \int_{\Omega_2} g\,\mathrm{d}\mathbb{P}_2. \tag{3.4.1}$$

Beweis. Wir beginnen mit den Messbarkeitsaussagen. Für $C \in \mathfrak{F}_1 \otimes \mathfrak{F}_2$ und $f = \mathbb{1}_C$ ist haben wir bereits im Beweis von Theorem 3.11 gesehen, dass

$$h(x) = \int_{\Omega_2} f(x,y)\mathbb{P}_2(\mathrm{d}y) = \mathbb{P}_2(C_x)$$

und

$$g(y) = \int_{\Omega_1} f(x,y)\mathbb{P}_1(\mathrm{d}x) = \mathbb{P}_1(C^y)$$

messbar sind wir behauptet. Wegen der Linearität des Integrals folgt dann dasselbe für jede einfache Funktion. Schliesslich stellen wir jede messbare positive Funktion als monotonen Limes von einfachen Funktionen dar und schliesst daraus das Resultat im allgemeinen Fall.

Gleichung (3.4.1) ist im Fall wo f Indikatorfunktion ist schon Teil des Theorems 3.11. Wieder folgt der Fall einfacher Funktionen aus der Linearität und der allgemeine Fall durch Approximation durch monotone Folgen von einfachen Funktionen und der (zweifachen) Anwendung des Satzes von der monotonen Konvergenz.

□

Als nächstes betrachten wir den Fall allgemeiner messbarer Funktionen.

Satz 3.13 (Fubini-Lebesgue). *Sei $f : (\Omega_1 \times \Omega_2, \mathfrak{F}_1 \otimes \mathfrak{F}_2) \to (\mathbb{R}, \mathfrak{B}(\mathbb{R}))$ absolut integrierbar bezüglich des Produktmasses $\mathbb{P}_1 \otimes \mathbb{P}_2$. Dann ist*

(i) $f(x,y)$ *für* $\mathbb{P}_1$*-fast-alle* x *absolut integrierbar bezüglich* $\mathbb{P}_2$*, und umgekehrt.*
(ii) Die Funktionen

$$h(x) \equiv \int_{\Omega_2} f(x,y)\mathbb{P}_2(\mathrm{d}y) \quad \textit{bzw.} \quad g(y) \equiv \int_{\Omega_1} f(x,y)\mathbb{P}_1(\mathrm{d}x)$$

sind wohldefiniert, ausser möglicherweise auf Mengen vom Maß Null bezüglich $\mathbb{P}_1$ *bzw.* $\mathbb{P}_2$*, und absolut integrierbar bezüglich dieser Maße.*
(iii) Es gilt, dass

$$\int_{\Omega_1\times\Omega_2} f \,\mathrm{d}(\mathbb{P}_1 \otimes \mathbb{P}_2) = \int_{\Omega_1} h(x)\mathbb{P}_1(\mathrm{d}x) = \int_{\Omega_2} g(y)\mathbb{P}_2(\mathrm{d}y). \tag{3.4.2}$$

Beweis. Indem wir den vorhergehenden Satz auf die Funktion $|f|$ anwenden, erhalten wir, dass

$$\int_{\Omega_1} \left(\int_{\Omega_2} |f(x,y)|\mathbb{P}_2(\mathrm{d}y) \right) \mathbb{P}_1(\mathrm{d}x) = \int_{\Omega_1\times\Omega_2} |f| \,\mathrm{d}(\mathbb{P}_1 \otimes \mathbb{P}_2) < \infty. \tag{3.4.3}$$

Daher folgt, dass $\int_{\Omega_2} |f(x,y)|\mathbb{P}_2(\mathrm{d}y)$ nur auf einer Menge vom $\mathbb{P}_1$-Maß null nicht endlich sein kann. Hieraus folgt die erste Behauptung.

Indem wir nun f in den positiven und negativen Teil zerlegen und wieder das Resultat von oben verwenden, finden wir sofort, dass $h(x)$ und $g(y)$ wie behauptet messbar sind (als Differenzen entsprechender messbarer Funktionen), wobei wir genau genommen diesen Funktionen einen beliebigen Wert, etwa 0 für diejenigen x (bzw. y) zuschreiben muss, an denen die die absolute Integrierbarkeit nicht gilt. Da dies Nullmengen sind, spielen sie keine Rolle.

Weiter ist

$$\int_{\Omega_1} |h(x)|\mathbb{P}_1(\mathrm{d}x) \le \int_{\Omega_1} \left(\int_{\Omega_2} |f(x,y)|\mathbb{P}_2(\mathrm{d}y) \right) \mathbb{P}_1(\mathrm{d}x) < \infty,$$

so dass auch die behauptete Integrierbarkeit bewiesen ist.

Um schliesslich den Punkt (iii) zu beweisen genügt es zu benutzen, dass

$$\int_{\Omega_1\times\Omega_2} f \,\mathrm{d}(\mathbb{P}_1 \otimes \mathbb{P}_2) = \int_{\Omega_1\times\Omega_2} f_+ \,\mathrm{d}(\mathbb{P}_1 \otimes \mathbb{P}_2) - \int_{\Omega_1\times\Omega_2} f_- \,\mathrm{d}(\mathbb{P}_1 \otimes \mathbb{P}_2)$$

gilt, und den Satz von Fubini-Tonnelli auf beide Terme anzuwenden. □

Anmerkung. In beiden vorgehenden Sätzen ist die Tatsache, dass wir es mit Wahrscheinlichkeitsmaßen zu tun haben nicht wesentlich. Sie gelten auch für allgemeine σ-endliche Maße.

Wenn man sich die Details des Beweises anschaut, sieht man, dass die absolute Integrierbarkeit von f wesentlich benutzt wird. Insbesondere ist andernfalls die Schlussfolgerung im Allgemeinen falsch.

Übung. Zeige, dass der Satz von Fubini für die Funktion $f(x,y) = 2e^{-2xy} - e^{-xy}$ auf $(0,\infty)\times(0,1)$ bezüglich des Lebesguemaßes nicht zutrifft.

3.5 Unendliche Produkte

Natürlich würden wir letztlich gerne von der Verteilung von "beliebig", also "unendlich" vielen Zufallsexperimenten, etwa Münzwürfen, sprechen. Ist das wirklich so schwierig? Wir könnten zunächst geneigt sein, diese Frage zu verneinen. Nehmen wir dazu als einfache Räume Ω_i endliche Mengen (etwa $\Omega_i = \{0,1\}$). Die Frage ist dann, was die geeignete σ-Algebra für den unendlichen Produktraum $\prod_{i=1}^{\infty}\Omega_i$ sein soll. Wir könnten uns vorstellen, wie im Falle endlicher Produkte, die Potenzmenge zuwählen. Ein wenig Nachdenken sollte uns aber skeptisch stimmen: es ist ja bekanntlich so, dass der Raum $\{0,1\}^{\mathbb{N}}$ isomorph zu dem Intervall $[0,1]$ ist (bekanntlich via der Abbildung $\omega \equiv (\omega_1,\omega_2,\ldots) \mapsto \sum_{i=1}^{\infty}\omega_i 2^{-i}$); insbesondere ist stets $\Omega^{\mathbb{N}}$ überabzählbar. Würden wir also einen Wahrscheinlichkeitsraum über $\Omega^{\mathbb{N}}$ mit der σ-Algebra der Potenzmenge konstruieren, so hätten wir implizit dasselbe für die reellen Zahlen getan, was aber auf die bekannten Schwierigkeiten stossen muss. Wir müssen also davon ausgehen, dass wir eine kleinere σ-Algebra konstruieren müssen, ähnlich der Borel σ-Algebra im reellen Fall (in der Tat könnte wir dies sogar via obiger Abbildung genau so tun).

Wir wollen uns bei unserem Vorgehen aber lieber von praktischen Erwägungen leiten lassen. Nun ist es ja so, dass wir auch wenn wir unendlich viele Münzwürfe durchführen wollen, uns stets zunächst für den Ausgang der ersten n davon interessieren, d.h. wie betrachten zunächst jeweils nur endlich viele auf einmal. Das heisst, dass unsere σ-Algebra sicher alle endlichen Produkte von Elementen der σ-Algebren der einfachen Mengen Ω_i enthalten soll. Wir können uns ohne weiteres auf den Standpunkt stellen, dass ausser diesen nur das Unvermeidliche noch dazugenommen werden soll, also dass die σ-Algebra $\mathfrak{B}(\prod_i \Omega_i)$ gerade die von diesen Mengen erzeugte σ-Algebra sein soll.

Definition 3.14. Seien $(\Omega_i,\mathfrak{F}_i)$, $i\in\mathbb{N}$, Messräume, $\widehat{\Omega} \equiv \prod_{i=1}^{\infty}\Omega_i$ der unendlich Produktraum. Dann definieren wir die Produkt-σ-Algebra, $\widehat{\mathfrak{F}}$, über $\widehat{\Omega}$ als die kleinste σ-Algebra, die alle Teilmengen von $\widehat{\Omega}$ der Form

$$A = \bigotimes_{i\in I} A_i \bigotimes_{j\notin I} \Omega_j \tag{3.5.1}$$

enthält, wo $A_i \in \mathfrak{F}_i$ und $I = (i_1,\ldots,i_k) \subset \mathbb{N}$ endlich ist. Die Mengen A der Form (3.5.1) heissen *Zylindermengen*.

Notation: Die Notation in (3.5.1) bedeutet

$$\bigotimes_{i\in I} A_i \bigotimes_{j\notin I} \Omega_j = B_1 \times B_2 \times B_3 \times \cdots \tag{3.5.2}$$

wobei $B_i = A_i$ falls $i \in I$ und $B_i = \Omega_i$ falls $i \notin I$.

Definition 3.15. Seien $(\Omega_i, \mathfrak{F}_i, \mathbb{P}_i)$ Wahrscheinlichkeitsräume. Dann definieren wir das unendliche Produktmaß, $\widehat{\mathbb{P}} \equiv \bigotimes_i \mathbb{P}_i$, auf $(\widehat{\Omega}, \widehat{\mathfrak{F}})$ dadurch, dass für alle Zylindermengen A der Form (3.5.1)

$$\widehat{\mathbb{P}}(A) = \prod_{i\in I} \mathbb{P}_i(A_i). \tag{3.5.3}$$

Die Produkt-σ-Algebra enthält eine äusserst reiche Klasse von Mengen, jedoch ist sie wieder, und zwar selbst in dem Fall, dass Ω endlich ist, kleiner als die Potenzmenge. In der Tat ist sie ihrer Natur nach der Borel'schen σ-Algebra vergleichbar. In der Tat gilt folgender Satz, den wir hier aber nicht beweisen wollen.

Satz 3.16. *Seien Ω_i, $i \in \mathbb{N}$, metrische Räume (etwa $\Omega_i = \mathbb{R}$), und $\mathfrak{B}(\Omega_i)$ die zugehörigen Borel'schen σ-Algebren. Dann kann der unendliche Produktraum $\widetilde{\Omega} \equiv \otimes_i \Omega_i$ mit einer Metrik versehen werden, so dass die Produkt-σ-Algebra die Borel'sche σ-Algebra bezüglich $\widehat{\Omega}$ ist, d.h. es ist die von den offenen Mengen bezüglich der metrischen Topologie erzeugte σ-Algebra.*

In anderen Worten, die Produkt-σ-Algebra enthält alle offenen Mengen (und somit auch alle abgeschlossenen Mengen) bezüglich der *Produktopologie* auf $\widehat{\Omega}$. Für unsere Zwecke heisst das letztlich einfach: keine Angst vor unendlichen Produkträumen, sie sind nicht schlimmer als die reellen Zahlen!

Übung. Benutze den Isomorphismus $I : \{0,1\}^{\mathbb{N}} \to [0,1]$, $I(\omega) = \sum_{i=1}^{\infty} \omega_i 2^{-i}$ und das Beispiel einer nicht-Borel'schen Menge aus Kapitel 2, um eine Menge in $\{0,1\}^{\mathbb{N}}$ zu konstruieren, die nicht in der Produkt-σ-Algebra enthalten ist.

Wir können mittels der Konstruktion unendlicher Produkträume nun unendliche Folgen von Zufallsvariablen konstruieren.

Definition 3.17. Sei $(\Omega, \mathfrak{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum. Dann heisst eine messbare Abbildung, $f : (\Omega, \mathfrak{F}) \to (\mathbb{R}^{\mathbb{N}}, \mathfrak{B}(\mathbb{R}^{\mathbb{N}}))$ eine *Zufallsfolge* oder ein *stochastischer Prozess* (mit diskreter Zeit).

Zur Notation. Ich werde ab sofort der verbreiteten Konvention folgen und das (unspezifierte) Wahrscheinlichkeitsmaß auf dem (abstrakten) Messraum $(\Omega, \mathfrak{F})$, auf dem alle unsere Zufallsvariablen definiert sind, mit $\mathbb{P}$ bezeichnen. Für eine Zufallsvariable auf $(\Omega, \mathfrak{F}, \mathbb{P})$ bezeichnet dann $\mathbb{P}(X \in B)$, die "Wahrscheinlichkeit, dass $X \in B$". Was die Verteilung von X im einzelnen ist, ist dann in der Konstruktion der Zufallsvariablen X kodiert. Im allgemeinen geben wir weder den Raum $(\Omega, \mathfrak{F})$ noch X als Abbildung von Ω nach $\mathbb{R}$ explizit an. Man stellt sich dann auf den Standpunkt, dass es einen Wahrscheinlichkeitsraum gibt, auf dem alle betrachteten Zufallsvariablen konstruiert werden können, so dass ihre gemeinsamen Verteilungen so wie vorgeschrieben sind.

Falls die Verteilung von f, $\mathbb{P} \circ f^{-1}$, ein Produktmaß auf $(\mathbb{R}^{\mathbb{N}}, \mathfrak{B}(\mathbb{R}^{\mathbb{N}}))$ ist, so heisst f eine Folge unabhängiger Zufallsvariablen. Sind die Verteilungen der Komponentenfunktionen darüber hinaus identisch, so heisst die Folge eine Folge unabhängiger, identisch verteilter Zufallsvariablen.

Unendliche Folgen unabhängiger Zufallsvariablen sind die wichtigsten Bausteine der Wahrscheinlichkeitstheorie. Mit ihrer Hilfe können wir insbesondere die Folge der Ergebnisse von (beliebig oft) wiederholten identischen Zufallsexperimenten modellieren, also etwa wiederholte Münzwürfe, Roulettespiele, etc.

3.6 Summen von unabhängigen Zufallsvariablen

Ein weiter Teil der Wahrscheinlichkeitstheorie behandelt die Eigenschaften von Funktionen von unabhängigen Zufallsvariablen. Insbesondere deren Summen, aber auch anderer, wie etwa der Maxima. In der Vorlesung werden wir uns im weiteren ebenfalls weitgehend darauf konzentrieren.

3.6.1 Die Irrfahrt

Gerne betrachten wir eine leichte Abwandlung der Summe S_n: wir wählen statt der Bernoulli-Variablen X_i die (manchmal[1]) sogenannten *Rademacher Variablen*, Y_i, mit der Eigenschaft, dass

$$\mathbb{P}[Y_i = 1] = 1 - \mathbb{P}[Y_i = -1] = p,$$

wobei der Fall $p = 1/2$ von besonderem Interesse ist. In diesem Fall nennen wir die Folge von Zufallsvariablen

$$S_n = \sum_{i=1}^{n} Y_i$$

die *einfache (falls $p = 1/2$ symmetrische) Irrfahrt* auf $\mathbb{Z}$. Beachte dass die Folge S_n, $n \in \mathbb{N}$ selbst wieder eine Zufallsfolge ist, allerdings natürlich keine unabhängigen. S_n ist unser erster *stochastische Prozess* neben unabhängigen Zufallsvariablen.

Das Interesse an S_n ist in natürlicher Weise dadurch begründet, dass es die Entwicklung des Gewinns (oder Verlustes) eines Spielers darstellt, der wiederholt auf den Ausgang von Münzwürfen wettet und dabei jeweils einen

[1] Oft werden auch die folgenden Rademacher Variablen als Bernoulli Variablen bezeichnet.

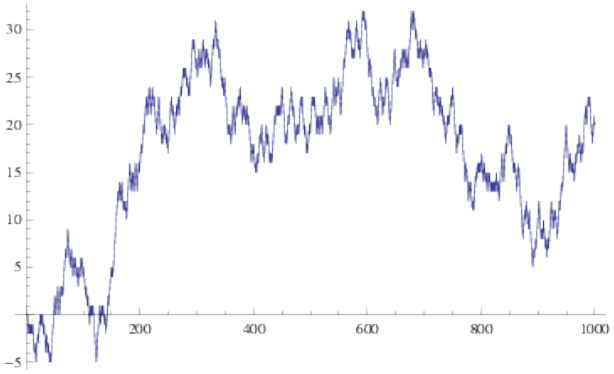


Abb. 3.1 Eine Realisierung der symmetrischen Irrfahrt: Abbildung von $\{(k, S_k), 0 \leq k \leq n = 1000\}$.

festen Betrag, 1, setzt, wobei die Bank ihm im Gewinnfalle den doppelten Betrag auszahlt (d.h., die Bank bewertet das Spiel so, als wäre die Münze fair, also $p = 1/2$).

Unser Formalismus, d.h. die Modellierung von wiederholten Spielen durch unabhängige Zufallsvariablen, erlaubt es uns nun nicht nur einzelne Spiele, sondern ganze Folgen von Spielen zu analysieren. An dieser Stelle ist es vielleicht interessant, zwei Beispiele von Resultaten, die wir damit erhalten können zu betrachten.

Beispiel: Strategien. Ein Spieler könnte versuchen, seine Gewinnchancen in einer Folge von Spielen zu verbessern, indem er in irgendeiner Weise statt immer auf Kopf zu setzen, wahlweise auf Kopf oder Zahl setzt. Eine solche Strategie ist dann gegeben durch eine Folge $a_i \in \{0, 1\}$, $i \in \mathbb{N}$. Gegeben eine solche Strategie ist die Auszahlung im i-ten Spiel

$$r(i) \equiv 2\mathbb{1}_{X_i = a_i} - 1. \tag{3.6.1}$$

Es ist klar, dass, wenn die Folge a_i von vorneherein festgesetzt wird, die $r(i)$ unabhängige Rademachervariablen sind, der akkumulierte Gewinn also die gleiche Verteilung für jede Wahl der Folge a_i hat. Nun könnte aber der Spieler seine Strategie dem Spielverlauf anpassen, d.h. a_k könnte als Funktion der Ausgänge der vorangegangenen Spiele gewählt werden (etwa $a_i = X_{i-1}$), d.h. $a_k = a_k(X_1, \ldots, X_{k-1})$. (Natürlich kann a_k von X_k nur dann abhängen, wenn der Spieler betrügt (bzw. "Insiderwissen" hat)). Interessanterweise ist auch damit nichts gewonnen, und die Auszahlungen $r(i)$ bleiben unabhängige Rademachervariablen.

Satz 3.18. *Sei a_k, $k \in \mathbb{N}$, eine Folge von bezüglich der von den Bernoulli Zufallsvariablen $X_1, \ldots, X_{k-1}$ erzeugten σ-Algebren (im weiteren $\mathfrak{F}_{k-1}$ genannt) messbaren Funktionen. Dann ist die durch (3.6.1) definierte Folge von Zufallsvariablen unabhängig.*

Beweis. Wir müssen nur zeigen, dass die Zufallsvariable $r(k)$ unabhängig von den durch die Zufallsvariablen $r(1), \dots, r(k-1)$ erzeugte σ-Algebra ist. Nun ist aber klar, dass $r(i)$ nur von X_i und a_i abhängt, welche wiederum nur von den $X_1, \dots, X_{i-1}$ abhängen. Damit ist die von $r(1), \dots, r(k-1)$ erzeugte σ-Algebra in $\mathfrak{F}_{k-1}$ enthalten. Sei nun $R_k \in \sigma(r(1), \dots, r(k-1))$. Dann ist

$$\begin{aligned} &\mathbb{P}[r(k) = 1 | R_k] = \mathbb{P}[X_k = a_k | R_k] \qquad (3.6.2)\\ &= \mathbb{P}[X_k = 0 | \{a_k = 0\} \cap R_k]\mathbb{P}[a_k = 0 | R_k] + \mathbb{P}[X_k = 1 | \{a_k = 1\} \cap R_k]\mathbb{P}[a_k = 1 | R_k] \\ &= \frac{1}{2}\mathbb{P}[a_k = 0 | B_k] + \frac{1}{2}\mathbb{P}[a_k = 1 | R_k] = \frac{1}{2} \end{aligned}$$

da nämlich das Ereignis $r(k) = 1$ nur von X_k und a_k abhängt und $\{X_k = i\}$ von $a_k = 1$ und B_k unabhängig sind. Genauso ist

$$\mathbb{P}[r(k) = -1 | R_k] = \mathbb{P}[X_k \neq a_k | R_k] = 1/2 \qquad (3.6.3)$$

was die Aussage beweist. □

3.6.2 Strategien 2. Optionspreise.

Wir kommen im Kontext der Irrfahrt wieder auf unser Problem der Bewertung von Wetten zurück. Dazu betrachten wir eine Summe, S_n, von unabhängigen Rademacher Zufallsvariablen, Y_n, mit Parameter p. Diese stelle den Logarithmus des Wertes einer Aktie zm Zeitpunkt n dar. Das heisst, der Kurs der Aktie

$$W_n = \exp\left(\delta \sum_{i=1}^{n} Y_i\right) = \exp(\delta S_n),$$

wo $\delta > 0$ ein Parameter ist. Eine (europäische) Option ist eine Wette auf den Wert, S_N (bzw W_N), zu einem festen Zeitpunkt N. Der Begeber der Option (etwa eine Bank) verpflichtet sich, dem Optionsinhaber, einen Betrag $f(x) \geq 0$ auszuzahlen, wenn $S_N = x$ (aus Bequemlichkeit denken wir lieber an f als Funktion von S_N). Das Problem besteht darin, zu bestimmen, was der *Wert* der Option ist, d.h. was der niedrigste Preis, V, ist, der es der Bank möglich macht, mit der Option die Option ohne Verlustrisiko verkauft werden kann.

Anmerkung. Klassische "call" bzw. "put" Optionen bestehen in dem Recht, zum Zeitpunkt N die Aktie zum Preis W_c zu kaufen, bzw. zum Preis W_p zu verkaufen. Man sieht, dass dies den Funktionen $F(S_N) = (W_N - W_c)_+$, bzw. $F(S_N) = -(W_p - W_N)_+$ entspricht. Die Theorie der Optionspreisbewertung hat dazu geführt, dass auch viel "exotischere" Optionen angeboten

werden. Dabei hofft der Optionsgeber, dem Kunden eine überteuerte Option verkaufen zu können.

Wie ist das überhhaupt möglich? Um risikofrei wetten zu können, müssen wir in der Lage sein, eine Zufallsvariable zu konstruieren, die mit Sicherheit grösser oder gleich dem Wert der Auzahlung der Option, $F(S_N)$ ist. Genauer gesagt, die Bank verkauft die Option zur Zeit $n = 0$ zum Preis V, und investiert einen Teil dieser Summe, a_0 in die Aktie. Am nächsten Zeitpunkt, $n = 1$, hat sie dann das Kapital $V_1 = V_0 - a_0 + a_0 e^{Y_1 \delta}$; von diesem wird wieder ein Teil, a_1 in die Aktie investiert, und so weiter. Dann entwickelt sich ein Anfangskapital V_0 mit der Zeit wie

$$V_n = V_0 + \sum_{i=1}^{n} a_{i-1}(e^{\delta Y_i} - 1). \tag{3.6.4}$$

Wenn wir also die Option zum Preis V_0 verkaufen, und sicherstellen können, durch geeignete Wahl der a_i am Ende $V_N \geq F(S_N)$ zu erzielen, dann können wir offenbar $F(S_N)$ bezahlen, und haben sogar noch den Betrag $V_N - F(S_N)$ als Gewinn übrig. Man bezeichnet eine solche Reproduktionsstragie auch gerne als *"hedging"*. Der minimale oder "faire" Preis der Option ergibt sich aus der Forderung, dass $V_N = F(S_N)$ gelten soll.

Dass so etwas möglich ist, wollen wir im einfachsten Fall, wo S_N die gewöhnliche Irrfahrt ist, nachprüfen. Wir wollen im Folgenden mit $V_n(x)$ als den "Wert" der Option zum Zeitpunkt n bezeichnen, wenn $S_n = x$ ist.

Dazu betrachten wir zunächst den letzten Zeitschritt. Sei zu diesem Zeitpunkt, $N-1$, sei $S_{N-1} = x$. Sei unser Kapital zu diesem Zeitpunkt K. Dann wollen wir einen Betrag a in die Aktie so investieren, dass unser Kapital zum Zeitpunkt N gerade $F(S_N)$ ist, und zwar unabhängig davon, ob im letzten Schritt die Aktie steigt oder fällt. Das heisst, K und a müssen so gewählt sein, dass

$$f(x+1) = K + a(e^{\delta} - 1), \quad \text{und} \quad f(x-1) = K + a(e^{-\delta} - 1) \tag{3.6.5}$$

gelten. Dieses Gleichungen sind aber leicht zu lösen, mit

$$\begin{aligned} a = a(x) &= \frac{1}{2}\left[f(x+1) - f(x-1)\right] / \sinh \delta \qquad (3.6.6)\\ K = K(x) &= \frac{1}{2}\left[f(x+1) + f(x-1)\right] - a(x)(\cosh \delta - 1) \\ &= \left[\frac{1 - e^{-\delta}}{e^{\delta} - e^{-\delta}} f(x+1) + \frac{e^{\delta} - 1}{e^{\delta} - e^{-\delta}} f(x-1)\right] \end{aligned}$$

$K(x)$ is dann der faire Preis der Option zum Zeitpunkt $N-1$, wenn $S_{N-1} = x$.

Als nächstes können wir berechnen, wieviel Kapital zum Zweitpunkt $N-2$ nötig ist, um zum Zeitpunkt $N-1$ den Betrag $V_{N-1}(S_{N-1})$ zur Verfügung zu haben, wenn wir wissen, dass $S_{N-2} = x$, unabhängig davon was im nächsten

Schritt passiert, d.h. wir müssen im Zeitpunkt $N-2$ eine Strategie fahren, die uns sicherstellt, dass wenn $Y_{N-2} = x$,

$$V_{N-1}(x \pm 1) = V_{N-2}(x) + a_{n-1}(x)(e^{\pm\delta} - 1). \tag{3.6.7}$$

Iterativ folgt, dass

$$a_{j-1}(x) = \frac{1}{2}\left[V_j(x-1) - V_j(x+1)\right] / \sinh\delta \tag{3.6.8}$$

$$V_{j-1}(x) = \left[\frac{1-e^{-\delta}}{e^{\delta}-e^{-\delta}} V_j(x+1) + \frac{e^{\delta}-1}{e^{\delta}-e^{-\delta}} V_j(x-1)\right] \tag{3.6.9}$$

bis wir schliesslich V_0 erreichen.

Beachte, dass die Rekursion für V_j geschlossen ist, und wir a_j nicht notwendig berechnen müssen. Wir können diese in der Form

$$V_{j-1}(x) = \mathbb{E}_{p^*} V_j(x + X_j) \tag{3.6.10}$$

wo $\mathbb{E}_{p^*}$ die Erwartung bezüglich einer neuen Verteilung der Zufallvaiablen X_j ist, für die

$$p^* = \mathbb{P}_{p^*}(X_1 = 1) = \frac{1-e^{-\delta}}{e^{\delta}-e^{-\delta}}, \quad \mathbb{P}_{p^*}(X_1 = -1) = 1 - p^*. \tag{3.6.11}$$

Damit können wir Schlussresultat in der Form

$$V_0 = \mathbb{E}_{p^*} F(S_N) \tag{3.6.12}$$

schreiben, wobei $S_N = \sum_{i=1}^N X_i$ und X_i unabhängige Zufallsvariablen mit Verteilung $\mathbb{P}_{p^*}$ sind.

Wie man leicht nachrechnet, ist diese neue Verteilung dadurch charakterisiert, dass $\mathbb{E}_{p^*} e^{\delta X_i} = 1$ gilt. Die Formel (3.6.12) heisst die *Black-Sholes* Formel in der Optionspreistheorie. Es mag vielleicht noch überraschender sein, dass wir die Formel (3.6.12) auch ohne viel zu rechnen herleiten können. Wir beobachten dazu, dass (3.6.4) mit Koeffizienten a_i die $\mathfrak{F}_i$, messbar sind, also nur von $Y_1, \ldots, Y_i$ abhängen, die einzigen zulässigen Investmentstrategien darstellen. Nehmen wir nun an, dass es möglich ist a_i so zu finden, dass

$$V_N = F(S_N)$$

gilt. Dann ist für jedes Produktmass $\mathbb{P}_p$ mit $\mathbb{P}_p(Y_i = 1) = p$ und $\mathbb{P}_p(Y_i = -1) = 1 - p$,

$$\mathbb{E}_p F(S_N) = \mathbb{E}_p V_N = V_0 + \sum_{i=1}^N \mathbb{E}_p(a_{i-1}) \mathbb{E}_p\left(e^{\delta Y_i} - 1\right).$$

Wählen wir nun $p = p^*$, so erhalten wir

$$\mathbb{E}_{p^*} F(S_N) = V_0.$$

Diese Beobachtung ist viel allgemeiner als unser spezielles Modell für den Aktienkurs. Sie sagt, dass, für jedes Modell mit unabhängigen Zuwächsen des Aktienkurses, für dass es eine zulässige Anlagestrategie gibt, die die Option zur Zeit N exakt reproduziert, gilt die Gleichung (3.6.12) für dasjenige Maß, unter dem die Zuwächse Erwartungswert Null haben. Das Maß $\mathbb{E}_{p^*}$ ist in der Optionspreistheorie als "äquivalentes Martingalmass" bekannt. Beachte, dass der Parameter p der ursprünglichen Verteilung der Zufallsvariablen Y_i nirgendwo eine Rolle gespielt hat!

In dieser zweiten Herleitung der Optionspreisformel wird die Hedging-Strategie a gar nicht mehr berechent. Allerdings setzten wir voraus, dass es eine solche Strategie gibt! Man bezeichnet ein Modell, in dem solche Strategien exisitieren als *vollständigen Märkt*.

Die Größen $V_j(x)$ sind die Werte der Option zum Zeitpunkt j, falls der Aktienkurs zu dieser Zeit gerade $e^{\delta x}$ ist. Wir können diese darstellen als

$$V_j(x) = \mathbb{E}_{p^*}[F(S_N)|S_j = x]. \tag{3.6.13}$$

Übung. Wir haben bisher angenommen, dass das nicht investierte Kapital mit einem Zinssatz Null verzinst wird. Wie ändern sich die obigen Resultate, wenn das nicht in die Aktie inverstierte Kapital mit einem Zinssatz q verzinst wird?

Das hier betrachtete Modell für W_n ist sehr unrealistisch. Tatsächlich aber ist das Grundprinzip, das wir hier dargelegt haben, die Grundlage der modernen Optionspreistheorie.

3.6.3 Das Ruin-Problem

Eine andere Form der Spielstrategie ist es, solange zu spielen, bis entweder ein festgesetzter Gewinn oder Verlust erreicht wird. Wir gehen davon aus, dass ein Spieler ein Anfangskapital $V > 0$ besitzt und nun solange spielt bis er entweder sein Kapital auf $G > V$ vermehrt hat, oder alles verloren hat und nicht mehr weiterspielen kann erreicht ist. Sei also $K(0) = V$ als das Anfangskapital des Spielers. Wir nehmen an, dass nach jedem Spiel das Kapital um einen Betrag $X_i \in \{-1, +1\}$ anwächst, wobei X_i unabhängige, identisch verteilte (Rademacher) Zufallsvariablen mit $\mathbb{P}[X_i = 1] = p = 1 - \mathbb{P}[X_i = -1]$ seien. Dann ist das Kapital des Spielers zum Zeitpunkt n gegeben durch die Zufallsvariable $K(n) = K(0) + S_n$, wo wieder $S_n = \sum_{i=1}^n X_i$.

In einem solchen Spiel können wir die Frage stellen, wie wahrscheinlich es ist, dass die Spielfolge mit dem Ruin des Spielers endet. Wir sehen dass hier die Anzahl der Spiele nicht von vorherein feststeht, wir also wirklich eine Frage im unendlichen Produktraum $\{-1, 1\}^{\mathbb{N}}$ stellen.

Wie können wir das gesuchte Ereignis formal beschreiben: Dazu legen wir zunächst den Wert, n, an dem das Spiel endet fest, und betrachten dann die Vereinigung über alle diese Werte. Wir setzen also

$$A_n = \{S_n = -V\} \bigcap_{k=1}^{n-1} \{-V < S_k < G - V\}$$

und unser gesuchtes Ereignis ist

$$A = \bigcup_{n=1}^{\infty} A_n.$$

Wir sehen sofort an der Konstruktion, dass $A \in \widehat{\mathfrak{F}}$ ist.

Es gibt allerdings eine in mancher Hinsicht einfachere Beschreibung desselben Ereignisses:

$$\begin{aligned} A &= \{\inf\{n : S_n = -V\} < \inf\{n : S_n = G - V\}\} \\ &= \{\inf\{n : K(n) = 0\} < \inf\{n : K(n) = G\}\}. \end{aligned}$$

Mathematisch formuliert sieht unsere Frage wie folgt aus: Was ist $\mathbb{P}[A]$?

Diese Frage sieht zunächst nach einem äusserst üblen kombinatorischen Problem aus. Zum Glück kann man sich das mühsame Zählen sparen, wenn man geschickt vorgeht.

Nun können wir zunächst einmal in Gedanken das erste Spiel ausführen. Mit Wahrscheinlichkeit von je p bzw. $1 - p$ ist nach dem ersten Spiel das Kapital, $K(1)$, des Spielers gleich $K(0)+1$ bzw. $K(0)-1$. Wenn $K(1) = 0$ ist, so ist das Spiel beendet, und A ist eingetreten, während im Falle $K(1) = G$, das Spiel ebenfalls beendet ist, aber A nicht eingetreten ist. In allen Anderen Fällen wird weitergespielt wie zuvor, nur dass jetzt das Anfangskapital $K(1)$ ist. Wir sehen daher, dass es sinnvoll ist, die Wahrscheinlichkeit von A als Funktion des Anfangskapitals einzuführen. Wir setzen dazu

$$h(K) = \mathbb{P}\left(\inf\{n : K(n) = 0\} < \inf\{n : K(n) = G\} \middle| K(0) = K\right), \quad (3.6.14)$$

falls $0 < K < G$; es wird zweckmässig sein $h(0) = 1$ und $h(G) = 0$ zu setzen. Dann ist die gesuchte Wahrscheinlichkeit gegeben durch

$$\mathbb{P}[A] = h(V). \quad (3.6.15)$$

Aus den obigen Überlegungen erhalten wir die Gleichung

$$\begin{aligned} h(K) &= (1-p)\mathbb{1}_{K=1} + (1-p)\mathbb{1}_{K>1}h(K-1) + p\mathbb{1}_{K<G-1}h(K+1) \\ &\quad + 0 \times p\mathbb{1}_{K=G-1} \\ &= (1-p)h(K-1) + ph(K+1), \end{aligned} \quad (3.6.16)$$

für $0 < K < G$. Da die "Randwerte" $h(0) = 1$ und $h(G) = 0$ festgelegt ist, stellt (3.6.16) eine *diskretes Randwertaufgabe* dar, die in Analogie zu der entsprechenden Differentialgleichung auch *Dirichletproblem* genannt wird.

Die Lösung dieser Aufgabe kann man leicht über eine Rekursion erhalten (Übung!). Im einfachsten Fall, wenn $p = 1/2$, ist

$$h(V) = 1 - V/G \tag{3.6.17}$$

wie man leicht nachprüft. Aus (3.6.15) folgt

$$\mathbb{P}[A] = (G - V)/G \quad \text{für } p = 1/2.$$

3.6.4 Das Arcussinusgesetz

Ein interessantes, weil nicht intuitives Resultat über die einfache Irrfahrt ist das sogenannte Arcussinusgesetz. Wir betrachten wieder die Irrfahrt, $S_n = \sum_{i=1}^{n} X_i$, wo X_i unabhängige Rademachervariablen mit Parameter $1/2$ sind. Die Frage, die wir uns stellen wollen ist die nach dem Verhältnis der Zeit, die eine solche Irrfahrt positiv, bzw. negativ ist. Man sollte denken, dass mit grosser Wahrscheinlichkeit diese Zeiten in etwa gleich sind. Tatsächlich aber gilt der folgende Satz.

Wir führen zunächst folgende Variable ein:

$$Y_i \equiv \begin{cases} 1, & \text{falls } S_i > 0 \text{ oder } S_{i+1} > 0, \\ 0, & \text{sonst.} \end{cases} \tag{3.6.18}$$

Wir interpretieren Y_i als Indikator dafür, im i-ten Spiel in der Gewinnzone zu sein.

Satz 3.19. *Sei S_n die einfache symmetrische Irrfahrt. Sei $p_{2k,2n}$ die Wahrscheinlichkeit, bis zur Zeit $2n$ $2k$-mal in der Gewinnzone zu sein, d.h.*

$$p_{2k,2n} = \mathbb{P}\left(\sum_{\ell=1}^{2n} Y_\ell = 2k\right). \tag{3.6.19}$$

Dann gilt

$$p_{2k,2n} = \binom{2k}{k} 2^{-2k} \binom{2n-2k}{n-k} 2^{-2n+2k}. \tag{3.6.20}$$

Beweis. Sei $0 < k < n$. Dann muss die Irrfahrt irgendwann die Null-Linie kreuzen, und dies insbesondere irgendwann zum ersten Mal tun. Sei f_{2r} die Wahrscheinlichkeit, dass die erste Rückkehr der Irrfahrt nach 0 zur Zeit $2r$ passiert,

$$f_{2r} = \mathbb{P}[\inf(i > 0 : S_i = 0) = 2r]. \tag{3.6.21}$$

Bis zu dieser Zeit ist S_i entweder stets positiv, oder stets negativ (ausser natürlich $S_0 = 0$). Beides tritt mit gleicher Wahrscheinlichkeit ein. Im Fall, dass sie bis $2r$ positiv bleibt, kann r nicht grösser sein als k, und im umgekehrten Fall nicht grösser als $n - k$. Nach der ersten Rückkehr nach Null sieht im weiteren alles so aus wie am Anfang, nur dass wir nur noch $2n - 2r$ Schritte zu tun haben. Also haben wir

$$p_{2k,2n} = \frac{1}{2}\sum_{r=1}^{k} f_{2r} p_{2k-2r,2n-2r} + \frac{1}{2}\sum_{r=1}^{n-k} f_{2r} p_{2k,2n-2r}.$$

Wir versuchen diese Rekursion lösen, ohne zunächst f_{2r} zu berechnen. Dazu bemerken wir zunächst, dass

$$\mathbb{P}[S_{2n} = 0] = \frac{1}{2^{2n}}\binom{2n}{n} \equiv u_{2n}.$$

Ausserdem ist

$$u_{2n} = \mathbb{P}[S_{2n} = 0] = \sum_{r=1}^{n} f_{2r}\mathbb{P}[S_{2n-2r} = 0] = \sum_{r=1}^{n} f_{2r} u_{2n-2r}. \tag{3.6.22}$$

Nun können wir unseren Satz per Induktion beweisen.

Wir nehmen an,

$$p_{2k,2m} = u_{2k} u_{2m-2k}$$

gelte für $m \leq n - 1$ und für alle $0 < k < m$. Dann folgt für $m = n$

$$p_{2k,2n} = \frac{1}{2} u_{2n-2k} \sum_{r=1}^{k} f_{2r} u_{2k-2r} + \frac{1}{2} u_{2k} \sum_{r=1}^{n-k} f_{2r} u_{2n-2k-2r},$$

wobei wir die noch unbewiesene Annahme $p_{0,2m} = p_{2m,2m} = u_{2m}$ gemacht haben. Wir werden dies später zeigen. Beide Summen können wir dann mittels (3.6.22) berechnen und erhalten

$$p_{2k,2n} = \frac{1}{2} u_{2k} u_{2n-2k} + \frac{1}{2} u_{2k} u_{2n-2k} = u_{2n-2k} u_{2k},$$

wie behauptet.

Wir müssen nun noch zeigen, dass $p_{0,2n} = p_{2n,2n} = u_{2n}$. Dazu brauchen wir f_{2r} zu berechnen. Die Gleichheit von $p_{0,2n}$ und $p_{2n,2n}$ folgt wegen der symmetrischen Definition der Variablen Y_i. Beachten wir zunächst, dass

$$\mathbb{P}\left[\forall_{1\leq k\leq 2n} S_k > 0\right] = \mathbb{P}\left[\forall_{1\leq k\leq 2n+1} S_k > 0\right], \tag{3.6.23}$$

da S zu einem ungeraden Zeitpunkt $2n + 1$ nicht in der Null sein kann. Andererseits sieht man leicht, dass

$$\mathbb{P}\left[\forall_{1\le k\le 2n+1} S_k > 0\right] = \frac{1}{2}\,\mathbb{P}\left[\forall_{1\le k\le 2n} S_k \ge 0\right], \tag{3.6.24}$$

so dass also

$$\begin{aligned} p_{2n,2n} &= \mathbb{P}\left[\forall_{1\le k\le 2n} S_k \ge 0\right] = 2\,\mathbb{P}\left[\forall_{1\le k\le 2n} S_k > 0\right] \\ &= \mathbb{P}\left[\inf(r>1 : S_r = 0) > 2n\right] = 1 - \sum_{r=1}^{n} f_{2r}. \end{aligned} \tag{3.6.25}$$

Wir müssen also doch f_{2r} berechnen. Dies ist natürlich auch von unabhängigem Interesse.

In Lemma 3.20 zeigen wir, dass

$$f_{2r} = u_{2r-2} - u_{2r}.$$

Dann setzen wir dieses Resultat in (3.6.25) ein, erhalten wir sofort $p_{0,2n} = p_{2n,2n} = u_{2n}$. Damit sind aber unsere Induktionshypothesen bewiesen und der Beweis des Satzes vollständig. □

Lemma 3.20. *Sei S eine symmetrische einfache Irrfahrt und f_{2r} definiert durch (3.6.21). Dann gilt*

$$f_{2r} = \frac{1}{2r} u_{2r-2} = u_{2r-2} - u_{2r}. \tag{3.6.26}$$

Beweis. Wir betrachten dazu zunächst die Wahrscheinlichkeit

$$g_{2n} \equiv \mathbb{P}\left[\forall_{1\le k\le 2n-1} S_k > 0 \wedge S_{2n} = 0\right]. \tag{3.6.27}$$

Es ist aber klar, dass $f_{2n} = 2g_{2n}$. Offenbar ist

$$g_{2n} = \frac{1}{2}\mathbb{P}\left[S_k > 0, \forall_{1\le k\le 2n-2} \wedge S_{2n-1} = 1\right]. \tag{3.6.28}$$

Weiter ist

$$\begin{aligned} &\mathbb{P}\left[S_k > 0, \forall_{1\le k\le 2n-2} \wedge S_{2n-1} = 1\right] \\ &= \mathbb{P}\left[S_1 = 1 \wedge S_{2n-1} = 1\right] \\ &-\mathbb{P}\left[S_1 = 1 \wedge \exists_{1<k\le 2n-2} : S_k \le 0 \wedge S_{2n-1} = 1\right]. \end{aligned} \tag{3.6.29}$$

Der erste Term auf der rechten Seite ist elementar zu berechenen:

$$\mathbb{P}\left[S_1 = 1 \wedge S_{2n-1} = 1\right] = 2^{-2n+1}\binom{2n-2}{n-1}. \tag{3.6.30}$$

Für den zweiten Term benutzen wir eine elementare geometrische Überlegung, die als *Reflektionsprinzip* bekannt ist (siehe Fig. 3.2):

$$\begin{aligned}&\mathbb{P}\left[S_1 = 1 \wedge S_{2n-1} = 1 \wedge \exists_{1<k\leq 2n-2} : S_k \leq 0\right] \\ &= \mathbb{P}\left[S_1 = 1 \wedge S_{2n-1} = -1 \wedge \exists_{1<k\leq 2n-2} : S_k \leq 0\right] \\ &= \mathbb{P}\left[S_1 = 1 \wedge S_{2n-1} = -1\right].\end{aligned} \tag{3.6.31}$$

(Hier ist es wichtig, dass wir den ersten Schritt nach eins festgelegt haben, da dies sicherstellt, dass alle Pfade die in der letzten Wahrscheinlichkeit beitragen, durch die Null gehen müssen!) Die letzte Wahrscheinlichkeit ist wieder

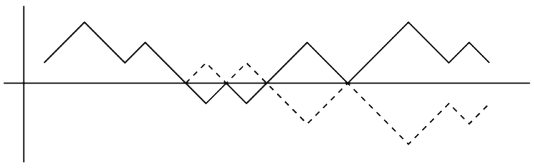


Abb. 3.2 Illustrations des Reflexionsprinzips.

elementar,

$$\mathbb{P}\left[S_1 = 1 \wedge S_{2n-1} = -1\right] = 2^{-2n+1}\binom{2n-2}{n},$$

so dass schliesslich

$$\begin{aligned}f_{2n} = 2g_{2n} &= 2^{-2n+1}\left(\binom{2n-2}{n-1} - \binom{2n-2}{n}\right) \\ &= 2^{-2n+2}\frac{1}{2n}\binom{2n-2}{n-1} = \frac{1}{2n}u_{2n-2}.\end{aligned} \tag{3.6.32}$$

Schliesslich ist $f_{2r} = u_{2r-2} - u_{2r}$ leicht nachzurechnen. □

Asymptotisches Verhältnis (d.h. für grösse n, k). Mittels der Approximation der Binomialkoeffizienten durch die Stirlingformel, d.h. $n! \sim \sqrt{2\pi n}\, n^n e^{-n}$, erhalten wir für grosse n und k

$$p_{2k,2n} \sim \frac{1}{\pi\sqrt{k}\sqrt{n-k}} = n^{-1}\frac{1}{\pi\sqrt{k/n}\sqrt{1-k/n}}.$$

Mithin ist die Wahrscheinlichkeit, dass k/n zwischen $1/2$ und α liegt

$$\begin{aligned}\sum_{n/2\leq k\leq \alpha n} p_{2k,2n} &\sim \frac{1}{\pi n}\sum_{n/2\leq k\leq \alpha n}\frac{1}{\sqrt{k/n}\sqrt{1-k/n}} \\ &\sim \pi^{-1}\int_{1/2}^{\alpha}\frac{\mathrm{d}x}{\sqrt{x(1-x)}} = \frac{2}{\pi}\arcsin\sqrt{\alpha} - \frac{1}{2}.\end{aligned} \tag{3.6.33}$$

So ist die asymptotische Verteilungsfunktion $F(\alpha)$ gegeben durch

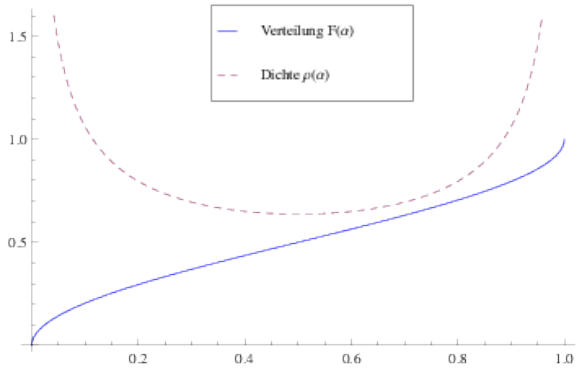


Abb. 3.3 Die Arcussinusverteilung.

$$F(\alpha) \equiv \lim_{n\to\infty} \mathbb{P}(S_k \leq \alpha n \text{ für alle } 1 \leq k \leq n) = \frac{2}{\pi} \arcsin \sqrt{\alpha}$$

und hat Wahrscheinlichkeitsdichte (siehe Fig. 3.3)

$$\rho(\alpha) = \frac{d}{d\alpha} F(\alpha) = \frac{1}{\pi\sqrt{\alpha(1-\alpha)}}.$$

Die Botschaft dieser Rechnung ist, dass die Irrfahrt mit hoher Wahrscheinlichkeit sehr einseitig ist, während der ausgeglichene Fall, halb positiv, halb negativ, kaum vorkommt (siehe Fig. 3.4)).

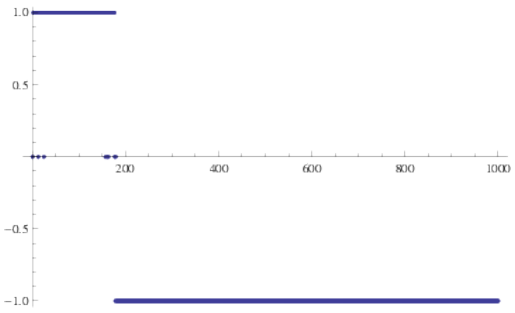


Abb. 3.4 Eine Realisierungen von $\mathrm{sign}\,(\mathrm{S_n})$.

3.6.5 Faltungen

Für die Verteilungsfunktion der Summe zweier unabhängiger Zufallsvariablen ergibt sich in einfacher Weise der folgende Ausdruck. Seien F_X, F_Y, F_{X+Y} die Verteilungsfunktionen der jeweiligen Variablen, dann ist

$$\begin{aligned} F_{X+Y}(a) &= \int_{\mathbb{R}^2} \mathbb{1}_{x+y\le a} \mathrm{d}P_X(x) \otimes \mathrm{d}P_Y(y) = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \mathbb{1}_{x\le a-y} \mathrm{d}P_X(x) \right) \mathrm{d}P_Y(y) \\ &= \int_{\mathbb{R}} F_X(a-y)\, \mathrm{d}P_Y(y) = \int_{\mathbb{R}} F_Y(a-x)\, \mathrm{d}P_X(x). \end{aligned} \tag{3.6.34}$$

Hier haben wir den Satz von Fubini-Tonelli benutzt um das Integral bezüglich des Produktmaßes sukzessive auszuführen. Die letzte Gleichung folgt indem wir die Integrationen bez. x und y in umgekehrter Reihenfolge ausführen.

Wir schreiben die *Faltung* zweier Verteilungsfunktionen F_X und F_Y mit $F_{X+Y} = F_X \star F_Y$.

Wenn die Zufallsvariablen X und Y Verteilungen mit Dichten ρ_X, ρ_Y haben, prüft man leicht nach, dass

$$\rho_{X+Y}(z) = \int_{\mathbb{R}} \rho_X(x)\rho_Y(z-x)\, \mathrm{d}x \tag{3.6.35}$$

gilt.

Man kann sich die Frage stellen, ob es Typen von Verteilungen gibt, die unter der Faltungsoperation invariant bleiben. Solche Verteilungen nennt man *stabil*. Wir werden diese Frage hier nicht im allgemeinen untersuchen, sondern nur ein wichtiges Beispiel betrachten.

Satz 3.21 (Stabilität der Gaußverteilung). *Seien X_1, X_2 zwei unabhängige Gauß'sche Zufallsvariablen mit Varianz σ_1^2, σ_2^2 und Mittelwerten m_1, m_2. Dann ist $X_1 + X_2$ Gaußverteilt mit Mittelwert $m_1 + m_2$ und Varianz $\sigma_1^2 + \sigma_2^2$.*

Beweis. Zum Beweis benutzen wir die Formel (3.6.35) für die Dichte der Faltung. Wir sehen dass

$$\rho_{X_1+X_2-m_1-m_2}(z) = \frac{1}{2\pi\sigma_1\sigma_2} \int_{\mathbb{R}} \mathrm{d}x \exp\left(-\frac{(z-x)^2}{2\sigma_2^2} - \frac{x^2}{2\sigma_1^2} \right). \tag{3.6.36}$$

Nun benutzen wir nur noch, dass

$$\begin{aligned}\frac{(z-x)^2}{\sigma_2^2}+\frac{x^2}{\sigma_1^2} &= \frac{z^2\sigma_1^2+x^2(\sigma_1^2+\sigma_2^2)-2xz\sigma_1^2}{\sigma_2^2\sigma_1^2}\\ &= \frac{z^2\left(\sigma_1^2-\frac{\sigma_1^4}{\sigma_1^2+\sigma_2^2}\right)+(\sigma_1^2+\sigma_2^2)\left(x-\frac{z\sigma_1^2}{\sigma_1^2+\sigma_2^2}\right)^2}{\sigma_1^2\sigma_2^2}\\ &= \frac{z^2}{\sigma_1^2+\sigma_2^2}+\frac{(\sigma_1^2+\sigma_2^2)\left(x-\frac{z\sigma_1^2}{\sigma_1^2+\sigma_2^2}\right)^2}{\sigma_1^2\sigma_2^2}.\end{aligned}$$

Wenn wir diese Gleichung in (3.6.36) einsetzen und die Integration über x ausführen, erhalten wir

$$\rho_{X_1+X_2-m_1-m_2}(z)=\frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}}\exp\left(-\frac{z^2}{2(\sigma_1^2+\sigma_2^2)}\right).$$

Dann wegen

$$\rho_{X_1+X_2}(z)=\rho_{X_1+X_2-m_1-m_2}(z-m_1-m_2)$$

erhalten wir die Dichte einer Gaußverteilung mit Varianz $\sigma_1^2+\sigma_2^2$ und Mittelwert m_1+m_2. □

Korollar 3.22. *Seien $X_i, i\in\mathbb{N}$ unabhängige Gauß'sche Zufallsvariablen mit Varianz σ^2 und Mittelwert 0. Dann hat $n^{-1/2}(X_1+\cdots+X_n)$ dieselbe Verteilung wie X_1.*

Anmerkung. Stabilität einer Klasse von Verteilungen lässt sich auch mit Hilfe der sogenannten characteristische Funktionen (die wir in Kapitel 6 einführen werden, siehe Beispiel 6.5) bestimmen.

Die Gaußverteilung ist nicht die einzige Verteilungsfunktion, die stabil bezüglich Faltung ist: Seien X und Y unabhängig, dann:

- $X\sim\mathrm{Poi}(\lambda_1)$ und $Y\sim\mathrm{Poi}(\lambda_2)\Rightarrow X+Y\sim\mathrm{Poi}(\lambda_1+\lambda_2)$
- $X\sim\mathrm{Bin}(n_1,p)$ und $Y\sim\mathrm{Bin}(n_2,p)\Rightarrow X+Y\sim\mathrm{Bin}(n_2+n_2,p)$
- $X\sim\mathrm{Cauchy}(a_1)$ und $Y\sim\mathrm{Cauchy}(a_2)\Rightarrow X+Y\sim\mathrm{Cauchy}(a_1+a_2)$

Dagegen ist die Exponentialverteilung nicht stabil.

Kapitel 4
Konvergenzbegriffe

Un des points les plus importants de la Théorie des Probabilités, et celui qui prète le plus aux illusions, est la manière dont les probabilités augmentent ou dimunuent par leurs combinaisons mutuelles[a].
Pierre Simon de Laplace, Théorie Analytique des Probabilités

[a] Einer der wichtigsten Punkte in der Wahrscheinlichkeitstheorie, und derjenige, der am meissten Anlass zu Irrglauben gibt, ist die Art, in der Wahrscheinlichkeiten aufgrund ihrer gegenseitigen Verknüpfungen anwachsen oder abnehmen.

Wie immer in der Analysis ist auch in der Wahrscheinlichkeitstheorie der Konvergenzbegriff ein ganz zentrales Konzept. Dabei gibt es einige Besonderheiten, und es ist sinnvoll, sich die Begrifflichkeiten von Anfang an klar zu machen. Wir werden in der Folge dann verschiedene wichtige Beispiele kennenlernen.

4.1 Konvergenz von Verteilungsfunktionen

Wahrscheinlichkeitsmaße waren die ersten Objekte die wie kennengelernt haben. Klarerweise ist die Konvergenz von Folgen von Wahrscheinlichkeitsmaßen nun auch das erste, was wir betrachten müssen.

Wir wollen dafür zunächst nur Wahrscheinlichkeitsmaße auf $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$, also Verteilungen von reellwertigen Zufallsvariablen, betrachten. Wir hatten gesehen, dass diese eindeutig durch ihre Verteilungsfunktionen charakterisiert sind. Daher können wir diese auch zur Definition von Konvergenz heranziehen.

Definition 4.1. Seien F_n, $n \in \mathbb{N}$ eine Folge von Verteilungsfunktionen. Dann konvergiert F_n *schwach* gegen eine Verteilungsfunktion F, genau dann wenn

$$F_n(c) \to F(c), \tag{4.1.1}$$

für alle $c \in \mathbb{R}$ für welche F stetig ist.

Die Einschränkung der Konvergenzforderung auf die Stetigkeitstellen der Funktion F mag zunächst überraschen. Doch wissen wir ja, dass die einzigen Unstetigkeiten von F Sprungstellen sind, an denen F rechtststetig ist. Nun kann man sich leicht Funktionenfolgen konstruieren, die an den

Unstetigkeitstellen nicht konvergieren, oder keinen rechtstetigen Limes haben. Zum Beispiel konvergiert die Folge von Verteilungsfunktionen $F_n(x) = (1 + \tanh(nx))/2$ gegen eine nicht-rechtstetige Funktion

$$\lim_{n\to\infty} F_n(x) = \begin{cases} 0, & \text{für } x < 0, \\ 1/2, & \text{für } x = 0, \\ 1, & \text{für } x > 0. \end{cases}$$

Dann würde man dennoch die rechtsstetige Variante als Limes akzeptieren wollen, d.h. F_n konvergiert schwach gegen $F(x) = \mathbb{1}_{x\geq 0}$.

Schwache Konvergenz von Verteilungsfunktionen ist äquivalent zur schwachen Konvergenz von Wahrscheinlichkeitsmaßen, die wie folgt definiert wird:

Definition 4.2. Sei Ω ein metrischer Raum und $\mathfrak{B}(\Omega)$ die Borel-σ-Algebra. Sei $\mathbb{P}_n$ eine Folge von Wahrscheinlichkeitsmaßen auf $(\Omega, \mathfrak{B}(\Omega))$. Dann konvergiert $\mathbb{P}_n$ *schwach* gegen ein Wahrscheinlichleitsmaß $\mathbb{P}$, genau dann wenn, für alle beschränkten *stetigen* Funktionen g,

$$\int_\Omega g \,\mathrm{d}\mathbb{P}_n \to \int_\Omega g \,\mathrm{d}\mathbb{P}. \tag{4.1.2}$$

Insbesondere gilt:

Satz 4.3. *Sei $\mathbb{P}_n$, $n \in \mathbb{N}$, eine Folge von Wahrscheinlichkeitsmaßen auf $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$ und seien F_n die zugehörigen Verteilungsfunktionen. Dann konvergiert $\mathbb{P}_n$* schwach *gegen ein Wahrscheinlichkeitsmaß $\mathbb{P}$ mit Verteilungsfunktion F genau dann, wenn die Folge F_n schwach gegen F konvergiert.*

Beweis. Wir zeigen zuerst, dass aus $\mathbb{P}_n$ schwach gegen $\mathbb{P}$ konvergiert folgt, dass $F_n(c) \to F(c)$, für alle $c \in \mathbb{R}$ an denen F stetig ist. Dazu definieren wir für jedes $\epsilon > 0$ eine stetige Funktion g_ϵ mit der Eigenschaft

$$\mathbb{1}_{x\leq c} \leq g_\epsilon(x) \leq \mathbb{1}_{x\leq c+\epsilon}$$

(zum Beispier durch lineare Interpolation). Dann gilt

$$F_n(c) \leq \int_{\mathbb{R}} g_\epsilon(x) \,\mathrm{d}\mathbb{P}_n(x) \to \int_{\mathbb{R}} g_\epsilon(x) \,\mathrm{d}\mathbb{P}(x) \leq F(c+\epsilon).$$

Daher ist für jedes $\epsilon > 0$, $\limsup_{n\to\infty} F_n(c) \leq F(c+\epsilon)$. Daraus folgt, da F bei c stetig ist, $\limsup_{n\to\infty} F_n(c) \leq F(c)$. Analog zeigt man, dass $\liminf_{n\to\infty} F_n(c) \geq F(c-\epsilon)$ für jedes $\epsilon > 0$, und so $\lim_{n\to\infty} F_n(c) = F(c)$.

Der Beweis des Umkehrschlusses folgt im Wesentlichen durch Approximation eier stetigen Funktion durch einfache Funktionen. Zunächst bestimmen wir, für beliebiges $\epsilon > 0$, ein beschränktes Intervall $[a, b]$ durch die Forderung $F(a) \leq \epsilon$ und $1 - F(b) \leq \epsilon$. Es gilt dann auch, dass für alle hinreichend grossen n, $F_n(a) \leq 2\epsilon$ und $1 - F_n(b) \leq 2\epsilon$.

Nun sei g ist stetig und daher auf dem beschränkten Intervall $[a,b]$ gleichmässig stetig. Für jedes $\delta > 0$ können wir dann ein $N = N(\delta)$ und Stetigheitsstellen von F, $a_1 = a < a_2 < \ldots < a_N = b$, finden, so dass $\sup_{x\in(a_k,a_{k+1}]} |g(x) - g(a_k)| \leq \delta$. Definiere

$$h(x) = \sum_{k=1}^{N} \mathbb{1}_{(a_k,a_{k+1}]}(x) g(a_k).$$

Dann ist

$$\int_{\mathbb{R}} h(x)\,\mathrm{d}\mathbb{P}_n(x) = \sum_{k=1}^{N} g(a_k)(F_n(a_{k+1}) - F_n(a_k))$$

und daher $\int_{\mathbb{R}} h(x)\mathrm{d}P_n(x) \to \int_{\mathbb{R}} h(x)\mathrm{d}P(x)$. Sei nun g beschränkt, also $|g(x)| \leq M$, für alle $x \in \mathbb{R}$.

$$\begin{aligned}\Big|\int_{\mathbb{R}} (g(x) - h(x))\,\mathrm{d}\mathbb{P}_n(x)\Big| &\leq \Big|\int_a^b (g(x) - h(x))\,\mathrm{d}\mathbb{P}_n(x)\Big| + 2M\mathbb{P}_n([a,b]^c)\\ &\leq \delta + 4M\epsilon\end{aligned}$$

und dasselbe gilt für $\mathbb{P}$ statt $\mathbb{P}_n$. Es folgt nun leicht, dass

$$\limsup_{n\to\infty} \Big|\int_{\mathbb{R}} g(x)\,\mathrm{d}\mathbb{P}_n(x) - \int_{\mathbb{R}} g(x)\,\mathrm{d}\mathbb{P}(x)\Big| \leq 2\delta + 8M\epsilon, \tag{4.1.3}$$

für alle $\epsilon, \delta > 0$. Daraus folgt aber die gewünschte Konvergenz. □

4.2 Konvergenz von Zufallsvariablen

Als nächstes betrachten wir nun die Frage der Konvergenz von Folgen von Zufallsvariablen. Hier ergeben sich interessante neue Begriffe.

4.2.1 Konvergenz in Verteilung

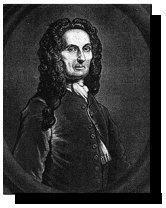
Definition 4.4. Sei $\{X_n\}_{n\in\mathbb{N}}$ eine Folge von (reellen) Zufallsvariablen, wobei X_n auf einem Wahrscheinlichkeitsraum $(\Omega_n, \mathfrak{F}_n, \mathbb{P}_n)$ definiert ist. Dann konvergiert die Folge X_n *in Verteilung* gegen eine Zufallsvariable X,

$$X_n \xrightarrow{\mathcal{D}} X,$$

genau dann, wenn die Verteilungsfunktionen, $F_n(x) \equiv \mathbb{P}(X_n \leq x)$, schwach gegen die Verteilungsfunktion $F(x) \equiv \mathbb{P}(X \leq x)$ der Zufallsvariablen X konvergieren.

Anmerkung. Die schwache Konvergenz einer Folge $X_1, X_2, \ldots$ von Zufallsvariablen gegen eine Zufallsvariable X erfordert nicht, dass diese auf demselben Wahrscheinlichkeitsraum definiert sind.

4.2.1.1 Beispiel: Der Satz von de Moivre-Laplace.



Wir können aus der Definition und der Rechnung, die wir schon bei der Betrachtung von Summen von Zufallsvariablen im Kapitel 3 ausgeführt haben, unsere erste Version des *zentralen Grenzwertsatzes* wie er im 17. Jahrhundert zuerst von de Moivre bewiesen wurde, erhalten.

Satz 4.5 (Der Satz von de Moivre-Laplace). *Seien X_i eine Folge von unabhängigen Bernoullivariablen mit Parameter p. Dann konvergiert die Folge $Z_n \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - p)$ in Verteilung gegen eine Gaußverteilte Zufallsvariable $\mathcal{N}(0, p(1-p))$.*

Beweis. Wir wählen ein Intervall $I = [a, b]$, $a < b \in \mathbb{R}$. Wir wollen zeigen, dass

$$\lim_{n\uparrow\infty} \mathbb{P}\left(Z_n \in I\right) = \frac{1}{\sqrt{2\pi p(1-p)}} \int_a^b e^{-\frac{x^2}{2p(1-p)}} dx. \tag{4.2.1}$$

Wir setzen $S_n \equiv \sum_{i=1}^{n} X_i$. Dann ist $Z_n = \frac{1}{\sqrt{n}}(S_n - pn)$ und

$$\mathbb{P}\left(Z_n \in I\right) = \sum_{k:\frac{1}{\sqrt{n}}(k-pn)\in I} \mathbb{P}\left(S_n = k\right). \tag{4.2.2}$$

Wir müssen also zunächst die Verteilung der Zufallsvariablen Zufallsvariablen S_n. genauer anschauen. Dies lässt sich einfach kombinatorisch lösen:

$$\begin{aligned}\mathbb{P}(S_n = k) &= \sum_{(i_1,\ldots,i_k)\subset(1,\ldots,n)} \mathbb{P}(\forall_{j=1}^{k} X_{i_j} = 1, \forall_{l\notin\{i_1,\ldots,i_k\}} X_l = 0) \\ &= p^k(1-p)^{n-k} \sum_{(i_1,\ldots,i_k)\subset(1,\ldots,n)} 1 = p^k(1-p)^{n-k}\binom{n}{k},\end{aligned} \tag{4.2.3}$$

d.h. S_n ist binomial verteilt mit Parametern n, p.

Für die Binomialkoeffizienten benutzen wir die Stirling'sche Approximation für die Fakuläten. Diese sagt, dass

$$\sqrt{2\pi} n^{n+1/2} e^{-n}(1+1/(12n)) \leq n! \leq \sqrt{2\pi} n^{n+1/2} e^{-n}(1+1/(12n-1)). \tag{4.2.4}$$

Damit gilt

$$\begin{aligned}\binom{n}{k} &= \frac{n!}{(n-k)!k!} = \frac{1}{\sqrt{2\pi}}\sqrt{\frac{n}{(n-k)k}}\,\frac{n^n}{(n-k)^{n-k}k^k} \\ &\quad\times(1+O(1/n)+O(1/k)+O(1/(n-k))) \\ &= \frac{1}{\sqrt{2\pi n}}\sqrt{\frac{1}{(1-k/n)k/n}}\,\frac{1}{(1-k/n)^{n-k}(k/n)^k} \\ &\quad\times(1+O(1/n)+O(1/k)+O(1/(n-k))) \\ &= \frac{1}{\sqrt{2\pi n}}\sqrt{\frac{1}{(1-k/n)k/n}}\left(\frac{1}{(1-k/n)^{1-k/n}(k/n)^{k/n}}\right)^n \\ &\quad\times(1+O(1/n)+O(1/k)+O(1/(n-k)))\,. \end{aligned} \tag{4.2.5}$$

Für die Werte von k, die in der Summe (4.2.2) auftreten sind sowohl k als $n-k$ von der Ordnung n. Daher sind alle Fehlerterme von der Ordnung $O(n^{-1})$.

Setzen wir nun $k/n = x$ und all dies in die Formel (4.2.3) für $\mathbb{P}(S_n = nx)$ ein, so ist

$$\begin{aligned}\mathbb{P}(S_n = nx) &= \frac{1}{\sqrt{2\pi n}}\sqrt{\frac{1}{(1-x)x}}\left(\frac{p^x(1-p)^{1-x}}{(1-x)^{1-x}(x)^x}\right)^n(1+O(n^{-1})) \\ &= \frac{1}{\sqrt{2\pi n}}\sqrt{\frac{1}{(1-x)x}}\exp\left(-nI(p,x)\right)(1+O(n^{-1})) \end{aligned} \tag{4.2.6}$$

wo

$$\begin{aligned} I(p,x) &= \ln\left((x/p)^x[(1-x)/(1-p)]^{1-x}\right) \\ &= x\ln(x/p) + (1-x)\ln((1-x)/(1-p)) \end{aligned} \tag{4.2.7}$$

Folgende einfache Sachverhalte sind leicht nachzuprüfen (Übung!):

(i) $I(p,p) = 0$
(ii) $I(p,x)$ is konvex als Funktion von $x \in (0,1)$ und nimmt ihr einziges Minimum $x = p$ an.
(iii) $\frac{\partial^2 I(p,x)}{\partial x^2} = \frac{1}{x} + \frac{1}{1-x} = \frac{1}{x(1-x)} \geq 4$.
(iv) $I(p,x)$ ist unendlich oft differenzierbar in $x \in (0,1)$.

Wir sehen an den obigen Rechnungen, dass $\mathbb{P}(S_n = nx)$ nur dann nicht exponentiell klein in n wird, wenn x sehr nahe bei p liegt.

Mittels der Taylorformel dritter Ordnung zeigt man nun leicht, dass für alle Werte von k, die in der Summe (4.2.2) auftreten,

$$\left|I(p,k) - \frac{(k/n-p)^2}{2p(1-p)}\right| \leq Cn^{-3/2},$$

wo die Konstante C nur von p, a, b abhängt. Weiter ist für diese Werte

$$\left|\sqrt{\frac{1}{(1-k/n)k/n}} - \sqrt{\frac{1}{p(1-p)}}\right| \leq Cn^{-1/2}.$$

Damit erhalten wir

$$\begin{aligned}
&\mathbb{P}(Z_n \in I) \qquad (4.2.8)\\
&= \sum_{k:\frac{1}{\sqrt{n}}(k-pn)\in I} \frac{1}{\sqrt{2\pi n}}\sqrt{\frac{1}{(1-p)p}}\exp\left(-n\frac{(k/n-p)^2}{2p(1-p)}(1+O(n^{-3/2}))\right)(1+O(n^{-1/2}))
\end{aligned}$$

Wir erkennen die Dichte der Gaußverteilung mit Varianz $\sigma^2 = (1-p)p$. Jetzt brauchen wir nur noch die Summe durch ein Integral zu ersetzen. Dazu bemerkt man wie üblich, dass

$$\begin{aligned}
&\frac{1}{n}\exp\left(-n\frac{(k/n-p)^2}{2p(1-p)}(1+O(n^{-3/2}))\right)(1+O(n^{-1/2})) \qquad (4.2.9)\\
&= \int_{k/n-p}^{(k+1)/n-p} \exp\left(-n\frac{y^2}{2p(1-p)}(1+O(n^{-3/2}))\right)(1+O(n^{-1/2}))\mathrm{d}y,
\end{aligned}$$

da sich der Integrand zwischen den Integrationsgrenzen nur um einen Faktor höchstens der Form $1+O(1/n)$ unterscheidet. Somit haben wir

$$\begin{aligned}
&\sum_{k:\frac{1}{\sqrt{n}}(k-pn)\in I} \frac{1}{\sqrt{2\pi n}}\sqrt{\frac{1}{(1-p)p}}\exp\left(-n\frac{(k/n-p)^2}{2p(1-p)}(1+O(n^{-3/2}))\right)(1+O(n^{-1/2}))\\
&= \int_{a/\sqrt{n}}^{b/\sqrt{n}} \frac{\sqrt{n}}{\sqrt{2\pi p(1-p)}}\exp\left(-n\frac{y^2}{2p(1-p)}(1+O(n^{-3/2}))\right)(1+O(n^{-1/2}))\mathrm{d}y\\
&= \int_a^b \frac{1}{\sqrt{2\pi p(1-p)}}\exp\left(-\frac{x^2}{2p(1-p)}(1+O(n^{-1/2}))\right)(1+O(n^{-1/2}))\mathrm{d}x\\
&\to \int_a^b \frac{1}{\sqrt{2\pi p(1-p)}}\exp\left(-\frac{x^2}{2p(1-p)}\right)\mathrm{d}x \qquad (4.2.10)
\end{aligned}$$

Da dies für jedes Intervall (a,b) gilt, folgt schliesslich auch die Konvergenz der Verteilungsfunktionen. Damit haben wir aber das behauptete Resultat bewiesen. □

Anmerkung. Die Abschätzungen, die wir im Beweis benutzen, sind sogar stärker als das Endresultat. So können wir auch genaue asymptotische Abschätzungen für die Masse von Intervallen geben, deren Länge mit n schrumpft.

4.2.2 Konvergenz in Wahrscheinlichkeit

Ein besonderer Fall liegt vor, wenn die Zufallsvariablen X_n gegen eine deterministische Zufallvariable, also eine Konstante konvergieren, wie wir es etwa im Gesetz der grossen Zahlen sehen werden. Hier benutzen wie gerne auch noch den Begriff der "Konvergenz in Wahrscheinlichkeit":

Definition 4.6. Eine Folge von Zufallsvariablen, $(X_n)_{n\geq 1}$, konvergiert *in Wahrscheinlichkeit* gegen eine Konstante, x, genau dann, wenn, für alle $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - x| > \epsilon) = 0. \tag{4.2.11}$$

Es ist leicht einzusehen, dass eine Zufallsvariable genau dann in Wahrscheinlichkeit gegen eine Konstante x konvergiert, wenn ihre Verteilung gegen die Dirac-Verteilung δ_x konvergiert.

Definition 4.7. Seien $X, X_n, n \in \mathbb{N}$ Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathfrak{F}, \mathbb{P})$. Die Folge $(X_n)_{n\geq 1}$ konvergiert *in Wahrscheinlichkeit* gegen X, falls für alle $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \epsilon) = 0. \tag{4.2.12}$$

4.2.3 Fast sichere Konvergenz

Ein wesentlich stärkerer Konvergenzbegriff für Zufallsvariablen ist allerdings der der sogenannten *fast sicheren* Konvergenz. Wir rufen uns ins Gedächtnis, dass eine Folge von Zufallsvariablen ja eine messbare Funktion von Ω in den Produktraum $\mathbb{R}^{\mathbb{N}}$ ist. Wir können uns also fragen, ob tatsächlich diese Folgen (fast) alle gegen den gleichen Wert x, bzw. eine Zufallsvariable X streben. Hier betrachten wir also wieder einmal Wahrscheinlichkeiten auf dem gesamten unendlichen Produktraum.

Definition 4.8. Sei X_n eine Folge von Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathfrak{F}, \mathbb{P})$. Dann sagen wir, dass X_n *fast sicher* (f.s.) gegen eine Zufallsvariable X konvergiert,

$$X_n \to X \quad \text{f.s.}, \tag{4.2.13}$$

genau dann, wenn

$$\mathbb{P}\left(\lim_{n\to\infty} X_n = X\right) \equiv \mathbb{P}\left(\omega \in \Omega : \lim_{n\uparrow\infty} X_n(\omega) = X(\omega)\}\right) = 1. \tag{4.2.14}$$

Anmerkung. Natürlich kann die Zufallsvariable X auch deterministisch sein, d.h. X kann eine Konstante x sein. Man beachte auch, dass wenn für zwei

Zufallsvariablen gilt, dass $\mathbb{P}(X = Y) = 1$ (d.h. X und Y sind fast sicher gleich, und wenn $X_n \to X$ f.s., dann gilt auch $X_n \to Y$ f.s..

Wir sollten als erstes nachprüfen, ob diese Definition sinnvoll ist, d.h. ob das Ereignis $\{\lim_{n\to\infty} X_n = X\}$ überhaupt in $\mathfrak{B}(\mathbb{R})$ liegt.

Dazu müssen wir das Ereignis $\{\lim_{n\to\infty} X_n = X\}$ unter Verwendung der Definition der Konvergenz ausschreiben:

$$\left\{\lim_{n\to\infty} X_n = X\right\} = \bigcap_{k=1}^{\infty} \bigcup_{n_0=1}^{\infty} \bigcap_{n=n_0}^{\infty} \{|X_n - X| \le 1/k\}\,. \tag{4.2.15}$$

Offenbar ist jeder Klammerausdruck $\{|X_n - X| \le 1/k\}$ eine Borelmenge, und somit auch die abzählbaren Durchschnitte und Vereinigungen davon, so dass also unsere Frage Sinn macht.

In Worten lautet die rechte Seite von (4.2.15): "Für alle $k \in \mathbb{N}$ ist, bis auf endlich viele Werte von n, $|X_n - X| \le 1/k$". Das komplementäre Ereignis ist dann "Es gibt k so, dass für unendlich viele Werte des Indexes n, $|X_n - X| > 1/k$ gilt". Damit ist

$$\mathbb{P}\left(\lim_{n\to\infty} X_n = X\right) = 1 - \mathbb{P}\left(\cup_k \{|X_n - X| > 1/k \text{ für unendlich viele } n\}\right) \tag{4.2.16}$$

Üblicherweise benutzt man die Notation

$$\{A_n, \text{ u.o.}\} \equiv \{A_n \text{ für unendlich viele } n\} \equiv \{\cap_{n_0<\infty} \cup_{n\ge n_0} A_n\}\,, \tag{4.2.17}$$

wo $A_n \in \mathfrak{F}$ eine Folge von Ereignissen ist. Somit ist $\mathbb{P}\left(\lim_{n\to\infty} X_n = X\right) = 1$ genau dann, wenn $\mathbb{P}\left(\cup_k \{|X_n - X| > 1/k, \text{ u.o.}\}\right) = 0$. Da aber

$$\begin{aligned} \sum_{k\in\mathbb{N}} \mathbb{P}\left(\{|X_n - X| > 1/k, \text{ u.o.}\}\right) &\ge \mathbb{P}\left(\cup_k \{|X_n - X| > 1/k, \text{ u.o.}\}\right) \\ &\ge \max_{k\in\mathbb{N}} \mathbb{P}\left(\{|X_n - X| > 1/k, \text{ u.o.}\}\right) \end{aligned} \tag{4.2.18}$$

sehen wir, dass folgendes Lemma gilt:

Lemma 4.9. *Sei X_n eine Folge von Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathfrak{F}, \mathbb{P})$. Dann ist*

$$\mathbb{P}\left(\lim_{n\to\infty} X_n = X\right) = 1 \quad \Longleftrightarrow \quad \forall k \in \mathbb{N},\, \mathbb{P}\left(\{|X_n - X| > 1/k, \text{ u.o.}\}\right) = 0. \tag{4.2.19}$$

Letztere Frage kann nun mit einem der wichtigsten Lemma der Wahrscheinlichkeitstheorie entschieden werden, dem sogenannten *Borel-Cantelli Lemmas.*

Lemma 4.10 (Erstes Borel-Cantelli Lemma). *Sei $(\Omega, \mathfrak{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum, und seien $A_n \in \mathfrak{F}$ eine Folge von Ereignissen. Wenn $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, dann gilt*

$$\mathbb{P}(A_n, \text{ u.o.}) = 0. \tag{4.2.20}$$

Lemma 4.11 (Zweites Borel-Cantelli Lemma). *Sei $(\Omega, \mathfrak{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum, und seien $A_n \in \mathfrak{F}$ eine Folge von* unabhängigen *Ereignissen. Wenn $\sum_{n=1}^{\infty} P(A_n) = +\infty$, dann gilt*

$$\mathbb{P}(A_n, \text{ u.o.}) = 1. \tag{4.2.21}$$

Beweis. Wir beweisen zunächst das wichtigere erste Borel-Cantelli Lemma. Wir haben

$$\mathbb{P}(A_n, \text{ u.o.}) = \mathbb{P}\left(\cap_{k=1}^{\infty} \cup_{n\geq k} A_n\right) = \lim_{k\to\infty} \mathbb{P}\left(\cup_{n\geq k} A_n\right) \leq \lim_{k\to\infty} \sum_{n=k}^{\infty} \mathbb{P}(A_n). \tag{4.2.22}$$

Nun ist nach Voraussetzung die Reihe $\sum_{n=1}^{\infty} \mathbb{P}(A_n)$ konvergent, woraus folgt, dass dass die Folge $r_k \equiv \sum_{n=k}^{\infty} \mathbb{P}(A_n)$ eine Nullfolge ist. Damit ist die Aussage des Lemma evident.

Beweisen wir nun noch das zweite Lemma. Wieder ist

$$\mathbb{P}(A_n, \text{ u.o.}) = \mathbb{P}\left(\cap_{k=1}^{\infty} \cup_{n\geq k} A_n\right) = \lim_{k\to\infty} \mathbb{P}\left(\cup_{n\geq k} A_n\right). \tag{4.2.23}$$

Aber

$$\begin{aligned} 0 \leq 1 - \mathbb{P}\left(\cup_{n\geq k} A_n\right) &= \mathbb{P}\left(\left(\cup_{n\geq k} A_n\right)^c\right) = \mathbb{P}\left(\cap_{n\geq k} A_n^c\right) \qquad (4.2.24)\\ &= \lim_{N\to\infty} \mathbb{P}\left(\cap_{N\geq n\geq k} A_n^c\right) \overset{\text{unab.}}{=} \lim_{N\to\infty} \prod_{n=k}^{N} \mathbb{P}\left(A_n^c\right)\\ &= \prod_{n=k}^{\infty} \left(1 - \mathbb{P}\left(A_n\right)\right) \leq \exp\Big(-\sum_{n=k}^{\infty} \mathbb{P}\left(A_n\right)\Big) = 0 \end{aligned}$$

da ja für jedes k, $\sum_{n=k}^{\infty} \mathbb{P}\left(A_n\right) = +\infty$ ist. Ausserdem haben wir hier noch die (auch sonst) sehr nützliche Abschätzung

$$1 - x \leq e^{-x} \tag{4.2.25}$$

benutzt. Damit ist für alle $k < \infty$ $\mathbb{P}\left(\cup_{n\geq k} A_n\right) = 1$ und somit auch $\lim_{k\uparrow\infty} \mathbb{P}\left(\cup_{n\geq k} A_n\right) = 1$. Daraus folgt (4.2.21). □

Wir können diese Lemmata sofort auf die Frage der fast sicheren Konvergenz anwenden.

Korollar 4.12. *Eine Folge von Zufallsvariablen X_n konvergiert fast sicher gegen eine Zufallsvariable X, wenn für alle $\epsilon > 0$*

$$\sum_{n=1}^{\infty} \mathbb{P}(|X_n - X| > \epsilon) < \infty. \tag{4.2.26}$$

Wenn X_n eine Folge von unabhängigen Zufallsvariablen ist, so ist die Bedingung (4.2.26) auch notwendig.

Beweis. Wir haben zu gesehen, dass X_n genau dann fast sicher gegen x konvergiert, wenn für alle $1 \leq k < \infty$, $\mathbb{P}(|X_n - x| > 1/k, \text{u.o.}) = 0$. Wegen dem ersten Borel-Cantelli Lemma gilt dies aber wegen (4.2.26). Die Notwendigkeit folgt aus dem zweiten Borel-Cantelli Lemma. □

Wir sehen aus dem Korollar leicht, dass es möglich ist, dass eine Folge von Zufallsvariablen in Wahrscheinlichkeit gegen eine Konstante x konvergiert, nicht aber fast sicher. Das einfachste Beispiel ist durch eine Folge von unabhängigen Zufallsvariables X_n gegeben, bei denen

$$\mathbb{P}(X_n = 0) = 1 - n^{-\alpha} \quad \text{und} \quad \mathbb{P}(X_n = 1) = n^{-\alpha}.$$

Diese Folge konvergiert für jedes $\alpha > 0$ in Wahrscheinlichkeit gegen 0, aber nur für $\alpha > 1$ tut sie das auch fast sicher.

Anmerkung. Die fast sichere Konvergenz ist die stärkste Konvergenzform: Wenn $X_n \to X$, f.s., dann konvergiert X_n auch in Wahrscheinlichkeit gegen X. Wenn X_n in Wahrscheinlichkeit gegen X konvergiert, so konvergiert X_n auch in Verteilung gegen X. Die umgekehrten Schlüsse gelten nicht.

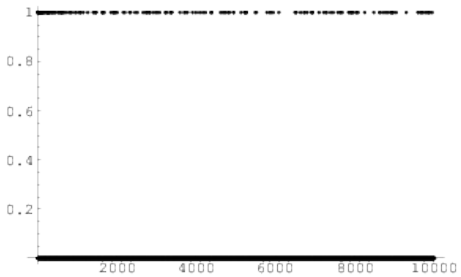


Abb. 4.1 Folge von Bernoullivariablen mit $p_n = 1/\sqrt{n}$.

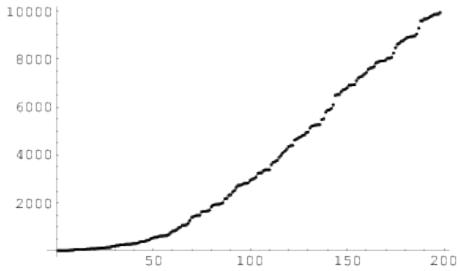


Abb. 4.2 Folge der Werte n mit $X_n = 1$, mit $p_n = 1/\sqrt{n}$.

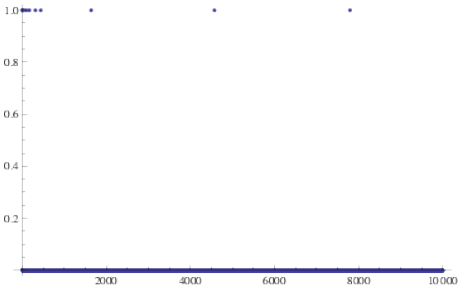


Abb. 4.3 Folge von Bernoullivariablen mit $p_n = n^{-1}$.

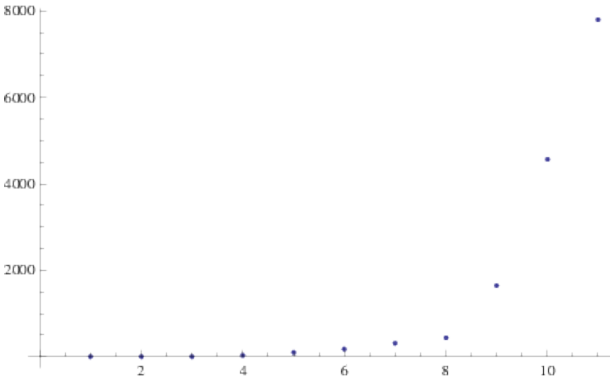


Abb. 4.4 Folge der Werte n mit $X_n = 1$, mit $p_n = n^{-1}$.

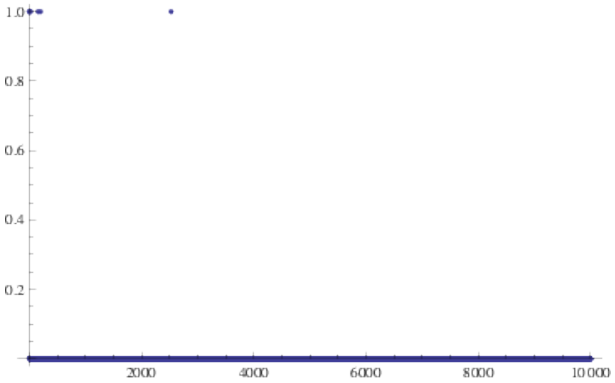


Abb. 4.5 Folge von Bernoullivariablen mit $p_n = n^{-1.1}$.

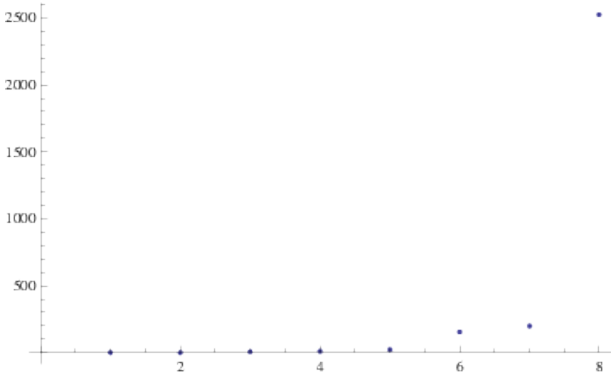


Abb. 4.6 Folge der Werte n mit $X_n = 1$, mit $p_n = n^{-1.1}$.

Kapitel 5
Das Gesetz der großen Zahlen.

Au milieu des causes variables et inconnues que nous comprenons sous le nom de hazard*, et qui rendent incertaine et irrégulière la marche des événements, on voit naître, à mesure qu'ils se multiplient, une régularité frappante, qui semble tenir à un dessein et que l'on a considérée comme une preuve de la providence*[a]*,*
Pierre Simon de Laplace, Théorie Analytique des Probabilités

[a] Inmitten der veränderlichen und unbekannten Ursachen, die wir unter dem Namen *Zufall* verstehen, und die den Ablauf der Ereignisse unsicher und irregulär machen, sieht man, während ihre Zahl verfielfacht eine frappierende Regularität zum Vorschein kommen, die sich an einem Plan zu halten scheint und die man als einen Beweis der Vorsehung betrachtet hat.

Das zentrale Anliegen dieser Sektion ist die Behandlung des wohl fundamentalsten Satzes der Wahrscheinlichkeitstheorie, des Gesetzes der großen Zahlen. Dieses begründet insbesondere den Zusammenhang zwischen Wahrscheinlichkeit und Frequenz, und erklärt die Bedeutung des Erwartungswertes als Mittel über wiederholte Zufallsexperimente. Im weiteren Sinne ist das Gesetz der großen Zahlen unsere erste Begegnung mit dem Prinzip, dass aus völlig zufälligen Ereignissen dennoch völlig deterministische Resultate folgen können.

5.1 Erwartungswert, Varianz, Momente

Sei X eine reelle Zufallsvariable auf $(\mathbb{R}, \mathfrak{B}, \mathbb{P})$ mit Verteilungsfunktion

$$F(x) \equiv \mathbb{P}\left(X \leq x\right).$$

Grundsätzlich haben wir ja gesehen, dass diese durch ihre Verteilungsfunktion die Zufallsvariable vollständig charakterisiert. Wir sind aber vielfach an alternativen, einfacheren Kenngrössen interessiert, und insbesondere für statistische Anwendungen möchten wir einige wenige bedeutungsvolle Parameter identifizieren, die die Eigenschaft einer Verteilung bestimmen.

Wir hatten bereits gesehen dass der Erwartungswert von X gegeben ist durch

$$\mathbb{E}X \equiv \int_{\mathbb{R}} x \, \mathrm{d}P_X(x). \tag{5.1.1}$$

wo $P_X \equiv \mathbb{P} \circ X^{-1}$ die Verteilung von X ist. Die Bedeutung der Erwartung ist ziemlich offensichtlich. Im weiteren möchte man natürlich wissen, wie sehr sich die Verteilung um diese Erwartung herum streut.

Die erste naheliegende Grösse ist die sogenannte *Varianz*,

$$\mathrm{var}(X) \equiv \mathbb{E}(X - \mathbb{E}X)^2 \tag{5.1.2}$$

Man bezeichnet im übrigen die Quadratwurzel der Varianz als *Standardabweichung*. Beachte, dass die Varianz einer Zufallsvariablen unendlich sein kann, auch wenn die Erwartung endlich ist.

Momente.

Eine naheliegende Verallgemeinerung der Varianz sind die sogenannten *Momente* eine Wahrscheinlichkeitsverteilung. Wir definieren

$$M_p \equiv \mathbb{E}X^p \tag{5.1.3}$$

Momente spielen auch deswegen eine äusserst wichtige Rolle, weil in vielen, *aber nicht allen* (!) Fällen die Kenntnis aller Momente einer Wahrscheinlichkeitsverteilung diese vollständig bestimmen. Ohne im Detail auf diese Fragen eingehen zu wollen, ist es nützlich folgendes Kriterium zu kennen:

Satz 5.1. *Sei $M_n \in \mathbb{R}$ eine Folge von Zahlen mit der Eigenschaft, dass für $p \in \mathbb{N}$ $M_{2p} \geq 0$ und es $a > 0$ gibt, so dass*

$$\sum_{p=1}^{\infty} M_{2p} \frac{a^{2p}}{(2p)!} < \infty. \tag{5.1.4}$$

Dann existiert höchstens ein Wahrscheinlichkeitsmaß auf $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$, so dass $M_n = \int_{\mathbb{R}} x^n dP$, für alle $n \in N$.

Anmerkung. Die Aussage von Satz 5.1 impliziert, dass fall die Momente M_p einer Zufallsvariablen die Bedingung (5.1.4) erfüllen, dann legen diese die Verteilung der Zufallsvariablen eindeutig fest.

Beispiel 5.2. Für $X \sim \mathcal{N}(0,1)$, $M_p = 0$ für ungerade p und sonst $M_{2p} = (2p)!/(2^p p!)$, $p \in \mathbb{N}$. Deshalb gilt (5.1.4) für alle $a \in \mathbb{R}$.

Erzeugende Funktionen.

Eng mit den Momenten verküpft, haüfig aber weit nützlicher, ist die sogenannte *Momenten erzeugende Funktion*, oder *Laplace Transformierte.* Diese

ist definiert durch

$$\psi(z) \equiv \mathbb{E}(e^{zX}). \tag{5.1.5}$$

Natürlich muss $\psi(z)$ für $z \neq 0$ nicht notwendig endlich sein. Wenn es $h > 0$ gibt, so dass $\psi(\pm h) < \infty$, dann existiert $\psi(z)$ für alle $|z| \leq h$, ist unendlich oft differenzierbar für $z < |h|$ und es gilt, dass

$$M_p = \frac{d^p}{dz^p}\psi(z = 0),$$

d.h. aus ψ können alle Momente berechnet werden.

Beispiel 5.3. Hier ist eine Liste von momentenerzeugende Funktionen wichtiger Verteilungen.

- Für $X \sim \mathcal{N}(m, \sigma^2)$, gilt $\psi(z) = \exp(\sigma^2 z^2/2 + zm)$.
- Für $X \sim \mathrm{Exp}(a)$, gilt $\psi(z) = 1/(1 - z/a)$ für $|z| < a$.
- Für $X \sim \mathrm{Poi}(\lambda)$, gilt $\psi(z) = \exp(-\lambda(e^z - 1))$.
- Für $X \sim \mathrm{Geo}(q)$, gilt $\psi(z) = (1 - q)/(1 - qe^z)$ für $|z| < \ln(1/q)$.
- Für $X \sim \mathrm{Bin}(n, p)$, gilt $\psi(z) = (1 - p + p\, e^z)^n$.
- Für $X \sim \mathrm{Cauchy}(a)$ ist $\psi(z) = \infty$ für alle $z \neq 0$.

5.2 Chebychev's Ungleichung

Die Bedeutung von Varianz, Momenten und erzeugenden Funktionen erschliesst sich zum Teil aus der sogenannten *Chebychev Ungleichung*.

Lemma 5.4. *Sei X eine reellwertige Zufallsvariable mit Verteilung $\mathbb{P}$. Dann gilt, für alle $x > 0$*

$$\mathbb{P}(|X - \mathbb{E}X| > x) \leq \frac{\mathrm{var}(X)}{x^2}. \tag{5.2.1}$$

Beweis. Wir können ohne Verlust der Allgemeinheit annehmen, dass $\mathbb{E}X = 0$. Dann ist, für alle $x > 0$,

$$\mathbb{P}(|X| > x) = \mathbb{E}(\mathbb{1}_{|X|>x}) \leq \mathbb{E}\left(\mathbb{1}_{|X|>x}\frac{X^2}{x^2}\right) \leq \mathbb{E}\left(\frac{X^2}{x^2}\right) = \frac{\mathrm{var}(X)}{x^2},$$

was zu beweisen war. □

Die Herleitung dieser Ungleichung mag diese auf den ersten Blick völlig absurd wirken lassen. Allerdings steht der Nutzen der Ungleichung in keinem Verhältnis zu der Schwierigkeit ihres Beweises. Der Punkt ist die große Universalität der Aussage, die wesentliche Informationen aus nur einer relative leicht berechenbaren Kenngrösse einer Verteilung zu ziehen erlaubt.

Der singulär einfache Beweis läd natürlich dazu ein, eine allgemeinere Ungleichung herzuleiten:

Lemma 5.5. *Sei X eine rellwertige Zufallsvariable mit Verteilung $\mathbb{P}$, und sein $f : \mathbb{R} \to \mathbb{R}_+$ eine monoton wachsende Funktion. Dann gilt für alle x,*

$$\mathbb{P}(X > x) \leq \frac{\mathbb{E}f(X)}{f(x)}. \tag{5.2.2}$$

Beweis. Für alle x,

$$\mathbb{P}(X > x) = \mathbb{E}\mathbb{1}_{X>x} \leq \mathbb{E}\mathbb{1}_{X>x}\frac{f(X)}{f(x)} \leq \frac{\mathbb{E}f(X)}{f(x)},$$

was zu beweisen war. □

Die allgemeinere Ungleichung ist natürlich nur dann nützlich, wenn $\mathbb{E}f(X)$ nicht nur endlich, sondern auch berechenbar ist. Typischerweise wird die Markov-Ungleichung für die Fälle $f(x) = |x|^p$ und $f(x) = \exp(tx)$ gerne verwendet. Insbesondere der letzte Fall ist von großer Wichtigkeit, und bildet die Grundlage der sogenannten *Theorie der großen Abweichungen.*

Korollar 5.6. *Sei X eine rellwertige Zufallsvariable. Dann gilt*

$$\mathbb{P}(X \geq x) \leq \inf_{t\geq 0} e^{-tx}\mathbb{E}(e^{tX}). \tag{5.2.3}$$

Diese Abschätzung ist natürlich nur dann nützlich, wenn $\mathbb{E}e^{tX}$ zumindest für kleine positive t endlich ist.

Die besondere Stärke dieser Ungleichung erweist sich wenn man Summen unabhängiger Zufallsvariablen betrachtet:

Korollar 5.7. *Sei X_i eine Familie unabhängiger Zufallsvariablen. Dann gilt*

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq x\right) \leq \inf_{t\geq 0} e^{-tx}\prod_{i=1}^{n}\mathbb{E}(e^{tX_i}). \tag{5.2.4}$$

Das Produkt ist dabei oft leicht zu berechnen. Insbesondere im Fall identisch verteilter Zufallsvariablen ergibt sich ein sehr einfacher Ausdruck. Betrachten wir als Beispiel unabhängige Rademachervariablen mit Parameter 1/2 (d.h. $\mathbb{P}(X = \pm) = 1/2$). Dann ist

$$\begin{aligned}\mathbb{P}\left(n^{-1}\sum_{i=1}^{n} X_i \geq x\right) &\leq \inf_{t\geq 0} e^{-txn}(\cosh t)^n \\ &= \left[\exp\left(\inf_{t\geq 0}(-tx + \ln\cosh(t))\right)\right]^n = e^{-nI(x)}\end{aligned}$$

wo $I(x) = \frac{(1-x)}{2}\ln(1-x) + \frac{(1+x)}{2}\ln(1+x)$. Um dieses Ergebnis zu erhalten bemerkt man, dass das Minimum der Funktion $-tx + \ln\cosh t$ angenommen wird, wenn $\tanh(t) = x$. Da $\tanh^{-1}(x) = \frac{1}{2}\ln\frac{1+x}{1-x}$ ist, folgt dies nach einigen elementaren Rechnungen. Man vergleiche mit dem exakten Wert!!

5.3 Das Gesetz der großen Zahlen

In diesem Abschnitt werden wir den vielleicht wichtigsten Satz der Wahrscheinlichkeitstheorie beweisen, das sogenannte *starke Gesetz der großen Zahlen*. Das Gesetz der großen Zahlen macht für den Fall des Modells von unabhängigen Zufallsvariablen den Zusammenhang zwischen Wahrscheinlichkeit und Frequenz mathematisch rigoros.

Unser Ziel ist es den folgenden Satz zu beweisen.

Satz 5.8 (Starkes Gesetz der großen Zahlen). *Seien X_i, $i \in \mathbb{N}$, unabhängige, identisch verteilte, integrierbare Zufallsvariablen mit Mittelwert $\mu = \mathbb{E}X_i$. Sei $S_n \equiv \sum_{i=1}^n X_i$. Dann ist*

$$\lim_{n\to\infty} n^{-1} S_n = \mu, \quad \text{f.s.} \tag{5.3.1}$$

Diese Formulierung ist sehr befriedigend, da sie an die Zufallsvariablem ausser der Abhängigkeit nur die Integrierbarkeit verlangt, was ja eine Mindestanforderung is damit überhaupt die rechte Seite exitiert. Der Beweis dieses Satzes ist nicht so einfach, was genau daran liegt, dass wir nur diese minimale Forderung stellen. Wir werden daher zunächst zwei einfachere Fälle betrachten.

5.3.1 Das schwache Gesetz unter Momentenannahmen.

Die erste Naheliegende Idee um ein Gesetz der großen Zahlen zu erhalten ist die Verwendung der Chebeychev Ungleichung. Wir können zunächst ohne Beschränkung der Allgemeinheit $\mu = 0$ annehmen. Nun sieht man schnell, dass man mit einer Abschätzung

$$\mathbb{P}\left(\left|n^{-1}\sum_{i=1}^n X_i\right| > x\right) \leq \frac{\mathbb{E}\left|\sum_{i=1}^n X_i\right|}{nx} \leq \frac{\mathbb{E}|X_1|}{x}$$

nicht weiterkommt, da diese die Tatsache, dass $\mathbb{E}X_i = 0$ ist nicht auszunutzen vermage. Die nächste Idee wäre es mit der Chebeychev Ungleichung der Ordnung zwei zu versuchen, nämlich

$$\mathbb{P}\left(\left|n^{-1}\sum_{i=1}^n X_i\right| > x\right) \leq \frac{\mathbb{E}\left(\sum_{i=1}^n X_i\right)^2}{n^2x^2}.$$

Wenn wir hier das Quadrat entwickeln, so sehen wir, das alle gemischten Terme $\mathbb{E}X_iX_j$, $i \neq j$ verschwinden, so dass wir die rechte Seite durch

$$\frac{\mathbb{E}X_1^2}{nx^2}$$

abschätzen können. Dies geht zumindest gegen Null, wenn $n \uparrow \infty$, falls denn $\mathbb{E}X_1^2 < \infty$. Wir brauchen also zwei Momente.

Diese Idee liefert schon ein Ergebnis, wenn auch nicht ganz das, was wir wollen.

Satz 5.9. *Seien X_i, $i \in \mathbb{N}$, identische verteilte und paarweise unkorrelierte Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathfrak{F}, \mathbb{P})$ mit endlicher Varianz σ^2. Sei $S_n \equiv \sum_{i=1}^n X_i$. Dann gilt*

$$\frac{1}{n}\sum_{i=1}^n X_i \to \mathbb{E}X_1 \quad \textit{in Wahrscheinlichkeit.} \tag{5.3.2}$$

Beweis. Der Beweis ist denkbar einfach. Wir haben wegen der Chebychev Ungleichung (5.2.1), dass

$$\begin{aligned}\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}X_1) > \epsilon\right) &\leq \frac{\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}X_1)\right)^2}{\epsilon^2} \\ &= \frac{n^{-2}\sum_{i=1}^n \mathbb{E}(X_i - \mathbb{E}X_1)^2}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}.\end{aligned} \tag{5.3.3}$$

Genauso gilt

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}X_1) < -\epsilon\right) \leq \frac{\sum_{i=1}^n \mathbb{E}(X_i - \mathbb{E}X_1)^2}{n^2\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}. \tag{5.3.4}$$

Da die rechten Seiten für jedes $\epsilon > 0$ nach Null konvergieren, folgt die Konvergenz wie behauptet sofort. □

Anmerkung. Beachte, dass wir hier keine Unabhängigkeit, sondern nur die schwächere Annahme der Unkorreliertheit gefordert haben!

5.3.2 Das starke Gesetz unter Momentenbedingungen

Die Schranke in (5.3.4) ist nicht über n summierbar, daher lässt sich hieraus nicht die fast sichere Konvergenz via Borel-Cantelli Lemma ableiten. Die naheliegende Idee ist nun diese Abschätzung zu verbessern, indem wir eine Chebychev-Ungleichung höherer Ordnung verwenden. Dies liefert z.B. folgende Aussage:

Proposition 5.10. *Seien X_i unabhängige, identisch verteilte Zufallsvariablen, und sei $\mathbb{E}X_i^4 < \infty$. Dann gilt dass*

$$\frac{S_n}{n} \equiv \frac{1}{n}\sum_{i=1}^{n} X_i \to \mathbb{E}X_1 \quad \text{f.s.} \tag{5.3.5}$$

Beweis. Wir können ohne Schaden annehmen, dass $\mathbb{E}X_1 = 0$. Unter Verwendung unseres Kriteriums aus Korollar 4.12 müssen wir nur zeigen, dass

$$\sum_{n=1}^{\infty} \mathbb{P}(|S_n/n| > \epsilon) < \infty. \tag{5.3.6}$$

Dies folgt aus der Chebychev-Ungleichung wenn wir zeigen, dass

$$\mathbb{E}(S_n/n)^4 \leq Cn^2,$$

für $C < \infty$. Nun ist aber

$$\mathbb{E}S_n^4 = \sum_{i_1,i_2,i_3,i_4=1}^{n} \mathbb{E}X_{i_1}X_{i_2}X_{i_3}X_{i_4}.$$

Wegen $\mathbb{E}X_i = 0$ tragen in dieser Summe nur Terme bei, in denen je zwei der Indizes gleich sind. Daher ist

$$\sum_{i_1,i_2,i_3,i_4=1}^{n} \mathbb{E}X_{i_1}X_{i_2}X_{i_3}X_{i_4} = (3n^2 - n)\mathbb{E}X_1^2 + n\mathbb{E}X_1^4.$$

Hieraus folgt aber das gewünschte Ergebnis sofort. □

Wir haben also ein starkes Gesetz, aber nur unter recht starken Momentenannahmen. Damit sind wir noch nicht zufrieden. Was wir aus dem Beweisen aber sehen, ist, dass wir mit der Chebeychev Ungleichung nicht weiterkommen. Wir brauchen eine bessere Ungleichung.

5.3.3 Kolmogorov's Ungleichung

Die gesuchte Verbesserung ist die folgende sog. *Kolmogorov Ungleichung*. Sie sagt etwas über das Maximum einer ganzen Familie $S_k, k \leq n$ aus.

Lemma 5.11. *Seien X_i, $i \in \mathbb{N}$, unabhängige Zufallsvariablen mit Mittelwerten $\mathbb{E}X_k = \mu_k$ und Varianzen σ_k^2. Sei $S_n = \sum_{k=1}^n X_k$, $m_n = \sum_{k=1}^n \mu_k$ und $s_n^2 \equiv \sum_{k=1}^n \sigma_k^2$. Dann ist für alle $t > 0$,*

$$\mathbb{P}\left(\exists_{k\leq n} \,:\, |S_k - m_k| \geq ts_n\right) \leq t^{-2}. \tag{5.3.7}$$

Beweis. O.b.d.A. nehmen wir an, dass $\mu_k = 0$, $k \geq 1$, so dass auch $m_n = 0$, $n \geq 1$.

Wir definieren die Zufallsvariablen

$$Y_k = \mathbb{1}_{|S_k| \geq t s_n} \prod_{\ell<k} \mathbb{1}_{|S_\ell| < t s_n} = \mathbb{1}_{\{k = \min\{\ell : S_\ell \geq t s_n\}}. \tag{5.3.8}$$

Offenbar kann nur höchstens eine der Variablen Y_k den Wert eins annehmen, so dass $Z_n \equiv \sum_{k=1}^n Y_k$ nur die Werte null und eins annimmt. Offenbar ist Z_n genau dann eins wenn das Ereignis in (5.3.7) eintritt. Daher ist auch $\mathbb{P}(Z_n = 1) = \mathbb{E}Z_n$. Ferner ist

$$Z_n S_n^2 \leq S_n^2,$$

und somit

$$\mathbb{E}Z_n S_n^2 = \sum_{k=1}^n \mathbb{E}Y_k S_n^2 \leq s_n^2. \tag{5.3.9}$$

Nun setzen wir

$$U_k \equiv S_n - S_k = \sum_{\ell=k+1}^n X_\ell.$$

Die letzte Gleichung macht deutlich, dass U_k nur von den Variablen X_ℓ mit $\ell > k$ abhängt, weswegen U_k von S_k und von Y_k unabhängig sind. Nun schreiben wir

$$S_n^2 = (U_k + S_k)^2,$$

und erhalten so

$$\begin{aligned} \mathbb{E}Y_k S_n^2 &= \mathbb{E}Y_k (U_k + S_k)^2 \\ &= \mathbb{E}Y_k S_k^2 + 2\mathbb{E}U_k Y_k S_k + \mathbb{E}U_k^2 Y_k. \end{aligned}$$

Wegen der angesprochenen Unabhängigkeit ist der zweite Term im letzten Ausdruck gleich $2\,\mathbb{E}U_k\,\mathbb{E}Y_k S_k = 0$, da die Erwartung von U_k verschwindet. Da zudem der letzte Term nicht negativ ist, erhalten wir

$$\mathbb{E}Y_k S_n^2 \geq \mathbb{E}Y_k S_k^2.$$

Da, wenn $Y_k \neq 0$ ist, $|S_k| \geq t s_n$, folgt weiter

$$\mathbb{E}Y_k S_n^2 \geq \mathbb{E}Y_k t^2 s_n^2.$$

Setzen wir diese Ungleichung in (5.3.9) ein folgt

$$\mathbb{E}Z_n t^2 s_n^2 \leq s_n^2,$$

was unmittelbar die Behauptung ergibt. □

Anmerkung. Wir sehen, dass die Aussage des Satzes die Chebychev-Ungleichung der Ordnung zwei für den Endpunkt S_n impliziert. Die Kolmogorov Unglei-

chung ist aber strikt schärfer, da sie ja das Maximum der S_k mit $k \leq n$ kontrolliert. In der Tat ist die erzielte Verbesserung signifikant.

5.3.4 Beweis des starken Gesetzes der großen Zahlen

Die Stärke der Kolmogorov'schen Ungleichung zeigt sich im folgenden Kriterium für das starke Gesetz für unabhängige, aber nicht identisch verteilte Zufallsvariablen.

Lemma 5.12. *Seien X_k, $k \in \mathbb{N}$ unabhängige Zufallsvariablen mit Varianzen σ_k^2 und Mittelwerten μ_k. Wenn*

$$\sum_{k=1}^{\infty} \frac{\sigma_k^2}{k^2} < \infty, \tag{5.3.10}$$

dann gilt

$$\frac{1}{n}\sum_{k=1}^{n}(X_k - \mu_k) \to 0, \quad \text{f.s.} \tag{5.3.11}$$

Beweis. Wir definieren die Ereignisse A_p durch

$$A_p = \bigcup_{2^{p-1}<n\leq 2^p} \{|S_n| \geq \epsilon\, n\}\,.$$

Wenn die Summe der Wahrscheinlichkeiten der A_p konvergiert, so folgt die fast sicher Konvergenz aus dem ersten Borel-Cantelli Lemma. Wir müssen also die Wahrscheinlichkeiten der A_p abschätzen. Nun impliziert das Ereignis A_p, dass für ein n zwischen $2^{p-1}+1$ und 2^p, $|S_n| \geq \epsilon 2^{p-1}$. Dies ist aber ein Ereignis, dessen Wahrscheinlichkeit durch die Kolmogorov'sche Ungleichung abgeschätzt werden kann. Nämlich

$$\begin{aligned}
\mathbb{P}(A_p) &\leq \mathbb{P}\left(\exists_{2^{p-1}<k\leq 2^p}\{|S_k| \geq \epsilon 2^{p-1}\}\right)\\
&\leq \mathbb{P}\left(\exists_{1\leq k\leq 2^p}\{|S_k| \geq \epsilon 2^{p-1} s_{2^p}^{-1} s_{2^p}\}\right)\\
&\leq 4\epsilon^{-2}2^{-2p}s_{2^p}^2.
\end{aligned}$$

Nun müssen wir nur noch summieren:

$$\begin{aligned}
\sum_{p=1}^{\infty}\mathbb{P}(A_p) &\leq \sum_{p=1}^{\infty} 4\epsilon^{-2}2^{-2p}s_{2^p}^2 = 4\epsilon^{-2}\sum_{p=1}^{\infty}2^{-2p}\sum_{k=1}^{2^p}\sigma_k^2 \\
&= 4\epsilon^{-2}\sum_{k=1}^{\infty}\sigma_k^2 \sum_{p:2^p\geq k} 2^{-2p} \leq 8\epsilon^{-2}\sum_{k=1}^{\infty}\sigma_k^2 k^{-2}
\end{aligned} \tag{5.3.12}$$

was nach Annahme endlich ist. Somit ist das Lemma bewiesen. □

Mit diesem Kriterium können wir nun den Beweis von Satz 5.8 führen.

Beweis. von Satz 5.8. Hier lernen wir noch eine wichtige Technik kennen, die der *Trunkation*. Im wesentlichen wollen wir unsere Variablen so aufspalten, dass wir einen Term erhalten, auf den wir das Lemma von oben anwenden können, während der Rest nach null konvergiert. Dazu setzen wir

$$U_k = X_k \mathbb{1}_{|X_k|<k}, \quad V_k = X_k \mathbb{1}_{|X_k|\geq k}.$$

Offenbar ist $X_k = U_k + V_k$. Nun erfüllen die U_k Kolmogorov's Kriterium. Dazu berechnen wir

$$\sigma_k^2 \equiv \mathrm{var}(U_k) \leq \mathbb{E}(U_k^2) \leq \sum_{\ell=1}^{k} \ell\, \mathbb{E}\left(|X_k|\{\mathbb{1}_{\ell-1\leq|X_k|<\ell\}}\right) \equiv \sum_{\ell=1}^{k} \ell a_\ell.$$

beachte, dass a_ℓ nicht von k abhängt, da die X_k gleichverteilt sind.Daher gilt

$$\sum_{k=1}^{\infty} \frac{\sigma_k^2}{k^2} \leq \sum_{k=1}^{\infty} \frac{1}{k^2} \sum_{\ell=1}^{k} \ell a_\ell = \sum_{\ell=1}^{\infty} \ell a_\ell \sum_{k=\ell}^{\infty} \frac{1}{k^2} \leq \sum_{\ell=1}^{\infty} a_\ell,$$

wobei wir benutzt haben, dass $\sum_{k=\ell}^{\infty} \frac{1}{k^2} \leq 2/\ell$ (für $\ell > 4$) ist; nun ist aber

$$\sum_{\ell=1}^{\infty} a_\ell = \mathbb{E}|X_k| < \infty,$$

nach Vorraussetzung. Somit ist in der Tat das Kolmogorov Kriterium erfüllt. Weiter ist

$$\mathbb{E}U_k = \mu - \mathbb{E}\left(X_k \mathbb{1}_{|X_k|\geq k}\right). \tag{5.3.13}$$

Aber

$$\left|\mathbb{E}\left(X_k \mathbb{1}_{|X_k|\geq k}\right)\right| \leq \mathbb{E}\left(|X_k| \mathbb{1}_{|X_k|\geq k}\right) = \sum_{\ell=k}^{\infty} a_\ell. \tag{5.3.14}$$

Nun wissen wir schon, dass die Reihe $\sum_{\ell=1}^{\infty} a_\ell$ konvergiert, also konvergiert die Folge $\sum_{\ell=k}^{\infty} a_\ell$ nach Null, wenn $\ell \uparrow \infty$.

Da wir leicht sehen, dass $\mathbb{E}U_k \to \mu$, liefert das vorhergehende Lemma, dass $n^{-1}\sum_{k=1}^{n}(U_k - \mathbb{E}U_k) \to 0$, fast sicher, und $n^{-1}\sum_{k=1}^{n} \mathbb{E}U_k \to \mu$, so dass $n^{-1}\sum_{k=1}^{n} U_k \to \mu$, fast sicher. Dmit konvergiert $\mathbb{E}U_k$ gegen μ. Daraus folgt aber auch, dass $n^{-1}\sum_{k=1}^{n} \mathbb{E}U_k \to \mu$, wenn $n \uparrow \infty$.

Wir müssen nur noch zeigen, dass V_n unwichtig ist. Die Gefahr an V_n ist ja, dass es sehr groß sein kann: dafür ist es aber auch meistens gleich Null. In der Tat wollen wir zeigen, dass es nur endlich oft von Null verschieden ist. Dazu schreiben wir

$$\mathbb{P}(V_n \neq 0) = \mathbb{E}\mathbb{1}_{|X_n|\geq n} \leq \sum_{\ell=n}^{\infty} \frac{a_{\ell+1}}{\ell}.$$

Dann ist

$$\sum_{n=1}^{\infty} \mathbb{P}(V_n \neq 0) \leq \sum_{n=1}^{\infty} \sum_{\ell=n}^{\infty} \frac{a_{\ell+1}}{\ell} = \sum_{\ell=1}^{\infty} \frac{a_{\ell+1}}{\ell} \sum_{n=1}^{\ell} 1 = \sum_{\ell=1}^{\infty} a_{\ell+1} < \infty \quad (5.3.15)$$

und das Ergebnis folgt aus dem ersten Borel-Cantelli Lemma. □

Kapitel 6
Der zentrale Grenzwertsatz

On peut facilement, au moyen de ces formules, déterminer les bénéfices des loteries[a].
Pierre Simon de Laplace, Théorie Analytique des Probabilités

[a] Man kann mittels dieser Formeln leicht den Gewinn von Lotterien berechnen.

Wir kommen nun zu dem zweiten wichtigen Satz der Wahrscheinlichkeitstheorie, dem nicht ohne Grund so genannten *zentralen Grenzwertsatz*. Seine Bedeutung liegt zum einen wieder in den Implikationen für die Statistik, denn er rechtfertigt in vielen Fällen die Annahme einer Gauß'schen Verteilung (bzw. derer Derivate) für Zufallsgrössen die auf komplizierte Art und Weise zustande kommen. Zum anderen ist er ein weiteres Beispiel dafür, wie spezifische Gesetzmässigkeiten aus zufälligem Geschehen folgen. Einen speziellen Fall des zentralen Grenzwertsatzes haben wir schon mit dem Satz von de Moivre-Laplace kennengelernt.

6.1 Grenzwertsätze

Der zentrale Grenzwertsatz kann als Verfeinerung des Gesetzes der großen Zahlen aufgefasst werden. Wir wissen, das für Summen, $S_n \equiv \sum_{i=1}^{n} X_i$, unabhängiger, identisch verteilter Zufallsvariablen, X_i, $n^{-1}S_n$ fast sicher gegen den Erwartungswert, $\mathbb{E}X_1$ konvergiert. Es liegt nun nahe, die Frage nach der Konvergenzgeschwindigkeit zu stellen. Dazu nehmen wir $n^{-1}S_n - \mathbb{E}X_1$ und blasen es mit einem n-abhängigen Faktor auf, der so gewählt ist, dass im Grenzwert etwa endliches übrig bleibt. Es liegt nahe, eine Potenz von n zu versuchen. Die Frage ist also: gibt es $\gamma > 0$, so dass

$$n^{\gamma}(n^{-1}S_n - \mathbb{E}X_1) \tag{6.1.1}$$

einen nicht-trivialen Limes hat. Dieser wird i.A. eine Zufallsvariable sein. Schon numerischen Simulationen zeigen dabei, dass die Konvergenz dabei bestenfalls in Verteilung zu erwarten ist. Unser Problem ist also die Berech-

nung der Verteilung des Limes von Summen unnabhängiger Zufallsvariablen nach geeigneter Reskalierung. Unsere Erfahrung mit dem speziellen Fall der Bernoulliverteilung legt dabei nahe, dass wohl $\gamma = 1/2$ gewählt werden sollte, und das der Grenzwert gerade die Gaußverteilung sein sollte; jedoch ist von vorneherein nicht auszuschliessen, dass all dies von der speziellen Wahl der Verteilungen abhängen kann.

Allgemein gesprochen, stellt sich die Aufgabe also wie folgt:

- Unter welchen Annahmen an die Zufallsvariablen X_i gibt es ein γ, so dass der Ausdruck in (6.1.1) in Verteilung gegen eine Zufallsvariable konvergiert?
- Was sind die möglichen Verteilungen der Grenzwerte?
- Welche Bedingungen an die Verteilungen der X_i charakterisieren die Verteilung des Grenzwertes?

Wir werden uns im folgenden auf den Fall beschränken, dass die Zufallsvariablen X_i endliche Varianz haben. Dann können wir sofort schliessen, dass $\gamma = 1/2$ sein muss, denn es ist dann

$$\mathbb{E}\left(n^{\gamma}\left(n^{-1}S_n - \mathbb{E}X_1\right)\right)^2 = n^{2\gamma-1}\mathrm{var}(X_1), \tag{6.1.2}$$

was nur für $\gamma = 1/2$ gegen einen von Null verschiedenen Grenzwert konvergieren kann. Es bleibt zu zeigen, dass für diese Wahl dann auch tatsächlich Konvergenz in Verteilung folgt.

6.2 Charakteristische Funktionen

Wir hatten gesehen, dass die Verteilungen als n-fache Faltungen der Verteilungen von X_i ausgedrückt werden können. Die entsprechenden Ausdrücke wirken allerdings im Allgemeinen unhandlich. Eine gute Methode, mit solchen Faltungen umzugehen ist die sogenannte Fouriertransformation.

Definition 6.1. Sei X eine reelle Zufallsvariable auf einem Wahrscheinlichkeitsraum $(\Omega, \mathfrak{F}, \mathbb{P})$, dann heisst

$$\phi(t) \equiv \phi_X(t) = \mathbb{E}e^{\mathbf{i}tX} \equiv \mathbb{E}\cos(Xt) + \mathbf{i}\mathbb{E}\sin(tX), \tag{6.2.1}$$

wo $t \in \mathbb{R}$ und $\mathbf{i} = \sqrt{-1}$ ist, die *charakteristische Funktion* von X bzw. die charakteristische Funktion der Verteilung, $\mathbb{P}_X \equiv \mathbb{P} \circ X^{-1}$, von X.

Anmerkung. Natürlich ist, wenn P_X die Verteilung von X ist,

$$\phi_X(t) = \int_{\mathbb{R}} e^{\mathbf{i}tx}\,\mathrm{d}P_X(x)$$

gerade die *Fouriertransformierte* des Masses P_X. Für ein Mass, μ, auf $\mathbb{R}$ schreiben wir auch ϕ_μ für $\int e^{\mathbf{i}tx}\,\mathrm{d}\mu(x)$ und nennen ϕ_μ die charakteristische

Funktion des Masses μ. In der Literatur wird häufig auch die Bezeichnung $\widehat{\mu} \equiv \phi_\mu$ benutzt.

Wir beobachten zuächst, dass $\phi_X(t)$ für alle $t \in \mathbb{R}$ existiert, da sowohl $\sin(xt)$ als auch $\cos(xt)$ beschränkt und messbar, also insbesondere integrierbar gegen jedes W-Maß sind. Weiterhin kann man zeigen, dass jede charakteristische Funktion stetig ist.

Lemma 6.2. *Jede charakteristische Funktion, ϕ, eines Wahrscheinlichkeitsmasses ist gleichmässig stetig auf $\mathbb{R}$.*

Beweis. Eine elementare Rechnung zeigt, dass

$$|\phi(t) - \phi(s)|^2 \leq 2\left(1 - \Re\left(\phi(t-s)\right)\right).$$

Es ist nämlich

$$\begin{aligned}
|\phi(t) - \phi(s)| &= \left|\mathbb{E}\left[e^{\mathbf{i}tX}\left(1 - e^{\mathbf{i}(s-t)X}\right)\right]\right| \\
&\leq \mathbb{E}\left[\left|1 - e^{\mathbf{i}(s-t)X}\right|\right] = \mathbb{E}\left[\sqrt{(1-\cos((s-t)X))^2 + \sin^2((s-t)X)}\right] \\
&= \mathbb{E}\left[\sqrt{2 - 2\cos((s-t)X)}\right] \leq \sqrt{2 - 2\mathbb{E}[\cos((s-t)X)]},
\end{aligned}$$

wo die letzte Ungleichung die Cauchy-Schwartz Ungleichung benutzt. Weiter gilt, für jedes $N < \infty$,

$$\begin{aligned}
1 - \Re\phi(u) &\leq \int_{\mathbb{R}} |1 - e^{\mathbf{i}ux}| \, \mathrm{d}\mathbb{P}(x) \\
&\leq \int_{|x|\leq N} |1 - e^{\mathbf{i}ux}| \, \mathrm{d}\mathbb{P}(x) + \int_{|x|>N} |1 - e^{\mathbf{i}ux}| \, \mathrm{d}\mathbb{P}(x) \\
&\leq \sup_{|x|\leq N} |1 - e^{\mathbf{i}ux}| + 2\mathbb{P}\left([-N, N]^c\right). \qquad (6.2.2)
\end{aligned}$$

Nun können wir für jedes $\epsilon > 0$ Zahlen $N \in \mathbb{N}$ und $u_0 > 0$ so finden, dass für alle $|u| \leq u_0$, sowohl der erste als auch der zweite Ausdruck kleiner als $\epsilon^2/2$ sind. Damit folgt aber die Stetigkeit, und sogar die gleichmässige Stetigkeit von ϕ. □

Wie schon die erzeugenden Funktionen sind die charakteristischen Funktionen mit den Momenten verknüpft.

Lemma 6.3. *Seien ϕ die charakteristische Funktion einer Zufallvariablen X und sei ferner $\mathbb{E}|X|^n < \infty$. Dann ist $\phi(t)$ n-mal differenzierbar und es gelten*

$$\phi(0) = 1, \qquad (6.2.3)$$

$$\phi^{(n)}(0) \equiv \frac{d^n}{\mathrm{d}t^n}\phi(t=0) = \mathbf{i}^n \mathbb{E}X^n, \qquad (6.2.4)$$

Beweis. Zunächst ist $\phi(0) = \mathbb{E}1 = 1$. Wir setzen $e(t;x) = e^{\mathbf{i}xt}$ und $e^{(n)}(t;x) \equiv \frac{\partial^n}{\partial t^n} e(t;x)$. Dann ist

Benutzen wir, dass

$$e(t;X) = e(0;X) + \int_0^t e^{(1)}(t_1;X)\,\mathrm{d}t_1,$$

und also

$$\phi(t) = \phi(0) + \mathbb{E}\left(\int_0^t e^{(1)}(t_1;X)\,\mathrm{d}t_1\right).$$

Nun ist $|e^{(1)}(t_1;X)| \leq |X|$ und daher unter der Annahme, dass $\mathbb{E}|X| < \infty$, nach dem Satz von Fubini-Lebesgue,

$$\mathbb{E}\left(\int_0^t e^{(1)}(t_1;X)\,\mathrm{d}t_1\right) = \int_0^t \mathbb{E}\left(e^{(1)}(t_1;X)\right)\,\mathrm{d}t_1.$$

Die rechte Seite ist nun explizit differenzierbar bezüglich t und daher

$$\phi'(t) = \mathbb{E}\left(e^{(1)}(t_1;X)\right) = i\mathbb{E}Xe^{\mathbf{i}tX},$$

und somit $\phi'(0) = \mathbf{i}\mathbb{E}X$.

Die Verallgemeinerung auf den Fall der n-ten Ableitung geht genauso, indem wir benutzen, dass

$$e^{\mathbf{i}tX} - \sum_{j=0}^{n-1} \frac{(\mathbf{i}tX)^j}{j!} = \mathbf{i}^n X^n \int_0^t \int_0^{t_n} \dots \int_0^{t_2} e^{\mathbf{i}t_1 X}\,\mathrm{d}t_1 \dots \mathrm{d}t_n.$$

Daher ist

$$\begin{aligned}\phi(t) - \sum_{j=0}^{n-1} \frac{(\mathbf{i}t)^j \mathbb{E}X^j}{j!} &= \mathbf{i}^n \mathbb{E}\left(X^n \int_0^t \int_0^{t_n} \dots \int_0^{t_2} e^{\mathbf{i}t_1 X}\,\mathrm{d}t_1 \dots \mathrm{d}t_n\right) \\ &= \mathbf{i}^n \int_0^t \int_0^{t_n} \dots \int_0^{t_2} \mathbb{E}\left(X^n e^{\mathbf{i}t_1 X}\right)\,\mathrm{d}t_1 \dots \mathrm{d}t_n.\end{aligned}$$

Hier haben wir wieder den Satz von Fubini-Lebesgue unter den Annahme dass $\mathbb{E}|X|^n < \infty$ ist benutzt um die Erwartung bez. X und die t-Integrale zu vertauschen. Jetzt können wir beide Seiten n-mal ableiten und t Null setzen um (6.2.3) zu erhalten. □

Die Nützlichkeit der charakteristischen Funktionen rührt unter anderem daher, dass sie eine sehr schöne Eigenschaft bezüglich der Faltung hat. Wir werden im folgenden stets Zufallsvariablen mit Mittelwert Null betracten, da wir uns durch Subtraktion des Mittelwertes immer auf triviale Weise auf diesen Fall zurückziehen können.

Lemma 6.4. *Seien X_ℓ, $\ell \in \mathbb{N}$ unabhängige Zufallsvariablen mit Erwartungswert $\mathbb{E}X_\ell$ und mit charakteristischen Funktionen $\phi_\ell(t) \equiv \phi_{X_\ell}(t)$. Sei $S_n = \sum_{\ell=1}^n X_\ell$. Dann ist*

$$\phi_{S_n}(t) = \prod_{\ell=1}^n \phi_\ell(t). \tag{6.2.5}$$

Weiter gilt, wenn

$$Z_n \equiv n^{-1/2} S_n, \tag{6.2.6}$$

$$\phi_{Z_n}(t) = \prod_{\ell=1}^n \phi_\ell(t/\sqrt{n}). \tag{6.2.7}$$

Beweis. Die Aussagen folge sofort aus Lemma 3.7 und (6.2.7). □

Beispiel 6.5. Vergleiche mit Bemerkung 3.6.5 und Beispiel 5.3.

- Für $X \sim \mathcal{N}(\mu, \sigma^2)$, gilt $\phi(t) = \exp(-\sigma^2 t^2/2 + it\mu)$.
- Für $X \sim \mathrm{Bin}(n, p)$, gilt $\phi(t) = (1 - p + p\, e^{it})^n$.
- Für $X \sim \mathrm{Poi}(\lambda)$, gilt $\phi(t) = \exp(-\lambda(e^{it} - 1))$.
- Für $X \sim \mathrm{Exp}(a)$, gilt $\phi(t) = 1/(1 - it/a)$.
- Für $X \sim \mathrm{Geo}(q)$, gilt $\phi(t) = (1 - q)/(1 - qe^{it})$.
- Für $X \sim \mathrm{Cauchy}(a)$, gilt $\phi(t) = e^{-i|t|a}$.

In der Welt der charakteristischen Funktionen sind also die Summen unnabhängier Zufallsvariablen einfach mit den Produkten verknüpft, was viel leichter zu handhaben ist als die Faltung. Was man also nur noch braucht, damit dies nützlich ist, ist ein Weg zurück aus der Welt der charakteristischen Funktionen in die der Verteilungen. Diesen liefert uns der folgende Satz von Lévy.

Satz 6.6. *Die charakteristische Funktion einer Zufallsvariablen legt deren Verteilung eindeutig fest.*

Beweis. Der Beweis benutzt den Gauss'schen Fall als Startpunkt. Wir beginnen daher mit folgendem Lemma.

Lemma 6.7. *Sei X eine Gauss'sche Zufallsvariable mit Mittelwert Null und Varianz σ^2. Dann ist*

$$\phi_X(t) = \exp\left(-\frac{\sigma^2 t^2}{2}\right). \tag{6.2.8}$$

Beweis. Man kann dieses Resultat auf verschiedene Arten zeigen. Wir gehen wie folgt vor. Aus dem Beweis von Satz 6.3 wissen wir schon, dass

$$\phi'_X(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} e^{-\frac{x^2}{2\sigma^2}} \mathbf{i} x e^{\mathbf{i}tx}\, \mathrm{d}x. \tag{6.2.9}$$

Nun ist

$$e^{-\frac{x^2}{2\sigma^2}}\mathbf{i}xe^{itx} = -\mathbf{i}\sigma^2\left(\frac{\mathrm{d}}{\mathrm{d}x}e^{-\frac{x^2}{2\sigma^2}}\right)e^{\mathbf{i}tx},$$

und daher erhalten wir durch partielle integration in (6.2.9),

$$\phi'_X(t) = -t\sigma^2\phi_X(t). \tag{6.2.10}$$

Da $\phi_X(0) = 1$ gelten muss, ist (6.2.8) die einzige Lösung dieser Differentialgleichung. □

Wir kommen nun zum Beweis des eigentlichen Satzes. Wir setzen

$$p_\sigma(x) \equiv \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{x^2}{2\sigma^2}}. \tag{6.2.11}$$

Sei μ ein Wahrscheinlichkeitsmaß auf $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$. Wir definieren

$$f_\sigma(x) \equiv \int_{\mathbb{R}} p_\sigma(x-y)\,\mathrm{d}\mu(y) \equiv p_\sigma \star \mu(x), \tag{6.2.12}$$

und $d\mu_\sigma(x) \equiv f_\sigma(x)\,\mathrm{d}x$.

Wir zeigen zunächst, dass μ_σ eindeutig durch ϕ_μ bestimmt ist. Dazu beobachten wir, dass

$$\sqrt{2\pi\sigma^2}p_\sigma(x) = e^{-\frac{x^2}{2\sigma^2}} = \int_{\mathbb{R}} e^{-\mathbf{i}tx}p_{1/\sigma}(t)\,\mathrm{d}t.$$

Darum haben wir

$$\begin{aligned} f_\sigma(x) &= \int_{\mathbb{R}} p_\sigma(x-y)\,\mathrm{d}\mu(y) = \frac{1}{\sqrt{2\pi\sigma^2}}\int_{\mathbb{R}}\int_{\mathbb{R}} e^{-\mathbf{i}t(x-y)}p_{1/\sigma}(t)\,\mathrm{d}t\,\mathrm{d}\mu(y) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\int_{\mathbb{R}} e^{-\mathbf{i}tx}p_{1/\sigma}(t)\left(\int_{\mathbb{R}} e^{\mathbf{i}ty}\,\mathrm{d}\mu(y)\right)\mathrm{d}t \\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\int_{\mathbb{R}} e^{-\mathbf{i}tx}p_{1/\sigma}(t)\phi_\mu(t)\,\mathrm{d}t. \end{aligned} \tag{6.2.13}$$

Hier haben wir den Satz von Fubini-Lebesgue in der ersten Gleichung verwendet und die Definition der charakteristichen Funktion in der zweiten. Im Ergebnis haben wir nun eine Formel für die Dichte des Maßes μ_σ in die nur die charakteristische Funktion von μ eingeht.

Schliesslich zeigen wir noch, dass für jede stetige und beschränkte Funktion, h,

$$\lim_{\sigma\downarrow 0}\int h(x)\,\mathrm{d}\mu_\sigma(x) = \int_{\mathbb{R}} h(x)\,\mathrm{d}\mu(x) \tag{6.2.14}$$

gilt. Zunächst sehen wir, dass, wieder unter Verwendung des Satzes von Fubini,

$$\int_{\mathbb{R}} h(x)\,\mathrm{d}\mu_\sigma(x) = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} h(x)\rho_\sigma(x-y)\mathrm{d}\mu(y) \right) \mathrm{d}x \tag{6.2.15}$$
$$= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \rho_\sigma(x-y)h(x)\mathrm{d}x \right) \mathrm{d}\mu(y) = \int_{\mathbb{R}} p_\sigma \star h(x)\,\mathrm{d}\mu(y).$$

Dabei haben wir die Faltung zweier Funktionen definiert als $h \star f(y) = \int_{\mathbb{R}} h(x-y)f(x)\mathrm{d}x$ und benutzt dass $\rho_\sigma(x) = \rho_\sigma(-x)$.

Dann benutzen wir die elementaren Eigenschaften der Gauss'schen Dichte,

$$\int_{\mathbb{R}} p_\sigma(x)\,\mathrm{d}x = 1 \tag{6.2.16}$$

$$\lim_{\sigma \downarrow 0} \int_{|x|>\epsilon} p_\sigma(x)\,\mathrm{d}x = 0, \quad \forall \epsilon > 0. \tag{6.2.17}$$

Dies impliziert für stetige und beschränkte Funktionen h, dass

$$\lim_{\sigma\downarrow 0} p_\sigma \star h(x) \equiv \lim_{\sigma\downarrow 0} \int_{\mathbb{R}} p_\sigma(x-y)h(y)\mathrm{d}y = h(x).$$

Da weiter $p_\sigma \star h(x) \leq \sup h(x) < \infty$, können wir den Satz von Lebesgue benutzen um zu zeigen, dass (6.2.14) gilt. Damit ist aber dass Maß μ eindeutig durch ϕ_μ festgelegt. □

Es ist also nicht verwunderlich, dass Konvergenz der charakteristischen Funktionen einer Folge von Zufallsvariablen deren Verteilung in Konvergenz impliziert. Auch dieser Satz geht auf Lévy zurück.

Satz 6.8. *Sei X_n, $n \in \mathbb{N}$, eine Folge von Zufallsvariablen und seien ϕ_n deren charakteristiche Funktionen. Wenn die charakteristischen Funktionen $\phi_n(t)$ gegen einen Grenzwert $\phi(t)$ auf $\mathbb{R}$ konvergieren, der die charakteristische Funktion einer Zufallsvariablem X ist, dann konvergieren die Zufallsvariablen X_n in Verteilung gegen X.*

Beweis. Es sei $\phi_n(t)$ eine Folge von charakteristischen Funktionen, die gegen eine charakteristishe Funktion ϕ konvergiert. Es seien μ_n, μ, die zugehörigen Wahrscheinlichkeitsmaße. Wir wollen zeigen, daß μ_n schwach gegen μ konvergiert. Sei dazu f eine stetige Funktion mit kompaktem Träger. Wir zeigen zunächst, dass $\int f\,\mathrm{d}\mu_n \to \int f\,\mathrm{d}\mu$. Wir zeigen dazu, dass für alle $\sigma > 0$,

$$\int_{\mathbb{R}} p_\sigma \star f\,\mathrm{d}\mu_n \to \int_{\mathbb{R}} p_\sigma \star f\,\mathrm{d}\mu. \tag{6.2.18}$$

Dazu benutzen wir, dass, wie wir schon sahen,

$$\int_{\mathbb{R}} p_\sigma \star f\,\mathrm{d}\mu_n = \int_{\mathbb{R}} f(x) \left(\frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} e^{-\mathbf{i}xt} p_{1/\sigma}(t)\phi_n(t)\,\mathrm{d}t \right) \mathrm{d}x. \tag{6.2.19}$$

Da ϕ_n punktweise konvergiert und die Integranden (bezüglich der t-Integration) $e^{-\mathrm{i}tx}p_{1/\sigma}(t)\phi_n(t)$ im Betrag durch die integrierbare Funktion $p_{1/\sigma}(t)$ beschränkt sind, folgt aus dem Satz von Lebesgue, dass die

$$\frac{1}{\sqrt{2\pi\sigma^2}}\int_{\mathbb{R}} e^{-\mathrm{i}xt}p_{1/\sigma}(t)\phi_n(t)\,\mathrm{d}t \to \frac{1}{\sqrt{2\pi\sigma^2}}\int_{\mathbb{R}} e^{-\mathrm{i}xt}p_{1/\sigma}(t)\phi(t)\,\mathrm{d}t,$$

und da diese im Betrag kleiner oder gleich 1 sind, können wir wieder den Satz von Lebesgue auf die x-Integration anwenden (da f beschränkt mit kompaktem Träger ist) und erhalten (6.2.18).

Schliesslich bemerken wir, dass, für jedes $\sigma > 0$,

$$\begin{aligned}\left|\int f\,\mathrm{d}\mu_n - \int f\,\mathrm{d}\mu\right| &\le \int |f - p_\sigma \star f|\,\mathrm{d}\mu_n \\ &\quad + \left|\int p_\sigma \star f\,\mathrm{d}\mu_n - \int p_\sigma \star f\,\mathrm{d}\mu\right| \\ &\quad + \int |p_\sigma \star f - f|\,\mathrm{d}\mu.\end{aligned} \tag{6.2.20}$$

Sei $\epsilon > 0$ beliebig; dann wählen wir σ so, dass $\sup_x |p_\sigma \star f(x) - f(x)| \le \epsilon/3$ und danach n so, dass $\left|\int p_\sigma \star f\,\mathrm{d}\mu_n - \int p_\sigma \star f\,\mathrm{d}\mu\right| \le \epsilon/3$ (das ist wegen (6.2.18) möglich). Dann folgt mit (6.2.20), dass für solche n,

$$\left|\int f\,\mathrm{d}\mu_n - \int f\,\mathrm{d}\mu\right| \le \epsilon,$$

und mithin die Konvergenz von $\int f\,\mathrm{d}\mu_n$ nach $\int f\,\mathrm{d}\mu$.

Zum Schluss müssen wir noch zeigen, dass die Konvergenz für alle stetigen Funktionen mit kompaktem Träger ausreicht, um die Konvergenz von $\int f\,\mathrm{d}\mu_n$ für alle beschränkten Funktionen zu zeigen. Sei dazu h_k eine Folge von stetigen Funktionen mit kompaktem Träger und $0 \le h_k(x) \le 1$, so dass $h_k \uparrow 1$. Dann ist fh_k ebenfalls stetig mit kompaktem Träger, und somit

$$\int h_k f\,\mathrm{d}\mu_n \to \int h_k f\,\mathrm{d}\mu.$$

Weiter ist

$$\begin{aligned}\left|\int f\,\mathrm{d}\mu_n - \int f h_k\,\mathrm{d}\mu_n\right| &\le \sup_x |f(x)|\left(1 - \int h_k\,\mathrm{d}\mu_n\right), \\ \left|\int f\,\mathrm{d}\mu - \int f h_k\,\mathrm{d}\mu\right| &\le \sup_x |f(x)|\left(1 - \int h_k\,\mathrm{d}\mu\right).\end{aligned}$$

Somit haben wir

$$\left|\int f\,\mathrm{d}\mu_n - \int f\,\mathrm{d}\mu\right| \le \left|\int f h_k\,\mathrm{d}\mu_n - \int f h_k\,\mathrm{d}\mu\right| + M\left(1 - \int h_k\,\mathrm{d}\mu_n\right) + M\left(1 - \int h_k\,\mathrm{d}\mu\right)$$

wobei $M = \sup_x |f(x)|$.

Da $\int h_k\,\mathrm{d}\mu_n$ nach $\int h_k\,\mathrm{d}\mu$ strebt, wenn $n \to \infty$, und $\int h_k\,\mathrm{d}\mu \uparrow 1$, wenn $k \to \infty$, folgt die Konvergenz von $\int f\,\mathrm{d}\mu_n$ für alle beschränkten stetigen Funktionen: Für jedes $\epsilon > 0$ wähle k, so dass $0 \le 1 - \int h_k\,\mathrm{d}\mu \le \epsilon/4M$, und dann n_0, so dass für $n \ge n_0$, $M\left|\int h_k\,\mathrm{d}\mu_n - \int h_k\,\mathrm{d}\mu\right| \le \epsilon/4$ und $\left|\int f h_k\,\mathrm{d}\mu_n - \int f h_k\,\mathrm{d}\mu\right| \le \epsilon/4$.

Dann folgt die schwache Konvergenz aus Satz 4.2. □

6.3 Der zentrale Grenzwertsatz

Der Satz 6.8 von Lévy gibt uns ein einfach zu handhabendes Kriterium an die Hand, um einen zentralen Grenzwertsatz zu beweisen. Es genügt danach offenbar, die charakteristische Funktion der Zufallsvariablen Z_n zu berechnen und deren Konvergenz nachzuweisen und den Grenzwert als charakteristische Funktion einer bekannten Zufallsvariable zu identifizieren. In Hinblick darauf, dass wir stets statt X_i die Variablen $X_i - \mathbb{E}X_i$ betrachten können, genügt es im Folgenden die Annahme $\mathbb{E}X_i = 0$ zu machen.

Aus Lemma 6.4 folgt sofort als Korollar:

Korollar 6.9. *Seien X_i unabhängige identisch verteilte Zufallsvariablen mit Erwartungswert 0 und charakteristischer Funktion ϕ, und sei Z_n wie in (6.2.6).*

$$\phi_{Z_n}(t) = \left[\phi(n^{-1/2}t)\right]^n. \tag{6.3.1}$$

Bleibt also nur zu zeigen, wann und wohin $\left[\phi(n^{-1/2}t)\right]^n$ konvergiert. Hierzu benutzen wir das folgende elementare Lemma.

Lemma 6.10. *Sei a_n eine Folge von reellen Zahlen so dass $\lim_{a_n} = a$. Dann gilt*

$$\lim_{n\uparrow\infty} (1 + a_n/n)^n = e^a. \tag{6.3.2}$$

Beweis. Offenbar ist $1 + a_n/n = \exp\left(\ln(1 + a_n/n)\right)$. Für hinreichend grosse n is dann auch $|a_n/n| \le 1/10$. Andererseits gibt es eine endliche Konstante C, so dass für alle $|x| \le 1/10$, $|\ln(1 + x) - x| \le Cx^2$. Mithin ist für hinreichend grosse n

$$(1 + a_n/n)^n \le \exp\left(a_n + C|a_n|/n\right) \tag{6.3.3}$$

$$(1 + a_n/n)^n \ge \exp\left(a_n - C|a_n|/n\right). \tag{6.3.4}$$

Hieraus folgt offensichtlich die Behauptung.

Wir können nun unsere Kernaussage formulieren.

Lemma 6.11. *Sei ϕ eine zweimal differenzierbare Funktion auf $\mathbb{R}$ mit $\phi(0) = 1$ und $\phi'(0) = 0$. Dann gilt*

$$\lim_{n\to\infty} \left[\phi(n^{-1/2}t)\right]^n = \exp\left(+\frac{t^2}{2}\phi''(0)\right). \tag{6.3.5}$$

Beweis. Wir setzen

$$R_2(s) \equiv \phi(s) - 1 - \frac{s^2}{2}\phi''(0). \tag{6.3.6}$$

Wenn $\phi(t)$ zweimal differenzierbar ist, so bedeutet dies, da $\phi(0) = 1$ und $\phi'(0) = 0$ ist, dass

$$\lim_{|s|\downarrow 0} s^{-2}R_2(s) = 0,$$

also für jedes $t \in \mathbb{R}$,

$$\lim_{n\to\infty} \frac{n}{t^2}R_2(tn^{-1/2}) = 0.$$

Nun ist aber

$$\left[\phi(n^{-1/2}t)\right]^n = \left[1 + \frac{t^2}{2n}\phi''(0) + R_2(n^{-1/2}t)\right]^n$$

Damit erfüllt $a_n \equiv \frac{t^2}{2}\phi''(0) + nR_n(\sqrt{n}t)$ die Voraussetzung von Lemma 6.10 mit $a = \frac{t^2}{2}\phi''(0)$, und wir erhalten

$$\lim_{n\to\infty} \left[1 + \frac{t^2}{2n}\phi''(0) + R_2(tn^{-1/2})\right]^n = \exp\left(+\frac{t^2}{2}\phi''(0)\right) \tag{6.3.7}$$

Damit ist das Lemma bewiesen. □

Damit können wir nun unser Hauptresultat sehr leicht herleiten.

Satz 6.12 (Zentraler Grenzwertsatz). *Seinen X_i, $i \in \mathbb{N}$ unabhängige identisch verteilte Zufallsvariable mit $\mathbb{E}X_i = \mu$ und $\mathrm{var}(X_i) = \sigma^2 < \infty$. Dann konvergiert*

$$Z_n \equiv \frac{\sum_{i=1}^n (X_i - \mu)}{\sqrt{n}}$$

in Verteilung gegen eine Gauß'sche Zufallsvariable mit Mittelwert 0 und Varianz σ^2.

Anmerkung. In dieser Allgemeinheit wurde der Zentrale Grenzwertsatz 1922 von Jarl Waldemar Lindeberg [11] bewiesen, nachdem Lyapunov eine Version unter stärkeren Bedingungen schon 1901 gezeigt hatte.

Beweis. Wir nehmen ohne Beschränkung der Allgemeinheit $\mu = 0$ an. Offenbar konvergiert nach dem Vorhergehenden die charakteristische Funktion von Z_n gegen $\exp(-t^2\sigma^2/2)$ weil $-\phi''_{X_k}(0) = \mathrm{var}(X_k) = \sigma^2$. Diese kennen wir schon als charakteristische Funktion der Gaußverteilung $\mathcal{N}(0, \sigma^2)$. □

6.4 Stabile Verteilungen

Die Tatsache, dass die Normalverteilung im zentralen Grenzwertsatz auftaucht kann man auch anders als über den oben gezeigten Beweis verstehen. Man kann sich nämlich die Frage stellen, welche Eigenschaften überhaupt Zufallsvariablen haben müssen, die als Limes von reskalierten Summen wie in (6.1.1) auftreten. Wir nehmen wieder an, dass $\mathbb{E}X_i = 0$.

Dazu schreiben $p < 1$, und $q = 1-p$. Wir setzen $n = [pn]+[qn]$. Dann ist in Verteilung $S_n = S_{[pn]}+S'_{[qn]}$ wobei wir $S_{[pn]} = \sum_{i=1}^{[pn]} X_i$ und $S'_{[pn]} = \sum_{i=1}^{[qn]} X'_i$, wobei die $X'_i \equiv X_{[pn]+i}$. Offenbar ist dann

$$\begin{aligned} Z_n &= n^{-\gamma}\left(S_{[pn]} + S'_{[qn]}\right) \\ &= n^{-\gamma}[np]^{\gamma}[np]^{-\gamma}S_{[pn]} + n^{-\gamma}[nq]^{\gamma}[nq]^{-\gamma}S'_{[qn]} \\ &\sim p^{\gamma}Z_{[pn]} + q^{\gamma}Z'_{[qn]}, \end{aligned} \tag{6.4.1}$$

wo $Z_m \equiv m^{-\gamma}\sum_{k=1}^{m} X_k$, und Z'_m von Z_m unabhängig ist und die gleiche Verteilugn hat. Wenn nun Z_n in Verteilung gegen eine Zufallsvariable Z konvergiert, so konvergieren natürlich die Verteilungen von $Z_{[pn]}$ und $Z'_{[qn]}$ gegen Zufallsvariablen mit derselben Verteilung. Dass, heisst, Z muss die Eigenschaft haben, dass

$$Z \stackrel{\mathcal{D}}{=} p^{\gamma}Z + q^{\gamma}Z', \tag{6.4.2}$$

wo Z und Z' unabhängig sind und die gleiche Verteilung haben. Wir hatten schon gesehen, dass für $\gamma = \frac{1}{2}$, die Gaußverteilung gerade diese Eigenschaft hat. Man kann zeigen, dass die Gaußverteilung die einzige Verteilung ist, die diese Eigenschaft mit $\gamma = 1/2$ hat. Damit ist die Gaußverteilung in diesem Fall schon ein klarer Favorit.

Im Fall, dass die Varianz von X_i nicht endlich ist, schlägt das Argument für $\gamma = 1/2$ natürlich nicht mehr, und man kann sich dann die Frage nach einem Verteilungslimes mit allgemeineren γ stellen. Aus den obigen Betrachtungen sehen wir dann, dass im Ergebnis in jedem Fall nur eine Zufallvariable herauskommen kann, die die Gleichung (6.4.2) erfüllt. Die Verteilungen solcher Zufallsvariablen nennt man auch *stabile* Verteilungen (im engeren Sinn). Mit Hilfe solcher Verteilungen kann man in der Tat Verallgemeinerungen des zentralen Grenzwertsatzes für Zufallsvariablen die keine endliche Varianz haben herleiten. Es würde hier allerdings zu weit gehen, dieses Thema auszuführen.

Kapitel 7
Anwendungen in der Statistik

La probabilité de la plupart des événements simples est inconnue : en la considérant a priori*, elle nous paraît susceptible de toutes les valeurs comprises entre zéro et l'unité; mais, si l'on a observé un résultat composé de plusieurs de ces événements, la manière dont ils y entrent rend quelques-unes de ces valeurs plus probables que les autres. Ainsi, à mesure que le résultat observé se compose par le développement des événements simples, leur vraie possibilité se fait de plus en plus connaître, et il devient de plus en plus probable qu'elle tombe dans les limites qui, se resserant sans cesse, finiraient par coincider, si le nombre des événements simples devenait infini*[a].
Pierre Simon de Laplace, Théorie Analytique des Probabilités

[a] Die Wahrscheinlichkeit des meissten einfachen Ereignisse ist unbekannt: indem wir sie *a priori* betrachten, erscheinen alle Werte zwischen null und eins möglich; wenn man aber ein Ergebnis beobachtet, dass aus mehreren dieser Ereignisse zusammengesetzt ist, so macht die Art, wie diese eintreten, einige dieser Werte wahrscheinlicher als andere. So lässt sich, sofern das beobachtete Resultat sich aus der Entwicklung der einfachen Ereignisse zusammensetzt, ihre wirkliche Möglichkeit mehr und mehr erkennen, und es wird immer wahrscheinlicher, dass sie zwischen Schranken fällt, die, indem sie sich immer mehr zusammenziehen schlussendlich zusammenfielen, wenn die Zahl der einfachen Ereignisse unendlich würde.

7.1 Statistische Modelle und Schätzer

Die Aufgabe der Statistik ist die Beschreibung von Beobachtungen von "Zufallsexperimenten" durch ein auf ein auf Zufallsvariablen basiertem *Modell*. Ganz allgemein gesprochen sieht das so aus. Gegeben sind eine Folge von *Beobachtungen* (= Ausgänge von Zufallexperimenten), $Z_1, \ldots, Z_n$. Der Statistiker möchte diese als Realisierungen von n Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathfrak{F}, \mathbb{P})$ interpretieren. Er interessiert sich für die gemeinsame Verteilung der entsprechenden n Zufallsvariablen, die er a priori nicht kennt, sondern aus den Beobachtungen Z_i (interpretiert als einer Realisierung $\omega \in \Omega$), bestimmen, bzw. im statistischen Sprachgebrauch, *schätzen*.

Ohne weiteres ist dies praktisch nicht möglich, und man wird aufgrund von zusätzlichen "a priori" Informationen weitere Annahmen (Hypothesen) an

die Zufallsvariablen machen. Im allgemeinen besteht ein statistisches Modell somit aus Modellannahmen und Modellparametern, wobei die Annahmen als wahr angesehen werden, und dir Parameter zunächst unbekannt sind. Um die unbekannten Parameter zu bestimmen konstruiert der Statistiker nun sogenannte Schätzer, d.h. Funktionen der beobachteten Größen X_i, die die Werte der "wahren" Parameter annähren sollen. Die Schätzer, a_n, hängen dabei von n und von den Beobachtungen X_i, $i \leq n$ ab.

Eine wichtige Eigenschaft, die man von Schätzern fordert, ist die *Konsistenz*

Definition 7.1. Sei $X_n, i \in \mathbb{N}$ eine Families von Zufallsvariablen mit gemeinsamer Verteilung, die durch Parameter $a \in \mathbb{R}^k$ parametrisiert ist. Dann heisst eine Funktion $a_n : \mathbb{R}^n \to \mathbb{R}$ ein *konsistenter Schätzer* für die Parameter a, falls die Zufallsvariablen

$$a_n(X_1(\omega), \ldots, X_n(\omega)) \to a, \text{ f.s.}, \tag{7.1.1}$$

wenn $n \to \infty$.

Wir betrachten jetzt einige wichtige Beispiele.

7.1.1 Frequenzen

Seien unsere Beobachtungen X_i die Ausgänge von stets gleichen und sich nicht beeinflussenden Zufallsexperimenten, etwa eine Folge von Glücksspielen. Dann ist es eine plausible Annahme, dass die X_i durch unabhängige, gleichverteilte Zufallsvariablen mit gemeinsamer Verteilung ν zu modellieren sind. Hier ist also die Unabhängigkeit eine Modellannahmen, während die Verteilung, ν, zunächst ein unbekannter "Parameter" ist. Wie können wir aus den Beobachtungen ν schätzen?

Das Gesetz der großen Zahlen erlaubt es uns auf die Frage nach der Konvergenz der Frequenzen, die schon im ersten Abschnitt angesprochen war genauer einzugehen. Wir erinnern uns, dass wir in einer Reihe von n "identischen" Spiele (Zufallsexperimente) die Frequenzen der Ausgänge $X_i \in A$ definiert hatten als

$$\nu_n(A) \equiv \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_A(X_i). \tag{7.1.2}$$

Wir hatten damals gesagt, dass falls dies Frequenzen konvergieren, der Limes das einzige für eine Spielbank akzeptable Wahrscheinlichkeitsmaß ist. Folgen unabhängiger, identisch verteilter Zufallsvariablen sind nun genau das statistische Modell für eine solche Folge identischer, sich nicht beeinflussender Zufallsexperimente. Das Gesetz der großen Zahlen sagt uns dann, dass die Annahme der Konvergenz in der Tat korrekt war. Es gilt nämlich:

Lemma 7.2. *Seien $X_i, i \in \mathbb{N}$, eine Folge rellwertiger, unabhängiger, identisch verteilter Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathfrak{F}, \mathbb{P})$ mit Verteilung ν. Dann gilt, mit ν_n definiert durch (7.1.2),*

(i) Für jedes $A \in \mathfrak{B}(\mathbb{R})$ gilt

$$\nu_n(A) \to \nu(A) \quad \mathbb{P} - \text{f.s.}, \tag{7.1.3}$$

und

(ii) ν ist die Wahrscheinlichkeitsverteilung von X_1, i.e. für alle $A \in \mathfrak{F}$ gilt

$$\nu(A) = \mathbb{P}[X_1 \in A].$$

Beweis. Der Beweis ist denkbar einfach: Die Funktionen $\mathbb{1}_A(X_i)$ sind selbst Zufallsvariablen, und zwar, wie man leicht nachprüft, unabhängige. Ihre Erwartung ist gerade

$$\mathbb{E}[\mathbb{1}_A(X_i)] = \mathbb{P}[X_i \in A] = \mathbb{P}[X_1 \in A].$$

Da diese endlich sind, folgen beide Aussagen des Lemmas aus dem starken Gesetz der großen Zahlen. □

Die Sammlung der $\nu_n(A)$ stellt für jede Realisierung der Zufallsvariablen X_i ein Wahrscheinlichkeitsmaß auf den reellen Zahlen dar. Wir können damit ν_n auch als eine Abbildung von Ω in die Menge der Wahrscheinlichkeitsmaße über $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$ auffassen. Man nennt so etwas manchmal auch eine *maßwertige* Zufallsvariable.

Satz 7.3. *Seien $X_i, i \in \mathbb{N}$, eine Folge rellwertiger, unabhängiger, identisch verteilter Zufallsvariablen mit Verteilungsfunktion F auf einem Wahrscheinlichkeitsraum $(\Omega, \mathfrak{F}, \mathbb{P})$. Seien ν_n die oben definierten empirische Maße, und F_n die zugehörigen Verteilungsfunktionen. Dann gibt es eine Menge $\widetilde{\Omega} \subset \Omega$, mit $\mathbb{P}[\widetilde{\Omega}] = 1$, so dass, für alle $\omega \in \widetilde{\Omega}$,*

$$F_n^\omega \xrightarrow{\mathcal{D}} F. \tag{7.1.4}$$

Beweis. Wir wissen, dass ν_n von den Zufallsvariablem X_i abhängt, mithin also eine Funktion auf Ω. Wir machen diese Abhängigkeit für die zugehörigen Verteilungsfunktionen F_n^ω durch den Superskript ω explizit.

Wir wissen aus Lemma 7.2, dass für jedes $x \in \mathbb{R}$ wenn F bei x stetig ist, eine Teilmenge, $\widetilde{\Omega}_x$, vom Maß eins existiert, so dass für alle $\omega \in \widetilde{\Omega}_x$,

$$\lim_{n\to\infty} F_n^\omega(x) = F(x). \tag{7.1.5}$$

Nun ist auch, $\mathbb{P}[\cap_{q\in\mathbb{Q}} \widetilde{\Omega}_q] = 1$, so dass es auch eine Teilmenge vom Maß eins gibt, auf der (7.1.5) simultan für alle $x \in \mathbb{Q}$ gilt. Aber eine monotone Funktion, die auf einer dichten Teilmenge von $\mathbb{R}$ gegen eine Funktion F konvergiert,

konvergiert an jeder Stetigkeitstelle von F und hat einen eindeutigen rechtsstetigen Limes. □

Also, im Rahmen des statistischen Modells, in dem die Ausgänge eines Zufallsexperiments unabhängige, gleichverteilte Zufallsvariablen sind, sind die empirischen Verteilungen, d.h. die Frequenzen, tatsächlich *Schätzer* für die gemeinsame Verteilung dieser Zufallsvariablen, und dieser Schätzer ist darüberhinaus konsistent.

Mit der Chebychev'schen Ungleichung erhalten wir sogar eine *Qualitätsabschätzung.*

Lemma 7.4. *Seien $X_i, i \in \mathbb{N}$, eine Folge rellwertiger, unabhängiger, identisch verteilter Zufallsvariablen mit Verteilungsfunktion F auf einem Wahrscheinlichkeitsraum $(\Omega, \mathfrak{F}, \mathbb{P})$. Dann gilt, für jede Borelmenge A, dass*

$$\mathbb{P}\left[|\nu_n(A) - \nu(A)| > c\nu(A)\right] \leq \frac{1}{nc^2\nu(A)}. \tag{7.1.6}$$

Beweis. Übung! □

Wie man an der Abschätzung sieht, sind die Schätzungen für Mengen kleiner Masse fehlerhafter als die von großer Masse. Dies ist nur natürlich: Ist $\nu(A)$ klein, so bedarf er vieler Experimente, bis überhaupt einmal ein Ergebnis in A fällt! Die Qualität des Schätzers hängt also von der erwarteten Zahl der Ereignisse, die in A fallen, eben $n\nu(A)$, direkt ab.

Anmerkung. Es ist natürlich nicht praktikabel, alle Werte von $F(q)$, $q \in \mathbb{Q}$ gleichzeitig zu schätzen.

7.1.2 Schätzen von Erwartungswert und Varianz

Wir haben gesehen, dass Erwartungswert und Varianz einer Zufallsvariable bereits wichtige Informationen über deren Verteilung enthalten. Es liegt also für einen Statistiker nahe, zunächst mal diese Kenngrößen zu schätzen, als gleich die ganze Verteilung. Das Gesetz der großen Zahlen liefert uns wieder Kandidaten für solche Schätzer sowie eine Rechtfertigung. Betrachten wir zunächst den Mittelwert einer Verteilung. Nach dem Gesetz der großen Zahlen konvergiert ja das *empirische Mittel*,

$$m_n \equiv n^{-1} \sum_{i=1}^{n} X_i \tag{7.1.7}$$

fast sicher gegen $\mu \equiv \mathbb{E}X_1$, falls die X_i unabhängige, identisch Verteilte Zufallsvariablen sind. Damit ist die Zufallsvariable m_n, gut geeignet, um als

Schätzer für den Mittelwert zu dienen. Darüber hinaus hat dieser Schätzer noch die Eigenschaft, dass

$$\mathbb{E}m_n = \mu.$$

Solche Schätzer nennt man in der Statistik *"erwartungstreu"*, oder *"unvoreingenommen"* (Englisch "un-biased"). Vielfach (aber nicht immer) wird diese Eigenschaft gefordert, um einem Schätzer vor anderen den Vorzug zu geben. Der Punkt ist dabei, dass wir zu jedem Schätzer (genauer gesagt einer Folge von Schätzern) noch eine Nullfolge dazu addieren können, und eine andere Familie von Schätzern zu bekommen, die auch gegen den gesuchten Schätzwert konvergiert. So könnten wir etwa alternativ zu m_n die Größe

$$\widetilde{m}_n \equiv \frac{1}{n-1}\sum_{i=1}^{n} X_i$$

wählen. Sicher konvergiert auch $\widetilde{m}_n = \frac{n}{n-1}m_n$ fast sicher gegen m, aber

$$\mathbb{E}\,\widetilde{m}_n = \frac{n}{n-1}\mu \neq \mu.$$

Dieser Schätzer hätte also die Tendenz, den Mittelwert leicht zu überschätzen. Betrachten wir nun wieder die Zuverlässigkeit des Schätzers. Wir begnügen uns mit dem Fall, dass die X_1 endliche zweite Momente haben. Dann liefert die Chebychev Ungleichung sofort:

Lemma 7.5. *Seien $X_i, i \in \mathbb{N}$, unabhängige, gleichverteilte Zufallsvariablen mit Mittelwert μ und mit endlicher Varianz σ^2. Dann ist m_n ein erwartungstreuer Schätzer für μ und es gilt*

$$\mathbb{P}[|m_n - \mu| > c\mu] \leq \frac{\sigma^2}{n\mu^2 c^2}. \tag{7.1.8}$$

Wir sehen, dass die Qualität des Schätzers erheblich von Verhältnis σ^2/μ^2 abhängt. In der Praxis will man sich ja eine gewisse Genauigkeit der Schätzung vorgeben, und dann n so wählen, dass diese erzielt wird. Dabei soll natürlich n so klein wie möglich sein, da in der Regel die Durchführung eines Zufallsexperimentes Kosten verursacht.

Nun kennen wir natürlich μ und σ^2 nicht, wir wollen μ ja gerade bestimmen. Was μ angeht, ist das nicht so tragisch, da wir ja zumindest den Schätzer m_n haben. Allerdings reicht das noch nicht aus, um eine "Stoppregel" für das benötigte n zu entwickeln, da wir dazu auch σ^2 brauchen. Also sollten wir besser auch gleich versuchen, einen Schätzer für die Varianz zu finden und gleich mitzuberechnen. Naheliegend ist wieder die *empirische Varianz*, d.h. die Varianz der empirischen Verteilung ν_n:

$$V_n \equiv \nu_n(X - \nu_n(X))^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - m_n)^2\,, \tag{7.1.9}$$

wobei $X = (X_1, \ldots, X_n)$. Wir zeigen zunächst, dass dieser Schätzer fast sicher gegen die Varianz konvergiert, falls σ^2 endlich ist.

Lemma 7.6. *Seien $X_i, i \in \mathbb{N}$, wie in Lemma 7.5 und sei $\mathrm{var}(X_i) = \sigma^2$. Dann konvergiert die Zufallsvariable V_n fast sicher gegen σ^2.*

Beweis. Zum Beweis schreiben wir V_n leicht um:

$$V_n = \frac{1}{n} \sum_{i=1}^{n} X_i^2 - m_n^2.$$

Nach Voraussetzung sind die X_i^2 unabhängige, gleichverteilte Zufallsvariablen mit endlicher Erwartung. Daher konvergiert die erste Summe, wegen dem starken Gesetz, fast sicher

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i^2 = \mathbb{E}X_1^2 \quad \text{f.s..}$$

Andererseits wissen wir, dass $m_n \to \mu$, f.s., und somit auch $m_n^2 \to \mu^2$, f.s.. Daraus folgt, dass

$$\frac{1}{n} \sum_{i=1}^{n} X_i^2 - m_n^2 \to \mathbb{E}X_1^2 - (\mathbb{E}X_1)^2 = \sigma^2 \quad \text{f.s.,}$$

was wir behauptet haben. □

Wir wollen noch nachprüfen, ob V_n erwartungstreu ist. Da man nachrechnet, dass

$$\mathbb{E}V_n = \frac{n-1}{n}\sigma^2,$$

ist dies offenbar nicht der Fall. Man findet natürlich leicht einen erwartungstreuen Schätzer für die Varianz, der ebenfalls fast sicher gegen σ^2 konvergiert, nämlich

$$V_n^* \equiv \frac{n}{n-1} V_n = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - m_n)^2 . \tag{7.1.10}$$

Dieser Ausdruck hat den Charme anzudeuten, dass nach einer Beobachtung die Varianz noch als unendlich geschätzt werden sollte (während eine einzige Beobachtung bereits einen endlichen erwartungstreuen Schätzer für das Mittelwert liefert. Natürlich ist dieser für praktische Belange ziemlich unbrauchbar). Die Forderung der Erwartungstreue ist ansonsten etwas willkürlich, und nicht oft sub-optimal. Wenn wir die Qualität des Schätzers für die Varianz bestimmen wollten, so könnten wir wie bei m_n vorgehen, benötigten dann aber wieder höhere Momente von X_1, die wiederum geschätzt werden müssten, etc.

Immerhin sehen wir, dass wir mit Hilfe unserer Schätzer m_n und V_n^* bereits ein praktisches Verfahren zur qualitätskontrollierten Schätzung des Mit-

telwertes haben. Dazu ersetzen wir in der Abschätzung (7.1.8) für die Wahrscheinlichkeit einer Abweichung des Schätzers m_n vom wahren Wert μ, die Größen μ und σ^2 durch ihre Schätzer. Dies liefert uns einen Schätzer für den wahren Fehler, der zumindest die gute Eigenschaft hat, fast sicher gegen eine obere Schranke zu konvergieren. Damit liegt folgende Strategie nahe: Wir suchen einen Schätzer für μ, der mit höchstens Wahrscheinlichkeit ϵ um mehr als $c\mu$ falsch liegt. Dann berechnen wir sukzessive m_n, V_n bis zu einem Wert n^* wo erstmals

$$\frac{V_{n^*}^2}{n^* m_{n^*}^2 c^2} < \epsilon.$$

7.2 Parameterschätzung

Wir hatten im vorigen Kapitel gesehen, wie das Gesetz der großen Zahlen verwendet werden kann um Schätzer sowohl für Wahrscheinlichkeitsverteilungen als auch Erwartungswert und Varianz zu konstruieren. Allerdings hatten wir auch gesehen, dass es schwierig und aufwendig ist, Wahrscheinlichkeitsverteilungen zu schätzen. Es wäre für praktische Zwecke wesentlich einfacher, wenn wir bereits a priori etwas über die Wahrscheinlichkeitsverteilung der zugrundeliegenden Zufallsvariablen wüssten, und nur noch einige wenige *Parameter* identifizieren müssten. Der zentrale Grenzwertsatz ist *ein* wesentliches Resultat, dass in gewissen Situationen solche von wenigen Parametern indizierten Klassen von Verteilungen suggeriert, hier nämlich gerade die *Gaußverteilung*. Nehmen wir etwa als Model an, dass X_i eine Familie von unabhängigen und identisch Gauß-verteilten Zufallvariablen sein, so bleiben als Parameter nur noch Mittelwert und Varianz zu schätzen, was wir bereit können.

Ein interessanteres Beispiel ist die sogenannte lineare Regression. Wir betrachten etwa einen zeitabhängigen Vorgang, $f(t) \in \mathbb{R}$, $t \in \mathbb{R}_+$, zu gewissen Zeiten $t_1 < t_2 < \cdots < t_n$. Jede Beobachtung liefert einen Messwert z_i. Idealerweise wäre $z_i = f(t_i)$, aber durch Fehler ist diese Gleichung verfälscht und wir sollen annehmen, dass die Differenz eine Zufallsvariable ist. Unsere Aufgabe ist, aus den Beobachtungen einen Schätzer für f zu gewinnen, und gleichzeitig eine Qualitätsabschätzung für den Schätzer, sowie einen Schätzer für die Verteilung der Fehler, finden.

Ohne weitere Vorabinformation ist dieses Problem praktisch unlösbar, da es unendlich viele Parameter involviert. Wir müssen also vereinfachende Annahmen machen. Zuächst betrachten wir den Fall, in dem wir annehmen, dass $f(t) = a + bt$ eine lineare Funktion ist, wobei a und b unbekannte, zu bestimmende Parameter sind. Weiter nehmen wir an, dass die Messfehler unabhängige, identisch verteilte Zufallsvariablen, X_i sind. Dann sind unsere Beobachtungen (im Rahmen des Modells) beschrieben als Zufallsvariablen

$$Z_i = a + bt_i + X_i. \tag{7.2.1}$$

Eine weitere Vereinfachung träte ein, wenn wie einschränkende Annahmen an die Verteilung der X_i machen könnten. Hier greift nun der zentrale Grenzwertsatz: wenn wir der Überzeugung sind, dass die Fehler X_i sich als Summen vieler kleiner "Elementarfehler", die unseren Messapparat beeinflussen, ergeben, dann liegt es nahe anzunehmen, dass die X_i gaußverteilt sind, mit unbekanntem Mittelwert, μ, und Varianz, σ^2. Wir haben also ein vier-parametriges *Modell* für unsere Beobachtungen, mit Parametern a, b, μ, σ^2 (wobei wir leicht sehen, dass wir in unserem Fall zwischen a und μ nicht unterscheiden können, und daher nur hoffen können, dass $\mu = 0$, d.h. dass unsere Messungen keinen systematischen Fehler aufweisen). Die Aufgabe der Statistik ist es nun, Schätzer für diese Parameter zu finden (also Familien von Zufallsvariablen, die, wenn die Z_i durch dieses Modell beschrieben werden), gegen diese Parameter konvergieren. Eine solche Familie von Schätzern nennt man *konsistent*. Letzlich ist dies eigentlich noch nicht genug: wir würden auch gerne wissen, ob unsere Modellannahmen plausibel waren!

7.2.1 Das Maximum-Likelihood Prinzip

Eine einleuchtende Idee zu solchen Schätzern zu kommen besteht darin, die Parameter so zu schätzen, dass den beobachteten Werten, X_i, die größte Wahrscheinlichkeit zukommt. Betrachten wir dazu zunächst ein sehr einfaches Beispiel: Wir beobachten eine Folge von Münzwürfen, $z_1, \ldots, z_n \in \{0, 1\}$. Wir wollen diese modellieren als Realisierung von unabhängigen, identisch verteilten Bernoulli Zufallsvariablen, X_i, mit Parameter p. Aus den Beobachtungen wollen wir nun den Wert von p schätzen. Das Maximum-likelihood Prinzip sagt, man schätze $p = p(z_1, \ldots, z_n)$, so dass die Wahrscheinlichkeit der Beobachtungen maximal wird, also dass

$$\begin{aligned} \varrho_n(p; z_1, \ldots, z_n) &\equiv \mathbb{P}[X_1 = z_1 \wedge X_2 = z_2 \wedge \cdots \wedge X_n = z_n] \qquad (7.2.2) \\ &= \prod_{i=1}^{n} p^{z_i}(1-p)^{1-z_i} \end{aligned}$$

maximal wird. Wir nennen $\varrho_n(p; z_1, \ldots, z_n)$ die *likelihood Funktion* für unser Modell.

Um dasjenige p zu bestimmen, dass $\varrho_n(p; z_1, \ldots, z_n)$ maximiert, suchen wir zunächst einen kritischen Punkt dieser Funktion, d.h. wir lösen die Gleichung

$$\begin{aligned} 0 = \frac{\mathrm{d}}{dp}\varrho_n(p; z_1, \ldots, z_n) &= \sum_{i=1}^{n} \left(\frac{z_i}{p} - \frac{1-z_i}{1-p} \right) \prod_{i=1}^{n} p^{z_i}(1-p)^{1-z_i} \\ &= \varrho_n(p; z_1, \ldots, z_n) \sum_{i=1}^{n} \left(\frac{z_i}{p(1-p)} - \frac{1}{1-p} \right). \end{aligned}$$

Diese Gleichung hat als einzige Lösung

$$p = p_n^* = p_n^*(z_1, \dots, z_n) = \frac{1}{n} \sum_{i=1}^{n} z_i.$$

Da $z_i \in \{0, 1\}$ liegen, ist $z_i = \mathbb{1}_{z_i=1}$, so dass der Maximum-Likelihood Schätzer für die Wahrscheinlichkeit von $\{X_i = 1\}$ gerade gleich der Frequenz des Auftretens von 1 ist, der uns ja schon als konsistenter Schätzer bekannt ist. In diesem Fall liefert das Maximum-likelihood Prinzip also nichts neues, gibt aber eine interessante alternative Interpretation des Schätzers.

Als nächstes betrachten wir das interessantere Beispiel der Regression in dem oben beschriebenen Gauß'schen Modell. Hier ist es allerdings so, dass wegen der Stetigkeit der Gaußverteilung die Wahrscheinlichkeit jeder Beobachtung gleich null ist. Es liegt aber nahe, als "likelihood Funktion" statt der Wahrscheinlichkeit der Beobachtung die Wahrscheinlichkeitsdichte zu wählen, also

$$\begin{aligned} \varrho_n(a, b, \sigma^2; z_1, \dots, z_n) &\equiv \prod_{i=1}^{n} \rho_{0,\sigma^2}(z_i - a - bt_i) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(z_i - a - bt_i)^2}{2\sigma^2}\right). \end{aligned} \tag{7.2.3}$$

Das maximum-likelihood Prinzip sagt nun, dass der maximum-likelihood Schätzer für a, b, σ^2, $a_n^*, b_n^*, (\sigma^2)_n^*$, dadurch gegeben ist, dass

$$\varrho_n(a_n^*, b_n^*, (\sigma^2)_n^*; z_1, \dots, z_n) \equiv \max_{a,b \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+} \varrho_n(a, b, \sigma^2; z_1, \dots, z_n) \tag{7.2.4}$$

Natürlich hängt der maximum-likelihood Schätzer von den Beobachtungen z_i ab, ist also eine Zufallsvariable.

In unserem Fall ist die Lösung des Maximierungsproblems recht einfach. Es empfiehlt sich, anstatt direkt ϱ_n zu maximieren, dessen Logarithmus,

$$\ln \varrho_n(a, b, \sigma^2; z_1, \dots, z_n) = -\frac{n}{2} \ln(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{(z_i - a - bt_i)^2}{2\sigma^2},$$

zu maximieren. Dies führt auf die drei Gleichungen

$$\frac{\partial \ln \varrho_n}{\partial a} = 0 \leftrightarrow \sum_{i=1}^{n}(z_i - a - bt_i)/\sigma^2 = 0,$$
$$\frac{\partial \ln \varrho_n}{\partial b} = 0 \leftrightarrow \sum_{i=1}^{n} t_i(z_i - a - bt_i)/\sigma^2 = 0,$$
$$\frac{\partial \ln \varrho_n}{\partial \sigma^2} = 0 \leftrightarrow \sum_{i=1}^{n}(z_i - a - bt_i)^2/2\sigma^4 - \frac{n}{2\sigma^2} = 0.$$

Es folgt

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(z_i - a - bt_i)^2 \tag{7.2.5}$$
$$a = \frac{1}{n}\sum_{i=1}^{n}(z_i - bt_i) \tag{7.2.6}$$
$$b = \frac{\sum_{i=1}^{n} t_i(z_i - a)}{\sum_{i=1}^{n} t_i^2} \tag{7.2.7}$$

und weiter, mit $T_n = \sum_{i=1}^{n} t_i$,

$$b_n^* = \frac{\sum_{i=1}^{n} t_i z_i - \frac{T_n}{n}\sum_{i=1}^{n} z_i}{\sum_{i=1}^{n} t_i^2 - \frac{T_n^2}{n}}. \tag{7.2.8}$$

Nachdem b explizit bekannt ist kann nun a und σ^2 ebenfalls exlizit durch Einsetzen ausgerechnet werden:

$$a_n^* = \frac{1}{n}\sum_{i=1}^{n}(z_i - b_n^* t_i), \tag{7.2.9}$$
$$(\sigma^2)_n^* = \frac{1}{n}\sum_{i=1}^{n}(z_i - a_n^* - b_n^* t_i)^2. \tag{7.2.10}$$

Wesentlich zu bemerken ist aber, dass die Gleichungen (7.2.6) und (7.2.7) besagen, dass a und b so gewählt werden müssen, dass der durch (7.2.5) gegebene Ausdruck für σ^2 als Funktion von a und b minimiert wird. Letzterer ist aber gerade die Summe der Quadrate der Abweichung des Beobachtung vom theoretischen Wert. Mit anderen Worten, die maximum-likelihood Methode liefert im Fall der Gaußverteilung gerade die Methode der *kleinsten Quadrate* für die Schätzung der Parameter a und b.

Wir wollen noch nachprüfen, ob bzw. wann unsere Schätzer gut sind, d.h., ob sie im Fall, dass unsere Modellannahme richtig war, d.h. ob, wenn die z_i durch die Zufallsvariablen (7.2.1) gegeben sind, $a_n^* \to a$, $b_n^* \to b$ und $(\sigma^2)_n^* \to \sigma^2$ konvergieren. Dazu stellen wir als erstes fest, dass unsere Schätzer für a und b erwartungstreu sind. Indem wir (7.2.1) in (7.2.8) einsetzen, sehen

wir nämlich leicht, dass

$$\begin{aligned}\mathbb{E}b_n^* &= \frac{\sum_{i=1}^n t_i \mathbb{E}Z_i - \frac{T_n}{n}\sum_{i=1}^n \mathbb{E}Z_i}{\sum_{i=1}^n t_i^2 - \frac{T_n^2}{n}} \\ &= \frac{\sum_{i=1}^n t_i(a+bt_i) - \frac{T_n}{n}\sum_{i=1}^n (a+bt_i)}{\sum_{i=1}^n t_i^2 - \frac{T_n^2}{n}} \\ &= \frac{b\sum_{i=1}^n t_i^2 + a\sum_{i=1}^n t_i - T_n a - b\frac{T_n^2}{n}}{\sum_{i=1}^n t_i^2 - \frac{T_n^2}{n}} \\ &= b.\end{aligned}$$

Weiter ist dann auch

$$\mathbb{E}a_n^* = \frac{1}{n}\sum_{i=1}^n \mathbb{E}(Z_i - b_n^* t_i) = \frac{1}{n}\sum_{i=1}^n (a + bt_i - \mathbb{E}b_n^* t_i) = a.$$

Der Schätzer für σ^2 wird dagegen nicht erwartungstreu sein, was uns aber hier nicht bekümmern soll.

Als nächstes fragen wir nach der Konsistenz. Wir betrachten dabei der Einfacheit halber nur den Fall $t_i = i/n$, womit dann $T_n = (n+1)/2$. Offenbar ist

$$\begin{aligned}b_n^* &= \frac{\sum_{i=1}^n t_i(a+bt_i+X_i) - \frac{T_n}{n}\sum_{i=1}^n (a+bt_i+X_i)}{\sum_{i=1}^n t_i^2 - \frac{T_n^2}{n}} \\ &= b + \frac{\sum_{i=1}^n t_i X_i - \frac{T_n}{n}\sum_{i=1}^n X_i}{\sum_{i=1}^n t_i^2 - \frac{T_n^2}{n}}.\end{aligned}$$

Wir wollen zeigen, dass der zweite Term nach null konvergiert. Dabei benutzen wir diesmal, dass die Variablen X_i gaußverteilt sind, und daher dasselbe für die hier auftretenden Summen gilt. Wir können zum Beispiel die exponentielle Markov-Ungleichung (Korollar 5.6) benutzen um zu zeigen, dass

$$\mathbb{P}\left[\left|\sum_{i=1}^n X_i\right| \geq C_n\sqrt{n}\right] \leq 2e^{-C_n^2/2\sigma^2} \tag{7.2.11}$$

und

$$\mathbb{P}\left[\left|\sum_{i=1}^n t_i X_i\right| \geq C_n\sqrt{\sum_{i=1}^n t_i^2}\right] \leq 2e^{-C_n^2/2\sigma^2} \tag{7.2.12}$$

(**Übung:** Beweise die Abschätzungen (7.2.11) und (7.2.12)!) Wenn wir $C_n = 2\sigma\sqrt{\ln n}$ wählen, so sind diese Wahrscheinlichkeiten summierbar, die betreffenden Ereignisse treten also mit Wahrscheinlichkeit 1 nur endlich oft auf. Daher haben wir fast sicher für alle bis auf endlich viele Werte von n,

$$
\begin{aligned}
|b_n^* - b| &\le C_n \frac{\sqrt{\sum_{i=1}^n t_i^2} + \frac{T_n}{\sqrt{n}}}{\sum_{i=1}^n t_i^2 - \frac{T_n^2}{n}} \qquad (7.2.13)\\
&= C_n n^{-1/2} \frac{\sqrt{(n+1)(2n+1)/6} + (n+1)/2}{(n+1)(n-1)/12n}\\
&\le C_n C n^{-1/2} = 2C\sigma n^{-1/2} \ln n \to 0 \text{ für } n \to \infty,
\end{aligned}
$$

mit C eine numerische Konstante (z.B. 25).

Weiter ist

$$a_n^* - a = \frac{1}{n}\sum_{i=1}^n X_i + \frac{1}{n}\sum_{i=1}^n t_i(b - b_n^*).$$

Der erste Term der rechten Seite konvergiert wegen dem Gesetz der großen Zahlen fast sicher gegen Null; der zweite ist wegen (7.2.13) fast sicher für alle bis auf endliche viele n kleiner als

$$CC_n T_n/n \le C' n^{-1/2} \ln n,$$

(mit einer numerischen Konstanten C') und konvergiert damit auch fast sicher gegen null. Damit sind also bereits b_n^* und a_n^* konsistente Schätzer. Schließlich bleibt noch $(\sigma^2)_n^*$ zu betrachten. Hier ist

$$
\begin{aligned}
(\sigma^2)_n^* &= \frac{1}{n}\sum_{i=1}^n (X_i + (a - a_n^*) + (b - b_n^*)t_i)^2 \qquad (7.2.14)\\
&= \frac{1}{n}\sum_{i=1}^n X_i^2\\
&+ \frac{1}{n}\sum_{i=1}^n \left(2X_i((a - a_n^*) + (b - b_n^*)t_i) + ((a - a_n^*) + (b - b_n^*)t_i)^2\right).
\end{aligned}
$$

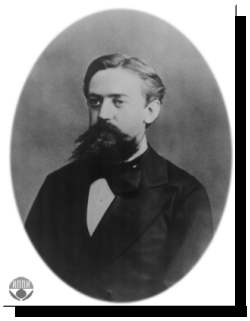
Der erste Term strebt fast sicher gegen σ^2 nach dem Gesetz der großen Zahlen, und die letzte Zeile konvergiert fast sicher gegen null, wie man unter Benutzung der bisherigen Abschätzungen mit einiger Rechnung zeigen kann.

Die maximum-likelihood Methode liefert uns also tatsächslich eine konsistente Familie von Schätzern. Ein großer Vorteil der Methode ist es, in sehr vielfältigen Situationen anwendbar zu sein.

Kapitel 8
Markov Prozesse

Un des grands avantages du Calcul des Probabilités est d'apprendre à se défier des premiers aperçus. Comme on reconnaît qu'ils trompent souvent lorsqu'on peut les soumettre cu calcul, on doit en conclure que sur d'autres objets il ne faut s'y livrer qu'avec une circonspection extrème[a].
Pierre Simon de Laplace, Théorie Analytique des Probabilités

[a] Ein großen Nutzen der Wahrscheinlichkeitsrechnung ist es uns zu lehren den ersten Eindrücken zu misstrauen. Da man feststellt, dass diese da wo man sie mit mit Berechnungen konfrontieren kann, oft täuschen, so muss man schliessen, dass man sich ihnen in anderen Gegenständen nur mit der äusserster Umsicht ausliefern darf.



In den bisherigen 7 Kapiteln haben wir die grundlegenden Begriffe der Wahrscheinlichkeitstheorie kennengelernt und insbesonders die zwei wichtigsten Sätze, das Gesetz der Großen Zahlen und den zentralen Grenzwertsatz hergeleitet. Dabei waren unabhängige Zufallsvariablen unser Grundbaustein, und alle unsere Resultate betrafen Objekte, die aus solchen konstruiert waren, insbesonder Summen und deren Grenzwerte.

In diesem Teil der Vorlesung wollen wir erstmals über unabhängige Zufallsvariablen hinausgehen und eine in vielen Anwendungen wichtige Klasse von *stochastischen Prozessen*, die sogenannten *Markov Prozesse* behandeln. Diese sind in vieler Hinsicht die wichtigsten stochastischen Prozesse überhaupt. Der Grund dafür ist, dass sie einerseits so vielseitig sind, dass sehr viele dynamischen Prozesse mit ihrer Hilfe modelliert werden können, andererseits aber mathematisch noch einigermaßen behandelbar sind. Wir werden in dieser Vorlesung natürlich nur einige wenige, einfache Beispiele dieser reichen Klasse betrachten. Markov Prozesse wurden von Andrey Andreyevich Markov (1856-1922) eingeführt.

8.1 Definitionen

Der Begriff des stochastischen Prozesses verallgemeinert den der Folgen unabhängiger Zufallsvariablen beziehungsweise der Summen solcher, wie wir sie

in den vorherigen Kapiteln betrachtet haben. Bausteine sind Familien von Zufallsvariable X_t, die für gegebenes t Werte in einem Raum S, dem sogenannten *Zustandsraum*, annehmen. In der Regel wird S eine Teilmenge von $\mathbb{R}$, oder von $\mathbb{R}^d$, $d \geq 1$ sein, man kann aber auch allgemeinere Räume zulassen. t nimmt Werte in einer sogenannten *Indexmenge*, I an. Die wichtigsten Beispiele sind $I = \mathbb{N}_0$ und $I = \mathbb{R}_+$, wobei wir uns hier auf den einfacheren Fall $I = \mathbb{N}_0$ einschränken wollen. Wir interpretieren den Index t als *Zeit*, und fassen X_t als Zustand eines Systems zur Zeit t auf. Der stochastische Prozess $\{X_t\}_{t\in I}$ ist als Familie von Zufallsvariablen definiert auf einem Wahrscheinlichkeitsraum $(\Omega, \mathfrak{F}, \mathbb{P})$ zu verstehen. Im Fall, dass $I = \mathbb{N}_0$ können wir natürlich $\Omega = S^{\mathbb{N}_0}$, und $\mathfrak{F} = \mathfrak{B}(S)^{\otimes \mathbb{N}_0}$, also den unendlichen Produktraum, wählen.

Alternativ zu der Definition 3.17 können wir einen stochastischen Prozess mit diskreter Zeit auch als eine messbare Abbildung mit Werten im Folgenraum $S^{\mathbb{N}_0}$ auffassen:

Definition 8.1. Sei $(\Omega, \mathfrak{F}, \mathbb{P})$ ein abstrakter Wahrscheinlichkeitsraum. Eine messbaren Abbildungen von $(\Omega, \mathfrak{F}) \to (S^{\mathbb{N}_0}, \mathfrak{B}(S)^{\otimes \mathbb{N}_0})$ heißt ein *Stochastischer Prozess* mit Zustandsraum S und Indexmenge $\mathbb{N}_0$.

Eine wichtige Größe ist selbstverständlich die Verteilung des Prozesses X, formal gegeben durch das Maß $P_X \equiv \mathbb{P} \circ X^{-1}$. P_X ist dann ein Wahrscheinlichkeitsmaß auf $(S^{\mathbb{N}_0}, \mathfrak{B}(S)^{\otimes \mathbb{N}_0})$.

Eine besonders wichtige Klasse von stochastischen Prozessen sind die sogenannten *Markovprozesse*. Sie stellen in gewisser Weise das stochastische Analogon zu dynamischen Systemen dar und spielen in der Modellierung des dynamischen Verhaltens vieler Systeme eine große Rolle. Wir werden in dieser Vorlesung nur eine spezielle Unterklasse von Markovprozessen, die sogenannten Markovketten mit diskreter Zeit, betrachten. Dabei ist der Zustandsraum eine zunächst eine endliche Menge.

Definition 8.2. Ein stochastischer Prozess mit diskreter Zeit und endlichem Zustandsraum S heißt eine *Markovkette*, genau dann, wenn, für alle $n \in \mathbb{N}_0$, und $t_1 < t_2 < \cdots < t_n$, $x_1, \ldots, x_n \in S$, so dass

$$\mathbb{P}\left[X_{t_{n-1}} = x_{n-1}, X_{t_{n-2}} = x_{n-2}, \ldots, X_{t_1} = x_1\right] > 0,$$

gilt

$$\begin{aligned} &\mathbb{P}\left[X_{t_n} = x_n | X_{t_{n-1}} = x_{n-1}, X_{t_{n-2}} = x_{n-2}, \ldots, X_{t_1} = x_1\right] \qquad (8.1.1)\\ &= \mathbb{P}\left[X_{t_n} = x_n | X_{t_{n-1}} = x_{n-1}\right]. \end{aligned}$$

Anmerkung. Dieselbe Definition kann auch im Fall abzählbarer Zustandsräume verwandt werden. Im allgmeineren Fall überabzählbarer Zustandsräume tritt aber das Problem auf, dass alle betrachteten Ereignisse Wahrscheinlichkeit Null haben könnten. Um dieses Problem zu lösen werden wir den Begriff der bedingten Wahrscheinlichkeit so verallgemeinern müssen, dass auch auf

Ereignisse mit Wahrscheinlichkeit Null bedingt werden kann. Dies wird aber erst Gegenstand der Vorlesung Stochastische Prozesse sein.

Aufgrund der Diskretheit der Zeit können wir in (8.1.1) natürlich $t_i = i$ wählen und erhalten dann, dass

$$\begin{aligned} &\mathbb{P}\left[X_n = x_n | X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, \ldots, X_1 = x_1\right] \\ &= \mathbb{P}\left[X_n = x_n | X_{n-1} = x_{n-1}\right] \equiv p_{n-1}(x_{n-1}, x_n). \end{aligned} \tag{8.1.2}$$

Satz 8.3. *Die Wahrscheinlichkeitsverteilung einer Markovkette mit diskreter Zeit ist eindeutig bestimmt durch die Angabe der Anfangsverteilung, $\pi_0(x)$, $x \in S$ und der Übergangswahrscheinlichkeiten $p_n(x, y)$, $n \in \mathbb{N}, x, y \in S$. Umgekehrt gibt es für jedes Wahrscheinlichkeitsmaß π_0 auf $(S, \mathfrak{B}(S))$ und einer Sammlung von Zahlen $p_n(x, y)$ mit der Eigenschaft, dass, für alle $n \in \mathbb{N}$ und alle $x \in S$,*

$$\sum_{y \in S} p_n(x, y) = 1, \tag{8.1.3}$$

eine Markovkette mir Übergangswahrscheinlichkeiten $p_n(x, y)$ und Anfangsverteilung π_0.

Anmerkung. Mann bezeichnet p_n auch als Übergangsmatrix. Eine Matrix mit der Eigenschaft (8.1.3) nennt man auch *stochastische Matrix*.

Beweis. Wir zeigen, dass die endlich dimensionalen Verteilungen festgelegt sind. Da wir auf einem endlichen Raum S arbeiten, genügt es offenbar für alle $n \in \mathbb{N}$, und alle $x_i \in S$, $i \leq n$, alle Wahrscheinlichkeiten der Form

$$\mathbb{P}[X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_1 = x_1, X_0 = x_0]$$

zu kennen. Nun ist aber wegen der Markoveigenschaft (5.5) und der Definition der bedingten Wahrscheinlichkeit

$$\begin{aligned} &\mathbb{P}[X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_1 = x_1, X_0 = x_0] \\ &= \mathbb{P}[X_n = x_n | X_{n-1} = x_{n-1}] \mathbb{P}[X_{n-1} = x_{n-1}, \ldots, X_1 = x_1, X_0 = x_0] \\ &= p_{n-1}(x_{n-1}, x_n) \mathbb{P}[X_{n-1} = x_{n-1}, \ldots, X_1 = x_1, X_0 = x_0] \\ &= p_{n-1}(x_{n-1}, x_n) p_{n-2}(x_{n-2}, x_{n-1}) \mathbb{P}[X_{n-2} = x_{n-2}, \ldots, X_1 = x_1, X_0 = x_0] \\ &= p_{n-1}(x_{n-1}, x_n) p_{n-2}(x_{n-2}, x_{n-1}) \ldots p_0(x_0, x_1) \mathbb{P}[X_0 = x_0] \\ &= p_{n-1}(x_{n-1}, x_n) p_{n-2}(x_{n-2}, x_{n-1}) \ldots p_0(x_0, x_1) \pi_0(x_0). \end{aligned} \tag{8.1.4}$$

Die Frage, ob es eine Verteilung des Prozesses gibt, die diese endlich dimensionalen Verteilungen besitzt, wollen wir hier noch nicht im Detail angehen. Dies wird in der Vorlesung "Stochastische Prozesse" getan werden. Wir bemerken lediglich, dass die so berechneten Verteilungen kompatibel sind in dem Sinne, dass

$$\begin{aligned}&\mathbb{P}[X_{n-1} = x_{n-1}, \dots, X_1 = x_1, X_0 = x_0] \\ &= \sum_{x_n \in S} \mathbb{P}[X_n = x_n, X_{n-1} = x_{n-1}, \dots, X_1 = x_1, X_0 = x_0]\end{aligned} \tag{8.1.5}$$

was aber aus der expliziten Formel (8.1.4) und der Eigenschaft (8.1.3) sogleich folgt. □

8.2 Markovketten mit stationären Überganswahrscheinlichkeiten

Nach diesem allgemeinen Bemerkungen wollen wir uns zunächst nur mit dem einfachsten, aber bereits interessanten Spezialfall befassen, in dem

(i) der Zustandsraum, S, eine endlich Menge ist, also $S = \{1, \dots, d\}$, $d \in \mathbb{N}$, und
(ii) die Übergangswahrscheinlichkeiten $p_{n-1}(x, y)$ nicht von n abhängen.

Man nennst solche Markovketten *zeitlich homogene* oder Markovketten oder Markovketten mit stationären Übergagnswahrscheinlichkeiten.

Beispiel. Ein sehr einfaches Beispiel für eine stationäre Markovkette ist folgendes (recht schlechtes) Klimamodell. Wir wollen dabei das Wetter auf die Grundfrage "Regen oder Sonnenschein" reduzieren. Das Wetter am Tag n soll also durch eine Zufallsvariable X_n die die Werte 0 (=Regen) und 1 (=Sonne) annimmt beschrieben werden. Versucht man diese durch unabhängige Zufallsvariablen zu beschreiben, stellt man fest, dass dies mit den Beobachtungen nicht kompatibel ist: längere Perioden mit konstantem Regen oder Sonnenschein treten in Wirklichkeit häufiger auf als das Modell vorhersagt. Man überlegt sich, dass es sinnvoll scheint, die Prognose des Wetters morgen davon abhängig zu machen, wie das Wetter heute ist (aber nicht davon wie es gestern und vorgestern war). Dies führt auf die Beschreibung durch eine Markovkette mit den Zuständen 0 und 1, und Übergangswahrscheinlichkeiten

$$\begin{aligned}p(0,1) = p_{0,1}, \quad p(0,0) = p_{0,0} = 1 - p_{0,1}, \\ p(1,0) = p_{1,0}, \quad p(1,1) = p_{1,1} = 1 - p_{1,0}.\end{aligned} \tag{8.2.1}$$

Zusammen mit der Anfangsverteilung $\pi(0) = p_0, \pi(1) = p_1 = 1 - p_0$ legt dies eine Markovkette fest. Wie sehen, dass wir nun 3 freie Parameter zur Verfügung haben, mit denen wir im Zweifel das Wetter besser fitten können.

Wir sehen, dass die Übergangswahrscheinlichkeiten einer stationären Markovkette eine $d \times d$ Matrix, P, bilden. Diese Matrix nennt man auch die Übergangsmatrix der Markovkette. Zusammen mit dem *Vektor* der Anfangsverteilung, π_0, legt diese die Wahrscheinlichkeitsverteilung einer Markovkette vollständig fest, d.h. Wahrscheinlichkeiten beliebiger Ereignisse lassen sich

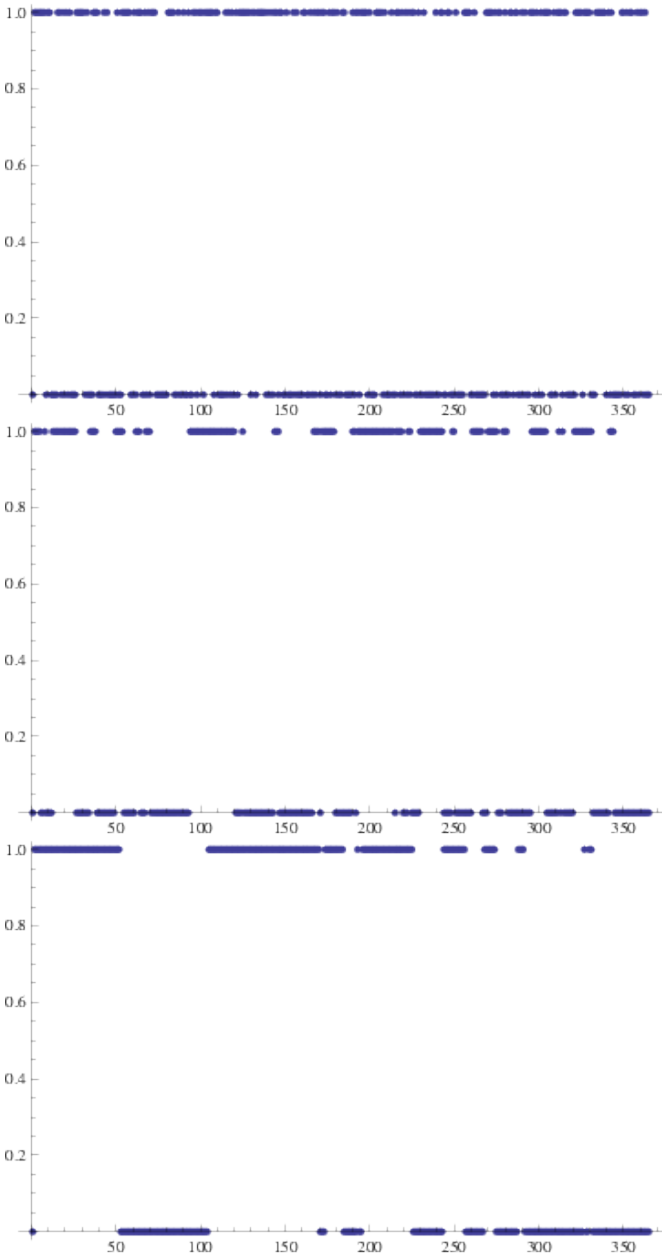


Abb. 8.1 Ein Jahresverlauf des "Wetters" in unserem Modell mit $p_{01} = p_{10} = 0.5$, 0.15, und 0.05.

durch diese Objekte ausdrücken. Durch diese Beobachtung begründet sich ein enger Zusammenhang zwischen Markovketten und der linearen Algebra.

Übergangsmatrizen sind freilich keine beliebigen Matrizen, sondern sie haben eine Reihe von wichtigen Eigenschaften.

Lemma 8.4. *Sei P die Übergangsmatrix einer stationären Markovkette mit Zustandsraum $S = \{1, \dots, d\}$. Seien p_{ij} die Elemente von P. Dann gilt:*

(i) Für alle $i, j \in S$ gilt $1 \geq p_{ij} \geq 0$.
(ii) Für alle $i \in S$ gilt $\sum_{j \in S} p_{ij} = 1$.

Umgekehrt gilt: Jede Matrix die (i) und (ii) erfüllt, ist die Übergangsmatrix einer Markovkette.

Beweis. Die beiden ersten Eigenschaften sind offensichtlich, weil ja für jedes i, $p_{i,\cdot} = \mathbb{P}[X_{n+1} = \cdot | X_n = i]$ eine Wahrscheinlichkeitsverteilung auf S ist. Der Umkehrschluss folgt aus Satz 8.3. □

Matrizen die die Eigenschaften (i) und (ii) aus Lemma 8.4 erfüllen heissen *stochastische Matrizen.* Wir wollen uns die Übergangsmatrizen für einige Beispiele von Markovketten ansehen.

- **Unabhängige Zufallsvariablen.** Schon eine Folge unabhängiger, identisch verteilter Zufallsvariablen ist eine Markovkette. Hier ist

$$p_{ij} = \mathbb{P}[X_n = j | X_{n-1} = i] = \mathbb{P}[X_0 = j] = \pi_0(j),$$

  d.h. alle Zeilen der Matrix P sind identisch gleich dem Vektor der die Anfangsverteilung der Markovkette angibt.
- **Irrfahrt mit Rand.** Auch Summen unabhängiger Zufallsvariablen sind Markovketten. Wir betrachten den Fall, dass X_i unabhängige Rademachervariablen mit Parameter p sind, also eine Irrfahrt. In der Tat ist

$$\mathbb{P}[S_n = j | S_{n-i} = i] = \begin{cases} p, & \text{falls } j = i+1 \\ 1-p, & \text{falls } j = i-1 \\ 0, & \text{sonst} \end{cases} \tag{8.2.2}$$

  allerdings ist in diesem Fall der Zustandsraum abzählbar unendlich, nämlich $\mathbb{Z}$. Wir können eine Variante betrachten, in dem die Irrfahrt angehalten wird, wenn sie auf den Rand des endlichen Gebiets $[-L, L]$ trifft. Dazu modifizieren wir die Überangswahrscheinlichkeiten aus (8.2.2) für den Fall $i = \pm L$, so dass

$$\mathbb{P}[S_n = j | S_{n-i} = \pm L] = \begin{cases} 1, & \text{falls } i = \pm L \\ 0, & \text{sonst} \end{cases} \tag{8.2.3}$$

  Die Übergangsmatrix hat dann folgende Gestalt:

$$P = \begin{pmatrix} 1 & 0 & 0 & \dots & \dots & \dots & 0 \\ 1-p & 0 & p & 0 & \dots & \dots & 0 \\ 0 & 1-p & 0 & p & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 1-p & 0 & p & 0 \\ 0 & \dots & \dots & 0 & 1-p & 0 & p \\ 0 & \dots & \dots & \dots & 0 & 0 & 1 \end{pmatrix}$$

- **Unser Wettermodell (8.2.1)**. Hier ist

$$P = \begin{pmatrix} 1-p_{0,1} & p_{0,1} \\ p_{1,0} & 1-p_{1,0} \end{pmatrix}$$

Das der Zusammenhang zwischen Markovketten und Matrizen nicht nur oberflächlich ist, zeigt sich daran, dass in der Berechnung verschiedener Wahrscheinlichkeiten tatsächlich Matrixoperationen auftauchen. So ist

$$\mathbb{P}[X_n = j | X_0 = i] = \sum_{i_1, i_2, \dots, i_{n-1}} p_{ii_1} p_{i_1 i_2} \dots p_{i_{n-2} i_{n-1}} p_{i_{n-1} j} = (P^n)_{ij} \,.$$

Man schreibt gelegentlich für die bedingte Wahrscheinlichkeit $\mathbb{P}[X_n = j | X_0 = i] = P_n(i,j)$ und nennt diesen Ausdruck den *Propagator*. Es folgt, dass

$$\pi_n(j) \equiv \mathbb{P}[X_n = j] = \sum_{i \in S} \pi_0(i) P_n(ij) = (\pi_0 P^n)_j \,. \tag{8.2.4}$$

Wir sehen also, dass die Verteilung der Markovkette zur Zeit n durch die Wirkung der Matrix P^n von links auf die Anfangsverteilung gegeben ist.

8.3 Invariante Verteilungen

Eine der ersten Fragen, die man sich stellen wird, ist, ob Verteilungen, π_0, gibt, die unter der Wirkung der Markovkette *invariant* sind.

Definition 8.5. Sei X eine Markovkette mit disketer Zeit, endlichem Zustandsraum S und stationären Übergangswahrscheinlichkeiten P. Dann heisst ein Wahrscheinlichkeitsmaß, π_0, *invariante Verteilung*, wenn für alle $n \in \mathbb{N}$ und alle $j \in S$,

$$\pi_n(j) = \pi_0(j), \tag{8.3.1}$$

gilt.

Offensichtlich ist wegen der Gleichung (8.2.4), die Frage nach invarianten Verteilungen äquivalent zur Frage nach links-Eigenwerten der Matrix P:

Lemma 8.6. *Sei P eine stochastische Matrix. Dann ist π_0 genau dann eine invariante Verteilung für eine stationäre Markovkette mit Übergangsmatrix P, wenn π_0 ein links-Eigenvektor von P zum Eigenwert 1 ist, mit $\pi_0(i) \geq 0$ und $\sum_{i\in S} \pi_0(i) = 1$.*

Beweis. Wir kombinieren (8.3.1) mit (8.2.4) und erhalten, dass π_0 invariant ist, wenn

$$\pi_0(i) = (\pi_0 P)_i \,. \tag{8.3.2}$$

Wenn andererseits ein Vektor mit positiven Komponenten deren Summe gleich eins ist die Gleichung (8.3.2) erfüllt, so liefert er eine invariante Anfangsverteilung. □

Satz 8.7. *Jede stationäre Markovkette mit endlichem Zustandsraum besitzt mindestens eine invariante Verteilung.*

Beweis. Der Beweis ist am einfachsten mit Hilfe eines tiefen Resultats der linearen Algebra, dem Perron-Frobenius Theorem zu führen. Dieses lautet wie folgt.

Satz 8.8 (Perron-Frobenius 2). *Sei $A \neq 0$ eine $d \times d$ Matrix mit nicht-negativen Einträgen. Sei λ_0 definiert als Supremum über all $\lambda \in \mathbb{R}$ für die es einen Vektor x mit nicht-negativen reellen Elementen gibt, so dass*

$$\sum_{i=1}^{d} x_i = 1, \quad \text{und} \quad (Ax)_i \geq \lambda x_i, \forall i = 1, \ldots, d. \tag{8.3.3}$$

Dann gilt

(i)λ_0 ist ein Eigenwert mit Eigenvektor x mit nicht-negativen Elementen.
(ii) Alle anderen Eigenwerte, λ, von A erfüllen $|\lambda| \leq \lambda_0$.
(iii) Wenn λ Eigenwert von A ist und $|\lambda| = \lambda_0$, dann ist $\lambda/\lambda_0 \equiv \eta$ eine Wurzel der Eins (d.h. es gibt $k \in \mathbb{N}$, so dass $\eta^k = 1$) und $\eta^m \lambda_0$ ist für alle $m \in \mathbb{N}$ ein Eigenwert von A.

Wir wollen diesen Satz nun auf den Fall anwenden, wo A die Übergangsmatrix, P, einer Markovkette ist. Da P die Voraussetzunges des Satzes von Perron-Frobenius erfüllt, exisitiert ein maximaler positiver Eigenwert λ_0 und ein zugehöriger (Links-) Eigenvektor v der nichtnegative Einträge hat und die Normierung $\sum_i v_i = 1$ erfüllt. Wir müssen nur noch zeigen, dass $\lambda_0 = 1$ gilt. Dazu schreiben wir die Eigenwertgleichung $(vP)_i = \lambda_0 v_i$, für $i = 1, \ldots, d$ und summieren über i. Da P stochastisch ist, gilt dann

$$\lambda_0 \sum_{i=1}^{d} v_i = \sum_{j=1}^{d} \sum_{i=1}^{d} v_j p_{ji} = \sum_{j=1}^{d} v_j. \tag{8.3.4}$$

Da $\sum_{i=1}^{d} v_i = 1$, folgt $\lambda_0 = 1$. v liefert damit eine invariante Verteilung. □

Nach der Existenz sind die Fragen der Eindeutigkeit und der Konvergenz naheliegend. Diese gestalten sich etwas komplexer.

8.3.1 Markovketten und Graphen. Klassifizierung der Zustände

Es erweist sich als instruktiv mit einer Übergangsmatrix einen gerichteten Graphen auf dem Zustandsraum S zu verbinden. Wir fassen die Menge S als Knotenmenge eines (gerichteten) Graphen, $(S, \mathcal{E})$ auf. Wir sagen, dass $\mathcal{E}$ die *Kante*, (i, j), $i \in S, j \in S$ enthält, $(i, j) \in \mathcal{E}$, wenn $p_{ij} > 0$. Graphisch stellen wir dies durch einen Pfeil dar.

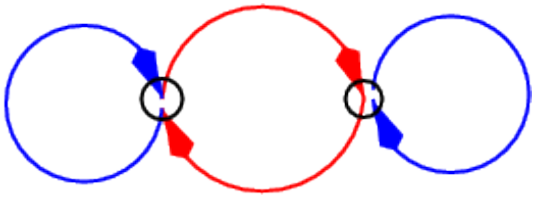


Abb. 8.2 Der Graph der Markovkette unseres Wettermodells

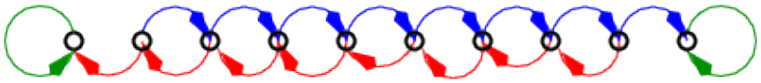


Abb. 8.3 Der Graph der am Rand gestoppten Irrfahrt

Definition 8.9. Ein *Pfad* γ in einem gerichteten Graphen $(S, \mathcal{E})$ ist eine Folge $\gamma = (e_1, e_2, \ldots, e_k)$ von Kanten $e_\ell \in \mathcal{E}$, so dass für jedes $\ell = 1, \ldots, k-1$ gilt, dass der Endpunkt von e_ℓ der Anfangspunkt von $e_{\ell+1}$ ist. γ verbindet i mit j falls der Anfangspunkt von e_1 i und der Endpunkt von e_k j ist.

Definition 8.10. Zwei Knoten, $i, j \in S$ einem gerichteten Graphen *kommunizieren*, wenn Pfade gibt, die i mit j verbinden und solche, die j mit i verbinden. WIr sagen auch, dass jeder Zustand mit sich selbst kommuniziert.

Man kann leicht nachprüfen, dass die Relation “kommunizieren” eine Äquivalenzrelation ist. Nun definiert eine Äquivalenzrelation eine Zerlegung der Menge S in Äquivalenzklassen. Wir bezeichnen die Äquivalenzklassen kommunizierender Zustände als *kommunizierde Klassen* oder einfach als *Klassen*.

Definition 8.11. Eine Markovkette heißt *irreduzibel* genau dann wenn der Zustandsraum aus einer einzigen Klasse besteht.

Anmerkung. Beachte, dass eine Markovkette deren Graph nicht zusammenhängend ist, auch nicht irreduzibel ist. Wenn der Graph einer Markovkette zusammenhängend ist, muss diese aber noch lange nicht irreduzibel sein.

Lemma 8.12. *Eine Markovkette ist genau dann irreduzibel, wenn es für jedes Paar, $(i,j) \in S \times S$, ein $k \in \mathbb{N}_0$ gibt, so dass $\left(P^k\right)_{i,j} > 0$.*

Beweis. Es gilt

$$\begin{aligned}\left(P^k\right)_{ij} &= \sum_{i_1,i_2,\ldots,i_{k-1}} p_{ii_1}p_{i_1i_2}\ldots p_{i_{k-1}j} \\ &= \sum_{\substack{\gamma:i\to j \\ |\gamma|=k}} p_{e_1}p_{e_2}\ldots p_{e_k} \qquad (8.3.5)\end{aligned}$$

Die rechte Seite ist offenbar genau dann positiv, wenn es einen solchen Weg gibt. Daraus folgt das Lemma direkt. □

Die Bedeutung der Aussage des letzten Lemmas erschließt sich aus dem sog. ersten Perron-Frobenius Theorem.

Satz 8.13 (Perron-Frobenius 1). *Sei A eine $d \times d$ Matrix mit* strikt *positiven Einträgen. Dann gibt es einen Vektor, $\mathbf{x}$, mit* strikt *positiven Komponenten, so dass $A\mathbf{x} = \lambda_0\mathbf{x}$. Der Eigenwert λ_0 ist einfach, und für alle anderen Eigenwerte, λ_i, von A, gilt $|\lambda_i| < \lambda_0$.*

Die Anwendung auf unsere Markovketten ist wie folgt:

Satz 8.14. *Sei P die Übergangsmatrix einer Markovkette mit endlichem Zustandsraum und es gebe $k \in \mathbb{N}$ so, dass die Matrix P^k nur strikt positive Einträge hat. Dann gibt es genau eine invariante Verteilung, μ, mit $\mu P = \mu$, und*

$$\lim_{n\to\infty} P^n = \Pi_0$$

existiert und ist eine stochastische Matrix vom Rang 1 deren Zeilen gerade durch den Vektor μ gegeben sind, d.h.

$$\Pi_0 = \begin{pmatrix} \mu(1) & \mu(2) & \ldots & \mu(d) \\ \mu(1) & \mu(2) & \ldots & \mu(d) \\ \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ \mu(1) & \mu(2) & \ldots & \mu(d) \end{pmatrix}. \qquad (8.3.6)$$

Insbesondere konvergiert für jede Anfangsverteilung π_0 die Verteilung $\pi_n = \pi_0 P^n$ gegen die einzige invariante Verteilung μ.

Anmerkung. Markovketten, für die die Aussage des Theorems 8.14 gilt, d.h. die eine einzige invariante Verteilung, μ, besitzen gegen welche die Verteilung π_t für jede Anfangsverteilung π_0 konvergiert, nennt man auch *ergodisch*. Die Aussage des Theorems kombiniert mit Lemma 8.19 ist dann, dass jede irreduzible, aperiodische Markovkette mit endlichem Zustandsraum ergodisch ist.

Beweis. Nach Voraussetzung erfüllt die Matrix $A = P^k$ die Voraussetzungen des ersten Perron-Frobenius Satzes (Satz 8.13). Insbesondere besitzt P^k einen einzigen maximalen Eigenwert 1 mit Eigenvektor μ, der strikt positive Einträge hat. Andererseits wissen wir, dass P mindestens einen maximalen Eigenwert 1 hat. Sei nun ν ein Eigenvektor von P mit Eigenwert λ und $|\lambda| = 1$. Dann gilt auch $\nu P^k = \lambda^k \nu$, und notwendig $\lambda^k = 1$. Somit muss $\nu = \mu$ sein. Damit gibt es aber nur einen Eigenwert von P der Betrag 1 hat, und daher folgt $\lambda = 1$. Alle anderen Eigenwerte sind im Betrag strikt kleiner als 1. Daher können wir P zerlegen als

$$P = \Pi_0 + Q, \tag{8.3.7}$$

wobei Π_0 der in (8.3.6)angegebene Projektor auf den eindimensionalen Eigenraum zum Eigenwert 1 (und zwar sowohl bezüglich der Wirkung nach rechts als nach links) ist, und Q bildet den dazu orthogonalen Unterraum auf sich ab. Nämlich:

(i) $\Pi_0^2 = \Pi_0$, und
(ii)$\Pi_0 Q = Q\Pi_0 = 0$.

Beide Aussagen folgen durch Nachrechnen.

Als nächstes zeigen wir, dass jeder Eigenwert der Matrix $Q \equiv P - \Pi_0$ im Betrag strikt kleiner als eins ist. Gilt nämlich $\mathbf{v}Q = \lambda\mathbf{v}$, so haben wir

$$\lambda\mathbf{v}\Pi_0 = \mathbf{v}Q\Pi_0 = 0. \tag{8.3.8}$$

und daher ist, falls nicht $\lambda = 0$, $\mathbf{v}\Pi_= $, und daher $\mathbf{v}P = \mathbf{v}(\Pi_0 + Q) = \mathbf{v}Q = \lambda\mathbf{v}$. Damit ist aber entweder $|\lambda| < 1$, oder $\lambda = 1$. Im letzteren Fall ist aber $\mathbf{v} = \mu$, und somit dann $\mathbf{v}Q = 0$, im Widerspruch zur Annahme $\lambda = 1$. Es bleibt also nur die Möglichkeit $|\lambda| < 1$.

Wir benötigen nun ein weiteres Resultat aus der linearen Analysis:

Lemma 8.15. *Sei B eine $d \times d$-Matrix. Dann besitzt B einen Eigenwert vom maximalen Betrag, r, und sei $\|\cdot\|$ eine Norm auf dem Raum der Matrizen (d.h. $\|B\| \equiv \sum_{\mathbf{v}\in\mathbb{R}^d} \frac{\|B\mathbf{v}\|}{\|\mathbf{v}\|}$, wo $\|\mathbf{v}\|$ eine beliebige Norm auf $\mathbb{R}^d$ ist). Dann gilt*

$$r = \limsup_{n\uparrow\infty} \|B^n\|^{1/n}. \tag{8.3.9}$$

Beweis. Jede Matrix B kann durch eine nicht-singuläre Transformation auf die Jordan-Normalform gebracht werden, d.h. es existiert eine invertierbare

Matrix U, so dass $U^{-1}BU = J$, wo J blockdiagonal ist und jeder Block entweder diagonal ist oder die Form

$$\begin{pmatrix} \lambda_i & 1 & 0 & 0 & \dots & 0 \\ 0 & \lambda_i & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & & \\ \dots & \dots & \dots & \dots & & \\ 0 & \dots & 0 & 0 & \lambda_i & 1 \\ 0 & \dots & 0 & 0 & 0 & \lambda_i \end{pmatrix} \tag{8.3.10}$$

hat, wo λ_i die Eigenwerte von B sind. Insbesondere ist J von der Form $J = D + N$, wo D diagonal ist, D und N kommutieren, und N nilpotent ist, d.h. $N^d = 0$. Daraus folgt, dass (für $n \geq d$)

$$J^n = \sum_{k=0}^{d-1} \binom{n}{k} D^{n-k} N^k, \tag{8.3.11}$$

und somit

$$\|J^n\| \leq \sum_{k=0}^{d-1} \|D\|^{n-k} \|N\|^k n^k = r^n \sum_{k=0}^{d-1} r^{-k} \|N\|^k n^k. \tag{8.3.12}$$

Wenn wir hier die n-te Wurzel ziehen und dann den Grenzwert $n \uparrow \infty$ betrachten, erhalten wir

$$\limsup_{n\uparrow\infty} \|J^n\|^{1/n} \leq r \lim_{n\uparrow\infty} \left(\sum_{k=0}^{d-1} r^{-k} \|N\|^k n^k \right)^{1/n} = r. \tag{8.3.13}$$

Da U und U^{-1} beschränkt sind, folgt auch dass

$$\limsup_{n\uparrow\infty} \|B^n\|^{1/n} \leq \lim_{n\uparrow\infty} \|J^n\|^{1/n} \|U\|^{1/n} \|U^{-1}\|^{1/n} = r. \tag{8.3.14}$$

Die Schranke in die umgekehrte Richtung ist einfacher. Wir benutzen nur, dass für jedes $n \geq 1$, und jeden Eigenwert λ mit Eigenvektor $\mathbf{v}$,

$$\|B^n\| \geq \frac{\|B^n \mathbf{v}\|}{\|\mathbf{v}\|} = |\lambda|^n, \tag{8.3.15}$$

also $\|B^n\|^{1/n} \geq \lambda$. $\square$

In unserem Fall ist aber $r < 1$. Dann folgt aus dem Lemma, dass für jedes $\epsilon > 0$, für alle hinreichend grossen n $\|Q^n\|_\infty \leq (r + \epsilon)^n$. Da wir so wählen können, dass $\rho + \epsilon < 1$, folgt das für alle $\mathbf{v}$

$$\lim_{n\uparrow\infty} \|Q^n\| = 0. \tag{8.3.16}$$

Da weiter $P^n = \Pi_0 + Q^n$, so folgt für alle Anfangsverteilungen π_0, dass

$$\lim_{n\uparrow\infty} \pi_0 P^n = \pi_0 \Pi_0 = \mu, \tag{8.3.17}$$

was der Behauptung entspricht. □

Anmerkung. Der Beweis von Satz 8.14 folgt dem Buch von Karlin und Taylor [7]. Man kann den Satz 8.14 auch ohne Verwendung der Sätze von Perron und Frobenius führen, siehe z.B. das Buch von Georgii [5]. Ich halte aber es aber für interessant und lehrreich, den Zusammenhang zwischen diesen Gebieten zu betonen. Insbesondere liefert der Beweis auch eine Kontrolle der Konvergenzgeschwindigkeit, nämlich $\|\pi_0 P^n - \mu\| \leq C|\lambda_1|^n$, wo λ_1 der Eigenwert von P mit zweitgrößtem Betrag ist.

Wir wollen uns nun Fragen, für welche Markovketten die Vorraussetzung des Satzes 8.14 gelten. Klar ist, dass irreduzibilität eine *notwendige* Bedingung ist, die aber noch nicht ausreicht.

Ein weiteres wichtiges Konzept ist die Periodizität.

Definition 8.16. Wir sagen, dass ein Zustand i Periode $d(i)$ hat, wenn $d(i)$ der größte gemeinsame Teiler aller Zahlen $n \in \mathbb{N}$ ist für die $(P^n)_{i,i} > 0$. Ein Zustand mit Periode 1 heißt *aperiodisch*.

Lemma 8.17. *Wenn $i, j \in S$ kommunizieren, dann ist $d(i) = d(j)$.*

Beweis. Wir wissen, das es n und m gibt, so dass $P^n_{j,i} > 0$ und $P^m_{i,j} > 0$. Sei nun $P^\ell_{i,i} > 0$. Dann ist auch

$$P^{n+\ell+m}_{j,j} \geq P^n_{j,i} P^\ell_{i,i} P^m_{i,j} > 0.$$

Da auch $P^{2\ell}_{i,i} > 0$, ist auch $P^{n+2\ell+m}_{j,j} > 0$, so dass $d(j)$ sowohl $n + m + \ell$ als auch $n + m + 2\ell$ teilt. Mithin teilt es auch die Differenz dieser Zahlen, nämlich ℓ. Das gilt für alle ℓ für die $P^\ell_{i,i} > 0$, deshalb ist $d(j) \leq d(i)$. Da wir das Argument auch umdrehen können, folgt genausogut, dass $d(i) \leq d(j)$, mithin die Behauptung. □

Lemma 8.18. *Wenn $i \in S$ Periode $d(i)$ hat, dann gibt es $N \in \mathbb{N}$, so dass für alle $n \geq N$, $\left(P^{n\,d(i)}\right)_{i,i} > 0$.*

Beweis. Die Behauptung folgt aus der zahlentheoretischen Tatsache, dass, wenn $n_1, \ldots, n_k$ natürliche Zahlen mit größtem gemeinsamen Teiler d sind, es ein $M \in \mathbb{N}$ gibt, so dass für alle $m \geq M$, $d\,m$ als Linearkombination der n_i geschrieben werden kann,

$$d\,m = \sum_{i=1}^{k} c_i n_i, \tag{8.3.18}$$

wo $c_i \in \mathbb{N}_0$ sind[1].

□

Lemma 8.19. *Eine irreduzible und aperiodische Markovkette mit endlichem Zustandsraum hat die Eigenschaft, dass es ein $k \in \mathbb{N}$ gibt, so dass für alle $i, j \in S$, $\left(P^k\right)_{i,j} > 0$.*

Beweis. Wegen der vorhergehenden Sätze wissen wir, dass existiert $M \in \mathbb{N}$ so dass für alle $m \geq M$, $P^m_{j,j} > 0$. Man kann M unabhängig von j nehmen, weil S endlich ist. Andererseits gibt es für jedes (i, j) ein $n_{i,j}$ so, dass

$$P^{n_{i,j}}_{i,j} > 0.$$

Wenn $P^m_{j,j} > 0$, was für alle großen m der Fall ist, ist dann auch

$$P^{n_{i,j}+m}_{i,j} > 0.$$

Deshalb gilt für $k \geq M + \max_{i,j} n_{i,j}$, dass $(P^k)_{i,j} > 0$. □

Irreduzible und aperiodische Markovketten sind in der Praxis von großer Wichtigkeit. Darüber hinaus kann man auch Resultate für diese Ergebnisse für den allgemeinen Fall zusammenbasteln.

Der Ergodensatz nutzt die Aperiodizät entscheidend aus. Er kann in dieser Form für periodische Markovketten auch nicht richtig sein. Es gilt aber für nur irreduzible Markovketten immer noch, dass Sie eine einzige invariante Verteilung besitzen.

Satz 8.20. *Sei P die Übergangsmatrix einer irreduziblem Markovkette mit endlichem Zustandsraum. Dann besitzt P genau eine invariante Verteilung μ und es gilt, dass für alle $i \in S$, $\mu(i) > 0$.*

Beweis. Der Beweis ist denkbar einfach. Wir definieren für $\epsilon > 0$ die Matrix $P_\epsilon \equiv \epsilon \mathbb{1} + (1 - \epsilon)P$. Dann haben wir folgende elementare Eigenschaften:

(i) P_ϵ ist eine stochastische Matrix.

[1] Der Beweis dieser Tatsache ist nicht sonderlich schwer: Es seine zunächst A die Menge aller Zahlen die durch die rechte Seite von (8.3.18) dargestellt werden können, und sodann B die Menge aller ganzzahligen Linearkombinationen aus Zahlen von A. Es sei dann d' die kleinste positive Zahl in B. Sei nun $N > 0$ eine Zahl in A die nicht durch d' teilbar ist. Dann sind $d' - N$ sowie $N - \ell d'$ für jedes $\ell \geq 1$ in B und ungleich Null. Aber eine dieser Zahlen muss dann kleiner als d' sein, weswegen d' gemeinsamer Teiler aller Zahlen aus A ist, inbesondere also auch alle n_i teilt. Ganz ähnlich zeigt man, dass es auch keinen grösseren gemeinsamen Teiler aller Zahlen aus A geben kann, und damit auch keinen grösseren gemeinsamen Teiler der n_i. Also ist $d' = d$. Indem man die Gleichung (8.3.18) durch d teilt kann man sich auf den Fall $d = 1$ zurückziehen. Es folgt dann aus dem bisher gesagten, dass es N_1, N_2 aus A gibt, so dass $N_1 - N_2 = 1$. Nun sei $m > N_2^2$, also $m = N_2^2 + \ell$, mit $\ell \in \mathbb{N}$. Dann ist $m = N_2^2 + bN_2 + j(N_1 - N_2)$ mit $0 \leq j < N_2$. Man kann sich nun davon überzeugen, dass dies die gewünschte Darstellung von m ergibt.

(ii) Die von P_ϵ erzeugte Markovkette ist irreduzibel und aperiodisch.
(iii) P und P_ϵ besitzen die gleichen Eigenvektoren.
(iv) $\mu P = \mu$ gilt genau dann wenn $\mu P_\epsilon = \mu$.

Nun wissen wir, dass P_ϵ einen einfachen Eigenwert 1 mit strikt positivem Eigenvektor besitzt. Damit folgt dasselbe auch für P, was zu beweisen war. □

Die einfachste periodische Kette ist die mit Übergangmatrix

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

Diese besitzt die Eigenwerte 1 und -1, und die invariante Verteilung $\mu = (1/, 1/2)$. Hier gibt es aber auch einen Eigenvektor, $\nu = (1, -1)$ mit Eigenwert -1. Man auch leicht, dass

$$P^n = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}^n = \begin{cases} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, & \text{wenn } n \text{ ungerade ist,} \\ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, & \text{wenn } n \text{ gerade ist.} \end{cases}$$

Hier konvergiert P^n also nicht. Klarerweise konvergiert dann auch $\pi_0 P^n$ für allgemeine Anfangsverteilungen aber nicht gegen die invariante Verteilung.

Wenn X eine ergodische Markovkette und μ ihre einzige invariante Verteilung ist, so bezeichnet man die Verteilung, $\mathbb{P}_\mu$, dieses Prozesses mit Anfangsverteilung $\pi_0 = \mu$ auch als stationäre Verteilung. Es gilt dann insbesondere, dass

$$\mathbb{P}_\mu[(X_0, X_1, \dots) \in A] = \mathbb{P}_\mu[(X_n, X_{n+1}, \dots) \in A],$$

für alle $n \in \mathbb{N}_0$ und alle $A \in \mathfrak{B}(S)^{\otimes \mathbb{N}_0}$. Es gilt in der Tat, dass die Verteilung einer ergodischen Markovkette gegen diese stationäre Verteilung konvergiert, in dem Sinne, dass für alle $A \in \mathfrak{B}(S)^{\otimes \mathbb{N}_0}$ und alle $x \in S$,

$$\lim_{n\to\infty} |\mathbb{P}_\mu[(X_0, X_1, \dots) \in A] - \mathbb{P}_x[(X_n, X_{n+1}, \dots) \in A]| = 0.$$

Der Beweis ist sehr einfach und nutzt die definierende Eigenschaft einer Markovkette:

$$\begin{aligned} & |\mathbb{P}_\mu[(X_0, X_1, \dots) \in A] - \mathbb{P}_x[(X_n, X_{n+1}, \dots) \in A]| \\ = & \left| \sum_{y\in S} (\mathbb{P}_x[X_n = y] - \mu(y))\, \mathbb{P}_y[(X_0, X_1, \dots) \in A] \right| \\ \leq & \sum_{y\in S} |\mathbb{P}_x[X_n = y] - \mu(y)| \to 0 \quad \text{wenn } n \to \infty. \end{aligned}$$

8.3.2 Die Sätze von Perron und Frobenius

Wie viele Dinge in der Theorie der Markov Ketten, sind die Sätze von Perron und Frobenius Gegenstand der linearen Algebra. Wegen ihrer Bedeutung geben wir hier trotzdem die Beweise an. Wir beginnen mit dem ersten Satz von Perron-Frobenius.

Beweis. (von Satz 8.8) Es ist nun A eine $n \times n$ Matrix mit reellen Einträgen. Wir betrachten die Menge

$$\Lambda \equiv \left\{ \lambda \in \mathbb{R} : \exists \mathbf{x} \in \mathbb{R}^n, \sum_{i=1}^{n} x_i = 1, x_i \geq 0 \; \forall_{i=1}^{n} A\mathbf{x} \geq \lambda \mathbf{x} \right\}. \tag{8.3.19}$$

Wir setzten $\lambda_0 = \sup\{\lambda \in \Lambda\}$. Es ist zunächst klar, dass $\lambda_0 > 0$ sein muss.

Des weitern existiert eine Folge γ_i die nach λ_0 konvergiert und Vektoren $\mathbf{x}^i$ mit nicht-negativen Einträgen (und mindestens einem strikt positiven Eintrag), so dass $A\mathbf{x}^i \geq \gamma_i \mathbf{x}^i$ und $\sum_j \mathbf{x}^i_j = 1$. Wegen der Kompaktheit des Raumes der betrachteten Vektoren existieren Folgen $k_j \uparrow \infty$, so dass

$$\lim_{j \uparrow \infty} \mathbf{x}^{k_j} = \mathbf{x}^0. \tag{8.3.20}$$

Dabei hat $\mathbf{x}^0$ dieselben Eigenschaften wie die $\mathbf{x}^i$. Darüberhinaus gilt auch, dass

$$A\mathbf{x}^0 \geq \lambda_0 \mathbf{x}^0. \tag{8.3.21}$$

Angenommen die Unleichung (8.3.21) wäre streng. Dann gilt auch

$$A^2 \mathbf{x}^0_i = \sum_j A_{j\ell} \sum_\ell A_{\ell m} \mathbf{x}^0_m. \tag{8.3.22}$$

Nun ist aber $\mathbf{y}_\ell \equiv \sum_m A_{\ell m} \mathbf{x}^0_m > 0$, für alle ℓ, und somit $\mathbf{y}$ ein Vektor mit strikt positiven Einträgen für den $A\mathbf{y} > \lambda_0 \mathbf{y}$ gilt. Durch Normierung folgt dann, dass es einen Vektor mit den in der Definition von Λ geforderten Eigenschaften gibt, für den diese Ungleichung gilt. Das ist ein Wiederspruch zur Definiton von λ_0. Damit ist aber λ_0 Eigenwert und $\mathbf{x}^0$ der zugehörige Eigenvektor. Offenbar muss dieser Eigenvektor strikt positive Einträge haben.

Sei nun $\lambda \neq \lambda_0$ ein Eigenwert von A mit Eigenvektor $\mathbf{z}$. Dann gilt

$$|\lambda| |\mathbf{z}_i| = |\lambda| \left| \sum_j A_{ij} \mathbf{z}_j \right| \leq \sum_j A_{ij} |\mathbf{z}_j| \leq \lambda_0 |z_i|, \tag{8.3.23}$$

woraus folgt, dass $|\lambda| \leq \lambda_0$. Um zu zeigen, dass $|\lambda| < \lambda_0$, bemerken wir, dass wir stets ein $\delta > 0$ finden können, so dass $A_\delta \equiv A - \delta \mathbb{1}$ noch immer strikt positive Einträge hat. Der grösste Eigenwert von A_δ ist aber $\lambda_0 - \delta$. Nun folgt $|\lambda - \delta| \leq \lambda_0 - \delta$. Damit folgt aber, dass $|\lambda| \leq |\lambda - \delta| + \delta \leq \lambda_0$. Damit

kann aber $|\lambda| = \lambda_0$ nur dann gelten, wenn δ reell und positive ist, und dann ist $\lambda = \lambda_0$.

Schliesslich zeigen wir, dass der Eigenraum von λ_0 eindimensional ist. Seine $\mathbf{x}, \mathbf{y}$ zwei nicht-kolineare Eigenvektoren zu diesem Eigenwert. Dann gilt dasselbe auch für $\mathbf{z} = \mathbf{x} - c\mathbf{y}$. Dabei kann man aber stets c so wählen, dass die Einträge dieses Vektors unterschiedliches Vorzeichen haben. Aber dann gilt

$$\lambda_0 |\mathbf{z}_i| = \left| \sum_j A_{ij} z_j \right| > \sum_j A_{ij} |z_j|, \tag{8.3.24}$$

for jedes j, und daraus folgt ein Widerspruch zur Maximalität von λ_0. □

Wir kommen nun zum Beweis des zweiten Satzes von Perron und Frobenius, Satz 8.13.

Beweis. Es sei $\mathbf{E}$ die Matrix mit Einträgen $\mathbf{E}_{ij} \equiv 1$. Das A nicht-negative Einträge hat, hat für jedes $\delta > 0$ die Matrix $A + \delta\mathbf{E}$ strikt positive Einträge. Sei $\mathbf{x} \neq 0$ ein Vektor mit nicht-negativen Einträgen und $\sum_i \mathbf{x}_i = 1$. Sei nun $\delta_2 > \delta_1 > 0$. Wenn $(A + \delta_1\mathbf{E})\mathbf{x} \geq \lambda\mathbf{x}$, so haben wir

$$(A + \delta_2\mathbf{E})\mathbf{x} \geq (\lambda + \delta_2 - \delta_1)\mathbf{x}. \tag{8.3.25}$$

Setzen wir als $\lambda_0(\delta)$ gleich dem größten Eigenwert von $(A + \delta\mathbf{E})$, so zeigt dies, dass $\lambda_0(\delta)$ in δ monoton wächst. Nun wissen wir aber wegen dem ersten Perron-Frobenius Satz, dass es für jedes $\delta > 0$ genau einen normierten Vektor $\mathbf{x}(\delta)$ mit strikt positiven Einträgen gibt der Eigenvektor von $(A + \delta\mathbf{E})$ zum Eingenwert $\lambda_0(\delta)$ ist. Wegen Kompaktheit gibt es wieder eine Folge $\delta_k \downarrow 0$ so dass $\mathbf{x}(\delta_k) \to \mathbf{x}(0)$ konvergiert. Ausserdem konvergiert $\lambda_0(\delta_j)$ wegen Monotonie gegen eine Zahl $\lambda' \geq \lambda_0$. Andererseits ist leicht zu sehen, dass

$$A\mathbf{x}(0) = \lambda'\mathbf{x}(0). \tag{8.3.26}$$

Damit muss aber $\lambda' \leq \lambda_0$ gelten. Damit ist $\lambda' = \lambda(0)$ und wir sehen, dass λ_0 Eigenwert mit Eigenvektor $\mathbf{x}_0$ der nur nicht-negative Einträge hat ist. Damit ist Teil (i) bewiesen. Der Beweis von (ii) folgt wieder, weil aus der Existenz eines Eigenwertes mit grösserem Betrag als λ_0 ein Widerspruch zur Definition von λ_0 folgt.

Den Beweis der Eigenschaft (iii) geben wir nicht, da diese nicht direkt genutzt wird. □

8.3.3 Wesentliche und unwesentliche Klassen.

Besitzt eine Markovkette mehrere Klassen, so kann man diese in zwei Gruppen einteilen: solche, aus denen man in eine andere Klasse austreten kann

(aber nicht wieder zurück kann), und solche aus denen man nicht in eine andere Klasse eintreten kann (in die man aber ggf. aus anderen eintreten kann). Erstere heissen "unwesentlich", letztere "wesentlich".

Anmerkung. Im Fall endlichen Zustandsraums können wir wesentliche Klassen auch als rekurrent, unwesentliche als transient bezeichnen. Im Fall von Markovketten mit unendlichem Zustandsraum sind diese Begriffe aber zu unterscheiden.

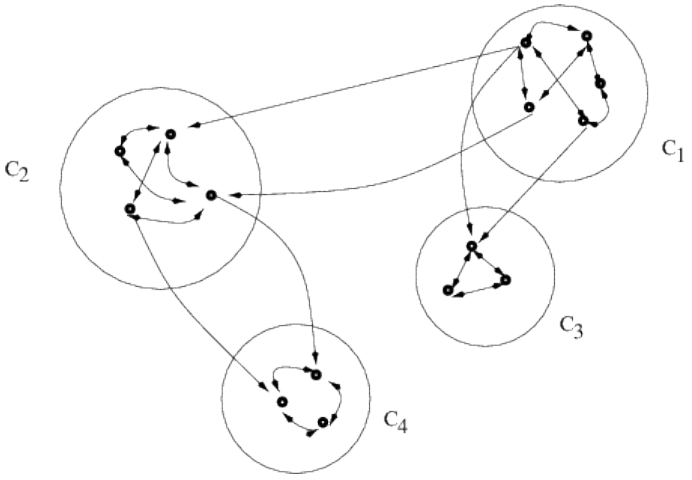


Abb. 8.4 Der Graph einer Markovkette mit vier Klassen C_1, C_2, C_3, C_4. Die Klassen C_1 und C_2 sind transient, C_3 und C_4 sind rekurrent.

Satz 8.21. *Sei X eine Markovkette mit Zustandsraum S. S zerfalle in die wesentlichen Klassen $C_1, \ldots, C_\ell$ und die unwesentlichen Klassen $D_1, \ldots, D_k$. Dann gibt es ℓ invariante Verteilungen $\mu_1, \ldots, \mu_\ell$ mit Träger auf den wesentlichen Klassen $C_1, \ldots, C_\ell$, und alle invarianten Verteilungen μ sind von der Form*

$$\mu = \sum_{i=1}^{\ell} \alpha_i \mu_i,$$

mit $\alpha_i \geq 0$ und $\sum_i \alpha_i = 1$.

Beweis. Es ist klar, dass es für jede wesentliche aperiodische Klasse genau eine invariante Verteilung gibt. Sei nämlich C eine wesentliche Klasse. Wenn die Anfangsverteilung π_0 so gewählt ist, dass für alle $i \notin C$, $\pi_0(i) = 0$, dann ist für alle Zeiten für solche i, $\pi_t(i) = 0$. Die Matrix P eingeschränkt auf den von den Zuständen $j \in C$ aufgespannten Unterraum ist aber die Übergangsmatrix einer irreduziblen aperiodischen Markovkette mit Zustandsraum C. Also gibt

es eine invariante Verteilung μ_C die C Maß eins gibt. Dies gilt für jede wesentliche Klasse separat.

Ebenso kann man sich leicht überzeugen, dass für jede invariante Verteilung μ und jede unwesentliche Klasse D gilt, dass $\mu(D) = \sum_{j\in D} \mu(j) = 0$. Sei nämlich $\mu(D) > 0$. Wir betrachten dazu zuächst solche unwesentliche Klassen, in die man aus keiner anderen Klasse eintreten kann (wegen der Endlichkeit des Zustandsraumes muss es mindestens eine solche geben). Sei D eine solche Klasse. Da μ invariant ist, muss $(\mu P)(D) = \mu(D)$ gelten. Nun ist aber

$$(\mu P)(D) = \sum_{j\in D}\sum_{i\in S} \mu(i)p_{i,j} = \sum_{j\in D}\sum_{i\in D} \mu(i)p_{i,j} + 0 \tag{8.3.27}$$

da ja für alle $j \in D$ und $i \notin D$, $p_{i,j} = 0$, gemäß unserer Annahme. Daher ist

$$(\mu P)(D) = \sum_{i\in D} \mu(i) \sum_{j\in D} p_{i,j} = \sum_{i\in D} \mu(i) - \sum_{i\in D} \mu(i) \sum_{j\notin D} p_{i,j} \leq \mu(D). \tag{8.3.28}$$

Dabei kann Gleichheit nur dann gelten, wenn für alle $i \in D$ für die es $j \in D^c$ gibt mit $p_{i,j} > 0$, $\mu(i) = 0$. Andererseits gilt für diese j dann

$$0 = \mu(i) = \sum_{j\in D} \mu(j)p_{j,i},$$

weswegen $\mu(j) = 0$ auch für alle Zustände in D gilt die mit i verbunden sind; indem wir dieses Argument iterieren, und benutzen, dass D eine kommunizierende Klasse ist, folgt $\mu(j) = 0$ für alle $j \in D$.

Nachdem wir wissen, dass $\mu(D) = 0$ für alle unwesentlichen Klassen, in die man nicht eintritt, kann man nun diese D aus dem Zustandsraum aussondern, und die Restriktion der Markovkette auf den verbleibenden Zustandsraum $S\backslash D$ betrachten. Wenn dieser noch unwesentliche Klassen enthält, so gibt es mindestens eine, in die man nicht mehr eintreten kann, und man sieht, dass auf diesen die invariante Verteilung auch Null ist. Durch Iteration folgt, dass μ auf allen unwesentlichen Klassen verschwindet.

Nutzt man nun diese Information, so verbleibben als Gleichungssystem für die invarianten Verteilungen nur noch entkpoppelte Systeme für jede der verbleibenden wesentlichen irreduziblem Klassen. Daraus folgt die behauptete Struktur der invarianten Maße sofort. □

Beispiele. Wir schauen uns die Klassenzerlegung und invarianten Verteilungen für unsere drei Beispiele von vorher an.

- **Unabhängige Zufallsvariablen**. Hier ist die Markovkette irreduzibel und aperiodisch. Darüber hinaus ist die Übergangsmatrix bereits ein Projektor auf die einzige invariante Verteilung π_0.
- **Irrfahrt mit Rand**. Hier gibt es offenbar drei Klassen: $C_1 \equiv \{-L+1, \ldots, L-1\}$, $C_2 = \{-L\}$ und $C_3 = \{L\}$. Dabei ist C_1 unwesentlich und C_2 und C_3 sind

wesentlich. Daher haben wir zwei invariante Verteilungen, μ_2 und μ_3, wobei

$$\mu_2(j) = \delta_{j,-L}, \qquad \mu_3(j) = \delta_{j,L}.$$

Natürlich sind auch alle konvexen Linearkombinationen dieser zwei Verteilungen invariante Verteilungen. Da für jede invariante Verteilung $\mu(C_1) = 0$ gilt, erschöpfen diese offenbar die invarianten Verteilungen dieser Markovkette.

- **Wettermodell**. Seien zunächst $p_{0,1}, p_{1,0} \in (0,1)$. Dann ist die Markovkette wieder irreduzibel und aperiodisch, und die einzige invariante Verteilung ist

$$\mu = \frac{1}{(p_{0,1} + p_{1,0})}(p_{1,0}, p_{0,1}).$$

Dasselbe gilt wenn einer der beiden Parameter gleich eins ist, der andere aber in $(0,1)$ liegt.
Wenn $p_{1,0}$ und $p_{0,1}$ gleich null sind, so gibt es zwei wesentliche Klassen mit den jeweils trivialen Verteilungen. Falls nur eine der beiden null ist, so gibt es eine wesentliche und eine unwesentliche Klasse.
Wenn $p_{0,1} = p_{1,0} = 1$ ist, haben wir eine irreduzible, aber nicht aperiodische Klasse. Die Markovkette hat dann Periode zwei, wie schon oben beschrieben.

8.4 Stoppzeiten und der starke Ergodensatz

In der Folge werden wir mit Erwartungen von Funktionen von Markovprozessen beschäftigt sein. Wir schreiben dazu für messbare Funktionen F auf $(S^{\mathbb{N}_0}, \mathfrak{B}(S)^{\otimes \mathbb{N}_0})$ und für $x \in S$,

$$\mathbb{E}_x F = \mathbb{E}[F(X_0, X_1, \dots, X_n, \dots) | X_0 = x].$$

Es ist in der Folge oft bequem, die Wahrscheinlichkeitsräume $(\Omega, \mathfrak{F}, \mathbb{P})$ und $(S_0^{\mathbb{N}}, \mathfrak{B}(S)^{\otimes \mathbb{N}_0}, P_X)$ zu identifizieren. Wir definieren die Zeittranslation θ_n durch

$$F \circ \theta_T(X_0, X_1, \dots, X_n, \dots) \equiv F(X_T, X_{T+1}, \dots, X_{T+n}, \dots).$$

8.4.1 Die starke Markoveigenschaft

Ein wesentliches Konzept in der Analyse von Markovprozessen ist das der *Stoppzeit*. Wir bezeichnen mit $\mathfrak{F}_n \equiv \sigma(X_0, \ldots, X_n)$ die σ-Algebra, die von den Zufallsvariablen $X_0, X_1, \ldots X_n$ erzeugt wird. Die Familie $\{\mathfrak{F}_n\}_{n\in\mathbb{N}_0}$ bezeichnet man auch als eine Filtrierung, bzw. die dem Markovprozess $\{X_n\}_{n\in\mathbb{N}_0}$ zugehörige Filtrierung der σ-Algebra $\mathfrak{F}$.

Definition 8.22. Eine Abbildung $T : \Omega \to \mathbb{N}_0$ heißt eine *Stoppzeit* genau dann, wenn für jedes $n \in \mathbb{N}_0$, das Ereignis $\{T = n\}$ in $\mathfrak{F}_n$ liegt.

Stoppzeiten sind also dadurch charakterisiert, dass man zu jedem Zeitpunkt, n, aus der Kenntnis des Verlaufs der Vergangenheit des Prozesses X entscheiden kann, ob diese Stoppzeit gerade erreicht ist.

Ein wichtiges Beispiel für Stoppzeiten sind die *ersten Eintrittszeiten* in Untermengen. Ist $D \subset S$, so definieren wir

$$\tau_D \equiv \inf\{n > 0 | X_n \in D\}. \tag{8.4.1}$$

Wir sehen, dass τ_D eine Stoppzeit ist:

$$\{\tau_D = n\} = \{\forall_{k<n}, X_k \notin D\} \cap \{X_n \in D\}.$$

Die rechte Seite ist manifest in $\mathfrak{F}_n$, weil sie nur von X_k mit $k \leq n$ abhängt.

Beispiel für eine interessante Größe, die keine Stoppzeit ist, ist die *letzten Austrittszeiten* aus Untermengen,

$$\sigma_D \equiv \sup\{n \geq 0 | X_n \in D\}.$$

Klarerweise können wir zu keinem Zeitpunkt wissen, ob der Prozess nicht nochmal nach D zurückkehrt, ohne in die Zukunft zu blicken.

Eine der wichtigsten Eigenschaften von Stoppzeiten ist die sogenannte *starke Markoveigenschaft*. Sie besagt, dass man die Erwartung bezüglich Verteilungen einer Markovkette an Stoppzeiten faktorisieren kann. Damit meinen wir das folgende.

Wir definieren zunächst für eine Stoppzeit T die σ-Algebra $\mathfrak{F}_T$ als die Menge aller Ereignisse, die nur von X_n mit $n \leq T$ abhängen. Formal ist

$$\mathfrak{F}_T \equiv \bigcup_{n\in\mathbb{N}_0} \mathfrak{F}_n \cap \{n \leq T\}.$$

Satz 8.23 (Starke Markoveigenschaft). *Sei T eine Stoppzeit und seien F und G $\mathfrak{F}$-messbare Funktionen auf Ω. Sei darüber hinaus F messbar bezüglich $\mathfrak{F}_T$. Dann gilt für jedes $x \in S$, dass*

$$\mathbb{E}_x\left[\mathbb{1}_{T<\infty} F\, G \circ \theta_T\right] = \mathbb{E}_x\left[\mathbb{1}_{T<\infty} F\, \mathbb{E}_{X_T}[G]\right], \tag{8.4.2}$$

d.h.,

$$\mathbb{E}\left[\mathbb{1}_{T(X)<\infty}F(X)\,(G\circ\theta_T)(X)|X_0=x\right] = \mathbb{E}\left[\mathbb{1}_{T(X)<\infty}F(X)\,\mathbb{E}[G(X')|X'_0=X_T]|X_0=x\right] \quad (8.4.3)$$

wo X' eine unabhängige Kopie von X ist.

Beweis. Man kann sich durch explizites Ausschreiben davon überzeugen, dass für jedes endliche n

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}_{T(X)=n}F(X)\,(G\circ\theta_T)(X)|X_0=x\right]\\
&=\sum_{y\in S}\mathbb{E}\left[\mathbb{1}_{T(X)=n}\mathbb{1}_{X_n=y}F(X)\,(G\circ\theta_n)(X)|X_0=x\right]\\
&=\sum_{y\in S}\mathbb{E}\left[\mathbb{1}_{T(X)=n}\mathbb{1}_{X_n=y}F(X)\,|X_0=x\right]\mathbb{E}[G(X')|X'_0=y]\\
&=\mathbb{E}\left[\mathbb{1}_{T(X)=n}F(X)\mathbb{E}[G(X')|X'_0=X_n]|X_0=x\right].
\end{aligned}$$

Nun summieren wir einfach über n und erhalten die Behauptung. □

Eine Anwendung der starken Markoveigenschaft liefert eine neue Interpretation der invarianten Verteilung.

Lemma 8.24. *Sei X eine irreduzible Markovkette mit endlichem Zustandsraum S. Sei μ die invariante Verteilung. Dann gilt, für $j,\ell\in S$,*

$$\mu(j)=\frac{\mathbb{E}_\ell\left[\sum_{t=1}^{\tau_\ell}\mathbb{1}_{X_t=j}\right]}{\mathbb{E}_\ell\tau_\ell} \quad (8.4.4)$$

wobei $\tau_\ell\equiv\inf\{n>0|X_n=\ell\}$.

Beweis. Wir zeigen zunächst, dass $\mathbb{E}_\ell[\tau_\ell]<\infty$, und somit der Ausdruck auf der rechten Seite von (8.6.8) Sinn macht.

Betrachten wir zunächst den Fall, dass unsere Markovkette apreiodisch ist. Dann wissen wir, dass es $k\in\mathbb{N}$ gibt, so dass für alle $i,j\in S$ $(P^k)_{i,j}\geq c>0$ ist. Dann ist aber

$$\mathbb{P}_\ell[\tau_\ell>t]\leq\mathbb{P}_\ell[X_{kn}\neq\ell,\forall kn\leq t]\leq\prod_{n:kn\leq t}(1-\min_{i\in S}(P^k)_{i,\ell})\lesssim(1-c)^{t/k}. \quad (8.4.5)$$

Damit ist dann natürlich $\mathbb{E}_\ell[\tau_\ell]=\sum_{t\geq0}\mathbb{P}_\ell[\tau_\ell>t]<\infty$.

Falls die Kette nur irreduzibel und nicht notwending aperiodisch ist, so muss das das obige Argument leicht verändert werden. Es gilt nun immer noch, dass es für jedes $j\in S$ ein $k_j<\infty$ gibt, so dass $P^{k_j}_{j\ell}>0$. Daher gibt es $k^*\equiv\max_{j\in S}k_j$ mit der Eigenschaft, dass für alle $j\in S$, $\exists_{k_j\leq k}$ so dass $P^{k_j}_{j\ell}>0$ Damit aber ist

$$\min_{j\in S}\mathbb{P}_j\left(X_t\neq\ell,\forall_{t\leq k}\right)\geq\min_{j\in S}\mathbb{P}_j\left(X_{k_j}\neq\ell\right)>0.$$

Indem wir diese Abschätzung in (8.4.5) verwenden erhalten wir dieselbe Schlussfolgerung.

Wir definieren $\nu_\ell(j) = \mathbb{E}_\ell\left[\sum_{t=1}^{\tau_\ell} \mathbb{1}_{X_t=j}\right]$. Wenn wir zeigen, dass $\nu_\ell(j)$ die Invarianzeigenschaft erfüllt, so tut dies auch μ, und nach Konstruktion ist μ eine Wahrscheinlichkeitsverteilung. Wir schreiben zunächst $1 = \sum_{m\in S} \mathbb{1}_{X_{t-1}=m}$, und

$$\begin{aligned}\nu_\ell(j) = \mathbb{E}_\ell\left[\sum_{t=1}^{\infty} \mathbb{1}_{X_t=j}\mathbb{1}_{t\le\tau_\ell}\right] &= \sum_{t=1}^{\infty} \mathbb{P}_\ell\left(X_t = j, t \le \tau_\ell\right) \\ &= \sum_{m\in S}\sum_{t=1}^{\infty} \mathbb{P}_\ell\left(X_{t-1} = m, X_t = j, t \le \tau_\ell\right).\end{aligned}$$

Nun ist das Ereignis $\{t \le \tau_\ell\} = \{\tau_\ell \le t-1\}^c \in \mathfrak{F}_{t-1}$. Daher können wir die Markov-Eigenschaft zur Zeit $t-1$ anwenden und erhalten

$$\begin{aligned}\mathbb{P}_\ell\left(X_{t-1} = m, X_t = j, t \le \tau_\ell\right) &= \mathbb{P}_\ell\left(X_{t-1} = m, t \le \tau_\ell\right)\mathbb{P}_m\left(X_1 = j\right) \\ &= \mathbb{P}_\ell\left(X_{t-1} = m, t \le \tau_\ell\right)p_{m,j}. \qquad (8.4.6)\end{aligned}$$

Damit ist aber

$$\nu_\ell(j) = \sum_{m\in S}\mathbb{E}_\ell\left[\sum_{t=1}^{\infty}\mathbb{1}_{X_{t-1}=m}\mathbb{1}_{t\le\tau_\ell}\right]p_{m,j} = \sum_{m\in S}\mathbb{E}_\ell\left[\sum_{t=1}^{\tau_\ell}\mathbb{1}_{X_{t-1}=m}\right]p_{m,j}.$$

Andererseits

$$\sum_{t=1}^{\tau_\ell}\mathbb{1}_{X_{t-1}=m} = \mathbb{1}_{X_0=m} + \sum_{t=1}^{\tau_\ell}\mathbb{1}_{X_t=m} - \mathbb{1}_{X_{\tau_\ell}=m} = \sum_{t=1}^{\tau_\ell}\mathbb{1}_{X_t=m}$$

weil $X_0 = X_{\tau_\ell}$. Somit ist aber

$$\nu_\ell(j) = \sum_{m\in S}\mathbb{E}_\ell\left[\sum_{t=1}^{\tau_\ell}\mathbb{1}_{X_t=m}\right]p_{m,j} = \sum_{m\in S}\nu_\ell(m)p_{m,j}.$$

Dies ist aber gerade die Gleichung für die invariante Verteilung. Daher ist $\nu_\ell(j)/\sum_{i\in S}\nu_\ell(i)$ eine invariante Wahrscheinlichkeitsverteilung, und wegen deren Eindeutigkeit ist $\nu_\ell = \mu$. Bemerke, dass ν_ℓ unabhängig von ℓ ist! Nun ist aber

$$\sum_{i\in S}\nu_\ell(i) = \sum_{i\in S}\mathbb{E}_\ell\left[\sum_{t=1}^{\tau_\ell}\mathbb{1}_{X_t=i}\right] = \mathbb{E}_\ell\left[\sum_{t=1}^{\tau_\ell}\mathbb{1}_{X_t\in S}\right] = \mathbb{E}_\ell[\tau_\ell]$$

woraus die Behauptung des Lemmas folgt. □

Korollar 8.25. *Für eine irreduzible Markovkette mit endlichem Zustandsraum gilt*

$$\mu(j) = \frac{1}{\mathbb{E}_j \tau_j}. \tag{8.4.7}$$

Beweis. Formel (8.6.8) gilt für jede Wahl von ℓ. Indem wir $\ell = j$ wählen und benutzen, dass

$$\nu_j(j) = \mathbb{E}_j[\sum_{t=1}^{\tau_j} \mathbb{1}_{X_t=j}] = 1$$

ist, weil aus der Definition von τ_j folgt $\mathbb{1}_{X_t=j} = \delta_{\tau_j,t}$ für $t = 1, \ldots, \tau_j$, erhlten wir (**??**). □

8.4.2 Der starke Ergodensatz

Wir sind nun in der Lage eine starke Form des Ergodensatzes für irreduzible Markovketten zu formulieren, die in gewisser Weise das Analogon des Gesetzes der grossen Zahlen für Markovketten ist.

Satz 8.26 (Starker Ergodensatz). *Sei X eine irreduzible Markovkette mit endlichem Zustandsraum S und invarianter Verteilung μ. Sei $f : S \to \mathbb{R}$ eine beschränkte messbare Funktion. Dann gilt*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(X_k) = \int_S f d\mu \quad \text{f.s.} \tag{8.4.8}$$

Anmerkung. Die Voraussetzungen an f sind angesichts der Endlichkeit des Zustandsraums natürlich trivial.

Beweis. Es genügt offenbar den Satz für Indikatorfunktionen $f = \mathbb{1}_i$, $i \in S$, zu beweisen. Sei nun t_ℓ eine Folge von Stoppzeiten definiert durch

$$\begin{aligned} t_0 &\equiv \inf\{k \geq 0 : X_k = i\}, \\ t_\ell &\equiv \inf\{k > t_{\ell-1} : X_k = i\}. \end{aligned} \tag{8.4.9}$$

Mit anderen Worten, die Zeiten t_ℓ sind genau die Zeiten, an denen X den Zustand i besucht. Offenbar ist dann

$$\sum_{k=1}^{n} f(X_k) = \sum_{k=1}^{n} \mathbb{1}_{X_k=i} = \max\{\ell : t_\ell \leq n\}. \tag{8.4.10}$$

Nun machen wir folgende wichtige Beobachtung: Setze $\sigma_\ell = t_\ell - t_{\ell-1}$. Dann sind für $\ell \geq 1$ die σ_ℓ unabhängige, identisch verteilte Zufallsvariablen. Das folgt aus der starken Markoveigenschaft, indem wir nachweisen, dass für beliebige integrierbare Funktionen, $g, h : \mathbb{N} \to \mathbb{R}$,

$$\mathbb{E}_\ell[g(\sigma_i)h(\sigma_j)] = \mathbb{E}_\ell[g(\sigma_i)]\,\mathbb{E}_\ell[h(\sigma_j)] \quad \text{für alle } i \neq j.$$

(Übung!). Es gilt $\mathbb{P}[\sigma_\ell \leq k] = \mathbb{P}[t_1 \leq k | X_0 = i] = \mathbb{P}_i[\tau_i \leq k]$. Wir wissen schon, dass $\mathbb{E}[\sigma_\ell] = \mathbb{E}_i[\tau_i] < \infty$. Daher gilt nach dem Gesetz der grossen Zahlen,

$$\lim_{n\to\infty} \frac{t_n}{n} = \mathbb{E}\left[t_1 | X_0 = i\right] = \mathbb{E}_i[\tau_i] \quad \text{f.s.} \tag{8.4.11}$$

Ausserdem ist für jedes ℓ,

$$\lim_{n\to\infty} \frac{\sigma_\ell}{n} = 0 \quad \text{f.s.}$$

Dann ist leicht einzusehen (Übung!), dass daraus folgt, dass

$$\lim_{n\to\infty} \frac{1}{n} \max\left\{\ell : t_\ell \leq n\right\} = \frac{1}{\mathbb{E}_i[\tau_i]} = \mu(i) \quad \text{f.s.} \tag{8.4.12}$$

□

Anmerkung. Wir sehen, dass wir für den starken Ergodensatz die Aperiodizität nicht voraussetzen müssen. Es folgt daraus auch, dass für irreduzible Markovketten gilt, dass

$$\lim_{n\uparrow\infty} \frac{1}{n} \sum_{k=1}^{n} \pi_0 P^k = \mu, \tag{8.4.13}$$

das heisst, die Verteilung einer irreduziblem Markovkette konvergiert im Cesaro-Mittel stets gegen die invariante Verteilung konvergiert.

8.4.3 Markovketten Monte-Carlo Verfahren.

Eine in der Praxis wesentliche Anwendung des Ergodensatzes für Markovketten ist die Möglichkeit, mit seiner Hilfe Integrale bezüglich einer gewünschten Verteilung numerisch approximativ zu berechnen.

Bei der Berechnung von Erwartungswerten trifft man in der Praxis of auf zwei Probleme: (1) Der Zustandsrausm ist sehr gross (und hochdimensional) (etwa etwa in der statistischen Mechanik, Maße nur "bis auf die Normierung" explizit gegeben, eta in der Form

$$\rho(x) = \frac{1}{Z} \exp\left(-\beta H(x)\right),$$

wo $H(x)$ eine einfach zu berechende Funktion ist, die Konstante Z aber nur als $\sum_{x\in S} \exp\left(-\beta H(x)\right)$ gegeben ist, also etwa so schwer zu berechnen ist wie das Integral selbst.

Hier kommen nun die Markovketten und der Ergodensatz ins Spiel. Angenommen, wir fänden eine ergodische Markovkette mit Zustandraum S derart, das die invariante Verteilung der Kette gerade ρ ist. Da die Normierung für die Invarianzgleichung keine Rolle spielt, kann man eine solche konstruieren, ohne Z zu kennen. Dann wissen wir, dass

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(X_k) \equiv \int_S f d\rho \quad \text{f.s.}$$

Um eine systematiche Approximation unseres Integrals zu bekommen, benötigen wir also nur eine Realisierungen der Zufallsvariablen $X_1, X_2, \ldots$. Dabei gewinnen wir natürlich nur dann etwas, wenn die entsprechenden bedingten Verteilungen, also die Übergangwahrscheinlichkeiten der Markovkette, finden können. Dazu muss man natürlich in der Lage sein, diese Zufallsvariablen in einfacher Weise numerisch zu konstruieren. Dazu ist es nützlich, die Markovkette so zu konstruieren, dass man von einem gegebenen Zustand aus nur sehr wenige Zustände erreichen kann; im obigen Beispiel $S = \{-1, 1\}^N$ wählt man die Markovkette etwa so, dass man in einem Schritt nur eine der Koordinaten des Vektors x ändern kann. Dann sind die Übergangswahrscheinlichkeiten effektiv Verteilungen auf nur N (statt 2^N) Zuständen, und somit viel leichter handhabbar. Im obigen Fall kann man z.B. die Übergangswahrscheinlichkeiten in der Form

$$p_{xy} = \frac{1}{N} \exp\left(-[H_N(y) - H_N(x)]_+\right), \text{wenn } |x - y| = 2,$$

$$p_{xx} = 1 - \sum_{y:|x-y|=2} p_{xy},$$

und null sonst, wählen (Übung!).

Damit dieses Verfahren funktioniert, sollte natürlich die Konvergenz gegen die invariante Verteilung schnell genug erfolgen, so dass man tatsächlich rasch gute Approximationen erhält. Dies zu quantifizieren ist im Allgemeinen ein schwieriges Problem. In vielen Fällen liefert dieses *Markovketten Monte-Carlo Verfahren* aber sehr gute Resultate. Monte-Carlo Verfahren sind ein wichtiges Hilfsmittel der stochastischen Numerik und werden in verschiedener Form sehr verbreitet eingesetzt.

8.5 Vorwärtsgleichungen, Eintrittswahrscheinlichkeiten und Zeiten.

Ein typisches Vorgehen zur Berechnung verschiedener Wahrscheinlichkeiten in Markovketten besteht in der Herleitung von linearen Gleichungen für diese. Als Beispiel betrachten wir eine Markovkette mit Zustandsraum S, die zwei

wesentliche Klassen C_1, C_2 sowie eine unwesentliche Klasse $C_3 = S\backslash(C_1 \cup C_2)$ habe. Wir interessieren uns dafür, mit welcher Wahrscheinlichkeit man, ausgehend von einem Zustand $x \in C_3$ in der wesentlichen Klasse C_1 endet. Diese können wir schreiben als

$$\mathbb{P}_x \left[\tau_{C_1} < \tau_{C_2}\right].$$

Um eine Gleichung für diese Wahrscheinlichkeit zu erhalten, betrachten wir zunächst alle möglichen ersten Schritte der Kette und wenden dann die Markoveigenschaft an. Wenn der erste Schritt bereits nach C_1 führt, so ist das Ereignis bereits realisiert und wir erhalten einen Beitrag 1; führt der erste Schritt nach C_2, so kann das Ereignis nicht eintreten, und wir erhalten einen Beitrag 0; wenn schliesslich der erste Schritt nach y in C_3 bleibt, ist der Beitrag gerade die Wahrscheinlichkeit, das Ereignis ausgehend von y zu realisieren. Dies liefert

$$\mathbb{P}_x \left[\tau_{C_1} < \tau_{C_2}\right] = \sum_{y \in C_1} p(x,y) + \sum_{y \in S\backslash(C_1 \cup C_2)} p(x,y) \mathbb{P}_y \left[\tau_{C_1} < \tau_{C_2}\right]. \quad (8.5.1)$$

Wir können diese Gleichung in einer geschlossenen Form schreiben, wenn wir die Funktion $h_{C_1,C_2}(x)$ definieren als

$$h_{C_1,C_2}(x) \equiv \begin{cases} \mathbb{P}_x \left[\tau_{C_1} < \tau_{C_2}\right], & \text{wenn } x \in S\backslash(C_1 \cup C_2), \\ 0, & \text{wenn } x \in C_2, \\ 1, & \text{wenn } x \in C_1. \end{cases}$$

Damit wird (8.5.1) in der Form

$$h_{C_1,C_2}(x) = \sum_{y \in S} p(x,y) h_{C_1,C_2}(y) = (P h_{C_1,C_2})(x) \quad (8.5.2)$$

schreibbar. Eine solche Gleichung nennt man auch *Vorwärtsgleichung*. Eine Funktion, die in einem Gebiet die Gleichung $f = Pf$ löst, wo P Übergangsmatrix einer Markovkette ist, nennt man auch eine *harmonische Funktion*. Die Funktion h_{C_1,C_2} heisst speziell auch *Gleichgewichtspotential*. Man kan diese als Lösung des Gleichungssystems

$$\begin{aligned} &h_{C_1,C_2}(x) = (P h_{C_1,C_2})(x), \quad x \in S\backslash(C_1 \cup C_2), \\ &h_{C_1,C_2}(x) = 1, \quad x \in C_1, \\ &h_{C_1,C_2}(x) = 0, \quad x \in C_2, \end{aligned} \quad (8.5.3)$$

erhalten. Gleichungen wie (8.5.3) bilden die Grundlage für eine sehr weitgehende und tiefe Beziehung zwischen der Theorie der Markovprozesse und der *Potentialtheorie*, mithin zwischen Stochastik und Analysis. Wir werden diese Thematik in fortgeschrittenen Vorlesungen zur W-Theorie wieder aufgreifen.

Hier wollen wir einige grundlegenden Ergebnisse im Fall endlichen Zustandsraumes betrachten. Die erste Frage die wir uns stellen müssen, ist, ob Gleichungen des Typs (8.5.3) eindeutige Lösungen haben.

Definition 8.27. Sei P eine Übergangsmatrix einer Markovkette mit Zustandsraum S und sei $D \subset S$. Eine Funktion $f : S \to \mathbb{R}$ heisst *harmonisch* (bez. P) auf D, falls für alle $x \in D$, $f(x) = Pf(x)$.

Die Eindeutigkeit der Lösung folgt dann aus folgenden Satz (wobei man $D^c = C_1 \cup C_2$ und f die Differenz von zwei Lösungen von (8.5.3) einsetzt).

Satz 8.28. *Sei P die Übergangsmatrix einer Markovkette mit endlichem Zustandsraum S. Sei $D \subset S$ so dass von jedem $x \in D$ die Menge $D^c \equiv S \backslash D$ längs des Graphen der Markovkette erreicht werden kann. Dann hat das Gleichungssystem*

$$\begin{aligned} Pf(x) &= f, \ \textit{wenn } x \in D, \\ f(x) &= 0, \ \textit{wenn } x \in D^c, \end{aligned} \tag{8.5.4}$$

die eindeutige Lösung $f(x) \equiv 0$.

Beweis. Der Beweis dieses Satzes beruht auf dem sogenannten *Maximumsprinzip* für harmonische Funktionen.

Lemma 8.29. *Seien P und D wie im vorigen Satz und sei h eine harmonische Funktion of D. Dann nimmt h ihr Maximum auf D^c an.*

Beweis. Sei $x \in D$ ein Maximum von h. Dann gilt

$$h(x) = \sum_y p_{xy} h(y). \tag{8.5.5}$$

Da $h(x) \geq h(y)$ für alle y in der Summe für die $p_{xy} > 0$, folgt dass $h(x) = h(y)$ für alle diese Punkte. Indem wir dieses argument iterieren, finden wir, dass es einen Weg längs Kanten des Graphen der Kette von x nach D^c gibt, längs dem h konstant den Wert $h(x)$ annimmt. □

Für unseren Fall ist f harmonisch auf D und $f = 0$ auf D^c. Daher ist $f(x) \leq 0$. Indem wir dasselbe Argument auf $-f$ anwenden, folgt auch, dass $f(x) \geq 0$. □

Übung. Sei eine Markovkette wie oben mit zwei wesentlichen und einer unwesentlichen Klasse gegeben. Seien die wesentlichen Klassen aperiodisch, und seien μ_1, μ_2 die invarianten Maße mit $\mu_i(C_i) = 1$. Dann gilt, für alle $x \in C_3$, wenn $\pi_0(y) = \delta_x(y)$,

$$\lim_{n\to\infty} \pi_n(z) = \mathbb{P}_x \left[\tau_{C_1} < \tau_{C_2}\right] \mu_1(z) + \mathbb{P}_x \left[\tau_{C_2} < \tau_{C_1}\right] \mu_2(z).$$

Neben den Eintreffwahrscheinlichkeiten in verschiedenen Klassen kann man auch nach der Verteilung der Eintrittszeiten fragen. So sei D eine beliebige Untermenge des Zustandsraums S. Was ist die Verteilung der Stoppzeit τ_D,

$$\mathbb{P}_x[\tau_D = t] \equiv f_D(x,t). \tag{8.5.6}$$

Wir können wieder eine Gleichung für $f_D(x,t)$ herleiten, indem wir uns zunächst den ersten Schritt der Kette ansehen. Falls $t = 1$, sehen wir dass (für $t = 1$ und $x \notin D$)

$$\mathbb{P}_x[\tau_D = 1] = \sum_{y \in D} p(x,y),$$

für $t > 1$ ist

$$\mathbb{P}_x[\tau_D = t] = \sum_{y \notin D} p(x,y)\mathbb{P}_y[\tau_D = t-1].$$

Diese Gleichung kann man in einer schöneren Form schreiben, wenn die Definition der Funktion f_D wie folgt ausweitet:

$$f_D(x,t) \equiv \begin{cases} \mathbb{P}_x[\tau_D = t], & \text{wenn } x \in D^c, t \geq 1, \\ 0, & \text{wenn } x \in D, t \geq 1, \\ 0, & \text{wenn } x \in D^c, t = 0, \\ 1, & \text{wenn } x \in D, t = 0. \end{cases}$$

Dann erhalten wir nämlich für all $t \geq 1$, $x \in D^c$,

$$f_D(x,t) = \sum_{y \in S} p(x,y) f_D(y, t-1).$$

Damit sieht man, dass man die gesuchte Wahrscheinlichkeit durch Lösung eines diskreten Rand-Anfangswertproblems erhalten kann, dass wie folgt aussieht:

$$\begin{aligned} f_D(x,t) - f_D(x,t-1) &= \sum_{y \in S \setminus x} p(x,y) f_D(y,t-1),\ x \in D^c, t \geq 1, \\ f_D(x,t) &= 0, \quad x \in D, t \geq 1, \\ f_D(x,0) &= 0, \quad x \in D^c, \\ f_D(x,0) &= 1, \quad x \in D. \end{aligned} \tag{8.5.7}$$

Mit Hilfe der Matrix $L \equiv P - \mathbb{1}$ können wir die Gleichung (8.5.8) noch in der Form

$$f_D(x,t) - f_D(x,t-1) = (L f_D)(x,t-1)$$

schreiben. Die Lösung dieser linearen Gleichungen sind also geeignet um die Wahrscheinlichkeitsverteilung von τ_D zu berechnen.

Übung. Zeige, dass die Funktion

$$w_D(x) \equiv \begin{cases} \mathbb{E}_x \tau_D, & x \in D^c, \\ 0, & x \in D, \end{cases}$$

die Gleichung

$$w_D(x) = \sum_{y \in S} p(x,y) w_D(y) + 1, \; x \in D^c, \tag{8.5.8}$$
$$w_D(x) = 0, \quad x \in D,$$

löst. Benutze dazu entweder die Gleichung (8.5.8) und die Beobachtung, dass $w_D(x) = \sum_{t=1}^{\infty} f_D(x,t)$, oder leite die Gleichung direkt analog zu der für f_D her.

8.6 Markovketten mit abzählbarem Zustandsraum

Wir wollen abschliessend unsere Betrachtung vom Markovketten noch auf den Fall von unendliche, aber abzählbare Zustandsräume ausdehnen. Ganz natürliche Prozesse, wie die Irrfahrt auf $\mathbb{Z}$ oder $\mathbb{Z}^d$, gehören dazu.

An den Definitionen einer Marlovkette ändert sich zunächst nicht. Ebenso können wir die Begriffe von kommunizierenden Klassen, Irreduzibilität, Periodizität ohne weiteres in diesem Kontext anwenden. Aus der Übergansmatrix wird nun eine unendlichdimensionale Übergansmatrix, P, mit Elementen p_{ij}, $i, j \in S$. Bei der Frage nach der Existenz und Eindeutigkeit einer invarianten Verteilung haben wir aber sehr stark auf der Theorie endlich dimensionaler Matrizen aufgebaut. Hier werden sich nun neue Fragen auftun.

Als erstes führen wir die Begriffe der *Rekurrenz* und *Transienz* ein.

Definition 8.30. Sei X eine irreduzible Markovkette mit abzählbarem Zustandsraum S.

(i) X heisst *transient*, wenn für jedes $i \in S$,

$$\mathbb{P}_i(\tau_i < \infty) < 1; \tag{8.6.1}$$

(ii)Andernfalls heisst X *rekurrent*.
(iii) X heisst *positiv rekurrent* falls für alle $i \in S$,

$$\mathbb{E}_i(\tau_i) < \infty. \tag{8.6.2}$$

Anmerkung. Man kann Transienz und Rekurrenz auch als Eigenschaft einzelner Zustände definieren. Diese Eigenschaften sind aber wieder Klasseneingenschaften, so dass sie für irreduzible Ketten Eigenschaften der Kette werden. Damit ist eine irreduzible Markovkette transient, rekurrent oder positiv

rekurrent, wenn es einen Zustnd gibt, für die entprechenden Eigenschaften gelten.

Wir haben die folgenden alternative Charakterisierung von Transienz:

Lemma 8.31. *Sei X ein irreduzible Markovkette mit abzählbarem Zustandsraum. Dann ist X transient genau dann, wenn für jeden Zustand $\ell \in S$,*

$$\mathbb{P}_\ell\left(X_t = \ell, \text{i.o.}\right) = 0. \tag{8.6.3}$$

Beweis. Sei X transient, also $\mathbb{P}_\ell(\tau_\ell < \infty) \equiv c < 1$. Wegen der starken Markoveigenschaft sind die sukzessiven Versuche, von ℓ nach ℓ in endlicher Zeit zurückzukommen unabhängig. Daher gilt

$$\mathbb{P}_\ell\left(X_t = \ell,\ n\text{-mal}\right) = \mathbb{P}_\ell(\tau_\ell < \infty)^n \mathbb{P}_\ell(\tau_\ell = \infty) = c^n(1-c). \tag{8.6.4}$$

Nun ist wegen dem ersten Borel-Cantelli Lemma (8.6.3) wahr, falls

$$\sum_t \mathbb{P}_\ell(X_t = \ell) < \infty. \tag{8.6.5}$$

Aber

$$\sum_t \mathbb{P}_\ell(X_t = \ell) = \mathbb{E}_\ell\left(\sum_t \mathbb{1}_{X_t = \ell}\right) = \sum_{n=1}^{\infty} n\mathbb{P}_\ell(X_t = \ell,\ n\text{-mal}). \tag{8.6.6}$$

Da die Summanden wegen (8.6.4) kleiner sind als c^n mit $c < 1$, konvergiert die Summe. Sei umgekehrt (8.6.3) wahr. Nun ist

$$\begin{aligned} 1 - \mathbb{P}_\ell\left(X_t = \ell, \text{i.o.}\right) &= \sum_t \mathbb{P}_\ell(X_t = \ell)\mathbb{P}_\ell(\tau_\ell = \infty) \\ &= \sum_{n=1}^{\infty} n\mathbb{P}_\ell(X_t = \ell,\ n\text{-mal}))\mathbb{P}_\ell(\tau_\ell = \infty). \end{aligned} \tag{8.6.7}$$

Wenn nun die linke Seite der Gleichung gleich 1 ist, so muss $\mathbb{P}_\ell(\tau_\ell < \infty) < 1$ sein. □

Diese Eigenschaft erklärt den Begriff “transient’: eine transiente Kette “verschwindet” fast sicher nach “unendlich” und kommt irgendwann einmal nie wieder zum Startpunkt zurück.

Positiv rekurrente Markovketten verhalten sich ähnlich wie irreduzible Markovketten mit endlichem Zustandsraum. Insbesondere besitzen sie eine einzige invariante Wahrsscheinlichkeitsverteilung. Dies ist der Inhalt des folgenden Satzes.

Satz 8.32. *Sei X eine positiv rekurrente Markovkette mit abzählbarem Zustandsraum S. Dann ist für jedes $j, \ell \in S$,*

$$\mu(j) \equiv \frac{\mathbb{E}_\ell\left(\sum_{t=1}^{\tau_\ell} \mathbb{1}_{X_t=j}\right)}{\mathbb{E}_\ell \tau_\ell}. \tag{8.6.8}$$

die eindeutige invariante (Wahrscheinlichkeits)Verteilung von X.

Beweis. Sei $\nu_\ell(j) = \mathbb{E}_\ell\left[\sum_{t=1}^{\tau_\ell} \mathbb{1}_{X_t=j}\right]$. Das ν_ℓ eine invariante Verteilung ist, haben wir schon in Lemma 8.32 gezeigt; die Tatsache, dass der Zustandsraum endlich ist wurde dort nur genutzt um zu zeigen, dass $\mathbb{E}_\ell\tau_\ell < \infty$; dies ist hier aber eine Annahme.

Wir müssen noch die Eindeutigkeit beweisen. Dazu zeigen wir zunächst folgendes:
Wenn X irreduzibel und μ ein invariantes Maß ist, und für irgendein $i \in S$ $\mu(i) = 0$ gilt, dann ist μ das Nullmaß.

Denn wenn für irgendein $j \in S$ $\mu(j) > 0$, dann gibt es ein endliches t so dass $p^t_{ji} > 0$, und somit wegen der Invarianzeigenschaft, dass

$$\mu(i) = \sum_k \mu(k)(t)_{ki} \geq \mu(j)p^t_{ji} > 0,$$

im Widerspruch zu der Annahme, dass $\mu(i) = 0$.

Wir werden zeigen, dass das oben definierte ν_ℓ das einzige invariante Mass ist, so dass $\nu(\ell) = 1$ gilt. Wegen der obigen Bemerkung folgt daraus die Eindeutigkeit. Gäbe es nämlich ein anderes invariantes Maß ν, dass nicht ein Vielfaches von ν_ℓ ist, so müsste ja $\nu(\ell) > 0$ sein, und daher $\nu' \equiv \nu/\nu(\ell)$ ein invariantes Maß mit $\nu'(\ell) = 1$!

Sei also ν ein invariantes Maß mit $\nu(\ell) = 1$. Wir werden zeigen, dass dann für alle Zustände $j \in S$, $\nu(j) \geq \nu_\ell(j)$. Dann aber wäre $\nu - \nu_\ell$ ein positives invariantes Maß, welches aber in ℓ verschände, weswegen folgen würde, dass $\nu = \nu_\ell$.

Nun gilt, da nach Voraussetzung $\nu(\ell) = 1$,

$$\nu(i) = \sum_{j\neq\ell} \nu(j)p_{ji} + p_{\ell i}. \tag{8.6.9}$$

Wir schreiben $p_{\ell i}$ als

$$p_{\ell i} = \mathbb{E}_\ell\left(\mathbb{1}_{\tau_\ell\geq 1}\mathbb{1}_{X_1=i}\right).$$

Nun können wir die Gleichung (8.6.9) für die Terme in der rechten Seite in sich selbst einsetzen. Es folgt

$$\begin{aligned}\nu(i) &= \sum_{j_1,j_2\neq\ell} p_{j_2j_1}p_{j_1i}\nu(j_2) + \sum_{j_1\neq\ell} p_{\ell j_1}p_{j_1 i} + \mathbb{E}_\ell\left(\mathbb{1}_{\tau_\ell\geq 1}\mathbb{1}_{X_1=i}\right)\\ &= \sum_{j_1,j_2\neq\ell} p_{j_2j_1}p_{j_1i}\nu(j_2) + \mathbb{E}_\ell\left(\sum_{s=1}^{2\wedge\tau_\ell} \mathbb{1}_{X_s=i}\right).\end{aligned} \tag{8.6.10}$$

Weitere Iteration liefert für jedes $n \in \mathbb{N}$

$$\nu(i) = \sum_{j_1, j_2, \ldots j_n \neq \ell} p_{j_n j_{n-1}} \ldots p_{j_2 j_1} p_{j_1 i} \nu(j_n) + \mathbb{E}_\ell \left(\sum_{s=1}^{n \wedge \tau_\ell} \mathbb{1}_{X_s = i} \right)$$
$$\geq \mathbb{E}_\ell \left(\sum_{s=1}^{n \wedge \tau_\ell} \mathbb{1}_{X_s = i} \right). \tag{8.6.11}$$

Da der letzte Ausdruck mit n gegen $\nu_\ell(i)$ konvergiert, folgt, wie angekündigt, $\nu(i) \geq \nu_\ell(i)$, und der Beweis ist abgeschlossen. □

Korollar 8.33. *Für positive rekurrente Markovketten gilt*

$$\mu(j) = \frac{1}{\mathbb{E}_j(\tau_j)}. \tag{8.6.12}$$

Beweis. Wähle $\ell = j$ in der Definition von $\mu(j)$, und beachte, dass

$$\nu_j(j) = \mathbb{E}_j(\sum_{t=1}^{\tau_j} \mathbb{1}_{X_t = x}) = 1.$$

□

Wir sehen, dass die positive Rekurrenz notwendig ist, um die Existenz eines normierbaren invarianten Maßes zu sichern. Wir wollen nun zeigen, dass unter der weiteren Annahme der Aperiodizität auch die Konvergenz gegen das invariante Wahrscheinlichkeitsmaß gegeben ist.

Zunächst zeigen wir, dass die Existenz eines strikt positiven invarianten Wahrscheinlichkeitsmasses positive Rekurrenz impliziert.

Lemma 8.34. *Sei X eine irreduzible Markovkette mit abzählbarem Zustandsraum. Wenn X ein invariantes Wahrscheinlichkeitsmaß μ besitzt, dann ist $\mu(i) = 1/\mathbb{E}_i\tau_i$, und X ist positiv rekurrent.*

Beweis. Da μ Wahrscheinlickeitsmass ist, so muss wegen der Irreduzibilität für jeden Zustand ℓ für geeignetes n gelten, dass $\mu(\ell) = \sum_{i \in S} \mu(i)(p^n)_{i\ell} > 0$. Dann ist $\lambda(j) \equiv \mu(j)/\mu(\ell)$ invariantes Maß mit $\lambda(i) = 1$. Dann haben wir aber im vorigen Beweis gesehen, dass $\lambda(k) \geq \nu_\ell(k)$. Daher gilt

$$\mathbb{E}_\ell \tau_\ell = \sum_{i \in S} \nu_\ell(i) \leq \sum_{i \in S} \frac{\mu(i)}{\mu(\ell)} = \frac{1}{\mu(\ell)} < \infty. \tag{8.6.13}$$

Daher ist X positiv rekurrent. □

Satz 8.35. *Sei X eine irreduzible, aperiodische und positiv rekurrente Markovkette mit abzählbarem Zustandsraum S, Übergangsmatrix P und invarianter Wahrscheinlichkeitsverteilung μ. Dann gilt für jede Anfangsverteilung π_0, dass für alle $i \in S$,*

$$\lim_{n \uparrow \infty} (\pi_0 P^n)_i = \mu(i). \tag{8.6.14}$$

Beweis. Der Beweis benutzt die sogennante "Kopplungsmethode". Sei π_0 die Anfangsverteilung unserer Kette X. Dann konstruieren wir eine zweite, von X unabhängige Markovkette mit derselben Übergangsmatrix aber mit Anfangsverteilung μ. Wir definieren eine Stoppzeit T bezüglich der Filtrierung $\mathfrak{F}_n \equiv \sigma(X_0, Y_0, X_1, Y_1, \ldots, X_n, Y_n)$ als

$$T \equiv \inf\{n : X_n = Y_n = i\}, \tag{8.6.15}$$

wo $i \in S$ ein beliebiger Zustand in S ist.

Wir zeigen zunächst, dass T fast sicher endlich ist. Dazu betrachten wir das Paar $W = (X, Y)$ als Markovkette mit Zustandsraum $S \times S$ und Übergangsmatrix $\widetilde{P}$ mit Elementen

$$\tilde{p}_{(ik)(jm)} \equiv p_{ij} p_{km}. \tag{8.6.16}$$

Die Anfangsverteilung dieser Kette ist $\tilde{\pi}_0((jk)) = \pi_0(j)\mu(k)$. Weil P irreduzibel und aperiodisch ist, so existiert für jedes i, j, k, ℓ ein n, so dass

$$\tilde{p}^n_{(ik)(jm)} = p^n_{ij} p^n_{km} > 0. \tag{8.6.17}$$

Daher ist W irreduzibel. Weiter ist offensichtlich, dass die invariante Verteilung $\tilde{\mu}$ der Kette W gegeben is durch

$$\tilde{\mu}((jk)) = \mu(j)\mu(k) > 0. \tag{8.6.18}$$

Daher ist W positiv rekurrent. Da $T = \inf\{n \geq 0 : W_n = (ii)\}$, ist $\mathbb{E}T < \infty$ und somit $\mathbb{P}(T < \infty) = 1$.

Nun konstruieren wir eine neue Markovkette Z mit Zustandsraum S, nämlich

$$Z_n = \begin{cases} X_n, & \text{wenn } n < T \\ Y_n, & \text{wenn } n \geq T. \end{cases} \tag{8.6.19}$$

Diese Markovkette hat aber dieselbe Verteilung wir X, was man formal mit der starken Markoveigenschaft beweist.

Daraus folgt nun aber

$$\begin{aligned} \mathbb{P}(X_n = i) &= \mathbb{P}(Z_n = i) \\ &= \mathbb{P}(Z_n = i \wedge \{n < T\}) + \mathbb{P}(Z_n = i \wedge \{n \geq T\}) \\ &= \mathbb{P}(X_n = i \wedge \{n < T\}) + \mathbb{P}(Y_n = i \wedge \{n \geq T\}) \\ &= \mathbb{P}(Y_n = i) + -\mathbb{P}(Y_n = i \wedge \{n < T\}) + \mathbb{P}(X_n = i | \{n < T\}) \\ &= \mu(i) + (\mathbb{P}(Y_n = i | n < T) - \mathbb{P}(X_n = i | n < T))\, \mathbb{P}(n < T). \end{aligned} \tag{8.6.20}$$

Nun ist aber der Ausdruck in der Klammer im Betrag kleiner als eins und der Koeffizient $\mathbb{P}(n < T)$ strebt nach Null, wenn $n \uparrow \infty$. Damit ist die Behauptung bewiesen. □

Anmerkung. Beachte, dass zum Beweis der Irreduzibilität der Kette W die Aperiodizität der Kette X notwendig war. So ist zum Beispiel im einfachsten Beispiel der deterministischen periodischen Kette mit Zustandsraum $\{1,2\}$ der Zustand $(1,2)$ nicht vom Zustant $(1,1)$ erreichbar. Der Zustandandsraum der Kette W zerfällt dann in die Klassen $C_1 = \{(12),(2,1)\}$ und $C_2 = \{(11),(22)\}$.

Anmerkung. Die Chebeychev Ungleichung liefert $\mathbb{P}(T > n) \leq \frac{\mathbb{E}T}{n}$. Damit liefert dieser Beweis für den Fall, dass der Zustandsraum endlich ist ein schwächeres Resultat. Allerdings könnte man dann auch zeigen, dass für $\lambda > 0$ klein genug, $\mathbb{E}\exp(\lambda T) < \infty$, woraus in dann exponentiel schnelle Konvergenz zum Gkeichgewicht folgt.

Wir wollen noch anmerken, dass für transiente Zustände, i, einer Markovkette gilt, dass für alle j und für jedes invariante Maß μ,

$$\lim_{n\uparrow\infty}(p^n)_{ji} = 0 = \mu(i).$$

Es gilt nämlich, dass wegen Lemma 8.6.14

$$\sum_{n=0}^{\infty}(p^n)_{ji} \leq \mathbb{E}_i\left(\sum_{n=0}^{\infty}\mathbb{1}_{X_n=i}\right) < \infty.$$

Daraus folgt aber die Behauptung sofort.

Abschliessend bemerken wir noch, dass der starke Ergodensatz (Satz 8.26) auch für positive rekurrente Markovketten mit abzählbarem Zustandsraum gilt. Um dies zu sehen, dass im Beweis dieses Satzen die Annahme endlichen Zustandstaumes nur benutzt wird um die Existenz und Eindeutigkeit einer invarianten Verteilung sowie die Endlichleit von $\mathbb{E}_\ell\tau_\ell$ sicherzustellen, was aber im positiv rekurrenten Fall auch gilt.

Literaturverzeichnis

1. P. Billingsley. *Probability and measure*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, 1995.
2. Yuan Shih Chow and Henry Teicher. *Probability theory*. Springer Texts in Statistics. Springer-Verlag, New York, third edition, 1997.
3. William Feller. *An introduction to probability theory and its applications. Vol. I*. Third edition. John Wiley & Sons Inc., New York, 1968.
4. William Feller. *An introduction to probability theory and its applications. Vol. II*. Second edition. John Wiley & Sons Inc., New York, 1971.
5. H.-O. Georgii. Spontaneous magnetization of randomly dilute ferromagnets. *J. Statist. Phys.*, 25(3):369–396, 1981.
6. Hans-Otto Georgii. *Stochastik*. de Gruyter Lehrbuch. Walter de Gruyter & Co., Berlin, 2002.
7. Samuel Karlin and Howard M. Taylor. *A first course in stochastic processes*. Academic Press [A subsidiary of Harcourt Brace Jovanovich, Publishers], New York-London, second edition, 1975.
8. G. Kersting and A. Wakolbinger. *Elementare Stochastik*. Birkhäuser, Basel, Boston, Berlin, 2008.
9. A. Klenke. *Wahrscheinlichkeitstheorie*. Springer-Verlag, New York, 2006.
10. Pierre Simon de Laplace. *Théorie Analytique des Probabilités*. V. Courcier, Paris, 1820. available online `http://gallica.bnf.fr/ark:/12148/bpt6k775950/f4`.
11. J. W. Lindeberg. Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung. *Math. Zeitschrift*, 15(4):211–225, 1922.

Glossary

$\mathfrak{A}$ Algebra

$\mathfrak{B}$ Borel'asche σ-Algebra

$\mathfrak{F}$ σ-Algebra

$\mathfrak{C}$ Mengen-System

$\mathcal{D}(\mathfrak{C})$ von $\mathfrak{C}$ erzeugtes Dynkinsystem

$\sigma(\mathfrak{C})$ von $\mathfrak{C}$ erzeugte σ-Algebra

Ω Menge

$\mathbb{P}$ Wahrscheinlichkeitsmaß, meißt auf einen abstrakten W-Raum

μ Maß

P_f Bildmaß von $\mathbb{P}$ unter f

X Zufallsvariable

$\mathbb{E}$ Erwartung bezüglich $\mathbb{P}$

$\mathbb{P}(A|B)$ Bedingte Wahrscheinlichkeit von A gegeben B

$\mathcal{N}(m, \sigma^2)$ Gaußverteilung mit Mittelwert m und Varianz σ^2

F Verteilungsfunktion

$\mathbb{1}_A$ Indikatorfunktion der Menge A

δ_x Diracmaß auf x

X, Y, Z Zufallsvariablen

$\sigma(X)$ von X erzeugte σ-Algebra

τ Stoppzeit

τ_D Erste Treffzeit von D.

Sachverzeichnis