Stochastik

LEIF DÖRING[1]

UNIVERSITÄT MANNHEIM

[1]Besten Dank an den Skriptschreiber (aka Johannes Nägele) fürs Tippen und die Erstellung der Graphiken! Auch vielen Dank für die Korrekturen (auch zukünftige) von Tippfehlern durch Teilnehmer der Vorlesung!

Inhaltsverzeichnis

Was machen wir, was nicht?

„Stochastik“ ist ein Oberbegriff für „Mathematik des Zufalls“. In Mannheim ist die Stochastik in Lehre und Forschung sehr ausgeprägt:

Modellierung und theoretische Untersuchung zufälliger Experimente

$\Updownarrow$

Wahrscheinlichkeitstheorie ($\rightsquigarrow$ Döring)

•

Anpassung der Modelle auf „echte“ zufällige Experimente

$\Updownarrow$

Mathematische Statistik ($\rightsquigarrow$ Schlather)

•

Ausführung der Modelle („Zufall erzeugen“)

$\Updownarrow$

Stochastische Numerik ($\rightsquigarrow$ Neuenkirch)

•

Anwendung auf Finanzmärkte

$\Updownarrow$

Finanzmathematik ($\rightsquigarrow$ Prömel)

•

Anwendung auf Wirtschaftsdaten

$\Updownarrow$

Ökonometrie ($\rightsquigarrow$ Trenkler, Rothe)

•

Zählen von Möglichkeiten $\rightsquigarrow$ Gleichverteilung (z. B. Lotto; Ziehen aus Urnen)

$\Updownarrow$

Kombinatorik (in Mannheim nicht vertreten)

Das Ziel dieser Vorlesung ist es, die Grundlagen der Stochastik zu legen. Das ist anfangs etwas trocken, ihr werdet aber im Verlauf des Studiums davon profitieren, dass alle Begriffe auf stabilen Fundament stehen. In den Vorlesungen Stochastik 2, Monte Carlo Methoden, Finanzmathematik, Ökonometrie und Wahrscheinlichkeitstheorie 1 werden die Grundlagen noch im Bachelor angewandt und in diversen Spezialisierungsrichtungen im Master erweitert.

Teil 1: Maß- und Integrationstheorie

Kapitel 1

Maßtheorie (Modellierung von Ereignissen und Wahrscheinlichkeiten)

Vorlesung 1

Maß- und Integrationstheorie bildet die formale Grundlage um zufällige Experimente zu modellieren. In diesem ersten Teil der Vorlesungen beweisen wir alle notwendigen Theoreme. Nicht alles wird später uneingeschränkt wichtig sein, das Arbeiten mit den neuen Begriffen wird sich in zukünftigen Vorlesungen aber auszahlen!

1.1 σ-Algebren und Maße - die Grundbegriffe der Stochastik

Im Prinzip sind die kommenden fünf Vorlesungen total elementar, wir brauchen nur Kenntnisse über Mengen, Folgen und Reihen. Die Vorlesung nutzt also nur Kenntnisse der Analysis 1. Dennoch wird euch der Inhalt zunächst schwer fallen weil wir Mengensysteme nicht visualisieren können und daher viel abstrakt denken müssen. Es wird sehr wichtig sein, die richtigen Beispiele im Kopf zu haben. Diese sollten nicht zu einfach sein, weil sonst der Großteil der Schwierigkeiten nicht erkannt werden kann. Für σ-Algebren sollten wir möglichst schnell die Borel-σ-Algebra als Standdardbeispiel im Kopf halten, für Maße das Lebesgue-Maß. Endliche Beispiele werden wir nur ganz kurz als Motivation der Maßtheorie für Stochastik betrachten (Würfeln, Münzwurf, etc.), solche Beispiele bringen leider nicht viel um die Konzepte der Wahrscheinlichkeitstheorie richtig zu verstehen.

Im Folgenden sei $\Omega \neq \emptyset$ immer eine beliebige Grundmenge. Für $A \subseteq \Omega$ bezeichnet A^C immer das Komplement von A in Ω, d. h. $A^C = \{w \in \Omega \mid w \notin A\}$. $\mathcal{P}(\Omega)$ bezeichnet die Potenzmenge von Ω (inklusive $\emptyset$ und Ω), eine Teilmenge von $\mathcal{P}(\Omega)$ ist also eine Menge von Mengen (man sagt auch Mengensystem).

Definition 1.1.1. $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ heißt σ-Algebra, falls

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

(ii) $A \in \mathcal{A} \Rightarrow A^C \in \mathcal{A}$, das nennt man auch abgeschlossen (oder stabil) unter Komplementbildung,

(iii) $A_1, A_2, ... \in \mathcal{A} \Rightarrow \bigcup_{k=1}^{\infty} A_k \in \mathcal{A}$, das nennt man auch abgeschlossen (oder stabil) unter abzählbarer Vereinigung.

Elemente von $\mathcal{A}$ heißen **messbare Mengen**. Ist $\mathcal{A} \subseteq \mathcal{B}$ und $\mathcal{A}, \mathcal{B}$ sind σ-Algebren, so nennt man $\mathcal{A}$ Unter-σ-Algebra von $\mathcal{B}$.

Die Stabilität für Vereinigungen gilt auch, wenn man nur endlich viele messbare Mengen vereinigen will. Dazu nutzt man einfach folgenden Trick: Wenn man messbare Mengen $A_1, ..., A_N$ vereinigen möchte, so setzt man $A_{N+1} = A_{N+2} = ... = \emptyset$ und beachtet, dass damit wegen der σ-Additivität $\cup_{k=1}^N A_k = \cup_{k=1}^\infty A_k \in \mathcal{A}$ gilt. Merkt euch solche kleinen Tricks, der selbe Trick taucht gleich nochmal auf.

Example 1.1.2. Ist $\Omega \neq \emptyset$ eine beliebige Grundmenge, so sind folgende Mengensysteme σ-Algebren:

- $\mathcal{A}_1 = \{\emptyset, \Omega\}$
- $\mathcal{A}_2 = \mathcal{P}(\Omega)$
- $\mathcal{A}_3 = \{\emptyset, \Omega, A, A^C\}$ für $A \subseteq \Omega$ beliebig
- $\mathcal{A}_4 = \{A \subseteq \Omega \mid A \text{ oder } A^C \text{ ist abzählbar}\}$

Da $\mathcal{A}_1, ..., \mathcal{A}_4$ Teilmengen der Potenzmenge sind, muss man jeweils nur die drei definierenden Eigenschaften einer σ-Algebra testen. Bei den ersten drei Beispielen ist das direkt, indem man alle Möglichkeiten ausprobiert. Im vierten Beispiel müssen wir nur bei der abzählbaren Vereinigung kurz nachdenken. Seien also $A_1, A_2, ...$ Teilmengen von Ω, die entweder abzählbar sind oder deren Komplemente abzählbar sind. Sind all diese Mengen abzählbar, so ist nach Analysis 1 auch die Vereinigung abzählbar, also ist die Vereinigung wieder in $\mathcal{A}_3$. Ist eine der Mengen nicht abzählbar, sagen wir A_j, so ist das Komplement A_j^C abzählbar. Doch dann ist wegen

$$\Big(\bigcup_{k=1}^\infty A_i\Big)^C \overset{\text{de Morgan}}{=} \bigcap_{k=1}^\infty A_i^C \subseteq A_j^C$$

das Komplement der Vereinigung nach Analysis 1 abzählbar. Also ist die Vereinigung in $\mathcal{A}_4$ und damit $\mathcal{A}_4$ abgeschlossen bezüglich Vereinigungen.

Lemma 1.1.3. Für jede σ-Algebra $\mathcal{A}$ gelten

(i) $\emptyset \in \mathcal{A}$

(ii) $A_1, A_2, ... \in \mathcal{A} \Rightarrow \bigcap_{k=1}^\infty A_k \in \mathcal{A}$

(iii) Aus $A, B \in \mathcal{A}$ folgt A $\backslash B := A \cap B^C \in A$ sowie $A \Delta B := (A \cap B^C) \cup (B \cap A^C) \in \mathcal{A}$.

Beweis. Die Strategie ist immer gleich: Man versucht die Behauptung aus den drei Regeln einer σ-Algebra herzuleiten. Da $\emptyset = \Omega^C$ gilt, gilt wegen den Eigenschaften (i) und (ii) einer σ-Algebra auch $\emptyset \in \mathcal{A}$. Mit de Morgan und den Eigenschaften (ii), (iii) der σ-Algebra gilt

$$\bigcap_{k=1}^\infty A_k = \underbrace{\Big(\underbrace{\bigcup_{k=1}^\infty \underbrace{A_k^C}_{\in \mathcal{A},\ (\text{ii})}}_{\in \mathcal{A},\ (\text{iii})}\Big)^C}_{\in \mathcal{A},\ (\text{ii})}.$$

Probiert die dritte Behauptung mal selber aus, eigentlich steht alles schon da. □

Genau wie bei Vereinigungen, gilt die Abgeschlossenheit auch für endliche Schnitte. Probiert das mal selber aus, bei dem Trick nutzt ihr aber Ω statt $\emptyset$.

Bemerkung. Wie in Analysis 1 nutzen wir die **erweiterte Zahlengerade**

$$\overline{\mathbb{R}} = [-\infty, +\infty] := \mathbb{R} \cup \{-\infty, +\infty\}.$$

Wir definieren

- $-\infty < a < +\infty$ für alle $a \in \mathbb{R}$,
- $+\infty + a = +\infty$ und $-\infty + a = -\infty$ für alle $a \in \mathbb{R}$,
- $x \cdot (+\infty) = +\infty$ und $x \cdot (-\infty) = -\infty$ für alle $x > 0$,
- $0 \cdot (+\infty) = 0$ und $0 \cdot (-\infty) = 0$,
- $+\infty + (+\infty) = +\infty$ und $-\infty + (-\infty) = -\infty$,
- $-\infty + (+\infty)$ wird nicht definiert.

Im Gegensatz zu $\mathbb{R}$ können wir aus $\overline{\mathbb{R}}$ keine sinnvolle algebraische Struktur formen, das soll uns aber nicht weiter stören. Sehr oft schreibt man ∞ statt $+\infty$. Wenn wir in dieser Vorlesung von den natürlichen Zahlen sprechen, meinen wir $\mathbb{N} = \{0, 1, ...\}$, die 0 soll also zu $\mathbb{N}$ gehören.

Definition 1.1.4. Für eine σ-Algebra $\mathcal{A}$ heißt eine Abbildung $\mu\colon \mathcal{A} \to [0, \infty]$ ein **Maß auf** $\mathcal{A}$, falls folgende Eigenschaften gelten:

(i) $\mu(\emptyset) = 0$

(ii) Sind $A_1, A_2, ... \in \mathcal{A}$ paarweise disjunkte Mengen (d. h. $A_i \cap A_j = \emptyset$ für alle $i \neq j$), so gilt

$$\mu\Big(\dot{\bigcup_{k=1}^{\infty}} A_k\Big) = \sum_{k=1}^{\infty} \mu(A_k).$$

Wir nenne diese Eigenschaft σ-Additivität, wobei sich das σ auf die unendliche Anzahl von Mengen bezieht.

Ein Maß μ heißt **endlich**, falls $\mu(\Omega) < \infty$. μ heißt **Wahrscheinlichkeitsmaß**, falls $\mu(\Omega) = 1$.

Natürlich impliziert die σ-Additivität auch die endliche Additivität $\mu\Big(\dot{\bigcup}_{k=1}^{N} A_k\Big) = \sum_{k=1}^{N} \mu(A_k)$. Dazu wird, wie unter der Definition der σ-Algebra, $A_{N+1} = A_{N+2} = ... = \emptyset$ gewählt. Der Begriff „Maß“ hat durchaus einen Sinn, der an Beispielen später viel klarer wird. Man misst in einem abstrakten Sinn die Größe der messbaren Mengen. Deswegen sind die zwei definierenden Eigenschaften auch klar. Malt euch einfach mal zwei Mengen in $\mathbb{R}^2$ hin und überlegt, warum die „Größe“ nur für disjunge Mengen die Summe der „Größen“ der einzelnen Mengen sein sollte.

Bemerkung 1.1.5. Oft werden Wahrscheinlichkeitsmaße mit $\mathbb{P}$ anstelle von μ geschrieben und **Verteilungen** oder **Wahrscheinlichkeitsverteilung** genannt.

Folgende Begrifflichkeiten werden wir ständig nutzen, um möglichst effizient formulieren zu können:

Definition 1.1.6.

- $(\Omega, \mathcal{A})$ heißt **messbarer Raum**
- $(\Omega, \mathcal{A}, \mu)$ heißt **Maßraum**
- $(\Omega, \mathcal{A}, \mathbb{P})$ heißt **Wahrscheinlichkeitsraum**
- $\mu(A)$ nennt man **Maß von** A oder Masse von A
- $\mu(\Omega)$ nennt man **Gesamtmasse von** μ

Bemerkung 1.1.7. Bei einem Wahrscheinlichkeitsraum spricht man von **Ereignissen** A statt messbaren Mengen. $\mathbb{P}(A)$ heißt **Wahrscheinlichkeit** von A. Einelementige messbare Mengen $A = \{a\}$ heißen in Wahrscheinlichkeitsräumen **Elementarereignisse**.

Um langsam in die Denkweise der Stochastik einzusteigen, werden wir wieder und wieder diskutieren, warum unsere formellen Modelle für die Modellierung echter zufälliger Experimente gut geeignet sind.

Diskussion 1.1.8. **[Stochastische Modellierung, Nr. 1]** Warum machen die Definitionen von Wahrscheinlichkeitsräumen $(\Omega, \mathcal{A}, \mathbb{P})$ für die Modellierung von zufälligen Experimenten Sinn? Wir interpretieren dazu

- $\Omega =$ „Das kann bei dem Experiment passieren", wir können aber vielleicht das Eintreten der Elementarereignisse nicht beobachten.
- $\mathcal{A} =$ „Ereignisse, deren Eintreten (oder Nichteintreten) beobachtet werden kann." Die σ-Algebra besteht also aus den Ereignissen des Experiments, die wir beobachten können.
- $A^C =$ „Gegenereignis", also „Ereigniss A trifft nicht ein".
- $A \in \mathcal{A} \Rightarrow A^C \in \mathcal{A}$ bedeutet „Wenn man das Eintreten von Ereignis A beobachten kann, dann kann man auch beobachten, dass A nicht eintritt."
- $A, B \in \mathcal{A} \Rightarrow A \cup B \in \mathcal{A}$ bedeutet „Wenn man das Eintreten sowohl von Ereignis A als auch von Ereignis B beobachten kann, dann kann auch beobachten, ob eines von beiden auftritt." Die Interpretation der Abgeschlossenheit bezüglich endlicher Vereinigungen ist analog („Man kann beobachten, ob eines der Ereignisse eingetreten ist"). Wir lassen hier offen, warum man für die Mathematik auch abzählbare Vereinigungen erlauben muss. Hier bleibt für den Moment nur zu sagen: Es würde nicht funktionieren.
- $\mathbb{P}(A) =$ „Wahrscheinlichkeit des Eintretens des Ereignisses A."
- $\mathbb{P}(A^C) = 1 - \mathbb{P}(A)$ bedeutet „Gegenereignis hat Gegenwahrscheinlichkeit." Die Gleichheit gilt, da wegen $A \,\dot\cup\, A^C = \Omega$ auch $\mathbb{P}(A) + \mathbb{P}(A^C) = \mathbb{P}(\Omega) = 1$ gilt.

Damit haben wir den Sinn der Definitionen einer σ-Algebra und eines Maßes hoffentlich großteils motiviert.

Als Beispiel modellieren wir den Wurf eines Würfels gemäß obiger Interpretation. Wir wählen $\Omega = \{1, ..., 6\}$ weil das zufällige Experiment (Würfel werfen) sechs Möglichkeiten hat. Die Zahlen spielen hier keine Rolle, es geht nur darum, dass es sechs Elementarereignisse des Experiments gibt. Wir könnten die Elementarereignisse zum Beispiel auch $\Omega = \{\omega_1, ..., \omega_6\}$ nennen. Als σ-Algebra der beobachtbaren Ereignisse nehmen wir $\mathcal{A} = \mathcal{P}(\Omega)$ weil wir alle Ereignisse des Würfelwurfs beobachten können. Ein Ereigniss $A \in \mathcal{A}$ bedeutet „Eine der Zahlen in A ist gewürfelt worden". Weil unser Würfel fair sein soll, legen wir $\mathbb{P}(\{1\}) = ... = \mathbb{P}(\{6\}) = \frac{1}{6}$ fest. Die Wahrscheinlichkeiten aller weiteren Ereignisse sind automatisch festgelegt, indem das Ereigniss in die disjunkten Elementarereignisse zerlegt wird, z. B. die Wahrscheinlichkeit eine gerade Zahl zu würfeln:

$$\mathbb{P}(\{2,4,6\}) = \mathbb{P}(\{2\} \,\dot\cup\, \{4\} \,\dot\cup\, \{6\}) \overset{\text{disj.}}{=} \mathbb{P}(\{2\}) + \mathbb{P}(\{4\}) + \mathbb{P}(\{6\}) = \frac{3}{6} = \frac{1}{2}.$$

Nicht jedes zufällige Experiment ist so einfach wie das Würfeln (endlich viele Möglichkeiten), für kompliziertere zufällige Experimente (z. B. die Temperatur morgen in Mannheim) brauchen wir leider viel kompliziertere Modelle. Gehen wir also zurück zu allgemeinen Maßräumen.

Zunächst eine kleine Folgerung der Definition des Maßes, die im Sinne von „$\mu(A) =$Größe von A" von Mengen total Sinn macht.

Lemma 1.1.9. **[Monotonie und Subadditivität]** Es sei μ ein Maß auf einer σ-Algebra $\mathcal{A}$, dann gelten:

(i) Sind $A, B \in \mathcal{A}$ mit $B \subseteq A$, so gilt $\mu(B) \leq \mu(A)$.

(ii) Sind $A_1, A_2, ... \in \mathcal{A}$, so gilt: $\mu(\cup_{k=1}^{\infty} A_k) \leq \sum_{k=1}^{\infty} \mu(A_k)$.

Beweis. (i) Mit den definierenden Eigenschaften des Maßes gilt:

$$\mu(B) \overset{\mu\geq 0}{\leq} \mu(B) + \mu(A\backslash B) = \mu(B \cup A\backslash B) = \mu(A),$$

wobei wir beide definierenden Eigenschaften des Maßes genutzt haben.

(ii) wird in den Tutorien oder der großen Übung diskutiert. □

Zu beachten ist, dass bei der Subadditivität nicht gefordert wird, dass die Mengen disjunkt sind (sonst würde schließlich Gleichheit gelten!). Die Ungleichheit entsteht dadurch, dass bei nicht-disjunkten Mengen der Schnitt mehrfach gezählt wird.

Um ein wenig mit der Definition des Maßes zu experimentieren, rechnet mal die folgende Bemerkung nach:

Bemerkung. Sind μ_1, μ_2 Maße auf $\mathcal{A}$ und $a, b \geq 0$, so ist auch die Summe

$$a\mu_1 + b\mu_2 : A \mapsto a\mu_1(A) + b\mu_2(A)$$

ein Maß auf $\mathcal{A}$. Summen von Maßen nennt man auch **Mischung**. Sind $\mathbb{P}_1$ und $\mathbb{P}_2$ Wahrscheinlichkeitsmaße und zusätzlich $a + b = 1$, so ist die Mischung $\mathbb{P} = a\mathbb{P}_1 + b\mathbb{P}_2$ wieder ein Wahrscheinlichkeitsmaß.

Kommen wir nun zu ein paar Beispielen:

Beispiel 1.1.10. **[endliche Gleichverteilung]** Sei $\#\Omega < \infty$ und $\mathcal{A} = \mathcal{P}(\Omega)$. Dann heißt das Maß $\mu(A) = \frac{\#A}{\#\Omega}$ Gleichverteilung auf Ω. Checkt mal selber, dass μ die Eigenschaften von Maßen erfüllt. Weil $\mu(\Omega) = 1$ gilt, würde man $\mathbb{P}$ statt μ schreiben. Der Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ ist ein Modell für das zufällige Experiment, in dem aus $\#\Omega$ vielen Elementen jedes Element mit der selben Wahrscheinlichkeit gezogen wird, zum Beispiel Lotto.

Beispiel 1.1.11. **[abzählbare Verteilungen, Zählmaß]** Sei Ω abzählbar, z. B. $\Omega = \mathbb{N}$. Wir wählen $\mathcal{A} = \mathcal{P}(\Omega)$ und eine Folge $(p_k)_{k\in\Omega}$ nicht-negativer Zahlen. Definieren wir

$$\mu(A) := \sum_{k\in A} p_k, \quad A \in \mathcal{A},$$

so ist μ ein Maß. Weil ein Maß per Definition nicht-negativ ist, muss natürlich $p_k \geq 0$ gelten für alle $k \in \mathbb{N}$ (um das einzusehen, wähle $A = \{k\}$). Zwei Spezialfälle:

- Damit μ ein Wahrscheinlichkeitsmaß ist, muss $\sum_{k\in\Omega} p_k = \mu(\Omega) = 1$ gelten. In dem Fall würden wir wieder $\mathbb{P}$ statt μ schreiben.
- Ist $p_k = 1$ für alle $k \in \Omega$, so heißt μ **Zählmaß** weil $\mu(A) = \#A$ die Anzahl der Elemente von A zählt.

Die p_k werden auch Gewichte, oder Wahrscheinlichkeitsgewichte, genannt.

Beispiel 1.1.12. **[Poissonverteilung auf $\mathbb{N}$]** Hier ist ein konkretes Beispiel zu der vorherigen Klasse von Beispielen, die Poissonverteilung. Für ein $\lambda > 0$ (λ nennt man Parameter der Verteilung) sei $p_k = e^{-\lambda}\frac{\lambda^k}{k!}$ für $k \in \mathbb{N}$. Es gelten dann

- $p_k \geq 0$ für alle $k \in \mathbb{N}$,
- $\sum\limits_{k=0}^{\infty} p_k = \sum\limits_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}\sum\limits_{k=0}^{\infty}\frac{\lambda^k}{k!} = e^{-\lambda}e^{\lambda} = e^{-\lambda+\lambda} = 1.$

Also definiert $\mathbb{P}(A) = e^{-\lambda}\sum_{k\in A}\frac{\lambda^k}{k!}$ ein Wahrscheinlichkeitsmaß auf $(\mathbb{N}, \mathcal{P}(\mathbb{N}))$. Man nennt $\mathbb{P}$ auch **Poissonverteilung mit Parameter λ**.

Beispiel 1.1.13. **[Diracmaß]** Sei $\mathcal{A}$ eine σ-Algebra auf Ω und $x \in \Omega$, so heißt

$$\delta_x(A) := \begin{cases} 1, & x \in A \\ 0, & x \notin A \end{cases}, \quad A \in \mathcal{A},$$

Diracmaß an der Stelle x. Die Eigenschaften eines Maßes kann man ganz einfach checken:

(i) Aufgrund der Definition gilt natürlich $\delta_x(\emptyset) = 0$.

(ii) Für paarweise disjunkte Mengen $A_1, A_2, ... \in \mathcal{A}$ gilt

$$\begin{aligned}\delta_x\Big(\dot{\bigcup_{k=1}^{\infty}} A_k\Big) &= \begin{cases} 1 & : x \in \dot{\bigcup}_{k=1}^{\infty} A_k \\ 0 & : x \notin \dot{\bigcup}_{k=1}^{\infty} A_k \end{cases} \\ &= \sum_{k=1}^{\infty} \delta_x(A_k),\end{aligned}$$

weil in der unendlichen Summe nur der Summand 1 sein kann, in dem x liegt.

Also ist das Diracmaß ein Maß auf $\mathcal{A}$.

Weitere wichtige Beispiele wie die geometrische Verteilung und die Binominalverteilung kommen auf dem Übungsblatt zum Ausprobieren. An dieser Stelle legen wir die Begrifflichkeiten der Stochastik wieder beiseite und beschäftigen uns für die nächsten Wochen nur mit allgemeinen Maßen. Zum Gewöhnen für später beachtet, dass endliche Maße und Wahrscheinlichkeitsmaße sehr eng beieinander liegen: Durch $\mathbb{P}(A) := \frac{\mu(A)}{\mu(\Omega)}$ kann ein endliches Maß immer zu einem Wahrscheinlichkeitsmaß „normiert“ werden.

Um uns mit den definierenden Eigenschaften weiter vertraut zu machen, beweisen wir eine wichtige Eigenschaft von Maßen: Vorlesung 2

Satz 1.1.14. **[Stetigkeit von Maßen]** Sei $(\Omega, \mathcal{A}, \mu)$ ein Maßraum und $(A_n)_{n\in\mathbb{N}}$ eine Folge messbarer Mengen, so gelten:

(i) Aus $A_n \uparrow A$ (d. h. $A_1 \subseteq A_2 \subseteq A_3 \subseteq ...$, $\bigcup_{n=1}^{\infty} A_n = A$) folgt $\lim_{n\to\infty} \mu(A_n) = \mu(A)$.

(ii) Aus μ endlich und $A_n \downarrow A$ (d. h. $A_1 \supseteq A_2 \supseteq ...$, $\bigcap_{n=1}^{\infty} A_n = A$) folgt $\lim_{n\to\infty} \mu(A_n) = \mu(A)$.

Beweis.

(i) Definiere

$$A_1' \coloneqq A_1, \quad A_2' \coloneqq A_2 \backslash A_1', \quad A_n' \coloneqq A_n \backslash A_{n-1}', \quad n \geq 3.$$

Malt euch auf jeden Fall eine Skizze, um die Bedeutung dieser Mengen zu verstehen! Weil die A_n' paarweise disjunkt sind und $A_n = \dot{\bigcup}_{k=1}^{n} A_k'$ gilt, folgt

$$\begin{aligned}\lim_{n\to\infty} \mu(A_n) &= \lim_{n\to\infty} \mu\Big(\dot{\bigcup_{k=1}^{n}} A_k'\Big) \overset{\text{Def. Maß}}{=} \lim_{n\to\infty} \sum_{k=1}^{n} \mu(A_k') \overset{\text{Def.}}{=} \sum_{k=1}^{\infty} \mu(A_k') \\ &\overset{\text{Def. Maß, disj.}}{=} \mu\Big(\dot{\bigcup_{k=1}^{\infty}} A_k'\Big) = \mu\Big(\bigcup_{k=1}^{\infty} A_k\Big) = \mu(A).\end{aligned}$$

(ii) Die Behauptung sofort aus (i) weil

$$A_n \downarrow A \quad \Leftrightarrow \quad A_n^C \uparrow A^C, \quad n \to \infty.$$

Weil μ endlich ist gilt für alle messbaren Mengen

$$\mu(\Omega) = \mu(A \cup A^C) \overset{\text{Maß}}{=} \mu(A) + \mu(A^C)$$

und damit auch $\mu(A) = \mu(\Omega) - \mu(A^C)$. Damit folgt mit (i)

$$\lim_{n\to\infty} \mu(A_n) = \lim_{n\to\infty} \big(\mu(\Omega) - \mu(A_n^C)\big) = \mu(\Omega) - \lim_{n\to\infty} \mu(A_n^C) = \mu(\Omega) - \mu(A^C) = \mu(A).$$

□

Für den Moment ist noch nicht so klar, wie wichtig die Stetigkeit von Maßen ist. Es wird aber kaum eine Vorlesung in der Stochastik 1 geben, in der die Stetigkeit von Maßen nicht auftaucht.

Beispiel 1.1.15. **[Gegenbeispiel zu (ii) mit $\mu(\Omega) = \infty$]** In dem Beweis haben wir die Endlichkeit des Maßes deutlich benutzt, sonst stände auf beiden Seiten ein unendlicher Summand. Folgendes Beispiel zeigt, dass die Aussage bei unendlichen Maßen tatsächlich schief gehen kann. Sei dazu $\Omega = \mathbb{N}, \mathcal{A} = \mathcal{P}(\mathbb{N})$ und μ das Zählmaß aus Beispiel 1.1.11, also

$$\mu(A) = \sum_{k\in A} 1 = \#A.$$

Wegen $\mu(\mathbb{N}) = \#\mathbb{N} = +\infty$ ist das Zählmaß unendlich. Mit $A_n = \{n, n+1, ...\}$ gilt $\mu(A_n) = +\infty$ für alle $n \in \mathbb{N}$ und $A_n \downarrow A = \emptyset$. Weil $\mu(\emptyset) = 0$ und $\mu(A_n) = +\infty$ für alle $n \in \mathbb{N}$ gilt, ist die Aussage von Satz 1.1.14 hier falsch.

1.2 Erzeuger von σ-Algebren und Dynkin-Systeme

Bisher waren die Beispiel sehr einfach. Die Angelegenheit wird aber sehr viel schwieriger, wenn wir überabzählbare Grundmengen Ω betrachten, $\mathbb{R}$ zum Beispiel. Wir werden in diesem Abschnitt überlegen, wie man trotzdem Maße auf sehr komplizierten σ-Algebren $\mathcal{A}$ auf Ω verstehen kann. Der Trick wird sein, die Maße nur auf relativ einfachen Teilmengen von $\mathcal{A}$ anzuschauen. Das ist ein wenig wie in der linearen Algebra, dort müssen lineare Abbildungen auch nur auf einer Basis definiert werden. Um das Konzept dieser einfachen Teilmengen (sogenannte Erzeuger) verstehen zu können, braucht es ein wenig Vorarbeit.

Satz 1.2.1. Der Durchschnitt einer beliebigen Menge von σ-Algebren über Ω ist eine σ-Algebra auf Ω.

Beweis. Ruft euch zunächst in Erinnerung, dass σ-Algebren über Ω Mengen sind, Mengen von Teilmengen von Ω. Mengen kann man schneiden, also macht die Aussage grundsätzlich Sinn. Sei nun $\mathcal{A}_i$, $i \in I$, eine Menge von σ-Algebren über Ω und $\mathcal{A} := \bigcap_{i\in I} A_i$. Wir checken die drei Eigenschaften einer σ-Algebra für $\mathcal{A}$:

(i) $\Omega \in \mathcal{A}$ ist klar weil $\Omega \in \mathcal{A}_i$ für alle $i \in I$ und damit ist Ω auch im Durchschnitt.

(ii) Sei $A \in \mathcal{A}$, also ist $A \in \mathcal{A}_i$ für alle $i \in I$. Weil alle $\mathcal{A}_i$ σ-Algebren sind, ist auch $A^C \in \mathcal{A}_i$ für alle $i \in I$ und damit ist A^C im Durchschnitt der $\mathcal{A}_i$. Folglich ist $\mathcal{A}$ abgeschlossen bezüglich Komplementbildung.

(iii) Sei (A_n) eine Folge von Mengen in $\mathcal{A}$, also $A_n \in \mathcal{A}_i$ für alle i und n. Weil das alles σ-Algebren sind, gilt

$$\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}_i$$

für alle $i \in I$. Damit ist die Vereinigung auch im Durchschnitt aller $\mathcal{A}_i$, also $\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$. Folglich ist $\mathcal{A}$ auch abgeschlossen bezüglich beliebigen Vereinigungen.

□

Wie an an anderen Stellen der Mathematik (z. B. bei Unterverktorräumen) ist der Schnitt oft strukturerhaltend. Die Vereinigung meistens nicht, das ist auch bei σ-Algebren so:

Bemerkung 1.2.2. Die Vereinigung von σ-Algebren ist nicht immer eine σ-Algebra. Für das Übungsblatt sollt ihr euch dazu Beispiele überlegen.

Korollar 1.2.3. Sei $\mathcal{E} \subseteq \mathcal{P}(\Omega)$, so existiert genau eine σ-Algebra $\mathcal{A}$ mit

(i) $\mathcal{E} \subseteq \mathcal{A}$

(ii) Ist $\mathcal{E} \subseteq \mathcal{B}$ und $\mathcal{B}$ ist eine σ-Algebra, so gilt $\mathcal{A} \subseteq \mathcal{B}$.

Dabei bedeutet (ii), dass $\mathcal{A}$ die kleinste σ-Algebra ist, die $\mathcal{E}$ enthält.

Beweis. Existenz:

$$\mathcal{A} := \bigcap_{\substack{\mathcal{E} \subseteq \mathcal{B}, \\ \mathcal{B}\ \sigma\text{-Alg.}}} \mathcal{B}$$

erfüllt die geforderten Eigenschaften.
Eindeutigkeit: Sei $\mathcal{A}'$ eine weitere solche σ-Algebra. Dann gilt $\mathcal{A} \subseteq \mathcal{A}'$ weil $\mathcal{A}$ der Schnitt über alle solche $\mathcal{A}'$ ist und der Schnitt von Mengen in jeder Menge enthalten ist, über die geschnitten wird. Weil $\mathcal{A}'$ die Eigenschaft (ii) erfüllt, ist auch $\mathcal{A}' \subseteq \mathcal{A}$. Damit ist $\mathcal{A} = \mathcal{A}'$ gezeigt und es gibt nur eine σ-Algebra mit den Eigenschaften (i) und (ii). □

Die σ-Algebra aus dem Korollar hat eine enorme Bedeutung in der Stochastik, ohne sie wäre der Rest dieses Abschnittes nicht machbar. Geben wir dieser σ-Algebra also einen Namen:

Definition 1.2.4. Für $\mathcal{E} \subseteq \mathcal{P}(\Omega)$ heißt

$$\sigma(\mathcal{E}) = \bigcap_{\substack{\mathcal{E} \subseteq \mathcal{B}, \\ \mathcal{B}\ \sigma\text{-Alg.}}} \mathcal{B}$$

die von $\mathcal{E}$ **erzeugte σ-Algebra**. Ist $\mathcal{A} = \sigma(\mathcal{E})$, so nennt man $\mathcal{E}$ einen Erzeuger von $\mathcal{A}$.

Warnung: Der Erzeuger einer σ-Algebra ist nicht eindeutig. Das wird gleich anhand der Borel-σ-Algebra deutlich werde.
In folgendem Beispiel wird eine offensichtliche kleine Beobachtung genutzt: Aus der Definition folgt sofort $\mathcal{E} \subseteq \sigma(\mathcal{E})$, $\sigma(\mathcal{E})$ ist schließlich die kleinste σ-Algebra, die $\mathcal{E}$ enthält. Merkt euch das, so wie in folgendem Beispiel wird das öfters benutzt.

Beispiel 1.2.5. Sei $\Omega \neq \emptyset$ und $\mathcal{E} = \{\{x\} : x \in \Omega\}$. Dann ist

$$\sigma(\mathcal{E}) = \{A \subseteq \Omega : A \text{ abzählbar oder } A^C \text{ abzählbar}\} =: \mathcal{B}.$$

Warum? Es gilt offensichtlich $\sigma(\mathcal{E}) \subseteq \mathcal{B}$ weil $\sigma(\mathcal{E})$ die kleinste σ-Algebra ist, die $\mathcal{E}$ enthält und $\mathcal{B}$ auch eine σ-Algebra ist (siehe Beispiel 1.1.2), die $\mathcal{E}$ enthält. Es gilt aber auch $\mathcal{B} \subseteq \sigma(\mathcal{E})$ weil jede abzählbare Menge als abzählbare Vereinigung von einelementigen Mengen wieder zu $\sigma(\mathcal{E})$ gehört und auch Komplemente abzählbarer Mengen wieder in $\sigma(\mathcal{E})$ enthalten sind (Definition σ-Algebra).

Beispiel 1.2.6. [**Das Beispiel - die Borel-σ-Algebra**] Wir kommen jetzt zu dem mit Abstand wichtigstem Beispiel einer σ-Algebra der Stochastik. Sei $\Omega = \mathbb{R}^d$ und $\mathcal{E} = \{O \subseteq \mathbb{R}^d : O \text{ offen}\}$. Dann heißt $\mathcal{B}(\mathbb{R}^d) := \sigma(\mathcal{E})$ die **Borel-σ-Algebra auf** $\mathbb{R}^d$. Die Mengen in $\mathcal{B}(\mathbb{R}^d)$ heißen **Borelmengen**.

Anhand der Borel-σ-Algebra spielen wir mal mit den neuen Begriffen rum. Wenn ihr folgende Überlegungen verstanden habt, habt ihr einen großen Schritt geschafft!

Ganz wichtige Übung: Die Borel-σ-Algebra hat viele verschiedene Erzeuger, z. B.

$$\begin{aligned}
\mathcal{E}_2 &= \{K \subseteq \mathbb{R}^d \colon K \text{ kompakt}\},\\
\mathcal{E}_3 &= \{Q \subseteq \mathbb{R}^d \colon Q \text{ Quader}\},\\
\mathcal{E}_4 &= \{(a_1, b_1) \times ... \times (a_d, b_d) \colon a_i, b_i \in \mathbb{R}\},\\
\mathcal{E}_5 &= \{(-\infty, b_1] \times ... \times (-\infty, b_d] \colon b_i \in \mathbb{R}\},\\
\mathcal{E}_6 &= \{A \subseteq \mathbb{R}^d \colon A \text{ abgeschlossen}\}.
\end{aligned}$$

Wir müssen folgendes verstehen: Wie zeigt man allgemein für zwei Mengensysteme $\mathcal{E}$, $\mathcal{E}' \subseteq \mathcal{P}(\Omega)$, dass $\sigma(\mathcal{E}) = \sigma(\mathcal{E}')$ gilt? **Trick:** Indem man zeigt, dass

$$\mathcal{E} \subseteq \sigma(\mathcal{E}') \quad \text{sowie} \quad \mathcal{E}' \subseteq \sigma(\mathcal{E}) \tag{1.1}$$

gelten. Warum reicht das? Dazu nutzen wir zwei Eigenschaften, die direkt aus der Definition der erzeugen σ-Algebra folgen:

(i) $\sigma(\sigma(\mathcal{E})) = \sigma(\mathcal{E})$, „Idempotenz“

(ii) $\mathcal{E} \subseteq \mathcal{E}' \Rightarrow \sigma(\mathcal{E}) \subseteq \sigma(\mathcal{E}')$, „Monotonie“

Aus (1.1) und (i), (ii) folgt

$$\sigma(\mathcal{E}) \subseteq \sigma(\sigma(\mathcal{E}')) = \sigma(\mathcal{E}')$$

sowie

$$\sigma(\mathcal{E}') \subseteq \sigma(\sigma(\mathcal{E})) = \sigma(\mathcal{E}),$$

also zusammen $\sigma(\mathcal{E}) = \sigma(\mathcal{E}')$.

Nun zurück zur Borel-σ-Algebra. Wir zeigen mit obigem Trick nur $\sigma(\mathcal{E}_4) = \mathcal{B}(\mathbb{R}^d)$, den Rest macht ihr in den Übungen. Es ist klar, dass $\mathcal{E}_4 \subseteq \sigma(\{O \subseteq \mathbb{R}^d : O \text{ offen}\})$, denn $\mathcal{E}_4$ enthält nur offene Mengen und die von einer Menge von Mengen erzeuge σ-Algebra enthält auch all die Mengen selbst. Umgekehrt existieren für jedes Element x einer offenen Menge $O \subseteq \mathbb{R}^d$ irgendwelche $a_1, ...a_d, b_1, ...b_d \in \mathbb{Q}$ mit $x \in (a_1, b_1) \times ... \times (a_d, b_d) \subseteq O$. Damit gilt für jede offene Menge $O \subseteq \mathbb{R}^d$

$$O = \bigcup_{x \in O} \{x\} \subseteq \bigcup_{\text{abz. viele}} (a_1, b_1) \times ... \times (a_d, b_d) \subseteq O,$$

also

$$O = \bigcup_{\text{abz. viele}} (a_1, b_1) \times ... \times (a_d, b_d) \in \sigma(\mathcal{E}_4).$$

weil abzählbare Vereinigungen von Mengen einer σ-Algebra (hier die offenen Quader) wieder in der σ-Algebra sind. Wir haben jetzt beide Richtungen von (1.1) gezeigt und damit $\sigma(\mathcal{E}_4) = \sigma(\{O \subseteq \mathbb{R}^d : O \text{ offen}\}) = \mathcal{B}(\mathbb{R}^d)$ bewiesen.

Warnung: Die Borel-σ-Algebra ist wahnsinnig groß! Sie enthält alle offenen Mengen, abgeschlosssene Mengen, kompakte Mengen, jegliche Arten von Quadern, alle abzählbare Vereinigungen von solchen und so weiter und so weiter. Es gibt keine Chance die Borel-σ-Algebra explizit hinzuschreiben, wir haben keine Ahnung, aus welchen Mengen die Borel-σ-Algebra wirklich besteht. Im Folgenden wollen wir deshalb, soweit möglich, Maße auf der Borel-σ-Algebra durch einen ihrer Erzeuger untersuchen, also indem wir Maße nur auf offene Mengen, kompakte Mengen oder verschiedene Quader untersuchen. Um zu zeigen, dass man das machen kann, müssen wir etwas arbeiten. Ein Hilfsmittel dafür sind sogenannte Dynkin-Systeme:

Definition 1.2.7. $\mathcal{D} \subseteq \mathcal{P}(\Omega)$ heißt **Dynkin-System**, falls

(i) $\Omega \in \mathcal{D}$

(ii) $A \in \mathcal{D} \Rightarrow A^C \in \mathcal{D}$

(iii) $A_1, A_2, ... \in \mathcal{D}$ **paarweise disjunkt** $\Rightarrow \dot{\bigcup}_{k=1}^{\infty} A_k \in \mathcal{D}$

Im Gegensatz zu einer σ-Algebra ist ein Dynkin-System also nur abgeschlossen bezüglich paarweise disjunkter Vereinigungen. Das erinnert natürlich an die σ-Additivität von Maßen, woraus sich auch das wichtigste (eigentlich auch das einzig relevante) Beispiel eines Dynkin-Systems ergibt.

Beispiel 1.2.8.

- Jede σ-Algebra ist ein Dynkin-System, die Definition einer σ-Algebra fordert mehr.
- Sind μ_1, μ_2 endliche Maße auf einem messbaren Raum $(\Omega, \mathcal{A})$ mit $\mu_1(\Omega) = \mu_2(\Omega)$, so ist $\mathcal{M} = \{A \in \mathcal{A} \colon \mu_1(A) = \mu_2(A)\}$ ein Dynkin-System. Warum?
  (i) Klar, das nehmen wir an.
  (ii) Ist $A \in \mathcal{M}$, so gilt $\mu_1(A^C) = \mu_1(\Omega) - \mu_1(A) = \mu_2(\Omega) - \mu_2(A) = \mu_2(A^C)$ wegen der Rechenregel für Maße und der Annahme an μ_1, μ_2. Damit ist auch $A^C \in \mathcal{M}$. Beachtet, dass hier die Annahme der Endlichkeit wie bei der Stetigkeit der Maße benutzt wird!
  (iii) Seien $A_1, A_2, ... \in \mathcal{M}$ paarweise disjunkt, es gilt also $\mu_1(A_n) = \mu_2(A_n)$ für alle $n \geq 1$. Damit folgt wegen der σ-Additivität von Maßen
  $$\mu_1\Big(\dot{\bigcup_{k=1}^{\infty}} A_k\Big) = \sum_{k=1}^{\infty} \mu_1(A_k) = \sum_{k=1}^{\infty} \mu_2(A_k) = \mu_2\Big(\dot{\bigcup_{k=1}^{\infty}} A_k\Big),$$
  also ist $\dot{\bigcup}_{k=1}^{\infty} A_k \in \mathcal{M}$.

Die Definitionen von σ-Algebra und Dynkin-System sind recht ähnlich. Um zu zeigen, dass ein Dynkin-System sogar eine σ-Algebra ist, nutzt man oft folgenden Proposition:

Proposition 1.2.9. Ein Dynkin-System $\mathcal{D}$ ist eine σ-Algebra genau dann, wenn $\mathcal{D}$ $\cap$-stabil ist (d. h. $A, B \in \mathcal{D} \Rightarrow A \cap B \in \mathcal{D}$).

Beweis. Die eine Richtung ist wegen Lemma 1.1.3 easy. Überlegt mal schnell.

Sei nun $\mathcal{D}$ ein $\cap$-stabiles Dynkin-System. Wir basteln etwas rum, bis wir die Vereinigungseigenschaft gezeigt haben. Das sieht vielleicht lästig aus, ist aber eine gute Möglichkeit, euch an die Rechentricks zu gewöhnen.

(i) Es gilt $A, B \in \mathcal{D},\ B \subseteq A \Rightarrow A \backslash B \in \mathcal{D}$, weil
$$A \backslash B = A \cap B^C \overset{\text{de Morg.}}{=} \underbrace{(\underbrace{A^C \,\dot{\cup}\, B}_{\in \mathcal{D},\text{ weil disj.}})^C}_{\in \mathcal{D},\text{ weil Kompl.}} \in \mathcal{D}.$$

(ii) Beliebige endliche Vereinigungen von Mengen aus $\mathcal{D}$ sind in $\mathcal{D}$: Seien dazu $A, B \in \mathcal{D}$. Da $A \cap B \in \mathcal{D}$ per Annahme gilt, bekommen wir mit (i)
$$A \cup B = \underbrace{A \,\dot{\cup}\, (\underbrace{B \backslash (A \cap B)}_{\in \mathcal{D},\text{ wegen (i)}})}_{\in \mathcal{D},\text{ weil disj.}} \in \mathcal{D}.$$

Per Induktion bekommt man aus der Vereinigung zweier Mengen auch die Vereinigung endlich vieler Mengen.

(iii) Seien $A_1, A_2, ... \in \mathcal{D}$. Definiere

$$B_n = \bigcup_{k=1}^{n} A_k \quad \text{sowie} \quad C_n = B_n \backslash B_{n-1}.$$

Aus (ii) folgt $B_n \in \mathcal{D}$ und dann mit (i) auch $C_n \in \mathcal{D}$. Mit der Definition der Dynkin-Systeme folgt dann

$$\bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty} B_n = \dot{\bigcup_{n=1}^{\infty}} C_n \in \mathcal{D}.$$

Also ist $\mathcal{D}$ abgeschlossen bezüglich abzählbarer Vereinigungen und damit ist $\mathcal{D}$ eine σ-Algebra.

□

Vorlesung 3

Wir haben letztes Mal gezeigt, dass der Schnitt von σ-Algebren wieder eine σ-Algebra ist. Exakt genauso zeigt man, dass auch der Schnitt von Dynkin-Systemen wieder ein Dynkin-System ist. Daraus motiviert definieren wir auch wieder von Teilmengen von $\mathcal{P}(\Omega)$ erzeugte Dynkin-Systeme:

Definition 1.2.10. Für ein Mengensystem $\mathcal{E} \subseteq \mathcal{P}(\Omega)$ heißt

$$d(\mathcal{E}) := \bigcap_{\substack{\mathcal{E} \subseteq \mathcal{D}, \\ \mathcal{D} \text{ Dynk.-S.}}} \mathcal{D}$$

das $\mathcal{E}$ erzeugtes Dynkin-System. Dass $d(\mathcal{E})$ das kleinste Dynkin-System ist das $\mathcal{E}$ enhält, zeigt man genauso wie für σ-Algebren.

Der nächste Satz sieht harmlos aus, ist aber ein enorm mächtiges Werkzeug. Deshalb bekommt er auch den großen Namen „Hauptsatz“. Tiefer werden wir nicht in die Theorie der Dynkin-Systeme eintauchen, nach dem Hauptsatz kommt die für uns relevante Anwendung.

Satz 1.2.11. **[Hauptsatz für Dynkin-Systeme]** Ist $\mathcal{E} \subseteq \mathcal{P}(\Omega)$ $\cap$-stabil, so gilt $d(\mathcal{E}) = \sigma(\mathcal{E})$.

Beweis. „$\subseteq$“: Die Richtung $d(\mathcal{E}) \subseteq \sigma(\mathcal{E})$ folgt sofort, denn jede σ-Algebra ist auch immer ein Dynkin-System, folglich gilt

$$d(\mathcal{E}) = \bigcap_{\substack{\mathcal{E} \subseteq \mathcal{D}, \\ \mathcal{D} \text{ Dynk.-S.}}} \mathcal{D} \subseteq \bigcap_{\substack{\mathcal{E} \subseteq \mathcal{B}, \\ \mathcal{B}\ \sigma\text{-Alg.}}} \mathcal{B} = \sigma(\mathcal{E}).$$

Dabei nutzten wir, dass der Schnitt über mehr Mengen natürlich kleiner ist als der Schnitt über weniger Mengen.

„$\supseteq$“: Für die Richtung $d(\mathcal{E}) \supseteq \sigma(\mathcal{E})$ nehmen wir mal kurz an, dass $d(\mathcal{E})$ $\cap$-stabil wäre. Denn in diesem Fall folgt nach Proposition 1.2.9, dass $d(\mathcal{E})$ eine σ-Algebra ist. Weil aber $\mathcal{E} \subseteq d(\mathcal{E})$ gilt, muss die kleinste σ-Algebra (was gerade $\sigma(\mathcal{E})$ ist) dann Teilmenge von $d(\mathcal{E})$ sein. Das war's.

Wir müssen also nur noch zeigen, dass auch $d(\mathcal{E})$ $\cap$-stabil, wenn $\mathcal{E}$ $\cap$-stabil ist:

(a) Definiere dazu zunächst

$$\mathcal{D}_D = \{\mathcal{A} \in \mathcal{P}(\Omega) : A \cap D \in d(\mathcal{E})\}$$

für ein beliebige $D \in d(\mathcal{E})$. Wir zeigen zunächst, dass $\mathcal{D}_D$ ein Dynkin-System ist:

(i) Weil per Annahme $D \in d(\mathcal{E})$ und $D = \Omega \cap D$ gilt, ist $\Omega \in \mathcal{D}_D$.

(ii) Sei $A \in \mathcal{D}_D$. Damit auch $A^C \in \mathcal{D}_D$ gilt, zeigen wir $A^C \cap D \in d(\mathcal{E})$. Da Dynkin-Systeme abgeschlossen bezüglich disjunkter Vereinigung sind, folgt aus den Regeln des Dynkin-Systems $d(\mathcal{E})$

$$A^C \cap D \overset{\text{de Morg.}}{=} (A \cup D^C)^C = (\underbrace{(A \cap D)}_{\in d(\mathcal{E})} \dot\cup D^C)^C \in d(\mathcal{E}).$$

(iii) Seien $A_1, A_2, ... \in \mathcal{D}_D$ paarweise disjunkt, dann gilt

$$\dot{\bigcup_{k=1}^{\infty}} A_k \cap D = \dot{\bigcup_{k=1}^{\infty}} \underbrace{(A_k \cap D)}_{\in d(\mathcal{E})} \in d(\mathcal{E}).$$

(b) Es gilt $d(\mathcal{E}) \subseteq \mathcal{D}_D$ für alle $D \in \mathcal{E}$. Warum? Sei $E \in \mathcal{E}$, so ist $E \cap D \in \mathcal{E}$, weil $\mathcal{E}$ nach Annahme $\cap$-stabil ist. Damit ist $E \in \mathcal{D}_D$ und folglich $\mathcal{E} \subseteq \mathcal{D}_D$. Dann gilt aber auch $d(\mathcal{E}) \subseteq d(\mathcal{D}_D) \overset{(a)}{=} \mathcal{D}_D$ weil das von einem Dynkin-System $\mathcal{D}$ erzeugte Dynkin-System gerade $\mathcal{D}$ ist.

(c) Weil $\mathcal{E} \subseteq d(\mathcal{E})$ immer gilt, gilt wegen (b) auch $\mathcal{E} \subseteq \mathcal{D}_D$ für alle $D \in d(\mathcal{E})$, d.h. $D \cap E \in d(\mathcal{E})$ für alle $E \in \mathcal{E}$.

(d) Aus (c) und (a) folgt $d(\mathcal{E}) \subseteq d(\mathcal{D}_D) = \mathcal{D}_D$ für alle $D \in d(\mathcal{E})$. Das ist aufgrund der Definition von $\mathcal{D}_D$ aber gerade die $\cap$-Stabilität von $d(\mathcal{E})$.

□

Klar, aus dem Beweis nimmt man nicht so richtig viel Verständniss mit. Aber bevor ihr den Beweis einfach weglasst: Nur durch das Durcharbeiten von solch komischen Argumenten bekommt ihr in Kombination mit den Übungsaufgaben „Rechenroutine“ mit den relevanten Mengenoperationen.

Nun kommen wir zu der wesentlichen Anwendung von Dynkin-Systemen, nur deshalb sprechen wir überhaupt über Dynkin-Systeme! Mit Dynkin-Systemen können wir ganz einfach zeigen, dass die Gleichheit von Maßen schon aus der Gleichheit der Maße auf einem $\cap$-stabilen Erzeuger folgt. Wenn wir zum Beispiel an die wahnsinnig große Borel-σ-Algebra $\mathcal{B}(\mathbb{R})$ denken, macht all das schnell Sinn. Um die Gleichheit von zwei Maßen auf der ganzen Borel-σ-Algebra zu zeigen reicht es, die Gleichheit auf einem der vielen Erzeuger zu checken.

Satz 1.2.12. Es sei $(\Omega, \mathcal{A})$ ein messbarer Raum und $\mathcal{E}$ ein $\cap$-stabiler Erzeuger von $\mathcal{A}$. Sind μ_1, μ_2 endliche Maße auf $\mathcal{A}$ und es gelten

- $\mu_1(\Omega) = \mu_2(\Omega)$,
- $\mu_1(A) = \mu_2(A)$ für alle $A \in \mathcal{E}$,

so gilt auch $\mu_1(A) = \mu_2(A)$ für alle $A \in \mathcal{A}$, d. h. $\mu_1 = \mu_2$.

Beweis. Wir nutzten zum ersten Mal den sogenannten **Trick der guten Mengen**. Dazu schreiben wir die Menge $\mathcal{M}$ der Mengen hin, für die die Aussage gelten soll. Das sind die guten Mengen. Das Ziel ist also, $\mathcal{M} = \mathcal{A}$ zu zeigen. Hierfür nutzen wir einen Dynkin-System Trick, den wir noch mehrmals sehen werden. Gezeigt haben wir schon, dass

$$\mathcal{M} = \{A \in \mathcal{A} \colon \mu_1(A) = \mu_2(A)\}$$

ein Dynkin-System ist. Nach Annahme ist $\mathcal{E} \subseteq \mathcal{M}$. Weil $\mathcal{M}$ ein Dynkin-System ist, gilt $d(\mathcal{E}) \subseteq d(\mathcal{M}) = \mathcal{M} \subseteq \mathcal{A}$. Weil nach Annahme $\mathcal{E}$ $\cap$-stabil ist, gilt nach dem Hauptsatz über Dynkin-Systeme $\sigma(\mathcal{E}) = d(\mathcal{E})$. Nach Annahme ist aber $\sigma(\mathcal{E}) = \mathcal{A}$. Alles zusammen ergibt

$$\mathcal{A} = \sigma(\mathcal{E}) = d(\mathcal{E}) \subseteq d(\mathcal{M}) = \mathcal{M} \subseteq \mathcal{A}.$$

Weil rechts und links das selbe steht, müssen alle Teilmengenrelationen sogar Gleichheiten sein. Damit gilt $\mathcal{M} = \mathcal{A}$ und das ist die Aussage des Satzes. □

Wir schauen uns noch einen Trick an, die Endlichkeitsannahme aus Satz 1.2.12 abzuschwächen.

Satz 1.2.13. **[Eindeutigkeitssatz]** Es sei $(\Omega, \mathcal{A})$ ein messbarer Raum, $\mathcal{E}$ ein $\cap$-stabiler Erzeuger von $\mathcal{A}$ und μ_1, μ_2 seien Maße auf $\mathcal{A}$. Zudem gelten:

(i) Es gibt eine Folge $(E_n) \subseteq \mathcal{E}$ mit $E_n \uparrow \Omega$, $n \to \infty$, und $\mu_i(E_n) < \infty$ für alle $n \in \mathbb{N}$, $i = 1, 2$.

(ii) $\mu_1(A) = \mu_2(A)$ für alle $A \in \mathcal{E}$.

Dann gilt $\mu_1 = \mu_2$, d. h. $\mu_1(A) = \mu_2(A)$ für alle $A \in \mathcal{A}$.

Geben wir der genutzten Erweiterung endlicher Maße einen Namen:

Definition 1.2.14. Ist $(\Omega, \mathcal{A}, \mu)$ ein Maßraum und es gibt eine Folge $(E_n) \subseteq \mathcal{A}$ mit $E_n \uparrow \Omega$, $n \to \infty$, und $\mu(E_n) < \infty$ für alle $n \in \mathbb{N}$, so nennt man μ ein **σ-endliches Maß**.

Die meisten Sätze für endliche Maße lassen sich mit dem Trick des folgenden Beweises auf σ-endliche Maße ausdehen. Beispiele folgen noch.

Beweis von Satz 1.2.13. Definiere dazu für $A \in \mathcal{A}$ und $n \in \mathbb{N}$

$$\begin{aligned} \mu_1^n(A) &:= \mu_1(A \cap E_n), \\ \mu_2^n(A) &:= \mu_2(A \cap E_n). \end{aligned}$$

Man rechnet sofort nach, dass auch die μ_i^n wieder Maße auf $\mathcal{A}$ sind. Des Weiteren sind μ_1^n, μ_2^n endlich, weil $\mu_i^n(\Omega) = \mu_i(\Omega \cap E_n) = \mu_i(E_n) \overset{\text{Ann.}}{<} \infty$. Nach Satz 1.2.12 gilt $\mu_1^n = \mu_2^n$ für alle $n \in \mathbb{N}$. Nun gilt wegen Stetigkeit von Maßen

$$\begin{aligned} \mu_1(A) = \mu_1(A \cap \Omega) = \mu_1\Big(A \cap \bigcup_{n=1}^{\infty} E_n\Big) = \mu_1\Big(\bigcup_{n=1}^{\infty} (A \cap E_n)\Big) &\overset{1.1.14}{=} \lim_{n\to\infty} \mu_1(A \cap E_n) \\ \overset{\text{Def.}}{=} \lim_{n\to\infty} \mu_1^n(A) = \lim_{n\to\infty} \mu_2^n(A) \overset{\text{Def.}}{=} \lim_{n\to\infty} \mu_2(A \cap E_n) &\overset{1.1.14}{=} \mu_2\Big(A \cap \bigcup_{n=1}^{\infty} E_n\Big) = \mu_2(A). \end{aligned}$$

□

So, nun endlich ein richtig konkretes Beispiel!

Beispiel 1.2.15. Sei $\mathbb{P}$ ein Wahrscheinlichkeitsmaß auf $\mathcal{B}(\mathbb{R})$. Dann heißt

$$F_{\mathbb{P}}(t) := \mathbb{P}((-\infty, t]), \quad t \in \mathbb{R},$$

die **Verteilungsfunktion** von $\mathbb{P}$. $F_{\mathbb{P}}$ erfüllt folgende Eigenschaften:

- $0 \leq F_{\mathbb{P}} \leq 1$,
- $F_{\mathbb{P}}$ ist nicht fallend,
- $\lim_{t\to+\infty} F_{\mathbb{P}}(t) = 1$,
- $\lim_{t\to-\infty} F_{\mathbb{P}}(t) = 0$.

Die ersten beiden Eigenschaften folgen aus der Definition von Wahrscheinlichkeitsmaßen und der Monotonie von Maßen. Die weiteren Eigenschaften folgen aus der Stetigkeit von Maßen, damit dürft ihr euch in den Übungen auseinandersetzen. Um die gerade bewiesenen Sätze anzuwenden, zeigen wir folgende Behauptung:

$$F_{\mathbb{P}_1}(t) = F_{\mathbb{P}_2}(t) \quad \text{für alle } t \in \mathbb{R} \quad \Longrightarrow \quad \mathbb{P}_1 = \mathbb{P}_2.$$

In Worten: Wahrscheinlichkeitsmaße auf $\mathcal{B}(\mathbb{R})$ sind durch ihre Verteilungsfunktion eindeutig festgelegt. Die Behauptung folgt aus Satz 1.2.12 mit $\mathcal{E} = \{(-\infty, t] : t \in \mathbb{R}\} \subseteq \mathcal{P}(\mathbb{R})$. Checken wir dazu die benötigten Eigenschaften:

- $\sigma(\mathcal{E}) = \mathcal{B}(\mathbb{R})$ ist aus den Übungen bekannt,
- $\mathcal{E}$ ist $\cap$-stabil, denn $(-\infty, s] \cap (-\infty, t] = (-\infty, \min\{s, t\}]$ für alle $s, t \in \mathbb{R}$,
- $\mathbb{P}_1(A) = \mathbb{P}_2(A)$ für alle $A \in \mathcal{E}$ weil das gerade die Gleichheit der Verteilungsfunktionen ist.

Genauso beweist man auch die Aussage des nächsten Beispiels:

Beispiel 1.2.16. Seien μ_1, μ_2 σ-endliche Maße auf $\mathcal{B}(\mathbb{R}^d)$ mit einer der folgenden Eigenschaften:

$$\begin{aligned}
&\mu_1(Q) = \mu_2(Q) \text{ für alle Quader } Q,\\
&\mu_1(K) = \mu_2(K) \text{ für alle kompakten Mengen } K,\\
&\mu_1(O) = \mu_2(O) \text{ für alle offenen Mengen } O,\\
&\mu_1(A) = \mu_2(A) \text{ für alle abgeschlossenen Mengen } A.
\end{aligned}$$

Dann gilt $\mu_1 = \mu_2$, die Maße stimmen also auf allen Borelmengen überein.

1.3 Konstruktion von Maßen

Gerade haben wir gesehen, dass endliche Maße auf $\mathcal{B}(\mathbb{R})$ schon auf den Intervallen (also durch die Verteilungsfunktionen) eindeutlich festgelegt sind. Das ist eine Eindeutigkeitsaussage. Jetzt drehen wir das Ganze um und untersuchen Existenzaussagen. Am Ende soll folgendes rauskommen: Wenn wir eine geeignete Mengenfunktion (also eine Funktion auf Mengen, so wie ein Maß) nur auf den Intervallen definieren, dann gibt es auch ein passendes Maß auf $\mathcal{B}(\mathbb{R})$. Dazu brauchen wir einiges an Handwerkszeugs.

Definition 1.3.1. $\mathcal{S} \subseteq \mathcal{P}(\Omega)$ heißt **Semiring**, falls

(i) $\emptyset \in \mathcal{S}$

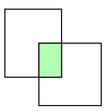
(ii) $A, B \in \mathcal{S} \Rightarrow A \cap B \in \mathcal{S}$, also ist $\mathcal{S}$ „$\cap$-stabil“

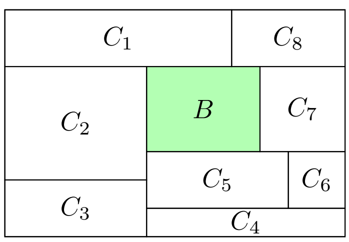
(iii) $A, B \in \mathcal{S} \Rightarrow$ es gibt paarweise disjunkte Mengen $C_1, ..., C_m \in \mathcal{S}$ mit $A \backslash B = \dot{\bigcup}_{k=1}^{m} C_k$.

Die Definition ist etwas komisch, verallgemeinert aber einfach nur das folgendes Beispiel, dass wir immer im Kopf halten und später auch hauptsächlich nutzen:

Beispiel. $\mathcal{Q} := \{Q \subseteq \mathbb{R}^2 : Q \text{ Quader}\}$ ist ein Semiring. Checken wir anschaulich die definierenden Eigenschaften:

(i) $\emptyset \in \mathcal{Q}$ weil $\emptyset = (0, 0) \times (0, 0)$

(ii) 

(iii) 

Natürlich kann man das formell sauber hinschreiben, das gibt uns aber keinen Verständnis Mehrwehrt.

Aus der zweiten Eigenschaft eines Semirings kann man durch Induktion sofort Abgeschlossenheit bezüglich endlich vielen Schnitten zeigen. Probiert's einfach aus, ist zu einfach als Übungsaufgabe. Auch die letzte Eigenschaft kann man verallgemeinern. Bei Quadern ist das anschaulich klar: Wenn man aus einem Quader endlich viele Quader entfernt, bleibt eine disjunkte Vereinigung von Quadern übrig. Mit Induktion kriegen wir das auch für allgemeine Semiringe hin:

Lemma 1.3.2. **[Eine kleine Indexschlacht]** Es gilt in einem Semiring auch (iii)$'$: Sind $B_1, ..., B_r, A \in \mathcal{S}$ mit $B_1, B_2, ..., B_r$ paarweise disjunkt, so existieren $C_1, ..., C_n \in \mathcal{S}$ paarweise disjunkt mit

$$A \backslash (B_1 \mathbin{\dot\cup} ... \mathbin{\dot\cup} B_r) = \dot{\bigcup_{k=1}^{n}} C_k.$$

Beweis. Vollständige Induktion bezüglich r.

IA: Für $A \backslash B_1$ folgt das direkt aus der Definition des Seminrings.

IV: Die Behauptung gelte für ein beliebiges aber festes $r \in \mathbb{N}$.

IS: Seien nun $B_1, B_2, ..., B_{r+1} \in \mathcal{S}$ paarweise disjunkt. Nach Induktionsvoraussetzung und Rechenregeln mit Schnitten von Mengen gilt mit Umklammern von Schnitten und Vereinigungen

$$\begin{aligned} A \backslash \dot{\bigcup_{i=1}^{r+1}} B_i &= A \cap \bigcap_{i=1}^{r+1} B_i^C \\ &= \Big(A \backslash \dot{\bigcup_{i=1}^{r}} B_i\Big) \cap B_{r+1}^C \\ &\overset{\text{IV}}{=} \Big(\dot{\bigcup_{k=1}^{m_r}} C_{r,k}\Big) \cap B_{r+1}^C = \dot{\bigcup_{k=1}^{m_r}} [C_{r,k} \backslash B_{r+1}]. \end{aligned}$$

Nun existieren für alle $1 \leq k \leq m_r$ jeweils endlich viele paarweise disjunkte Mengen $C_{r,k,1}, ..., C_{r,k,l_{r,k}} \in \mathcal{S}$, mit

$$C_{r,k} \backslash B_{r+1} = \dot{\bigcup_{m=1}^{l_{r,k}}} C_{r,k,m}.$$

Also gilt

$$A \backslash \bigcup_{i=1}^{r+1} B_i = \dot{\bigcup_{k=1}^{m_r}} \dot{\bigcup_{m=1}^{l_{r,k}}} C_{r,k,m}.$$

Da die endlich vielen Mengen $C_{r,k,m}$ paarweise disjunkt sind, haben wir die gewünschte Darstellung für $A \backslash (B_1 \mathbin{\dot\cup} ... \mathbin{\dot\cup} B_{r+1})$ gefunden.

$\square$

Vorlesung 4

Es kommen jetzt ein paar schmerzhafte Vorbereitungen für den wichtigsten Satz dieses Abschnitts, den Fortsetztungssatz von Carathéodory. Wir werden zunächst zeigen, dass die Monotonie und Subadditivität (siehe Lemma 1.1.9) auch für „Maße" auf Semiringen gilt. Warum steht hier Maß in Anführungsstrichen? Per Definition ist ein Maß immer auf einer σ-Algebra definiert, der Begriff macht also auf Semiringen gar keinen Sinn. Die genaue Aussage ist also, dass eine Mengenfunktion auf Semiringen mit Eigenschaften die einem Maß ähneln, ebenfalls Monotonie und Subadditivität erfüllen.

Lemma 1.3.3. **[Eine ziemlich große Indexschlacht]** Sei $\mathcal{S}$ ein Semiring und $\mu: \mathcal{S} \to [0, \infty]$ eine Mengenfunktion mit

- $\mu(\emptyset) = 0$

- μ ist **σ-additiv** (d.h. sind $A_1, A_2, ... \in S$ paarweise disjunkt mit $A := \dot\bigcup_{k=1}^{\infty} A_k \in \mathcal{S}$, so gilt $\mu(A) = \sum_{k=1}^{\infty} \mu(A_k)$).

Dann gilt

(i) Monotonie: $\mu(A) \leq \mu(B)$ für alle $A, B \in \mathcal{S}$ mit $A \subseteq B$.

(ii) „**Subadditivität**“: Sind $A, A_1, A_2, ... \in \mathcal{S}$ und $A \subseteq \bigcup\limits_{k=1}^{\infty} A_k$, so gilt $\mu(A) \leq \sum\limits_{k=1}^{\infty} \mu(A_k)$.

Man beachte, dass die Eigenschaften der σ-Additivität etwas komisch sind. Da wir nicht fordern, dass Semiringe abgeschlossen bezüglich Vereinigungen sind (sie sind es auch meistens nicht, man denke nur an den Semiring der Quader), muss immer gefordert werden, dass die Vereinigungen wieder in $\mathcal{S}$ liegen. Sonst wäre $\mu(A)$ schließlich gar nicht definiert! Auch zu beachten ist, dass Vereinigungen und Komplemente nicht automatisch in $\mathcal{S}$ liegen. Daher sind einfache Eigenschaften für Maße auf σ-Algebren, nicht so einfach für Mengenfunktionen auf Semiringen.

Beweis.

(i) Es gibt wegen der Eigenschaften eines Semirings Mengen $C_1, ..., C_m \in \mathcal{S}$ mit

$$B \backslash A = \dot\bigcup_{k=1}^{m} C_k.$$

Damit gilt wegen der geforderten Additivität von μ

$$\mu(B) = \mu(A \,\dot\cup\, (B\backslash A)) = \mu(A \,\dot\cup\, C_1 \,\dot\cup\, ... \,\dot\cup\, C_m) = \mu(A) + \mu(C_1) + ... + \mu(C_m) \geq \mu(A).$$

(ii) Erst machen wir (wie immer wieder!) die A_n disjunkt:

$$\begin{aligned} A_1' &:= A_1, \\ A_2' &:= A_2 \backslash A_1', \\ A_n' &:= A_n \backslash (A_1' \,\dot\cup\, ... \,\dot\cup\, A_{n-1}'), \quad n \geq 3. \end{aligned}$$

Beachte: Die A_n' müssen nicht in in $\mathcal{S}$ sein. Weil die A_n' die Form $A_n \backslash ...$ haben, gibt es wegen Lemma 1.3.2 allerdings paarweise disjunkte $C_{n,j} \in \mathcal{S}$ mit

$$A_n' = \dot\bigcup_{j=1}^{l_n} C_{n,j} \tag{1.2}$$

für alle $n \in \mathbb{N}$. Darum gibt es wegen Lemma 1.3.2 Mengen $D_{n,k} \in \mathcal{S}$ mit

$$A_n \backslash A_n' = \dot\bigcup_{k=1}^{m_n} D_{n,k}. \tag{1.3}$$

Damit gelten

- $$A = \dot\bigcup_{n=1}^{\infty} A_n' \cap A = \dot\bigcup_{n=1}^{\infty} \dot\bigcup_{j=1}^{l_n} A \cap C_{n,j}$$

  weil $A \subseteq \bigcup\limits_{n=1}^{\infty} A_n = \dot\bigcup\limits_{n=1}^{\infty} A_n'$ angenommen wurde,

- $$A_n = A_n' \,\dot\cup\, (A_n \backslash A_n') \overset{(1.2)}{\underset{(1.3)}{=}} \dot\bigcup_{j=1}^{l_n} C_{n,j} \,\dot\cup\, \dot\bigcup_{n=1}^{m_n} D_{n,k}.$$

Alles zusammen ergibt die Subadditivität:

$$
\begin{aligned}
\mu(A) &= \mu\Big(\dot{\bigcup_{n=1}^{\infty}} \dot{\bigcup_{j=1}^{l_n}} \underbrace{A \cap C_{n,j}}_{\substack{\in \mathcal{S}, \text{ weil } \mathcal{S} \\ \cap\text{-stabil}}}\Big) \\
&\overset{\sigma\text{-add.}}{=} \sum_{n=1}^{\infty} \sum_{j=1}^{l_n} \mu(A \cap C_{n,j}) \\
&\overset{\text{Monotonie}}{\leq} \sum_{n=1}^{\infty} \sum_{j=1}^{l_n} \mu(C_{n,j}) \\
&\overset{\mu \geq 0}{\leq} \sum_{n=1}^{\infty} \Big(\sum_{j=1}^{l_n} \mu(C_{n,j}) + \sum_{k=1}^{m_n} \mu(D_{n,k}) \Big) \\
&\overset{\sigma\text{-add.}}{=} \sum_{n=1}^{\infty} \mu\Big(\dot{\bigcup_{n=1}^{l_n}} C_{n,j} \,\dot{\cup}\, \dot{\bigcup_{j=1}^{m_n}} D_{n,k} \Big) \\
&= \sum_{n=1}^{\infty} \mu(A_n).
\end{aligned}
$$

□

Definition 1.3.4. $\mu^* : \mathcal{P}(\Omega) \to [0, \infty]$ heißt **äußeres Maß**, falls

(i) $\mu^*(\emptyset) = 0$

(ii) $A \subseteq B \subseteq \Omega \Rightarrow \mu^*(A) \leq \mu^*(B)$

(iii) $A_1, A_2, ... \subseteq \Omega \Rightarrow \mu^*\Big(\bigcup_{k=1}^{\infty} A_k\Big) \leq \sum_{k=1}^{\infty} \mu^*(A_k)$

Für den Moment bleibt es unklar, weshalb wir dieser abstrakten Definition den Namen „äußeres Maß" geben. Das wird aber in dem Beweis des Carathéodory Fortsetzungssatzes klar werden. Das dort definierte äußere Maß hat eine klare Interterpretation.

Definition 1.3.5. Sei μ^* ein äußeres Maß auf Ω. Dann heißt $A \subseteq \Omega$ **μ^*-messbare Menge**, falls für alle $Z \subseteq \Omega$

$$\mu^*(Z) = \mu^*(Z \cap A) + \mu^*(Z \cap A^C)$$

gilt. Die Menge der μ^*-messbaren Mengen heißt $\mathcal{A}_{\mu^*}$.

Proposition 1.3.6.

(i) $\mathcal{A}_{\mu^*}$ ist eine σ-Algebra.

(ii) μ^* eingeschränkt auf $\mathcal{A}_{\mu^*}$ ist ein Maß.

Beweis. Wir zeigen nacheinander (a) $\mathcal{A}_{\mu^*}$ ist eine σ-Algebra und (b) μ^* ist ein Maß auf $\mathcal{A}_{\mu^*}$.
(a) Zu zeigen sind die definierenden Eigenschaften einer σ-Algebra:

(i) Für $Z \subseteq \Omega$ gilt

$$\mu^*(Z) = \mu^*(Z) + 0 = \mu^*(Z \cap \Omega) + \mu^*(Z \cap \underbrace{\Omega^C}_{=\emptyset}).$$

Damit ist $\Omega \in \mathcal{A}_{\mu^*}$ gezeigt.

(ii) Sei $A \in \mathcal{A}_{\mu^*}$, dann erfüllt (+ kommutieren und $(A^C)^C = A$ nutzen) auch A^C die definierende Eigenschaft von $\mathcal{A}_{\mu^*}$. Also ist $\mathcal{A}_{\mu^*}$ abgeschlossen unter Komplementbildung.

(iii) Wir zeigen zunächst die Abgeschlossenheit bezüglich Vereinigungen von zwei Mengen. Seien also $A_1, A_2 \in \mathcal{A}_{\mu^*}$. Sei $Z \subseteq \Omega$ beliebig, dann folgt mit $Z' := Z \cap (A_1 \cup A_2)$

$$\begin{aligned}\mu^*(Z \cap (A_1 \cup A_2)) &\overset{A_1 \in \mathcal{A}_{\mu^*}}{=} \mu^*(Z \cap (A_1 \cup A_2) \cap A_1) + \mu^*(Z \cap (A_1 \cup A_2) \cap A_1^C) \\ &= \mu^*(Z \cap A_1) + \mu^*(Z \cap A_2 \cap A_1^C).\end{aligned}$$

Folglich gilt auch

$$\begin{aligned}&\mu^*(Z \cap (A_1 \cup A_2)) + \mu^*(Z \cap (A_1 \cup A_2)^C) \\ &= \mu^*(Z \cap A_1) + \mu^*(Z \cap A_2 \cap A_1^C) + \mu^*(Z \cap (A_1 \cup A_2)^C) \\ &= \mu^*(Z \cap A_1) + \mu^*(\underbrace{Z \cap A_1^C}_{=:Z''} \cap A_2) + \mu^*(\underbrace{Z \cap A_1^C}_{=:Z''} \cap A_2^C) \\ &\overset{A_2 \in \mathcal{A}_{\mu^*}}{=} \mu^*(Z \cap A_1) + \mu^*(Z \cap A_1^C) \\ &\overset{A_1 \in \mathcal{A}_{\mu^*}}{=} \mu^*(Z)\end{aligned}$$

und damit ist $A_1 \cup A_2 \in \mathcal{A}_{\mu^*}$.

(iv) Per Induktion folgt aus (iii) die Abgeschlossenheit bezüglich Vereinigungen endlich vieler Mengen.

(v) Es fehlt jetzt noch die Abgeschlossenheit bezüglich abzählbar unendlicher Vereinigungen. Seien also $A_1, A_2, ... \in \mathcal{A}_{\mu^*}$, zu zeigen ist $\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}_{\mu^*}$. Zuerst nutzen wir den schon bekannten Trick, der uns erlaubt, ohne Einschränkung der Allgemeinheit anzunehmen, dass die Mengen diskjunkt sind. Dazu definieren wir die paarweise disjunkten Mengen

$$A_1' = A_1, \quad A_2' = A_2 \backslash A_1', \quad \text{und} \quad A_n' = A_n \backslash (A_1' \,\dot\cup\, ... \,\dot\cup\, A_{n-1}'), \quad n \geq 3,$$

und beachten, dass damit $\bigcup_{n=1}^{\infty} A_n = \dot\bigcup_{n=1}^{\infty} A_n'$ gilt. Wenn also die Vereinigung der disjunkten A_n' wieder in $\mathcal{A}_{\mu^*}$ ist, ist auch die Vereinigung über die A_n in $\mathcal{A}_{\mu^*}$. Es reicht also die Aussage für disjunkte Mengen zu beweisen. Damit die Rechnungen lesbarer bleiben, nehmen wir also ohne Beschränkung der Allgemeinheit an, dass die Mengen $A_1, A_2, ...$ paarweise disjunkt sind, so sparen wir uns die $'$ in den Gleichungen. Aufgrund der Definition von $\mathcal{A}_{\mu^*}$ wählen wir ein $Z \subseteq \Omega$ beliebig. Wir zeigen erstmal induktiv

$$\mu^*(Z) = \sum_{k=1}^{n} \mu^*(Z \cap A_k) + \mu^*\Big(Z \cap \bigcap_{k=1}^{n} A_k^C\Big), \quad \forall n \in \mathbb{N}. \tag{1.4}$$

IA: Für $n = 1$ gilt die Behauptung, weil $A_1 \in \mathcal{A}_{\mu^*}$.

IV: Es gelte (1.4) für ein beliebiges, aber festes $n \in \mathbb{N}$.

IS: Eine kleine Runde Kampfrechnen mit Mengen. Weil nach Annahme $A_{n+1} \in \mathcal{A}_{\mu^*}$ und die A_n paarweise disjunkt sind, gilt

$$\begin{aligned}\mu^*\Big(Z \cap \bigcap_{k=1}^{n} A_k^C\Big) &= \mu^*\Big(Z \cap \bigcap_{k=1}^{n} A_k^C \cap A_{n+1}\Big) + \mu^*\Big(Z \cap \bigcap_{k=1}^{n} A_k^C \cap A_{n+1}^C\Big) \\ &= \mu^*(Z \cap A_{n+1}) + \mu^*\Big(Z \cap \bigcap_{k=1}^{n+1} A_k^C\Big).\end{aligned}$$

Einsetzen in die Induktionsvoraussetzung gibt

$$\mu^*(Z) = \sum_{k=1}^{n+1} \mu^*(Z \cap A_k) + \mu^*\Big(Z \cap \bigcap_{k=1}^{n+1} A_k^C\Big).$$

Damit ist der Induktionsschritt gezeigt.

Zurück zur Vereinigung: Wegen der angenommenen Monotonie von μ^* folgt aus (1.4)

$$\mu^*(Z) \geq \sum_{k=1}^{n} \mu^*(Z \cap A_k) + \mu^*\Big(Z \cap \bigcap_{k=1}^{\infty} A_k^C\Big), \quad \forall n \in \mathbb{N}.$$

Mit $n \to \infty$ folgt (Monotonie von Grenzwertbildung aus Analysis 1)

$$\begin{aligned}\mu^*(Z) &\geq \sum_{k=1}^{\infty} \mu^*(Z \cap A_k) + \mu^*\Big(Z \cap \bigcap_{k=1}^{\infty} A_k^C\Big)\\ &\geq \mu^*\Big(\bigcup_{k=1}^{\infty} A_k \cap Z\Big) + \mu^*\Big(Z \cap \Big(\bigcup_{k=1}^{\infty} A_k\Big)^C\Big)\\ &\geq \mu^*\Big(\Big(Z \cap \bigcup_{k=1}^{\infty} A_k\Big) \cup \Big(Z \cap \Big(\bigcup_{k=1}^{\infty} A_k\Big)^C\Big)\Big)\\ &= \mu^*\Big(Z \cap \underbrace{\Big(\bigcup_{k=1}^{\infty} A_k \cup \Big(\bigcup_{k=1}^{\infty} A_k\Big)^C\Big)}_{\Omega}\Big) = \mu^*(Z).\end{aligned} \tag{1.5}$$

Für die letzten beiden Ungleichungen haben wir die angenommene Subadditivität (Ungleichung andersrum als üblich) genutzt. Weil die linke und rechte Seite der Kette von Ungleichungen identisch sind, sind die Ungleichungen alles Gleichungen, also gilt

$$\mu^*(Z) = \mu^*\Big(Z \cap \bigcup_{k=1}^{\infty} A_k\Big) + \mu^*\Big(Z \cap \Big(\bigcup_{k=1}^{\infty} A_k\Big)^C\Big).$$

Weil Z beliebig war, ist damit

$$\bigcup_{k=1}^{\infty} A_k \in \mathcal{A}_{\mu^*}$$

aufgrund der Definition von $\mathcal{A}_{\mu^*}$. Damit ist die Abgeschlossenheit bezüglich abzählbarer Vereinigungen gezeigt und folglich ist $\mathcal{A}_{\mu^*}$ eine σ-Algebra.

(b) Aufgrund der Gleichheiten in (1.5) gilt auch

$$\mu^*(Z) = \sum_{k=1}^{\infty} \mu^*(Z \cap A_k) + \mu^*\Big(Z \cap \bigcap_{k=1}^{\infty} A_k^C\Big).$$

Wählen wir $Z = \bigcup\limits_{k=1}^{\infty} A_k$, so gilt wegen $\mu(\emptyset) = 0$

$$\mu^*\Big(\bigcup_{k=1}^{\infty} A_k\Big) = \sum_{k=1}^{\infty} \mu^*(A_k) + 0$$

und das ist gerade die σ-Additivität von μ^*. Also ist μ^* auch ein Maß auf der σ-Algebra $\mathcal{A}_{\mu^*}$. □

Vorlesung 5

Kommen wir endlich zum Höhepunkt der ersten Wochen. Nach dem Beweis sind wir auch endlich mitten in der Stochastik angelangt! Der Beweis enthält tatsächlich viele Informationen. Insbesondere wird klar, wo der Begriff äußeres Maß herkommt.

Satz 1.3.7. **[Fortsetzungssatz von Carathéodory]** Sei $\mathcal{S}$ ein Semiring und $\mu\colon \mathcal{S} \to [0, \infty]$ eine Mengenfunktion mit

- $\mu(\emptyset) = 0$,
- μ ist **σ-additiv** (d.h. sind $A_1, A_2, ... \in S$ paarweise disjunkt mit $A := \bigcup_{k=1}^{\infty} A_k \in \mathcal{S}$, so gilt $\mu(A) = \sum_{k=1}^{\infty} \mu(A_k)$).

Dann existiert ein Maß $\bar{\mu}$ auf $\sigma(\mathcal{S})$ mit $\mu(A) = \bar{\mu}(A)$ für alle $A \in \mathcal{S}$.

Man sagt, dass die Mengenfunktion μ von $\mathcal{S}$ nach $\sigma(\mathcal{S})$ „fortgesetzt“ wird.

Beweis. Für $A \in \mathcal{P}(\Omega)$ definieren wir

$$\mu^*(A) = \inf\Big\{ \sum_{k=1}^{\infty} \mu(A_k) \colon A_1, A_2, ... \in \mathcal{S} \text{ mit } A \subseteq \bigcup_{k=1}^{\infty} A_k \Big\}.$$

Beachte: Weil per Definition (Analysis 1) $\inf \emptyset = +\infty$ gilt, ist $\mu * (A)$ auch definiert, wenn man A nicht durch Mengen aus $\mathcal{S}$ überdecken kann.

Wir zeigen nun nacheinander (a) μ^* ist ein äußeres Maß, (b) $\mathcal{S} \subseteq \mathcal{A}_{\mu^*}$ und (c) $\mu^*(A) = \mu(A)$ für alle $A \in \mathcal{S}$. Der Beweis ist dann vollendet, weil nach dem vorherigen Satz $\mathcal{A}_{\mu^*}$ eine σ-Algebra ist und μ^* ein Maß auf $\mathcal{A}_{\mu^*}$ ist. Wegen (b) gilt $\sigma(S) \subseteq \mathcal{A}_{\mu^*}$, weil die kleinste σ-Algebra die $\mathcal{S}$ enthält, auch Teilmenge von allen σ-Algebras ist, die $\mathcal{S}$ enthalten. Damit ist auch die Einschränkung von μ^* auf $\sigma(S)$ ein Maß (wir setzen dann $\bar{\mu} := \mu^*_{|\sigma(\mathcal{S})}$) und wegen (c) ist $\bar{\mu}$ eine Fortsetzung von μ.

(a) Wir checken die definierenden Eigenschaften eines äußeren Maßes:

(i) $\mu^*(\emptyset) = 0$ ist klar, weil $\emptyset \in \mathcal{S}$ und $\mu(\emptyset) = 0$.

(ii) Monotonie folgt direkt aus der Definition.

(iii) Nun zur Subadditivität. Seien dazu $A_1, A_2, ... \in \mathcal{P}(\Omega)$ und $A = \bigcup_{k=1}^{\infty} A_k$. Wir können annehmen, dass $\mu^*(A_k) < \infty$ für alle $k \in \mathbb{N}$ gilt (sonst gilt die Ungleichung sowieso). Sei nun $\varepsilon > 0$ beliebig. Für jedes $k \in \mathbb{N}$ existiert qua Definition (Infimum ist die größte untere Schranke) eine Folge von Mengen $A_{k,1}, A_{k,2}, ... \in \mathcal{S}$ mit

$$A_k \subseteq \bigcup_{j=1}^{\infty} A_{k,j} \quad \text{und} \quad \sum_{j=1}^{\infty} \mu(A_{k,j}) \leq \mu^*(A_k) + \frac{\varepsilon}{2^k}.$$

Wem das nicht klar ist, der schaue bitte in den Analysis 1 Mitschrieb! Weil

$$A \overset{\text{Def.}}{=} \bigcup_{k=1}^{\infty} A_k \subseteq \bigcup_{k=1}^{\infty} \bigcup_{j=1}^{\infty} A_{k,j}$$

gilt, folgt (das Infimum einer Menge ist kleiner gleich jedem Element der Menge)

$$\mu^*(A) \underset{\text{als inf}}{\overset{\text{Def. } \mu^*}{\leq}} \sum_{k=1}^{\infty} \sum_{j=1}^{\infty} \mu(A_{k,j}) \leq \sum_{k=1}^{\infty} \Big(\mu^*(A_k) + \frac{\varepsilon}{2^k} \Big) = \sum_{k=1}^{\infty} \mu^*(A_k) + \underbrace{\sum_{k=1}^{\infty} \frac{\varepsilon}{2^k}}_{\varepsilon}.$$

Weil ε beliebig gewählt wurde, gilt damit die Subadditivität von μ^*. An dieser Stelle eine kleine Anmerkung: Eine Überdeckung durch eine abzählbare Vereinigung von abzählbaren Vereinigungen ist gleichbedeutend zu nur einer abzählbaren Vereinigung. Das Stichwort ist Cantors Diagonalverfahren, siehe Analysis 1 (oder irgendeine andere Quelle). Wir können Paare von natürlichen Zahlen auf verschiedenen Weisen zählen: Zeilenweise, Spaltenweise, oder wie eine Schlange im Diagonalverfahren.

(b) Wir zeigen $\mathcal{S} \subseteq \mathcal{A}_{\mu^*}$. Dazu müssen wir

$$\mu^*(Z) = \mu^*(Z \cap S) + \mu^*(Z \cap S^C), \quad \text{für alle } Z \subseteq \Omega, S \in \mathcal{S},$$

nachrechnen.

„$\leq$“: $\mu^*(Z) = \mu^*((Z \cap S) \cup (Z \cap S^C)) \leq \mu^*(Z \cap S) + \mu^*(Z \cap S^C)$ gilt aufgrund der gezeigten Subadditivität von μ^*.

„$\geq$“: Seien $A_1, A_2, ... \subseteq \mathcal{S}$ mit $Z \subseteq \bigcup_{k=1}^{\infty} A_k$. Weil $\mathcal{S}$ ein Semiring ist, existieren $C_{k,1}, ..., C_{k,m_k} \in \mathcal{S}$ mit

$$A_k \cap S^C = A_k \backslash S = \dot{\bigcup_{j=1}^{m_k}} C_{k,j}.$$

Es gelten

$$Z \cap S \subseteq \Big(\bigcup_{k=1}^{\infty} A_k \Big) \cap S = \bigcup_{k=1}^{\infty} (A_k \cap S)$$

sowie analog

$$Z \cap S^C \subseteq \Big(\bigcup_{k=1}^{\infty} A_k \Big) \cap S^C = \bigcup_{k=1}^{\infty} \big(A_k \cap S^C\big) = \bigcup_{k=1}^{\infty} \dot{\bigcup_{j=1}^{m_k}} C_{k,j}.$$

Mit den Definitionen folgt

$$\begin{aligned} \mu^*(Z \cap S) + \mu^*(Z \cap S^C) &\leq \sum_{k=1}^{\infty} \Big(\mu(A_k \cap S) + \sum_{j=1}^{m_k} \mu(C_{k,j}) \Big) \\ &\overset{\mu\ \sigma\text{-add.}}{=} \sum_{k=1}^{\infty} \mu\Big((A_k \cap S) \,\dot\cup\, \dot{\bigcup_{j=1}^{m_k}} C_{k,j} \Big) \\ &= \sum_{k=1}^{\infty} \mu\big((A_k \cap S) \,\dot\cup\, (A_k \cap S^C) \big) \\ &= \sum_{k=1}^{\infty} \mu(A_k). \end{aligned}$$

Daraus folgt $\mu^*(Z \cap S) + \mu^*(Z \cap S^C) \leq \mu^*(Z)$, weil

$$\mu^*(Z) = \inf \left\{ \sum_{k=1}^{\infty} \mu(A_k) \colon A_1, A_2, ... \in \mathcal{S} \text{ mit } Z \subseteq \bigcup_{k=1}^{\infty} A_k \right\}.$$

Somit ist „$\geq$“ gezeigt. Also ist jedes $S \in A_{\mu^*}$ und damit gilt $\mathcal{S} \subseteq A_{\mu^*}$.

(c) Fehlt noch $\mu^*(A) = \mu(A)$ für alle $A \in \mathcal{S}$. Im Prinzip ist das Lemma 1.3.3.

„$\leq$“:

$$\mu^*(A) = \inf \left\{ \sum_{k=1}^{\infty} \mu(A_k) \colon A_1, A_2, ... \in \mathcal{S} \text{ mit } A \subseteq \bigcup_{k=1}^{\infty} A_k \right\} \leq \mu(A)$$

für alle $A \in \mathcal{S}$.

„$\geq$“: Ist $A \subseteq \bigcup_{k=1}^{\infty} A_k$ für $A_1, A_2, ... \in \mathcal{S}$, so gilt

$$\mu(A) \overset{1.3.3}{\leq} \sum_{k=1}^{\infty} \mu(A_k).$$

Folglich gilt

$$\mu(A) \leq \inf \left\{ \sum_{k=1}^{\infty} \mu(A_k) \colon A_1, A_2, ... \in \mathcal{S} \text{ mit } A \subseteq \bigcup_{k=1}^{\infty} A_k \right\} = \mu^*(A).$$

$\square$

Wir fassen nun den Existenz- und den Eindeutigkeitssatz zusammen, das gibt folgendes zentrale Theorem:

Satz 1.3.8. **[Existenz und Eindeutigkeit von Maßen]** Es sei $(\Omega, \mathcal{A})$ ein messbarer Raum, $\mathcal{E}$ ein Semiring mit $\sigma(\mathcal{E}) = \mathcal{A}$ und $\mu\colon \mathcal{E} \to [0, \infty]$ eine Mengenfunktion mit

- $\mu(\emptyset) = 0$
- μ ist σ-additiv
- es gibt eine Folge $E_1, E_2, ... \in \mathcal{E}$ mit $E_n \uparrow \Omega$ und $\mu(E_n) < \infty$ für alle $n \in \mathbb{N}$.

Dann existiert genau ein Maß $\bar{\mu}$ auf $\mathcal{A} = \sigma(\mathcal{E})$, so dass $\bar{\mu}(A) = \mu(A)$ für alle $A \in \mathcal{E}$.

Beweis. Existenz folgt direkt aus 1.3.7, Eindeutigkeit folgt direkt aus Satz 1.2.13. □

Merkt euch für immer folgendes, das hilft euch, die verschiedenen Konzepte einzuordnen:

Existenz ist Carathéodory mit äußeren Maßen, Eindeutigkeit folgt aus Dynkin-Systemen!

Anschließend an die abstrakte Theorie wollen wir nun als Beispiel die Borel-σ-Algebra auf $\mathbb{R}$ diskutieren. Hier wird alles viel klarer werden und das ist gerade die Anwendung, die wir für die Stochastik brauchen.

1.4 Das Beispiel - Wahrscheinlichkeitsmaße auf $\mathcal{B}(\mathbb{R})$ aus Verteilungsfunktionen

Das Konzept von Verteilungsfunktionen von Wahrscheinlichkeitsmaßen auf $\mathcal{B}(\mathbb{R})$ ist euch bereits bei Dynkin-Systemen und in den Übungen über den Weg gelaufen. Definieren wir nun abstrakt, welche reellen Funktionen wir Verteilungsfunktionen nennen wollen:

Definition 1.4.1. $F\colon \mathbb{R} \to \mathbb{R}$ heißt **Verteilungsfunktion**, falls

(i) $0 \leq F(t) \leq 1$ für alle $t \in \mathbb{R}$,

(ii) F ist nicht fallend,

(iii) F ist rechtsstetig, d. h. $\lim_{s \downarrow t} F(s) = F(t)$,

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

Der Hauptsatz über Wahrscheinlichkeitsmaße auf $\mathcal{B}(\mathbb{R})$ ist nun folgender bijektiver Zusammenhang zu Verteilungsfunktionen:

Satz 1.4.2. **[Wahrscheinlichkeitsmaße aus Verteilungsfunktionen]** Für jede Verteilungsfunktion F gibt es **genau ein** Wahrscheinlichkeitsmaß $\mathbb{P}_F$ und $\mathcal{B}(\mathbb{R})$ mit

$$\mathbb{P}_F((-\infty, t]) = F(t), \quad t \in \mathbb{R}.$$

Man sagt dann, „$\mathbb{P}_F$ ist gemäß F verteilt“ oder „$\mathbb{P}_F$ hat Verteilung F“ und schreibt $\mathbb{P}_F \sim F$.

Beweis. **Eindeutigkeit**: Das haben wir uns schon in Beispiel 1.2.15 überlegt.

Existenz: Um den Fortsetzungssatz von Carathéodory zu nutzen, müssen wir zunächst einen Semiring wählen, der die Borel-σ-Algebra erzeugt. Wir wissen bereits, dass alle möglichen Arten von Intervallen $\mathcal{B}(\mathbb{R})$ erzeugt, die meisten sind aber keine Semiringe. Wir nehmen

$$\mathcal{S} = \{(a, b] : a \leq b\},$$

und stellen sofort fest (Eigenschaften checken), dass $\mathcal{S}$ ein Semiring mit $\sigma(\mathcal{S}) = \mathcal{B}(\mathbb{R})$ ist. Als Mengenfunktion auf $\mathcal{S}$ definieren wir

$$\mu((a,b]) := F(b) - F(a), \quad a \leq b.$$

Weil F nicht-fallend ist und $0 \leq F \leq 1$ gilt, bildet μ nach $[0,1]$ ab. Checken wir als nächstes die Voraussetzungen vom Fortsetzungssatz:

(i) $\mu(\emptyset) = \mu((a,a]) = F(a) - F(a) = 0$

(ii) Für die σ-Additivität von μ auf $\mathcal{S}$ seien $(a_n, b_n] \in \mathcal{S}$ paarweise disjunkt mit

$$\dot{\bigcup_{k=1}^{\infty}} (a_k, b_k] \in \mathcal{S}, \qquad \text{also } \dot{\bigcup_{k=1}^{\infty}} (a_k, b_k] =: (a,b],$$

für geeignete $a, b \in \mathbb{R}$. Als Anschauungsbeispiel haltet ihr am besten das konkrete Beispiel $(0,1] = \dot{\bigcup}_{k=1}^{\infty} (\frac{1}{k+1}, \frac{1}{k}]$ im Kopf. Um den Beweis besser zu verstehen, schauen wir uns erstmal den endlichen Fall an, d. h.

$$(a,b] = \dot{\bigcup_{k=1}^{N}} (a_k, b_k],$$

für ein $N \in \mathbb{N}$ mit geordneten Intervallen $a = a_1 < b_1 = a_2 < ... < b_N = b$. Dann bekommen wir die σ-Additivität sofort:

$$\mu((a,b]) \overset{\text{Def.}}{=} F(b) - F(a) \overset{\text{Teleskop}}{=} \sum_{k=1}^{N} (F(b_k) - F(a_k)) = \sum_{k=1}^{N} \mu((a_k, b_k]),$$

wobei wir $F(a_k) = F(b_{k-1})$ genutzt haben. Nun aber zurück zum allgemeinen Fall: Wir zeigen Vorlesung 6

$$F(b) - F(a) = \sum_{k=1}^{\infty} (F(b_k) - F(a_k)), \tag{1.6}$$

denn das ist gerade die σ-Additivität

$$\mu\Big(\dot{\bigcup_{k=1}^{\infty}} (a_k, b_k]\Big) = \sum_{k=1}^{\infty} \mu((a_k, b_k])$$

für $(a,b] = \dot{\bigcup}_{k=1}^{\infty} (a_k, b_k]$. Für die Gleichheit (1.6) zeigen wir beide Ungleichungen:

„$\geq$“: Weil F nicht-fallend ist, folgt

$$F(b) - F(a) \geq F(b_N) - F(a_1) \overset{\text{Teleskop}}{=} \sum_{k=1}^{N} (F(b_k) - F(a_k))$$

für alle $N \in \mathbb{N}$. Wegen der Monotonie von Folgengrenzwerten gilt

$$F(b) - F(a) \geq \lim_{N \to \infty} \sum_{k=1}^{N} (F(b_k) - F(a_k)) = \sum_{k=1}^{\infty} (F(b_k) - F(a_k)).$$

„$\leq$“: Sei $\varepsilon > 0$ und seien $b_n < \tilde{b}_n$, so dass

$$0 \leq F(\tilde{b}_n) - F(b_n) < \frac{\varepsilon}{2^n} \tag{1.7}$$

für alle $n \in \mathbb{N}$. Die $\tilde{b}_n$ existieren weil F rechtsstetig ist (schreibt mal die Definition der Stetigkeit mit $\frac{\varepsilon}{2^n}$ statt ε hin). Weil

$$(a,b] = \bigcup_{k=1}^{\infty}(a_k,b_k] \overset{b_k<\tilde{b}_k}{\subseteq} \bigcup_{k=1}^{\infty}(a_k,\tilde{b}_k)$$

gilt, gilt auch

$$[a+\varepsilon,b] \subseteq \bigcup_{k=1}^{\infty}(a_k,\tilde{b}_k).$$

Nach Heine-Borel ist $[a+\varepsilon,b]$ kompakt. Aufgrund der Definition der Kompaktheit reichen endlich viele $(a_k,\tilde{b}_k)$, um $[a+\varepsilon,b]$ zu überdecken. Also gibt es ein $N \in \mathbb{N}$ mit

$$[a+\varepsilon,b] \subseteq \bigcup_{k=1}^{N}(a_k,\tilde{b}_k).$$

Daraus folgt dann

$$\begin{aligned} F(b)-F(a+\varepsilon) &\overset{\text{F monoton}}{\leq} \sum_{k=1}^{N}(F(\tilde{b}_k)-F(a_k)) \\ &\overset{(1.7)}{\leq} \sum_{k=1}^{\infty}\Big(F(b_k)+\frac{\varepsilon}{2^k}-F(a_k)\Big) = \sum_{k=1}^{\infty}(F(b_k)-F(a_k))+\varepsilon, \end{aligned}$$

wobei wir im letzten Schritt die geometrische Reihe $\sum_{k=1}^{\infty}\frac{1}{2^k}=1$ genutzt haben. Wegen der Rechtsstetigkeit von F folgt damit

$$\begin{aligned} F(b)-F(a) &= \lim_{\varepsilon\downarrow 0}(F(b)-F(a+\varepsilon)) \\ &\leq \lim_{\varepsilon\downarrow 0}\Big(\sum_{k=1}^{\infty}(F(b_k)-F(a_k))+\varepsilon\Big) = \sum_{k=1}^{\infty}(F(b_k)-F(a_k)). \end{aligned}$$

Der Fortsetzungssatz impliziert nun die Existenz eines Maßes $\mathbb{P}_F$ auf $\mathcal{B}(\mathbb{R})$ mit

$$\mathbb{P}_F((a,b]) = \mu((a,b]) = F(b)-F(a), \quad a<b.$$

Das Maß ist nicht automatisch ein Wahrscheinlichkeitsmaß, das folgt aber direkt aus der Stetigkeit von Maßen und der Charakterisierung von $\mathbb{P}_F$ auf den Intervallen:

$$\begin{aligned} \mathbb{P}_F(\mathbb{R}) &= \mathbb{P}_F\Big(\bigcup_{k=1}^{\infty}(-k,k]\Big) \\ &\overset{\text{Stet. Maße}}{=} \lim_{n\to\infty}\mathbb{P}_F((-n,n]) \\ &\overset{\text{Def.}}{=} \lim_{n\to\infty}(F(n)-F(-n)) \\ &= \lim_{n\to\infty}F(n) - \lim_{n\to\infty}F(-n) = 1. \end{aligned}$$

Achtung, das Argument werden wir jetzt immer wieder nutzen!

Ganz ähnlich zeigen wir den Zusammenhang von F und $\mathbb{P}_F$ auf unendlichen Intervallen, wie im Satz behauptet wird:

$$\mathbb{P}_F((-\infty,t]) = \mathbb{P}_F\Big(\bigcup_{k=\lceil |t|\rceil}^{\infty}(-k,t]\Big) = \lim_{k\to\infty}\mathbb{P}_F((-k,t]) = F(t) - \lim_{k\to\infty}F(-k) = F(t),$$

wobei $\lceil |t| \rceil$ die obere Gaußklammer von $|t|$ ist, also $|t|$ aufgerundet.

□

Bemerkung 1.4.3. Es gibt ganz analog eine Definition für Verteilungsfunktionen auf dem $\mathbb{R}^d$, sogenannte „multivariate Verteilungsfunktionen“ – das machen wir später.

Bevor wir zu Beispielen von Wahrscheinlichkeitsmaßen auf $\mathcal{B}(\mathbb{R})$ kommen, hier noch das zweite wichtige Maßes auf der Borel-σ-Algebra - das Lebesgue-Maß (Volumen) auf $\mathcal{B}(\mathbb{R}^d)$.

Satz 1.4.4. **[Lebesgue-Maß auf $\mathbb{R}^d$]** Es gibt ein eindeutiges Maß λ auf $\mathcal{B}(\mathbb{R}^d)$ mit $\lambda(Q) = \text{Volumen(Q)}$ für alle Quader $Q \subseteq \mathbb{R}^d$. λ heißt Lebesgue-Maß auf $\mathbb{R}^d$ und ist ein unendliches Maß.

Beweis. Übung, ziemlich analog zum vorherigen Beweis. Wir checken die Voraussetzungen von Satz 1.3.8, hier ist eine Skizze: Betrachte

$$\mathcal{S} := \big\{(a_1, b_1] \times \cdots \times (a_n, b_n] : a_1, \ldots, a_d, b_1, \ldots, b_d \in \mathbb{R}\big\},$$

$\mathcal{S}$ ist ein Semiring.

$$\mu(Q) := \text{Volumen(Q)} = \prod_{k=1}^{d} (b_k - a_k)$$

ist eine σ-additive Mengenfunktion auf $\mathcal{S}$ (die σ-Additivität zeigt man mit Kompaktheit im $\mathbb{R}^d$, genau wie im letzten Satz). Mit der Folge $E_n := (-n, n] \times \ldots (-n, n] \in \mathcal{S}$ haben wir eine Folge mit endlichem Volumen, die gegen $\mathbb{R}^d$ wächst. Das ist die dritte Eigenschaft von Satz 1.3.8, es gibt also eine eindeutige Fortsetzung von μ auf $\mathcal{B}(\mathbb{R}^d)$. Das Maß nennen wir λ. Es fehlt noch, dass λ ein unendliches Maß ist. Aber auch das geht mit den Argumenten des vorherigen Beweises:

$$\lambda(\mathbb{R}^d) \overset{\text{Stet. Maße}}{=} \lim_{n \to \infty} \lambda((n, -n] \times ... \times (n, -n]) = \lim_{n \to \infty} (2n)^d = \infty.$$

□

Meistens betrachten wir das Lebesguemaß auf $\mathbb{R}$, das im Prinzip die „Länge“ einer Menge misst, zumindest gilt das für Intervalle (oder disjunkte Vereinigungen von Intervallen).

Bemerkung 1.4.5. Auf den Übungsblättern diskutieren wir das Lebesgue-Maß auf Teilmengen vom $\mathbb{R}^d$, insbesondere auf Intervalle oder Quadern. Beispielsweise sind dann die messbaren Mengen

$$\mathcal{B}([0,1]) := \big\{B \subseteq [0,1] : B \in \mathcal{B}(\mathbb{R})\big\} = \sigma(\{[a,b] : 0 \leq a < b \leq 1\})$$

die Borel-messbaren Teilmengen von $[0,1]$ und $\lambda_{[0,1]}$ das eindeutige Maß auf $\mathcal{B}([0,1])$ mit $\lambda_{[0,1]}(B) = \lambda(B)$ für Borel-messbare Teilmengen von $[0,1]$. Das Lebesgue-Maß auf $[0,1]$ ist das eindeutige Maß auf den Borel-messbaren Teilmengen von $[0,1]$, so dass das Maß von Intervallen die Länge ist.

Jetzt kommen wir zu konkreten Beispielen von Verteilungsfunktionen, die uns erneut in der Stochastik begegnen werden. Im Folgenden werden wir regelmäßig **Indikatorfunktionen** benutzen:

$$\mathbf{1}_A(x) := \begin{cases} 1 & : x \in A \\ 0 & : x \notin A \end{cases}, \quad x \in \Omega,$$

die euch in Analysis 2 (vielleicht in anderer Schreibweise) im Rahmen der Integrationstheorie vermutlich schon über den Weg gelaufen sind.

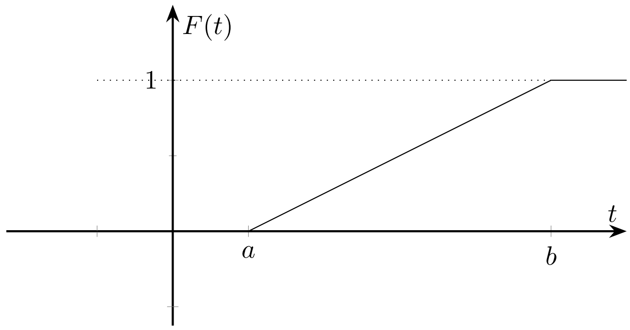
Beispiel 1.4.6. Für $a < b$ sei

$$F(t) = \frac{t-a}{b-a} \mathbf{1}_{[a,b]}(t) + \mathbf{1}_{(b,\infty)}(t), \quad t \in \mathbb{R},$$

oder anders geschrieben als

$$F(t) = \begin{cases} 0 & : t < a \\ \frac{t-a}{b-a} & : t \in [a,b] \\ 1 & : t > b \end{cases}.$$

Natürlich erfüllt F die Eigenschaften einer Verteilungsfunktion, das zugehörige Maß $\mathbb{P}_F$ nennt man **Gleichverteilung** auf $[a,b]$ und man schreibt $\mathbb{P}_F \sim \mathcal{U}([a,b])$.



Man nennt das Maß auch $\mathcal{U}([a,b])$, $\mathcal{U}$ steht dabei für uniform.

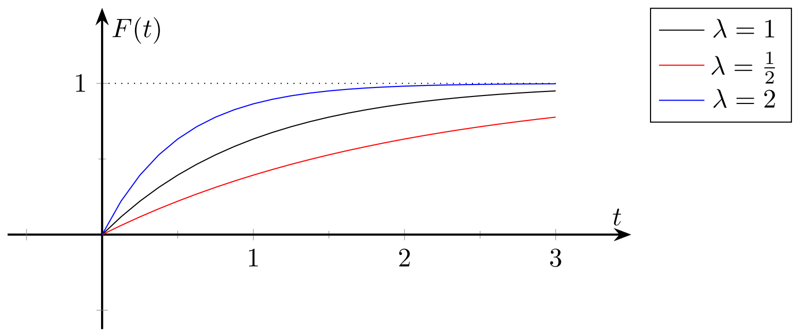
Beispiel 1.4.7. Für $\lambda > 0$ sei

$$F(t) = (1 - e^{-\lambda t})\mathbf{1}_{[0,\infty)}(t), \quad t \in \mathbb{R},$$

oder anders geschrieben als

$$F(t) = \begin{cases} 0 & : t \leq 0 \\ 1 - e^{-\lambda t} & : t > 0 \end{cases}.$$

Aufgrund der Eigenschaften der Exponentialfunktion erfüllt F die Eigenschaften der Exponentialfunktion, das zugehörige Maß $\mathbb{P}_F$ nennt man **Exponentialverteilung mit Parameter** $\lambda > 0$ und schreibt $\mathbb{P}_F \sim \operatorname{Exp}(\lambda)$.



Man nennt das Maß auch $\operatorname{Exp}(\lambda)$. In der Graphik ist $\operatorname{Exp}(\lambda)$ für drei verschiedene λ geplotted.

Definition 1.4.8. Ist $f : \mathbb{R} \to [0,\infty)$ integrierbar mit $\int_{\mathbb{R}} f(x)\mathrm{d}x = 1$, dann heißt f **Dichtefunktion** der Verteilungsfunktion

$$F(t) = \int_{-\infty}^{t} f(x)\mathrm{d}x, \quad t \in \mathbb{R}. \tag{1.8}$$

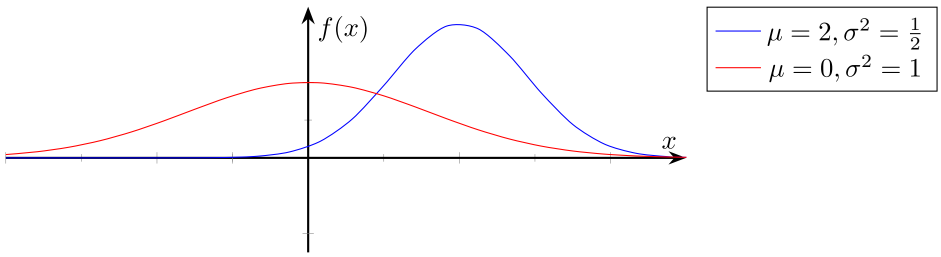
Beachte: Solch eine Integralfunktion F erfüllt automatisch die Eigenschaften einer Verteilungsfunktion (siehe große Übung)! Ist umgekehrt F von der Form (1.8), so heißt f **Dichte** von F. Verteilungsfunktionen mit Dichten nennt man auch **absolutstetig**, Wahrscheinlichkeitsmaße auf $\mathcal{B}(\mathbb{R})$ mit absolutstetiger Verteilungsfunktion nennt man **absolutstetige Maße**.

Zwei Beispiele haben wir schon gesehen: $\mathcal{U}([a,b])$ und $\operatorname{Exp}(\lambda)$ haben beide absolutstetige Verteilungsfunktionen. Die zugehörigen Dichten berechnet ihr in den Übungsaufgaben. Aber wie findet man die Dichten von absolutstetigen Verteilungsfunktionen? Ableiten! Das ist, zumindest für stetige Dichten, der Hauptsatz der Integral- und Differentialrechnung. Probiert das bei den zwei Beispielen mal aus. Ableiten ohne nachzudenken erlaubt es die Dichte f zu erraten, wenn man dann durch Integrieren $F(t) = \int_{-\infty}^t f(x)\mathrm{d}x$ nachrechnen kann, so ist f eine Dichte von F. Warum es praktisch ist eine absolutstetige Verteilungsfunktion zu haben, wird zum Beispiel in Diskussion 1.4.13 klarer. Man kann direkt wichtige Eigenschaften des Maßes $\mathbb{P}_F$ aus der Dichte f ablesen.

Beispiel 1.4.9. Die schönste Anwendung von Polarkoordinaten und Fubini (siehe Analysis 2) ist die Berechnung des Integrals $\int_{\mathbb{R}} e^{-\frac{x^2}{x}} dx = \sqrt{2\pi}$. Damit ist $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ eine Dichtefunktion. Man nennt die zugehörige Verteilungsfunktion

$$F(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \mathrm{d}x, \quad t \in \mathbb{R},$$

Verteilungsfunktion der **(standard) Normalverteilung**. Das Maß $\mathbb{P}_F$ nennt man dann auch (standard) normalverteilt und man schreibt $\mathbb{P}_F \sim \mathcal{N}(0,1)$. In der großen Übung wird diskutiert, dass für $\mu \in \mathbb{R}$ und $\sigma^2 \geq 0$ auch $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ eine Dichtefunktion ist. Die zugehörige Verteilung nennt man auch normalverteilt und schreibt $\mathcal{N}(\mu, \sigma^2)$.



Die Bedeutung von μ und σ^2 diskutierten wir später. Warnung: Warum schreiben wir nur eine Formel für die Dichte f, jedoch nicht für die Verteilungsfunktion F hin? Es gibt einfach keine Formel für das Integral $\int_{-\infty}^t e^{-x^2/2} dx$! Aufgrund der Form der Kurve spricht man auch von der Glockenkurve und weil diese von Gauß entdeckt wurde, von der Gausschen Glockenkurve.

Das Gegenstück zu absolutstetigen Verteilungen sind sogenannte diskrete Verteilungen:

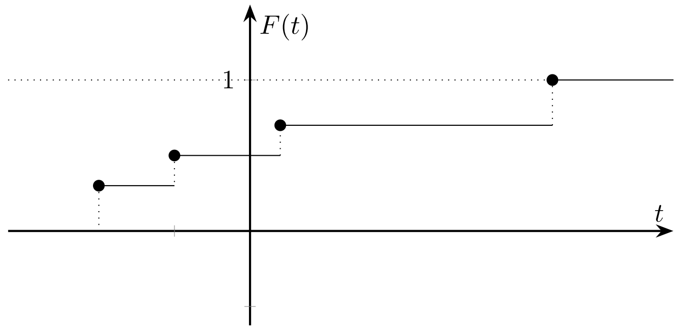
Beispiel 1.4.10. Für $a_1, ..., a_N \in \mathbb{R}$, $N \in \mathbb{N}$ oder $N = +\infty$, mit $p_1, ..., p_N \geq 0$ und $\sum_{k=1}^{N} p_k = 1$ ist

$$F(t) := \sum_{k=1}^{N} p_k \mathbf{1}_{[a_k,\infty)}(t) = \sum_{a_k \leq t} p_k, \quad t \in \mathbb{R},$$

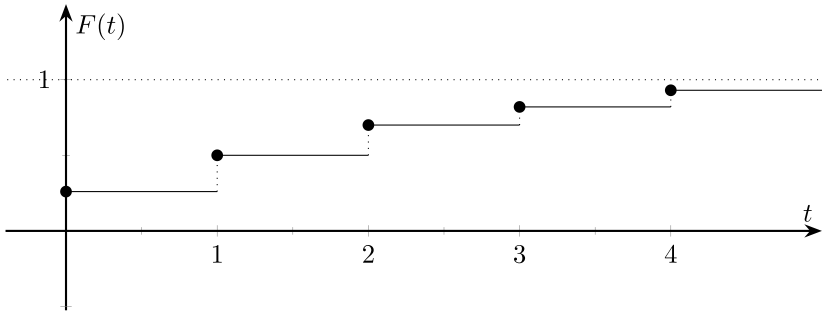
eine Verteilungsfunktion. Die zugehörigen Maße $\mathbb{P}_F$ werden **(endliche) diskrete Verteilungen** genannt. In den Übungen zeigt ihr, dass die Maße im diskreten Fall ganz einfach angegeben werden können, es sind Mischungen aus Dirac-Maßen an den Stellen $a_1, ..., a_N$:

$$\mathbb{P}_F = \sum_{k=1}^{N} p_k \delta_{a_k}.$$

Wie zeigt man das? Einfach die Menge $(-\infty, t]$ in das Maß einsetzen, das gibt das gewünschte F.



Ganz konkret heißt $\mathbb{P}_F$ für $a_k = k$ und $p_k = e^{-\lambda}\frac{\lambda^k}{k!}$, $k \in \mathbb{N}$, **Poissonverteilung mit Parameter** $\lambda > 0$ auf $\mathcal{B}(\mathbb{R})$. Beachte: Weil wir die Poissonverteilung bereits auf $\mathcal{P}(\mathbb{N})$ definiert haben gibt es eine gewisse Doppeldeutigkeit. Mit der Diskussion der nächsten Vorlesung wird aber klar, dass beide Maße das gleiche beschreiben, nämlich die Verteilung einer Einheit Masse auf $\mathbb{N}$ mit den Wahrscheinlichkeiten p_k für die die natürliche Zahl k. Die Poissonverteilung mit Parameter λ wird auch als **Poi**(λ) genannt.



Vorlesung 7

Manche werden sich fragen, wo denn jetzt die Stochastik geblieben ist. Wir haben schließlich gerade Begriffe der Stochastik benutzt, z. B. den Begriff der Uniformverteilung, auch die Gaußsche Glockenfunktion ist bereits aufgetaucht, über zufällige Experimente haben wir aber schon länger nicht gesprochen. Als konkrete Motivation zur Nutzung der abstrakten Theorie zur Modellierung zufälliger Experimente, schauen wir uns das uniforme Ziehen aus $[0,1]$ an.

Diskussion 1.4.11. **[Stochastische Modellierung, Nr. 2]** Das Modellieren von endlich vielen Möglichkeiten ist relativ einfach, siehe Diskussion 1.1.8. Man kommt recht natürlich auf die Eigenschaften der σ-Algebra und des Maßes. Zur Erinnerung war das gleichverteilte Ziehen aus einer endlichen Menge modelliert durch den endlichen Zustandsraum Ω (=Möglichkeiten zum Ziehen), $\mathcal{A} = \mathcal{P}(\Omega)$ und der diskreten Gleicherverteilung $\mathbb{P}(A) = \frac{\#A}{\#\Omega}$.
Das Modellieren von Experimenten mit unendlich vielen Möglichkeiten ist dagegen schwieriger. Wie modelliert man zum Beispiel das Ziehen aus dem Intervall $[0,1]$, sodass kein Bereich von $[0,1]$ bevorteilt wird? Wenn wir beobachten wollen, ob eine feste Zahl gezogen wurde oder nicht, müssen die einelementigen Mengen $\{t\}$ in der σ-Algebra sein. Wenn kein Element bevorzugt werden soll, also $\mathbb{P}(\{t\})$ für alle t gleich sein soll, führt die Unendlichkeit automatisch zu $\mathbb{P}(\{t\}) = 0$ für alle $t \in [0,1]$. Warum das? Wenn man irgendeine Folge (a_n) unterschiedlicher Zahlen in $[0,1]$ wählt, z. B. $a_n = \frac{1}{n}$, und $\mathbb{P}(\{t\}) =: c$ für alle t setzt, so gilt wegen der σ-Addititivität von Maßen

$$1 \geq \mathbb{P}\Big(\dot{\bigcup_{k=1}^{\infty}}\{a_k\}\Big) = \sum_{k=1}^{\infty}\mathbb{P}(\{a_k\}) = \sum_{k=1}^{\infty} c,$$

also $c = 0$. Hier sehen wir deutlich den Unterschied zur Gleichverteilung auf endlichen Mengen, die einfache Definition durch Einpunktmengen führt zu nichts! Im Gegensatz zum endlichen Fall legen wir für gleichverteilten Zufall in $[0,1]$ jetzt fest, dass die Wahrscheinlichkeit von Teilintervallen von $[0,1]$ nur von der Länge abhängen soll. Das führt zur Forderung $\mathbb{P}((a,b]) = b - a = F(b) - F(a)$ für $a < b$ aus $[0,1]$, wobei F die Verteilungsfunktion aus Beispiel 1.4.6 ist. Da wir als mathematisches

Modell des zufälligen Ziehens eine σ-Algebra und ein Maß haben wollen, wählen wir nun die kleinste σ-Algebra die all diese Intervalle enthält (die Borel-σ-Algebra) und darauf ein Maß, das den Intervallen die geforderten Wahrscheinlichkeiten gibt. Aufgrund des Fortsetzungssatzes gibt es so ein Maß, das ist gerade $\mathcal{U}([0,1])$.

Hoffentlich ist jetzt einsichtig, warum die Modellierung von komplizierten reellen zufälligen Experimenten mit der Borel-σ-Algebra Sinn macht. Eine Frage bleibt aber noch: Warum nehmen wir nicht einfach die ganze Potenzmenge auf $\mathbb{R}$ als Modell, so wie beim zufälligen Ziehen in endlichen Mengen?

Bemerkung 1.4.12.

(i) $\mathcal{B}(\mathbb{R})$ funktioniert wunderbar! Insbesondere weil wir sehr handliche Erzeuger haben (z. B. verschiedene Arten von Intervallen) und deshalb aufgrund der bewiesenen Theoreme (fast) nur mit Intervallen arbeiten müssen.

(ii) $\mathcal{P}(\mathbb{R})$ ist zu groß, z. B. das Lebesgue-Maß oder die Normalverteilung kann zwar auf $\mathcal{B}(\mathbb{R})$, aber nicht auf $\mathcal{P}(\mathbb{R})$ definiert werden ($\rightsquigarrow$ Vitali-Menge). Es gilt tatsächlich $\mathcal{B}(\mathbb{R}) \subsetneq \mathcal{P}(\mathbb{R})$, ganz einfache Beispiele für nicht Borel-messbare Mengen gibt es aber nicht.

Die nächste Runde der Modellierung zufälliger Experimente findet erst in ein paar Wochen statt. Bis dahin könnt ihr die Ideen sacken lassen und euch wieder an der abstrakten Theorie erfreuen.

Das Umschalten im Kopf von Verteilungsfunktionen auf Maße ist anfangs extrem schwierig. Wir wissen zwar abstrakt, dass es für jede Verteilungsfunktion genau ein Maß auf $\mathcal{B}(\mathbb{R})$ gibt und andersrum für jedes Maß eine eindeutige Verteilungsfunktion, aber was bedeutet das konkret? Das versteht man am besten, wenn man Eigenschaften von F in Eigenschaften von $\mathbb{P}_F$ übersetzt:

Diskussion 1.4.13. Wir starten mit einer nicht sehr rigorosen aber dennoch hilfreichen Interpretation:

„F beschreibt, wie durch $\mathbb{P}_F$ eine Einheit Zufall auf $\mathbb{R}$ verteilt wird.“

Dazu sei $F(b) - F(a)$ der Anteil des gesamten Zufalls ($F(b) - F(a)$ ist immer zwischen 0 und 1), der in $(a, b]$ gelandet ist. Man spricht auch statt „Anteil“ von der „Masse“ Zufall in $(a, b]$.

Wir schauen uns jetzt an, was drei Eigenschaften von F (stetig, konstant, stark wachsend) für die Verteilung der Masse bedeuten.

Stetigkeit vs. Sprünge: Zunächst berechnen wir die Masse einer Einpunktmenge $\{t\}$ aus den bekannten Eigenschaften von Maßen und Verteilungsfunktionen. Wie immer versuchen wir die gesuchte Menge durch Mengen der Form $(a, b]$ auszudrücken, weil wir für diese Mengen eine Verbindung zwischen F und $\mathbb{P}_F$ haben:

$$\begin{aligned}
\mathbb{P}_F(\{t\}) &= \mathbb{P}_F\Big(\bigcap_{n=1}^{\infty}\Big(t-\frac{1}{n},t\Big]\Big)\\
&\overset{\text{Stet. Maße}}{=} \lim_{n\to\infty}\mathbb{P}_F\Big(\Big(t-\frac{1}{n},t\Big]\Big)\\
&\overset{\text{Def. } \mathbb{P}_F}{=} \lim_{n\to\infty}\Big(F(t)-F\Big(t-\frac{1}{n}\Big)\Big)\\
&= F(t) - \lim_{n\to\infty}F\Big(t-\frac{1}{n}\Big) = F(t)-F(t-),
\end{aligned}$$

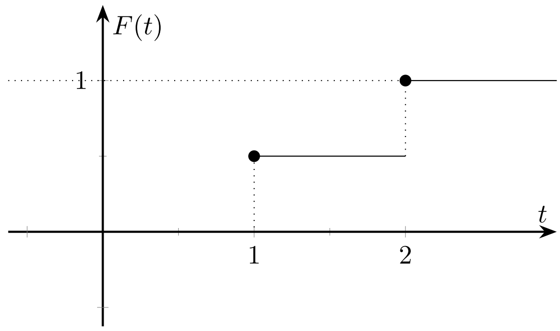
wobei $F(t-) := \lim_{s\uparrow t} F(s)$ der Linksgrenzwert aus der Analysis ist. Konsequenz: Ist F stetig in t, so hat die Einpunktmenge $\{t\}$ keine Masse. Hierzu beachte man, dass F an jeder Stelle rechtsstetig ist, die Stetigkeit somit äquivalent zu $F(t) = F(t-)$ ist. Insbesondere haben alle einpunktigen Mengen keine Masse, sofern F eine stetige Funktion ist (z. B. bei $\mathcal{U}([a,b])$, $\operatorname{Exp}(\lambda)$, $\mathcal{N}(\mu,\sigma^2)$).

Klingt komisch, oder? Ist es aber nicht. Hier sehen wir, warum Maße erst auf überabzählbaren Mengen wirklich spannend werden:

$$\mathbb{P}_F((a,b]) = \mathbb{P}_F\Big(\dot{\bigcup_{t\in(a,b]}} \{t\}\Big) \neq \sum_{t\in(a,b]} \mathbb{P}_F(\{t\}),$$

weil σ-Additivität nur für Vereinigungen abzählbar vieler Mengen gilt. Was sollte die überabzählbare Summe auf der rechten Seite auch bedeuten?

F **konstant:** Überlegen wir nun, was es für $\mathbb{P}_F$ bedeutet, wenn F auf einem Intervall konstant ist. Schauen wir dazu zunächst ein Beispiel an. Betrachten wir folgende einfache Verteilungsfunktion



aus der Klasse der diskreten Verteilungen. Nach der Diskussion zur Stetigkeit wissen wir, dass das zugehörige Maß $\mathbb{P}_F$ folgendes erfüllt: $\mathbb{P}_F(\{1\}) = \mathbb{P}_F(\{2\}) = \frac{1}{2}$. Wegen der σ-Additivität folgt natürlich (es gibt insgesamt nur eine Einheit Zufall zu verteilen), dass $\mathbb{P}_F(A) = 0$ für alle Borelmengen A mit $1, 2 \notin A$. Das Maß $\mathbb{P}_F$ hat also keine Masse außerhalb der Menge $\{1, 2\}$. Schauen wir uns F an, so sehen wir also, dass $\mathbb{P}_F$ keine Masse in den konstanten Bereichen hat. Für Intervalle $(a, b]$ folgt das allgemein natürlich aus $\mathbb{P}_F((a, b]) = F(b) - F(a)$ was gerade 0 ist, wenn F zwischen a und b konstant ist:

„$\mathbb{P}_F$ hat keine Masse dort, wo F konstant ist.“

Wenn wir die Beobachtung auf $\text{Poi}(\lambda)$ aus Beispiel 1.4.10 anwenden, so sehen wir, dass das zugehögige Maß $\mathbb{P}_F$ nur Masse auf $\mathbb{N}$ hat. Damit kann man ein $\text{Poi}(\lambda)$-verteilte Maß auf $\mathcal{B}(\mathbb{R})$ mit der Definition aus Beispiel 1.1.12 identifizieren, wir verteilen eine Einheit Zufall jeweils auf $\mathbb{N}$ (einmal wird die Einheit Zufall direkt auf $\mathbb{N}$ verteilt, einmal auf $\mathbb{N}$ als Teilmenge von $\mathbb{R}$).

F **stark wachsend:** Wir wissen nun wieviel Masse an Sprungstellen liegt und auch, dass keine Masse in konstanten Bereichen liegt. Fragt sich also, wo die Masse sonst noch zu finden ist:

„$\mathbb{P}_F$ hat viel Masse dort, wo F am stärksten wächst.“

Formell folgt das natürlich aus $\mathbb{P}_F((a, b]) = F(b) - F(a)$ weil dann auf ein kleines Intervall $(a, b]$ viel Masse verteilt wird, wenn $F(b)$ deutlich größer als $F(a)$. Ist a nah an b, so bedeutet das natürlich, dass F dort stark wächst. Schauen wir uns wieder ein passendes Beispiel an, die Exponentialverteilung $\text{Exp}(\lambda)$ für verschiedene $\lambda > 0$. Am Bildchen in Beispiel 1.4.7 ist zu erkennen, dass viel Masse nah bei der 0 liegt wenn λ groß ist, die Verteilungsfunktion bei 0 also steil ist. Natürlich sehen wir das auch formell aus der Verteilungsfunktion weil für alle $\varepsilon > 0$

$$\mathbb{P}_F((0, \varepsilon]) = F(\varepsilon) - F(0) = (1 - e^{-\lambda\varepsilon}) - (1 - e^{-\lambda 0}) = 1 - e^{-\lambda\varepsilon},$$

was monoton wachsend in λ ist.

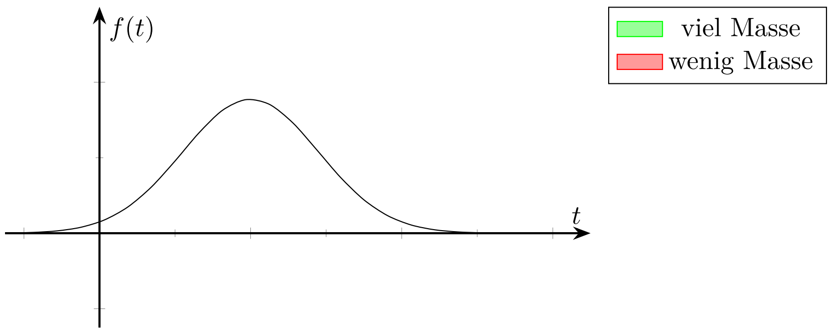
Der Fall mit Dichten: Die obige Diskussion können wir für Verteilungsfunktionen mit Dichten noch konkretisieren. Sei dazu F eine Verteilungsfunktion mit Dichte f, also $F(t) = \int_{-\infty}^{t} f(x)\mathrm{d}x$. Weil F stetig ist, haben alle einpunktigen Mengen keine Masse. Aber wie können wir an f direkt sehen, wo die Masse verteilt ist? Ist f stetig, so folgt aus dem Hauptsatz der Analysis $F'(t) = f(t)$ für alle $t \in \mathbb{R}$. Folglich impliziert ein an der Stelle t großes f ein in t stark wachsendes F und damit viel Masse um t. Andersrum impliziert ein an der Stelle t kleines f ein in t wenig wachsendes F und damit wenig Masse um t. Im Extremfall impliziert natürlich $f = 0$ in $(a, b]$ auch F konstant in $(a, b]$ und damit wird keine Masse auf $(a, b]$ verteilt. Wir merken uns grob

„Hat F eine Dichte, so ist viel Masse dort, wo f groß ist.“

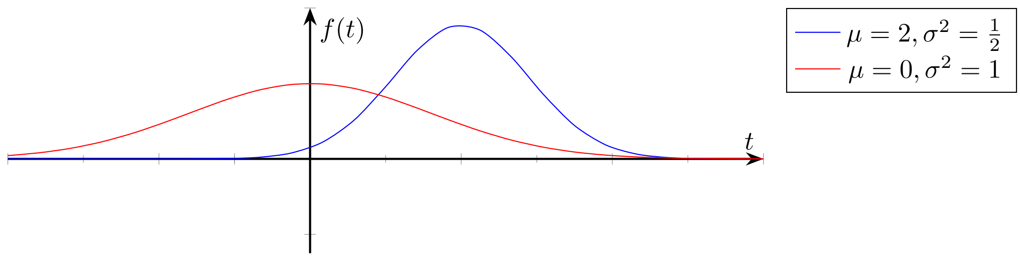
Die nützlichste Interpretation ist durch den Flächeninhalt zwischen Graphen von f und der x-Achse gegeben. Wegen

$$\mathbb{P}_F((a,b]) = F(b) - F(a) = \int_{-\infty}^{b} f(x)\mathrm{d}x - \int_{-\infty}^{a} f(x)\mathrm{d}x = \int_{a}^{b} f(x)\mathrm{d}x,$$

ist die Masse in $(a,b]$ gerade die Fläche unter f zwischen a und b. Dazu ist zu beachten, dass nach Annahme die Gesamtfläche zwischen Graphen und x-Achse 1 ist.
In folgendem Beispiel ist die Dichte von $\mathcal{N}(2,1)$ geplottet:



Wir sehen also, dass viel Masse des Maßes $\mathcal{N}(2,1)$ um die 2 herum verteilt ist und sehr wenig Masse weit weg von der 2 verteilt ist. Der grüne Bereich ist gerade so gewählt, dass dieser Flächeninhalt $\frac{1}{3}$ ist. Ein Drittel der Masse von $\mathcal{N}(2,1)$ liegt also im Schnittbereich des grünen Bereichs mit der x-Achse, sehr nah an der 2. Man sagt, die Verteilung ist um 2 konzentriert.
Wenn wir zwei verschiedene Normalverteilungen vergleichen, sieht es wie im folgenden Beispiel aus:



Der Inhalt der grünen Flächen ist wieder $\frac{1}{3}$, die zugehörigen normalverteilten Maße auf $\mathcal{B}(\mathbb{R})$ haben deshalb Masse $\frac{1}{3}$ im jeweiligen Schnittbereich mit der x-Achse. Wir sehen schon an dem Bild, dass niedrigeres σ dafür sorgt, dass die Verteilung mehr Masse nah an μ hat. Darauf gehen wir in ein paar Wochen noch viel ausführlicher ein.

Kapitel 2

Abbildungen zwischen messbaren Räumen

Vorlesung 8

Bevor wir messbare Abbildungen definieren, erinnern wir kurz an bereits bekannte Konzepte in der Mathematik. Wir betrachten immer Objekte und Abbildungen zwischen Objekten, die auf eine gewisse Art „natürlich“ (strukturerhaltend) sind:

Mengen	Abbildungen
Gruppen	Homomorphismen
Vektorräume	Lineare Abbildungen
Metrische Räume	stetige Abbildungen

Passend dazu diskutieren wir jetzt die strukturerhaltenden Abbildungen zwischen messbaren Räumen, sogenannte messbare Abbildungen.

2.1 Messbare Abbildungen

Definition 2.1.1. Seien $(\Omega, \mathcal{A})$, $(\Omega', \mathcal{A}')$ messbare Räume und $f\colon \Omega \to \Omega'$. f heißt **messbar**, falls Urbilder messbarer Mengen messbar sind; in Formeln

$$A' \in \mathcal{A}' \;\Rightarrow\; f^{-1}(A') \in \mathcal{A}.$$

Es gibt verschiedene Notationen für messbare Abbildungen. Man nutzt synonym

- $f : \Omega \to \Omega'$ ist $(\mathcal{A}, \mathcal{A}')$-messbar,
- $f : (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')$ ist messbar,
- $f : \Omega \to \Omega'$ ist messbar bezüglich $\mathcal{A}$ und $\mathcal{A}'$.

Genau wie Stetigkeit zwischen metrischen Räumen von den gewählten Metriken abhängt, hängt auch die Messbarkeit von den gewählten σ-Algebren ab. Wenn klar ist, welche σ-Algebren gewählt sind, redet man trotzdem einfach nur von messbaren Abbildungen.

Bemerkung 2.1.2. Die Definition der Messbarkeit ist analog zur Stetigkeit zwischen metrischen Räumen, dabei werden messbare Mengen durch offene Mengen ersetzt.

Definition 2.1.3. Ist $(\Omega', \mathcal{A}') = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, dann nennt man eine messbare Abbildung auch **Zufallsvariable** und schreibt X statt f.

Wie bei der Konstruktion von Maßen haben wir das Problem, dass wir alle messbaren Mengen testen müssen. Das ist gerade bei der Borel-σ-Algebra unmöglich, wir kennen die Mengen nicht alle. Zum Glück ist es wie im Kapitel zuvor, es reicht einen Erzeuger zu betrachten:

Proposition 2.1.4. Ist $\mathcal{E}'$ ein Erzeuger von $\mathcal{A}'$ und $f\colon \Omega \to \Omega'$. Dann ist f messbar bzgl. $\mathcal{A}$ und $\mathcal{A}'$ genau dann, wenn

$$A' \in \mathcal{E}' \Rightarrow f^{-1}(A') \in \mathcal{A}.$$

Beweis.

„$\Rightarrow$“: ✓ weil $\mathcal{E}' \subseteq \mathcal{A}'$

„$\Leftarrow$“: Mal wieder der Trick der guten Mengen. Sei dazu

$$\mathcal{F}' := \{A' \in \mathcal{A}' \colon f^{-1}(A') \in \mathcal{A}\},$$

wir zeigen $\mathcal{F}' = \mathcal{A}'$. Nach Annahme gilt $\mathcal{E}' \subseteq \mathcal{F}'$. Wenn $\mathcal{F}'$ eine σ-Algebra ist, dann sind wir fertig, weil dann

$$\mathcal{A}' = \sigma(\mathcal{E}') \subseteq \sigma(\mathcal{F}') = \mathcal{F}' \subseteq \mathcal{A}'$$

und folglich $\mathcal{A}' = \mathcal{F}'$ gilt. Doch wenn man die Definition von $\mathcal{F}'$ anschaut, ist das gerade die Messbarkeit.

Wir überprüfen die definierenden Eigenschaften einer σ-Algebra und können dazu auf elementare Eigenschaften des Urbildes von Abbildungen in Analysis 1 zurückgreifen:

(i) $\emptyset \in \mathcal{F}'$, weil $f^{-1}(\emptyset) = \emptyset \in \mathcal{A}$

(ii) Ist $A' \in \mathcal{F}'$, so gilt

$$f^{-1}((A')^C) = (f^{-1}(A'))^C \in \mathcal{A}$$

weil $A \in \mathcal{F}'$ ist und $\mathcal{A}$ als σ-Algebra abgeschlossen bezüglich Komplementbildung ist.

(iii) Sind $A_1', A_2', ... \in \mathcal{F}'$, so gilt

$$f^{-1}\Big(\bigcup_{n=1}^{\infty} A_n'\Big) = \bigcup_{n=1}^{\infty} f^{-1}(A_n') \in \mathcal{A}$$

weil die Mengen in $\mathcal{F}'$ sind und $\mathcal{A}$ als σ-Algebra abgeschlossen bezüglich Vereinigungen ist.

□

Definition 2.1.5. Ist $f\colon (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)) \to (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^{d'}))$ messbar, so heißt f **Borel-messbar**.

Beispiel 2.1.6.

- Jede stetige Abbildung $f\colon \mathbb{R} \to \mathbb{R}$ ist auch Borel-messbar. Warum? Wir nutzen Proposition 2.1.4, angewandt auf $\sigma(\{O \subseteq \mathbb{R} \colon O \text{ offen}\}) = \mathcal{B}(\mathbb{R})$ mit der Erinnerung, dass Urbilder offener Mengen unter stetigen Abbildungen offen (insbesondere Borel-messbar) sind.

- Indikatorfunktionen

$$\mathbf{1}_A : \Omega \to \mathbb{R},\ \mathbf{1}_A(\omega) = \begin{cases} 1 & : \omega \in A \\ 0 & : \omega \notin A \end{cases}$$

sind $(\mathcal{A}, \mathcal{B}(\mathbb{R}))$-messbar genau dann, wenn A messbar ist. Das zu prüfen ist relativ simpel, weil wir alle möglichen Urbilder direkt hinschreiben können:

$$\mathbf{1}_A^{-1}(B) = \{\omega \in \Omega \,:\, \mathbf{1}_A(\omega) \in B\} = \begin{cases} A & : 1 \in B,\ 0 \notin B \\ A^C & : 1 \notin B,\ 0 \in B \\ \mathbb{R} & : 1, 0 \in B \\ \emptyset & : 1, 0 \notin B \end{cases}.$$

Wie für stetige Abbildungen zeigt man, dass die Verknüpfung messbarer Abbildungen wieder messbar ist. Auch das ist eine kleine Übungsaufgabe.

Bemerkung 2.1.7. Wir erinnern daran, dass

$$\mathcal{B}(\mathbb{R}) = \sigma(\{(-\infty, t] \colon t \in \mathbb{R}\}) = \sigma(\{(-\infty, t) \colon t \in \mathbb{R}\}) = \sigma(\{(a, b) \colon a < b\}).$$

Wegen Proposition 2.1.4 ist deshalb $f \colon \Omega \to \mathbb{R}$ $(\mathcal{A}, \mathcal{B}(\mathbb{R}))$-messbar genau dann, wenn

$$f^{-1}((-\infty, t]) = \{\omega \in \Omega \colon f(\omega) \leq t\} =: \{f \leq t\} \in \mathcal{A}$$

für alle $t \in \mathbb{R}$. Die Kurzschreibweise $\{f \leq t\}$ ist etwas ungewohnt, wird ab jetzt aber oft genutzt. Analog ist auch f messbar genau dann, wenn

$$f^{-1}((-\infty, t)) = \{\omega \in \Omega \colon f(\omega) < t\} =: \{f < t\} \in \mathcal{A}$$

für alle $t \in \mathbb{R}$ oder

$$f^{-1}((a, b)) = \{\omega \in \Omega \colon f(\omega) \in (a, b)\} =: \{f \in (a, b)\} \in \mathcal{A}$$

für alle reellen Zahlen $a < b$. Analog kann man auch halb-offene Intervalle, abgeschlossene Mengen, kompakte Mengen, offene Mengen und so weiter nutzen, jeder Erzeuger von $\mathcal{B}(\mathbb{R})$ gibt eine Möglichkeit um Messbarkeit zu prüfen.

Definition 2.1.8. Sei $f : \Omega \to \Omega'$ für einen messbaren Raum $(\Omega', \mathcal{A}')$ ist. Dann ist die Menge aller Urbilder

$$\mathcal{A} := \left\{f^{-1}(A') \colon A' \in \mathcal{A}'\right\}$$

eine σ-Algebra und $\mathcal{A}$ natürlich ist die kleinste σ-Algebra auf Ω, für die f $(\mathcal{A}, \mathcal{A}')$-messbar. Wir nennen die σ-Algebra $\mathcal{A}$ auch die **von f erzeugte σ-Algebra** und schreiben $\sigma(f)$.

Schreibt euch das nächste Beispiel einmal selber hin:

Beispiel 2.1.9.

- $\sigma(\mathbf{1}_A) = \{\emptyset, \Omega, A, A^C\}$
- Sei $f \colon \mathbb{R} \to \mathbb{R}$, $f \equiv c$ eine konstante Funktion, dann ist $\sigma(f) = \{\emptyset, \mathbb{R}\}$.

Die nächste Definition verallgemeinert die erzeugte σ-Algebra von einer auf beliebig viele Funktionen. Der Begriff ist für die Stochastik 1 nicht besonders wichtig, ist für die Finanzmathematik aber essentiell. Mit der Definition wird das Konzept von Information mathematisiert.

Definition 2.1.10. Seien $(\Omega_i', \mathcal{A}_i')$ messbare Räume und $f_i \colon \Omega \to \Omega_i'$, $i \in I$, für eine beliebige Indexmenge. Dann ist

$$\sigma(f_i,\ i \in I) := \sigma\Big(\bigcup_{i \in I} \sigma(f_i)\Big) = \sigma(\{f_i^{-1}(A_i) \colon A_i' \in \mathcal{A}_i',\ i \in I\})$$

die kleinste σ-Algebra auf Ω, bezüglich derer alle f_i messbar sind. Man spricht auch hier von der **von den $f_i, i \in I$, erzeugte σ-Algebra**.

2.2 Bildmaße oder „push-forward“ eines Maßes

Wir nutzten die Messbarkeit einer $(\mathcal{A}, \mathcal{A}')$-messbaren Abbildung $f : \Omega \to \Omega'$, um ein Maß μ auf $\mathcal{A}$ auf ein Maß μ_f auf $\mathcal{A}'$ rüberzuschieben (deshalb „push-forward“). In dem Stochastikteil werden wir noch sehen, dass der push-forward extrem wichtig ist.

Satz 2.2.1. Sei $f : \Omega \to \Omega'$ $(\mathcal{A}, \mathcal{A}')$-messbar und μ ein Maß auf $\mathcal{A}$. Dann ist

$$\mu_f(B) := \mu\left(f^{-1}(B)\right), \quad B \in \mathcal{A}',$$

ein Maß auf $\mathcal{A}'$. Dieses Maß heißt „Bildmaß" oder „push-forward" von f.

Beweis. μ_f ist wohldefiniert weil f messbar ist und daher $f^{-1}(B) \in \mathcal{A}$ gilt. Auf $\mathcal{A}$ ist μ definiert, also macht die Definition von μ_f Sinn. Die Positivität von μ_f folgt natürlich direkt aus der Positivität von μ. Checken wir noch die zwei definierenden Eigenschaften eines Maßes:

(i) $\mu_f(\emptyset) = \mu\left(f^{-1}(\emptyset)\right) = \mu(\emptyset) = 0$

(ii) Seien $B_1, B_2, ... \in \mathcal{A}'$ paarweise disjunkt, dann folgt aus der Definition und den Maßeigenschaften von μ

$$\begin{aligned}
\mu_f\Big(\dot{\bigcup_{k=1}^\infty} B_k\Big) &\overset{\text{Def.}}{=} \mu\Big(f^{-1}\Big(\dot{\bigcup_{k=1}^\infty} B_k\Big)\Big) \\
&\overset{\text{Urbild}}{=} \mu\Big(\dot{\bigcup_{k=1}^\infty} f^{-1}(B_k)\Big) \\
&\overset{\mu \text{ Maß}}{=} \sum_{k=1}^\infty \mu\Big(f^{-1}(B_k)\Big) \overset{\text{Def.}}{=} \sum_{k=1}^\infty \mu_f(B_k).
\end{aligned}$$

Damit ist μ_f auch σ-additiv.

□

Beispiel 2.2.2. Sei $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x + a$. f ist Borel-messbar weil f stetig ist. Sei $\mu := \lambda$ das Lebesgue-Maß auf $\mathcal{B}(\mathbb{R})$, was ist dann der push-forward μ_f? μ_f ist laut Satz 2.2.1 ein Maß, aber welches? Es gilt tatsächlich, dass der Push-forward das gleiche Maß ist: $\mu_f = \lambda$.
Warum gilt das? Berechnen wir dazu μ_f auf einem $\cap$-stabilen Erzeuger von $\mathcal{B}(\mathbb{R})$:

$$\mu_f((c,d]) \overset{\text{Def.}}{=} \mu(f^{-1}((c,d])) = \lambda((c-a, d-a]) = (d-a) - (c-a) = d - c = \lambda((c,d]).$$

Weil $\mathcal{E} = \{(c,d] : c < d\}$ $\cap$-stabil ist mit $\sigma(\mathcal{E}) = \mathcal{B}(\mathbb{R})$, gilt aufgrund von Folgerung 1.2.13 auch $\lambda = \mu_f$ (wir wählen dabei $E_n = (-n, n]$). Weil a beliebig war, gilt also

$$\lambda(B) = \lambda(B + a), \quad \forall a \in \mathbb{R}, B \in \mathcal{B}(\mathbb{R}),$$

wobei $B + a := \{b + a : b \in B\}$ die um a verschobene Menge ist. Man sagt, das Lebesgue-Maß ist **translationsinvariant**, Verschiebungen von Mengen ändert ihr Maß (die „Größe") nicht. Diese Eigenschaft gilt natürlich nicht für alle Maße. Mehr noch, bis auf triviale Modifikationen (Konstanten addieren) ist das Lebesgue-Maß das einzige translationsinvariante Maß auf $\mathcal{B}(\mathbb{R})$.

2.3 Messbare numerische Funktionen

Wir nutzen wie in Kapitel 1 die erweiterte Zahlengerade $\overline{\mathbb{R}} = [-\infty, +\infty]$. Dabei nutzen wir die definierten „Rechenregeln" aus Kapitel 1 und auch die Konvergenzen am Rand:

$$a_n \to +\infty, n \to \infty, \qquad \text{und} \qquad a_n \to -\infty, n \to \infty,$$

wie in Analysis 1 definiert. Oft schreiben wir ∞ statt $+\infty$.

Definition 2.3.1. Auf $\overline{\mathbb{R}}$ definieren wir die erweiterte Borel-σ-Algebra:

$$\mathcal{B}(\overline{\mathbb{R}}) := \big\{B \subseteq \overline{\mathbb{R}} : B \cap \mathbb{R} \in \mathcal{B}(\mathbb{R})\big\}.$$

Kurz überlegen zeigt uns, dass $\mathcal{B}(\overline{\mathbb{R}})$ folgende Mengen enthält: alle $B \in \mathcal{B}(\mathbb{R})$, sowie $B \cup \{+\infty\}$, $B \cup \{-\infty\}$ und $B \cup \{-\infty, +\infty\}$.

Definition 2.3.2. Für einen messbaren Raum $(\Omega, \mathcal{A})$ heißt $f\colon \Omega \to \overline{\mathbb{R}}$ **messbare numerische Funktion**, falls f $(\mathcal{A}, \mathcal{B}(\overline{\mathbb{R}}))$-messbar ist.

In der Stochastik 1 spielen numerische Funktionen noch keine besonders wichtige Rolle. Ihr solltet euch nicht erschrecken lassen, bei (fast) allen Argumenten spielt es keine Rolle, ob eine Funktion reell oder numerisch ist. Numerische Funktionen sind einfach nur eine etwas größere Klasse von Funktionen, die reelle Funktionen enthalten. Gewöhnt euch einfach direkt daran, dass unsere messbaren Funktionen auch die Werte $+\infty$ oder $-\infty$ annehmen dürfen.

Bemerkung 2.3.3.

(i) Jede $(\mathcal{A}, \mathcal{B}(\mathbb{R}))$-messbare Funktion $f\colon \Omega \to \mathbb{R}$ ist auch eine messbare numerische Funktion, denn $f^{-1}(A \cup B) = f^{-1}(B) \in \mathcal{A}$ für alle $B \in \mathcal{B}(\mathbb{R})$ und $A \in \{\{+\infty\}, \{-\infty\}, \{+\infty, -\infty\}\}$.

(ii) Aussagen für messbare reelle Funktionen gelten ganz analog für messbare numerische Funktionen. So gilt etwa: $f\colon \Omega \to \overline{\mathbb{R}}$ ist $(\mathcal{A}, \mathcal{B}(\overline{\mathbb{R}}))$-messbar genau dann, wenn $\{f \leq t\} \in \mathcal{A}$ für alle $t \in \overline{\mathbb{R}}$. Das folgt auch aus Proposition 2.1.4 weil $\mathcal{E} = \{[-\infty, t] \colon t \in \overline{\mathbb{R}}\}$ die σ-Algebra $\mathcal{B}(\overline{\mathbb{R}})$ erzeugt (überlegt mal, warum das stimmt).

Definition 2.3.4. Für $a, b \in \overline{\mathbb{R}}$ definieren wir

$$a \wedge b := \min\{a, b\} \quad \text{und} \quad a \vee b := \max\{a, b\}$$

sowie

$$a^+ := \max\{0, a\} \quad \text{und} \quad a^- := -\min\{0, a\}.$$

Für numerische Funktionen werden entsprechend punktweise $f \wedge g$, $f \vee g$, f^+, f^- definiert. f^+ heißt **Positivteil** von f und f^- **Negativteil** von f.

Beachte: Postivteil und Negativteil sind beide positiv aufgrund des zusätzlichen Minus in der Definition des Negativteils.
Es gelten direkt aus der Definition folgende wichtige Identitäten

$$f = f^+ - f^- \quad \text{und} \quad |f| = f^+ + f^-,$$

die uns zeigen, weshalb es oft reicht f^+ und f^- zu untersuchen.

Lemma 2.3.5. Sind $f, g\colon \Omega \to \overline{\mathbb{R}}$ $(\mathcal{A}, \mathcal{B}(\overline{\mathbb{R}}))$-messbar, so sind die Mengen

$$\{f < g\}, \quad \{f \leq g\}, \quad \{f = g\} \quad \text{und} \quad \{f \neq g\}$$

messbar, also in $\mathcal{A}$.

Beweis. Der Trick ist es, die Mengen als abzählbare Vereinigungen, Komplemente, Schnitte, etc. von messbaren Mengen zu schreiben. Weil f und g messbar sind, führen wir also auf Urbilder offener Mengen von f und g zurück. Als erstes schreiben wir

$$\{f < g\} \overset{\text{Trick!}}{=} \bigcup_{t \in \mathbb{Q}} \{f < t < g\} = \underbrace{\bigcup_{t \in \mathbb{Q}} \underbrace{\{f < t\}}_{\in \mathcal{A}} \cap \underbrace{\{t < g\}}_{\in \mathcal{A}}}_{\in \mathcal{A}}.$$

Der wesentliche Trick war natürlich die erste Gleichheit. Genauso zeigt man auch $\{f > g\} \in \mathcal{A}$. Weil $\{f = g\} = (\{f < g\} \cup \{f > g\})^C$ und $\{f \neq g\} = \{f = g\}^C$ gelten, sind auch die letzten beiden Mengen in $\mathcal{A}$. Die zweite Menge schreiben wir als $\{f \leq g\} = \{f < g\} \cup \{f = g\}$, die rechte Seite ist in $\mathcal{A}$. $\square$

Lemma 2.3.6. Sind $f, g : \Omega \to \overline{\mathbb{R}}$ $(\mathcal{A}, \mathcal{B}(\overline{\mathbb{R}}))$-messbar, so sind auch $f+g$, αf für $\alpha \in \mathbb{R}$, $f \cdot g$, $f \wedge g$, $f \vee g$, und $|f|$ messbar.

Beweis. Tricks aus dem letzten Beweis ausprobieren, und in Übungen/Übungsaufgaben üben! □

Eine kleine Warnung: Wir müssen beim Addieren von numerischen Funktionen aufpassen, dass die Addition wohldefiniert ist. Es darf niemals $+\infty + (-\infty)$ auftauchen, das ist nicht definiert worden. Man sollte also immer schreiben, „$f+g$ (wenn die Addition wohldefiniert ist)“. Weil solche Probleme in der Stochastik 1 keine ernsthafte Rolle spielen, sind wir hier bewusst etwas unsauber, um nicht von den wichtigsten Punkten abzulenken.

Auch sehr wichtig ist, dass punktweise Grenzwerte von Folgen messbarer numerischer Funktionen wieder messbar sind:

Proposition 2.3.7. Es sei $f_1, f_2, ... : \Omega \to \overline{\mathbb{R}}$ eine Folge $(\mathcal{A}, \mathcal{B}(\overline{\mathbb{R}}))$-messbarer numerischer Funktionen.

(i) Dann sind auch die punktweise definierten Funktionen

- $g_1(\omega) := \inf_{n\in\mathbb{N}} f_n(\omega), \quad \omega \in \Omega,$
- $g_2(\omega) := \sup_{n\in\mathbb{N}} f_n(\omega), \quad \omega \in \Omega,$
- $g_3(\omega) := \limsup_{n\to\infty} f_n(\omega), \quad \omega \in \Omega,$
- $g_4(\omega) := \liminf_{n\to\infty} f_n(\omega), \quad \omega \in \Omega,$

messbare numerische Funktionen. Beachte: Weil wir über numerische Funktionen reden, sind alle Ausdrücke wohldefiniert, die Werte $+\infty$ und $-\infty$ dürfen auftauchen.

(ii) Existieren die Grenzwerte in $\overline{\mathbb{R}}$ für alle $\omega \in \Omega$, so ist auch die punktweise definierte Funktion

$$g(\omega) := \lim_{n\to\infty} f_n(\omega), \quad \omega \in \Omega,$$

messbar.

Beweis. Der Beweis wird in der großen Übung diskutiert, hier nur für g_1. Wegen Bemerkung 2.1.7 reicht es, für alle $t \in \mathbb{R}$, $\{g_1 < t\} \in \mathcal{A}$ zu zeigen. Die Mengen $\{g_1 < t\}$ werden wieder geschrieben als abzählbare Vereinigungen, Komplemente, Schnitte, etc. von aufgrund der Voraussetzung messbaren Mengen:

$$\begin{aligned}\{g_1 < t\} &= \Big\{\omega \in \Omega : \inf_{n\in\mathbb{N}} f_n(\omega) < t\Big\}\\ &= \{\omega \in \Omega : f_n(\omega) < t \text{ für ein n} \in \mathbb{N}\}\\ &= \bigcup_{n\in\mathbb{N}} \{\omega \in \Omega : f_n(\omega) < t\}\\ &= \underbrace{\bigcup_{n\in\mathbb{N}} \underbrace{\{f_n < t\}}_{\in\mathcal{A}}}_{\in\mathcal{A}}.\end{aligned}$$

Für $g_2, ..., g_4$ muss man sich überlegen, was $\{g_i < t\}$ eigentlich bedeutet und das dann in abzählbare Vereinigungen, Komplemente, Schnitte, etc. messbarer Mengen der Form $\{f_n \in ...\}$ umschreiben. Probiert es aus! Jetzt ist ein guter Moment zu wiederholen, wie $\liminf$, $\limsup$ definiert sind. □

An dieser Stelle ist noch nicht so klar, warum Messbarkeit nützlich ist. Die gerade gezeigten Aussagen sind der Grund, weshalb die im Anschluss zu entwickelnde Lebesgue Integrationstheorie so erfolgreich ist: Alle möglichen Manipulationen mit messbaren Funktionen bleiben messbar.

Kapitel 3

Integrationstheorie

Vorlesung 9

Im Folgenden entwickeln wir die Integrationstheorie im Sinne von Henry Lebesgues. An und für sich hat das nichts mit Stochastik zu tun und gehört eher in den Bereich der Analysis. Im Sinne der abstrakten Modellierung eines zufälligen Experimentes durch einen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ wollen wir aber Lebesgue Integrale nutzen, um Begriffe wie Erwartungswert und Varianz als Integrale über $(\Omega, \mathcal{A}, \mathbb{P})$ definieren. Damit kann man dann in der Stochastik die Massenverteilung zufälliger Experimente genauer untersuchen.

3.1 Das (allgemeine) Lebesgue Integral

In diesem Abschnitt werden wir Integrale der Form $\int_\Omega f \mathrm{d}\mu$ für beliebige Maßräume $(\Omega, \mathcal{A}, \mu)$ und messbare numerische Funktionen $f : \Omega \to \bar{\mathbb{R}}$ definieren. Das Maß kann endlich sein (z. B. ein Wahrscheinlichkeitsmaß) oder unendlich sein (z. B. das Lebesguemaß auf $\mathcal{B}(\mathbb{R})$).

Bevor wir mit der Konstruktion starten, diskutieren wir ganz kurz den Zusammenhang zum Riemann Integral, das ihr vermutlich aus der Schule oder den Analysis Vorlesungen kennt. Dort habt ihr für reelle Funktionen Integrale $\int_a^b f(x)\,\mathrm{d}x$ definiert, indem Treppenfunktionen (stückweise konstant auf einer Zerlegung von $[a, b]$ in kleine Intervalle) über und unter den Graphen von f gelegt wurden. Wenn die Treppenfunktionen von oben und unten immer feiner an f angenähert werden und dabei die Integrale (Ober- und Untersummen) im Grenzwert gleich sind, so heißt f Riemann integrierbar und das Integral von f ist als dieser Grenzwert definiert (wer alles vergessen hat, kann mal schnell die Bildchen bei Wikipedia zum Riemann Integral anschauen). Die Interpretation des Integrals als Flächeninhalt wird dadurch visuell klar. Anschließend habt ihr das uneigentliche Riemann Integral $\int_\mathbb{R} f(x)\,\mathrm{d}x$ als Grenzwert des (eigentlichen) Riemann Integrals $\lim_{n\to\infty} \int_{-n}^n f(x)\,\mathrm{d}x$ definiert, falls der Grenzwert existiert. Wenn ihr also eigentliche Riemann Integrale durch Stammfunktionen, partielle Integration oder Substitution berechnen könnt, so könnt ihr auch uneigentliche Riemann Integrale berechnen.
Wenn wir jetzt für $f : \Omega \to \mathbb{R}$ statt für $f : \mathbb{R} \to \mathbb{R}$ genauso vorgehen wollen, haben wir ein Problem: Wie zerlegen wir Ω in kleine Intervalle? Das geht nicht einfach so, Ω ist schließlich eine völlig beliebige Menge! Was ist aber in beiden Fällen gleich? Der Bildbereich! Der Trick beim Lebesgue Integral ist deshalb, nicht das Urbild in Intervalle zu zerlegen, sondern den Bildbereich in Intervalle zu zerlegen! Warum dafür gerade messbare Funktionen geeignet sind, wird in Satz 3.1.6 deutlich werden. Die Idee von Lebesgue den Bildbereich zu zerteilen, wird es uns daher später erlauben, Erwartungswerte, Varianzen, etc. für beliebige zufällige Experimente zu definieren.

Um leichter zu folgen, kann es nützliche sein, den Spezialfall $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$ im Kopf zu halten, denn da können wir besser zeichnen. Wir schreiben in dem Fall statt $\int_\mathbb{R} f\,\mathrm{d}\lambda$ auch $\int_\mathbb{R} f(x)\,\mathrm{d}x$, um klar zu stellen, dass das Lebesgue Integral für viele „nette“ Integranden $f : \mathbb{R} \to \mathbb{R}$ das gleiche ist, wie das (uneigentliche) Riemann Integral. Merkt euch für später schon mal, dass beispielsweise für

- nicht-negative stückweise stetige Integranden,

- stückweise stetige Integranden, die außerhalb eines Intervalls 0 sind,

das neue Lebesgue Integral und das schon bekannte uneigentliche Riemann Integral gleich sind. Später wird das nützlich sein, weil ihr dann die Rechenregeln der Analysis 1 (oder Schule) nutzen könnt, um $\int_{\mathbb{R}} f \,\mathrm{d}\lambda$ als Grenzwert von $\int_{-n}^{n} f(x)\,\mathrm{d}x$ auszurechnen. Alternativ könnten wir die Rechenregeln aus der Analysis nochmal für das Lebesgue Integral nachrechnen, aber das wäre vielleicht etwas langweilig.

Genug der Vorrede, kommen wir nun zum Lebesgue Integral:

Definition 3.1.1. Eine messbare Abbildung $f\colon \Omega \to \overline{\mathbb{R}}$ heißt **einfach** (alternativ **elementar**, manchmal auch **Treppenfunktion**), falls f nur endlich viele Werte annimmt. Wir definieren auch noch

$$\begin{aligned} \mathcal{E} &= \{f : \Omega \to \bar{\mathbb{R}} \mid f \text{ einfache Funktion}\}, \\ \mathcal{E}^+ &= \{f : \Omega \to \bar{\mathbb{R}} \mid \; f \text{ einfache Funktion, } f \geq 0\}. \end{aligned}$$

Eine Darstellung der Form

$$f = \sum_{k=1}^{n} \alpha_k \mathbf{1}_{A_k} \tag{3.1}$$

nennen wir disjunkte Darstellung, wenn $\alpha_1, ..., \alpha_n \in \overline{\mathbb{R}}$ und $A_1, ..., A_n \in \mathcal{A}$ paarweise disjunkt sind. Nimmt eine einfache Funktion die Werte $\alpha_1, ..., \alpha_n$ an, so gilt (3.1) zum Beispiel mit den messbaren Mengen

$$A_k = \{f = \alpha_k\} = \{\omega\colon f(\omega) = \alpha_k\} = f^{-1}([\alpha_k, \alpha_k]) \in \mathcal{A}.$$

Bemerkung. Wenn wir von einfachen Funktionen sprechen, meinen wir also immer, dass entweder f endlich viele Werte annimmt, oder f die obige Darstellung als Summe von Indikatorfunktionen hat. Meistens nutzen wir aber die disjunkten Darstellungen weil die für Integrale benötigt werden.

Disjunkte Darstellungen messbarer Funktionen sind nicht eindeutig, z. B. gilt

$$\mathbf{1}_{[-2,-1]} + 2 \cdot \mathbf{1}_{[1,2]} = \mathbf{1}_{[-2,-3/2]} + \mathbf{1}_{(-3/2,-1]} + 2 \cdot \mathbf{1}_{[1,2]}.$$

Wir definieren im Folgenden das Integral nicht-negativer einfacher Funktionen, dann durch Approximation das Integral nicht-negativer messbarer Funktionen und schließlich durch die Zerlegen $f = f^+ - f^-$ das Integral beliebiger messbarer Funktionen.

Integrale nicht-negativer einfacher Funktionen

Definition 3.1.2. Für $f \in \mathcal{E}^+$ definiert man

$$\int_\Omega f \,\mathrm{d}\mu := \sum_{k=1}^{n} \alpha_k \mu(A_k) \in [0, +\infty],$$

wenn f die disjunkte Darstellung $f = \sum_{k=1}^{n} \alpha_k \mathbf{1}_{A_k}$ hat. $\int_\Omega f \mathrm{d}\mu$ heißt Integral von f bezüglich μ, f heißt Integrand. Weil $\alpha_k = +\infty$ sowie $\mu(A_k) = +\infty$ möglich sind, muss festgelegt werden, wie $+$ und $\cdot$ mit ∞ geht. Siehe dazu die Definition in Abschnitt 1.1.

Beispiel. Weil das Integral den „Flächeninhalt“ zwischen Graphen und Achse beschreiben soll, sind die Rechenregeln $+$ und $\cdot$ mit ∞ durchaus sinnvoll definiert worden:

$$\int_{\mathbb{R}} 0 \cdot \mathbf{1}_{\mathbb{R}} \, \mathrm{d}\lambda = 0 \cdot \lambda(\mathbb{R}) = 0 \cdot (+\infty) = 0,$$

$$\int_{\mathbb{R}} \mathbf{1}_{\mathbb{R}} \, \mathrm{d}\lambda = 1 \cdot \lambda(\mathbb{R}) = 1 \cdot (+\infty) = +\infty,$$

$$\int_{\mathbb{R}} (+\infty) \cdot \mathbf{1}_{[a,b]} \, \mathrm{d}\lambda = +\infty \cdot \lambda([a,b]) = +\infty \cdot (b-a) = +\infty,$$

$$\int_{\mathbb{R}} 3 \cdot \mathbf{1}_{[0,1]} \, \mathrm{d}\lambda = 3 \cdot \lambda([0,1]) = 3.$$

Was sollten die Integrale auch sonst sein?

Rechnen wir noch nach, dass das Integral einer nicht-negativen einfachen Funktion nicht von der disjunkten Darstellung abhängt. Weil es verschiedene disjunkte Darstellungen für die gleiche Funktion gibt, würde die Definition sonst keinen Sinn machen.

Lemma 3.1.3. Es gelte

$$\sum_{k=1}^{n} \alpha_k \mathbf{1}_{A_k} = \sum_{l=1}^{m} \beta_l \mathbf{1}_{B_l}$$

mit $\alpha_k, \beta_l \geq 0$ und paarweise disjunkten $A_1, ..., A_n, B_1, ..., B_m$, so gilt

$$\sum_{k=1}^{n} \alpha_k \mu(A_k) = \sum_{l=1}^{m} \beta_l \mu(B_l).$$

Beweis. Ohne Einschränkung der Allgemeinheit seien alle $\alpha_k, \beta_l \neq 0$. Wegen $\dot\bigcup_{k=1}^{n} A_k = \{f > 0\} = \dot\bigcup_{l=1}^{m} B_k$ und der σ-Additivität von μ gilt dann

$$\begin{aligned}\sum_{k=1}^{n} \alpha_k \mu(A_k) &\overset{\sigma\text{-add.}}{=} \sum_{k=1}^{n} \alpha_k \sum_{l=1}^{m} \mu(A_k \cap B_l) = \sum_{k=1}^{n} \sum_{l=1}^{m} \alpha_k \mu(A_k \cap B_l) \\ &\overset{(\star)}{=} \sum_{l=1}^{m} \sum_{k=1}^{n} \beta_l \mu(B_l \cap A_k) = \sum_{l=1}^{m} \beta_l \mu(B_l).\end{aligned}$$

$(\star)$ gilt, weil entweder $\mu(A_k \cap B_l) = 0$ oder $\mu(A_k \cap B_l) > 0$ gilt. Im ersten Fall gilt $\alpha_k \mu(A_k \cap B_l) = 0 = \beta_l \mu(A_k \cap B_l)$ trivialerweise, im zweiten Fall impliziert $\mu(A_k \cap B_l) > 0$ schon $A_k \cap B_l \neq \emptyset$ und damit $\alpha_k = \beta_l$ weil die beiden Darstellungen disjunkt sind. $\square$

Lemma 3.1.4. Für $f, g \in \mathcal{E}^+$, $\alpha \geq 0$ und $A \in \mathcal{A}$ gelten

(i) $\mathbf{1}_A \in \mathcal{E}^+$ und $\int_\Omega \mathbf{1}_A \, \mathrm{d}\mu = \mu(A)$.

(ii) $\alpha f \in \mathcal{E}^+$ und $\int_\Omega \alpha f \, \mathrm{d}\mu = \alpha \int_\Omega f \, \mathrm{d}\mu$,

(iii) $f + g \in \mathcal{E}^+$ und $\int_\Omega (f+g) \, \mathrm{d}\mu = \int_\Omega f \, \mathrm{d}\mu + \int_\Omega g \, \mathrm{d}\mu$,

(iv) $f \leq g \Rightarrow \int_\Omega f \, \mathrm{d}\mu \leq \int_\Omega g \, \mathrm{d}\mu$.

Beweis.

(i) ✓

(ii) αf nimmt auch nur endlich viele Werte an (ist also eine einfache Funktion) und hat die disjunkte Darstellung

$$\alpha f = \alpha \sum_{k=1}^{n} \alpha_k \mathbf{1}_{A_k} = \sum_{k=1}^{n} (\alpha\alpha_k)\mathbf{1}_{A_k}.$$

Das Integral berechnet sich aus der Definition:

$$\int_\Omega (\alpha f)\,\mathrm{d}\mu \overset{\text{Def.}}{=} \sum_{k=1}^{n} (\alpha\alpha_k)\,\mu(A_k) = \alpha \sum_{k=1}^{n} \alpha_k \mu(A_k) \overset{\text{Def.}}{=} \alpha \int_\Omega f\,\mathrm{d}\mu$$

(iii) Wir nehmen an, dass f und g die disjunkten Darstellungen

$$f = \sum_{k=1}^{n} \alpha_k \mathbf{1}_{A_k} \quad \text{und} \quad g = \sum_{l=1}^{m} \beta_l \mathbf{1}_{B_l}$$

haben. Ohne Einschränkung gelte $\dot{\bigcup}_{k=1}^{n} A_k = \Omega = \dot{\bigcup}_{l=1}^{m} B_k$. Wäre das nicht der Fall, so würden wir $A_{n+1} := (\bigcup_{k=1}^n A_k)^C$ und $\alpha_{n+1} = 0$ wählen (analog für g). Damit ist dann wegen $1 = \mathbf{1}_\Omega = \sum_{k=1}^n \mathbf{1}_{A_k} = \sum_{l=1}^m \mathbf{1}_{B_k}$ und $\mathbf{1}_{A_k} \cdot \mathbf{1}_{B_l} = \mathbf{1}_{A_k \cap B_l}$ auch

$$\begin{aligned} f + g &= \sum_{k=1}^{n} \alpha_k \mathbf{1}_{A_k} + \sum_{l=1}^{m} \beta_l \mathbf{1}_{B_l} \\ &= \sum_{k=1}^{n} \alpha_k \sum_{l=1}^{m} \mathbf{1}_{A_k\cap B_l} + \sum_{l=1}^{m} \beta_l \sum_{k=1}^{n} \mathbf{1}_{A_k\cap B_l} \\ &= \sum_{k=1}^{n}\sum_{l=1}^{m} (\alpha_k + \beta_l)\mathbf{1}_{A_k\cap B_l}. \end{aligned}$$

Damit haben wir eine disjunkte Darstellung für die einfache Funktion $f+g$ und es gilt

$$\begin{aligned} \int_\Omega (f+g)\,\mathrm{d}\mu &\overset{\text{Def.}}{=} \sum_{k=1}^{n}\sum_{l=1}^{m} (\alpha_k + \beta_l)\mu(A_k \cap B_l) \\ &\overset{\sigma\text{-add.}}{=} \sum_{k=1}^{n} \alpha_k \mu\Big(\dot{\bigcup_{l=1}^{m}} (A_k \cap B_l)\Big) + \sum_{l=1}^{l} \beta_l \mu\Big(\dot{\bigcup_{k=1}^{n}} (A_k \cap B_l)\Big) \\ &= \sum_{k=1}^{n} \alpha_k \mu(A_k) + \sum_{l=1}^{l} \beta_l \mu(B_l) \\ &\overset{\text{Def.}}{=} \int_\Omega f \mathrm{d}\mu + \int_\Omega g \mathrm{d}\mu, \end{aligned}$$

und damit die Behauptung.

(iv) Monotonie ✓, folgt direkt aus der Definition.

□

Integral nicht-negativer messbarer numerischer Funktionen

Weiter geht's für nicht-negative Integranden. Wir definieren zunächst einen sofort wohldefinierten Ausdruck und zeigen danach, dass dies auch der Grenzwert beliebiger wachsender Folgen von unten ist.

Definition 3.1.5. Für $(\mathcal{A}, \mathcal{B}(\overline{\mathbb{R}}))$-messbares $f : \Omega \to \overline{\mathbb{R}}$ mit $f \geq 0$ definieren wir

$$\int\limits_\Omega f \,\mathrm{d}\mu := \sup\left\{ \int\limits_\Omega g \,\mathrm{d}\mu \,:\, 0 \leq g \leq f,\ g \in \mathcal{E}^+ \right\}.$$

$\int_\Omega f \mathrm{d}\mu$ heißt wieder **Integral von f bezüglich μ**, f heißt **Integrand**. Wie bei nicht-negativen einfachen Funktionen ist $\int_\Omega f \,\mathrm{d}\mu = +\infty$ ausdrücklich erlaubt!

Jetzt wollen wir diese komplizierte Definition (wie soll man damit irgendwas zeigen?) durch eine handlichere äquivalente Darstellung ersetzen.

Satz 3.1.6. **[Darum sind messbare Funktionen so wichtig!!!]** Für jede nicht-negative messbare numerische Funktion existiert eine wachsende Folge von Treppenfunktionen $(f_n)_{n\in\mathbb{N}} \subseteq \mathcal{E}^+$ mit $f_n \uparrow f$ punktweise für $n \to \infty$.

Beweis. Wir definieren

$$f_n = \underbrace{\sum_{k=0}^{n\cdot 2^n-1} \frac{k}{2^n} \mathbf{1}_{\underbrace{f^{-1}\left(\left[\frac{k}{2^n}, \frac{k+1}{2^n}\right)\right)}_{\in\mathcal{A}}} + n\mathbf{1}_{\underbrace{f^{-1}([n,+\infty])}_{\in\mathcal{A}}}}_{\text{messbar}}.$$

Fürs bessere Verständnis zeichne man die Folge f_n für das Beispiel $f : \mathbb{R} \to \overline{\mathbb{R}}$ mit $f(x) = +\infty \cdot \mathbf{1}_{(-\infty,0]} + \frac{1}{x} \cdot \mathbf{1}_{(0,+\infty)}$ hin! Weil f messbar ist, sind die A_k messbare Mengen. Also sind die f_n einfache Funktionen. Aufgrund der Definition gelten sofort die geforderten Eigenschaften:

- $0 \leq f_n \leq f$ für alle $n \in \mathbb{N}$.
- Die Folge (f_n) ist punktweise wachsend.
- Die Folge (f_n) konvergiert punktweise gegen f.

□

Lemma 3.1.7. **[Montone Konvergenz Theorem (MCT) für einfache Funktionen]** Sei $(f_n) \subseteq \mathcal{E}^+$ mit $f_n \uparrow f$, $n \to \infty$, für eine nicht-negative messbare numerische Funktion f. Dann gilt

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

wobei in der Gleichheit $+\infty = +\infty$ möglich ist.

Für monoton wachsende Folgen einfacher Funktionen darf der Limes also in das Integral getauscht werden.

Beweis. Die Folge $(\int_\Omega f_n \,\mathrm{d}\mu)_{n\in\mathbb{N}}$ wächst (Monotonie des Integrals für einfache Integranden) und konvergiert also in $[0, +\infty]$.

„$\leq$“: Folgt direkt aus der Definition

$$\int_\Omega f \,\mathrm{d}\mu = \sup\left\{ \int_\Omega g \,\mathrm{d}\mu \colon g \leq f,\ g \in \mathcal{E}^+ \right\},$$

weil das Supremum einer Menge eine obere Schranke der Menge ist und $f_n \leq f$ nach Voraussetzung.

„$\geq$“: Wir behaupten: Ist $g \in \mathcal{E}^+$ mit $g \leq f$, so gilt

$$\lim_{n\to\infty} \int_\Omega f_n \, \mathrm{d}\mu \geq \int_\Omega g \, \mathrm{d}\mu. \tag{3.2}$$

Weil das Supremum einer Menge M die *kleinste* obere Schranke ist, sind wir dann fertig weil aufgrund von (3.2) auch $\lim_{n\to\infty} \int_\Omega f_n \, \mathrm{d}\mu$ eine obere Schranke ist.

Warum gilt die Behauptung? Sei $\varepsilon \in (0,1)$ beliebig und sei

$$g = \sum_{k=1}^r \gamma_k \mathbf{1}_{C_k} \in \mathcal{E}^+ \quad \text{mit} \quad g \leq f.$$

Wegen $f_n \uparrow f$ gilt $A_n \uparrow \Omega$, $n \to \infty$, für

$$A_n := \big\{ f_n \geq (1-\varepsilon) g \big\} = \big\{ \omega \colon f_n(\omega) \geq (1-\varepsilon) g(\omega) \big\}.$$

Weil aufgrund der Definition der Mengen A_n und $f \geq 0$

$$f_n(\omega) \geq f_n(\omega) \mathbf{1}_{A_n}(\omega) \geq (1-\varepsilon) g(\omega) \mathbf{1}_{A_n}(\omega)$$

für alle $\omega \in \Omega$ gilt (man teste die zwei Möglichkeiten $\omega \in A_n$ und $\omega \notin A_n$), folgt

$$\begin{aligned}
\int_\Omega f_n \, \mathrm{d}\mu &\overset{\text{Mon.}}{\geq} \int_\Omega f_n \mathbf{1}_{A_n} \, \mathrm{d}\mu \\
&\overset{\text{Mon.}}{\geq} \int_\Omega (1-\varepsilon) g \mathbf{1}_{A_n} \, \mathrm{d}\mu \\
&\overset{\text{Lin.}}{=} (1-\varepsilon) \int_\Omega \Big(\sum_{k=1}^r \gamma_k \mathbf{1}_{C_k} \Big) \mathbf{1}_{A_n} \, \mathrm{d}\mu \\
&= (1-\varepsilon) \int_\Omega \sum_{k=1}^r \gamma_k \mathbf{1}_{A_n \cap C_k} \, \mathrm{d}\mu \\
&\overset{\text{Def.}}{=} (1-\varepsilon) \sum_{k=1}^r \gamma_k \mu(A_n \cap C_k).
\end{aligned}$$

Wegen Stetigkeit von Maßen gilt

$$\lim_{n\to\infty} \mu(A_n \cap C_k) = \mu\Big(\bigcup_{n=1}^\infty (A_n \cap C_k) \Big) = \mu\Big(\underbrace{\Big(\bigcup_{n=1}^\infty A_n \Big)}_{=\Omega} \cap C_k \Big) = \mu(C_k),$$

also gilt zusammen

$$\lim_{n\to\infty} \int_\Omega f_n \, \mathrm{d}\mu \geq (1-\varepsilon) \sum_{k=1}^r \gamma_k \mu(C_k) = (1-\varepsilon) \int_\Omega g \, \mathrm{d}\mu.$$

Weil ε beliebig gewählt war folgt die Hilfsbehauptung und damit ist der Beweis fertig.

□

Vorlesung 10

Warum war das Lemma so wichtig? Die Definition des Integrals als Supremum ist sehr unhandlich. Es hat natürlich den Vorteil, dass das Integral sofort sinnvoll definiert ist, dafür können wir mit der Definition nichts anstellen. Schauen wir uns als Beispiel die Beweise der folgenden elementaren Rechenregeln an. Per Approximation durch einfache Funktionen sind die Argumente sehr einfach, per Definition als Supremum wären die Argumente ziemlich fies.

Lemma 3.1.8. Für $f, g \colon \Omega \to [0, +\infty]$ messbar und $\alpha \geq 0$ gelten

(i) $\int_\Omega \alpha f \,\mathrm{d}\mu = \alpha \int_\Omega f \,\mathrm{d}\mu$,

(ii) $\int_\Omega (f+g)\,\mathrm{d}\mu = \int_\Omega f\,\mathrm{d}\mu + \int_\Omega g\,\mathrm{d}\mu$,

(iii) $f \leq g \Rightarrow \int_\Omega f\,\mathrm{d}\mu \leq \int_\Omega g\,\mathrm{d}\mu$.

Beweis. Wir zeigen nur (ii), (i) geht analog und (iii) folgt direkt aus der Definition als Supremum. Beachtet dabei folgende Eigenschaften vom Supremum: $M \subseteq N$ impliziert natürlich $\sup M \leq \sup N$.

Seien $(f_n), (g_n) \subseteq \mathcal{E}^+$ mit $f_n \uparrow f$, $g_n \uparrow g$, $n \to \infty$. Weil dann auch $f_n + g_n \in \mathcal{E}^+$ und $f_n + g_n \uparrow f + g$ gelten, folgt mit Lemma 3.1.7 und der Linearität des Integrals für einfache Funktionen

$$\int_\Omega f\,\mathrm{d}\mu + \int_\Omega g\,\mathrm{d}\mu = \lim_{n\to\infty}\Big(\int_\Omega f_n\,\mathrm{d}\mu + \int_\Omega g_n\,\mathrm{d}\mu\Big) = \lim_{n\to\infty}\int_\Omega (f_n+g_n)\,\mathrm{d}\mu = \int_\Omega (f+g)\,\mathrm{d}\mu.$$

□

Integral messbarer numerischer Funktionen

Im letzten Schritt wollen wir noch die Annahme der Nichtnegativität weglassen. Sei dazu $f\colon \Omega \to \overline{\mathbb{R}}$ $(\mathcal{A}, \mathcal{B}(\overline{\mathbb{R}}))$-messbar. Um f auf nicht-negative Funktionen zurückzuführen, erinnern wir an die Zerlegung von f in Postiv- und Negativteil

$$f = f^+ - f^- \quad \text{und} \quad |f| = f^+ + f^-$$

aus dem Kapitel über messbare Abbidungen.

Definition 3.1.9. Sei $f : \Omega \to \overline{\mathbb{R}}$ messbar und

$$\int_\Omega f^+\,\mathrm{d}\mu < \infty \quad \underline{\text{oder}} \quad \int_\Omega f^-\,\mathrm{d}\mu < \infty.$$

Dann definieren wir

$$\int_\Omega f\,\mathrm{d}\mu = \int_\Omega f^+\,\mathrm{d}\mu - \int_\Omega f^-\,\mathrm{d}\mu \in [-\infty, +\infty]$$

und sagen, das Integral $\int f\,\mathrm{d}\mu$ ist wohldefiniert. Ist $\int f\,\mathrm{d}\mu \in \mathbb{R}$, d.h. die Integrale über Positiv- und Negativteil sind beide endlich, so heißt f **μ-integrierbar** und wir sagen, das Integral existiert. Existiert bedeutet also wohldefiniert und endlich. Zur Notation: Man schreibt statt $\int_\Omega f\,\mathrm{d}\mu$ auch

$$\int_\Omega f(\omega)\,\mathrm{d}\mu(\omega) \quad \text{oder} \quad \int_\Omega f(\omega)\,\mu(\mathrm{d}\omega),$$

oft lässt man aus Faulheit auch Ω unter dem Integral weg.

Ohne die Einschränkung, dass eines der Integrale endlich ist, könnten wir das Integral nicht sinnvoll definieren. Das liegt daran, dass $+\infty - (+\infty)$ nicht sinnvoll definierbar ist.

Definition 3.1.10. Ist $f\colon \Omega \to \overline{\mathbb{R}}$ messbar und $A \in \mathcal{A}$, so definiert man

$$\int_A f\,\mathrm{d}\mu := \int_\Omega f \mathbf{1}_A\,\mathrm{d}\mu,$$

wenn die rechte Seite wohldefiniert ist. Alternativ schreibt man auch hier

$$\int_A f(\omega)\,\mathrm{d}\mu(\omega) \quad \text{oder} \quad \int_A f(\omega)\,\mu(\mathrm{d}\omega)$$

und wenn $\Omega = \mathbb{R}$ ist, auch $\int_{\mathbb{R}} f(x)\mathrm{d}\mu(x)$. Für den Spezialfall des Lebesgue-Maßes auf $\mathcal{B}(\mathbb{R})$ schreiben wir wieder

$$\int_a^b f(x)\,\mathrm{d}x \quad \text{statt} \quad \int_{[a,b]} f\,\mathrm{d}\lambda,$$

damit die Analogie zur Analysis nicht verloren geht. Weil das Integral nur für messbare Funktionen definiert ist, ist es ganz essentiell, dass auch $f\mathbf{1}_A$ eine messbare Funktion ist. Das liegt an der großen Flexibilität von messbaren Funktionen: $\mathbf{1}_A$ ist messbar, weil A messbar ist und das Produkt messbarer Funktionen ist messbar.

Lemma 3.1.11. Für $f, g\colon \Omega \to \overline{\mathbb{R}}$ μ-integrierbar und $\alpha \in \mathbb{R}$ gelten

(i) αf ist μ-integrierbar und
$$\int_\Omega \alpha f\,\mathrm{d}\mu = \alpha \int_\Omega f\,\mathrm{d}\mu.$$

(ii) Wenn $f+g$ sinnvoll definiert ist (d. h. kein $+\infty + (-\infty)$), so ist $f+g$ μ-integrierbar und
$$\int_\Omega (f+g)\,\mathrm{d}\mu = \int_\Omega f\,\mathrm{d}\mu + \int_\Omega g\,\mathrm{d}\mu.$$

(iii)
$$f \le g \quad \Rightarrow \quad \int_\Omega f\,\mathrm{d}\mu \le \int_\Omega g\,\mathrm{d}\mu$$

(iv) $\triangle$-Ungleichung:
$$\Big|\int_\Omega f\,\mathrm{d}\mu\Big| \le \int_\Omega |f|\,\mathrm{d}\mu$$

Beweis.

(i) Für $\alpha \ge 0$ gelten
$$(\alpha f)^+ = \alpha f^+ \quad \text{und} \quad (\alpha f)^- = \alpha f^-.$$
Damit ist αf μ-integrierbar, weil mit Lemma 3.1.8(i)
$$\int_\Omega \alpha f^+\,\mathrm{d}\mu = \alpha \int_\Omega f^+\,\mathrm{d}\mu < \infty \quad \text{und} \quad \int_\Omega \alpha f^- \mathrm{d}\mu = \alpha \int_\Omega f^-\,\mathrm{d}\mu < \infty$$
gelten. Es gilt dann per Definition des Integrals als Differenz der Integrale über Positiv- und Negativteil
$$\int_\Omega \alpha f\,\mathrm{d}\mu \overset{\text{Def.}}{=} \int_\Omega \alpha f^+\,\mathrm{d}\mu - \int_\Omega \alpha f^-\,\mathrm{d}\mu \overset{3.1.8}{=} \alpha \int_\Omega f^+\,\mathrm{d}\mu - \alpha \int_\Omega f^-\,\mathrm{d}\mu = \alpha \int_\Omega f\,\mathrm{d}\mu.$$
Der Fall $\alpha < 0$ geht genauso, wir nutzen hierbei $(\alpha f)^+ = -\alpha f^-$ und $(\alpha f)^- = -\alpha f^+$ und gehen dann genauso vor.

(ii) Die Summe ist bei Integralen immer der delikate Teil. Es gelten zunächst punktweise (Fallunterscheidungen)
$$0 \le (f+g)^+ \le f^+ + g^+ \quad \text{und} \quad 0 \le (f+g)^- \le f^- + g^-.$$
Damit gelten
$$\int_\Omega (f+g)^+\,\mathrm{d}\mu \overset{3.1.8}{\le} \int_\Omega (f^+ + g^+)\,\mathrm{d}\mu \overset{3.1.8}{=} \int_\Omega f^+\,\mathrm{d}\mu + \int_\Omega g^+\,\mathrm{d}\mu < \infty$$

und

$$\int_\Omega (f+g)^- \mathrm{d}\mu \overset{3.1.8}{\leq} \int_\Omega (f^- + g^-)\mathrm{d}\mu \overset{3.1.8}{=} \int_\Omega f^- \mathrm{d}\mu + \int_\Omega g^- \mathrm{d}\mu < \infty.$$

Also ist gemäß Definition $f+g$ μ-integrierbar. Die Berechnung des Integrals von $f+g$ ist clever. Wir kennen die Linearität bisher nur für nicht-negative Funktionen. Führen wir die Behauptung also auf den Fall zurück, indem wir wie folgt $f+g$ auf zwei Arten in Positiv- und Negativteil zerlegen:

$$\underbrace{(f+g)^+}_{\geq 0} - \underbrace{(f+g)^-}_{\geq 0} = f+g = (\underbrace{f^+}_{\geq 0} - \underbrace{f^-}_{\geq 0}) + (\underbrace{g^+}_{\geq 0} - \underbrace{g^-}_{\geq 0}).$$

Umformen ergibt

$$(f+g)^+ + f^- + g^- = (f+g)^- + f^+ + g^+.$$

Weil jetzt nur noch Summen nicht-negativer Funktionen auftauchen, können wir die bereits bekannte Linearität des Integrals aus Lemma 3.1.8 nutzen:

$$\int_\Omega (f+g)^+ \,\mathrm{d}\mu + \int_\Omega f^- \,\mathrm{d}\mu + \int_\Omega g^- \,\mathrm{d}\mu = \int_\Omega (f+g)^- \,\mathrm{d}\mu + \int_\Omega f^+ \,\mathrm{d}\mu + \int_\Omega g^+ \,\mathrm{d}\mu.$$

Erneutes Auflösen ergibt

$$\int_\Omega (f+g)^+ \,\mathrm{d}\mu - \int_\Omega (f+g)^- \,\mathrm{d}\mu = \int_\Omega f^+ \,\mathrm{d}\mu - \int_\Omega f^- \,\mathrm{d}\mu + \int_\Omega g^+ \,\mathrm{d}\mu - \int_\Omega g^- \,\mathrm{d}\mu$$

und Ausnützen der Definition des Integrals als Differenz der Positiv- und Negativteile

$$\int_\Omega (f+g) \,\mathrm{d}\mu = \int_\Omega f \,\mathrm{d}\mu + \int_\Omega g \,\mathrm{d}\mu.$$

(iii) Natürlich gilt $f \leq g \Leftrightarrow g - f \geq 0$. Weil die Nullfunktion sowie $g-f$ nicht-negativ sind, gilt wegen der Linearität und der Definition des Integrals für einfache Funktionen (die Nullfunktion)

$$0 = \int_\Omega 0 \,\mathrm{d}\mu \leq \int_\Omega (g-f) \,\mathrm{d}\mu \overset{\text{(i),(ii)}}{=} \int_\Omega g \mathrm{d}\mu - \int_\Omega f \mathrm{d}\mu.$$

Umformen gibt die Behauptung.

(iv) Die Dreicksungleichung für Integrale folgt aus der Dreiecksungleichung in $\mathbb{R}$:

$$\begin{aligned}
\Big|\int_\Omega f \,\mathrm{d}\mu\Big| &\overset{\text{Def.}}{=} \Big|\int_\Omega f^+ \,\mathrm{d}\mu - \int_\Omega f^- \,\mathrm{d}\mu\Big| \\
&\overset{\triangle}{\leq} \Big|\int_\Omega f^+ \,\mathrm{d}\mu\Big| + \Big|\int_\Omega f^- \,\mathrm{d}\mu\Big| \\
&\overset{\geq 0}{=} \int_\Omega f^+ \,\mathrm{d}\mu + \int_\Omega f^- \,\mathrm{d}\mu \\
&\overset{\text{(ii)}}{=} \int_\Omega (f^+ + f^-) \,\mathrm{d}\mu = \int_\Omega |f| \,\mathrm{d}\mu.
\end{aligned}$$

□

Eine Konsequenz der gerade genutzten Linarität kombiniert mit $|f| = f^+ + f^-$ ist folgendes Äquivalenz:

$$f \text{ ist } \mu\text{-integrierbar} \quad \Longleftrightarrow \quad \int_\Omega f \,\mathrm{d}\mu \text{ existiert} \quad \Longleftrightarrow \quad \int_\Omega |f| \,\mathrm{d}\mu < \infty. \tag{3.3}$$

Die rechte Seite sieht sehr viel nützlicher aus als die eigentliche Definition von μ-Integrierbarkeit, ist aber nur eine recht triviale Modifikation.

Beispiel 3.1.12.

(i) Für den wichten Spezialfall $(\Omega, \mathcal{A}, \mu) = (\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$ schauen wir uns mal ein paar Beispiele an:

(a) Ist f stückweise stetig, so ist f Riemann integrierbar auf Intervallen $[a, b]$ und gleich dem Lebesgue Integral $\int_{[a,b]} f \,\mathrm{d}\lambda$. Das erklärt auch wieder, warum wir auch für das Lebesgue integral die $\mathrm{d}x$-Notation nutzen und $\int_{[a,b]} f \,\mathrm{d}\lambda = \int_a^b f(x)\,\mathrm{d}x$ schreiben. Ihr könnt also für nette Integranden $\int_{[a,b]} f \,\mathrm{d}\lambda$ mit den Rechnenregeln aus der Schule und Analysis berechnen (Stammfunktionen, partielle Integration, Substitution). Stückweise stetig ist eigentlich nicht die richtige Annahme, die richtige Formulierung ist folgende (Lebesgue'schen Kriterium für Riemann-Integrierbarkeit): Eine Funktion ist Riemann integrierbar auf $[a, b]$ genau dann, wenn die Unstetigkeitsstellen eine Nullmenge sind. Solch eine Funktion ist auch Lebesgue integrierbar und die Integrale sind gleich. Weil wir uns in dieser Vorlesung nicht für das Riemann Integral interessieren, führen wir das nicht weiter aus.

(b) Ist $f \geq 0$ und stückweise stetig, so stimmt $\int_{\mathbb{R}} f \,\mathrm{d}\lambda$ mit dem uneigentlichen Riemann Integral überein. Das sehen wir später mit dem monotonen Konvergenz Theorem und (a). Ihr dürft das Integral in dem Fall als

$$\int_{\mathbb{R}} f \,\mathrm{d}\lambda = \lim_{n\to\infty} \int_{-n}^{n} f(x)\,\mathrm{d}x \tag{3.4}$$

schreiben, und die rechte Seite mit den Tricks der Analysis berechnen. Deshalb schreiben wir auch wie in Analysis $\int_{\mathbb{R}} f(x)\,\mathrm{d}x$ statt $\int_{\mathbb{R}} f \,\mathrm{d}\lambda$. Warum das gilt, schauen wir uns nach Satz 3.2.1 nochmal an.

(c) Warnung: Den Rechentrick aus (3.4) dürft ihr nicht immer nutzen, wir haben schließlich angenommen, dass f nicht-negativ ist. Hier sind zwei Gegenbeispiele:

$$f(x) = \frac{\sin(x)}{x}, \quad x \in \mathbb{R},$$
$$f(x) = \sum_{k=1}^{\infty} (-1)^k \frac{1}{k} \mathbf{1}_{[k-1,k)}(x), \quad x \in \mathbb{R}.$$

Bei beiden Beispielen sind Positiv- und Negativteil nicht integrierbar (das kann man mit (b) nachrechnen), $\int_{\mathbb{R}} f \,\mathrm{d}\lambda$ ist also nicht einmal wohldefiniert, der Grenzwert $\lim_{n\to\infty} \int_{-n}^{n} f(x)\,\mathrm{d}x$ existiert jedoch. Für das Riemann Integral (auf einem Interval) gilt zwar, „Riemann integrierbar impliziert Lebesgue integrierbar und die Integrale sind gleich", für das uneigentliche Riemann Integral gilt das weder der Beispiele jedoch nicht!

(d) In (a) haben wir gesagt, dass Riemann integrierbare Funktionen auch Lebesgue integrierbar sind. Hier ist ein Beispiel, das zeigt, dass die Umkehrung nicht immer gilt: $f = \mathbf{1}_{\mathbb{Q}\cap[0,1]}$, die Funktion die nur den Wert 1 für rationale Zahlen in $[0, 1]$ annimmt. Weil f eine einfache Funktion ist, $f = 1 \cdot \mathbf{1}_A$ für die Borel-messbare Menge $A = \mathbb{Q}\cap[0, 1]$, ist sie Lebesgue integrierbar mit Integral $\int_{[0,1]} \mathbf{1}_{\mathbb{Q}} \,\mathrm{d}\lambda = 1\cdot\lambda(\mathbb{Q}\cap[0, 1]) = 0$. Die Funktion ist aber nicht Riemann integrierbar: Jede Treppenfunktion über kleine Intervalle, die über f liegt, ist auf $[0, 1]$ größer als 1. Liegt eine Treppenfunktion unter f, so ist sie kleiner als 0 auf $[0, 1]$. Damit können die Ober- und Untersummen nicht gegen den gleichen Wert konvergieren.

(ii) Schauen wir uns jetzt das Beispiel $(\Omega, \mathcal{A}, \mu) = (\mathbb{N}, \mathcal{P}(\mathbb{N}), \mu)$, $\mu(A) = \#A$, an. Ganz am Anfang der Vorlesung hatten wir dieses Maß Zählmaß genannt. An diesem Beispiel lernen wir: Summen sind auch nur Integrale! Warum? Für $f : \mathbb{N} \to [0, \infty)$ gilt

$$\int_{\mathbb{N}} f \,\mathrm{d}\mu = \sum_{k=0}^{\infty} f(k).$$

Das folgt direkt aus Lemma 3.1.7 weil f als $f = \sum_{k=0}^{\infty} f(k)\mathbf{1}_{\{k\}}$ geschrieben werden kann (setzt mal m in die rechte Seite ein, es ist immer nur ein Summand ungleich 0!) und damit $f_n \uparrow f$ gilt, mit der Folge $f_n := \sum_{k=1}^{n} f(k)\mathbf{1}_{\{k\}}$ einfacher Funktionen. Zusammen gibt das aufgrund der Definition des allgemeinen Lebesgue Integrals für einfache Integranden

$$\int_{\mathbb{N}} f \,\mathrm{d}\mu \overset{(3.1.7)}{=} \lim_{n\to\infty} \int_{\mathbb{N}} f_n \,\mathrm{d}\mu \overset{\text{Def.}}{=} \lim_{n\to\infty} \sum_{k=0}^{n} f(k)\,\mu(\{k\}) = \lim_{n\to\infty} \sum_{k=0}^{n} f(k) = \sum_{k=0}^{\infty} f(k).$$

Vielleicht habt ihr in der Analysis 2 schon über Nullmengen gesprochen. Dann habt ihr schon gelernt, dass Nullmengen bei Integralen keine Rolle spielen. Wenn nicht, lernt ihr Nullmengen und ihre Bedeutung für Integrale jetzt kennen:

Definition 3.1.13. $N \in \mathcal{A}$ heißt μ-**Nullmenge**, falls $\mu(N) = 0$. [1]

Aufgrund der Subadditivität von Maßen (folgt aus der σ-Additivität) folgt sofort, dass abzählbare Vereinigungen von Nullmengen wieder Nullmengen sind (kleine Übung).

Definition 3.1.14.

(i) Gilt eine Eigenschaft für alle $\omega \in \Omega$ außer einer μ-Nullmenge (d. h. für $N := \{\omega \in \Omega :$ Eigenschaft gilt nicht$\}$ gilt $\mu(N) = 0$), so gilt die Eigenschaft μ-fast überall. Man schreibt kurz auch μ-f.ü.

(ii) Ist μ ein Wahrscheinlichkeitsmaß, so sagt man anstelle von „μ-fast überall" auch „μ-fast sicher" oder kurz μ-f.s.

Wenn klar ist über welches Maß μ gesprochen wird, sagt man auch nur „fast überall" oder „fast sicher". Vorlesung 11

Weil aufgrund der Definition des Integrals für Indikatorfunktionen $f = \mathbf{1}_N$ über Nullmengen $\int_\Omega f \mathrm{d}\mu = 1 \cdot \mu(N) = 0$, ist es nicht überraschend, dass Nullmengen beim Integrieren keine Rolle spielen. Das wird in (ii) des folgendes sehr wichtigen Satzes deutlich:

Satz 3.1.15. Sind $f, g\colon \Omega \to \overline{\mathbb{R}}$ μ-integrierbar, so gelten:

(i) f ist μ-fast überall endlich.

(ii) $f = g$ μ-fast überall impliziert $\int_\Omega f \,\mathrm{d}\mu = \int_\Omega g \,\mathrm{d}\mu$.

(iii) $f \geq 0$ und $\int_\Omega f \,\mathrm{d}\mu = 0$ impliziert $f = 0$ μ-fast überall.

Beweis.

(i) Sei $A := \{|f| = \infty\} = f^{-1}(\{-\infty, +\infty\}) \in \mathcal{A}$. Weil $n\mathbf{1}_A \leq |f|$ für alle $n \in \mathbb{N}$ gilt, folgt

$$n\mu(A) \overset{\text{Def.}}{=} \int_\Omega n\mathbf{1}_A \,\mathrm{d}\mu \overset{\text{Mon.}}{\leq} \int_\Omega |f| \,\mathrm{d}\mu = \int_\Omega (f^+ + f^-)\mathrm{d}\mu \underset{\mu\text{-int.}}{\overset{f}{<}} \infty$$

für alle $n \in \mathbb{N}$. Weil $\mu(A) > 0$ einen Widerspruch gibt (dann wäre $n\mu(A)$ unbeschränkt aber $\int |f| \,d\mu$ ist eine obere Schranke), folgt die Behauptung.

(ii) Für $N := \{f \neq g\} \in \mathcal{A}$ gilt aufgrund der Voraussetzung $\mu(N) = 0$. Weil aus $f = g$ fast überall auch $f^+ = g^+$ und $f^- = g^-$ fast überall folgt, impliziert die Definition des Integrals als $\int_\Omega f \,\mathrm{d}\mu = \int_\Omega f^+ \,\mathrm{d}\mu - \int_\Omega f^- \,\mathrm{d}\mu$ bzw. $\int_\Omega g \,\mathrm{d}\mu = \int_\Omega g^+ \,\mathrm{d}\mu - \int_\Omega g^- \,\mathrm{d}\mu$, dass die Aussage nur für $f, g \geq 0$ gezeigt werden muss (wende Aussage dann auf Positiv- und Negativteil an). Seien also $f, g \geq 0$ und

$$N = \{f \neq g\} = \{\omega\colon f(\omega) \neq g(\omega)\}$$

[1] Eigentlich macht man das Allgemeiner: Eine Teilmenge N von Ω heißt Nullmenge, falls es eine messbare Menge A mit $N \subseteq A$ und $\mu(A) = 0$ gibt. Weil in der Stochastik 1 alle von uns benötigten Nullmengen selber messbar sind, ignorieren wir das. Das Thema wird erst relevant, wenn ihr die Brownsche Bewegung kennenlernt.

die Nullmenge, auf der f und g nicht übereinstimmen. Dann gilt aufgrund der Monotonie und der Definition des Integrals

$$0 \leq \int_\Omega f\mathbf{1}_N \,\mathrm{d}\mu \leq \int_\Omega (+\infty)\mathbf{1}_N \,\mathrm{d}\mu = +\infty \cdot \mu(N) = +\infty \cdot 0 = 0.$$

Für die letzte Gleichung haben wir die Konvention $+\infty \cdot 0 = 0$ genutzt. Genauso gilt $\int_\Omega g\mathbf{1}_N \,\mathrm{d}\mu = 0$. Wenn wir jetzt $1 = \mathbf{1}_N + \mathbf{1}_{N^C}$ schreiben, ergibt sich mit der Linearität des Integrals

$$\int_\Omega f \,\mathrm{d}\mu = \int_\Omega f\mathbf{1}_N \,\mathrm{d}\mu + \int_\Omega f\mathbf{1}_{N^C} \,\mathrm{d}\mu = \int_\Omega f\mathbf{1}_{N^C} \,\mathrm{d}\mu = \int_\Omega g\mathbf{1}_{N^C} \,\mathrm{d}\mu = \int_\Omega g \,\mathrm{d}\mu.$$

(iii) Seien $A_n = \{f > \frac{1}{n}\} = \{\omega\colon f(\omega) > \frac{1}{n}\}$ für $n \in \mathbb{N}$. Damit ist mit der Monotonie des Integrals

$$0 \overset{\text{Ann.}}{=} \int_\Omega f \,\mathrm{d}\mu \overset{\text{Mon.}}{\geq} \int_\Omega f\mathbf{1}_{A_n} \,\mathrm{d}\mu \overset{\text{Mon.}}{\geq} \int_\Omega \frac{1}{n}\mathbf{1}_{A_n} \,\mathrm{d}\mu \overset{\text{Def.}}{=} \frac{1}{n}\mu(A_n) \geq 0,$$

weil $\frac{1}{n}\mathbf{1}_{A_n} \leq f\mathbf{1}_{A_n} \leq f$. Also ist $\mu(A_n) = 0$ für alle $n \in \mathbb{N}$. Damit gilt

$$0 \leq \mu(\{\omega\colon f(\omega) > 0\}) = \mu\Big(\bigcup_{k=1}^\infty A_k\Big) \overset{\text{subadd.}}{\leq} \sum_{k=1}^\infty \mu(A_k) = 0.$$

Es gilt also $f = 0$ μ-fast überall.

$\square$

Für spätere Verwendungen noch ein Satz zur Transformation von Integralen. In Analysis 2 habt ihr den schon in konkreter Form im $\mathbb{R}^d$ kennengelernt.

Satz 3.1.16. **[abstrakter Transformationssatz]** Seien $(\Omega, \mathcal{A})$, $(\Omega', \mathcal{A}')$ messbare Räume, μ ein Maß auf $\mathcal{A}$, $f\colon \Omega \to \Omega'$ messbar, $g\colon \Omega' \to \overline{\mathbb{R}}$ messbar und $g \geq 0$. Dann gilt

$$\int_{\Omega'} g \,\mathrm{d}\mu_f = \int_\Omega g \circ f \,\mathrm{d}\mu,$$

wobei $+\infty = +\infty$ möglich ist. Dabei ist μ_f der push-forward (Bildmaß) von μ.

$$\begin{tikzcd}
(\Omega, \mathcal{A}, \mu) \arrow[r, "f"'] \arrow[dr, "g\circ f", "\int_\Omega g\circ f \,\mathrm{d}\mu"'] & (\Omega', \mathcal{A}', \mu_f) \arrow[d, "g"', "\int_{\Omega'} g \,\mathrm{d}\mu_f"] \\
& (\overline{\mathbb{R}}, \mathcal{B}(\overline{\mathbb{R}}))
\end{tikzcd}$$

Beweis. Einmal durch die „Gebetsmühle“ der Integrationstheorie (d. h. zeige die Aussage für einfache Integranden und nehme dann den Grenzwert):

(A) Sei erstmal

$$g = \sum_{k=1}^n \alpha_k \mathbf{1}_{A_k} \geq 0$$

eine nicht-negative einfache Funktion. Wir nehmen die Darstellung mit $A_k = g^{-1}(\{\alpha_k\})$. Dann gilt

$$g \circ f = \sum_{k=1}^n \alpha_k \mathbf{1}_{f^{-1}(A_k)} \geq 0,$$

$g \circ f$ ist also auch eine einfache Funktion. Weil nach Definition des push-forwards $\mu_f(A_k) = \mu(f^{-1}(A_k))$ gilt, bekommen wir

$$\int_\Omega g \circ f \,\mathrm{d}\mu \overset{\text{Def.}}{=} \sum_{k=1}^n \alpha_k \mu(f^{-1}(A_k)) = \sum_{k=1}^n \alpha_k \mu_f(A_k) \overset{\text{Def.}}{=} \int_{\Omega'} g \,\mathrm{d}\mu_f.$$

Damit gilt die Behauptung für einfache Funktionen.

(B) Weil g messbar ist, existiert eine Folge $(g_n) \subseteq \mathcal{E}^+$ mit $g_n \uparrow g$, $n \to \infty$. Also gilt

$$\int_{\Omega'} g \,\mathrm{d}\mu_f \overset{3.1.7}{=} \lim_{n\to\infty} \int_{\Omega'} g_n \,\mathrm{d}\mu_f \overset{(A)}{=} \lim_{n\to\infty} \int_\Omega g_n \circ f \,\mathrm{d}\mu \overset{3.1.7}{=} \int_\Omega g \circ f \,\mathrm{d}\mu.$$

Im letzten Schritt haben wir genutzt, dass auch $g_n \circ f$ einfach ist (siehe (A)) und $g_n \circ f \uparrow g \circ f$, $n \to \infty$, gilt.

□

Wir hätten den letzten Satz auch direkt ohne die Nichtnegativität formulieren können. Aus didaktischen Gründen zerlegen wir die Aussage in den Satz und das folgende Korollar. Der nicht-negative Fall ist einfacher zu formulieren weil Integrale über nicht-negative Funktionen immer definiert sind. Bei allgemeinen Integranden muss man immer aufpassen, dass nicht $+\infty + (-\infty)$ auftaucht, das haben wir Wohldefiniertheit genannt. Die Formulierung des Satzes muss also die Wohldefiniertheit beinhalten. Um sich das Leben leichter zu machen, nimmt man meistens sogar die Integrierbarkeit (Wohldefiniert und endlich) an, das reicht in den Anwendungen ohnehin meistens aus.

Korollar 3.1.17. Unter den Voraussetzungen von Satz 3.1.16 gelte jetzt nur noch $g\colon \Omega' \to \overline{\mathbb{R}}$. Dann ist g μ_f-integrierbar genau dann, wenn $g \circ f$ μ-integrierbar ist. Ist eine dieser Aussagen erfüllt, so gilt ebenfalls die Transformationsformel

$$\int_{\Omega'} g \,\mathrm{d}\mu_f = \int_\Omega g \circ f \,\mathrm{d}\mu.$$

Beweis. Wegen $g^+ \circ f = (g \circ f)^+$ und $g^- \circ f = (g \circ f)^-$ folgt aus dem Transformationssatz für nicht-negative Funktionen

$$\begin{aligned} g\ \mu_f\text{-integrierbar} &\overset{\text{Def.}}{\Leftrightarrow} \int_{\Omega'} g^+ \,\mathrm{d}\mu_f < \infty \quad \text{und} \quad \int_{\Omega'} g^- \,\mathrm{d}\mu_f < \infty \\ &\Leftrightarrow \int_\Omega g^+ \circ f \,\mathrm{d}\mu < \infty \quad \text{und} \quad \int_\Omega g^- \circ f \,\mathrm{d}\mu < \infty \\ &\overset{\text{Def.}}{\Leftrightarrow} g \circ f\ \mu\text{-integrierbar}. \end{aligned}$$

Nun zur Berechnung der Integrale:

$$\begin{aligned} \int_{\Omega'} g \,\mathrm{d}\mu_f &\overset{\text{Def.}}{=} \int_{\Omega'} g^+ \,\mathrm{d}\mu_f - \int_{\Omega'} g^- \,\mathrm{d}\mu_f \\ &\overset{3.1.16}{=} \int_\Omega g^+ \circ f \,\mathrm{d}\mu - \int_\Omega g^- \circ f \,\mathrm{d}\mu \\ &= \int_\Omega (g \circ f)^+ \,\mathrm{d}\mu - \int_\Omega (g \circ f)^- \,\mathrm{d}\mu \\ &\overset{\text{Def.}}{=} \int_\Omega g \circ f \,\mathrm{d}\mu. \end{aligned}$$

□

3.2 Konvergenzsätze

Im folgenden sei $(\Omega, \mathcal{A}, \mu)$ ein fester Maßraum. Gezeigt haben wir schon

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

wenn $(f_n)_{n\in\mathbb{N}}$ eine Folge nicht-negativer einfacher Funktionen ist, die wachsend gegen f konvergieren. Wir wollen nun die gleiche Aussage für beliebige nicht-negative wachsende Folgen zeigen.

Satz 3.2.1. **[Monotone Konvergenz Theorem (MCT)]** Seien $f, f_1, f_2, ...\colon \Omega \to \overline{\mathbb{R}}$ messbar und es gelte $0 \leq f_1 \leq f_2 \leq ... \leq f$ sowie $f = \lim\limits_{n\to\infty} f_n$ μ-f.ü. Dann gilt

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

wobei $+\infty = +\infty$ möglich ist.

Für monoton wachsende Funktionenfolgen darf der Limes also in das Integral getauscht werden.

Beweis.

(i) Wir nehmen an, dass $f_1(\omega) \leq f_2(\omega) \leq ... \leq f(\omega)$ und $f(\omega) = \lim\limits_{n\to\infty} f_n(\omega)$ nicht nur fast überall, sondern für alle $\omega \in \Omega$ (also punktweise) gelten.

„$\leq$“: Wegen der Monotonie des Integrals gilt

$$\int_\Omega f_n \,\mathrm{d}\mu \leq \int_\Omega f \,\mathrm{d}\mu$$

für alle $n \in \mathbb{N}$. Weil die Folge der Integrale aufgrund der Monotonie wachsend ist, existiert der Grenzwert ($+\infty$ ist möglich). Warum? Entweder die Folge divergiert nach $+\infty$, oder sie ist beschränkt. Im ersten Fall haben wir die Konvergenz gegen $+\infty$, im zweiten Fall haben wir die Konvergenz gegen eine reelle Zahl (beschränkte monotone Folgen konvergieren nach Analysis 1). Wieder nach Analysis 1, Vergleichssatz für Folgen, gilt damit

$$\lim_{n\to\infty} \int_\Omega f_n \,\mathrm{d}\mu \leq \int_\Omega f \,\mathrm{d}\mu.$$

„$\geq$“: Weil alle f_n messbar sind, existieren Folgen $(g_{n,k}) \subseteq \mathcal{E}^+$ mit $g_{n,k} \uparrow f_n$, $k \to \infty$. Sei $h_n = g_{1,n} \vee ... \vee g_{n,n} = \max\{g_{1,n}, ..., g_{n,n}\}$. Die h_n sind einfache Funktionen, für die zwei Eigenschaften gelten:

(a) $h_n \leq f_n$

(b) $h_n \uparrow f$

Zu (a): Es gilt $g_{i,n} \leq f_i \leq f_n$ für alle $i \leq n$, also ist auch das punktweise Maximum kleiner als f_n. Zu (b): Weil $h_n \geq g_{i,n}$ für alle $i = 1, ..., n$ gilt, ist auch

$$\lim_{n\to\infty} h_n \geq \lim_{n\to\infty} g_{i,n} = f_i$$

für alle festen $i \in \mathbb{N}$. Weil aber $\lim\limits_{i\to\infty} f_i = f$ gilt, ist $\lim\limits_{n\to\infty} h_n \geq f$, erneut nach dem Vergleichssatz für Folgen aus Analysis 1. Weil auch noch $h_n \leq f_n \leq f$ gilt, folgt mit der letzten Aussage $\lim\limits_{n\to\infty} h_n = f$ punktweise. Die Folge (h_n) ist also eine wachsende Folge von einfachen Funktionen, so dass (f_n) zwischen (h_n) und f liegt und (h_n) punktweise gegen f konvergiert. Damit bekommen wir aus dem Monotone Konvergenz Theorem für einfache Funktionen durch Einschachtelung die Behauptung:

$$\lim_{n\to\infty} \int_\Omega f_n \,\mathrm{d}\mu \underset{\text{Mon.}}{\overset{\text{(a)}}{\geq}} \lim_{n\to\infty} \int_\Omega h_n \,\mathrm{d}\mu \overset{3.1.7}{=} \int_\Omega \lim_{n\to\infty} h_n \,\mathrm{d}\mu \overset{\text{(b)}}{=} \lim_{n\to\infty} \int_\Omega f \,\mathrm{d}\mu.$$

(ii) Sei N die Nullmenge, auf der die Annahme aus (i) nicht gilt, also

$$N = \{\omega \in \Omega : f_n(\omega) \not\to f(\omega)\}.$$

Es gilt $f_n \mathbf{1}_{N^C} \uparrow f\mathbf{1}_{N^C}$, $n \to \infty$, punktweise, weil für alle $\omega \in N$ die Folge konstant 0. Wegen (i) gilt dann

$$\lim_{n\to\infty} \int_\Omega f_n \,\mathrm{d}\mu \overset{3.1.15}{=} \lim_{n\to\infty} \int_\Omega f_n \mathbf{1}_{N^C} \,\mathrm{d}\mu \overset{(i)}{=} \int_\Omega f \mathbf{1}_{N^C} \,\mathrm{d}\mu \overset{3.1.15}{=} \int_\Omega f \,\mathrm{d}\mu,$$

weil Integrale zweier Funktionen gleich sind, wenn sie nur auf Nullmengen unterschiedlich sind.

□

Kommen wir zu einer Anwendung, die für die Stochastik sehr wichtig ist. Gerade in der Finanzmathematik wird folgender „Maßwechsel“ essentiell sein!

Anwendung 3.2.2. Sei $(\Omega, \mathcal{A}, \mu)$ ein Maßraum, $f\colon \Omega \to \overline{\mathbb{R}}$ messbar und nicht-negativ und μ ein Maß auf $\mathcal{A}$. Dann ist

$$\nu(A) := \int_A f \,\mathrm{d}\mu = \int_\Omega f \mathbf{1}_A \,\mathrm{d}\mu$$

ein Maß auf $\mathcal{A}$.

Beweis.

$\nu \geq 0$ ✓ wegen Integral über nicht-negative Funktion.

$\nu(\emptyset) = 0$ ✓ weil Integral über die Nullfunktion 0 ist.

σ-Additivität: Seien $A_1, A_2, ... \in \mathcal{A}$ paarweise disjunkt. Dann gilt

$$\begin{aligned}
\nu\Big(\dot{\bigcup_{k=1}^{\infty}} A_k\Big) &\overset{\text{Def.}}{=} \int_\Omega f \mathbf{1}_{\dot{\bigcup}_{k=1}^{\infty} A_k} \,\mathrm{d}\mu \\
&= \int_\Omega f \cdot \Big(\sum_{k=1}^{\infty} \mathbf{1}_{A_k}\Big) \,\mathrm{d}\mu \\
&\overset{\text{Lin. Reihe}}{=} \int_\Omega \Big(\sum_{k=1}^{\infty} f \cdot \mathbf{1}_{A_k}\Big) \,\mathrm{d}\mu \\
&\overset{\text{Def. Reihe}}{=} \int_\Omega \lim_{n\to\infty} \sum_{k=1}^{n} f \mathbf{1}_{A_k} \,\mathrm{d}\mu \\
&\overset{3.2.1}{=} \lim_{n\to\infty} \int_\Omega \sum_{k=1}^{n} f \mathbf{1}_{A_k} \,\mathrm{d}\mu \\
&\overset{\text{Lin. Integral}}{=} \lim_{n\to\infty} \sum_{k=1}^{n} \int_\Omega f \mathbf{1}_{A_k} \,\mathrm{d}\mu \\
&\overset{\text{Def.}}{=} \lim_{n\to\infty} \sum_{k=1}^{n} \nu(A_k) \overset{\text{Def. Reihe}}{=} \sum_{k=1}^{\infty} \nu(A_k).
\end{aligned}$$

Weil hier die Folge $\Big(\sum_{k=1}^{n} f \mathbf{1}_{A_k}\Big)_{n\in\mathbb{N}} \not\subseteq \mathcal{E}^+$, reicht die einfache Version der monotonen Konvergenz aus 3.1.7 nicht aus, wir brauchen monotone Konvergenz wirklich für allgemeine messbare Funktionen. □

Anschließend an den Maßwechsel noch eine Bemerkung, die für diese Stochastik 1 Vorlesung nicht wichtig ist. Da das Thema zum Beispiel in der Finanzmathematik extrem relevant ist, schreiben wir die Notation zum Gewöhnen schon einmal auf:

Bemerkung 3.2.3.

(i) Man schreibt mit ν aus vorheriger Anwendung auch

$$\frac{\mathrm{d}\nu}{\mathrm{d}\mu} = f$$

und nennt f die **Radon-Nikodým-Ableitung** oder **-Dichte von ν bezüglich** μ. Man sagt auch, dass ν absolutstetig bezüglich μ ist. ν ist ein Wahrscheinlichkeitsmaß (d. h. $\nu(\Omega) = 1$) genau dann, wenn $\int_\Omega f \,\mathrm{d}\mu = 1$. Das kennen wir ja schon!

(ii) Wir kennen den Begriff der Absolutstetigkeit schon für Verteilungsfunktionen (siehe Definition 6.3.1). In der Tat passt das genau zu dem neuen Begriff der Absolutstetigkeit für Maße: Ist $\mathbb{P}_F$ ein Wahrscheinlichkeitsmaß auf $\mathcal{B}(\mathbb{R})$ mit Verteilungsfunktion F und Dichte f, so gilt aufgrund der Definition von $\mathbb{P}_F$ (Maße sind auf einem $\cap$-stabilen Erzeuger eindeutig festgelegt!), dass

$$\frac{\mathrm{d}\mathbb{P}_F}{\mathrm{d}\lambda} = f, \quad \lambda = \text{Lebesguemaß}.$$

Ist also F absolutstetig mit Dichte f, so ist das Maß $\mathbb{P}_F$ absolutstetig bezüglich dem Lebesguemaß mit Dichte f. Die zwei Begriffe der Absolutstetigkeit für Verteilungsfunktionen und Maße passen damit zusammen.

Ganz analog funktioniert das für diskrete Verteilungen. Ist F diskret mit Sprungstellen $(a_k)_{k=1,\dots,N}$ und Sprunghöhen $(p_k)_{k=1,\dots,N}$, so gilt

$$\frac{\mathrm{d}\mathbb{P}_F}{\mathrm{d}\mu} = \sum_{k=1}^N p_k \mathbf{1}_{\{a_k\}}, \quad \mu = \sum_{k=1}^N \delta_{a_k}.$$

Das ist der Grund dafür, weshalb bei diskreten Verteilungen die Wahrscheinlichkeiten $(p_k)_{k=1,\dots,N}$ manchmal auch Zähldichte genannt werden. Die Formel ist natürlich furchtbar, sie hat aber eine einfache Bedeutung: Das Maß μ wird durch die Dichte „umgewichtet". Die Masse liegt auf den gleichen Werten $a_1, \dots, a_N$, $\mathbb{P}_F$ hat aber eine andere Verteilung der Masse auf $a_1, \dots, a_N$: Auf den Werten a_k liegt nun nicht Masse 1, sondern Masse p_k.

Vorlesung 12

Satz 3.2.4. **[Lemma von Fatou]** Seien $f_1, f_2, \dots : \Omega \to \overline{\mathbb{R}}$ messbar und nicht-negativ, die Folge (f_n) muss dabei nicht konvergieren. Dann gilt

$$\int_\Omega \liminf_{n\to\infty} f_n \,\mathrm{d}\mu \leq \liminf_{n\to\infty} \int_\Omega f_n \,\mathrm{d}\mu,$$

$+\infty \leq +\infty$ ist dabei möglich.

Wenn (f_n) sogar μ-f.ü. konvergiert und $(\int_\Omega f_n \,\mathrm{d}\mu)$ konvergiert, gilt damit

$$\int_\Omega \lim_{n\to\infty} f_n \,\mathrm{d}\mu \leq \lim_{n\to\infty} \int_\Omega f_n \,\mathrm{d}\mu$$

weil für konvergente Folgen $\liminf_{n\to\infty} a_n = \lim_{n\to\infty} a_n$ gilt. Die Ungleichung „$\leq$" gilt in den Konvergenzsätzen also mit weniger Annahmen als Satz 3.2.1 (und anschliessend in Satz 3.2.5).

Beweis. Definieren wir $g_n := \inf\limits_{k\geq n} f_k$, so ist g_n messbar für alle $n \in \mathbb{N}$, wachsend in n und erfüllt $g_n \leq f_n$, $n \in \mathbb{N}$, sowie punktweise $\lim\limits_{k\to\infty} g_k = \liminf\limits_{n\to\infty} f_n$ (das ist eine der äquivalenten Definition des Limes Inferiors). Satz 3.2.1 und Monotonie von Integralen und Folgengrenzwerten gibt dann die Aussage:

$$\int_\Omega \liminf_{n\to\infty} f_n \,\mathrm{d}\mu = \int_\Omega \lim_{n\to\infty} g_n \,\mathrm{d}\mu = \lim_{n\to\infty} \int_\Omega g_n \,\mathrm{d}\mu = \liminf_{n\to\infty} \int_\Omega g_n \,\mathrm{d}\mu \leq \liminf_{n\to\infty} \int_\Omega f_n \,\mathrm{d}\mu.$$

Das dritte Gleichheitszeichen gilt weil g_n (und damit das Integral) wachsend in n ist. $\square$

Satz 3.2.5. **[Dominierte Konvergenz Theorem (DCT)]** Seien $f, f_1, f_2, \ldots : \Omega \to \overline{\mathbb{R}}$ messbar. Es sollen gelten

(a) $\lim_{n\to\infty} f_n = f$ μ-fast überall,

(b) $|f_n| \leq g$ μ-fast überall für alle $n \in \mathbb{N}$, für eine beliebige μ-integrierbare nicht-negative messbare numerische Funktion g.

Dann sind $f, f_1, f_2, \ldots$ μ-integrierbar und

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

Die Funktion g spielt keine große Rolle (sie muss nur existieren) und wird integrierbare Majorante für die Folge (f_n) genannt.

Beweis. Die behauptete Integrierbarkeit von $f, f_1, f_2, \ldots$ folgt aus (3.3) weil $|f_n| \leq g$ μ-fast überall angenommen wird (das gilt dann auch für den Grenzwert). Wie beim Beweis der monotonen Konvergenz nehmen wir zunächst an, dass die Konvergenz sogar für alle $\omega \in \Omega$ gilt. In einem zweiten Schritt kann man dann wie bei monotoner Konvergenz mit der Hilfsfolge $(f_n \mathbf{1}_{N^C})$ für die Nullmenge $N = \{\omega : f_n(\omega) \nrightarrow f(\omega)\}$ den Fall der μ-f.ü. Konvergenz zeigen.

Der Beweis beruht auf einer elementaren Erkenntniss: Wenn $|f_n| \leq g$ gilt, so gilt $f_n \leq g$ und $-f_n \leq g$ oder umgeformt auch $0 \leq f_n + g$ und $0 \leq g - f_n$. In anderen Worten: Wir können die f_n so geschickt verschieben, dass wir nicht-negative Funktionen bekommen und Fatou anwenden können.

(i)

$$\begin{aligned}
\int_\Omega f \,\mathrm{d}\mu + \int_\Omega g \,\mathrm{d}\mu &\overset{\text{Lin.}}{=} \int_\Omega (f+g) \,\mathrm{d}\mu \\
&\overset{\text{Ann.}}{=} \int_\Omega \big(\lim_{n\to\infty} f_n + g\big) \,\mathrm{d}\mu \\
&= \int_\Omega \big(\liminf_{n\to\infty} f_n + g\big) \,\mathrm{d}\mu \\
&\overset{3.2.4}{\leq} \liminf_{n\to\infty} \int_\Omega (f_n + g) \,\mathrm{d}\mu \\
&\overset{\text{Lin.}}{=} \liminf_{n\to\infty} \Big(\int_\Omega f_n \,\mathrm{d}\mu + \int_\Omega g \,\mathrm{d}\mu \Big) \\
&= \liminf_{n\to\infty} \int_\Omega f_n \,\mathrm{d}\mu + \int_\Omega g \,\mathrm{d}\mu.
\end{aligned}$$

Wenn wir nun auf beiden Seiten das Integral über g abziehen, dann bekommen wir

$$\int_\Omega f \,\mathrm{d}\mu \leq \liminf_{n\to\infty} \int_\Omega f_n \,\mathrm{d}\mu.$$

(ii) Das selbe Argument wenden wir auf $0 \leq g - f_n$ an. Dieselbe Rechnung gibt

$$\int_\Omega -f \,\mathrm{d}\mu \leq \liminf_{n\to\infty} \int_\Omega -f_n \,\mathrm{d}\mu \overset{\text{Lin.}}{=} \liminf_{n\to\infty} - \int_\Omega f_n \,\mathrm{d}\mu = -\limsup_{n\to\infty} \int_\Omega f_n \,\mathrm{d}\mu,$$

also

$$\limsup_{n\to\infty} \int_\Omega f_n \,\mathrm{d}\mu \leq \int_\Omega f \,\mathrm{d}\mu.$$

Beide Schritte zusammen ergeben

$$\int_\Omega f \,\mathrm{d}\mu \leq \liminf_{n\to\infty} \int_\Omega f_n \,\mathrm{d}\mu \leq \limsup_{n\to\infty} \int_\Omega f_n \,\mathrm{d}\mu \leq \int_\Omega f \,\mathrm{d}\mu.$$

Also stimmen lim inf und lim sup überein und geben nach Analysis 1 den Grenzwert

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

□

Beispiel. In beiden Beispielen sei $(\Omega, \mathcal{A}, \mu) = (\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$.

(i) Schauen wir uns mal ein Gegenbeispiel für die Konvergenzsätze an, dafür nehmen wir die Folge $f_n = n\mathbf{1}_{(0,\frac{1}{n})}$. Das sind Indikatorfunktionen, deren Breite kleiner wird, die Höhe allerdings größer wird. Es gilt punktweise $\lim_{n\to\infty} f_n = f$, wobei f die Nullfunktion ist. Wegen

$$\lim_{n\to\infty}\int_\mathbb{R} f_n \,\mathrm{d}\lambda \overset{\text{Höhe mal Breite}}{=} \lim_{n\to\infty} 1 = 1 \neq 0 = \int_\mathbb{R} f \,\mathrm{d}\lambda,$$

scheinen MCT und DCT nicht zu funktionieren. Warum? Die Annahme von MCT ist nicht erfüllt weil die Folge f_n nicht punktweise wächst. Die Folge ist auch nicht beschränkt durch „die selbe“ integrierbare Funktion g, daher ist die Annahme von DCT nicht erfüllt.

(ii) Für reelle Integrale können wir jetzt einmal schnell den Berechnungsweg

$$\int_\mathbb{R} f \,\mathrm{d}\lambda = \lim_{n\to\infty}\int_{[-n,n]} f \,\mathrm{d}\lambda$$

für $f \geq 0$ begründen. Das folgt direkt aus MCT mit der Wahl $f_n = f\mathbf{1}_{[-n,n]}$ (f wird also außerhalb von $[-n, n]$ auf 0 gesetzt) weil $f_n \uparrow f$ (das gilt nur für $f \geq 0$!). In dx-Notation steht da also, wie in Beispiel 3.1.12 (b) behauptet,

$$\int_\mathbb{R} f(x) \,\mathrm{d}x = \lim_{n\to\infty}\int_{-n}^{n} f(x) \,\mathrm{d}x.$$

Warnung: Gegenbeispiele wie $\frac{\sin(x)}{x}$ aus Beispiel 3.1.12 (c) zeigen, dass das Integral über ganz $\mathbb{R}$ nicht umbedingt der Grenzwert der Integrale von $-n$ bis n sein muss! Weil die Folge (f_n) in dem Fall nicht wachsend ist, kann in dem Fall MCT auch nicht angewandt werden.

Eine ganz wichtige Folgerung für die spätere Stochastik ist der Spezialfall, wenn μ ein endliches Maß (insbesondere ein Wahrscheinlichkeitsmaß) ist:

Korollar 3.2.6. Ist μ ein endliches Maß (z. B. ein Wahrscheinlichkeitsmaß) und $|f_n| \leq C$ für alle $n \in \mathbb{N}$ μ-f.ü., d. h. alle f_n sind *beschränkt* durch das gleiche C, und $\lim_{n\to\infty} f_n = f$ μ-f.ü., so gilt

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

Beweis. Das ist dominierte Konvergenz mit der Majorante $g \equiv C$. Als Indikatorfunktion ist die Majorante integrierbar, weil

$$\int_\Omega g \,\mathrm{d}\mu = \int_\Omega C\mathbf{1}_\Omega \,\mathrm{d}\mu \overset{\text{Def.}}{=} C\mu(\Omega) < \infty.$$

□

3.3 Das Beispiel - Integrale über Wahrscheinlichkeitsmaße auf $\mathcal{B}(\mathbb{R})$

Weil in dem Spezialfall $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P}_F)$ die Integrale später in der Stochastik unter dem Namen „Erwartunswerte" eine enorm wichtige Rolle spielen werden, schauen wir uns den Spezialfall jetzt schon mal in Ruhe an. Das gibt euch genug Zeit, die wichtigsten Rechnungen über das Semester mehrfach zu üben.

Kurze Erinnerung: Ein Wahrscheinlichkeitsmaß $\mathbb{P}_F$ auf $\mathcal{B}(\mathbb{R})$ mit Verteilungsfunktion F beschreibt ein reellwertiges Zufallsexperiment, bei dem die „gezogene Zahl" mit Wahrscheinlichkeit $\mathbb{P}_F((a,b]) = F(b) - F(a)$ in $(a,b]$ liegt. Hat F eine Dichte f, so gilt

$$F(t) = \int_{-\infty}^{t} f(x)\,\mathrm{d}x, \quad \text{also } \mathbb{P}_F((a,b]) = \int_a^b f(x)\,\mathrm{d}x.$$

Ist F diskret mit Werten $a_1, ..., a_N$ und Wahrscheinlichkeiten $p_1, ..., p_N$, so gilt

$$F(t) = \sum_{k=1}^{N} p_k \mathbf{1}_{[a_k,+\infty)} = \sum_{a_k \leq t} p_k, \quad \text{also } \mathbb{P}_F((a,b]) = \sum_{a < a_k \leq b} p_k.$$

Wir summieren also die Wahrscheinlichkeiten der Werte in $(a,b]$.

Ein paar konkreten Integralen geben wir jetzt Namen und überlegen uns anschließend, wie man die Integrale in vielen Beispielen berechnen kann.

Definition 3.3.1.

(i) Für $k \in \mathbb{N}$ heißt

$$\int_{\mathbb{R}} x^k \,\mathrm{d}\mathbb{P}_F(x)$$

k-tes Moment von $\mathbb{P}_F$ (oder k-tes Moment der Verteilungsfunktion F), wenn das Integral wohldefiniert ist.

(ii) Für $\lambda \in \mathbb{R}$ heißt

$$\int_{\mathbb{R}} e^{\lambda x} \,\mathrm{d}\mathbb{P}_F(x)$$

heißt **exponentielles Moment** von $\mathbb{P}_F$ (oder exponentielles Moment der Verteilungsfunktion F).

Allgemein betrachten wir für messbare Abbildungen $g : \mathbb{R} \to \overline{\mathbb{R}}$ auch die Integrale

$$\int_{\mathbb{R}} g(x) \,\mathrm{d}\mathbb{P}_F(x),$$

jedoch ohne ihnen extra einen Namen zu geben.

Beachte: Alle Integrale über nicht-negative messbare Integranden sind wohldefiniert, das Integral könnte aber $+\infty$ sein. Damit sind exponentielle Momente und gerade Momente immer in $[0, +\infty]$ definiert, existieren nach unserer Konvention aber nur, wenn sie endlich sind.

Satz 3.3.2. **[Integrale für absolutstetige Verteilungen]** Sei $\mathbb{P}_F$ ein Wahrscheinlichkeitsmaß auf $\mathcal{B}(\mathbb{R})$ und F habe Dichte f. Dann gilt für $g\colon \mathbb{R} \to \overline{\mathbb{R}}$ Borel-messbar:

$$\int_{\mathbb{R}} g \,\mathrm{d}\mathbb{P}_F \text{ ist wohldefiniert} \quad \Leftrightarrow \quad \int_{\mathbb{R}} g(x) f(x)\,\mathrm{d}x \text{ ist wohldefiniert} \tag{3.5}$$

und, wenn die Integrale wohldefiniert sind, gilt die Rechenregel

$$\int_{\mathbb{R}} g \,\mathrm{d}\mathbb{P}_F = \int_{\mathbb{R}} g(x) f(x)\,\mathrm{d}x.$$

Der Satz besagt, dass wir die abstrakten Integrale $\int_{\mathbb{R}} g \, d\mathbb{P}_F$ durch sehr viel weniger abstrakte Integrale behandeln können. Beachtet, dass mit $\int_{\mathbb{R}} g(x)f(x) \, dx$ das Lebesgue Integral $\int_{\mathbb{R}} gf \, d\lambda$ gemeint ist. Nach dem Beweis diskutieren wir nochmal ausführlich, wie ihr das mit den Tricks aus der Analysis berechnen könnt.

Beweis.

(i) Für $g \geq 0$ starten wir die Gebetsmühle der Integrationstheorie:

- Sei zunächst $g = \mathbf{1}_A$ für $A \in \mathcal{B}(\mathbb{R})$. Es gilt

$$\int_{\mathbb{R}} g \, d\mathbb{P}_F \overset{\text{Def.}}{=} 1 \cdot \mathbb{P}_F(A) \overset{(*)}{=} \int_A f(x) \, dx = \int_{\mathbb{R}} \mathbf{1}_A(x) f(x) \, dx = \int_{\mathbb{R}} g(x) f(x) \, dx.$$

  Warum gilt $(*)$, also $\mathbb{P}_F(A) = \int_A f(x) \, dx$? Ist $A = (a, b]$, so gilt $\mathbb{P}_F((a, b)) = \int_a^b f(x) \, dx$ weil F Dichte f hat. In Anwendung 3.2.2 haben wir gezeigt, dass $\nu(A) = \int_A f(x) \, dx$ ein Maß auf $\mathcal{B}(\mathbb{R})$ ist. Es gilt also $\mathbb{P}_F = \nu$ auf einem $\cap$-stabilen Erzeuger von $\mathcal{B}(\mathbb{R})$ und damit aufgrund von Korollar 1.2.12 auch auf $\mathcal{B}(\mathbb{R})$. Also gilt $(*)$ für alle $A \in \mathcal{B}(A)$.

- Ist g eine einfache Funktion, so folgt nun

$$\int_{\mathbb{R}} g \, d\mathbb{P}_F = \int_{\mathbb{R}} g(x) f(x) \, dx$$

  aufgrund der Linearität des Integrals.

- Ist $g \geq 0$, wählen wir eine Folge $(g_n) \subseteq \mathcal{E}^+$ mit $g_n \uparrow g$, $n \to \infty$. Die Folge existiert weil g messbar ist. Mit dem Monotone Konvergenz Theorem und dem gerade Gezeigen für einfache Funktionen folgt

$$\int_{\mathbb{R}} g \, d\mathbb{P}_F \overset{3.2.1}{=} \lim_{n \to \infty} \int_{\mathbb{R}} g_n \, d\mathbb{P}_F = \lim_{n \to \infty} \int_{\mathbb{R}} g_n(x) f(x) \, dx \overset{3.2.1}{=} \int_{\mathbb{R}} g(x) f(x) \, dx.$$

(ii) Für g beliebig zerlegt man g in $g^+ - g^-$ und wende (i) auf die Integrale der Positiv- und Negativteile an. Vergleiche den Anfang des Beweises von Korollar 3.1.17.

$\square$

Weil ihr mit dem Satz in dieser Vorlesung viel rechnen werdet, verbinden wir den Satz noch schnell mit Beispiel 3.1.12. Das Integral auf der rechten Seite der Äquivalenzen ist das Lebesgue Integral $\int_{\mathbb{R}} gf \, d\lambda$ in der weniger angstverbreitenden dx-Notation. Für nette Integranden (stückweise stetig) ist das Integral aufgrund der Diskussion aus Beispiel 3.1.12 wie folgt zu berechnen ist:

- Ist $g \geq 0$ (z. B. $g(x) = e^x$), so ist alles einfach: Weil f als Dichte immer nicht-negativ ist, ist auch das Produkt gf nicht-negativ. Damit ist das Integral <u>immer</u> wohldefiniert (es könnte aber unendlich sein). Daher könnt ihr den ersten Teil im Satz ignorieren und sofort

$$\int_{\mathbb{R}} g \, d\mathbb{P}_F = \int_{\mathbb{R}} g(x) f(x) \, dx \overset{\text{MCT}}{=} \lim_{n \to \infty} \int_{-n}^{n} g(x) f(x) \, dx$$

  mit den Tricks der Analysis ausrechnen.

- Habt ihr Pech, so gilt $g \geq 0$ nicht (z. B. für $g(x) = x$). Dann überlegt ihr euch, was Positivteil g^+ ist und was Negativteil g^- ist. Weil beide nicht-negativ sind, macht ihr 2x das gerade beschriebene, ihr berechnet also $\int_{\mathbb{R}} g^+(x) f(x) \, dx$ und $\int_{\mathbb{R}} g^-(x) f(x) \, dx$. Wenn eines von beiden (oder beide) endlich ist, so ist nach dem Satz $\int_{\mathbb{R}} g \, d\mathbb{P}_F$ wohldefiniert und ihr habt den Wert

$$\int_{\mathbb{R}} g \, d\mathbb{P}_F = \int_{\mathbb{R}} g^+(x) f(x) \, dx - \int_{\mathbb{R}} g^-(x) f(x) \, dx$$

  schon ausgerechnet. Wir haben dabei $(gf)^+ = g^+ f$ und $(gf)^- = g^- f$ benutzt, weil f als Dichte nicht-negativ ist. Vorsicht: In diesem zweiten Fall dürft ihr <u>nicht</u> einfach $\int_{\mathbb{R}} g \, d\mathbb{P}_F = \lim_{n \to \infty} \int_{-n}^{n} g(x) f(x) \, dx$ nutzen! Nach Beispiel 3.1.12 (iii) kann das falsch sein! Wenn ihr neugierig seid, könnt ihr für ein konkretes Beispiel zu Warnung 3.3.7 vorblättern.

Angelehnt an (3.3) können wir auch folgendes formulieren, wenn wir uns nur für die Endlichkeit des Integrals interessieren:

$$g \text{ ist } \mathbb{P}_F\text{-integrierbar} \quad \Longleftrightarrow \quad \int_{\mathbb{R}} g \,\mathrm{d}\mathbb{P}_F \text{ existiert} \quad \Leftrightarrow \quad \int_{\mathbb{R}} |g(x)| f(x) \,\mathrm{d}x < \infty. \tag{3.6}$$

Vorlesung 13

Im diskreten Fall nehmen wir mal diese Formulierung:

Satz 3.3.3. **[Integrale für diskrete Verteilungen]** Sei $\mathbb{P}_F$ ein Maß auf $\mathcal{B}(\mathbb{R})$ und F sei diskret, d. h.

$$F(t) = \sum_{k=1}^{N} p_k \mathbf{1}_{[a_k,\infty)}(t), \quad t \in \mathbb{R},$$

mit $N \in \mathbb{N} \cup \{+\infty\}$, $a_1, ..., a_N \in \mathbb{R}$ und $\sum_{k=1}^{N} p_k = 1$. Dann gilt für $g : \mathbb{R} \to \overline{\mathbb{R}}$ messbar:

$$g \ \mathbb{P}_F\text{-integrierbar} \quad \Leftrightarrow \quad \sum_{k=1}^{N} |g(a_k)| p_k < \infty$$

und, wenn g $\mathbb{P}_F$-integrierbar ist, gilt

$$\int_{\mathbb{R}} g \,\mathrm{d}\mathbb{P}_F = \sum_{k=1}^{N} g(a_k) p_k.$$

Zu beachten ist, dass in vielen diskreten Modellen N endlich ist und g die Werte $+\infty$ und $-\infty$ nicht annimmt, dann ist das Integral $\int_{\mathbb{R}} g \,\mathrm{d}\mathbb{P}_F$ natürlich immer definiert und ihr merkt euch einfach die Rechenregel $\int_{\mathbb{R}} g \,\mathrm{d}\mathbb{P}_F = \sum\limits_{k=1}^{N} g(a_k) p_k$.

Beweis.

(i) Sei zunächst $g \geq 0$:

(a) Am einfachsten ist der Fall $N \in \mathbb{N}$, denn dann sprechen wir nur von endlichen Summen. Weil das Maß $\mathbb{P}_F$ von der Form $\mathbb{P}_F = \sum_{k=1}^{N} p_k \delta_{a_k}$ ist, gilt $g = g\mathbf{1}_{\{a_1,...,a_N\}}$ $\mathbb{P}_F$-fast überall. Es folgt dann

$$\begin{aligned}
\int_{\mathbb{R}} g \,\mathrm{d}\mathbb{P}_F &\overset{\text{Satz 3.1.15}}{=} \int_{\mathbb{R}} g \mathbf{1}_{\{a_1,...,a_N\}} \,\mathrm{d}\mathbb{P}_F \\
&= \int_{\mathbb{R}} g \sum_{k=1}^{N} \mathbf{1}_{\{a_k\}} \,\mathrm{d}\mathbb{P}_F \\
&= \int_{\mathbb{R}} \sum_{k=1}^{N} g(a_k) \mathbf{1}_{\{a_k\}} \,\mathrm{d}\mathbb{P}_F \\
&\overset{\text{Lin.}}{=} \sum_{k=1}^{N} \int_{\mathbb{R}} \underbrace{g(a_k) \mathbf{1}_{\{a_k\}}}_{\text{einfach}} \,\mathrm{d}\mathbb{P}_F \\
&\overset{\text{Def. Int.}}{=} \sum_{k=1}^{N} g(a_k) \mathbb{P}_F(\{a_k\}) \overset{\text{Def. Maß}}{=} \sum_{k=1}^{N} g(a_k) p_k.
\end{aligned}$$

(b) $N = +\infty$ funktioniert im Prinzip genauso, wir müssen nur einmal monotone Konvergenz für die wachsende Folge von messbaren Funktionen $g_n := \sum_{k=1}^{n} g(a_k) \mathbf{1}_{\{a_k\}}$

nutzen, um Integral und Summe zu vertauschen:

$$\begin{aligned}
\int_{\mathbb{R}} g \,\mathrm{d}\mathbb{P}_F &= \int_{\mathbb{R}} g \mathbf{1}_{\{a_1,a_2,\dots\}} \,\mathrm{d}\mathbb{P}_F \\
&= \int_{\mathbb{R}} g \sum_{k=1}^{\infty} \mathbf{1}_{\{a_k\}} \,\mathrm{d}\mathbb{P}_F \\
&\overset{\text{Def. Reihe}}{=} \int_{\mathbb{R}} \lim_{n\to\infty} \sum_{k=1}^{n} g(a_k)\mathbf{1}_{\{a_k\}} \,\mathrm{d}\mathbb{P}_F \\
&\overset{3.2.1}{=} \lim_{n\to\infty} \int_{\mathbb{R}} \sum_{k=1}^{n} g(a_k)\mathbf{1}_{\{a_k\}} \,\mathrm{d}\mathbb{P}_F \\
&\overset{\text{Lin.}}{=} \lim_{n\to\infty} \sum_{k=1}^{n} \int_{\mathbb{R}} \underbrace{g(a_k)\mathbf{1}_{\{a_k\}}}_{\text{einfach}} \,\mathrm{d}\mathbb{P}_F \\
&\overset{\text{Def. Int.}}{=} \sum_{k=1}^{\infty} g(a_k)\mathbb{P}_F(\{a_k\}) \overset{\text{Def. Maß}}{=} \sum_{k=1}^{\infty} g(a_k)p_k.
\end{aligned}$$

(ii) Sei nun g messbar, aber nicht mehr nicht-negativ. Es gilt also wegen (3.3) und Teil (i)

$$\int_{\mathbb{R}} g \,\mathrm{d}\mathbb{P}_F \text{ existiert} \quad\Leftrightarrow\quad \int_{\mathbb{R}} |g| \,\mathrm{d}\mathbb{P}_F < \infty \quad\Leftrightarrow\quad \sum_{k=1}^{N} |g(a_k)|p_k < \infty.$$

Wenn das Integral existiert, gilt wegen (i)

$$\begin{aligned}
\int_{\mathbb{R}} g \,\mathrm{d}\mathbb{P}_F &\overset{\text{Def.}}{=} \int_{\mathbb{R}} g^+ \,\mathrm{d}\mathbb{P}_F - \int_{\mathbb{R}} g^- \,\mathrm{d}\mathbb{P}_F \\
&\overset{\text{(i)}}{=} \sum_{k=1}^{N} g^+(a_k)p_k - \sum_{k=1}^{N} g^-(a_k)p_k \\
&= \sum_{k=1}^{N} \big(g^+(a_k) - g^-(a_k)\big)p_k = \sum_{k=1}^{N} g(a_k)p_k.
\end{aligned}$$

□

Wir werden in der Stochastik das 1.te Moment Erwartungswert nennen, das soll also so etwas wie der Mittelwert über Versuchsausführungen sein (In Vorlesung 26 wird das mit dem Gesetz der großen Zahlen klar). Um mal zwei Beispiele konkret mit den gerade gezeigten Rechenregeln auszurechnen, schauen wir uns die Erwartungswerte (1.te Momente) vom Würfelexperiment und vom gleichverteilten Ziehen aus $[0,1]$ mal an:

Beispiel 3.3.4.

- Seien $a_1 = 1, \dots, a_6 = 6$ und $p_1 = \dots = p_6 = \frac{1}{6}$. Dann gilt

$$\int_{\mathbb{R}} x \,\mathrm{d}\mathbb{P}_F(x) = \sum_{k=1}^{6} k\frac{1}{6} = 3,5.$$

  Dies ist so ein Beispiel, bei dem die Summe natürlich endlich ist (N endlich, g endlich), wir uns also gar keine Gedanken um Wohldefiniertheit und Endlichkeit machen müssen. Wir können Satz 3.3.3 also als einfache Rechenregeln benutzen.

- Sei jetzt $\mathbb{P}_F$ absolutstetig mit Dichte $f = \mathbf{1}_{[0,1]}$, dafür nutzen wir Satz 3.3.2. Der Integrand von $\int_{\mathbb{R}} xf(x)\,\mathrm{d}x$ ist nicht-negativ, also sind alle Integrale wohldefiniert (könnten aber den Werte $+\infty$ annehmen). Wir können also direkt die Rechenregel aus dem Satz benutzen:

$$\int_{\mathbb{R}} x \,\mathrm{d}\mathbb{P}_F(x) = \int_{\mathbb{R}} xf(x)\,\mathrm{d}x = \int_{\mathbb{R}} x\mathbf{1}_{[0,1]}(x)\,\mathrm{d}x = \int_0^1 x\,\mathrm{d}x = \Big[\frac{1}{2}x^2\Big]_0^1 = \frac{1}{2}.$$

Falls ihr eine grobe Vorstellung von dem Begriff Erwartungswert habt, so sollten diese zwei Beispiele dazu passen. In ein paar Wochen wird euch das auch klar sein.

Warnung 3.3.5. Es ist nicht immer der Fall, dass eine Verteilungsfunktion diskret oder absolutstetig ist. In diesen Fällen gibt es keine einfach Formel für $\int_{\mathbb{R}} g \, \mathrm{d}\mathbb{P}_F$! Es gibt drei Typen von Verteilungsfunktionen:

- F ist absolutstetig,
- F ist diskret,
- F ist „stetigsingulär“ (stetig, hat aber keine Dichte).

Jede stetige Verteilungsfunktion lässt sich zerlegen in absolutstetigen und stetigsingulären Anteil (Satz von Lebesgue). Das geht hier aber zu weit, Beispiele für stetigsinguläre Verteilungen sind tricky (z. B. die Cantorverteilung).

In vielen Beispielen müssen wir gar nicht rechnen, sondern sehen das Ergebniss direkt. Kurze Erinnerung an die Schule: f heißt punktsymmetrisch, falls $f(x) = -f(-x)$ und achsensymmetrisch, falls $f(x) = f(-x)$ für alle $x \in \mathbb{R}$. Für integrierbare punktsymmetrische Funktionen gilt $\int_{\mathbb{R}} f(x) \, \mathrm{d}x = 0$. Das können wir direkt ausnutzen, um viele Momente direkt als 0 zu erkennen:

Lemma 3.3.6. Ist F absolutstetig mit Dichte f und ist das $(2k+1)$-te Moment von F wohldefiniert, so gilt

$$f \text{ achsensymmetrisch} \quad \Rightarrow \quad \int_{\mathbb{R}} x^{2k+1} \, \mathrm{d}\mathbb{P}_F(x) = 0.$$

Beweis. Ist f achsensymmetrisch, so ist $h(x) := x^{2k+1} f(x)$ punktsymmetrisch weil dann $h(-x) = (-x)^{2k+1} f(-x) = -x^{2k+1} f(x) = -h(x)$. Wegen 3.3.2 ist also

$$\int_{\mathbb{R}} x^{2k+1} \, \mathrm{d}\mathbb{P}_F(x) = \int_{\mathbb{R}} x^{2k+1} f(x) \, \mathrm{d}x \overset{\text{punktsym.}}{=} 0.$$

□

Warnung 3.3.7. **[Cauchyverteilung]** Wir müssen in Lemma 3.3.6 auf jeden Fall annehmen, dass die Momente existieren. Beispielsweise hat die Cauchyverteilung eine achsensymmetrische Dichte, das erste Moment ist aber gar nicht definiert, damit insbesondere nicht 0! Dies ist einer der fiesen $\infty - \infty$ Fälle. Rechnen wir das mal nach. Wir betrachten also die Cauchy-Dichte $f(x) = \frac{1}{\pi} \frac{1}{1+x^2}$ und berechnen zunächst den Positivteil, wobei wir die konkrete Form des Positivteils x^+ einsetzen:

$$\begin{aligned}
\int_{\mathbb{R}} x^+ \, \mathrm{d}\mathbb{P}_F(x) &\overset{3.3.2, x^+ \geq 0}{=} \frac{1}{\pi} \int_{\mathbb{R}} x^+ \frac{1}{1+x^2} \, \mathrm{d}x \\
&= \frac{1}{\pi} \int_{-\infty}^{\infty} x \mathbf{1}_{[0,\infty)}(x) \frac{1}{1+x^2} \, \mathrm{d}x \\
&= \frac{1}{\pi} \int_0^{\infty} \frac{1}{1/x + x} \, \mathrm{d}x \\
&\geq \frac{1}{\pi} \int_1^{\infty} \frac{1}{1+x} \, \mathrm{d}x \\
&\overset{3.2.1}{=} \frac{1}{\pi} \lim_{N \to \infty} \int_1^N \frac{1}{1+x} \, \mathrm{d}x \\
&\overset{\text{Hauptsatz}}{=} \frac{1}{\pi} \lim_{N \to \infty} \Big[\ln(1+x) \Big]_1^N = +\infty.
\end{aligned}$$

Genauso gibt die selbe Rechnung, mit der konkreten Form des Negativteils x^-,

$$\int_{\mathbb{R}} x^- \,\mathrm{d}\mathbb{P}_F(x) \overset{3.3.2, x^+ \geq 0}{=} \frac{1}{\pi} \int_{-\infty}^{0} -x \frac{1}{1+x^2} \,\mathrm{d}x = \frac{1}{\pi} \int_0^\infty x \frac{1}{1+x^2} \,\mathrm{d}x = +\infty.$$

Damit ist $\int_{\mathbb{R}} x \,\mathrm{d}\mathbb{P}_F(x) = \int_{\mathbb{R}} x^+ \,\mathrm{d}\mathbb{P}_F(x) - \int_{\mathbb{R}} x^- \,\mathrm{d}\mathbb{P}_F(x) = +\infty - (+\infty)$ nicht wohldefiniert. Für die Cauchyverteilung ist das erste Moment also nicht wohldefiniert!

Die Abschätzung für die Cauchyverteilung war nicht so einfach. Daher wäre es nützlich, den Integralen direkt anzusehen, ob sie existieren, oder nicht. Wenn man weis was zu tun ist, dann ist die formelle Rechnung viel leichter. Hier ist eine grobe Heuristik:

Bemerkung 3.3.8. **[Heuristik mit Dichten]** Wie „sieht“ man, ob ein Integral existiert? Man vergleicht mit bekannten Integralen. Erinnern wir uns kurz an die Analysisvorlesung:

$$\begin{aligned}
\int_1^\infty \frac{1}{x^a} \,\mathrm{d}x &\overset{3.2.1}{=} \lim_{N\to\infty} \int_1^N \frac{1}{x^a} \,\mathrm{d}x \\
&= \lim_{N\to\infty} \begin{cases} \Big[\ln(x)\Big]_1^N & : a = 1 \\ \frac{1}{1-a}\Big[x^{1-a}\Big]_1^N & : a \neq 1 \end{cases} \\
&= \lim_{N\to\infty} \begin{cases} \ln(N) & : a = 1 \\ \frac{1}{1-a}(N^{1-a} - 1) & : a \neq 1 \end{cases} \\
&= \begin{cases} +\infty & : a \leq 1 \\ \frac{1}{a-1} < \infty & : a > 1 \end{cases}.
\end{aligned}$$

Unsere Heuristik für die Integrierbarkeit bei unendlich ist es, grob mit der Funktion $\frac{1}{x}$ zu vergleichen. Fällt ein Integrand f deutlich schneller gegen 0, ist vermutlich $\int_1^\infty f(x) \,\mathrm{d}x < \infty$. Fällt der Integrand langsamer als $\frac{1}{x}$ gegen 0, so ist sicherlich $\int_1^\infty f(x) \,\mathrm{d}x = +\infty$.
Hier sind ein paar Beispiele:

(i) Nochmal die Cauchyverteilung: $\frac{1}{\pi} x \frac{1}{1+x^2}$ ist bei ∞ ungefähr wie $\frac{1}{x}$, das ist aber *nicht* integrierbar. Die Heuristik sagt uns also auf einen Blick, dass etwas schief gehen sollte. Um daraus ein sauberes Argument zu machen, müssen wir leider doch die Abschätzung aus 3.3.7 durchgehen.

(ii) Für welches β hat $\mathrm{Exp}(\lambda)$ ein endliches exponentielle Moment, wann ist also

$$\lambda \int_{\mathbb{R}} e^{\beta x} e^{-\lambda x} \mathbf{1}_{[0,\infty)}(x) \,\mathrm{d}x = \lambda \int_0^\infty e^{x(\beta - \lambda)} \,\mathrm{d}x$$

endlich? Weil e^{ax} für $a > 0$ schneller als jedes Polynom wächst, fällt e^{-ax} ($\lambda > 0$) viel schneller als jedes Polynom gegen 0. Also sind alle exponentiellen Momente genau dann endlich, wenn $\beta < \lambda$. In dem Fall können wir alles natürlich sofort ausrechnen, weil der Integrand eine einfache Stammfunktion hat.

(iii) Für welches β hat $\mathcal{N}(\mu, \sigma^2)$ ein endliches exponentielles Moment, wann ist also

$$\frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} e^{\beta x} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \,\mathrm{d}x$$

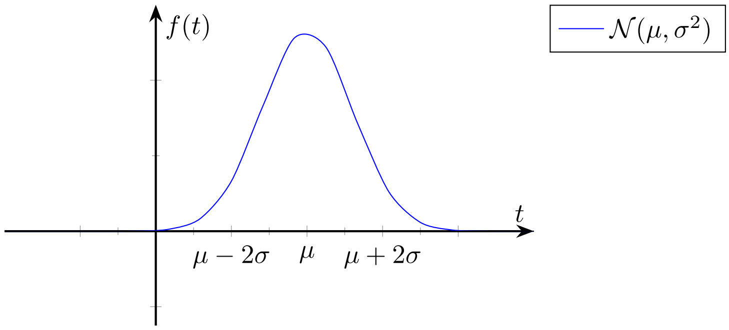
definiert? Natürlich geht $e^{\beta x} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ viel schneller als $\frac{1}{x}$ gegen 0, weil x^2 schneller wächst als x, und die Exponentialfunktion alles polynomielle dominiert.

Um einen ersten Eindruck zu bekommen, warum Momente überhaupt nützlich sind, schauen wir uns eine Variante der Markov-Ungleichung an. Wir sehen hier, dass wir mit den Momenten etwas über die Verteilung der Masse aussagen können. Schauen wir uns dazu die Konzentration der Masse der Normalverteilung um μ an:

Beispiel 3.3.9. **[Konzentrationsungleichung]** Wir kennen bereits die Normalverteilung $\mathcal{N}(\mu, \sigma^2)$. Der Parameter μ verschiebt die Stelle des Maximalpunktes der Dichte an die Stelle μ, der Parameter σ^2 ist für die Stauchung zuständig. Für kleines σ^2 ist die Dichte „spitzer", für großes σ^2 flacher. Qualitativ wissen wir schon, dass viel Masse nah bei μ liegt, dort ist die Dichte schließlich groß (vergleiche die Diskussion 1.4.13). Können wir das auch genauer machen? Wenn jemand fragt, wie weit man von μ weggehen muss, so dass beispielsweise 99,7 Prozent der Masse in $[\mu - a, \mu + a]$ liegt, was antworten wir? Schauen wir das Bildchen an: Reicht der grüne Bereich, um $\mathbb{P}([\mu - a, \mu + a]) \approx 0,997$ zu erreichen? Vermutlich nicht. Grün, rot und orange zusammen könnten vom Bild her aber hinhauen. Solche Fragen durch Ungleichungen möglichst gut zu beantworten, sind das Themengebiet der **Konzentrationsungleichungen**, also Ungleichungen der Art

$$\mathbb{P}([a, b]) \leq ... \quad \text{oder} \quad \mathbb{P}([a, b]) \geq ...$$

Das ist natürlich einfach zu beantworten, wenn die Dichte F von $\mathbb{P}$ explizit ist, weil dann $\mathbb{P}_F([a, b]) = F(b) - F(a)$ gilt. Bei der Exponentialverteilung braucht man z. B. nichts abzuschätzen weil F eine einfache Formel hat. Für die Normalverteilung geht das allerdings nicht, F ist als Integral gegeben und das Integral kann nicht ausgerechnet werden.



Das bunte Bildchen visualisiert die sogenannte 1-2-3-σ-Regel für $\mathcal{N}(\mu, \sigma^2)$. Die Regel besagt, dass

- in $[\mu - \sigma, \mu + \sigma]$ ungefähr 68 Prozent (also $0,68$) der Masse liegt,
- in $[\mu - 2\sigma, \mu + 2\sigma]$ ungefähr 95 Prozent (also $0,95$) der Masse liegt,
- in $[\mu - 3\sigma, \mu + 3\sigma]$ ungefähr 99,7 Prozent (also $0,997$) der Masse liegt.

Weil σ auch Standardabweichung genannt wird, heißt das in Worten: „Um zwei Standardabweichungen nach links und rechts von μ liegt 95 Prozent der Masse von $\mathcal{N}(\mu, \sigma^2)$".

Wir zeigen jetzt mit der Markov-Ungleichung unsere erste Integralabschätzung und versuchen uns danach an einem Teil der 1-2-3-σ-Regel.

Proposition 3.3.10. **[Markov-Ungleichung für Polynome]** Sei $\mathbb{P}_F$ ein Wahrscheinlichkeitsmaß auf $\mathcal{B}(\mathbb{R})$, sodass für eine gerade natürliche Zahl $2k$ das $2k$-te Moment existiert ist. Dann gilt für alle $a > 0$

$$\mathbb{P}_F\big([-a, a]\big) \geq 1 - \frac{\int_{\mathbb{R}} x^{2k} \, \mathrm{d}\mathbb{P}_F(x)}{a^{2k}}.$$

Gleichbedeutend (Gegenereignis) gilt

$$\mathbb{P}_F\big([-a, a]^C\big) \leq \frac{\int_{\mathbb{R}} x^{2k} \, \mathrm{d}\mathbb{P}_F(x)}{a^{2k}}.$$

Beweis. Der Beweis ist tatsächlich sehr einfach und basiert auf dem kleinen Trick, den wir schon ein paar mal gesehen haben. Wir mogeln einen Indikator über eine Menge in das Integral und schätzen auf dem Indikator die Funktion ab. Das geht natürlich nur, wenn der Indikator über eine messbare Menge so gewählt wird, dass die Menge etwas mit dem Integranden zu tun hat. Wir nehmen dazu den Integranden $g(x) = x^{2k}$ und die Menge $[-a,a]^C$. Für $x \in [-a,a]^C$ gilt natürlich $|x| > a$ und damit, $2k$ ist gerade, $x^{2k} = |x|^{2k} \geq a^{2k}$. Dann müssen wir nur noch die Monotonie vom Integral nutzen:

$$\begin{aligned}\int_{\mathbb{R}} x^{2k}\,\mathrm{d}\mathbb{P}_F(x) &\overset{\text{Mon.}}{\geq} \int_{\mathbb{R}} x^{2k}\mathbf{1}_{[-a,a]^C}(x)\,\mathrm{d}\mathbb{P}_F(x)\\ &\geq \int_{\mathbb{R}} \underbrace{a^{2k}\mathbf{1}_{[-a,a]^C}(x)}_{\text{einfach}}\,\mathrm{d}\mathbb{P}_F(x)\\ &\overset{\text{Def.}}{=} a^{2k}\mathbb{P}_F([-a,a]^C).\end{aligned}$$

Auflösen gibt

$$\mathbb{P}_F([-a,a]^C) \leq \frac{\int_{\mathbb{R}} x^{2k}\,\mathrm{d}\mathbb{P}_F(x)}{a^{2k}}.$$

Die zweite Ungleichung ist die Gegenwahrscheinlichkeit, weil für Wahrscheinlichkeitsmaße $\mathbb{P}_F(B) = 1 - \mathbb{P}_F(B^C)$ gilt. □

Kommen wir zurück zur Konzentration der Normalverteilung um μ.

Beispiel 3.3.11. Aufgabe: Sei $\mathbb{P}$ normalverteilt mit Parametern $\mu = 0$ und $\sigma^2 > 0$. Finde ein $a > 0$ mit $\mathbb{P}([-a,a]) \approx 0.997$. Damit uns die Markov-Ungleichung explizite Zahlen gibt, brauchen wir Formel für gerade Momente, wir nehmen einfach mal das 2.te und das 8.te. Für 2.te und 8.te Momente der Normalverteilungen gelten, das sehen wir später (oder ihr rechnet die Integrale von Hand aus),

$$\int_{\mathbb{R}} x^2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,\mathrm{d}x = \sigma^2 \quad \text{und} \quad \int_{\mathbb{R}} x^8 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,\mathrm{d}x = 105\sigma^8.$$

Um die Konzentration der Normalverteilung in $[-a,a]$ abzuschätzen, verwenden wir Proposition 3.3.10 einmal mit $2k = 2$ und einmal mit $2k = 8$:

$$\mathbb{P}([-a,a]) \geq 1 - \frac{\sigma^2}{a^2} \quad \text{sowie} \quad \mathbb{P}([-a,a]) \geq 1 - \frac{105\sigma^8}{a^8}.$$

Wir probieren jetzt mal aus, welche Abschätzung besser ist. Einsetzen, umformen und in Taschenrechner einsetzen, gibt beim zweiten Moment ungefähr

$$1 - \frac{\sigma^2}{a^2} \geq 0.997 \quad \Leftrightarrow \quad a \geq 18,26 \cdot \sigma$$

und beim 8.ten Moment ungefähr

$$1 - \frac{105\sigma^8}{a^8} \geq 0.997 \quad \Leftrightarrow \quad a \geq 3,69 \cdot \sigma.$$

Das ist jetzt zwar nicht ganz an an der richtigen Lösung aus Beispiel 3.3.9, aber für das achte Moment auch nicht ganz weit davon entfernt.

Vorlesung 14

3.4 Integralabschätzungen und L^p-Räume

Sei jetzt wieder $(\Omega, \mathcal{A}, \mu)$ ein beliebiger messbarer Raum und $f\colon \Omega \to \overline{\mathbb{R}}$ sei $(\mathcal{A}, \mathcal{B}(\overline{\mathbb{R}}))$-messbar. Wir werden in diesem Kapitel mehrfach nutzen, dass wegen Satz 3.1.15 folgende Äquivalenz gilt:

$$\int_{\Omega} |f|\,\mathrm{d}\mu = 0 \quad \Leftrightarrow \quad |f| = 0 \ \mu\text{-fast überall} \quad \Leftrightarrow \quad f = 0 \ \mu\text{-fast überall}.$$

Satz 3.4.1. **[Hölder-Ungleichung]** Seien $p, q > 1$ mit $\frac{1}{p} + \frac{1}{q} = 1$. Dann gilt

$$\int_\Omega |fg| \, \mathrm{d}\mu \leq \Big(\int_\Omega |f|^p \, \mathrm{d}\mu \Big)^{\frac{1}{p}} \Big(\int_\Omega |g|^q \, \mathrm{d}\mu \Big)^{\frac{1}{q}}.$$

Beweis. Alle auftretenden Integranden sind messbar und nicht-negativ, also sind alle Integrale definiert, $+\infty = +\infty$ ist aber möglich. Wir erinnern an die Young-Ungleichung aus Analysis 2 (das ist gerade die Konkavität des Logarithmus): Für $\alpha, \beta \geq 0$ gilt

$$\alpha\beta \leq \frac{\alpha^p}{p} + \frac{\beta^q}{q}.$$

Ist ein Faktor der rechten Seite der Hölder-Ungleichung 0 oder $+\infty$, so ist nichts zu zeigen. Das ist sofort klar für $+\infty$, aber auch der Fall 0 ist nicht schwer: Wenn nämlich $(\int_\Omega |f|^p \, \mathrm{d}\mu)^{1/p} = 0$ gilt, so muss $f = 0$ μ-fast überall gelten. Also ist auch $|fg| = 0$ μ-fast überall und damit ist auch die linke Seite 0. Die Ungleichung ergibt dann also $0 \leq 0$ und das ist richtig. Also nehmen wir an, beide Faktoren sind positiv. Wir definieren

$$\sigma = \Big(\int_\Omega |f|^p \, \mathrm{d}\mu \Big)^{\frac{1}{p}} > 0 \quad \text{und} \quad \tau = \Big(\int_\Omega |g|^q \, \mathrm{d}\mu \Big)^{\frac{1}{q}} > 0$$

sowie die messbaren Abbildungen

$$\alpha = \frac{|f|}{\sigma} \quad \text{und} \quad \beta = \frac{|g|}{\tau}.$$

Mit Young folgt

$$\frac{|f(\omega)g(\omega)|}{\sigma\tau} \leq \frac{|f(\omega)|^p}{\sigma^p p} + \frac{|g(\omega)|^q}{\tau^q q}, \quad \forall \omega \in \Omega.$$

Integrieren beider Seiten gibt wegen der Monotonie des Integrals

$$\frac{1}{\sigma\tau} \int_\Omega |fg| \, \mathrm{d}\mu \leq \frac{\int_\Omega |f|^p \, \mathrm{d}\mu}{\sigma^p p} + \frac{\int_\Omega |g|^q \, \mathrm{d}\mu}{\tau^q q} = \frac{1}{p} + \frac{1}{q} = 1.$$

Durchmultiplizieren gibt die Höldersche Ungleichung. □

Der Fall $p = q = 2$ heißt auch Cauchy-Schwarz:

$$\Big(\int_\Omega |fg| \, \mathrm{d}\mu \Big)^2 \leq \int_\Omega |f|^2 \, \mathrm{d}\mu \int_\Omega |g|^2 \, \mathrm{d}\mu.$$

Korollar 3.4.2. Ist $\mathbb{P}$ ein Wahrscheinlichkeitsmaß auf $\mathcal{A}$, so gilt

$$\left(\int_\Omega |f| \, \mathrm{d}\mathbb{P} \right)^p \leq \int_\Omega |f|^p \, \mathrm{d}\mathbb{P}$$

für alle $p > 1$.

Beweis. Wegen $\mathbb{P}(\Omega) = 1$ gilt mit $q = \frac{1}{1-p}$ wegen der Hölder Ungleichung

$$\int_\Omega |f| \, \mathrm{d}\mathbb{P} = \int_\Omega |f \cdot 1| \, \mathrm{d}\mathbb{P} \leq \left(\int_\Omega |f|^p \, \mathrm{d}\mathbb{P} \right)^{\frac{1}{p}} \left(\int_\Omega \underbrace{|1|^q}_{=1 \cdot \mathbf{1}_\Omega} \, \mathrm{d}\mathbb{P} \right)^{\frac{1}{q}} = \left(\int_\Omega |f|^p \, \mathrm{d}\mathbb{P} \right)^{\frac{1}{p}} \cdot 1,$$

also die Behauptung. □

Satz 3.4.3. **[Minkowski-Ungleichung]** Sei $p \geq 1$, so gilt

$$\Big(\int_\Omega |f + g|^p \, \mathrm{d}\mu \Big)^{1/p} \leq \Big(\int_\Omega |f|^p \, \mathrm{d}\mu \Big)^{1/p} + \Big(\int_\Omega |g|^p \, \mathrm{d}\mu \Big)^{1/p}.$$

Beide Seiten können den Wert $+\infty$ annehmen.

Beweis. Wie in Analysis 2, folgt aus Hölder und der Young Ungleichung. Wir zeigen die stärkere Ungleichung

$$\left(\int_\Omega (|f|+|g|)^p \,\mathrm{d}\mu\right)^{1/p} \leq \left(\int_\Omega |f|^p \,\mathrm{d}\mu\right)^{1/p} + \left(\int_\Omega |g|^p \,\mathrm{d}\mu\right)^{1/p}. \tag{3.7}$$

Tatsächlich impliziert (3.7) Minkowski weil die linke Seite von Minkowski kleiner ist als die linke Seite von (3.7). Das folgt direkt aus der Monotonie des Integrals und weil $|f+g| \leq |f|+|g|$ gilt.

(a) Für $p=1$ gilt wegen der Linearität des Integrals (3.7) natürlich mit Gleichheit.

(b) Sei nun $p>1$. Ist die rechte Seite $+\infty$, so gilt (3.7). Also nehmen wir an, dass beide Integrale der rechten Seite endlich sind. Dann ist aber auch die linke Seite wegen der elementaren Abschätzung

$$(|f|+|g|)^p \leq (2|f| \vee |g|)^p = 2^p(|f^p| \vee |g^p|) \leq 2^p(|f|^p + |g|^p)$$

und der Monotonie des Integrals endlich. Damit nun zum Beweis von (3.7):

$$\begin{aligned}
\int_\Omega (|f|+|g|)^p \,\mathrm{d}\mu &= \int_\Omega (|f|+|g|)^{p-1}(|f|+|g|) \,\mathrm{d}\mu \\
&\overset{\text{ausm., Lin.}}{=} \int_\Omega (|f|+|g|)^{p-1}|f| \,\mathrm{d}\mu + \int_\Omega (|f|+|g|)^{p-1}|g| \,\mathrm{d}\mu \\
&\overset{2\times \text{ Hölder}}{\leq} \left(\int_\Omega |f|^p \,\mathrm{d}\mu\right)^{\frac{1}{p}} \left(\int_\Omega (|f|+|g|)^{(p-1)q} \,\mathrm{d}\mu\right)^{\frac{1}{q}} \\
&\quad + \left(\int_\Omega |g|^p \,\mathrm{d}\mu\right)^{\frac{1}{p}} \left(\int_\Omega (|f|+|g|)^{(p-1)q} \,\mathrm{d}\mu\right)^{\frac{1}{q}} \\
&\overset{\text{auskl.}}{=} \left(\left(\int_\Omega |f|^p \,\mathrm{d}\mu\right)^{\frac{1}{p}} + \left(\int_\Omega |g|^p \,\mathrm{d}\mu\right)^{\frac{1}{p}}\right) \left(\int_\Omega |f|+|g| \,\mathrm{d}\mu\right)^{1-\frac{1}{p}}.
\end{aligned}$$

In der letzten Gleichung haben wir genutzt, dass $1-\frac{1}{p}=\frac{1}{q}$ und $(p-1)q=p$ aufgrund der Voraussetzung an p und q gelten. Rübermultiplizieren des zweiten Faktors gibt dann (3.7).

□

Definition 3.4.4. $f\colon \Omega \to \overline{\mathbb{R}}$ messbar heißt ***p*-fach integrierbar** (oder p-fach μ-integrierbar), falls $\int_\Omega |f|^p \,\mathrm{d}\mu < \infty$. Statt 2-fach integrierbar sagt man auch **quadratintegrierbar** (oder μ-quadratintegrierbar), statt 1-fach integrierbar sagt man **integrierbar** (oder μ-integrierbar).

Das μ lässt man bei den Begrifflichkeiten oft aus Faulheit weg, die Abhängigkeit von μ ist aber essentiell wichtig. Da wir den Begriff der μ-Integrierbarkeit in Definition 3.1.9 schon definiert haben, wäre es besser, wenn die Definition 3.4.4 mit der alten Definition übereinstimmt. Das ist natürlich der Fall:

$$f\ \mu\text{-int.} \underset{\text{Def.}}{\overset{\text{alte}}{\Leftrightarrow}} \int_\Omega f^+ \,\mathrm{d}\mu < \infty, \int_\Omega f^- \,\mathrm{d}\mu < \infty \underset{\text{siehe (3.3)}}{\overset{\text{Übung}}{\Leftrightarrow}} \int_\Omega |f| \,\mathrm{d}\mu < \infty \underset{\text{Def.}}{\overset{\text{neue}}{\Leftrightarrow}} f\ \mu\text{-int.}$$

Schauen wir uns ein ganz konkretes Beispiel für die neue Definition an.

Beispiel 3.4.5.

- Für $\Omega=[1,\infty)$, $\mathcal{A}=\mathcal{B}([1,\infty))$, $\mu=\lambda_{|[1,\infty)}$ ist $\int_\Omega f \,\mathrm{d}\mu = \int_1^\infty f(x)\,\mathrm{d}x$. Für $f(x)=\frac{1}{x^a}$ ist f p-fach integrierbar genau dann, wenn $p>\frac{1}{a}$.

- Für $\Omega=[0,1]$, $\mathcal{A}=\mathcal{B}([0,1])$, $\mu=\lambda_{[0,1]}$ ist $\int_\Omega f \,\mathrm{d}\mu = \int_0^1 f(x)\,\mathrm{d}x$. Für

$$f(x)=\begin{cases} \frac{1}{x^a} & : x\in(0,1] \\ +\infty & : x=0 \end{cases}$$

ist f p-fach integrierbar genau dann, wenn $p < \frac{1}{a}$. Für das Integral ist der Funktionswert $+\infty$ unproblematisch, da die Menge $\{0\}$ eine μ-Nullmenge ist.

In der Mathematik wollen wir aus allen Objekten möglichst nützliche Strukturen schaffen, in diesem Fall einen Vektorraum.

Definition 3.4.6.

$$\mathcal{L}^p(\mu) := \left\{ f\colon \Omega \to \overline{\mathbb{R}} \text{ messbar} \,\middle|\, \int_\Omega |f|^p \,\mathrm{d}\mu < \infty \right\}.$$

Manachmal schreibt man auch $\mathcal{L}^p(\Omega, \mathcal{A}, \mu)$ oder nur Ł^p.

Lemma 3.4.7. Mit punktweiser Addition und Skalarmultiplikation ist $\mathcal{L}^p(\mu)$ ein reeller Vektorraum, sogar ein Untervektorraum der messbaren Funktionen, d. h.

(i) $0 \in \mathcal{L}^p(\mu)$, wobei 0 die konstante Nullfunktion ist.

(ii) $f, g \in \mathcal{L}^p(\mu) \Rightarrow f + g \in \mathcal{L}^p(\mu)$,

(iii) $\alpha \in \mathbb{R}$, $f \in \mathcal{L}^p(\mu) \Rightarrow \alpha f \in \mathcal{L}^p(\mu)$,

Beweis. Messbare Funktionen mit punktweiser Addition und skalarer Multiplikation geben einen Vektorraum. Die Eigenschaften (ii)-(i) bedeuten, dass $\mathcal{L}^p(\mu)$ ein Untervektorraum ist. Wir prüfen also nur die dafür benötigten Eigenschaften:

(i) $\int_\Omega |0|^p \,\mathrm{d}\mu = 0 < \infty$

(ii) $\int_\Omega |\alpha f|^p \,\mathrm{d}\mu \overset{\text{Lin.}}{=} |\alpha|^p \int_\Omega |f|^p \,\mathrm{d}\mu < \infty$ weil $f \in \mathcal{L}^p(\mu)$ angenommen wurde. Also gilt auch $\alpha f \in \mathcal{L}^p(\mu)$.

(iii) Wegen Minkowski gilt

$$\int_\Omega |f+g|^p \,\mathrm{d}\mu \le \left(\underbrace{\Big(\int_\Omega |f|^p \,\mathrm{d}\mu \Big)^{\frac{1}{p}}}_{<\infty} + \underbrace{\Big(\int_\Omega |g|^p \,\mathrm{d}\mu \Big)^{\frac{1}{p}}}_{<\infty} \right)^p < \infty.$$

Also ist $f + g \in \mathcal{L}^p(\mu)$.

□

Aus der Analysis wisst ihr schon, das man aus einem Vektorraum immer gerne einen normierten Raum machen möchte. Dann kann man über Folgenkonvergenz und Stetigkeit sprechen. Können wir also aus $\mathcal{L}^p(\mu)$ einen normierten Raum machen? Nein!

Lemma 3.4.8.

$$||f||_p := \Big(\int_\Omega |f|^p \,\mathrm{d}\mu \Big)^{\frac{1}{p}}$$

ist eine **Halbnorm** auf $\mathcal{L}^p(\mu)$, d. h. es gelten für $f, g \in \mathcal{L}^p(\mu)$ und $\alpha \in \mathbb{R}$

(i) $0 \le ||f||_p < \infty, ||0||_p = 0$ (Definitheit fehlt)

(ii) $||\alpha f||_p = |\alpha| ||f||_p$

(iii) $||f + g||_p \le ||f||_p + ||g||_p$.

Beweis. Die ersten zwei Eigenschaften sind klar. Die Dreiecksungleichung in $\mathcal{L}^p(\mu)$ ist gerade die Minkowski Ungleichung! □

Warnung: $||\cdot||_p$ ist *keine* Norm! Jedes f mit $\mu(\{f \neq 0\}) = 0$ erfüllt

$$||f||_p = \Big(\int_\Omega |f|^p \,\mathrm{d}\mu\Big)^{\frac{1}{p}} = 0.$$

Wenn f auf einer μ-Nullmenge ungleich 0 ist, so ist aber $f \neq 0$. Also ist die Definitheit nicht erfüllt. Mit einem Trick kann man $\mathcal{L}^p(\mu)$ zu einem normierten Vektorraum ändern, indem man das Problem der Definitheit wegdefiniert. Dazu betrachtet man den Quotientenraum bezüglich der Äquivalenzrelation

$$f \sim g \quad :\Leftrightarrow \quad f = g \;\; \mu\text{-fast überall},$$

der aus den Äquivalenzklassen

$$[f] := \big\{g \in \mathcal{L}^p(\mu) \,:\, f \sim g\big\} = \big\{g \in \mathcal{L}^p(\mu) \,:\, f = g \;\mu\text{-fast überall}\big\}$$

besteht. Die Operationen und die Norm werden durch beliebige Repräsentanten der Äquivalenzklassen definiert:

$$[f] + [g] := [f+g], \quad \alpha[f] := [\alpha f] \quad \text{ und } \quad ||[f]||_p := ||f||_p.$$

Der Quotientenraum wird als $L^p(\mu) = \{[f] : f \in \mathcal{L}^p(\mu)\}$ bezeichnet. Gemeinsam mit den Operationen und der Norm auf den Elementen (Äquivalenzklassen) ist $L^p(\mu)$ ein normierter Vektorraum:

Satz 3.4.9. Für $p \geq 1$ ist $L^p(\mu)$ mit den gerade definierten Operationen ein normierter Vektorraum und mit $||\cdot||_p$ ein vollständiger normierter Raum (Banachraum).

Beweis. Wir zeigen, dass $L^p(\mu)$ mit den definierten Operationen ein normierter Vektorraum ist.

(a) Vektorraum aus Lineare Algebra 1.

(b) Die Eigenschaften der Halbnorm folgen direkt aus der Definition weil $||\cdot||_p$ eine Halbnorm auf $\mathcal{L}^p(\mu)$ ist. Es fehlt also nur noch die Definitheit. Sei dazu $f \in L^p(\mu)$, so gilt:

$$\begin{aligned}
||[f]||_p = 0 \quad &\overset{\text{Def.}}{\Leftrightarrow} \quad ||f||_p = 0 \\
&\overset{\text{Def.}}{\Leftrightarrow} \quad \int_\Omega |f|^p \,\mathrm{d}\mu = 0 \\
&\overset{3.1.15}{\Leftrightarrow} \quad |f|^p = 0 \;\mu\text{-fast überall} \\
&\Leftrightarrow \quad f = 0 \;\mu\text{-fast überall} \\
&\overset{\text{Def. }[0]}{\Leftrightarrow} \quad [f] = [0]
\end{aligned}$$

Damit ist $||\cdot||_p$ eine Norm.

Was wir hier nicht zeigen werden, ist die Vollständigkeit von $\mathcal{L}^p(\mu)$. Aber irgendwas sollten wir ja für die Funktionalanalysis übrig lassen! □

Vorlesung 15

3.5 Produktmaße und Satz von Fubini

Jetzt wird es noch einmal so richtig dreckig, bevor wir die wunderbare Welt der Stochastik erobern können. Sobald wir uns durch die Konstruktion des Produktmaßes und des Satzes von Fubini gequält haben, fällt uns alles andere aber sofort vor die Füße.

Im Folgenden seien $(\Omega_1, \mathcal{A}_1, \mu_1)$ und $(\Omega_2, \mathcal{A}_2, \mu_2)$ Maßräume und

$$\Omega = \Omega_1 \times \Omega_2 = \{(\omega_1, \omega_2) \colon \omega_1 \in \Omega_1,\; \omega_2 \in \Omega_2\}$$

das kartesische Produkt aus Analysis 1. Wir wollen auf Ω eine σ-Algebra und darauf ein Maß mit einer schönen Produkteigenschaft definieren (deshalb wird das Maß Produktmaß heißen).

Definition 3.5.1.

(i) Die σ-Algebra
$$\mathcal{A}_1 \otimes \mathcal{A}_2 = \sigma(\{A_1 \times A_2 \colon A_1 \in \mathcal{A}_1,\ A_2 \in \mathcal{A}_2\})$$
heißt **Produkt-σ-Algebra** auf $\Omega_1 \times \Omega_2$.

(ii) Ein Maß auf $\mathcal{A}_1 \otimes \mathcal{A}_2$ heißt **Produktmaß**, falls
$$\mu(A_1 \times A_2) = \mu_1(A_1) \cdot \mu_2(A_2) \tag{3.8}$$
für alle Mengen $A_1 \in \mathcal{A}_1, A_2 \in \mathcal{A}_2$ gilt.

Natürlich wäre es schön, wenn die Produkt-σ-Algebra einfach nur aus allen Mengen $A_1 \times A_2$ bestehen würde. Leider gibt das keine σ-Algebra (die Komplementbildung geht schief), die Mengen geben nur einen $\cap$-stabilen Erzeuger von $\mathcal{A}_1 \otimes \mathcal{A}_2$. Die Eigenschaft des Produktmaßes legt das Produktmaß also nur auf einem $\cap$-stabilen Erzeuger fest. Mit unseren Kenntnissen der Maßtheorie, könnten wir also reflexhaft sagen: Kein Problem, mit Carathéodory können wir ein Produktmaß aus den Werten auf dem Erzeuger konstruieren und wegen Dynkin-Systemen kann es nur ein Maß mit der Produktmaßeigenschaft bekommen. Genau richtig - das funktioniert! Wir quälen uns im nächsten Satz aber gewaltig mehr. Wir konstruieren das Produktmaß nicht mit Carathéodory, sondern schreiben eine Formel hin. Das ist in der Tat viel komplizierter, der Vorteil ist aber, dass wir damit den darauf folgenden Satz von Fubini schon fast bewiesen haben. Würden wir das Produktmaß mit Carathéodory konstruieren, würde der ganze Aufwand in den Beweis von Fubini verschoben.

Satz 3.5.2. **[Konstruktion Produktmaß]** Sind μ_1, μ_2 σ-endliche Maße auf $\mathcal{A}_1, \mathcal{A}_2$, so existiert ein eindeutiges Maß $\mu_1 \otimes \mu_2$ auf $\mathcal{A}_1 \otimes \mathcal{A}_2$ mit
$$\mu_1 \otimes \mu_2(A_1 \times A_2) = \mu_1(A_1) \cdot \mu_2(A_2)$$
für alle Mengen $A_1 \in \mathcal{A}_1, A_2 \in \mathcal{A}_2$.

Beweis. **Eindeutigkeit**: Für die Eindeutigkeit nutzen wir wieder den Eindeutigkeitssatz 1.2.13, der auf Dynkin-Systemen beruht. Sei
$$\mathcal{S} = \{A_1 \times A_2 \colon A_1 \in \mathcal{A}_1,\ A_2 \in \mathcal{A}_2\},$$
dann ist $\mathcal{S}$ ein $\cap$-stabiler Erzeuger von $\mathcal{A}_1 \otimes \mathcal{A}_2$. Um den Eindeutigkeitssatz anzuwenden, müssen wir noch zeigen, dass Produktmaße σ-endlich sein müssen, wir brauchen also eine wachsende Folge in $\mathcal{S}$, die endliches Maß hat und Ω ausfüllt. Weil μ_1, μ_2 σ-endliche Maße sind, existieren Folgen $(E_n^1)_{n\in\mathbb{N}} \subseteq \mathcal{A}_1$, $(E_n^2)_{n\in\mathbb{N}} \subseteq \mathcal{A}_2$ mit $E_n^1 \uparrow \Omega_1$, $E_n^2 \uparrow \Omega_2$ und $\mu_1(E_n^1) < \infty$, $\mu_2(E_n^2) < \infty$ für alle $n \in \mathbb{N}$. Sei nun $E_n := E_n^1 \times E_n^2$, dann gelten
$$E_n \uparrow \Omega,\ n \to \infty, \quad \text{und} \quad \mu_1(E_n^1) \cdot \mu_2(E_n^2) < \infty, \quad \forall n \in \mathbb{N}.$$
Aus dem Eindeutigkeitssatz folgt also, dass zwei Maße μ und $\bar{\mu}$, die die Definition des Produktmaßes erfüllen, gleich sind. Es kann also nur ein Produktmaß auf $\mathcal{A}_1 \otimes \mathcal{A}_2$ geben. Ob es so ein Maß gibt, ist natürlich noch nicht klar.

Existenz: Anstatt das Produktmaß mit Carathéodory zu konstruieren, schreiben wir es einfach hin. Hier ist es:
$$\mu(A) := \int_{\Omega_1} \mu_2(A_{\omega_1})\,\mathrm{d}\mu_1, \quad A \in \mathcal{A}_1 \otimes \mathcal{A}_2,$$
wobei $A_{\omega_1} = \{\omega_2 \in \Omega_2 \colon (\omega_1, \omega_2) \in A\}$. Wir machen sogar noch mehr, wir schreiben noch ein zweites Produktmaß hin:
$$\overline{\mu}(A) := \int_{\Omega_2} \mu_1(A_{\omega_2})\,\mathrm{d}\mu_2, \quad A \in \mathcal{A}_1 \otimes \mathcal{A}_2,$$

wobei jetzt $A_{\omega_2} = \{\omega_1 \in \Omega_1 : (\omega_1, \omega_2) \in A\}$. Wir zeigen nun, dass μ ein Produktmaß auf $\mathcal{A}_1 \otimes \mathcal{A}_2$ ist. Mit exakt demselben Beweis zeigt man auch, dass $\overline{\mu}$ ein Produktmaß ist, weshalb dann wegen der Eindeutigkeit auch $\mu = \overline{\mu}$ gilt. Zeigen wir also die Eigenschaften eines Maßes sowie die definierende Eigenschaft des Produktmaßes:

(i) $\mu \colon \mathcal{A}_1 \otimes \mathcal{A}_2 \to [0, \infty]$ gilt, weil μ_2 ein Maß ist (deshalb nicht-negativ) und Integrale über nicht-negative Funktionen nicht-negativ sind.

(ii) $\mu(\emptyset) = 0$ gilt, weil $\emptyset_{\omega_1}$ auch die leere Menge ist und Maße der leeren Menge 0 sind.

(iii) Nun zur σ-Additivität. Seien dazu $A^1, A^2, ... \in \mathcal{A}_1 \otimes \mathcal{A}_2$ paarweise disjunkt und sei

$$A := \dot{\bigcup_{k=1}^{\infty}} A^k.$$

Dann gilt $A_{\omega_1} = \dot\bigcup_{k=1}^{\infty} A^k_{\omega_1}$ und mit den Maßeigenschaften sowie monotoner Konvergenz

$$\begin{aligned}
\mu(A) &\overset{\text{Def.}}{=} \int_{\Omega_1} \mu_2\Big(\Big(\dot{\bigcup_{k=1}^{\infty}} A^k\Big)_{\omega_1}\Big) \, \mathrm{d}\mu_1(\omega_1) \\
&= \int_{\Omega_1} \mu_2\Big(\dot{\bigcup_{k=1}^{\infty}} A^k_{\omega_1}\Big) \, \mathrm{d}\mu_1(\omega_1) \\
&\overset{\mu_2\ \sigma\text{-add.}}{=} \int_{\Omega_1} \sum_{k=1}^{\infty} \mu_2(A^k_{\omega_1}) \, \mathrm{d}\mu_1(\omega_1) \\
&\overset{3.2.1}{=} \sum_{k=1}^{\infty} \int_{\Omega_1} \mu_2(A^k_{\omega_1}) \, \mathrm{d}\mu_1(\omega_1) \overset{\text{Def.}}{=} \sum_{k=1}^{\infty} \mu(A^k).
\end{aligned}$$

(iv) μ ist also ein Maß auf $\mathcal{A}_1 \otimes \mathcal{A}_2$. Wir müssen noch die definierende Eigenschaft auf dem Erzeuger zeigen. Sei dazu $A = A_1 \times A_2$. Weil aufgrund der Definitionen

$$(A_1 \times A_2)_{\omega_1} = \begin{cases} \emptyset & : \omega_1 \notin A_1 \\ A_2 & : \omega_1 \in A_1 \end{cases}$$

gilt, folgt aufgrund der Definition des Integrals für einfache Integranden

$$\begin{aligned}
\mu(A_1 \times A_2) &= \int_{\Omega_1} \mu_2((A_1 \times A_2)_{\omega_1}) \, \mathrm{d}\mu_1(\omega_1) \\
&= \int_{\Omega_1} \mu_2(A_2) \mathbf{1}_{A_1}(\omega_1) \, \mathrm{d}\mu_1(\omega_1) \\
&\overset{\text{Linear}}{=} \mu_2(A_2) \int_{\Omega_1} \mathbf{1}_{A_1}(\omega_1) \, \mathrm{d}\mu(\omega_1) = \mu_2(A_2) \cdot \mu_1(A_1).
\end{aligned}$$

Also ist μ ein Produktmaß.

Eigentlich könnte alles so schön sein, und der Beweis ist hier zu Ende. Leider haben wir geschummelt. Warum ist μ überhaupt sinnvoll definiert? Klingt blöd, ist aber gar nicht so klar. Warum ist der Integrand überhaupt definiert, d. h. warum gilt $A_{\omega_1} \in \mathcal{A}_2$? Unklar. Warum ist $\omega_1 \mapsto \mu_2(A_{\omega_1})$ überhaupt $(\mathcal{A}_1, \mathcal{B}(\overline{\mathbb{R}}))$-messbar, warum macht das Integral also überhaupt Sinn? Unklar. Wir sollten also beides noch checken.

Wir zeigen jetzt nacheinander

(a) $A_{\omega_1} \in \mathcal{A}_2$ für alle $A \in \mathcal{A}_1 \otimes \mathcal{A}_2$

(b) $\omega_1 \mapsto \mu_2(A_{\omega_1})$ ist $(\mathcal{A}_1, \mathcal{B}(\overline{\mathbb{R}}))$-messbar.

Zu (a): Wir folgen dem Trick der guten Mengen. Seien dazu die guten Mengen

$$\mathcal{F} = \{A \in \mathcal{A}_1 \otimes \mathcal{A}_2 \colon A_{\omega_1} \in \mathcal{A}_2\}.$$

Wir zeigen: $\mathcal{F} = \mathcal{A}_1 \otimes \mathcal{A}_2$. Dazu zeigen wir zunächst, dass $\mathcal{F}$ eine σ-Algebra ist:

- $\Omega \in \mathcal{F}$ gilt, weil $\Omega_{\omega_1} = \Omega_2 \in \mathcal{F}$ gilt.
- Sei $A \in \mathcal{F}$, dann ist $A^C \in \mathcal{F}$ weil $(A^C)_{\omega_1} = (A_{\omega_1})^C \in \mathcal{A}_2$, da $\mathcal{A}_2$ als σ-Algebra abgeschlossen unter Komplementbildung ist.
- Genauso mit abzählbaren Vereinigungen: Wegen der Abgeschlossenheit der σ-Algebra $\mathcal{A}_2$ unter Bildung von abzählbaren Vereinigungen, gilt für $A^1, A^2, ... \in \mathcal{F}$

$$\Big(\bigcup_{k=1}^{\infty} A^k\Big)_{\omega_1} = \bigcup_{k=1}^{\infty} \underbrace{A^k_{\omega_1}}_{\in \mathcal{A}_2} \in \mathcal{A}_2.$$

  Damit ist $\bigcup_{k=1}^{\infty} A^k \in \mathcal{F}$.

Folglich ist $\mathcal{F}$ eine σ-Algebra. Weil $\mathcal{S} \subseteq \mathcal{F}$ und $\sigma(\mathcal{S}) \overset{\text{Def.}}{=} \mathcal{A}_1 \otimes \mathcal{A}_2$ gilt, bekommen wir zusammen:

$$\mathcal{A}_1 \otimes \mathcal{A}_2 = \sigma(\mathcal{S}) \subseteq \sigma(\mathcal{F}) = \mathcal{F} \subseteq \mathcal{A}_1 \otimes \mathcal{A}_2$$

Weil links und rechts das gleiche steht, bekommen wir überall Gleichheiten, es gilt also $\mathcal{F} = \mathcal{A}_1 \otimes \mathcal{A}_2$ und die Behauptung folgt. Das war wieder unser „standard“ gute Mengen Trick in der einfacheren Situation mit σ-Algebren.

Nun zu (b): Wir checken erst den Fall $\mu_2(\Omega_2) < \infty$ und schieben die Aussage dann mit der σ-Endlichkeit auf den allgemeinen Fall. Sei also erstmal $\mu_2(\Omega_2) < \infty$. Wir zeigen, wieder mit dem gute Mengen Trick (aber in der Dynkin-System Variante), $\mathcal{A}_1 \otimes \mathcal{A}_2 = \mathcal{F}$, mit der Menge

$$\mathcal{F} := \{A \in \mathcal{A}_1 \otimes \mathcal{A}_2 \colon \omega_1 \mapsto \mu_2(A_{\omega_1}) \text{ messbar}\}$$

der guten Mengen. Es gilt $\mathcal{S} \subseteq \mathcal{F}$, da

$$\omega_1 \mapsto \mu_2((A_1 \times A_2)_{\omega_1}) = \underbrace{\mu_2(A_2)\mathbf{1}_{A_1}(\omega_1)}_{\text{messbar}}$$

als Produkt messbarer Abbildungen messbar ist. Um wie in (a) zu argumentieren, zeigen wir nun, dass $\mathcal{F}$ ein Dynkin-System ist:

- $\Omega \in \mathcal{F}$ ist klar, weil $\omega_1 \mapsto \mu_2(\Omega_{\omega_1}) = \mu_2(\Omega_2) < \infty$ konstant und damit messbar ist.
- Sei $A \in \mathcal{F}$, dann gilt

$$\omega_1 \mapsto \mu_2((A^C)_{\omega_1}) = \mu_2(A^C_{\omega_1}) \overset{\text{endl. Maß}}{=} \underbrace{\underbrace{\mu_2(\Omega_2)}_{\text{messbar}} - \underbrace{\mu_2(A_{\omega_1})}_{\text{messbar}}}_{\text{messbar}}.$$

  Also gilt $A^C \in \mathcal{F}$ und damit ist $\mathcal{F}$ abgeschlossen bezüglich Komplementbildung.

- Seien nun $A^1, A^2, ... \in \mathcal{F}$ paarweise disjunkt, dann gilt

$$\begin{aligned}\omega_1 \mapsto \mu_2\Big(\Big(\dot{\bigcup_{k=1}^{\infty}} A^k\Big)_{\omega_1}\Big) &= \mu_2\Big(\dot{\bigcup_{k=1}^{\infty}} A^k_{\omega_1}\Big)\\ &\overset{\sigma\text{-add.}}{=} \sum_{k=1}^{\infty} \underbrace{\mu(A^k_{\omega_1})}_{\text{messbar}} = \underbrace{\lim_{m\to\infty} \underbrace{\sum_{k=1}^{m} \mu(A^k_{\omega_1})}_{\text{messbar}}}_{\text{messbar}},\end{aligned}$$

  weil Grenzwerte messbarer Abbildungen wieder messbar sind.

$\mathcal{F}$ ist also ein Dynkin-System. Nun folgt wie immer beim Trick der guten Mengen

$$\mathcal{A}_1 \otimes \mathcal{A}_2 \overset{\text{Def.}}{=} \sigma(\mathcal{S}) = d(\mathcal{S}) \subseteq d(\mathcal{F}) = \mathcal{F} \subseteq \mathcal{A}_1 \otimes \mathcal{A}_2,$$

wobei wir den Hauptsatz für Dynkin-Systeme (Satz 1.2.11) für $\mathcal{S}$ genutzt haben. Also gilt wieder $\mathcal{F} = \mathcal{A}_1 \otimes \mathcal{A}_2$ und die Behauptung folgt.
Jetzt fehlt nur noch der Fall $\mu_2(\Omega_2) = \infty$. Sei dazu $(E_n^2) \subseteq \mathcal{A}_2$ mit $E_n^2 \uparrow \Omega$ und $\mu_2(E_n^2) < \infty$ für alle $n \in \mathbb{N}$. Wir definieren

$$\mu_2^n(A) = \mu_2(A \cap E_n^2), \quad n \in \mathbb{N},$$

wie wir schon mehrfach gemacht haben (siehe z. B. den Beweis von 1.2.13). Dann sind die μ_2^n endliche Maße auf $(\Omega_2, \mathcal{A}_2)$. Aus dem ersten Schritt folgt, dass $\omega_1 \mapsto \mu_2^n(A_{\omega_1})$ messbar ist. Weil wegen der Stetigkeit von Maßen

$$\lim_{n\to\infty} \mu_2^n(A_{\omega_1}) = \lim_{n\to\infty} \mu_2(A_{\omega_1} \cap E_n^2) = \mu_2(A_{\omega_1})$$

gilt, ist auch die Abbildung $\omega_1 \mapsto \mu_2(A_{\omega_1})$ als punktweiser Grenzwert von messbaren Funktionen messbar. Das war's! □

Natürlich kann man per Induktion von $n = 2$ auf $n \in \mathbb{N}$ schließen:

Korollar 3.5.3. Sind $(\Omega_i, \mathcal{A}_i, \mu_i)_{i=1,\dots,n}$ σ-endliche Maßräume, so existiert genau ein Maß $\mu_1 \otimes \dots \otimes \mu_n$ auf der n-fachen Produkt-σ-Algebra

$$\mathcal{A}_1 \otimes \dots \otimes \mathcal{A}_n := \sigma(\{A_1 \times \dots \times A_n : A_i \in \mathcal{A}_i\})$$

auf $\Omega_1 \times \dots \times \Omega_n$ mit

$$\mu_1 \otimes \dots \otimes \mu_n(A_1 \times \dots \times A_n) = \mu_1(A_1) \cdot \dots \cdot \mu_n(A_n)$$

für alle Mengen $A_1 \in \mathcal{A}_1, \dots, A_n \in \mathcal{A}_n$.

Beweis. Induktion. □

Definition 3.5.4. Sind alle $(\Omega_i, \mathcal{A}_i, \mu_i)$ identisch, so schreibt man $\mu^{\otimes n}$ statt $\mu \otimes \dots \otimes \mu$. $\mu^{\otimes n}$ heißt dann **n-faches Produktmaß** von μ.

Vorlesung 16

Satz 3.5.5. **[Satz von Fubini für $f \geq 0$]** Seien $(\Omega_1, \mathcal{A}_1, \mu_1)$, $(\Omega_2, \mathcal{A}_2, \mu_2)$ σ-endliche Maßräume und $f\colon \Omega_1 \times \Omega_2 \to [0, +\infty]$ sei $(\mathcal{A}_1 \otimes \mathcal{A}_2, \mathcal{B}(\overline{\mathbb{R}}))$-messbar. Dann gelten:

(i) Die Integranden nach einer Variablen sind messbar, d. h.

$$\omega_2 \mapsto f_{\omega_1}(\omega_2) := f(\omega_1, \omega_2) \text{ ist } (\mathcal{A}_2, \mathcal{B}(\overline{\mathbb{R}}))\text{-messbar für alle } \omega_1 \in \Omega_1 \text{ fest},$$
$$\omega_1 \mapsto f_{\omega_2}(\omega_1) := f(\omega_1, \omega_2) \text{ ist } (\mathcal{A}_1, \mathcal{B}(\overline{\mathbb{R}}))\text{-messbar für alle } \omega_2 \in \Omega_2 \text{ fest}.$$

(ii) Die Integralfunktionen nach einer Variablen sind messbar, d. h.

$$\omega_2 \mapsto \int_{\Omega_1} f_{\omega_2}(\omega_1)\, \mathrm{d}\mu_1(\omega_1) \text{ ist } (\mathcal{A}_2, \mathcal{B}(\overline{\mathbb{R}}))\text{-messbar},$$
$$\omega_1 \mapsto \int_{\Omega_2} f_{\omega_1}(\omega_2)\, \mathrm{d}\mu_2(\omega_2) \text{ ist } (\mathcal{A}_1, \mathcal{B}(\overline{\mathbb{R}}))\text{-messbar}.$$

(iii) Es gilt Fubini und der Fubini-Flip:

$$\begin{aligned}\int_{\Omega_1 \times \Omega_2} f \,\mathrm{d}\mu_1 \otimes \mu_2 &\overset{\text{Fubini}}{=} \int_{\Omega_1} \Big(\int_{\Omega_2} f_{\omega_1}(\omega_2)\, \mathrm{d}\mu_2(\omega_2) \Big)\, \mathrm{d}\mu_1(\omega_1) \\ &\overset{\text{Fubini-Flip}}{=} \int_{\Omega_2} \Big(\int_{\Omega_1} f_{\omega_2}(\omega_1)\, \mathrm{d}\mu_1(\omega_1) \Big)\, \mathrm{d}\mu_2(\omega_2)\end{aligned}$$

Ganz wichtig: (i) und (ii) besagen lediglich, dass alle Integrale in der wesentlichen Aussage (iii) Sinn machen. In (iii) kann auch auf allen Seiten der Gleichheiten $+\infty$ stehen.

Beweis. Weil f messbar ist, existiert eine Folge $(f^n)_{n\in\mathbb{N}}$ einfacher Funktionen, die punktweise (also für alle (ω_1,ω_2)), gegen f wachsen. Dann gilt auch punktweise $f^n_{\omega_1} \uparrow f_{\omega_1}$, $f^n_{\omega_2} \uparrow f_{\omega_2}$, $n\to\infty$, weil dort einfach eine Variable festgehalten wird. Da $(f^n_{\omega_1})_{n\in\mathbb{N}}$, $(f^n_{\omega_2})_{n\in\mathbb{N}}$ Folgen einfacher Funktionen (in jeweils einer Koordinate) sind, sind f_{ω_1}, f_{ω_2} als punkteweise Grenzwerte messbarer Funktionen auch messbar. Also gilt (i). Weil mit monotoner Konvergenz für alle $\omega_1\in\Omega_1$

$$\lim_{n\to\infty}\int_{\Omega_2} f^n_{\omega_1}(\omega_2)\,\mathrm{d}\mu_2(\omega_2) \overset{3.2.1}{=} \int_{\Omega_2} f_{\omega_1}(\omega_2)\,\mathrm{d}\mu_2(\omega_2)$$

gilt, ist

$$\omega_1 \mapsto \int_{\Omega_2} f_{\omega_1}(\omega_2)\,\mathrm{d}\mu_2(\omega_2)$$

als punktweiser Grenzwert messbarer Funktionen auch messbar. Man beachte dabei, dass

$$\omega_1 \mapsto \int_{\Omega_2} f^n_{\omega_1}(\omega_2)\,\mathrm{d}\mu_2(\omega_2)$$

eine endliche Linearkombination von Funktionen der Form $\omega_1 \mapsto \mu_2(A_{\omega_1})$ ist. Diese Abbildungen sind, wie im Beweis von 3.5.3 gezeigt, messbar. Damit gilt also auch (ii).

Nun zur eigentlichen Aussage, (iii). Nicht überraschend, erst für Indikatoren, dann die Gebetsmühle. Sei also zunächst $f=\mathbf{1}_A$ für ein $A\in\mathcal{A}_1\otimes\mathcal{A}_2$. Dann gilt aufgrund der expliziten Konstruktion des Produktmaßes

$$\begin{aligned}\int_{\Omega_1\times\Omega_2} f\,\mathrm{d}\mu_1\otimes\mu_2 &\overset{\text{Def. Int.}}{=} \mu_1\otimes\mu_2(A)\\ &\overset{\text{Produktmaß}}{=} \int_{\Omega_1}\mu_2(A_{\omega_1})\,\mathrm{d}\mu_1(\omega_1)\\ &= \int_{\Omega_1}\Big(\int_{\Omega_2} f_{\omega_1}(\omega_2)\,\mathrm{d}\mu_2(\omega_2)\Big)\,\mathrm{d}\mu_1(\omega_1),\end{aligned}$$

weil $f_{\omega_1}=\mathbf{1}_{A_{\omega_1}}$ gilt. Genau analog ergibt sich

$$\int_{\Omega_1\times\Omega_2} f\,\mathrm{d}\mu_1\otimes\mu_2 = \int_{\Omega_2}\Big(\int_{\Omega_1} f(\omega_1,\omega_2)\,\mathrm{d}\mu_1(\omega_1)\Big)\,\mathrm{d}\mu_2(\omega_2),$$

weil wir am Anfang der Konstruktion des Produktmaßes auch schon die Identität gesehen haben:

$$\mu_1\otimes\mu_2(A) = \int_{\Omega_2}\mu_1(A_{\omega_2})\,\mathrm{d}\mu_2(\omega_2).$$

Das hatten wir in dem Beweis $\bar{\mu}$ genannt. Damit haben wir beide Gleichheiten in (iii) für Indikatorfunktionen $f=\mathbf{1}_A$ bewiesen. Nun noch durch die Gebetsmühle der Integrationstheorie: Für einfache Funktionen folgt (iii) durch Linearität des Integrals und monotone Konvergenz folgt die Aussage auch für alle messbaren $f\geq 0$. $\square$

Warnung 3.5.6. Meistens wird nur der „Fubini-flip" genutzt, also die zweite Gleichheit. Die gemeinsame $(\mathcal{A}_1\otimes\mathcal{A}_2,\mathcal{B}(\bar{\mathbb{R}}))$-Messbarkeit in **beiden** Variablen muss für f trotzdem gecheckt werden, auch wenn es in vielen Büchern und Skripten ignoriert wird. Messbarkeit in jeweils einer Koordinate reicht nicht aus! Das ist wie in Analysis 2, dort war partielle Stetigkeit und Stetigkeit einer Abbildung $f:\mathbb{R}^2\to\mathbb{R}$ nicht äquivalent. Wer mal ausprobieren will was schief gehen kann, sollte sich folgendes Beispiel $f:\mathbb{R}\times\mathbb{R}\to\mathbb{R}$ anschauen: Sei $E\notin\mathcal{B}(\mathbb{R})$ und

$$f(x,y)=\begin{cases}1 & : x=y, x\in E\\ 0 & : \text{sonst}\end{cases}.$$

Dann ist f zwar partiell messbar (also messbar in einer Variablen für feste andere Variable), jedoch nicht messbar als Funktion in zwei Variablen.

Merkt euch trotzdem folgende ganz grobe Behauptung: „Fubini funktioniert immer“. Mathematisch ist das natürlich nicht korrekt, man muss schließlich Voraussetzungen prüfen. Im Gegensatz zu monotoner oder dominierter Konvergenz ist das aber eigentlich nie ein Problem. Integrale vertauschen ist selten ein Problem, Grenzwerte und Integrale zu vertauschen ist dagegen oft schwierig!

Beispiel 3.5.7.

(i) Der Satz von Fubini aus Analysis 2 ist gerade der Spezialfall für $\mu_1 = \mu_2 = \lambda$ und liest sich als

$$\int_{\mathbb{R}^2} f(x_1,x_2)\,\mathrm{d}(x_1,x_2) = \int_{\mathbb{R}} \Big(\int_{\mathbb{R}} f(x_1,x_2)\,\mathrm{d}x_1 \Big)\,\mathrm{d}x_2 = \int_{\mathbb{R}} \Big(\int_{\mathbb{R}} f(x_1,x_2)\,\mathrm{d}x_2 \Big)\,\mathrm{d}x_1$$

weil das zweidimensionale Lebesgue-Maß gerade das Produktmaß $\lambda \otimes \lambda$ ist. Wie immer schreiben wir lieber $\mathrm{d}x$ (bzw. $\mathrm{d}(x_1,x_2)$) statt $\mathrm{d}\lambda$, das ist aber nur eine Notationsfrage!

(ii) Als Übungsaufgabe zeigt ihr, dass wir bei Doppelreihen die Reihenfolge immer ändern dürfen, wenn alle Koeffizienten nicht-negativ sind:

$$\sum_{k=1}^{\infty} \sum_{n=1}^{\infty} a_{k,n} = \sum_{n=1}^{\infty} \sum_{k=1}^{\infty} a_{k,n}.$$

Das ist tatsächlich nur Fubini weil Reihen gerade Integrale mit Zählmaßen sind, vergleiche Beispiel 3.1.12. Vermutlich kennt ihr das Drehen von Reihen schon aus der Analysis 1, jetzt nochmal mit Fubini.

In den meisten Anwendungen reicht Fubini für nicht-negative messbare Funktionen, darum haben wir folgende Verallgemeinerung in der Vorlesung weggelassen. Zur Vollständigkeit wollen wir aber wieder durch Zerlegung in Positiv- und Negativteil eine Variante von Fubini für reell-wertige Integranden formulieren:

Satz 3.5.8. (Fubini für allgemeines integrierbares f)
Unter den selben Voraussetzungen von Satz 3.5.5 sei nun $f : \Omega_1 \times \Omega_2 \to \overline{\mathbb{R}}$ integrierbar. Dann gelten die Aussagen (i), (ii) und (iii).

Beweis. Wir schreiben $f = f^+ - f^-$, und wenden auf $f^+ \geq 0$ und $f^- \geq 0$ Satz 3.5.5 an. Die Messbarkeitsaussagen folgen direkt aus der Messbarkeit von Differenzen messbarer Abbildungen, Aussage (iii) folgt durch Zerlegung der Integrale. Die erste Gleichheit von (iii) zeigen wir wie folgt:

$$\begin{aligned}
\int_{\Omega_1\times\Omega_2} f(\omega_1,\omega_2)\,\mathrm{d}\mu_1\otimes\mu_2(\omega_1,\omega_2) &\overset{\text{Def.}}{=} \int_{\Omega_1\times\Omega_2} f^+(\omega_1,\omega_2)\,\mathrm{d}\mu_1\otimes\mu_2(\omega_1,\omega_2) \\
&\quad - \int_{\Omega_1\times\Omega_2} f^-(\omega_1,\omega_2)\,\mathrm{d}\mu_1\otimes\mu_2(\omega_1,\omega_2) \\
&\overset{\text{2x Fubini}}{=} \int_{\Omega_1} \Big(\int_{\Omega_2} f^+(\omega_1,\omega_2)\,\mathrm{d}\mu_2(\omega_2) \Big)\,\mathrm{d}\mu_1(\omega_1) \\
&\quad - \int_{\Omega_1} \Big(\int_{\Omega_2} f^-(\omega_1,\omega_2)\,\mathrm{d}\mu_2(\omega_2) \Big)\,\mathrm{d}\mu_1(\omega_1) \\
&\overset{\text{2x Lin.}}{=} \int_{\Omega_1} \Big(\int_{\Omega_2} (f^+ - f^-)(\omega_1,\omega_2)\,\mathrm{d}\mu_2(\omega_2) \Big)\,\mathrm{d}\mu_1(\omega_1) \\
&= \int_{\Omega_1} \Big(\int_{\Omega_2} f(\omega_1,\omega_2)\,\mathrm{d}\mu_2(\omega_2) \Big)\,\mathrm{d}\mu_1(\omega_1).
\end{aligned}$$

Die zweite Gleichheit von (iii) folgt genauso aus Satz 3.5.5. □

Kapitel 4

Stochastik

An dieser Stelle beenden wir die allgemeine Maß- und Integrationstheorie und wenden uns endlich vollständig der Stochastik zu. Wir haben bereits zufällige Experimente durch Wahrscheinlichkeitsräume modelliert, die Verteilung „einer Masse Zufall“ auf die reellen Zahlen durch Verteilungsfunktionen diskutiert und die Konzentration des Zufalls unter $\mathbb{P}_F$ abgeschätzt. Jetzt kommt noch ein letzter Modellierungsschritt, wir wenden uns sogenannten Zufallsvariablen zu.

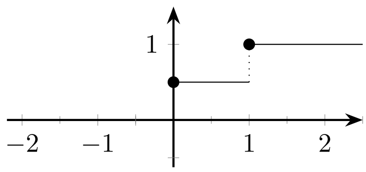
4.1 Zufallsvariablen

Diskussion 4.1.1. Schauen wir uns nochmal die Modellierung sehr einfacher zufälliger Experimente an. Beispielsweise für den Münzwurf haben wir zwei Modellierungsvarianten diskutiert:

Variante 1: $\Omega = \{\text{Kopf}, \text{ Zahl}\}$, $\mathcal{A} = \mathcal{P}(\Omega)$ und das Maß

$$\mathbb{P}(\{\text{Kopf}\}) = \mathbb{P}(\{\text{Zahl}\}) = \frac{1}{2}, \quad \mathbb{P}(\Omega) = 1, \quad \mathbb{P}(\emptyset) = 0.$$

Variante 2: $\Omega = \mathbb{R}$, $\mathcal{A} = \mathcal{B}(\mathbb{R})$, $\mathbb{P} = \mathbb{P}_F$, wobei $\mathbb{P}_F$ das Wahrscheinlichkeitsmaß mit diskreter Verteilungsfunktion F ist:



Beide Modelle modellieren ein Experiment, bei dem zwei Elementarereignisse jeweils mit Wahrscheinlichkeit $\frac{1}{2}$ auftreten. Das zweite Modell ist natürlich viel zu kompliziert, hat aber den Vorteil, dass wir reelle Zahlen bekommen. Der Vorteil ist, dass wir so zum Beispiel so etwas wie einen Mittelwert (heißt später Erwartungswert) definieren können. Da wir den Ereignissen die Werte 0 und 1 geben, hättet ihr ohne viel nachzudenken sicherlich als umgangssprachlichen Erwartungswert $\frac{1}{2}$ vorgeschlagen.
Was unterscheidet Variante 1 von Variante 2? In der ersten Variante haben wir nur das zufällige Experiment Würfeln modelliert in der zweiten Variante haben wir zwei Dinge auf einmal modelliert:

- Was passiert? (Welches zufällige Ereigniss passiert beim Würfeln)
- Was wird ausgezahlt? (Was wird für Kopf bzw. Zahl ausgezahlt)

Wenn wir bei Variante 1 auch über Auszahlungen sprechen wollen, müssen wir die möglichen Elementarereignisse noch in Auszahlungen übersetzen. Dafür kommen messbare Abbildungen $X : \Omega \to \mathbb{R}$ ins Spiel.

Ab jetzt: Trenne in der Modellierung „Was passiert?“ (also Ereignisse und Wahrscheinlichkeiten) von „Was wird beobachtet/ausgezahlt?“.

Definition 4.1.2. Ein **stochastisches Modell** besteht aus

(i) einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$,

(ii) einer $(\mathcal{A}, \mathcal{B}(\mathbb{R}))$-messbaren Abbildung $X\colon \Omega \to \mathbb{R}$.

Dabei beschreibt (i) das zufällige Experiment, (ii) beschreibt die „Beobachtung/Auszahlung“. X wird auch **Zufallsvariable (ZV)** genannt. Eine konkrete Ausführung $X(\omega)$ nennt man auch **Realisierung** der Zufallsvariablen.

Die Übersetztung des Elementarereignisses ω in die Beobachtung $X(\omega)$ ist dabei deterministisch (nicht zufällig), der Zufall steckt ausschließlich in dem Auftreten von ω. Irgendjemand unbekanntes (die Physik, ein Computer, ein Gott, etc.) entscheidet über die Wahl des zufälligen ω, das wird dann in die Zufallsvariable X eingesetzt und wir beobachten den Wert $X(\omega)$. In dieser Vorlesung sprechen wir nicht weiter über die „Ausführung von Zufall“, wir modellieren Wahrscheinlichkeiten. Für die Ausführung von Zufall (wir sagen die Realisierung der Zufallsvariablen) verweisen wir auf Vorlesungen über Monte Carlo Methoden oder Philosophie.

Definition 4.1.3. Sei X eine Zufallsvariable auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$.

(i) Die **Verteilung der Zufallsvariablen** X ist definiert als

$$\mathbb{P}_X(B) := \mathbb{P}(X \in B) \overset{\text{Notation}}{=} \mathbb{P}(\{\omega \in \Omega\colon X(\omega) \in B\}) = \mathbb{P}(X^{-1}(B)), \quad B \in \mathcal{B}(\mathbb{R}).$$

Unabhängig von dem zugrundeliegenden Wahrscheinlichkeitsraums ist $\mathbb{P}_X$ also ein Maß auf $\mathcal{B}(\mathbb{R})$ und zwar der push-forward von $\mathbb{P}$ unter X.

(ii) Die **Verteilungsfunktion der Zufallsvariablen** X ist definiert als

$$F_X(t) := \mathbb{P}_X((-\infty, t]) \overset{\text{Def.}}{=} \mathbb{P}(X \leq t), \quad t \in \mathbb{R}.$$

Wir schreiben $X \sim F_X$ und $X \sim \mathbb{P}_X$ und sagen „X ist verteilt gemäß F_X bzw. X ist verteilt gemäß $\mathbb{P}_X$“.

Weil so viele Indizes natürlich nerven, werden wir immer $X \sim F$ schreiben, wenn wir meinen, dass X gemäß F verteilt ist. Wir sagen dann auch kurz „X hat Verteilungsfunktion F“. Beachte: Aufgrund der Definition ist natürlich $\mathbb{P}_X = \mathbb{P}_{F_X}$, das werden wir immer mal wieder nutzen.

Wir haben uns schon in der Maßtheorie langsam an den Begriff $\mu(\{f \leq t\})$ als Abkürzung für $\mu(\{\omega : f(\omega) \leq t\})$ gewöhnen müssen. In der Stochastik gehen wir noch einen Schritt weiter und lassen die Klammern auch noch weg. Wir schreiben daher immer abkürzend

$$\mathbb{P}(X < t), \quad \mathbb{P}(X \in (a, b]), \quad \mathbb{P}(X = a)$$

und so weiter. Das liest sich als „Wahrscheinlichkeit, dass X kleiner als t ist“ auch ziemlich natürlich.

Definition 4.1.4. Zufallsvariablen X und Y, möglicherweise auf verschiedenen Wahrscheinlichkeitsräumen, heißen **identisch verteilt**, falls $F_X = F_Y$ bzw. $\mathbb{P}_X = \mathbb{P}_Y$. Man schreibt dann $X \sim Y$. Haben zwei Zufallsvariablen die gleiche Verteilungsfunktion, so sehen wir sie als gleichwertig an.

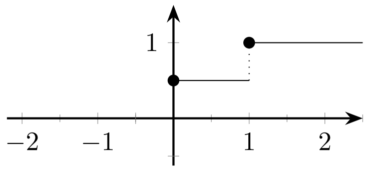
Zurück zum Münzwurf mit Auszahlung 1 für Zahl und 0 für Kopf. Wir schreiben dazu wieder zwei stochastische Modelle hin. Zum einen sei $\Omega = \{\text{Kopf}, \text{ Zahl}\}$, $\mathcal{A} = \mathcal{P}(\Omega)$ und

$$\mathbb{P}(\{\text{Kopf}\}) = \mathbb{P}(\{\text{Zahl}\}) = \frac{1}{2}, \quad \mathbb{P}(\Omega) = 1, \quad \mathbb{P}(\emptyset) = 0,$$

mit Zufallsvariable (Auszahlungsfunktion) definiert als

$$X(\text{Kopf}) = 0, \quad X(\text{Zahl}) = 1.$$

Zum anderen sei $\Omega = \mathbb{R}$, $\mathcal{A} = \mathcal{B}(\mathbb{R})$ und $\mathbb{P} = \mathbb{P}_F$ mit der Verteilungsfunktion



Da wir diesmal die Auszahlung schon direkt im Modell modelliert haben, zahlen wir genau den Betrag aus, der zufällig gezogen wird. Das machen wir mit der Zufallsvariablen $Y(\omega) = \omega$. Berechnen wir nun die Verteilungen der Zufallsvariablen X und Y:

$$\begin{aligned}\mathbb{P}_X(B) \overset{\text{Def.}}{=} \mathbb{P}(X \in B) &= \begin{cases} \frac{1}{2} & : 0 \in B, 1 \notin B \text{ oder } 1 \in B, 0 \notin B \\ 1 & : 0, 1 \in B \\ 0 & : 0 \notin B, 1 \notin B \end{cases} \\ &= \frac{1}{2}\delta_0(B) + \frac{1}{2}\delta_1(B)\end{aligned}$$

sowie

$$\mathbb{P}_Y(B) \overset{\text{Def.}}{=} \mathbb{P}(Y \in B) = \mathbb{P}(\{\omega \in \mathbb{R} : Y(\omega) \in B\}) = \mathbb{P}(\{\omega \in \mathbb{R} : \omega \in B\}) = \mathbb{P}(B) \overset{\text{Def.}}{=} \mathbb{P}_F(B).$$

Weil $\mathbb{P}_F = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$, gilt also $\mathbb{P}_X = \mathbb{P}_Y$ bzw. $F_X = F_Y$. Damit sind X und Y identische verteilte Zufallsvariablen und wir sehen die beiden stochastischen Modelle als gleichwertige Modelle für den Münzwurf an. Vorlesung 17

Definition 4.1.5.

(i) Eine Zufallsvariable X heißt **absolutstetig** mit Dichte f, falls die Verteilungsfunktion F_X von X die Dichte f hat, es gilt also

$$\mathbb{P}(X \in (a,b]) = \mathbb{P}_X((a,b]) = F_X(b) - F_X(a) = \int_a^b f(x)\,\mathrm{d}x, \quad a < b.$$

(ii) Eine Zufallsvariable X heißt **diskret**, falls F_X eine diskrete Verteilungsfunktion ist (oder $\mathbb{P}_X$ ein diskretes Maß ist). In anderen Worten: X nimmt nur abzählbar viele Werte $a_1, ..., a_N$ mit Wahrscheinlichkeiten $p_1, ..., p_N$ an:

$$\mathbb{P}(X = a_k) = \mathbb{P}_X(\{a_k\}) = p_k, \quad k = 1, ..., N.$$

Ihr merkt euch am besten folgende Rechenformeln:

$$\mathbb{P}(X \in A) = \begin{cases} \int_A f(x)\,\mathrm{d}x & : X \text{ absolut stetig} \\ \sum_{a_k \in A} p_k & : X \text{ diskret} \end{cases},$$

man integriert oder summiert über die Werte, die X annehmen soll. Dumm ist nur, wenn wir nicht wissen, dass X absolutstetig oder diskret ist. Dann ist eine Zufallsvariable halt irgendeine reellwertige messbare Abbildung auf irgendeinem Wahrscheinlichkeitsraum mit irgendeiner Verteilungsfunktion F.

Beispiel 4.1.6. Wir kennen die meisten wichtigen Beispiele schon, manche kommen noch dazu:

Diskrete Zufallsvariablen	Absolutstetige Zufallsvariablen
$X \sim \mathrm{Poi}(\lambda)$	$X \sim \mathcal{N}(\mu, \sigma^2)$
$X \sim \mathrm{Bin}(n, p)$	$X \sim \mathrm{Cauchy}(s, t)$
$X \sim \mathrm{Ber}(p)$	$X \sim \mathrm{Exp}(\lambda)$
$X \sim \mathrm{Geo}(\lambda)$	$X \sim \mathcal{U}([a, b])$
$X \sim$ Gleichverteilt auf endlichem Ω	$X \sim \Gamma(\alpha, \beta)$
	$X \sim \mathrm{Pareto}(k, a)$

Im Appendix gibt es eine Übersicht über die wichtigsten Verteilungen.

Um die Begriffe auszuprobieren, schauen wir uns eine kleine Rechnung an. Wir behaupten, dass $Y \sim \mathrm{Exp}(1)$, wenn $Y = -\ln(U)$ mit $U \sim \mathcal{U}([0, 1])$. Probieren wir die Begriffe aus und berechnen definitionsgemäß die Verteilungsfunktion von Y durch Auflösen:

$$F_Y(t) = \mathbb{P}(Y \leq t) \overset{\text{Def.}}{=} \mathbb{P}(-\ln(U) \leq t) = \mathbb{P}(U \geq \exp(-t)) = 1 - \mathbb{P}(U \leq \exp(-t)).$$

Wenn wir jetzt die Verteilungsfunktion von $\mathcal{U}([0, 1])$ einsetzen, bekommen wir $F_Y(t) = (1 - e^{-t})\mathbf{1}_{[0,\infty)}(t)$ und das ist die Verteilungsfunktion der Exponentialverteilung.

Wir waren ein klein wenig ungenau weil Y den Wert $+\infty$ annehmen kann, eine Zufallsvariable per Definition aber nur reelle Werte annimmt. Das kann man einfach reparieren, indem man zum Beispiel $\log(0) = 0$ umdefiniert. Alternativ nimmt man $U \sim \mathcal{U}((0, 1))$ statt $U \sim \mathcal{U}([0, 1])$, das diskutieren wir ausführlich in Abschnitt 4.3.1.

Es ist jetzt hoffentlich klar, was ein stochastisches Modell sein soll. Aber gibt es für jede Verteilungsfunktion überhaupt solch Modell? Ja! Die in folgendem Beweis auftauchende Konstruktion heißt „kanonische Konstruktion" und wird in der Stochastik in diversen Kontexten immer wieder benutzt.

Satz 4.1.7. **[Existenz stochastischer Modelle]** Für jede Verteilungsfunktion F existiert eine Zufallsvariable X mit $X \sim F$. Genauer: Es existiert ein stochastisches Modell, d. h. ein Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ und eine Zufallsvariable X auf $(\Omega, \mathcal{A}, \mathbb{P})$, mit $X \sim F$.

Beweis. Als Wahrscheinlichkeitsraum definieren wir $\Omega = \mathbb{R}$, $\mathcal{A} = \mathcal{B}(\mathbb{R})$, $\mathbb{P} = \mathbb{P}_F$ und darauf die Zufallsvariable $X(\omega) = \omega$. Beachte: $X(\omega) = \omega$ ist eine stetige Abbildung von $\mathbb{R}$ nach $\mathbb{R}$ und damit auch Borel-messbar. Berechnen wir die Verteilungsfunktion dieser konkreten Zufallsvariablen auf dem konkreten Wahrscheinlichkeitsraum:

$$F_X(t) = \mathbb{P}(X \leq t) = \mathbb{P}_F(\{\omega : X(\omega) \leq t\}) = \mathbb{P}_F(\{\omega : \omega \leq t\}) = \mathbb{P}_F((-\infty, t]) = F(t), \quad t \in \mathbb{R}.$$

Also gilt $X \sim F$, das war es schon! Zu beachten ist, dass die Konstruktion weit von trivial ist. Die Existenz von $\mathbb{P}_F$ benötigt den Satz von Carathéodory und damit die komplette Maßtheorie. □

Bemerkung 4.1.8. Wenn wir uns nur für diskrete Zufallsvariablen interessieren würden, kämen wir komplett *ohne* Maßtheorie aus! Hier ist eine viel einfachere Konstruktion, die für alle diskrete Zufallsvariablen funktioniert. Sei dazu F eine diskrete Verteilungsfunktion, die an N-vielen Stellen a_k um p_k nach oben springt. Sei nun $\Omega = \{\omega_1, ..., \omega_N\}$ eine beliebige Menge mit N Elementen (z.B. $\Omega = \{\mathrm{Kopf}, \mathrm{Zahl}\}$ beim Würfeln). Auf Ω wählen wir als σ-Algebra $\mathcal{A} = \mathcal{P}(\Omega)$. Das Maß definieren wir, indem wir es auf den Elementarereignissen als $\mathbb{P}(\{\omega_k\}) = p_k$ definieren und mit der σ-Additivität für beliebiges $A \in \mathcal{A}$ fortsetzen, d. h. $\mathbb{P}(A) = \sum_{\omega_k \in A} \mathbb{P}(\{\omega_k\}) = \sum_{\omega_k \in A} p_k$. An dieser Stelle nutzen wir die Abzählbarkeit. Als Zufallsvariable wählen wir die Abbildung $X(\omega_k) := a_k$. Weil auf dem Urbildraum die Potenzmenge gewählt wurde, ist natürlich jede Abbildung nach $\mathbb{R}$ auch $(\mathcal{A}, \mathcal{B}(\mathbb{R}))$-messbar. Damit gilt $\mathbb{P}(X = a_k) = p_k$, also ist X gemäß F verteilt. Diese Konstruktion funktioniert nur für diskrete Verteilungsfunktionen so einfach (probiert es einfach mal für absolutstetige Verteilungsfunktionen aus, ihr werdet schnell den Fortsetzungssatz von Carathéodory brauchen). Im Allgemeinen kommen wir nicht umher, die Maßtheorie wie im Beweises von Satz 4.1.7 zu nutzen weil man Maße auf überabzählbaren Mengen nicht einfach auf den Elementarereignissen definieren kann.

Hier sind zwei konkrete Beispiele: Wenn jemand von euch ein stochastisches Modell für eine $\mathcal{N}(0,1)$-verteilte Zufallsvariable verlangt, so entgegnet ihr

$$\Omega = \mathbb{R}, \quad \mathcal{A} = \mathcal{B}(\mathbb{R}), \quad \mathbb{P} = \mathbb{P}_F, \quad X(\omega) = \omega,$$

wobei $\mathbb{P}_F$ das eindeutige Maß auf $\mathcal{B}(\mathbb{R})$ mit Verteilungsfunktion $F \sim \mathcal{N}(0,1)$ aus Carathéodory ist. Will jemand das gleiche für eine Poi(λ)-verteilte Zufallsvariable haben, so entgegnet ihr entweder das gleiche, oder

$$\Omega = \mathbb{N}, \quad \mathcal{A} = \mathcal{P}(\mathbb{N}), \quad \mathbb{P}(\{k\}) = e^{-\lambda}\frac{\lambda^k}{k!}, \quad X(\omega) = \omega.$$

Um ganz deutlich zu sein: Diskret ist einfach, weil die Potenzmenge von abzählbaren Mengen noch klein genug ist und Maße durch die abzählbare σ-Additivität von Maßen nur auf einpunktigen Mengen definiert werden müssen! Weil all das für überabzählbare Mengen wie $\mathbb{R}$ schief geht, haben wir Maßtheorie gemacht. Auf die Normalverteilung können wir einfach nicht verzichten!

Definition 4.1.9. Ist X eine Zufallsvariable auf $(\Omega, \mathcal{A}, \mathbb{P})$, so heißen, falls die Integrale wohldefiniert sind,

(i)
$$\mathbb{E}[X] := \int_\Omega X(\omega)\,\mathrm{d}\mathbb{P}(\omega) \qquad \textbf{Erwartungswert} \text{ von X},$$

(ii)
$$\mathbb{E}[X^k] := \int_\Omega X^k(\omega)\,\mathrm{d}\mathbb{P}(\omega) \qquad k\textbf{-tes Moment} \text{ von X für } k \in \mathbb{N},$$

(iii)
$$\mathbb{V}[X] := \mathbb{E}[(X - \mathbb{E}[X])^2] \qquad \textbf{Varianz} \text{ von X},$$

(iv)
$$\mathbb{E}[e^{\lambda X}] := \int_\Omega e^{\lambda X(\omega)}\,\mathrm{d}\mathbb{P}(\omega) \qquad \textbf{exponentielles Moment} \text{ von X für } \lambda \in \mathbb{R}.$$

Allgemein betrachten wir für $g : \mathbb{R} \to \overline{\mathbb{R}}$ messbar die Erwartungswerte

$$\mathbb{E}[g(X)] := \int_\Omega g(X(\omega))\,\mathrm{d}\mathbb{P}(\omega),$$

falls das Integral wohldefiniert ist. Wir sagen die Erwartungswerte existieren, falls die Integrale existieren (also endlich sind).

Wir erinnern daran, dass ein wohldefiniertes Integral auch die Werte $+\infty$ oder $-\infty$ annehmen darf. Meistens werden wir aber davon sprechen, dass die Integrale existieren, also endlich sind. Wegen der allgemeinen Äquivalenzen

$$\int f\,\mathrm{d}\mu \text{ existiert} \quad \overset{\text{Def.}}{\Leftrightarrow} \quad \int f^+\,\mathrm{d}\mu < \infty, \int f^-\,\mathrm{d}\mu < \infty \quad \Leftrightarrow \quad \int |f|\,\mathrm{d}\mu < \infty,$$

schreiben wir meistens bequemer „$\mathbb{E}[|g(X)|] < \infty$“ anstelle von „$\mathbb{E}[g(X)]$ existiert“. Bei Erwartungswerten sagen wir also „der Erwartungswert von X existiert“, wenn der Erwartungswert wohldefiniert und endlich ist. Das wird oft in der Literatur genauso gehandhabt, manchmal aber auch anders. Manche sagen, der Erwartungswert existiert, wenn $\mathbb{E}[X]$ wohldefiniert ist (aber vielleicht unendlich) ist.

Bemerkung. Die Notation $\mathbb{E}[g(X)]$ ist etwas unglücklich weil das Integral nicht nur von g und X abhängt, sondern auch von $\mathbb{P}$. Daher sollte man eher $\mathbb{E}_\mathbb{P}[g(X)]$ schreiben, man lässt aber üblicherweise das $\mathbb{P}$ aus Faulheit weg.

Lemma 4.1.10. Ist die Zufallsvariable X gemäß F verteilt, d. h. $X \sim F$, so gilt unabhängig von dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ auf dem X definiert ist,

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}} g(x) \, \mathrm{d}\mathbb{P}_F(x)$$

für messbare numerische Abbildungen $g \colon \mathbb{R} \to \overline{\mathbb{R}}$. Wie immer gilt die Gleichheit, wenn eine Seite (und damit die die andere Seite) wohldefiniert ist.

Beweis. Mit dem Transformationssatz 3.1.16 (bzw. Korollar (3.1.17) gilt in einem Schaubild

$$\begin{tikzcd}
(\Omega, \mathcal{A}, \mathbb{P}) \arrow[r, "X"] \arrow[dr, "g \circ X"'] & (\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P}_X) \arrow[d, "g"] \\
& (\overline{\mathbb{R}}, \mathcal{B}(\overline{\mathbb{R}}))
\end{tikzcd}$$

wobei $\mathbb{P}_X$ der push-forward von X ist. Weil definitionsgemäß $\mathbb{P}_X = \mathbb{P}_F$ ist, gibt das sauber ausgeschrieben

$$\mathbb{E}[g(X)] \overset{\text{Def.}}{=} \int_{\Omega} g(X(\omega)) \, \mathrm{d}\mathbb{P}(\omega) \overset{3.1.16}{=} \int_{\mathbb{R}} g(x) \, \mathrm{d}\mathbb{P}_X(x) \overset{\mathbb{P}_X = \mathbb{P}_F}{=} \int_{\mathbb{R}} g(x) \, \mathrm{d}\mathbb{P}_F(x).$$

□

Die Konsequenz ist natürlich, dass für beliebiges g, $\mathbb{E}[g(X)]$ gar nicht von dem kompletten Modell $(\Omega, \mathcal{A}, \mathbb{P}, X)$ abhängt, $\mathbb{E}[g(X)]$ hängt einfach nur von der Verteilung von X ab. Das ist der Grund, warum man sich üblicherweise nur für die Verteilungsfunktion von X, jedoch nicht für $(\Omega, \mathcal{A}, \mathbb{P})$ interessiert. Fassen wir die Beobachtung zusammen:

Bemerkung. Sind X, Y identisch verteilte Zufallsvariablen, die auf irgendwelchen Wahrscheinlichkeitsräumen definiert sind, so gilt

$$\mathbb{E}[g(X)] = \mathbb{E}[g(Y)]$$

für alle $g \colon \mathbb{R} \to \overline{\mathbb{R}}$ messbar.

Jetzt wissen wir auch schon, wie wir $\mathbb{E}[g(X)]$ berechnen können, das haben wir nämlich schon gemacht. Vieles was jetzt kommt sind Wiederholungen, indem wir Sätze für allgemeine Integrale nochmal für Erwartungswerte hinschreiben.

Satz 4.1.11. **[Berechnungsregeln]** Sei X eine Zufallsvariable auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ und $g : \mathbb{R} \to \overline{\mathbb{R}}$ messbar, so gelten:

(i) Ist X absolutstetig mit Dichte f, so gilt

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

(ii) Ist X diskret und nimmt die Werte $a_1, ..., a_N \in \mathbb{R}$ mit Wahrscheinlichkeiten $p_1, ..., p_N$ an, so gilt

$$\mathbb{E}[g(X)] = \sum_{k=1}^{N} p_k \, g(a_k) = \sum_{k=1}^{N} \mathbb{P}(X = a_k) g(a_k).$$

Dabei ist wieder eine Seite genau dann wohldefiniert, wenn es die andere Seite ist.

Beweis. Dazu kombinieren wir nur die Formel aus Satz 4.1.10 mit den Formeln aus Satz 3.3.2 und Satz 3.3.3. □

Beispiel 4.1.12.

- Für $X \sim \mathcal{N}(\mu, \sigma^2)$ gilt $\mathbb{E}[X] = \int_{\mathbb{R}} x \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \,\mathrm{d}x = \mu$.
- Für $X \sim \mathrm{Ber}(p)$ gilt $\mathbb{E}[X] = 1 \cdot \mathbb{P}(X = 1) + 0 \cdot \mathbb{P}(X = 0) = p$.
- Für $X \sim \mathrm{Poi}(\lambda)$ gilt $\mathbb{E}[X] = \sum_{k=0}^{\infty} k \cdot \mathbb{P}(X = k) = \sum_{k=0}^{\infty} k e^{-\lambda} \frac{\lambda^k}{k!} = \lambda$.

Wir sehen also: Ein großer Teil der Stochastik besteht aus dem Berechnen von Integralen und Summen bzw. Reihen.

Vorlesung 18

Beispiel 4.1.13. Eine Webseite wird im Mittel pro Stunde zweimal geklickt. Wie groß ist die Wahrscheinlichkeit, dass die Webseite in einer Stunde mindestens fünfmal geklickt wird?

Um das Beispiel stochastisch zu behandeln, müssen wir zunächst ein Modell annehmen. Welche uns bekannte Zufallsvariable könnte die Anzahl der Klicks pro Stunde modellieren? Da die Ergebnisse der Zufallszahlen natürliche Zahlen sind, muss die Verteilung diskret sein. Nun ist die Anzahl nicht beschränkt, es sollte also eine Zufallsvariable sein, die alle Werte in $\mathbb{N}$ annehmen kann. Dazu kennen wir bisher nur die geometrische- oder die Poissonverteilung. An der jetzigen Stelle können wir ohne weitere Annahmen keine von beiden ausschließen. Nehmen wir einfach mal an, dass $X \sim \mathrm{Poi}(\lambda)$ ein gutes Modell ist. Aber was ist λ? Weil wir wissen, dass $\mathbb{E}[X] = \lambda$ ist, schließen wir $\lambda = 2$ aus der Vorinformation (Das Gesetz der großen Zahlen wird uns in Satz 4.6.8 des letzten Kapitel die Rechtfertigung geben, warum wir aus dem Mittel auf den Erwartungswert schließen.) Um nun die Aufgabe zu lösen, müssen wir für eine Poi(2)-verteilte Zufallsvariable $\mathbb{P}(X \geq 5)$ berechnen. Das geht ganz einfach:

$$\begin{aligned}
\mathbb{P}(X \geq 5) &= 1 - \mathbb{P}(X \leq 4) \\
&= 1 - \sum_{k=0}^{4} \mathbb{P}(X = k) \\
&= 1 - e^{-2} \left(\frac{1}{1} + \frac{2}{1} + \frac{4}{2} + \frac{8}{6} + \frac{16}{24} \right) \approx 0{,}053.
\end{aligned}$$

Hierbei haben wir genutzt, dass für eine diskrete Zufallsvariable immer

$$\mathbb{P}(X \in A) = \sum_{a_k \in A} \mathbb{P}(X = a_k)$$

gilt. Das folgt natürlich aus der σ-Additivität von Maßen weil $A \mapsto \mathbb{P}(X \in A)$ ein Maß ist (die Verteilung von X).

Proposition 4.1.14. **[Rechenregeln für den Erwartungswert]** Seien X, Y Zufallsvariablen auf $(\Omega, \mathcal{A}, \mathbb{P})$ mit $\mathbb{E}[|X|], \mathbb{E}[|Y|] < \infty$ und $\alpha, \beta \in \mathbb{R}$, so gelten:

(i) $\mathbb{E}[\alpha X + \beta Y] = \alpha \mathbb{E}[X] + \beta \mathbb{E}[Y]$

(ii) $X \geq 0$ $\mathbb{P}$-f.s. $\Rightarrow \mathbb{E}[X] \geq 0$ und $X \geq Y$ $\mathbb{P}$-f.s. $\Rightarrow \mathbb{E}[X] \geq \mathbb{E}[Y]$

(iii) Ist $X = \alpha$ $\mathbb{P}$-f.s., so ist $\mathbb{E}[X] = \alpha$.

(iv) $\mathbb{P}(X \in A) = \mathbb{E}[\mathbf{1}_A(X)]$, insbesondere gilt $F_X(t) = \mathbb{E}[\mathbf{1}_{(-\infty,t]}(X)]$, $t \in \mathbb{R}$.

Beweis. Wir müssen nur beachten, dass Erwartungswerte per Definition Integrale sind. Dann können wir die Rechenregeln für Integrale direkt anwerden. Zu beachten ist, dass wegen Satz 3.1.15 Änderungen auf Nullmengen Integrale nicht ändern.

(i) Linearität von Integralen

(ii) Monotonie von Integralen (die Nullmengen spielen keine Rolle)

(iii) Nach Annahme gilt $X = \alpha \mathbf{1}_\Omega$ $\mathbb{P}$-f.s. Wegen Satz 3.1.15 können wir sofort die Definition des Integrals einsetzen:

$$\mathbb{E}[X] = \int_\Omega X(\omega)\,\mathrm{d}\mathbb{P}(\omega) = \int_\Omega \underbrace{\alpha \mathbf{1}_\Omega}_{\text{einfach}}\,\mathrm{d}\mathbb{P} \overset{\text{Def.}}{=} \alpha \mathbb{P}(\Omega) = \alpha$$

(iv) Hier müssen wir nur die Definitionen im Kopf klar bekommen:

$$\mathbb{E}[\mathbf{1}_A(X)] = \int_\Omega \mathbf{1}_A(X(\omega))\,\mathrm{d}\mathbb{P}(\omega) \overset{\text{Trafo}}{=} \int_\mathbb{R} \underbrace{\mathbf{1}_A(x)}_{\text{einfach}}\,\mathrm{d}\mathbb{P}_X(x) = \mathbb{P}_X(A) \overset{\text{Def.}}{=} \mathbb{P}(X \in A)$$

□

Natürlich ist eine Zufallsvariable, die fast sicher den selben Wert annimmt, gar keine interessante Zufallsvariable! Das modellierte Zufallsexperiment ist gar nicht zufällig, es passiert immer das gleiche! Beispiel: Jeden Tag um 7 Uhr wird die Zeit (Stunde) angeschaut. Es kommt immer 7 dabei raus, die beschreibende Zufallsvariable erfüllt also $\mathbb{P}(X = 7) = 1$, also $X = 7$ $\mathbb{P}$-fast sicher. Viel interessanter wäre zum Beispiel, jeden Tag um 7 Uhr die Temperatur zu messen. Die entsprechende Zufallsvariable wäre nicht fast sicher konstant.

Korollar 4.1.15. **[Rechenregeln für die Varianz]** Sei X eine Zufallsvariable mit $\mathbb{E}[X^2] < \infty$ (wir sagen auch, X ist quadratintegrierbar), so gelten:

(i) Es gilt $\mathbb{V}[X] < \infty$ und
$$\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

(ii) Es gilt $\mathbb{V}[X] = 0$ genau dann, wenn X fast sicher den gleichen Wert annimmt, dieser ist dann $\mathbb{E}[X]$.

(iii) Für $a \in \mathbb{R}$ gilt $\mathbb{V}[aX] = a^2\mathbb{V}[X]$ und $\mathbb{V}[a + X] = \mathbb{V}[X]$.

Beweis. Übung. Beachte dazu: Ist $Y \geq 0$ $\mathbb{P}$-fast sicher und $\mathbb{E}[Y] = 0$, so ist $Y \equiv 0$ $\mathbb{P}$-fast sicher. Das gilt wegen Satz 3.1.15, der Erwartungswert ist schließlich ein Integral! □

Die Varianz misst als Kenngröße die Variabilität von Zufallsvariablen. Ist die Varianz null, so gibt es gar keine Variabilität, ist die Varianz groß, so nimmt X auch Werte an, die weit von dem Erwartungswert entfernt sind. Das ist mathematisch natürlich keine saubere Formulierung (dafür gibt es schließlich die Varianz), gibt aber das richtige Gefühl. Daher ist auch einleuchtend, dass sich die Variabilität beim Verschieben um einen festen Wert sich nicht ändert.

Satz 4.1.16. **[Konvergenzsätze für Zufallsvariablen]** Seien $Y, X, X_1, X_2, ...$ Zufallsvariablen auf $(\Omega, \mathcal{A}, \mathbb{P})$ mit $\lim_{n\to\infty} X_n = X$ $\mathbb{P}$-fast sicher.

(i) MCT: Gilt $0 \leq X_1 \leq X_2 \leq ... \leq X$ $\mathbb{P}$-fast sicher, so gilt
$$\lim_{n\to\infty} \mathbb{E}[X_n] = \mathbb{E}[X].$$

(ii) DCT: Gilt $|X_n| \leq Y$ $\mathbb{P}$-fast sicher für alle $n \in \mathbb{N}$ mit $\mathbb{E}[|Y|] < \infty$, so gilt
$$\lim_{n\to\infty} \mathbb{E}[X_n] = \mathbb{E}[X].$$

(iii) Gilt $|X_n| \leq C$ $\mathbb{P}$-fast sicher für alle $n \in \mathbb{N}$ für ein $C > 0$, so gilt
$$\lim_{n\to\infty} \mathbb{E}[X_n] = \mathbb{E}[X].$$

Beweis. Wegen $\mathbb{E}[X_n] \overset{\text{Def.}}{=} \int_\Omega X_n\,\mathrm{d}\mathbb{P}$ und $\mathbb{E}[X] \overset{\text{Def.}}{=} \int_\Omega X\,\mathrm{d}\mathbb{P}$ ist das gerade Satz 3.2.1, Satz 3.2.5 und Korollar 3.2.6. □

Definition 4.1.17. Für eine Zufallsvariable X heißt

$$\mathcal{M}_X(t) := \mathbb{E}[e^{tX}], \quad t \in \mathbb{R},$$

die **momenterzeugende Funktion**. $\mathcal{M}_X$ ist nur für die t definiert, für die $\mathbb{E}[e^{tX}] < \infty$ gilt.

Die momenterzeugenden Funktionen sind auf ihrem Definitionsbereich ganz normale Funktionen von $\mathbb{R}$ nach $\mathbb{R}$. Wir können also über Ableitungen sprechen, Monotonie, und so weiter. In vielen Beispielen ist M_X eine ganz harmlose Funktion, manchmal ist $\mathcal{M}_X$ aber auch gar nicht definiert.

Beispiel 4.1.18.

- Sei $X \sim \text{Cauchy}(s,t)$, so ist $\mathcal{M}_X$ nirgends definiert!
- Sei $X \sim \mathcal{N}(\mu, \sigma^2)$, so ist $\mathcal{M}_X(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right)$ für $t \in \mathbb{R}$, siehe Übungsaufgabe.
- Sei $X \sim \text{Poi}(\lambda)$, so ist

  $$\mathcal{M}_X(t) = \mathbb{E}[e^{tX}] = \sum_{k=0}^{\infty} e^{tX}\mathbb{P}(X=k) = \sum_{k=0}^{\infty} e^{tk}e^{-\lambda}\frac{\lambda^k}{k!} = e^{\lambda(e^t-1)}$$

  für alle $t \in \mathbb{R}$.

Noch viel mehr explizite Beispiele sind im Appendix gesammelt.

Das ganze ist ein so nützliches Konzept, weil wir viele Beispiele explizit ausrechnen können und mit dem nächsten Satz gleich noch alle Momente durch Ableiten ausrechnen können:

Satz 4.1.19. Sei X eine Zufallsvariable, für die für irgendein $\varepsilon > 0$ die Momenterzeugende Funktion $\mathcal{M}_X$ auf $(-\varepsilon, \varepsilon)$ definiert ist. Dann ist $\mathcal{M}_X$ an der Stelle 0 unendlich oft differenzierbar und es gilt für alle $n \in \mathbb{N}$

$$\mathbb{E}[X^n] = \mathcal{M}_X^{(n)}(0),$$

wobei $\mathcal{M}_X^{(n)}(0)$ die n-te Ableitung an der Stelle 0 ist.

Beweis.

(a) Wir zeigen zunächst, dass $\mathcal{M}_X$ in $(-\varepsilon, \varepsilon)$ eine Potenzreihe ist. Nach Analysis 1 gilt punktweise (also für jedes $\omega \in \Omega$)

$$e^{tX(\omega)} = \sum_{k=0}^{\infty} \frac{(tX(\omega))^k}{k!} = \lim_{m\to\infty} \sum_{k=0}^{m} \frac{(tX(\omega))^k}{k!} =: \lim_{m\to\infty} S_m(\omega).$$

Die Zufallsvariable e^{tX} kann also als punktweiser Grenzwert der Folge $(S_m)_{m\in\mathbb{N}}$ von Zufallsvariablen geschrieben werden. Um gleich dominierte Konvergenz zu nutzen, brauchen wir eine integrierbare Majorante S für die Folge (S_m). Das ist gar nicht so schwer:

$$|S_m| \overset{\text{Def.}}{=} \left|\sum_{k=0}^{m} \frac{(tX)^k}{k!}\right| \overset{\triangle}{\leq} \sum_{k=0}^{m} \left|\frac{(tX)^k}{k!}\right| \leq \sum_{k=0}^{\infty} \left|\frac{(tX)^k}{k!}\right| =: S.$$

Wegen $S = e^{|tX|} \leq e^{tX} + e^{-tX}$ ist S eine integrierbare Majorante:

$$\mathbb{E}[|S|] = \mathbb{E}[e^{|tX|}] \underset{\text{Lin.}}{\overset{\text{Mon.}}{\leq}} \mathbb{E}[e^{tX}] + \mathbb{E}[e^{-tX}] \overset{\text{Def.}}{=} \mathcal{M}_X(t) + \mathcal{M}_X(-t) < \infty$$

für $t \in (-\varepsilon, \varepsilon)$ nach Annahme. Jetzt kann also dominierte Konvergenz angewandt werden. Es gilt damit für $t \in (-\varepsilon, \varepsilon)$, dass

$$M_X(t) = \mathbb{E}[\lim_{m\to\infty} S_m] \overset{\text{DCT}}{=} \lim_{m\to\infty} \mathbb{E}[S_m] \overset{\text{Lin.}}{=} \lim_{m\to\infty} \sum_{k=0}^{m} \frac{t^k \mathbb{E}[X^k]}{k!} = \sum_{k=0}^{\infty} \frac{t^k \mathbb{E}[X^k]}{k!}.$$

Damit ist M_X in $(-\varepsilon, \varepsilon)$ eine Potenzreihe mit Koeffizienten $a_k = \frac{\mathbb{E}[X^k]}{k!}$ und Entwicklungspunkt $x_0 = 0$.

(b) Aus Analysis 1 wisst ihr (so habt ihr die Taylor-Koeffizienten bestimmt!), dass M_X in $(-\varepsilon, \varepsilon)$ unendlich oft differenzierbar ist und die Reihe gliedweise differenziert werden kann. n-faches Ableiten und den Entwicklungspunkt $x_0 = 0$, einsetzen gibt dann $M_X^{(n)}(0) = \mathbb{E}[X^n]$.

□

Beispiel 4.1.20.

- Für die Normalverteilungen $X \sim \mathcal{N}(\mu, \sigma^2)$ ergibt die explizite Formel der momenterzeugenden Funktion mit dem Satz

$$\begin{aligned}\mathbb{E}[X] &= \mathcal{M}'_X(0) = \exp\Big(\mu \cdot 0 + \frac{\sigma^2 t^2}{2}\Big) \cdot (\mu + \sigma^2 \cdot 0) = \mu,\\ \mathbb{E}[X^2] &= \mathcal{M}''_X(0) = \mu^2 + \sigma^2.\end{aligned}$$

  Beides zusammen ergibt $\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mu^2 + \sigma^2 - \mu^2 = \sigma^2$. Die zwei Parameter der Normalverteilung sind also gerade Erwartungswert μ und Varianz σ^2.

- Für $X \sim \text{Poi}(\lambda)$ ergibt die explizite Formel der momenterzeugenden Funktion mit dem Satz

$$\mathbb{E}[X] = \mathcal{M}'_X(0) = \lambda \quad \text{und} \quad \mathbb{E}[X^2] = \mathcal{M}''_X(0) = \lambda^2 + \lambda.$$

  Beides zusammen ergibt $\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \lambda$.

Proposition 4.1.21. **[Hölder und Jensen'sche Ungleichung]**

(i) Für $p, q > 1$ mit $\frac{1}{p} + \frac{1}{q} = 1$ und Zufallsvariablen X, Y auf $(\Omega, \mathcal{A}, \mathbb{P})$ gilt

$$\mathbb{E}[|XY|] \leq \big(\mathbb{E}[|X|^p]\big)^{1/p} \big(\mathbb{E}[|Y|^q]\big)^{1/q}.$$

(ii) Ist X eine Zufallsvariable mit $\mathbb{E}[|X|] < \infty$ und $\varphi\colon \mathbb{R} \to \mathbb{R}$ *konvex* mit $\mathbb{E}[|\varphi(X)|] < \infty$, so gilt

$$\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)].$$

Beweis. (i) Weil Erwartungswerte Integrale sind, ist das nur ein Spezialfall von Satz 3.4.1.

(ii) Wir geben den Beweis nur für differenzierbares φ. Wegen der Konvexität gibt es für jedes feste $x_0 \in \mathbb{R}$ ein $b \in \mathbb{R}$ mit

- $\varphi'(x_0)x + b \leq \varphi(x)$ für alle $x \in \mathbb{R}$,
- $\varphi'(x_0)x_0 + b = \varphi(x_0)$.

Wir wählen $x_0 = \mathbb{E}[X]$. Mit der zweiten Eigenschaft schreiben wir $\varphi(\mathbb{E}[X])$ wie folgt um:

$$\varphi(\mathbb{E}[X]) = \varphi(x_0) = \varphi'(x_0)x_0 + b = \varphi'(x_0)\mathbb{E}[X] + b.$$

Weil der Erwartungswert linear und monoton ist, sowie der Erwartungswert einer konstanten Zufallsvariable gerade die Konstante ist, können wir die rechte Seite wie folgt behandeln:

$$\varphi'(x_0)\mathbb{E}[X] + b = \mathbb{E}[\varphi'(x_0)X] + \mathbb{E}[b] = \mathbb{E}[\varphi'(x_0)X + b] \overset{\text{Mon.}}{\leq} \mathbb{E}[\varphi(x)].$$

Zusammen folgt die Behauptung. □

Beispiel. Als Merkregel für das „≤" in Proposition 4.1.21 nimmt man $\varphi(x) = x^2$. Weil

$$0 \leq \mathbb{V}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] \overset{\text{Üb.}}{=} \mathbb{E}[X^2] - \mathbb{E}[X]^2,$$

muss $\mathbb{E}[X]^2 \leq \mathbb{E}[X^2]$ gelten. Also muss in 4.1.21 „≤" und nicht „≥" stehen.

Vorlesung 19

Zum Abschluss nochmal die Markov- und Tschebyscheff-Ungleichungen, die wir für Integrale über beliebige Maße schon angeschaut haben. Weil Erwartungswerte Integrale sind, geht das in diesem Spezialfall natürlich genauso:

Satz 4.1.22. **[Markov- und Tschebyscheff-Ungleichung]** Sei X eine Zufallsvariable, dann gelten für alle $a > 0$ folgende Ungleichungen:

(i) Für $h\colon \mathbb{R} \to (0, \infty)$ wachsend gilt

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[h(X)]}{h(a)} \qquad \text{(Markov-Ungleichung)}$$

(ii) Für $h\colon [0, \infty) \to (0, \infty)$ wachsend gilt

$$\mathbb{P}(|X| \geq a) \leq \frac{\mathbb{E}[h(|X|)]}{h(a)} \qquad \text{(Markov-Ungleichung)}$$

(iii)

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq a) \leq \frac{\mathbb{V}[X]}{a^2} \qquad \text{(Tschebyscheff-Ungleichung)}$$

Beweis.

(i) Definiere $A = [a, \infty)$, dann gilt weil h wachsend ist

$$\begin{aligned}\mathbb{E}[h(X)] &\overset{\text{Mon.}}{\geq} \mathbb{E}[h(X) \cdot \mathbf{1}_A(X)] \\ &\geq \mathbb{E}[h(a) \cdot \mathbf{1}_A(X)] \\ &\overset{\text{Lin.}}{=} h(a) \cdot \mathbb{E}[\mathbf{1}_A(X)] \\ &\overset{4.1.14,(iv)}{=} h(a) \cdot \mathbb{P}(X \geq a).\end{aligned}$$

Durchteilen gibt die Abschätzung. Hier haben wir den kleinen Trick genutzt, dass $1 \equiv \mathbf{1}_\Omega \equiv \mathbf{1}_A + \mathbf{1}_{A^C} \geq \mathbf{1}_A$ gilt. Der Trick wird jetzt immer wieder kommen!

(ii) Genau wie (i).

(iii) Benutze die Markov Ungleichung mit $h(x) = x^2$ und der „zentrierten" Zufallsvariablen $X - \mathbb{E}[X]$.

□

Wie bei den Konzentrationsungleichungen für Maße, können wir durch Bildung von Gegenwahrscheinlichkeiten sofort Ungleichungen für $\mathbb{P}(X < a)$ oder $\mathbb{P}(|X| < a)$ bekommen. Hier ein konkretes Beispiel: Ist $X \sim \mathcal{N}(\mu, \sigma^2)$, so gilt $\mathbb{P}(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}$.

4.2 Zufallsvektoren

Nachdem Zufallsvariablen jetzt hoffentlich einigermaßen klar geworden sind, gehen wir jetzt weiter zu Zufallsvektoren. Das sind Zufallsvariablen, deren Werte nicht reell, sondern aus dem $\mathbb{R}^d$ sind. Weil wir den Namen Zufallsvariablen nur für den Fall $d = 1$ definiert haben, sprechen wir nun von Zufallsvektoren. Das Kapitel ist in großen Teilen eine Wiederholung, wir gehen durch die gleichen Schritte, die Notationen werden nur einen Tick aufwendiger. Das Vorgehen ist genau wie für reelle (eindimensionale) Zufallsvariablen:

- Lege σ-Algebra auf $\mathbb{R}^d$ fest und zeige wichtige Eigenschaften für später.
- Charakterisiere Maße auf $\mathbb{R}^d$ durch Verteilungsfunktionen.

- Definiere Zufallsvektoren als messbare Abbildungen und verbinde diese zu Maßen und Verteilungsfunktionen.
- Definiere Erwartungswerte und zeige Rechnenregeln für diskrete und absolutstetige Zufallsvektoren.
- Rechnen, rechnen, rechnen.

(A) Borel-σ-Algebra auf $\mathbb{R}^d$

Definition 4.2.1. Wir wählen die Produkt-σ-Algebra auf dem $\mathbb{R}^d$, die aus d Kopien der Borel-σ-Algebra besteht:

$$\mathcal{B}(\mathbb{R}^d) := \underbrace{\mathcal{B}(\mathbb{R}) \otimes ... \otimes \mathcal{B}(\mathbb{R})}_{d\text{-viele}} \overset{\text{Def.}}{=} \sigma(\{B_1 \times ... \times B_d : B_i \in \mathcal{B}(\mathbb{R})\})$$

Wir hatten im ersten Kapitel schon erwähnt, dass die Definition der Borel-σ-Algebra als kleinste σ-Algebra erzeugt durch offene Mengen auch im $\mathbb{R}^d$ funktioniert. Das ist in der Tat das selbe wie die gerade definierte Produkt-σ-Algebra $\mathcal{B}(\mathbb{R}^d)$.

Lemma 4.2.2. Es gilt

$$\begin{aligned}\mathcal{B}(\mathbb{R}^d) &= \sigma(\{O \subseteq \mathbb{R}^d : O \text{ offen}\})\\ &= \sigma(\{(-\infty, t_1] \times ... \times (-\infty, t_d] : t_i \in \mathbb{R}\})\\ &= \sigma(\{(a_1, b_1] \times ... \times (a_d, b_d] : a_i, b_i \in \mathbb{R}\})\\ &= \sigma(\{(a_1, b_1) \times ... \times (a_d, b_d) : a_i, b_i \in \mathbb{R}\})\\ &= ...,\end{aligned}$$

wobei ... bedeutet, dass ihr wie für $d = 1$ alle vorstellbaren Kombinationen von Intervallen nutzen könnt.

Beweis. Übungsaufgabe. □

Bemerkung 4.2.3. Wie in Dimension 1 gilt auch jetzt wieder, dass jede stetige Abbildung $f : \mathbb{R}^n \to \mathbb{R}^m$ $(\mathcal{B}(\mathbb{R}^n), \mathcal{B}(\mathbb{R}^m))$-messbar ist. Das gilt wieder weil wegen Proposition 2.1.4 Messbarkeit nur auf einem beliebigen Erzeuger getestet werden muss und für stetige Abbildungen Urbilder offener Mengen offen sind.

(B) Maße auf $\mathcal{B}(\mathbb{R}^d)$ und multivariate Verteilungsfunktionen

Wie für $d = 1$ definieren wir für Maße auf $\mathcal{B}(\mathbb{R}^d)$ Verteilungsfunktionen, der Unterschied ist nur die Anzahl der Variablen, d viele statt einer:

Definition 4.2.4. Ist $\mathbb{P}$ ein Wahrscheinlichkeitsmaß auf $\mathcal{B}(\mathbb{R}^d)$, so heißt

$$F_{\mathbb{P}}(t_1, ..., t_d) = \mathbb{P}((-\infty, t_1] \times \cdots \times (-\infty, t_d]), \quad t_1, ..., t_d \in \mathbb{R},$$

(multivariate) Verteilungsfunktion von $\mathbb{P}$.

Für die Vorstellung nehmen wir immer den Fall $d = 2$. Dann ist $F(t_1, t_2)$ gerade das Maß des „unendlichen Rechtecks unten links“ unter dem Punkt (t_1, t_2), also das Maß von $(\infty, t_1] \times (-\infty, t_2]$.
Wie für $d = 1$ (nichtfallend, rechtsstetig, Grenzwerte 1 und 0 bei unendlich) können wir aus den Eigenschaften des Maßes Eigenschaften der Verteilungsfunktion ableiten:

Proposition 4.2.5. Ist F die Verteilungsfunktion eines Wahrscheinlichkeitsmaßes auf $\mathcal{B}(\mathbb{R}^d)$, so gelten

(i) $F : \mathbb{R}^d \to [0, 1]$

(ii) F konvergiert gegen 0, wenn eine Koordinate nach $-\infty$ läuft:

$$\lim_{t_1\to-\infty} F(t_1,...,t_d) = ... = \lim_{t_d\to-\infty} F(t_1,...,t_d) = 0.$$

(iii) F konvergiert gegen 1, wenn alle Koordinaten gemeinsam nach $+\infty$ laufen:

$$\lim_{t_i\to-\infty, i=1,\dots,d} F(t_1,...,t_d) = 1.$$

(iv) F ist **rechtsstetig** in jeder Koordinate.

(v) F ist **rechtecksmonoton**, d. h. für alle $a^1, a^2 \in \mathbb{R}^d$ mit $a^1 \leq a^2$ (d. h. $a_1^1 \leq a_1^2, ..., a_d^1 \leq a_d^2$) gilt

$$\Delta_{a^1}^{a^2} F := \sum_{i_1,\dots,i_d\in\{1,2\}} (-1)^{i_1+\dots+i_d} F(a_1^{i_1}, ..., a_d^{i_d}) \geq 0.$$

Eine Funktion mit den Eigenschaften (i)-(v) nennt man (multivariate) Verteilungsfunktion.

Beweis. Sei $\mathbb{P}$ ein Wahrscheinlichkeitsmaß auf $\mathcal{B}(\mathbb{R}^d)$ und $F = F_{\mathbb{P}}$ die zugehörige Verteilungsfunktion. Die erste Eigenschaften von F ist klar, die weiteren drei Eigenschaften folgen aus der Stetigkeit von Maßen, genau wie für $d = 1$. Interessanter ist die Rechtecksmonotonie, die wir uns nur für $d = 1$ und $d = 2$ veranschaulichen.

$d = 1$: Einsetzen gibt hier $F(a^2) - F(a^1) \geq 0$ für $a^2 \geq a^1$, und das ist gerade die Monotonie, die wir für schon kennen aus der Diskussion von Verteilungsfunktionen in einer Variablen.

$d = 2$: Einsetzen in die Formel (es gibt $2^d = 4$ Summanden) ergibt

$$F(a_1^2, a_2^2) - F(a_1^2, a_2^1) - F(a_1^1, a_2^2) + F(a_1^1, a_2^1) \geq 0.$$

Doch was soll das bedeuten? Dazu ist zu beachten, dass die zwei Punkte $a^1 \leq a^2$ ein Rechteck R „aufspannen". Die Eckpunkte von R sind gerade (siehe Bildchen)

- (a_1^1, a_2^1), unten links
- (a_1^2, a_2^1), unten rechts
- (a_1^2, a_2^2), oben rechts
- (a_1^1, a_2^2), oben links

Weil $\mathbb{P}$ ein Maß ist, gilt $\mathbb{P}(R) \geq 0$. Jetzt schreiben wir durch Zerlegung von R in „unendliche Rechtecke unten links", unter Berücksichtigung der σ-Additivität von $\mathbb{P}$, $\mathbb{P}(R)$ als

$$\begin{aligned}\mathbb{P}(R) &= \mathbb{P}((-\infty, a_1^2] \times (-\infty, a_2^2]) - \mathbb{P}((-\infty, a_1^1] \times (-\infty, a_2^2]) \\ &\quad - \mathbb{P}((-\infty, a_1^2] \times (-\infty, a_2^1]) + \mathbb{P}((-\infty, a_1^1] \times (-\infty, a_2^1]) \\ &\overset{\text{Def. } F}{=} F(a_1^2, a_2^2) - F(a_1^1, a_2^2) - F(a_1^2, a_2^1) + F(a_1^1, a_2^1).\end{aligned}$$

Die Bedingung $\Delta_{a_1}^{a_2} F \geq 0$ gilt also weil $\Delta_{a_1}^{a_2} F$ nur ein komplizierter Ausdruck für die Wahrscheinlichkeit des von a^1 und a^2 aufgespannten Rechtecks ist! □

In Analogie zum eindimensionalen Fall fragen wir nun, ob die Umkehrung auch gilt. Gibt es also für jede multivariate Verteilungsfunktion F ein Wahrscheinlichkeitsmaß $\mathbb{P}_F$ auf $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, dessen Verteilungsfunktion F ist. In anderen Worten, gibt es eine bijektive Abbildung zwischen den multivariaten Verteilungsfunktionen und den Wahrscheinlichkeitsmaßen auf $\mathcal{B}(\mathbb{R}^d)$?

Satz 4.2.6. **[Analogie zu 1.4.2]** Für jede multivariate Verteilungsfunktion $F : \mathbb{R}^d \to \mathbb{R}$ gibt es genau ein Wahrscheinlichkeitsmaß $\mathbb{P}_F$ auf $\mathcal{B}(\mathbb{R}^d)$ mit Verteilungsfunktion F, d. h.

$$\mathbb{P}_F((-\infty, t_1]) \times ... \times (-\infty\, t_d]) = F(t_1, ..., t_d), \quad t_i \in \mathbb{R}. \tag{4.1}$$

Man sagt wieder „$\mathbb{P}$ ist gemäß F verteilt."

Beweis. Wir führen den Beweis nicht vollständig aus, die Argumente gehen im Prinzip wie für $d = 1$.

Eindeutigkeit: Wie immer nutzten wir für die Eindeutigkeit Dynkin-Systeme. Weil (4.1) das Maß auf $\cap$-stabilem Erzeuger festlegt, kann es aufgrund von Satz 1.2.12 nur ein Wahrscheinlichkeitsmaß mit der Eigenschaft (4.1) geben.

Existenz: Zur Konstruktion haben wir den Fortsetzungssatz von Carathéodory, Satz 1.3.7. Hier nur eine Skizze, die für das Verständnis der komischen Rechtecksmonotonie hilfreich ist, formuliert für $d = 2$. Zunächst muss eine σ-additive Mengenfunktion auf einem Erzeuger definiert werden. Dazu nehmen wir die Rechtecke der Form $(a_1^1, a_1^2] \times (a_2^1, a_2^2]$ und definieren

$$\mu((a_1^1, a_1^2] \times (a_2^1, a_2^2]) := \Delta_{a^1}^{a^2} F \geq 0.$$

Die Definition ist motiviert durch den Beweis von Proposition 4.2.5, $\Delta_{a^1}^{a^2} F$ war dort ja gerade die Wahrscheinlichkeit des Rechtecks $(a_1^1, a_1^2] \times (a_2^1, a_2^2]$. Nun muss man wie für $d = 1$ zeigen, dass μ eine σ-additive Mengenfunktion auf den Rechtecken ist. Das ist wieder etwas hässlich, funktioniert aber wie im Beweis von Satz 1.4.2. Hat man das geschafft, so existiert eine Fortsetzung von μ auf $\mathcal{B}(\mathbb{R}^d)$ und die tut es. □

Vorlesung 20

Proposition 4.2.7. **[Spezialfall Produktmaß]** Sind $F_1, ..., F_d$ reelle Verteilungsfunktionen, so ist

$$F(t_1, ..., t_d) := F_1(t_1) \cdot ... \cdot F_d(t_d), \quad t_i \in \mathbb{R},$$

eine multivariate Verteilungsfunktion. Es gilt: $\mathbb{P}_F = \mathbb{P}_{F_1} \otimes ... \otimes \mathbb{P}_{F_d}$.

Beweis. Variante 1: F ist eine multivariate Verteilungsfunktion $\rightsquigarrow$ Große Übung. Das ist ein gutes Beispiel, um die Eigenschaften mal selber nachzurechnen. Mit Satz 4.2.6 gibt es ein dazugehöriges Maß auf $\mathcal{B}(\mathbb{R}^d)$. Um zu checken, dass dieses Maß das Produktmaß ist, berechnet man mit der Formel für $\Delta_{a_1}^{a_2} F$ (Produktform von F einsetzen!) die Wahrscheinlichkeit von Quadern aus und findet gerade die benötigte Formel (3.8) aus der Definition des Produktmaßes.

Variante 2: Das Produktmaß $\mathbb{P}_{F_1} \otimes ... \otimes \mathbb{P}_{F_d}$ existiert auf $\mathcal{B}(\mathbb{R}^d)$ nach Korollar 3.5.3. Behauptung: F ist die (multivariate) Verteilungsfunktion von $\mathbb{P}_{F_1} \otimes ... \otimes \mathbb{P}_{F_d}$. Checken wir also, dass das Produktmaß die richtige Verteilungsfunktion hat:

$$\begin{aligned} \mathbb{P}_{F_1} \otimes ... \otimes \mathbb{P}_{F_d}((-\infty, t_1] \times ... \times (-\infty, t_d]) &\overset{\text{Def.}}{=} \mathbb{P}_{F_1}((-\infty, t_1]) \cdot ... \cdot \mathbb{P}_{F_d}((-\infty, t_d]) \\ &= F_1(t_1) \cdot ... \cdot F_d(t_d) \\ &= F(t_1, ..., t_d), \quad t_i \in \mathbb{R}. \end{aligned}$$

□

Genau wie für reellwertige Zufallsvariablen gibt es absolutstetige und diskrete Wahrscheinlichkeitsmaße auf $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$:

Definition 4.2.8.

(i) Eine multivariate Verteilungsfunktion F (bzw. das zugehörige Maß $\mathbb{P}_F$) heißt **absolutstetig** mit Dichte $f\colon \mathbb{R}^d \to [0, +\infty]$, falls f messbar ist und

$$F(t_1, ..., t_d) = \int_{(-\infty, t_1] \times ... \times (-\infty, t_d]} f(x) \, \mathrm{d}x, \quad t_i \in \mathbb{R}.$$

(ii) Ein multivariate Verteilungsfunktion F (bzw. das zugehörige Maß $\mathbb{P}_F$) heißt **diskret**, falls für ein $N \in \mathbb{N} \cup \{+\infty\}$ Vektoren $a_1, ..., a_N \in \mathbb{R}^d$ und Wahrscheinlichkeitsgewichte $p_1, ..., p_N \geq 0$ existieren, sodass

$$F(t_1, ..., t_d) = \sum_{k=1}^{N} p_k \mathbf{1}_{[a_{k,1}, \infty) \times ... \times [a_{k,d}, \infty)}(t_1, ..., t_d), \quad t_i \in \mathbb{R}.$$

Die Definition ist wie für Zufallsvariablen, nur aufgrund mehrerer Koordinaten unübersichtlicher. Im diskreten Fall merkt ihr euch wie im eindimensionalen Fall einfach, dass $\mathbb{P}_F$ Masse auf den Vektoren $a_1, ..., a_N$ hat und die Wahrscheinlichkeiten $p_1, ..., p_N$ sind.
Wenn wir besonders Wert auf die einzelnen Koordinaten legen wollen, schreiben wir das Integral auch als $\int_{(-\infty,t_1]\times...\times(-\infty,t_d]} f(x_1,...,x_d)\,\mathrm{d}(x_1,...,x_d)$. Wir meinen mit beiden Notationen immer das Lebesgue Integral bezüglich des d-dimensionalen Lebesgue-Maßes λ auf $\mathcal{B}(\mathbb{R}^d)$, also $\int_{(-\infty,t_1]\times...\times(-\infty,t_d]} f\,\mathrm{d}\lambda$. Aufgrund von Fubini gilt für absolutstetige Maße immer auch die Darstellung durch Mehrfachintegrale

$$F(t_1,...,t_d) = \int_{-\infty}^{t_1} ... \Big(\int_{-\infty}^{t_d} f(x_1,...,x_d)\,\mathrm{d}x_d \Big) ...\,\mathrm{d}x_1, \quad t_i \in \mathbb{R}$$

und das werden wir zum Rechnen auch meistens benutzen. Das einfachste Beispiel solche iterierten Integrale zu berechnen, tritt auf, wenn f faktorisiert, d. h. die einzelnen Koordinaten sich nur durch Produkte bedingen. In dem Fall wird die Verteilungsfunktion auch faktorisiert:

Beispiel 4.2.9. Sind $f_1, ..., f_d$ Dichten von reellen Verteilungsfunktionen $F_1, ..., F_d$. Dann ist $f(x) = f(x_1, ..., x_d) := f_1(x_1) \cdot ... \cdot f_d(x_d)$ eine Dichte von $F = F_1 \cdot ... \cdot F_d$ aus Proposition 4.2.7. Das können wir sofort mit Fubini zeigen:

$$\begin{aligned}
F(t_1,...,t_d) &\overset{\text{Def.}}{=} F_1(t_1)\cdot ... \cdot F_d(t_d)\\
&\overset{\text{Def.}}{=} \int_{-\infty}^{t_1} f_1(x_1)\,\mathrm{d}x_1 \cdot ... \cdot \int_{-\infty}^{t_d} f_d(x_d)\,\mathrm{d}x_d\\
&\overset{\text{Lin.}}{=} \int_{-\infty}^{t_1} ... \Big(\int_{-\infty}^{t_d} f_1(x_1)\cdot ... \cdot f_d(x_d)\,\mathrm{d}x_d \Big) ...\,\mathrm{d}x_1\\
&\overset{3.5.8}{=} \int_{(-\infty,t_1]\times...\times(-\infty,t_d]} f_1(x_1)\cdot ... \cdot f_d(x_d)\,\mathrm{d}(x_1,...,x_d)\\
&\overset{\text{Notation}}{=} \int_{(-\infty,t_1]\times...\times(-\infty,t_d]} f(x)\,\mathrm{d}x.
\end{aligned}$$

Natürlich müssen wir für Fubini die Messbarkeit von f zeigen. Zum Glück ist f das Produkt messbarer Funktionen und damit wieder messbar.

(C) Zufallsvektoren

Nachdem wir die Maße auf $\mathcal{B}(\mathbb{R}^d)$ verstanden haben, kommen wir jetzt analog zum reellen Fall zu den Zufallsvektoren.

Definition 4.2.10. Ist $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum, so heißt eine $(\mathcal{A}, \mathcal{B}(\mathbb{R}^d))$-messbar Abbildung $X : \Omega \to \mathbb{R}^d$ **Zufallsvektor**.

Proposition 4.2.11.

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_d \end{pmatrix} : \Omega \to \mathbb{R}^d$$

ist ein Zufallsvektor genau dann, wenn $X_1, ..., X_d : \Omega \to \mathbb{R}$ Zufallsvariablen sind.

Die Proposition ist eine reine Messbarkeitseigenschaft. Sie besagt nur, dass eine vektorwertige Abbildung messbar ist, genau dann, wenn jede Koordinatenabbildung messbar ist. Das ist ein wenig wie in Analysis 2, als immer $f : \mathbb{R}^n \to \mathbb{R}^m$ auf die Koordinatenabbildungen $f_i : \mathbb{R}^n \to \mathbb{R}$ reduziert wurde.

Beweis. „$\Rightarrow$“: Für $B \in \mathcal{B}(\mathbb{R})$ gilt

$$X_k^{-1}(B) = X^{-1}(\underbrace{\mathbb{R} \times ... \times \mathbb{R} \times B \times \mathbb{R} \times ... \times \mathbb{R}}_{k\text{-te Stelle}}) \in \mathcal{A}.$$

„$\Leftarrow$“: Messbarkeit muss nur auf einem Erzeuger gezeigt werden, wir wählen dazu $\mathcal{S} = \{B_1 \times ... \times B_d \colon B_i \in \mathcal{B}(\mathbb{R})\}$:

$$X^{-1}(B_1 \times ... \times B_d) = \left\{ \omega \in \Omega \colon \begin{pmatrix} X_1(\omega) \\ \vdots \\ X_d(\omega) \end{pmatrix} \in B_1 \times ... \times B_d \right\}$$
$$= \bigcap_{k=1}^{d} \{\omega \in \Omega \colon X_k(\omega) \in B\} \in \mathcal{A}$$

Damit ist X messbar.

$\square$

Diskussion 4.2.12. Wegen Proposition 4.2.11 gibt es jetzt zwei Interpretationen von Zufallsvektoren:

(i) Ein Zufallsvektor beschreibt d-viele Eigenschaften („Feature-Vektor“) **einer** zufälligen Beobachtung.

(ii) X beschreibt die Beobachtungen von **d-vielen** zufälligen eindimensionalen Experimenten.

Wir verbalisieren (i) und (ii) unterschiedlich auch wenn es sich mathematisch eigentlich um das gleiche Objekt handelt:

(i) „Sei $X = \begin{pmatrix} X_1 \\ \vdots \\ X_d \end{pmatrix}$ ein Zufallsvektor auf $(\Omega, \mathcal{A}, \mathbb{P})$.“

(ii) „Seien $X_1, ..., X_d$ Zufallsvariablen auf $(\Omega, \mathcal{A}, \mathbb{P})$.“

Weiter geht's mit der Verallgemeinerung von Verteilungsfunktionen von Zufallsvariablen auf Zufallsvektoren. Analog zum eindimensionalen Fall definieren wir die gleichen Begriffe:

Definition 4.2.13.

(i) Für einen Zufallsvektor X auf $(\Omega, \mathcal{A}, \mathbb{P})$ heißt

$$F_X(t_1, ..., t_d) := \mathbb{P}(X_1 \leq t_1, ..., X_d \leq t_d), \quad t_i \in \mathbb{R},$$

Verteilungsfunktion von X. Dabei steht das Komma für „und“ (also Durchschnitt), wir lesen also „Wahrscheinlichkeit, dass $X_1 \leq t_1$ und .... und $X_d \leq t_d$“, formell steht da aber der schlechter lesbare Ausdruck $\mathbb{P}(\cap_{k=1}^d \{X_k \leq t_k\})$ oder noch schlimmer $\mathbb{P}(\cap_{k=1}^d \{\omega \in \Omega : X_k(\omega) \leq t_k\})$. F heißt auch **gemeinsame Verteilungsfunktion** der Zufallsvariablen $X_1, ..., X_d$. Wir nutzen wieder die Schreibweise $X \sim F$, wenn F die Verteilungsfunktion von X ist.

(ii) Das Bildmaß

$$\mathbb{P}_X(B) := \mathbb{P}(X \in B), \quad B \in \mathcal{B}(\mathbb{R}^d),$$

heißt **Verteilung von** X oder die **gemeinsame Verteilung** der Zufallsvariablen $X_1, ..., X_d$. $\mathbb{P}_X$ ist ein Maß auf $\mathcal{B}(\mathbb{R}^d)$.

(iii) Zwei Zufallsvektoren X und Y heißen **identisch verteilt**, falls $F_X = F_Y$ gilt.

(iv) Für $k = 1, ..., d$ heißt

$$\mathbb{P}_{X_k}(B) = \mathbb{P}(X_k \in B) = \mathbb{P}(\{\omega \colon X_k(\omega) \in B\}), \quad B \in \mathcal{B}(\mathbb{R}),$$

die **(eindimensionale) Randverteilung** von X_k und

$$F_{X_k}(t) = \mathbb{P}(X_k \leq t), \quad t \in \mathbb{R},$$

die **(eindimensionale) Randverteilungsfunktion** von X_k.

Wir hatten im eindimensionalen Fall erst die Verteilung $\mathbb{P}_X$ und dann daraus die Verteilungsfunktion F_X definiert. Hier haben wir die Verteilungsfunktion direkt definiert und anschließend die Verteilung $\mathbb{P}_X$. Um mit den Definitionen zu spielen überlegt mal, warum wieder $\mathbb{P}_X = \mathbb{P}_{F_X}$ gilt. Wegen der Gleichheit ist es egal, ob wir die Begriffe wie hier oder wie in Definition 4.1.3 einführen.

Die Notationen werden hier etwas unübersichtlich, diskutieren wir sie also ein wenig.

Bemerkung 4.2.14. Weil $\mathbb{P}(X_i \leq t) = \mathbb{P}(X_1 \in \mathbb{R}, ..., X_i \leq t, ..., X_d \in \mathbb{R})$ gilt, folgt aus der Stetigkeit von Maßen

$$F_{X_i}(t_i) = \lim_{\substack{t_k \to \infty, \\ \forall k \neq i}} F(t_1, ..., t_d) =: F_X(+\infty, ..., t_i, ..., +\infty).$$

Mit dieser Formel ist klar, wie aus der gemeinsamen Verteilungsfunktion aller X_i die Verteilungsfunktion eines einzelnen X_i berechnet werden kann: Man schickt einfach in der gemeinsamen Verteilungsfunktion alle anderen Variablen t_k nach $+\infty$.

Analog zum eindimensionalen Fall jetzt auch noch die kanonische Konstruktion von Zufallsvektoren, die funktioniert fast wörtlich wie die Konstruktion im Beweis von Satz 4.1.7.

Satz 4.2.15. [**Kanonische Konstruktion von Zufallsvektoren**] Für jede multivariate Verteilungsfunktion F gibt es einen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ und einen Zufallsvektor $X : \Omega \to \mathbb{R}^d$ mit $X \sim F$.

Beweis. Als Wahrscheinlichkeitsraum definieren wir $\Omega = \mathbb{R}^d$, $\mathcal{A} = \mathcal{B}(\mathbb{R}^d)$, $\mathbb{P} = \mathbb{P}_F$ aus Satz 4.2.6 und darauf den Zufallsvektor $X(\omega) = \omega$, also $X_i(\omega) = \omega_i$. Beachte: Die Identitätsabbildung $X(\omega) = \omega$ ist eine stetige Abbildung von $\mathbb{R}^d$ nach $\mathbb{R}^d$ und damit auch messbar. Berechnen wir die Verteilungsfunktion dieses konkreten Zufallsvektors:

$$\begin{aligned}
\mathbb{P}(X_1 \leq t_1, ..., X_d \leq t_d) &= \mathbb{P}_F(\{\omega \in \mathbb{R}^d : X_1(\omega) \leq t_1, ..., X_d(\omega) \leq t_d\}) \\
&= \mathbb{P}_F(\{\omega \in \mathbb{R}^d : \omega_1 \leq t_1, ..., \omega_d \leq t_d\}) \\
&= \mathbb{P}_F((-\infty, t_1] \times ... \times (-\infty, t_d]) \\
&= F(t_1, ..., t_d), \quad t_i \in \mathbb{R}.
\end{aligned}$$

Das war es schon! Zu beachten ist, dass die Konstruktion weit von trivial ist. Die Existenz von $\mathbb{P}_F$ benötigt den Satz von Carathéodory und damit die komplette Maßtheorie. □

Ab jetzt werden wir immer die Interpretation eines Zufallsvektors als d-viele Zufallsvariablen nutzen, damit wir uns langsam an Folgen von Zufallsvariablen gewöhnen.

Definition 4.2.16. Seien $X_1, ..., X_d$ Zufallsvariablen auf $(\Omega, \mathcal{A}, \mathbb{P})$.

(i) $X_1, ..., X_d$ haben die **gemeinsame Dichte** f, falls die gemeinsame Verteilungsfunktion F_X absolutstetig ist und Dichte f hat.

(ii) $X_1, ..., X_d$ heißen **diskret**, falls die gemeinsame Verteilungsfunktion F_X diskret ist.

Der diskrete Fall ist viel einfacher, als es aussieht. Die Definition bedeutet einfach nur, dass der Vektor X abzählbar viele Werte $a_1, ..., a_N \in \mathbb{R}^d$ mit Wahrscheinlichkeiten $p_1, ..., p_N$ annimmt. Oder, anders ausgedrückt, dass alle Zufallsvariablen $X_1, ..., X_d$ nur abzählbar viele Werte annehmen.

Aus der Definition folgt sofort, dass eine gemeinsame Dichte nicht-negativ und messbar ist, sowie $\int_{\mathbb{R}^d} f(x)\,\mathrm{d}x = 1$ erfüllt. Andersrum zeigt ihr in den Übungsaufgaben, dass für solch eine Funktion $F(t_1, .., , t_d) := \int_{(-\infty,t_1]\times...\times(-\infty,t_d]} f(x)\,\mathrm{d}x$ die Eigenschaften einer multivariaten Verteilungsfunktion erfüllt.

Proposition 4.2.17. Seien $X_1, ..., X_d$ Zufallsvariablen auf $(\Omega, \mathcal{A}, \mathbb{P})$.

(i) Haben $X_1, ..., X_d$ die gemeinsame Dichte f, so haben $X_1, ..., X_d$ Dichten $f_1, ..., f_d$ und es gilt

$$f_i(x_i) = \underbrace{\int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty}}_{(d-1)\text{-viele}} \underbrace{f(x_1, ..., x_d)}_{x_i \text{ fest}} \underbrace{\mathrm{d}x_1 ... \mathrm{d}x_d}_{\text{ohne } x_i}, \quad x_i \in \mathbb{R},$$

ist eine Dichte von X_i für $i = 1, ..., d$. In Worten: Ist X absolutstetig, so sind alle X_i absolutstetig und die Dichten der X_i entstehen durch Ausintegrieren aller anderen Variablen.

(ii) Die Rückrichtung gilt im Allgemeinen nicht. Es gibt also absolutstetige Zufallsvariablen, die keine gemeinsame Dichte haben.

Beweis. (i) Rechnen wir die Verteilungsfunktion von X_i aus:

$$\begin{aligned}
F_{X_i}(t_i) &\overset{\text{Def.}}{=} \mathbb{P}(X_i \leq t_i) \\
&\overset{\text{Trick}}{=} \mathbb{P}(X_1 \in \mathbb{R}, ..., X_i \leq t_i, ..., X_d \in \mathbb{R}) \\
&\overset{\text{Stet. Maße}}{=} \lim_{\substack{t_k \to \infty, \\ k \neq i}} \mathbb{P}(X_1 \leq t_1, ..., X_d \leq t_d) \\
&\overset{\text{Dichte}}{=} \lim_{\substack{t_k \to \infty, \\ k \neq i}} \underbrace{\int_{-\infty}^{t_1} ... \int_{-\infty}^{t_d}}_{d\text{-mal}} f(x_1, ..., x_d) \, \mathrm{d}x_d ... \mathrm{d}x_1 \\
&\overset{3.2.1}{=} \int_{-\infty}^{t_i} f_i(x_i) \, \mathrm{d}x_i.
\end{aligned}$$

Im letzten Gleichheitszeichen haben wir fröhlich die Reihenfolge der iterierten Integrale getauscht, das war natürlich der Satz von Fubini.

(ii) Als Beispiel kann man $X_1 \sim \mathcal{U}([0,1])$ und $X_2 = X_1$ betrachten. Der Vektor (X_1, X_2) nimmt nur Werte in $A = \{(x_1, x_2) \in [0,1] \times [0,1] : x_1 = x_2\}$ an und das ist eine Lebesgue-Nullmenge. In den Übungen sollt ihr euch überlegen, dass in so einem Fall keine Dichte existieren kann. □

Die Situation ist einfacher für diskrete Zufallsvariablen. Wir sehen hier sehr deutlich, warum diskrete Stochastik so viel einfacher ist.

Proposition 4.2.18. Seien $X_1, ..., X_d$ Zufallsvariablen auf $(\Omega, \mathcal{A}, \mathbb{P})$, dann gilt:

$$X \text{ ist ein diskreter Zufallsvektor} \quad \Longleftrightarrow \quad X_1, ..., X_d \text{ sind diskrete Zufallsvariablen}$$

Beweis. „$\Rightarrow$“: Wenn der Vektor nur abzählbar viele Werte im $\mathbb{R}^d$ annimmt, kann natürlich auch jeder Eintrag nur abzählbar viele Werte (die Koordinaten der abzählbar vielen Werte) annehmen. Damit sind die Koordinaten $X_1, ..., X_d$ diskrete Zufallsvariablen. Wie im absolutstetigen Fall ergeben sich die Wahrscheinlichkeiten durch Ausintegrieren, was hier Aussummieren bedeutet. Wenn X die Vektoren $a_1, ..., a_N$ annimmt, so gilt

$$\mathbb{P}(X_i = a_{k,i}) = \underbrace{\sum_{k_1=1}^{N} ... \sum_{k_d=1}^{N}}_{(d-1)\text{-viele}} \mathbb{P}(X_1 = a_{k_1,1}, ..., X_i = a_{k,i}, ..., X_d = a_{k_d,d}),$$

wobei nur $(d-1)$-viele Koordinaten aussummiert werden.
„$\Leftarrow$“: Wenn die Zufallsvariablen X_i die reellen Werte $a_{1,i}, ..., a_{N_i,i}$ annehmen, so kann der Zufallsvektor $X = (X_1, ..., X_d)$ als Werte nur die $N_1 \cdot ... \cdot N_d$ Vektoren annehmen, die sich aus den Kombinationen der Werte ergeben. Die $N_1, ..., N_d$ können endlich oder unendlich sein. Sind alle endlich, so nimmt X nur endlich viele Werte an, andernfalls abzählbar unendlich viele Werte. In beiden Fällen ist X ein diskreter Zufallsvektor. □

Vorlesung 21

Bisher ist in diesem Kapitel kaum neues passiert. Nur die Rechtecksmonotonie einer multivariaten Verteilungsfunktion ist als neue Idee hinzugekommen. Das ändert sich jetzt allerdings mit dem Konzept der Unabhängigkeit. Euch ist sicher intuitiv von irgendwo bekannt, was zum Beispiel die Unabhängigkeit von drei Würfelwürfeln sein soll, insbesondere werdet ihr alle sofort sagen, dass Wahrscheinlichkeiten durch Produktbildung von Wahrscheinlichkeiten entstehen. Genau das machen wir jetzt mathematisch präzise:

Definition 4.2.19. Seien $X_1, ..., X_d$ Zufallsvariablen auf $(\Omega, \mathcal{A}, \mathbb{P})$.

(i) $X_1, ..., X_d$ heißen **unabhängig**, falls die gemeinsame Verteilungsfunktion in die Randverteilungsfunktionen faktorisiert, d. h.

$$F_X(t_1, ..., t_d) = F_{X_1}(t_1) \cdot ... \cdot F_{X_d}(t_d), \quad t_i \in \mathbb{R}$$

oder mit Wahrscheinlichkeiten geschrieben

$$\mathbb{P}(X_1 \leq t_1, ..., X_d \leq t_d) = \mathbb{P}(X_1 \leq t_1) \cdot ... \cdot \mathbb{P}(X_d \leq t_d), \quad t_i \in \mathbb{R}.$$

(ii) $X_1, ..., X_d$ heißen **abhängig**, falls sie nicht unabhängig sind.

(iii) $X_1, ..., X_d$ heißen **unabhängig und identisch verteilt** (**u.i.v.**), falls sie unabhängig und identisch verteilt ($F_{X_1} = ... = F_{X_d}$) sind. Weil die gemeinsame Verteilungsfunktion F bei u.i.v. Zufallsvariablen schon eindeutig durch jede Randverteilungsfunktion festgelegt ist, gibt man oft nur die Verteilung von X_1 an.

Was soll das abstrakte Konzept der Unabhängigkeit eigentlich bedeuten? Unabhängigkeit ist die mathematische Formulierung der Idee, dass der Wert von einer Zufallsvariablen keinen Einfluss auf den Wert der anderen Zufallsvariablen hat. Die Temperaturen in Heidelberg und Mannheim morgen um 12 Uhr sind vermutlich nicht unabhängig (ist es in Heidelberg kalt, so ist es vermutlich auch in Mannheim kalt). Andererseits hat die Temperatur morgen in Peking vermutlich keinen Einfluss darauf, wie groß der Kaffeefleck auf meiner Hose übermorgen ist.

Klingt vielleicht blöd, aber warum gibt es überhaupt u.i.v Zufallsvariablen? Natürlich wegen des Produktmaßes!

Satz 4.2.20. Nehmt eure Lieblingsverteilungsfunktion F, so gibt es u.i.v $X_1, ..., X_d$ auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ mit $X_1 \sim F$. Es gilt $\mathbb{P}_X = \mathbb{P}_F \otimes ... \otimes \mathbb{P}_F$.

Beweis. Zunächst sieht man sofort (oder nutzt 4.2.7), dass $\bar{F}(t_1, ..., t_d) = F(t_1) \cdot ... \cdot F(t_d)$ für $t_i \in \mathbb{R}$ eine Verteilungsfunktion ist. Mit der kanonischen Konstruktion 4.2.15 gibt es also Zufallsvariablen $X_1, ..., X_d$, deren gemeinsame Verteilungsfunktion $\bar{F}$ ist und deren gemeinsame Verteilung ist nach Proposition 4.2.7 gerade das Produktmaß. Schauen wir uns nun die Randverteilung der X_i an. Wegen Bemerkung 4.2.14 gilt also

$$\bar{F}_{X_i}(t_i) = \lim_{\substack{t_k \to \infty, \\ \forall k \neq i}} F(t_1) \cdot ... \cdot F(t_d) = F(t_i), \quad t_i \in \mathbb{R},$$

weil für Verteilungsfunktionen $\lim_{t \to +\infty} F(t) = 1$ gilt. Damit sind $X_1, ..., X_d$ also identisch verteilt mit $X_1 \sim F$. Die Unabhängigkeit folgt damit direkt aus der Definition von $\bar{F}$:

$$F_X(t_1, ..., t_d) = \bar{F}(t_1, ..., t_d) = F(t_1) \cdot ... \cdot F(t_d) = F_{X_1}(t_1) \cdot ... \cdot F_{X_d}(t_d) \quad t_i \in \mathbb{R}.$$

□

Beispiel 4.2.21. Sei $X_1 \sim \mathcal{N}(0, 1)$ und $X_2 := -X_1$. Dann sind X_1, X_2 identisch verteilt, jedoch nicht unabhängig. Bestimmen wir dazu zunächst die Verteilungsfunktionen:

$$F_{X_1}(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \, dx$$

und

$$F_{X_2}(t) = \mathbb{P}(-X_1 \leq t) = \int_{-t}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \,\mathrm{d}x \overset{\text{subst.}}{=} \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \,\mathrm{d}x.$$

Die letzte Gleichheit gilt natürlich weil $\int_{-\infty}^{t} f(x)\,\mathrm{d}x = \int_{-t}^{+\infty} f(x)\,\mathrm{d}x$ für jede symmetrische integrierbare Funktion gilt. Also gilt $X_1 \sim \mathcal{N}(0,1)$, $X_2 \sim \mathcal{N}(0,1)$ und damit sind X_1, X_2 identisch verteilt. Um zu zeigen, dass sie nicht unabhängig sind, berechnen wir die gemeinsame Verteilungsfunktion an einer Stelle und zeigen, dass diese nicht faktorisiert. Es gelten

$$F_X(0,0) = \mathbb{P}(X_1 \leq 0, X_2 \leq 0) = \mathbb{P}(X_1 \leq 0, X_1 \geq 0) = \mathbb{P}(X_1 = 0) \overset{\text{abs. st.}}{=} 0$$

und

$$F_{X_1}(0) = F_{X_2}(0) = \mathbb{P}(X_1 \leq 0) = \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \,\mathrm{d}x = \frac{1}{2},$$

also gilt

$$F_X(0,0) = 0 \neq \frac{1}{4} = F_{X_1}(0) \cdot F_{X_2}(0)$$

und damit sind X_1, X_2 abhängig. Natürlich war intuitiv sowieso klar, dass X_1, X_2 nicht unabhängig sind. Unabhängig bedeutet schließlich, dass X_1 keinen Einfluss auf X_2 hat. Bei der Beziehung $X_1 = -X_2$ haben wir natürlich eine extreme Abhängigkeit: Kennen wir den Wert von X_1, so kennen wir auch den Wert von X_2.

Um mit gemeinsamen Verteilungen rumzurechnen, ist das nächste Korollar nützlich.

Korollar 4.2.22. Sind $X_1, ..., X_d$ Zufallsvariablen mit gemeinsamer Dichte f, dann gilt:

$$X_1, ..., X_d \text{ sind unabhängig} \quad \Leftrightarrow \quad f(x) = f_1(x_1) \cdot ... \cdot f_d(x_d) \quad \text{Lebesgue-fast überall},$$

wobei $f_1, ..., f_d$ Dichten von $X_1, ..., X_d$ sind.

Beweis. Zuerst erinnern wir daran, dass die Existenz der gemeinsamen Dichte die Absolutstetigkeit der einzelnen Zufallsvariablen impliziert (nicht andersrum).

„$\Leftarrow$“: Um die Unabhängigkeit zu prüfen, rechnen wir die Verteilungsfunktion aus und zeigen dabei, dass sie faktorisiert:

$$\begin{aligned} F_X(t_1, ..., t_d) &\overset{\text{Annahme}}{=} \int_{-\infty}^{t_1} ... \int_{-\infty}^{t_d} f_1(x_1) \cdot ... \cdot f_d(x_d) \,\mathrm{d}x_d ... \mathrm{d}x_1 \\ &\overset{\text{Lin.}}{=} \int_{-\infty}^{t_1} f_1(x_1)\,\mathrm{d}x_1 \cdot ... \cdot \int_{-\infty}^{t_d} f_d(x_d)\,\mathrm{d}x_d \\ &\overset{\text{Def.}}{=} F_{X_1}(t_1) \cdot ... \cdot F_{X_d}(t_d), \quad t_i \in \mathbb{R}. \end{aligned}$$

Also sind $X_1, ..., X_d$ nach Definition unabhängig.

„$\Rightarrow$“: Rechnen wir andersrum mit der gemeinsamen Verteilungsfunktion los:

$$\begin{aligned} \int_{(-\infty,t_1] \times ... \times (-\infty,t_d]} f(x)\,\mathrm{d}x &\overset{\text{Dichte}}{=} F_X(t_1, ..., t_d) \\ &\overset{\text{Ann.}}{=} F_{X_1}(t_1) \cdot ... \cdot F_{X_d}(t_d) \\ &= \int_{-\infty}^{t_1} f_1(x_1)\,\mathrm{d}x_1 \cdot ... \cdot \int_{-\infty}^{t_d} f_d(x_d)\,\mathrm{d}x_d \\ &\overset{\text{Lin.}}{=} \int_{-\infty}^{t_1} ... \int_{-\infty}^{t_d} f_1(x_1) \cdot ... \cdot f_d(x_d)\,\mathrm{d}x_d ... \mathrm{d}x_1 \\ &\overset{\text{Fubini}}{=} \int_{(-\infty,t_1] \times ... \times (-\infty,t_d]} f_1(x_1) \cdot ... \cdot f_d(x_d)\,\mathrm{d}(x_1, ..., x_d). \end{aligned}$$

Beachtet wieder, dass wir sowohl mit $\mathrm{d}x$ als auch mit $\mathrm{d}(x_1, ..., dx_d)$ das Lebesguemaß meinen. Wir haben also gezeigt, dass sowohl f als auch das Produkt der f_i Dichten von F sind.

Es fehlt jetzt noch die Aussage, dass zwei Dichten automatisch fast überall gleich sind. Das ist ein bisschen hübsche Integrationstheorie. Hier ist ein kleiner Hinweis, die Details können diejenigen zusammen basteln, die aktuell noch genug Kraft übrig haben.

- Wir haben schon gesehen, dass $\nu(A) = \int_A f(x_1, ..., x_d) d(x_1, ..., x_d)$ und $\mu(A) = \int_A f_1(x_1) \cdot ... \cdot f_d(x_d) d(x_1, ..., x_d)$ Wahrscheinlichkeitsmaße auf $\mathcal{B}(\mathbb{R}^d)$ sind. Das Stichwort ist MCT.
- Nach obiger Rechnung gilt $\nu(A) = \mu(A)$ für die Rechteckmengen. Weil die Rechteckmengen ein schnittstabiler Erzeuger sind, gilt also aufgrund des Eindeutigkeitssatzes $\mu = \nu$ auf $\mathcal{B}(\mathbb{R}^d)$.
- Wir wissen aus einer Übungsaufgabe, dass für Integrale auch strikte Monotonie gilt: Wenn $h < g$ fast überall gilt, so ist das Integral über h auch strikt kleiner als das Integral über g.
- Durch die Wahl der messbaren Mengen $A := \{f \neq g\} = \{f > g\} \cup \{f < g\} =: A_1 \cup A_2$ können wir aus den letzten zwei Schritten direkt die Aussage bekommen. Warum?

□

Wir kennen jetzt Zufallsvektoren und deren Verteilungen, fehlen noch Erwartungswerte von Zufallsvektoren.

(D) Erwartungswerte

Die Definition ist analog zu der Definition für eine Zufallsvariable:

Definition 4.2.23. Seien $X_1, ...X_d$ Zufallsvariablen auf $(\Omega, \mathcal{A}, \mathbb{P})$ und $g \colon \mathbb{R}^d \to \overline{\mathbb{R}}$ $(\mathcal{A}, \mathcal{B}(\overline{\mathbb{R}}))$-messbar. Dann sei

$$\mathbb{E}[g(X_1, ..., X_d)] := \int_\Omega g(X_1(\omega), ..., X_d(\omega))\, \mathrm{d}\mathbb{P}(\omega),$$

falls das Integral wohldefiniert ist. Wir sprechen von $\mathbb{E}[g(X_1, ..., X_d)]$ als Erwartungswert, weil $Y := g(X_1, ..., X_d)$ eine Zufallsvariable ist (zumindest wenn g endlich ist).

Die Berechnungstheorie geht jetzt komplett analog zu dem Fall einer Zufallsvariablen. Erst der Trafosatz, dann die Rechenregeln für absolutstetige und diskrete Zufallsvektoren.

Lemma 4.2.24. Seien $X_1, ...X_d$ Zufallsvariablen auf $(\Omega, \mathcal{A}, \mathbb{P})$ und sei $\mathbb{P}_X$ die gemeinsame Verteilung von $X = (X_1, ..., X_d)$. Dann gilt

$$\mathbb{E}[g(X_1, ..., X_d)] = \int_{\mathbb{R}^d} g(x_1, ..., x_d)\, \mathrm{d}\mathbb{P}_X(x_1, ..., x_d),$$

wobei eine Seite wohldefiniert ist, wenn es die andere Seite ist.

Beweis. Das ist nichts anderes als der Trafosatz, genau wie in Lemma 4.1.10:

$$\begin{tikzcd}
(\Omega, \mathcal{A}, \mathbb{P}) \arrow[r, "X"] \arrow[dr, "g \circ X"'] & (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \mathbb{P}_X) \arrow[d, "g"] \\
& (\overline{\mathbb{R}}, \mathcal{B}(\overline{\mathbb{R}}))
\end{tikzcd}$$

□

Wie für eine Zufallsvariable in Satz 4.1.11 kommen nun Rechenregeln für Wahrscheinlichkeiten und Integrale. Zusätzlich zu den diskreten und absolutstetigen Fällen gibt es jetzt auch noch Regeln für unabhängige Zufallsvariablen.

Satz 4.2.25. [**Berechnungsregeln, allgemeiner Fall**] Sind $X_1, ..., X_d$ Zufallsvariablen auf $(\Omega, \mathcal{A}, \mathbb{P})$, so gelten:

(i) Für $A \in \mathcal{B}(\mathbb{R}^d)$ gilt $\mathbb{E}[\mathbf{1}_A(X)] = \mathbb{P}(X \in A)$.

(ii) Haben $X_1, ..., X_d$ eine gemeinsame Dichte f, so gilt

$$\mathbb{E}[g(X_1, ..., X_d)] = \int_{\mathbb{R}^d} g(x_1, ..., x_d) f(x_1, ..., x_d) \, \mathrm{d}(x_1, ..., x_d).$$

(iii) Sind $X_1, ..., X_d$ diskret und nimmt der Zufallsvektor $X = (X_1, ..., X_d)$ die Vektoren $a_1, ..., a_N \in \mathbb{R}^d$ mit Wahrscheinlichkeiten $p_1, ..., p_N$ an, so gilt

$$\mathbb{E}[g(X_1, ..., X_d)] = \sum_{k=1}^{N} p_k g(a_k) = \sum_{k=1}^{N} \mathbb{P}(X = a_k) g(a_k).$$

Wenn N endlich ist, bekommen wir endliche Summen (einfach!), für $N = +\infty$ unendliche Reihen.

Wie für $d = 1$ gilt in (ii) und (iii), dass die Erwartungswerte wohldefiniert sind (oder existieren), genau dann, wenn die Integrale bzw. Summen wohldefiniert (oder endlich) sind.

Beweis. (i) Wir müssen dazu nur die Transformationsformel und die Definition der gemeinsam Verteilung $\mathbb{P}_X$ nutzen:

$$\mathbb{E}[\mathbf{1}_A(X)] = \int_{\mathbb{R}^d} \mathbf{1}_A(x_1, ..., x_d) \mathbb{P}_X(x_1, ..., x_d) = \mathbb{P}_X(A) = \mathbb{P}(X \in A).$$

(ii) und (iii) beweisen wir nicht. Der Beweis ist etwas länglich, aber exakt wie im eindimensionalen Fall bewiesen, siehe Beweis von Satz 4.1.11. □

Nach diesen allgemeinen Regeln, wenden wir uns jetzt konkret dem Fall von unabhängigen Zufallsvariablen zu. Hier werden viele Rechnungen einfacher weil Dichten und Wahrscheinlichkeiten faktorisieren.

Satz 4.2.26. Sind $X_1, ..., X_d$ unabhängige Zufallsvariablen auf $(\Omega, \mathcal{A}, \mathbb{P})$, so gilt

$$\mathbb{E}[g_1(X_1) \cdot ... \cdot g_d(X_d)] = \mathbb{E}[g_1(X_1)] \cdot ... \cdot \mathbb{E}[g_d(X_d)]$$

für alle messbaren $g_1, ..., g_d : \mathbb{R} \to \overline{\mathbb{R}}$. Insbesondere gilt auch

$$\mathbb{P}(X_1 \in A_1, ..., X_d \in A_d) = \mathbb{E}[\mathbf{1}_{A_1}(X_1)] \cdot ... \cdot \mathbb{E}[\mathbf{1}_{A_d}(X_d)] = \mathbb{P}(X_1 \in A_1) \cdot ... \cdot \mathbb{P}(X_d \in A_d)$$

für alle $A_1, ..., A_d \in \mathcal{B}(\mathbb{R})$.

Beweis. Wir schreiben den Beweis nur für $d = 2$, sonst wird die Notation zu hässlich. Wir wissen schon, dass Unabhängigkeit gerade bedeutet, dass die gemeinsame Verteilung ein Produktmaß der Randverteilungen ist, d. h. $\mathbb{P}_X = \mathbb{P}_{X_1} \otimes \mathbb{P}_{X_2}$. Berechnen wir damit den Erwartungswert mit dem Trafosatz und der Funktion $g(x_1, x_2) := g_1(x_1)g_2(x_2)$:

$$\begin{aligned}
\mathbb{E}[g_1(X_1) \cdot g_2(X_2)] &= \mathbb{E}[g(X_1, X_2)] \\
&\overset{4.2.24}{=} \int_{\mathbb{R}^2} g(x_1, x_2) \,\mathrm{d}\mathbb{P}_{(X_1,X_2)}(x_1, x_2) \\
&= \int_{\mathbb{R}^2} g(x_1, x_2) \,\mathrm{d}\mathbb{P}_{X_1} \otimes \mathbb{P}_{X_2}(x_1, x_2) \\
&\overset{\text{Fubini}}{=} \int_{\mathbb{R}} \Big(\int_{\mathbb{R}} g_1(x_1)g_2(x_2) \,\mathrm{d}\mathbb{P}_{x_1}(x_1) \Big) \,\mathrm{d}\mathbb{P}_{x_2}(x_2) \\
&\overset{\text{Lin.}}{=} \int_{\mathbb{R}} g_1(x_1) \,\mathrm{d}\mathbb{P}_{X_1}(x_1) \int_{\mathbb{R}} g_2(x_2) \,\mathrm{d}\mathbb{P}_{X_2}(x_2) \\
&\overset{2\text{x } 4.1.10}{=} \mathbb{E}[g_1(X_1)] \cdot \mathbb{E}[g_2(X_2)].
\end{aligned}$$

Die zweite Aussage folgt aus der ersten mit den messbaren Abbildungen $g_1 = \mathbf{1}_{A_1}, ..., g_d = \mathbf{1}_{A_d}$ sowie die wichtige Verbindung von Wahrscheinlichkeiten und Erwartungswerten aus Satz 4.2.25 mit der Menge $A = A_1 \times ... \times A_d$ beziehungsweise Satz 4.1.14 (iv). □

Der diskrete Fall (z. B. drei Mal Würfeln) geht in der allgemeinen Theorie leicht unter, daher schreiben wir es sicherheitshalber explizit hin:

Bemerkung. Sind $X_1, ..., X_d$ diskret und unabhängig, so gilt

$$\mathbb{P}(X_1 = a_1, ..., X_d = a_d) = \mathbb{P}(X_1 = a_1) \cdot ... \cdot \mathbb{P}(X_d = a_d).$$

Das ist einfach nur der vorherige Satz mit den Einpunktmengen $A_i = \{a_{k,i}\}$. Wenn wir also zweimal Würfeln, stände da zum Beispiel

$$\mathbb{P}(X_1 = 3, X_2 = 2) = \mathbb{P}(X_1 = 3)\mathbb{P}(X_2 = 2) = \frac{1}{6}\frac{1}{6} = \frac{1}{36}.$$

Zum Abschluss können wir die Faktorisierungen von Dichten und Wahrscheinlichkeiten noch in die Berechnungsregeln einsetzten, um einfachere Formen im Fall unabhängiger Zufallsvariablen zu bekommen. Im Prinzip ist der Satz schon in obigen Formeln enthalten, wir wollen ihn aber nochmal deutlich hinschreiben. Diese Formeln werden im nächsten Abschnitt immer wieder zum konkreten Rechnen mit Zufallsvariablen benutzt.

Satz 4.2.27. **[Berechnungsregeln, unabhängiger Fall]** Sind $X_1, ..., X_d$ unabhängige Zufallsvariablen auf $(\Omega, \mathcal{A}, \mathbb{P})$, so gelten:

(i) Haben $X_1, ..., X_d$ Dichten $f_1, .., f_d$, so gilt

$$\mathbb{E}[g(X_1, ..., X_d)] = \int_{\mathbb{R}^d} g(x_1, ..., x_d) f_1(x_1) \cdot ... \cdot f_d(x_d) \,\mathrm{d}(x_1, ..., x_d).$$

(ii) Sind $X_1, ..., X_d$ diskret und nimmt der Zufallsvektor $X = (X_1, ..., X_d)$ die Vektoren $a_1, ..., a_N \in \mathbb{R}^d$ an, so gilt

$$\mathbb{E}[g(X_1, ..., X_d)] = \sum_{k=1}^{N} \mathbb{P}(X_1 = a_{k,1}) \cdot ... \cdot \mathbb{P}(X_d = a_{k,d}) g(a_{k,1}, ..., a_{k,d}).$$

Wie für $d = 1$ gilt in (i) und (iii), dass die Erwartungswerte wohldefiniert sind (oder existieren), genau dann, wenn die Integrale bzw. Summen wohldefiniert (oder endlich) sind.

Eine direkte Folgerung aus obigen Regeln ist die Folgerung, dass Unabhängigkeit erhalten bleibt, wenn messbare Abbildungen angewandt werden.

Korollar 4.2.28. Sind $X_1, ...X_d$ unabhängige Zufallsvariablen auf $(\Omega, \mathcal{A}, \mathbb{P})$ und $f_1, ..., f_d : \mathbb{R} \to \mathbb{R}$ messbar. Dann sind auch $f_1(X_1), ..., f_d(X_d)$ unabhängige Zufallsvariablen.

Beweis. Zunächst sind die $f_i(X_i)$ auch Zufallsvariablen weil die Verknüpfung messbarer Abbildungen wieder messbar ist. Wir schreiben den Beweis nur für $d = 2$, sonst wird die Notation zu hässlich. Mit den vorherigen Sätzen (geht die Rechnung durch und sucht nach den passenden Sätzen als Wiederholung!) folgt

$$\begin{aligned}
\mathbb{P}(f_1(X_1) \leq t_1, f_2(X_2) \leq t_2) &= \mathbb{P}(X_1 \in f_1^{-1}((-\infty, t_1]), X_2 \in f_2^{-1}((-\infty, t_2])) \\
&= \mathbb{E}[\mathbf{1}_{f_1^{-1}((-\infty,t_1])}(X_1)\mathbf{1}_{f_2^{-1}((-\infty,t_2])}(X_2)] \\
&= \mathbb{E}[\mathbf{1}_{f_1^{-1}((-\infty,t_1])}(X_1)]\mathbb{E}[\mathbf{1}_{f_2^{-1}((-\infty,t_2])}(X_2)] \\
&= \mathbb{P}(X_1 \in f_1^{-1}((-\infty, t_1]))\mathbb{P}(X_2 \in f_2^{-1}((-\infty, t_2])) \\
&= \mathbb{P}(f_1(X_1) \leq t_1)\mathbb{P}(f_2(X_2) \leq t_2).
\end{aligned}$$

Also sind $f_1(X_1)$ und $f_2(X_2)$ unabhängig. □

Wir schließen unsere Diskussion von mehreren Zufallsvariablen mit einer wichtigen Kenngröße, die in einem gewissen Sinne die Abhängigkeit zweier Zufallsvariablen misst. In der Wahrscheinlichkeitstheorie spielen die Begriffe keine sehr große Rolle, in der Statistik jedoch eine sehr große.

Definition 4.2.29.

(i) Sind X, Y quadratintegrierbare Zufallsvariablen auf $(\Omega, \mathcal{A}, \mathbb{P})$, d. h. $\mathbb{E}[X^2], \mathbb{E}[Y^2] < \infty$, dann heißt
$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$
Kovarianz von X und Y.

(ii) Sind $\mathbb{V}[X], \mathbb{V}[Y] \neq 0$, das bedeutet X und Y sind nicht fast sicher konstant, so heißt
$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathbb{V}[X]\mathbb{V}[Y]}}$$
Korrelation von X, Y.

(iii) Ist $\rho(X, Y) = 0$, so heißen X, Y **unkorreliert**.

In der großen Übung wurden folgende Eigenschaften diskutiert:

Bemerkung 4.2.30.

- Fall X und Y endliche zweite Momente haben, so existiert die Kovarianz und es gilt $\rho(X, Y) \in [-1, 1]$. Das ist einfach nur Cauchy-Schwarz (d. h. Hölder mit $p = q = 2$):
$$\begin{aligned}
\mathrm{Cov}(X, Y) &= \mathbb{E}[(X - \mathbb{E}[X])(Y - E[Y])] \\
&\leq \sqrt{\mathbb{E}[(X - \mathbb{E}[X])^2]}\sqrt{\mathbb{E}[(Y - E[Y])^2]} = \sqrt{\mathbb{V}[X]\mathbb{V}[Y]}.
\end{aligned}$$
Durchteilen gibt $\rho(X, Y) \in [-1, 1]$.
- Sind X und Y unabhängig, so gilt $\mathrm{Cov}(X, Y) = \rho(X, Y) = 0$. Unabhängige Zufallsvariablen sind also auch unkorelliert! Das folgt sofort aus Satz 4.2.26 angewandt auf die Funktionen $g_1(x) = x - \mathbb{E}[X]$ und $g_2(y) = y - \mathbb{E}[Y]$ und Ausmultiplizieren.
- Die Korrelation wird in der Statistik genutzt, um Abhängigkeiten zu beschreiben. Positive Korrelation bedeutet, dass X und Y eher gleiches Vorzeichen haben, negative Korrelation bedeutet, dass X und Y eher ungleiches Vorzeichen haben. Je näher $\rho(X, Y)$ an ± 1 ist, desto stärker ist dieser Effekt. Je näher $\rho(X, Y)$ bei 0 ist, desto weniger wissen wir über den Zusammenhang von X und Y. Am besten sieht man das an den Extremfällen: Für $X = Y$ gilt $\rho(X, Y) = 1$, für $X = -Y$ gilt $\rho(X, Y) = -1$, für unabhängige Zufallsvariablen gilt $\rho(X, Y) = 0$.

Satz 4.2.31. **[Bienaymé]** Sind $X_1, ..., X_d$ Zufallsvariablen auf $(\Omega, \mathcal{A}, \mathbb{P})$ mit $\mathbb{E}[X_1^2], ..., \mathbb{E}[X_d^2] < \infty$, so gelten:

(i)
$$\mathbb{V}\Big[\sum_{k=1}^d X_k\Big] = \sum_{k=1}^d \mathbb{V}[X_k] + \sum_{\substack{i,j=1\\ i\neq j}}^d \operatorname{Cov}(X_i, X_j).$$

(ii)
$$\mathbb{V}\Big[\sum_{k=1}^d X_k\Big] = \sum_{k=1}^d \mathbb{V}[X_k],$$

falls $X_1, ..., X_d$ paarweise unkorreliert sind, d. h. wenn $\operatorname{Cov}(X_i, X_j) = 0$ für alle $i, j = 1, ..., d$, $i \neq j$.

Beweis. Übung, das ist einfach nur Ausmultiplizieren und Definitionen einsetzen. □

4.3 Rechnen mit Zufallsvariablen

Vorlesung 22

In diesem Abschnitt wollen wir endlich mal mit Zufallsvariablen konkret rechnen. Wir schauen uns an, wie man eine Zufallsvariable zu einer anderen transformiert, wie man in konkreten Beispielen mehrere Zufallsvariablen zu einer neuen transformiert und dann noch, was mit Summen von unabhängigen Zufallsvariablen passiert.

4.3.1 Inverse Transformations Methode

Schauen wir uns als erstes an, wie man einzelne Zufallsvariablen zu anderen Zufallsvariablen transformieren kann. Den Trick haben wir schon gesehen, als wir nach Beispiel 4.1.6 eine uniforme Zufallsvariable in eine exponentielle Zufallsvariable transformiert haben. Das geht auch allgemeiner. Wir brauchen dazu die Pseudoinverse, die auch in Stochastik 2 und Monte Carlo Methoden sehr prominent auftauchen wird.

Definition 4.3.1. Ist F eine Verteilungsfunktion, so heißt

$$F^{-1}(x) := \inf\{s \in \mathbb{R} \colon F(s) \geq x\}, \quad x \in [0,1],$$

Pseudoinverse (oder **verallgemeinerte Inverse**) von F. Mit $\inf(\emptyset) := +\infty$ und $\inf(\mathbb{R}) := -\infty$ ist F^{-1} eine Abbildung von $[0,1]$ nach $\bar{\mathbb{R}}$.

Der Begriff Pseudoinverse taucht auf, weil die Umkehrfunktion (inverse Funktion) nur für bijektive Funktionen definiert ist. Für Verteilungsfunktionen muss F allerdings nicht surjektiv sein (Sprünge) oder nicht injektiv sein (stückweise konstant). Wenn F stetig und streng monoton wachsend ist (z. B. wenn F eine positive Dichte hat), dann ist die Pseudoinverse einfach nur die Umkehrfunktion (Spiegelung an der Winkelhalbierenden) aus Analysis 1!

Bemerkung. Sprünge in F werden konstante Stücke in F^{-1}, konstante Stücke in F werden Sprünge in F^{-1}.

Folgende elementare Eigenschaften sind essentiell, um mit F^{-1} zu arbeiten:

Lemma 4.3.2.

(i) Ist F bijektiv, so ist F^{-1} die Umkehrfunktion aus Analysis 1.

(ii) F^{-1} ist nicht-fallend.

(iii) Es gilt $F^{-1}(y) \leq t \Leftrightarrow y \leq F(t)$ für alle $t \in \mathbb{R}$ und $y \in (0,1)$.

Beweis.

(i) Definition.

(ii) Definition.

(iii) Nach Definition gilt

$$F^{-1}(y) \leq t \quad \Leftrightarrow \quad \inf\{s\colon F(s) \geq y\} \leq t \quad \Leftrightarrow \quad F(t) \geq y.$$

□

Der Trick an der verallgemeinerten Inversen ist, dass man damit alle (ja, wirklich alle!) Zufallsvariablen aus einer uniformen Zufallsvariablen basteln kann, indem man diese in die verallgemeinerte Inverse einsetzt. Für die Theorie ist das unglaublich nützlich.

Satz 4.3.3. **[Inverse Transformation Methode]** Ist $U \sim \mathcal{U}((0,1))$ und F eine beliebige Verteilungsfunktion, so gilt $X := F^{-1}(U) \sim F$.

Beweis. Um die Verteilungsfunktion von X zu berechnen, setzten wir einfach die Definition ein und führen die Wahrscheinlichkeit mit (iii) aus dem letzten Lemma auf die uniforme Verteilung zurück:

$$F_X(t) = \mathbb{P}(X \leq t) = \mathbb{P}(F^{-1}(U) \leq t) = \mathbb{P}(U \leq F(t)) = F(t), \quad t \in \mathbb{R}.$$

Im letzten Schritt haben wir die Verteilungsfunktion von $\mathcal{U}((0,1))$ eingesetzt (siehe Übungsaufgaben). □

Eine kleine Bemerkung zur uniformen Verteilung in der Trafomethode. Oft sieht man $V \sim \mathcal{U}([0,1])$ statt $U \sim \mathcal{U}((0,1))$. Im Prinzip ist es egal, was man nimmt, die beiden Zufallsvariablen U und V sind nämlich identisch verteilt (siehe Übungsaufgaben), der Unterschied liegt auf einer Nullmenge. Weil aber $F^{-1}(0) = -\infty$ und $F^{-1}(1) = +\infty$ sein kann, müssten wir uns dann mit V Gedanken über die Definition einer Zufallsvariablen machen. Die nehmen gemäß unserer Definition nur endliche reelle Werte an.

Die Trafomethode sieht auf den ersten Blick klasse aus. Wenn man eine uniforme Zufallsvariable simulieren kann (das nennt man auch samplen), so kann man automatisch alle Zufallsvariablen simulieren! In der Monte Carlo Vorlesung wird der erste Simulationsalgorithmus für Zufallsvariablen daher wie folgt funktionieren. Simuliere eine $\mathcal{U}((0,1))$ Zufallsvariable (das ist eigentlich Zahlen-theorie) und setzte diese in F^{-1} ein. Das gibt dann eine Simulation einer Zufallsvariablen mit Verteilungsfunktion F. Leider ist das zu schön, um wahr zu sein. Der Grund ist, dass leider F^{-1} selten explizit bekannt ist, dann bringt die Methode natürlich nicht viel für praktische Anwendungen. Es gibt aber einige Beispiele, in denen F^{-1} explizit bekannt ist.

Beispiel 4.3.4.

- Für den diskreten Fall solltet ihr euch Beispiele aufs Papier malen. Ist F eine diskrete Verteilungsfunktion mit Werten $a_1 < ... < a_N$ und Wahrscheinlichkeiten $p_1, ..., p_N$, so ist F^{-1} eine Treppenfunktion. Die Trafomethode gibt folgenden intuitiven Algorithmus, eine diskrete Zufallsvariable zu simulieren: Zerlege das Intervall $(0,1)$ in die Teilintervalle $I_1 = (0, p_1], I_2 = (p_1, p_1 + p_2]$ bis $I_N = (p_1 + ... + p_{N-1}, 1)$ und ziehe uniform eine Zahl U aus $(0,1)$, U liegt also in einem der Intervalle I_k. Liegt U in I_k, so gebe den Wert a_k aus. Die so gewonnene Zufallsvariable nimmt die Werte a_k mit Wahrscheinlichkeiten p_k an, ist also diskret mit Verteilungsfunktion F.

- Für die Exponentialverteilung Exp(1) ist $F^{-1}(t) = -\log(1-t), t \in [0,1]$. Also gilt $X = -\ln(1-U) \sim \text{Exp}(1)$, sofern U eine uniforme Zufallsvariable ist.

- Für die Normalverteilung funktioniert die Methode nicht. Wir haben keine explizite Formel für F, also auch nicht für F^{-1}. Die Methode kann aber trotzdem für theoretische Betrachtungen genutzt werden, wie wir bei der Konvergenz von Zufallsvariablen sehen werden.

Mehr explizite Beispiele kennen wir leider noch nicht.

4.3.2 Ein paar konkrete Beispiele

Als nächstes wollen wir an ein paar Beispielen zeigen, wie man aus verschiedenen Zufallsvariablen neue Zufallsvariablen basteln kann. Ihr sollt damit ein besseres Gefühl für verschiedene Zufallsvariablen bekommen und lernen, wie man mit den Erwartungswerten von unabhängigen Zufallsvariablen rechnet. Im nächsten Abschnitte betrachten wir dann den Spezialfall von Summen von Zufallsvariablen.

Beispiel 4.3.5. [**gespiegelte Zufallsvariable**] Eine Zufallsvariable X heißt symmetrisch, wenn $\mathbb{P}(X \leq t) = \mathbb{P}(X \geq -t)$ für alle $t \in \mathbb{R}$ gilt. Das bedeutet, dass X gleichwahrscheinlich Werte in an 0 gespiegelten Mengen annimmt. Eine äquivalente Formulierung ist, dass $X \sim -X$ gilt. Checken wir dafür einmal schnell die Verteilungsfunktionen:

$$F_{-X}(t) = \mathbb{P}(-X \leq t) = \mathbb{P}(X \geq -t) = \mathbb{P}(X \leq t) = F_X(t), \quad t \in \mathbb{R}.$$

Symmetrie ist einfach im diskreten oder absolutstetigen Fall zu erkennen. Im diskreten bedeutet dies, dass die „Zähldichte“ immer an Werten a_k und $-a_k$ die gleichen Werte p_k annimmt. Im absolutstetigen Fall muss die Dichte achsensymmetrisch sein, so wie zum Beispiel bei $\mathcal{N}(0,1)$.

Beispiel 4.3.6. [**Uniform zu uniform**] Ist $U \sim \mathcal{U}([0,1])$ eine uniform verteilte Zufallsvariable, so ist auch $V := 1 - U \sim \mathcal{U}([0,1])$. Dazu berechnen wir einfach die Verteilungsfunktion von V, weil wir die Verteilungsfunktion von U kennen:

$$F_V(t) = \mathbb{P}(V \leq t) = \mathbb{P}(U \geq 1-t) = 1 - \mathbb{P}(U \leq 1-t) = \begin{cases} 0 & : t < 0 \\ 1-(1-t) & : t \in (0,1) \\ 1 & : t > 1 \end{cases},$$

für $t \in \mathbb{R}$. Damit hat V die Verteilungsfunktion einer $\mathcal{U}([0,1])$-verteilten Zufallsvariable. Beachtet, dass wir $\mathbb{P}(U \leq 1-t) = \mathbb{P}(U < 1-t)$, also $\mathbb{P}(U = 1-t) = 0$ genutzt haben. Das liegt daran, dass U eine absolutstetige Zufallsvariable ist.

Beispiel 4.3.7. [**Minimum von** Exp-**verteilten Zufallsvariablen**] Sind $X \sim \operatorname{Exp}(\alpha)$ und $Y \sim \operatorname{Exp}(\beta)$ unabhängige Zufallsvariablen, so gilt $\min\{X,Y\} \sim \operatorname{Exp}(\lambda + \beta)$. Wir beachten dazu, dass das Minimum zweier messbarer Abbildungen wieder messbar ist, also ist auch $Z := \min\{X,Y\}$ eine Zufallsvariable auf $(\Omega, \mathcal{A}, \mathbb{P})$. Berechnen wir nun die Verteilungsfunktion von Z (das Minimum ist größer als t genau dann, wenn beide größer als t sind):

$$\begin{aligned} \mathbb{P}(Z \leq t) &= 1 - \mathbb{P}(\min\{X,Y\} > t) \\ &\overset{\text{Unab.}}{=} 1 - \mathbb{P}(X > t, Y > t) \\ &= 1 - \mathbb{P}(X > t)\mathbb{P}(X > y) \\ &= 1 - e^{-\alpha t}e^{-\beta t} = 1 - e^{-(\alpha+\beta)t}, \quad t > 0. \end{aligned}$$

Wenn man $t \leq 0$ einsetzt, kommt überall 0 raus weil X und Y nicht-negative Zufallsvariablen sind. Damit hat Z die Verteilungsfunktion von $\operatorname{Exp}(\alpha + \beta)$. Exakt die gleiche Rechnung funktioniert auch mit geometrischen Zufallsvariablen, das Minimum zweier unabhängiger geometrischer Zufallsvariablen ist wieder geometrisch - probiert das mal aus!

Beispiel 4.3.8. Seien $X, Y \sim \operatorname{Exp}(1)$ unabhängig, so gilt

$$\frac{X}{X+Y} \sim \mathcal{U}([0,1]).$$

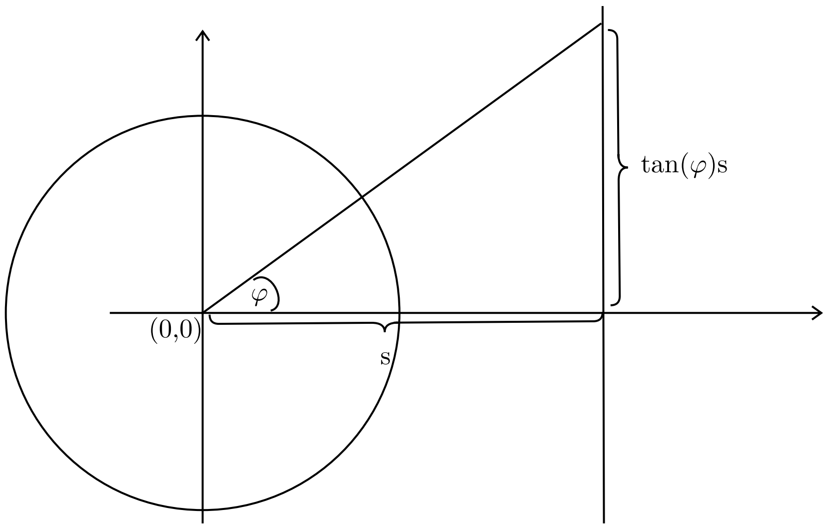
Diese Übungsaufgabe braucht ein gutes Verständniss vom Rumrechnen mit Zufallsvektoren, daher ein Tipp. Zu berechnet ist die Verteilungsfunktion $\mathbb{P}(\frac{X}{X+Y} \leq t)$. Die einzige Information ist die gemeinsame Dichte von (X, Y), die aufgrund der Unabhängigkeit faktorisiert. Wir schreiben die Verteilungsfunktion mit geeignetem g wieder als $\mathbb{E}[g(X, Y)]$ um, und nutzen dann die Berechnungsregel im absolutstetigen Fall. In diesem Fall schreibt man

$$\mathbb{P}\Big(\frac{X}{X+Y} \leq t\Big) = \mathbb{P}\Big(X \leq Y\frac{t}{1-t}\Big) = \mathbb{E}\Big[\mathbf{1}_{(-\infty, Y\frac{t}{1-t}]}(X)\Big].$$

Die rechte Seite kann man durch Einsetzen ausrechnen, auf geht's!

Einen Haufen weiterer Verbindungen verschiedener Verteilungen kann man hier (klicken) verlinkt finden. Sollte der Link bei eurem pdf-viewer nicht funktionieren, sucht einfach nach „Wikipedia relationships among probability distributions. Auch witzig ist zum Beispiel folgende Aussage: Sind $X, Y \sim \mathcal{N}(0, 1)$ und unabhängig, so ist $Z := \frac{X}{Y}$ Cauchyverteilt!

Beispiel 4.3.9. **[Discokugel]** Was hat eine „Discokugel“ (blinkende Kugel in der Mitte eines Raumes, die mittels kleiner Laser bunte Punkte an die Wand wirft) mit der Cauchyverteilung zu tun? Die Punkte an der Wand sind Realisierungen einer Cauchyverteilung! Schauen wir uns zunächst eine Skizze an:



Der Kreis ist ein zweidimensionaler Schnitt der Kugel. Es wird ein Punkt auf der rechten Hälfte des Kreisrandes uniform gezogen (d. h. der Winkel φ wird uniform aus $(-\frac{\pi}{2}, \frac{\pi}{2})$ gezogen, φ hat also Dichte $\frac{1}{\pi}\mathbf{1}_{(-\frac{\pi}{2}, \frac{\pi}{2})}$). Dann wird der Laser vom Ursprung in Richtung des gezogenen Punktes auf dem Kreisrand geschossen und bis zur Mauer verlängert, wo dann ein bunter Punkt erscheint. Aus der Schule sollte noch bekannt sein, dass der Treffpunkt der Mauer $(s, \tan(\varphi)s)$ ist. Berechnen wir nun die Verteilung des Treffpunktes Y (y-Achsen Abstand) auf der Mauer. Weil $\varphi \sim \mathcal{U}((-\frac{\pi}{2}, \frac{\pi}{2}))$ ist, gilt

$$\mathbb{P}(Y \leq t) = \mathbb{P}(\tan(\varphi)s \leq t) = \mathbb{P}(\varphi \leq \arctan(t/s)) = \frac{1}{2} + \frac{1}{\pi}\arctan(t/s),$$

wobei in der letzten Gleichheit die Verteilungsfunktion von $\mathcal{U}((-\frac{\pi}{2}, \frac{\pi}{2}))$ eingesetzt wurde (beachte, dass $\arctan(t/s) \in (-\frac{\pi}{2}, \frac{\pi}{2})$ für alle $t, s \in \mathbb{R}$). Damit ist die Verteilungsfunktion von Y gerade die Verteilungsfunktion von Cauchy$(s, 0)$.

Wir hatten angemerkt, dass die inverse Transformations Methode für die Normalverteilung nicht funktioniert weil F^{-1} nicht explizit bekannt ist. Das nächste Beispiel ist daher ziemlich verblüffend, die sogenante Box-Muller Methode. Man kann aus einer uniformen Zufallsvariable zwar nicht einfach eine normalverteilte Zufallsvariable bekommen, man kann aber ganz einfach aus zwei uniformen Zufallsvariablen zwei (und damit auch eine) normalverteilte bekommen! Sind U_1, U_2 unabhängige $\mathcal{U}([0,1])$ verteilte Zufallsvariablen, so sind

$$X_1 = \cos(2\pi U_1)\sqrt{-\log(2U_2)} \quad \text{und} \quad X_2 = \sin(2\pi U_1)\sqrt{-\log(2U_2)}$$

zwei unabhängige $\mathcal{N}(0,1)$ Zufallsvariablen. Wer gerade noch Energie übrig hat, kann mal versuchen, $\mathbb{P}(X_1 \leq t_1, X_2 \leq t_2)$ mit Polarkoordinaten auszurechnen (alle anderen können sich das in der Monte Carlo Methoden Vorlesung anschauen). Um die Verteilung zu berechnen, solltet ihr euch an Korollar 4.2.22 erinnern. Das Korollar besagt, dass U_1, U_2 und X_1, X_2 gemeinsame Dichten haben, nämlich jeweils das Produkt der einzelnen, also $\mathbf{1}_{[0,1]\times[0,1]}$ bzw. $\frac{1}{2\pi}e^{-(x_1^2+x_2^2)/2}$. Damit wisst ihr, wie ihr $\mathbb{P}(X_1 \leq t_1, X_2 \leq t_2)$ durch Einsetzen der Definition ausrechnet und was rauskommen sollte.

4.3.3 Summen von unabhängigen Zufallsvariablen

Wir berechnen nun die Verteilungen von Summen unabhängiger Zufallsvariablen. Was ist zum Beispiel die Verteilung der Summe zweier exponentialverteilter Zufallsvariablen? Oder wie ist die Summe zweier Normalverteilungen verteilt? Allgemein ist das ziemlich schwierig zu sagen, wenn die Zufallsvariablen aber unabhängig sind, gibt es schöne Rechentricks.

Starten wir mit dem diskreten Fall, der ist einfacher. Die Idee ist einfach: Durch welche Kombination von Werten, kann die Summe $X+Y$ einen gegebenen Wert annehmen? Natürlich indem Y irgendeinen Wert annimmt, und X gerade die Differenz zum gegebenen Wert. Wenn man alle solche Möglichkeiten addiert (und die Unabhängigkeit ausnutzt), bekommt man die diskrete Faltungsformel:

Satz 4.3.10. **[Diskrete Faltungsformel]** Sind X, Y diskrete Zufallsvariablen auf $(\Omega, \mathcal{A}, \mathbb{P})$, dann ist auch $X+Y$ diskret mit möglichen Werten in $X(\Omega)+Y(\Omega) := \{a+b : a \in X(\Omega), b \in Y(\Omega)\}$. Dabei bezeichnen wir mit $X(\Omega), Y(\Omega)$ die möglichen Werte von X, Y sind. Sind X und Y unabhängig, so können die Wahrscheinlichkeiten mit der Faltungsformel berechnet werden:

$$\mathbb{P}(X+Y=k) = \sum_{b\in Y(\Omega)} \mathbb{P}(X=k-b)\mathbb{P}(Y=b), \quad k \in X(\Omega)+Y(\Omega).$$

Lasst euch nicht von der komischen Notation mit den $X(\Omega)$ und $Y(\Omega)$ irritieren, man kann die Formel einfach nicht schön hinschreiben. Sobald ihr euch die folgenden Beispiele angeschaut habt, ist ganz schnell alles klar.

Beweis. Dass die Summe diskret ist und als Werte gerade Summen der Werte von X und Y annehmen kann, ist hoffentlich klar (denkt an zwei Würfel). Der Trick für die Faltungsformel ist, ein diskretes Ereigniss in eine disjunkte Vereinigung von Teilereignissen (alle Möglichkeiten) zu zerlegen und darauf die σ-Additivität anzuwenden. Wir schreiben den Beweis einmal knapp auf, so schreibt man Argumente mit diskreten Zufallsvariablen fast immer auf, und dann noch einmal super ausführlich, um die Maßtheorie dahinter zu erkennen:

$$\begin{aligned}\mathbb{P}(X+Y=k) &\overset{\sigma\text{-add.}}{=} \sum_{b\in Y(\Omega)} \mathbb{P}(X+Y=k, Y=b)\\ &= \sum_{b\in Y(\Omega)} \mathbb{P}(X=k-b, Y=b)\\ &\overset{\text{unab.}}{=} \sum_{b\in Y(\Omega)} \mathbb{P}(X=k-b)\mathbb{P}(Y=b).\end{aligned}$$

Die zweite Gleichheit nutzt σ-Additivität weil alle Möglichkeiten von Y ausprobiert wurden.

Ganz ausführlich sieht das gleiche Argument wie folgt aus:

$$\begin{aligned}
\mathbb{P}(X+Y=k) &\overset{\text{Notation}}{=} \mathbb{P}(\{\omega\colon X(\omega)+Y(\omega)=k\})\\
&\overset{\text{Trick}}{=} \mathbb{P}\Big(\dot{\bigcup_{b\in Y(\Omega)}}\{\omega\colon X(\omega)+Y(\omega)=k, Y(\omega)=b\}\Big)\\
&\overset{\sigma\text{-add.}}{=} \sum_{b\in Y(\Omega)} \mathbb{P}(\{\omega\colon X(\omega)+Y(\omega)=k, Y(\omega)=b\})\\
&= \sum_{b\in Y(\Omega)} \mathbb{P}(\{\omega\colon X(\omega)=k-b, Y(\omega)=b\})\\
&\overset{\text{Notation}}{=} \sum_{b\in Y(\Omega)} \mathbb{P}(X=k-b, Y=b)\\
&\overset{\text{unab.}}{=} \sum_{b\in Y(\Omega)} \mathbb{P}(X=k-b)\mathbb{P}(Y=b).
\end{aligned}$$

Wir müssen in der Mathematik immer vorsichtig sein, einfache Dinge nicht zu sehr zu verkomplizieren. Diskrete Zufallsvariablen sind ganz sicher so ein Beispiel! □

Als Anmerkung zum Freuen auf die strahlende Zukunft: Genau wegen dieses kleinen Tricks, kann man so schön mit Markovketten rumrechnen!

Die Formel sieht vielleicht abstrakt und unhandlich aus, wie können aber wirklich ganz einfach in konkrete Beispielen damit rechnen:

Beispiel 4.3.11. Sind $X_1, ..., X_n$ u.i.v mit $X_1 \sim \text{Ber}(p)$ für ein $p \in (0,1)$, dann gilt $X_1 + ... + X_n \sim \text{Bin}(n,p)$. Interpretation: Wenn $\text{Ber}(p)$ die erfolgreiche Ausführung eines Versuchs (1=Erfolg, 0=Misserfolg) beschreibt, so beschreibt $\text{Bin}(n,p)$ die Anzahl der erfolgreichen Ausführungen von n unabhängigen Versuchen.

Beweis. Induktion über n:

IA: $n = 1$: ✓ $\text{Ber}(p) = \text{Bin}(1,p)$ nach Definition beider Verteilungen.

IV: Die Behauptung gelte für ein *beliebiges*, aber festes $n \in \mathbb{N}$.

IS: Indem wir $X_1 + ... + X_{n+1} = (X_1 + ... + X_n) + X_{n+1}$ klammern, wenden wir die diskrete Faltungsformel an und nutzen dann die Induktionsvoraussetzung:

$$\begin{aligned}
\mathbb{P}(X_1+...+X_{n+1}=k) &\overset{4.3.10}{=} \mathbb{P}(X_1+...+X_n=k-1)\mathbb{P}(X_{n+1}=1)\\
&\quad+\mathbb{P}(X_1+...+X_n=k)\mathbb{P}(X_{n+1}=0)\\
&= \binom{n}{k-1}p^{k-1}(1-p)^{n-k+1}p + \binom{n}{k}p^k(1-p)^{n-k}(1-p)\\
&= \left(\binom{n}{k-1}+\binom{n}{k}\right)p^k(1-p)^{n-k+1}\\
&= \binom{n+1}{k}p^k(1-p)^{n-k+1}.
\end{aligned}$$

Im letzten Schritt haben wir eine Rechnenregel für Binomialkoeffizienten benutzt, die kennt ihr vermultich aus Analysis 1. Also ist $X_1 + ... + X_n \sim \text{Bin}(n,p)$ gezeigt.

□

Beispiel 4.3.12. Seien $X \sim \text{Poi}(\lambda), Y \sim \text{Poi}(\beta)$ unabhängig. Dann ist auch $X+Y$ Poisson-verteilt, und zwar mit Parameter $\lambda+\beta$. In den Übungen setzt ihr einfach in die Faltungsformel ein, um $X+Y \sim \text{Poi}(\lambda+\beta)$ zu zeigen. Easy.

Definition 4.3.13.

(i) Sind $\mu_1, ..., \mu_n$ Wahrscheinlichkeitsmaße auf $\mathcal{B}(\mathbb{R})$, so heißt das Bildmaß vom Produktmaß $\mu_1 \otimes ... \otimes \mu_n$ unter der Abbildung $h_d(x_1, ..., x_d) = x_1 + ... + x_d$ **Faltung** der Maße. Wir schreiben $\mu_1 * ... * \mu_n$ für die Faltung.

(ii) Sind $X_1, ..., X_d$ unabhängige Zufallsvariablen auf $(\Omega, \mathcal{A}, \mathbb{P})$, so heißt $\mathbb{P}_{X_1} * ... * \mathbb{P}_{X_d}$ **Faltung** von $X_1, ..., X_d$.

Bemerkung. Natürlich ist die Definition der Faltung abstrakt, andererseits ist sie auch konkret. Die Faltung ist nichts weiter als die Verteilung der Summe unabhängiger Zufallsvariablen:

$$X_1 + ... + X_n \sim \mathbb{P}_{X_1} * ... * \mathbb{P}_{X_d}.$$

Wir müssen uns dafür nur daran erinnern, dass die Verteilung unabhängiger Zufallsvariablen das Produktmaß ist:

$$\begin{aligned}\mathbb{P}(X_1 + ... + X_d \in B) &= \mathbb{P}(X_1, ..., X_d \in h_d^{-1}(B))\\ &\overset{\text{unab.}}{=} \mathbb{P}_{X_1} \otimes ... \otimes \mathbb{P}_{X_d}(h_d^{-1}(B))\\ &\overset{\text{Def. push-forw.}}{=} \mathbb{P}_{X_1} * ... * \mathbb{P}_{X_d}(B).\end{aligned}$$

Mit der Definition der Faltung können wir erstmal nicht viel anstellen, wir haben die Faltung schließlich einfach als das definiert, was wir berechnen wollen. Was wir gerne hätten, wäre ein analog zu der diskreten Faltungsformel, weil wir damit in konkreten Beispielen rechnen können. Hier ist die allgemeine Formel, danach konkretisieren wir diese für den Fall mit Dichten.

Proposition 4.3.14. Seien μ_1, μ_2 Wahrscheinlichkeitsmaße auf $\mathcal{B}(\mathbb{R})$ mit Verteilungsfunktionen F_1, F_2, dann gelten:

(i) Mit $B - y := \{x - y \colon x \in B\}$, d. h. die Verschiebung der Menge B um y nach links, gilt

$$\mu_1 * \mu_2(B) = \int_{\mathbb{R}} \mu_1(B - y) \,\mathrm{d}\mu_2(y)$$

für alle $B \in \mathcal{B}(\mathbb{R})$.

(ii) Für alle $t \in \mathbb{R}$ gilt

$$F_{\mu_1 * \mu_2}(t) = \int_{\mathbb{R}} F_1(t - y) \,\mathrm{d}\mu_2(y).$$

Überlegt mal kurz zum Spaß, warum die verschobenen Mengen $B - y$ wieder messbar sind. [1]

Beweis.

(i) Wegen der Manipulation

$$\mathbf{1}_{B-y}(x) = \begin{cases} 1 & : x \in B - y \\ 0 & : \text{sonst} \end{cases} = \begin{cases} 1 & : x + y \in B \\ 0 & : \text{sonst} \end{cases} = \mathbf{1}_{h_2^{-1}(B)}(x, y)$$

folgt

$$\begin{aligned}\mu_1 * \mu_2(B) &\overset{\text{Def.}}{=} \mu_1 \otimes \mu_2(h_2^{-1}(B))\\ &= \int_{\mathbb{R}^2} \mathbf{1}_{h_2^{-1}(B)}(x, y) \,\mathrm{d}\mu_1 \otimes \mu_2(x, y)\\ &= \int_{\mathbb{R}^2} \mathbf{1}_{B-y}(x) \,\mathrm{d}\mu_1 \otimes \mu_2(x, y)\\ &\overset{\text{Fubini}}{=} \int_{\mathbb{R}} \Big(\int_{\mathbb{R}} \mathbf{1}_{B-y}(x) \,\mathrm{d}\mu_1(x) \Big) \mu_2(y)\\ &= \int_{\mathbb{R}} \mu_1(B - y) \,\mathrm{d}\mu_2(y).\end{aligned}$$

Das ist die erste Aussage.

[1] Weil $B - y$ das Urbild von B unter der messbaren (weil stetig) Abildung $z \mapsto z + y$ ist.

(ii) Wir setzten in (i) die Mengen $B = (-\infty, t]$ mit $B - y = (-\infty, t - y]$ ein, und nutzen die Definition der Verteilungsfunktion.

□

Nun zur konkreten Rechenvorschrift, um die Verteilung der Summe unabhängiger Zufallsvariablen mit Dichten zu berechnen:

Satz 4.3.15. **[Stetige Faltungsformel]** Sind X, Y unabhängige absolutstetige Zufallsvariablen mit Dichten f_X, f_Y, dann ist auch $X + Y$ absolutstetig und hat Dichte

$$f_{X+Y}(x) = \int_{\mathbb{R}} f_X(x-y) f_Y(y)\,\mathrm{d}y, \quad x \in \mathbb{R}.$$

Beweis. Mit vorheriger Proposition kennen wir die Verteilungsfunktion der Summe. Wir rechnen diese mit der Formel aus und setzen dabei die bekannten Dichten ein:

$$\begin{aligned}
\mathbb{P}_{X+Y}((-\infty, t]) &\overset{\text{Def. Faltung}}{=} \mathbb{P}_X * \mathbb{P}_Y((-\infty, t]) \\
&\overset{4.3.14}{=} \int_{\mathbb{R}} \mathbb{P}_X((-\infty, t-y])\,\mathrm{d}\mathbb{P}_Y(y) \\
&\overset{3.3.2}{=} \int_{\mathbb{R}} \int_{-\infty}^{t-y} f_X(x)\,\mathrm{d}x f_Y(y)\,\mathrm{d}y \\
&\overset{\text{Subst.}}{=} \int_{\mathbb{R}} \int_{-\infty}^{t} f_X(x-y)\,\mathrm{d}x f_Y(y)\,\mathrm{d}y \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} \mathbf{1}_{(-\infty,t]}(x) f_X(x-y) f_Y(y)\,\mathrm{d}x\,\mathrm{d}y \\
&\overset{\text{Fubini}}{=} \int_{-\infty}^{t} \int_{\mathbb{R}} f_X(x-y) f_Y(y)\,\mathrm{d}y\,\mathrm{d}x \\
&= \int_{-\infty}^{t} f_{X+Y}(x)\,\mathrm{d}x.
\end{aligned}$$

Also ist $X + Y$ absolutstetig mit behaupteter Dichte f_{X+Y}. □

Das Konzept der Faltung kommt nicht aus der Stochastik, daher können wir mit dem Begriff „Faltung“ auch nichts anfangen. Die Faltung zweier integrierbarer Funktionen wird in der Fourieranalysis als

$$f * g(x) = \int_{\mathbb{R}} f(x-y) g(y)\,\mathrm{d}y, \quad x \in \mathbb{R},$$

definiert und zum Beispiel in der Signalverarbeitung studiert. Dass die Faltung bei uns für Summen unabhängiger Zufallsvariablen auftaucht, ist ein Zufall der Mathematik. Es gibt also keinen Grund sich Gedanken über die Begrifflichkeit Faltung zu machen, für uns ist es einfach nur eine Berechnungsformel. Vorlesung 23

Beispiel 4.3.16.

(i) Sind $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ und $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ unabhängig, dann ist auch die Summe wieder normalverteilt, genauer, es gilt $X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. Das kann man mit der stetigen Faltungsformel ausrechnen:

$$\begin{aligned}
f_{X_1+X_2}(x) &= \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(x-y-\mu_1)^2}{2\sigma_1^2}} \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{(y-\mu_2)^2}{2\sigma_2^2}}\,\mathrm{d}y \\
&= \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}} e^{-\frac{(x-(\mu_1+\mu_2))^2}{2(\sigma_1^2+\sigma_2^2)}}, \quad x \in \mathbb{R}.
\end{aligned}$$

Auf den ersten Blick ist nicht so klar, ob die Berechnung einfach oder nicht so einfach ist. Im Prinzip muss man nur richtig quadratisch ergänzen und Substituieren. In der Tat ist die Rechnung ziemlich böse, klappt aber. Weil wir gleich eine viel einfachere Methode kennenlernen, lassen wir die Rechnung weg.

(ii) Sind $X_1 \sim \operatorname{Exp}(\lambda), X_2 \sim \operatorname{Exp}(\lambda)$ unabhängig, so ist $X_1 + X_2 \sim \Gamma(2,\lambda)$. Das kann man direkt mit der stetigen Faltungsformel ausrechnen:

$$\begin{aligned} f_{X_1+X_2}(x) &= \int_{\mathbb{R}} f_{X_1}(x-y) f_{X_2}(y)\,\mathrm{d}y \\ &= \int_{\mathbb{R}} \mathbf{1}_{[0,\infty)}(x-y)\lambda e^{-\lambda(x-y)} \mathbf{1}_{[0,\infty)}(y)\lambda e^{-\lambda y}\,\mathrm{d}y \\ &= \lambda^2 \int_0^x e^{-\lambda x}\,\mathrm{d}y = \lambda^2 x e^{-\lambda x}. \end{aligned}$$

Jetzt noch eine ganz andere Methode zur Berechnung der Verteilung der Summe von unabhängigen Zufallsvariablen. Dafür nutzen wir einen Satz der Wahrscheinlichkeitstheorie, den Eindeutigkeitssatz für momenterzeugende Funktionen. Der Satz besagt, dass Verteilungen eindeutig durch ihre momenterzeugenden Funktionen festgelegt sind, falls diese um 0 existieren.

Satz 4.3.17. Seien X und Y Zufallsvariablen für die die momenterzeugenden Funktionen $\mathcal{M}_X, \mathcal{M}_Y$ in $(-\varepsilon,\varepsilon)$ existieren für ein $\varepsilon > 0$. Falls $\mathcal{M}_X(t) = \mathcal{M}_Y(t)$ für alle $t \in (-\varepsilon,\varepsilon)$ gilt, so gilt $X \sim Y$.

Beweis. Siehe fortgeschrittene Stochastikvorlesung - hart! □

Für Verteilungen mit expliziten momenterzeugenden Funktionen ist diese ein extrem nützliches Hilfsmittel.

Beispiel 4.3.18. Mit der momenterzeugenden Funktion können wir ganz einfach die Skalierungseigenschaft der Normalverteilung zeigen: Ist $X \sim \mathcal{N}(0,1)$, $\mu \in \mathbb{R}$ und $\sigma^2 > 0$, so gilt $Y := \sigma X + \mu \sim \mathcal{N}(\mu,\sigma^2)$. Rechnen wir dazu die momenterzeugende Funktion von Y aus:

$$M_Y(t) = \mathbb{E}[e^{(\sigma X+\mu)t}] = \mathbb{E}[e^{\sigma t X}]e^{\mu t} = M_X(\sigma t)\cdot e^{\mu t} = e^{\frac{\sigma^2 t^2}{2}+\mu t}, \quad t \in \mathbb{R},$$

und die rechte Seite ist gerade die momenterzeugende Funktion einer $\mathcal{N}(\mu,\sigma^2)$-verteilten Zufallsvariable. Also folgt die Behauptung aus Satz 4.3.17.

Gemeinsam mit folgender Proposition ist Satz 4.3.17 ein mächtiges Hilfsmittel um die Verteilung Summen unabhängiger Zufallsvariablen zu berechnen:

Proposition 4.3.19. Sind X und Y unabhängige Zufallsvariablen mit $\mathcal{M}_X(t), \mathcal{M}_Y(t) < \infty$ für ein $t \in \mathbb{R}$, so gilt auch $\mathcal{M}_{X+Y}(t) = \mathcal{M}_X(t)\mathcal{M}_Y(t)$.

Beweis. Das folgt direkt daraus, dass Erwartungswerte von Produkten unabhängiger Zufallsvariablen faktorisieren, siehe Satz 4.2.26:

$$\mathcal{M}_{X+Y}(t) = \mathbb{E}[e^{t(X+Y)}] = \mathbb{E}[e^{tX}\cdot e^{tY}] = \mathbb{E}[e^{tX}]\cdot\mathbb{E}[e^{tY}] = \mathcal{M}_X(t)\mathcal{M}_Y(t).$$

□

Die Anwendung der momenterzeugenden Funktion zur Bestimmung der Verteilung von Summen unabhängiger Zufallsvariablen versteht man am besten mit folgendem einfachen Beispiel. Beachte: Das Beispiel ist einfach, die Aussage aber nicht. Hätten wir nicht die Kanone aus der Wahrscheinlichkeitstheorie ausgepackt, hätten wir das mit der stetigen Faltungsformel nachrechnen müssen.

Beispiel 4.3.20. Kommen wir zurück zu Beispiel 4.3.16 (i) und berechnen mit der Proposition die momenterzeugende Funktion der Summe der zwei unabhängigen Normalverteilungen. Wegen Proposition 4.3.19 und der in den Übungen berechneten Formel für die momenterzeugende Funktion einer Normalverteilung gilt

$$\mathcal{M}_{X_1+X_2}(t) = \mathcal{M}_{X_1}(t)\mathcal{M}_{X_2}(t) = e^{\mu_1 t + \frac{\sigma_1^2}{2}t^2} e^{\mu_2 t + \frac{\sigma_2^2}{2}t^2} = e^{(\mu_1+\mu_2)t + \frac{\sigma_1^2+\sigma_2^2}{2}t^2}, \quad t \in \mathbb{R}.$$

Nun wissen wir aber auch, dass

$$\mathcal{M}_Y(t) = e^{(\mu_1+\mu_2)t + \frac{\sigma_1^2+\sigma_2^2}{2}t^2}, \quad t \in \mathbb{R},$$

für $Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. Also gilt $X_1 + X_2 \sim Y$ nach Satz 4.3.17 und damit gilt für die Summe $X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

Mit ähnlichen Rechnungen kann man (siehe Übungen) mit momenterzeugenden Funktionen auch zeigen, dass

- die Summe unabhängiger Exponentialverteilungen gammaverteilt ist,
- die Summe unabhängiger Gammaverteilungen wieder gammaverteilt ist,
- die Summe unabhängiger Bernoulliverteilungen binomialverteilt ist (haben wir oben schon mit der Faltungsformel berechnet),

Das Argument ist immer identisch: Berechne die momenterzeugende Funktion der Summe als Produkt der momenterzeugenden Funktionen der einzelnen (Unabhängigkeit) und hoffe, dass das Produkt eine bereits bekannt momenterzeugende Funktion ist. Dann kann die Verteilung der Summe mittels Satz 4.3.17 identifiziert werden.

Bemerkung 4.3.21.

- Der Trick mit der Normalverteilung funktioniert aufgrund der Faktorisierungseigenschaft der Exponentialfunktion. Wir sehen also, dass der Trick auf Situationen beschränkt ist, wenn in einer nützlichen Form Potenzen auftauchen.
- Die Vorlesung ist hier etwas irreführend. Satz 4.3.17 ist kein Allheilmittel. In der (mathematischen) Realität sind exponentielle Momente sehr oft unendlich und wir können nicht mit der momenterzeugenden Funktion arbeiten. In dieser Vorlesung sind alle Verteilungen außer Cauchy und Exponentiell unproblematisch, weil alle exponentiellen Momente existieren. In der Wahrscheinlichkeitstheorie werden wir das Problem lösen, indem wir die Momenterzeugende Funktionen durch sogenannte charakteristische Funktionen $\varphi(t) = \mathbb{E}[e^{itX}]$ ersetzen.

4.4 Bedingte Wahrscheinlichkeiten und Unabhängigkeit

Wir kommen nun zu bedingten Wahrscheinlichkeiten, die aus der Schule vielleicht bekannt sind. In dieser Vorlesung spielen bedingte Wahrscheinlichkeiten noch keine zentrale Rolle, das ändert sich in späteren Vorlesungen sehr, wenn zum Beispiel Markovketten angeschaut werden.

Definition 4.4.1. Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $A, B \in \mathcal{A}$ mit $\mathbb{P}(B) > 0$. Dann heißt

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

bedingte Wahrscheinlichkeit von A **gegeben** B.

Lemma 4.4.2. Für $B \in \mathcal{A}$ mit $\mathbb{P}(B) > 0$ ist $A \mapsto \mathbb{P}(A|B) =: \mathbb{P}_B(A)$ ein Maß auf $(\Omega, \mathcal{A})$.

Beweis. Wir rechnen die definierenden Eigenschaften eines Maßes nach:

- $\mathbb{P}_B(A) \geq 0$ für alle $A \in \mathcal{A}$ ist klar, weil $\mathbb{P}$ ein Maß ist.
- Weil $\mathbb{P}$ ein Maß ist, gilt auch
$$\mathbb{P}_B(\emptyset) = \mathbb{P}(\emptyset|B) = \frac{\mathbb{P}(\emptyset \cap B)}{\mathbb{P}(B)} = 0.$$
- σ-Additivität: Seien $A_1, A_2, ... \in \mathcal{A}$ disjunkt, so gilt wegen der σ-Additivität von $\mathbb{P}$
$$\begin{aligned}\mathbb{P}_B\Big(\dot{\bigcup_{k=1}^{\infty}} A_k\Big) &\overset{\text{Def.}}{=} \frac{\mathbb{P}\big(\dot\bigcup_{k=1}^{\infty} A_k \cap B\big)}{\mathbb{P}(B)}\\ &= \frac{\mathbb{P}\big(\dot\bigcup_{k=1}^{\infty}(A_k \cap B)\big)}{\mathbb{P}(B)}\\ &\overset{\sigma\text{-Add.}}{=} \frac{\sum_{k=1}^{\infty} \mathbb{P}(A_k \cap B)}{\mathbb{P}(B)} = \sum_{k=1}^{\infty} \mathbb{P}_B(A_k).\end{aligned}$$

□

Wir kommen nun zu extrem wichtigen Rechenregeln, obwohl diese ganz einfach aus der Definition folgen:

Satz 4.4.3.

(i) **Multiplikationsregel**: Für $A_1, ..., A_n \in \mathcal{A}$ mit $\mathbb{P}(\bigcap_{k=1}^n A_k) > 0$ gilt
$$\mathbb{P}\Big(\bigcap_{k=1}^{n} A_k\Big) = \mathbb{P}(A_1) \cdot \mathbb{P}(A_2|A_1) \cdot ... \cdot \mathbb{P}\Big(A_n\Big|\bigcap_{k=1}^{n-1} A_k\Big).$$

(ii) **Formel der totalen Wahrscheinlichkeit**: Ist $B_1, ..., B_n \in \mathcal{A}$ eine disjunkte Zerlegung von Ω, d. h. $\dot\bigcup_{k=1}^n B_k = \Omega$, mit $\mathbb{P}(B_k) > 0$ für alle $k = 1, ..., n$, so gilt
$$\mathbb{P}(A) = \sum_{k=1}^{n} \mathbb{P}(B_k)\mathbb{P}(A|B_k), \quad \forall A \in \mathcal{A}.$$

(iii) **Bayes-Formel**: Mit $B_1, ..., B_n$ aus (ii) gilt für $A \in \mathcal{A}$ mit $\mathbb{P}(A) > 0$
$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B) \cdot \mathbb{P}(B)}{\mathbb{P}(A)}, \quad \forall B \in \mathcal{A},$$
oder
$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B) \cdot \mathbb{P}(B)}{\sum_{k=1}^{n} \mathbb{P}(B_k)\mathbb{P}(A|B_k)}, \quad \forall B \in \mathcal{A}.$$

Beweis. (i) Induktion über n:

IA: $n = 2$ folgt direkt aus der Definition der bedingten Wahrscheinlichkeit.

IV: Die Behauptung gelte für ein *beliebiges*, aber festes $n \in \mathbb{N}$.

IS: Wenn wir nun die Induktionsvoraussetzung und den Fall $n = 2$ nutzen, so bekommen wir
$$\begin{aligned}\mathbb{P}\Big(\bigcap_{k=1}^{n+1} A_k\Big) &= \mathbb{P}\Big(\bigcap_{k=1}^{n} A_k \cap A_{n+1}\Big)\\ &= \mathbb{P}\Big(\bigcap_{k=1}^{n} A_k\Big)\mathbb{P}\Big(A_{n+1}\Big|\bigcap_{k=1}^{n} A_k\Big)\\ &\overset{\text{IV}}{=} \mathbb{P}(A_1) \cdot \mathbb{P}(A_2|A_1) \cdot ... \cdot \mathbb{P}\Big(A_{n+1}\Big|\bigcap_{k=1}^{n} A_k\Big)\end{aligned}$$

und das ist gerade die Aussage.

(ii) Zunächst folgt direkt aus der Definition

$$\mathbb{P}(B_k)\mathbb{P}(A|B_k) = \mathbb{P}(B_k)\frac{\mathbb{P}(A \cap B_k)}{\mathbb{P}(B_k)} = \mathbb{P}(A \cap B_k).$$

Setzen wir dies in folgende Rechnung ein, so bekommen wir die Aussage:

$$\mathbb{P}(A) = \mathbb{P}(A \cap \Omega) = \mathbb{P}\Big(A \cap \dot{\bigcup_{k=1}^{n}} B_k\Big) = \mathbb{P}\Big(\dot{\bigcup_{k=1}^{n}}(A \cap B_k)\Big) \overset{\sigma\text{-Add.}}{=} \sum_{k=1}^{n} \mathbb{P}(A \cap B_k).$$

(iii) Die einfache Bayes-Formel folgt direkt durch Einsetzen der Definition und Erweitern:

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

Ganz Aufmerksame werden merken, dass wir bei $\mathbb{P}(B) = 0$ durch null geteilt haben. Das geht natürlich nicht! Ist aber kein Problem, weil die Bayes-Formel in dem Fall als $0 = 0$ ohnehin gilt. Wir können also ohne Einschränkung $\mathbb{P}(B) > 0$ annehmen und dann taucht das Problem nicht auf.

Die zweite Formel folgt aus der ersten Formel durch Ersetzen des Nenners durch die Formel der totalen Wahrscheinlichkeit.

$\square$

Kommen wir nun zu zwei klassischen Anwendungen. Hier ist allerdings etwas Vorsicht angesagt, das ist alles etwas wild (funktioniert aber). Gemäß Definition brauchen wir für bedingte Wahrscheinlichkeiten einen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$. In vielen Anwendungen außerhalb der Mathematik werden die Formeln allerdings auch genutzt, ohne einen Wahrscheinlichkeitsraum hinzuschreiben. Nennen wir das vielleicht „heuristischen Gebrauch von Wahrscheinlichkeiten". Das ist natürlich nicht sehr schön, allerdings sind solche Aussagen extrem wichtig.

Beispiel 4.4.4.

(i) Wir ziehen aus vier blauen und drei weißen Kugeln zweimal ohne Zurücklegen. Was ist die Wahrscheinlichkeit, zwei blaue zu ziehen? Machen wir das ganze zunächst „heuristisch". Das ist ein zweistufiges Experiment. Im ersten Versuch ziehen wir blau mit Wahrscheinlichkeit $\frac{4}{7}$. Mit dem Wissen im ersten Zug blau gezogen zu haben, ist das „bedingte Ziehen" im zweiten Schritt ein Ziehen aus drei blauen und drei weißen Kugeln. Die bedingte Wahrscheinlichkeit im zweiten Schritt blau zu ziehen, gegeben das erste Ziehen gab eine blaue Kugel, ist also $\frac{3}{6}$. Mit der Multiplikationsregel folgt

$$\begin{aligned} &\mathbb{P}(\text{beide Kugeln blau}) \\ &= \mathbb{P}(\text{erste Kugel blau}) \cdot \mathbb{P}(\text{zweite Kugel blau} \,|\, \text{erste Kugel blau}) \\ &= \frac{4}{7} \cdot \frac{3}{6} = \frac{2}{7}. \end{aligned}$$

Das funktioniert ganz entspannt, ist aber schon etwas kritisch. Wir nutzen hier ganz intuitiv den Begriff der bedingten Wahrscheinlichkeit in einem interpretierten Sinn (Änderung des Modells, eine Kugel weniger). Dann haben wir die Multiplikationsregel genutzt, die aufgrund unseres Beweises und damit aufgrund der mathematischen Definition der bedingten Wahrscheinlichkeit stimmt. Nun ist allerdings nicht so ganz klar, warum der intuitive Begriff der bedingten Wahrscheinlichkeit (geändertes Experiment) mit der mathematischen Definition $\mathbb{P}(A \,|\, B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$ überhaupt etwas zu tun hat. Machen wir das ganze zur Beruhigung also nochmal mathematisch penibel genau. Schreiben wir zunächst ein Modell hin, das wir als Modell für das zweifache Würfeln plausibel finden. Dazu sei

$$\Omega = \{(\omega_1, \omega_2) \colon \omega_1, \omega_2 \in \{\text{blau, weiß}\}\}$$

und $\mathcal{A} = \mathcal{P}(\Omega)$. Als Wahrscheinlichkeiten definieren wir

$$\mathbb{P}(\{\omega_1, \omega_2\}) = \begin{cases} \frac{4}{7} \cdot \frac{3}{6} & : \omega_1 = \omega_2 = \text{blau} \\ \frac{3}{7} \cdot \frac{2}{6} & : \omega_1 = \omega_2 = \text{weiß} \\ \frac{4}{7} \cdot \frac{3}{6} & : \omega_1 = \text{blau},\ \omega_2 = \text{weiß} \\ \frac{3}{7} \cdot \frac{4}{6} & : \omega_1 = \text{weiß},\ \omega_2 = \text{blau} \end{cases}.$$

Mit den Ereignissen $A = \{(\omega_1, \omega_2) \colon \omega_1 = \text{blau}\}$, $B = \{(\omega_1, \omega_2) \colon \omega_2 = \text{blau}\}$ wollen wir die Wahrscheinlichkeit von $A \cap B$ bestimmen. Mit der Multiplikationsregel folgt

$$\mathbb{P}(\text{ziehe zweimal blau}) = \mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B|A) = \frac{4}{7} \cdot \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \frac{2}{7}.$$

Das macht natürlich auch wieder nicht so richtig viel Sinn, wir hätten die Wahrscheinlichkeit von $A \cap B$ auch ohne die Multiplikationsregel „ablesen“ können. Dennoch ist das vielleicht eine gute Beruhigung: Wenn wir wollen, können wir ein rigoroses Modell hinschreiben, in dem die Multiplikationsformel rigoros gemacht werden kann. Für die Schulanwendungen ist das natürlich viel zu kompliziert.

(ii) Nun ein Beispiel für die Bayes-Formel, ein medizinischer Test, z. B. ein Aidstest. Wir machen das jetzt wieder in der „heuristischen“ Art, wer will, kann sich wieder ein sauberes Modell definieren. Wir nehmen an, dass die Wahrscheinlichkeiten folgender Ereignisse bekannt sind:

- 1% der Bevölkerung ist tatsächlich krank.
- Test ist mit 98% positiv, wenn eine Person krank ist.
- Test ist mit 5% positiv, wenn eine Person nicht krank ist, das nennt mann false-positive (Fehlalarm).

Gesucht ist die Wahrscheinlichkeit gesund zu sein, obwohl der Test positiv ist. Mit der Bayes-Formel gilt

$$\begin{aligned} &\mathbb{P}(\text{krank} \mid \text{Test positiv}) \\ &= \mathbb{P}(\text{Test positiv} \mid \text{krank}) \cdot \frac{\mathbb{P}(\text{krank})}{\mathbb{P}(\text{Test positiv})} \\ &= \frac{\mathbb{P}(\text{Test positiv} \mid \text{krank})\mathbb{P}(\text{krank})}{\mathbb{P}(\text{Test positiv} \mid \text{krank})\mathbb{P}(\text{krank}) + \mathbb{P}(\text{Test positiv} \mid \text{gesund})\mathbb{P}(\text{gesund})} \\ &= \frac{0,98 \cdot 0,01}{0,98 \cdot 0,01 + 0,05 \cdot 0,99} \\ &= 0,165. \end{aligned}$$

Das ergibt dann

$$\mathbb{P}(\text{gesund} \mid \text{Test positiv}) = 1 - 0,165 = 0,835,$$

was erschreckend hoch ist! Die wichtige take-home message ist also: Wird etwas sehr unwahrscheinliches getestet, dominieren falsche Tests und man muss bei positiven Tests unbedingt weitere Tests machen. Ein weiteres sehr wichtiges Beispiel sind pränatale Tests bei ungeborenen Kindern, z. B. Tests auf Trisomie 21 (auf die moralische Frage wollen wir hier natürlich nicht eingehen!).

Weiter geht es jetzt mit einer mathematischen Definition von Unabhängigkeit von Ereignissen. Ihr habt sicherlich eine naive Vorstellung „Hat nichts mit einander zu tun“. Beispielsweise wären die Ereignisse „Meine Kaffeemaschiene geht morgen kaputt“ und „Es regnet morgen in Thailand“ vermutlich unabhängig. Hier ist eine Definition:

Definition 4.4.5. Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und seien $A, B \in \mathcal{A}$. Die Ereignisse A und B heißen **unabhängig**, falls $\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$.

Aufgrund der Definition der bedingten Wahrscheinlichkeit sind, sofern $\mathbb{P}(B) > 0$ gilt, A und B unabhängig genau dann, wenn $\mathbb{P}(A|B) = P(A)$. Das passt also zur Begriffsbildung: Zwei Ereignisse sind genau dann unabhängig, falls die Wahrscheinlichkeit des einen sich nicht ändert, wenn das Eintreten des anderen gegeben ist.
Oft braucht man auch die Unabhängigkeit mehrerer Ereignisse. Für endlich viele ist klar was man macht, die Wahrscheinlichkeit des endlichen Schnittes soll natürlich das endlich Produkt sein. Für unendlich viele ist das problematisch, darum führt man alles auf endlich viele zurück:

Definition 4.4.6. Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum, $A_i \in \mathcal{A}$, $i \in I$, und I eine beliebige Indexmenge.

(i) Die Ereignisse $(A_i)_{i \in I}$ heißen **unabhängig**, falls
$$\mathbb{P}\Big(\bigcap_{i \in J} A_i\Big) = \prod_{i \in J} \mathbb{P}(A_i), \quad \forall J \subseteq I \text{ mit } \#J < \infty.$$

(ii) Die Ereignisse $(A_i)_{i \in I}$ heißen **paarweise unabhängig**, falls
$$\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i) \cdot \mathbb{P}(A_j), \quad \forall i \neq j.$$

Selbstverständlich impliziert Unabhängikgkeit die paarweise Unabhängigkeit, statt aller endlicher Teilmengen J von I werden schließlich nur alle Teilmengen mit $\#J = 2$ gewählt. Die Umkehrung gilt nicht:

Warnung 4.4.7. Paarweise Unabhängigkeit impliziert im Allgemeinen nicht Unabhängigkeit. Kleine Mengen reichen schon aus, um Gegenbeispiel anzugeben, siehe Übungsblatt.

In Worten ausgedrückt heißt die Unabhängigkeit auch „Die Ereignisse $(A_i)_{i \in I}$ sind unabhängig, falls jede Wahl von endlich vielen Ereignissen $A_{i_1}, ..., A_{i_n}$ unabhängig ist“.
Als nächstes wollen wir die Unabhängigkeit von σ-Algebren und Zufallsvariablen thematisieren. Dazu zunächst eine allgemeinere Definition:

Definition 4.4.8. Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $(\mathcal{E}_i)_{i \in I}$ eine Familie von Teilmengen $\mathcal{E}_i \subseteq \mathcal{A}$ der σ-Algebra für eine beliebige Indexmenge I. Dann heißen die $(\mathcal{E}_i)_{i \in I}$ unabhängig, falls die Ereignisse $(A_i)_{i \in I}$ unabhängig sind, und zwar für alle $A_i \in \mathcal{E}_i$.

Die Definition wird insbesondere für σ-Algebren verwendet. Wie bei der Messbarkeit fragen wir uns hier, ob wir die Unabhängigkeit auf Erzeuger reduzieren können. Wie immer funktioniert das, zumindest wenn der Erzeuger $\cap$-stabil ist:

Proposition 4.4.9. Ist $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $\mathcal{E}_i \subseteq \mathcal{A}$ für für alle $i \in I$. Sind alle $\mathcal{E}_i$ $\cap$-stabil, so gilt:
$$(\mathcal{E}_i)_{i \in I} \text{ unabhängig} \quad \Leftrightarrow \quad (\sigma(\mathcal{E}_i))_{i \in I} \text{ unabhängig.}$$

Beweis.

„$\Leftarrow$“: Klar nach Definition, weil $\mathcal{E}_i \subseteq \sigma(\mathcal{E}_i)$ gilt. Die Unabhängigkeit von Mengensystemen bedeutet schließlich, dass die Unabhängigkeit für alle Auswahlen von Teilmengen gilt. Gilt dies für mehr Möglichkeiten, so natürlich auch für weniger Möglichkeiten.

„$\Rightarrow$“: Ohne Einschränkung sei I endlich weil die Definition der Unabhängkigkeit nur auf endlichen Teilmengen $J \subseteq I$ beruht. Nennen wir die Mengen $\mathcal{E}_1, ..., \mathcal{E}_n$. Wir zeigen, dass dann auch $\sigma(\mathcal{E}_1), ..., \sigma(\mathcal{E}_n)$ unabhängig sind. Dazu zeigen wir zunächst:
$$\mathcal{D} := \big\{E \in \mathcal{A} \colon \{E\}, \mathcal{E}_2, ..., \mathcal{E}_n \text{ sind unabhängig}\big\} \text{ ist ein Dynkin-System.}$$

Um das zu zeigen, checken wir die definierenden Eigenschaften eines Dynkin-Systems:

(i) Aufgrund der Definition 4.4.8 müssen wir zeigen, dass für alle $A_2 \in \mathcal{E}_2, ..., A_n \in \mathcal{E}_n$ die Mengen $\Omega, A_2, ..., A_n$ unabhängig sind, die Wahrscheinlichkeit vom Schnitt also zur Wahrscheinlichkeiten der einzelnen Ereignisse faktorisiert. Das folgt aber direkt aus der angenommenen Unabhängigkeit von $\mathcal{E}_1, ..., \mathcal{E}_n$:

$$\begin{aligned}\mathbb{P}(\Omega \cap A_2 \cap ... \cap A_n) &= \mathbb{P}(A_2 \cap ... \cap A_n)\\ &= \mathbb{P}(A_2) \cdot ... \cdot \mathbb{P}(A_n)\\ &= 1 \cdot \mathbb{P}(A_2) \cdot ... \cdot \mathbb{P}(A_n) = \mathbb{P}(\Omega) \cdot \mathbb{P}(A_2) \cdot ... \cdot \mathbb{P}(A_n).\end{aligned}$$

Damit ist $\Omega \in \mathcal{D}$.

(ii) Nun zur Abgeschlossenheit bezüglich Komplementbildung. Wir argumentieren wie im ersten Schritt. Sei dazu $E \in \mathcal{D}$ und seien $A_2 \in \mathcal{E}_2, ..., A_n \in \mathcal{E}_n$ beliebig. Weil $\Omega = E \cup E^C$ ergibt sich

$$\begin{aligned}&\mathbb{P}(E^C \cap A_2 \cap ... \cap A_n)\\ &\overset{\sigma\text{-Add.}}{=} \mathbb{P}(\Omega \cap A_2 \cap ... \cap A_n) - \mathbb{P}(E \cap A_2 \cap ... \cap A_n)\\ &\overset{\Omega, E\in\mathcal{D}}{=} \mathbb{P}(\Omega) \cdot \mathbb{P}(A_2) \cdot ... \cdot \mathbb{P}(A_n) - \mathbb{P}(E) \cdot \mathbb{P}(A_2) \cdot ... \cdot \mathbb{P}(A_n)\\ &= (\mathbb{P}(\Omega) - \mathbb{P}(E)) \cdot \mathbb{P}(A_2) \cdot ... \cdot \mathbb{P}(A_n)\\ &= \mathbb{P}(E^C) \cdot \mathbb{P}(A_2) \cdot ... \cdot \mathbb{P}(A_n).\end{aligned}$$

Also sind $E^C, A_2, ..., A_n$ unabhängige Ereignisse und damit ist $E^C \in \mathcal{D}$.

(iii) Das Argument für die Vereinigungen geht genau wie für die Komplemente.

Jetzt beenden wir den Beweis. Aufgrund der Annahme $\mathcal{E}_1, ..., \mathcal{E}_n$ unabhängig, gilt für jede Menge $E \in \mathcal{E}_1$ auch $E \in \mathcal{D}$. Also gilt $\mathcal{E}_1 \subseteq \mathcal{D}$. Daraus folgt mit dem Hauptsatz für Dynkinsysteme, Satz 1.2.11, wie immer (Bildung des kleinsten Dynkin-Systems ist monoton)

$$\sigma(\mathcal{E}_1) \overset{\cap\text{-stabil}}{=} d(\mathcal{E}_1) \subseteq d(\mathcal{D}) = \mathcal{D}.$$

Weil also $\sigma(\mathcal{E}_1) \subseteq \mathcal{D}$ gilt, folgt aus der Definition von $\mathcal{D}$, dass $\sigma(\mathcal{E}_1), \mathcal{E}_2, ..., \mathcal{E}_n$ unabhängig sind. Iterativ ersetzen wir nun Schritt für Schritt mit einem analogen Argument ein $\mathcal{E}_k$ nach dem anderen durch $\sigma(\mathcal{E}_k)$, indem wir genau wie oben zeigen, dass alle

$$\mathcal{D}_k := \big\{E \in \mathcal{A} \colon \sigma(\mathcal{E}_1), \sigma(\mathcal{E}_2), ..., \sigma(\mathcal{E}_{k-1}), \{E\}, \mathcal{E}_{k+1}, ..., \mathcal{E}_n \text{ sind unabhängig}\big\}$$

Dynkin-Systeme sind und daraus $\sigma(\mathcal{E}_k) \subseteq \mathcal{D}_k$ folgern.

□

Vorlesung 24

Nach der Unabhängigkeit von Ereignissen und Zufallsvariablen kommen wir nun zu einer alternativen Definition der Unabhängigkeit von Zufallsvariablen. Anstatt die Faktorisierung der gemeinsamen Verteilung zu fordern, kann man auch fordern, dass die erzeugten σ-Algebren (siehe Definition 2.1.8) unabhängig sind:

Definition 4.4.10. Für Zufallsvariablen $(X_i)_{i\in I}$ auf $(\Omega, \mathcal{A}, \mathbb{P})$ definiert man: $(X_i)_{i\in I}$ sind unabhängig, falls die erzeugten σ-Algebren $(\sigma(X_i))_{i\in I}$ unabhängig sind.

Weil $\sigma(X_i) \overset{\text{Def.}}{=} \{X_i^{-1}(B) : B \in \mathcal{B}(\mathbb{R})\} = \sigma(\{X_i^{-1}((-\infty, t]) \colon t \in \mathbb{R}\})$, können wir direkt folgern, dass die Unabhängigkeit auch durch die gemeinsame Verteilungsfunktion definiert werden kann.

Korollar 4.4.11. Für Zufallsvariablen $X_1, ..., X_d$ auf $(\Omega, \mathcal{A}, \mathbb{P})$ stimmt die neue Definition der Unabhängigkeit mit der alten überein. Wir können Unabhängigkeit also auf verschiedene

Arten beschreiben:

$$\begin{aligned} & X_1, ..., X_d \text{ sind unabhängig} \\ \Leftrightarrow \quad & \sigma(X_1), ..., \sigma(X_d) \text{ sind unabhängige } \sigma\text{-Algebren} \\ \Leftrightarrow \quad & F_X(t_1, ..., t_d) = F_{X_1}(t_1) \cdot ... \cdot F_{X_d}(t_d), \quad \forall t_i \in \mathbb{R} \\ \Leftrightarrow \quad & \mathbb{P}(X_1 \in A_1, ..., X_d \in A_d) = \mathbb{P}(X_1 \in A_1) \cdot ... \cdot \mathbb{P}(X_d \in A_d), \quad \forall A_i \in \mathcal{B}(\mathbb{R}). \end{aligned}$$

Beweis. Um die vorherige Proposition anzuwenden, seien $\mathcal{E}_i := \{\{X_i \leq t\} \colon t \in \mathbb{R}\}$ für $i = 1, ..., d$. Also gilt $\sigma(\mathcal{E}_i) = \sigma(X_i)$ und die $\mathcal{E}_i$ sind $\cap$-stabil. Überlegen wir einmal schnell, warum $\sigma(\mathcal{E}_i) = \sigma(X_i)$ gilt. Die Richtung „$\subseteq$" gilt, weil $\mathcal{E}_i \subseteq \sigma(X_i)$ aufgrund der Definition von $\sigma(X_i)$. Die Richtung „$\supseteq$" gilt, weil wegen Proposition 2.1.4 X_i $(\sigma(\mathcal{E}_i), \mathcal{B}(\mathbb{R}))$-messbar ist und $\sigma(X_i)$ die kleinste σ-Algebra ist, bezüglich derer X_i messbar ist.

Damit gilt aufgrund der Definitionen der Unabhängigkeit von Mengensystemen und Ereignissen

$$\begin{aligned} & \sigma(X_1), ..., \sigma(X_d) \text{ unabhängig} \\ \overset{4.4.9}{\Leftrightarrow} \quad & \mathcal{E}_1, ..., \mathcal{E}_d \text{ unabhängig} \\ \Leftrightarrow \quad & E_1, ..., E_d \text{ unabhängig für alle } E_1 \in \mathcal{E}_1, ..., E_d \in \mathcal{E}_d \\ \Leftrightarrow \quad & \mathbb{P}(\{X_1 \leq t_1\} \cap ... \cap \{X_d \leq t_d\}) = \mathbb{P}(\{X_1 \leq t_1\}) \cdots \mathbb{P}(\{X_d \leq t_d\}), \quad t_1, ..., t_d \in \mathbb{R} \\ \overset{\text{Notation}}{\Leftrightarrow} \quad & \mathbb{P}(X_1 \leq t_1, ..., X_d \leq t_d) = \mathbb{P}(X_1 \leq t_1) \cdots \mathbb{P}(X_d \leq t_d), \quad t_1, ..., t_d \in \mathbb{R} \\ \overset{\text{Def. VF}}{\Leftrightarrow} \quad & F_X(t_1, ..., t_d) = F_{X_1}(t_1) \cdots F_{X_d}(t_d), \quad t_1, ..., t_d \in \mathbb{R}. \end{aligned}$$

□

Bitte beachtet die Notation im letzten Beweis. In der Stochastik bevorzugen wir immer die Notation $\mathbb{P}(X_1 \leq t_1, ..., X_d \leq t_d)$, wir nutzen praktisch nie die ausführliche Schreibweise $\mathbb{P}(\{X_1 \leq t_1\} \cap ... \cap \{X_d \leq t_d\})$. Das liegt einfach nur daran, dass sich die erste Notation viel natürlicher lesen lässt. Am besten gewöhnt ihr euch direkt die kompakte Schreibweise an.

Im nächsten Abschnitt besprechen wir Konvergenzen von Folgen von Zufallsvariablen. Wir werden insbesondere Folgen unabhängiger Zufallsvariablen nutzen. Unsere ursprüngliche Definition war nur für endlich viele Zufallsvariablen, die Definition dieses Abschnittes funktioniert auch für unendlich viele Zufallsvariablen (man testet die Eigenschaft einfach für alle endlichen Teilmengen). Genau so wollen wir ab jetzt die Unabhängigkeit einer ganzen Folgen von Zufallsvariablen definieren.

Definition 4.4.12. Ist $X_1, X_2, ...$ eine (unendliche) Folge von Zufallsvariablen auf $(\Omega, \mathcal{A}, \mathbb{P})$, so heißt die Folge unabhängig, falls eine der äquivalenten Eigenschaften gilt:

(i) Für alle $n \in \mathbb{N}$ sind $\sigma(X_1), ..., \sigma(X_n)$ unabhängige σ-Algebren.

(ii) Für alle $n \in \mathbb{N}$ gilt: $\mathbb{P}(X_1 \leq t_1, ..., X_n \leq t_n) = \prod_{k=1}^n \mathbb{P}(X_k \leq t_k), \quad t_1, ..., t_n \in \mathbb{R}$.

Wie für Zufallsvariablen und Zufallsvektoren ist es auch für Folgen von Zufallsvariablen nicht klar, dass es diese überhaupt gibt. In der Tat kann man auch in diesem Fall eine kanonische Konstruktion angeben, die uns die Existenz von Folgen unabhängiger Zufallsvariablen gibt. Der kanonische Wahrscheinlichkeitsraum besteht ganz analog aus den Werten, die angenommen werden. Dies war zunächst $\mathbb{R}$, dann $\mathbb{R}^d$ und ist nun $\mathbb{R}^\infty$ (die Menge der reellen Folgen). Die kanonische σ-Algebra ist die passende „Borel"-σ-Algebra und die Folge der Zufallsvariablen ist durch die Identitätsabbildung gegeben. Das Thema gehört eigentlich nicht in die Stochastik 1 sondern in die Wahrscheinlichkeitstheorie 1. Daher skizzieren wir die Konstruktion nur ganz kurz. Ihr solltet euch jedoch merken, dass es eine kanonische Konstruktion gibt und insbesondere Folgen von unabhängigen Zufallsvariablen existieren

Satz 4.4.13. **[Kanonische Konstruktion von Folgen unabhängiger Z.V.]** Seien $F_1, F_2, ...$ Verteilungsfunktionen, so existieren ein Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ und eine Folge *unabhängiger* Zufallsvariablen $X_1, X_2, ...$ auf $(\Omega, \mathcal{A}, \mathbb{P})$ mit $X_i \sim F_i$.

Beweis. Man kann ganz analog zu $\mathbb{R}$ und zum $\mathbb{R}^d$ eine kanonische Konstruktion angeben. Hier nur eine Skizze für die übermotivierten Studis, die Konstruktion wird in Ruhe in einer der Vorlesungen im Master thematisiert. Alle anderen merken sich aber bitte die Aussage des Satzes, sonst wäre die Vorlesung an dieser Stelle vorbei!

- $\Omega := \{(\omega_n)_{n\in\mathbb{N}} \colon \omega_n \in \mathbb{R}\}$, die Menge der „reelle Folgen". Die ω sind also Folgen, oder unendlich lange Vektoren. Als Analogie zum $\mathbb{R}^d$ schreibt man auch $\mathbb{R}^\infty$.
- $\mathcal{A} := \mathcal{B}(\mathbb{R}^\infty) := \sigma(\{B_1 \times ... \times B_d \times \mathbb{R} \times \mathbb{R} \times ... : d \in \mathbb{N}, B_1, ..., B_d \in \mathcal{B}(\mathbb{R})\})$
- $\mathbb{P} := \mathbb{P}_{F_1} \otimes \mathbb{P}_{F_2} \otimes ...$ sei das unendliche Produktmaß, das auf dem Erzeuger von $\mathcal{B}(\mathbb{R}^\infty)$ festgelegt ist durch $\mu(B_1 \times ... \times B_d \times \mathbb{R} \times ...) = \mathbb{P}_{F_1}(B_1) \cdots \mathbb{P}_{F_d}(B_d)$.
- $(X_1(\omega), X_2(\omega), ...) := (\omega_1, \omega_2, ...) = \omega$

Um zu zeigen, dass es ein unendliches Produktmaß auf $(\mathbb{R}^\infty, \mathcal{B}(\mathbb{R}^\infty))$ auch gibt, kann man den Fortsetzungssatz von Carathéodory anwenden. Das ist etwas hässlich. Wir haben nun also einen Wahrscheinlichkeitsraum und eine Folge von Zufallsvariablen, die einfach nur für ein ω die Koordinaten ausgibt (genau wie im $\mathbb{R}^d$, vergleiche den Beweis von Satz 4.2.15). Die Unabhängigkeit der konstruierten Folge folgt direkt aus der Produkteigenschaft des Produktmaßes, weil aufgrund der Definition der Zufallsvariablen als Koordinatenabbildungen

$$\begin{aligned}\mathbb{P}(X_1 \le t_1, ..., X_n \le t_n) &\overset{\text{Def } X}{=} \mathbb{P}((-\infty, t_1] \times ... \times (-\infty, t_n] \times \mathbb{R} \times \mathbb{R} \times ...)\\ &\overset{\text{Produktmaß}}{=} \mathbb{P}_{F_1}((-\infty, t_1]) \cdots \mathbb{P}_{F_n}((-\infty, t_n])\\ &= \mathbb{P}(X_1 \le t_1) \cdots \mathbb{P}(X_n \le t_n).\end{aligned}$$

Auch sofort folgt durch Einsetzen, dass die Randverteilungen $\mathbb{P}(X_i \le t) = F_{X_i}(t)$ erfüllen. □

Definition 4.4.14. Sind alle F_i gleich, so nennen wir die Folge unabhängiger Zufallsvariablen $X_1, X_2, ...$ aus Satz 4.4.13 eine u.i.v. Folge mit $X_1 \sim F_1$.

In den nächsten Kapitel schauen wir uns Konvergenzeigenschaften von u.i.v Folgen an, insbesondere das Gesetz der großen Zahlen und den zentralen Grenzwertsatz.

4.5 Konvergenz von Folgen von Zufallsvariablen

Bevor wir zu den zentralen Konvergenzsätzen kommen, müssen wir uns überlegen, was Konvergenz von Folgen von Zufallsvariablen überhaupt bedeutet. Das ist in der Tat gar nicht so klar, es gibt verschiedene Begriffe:

Definition 4.5.1. **[Vier Konvergenzarten in der Stochastik]** Für eine Zufallsvariable X und eine Folge $X_1, X_2, ...$ von Zufallsvariablen auf $(\Omega, \mathcal{A}, \mathbb{P})$ definiert man

(i) „X_n konvergiert **stochastisch** (oder **in Wahrscheinlichkeit**) gegen X", man schreibt

$$X_n \xrightarrow{P} X, \quad n \to \infty,$$

falls für alle $\varepsilon > 0$

$$\mathbb{P}(|X_n - X| > \varepsilon) \to 0, \quad n \to \infty.$$

(ii) „X_n konvergiert **in $\mathcal{L}^p$** gegen X" (oder **im p-ten Mittel**) für $p \ge 1$, man schreibt

$$X_n \xrightarrow{\mathcal{L}^p} X, \quad n \to \infty,$$

falls

$$\mathbb{E}[|X_n - X|^p] \to 0, \quad n \to \infty.$$

(iii) „X_n konvergiert **fast sicher** gegen X“, man schreibt

$$X_n \xrightarrow{\text{f.s.}} X, \quad n \to \infty,$$

falls

$$\mathbb{P}(X_n \to X) := \mathbb{P}(\{\omega\colon X_n(\omega) \to X(\omega),\, n \to \infty\}) = 1.$$

(iv) „X_n konvergiert **in Verteilung** (oder **schwach**) gegen X“, man schreibt

$$X_n \xrightarrow{(d)} X, \quad n \to \infty,$$

falls für alle $f : \mathbb{R} \to \mathbb{R}$ stetig und beschränkt

$$\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)], \quad n \to \infty.$$

Nur die ersten drei Konvergenzbegriffe benötigen wirklich, dass die Zufallsvariablen auf dem selben Wahrscheinlichkeitsraum definiert sind. Die Konvergenz in Verteilung ist strukturell anders weil die Zufallsvariablen X_n nicht direkt mit der Grenzzufallsvariablen X „verglichen“ werden, es wird nicht $|X_n(\omega) - X(\omega)|$ berechnet. Die Konvergenz in Verteilung hängt nur von den Verteilungen ab, vergleiche dazu die Berechnungsformel des Erwartungswertes mit dem Transformationssatz, Lemma 4.1.10.

Bemerkung 4.5.2.

(i) Warnung: Die Konvergenzen sind nicht durch Metriken definiert worden, d. h. übliche Tricks aus der Analysis (z. B. $\triangle$-Ungleichung, Eindeutigkeit von Grenzwerten, ...) gelten nicht einfach so! Konkretes Beispiel: Nur wenn man einen Quotientenraum mit fast sicher gleichen Zufallsvariablen bildet, ist Konvergenz im p-ten Mittel eine Normenkonvergenz und die Eindeutigkeit von Grenzwerten gilt (siehe Übungsaufgaben).

(ii) Zwei Konvergenzarten sind uns schon bekannt:

- fast sichere Konvergenz ist schon von messbaren Funktionen bekannt,
- p-tes Mittel ist schon von $(\mathcal{L}^p, ||\cdot||_p)$ bekannt.

Um ein Gefühl für die Konvergenzarten zu bekommen, sind Beispiele äusserst nützlich. Viele nützliche Beispiele können ganz explizit hingeschrieben werden.

Beispiel 4.5.3. Seien $X_1, X_2, ...$ Zufallsvariablen auf irgendeinem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ mit

$$\mathbb{P}(X_n = e^n) = \frac{1}{n}, \quad \mathbb{P}(X_n = 0) = 1 - \frac{1}{n},$$

so gelten:

$\mathbf{X_n \xrightarrow{P} 0, n \to \infty}$, weil

$$\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{P}(X_n = e^n) = \frac{1}{n} \to 0, \quad n \to \infty.$$

$\mathbf{X_n \overset{\mathcal{L}^p}{\not\longrightarrow} 0, n \to \infty}$, für alle $p \geq 1$, weil

$$\mathbb{E}[|X_n - X|^p] = \mathbb{E}[|X_n|^p] = e^{pn}\frac{1}{n} + 0^p\Big(1 - \frac{1}{n}\Big) = \frac{e^{pn}}{n} \to +\infty, \quad n \to \infty.$$

Das nächste Beispiel ist sehr anschaulich. Dazu beachten wir, dass die Einschränkung des Lebesguemaßes auf $[0, 1]$ ein Wahrscheinlichkeitsmaß ist. Wir können also viele Beispiele basteln, wenn wir uns als Zufallsvariablen einfach messbare Funktionen (z. B. Indikatorfunktionen oder stetige Funktionen) auf $[0, 1]$ wählen. Das hat den Vorteil, dass die Begriffe durch Skizzen sehr anschaulich gemacht werden können. So ist zum Beispiel die fast sichere Konvergenz die punktweise Konvergenz (bis auf eine Nullmenge) und die $\mathcal{L}^p$-Konvergenz ist die Konverenz der Flächeninhalte der Differenzfunktion (hoch p) weil $\mathbb{E}[X] = \int_0^1 X(\omega)\,\mathrm{d}\omega$. Sieht wegen $\mathrm{d}\omega$ vielleicht blöd aus, ist aber einfach nur das ganz normale Integral auf $[0, 1]$.

Beispiel 4.5.4. Sei $\Omega = [0,1]$, $\mathcal{A} = \mathcal{B}([0,1])$ und $\mathbb{P} = \lambda_{[0,1]}$ das Lebesgue Maß auf $[0,1]$. Schauen wir uns als Beispiel $X \equiv 0$ (Nullfunktion) und die Folge

$$X_n = \mathbf{1}_{(\frac{m}{2^k}, \frac{m+1}{2^k}]}$$

an, wobei $m, k \in \mathbb{N}$ die eindeutigen natürlichen Zahlen mit $n = 2^k + m$ und $m < 2^k$ sind. In Worten (am besten skizziert ihr die Funktionen) schieben wir für wachsendes n einfach nur Indikatorfunktionen von links nach rechts durch $[0,1]$, wobei die Breite der Indikatorfunktionen schmaler wird: $\mathbf{1}_{(0,1]}, \mathbf{1}_{(0,\frac{1}{2}]}, \mathbf{1}_{(\frac{1}{2},1]}, \mathbf{1}_{(0,\frac{1}{4}]}, \mathbf{1}_{(\frac{1}{2},\frac{1}{4}]}$, ... Mit dieser Folge gelten:

$\mathbf{X_n \xrightarrow{\mathcal{L}^p} 0, n \to \infty}$, weil

$$\mathbb{E}[|X_n - X|^p] = \mathbb{E}[|X_n|^p] = \int_\Omega X_n^p(\omega)\,\mathrm{d}\mathbb{P}(\omega) = \int_0^1 \mathbf{1}_{(\frac{m}{2^k}, \frac{m+1}{2^k}]}(\omega)\,\mathrm{d}\omega = \frac{1}{2^k}.$$

Weil mit $n \to \infty$ auch $k \to \infty$ gilt, konvergiert die Folge also im p-ten Mittel gegen 0. Warum ist das auch anschaulich klar? $\mathbb{E}[|X_n|^p]$ ist der Flächeninhalt zwischen der Indikatorfunktion und der x-Achse. Weil die Breite des Indikators gegen 0 konvergiert, konvergiert das Integral und damit (in diesem Wahrscheinlichkeitsraum) der Erwartungswert gegen 0.

$\mathbf{X_n \overset{\text{f.s.}}{\not\longrightarrow} 0, n \to \infty}$, das ist klar, weil

$$\mathbb{P}(\{\omega \colon X_n(\omega) \to 0\}) = 0.$$

Man beachte: In diesem Wahrscheinlichkeitsraum ist die fast sichere Konvergenz gerade die punktweise Konvergenz auf einer Menge mit Maß 1. Weil die Folge (X_n) ausgewertet in beliebigem $\omega \in [0,1]$ unendlich oft zwischen 0 und 1 wechselt (1 wenn das kleine Intervall ω enthält, 0 sonst), konvergiert sie fast sicher nicht.

Beispiel 4.5.5. Wie im vorherigen Beispiel sei $\Omega = [0,1]$, $\mathcal{A} = \mathcal{B}([0,1])$ und $\mathbb{P} = \lambda_{|[0,1]}$, das Lebesguemaß auf $[0,1]$. Wir schauen uns die konkrete Folge

$$X_n = n \cdot \mathbf{1}_{[0,\frac{1}{n}]}$$

an, und schauen, in welchem Sinne sie gegen die Grenzzufallsvariable $X = 0$ (Nullfunktion) konvergiert.

$\mathbf{X_n \xrightarrow{\text{f.s.}} 0, n \to \infty}$: Das ist klar, weil in diesem Beispiel die fast sichere Konvergenz einfach nur die übliche punktweise Konvergenz von Funktionen auf einer Menge von Maß 1 bedeutet und unsere Folge auf $(0,1]$ punktweise gegen 0 konvergiert. Weil ein einzelner Punkt im Lebesguemaß eine Nullmenge ist, konvergiert die Folge fast sicher:

$$\mathbb{P}(X_n \to 0) = \mathbb{P}((0,1]) = 1.$$

$\mathbf{X_n \xrightarrow{P} 0, n \to \infty}$: Die stochastische Konvergenz gilt, weil für $\varepsilon > 0$

$$\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{P}(|X_n| > \varepsilon) = \mathbb{P}(n\mathbf{1}_{[0,\frac{1}{n}]} > \varepsilon) = \frac{1}{n} \to 0, \quad n \to \infty.$$

$\mathbf{X_n \overset{\mathcal{L}^p}{\not\longrightarrow} 0, n \to \infty}$: Weil wir nur Erwartungswerte von Indikatoren berechnen müssen, folgt alles direkt aus den Rechenregeln für Erwartungswerte:

$$\mathbb{E}[X_n - X|^p] = \mathbb{E}[n^p \cdot \mathbf{1}^p_{[0,\frac{1}{n}]}] = n^p\mathbb{E}[\mathbf{1}_{[0,\frac{1}{n}]}] = n^p\frac{1}{n} = n^{p-1} \not\to 0, \quad n \to \infty,$$

weil wir bei $\mathcal{L}^p$-Konvergenz immer $p \geq 1$ annehmen.

Beispiel 4.5.6. Sei $X \sim \mathcal{N}(0,1)$ und $X_n = (-1)^n X$ für $n \in \mathbb{N}$. Es gilt aufgrund der Symmetrie der Normalverteilung $X_n \sim \mathcal{N}(0,1)$ für alle $n \in \mathbb{N}$. Weil Erwartungswerte nur von der Verteilung abhängen, gilt also $\mathbb{E}[f(X)] = \mathbb{E}[f(X_n)]$ für alle $n \in \mathbb{N}$, die Folge der Erwartungswerte ist also konstant und konvergiert daher gegen $\mathbb{E}[f(X)]$. Also gilt $X_n \xrightarrow{(d)} X, n \to \infty$. Andere Konvergenzarten gelten für diese Folge nicht.

Beispiel 4.5.7. Seien $X_1, X_2, \ldots$ unabhängige Zufallsvariablen mit

$$\mathbb{P}(X_n = 1) = \frac{1}{n}, \quad \mathbb{P}(X_n = 0) = 1 - \frac{1}{n},$$

d. h. $X_n \sim \mathrm{Ber}(\frac{1}{n})$, $n \in \mathbb{N}$. Man stelle sich z. B. unabhängige Münzwürfe vor (1 bedeutet „Zahl“, 0 bedeutet „Kopf“), bei denen die Wahrscheinlichkeit für „Zahl“ immer kleiner wird. Alternativ kann man sich irgendwelche unabhängigen Versuche vorstellen, bei denen 1 „Erfolg“ und 0 „Misserfolg“ bedeutet. Mit dieser Folge gelten:

$\mathbf{X_n \xrightarrow{P} 0, n \to \infty}$: Für $\varepsilon > 0$ gilt

$$\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{P}(X_n > \varepsilon) = \begin{cases} \mathbb{P}(X_n = 1) & : \varepsilon \leq 1 \\ 0 & : \varepsilon > 1 \end{cases} = \begin{cases} \frac{1}{n} & : \varepsilon \leq 1 \\ 0 & : \varepsilon > 1 \end{cases} \to 0, \quad n \to \infty.$$

$\mathbf{X_n \overset{\text{f.s.}}{\not\longrightarrow} 0, n \to \infty}$: Das Argument ist nicht einfach, taucht aber im nächsten Kapitel mehrfach auf. Wir nutzen dazu eine Umformulierung der Konvergenz in Schnitte und Vereinigungen: Unter der Beachtung, dass eine Folge mit den Werten 0 und 1 nur gegen 0 konvergiert, wenn sie irgendwann nur noch den Wert 0 annimmt, ist das

$$\begin{aligned} \{\omega \in \Omega \,|\, X_n(\omega) \to 0\} &= \{\omega \in \Omega \,|\, \exists n_0 \in \mathbb{N} \colon X_n(\omega) = 0 \; \forall n \geq n_0\} \\ &= \bigcup_{n_0=1}^{\infty} \bigcap_{n \geq n_0} \{\omega \in \Omega \,|\, X_n(\omega) = 0\}. \end{aligned} \tag{4.2}$$

Zur Erinnerung, wie hatten wir Vereinigungen und Schnitte in Analysis 1 definiert?

$$\begin{aligned} \bigcup_{i \in I} A_i &:= \{\omega \in \Omega \,|\, \exists i \in I : \omega \in A_i\}, \\ \bigcap_{i \in I} A_i &:= \{\omega \in \Omega \,|\, \forall i \in I : \omega \in A_i\}. \end{aligned}$$

Diese kleine Überlegung ist unglaublich wichtig. Sie wird später der Weg sein, das starke Gesetz der großen Zahlen zu beweisen. Damit kann fast sichere Konvergenz immer in Vereinigungen über Schnitte umformuliert werden und diese können wir immer mit Subadditivität und Stetigkeit von Maßen attackieren. Mit (4.2) bekommen wir also

$$\begin{aligned} \mathbb{P}(X_n \to 0) &= \mathbb{P}\Big(\bigcup_{n_0=1}^{\infty} \bigcap_{n \geq n_0} \{\omega \in \Omega \,|\, X_n(\omega) = 0\} \Big) \\ &\overset{\text{Subadd.}}{\leq} \sum_{n_0=1}^{\infty} \mathbb{P}\Big(\bigcap_{n \geq n_0} \{\omega \in \Omega \,|\, X_n(\omega) = 0\} \Big) \\ &\overset{\text{Stet. Maße}}{=} \sum_{n_0=1}^{\infty} \lim_{m \to \infty} \mathbb{P}\Big(\bigcap_{n=n_0}^{m} \{\omega \in \Omega \,|\, X_n(\omega) = 0\} \Big) \\ &\overset{\text{unab.}}{=} \sum_{n_0=1}^{\infty} \lim_{m \to \infty} \prod_{n=n_0}^{m} \mathbb{P}(X_n = 0) \\ &= \sum_{n_0=1}^{\infty} \lim_{m \to \infty} \Big(1 - \frac{1}{n_0}\Big) \cdots \Big(1 - \frac{1}{m}\Big) \\ &= \sum_{n_0=1}^{\infty} \lim_{m \to \infty} \Big(\frac{n_0 - 1}{n_0}\Big) \cdots \Big(\frac{m-1}{m}\Big) \\ &\overset{\text{Kürzen}}{=} \sum_{n_0=1}^{\infty} \lim_{m \to \infty} \frac{n_0 - 1}{m} = 0. \end{aligned}$$

Folglich ist die Wahrscheinlichkeit der Konvergenz sogar 0 und daher gilt fast sichere Konvergenz nicht. Genau das gleiche Argument wird später bei dem Borel-Cantelli Lemma noch einmal auftauchen.

Mit den Beispielen haben wir schon ein paar Beispiele gesammelt, die zeigen, dass die Konvergenzarten nicht äquivalent sein können. Den Zusammenhang der Konvergenzarten wollen wir nun genauer beleuchten.

Vorlesung 25

Zunächst schauen wir die Konvergenz in Verteilung etwas genauer an. Dazu zunächst ein Hilfslemma:

Lemma 4.5.8. Seien $F, F_1, F_2, \ldots$ Verteilungsfunktionen. Wenn $\lim_{n\to\infty} F_n(t) = F(t)$ für alle Stetigkeitsstellen t von F gilt, so gilt $\lim_{n\to\infty} F_n^{-1}(y) = F^{-1}(y)$ für alle Stetigkeitsstellen von F^{-1}.

Beweis. Der Beweis basiert hauptsächlich auf einer Eigenschaft der Pseudoinversen, die wir aus Lemma 4.3.2 (iii) schon kennen. Wir nutzen die Eigenschaft aber auch als strikte Abschätzung in die andere Richtung:

$$a < F^{-1}(u) \leq b \Leftrightarrow F(a) < u \leq F(b) \quad \text{für alle } a, b \in \mathbb{R} \text{ und } u \in (0,1). \tag{4.3}$$

Die Aussage folgt genau wie in Lemma 4.3.2 direkt aus der Definition von F^{-1}. Wir wählen jetzt Stetigkeitsstellen a und b von F mit $a < F^{-1}(y) < b$ und ein $v \in (y, 1)$ mit $F^{-1}(v) < b$. Damit es so ein v gibt, nutzen wir die Stetigkeit von F^{-1} an der Stelle y (das ist die $\varepsilon - \delta$ Definition für $\varepsilon = b - F^{-1}(y)$). Jetzt formen wir mit (4.3) um und nutzen die Definition der Folgenkonvergenz:

$$\begin{aligned} & a < F^{-1}(y) \leq F^{-1}(v) < b \\ \overset{(4.3)}{\Longrightarrow} \quad & F(a) < y < v \leq F(b) \\ \Longrightarrow \quad & \exists N \in \mathbb{N} : F_n(a) < y < F_n(b) \text{ für alle } n \geq N \\ \overset{(4.3)}{\Longrightarrow} \quad & \exists N \in \mathbb{N} : a < F_n^{-1}(y) \leq b \text{ für alle } n \geq N \end{aligned}$$

Wir wissen noch nicht, dass die Folge $(F_n^{-1}(y))_{n\in\mathbb{N}}$ konvergiert, nutzen also $\liminf$ und $\limsup$, um sauber zu formulieren: Da steht also

$$a \leq \liminf_{n\to\infty} F_n^{-1}(y) \leq \limsup_{n\to\infty} F_n^{-1}(y) \leq b.$$

Weil a und b beliebig nah an $F^{-1}(y)$ gewählt werden können (es gibt nur abzählbar viele Unstetigkeitsstellen von F^{-1}), gilt also $\liminf_{n\to\infty} F_n^{-1}(y) = \limsup_{n\to\infty} F_n^{-1}(y) = F^{-1}(y)$. Damit existiert der Grenzwert und es gilt $\lim_{n\to\infty} F_n^{-1}(y) = F^{-1}(y)$. □

Das Lemma ist etwas technisch, daraus folgt aber folgende äußerst wichtige Aussage. Wir hätten die Konvergenz in Verteilung auch durch Konvergenz der Verteilungsfunktionen definieren können:

Satz 4.5.9. Seien $X, X_1, X_2, \ldots$ Zufallsvariablen mit $X \sim F$ und $X_n \sim F_n$, $n \in \mathbb{N}$, so gilt:

$$X_n \xrightarrow{(d)} X, \quad n \to \infty \quad \Longleftrightarrow \quad \lim_{n\to\infty} F_n(t) = F(t) \quad \text{für alle Stetigkeitsstellen von } F.$$

Beweis.

„⇒“: Weil die Konvergenz in Verteilung gerade $\lim_{n\to\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)]$ bedeutet, sind wir schon fast fertig. Wir würden gerne $f = \mathbf{1}_{(-\infty,t]}$ nutzen, denn dann hätten wir schon die Konvergenz der Verteilungsfunktionen an der Stelle t. Leider geht das nicht, f ist zwar beschränkt, jedoch nicht stetig. Um das zu Umschiffen, wählen wir für beliebiges $\delta > 0$ folgende Modifikation: Wir nehmen die Funktionen f_+ und f_-, die wie folgt definiert sind: f_+ ist der Indikator auf $(-\infty, t]$ und verbindet dann linear 1 und 0 auf $[t, t+\delta]$, rechts von

δ ist f_+ null. f_- ist ähnlich, verbindet aber auf $[t-\delta, t]$ linear die 1 und die 0 (Bildchen malen!). Es gilt also

$$\mathbf{1}_{(-\infty,t-\delta]} \leq f_- \leq f \leq f_+ \leq \mathbf{1}_{(-\infty,t+\delta]}$$

und f_-, f_+ sind beschränkt und stetig weil es einfach Indikatorfunktionen sind, die die Unstetigkeitsstelle linear stetig machen. Wir können also die angenommene Konvergenz für f_+ und f_- anwenden.

Mit Monotonie des Erwartungswertes gilt

$$F_n(t) = \mathbb{E}[\mathbf{1}_{(-\infty,t]}(X_n)] \leq \mathbb{E}[f_+(X_n)] \overset{f_+ \text{ stet.}}{\longrightarrow} \mathbb{E}[f_+(X)] \leq \mathbb{E}[\mathbf{1}_{(-\infty,t+\delta]}(X)] = F(t+\delta).$$

Analog mit f_- schätzen wir nach unten ab:

$$F_n(t) = \mathbb{E}[\mathbf{1}_{(-\infty,t]}(X_n)] \geq \mathbb{E}[f_-(X_n)] \overset{f_- \text{ stet.}}{\longrightarrow} \mathbb{E}[f_-(X)] \geq \mathbb{E}[\mathbf{1}_{(-\infty,t-\delta]}(X)] = F(t-\delta).$$

Weil $\delta > 0$ beliebig war, gilt also $\lim_{n\to\infty} F_n(t) = F(t)$. An dieser letzten Stelle haben wir die Annahme benutzt, dass t eine Stetigkeitsstelle von F ist!

„$\Leftarrow$“: Sei $U \sim \mathcal{U}((0,1))$ durch die kanonische Konstruktion konstruiert. Der Wahrscheinlichkeitsraum ist also $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathcal{U}((0,1))$ und die Zufallsvariable die Identitätsabbildung. Definiere

$$Y = F^{-1}(U) \quad \text{und} \quad Y_n = F_n^{-1}(U), \quad n \in \mathbb{N}.$$

Wegen der inversen Transformations Methode gilt

$$Y \sim F \quad \text{und} \quad Y_n \sim F_n, \quad n \in \mathbb{N}.$$

Aufgrund der Voraussetzung und Lemma 4.5.8 konvergiert $F_n^{-1}(t)$ für alle Stetigkeitsstellen von F^{-1} gegen $F^{-1}(t)$. Die Unstetigkeitsstellen der nicht-fallenden Funktion F^{-1} sind abzählbar, also eine Nullmenge in dem gewählten Wahrscheinlichkeitsraum (abzählbare Mengen sind Nullmengen für das Lebesguemaß). Weil U die Identitätsabbildung ist, gilt also

$$Y_n = F_n^{-1}(U) \overset{\text{f.s.}}{\to} F^{-1}(U) = Y, \quad n \to \infty.$$

Wir können in diesem Fall sogar das Ereigniss von Maß 1 benennen, auf dem die Konvergenz gilt: Das sind gerade die Stetigkeitsstellen von F^{-1} in $(0,1)$. Weil Erwartungswerte von identisch verteilten Zufallsvariablen gleich sind (Lemma 4.1.10 und die Bemerkung danach), gilt also für f stetig und beschränkt

$$\begin{aligned}\lim_{n\to\infty} \mathbb{E}[f(X_n)] &\overset{X_n \sim Y_n}{=} \lim_{n\to\infty} \mathbb{E}[f(Y_n)] \overset{3.2.5}{=} \mathbb{E}[\lim_{n\to\infty} f(Y_n)] \\ &\overset{Y_n \to Y \text{ f.s.}}{=} \mathbb{E}[f(Y)] \overset{X \sim Y}{=} \mathbb{E}[f(X)].\end{aligned}$$

Damit ist die Konvergenz in Verteilung bewiesen.

□

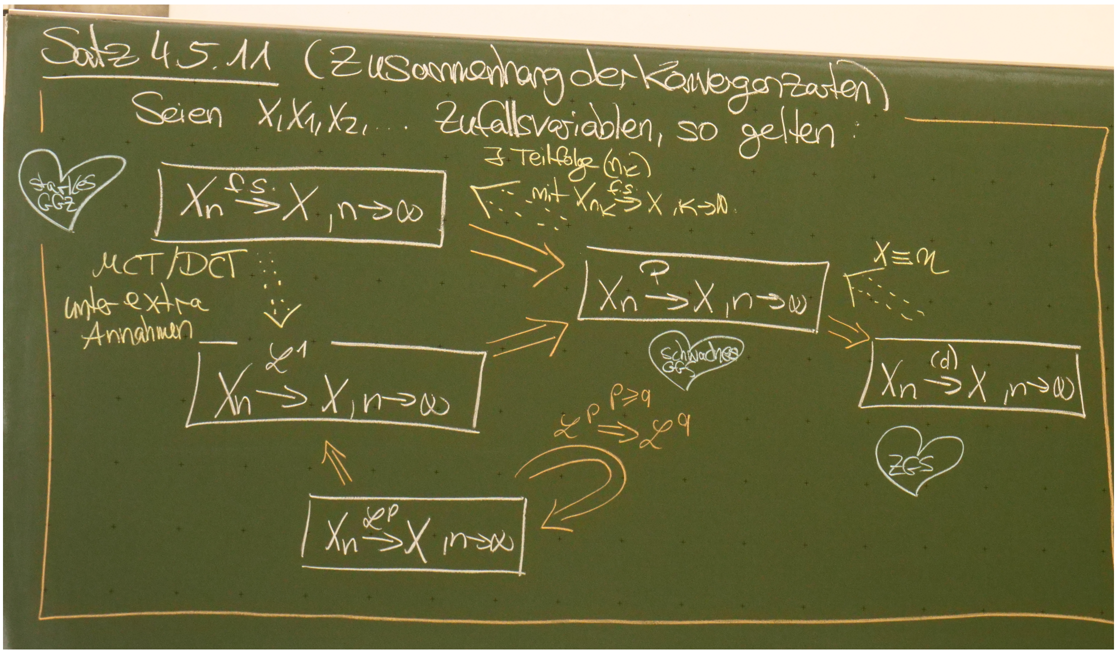
Im Prinzip sagt die Aussage, dass schwache Konvergenz (fast) das gleiche ist, wie punktweise Konvergenz der Verteilungsfunktionen. Die Stetigkeitsstellen sind nur für die Grenzverteilungsfunktion gefordert. In vielen Fällen, z. B. wenn die Grenzverteilung eine Normalverteilung ist, ist F stetig. In dem Fall ist schwache Konvergenz nichts anderes als die punktweise Konvergenz der Verteilungsfunktionen. Das taucht später zum Beispiel beim zentralen Grenzwertsatz noch auf.

Bemerkung 4.5.10. Die Konstruktion im Beweis heißt Skorokhod-Kopplung: Für jede Folge $X_1, \ldots$ von Zufallsvariablen (auf beliebigen Wahrscheinlichkeitsräumen), die schwach gegen eine Zufallsvariable X konvergiert, gibt es einen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ mit Zufallsvariablen $Y, Y_1, \ldots$, so dass

- $X \sim Y$, $X_n \sim Y_n$ für alle $n \in \mathbb{N}$,
- $Y_n \xrightarrow{\text{f.s.}} Y$, $n \to \infty$.

Die Skorokhod-Kopplung ist besonders spektakulär, wenn man mit dem nächsten Satz vergleicht, in dem der Zusammenhang der Konvergenzarten beleuchtet wird:

Satz 4.5.11. **[Zusammenhang Konvergenzarten]**



Beweis. **Konvergenz in $\mathcal{L}^p$ $\Rightarrow$ Konvergenz in $\mathcal{L}^q$ für $q < p$:**
Wir wissen schon, wie aus Hölder $\mathbb{E}[|X|^q] \leq \mathbb{E}[|X|^p]$ folgt, man fügt einfach eine 1 hinzu. Als Wiederholung machen wir das nochmal: Definiere dazu $r = \frac{p}{q}$ und $r' = \frac{r}{r-1}$, also gilt

$$\mathbb{E}[|X_n - X|^q] = \mathbb{E}[|X_n - X|^q \cdot 1] \leq (\mathbb{E}[|X_n - X|^{qr}])^{1/r}(\mathbb{E}[1^{r'}])^{1/r'} = \mathbb{E}[|X_n - X|^p]^{q/p} \cdot 1.$$

Wenn die rechte Seite gegen 0 konvergiert, konvergiert also auch die linke Seite gegen 0. Das ist genau das, was wir zeigen wollten.
Konvergenz in $\mathcal{L}^1$ $\Rightarrow$ Stochastische Konvergenz:
Mit der Markov Ungleichung gilt:

$$\mathbb{P}(|X_n - X| > \varepsilon) \leq \frac{\mathbb{E}[|X_n - X|]}{\varepsilon} \overset{\text{Vor.}}{\to} 0, \quad n \to \infty.$$

Das war es schon.
Fast sichere Konvergenz $\Rightarrow$ Stochastische Konvergenz:
Der Trick ist es, die Definition der Konvergenz aus Analysis 1 als Schnitte und Vereinigungen von Mengen zu schreiben:

$$\begin{aligned} 1 &= \mathbb{P}(X_n \to X) \\ &\overset{\text{Notation}}{=} \mathbb{P}(\{\omega \colon X_n(\omega) \to X(\omega), n \to \infty\}) \\ &\overset{\text{Def.}}{=} \mathbb{P}(\{\omega \colon \forall \varepsilon > 0 \exists N \in \mathbb{N} \colon |X_n(\omega) - X(\omega)| < \varepsilon \,\forall n \geq N\}) \\ &= \mathbb{P}\Big(\bigcap_{\varepsilon > 0} \bigcup_{N \in \mathbb{N}} \bigcap_{n \geq N} \{\omega \colon |X_n(\omega) - X(\omega)| < \varepsilon\}\Big). \end{aligned}$$

Mit Komplementbildung und de Morgan'schen Regeln folgt daraus

$$\begin{aligned}
0 &= \mathbb{P}\Big(\Big(\bigcap_{\varepsilon>0}\bigcup_{N\in\mathbb{N}}\bigcap_{n\geq N}\{\omega\colon |X_n(\omega)-X(\omega)|<\varepsilon\}\Big)^C\Big)\\
&= \mathbb{P}\Big(\bigcup_{\varepsilon>0}\bigcap_{N\in\mathbb{N}}\bigcup_{n\geq N}\{\omega\colon |X_n(\omega)-X(\omega)|\geq\varepsilon\}\Big)\\
&\overset{\text{Mon. } \varepsilon \text{ fest}}{\geq} \mathbb{P}\Big(\bigcap_{N\in\mathbb{N}}\bigcup_{n\geq N}\{\omega\colon |X_n(\omega)-X(\omega)|\geq\varepsilon\}\Big)\\
&\overset{\text{Mon. Maße}}{=} \lim_{N\to\infty}\mathbb{P}\Big(\bigcup_{n\geq N}\{\omega\colon |X_n(\omega)-X(\omega)|\geq\varepsilon\}\Big)\\
&\overset{\text{Mon.}}{\geq} \lim_{N\to\infty}\mathbb{P}(\{\omega\colon |X_n(\omega)-X(\omega)|>\varepsilon\})\geq 0.
\end{aligned}$$

Also gilt $\lim_{n\to\infty}\mathbb{P}(|X_n-X|>\varepsilon)=0$ für alle $\varepsilon>0$. Das ist gerade die Definition der stochastische Konvergenz und wir sind fertig.

Stochastische Konvergenz $\Rightarrow$ Konvergenz in Verteilung:
Wir zeigen die Aussage in zwei Schritten, erst für gleichmäßig stetiges f, dann für stetiges f.

(i) Sei f gleichmäßig stetig, d. h.

$$\forall\varepsilon>0\exists\delta>0\colon |x-x'|<\delta\Rightarrow|f(x)-f(x')|<\varepsilon. \tag{4.4}$$

Sei nun $\varepsilon>0$ beliebig und δ dazugehörig aus (4.4). Definiere $A_n=\{\omega\colon |X_n(\omega)-X(\omega)|\geq\delta\}$. Damit gilt, mit $||f||_\infty=\sup_{x\in\mathbb{R}}|f(x)|$,

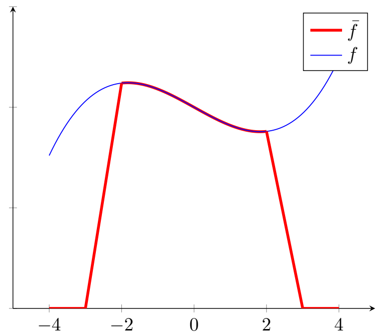
$$\begin{aligned}
0 &\leq |\mathbb{E}[f(X_n)]-\mathbb{E}[f(X)]|\\
&\overset{\Delta}{\leq} \mathbb{E}[|f(X_n)-f(X)|]\\
&= \mathbb{E}[|f(X_n)-f(X)|\cdot\underbrace{1}_{\mathbf{1}_{A_n}+\mathbf{1}_{A_n^C}}|]\\
&= \mathbb{E}[|f(X_n)-f(X)|\cdot\mathbf{1}_{A_n}]+\mathbb{E}[|f(X_n)-f(X)|\cdot\mathbf{1}_{A_n^C}]\\
&\overset{(4.4),\text{ Def. } A_n}{\leq} \mathbb{E}[2||f||_\infty\mathbf{1}_{A_n}]+\mathbb{E}[\varepsilon\mathbf{1}_{A_n^C}]\\
&= 2||f||_\infty\mathbb{P}(A_n)+\varepsilon\mathbb{P}\big(A_n^C\big)\\
&\leq 2||f||_\infty\mathbb{P}(A_n)+\varepsilon\\
&= 2||f||_\infty\mathbb{P}(|X_n-X|>\delta)+\varepsilon\\
&\overset{\text{Vor.}}{\to} \varepsilon,\quad n\to\infty.
\end{aligned}$$

Weil $\varepsilon>0$ beliebig, gilt damit $\lim_{n\to\infty}|\mathbb{E}[f(X_n)]-\mathbb{E}[f(X)]|=0$ und somit

$$\lim_{n\to\infty}\mathbb{E}[f(X_n)]=\mathbb{E}[f(X)].$$

Damit ist die Definition der Konvergenz in Verteilung für gleichmäßig stetige beschränkte Funktionen gezeigt.

(ii) Sei jetzt f eine beliebige stetige beschränkte Funktion und $\varepsilon>0$ fest. Für Intervalle $[-k,k]$ gilt wegen der Stetigkeit von Maßen $\mathbb{P}(X\notin[-k,k])\to 0$, $k\to\infty$. Daher können wir ein $k\in\mathbb{N}$ mit $\mathbb{P}(X\notin[-k+1,k-1])<\varepsilon$ wählen. Wir definieren jetzt die stetige Funktion $\bar{f}$ wie in dem Bildchen, in dem $k=2$ gewählt ist.



In Worten: $\bar{f}$ ist gleich f in $[-k,k]$, null außerhalb von $[-k-1,k+1]$ und verbindet dazwischen linear 0 und $f(k)$ bzw. $f(-k)$. Die wichtige dazu gewonnene Information ist, dass $\bar{f}$ gleichmäßig stetig ist. Das gilt, weil stetige Funktionen auf kompakten Mengen gleichmäßig stetig sind (siehe Analysis 1). Jetzt kommt ein mehrfach genutzter Trick aus der Analysis. Wir addieren zwei Mal 0 und nutzen die Dreicksungleichung

$$|\mathbb{E}[f(X_n)] - \mathbb{E}[f(X)]| \overset{\triangle}{\leq} \underbrace{\mathbb{E}[|f(X_n) - \bar{f}(X_n)|]}_{:=\,\mathrm{I}_n} + \underbrace{\mathbb{E}[|\bar{f}(X_n) - \bar{f}(X)|]}_{:=\,\mathrm{II}_n} + \underbrace{\mathbb{E}[|\bar{f}(X) - f(X)|]}_{:=\,\mathrm{III}_n} \quad (4.5)$$

und betrachten einzeln die Grenzwerte der drei Summanden. Aus (i), und weil $\bar{f}$ gleichmäßig stetig, wissen wir, dass $\mathrm{II}_n \to 0$ für $n \to \infty$. Für den dritten Summanden gilt

$$\begin{aligned}
\mathrm{III}_n &= \mathbb{E}[|f(X) - \bar{f}(X)| \cdot (\mathbf{1}_{[-k,k]}(X) + \mathbf{1}_{[-k,k]^C}(X))] \\
&\leq 0 + \mathbb{E}[2||f||_\infty \mathbf{1}_{[-k,k]^C}(X)] \\
&= 2||f||_\infty \mathbb{P}(X \in [-k,k]^C) \\
&= 2||f||_\infty \mathbb{P}(X \notin [-k,k]) \\
&\overset{\text{Mon.}}{\leq} 2||f||_\infty \mathbb{P}(X \notin [-k+1,k-1]) < 2||f||_\infty \varepsilon,
\end{aligned}$$

wegen der Wahl von k und weil $f(x) - \bar{f}(x) = 0$ für $x \in [-k,k]$. Schließlich noch der erste Summand. Wegen der angenommenen stochastischen Konvergenz gilt

$$0 \leq \mathbb{P}(X_n \notin [-k,k], X \in [-k+1,k-1]) \leq \mathbb{P}(|X_n - X| > 1) \to 0, \quad n \to \infty.$$

Damit zerlegen wir wie folgt:

$$\begin{aligned}
\mathrm{I}_n &= \mathbb{E}[|f(X_n) - \bar{f}(X_n)| \cdot (\mathbf{1}_{[-k,k]}(X_n) + \mathbf{1}_{[-k,k]^C}(X_n)] \\
&\leq 0 + \mathbb{E}[2||f||_\infty \mathbf{1}_{[-k,k]^C}(X_n)] \\
&= 2||f||_\infty \mathbb{P}(X_n \notin [-k,k]) \\
&= 2||f||_\infty \big(\mathbb{P}(X_n \notin [-k,k], X \in [-k+1,k-1]) + \mathbb{P}(X_n \notin [-k,k], X \notin [-k+1,k-1])\big) \\
&\leq 2||f||_\infty \big(\mathbb{P}(|X_n - X| \geq 1) + \mathbb{P}(X \notin [-k+1,k-1])\big) \\
&\leq 2||f||_\infty \big(\mathbb{P}(|X_n - X| \geq 1) + \varepsilon\big).
\end{aligned}$$

Die rechte Seite konvergiert dann gegen $2||f||_\infty \varepsilon$. In der Rechnung haben wir viele kleine Eigenschaften benutzt, checkt es mal selber Zeile für Zeile: Monotonie und Linearität von Erwartungswerten, dass Erwartungswerte von Indikatoren Wahrscheinlichkeiten sind (siehe Proposition 4.1.14), sowie die σ-Additivität und Monotonie von Maßen.

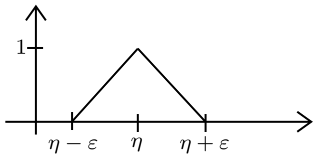
Die drei einzelnen Betrachtungen zusammen ergeben wegen (4.5)

$$0 \leq \limsup_{n\to\infty} |\mathbb{E}[f(X_n)] - \mathbb{E}[f(X)]| \leq 2||f||_\infty \varepsilon + 0 + 2||f||_\infty \varepsilon.$$

Weil ε beliebig war, folgt daraus $\limsup_{n\to\infty} |\mathbb{E}[f(X_n)] - \mathbb{E}[f(X)]| = 0$ und damit $\lim_{n\to\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)]$. Das ist die Konvergenz in Verteilung.

Die Hauptrichtungen sind nun vollständig bewiesen. Wir zeigen jetzt noch, dass eine Umkehrung gilt, wenn der Grenzwert eine konstante Zufallsvariable ist:

Konvergenz in Verteilung gegen eine <u>konstante</u> Zufallsvariable $\Rightarrow$ Stochastische Konvergenz:
Sei nun (X_n) eine Folge von Zufallsvariablen, die gegen eine fast sicher konstante Zufallsvariable X in Verteilung konvergiert. Sei $\eta \in \mathbb{R}$ der Wert, den X $\mathbb{P}$-fast sicher annimmt und sei $\varepsilon > 0$ beliebig. Sei nun $I = [\eta - \varepsilon, \eta + \varepsilon]$ und f eine stetige Funktion mit $f \leq \mathbf{1}_I$ sowie $f(\eta) = 1$. Natürlich gibt es so ein f, zum Beispiel das aus folgendem Bildchen:



Wegen der angenommenen Konvergenz in Verteilung gilt damit

$$\begin{aligned} 1 &= f(\eta) \\ &\overset{4.1.14\text{ (iii)}}{=} \mathbb{E}[f(X)] \\ &= \lim_{n\to\infty} \mathbb{E}[f(X_n)] \\ &\overset{\text{Monotonie}}{\leq} \limsup_{n\to\infty} \mathbb{E}[\mathbf{1}_I(X_n)] \\ &\overset{4.1.14\text{ (iv)}}{=} \limsup_{n\to\infty} \mathbb{P}(X_n \in I) \leq 1, \end{aligned}$$

weil Wahrscheinlichkeiten immer durch 1 dominiert sind. Weil obere und untere Schranke gleich sind, bekommen wir $\lim_{n\to\infty} \mathbb{P}(X_n \in I) = 1$ und damit

$$\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{P}(|X_n - \eta| > \varepsilon) = \mathbb{P}(X_n \notin I) = 1 - \mathbb{P}(X_n \in I) \to 0, \quad n \to \infty.$$

Warum tauchte gerade der Limes superior auf? Das liegt einfach nur daran, dass wir nicht wissen, ob der Grenzwert der Erwartungswerte existiert. Wir wissen das zwar für stetige Funktionen, aber der Indikator ist nicht stetig.

Stochastische Konvergenz $\Rightarrow$ Fast sichere Konvergenz einer Teilfolge:
Den Teil liefern wir nach dem Borel-Cantelli Lemma nach, siehe die Bemerkung 4.6.6. □

Aus dem Konvergenzdiagramm folgt natürlich, dass fast sichere Konvergenz auch Konvergenz in Verteilung impliziert. Überlegt doch mal, warum diese Aussage sehr viel schneller auch ohne den Umweg über die stochastische Konvergenz folgt (Stichwort dominierte Konvergenz).

Das Ziel ist immer, die „starken" Konvergenzen (links im Bildchen) zu zeigen. Das geht leider nicht immer! Um die Begriffe mit Leben zu füllen, zeigen wir drei berühmte Sätze, je einen für stochastische Konvergenz, fast sichere Konvergenz und Konvergenz in Verteilung:

- schwaches Gesetz der großen Zahlen (stochastische Konvergenz)
- starkes Gesetz der großen Zahlen (fast sichere Konvergenz)
- Zentraler Grenzwertsatz (Konvergenz in Verteilung)

Beim schwachen Gesetz der großen Zahlen werden wir merken, dass Konvergenz in $\mathcal{L}^p$ gerade für $p = 1$ oder $p = 2$ ein extrem nützliches Werkzeug ist. Der Grund ist, dass man mit Momenten gut rumrechnen kann (Ausmultiplizieren, Linearität, ...). Folgendes Resultat ist nicht sehr kompliziert, kann aber schon nützlich sein. Sehr viel nützlicher wird nächste Woche aber das starke Gesetzt der großen Zahlen sein!

Satz 4.5.12. **[Schwaches Gesetz der großen Zahlen]**

(i) Klassische Variante: Sind $X_1, X_2, \ldots$ u.i.v. mit $\mathbb{E}[X_1^2] < \infty$, so gilt

$$\frac{1}{n}\sum_{k=1}^n X_k \xrightarrow{P} \mathbb{E}[X_1], \quad n \to \infty.$$

(ii) Variante mit schwächeren Annahmen: Sind $X_1, X_2, \ldots$ quadratintegrierbar, paarweise unkorreliert (z. B. paarweise unabhängig) mit identischen Erwartungswert und

$$\frac{1}{n^2}\sum_{k=1}^n \mathbb{V}(X_k) \to 0, \quad n \to \infty, \tag{4.6}$$

so gilt

$$\frac{1}{n}\sum_{k=1}^n X_k \xrightarrow{P} \mathbb{E}[X_1], \quad n \to \infty.$$

Fürs bessere Verständniss kann man überlegen, wie die Annahme „identischer Erwartunswert“ in (ii) ersetzt werden kann, da gibt es verschiedene Möglichkeiten.

Beweis. (i) Wenn man den Beweis gesehen hat, ist alles ziemlich simple: **Erst Tschebycheff, dann Bienaymé**. Mit Tschebycheff und Bienaymé gilt für beliebiges $\varepsilon > 0$ aufgrund der u.i.v. Annahme

$$\begin{aligned}\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{k=1}^n X_k - \mathbb{E}[X_1]\Big| \geq \varepsilon\Big) &= \mathbb{P}\Big(\Big|\sum_{k=1}^n X_k - n\mathbb{E}[X_1]\Big| \geq n\varepsilon\Big)\\ &\overset{4.1.22}{\leq} \frac{\mathbb{V}\big[\sum_{k=1}^n X_k\big]}{n^2\varepsilon^2}\\ &\overset{4.2.31}{=} \frac{\sum\limits_{k=1}^n \mathbb{V}[X_k]}{n^2\varepsilon^2}\\ &= \frac{n\cdot\mathbb{V}[X_1]}{n^2\varepsilon^2} \to 0, \quad n\to\infty.\end{aligned}$$

Also gilt $\frac{1}{n}\sum_{k=1}^n X_k \xrightarrow{P} \mathbb{E}[X_1]$ für $n \to \infty$. Beachtet dabei, dass wir für Tschebycheff $\mathbb{E}[\sum_{k=1}^n X_k] = n\mathbb{E}[X_1]$ genutzt haben. Um ganz genau zu sein (und wir wollen natürlich genau sein!), merken wir noch an, dass „$>$“ oder „$\geq$“ in der Definition der schwachen Konvergenz natürlich keine Rolle spielt, ε ist schließlich beliebig.

Man kann das Argument auch anders hinschreiben. Benutzt zum Üben doch mal statt Tschebycheff die Markov Ungleichung mit $h(x) = x^2$ plus die Verschiebungsformel für die Varianz.

(ii) Um die schwächeren Annahmen von (ii) zu verstehen, brauchen wir nur in den vier Zeilen des Arguments zu schauen, wie wir die Unabhängigkeit abschwächen können, so dass die Konvergenz immer noch folgt. Wir sehen dann sofort, dass für Bienaymé nur paarweise unkorreliert gebraucht wird, dann aber für die Konvergenz noch (4.6) gefordert werden muss (es wird nicht mehr identisch verteilt angenommen, es gilt also nicht $\mathbb{V}[X_k] = \mathbb{V}[X_1]$). Schaut euch das in Ruhe an, um die Idee „erst Tschebyscheff, dann Bienaymé“ einzubrennen. □

Zum besseren Verständnis kann man mal ausprobieren, ob man das schwache Gesetz genauso mit der Annahme $\mathbb{E}[|X_1|] < \infty$ beweisen könnte. Warum funktioniert der Beweis nicht, wenn man die Markovungleichung für das erste Moment statt für das zweite Moment ausprobiert? Der Trick an dem Beweis mit zweiten Momenten ist, dass die Varianz der Summe, die eigentlich aus n^2 vielen Summanden besteht, sich aufgrund der Annahme (unabhängig oder unkorreliert) zu einer Summe aus nur n vielen Summanden reduziert (im Beweis von Bienaymé kürzen sich die meisten Terme raus). Damit dominiert der Nenner mit n^2 und die obere Schranke konvergiert gegen 0. Der Effekt passiert beim ersten Moment nicht, weil keine „gemischten Terme“ $\mathbb{E}[X_iX_j] = 0$

auftauchen. Deshalb stünde sowohl in Zähler als auch in Nenner etwas mit n, die obere Schranke würde also nicht gegen 0 konvergieren, in Formeln (für $\mathbb{E}[X_1] = 0$):

$$\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{k=1}^{n} X_k\Big| \geq \varepsilon\Big) \overset{4.1.22,\, h(x)=|x|}{\leq} \frac{\mathbb{E}\big[\big|\sum_{k=1}^{n} X_k\big|\big]}{n\varepsilon} \leq \frac{\sum_{k=1}^{n}\mathbb{E}[|X_k|]}{n\varepsilon} = \frac{n\mathbb{E}[|X_1|]}{n\varepsilon} \not\to 0,$$

für $n \to \infty$. Genau den selben Effekt werden wir beim Beweis des starken Gesetzes der großen Zahlen sehen, bei dem wir endliche 4.te Momente annehmen und bei der Markovungleichung mit $h(x) = x^4$ genug Summanden verschwinden, dass der Nenner dominiert.

4.6 Starkes Gesetz der großen Zahlen

Vorlesung 26

Was bedeutet die stochastische Konvergenz, bzw. warum ist sie schwach? Seien als Beispiel $X_1, X_2, ...$ u.i.v. Würfel, also diskret gleichverteilt auf $\{1, ..., 6\}$. Wegen $\mathbb{E}[X_1] = 3,5$ bedeutet das schwache Gesetz der großen Zahlen (wähle zum Beispiel $\varepsilon = 0.01$), dass

$$\mathbb{P}\Big(3,49 < \frac{1}{n}\sum_{i=1}^{n} X_i < 3,51\Big) = \mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X_1]\Big| < 0,01\Big) \to 1, \quad n \to \infty.$$

In Worten steht hier: „Der Mittelwert wird mit hoher Wahrscheinlichkeit nah bei 3,5 liegen, wenn n groß ist.“ Es wird damit nicht ausgeschlossen, dass der Mittelwert mit kleiner Wahrscheinlichkeit weit vom Erwartungswert entfernt liegt. Genau hier liegt der Unterschied zum starken Gesetz der großen Zahlen, das wir als nächstes beweisen wollen. Hier wird die stochastische Konvergenz durch fast sichere Konvergenz ersetzt. Weil in der Definition der fast sicheren Konvergenz alle n gemeinsam *innerhalb* der Wahrscheinlichkeit auftauchen, $\mathbb{P}(X_n \to X) = 1$, kann der Effekt nicht auftreten.

Als Hilfsmittel für den Beweis diskutieren wir zunächst das Borel-Cantelli Lemma. Dazu zunächst ein paar Definitionen:

Definition 4.6.1. Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und seien $A_1, A_2, ... \in \mathcal{A}$ beliebige Ereignisse.

(i)

$$\begin{aligned}\limsup_{n\to\infty} A_n &:= \bigcap_{n=1}^{\infty}\bigcup_{k=n}^{\infty} A_k \\ &= \{\omega \in \Omega \colon \omega \in A_n \text{ für unendlich viele } n\} \\ &\overset{\text{Notation}}{=} \{A_n \text{ unendlich oft}\}\end{aligned}$$

heißt **Limes superior** der Folge (A_n) von Ereignissen.

(ii)

$$\begin{aligned}\liminf_{n\to\infty} A_n &:= \bigcup_{n=1}^{\infty}\bigcap_{k=n}^{\infty} A_k \\ &= \{\omega \in \Omega \colon \omega \in A_n \text{ schließlich immer}\} \\ &\overset{\text{Notation}}{=} \{A_n \text{ schließlich immer}\}\end{aligned}$$

Limes inferior der Folge (A_n) von Ereignissen.

Weil aufgrund der Definition einer σ-Algebra abzählbare Schnitte und Vereinigungen wieder in $\mathcal{A}$ sind, sind auch $\limsup_{n\to\infty} A_n$ und $\liminf_{n\to\infty} A_n$ in $\mathcal{A}$. Wem der Begriff „schließlich immer“ suspekt ist, der oder die schaue einfach die formelle Definition an. Diese besagt, dass es eine natürliche Zahl n gibt, so dass das Ereigniss danach *immer* eintritt (der Durchschnitt von Mengen enthält alle Elemente, die in allen Mengen enthalten sind).

Tatsächlich haben die neuen Begriffe lim inf und lim sup für Mengen auch etwas mit den uns bekannten Begriffen lim inf und lim sup für Folgen zu tun:

Lemma 4.6.2. Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und seien $A_1, A_2, ... \in \mathcal{A}$ beliebige Ereignisse. Dann gelten:

(i) $\liminf_{n\to\infty} A_n \subseteq \limsup_{n\to\infty} A_n$,

(ii) $(\liminf_{n\to\infty} A_n)^C = \limsup_{n\to\infty} A_n^C$,

(iii) $\limsup_{n\to\infty} \mathbf{1}_{A_n}(\omega) = \mathbf{1}_{\limsup_{n\to\infty} A_n}(\omega), \quad \forall \omega \in \Omega$,

(iv) $\liminf_{n\to\infty} \mathbf{1}_{A_n}(\omega) = \mathbf{1}_{\liminf_{n\to\infty} A_n}(\omega), \quad \forall \omega \in \Omega$.

Beweis. Denkt einfach mal kurz darüber nach, was $\liminf_{n\to\infty} a_n = 1$ oder $\liminf_{n\to\infty} a_n = 0$ für eine reelle Folge (a_n) bedeutet, wenn diese nur die Werte 0 und 1 annimmt. Übung. □

Das Borel-Cantelli Lemma gibt uns gleich ein Kriterium, ob die Wahrscheinlichkkeit des $\limsup A_n$ null ist oder (mit einer stärkeren Annahme) 1 ist. Dafür basteln wir mit Zufallsvariablen rum, alles basiert auf folgender Bemerkung:

Bemerkung 4.6.3. Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und seien $A_1, A_2, ... \in \mathcal{A}$ beliebige Ereignisse. Dann gilt

$$A_1, A_2, ... \text{ sind unabhängig} \quad \Leftrightarrow \quad \mathbf{1}_{A_1}, \mathbf{1}_{A_2}, ... \text{ sind unabhängig},$$

wobei wir auch die Unabhängigkeit durch paarweise Unabhängigkeit ersetzten können. Beachte: $A_1, A_2, ...$ ist ein Folge von Ereignissen, wohingegen $\mathbf{1}_{A_1}, \mathbf{1}_{A_2}, ...$ eine Folge von Zufallsvariablen ist. Das ist also ein guter Moment, in Kapitel 4.4 die Definitionen von Unabhängigkeit von Ereignissen und Zufallsvariablen nochmal zu vergleichen! Warum gilt die Äquivalenz? Checken wir die Definitionen:

$$\begin{aligned} \mathbf{1}_{A_1}, \mathbf{1}_{A_2}, ... \text{ unabhängig} &\overset{\text{Def. 4.4.10}}{\Leftrightarrow} \sigma(\mathbf{1}_{A_1}), \sigma(\mathbf{1}_{A_2}), ... \text{ unabhängig} \\ &\overset{2.1.9}{\Leftrightarrow} \{\emptyset, \Omega, A_1, A_1^C\}, \{\emptyset, \Omega, A_2, A_2^C\}, ... \text{ unabhängig} \\ &\Leftrightarrow A_1, A_2, ... \text{ unabhängig.} \end{aligned}$$

Für die dritte Äquivalenz haben wir Definition 4.4.8 genutzt, sowie die Eigenschaft, dass Unabhängigkeit von Ereignissen sich auch auf die Komplemente überträgt (Übung).

Satz 4.6.4. **[Borel-Cantelli-Lemma]** Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und seien $A_1, A_2, ... \in \mathcal{A}$ beliebige Ereignisse, so gelten:

(i)
$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty \quad \Rightarrow \quad \mathbb{P}\big(\limsup_{n\to\infty} A_n\big) = 0.$$

(ii) Sind die $A_1, A_2, ...$ *zusätzlich* paarweise unabhängig, so gilt
$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty \quad \Rightarrow \quad \mathbb{P}\big(\limsup_{n\to\infty} A_n\big) = 1.$$

Damit kennt ihr nun euer erstes „0-1-Gesetz“: Sind die Ereignisse $A_1, A_2, ...$ paarweise unabhängig, so gilt automatisch $\mathbb{P}(\limsup_{n\to\infty} A_n) \in \{0,1\}$ und

$$\mathbb{P}(\limsup_{n\to\infty} A_n) = 1 \quad\Longleftrightarrow\quad \mathbb{P}(\limsup_{n\to\infty} A_n) > 0 \quad\Longleftrightarrow\quad \sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty.$$

Das nützliche an 0-1-Gesetzen ist, dass man nur zeigen muss, dass etwas strikt positive Wahrscheinlichkeit hat, um sogar Wahrscheinlichkeit 1 zu schließen.

Beweis.

(i) „triviale Rechnung“: Die einfache Richtung folgt aus einfachen Manipulationen mit Mengen und Eigenschaften von Maßen:

$$\begin{aligned}\mathbb{P}(\limsup_{n\to\infty} A_n) &= \mathbb{P}\Big(\bigcap_{n=1}^{\infty}\bigcup_{k=n}^{\infty} A_k\Big)\\ &\overset{\text{Stet. Maße}}{=} \lim_{N\to\infty} \mathbb{P}\Big(\bigcap_{n=1}^{N}\bigcup_{k=n}^{\infty} A_k\Big)\\ &\overset{\text{Monotonie}}{\leq} \lim_{N\to\infty} \mathbb{P}\Big(\bigcup_{k=N}^{\infty} A_k\Big)\\ &\overset{\text{Subadd.}}{\leq} \lim_{N\to\infty} \sum_{k=N}^{\infty} \mathbb{P}(A_k) = 0.\end{aligned}$$

Die letzte Gleichheit gilt nach Annahme und Analysis 1 (Eigenschaft konvergenter Reihen), weil $\sum_{k=1}^{\infty} \mathbb{P}(A_k) < \infty$ angenommen wurde.

(ii) Die Rückrichtung ist deutlich schwieriger, wir nutzen die sogenannte „zweite-Momente-Methode“. Dafür werden erstes und zweites Moment einer geeigneten Zufallsvariable miteinander verglichen. Wir betrachten im Folgenden die Zufallsvariablen $\mathbf{1}_{A_n}$, die aufgrund von Bemerkung 4.6.3 unabhängig sind. Weil die Zufallsvariablen nur die Werte 0 und 1 annehmen, sind sie Bernoulli-verteilt. Genauer, es gilt $\mathbf{1}_{A_n} \sim \mathrm{Ber}(p_n)$, wobei $p_n = \mathbb{P}(A_n)$ die Wahrscheinlichkeit für den Wert 1 ist. Kleine Erinnerung: Für $\mathrm{Ber}(p)$-verteilte Zufallsvariablen ist der Erwartungswert p und die Varianz $p(1-p)$. Das werden wir im Folgenden ausnutzen.

Wir betrachten die Folge

$$Z_k = \sum_{n=1}^{k} \mathbf{1}_{A_n}, \quad k \in \mathbb{N},$$

weil mit dieser Folge $\limsup_{n\to\infty} A_n$ beschrieben werden kann:

$$\omega \in \limsup_{n\to\infty} A_n \quad\Leftrightarrow\quad +\infty = \sum_{n=1}^{\infty} \mathbf{1}_{A_n}(\omega) = \lim_{k\to\infty} Z_k(\omega). \tag{4.7}$$

Das gilt natürlich weil die Reihe nur aus Summanden 0 oder 1 besteht und daher unendlich ist genau dann, wenn unendlich viele Summanden 1 sind, also wenn ω in unendlich vielen A_n ist. Für Erwartungswert und Varianz gelten

$$\mathbb{E}[Z_k] = \sum_{n=1}^{k} \mathbb{E}[\mathbf{1}_{A_n}] = \sum_{n=1}^{k} \mathbb{P}(A_n) \overset{\text{Ann.}}{\longrightarrow} +\infty, \quad k \to \infty,$$

sowie

$$\mathbb{V}[Z_k] \underset{4.6.3}{\overset{4.2.31}{=}} \sum_{n=1}^{k} \mathbb{V}[\mathbf{1}_{A_n}] = \sum_{n=1}^{k} \mathbb{P}(A_n)(1-\mathbb{P}(A_n)) \leq \sum_{n=1}^{k} \mathbb{P}(A_n) = \mathbb{E}[Z_k],$$

weil unabhängige Zufallsvariablen auch unkorrelliert sind. Jetzt benutzen wir Tschebyscheff mit den Formeln für Erwartungswert und Varianz:

$$\mathbb{P}\Big(|Z_k - \mathbb{E}[Z_k]| \geq \frac{\mathbb{E}[Z_k]}{2}\Big) \overset{4.1.22}{\leq} \frac{\mathbb{V}[Z_k]}{\frac{1}{4}\mathbb{E}[Z_k]^2} \leq \frac{4\mathbb{E}[Z_k]}{\mathbb{E}[Z_k]^2} = \frac{4}{\mathbb{E}[Z_k]} \to 0, \tag{4.8}$$

für $k \to \infty$. Sei nun $\lambda > 0$ beliebig. Dann existiert wegen der Divergenz der Erwartungswerte gegen unendlich ein $k_0 \in \mathbb{N}$ mit $\frac{\lambda}{\mathbb{E}[Z_k]} < \frac{1}{2}$ für alle $k \geq k_0$. Weil aufgrund der Definition von Z_k auch $Z_k \leq \sum_{n=1}^{\infty} \mathbf{1}_{A_n}$ gilt, bekommen wir für beliebiges $k \geq k_0$

$$\begin{aligned}
\mathbb{P}\Big(\sum_{n=1}^{\infty} \mathbf{1}_{A_n} < \lambda\Big) &\leq \mathbb{P}(Z_k < \lambda) \\
&= \mathbb{P}\Big(\frac{Z_k}{\mathbb{E}[Z_k]} < \frac{\lambda}{\mathbb{E}[Z_k]}\Big) \\
&\leq \mathbb{P}\Big(\frac{Z_k}{\mathbb{E}[Z_k]} < \frac{1}{2}\Big) \\
&= \mathbb{P}\Big(\frac{Z_k}{\mathbb{E}[Z_k]} - 1 < -\frac{1}{2}\Big) \\
&\leq \mathbb{P}\Big(\Big|\frac{Z_k}{\mathbb{E}[Z_k]} - 1\Big| \geq \frac{1}{2}\Big) \\
&\overset{\text{Vor.}}{=} \mathbb{P}\Big(\big|Z_k - \mathbb{E}[Z_k]\big| \geq \frac{\mathbb{E}[Z_k]}{2}\Big) \overset{(4.8)}{\longrightarrow} 0, \quad k \to \infty.
\end{aligned}$$

Also gilt $\mathbb{P}\big(\sum_{n=1}^{\infty} \mathbf{1}_{A_n} < \lambda\big) = 0$. Weil λ beliebig gewählt war, folgt aufgrund der Stetigkeit von Maßen auch $\mathbb{P}\big(\sum_{n=1}^{\infty} \mathbf{1}_{A_n} < \infty\big) = \lim_{N\to\infty} \mathbb{P}\big(\sum_{n=1}^{\infty} \mathbf{1}_{A_n} < N\big) = 0$. Wegen (4.7) gilt nun

$$\mathbb{P}\big(\limsup_{n\to\infty} A_n\big) = \mathbb{P}\Big(\sum_{n=1}^{\infty} \mathbf{1}_{A_n} = +\infty\Big) = 1.$$

□

Kleine Anmerkung an dieser Stelle: Wenn im zweiten Teil statt paarweiser Unabhängigkeit sogar Unabhängigkeit angenommen wird, dann gibt es einen einfacheren Beweis für Teil (ii):

$$\begin{aligned}
\mathbb{P}\big((\limsup_{n\to\infty} A_n)^C\big) &\overset{\text{de Morgan}}{=} \mathbb{P}\Big(\bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k^C\Big) \\
&\overset{\text{Subadd.}}{\leq} \sum_{n=1}^{\infty} \mathbb{P}\Big(\bigcap_{k=n}^{\infty} A_k^C\Big) \\
&\overset{\text{Stet. Maße}}{\leq} \sum_{n=1}^{\infty} \lim_{N\to\infty} \mathbb{P}\Big(\bigcap_{k=n}^{N} A_k^C\Big) \\
&\overset{\text{Unab.}}{=} \sum_{n=1}^{\infty} \lim_{N\to\infty} \prod_{k=n}^{N} (1 - \mathbb{P}(A_n)) \\
&\overset{1-x\leq e^{-x}}{\leq} \sum_{n=1}^{\infty} \lim_{N\to\infty} \prod_{k=n}^{N} e^{-\mathbb{P}(A_n)} \\
&= \sum_{n=1}^{\infty} \lim_{N\to\infty} e^{-\sum_{k=n}^{N} \mathbb{P}(A_n)} \overset{\text{Vor.}}{=} \sum_{n=1}^{\infty} 0 = 0.
\end{aligned}$$

Checkt mal genau, warum dieser Beweis Unabhängigkeit statt nur paarweise Unabhängigkeit benutzt! Die komplizierte Variante wurde hauptsächlich aus didaktischen Gründen gewählt, um Argumente mit der Markovungleichung und Bernoulli-Zufallsvariablen zu wiederholen.

Nach diesem sehr abstrakten Highlight, nun zurück zur fast sicheren Konvergenz und dem starken Gesetz der großen Zahlen.

Korollar 4.6.5. Seien $X, X_1, X_2, ...$ Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$, dann gelten:

(i)
$$\sum_{n=1}^{\infty} \mathbb{P}(|X_n - X| > \varepsilon) < \infty \text{ für alle } \varepsilon > 0 \quad \Longrightarrow \quad X_n \overset{\text{f.s.}}{\to} X, \quad n \to \infty.$$

(ii) Unter der zusätzlich Annahme $X_1 - X, X_2 - X, ...$ sind paarweise unabhängig, gilt:
$$\sum_{n=1}^{\infty} \mathbb{P}(|X_n - X| > \varepsilon) = +\infty \text{ für ein } \varepsilon > 0 \quad \Longrightarrow \quad X_n \overset{\text{f.s.}}{\not\to} X, \quad n \to \infty.$$

Beweis. Beide Aussagen folgen sofort aus Borel-Cantelli:

(i) Mit $\varepsilon = \frac{1}{k}$ impliziert das Borel-Cantelli-Lemma $\mathbb{P}(|X_n - X| > \frac{1}{k}$ unendlich oft$) = 0$ bzw. $\mathbb{P}(|X_n - X| > \frac{1}{k}$ nur endlich oft$) = 1$ für alle $k \in \mathbb{N}$. Daraus folgt
$$\begin{aligned}\mathbb{P}(X_n \to X) &= \mathbb{P}\Big(\Big\{\omega : \forall k \in \mathbb{N}\, \exists N \in \mathbb{N} : |X_n(\omega) - X(\omega)| \leq \frac{1}{k}\, \forall n \geq N\Big\}\Big)\\ &= \mathbb{P}\Big(\bigcap_{k=1}^{\infty} \Big\{\omega : \exists N \in \mathbb{N} : |X_n(\omega) - X(\omega)| \leq \frac{1}{k}\, \forall n \geq N\Big\}\Big)\\ &= \mathbb{P}\Big(\Big\{\omega : |X_n(\omega) - X(\omega)| > \frac{1}{k} \text{ endlich oft}\Big\}\Big) = 1,\end{aligned}$$
weil der Durchschnitt abzählbar vieler Mengen von Maß 1 auch Maß 1 hat (wegen Komplementbildung, abzählbare Vereinigungen von Nullmengen sind Nullmengen). Denkt an dieser Stelle nochmal daran, dass wir manchmal in Wahrscheinlichkeiten die ω bei Zufallsvariablen weglassen weil es sich besser liest.

(ii) Borel-Cantelli impliziert $\mathbb{P}(|X_n - X| > \varepsilon$ unendlich oft$) = 1$, also ist die Wahrscheinlichkeit, dass X_n gegen X konvergiert, sogar 0.

□

Wegen Borel-Cantelli können wir den Unterschied von stochastischer und fast sicherer Konvergenz nun besser verstehen:

Bemerkung 4.6.6. **[Stochastische Konvergenz vs. fast sichere Konvergenz]**

- Um stochastische Konvergenz zu zeigen, muss $a_n := \mathbb{P}(|X_n - X| > \varepsilon)$ für beliebiges $\varepsilon > 0$ eine Nullfolge sein. Konvergiert die Nullfolge so schnell gegen 0, dass auch der Reihengrenzwert $\sum_{n=1}^{\infty} a_n$ endlich ist, so konvergiert nach Korollar 4.6.5 die Folge (X_n) fast sicher gegen X. Wenn wir uns an Analysis 1 erinnern, ist es ein großer Unterschied, ob die Reihe über eine Folge konvergiert oder die Folge eine Nullfolge ist. Beispielsweise reicht für stochastische Konvergenz die Abschätzung $\mathbb{P}(|X_n - X| > \varepsilon) \leq \frac{1}{n}$ für fast sichere Konvergenz nicht!

- Wir müssen noch einen Teil des Beweises von Satz 4.5.11 nachliefern. Das folgt aus dem ersten Teil des Korollars mit der Wahl der Teilfolge n_k, die
$$\mathbb{P}\Big(|X_{n_k} - X| > \frac{1}{k}\Big) < \frac{1}{2^k}$$
erfüllt.

Beispiel 4.6.7. Schauen wir uns das Beispiel 4.5.7 nochmal etwas allgemeiner an, um ein konkretes Beispiel für Korollar 4.6.5 zu haben: Seien $X_1, X_2, ...$ unabhängige Zufallsvariablen mit

$$\mathbb{P}(X_n = 1) = \frac{1}{n^p}, \quad \mathbb{P}(X_n = 0) = 1 - \frac{1}{n^p},$$

also $X_n \sim \text{Ber}(\frac{1}{n^p})$, $n \in \mathbb{N}$. Man stelle sich wieder unabhängige Versuche vor (1 bedeutet „Erfolg", 0 bedeutet „Misserfolg"), bei denen die Wahrscheinlichkeit für „Erfolg" immer kleiner wird. Fragen wir uns wieder: Konvergiert die Folge fast sicher gegen die Zufallsvariable $X = 0$? Wegen der angenommenen Unabhängigkeit gilt mit Korollar 4.6.5

$$X_n \overset{\text{f.s.}}{\to} 0, \quad n \to \infty \quad \Leftrightarrow \quad \sum_{n=1}^{\infty} \mathbb{P}(X_n = 1) = \sum_{n=1}^{\infty} \frac{1}{n^p} < +\infty \quad \Leftrightarrow \quad p > 1.$$

Ausformuliert ist die Aussage noch spektakulärer weil die Konvergenz gegen 0 einer Folge mit Werten 0 oder 1 bedeutet, dass die Folge irgendwann nur noch den Wert 0 annimmt. Ist $p \leq 1$, so ist der Versuch also immer mal wieder erfolgreich, wohingegen für $p > 1$ der Versuch nur endlich oft erfolgreich ist. Wenn man bedenkt, dass der Versuch unendlich oft ausgeführt wird und jedes Mal positive Wahrscheinlichkeit für Erfolg besteht, ist der Fall $p > 1$ schon überraschend! Hier ist ein ganz konkretes Anwendungsbeispiel. Stellt euch vor, ihr schreibt jedes Jahr die Stochastik 1 Klausur, ohne euch mit dem Inhalt zu beschäftigen. Weil ihr den Inhalt mit der Zeit vergesst, wird die Wahrscheinlichkeit des Bestehens immer kleiner. Die Frage ist nun: Wenn ihr immer wieder probiert, werdet ihr irgendwann bestehen? Das Beispiel zeigt, dass das davon abhängt, wie schnell die Bestehenswahrscheinlichkeit fällt, oder anders formuliert, wie schnell ihr vergesst.

Jetzt zum starken Gesetz der großen Zahlen:

Satz 4.6.8. **[Starkes Gesetz der großen Zahlen]** Sind $X_1, X_2, ...$ u.i.v. Zufallsvariablen auf $(\Omega, \mathcal{A}, \mathbb{P})$ mit $\mathbb{E}[|X_1|] < \infty$, so gilt

$$\frac{1}{n} \sum_{k=1}^{n} X_k \xrightarrow{\text{f.s.}} \mathbb{E}[X_1], \quad n \to \infty.$$

Beweis. Wir beweisen den Satz nur unter der zusätzlichen Annahme $\mathbb{E}[X_1^4] < \infty$. Die Aussage gilt auch unter der schwachen Annahme $\mathbb{E}[|X_1|] < \infty$, den Beweis werden wir aber erst in einer weiterführenden Vorlesung diskutieren. Wir nehmen ohne Einschränkung $\mathbb{E}[X_1] = 0$ an (sonst mit $\bar{X}_k := X_k - \mathbb{E}[X_k]$ verschieben, man nennt das Zentrieren). Wir wenden jetzt Korollar 4.6.5 an:

$$\begin{aligned}
\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{k=1}^{n} X_k - 0\Big| \geq \varepsilon\Big) &\overset{4.1.22, h(x)=x^4}{\leq} \frac{\mathbb{E}\Big[\Big|\frac{1}{n}\sum_{k=1}^{n} X_k\Big|^4\Big]}{\varepsilon^4} \\
&= \frac{1}{\varepsilon^4 n^4}\mathbb{E}\Big[\sum_{k_1,k_2,k_3,k_4=1}^{n} X_{k_1}X_{k_2}X_{k_3}X_{k_4}\Big] \\
&\overset{\text{unabh., zentr.}}{=} \frac{1}{\varepsilon^4 n^4}\Big(\sum_{k=1}^{n}\mathbb{E}[X_k^4] + \sum_{i\neq l=1}^{n}\mathbb{E}[X_i^2 X_l^2]\Big) \\
&\overset{\text{ident. vert.}}{=} \frac{1}{\varepsilon^4 n^4}\Big(n\mathbb{E}[X_1^4] + \frac{n(n-1)}{2}\mathbb{E}[X_1^2]\mathbb{E}[X_1^2]\Big) \\
&\leq \frac{C}{n^2},
\end{aligned}$$

mit $C = \frac{\mathbb{E}[X_1^4]}{\varepsilon^4} + \frac{\mathbb{E}[X_1^2]^2}{2\varepsilon^4}$. Damit haben wir die Voraussetzung von Korollar 4.6.5 (i) gecheckt und die Aussage folgt. Wie beim schwachen Gesetzt gilt auch hier wieder: Weil ε beliebig ist, spielt „>" oder „≥" keine Rolle.

Der Trick ist die dritte Gleichheit. Eigentlich sollten bei Tschebyscheff mit vierter Potenz n^4 Summanden im Zähler auftauchen. Wegen der Unabhängigkeit fallen von den Summanden $\mathbb{E}[X_{i_1}X_{i_2}X_{i_3}X_{i_4}]$ jedoch alle als 0 weg, bei denen eine der Zufallsvariablen nur einmal auftaucht (es gilt wegen der Unabhängigkeit zum Beispiel $\mathbb{E}[X_1X_1X_4X_5] = \mathbb{E}[X_1^2]\mathbb{E}[X_3]\mathbb{E}[X_4] = \mathbb{E}[X_1^2]0 = 0$). Es bleiben also nicht n^4 viele Summanden stehen, sondern nur die, bei denen entweder immer die gleiche oder nur zwei verschiedene Zufallsvariablen auftauchen. Das sind aber nur etwa n^2 viele und damit bringt der Nenner n^4 die Reihe zum konvergieren. □

Bemerkung 4.6.6 zeigt uns ganz genau den Unterschied zwischen unseren Beweisen der schwachen und dem starken Gesetze der großen Zahlen: Im Beweis des schwachen Gesetzes haben wir mit zweiten Momenten die obere Schranke $\frac{\mathbb{V}[X_1]}{\varepsilon n}$ hergeleitet. Das reichte für stochastische Konvergenz, aber nicht für fast sichere Konvergenz weil die harmonische Reihe divergiert. Das Tschebyscheff Argument mit vierten Momenten hingeben gibt die summierbare obere Schranke $\frac{C}{n^2}$. Dritte Momente funktionieren übrigens auch nicht, da stört der Betrag. Der „richtige“ Beweis (nur unter der Voraussetzung $\mathbb{E}[|X_1|] < \infty$ funktioniert anders, Tschebyscheff ist einfach keine gute Abschätzung.

Bemerkung 4.6.9. Unabhängig von den Annahmen an die Folge $X_1, X_2, ...$ spricht man immer von einem „starken Gesetz“, wenn fast sichere Konvergenz vorliegt. Man spricht von einem „schwachen Gesetz“, wenn stochastische Konvergenz vorliegt. Weil fast sichere Konvergenz die stochastische Konvergenz impliziert, impliziert ein starkes Gesetz immer ein schwaches Gesetz. Wir haben also das schwache Gesetz der großen Zahlen für u.i.v. Folgen mit endlichen zweiten Momenten gezeigt, das starke Gesetz der großen Zahlen für u.i.v. Folgen mit endlichen ersten Momenten. Weil $\mathbb{E}[X_1^2] < \infty$ auch $\mathbb{E}[|X_1|] < \infty$ impliziert (Hölder mit 1!), ist die Annahme $\mathbb{E}[X_1^2] < \infty$ im schwachen Gesetz natürlich viel zu stark, es gilt schließlich mit der schwächeren Annahme $\mathbb{E}[|X_1|] < \infty$ die stärkere Aussage der fast sicheren Konvergenz! Teil (i) in Satz 4.5.12 ist also im Prinzip überflüssig und wurde nur aus didaktischen Gründen behandelt. Interessanter ist eigentlich Teil (ii), denn unter diesen schwächeren Annahmen muss das starke Gesetz der großen Zahlen nicht gelten. Ein Beispiel ist folgendes: Ist $X_1, X_2, ...$ eine Folge von unabhängigen (nicht identisch verteilten) Zufallsvariablen mit

$$\begin{aligned}\mathbb{P}(X_n = n) &= \frac{1}{2n\log(n+1)},\\ \mathbb{P}(X_n = -n) &= \frac{1}{2n\log(n+1)},\\ \mathbb{P}(X_n = 0) &= 1 - \frac{1}{n\log(n+1)},\end{aligned}$$

so gilt das schwache Gesetz, aber nicht das starke Gesetz. Es gibt also durchaus einen Grund für Teil (ii) von Satz 4.5.12.

Am Ende des Kapitels noch eine kleine Bonus-Anwendung von Borel-Cantelli. Der Satz kommt in der Vorlesung nach dem Abspann weil die Aussage zwar nützlich, aber nicht notwenig für euer Verständniss ist (also nicht Klausurrelevenat ist). Wir zeigen, dass Mittelwerte $\frac{1}{n}\sum_{k=1}^n X_k$ von u.i.v. Folgen von Zufallsvariablen nur für $\mathbb{E}[|X_1|] < \infty$ fast sicher konvergieren können und (dann gilt das starke Gesetz der großen Zahlen) daher nur gegen den Erwartungswert konvergieren können!

Satz 4.6.10. Es seien $X_1, X_2, ...$ unabhängig und identisch verteilt, so dass $\frac{1}{n}\sum_{k=1}^n X_k$ fast sicher für $n \to \infty$ gegen eine Zufallsvariable X konvergiert. Dann gilt $\mathbb{E}[|X_1|] < \infty$ und

$$\frac{1}{n}\sum_{k=1}^n X_k \xrightarrow{\text{f.s.}} \mathbb{E}[X_1], \quad n \to \infty.$$

Beweis. Elementare Umformungen geben

$$\frac{X_n}{n} = \frac{\sum_{k=1}^n X_k}{n} - \frac{\sum_{k=1}^{n-1} X_k}{n} = \frac{\sum_{k=1}^n X_k}{n} - \frac{n-1}{n}\frac{\sum_{k=1}^{n-1} X_k}{n-1} \overset{\text{f.s.}}{\to} 0, \quad n \to \infty,$$

weil beide Summanden der rechten Seite nach Annahme gegen X konvergieren. Damit gilt dann

$$\mathbb{P}\Big(\frac{|X_n|}{n} > 1 \text{ unendlich oft}\Big) = 0$$

und mit $A_n := \{|X_n| > n\}$ folgt mit der zweiten Aussage von Borel-Cantelli $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$. Der Vergleich von Integralen mit Reihen aus Analysis 1 (für die monoton fallende Funktion $f(t) = \mathbb{P}(|X_1| > t)$ mit $f(n) = \mathbb{P}(|X_1| > n) \overset{\text{u.i.v.}}{=} \mathbb{P}(|X_n| > n) = \mathbb{P}(A_n)$) impliziert damit

$$\mathbb{E}[|X_1|] = \int_0^{\infty} \mathbb{P}(|X_1| > t)\,\mathrm{d}t < \infty.$$

Zur Erinnerung: Der Satz der Analysis besagt

$$\sum_{n=1}^{\infty} f(n) < \infty \quad \Leftrightarrow \quad \int_1^{\infty} f(x)\,\mathrm{d}x < \infty.$$

Dabei haben wir eine Übungsaufgabe benutzt: Es gilt nämlich für jede nicht-negative Zufallsvariable Y die Identität $\mathbb{E}[Y] = \int_0^{\infty} \mathbb{P}(Y > t)\,\mathrm{d}t$. Das folgt aus Fubini, wenn man die Wahrscheinlichkeit als Erwartungswert $\mathbb{E}[\mathbf{1}_{(t,+\infty)}(Y)]$ schreibt.

Damit ist die erste Aussage gezeigt. Weil nun $\mathbb{E}[|X_1|] < \infty$ gilt, ist die Voraussetzung von Satz 4.6.8 erfüllt und die zweite Aussage folgt. $\square$

Vorlesung 27

Anwendung 4.6.11. **[Empirisches Gesetz der großen Zahlen]** Seien $X_1, X_2, ...$ unabhängig und identisch verteilte Zufallsvariablen, so gilt

$$\frac{1}{n}\#\{k \leq n \colon X_k \in A\} \xrightarrow{\text{f.s.}} \mathbb{P}(X_1 \in A), \quad n \to \infty,$$

für $A \in \mathcal{B}(\mathbb{R})$. Insbesondere gilt mit der Wahl $A = (-\infty, t]$

$$\frac{1}{n}\#\{k \leq n \colon X_k \leq t\} \xrightarrow{\text{f.s.}} F(t), \quad n \to \infty,$$

und mit der Wahl $A = \{a_l\}$ für diskrete Zufallsvariablen mit Werten $\{a_1, ..., a_N\}$ und Wahrscheinlichkeiten $p_1, ..., p_N$

$$\frac{1}{n}\#\{k \leq n \colon X_k = a_l\} \xrightarrow{\text{f.s.}} p_l, \quad n \to \infty.$$

Beweis. Wir definieren $Y_k := \mathbf{1}_A(X_k)$. Die Y_i sind u.i.v. (siehe Korollar 4.2.28) mit endlichem Erwartungswert (sie sind durch 1 beschränkt). Berechnen wir noch den Erwartungswert: $\mathbb{E}[Y_1] = \mathbb{E}[\mathbf{1}_A(X_1)] = \mathbb{P}(X_1 \in A)$. Also kann das Gesetz der großen Zahlen angewandt werden und es gilt, weil Y_k nur die Werte 0 und 1 annehmen,

$$\frac{1}{n}\#\{k \leq n \colon X_k \in A\} = \frac{1}{n}\sum_{k=1}^{n} Y_k \xrightarrow{\text{f.s.}} \mathbb{E}[Y_1] = \mathbb{P}(X_1 \in A), \quad n \to \infty$$

$\square$

In der Statistik nennt man $F_n(t) := \frac{1}{n}\#\{k \leq n \colon X_k \leq t\}, t \in \mathbb{R}$, empirische Verteilungsfunktion der Stichprobe $X_1, ..., X_n$. Das Korollar besagt, dass die empirische Verteilungsfunktion punktweise gegen die Verteilungsfunktion konvergiert, wenn die Beobachtungsgröße wächst. Wozu ist das nützlich? Stellen wir uns vor, wir könnten ein Experiment beobachten, kennen aber nicht die Verteilungsfunktion der beschreibenden Zufallsvariablen. Wenn wir die Verteilungsfunktion $F(t)$ „schätzen“ wollen, beobachten wir also möglichst viele unabhängige Ausführungen des Experiments und nehmen $F_n(t)$ als Schätzwert für $F(t)$. Fragen dieser Art werden in Vorlesungen der Statistik und Ökonometrie thematisiert.

Anwendung 4.6.12. [**Monte-Carlo-Methode**] Numerische Methoden die darauf basieren, eine „gesuchte Größe“ μ als Erwartungswert irgendeiner Zufallsvariablen zu schreiben und diese mit dem Gesetz der Großen Zahlen als $\frac{1}{N}\sum_{k=1}^N X_k$ zu approximieren, heißen **Monte-Carlo Methoden**. Als Beispiel schauen wir uns die Berechnung von Integralen $\int\limits_0^1 f(x)\,\mathrm{d}x$ mit einer Monte-Carlo Methode an. Sei f integrierbar und $U \sim \mathcal{U}([0,1])$. Dann ist die Zufallsvariable $f(U)$ integrierbar mit

$$\mathbb{E}[f(U)] = \int_{\mathbb{R}} f(x) \cdot \mathbf{1}_{[0,1]}(x)\,\mathrm{d}x = \int\limits_0^1 f(x)\,\mathrm{d}x.$$

Der Monte-Carlo Ansatz zur Berechnung funktioniert nun wie folgt. Sind $U_1, U_2, ...$ u.i.v. mit $U_1 \sim \mathcal{U}([0,1])$, so gilt

$$\frac{1}{n}\sum_{k=1}^n f(U_k) \xrightarrow{\text{f.s.}} \int\limits_0^1 f(x)\,\mathrm{d}x, \quad n \to \infty.$$

Weil die Konvergenz fast sicher ist, müssen wir also „nur“ auf dem Computer eine möglichst lange Realisierung $U_1(\omega), ..., U_N(\omega)$ uniformer Zufallsvariablen erzeugen und $\frac{1}{N}\sum\limits_{k=1}^N f(U_k(\omega))$ als Approximation von $\int_0^1 f(x)\,\mathrm{d}x$ nehmen. Wie man an solch eine Realisierung $U_1(\omega), ..., U_N(\omega)$ im Computer rankommt, lernt ihr am Anfang der Vorlesung Monte Carlo Methoden.

Anwendung 4.6.13. [**Momentenschätzer**] Eine Grundfrage der Statistik ist folgende: Gegeben sei eine Verteilung einer parametrischen Klasse von Verteilungen $\{F_\theta : \theta \in \Theta\}$. Wir kennen den Parameter nicht, können aber unabhängige Zufallsvariablen unserer Verteilung beobachten. Stellt euch vor, ihr habt ein unfaire Münze, kennt aber die Wahrscheinlichkeit für Kopf nicht. Ihr habt also eine Verteilung aus der parametrischen Klasse $\{\mathrm{Ber}(p) : p \in (0,)]\}$ und wüsstet gerne den Parameter p. In manchen Fällen hilft das GGZ, und zwar dann, wenn der unbekannte Parameter θ mit dem Erwartungswert zusammenhängt. In dem Beispiel der Münze gilt beispielsweise $p = \mathbb{E}[X_1]$. Wenn ihr also eine u.i.v. Folge der Verteilung beobachten könnt (in dem Beispiel werft ihr immer wieder die Münze), so gibt das starke GGZ euch den Parameter der Verteilung: $p = \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^n X_k$. Für ein festes N nennt man deshalb $\hat{p}_N := \frac{1}{N}\sum_{k=1}^N X_k$ daher einen Schätzer (Schätzwert) von p. Ganz analog geht man zum Beispiel vor, wenn man von vornherein weiß, was die unbekannte Verteilung $\mathcal{N}(\mu, \sigma^2)$ ist, für einen unbekannten Parameter μ. Dann wäre $\hat{\mu}_N = \frac{1}{N}\sum_{k=1}^N X_k$ ein Schätzer von μ. Fragen dieser Art lernt ihr in der Stochastik 2 Vorlesung viel genauer kennen!

Die letzten zwei Anwendungen kommen aus zwei verschiedenen Gebieten, der Numerik und der Statistik. Ihr habt vielleicht gemerkt, dass die Fragestellungen in gewisser Art sehr ähnlich sind und in diesen einfachen Beispielen aus dem GGZ motiviert sind. In der stochastischen Numerik spricht man von Realisierungen, in der Statistik eher von Beobachtungen, gemeint ist immer $X(\omega)$. Der strukturelle Unterschied von stochastischer Numerik und Statistik ist eigentlich nur folgender: In der Statistik wird angenommen, dass man aus dem echten Leben eine Beobachtung von Zufallsvariablen hat (z. B. ökonomische Kenngrößen), in der stochastischen Numerik wird eine Realisierung der Zufallsvariablen selber erzeugen. Mit beiden Sichtweisen kann man mittels GGZ etwas über die gleiche Kenngröße aussagen (den Erwartungswert der Zufallsvariablen), die Anwendungen haben jedoch völlig unterschiedliche Motivationen.

4.7 Zentraler Grenzwertsatz

In diesem letzten Abschnitt besprechen wir den zentralen Grenzwertsatz (ZGS). Als Motivation stellen wir uns folgende Frage: Wir wollen wie im letzten Beispiel das Integral $\int_0^1 f(x)\,\mathrm{d}x$ numerisch approximieren, indem wir auf dem Computer viele uniforme Zufallsvariablen erzeugen.

In der Realität können wir n nicht gegen unendlich schicken, sagen wir also N ist eine große feste Zahl, z. B. $N = 10000$. Wir fragen nun:

$$\text{Wie gut ist die Näherung} \quad \frac{1}{10000}\sum_{k=1}^{10000} f(U_k) \quad \text{von} \quad \int_0^1 f(x)\,\mathrm{d}x \quad ?$$

Die Antwort ist leider: nicht so gut. Schauen wir uns dazu zunächst den zentralen Grenzwertsatz an:

Satz 4.7.1. **[Zentraler Grenzwertsatz]** Sind $X_1, X_2, ...$ u.i.v. Zufallsvariablen mit $\mathbb{E}[X_1] = \mu$ und endlicher Varianz $\sigma^2 := \mathbb{V}[X_1] > 0$. Dann gilt

$$\frac{\sum_{k=1}^n X_k - n\mu}{\sqrt{n}\cdot\sigma} \xrightarrow{(d)} Y, \quad n \to \infty,$$

mit $Y \sim \mathcal{N}(0,1)$.

Oft sieht man den zentralen Grenzwertsatz auch mit Wahrscheinlichkeiten geschrieben als

$$\mathbb{P}\Big(\frac{\sum_{k=1}^n X_k - n\mu}{\sqrt{n}\cdot\sigma} \leq t\Big) \to \frac{1}{\sqrt{2\pi}}\int_{-\infty}^t e^{-\frac{x^2}{2}}\,\mathrm{d}x =: \Phi(t), \quad n \to \infty,$$

oder

$$\mathbb{P}\Big(a \leq \frac{\sum_{k=1}^n X_k - n\mu}{\sqrt{n}\cdot\sigma} \leq b\Big) \to \frac{1}{\sqrt{2\pi}}\int_a^b e^{-\frac{x^2}{2}}\,\mathrm{d}x = \Phi(b) - \Phi(a), \quad n \to \infty.$$

Diese Umformulierungen folgen aus Satz 4.5.9 und der Stetigkeit der Verteilungsfunktion Φ der Normalverteilung.

Beweis. Wir geben den Beweis nur unter der stärkeren Annahme $\mathbb{E}[|X_1^3|] < \infty$. Das Argument funktioniert auch für $\mathbb{E}[X_1^2] < \infty$ (was gerade $\mathbb{V}[X_1] < \infty$ entspricht), wäre aber länger und komplizierter. Um das Hauptargument kompakt zu halten, besprechen wir zunächst vier Zutaten:

(i) Standardisierung: Ähnlich wie im Beweis des starken GGZ (dort haben wir zentriert) können wir ohne Einschränkung $\mu = 0$ und $\sigma = 1$ annehmen, um die Notation zu vereinfachen. Dazu definieren wir $Z_k = \frac{X_k - \mu}{\sigma}$ und bemerken, dass Z standardisiert ist (Linearität des Erwartungswertes und Verschiebungsformel der Varianz) sowie $\frac{\sum_{k=1}^n X_k - n\mu}{\sqrt{n}\cdot\sigma} = \frac{\sum_{k=1}^n Z_k}{\sqrt{n}}$ gilt.

(ii) Wir erinnern an die Skalierungs und Faltungseigenschaften der Normalverteilung: Sind $Y_1, ..., Y_n$ u.i.v. $\mathcal{N}(0,1)$, so gilt $\frac{\sum_{k=1}^n Y_k}{\sqrt{n}} \sim \mathcal{N}(0,1)$, siehe 4.3.16 und 4.3.18.

(iii) Statt die Konvergenz in Verteilung für alle beschränkten stetigen Funktionen zu zeigen, zeigen wir sie nur für glatte f weil wir dafür Taylor nutzen können. Es reicht, die Konvergenz für $f \in C^3(\mathbb{R})$ mit beschränkten f', f'', f''' zu zeigen. Warum? Das Argument haben wir im Beweis von Satz 4.5.9 schon gesehen: Wir approximieren die Indikatorfunktionen $\mathbf{1}_{(-\infty,t]}$ durch Funktionen f_+ und f_-, dieses mal aber nicht stückweise linear sondern glatt. Genau wie im Beweis der Hinrichtung von Satz 4.5.9 bekommen wir also die punktweise Konvergenz der Verteilungsfunktionen. Weil die Grenzverteilungsfunktion Φ stetig ist, ist das nach der Rückrichtung von Satz 4.5.9 gerade die Konvergenz in Verteilung.

(iv) Wir werden auf die Funktionen in (iii) Taylor mit der Restglieddarstellung aus dem Mittelwert benutzen. Für $f \in C^3(\mathbb{R})$ gilt, für $x, x_0 \in \mathbb{R}$,

$$f(x + x_0) = f(x_0) + f'(x_0)x + \frac{f''(x_0)}{2}x^2 + \frac{f'''(\tilde{x})}{6}x^3$$

für ein $\tilde{x}$ zwischen x und x_0.

Auf geht's: Es sei nun X_1 zentriert, f wie in (iii) und $Y, Y_1, Y_2, ...$ eine u.i.v. Folge von Zufallsvariablen mit $Y \sim \mathcal{N}(0,1)$, die unabhängig von der u.i.v. Folge von Zufallsvariablen $X_1, X_2, ...$ ist. Dann gilt mit einem coolen Teleskop Trick:

$$\begin{aligned}
&\Big|\mathbb{E}\Big[f\Big(\frac{\sum_{k=1}^n X_k}{\sqrt{n}}\Big)\Big] - \mathbb{E}\big[f(Y)\big]\Big| \\
\overset{\text{(ii)}}{=}\ &\Big|\mathbb{E}\Big[f\Big(\frac{\sum_{k=1}^n X_k}{\sqrt{n}}\Big) - f\Big(\frac{\sum_{k=1}^n Y_k}{\sqrt{n}}\Big)\Big]\Big| \\
\overset{\text{Teleskop}}{=}\ &\Big|\sum_{i=1}^n \Big(\mathbb{E}\Big[f\Big(\frac{\sum_{k=1}^{i-1} X_k + X_i + \sum_{k=i+1}^n Y_k}{\sqrt{n}}\Big) - f\Big(\frac{\sum_{k=1}^{i-1} X_k + Y_i + \sum_{k=i+1}^n Y_k}{\sqrt{n}}\Big)\Big]\Big)\Big|.
\end{aligned}$$

Auf alle $2n$ auftretenden Funktionen f wenden wir jetzt Taylor (ω-weise) wie in (iii) an, wobei wir jedes Mal $X_0 := \sum_{k=1}^{i-1} X_k + \sum_{k=i+1}^n Y_k$ wählen. Exemplarisch stehen dort ω-weise

$$f(X_0(\omega)) + f'(X_0(\omega))X_i(\omega) + f''(X_0(\omega))\frac{X_i(\omega)^2}{2} + f'''(\tilde{X}_i(\omega))\frac{X_i(\omega)^3}{6}$$

mit einem $\tilde{X}_i(\omega) \in \mathbb{R}$ aus dem Taylor-Restglied. Als Erwartungswert geschrieben, bekommen wir lauter Summanden der Form

$$\mathbb{E}\Big[f(X_0) + f'(X_0)X_i + f''(X_0)\frac{X_i^2}{2} + f'''(\tilde{X}_i)\frac{X_i^3}{6}\Big]$$

bzw.

$$\mathbb{E}\Big[f(X_0) + f'(X_0)Y_i + f''(X_0)\frac{Y_i^2}{2} + f'''(\tilde{Y}_i)\frac{Y_i^3}{6}\Big].$$

In der Differenz fallen die $f(X_0)$-Terme weg. Die f'-Terme fallen weg weil alle Zufallsvariablen zentriert sind und daher aufgrund der angenommenen Unabhängigkeit

$$\mathbb{E}[f'(X_0)X_k] = \mathbb{E}[f'(X_0)]\mathbb{E}[X_k] = 0$$

gilt (genauso für die Y_k). Auch die $f''(X_0)X_0^2/2$-Terme fallen in der Differenz weg, weil wieder aufgrund der Zentrierung und Unabhängigkeit $\mathbb{E}[f''(X_0)X_k^2] = \mathbb{E}[f''(X_0)]\mathbb{E}[X_k^2] = \mathbb{E}[f''(X_0)]$ und $\mathbb{E}[f''(X_0)Y_k^2] = \mathbb{E}[f''(X_0)]\mathbb{E}[Y_k^2] = \mathbb{E}[f''(X_0)]$ gelten. Wenn wir nun alles einsetzen und zurechtkürzen, bekommen wir

$$\begin{aligned}
&\Big|\mathbb{E}\Big[f\Big(\frac{\sum_{k=1}^n X_k}{\sqrt{n}}\Big)\Big] - \mathbb{E}\big[f(Y)\big]\Big| \\
&= \Big|\mathbb{E}\Big[\sum_{i=1}^n \Big(\frac{f'''(\tilde{X}_i)}{6}\frac{X_i^3}{n^{3/2}} - \frac{f'''(\tilde{Y}_i)}{6}\frac{Y_i^3}{n^{3/2}}\Big|\Big)\Big] \\
&\overset{\Delta}{\leq} \mathbb{E}\Big[\sum_{i=1}^n \Big(\Big|\frac{f'''(\tilde{X}_i)}{6}\frac{X_i^3}{n^{3/2}}\Big| + \Big|\frac{f'''(\tilde{Y}_i)}{6}\frac{Y_i^3}{n^{3/2}}\Big|\Big)\Big] \\
&\leq \frac{\sup_{x\in\mathbb{R}} |f'''(x)|}{6} \sum_{i=1}^n \frac{\mathbb{E}[|X_i|^3] + \mathbb{E}[|Y_i|^3]}{n^{3/2}} \\
&\overset{\text{u.i.v.}}{=} \frac{\sup_{x\in\mathbb{R}} |f'''(x)|}{6} \frac{\mathbb{E}[|X_1|^3] + \mathbb{E}[|Y_1|^3]}{n^{1/2}} \\
&\to 0, \quad n \to \infty.
\end{aligned}$$

Aufgrund von (iii) ist der Beweis damit beendet. □

Der Beweis war nicht sehr lang, aber äußerst komplex weil viel Verständniss verschiedener Objekte der Vorlesung nötig sind. In anderen Worten: Der Beweis lohnt sich, um verschiedene Stellen der Vorlesungen besser zu verstehen!

Bemerkung 4.7.2. Wegen Satz 4.7.1 bezeichnet man $\mathcal{N}(0,1)$ als **universelle Verteilung**. Wir haben nur $\mathbb{V}[X_1] < \infty$ angenommen – nichts weiter über die Verteilung. Dennoch kommt immer die Normalverteilung als Grenzwert raus!

In den Anwendungen des starken Gesetz der großen Zahlen approximieren wir jeweils einen festen Wert durch eine Folge von Zufallsvariablen. Typischerweise wird für großes N zwar $\frac{1}{N}\sum_{k=1}^N X_k \approx \mathbb{E}[X_1]$ gelten, jedoch nicht $\frac{1}{N}\sum_{k=1}^N X_k = \mathbb{E}[X_1]$, die Approximation $\frac{1}{N}\sum_{k=1}^N X_k$ ist schließlich eine Zufallsvariable. Es fragt sich also, wie stark die normierte Summe um $\mathbb{E}[X_1]$ konzentriert ist. Dabei hilft uns der zentrale Grenzwertsatz in der anders geklammerten Form

$$\frac{\sqrt{n}}{\sigma}\Big(\frac{1}{n}\sum_{k=1}^n X_k - \mathbb{E}[X_1]\Big) \overset{(d)}{\to} Y, \quad n \to \infty.$$

Wenn wir so tun, als ob für ein großes N bei der Konvergenz ungefähr (mathematisch nicht präzise!) Gleichheit eingetreten ist, und wir die Gleichung auflösen können, so gibt die Skalierungseigenschaft der Normalverteilung

$$\frac{1}{N}\sum_{k=1}^N X_k \approx \frac{\sigma}{\sqrt{N}}Y + \mathbb{E}[X_1] \sim \mathcal{N}\Big(\mathbb{E}[X_1], \frac{\sigma^2}{N}\Big). \tag{4.9}$$

Wenn wir jetzt an unsere Konzentrationsungleichungen aus Beispiel 3.3.9 zurückdenken, so gilt also grob: Mit Wahrscheinlichkeit $0,997$ ist die Abweichung $|\frac{1}{N}\sum_{k=1}^N X_k - \mathbb{E}[X_1]|$ kleiner als $\frac{3\sigma}{\sqrt{N}}$. Der zentrale Grenzwertsatz hilft uns also die Konvergenz im Gesetz der großen Zahlen genauer zu verstehen.

Anwendung 4.7.3. [**Monte Carlo Methoden**] Für die stochastische Numerik besagt die obige Diskussion, dass die Konvergenzordnung von Monte Carlo Verfahren (siehe Anwendung 4.6.12) leider nur $\frac{1}{2}$ ist, der Approximationsfehler also in der Wurzel der Simulationsgröße gegen 0 konvergiert. Das ist extrem langsam, $\frac{1}{\sqrt{10000}} = 0,01$ ist nicht sehr klein! Monte Carlo Verfahren sind daher grundsätzlich nicht sehr gut, funktionieren allerdings auch in schwierigen Situation immer (!) mit der Konvergenzordnung $\frac{1}{2}$. Wir sehen aus (4.9) auch, dass die Varianz der gewählten Approximationsfolge eine Rolle spielt, kleines σ gibt mehr Konzentration um den gesuchten Wert $\mu = \mathbb{E}[X_1]$. Tricks, die aus einer gegebenen u.i.v. Folge $X_1, X_2, ...$ mit $\mathbb{E}[X_1] = \mu$ eine neue neue u.i.v. Folge $\hat{X}_1, \hat{X}_2, ...$ mit $\mathbb{E}[\hat{X}_1] = \mu$ und kleinerer Varianz machen, nennt man Varianzreduktionsmethoden. Solche Tricks könnt ihr in der Monte Carlo Methoden Vorlesung kennenlernen, sie spielen auch im maschinellen Lernen eine große Rolle.

Beispiel 4.7.4. Schauen wir uns den Zusammenhang zur Binomialverteilung an, sagen wir $n = 20$ und $p = \frac{1}{2}$. Was ist $\mathbb{P}(X \leq 12)$ für $X \sim \mathrm{Bin}(20, \frac{1}{2})$? Die Wahrscheinlichkeit können wir natürlich als $\sum_{k=0}^{12} p_k$ mit $p_k = \binom{20}{k}\frac{1}{2^{20}}$ durch Einsetzen mühsam von Hand ausrechnen, das gibt $\frac{910596}{2^{20}} \approx 0,8684$. Alternativ machen wir das mit dem ZGS weil $\mathrm{Bin}(20, \frac{1}{2})$ sich auch als Summe von 20 unabhängigen $\mathrm{Ber}(\frac{1}{2})$ Zufallsvariablen $X_1, ..., X_{20}$ schreiben lässt. Diese haben Erwartungswert $\mu = \frac{1}{2}$ und Varianz $\sigma^2 = \frac{1}{4}$, also bekommen wir durch Erweitern

$$\mathbb{P}(X \leq 12) = \mathbb{P}\Big(\sum_{k=1}^{20} X_k \leq 12\Big) = \mathbb{P}\Big(\frac{\sum_{k=1}^{20} X_k - 20 \cdot \frac{1}{2}}{\sqrt{20} \cdot \frac{1}{2}} \leq \frac{12 - 20 \cdot \frac{1}{2}}{\sqrt{20} \cdot \frac{1}{2}}\Big) \overset{\text{ZGS}}{\approx} \Phi(0,8944).$$

Jetzt sucht ihr euch eine Tabelle für die Verteilungsfunktion Φ der Standardnormalverteilung (oder eine App) und findet etwa $0,8133$. So richtig gut ist das nicht, aber 20 ist auch nicht sonderlich groß und wie oben besprochen ist die Konvergenz im ZGS langsam.

Ganz zum Schluss noch ein Beispiel, bei dem der zentrale Grenzwertsatz (also die Stochastik) mit der Zahlentheorie/Kombinatorik zusammentrifft. Mega!

Beispiel 4.7.5. [**Stirlingformel für Fakultäten**] Habt ihr euch schon mal gefragt, wie groß $n!$ oder konkret $1000!$ ist? Tatsächlich riesig, aber wie riesig? $n!$ ist für großes n ziemlich genau $n^n e^{-n}\sqrt{2\pi n}$. Genauer, es gilt

$$\frac{n!}{n^n e^{-n}\sqrt{2\pi n}} \to 1, \quad n \to \infty. \tag{4.10}$$

Warum, was hat das mit dem ZGS zu tun? Man nimmt dazu $X_1, X_2, \ldots$ u.i.v. mit $X_1 \sim \operatorname{Poi}(1)$, es gilt mit der diskreten Faltungsformel also $\sum_{k=1}^n X_k \sim \operatorname{Poi}(n)$. Mit $f(x) = x^+$ benutzt man nun den ZGS:

$$\mathbb{E}\Big[f\Big(\frac{\sum_{k=1}^n X_k - n}{\sqrt{n}}\Big)\Big] \to \mathbb{E}[f(Y)] = \frac{1}{\sqrt{2\pi}}\int_0^\infty x e^{-\frac{x^2}{2}}\,\mathrm{d}x, \quad n \to \infty.$$

Das uneigentliche Integral auf der rechten Seite ist 1 (Substitution) und die Summe auf der linken Seite mit der Berechnungsformel aus Satz 4.1.11 gerade

$$\begin{aligned}
\mathbb{E}\Big[f\Big(\frac{\sum_{k=1}^n X_k - n}{\sqrt{n}}\Big)\Big] &\overset{\operatorname{Poi}(n)}{=} \sum_{k=1}^\infty f\Big(\frac{k-n}{\sqrt{n}}\Big)e^{-n}\frac{n^k}{k!}\\
&= \sum_{k=n+1}^\infty \Big(\frac{k-n}{\sqrt{n}}\Big)e^{-n}\frac{n^k}{k!}\\
&= \frac{e^{-n}}{\sqrt{n}}\Big(\sum_{k=n+1}^\infty k\frac{n^k}{k!} - \sum_{k=n+1}^\infty n\frac{n^k}{k!}\Big)\\
&= \frac{e^{-n}}{\sqrt{n}}\Big(\sum_{k=n+1}^\infty \frac{n^{k-1}n}{(k-1)!} - \sum_{k=n+1}^\infty n\frac{n^k}{k!}\Big)\\
&= \frac{e^{-n}n}{\sqrt{n}}\Big(\sum_{k=n}^\infty \frac{n^k}{k!} - \sum_{k=n+1}^\infty \frac{n^k}{k!}\Big) = \frac{e^{-n}\sqrt{n}n^n}{n!}.
\end{aligned}$$

Das war es schon, wenn man $\sqrt{2\pi}$ vom Grenzwert rüberzieht. Eine kleine Warnung: Die Funktion $f(x) = x^+$ ist zwar stetig, aber nicht beschränkt, passt also nicht zur Definition der Konvergenz in Verteilung. Wenn wir aber in unseren Beweis des ZGS schauen, bekommen wir die Konvergenz auch für x^+ hin, indem wir mit einer glatten Funktion f den Knick approximieren.

ENDE

Teil I

Lecture Stochastic Processes

Kapitel 5

Conditional expectation

Lecture 1

Some people say advanced probability theory is measure theory plus conditioning. While elementary conditioning on events with positive probability is not a big deal this chapter covers conditioning on σ-algebras and random variables, topics which are central in statistics and probability theory. We will discuss several constructions

- $\mathbb{E}[X|\mathcal{F}]$ - used in probability theory and mathematical finance for martingales,
- $\mathbb{E}[X|Y]$ and $\mathbb{E}[X|Y=y]$ - mostly used in statistics for regression,
- $\mathbb{P}(X \in A|Y)$ and $\mathbb{P}(X \in A\,|\,Y=y)$ - used to define a Markov process in probability theory.

One can already guess why many students struggle with conditional expectation. Similar objects appear in different contexts with different interpretations. On top, a bit like for the classical expectation, conditional expectation has concrete formulas for discrete and absolutely continuous random variables but has an abstract theoretical definition in the general case.

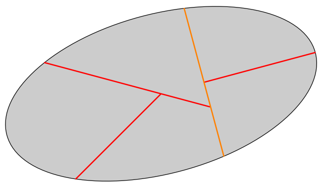
5.1 A hopefully gentle introduction

The main point in applications of conditional expectation is about information that is usually modelled using σ-algebras and measurability. In conditional expectation we will try to approximate random variables with other random variables that carry less "information" - which might already sound familiar from regression in statistics. For this sake, let X be a real-valued random variable on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Recall from the very beginning of Stochastik 1 (Discussion 1.1.8) the modelling idea of a measurable space in probability. The entire "knowledge of some abstract universe" is encoded in Ω, ω refers to one particular instance of randomness in the universe, for instance some atomic collision. The σ-algebra $\mathcal{A}$ contains as sets the events in the universe which occurrence can be observed (e.g. an apple is falling down a tree). So far we fixed such "information" and considered observations of real values using the concept of random variables. The concept of "information" behind the random variable did not play a major role so far as computing expectations, variances, etc. only depends on the distribution function. We will now try to get a better feeling for "information" by comparing different σ-algebras and by discussing the "information" given by random variables.

To keep things simple suppose X is a discrete random variable taking values $a_1, ..., a_N \in \mathbb{R}$. So far we only think about the values of X, but we know more. We also know the events on which X takes its values, namely $A_k := \{\omega \in \Omega : X(\omega) = a_k\} \in \mathcal{A}$. In terms of information this is a partition of Ω (i.e. $\Omega = \cup_{k=1}^N A_k$) into the observable events that can be distinguished through the different outcomes of X. Note that from this point of view the values of X are irrelevant, they only need to be different.

For $A_1, ..., A_N \in \mathcal{A}$ describe explicitly the smallest σ-algebra containing all these events. If the A_k are the events on that a discrete random variable takes its values, then this is nothing but $\sigma(X)$, the σ-algebra generated by X defined in 2.1.8.

Now suppose a second discrete random variable Y is given. How should we compare the "information" given by X and Y? How should we formalise that Y carries more "information" than X? It seems quite plausible to say that Y contains more "information" than X if the outcomes of the random variable X can be derived from the outcomes of the random variable Y. In more formal words, if the partition of Ω in "information" given by Y is finer (see the picture) than the partition given by X.



Two partitions of Ω into two yellow or five red sets

Formulated in terms of σ-algebras the notion of more information for random variables is nothing else but saying $\sigma(X) \subseteq \sigma(Y)$. A good example to keep in mind is the random variable Y that measures some temperature in tenth of a degree celsius and the random variable X that measures the same temperature only in degrees.

Draw a picture of the partitions of Ω for the simple temperature example. What are $\sigma(X)$ and $\sigma(Y)$? Make sure you understand why $\sigma(X) \subseteq \sigma(Y)$ holds.

In mathematics it is important to chose examples which are not too simple and not too complicated in order to understand definitions. The best example to understand the concept of information of σ-algebras is probably the sequence of σ-algebras induced by a sequence of random variables.

Example 5.1.1. Suppose $X_1, X_2, ...$ is a sequence of random variables on $(\Omega, \mathcal{A}, \mathbb{P})$, which we will later call a stochastic process. Then define

$$\mathcal{F}_n := \sigma(X_1, ..., X_n) := \{X_k^{-1}(B) : k \leq n, B \in \mathcal{B}(\mathbb{R}^n)\}, \qquad \text{for } n \in \mathbb{N}.$$

It is clear from the definition that the sequence of σ-algebras $\mathcal{F}_n$ is increasing in the sense that $\mathcal{F}_n \subseteq \mathcal{F}_{n+1}$, thus, carries more and more "information". Increasing sequences of σ-algebras will later-on be called filtrations. Let us assume, all X_n are discrete random variables taking values in $\mathbb{Z}$ and check what "information" is carried by $\mathcal{F}_n$? Since everything is discrete we can write down all preimages to obtain simple generators of the σ-algebras:

$$\begin{aligned}\mathcal{F}_n &= \sigma(\{\{X_1 \in A_1, ..., X_n \in A_n\} : A_i \subseteq \mathbb{Z}\})\\ &= \sigma(\{\{X_1 = i_1, ..., X_n = i_n\} : i_k \in \mathbb{Z}\}).\end{aligned}$$

Formulated in words, $\mathcal{F}_n$ is generated by events under which the process follows a given path, or, in other words, $\mathcal{F}_n$ contains all events which occurrence can be decided by looking at the first n steps of the stochastic process. Increasing the time of the process thus leads to a larger set of observable events, thus, to more "information". As an example let us check that the event "0 was hit before time n" belongs to $\mathcal{F}_n$ (but not to $\mathcal{F}_k$ for $k < n$):

$$\begin{aligned}
&\{0 \text{ was hit before time } n\} \\
&= \{0 \text{ was hit at time } 1\} \cup ... \cup \{0 \text{ was hit at time } n\} \\
&= \bigcup_{i_2,i_3,...,i_n \in \mathbb{Z}} \{X_1 = 0, X_2 = i_2, ..., X_n = i_n\} \cup ... \cup \bigcup_{i_1,...,i_{n-1} \in \mathbb{Z}} \{X_1 = i_1, ..., X_{n-1} = i_{n-1}, X_n = 0\} \\
&\in \mathcal{F}_n.
\end{aligned}$$

It is a good exercise to think of some other event and try to check that they belong (or do not belong) to the information of the first n steps $\mathcal{F}_n$.

These first thoughts comparing finitely many events break down immediately when the random variables fail to be discrete. To generalise we can only work with the generated σ-algebras

$$\sigma(X) = \{X^{-1}(B) : B \in \mathcal{B}(\mathbb{R})\}.$$

Recall that $\sigma(X)$ is the smallest σ-algebra on Ω so that X is measurable. If X is discrete $\sigma(X)$ only contains the events $\{X = a_k\}$ and their unions and complements. The discussion above suggests to say that Y carries more "information" than X if $\sigma(X) \subseteq \sigma(Y)$. To go even further, we can remove the concept of a random variable from the discussion. Similarly as above, we extend the idea of "information" to general σ-algebras by saying an σ-algebra $\mathcal{F}$ carries the "information" of its events. If $\mathcal{B}$ and $\mathcal{F}$ are σ-algebras on Ω we say that $\mathcal{B}$ carries more "information" than $\mathcal{F}$ if $\mathcal{F} \subseteq \mathcal{B}$. Recalling the motivation of σ-algebras comprising the observable events of the universe that makes perfect sense.

The information that two σ-algebras are generated by random variables is extremely valuable. It allows to formalise the concept of more/less information in the following way:

Lemma 5.1.2. (Doob's factorisation lemma)
Let X, Y two random variables on $(\Omega, \mathcal{A}, \mathbb{P})$. If X is measurable with respect to $\sigma(Y)$, i.e. $\sigma(X) \subseteq \sigma(Y)$, then there is a Borel-measurable mapping $h : \mathbb{R} \to \mathbb{R}$ such that $X = h(Y)$.

Proof. Exercise. First understand the statement if Y (and thus X) is discrete, h can easily be written down. You can see that your h can be defined arbitrarily for values that Y does not take. Then approximate Y using a sequence of step functions as in Theorem 3.1.6 and use the appearing sequence of mappings h_n to define h as their pointwise limit. □

The function h is not unique! Changing h on values that Y does not take will not change $h(Y)$.

A random variable X carrying less "information" than another random variable Y is thus only a transformation of Y and as such, not more complicated than Y. In the example of measuring temperatures with different thermometers, depending on the thermometers, the mapping h might just erase the decimals. The factorisation h simplifies the more fine information $\{Y = 5, 0\}$, ..., $\{Y = 5, 9\}$ into one $\{X = 5\}$. The reverse statement does not hold. In the setting of the theorem there will usually not be a mapping g such that $Y = g(X)$, the random variable X cannot explain the random variable Y without further information, think of measuring the temperature!

Now suppose we are in a situation in which the factorisation does not apply but we would still like to express Y as good as we can through X. Imagine Y is a feature vector that we are interested in but we can only observe some simpler feature vector X of which we believe X should be able to explain Y reasonably well. In statistics or machine learning you might have seen the idea of regressing Y on X in different contexts, such as finding a neural network functions such that $Y \approx f(X)$. In the following we discuss the measure theoretic version of this problem which is motivated from the factorisation lemma. If Y cannot be written as $h(X)$ then at least we aim at finding a measurable function h such that $Y \approx h(X)$ with a rigorously defined meaning of $\approx$. If we formalise $\approx$ through the L^2-distance of random variables, then - at least for square-integrable random variables - we can solve the minimisation problem using so-called conditional expectations.

Before motivating conditional expectations through best approximation let us recall an elementary but important fact of expectations that should be familiar from Stochastik 2:

Given a random variable X, how would we best approximate X by a constant value? As an instructive example, without any further thought, how would we approximate a $\mathcal{N}(\mu, \sigma^2)$-random variable best with a constant random variable? Of course using $Y \equiv \mu$, what else? To get the intuition straight let's check the L^2-approximation property of the expectation by solving

$$\min_{\theta \in \mathbb{R}} \mathbb{E}[(X-\theta)^2].$$

Expanding the square and minimising the quadratic function the solution to this simple approximation problem gives $\theta^* = \mathbb{E}[X]$.

Here is a simple visualisation that will be useful for conditional expectations in a bit:

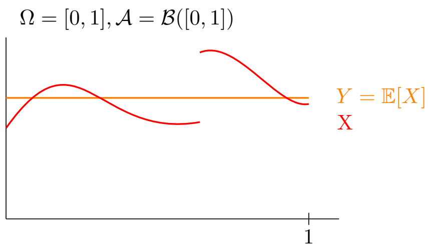


Illustration of best L^2-approximation of X through the expectation

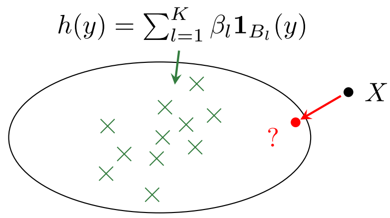
Warning: People say "the best estimation of a random variable without further information is the expectation" but this is only correct if we talk about minimisation of the L^2-distance. Minimising $\mathbb{E}[|X-\theta|]$ leads to the median, not the expectation!

Let us now return to the approximation of random variables through random variables carrying less information, using the L^2-distance. Suppose we have a random variable X on $(\Omega, \mathcal{A}, \mathbb{P})$ and a discrete σ-algebra $\mathcal{F} \subsetneq \sigma(X)$, let's say $\mathcal{F} = \sigma(B_1, ..., B_K)$ or, alternatively, $\mathcal{F} = \sigma(Y)$ for some random variable of the form $Y = \sum_{l=1}^K b_l \mathbf{1}_{B_l}$. Now assume we would like to explain X as good as we can using only the "information" from $\mathcal{F}$ (or alternatively from Y). It is not possible to write $X = h(Y)$ as all $\mathcal{F}$-measurable functions take the form

$$h(Y) = \sum_{l=1}^K h(b_l) \mathbf{1}_{B_l}$$

so that $X = h(Y)$ would contradict $\mathcal{F} \subsetneq \sigma(X)$. Instead, we try to best approximate X by all $\mathcal{F}$-measurable random variables $\sum_{l=1}^K \beta_l \mathbf{1}_{B_l}$, again in the L^2-sense as this is the easiest for computations.
Let us solve the minimisation problem

$$\min_{Z\ \mathcal{F}\text{-measurable}} \mathbb{E}[(X-Z)^2]$$



Approximating X best from the set of all $\mathcal{F}$-measurable random variables

without thinking much by just rewriting

$$\begin{aligned}
\min_{Z\ \mathcal{F}\text{-measurable}} \mathbb{E}[(X-Z)^2] &= \min_{\beta_1,\dots,\beta_K} \mathbb{E}\Big[\Big(X - \sum_{l=1}^K \beta_l \mathbf{1}_{B_l}\Big)^2\Big] \\
&= \min_{\beta_1,\dots,\beta_K} \mathbb{E}\Big[\Big(\sum_{l=1}^K (X-\beta_l)\mathbf{1}_{B_l}\Big)^2\Big] \\
&\overset{B_i \cap B_j = \emptyset}{=} \min_{\beta_1,\dots,\beta_K} \sum_{l=1}^K \mathbb{E}\big[(X-\beta_l)^2 \mathbf{1}_{B_l}\big]
\end{aligned}$$

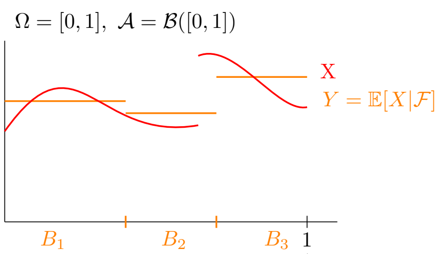
Minimising the right-hand side over the vector β gives

$$\beta = \Big(\frac{\mathbb{E}[X\mathbf{1}_{B_1}]}{\mathbb{P}(B_1)}, \dots, \frac{\mathbb{E}[X\mathbf{1}_{B_K}]}{\mathbb{P}(B_K)}\Big).$$

Introducing the notation $\mathbb{E}[X|A] := \frac{\mathbb{E}[X\mathbf{1}_A]}{\mathbb{P}(A)}$ we get the following solution of the L^2-approximation problem:

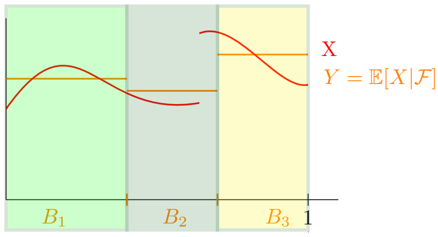
$$Z^* = \sum_{l=1}^K \mathbb{E}[X|B_l]\mathbf{1}_{B_l}. \tag{5.1}$$

We will denote this random variable by $\mathbb{E}[X|\mathcal{F}]$ and call it the conditional expectation of X given $\mathcal{F}$. $\mathbb{E}[X|\mathcal{F}]$ is the most we can say about X (in the L^2-sense) with the information given by $\mathcal{F}$.



Best L^2-approximation in the light of approximation through simple functions

The computation above did not shed much light on what is going on. Could we have guessed the formula (5.1) without computations? For that sake let us recall the interpretation of elementary conditional probability measures $\mathbb{P}(\cdot \,|\, B) := \frac{\mathbb{P}(\cdot\, \cap B)}{\mathbb{P}(B)}$. The original random experiment is restricted to a subexperiment on B. All knowledge from $\Omega \backslash B$ is forgotten for the subexperiment, the relative probabilities of events $A \subseteq B$ remain unchanged by normalising all events with the same factor $\mathbb{P}(B)$. If X was a random variable on Ω then the restriction to B is a random variable on the restriction $(B, \mathcal{A}_{|B}, \mathbb{P}_{|B})$ with $\mathbb{E}_{|B}[X] = \mathbb{E}[X|B]$. We also call $\mathbb{E}_{|B}$ elementary conditional expectation. With this interpretation of elementary conditioning in mind formula (5.1) is nothing



Best L^2-approximation in the light of constant approximation on subexperiments

but K-times elementary L^2-approximation with constants for the K restricted random variables $X_{|B}$.
Writing these thoughts formally by introducing $1 = \sum_{l=1}^K \mathbf{1}_{B_l}$ for the splitting the experiment into the K subexperiments gives

$$\mathbb{E}[(X-Y)^2] = \mathbb{E}\Big[\Big(\sum_{l=1}^K (X-\beta_l)\mathbf{1}_{B_l}\Big)^2\Big] = \sum_{l=1}^K \mathbb{E}[(X-\beta)^2\mathbf{1}_{B_l}] = \sum_{l=1}^K \mathbb{E}_{|B_l}[(X-\beta_l)^2]\mathbb{P}(B_l).$$

Best L^2-approximating all K subexperiments by constants gives the same solution β that we have found above by direct computation. The advantage is our better understanding of the appearing factors as elementary conditional expectations $\mathbb{E}_{|B_l}[X]$.
In many textbooks on probability theory the motivation of conditional expectations is less formal, taking for granted that the best approximation of a random variable without extra information is the expectation. The intuitive reasoning is that the conditional expectation is the expectation given prior knowledge. Even though the statement "given" is very imprecise (the true formulation is L^2-minimisation as above) it works intuitively quite well.

Example 5.1.3. Suppose X denotes a dice taking values 1, ..., 6 with probabilities $\frac{1}{6}$. What is our best guess for X if we know that the dice is even (or odd)? Keeping the L^2-minimisation property of expectations in mind we should compute the expectations of the conditional experiments, the dice conditioned on being even (or odd). Intuitively, this is 4 in the even and 3 in the odd case. Denoting by $\mathcal{F} = \{\Omega, \emptyset, \{1,3,5\}, \{2,4,6\}\}$ the additional information we have this is nothing but a sketch of the rigorous formula

$$\mathbb{E}[X|\mathcal{F}] = \mathbb{E}[X|X \in \{1,3,5\}]\mathbf{1}_{X\in\{1,3,5\}} + \mathbb{E}[X|X \in \{2,4,6\}]\mathbf{1}_{X\in\{2,4,5\}}$$

from above. Plug-in the definition of the elementary conditional expectations to check they give 3 and 4!

There is a major challenge remaining. How do we generalise the above reasoning to general random variables and general σ-algebras if we cannot easily compute with finite sums of indicators? We follow the axiomatic approach of Kolmogorov in which the conditional expectations $\mathbb{E}[X|\mathcal{F}]$ and $\mathbb{E}[X|Y]$ are defined through the two axiomatic properties

- $\mathbb{E}[X|\mathcal{F}]$ is $\mathcal{F}$-measurable,
- $\mathbb{E}\big[\mathbb{E}[X|\mathcal{F}]\mathbf{1}_A\big] = \mathbb{E}[X\mathbf{1}_A]$ holds for all $A \in \mathcal{F}$.

Before discussing for general random variables how these properties define a reasonable object let us check that our discrete conditional expectation satisfies these properties. The measurability is a matter of our approach, we only minimised the distance over $\mathcal{F}$-measurable random variables. The more strange second property can be checked through the following quick computation. Let us first assume A is equal to precisely one of the B_l, thus, disjoint from all the others:

$$\mathbb{E}\big[\mathbb{E}[X|\mathcal{F}]\mathbf{1}_A\big] = \mathbb{E}\Big[\sum_{l=1}^K \mathbb{E}[X|B_l]\mathbf{1}_{B_l}\mathbf{1}_A\Big] = \mathbb{E}\big[\mathbb{E}[X|A]\mathbf{1}_A\big] = \mathbb{E}[X|A]\mathbb{E}[\mathbf{1}_A] \overset{\text{def}}{=} \mathbb{E}[X\mathbf{1}_A].$$

Since all $A \in \mathcal{F}$ can be written as finite union of the B_k the general claim follows from linearity. The magic of conditional expectation is that the preceding two properties are all we need for a powerful definition of general conditional expectation with amazing consequences!

5.2 The axiomatic approach of Kolmogorov

We shall now turn to a general setup, dropping the assumption of discrete random variables and finite σ-algebras. We will reverse the story and start with an axiomatic definition of conditional expectation from which we then derive abstract existence and identify key properties (such as L^2-error minimization) that reflect the motivation given before.

Definition 5.2.1. Let $(\Omega, \mathcal{A}, \mathbb{P})$ a probability space, $X \in \mathcal{L}^1(\Omega, \mathcal{A}, \mathbb{P})$, and $\mathcal{F}$ a sub-σ-Algebra of $\mathcal{A}$. Then a random variable Z on $(\Omega, \mathcal{A}, \mathbb{P})$ is called the **conditional expectation of X given $\mathcal{F}$** if

- Z is $\mathcal{F}$-measurable,
- $\mathbb{E}[Z\mathbf{1}_A] = \mathbb{E}[X\mathbf{1}_A]$ for all $A \in \mathcal{F}$.

If such a random variable Z exists, then we write $Z = E[X|\mathcal{F}]$. If $\mathcal{F} = \sigma(Y)$ for a random variable Y then we write $Z = \mathbb{E}[X|Y]$ and call Z the **conditional expectation of X give Y**.

Choosing $A = \Omega$ in the second condition it is clear that the integrability assumption on X cannot be avoided. In contrast to the section before we cannot use L^2-distances, second moments are not assumed to be finite. It is important to note that the very definition of $\mathbb{E}[X|\mathcal{F}]$ shows that conditional expectation is not defined uniquely. This is caused by the second property as expectations do not change if random variables are changed on zero sets. Any other $\mathcal{F}$-measurable random variable $\bar{Z}$ that equals Z almost surely will automatically satisfy the second condition.

Definition 5.2.2. Suppose Z and $\bar{Z}$ are random variables on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. We call $\bar{Z}$ a **version** of Z if $\mathbb{P}(Z = \bar{Z}) = 1$.

We will see later that it can be useful to choose particularly versions of the conditional expectation that satisfies additional measurability properties.

Theorem 5.2.3. Let X an integrable random variable on $(\Omega, \mathcal{A}, \mathbb{P})$ and $\mathcal{F}$ a sub-σ-algebra of $\mathcal{A}$. Then the conditional expectation $\mathbb{E}[X|\mathcal{F}]$ exists and is almost surely unique, i.e. two random variables satisfying the two defining properties are versions of each other.

Proof. **Uniqueness:** Suppose Z and $\bar{Z}$ are random variables fulfilling the two defining properties of Definition 5.2.1 and let $A = \{\bar{Z} < Z\} \in \mathcal{F}$. Using the second property of conditional expectation, it holds that

$$0 = \mathbb{E}[\mathbf{1}_A Z] - \mathbb{E}[\mathbf{1}_A \bar{Z}] = \mathbb{E}[(Z - \bar{Z})\mathbf{1}_A] \geq 0.$$

It follows that $\mathbf{1}_A(Z - \bar{Z}) = 0$ almost surely. Similarly, it holds that $\mathbb{1}_{A^c}(Z - \bar{Z}) = 0$ almost surely. Taking the unions of the nullsets gives $Z = \bar{Z}$ almost surely.

Existence: To prove the existence we use the Radon-Nykodyn from functional analysis. The theorem states that a σ-finite measure ν has a density with respect to another σ-finite measure μ if and only if $\nu \ll \mu$. Here we use the absolutely continuity notion $\mu \ll \mu$ if every zero set for μ is also a zero set for ν. The interested reader is referred to **??** Corollary 7.34. Now let

$$Q^+ : \mathcal{F} \to [0,1], A \mapsto \mathbb{E}[X^+\mathbf{1}_A] \quad \text{and} \quad Q^- : \mathcal{F} \to [0,1], A \mapsto \mathbb{E}[X^-\mathbf{1}_A], \tag{5.2}$$

which are both σ-finite measures on $\mathcal{F}$ with $Q^+ \ll \mathbb{P}$ and $Q^- \ll \mathbb{P}$. Hence, there are $\mathcal{F}$-measurable densities Z^+, Z^- with

$$Q^+(A) = \int_A Z^+ \, \mathrm{d}\mathbb{P} \quad \text{and} \quad Q^-(A) = \int_A Z^- \, \mathrm{d}\mathbb{P}$$

for all $A \in \mathcal{A}$. Let $Z := Z^+ - Z^-$, then Y is $\mathcal{F}$ measurable and

$$\begin{aligned}
\mathbb{E}[\mathbf{1}_A Z] &= \mathbb{E}[\mathbf{1}_A Z^+] - \mathbb{E}[\mathbf{1}_A Z^-] \\
&= \int_A Z^+ \, \mathrm{d}\mathbb{P} - \int_A Z^+ \, \mathrm{d}\mathbb{P} \\
&\overset{(5.2)}{=} Q^+(A) - Q^-(A) \\
&= \mathbb{E}[\mathbf{1}_A X^+] - \mathbb{E}[\mathbf{1}_A X^-] \\
&= E[\mathbf{1}_A X].
\end{aligned}$$

Therefore, the random variable Z fullfills the properties of the conditional expectation of X given $\mathcal{F}$. □

Lecture 2

Just as for typical expectations we continue with a set of properties fulfilled by the conditional expectation. Some properties look familiar to old properties of expectations, but always keep in mind: conditional expectations are random variables!

Theorem 5.2.4. (Standard properties of conditional expectation)
Let $X, Y \in \mathcal{L}^1(\Omega, \mathcal{A}, \mathbb{P})$, $\lambda \in \mathbb{R}$, and $\mathcal{G} \subseteq \mathcal{F} \subseteq \mathcal{A}$ sub-σ-Algebras. Then the following properties hold:

(i) $\mathbb{E}\big[\mathbb{E}[X \,|\, \mathcal{F}]\big] = \mathbb{E}[X]$ and $\mathbb{E}[1 \,|\, \mathcal{F}] = 1$ a.s.

(ii) $\mathbb{E}[\,\lambda X + Y \,|\, \mathcal{F}\,] = \lambda \mathbb{E}[\,X \,|\, \mathcal{F}\,] + \mathbb{E}[Y \,|\, \mathcal{F}]$ a.s.

(iii) $X \geq Y$ a.s. $\Rightarrow \mathbb{E}[X \,|\, \mathcal{F}] \geq \mathbb{E}[Y \,|\, \mathcal{F}]$ a.s.

(iv) If $\mathbb{E}[|XY|] < \infty$ and Y is $\mathcal{F}$-measurable, then

$$\mathbb{E}[XY \,|\, \mathcal{F}] = Y\mathbb{E}[X \,|\, \mathcal{F}] \text{ a.s.} \quad \text{and} \quad \mathbb{E}[Y \,|\, \mathcal{F}] = Y \text{ a.s.}$$

(v) $\mathbb{E}\big[\mathbb{E}[X \,|\, \mathcal{F}] \,\big|\, \mathcal{G}\big] = \mathbb{E}\big[\mathbb{E}[X \,|\, \mathcal{G}] \,\big|\, \mathcal{F}\big] = \mathbb{E}[X \,|\, \mathcal{G}]$ a.s.

(vi) $\big|\mathbb{E}[X \,|\, \mathcal{F}]\big| \leq \mathbb{E}\big[|X| \,\big|\, \mathcal{F}\big]$ a.s.

(vii) If $\sigma(X)$ and $\mathcal{F}$ are independent, then $\mathbb{E}[X \,|\, \mathcal{F}] = \mathbb{E}[X]$ a.s.

(viii) If $\mathbb{P}(A) \in \{0, 1\}$ for all $A \in \mathcal{F}$, then $\mathbb{E}[X \,|\, \mathcal{F}] = \mathbb{E}[X]$ a.s.

(ix) If $\mathbb{E}[X|\mathcal{F}]$ is a.s. constant, then $\mathbb{E}[X|\mathcal{F}] = \mathbb{E}[X]$ a.s.

(x) If $\mathbb{E}[X\mathbf{1}_A] = \mathbb{E}[Y\mathbf{1}_A]$ for all $A \in \mathcal{F}$, then $\mathbb{E}[X|\mathcal{F}] = \mathbb{E}[Y|\mathcal{F}]$ a.s.

(xi) Suppose $|X_n| \leq Y$ a.s. for $Y \in \mathcal{L}^1(\Omega, \mathcal{A}, \mathbb{P})$ and $\lim_{n\to\infty} X_n = X$ a.s., then

$$\lim_{n\to\infty} \mathbb{E}[X_n \,|\, \mathcal{F}] = \mathbb{E}[X \,|\, \mathcal{F}] \text{ a.s and in } L^1.$$

(xii) Suppose $X_n \geq 0$ is an increasing sequence of random variables with $\lim_{n\to\infty} X_n = X$ a.s., then

$$\lim_{n\to\infty} \mathbb{E}[X_n \,|\, \mathcal{F}] = \mathbb{E}[X \,|\, \mathcal{F}] \text{ a.s.}$$

Proof. (i) Exercise

(ii) The trick is always the same: If we intend to prove $\mathbb{E}[...|...] = Z$ a.s., we need to check for Z the two defining properties of the claimed conditional expectation. Uniqueness of conditional expectations then implies that Z is a version of the conditional expectation. Following once in detail this routine let us denote the righthand side by Z. Since Z is a linear combination of two $\mathcal{F}$-measurable random variables, Z is $\mathcal{F}$-measurable, which is the first condition of conditional expectation. For the expectation condition, with $A \in \mathcal{F}$, we obtain, using properties of the usual expectation (linearity and (ii) of Theorem 3.1.15) and the second property of the conditional expectations $\mathbb{E}[X \mid \mathcal{F}]$, $\mathbb{E}[Y \mid \mathcal{F}]$,

$$\begin{aligned}\mathbb{E}[\mathbf{1}_A ZF] &= \mathbb{E}[\mathbf{1}_A(\lambda\mathbb{E}[\,X \mid \mathcal{F}\,] + \mathbb{E}[Y \mid \mathcal{F}])] \\ &= \lambda\mathbb{E}[\mathbf{1}_A\mathbb{E}[X \mid \mathcal{F}]] + \mathbb{E}[\mathbf{1}_A\mathbb{E}[Y \mid \mathcal{F}]] \\ &= \lambda\mathbb{E}[\mathbf{1}_A X] + \mathbb{E}[\mathbf{1}_A Y] \\ &= \mathbb{E}[\mathbf{1}_A(\lambda X + Y)].\end{aligned}$$

Hence, Z satisfies the defining conditions of $\mathbb{E}[\lambda X + Y \mid \mathcal{F}]$.

(iii) Lemma **??** tells us that $A := \{\mathbb{E}[X|\mathcal{F}] < \mathbb{E}[Y|\mathcal{F}]\} \in \mathcal{F}$. Using monotonicity of expectations and linearity of conditional expectations from (ii) gives

$$0 \geq \mathbb{E}[\mathbf{1}_A(\mathbb{E}[X \mid \mathcal{F}] - \mathbb{E}[X \mid \mathcal{F}])] = \mathbb{E}[\mathbf{1}_A\mathbb{E}[X - Y \mid \mathcal{F}]] = \mathbb{E}[\mathbf{1}_A(X - Y)] \overset{\text{ass.}}{\geq} 0.$$

Since $\mathbf{1}_A(\mathbb{E}[X \mid \mathcal{F}] - \mathbb{E}[Y \mid \mathcal{F}]) \leq 0$ by definition of A we find $\mathbf{1}_A(\mathbb{E}[X \mid \mathcal{F}] - \mathbb{E}[Y \mid \mathcal{F}]) = 0$ a.s. (Theorem 3.1.15, (iii)) which gives $\mathbf{1}_A = 0$ a.s. or, equivalently, $\mathbb{E}[X \mid \mathcal{F}] \geq \mathbb{E}[Y \mid \mathcal{F}]$ a.s.

(iv) The proof works through discretisation and linearity. First assume $X, Y \geq 0$. Define $Y_n = \frac{1}{2^n} \cdot \lfloor 2^n \cdot Y \rfloor$ (draw a picture to undestand it!) so that almost surely $Y_n \nearrow Y$ and

$$Y_n\mathbb{E}[X \mid \mathcal{F}] \nearrow Y\mathbb{E}[X \mid \mathcal{F}].$$

MCT for usual expectations implies $\lim\limits_{n\to\infty} \mathbb{E}[\mathbf{1}_A Y_n \mathbb{E}[X \mid \mathcal{F}]] = \mathbb{E}[\mathbf{1}_A Y \mathbb{E}[X \mid \mathcal{F}]]$ for $A \in \mathcal{F}$. Now we compute the left-hand side:

$$\begin{aligned}\mathbb{E}\left[\mathbf{1}_A Y_n \mathbb{E}[X \mid \mathcal{F}]\right] &= \mathbb{E}\left[\mathbf{1}_A \sum_{k=0}^{\infty} \frac{k}{2^n} \mathbf{1}_{Y_n = \frac{k}{2^n}} \mathbb{E}[X \mid \mathcal{F}]\right] \\ &\overset{\text{MCT}}{=} \sum_{k=0}^{\infty} \mathbb{E}\left[\mathbf{1}_A \frac{k}{2^n} \mathbf{1}_{Y_n = \frac{k}{2^n}} \mathbb{E}[X \mid \mathcal{F}]\right] \\ &= \sum_{k=0}^{\infty} \mathbb{E}\left[\mathbf{1}_A \frac{k}{2^n} \mathbf{1}_{Y_n = \frac{k}{2^n}} X\right] \overset{\text{MCT}}{=} \mathbb{E}[\mathbf{1}_A Y_n X]\end{aligned}$$

Again MCT yields $\mathbb{E}[\mathbf{1}_A YX] = \mathbb{E}[\mathbf{1}_A Y \mathbb{E}[X \mid \mathcal{F}]]$ which shows the expectation condition

$$\mathbb{E}[YX \mid \mathcal{F}] = Y\mathbb{E}[X \mid \mathcal{F}].$$

The $\mathcal{F}$-measurability of $Y\mathbb{E}[X|\mathcal{F}]$ follows from measurability of products. For the general case we proceed as usually, writing $X = X^+ - X^-$,$Y = Y^+ - Y^-$ and then using linearity from (ii). The second claim follows from (i) using $X = 1$.

(v) Exercise

(vi) The proof is exactly the same that we have already encountered for the classical expectation in Lemma **??**:

$$\begin{aligned}\big|\mathbb{E}[X\,|\,\mathcal{F}]\big| &= \big|\mathbb{E}[X^+\,|\,\mathcal{F}] - \mathbb{E}[X^-\,|\,\mathcal{F}]\big|\\ &\overset{\Delta}{\le} \big|\mathbb{E}[X^+\,|\,\mathcal{F}]\big| + \big|\mathbb{E}[X^-\,|\,\mathcal{F}]\big|\\ &= \mathbb{E}[X^+\,|\,\mathcal{F}] + \mathbb{E}[X^-\,|\,\mathcal{F}]\\ &= \mathbb{E}[X^+ + X^-\,|\,\mathcal{F}]\\ &= \mathbb{E}\big[|X|\,\big|\,\mathcal{F}\big]\end{aligned}$$

(vii) First recall that one (of several equivalent) ways to state the independence is to say that X is independent of $\mathbf{1}_A$ for all $A \in \mathcal{F}$. Using the factorisation of expectations of independent random variables gives

$$\mathbb{E}[\mathbf{1}_A X] \overset{\text{ind.}}{=} \mathbb{E}[X]\mathbb{E}[\mathbf{1}_A] \overset{\text{lin.}}{=} \mathbb{E}[\mathbb{E}[X]\mathbf{1}_A]$$

for all $A \in \mathcal{F}$. Since additionally all constant random variables are $\mathcal{F}$-measurable, $\mathbb{E}[X]$ is a version of $\mathbb{E}[X\,|\,\mathcal{F}]$.

(viii) The "trivial" σ-algebra $\mathcal{F}$ is independent of all sub-σ-algebras $\mathcal{A}$, in particular of $\sigma(X)$. Now use (vii).

(ix) Suppose $\mathbb{E}[X|\mathcal{F}] = c$ almost surely. The defining property yields $\mathbb{E}[X\mathbf{1}_A] = \mathbb{E}[c\,\mathbf{1}_A] = c\,\mathbb{P}(A)$ for all $A \in \mathcal{F}$. Now choose $A = \Omega$ and the claim follows.

(x) We show that $\mathbb{E}[Y|\mathcal{F}]$ satisfies the defining properties of $\mathbb{E}[X|\mathcal{F}]$ and then apply the uniqueness. Measurability follows from the measurability of conditional expectations, the expectation property as follows:

$$\mathbb{E}[X\mathbf{1}_A] = \mathbb{E}[Y\mathbf{1}_A] \overset{(i)}{=} \mathbb{E}\big[\mathbb{E}[Y\mathbf{1}_A|\mathcal{F}]\big] \overset{(iv)}{=} \mathbb{E}\big[\mathbb{E}[Y|\mathcal{F}]\mathbf{1}_A\big], \quad \forall A \in \mathcal{F}.$$

(xi) Define $Z_n = \sup_{k\ge n}|X_k - X|$ so that $0 \le Z_n \le 2Y$ and $Z_n \to 0, n \to \infty$. By usual DCT we find $\mathbb{E}[Z_n] \to 0$ for $n \to \infty$. The L^1-convergence can now be deduced as

$$\mathbb{E}\big[|\mathbb{E}[X_n\,|\,\mathcal{F}] - \mathbb{E}[X\,|\,\mathcal{F}]|\big] \overset{\Delta}{\le} \mathbb{E}\big[\mathbb{E}[|X_n - X|\,|\,\mathcal{F}]\big] = \mathbb{E}[|X_n - X|] \to 0, \quad n \to \infty$$

Next, towards the almost sure convergence. Since Z_n is decreasing in n the monotonicity implies that $\mathbb{E}[Z_n\,|\,\mathcal{F}]$ is decreasing. Denote the limit by M. With Fatou we get

$$0 \le \mathbb{E}[M] \le \liminf_{n\to\infty} \mathbb{E}[\mathbb{E}[Z_n\,|\,\mathcal{F}]] = \liminf_{n\to\infty} \mathbb{E}[Z_n] = 0.$$

Hence, $M = 0$ a.s. Finally, we can conclude

$$0 \le \lim_{n\to\infty}|\mathbb{E}[X_n\,|\,\mathcal{F}] - \mathbb{E}[X\,|\,\mathcal{F}]| \le \lim_{n\to\infty} \mathbb{E}[Z_n\,|\,\mathcal{F}] = M = 0.$$

(xii) Using the monotonicity yields that $0 \le \mathbb{E}[X_n|\mathcal{F}] \le \mathbb{E}[X_{n+1}|\mathcal{F}] \le ... \le \mathbb{E}[X|\mathcal{F}]$ so there is an almost sure limit $V := \lim_{n\to\infty} \mathbb{E}[X_n|\mathcal{F}]$ and $V \le \mathbb{E}[X|\mathcal{F}]$. Define $B = \{V < \mathbb{E}[X|\mathcal{F}]\}$, then

$$\begin{aligned}\mathbb{E}\big[\mathbb{E}[X|\mathcal{F}]\mathbf{1}_B\big] &\overset{B\in\mathcal{F}}{=} \mathbb{E}[X\mathbf{1}_B]\\ &\overset{\text{MCT}}{=} \lim_{n\to\infty} \mathbb{E}[X_n\mathbf{1}_B]\\ &\overset{(i)}{=} \lim_{n\to\infty} \mathbb{E}\big[\mathbb{E}[X_n\mathbf{1}_B|\mathcal{F}]\big]\\ &\overset{B\in\mathcal{F}}{=} \lim_{n\to\infty} \mathbb{E}\big[\mathbb{E}[X_n|\mathcal{F}]\mathbf{1}_B\big]\\ &\overset{\text{MCT}}{=} \mathbb{E}[V\mathbf{1}_B]\end{aligned}$$

But then $\mathbb{P}(B)$ must be 0 as otherwise the equalities could not hold.

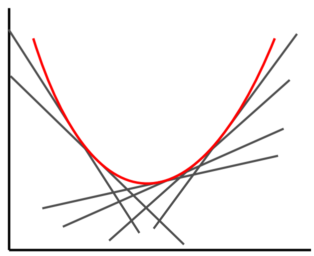
$\square$

Theorem 5.2.5. (Jensen's inequality for conditional expectation) Suppose φ is convex with $X, \varphi(X) \in \mathcal{L}^1(\Omega, \mathcal{A}, \mathbb{P})$, then

$$\varphi\big(\mathbb{E}[X \mid \mathcal{F}]\big) \leq \mathbb{E}[\varphi(X) \mid \mathcal{F}] \quad \text{a.s.}$$

Proof. Let $E_\varphi = \{(a,b) \in \mathbb{R}^2 : \varphi(x) \geq ax + b, \forall x\}$ the set of all subtangents. Then one (of many) ways to express the convexity is the expression

$$\varphi(x) = \sup_{(a,b)\in E_\varphi} (ax+b) = \sup_{(a,b)\in E_\varphi \cap \mathbb{Q}^2} (ax+b)$$



Representation of convex function through subtangents

Then,

$$\begin{aligned}
\mathbb{E}[\varphi(X) \mid \mathcal{F}] &= \mathbb{E}\Big[\sup_{(a,b)\in E_\varphi \cap \mathbb{Q}^2} (aX+b) \,\Big|\, \mathcal{F}\Big] \\
&\overset{\text{monotonicity}}{\geq} \sup_{(a,b)\in E_\varphi \cap \mathbb{Q}^2} \mathbb{E}[(aX+b) \mid \mathcal{F}] \\
&\overset{\text{lin.}}{=} \sup_{(a,b)\in E_\varphi \cap \mathbb{Q}^2} \big(a\mathbb{E}[X \mid \mathcal{F}] + b\big) \\
&= \varphi\big(\mathbb{E}[X \mid \mathcal{F}]\big).
\end{aligned}$$

There is a very important point to make in this calculation. Applying the calculation rules result in a.s. statements for all applications of the rules. The chain of equalities and inequalities holds on the intersection of those events of probability one. Since also their intersection has probability one, the chain and the statement holds almost surely. $\square$

We can now turn back towards our original interpretation of approximating random variables with random variables that carry less information. In general our approach to minimise $\mathbb{E}[(X-Y)^2]$ is not suitable as the expectation could be infinite. This is one of the reasons why in probability theory we tend to work with the abstract definition instead of an L^2-minimisation definition. Still, if we impose the extra assumption that X is square-integrable we indeed have the L^2-minimisation property:

Theorem 5.2.6. If X is square-integrable, then $\mathbb{E}[X|\mathcal{F}]$ is the orthogonal projection of X to $\mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$. Equivalently, the L^2-minimisation property holds:

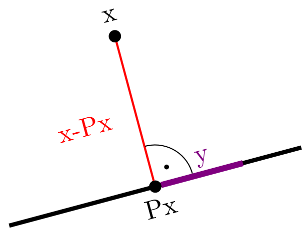
$$\mathbb{E}[(X-Y)^2] \geq \mathbb{E}[(X - \mathbb{E}[X \mid \mathcal{F}])^2], \quad \forall Y \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P}),$$

with equality if and only if $Y = \mathbb{E}[X|\mathcal{F}]$.

Proof. Let us recall from Functional Analysis the concept of an orthogonal projection. If $(H, \langle \cdot, \cdot \rangle)$ is a Hilbert space, G a closed subspace and $P : H \to G$ a linear operator. Then P is called an orthogonal projection if either

- $\langle y, x - Px \rangle = 0$ for all $y \in G$, or,
- $||x - y||_H \geq ||x - Px||_H$ for all $y \in G$.

Both properties can be best understood in a picture.



Now recall that $H := \mathcal{L}^2(\Omega, \mathcal{A}, \mathbb{P})$ (more precisely, their equivalence classes $L^2(\Omega, \mathcal{A}, \mathbb{P})$, see Theorem **??** and the discussion around) is a Hilbert space with $\langle X, Y \rangle = \mathbb{E}[XY]$. As a subspace we choose $G := \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ which is closed as limits of measurable maps do not loose the measurability. We first check that $P : X \mapsto \mathbb{E}[X|\mathcal{F}]$ is a mapping from H to G. The measurability is clear from the first property of conditional expectation, so we need to check the square-integrability. Jensen's inequality gives $\mathbb{E}[X|\mathcal{F}]^2 \leq \mathbb{E}[X^2|\mathcal{F}]$ a.s. so that monotonicity of expectations yields

$$\mathbb{E}\big[\mathbb{E}[X|\mathcal{F}]^2\big] \leq \mathbb{E}\big[\mathbb{E}[X^2|\mathcal{F}]\big] = \mathbb{E}[X^2] < \infty.$$

In order to prove the claimed minimisation property we proof the equivalent orthogonality property. This is easier as we can manipulate with linearity. Let $Y \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$, so that $\mathbb{E}[|XY|] < \infty$ by Cauchy-Schwarz. Then we can use the properties for expectations and conditional expectations to deduce

$$\begin{aligned}
\langle Y, X - \mathbb{E}[X|\mathcal{F}] \rangle &\overset{\text{lin., def.}}{=} \mathbb{E}[YX] - \mathbb{E}[Y\mathbb{E}[X|\mathcal{F}]] \\
&\overset{Y\ \mathcal{F}\text{-meas.}}{=} \mathbb{E}[YX] - \mathbb{E}[\mathbb{E}[YX|\mathcal{F}]] \\
&= \mathbb{E}[YX] - \mathbb{E}[YX] \\
&= 0.
\end{aligned}$$

□

We finish the abstract theory with an application to independence. Recall that two σ-algebras $\mathcal{F}_1, \mathcal{F}_2$ are called independent if all choices of pairs A, B from the σ-algebras are independent: $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. Equivalently, all pairs random variables that are measurable with respect to the σ-algebras are independent random variables.

Proposition 5.2.7. Two sub-σ-Algebras $\mathcal{F}_1$ and $\mathcal{F}_2$ of $\mathcal{A}$ are independent if and only if $\mathbb{E}[X \mid \mathcal{F}_1] = \mathbb{E}[X]$ a.s. for all $\mathcal{F}_2$-measurable $X \geq 0$

Proof. "⇒": Follows from Theorem 5.2.4 (vii).

"⇐": Taking $A \in \mathcal{F}_2$ and $B \in \mathcal{F}_1$ we need to show $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. All we use is the non-negative random variable $X = \mathbf{1}_A$. Using elementary properties of the expectation and Theorem 5.2.4 (iv)

$$\mathbb{P}(A \cap B) = \mathbb{E}[\mathbf{1}_{A \cap B}] = \mathbb{E}[\mathbb{E}[\mathbf{1}_A \mathbf{1}_B|\mathcal{F}_1]] \overset{B \in \mathcal{F}_1}{=} \mathbb{E}[\mathbf{1}_B \mathbb{E}[\mathbf{1}_A|\mathcal{F}_1]] \overset{\text{ass.}}{=} \mathbb{E}[\mathbf{1}_B \mathbb{E}[\mathbf{1}_A]] = \mathbb{P}(B)\mathbb{P}(A).$$

□

Example 5.2.8. Let $X_1, ..., X_n$ iid random variables on $(\Omega, \mathcal{A}, \mathbb{P})$ and $X = \sum_{k=1}^n X_k$. Then

$$\mathbb{E}[X_k|X] = \frac{X}{n} \text{ a.s.} \quad \text{and} \quad \mathbb{E}[X|X_1] = (n-1)\mathbb{E}[X_1] + X_1 \text{ a.s.}$$

Before we proof the identity by checking the definition we intuitively derive the results. If we know nothing about the X_k but the value of the sum then the best constant guess for each of the summands (using iid) is the value of the sum divided by n. The second is easier to guess. Fixing the value of X_1 the best constant guess of the sum is the expectation of the sum of $n-1$ copies plus the fixed values. Let us now check the claims rigorously. The second claim is just the linearity of conditional expectation combined with the elementary properties (iv) and (i). For the first claim first note that

$$\mathbb{E}[X_1|X] = ... = \mathbb{E}[X_n|X] \quad \text{a.s.} \tag{5.3}$$

Why? If $A \in \sigma(X)$, then we can write $\mathbf{1}_A = h(X) = f(X_1, ..., X_n)$ so that

$$\mathbb{E}[X_1 \mathbf{1}_A] = ... = \mathbb{E}[X_n \mathbf{1}_A],$$

because the iid assumptions yields $\mathbb{E}[g(X_1, ..., X_n)] = \mathbb{E}[g(X_{\sigma(1)}, ..., X_{\sigma(n)})]$ for all permutations σ. But then (5.3) follows from property (x) of conditional expectation. Now linearity used for the sum X implies $\mathbb{E}[X_1|X] = ... = \mathbb{E}[X_n|X] = \frac{1}{n}\mathbb{E}[X|X] = \frac{1}{n}X$.

5.3 Conditional expectation for random variables

The most important special case with many applications in statistics is $\mathbb{E}[X|Y]$, sometimes also more generally $\mathbb{E}[h(X,Y)|Y]$, the conditional expectation of an integrable random variable with respect to another random variable. Interestingly, we can use this conditional expectation to gain a much deeper understanding of random vectors than we have so far. Let us first recall from Stochastik 1 some definitions on pairs of random variables.

If (X, Y) is a random vector of two random variables X and Y then $\mathbb{P}_{(X,Y)}(A \times B) = \mathbb{P}(X \in A, Y \in B)$ was called the law of (X, Y). The law is a probability measure on the product space $(\mathbb{R} \times \mathbb{R}, \mathcal{B}(\mathbb{R}) \otimes \mathcal{B}(\mathbb{R}))$ which is uniquely determined by the joint distribution function $F_{(X,Y)}(t_1, t_2) = \mathbb{P}_{(X,Y)}((-\infty, t_1) \times (-\infty, t_2))$. In analogy to the case of one random variable, expectations $\mathbb{E}[h(X,Y)]$ were defined by $\int_\Omega h(X,Y)\,\mathrm{d}\mathbb{P}$ which, using the transformation formula and Fubini, equals $\int_\mathbb{R}\int_\mathbb{R} h(x,y)\mathbb{P}_{(X,Y)}(\mathrm{d}x, \mathrm{d}y)$. The formula simplifies a lot for independent random variables for which the law $\mathbb{P}_{(X,Y)} = \mathbb{P}_X \otimes \mathbb{P}_Y$ is a product measure and we can integrate the two coordinates separately. An important technical point was that it suffices to define measures on $\mathcal{B}(\mathbb{R}) \otimes \mathcal{B}(\mathbb{R})$ on rectangular sets $A \times B$ to obtain a measure on the entire σ-algebra, see the proof of Theorem 1.4.2. The key words are Dynkin-systems for uniqueness and Carathéodory for the existence.

In Stochastik 1 the structure of dependent random variables remained widely open, we only motivated dependence informally as "X influences Y or Y influences X" and defined dependent as not being independent. In this section conditional expectation are used to understand properly the dependence of random variables and to get a handy formula to computing conditional expectations.

Definition 5.3.1. Let $(R, \mathcal{R})$ and $(S, \mathcal{S})$ measurable spaces. Then a mapping $\kappa : R \times \mathcal{S} \to [0, \infty]$ is called a **Markov kernel** (or **transition kernel**) on $R \times \mathcal{S}$ if

(i) $y \mapsto \kappa(y, A)$ is $(\mathcal{R}, \mathcal{B}(\bar{\mathbb{R}}))$-measurable for all $A \in \mathcal{S}$,

(ii) $A \mapsto \kappa(y, A)$ is a probability measure for all $y \in R$.

We think of a kernel to be a measure parametrised by a real parameter, such as the law of the exponential distribution which is parametrised by λ. The defining properties of a kernel are best understood through the next proposition. The definition of the integral needs the measurability of the integrand and the measure property of the left hand side needs the measure property of the kernels.

Proposition 5.3.2. Suppose Y is a random variable and κ is a transition kernel on $\mathbb{R} \times \mathcal{B}(\mathbb{R})$, then there is a another random variable X such that the joint law of X and Y satisfies

$$\mathbb{P}_{(X,Y)}(A \times B) = \int_B \kappa(y, A)\, \mathbb{P}_Y(\mathrm{d}y), \quad A, B \in \mathcal{B}(\mathbb{R}). \tag{5.4}$$

Proof. All we need to do is to use the right hand side as a definition of a probability measure μ on $\mathcal{B}(\mathbb{R}) \otimes \mathcal{B}(\mathbb{R})$ through the distribution function

$$F(t_1, t_2) := \int_{(-\infty, t_2]} \kappa(y, (-\infty, t_1])\mathbb{P}_Y(\mathrm{d}y)$$

for all $t_1, t_2 \in \mathbb{R}$.

Check the properties of a multivariabe distribution function using dominated convergence and linearity/monotonicity of integrals.

Then Theorem 4.2.15 gives us a pair (X, Y) of random variables with distribution function F, hence, the joint law fulfills Equation (5.4). $\square$

Lecture 3

Kernels are not unique! If two kernels are only $\mathbb{P}_Y$-almost surely equal, then the integrals in (5.4) are the same.

Laws $\mathbb{P}_{(X,Y)}$ on the product space defined through kernels are the natural generalisation of product measures $\mathbb{P}_X \otimes \mathbb{P}_Y$, the laws of independent random variables. Indeed, if $\kappa(y, \cdot) = \mathbb{P}_X$ is independent of y, then (5.4) simplifies to

$$\mathbb{P}_{(X,Y)}(A \times B) = \int_B \mathbb{P}_X(A)\mathbb{P}_Y(\mathrm{d}y) = \mathbb{P}_X(A)\mathbb{P}_Y(B),$$

the formula describing independent random variables. The entire point of kernels is to understand better the concept of dependent random variables. A good probabilistic interpretation of random vectors (X, Y) with distribution (5.4) goes as two-stage experiment. First sample from Y and given the value of Y sample from X which distribution depends on the value y of Y.

Definition 5.3.3. We call two random variables X and Y a **two-stage experiment** with transition kernel κ if $\mathbb{P}_{(X,Y)}$ can be written in the disintegration form of (5.4).

The wording transition kernel makes much more sense with the notion of two-stage experiments in mind as κ describes the transition from the first stage to the second stage of the experiment.

Here is an instructive example. First choose uniformly p from $[0, 1]$ and then toss a coin with probability of success p. This simple two-stage experiment is modelled through

$$\mathbb{P}_Y \sim \mathcal{U}([0, 1]), \quad \kappa(p, A) = \big(p\delta_{\{1\}}(A) + (1 - p)\delta_{\{0\}}(A)\big)\mathbf{1}_{[0,1]}(p).$$

The two properties of a kernel are obviously fulfilled in this example.

It might be surprising, but **every** pair of random variables can be seen as a two-stage experiment!

Theorem 5.3.4. (Disintegration of (X,Y) into Y and a kernel)
Suppose X and Y are random variables on $(\Omega, \mathcal{A}, \mathbb{P})$, then there is a kernel κ on $\mathbb{R} \times \mathcal{B}(\mathbb{R})$ such that

$$\mathbb{P}_{(X,Y)}(A \times B) = \int_B \kappa(y, A)\, \mathbb{P}_Y(\mathrm{d}y), \quad A, B \in \mathcal{B}(\mathbb{R}). \tag{5.5}$$

The kernel is unique up to null-sets of $\mathbb{P}_Y$.

The formula appearing in the theorem is called a disintegration (‚Zerlegegung'), meaning the inverse of producing one measure from several measures as seen in Proposition 5.4. Again, keep in mind that κ is not unique! Changing κ on zero-sets of Y gives other disintegrations.

Proof. **Uniqueness:** Suppose there are two kernels $\kappa, \bar{\kappa}$. Setting $A = (-\infty, t]$ yields

$$\int_B \kappa(y, (-\infty, t])\, \mathbb{P}_Y(\mathrm{d}y) = \int_B \bar{\kappa}(y, (-\infty, t])\, \mathbb{P}_Y(\mathrm{d}y), \quad t \in \mathbb{Q}, B \in \mathcal{B}(\mathbb{R}).$$

This implies that there are sets $\mathcal{M}_t$ with $\mathbb{P}_Y(\mathcal{M}_t) = 1$ so that $\kappa(y, (-\infty, t]) = \bar{\kappa}(y, (-\infty, t])$ for all $z \in \mathcal{M}_t$. Defining $\mathcal{M} := \cap_{t \in \mathbb{Q}} \mathcal{M}_t$ yields equality of $\kappa(y, \cdot)$ and $\bar{\kappa}(y, \cdot)$ on an $\cap$-stable generator of $\mathcal{B}(\mathbb{R})$ for all $y \in \mathcal{M}$. Since equality on an $\cap$-stable generator implies equality of measures (see Theorem 1.2.12) we obtain equality $\kappa(y, \cdot) = \bar{\kappa}(y, \cdot)$ for all $y \in \mathcal{M}$. Since $\mathcal{M}$ is the intersection of countably many events of probability 1 the claim follows.

Existence: For every $r \in \mathbb{Q}$ let g_r the measurable mappings from the factorisation lemma so that

$$g_r(Y) := \mathbb{E}[\mathbf{1}_{(-\infty, r]}(X)|Y] \quad \text{a.s.}$$

From the monotonicity of conditional expectations (countably many!) we deduce

$$\mathbb{P}(g_r(Y) \leq g_s(Y)\ \forall r \leq s) = 1.$$

In other words, the set $E_1 := \{y \in \mathbb{R} : g_r(y) \leq g_s(y)\ \forall r \leq s\}$ has measure 1 under $\mathbb{P}_Y$. Similarly, using DCT for conditional expectations, the sets $E_2 := \{y \in \mathbb{R} : \inf_{r \in \mathbb{Q}} g_r(y) = 0\}$ and $E_3 := \{y \in \mathbb{R} : \sup_{r \in \mathbb{Q}} g_r(y) = 1\}$ have measure 1 under $\mathbb{P}_Y$. Now define

$$E := E_1 \cap E_2 \cap E_3$$

which again has measure 1 under $\mathbb{P}_Y$. What does this mean? For all $y \in E$ the mapping

$$r \mapsto g_r(y), \quad r \in \mathbb{Q},$$

is increasing and has the limits of a cumulative distribution function (CDF). Now we extend this to a family of CDFs. First choose the CDF $\mathbf{1}_{[0,\infty)}$ and define

$$F_y(t) := \begin{cases} \inf\{g_r(y) : r > t, r \in \mathbb{Q}\} & : y \in E \\ \mathbf{1}_{[0,\infty)}(t) & : y \notin E \end{cases}, \quad t \in \mathbb{R}.$$

We could have chosen any other CDF instead of $\mathbf{1}_{[0,\infty)}$ but this particular choice leads to the simple form $\kappa(y, A) = 0$ on the $\mathbb{P}_Y$-zero set in the formulas below. Now, for every $y \in \mathbb{R}$ the mapping $t \mapsto F_y(t)$ is increasing, with limits 0 and 1 and $-\infty$ and $+\infty$, and right-continuous. Writing down carefully these arguments requires a bit tedious analysis that does not provide deeper understanding, we prefer to skip them. Hence, the family $(F_y)_{y \in \mathbb{R}}$ is a family of CDFs. Additionally, for fixed $t \in \mathbb{R}$, the mapping $t \mapsto F_y(t)$ is measurable as it can be written as

$$y \mapsto F_y(t) = \underbrace{\mathbf{1}_{[0,\infty)}(t)\mathbf{1}_{E^C}(y)}_{\text{measurable}} + \underbrace{\mathbf{1}_E(y)}_{\text{measurable}} \underbrace{\inf\{g_r(y) : t \leq r, r \in \mathbb{Q}\}}_{\text{measurable}},$$

as a countable infimum of measurable functions is measurable. Now we are in a position to define the kernel. For every $y \in \mathbb{R}$ fixed we use the CDFs F_y to define the corresponding probability measure $\kappa(y,\cdot)$ as usually (compare Theorem 1.4.2):

$$\kappa(y,(a,b]) := F_y(b) - F_y(a), \quad a < b.$$

It is clear from this construction that $A \mapsto \kappa(y,A)$ are probability measures for all $y \in \mathbb{R}$. To prove that $y \mapsto \kappa(y,A)$ is measurable we use the trick of good sets (compare for instance the proof of Theorem 1.2.12). Define

$$\mathcal{M} := \big\{A \in \mathcal{A} \,|\, y \mapsto \kappa(y,A) \text{ is measurable}\big\}.$$

Then $\mathcal{M}$ contains an $\cap$-stable generator of $\mathcal{B}(\mathbb{R})$, namely the set $\mathcal{E} = \{(a,b] : a \leq b\}$, because $y \mapsto \kappa(y,(a,b]) = F_y(b) - F_y(a)$ is measurable as a difference of two measurable functions. Additionally, $\mathcal{M}$-is also a Dynkin-system. The complement property holds as $y \mapsto \kappa(y,A^C) = 1-\kappa(y,A)$ is measurable as a difference of two measurable functions. Now let $A_1,\dots$ be disjoint sets from $\mathcal{M}$. Then we see that

$$y \mapsto \kappa\Big(y, \bigcup_{k=1}^{\infty} A_k\Big) = \sum_{k=1}^{\infty} \kappa(y,A_k) = \lim_{n\to\infty} \sum_{k=1}^{n} \kappa(y,A_k)$$

is measurable as a limit of measurable maps. Hence, $\mathcal{M}$ is closed under disjoint unions. Now we can use the "trick of good sets":

$$\mathcal{A} = \sigma(\mathcal{E}) \overset{1.2.11}{=} d(\mathcal{E}) \subseteq d(\mathcal{M}) \subseteq \mathcal{A}$$

which implies $\mathcal{M} = \mathcal{A}$. In other words, $y \mapsto \kappa(y,A)$ is measurable for all $A \in \mathcal{A}$.

To finish the proof we only need to check the identity (5.5) on some $\cap$-stable generator of $\mathcal{B}(\mathbb{R}) \otimes \mathcal{B}(\mathbb{R})$. We use the set of infinite rectangles $(-\infty,t] \times (-\infty,s]$ with rational end-points:

$$\begin{aligned}
\int_{(-\infty,s]} \kappa(y,(-\infty,t])\mathbb{P}_Y(\mathrm{d}y) &= \int_{\mathbb{R}} \kappa(y,(-\infty,t])\mathbf{1}_{(-\infty,s]}(y)\mathbb{P}_Y(\mathrm{d}y) \\
&= \mathbb{E}\left[\kappa(Y,(-\infty,t])\mathbf{1}_{(-\infty,s]}(Y)\right] \\
&\overset{\text{def. } \kappa}{=} \mathbb{E}\left[\mathbb{E}[\mathbf{1}_{(-\infty,t]}(X)|Y]\mathbf{1}_{(-\infty,s]}(Y)\right] \\
&\overset{\text{meas.}}{=} \mathbb{E}\left[\mathbb{E}[\mathbf{1}_{(-\infty,t]}(X)\mathbf{1}_{(-\infty,s]}(Y)|Y]\right] \\
&\overset{\mathbb{E} \text{ of cond. exp.}}{=} \mathbb{E}[\mathbf{1}_{(-\infty,t]}(X)\mathbf{1}_{(-\infty,s]}(Y)] \\
&= \mathbb{P}(X \leq t, Y \leq s) \\
&= \mathbb{P}_{(X,Y)}((-\infty,t] \times (-\infty,s]).
\end{aligned}$$

□

There are several situations in which the kernels κ can be guessed and have a nice form. The most important examples are absolutely continuous random vectores:

If X,Y have a joint density f, then

$$\kappa(y,A) = \int_A \frac{f(x,y)}{f_y(y)} \mathbf{1}_{f_y(y)>0}\,\mathrm{d}x, \quad y \in \mathbb{R}, A \in \mathcal{B}(\mathbb{R}),$$

where $f_y(y) = \int_{\mathbb{R}} f(x,y)\,\mathrm{d}x$.

The formula follows directly from the standard calculation rules of probability theory:

$$\begin{aligned}
\mathbb{P}_{(X,Y)}(A \times B) = \int_{A\times B} f(x,y)\,\mathrm{d}(x,y) &= \int_B \Big(\int_A \frac{f(x,y)}{f_y(y)}\mathbf{1}_{f_y(y)>0}\,\mathrm{d}x\Big) f_y(y)\,\mathrm{d}y \\
&= \int_B \kappa(y,A)\mathbb{P}_Y(\mathrm{d}y).
\end{aligned}$$

If Y is discrete with values $a_1, ..., a_N$, then

$$\begin{aligned}\kappa(y, A) &= \sum_{k=1}^{N} \mathbb{P}(X \in A|Y = y)\mathbf{1}_{\{a_k\}}(y) \\ &= \begin{cases} \mathbb{P}(X \in A|Y = a_k) & : y = a_k \\ 0 & : \text{otherwise} \end{cases}, \quad y \in \mathbb{R}, A \in \mathcal{B}(\mathbb{R}).\end{aligned}$$

The formula is a direct consequence of the formula of total probability:

$$\begin{aligned}\mathbb{P}_{(X,Y)}(A \times B) &= \mathbb{P}(X \in A, Y \in B) \\ &= \sum_{a_k \in B} \mathbb{P}(X \in A|Y = a_k)\mathbb{P}(Y = a_k) \\ &\overset{3.3.3}{=} \int_{\mathbb{R}} \kappa(y, A)\, \mathbb{P}_Y(\mathrm{d}y).\end{aligned}$$

With the powerful disintegration theorem in hands we can understand how to compute with conditional expectations for random variables. Keeping in mind the explicit formulas in special cases above, the theorem allows us to perform many explicit computations.

Theorem 5.3.5. Suppose X and Y are random variables, X integrable, and $h : \mathbb{R}^2 \to \mathbb{R}$ is measurable, then

(i) $\mathbb{E}[h(X, Y)|Y] = \int_{\mathbb{R}} h(x, Y)\, \kappa(Y, \mathrm{d}x)$ a.s.,

(ii) $\mathbb{E}[X|Y] = \int_{\mathbb{R}} x\, \kappa(Y, \mathrm{d}x)$ a.s.,

where κ is the kernel from Theorem 5.3.4

Proof. We only need to prove the first claim, the second follows by choosing $h(x, y) = x$. As usually we check the two defining properties of conditional expectation. The righthand side can be written as $\phi(Y)$ with

$$\phi(z) = \int_{\mathbb{R}} h(x, z)\, \kappa(z, \mathrm{d}x).$$

As ϕ is measurable (approximate the integral by sums) we immediately get that $\phi(Y)$ is $\sigma(Y)$-measurable. For the expectation property fix some $A \in \sigma(Y)$. According to the factorisation lemma there is a Borel map g so that $\mathbf{1}_A = g(Y)$. Now we compute:

$$\begin{aligned}\mathbb{E}[\mathbf{1}_A \phi(Y)] &= \mathbb{E}\Big[g(Y) \int_{\mathbb{R}} h(x, Y)\, \kappa(Y, \mathrm{d}x)\Big] \\ &= \mathbb{E}\Big[\int_{\mathbb{R}} g(Y) h(x, Y)\, \kappa(Y, \mathrm{d}x)\Big] \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} g(y) h(x, y)\, \kappa(y, \mathrm{d}x)\, \mathbb{P}_Y(dy) \\ &\overset{(*)}{=} \int_{\mathbb{R}} \int_{\mathbb{R}} g(y) h(x, y) \mathbb{P}_{(X,Y)}(\mathrm{d}x, \mathrm{d}y) \\ &= \mathbb{E}[g(Y) h(X, Y)] \\ &= \mathbb{E}[\mathbf{1}_A h(X, Y)].\end{aligned}$$

The equality $(*)$ needs some clarification. For indicator functions $f = \mathbf{1}_{A \times B}$ of rectangle sets the equality

$$\int_{\mathbb{R}} \int_{\mathbb{R}} f(x, y)\, \kappa(y, \mathrm{d}x)\, \mathbb{P}_Y(dy) = \int_{\mathbb{R}} \int_{\mathbb{R}} f(x, y)\, \mathbb{P}_{(X,Y)}(\mathrm{d}x, \mathrm{d}y) \tag{5.6}$$

holds by the definition of the kernel. Now define $\mathcal{M} := \{M \in \mathcal{B}(\mathbb{R}^2) : (5.6) \text{ holds for } f = \mathbf{1}_M\}$. Using that indicators over disjoint unions are sums over indicators and monotone convergence we see that $\mathcal{M}$ is a Dynkin-system containing the $\cap$-stable generator $\mathcal{E} := \{A \times B : A, B \in \mathcal{B}(\mathbb{R})\}$ of the Borel-σ-algebra. But then, as usually

$$\mathcal{B}(\mathbb{R}^2) = \sigma(\mathcal{E}) \overset{1.2.11}{=} d(\mathcal{E}) \subseteq d(\mathcal{M}) = \mathcal{M} \subseteq \mathcal{B}(\mathbb{R}^2).$$

Hence, Equality (5.6) holds for all indicators over $M \in \mathcal{B}(\mathbb{R}^2)$. But then it holds for all simple functions, with monotone convergence for all non-negative measurable functions and by splitting $f = f^+ - f^-$ for all measurable functions. Finally, we apply (5.6) with the function $f(x,y) = g(y)h(x,y)$. $\square$

The theorem can be used for abstract considerations but also for explicit computations. To perform explicit computations the kernel κ needs to be known. We collected several examples of kernels above, they cover most of the relevant situations in applications. Here are some examples to try:

(i) $\mathbb{E}[(Y-X)^+|X]$, where X, Y are independent uniform random variables on $[0,1]$.

(ii) $\mathbb{E}[X|X+Y]$, where X, Y are independent Poisson random variables with parameters λ and μ respectively.

(iii) $\mathbb{E}[X|Y]$, where X, Y has a joint density

$$f(x,y) = 4y(x-y)e^{-(x+y)}\mathbf{1}_{0<y<x}.$$

The kernel κ can be used to give a rigorous discussion of regular conditional distributions, the distributions behind the conditional expectations. The wording regular refers to the measurability in y of the kernel.

Definition 5.3.6. (Regular conditional distributions)
Let X and Y be random variables and $A \in \mathcal{B}(\mathbb{R})$. If κ is the kernel from Theorem 5.3.4, then we define

$$\mathbb{P}(X \in A|Y) = \kappa(Y, A), \quad A \in \mathcal{B}(\mathbb{R}),$$

and

$$\mathbb{P}(X \in A|Y = y) := \kappa(y, A), \quad y \in \mathbb{R}, A \in \mathcal{B}(\mathbb{R}).$$

Recall that kernels are not unique as they can differ on zero sets of $\mathbb{P}_Y$. Hence, also $\mathbb{P}(X \in A|Y = y)$ is only well-defined as a function in y up to null-sets of $\mathbb{P}_Y$. For the probabilistic intuition this perfectly makes sense. If Y does not take the value y why should we care about $\mathbb{P}(X \in A|Y = y)$?

It is always instructive to check some examples. Use the discrete and absolutely formulas for κ to compute the following conditional laws!

(i) Let Z_1, Z_2 be independent Poisson-distributed random variables with parameter $\lambda_1, \lambda_2 > 0$. Check that

$$\mathbb{P}\big(Z_1 = k \,\big|\, Z_1 + Z_2 = n\big) = b_{n,p}(k), \quad k = 0, 1, ...,$$

with $b_{n,p}(k) \sim \mathrm{Bin}(n,p)$ and $p = \frac{\lambda_1}{\lambda_1+\lambda_2}$.

(ii) What is

$$\mathbb{P}\big(Z_1 \in \cdot \,|\, Z_1 + Z_2 = x\big)$$

if Z_1 and Z_2 are independent standard Gaussians? Define $X = Z_1, Y = Z_1+Z_2$, find $f_{x,y}$ and then compute with the formulas from above. There is also a more clever way to avoid computations somewhere hidden in these notes!

Let us compare the definition with Theorem 5.3.5 in the special case $h(x, y) = \mathbf{1}_A(y)$. Then we see immediately the connection

$$\mathbb{E}[\mathbf{1}_A(X)|Y] = \kappa(Y, A) = \mathbb{P}(X \in A|Y) \quad \text{a.s.},$$

which is exactly the classical connection $\mathbb{E}[\mathbf{1}_A(X)] = \mathbb{P}(X \in A)$ between expectations and probabilities. Hence, the notion $\mathbb{P}(X \in A|Y)$ makes a lot of sense. There are other ways of construction $\mathbb{P}(X \in A|Y)$ more directly using the conditional expectation. One could for instance define $\mathbb{P}(X \in A|Y) := \mathbb{E}[\mathbf{1}_A(X)|Y]$ but quickly runs into problems with zero sets that depend on A. Similar to the proof of Theorem 5.3.4 the issue is resolved by defining the measures $\mathbb{P}(X \in (a, b]|Y)$ for rational end-points combined with a suitable extension to the reals. We do this in the next section to define $\mathbb{P}(X \in A|\mathcal{F})$ for general σ-algebras.

Due to the definitions through the kernel the function $y \mapsto \mathbb{P}(X \in A|Y = y)$ is the measurable mapping from the factorisation lemma that satisfies $h(Y) = \mathbb{P}(X \in A|Y)$. It is common practice to use this connection as an abstract definition of $\mathbb{P}(X \in A|Y = y)$. We prefer the construction through disintegration as the disintegration formula with $B = \mathbb{R}$ directly translates into the meaningful formula

$$\mathbb{P}(X \in A) = \int_{\mathbb{R}} \mathbb{P}(X \in A|Y = y)\mathbb{P}_Y(\mathrm{d}y), \quad A \in \mathcal{B}(\mathbb{R}),$$

a formula that we understand very well in the discrete setting through the formula of total probabilities (recall Theorem **??**):

$$\mathbb{P}(X \in A) = \sum_{k=1}^{N} \mathbb{P}(X \in A|Y = a_k)\mathbb{P}(Y = a_k).$$

It also makes sense to replace the kernel in Theorem 5.3.5 with the newly defined conditional probability to get a more intuitive feeling for the formulae:

$$\mathbb{E}[h(X, Y)|Y] = \int_{\mathbb{R}} h(x, Y)\, \mathbb{P}(X \in \mathrm{d}x|Y) \quad \text{a.s.}$$
$$\mathbb{E}[X|Y] = \int_{\mathbb{R}} x\, \mathbb{P}(X \in \mathrm{d}x|Y) \quad \text{a.s.}$$

Just as $\mathbb{P}(X \in A|Y)$ is related through integration to $\mathbb{E}[h(X, Y)|Y]$ we can ask if $\mathbb{P}(X \in A|Y = y)$ is related to an object $\mathbb{E}[h(X, Y)|Y = y]$. Indeed, this is the case in exactly the way we have seen before.

Theorem 5.3.7. Suppose X and Y are random variables, X integrable, and $h : \mathbb{R}^2 \to \mathbb{R}$ is measurable, then

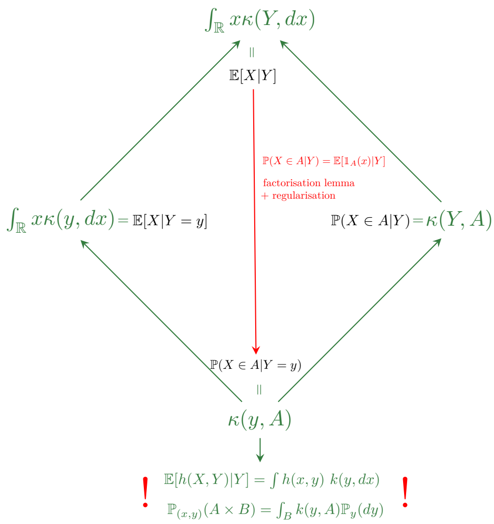
(i) $\mathbb{E}[h(X, Y)|Y = y] := \int_{\mathbb{R}} h(x, y)\, \mathbb{P}(X \in \mathrm{d}x|Y = y)$,

(ii) $\mathbb{E}[X|Y = y] := \int_{\mathbb{R}} x\, \mathbb{P}(X \in \mathrm{d}x|Y = y)$

are the measurable functions from the factorisation lemma that turn Y into $\mathbb{E}[h(X, Y)|Y]$ and $\mathbb{E}[X|Y]$.

Proof. The right hand sides are measurable functions in y and if we plug-in Y then we obtain the conditional expectations by Theorem 5.3.5. ☐

We have defined rigorously a couple of objects: $\mathbb{E}[X|Y]$, $\mathbb{E}[X|Y=y]$, $\mathbb{P}(X \in A|Y)$, and $\mathbb{P}(X \in A|Y=y)$ that are connected through integration and the factorisation lemma. Everything was based on the "most fine" object $\kappa(y,\cdot)$ from which we could reconstruct in a compact way all objects that relate to X and Y. Explicit formulas in the most important special case allow explicit computations with conditional expectations and probabilities. The following diagram should give an overview over the relations.



The logic behind the disintegration kernel and conditional expectation

We finish the section with a discussion on conditioning on zero sets. A bit of care is needed about the interpretation of $\mathbb{P}(X \in A|Y=y)$ as typically the event $\{Y=y\}$ has probability zero. One can easily get confused about the notation as the elementary conditional probability $\mathbb{P}(A|B) = \frac{\mathbb{P}(A\cap B)}{\mathbb{P}(B)}$ is not defined for null-sets B. The point is that we did not use the elementary definition of conditional probability. Instead, we constructed in some obscure way a mapping $y \mapsto \mathbb{P}(X \in A|Y=y)$ that satisfies a general total probability rule

$$\mathbb{P}(X \in A) = \int_{\mathbb{R}} \mathbb{P}(X \in A|Y=y)\, \mathbb{P}_Y(\mathrm{d}y), \quad A \in \mathcal{B}(\mathbb{R}).$$

In general there is no reason to expect any relation to elementary conditioned probabilities. Nonetheless, there are two situation in which the definition $\mathbb{P}(X \in A|Y=y) = \kappa(y,A)$ coincides with elementary conditioning and that's why we abuse the notation of conditional probability.

If Y is discrete, then the above formulas give

$$\mathbb{P}(X \in A|Y=y) = \begin{cases} \mathbb{P}(X \in A|Y=a_k) & : y = a_k \\ 0 & : \text{otherwise} \end{cases}$$

with elementary conditional probability on the right hand side.

If X and Y are jointly absolutely continuous with density $f > 0$ there is another way to define $\mathbb{P}(X \in A|Y=y)$ through a limiting procedure. The result is exactly the same:

$$\begin{aligned} \mathbb{P}(X \in A|Y=y) &:= \lim_{\varepsilon\to 0} \mathbb{P}(X \in A|Y \in (y-\varepsilon, y+\varepsilon)) \\ &= \lim_{\varepsilon\to 0} \frac{\int_{y-\varepsilon}^{y+\varepsilon} \int_A f(x,y)\,\mathrm{d}x\,\mathrm{d}y}{\int_{y-\varepsilon}^{y+\varepsilon} f_y(y)\,\mathrm{d}x} \\ &\overset{\text{l'Hospital}}{=} \frac{\int_A f(x,y)\,\mathrm{d}x}{f_y(y)} = \kappa(y,A), \end{aligned}$$

with the kernel for absolutely continuous random vectors.

5.4 General regular conditional distributions

Not part of the course

We finish the discussion of conditional expectation with a generalisation of Theorem 5.3.5 towards conditional expectations on a general σ-algebra $\mathcal{F}$. We want to provide a computation rule

$$\mathbb{E}[h(X)|\mathcal{F}] = \int_{\mathbb{R}} h(x)\,\mathbb{P}(X \in \mathrm{d}x|\mathcal{F}) \quad \text{a.s.} \tag{5.7}$$

for all measurable h. For $\mathcal{F} = \sigma(Y)$ we proved such a formula using the "random" measures $\mathbb{P}(X \in \mathrm{d}x|Y) = \kappa(Y, \mathrm{d}x)$. The general setting does not provide a kernel $\mathbb{P}(X \in A|Y = y) = \kappa(y, A)$ to work with but still allows us to construct a "random" measure satisfying (5.7)

Definition 5.4.1. Suppose $(\Omega, \mathcal{A}, \mathbb{P})$ is a probability space. A Markov kernel μ on $\Omega \times \mathcal{B}(\mathbb{R})$ is called a **random measure**.

The simplest example of a random measure already appeared in the previous section. If κ is a kernel and Y is a random variable, then $\kappa(Y, \cdot)$ is a random measure. The measure property is clear, the measurability follows as the concatenation of Y and κ is measurable again.

Random measures can be found at very different places in statistics and probability theory under different names. For conditional expectations we use random measures to define regular conditional distributions:

Definition 5.4.2. Let X be an integrable random variable on $(\Omega, \mathcal{A}, \mathbb{P})$ and $\mathcal{F} \subseteq \mathcal{A}$ a sub-σ-algebra. A random measure μ on $\Omega \times \mathcal{B}(\mathbb{R})$ is called a regular version of the conditional distribution (or a **regular conditional distribution**) of X given $\mathcal{F}$ if

$$\mu(\omega, B) = \mathbb{E}\big[\mathbf{1}_B(X)\big|\mathcal{F}\big](\omega) \quad \text{a.s.}$$

for all $B \in \mathcal{B}(\mathbb{R})$. We will write $\mathbb{P}(X \in B|\mathcal{F})(\omega)$, $\mathbb{P}_{X|\mathcal{F}}(B)(\omega)$, or $\mathbb{E}[\mathbf{1}_B(X)\big|\mathcal{F}](\omega)$ instead of $\mu(\omega, B)$ but often skip the dependence of ω just as we usually do for random variables.

Going back again to the special case $\mathcal{F} = \sigma(Y)$ from the previous section, we obtain our first example. Theorem 5.3.5 showed that

$$\mathbb{E}[\mathbf{1}_B(X)|Y] = \kappa(Y, B) \overset{\text{Notation}}{=} \mathbb{P}(X \in B|Y),$$

which is exactly the formula we want to generalise for $\mathcal{F}$.

Let us now prove the existence of a regular conditional expectation for general σ-algebras:

Theorem 5.4.3. If X is an integrable random variable on $(\Omega, \mathcal{A}, \mathbb{P})$ and $\mathcal{F} \subseteq \mathcal{A}$ is a sub-σ-algebra, then there exists a regular conditional distribution μ of X given $\mathcal{F}$.

Proof. Our strategy is as follows:

(i) Define a candidate measure $\mu(\omega, \cdot)$.

(ii) Check that μ is a random measure on $\Omega \times \mathcal{B}(\mathbb{R})$.

(iii) Check for all $B \in \mathcal{B}(\mathbb{R})$ that $\mu(\cdot, B)$ is a version of $\mathbb{E}[\mathbf{1}_B(X)|\mathcal{F}]$.

Step (i): The main idea of the argument is that measures on $\mathcal{B}(\mathbb{R})$ are uniquely determined by their commulative distribution function (Dynkin-Systems + Carathéodory, see Theorem 1.4.2). For fixed $t \in \mathbb{Q}$ define the $\mathcal{F}$-measurable random variable

$$F_\omega(t) := \mathbb{E}\big[\mathbf{1}_{(-\infty,t]}(X)|\mathcal{F}\big](\omega)$$

as an arbitrary version of the conditional expectation which exists by Theorem 5.2.3. Using monotonicity of conditional expectations and the countability of $\mathbb{Q}$, there is a measurable set $\mathcal{M}^1 \in \mathcal{A}$ with $\mathcal{P}(\mathcal{M}^1) = 1$ so that

$$t \mapsto F_\omega(t) \quad \text{is increasing on } \mathbb{Q} \text{ for all } \omega \in \mathcal{M}^1.$$

The set $\mathcal{M}^1$ is the intersection of the countably many sets $\mathcal{M}_{t,s}$ of measure 1 on which Theorem 5.2.4 (iii) gives $F_\omega(s) \leq F_\omega(t)$ for $s < t$. Using dominated convergence for conditional expectations (Theorem 5.2.4 (ix)) we find another measurable set $\mathcal{M}^2$ with $\mathbb{P}(\mathcal{M}^2) = 1$ so that

$$\lim_{t \to +\infty, t \in \mathbb{Q}} F_\omega(t) = 1 \quad \text{and} \quad \lim_{t \to -\infty, t \in \mathbb{Q}} F_\omega(t) = 0 \quad \text{for all } \omega \in \mathcal{M}^2.$$

Also with dominated convergence we find another measurable set $\mathcal{M}^3$ with $\mathbb{P}(\mathcal{M}^3) = 1$ so that right-continuity holds on $\mathbb{Q}$:

$$\lim_{t \downarrow s, t \in \mathbb{Q}} F_\omega(t) = F_\omega(s) \quad \text{for all } s \in \mathbb{Q} \text{ and } \omega \in \mathcal{M}^3.$$

If now we fix some arbitrary distribution function $\tilde{F}$ and define $F_\omega(t) := \tilde{F}(t)$ for $\omega \notin \mathcal{M}^3$, then we have defined $F_\omega(t)$ for all $t \in \mathbb{Q}$ and all $\omega \in \Omega$ such that $t \mapsto F_\omega(t)$ satisfies the properties of distribution functions on $\mathbb{Q}$ for all $\omega \in \Omega$ and $F_\cdot(t)$ is a version of $\mathbb{E}[\mathbf{1}_{(-\infty,t]}(X)|\mathcal{F}]$ for all $t \in \mathbb{Q}$.

Now we extend $\mathbb{Q}$ to $\mathbb{R}$ by defining

$$\overline{F}_\omega(s) := \inf\left\{F_\omega(t)\middle|t > s\,,\, t \in \mathbb{Q}\right\}, \quad s \in \mathbb{R}, \omega \in \Omega.$$

It is a bit tedious (basic analysis) to show that $t \mapsto \overline{F}_\omega(t)$ inherits the properties of a distribution function from F_ω. Hence, by Theorem 1.4.2 there are probability measures $\mu(\omega, \cdot)$ on $\mathcal{B}(\mathbb{R})$ for all $\omega \in \Omega$.

Steps (ii) and (iii): We use our favorite argument of "good sets" from measure theory (compare the proof of Theorem 1.2.12). Let

$$\mathcal{B} = \{A \in \mathcal{A} : \mu(\cdot, A) \text{ is a version of } \mathbb{E}[\mathbf{1}_A(X)|\mathcal{F}] \text{ and } \omega \mapsto \mu(\omega, A) \text{ is a random measure}\}.$$

We have to prove that $B = \mathcal{B}(\mathbb{R})$ is a Dynkin-system and contains an $\cap$-stable generator of $\mathcal{B}(\mathbb{R})$. This works essentially as in the proof of Theorem 5.3.4.

□

We can now come back to the motivation and compute different conditional expectations just the way we are used to for the classical expectation:

Proposition 5.4.4. Suppose $h : \mathbb{R} \mapsto \mathbb{R}$ is Borel-measurable such that $h(X) \in \mathcal{L}^1(\Omega, \mathcal{A}, \mathbb{P})$, then

$$\mathbb{E}[h(X)|\mathcal{F}] = \int_{\mathbb{R}} h(x)\, \mathbb{P}_{X|\mathcal{F}}(\mathrm{d}x) \quad \text{a.s.}$$

Proof. If $f = \mathbf{1}_B$, $B \in \mathcal{B}(\mathbb{R})$, the statement follows from the definition of $\mathbb{P}_{X|\mathcal{F}}$. Now let $h \geq 0$

measurable and $h_n = \sum_{k=1}^{n} \alpha_k \mathbb{1}_{B_k}$ a sequence of simple functions with $h_n \uparrow h$. Then, for $A \in \mathcal{F}$,

$$
\begin{aligned}
\mathbb{E}\big[\mathbb{1}_A h(X)\big] &\overset{\text{MCT}}{=} \lim_{n\to\infty} \mathbb{E}\big[\mathbb{1}_A h_n(X)\big] \\
&\overset{\text{lin.}}{=} \lim_{n\to\infty} \sum_{k=1}^{n} \alpha_k \mathbb{E}\big[\mathbb{1}_A \mathbb{1}_{B_k}(X)\big] \\
&\overset{A\in\mathcal{F}}{=} \lim_{n\to\infty} \sum_{k=1}^{n} \alpha_k \mathbb{E}\big[\mathbb{1}_A \mathbb{E}[\mathbb{1}_{B_k}(X) \,|\, \mathcal{F}]\big] \\
&= \lim_{n\to\infty} \sum_{k=1}^{n} \alpha_k \mathbb{E}\big[\mathbb{1}_A \mathbb{P}_{X|\mathcal{F}}(B_k)\big] \\
&\overset{\text{lin.}}{=} \lim_{n\to\infty} \mathbb{E}\Big[\mathbb{1}_A \sum_{k=1}^{n} \alpha_k \mathbb{P}_{X|\mathcal{F}}(B_k)\Big] \\
&= \lim_{n\to\infty} \mathbb{E}\Big[\mathbb{1}_A \int_{\mathbb{R}} h_n(x)] P_{X|\mathcal{F}}(\mathrm{d}x)\Big] \\
&\overset{\text{2x MCT}}{=} \mathbb{E}\Big[\mathbb{1}_A \int_{\mathbb{R}} h(x) \mathbb{P}_{X|\mathcal{F}}(\mathrm{d}x)\Big].
\end{aligned}
$$

This proves the expectation property of conditional expectation. With the same approximation we also obtain the measurability property as $\int_{\mathbb{R}} h(x)\mu(\cdot, \mathrm{d}x)$ is a pointwise limit of $\sum_{k=1}^{n} \alpha_k P_{X|\mathcal{F}}(B_k)$ which are all $\mathcal{F}$-measurable. Hence, the right-hand side satisfies both defining properties of $\mathbb{E}[\big[h(X)|\mathcal{F}\big]$ and as such is a version of the conditional expectation. For general h we write $h = h^+ - h^-$ and use linearity as usual. $\square$

$$
\begin{tikzcd}
\mathbb{E}[X|\mathcal{F}] \arrow[d, "{\mathbb{P}(X \in A|\mathcal{F}) = \mathbb{E}[\mathbb{1}_A(X)|\mathcal{F}] + \text{regularisation}}"] \\
\mathbb{P}(X \in A|\mathcal{F}) \arrow[d, "\text{approximate by simple functions}"] \\
\mathbb{E}[h(X)|\mathcal{F}] = \int_{\mathbb{R}} h(x)\mathbb{P}(X \in dx|\mathcal{F})
\end{tikzcd}
$$

The logic behind the regular conditional probability

As soon as you attend a lecture on general Markov processes, the object of regular conditional expectation will be central to define the Markov property as

$$\mathbb{P}(X_{t+s} \in A|\mathcal{F}_t) = \mathbb{P}(X_{t+s} \in A|X_t) \quad \text{a.s.},$$

where $\mathcal{F}_t = \sigma(X_s : s \leq t)$ is the information of the process up to time t. This is a way of formalising the idea of the future distribution given the past is equal to the future distribution given the entire past.

It is typically a good exercise to check Markov inequalities as the proofs require the standard tools of probability theory. Do it!

Let $h : [0, \infty) \to [0, \infty)$ increasing, then

$$\mathbb{P}(|X| > \epsilon | \mathcal{F}) \leq \frac{\mathbb{E}[h(X)|\mathcal{F}]}{h(\epsilon)} \quad \text{a.s.}$$

Kapitel 6

Martingale theory

Lecture 4

Most probabilists will agree to say their favorite application of conditional expectation is martingales. In this chapter we will discuss martingale theory in discrete time and, as a rather simple application, give a proof of the strong law of large numbers from Section **??**.

6.1 Introduction to discrete-time stochastic processes

Before turning to the special class of martingales we will fix some notation of stochastic processes and prove some elementary facts about stopping times.

Definition 6.1.1. Suppose $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space, $(E, \mathcal{E})$ is a measurable space, and I is an index set. Then a family of random variables $X = (X_t)_{t\in I}$ on $(\Omega, \mathcal{F}, \mathbb{P})$ with values in $(E, \mathcal{E})$ is called a **stochastic process** with values in E, called the state-space.

This is for the first time that we use the name random variable more freely. In Chapter **??** a random variable was defined to be a real-valued (or $\bar{\mathbb{R}}$-valued) measurable mapping, a vector-valued random variable was called random vector. From now on we will call all E-valued $(\mathcal{F}, \mathcal{E})$-measurable mappings random variables in order to avoid the use of even more notation. Authors prefer to use the wording random element if the state-space is different from $\mathbb{R}$.

Most of the time we see the ordered index sets

$$I = \mathbb{R}, \quad I = [0, \infty), \quad I = [0, T]$$

or

$$I = \mathbb{Z}, \quad I = \mathbb{N}_0 = \{0, 1, 2, ...\}, \quad I = -\mathbb{N}_0 = \{..., -1, 0\}, \quad I = \mathbb{N}, \quad I = \{1, ..., N\}$$

which are interpreted as time. Most of the time there is a first element such as 0 and we interpret the time running forwards. If there is a last element such as 0, we speak of a backwards process which is running from the past to present time.

Definition 6.1.2. If I is a discrete set (there is no accumulation point), a stochastic process indexed by I is called a **discrete-time stochastic process**.

Just as for random vectors there are two ways of seeing a stochastic process. Either as a sequence of random variables or as a function-valued random variable:

Definition 6.1.3. If X is an E-valued stochastic process indexed by I, then a realisation

$$X(\omega) : t \mapsto X_t(\omega)$$

is called a path (or sample **path**, or trajectory) of X. The set of all paths (functions from I to E) is denoted by $E^I := \{f : I \to E\}$. We always equip E^I with the **σ-algebra of so-called cylinder** sets generated by the finite projections. That is,

$$\mathcal{E}^I := \sigma(\{\pi_\alpha^{-1}(B) : \alpha \subseteq I, |\alpha| < \infty, B \in \mathcal{E}^{|\alpha|}\}),$$

where $\pi_\alpha(f) = (f(\alpha_1), ..., f(\alpha_{|\alpha|}))$ takes a path and gives the values of the path at finitely many given time points.

To get a feeling of the cylinder sets let us fix a vector $\alpha = (\alpha_1, ..., \alpha_4)$ of four time points, a Borel-set in $\mathbb{R}^4$ (let's say a four dimensional cube $I_1 \times ... \times I_4$) and visualise the corresponding cylinder set: [1]

It is important for the understanding to keep in mind that $\mathbb{R}^I$ is nothing but $\mathbb{R}^d$ if $|\alpha| = d$. There is no difference between a vector and a mapping with finitely many variables! We also like to use the notation $\mathbb{R}^\infty$ instead of $\mathbb{R}^\mathbb{N}$ and keep in mind that this is just a different notation for the set of real sequences.

The path space $(E^I, \mathcal{E}^I)$ is a nice measurable space if I is discrete but not nice at all at all if I is not discrete. We will get back to the continuous case only when we start to discuss the Brownian motion.

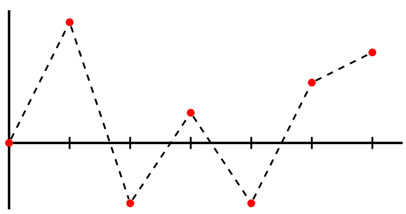
In the discrete-time case one can check just as in Proposition 4.2.11 that the notion of a stochastic process as a family of random variables is equivalent to that as a function-valued random variable:

$$X \text{ is } (\mathcal{F}, \mathcal{E}^I)\text{-measurable} \quad \Leftrightarrow \quad X_t \text{ is } (\mathcal{F}, \mathcal{E})\text{-measurable for all } t \in I.$$

One can prove this fact by checking the measurability on a generator of $\mathcal{E}^I$, namely the 1-cylinder sets. To see that the 1-cylinder sets generate $\mathcal{E}^I$ one only needs to note that all cylinder sets can be obtained by intersecting 1-cylinder sets.

Here are some examples of stochastic processes:

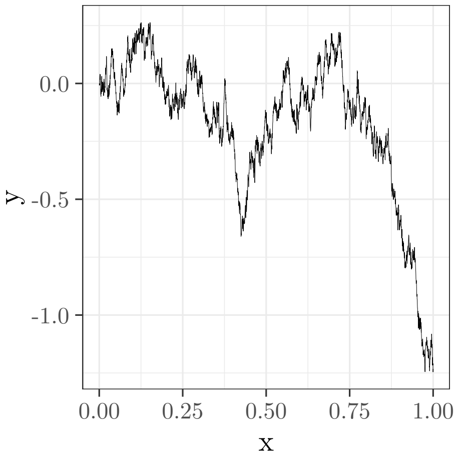
Example 6.1.4. (i) Every sequence of random variables (e.g. iid) $X_1, X_2, ...$ defines a stochastic process indexed by $\mathbb{N}$. Since there is no dependence between the values at different



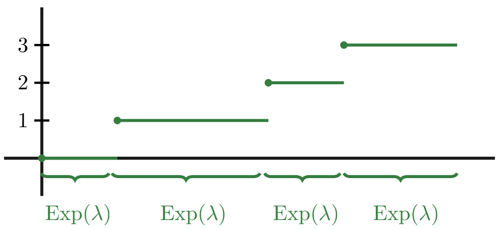
times the paths can look completely wild.

(ii) Here is a path of the so-called Brownian motion, indexed by $I = [0, \infty)$ that we will get to know later in this course.

[1] bild fehlt



(iii) A Poisson process is indexed by $I = [0, \infty)$ but mostly acts like a discrete-time process. The process jumps up by 1 at independent exponentially distributed random variables of parameter $\lambda > 0$. The parameter ist called jump-intensity as a change in λ result in more/less frequent jumps.



The Poisson process is called Poisson process as the so-called one-dimensional distributions X_t are Poi(λt)-distributed for all $t > 0$.

(iv) Markov chains are stochastic processes indexed by $\mathbb{N}$ or $\mathbb{N}_0$ (depending on taste).

We come now back to the discussion from Section 5.1 about σ-algebras and information in the context of stochastic processes.

Definition 6.1.5. Suppose $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and I an index set.

(i) An increasing family $\{\mathcal{F}_t : t \in I\}$ of sub-σ-algebras of $\mathcal{F}$, i.e $\mathcal{F}_t \subseteq \mathcal{F}_s$ for $t \leq s$, is called a **filtration**.

(ii) A tuple $\big(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in I}, \mathbb{P}\big)$ is called a **filtered probability space**.

(iii) We will frequently use the notation $\mathcal{F}_\infty := \sigma\big(\bigcup_{t\in I} \mathcal{F}_t\big) \subseteq \mathcal{F}$.

There are many filtrations for which we always use the interpretation that $\mathcal{F}_t$ contains the information someone gives us up to time t. The most important example in the study of stochastic processes is the natural filtration induced by a stochastic process:

Definition 6.1.6. Suppose X is a stochastic process indexed by I, then

$$\mathcal{F}_t := \sigma(X_s : s \leq t\}, \quad t \in I,$$

is called the **natural filtration of** X or the filtration generated by X.

Recall the discussion from Section 5.1 if time and space E are discrete. Since everything is countable we can write down $\mathcal{F}_n$ explicitly as

$$\mathcal{F}_n = \sigma\big(\{\{X_1 \in A_1, ..., X_n \in A_n\} : A_i \subseteq E\}\big) = \sigma\big(\{\{X_1 = i_1, ..., X_n = i_n\} : i_k \in E\}\big).$$

We always interpret $\mathcal{F}_n$ as the information of X up to time n as precisely those events $A \in \mathcal{F}$ belong to $\mathcal{F}_n$ that can be written in terms of paths up to time n.

Definition 6.1.7. A stochastic process $(X_t)_{t\in I}$ is **adapted to a filtration** $(\mathcal{F}_t)_{t\in I}$ if X_t is $\mathcal{F}_t$-measurable for all $t \in I$.

If we recall from Section 5.1 the interplay of σ-algebras, measurability, and information the following interpretation should be kept in mind. If X is adapted to a filtration, then the information of the filtration is enough to know X. A stochastic process is always adapted to it's natural filtration (it's own information) and to all bigger filtrations (even more information).

From now on we will discuss discrete-time stochastic processes only equipped with the power set as σ-algebra. We will return to continuous-time processes when we introduce the Brownian motion.

While in discrete time the measure theory works analogously to finite time (finite product σ-algebras) the measure theory becomes very involved in continuous time.

Definition 6.1.8. Let $\big(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\in I}, \mathbb{P}\big)$ be a filtered probability space.

(i) A random variable $\tau\colon \Omega \to I \cup \{+\infty\}$ is called an $(\mathcal{F}_n)$**-stopping time** if

$$\{\tau \le n\} \in \mathcal{F}_n \quad \forall n \in I.$$

(ii) τ is called a **finite stopping time** if $\tau < \infty$ a.s.

We usually think of a stopping time to model that something of interest happens at that time, for instance a given state is hit by the process. The idea of a stopping time is then easy to grasp: a random variable is called a stoping time if we only need information up to time n to decided if τ happened before time n or not.

Remark 6.1.9. (i) It is important to allow $\tau = \infty$ with the interpretation "the thing of interest did not happen".

(ii) Since here we only work in discrete time we could also define a stopping time by asking $\{\tau = n\} \in \mathcal{F}_n$ for all $n \in I$. This follows easily from

$$\{\tau \le n\} = \bigcup_{k\le n} \{\tau = k\}$$

because $\{\tau = k\} \in \mathcal{F}_k \subseteq \mathcal{F}_n$.

(iii) A stopping time τ is not only $\mathcal{F}$-measurable but even $\mathcal{F}_\infty$-measurable:

$$\{\tau = n\} = \{\tau \le n\} \cap \{\tau \le n-1\}^C \in \mathcal{F}_n \subseteq \mathcal{F}_\infty$$

and

$$\{\tau = \infty\} = \{\tau \ne \infty\}^C = \Big(\bigcup_{n\in\mathbb{N}} \{\tau = n\}\Big)^C \in \mathcal{F}_\infty.$$

Even though we could formulate everything with $\{\tau = n\}$ we prefer to use $\{\tau \leq n\}$ in order to slowly get acquainted to a notion which cannot be avoided in continuous time.

Here are the most important (and most simplistic) examples:

Example 6.1.10. (i) Every constant $\tau = N$ is a stopping time as $\{N \leq n\} \in \{\emptyset, \Omega\} \in \mathcal{F}_n$.

(ii) Fix some $B \in \mathcal{E}$, then the **first hitting time** $\tau_B := \inf\{n \in I : | X_n \in B\}$ is a stopping time as

$$\{\tau_B \leq n\} = \bigcup_{k \leq n} \underbrace{\{X_k \in B\}}_{\in \mathcal{F}_k \subseteq \mathcal{F}_n} \in \mathcal{F}_n.$$

Intuitively it is clear that the minimum of two stopping times ("one of the two events happened") is a stopping time again. Please prove this as a short exercise:

If $T_1, T_2, ...$ is a sequence of $(\mathcal{F}_n)$-stopping times, then

$$\inf_{k \in \mathbb{N}} T_k, \ \sup_{k \in \mathbb{N}} T_k, \ \liminf_{k \to \infty} T_k, \ \text{and} \ \limsup_{k \to \infty} T_k$$

are stopping times as well.

As mentioned above we interpret the natural filtration as information of the process, $\mathcal{F}_n$ as the information up to time n. We now generalise towards stopping times and give a mathematical definition of the information up to a stopping time.

Definition 6.1.11. Let τ be an $(\mathcal{F}_n)$-stopping time. Then

$$\mathcal{F}_\tau := \{A \in \mathcal{F}_\infty : A \cap \{\tau \leq n\} \in \mathcal{F}_n \ \forall n \in I\}$$

is called the **σ-algebra generated by** τ

It will need a bit of time to get used to $\mathcal{F}_\tau$. As always it is most instructive to have some examples in mind. Fix the first stopping time τ_B of a set B and check by hands that for some other set A the event "A was hit before B" is in $\mathcal{F}_{\tau_B}$. This should intuitively be clear as only the process until first hitting B is needed to decide if A was already hit.

Proposition 6.1.12. Suppose τ is an $(\mathcal{F}_n)$-stopping time.

(i) $\mathcal{F}_\tau$ is a σ-algebra on Ω.

(ii) τ is $\mathcal{F}_\tau$-measurable.

Proof. (i) Let us check the defining properties of a σ-algebra:

- $\Omega \cap \{\tau \leq n\} \in \mathcal{F}_n$, hence, $\Omega \in \mathcal{F}_\tau$.
- Let $A \in \mathcal{F}_\tau$, then

$$A^C \cap \{\tau \leq n\} = \{\tau \leq n\} \cap \big(A \cap \{\tau \leq n\}\big)^C \in \mathcal{F}_n$$

so that $A^C \in \mathcal{F}_\tau$.

- Let $A_1, A_2, \ldots \in \mathcal{F}_\tau$, then

$$\bigcup_{k=1}^{\infty} A_k \cap \{\tau \le n\} = \bigcup_{k=1}^{\infty} \underbrace{(A_k \cap \{\tau \le n\})}_{\in \mathcal{F}_n} \in \mathcal{F}_n$$

  so that $\cup_{k=1}^{\infty} A_k \in \mathcal{F}_\tau$.

(ii) Exercise

$\square$

Stopping times are useful as many properties of deterministic times also hold for stopping times.

Proposition 6.1.13. Suppose S, T are $(\mathcal{F}_n)$-stopping times.

(i) $S \le T$ a.s. $\Rightarrow \mathcal{F}_S \subseteq \mathcal{F}_T$

(ii) $\mathcal{F}_{S \wedge T} = \mathcal{F}_S \cap \mathcal{F}_T$

(iii) $\{S \le T\} \in \mathcal{F}_{S \wedge T}$ and $\{S = T\} \in \mathcal{F}_{S \wedge T}$

Before checking the proofs have a quick thought why those statements should intuitively be true with the interpretation of stopping times and the information given by a stopping time.

Proof. (i)

$$\begin{aligned} A \in \mathcal{F}_S \quad &\Rightarrow \quad A \cap \{S \le n\} \in \mathcal{F}_n \ \forall n \in I \\ &\Rightarrow \quad A \cap \{T \le n\} = A \cap \{S \le n\} \cap \{T \le n\} \in \mathcal{F}_n \ \forall n \in I \\ &\Rightarrow \quad A \in \mathcal{F}_T \end{aligned}$$

(ii) We have seen above that $S \wedge T$ is an $(\mathcal{F}_n)$-stopping time again. Now towards the generalised σ-algebras:

"$\supseteq$": Let $A \in \mathcal{F}_S \cap \mathcal{F}_T$ and $n \in I$, then

$$A \cap \{S \wedge T \le n\} = A \cap \big(\{S \le n\} \cup \{T \le n\}\big) = \big(A \cap \{S \le n\}\big) \cup \big(A \cap \{T \le n\}\big) \in \mathcal{F}_n$$

Hence, $A \in \mathcal{F}_{S \wedge T}$.

"$\subseteq$": This follows from the monotonicity proved in (i): $\mathcal{F}_{S \wedge T} \subseteq \mathcal{F}_s$, $\mathcal{F}_{S \wedge T} \subseteq \mathcal{F}_T$

(iii) Try yourself!

$\square$

The definitions of stopping times and adapted processes work nicely together, here is an example:

Proposition 6.1.14. If X is adapted to the filtration $(\mathcal{F}_n)$ and τ is a finite $(\mathcal{F}_n)$-stopping time, then X_τ is $\mathcal{F}_\tau$-measurable.

Proof. The finiteness of τ was only assumed to make X_τ well-defined as we did not define X_∞. Let $A \in \mathcal{E}$, we show $\{X_\tau \in A\} \in \mathcal{F}_\tau$. Let $n \in I$, then

$$\{X_\tau \in A\} \cap \{\tau \le n\} = \bigcup_{m \le n} \underbrace{\{X_m \in A\} \cap \{\tau = m\}}_{\in \mathcal{F}_m \subseteq \mathcal{F}_n} \in \mathcal{F}_n$$

$\square$

Lecture 5

6.2 Basics of martingales

All processes in this chapter are indexed by discrete ordered sets such as $I = \mathbb{N}_0$, $I = \mathbb{N}$, $I = \mathbb{Z}$, $I = -\mathbb{N}$, or $I \subseteq \mathbb{N}$ and we write n instead of t. We start the discussion of martingales with definitions and some first properties.

Definition 6.2.1. Let I a discrete ordered index-set, $\big(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\in I}, \mathbb{P}\big)$ a filtered probability space, and $X = (X_n)_{n\in I}$ an $(\mathcal{F}_n)_{n\in I}$-adapted process with $\mathbb{E}[|X_n|] < \infty$ for all $n \in I$. Then X is called an

(i) $(\mathcal{F}_n)_{n\in I}$-**martingale** if $\mathbb{E}[X_{n+1} \,|\, \mathcal{F}_n] = X_n$ a.s. for all $n \in I$,

(ii) $(\mathcal{F}_n)_{n\in I}$-**supermartingale** if $\mathbb{E}[X_{n+1} \,|\, \mathcal{F}_n] \leq X_n$ a.s. for all $n \in I$,

(iii) $(\mathcal{F}_n)_{n\in I}$-**submartingale** if $\mathbb{E}[X_{n+1} \,|\, \mathcal{F}_n] \geq X_n$ a.s. for all $n \in I$.

If $I = -\mathbb{N}$ or $I = -\mathbb{N}_0$ we will speak of a **backwards martingale**, for $I = \mathbb{N}$, $I = \mathbb{N}_0$, or $I = \{0, ..., N\}$ of a **forwards martingale** but we will always skip the supplement forwards.

The interpretation of a martingale is that of a fair game (this is where the name "martingale" comes from). Given the past value the expectation of profit in the next step is 0. Analogously, a supermartingale is seen as an unfavourable game (we will loose in expectation) and submartingales are seen as favourable games. The power of martingales is astonishing. Most of the time they appear from nowhere in a context that does not look like a typical martingale setting and their powerful convergence theorems yield strong results. We will see as an example a proof of the law of large numbers without any additional assumption.

Please check the following easy properties yourself!

(i) The (sub)(super)martingale property also holds over several time steps:

$$\mathbb{E}[X_m \,|\, \mathcal{F}_n] \begin{cases} = X_n & : X \text{ martingale} \\ \leq X_n & : X \text{ supermartingale} \\ \geq X_n & : X \text{ supermartingale} \end{cases},$$

for all $m \geq n$.

(ii) Expectations (increase)(decrease)stay constant depending on X being a (sub)(super)martingale:

$$\mathbb{E}[X_m] \begin{cases} = \mathbb{E}[X_n] & : X \text{ martingale} \\ \leq \mathbb{E}[X_n] & : X \text{ supermartingale} \\ \geq \mathbb{E}[X_n] & : X \text{ submartingale} \end{cases},$$

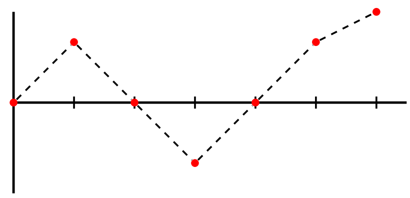
for all $m \geq n$.

(iii) X is a supermartingale $\Leftrightarrow$ $-X$ is a submartingale.

The final property allows us to prove theorems in most cases for either sub- or supermartingales and then transfer the theorems to the other.

Even though the (super)(sub)martingale properties seems artificial many stochastic processes share this property:

Example 6.2.2. The most prominent stochastic process in discrete time is the random walk which is a Markov chain and also a martingale.



A trajectory of the simple random walk

Let $Y_1, Y_2, Y_3, ..$ be iid real-valued integrable random variable. The **random walk with jump sizes** Y_k is defined by $X_0 := x \in \mathbb{R}$ and

$$X_n = x + \sum_{k=1}^{n} Y_k, \quad n \in \mathbb{N}.$$

If $\mathbb{P}(Y_1 = 1) = p$ and $\mathbb{P}(Y_1 = -1) = 1-p$ the random walk is called **simple random walk**, for $p = \frac{1}{2}$ symmetric simple random walk.

If we define $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and $\mathcal{F}_n = \sigma(Y_1, ..., Y_n)$ then the random walk is an

- $(\mathcal{F}_n)$-martingale if $\mathbb{E}[Y_1] = 0$,
- $(\mathcal{F}_n)$-supermartingale if $\mathbb{E}[Y_1] \leq 0$,
- $(\mathcal{F}_n)$-submartingale if $\mathbb{E}[Y_1] \geq 0$.

To see why, let us check the definition. Adaptivity and integrability (Δ-inequality) is clear, the martingale property is deduced using properties of the conditional expectation:

$$\begin{aligned}
\mathbb{E}[X_{n+1} \,|\, \mathcal{F}_n] &= \mathbb{E}\Big[\sum_{k=1}^{n+1} Y_k \,\Big|\, \mathcal{F}_n\Big] \\
&= \sum_{k=1}^{n+1} \mathbb{E}[Y_k \,|\, \mathcal{F}_n] \\
&= \sum_{k=1}^{n} Y_k \; + \; \mathbb{E}[Y_{n+1} \,|\, \mathcal{F}_n] \\
&= X_n + \mathbb{E}[Y_1],
\end{aligned}$$

using in the third equality the measurability and independence assumption on the jump sizes.

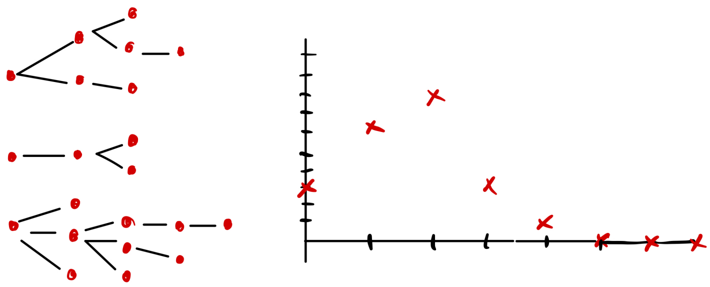
Example 6.2.3. Another famous class of discrete time stochastic processes are so-called branching processes ("Verzweigungsprozesse").

Let ξ_i^n, $i, n \in \mathbb{N}$, be integrable, non-negative discrete iid random variables with $\mathbb{P}(\xi_1^n = k) = p_k$, $k \in \mathbb{N}_0$. The classical **branching process** (or **Galton-Watson process**) with offspring distribution ξ is defined by

$$X_0 := m, \quad X_{n+1} := \sum_{k=1}^{X_n} \xi_k^n, \quad n \in \mathbb{N}.$$

We will call X_n the number of individuals of a population at time n.

The interpretation goes as follows. At time zero there are m individuals, plants for examples. At every time-unit, once a year for plants, every existing individual gets offspring. The number is independent from all offspring of other individuals in the same generation but also independent



Branching process with $m = 3$ and extinction time 5

of the past. The genealogical picture is typically represented by a graph, X_n counts the number of individuals at time n. We say the branching process gets extinct if $X_n = 0$, after the first extinction time the process stays extinct forever. We always assume $p_0, p_1 \neq 1$ as otherwise the branching process either dies out immediately ($p_0 = 1$ forces all initial individuals to have zero offspring) or stays constant ($p_1 = 1$ forces all individuals to have exactly one offspring so that $X_n = m$ for all $n \in \mathbb{N}$). Typical questions concern the probability of extinction or the rate of growth (exponential, subexponential). A critical feature is the mean number of offspring $\mu = \mathbb{E}[\xi_1^n] = \sum_{k=0}^{\infty} k \cdot p_k$ which is the main driver for the longtime behavior.

If we define $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and $\mathcal{F}_n = \sigma(\xi_i^k : i \in \mathbb{N}, k \leq n)$ then the branching process X is an

- $(\mathcal{F}_n)$-martingale if $\mu = 1$, called the **critical case**,
- $(\mathcal{F}_n)$-supermartingale if $\mu < 1$, called the **subcritical case**,
- $(\mathcal{F}_n)$-submartingale if $\mu > 1$, called the **supercritical case**.

Let us check the definition which is similar to the computation for the random walk except we need to deal with the random delimiter in the sums for which we need the Wald-identity:

Suppose $N, Y_1, Y_2, \ldots$ are independent, $\mathbb{E}[N] < \infty$, and $Y_1, Y_2, \ldots$ are identically distributed, then

$$\mathbb{E}\Big[\sum_{k=1}^{N} Y_k\Big] = \mathbb{E}[N] \cdot \mathbb{E}[Y_1]. \tag{6.1}$$

Integrability is deduced immediately as the Wald-identity gives $\mathbb{E}[X_n] = \mathbb{E}[X_{n-1}] \cdot \mathbb{E}[\xi_1^1]$ so that $\mathbb{E}[X_n] = m\mu^n < \infty$ by a simple induction. Adaptivity follows direction from the definition of $\mathcal{F}_n$. The martingale property is deduced using properties of the conditional expectation:

$$\begin{aligned}
\mathbb{E}[X_{n+1} \,|\, \mathcal{F}_n] &= \mathbb{E}\Big[\sum_{k=1}^{\infty} \mathbf{1}_{k \leq X_n} \cdot \xi_k^{n+1} \,\Big|\, \mathcal{F}_n\Big] \\
&= \sum_{i=1}^{\infty} \mathbf{1}_{j \leq X_n} \mathbb{E}\big[\xi_j^{n+1} \,|\, \mathcal{F}_n\big] \\
&= \sum_{i=1}^{\infty} \mathbf{1}_{j \leq X_n} \mathbb{E}\big[\xi_j^{n+1}\big] = \mu \cdot X_n,
\end{aligned}$$

using monotone convergence, measurability and the independence of conditional expectation. Interestingly, there is another martingale appearing in the branching process. If we define $M_n := \frac{1}{\mu^n} \cdot X_n$, $n \in \mathbb{N}$, then M is a martingale in all three regimes! Adaptivity and measurability is clear, the martingale property follows with the same calculation as above with an additional cancellation:

$$\mathbb{E}[M_{n+1} \,|\, \mathcal{F}_n] = \frac{1}{\mu^{n+1}} \cdot \mathbb{E}[X_{n+1} \,|\, \mathcal{F}_n] = \frac{1}{\mu^{n+1}} \cdot \mu \cdot X_n = M_n.$$

We already get a first impression of what is going by checking the expectations, which are constant for the martingale M. Then the expectations of X_n remain constant for $\mu = 1$ grow exponentially as $\mu^n = e^{\log(\mu)n}$ for $\mu > 1$ and decay exponentially for $\mu < 1$.

Example 6.2.4. If Z is an integrable random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ and $(\mathcal{F}_n)$ is a filtration, then

$$X_n := \mathbb{E}[Z \,|\, \mathcal{F}_n], \quad n \in \mathbb{N},$$

is a martingale. The argument is simple but very important:

$$\mathbb{E}[X_{n+1} \,|\, \mathcal{F}_n] = \mathbb{E}\big[\mathbb{E}[Z \,|\, \mathcal{F}_{n+1}] \,\big|\, \mathcal{F}_n\big] \overset{\text{tower prop.}}{=} \mathbb{E}[Z \,|\, \mathcal{F}_n] = X_n \quad \text{a.s.}$$

Every martingale $(X_n)_{n \in \mathbb{N}}$ that can be written as $\mathbb{E}[Z \,|\, \mathcal{F}_n]$ for some integrable random variable Z is called a **closed martingale** or **Doob martingale**. It is best to think of a Doob martingale as a martingale on finite time horizon $\{0, ..., N\}$ as in both cases one random variable (Z for a Doob martinglae, X_N for a finite-time martingale) determines the entire martingale. In the end of Section 6.3.3 it will be shown that closed martingales (or, equivalently, uniformly integrable martingales) indeed share important properties of finite-time martingales.

Proposition 6.2.5. Let $\varphi \colon \mathbb{R} \to \mathbb{R}$ be a convex function, and $(X_n)_{n \in \mathbb{N}}$ a stochastic process with $\mathbb{E}[|\varphi(X_n)|] < \infty$.

(i) If $(X_n)_{n \in \mathbb{N}}$ is an $(\mathcal{F}_n)$-martingale, then $\big(\varphi(X_n)\big)_{n \in \mathbb{N}}$ is an $(\mathcal{F}_n)$-submartingale.

(ii) If $(X_n)_{n \in \mathbb{N}}$ is an $(\mathcal{F}_n)$-submartingale and φ is increasing, then $\big(\varphi(X_n)\big)_{n \in \mathbb{N}}$ is an $(\mathcal{F}_n)$-submartingale.

Beweis. We use Jensen's inequality for conditional expectation:

(i) $\mathbb{E}\big[\varphi(X_{n+1}) \,|\, \mathcal{F}_n\big] \geq \varphi\big(\mathbb{E}[X_{n+1} \,|\, \mathcal{F}_n]\big) = \varphi(X_n)$ a.s. The equality uses the martingale property.

(ii) $\mathbb{E}\big[\varphi(X_{n+1}) \,|\, \mathcal{F}_n\big] \geq \varphi\big(\mathbb{E}[X_{n+1} \,|\, \mathcal{F}_n]\big) \geq \varphi(X_n)$ a.s. The second inequality uses that φ is increasing and the submartingale property.

□

As usual, the simplest examples are the most useful ones. One can regularly see the use of $\varphi(x) = |x|$, $\varphi(x) = (x - a)^+$, and powers $\varphi(x) = |x|^p$ for $p \geq 1$.

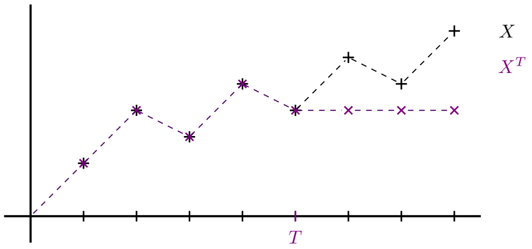
Most importantly, $(X_n^2)_{n \in \mathbb{N}}$ is a submartingale if $(X_n)_{n \in \mathbb{N}}$ is a martingale with $\mathbb{E}[X_n^2] < \infty$ for all $n \in \mathbb{N}$.

Here is something simple to check yourself.

Suppose $(X_n)_{n \in \mathbb{N}}$ and $(Y_n)_{n \in \mathbb{N}}$ are $(\mathcal{F}_n)$-martingales. Then the sum $(X_n + Y_n)_{n \in \mathbb{N}}$ is an $(\mathcal{F}_n)$-martingale and the maximum $(X_n \vee Y_n)_{n \in \mathbb{N}}$ is an $(\mathcal{F}_n)$-submartgingale.

We come to the first theorem on martingales where we relate martingales and stopping times. If T is a stopping time then we define the **stopped process** X^T as

$$X_n^T(\omega) := X_{n \wedge T(\omega)}(\omega), \quad n \in \mathbb{N}.$$



Path of a stopped process X^T

The path of the stopped process is the same as the original process up to time T and stays constant at the value X_T after time T. It might be instructive to realise that stopping at deterministic times $T = N$ shows how to relate infinite time-horizon martingales to finite time-horizon martingales on $\{0, ..., N\}$.

The optional stopping theorem states that a martingale stopped at a stopping time (i.e. without future information) remains a martingale. We can derive as an application the so-called optional sampling theorem. The theorem formalises the idea of a fair game that without future information it is impossible to reach a gain in expectation by stopping. While we will see after the theorem an example that this statement is a bit too optimistic it does hold for bounded stopping times:

Theorem 6.2.6. (Optional Stopping/Optional Sampling Theorem) Suppose $(X_n)_{n\in\mathbb{N}}$ is an $(\mathcal{F}_n)$-martingale and T is an $(\mathcal{F}_n)$-stopping time.

(i) The **stopped process** $(X_n^T)_{n\in\mathbb{N}}$ is an $(\mathcal{F}_n)$-martingale.

(ii) If T is a **bounded** $(\mathcal{F}_n)$-stopping time, i.e. $\mathbb{P}(T \leq K) = 1$ for some $K \in \mathbb{N}$, then X_T is an integrable random variable with $\mathbb{E}[X_T] = \mathbb{E}[X_1]$.

The optional sampling theorem in particular applies to martingales on finite time-horizon $\{0, ..., N\}$ when the optional sampling theorem is applied to the stopped martingale X^N.

Proof. We mainly prove the optional stopping theorem, optional sampling is a direct consequence.

Optional Stopping Theorem: We check the three defining properties (adapted, integrable, martingale property):

- First note that $n \wedge T$ is a stopping time (minimum of two stopping times), hence, $X_{n\wedge T}$ is $\mathcal{F}_{n\wedge T}$-measurable by Proposition 6.1.14. Since $\mathcal{F}_{n\wedge T} \subseteq \mathcal{F}_n$ by Proposition 6.1.13, X^T is $(\mathcal{F}_n)$-adapted.
- The integrability of X^T follows by splitting on the possible values of T:

$$\begin{aligned}
\mathbb{E}\big[|X_n^T|\big] &= \mathbb{E}\Big[\,|X_{n\wedge T}|\sum_{k=1}^{\infty}\mathbf{1}_{T=k}\Big]\\
&\leq \mathbb{E}\Big[\sum_{k=1}^{n}|X_k|\,\mathbf{1}_{T=k}\Big] + \mathbb{E}\Big[\sum_{k=n+1}^{\infty}|X_n|\,\mathbf{1}_{T=k}\Big]\\
&\leq \mathbb{E}\Big[\sum_{k=1}^{n}|X_k|\Big] + \mathbb{E}\big[\,|X_n|\,\mathbf{1}_{T\geq n+1}\big]\\
&\leq \sum_{k=1}^{n}\mathbb{E}[|X_k|] + \mathbb{E}[|X_n|] < \infty.
\end{aligned}$$

- To show the martingale property we use a trick that will return frequently. To show the martingale property it is enough to show that the differences are so-called martingale differences:

Trick: $(X_n)_{n\in\mathbb{N}}$ is an $(\mathcal{F}_n)$-martingale iff $\mathbb{E}[X_{n+1} - X_n \,|\, \mathcal{F}_n] = 0$ a.s. for all $n \in \mathbb{N}$.

To justify the martingale difference trick one only needs to use the linearity of conditional expectation and that X is adapted.

Let's check that the differences of X^T are martingale differences:

$$\begin{aligned}\mathbb{E}\big[X^T_{n+1} - X^T_n \,\big|\, \mathcal{F}_n\big] &= \mathbb{E}\big[(X^T_{n+1} - X^T_n)(\mathbf{1}_{T\le n} + \mathbf{1}_{T>n}) \,\big|\, \mathcal{F}_n\big]\\ &= \mathbb{E}\big[(X_{n+1} - X_n)\mathbf{1}_{T>n} \,\big|\, \mathcal{F}_n\big]\\ &= \mathbf{1}_{T>n}\mathbb{E}[X_{n+1} - X_n \,|\, \mathcal{F}_n] = 0\end{aligned}$$

where we used that

- $X^T_{n+1} = X^T_n$ on the event $\{T \le n\}$,
- $X^T_{n+1} = X_{n+1}, X^T_n = X_n$ on $\{T > n\}$,
- $\mathbf{1}_{T>n}$ is measurable as $\{T > n\} = \{T \le n\}^C \in \mathcal{F}_n$ by the stopping time property.

Optional Sampling Theorem: Using that the stopped martingale is again a martingale and that martingales have constant expectations we obtain the theorem:

$$\mathbb{E}[X_T] \overset{T\le K}{=} \mathbb{E}[X_{K\wedge T}] = \mathbb{E}[X^T_K] = \mathbb{E}[X^T_1] = \mathbb{E}[X_{T\wedge 1}] = \mathbb{E}[X_1].$$

□

The boundedness of T can be weakend but not generally be removed! Always keep in mind the random walk example below as a counter example!

Remark 6.2.7. (i) The boundedness assumption on T can be removed if X is bounded! If T is a finite stopping time, then

$$\mathbb{E}[X_T] = \mathbb{E}\big[\lim_{n\to\infty} X_{T\wedge n}\big] \overset{\text{DCT}}{=} \lim_{n\to\infty} \mathbb{E}[X_{T\wedge n}] \overset{\text{mart.}}{=} \lim_{n\to\infty} \mathbb{E}[X_1] = \mathbb{E}[X_1].$$

(ii) The boundedness assumption on T can be weakend to $\mathbb{E}[T] < \infty$ if the martingale differences are almost surely bounded, i.e. $|X_n - X_{n+1}| \le K$ almost surely for all $n \in \mathbb{N}$. In that case it holds that

$$|X_T - X_{T\wedge n}| \overset{\text{telescope}}{=} \Big|\sum_{k=T\wedge n}^{T-1} (X_{k+1} - X_k)\Big| \le K \cdot T$$

so that, again by dominated convergence, we obtain

$$\lim_{n\to\infty} \big(\mathbb{E}[X_T] - \mathbb{E}[X_{\wedge n}]\big) = \mathbb{E}\big[\lim_{n\to\infty} (X_T - X_{T\wedge n})\big] = \mathbb{E}[0] = 0.$$

Since $\mathbb{E}[X_{T\wedge n}] = \mathbb{E}[X_1]$ by optional stopping, optional sampling follows.

(iii) As a counter example to the optional sampling theorem when the boundedness of T is violated let us consider the symmetric simple random walk. Let $T = \inf\{n \in \mathbb{N} \colon X_n = 1\}$ and $X_0 = 0$. Then $T < \infty$ almost surely but

$$\mathbb{E}[X_T] = 1 \ne 0 = \mathbb{E}[X_0] = \mathbb{E}[X_n]$$

for $n \in \mathbb{N}$. Since the martingale differences are bounded (they are $+1$ or -1), (ii) shows that $\mathbb{E}[T] = \infty$.

The random walk example is the first time we can feel the power of martingales. It was quite easy to prove $\mathbb{E}[T] = \infty$ via the option sampling theorem. But how can you prove this by hands? Just try to compute $\mathbb{P}(T = k)$ for some k and you will see that you run quickly into combinatorics. Of course you could try to do this by hands by

Definition 6.2.8. A stochastic process (H_n) is called **previsible** (or **predictable**) if H_n is $\mathcal{F}_{n-1}$-measurable. („H_n only depends on information up to $n-1$").

[2] Previsibility is actually quite a natural concept. If for instance H_n could be the amount you want to invest at day n based upon the price of the day $n-1$ before. Such concepts are clearly important in mathematical finance but also in probability theory.

Definition 6.2.9. Suppose $(X_n)_{n\in\mathbb{N}_0}$ is an $(\mathcal{F}_n)$-adapted stochastic process and (H_n) is $(\mathcal{F}_n)$-predictable, then we define

$$\begin{aligned}(H\cdot X)_0 &:= 0,\\ (H\cdot X)_n &:= \sum_{k=1}^n H_k\cdot(X_k-X_{k-1}).\end{aligned}$$

Since $H\cdot X$ is the discrete analog of stochastic integral $\int_0^t H_s\,\mathrm{d}X_s$ we call $H\cdot X$ the **discrete stochastic integral of H agains X**.

The interpretation of $H\cdot X$ in mathematical finance is the wealth obtained by trading the asset X using the trading strategy H. Here is bad news: If your favorit asset is a martingale and you only have a bounded amount of money to invest, there is no way of making money by clever stopping (in your life time) your investment. Unfortunately, $H\cdot X$ will always be a martingale so that the expectation will be constant at all bounded stopping times by the optional sampling theorem. Lecture 6

Theorem 6.2.10. (Sorry, but you really cannot beat the system.) Suppose (H_n) is predictable and bounded (i.e. $|H_n|\leq K$ a.s. for all $n\in\mathbb{N}$).

(i) If X is a martingale, then $H\cdot X$ is a martingale.

(ii) If X is a supermartingale and $H\geq 0$, then $H\cdot X$ is a supermartingale.

(iii) If X is a submartingale and $H\geq 0$, then $H\cdot X$ is a submartingale.

Beweis. (i) We need to check the three defining properties of a martingale.

- $(H\cdot X)_n$ is $\mathcal{F}_n$-adapted by definition
- Since H is bounded by assumption and X is integrable as a martingale we obtain

$$\mathbb{E}\big[|(H\cdot X)_n|\big]\overset{\Delta}{\leq}\sum_{k=1}^n\mathbb{E}\big[|H_k\cdot(X_k-X_{k-1})|\big]\leq K\sum_{k=1}^n(\mathbb{E}[|X_k|]+\mathbb{E}[|X_{k-1}|])<\infty.$$

- The martingale property follows from a direct computation using the assumed measurability properties to simplify the conditional expectations:

$$\begin{aligned}\mathbb{E}\big[(H\cdot X)_{n+1}\,\big|\,\mathcal{F}_n\big]&=\mathbb{E}\Big[\sum_{k=1}^{n+1}H_k(X_k-X_{k-1})\,\Big|\,\mathcal{F}_n\Big]\\&=\sum_{k=1}^n H_k(X_k-X_{k-1})+\mathbb{E}\big[H_{n+1}(X_{n+1}-X_n)\,\big|\,\mathcal{F}_n\big]\\&=(H\cdot X)_n+0,\quad\text{a.s.}\end{aligned}$$

The last equation holds, because H is predictable and $\mathbb{E}[X_{n+1}-X_n\,|\,\mathcal{F}_n]=0$ almost surely by the martingale property.

[2]Leif: Indexmengen aufraeumen

(ii) Since $\mathbb{E}[X_{n+1} - X_n \,|\, \mathcal{F}_n] \leq 0$ a.s. for a supermartingale and $H \geq 0$ we can replace the last equation from our calculation above with $\leq$.

(iii) Just like (ii).

$\square$

In this probability theory lecture we will not care about mathematical finance, but nonetheless, the discrete stochastic integral will be a massively useful tool for us! We can for instance give an alterntive proof for optional stopping by using the previous theorem with $H_n := \mathbf{1}_{T \geq n} = 1 - \mathbf{1}_{T<n}$. H is previsible and $X_{n\wedge T}$ can be rewritten as

$$X_{n\wedge T} = (H \cdot X)_n + X_0$$

because

$$(H \cdot X)_n = \sum_{k=1}^{n} \mathbf{1}_{T\geq k}(X_k - X_{k-1}) = X_{n\wedge T} - X_0$$

Hence, $(X_{n\wedge T})_{n\in\mathbb{N}}$ is a martingale. Similar tricks of playing with the discrete stochastic integrals will appear later. To prove the almost sure martingale convergence theorem and the optional sampling theorem for uniformly integrable martingales.

6.3 Martingale convergence theorems

The most striking feature of martingales is the rich convergence theory. Under very mild assumptions we will prove convergence

$$X_n \to X_\infty, \quad n \to \infty,$$

to some limiting random variable X_∞, where the mode of convergence depends on the assumptions on X. As an example, without any further knowledge a non-negative martingale converges almost surely to a limit X_∞. In the following sections we will discuss almost sure, L^p, and L^1 limit theorems.

For the convergence theorems we will need a first element and an open end in the forwards direction for the limit to be interesting. The theorems work equally with $\mathbb{N}$ or $\mathbb{N}_0$, to have a consistent notation we will work with $\mathbb{N}_0$.

6.3.1 Almost sure martingale convergence theorem

We start with the most prominent martingale convergence theorem, the almost sure convergence theorem:

Theorem 6.3.1. (Almost sure martingale convergence theorem)
If $(X_n)_{n\in\mathbb{N}_0}$ is a (sub)martingale with $\sup_{n\in\mathbb{N}_0} \mathbb{E}[X_n^+] < \infty$, then almost surely

$$X_\infty := \lim_{n\to\infty} X_n$$

exists, is $\mathcal{F}_\infty$ measurable and almost surely finite with $\mathbb{E}[|X_\infty|] < \infty$.

The most useful application is towards non-negative martingales as they always satisfy the assumption of the almost sure martingale convergence theorem:

Corollary 6.3.2. If $(X_n)_{n\in\mathbb{N}_0}$ is a non-negative martingale, i.e. $X_n \geq 0$ a.s. for all $n \in \mathbb{N}_0$, then almost surely X_n converges to a limit X_∞ with $0 \leq \mathbb{E}[X_\infty] \leq \mathbb{E}[X_0]$.

Beweis. The martingale property yields $0 \leq \mathbb{E}[X_n^+] = \mathbb{E}[X_n] = \mathbb{E}[X_0] < \infty$ for all $n \in \mathbb{N}_0$ so that the almost sure martingale convergence theorem applies. Using Fatou's lemma then gives

$$\mathbb{E}[X_\infty] = \mathbb{E}\big[\lim_{n\to\infty} X_n\big] \leq \liminf_{n\to\infty} \mathbb{E}[X_n] = \mathbb{E}[X_0].$$

□

Here is the idea of the proof.

In order to have convergence of a real-valued sequence (a_n) to a real number or infinity it is enough to prove that all intervals $[a,b]$ with rational endpoints are crossed only finitely many times.

The case of convergence towards infinity is clear, for a finite limit the claim is best seen by its contraposition. If a sequence does not converge, there are $a, b \in \mathbb{Q}$ such that $\liminf_n a_n < a < b < \limsup_n a_n$. Hence, the sequence (a_n) takes infinitely many values larger than b and smaller than a. But then the interval is crossed infinitely often. The martingale convergence theorem is proved by showing that the expected number of crosses through arbitrary intervals is finite, hence, the crossing number is almost surely finite. In fact, to prove that there are finitely many crossings it is enough to show that there are only finitely many upcrossings from below a to above b. More formally, we define the crossing times by

$$\begin{aligned} S_1 &:= 0 \\ T_1 &:= \min\{n \in \mathbb{N}_0 : X_n \geq b\} \\ S_{k+1} &:= \min\{n > T_k : X_n \leq a\} \\ T_{k+1} &:= \min\{n > S_k : X_n \geq b\} \end{aligned}$$

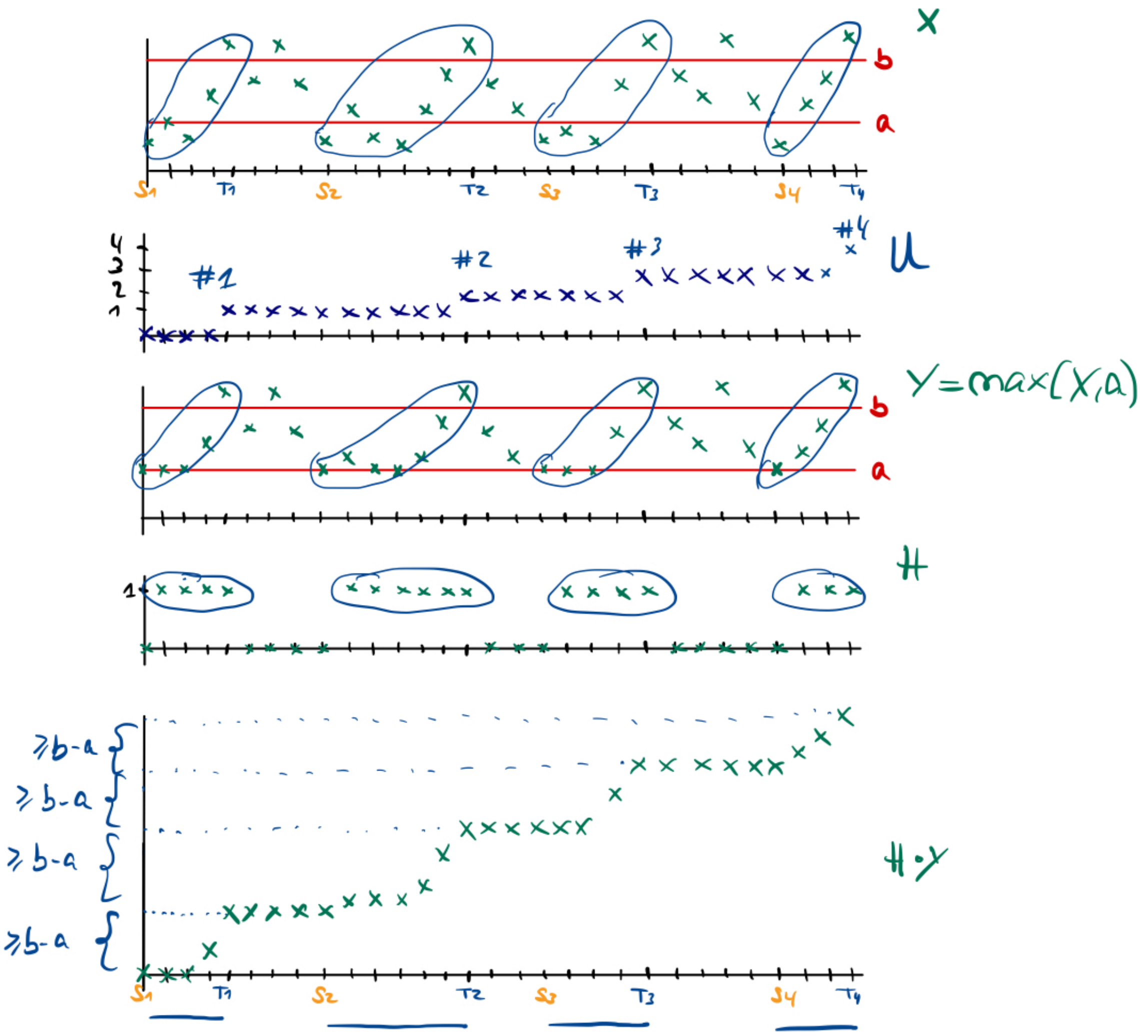
and set

$$U_n[a,b] := \sum_{k=1}^{\infty} \mathbf{1}_{T_k \leq n},$$

which is the number of finished upcrosings up to time n. The main ingredient towards the convergence theorem is the following estimate for the number of upcrossings:

Lemma 6.3.3. (Doob's upcrossing inequality)
If $(X_n)_{n\in\mathbb{N}_0}$ is a (sub)martingale and $a < b$, then

$$\mathbb{E}\big[U_n[a,b]\big] \leq \frac{\mathbb{E}\big[(X_n - a)^+ - (X_0 - a)^+\big]}{b-a}.$$



Realisation of X with four upcrossings and the counting using H, Y, and $H \cdot Y$

Proof. In order to count the number of upcrossings we introduce

$$Y_n := \max\{X_n, a\} = (X_n - a)^+ + a,$$

$$H_n := \sum_{k=1}^{\infty} \mathbf{1}_{\{S_k < n \leq T_k\}} = \begin{cases} 1 & : n \in \{S_k + 1, ..., T_k\} \text{ for some } k \\ 0 & : n \in \{T_k + 1, ..., S_k\} \text{ for some } k \end{cases}.$$

Additionally, we are interested in the discrete stochastic integral $H \cdot Y$. The idea of the proof is best understood through the illustration of the processes. We have chosen H such that H only takes the values 1 and 0 so that $H \cdot Y$ either stays constant (in intervals with $H = 0$) or sums up the increments of Y. Every upcrossing of X yields an increase of $H \cdot Y$ of at least $b - a$ so that $H \cdot Y$ counts (up to a factor $b - a$) the number of upcrossings.

Let's have a more formal look a the definitions. The telescopic sum property is best seen at the endpoints of upcrossings by removing all 0 summmands:

$$\begin{aligned}(H \cdot Y)_{T_n} &= \sum_{k=1}^{T_n} H_k (Y_k - Y_{k-1}) \\ &= \sum_{k=1}^{n} \sum_{j=S_k+1}^{T_k} 1 \cdot (Y_j - Y_{j-1})\end{aligned}$$

$$\begin{aligned} &\overset{\text{telescope}}{=} \sum_{k=1}^{n}(Y_{T_k} - Y_{S_k}) \\ &\geq n \cdot (b-a) \end{aligned}$$

because $Y_{T_k} \geq b$ and $Y_{S_k} \leq a$. Now let us have a look at the values of $H \cdot Y$ in the two different ranges of the definition of H:

$$(H \cdot Y)_j \overset{\text{adding } 0s}{=} (H \cdot Y)_{T_n} \geq n(b-a), \quad \forall j \in \{T_n, \cdots, S_{n+1}\}$$

and, using this equality again,

$$(H \cdot Y)_j \overset{\text{telescope}}{\geq} (H \cdot Y)_{S_{n+1}} \overset{\text{adding } 0s}{=} (H \cdot Y)_{T_n} \geq n(b-a), \quad \forall j \in \{S_{n+1}, \cdots, T_{n+1}-1\}.$$

Noting that $j \geq T_n$ is the same as $n \geq U_j[a,b]$ (recall the definitions of T_n and $U_j[a,b]$) we obtain

$$(H \cdot Y)_n \geq (b-a)U_n[a,b], \quad n \in \mathbb{N}_0.$$

We are now close to finishing the proof using a submartingale argument. First note that Y is a submartingale by Proposition 6.2.5 as $x \mapsto (x-a)^+ + a$ is a convex function. Next, H_n is previsible as $\mathbf{1}_{\{S_k<n\leq T_k\}} = \mathbf{1}_{\{S_k<n\}} \cdot \mathbf{1}_{\{T_k<n\}^C}$ and sums and limits of measurable functions are measurable (compare 2.3.6 and **??**). Hence, also $1-H_n$ is previsible so that the discrete integral $(1-H) \cdot X$ is a submartingale by Theorem 6.2.10 because $1-H_n$ is bounded by 1 and non-negative. But then, using that submartingales have increasing expectation, we obtain the desired bound:

$$\begin{aligned} \mathbb{E}[Y_n - Y_0] &\overset{\text{telescope}}{=} \mathbb{E}[(1 \cdot Y)_n] \\ &= \mathbb{E}[(H \cdot Y)_n] + \mathbb{E}[((1-H) \cdot Y)_n] \\ &\overset{\text{see above}}{\geq} (b-a)\mathbb{E}\big[U_n[a,b]\big] + \mathbb{E}[((1-H) \cdot Y)_n] \\ &\overset{(1-H)\cdot Y \text{ submart.}}{\geq} (b-a)\mathbb{E}\big[U_n[a,b]\big] + \mathbb{E}\big[((1-H) \cdot Y)_0\big] \\ &= (b-a)\mathbb{E}\big[U_n[a,b]\big] + 0. \end{aligned}$$

Dividing by $b-a$ and plugging-in the definition of Y yields the upcrossing inequality. □

With the upcrossing lemma we can quickly finish the proof of the martingale convergence theorem. The proof looks much worse than it is!

Proof of the almost sure martingale convergence theorem. Let $a<b$. Since $(X_n - a)^+ \leq |a| + X_n^+$ the upcrossing inequality gives

$$\mathbb{E}\big[U_n[a,b]\big] \leq \frac{a + \mathbb{E}[X_n^+]}{b-a}$$

and the right hand side is bounded by some $C<\infty$ due to the assumption on the submartingale. Now define

$$U[a,b] := \lim_{n\to\infty} U_n[a,b] \in [0,\infty],$$

which is the total number of upcrossings through $[a,b]$. The limit exists as monotone sequences converge (with possible limit $+\infty$). Since the limit is monotone in n we can use the monotone convergence theorem to obtain

$$\mathbb{E}\big[U_n[a,b]\big] = \lim_{n\to\infty} \mathbb{E}\big[U_n[a,b]\big] \leq C.$$

Since non-negative random variables with finite expectation are finite almost surely, we proved that

$$\mathbb{P}(U[a,b]<\infty) = 1, \quad \text{for all } a<b.$$

Hence, we proved that almost surely the submartingale only crosses $[a,b]$ finitely often. If we define

$$C^{a,b} := \{\omega : U[a,b](\omega) = \infty\} \quad \text{and} \quad C := \bigcup_{a,b\in\mathbb{Q},\, a<b} C^{a,b},$$

then the above shows that $\mathbb{P}(C^{a,b}) = 0$ and, hence, $\mathbb{P}(C) = 0$. Since C^C is the event that X does not cross any interval infinitely often (equivalently, X_n converges) and $\mathbb{P}(C^C) = 1$ we proved that $\lim_{n\to\infty} X_n$ exists almost surely (with a possibly infinite limit).

Now define $X_\infty := \lim_{n\to\infty} X_n$. Since all X_n are $\mathcal{F}_\infty$-measurable, X_∞ is $\mathcal{F}_\infty$-measurable as a limit. If we can show that $\mathbb{E}[|X_\infty|] < \infty$, then X_∞ is finite almost surely. To show this we use Fatou's lemma twice. First,

$$\mathbb{E}[X_\infty^-] \leq \liminf_{n\to\infty} \mathbb{E}[X_n^-] = \liminf_{n\to\infty} \big(\mathbb{E}[X_n^+] - \mathbb{E}[X_n]\big) \leq \sup_{n\in\mathbb{N}_0} \mathbb{E}[X_n^+] - \mathbb{E}[X_0] < \infty$$

and, secondly,

$$\mathbb{E}[X_\infty^+] \leq \liminf_{n\to\infty} \mathbb{E}[X_n^+] < \infty.$$

Hence, $\mathbb{E}[|X_\infty|] < \infty$ and in particular $|X_\infty| < \infty$ a.s. □

A typical example for the martingale convergence theorem is a better understanding of the regimes in the branching process from Example 6.2.3.

Example 6.3.4. Recall the martingale M obtained from the branching process which is actually a non-negative martingale. Hence, by Corollary 6.3.2 there is a finite almost sure limit M_∞ of M_n which directly translates into knowledge on X. Here is an important fact that we use: If a sequence with values on $\mathbb{N}_0$ converges, then the sequence must ultimately be constant.

(i) $\mu < 1$ (subcritical case):
$\frac{1}{\mu^n} \to \infty$, so that $X_n = \mu^n M_n \to 0$, $n \to \infty$. But then the population modelled by X suffers extinction in finite time almost surely, that is, almost surely gets absorbed at 0 after some (unknown) random time $N(\omega)$.

(ii) $\mu = 1$ (critical case):
$(X_n)_{n\in\mathbb{N}}$ itself is a non-negative martingale so that $X_n \to X_\infty$ a.s. But how could the branching process stop moving ultimately? Right, only by almost sure extinct of the population in finite time (check the definition).

(iii) $\mu > 1$ (supercritical case):
Again the martingale convergence theorem implies

$$\frac{1}{\mu^n} \cdot X_n \to M_\infty \in [0,\infty).$$

Unfortunately, so far do not know anything about M_∞ except being finite. On the event $E := \{\omega : M_\infty(\omega) > 0\}$ we observe exponential growth of the population, namely,

$$X_n(\omega) \sim M_\infty(\omega)\mu^n = M_\infty(\omega) e^{\log(\mu)\cdot n}.$$

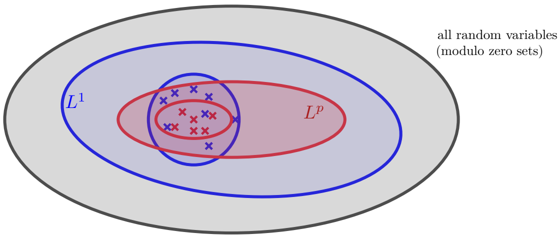
So far we cannot say if E has positive probability. Under a square-integrability condition we will solve this question using the L^2-martingale convergence theorem below.

What we see is an effect everyone learnt during the Covid pandemic. Sick people infecting on average more than 1 person can lead to exponential growth, infecting on average less than 1 person leads to extinction of the disease. Of course, this model is very simplistic due to the iid assumption on the offspring.

Lecture 7

6.3.2 L^p-martingale convergence theorem for $p > 1$

After the almost sure convergence we now look for stronger conditions on martingales that additionally ensure convergence of X_n towards X_∞ in L^p, i.e. $\lim_{n\to\infty} \mathbb{E}[|X_n - X_\infty|^p] = 0$. If you forgot about the L^p-spaces of random variables (modulo zero sets) with the norms $||X||_p = \mathbb{E}[|X|^p]^{1/p}$ please check Theorem **??** and the discussion around. In this and the following section bounded sets in L^p and L^1 will play a crucial role. Recall from analysis that bounded subsets of a normed space are subsets that lie in a ball with respect to the norm. In a sketchy picture (ignoring the linear structure) the situation looks as follows. In this section we will show



A schematic drawing of bounded sets of random variables in L^p and L^1

that martingales that are bounded in L^p automatically converge in L^p to a limit X_∞. This feature of martingales is very special as typically one should not hope for convergence of a sequence only from knowing it is bounded. Martingales are just amazing!

The estimates we develop are almost more important than the theorem itself, most importantly, Doob's inequalities for the so-called running supremum $X^* := \max_{k\le n} X_k$ appears in many places of probability theory. We start with a continuation of the optional sampling theorem:

Lemma 6.3.5. Let $(X_n)_{n\in\mathbb{N}_0}$ be an $(\mathcal{F}_n)$-(sup)martingale and S, T bounded $(\mathcal{F}_n)$-stopping times with $S \le T$ almost surely. Then

(i) $\mathbb{E}[X_S] \le \mathbb{E}[X_T]$

(ii) $\mathbb{E}[X_T \,|\, \mathcal{F}_s] \ge X_s$, i.e. the (sup)martingale property also holds at bounded stopping times.

Beweis. (i) The first claim is left as an exercise:

Do you remember the proof of the optional sampling theorem sketched below Theorem 6.2.10? The same trick can be used here using $H_n = \mathbf{1}_{S<n\le T}$. Do it!

(ii) First recall a general fact: If X is a random variable on $(\Omega, \mathcal{A}, \mathbb{P})$ with $\int_A X \,\mathrm{d}\mathbb{P} \ge 0$ for all $A \in \mathcal{A}$, then $X \ge 0$ a.s. This follows directly by using the sets $A = \{X \le 0\}$ and Theorem 3.1.15 (iii).

To use this fact we show $*$ in

$$\mathbb{E}\big[\mathbb{E}[X_T \,|\, \mathcal{F}_S]\mathbf{1}_A\big] \overset{\text{cond. exp.}}{=} \mathbb{E}[X_T \mathbf{1}_A] \overset{*}{\ge} \mathbb{E}[X_S \mathbf{1}_A], \quad \forall A \in \mathcal{F}_S$$

because this implies $\mathbb{E}[(\mathbb{E}[X_T \,|\, \mathcal{F}_S] - X_S)\mathbf{1}_A] \ge 0$ for all $A \in \mathcal{F}_S$ and thus, using the fact above, $\mathbb{E}[X_T \,|\, \mathcal{F}_S] \ge X_S$ a.s.

To show $*$ fix $A \in \mathcal{F}_S$ and define $\tau_A = S \cdot \mathbf{1}_A + T \cdot \mathbf{1}_{A^C}$. Then, τ_A is an $(\mathcal{F}_n)$-stopping time (check the definition for yourself!). Finally, checking cases $\omega \in A$ and $\omega \in A^C$ for the first equality yields

$$\mathbb{E}[X_T] \overset{\text{(i)}}{\ge} \mathbb{E}[X_{\tau_A}] = \mathbb{E}[X_T - X_T\mathbf{1}_A + X_S\mathbf{1}_A] = \mathbb{E}[X_T] - \mathbb{E}[X_T\mathbf{1}_A] + \mathbb{E}[X_S\mathbf{1}_A]$$

which yields * by rearranging the inequality.

□

We can use the lemma to prove an important theorem on martingales (or submartingales). Tail probabilities of the running maximum $X^* = \max_{k\leq n} X_n$ can be estimated with last element X_n of the maximum. The inequality is not only important to prove limit theorems but appears at many places in probability theory or mathematical finance.

Theorem 6.3.6. (Doobs's maximal inequality)
Let $(X_n)_{n\in\mathbb{N}_0}$ be an $(\mathcal{F}_n)$-(sub)martingale and $\lambda > 0$. If $X_n^* := \max_{k\leq n} X_k$ denotes the running maximum process, then the following inequalities hold:

$$\lambda \cdot \mathbb{P}(X_n^* \geq \lambda) \leq \mathbb{E}[X_n \mathbf{1}_{\{X_n^* \geq \lambda\}}] \leq \mathbb{E}[X_n^+]$$

To understand better the formula it might be useful to compare with the Markov inequality. Using the Markov inequality with $h(x) = x^+$ would yield $\mathbb{E}[(X_n^*)^+]$ on the right hand side which is potentially much bigger than the expectation of only the last random variable.

Proof. Let $T := \min\{n \in \mathbb{N}_0 : X_n \geq \lambda\}$ which is an $(\mathcal{F}_n)$-stopping time so that

$$A := \{X_n^* \geq \lambda\} = \{T \leq n\} \in \mathcal{F}_n$$

and $\mathbb{E}[X_{T\wedge n}] \leq \mathbb{E}[X_n]$ by Lemma 6.3.5. Now we write

$$X_{T\wedge n} = X_T \mathbf{1}_{T\leq n} + X_n \mathbf{1}_{T>n} \geq \lambda \mathbf{1}_{T\leq n} + X_n \mathbf{1}_{T>n}$$

to get

$$\mathbb{E}[X_n] \geq \mathbb{E}[X_{T\wedge n}] \geq \lambda \cdot \mathbb{P}(T \leq n) + \mathbb{E}[X_n \mathbf{1}_{T>n}] = \lambda\, \mathbb{P}(T \leq n) + \mathbb{E}[X_n] - \mathbb{E}[X_n \mathbf{1}_{T\leq n}]$$

which implies the first inequality:

$$\lambda\, \mathbb{P}(X_n^* \geq \lambda) = \lambda\, \mathbb{P}(T \leq n) \leq \mathbb{E}[X_n \cdot \mathbf{1}_{X_n^* \geq \lambda}]$$

The second inequality follows immediately from the first:

$$\mathbb{E}[X_n \mathbf{1}_{X_n^* \geq \lambda}] \leq \mathbb{E}[X_n^+ \mathbf{1}_{X_n^* \geq \lambda}] \leq \mathbb{E}[X_n^+]$$

□

We now turn the tail estimates into moment estimates by writing the moments as integrals over the tail probabilities.

Theorem 6.3.7. (Doob's L^p-maximum inequality)
Suppose p is a constant that is strictly larger than 1.

(i) If $(X_n)_{n\in\mathbb{N}_0}$ is a non-negative $(\mathcal{F}_n)$-(sub)martingale, then

$$\mathbb{E}\big[\max_{k\leq n} X_n^p\big] \leq \Big(\frac{p}{p-1}\Big)^p \mathbb{E}[X_n^p], \quad n \in \mathbb{N}_0.$$

(ii) If $(Y_n)_{n\in\mathbb{N}_0}$ is an $(\mathcal{F}_n)$-martingale, then

$$\mathbb{E}\big[\max_{k\leq n} |Y_n|^p\big] \leq \Big(\frac{p}{p-1}\Big)^p \mathbb{E}[|Y_n|^p], \quad n \in \mathbb{N}_0.$$

Keep your eyes open to see why the proof cannot be modified in any way for $p = 1$. The corresponding theorem for $p = 1$ will be proved in the next section and is much harder.

Beweis. (i) The proof is just a clever computation keeping in mind that expectations can always be written as integrals over tail probabilities (Fubini, compare also proof of Theorem **??** or the exercises of Stochastik 1):

$$\begin{aligned}
\frac{1}{p}\mathbb{E}\big[(X_n^*)^p\big] &= \mathbb{E}\Big[\int_0^{X_n^*}\lambda^{p-1}\,\mathrm{d}\lambda\Big]\\
&\overset{\text{Fubini}}{=} \int_0^\infty \lambda^{p-2}\cdot\lambda\cdot\mathbb{P}(X_n^*\geq\lambda)\,\mathrm{d}\lambda\\
&\overset{6.3.6}{\leq} \int_0^\infty \lambda^{p-2}\mathbb{E}[X_n\mathbf{1}_{X_n^*\geq\lambda}]\,\mathrm{d}\lambda\\
&\overset{\text{Fubini}}{=} \mathbb{E}\Big[X_n\int_0^\infty\lambda^{p-2}\mathbf{1}_{X_n^*\geq\lambda}\,\mathrm{d}\lambda\Big]\\
&= \mathbb{E}\Big[X_n\int_0^{X_n^*}\lambda^{p-2}\,\mathrm{d}\lambda\Big]\\
&= \frac{1}{p-1}\mathbb{E}\big[X_n\cdot(X_n^*)^{p-1}\big]\\
&\overset{\text{Hölder } q=\frac{p}{p-1}}{\leq} \frac{1}{p-1}\mathbb{E}\big[(X_n)^p\big]^{\frac{1}{p}}\mathbb{E}\big[(X_n^*)^p\big]^{\frac{p-1}{p}}
\end{aligned}$$

Dividing both sides gives the result.

(ii) follows from (i) as $X_n := |Y_n|$ is a submartingale

□

As announced at the beginning of this section we will need to impose a stronger assumption on martingales in order to strengthen the almost sure convergence to L^p-convergence. A good condition is uniform L^p-boundedness:

Definition 6.3.8. A martingale is called L^p-martingale if $\sup_{n\in\mathbb{N}_0}\mathbb{E}[|X_n|^p]<\infty$. Most importantly, for $p=2$ we speak of square-integrable martingales.

It is very important to keep in mind that the definition does not require finiteness of all pth moments but uniform boundedness of pth moments. Modulo zero sets this means the random variables of the martingale form a bounded set in the normed space $(L^p, ||\cdot||_p)$. Since $(|X_n|^p)_{n\in\mathbb{N}_0}$ is a submartingale we know that $n\mapsto\mathbb{E}[X_n^p]$ is increasing, possibly towards $+\infty$. For L^p-martingales the sequence of pth moments does not grow to $+\infty$ but has a finite limit.

We can now prove that L^p-boundedness is not only a necessary but also a sufficient condition for the L^p-convergence of martingales. The equivalence is not true for other sequences of random variables, we heavily use the martingale property.

Theorem 6.3.9. **(L^p-martingale convergence theorem)**
Suppose p is a constant that is strictly larger than 1 and $(X_n)_{n\in\mathbb{N}_0}$ is an L^p-martingale.

(i) There is a (finite) limiting random variable $X_\infty\in\mathcal{L}^p$ such that

- $X_n\overset{\text{a.s.}}{\to}X_\infty$ for $n\to\infty$,
- $X_n\overset{L^p}{\to}X_\infty$ for $n\to\infty$.

(ii) $\mathbb{E}\big[|X_\infty|^p\big]=\lim_{n\to\infty}\mathbb{E}\big[|X_n|^p\big]=\sup_{n\in\mathbb{N}_0}\mathbb{E}\big[|X_n|^p\big]<\infty$

(iii) $\mathbb{E}\big[\sup_{n\in\mathbb{N}_0}|X_n|^p\big]\leq\big(\frac{p}{p-1}\big)^p\cdot\mathbb{E}\big[|X_\infty|^p\big]<\infty$

Proof. (i) The first step is simple, using the simple estimate

$$x^+ \leq |x|^p + 1, \quad x \in \mathbb{R}.$$

Hence, every L^p-martingale satisfies the condition $\sup_{n\in\mathbb{N}_0} \mathbb{E}[X_n^+] < \infty$ of the martingale convergence theorem. Then there is a (finite) almost sure limit X_∞. It remains to show that $X_\infty \in \mathcal{L}^p$ and the L^p-convergence. There is a simple thought to keep in mind, which also helps to appreciate Doob's inequalities:

If $\sup_{k\in\mathbb{N}_0} |X_k|^p$ is integrable, than this is the perfect upper bound for dominated convergence.

To check the integrability of the upper bound we use Theorem 6.3.7

$$\mathbb{E}\big[|X_k^*|^p\big] \leq \Big(\frac{p}{p-1}\Big)^p \mathbb{E}\big[|X_k|^p\big] \leq \Big(\frac{p}{p-1}\Big)^p \sup_{n\in\mathbb{N}_0} \mathbb{E}\big[|X_n|^p\big], \quad \forall k \in \mathbb{N}_0,$$

and monotone convergence:

$$\mathbb{E}\big[|\sup_{k\in\mathbb{N}_0} X_k|^p\big] = \mathbb{E}\big[\lim_{k\to\infty} |X_k^*|^p\big] \leq \lim_{k\to\infty} \Big(\frac{p}{p-1}\Big)^p \mathbb{E}\big[|X_k|^p\big] < \infty.$$

Hence, $\sup_{k\in\mathbb{N}_0} X_k \in \mathcal{L}^p$ from which we immediately deduce $X_\infty \in \mathcal{L}^p$ because $|X_\infty| \leq \sup_{k\in\mathbb{N}_0} |X_k|$. For the L^p-convergence first note that

$$|X_n - X_\infty|^p \leq (|X_n| + \lim_{n\to\infty} |X_n|)^p \leq 2^p \sup_{k\in\mathbb{N}_0} |X_k|^p \quad \text{for all } n \in \mathbb{N}_0,$$

and the right hand side is integrable. Then we can apply dominated convergence to the sequence of the left hand side:

$$\lim_{n\to\infty} \mathbb{E}\big[|X_n - X_\infty|^p\big] = \mathbb{E}\big[\lim_{n\to\infty} |X_n - X_\infty|^p\big] = \mathbb{E}[0] = 0.$$

(ii) The first equality follows from basic functional analysis as convergence in a normed space implies convergence of norms: $\lim_{n\to\infty} ||X_n - X|| = 0 \Rightarrow \lim_{n\to\infty} ||X_n|| = ||X||$ because norms are continuous functions. Keeping in mind that L^p is a normed space (modulo zero sets) we immediately find

$$\mathbb{E}\big[|X_\infty|^p\big] = \lim_{n\to\infty} \mathbb{E}\big[|X_n|^p\big].$$

The second equality follows from the monotonicity of $n \mapsto \mathbb{E}[|X_n|^p]$ as $(|X_n|^p)_{n\in\mathbb{N}_0}$ is a submartingale.

(iii) The inequality follows from monotone convergence and Doob's L^p-maximal inequality:

$$\mathbb{E}\Big[\sup_{n\in\mathbb{N}_0} |X_n|^p\Big] = \lim_{n\to\infty} \mathbb{E}\Big[\sup_{k\leq n} |X_k|^p\Big] \leq \lim_{n\to\infty} \Big(\frac{p}{p-1}\Big)^p \mathbb{E}\big[|X_n|^p\big] \overset{(ii)}{=} \Big(\frac{p}{p-1}\Big)^p \mathbb{E}\big[|X_\infty|^p\big]$$

□

We come back to the branching processes as prime example for the convergence of martingales. The almost sure convergence of M towards a finite limit M_∞ was already checked. In the supercritical case we left open if M_∞ is trivial, i.e. almost surely equal to 0, or not. We can solve this question under the additional assumption that the offspring number has finite second moment.

Example 6.3.10. Let us additionally assume $\sigma^2 := \mathbb{V}[\xi_1^1] < \infty$, i.e. the offspring numbers have finite second moments. To keep things easy we assume $m = 1$, there is one individual at time 0. Wald's identity was used to compute first moments, to compute second moments of random sums one can first prove the Blackwell-Girshick identity:

Suppose $N, Y_1, Y_2, ...$ are independent, $\mathbb{E}[N] < \infty$, and $Y_1, Y_2, ...$ are identically distributed, then

$$\mathbb{E}\Big[\Big(\sum_{k=1}^{N} Y_k\Big)^2\Big] = \mathbb{E}[N]\mathbb{V}[Y_1] + \mathbb{E}[N^2]\mathbb{E}[Y_1]^2. \tag{6.2}$$

Hence, we can compute the second moments of the martingale $M_n = \mu^{-n} X_n$:

$$\mathbb{E}\big[M_n^2\big] = \frac{1}{\mu^{2n}}\big(\mu^{n-1}\sigma^2 + \mathbb{E}\big[X_{n-1}^2\big]\mu^2\big) = \frac{\sigma^2}{\mu^{n+1}} + \mathbb{E}\big[M_{n-1}^2\big]$$

By induction this iteration gives

$$\mathbb{E}\big[M_n^2\big] = \frac{\sigma^2}{\mu}\sum_{k=1}^{n}\mu^{-k} + 1$$

which increases for $\mu > 1$ (geometric series) to a finite positive number. Hence, for $\mu > 1$ we deduce L^2-convergence and, most importantly, $\mathbb{E}[M_\infty^2] = \lim_{n\to\infty}\mathbb{E}[M_n^2] > 0$. But this implies $M_\infty > 0$ with positive probability. If now we compare with Example 6.3.4 we proved that the supercritical branching processes can increase exponentially fast with positive probability.

6.3.3 L^1-martingale convergence theorem

Lecture 8

We proved that L^p-bounded martingales converge automatically almost surely and in L^p to a limit X_∞. How about the same theorem for $p = 1$? Unfortunately, the story is more complicated. The right condition for convergence in L^1 is not L^1-boundedness but a slightly stronger condition, so-called uniform integrability. To be honest, the world would be too simple if L^1-boundedness would be enough as otherwise every non-negative martingale would not only converge almost surely but also in L^1 as expectations of martingales are constant. But this cannot be as $\mathbb{E}[X_\infty] = \mathbb{E}[0] = 0 \neq \mathbb{E}[X_0]$ for the critical branching processes.

Definition 6.3.11. A familiy $(X_\alpha)_{\alpha\in I}$ of random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ indexed by a set $I \neq \emptyset$ is called **uniformly integrable** if

- $\mathbb{E}[|X_\alpha|] < \infty$ for all $\alpha \in I$,
- $\lim_{M\to\infty}\Big(\sup_{\alpha\in I}\mathbb{E}\left[|X_\alpha| \cdot \mathbf{1}_{|X_\alpha|\geq M}\right]\Big) = 0.$

All the trouble about understanding this section is that we do not have simple examples to keep in mind. Everything is about abstract integration theory.

Example 6.3.12. At least we have some classes of random variables that we are used to work with:

(i) Families consisting of only one integrable random variable $\{X\}$ are uniformly integrable. This is easy:

$$\lim_{M\to\infty}\mathbb{E}[|X|\mathbf{1}_{|X|\geq M}] \overset{\text{DCT}}{=} 0$$

The same holds for finite families:

Check that every finite number of integrable random variables forms a uniformly integrable family.

(ii) Every family of random variables that is dominated by an integrable random variable is uniformly integrable. We know this situation very well from the dominated convergence theorem! Suppose $|X_\alpha| \leq Z$ a.s. for all $\alpha \in I$ and $\mathbb{E}[|Z|] < \infty$, then

$$\lim_{M\to\infty} \sup_{\alpha\in I} \mathbb{E}\big[|X_\alpha| \cdot \mathbf{1}_{|X_\alpha|\geq M}\big] \leq \lim_{M\to\infty} \mathbb{E}[Z \cdot \mathbf{1}_{Z\geq M}] \overset{\text{DCT}}{=} 0, \quad M \to \infty$$

(iii) If $(X_\alpha)_{\alpha\in I}$ is bounded in L^p for some $p > 1$, i.e. $\sup_{\alpha\in I} \mathbb{E}[|X_\alpha|^p] < C$, then $(X_\alpha)_{\alpha\in I}$ is also uniformly integrable. We will later use this situation to compare the L^1-convergence theorem to the L^p-convergence theorem of the previous section. To proof the claim let us first use the Hölder inequality with p and M fixed:

$$\mathbb{E}[|X_\alpha|\mathbf{1}_{|X_\alpha|\geq M}] \leq \big(\mathbb{E}[|X_\alpha|^p]\big)^{\frac{1}{p}} \big(\mathbb{E}[\mathbf{1}_{|X_\alpha|\geq M}]\big)^{\frac{1}{q}} \leq C^{1/p} \big(\mathbb{P}(|X_\alpha| \geq M)\big)^{\frac{1}{q}}. \tag{6.3}$$

We use the inequality to argue indirectly and assume the right hand side does not converge to zero uniformly in α. If there is $\varepsilon > 0$ with $\liminf_{M\to\infty} \sup_{\alpha\in I} \mathbb{E}[|X_\alpha|\mathbf{1}_{|X_\alpha|\geq M}] > \varepsilon$, then there is a sequence α_M with $\mathbb{P}(|X_{\alpha_M}| \geq M) > (\frac{\varepsilon}{2C^{1/p}})^q$. But then

$$\mathbb{E}\big[|X_{\alpha_M}|^p\big] \geq \mathbb{E}\big[|X_{\alpha_M}|^p \mathbf{1}_{|X_{\alpha_M}|\geq M}\big] \geq \Big(\frac{\varepsilon}{2C^{1/p}}\Big)^q \cdot M^p \to \infty, \quad M \to \infty,$$

which is a contradiction to L^p-boundedness. Hence, the right hand side of (6.3) goes to zero for all M. But then also the left hand side goes to zero for all M, which is exactly the uniform integrability.

Definition 6.3.13. A martingale is called uniformly integrable martingale if $(X_n)_{n\in\mathbb{N}_0}$ is a uniformly integrable family of random variables.

Most importantly, the previous example shows that all L^p-martingales are also uniformly integrable martingales.

If we compare with the first example we see clearly the point of uniform integrability. Integrability of X_α means that $f_M(\alpha) := \mathbb{E}\left[|X_\alpha|\mathbf{1}_{\{|X_\alpha|\geq M\}}\right]$ vanishes as $M \to \infty$, there is not too much mass near infinity. Uniform integrability means that $f_M(\alpha)$ vanishes uniformly in α, all random variables have equally little mass at infinity. With this intuition it should be clear that the following gives good counter examples, check it!

If $X_\alpha \sim \delta_{x_\alpha}$, then $(X_n)_{\alpha\in I}$ is uniformly bounded if and only if $\{x_\alpha : \alpha \in I\} \subseteq \mathbb{R}$ is bounded.

We could ask ourselves how close uniform integrability is to L^1-boundedness, that is boundedness of the set $(X_\alpha)_{\alpha\in I}$ seen as a subset of L^1, in formulas $\sup_{\alpha\in I} ||X_\alpha||_1 < \infty$. In fact, it is easy to see that uniformly integrable families are also bounded as subsets of L^1:

$$\sup_{\alpha\in I} ||X_\alpha||_1 \leq \sup_{\alpha\in I} \mathbb{E}\left[|X_\alpha|\mathbf{1}_{|X_\alpha|\leq M}\right] + \sup_{\alpha\in I} \mathbb{E}\left[|X_\alpha|\mathbf{1}_{|X_\alpha|> M}\right] \leq M + 1 \tag{6.4}$$

for some M large enough. The next proposition gives a criterion which bounded subsets of L^1 are actually the uniformly integrable sets.

Proposition 6.3.14. Suppose the family $(X_\alpha)_{\alpha\in I}$ is bounded as a subset of L^1, i.e. $\sup_{\alpha\in I} \mathbb{E}[|X_\alpha|] < \infty$. Then the following are equivalent:

(i) $(X_\alpha)_{\alpha\in I}$ is uniformly integrable.

(ii) $\forall \varepsilon > 0\, \exists \delta > 0$: $\sup\limits_{\alpha\in I} \mathbb{E}[|X_\alpha|\mathbf{1}_A] < \varepsilon \;\forall A \in \mathcal{F}$ with $\mathbb{P}(A) < \delta$

Without any doubt the criterion does not look useful at all but it will be applicable for the most important example to follow, families that remind us of Doob martingales. The advantage is that the sets A do not depend on α compared to the sets $\{|X_\alpha| > M\}$ that appear in the definition of uniform integrability.

Proof. (i) $\Rightarrow$ (ii): Fix $\varepsilon > 0$ and choose M large enough so that $\sup_{\alpha\in I}\mathbb{E}[|X_\alpha|\mathbf{1}_{|X_\alpha|\geq M}] < \frac{\varepsilon}{2}$. Such an M exists by the definition of uniform integrability. Now we set $\delta := \frac{\varepsilon}{2M}$ and check the inequality for all $A \in \mathcal{F}$ with $\mathbb{P}(A) < \delta$:

$$\begin{aligned}\sup_{\alpha\in I}\mathbb{E}[|X_\alpha|\mathbf{1}_A] &\leq \sup_{\alpha\in I}\mathbb{E}[|X_\alpha|\mathbf{1}_A\mathbf{1}_{|X_\alpha|\geq M}] + \sup_{\alpha\in I}\mathbb{E}[|X_\alpha|\mathbf{1}_A\mathbf{1}_{|X_\alpha|<M}]\\ &\leq \frac{\varepsilon}{2} + M\cdot\mathbb{P}(A) = \varepsilon,\end{aligned}$$

where we got rid of the indicators using monotonicity of expectations.

(ii) $\Rightarrow$ (i): Exercise! □

Lemma 6.3.15. Suppose $Z \in \mathcal{L}^1(\Omega,\mathcal{F},\mathbb{P})$.

(i) The family

$$\big\{\mathbb{E}[Z|\mathcal{G}] \colon \mathcal{G} \text{ sub-}\sigma\text{-Algebra of } \mathcal{F}\big\}$$

is uniformly integrable family of random variables.

(ii) If $(\mathcal{F}_n)_{n\in\mathbb{N}_0}$ is a filtration, then $X_n := \mathbb{E}[X|\mathcal{F}_n]$ is a uniformly integrable martingale.

Beweis. (ii) follows immediately from (i), the claimed martingale property was proved in Example 6.2.4.

To prove (i) we will use the characterisation from Proposition 6.3.14 for the uniformly integrable family $\{Z\}$ to deduce the definition of uniform integrability for the family of conditional expectations. Recalling the Markov inequality yields

$$\mathbb{P}\big(|\mathbb{E}[Z|\mathcal{G}]| > M\big) \leq \frac{\mathbb{E}\big[|\mathbb{E}[Z|\mathcal{G}]|\big]}{M} = \frac{\mathbb{E}[|Z|]}{M} \tag{6.5}$$

for all $G \subseteq \mathcal{F}$. The important point is that the upper bound is independent of G! We now fix $\varepsilon > 0$, choose δ from Proposition 6.3.14 and M large enough so that $\frac{\mathbb{E}[|Z|]}{M} < \delta$ and use the sets $A := \{|\mathbb{E}[Z|\mathcal{G}] \geq M\}$. Then the inequality

$$\mathbb{E}\big[|Z|\cdot\mathbf{1}_{|\mathbb{E}[Z|\mathcal{G}]|\geq M}\big] < \varepsilon$$

applies for all such M and all $\mathcal{G}$. Using properties of conditional expectation and the above estimate yields

$$\begin{aligned}\mathbb{E}\big[|\mathbb{E}[Z|\mathcal{G}]|\cdot\mathbf{1}_{|\mathbb{E}[Z|\mathcal{G}]|\geq M}\big] &\leq \mathbb{E}\big[\mathbb{E}[|Z||\mathcal{G}]\cdot\mathbf{1}_{|\mathbb{E}[Z|\mathcal{G}]|\geq M}\big]\\ &\overset{\text{meas.}}{=} \mathbb{E}\big[\mathbb{E}[|Z|\cdot\mathbf{1}_{|\mathbb{E}[Z|\mathcal{G}]|\geq M}|\mathcal{G}]\big]\\ &= \mathbb{E}\big[|Z|\cdot\mathbf{1}_{|\mathbb{E}[Z|\mathcal{G}]|\geq M}\big] < \varepsilon\end{aligned}$$

for all sub-σ-algebras $\mathcal{G}$. But this is exactly the definition of uniform integrability written in ε-M-notation. □

So far it is completely unclear why we are discussing uniformly integrable families of random variables. So far you only learnt sufficient conditions under which limits and expectation can be exchanged for almost surely converging sequences of random variables. But how about necessary and sufficient conditions?

Theorem 6.3.16. (Generalized DCT)
Suppose $(X_n)_{n\in\mathbb{N}_0}$ is a sequence of integrable random variables and there is a random variable X_∞ such that $X_n \overset{\text{a.s.}}{\to} X_\infty$ for $n\to\infty$.
Then the following conditions are equivalent:

(i) $(X_n)_{n\in\mathbb{N}_0}$ is uniformly integrable,

(ii) $X_n \overset{L^1}{\to} X_\infty$ for $n\to\infty$,

(iii) $\lim_{n\to\infty}\mathbb{E}[|X_n|] = \mathbb{E}[|X_\infty|]$.

In all cases it holds that $\lim_{n\to\infty}\mathbb{E}[X_n] = \mathbb{E}[X_\infty]$.

Proof. The equivalence of (ii) and (iii) is called Scheffé's lemma and will be covered in the exercises.

(ii) $\Rightarrow$ (i): Any finite family of integrable random variables is uniformly integrable, so by Proposition 6.3.14 applied to $(X_n)_{n\le K}$, for every $\varepsilon'>0$ there are constants $\delta(K,\varepsilon')>0$ with

$$\sup_{n\le K}\mathbb{E}[|X_n|\mathbf{1}_A] < \frac{\varepsilon'}{4} \quad \forall A\in\mathcal{F} \text{ with } \mathbb{P}(A)<\delta(K,\varepsilon'). \tag{6.6}$$

To apply 6.3.14 for the entire sequence $(X_n)_{n\in\mathbb{N}_0}$ (but in the other direction) we need a version of (6.6) where δ is independent of K. To do so, fix $\varepsilon>0$. Since L^1 is complete (compare Theorem **??**, for a proof we refer to functional analysis) there is an $N\in\mathbb{N}$ so that

$$\mathbb{E}\big[|X_N - X_n|\big] < \frac{\varepsilon}{2} \quad \forall n\ge N. \tag{6.7}$$

This is nothing but the definition of a Cauchy-sequence for the norm $||X||_1 = \mathbb{E}[|X|]$. Hence, using this N and $\delta := \delta(N,\varepsilon)$ from above, we get

$$\begin{aligned}
\sup_{n\in\mathbb{N}_0}\mathbb{E}\big[|X_n|\mathbf{1}_A\big] &\le \sup_{n<N}\mathbb{E}\big[|X_n|\mathbf{1}_A\big] + \sup_{n\ge N}\mathbb{E}\big[|X_n|\mathbf{1}_A\big]\\
&\overset{\Delta}{\le} \sup_{n<N}\mathbb{E}\big[|X_n|\mathbf{1}_A\big] + \mathbb{E}\big[|X_N|\mathbf{1}_A\big] + \sup_{n\ge N}\mathbb{E}\big[|X_n - X_N|\mathbf{1}_A\big]\\
&\le 2\sup_{n\le N}\mathbb{E}\big[|X_n|\mathbf{1}_A\big] + \sup_{n\ge N}\mathbb{E}\big[|X_n - X_N|\big]\\
&\overset{(6.6),(6.7)}{\le} \frac{\varepsilon}{2}+\frac{\varepsilon}{2} = \varepsilon
\end{aligned}$$

for all $A\in\mathcal{F}$ with $\mathbb{P}(A)<\delta$. But then (ii) of Proposition 6.3.14 is justified.

(i) $\Rightarrow$ (ii): First note that from 6.3.14 and

$$\mathbb{E}\big[|X_n - X_m|\mathbf{1}_A\big] \overset{\Delta}{\le} \mathbb{E}\big[|X_n|\mathbf{1}_A\big] + \mathbb{E}\big[|X_m|\mathbf{1}_A\big]$$

also the family $(|X_n - X_m|)_{n,m\in\mathbb{N}_0}$ is uniformly integrable (choose $\frac{\varepsilon}{2}$ in (ii) of 6.3.14). Hence, using the definition of uniform integrability, for all $\varepsilon>0$ there is some $M>0$ such that

$$\mathbb{E}\big[|X_n - X_m|\cdot\mathbf{1}_{|X_n - X_m|\ge M}\big] < \varepsilon \;\forall n,m$$

We now combine this estimate with the convergence in probability to show that $(X_n)_{n\in\mathbb{N}}$ is a

Cauchy sequence in L^1:

$$
\begin{aligned}
&\mathbb{E}\big[|X_n - X_m|\big]\\
&\leq \underbrace{\mathbb{E}\big[|X_n - X_m|\mathbf{1}_{|X_n - X_m|\geq M}\big]}_{\leq \varepsilon\ \forall n,m} + \underbrace{\mathbb{E}\big[|X_n - X_m|\mathbf{1}_{\varepsilon<|X_n - X_m|<M}\big]}_{\leq M\mathbb{E}[\mathbf{1}_{\varepsilon<|X_n-X_m|<M}]\ \forall n,m} + \underbrace{\mathbb{E}\big[|X_n - X_m|\mathbf{1}_{|X_n - X_m|\leq \varepsilon}\big]}_{\leq \varepsilon\cdot 1\ \forall n,m}\\
&\leq \varepsilon + M\cdot \mathbb{P}(|X_n - X_m| > \varepsilon) + \varepsilon\\
&\overset{\Delta,\mathbb{P}\text{ monot.}}{\leq} 2\varepsilon + M\cdot\mathbb{P}(|X_n - X_\infty| + |X_m - X_\infty| > \varepsilon)\\
&\overset{\mathbb{P}\text{ monot.}}{\leq} 2\varepsilon + \mathbb{P}(\{|X_n - X_\infty| > \varepsilon/2\}\cup\{|X_m - X_\infty| > \varepsilon/2\})\\
&\overset{\text{subadd.}}{\leq} 2\varepsilon + \mathbb{P}(\{|X_n - X_\infty| > \varepsilon/2) + \mathbb{P}(\{|X_m - X_\infty| > \varepsilon/2\})\\
&\to 2\varepsilon, \quad n,m\to\infty,
\end{aligned}
$$

since $X_n \xrightarrow{P} X_\infty$. Hence, $(X_n)_{n\in\mathbb{N}}$ is Cauchy in L^1 and thus converges to some limit X. Since convergence in L^1 also implies convergence in probability we also find that $X_n \xrightarrow{P} X$ and we also have that $X_n \xrightarrow{P} X_\infty$ by assumption. Using that convergence in probability has unique limits we can deduce $X = X_\infty$ which proves the L^1 convergence towards X_∞.

The final claim follows from the above as follows. Using $|X_n^+| \leq |X_n|$ and $|X_n^-| \leq |X_n|$ one can check readily that uniform integrability of (X_n) implies uniform integrability of (X_n^+) and (X_n^-). Then the above is applied to positive- and negative part and the claim follows from linearity. □

By the way, can you prove the claim used in the proof?

> If $X_n \xrightarrow{P} X$ and $X_n \xrightarrow{P} Y$, then $X = Y$ almost surely. The easiest way to check the exercise is to carefully look at Theorem **??** and use that limits are unique in normed spaces.

Before turning our attention towards the L^1-martingale convergence theorem let us check how the previous theorem relates to dominated convergence for non-negative sequences. Sequences dominated by an integrable random variable are uniformly integrable, hence, the previous theorem implies the dominated convergence theorem. Lecture 9

Now towards the main theorem of this section:

> **Theorem 6.3.17.** **(L^1-martingale convergence theorem)**
> Suppose $(X_n)_{n\in\mathbb{N}_0}$ is a martingale on $(\Omega,\mathcal{F},\mathbb{P})$. Then the following statements are equivalent:
>
> (i) There is a random variable X_∞ with $X_n \to X_\infty$ a.s. and in L^1 for $n\to\infty$.
>
> (ii) $(X_n)_{n\in\mathbb{N}_0}$ is uniformly integrable.
>
> (iii) $(X_n)_{n\in\mathbb{N}_0}$ is a closed martingale.
>
> In all cases $(X_n)_{n\in\bar{\mathbb{N}}_0}$ satisfies the martingale property on $\bar{\mathbb{N}}_0 = \mathbb{N}_0\cup\{\infty\}$ with terminal element X_∞ and $\mathbb{E}[X_\infty] = \lim_{n\to\infty}\mathbb{E}[X_n]$.

Recall from Example 6.2.4 the notion of a closed martingale (or Doob martingale). There is an integrable random variable Z such that $X_n = \mathbb{E}[Z|\mathcal{F}_n]$ almost surely for all $n\in\mathbb{N}_0$.

Proof. (i) $\Rightarrow$ (ii): Follows from Theorem 6.3.16.

(ii) $\Rightarrow$ (iii): We start with a convergence property of conditional expectation:

$$X_n \xrightarrow{L^1} X_\infty, \quad n\to\infty \quad\Longrightarrow\quad \mathbb{E}[X_n|\mathcal{G}] \xrightarrow{L^1} \mathbb{E}[X_\infty|\mathcal{G}], \quad n\to\infty \tag{6.8}$$

holds for all sub-σ-algebras $\mathcal{G} \subseteq \mathcal{F}$. Looking at the definition of L^1-convergence we see immediately what needs to be done:

$$\mathbb{E}\big[|\mathbb{E}[X_n|\mathcal{G}] - \mathbb{E}[X_\infty|\mathcal{G}]|\big] \leq \mathbb{E}\big[\mathbb{E}[|X_n - X_\infty||\mathcal{G}]\big] = \mathbb{E}[|X_n - X_\infty|] \to 0, \quad n \to \infty.$$

We can now combine everything we learnt so far. First recall from (6.4) that uniformly integrable implies L^1-bounded and, hence, $\sup_{n\in\mathbb{N}_0} \mathbb{E}[X_n^+] \leq \sup_{n\in\mathbb{N}_0} \mathbb{E}[|X_n|] < \infty$. But this is the assumption of the almost sure martingale convergence theorem 6.3.1 so we get the existence of an almost sure limit X_∞. Since almost sure convergence implies convergence in probability Theorem 6.3.16 implies the L^1-convergence of X_n towards X_∞. To see that $(X_n)_{n\in\mathbb{N}_0}$ is a closed martingale with $Z := X_\infty$ we use (6.8) for fixed n:

$$X_n \overset{m\geq n \text{ arbitrary}}{=} \mathbb{E}[X_m \,|\, \mathcal{F}_n] \overset{L^1}{\to} \mathbb{E}[X_\infty \,|\, \mathcal{F}_n], \quad m \to \infty$$

Hence, X_n and $\mathbb{E}[X_\infty|\mathcal{F}_n]$ coincides in L^1, which means $X_n = \mathbb{E}[X_\infty \,|\, \mathcal{F}_n]$ almost surely. But this is nothing but $(X_n)_{n\in\mathbb{N}_0}$ is a closed martingale with $Z = X_\infty$.

(iii) $\Rightarrow$ (i): Lemma 6.3.15 implies that $(X_n)_{n\in\mathbb{N}}$ is uniformly integrable, thus, $(X_n)_{n\in\mathbb{N}}$ is bounded in L^1 by 6.4. In particular $\sup_{n\in\mathbb{N}_0} \mathbb{E}[X_n^+] \leq \sup_{n\in\mathbb{N}_0} \mathbb{E}[|X_n|] < \infty$ so that the almost sure martingale convergence theorem 6.3.1 implies the existence of an almost sure limit X_∞. Since almost sure convergence implies convergence in probability, the L^1 convergence follows from Theorem 6.3.16.

The additional statement follows from the proof of (ii) $\Rightarrow$ (iii) and Theorem 6.3.16. $\square$

There is an interesting detail concerning Doob martingales. Suppose Z is not $\mathcal{F}_\infty$-measurable and $X_n = \mathbb{E}[Z|\mathcal{F}_n]$ is the corresponding closed martingale to which we apply the theorem. Then it does not hold that $X_n \to Z$ almost surely, as otherwise Z would be $\mathcal{F}_\infty$-measurable. On the other hand, in the proof (ii) $\Rightarrow$ (iii) the random variable that closes the martingale was constructed as the almost sure limit X_∞. What looks like a contradiction only reflects the fact that a martingale can be closed by different random variables. Define $\bar{Z} := \mathbb{E}[Z|\mathcal{F}_\infty]$, then

$$\mathbb{E}[\bar{Z}|\mathcal{F}_n] = \mathbb{E}\big[\mathbb{E}[Z|\mathcal{F}_\infty]\big|\mathcal{F}_n\big] = \mathbb{E}[Z|\mathcal{F}_n]$$

so that Z and $\bar{Z}$ are different but induce the same Doop martingale. Here is a question: Can we still identify the limit X_∞ of a Doob martingale? Yes, the limit is $\mathbb{E}[Z|\mathcal{F}_\infty]$, which equals Z if Z is $\mathcal{F}_\infty$-measurable.

Proposition 6.3.18. Let $Z \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ and $(\mathcal{F}_n)_{n\in\mathbb{N}_0}$ an increasing family of σ-algebras, then

$$\mathbb{E}[Z|\mathcal{F}_n] \overset{\text{a.s}/L^1}{\to} \mathbb{E}[Z|\mathcal{F}_\infty], \quad n \to \infty.$$

Beweis. Let us first assume $Z \geq 0$. If we define $X_n := \mathbb{E}[Z|\mathcal{F}_n]$, then X is a uniformly integrable martingale so there is a limit X_∞ (a.s. and in L^1) satisfying $X_n = \mathbb{E}[X_\infty|\mathcal{F}_n]$ a.s. Now take $A \in \mathcal{F}_n \subseteq \mathcal{F}_\infty$. Then

$$V_{X_\infty}(A) := \mathbb{E}[X_\infty \cdot \mathbf{1}_A] \overset{A\in\mathcal{F}_n}{=} \mathbb{E}[X_n \cdot \mathbf{1}_A] = \mathbb{E}[\mathbb{E}[Z|\mathcal{F}_n] \cdot \mathbf{1}_A] \overset{A\in\mathcal{F}_n}{=} \mathbb{E}[Z \cdot \mathbf{1}_A] =: V_Z(A)$$

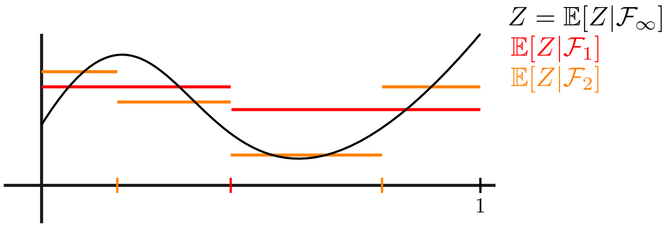
from the properties of conditional expectation. Now recall from 3.2.2 that $V_{X_\infty}(A) := \mathbb{E}[X_\infty \cdot \mathbf{1}_A]$ and $V_Z(A) := \mathbb{E}[Z \cdot \mathbf{1}_A]$ are measures on $\mathcal{F}_\infty$ for which we just proved equality on the $\cap$-stable generator $\cup_{k=1}^\infty \mathcal{F}_k$. Hence, by Theorem 1.2.12, both measures are equal on $\mathcal{F}_\infty$, i.e. $\mathbb{E}[X_\infty \cdot \mathbf{1}_A] = \mathbb{E}[Z \cdot \mathbf{1}_A]$ for all $A \in \mathcal{F}_\infty$. But then $X_\infty = \mathbb{E}[Z|\mathcal{F}_\infty]$ almost surely.

Finally, splitting $Z = Z^+ - Z^-$ and applying the above to both summands yields the claim. $\square$

Here is a fun application that completes the story of interpreting conditional expectation as approximation with information given by the σ-algebra. Recall the pictures from Section 5.1 where we best (in the sense of L^2) approximated a Borel function $Z : [0,1] \to \mathbb{R}$ by simple functions. Now we use the martingale convergence theorems to get a convergence theorem. Let $\Omega = [0,1]$, $\mathcal{F} = \mathcal{B}([0,1])$, and

$$\mathcal{F}_n = \sigma\Big(\Big\{\Big[0, \frac{1}{2^n}\Big), \Big[\frac{1}{2^n}, \frac{2}{2^n}\Big), ..., \Big[\frac{2^n-1}{2^n}, 1\Big]\Big\}\Big),$$

that is, we partition the interval $[0,1]$ finer and finer by dividing the intervals successively in half.



[3]

Then $(\mathcal{F}_n)$ is increasing and it is not hard to see that $\mathcal{F}_\infty = \mathcal{B}([0,1])$ as $\mathcal{F}_\infty$ contains all intervals with rational end points. But then,

$$\mathbb{E}[Z|\mathcal{F}_n] \overset{\text{a.s}/L^1}{\to} \mathbb{E}[Z|\mathcal{B}([0,1])] \overset{\text{meas.}}{=} Z, \quad n \to \infty.$$

If $Z \in \mathcal{L}^p$, then $\mathbb{E}[|\mathbb{E}[Z|\mathcal{F}_n]|^p] \leq \mathbb{E}[\mathbb{E}[|Z^p| \,|\, \mathcal{F}_n] = \mathbb{E}[|Z|^p] < \infty$ also ensure L^p-convergence due to the L^p-martingale convergence theorem.

A further application of the L^1-martingale convergence theorem gives us another version of optional sampling of martingales. Recall from Theorem 6.2.6 that stopping at bounded stopping times does not change the expectation of a martingale while in general this must not be the case as we have seen for the simple random walks. Uniformly integrable martingales can be seen as similar to finite time-horizon and as such allow for optional sampling with arbitrary stopping times:

Theorem 6.3.19. (optional sampling revisited)
Suppose $(X_n)_{n\in\mathbb{N}_0}$ is a uniformly integrable $(\mathcal{F}_n)$-martingale with limit X_∞ and let S, T be $(\mathcal{F}_n)$-stopping times with $S \leq T$ almost surely. Then

(i) $\mathbb{E}[X_\infty \,|\, \mathcal{F}_T] = X_T$ a.s.

(ii) $\mathbb{E}[X_T] = \mathbb{E}[X_\infty] = \mathbb{E}[X_n]$ for all $n \in \mathbb{N}$

(iii) $X_S = \mathbb{E}[X_T \,|\, \mathcal{F}_S]$ a.s.

Keep in mind that all L^p-martingales are uniformly integrable, thus, the theorem can for instance be applied to the supercritical branching process with finite second moment offspring numbers from Example 6.3.10.

[3]Leif: Bild schlecht, Punkte bei 1/4, 1/2, 3/4

Beweis. Let us first check that X_T is integrable using that $(X_n)_{n\in\mathbb{N}_0}$ can be extended to a martingale indexed by $\mathbb{N}_0 \cup \{\infty\}$ (i.e. is closed by X_∞):

$$\begin{aligned}
\mathbb{E}[|X_T|] &= \mathbb{E}\Big[|X_T| \cdot \big(\sum_{k=1}^{\infty} \mathbf{1}_{T=k} + \mathbf{1}_{T=\infty}\big)\Big]\\
&\overset{\text{MCT}}{=} \sum_{k=0}^{\infty} \mathbb{E}\big[\mathbf{1}_{T=k}|X_k|\big] + \mathbb{E}\big[\mathbf{1}_{T=\infty}|X_\infty|\big]\\
&= \sum_{k=0}^{\infty} \mathbb{E}\big[\underbrace{\mathbf{1}_{T=k}}_{\{T=k\}\in\mathcal{F}_k} |\mathbb{E}[X_\infty \,|\, \mathcal{F}_k]|\big] + \mathbb{E}\big[\mathbf{1}_{T=\infty}|X_\infty|\big]\\
&\overset{\Delta,\text{ cond. exp.}}{\leq} \sum_{k=0}^{\infty} \mathbb{E}\big[\mathbb{E}[\mathbf{1}_{T=k}|X_\infty| \,|\, \mathcal{F}_k]\big] + \mathbb{E}\big[\mathbf{1}_{T=\infty}|X_\infty|\big]\\
&\overset{\text{cond. exp.}}{=} \sum_{k=0}^{\infty} \mathbb{E}\big[\mathbf{1}_{T=k}|X_\infty| + \mathbf{1}_{T=\infty}|X_\infty|\big]\\
&\overset{\text{MCT}}{=} \mathbb{E}[|X_\infty|] < \infty
\end{aligned}$$

To prove (i), as always, we check that X_T satisfies the defining properties of the conditional expectation. The $\mathcal{F}_T$-measurability of X_T was proved in Proposition 6.1.12. Now let $A \in \mathcal{F}_T$, i.e. $A \cap \{T = n\} \in \mathcal{F}_n$ for all $n \in \mathbb{N}_0$. Then, using the same arguments as above,

$$\begin{aligned}
\mathbb{E}\big[\mathbf{1}_A X_T\big] &\overset{\text{MCT}}{=} \sum_{k=0}^{\infty} \mathbb{E}\big[\mathbf{1}_{A\cap\{T=k\}} \cdot X_k\big] + \mathbb{E}\big[\mathbf{1}_{A\cap\{T=\infty\}} X_\infty\big]\\
&= \sum_{k=0}^{\infty} \mathbb{E}\big[\mathbf{1}_{A\cap\{T=k\}}\mathbb{E}[X_\infty|\mathcal{F}_n]\big] + \mathbb{E}\big[\mathbf{1}_{A\cap\{T=\infty\}} X_\infty\big]\\
&= \sum_{k=0}^{\infty} \mathbb{E}\big[\mathbf{1}_{A\cap\{T=k\}} X_\infty\big] + \mathbb{E}\big[\mathbf{1}_{A\cap\{T=\infty\}} X_\infty\big]\\
&\overset{\text{MCT}}{=} \mathbb{E}\big[\mathbf{1}_A X_\infty\big]
\end{aligned}$$

(ii) follows by taking expectations und using Theorem 6.3.17.

(iii) follows from (i), Proposition 6.1.13, and the tower property of conditional expectation:

$$X_S = \mathbb{E}[X_\infty \,|\, \mathcal{F}_S] \overset{\mathcal{F}_S \subseteq \mathcal{F}_T}{=} \mathbb{E}\big[\mathbb{E}[X_\infty \,|\, \mathcal{F}_T] \,|\, \mathcal{F}_S\big]$$

□

We finish the section with a discussion of the three martingale convergence theorems, a.s.-, L^p-, and L^1-convergence. The L^p-boundedness assumption implies the uniform integrability (in particular, L^1-boundedness) and, using $x^+ \leq |x|$, L^1-boundedness implies the assumption of the almost sure martingale convergence theorem. If possible we will always try to check the L^p-boundedness but keep in mind that the assumption is very strong! Most applications will use the almost sure martingale convergence theorem for non-negative martingales, no assumption needs to be checked, the entire magic of martingales unfolds! The second most important class of applications uses L^2-boundedness as second moments are the best for manipulations (compare the branching process!). For L^p-convergence one most hope for some clever Hölder trick, checking uniform integrability without proving L^p-boundedness is always super hard!

The general discussion can be made more clear in the prime example of martingales, the Galton-Watson branching process:

Example 6.3.20. Recall the definition of the branching process and the martingale M from Example 6.2.3. Since M is non-negative, the almost sure convergence to a finite limit M_∞ comes for free (martingale magic!) and $M_\infty > 0$ corresponds to exponential growth of X. The non-triviality $M_\infty \neq 0$ does not come for free and, in fact, M_∞ is trivial if $\mu \leq 1$. In the supercritical case $\mu > 1$ we used the L^2-martingale convergence theorem to prove that $\mathbb{P}(M_\infty > 0) > 0$ as soon as the offspring distribution has finite second moments. This was possible as the Blackwell-Girshick formula allows to compute second moments. Since L^2-boundedness implies uniform integrability also the properties from Theorem 6.3.17 apply to the branching process. There is no obvious way of how to get rid of the additional second moment assumption using L^p-convergence for $p < 2$ as we have no clue how to simplify a pth power of a sum for $p < 2$. To use the L^1-convergence theorem uniform integrability is needed but we have no clue how to do this. Without even sketching a proof the most famous theorem on branching processes should at least be mentioned:

$$\mathbb{P}(M_\infty > 0) > 0 \quad \Longleftrightarrow \quad \mathbb{E}[\xi_1^1 \log(\xi_1^1)] < \infty,$$

according to the celebrated Kesten-Stigum theorem. There is a counter intuitive point around the Kesten-Stigum theorem. The theorem says that exponential growth is possible for X (with positive probability) if the offspring distribution does not have too much mass at infinity in the sense that $\mathbb{E}[\xi_1^1 \log(\xi_1^1)] < \infty$. This seems counter intuitive as one should believe that more mass as infinity (i.e. more offspring) should give stronger growth, not weaker growth. The reason is the following: If the expectation μ is fixed, then more mass at infinity must be compensated by more mass at 0 to keep the expectation at μ. But more mass at 0 leads to more likely extinction!

6.4 Backward martingales

So far we discussed forwards in time martingales with time indexed by $I = \mathbb{N}_0$. Now we choose $I = -\mathbb{N}_0$, i.e. $\{\cdots, -3, -2, -1, 0\}$. Stochastic processes indexed by $-\mathbb{N}_0$ have been running forever, we call them backwards processes. As an example think of a time series of climate data from the past till the present day. The definition of martingales was initially given for generic ordered sets I but let us quickly recall the definition of a backwards martingale.

Definition 6.4.1. An $(\mathcal{F}_n)_{-\mathbb{N}_0}$ martingale $(X_n)_{n\in-\mathbb{N}_0}$ on $(\Omega, \mathcal{F}, \mathbb{P})$ is called **backwards martingale**. The filtration $(\mathcal{F}_n)_{n\in-\mathbb{N}_0}$ is called a **backwards filtration**.

There is a huge difference between (forward) martingales and backwards martingales. Those are not symmetric concepts as backwards process do not run in the negative time direction. The filtration grows forwards in time and the martingale property also holds forwards in time. In the light of the previous section we get a particularly useful property for backwards martingales, there is always a last element X_0 which closes the entire backwards martingale as $X_n = \mathbb{E}[X_0 \,|\, \mathcal{F}_n]$ for all $n \in -\mathbb{N}_0$. Of course, this should remind us of Doob martingales and, in fact, backwards martingales are as useful as Doob martingales are.

Proposition 6.4.2. Every backward martingale is uniformly integrable.

Proof. Since $X_n = \mathbb{E}[X_0 \,|\, \mathcal{F}_n]$ for all $n \in -\mathbb{N}_0$ and $\mathbb{E}[|X|] < \infty$, this follows from Lemma 6.3.15. □

Since uniform integrability automatically holds, almost sure and L^1 convergence to some $X_{-\infty}$ should always hold! Lecture 10

Theorem 6.4.3. (Backwards martingale convergence theorem) Let $(X_n)_{n\in-\mathbb{N}_0}$ be an $(\mathcal{F}_n)_{n\in-\mathbb{N}_0}$-backwards martingale and let $\mathcal{F}_{-\infty} := \bigcap_{n=0}^\infty \mathcal{F}_{-n}$. Then there is a finite $\mathcal{F}_{-\infty}$-measurable integrable random variable $X_{-\infty}$ with

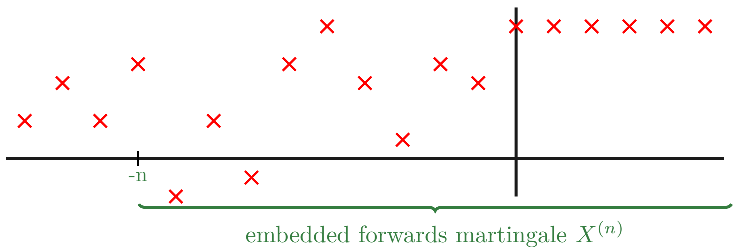
$X_{-\infty} = \mathbb{E}[X_0|\mathcal{F}_{-\infty}]$ and

$$X_n \overset{\text{a.s.}/L^1}{\to} X_{-\infty}, \quad n \to -\infty.$$

Before we give the proof please recall that such a theorem does not come for free for forwards martingales. For forwards martingales we must additionally assume the martingale is uniformly integrable which is a strong assumption!

Proof. The proof follows the same upcrossing idea that we know from the almost sure (forwards) martingale convergence theorem. Since all backwards martingales the L^1-convergences follows for free from Proposition 6.4.2 and Theorem 6.3.16.

For $a < b$ and $n \in -\mathbb{N}_0$ let us define $U_n[a,b]$ to be the number of upcrossings of $X_n, ..., X_0$ through $[a,b]$ and $U[a,b]$ the total number of upcrossings of the backwards martingale through $[a,b]$. Arguing similarly to the proof of Theorem 6.3.1 we need to derive an upper bound of $\mathbb{E}[U_n[a,b]]$ that is independent of n. Here is the trick: we interpret $X_n, ..., X_0$ as the first steps of a (forwards) martingale $X^{(n)}$ that is stopped at time $-n$, compare the picture. Formally, we define



$$X_k^{(n)} := \begin{cases} X_{n+k} & : k \in \{0, \cdots, -n\} \\ X_0 & : k > -n \end{cases}$$

and the filtrations

$$\mathcal{F}_k^{(n)} := \begin{cases} \mathcal{F}_{n+k} & : k \in \{0, \cdots, -n\} \\ \mathcal{F}_0 & : k > -n \end{cases}.$$

All these processes $(X_k^{(n)})_{k\in\mathbb{N}_0}$ are forwards $(\mathcal{F}_k^{(n)})_{k\in\mathbb{N}_0}$-martingales (straight from the definition) that are embedded in our backwards martingale. Since $U_n[a,b] = U_{-n}^{(n)}[a,b]$, we can use the upcrossing inequality of Lemma 6.3.3:

$$\mathbb{E}\big[U_n[a,b]\big] = \mathbb{E}\big[U_{-n}^{(n)}[a,b]\big] \leq \frac{\mathbb{E}\big[(X_{-n}^{(n)} - a)^+\big]}{b-a} = \frac{\mathbb{E}\big[(X_0 - a)^+\big]}{b-a} \leq \frac{\mathbb{E}[|X_0|] + |a|}{b-a}.$$

Hence, as in the proof of Theorem 6.3.1, the expected total number of upcrossings is finite:

$$\mathbb{E}\big[U[a,b]\big] \overset{\text{MCT}}{=} \lim_{n\to-\infty} \mathbb{E}\big[U_n[a,b]\big] < \infty.$$

Also as in the proof of Theorem 6.3.1 almost sure finiteness of the total upcrossing number through all intervals with rational end-points implies the almost sure existence of the limit $X_{-\infty} := \lim_{n\to-\infty} X_n$. Since the backwards martingale $(X_n)_{n\in-\mathbb{N}_0}$ is automatically uniformly integrable, Theorem 6.3.16 implies the L^1-convergence. In particular, $X_{-\infty}$ is almost surely finite.

We have to work a bit for the representation $X_{-\infty} = \mathbb{E}[X_0 \,|\, \mathcal{F}_\infty]$ of the limit. As usually we verify the two defining properties of conditional expectation:

- For the measurability condition we use Proposition 2.1.4. It is enough to prove that $\{X_{-\infty} \in (a,b)\} \in \mathcal{F}_{-\infty}$ as the open intervals generate the Borel-σ-algebra. It follows directly from the definition of convergence that $X_{-\infty} \in (a,b)$ if and only if $X_{-n} \in (c,d)$ for all n large enough. In formulas: For all $k \in -\mathbb{N}_0$

$$\big\{X_{-\infty} \in (c,d)\big\} = \bigcup_{N=k}^{-\infty} \underbrace{\bigcap_{n=N}^{-\infty} \underbrace{\big\{X_n \in (c,d)\big\}}_{\in \mathcal{F}_n}}_{\in \mathcal{F}_N \subseteq \mathcal{F}_k} \in \mathcal{F}_k,$$

  hence, $\big\{X_{-\infty} \in (c,d)\big\} \in \cap_{k=0}^{-\infty} \mathcal{F}_k = \mathcal{F}_{-\infty}$.

- Now let $A \in \mathcal{F}_{-\infty}$. Since $\mathcal{F}_{-\infty} = \cap_{n \in -\mathbb{N}_0} \mathcal{F}_n$ it holds that $A \in \mathcal{F}_n$ for all $n \in -\mathbb{N}_0$, so that the expectation condition of conditional expectation holds:

$$\mathbb{E}[X_\infty \cdot \mathbf{1}_A] = \lim_{n \to -\infty} \mathbb{E}[X_n \cdot \mathbf{1}_A] \overset{X_n = \mathbb{E}[X_0 | \mathcal{F}_n]}{=} \lim_{n \to -\infty} \mathbb{E}[X_0 \cdot \mathbf{1}_A] = \mathbb{E}[X_0 \cdot \mathbf{1}_A].$$

  The first equality holds as the L^1-convergence implies $\mathbb{E}[|X_n - X_\infty|\mathbf{1}_A] \leq \mathbb{E}[|X_n - X_\infty|] \to 0$ so that $|\mathbb{E}[X_n \mathbf{1}_A] - \mathbb{E}[X_\infty \mathbf{1}_A]| \to 0$ by the triangle inequality for expectations.

□

As an application we can derive a complement to Proposition 6.3.18 but now for decreasing σ-algebras.

Proposition 6.4.4. Suppose Z is an integrable random variable and $(\mathcal{G}_n)_{n \in \mathbb{N}_0}$ a decreasing family of σ-algebras, then

$$\mathbb{E}[Z \,|\, \mathcal{G}_n] \overset{\text{a.s}/L^1}{\to} \mathbb{E}[Z \,|\, \mathcal{G}_\infty], \quad n \to \infty,$$

where $\mathcal{G}_\infty = \bigcap_{n=0}^\infty \mathcal{G}_n$.

Proof. For $n \in \mathbb{N}_0$ define $X_{-n} = \mathbb{E}[Z \,|\, \mathcal{G}_n]$ and $\mathcal{F}_{-n} = \mathcal{G}_n$. Then $(X_n)_{n \in -\mathbb{N}_0}$ is an $(\mathcal{F}_n)$-backwards martingale:

- X_n is $\mathcal{F}_n$-measurable for all $n \in -\mathbb{N}_0$ as a conditional expectation,
- $\mathbb{E}[|X_{-n}|] \overset{\Delta}{\leq} \mathbb{E}\big[\mathbb{E}[|Z| \,|\, \mathcal{G}_n]\big] = \mathbb{E}[|Z|] < \infty$,
- $\mathbb{E}[X_{-n} | \mathcal{F}_{-m}] = \mathbb{E}\big[\mathbb{E}[Z \,|\, \mathcal{G}_n] \,\big|\, \mathcal{G}_m\big] \overset{\text{tower prop.}}{=} \mathbb{E}[Z \,|\, \mathcal{G}_m] = X_{-m}$ for all $n \leq m$.

Then Theorem 6.4.3 implies convergence of $X_{-n} = \mathbb{E}[Z | \mathcal{G}_n]$ for $n \to \infty$ almost surely and in L^1 to

$$X_{-\infty} = \mathbb{E}[X_0 | \mathcal{F}_{-\infty}] = \mathbb{E}\big[\mathbb{E}[Z \,|\, \mathcal{G}_0] \,|\, \mathcal{G}_\infty\big] \overset{\text{tower}}{=} \mathbb{E}[Z \,|\, \mathcal{G}_\infty].$$

□

6.5 Application: Proof of the strong law of large numbers

We want to finish the chapter on martingales with a typical application. But what is a typical application of martingales? The only common feature of (almost) all famous applications is their magic connection to martingales for questions that seem completely unrelated to martingales[4]. In that sense, the following proof of the strong law of large numbers is a typical application. Who would guess to derive the law of large number from the backwards martingale convergence theorem?

[4] Check out this article: Yuval Pères: The Unreasonable Effectiveness of Martingales

Definition 6.5.1. For a stochastic process Y on $(\Omega, \mathcal{F}, \mathbb{P})$ one defines the **tail-σ-algebra** $\tau := \bigcap_{n=1}^{\infty} \tau_n$, where $\tau_n := \sigma(Y_n, Y_{n+1}, ...)$.

The tail-σ-algebra is a sub-σ-algebra of the underlying σ-algebra $\mathcal{F}$ (and also of $\mathcal{F}_\infty$) containing precisely the events that which do not depend on finitely many of the Y_n. Typical examples are events describing convergence properties, finiteness of sum, etc. Here is an example to see a typical computation from the context of the Borel-Cantelli Lemma 4.6.4. For all $N \in \mathbb{N}_0$

$$A := \{Y_n \geq \lambda \text{ infinitely often}\} = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} \{Y_k \geq \lambda\} = \bigcap_{n=N}^{\infty} \underbrace{\bigcup_{k=n}^{\infty} \{Y_k \geq \lambda\}}_{\in \tau_n \subseteq \tau_N} \in \tau_N$$

by monotonicity of the inner union in n. But then A is in the tail-σ-algebra τ.

Theorem 6.5.2. (Kolmogorov 0-1 law)
Let $Y_1, Y_2, ...$ be independent random variables. Then τ is trivial, i.e.

$$\mathbb{P}(A) \in \{0, 1\} \quad \text{ for all } A \in \tau.$$

Proof. Take $A \in \tau$ and let $\mathcal{F}_n = \sigma(Y_1, ..., Y_n)$. By the independence assumption $\mathcal{F}_n$ is independent of τ_{n+1} for all $n \in \mathbb{N}_0$, and hence also independent of the sub-σ-algebra τ (recall the definition 4.4.8 of independence of σ-algebras). Using the (forwards) martingale convergence theorem (more precisely Proposition 6.3.18) then yields

$$\mathbb{P}(A) = \mathbb{E}[\mathbf{1}_A] \overset{\text{ind.}}{=} \mathbb{E}[\mathbf{1}_A | \mathcal{F}_n] \overset{n\to\infty}{\longrightarrow} \mathbb{E}[\mathbf{1}_A | \mathcal{F}_\infty] \overset{\tau \subseteq \mathcal{F}_\infty}{=} \mathbf{1}_A \text{ a.s.}$$

This shows $\mathbb{P}(A) = \mathbf{1}_A$ almost surely. The left hand side is a constant and the right hand only takes values 0 and 1, that's it. □

If $X_1, ...$ is an iid sequence, then

$$A := \Big\{\omega \in \Omega : \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} X_k(\omega) \text{ exists}\Big\}$$

is in the tail-σ-algebra as changing the values of finitely many summands does not influence the convergence. Using the Kolmogorov 0-1 law we find that convergence in the strong law of large numbers must happen with probability 0 or 1. Hence, in order to prove the strong law of large numbers it is enough to prove the probability of convergence cannot be zero and identify the limit.

Theorem 6.5.3. (Strong law of large numbers)
Let $X_1, X_2, ...$ be an iid sequcence with $\mathbb{E}[|X_1|] < \infty$, then

$$\frac{1}{n} \sum_{k=1}^{n} X_k \overset{\text{a.s.}}{\longrightarrow} \mathbb{E}[X_1], \quad n \to \infty.$$

Proof. Here is the main trick that relates the strong law to martingales. Let $S_n := \sum_{k=1}^{n} X_k$, then the normalised sum can be written as a backwards martingale:

$$\mathbb{E}\big[X_1 | \sigma(S_n, S_{n+1}, ..)\big] = \frac{S_n}{n}, \quad \text{for all } n \in \mathbb{N}$$

To check the claim we use properties of conditional expectations and a trick:

$$
\begin{aligned}
\mathbb{E}\big[X_1|\sigma(S_n, S_{n+1}, ..)\big] &= \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}\big[X_k \,|\, \sigma(S_n, S_{n+1}, ...)\big] \\
&= \mathbb{E}\Big[\frac{1}{n}\sum_{k=1}^{n} X_k \,\Big|\, \sigma(S_n, S_{n+1}, ...)\Big] \\
&= \mathbb{E}\Big[\frac{S_n}{n} \,\Big|\, \sigma(S_n, S_{n+1}, ...)\Big] = \frac{S_n}{n}.
\end{aligned}
\tag{6.9}
$$

Only the first equality needs extra explanation, but we have already seen the argument in Example 5.2.8. The intuitive reason is the same that was mentioned in the example: Given the values of all sums after n the best guess for the first n summands is the same for each summand as they are iid and equally influence the values of the sums. Writing a formal proof is a bit messy. Using property (x) of Theorem 5.2.4 it suffices to check

$$
\mathbb{E}[X_1 \mathbf{1}_A] = ... = \mathbb{E}[X_n \mathbf{1}_A], \quad \forall A \in \sigma(S_n, ...) \tag{6.10}
$$

and then take the average. To check (6.10) note that the iid assumption gives $\mathbb{E}[F(X_1, ...)] = \mathbb{E}[F(X_{\sigma(1)}, ...)]$ for all finite permutations and all bounded measurable functions. Choosing $A \in \sigma(S_n, ...)$ there is a measurable mapping h such that $\mathbf{1}_A = h(S_n, ...)$. Hence, there is also a measurable mapping g such that $\mathbf{1}_A = g(X_1, ...)$ and g does not change by permuting the first n entries. Using $F(x_1, ...) = x_1 g(x_1, ...)$ and the permutation that only exchanges two integers yields (6.10).

So how do we finish the proof? We first show that $\liminf_{n\to\infty} \frac{S_n}{n}$ is almost surely constant. To see this first note that, for all $\lambda \in \mathbb{R}$,

$$
\Big\{\liminf_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} X_k \le \lambda\Big\} = \Big\{\frac{1}{n}\sum_{k=1}^{n} X_k \le \lambda \text{ i.o.}\Big\} \in \tau.
$$

Thus, by Proposition 2.1.4, the random variable $\liminf_{n\to\infty} \frac{S_n}{n}$ is measurable with respect to the tail-σ-algebra. Now we use the following simple fact:

If $\mathcal{A}$ is a trivial σ-algebra on Ω, then all $\mathcal{A}$-measurable random variables are constant.

Applying Proposition 6.4.4 with $G_n := \sigma(S_n, S_{n+1}, ...)$ to the left hand side of (6.9) shows that the limits of both sides exist almost surely. If the limit of the right hand side exists it equals the limit inferior which, as we have seen above, is constant. Hence, both sides of

$$
\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} X_k = \mathbb{E}\big[X_1|\mathcal{G}_\infty\big]
$$

are constant. But then Theorem 5.2.4 (ix) implies that the right hand side equals $\mathbb{E}[X_1]$ almost surely. That's it! □

Here is a question: Do we really need the iid assumption in the proof of the strong law of large numbers? Actually, not quite. Ommitting the use of Kolmogorov's 0-1 law gives the law of large numbers for so-called exchangeable sequences (finite permutations do not change the distribution of the sequence). In that case the proof only gives the almost sure onvergence of the normalised sum towards the (non-constant) random variable $\mathbb{E}[X_1|\mathcal{G}_\infty]$. This, and more on exchangeable sequences can be found in a beautiful article of Kingman [5].

[5] J. F. C. Kingman, "Uses of Exchangeability", Annals of Probability, 1978, Vol. 6, pp. 183-197

Kapitel 7

Convergence of Measures

Lecture 11

In this chapter the technical tools for proving Donsker's famous invariance principle will be developed. This main result of Chapter 8 will be to prove weak convergence of scaled random walks with second moment jumps towards the Brownian motion. But what does weak convergence of a sequence of stochastic processes mean? The convergence will be defined to be weak convergence of the processes interpreted as path-valued random variables. We already met the notion of weak convergence of real-valued random variables in Section **??**, in this chapter weak convergence will be generalised to much more general state spaces. To do so, a bit of Functional Analysis needs to be combined with measure theory and a bit of probability. Along the way we will also prove a couple of theorems on the characterisation of random variables through moments.

7.1 A bit of topology, measure, and integration theory

To get started let us recall from analysis some topological concepts that will lead us naturally to general Borel-σ-algebras. We will always work with metric or normed spaces E, metrics will typically be denoted by d and norms $||\cdot||$. A central object of basic topology are open and closed sets:

Definition 7.1.1. Suppose (E,d) is a metric space and

$$B_\varepsilon(x) := \{y \in E : d(x,y) < \varepsilon\}, \quad \varepsilon > 0,$$

are the ε-balls around x.

- $A \subseteq E$ is called **open** if for all $x \in A$ there is some $\varepsilon > 0$ such that $B_\varepsilon(x) \subseteq A$.
- $A \subseteq E$ is called **closed** if A^c is open.

The set of all open sets is also the **topology** of E and denoted by τ. Any open set containing a ball $B_\varepsilon(x)$ is called an (open) **neighbourhood** of x.

- $A \subseteq E$ is called **bounded** if there is some $r > 0$ such that $d(x,y) \le r$ for all $x, y \in A$.

To study various properties of subsets of metric spaces very quickly one needs to think about more fine properties. Since there are many equivalent ways of redefining the following definitions it is likely that you might have seen different notions in your basic analysis lectures.

Definition 7.1.2. Suppose (E,d) is a metric space and $A \subseteq E$.

- $x \in A$ is called an **inner point** if there is $\varepsilon > 0$ such that $B_\varepsilon(x) \subseteq A$.

- The set of inner points of A is denoted by $\dot{A}$ and is called **interior of A**.
- The **closure** $\bar{A}$ is the smallest closed set containing A (the intersection of all closed sets containing A).
- $\partial A := \bar{A} \backslash \dot{A}$ is called the **boundary** of A.
- A is called **dense** in E if $\bar{A} = \bar{E}$.

Please keep in mind the important facts that arbitrary unions of open sets are open and finite intersections of open sets are open. The empty set and the entire space are both open and closed.

Definition 7.1.3. Suppose (E, d) is a metric space.

- A sequence $(x_n)_{n\in\mathbb{N}} \subseteq E$ **converges** in E towards x if $\lim_{n\to\infty} d(x_n, x) = 0$, i.e. for every ε there is some $N \in \mathbb{N}$ such that $d(x_n, x) < \varepsilon$ for all $n \geq N$.
- A sequence $(x_n)_{n\in\mathbb{N}} \subseteq E$ is called a **Cauchy-sequence** in E if for all $\varepsilon > 0$ there is some $N \in \mathbb{N}$ such that $d(x_n, x_m) < \varepsilon$ for all $n, m \geq N$.
- If all Cauchy-sequences in E converge, then E is called **complete**.

This section is mostly about convergence so let us recall the important connection of closeness and convergence. A set A is closed if and only if all converging sequences $(x_n) \subseteq A$ converge to some $x \in A$. Formulated differently, $\bar{A}$ consists precisely of those elements which are the limits of sequences in A. Just check yourself some examples for intervals in $\mathbb{R}$!

In the discussion of conditional expactations $\mathbb{E}[X|Y]$ we have strongly used that $\mathbb{R}$ contains the countable dense subset $\mathbb{Q}$. Once we combine general metric spaces with measures theory it won't come as a surprise (think of the definition of a measure and σ-algebras) that the existence of countable dense subsets should be useful.

Definition 7.1.4. A metric space (E, d) is called **separable** if there is a countable and dense subset of E.

Here is a small but useful fact from topology. A metric space (E, d) is separable if and only if the topology τ has a countable base $\mathcal{B}$. A base $\mathcal{B}$ is a subset of open sets such that all open sets can be written as a union of sets from the base. The concept of a base is a bit similar to a generator of a σ-algebra but somewhat simpler as we are only allowed to take countable unions of events in σ-algebras but arbitrary unions of open sets for a base.

The set of all intervals with rational end-points is a countable base of the topology of $\mathbb{R}$ induced by the usual norm.

There is a special class of metric spaces that turned out to be most useful in Functional Analysis and probability theory. This is the typical setting in which the general theory of Markov processes can be developed.

Definition 7.1.5. A **Polish metric space** is a complete and separable metric space.

The word Polish in the definition honours the school of Polish mathematicians from the 20th century that was responsible for the development of almost all tools of Functional Analysis. There are a few prime examples of Polish metric spaces that we will encounter again and again:

$$(\mathbb{R}, |\cdot|), \quad (\mathbb{R}^d, |\cdot|), \quad (\mathbb{C}, |\cdot|), \quad \text{and} \quad (C([0,1]), ||\cdot||_\infty),$$

where $C([0,1])$ are the continuous real-valued functions on $[0,1]$.

Definition 7.1.6. Suppose (E,d) is a metric space and $A \subseteq E$.

- A is called **compact** if every covering of A by open sets has a finite subcovering, A is called **sequentially compact** if every sequence in A has a subsequence that converges to a limit in A.
- A is called **relatively compact** if $\bar{A}$ is compact, A is called **relatively sequentially compact** if all sequences in A have a subsequence with limit in $\bar{A}$.

Different characterisations of compactness have been discussed for $(\mathbb{R}^d, |\cdot|)$ in analysis, in particular, compactness and sequential compactness are equivalent and compact sets are precisely the closed and bounded sets (Heine-Borel Theorem). The Heine-Borel equivalence fails in most other metric spaces but we still know from analysis (check it!) that

$$A \text{ is (relatively) compact} \quad \Longleftrightarrow \quad A \text{ is (relatively) sequentially compact},$$

holds in all metric spaces. In complete metric spaces a more general version of Heine-Borel can be formulated using the concept of total boundedness:

Definition 7.1.7. A set $A \subseteq E$ is called **totally bounded** if for all $\varepsilon > 0$ there are finitely many points $x_1, ..., x_n \in A$ with $A \subseteq \bigcup_{k=1}^{n} B_\varepsilon(x_k)$.

Using the triangle inequality it is easy to see that totally bounded sets are always bounded, i.e. are covered by a large single ball. It is not generally the case that bounded sets are also totally bounded.

Check that boundedness and totally boundedness are equivalent in $(\mathbb{R}^d, |\cdot|)$ but are not equivalent in all infinite sets with the discrete metric (i.e. $d(x,y) = \frac{1}{2}$ for all $x \neq y$).

We will see later that totally-bounded and bounded are not the same in $(C([0,1], ||\cdot||_\infty)$.[1]

Proposition 7.1.8. Suppose (E,d) is a complete metric space and $A \subseteq E$, then

$$A \text{ is compact} \quad \Longleftrightarrow \quad A \text{ is closed and totally bounded.}$$

Proof. "$\Rightarrow$": A totally bounded follows from the compactness definition by covering A with ε-balls around all elements, A closed follows from sequential compactness since in metric spaces all subsequences converge to the same limit as their convergent sequences.

"$\Leftarrow$": Take a sequence $(x_n) \subseteq A$. For each $m \in \mathbb{N}$ take a finite covering $B_{\frac{1}{m}}(y_1^m), ...$ of A. By the finiteness there must be a subsequence which lies eventually in one of the balls $B_1(y_k^1)$. Similarly, from this subsequence we extract a further subsequence which lies eventually in one of the $B_{\frac{1}{2}}(y_k^2)$. A diagonal argument gives a subsequence with $x_n \subseteq B_{\frac{1}{n}}(y_k^n)$. This is Cauchy and converges by completeness. Hence, A is sequentially compact. ☐

It is not too hard to come up with counter examples to the second equivalence in non-complete spaces. For instance, take the normed space $(\mathbb{Q}, |\cdot|)$ and $A = (-\sqrt{2}, \sqrt{2}) \cap \mathbb{Q}$. Then A is closed, totally bounded but not compact.

[1] nicht vergessen

As always in mathematics we are interested in the natural mappings between objects, mappings that respect the structure of the objects. For σ-algebra these were the measurable mappings, for metric spaces these are the continuous mappings:

Definition 7.1.9. A mapping between two metric spaces (E, d_E) and (F, d_F) is called **continuous** if preimages of all open sets in F are open in E. We use the notation

- $C(E, F) := \{f : E \to F \,|\, f \text{ continuous}\}$
- $C(E) := \{f : E \to \mathbb{R} \,|\, f \text{ continuous}\}$
- $C_b(E) := \{f : E \to \mathbb{R} \,|\, f \text{ continuous, bounded}\}$
- $C_c(E) := \{f : E \to \mathbb{R} \,|\, f \text{ continuous, compact support}\}$

A compactly supported function is defined as in basic analysis as a function such that $\{x \in E : f(x) = 0\}$ is contained in a compact set.

Recall that there are other equivalent ways of defining continuity via ε-δ formalism

$$\forall \varepsilon > 0 \exists \delta > 0 : d_E(x, y) < \delta \;\Rightarrow\; d_F(f(x), f(y)) < \varepsilon$$

and sequences

$$x_n \to x,\, n \to \infty \quad \Longrightarrow \quad f(x_n) \to f(x),\, n \to \infty.$$

We will always use the most convenient formulation. Here is a not so simple exercise:

$(C_b([0, \infty)), ||\cdot||_\infty)$ is not separable but $(C_c([0, \infty)), ||\cdot||_\infty)$ is separable.

Now we come to the fun part, extending the concepts from measure theory on the Borel-σ-algebra of $\mathbb{R}$ to general metric spaces. Recall that $\mathcal{B}(\mathbb{R})$ was defined to be the smallest σ-algebra containing all open sets (or closed sets, or intervals, etc.), hence, it is quite clear how we should proceed for general metric spaces.

Definition 7.1.10. For a metric space (E, d) we call

$$\mathcal{B}(E) = \sigma(\{O \subseteq E : O \text{ open}\})$$

the **Borel-σ-algebra** on E.

Of course, since σ-algebras are closed under taking complements $\mathcal{B}(E)$ is also generated by all closed sets. It is also instructive to check the following exercise to link topology and measure theory. The proof is precisely the same that shows that $\mathcal{B}(\mathbb{R})$ is also generated by all intervals.

If (E, d) is separable, then $\mathcal{B}(E) = \sigma(\{B_\varepsilon(x) : x \in E, \varepsilon > 0\})$.

Measurable mappings between two metric spaces will always be with respect to the corresponding Borel-σ-algebras. Not surprisingly we will call such measurable mappings **Borel-measurable**. From the definition of continuity and Proposition 2.1.4 it is clear that all continuous functions between metric spaces are Borel-measurable.

Definition 7.1.11. Let (E, d) be a metric space, then

- $\mathcal{M}_f(E) := \mathcal{M}_f := \{\mu : \mu \text{ is a finite measures on } \mathcal{B}(E)\}$

- $\mathcal{M}_1(E) := \mathcal{M}_1 := \{\mu : \mu$ is a probability measures on $\mathcal{B}(E)\}$
- $\mathcal{M}_{\leq 1}(E) := \mathcal{M}_{\leq 1} := \{\mu : \mu$ is a sub-probability measures one $\mathcal{B}(E)\}$

After all these definition let us prove a first proposition on measures on Polish spaces. Before checking the proof have a quick thought how you would prove the statement on $\mathcal{B}(\mathbb{R})$ using continuity of measures with $(-n, n)$ and you will immediately appreciate a useful property of $\mathbb{R}$, namely, to be able to fill $\mathbb{R}$ from the inside by increasing intervals.

Proposition 7.1.12. Suppose (E, d) is Polish and $\mu \in \mathcal{M}_f$. Then, for all $\varepsilon > 0$, there is a compact set K with $\mu(K^c) < \varepsilon$.

Proof. The trick is to replace the increasing intervals $(-n, n)$ in $\mathbb{R}$ by the right substitute and then try to argue with continuity of measures. Let $x_1, x_2, ...$ the countable dense subset of E and $n \in \mathbb{N}$. Then $E = \bigcup_{k=1}^{\infty} B_{\frac{1}{n}}(x_k)$ for all $n \in \mathbb{N}$. Using continuity of measures (this needs $\mu \in \mathcal{M}_f$, compare Theorem 1.1.14) fix $N_n \in \mathbb{N}$ with

$$\mu\Big(E \backslash \bigcup_{k=1}^{N_n} B_{\frac{1}{n}}(x_k)\Big) < \frac{\varepsilon}{2^n}$$

Now define $A := \bigcap_{n=1}^{\infty} \bigcup_{k=1}^{N_n} B_{\frac{1}{n}}(x_k) \in \mathcal{B}(E)$. A is totally bounded as $A \subseteq \bigcup_{k=1}^{N_n} B_{\frac{1}{n}}(x_k)$ for all $n \in \mathbb{N}$, hence, $\bar{A}$ is compact as E is complete. If we choose $K := \bar{A}$, then

$$\mu(K^c) = \mu(E \backslash K) \overset{\text{mon.}}{\leq} \mu(E \backslash A) = \mu\Big(\bigcup_{n=1}^{\infty} \bigcap_{k=1}^{N_n} B^c_{\frac{1}{n}}(x_k)\Big).$$

Using sub-additivity we can continue the chain of inequalities with

$$\sum_{n=1}^{\infty} \mu\Big(\bigcap_{k=1}^{N_n} B^c_{\frac{1}{n}}(x_k)\Big) = \sum_{n=1}^{\infty} \mu\Big(E \backslash \bigcup_{k=1}^{N_n} B_{\frac{1}{n}}(x_k)\Big) \leq \sum_{n=1}^{\infty} \frac{\varepsilon}{2^n} = \varepsilon$$

which finishes the proof. □

We can now turn towards a crucial topic of this chapter. Is it possible to characterise measures using only integrals over certain functions? [2]

Definition 7.1.13. Let (E, d) a metric space and $F \subseteq \mathcal{M}_f(E)$ a family of measures. A family C of measurable mappings $E \to \mathbb{R}$ is called **separating family for** F if for all $\mu, \nu \in F$

$$\int_E f \, \mathrm{d}\mu = \int_E f \, \mathrm{d}\nu \quad \forall f \in C \cap L^1(\mu) \cap L^1(\nu) \;\Rightarrow\; \mu = \nu$$

The most simplistic (and least useful) family is the family of all measurable indicator functions $C := \{\mathbf{1}_A : A \in \mathcal{B}(E)\}$ which trivially separates all families of measures as $f = \mathbf{1}_A$ yields

$$\nu(A) = \int_E f \, \mathrm{d}\nu = \int_E f \, \mathrm{d}\mu = \mu(A).$$

Since the set of all indicators on measurable set is equally big as the Borel-σ-algebra there is a big desire to finde more approachable sets such as all exponential functions or all polynomial

[2] brauchen wir wirklich $\mathcal{M}_f$?

functions. To understand the background of separating families in probability let us recall Theorem 4.3.17, which in Stochastik 1 we stated without a proof:

$$\int_{\mathbb{R}} e^{tx}\,\mathrm{d}\mathbb{P}_X(x) = M_X(t) = M_Y(t) = \int_{\mathbb{R}} e^{tx}\,\mathrm{d}\mathbb{P}_Y(x), \quad \forall t \in [-\varepsilon, \varepsilon] \quad \Longrightarrow \quad \mathbb{P}_X = \mathbb{P}_Y.$$

If we reformulate, Theorem 4.3.17 states that exponential functions are separating for probability measures with exponential moments. A more precise statement will be proved towards the end of this chapter.

During the course of this chapter we will get to know which families of measurable mappings are separating for different classes of measures. We start with a first smaller step and prove that Lipschitz continuous functions are separating for finite measures.

Definition 7.1.14. Let (E, d_E) and (F, d_F) be metric spaces.

- $f\colon E \to F$ is called Lipschitz continuous (with constant K) if

$$d_F\big(f(x), f(y)\big) \leq K \cdot d_E(x,y) \quad \forall x, y \in E$$

- $\mathrm{Lip}_K(E,F) := \{f\colon E \to F \,|\, \text{Lipschitz with constant } K\}$
- $\mathrm{Lip}(E,F) := \{f\colon E \to F \,|\, \text{Lipschitz}\}$
- $\mathrm{Lip}_K(E) := \mathrm{Lip}_K(E, \mathbb{R})$
- $\mathrm{Lip}(E) := \mathrm{Lip}(E, \mathbb{R})$

As promised the Lipschitz continuous functions are separating the finite measures on E. In fact, the same proof also shows that bounded and compactly supported continuous functions separate $\mathcal{M}_f(E)$.

Proposition 7.1.15. In metric spaces $\mathrm{Lip}_1(E)$, $C_b(E)$, and $C_c(E)$ are separating for $\mathcal{M}_f(E)$.

Proof. First of all, note that $f \in \mathrm{Lip}_K(E)$ implies $\frac{1}{K} f \in \mathrm{Lip}_1(E)$. Hence, if we assume $\int_E f\,\mathrm{d}\mu = \int_E f\,\mathrm{d}\nu$ for all $f \in \mathrm{Lip}_1(E)$ we can also use

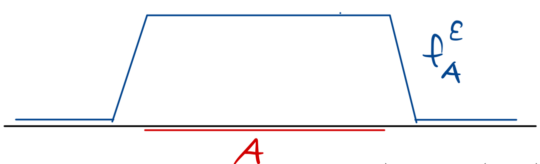
$$\int_E f\,\mathrm{d}\mu = \int_E f\,\mathrm{d}\nu, \quad \forall f \in \mathrm{Lip}(E), \tag{7.1}$$

whenever the integrals are finite. We will show that (7.1) implies $\nu(A) = \mu(A)$ for all A closed. Since the closed sets form an $\cap$-stable generator of $\mathcal{B}(E)$ we can then deduce $\mu = \nu$ from Theorem 1.2.12 (this uses the finiteness of μ and ν).

We now fix $\mu, \nu \in \mathcal{M}_f(E)$ and $A \subseteq E$ closed. Then define the functions

$$f_A^\varepsilon(x) := \Big(1 - \frac{1}{\varepsilon} \cdot d(x, A)\Big)^+, \quad x \in E, \varepsilon > 0,$$

with $d(x,A) = \inf\{d(x,y) : y \in A\}$. Remember from basic analysis (or quickly check yourself by



using that $\bar{A}$ are precisely the limits of sequences in A) that $d(x,A) = 0$ if and only if $x \in \bar{A}$. The functions f_A^ε have the following properties:

- $f_A^\varepsilon = 1$ on A
- $f_A^\varepsilon \to \mathbf{1}_A$, because for closed sets $d(x, A) = 0$ if and only if $x \in A$,
- $f_A^\varepsilon \leq 1$

Since the idea of using the functions f_A^ε is (almost) the entire point of the proof it is useful to check the following yourself:

$f_A^\varepsilon \in \mathrm{Lip}_{\frac{1}{\varepsilon}}(E)$ and, obviously, $f_A^\varepsilon \in C_b(E)$, $f_A^\varepsilon \in C_c(E)$.

We can now use (7.1) to prove that μ and ν coincide on the closed set A:

$$\mu(A) = \int_E \mathbf{1}_A \,\mathrm{d}\mu \overset{\text{DCT}}{=} \lim_{\varepsilon \to \infty} \int_E f_A^\varepsilon \,\mathrm{d}\mu \overset{\text{ass.}}{=} \lim_{\varepsilon \to \infty} \int_E f_A^\varepsilon \,\mathrm{d}\nu \overset{\text{DCT}}{=} \int_E \mathbf{1}_A \,\mathrm{d}\nu = \nu(A).$$

As explained above we proved that (7.1) implies $\nu = \mu$, or, in other words, that $\mathrm{Lip}_1(E)$ is separating for $\mathcal{M}_f(E)$. The claims about $C_b(E)$ and $C_c(E)$ follow in exactly the same way as $f_A^\varepsilon \in C_b(E)$ and $f_A^\varepsilon \in C_c(E)$. □

More explicit classes of functions appear later in the course of this chapter.

7.2 Weak convergence of measures - the basics

As announced the ultimate goal is to prove the convergence of scaled random walks towards the Brownian motion, both seen as function-valued random variables. Let us recall from Definition **??** the notion of weak convergence of real-valued random variables

$$\int_{\mathbb{R}} f \,\mathrm{d}\mathbb{P}_{X_n} = \mathbb{E}[f(X_n)] \overset{n\to\infty}{\to} \mathbb{E}[f(X)] = \int_{\mathbb{R}} f \,\mathrm{d}\mathbb{P}_X, \quad n \to \infty, \tag{7.2}$$

which is actually a notion of convergence for the sequence of probability measures $(\mathbb{P}_{X_n})_{n\in\mathbb{N}}$ on $\mathcal{B}(\mathbb{R})$. In order define a notion of convergence of stochastic processes we will introduce the law of a stochastic process (or a random variable with general state-space) and then introduce a general notion of weak convergence.

Definition 7.2.1. Let X be a random variable on $(\Omega, \mathcal{A}, \mathbb{P})$ with values in a metric space (E, d), then **the law** $\mathbb{P}_X$ **of** X is the probability measure

$$\mathbb{P}_X(A) := \mathbb{P}\big(X^{-1}(A)\big) = \mathbb{P}\big(\{\omega \in \Omega \colon X(\omega) \in A\}\big), \quad A \in \mathcal{B}(E).$$

As for real-valued random variables we also use the notion $\mathbb{P}\big(X \in A\big)$ as this can be read more naturally as "probability of X in A".

For the study of stochastic processes we will always keep in mind the example $E = C([0, 1])$ or $E = C([0, \infty))$ which are the state-spaces of stochastic processes indexed by $[0, 1)$ or $[0, \infty)$, respectively, reinterpreted as function-valued random variables.

In order to speak of convergence of processes we will thus have to define a notion of convergence on measures on more general state-spaces than $\mathbb{R}$. This leads us to the general notion of weak convergence of measures on metric spaces:

Definition 7.2.2. Let (E, d) a metric space and $\mu, \mu_1, \mu_2, ... \in \mathcal{M}_f(E)$. We say $(\mu_n)_{n\in\mathbb{N}}$ **converges weakly to** μ if

$$\lim_{n\to\infty} \int_E f \,\mathrm{d}\mu_n = \int_E f \,\mathrm{d}\mu, \quad \text{for all } f \in C_b(E),$$

and write $\mu_n \xrightarrow{(w)} \mu$ or $\mu_n \Rightarrow \mu$ or $\mu = \text{w-}\lim_{n\to\infty} \mu_n$.

Before returning to probability theory let us discuss important properties that help us to link the general weak convergence theory (a field of Functional Analysis) to properties from Section **??** in the particular case $E = \mathbb{R}$. There are many other ways of defining convergence of measures through distances. They are usually stronger (less sequences converge) which is one of the reasons to speak of weak convergence.

Let $(\delta_{x_n})_{n\in\mathbb{N}}$ a sequence of Dirac-measures on E such that $x_n \to x$ for $n \to \infty$. Show that $(\delta_{x_n})_{n\in\mathbb{N}}$ converges weakly in $\mathcal{M}_1(E)$ to δ_x.

If you are familiar with Functional Analysis a bit of care is needed as the wording does not match. In the standard terminology of Functional Analysis this is not weak convergence but weak-$*$-convergence on $\mathcal{M}_f(E)$ using that $\mathcal{M}_f(E)$ is the dual-space of $C_b(E)$.

Remark 7.2.3. If (E, d) is separable then weak convergence in $\mathcal{M}_f(E)$ can be metrized: Defining the **Prohorov metric** (here the definition for $\nu, \mu \in \mathcal{M}_1(E)$)

$$d_p(\mu, \nu) := \inf \big\{\varepsilon > 0 \colon \mu(B) \le \nu(B_\varepsilon(0)) + \varepsilon \;\forall B \in \mathcal{B}(E)\big\}$$

one can show that d_p is a metric on $\mathcal{M}_f(E)$ and

$$\mu_n \xrightarrow{(w)} \mu,\; n \to \infty \quad \Leftrightarrow \quad d_p(\mu_n, \mu) \to 0,\; n \to \infty.$$

This is important as all properties of convergence in metric spaces (such as uniqueness of limits) hold for weak convergence. Taking $f \equiv 1$ shows that also the total masses must converge. In particular, the weak limit of a sequence of probability measures is a probability measure, no mass gets lost or appears. In other words of topology, $\mathcal{M}_1(E)$ is a closed subset of $\mathcal{M}_f(E)$ with respect to the Prohorov metric.

Lecture 12

We approach weak convergence of measures in two steps. First, we will derive some equivalent conditions, usually referred to Portemantau theorem, via elementary (but tedious) manipulations with measures and integrals. In the next section we start to understand weak convergence from the point of view of convergence in the metric space $(\mathcal{M}_f, d_P)$ with tools from metric space theory. The metric space approach is necessary in order to derive handy criteria that depend strongly on the corresponding underlying space (E, d).

Here is the basic Portemanteau theorem:

Theorem 7.2.4. (Portemanteau theorem)
Let (E, d) a metric space and $\mu, \mu_1, \mu_2, ... \in \mathcal{M}_1(E)$, then the following are equivalent:

(i) $\mu_n \xrightarrow{(w)} \mu,\; n \to \infty$

(ii) $\lim_{n\to\infty} \int_E f \,\mathrm{d}\mu_n = \int_E f \,\mathrm{d}\mu \quad \forall f$ bounded, Lipschitz continuous

(iii) $\limsup\limits_{n\to\infty} \mu_n(F) \le \mu(F) \quad \forall$ closed $F \subseteq E$

(iv) $\liminf\limits_{n\to\infty} \mu_n(G) \ge \mu(F) \quad \forall$ open $G \subseteq E$

(v) $\lim\limits_{n\to\infty} \mu_n(A) = \mu(A) \quad \forall A$ with $\mu(\partial A) = 0$

To understand similarities it is instructive to recall the statment and the proof of Theorem 4.5.9

for a sequence $\mu_n = \mathbb{P}_{X_n}$ of probability measures on $\mathcal{B}(\mathbb{R})$. In that simpler setting property (v) holds with the closed sets $A = (-\infty, t]$ because $\mathbb{P}_X(\{t\}) = 0$ is equivalent to t being a point of continuity of the distribution function F_X. In the next section we will derive a much more useful statement if (E, d) is even a Polish metric space:

(vi) tightness + $\lim_{n\to\infty} \int_E f \,\mathrm{d}\mu_n = \int_E f \,\mathrm{d}\mu$ for some separating family $C \subseteq C_b(E)$ of $\mathcal{M}_1(E)$

More concrete criteria for special cases such as $E = [a, b]$, $E = [0, \infty)$, $E = \mathbb{R}^d$, or $E = C([0, 1])$ will follow below.

Proof. (i) $\Rightarrow$ (ii): trivial (Lipschitz is continuous)

(ii) $\Rightarrow$ (iii): Let F be closed and f_F^ε from the proof of Proposition 7.1.15. Then

$$\limsup_{n\to\infty} \mu_n(F) \overset{\mathbf{1}_F \le f_F^\varepsilon}{\le} \limsup_{n\to\infty} \int_E f_F^\varepsilon \,\mathrm{d}\mu_n, \quad \forall \varepsilon > 0,$$

so that

$$\begin{aligned}
\limsup_{n\to\infty} \mu_n(F) &\le \inf_{\varepsilon>0} \limsup_{n\to\infty} \int_E f_F^\varepsilon \,\mathrm{d}\mu_n \\
&\overset{\text{Limit exists}}{=} \inf_{\varepsilon>0} \lim_{n\to\infty} \int_E f_F^\varepsilon \,\mathrm{d}\mu_n \\
&= \inf_{\varepsilon>0} \int_E f_F^\varepsilon \,\mathrm{d}\mu \overset{\text{DCT}}{=} \mu(F).
\end{aligned}$$

(iii) $\Leftrightarrow$ (iv): This follows by taking complements as F closed $\Leftrightarrow$ $G = F^c$ open and $\mu_n(F) = 1 - \mu_n(F^c)$ and $\liminf_{n\to\infty}(-a_n) = -\limsup_{n\to\infty}(a_n)$.

(iii) + (iv) $\Rightarrow$ (v): Let $A \in \mathcal{B}(E)$ with $\mu(\partial A) = 0$. First note that

$$\limsup_{n\to\infty} \mu_n(A) \overset{\text{monot.}}{\le} \limsup_{n\to\infty} \mu_n(\bar{A}) \overset{\text{(iii)}}{\le} \mu(\bar{A})$$

and

$$\liminf_{n\to\infty} \mu_n(A) \overset{\text{monot.}}{\ge} \liminf_{n\to\infty}(\dot{A}) \overset{\text{(iv)}}{\ge} \mu(\dot{A})$$

Since $\overline{A} = \dot{A} \cup \partial A$ we have $\mu(\dot{A}) + \mu(\partial A) \overset{\text{ass.}}{=} \mu(\bar{A})$ so that $\mu(A) = \mu(\bar{A}) = \mu(\dot{A})$ so that the above yields

$$\limsup_{n\to\infty} \mu_n(A) \le \mu(A) \le \liminf_{n\to\infty} \mu_n(A)$$

which implies that the limit exists and is equal to $\mu(A)$.

(v) $\Rightarrow$ (iii): Let $F \subseteq E$ closed and enlarge F by δ: $F^\delta := \{x \in E : d(x, F) \le \delta\}$. Since $\partial F^\delta \subseteq \{x \in E : d(x, E) = \delta\}$ the sets ∂F^δ are disjoint for different δ. Now we use that for a probability measure there cannot be uncountably many disjoint events with positive probability. So there must be a sequence $\delta_k \to 0$ along which $\mu(\partial F^{\delta_k}) = 0$. Hence, we can use (v) for F^{δ_k} for all $k \in \mathbb{N}$. Thus,

$$\limsup_{n\to\infty} \mu_n(F) \overset{\text{monot.}}{\le} \limsup_{n\to\infty} \mu_n(F^{\delta_k}) \overset{\text{(v)}}{=} \lim_{n\to\infty} \mu_n(F^{\delta_k}) \overset{\text{(v)}}{=} \mu(F^{\delta_k}), \quad k \in \mathbb{N},$$

where we used that limit and limit superior coincide if a limit exists. Now we take limits in k on both sides. Since $F^{\delta_k} \downarrow F$ and the left hand side is independent of k, we obtain form monotonicity of measures the desired inequality $\limsup_{n\to\infty} \mu_n(F) \le \mu(F)$.

(iii) ⇒ (i): Let $f \in C_b(E)$. Without loss of generality it can be assumed that $0 < f < 1$. If not, we choose $a > 0$ and $b \in \mathbb{R}$ so that $\bar{f} := a \cdot f + b \in [0,1)$ and the argument is continued using $\bar{f}$. Now define the closed sets $F_i^{(k)} := \{x \in E \colon f(x) \geq \frac{i}{k}\}$ for $k \in \mathbb{N}$, $0 \leq i \leq k$. For $\mu \in \mathcal{M}_1(E)$ monotonicity of integrals and the definition of the integral for simple functions yields

$$\sum_{i=1}^{k} \frac{i-1}{k}\Big(\mu\big(F_{i-1}^{(k)}\big) - \mu\big(F_i^{(k)}\big)\Big) \leq \int_E f \,\mathrm{d}\mu \leq \sum_{i=1}^{k} \frac{i}{k}\Big(\mu\big(F_{i-1}^{(k)}\big) - \mu\big(F_i^{(k)}\big)\Big).$$

Warning: The inequalities look a bit strange. Typically one would partition the image space $[0,1)$ into the disjoint sets $E_i^{(k)} = \{x \in E \colon \frac{i+1}{k} > f(x) \geq \frac{i}{k}\}$ and work with the inequalities $\sum \frac{i-1}{k}\mu(E_{i-1}^{(k)}) \leq \int_E f \,\mathrm{d}\mu \leq \sum \frac{i}{k}\mu(E_{i-1}^{(k)})$, but those sets are not closed. If we use the closed sets $F_i^{(k)}$, then we double count and need to subtract the pieces counted twice. Since $F_k^{(k)} = \emptyset$ and writing out the summand this simplifies to

$$\frac{1}{k}\sum_{i=1}^{k-1} \mu(F_i^{(k)}) \leq \int_E f \,\mathrm{d}\mu \leq \frac{1}{k}\sum_{i=1}^{k} \mu(F_{i-1}^{(k)}).$$

This looks more complicated than it is and only comes from looking closely to see the simplifications $\frac{i}{k}\mu(F_{i-1}^{(k)}) - \frac{i-1}{k}\mu(F_{i-1}^{(k)}) = \frac{1}{k}\mu(F_{i-1}^{(k)})$ which cancel half of the summands. Note that the estimates also hold for integrals with μ_n instead of μ, the argument was only based on splitting the function f. Using (iii) for the closed sets then gives

$$\begin{aligned}
\limsup_{n\to\infty} \int_E f \,\mathrm{d}\mu_n &\leq \limsup_{n\to\infty} \frac{1}{k}\sum_{i=1}^{k} \mu_n(F_{i-1}^{(k)}) \\
&\leq \frac{1}{k}\sum_{i=1}^{k} \limsup_{n\to\infty} \mu_n(F_{i-1}^{(k)}) \\
&\overset{\text{(iii)}}{\leq} \frac{1}{k}\sum_{i=1}^{k} \mu(F_{i-1}^{(k)}) \\
&\overset{\mu\leq 1}{\leq} \frac{1}{k} + \frac{1}{k}\sum_{i=1}^{k-1} \mu(F_{i-1}^{(k)}) \\
&\leq \frac{1}{k} + \int_E f \,\mathrm{d}\mu.
\end{aligned}$$

Since k was arbitrary, we obtain $\limsup\limits_{n\to\infty} \int_E f \,\mathrm{d}\mu_n \leq \int_E f \,\mathrm{d}\mu$. The same reasoning for $-f$ gives $\liminf\limits_{n\to\infty} \int_E f \,\mathrm{d}\mu_n \geq \int_E f \,\mathrm{d}\mu$. Both inequalities prove (i). □

For the next theorem recall from Section **??** the notion of the push-forward of a measure under a measurable map. If $g : \Omega \to \Omega'$ is $(\mathcal{A}, \mathcal{A}')$-measurable and μ is a measures on $\mathcal{A}$, then we defined

$$\mu_g(A) := \mu(g^{-1}(A)), \quad A \in \mathcal{A}',$$

also written $\mu \circ g$ (more useful if there is a further index), is a measure on $\mathcal{A}'$. The measure is called the push-forward of μ under g or the image measure. If we think of converging sequences of measures it is natural to also ask for convergence of the sequence of push-forwards for a given mapping. If g is continuous, this can be proved easily:

Theorem 7.2.5. (continuous mapping theorem)
Let (E_1, d_1), (E_2, d_2) be metric spaces and $g \colon E_1 \to E_2$ continuous. If $\mu, \mu_1, \mu_2, \ldots$

are finite measures on $\mathcal{B}(E_1)$, then

$$\mu_n \overset{(w)}{\to} \mu,\ n \to \infty \quad \Longrightarrow \quad \mu_n \circ g \overset{(w)}{\to} \mu \circ g,\ n \to \infty.$$

In fact, continuity is not really needed, the assumption $\mu\big(\{x \in E_1 \,|\, g \text{ not continuous in } x\}\big) = 0$ is enough for the theorem to hold, but the proof is a bit lengthy.

Proof. Let $f \in C_b(E_2)$, then $f \circ g \in C_b(E_1)$. Using the transformation formula for integrals twice (see Theorem 3.1.16) then gives

$$\int_{E_2} f \,\mathrm{d}\mu_n \circ g = \int_{E_1} f \circ g \,\mathrm{d}\mu_n \overset{n\to\infty}{\longrightarrow} \int_{E_1} f \circ g \,\mathrm{d}\mu = \int_{E_2} f \,\mathrm{d}\mu \circ g.$$

Hence, $\mu_n \circ g \overset{(w)}{\to} \mu \circ g$. □

We finish the section by writing down the generalised notion of convergence that was announced at the beginning of this section:

Definition 7.2.6. Let $X, X_1, X_2, ...$ be random variables with values in a metric space (E,d). Then we say that $(X_n)_{n\in\mathbb{N}}$ **converges to** X **in distribution** if $\mathbb{P}_{X_n} \overset{(w)}{\longrightarrow} \mathbb{P}_X$, $n \to \infty$. One usually writes $X_n \overset{(d)}{\to} X$, $X_n \Rightarrow X$ or $X_n \Rightarrow \mathbb{P}_X$ for $n \to \infty$ and also says $(X_n)_{n\in\mathbb{N}}$ converges weakly to X.

Realising that $f(X_n)$ is a real-valued random variable we find that convergence in distribution can be rewritten in the more probabilistic language

$$X_n \overset{(d)}{\longrightarrow} X,\ n \to \infty \quad \Leftrightarrow \quad \lim_{n\to\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)], \quad \forall f \in C_b(E).$$

In the case of real-valued random variables this is nothing else but the convergence in distribution from (7.2) that was used to formulate the central limit theorem. This also explains why sometimes convergence in distribution of random variables is called weak convergence. For random variables the continuous mapping theorem states

$$X_n \overset{(d)}{\longrightarrow} X,\ n \to \infty \quad \Longrightarrow \quad g(X_n) \overset{(d)}{\longrightarrow} g(X),\ ,n \to \infty$$

for all continuous g, a statement that is used a lot!

We will now turn towards the study of weak convergence of measures seen as metric convergence in $(\mathcal{M}_f(E), d_P)$ which is based on the following characterisation of convergence:

Proposition 7.2.7. Let (E,d) a metric space and $(x_n)_{n\in\mathbb{N}}$ a sequence in E. Then the following are equivalent:

- $x_n \to x, n \to \infty$,
- (i) $A := \{x_n : n \in \mathbb{N}\}$ is relatively sequentially compact.
  (ii) $x_{n_k} \to x, k \to \infty$, for all converging subsequences.

Proof. "$\Rightarrow$": Any subsequences of converging sequences converge to the same limit, hence, the second limit follows trivially and also the relative sequential compactness follows.

"$\Leftarrow$": If (x_n) does not converge towards x then there is some $\varepsilon > 0$ and a subsequence with $d(x_{n_k}, x) > \varepsilon$. Since also the subsequence $(x_{n_k}) \subseteq (x_n)$ is relatively sequentially compact there is another subsequence $(x_{n'_k})$ of (x_{n_k}) that converges. Since the limit must be x by assumption we find $d(x_{n'_k}, x) < \varepsilon$ for k large enough. But this is a contradiction. □

In order to use the proposition for weak convergence of measures we will proceed in three steps. In Section 7.3 a characteristation of relative sequential compactness will be proved. The so-called tightness gives a condition that can be checked if compact sets of E are sufficiently well understood. Section 7.4 adresses the problem of identifying limits of subsequences. Finally, combining both sections we can derive good conditions for convergence of measures on different Polish spaces (E, d). Lecture 13

7.3 Relative sequential compacteness of measures

In the light of Proposition 7.2.7. The aim of this section is to derive a well accessible equivalent notion to relative sequential compactness in this particular metric space.

Definition 7.3.1. A family $\mathcal{F} \subseteq \mathcal{M}_f(E)$ is called **tight** if for every $\varepsilon > 0$ there is a compact set $K \subseteq E$ with

$$\sup_{\mu \in \mathcal{F}} \mu\big(K^c\big) < \varepsilon$$

To get a visual idea of tightness recall the interpretation of a measure that describes how mass is distributed over the state-space E. Tightness means that no mass gets lost towards infinity, the mass of all measures is (uniformly) concentrated in compact sets. Before we go into Prohorov's

Tightness of four measures visualised in terms of distribution of mass

theorem let us check some examples to get some feeling of tight families.

Example 7.3.2.

- By Proposition 7.1.12 every family consisting of a single measure on a Polish space is tight
- $(\delta_{x_n})_{n \in \mathbb{N}}$ is a tight sequence for a Polish space (E, d) if and only if $(x_n)_{n \in \mathbb{N}}$ is totally bounded in E. For instance if $(x_n)_{n \in \mathbb{N}}$ is bounded for $E = \mathbb{R}$.
- If a family $(X_\alpha)_{\alpha \in I}$ of real-valued random variables is bounded in L^1, then the family of the laws $\{\mathbb{P}_{X_\alpha} : \alpha \in I\}$ is tight. This follows from our beloved Markov inequality as follows. If C is the L^1-bound and $\varepsilon > 0$, then $K := [-C/\varepsilon, C/\varepsilon]$ does the job:

$$\mathbb{P}_{X_\alpha}(K^c) = \mathbb{P}\big(|X_\alpha| > C/\varepsilon\big) \overset{\text{Markov}}{\leq} \frac{\mathbb{E}[|X_\alpha|]}{C/\varepsilon} \leq \varepsilon$$

- $\{U([-n, n]) : n \in \mathbb{N}\}$ is not tight in $\mathcal{M}_f(\mathbb{R})$ as mass disappears towards infinity.
- If K is a compact set and $(\mu_n)_{n \in \mathbb{N}}$ is a sequence in $\mathcal{M}_1(E)$ with $\mu_n(K) = 1$ for all $n \in \mathbb{N}$, then $(\mu_n)_{n \in \mathbb{N}}$ is tight.

The final example shows that a good understanding of compact sets of E can be extremely useful to understand tightness. We will return to this observation once we discuss in more detail $E = C([0, 1])$ in which for instance sets of bounded Hölder continuous functions are compact.

The importance of tightness becomes clear with Prohorov's famous theorem in combination with Proposition 7.2.7.

Theorem 7.3.3. (Prohorov)
Let (E,d) be a Polish metric space and $\mathcal{F} \subseteq \mathcal{M}_1(E)$, then

$$\mathcal{F} \text{ is weakly relatively sequentially compact} \quad \Leftrightarrow \quad \mathcal{F} \text{ is tight.}$$

The formulation of Prohorov's theorem is frightening on first sight, weakly relatively sequentially compact means that the subset $\mathcal{F}$ of $\mathcal{M}_1(E)$ is relatively compact (the closure is compact) with respect to the topology induced by the Prohorov metric. Using properties of closures and the sequential compactness this means that every sequence (μ_n) in $\mathcal{F}$ has a subsequence that converges to some limit in $\mathcal{M}_f$ but the limit does not necessarily belong to $\mathcal{F}$. The entire point is that Prohorov's theroem connects a measure property (tightness) with a topological property (compactness).

Proof. There is a relatively simple and a very hard direction. We start with the simpler one and discuss the hard direction only for $E = \mathbb{R}$. A sketch for the general case is given after the proof.

"$\Leftarrow$": We start with arguments similar to the ones from the proof of Proposition 7.1.12. Let $x_1, x_2, \ldots$ be dense in E and

$$A_{N,n} := \bigcup_{k=1}^{N} B_{\frac{1}{n}}(x_k) \ \uparrow E, \quad N \to \infty.$$

We first prove that

$$\lim_{N\to\infty} \inf_{\mu\in\mathcal{F}} \mu(A_{n,N}) = 1, \quad \forall n \in \mathbb{N}. \tag{7.3}$$

Suppose (7.3) does not hold for some $n \in \mathbb{N}$. Then there is some $c < 1$, an increasing sequence (N_j) of natural numbers and a sequence of measures $\mu_j \in \mathcal{F}$ with

$$\lim_{j\to\infty} \mu_j(A_{n,N_j}) \leq c < 1.$$

Since the sequence is weakly relative sequentially compact there is a subsequence (μ_{j_k}) that converges to some $\mu \in \mathcal{M}_1(E)$ (not necessarily in $\mathcal{F}$, μ could be in $\bar{\mathcal{F}}$). Since the $A_{n,N}$ are open, Portemanteau gives for all N

$$\mu(A_{n,N}) \leq \limsup_{k\to\infty} \mu_{j_k}(A_{n,N}) \overset{\text{mon.}}{\leq} \limsup_{k\to\infty} \mu_{j_k}(A_{n,N_{j_k}}) \leq c < 1.$$

But this gives a contradiction: $1 = \mu(E) \overset{\text{cont.}}{=} \lim_{k\to\infty} \mu(A_{n,N_{j_k}}) \leq c < 1$.
Now we use (7.3) to deduce the tightness. There are N, n_N with $\mu(A_{n_N,N}) \geq 1 - \frac{\varepsilon}{2^N}$ for all $\mu \in \mathcal{F}$. Defining $K := \bigcap_{N=1}^\infty \bar{A}_{n_N,N}$, K is closed and totally bounded, hence, K is compact as E is complete. Additionally,

$$\begin{aligned}
\mu(K^c) &= \mu\Big(\bigcup_{N=1}^\infty \bar{A}^c_{n_N,N}\Big) \\
&\overset{\text{sub. add.}}{\leq} \sum_{N=1}^\infty \mu(\bar{A}^c_{n_N,N}) \\
&= \sum_{N=1}^\infty \big(1 - \mu(\bar{A}_{n_N,N})\big) \\
&\overset{\text{mon.}}{\leq} \sum_{N=1}^\infty \big(1 - \mu(A_{n_N,N})\big) \leq \varepsilon
\end{aligned}$$

for all $\mu \in \mathcal{F}$. Hence, $\mathcal{F}$ is tight.

"$\Rightarrow$": Considering only $E = \mathbb{R}$ we have the big advantage that one can argue using the cumulative distribution function since

$$\mu_n \xrightarrow{(w)} \mu,\ n \to \infty \quad \Leftrightarrow \quad F_{\mu_n}(t) \to F_\mu(t),\ n \to \infty$$

for all points of continuity of F. Recall Theorem 4.5.9 and keep in mind that convergence in distribution of a sequence of random variables $X_n \sim F_n$ is equivalent to weak convergence of the corresponding sequence of probability measures $\mathbb{P}_{F_n}$. Here we use the measures μ_n and denote the corresponding CDFs by $F_n := \mu_n\big((-\infty, t]\big)$, $t \in \mathbb{R}$. Reformulated this way, Prohorov's theorem is only a theorem on CDFs: Every tight sequence of probability measures has a subsequence for which the CDFs converge to a limiting CDF at all points of continuity.

Let $(\mu_n)_{n\in\mathbb{N}}$ be a sequence in $\mathcal{F}$ and $(F_n)_{n\in\mathbb{N}}$ the corresponding sequence of CDFs. Since $\big(F_n(t)\big)_{n\in\mathbb{N}}$ is bounded for all $t \in \mathbb{R}$ there is a convergent subsequence. Using diagonalisation [3] there is a subsequence $(n_k)_{k\in\mathbb{N}}$ so that $F_{n_k}(q) \to \tilde{F}(q)$, $q \in \mathbb{Q}$, for some function $\tilde{F}$. We will check that

(i) $F(t) := \inf\big\{\tilde{F}(q)\colon q > t,\ q \in \mathbb{Q}\big\}$, $t \in \mathbb{R}$, is a cumulative distribution function.

(ii) $\mu_n \xrightarrow{(w)} \mu$, $n \to \infty$, where $\mu \sim F$.

Both claims are mostly technical and not too surprising, the interesting point is how the tightness comes in. The tightness is needed to prove that F does not loose mass at infinity, i.e. $\lim_{t\to+\infty} F(t) = 1$.

(i) Let us check the defining properties of a cumulative distribution function. We are a bit sloppy for the claims that obviously hold even though writing the details is a bit tedious (playing with ε-N and the definition of the infimum), those details do not lead to any deeper understanding. We actually skipped the same arguments before in the proof of Theorem 5.3.4.

- $F\colon\ \mathbb{R} \to [0,1]$, since $F_n\colon \mathbb{R} \to [0,1]$.
- $\tilde{F}$ is increasing on $\mathbb{Q}$ as all F_n are increasing, then F inherits the property construction.
- F is right-continuous by definition (here one should work careful with the definition of the infimum).
- $\lim_{t\to+\infty} F(t) = 1$, $\lim_{t\to-\infty} F(t) = 0$ is the interesting part. This is tightness, no mass gets lost. The implications look complicated but the argument is simple:

$$\begin{aligned}
\text{Tightness} \quad &\Rightarrow \quad \forall \varepsilon > 0\, \exists M \in \mathbb{Q}\colon \mu\big([-M,+M]^c\big) < \varepsilon\ \forall \mu \in \mathcal{F}\\
&\Rightarrow \quad \forall \varepsilon > 0\, \exists M \in \mathbb{Q}\colon \mu_n\big([-M,+M]^c\big) < \varepsilon\ \forall n \in \mathbb{N}\\
&\Rightarrow \quad \forall \varepsilon > 0\, \exists M \in \mathbb{Q}\colon 1 - F_n(M) + F_n(-M) < \varepsilon\ \forall n \in \mathbb{N}\\
&\Rightarrow \quad \forall \varepsilon > 0\, \exists M \in \mathbb{Q}\colon F_n(M) \geq 1 - \varepsilon, F_n(-M) \leq \varepsilon\ \forall n \in \mathbb{N}\\
&\Rightarrow \quad \forall \varepsilon > 0\, \exists M \in \mathbb{Q}\colon F(M) \geq 1 - \varepsilon, F(-M) \leq \varepsilon\\
&\Rightarrow \quad \lim_{t\to+\infty} F(t) = 1,\ \lim_{t\to-\infty} F(t) = 0.
\end{aligned}$$

(ii) As explained above we can apply Theorem 4.5.9 and prove the pointwise convergence of the CDFs F_n at all points of continuity of F. Let t be a point of continuity and $\varepsilon > 0$. [4] Then there are $q_i \in \mathbb{Q}$ with $q_1 < q_2 < t < q_3$ with $F(q_3) - F(q_1) < \varepsilon$. Using monotonicity we have

$$F_{n_k}(q_2) \leq F_{n_k}(t) \leq F_{n_k}(q_3),$$

[3] explain
[4] Bild

which gives

$$\tilde{F}(q_2) = \liminf F_{n_k}(q_2) \leq \liminf F_{n_k}(t) \leq \limsup F_{n_k}(t) \leq \limsup F_{n_k}(q_3) = \tilde{F}(q_3)$$

using the convergence on $\mathbb{Q}$. Hence,

$$F(q_1) \leq \tilde{F}(q_2) \leq \liminf F_{n_k}(t) \leq \limsup F_{n_k}(t) \leq \tilde{F}(q_3) \leq F(q_3).$$

But this implies that

$$\liminf F_{n_k}(t), \limsup F_{n_k}(t) \in [F(t) - \varepsilon, F(t) + \varepsilon].$$

Since ε is arbitrary we proved that $\lim F_{n_k}(t)$ exists and is equal to $F(x)$.

□

Here is a sketch on how the complicated "⇒" direction of Prohorov's theorem is proved for general Polish spaces:

(i) Since E is separable there is a countable base $\mathcal{U} \subseteq \tau$ for the topology, i.e. $O = \bigcup_{U \in \mathcal{U}, U \subseteq O} U$ for all O open.

(ii) Define a possible limit measure on $\mathcal{U}$: $\big(\mu_n(U)\big)_{n\in\mathbb{N}}$ is a bounded sequence, hence, has a converging subsequence. Since there are countably many U_1, U_2,.... $\in \mathcal{U}$ the same diagonalisation argument gives a subsequence such that $\big(\mu_{n_k}(U)\big)_{k\in\mathbb{N}}$ converges for all U.

(iii) Define $\mu(U) := \lim_{k\to\infty} \mu_{n_k}(U)$ for all $U \in \mathcal{U}$.

(iv) Main step: Carathéodory extension style construction of a probability measure $\bar{\mu}$ on $\mathcal{B}(E)$ with $\bar{\mu}(U) = \mu(U)$ for all $U \in \mathcal{U}$.

(v) Show $\bar{\mu}(O) \leq \liminf_{k\to\infty} \mu_{n_k}(O)$ for all O open.

(vi) Portemanteau implies $\mu_{n_k} \xrightarrow{(w)} \mu$, $k \to \infty$.

Hence, there is a weakly converging subsequence (not necessarily to a limit in $\mathcal{F}$) and this is precisely the relative compactness with respect to weak convergence.

A first simple consequence of Prohorov's characterisation is the following reformulation of Proposition 7.2.7 for weak convergence of probability measures:

Proposition 7.3.4. (A useful characterization of weak convergence) Let (E, d) be a Polish metric space and $\mu, \mu_1, \mu_2,... \in \mathcal{M}_1(E)$. Then weak convergence of μ_n to μ is equivalent to

(i) $\{\mu_n : n \in \mathbb{N}\}$ is tight,

(ii) $\lim_{n\to\infty} \int_E f \,\mathrm{d}\mu_n = \int_E f \,\mathrm{d}\mu$ for <u>some</u> separating family $C \subseteq C_b(E)$ of $\mathcal{M}_1(E)$.

Proof. "⇒": Converging sequences are relatively compact sets (in metric spaces all subsequences have the same limit), hence, tight by Prohorov's theorem. The convergence of the integrals holds by definition of weak convergence (even for all $f \in C_b(E)$).

"⇐": We use Proposition 7.2.7. Tightness implies the sequential relative compacteness and we only need to identify μ as limit of all converging subsequences. If (μ_{n_k}) is a subsequences of (μ_n) with limit ν, then

$$\int_E f \,\mathrm{d}\mu = \lim_{n\to\infty} \int_E f \,\mathrm{d}\mu_n = \lim_{k\to\infty} \int_E f \,\mathrm{d}\mu_{n_k'} = \int_E f \,\mathrm{d}\nu, \quad f \in C.$$

The first equality holds for all $f \in C$ by assumption, the second as (μ_{n_k}) is a subsequence, and the third even for all $f \in C_b(E)$. Since C was assumed to be separating we proved $\nu = \mu$. □

In order to prove weak convergence we will often refer to Proposition 7.3.4, but there are two drawbacks that depend crucially on the underlying space E. One needs a good understanding of tightness (i.e. of compact sets of E) and one needs a good separating family. For instance for $E = C([0,1])$ it will turn out to be more useful to circumvent the formulation of Proposition 7.3.4 and argue a bit more directly. Lecture 14

7.4 Identification of probability laws on $\mathbb{R}^d$

The previous section showed how to reformulate relative sequential compactness in terms of tightness. In this section we will deal more closely with the second property from Proposition 7.2.7, the identification of limits of converging subsequences. We will give precise examples that help to apply Proposition 7.3.4 for particularly simple E. As a side product we derive ways of determining the law of a random variable through "enough" expectations. Before doing so let us dive a bit into approximation theory.

Theorem 7.4.1. (Weierstraß approximation theorem)
Let $f\colon [0,1] \to \mathbb{R}$ be continuous, then there is a sequence $(f_n)_{n\in\mathbb{N}}$ of polynomials on $[0,1]$ with $\|f_n - f\|_\infty \to 0$, $n \to \infty$.

In words: The polynomials are dense in $(C([0,1]), ||\cdot||_\infty)$.

The proof we give is awesome, as it is a probabilistic proof for an analytic theorem. The sequence of polynomials is actually given explicitly, the so-called Bernstein polynomials.

Proof. Let $X_1, ..., X_n$ be iid Ber(p)-distributed random variables, hence, $S_n = \sum_{k=1}^n X_k$ is Bin(n,p)-distributed. Now recall the proof of the weak law of large numbers (4.5.12) - Tschebycheff and Bienaymé:

$$\mathbb{P}\Big(\Big|\frac{S_n}{n} - p\Big| > \varepsilon\Big) \leq \frac{\mathbb{V}(S_n)}{n^2\varepsilon^2} = \frac{n\mathbb{V}[X_1]}{n^2\varepsilon^2} = \frac{p(1-p)}{n\varepsilon^2}$$

Next, computing the discrete expectation for Bin(n,p) yields

$$\mathbb{E}\Big[f\Big(\frac{S_n}{n}\Big)\Big] = \sum_{k=0}^n f\Big(\frac{k}{n}\Big)\binom{n}{k}p^k(q-p)^{n-k} = f_n(p)$$

with the so-called Bernstein polynomial

$$f_n(x) = \sum_{k=0}^n f\Big(\frac{k}{n}\Big)\binom{n}{k}x^k(1-x)^{n-k},\ \ x \in [0,1].$$

Now fix $\varepsilon > 0$. Since f is uniformly continuous ($[0,1]$ is compact) there is some $\delta > 0$ with

$$|x - x'| < \delta \ \Rightarrow\ |f(x) - f(x')| < \varepsilon,$$

so that we can estimate

$$\Big|f\Big(\frac{S_n}{n}\Big) - f(p)\Big| \leq \varepsilon + \mathbf{1}_{|\frac{S_n}{n} - p| \geq \delta} 2\,\|f\|_\infty\,.$$

This gives

$$\begin{aligned}
|f_n(p) - f(p)| &= \Big|\mathbb{E}\Big[f\Big(\frac{S_n}{n}\Big)\Big] - \mathbb{E}\big[f(p)\big]\Big|\\
&\leq \mathbb{E}\Big[\Big|f\Big(\frac{S_n}{n}\Big) - f(p)\Big|\Big]\\
&= \varepsilon + 2\cdot\|f\|_\infty\,\mathbb{P}\Big(\Big|\frac{S_n}{n} - p\Big| \geq \delta\Big)\\
&\leq \varepsilon + 2\cdot\|f\|_\infty \cdot \frac{p(1-p)}{n\cdot\delta}.
\end{aligned}$$

Putting together what we have so far we obtain

$$\|f_n - f\|_\infty = \sup_{p\in[0,1]} |f_n(p) - f(p)| \le \varepsilon + 2\cdot\|f\|_\infty \cdot \frac{1}{n\delta^2} \to \varepsilon, \quad n\to\infty.$$

Since ε was arbitrary, we proved that $\|f_n - f\|_\infty \to 0$, $n\to\infty$. □

It turns out that classes of complex-valued functions are more useful than only real-valued functions. For that sake let us recall some definitions and facts for the complex numbers:

As a set the complex numbers are identical to $\mathbb{R}^2$ with a different notation for the vectors:

$$\mathbb{C} := \{z = u + iv \,|\, u, v \in \mathbb{R}\}.$$

The following properties will be used:

- $|z| = \sqrt{u^2+v^2}$
- $\mathcal{R}e(z) = u$, $\mathcal{I}m(z) = v$
- Seen as a metric space $(\mathbb{C}, |\cdot|)$ is a Polish metric space and identical to $(\mathbb{R}^2, |\cdot|)$. In particular $\mathcal{B}(\mathbb{C}) = \mathcal{B}(\mathbb{R}^2)$ and in terms of measure theory all results from Section **??** apply. As an example, using Proposition 4.2.11, a $\mathbb{C}$-valued random variable $X\colon \Omega\to\mathbb{C}$ is measurable if and only if $\mathcal{R}e(X)$ and $\mathcal{I}m(X)$ are measurable.
- Complex conjugation: $\bar{z} = u - i\cdot v$
- Polar coordinates: $z = |z|\cdot e^{i\varphi}$ for an angle $\varphi\in[0,2\pi)$.
- Multiplication $z_1\cdot z_2 := (u_1u_2 + u_2v_2) + i(u_1v_2 + v_2u_1)$ so that $i^2 = -1$, or in polar coordinates: $z_1\cdot z_2 = |z_1|\,|z_2|e^{i(\varphi_1+\varphi_2)}$, which corresponds to rotation and expanding.
- Addition: $z_1 + z_2 = (u_1+u_2) + i(v_1+v_2)$
- Field with $+, \cdot$, $0 = (0,0)$, $1 = (1,0)$
- $\mathcal{R}e(z) = \frac{z+\bar z}{2}$, $\mathcal{I}m(z) = \frac{z-\bar z}{2i}$
- Exponential function $\exp\colon\mathbb{C}\to\mathbb{C}$

  $$\sum_{k=0}^\infty \frac{z^k}{k!} = \exp(z) = \exp(u+iv) = \exp(u)\exp(iv)$$

  with

  $$\exp(z_1+z_2) = \exp(z_1)\exp(z_2).$$

  and Euler formula

  $$\exp(iv) = \cos(u) + i\sin(v)$$

  that implies $|e^{iv}| = 1$ for all $v\in\mathbb{R}$.
- $\cos(u) = \frac{e^{iu}+e^{-iu}}{2}$, $\sin(u) = \frac{e^{iu}-e^{-iu}}{2i}$

Complex integration of functions $f:\mathbb{C}\to\mathbb{C}$ is covered in complex analysis lectures with all magic tricks to compute such integrals. We will not touch upon this topic and only define the Lebesgue integral for measurable $f:\Omega\to\mathbb{C}$ by integrating separately real- and imaginary part:

$$\int_\Omega f\,\mathrm{d}\mu := \int_\Omega \mathcal{R}e(f)\,\mathrm{d}\mu + i\int_\Omega \mathcal{I}m(f)\,\mathrm{d}\mu \in \mathbb{C}$$

if both (real-valued) integrals exist. Rules for the integral can be deduce from rules for both (real-valued) integrals. Expectations of complex-valued random variables are defined as follows:

$$\mathbb{E}[X] := \int_\Omega X(\omega)\,\mathrm{d}\mathbb{P}(\omega) = \int_\Omega \mathcal{R}e(X)(\omega)\,\mathrm{d}\mathbb{P}(\omega) + i\int_\Omega \mathcal{I}m(X)(\omega)\,\mathrm{d}\mathbb{P}(\omega) \in \mathbb{C}$$

which can be computed with typical tools for real-valued random variables since

$$\mathbb{E}[X] = \mathbb{E}[\mathcal{R}e(X)] + i\,\mathbb{E}[\mathcal{I}m(X)].$$

An important special case appears when X is a real-random variable and $g:\mathbb{R}\to\mathbb{C}$. Expectations can be calculated in the usual way as

$$\mathbb{E}[g(X)] = \begin{cases} \int_{\mathbb{R}} g(x)f(x)\,\mathrm{d}x & : X \text{ is absolutely continuous with density } f \\ \sum_{k=1}^N g(a_k)p_k & : X \text{ is discrete with values } a_k \text{ and probabilities } p_k \end{cases}. \tag{7.4}$$

To see why just split the expectations into two real-valued expectations, use the standard formulas and put them together. The most important example that we will discuss below is $\mathbb{E}[e^{itX}]$ for real-valued random variables.

Definition 7.4.2. Let (E,d) be a metric space and $K=\mathbb{R}$ or $K=\mathbb{C}$. A subset $C\subseteq C_b(E,K)$ is called an **algebra** of functions if

- $1\in C$,
- $f,g\in C \;\Rightarrow\; f\cdot g,\; f+h\in C$,
- $f\in C,\ \alpha\in K \;\Rightarrow\; \alpha\cdot f\in C$,

If $K=\mathbb{C}$ we always assume C is closed under complex conjugation. We say the **algebra** C **separates points** if for all $y,x\in E$ there is some $f\in C$ with $f(x)\neq f(y)$.

There are many examples of algebras, for instance the set of polynomials or all exponential functions. We will discuss the examples in the upcoming sections but first deal with an important generalisation of the Weierstraß approximation theorem. The Stone-Weierstraß approximation theorem allows to generalise the $[0,1]$ to some compact metric space and the polynomials to any algebra of functions:

Theorem 7.4.3. (Stone-Weierstraß approximation theorem)
Let (E,d) a compact metric space, $K=\mathbb{R}$ or $K=\mathbb{C}$, and $C\subseteq C_b(E,K)$ an algebra of functions that separates points. Then C is dense in $\big(C_b(E,K),||\cdot||_\infty\big)$.

To see why the algebra should separate points take as an example the set of all constant functions. This forms an algebra, does not separate points and is clearly not large enough to approximate all bounded continuous functions.

Proof. Let us first consider the case $K=\mathbb{R}$.

(i) First note that also $\bar{C}$ is an algebra, as the defining properties rely on continuous operations that transfer through taking limits (recall that $\bar{C}$ consists of all limits of sequences from C). By the is Weierstraß approximation theorem there is a sequence $(p_n)_{n\in\mathbb{N}}$ of polynomials with $p_n\to p, n\to\infty$, uniformly on $[0,1]$, where $p(x)=\sqrt{x}$. If $f\in\bar{C}$, then $|f|\in\bar{C}$ because

$$|f| = \|f\|_\infty\cdot \lim_{n\to\infty} \underbrace{p_n\Big(\underbrace{\frac{f^2}{\|f\|_\infty^2}}_{\in[0,1]}\Big)}_{\in\bar{C}\text{ as }\bar{C}\text{ is an algebra}} \in \bar{C},$$

as $\bar{C}$ is closed. Using $f \vee g = \frac{1}{2}(f + g + |f - g|)$ and $f \wedge g = \frac{1}{2}(f + g - |f - g|)$ we see that the algebra $\bar{C}$ (operations transfer to limit) is also closed under taking pointwise maxima and minima.

(ii) Now fix $\varepsilon > 0$. Let $f \in C_b(E, \mathbb{R})$ and $x \in E$. Then there is $g_x \in \bar{C}$ with

- $g_x(x) = f(x)$,
- $g_x(y) \leq f(y) + \varepsilon$, $\forall y \in E$.

To see why note that C separates points, hence, for all $z \in E \setminus \{x\}$ there is a function $H_z \in C$ with $H_z(z) \neq H_z(x)$. By adding a constant we can assume that $H_z(x) = 0$. Next, define $h_x = f$ and, for $z \neq x$,

$$h_z(y) := f(z) + \frac{f(x) - f(z)}{H_z(x)} \cdot H_z(y), \quad y \in E,$$

which is a function in C that coincides with f in x and z. Since f and h_z are continuous, for all $z \in E$ there is a neighbourhood U_z of z with $h_z \leq f + \varepsilon$ on U_z. Using compactness there is a finite covering $U_{z_1}, ..., U_{z_n}$ of E of such neighbourhoods. If finally we define $g_x := \min\{h_{z_1}, ..., h_{z_n}\}$, then $g \in \bar{C}$ by (i) and g satisfies the two claimed properties by the construction.

(iii) Since f and g_x are continuous and $f(x) = g_x(x)$ there are neighborhoods U_x of x with $g_x \geq f - \varepsilon$ on V_x. By compactness finitely many $V_{x_1}, ..., V_{x_k}$ cover E. Then define $g := \max\{g_{x_1}, ..., g_{x_k}\} \in \bar{C}$. By construction this gives $f + \varepsilon \geq g \geq f - \varepsilon$ or $\|f - g\|_\infty < \varepsilon$. Since ε is arbitrary we can find a sequence in $\bar{C}$ that converges uniformly to f. In other words, $\bar{C} = C_b(E, \mathbb{R})$.

It remains to consider the case $K = \mathbb{C}$. If $f = \mathcal{R}e(f) + \mathcal{I}m(f) \in C$, then real- and imaginary-part are in C. This follows from the assumption on C by writing $\mathcal{R}e(f) = \frac{f+\bar{f}}{i}$ and $\mathcal{I}m(f) = \frac{f-\bar{f}}{2i}$. Hence,

$$C_{\mathcal{R}} := \{\mathcal{R}e(f) \colon f \in C\} \subseteq C \quad \text{and} \quad C_{\mathcal{I}} := \{\mathcal{I}m(f) \colon f \in C\} \subseteq C$$

and, thus, both form algebras of real functions. Both sets are also separating according to the following trick. Suppose x and y are separated by $f \in C$, that is $f(x) \neq f(y)$. Since C is closed under adding 1 and multiplication with constants (which is rotation and stretching) there are functions $h, g \in C$ such that

$$h(x) = ia_1,\, h(y) = a_2 + ia_3 \quad \text{and} \quad g(x) = b_1 + ib_2,\, g(y) = b_3$$

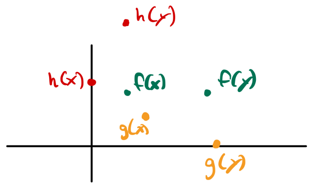
for some $a_i, b_i \neq 0$. Thus,

$$\mathcal{R}e(h)(x) = 0 \neq b_1 = \mathcal{R}e(g)(x) \quad \text{and} \quad \mathcal{I}m(h)(y) = a_3 \neq 0 = \mathcal{I}m(g)(y).$$

Hence, $\mathcal{R}e(h) \in C_{\mathcal{R}}$ and $\mathcal{I}m(g) \in C_{\mathcal{I}}$ both separate x and y. This proves that $C_{\mathcal{R}}$ and $C_{\mathcal{I}}$ are separating algebras of real functions. Why did we need to involve this little trick? It is possible that the imaginary parts and/or real parts of $f(x)$ and $f(y)$ coincide. If they do, then the imaginary parts and/or real parts of f do not separate x and y. To avoid this problem we can multiply f by a constant z' to rotate (and stretch) the complex numbers $f(x)$ and $f(y)$ to ensure that the real and imaginary parts of $h(z) := z'f(z) \in C$ differ and thus separate points. The trick is best understood in a picture:
Therefore, from the above, $\bar{C}_{\mathcal{R}} = \bar{C}_{\mathcal{I}} = C_b(E, \mathbb{R})$. Since $C_b(E, \mathbb{C}) = C_b(E, \mathbb{R}) + i \cdot C_b(E, \mathbb{R})$ we find that $C = C_{\mathcal{R}} + i \cdot C_{\mathcal{I}}$ is dense in $C_b(E, \mathbb{C})$. □

What is the magic of Stone-Weierstraß? Proving that families of functions are dense in other families is typically hard, a purely analytic ε-N story. Stone-Weierstraß allows us to prove the density in a completely different way, only checking very simple algebraic properties! If we think about polynomials or exponential functions the algebraic properties are easily checked. Lecture 15

The situation becomes even simpler if integrals are involved as linearity of integrals fits nicely to the properties of algebras.



The rotation trick to separate points

Corollary 7.4.4. Let (E, d) be a compact metric space and $C \subseteq C_b(E, K)$ a family that separates points, is closed under multiplication and contains 1 (constant 1 function). Then C is separating for $\mathcal{M}_1(E)$.

The examples that will appear in the sequal are typically polynomials or exponential functions on compact subsets of $\mathbb{R}$.

Proof. We start with a little trick on separating functions and algebras that relies on the linearity of integrals. Let μ, $\nu \in \mathcal{M}_1(E)$ with $\int_E g \, \mathrm{d}\mu = \int_E g \, \mathrm{d}\mu$ for all $g \in C$. Taking all linear combinations of functions in C and calling them C' then by linearity of integrals the equality also holds for all functions in C' and C' is an algebra of functions. Using the Stone-Weierstraß theorem, the equality of integrals holds for a dense subset of $C_b(E)$.
Fix $\varepsilon > 0$. Then for all $f \in C_b(E)$ there exists $g \in C'$ with $\|f - g\|_\infty < \varepsilon$ so that

$$\begin{aligned}
&\Big| \int_E f \, \mathrm{d}\mu - \int_E f \, \mathrm{d}\nu \Big| \\
&\leq \Big| \int_E f \, \mathrm{d}\mu - \int_E g \, \mathrm{d}\nu \Big| + \Big| \int_E g \, \mathrm{d}\nu - \int_E g \, \mathrm{d}\mu \Big| + \Big| \int_E g \, \mathrm{d}\mu - \int_E f \, \mathrm{d}\nu \Big| \\
&\leq \|f - g\|_\infty \cdot \mu(E) + 0 + \|f - g\|_\infty \cdot \nu(E) = 2\varepsilon.
\end{aligned}$$

Since this works for all $\varepsilon > 0$ it follows that

$$\int_E f \, \mathrm{d}\mu = \int_E f \, \mathrm{d}\nu, \quad \forall f \in C_b(E).$$

Recalling from Proposition 7.1.15 that $C_b(E)$ is separating for $\mathcal{M}_1(E)$ we proved $\mu = \nu$. Hence, C is separating for $\mathcal{M}_1(E)$. $\square$

Theorem 7.4.5. (i) Every measures $\mu \in \mathcal{M}_1([a, b])$ is uniquely determined by all integrals

$$\int_{[a,b]} x^m \, \mathrm{d}\mu(x), \quad m \in \mathbb{N}_0.$$

(ii) The law of a <u>bounded</u> real-valued random variable is determined by all it's moments, i.e. if $\mathbb{E}[X^m] = \mathbb{E}[Y^m]$ for all $m \in \mathbb{N}$, then $X \sim Y$.

Proof. All we need to do is to apply Corollary 7.4.4.
First note that $([a, b], |\cdot|)$ is a compact metric space. Define C to be the family of monomials on $[a, b]$, that is $C = \{x^m : m \in \mathbb{N}_0\}$. Then C is closed under multiplication, separates points in $[a, b]$ and contains the constant 1 function. Hence, C is a separating family for $\mathcal{M}_1([a, b])$. But then, recalling the definition of a separating family,

$$\int_{[a,b]} x^m \, \mathrm{d}\mu(x) = \int_{[a,b]} x^m \, \mathrm{d}\nu(x), \quad \forall m \in \mathbb{N}_0,$$

implies $\nu = \mu$.
The second claim follows from the first claim applied to the laws $\mathbb{P}_X$. $\square$

Theorem 7.4.6. (i) Every measures $\mu \in \mathcal{M}_1([0,\infty)])$ is uniquely determined by all integrals

$$\int_{[0,\infty)} e^{-\lambda x}\,\mathrm{d}\mu(x), \quad \lambda > 0.$$

(ii) The law of a <u>non-negative</u> random variable is determined by it's Laplace transformation $L_X(\lambda) := \mathbb{E}\big[e^{-\lambda X}\big]$, $\lambda \geq 0$, i.e. if $L_X = L_Y$ then $X \sim Y$.

Of course part (ii) of the theorem also holds for non-positive random variables replacing the Laplace transformation by the moment generating function $M_X(t) = \mathbb{E}[e^{tX}]$ for all $t > 0$.

Proof. As in the previous proof we would like to apply Corollary 7.4.4 to the family of exponential functions, but there is a problem as $[0,\infty)$ is not compact. We apply a trick from Functional Analysis and use $E = [0,\infty]$ instead (the so-called one-point compactification).

$([0,\infty], d)$ is a compact metric space with $d(x,y) := |e^{-x} - e^{-y}|$. The metric is compatible with the usual convergence in $[0,\infty]$.

Note that convergence towards ∞ is divergence from basic analysis. Continuity of functions $f : [0,\infty] \to \mathbb{R}$ is best understood through sequences, $\lim_{n\to\infty} f(x_n) = f(x)$ and this takes in particular sequences that diverge to ∞. Hence, we can extend continuous functions on $[0,\infty)$ to continuous functions on $[0,\infty]$ if and only if $\lim_{x\to\infty} f(x)$ exists and in that case we must set $f(\infty) = \lim_{x\to\infty} f(x)$. For the exponential functions we thus define

$$f_\lambda(x) := \begin{cases} e^{-\lambda x} & : x > 0 \\ 0 & : x = \infty, \lambda > 0 \\ 1 & : x = \infty, \lambda = 0 \end{cases}.$$

Then $C := \{f_\lambda : \lambda > 0\} \subseteq C_b([0,\infty])$ separates points, $f_0 \equiv 1 \in C$, and $f_\mu \cdot f_\lambda = f_{\mu+\lambda}$. Hence, C separates probability measures on $\mathcal{B}([0,\infty])$ and in particular measures on $\mathcal{B}([0,\infty))$ which can be extended to $\mathcal{B}([0,\infty])$ by extending a measure μ with $\bar{\mu}(\{+\infty\}) = 0$. That is, for $\nu, \mu \in \mathcal{M}([0,\infty])$,

$$\begin{aligned} & \int_{[0,\infty)} e^{-\lambda x}\,\mathrm{d}\mu(x) = \int_{[0,\infty)} e^{-\lambda x}\,\mathrm{d}\nu(x), \quad \forall \lambda \geq 0 \\ \Rightarrow \quad & \int_{[0,\infty]} f\,\mathrm{d}\bar{\mu} = \int_{[0,\infty]} f\,\mathrm{d}\bar{\nu}, \quad \forall f \in C \\ \Rightarrow \quad & \bar{\mu} = \bar{\nu} \\ \Rightarrow \quad & \mu = \nu. \end{aligned}$$

The second claim follows from the first claim applied to the laws $\mathbb{P}_X$. □

Could we extend the same argument also to all measures on $\mathbb{R}$ (i.e. all random variables)? No! To do so we needed a family that is closed under multiplication (that leads to polynomials or exponentials) that converge at $+\infty$ and $-\infty$. Trying to find such a family is a good exercise to better understand the previous proofs. The trick that we get to know below is to replace the exponential functions by complex exponential functions.

A little induction shows that all moments of $U \sim \mathcal{U}([a,b])$ are given by

$$\mathbb{E}[U^n] = \frac{1}{n+1}\sum_{k=0}^{n} a^k b^{n-k} = \frac{b^{n+1} - a^{n+1}}{(n+1)(b-a)} \tag{7.5}$$

If for some reason one can show that a random variable has the moments from (4.3.18), then the random variable must be uniform.

It is important to have an example in mind that shows that all moments do not uniquely determine the law of unbounded random variables. Here is nice counterexample that appears in the Black-Scholes theory of time-continuous financial markets. A random variable X is called **log-normal** distributed if $X = \exp(Z)$, with $Z \sim \mathcal{N}(0,1)$. Using substitution in $\mathbb{P}(X \leq t) = \mathbb{P}(Z \leq \log(t)) = \int_{-\infty}^{\log(t)} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \,\mathrm{d}x$ shows that X has density

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \frac{1}{x} e^{-\frac{1}{2}\log(x)^2} \mathbf{1}_{[0,\infty)}(x), \quad x \in \mathbb{R}.$$

From the moment generating function of the standard Gaussian distribution (see **??**) we obtain a formula for the moments: $\mathbf{E}\big[X^n\big] = \mathbf{E}\big[e^{n\cdot Z}\big] = e^{\frac{1}{2}n^2}$. Next, we will write down an entire family of density functions that give the same moments but are log-normal densities. For $\alpha \in [-1,1]$ define

$$f_\alpha(x) := f_X(x) \cdot \big(1 + \alpha \cdot \sin(2\pi \log(x)\big) \mathbf{1}_{[0,\infty)}(x), \quad x \in \mathbb{R}.$$

Those functions are clearly non-negative and the computation below for $n = 0$ shows they integrate to 1, hence, they are densities. To prove the moments are identical to those of the log-normal distribution it suffices to show that

$$m(n) := \int_0^\infty x^n \cdot f_X(x) \cdot \sin(2\pi \log(x)) \,\mathrm{d}x = 0, \quad \forall n \in \mathbb{N}_0,$$

and combine with $\int_{\mathbb{R}} x^n \cdot f_X(x) \,\mathrm{d}x = \mathbb{E}[X^n] = e^{\frac{1}{2}n^2}$. Here is the computation:

$$\begin{aligned} m(n) &= \int_{-\infty}^\infty e^{y\cdot n} \cdot f_X(e^y) \cdot \sin(2\pi y) \cdot e^y \,\mathrm{d}y \\ &\overset{\text{subst. } y=z+n}{=} \int_{-\infty}^\infty e^{zn+n^2} f_X(e^{z+n}) \sin(2\pi z + 2\pi n) e^{z+n} \,\mathrm{d}z \\ &= \frac{1}{\sqrt{2\pi}} e^{n^2} \int_{-\infty}^\infty e^{-\frac{1}{2}z^2 - \frac{1}{2}n^2} \sin(2\pi z) \,\mathrm{d}z \\ &= \frac{1}{\sqrt{2\pi}} e^{\frac{1}{2}n^2} \underbrace{\int_{-\infty}^\infty e^{-\frac{1}{2}z^2} \sin(2\pi z) \,\mathrm{d}z}_{=0,\text{ odd function and integrable}} = 0. \end{aligned}$$

That's it.

No doubt, the most exciting random variables (such as the Gaussian) are neither concentrated on a bounded interval nor are positive (or negative). In general the law is not determined only by moments as $\mathbb{R}$ (or $\mathbb{R}^d$) cannot be compactified such that enough functions are continuous on the compactification. Hence, an approach different from the Stone-Weierstraß theory is needed. The alternative approach uses the periodic structure of e^{it}:

Definition 7.4.7. For $\mu \in \mathcal{M}_1(\mathbb{R}^d)$ the function $\varphi_\mu : \mathbb{R}^d \to \mathbb{C}$ defined by

$$\varphi_\mu(t) := \int_{\mathbb{R}^d} e^{\langle t,x\rangle} \,\mathrm{d}\mu(x) \quad t \in \mathbb{R}^d,$$

is called the **characteristic function** (or Fourier transform) of μ.

The wording Fourier transformation comes from Fourier Analysis where integrable functions are studied through their Fourier transform $\hat{f}(x) := \int e^{i\langle x,y\rangle} f(y) \,\mathrm{d}y$. In probability theory people prefer to use the name characteristic function as a random variable is uniquely characterised through the characteristic functions.

As always we can either define objects for probability measures or random variables with the corresponding law. Since the characteristic function is such an important object let us fix the other notation as well:

Definition 7.4.8. For a random vector X the function $\varphi_X : \mathbb{R}^d \to \mathbb{C}$ defined by

$$\varphi_X(t) := \mathbb{E}[e^{i\langle t,X\rangle}] = \varphi_{\mathbb{P}_X}(t), \quad t \in \mathbb{R}^d,$$

is called the textbfcharacteristic function of X.

The characteristic function plays exactly the same role as moment generating function $\mathcal{M}_X$, or the distribution function F_X. It is just a function with certain properties (see below) that can be used to study more complicated mathematical objects such as measures or random variables. The major advantage of the characteristic function is that φ is always well-defined because φ_X is always defined as $|e^{ix}| = 1$.

Here are some properties that indicate that characteristic functions are useful to work with:

Lemma 7.4.9. Let X be a random vector and $\varphi_X : \mathbb{R}^d \to \mathbb{C}$ the characteristic function.

(i) $\varphi_X(0) = 1$ and $|\varphi_X(t)| \le 1$ for all $t \in \mathbb{R}^d$.

(ii) $\varphi_{aX+b}(t) = \varphi_X(at)e^{i\langle b,t\rangle}$ for all $a \in \mathbb{R}$ and $b \in \mathbb{R}^d$.

(iii) φ is real-valued if and only if X is symmetric, i.e. $\mathbb{P}_X = \mathbb{P}_{-X}$.

(iv) $\varphi_{X+Y} = \varphi_X \cdot \varphi_Y$ if X and Y are independent.

Proof. If you are not used to integrals for complex-valued functions the arguments are a bit more complicated than one might think.

To get started please check the linearity $\int(\alpha f + \beta g)\,\mathrm{d}\mu = \alpha\int f\,\mathrm{d}\mu + \beta\int f\,\mathrm{d}\mu$ for $\alpha, \beta \in \mathbb{C}$ of complex integrals by using the definition and properties of real integrals.

(i) Let us first assume the usual triangle inequality also for complex integrals. Then we obtain

$$|\varphi_X(t)| = \big|\mathbb{E}\big[e^{i\langle t,X\rangle}\big]\big| = \Big|\int_\Omega e^{i\langle t,X(\omega)\rangle}\,\mathrm{d}\mathbb{P}(\omega)\Big| \overset{\Delta}{\le} \int_\Omega \big|e^{i\langle t,X(\omega)\rangle}\big|\,\mathrm{d}\mathbb{P}(\omega) = \int_\Omega 1\,\mathrm{d}\mathbb{P}(\omega) = 1.$$

The triangle inequality $|\int_\Omega f\,\mathrm{d}\mu| \le \int_\Omega |f|\,\mathrm{d}\mu$ is more complicated than one might think. Suppose φ is the angle from the polar coordinate representation of $\int_\Omega f\,\mathrm{d}\mu$. Then

$$\Big|\int_\Omega f\,\mathrm{d}\mu\Big| = e^{-i\varphi}\int_\Omega f\,\mathrm{d}\mu = \int_\Omega e^{-i\varphi} f\,\mathrm{d}\mu = \int_\Omega \mathcal{R}e(e^{-i\varphi}f)\,\mathrm{d}\mu \le \int_\Omega |e^{-i\varphi}f|\,\mathrm{d}\mu = \int_\Omega |f|\,\mathrm{d}\mu.$$

The first equality holds as multiplying with $e^{-i\varphi}$ is a rotation, the second is linearity, the third holds as $|\cdot| \in \mathbb{R}$ so that the imaginary part of the integral vanishes, and the fourth is monotonicity for real-valued integrals. Finally, the fifth equality is $|e^{-i\varphi}f| = |e^{-i\varphi}||f| = |f|$.

(ii) Write it down yourself!

(iii) Recall that $e^{-i\theta} = \cos(-\theta) + i\sin(-\theta) = \cos(\theta) - i\sin(\theta) = \overline{e^{i\theta}}$, hence,

$$\varphi_{-X}(t) = \mathbb{E}\big[e^{-i\langle X,t\rangle}\big] = \mathbb{E}\big[\overline{e^{i\langle X,t\rangle}}\big] = \overline{\mathbb{E}[e^{i\langle X,t\rangle}]} = \overline{\varphi_X(t)},$$

where we used

$$\overline{\int f \,\mathrm{d}\mu} = \overline{\int \mathcal{R}e(f)\,\mathrm{d}\mu + i\int \mathcal{I}m(f)\,\mathrm{d}\mu} = \int \mathcal{R}e(f)\,\mathrm{d}\mu - i\int \mathcal{I}m(f)\,\mathrm{d}\mu = \int \overline{f}\,\mathrm{d}\mu.$$

The claim now follows immediately.

(iv) The proof is essentially the same that we have seen for moment generating functions in Proposition 4.3.19. All we need to do is to extend Theorem 4.2.26 to complex-valued functions. But this is simple using the linearity from above:

Check that $\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\mathbb{E}[g(Y)]$ holds for independent random variables X and Y.

But then

$$\varphi_{X+Y}(t) = \mathbb{E}[e^{i\langle t,X+Y\rangle}] = \mathbb{E}[e^{i\langle t,X\rangle}]\cdot\mathbb{E}[e^{i\langle t,Y\rangle}] = \varphi_X(t)\cdot\varphi_Y(t),$$

for all $t\in\mathbb{R}$.

□ Lecture 16

To get a feeling let us check some examples.

Example 7.4.10. Using the discrete computation rule from (7.4) yields

$$\varphi_{\mathrm{Poi}(\lambda)}(t) = \mathbb{E}\big[e^{itX}\big] = \sum_{k=0}^{\infty} e^{-\lambda}e^{itk}\frac{\lambda^k}{k!} = e^{\lambda(e^{it}-1)}e^{\lambda(e^{it}-1)}, \quad t\in\mathbb{R},$$

and

$$\begin{aligned}\varphi_{\mathrm{Bin}(n,p)}(t) &= \mathbb{E}\big[e^{itX}\big]\\ &= \sum_{k=0}^{n}\binom{n}{k}e^{itk}p^k(1-p)^{n-k}\\ &= \sum_{k=0}^{n}\binom{n}{k}(pe^{it})^k(1-p)^{n-k} = \big(1-p+pe^{it}\big)^n, \quad t\in\mathbb{R}.\end{aligned}$$

The second example can also be computed from (iv) of Lemma 7.4.9 using

$$\varphi_{\mathrm{Ber}(p)}(t) = \mathbb{E}[e^{itX}] = e^{it}p + (1-p)e^{it0}, \quad t\in\mathbb{R},$$

and that a binomial random variable is a sum of n independent Bernoulli random variables.

Example 7.4.11. Using the computation rule from (7.4) yields

$$\varphi_{\mathcal{U}([0,a])}(t) = \frac{e^{iat}-1}{iat}, \quad t\in\mathbb{R}.$$

The computation is straight forward using the definition of the complex integral:

$$\begin{aligned}\mathbb{E}\big[e^{itX}\big] &= \int_0^a e^{itx}\frac{1}{a}\,\mathrm{d}x\\ &= \frac{1}{a}\int_0^a \cos(tx)\,\mathrm{d}x + \frac{1}{a}i\int_0^a \sin(tx)\,\mathrm{d}x\\ &= \frac{1}{at}\big([\sin(tx)]_0^a - i[\cos(tx)]_0^a\big)\\ &= \frac{1}{at}\big(\sin(at) - i\cos(ax) + i\big)\\ &\overset{i^2=-1}{=} \frac{1}{at}\frac{1}{i}\big(\cos(at) + i\sin(at) + i^2\big) = \frac{e^{iat}-1}{ait}\end{aligned}$$

Unfortunately, it is not always the case that the complex integrals can be computed that easily. In many instances countour integrals $\int_\gamma f(z)\,\mathrm{d}z$ need to used in order to compute $\int_{\mathbb{R}} f(x)\,\mathrm{d}x$. If you know about contour integration this is a good exercise, otherwise it is a good reason to learn about contour integration!

Example 7.4.12.

$$\varphi_{\mathcal{N}(\mu,\sigma^2)}(t) = e^{i\mu t} e^{-\frac{\sigma^2}{2}t^2}, \quad t \in \mathbb{R},$$

and

$$\varphi_{\mathrm{Exp}(\lambda)}(t) = \frac{\lambda}{\lambda - it}, \quad t \in \mathbb{R}.$$

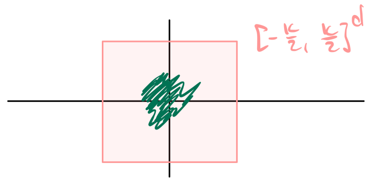
Apart from being nice to compute with there must be a deeper reason to study characteristic functions. The fundamental theorem on characteristic functions states that the complex exponential functions are separating for $\mathcal{M}_1(\mathbb{R}^d)$ or, formulated in the langueage of random variables, the law of a random variable is uniquely determined by the characteristic function. Indeed, this is the justification for the name, the law is characterised by the characteristic function.

Theorem 7.4.13. (i) Every measures $\mu \in \mathcal{M}_1(\mathbb{R}^d)$ is uniquely determined by all integrals

$$\int_{\mathbb{R}^d} e^{i\langle t,x\rangle}\,\mathrm{d}\mu(x), \quad t \in \mathbb{R}^d.$$

(ii) The law of a random vector X is determined by it's characteristic function, i.e. if $\varphi_X = \varphi_Y$ then $X \sim Y$.

Proof. Since $C_c(\mathbb{R}^d)$ is separating by Proposition 7.1.15 it suffices to prove that the complex exponentials are dense in $C_c(\mathbb{R}^d)$, then we can proceed similarly (but with an extra trick) to the proof of Corollary 7.4.4. The main idea of the proof is best understood in pictures. Let us fix two measures $\mu_1, \mu_2 \in \mathcal{M}_1(\mathbb{R}^d)$, $f \in C_c(\mathbb{R}^d)$ and suppose f that vanishes outside of a box $[-\frac{k}{2}, \frac{k}{2}]^d$. By enlarging k we can choose k such that $\mu_i\big(\mathbb{R}^d \setminus (k,k)^d\big) < \varepsilon$ for a fixed $\varepsilon > 0$ (continuity of



The support of f

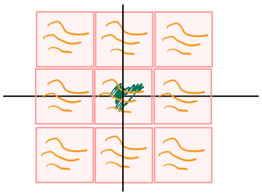
measures or Proposition 7.1.12). Next we define

$$g_m \colon \mathbb{R}^d \to \mathbb{C}, \quad g_m(x) = e^{i\langle \frac{\pi m}{k}, x\rangle}, \quad m \in x, \mathbb{Z}^d,$$

and C_k as the algebra of finite linear combinations of the g_m. Here is the important point of the functions g_m, they are periodic. For $n \in \mathbb{Z}^d$ they satisfy

$$g_m(x + 2kn) = e^{i\langle \frac{\pi m}{k}, x\rangle} \cdot \underbrace{e^{i\langle \frac{\pi m}{k}, 2kn\rangle}}_{=1} = g_m(x),$$

as $e^{2\pi l} = 1$ for all $l \in \mathbb{Z}$. Hence, the entire family C_k is periodic!
This is important to us as bounding $g \in C_k$ on $[-k,k]^d$ will automatically bound g everywhere. Restricting C_k to $[-k,k]^d$ the class is separating (this needs the $1/k$ in the scalar product as without it the functions would be periodic for a smaller box taking the same values on shifts of



Periodic function $g \in C_k$

the smaller boxes in the larger box). The Stone-Weierstraß approximation theorem now implies that there is $g \in C_k$ with

$$\sup_{x \in [-k,k]^d} |f(x) - g(x)| < \varepsilon.$$

Since g is close to f on $[-k,k]^d$ we can also use the periodicity to bound g on $\mathbb{R}^d$:

$$\sup_{x \in \mathbb{R}^d} |g(x)| = \sup_{x \in [-k,k]^d} |g(x)| \leq \sup_{x \in [-k,k]^d} |f(x) - g(x)| + \sup_{x \in [-k,k]^d} |f(x)| \leq \varepsilon + \|f\|_\infty .$$

Putting everything together, the above thoughts yields (please compare the proof of Corollar 7.4.4)

$$\begin{aligned}
&\Big|\int_{\mathbb{R}^d} f \,\mathrm{d}\mu_1 - \int_{\mathbb{R}^d} f \,\mathrm{d}\mu_2\Big| \\
&\leq \Big|\int_{\mathbb{R}^d} f \,\mathrm{d}\mu_1 - \int_{\mathbb{R}^d} g \,\mathrm{d}\mu_1\Big| + \underbrace{\Big|\int_{\mathbb{R}^d} g \,\mathrm{d}\mu_1 - \int_{\mathbb{R}^d} g \,\mathrm{d}\mu_2\Big|}_{=0 \text{ by assumption}} + \Big|\int_{\mathbb{R}^d} g \,\mathrm{d}\mu_2 - \int_{\mathbb{R}^d} f \,\mathrm{d}\mu_2\Big| \\
&\leq \int_{[k,k]^d} |f-g| \,\mathrm{d}\mu_1 + \int_{([k,k]^d)^c} |f-g| \,\mathrm{d}\mu_1 + \int_{[k,k]^d} |f-g| \,\mathrm{d}\mu_2 + \int_{([k,k]^d)^c} |f-g| \,\mathrm{d}\mu_2 \\
&\leq \mu_1\big([-k,k]^d\big) \cdot \varepsilon + \mu_1\big(([-k,k]^d)^c\big) \cdot (\varepsilon + \|f\|_\infty) + \mu_2\big([-k,k]^d\big) + \mu_2\big(([-k,k]^d)^c\big) \cdot (\varepsilon + \|f\|_\infty) \\
&\leq 2\varepsilon + 2\varepsilon\big(\varepsilon + \|f\|_\infty\big).
\end{aligned}$$

For the inequalities we used that μ_i are probability measures, $|f-g| \leq |f| + |g|$, and that μ_i only have mass at most ε outside of the chosen box. Since ε was arbitrary we proved $\int f \,\mathrm{d}\mu_1 = \int f \,\mathrm{d}\mu_2$. The second claim follows from the first claim applied to the laws $\mathbb{P}_X$. □

Of course one might ask why one might be interested in computing moments (for bounded random variables) or Laplace transformations (for non-negative random variables) if there is one general theorem that states that characteristic functions uniqueley determine all real random variables. This is because it is easier to compute moments or Laplace transformations these are the preferred tools in the special situations when they are sufficiently powerful.

So far we are used to describe random variables through their cummulative distribution function, density, or discrete probability weights. An alternative way is to describe random variables directly in terms of their chracteristic functions, which, according to the previous theorem, is as good as using the CDF.

The drawback of using characteristic functions is that one needs to compute expectations (integrals) for complex functions. At the same time this is the big advantage, too, as complex integration has wonderful tricks to offer.

In many examples the uniquneness is applied as follows. In order to identify the distribution of a random variable (such as a sum of other random variables) one tries to compute the characteristic function. If that characteristic function is already known we can identify the law. We already know this concept from Proposition 4.3.19 and the example below but always used the trick without a proof. Here is another little exercise to play with the examples, Lemma 7.4.9 and uniqueness theorem.

Check that

- If $Y \sim \mathcal{N}(0,1)$, then $Y := \sigma X + Y \sim \mathcal{N}(\mu, \sigma^2)$.
- If $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ are independent, then

$$X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

- If $X \sim \mathrm{Poi}(\lambda)$ and $Y \sim \mathrm{Poi}(\beta)$ are independent, then $X + Y \sim \mathrm{Poi}(\lambda + \beta)$.

7.5 Weak convergence on $\mathcal{B}(\mathbb{R}^d)$

We can finally return to the question of weak convergence. Recall from Propositon 7.3.4 that a sequence of probability measures (μ_n) converges weakly to μ (equivalently a sequence of random variables (X_n) converges in distribution to X) if and only if

(i) $\{\mu_n : n \in \mathbb{N}\}$ is tight,

(ii) $\lim_{n\to\infty} \int_E f \,\mathrm{d}\mu_n = \int_E f \,\mathrm{d}\mu$ for some separating family $C \subseteq C_b(E)$ of $\mathcal{M}_1(E)$.

We have identified separating families in the previous section. It now remains to identify situations in which the tightness trivially holds (bounded and non-negative) and give a handable criterion to check the tightness (convergence of characteristic functions). We keep the order of the previous section and first deal with the simple cases:

Theorem 7.5.1. (i) Suppose that $\mu, \mu_1, \ldots \in \mathcal{M}_1([a,b])$, then

$$\mu_n \xrightarrow{(\mathrm{w})} \mu,\ n \to \infty \quad \Leftrightarrow \quad \lim_{n\to\infty} \int_{[a,b]} x^m \,\mathrm{d}\mu_n(x) = \int_{[a,b]} x^m \,\mathrm{d}\mu(x), \quad \forall m \in \mathbb{N}_0.$$

(ii) Suppose that $X, X_1, \ldots$ are random variables with values in $[a,b]$, then

$$X_n \xrightarrow{(\mathrm{d})} X,\ n \to \infty \quad \Leftrightarrow \quad \lim_{n\to\infty} \mathbb{E}[X_n^m] = \mathbb{E}[X^m], \quad \forall m \in \mathbb{N}_0.$$

Proof. (ii) follows from (i) by the definition of convergence in distribution with the measures $\mu_n := \mathbb{P}_{X_n}$ using that $\mathbb{E}[X_n^m] = \int x^m \,\mathrm{d}\mathbb{P}_{X_n}(x)$.
(i) The situation is simple as $([a,b], |\cdot|)$ is compact. Hence, every sequence of measures is automatically tight (choose $K = [a,b]$ in the definition of tightness). According to Proposition 7.3.4 only convergence of integrals needs to be checked for a separating family. Since the family of monomials $C = \{x^m : m \in \mathbb{N}_0\}$ is separating according to Theorem 7.4.5 the proof is complete. $\square$

The situation is very similar for non-negative (or non-positive) sequences of random variables. In contrast to bounded sequences we now need convergence of the exponential moments (Laplace transformation):

Theorem 7.5.2. (i) Suppose that $\mu, \mu_1, \ldots \in \mathcal{M}_1([0,\infty))$, then

$$\mu_n \xrightarrow{(\mathrm{w})} \mu,\ n \to \infty \quad \Leftrightarrow \quad \lim_{n\to\infty} \int_{[0,\infty)} e^{-\lambda x} \,\mathrm{d}\mu_n(x) = \int_{[0,\infty)} e^{-\lambda x} \,\mathrm{d}\mu(x),\ \forall \lambda > 0.$$

(ii) Suppose that $X, X_1, \ldots$ are non-negative random variables, then

$$X_n \xrightarrow{(d)} X,\ n \to \infty \quad \Leftrightarrow \quad \lim_{n\to\infty} L_{X_n}(\lambda) = L_X(\lambda), \quad \forall \lambda > 0.$$

Proof. (ii) follows from (i) by the definition of convergence in distribution with the measures $\mu_n := \mathbb{P}_{X_n}$ using that $L_{X_n}(\lambda) = \int e^{-\lambda x}\, \mathrm{d}\mathbb{P}_{X_n}(x)$.
(i) We argue as in the proof of Theorem 7.4.6. Compactifying $[0, \infty]$ and extending the measures trivially to measures $\bar{\mu}_n$ yields a sequence of measures in a compact space. As in the previous proof the sequence is automatically tight (choose $K = [0, \infty]$). According to Proposition 7.3.4 only convergence of integrals needs to be checked for a separating family. Since the family of exponentials $C := \{f_\lambda \colon \lambda > 0\}$ is separating according to Theorem 7.4.6 the proof is complete. □

The most general theorem is Lévy's continuity theorem. Without any further assumption weak convergences is equivalent to convergence of the characteristic functions. Lecture 17

Theorem 7.5.3. (Lévy's continuity theorem) (i) Suppose that $\mu, \mu_1, \ldots \in \mathcal{M}_1(\mathbb{R}^d)$, then

$$\mu_n \xrightarrow{(w)} \mu,\ n \to \infty \quad \Leftrightarrow \quad \lim_{n\to\infty} \varphi_{\mu_n} = \varphi_\mu(t), \quad \forall t \in \mathbb{R}^d.$$

(ii) Suppose that $X, X_1, \ldots$ are $\mathbb{R}^d$-valued random variables, then

$$X_n \xrightarrow{(d)} X,\ n \to \infty \quad \Leftrightarrow \quad \lim_{n\to\infty} \varphi_X(t) = \varphi_X(t), \quad \forall t \in \mathbb{R}^d.$$

Proof. As earlier we only prove (i) as (ii) follows directly using $\mu_n = \mathbb{P}_{X_n}$.
"$\Rightarrow$": Since $g = e^{i\langle t, \cdot\rangle} = \cos(\langle t, \cdot\rangle) + i \sin(\langle t, \cdot\rangle)$ and $\cos, \sin \in C_b(\mathbb{R}^d)$ the pointwise convergence follows from the definition of weak convergence.
"$\Leftarrow$": We will actually prove a stronger statement than claimed, the so-called general continuity theorem:

Suppose $\lim_{n\to\infty} \varphi_{\mu_n}(t) = f(t)$, for all $t \in \mathbb{R}^d$, where $f \colon \mathbb{R}^d \to \mathbb{C}$ is continuous at 0. Then there is a probability measure Q on $\mathcal{B}(\mathbb{R}^d)$ with $f = \varphi_Q$ and $\mu_n \xrightarrow{(w)} Q, n \to \infty$.

Once we proved the general continuity theorem we immediately get the special continuity theorem by chosing $f = \varphi_\mu$ which is continuous at 0 by dominated convergence ($|e^{i\langle t, X\rangle}| = 1$).
Comparing with the proofs of the previous two theorems the strategy is as follows:

- deduce tightness of (μ_n),
- find a separating sequence for which the integrals converge,

because then the claim follows from Proposition 7.3.4. Chosing $C := \{e^{i\langle t, x\rangle} : t \in \mathbb{R}\}$ we already proved in Theorem 7.4.13 that C is separating. Hence, we need to prove that the assumed pointwise convergence of φ_{μ_n} implies tightness of (μ_n).
To do so let us first check that without loss of generality we may assume $d = 1$. If $\pi_k(x) = x_k$ denotes the projection on the kth coordinate, then we define $\mu_n^k := \mu_n \circ \pi_k$, the push-forward of the measures on the coordinates. Their characteristic functions of μ_n^k can be expresses through the characteristic functions of μ_n:

$$\varphi_{\mu_n^k}(t) = \int_{\mathbb{R}} e^{ixt}\, \mathrm{d}\mu_n^k(x) = \int_{\mathbb{R}} e^{i\langle x, te_k\rangle}\, \mathrm{d}\mu_n(x) = \varphi_{\mu_n}(te_k), \quad t \in \mathbb{R}.$$

Hence, the assumed pointwise convergence of φ_{μ_n} implies the convergence of all $\varphi_{\mu_n^k}$. If now we can prove that this implies tightness for all $(\mu_n^k)_{k\in\mathbb{N}}$, then, using Prohorov's characterisation of tightness through sequences, we obtain tightness of $(\mu_n)_{n\in\mathbb{N}}$. BECUASE.

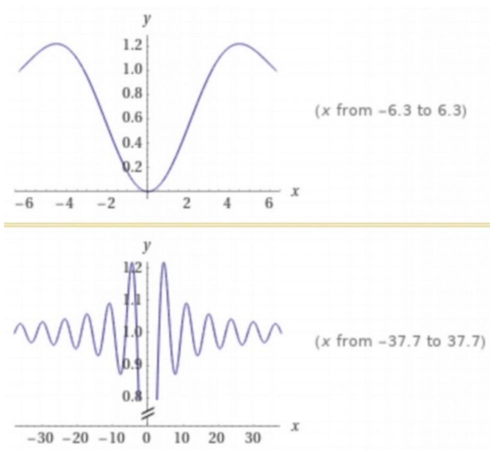
From now on we assume $d = 1$ and that $(\varphi_{\mu_n})_{n\in\mathbb{N}}$ converges pointwise to a function f that is continuous at 0.
First recall that

$$\varphi_\mu(t) = \int e^{itx}\,\mathrm{d}\mu(x) = \int \cos(tx)\,\mathrm{d}\mu(x) + i\int \sin(tx)\,\mathrm{d}\mu(x)$$

We want to relate tightness (probabilities) to integrals (characteristic functions). Usually we took closed sets A and approximated $\mathbf{1}_A$ with f_A^ε. Here we use a different trick by estimating suitiable indicators from above by the sine/cosine functions. Define $h\colon \mathbb{R}\to[0,\infty)$ as

$$h(x) = \begin{cases} 1 - \frac{\sin(x)}{x} & : x \neq 0 \\ 0 & : x = 0 \end{cases},$$

which is non-negative and continuous on $\mathbb{R}$ (Analysis):



Now define

$$\alpha := \inf\{h(x)\colon |x| \geq 1\} = 1 - \sin(1) > 0$$

so that $\frac{h(x)}{\alpha} \geq 1$ for $|x| > 1$. Then, for $k \geq 1$,

$$\begin{aligned}
\mu_n\big([-k,k]^c\big) &\overset{\text{mon.}}{\leq} \int_{[-k,k]^c} \frac{1}{\alpha} h\Big(\frac{x}{k}\Big)\,\mathrm{d}\mu_n(x)\\
&\leq \frac{1}{\alpha}\int_{\mathbb{R}} h\Big(\frac{x}{k}\Big)\,\mathrm{d}\mu_n(x)\\
&= \frac{1}{\alpha}\int_{\mathbb{R}} \underbrace{\bigg(\int_0^1 \Big(1 - \cos\Big(\frac{tx}{k}\Big)\Big)\,\mathrm{d}t\bigg)}_{=1-\sin(\frac{x}{k})\frac{k}{x}=h(\frac{x}{k})}\,\mathrm{d}\mu_n(x)\\
&\overset{\text{Fubini}}{=} \frac{1}{\alpha}\int_0^1\int_{\mathbb{R}} \Big(1 - \cos\Big(\frac{tx}{k}\Big)\Big)\,\mathrm{d}\mu_n(x)\,\mathrm{d}t\\
&= \frac{1}{\alpha}\int_0^1 \Big(1 - \mathcal{R}e\Big(\varphi_{\mu_n}\Big(\frac{t}{k}\Big)\Big)\Big)\,\mathrm{d}t.
\end{aligned}$$

Now we use dominated convergence to obtain

$$\begin{aligned}
\limsup_{n\to\infty} \mu_n\big([-k,k]^c\big) &\leq \frac{1}{\alpha} \lim_{n\to\infty} \int_0^1 \Big(1 - \mathcal{R}e\Big(\varphi_{\mu_n}\Big(\frac{t}{k}\Big)\Big)\Big)\,\mathrm{d}t \\
&\overset{\text{DCT}}{=} \frac{1}{\alpha} \int_0^1 \lim_{n\to\infty} \Big(1 - \mathcal{R}e\Big(\varphi_{\mu_n}\Big(\frac{t}{k}\Big)\Big)\Big)\,\mathrm{d}t \\
&= \frac{1}{\alpha} \int_0^1 \Big(1 - \mathcal{R}e\Big(f\Big(\frac{t}{k}\Big)\Big)\Big)
\end{aligned}$$

In the last step we have used that convergence of a sequence of complex numbers implies convergence of real- and imaginary-parts. Finally, we use the continuity of f (and thus $\mathcal{R}e(f)$) at 0. For all $\varepsilon > 0$ there is a $\delta > 0$ such that $|0 - \frac{t}{k}| < \delta$ implies $|1 - \mathcal{R}e(f(\frac{t}{k}))| = |\mathcal{R}e(f(0)) - \mathcal{R}e(f(\frac{t}{k}))| \leq \varepsilon$. Hence, there is some k so that $|1 - \mathcal{R}e(f(\frac{t}{k}))| < \varepsilon$ for all $t \in [0,1]$. But then there is some k so that $\limsup_{n\to\infty} \mu_n\big([-k,k]^c\big) < \varepsilon$.
From this the tightness follows as every single probability measure on $\mathbb{R}^d$ is tight. ARGUMENT HINSCHREIBEN □

We finish the section with a short discussion of the usefulness of the generalised version of the continuity theorem. Let us recall the different ways of characterising random variables:

$$\text{random variable } X \quad \Leftrightarrow \quad \text{law } \mathbb{P}_X \quad \Leftrightarrow \quad \text{CDF } F_X.$$

In fact, just as CDFs are functions with a set of properties one can ask if also characteristic functions are functions with a set of axiomatic properties so that it is reasonable to extend

$$\text{random variable } X \quad \Leftrightarrow \quad \text{law } \mathbb{P}_X \quad \Leftrightarrow \quad \text{CDF } F_X \quad \Leftrightarrow \quad \text{characteristic function } \varphi_X.$$

Indeed, this is the case but unfortunately the appearing property of positive definitness is hard to check:

Theorem 7.5.4. (Bochner)
A function $\varphi\colon \mathbb{R}^d \to \mathbb{C}$ is the characteristic function of a random vector (or a probability measure on $\mathcal{B}(\mathbb{R}^d)$) if and only if

- $\varphi(0) = 1$
- φ is continuous at 0
- φ is **positive semidefinite**, i.e.

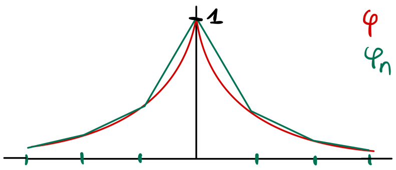
$$\sum_{k,l=1}^n y_k \bar{y}_l \varphi(t_k - t_l) \geq 0, \quad \forall n \in \mathbb{N},\ t_i \in \mathbb{R}^d,\ y_i \in \mathbb{C}.$$

Proof. "⇒": Suppose $\varphi(t) = \mathbb{E}[e^{itX}]$ for some random variable X. Then $\varphi(0) = 1$ is clear, continuity at 0 follows from dominated convergence, and

$$\begin{aligned}
\sum_{k,l=1} y_k \bar{y}_l \varphi(t_k - t_l) &= \sum_{k,l=1} y_k \bar{y}_l \int_{\mathbb{R}^d} e^{i\langle x, t_k - t_l\rangle}\,\mathrm{d}\mu(x) \\
&= \int_{\mathbb{R}^d} \sum_{k,l=1}^n y_k e^{i\langle x,t_k\rangle} \overline{y_l e^{i\langle x,t_l\rangle}}\,\mathrm{d}\mu(x) \\
&= \int_{\mathbb{R}^d} \Big|\sum_{k=1} y_k e^{i\langle x,t_k\rangle}\Big|^2\,\mathrm{d}\mu(x) \geq 0.
\end{aligned}$$

"⇐": Too hard □

There are many classes of characteristic functions that can be understood more directly using the general version of the continuity theorem. If for some reason we have a sequence of characteristic functions φ_n that converges pointwise to a function φ that is continuous at 0, then φ is a characteristic function. For all even continuous functions $\varphi\colon \mathbb{R} \to [0,1]$ with $\varphi(0) = 1$ that are convex on $[0,\infty)$ this can be done (Pólya's theorem). One can in fact construct simple explicit random variables which characteristic functions φ_n that are approximations of such such functions and converge pointwise to φ.



Example 7.5.5. For $\alpha \in (0,1)$ and $\lambda > 0$ the function $\varphi_\alpha(t) = e^{-\lambda|t|^\alpha}$ is even and convex on $[0,\infty)$. By Pólya's theorem this is a characteristic function of a random variable X. The random variable is called α-stable. In fact, the assumption $\alpha \in (0,1)$ is only needed to apply Pólya's theorem, actually φ_α is a characteristic function if and only if $\alpha \in [0,2]$. All such random variables are called α-stable and generalise the Gaussian distribution which appears for $\alpha = 2$.

Lecture 18

7.6 Applications

Recall from Theorem 4.1.19 the relation $\mathbb{E}[X^n] = \frac{\mathcal{M}_X^{(n)}(0)}{n!}$ between derivatives of the moment generating function $\mathcal{M}_X(t) = \mathbb{E}[e^{tX}]$ and the moments. The relation is useful as it allows to compute moments easily for a couple of examples (such as the Gaussian). In principle the idea is very simple, here for the first moment:

$$\mathcal{M}_X'(t) = \frac{d}{dt}\mathbb{E}[e^{tX}] = \mathbb{E}\Big[\frac{d}{dt}e^{tX}\Big] = \mathbb{E}[Xe^{tX}]$$

and then plugging-in $t = 0$. What makes the proof tricky is the interchange of differentiation and expectation for which dominated convergence needs to be applied in the right way and this forced us to assume finiteness of $\mathcal{M}_X$ in some interval $(-\varepsilon,\varepsilon)$. Since finiteness of $\mathcal{M}_X(t)$ means existence of exponential moments the theorem looks much better than it is, the assumption is extremely strong! We will now repeat the same story using the complex exponential. Before doing so we should first say a word on differentiation for functions $f : \mathbb{R} \to \mathbb{C}$. This is either defined by interpreting $\mathbb{C}$ as $\mathbb{R}^2$ and then the rewriting the derivative (a vector) in terms of polar coordinates or directly by taking limits in $\mathbb{C}$: $\frac{d}{dt}f(t) = f'(t) = \lim_{h\to 0}\frac{f(t+h)-f(t)}{h}$. Writing

$$\lim_{h\to 0}\frac{f(t+h)-f(t)}{h} = \lim_{h\to 0}\frac{\mathcal{R}ef(t+h)-\mathcal{R}ef(t)}{h} + i\lim_{h\to 0}\frac{\mathcal{I}mf(t+h)-\mathcal{I}mf(t)}{h}$$

shows that both approaches give exactly the same. As always it is important to keep in mind that we can freely ignore the formal definition and just compute in $\mathbb{C}$ as we are used to in $\mathbb{R}$ with the convention $i^2 = -1$. Most importantly, it holds that $\frac{d}{dt}e^{itx} = ixe^{itx}$. Higher order derivatives are definded recursively and denoted as usually by $f^{(n)}$. Now suppose, as for moment generating functions, the differentiation can be switched into the expectation, then we should get $\varphi_X^{(n)}(t) = \mathbb{E}\big[i^nX^ne^{itX}\big]$ so that plugging-in $t = 0$ gives again a moment formula $\mathbb{E}[X^n] = \frac{\varphi_X^{(n)}(0)}{i^n}$. Here again the magic of the complex exponential occurs. It is equally powerful in terms of what one can get from it but it is much more friendly because it is bounded. In essense the following theorem shows how to replace the real-exponential (moment generating function) by the complex-exponential (characteristic function) to obtain the same kind of results with minimal assumptions on the random variable.

Theorem 7.6.1. **(Moments of X and differentiability of φ_X)**
Let X be a real-valued random variable with characteristic function φ_X.

(i) If $\mathbb{E}[|X^n|] < \infty$, then φ_X is n-times continuously differentiable with

$$\varphi_X^{(n)}(t) = \mathbb{E}\big[e^{itX} i^n X^n\big], \quad t \in \mathbb{R}.$$

In particular, $\mathbb{E}[X^n] = \frac{\varphi_X^{(n)}(0)}{i^n}$.

(ii) If $\mathbb{E}[|X^n|] < \infty$ for some $n \in \mathbb{N}$, then φ_X satisfies the Taylor approximation

$$\varphi_X(t) = \sum_{k=0}^{n} \frac{i^k \mathbb{E}[X^k]}{k!} t^k + h_n(t)t^2, \quad t \in \mathbb{R}, \tag{7.6}$$

with a residual term satisfying $\lim\limits_{t\to 0} h_n(t) = 0$.

(iii) If $\lim\limits_{k\to\infty} \frac{t^k \cdot \mathbb{E}[|X^k|]}{k!} = 0$ for some $t \in \mathbb{R}$, then $\lim_{n\to\infty} h_n(t) = 0$. In particular, φ_X has the power series representation

$$\varphi_X(t) = \sum_{k=0}^{\infty} \frac{i^k \mathbb{E}[X^k]}{k!} t^k, \quad t \in \mathbb{R}.$$

The power series representation of φ_X is not surprsing at all. Writing the complex exponential as a power series and exchanging freely expectation and infinite sum this nothing but

$$\varphi_X(t) = \mathbb{E}\Big[\sum_{k=0}^{\infty} \frac{i^k t^k X^k}{k!}\Big] = \sum_{k=0}^{\infty} \frac{i^k \mathbb{E}[X^k]}{k!} t^k.$$

Of course, the interchange is non-trivial and there are essentially two ways to go. Either, arguing as in the proof of Theorem 4.1.19 one takes the limit of the partial sums, justifies dominated convergence, and then gets the moment formual by differentiating the power series, or, as we argue below, one first identifies the derivatives and then refers to Taylor's theorem to derive the power series.

Proof. In order to swich differentiation and expectation we will use the differential quotient and use dominated convergence to justify the change of expectation and limit in ε. We can do this since the differences of complex exponentials have useful bounds. Let us first check the basic estimate $|e^{itx} - e^{isx}| \leq |t-s| \cdot |x|$. If $s < t$ and $x > 0$, then expanding the exponential and using the Euler formula yields

$$\begin{aligned}
|e^{itx} - e^{isx}| &= |e^{is\frac{x}{2}}| \cdot |e^{it\frac{x}{2}}| \cdot |e^{i(t-s)\frac{x}{2}} - e^{-i(t-s)\frac{x}{2}}| \\
&= 2|\sin((t-s)x/2)| \\
&= 2\Big|\int_0^{(t-s)\frac{x}{2}} \cos(u)\,\mathrm{d}u\Big| \overset{|\cos|\leq 1}{\leq} |(t-s)x|
\end{aligned}$$

and the other cases are treated similarly.
(i) The first derivative is now simple:

$$\lim_{h\to 0} \frac{\varphi_X(t) - \varphi_X(t+h)}{h} = \lim_{h\to 0} \mathbb{E}\Big[\frac{e^{itx} - e^{i(t+h)x}}{h}\Big] \overset{\text{DCT}}{=} \mathbb{E}\Big[\lim_{h\to 0} \frac{e^{itx} - e^{i(t+h)x}}{h}\Big] = \mathbb{E}\big[iXe^{itX}\big]$$

Changing limits and expectation was justified by the integrable (assumption) upper bound

$$\Big|\frac{e^{itx} - e^{i(t+h)x}}{h}\Big| \leq \frac{|h \cdot x|}{|h|} = |x|$$

which was justified above. Hence, $\varphi_X'(t) = \mathbb{E}\big[e^{itX} iX\big]$ exists. Inductively we proceed in exactly the same way to differentiate $\varphi_X^{(k)}(t)$:

$$\begin{aligned}
\lim_{h\to 0} \frac{\varphi_X^{(k)}(t) - \varphi_X^{(k)}(t+h)}{h} &= \lim_{h\to 0} \frac{\mathbb{E}\big[e^{itX} i^k X^k\big] - \mathbb{E}\big[e^{i(t+h)X} i^k X^k\big]}{h} \\
&= \lim_{h\to 0} \mathbb{E}\bigg[\frac{\big(e^{itX - e^{i(t+h)X}}\big) i^k X^k}{h}\bigg] \\
&\overset{\text{DCT}}{=} \mathbb{E}\bigg[\lim_{h\to 0} \frac{\big(e^{itX - e^{i(t+h)X}}\big) i^k X^k}{h}\bigg] = \mathbb{E}\big[i^{k+1} X^{k+1} e^{itX}\big]
\end{aligned}$$

The interchange of limit and expectation is again justified by the upper bound

$$\frac{|(e^{itx} - e^{i(t+h)x}) i^k x^k|}{h} \leq |x| \cdot |x^k| = |x^{k+1}|.$$

The proof shows very clearly that φ_X can be differentiated as long as there are enough finite moments of X. But this is no additional assumption as otherwise the formula does not make sense anyways!

(ii) Since all derivatives at 0 are known from (i) the complex version of Taylor's theorem for $x_0 = 0$ gives

$$\varphi_X(t) = 1 + it\mathbb{E}[X] - \frac{1}{2}t^2\mathbb{E}[X^2] + \ldots + \frac{i^n t^n \mathbb{E}[X^n]}{n!} + R_n(t)t^n,$$

with a remainder term satisfying $\lim_{t\to 0} h_n(t) = 0$.

(iii) We need to show that, for fixed t, the residual $h_n(t)$ vanishes as n tends to infinity. It is most practical to use the integral representation for $R_n(t) := h_n(t)t^n$:

$$|R_n(t)| = \bigg|\int_0^t \frac{\varphi_X^{(n+1)}(s)}{(n+1)!}(t-s)^n \,\mathrm{d}s\bigg| \leq \int_0^t \frac{\mathbb{E}[|X^{n+1}|]}{(n+1)!}(t-s)^n \,\mathrm{d}s = \frac{t^{n+1}}{(n+1)} \frac{\mathbb{E}[|X^{n+1}|]}{(n+1)!}$$

Here we used the formula from (i) and that $|e^{itX} i^n| = 1$. But then $h_n(t) \to 0$ for $n \to \infty$. $\square$

Just as we used the moment generating functions to compute moments for random variables with exponential moments we can also use the characteristic functions:

Compute the first few moments of $\mathcal{N}(0,1)$, $\mathrm{Poi}(\lambda)$, and $\mathcal{U}([0,1])$.

As an application we prove what is sometimes called the **method of moments** and, as a special case, we can finally give a proof of Theorem 4.3.17.

Corollary 7.6.2. (i) Let X be a real-valued random variable such that there is a constant C with $\frac{1}{n}\big(\mathbb{E}[|X^n|]\big)^{\frac{1}{n}} < C$ for all $n \in \mathbb{N}$. Then the law of X is uniquely determined by all it's moments.

(ii) In particular, if $\mathcal{M}_X(t) < \infty$ for $t \in (-\varepsilon, \varepsilon)$ for some $\varepsilon > 0$, then $\mathcal{M}_X$ uniquely determines the law of X.

The corollary is a significant extension of Theorem 7.4.5. If $|X|$ is bounded by some C, then $\mathbb{E}[|X^n|]$ is bounded by C^n so that the assumption of the corollary is clearly satisfied. The corollary states that also for other (rather special) random variables which moments increase slowly enough the same statement holds.

Proof. (i) For $|t| < \frac{1}{3C}$ we have

$$\limsup_{n\to\infty} \frac{\mathbb{E}[|X^n|]\cdot |t|^n}{n!} \overset{\text{Sterling}}{=} \limsup_{n\to\infty} \Big(\mathbb{E}[|X^n|]^{\frac{1}{n}} |t| \frac{e}{n}\Big)^n \sqrt{2\pi n} \leq \limsup_{n\to\infty} \Big(\frac{e}{3}\Big)^n \sqrt{2\pi n} = 0$$

Hence, for all $t \in (-\frac{1}{3C}, \frac{1}{3C})$ we can use Theorem 7.6.1 (iii) to express φ_X as a power series. Since the coefficients are the moments, φ_X is determined on $(-\frac{1}{3C}, \frac{1}{3C})$ through the moments. Since a power series is uniquely determined by the values on some interval φ_X is also uniquely determined on $\mathbb{R}$ by all the moments. Finally, since the law of X is uniquely determined by φ_X according to Theorem 7.4.13 the proof is complete.

(ii) We argue as in the proof of Theorem 4.1.19:

$$\sum_{k=0}^{\infty} \frac{t^k \mathbb{E}[|X|^k]}{k!} \overset{\text{MCT}}{=} \mathbb{E}[e^{t|X|}] \leq \mathbb{E}[e^{tX}] + \mathbb{E}[e^{-tX}] = \mathcal{M}_X(t) + \mathcal{M}_X(-t) < \infty, \quad t \in (-\varepsilon, \varepsilon).$$

Now we can use the proof of (i) to finish the proof. $\square$

Theorem 7.6.3. (Central Limit Theorem)
Let $X_1, X_2, ...$ be iid with $\mu := \mathbb{E}[X_1]$ and $\sigma^2 := \mathbb{V}[X_1] < \infty$. Then

$$\frac{\sum_{k=1}^n X_k - n\mu}{\sqrt{\sigma^2 n}} \overset{(\mathrm{d})}{\longrightarrow} \mathcal{N}(0,1), \quad n \to \infty.$$

Proof. Without loss of generality we may assume $\mu = 0$ and $\sigma^2 = 1$ as otherwise we can consider the standardised random variables $\tilde{X} := \frac{X-\mu}{\sigma}$. Writing $S_n = \sum_{k=1}^n X_k$, Lévy's continuity theorem we only need to prove that

$$\varphi_{\frac{S_n}{\sqrt{n}}}(t) \longrightarrow e^{-\frac{1}{2}t^2} = \varphi_{\mathcal{N}(0,1)}(t), \quad \forall t \in \mathbb{R}.$$

Using Lemma 7.4.9 and the iid assumption shows that we only need to prove

$$\varphi_{\frac{S_n}{\sqrt{n}}}(t) = \Big(\varphi_{X_1}\Big(\frac{t}{\sqrt{n}}\Big)\Big)^n \longrightarrow e^{-\frac{1}{2}t^2}, \quad n \to \infty.$$

The trick is to replace the exponential by the first two summands of it's Taylor expansion (7.6) and then to use

$$\lim_{n\to\infty} \Big(1 - \frac{t^2}{2n}\Big)^n = e^{-\frac{1}{2}t^2}, \quad t \in \mathbb{R}.$$

Do do this rigorously first check by a quick induction that

$$|u^n - v^n| \leq |u - v| \cdot n \cdot \max(|u|, |v|)^{n-1}, \quad \forall u, v \in \mathbb{C}.$$

Using that $|\varphi_{X_1}|, |1 - \frac{t^2}{2n}| \leq 1$ for n large enough then yields

$$\begin{aligned} \left| \Big(1 - \frac{t^2}{2n}\Big)^n - \Big(\varphi_{X_1}\big(\frac{t}{\sqrt{n}}\big)\Big)^n \right| &\leq \left| \Big(1 - \frac{t^2}{2n}\Big)^n - \Big(1 - \frac{1}{2}\frac{t^2}{n} + h_2\Big(\frac{t}{\sqrt{n}}\Big)\frac{t^2}{n}\Big)^n \right| \\ &\leq n \cdot \left| h_2\Big(\frac{t}{\sqrt{n}}\Big)\frac{t^2}{n} \right| \\ &= t^2 \cdot \left| h_2\Big(\frac{t}{\sqrt{n}}\Big) \right| \to 0, \quad n \to \infty. \end{aligned}$$

$\square$

In many textbooks one can see the formulation

$$\lim_{n\to\infty} \mathbb{P}\Big(\frac{\sum_{k=0}^n X_k - n\cdot\mu}{\sqrt{n\sigma^2}} \in [a, b]\Big) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2}}\, \mathrm{d}x, \quad \forall a < b,$$

which holds due to Portemanteau because $\mathbb{P}_{\mathcal{N}(0,1)}(\partial[a, b]) = 0$.

Kapitel 8

Brownian Motion